Gene Order Evolution and Genomic Analyses of the Model Eukaryote,

Saccharomyces cerevisiae, and Other Yeast Species.

Cathal Seoighe

Ph.D. Thesis

University of Dublin

1999.

Declaration

This thesis is submitted by the undersigned for the degree of Doctor in Philosophy at the University of Dublin. It has not been submitted as an exercise for a degree at any other university.

Apart from the advice, assistance and joint effort mentioned in the acknowledgements and in the text, this thesis is entirely my own work.
I agree that the library may lend or copy this thesis freely on request.

Cathal Seoighe

Cathal Seoighe

October 1999

Acknowledgements

To my parents, Michael and Charlotte, and to my brother, David.

I owe a debt of gratitude to Ken Wolfe, for the huge amount of work that has gone into the supervision of this thesis and for his continued faith in me and encouragement throughout. This thesis would have been much harder for me, if not impossible, without the help and encouragement of the other people who have been in the group. Luce Skrabanek has been great help since the beginning and invaluable for her ability to see to the heart of even my most obscurely written C-code, and fix it. To Amanda Lohan for her pedagogical approach to knocking some biology into me, and for knocking me over the head whenever I thought to rise above my station, many thanks. Colin Semple, the father figure, who never tired of explaining basics. Thanks to Aoife for much more personable answers to all queries than you could possibly hope to find for yourself, and to Karsten for patience and help whenever called upon. Andrew Lloyd for countless suggestions and anecdotes and a door that was always open, even if he wished it hadn't been. Thanks to Denis Shields, for discussions on the maths in Chapter 2 and to Mike McLean for help with the programmes that were used to produce the graphic in Chapter 3 and Dan Bradley for comments and encouragement. Thanks also to Keith Bradnam for the collaboration that produced the results in Chapter 4.

I would like to acknowledge the contribution of family and friends to this thesis, by their encouragement and understanding. Thanks to my parents, brothers and sisters and to Alison, Niamh, Kieran, Michael, Treasa and Marie and to Richard and Suzanne and to all the people I have met over the last three years, both because of, and in spite of, this work.

Contents

Summary

The complete genome sequence of the yeast *Saccharomyces cerevisiae* can be used to form a clearer picture of the evolution of the diverse species of yeast and to construct a model of eukaryote genome evolution. In this study changes in the order of genes on yeast chromosomes were assessed and quantified by comparison with available sequence data from other yeast species and by the adaptation of comparative gene order techniques to the case of intraspecific comparative mapping within the *S. cerevisiae* genome. This is suggested by the hypothesis of genome duplication in the lineage leading to *S. cerevisiae*. Methods were developed to optimise the detection and display of cluster homology regions in an ancient tetraploid. A fresh look was taken at outstanding questions of genome organisation that have remained unresolved since the beginning of the yeast sequencing project. Compositional variation in yeast was shown to be explainable by short range correlation in base composition and no evidence was found for periodic variation in base composition over long distances. The available contiguous sequences from the pathogenic yeast *Candida albicans*, spanning most of the genome, were compared to *S. cerevisiae* and a large number of small inversions of gene order were found. The excess of small rearrangements of gene order has implications for comparative mapping and for the quantitative methods that have been applied to gene order comparison.

Abbreviations

bp	base pair
DSB	double strand break
EMBL	European molecular biology laboratory
GC3s	G+C content at third coding position silent sites
kb	kilobase
Mb	megabase
MIPS	Munich information centre for protein sequences
MYA	million years ago
Myr	million years
ORF	open reading frame
PFGE	pulsed field gel electrophoresis
RFPL	restriction fragment length polymorphism
YPD	yeast proteome database

Chapter 1

Introduction

1.1 Genomes of the ascomycetous yeasts

The term *yeast* is without taxonomic standing and is used to describe species of the fungal families ascomycetes, basidiomycetes and the imperfect fungi that tend to be unicellular for the greater part of their life cycle and reproduce by budding (Online Medical Dictionary; Alexopoulos *et al.*, 1996). Yeasts from the phylum Ascomycota, sometimes known as the "true" yeasts, have been divided into three classes that have been confirmed by molecular data: Archiascomycetes, Euascomycetes and Hemiascomycetes (Kurtzman, 1998a). All of the species studied in detail in this thesis belong to the yeasts of the *Saccharomycetales* family of the class Hemiascomycetes. Several species of *Saccharomycetales* are of economic importance and have been of significant benefit to human society and quality of life since the beginning of civilisation. The first documented evidence of the use of yeast dates to 6000 BC (Alexopoulos *et al.*, 1996). Today industrial use of yeasts involves millions of tons each year, making it the most important micro-organism used by humans (Lyons *et al.*, 1993) with applications ranging from brewing and baking to food and food supplement production, animal fodder production, tools for biotechnology, waste recycling and energy production in the form of fuel-ethanol (Hinchliffe and Kenny, 1993; Ingledew, 1993). There are also animal and plant pathogens among the *Saccharomycetales* species, including *Candida albicans* that exists in the gastrointestinal and urogenital tracts of many animal species, and *Ashbya gossypii,* a pathogen of cotton (Saltarelli, 1989). The *Saccharomycetales* yeasts include some species, such as *Saccharomyces cerevisiae*, that exist primarily in the yeast, or single-celled, form and others that are pleomorphic, having a filamentous state as well as yeast-like growth (e.g. *Candida albicans*). An intermediate state consisting of individual cells existing as pseudohyphae can also be found in several species (including *S. cerevisiae*; Gimeno *et al.*, 1992).

1.1.1 Sexual reproduction and morphology

There are both teleomorphic (sexual) and anamorphic (asexual) *Saccharomycetales* species. Sexual reproduction has been studied in detail in *Saccharomyces cerevisiae* and involves the fusion of two phenotypically distinct haploid cell types, **a** and α, to form an **a**/α diploid. The diploid cell may continue to reproduce through mitosis but if deprived of nutrients will undergo sporulation and give rise again to haploid **a** and α cells (Sprague, 1995). After a period of asexual growth homothallic *S. cerevisiae* colonies derived from single **a** or α cells are found to contain both the **a** and the α cell types due to the occurrence of mating-type switching. The switching of mating-type requires a very efficient programmed genetic rearrangement that involves the incorporation of information from the silent mating-type cassettes located at *HMR* and *HML* (on the right and left ends of chromosome III respectively) into the mating-type locus (Herskowitz *et al.*, 1992). In *Candida albicans*, a diploid anamorphic hemiascomycete, a region homologous to the MAT locus of *Saccharomyces cerevisiae* has recently been identified (Hull and Johnson, 1999). *C. albicans* is heterozygous at this locus, suggesting that it derived from a sexual ancestor that may have had a similar sexual cycle to *S. cerevisiae.* A system of mating and mating-type switching similar to the one outlined above for *S. cerevisiae* is also found in the distantly related species *Schizosaccharomyces pombe* (Sprague, 1995) indicating that it may have been an early development in yeast evolution. However, the alternative view, that similar mating-type switching strategies in *S. cerevisiae* and *Schiz. pombe* are a result of convergent evolution, is suggested by the fact that different genes are involved in mating-type switching in the two organisms (Haber, 1998).

Diploid **a**/α cells of *S. cerevisiae* may undergo a dimorphic transition in response to starvation. The cells become elongated and cell division is unipolar so that the cells grow in long, thin formations called pseudohyphae. The pseudohyphae grow away from the colony, allowing the cells to forage for scarce food supplies (Gimeno *et al.*, 1992). The formation of pseudohyphae is restricted to diploid cells in *S. cerevisiae*.

1.1.2 The genome of *Saccharomyces cerevisiae*

The task of sequencing the entire genome of the yeast *Saccharomyces cerevisiae* had been completed by 1996 (Dujon, 1996; Goffeau *et al.*, 1997), the starting date of the present work. As the first eukaryotic genome to be sequenced it represented a major advance and remains an invaluable resource, while continuing to pose significant challenges for the research community. The *S. cerevisiae* genome is 12 Mb in length (excluding large tandem repeats) and contains approximately 6000 protein coding genes (Dujon, 1996) in 16 chromosomes ranging from 0.2 – 1.5 Mb in length. It is a highly compact genome with almost 70% of the sequence in open reading frames and just 4% of genes containing introns, compared to about one in three genes which are intron-containing in *Schizosaccharomyces pombe* (Chen and Zhang, 1998). There are 274 intact nuclear tRNA genes in the yeast genome that can be grouped into 42 isoacceptor families (Hani and Feldmann, 1998) and there are 40 genes encoding small nuclear RNAs (snRNAs, Goffeau *et al.*, 1996). 51 ‘canonical’ retrotransposons of at least four different types (Ty1 – Ty4) have been found in the yeast genome, often at the 5` end of tRNA genes (Hani and Feldmann, 1998). Yeast retrotransposons contain 5.3 – 5.7 kb of internal DNA surrounded by approximately 350 bp long terminal repeats (Kaback, 1995). Complete sequence information from *S. cerevisiae* has greatly aided the understanding of the role played by retrotransposons in genome structure and evolution (Kaback, 1995).

Compositional variation in *S. cerevisiae*

The sequence of yeast chromosome III, the first eukaryotic chromosome to be completely sequenced, revealed non-uniformity of ORF G+C content along the chromosome (Sharp and Lloyd, 1993). High G+C 'peaks' appeared to occur periodically along chromosomes. This periodicity in base composition was observed for several subsequent chromosomes and was reported to be correlated to periodic variations in coding density for chromosome XI (Dujon *et al.*, 1994). Correlation between the positions of G+C peaks and gene dense regions appeared to link the base composition variation in yeast even more closely to the gene-dense heavy isochores that had been observed in warm blooded vertebrates (Bernardi *et al.*, 1985). The periodic variation of G+C content could only be observed with complete chromosome sequences and, as a result, it received considerable attention with each chromosome sequence report. Several chromosomes appeared to show periodicity while others lacked any G+C content structure. However by the end of the sequencing project questions concerning the existence of base composition variations and their relationship to the isochores of mammalian genomes were left unresolved (see Chapter 4).

Cluster Homology Regions

One of the most remarkable features of the yeast genome was the number and distribution of large, homologous regions of the genome, known as cluster homology regions (Coissac *et al.*, 1997; Mewes *et al.*, 1997; Wolfe and Shields, 1997). These regions normally occur in exactly two copies and led Wolfe and Shields to the hypothesis that *S. cerevisiae* is a degenerate tetraploid that underwent genome duplication approximately 10^8 years ago (Wolfe and Shields, 1997). As evidence to support their conclusions Wolfe and Shields pointed out (i) that the transcriptional orientation of duplicated gene pairs in yeast is almost always the same, either towards or away from the centromere; (ii) that the large 'sister' duplicated sections do not overlap with one another; and (iii) that gene order in the related species *Kluyveromyces lactis* (see below) is the same as what would be expected for a species that diverged from *S. cerevisiae* before genome duplication occurred in the *S. cerevisiae* lineage. These observations are not compatible with the alternative hypothesis of multiple independent duplications of sections of chromosomes (Wolfe and Shields, 1997; Keogh *et al.*, 1998). Fig. 1.1 illustrates the model, originally proposed by Wolfe and Shields, to explain how genome duplication followed by extensive gene deletion and reciprocal translocation gave rise to the cluster homology regions observable in the yeast genome. The hypothesis is also strongly supported by the recent report of the physical mapping of chromosome I of the unduplicated ascomycete *Ashbya gossypii* (Dietrich *et al.*, 1999).

Complete or partial duplication of the genome may have had a profound impact on many organisms' evolution (Ohno, 1970). In addition to altering the karyotype and increasing the number of genes, duplication brings about a reorganisation of local gene order through differential gene loss (Lundin, 1993) and may also increase the likelihood that large-scale chromosomal rearrangements will be fixed (Ahn and Tanksley, 1993; Ryu *et al.*, 1998).

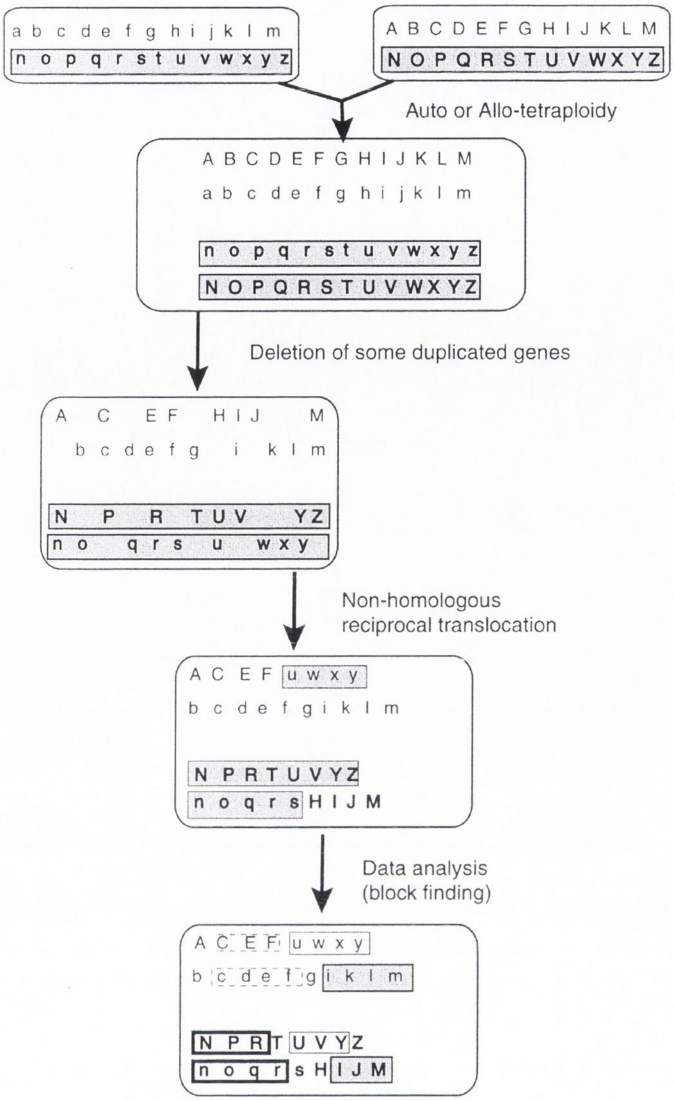
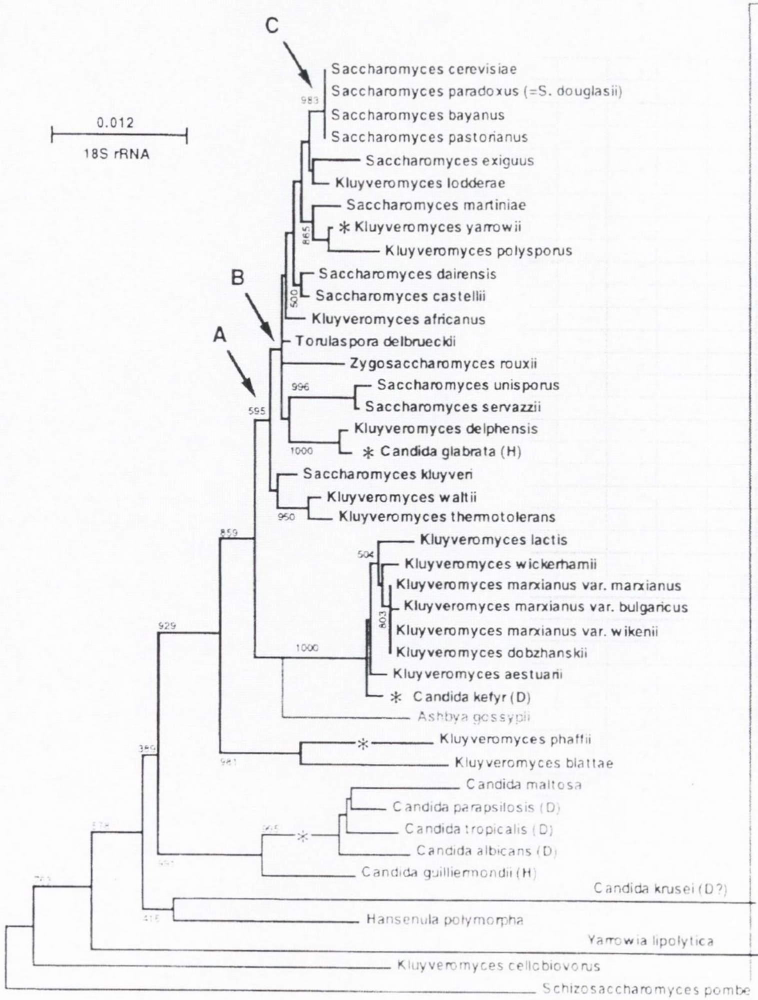


Figure 1.1 Schematic illustration of the model proposed by Wolfe and Shields to explain cluster homology regions in the yeast genome (see Keogh *et al.*, 1998). There are two chromosomes in the unduplicated genome, with a total of 26 genes. Tetraploidy followed by gene deletion results in a genome of four chromosomes and 35 genes. Reciprocal translocation breaks up the duplicated chromosomes and gives rise to disperse, partially duplicated regions which are the clustered homology regions.

1.1.3 Species in the genus *Saccharomyces*

There are several well-characterised yeast species closely related to *Saccharomyces cerevisiae* (see Fig. 1.2), providing information about species diversity and the more recent evolution of the *S. cerevisiae* genome. The genus *Saccharomyces* has been subdivided into two groups, *Saccharomyces sensu stricto* and *Saccharomyces sensu lato*. The *sensu stricto* group contains four closely related species: *S. cerevisiae*, *S. paradoxus*, *S. bayanus* (*S. uvarum*) and *S. pastorianus* (*S. carlsbergensis*), which is a hybrid of *S. cerevisiae* and *S. bayanus*, or an *S. bayanus* like species such as *S. monacencis* (Kielland-Brandt *et al.*, 1995). The species *S. castellii, S. dairenensis, S. servazzii, S. unisporus* and *S. kluyveri* are among the species that have often been included in the *sensu lato* group, although recent opinion would probably place *S. kluyveri* in a third *Saccharomyces* group if it is to remain within the genus *Saccharomyces* (Piskur *et al.*, 1998).

S. cerevisiae has 16 chromosomes. Karyotyping of *S. castellii and S. dairenensis* suggests that they contain 9 variably sized chromosomes (Petersen *et al.*, in press-b). Other *Saccharomyces* species have around 12 chromosomes except *S. kluyveri*, which has 5-7 chromosomes (Vaughan-Martini and Martini, 1998; Petersen *et al.*, in press-b), indicating that it is likely to be unduplicated. The genome duplication event was estimated to have occurred somewhere between points A and C on Fig. 1.2 (Keogh *et al.*, 1998). If *S. kluyveri* is unduplicated this can be narrowed to the region between points B and C. The karyotype of *Candida glabrata* (Fig. 1.2) suggests that it is also an ancient tetraploid. If it shares the same genome duplication as *S. cerevisiae* the duplication must have occurred at point B in Fig. 1.2 if the tree is correct. *S. castellii* and *S. dairenensis* contain the lowest number of chromosomes of the remaining *sensu stricto* species. They also cluster together according to phylogenetic analysis based on both 18S and 26S rDNA (Keogh *et al.*, 1998; Kurtzman and Robnett, 1998). If it is confirmed that these two species have as few as nine chromosomes (as suggested by Petersen *et al.*, in press-b) it would be reasonable to propose that the genome duplication occurred after the divergence of this pair of species from the rest of the *sensu lato* group and that *C. glabrata* was duplicated separately, assuming again that the topology of the tree in Fig. 1.2 is correct. Several *Saccharomyces* species are among the thirteen ascomycetes that are currently the subject of a French-based sequencing project (Genoscope web page). This project should greatly improve the understanding of *Saccharomyces* genome evolution. Unfortunately *S. dairenensis* and *S. castellii* are not among the species to be sequenced. Relatively fast evolving features, such as mitochondrial gene order evolution are also useful for resolving the relationships within closely related groups, and the mitochondrial genomes of *Saccharomyces* species have been studied extensively (as discussed below).



	Linkage conservation	PFGE estimates of number of chromosomes							PFGE estimates of genome size (Mb)				Type strain
	%	Naumov	Sor	Vaughan-Martini	Philippsen	Cardinali	Doi	Other	Vaughan-Martini	Sor	Doi	Other	CBS
Saccharomyces cerevisiae	100	16	13	12	12	13			11.085				1171
Saccharomyces paradoxus (=S. douglasii)	(100)	16		15		13			13.105				432
Saccharomyces bayanus	(100)	16		17	15	17			14.630				380
Saccharomyces pastorianus	100	16		14		13			11.750				1538
Saccharomyces exiguus				11	11				10.800				379
Kluyveromyces lodderae			13							15.5			2757
Saccharomyces martiniae		~10		10									6334
Kluyveromyces yarrowii			15							15.0			8242
Kluyveromyces polysporus			9							11.2			2163
Saccharomyces dairensis		7		6	8			7	7.920				421
Saccharomyces castellii				8				8-12	9.415				4309
Kluyveromyces africanus			14							17.0			2517
Torulaspora delbrueckii								6				13.0	1146
Zygosaccharomyces rouxii	(100)							~8					732
Saccharomyces unisporus		~16		11	10				11.375				398
Saccharomyces servazzii		~16		9	10				10.430				4311
Kluyveromyces delphensis			9							13.2			2170
Candida glabrata (H)	(100)						14	14			14.4	13.5	138
Saccharomyces kluyveri	(75)			7	5			7	9.550				3082
Kluyveromyces waltii			5							7.2			6430
Kluyveromyces thermotolerans			7							10.6			6340
Kluyveromyces lactis	74		6					6		12.0		12.3	683
Kluyveromyces wickerhamii			8							13.6		12.5	2745
Kluyveromyces marxianus var. marxianus	(83)		6-12							8.2-18.0			712
Kluyveromyces marxianus var. bulgaricus			10							13.2			2762
Kluyveromyces marxianus var. wikenii			7							10.0			5671
Kluyveromyces dobzhanskii			6							8.8			2104
Kluyveromyces aestuarii			7							10.2			4438
Candida kefyr (D)			12-14				17-19			21.9-27.2	27.4-31.1		607
Ashbya gossypii	(100)				7							9	
Kluyveromyces phaffii			11							10.6			4417
Kluyveromyces blattae			4-5							5.5-7.0			6284
Candida maltosa	(33)												5611
Candida parapsilosis (D)	(0)						14				25.8-26.4		604
Candida tropicalis (D)	(0)						12				29.5-31.4		2310
Candida albicans (D)	13						16	16			29.3-32.6	32.5	562
Candida guilliermondii (H)	(0)						8				11.5-11.6		566
Candida krusei (D?)							7-9				19.3-20.2		573
Hansenula polymorpha	18												4732
Yarrowia lipolytica	(20)							5				17.7	6124
Kluyveromyces cellobiovorus			6							10.6			7153
Schizosaccharomyces pombe	0							3				14	

Figure 1.2 Phylogenetic tree of some ascomycete species drawn from 18S RNA (taken from Keogh *et al.*, 1998). The tree was produced using the neighbour-joining method from a ClustalW alignment. Bootstrap values (1000 replicates) are shown if greater than 500. The

percentage of adjacent genes that remain adjacent in both species is given in the linkage conservation panel. Values from fewer than ten adjacent pairs are in parentheses. Pulsed field gel electrophoresis (PFGE) estimates are given for the number of chromosomes and the total genome size.

1.1.4 Evolution of the *Saccharomyces* mitochondrial genome

Yeast has played an important role in the study of mitochondrial inheritance (Piskur, 1993). Although the set of genes encoded in the mitochondria is the same in almost all species the order of these genes and the quantity of intergenic DNA is evolving rapidly, such that even closely related species like the *Saccharomyces* genus, contain a great deal of heterogeneity in quantity and organisation of mtDNA (Cardazzo *et al.*, 1998). Piskur *et al.* (1998) have carried out an analysis of mitochondrial genomes of several species of the genus *Saccharomyces*. They studied spontaneous and chemically induced petite mutants (mitochondrial mutations causing respiratory deficiency), as well as mitochondrial genome organisation and described three groups of species within the *sensu stricto* and *sensu lato* groups. *S. kluyveri* has a highly rearranged mitochondrial genome with respect to other *Saccharomyces* species and is the only species in the genus that is completely petite negative. It forms a separate group in Piskur's analysis and, as a result, they suggest that it does not belong in the genus *Saccharomyces*. They propose that the petite-positive character, a rare characteristic outside the *Saccharomyces* genus, arose after the separation of *S. kluyveri* from the common ancestor of *sensu stricto* and *sensu lato* yeasts.

1.1.5 Non-linear genome evolution in *Saccharomyces*

Karyotypic analysis continues to reveal natural hybrids among *Saccharomyces* species, including, most recently, *Saccharomyces* sp. CID1 which appears to contain chromosomes from two parent species and mitochondria from a third (Masneuf *et al.*, 1998; Groth *et al.*, in press). The *sensu stricto* yeast *S. pastorianus* has long been recognised as a hybrid between *S. cerevisiae* and *S. bayanus* (or an *S. bayanus*-like species such as *S. monacensis*). Further analysis reveals that both *S. cerevisiae*-type and *S. bayanus*-type chromosomes are present in *S. pastorianus* indicating that this species is an allotetraploid (Kielland-Brandt *et al.*, 1995; Tamai *et al.*, 1998; Petersen *et al.*, in press-a).

Despite the frequency of natural yeast hybrids polyploidy appears to have been an extremely rare event in the history of *S. cerevisiae*. The ascospores of *S. pastorianus* are usually not viable (Hansen, 1998). The lack of triplicated regions in the genome suggests that tetraploidy occurred only once in approximately 10^8 years (Wolfe and Shields, 1997). This suggests that the probability of allotetraploids such as *S. pastorianus* becoming fully fertile and surviving as a species is small. It is possible that horizontal transfer of single genes occurs more frequently while fixation of allotetraploidy is a very rare occurrence. If this is the case then karyotypic diversity among closely related species at a given time may give a false impression of the evolution of the yeast genome over long time spans.

1.1.6 The *Candida* genus

The species classified as *Candida* do not form a monophyletic group. Instead the genus *Candida* has become a kind of repository for all asexual ascomycetous yeasts that can not easily be placed in another genus (Meyer *et al.*, 1998). A clade of *Candida* species shares a deviation from the standard genetic code, with CUG-leu having been replaced by CUG-ser, probably some time close to the divergence of *C. albicans* from the lineage leading to *C. cylindracea* (Pesole *et al.*, 1995; Santos and Tuite, 1995; Kurtzman and Robnett, 1997). As a result it is difficult to express foreign DNA in *C. albicans*. However the CUG codon occurs with low frequency in *C. albicans* and has not had much of an impact on G+C content or on sequence divergence between *C. albicans* and *S. cerevisiae*. As a result the deviation from the standard genetic code is of limited significance for the present work.

The *Candida albicans* genome and sequencing project

The genome of *C. albicans* is slightly larger (~32 Mb of DNA per diploid genome; Magee, 1998) and more A+T-rich (~35% ORF G+C compared to ~40% ORF G+C content; Lloyd and Sharp, 1992) than the genome of *S. cerevisiae*. *C. albicans* is diploid with no known sexual cycle, a factor which complicates genetic analysis (Poulter, 1995; Tait *et al.*, 1997).

There is a high level of chromosome instability in *C. albicans* producing frequent deviations from the 2n = 16 complement of chromosomes. There is also considerable evidence to suggest that aneuploidy may provide a novel and elaborate means of gene regulation in *C. albicans* (Janbon *et al.*, 1998), and permit the harnessing of the resource of variable gene expression that can, for example, regulate the use of food supplies or produce resistance to the anti-fungal agent fluconazole (Janbon *et al.*, 1998; Perepnikhatka *et al.*, 1999). It is not clear how long *C. albicans* has been in its current imperfect (asexual) diploid state but in general *C. albicans* is homozygous at most loci (Hull and Johnson, 1999). Thus *C. albicans* must be a recent anamorph or possess an efficient means of gene homogenisation. Homogenisation could involve gene conversion or could be brought about by frequent aneuploidy involving chromosome loss followed, at some later stage, by duplication of the remaining chromosome. The latter could allow genomic rearrangements, particularly intrachromosomal rearrangements, to be fixed. If some form of chromosomal homogenisation has not occurred during *C. albicans* evolution then the high level of similarity between sister chromosomes in *C. albicans* implies that there has been little, if any, gene order rearrangement since *C. albicans* became an imperfect diploid.

C. albicans belongs to a clade of mostly anamorphic species (Kurtzman and Robnett, 1998,and see Fig. 1.2) and for this reason it is tempting to suggest that the anamorphic condition arose early in the evolution of *C. albicans*. However the existence of the teleomorph *Lodderomyces elongisporus* within the *C. albicans* clade (see Figure 9 in Kurtzman and Robnett, 1998) suggests that either the *Candida* species of this clade have all recently become anamorphic (possibly with several closely related teleomorphic species as yet undiscovered) or else that it was possible for *Lodderomyces elongisporus* (Kurtzman, 1998b) to revert to the perfect state. The balance of evidence, taking into account the lack of diversity among isolates suggests that *C. albicans* is a recent imperfect obligate diploid (Magee, 1998).

Because of its increasing importance as a human pathogen, particularly in immunocompromised patients (Lott *et al.*, 1999), *Candida albicans* is currently the subject of a genome sequencing project using a whole-genome shotgun sequencing approach. The project, which is based at Stanford University (C. albicans Sequencing Project), has already produced 1631 contigs greater than 2kb in size accounting for 14.9Mb of the ~16Mb genome. Several cosmids have also been sequenced at the Sanger Centre and a physical map of *C. albicans* chromosome 7 has been published (Chibana *et al.*, 1998). Preliminary data from these sources formed the basis of gene order comparisons between *C. albicans* and *S. cerevisiae* that are the subject of Chapter 5 of this thesis.

1.1.7 *Kluyveromyces lactis, Ashbya gossypii, Candida glabrata* and other related yeast species

Limited gene order data from *Kluyveromyces lactis* was a valuable early source of evidence for the hypothesis of ancient genome duplication in *Saccharomyces cerevisiae* (Wolfe and Shields, 1997; Keogh *et al.*, 1998). Because of the number of chromosomes in *K. lactis* and its location outside the *Saccharomyces* clade in the phylogenetic tree of Fig. 1.2, *K. lactis* was thought to be unduplicated. In a study involving 37 pairs of adjacent *K. lactis* genes that have homologues in *S. cerevisiae* we defined three categories of gene order conservation (Keogh *et al.*, 1998): complete conservation of an adjacent pair, conservation when duplicate blocks are taken into account and no evidence of adjacent pair conservation (see Fig. 1.3).

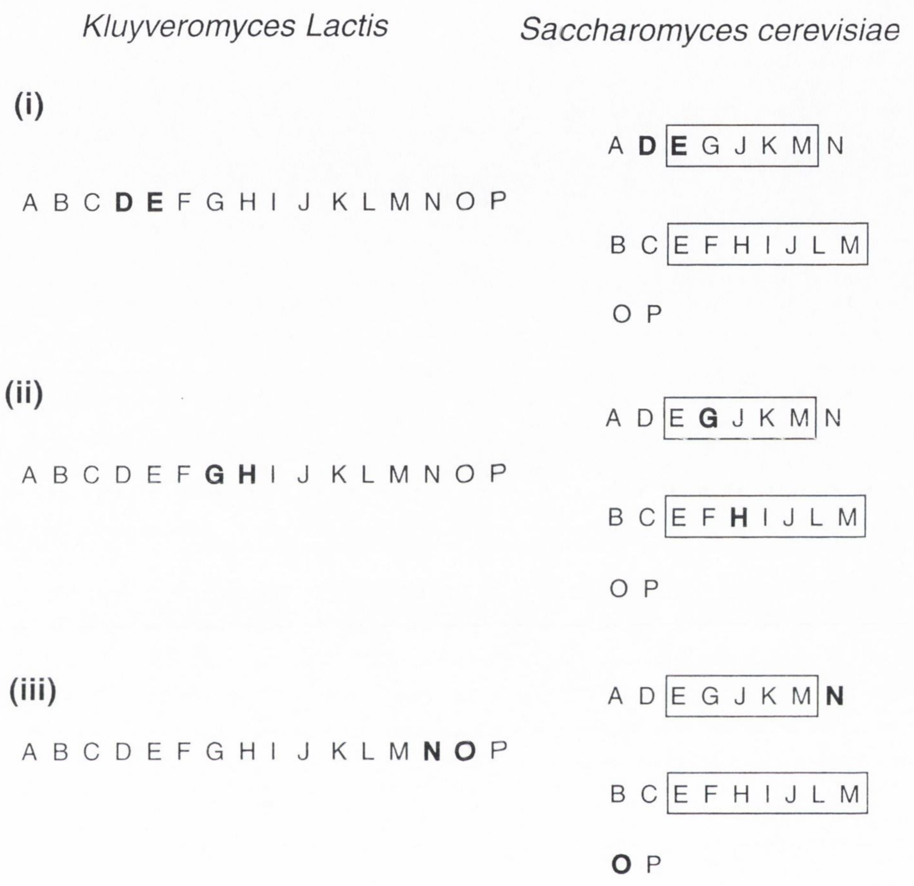


Figure 1.3 Schematic illustration of a single *Kluyveromyces lactis* chromosome with orthologues on three different *Saccharomyces cerevisiae* chromosomes. The adjacent pair D E of *K. lactis* genes (in bold) has remained adjacent in *S. cerevisiae,* (i), the pair G H are adjacent when the duplicated block (rectangle) is taken into account, (ii), and the pair N O shows no evidence of conservation, (iii).

We constructed a model to predict the number of pairs in each of the above categories, taking into account the following factors: (i) the incompleteness of the map of duplicated regions in the yeast genome; (ii) the disruption of adjacencies caused by reciprocal translocations; and (iii) the presence of duplicated genes in *S. cerevisiae*, which will increase the number of apparent conserved adjacencies. Assuming random deletion of single genes and random distribution of chromosomal rearrangements the predicted probability of conservation of an adjacency is

$$P_{adj} = t\{1 - 0.5((1 - d)/(1+d))^2\}$$

and the probability of block conservation is

$$P_{block} = \{b0.5((1 - d)/(1+d))^2\}$$

where d is the proportion of the pre-duplication genes that are retained in duplicate; b is the fraction of the genome covered by the map of duplicated blocks; and t is the probability that two genes that were originally adjacent have not been separated by a reciprocal translocation. The formulae above have not been published previously and are slightly corrected versions of the ones in our original publication (Keogh *et al.*, 1998). The observed values for the above parameters from the 37 pairs in the *K. lactis/S. cerevisiae* dataset used in Keogh *et al.* (1998) were $b = 0.68$, $d = 0.08$ and $t = 0.91$, giving $P_{adj} = 0.58$ and $P_{block} = 0.25$ using the formulae above. These predicted values are in relatively close agreement with the observed values; $P_{adj} = 0.59$, $P_{block} = 0.16$, and support the hypothesis of whole genome duplication. More extensive data from Ozier-Kalogeropoulos *et al.* (1998) were also in reasonably close agreement with the hypothesis (see Chapter 3).

A low-redundancy sequence map of chromosome I of *Ashbya gossypii* has been completed (Dietrich *et al.*, 1999) and also provides strong evidence for the hypothesis of genome duplication (P. Philippsen, personal communication). *A. gossypii* probably belongs to the same clade as *K. lactis* (Fig. 1.2) and it is the first of this group of species for which the gene map of a whole chromosome is available. Evidence for genome duplication can be seen as regions in chromosome I that share homology with two separate regions of *S. cerevisiae*.

Candida glabrata has a similar genome size (14 Mb) and number of chromosomes (14) to *S. cerevisiae* and may have undergone genome duplication (Seoighe and Wolfe, 1999b). However to establish from gene order data whether *C. glabrata* shares the same genome duplication as *S. cerevisiae* extensive gene order information would be required. It can be established that *C. glabrata* was duplicated in a separate duplication event if an undisrupted paired region exists in *C. glabrata* that contradicts the ancestral gene order in pre-duplication *S. cerevisiae* (established by reference to a third, unduplicated species). This would be evidence for a gene order rearrangement that occurred after the divergence of *S. cerevisiae* and *C. glabrata* but before genome duplication in *C. glabrata.* If the two species have a shared duplication and have a shared history for at least a part of the gene loss process then an excess of cases in which the same copy of a single-copy gene has been deleted in both species should be observed. It is likely to be much easier to establish whether *C. glabrata* and *S. cerevisiae* share a duplication event from accurate phylogenetic information and karyotyping of related species. If it is proven that *S. castellii* and *S. dairenensis* are unduplicated and the topology of the phylogenetic tree in Fig. 1.2 is confirmed then it is likely that the genome of *C. glabrata* has been duplicated independently to *S. cerevisiae* (if it has been duplicated at all).

1.2 Methods in gene order evolution

With the exponential increase in mapping and sequence data in recent years, there has been increasing interest in the branch of genomics concerned with the investigation of gene order evolution. Knowledge of gene order evolution can facilitate the use of comparative mapping to transfer information concerning the locations of disease genes and quantitative trait loci from densely mapped regions or species to species for which there is less map data available (Edwards, 1994; O'Brien *et al.*, 1999). Comparison of gene order between species can yield information on the phylogenetic relationships between them as well as on their evolutionary histories. Gene order comparison is sometimes useful even when phylogenetic inference from sequence comparison is difficult or uncertain. Rare gene order changes (rearrangements) in otherwise stable genomes can provide powerful cladistic characters for phylogenetic analysis (e.g. Boore, 1999; O'Brien *et al.*, 1999). The relative rates of evolution at the sequence and gene order levels vary greatly among species and genomes (Palmer and Herbon, 1987; Palmer and Herbon, 1988). For example, the order of genes from mitochondrial genomes has been shown to be of use in phylogenetic analysis when inference directly from sequence data is difficult due to saturation in point mutation (Sankoff *et al.*, 1992) or lack of point mutation in the case of plant mtDNA (Palmer and Herbon, 1988). The combination of sequence comparison and gene order information can help to improve the resolution of phylogenetic relationships.

1.2.1 Gene order versus sequence alignment

The importance of nucleotide and amino acid sequence alignments for the inference of phylogenetic relationships among species and taxa has been enormous. Using the still controversial molecular clock hypothesis and appropriate calibration dates it has also been possible to estimate dates of sequence divergence through sequence comparison (references in Li, 1997). Much effort has been devoted to developing efficient methods of comparing pairs of gene or protein sequences so that the most parsimonious account of the evolution of the sequence pair can be constructed from compositions of simple evolutionary operations. These operations involve single character substitutions as well as deletion or insertion of sequential subsets of characters. It is possible to place strings, related by these operations, in simple alignments in which they are side by side, with the identical characters connected by non-intersecting lines (Sankoff and Goldstein, 1989; Sankoff *et al.*, 1992). Methods have been developed for optimising these simple alignments so that the evolutionary "distance" separating the strings is minimised. Penalties are assigned to each of the evolutionary operations and the distance is defined as the sum of the penalties associated with the inferred operations. Sequence alignment methods can be broadly categorised into two varieties: similarity based methods, that maximise the number of matched pairs in the sequence, and difference based methods, that minimise the number of mismatched pairs (references in Li, 1997). The condition of non-intersection of the connecting lines has been important for sequence alignment techniques (Sankoff *et al.*, 1990). The types of operations that could disrupt this direct alignment, like inversions or transpositions, do not normally occur within gene sequences. Attempts have been made to develop string alignment techniques that allow for transpositions but this has proven to be a difficult problem mathematically (Sankoff and Goldstein, 1989).

A genome, in the simplest abstraction, can be thought of as a one dimensional array of objects called genes, interspersed with non-coding regions, centromeres and chromosome concatenation points (the telomeres). It is desirable to develop some method of species comparison at the level of the organisation of this array (Hannenhalli *et al.*, 1995). Although the analogy with sequence comparison is clear there are many qualitative differences between sequence and genome comparisons (Sankoff and Goldstein, 1989; Bafna and Pevzner, 1995). At the level of the genome, reciprocal translocations, inversions and transpositions are important. They are the major forms of gene order change and because of them genomes cannot be aligned in such a way as to connect homologous genes with non-intersecting lines, as is the case with sequence alignments. Insertions, deletions and substitutions, the relevant transformations at the gene or protein sequence level, tend not to be as important at the level of gene order. Unlike single bases or amino acids (the characters in sequence alignments) the genomic fragments themselves contain information about the similarity of the genomes and their divergence time. Lengths of fragments were used by Nadeau and Taylor (1984) to determine the extent of genome re-arrangement since the divergence of mouse and human.

1.2.2 The Nadeau and Taylor method

Nadeau and Taylor in 1984 pioneered the approach to gene order evolution based on the number and the length of genome fragments conserved between a species pair. Because the characters (conserved chromosome fragments) are created by evolutionary operations there is a correspondence between the number of characters and the number of evolutionary operations that have taken place. Nadeau and Taylor studied the mouse and human genome comparison. In this case, due to insufficient mapping data (equivalent to throwing away most of the characters in the string comparison) the number of characters could not be counted. Instead a probability based technique was developed to determine the average length of the conserved segments given the average length of conserved segments that had been observed, and adjusting for the fact that most of the observed conserved segments were incomplete and that there was an increased probability of observation for longer segments. In a very influential paper Nadeau and Taylor (1984) estimated that there have been about 180 linkage disruptions since the divergence of mouse and man.

A maximum likelihood approach to the same problem was devised by Sankoff *et al.* (1997b). From combinatorial arguments he has shown that for a genome with n gene order breakpoints (relative to some other genome and including chromosomal concatenation points) and m markers the probability that a non-empty segments will be observed reduces to the remarkably simple form

$$P(a,m,n) = \frac{\binom{m-1}{a-1}\binom{n+1}{a}}{\binom{n+m}{m}}$$

where a segment is defined as the region between two breakpoints. To arrive at a maximum likelihood estimate *n* is adjusted until the probability of observing *a* non-empty segments is a maximum.

1.2.3 Determining conserved segments from synteny data alone

Sankoff and Nadeau have recently extended the work of Nadeau and Taylor to estimate the number of breakpoints given synteny data alone (Ferretti *et al.*, 1996; Sankoff and Nadeau, 1996). This is useful because many genes have been mapped to chromosomes without any additional resolution, particularly in mammals other than human or mouse. Two genes are syntenous if they are on the same chromosome. Synteny is conserved between two genomes for a pair of genes if they are syntenous in both species. In the most recent version (Erlich *et al.*, 1997), the following function is derived for the probability of finding *r* genes in any segment: $P(r)=\frac{(n+c-1)m!(n+c+m-r-2)}{(n+c-1+m))!(m-r)!}$, where *c* is the number of chromosomes in a genome containing *m* markers and *n* breakpoints. Sankoff and Nadeau (1996) constructed a likelihood expression from this for the number of synteny sets containing a given number of markers. The only unknown variable in their expression is the number of segments, *n*. By maximising the likelihood of the observed frequency distribution of synteny sets the value of *n* can be estimated. Erlich *et al.* (1997) have shown that this method gives a realistic estimate (~140) for the number of conserved syntenies between mouse and man. This is smaller than estimates of the number of conserved segments (181). The difference can be explained by intrachromosomal rearrangements that are not accounted for in the synteny-based analysis. By estimating the number of conserved syntenies between mouse and man Erlich *et al.* provide an estimate of the relative rate of intra- versus interchromosomal rearrangement. However because estimates of the number of conserved linkages between mouse and man tend to miss small conserved linkages it is likely that this estimate of intrachromosomal rearrangement does not take account of small inversions (see Chapter 5). A cursory examination of the conserved orthologous segments in mouse and man (as presented by DeBry and Seldin, 1996) seems to indicate that the orientation with respect to the centromere of segments conserved between mouse and man is not conserved. Unlike in the case of yeast (Wolfe and Shields, 1997) large intrachromosomal rearrangements may have been common in mouse and human.

Constructing conserved segments

A method has been developed by Sankoff *et al.* (1997a) to identify the most likely set of conserved segments given limited map data. They have defined what they call the objective function, $D = \sum D_i$,

$$D_i = \gamma \max_{x,y \in i(1)} |x - y| + \alpha s[i(1)] + \gamma \max_{x,y \in i(2)} |x - y| + \alpha s[i(2)] - \beta m(i)$$

where α, β and γ are variable parameters, $x \in i(j)$ is the map position of a gene in segment *i* in species *j*, $m(i)$ is the number of orthologues in segment *i* and $s[i(j)]$ is the number of other segments that overlap with segment *i* in species *j*. Sankoff has developed a simple algorithm, *CONSEG,* which begins by considering each orthologue in a conserved synteny as a separate segment (Sankoff *et al.*, 1997a). It joins pairs of these segments together to form larger segments starting with the pair that produces a maximal decrease in *D*. When the final segments are produced they must satisfy the criterion that the number of species 2 chromosomes represented on each chromosome of species 1 be equivalent to what we would expect if the segments on species 1 were randomly distributed over the chromosomes of species 2. Departures from this might indicate that neighbouring segments on species 1 are segregating jointly onto the chromosomes of species 2. Jointly occurring segments might indicate that larger segments have been missed. The parameters are varied so that the maximum number of segments is produced but the segments on a given chromosome are still randomly distributed on the other species. This method is likely to be useful for certain kinds of data but can not easily be adapted for finding sister regions in the duplicated yeast genome. The yeast data contains additional noise not found in cross-species comparison due to the existence of duplicated genes arising from duplication events other than whole genome duplication.

1.2.4 Constructing the most parsimonious paths for gene order evolution

Sridhar Hannenhalli reformulated the problem of determining the most parsimonious way in which a subject genome can be transformed into a target genome by translocation (Hannenhalli *et al.*, 1995; Hannenhalli, 1996). He displays all the genome fragments as pairs of vertices of a graph. Each fragment, x, is represented by the vertices x^t and x^h representing head and tail so that orientation is taken into account. Vertices which are neighbours in the subject species (species a) are joined by grey lines and vertices which are neighbours in the target species (species b) are joined by black lines. x^h and x^t are never connected (see Fig. 1.3). Since each vertex is connected to the rest by exactly one black edge and exactly one grey edge the graph may be decomposed into disjoint cycles. A 2-cycle (e.g. vertices 5h and 6t in Fig 1.3) indicates that two genes which are adjacent in species a are also adjacent in species b. When all cycles are two cycles (or equivalently when c_A, the number of cycles, is a maximum) then gene order in species a and b is the same. Hannenhalli defines a sub-permutation as an interval, *I,* within a chromosome such that there exists no edge connecting a vertex within the interval to a vertex outside and such that there is at least one cycle of size greater than size two within *I.* A minimal sub-permutation is defined as a permutation that contains no other sub-permutations. Hannenhalli proves that for the target genome *B*, and the subject genome, *A,*

$\mathrm{d}(A\ ,\ B) = n - N - c_A + s_A + 1$ if A defines an odd number of minimal sub-permutations

$\mathrm{d}(A\ ,\ B) = n - N - c_A + s_A$ if A defines an even number of minimal sub-permutations

or $\mathrm{d}(A\ ,\ B) = n - N - c_A + s_A + 2$ in a well-defined special case

where $\mathrm{d}(A\ ,\ B)$ is the translocation distance to the target genome, n is the number of genes in A, N is the number of chromosomes, c_A is the number of cycles and s_A the number of minimal sub-permutations.

Hannenhalli has also developed an efficient algorithm for finding a set of operations for transforming the subject genome to the target genome. This algorithm was applied to the example of the herpes virus family (Hannenhalli *et al.*, 1995) to find ancestral gene orders. It is not clear that the ‘ancestral’ genomes found by these methods remain meaningful for more complicated examples or when the distance separating the genomes is increased.

Sankoff *et al.* (1992), Kececioglu and Sankoff (1993) and Bafna and Pevzner (1995) undertook an analysis of sorting using reversals only. They produced upper and lower bounds for the sort distance, going by reversals only, between two genomes. A good example of the usefulness of these techniques was given by Sankoff *et al.* (1992), who constructed a phylogenetic tree for eukaryotes based on the "edit distance" between mitochondrial genomes. Most of the applications, so far, have been restricted to organelle and viral genomes although work has also been carried out on the human X chromosome, which, as per Ohno's law (Ohno, 1970), is highly conserved (Bafna and Pevzner, 1995). A generally available programme, DERANGE, was also produced to find this distance. It is not clear whether putative ancestral gene orders produced by the most parsimonious account of genome evolution are meaningful. The example given by Kececioglu and Sankoff (1993) of flatworm and mammalian mitochondrial genomes does not appear to be useful because there have been too many rearrangements since divergence and the probability of repeated use of breakpoints is high.

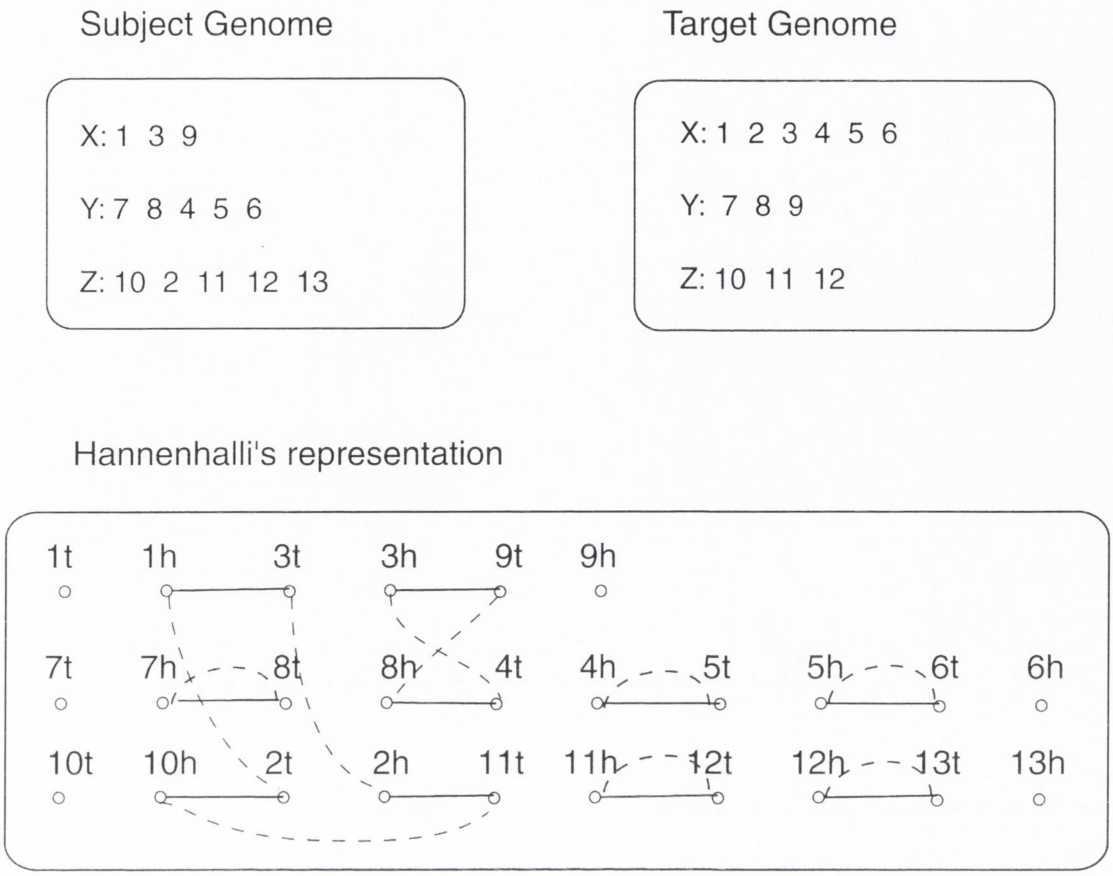


Figure 1.3 Hannenhalli's formalism. X, Y and Z represent chromosomes and numbers represent genes. Vertices that are adjacent in the target genome are connected by dashed lines. Adjacent vertices in the subject genome are connected by full lines. Each internal vertex is connected to one dashed and one full line and so cycles are composed of alternating dashed and full lines.

1.2.5 "Genome Halving", a maximum parsimony approach for the duplicated yeast genome

Nadia El-Mabrouk *et al.* (1998; 1999) tackled the problem of developing an analytical estimate of the minimum number of reciprocal translocations after genome duplication in an ancient tetraploid in which gene order has evolved primarily through reciprocal translocation. The approach was based on the formalism of Hannenhalli (1996) and, as an example, the results were applied to the map of duplicated blocks in *S. cerevisiae* from Wolfe and Shields (1997). The minimum number of reciprocal translocations required to produce two separate copies of an unduplicated genome was 45 and one possible ancestral unduplicated genome was suggested. However no unique ancestral gene order can be inferred from maximum parsimony methods based on "undoubling" the map of duplicated regions (see Chapter 2). No mention was made by El Mabrouk *et al.* of the problem of non-uniqueness of the most parsimonious unduplicated ancestral genome. As yet nothing has been established about the class of possible ancestral gene orders that satisfy the condition of minimum number of reciprocal translocations after genome duplication.

1.2.6 Multiple genome rearrangement and breakpoint phylogeny

David Sankoff's group has produced methods of ancestral genome reconstruction that are based on breakpoint analysis alone. If gene *a* is adjacent to gene *b* in genome *A* but not in genome *B* then the edge *ab* defines a breakpoint for genomes *A* and *B*. Sankoff and Blanchette (1998) have shown that the problem of determining a consensus ancestral genome for three or more known gene orders reduces easily to an instance of the Travelling Salesperson Problem[a] and can be solved algorithmically. They have shown through simulation that the uniqueness of the resulting ancestral genome depends on the number of rearrangements and on the number of genomes being compared. This method can be adapted to determine ancestral gene order based on a tree of fixed topology. They do not solve this version of the problem exactly but produce a solution that depends to some degree on the initialising strategies used.

The above methods have been applied by Blanchette *et al.* (1999) to the problem of determining phylogenetic trees using metazoan mitochondrial gene order. Minimum breakpoint ancestral mitochondrial gene orders were produced based on every possible tree of the major metazoan groupings. In a maximum parsimony approach, trees containing the smallest numbers of breakpoints were favoured. This was found to be the most effective way of using breakpoint data to construct phylogenetic trees from mitochondrial gene order and outperformed Neighbour-joining and Fitch-Margoliash routines based on a breakpoint distance matrix of mitochondrial gene orders. The exact topology of the metazoan phylogenetic tree is still a subject of debate. The most parsimonious trees produced by this method were not in exact agreement with any of the main theories of metazoan phylogeny and were of limited use for distinguishing between competing theories. It is also unclear how a maximum parsimony based approach to choosing between different phylogenetic trees is superior to a cladistic approach based on shared fragments of gene order. Methods based on gene order reconstruction have the advantage of estimating ancestral gene order

[a] A travelling salesperson must visit *n* towns, once each, travelling the least possible distance and ending up at the starting point. In this case each vertex of the complete graph containing the full set of genes must be visited and the distance between pairs of vertices of the graph is weighted by the number of times that the pair of genes is not adjacent in the genomes for which a median is being sought.

but this is normally not uniquely determined (although small fragments of gene order can be reconstructed uniquely).

1.2.7 Validity and limitations of statistical methods

In cases where there is incomplete knowledge of conserved segments as a result of insufficient data (Nadeau and Taylor, 1984; Sankoff *et al.*, 1997a) or gene deletion (Seoighe and Wolfe, 1998) statistical or probability based methods must be used to estimate the number of genome rearrangements since an evolutionary event. These methods normally rely on the random distribution of breakpoints and genetic markers. Whether breakpoints are randomly distributed is open to debate (Nadeau and Sankoff, 1998, see also Chapter 5) and may depend on whether regional chromosomal organisation is critical for the activity of genes in some cases. Rearrangements that effect chromatin structure may also influence levels of gene expression since supercoiled DNA is expressed at a much higher level than linearized DNA (Lundin, 1993; Sankoff and Nadeau, 1996). This could produce selection against certain rearrangements. Reciprocal translocations that produce excessively long or short chromosomes may also be selected against (Schubert and Oud, 1997). The effects of non-random distributions of genetic markers in the genome have been discussed by Sankoff *et al.* (1997b). Most of the statistical methods discussed above have been developed for a particular kind of data. They work best with short randomly distributed markers. Modern approaches to sequencing do not always result in short sequences scattered throughout the genome. Large-scale sequencing projects, which are now commonplace, produce clear information about chromosomal rearrangement. Assembled contig data allows simple estimation of the number of short-range chromosomal rearrangements but requires new techniques for the estimation of interchromosomal events.

1.3 Gene duplication

Gene duplication is required for the evolution of novel biological functions. It is clear that with increasing complexity the number of cell types increases as does the number of genes required to completely specify an organism (Miklos and Rubin, 1996). For example, vertebrates have approximately four times the number of genes in *Drosophila* (Sidow, 1996). New genes are produced by modification of existing genes through mutation. Modifications in critical sites of proteins with essential functions cannot be fixed however unless the essential function can be maintained (Ohno, 1970). If the essential gene is duplicated one copy may become redundant and mutation in that gene can often be neutral or nearly neutral. This leaves the additional gene free to evolve by random genetic drift (Kimura, 1983). There are two likely evolutionary reasons for retaining both copies of a duplicated gene: selection for increased levels of expression, or divergence of gene function. Functional divergence can be produced through complementary degeneration (Force *et al.*, 1999), where each daughter gene retains only a subset of the functions of the parent, or (perhaps more rarely) if one daughter acquires a new function.

1.3.1 Types of gene duplication

Gene duplication can occur through the duplication of part or all of a single linkage group or through the simultaneous duplication of all linkage groups (Ohno, 1970). Duplications involving a small number of genes (usually one) can be tandem duplications or duplications resulting from (retro)transposition. Tandem duplications are due to unequal cross-over during cell division. Duplicates are located side by side on the chromosome and frequently remain linked due to their proximity. Retrotransposition gives rise to a duplicate of a single gene, and usually produces processed pseudogenes. The duplication is mediated by an RNA intermediate and the duplicate gene can be located anywhere in the genome and lacks introns. Duplication involving all of the genes of the genome is called whole genome duplication or polyploidy.

1.3.2 Whole genome duplication

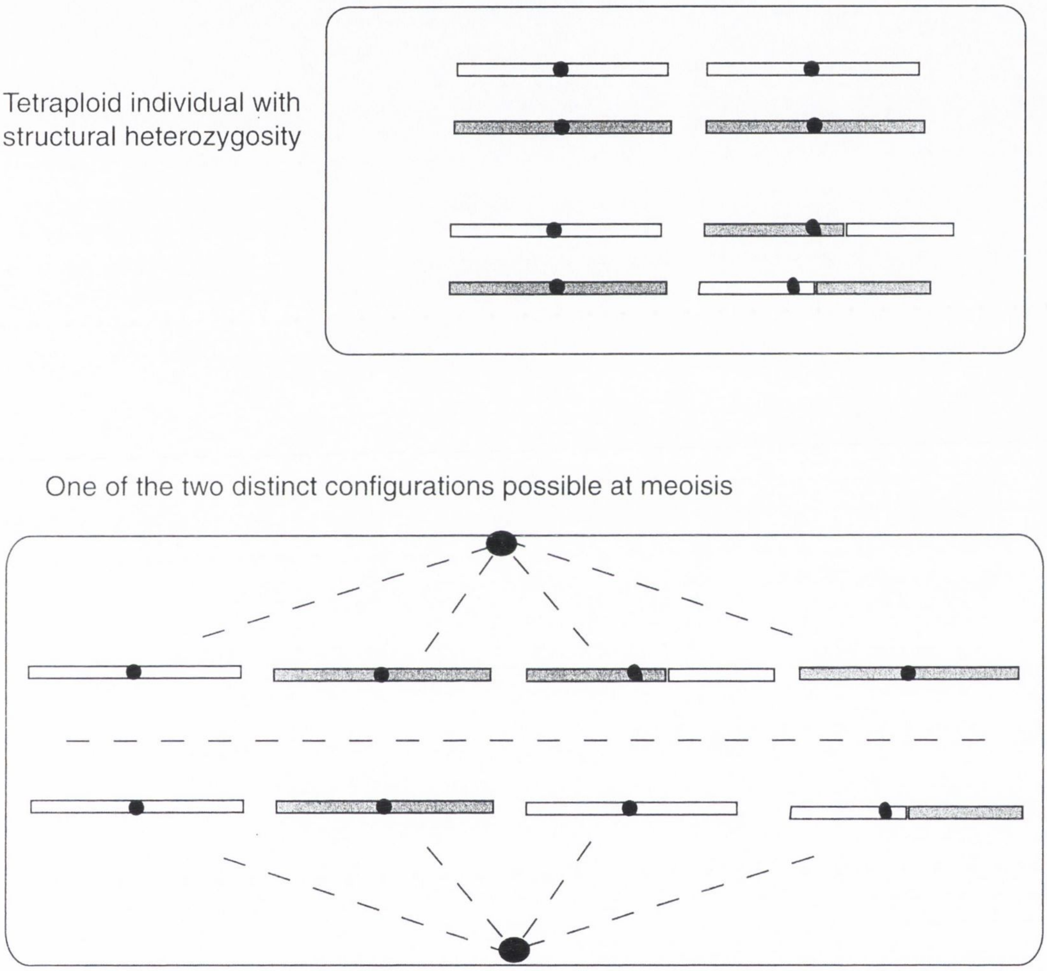
Whole genome duplication is believed to have played an important role in the evolution of complex organisms. Evidence of recent genome duplications can be seen in many plants. For example three species of the cereal Sorghum, *S. versicolor*, *S. sudanese* and *S. halpense* have 10, 20 and 40 chromosomes respectively and are diploid, tetraploid and octoploid. Recent polyploids are scarce among vertebrates and invertebrate animals due to the incompatibility of tetraploidy with the mechanism of sex determination in most animals (Ohno, 1970). However one example of a viable tetraploid mammal has recently been reported (Gallardo *et al.*, 1999). Tetraploids that have descended from a single ancestral diploid species are referred to as autotetraploids. Tetraploids derived from interspecific hybrids are known as allotetraploids. Plant hybrids can be infertile due to chromosomal rearrangement in one of the parent species that prevents homologous pairing of chromosomes during meoisis. Allotetraploidy may be an important means by which fertility can be restored. In certain cases some chromosomes exhibit tetravalent formation during meiosis while others assume bivalent formation. Such species are referred to as segmental allotetraploids. In ancient segmental allotetraploids duplicated genes resulting from polyploidy may have very different times of divergence. The ancestor of maize appears to have been a segmental allotetraploid (Gaut and Doebley, 1997).

There is considerable evidence to suggest that *Saccharomyces cerevisiae* is an ancient tetraploid that underwent auto or allo-tetraploidy approximately 10^8 years ago (Wolfe and Shields, 1997, and see section 1.1 above). The estimated date of polyploidy in *S. cerevisiae* coincides with the time that fruit-bearing plants became abundant in the Earth's flora. Species of *Saccharomyces sensu stricto* and *sensu lato* are Crabtree positive, that is, they ferment glucose vigorously under aerobic conditions if the glucose concentration is sufficiently high. Polyploidy may have been crucial to the evolution of this ability. Several of the genes with homologues believed to be attributable to genome duplication are regulated differently under aerobic and anaerobic conditions, for example *COX5A/5B* and *CYC1/*CYC7. *Kluyveromyces lactis,* which diverged from *Saccharomyces* before duplication as well as *Saccharomyces kluyveri* (which also appears to be unduplicated; Langkjaer *et al.*, pers. comm.) are Crabtree negative. Other yeasts may also have undergone whole genome duplication independently to *S. cerevisiae* (Keogh *et al.*, 1998).

1.4 Genome evolution after whole genome duplication

Diploidisation is the process whereby a tetraploid species resumes completely disomic inheritance. Genes on chromosomes that continue to exhibit tetravalent formation during meoisis cannot develop new functions since the four copies of the gene will be distributed randomly among the daughter cells (Gaut and Doebley, 1997). Structural heterozygosity can cause a unique pairing of the four chromosomes during meoisis. Inversions, transpositions and reciprocal translocations thus play an important role in the process of diploidisation. There may have been an increase in the rate of reciprocal translocations in the wake of genome duplication in some species, for instance maize (Ahn and Tanksley, 1993). Locations of chromosomal rearrangements that occur after genome duplication can be detected as disruptions in duplicated sections of the genome. Reciprocal translocations subsequent to genome duplication are expected to occur most frequently between homologous regions of the duplicated genome. However only reciprocal translocations involving non-homologous parts of the genome can be easily detected as disruptions in duplicated sister regions. Deletion of genes from a chromosome can also bring about a unique pairing at meoisis.

In some cases it may be possible for diploidisation through interchromosomal rearrangements in sexually reproducing species to give rise to triplicate regions. This is the case because individuals heterozygous for an interchromosomal rearrangement in a tetraploid can produce gametes with three copies of translocated genome fragments, as shown in Fig. 1.4. This implies that the occurrence of some triplicate regions may not necessarily imply that any other form of large-scale duplication has occurred besides the genome duplication. When triplicate regions are found they may point, however, to chromosomal rearrangement rather than gene deletion as being the key process that mediated the diploidisation.



Tetraploid individual with structural heterozygosity

One of the two distinct configurations possible at meoisis

Figure 1.4 Illustration of the theoretical possibility of triplicate regions in a recent tetraploid that has undergone chromosomal rearrangement. The example involves a tetraploid genome with two distinct chromosomes.

1.4.1 The fate of duplicated genes

After diploidisation has taken place the two copies of each gene begin to diverge. Frequently one copy acquires null mutations and becomes a pseudogene or is deleted altogether. Kimura estimated that the half-life of completely redundant gene pairs in vertebrates is 50 million years (Kimura, 1983). In catostomid and salmonid fishes about 50% of the genes remain in duplicate since a polyploidization event about 50 MYA (references in Lundin, 1993). Many of these genes may remain completely redundant. We have estimated that only approximately 8% of the genes duplicated in a round of polyploidy in *Saccharomyces cerevisiae* about 100 MYA have survived (Seoighe and Wolfe, 1998, see Chapter 2).

Differences in the rate of gene loss among species may be as a result of differences in the rate of resolution of gene redundancy as well as differences in generation time. It has been suggested that redundancy in genes involved in development can be maintained for longer than redundancy in "house-keeping" genes and can, in fact, be evolutionarily stable (Cooke *et al.*, 1997; Nowak *et al.*, 1997; Gibson and Spring, 1998). If so, then it should be no surprise that duplicated genes can persist for longer periods in more complex organisms. Redundancy may also be maintained by appropriately balanced mutation rates (Nowak *et al.*, 1997). The number of extra genes contributed by the ancient round of polyploidy in *S. cerevisiae* to the current number is likely to depend on the relative probabilities of fixation of a null allele and evolution of a novel function in duplicated genes. This rate in turn depends on the effective population size around the time of genome duplication (Walsh, 1995). Most of the genes retained in duplicate since genome duplication in *S. cerevisiae* have diverged significantly from their homologues with a mean amino acid identity between pairs of about 63% (Wolfe and Shields, 1997). The number of completely identical amino acid sequences is small.

1.5 Comparative mapping

Comparative mapping is important for the study of models of human disease-related loci in animals as well as for combining research efforts into relevant loci in agriculturally important organisms (Van Deynze *et al.*, 1995a; Carver and Stubbs, 1997). One of the aims is to identify orthologous segments of the genomes that have remained linked since species divergence. Much effort has gone into the identification of linkage groups conserved between human and the laboratory mouse. At least 181 such segments have been identified (DeBry and Seldin, 1996). However distinguishing between conserved linkage groups and coincidental syntenies can be a difficult problem, particularly if both intra- as well as inter-chromosomal rearrangements have taken place. The problem of devising a systematic approach to conserved segment identification has also been tackled by Sankoff *et al.* (1997a). They approach the problem by identifying each pair of orthologues as a conserved linkage segment and then fusing the segments that give the greatest increase in the "integrity" of the collection of segments until no further increase can be achieved (see section 1.2).

1.5.1 Intraspecific comparative mapping

Some of the techniques described above (section 1.2) can be adapted for use in a single organism by treating the duplicated genome as two genomes that diverged at the time of duplication. For example, in Chapter 2, we have adapted Nadeau and Taylor's method of calculating the number of rearrangements since the divergence of mouse and man to approximate the number of chromosomal rearrangements since genome duplication in *S. cerevisiae* (Seoighe and Wolfe, 1998). Methods developed to systematically identify orthologous segments between species may also be adapted for use in identifying duplicated segments conserved since genome duplication. Conversely lessons learned in the systematic identification of duplicated segments in an ancient tetraploid may, in turn, be applied to the task of identifying orthologous segments. Comparative mapping is complicated by genome duplication because of the phenomenon of differential silencing of genes duplicated since polyploidy (Lundin, 1993; Kurata *et al.*, 1994). A proper understanding of the evolution of genome organisation after genome duplication is a pre-requisite for understanding the comparative locations of orthologous genes in species separated by a stage of polyploidy (Keogh *et al.*, 1998). Joseph Nadeau (1991) has given a good account of the importance of the study of genome duplication particularly for comparative gene mapping and the analysis of genome organisation and evolution as well as some of the difficulties involved. He uses the term *intraspecific comparative gene mapping* to refer to the identification of members of gene families according to their positions within duplicated regions.

1.6 The Cereals, genome duplication and comparative mapping

The plant family Gramineae contains about 10,000 species and includes the cereals that form the staple diet of most of the world's population (Ahn and Tanksley, 1993; Moore *et al.*, 1995a). The study of cereal genetics is clearly of great agronomic importance and individual crop species have been studied extensively for the past 50 years. In the 1980's it became possible to make comparisons between species. Although the sizes of the grass genomes vary by as much as a factor of 40, gene order at a gross level is found to be relatively well conserved (Ahn *et al.*, 1993). The enormous range of genome sizes in the cereals, from the 16,000 Mb genome of wheat to 400Mb in rice is accounted for by differences in the amount of intergenic DNA accumulated (Moore, 1995; SanMiguel *et al.*, 1996; Bennetzen and Freeling, 1997) as well as gene and genome duplication. It is clearly useful to exploit gene order conservation in the cereals so that information about the location of genes or gene families with important functions can be imported from one cereal genome to another. Rice is currently the subject of a multinational sequencing project and the whole genome is likely to be sequenced within five years (Bevan and Murphy, 1999). Using sequence information, as it emerges, from rice and functional characterisation from experimental species such as *Arabidopsis thaliana* it should be possible to locate genes responsible for important genetic traits in several cereal species (Bevan and Murphy, 1999).

Cereal comparative genomics

Genomic maps of the cereals are compared using the technique of 'rice linkage segment analysis' (Moore *et al.*, 1995a). This approach clarifies the relationships between the gene orders found in different grass species by comparing gene order in each species to the order found in apparently conserved segments in rice. The cereal genomes are represented as concentric circles with the twelve rice chromosomes arranged in the centre (Fig. 1.5). The initial suggestion of Graham Moore *et al.* (1995b), that this circular arrangement of rice chromosomes could represent the single ancestral chromosome of the grasses no longer features in the literature. Although there has been repeated demonstration of the usefulness of this model for reducing the complexity of comparative mapping in the cereals (Moore *et al.*, 1995a) it is not based on a phylogenetic approach. Some attempts to reconstruct the evolutionary history of cereal genomes have made use of the rice segments as a basis for their analysis (e.g. Wilson *et al.*, 1999). However the representation of cereal genomes as mosaics of rice chromosomal segments is tied to the assumption that the gene order of rice resembles the ancestral order for the cereals. Arguments in favour of this possibility have been unconvincing. The circular representation of the cereal genomes may be a useful simplification of cereal gene order relationships but should be applied with caution to the study of cereal genome evolution. The criteria by which conserved rice linkage segments are determined are not clearly defined. By comparing conserved rice linkage maps produced for the maize genome by Moore *et al.* (1995a) and Wilson *et al.* (1999) it is clear that are many changes brought about by improvements in the maize genetic map.

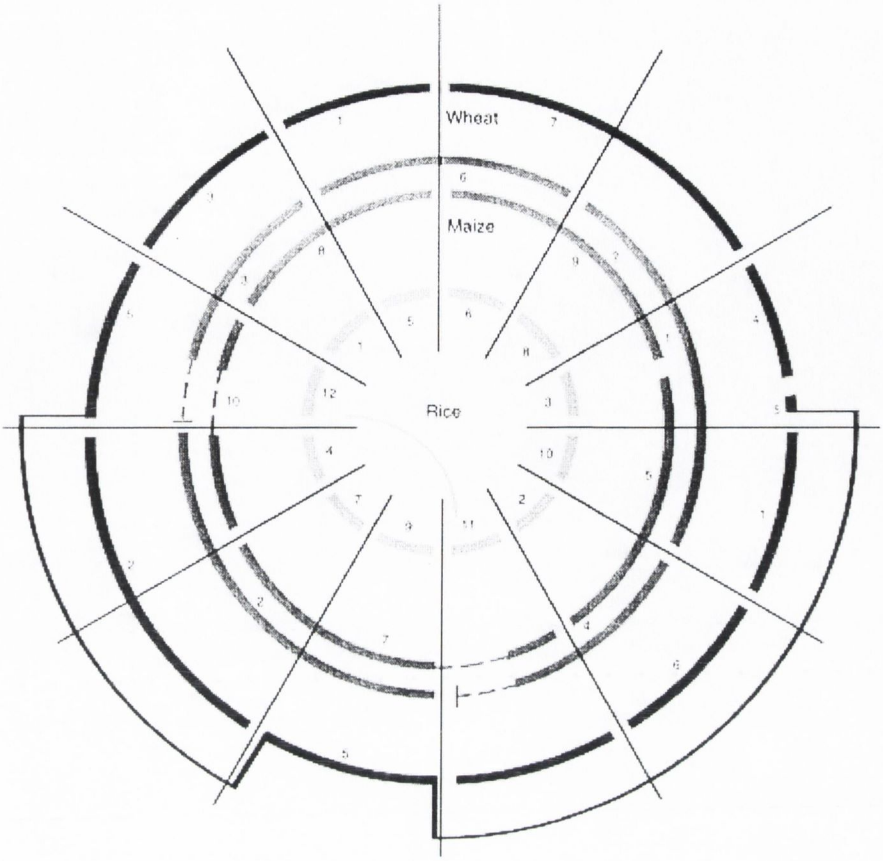


Figure 1.5 Simplified version of the concentric representation of cereal chromosomes around the chromosomes of rice (taken from Bennetzen and Freeling, 1997). The genomes of rice, wheat and allotetraploid maize are represented in this example.

It is estimated that approximately 50% of angiosperms are polyploid (Moore, 1995). Polyploidy is a common feature of domesticated plant genomes including the cereals, which belong to the plant family Gramineae. Wheat, for example has a hexaploid genome (Van Deynze *et al.*, 1995b). There is also considerable evidence of ancient tetraploid events in the cereal genomes. The genome of soyabean may have undergone two rounds of polyploidisation (Shoemaker *et al.*,

1996) during its history and may make an interesting comparison with the human genome. Maize is an ancient tetraploid which underwent genome duplication approximately 20 million years ago (Gaut and Doebley, 1997). Recent research into the process of diploidisation in maize, indicates that maize may have been a segmental allotetraploid and the return to completely disomic inheritance may have taken place over a prolonged period of time (Gaut and Doebley, 1997). This would explain the clusters of genes showing different divergence times in maize.

Gene order evolution after genome duplication in plants

The rate of accumulation of chromosomal rearrangements increases following a genome duplication, because of the relaxation of selection pressures brought about by having two copies of the complete genome (Ahn *et al.*, 1993; Ahn and Tanksley, 1993). Song *et al.* (1995) produced synthetic allotetraploids of *Brassica* species and used RFLP markers to detect genomic changes after several generations of self-pollination. They found an increase in the rate of loss and gain of restriction fragments following allotetraploidy with the greatest increase in polyploids created from more distantly related parents. This kind of chromosomal rearrangement may play an important role in the process of diploidisation. Ahn has suggested that comparative mapping carried out by his laboratory indicates that nearly as many chromosomal rearrangements have taken place within the maize genome since polyploidisation as have occurred between maize, wheat and rice since divergence from their last common ancestor (Ahn *et al.*, 1993). The rice conserved linkage segments also appear to exhibit greater fragmentation in the duplicated maize genome than in other cereals (Wilson *et al.*, 1999).

Diploidisation

In wheat (hexaploid) homeologous pairing between chromosomes derived from constituent genomes may be influenced by deletion of the single locus Ph1 (Moore, 1995). Since gene order is well conserved in wheat, chromosomal pairing may be determined by deletion of genes. However, some evidence of chromosomal rearrangement has also been found in the wheat genome (Ahn *et al.*, 1993). Further research is required to uncover more detail about the factors governing chromosomal pairing at meiosis. This will allow a better understanding of how natural polyploids revert to the pre-duplication mode of inheritance.

Moore *et al.* refer to the conserved rice linkage segments as "lego" blocks, implying that gene cereal gene order changes have come about by shuffling of fixed blocks of genes (Moore, 1995). However, if we look at Moore's comparisons of wheat, maize, sugar cane, foxtail millet, sorghum and rice (Moore *et al.*, 1995a), we see that some chromosomal rearrangements seem to have been unique to one or other of the duplicated maize chromosomes. An interesting feature of diploidisation in maize is the reduction in chromosome number from sixteen to ten if the ancestral maize genome constructed by Wilson *et al.* (1999) is correct. More detailed study of the chromosomal rearrangements since duplication in maize may reveal further information about possible processes through which chromosome number can be reduced such as Robertsonian translocations.

1.7 Yeast research after whole genome sequencing

Following complete genome sequencing of yeast there remains the even greater challenge of establishing a complete understanding of yeast regulatory mechanisms and assigning functions to novel genes that have been discovered. Over one third of the genes of *S. cerevisiae* still have no experimentally characterised functions and no close homologues whose functions have been characterised (Hodges *et al.*, 1999). Some of these genes were not essential for yeast growth under any known conditions but may have marginal fitness effects that are difficult to characterise (Thatcher *et al.*, 1998). An enormous aid to further understanding of yeast biology and a logical next step after complete genome sequencing is the use of DNA chip technology to provide whole-genome expression data for all of the genes in the genome simultaneously under different conditions (Lockhart *et al.*, 1996; Holstege *et al.*, 1998). The number of copies per cell of the mRNA corresponding to each yeast gene is now available on the world wide web (e.g. http://web.wi.mit.edu/young/expression /transcriptome.html). The expression data will assist efforts to determine the functions of genes that may function only under certain conditions as well as improving the understanding of the regulatory mechanisms involved in the cell cycle clock and response to environmental conditions (Holstege *et al.*, 1998). Using statistical analysis it is possible to cluster genes with similar expression profiles, thereby identifying the genes involved in a single pathway (Eisen *et al.*, 1998). Fresh insight is added by the expression data into the evolution of function and expression of duplicated genes (see Chapter 3).

Chapter 2

Evolution of *Saccharomyces cerevisiae* Gene Order After Genome Duplication

2.1 Introduction

Comparison of gene order among genomes can be used for two purposes: inferring the phylogenetic relationships of species, and estimating the number and type of genomic rearrangements that have occurred since two genomes last shared a common ancestor. Three mechanisms of rearrangement are usually considered: inversion, transposition and reciprocal translocation (Nadeau and Taylor, 1984; Sankoff, 1993; Blanchette *et al.*, 1996). Gene order comparisons have been made on sequenced organelle and viral genomes (Palmer *et al.*, 1988; Sankoff *et al.*, 1992; Bafna and Pevzner, 1995; Boore *et al.*, 1995; Hannenhalli *et al.*, 1995), and on more sparsely mapped mammalian and plant nuclear chromosomes (Nadeau and Taylor, 1984; Nadeau, 1991; Bafna and Pevzner, 1995; Paterson *et al.*, 1996).

The genome of yeast (*Saccharomyces cerevisiae*) contains approximately 55 large duplicated chromosomal regions, as described by our laboratory (Wolfe and Shields, 1997), Mewes *et al.* (Mewes *et al.*, 1997)) and Coissac *et al.* (Coissac *et al.*, 1997). Wolfe and Shields (1997) proposed that these duplicated regions ("blocks") are traces of ancient tetraploidy in *S. cerevisiae* that remain detectable after widespread deletion of superfluous duplicate genes, sequence divergence of the remaining duplicates, and successive genomic rearrangements (see Chapter 1). Patterns and characteristics of the duplicated blocks should contain information about the original order of the blocks and the number of rearrangements that have taken place since genome duplication, as well as information about the extent of gene retention versus deletion in the wake of the original genome duplication.

In this study we tried to estimate properties of the yeast genome prior to the whole-genome duplication, and to reconstruct gene order evolution in its aftermath. We assumed the conclusions laid out in the hypothesis of whole genome duplication. Our aim was to estimate the number of reciprocal translocations that occurred, the original number of genes in the genome, and the original order of the duplicated blocks that are now scattered throughout the genome. The methods used are based on comparative genomics, but they differ from previous gene order studies because the two genomes we are comparing are not distinct but are indistinguishable, fragmented, and fused within the same nucleus. Comparison of gene order in duplicated regions within a single genome has been called intra-specific comparative mapping by Nadeau (1991).

We began by making computer simulations to model yeast genome evolution. A genome was duplicated, genes were deleted at random, and reciprocal translocations were made between chromosomes. An algorithm equivalent to that used to find duplicated blocks in the real yeast data (Wolfe and Shields, 1997) was applied to the simulated genomes. These sets of blocks were then analysed in two ways. The first method involved reversing reciprocal translocations to bring the genome back to a symmetrical configuration (as would be expected immediately after genome duplication), and using parsimony to choose between alternative series of translocations. The simulations showed that this method cannot regenerate the original block order nor provide an accurate estimate of the number of translocations when this number is large. The second method involved adjusting the parameters of the simulation (number of duplicate genes retained, and number of translocations) to find the parameter ranges that yielded simulated genomes that were similar to the yeast data in terms of number of blocks, extent of the genome placed inside blocks, and number of duplicate genes identified in blocks. It was possible to find parameters for the simulations that produced duplicated block patterns very similar to those in the real genome.

An analytic approach was developed based on Nadeau and Taylor's method (Nadeau and Taylor, 1984) for estimating the number of rearrangements between the human and mouse genetic maps. This was used to estimate the number of reciprocal translocations in the real yeast data, given the proportion of the genome that is spanned by known duplicated chromosomal blocks. The estimate produced by this approach falls within the range of estimates produced independently by simulation. This in turn permits estimation of a rate of chromosomal translocation in yeast and its comparison with other species. Lastly, we investigated whether genome data from additional species would allow us to determine the original order of genes in yeast.

2.2 Methods

Unit of length

We took the distance between two genes to be the number of genes located between them, rather than the actual distance in kilobases, despite the fact that complete sequence data is available for the genome. The number of genes is a more natural unit when discussing the distribution of reciprocal translocation sites because the probability of a translocation event having been fixed between two points is likely to be influenced most by the amount of noncoding DNA in the interval, which is expected to be correlated more strongly with the number of genes than with the physical separation of the points along the chromosome. This unit is also more natural when discussing the distribution of duplicates that have been retained after diploidisation since the probability of deletion of a gene should not be strongly influenced by its physical size.

2.2.1 Simulations and identification of duplicated segments

In the simulations we assumed that there were no inversions, transpositions, or any other type of rearrangement except reciprocal translocations; that translocations occur at random intergenic locations; that gene deletion occurs by random deletion of single genes; that sequence similarity is only detected between genes duplicated during the tetraploidisation; and that natural selection does not impose any functional constraints on gene order. In our model an original genome with eight chromosomes was duplicated and genes were deleted randomly until the current configuration (5790 genes on 16 chromosomes) remained. This is a rough approximation of the process associated with genome duplication and subsequent diploidisation (Ohno, 1970). Reciprocal translocations were then made between randomly chosen points in the genome and blocks of duplicated genes were identified using criteria similar to the original study (Wolfe and Shields, 1997). It was not difficult to fully automate the block-finding process because all the duplicate genes in the simulated data resulted from genome duplication (there were no multigene families) and, as a result, blocks were easily identifiable by a simple program. The blocks produced were not very sensitive to the value chosen for the maximum distance between intervening genes once this was greater than about 20 genes. A maximum distance of 45 was used in practice. A minimum of three retained duplicates was required for the identification of a block. The program used to locate the blocks in the simulated data was adapted for use on the real data to permit direct comparison of the results with those in Wolfe & Shields (1997). Subtelomeric regions were ignored altogether because the level of noise was too great for the identification of blocks within these regions by this simple method. The threshold for identifying duplicate genes in the real data was a BLASTP score of 200. The resulting blocks were almost identical to the blocks reported by Wolfe and Shields (1997), which were identified using a criterion of three duplicate genes per 50 kilobases.

2.2.2 Transformation of the genome to a symmetrical configuration

By "symmetrical configuration" we mean a configuration of blocks in which the chromosomes can be grouped into two identical sets. The computer program written to transform the arrays of blocks to a symmetrical configuration is based on a simple search method in which a symmetry improving operation is chosen at each step. It does not find the shortest or most parsimonious path by which a symmetrical configuration can be achieved. Each point in Fig. 2.1 was constructed by choosing the shortest of just 10 such paths to symmetry. It is unnecessary to search further since from Fig. 2.1 we can see that we are already achieving symmetry in fewer steps than were involved in the simulation. The most parsimonious path tells us little about the actual evolutionary path taken.

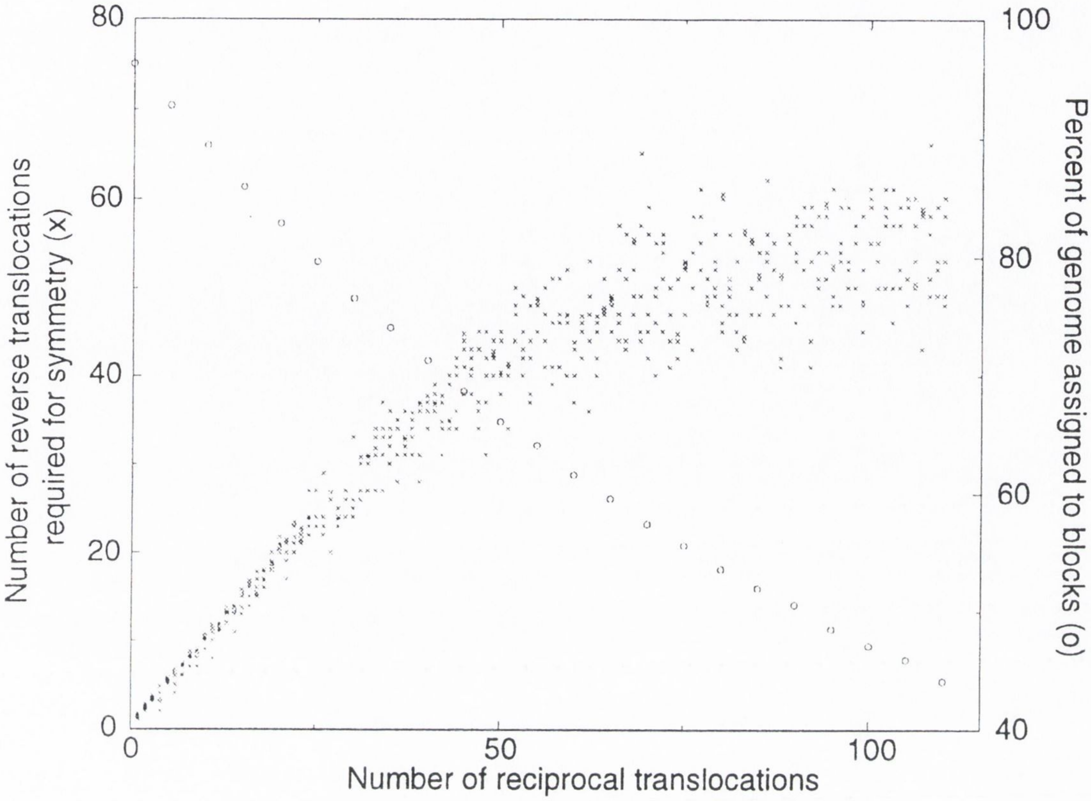


Figure 2.1 Simulations of rearranging a duplicated yeast genome and then reconstructing its original structure. The number of steps taken by our programme to bring about symmetry in a configuration of blocks is plotted against the number of reciprocal translocations in the simulation which brought about the original block configuration. Each point represents the shortest of 10 simulations of a 5790-gene genome with 446 pairs of retained paralogues. Five runs were carried out for each value on the X-axis. Circles indicate the average fraction of the genome that could be assigned to duplicated blocks in simulations (using a minimum of three duplicated genes per block); this fraction declines as more reciprocal translocations are made.

2.3 Results

2.3.1 Making the genome symmetrical by reversing reciprocal translocations

Inspection of the map of duplicated regions (Wolfe and Shields, 1997) shows three points where the symmetry of the map could be increased by reversing apparent reciprocal translocations. These points involve duplicated chromosomal blocks 14/23/37/50, 38/39/50/52, and 5/6/32/33 (Wolfe and Shields, 1997). In each of these cases four blocks can be reduced to two larger blocks by undoing a translocation. This suggests that it might be possible to "unscramble" the yeast genome by making a series of reversals of reciprocal translocations until a completely symmetrical genome remains. We speculated that the shortest series of reverse translocations leading to symmetry might correspond to the evolutionary path taken by the yeast genome after its duplication, and investigated this by computer simulation. The problem of finding the minimal number of translocations to transform the gene order of one genome into another has been studied extensively (Sankoff *et al.*, 1992; Bafna and Pevzner, 1995, see also Chapter 1; Hannenhalli *et al.*, 1995). Here, rather than calculating the translocation distance between two genomes, we wish to examine sets of translocations that relocate the paralogous blocks within a single genome so that the chromosomes form two identical sets.

Genomes were simulated undergoing duplication, gene deletion and multiple reciprocal translocations. Duplicated blocks (containing three or more duplicate genes) in the simulated genomes were then identified, and a search was made for series of reciprocal translocations that would rearrange these blocks into a symmetrical configuration. A search routine in which translocations were chosen by a hill-climbing approach (continually increasing the symmetry of the genome) was developed. In simulations with 20 or fewer translocations this search method usually returned the blocks to a perfectly symmetrical configuration in the same number of steps as were used to bring about the configuration (Fig. 2.1). As the number of translocations in the simulation is increased the number of steps required to bring about symmetry levels off and begins to fluctuate widely. The fraction of the genome that can be placed in duplicated blocks decreases as the number of translocations increases, and many smaller blocks are not detected (Fig. 2.1). The effect of failing to detect some blocks (or deleting some blocks from a data set) is to reduce the number of steps required to return the remaining blocks to a symmetric configuration (Ferretti *et al.*, 1996). It then becomes possible to return to a symmetrical genome in fewer steps than the original number of translocations.

The shortest solution we found for the real yeast data (in a non-exhaustive search) returned the blocks to a perfectly symmetrical configuration in 41 steps (after three initial inversions to correct the five blocks whose orientation with respect to the centromere is opposite to that of their copies (Wolfe and Shields, 1997) and without associating duplicated chromosome arms). Forty-one reciprocal translocations would give rise to $2R + C = 90$ pairs of duplicated chromosomal regions, where R is the number of reciprocal translocations and C is the original pre-duplication number of chromosomes (eight). Since only 55 duplicated blocks have been discovered and since only half of the genome is placed in blocks (Wolfe and Shields, 1997) we can be confident that there are many smaller duplicated regions that have not been discovered. Because the effect of deleting blocks only decreases the number of steps required to return to a symmetrical configuration we can deduce that it is likely that there have been more than 41 reciprocal translocations after genome duplication.

2.3.2 Numerical estimate of the number of reciprocal translocations after duplication

Even when the number of reciprocal translocations in simulations is so large that saturation has been reached for the number of reverse steps required to achieve symmetry (Fig. 2.1), the fraction of the genome that is assigned to duplicated blocks continues to decrease almost linearly. This suggests that approaches based on the latter measure might be more effective ways to estimate the number of translocations than the reverse-translocation approach taken above, when the number of translocations is large.

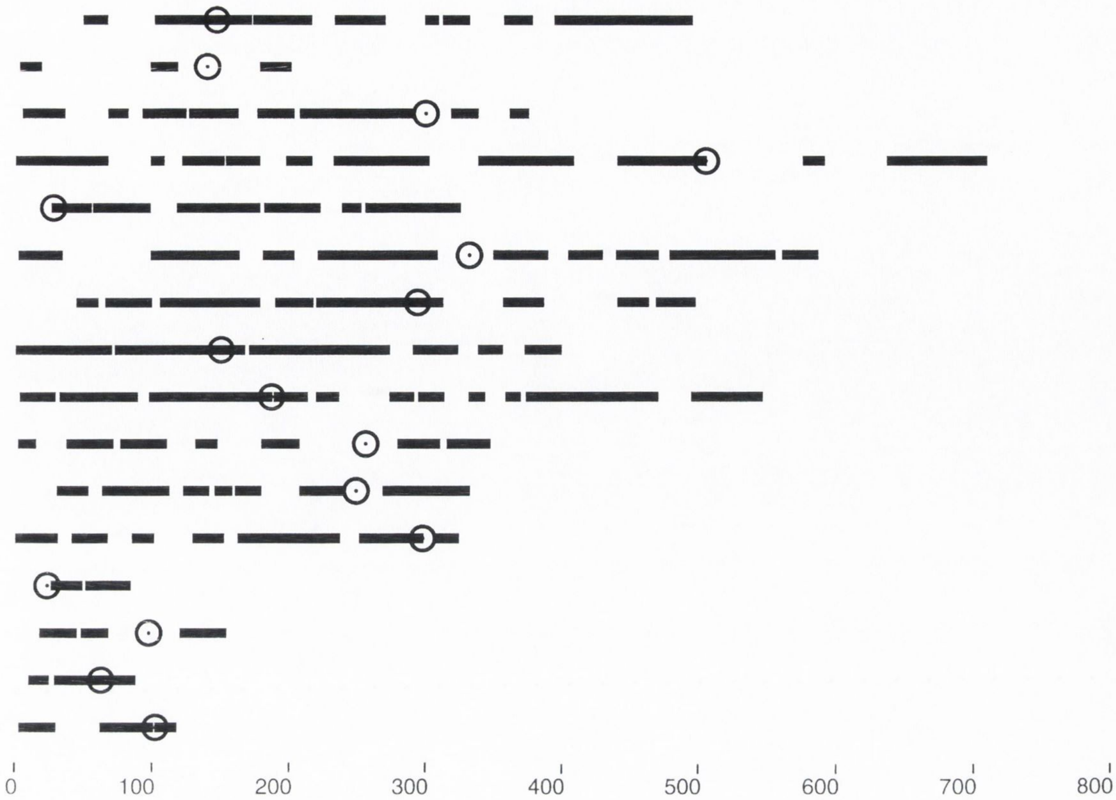


Figure 2.2 The duplicated chromosomal regions in a simulated genome with 446 pairs of paralogues retained and 75 reciprocal translocations since duplication. These simulations gave rise to patterns and densities of duplicated blocks that are similar to those mapped in the real data (Wolfe and Shields, 1997). Circles indicate centromeres; bars show duplicated blocks. The scale indicates numbers of genes.

We repeated the simulations varying two parameters in order to reproduce the observed state of the yeast genome. These simulations used reciprocal translocation as the sole mechanism of chromosomal rearrangement, and the block layouts they produced were similar to the structure of the real genome (Fig. 2.2). The parameters varied were the number of reciprocal translocations fixed since whole genome duplication, and the number of genes retained in duplicate (paralogues*) after genome duplication (Fig. 2.3). We do not have an exact value for the number of paralogues in the whole (real) yeast genome because similar genes can be identified as paralogues only by their occurrence in the correct position within a regional chromosomal duplication and we do not have a duplication map for all parts of the genome. Similarly, in the simulated genomes, the number of pairs of paralogues recovered in blocks is less than the actual number of paralogues present (Fig. 2.3).

Number (%) of duplicate genes retained	Number of reciprocal translocations									
	55	60	65	70	75	80	85	90	95	100
398 (7.4%)	54	56	57	58	58	59	58	59	59	59
	0.60	0.58	0.55	0.54	0.51	0.49	0.48	0.46	0.44	0.43
	343	337	330	324	317	309	304	298	291	286
414 (7.7%)	56	57	58	59	60	61	61	61	61	61
	0.61	0.59	0.57	0.55	0.52	0.51	0.49	0.47	0.46	0.44
	360	353	346	340	332	327	320	313	309	302
430 (8.0%)	57	59	60	61	62	63	63	63	63	63
	0.63	0.60	0.58	0.56	0.54	0.52	0.50	0.49	0.47	0.46
	377	370	363	356	350	343	334	329	324	316
446 (8.3%)	58	60	62	63	64	65	65	65	66	65
	0.64	0.62	0.60	0.57	0.56	0.54	0.52	0.51	0.48	0.47
	393	386	380	374	367	360	353	347	342	334
462 (8.7%)	60	62	63	64	65	66	67	68	67	68
	0.65	0.63	0.61	0.59	0.57	0.55	0.53	0.52	0.50	0.48
	411	404	397	391	384	378	370	364	356	351

Figure 2.3 Genome structure simulations in which the number of reciprocal translocations and the number of retained paralogues were varied. Each cell shows values for the number of blocks discovered in the genome (top), the percentage of the genome that is in blocks (middle) and the number of pairs of duplicated genes discovered within blocks (bottom). Mean values among 200

[*] We use the word "paralogues" here specifically to refer to duplicate genes produced by whole-genome duplication, and not to any other sort of paralogues (Fitch, 1970). Spring (1997) proposed "tetralogues" as a name for the four-member gene sets resulting from putative ancient octoploidy of vertebrate genomes, and "homeologues" has also been used (Morizot, 1990).

replicates are shown. The standard error on the number of blocks was = 4; on the fraction of genome in blocks was = 0.03; on the number of paralogues identified was = 9. Shaded cells are within two standard errors of the value for the yeast data. Fifty five blocks covering 51% of the genome and containing 365 pairs of paralogues have been mapped in the yeast data (Wolfe and Shields, 1997).

Each cell of Fig. 2.3 shows characteristics of the genomes produced from 200 simulations for a given combination of input parameters. The values, in the yeast data at the time of this study, of the three genome characteristics shown in Fig. 2.3 were: 55 blocks, 0.51 of the genome in blocks, and 365 pairs of paralogues in blocks. Updates of the map of duplicated regions have not altered these values greatly. Only input parameters in the region of 8% of duplicate genes retained (400 - 450 pairs) and 70 - 100 translocations give results similar to the real data.

2.3.4 Analytic estimate of the number of reciprocal translocations since duplication

It is also possible to convert the fraction of the genome in blocks (Fig. 2.1) into an estimate of the number of translocations, without computer simulation, using an analytic method analogous to the approach of Nadeau and Taylor (1984, and see also Chapter 1). In their approach to the similar problem of determining the combined rate of rearrangements in mouse and human, Nadeau and Taylor examined the average lengths of conserved linkage groups. In yeast, because we have complete sequence information, we can use the fraction of the genome that is spanned by paralogous chromosomal blocks instead.

We wish to estimate the underlying number of chromosomal regions ("Segments") that were demarcated by reciprocal translocations, rather than the number of duplicated regions that can now be identified ("Blocks"). Since we have assumed that paralogues are scattered randomly throughout the genome the number contained in a Segment of length x is described by a Poisson distribution. The probability that a Segment of length x contains three, or more, paralogues and so would be reported as a duplicated chromosomal Block is

$$\sum_{k=3}^{\infty} \frac{(Dx)^k}{k!} e^{-Dx}$$
$$= 1 - e^{-Dx} - Dxe^{-Dx} - \frac{(Dx)^2}{2} e^{-Dx}$$

Where D is the density of paralogues in the whole genome. We do not know the value of D exactly since only the paralogues that occur in the correct position within a duplicated chromosomal Block can be identified as paralogues. We have a lower limit on D since we know the number of paralogues that are contained in Blocks. We can estimate the correct value of D in two ways. We can use the simulations (Fig. 2.3) and note that there is a relatively small window of densities for which our parameter values come close to modelling the real data. The simulations suggest a density of about 0.15 paralogues per gene (*i.e.*, 2 x 7.5%), but this method has the undesirable effect of linking the analytic and simulative methods of calculating the result. To avoid this we can examine the number of cases where two genes that are homologues (i.e., a significant "simple" BLASTP pair as defined in Wolfe and Shields, 1997) are both located anywhere in the half of the genome that has been mapped into Blocks, but their locations are such that they are not considered to be paralogues. The number of such internal duplicates should be approximately the same as the number of strictly external non-paralog hits (*i.e.*, with both genes occurring outside Blocks) since the areas inside Blocks and outside Blocks are approximately the same in extent. Any excess in hits outside Blocks represents likely paralogues that have not been identified because they are not contained in Blocks of three or more. This method yields a density of 0.155 paralogues per gene.

Since the probability of having a region of length x with no translocation point is $e^{-x/L}$ where L is the average length of all Segments, the probability density of discovered Segment lengths is

$$\frac{N}{L}\left(1-e^{-Dx}-Dxe^{-Dx}-\frac{(Dx)^2}{2}e^{-Dx}\right)e^{-x/L}$$

The constant $\frac{N}{L}$ is introduced for normalisation, where N is the total number of Segments.

The total amount of the genome covered by discovered Segments, F, is then expected to be

$$F=\int_0^G\frac{N}{L}\left(1-e^{-Dx}-Dxe^{-Dx}-\frac{(Dx)^2}{2}e^{-Dx}\right)e^{-x/L}xdx$$

where G is the total length of the genome.

Since if L is small compared to G the integral evaluated at G approaches 0 we evaluate the integral at 0 only:

$$F = N\left(L - \frac{1}{L\left(D + L^{-1}\right)^{2}} - \frac{2D}{L\left(D + L^{-1}\right)^{3}} - \frac{3D^{2}}{L\left(D + L^{-1}\right)^{4}} \right)$$

$$F = \left(\frac{5790}{L}\right)\left(L - \frac{1}{L\left(D + L^{-1}\right)^{2}} - \frac{2D}{L\left(D + L^{-1}\right)^{3}} - \frac{3D^{2}}{L\left(D + L^{-1}\right)^{4}} \right) \qquad \textbf{[1]}$$

which is the amount of the genome covered by Segments that contain the discovered Blocks (5790 is the number of genes in the genome).

If m is the expected length of a Segment that contains n paralogues separated by a total distance r then $m = r(n+1)/(n-1)$ (as in Nadeau and Taylor, 1984). We can modify our figure for the fraction of the genome covered by Blocks to approximate the fraction of the genome covered by those Segments that contain the discovered Blocks. We calculate the expected length of each Segment from the range of the paralogues it contains, and sum the Segments. The fraction of the genome covered by Blocks is 0.496, and the expected value for the amount of the genome covered by the Segments containing these Blocks is 0.686 (not including some telomeric genes which could not be confidently placed in Blocks or outside the blocked region due to a high level of intertelomeric similarity). The value of L required to give this result in equation **[1]** is 16.45 genes. This gives $N = 5790 / 16.45 = 352$ Segments (organised as 176 pairs). From $2R + C =$ 176 pairs of Segments, and $C = 8$ chromosomes, the number of reciprocal translocations (R) is approximately 84. In simulations the standard deviation of the fraction of the genome under Blocks was = 0.03. This gives us an estimate of 84 ± 15 (for approximately two standard deviations) for the number of reciprocal translocations that have been fixed in yeast since genome duplication.

We can make a prediction of the number of additional Blocks that we would expect to find if we relaxed the block-finding criteria to include Segments containing only two paralogues. The probability of a Segment of length x containing y paralogues is $\frac{(Dx)^y}{y!}e^{-Dx}$. The expected number of Segments of length x is $\frac{1}{L}e^{-x/L}N$. Therefore the expected number of Segments containing y paralogues is

$$\int_0^c \frac{N}{L}\left(\frac{(Dx)^y}{y!}e^{-Dx}\right)e^{-x/L}dx$$
$$= \frac{D^y}{L(D+\frac{1}{L})^{(y+1)}}$$

On the basis of a model with 446 pairs of retained duplicates and 84 reciprocal translocations the expected value of the number of Segments containing two paralogues is 26 ± 5 (The error was calculated from simulations; Table 2.1). The number of additional two-member Blocks found in the real data is 34. This high value leads us to suspect that it could be difficult to distinguish between genuine small Blocks and statistical noise. The predicted number of one-member blocks is 36 (Table 2.1).

Table 2.1

Number of Paralogues (*P*)	Number of blocks having *P* paralogues Theoretical prediction	Simulation	Real Data
0	49.6	49.4	n/a
1	35.6	35.9 ± 5.9	n/a
2	25.6	26.2 ± 5.2	n/a
3	18.4	18.5 ± 4.3	10
4	13.2	13.2 ± 3.5	10
5	9.5	9.6 ± 3.1	6
6	6.8	6.8 ± 2.5	4
7	4.9	4.8 ± 2.1	6
8	3.5	3.3 ± 1.7	6
9	2.5	2.4 ± 1.4	1
10	1.8	1.8 ± 1.2	4
11	1.3	1.3 ± 1.1	2
12	0.9	0.9 ± 1.0	1
13	0.7	0.6 ± 0.7	4
14	0.5	0.4 ± 0.6	0
15	0.3	0.3 ± 0.6	0
16 - 20	0.8	0.6 ± 2.6	1
21 - 25	0.1	0.0	0

Table 2.1 Theoretical predictions, results of simulations, and values from the real data, of the number of blocks containing a given number of detectable paralogues. The simulation results

(mean ± 2 s.d. from 2000 replicates) are from a model with 446 retained paralogues and 84 reciprocal translocations.

Possible Clustering of Duplicates

The approach of trying to reverse reciprocal translocations produced symmetrical arrangements of the 55 blocks in 41 steps (after three initial inversions). In Fig. 2.1, 41 reverse steps is on the lower extreme of the scatter when the graph has become saturated, which is the region of interest because the other methods show that there have been approximately 84 reciprocal translocations since genome duplication. However, in other simulations in which the number of blocks was fixed from the outset at 55 (results not shown), 41 was close to the mean number of operations required for the return to a symmetrical configuration. This discrepancy arises because the simulations that produced the closest match to the real values of genome parameters (Fig. 2.3) tended to have slightly larger numbers of duplicated blocks than were discovered in the yeast data. In many cases this difference was only about 1 - 2 standard deviations and may not be systematic but a non-random distribution of paralogues could also explain the shortage of discovered blocks (Sankoff *et al.*, 1997b). If paralogues tend to be located in clusters more Segments than might be expected could fail to contain three paralogues, the requirement for identification. It was not possible to test for clustering of paralogues over the whole genome because they cannot be identified outside blocks. No clustering was discovered when all duplicates were taken into account. Some clustering of duplicates may occur for functional reasons, for example the frequent duplication of pairs of adjacent ribosomal protein genes transcribed divergently from a shared promoter. If there is an excess of undiscovered blocks due to clustering of retained duplicates we would expect this to affect the numbers of smaller blocks. Using the analytic method to predict the number of blocks of a given size that we would expect to find in the data, we find that the expectation is consistently higher than the actual value for blocks containing less than 7 paralogues (Table 2.1). This is what we would expect if clustering of duplicates is preventing the discovery of some smaller blocks.

2.3.5 Establishing the original gene order

We considered the possibility that the approach of finding the most parsimonious path might reveal some aspects of the original gene order, even though the number of steps in this series is too few, as explained above. However, there are a great many equally parsimonious paths returning the data to a symmetrical configuration in the shortest number of steps. This degeneracy is intrinsic. The two operations in Fig. 2.4 have the same effect on symmetry. Since reverse translocations are commutative (except possibly those involving more than one operation on a single chromosome arm) whole sets of stepwise equivalent operations can give rise to vastly different configurations of blocks in the same number of steps. For example, almost any series of 20 reverse translocations, each of which maximally improves the symmetry at each step, could be used to return a simulation involving 20 reciprocal translocations to symmetry. Since each of these translocations has an alternative that improves the symmetry to exactly the same degree (Fig. 2.4) we have 2^{20} sets of possible final block orders brought about by 20 operations on the data. We cannot distinguish between these block orders without further information.

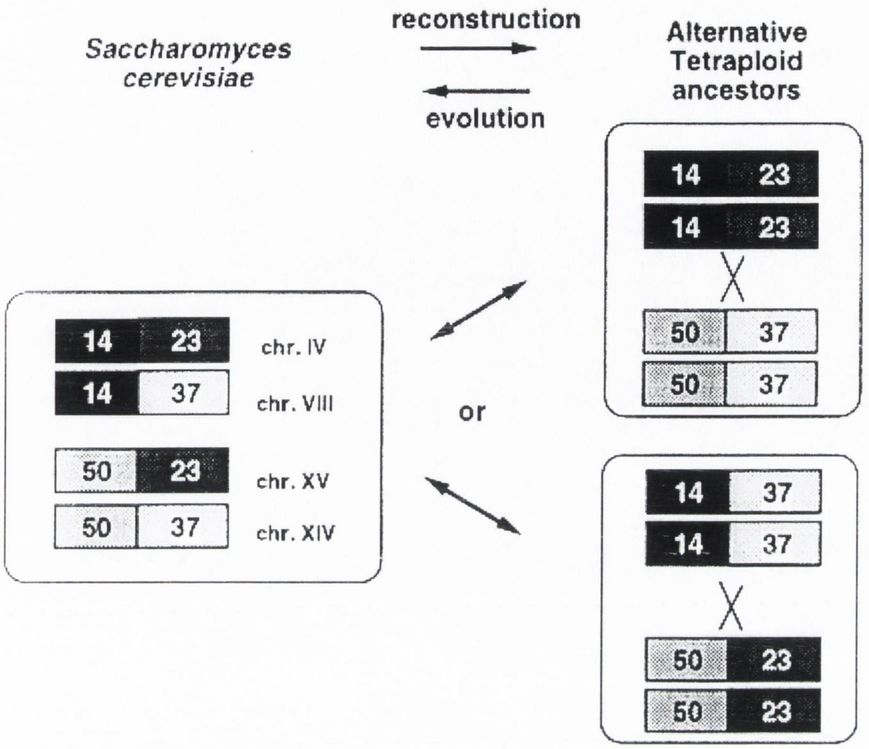


Figure 2.4 An example of the two indistinguishable solutions to the problem of reversing a single reciprocal translocation in *Saccharomyces cerevisiae*. The translocation involves duplicated chromosomal blocks 14, 23, 50 and 37 (Wolfe and Shields, 1997).

This degeneracy, in the case of yeast, could in principle be resolved by the inclusion of information from one other species that diverged from *S. cerevisiae* at around the time of genome duplication. We find in simulations that we can completely reconstruct the order of the blocks in a duplicated yeast-like genome, using a second species that diverged from it immediately prior to duplication as an outgroup, if fewer than 40 reciprocal translocations have been fixed in the duplicated genome (results not shown). Above this number of translocations the solution begins to decay, because if both copies of a block have been shifted by reciprocal translocation we no longer have any information from the duplicated genome about their original locations. In simulations with a realistic number (75) of reciprocal translocations approximately two thirds of the duplicated chromosomal blocks that contained sufficient numbers of paralogues for identification could be placed in their original order using a species that diverged shortly before genome duplication (results not shown).

Similar gene order reconstructions are possible even without genome duplication but sequence information from a third species is required. Using sequence information from three species that diverged at around the same time to determine ancestral gene order is more reliable than the method described above because in this case all the genes in the genome, not just the paralogues making up blocks, can be used to infer the original gene order.

2.4 Discussion

Our simulations involved several assumptions. Chromosomal inversions and gene transposition were ignored as possible mechanisms of gene order change. This assumption is reasonable because inversions and transpositions on a scale large enough to produce blocks containing at least three paralogues are evidently uncommon (Wolfe and Shields, 1997). We have assumed also that reciprocal translocations are evenly distributed even though this is open to debate (Lundin, 1993; Sankoff and Ferretti, 1996). If reciprocal translocations do not take place at random intergenic sites our result for the number of reciprocal translocations since genome duplication is likely to be an underestimate. Our estimates of the proportion of genes retained in duplicate, and the number of genes in the original genome, are sensitive to the sequence similarity threshold used in the analysis (BLASTP > 200, which is quite stringent). The choice of similarity threshold should not, however, affect the estimate of the number of translocations; detecting additional paralogues in yeast is analogous to mapping additional genes in humans and mice, and the inclusion of these extra pieces of data should not substantially alter the estimates of the extent of rearrangement (Nadeau and Taylor, 1984; Copeland *et al.*, 1993).

The 70 - 100 reciprocal translocations estimated to have occurred would have produced 148 - 208 paired duplicated chromosomal blocks if each breakpoint was unique. One-third of these (55 blocks) were large enough to be detected in the original study (Wolfe and Shields, 1997), and the remainder must correspond to blocks containing two, one or zero duplicated genes. We estimate that 36 one-member blocks and 26 two-member blocks exist (Table 2.1), but it will be difficult to identify them because of statistical noise. The map of duplicated regions has been revised using Smith-Waterman sequence similarity cut-offs and with the addition of tRNA genes (Seoighe and Wolfe, 1999a). The revised map was based on parameters which were optimised on the yeast data using the assumption of genome duplication followed by reciprocal translocation but the results have not altered significantly. The most significant change in the map has been the classification of duplicated regions as 'probable' and 'possible' paralogous regions with the inclusion of many smaller and less certain regions in the latter category (see Chapter 3). The analysis presented here has not been updated to reflect the updated map of the yeast genome, partly because of the difficulty of adapting the analysis to take account of 'probable' and 'possible' paralogous regions. In the previous analysis Wolfe and Shields (1997) made the inaccurate assumption that, because they discovered about 400 paralogues in half the genome, there would be 800 in the whole genome. In fact, the block-finding approach preferentially finds the most duplicate-rich regions of the genome. As shown in Fig. 2.3, we now envisage that the ancestral yeast genome had about 5350 - 5400 protein-coding genes, not 5000 (Wolfe and Shields, 1997). We have identified only 1/3 of the blocks but these contain about 80% of the paralogues (Fig. 2.3).

The most effective way to study how the yeast genome has evolved after its duplication would be to sequence the genome of a second, closely related, ascomycete species. A genome sequence from a second species would reveal most of the original order of the duplicated yeast blocks. It should also enable us to identify the 49 anticipated "zero-membered" blocks (Table 2.1). These are Segments of the yeast genome that are "sisters" derived from genome duplication, but where no paralogous genes have been retained. Genome sequencing projects, or "single-pass" sequencing surveys, are in progress for *Schizosaccharomyces pombe*, *Candida albicans*, *Kluyveromyces lactis*, *Ashbya gossypii* (Altmann-Jöhl and Philippsen, 1996) and several species in the *Saccharomyces* group (Genoscope). The extent of gene order conservation between yeast and either *Sch. pombe* or *C. albicans* is probably too low to permit reconstruction of much of the original yeast genome (Keogh *et al.*, 1998), but the others should be useful. Because the *Saccharomyces sensu stricto* species share the genome duplication (Keogh *et al.*, 1998) it may be possible to determine their phylogenetic relationships using gene order information alone. For example (see Fig. 2.4) if we can determine the ancestral order of blocks 14, 23, 50 and 37 using, say, information from *K. lactis* we can identify which pair of adjacent blocks represents the derived state and then search the other *sensu stricto* yeasts for synapomorphy.

Our estimate that 70 - 100 reciprocal translocations have occurred in roughly 100 Myr (Wolfe and Shields, 1997) since yeast genome duplication results in an estimate of the rate of genomic rearrangement in yeast that is quite similar to the rate in human/mouse comparisons (about 100 - 180 rearrangements, also in approximately 100 Myr ; Nadeau and Taylor, 1984; Copeland *et al.*, 1993; DeBry and Seldin, 1996; Sankoff *et al.*, 1997a). This is surprising given their very different genome sizes (12 Mb in yeast; 3000 Mb in human) and rates of homologous recombination (1 centimorgan corresponds to ~ 3 kb in yeast but ~ 1 Mb in human). Since the two organisms have similar genome sizes in centimorgans this suggests that the ratio (expressed in terms of rearrangements per centimorgan per year) between rates of translocation and homologous recombination may be similar in the two taxa. Estimates of rates of genomic rearrangement in plants indicate that they too may be similar to mammals (Paterson *et al.*, 1996; Wolfe, 1996), but whether there is really a molecular clock for chromosomal rearrangement as proposed by (Paterson *et al.*, 1996) will not be clear without better maps for many taxa.

Chapter 3

Updated Map of Duplicated Regions in the Yeast Genome

3.1 Introduction

This section describes a computational approach that was developed to extend and update the map of duplicated chromosomal segments in the *Saccharomyces cerevisiae* genome, originally published by Wolfe and Shields in 1997. The task of assigning chromosomal regions to duplicated blocks was largely automated and based on parameters that were applied consistently throughout the genome. Most of the contents of this chapter has been published in *Gene* (Seoighe and Wolfe, 1999a).

The genome of the yeast *Saccharomyces cerevisiae* contains many large paired chromosomal regions, consisting of duplicated gene pairs arranged in the same order on two chromosomes, interspersed with many unique genes (e.g. Lalo *et al.*, 1993; Goffeau *et al.*, 1996; Coissac *et al.*, 1997; Mewes *et al.*, 1997). Wolfe and Shields proposed in 1997 that these regions are the result of a single, ancient, duplication of the entire genome (which was subsequently fragmented by reciprocal translocations among chromosomes) rather than numerous successive independent duplication events (Wolfe and Shields, 1997, and see Chapter 1). Our model of yeast chromosome evolution is based on the hypothesis that the entire genome was duplicated, increasing the number of genes to 200% of its original value, but then that numerous deletions of redundant duplicate copies of genes reduced this figure to 108% (i.e. 2 X 8% in pairs and 92% single copy).

The parameters used to identify duplicated chromosomal regions were optimised such as to maximise the amount of the genome placed into paired regions, under the assumption that the hypothesis that the entire genome was duplicated in a single event is correct. The approach was to construct a core map of 'probable' sister regions (that satisfy quite strict criteria) and then to overlay this map with 'possible' regions that may also be sisters, but for which the evidence is less convincing. In doing this the approach has been more methodical than that taken in the earlier study of Wolfe and Shields. The core of the new map, with 52 pairs of regions containing three or more duplicated genes, is largely unchanged from the original map. 39 tRNA gene pairs and one snRNA pair were added. To find additional pairs of genes that may have been formed by whole genome duplication, we searched through the parts of the genome that are not covered by this map, looking for putative duplicated chromosomal regions containing only two duplicate genes instead of three (the criterion used for the original map), or having lower-scoring gene pairs. This approach identified a further 32 candidate paired regions, bringing the total number of protein-coding genes on the duplicated map to 905 (16% of the proteome). Results from the updated map also suggest that a second copy of the ribosomal DNA array has been deleted from chromosome IV. The current analysis is based on the sensitive Smith-Waterman search method instead of BLAST and the available gene order data from *K. lactis* have also been integrated with the map of *S. cerevisiae* duplications.

3.2 Data and Methods

The sequences used were the same 5790 proteins as in Wolfe and Shields (1997) and are available on our website (http://acer.gen.tcd.ie/~khwolfe/yeast). Subtelomeric repeat regions were excluded as in Chapter 3. Gene names were updated to those in version 7.1 of the Yeast Protein Database (YPD; http://www.proteome.com). The tRNA and snRNA genes analysed were those listed by the Saccharomyces Genome Database (SGD: http://genome-www.stanford.edu). All-against-all Smith-Waterman searches (Smith and Waterman, 1981) were done using the SSEARCH program in the FASTA package (Pearson and Lipman, 1988), using the BLOSUM62 matrix (Henikoff and Henikoff, 1992) and the **seg** filter (Wootton and Federhen, 1996). Computation time for these searches on a high performance parallel computer was provided by Compaq Computer Corporation. Duplicated chromosomal regions were identified by analysing these results using computer programs written in the C and Perl programming languages. The map of Fig. 3.2 was produced by a program written in Microsoft Visual Basic.

3.2.1 Optimising the parameters for defining duplicated chromosomal blocks

In our previous version of the map of sister chromosomal regions, pairs of homologues with BLASTP scores (Altschul *et al.*, 1990) in excess of 200 were included. The Smith-Waterman algorithm (Smith and Waterman, 1981) has been used instead of BLASTP for the revised map. Much work has been done on the relative merits of different algorithms and techniques for searching databases to find homologues of a query sequence. Smith-Waterman is generally accepted as the best method currently available in terms of sensitivity and specificity (Shpaer *et al.*, 1996), but requires much more computer time than does BLAST. We used the SSEARCH Smith Waterman program (Pearson and Lipman, 1988) with log-length normalisation following Shpaer *et al.* (1996). Raw scores from the Smith-Waterman algorithm are dependent upon the lengths of the sequences being compared, but dividing by the product of the logarithms of the sequence lengths removes this dependence and greatly improves selectivity.

When searching for sister chromosomal regions we are not interested in all duplicated proteins, but only those proteins that were duplicated as part of the whole-genome duplication. Paralogues that existed before that time, or that were formed more recently, are of no use in determining the map of sister regions. We did not consider it feasible to use either a molecular clock approach or a phylogenetic approach (Yuan *et al.*, 1998) to identify the set of paralogues that were duplicated simultaneously, because (i) there are no closely related outgroup sequences for many of the yeast gene pairs, and (ii) molecular clock analysis of a small number of tetraploidy-derived paralogues yielded a considerable range of date estimates, possibly due to gene conversion (Wolfe and Shields, 1997; Skrabanek and Wolfe, 1999).

Instead, we followed the logic that under the hypothesis of genome duplication, followed predominantly by reciprocal translocation, there should be no overlapping blocks (sister chromosomal regions). The fraction of the genome placed in overlapping blocks (with each block containing three or more duplicated genes, as in Wolfe and Shields (1997) was plotted by a computer programme for different cut-off values of similarity score (Fig. 3.1a). Very high cut-offs did not yield any duplicated blocks, whereas very low cut-offs generated many overlapping blocks. A cut-off of 17.5 (log-length normalised Smith-Waterman score) was chosen as the lowest similarity score that did not produce overlapping blocks.

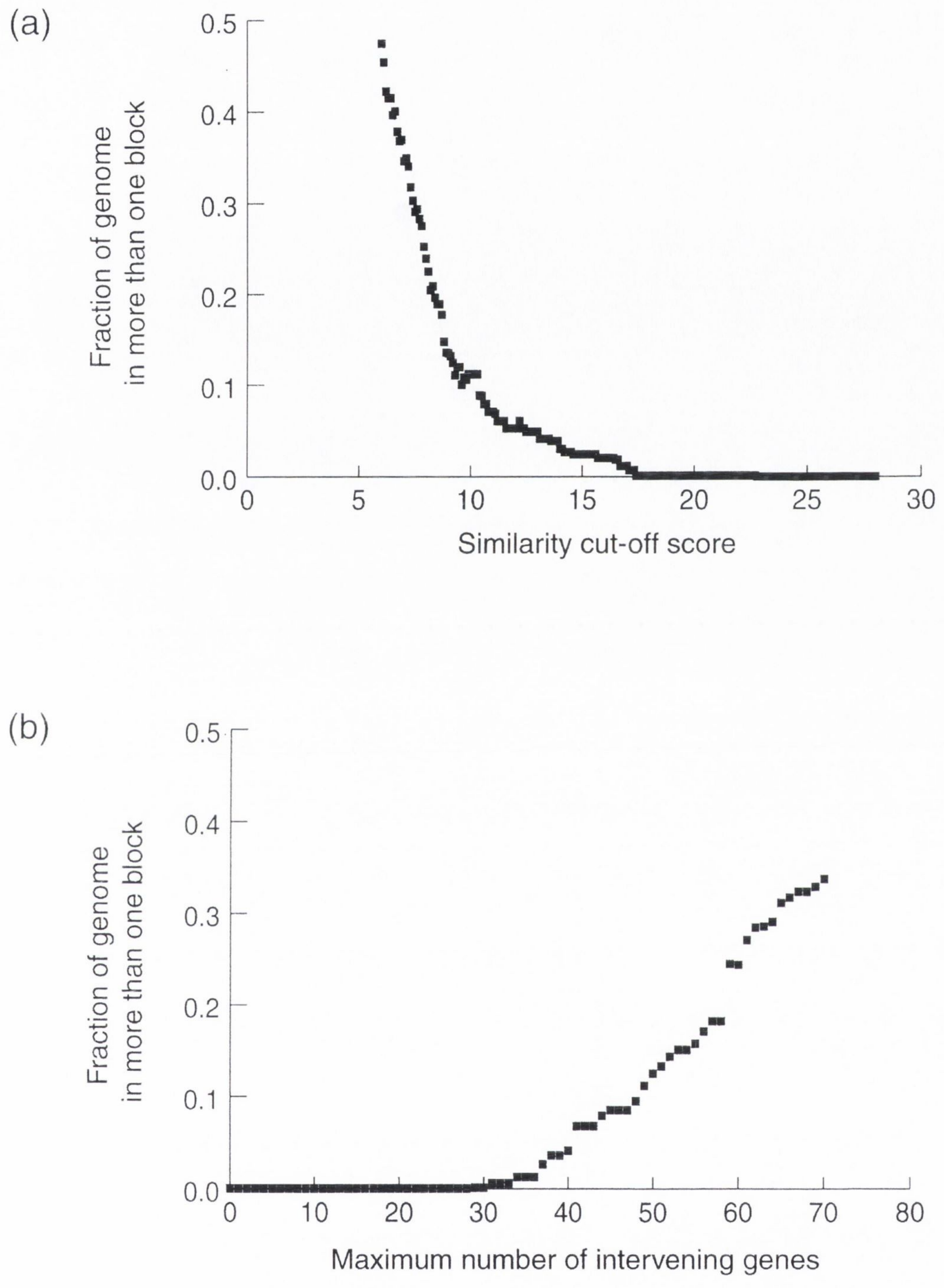


Figure 3.1 Optimisation of parameters used to construct the duplication map. (a) Fraction of the yeast genome simultaneously paired with more than one sister block, plotted as a function of the sequence similarity cut-off score used to define paralogues. (b) Fraction of the yeast genome simultaneously paired with more than one sister block, as a function of the maximum physical distance allowed (number of intervening non-duplicated genes) between successive paralogues making up a block.

Wolfe and Shields (1997) used an arbitrary limit of 50 kilobases (kb) as the maximum permitted gap between duplicated genes making up a block; this corresponds to approximately 25 genes. In Fig 3.1b the fraction of the genome assigned to overlapping blocks is plotted against the maximum number of intervening genes allowed between neighbouring paralogues. The automatic method of constructing blocks begins to report overlapping blocks at a cut off distance of about 30 genes. From this result we chose 30 genes as the cut-off for the maximum number of genes between consecutive paralogues in a block.

3.3 Results

3.3.1 The updated map of 'sister' regions

The updated map (Fig. 3.2) is organised into two levels: a core framework of duplicated chromosomal blocks that are 'probable' products of genome duplication, and a second level of 'possible' paralogues and regions for which the evidence is weaker. The map was constructed by first identifying the 'probable' regions using stringent criteria, and then relaxing the criteria both to add extra 'possible' genes to the blocks already identified, and to find additional 'possible' blocks. These 'possible' genes and blocks were only added to the map where they were not in conflict with the 'probable' framework. The 'possible' genes shown in Fig. 3.2 are thus a selective representation of the data, but one that maximises the biological information that can be extracted from the map when the genome duplication hypothesis is assumed to be correct.

Figure 3.2 (overleaf) Updated map of duplicated regions in the yeast genome. A web version of this map with links to information about each gene is at http://biotech.bio.tcd.ie/~ferdia/Yeast. Coloured rectangles adjacent to the vertical chromosome lines are 'probable' duplicated regions associated with genome duplication, containing three or more duplicated genes. Gene names written to the right of the chromosomes indicate the genes making up these 'probable' blocks. Coloured rectangles displaced to the left are 'possible' additional duplicated regions. Large numerals (1-55) show block numbers from Wolfe and Shields (1997) and large letters (A-C) show new blocks that are mentioned in the text. Numbers after gene names indicate the chromosome on which the duplicate copy is located; 'm' indicates genes with paralogues on multiple other chromosomes. '@' symbols before gene names indicate that the orientations of a pair of genes are not consistent with the orientations of the rest of the genes in the blocks in which they lie. '(L)' before a gene name indicates a low scoring match (log-length normalised Smith-Waterman score below 17.5). '#' before gene names indicate genes that appeared on the original map (Wolfe and Shields, 1997) but which would not otherwise appear on the updated map using the current criteria. tRNA genes are indicated by names such as P{AGG}CR (indicating a proline tRNA with anticodon AGG on the right arm of chromosome III). *K. lactis* gene order information from Table 3.1 is shown in red or blue lettering (with the prefix *K.l.*). Red lettering indicates *K. lactis* neighbouring pairs that support the block structure; blue lettering indicates those that are either neutral or conflict with the block structure. Cases of complete gene order conservation between *K. lactis* and *S. cerevisiae* (left-hand column in Table 3.1) are not shown.

The paralogous gene pairs that form the ‘probable’ duplicated blocks are shown as thick coloured bars with gene names written to the right of chromosomes in Fig. 3.2. There are 52 ‘probable’ blocks and 45.5% of the genes in the genome are located inside them. These blocks contain 655 ‘probable’ paralogues (this is not an even number because, as well as simple gene pairs, it includes a few cases where a gene in a block has two tandemly duplicated paralogues in the sister block). For only 11 pairs among these, the transcriptional orientation of one gene appears to be inverted as compared to the other (relative to the rest of the block that contains them), indicating a DNA inversion that occurred after the whole genome duplication. However, this result is influenced significantly by the fact that genes at the ends of blocks were not included in the ‘probable’ map if their orientations did not match those of the rest of the genes in the block, i.e. orientation was one of the criterion on which the ‘probable’ map was based. The inverted genes are marked with ‘@’ symbols and named to the left of the chromosomes in Fig. 3.2. Seven of these inverted genes result from three multi-gene inversions in blocks 27, 37 and 41.

A further 34 pairs of paralogues are included as ‘possible’ additional genes within the ‘probable’ blocks. These do not have similarity scores greater than the cut-off value but they are otherwise consistent with the rest of the map. These ‘possibles’ are named to the left in Fig. 3.2, and marked ‘(L)’ for low-scoring. Transcriptional orientation, relative to the rest of the block, is conserved for 31 of these 34 pairs, which indicates that the majority (probably about 25-30) of these are true paralogues. The ends of some of the ‘probable’ blocks can be extended by including ‘possible’ paralogues (i.e. gene pairs that are either inverted or low-scoring), and these extensions are shown as narrower coloured bars on the map (Fig. 3.2).

There are 117 additional smaller ‘possible’ blocks. Of these, 32 have both copies in genomic regions outside ‘probable’ blocks (excluding any extensions as described above), while 11 have both copies completely inside ‘probable’ blocks. This indicates that approximately 21 of the 32 two-membered blocks are genuine sister regions (the other 11 being artefacts), which is in good agreement with the theoretical prediction for the number of two-membered blocks in yeast in Chapter 2. Only the 32 two-membered blocks that are outside the ‘probable’ blocks in both copies are shown in Fig. 3.2. It should be noted that approximately 11 of these are expected to be artefactual.

The revised map includes 39 tRNA gene pairs as well as one snRNA gene pair (SNR17A/SNR17B; Hughes *et al.*, 1987). tRNAs could not be used in the construction of the map in Fig. 3.2 because there are only 42 families of tRNA genes in yeast (see Chapter 1) and most tRNAs have multiple high BLASTN hits. A tRNA was included in the map if it occurred within a block and had a homologue located in the sister block, in the equivalent interval between protein paralogues. RNA genes are named on the left of the map in Fig. 3.2. We used a BLASTN score > 200 as the cut-off for identifying tRNA genes as homologues. This is not entirely satisfactory since it is a length sensitive cut-off, but in the majority of cases tRNA BLASTN scores were clearly separated into high and low scoring groups (probably reflecting tRNA families).

3.3.2 Comparison with the original map

52 of the 55 blocks on our earlier map appear as 'probable' blocks in Fig. 3.2, where they are numbered using the same scheme as used by Wolfe and Shields (1997). Blocks 1 and 36 were rejected because they are very close to telomeres (on chromosomes I/VIII and VI/VII, respectively). Block 52 (on chromosomes XI/XV) is reduced to 'possible' status because the three pairs of paralogues in the centre of the block are low-scoring. To facilitate comparison with the earlier map, all genes that were on that map but which would not otherwise have been included in the revised map, are shown to the left in Fig. 3.2 and marked by hash symbols ('#'). The total numbers of genes marked in Fig. 3.2 are: 655 'probable', 250 'possible', 78 tRNA and two snRNA, as well as 71 withdrawn ('#' symbols). This compares to 743 protein genes in Wolfe and Shields (1997). The fraction of the proteome involved in the whole-genome duplication is approximately 16% (905 proteins on the updated map/5523 proteins encoded by non-telomeric regions of the genome).

The most remarkable change in the updated map is that block 16 has been extended so that it spans the ribosomal DNA on chromosome XII, pairing it with part of chromosome IV. On chromosome IV, *SDH4* and *Q(TTG)DR3* (a glutamine tRNA gene) are about 15 kb apart, but their paralogues on chromosome XII (*YLR165W* and *Q(TTG)LR*) are separated by approximately 1 Mb (100-200 copies of the 9137 base-pair ribosomal DNA repeat; Johnston *et al.*, 1997). A second copy of the rDNA array seems to have been deleted without trace from this section of chromosome IV. A similar deletion of an rDNA array may have occurred during formation of the allopolyploid species *S.*

pastorianus, which is a hybrid between *S. cerevisiae* and an *S. bayanus*-like species, but which contains only *S. bayanus*-like rDNA (James *et al.*, 1997; Kurtzman and Robnett, 1998).

A large new 'possible' duplicated block was discovered between chromosomes VII and X (labelled as block B in Fig. 3.2). It includes *RNR4/RNR2* (encoding a ribonucleotide reductase subunit), *BUB1/MAD3* (spindle-assembly checkpoint kinases), *TDH3/TDH2* (glyceraldehyde-3-phosphate dehydrogenase), *SNG1/YJR015W* (transport proteins), and two tRNA genes. Curiously, this block spans the centromere of chromosome X but not chromosome VII.

The updated map includes several well-known duplicated gene pairs that did not appear in the previous map. These include *PDR1/PDR3* (transcription factors), *IRA1/IRA2* (GTPase activating proteins), *HTA1/HTA2* and *HTB1/HTB2* (histones), *CLB3/CLB4* (cyclins), and *NTG1/NTG2* (glycosylases). Some other gene families are not resolved into pairs and remain in competing alternative 'possible' blocks, for example *ADH1/ADH2/ADH5* (alcohol dehydrogenases) and *TUB1/TUB3/TUB4* (tubulins).

3.3.3 Intron losses

The set of gene pairs retained in duplicate includes 49 pairs in which at least one gene contains an intron. Of these, 11 pairs are missing the intron in one copy. By comparison to the available nucleotide sequences from *C. albicans,* we conclude that in almost all cases the intron was present in the ancestral gene, so that one intron was lost in *S. cerevisiae* after the genome duplication. The single exception to this is the gene pair *SEC14/YKL091C*. In this case the intron, present in *S. cerevisiae SEC14* (Bankaitis *et al.*, 1989), is missing from *YKL091C* and all the available orthologues of *SEC214* (*K. lactis, C. glabrata, C. albicans*). Further analysis suggests that this pair of genes should not have been attributed to the genome duplication, despite their similar sequences and paired genomic locations (Wolfe and Shields, 1997), because in phylogenetic trees the hemiascomycete *SEC14* sequences cluster together with *YKL091C* as an outgroup. Nonetheless, the intron may have been gained in *S. cerevisiae SEC14* after its divergence from *C. glabrata* and other ascomycetes. The above data provide an idea of the rate at which introns are lost in *S. cerevisiae* (10 introns lost out of 96 in ~ 10^8 years, ignoring possible parallel loss).

3.3.4 Comparison with *Kluyveromyces lactis*

The limited gene order information that is available from related species can provide useful information about the location of new sister regions, as well as serving as a check on existing regions. In earlier work the locations of gene pairs that were adjacent in the yeast *Kluyveromyces lactis* were compared with the locations of their orthologues in *S. cerevisiae* (Keogh *et al.*, 1998). The *K. lactis* genome appears not to be duplicated, based on gene order data, number of chromosomes, and phylogenetic analysis of duplicated gene sequences (Wolfe and Shields, 1997; Keogh *et al.*, 1998). With extensive additional data from *K. lactis* (Ozier-Kalogeropoulos *et al.*, 1998) and a revised map of the duplicated regions in *S. cerevisiae*, it was worthwhile to re-examine adjacent gene pairs in *K. lactis*.

Table 3.1 lists 84 pairs of adjacent *K. lactis* genes and groups them into three categories of gene order conservation (see also Chapter 1). The genes listed in the middle column of Table 3.1 ('conserved between blocks') are labelled in red in Fig. 3.2; these are 19 cases where gene order in *K. lactis* resembles the gene order that existed in an ancestor of *S. cerevisiae* prior to genome duplication followed by differential gene loss. The gene pairs listed in the right-hand column in Table 3.1 are labelled in blue in Fig. 3.2; these are 19 cases where the gene order in *K. lactis* does not appear related to the known block structure in *S. cerevisiae*. Where both of these blue labels occur in unpaired parts of the genome, they may indicate previously undetected (highly fragmented) blocks, for example the genes *ADH4* and *URA1* which are adjacent in *K. lactis* and near the telomeres of chromosomes VII and XI in *S. cerevisiae*. Other blue labels conflict with the 'probable' framework and indicate either interspecies rearrangements (translocations in *K. lactis* or transpositions in either species) or mistakes in the map. Four of these cases involve genes located in duplicated block 53 on chromosome XII (Fig. 3.2). Four rearrangements in the small region occupied by block 53 seems unlikely, so this block is probably spurious. It contained the minimum number of paralogues (just three) for inclusion in the original map, and two paralogues (*YLL024W/YLR037C*) are members of the large PAU multigene family.

Table 3.1

	Gene-pairs adjacent in *K. lactis* that remain adjacent in *S. cerevisiae*	Gene-pairs conserved between duplicated blocks	Gene-pairs adjacent in *K. lactis* and not conserved in *S. cerevisiae*
Predicted	58%	25%	17%
Observed	54% RFT1 HAP3 TKL2 LYS2 MRK1 THI3 ERD1 YDR412W APA2 QCR7 CDC68 CHC1 ERG20 QCR8 GAL80 YML050W URA5 SEC65 RPL41A YNL161W GAL11 GSH2 YOL119C RPL18A GAL1 GAL10 GAL10 GAL7 ZWF1 YNL240C YNL240C KEX2 KEX2 YTP1 YTP1 SIN4 SPT4 COX18 PEX3 SKP1 YIR003W DJP1 YCL036W YCL035C RAD16 LYS2 ABD1 PRP5 YBR238C YBR239C YPL112C CAR1 NOP4 SSN3 YGR046W TFC4 YGR117C RPS23A GPD2 ARG1 RPO31 RPT5 YLL035W YLL034C SMC4 YLR087C YLR386W YLR387C YDR387C RVS167 YNL217W RAP1 SAM1 YLR181C YLR181C SWI6 UME1 YPL138C YJL082W YJL083W RPL32 RPL24A RFA2 YNL308C MET6 YER093C SDH3 CTK1 YOR294W YOR296W YKL006CA CAP1	23% HHT1 TRP1 TRP1 IPP1 RLP7 LEU2 RAP1 GYP7 GAL4 SGS1 PDA1 YDR101C UBP2 YDR372C ARG8 KRE1 YDR421W YML006C SFA1 GIM1 YDR430C YML011C YGR111W AXL1 YGR196C YJR013W APM2 YKL040C RED1 GLN4 SPF1 YJR046W PTA1 YOR359W YLR192C DLD1 RRN6 TRP5	23% LAG2 PGK1 KIN28 MRF1 MET17 YLL015W GAP1 ADH1 CTF18 CBF1 GLO1 PFK2 THI1 CYC1 YBR287W SCP1 MAK32 VAC8 YDR407C MOT1 SPP41 KRE6 ADH4 URA1 YGL036W KNS1 PRP38 DPS1 CPS1 YJL066C YLR455W VPS4 HGH1 YLL013C GAL7 NAT1 SEC31 YLR218C

Table 3.1 *S. cerevisiae* orthologues of adjacent pairs of *K. lactis* genes in three gene order conservation categories.

Adjacent *K. lactis* genes that map to locations near three 'possible' sister regions add weight to these new candidate blocks (blocks A, B and C; Table 3.1 and red labels in Fig. 3.2). These examples illustrate how complete mapping of *K. lactis* (or *A. gossypii*) would provide a much clearer picture of the sister regions in *S. cerevisiae* and of the evolution of gene order after genome duplication. Another example of the utility of *K. lactis* information is the relationship between block 49 (chromosomes XIV and XV) and the genes *KRE1* and *ARG8* which are adjacent in *K. lactis*. The positions of *KRE1* and *ARG8* in *S. cerevisiae* are incompatible with the possible extension of block 49 to include the gene pair *HXT14/HXT11*, so the *HXT* pair is probably artefactual.

In Table 3.1, the 'predicted' values for the percentage of gene pairs in three columns are based on the original map of duplicated regions (Wolfe and Shields, 1997). We did not adapt the analysis for the updated map because it is not clear how to include uncertain ('possible') regions in the analysis. Also, the results of Ozier-Kalogeropoulos *et al.* (1998) are based on 'genome survey' sequencing of both ends of plasmid clones, and in some cases their paired *K. lactis* sequences correspond to *S cerevisiae* genes that are separated by small number of intervening genes; these data are awkward to analyse. However, the difference between the maps is not significant and the observations from *K. lactis* (Table 3.1) remain close to the predictions in Chapter 1 and Keogh *et al.* (1998).

3.4 Discussion

3.4.1 Why Keep Duplicated Genes?

After gene duplication one member of a gene pair may accumulate deleterious mutations and be lost or both copies of the gene may be retained. There are two likely evolutionary reasons for retaining both copies: selection for increased levels of expression, or divergence of gene function. Functional divergence can be produced through complementary degeneration (Force *et al.*, 1999), where each daughter gene retains only a subset of the functions of the parent, or (perhaps more rarely) if one daughter acquires a new function. Degenerative tetraploidy provides an opportunity to study the evolution of many duplicated pairs of genes, which were all formed simultaneously (Seoighe and Wolfe, 1999b).

We estimated above that approximately 8% of the genes in the pre-duplication *Saccharomyces* genome were retained in duplicate, so that duplicate pairs formed by polyploidy account for approximately 16% of the current *S. cerevisiae* gene set. We have identified 12.9% of *S. cerevisiae*'s genes as polyploidy-derived duplicates, so most of the pairs formed by this event have already been found. The remainder lie in regions of the genome that were heavily fragmented by rearrangements. Compared to the average for the genome (12.9%), genes classified as essential are significantly under-duplicated, and non-essential genes are significantly over-duplicated (2.7% and 16.6%, respectively; Table 3.2).

Table 3.2 Fraction of *S. cerevisiae* proteins in different functional categories (Hodges *et al.*, 1999) that have been retained in duplicate since genome duplication.

Protein category	Number of proteins in category	Percent retained in duplicate	$\chi^{2\,*}$
All proteins	5792	12.9	
Essential proteins	731	2.7	59
Non-essential proteins	2255	16.6	24
YPD Functional Categories			
Cyclins	22	54.5	30
Protein phosphatases	40	32.5	12
Heat shock proteins	32	31.3	8
Protein kinases	123	29.3	26
GTPase-activating proteins	19	26.3	
Glucose metabolism	223	26.0	30
Guanine nucleotide exchange factors	23	21.7	
GTPases	55	18.2	
Amino acid metabolism	189	12.7	
Transcription factors	261	12.3	
tRNA synthetases	42	11.9	
ABC cassette proteins	30	10.0	
Proteases (non-proteasomal)	72	9.7	
Ubiquitin-conjugating proteins	24	8.3	
Proteasome subunits	34	2.9	
Serine-rich protein	10	0.0	
AAA ATPase domain proteins	16	0.0	
Ribosomal proteins	209	39.2	112
Mitochondrial Ribosomal proteins	44	0.0	6
Nucleic Ribosomal proteins	165	50.3	179

*χ^2 values with one degree of freedom are shown if significant at the 5% level.

This illustrates the apparent genetic redundancy of many duplicated genes, although Thatcher *et al.* (1998) reported that yeast genes that were previously classified as non-essential may in fact make a small contribution to evolutionary fitness. It is apparent from the functional classifications of duplicated proteins and from the excess of duplicated genes classified as non-essential (Table 3.2) that genes that were retained in duplicate continue to perform closely related functions. Of the 280 duplicated pairs for which the Yeast Proteome Database (YPD; Hodges *et al.*, 1999) lists a functional category for both proteins, the categories are different for only 26 pairs and most of these differences do not appear significant when examined more closely. Taken together these observations suggest that in many cases duplicated genes were retained to increase specificity and thereby to improve the efficiency with which important existing functions were carried out.

The genes that have been retained in duplicate in *S. cerevisiae* are also not distributed evenly among YPD functional categories (Table 3.2), indicating some non-randomness or predetermination of the fates of duplicated genes. Some functional categories are over-duplicated, including cyclins and much of the signal transduction apparatus (protein kinases and phosphatases, GTPases, GTPase-activating proteins and guanine nucleotide exchange factor, but not transcription factors). Many cytosolic ribosomal protein genes, but no mitochondrial ribosomal protein genes, are duplicated. For cytosolic ribosomal proteins, 50.3% of the genes are mapped to duplicated chromosomal regions and can be attributed to genome duplication. Many of the remainder are also duplicated (Planta and Mager, 1998) but do not belong to identified paired regions.

Many of the over-duplicated functional categories (Table 3.2) include very highly expressed genes, such as heat shock, glucose metabolism, and cytosolic ribosomal proteins. The correlation between the expression level of a gene and its likelihood of being retained after whole-genome duplication was explored further using whole-genome transcription data from Holstege *et al.* (1998). The tendency to retain high-expression genes in duplicate is not confined to the highest categories of gene expression but extends down to expression levels of about 10 mRNA molecules per cell (Fig. 3.3). Thus it appears that increased gene expression (and consequent rapid growth) was a significant concern in the sorting-out of which genes were retained and which were lost. It must, however, be noted from Fig. 3.3 that a majority of duplicated genes have expression levels below 10 molecules per cell, and that selection for diversification of gene function may have been important for these genes. It will be of interest to see whether the criteria for sorting-out were the same in other lineages such as *C. glabrata*.

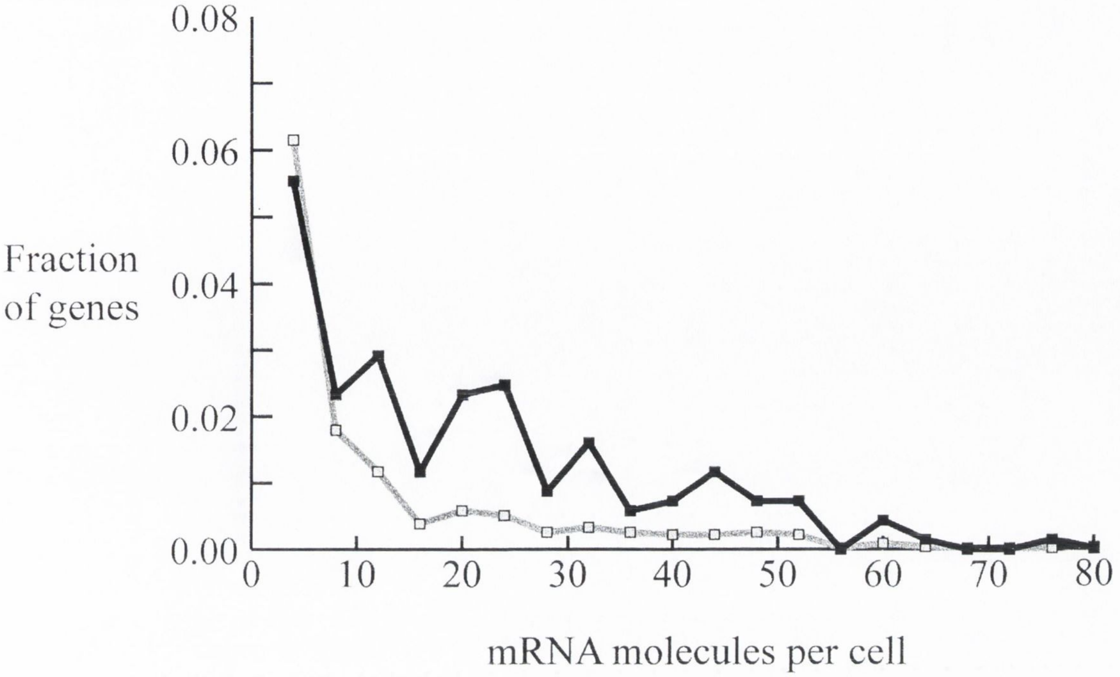


Figure 3.3 Fraction of all genes that are expressed at a given level (given as mRNA molecules per cell) plotted against the expression level in bins of size four molecules per cell (grey line). Fraction of those genes that have been retained in duplicate since genome duplication plotted against expression level (black line).

Chapter 4

G+C Content Variation Along and Among Yeast Chromosomes

4.1 Introduction

The variation of G+C content along chromosomes was discussed in the primary publications of most of the *S. cerevisiae* chromosome sequences (e.g. Bowman *et al.*, 1997; Johnston *et al.*, 1997). The reason for the high level of interest in compositional variation along yeast chromosomes may be connected to extensive research into isochores in vertebrates in the years preceding the *S. cerevisiae* sequencing project (e.g. Bernardi, 1993; Kadi *et al.*, 1993; Sabeur *et al.*, 1993) and to the report that chromosome III (the first chromosome to be sequenced) showed G+C content variation (Sharp and Lloyd, 1993). By the end of the sequencing project interest in compositional variation appears to have decreased and no comprehensive overview of G+C content was produced that took account of the data from the whole genome sequence.

There are several reasons for revisiting the question of variation of G+C content of *S. cerevisiae* chromosomes in the context of the present work. Besides the value of establishing a complete picture of compositional variation, using the whole genome sequence, it is useful to view patterns of G+C content in relation to the structural evolution of chromosomes. The set of chromosomes in a genome is continually undergoing change through rearrangement, both within and between chromosomes. If a pattern exists in G+C content, and is widespread, then the mutational process that causes it must be rapid compared to the rate of chromosomal rearrangement. However, patterns in the G+C content were not reported for all of the chromosomes in yeast. The strong peaks of G+C content that were observed in the sequence data from chromosome III (Sharp and Lloyd, 1993) have never been equalled by the variations in G+C content reported in sequence data from subsequent whole chromosome sequences and many authors reported no structure in base composition along the yeast chromosomes that they assessed (e.g. Johnston *et al.*, 1997; Tettelin *et al.*, 1997). It is not clear why there is more apparent structure in some chromosomes rather than others but it appeared to be possible that stability against rearrangement over time plays an important part in determining the extent of compositional heterogeneity (Bradnam *et al.*, 1999). Again the importance of studying the patterns of G+C content in relation to conserved regions of gene order is clear. Several authors reported periodicity in the patterns of GC3s content but this periodicity was not universal in yeast chromosomes. Slowly evolving periodic variations in G+C content, if they exist, could be used as a guide to distinguish between alternative ancestral block orders and help to tackle the reconstruction problems in Chapter 2. Chromosomal rearrangement would have a destructive effect on periodicity and so regions exhibiting stronger periodicity would be more likely to represent ancestral block orders.

Work described in this section was undertaken together with Keith Bradnam from the research group of Paul Sharp in Nottingham and has been published in *Molecular Biology and Evolution* (Bradnam *et al.*, 1999).

4.1.2 Review

DNA centrifugation in Cs_2SO_4-Ag+ density gradients divides high molecular weight DNA from the higher organisms into different components with distinct molecular weights (Filipski *et al.*, 1973), corresponding to large regions (>>300kb) of chromosomes that are homogeneous in G+C content (Bernardi, 1993). Five families of DNA fragments have been recognised, including two light families, L1 and L2, and three heavy families, H1, H2, H3. The chromosomes of warm-blooded vertebrates have been described as a mosaic of such components known as isochores (Bernardi *et al.*, 1985; Duret *et al.*, 1995). That this fractionation of the genome is biologically relevant is suggested by the fact that large regions around genes are normally compositionally homogeneous (Ikemura and Wada, 1991). Furthermore the coding density is very highly correlated with isochore class, with isochores in the H3 family sixteen times more gene dense than isochores in the light isochore families (Bernardi *et al.*, 1985). The importance of isochores is further increased by the distinct isochore composition of different chromosome bands (Saccone *et al.*, 1993). It is also significant, for evolutionary studies, that isochore patterns appear to be highly conserved between mammalian species. In studies involving nine orders of mammals six orders showed a general isochore pattern (Sabeur *et al.*, 1993; Caccio *et al.*, 1994). Eight orders of birds studied were also very compositionally similar (Kadi *et al.*, 1993). When homologous genes from species sharing the same isochore pattern are compared G+C values are highly correlated with slopes of regression lines close to unity. Comparisons between species with the general pattern and species showing specific isochore patterns (such as pangolin and shrew) continue to produce strong correlations although the regression lines have slopes different from unity, indicating that, although G+C contents have changed, the rank of genes remains the same (Caccio *et al.*, 1994).

It has frequently been suggested by Giorgio Bernardi and co-researchers that the higher body temperatures of warm-blooded vertebrates may be the reason for the greater compositional heterogeneity in warm-blooded vertebrates compared to cold-blooded vertebrates. Higher G+C content leads to greater stability against denaturation at high temperatures and the high G+C content isochores could act as a kind of genomic glue to ensure the integrity of important gene-dense regions in the genome. However if this is the correct explanation for the isochore structure then it is not clear why increases in the G+C content of G+C-rich regions have not been accompanied by at least some small increases in the G+C content of G+C-poor regions also.

The much smaller genomes of prokaryotes and unicellular eukaryotes do not contain the same kind of compositional heterogeneity observed in higher organisms. If isochores as large as the isochores in vertebrates (>300kb) were present in *Saccharomyces cerevisiae*, for example, they would frequently span entire chromosomes. In addition the variation in coding density is necessarily much smaller in more compact genomes such as *S. cerevisiae* and the kind of coding density variation that is linked with isochores is not seen. If any phenomenon analogous to isochores exists in the yeast genome then it must be on a different scale to isochores observed in vertebrates. When Sharp and Lloyd (1993) observed significant non-randomness in the G+C contents of genes along chromosome III of *S. cerevisiae* it seemed likely that isochores related to a more general phenomenon than previously thought. Unlike vertebrate isochores the variation in chromosome III was only observed in coding regions (and most strongly in the third codon position) leading the authors to postulate that intergenic regions are under greater constraint. When the weighted average of sliding windows of fifteen adjacent sequences (coding or intergenic regions) were taken clusters of high-G+C content genes were seen as peaks, with one such peak either arm of the chromosome (Sharp and Lloyd, 1993). The 'peaks' were strongest in the case of G+C content of third codon position silent sites (GC3s) of open reading frames. This analysis was repeated in the primary publication of chromosome XI, the second fully sequenced chromosome, but the authors carelessly referred to the clusters, made apparent by this method, as '(G+C)-rich peak[s]' (Dujon *et al.*, 1994). In fact there was never any real indication of peaks of G+C content, although the term was quite commonly and explicitly used in the primary publications of many of the *S. cerevisiae* chromosomes (e.g. Bowman *et al.*, 1997; Churcher *et al.*, 1997; Johnston *et al.*, 1997) and substituted in one case by the term 'waves' (Galibert *et al.*, 1996). The fact that compositional variation was normally viewed using the sliding window technique gave the appearance of peaks or waves of G+C content. Peaks in the sliding window plot of G+C content along chromosome XI led Dujon *et al.* (1994) to suggest that the distribution of G+C rich regions was periodic along the chromosome, with a period of about 100 kb. The periodicity of G+C content appeared also to coincide with periodicity in coding density, furthering the analogy with isochores in vertebrates. However correlations between 'peaks' of G+C content and coding density observed in chromosomes XI, III, II, VII, X, XIII and XV were often quite subjective and correlation was reported absent or not reported in the remaining chromosomes. In general little evidence was presented for a correlation except in the case of chromosome VII (Tettelin *et al.*, 1997).

4.2 Data and Methods

DNA sequences and details of ORFs as downloaded by our collaborators in Nottingham (January 1998) from the Saccharomyces Genome Database (Cherry *et al.*, 1998) were used in this study. Chromosome sequences were downloaded from ftp://genome-ftp.stanford.edu/pub/yeast/genome_seq/, and ORF locations from ftp://genome-ftp.stanford.edu/pub/yeast/tables/ORF-Locations/. After removal of ORFs that were completely contained within larger ORFs and Ty elements there were 6145 ORFs remaining in the dataset. In this section the duplicated blocks used were from the original study (Wolfe and Shields, 1997).

Keith Bradnam (1999) devised a method to delimit clusters of ORFs with similar GC3s content. The method involved examining all possible sets of adjacent genes in a chromosome, containing less than half of the total number of genes in the chromosome. Student's t-tests were then carried out to compare the mean of the set with the mean value of the rest of the chromosome. Adjacent sets of genes having mean GC3s content significantly different from the mean of the rest of the chromosome were recorded. Finally, the ORF locations were shuffled randomly and the analysis was repeated on the shuffled chromosome to test whether the apparently significant differences were the result of the large number of adjacent sets of genes that were compared. This method gives an empirical significance to the clusters that have been delimited using t-tests.

To construct a genome in which individual ORF GC3s content was influenced by the GC3s of the preceding ORF we introduced a modified shuffling method (see section 4.3.6 for application of this method). In this shuffling scheme, randomly chosen ORFs were sequentially accepted or rejected with a probability based on the difference between their GC3s content and the GC3s content of the preceding ORF. For example, if this difference in GC3s content was 0.02 and if the fraction of ORFs in the real data that had 0.02 of a difference in GC3s content as compared with their nearest neighbour was 0.1 then that randomly chosen ORF would be accepted with a probability of 0.1.

4.3 Results

The first step in reassessing the GC3s content pattern was to plot the GC3s variation along each of the 16 chromosomes using the same sliding window technique that was used in most of the original publications. The GC3s variation was plotted using a sliding window of 15 adjacent genes and the resultant graphs were drawn on the same scale for ease of comparison between chromosomes (Fig. 4.1; Bradnam *et al.*, 1999). Chromosome III shows the highest 'peaks' of GC3s content but there are also significant clusters on most of the other chromosomes, indicated by peaks on the sliding window plots. Some chromosomes, such as chromosome XI appear to show regular spacing between GC3s-rich clusters, however, overall it is difficult to see any consistent pattern in the spacing of GC3s clusters among all 16 chromosomes.

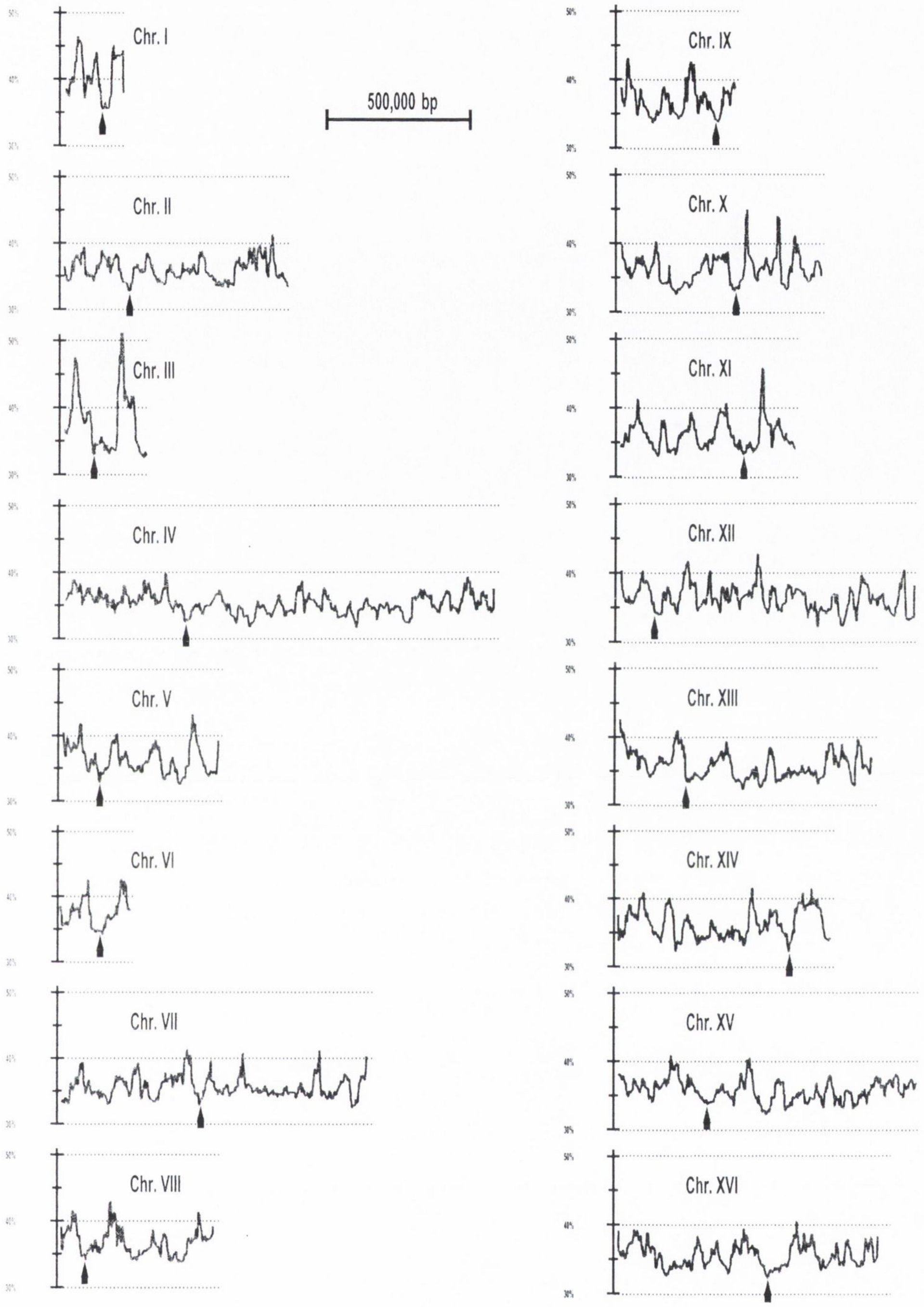


Figure 4.1 Variation in silent-site G+C content (GC3s) along 16 yeast chromosomes. GC3s was calculated as in Sharp and Lloyd (1993) using a sliding window of 15 open reading frames (ORFs), but plotted as a line rather than as a series of points for ease of presentation. All chromosomes are drawn to the same scale. Arrows denote approximate positions of centromeres. Dotted lines denote 30%, 40% and 50% GC3s.

4.3.1 GC3s Variation in Duplicated Chromosomal Regions

The peaks of high GC3s that appear in Fig. 4.1 are not universal. It is possible that these peaks occur in regions that have not undergone chromosomal rearrangement in their recent history. These peaks, which span up to approximately 40 ORFs, might be produced by long-range effects that require stability of a chromosome segment over a long period of time. If this is the case, then large regions that have been undisrupted by chromosomal rearrangements might be expected to contain some of the largest peaks. The largest duplicated blocks identified in Wolfe and Shields (1997) provide substantial regions that have not undergone large-scale rearrangement since genome duplication about 10^8 years ago. Locations of GC3s peaks exceeding 40% (see dotted line in Fig. 4.1) were examined against the co-ordinates of the 22 largest undisrupted segments in the yeast genome, spanning about 17% of the genome. Of 63 GC3s peaks only 10 were found to be even partially within the large duplicated blocks. This was not more than would be expected by chance. The conclusion is that 'peaks' of GC3s content are not preferentially located in stable regions of chromosomes. The corollary to this is that compositional heterogeneity is unlikely to be of any use in determining the original order of duplicated blocks in *S. cerevisiae*. Analysis of GC3s in pairs of genes that remain in duplicate since genome duplication (Fig. 4.2) shows only a very weak, though significant, correlation ($r = 0.34$, $N = 406$, $P < 0.01$) indicating that change of GC3s content is rapid compared to the time estimated for genome duplication.

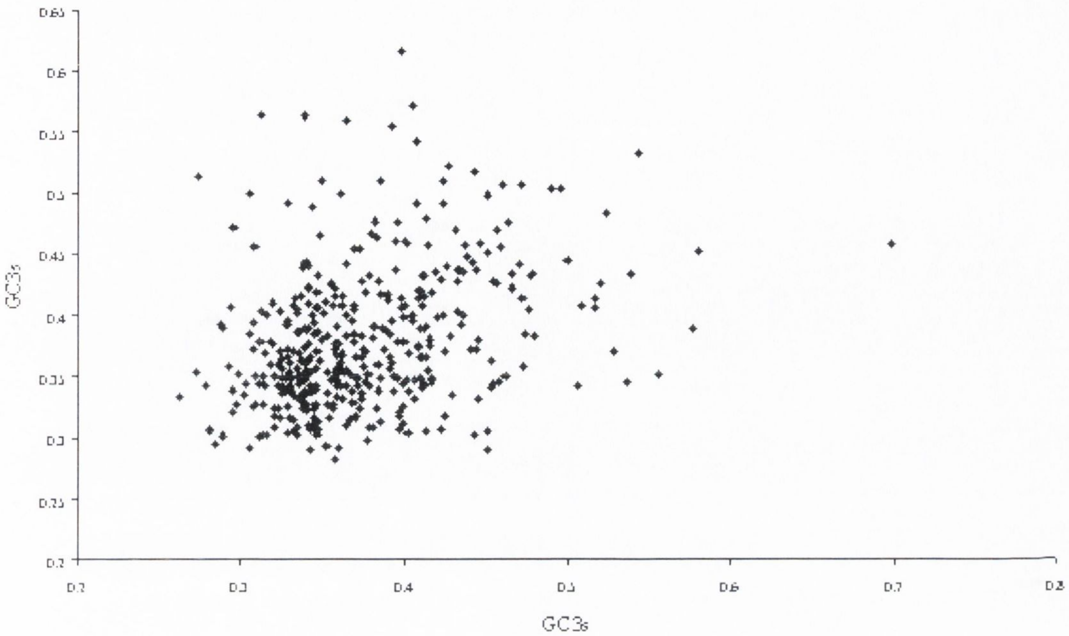


Figure 4.2 G+C content at silent sites for paralogous pairs of genes that have been retained since duplication of the yeast genome.

4.3.2 Clustering of ORFs with similar GC3s content

As already mentioned the methodology used to display GC3s variation in Fig. 4.1 has a natural bias towards showing peaks in the data and it is perhaps better to consider GC3s variation in terms of the clustering of ORFs of similar GC3s values. With the method that was devised to delimit such clusters (see section 4.2). Keith Bradnam found that most chromosomes had only a few of these clusters; some chromosomes (I and II) had none at all (Bradnam *et al.*, 1999). Both high-GC3s and low-GC3s clusters were found. The high GC3s clusters were typically small, less than ten ORFs, whereas the low GC3s clusters were usually much longer, up to 200 ORFs. Chromosomes X and XI produced many high GC3s clusters, but it was chromosome III that again stood out. Most of chromosome III is occupied by two sets of overlapping clusters, one of which includes a window of six GC3s-rich ORFs on the right arm that is the most significant cluster in the entire genome and the other is a large GC3s-poor area (data not shown). The conclusion is that GC3s content is not entirely random along the chromosomes. There are regions, varying in size, that have a significantly higher average GC3s content than the rest of the chromosome on which they are (particularly chromosome III).

4.3.4 Periodicity in GC3s

Time series analysis was used to test whether there was any periodicity in the GC3s content along chromosomes in the yeast genome (e.g. see Chatfield, 1989, chapter 2). Fig. 4.3 shows separate correlograms for each of the sixteen chromosomes and a single correlogram for all sixteen chromosomes taken together. These correlograms show the correlation (as measured by the autocorrelation coefficient, r_k) between the GC3s values of any two ORFs as a function of the distance (measured in ORFs) between the two ORFs. With completely random data a correlogram would be expected to decrease rapidly and fluctuate randomly about $r_k = 0$. Periodicity in the data should produce regular significant autocorrelations at large distances. For most chromosomes there was no significant autocorrelation at large distances. In each chromosome however there was a significant short-range autocorrelation normally extending only to nearest neighbouring ORFs (k = 1) or sometimes to second nearest neighbours (k = 2). The autocorrelation coefficient of neighbouring ORFs in the whole genome was 0.36 ($P << 0.001$). The nearest neighbour correlation is clearly highly statistically significant. In 10,000 random permutations of the data this level of correlation was not reproduced. It appears that the non-randomness in the distribution of GC3s is a very short-range phenomenon. However, chromosome III again is different from the other chromosomes and seems to leap out of Fig. 4.3. It is the only chromosome to exhibit significant long-range correlations, and short-range correlations (over distances of 1-5 ORFs) are by far the strongest in this chromosome. This very strong short-range correlation may be sufficient to give the impression of longer-range effects.

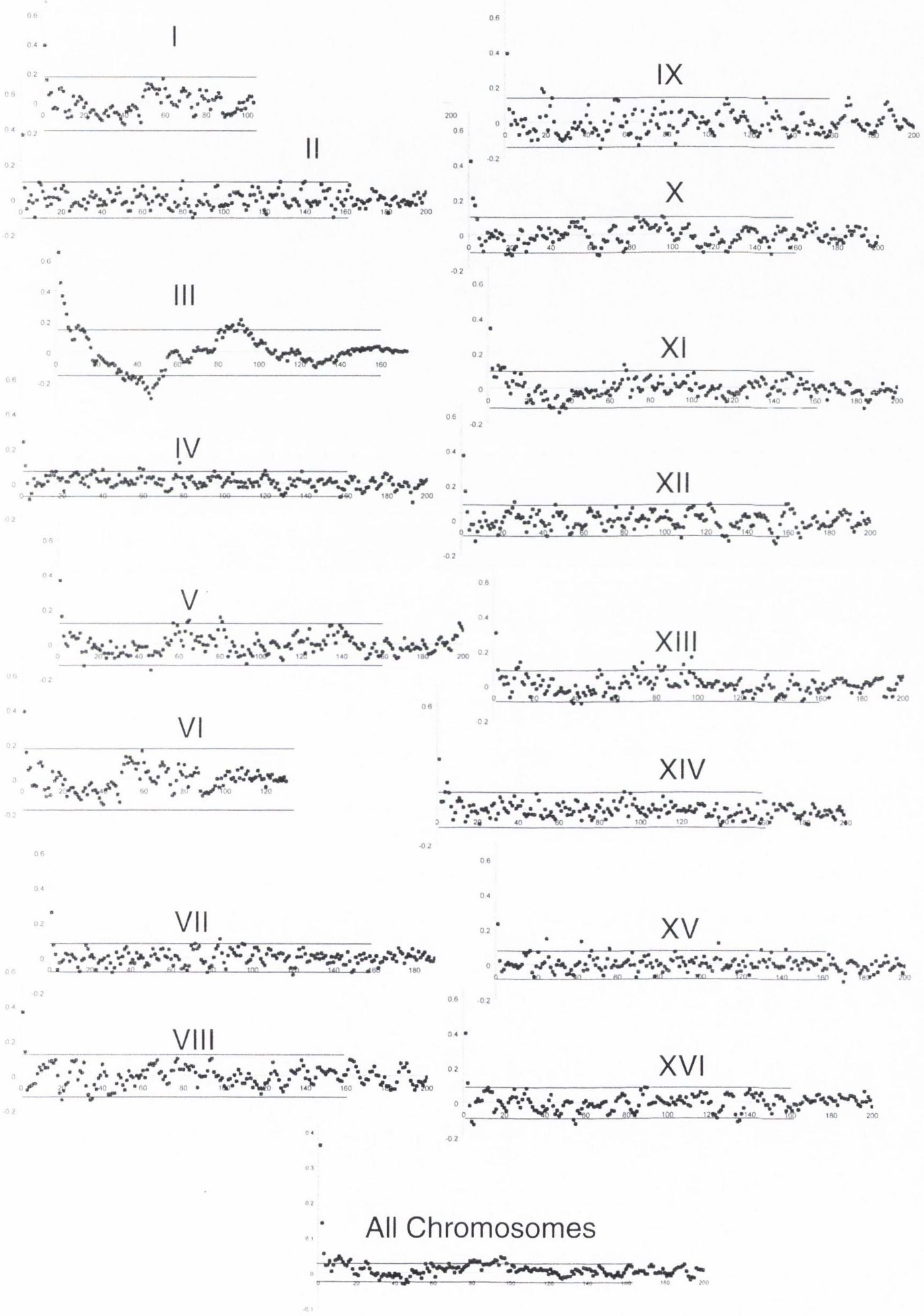


Figure 4.3 Correlograms for yeast chromosomes. The Y-axis in each plot shows the autocorrelation coefficient, r_k,, which is a measure of the correlation between the GC3s values of

all ORFs that are a distance *k* ORFs apart on the chromosome in question. Distances (*k*) between ORFs are shown on the X-axis for values up to $k = 200$. The 5% significance levels for autocorrelation coefficients are shown as two solid lines. Points lying outside these lines represent significant correlations. The bottom panel shows the pooled result when all chromosomes are considered together.

4.3.5 Nature of short-range correlations

The short-range correlations in GC3s between ORFs appears to be independent of the strand on which the neighbouring ORFs are located. The strength of correlation between neighbouring ORFs depends strongly on their distance of separation, measured in base pairs. In the whole genome the 5% of neighbouring ORFs that are most distantly separated have $r_1 = 0.19$; for the 5% of closest ORFs, $r_1 = 0.43$. There is no significant positive correlation between the G+C content of adjacent non-coding regions (even when taking into account the fact that neighbouring non-coding regions tend to be further apart than neighbouring ORFs). However, if non-coding regions are split in two, then there does appear to be a correlation in the G+C content of the two halves (r = 0.27, N = 6300, $P < 0.01$). This suggests that there is a very short-range correlation in the base composition of non-coding regions. Correlation between neighbouring ORFs is much stronger for ORF-pairs of high GC content. Nearest neighbour correlation is largely a phenomenon affecting ORFs with high G+C content. The correlation between the 1800 neighbouring ORF-pairs with high G+C content (both ORFs above 36% G+C) was strong (r_1= 0.30) compared to the correlation of the 1900 low (both below 36% G+C) G+C content ORF-pairs (r_1= 0.08). We checked to see whether the correlation between adjacent ORFs was a result of correlated levels of expression in neighbouring genes. Highly expressed yeast genes have a biased codon usage, and use a subset of 'optimal' codons (Bennetzen and Hall, 1982; Sharp and Cowe, 1991). No correlation was detected between the frequency of these optimal codons in ORFs and their nearest neighbours. This suggests that clustering of ORFs according to their level of expression is not a feature of the yeast genome, nor the cause of the nearest neighbour correlation.

4.3.6 Using short-range correlations to explain the observed GC3s heterogeneity

We investigated whether the significant clusters of ORFs with similar GC3s values can be explained by the very short-range correlations between neighbouring ORFs. To do this, the multiple t-test methodology outlined earlier (see section 4.2) was repeated, but a bias was introduced into the way the ORFs on a chromosome were shuffled. Chromosome III was chosen because it contained the most pronounced clustering of high-GC3s ORFs. Rather than shuffling ORFs randomly, ORFs were shuffled taking into account the short-range correlations mentioned above. To do this, we first observed the range of differences in the GC3s values of adjacent ORFs in the real, unshuffled data. From this we can determine how likely it is that two ORFs with a given difference in GC3s will be adjacent to each other, and so shuffled datasets can be produced that have short-range correlation profiles similar to the real data (see section 4.2). Using this modified shuffling technique, all of the significant clusters of ORFs on chromosome III could be easily reproduced. Therefore, the nearest neighbour correlation appears to be sufficient to explain the significant clustering of ORFs of similar GC3s values. Correlograms based on chromosome

III after this kind of shuffling resembled the correlogram of chromosome III in Fig. 4.3 and often give the impression of long-range periodic effects. Because the observed clustering and autocorrelations can be explained solely by short-range effects we conclude that there is little or no evidence for long-range order in the GC3s content of yeast chromosomes.

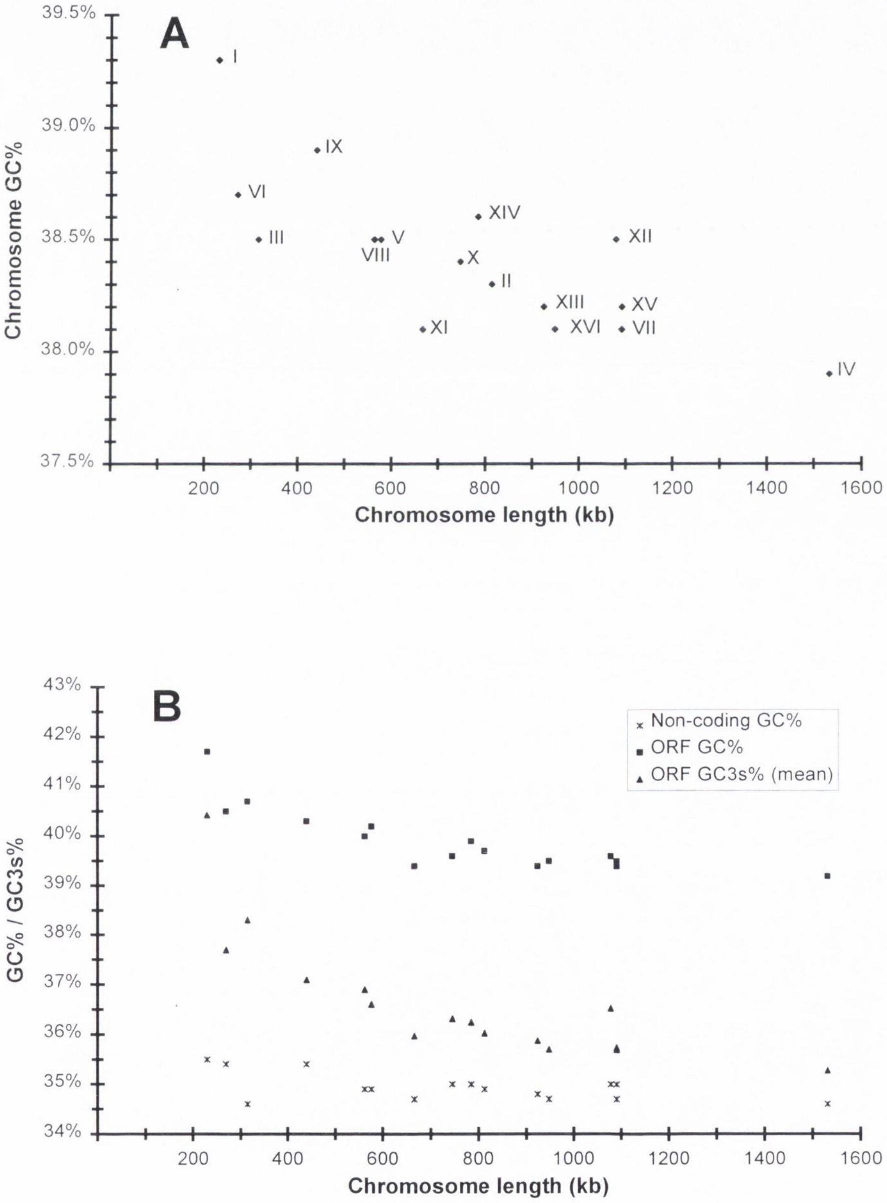


Figure 4.4 Figure taken from Bradnam *et al.* (1999), showing the relationships between chromosome length and G+C content. *A,* chromosome G+C% against chromosome length.

Roman numerals indicate chromosome number. *B,* non-coding G+C%, ORF G+C% and ORF GC3s% against chromosome length.

4.3.7 Chromosome length and G+C content

As well as compositional heterogeneity along chromosomes there are differences in GC3s (and bulk G+C) content between chromosomes. Keith Bradnam observed that chromosome average GC3s content was negatively correlated with chromosome length ($r = -0.81$, $P < 0.1$, see Fig. 4.4; Bradnam *et al.*, 1999). When modal values instead of average values of GC3s content were assessed this negative correlation disappeared, indicating that the distribution of ORFs with extreme GC3s contents was the cause of the negative correlation between G+C content and chromosome length. The frequency distribution of GC3s values (Fig. 4.5) is significantly asymmetrical, showing a tail towards high GC3s values. When the fraction of ORFs on a given chromosome with high GC3s content (above 0.38) is plotted against chromosome length a negative correlation of almost the same magnitude can be observed ($r = -0.82$, $P < 0.1$), indicating that it is the non-random distribution of high-GC3s ORFs that causes the correlation between chromosome length and average GC3s content.

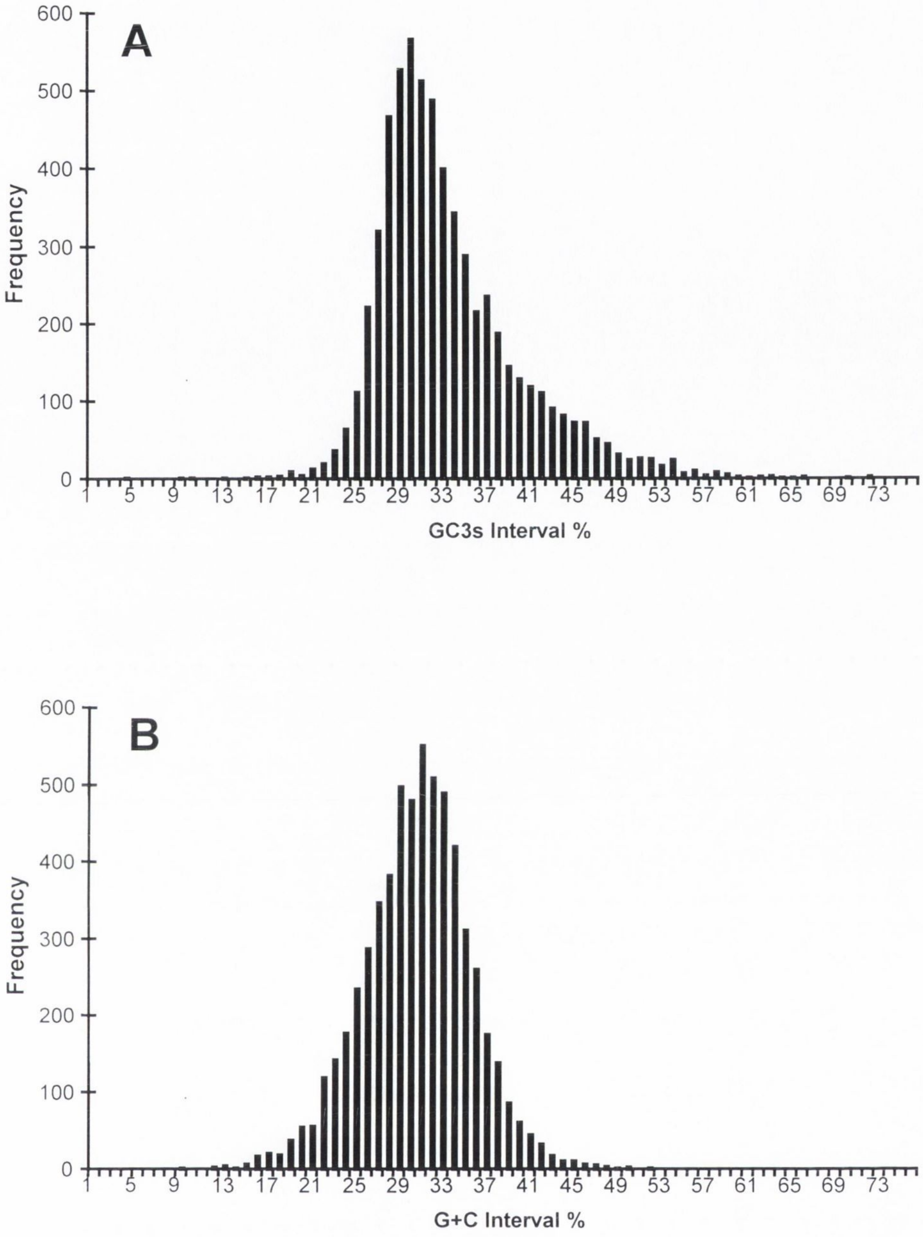


Figure 4.5 Distribution of G+C and GC3s values in the yeast genome (interval size, 1% G+C or GC3s). *A*, Distribution of all ORF GC3s values from 6145 ORFs. *B,* Distribution of G+C values from 6004 noncoding regions (regions of less than 75 bp were excluded from the analysis).

4.3.8 Interspecies comparisons

From a dataset consisting of *Candida albicans* contigs (see Chapter 5) the 2587 putative *C. albicans* ORFs that were contained completely within sequenced contigs were selected (*Candida albicans* ORFs were identified on the basis of similarity to *S. cerevisiae* proteins using TBLASTN (Altschul *et al.*, 1990) with seg filter (Wootton and Federhen, 1996) and a cut-off P-value of 10^{-10}). The contigs were not annotated and the problem of assessing third codon position G+C content was not tackled. However even this cursory examination of the G+C content of *C. albicans* and comparison to that of *S. cerevisiae* reveals clear differences between the two genomes (see Fig. 4.6A). The genome of *C. albicans* is known to be more A+T-rich than *S. cerevisiae* (Lloyd and Sharp, 1992). The average fraction of G or C bases in the *C. albicans* genes in our dataset was 0.353. The corresponding fraction in the *S. cerevisiae* homologues of this group of genes was 0.405. However, whereas the frequency distribution of G+C in *S. cerevisiae* shows a clear positive tail the distribution in *C. albicans* shows a tail towards the low end of G+C content (Fig. 4.6).

In *C. albicans* the correlation between nearest neighbour G+C contents in 2256 pairs of nearest neighbours was 0.41. In a randomly selected set of 2256 pairs of *S. cerevisiae* nearest neighbours the correlation co-efficient was 0.28. The stronger correlation in *C. albicans* nearest neighbours was in spite of the fact that they are located over twice as far apart as their *S. cerevisiae* homologues (possibly due to the existence of intervening genes in *C. albicans* that do not have *S. cerevisiae* homologues; see Methods, Chapter 5). If G+C content is more strongly correlated in *C. albicans* it is likely that there is some compositional clustering of genes in this genome also. The lower value of G+C content in *C. albicans* may be largely an effect of the third codon position. In a sample of 324 genes downloaded from the EMBL database the average value of

GC3s was 27.8%, much lower than the corresponding figure in *S. cerevisiae* (37.4%).

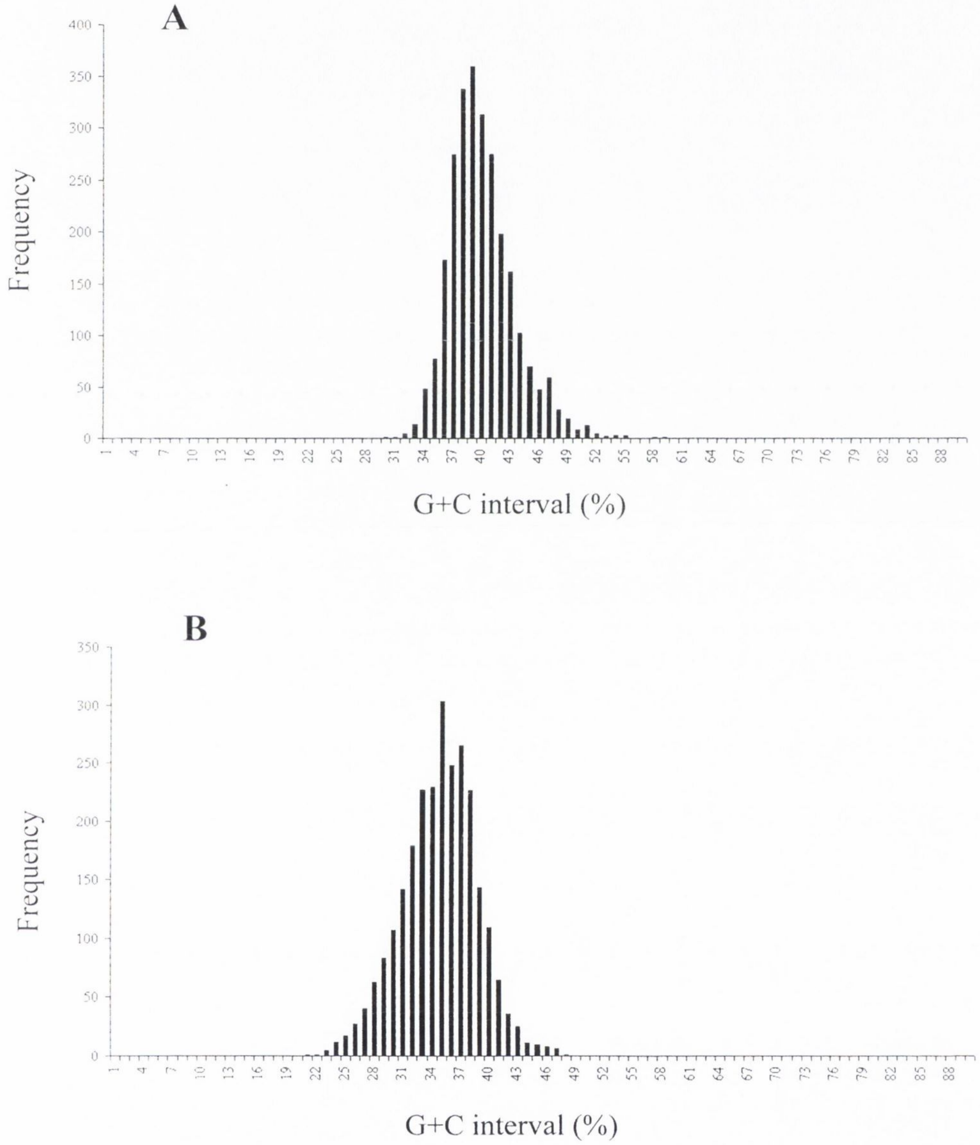


Figure 4.6 Distribution of G+C values (all codon positions combined) of 2600 *Saccharomyces cerevisiae* genes that have apparent homologues contained within the sequenced contigs of *Candida albicans* (A). Distribution of G+C values of the corresponding *C. albicans* genes (B).

4.4 Discussion

4.4.1 Nearest-Neighbour Effects

Regional variation in base composition has largely been inferred to be due to regional variation in mutation patterns (Filipski, 1987; Sharp and Lloyd, 1993). One possible cause of mutation pattern variation is that different regions of the genome are replicated at different times (Wolfe *et al.*, 1989; Eyre-Walker, 1992). A second possibility is raised by the discovery that the genome is partitioned into distinct replicational and transcriptional domains in the nucleus during S-phase (Wei *et al.*, 1998). If these domains are set up anew during each cell cycle, then neighbouring genes may tend to experience similar chemical environments during their evolution, whereas genes that are not close together may not have this shared history.

A third possibility is that regional mutation patterns reflect differences in the local frequency of recombination. Recombination involves DNA repair, a process known to be biased toward G+C-richness in mammals (Brown and Jiricny, 1988). It might therefore be expected that recombination hot spots would have elevated G+C content, and this is true at least for chromosome III, where hot spots for double-strand breaks (DSBs) coincide with G+C-rich areas of the chromosome (Baudat and Nicolas, 1997). DSBs tend to be located in intergenic sequences, so that the ensuing DNA repair may affect the ORFs on each side of the DSB and thus contribute to the correlation of GC3s in the neighbouring genes. Because the nearest neighbour correlation largely associated with neighbouring gene-pairs with high values of G+C content this third possibility is by far the most likely. The frequency distribution of G+C content of coding regions is skewed towards high values of G+C content. This skew is absent (or possibly reversed) in non-coding regions. If non-coding regions are under greater constraint than silent sites in coding regions as suggested by Sharp and Lloyd (1993) then the shape of this distribution and the absence of correlation between the G+C content of neighbouring non-coding regions are likely to be connected and related to the greater resistance in non-coding regions to mutational pressure from DSB repair.

4.4.2 Difference in G+C content between chromosomes

The disproportionate concentration of high-GC3s ORFs on shorter chromosomes gives rise to the negative correlation between chromosome length and chromosome G+C content. A negative correlation has also been reported for the relationship between chromosome length and genetic map length per kilobase (Mortimer *et al.*, 1992). It is a requirement for meiosis that there be at least one chiasma per chromosome, and this results in a higher chiasma density and a longer map length per kilobase on shorter chromosomes. High chiasmata density is associated with high G+C content in humans (Ikemura and Wada, 1991), so the differences in recombination rates per kilobase between chromosomes could cause the observed phenomenon. The relationship between chromosome length might therefore have been anticipated. The relationship is produced by ORFs of high GC3s content and not reflected in a difference in the modal GC3s content. Baudat and Nicolas (1997) have shown that DSBs occur in specific regions along chromosome III and, in general the G+C content of just a subset of the ORFs in the genome may be raised by DSBs.

4.4.3 The Paradox of Chromosome III

Chromosome III has consistently shown the strongest clustering effects under the objective criteria that have been applied in this study. No other chromosome displays such pronounced regional variation in GC3s (Fig. 4.1) or such large autocorrelations at either short or long distances (Fig. 4.3). To attempt to explain why chromosome III stands out it is important to establish what differentiates it from the other chromosomes. The chromosome III sequence was published in 1992 and is widely thought to be less accurate than other yeast chromosome sequences. Frameshift sequence errors could increase GC3s values for some genes, but such errors seem unlikely to produce the strong GC3s correlations between neighbouring genes seen in Fig. 4.3.

It could also be that the source of the sequence of chromosome III is derived from yeast species other than *S. cerevisiae*. The laboratory yeast strain (S288C) whose genome was sequenced is derived largely from a single natural isolate of *S. cerevisiae* (EM93), but small fractions of its genome (probably less than 5% in total) come from two other species: "*S. microellipsoides*" strain NRRL-210 (which is possibly *Zygosaccharomyces microellipsoideus*) and the lager yeast *Saccharomyces carlsbergensis* (Mortimer and Johnston, 1986). At present there is no information about the location of this "foreign" DNA and no reason to suspect that there is more of it on chromosome III than elsewhere and we do not consider it a likely explanation of the distinguishing characteristics of this chromosome.

Possibly the most significant distinguishing feature of chromosome III is the fact that it contains the mating-type loci. These comprise the *MAT* locus and the two silent mating-type cassettes (*HML* and *HMR*) located near the two ends of the chromosome (see Chapter 1). Homothallic yeast cells can switch mating type efficiently, with **a**-type cells selecting the *HML* cassette (which contains a silent copy of the α gene) as a donor 80% of the time (Haber, 1998). Donor preference in *MAT***a** cells appears to be achieved by the activation of a >40 kb region on the left arm of chromosome III for recombination (Wu and Haber, 1995). The recombination enhancer on the left arm is silenced in *MAT*α cells and the left arm and a part of the right arm of chromosome III become unavailable for use as donors for the *MAT* locus (Wu and Haber, 1995). Because this system of donor preference regulation involves altering the rate of recombination in large regions in chromosome III it may have been the cause of the regions of very high G+C content in either arm of this chromosome. The recombination enhancer is likely to be silenced in the *MAT*α/*MAT***a** diploids, concentrating most of the recombination during meoisis in the right arm of the chromosome. The high values of G+C content in the right arm of the chromosome could be explained by the requirement of recombination during meoisis. The required recombination could be restricted to a relatively small region on the right of chromosome III by the unavailability of the rest of the chromosome for recombination when the recombination enhancer is silenced.

Another consequence of the location of the mating-type loci on chromosome III may be a greater stability of this chromosome against rearrangement. If one of the silent cassettes is relocated to a different chromosome, mating-type switching still occurs, but its efficiency is greatly reduced, because the bias in donor selection is lost (Weiler *et al.*, 1995). It is likely, therefore, that there is selective pressure to preserve mating-type switching as an intrachromosomal reaction, and so to keep most of chromosome III (between *HML* and *HMR*) intact. If chromosome III has been largely free from structural disruption, then its pattern of GC3s variation may represent a fundamental pattern of mutation. Other chromosomes, which are not so constrained, may never be able to reveal such clear trends (Bradnam *et al.*, 1999). However there is no reason to think that correlation of neighbouring ORFs takes place on a time-scale comparable to the time-scale of chromosomal rearrangements. In fact regions undisrupted since genome duplication showed no increase in nearest neighbour correlation (see section 4.3).

4.4.4 Analogy with isochores in mammals

The observed features of G+C content in *S. cerevisiae* show some analogy to isochores in mammals. Isochores appear to have been produced in mammals through the increase of G+C content of regions that are now G+C-rich (Bernardi *et al.*, 1985) and are associated with increased rates of recombination (Bernardi, 1993). Similarly compositional heterogeneity along yeast chromosomes is likely to be the result of strong correlation in G+C content between G+C-rich genes, probably caused by DSB repair during recombination. It is not clear whether increased rate of recombination is one of the causes of heavy isochores in mammals. Heavy isochores are gene-rich in mammals, but there appears to be no relationship between coding density and G+C content in yeast (Bradnam *et al.*, 1999). There is far less variation in coding density in the more compact genome of *S. cerevisiae* than in mammalian genomes and heterogeneity of coding density is not as significant as it is in mammals.

A striking similarity between *S. cerevisiae* ORFs of high GC3s content and mammalian isochores in the relationship between gene length and G+C content. Mammalian genes in G+C-poor regions of the genome tend to be longer than average. Genes that are longer than the average (~500aa) are 1.9 times more frequent in G+C-poor isochores than in the rest of the mammalian genome (Duret *et al.*, 1995; Gardiner, 1996). In *S. cerevisiae* 50% of the G+C-poor (lower than median GC3s, i.e. < 36%) ORFs are longer than average (>470aa) compared to 27% of G+C-rich ORFs (personal observation, unpublished). Duret *et al.* (1995) refer to an increased risk of gene breakage by ectopic recombination with increasing gene length as a possible cause for this relationship in mammals. The role of recombination in producing G+C-rich genes is further reinforced if this is the cause of the analogous inverse correlation between G+C content and ORF length in yeast.

4.4.5 Comparison with *C. albicans*

Comparison between *S. cerevisiae* and *C. albicans* indicates fundamental differences in the organisation of the two genomes. Further investigation of the differences will be made easier as more of the available sequence from *C. albicans* is annotated. *C. albicans* does not make use of the universal genetic code (Santos and Tuite, 1995), further complicating the comparison of G+C content between the two species. Preliminary examination of the length of annotated *C. albicans* sequences from the EMBL database did not indicate any correlation between gene length and nucleotide content. It will be interesting to investigate fully whether this correlation is absent in *C. albicans*. The existence of a G+C content correlation between physically close genes in *C. albicans* allows us to predict that similar "peaks" of GC3s content will appear if the equivalent analysis is performed on *C. albicans* chromosomes using sliding windows.

Chapter 5

Prevalence of Small Inversions in Ascomycete Gene Order Evolution

5.1 Introduction

A sequencing project for the genome of *Candida albicans* was launched at Stanford University in October 1996 using a whole-genome shotgun sequencing approach. *C. albicans* is an asexual pleomorphic fungus with a 32 Mb diploid (2 x 16 Mb) genome (Ahearn, 1998), belonging to a clade of anamorphic yeasts (Meyer *et al.*, 1998) that share a deviation from the universal genetic code (CUG-leu replaced by CUG-ser; Pesole *et al.*, 1995). It is a common human pathogen that exists commensally in the gastrointestinal and urogenital tracts of 40 to 60% of the adult human population and is ranked third most commonly encountered isolate in clinical studies of infectious diseases (Magee, 1998). The time of divergence between *C. albicans* and *S. cerevisiae* is greater than the time since the radiation of the mammalian orders and has been alternatively estimated at approximately 300 million years (Pesole *et al.*, 1995) and approximately 140 million years (Berbee and Taylor, 1993).

The initial aim of the sequencing project was to achieve 1.5X coverage by summer 1998 and this target was subsequently extended to 10X coverage. By July 1999 1631 contigs greater than 2 kb in size had been assembled by the group at Stanford accounting for 14.9 Mb of the ~16 Mb haploid genome of *C. albicans*. The Stanford contigs had not been annotated but it was possible to approximate the position of a large number of probable *C. albicans* genes with putative homology to *Saccharomyces cerevisiae* genes using similarity searches. Gene density in *C. albicans* is approximately 0.4 – 0.5 kb^{-1} (based on the annotated cosmids in Fig. 5.5) and most of the contigs include sequence from two or more *C. albicans* genes. This allows the possibility of determining local gene order and orientations of a large number of genes in *C. albicans*, and provides the opportunity to study in detail the disruption of microcolinearity in distantly related eukaryote genomes for the first time.

Gene order change occurs by means of reciprocal translocation, inversion and transposition. Many of the methods that have been devised for the study of chromosomal rearrangements (see Chapter 1) have concentrated on using statistical methods based on partial mapping data and synteny data available for mammals (e.g. Nadeau and Taylor, 1984; Sankoff and Goldstein, 1989; Sankoff and Ferretti, 1996). Furthermore because available sequences were distributed randomly in the genomes studied, only rearrangements involving large regions of chromosomes were detected. However it is clear from sets of contiguous genes from several species and from the gene order of duplicates from the completely sequenced genome of *S. cerevisiae* that small disruptions of gene order have also occurred, involving just one or several adjacent genes (see below). We have used the extensive and detailed data, available for the first time for two eukaryotes, to investigate the extent of small rearrangements of gene order between the genomes of *Saccharomyces cerevisiae* and *Candida albicans* and propose that small rearrangements are a key feature of eukaryote gene order evolution and should be taken into account in comparative mapping between distantly related eukaryotes.

5.1.1 Review

Detailed examination of the gene order in mapped duplicated regions in yeast reveals that a minority of genes have opposite orientation to what is predicted by the hypothesis of genome duplication followed by reciprocal translocation (Wolfe and Shields, 1997; Seoighe and Wolfe, 1998). The regions of opposite transcriptional orientation may be large and comprise a whole duplicated block or smaller, involving just a single gene or a small number of adjacent genes. In the most recent version of the map of duplicated regions in *S. cerevisiae* 11 pairs of duplicated genes showed evidence of small inversions in *S. cerevisiae* since duplication of the whole genome (see Chapter 3). This is over 3% of the genes identified as duplicates in the genome. Studies of gene order conservation in ascomycete species have been carried out by comparing the *S. cerevisiae* genome sequence to DNA from related species, either by randomly sequencing both ends of small clones (Altmann-Jöhl and Philippsen, 1996; Hartung *et al.*, 1998; Ozier-Kalogeropoulos *et al.*, 1998) or using existing EMBL database sequences (Keogh *et al.*, 1998). Only one case of conserved gene order and orientation has so far been reported between *C. albicans* and *S. cerevisiae* (*STE6 – UBA1*; Raymond *et al.*, 1998) whereas there are three reported cases of gene pairs that have remained adjacent in both species but one gene has been inverted (*RAD16 – LYS2, NFS1 – LEU2*, and *RPS31 – SEC10*; Roig *et al.*, 1997; Hartung *et al.*, 1998; Keogh *et al.*, 1998; Plant *et al.*, 1998; Suvarna *et al.*, 1998). A much larger data set was required to test whether this apparent high rate of inversion was a general phenomenon between the genomes.

5.2 Data and Methods

Contigs were downloaded by anonymous ftp from ftp//cycle.stanford.edu/pub/projects/candida/ on July 15 1999. Analysis was performed using programmes written in the PERL programming language. Sequence similarity between the contigs and *S. cerevisiae* protein sequences was based on TBLASTN (Altschul *et al.*, 1997) using the **seg** filter (Wootton and Federhen, 1996) and a cut off E-value of 10^{-10}. Re-analysis of the data using different cut-off values (10^{-6}, 10^{-20}) did not significantly alter the results. Gene locations on the contigs from the Stanford data were based solely on similarity to *S. cerevisiae* genes and frequently involved extrapolation from regions of high similarity to determine the endpoints of the *C. albicans* genes.

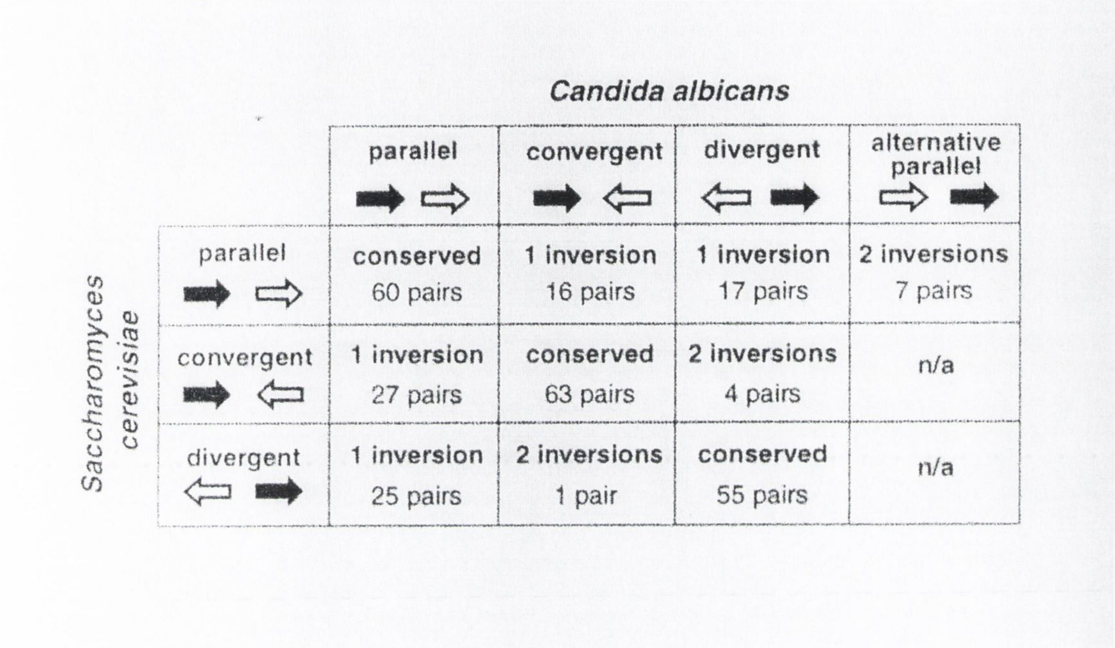
Complete annotation was available from GenBank for the cosmids shown in Fig. 5.5 that were sequenced at the Sanger Centre. The cut-off value used for the contig data was not applied and *C. albicans* genes without *S. cerevisiae* homologues, ignored in the contig data, are shown for the annotated cosmids.

The results of a TBLASTN similarity search comparing all yeast proteins to a database constructed from the contigs downloaded from Stanford were analysed using a programme written in PERL. The programme mapped the *S. cerevisiae* proteins onto the *C. albicans* contigs and selected the protein with the lowest E-value from each set of overlapping high-scoring hits. By this method orthologues of *S. cerevisiae* genes were delineated approximately on the *C. albicans* contigs. Contigs containing just one such putative orthologue were discarded. Consecutive *S. cerevisiae* orthologues were taken to represent adjacent genes in *C. albicans*.

5.3 Results

5.3.1 The extent of single-gene inversion

The contigs contained 2998 pairs of genes that appear to be adjacent in *C. albicans* (*i.e.*, either they are adjacent, or any intervening genes do not have *S. cerevisiae* orthologues). For 275 pairs (9%), the *S. cerevisiae* orthologues are also adjacent. Despite remaining as neighbours, 97 of these pairs (35%) have different gene orientation or order in the two species. Eighty-five pairs can be explained by inversions of one gene, and 12 pairs require two inversions each.



		Candida albicans			
		parallel ➡ ⇨	convergent ➡ ⇦	divergent ⇦ ➡	alternative parallel ⇨ ➡
Saccharomyces cerevisiae	parallel ➡ ⇨	conserved 60 pairs	1 inversion 16 pairs	1 inversion 17 pairs	2 inversions 7 pairs
	convergent ➡ ⇦	1 inversion 27 pairs	conserved 63 pairs	2 inversions 4 pairs	n/a
	divergent ⇦ ➡	1 inversion 25 pairs	2 inversions 1 pair	conserved 55 pairs	n/a

Figure 5.1 Order and orientation relationships between 275 gene pairs that are adjacent in both *S. cerevisiae* and *C. albicans*. All ten possible relationships between two adjacent genes are shown, with the number of inversions needed to convert any combination into any other. The names of gene pairs in each category are listed at http://acer.gen.tcd.ie/~khwolfe/candida. The categories labelled as "2 inversions" could also be explained by one gene leapfrogging over the other, but we consider this unlikely.

The four states of relative orientation possible for two genes for which the order is known are show in Fig. 5.2 as well as the transitions that are possible between the states by single-gene inversion. The four states have been labelled A, A`, B, B`. It is clear that all transitions between class A and B are allowed but transitions within the class A or B can not be made with a single inversion. If, in the course of evolution, a gene-pair in state A changes relative orientation then it must change to state B or B`. Another change in relative orientation must return the pair to state A or A`, with equal probability. The probability that the pair is in the original state (A) after 2 transitions is exactly 0.5.

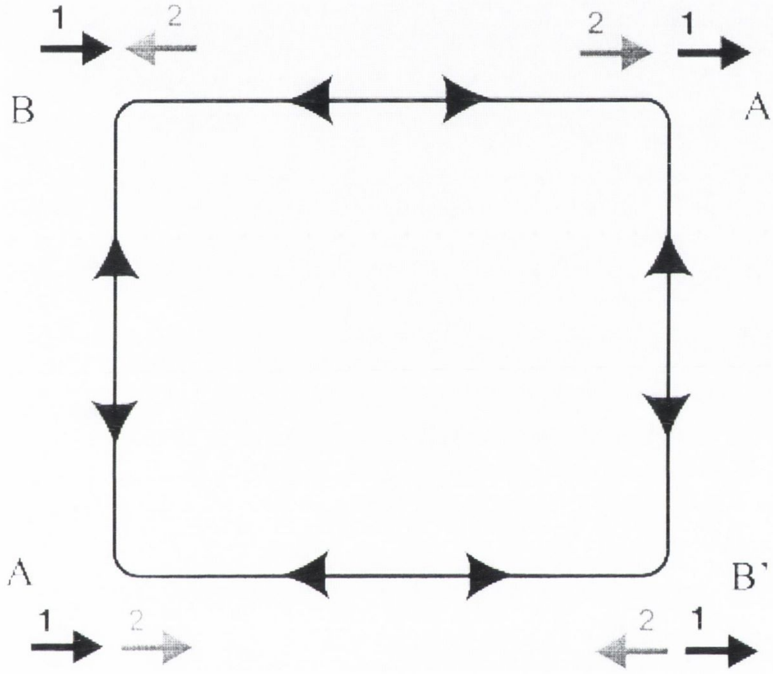


Figure 5.2 The four distinct relative gene orientations of an adjacent pair of genes, labelled A, A`, B, B`. Arrows along the rectangle show the possible transitions between the different states using a single inversion of one gene. Transitions across the diagonal can not be achieved using a single inversion.

In the current study we are interested in small inversions that have taken place since the divergence of two species. Since states A and B are symmetric no information is lost in a model if we assume that all of the inversions are taking place in one of the species.

We denote by $P_0(t)$, the number of gene-pairs that remain adjacent and that have not undergone a transition from their original relative orientation state, $P_1(t)$, the number of adjacent gene-pairs that have undergone exactly 1 transition etc. The number of adjacent gene-pairs that undergo a transition from state A to state B in a period of time, Δt, is likely to be proportional to Δt, as well as to the number of gene-pairs in state A. Therefore:

$$dP_0\,(t) = -\lambda P_0(t)dt$$

$$P_0(t) = Ce^{-\lambda t}$$

At time $t = 0$ all of the gene-pairs are in state P_0. Therefore $P_0(t)= Te^{-\lambda t}$, where T is the total number of gene-pairs in the study and λ is some rate constant.

$P_1(t)$ decays in a similar manner to $P_0(t)$ except that $P_1(t)$ also has a positive term (corresponding to the gene-pairs that decay from $P_0(t)$).

$$dP_1\,(t) = +\lambda P_0(t)dt - \lambda P_1(t)dt$$

$$= \lambda\; Te^{-\lambda t}\,dt - \lambda P_1(t)dt$$

This leads to the ordinary differential equation ${dP_1}/_{dt} = \lambda Te^{-\lambda t} - \lambda P_1(t)$ which is a typical differential equation for populations along a decay chain. The solution to the equation is $P_1(t) = T\lambda te^{-\lambda t} + Ce^{-\lambda t}$.

Since $P_1(0) = 0$ $C = 0$ and the solution becomes $P_1(t) = T\lambda te^{-\lambda t}$.

In the *C. albicans* data there are 275 pairs of genes that have remained adjacent. Of the adjacent gene-pairs 178 have the same relative orientation as the corresponding pair in *S. cerevisiae*, and 85 pairs require exactly one inversion to match the relative orientation of the corresponding pair in *S. cerevisiae* (Fig 5.1). We make the assumption that the number of gene-pairs that have undergone 2 inversions is small compared to the number that has not been inverted. Then $P_0(t) = 178$ and $P_1(t) = 85$. Since $\frac{P_0(t)}{P_1(t)} = \lambda t$, $\lambda t \approx 0.48$. The model then predicts that $P_0(t) \approx 170$, $P_1(t) \approx 82$ and $P_2(t) \approx 23$. Because half of the gene-pairs that undergo two inversions should be returned to the original state and the other half should be in a state requiring two inversions to return to the original state the model predicts that there should be 12 gene-pairs requiring two inversions to be returned to the original state, which is in good agreement with the observed population of this category.

From these observations, i.e. 128 ($P_1 + 2 \times P_2$) single gene inversions among 275 intergenic links, we estimate that the total number of single-gene inversions that have occurred in the genomes of *C. albicans* and *S. cerevisiae* since speciation to be about 1400 (= 6000 genes in the genome x (128/275) / 2 orientation conservations broken per inversion).

The set of 275 adjacent pairs includes one run of four genes, and 21 runs of three genes, that have conserved gene order in the two species. Among these, 13 examples of apparent single-gene inversions are seen (Fig. 5.3). The most dramatic example is the cluster *SLU7-RRP1-SSS1*, where the order is conserved but all three genes have reversed orientations. This could be explained either by three independent single-gene inversions, or by two short-distance transpositions, both of which seem quite improbable.

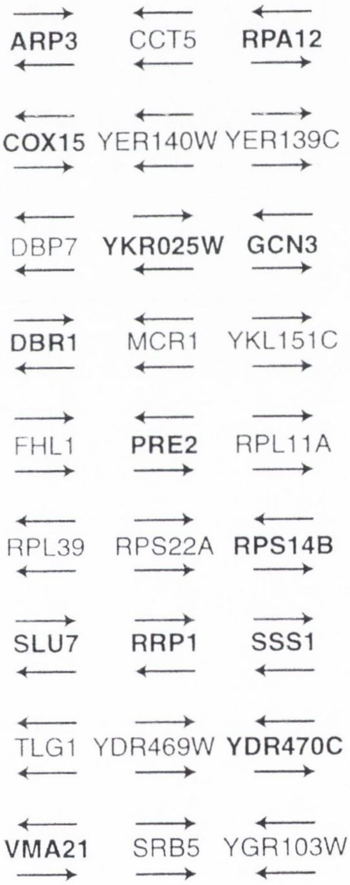


Figure 5.3 Examples of single-gene inversions. The three genes in each set are adjacent (ignoring any *C. albicans* genes without homologues) and in the same order in the two species. Directions of transcription in *S. cerevisiae* and *C. albicans* are shown above and below gene names, respectively.

5.3.2 Adjacent genes in *C. albicans* separated by a small number of genes in *S. cerevisiae*

Other pairs of adjacent *C. albicans* genes have *S. cerevisiae* orthologues that are physically close to each other but are not immediate neighbours (Fig. 5.4). The Stanford contig data includes 97 pairs of adjacent *C. albicans* genes whose *S. cerevisiae* orthologues are separated by 1 – 5 intervening genes. Gene orientation and relative order are conserved in 28 of these pairs, which is only slightly more than the 24.25 expected by chance. This suggests that multi-gene inversions may have occurred, moving genes over short distances. To study this further we examined some *C. albicans* cosmids (Tait *et al.*, 1997) that were completely sequenced at the Sanger Centre (Fig. 5.5). These sequence comparisons point to numerous rearrangements, both inter-chromosomal (translocations) and intra-chromosomal (small inversions). Most of the long *C. albicans* sequences contain small clusters of genes whose *S. cerevisiae* orthologues are also physically clustered (Fig. 5.5). These clusters are generally shorter than 10 genes in *C. albicans* and they are often interspersed with genes from other *S. cerevisiae* chromosomes. The ends of the clusters probably correspond approximately to sites of chromosomal translocations (Wolfe and Shields, 1997; Keogh *et al.*, 1998; Seoighe and Wolfe, 1998). In some cases a cluster of genes in *C. albicans* is related to two *S. cerevisiae* genomic regions (blocks) that are paired by whole-genome duplication in the *S. cerevisiae* lineage (Wolfe and Shields, 1997), as predicted by the model of genome duplication followed by chromosomal rearrangement (Keogh *et al.*, 1998; Seoighe and Wolfe, 1998).

The relationships shown in Fig. 5.5 comprise 32 orthologous genes and at least 11 independent inversions. It is not possible to estimate the exact sizes of these inversions (*i.e.*, the numbers of genes involved) because, in all cases, the genes immediately upstream and downstream of the inverted ones are different in the two species. For example, the inversion of *YLR423C* in cosmid Ca49C10 might have included some of the four genes downstream of it in *S. cerevisiae*. However, the inversions must be relatively small because gene order is conserved at a coarser level (e.g., *YLR423C* is in-between *YLR418C* and *YLR424W* in both species). The locations of *S. cerevisiae* homologues on *C. albicans* contigs in the July 1999 release from Stanford can be assessed using the web-interface programme at http://biotech.bio.tcd.ie/~ferdia/Candida.html. Some of the contigs are larger than the cosmids from the Sanger Centre and also show several interesting examples of small inversions.

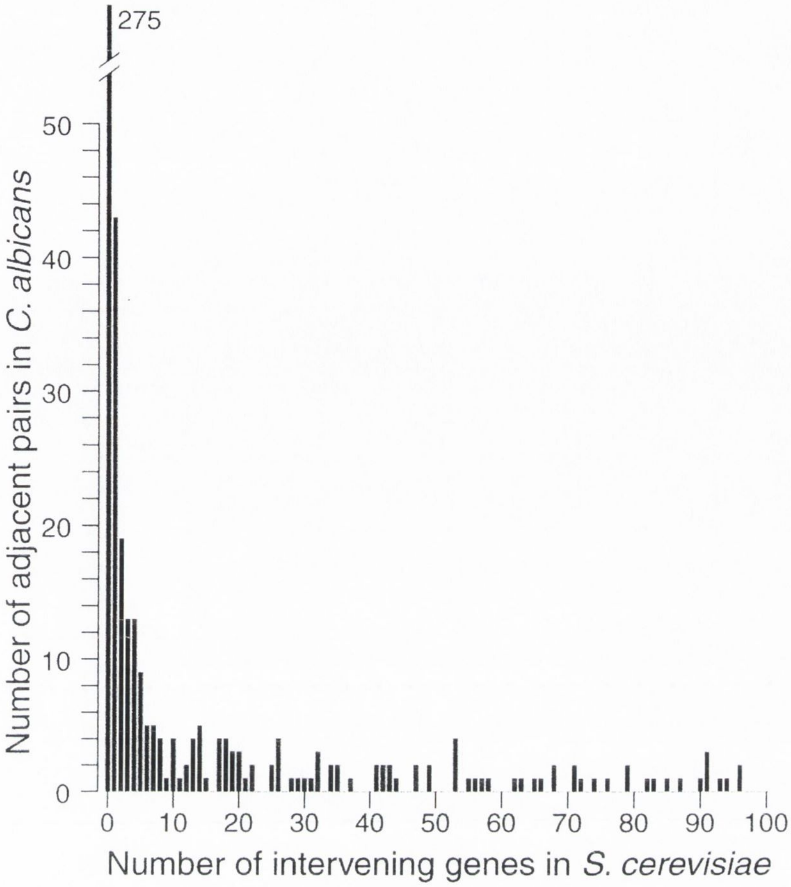


Figure 5.4 Histogram showing the distance apart in *S. cerevisiae* of the orthologues of gene pairs that are adjacent in *C. albicans*.

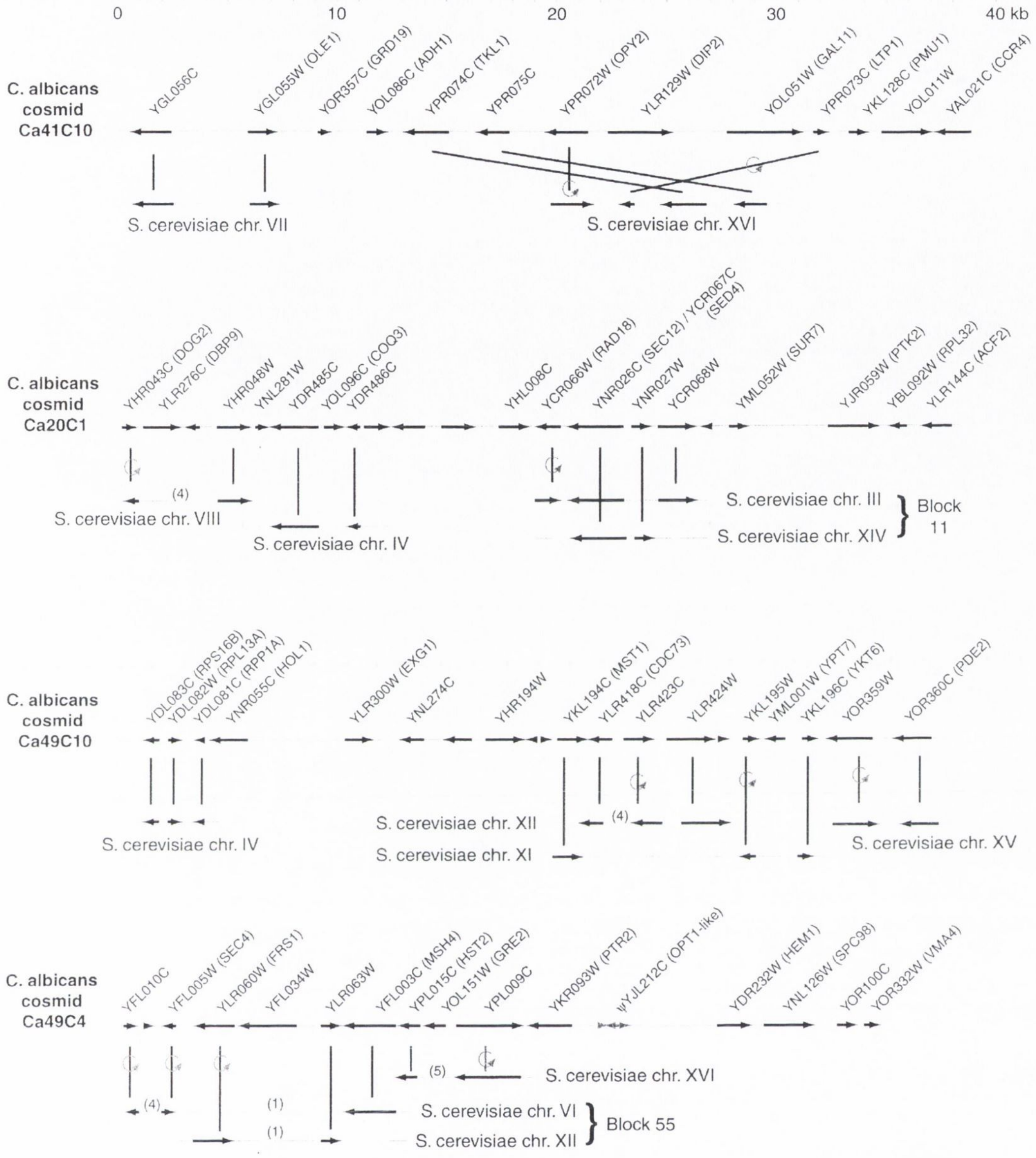


Figure 5.5 Relationships between *C. albicans* and *S. cerevisiae* chromosomal regions. Vertical lines connect orthologous genes. Curved arrows indicate genes with inverted orientations. *C. albicans* genes are named after their *S. cerevisiae* orthologues; unnamed genes have no close relative in *S. cerevisiae*. Numbers in parentheses indicate numbers of intervening genes in *S. cerevisiae* that are not shown. *S. cerevisiae* regions in or near duplicated blocks are labelled. The scale at the top refers to *C. albicans* only.

The conservation of small neighbourhoods of genes, without absolute conservation of order or orientation, suggests that small DNA inversions have contributed significantly to the evolution of ascomycete genomes. A further example is seen in cosmid Ca49C4 (Fig. 5.5), which contains a pseudogene related both to the *C. albicans* oligopeptide transporter gene *OPT1* (Lubkowitz *et al.*, 1997) and to its *S. cerevisiae* homologue *YJL212C*. The pseudogene has 98% DNA sequence identity over 2 kb to part of *OPT1,* but a 0.3 kb internal segment has been inverted relative to *OPT1* and other members of this gene family.

5.3.3 Relative rates of intra- versus inter-chromosomal rearrangements

Small rearrangements keep genes within a local neighbourhood, so we can use the *C. albicans/S. cerevisiae* comparisons to estimate the rate of small rearrangements relative to large rearrangements (translocations, larger inversions, and long-distance transpositions). Even if there had been no other chromosomal rearrangements, we would expect about half of the links between immediate neighbours in *S. cerevisiae* and *C. albicans* to have been broken by the process of random gene loss due to differential silencing after genome duplication in the *S. cerevisiae* lineage (Lundin, 1993; Keogh *et al.*, 1998). The remaining breaks are the combined result of inversions, translocations and transpositions. The fraction of links that has been conserved is under 10%, but this fraction has been reduced by a factor of two by genome duplication in *S. cerevisiae*. Consequently, chromosomal rearrangements are responsible for breaking over 80% of the links between neighbours. Assuming that breakpoints are made randomly with a Poisson distribution (as in the Jukes-Cantor multiple hits correction formula), this implies that there have been an average of 1.6

breaks per link, or approximately 10,000 breakpoints in total since speciation. This argument assumes that the *S. cerevisiae* genome duplication occurred recently, but an identical conclusion is reached if the genome duplication is assumed to have occurred shortly after speciation. It also assumes that no other genome duplications have occurred in either lineage.

Statistical methods have previously been developed to estimate relative numbers of intra- and inter-chromosomal rearrangements between species (Erlich *et al.*, 1997), but these methods are not adaptable to the current problem because the kind of data being considered is local (the contigs are short relative to chromosomes) and because the number of rearrangements is close to saturation. It is problematic to model the small inversions directly because not enough is know about their size distribution. Instead, to model the combined processes of large and small chromosomal rearrangements, adjacent genes in *C. albicans* having orthologues on the same chromosome in *S. cerevisiae* were divided into two categories: gene pairs that are also adjacent in *S. cerevisiae* (state A); and gene pairs that are "near-neighbours" (syntenous but separated by a small number of genes) in *S. cerevisiae* (state B).

The number of gene pairs in the sequenced sample that are in state A is $\mathrm{P_A}$. In a time interval Δt the change in population of state A is

$$\Delta \mathrm{P_A} = \mathrm{P_A}(L+S)\Delta t.....(1)$$

where L and S are the rates at which single intergenic links are broken by large and small rearrangements, respectively. Let I be the mean number of intervening genes for gene pairs that are near-neighbours in *S. cerevisiae*, so that

$I+1$ is the mean number of intervening links. If we make the assumption that the average separation of this category of gene pair has been similar throughout the evolutionary history then

$$\Delta \mathrm{P}_{\mathrm{B}} = S\mathrm{P}_{\mathrm{A}}\Delta t - (I+1)L\mathrm{P}_{\mathrm{B}}\Delta t.....(2)$$

This assumption is justified because gene pairs in state B are unlikely to drift too far apart before their linkage is broken by a translocation. Large rearrangements (translocations) are taken to be the only way in which gene pairs leave state B because the number of gene pairs that are syntenous but not near-neighbours is small (Fig. 5.4). Using the above assumptions the problem has again been transformed into the problem of determining the populations along a decay chain, similar to the problem in the previous section.

Equations **1** and **2** above can be treated as differential equations and solved using an integrating factor ($F = e^{(I+1)Lt}$) to obtain

$$\mathrm{P}_{\mathrm{A}} = C_1 e^{-(L+S)t}.....(3)$$

$$\mathrm{P}_{\mathrm{B}} = C_2 e^{-(I+1)Lt} + \frac{SC_1}{IL-S} e^{-(L+S)t}.....(4)$$

At time 0 $P_A = J/2$, where J is the number of gene pairs in the sample that are adjacent in *C. albicans* and have orthologues in *S. cerevisiae*, because *S. cerevisiae* has undergone genome duplication followed by differential silencing.

Therefore $C_1 = J/2$. At time 0 $P_B = 0$, therefore $C_2 = -\frac{1}{2}\frac{SJ}{(IL-S)}$.

Equations **3** and **4** provide an estimate of the proportion of all rearrangements that are small ($S / S{+}L$), given values for the number of conserved adjacent gene pairs (P_B , which is 275), the number of pairs that are adjacent in *C. albicans* but near-neighbours in S. cerevisiae$\left(P_B\right)$ and the average spacing between near-neighbours (I). The values of P_B and I can be calculated from the data in Fig 5.4 but depend on the maximum number of intervening genes that is permitted in the definition of near-neighbours ($I_{\max}$). In Fig. 5.4 there appears to be an excess of conserved linkages over short distances, up to a limit of at least 5 intervening genes and possibly as far as 20. The relationship between the estimated proportion of small rearrangements and $I_{\max}$ is shown in Fig. 5.6. Allowing a maximum of 5 genes between near-neighbours, 41% of broken links are attributed to small rearrangements. This increases to 68% for $I_{\max} = 20$ genes. These results suggest that approximately equal numbers of linkages have been broken by small and large rearrangements.

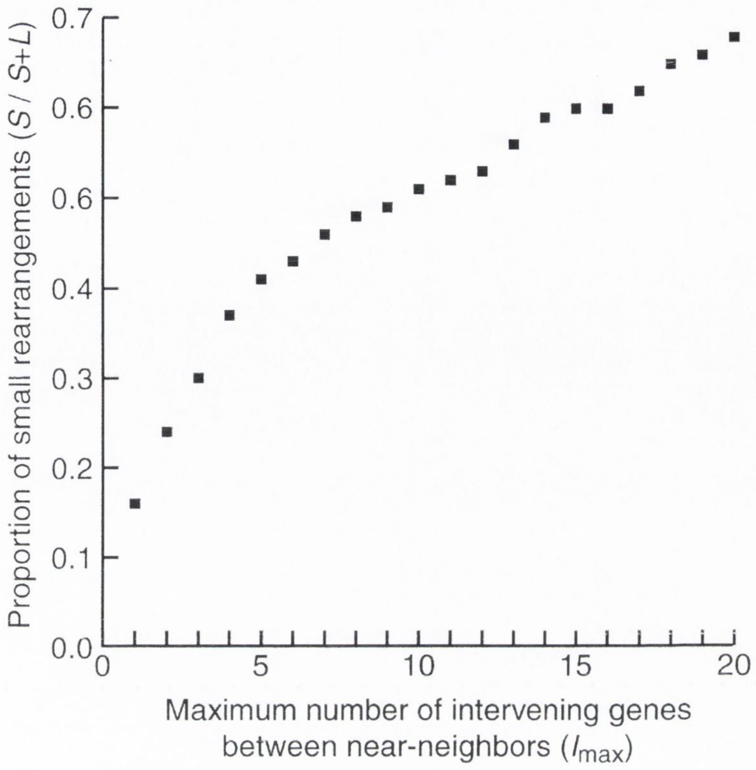


Figure 5.6 Relationship between the maximum permitted number of intervening genes (I_{max}) between near-neighbours (gene pairs in state B), and the estimate of the proportion of rearrangements that are small ($S / S + L$). Calculated numerically from equations **3** and **4** using data for P_B at different values of I_{max} from Fig. 5.4.

A limit of $I_{\max} = 5$ was also suggested by an experiment in which we compared the number of adjacent pairs in *C. albicans* whose homologues are syntenous in *S. cerevisiae*, to those whose homologues are located on specific pairs of different chromosomes, as a way of estimating the "background" level of random gene associations in Fig. 5.4. Chromosomes were paired in order of size to reduce the effect of size differences between chromosomes on the result.

5.4 Discussion

Our results suggest that successive random small inversions frequently cause a gene's chromosomal position and orientation to drift during its evolution. This process would alter gene order and orientation without moving any genes very far from their starting points (although the probability of interchromosomal rearrangement breaking the synteny of the pair is increased by the increased distance between the pair). It would also tend to blur the endpoints of interchromosomal translocations. The mechanism by which small inversions occur is unknown, and our data are uninformative in this regard because intergenic sequences are not conserved between *C. albicans* and *S. cerevisiae*. Our results also suggest that gene order in yeasts is relatively unconstrained by natural selection. The orientations of some pairs of adjacent genes, particularly those that are transcribed divergently from a shared regulatory region (such as the histone pair *HTA1-HTB1*) may be under selection, but the high frequency of rearrangement indicates that this type of constraint is the exception rather than the rule.

In our analysis we made an arbitrary distinction between small and large rearrangements, using a limit of 5 or 20 intervening genes based on inspection of Fig. 5.4. The size distribution of inversions during evolution is unknown but it seems likely that there is a skewed distribution with a bias towards smaller sizes, either due to mechanistic reasons or due to natural selection. This is illustrated by the large number of single-gene inversions inferred (approximately 1400). A more accurate description of the size distribution is clearly needed but will require comparisons between more closely related yeast species. However the very fact that results presented here indicate that there is a bias towards small inversions has implications for the statistical methods that have been developed to model gene order evolution through chromosomal rearrangement. Most of these methods rely explicitly on the assumption that the distribution of breakpoints of gene order conservation is random (e.g. Nadeau and Taylor, 1984; Nadeau, 1989; Sankoff *et al.*, 1997b), although Nadeau and Sankoff have acknowledged that recent gene order information from detailed comparative maps casts doubt on the random distribution hypothesis (Nadeau and Sankoff, 1998). The distribution of lengths of conserved segments discovered between mouse and man is in good agreement with the model of random distribution of breakpoints (Nadeau, 1989; Nadeau and Sankoff, 1998). However, studies carried out up to now have relied on relatively sparse genetic maps. As detailed gene order and orientation data emerge for large contiguous regions of the mouse and human genomes estimates of the number and size of undiscovered conserved segments (Nadeau and Sankoff, 1998) may prove to be incorrect, if the human and mouse genomes have also been affected to some extent by an excess of small inversions over what is predicted by the hypothesis of random distribution of breakpoints.

There is evidence for an excess of small rearrangements from several vertebrate species. By mapping a large region around the bovine *mh* locus Stonestegard *et al.* (1998) have increased the resolution of conserved syntenies between bovine chromosome 2 and human chromosome 2 and pointed to the existence of at least one small inversion. They used the term "microrearrangements" to describe these additional gene order changes in this syntenous region. Yang and Womack

(1998) have reported gene order rearrangements within the entirely syntenous chromosomes human 17 and bovine 19 and refer to the need to acquire a better understanding of rearrangements of gene order for effective transfer of mapping information between the human and bovine genomes. Comparative mapping of the DiGeorge syndrome region between mouse and man shows three gene order rearrangements within twenty genes in this region (Botta *et al.*, 1997; Lindsay *et al.*, 1999). The probability of having six breakpoints within a region of this size by the random breakpoints model is vanishingly small.

Several examples of small gene order rearrangements within conserved syntenies have been observed between *Fugu* and human or *Fugu* and mouse (Armes *et al.*, 1997; Gilley and Fried, 1999; Kehrer-Sawatzki *et al.*, 1999), with one example in the Surfeit region of gene order conservation without conserved orientation. Gilley and Fried recently proposed that small gene order differences between *Fugu* and human may have been caused by inversions (Gilley and Fried, 1999). There have also been several reports of conserved synteny, but not gene order, between *Caenorhabditis elegans* and *Drosophila* or mammals (Ruddle *et al.*, 1994; Trachtulec *et al.*, 1997; Pebusque *et al.*, 1998; Ruvkun and Hobert, 1998). Moreover, in *C. elegans* some gene families are unevenly distributed among chromosomes, with a statistical excess of within-chromosome duplications (even at large distances) over between-chromosome duplications (Ruvkun and Hobert, 1998; Semple and Wolfe, 1999). This could be caused by duplicate genes arising as tandem repeats (which are common in *C. elegans* but rare in yeast; Goffeau *et al.*, 1997; *Consortium*, 1998) and then dispersing along chromosomes by inversion (Ruvkun and Hobert, 1998). This model is supported by the negative correlation between the physical distances between duplicated gene pairs on the same chromosome in *C. elegans*, and the frequency at which they are found to be in the same transcriptional orientation (Semple and Wolfe, 1999).

Chapter 6

Conclusion

6.1 A model of gene order evolution

Yeast species are unicellular, can be grown on a defined medium and are ideally suited to classical genetic analysis making them some of the most important model organisms for the study of eukaryotic genetics (Goffeau *et al.*, 1996). In addition yeast species are widespread and of great economic and environmental significance (see Chapter 1). This thesis has made use of genomic data from yeast species, to tackle general questions related to eukaryote evolution, using yeast as a model, as well as exploring the implications of the available data for the evolution of the genomes of the yeast species themselves.

The proposal of Wolfe and Shields (1997) that the genome of *Saccharomyces cerevisiae* was duplicated approximately 10^8 years ago is strongly supported by the subsequent analysis of sequence data from related species. Genome duplication, first proposed by Susumo Ohno as a key evolutionary event in several lineages (Ohno, 1970), continues to generate substantial interest in relation to the genomes of several important organisms including human. The hypothesis of genome duplication in *S. cerevisiae* provided the first opportunity to study the effect of genome duplication on the organisation of a eukaryote genome using complete sequence information. More general questions concerning duplicated gene evolution were also made accessible by the large number of genes that were duplicated simultaneously as a part of the whole genome duplication. It is possible to investigate what factors affect whether a duplicated gene is retained and which genes are retained in duplicate. Similar factors may be important in determining gene loss following gene duplication in other eukaryotes. Because the *Saccharomyces* genus includes several species that share a common genome duplication, current *Saccharomyces* sequencing projects (see Chapter 1) will open up the possibility of studying different patterns of gene loss in species descended from the same duplication event. This will provide insight into the rate at which duplicated genes are lost. Similarities and differences in the sets of genes that are retained in duplicate in different species may suggest to what extent the set of genes that are retained is decided by chance or to what extent the fate of duplicate genes is determined by their function and adaptational requirements.

6.1.2 Gene order comparison in distantly related genomes

The comparison of gene order in distantly related species is likely to become more important as map data improves. *Fugu rubripes* has recently been proposed as a model organism for the study of the human genome (Brenner *et* *al.*, 1993). Consequently there has been increasing interest in the comparison of gene order between human and *Fugu*. Recent reports have cited examples of conserved and non-conserved local gene order (e.g. Gilley and Fried, 1999; McLysaght *et al.*, manuscript submitted). Because the genome of the zebrafish contains seven copies of the *HOX* gene-cluster, compared to four in mammals, and pairs of genes orthologous to single mammalian genes, it has been suggested that there has been an additional genome duplication event that occurred in fish after their divergence from other vertebrate lineages (Amores *et al.*, 1998; Gates *et al.*, 1999). If fish are separated from the majority of vertebrates by a duplication of the whole genome then comparison between the genomes of human and *Fugu* or zebrafish will need to take into account the impact of genome duplication on gene order conservation. For this reason the study of duplicated and unduplicated yeast genomes in the preceding chapters should provide a useful model for exploring the impact of genome duplication on comparative gene order evolution.

The impact of small rearrangements

Detailed study of gene order in distantly related species has also given rise to an awareness of small rearrangements, so that sparsely mapped regions that once appeared to be largely conserved have turned out to contain several micro-rearrangements of gene order, often involving just two, or a small number of genes. These small rearrangements of gene order were noticed in the map of duplicated regions in *S. cerevisiae* and their extent was measured here by comparing the genome of *S. cerevisiae* with available sequenced contigs from *Candida albicans* (see Chapter 5; Seoighe *et al.*, in press). The disproportionate number of small gene order rearrangements must be taken into account when the level of gene order rearrangement between species is determined. Preliminary steps towards developing a method of determining the relative number of small and large rearrangements have been taken in this thesis. In the near future comparison of small-scale gene order between diverse species should reveal whether the prevalence of small gene order rearrangements is ubiquitous in eukaryotic genomes or particular to ascomycetes.

6.2 A yeast genome resource

A resource has been provided, and added to in the present work, that should facilitate efforts of the yeast research community to relate the order of genes observed in different yeast species to the genome duplication event in the *S. cerevisiae* lineage. *S. cerevisiae* will, without doubt, remain the point of reference for the study of other yeast genomes for the foreseeable future and the outline that we have provided of duplicated regions should continue to be useful (see Chapter 3). The strategy adopted in the revised version of the yeast map was to maximise the usefulness of the map of duplicated regions by assigning as much as possible of the genome to putative duplicated regions. As the map data increase for related yeast species, particularly unduplicated yeasts of the *Saccharomyces sensu lato* it should be possible to confirm some of the small blocks that were included as "possible" duplicated regions in the revised map. Further revision of the map with the eventual assignment of most of the genome to sister regions should be possible if complete sequence data become available for an unduplicated yeast closely related to *S. cerevisiae* (such as *S. kluyveri* or

Kluyveromyces lactis). Further revision of the map of duplicated regions in *S. cerevisiae* will then be required.

6.2.1 Making use of duplication

The hypothesis of genome duplication was used in the present work to produce an estimate of the number of large chromosomal rearrangements in yeast since genome duplication. Estimates of the extent of chromosomal rearrangement in a single species in a given period of time normally require map data from three related species. Through the adaptation of techniques developed for use with partial map data from different species to the intraspecific genome comparison of an ancient tetraploid, it was possible to tackle this problem with data from *S. cerevisiae* alone. Estimates of the rates of inter-chromosomal rearrangement in yeast have shown that the number of rearrangements per million years appears to be similar in yeast and mammals (see Chapter 2). It will be interesting to investigate whether the rate of large-scale gene order rearrangement is similar in other lineages.

6.2.2 Variation in base composition

As whole chromosomes are sequenced in other species it will also be possible to form an overview of the variation of G+C content along chromosomes in different kinds of organisms. The analysis of G+C content variation along yeast chromosomes and possible causes presented here form a useful contribution to the debate over the nature and causes of compositional heterogeneity along chromosomes that has been observed, in varying forms, in species from mycoplasma to vertebrates. Interesting differences in the nature of this heterogeneity among vertebrate genomes have been explored and differences in patterns of base composition variation among yeast species may soon become apparent (see Chapter 4).

6.3 Future prospects for yeast genome research

In the medium-term the opportunities for whole-genome comparison of closely related and distantly related species are likely to increase. Comparison between *Caenorhabditis elegans* and *S. cerevisiae* has revealed that shared proteins normally fulfil core metabolic functions whereas genes unique to one organism are associated with organism-specific functions or pathways (Chervitz *et al.*, 1998). This type of comparative proteomics approach will become increasingly important as the sequences of other fungal genomes reach completion (e.g. *Schizosaccharomyces pombe*). Ultimately the differences in biology among species must be explicable in terms of differences in their genomes. Biological changes in which differences in the copy number of genes are implicated will provide particularly interesting information about the nature of evolution through gene duplication. Over large time-scales the role of gene duplication in producing larger, more complex genomes from the smaller genomes of simpler forms of life is clear. However it is not always obvious what the role, if any, of more recent duplications has been in organism evolution, and particularly in the differentiation between recently diverged species that are separated by gene duplications. Comparison of the functions of genes from gene families that have been affected by recent duplications in, for example, *Saccharomyces cerevisiae* and *Candida albicans* will be a useful model in this regard. This will be possible in the very near future as the *Candida albicans* sequencing project nears completion and the rapid accumulation of sequence information from other related yeast species continues apace.

References

Genoscope web page. http://www.genoscope.cns.fr/externe/English/Projets/Projet_AR/AR.html

The On-line Medical Dictionary. http://www.graylab.ac.uk/omd/index.html

Sequencing of *Candida albicans* at Stanford's DNA Sequencing and Technology Center on http://www-sequence.stanford.edu/group/candida/

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

Ahn, S., Anderson, J.A., Sorrells, M.E. and Tanksley, S.D.: Homoeologous relationships of rice, wheat and maize chromosomes. Mol Gen Genet 241 (1993) 483-490.

Ahn, S. and Tanksley, S.D.: Comparative linkage maps of the rice and maize genomes. Proc Natl Acad Sci U S A 90 (1993) 7980-7984.

Alexopoulos, C.J., Mims, C.W. and Blackwell, M.: Introductory Mycology, Fourth Edition ed. John Wiley & Sons, INC., New York, 1996.

Altmann-Jöhl, R. and Philippsen, P.: *AgTHR4*, a new selection marker for transformation of the filamentous fungus *Ashbya gossypii*, maps in a four-gene cluster that is conserved between *A. gossypii* and *Saccharomyces cerevisiae*. Mol Gen Genet 250 (1996) 69-80.

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

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

Amores, A., Force, A., Yan, Y.L., Joly, L., Amemiya, C., Fritz, A., Ho, R.K., Langeland, J., Prince, V., Wang, Y.L., Westerfield, M., Ekker, M. and Postlethwait, J.H.: Zebrafish hox clusters and vertebrate genome evolution. Science 282 (1998) 1711-1714.

Armes, N., Gilley, J. and Fried, M.: The comparative genomic structure and sequence of the surfeit gene homologs in the puffer fish Fugu rubripes and their association with CpG-rich islands. Genome Res 7 (1997) 1138-1152.

Bafna, V. and Pevzner, P.A.: Sorting by Reversals: Genome Rearrangements in Plant Organelles and Evolutionary History of X Chromosome. Mol Biol Evol 12 (1995) 239-246.

Bankaitis, V.A., Malehorn, D.E., Emr, S.D. and Greene, R.: The *Saccharomyces cerevisiae* SEC14 gene encodes a cytosolic factor that is required for transport of secretory proteins from the yeast Golgi complex. J Cell Biol 108 (1989) 1271-1281.

Baudat, F. and Nicolas, A.: Clustering of meiotic double-strand breaks on yeast chromosome III. Proc Natl Acad Sci U S A 94 (1997) 5213-5328.

Bennetzen, J.L. and Freeling, M.: The unified grass genome: synergy in synteny. Genome Res 7 (1997) 301-306.

Bennetzen, J.L. and Hall, B.D.: Codon selection in yeast. J Biol Chem 257 (1982) 3026-3031.

Berbee, M.L. and Taylor, J.W.: Ascomycete Relationships: Dating the Origin of Asexual Lineages with 18S Ribosomal RNA Gene Sequence Data. In: Reynolds, D.R. and Taylor, J.W. (Eds.), The Fungal Holomorph: Mitotic, Meiotic and Pleomorphic Speciation in Fungal Systematics. CAB International, Wallingford, 1993, pp. 67-78.

Bernardi, G.: The vertebrate genome: isochores and evolution. Mol Biol Evol 10 (1993) 186-204.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F.: The Mosaic Genome of Warm-Blooded Vertebrates. Science 228 (1985) 953-958.

Bevan, M. and Murphy, G.: The small, the large and the wild: the value of comparison in plant genomics. Trends Genet 15 (1999) 211-214.

Blanchette, M., Kunisawa, T. and Sankoff, D.: Parametric genome rearrangement. Gene 172 (1996) GC11-7.

Blanchette, M., Kunisawa, T. and Sankoff, D.: Gene order breakpoint evidence in animal mitochondrial phylogeny. J Mol Evol 49 (1999) 193-203.

Boore, J.L.: Animal mitochondrial genomes. Nucleic Acids Res 27 (1999) 1767-1780.

Boore, J.L., Collins, T.M., Stanton, D., Daehler, L.L. and Brown, W.M.: Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. Nature 376 (1995) 163-165.

Botta, A., Lindsay, E.A., Jurecic, V. and Baldini, A.: Comparative mapping of the DiGeorge syndrome region in mouse shows inconsistent gene order and differential degree of gene conservation [published erratum appears in Mamm Genome 1998 Apr;9(4):344]. Mamm Genome 8 (1997) 890-895.

Bowman, S., Churcher, C., Badcock, K., Brown, D., Chillingworth, T., Connor, R., Dedman, K., Devlin, K., Gentles, S., Hamlin, N., Hunt, S., Jagels, K., Lye, G., Moule, S., Odell, C., Pearson, D., Rajandream, M., Rice, P., Skelton, J., Walsh, S., Whitehead, S. and Barrell, B.: The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XIII. Nature 387 (1997) 90-93.

Bradnam, K.R., Seoighe, C., Sharp, P.M. and Wolfe, K.H.: G+C Content Variation Along and Among *Saccharomyces cerevisiae* Chromosomes. Mol Biol Evol 16 (1999) 666-675.

Brenner, S., Elgar, G., Sandford, R., Macrae, A., Venkatesh, B. and Aparicio, S.: Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. Nature 366 (1993) 265-268.

Brown, T.C. and Jiricny, J.: Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. Cell 54 (1988) 705-711.

Caccio, S., Perani, P., Saccone, S., Kadi, F. and Bernardi, G.: Single-copy sequence homology among the GC-richest isochores of the genomes from warm-blooded vertebrates. J Mol Evol 39 (1994) 331-339.

Cardazzo, B., Minuzzo, S., Sartori, G., Grapputo, A. and Carignani, G.: Evolution of mitochondrial DNA in yeast: gene order and structural organization of the mitochondrial genome of *Saccharomyces uvarum*. Curr Genet 33 (1998) 52-59.

Carver, E.A. and Stubbs, L.: Zooming in on the Human-Mouse Comparative Map: Genome Conservation Re-examined on a High-Resolution Scale. Genome Res 7 (1997) 1123-1137.

Chatfield, C: The Analysis of Time Series : An Introduction. Chapman and Hall, 1989.

Chen, T. and Zhang, M.Q.: Pombe: a gene-finding and exon-intron structure prediction system for fission yeast. Yeast 14 (1998) 701-710.

Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S. and Botstein, D.: SGD: *Saccharomyces* Genome Database. Nucleic Acids Res 26 (1998) 73-79.

Chervitz, S.A., Aravind, L., Sherlock, G., Ball, C.A., Koonin, E.V., Dwight, S.S., Harris, M.A., Dolinski, K., Mohr, S., Smith, T., Weng, S., Cherry, M.J. and Botstein, D.: Comparison of the Complete Protein Sets of Worm and Yeast. Orthology and Divergence. Science 282 (1998) 2022-2028.

Chibana, H., Magee, B.B., Grindle, S., Ran, Y., Scherer, S. and Magee, P.T.: A physical map of chromosome 7 of *Candida albicans*. Genetics 149 (1998) 1739-1752.

Churcher, C., Bowman, S., Badcock, K., Bankier, A., Brown, D., Chillingworth, T., Connor, R., Devlin, K., Gentles, S., Hamlin, N., Harris, D., Horsnell, T., Hunt, S., Jagels, K., Jones, M., Lye, G., Moule, S., Odell, C., Pearson, D., Rajandream, M., Rice, P., Rowley, N., Skelton, J., Smith, V., Barrell, B. *et al.*: The nucleotide sequence of *Saccharomyces cerevisiae* chromosome IX. Nature 387 (1997) 84-87.

Coissac, E., Maillier, E. and Netter, P.: A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. Mol Biol Evol 14 (1997) 1062-1074.

Consortium, C.e.S.: Genome Sequence of the Nematode *C. elegans*: A platform for Investigating Biology. Science 282 (1998) 1945-2140.

Cooke, J., Nowak, M.A. and Boerlijst, M.: Evolutionary origins and maintainance of redundant gene expression during metazoan development. Trends Genet 13 (1997) 360-364.

Copeland, N.G., Jenkins, N.A., Gilbert, D.J., Eppig, J.T., Maltais, L.J., Miller, J.C., Dietrich, W.F., Weaver, A., Lincoln, S.E., Steen, R.G., Stein, L.D., Nadeau, J.H. and Lander, E.S.: A genetic linkage map of the mouse: current applications and future prospects. Science 262 (1993) 57-66.

DeBry, R.W. and Seldin, M.F.: Human/mouse homology relationships. Genomics 33 (1996) 337-351.

Dietrich, F.S., Voegeli, S., Gaffney, T., Mohr, C., Rebishung, C., Wing, R., Choi, S., Goff, S. and Philippsen, P.: Gene map of chromosome I of *Ashbya Gosypii*. Curr Genet 35 (1999) 233.

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

Dujon, B., Alexandraki, D., Andre, B., Ansorge, W., Baladron, V., Ballesta, J.P., Banrevi, A., Bolle, P.A., Bolotin-Fukuhara, M., Bossier, P. and *et al.*: Complete DNA sequence of yeast chromosome XI. Nature 369 (1994) 371-378.

Duret, L., Mouchiroud, D. and Gautier, C.: Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J Mol Evol 40 (1995) 308-317.

Edwards, J.H.: Comparative genome mapping in mammals. Curr Opin Genet Dev 4 (1994) 861-867.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95 (1998) 14863-14868.

El-Mabrouk, N., Bryant, D. and Sankoff, D.: Reconstructing the Pre-Doubling Genome. Proceedings of the Third Annual International Conference on Computational Molecular Biology (1999) 154-163.

El-Mabrouk, N., Nadeau, J.H. and Sankoff, D.: Genome halving. In: Farach-Colton, M. (Ed.), Combinatorial Pattern Matching, 9th Annual Symposium. Springer-Verlag, Berlin, 1998, pp. 235-250.

Erlich, J., Sankoff, D. and Nadeau, J.: Synteny Conservation and Chromosome Rearrangements During Mammalian Evolution. Genetics 147 (1997) 289-296.

Eyre-Walker, A.: The role of DNA replication and isochores in generating mutation and silent substitution rate variance in mammals. Genet Res 60 (1992) 61-67.

Ferretti, V., Nadeau, J.H. and Sankoff, D.: Original synteny. In: Hirschberg, D. and Myers, G. (Eds.), Combinatorial Pattern Matching, 7th Annual Symposium. Springer-Verlag, Berlin, 1996, pp. 159-167.

Filipski, F., Thiery, J. and Bernardi, G.: An Analysis of the Bovine Genome by Cs2SO4-Ag+ Density Gradient Centrifugation. J Mol Biol 80 (1973) 177-197.

Filipski, J.: Correlation between molecular clock ticking, codon usage fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. FEBS Lett 217 (1987) 184-186.

Fitch, W.M.: Distinguishing homologous from analogous proteins. Syst Zool 19 (1970) 99-113.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y. and Postlethwait, J.: Preservation of Duplicate Genes by Complementary, Degerative Mutations. Genetics 151 (1999) 1531-1545.

Galibert, F., Alexandraki, D., Baur, A., Boles, E., Chalwatzis, N., Chuat, J.C., Coster, F., Cziepluch, C., De Haan, M., Domdey, H., Durand, P., Entian, K.D., Gatius, M., Goffeau, A., Grivell, L.A., Hennemann, A., Herbert, C.J., Heumann, K., Hilger, F., Hollenberg, C.P., Huang, M.E., Jacq, C., Jauniaux, J.C., Katsoulou, C., Karpfinger-Hartl, L. and *et al.*: Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome X. EMBO J. 15 (1996) 2031-2049.

Gallardo, M.H., Bickham, J.W., Honeycutt, R.L., Ojeda, R.A. and Kohler, N.: Discovery of tetraploidy in a mammal. Nature 40 (1999) 341-341.

Gardiner, K.: Base composition and gene distribution: critical patterns in mammalian genome organization. Trends Genet 12 (1996) 519-524.

Gates, M.A., Kim, L., Egan, E.S., Cardozo, T., Sirotkin, H.I., Dougan, S.T., Lashkari, D., Abagyan, R., Schier, A.F. and Talbot, W.S.: A genetic linkage map for zebrafish: comparative analysis and localization of genes and expressed sequences. Genome Res 9 (1999) 334-347.

Gaut, B.S. and Doebley, J.F.: DNA sequence evidence for the segmental allotetraploid origin of maize. Proc Natl Acad Sci U S A 94 (1997) 6809-6814.

Gibson, T.J. and Spring, J.: A model for massive genetic redundancy in vertebrates: polyploidy followed by persistence of genes encoding multidomain proteins. Trends Genet 16 (1998) 46-49.

Gilley, J. and Fried, M.: Extensive gene order differences within regions of conserved synteny between the fugu and human genomes: implications for chromosomal evolution and the cloning of disease genes. Hum Mol Genet 8 (1999) 1313-1320.

Gimeno, C.J., Ljungdahl, P.O., Styles, C.A. and Fink, G.R.: Unipolar cell divisions in the yeast *S. cerevisiae* lead to filamentous growth: regulation by starvation and RAS. Cell 68 (1992) 1077-1090.

Goffeau, A., Aert, R., Agostini-Carbone, M.L., Ahmed, A., Aigle, M. and *et al.*: The Yeast Genome Directory. Nature 387 (1997) (suppl.) 5-105.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S.G.: Life with 6000 genes. Science 274 (1996) 546, 563-7.

Groth, C., Hansen, J. and Piskur, J.: A natural chimeric yeast containing genetic material from three species. Int J Syst Bacteriol 49 (1999) 1933-1938.

Haber, J.E.: Mating-Type Gene Switching in *Saccharomyces cerevisiae*. Annu Rev Genet 32 (1998) 561-599.

Hani, J. and Feldmann, H.: tRNA genes and retroelements in the yeast genome. Nucleic Acids Res 26 (1998) 689-696.

Hannenhalli, S.: Polynomial-time algorithm for computing translocation distance between genomes. In: Galil, Z. and Ukkonen, E. (Eds.), Combinatorial Pattern Matching, 6th Annual Symposium. Springer-Verlag, Berlin, 1996, pp. 162-176.

Hannenhalli, S., Chappey, C., Koonin, E.V. and Pevzner, P.A.: Genome sequence comparison and scenarios for gene rearrangements: a test case. Genomics 30 (1995) 299-311.

Hansen, E.C.: *Saccharomyces pastorianus*. In: Kurtzman, C.P. and Fell, J.W. (Eds.), The Yeasts, A Taxonomic Study. Elsevier, Amsterdam, 1998, pp. 367-368.

Hartung, K., Frishman, D., Hinnen, A. and Wolfl, S.: Single-read sequence tags of a limited number of genomic DNA fragments provide an inexpensive tool for comparative genome analysis. Yeast 14 (1998) 1327-1332.

Henikoff, S. and Henikoff, J.G.: Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89 (1992) 10915-10919.

Herskowitz, I., Rine, J. and Strathern, J.: Mating-type Determination and Mating-type Interconversion in *Saccharomyces cerevisiae*. In: Jones, E.W., Pringle, J.R. and Broach, J.R. (Eds.), The Molecular And Cellular Biology Of The Yeast *Saccharomyces*. Cold Spring Harbour Laboratory Press, New York, 1992, pp. 583-656.

Hinchliffe, E. and Kenny, E.: Yeast as a Vehicle for the Expression of Heterologous Genes. In: Rose, A.H. and Harrison, J.S. (Eds.), The Yeasts. Academic Press, London, 1993, pp. 325-356.

Hodges, P.E., McKee, A.H.Z., Davis, B.P., Payne, W.E. and Garrels, J.I.: Yeast Protein Database (YPD): a model for the organization and presentation of genome-wide functional data. Nucleic Acids Res 27 (1999) 69-73.

Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S. and Young, R.A.: Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95 (1998) 717-728.

Hughes, J.M., Konings, D.A. and Cesareni, G.: The yeast homologue of U3 snRNA. EMBO J. 6 (1987) 2145-55.

Hull, C.M. and Johnson, A.D.: Identification of a mating type-like locus in the asexual pathogenic yeast candida albicans. Science 285 (1999) 1271-1275.

Ikemura, T. and Wada, K.: Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers; relation between nucleotide sequence data and cytogenetic data. Nucleic Acids Res 19 (1991) 4333-4339.

Ingledew: Yeasts for Production of Fuel Ethanol. In: Rose, A.H. and Harrison, J.S. (Eds.), The Yeasts. Academic Press, Amsterdam, 1993, pp. 245-291.

James, S.A., Cai, J., Roberts, I.N. and Collins, M.D.: A phylogenetic analysis of the genus *Saccharomyces* based on 18S rRNA gene sequences: description of *Saccharomyces kunashirensis* sp. nov. and *Saccharomyces martiniae* sp. Int J Syst Bacteriol 47 (1997) 453-460.

Janbon, G., Sherman, F. and Rustchenko, E.: Monosomy of a specific chromosome determines L-sorbose utilization: a novel regulatory mechanism in *Candida albicans*. Proc Natl Acad Sci U S A 95 (1998) 5150-5155.

Johnston, M., Hillier, L., Riles, L., Albermann, K., Andre, B., Ansorge, W., Benes, V., Bruckner, M., Delius, H., Dubois, E., Dusterhoft, A., Entian, K.D., Floeth, M., Goffeau, A., Hebling, U., Heumann, K., Heuss-Neitzel, D., Hilbert, H., Hilger, F., Kleine, K., Kotter, P., Louis, E.J., Messenguy, F., Mewes, H.W., Hoheisel, J.D. and *et al.*: The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XII. Nature 387 (1997) 87-90.

Kaback, D.B.: Yeast Genome Structure. In: Rose, A.H., Wheals, A.E. and Harrison, J.S. (Eds.), The Yeasts. Academic Press, London, 1995, pp. 179-222.

Kadi, F., Mouchiroud, D., Sabeur, G. and Bernardi, G.: The Compositional Patterns of the Avian Genomes and Their Evolutionary Implications. J Mol Evol 37 (1993) 544-551.

Kececioglu, J. and Sankoff, D.: Exact and approximate algorithms for the inversion distance between two chromosomes. In: Apostolico, A., Crochemore, M., Galil, Z. and Manber, U. (Eds.), Combinatorial Pattern Matching, 4th Annual Symposium. Springer-Verlag, Berlin, 1993.

Kehrer-Sawatzki, H., Maier, C., Moschgath, E., Elgar, G. and Krone, W.: Characterization of three genes, AKAP84, BAW and WSB1, located 3' to the neurofibromatosis type 1 locus in fugu rubripes. Gene 235 (1999) 1-11.

Keogh, R.S., Seoighe, C. and Wolfe, K.H.: Evolution of gene order and chromosome number in Saccharomyces, Kluyveromyces and related fungi. Yeast 14 (1998) 443-457.

Kielland-Brandt, M., Nielsson-Tillgren, T., Gjermansen, C., Holmberg, S. and Pedersen, M.B.: Genetics of Brewing Yeasts. In: Rose, A.H., Wheals, E.

and Harrison, J.S. (Eds.), The Yeasts. Academic Press, London, 1995, pp. 223-254.

Kimura, M.: The neutral theory of molecular evolution. Cambridge University Press, Cambridge, 1983.

Kurata, N., Moore, G., Nagamura, Y. and Foote, T.: Conservation of Genome Structure Between Rice and Wheat. Biotechnology 12 (1994) 276-278.

Kurtzman, C.P.: Discussion of teleomorphic and anamorphic ascomycetous yeasts and a key to genera. In: Kurtzman, C.P. and Fell, J.W. (Eds.), The Yeasts, A Taxonomic Study. Elsevier, Amsterdam, 1998a, pp. 111-121.

Kurtzman, C.P.: *Lodderomyces* van der Walt. In: Kurtzman, C.P. and Fell, J.W. (Eds.), The Yeasts, A Taxonomic Study. Elsevier, Amsterdam, 1998b, pp. 254-255.

Kurtzman, C.P. and Robnett, C.J.: Identification of clinically important ascomycetous yeasts based on nucleotide divergence in the 5' end of the large-subunit (26S) ribosomal DNA gene. J Clin Microbiol 35 (1997) 1216-1223.

Kurtzman, C.P. and Robnett, C.J.: Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. Antonie Van Leeuwenhoek 73 (1998) 331-371.

Lalo, D., Stettler, S., Mariotte, S., Slonimski, P.P. and Thuriaux, P.: Une duplication fossile entre les régions centromériques de deux chromosomes chez la levure. C R Acad Sci Paris 316 (1993) 367-373.

Li, W.-H.: Molecular Evolution. Sinauer Associates, Inc., Sunderland, 1997.

Lindsay, E.A., Botta, A., Jurecic, V., Carattini-Rivera, S., Cheah, Y.C., Rosenblatt, H.M., Bradley, A. and Baldini, A.: Congenital heart disease in mice deficient for the DiGeorge syndrome region. Nature 401 (1999) 379-83.

Lloyd, A.T. and Sharp, P.M.: Evolution of codon usage patterns: the extent and nature of divergence between *Candida albicans* and *Saccharomyces cerevisiae*. Nucleic Acids Res 20 (1992) 5289-5295.

Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E.L.: Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14 (1996) 1675-1680.

Lott, T.J., Holloway, B.P., Logan, D.A., Fundyga, R. and Arnold, J.: Towards understanding the evolution of the human commensal yeast Candida albicans. Microbiology 145 (1999) 1137-1143.

Lubkowitz, M.A., Hauser, L., Breslav, M., Naider, F. and Becker, J.M.: An oligopeptide transport gene from Candida albicans. Microbiology 143 (1997) 387-396.

Lundin, L.G.: Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. Genomics 16 (1993) 1-19.

Lyons, T.P., Jacques, K.A. and Dawson, K.A.: Miscellaneous Products from Yeast. In: Rose, A.H. and Harrison, J.S. (Eds.), The Yeasts. Academic Press, London, 1993, pp. 293-324.

Magee, P.T.: Analysis of the *Candida albicans* Genome. Meth Microbiol 26 (1998) 395-415.

Masneuf, I., Hansen, J., Groth, C., Piskur, J. and Dubourdieu, D.: New Hybrids between *Saccharomyces* Sensu Stricto Yeast Species Found among Wine and Cider Production Strains. Appl Environ Microbiol 64 (1998) 3887-3892.

McLysaght, A., Enright, A., Skrabanek, L. and Wolfe, K.H.: Estimation of Synteny Conservation and Genome Compaction between Pufferfish (Fugu) and Human. Comparative and Functional Genomics (*In Press*) .

Mewes, H.W., Albermann, K., Bähr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S.G., Pfeiffer, F. and Zollner, A.: Overview of the yeast genome. Nature 387 (1997) (suppl.) 7-65.

Meyer, S.A., Payne, R.W. and Yarrow, D.: *Candida* Berkhout. In: Kurtzman, C.P. and Fell, J.W. (Eds.), The Yeasts, A Taxonomic Study. Elsevier, Amsterdam, 1998, pp. 454-573.

Miklos, G.L. and Rubin, G.M.: The role of the genome project in determining gene function: insights from model organisms. Cell 86 (1996) 521-529.

Moore, G.: Cereal genome evolution: pastoral pursuits with 'Lego' genomes. Current Biology 5 (1995) 717-724.

Moore, G., Devos, K.M., Wang, Z. and Gale, M.D.: Grasses, line up and form a circle. Current Biology 5 (1995a) 737-739.

Moore, G., Foote, T., Helentjaris, T., Devos, K., Kurata, N. and Gale, M.: Was there a single ancestral cereal chromosome? [letter]. Trends Genet 11 (1995b) 81-82.

Morizot, D.C.: Use of fish gene maps to predict ancestral vertebrate genome organization. In: Ogita, Z.-I. and Markert, C.L. (Eds.), Isozymes: Structure, Function, and Use in Biology and Medicine. Wiley-Liss, New York, 1990, pp. 207-234.

Mortimer, R.K., Contopoulou, C.R. and King, J.S.: Genetic and physical maps of *Saccharomces cerevisiae,* edition 11. Yeast 8 (1992) 817-902.

Mortimer, R.K. and Johnston, J.R.: Genealogy of principal strains of the yeast genetic stock center. Genetics 113 (1986) 35-43.

Nadeau, J.H.: Maps of linkage and synteny homologies between mouse and man. Trends Genet 5 (1989) 82-86.

Nadeau, J.H.: Genome duplication and comparative gene mapping. In: Adolph, K.W. (Ed.), Advanced Techniques in Chromosome Research. Marcel Dekker, New York, 1991, pp. 269-296.

Nadeau, J.H. and Sankoff, D.: The lengths of undiscovered conserved segments in comparative maps. Mamm Genome 9 (1998) 491-495.

Nadeau, J.H. and Taylor, B.A.: Lengths of chromosomal segments conserved since divergence of man and mouse. Proc Natl Acad Sci U S A 81 (1984) 814-818.

Nowak, M.A., Boerlijst, M.C., Cooke, J. and Maynard Smith, J.: Evolution of genetic redundancy. Nature 388 (1997) 167-171.

O'Brien, S.J., Menotti-Raymond, M., Murphy, W.J., Nash, W.G., Wienberg, J., Stanyon, R., Copeland, N.G., Jenkins, N.A., Womack, J.E. and Marshall Graves, J.A.: The Promise of Comparative Genomics in Mammals. Science 286 (1999) 458-481.

Ohno, S.: Evolution by Gene Duplication. George Allen and Unwin, London, 1970.

Ozier-Kalogeropoulos, O., Malpertuy, A., Boyer, J., Tehaia , F. and Dujon, B.: Random exploration of the *Kluyveromyces lactis* genome and comparison with that of *Saccharomyces cerevisiae*. Nucleic Acids Res 26 (1998) 5511-5524.

Palmer, J.D. and Herbon, L.A.: Unicircular structure of the *Brassica hirta* mitochondrial genome. Curr Genet 11 (1987) 565-570.

Palmer, J.D. and Herbon, L.A.: Plant Mitochondrial DNA Evolves Rapidly in Structure, but Slowly in Sequence. J Mol Evol 28 (1988) 87-97.

Palmer, J.D., Osorio, B. and Thompson, W.F.: Evolutionary significance of inversions in legume chloroplast DNAs. Curr Genet 14 (1988) 65-74.

Paterson, A.H., Lan, T.H., Reischmann, K.P., Chang, C., Lin, Y.R., Liu, S.C., Burow, M.D., Kowalski, S.P., Katsar, C.S., DelMonte, T.A., Feldmann, K.A., Schertz, K.F. and Wendel, J.F.: Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence. Nat Genet 14 (1996) 380-382.

Pearson, W.R. and Lipman, D.J.: Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 85 (1988) 2444-2448.

Pebusque, M.J., Coulier, F., Birnbaum, D. and Pontarotti, P.: Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. Mol Biol Evol 15 (1998) 1145-1159.

Perepnikhatka, V., Fischer, F.J., Niimi, M., Baker, R.A., Cannon, R.D., Wang, Y.K., Sherman, F. and Rustchenko, E.: Specific chromosome alterations in fluconazole-resistant mutants of Candida albicans. J Bacteriol 181 (1999) 4041-4049.

Pesole, G., Lotti, M., Alberghina, L. and Saccone, C.: Evolutionary origin of nonuniversal CUGSer codon in some Candida species as inferred from a molecular phylogeny. Genetics 141 (1995) 903-907.

Petersen, R.F., Marinoni, G., Nielsen, M.L. and Piskur, J.: Molecular approaches for analyzing diversity and phylogeny among yeast species. (in press-a) .

Petersen, R.F., Nilsson-Tillgren, T. and Piskur, J.: Karyotypes of *Saccharomyces* sensu lato species. (in press-b) .

Piskur, J.: Inheritance of the Yeast Mitochondrial Genome. Plasmid 31 (1993) 229-241.

Piskur, J., Smole, S., Groth, C., Petersen, R.F. and Pedersen, M.B.: Structure and genetic stability of mitochondrial genomes vary among yeasts of the genus *Sacharomyces*. Int J Syst Bacteriol 48 (1998) 1015-1024.

Plant, E.P., Becher, D. and Poulter, R.T.: The SPL1 tRNA splicing gene of Candida maltosa and Candida albicans. Yeast 14 (1998) 287-295.

Planta, R.J. and Mager, W.H.: The list of cytoplasmic ribosomal proteins of *Saccharomyces cerevisiae*. Yeast 14 (1998) 471-477.

Poulter, R.T.M.: Genetics of *Candida* Species. In: Rose, A.H., Wheals, A.E. and Harrison, J.S. (Eds.), The Yeasts. Academic Press, London, 1995, pp. 285-308.

Raymond, M., Dignard, D., Alarco, A.M., Mainville, N., Magee, B.B. and Thomas, D.Y.: A Ste6p/P-glycoprotein homologue from the asexual yeast Candida albicans transports the a-factor mating pheromone in *Saccharomyces cerevisiae*. Mol Microbiol 27 (1998) 587-598.

Roig, P., Martinez, J.P. and Go\albo, D.: (1997) GenBank/EMBL/DDBJ database accession number Y15608.

Ruddle, F.H., Bentley, K.L., Murtha, M.T. and Risch, N.: Gene loss and gain in the evolution of the vertebrates. Dev Suppl (1994) 155-161.

Ruvkun, G. and Hobert, O.: The taxonomy of developmental control in Caenorhabditis elegans. Science 282 (1998) 2033-2041.

Ryu, S., Murooka, Y. and Kaneko, Y.: Reciprocal translocation at duplicated *RPL2* loci might cause speciation of *Saccharomyces bayanus* and *Saccharomyces cerevisiae*. Curr Genet 33 (1998) 345-351.

Sabeur, G., Macaya, G., Kadi, F. and Bernardi, G.: The isochore patterns of mammalian genomes and their phylogenetic implications [published erratum appears in J Mol Evol 1994 May;38(5):547]. J Mol Evol 37 (1993) 93-108.

Saccone, S., De Sario, A., Wiegant, J., Raap, A.K., Della Valle, G. and Bernardi, G.: Correlations between isochores and chromosomal bands in the human genome. Proc Natl Acad Sci U S A 90 (1993) 11929-11933.

Saltarelli, C.G.: *Candida albicans*: The Pathogenic Fungus. Hemisphere publishing company, New York, 1989.

Sankoff, D.: Analytical approaches to genomic evolution. Biochimie 75 (1993) 409-413.

Sankoff, D. and Blanchette, M.: Multiple genome rearrangement and breakpoint phylogeny. J Comput Biol 5 (1998) 555-570.

Sankoff, D., Cedergren, R. and Abel, Y.: Genomic Divergence through Gene Rearrangement. Methods Enzymol 183 (1990) 428-438.

Sankoff, D. and Ferretti, V.: Karyotype distributions in a stochastic model of reciprocal translocation. Genome Res 6 (1996) 1-9.

Sankoff, D., Ferretti, V. and Nadeau, J.H.: Conserved segment identification, RECOMB 97. Proceedings of the First Annual International Conference on Computational Molecular Biology. ACM Press, New York, 1997a, pp. 252-256.

Sankoff, D. and Goldstein, M.: Probabilistic models of genome shuffling. Bull Math Biol 51 (1989) 117-124.

Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B.F. and Cedergren, R.: Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. Proc Natl Acad Sci U S A 89 (1992) 6575-6579.

Sankoff, D. and Nadeau, J.H.: Conserved synteny as a measure of genomic distance. Discrete Applied Mathematics 71 (1996) 247-257.

Sankoff, D., Parent, M.-N., Marchand, I. and Ferretti, V.: On the Nadeau-Taylor theory of conserved chromosome segments. In: Apostolico, A. and Hein, J. (Eds.), Combinatorial Pattern Matching, 8th Annual Symposium. Springer-Verlag, Berlin, 1997b, pp. 262-274.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z. and Bennetzen, J.L.: Nested retrotransposons in the intergenic regions of the maize genome. Science 274 (1996) 765-768.

Santos, M.A. and Tuite, M.F.: The CUG codon is decoded in vivo as serine and not leucine in Candida albicans. Nucleic Acids Res 23 (1995) 1481-1486.

Schubert, I. and Oud, J.L.: There is an Upper Limit of Chromosome Size for Normal Development of an Organism. Cell 88 (1997) 515-520.

Semple, C. and Wolfe, K.H.: Gene duplication and gene conversion in the Caenorhabditis elegans genome. J Mol Evol 48 (1999) 555-564.

Seoighe, C., Federspiel, N., Jones, T., Hansen, N., Bivolarovic, V., Surzycki, R., Tamse, R., Komp, C., Huizar, L., Davis, R., Scherer, S., Tait, E., Shaw, D.J., Harris, D., Murphy, L., Oliver, K., Taylor, K., Rajandream, M., Barrell, B.G. and Wolfe, K.H.: Prevalence of small inversions in yeast gene order evolution. Science (submitted) .

Seoighe, C. and Wolfe, K.H.: Extent of genomic rearrangement after genome duplication in yeast. Proc Natl Acad Sci U S A 95 (1998) 4447-4452.

Seoighe, C. and Wolfe, K.H.: Updated map of duplicated regions in the yeast genome. Gene 238 (1999a) 253-261.

Seoighe, C. and Wolfe, K.H.: Yeast genome evolution in the post-genome era. Curr Opin Microbiol. 2 (1999b) 548-554.

Sharp, P.M. and Cowe, E.: Synonymous codon usage in *Saccharomyces cerevisiae*. Yeast 7 (1991) 657-678.

Sharp, P.M. and Lloyd, A.T.: Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure. Nucleic Acids Res 21 (1993) 179-183.

Shoemaker, R.C., Polzin, K., Labate, J., Specht, J., Brummer, E.C., Olson, T., Young, N., Concibido, V., Wilcox, J., Tamulonis, J.P., Kochert, G. and Boerma, H.R.: Genome duplication in soybean (Glycine subgenus soja). Genetics 144 (1996) 329-338.

Shpaer, E.G., Robinson, M., Yee, D., Candlin, J.D., Mines, R. and Hunkapiller, T.: Sensitivity and Selectivity in Protein Similarity Searches: A Comparison of Smith-Waterman in Hardware to BLAST and FASTA. Genomics 38 (1996) 179-191.

Sidow, A.: Gen(om)e duplications in the evolution of early vertebrates. Curr Opin Genet Dev 6 (1996) 715-722.

Skrabanek, L. and Wolfe, K.H.: Eukaryote genome duplication - where's the evidence? Curr Opin Genet Dev 8 (1999) 694-700.

Smith, T.F. and Waterman, M.S.: Identification of common molecular subsequences. J Mol Biol 147 (1981) 195-197.

Song, K., Lu, P., Tang, K. and Osborn, T.C.: Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. Proc Natl Acad Sci U S A 92 (1995) 7719-7723.

Sonstegard, T.S., Kappes, S.M., Keele, J.W. and Smith, T.P.: Refinement of bovine chromosome 2 linkage map near the mh locus reveals rearrangements between the bovine and human genomes. Anim Genet 29 (1998) 341-347.

Sprague, G.F.: Mating and Mating-type Interconversion in *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. In: Rose, A.H., Wheals, A.E. and Harrison, J.S. (Eds.), The Yeasts. Academic Press, London, 1995, pp. 411-459.

Spring, J.: Vertebrate evolution by interspecific hybridisation--are we polyploid? FEBS Lett 400 (1997) 2-8.

Suvarna, K., Seah, L., Bhattacherjee, V. and Bhattacharjee, J.K.: Molecular analysis of the LYS2 gene of Candida albicans: homology to peptide antibiotic synthetases and the regulation of the alpha- aminoadipate reductase. Curr Genet 33 (1998) 268-275.

Tait, E., Simon, M.C., King, S., Brown, A.J., Gow, N.A.R. and Shaw, D.J.: A *Candida albicans* Genome Project: Cosmid Contigs, Physical Mapping, and Gene Isolation. Fungal Genetics and Biology 21 (1997) 308-314.

Tamai, Y., Momma, T., Yoshimoto, H. and Kaneko, Y.: Co-existence of Two Types of Chromosome in the Bottom Fermenting Yeast, *Saccharomyces pastorianus*. Yeast 14 (1998) 923-933.

Tettelin, H., Agostoni Carbone, M.L., Albermann, K., Albers, M., Arroyo, J., Backes, U., Barreiros, T., Bertani, I., Bjourson, A.J., Bruckner, M., Bruschi, C.V., Carignani, G., Castagnoli, L., Cerdan, E., Clemente, M.L., Coblenz, A., Coglievina, M., Coissac, E., Defoor, E., Del Bino, S., Delius, H., Delneri, D., de Wergifosse, P., Dujon, B., Kleine, K. and *et al.*: The nucleotide sequence of *Saccharomyces cerevisiae* chromosome VII. Nature 387 (1997) 81-84.

Thatcher, J.W., Shaw, J.M. and Dickinson, W.J.: Marginal fitness contributions of nonessential genes in yeast. Proc Natl Acad Sci U S A 95 (1998) 253-257.

Trachtulec, Z., Hamvas, R.M., Forejt, J., Lehrach, H.R., Vincek, V. and Klein, J.: Linkage of TATA-binding protein and proteasome subunit C5 genes in mice and humans reveals synteny conserved between mammals and invertebrates. Genomics 44 (1997) 1-7.

Van Deynze, A.E., Nelsno, J.C. and Yglesias, E.S.: Comparative mapping in grasses, Wheat relationships. Mol Gen Genet 248 (1995a) 744-754.

Van Deynze, A.E., Nelson, J.C., O'Donoughue, L.S., Ahn, S.N., Siripoonwiwat, W., Harrington, S.E., Yglesias, E.S., Braga, D.P., McCouch, S.R. and Sorrells, M.E.: Comparative mapping in grasses. Oat relationships. Mol Gen Genet 249 (1995b) 349-356.

Vaughan-Martini, A. and Martini, A.: *Saccharomyces Meyen ex Reess*. In: Kurtzman, C.P. and Fell, J.W. (Eds.), The Yeasts, A Taxonomic Study. Elsevier, Amsterdam, 1998.

Walsh, J.B.: How Often Do Duplicated Genes Evolve New Functions? Genetics 139 (1995) 421-428.

Wei, X., Samarabandu, J., Devdhar, R.S., Siegel, A.J., Acharya, R. and Berezney, R.: Segregation of transcription and replication sites into higher order domains. Science 281 (1998) 1502-1506.

Weiler, K.S., Szeto, L. and Broach, J.R.: Mutations affecting donor preference during mating type interconversion in *Saccharomyces cerevisiae*. Genetics 139 (1995) 1495-1510.

Wilson, W.A., Harrington, S.E., Woodman, W.L., Lee, M., Sorrells, M.E. and McCouch, S.R.: Inferences on the genome structure of progenitor maize through comparative analysis of rice, maize and the domesticated panicoids. Genetics 153 (1999) 453-473.

Wolfe, K.H.: Molecular evolution of plants: more genomes, fewer generalities. In: Barber, J. and Andersson, B. (Eds.), Molecular Genetics of Photosynthesis. IRL Press, Oxford, 1996, pp. 45-57.

Wolfe, K.H., Sharp, P.M. and Li, W.H.: Mutation rates differ among regions of the mammalian genome. Nature 337 (1989) 283-285.

Wolfe, K.H. and Shields, D.C.: Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387 (1997) 708-713.

Wootton, J.C. and Federhen, S.: Analysis of compositionally biased regions in sequence databases. Methods Enzymol 266 (1996) 554-571.

Wu, X. and Haber, J.E.: MATa donor preference in yeast mating-type switching: activation of a large chromosomal region for recombination. Genes Dev 9 (1995) 1922-1932.

Yang, Y.P. and Womack, J.E.: Parallel radiation hybrid mapping: a powerful tool for high-resolution genomic comparison. Genome Res 8 (1998) 731-736.

Yuan, Y.P., Eulenstein, O., Vingron, M. and Bork, P.: Towards detection of orthologues in sequence databases. Bioinformatics 14 (1998) 285-289.