Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*

Manolis Kellis[1,2], **Bruce W. Birren**[1] **& Eric S. Lander**[1,3]

[1]*The Broad Institute, Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02138, USA*
[2]*MIT Computer Science and Artificial Intelligence Laboratory, and* [3]*Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02139, USA*

Whole-genome duplication followed by massive gene loss and specialization has long been postulated as a powerful mechanism of evolutionary innovation. Recently, it has become possible to test this notion by searching complete genome sequence for signs of ancient duplication. Here, we show that the yeast *Saccharomyces cerevisiae* arose from ancient whole-genome duplication, by sequencing and analysing *Kluyveromyces waltii*, a related yeast species that diverged before the duplication. The two genomes are related by a 1:2 mapping, with each region of *K. waltii* corresponding to two regions of *S. cerevisiae*, as expected for whole-genome duplication. This resolves the long-standing controversy on the ancestry of the yeast genome, and makes it possible to study the fate of duplicated genes directly. Strikingly, 95% of cases of accelerated evolution involve only one member of a gene pair, providing strong support for a specific model of evolution, and allowing us to distinguish ancestral and derived functions.

Genomic duplication has been proposed as an advantageous path to evolutionary innovation[1], because duplicated genes can supply genetic raw material for the emergence of new functions through the forces of mutation and natural selection. Such duplication can involve individual genes, genomic segments or whole genomes. Whole-genome duplication (WGD) is a particularly intriguing but poorly understood situation. In principle, coordinate duplication of an entire genome may allow for large-scale adaptation to new environments. However, polyploidy comes at the cost of major genomic instability[2], which persists until the genome returns to functionally normal ploidy through mutation, gene loss and genomic rearrangements. Direct study of such a cataclysmic genomic event may provide major insights into the dynamics of genome evolution and the emergence of new functions.

In the yeast *S. cerevisiae*, subtle analysis of the genomic locations of paralogues revealed the existence of ancestral duplication blocks, but their origin has remained controversial[3–17]. Wolfe and colleagues interpreted the presence and distribution of such regions in the *S. cerevisiae* genome as supporting a model of WGD[3,4]. Others have challenged this conclusion, noting that the WGD model is based on a small set of yeast genes (~8%), explains only a minority of the gene redundancy in the genome and defines putative sister regions covering less than 50% of the genome. An alternative model that has been proposed is that the duplicated segments arose via independent local duplication events[8–15]. Recent reports assert that low-coverage sequence data from related yeast species supports this alternative model[14,15]; Wolfe and colleagues have in turn challenged this analysis[16]. In addition, Wolfe and colleagues postulated that WGD occurred after the divergence of *S. cerevisiae* and *K. lactis*, while Langkjaer *et al.* have recently presented evidence that any putative WGD would necessarily predate this divergence[17].

Here we provide direct evidence of WGD in yeast, by sequencing and analysing a related species whose divergence precedes the duplication event. We show that *S. cerevisiae* arose from complete duplication of eight ancestral chromosomes, and subsequently returned to functionally normal ploidy by massive loss of nearly 90% of duplicated genes in small deletions. These were balanced and complementary in paired regions, preserving at least one copy of virtually each gene in the ancestral gene set. We identify 145 paired regions in *S. cerevisiae*, tiling 88% of the genome and containing 457 duplicated gene pairs.

We then analyse the post-duplication divergence of gene pairs, and show evidence of accelerated evolution in many cases. Strikingly, 95% of cases of accelerated evolution involve only one member of a gene pair, providing strong support for a specific model of evolution[1], and allowing ancestral and derived functions to be distinguished. We find that derived genes tend to be specialized in function, expression and localization, and lose essential aspects of their ancestral function. In addition, we find striking examples of neofunctionalization, including the emergence of silencing from origin-of-replication binding, and the emergence of viral defence mechanisms from translation elongation.

Evidence of whole-genome duplication

The expected signature of genome duplication is illustrated in Fig. 1. Following duplication, sister regions would undergo progressive gene loss by random local deletion: one or the other of the two paralogous copies of each gene would be lost in most cases, with both paralogues being retained if they happen to acquire distinct functions. Eventually, the only residual signature to show that two regions arose from ancestral duplication is the presence of a few paralogous genes in the same order and orientation scattered in a sea of otherwise unrelated genes. Paired regions containing such signatures will be relatively short, because chromosomal rearrangements would disrupt global gene order over time, leaving only weak evidence of ancestral duplication.

The clearest way to prove the existence of an ancient WGD would be to find a yeast species, denoted as Y, that descends directly from a common ancestor along a lineage that diverged before the duplication (Fig. 1). In principle, species Y and *S. cerevisiae* would be related by a 1:2 mapping satisfying several properties: (1) nearly every region in species Y would correspond to two sister regions in *S. cerevisiae*; (2) the two sister regions in *S. cerevisiae* would each contain an ordered subsequence of the genes in the corresponding region of species Y, with each containing roughly half of the genes and the two subsequences interleaving to account for nearly all of the genes; and (3) nearly every region of *S. cerevisiae* would correspond to one region of species Y, and thus be paired to a sister

region in *S. cerevisiae*. Below, we show that *K. waltii* matches these criteria.

Genome sequencing and alignment

We sequenced the genome of *K. waltii*, as part of a larger fungal genome programme. We note that *Kluyveromyces* is a polyphyletic genus, and in particular, *K. waltii* is closer to *S. cerevisiae* than the previously studied *K. lactis*. In light of this, *K. waltii* has been proposed to be renamed *Lachancea waltii*[18], although we use the more familiar name here. We used whole-genome shotgun sequencing to obtain eightfold sequence coverage and 42-fold physical coverage, and assembled the information into essentially eight complete chromosomes of total length 10.7 Mb (see Methods).

We annotated the draft genome sequence to identify 5,230 likely protein-coding genes (see Methods), notably fewer than the 5,714 genes in *S. cerevisiae*[19]. The genome also contains 240 transfer RNA genes (36 fewer than *Saccharomyces* species) and 60 Ty elements (primarily TyB). It also contains several telomeric gene families that are probably involved in adaptation to changing environments, including families of membrane proteins (51 members), hexose transporters (20 members), and flocculins (50 members, a notable expansion relative to *S. cerevisiae*). Finally, 7% of genes show no detectable protein similarity to *S. cerevisiae*. The genome sequence has been deposited in public databases (GenBank/DDBJ/EMBL), and the sequence and annotation are available from our website (http://www.broad.mit.edu/seq/YeastDuplication) and as Supplementary Information.

We compared the genomes of *K. waltii* and *S. cerevisiae* to identify orthologous regions. We first identified closely matching genes between the species and then identified blocks of conserved synteny—that is, regions containing numerous matching genes in the same order across the two species (see Methods). In contrast to the 1:1 mapping seen for close relatives[19], most local regions in *K. waltii* mapped to two regions in *S. cerevisiae*, with each containing matches to only a subset of the *K. waltii* genes.

To quantify this observation, we developed a computational algorithm to parse *K. waltii* into blocks of doubly conserved synteny (DCS)—defined as maximal regions that map across their entire length to two distinct regions in *S. cerevisiae* (see Methods). Briefly, we labelled the genes within each synteny block, and used the genes that belonged to multiple synteny blocks to detect regions of duplicate synteny. We then used the duplicate mapping in DCS blocks to define sister regions in the *S. cerevisiae* genome (see Methods).

Complete duplication

We identified a total of 253 DCS blocks, containing 75% of *K. waltii* genes and 81% of *S. cerevisiae* genes. The cross-species mapping of DCS blocks is shown in Fig. 2a, with a detailed view of a region of chromosome 1. DCS blocks tile 85% of each *K. waltii* chromosome in the pattern expected for WGD. Detailed views of all regions are shown in Supplementary Figs S5.

The typical gene falls in a DCS block containing 27 genes, and the

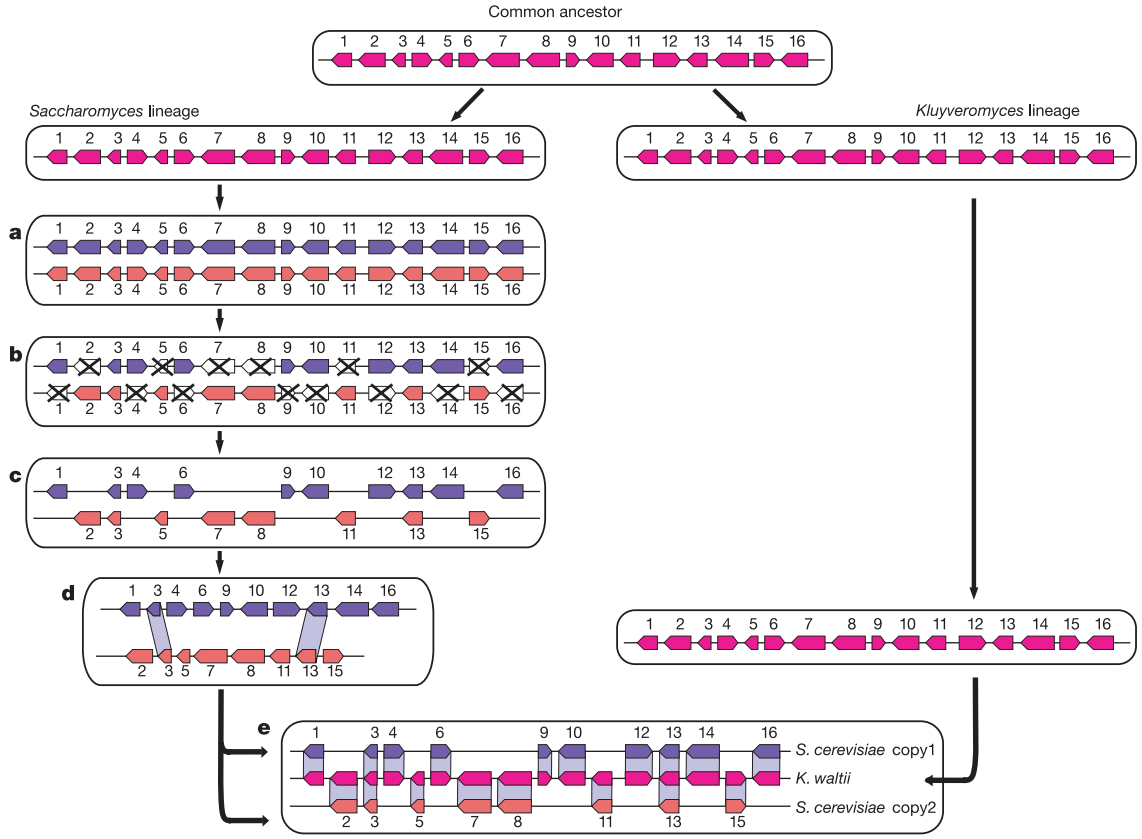

Figure 1 Model of WGD followed by massive gene loss predicts gene interleaving in sister regions. **a**, After divergence from *K. waltii*, the *Saccharomyces* lineage underwent a genome duplication event, creating two copies of every gene and chromosome. **b**, The vast majority of duplicated genes underwent mutation and gene loss. **c**, Sister segments retained different subsets of the original gene set, keeping two copies for only a small minority of duplicated genes, which were retained for functional purposes. **d**, Within *S. cerevisiae*, the only evidence comes from the conserved order of duplicated genes (numbered 3 and 13) across different chromosomal segments; the intervening genes are unrelated. **e**, Comparison with *K. waltii* reveals the duplicated nature of the *S. cerevisiae* genome, interleaving genes from sister segments on the basis of the ancestral gene order.

largest such block contains 81 genes. DCS blocks are separated by small segments (~3 genes on average) that appear to match only one conserved region in *S. cerevisiae*; it is likely that the second paralogous segment cannot be confidently recognized because it has lost too many genes or undergone rearrangement (see Methods). Finally, ~1% of the *K. waltii* genome lies in segments that match three or more regions in *S. cerevisiae*; the additional matching regions result mostly from slight local rearrangements (see Methods).

Complete interleaving of genes

Within DCS blocks, the interleaving of genes shows essentially

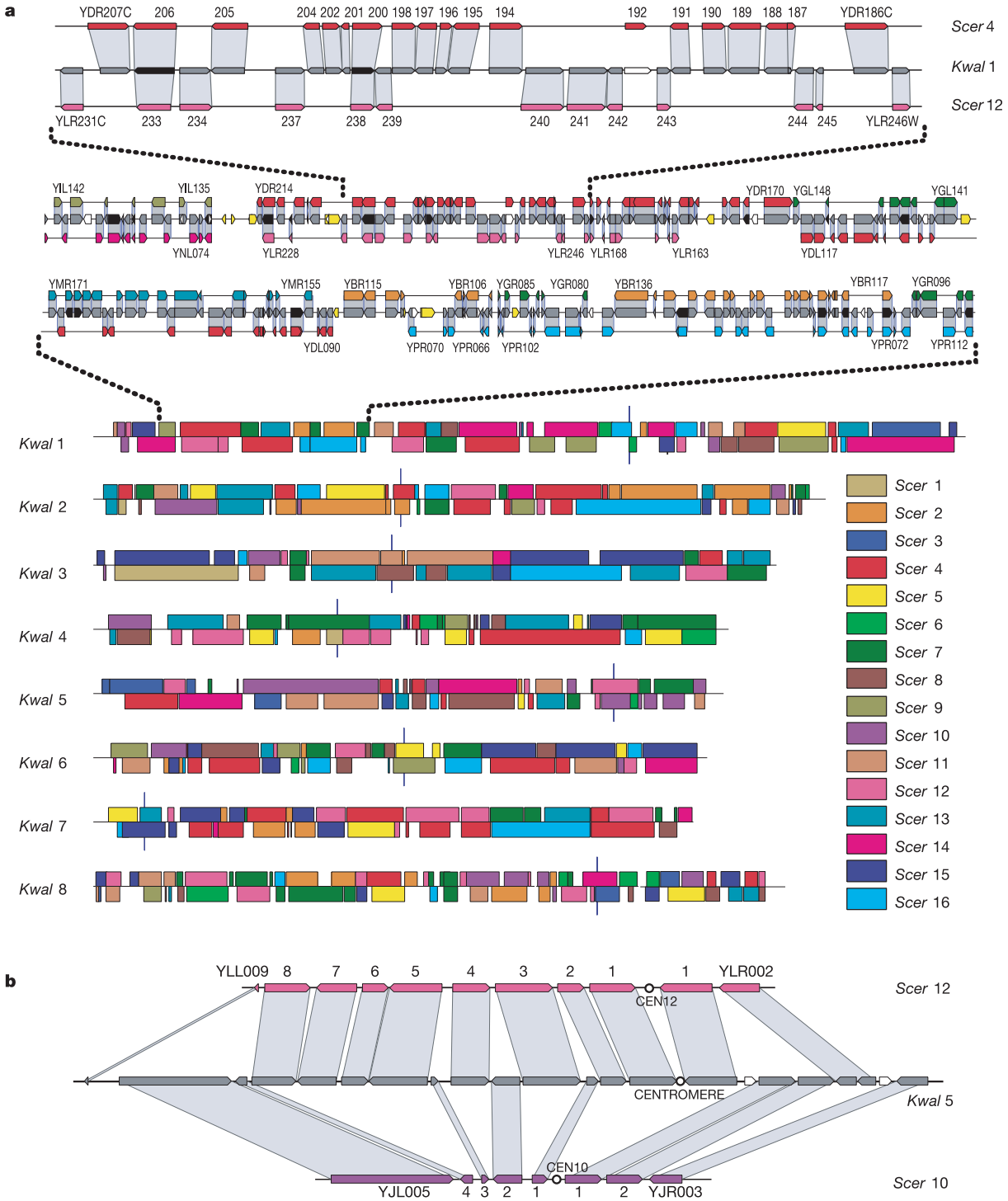

Figure 2 Gene and region correspondence with *K. waltii* reveals WGD. **a**, Each region of *K. waltii* (coloured by number of matches: white, 0; grey, 1; black, 2; yellow, > = 3) shows conserved gene order with two regions of *S. cerevisiae* (coloured by chromosome number). Spacing between *S. cerevisiae* genes is set to match *K. waltii* chromosomal positions. Vertical blue bars denote centromeres. **b**, DCS region showing duplicate mapping of centromeres (black circles). All sixteen *S. cerevisiae* centromeres show such duplicate mapping with *K. waltii* centromeres. This DCS region also illustrates that we can reliably recognize anciently duplicated segments even in the absence of any remaining two-copy genes. Evidence of WGD comes from gene interleaving and 2:1 mapping with orthologous *K. waltii* segments. The segments containing intervening genes are deleted, resulting in condensed sister regions. *Kwal*, *K. waltii*; *Scer*, *S. cerevisiae*.

complete duplicate mapping (Fig. 2b and Supplementary Information). In a typical DCS block, 90% of *K. waltii* genes have matches in at least one of the two *S. cerevisiae* regions (and on average ~12% have matches in both regions). Similarly, in each DCS block, 90% of genes from both *S. cerevisiae* regions have matches within the single corresponding region of *K. waltii*, and they are interleaved onto *K. waltii* while preserving gene order and orientation. The mapping fits the pattern expected for WGD followed by massive gene loss.

Complete pairing of *S. cerevisiae*

Based on the cross-species correspondence, we identified sister regions within *S. cerevisiae* (Fig. 3). We used the DCS blocks to define 253 initial sister regions for which the ancestral gene order was preserved in all three matching segments (*K. waltii* and the two sister regions in *S. cerevisiae*). We then merged consecutive sister regions when both copies in *S. cerevisiae* preserved the ancestral gene order, but were apparently separated owing to a rearrangement in *K. waltii*. This resulted in 145 sister regions in *S. cerevisiae* covering 88% of the genome.

Many of the paired regions could not be recognized without the relationship to *K. waltii*. For example, 47 DCS regions contain no duplicated genes whatsoever and the evidence of duplication comes solely from the interleaving of genes (Fig. 2b).

The chromosome structure of the two genomes provides further evidence of WGD. Indeed, the sixteen centromeres of *S. cerevisiae* are paired in a 2:1 manner with the eight centromeres of *K. waltii* (Fig. 2b and Methods).

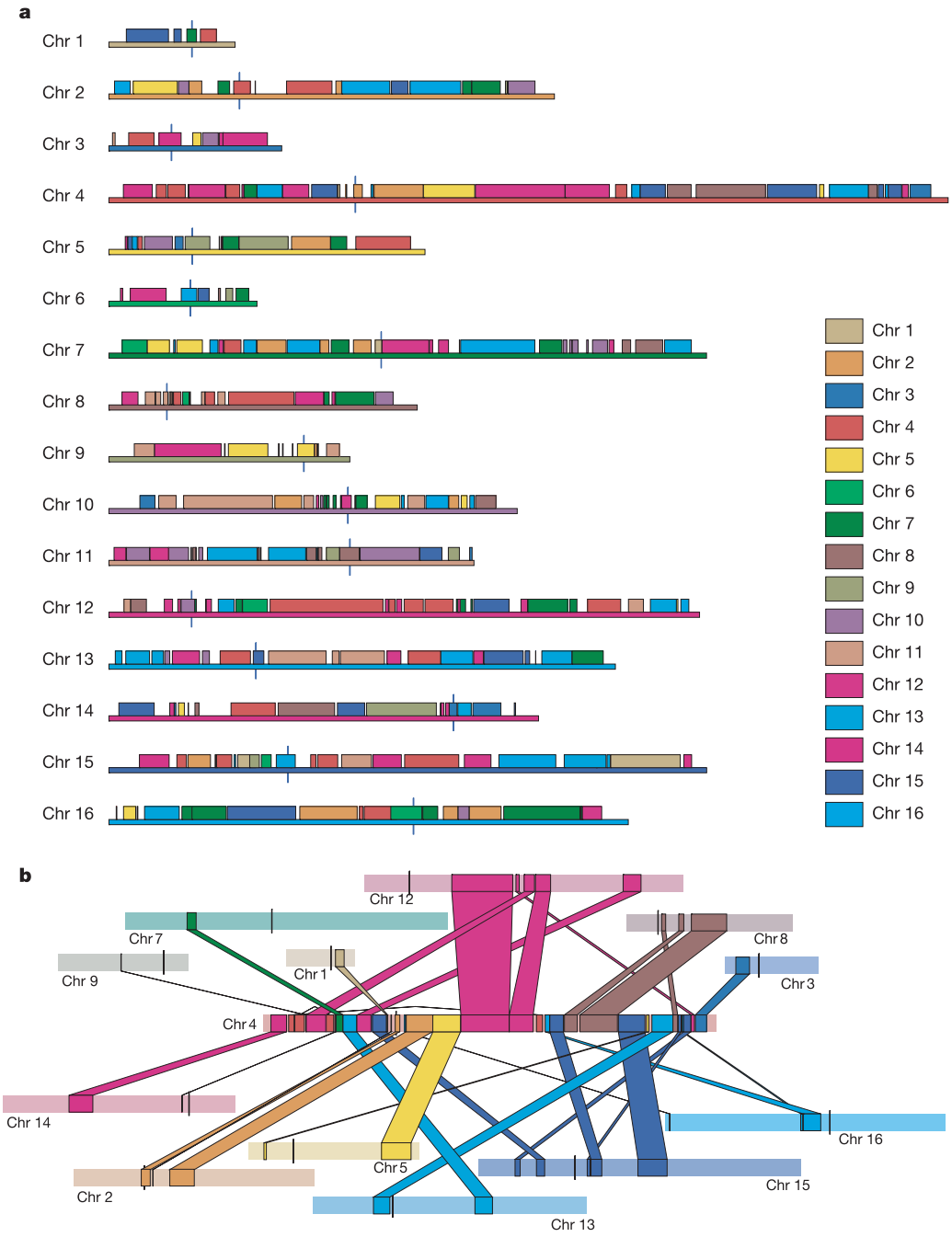

Figure 3 Duplicated blocks in *S. cerevisiae*. **a**, The duplicate mapping in sister regions extends to tile each *S. cerevisiae* chromosome, revealing complete duplication. Blue vertical bars denote centromeres. **b**, Detailed mapping of chromosome 4 with sister regions in other chromosomes.

NATURE | doi:10.1038/nature02424 | www.nature.com/nature

Summary

We conclude that a WGD event occurred in the *Saccharomyces* lineage after the divergence from *K. waltii* and that *S. cerevisiae* is therefore a degenerate tetraploid. WGD may have occurred either at the haploid or the diploid stage and either by endo-duplication (auto-polyploidy) or fusion of two close relatives (allo-polyploidy).

Evolutionary analysis

The relationship between *K. waltii* and *S. cerevisiae* provides the first comparison across an ancient WGD event and offers the opportunity to study the long-term fate of a genome after duplication.

Pattern of gene loss

The WGD event doubled the number of chromosomes in the *Saccharomyces* lineage, but subsequent gene-loss events led to the current *S. cerevisiae* genome, which is only 13% larger than *K. waltii* and contains only 10% more genes. The polyploid genome returned to functional normal ploidy, not by meiosis or chromosomal loss, but instead by a large number of deletion events. We can infer that 12% of the paralogous gene pairs were retained in each DCS block, and the remaining 88% of paralogous genes were lost.

In principle, gene loss can occur by large segmental deletions or individual gene deletions, and can be balanced between the two sisters or act primarily on one of them. Analysis of DCS blocks revealed that gene loss occurred by many small deletions (the average size of a lost segment is two genes), and was typically balanced between the two sister regions (average balance 57% to 43%). Additionally, a large number of rearrangement events have shuffled the duplicated chromosomes; the intergenic regions at the breakpoints seem to be associated with tRNA genes.

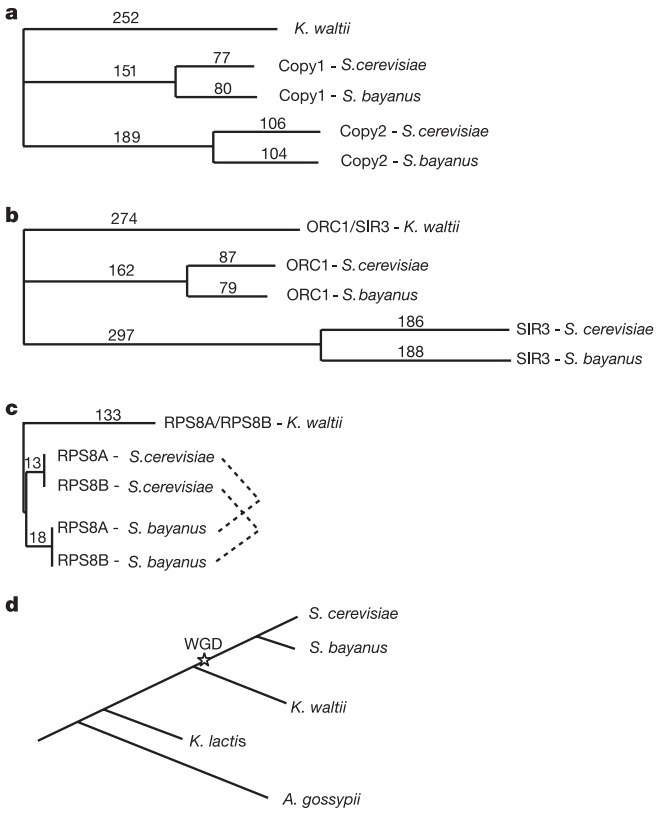

Figure 4 Divergence of duplicated gene pairs. Branches show number of substitutions per thousand amino acids. Evolutionary trees are rooted at the divergence of three species. **a**, Average protein divergence of all 457 gene pairs that arose by WGD. The faster-evolving paralogue is arbitrarily designated as copy 2 for each pair. **b**, Example of a gene showing accelerated protein divergence (1 of 72 cases). Ancestral and derived gene function can be inferred by comparison to *K. waltii*. In this case, the origin-of-replication recognition complex protein Orc1 is inferred to be ancestral and the silencing protein Sir3 is inferred to be derived. **c**, Example of duplicated gene pairs that have undergone recent gene conversion (1 of 60 cases). Comparison with *S. bayanus* shows that recent gene conversion events occurred both in *S. cerevisiae* and in *S. bayanus* lineages. Dotted lines connect orthologous genes in *S. cerevisiae* and in *S. bayanus*. **d**, Phylogeny and relative time of WGD. Estimated tree lengths are as reported in ref. 45.

Accelerated protein divergence

Most interestingly, we can study the evolution of the 457 gene pairs that arose by WGD. The synteny information allows us to unambiguously establish the ancestry for these genes pairs and distinguish them from many gene pairs in the yeast genome that arose by local duplication events (see Methods). This allows us to study their evolution with respect to each other and the non-duplicated *K. waltii* orthologue.

Two alternatives have been proposed for post-duplication divergence of duplicated gene pairs. Ohno has hypothesized that after duplication, one copy would preserve the original function and the other copy would be free to diverge[1]. However, this model has been contested by others, who argued that both copies would diverge more rapidly and acquire new functions[20–22]. We calculated divergence rates for the 457 pairs, using sequence information from *K. waltii*, *S. cerevisiae* and the related yeast *S. bayanus* (Fig. 4d). We searched for instances of unusually rapid or slow evolution.

We found that 76 of the 457 gene pairs (17%) show accelerated protein evolution relative to *K. waltii*, defined as instances in which the amino acid substitution rate along one or both of the *S. cerevisiae* branches was at least 50% faster than the rate along the *K. waltii* branch (Tables 1–3). We note that this calculation is conservative and may miss some cases of accelerated evolution (see Methods). These genes were heavily biased towards protein kinases (13 pairs, *P*-value 10^{-8}) and regulatory proteins (8 pairs, including the cell-cycle transcription factors Swi5 and Ace2, and the filamentation factors Phd1 and Sok2), and were generally involved in metabolism (38 pairs, *P*-value 10^{-18}) and cell growth (32 pairs, *P*-value 10^{-10}).

Ancestral and derived functions

Strikingly, in nearly every case (95%), accelerated evolution was confined to only one of the two paralogues. This strongly supports the model in which one of the paralogues retained an ancestral function while the other, relieved of this selective constraint, was free to evolve more rapidly[1].

For these 76 gene pairs showing accelerated evolution with respect to *K. waltii*, we can infer that the slowly evolving paralogue has probably retained the ancestral gene function and that the rapidly evolving paralogue probably acquired a derived function after duplication. In fact, the inference of ancestral and derived status can be extended to a total of 115 gene pairs consisting of those for which one *S. cerevisiae* paralogue has evolved at least 50% faster than the other *S. cerevisiae* paralogue.

This analysis thus identifies newer functions that evolved from older ones. We can infer, for example, that the silencing function of Sir3 (which acts in telomeres and mating type cassettes)[23] derived from the origin-of-replication binding function of Orc1 (ref. 24). Similarly, the anti-viral function of Ski7 that recognizes and represses non-polyadenylated messenger RNA was probably derived from the translation-elongation function of Hbs1 (ref. 25).

In many cases, the derived paralogues are specialized in their cellular localization or temporal expression. Acc1 (acetyl-CoA carboxylase) gave rise to Hfa1, which now functions in the mitochondrion[26]. Skt5, involved in cell-wall biosynthesis, gave rise to Shc1, whose function is restricted to spore formation[27]. And the pleiotropic transcription factor Skn7, involved in a number of stress responses, gave rise to the transcription factor Hms2, involved in pseudohyphal growth under nitrogen depletion[28].

Table 1 **Rapid protein divergence**

Slowly evolving gene (ancestral function)		Rapidly evolving gene (derived function)		Kwal rate	Relative rates	
Gene 1	Description 1	Gene 2	Description 2		Rate1 (%)	Rate2 (%)
RNR2	Small subunit of ribonucleotide reductase	*RNR4*	Ribonucleotide reductase	0.069	146	519
MCK1	Serine/threonine/tyrosine protein kinase	*YGK3*	Yeast homologue of mammalian glycogen synthase kinase 3	0.102	110	479
SEC14	Phosphatidylinositol transfer protein	*SFH1*	Sec14p homologue with highest sequence similarity but lowest functional redundancy	0.061	149	462
YSH1	Subunit of polyadenylation factor I (PF I)	*SYC1*	C-terminal portion of YSH1	0.130	112	380
GCS1	ADP-ribosylation factor GTPase-activating protein (ARF GAP)	*SPS18*	Transcription factor	0.172	112	309
FEN1	Probable subunit of 1,3-beta-glucan synthase	*ELO1*	Elongation enzyme 1	0.111	106	293
CET1	mRNA capping enzyme beta subunit (80 kDa)	*CTL1*	RNA triphosphatase, no capping function	0.231	76	263
EGD1	Regulator of pol II-transcribed genes	*BTT1*	Negative regulator of RNA polymerase II	0.146	109	240
SKN7	Transcription factor involved in oxidative stress	*HMS2*	Heat-shock transcription factor homologue	0.243	103	229
HBS1	Translation elongation factor	*SKI7*	Anti-viral protein	0.241	116	223
SEC24	Vesicle coat component	*SFB2*	Putative zinc finger protein	0.130	114	203
IFH1	Overexpression interferes with silencing at telomeres / HM, Interacts with FHL1	YDR223W	Hypothetical ORF	0.269	91	189
SKT5	Probable Ca^{2+} binding membrane protein (prenylated)	*SHC1*	Sporulation-specific homologue of csd4	0.242	93	188
ORC1	Largest subunit of origin recognition complex	*SIR3*	Silencing regulator HML, HMR, telomeres	0.274	91	177
BUB1	Serine/threonine protein kinase required for arrest in loss of microtubule function	*MAD3*	Spindle checkpoint complex subunit	0.249	133	172
ECM18	Involved in cell wall biogenesis	*ICT1*	Increased copper tolerance	0.224	98	161

Accelerated protein divergence of duplicated gene pairs (16 of 76 pairs). Kwal, amino acid substitution rate along *K. waltii* lineage. Rate1 and Rate2, substitution rates for Gene1 and Gene2 as a percentage of Kwal.

The functional distinction between the derived and ancestral paralogue is confirmed by the phenotype of deletion mutants when grown in rich medium conditions. For each of the 115 gene pairs, we examined the reported results of systematic deletion experiments[29,30]. Strikingly, deletion of the ancestral paralogue was lethal in 18% of cases (similar to the overall average for all *S. cerevisiae* genes), whereas deletion of the derived paralogue was never lethal. The derived paralogue is thus not essential under these conditions, either because it does not function in rich medium or because the ancestral paralogue can complement its function. Moreover, we infer that along with possibly gaining a new function, the derived copy has lost some essential aspect of its function, and cannot typically complement deletion of the ancestral gene.

Changes in codon usage

We also found that 32 of the 457 gene pairs show accelerated nucleotide, but not protein, evolution (Table 2). For example, the transcription elongation genes *YEF3* and *HEF3* encode very similar proteins and show a rate of protein change similar to that of *K. waltii*. However, *HEF3* shows twofold acceleration in degenerate third-codon positions (relative to both *YEF3* and *K. waltii*), an unusual pattern of codon usage (codon adaptation index[31] is 0.17 as compared to 0.78 for *YEF3*) and a much lower level of expression[32]. The same phenomenon is seen for the pyruvate kinase genes *CDC19* and *PYK2*. The accelerated divergence appears to reflect a relief from selection in the codon usage of the accelerated copy, and the presence of both gene copies may allow the cell to sense changes in metabolic state. Interestingly, *PYK2* is expressed preferentially in conditions of low glycolytic flux[33].

Gene conversion

At the other end of the spectrum, 60 of the 457 pairs show decelerated protein evolution, defined as instances where one or both copies evolved at least 50% slower than *K. waltii* (Table 3). These cases often involve proteins known to be highly constrained, such as ribosomal proteins (25 pairs), histone proteins (2 pairs) and translation initiation/elongation factors (4 pairs). In the vast majority of cases (90%), the paralogues both show decelerated evolution and tend to be very similar (98% amino acid identity versus 55% for all pairs), suggesting that they may be subject to periodic gene conversion. Indeed, the strong nucleotide similarity in fourfold degenerate codon positions that are not under the same selective pressures for conservation provides evidence supporting gene conversion (>90% nucleotide identity versus 41% for all pairs). Furthermore, in about half of these cases, the two paralogues in *S. cerevisiae* are closer in sequence to each other than either is to its syntenic orthologue in *S. bayanus*, showing that gene conversion has occurred after the relatively recent divergence of the two *Saccharomyces* lineages (Fig. 4c). For strongly conserved gene pairs, the two copies may confer an evolutionary benefit by serving as a repository for gene conversion and thus buffering deleterious changes in either copy.

Pairs with similar rates

The remaining 321 of the 457 gene pairs did not meet our stringent criteria for accelerated or decelerated evolution. In some cases, the functional changes may be similar to those above but subtler. In other cases, the primary functional divergence may be in regulatory sequences. This seems a likely possibility because the extensive gene

Table 2 **Rapid nucleotide divergence**

Gene 1	Gene 2	Description	Codon usage		Protein identity (%)	Rate in fourfold degenerate sites		
			CAI1	CAI2		Rate1	Rate2	Acceleration (%)
CDC19	*PYK2*	Pyruvate kinase	0.89	0.13	93	0.22	0.63	283
ADH1	*ADH5*	Alcohol dehydrogenase	0.81	0.25	91	0.35	0.57	164
YEF3	*HEF3*	Transcription elongation factor	0.78	0.17	92	0.34	0.62	181
GND1	*GND2*	Phosphogluconate dehydrogenase	0.62	0.20	96	0.39	0.64	166
GDH1	*GDH3*	NADP-specific glutamate dehydrogenase	0.59	0.16	95	0.41	0.62	150

Accelerated nucleotide divergence of duplicated gene pairs. CAI, codon adaptation index[31], a measure of typical codon usage between 0 and 1. Protein identity was measured between Gene1 and Gene2. Rate1 and Rate2, nucleotide substitution rate in fourfold degenerate sites for conserved amino acid residues. Acceleration, relative acceleration of Rate2 compared to Rate1.

Table 3 **Slowly evolving genes**

Gene 1	Gene 2	Description	Within *S. cerevisiae*		To *K. waltii*	
			All (%)	Fourfold (%)	All (%)	Fourfold (%)
TEF2	*TEF1*	Translational elongation factor EF-1 alpha	100	100	87	74
TIF2	*TIF1*	Translation initiation factor eIF4A	100	99	83	68
RPL1B	*RPL1A*	Ribosomal protein L1B	99	97	88	67
YCL073C	YKR106W	Hypothetical ORF	99	97	51	25
HHF1	*HHF2*	Histone H4 (HHF1 and HHF2 code for identical proteins)	97	92	89	71

Gene conversion events (5 of 60 cases). Nucleotide identity is shown within *S. cerevisiae* (between Gene1 and Gene2) and to *K. waltii* (between Gene2 and *K. waltii*). All, overall nucleotide divergence. Fourfold, nucleotide divergence in fourfold degenerate sites encoding conserved amino-acid residues. Supplementary Information contains complete divergence information for all gene pairs.

loss following WGD would probably juxtapose new combinations of intergenic and genic regions. In still other cases, the gene pairs may have been retained primarily to increase gene dosage, as with the slowly evolving genes.

Discussion

WGD followed by massive gene loss and gene specialization offers an important path for large-scale evolutionary innovation. Compared to multiple independent duplications and divergence of individual genes or segments, WGD may be more efficient and may offer great opportunities for coordinated evolution. Although organisms clearly can undergo WGD resulting in complete polyploids (as evidenced by existing tetraploid species of plants and animals[34–37]), it has been unclear whether WGD can then be followed by massive genome reshaping to yield diploids with expanded gene content.

Our analysis of *K. waltii* definitively proves that *S. cerevisiae* is the descendant of an ancient WGD, as originally proposed by Wolfe[3] on the basis of subtle genomic patterns. Comparison with *K. waltii* allows rigorous study of the many evolutionary innovations that arose in this dramatic event. We note that a similar analysis of the *Ashbya gossypii* genome has been carried out (P. Philippsen, personal communication). The results here suggest that it may also be fruitful to search for similar genomic signatures of ancient WGD in other organisms[38–40]. It will be interesting to see just how far such distant echoes of genomic upheaval may be traced. □

Methods

See the Supplementary Information for assembly contigs and scaffolds (Supplementary Table S1), predicted protein-coding genes (Supplementary Information S2), complete gene correspondence (Supplementary Information S3), a table of genes in syntenic segments (Supplementary Tables S4), gene-by-gene mapping for each chromosome (Supplementary Figs S5), detailed visualization of 253 DCS regions (Supplementary Figs S6), 145 paired regions in *S. cerevisiae* by chromosome (Supplementary Figs S7), alignments of duplicated gene pairs (Supplementary Information S8), protein and nucleotide divergence rates (Supplementary Tables S9) and mapping of centromeres (Supplementary Figs S10), also available from http://www.broad.mit.edu/seq/YeastDuplication.

Strains, sequencing and assembly

The sequence and annotation for *S. cerevisiae* reference strain S288C was obtained from http://www.yeastgenome.org, and updated according to recent comparative work[19]. *K. waltii* strain NCYC2644 was provided by I. Roberts (Norwich Research Park, UK). The WGS sequence was obtained at the Whitehead Institute/MIT Center for Genome Research. We used paired-end sequencing with 7.4-fold coverage in 4-kb plasmid clones and 0.6-fold coverage in 40-kb fosmid clones, providing eightfold sequence coverage and 42-fold physical coverage, with laboratory protocols as described at http://www.broad.mit.edu. We used the Arachne[41,42] program to assemble the resulting sequence information into contigs (continuous blocks of uninterrupted sequence) and scaffolds (contigs linked by paired forward-reverse reads from the same clone). The resulting draft genome sequence has a total length of 10.7 megabases, with ~96% contained in ten large scaffolds containing relatively short sequence gaps (average size 144 base pairs, together comprising only 0.5% of total scaffold length). Seven of the scaffolds appear to correspond to complete chromosomes, with recognizable centromeric and telomeric regions, and the remaining three scaffolds appear to belong to a single chromosome, with the order and orientation of the scaffolds readily inferred from the sequence (see below).

Centromeres and telomeres

The conservation of gene order in syntenic regions surrounding the centromeres allowed us to pinpoint candidate centromere-containing intergenic intervals within *K. waltii*. Sequence analysis of these intervals revealed a centromere structure similar to that of *S. cerevisiae*, containing three elements CDEI–CDEII–CDEIII, where CDEI is TCACCTG|TCANGTG, CDEII is 100–115 nucleotides with AT > 78%, and CDEIII is TTCCGA. This sequence occurs specifically only at the intervals we determined. Each *K. waltii* centromere maps to exactly two *S. cerevisiae* centromeres, as follows: Cen1:Cen7, Cen2:Cen4, Cen3:Cen14, Cen5:Cen9, Cen6:Cen16, Cen8:Cen11, Cen10:Cen12 and Cen13:Cen15. This pairing agrees with previous reports[16]. Telomeric repeat elements were identified in the ends of all scaffolds, except where joins are proposed. On the basis of centromere and telomere locations, paired fosmid ends, and synteny with *S. cerevisiae*, we infer that the three scaffolds together form the eighth *K. waltii* chromosome.

Annotation and gene correspondence

We determined the complete set of all predicted open reading frames (pORFs) in *K. waltii* that start with a methionine and have the potential to encode at least 50 amino acids. We found 6,753 non-overlapping ORFs, of which 5,230 are likely to represent functional protein-coding genes. These contain 4,711 ORFs of at least 150 amino acids, of which 92% show protein similarity to *S. cerevisiae*. From the remaining 2,042 ORFs between 50 and 150 amino acids, 519 show protein similarity to *S. cerevisiae*, and we expect these to encode functional proteins. We estimate that only an additional 45 ORFs at this range have been missed (0.8% of the total), on the basis of the percentage of longer ORFs that show similarity. We then constructed a bipartite graph of protein similarities across the two species, and resolved the gene correspondence based on protein similarity and gene order information using the BUS algorithm[43].

Doubly conserved synteny regions

We established blocks of conserved synteny by clustering matches of neighbouring genes that are less than 20 kb apart in each genome. We found 353 synteny blocks containing at least three conserved genes each, and matching a total of 4,590 *S. cerevisiae* genes and 4,139 *K. waltii* ORFs. We treated every synteny block as defining an interval on the *K. waltii* genome, and we marked all genes whose coordinates were contained within that interval. We then searched for *K. waltii* genes that were marked as belonging to multiple blocks, thus defining regions of overlap between synteny blocks (DCS regions).

Establishing ancestry of duplicated genes

We determined the ancestry of duplicated gene pairs solely on the basis of gene order information in DCS regions. The difficulty of reliably establishing ancestry based on amino acid divergence, especially in the presence of accelerated evolution, is illustrated by previous reports that have erroneously established WGD as preceding divergence from *K. lactis*[17] (the correct topology is shown in Fig. 4d), or erroneously inferred that GRS2 would represent the ancestral form of the GRS1/GRS2 pair[44] (in fact, GRS2 is the derived copy, and shows fourfold amino acid acceleration). In both cases, variation in protein divergence rates has misrepresented the true phylogenetic relationships.

Establishing accelerated divergence

Having established gene ancestry based on gene order information, we used amino acid and nucleotide divergence rates to detect cases of accelerated evolution. In seven of the 457 gene pairs, one or both copies in *S. cerevisiae* are split into multiple ORFs and were not used. We used CLUSTALW to construct protein, nucleotide, and codon-aware nucleotide alignments. We rooted three-way trees at the central node, whereas the true root lies on the *K. waltii* branch, and hence our criteria for accelerated divergence are conservative because they overestimate the *K. waltii* branch length. This is especially true in the case of gene-conversion events, and additional pre-duplication relatives will be necessary to estimate divergence between duplication and gene conversion.

Received 17 December 2003; accepted 19 January 2004; doi:10.1038/nature02424.
Published online 7 March 2004.

1. Ohno, S. *Evolution by Gene Duplication* (Allen and Unwin, London, 1970).
2. Mayer, V. W. & Aguilera, A. High levels of chromosome instability in polyploids of *Saccharomyces cerevisiae*. *Mutat. Res.* **231,** 177–186 (1990).
3. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387,** 708–713 (1997).
4. Seoighe, C. & Wolfe, K. H. Updated map of duplicated regions in the yeast genome. *Gene* **238,** 253–261 (1999).
5. Melnick, L. & Sherman, F. The gene clusters ARC and COR on chromosomes 5 and 10, respectively, of *Saccharomyces cerevisiae* share a common ancestry. *J. Mol. Biol.* **233,** 372–388 (1993).
6. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 546, 563–567 (1996).

7. Philippsen, P. *et al.* The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XIV and its evolutionary implications. *Nature* **387,** 93–98 (1997).
8. Mewes, H. W. *et al.* Overview of the yeast genome. *Nature* **387,** 7–65 (1997).
9. Coissac, E., Maillier, E. & Netter, P. A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. *Mol. Biol. Evol.* **14,** 1062–1074 (1997).
10. Friedman, R. & Hughes, A. L. Gene duplication and the structure of eukaryotic genomes. *Genome Res.* **11,** 373–381 (2001).
11. Hughes, T. R. *et al.* Widespread aneuploidy revealed by DNA microarray expression profiling. *Nature Genet.* **25,** 333–337 (2000).
12. Piskur, J. Origin of the duplicated regions in the yeast genomes. *Trends Genet.* **17,** 302–303 (2001).
13. Koszul, R., Caburet, S., Dujon, B. & Fischer, G. Eukaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J.* **23,** 234–243 (2004).
14. Llorente, B. *et al.* Genomic exploration of the hemiascomycetous yeasts: 18. Comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*. *FEBS Lett.* **487,** 101–112 (2000).
15. Llorente, B. *et al.* Genomic exploration of the hemiascomycetous yeasts: 20. Evolution of gene redundancy compared to *Saccharomyces cerevisiae*. *FEBS Lett.* **487,** 122–133 (2000).
16. Wong, S., Butler, G. & Wolfe, K. H. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc. Natl Acad. Sci. USA* **99,** 9272–9277 (2002).
17. Langkjaer, R. B., Cliften, P. F., Johnston, M. & Piskur, J. Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. *Nature* **421,** 848–852 (2003).
18. Kurtzman, C. P., Lynch, M. & Force, A. Phylogenetic circumscription of *Saccharomyces, Kluyveromyces* and other members of the Saccharomycetaceae, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygotorulaspora*. *FEMS Yeast Res.* **4,** 233–245 (2003).
19. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423,** 241–254 (2003).
20. Lynch, M. & Force, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154,** 459–473 (2000).
21. Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151,** 1531–1545 (1999).
22. Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Selection in the evolution of gene duplications. *Genome Biol.* **3,** RESEARCH0008.1–0008.9 (2002).
23. Shore, D., Squire, M. & Nasmyth, K. A. Characterization of two genes required for the position-effect control of yeast mating-type genes. *EMBO J.* **3,** 2817–2823 (1984).
24. Bell, S. P., Kobayashi, R. & Stillman, B. Yeast origin recognition complex functions in transcription silencing and DNA replication. *Science* **262,** 1844–1849 (1993).
25. Benard, L., Carroll, K., Valle, R. C., Masison, D. C. & Wickner, R. B. The ski7 antiviral protein is an EF1-alpha homolog that blocks expression of non-Poly(A) mRNA in *Saccharomyces cerevisiae*. *J. Virol.* **73,** 2893–2900 (1999).
26. Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425,** 686–691 (2003).
27. Sanz, M., Trilla, J. A., Duran, A. & Roncero, C. Control of chitin synthesis through Shc1p, a functional homologue of Chs4p specifically induced during sporulation. *Mol. Microbiol.* **43,** 1183–1195 (2002).
28. Lorenz, M. C. & Heitman, J. Regulators of pseudohyphal differentiation in *Saccharomyces cerevisiae* identified through multicopy suppressor analysis in ammonium permease mutant strains. *Genetics* **150,** 1443–1457 (1998).
29. Winzeler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285,** 901–906 (1999).
30. Gu, Z. *et al.* Role of duplicate genes in genetic robustness against null mutations. *Nature* **421,** 63–66 (2003).
31. Sharp, P. M. & Li, W. H. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15,** 1281–1295 (1987).
32. Sarthy, A. V. *et al.* Identification and kinetic analysis of a functional homolog of elongation factor 3, YEF3 in *Saccharomyces cerevisiae*. *Yeast* **14,** 239–253 (1998).
33. Boles, E. *et al.* Characterization of a glucose-repressed pyruvate kinase (Pyk2p) in *Saccharomyces cerevisiae* that is catalytically insensitive to fructose-1,6-bisphosphate. *J. Bacteriol.* **179,** 2987–2993 (1997).
34. Hughes, M. K. & Hughes, A. L. Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol. Biol. Evol.* **10,** 1360–1369 (1993).
35. Gallardo, M. H., Bickham, J. W., Honeycutt, R. L., Ojeda, R. A. & Kohler, N. Discovery of tetraploidy in a mammal. *Nature* **401,** 341 (1999).
36. Bailey, G. S., Poulter, R. T. & Stockwell, P. A. Gene duplication in tetraploid fish: model for gene silencing at unlinked duplicated loci. *Proc. Natl Acad. Sci. USA* **75,** 5575–5579 (1978).
37. Otto, S. P. & Whitton, J. Polyploid incidence and evolution. *Annu. Rev. Genet.* **34,** 401–437 (2000).
38. Blanc, G., Barakat, A., Guyot, R., Cooke, R. & Delseny, M. Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell* **12,** 1093–1101 (2000).
39. Sidow, A. Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.* **6,** 715–722 (1996).
40. Pebusque, M. J., Coulier, F., Birnbaum, D. & Pontarotti, P. Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol. Biol. Evol.* **15,** 1145–1159 (1998).
41. Batzoglou, S. *et al.* ARACHNE: a whole-genome shotgun assembler. *Genome Res.* **12,** 177–189 (2002).
42. Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: arachne 2. *Genome Res.* **13,** 91–96 (2003).
43. Kellis, M., Patterson, N., Birren, B., Berger, B. & Lander, E. S. Methods in comparative genomics: genome correspondence, gene identification, regulatory motif discovery. *J. Comput. Biol.* (in the press).
44. Turner, R. J., Lovato, M. & Schimmel, P. One of two genes encoding glycyl-tRNA synthetase in *Saccharomyces cerevisiae* provides mitochondrial and cytoplasmic functions. *J. Biol. Chem.* **275,** 27681–27688 (2000).
45. Kurtzman, C. P. & Robnett, C. J. Phylogenetic relationships among yeasts of the 'Saccharomyces complex' determined from multigene sequence analyses. *FEMS Yeast Res.* **3,** 417–432 (2003).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank I. Roberts for providing strains; M. Endrizzi, L.-J. Ma, D. Qi and the staff of the MIT/Whitehead Institute Center for Genome Research who generated the shotgun sequence from *Kluyveromyces waltii*; J. Butler and the Arachne assembly team who generated the genome assembly; D. Bartel, S. Calvo, J. Galagan, D. Jaffe and J. Vinson for discussions and comments on the manuscript; and G. Fink and A. Murray for discussions and advice.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.K. (manoli@mit.edu) and E.S.L. (lander@broad.mit.edu). The GenBank accession number for *K. waltii* is AADM01000000.