nature genetics

OPEN

Contrasting evolutionary genome dynamics between domesticated and wild yeasts

Jia-Xing Yue[1], Jing Li[1], Louise Aigrain[2], Johan Hallin[1], Karl Persson[3], Karen Oliver[2], Anders Bergström[2], Paul Coupland[2,5], Jonas Warringer[3], Marco Cosentino Lagomarsino[4], Gilles Fischer[4], Richard Durbin[2] & Gianni Liti[1]

Structural rearrangements have long been recognized as an important source of genetic variation, with implications in phenotypic diversity and disease, yet their detailed evolutionary dynamics remain elusive. Here we use long-read sequencing to generate end-to-end genome assemblies for 12 strains representing major subpopulations of the partially domesticated yeast *Saccharomyces cerevisiae* and its wild relative *Saccharomyces paradoxus*. These population-level high-quality genomes with comprehensive annotation enable precise definition of chromosomal boundaries between cores and subtelomeres and a high-resolution view of evolutionary genome dynamics. In chromosomal cores, *S. paradoxus* shows faster accumulation of balanced rearrangements (inversions, reciprocal translocations and transpositions), whereas *S. cerevisiae* accumulates unbalanced rearrangements (novel insertions, deletions and duplications) more rapidly. In subtelomeres, both species show extensive interchromosomal reshuffling, with a higher tempo in *S. cerevisiae*. Such striking contrasts between wild and domesticated yeasts are likely to reflect the influence of human activities on structural genome evolution.

Understanding how genetic variation translates into phenotypic diversity is a central theme in biology. With the rapid advancement of sequencing technology, genetic variation in large natural populations has been explored extensively for humans and several model organisms[1–9]. However, current knowledge of natural genetic variation is heavily biased toward single nucleotide variants (SNVs). Large-scale structural variants (SVs) such as inversions, reciprocal translocations, transpositions, novel insertions, deletions and duplications are not as well characterized owing to technical difficulties in detecting them with short-read sequencing data. This is a critical problem to address given that SVs often account for a substantial fraction of genetic variation and can have significant implications in adaptation, speciation and disease susceptibility[10–12].

The long-read sequencing technologies from Pacific Biosciences (PacBio) and Oxford Nanopore offer powerful tools for high-quality genome assembly[13]. Their recent applications provided highly continuous genome assemblies with many complex regions correctly resolved, even for large mammalian genomes[14,15]. This is especially important in characterizing SVs, which are frequently embedded in complex regions. For example, eukaryotic subtelomeres, which contribute to genetic and phenotypic diversity, are known hot spots of SVs due to rampant ectopic sequence reshuffling[16–19].

Baker's yeast, *S. cerevisiae*, is a leading biological model system with great economic importance in agriculture and industry. Discoveries in *S. cerevisiae* have helped shed light on almost every aspect of molecular biology and genetics. It was the first eukaryote to have its genome sequence, population genomics and genotype–phenotype map extensively explored[1,20,21]. Here we applied PacBio sequencing to 12 representative strains of *S. cerevisiae* or its wild relative *S. paradoxus* and identified notable interspecific contrasts in structural dynamics across their genomic landscapes. This study brings long-read sequencing technologies to the field of population genomics, studying genome evolution using multiple reference-quality genome sequences.

RESULTS

End-to-end population-level genome assemblies

We applied deep PacBio (100×–300×) and Illumina (200×–500×) sequencing to seven *S. cerevisiae* and five *S. paradoxus* strains representing evolutionarily distinct subpopulations of both species[1,6] (**Supplementary Tables 1** and **2**). The raw PacBio *de novo* assemblies of both nuclear and mitochondrial genomes showed compelling completeness and accuracy, with most chromosomes assembled into single contigs, and highly complex regions accurately assembled (**Supplementary Fig. 1**). After manual gap filling and Illumina-read-based error correction (Online Methods), we obtained end-to-end assemblies for almost all the 192 chromosomes, with only the rDNA array on chromosome XII and 26 of 384 (6.8%) chromosome ends remaining not fully assembled. We estimate that only 45–202 base-level sequencing errors remain across each 12-Mb nuclear genome (**Supplementary Tables 3** and **4**). For each assembly, we annotated

[1]Université Côte d'Azur, CNRS, INSERM, IRCAN, Nice, France. [2]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. [3]Department of Chemistry and Molecular Biology, Gothenburg University, Gothenburg, Sweden. [4]Laboratory of Computational and Quantitative Biology, Institut de Biologie Paris-Seine, UPMC University Paris 06, Sorbonne Universités, CNRS, Paris, France. [5]Present address: Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Cambridge, UK. Correspondence should be addressed to G.L. (gianni.liti@unice.fr).

Received 20 October 2016; accepted 22 March 2017; published online 17 April 2017; doi:10.1038/ng.3847

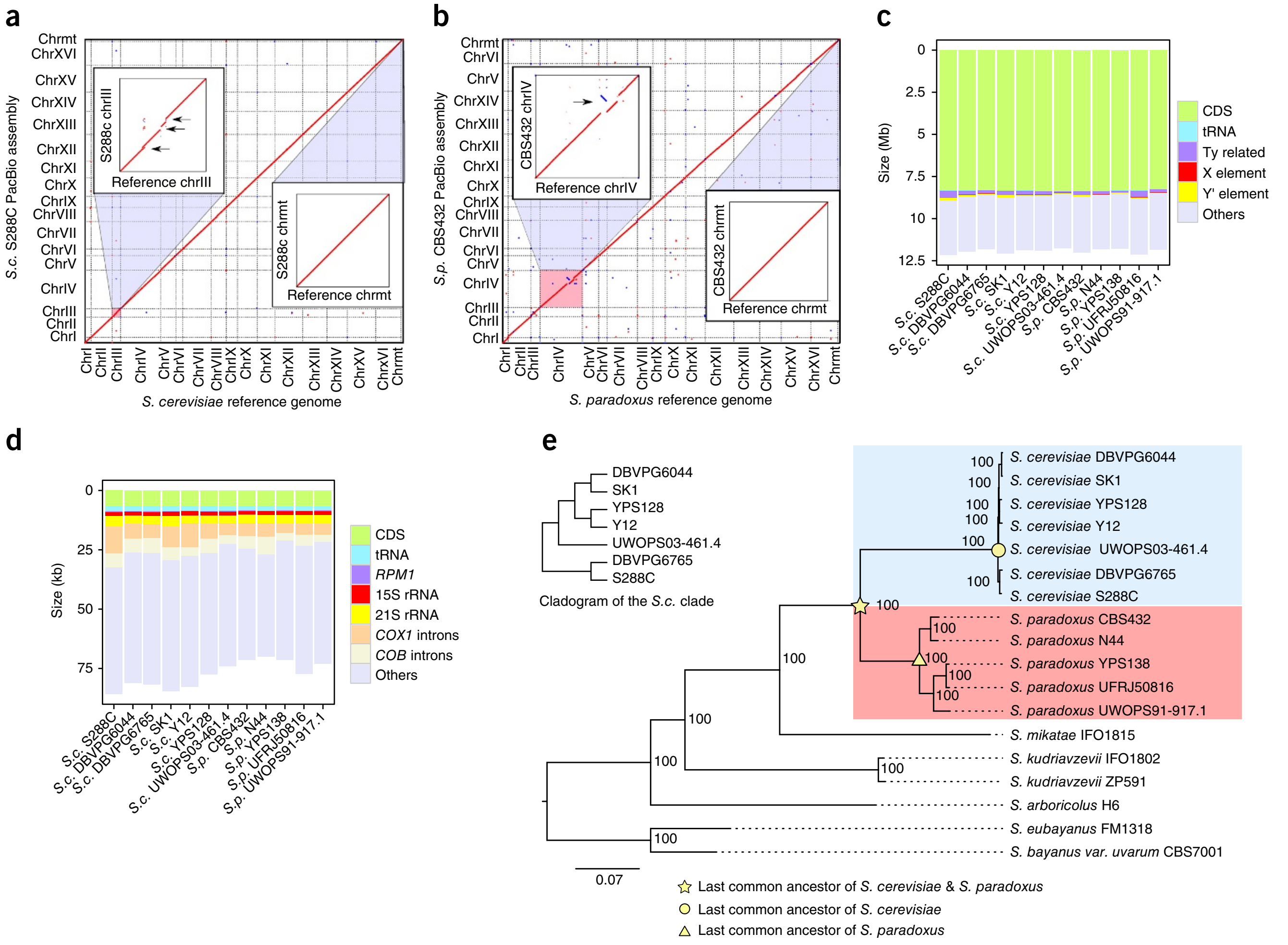

Figure 1 End-to-end genome assemblies and phylogenetic framework. (**a**) Comparison of the *S. cerevisiae* reference genome (strain S288C) and our S288C PacBio assembly. Sequence homology signals are indicated in red (forward match) or blue (reverse match). Insets, zoomed-in comparisons for chromosome III (chrIII) and the mitochondrial genome (chrmt). Black arrows indicate Ty-containing regions missing in the *S. cerevisiae* reference genome. (**b**) Comparison of the *S. paradoxus* reference genome (strain CBS432) and our CBS432 PacBio assembly, color coded as in **a**. Insets, zoomed-in comparison for chromosome IV (chrIV) and chrmt. Black arrow indicates the misassembly on chromosome IV in the *S. paradoxus* reference genome. (**c**,**d**) Cumulative lengths of annotated genomic features relative to the overall assembly size of the nuclear (**c**) and mitochondrial genome (**d**). CDS, coding sequence. (**e**) Phylogenetic relationships of the seven *S. cerevisiae* strains (blue) and five *S. paradoxus* strains (red) sequenced in this study. Six strains from other closely related *Saccharomyces* species were used as outgroups. All internal nodes have 100% fast-bootstrap supports. Inset, detailed relationships of the *S. cerevisiae* strains.

centromeres, protein-coding genes, tRNAs, Ty retrotransposable elements, core X elements, Y′ elements and mitochondrial RNA genes (**Supplementary Tables 5**–**7**). Chromosomes were named according to their encompassed centromeres.

When evaluated against the current *S. cerevisiae* and *S. paradoxus* reference genomes, our PacBio assemblies of the same strains (S288C and CBS432, respectively) show clean collinearity for both nuclear and mitochondrial genomes (**Fig. 1a**,**b**) with only a few discrepancies at finer scales, which were caused by assembly problems in the reference genomes. For example, we found five nonreference Ty1 insertions on chromosome III in our S288C assembly (**Fig. 1a**, inset), which were corroborated by previous studies[22–24] as well as our own long-range PCR amplifications. Likewise, we found a misassembly on chromosome IV (**Fig. 1b**, inset) in the *S. paradoxus* reference genome, which was confirmed by Illumina and Sanger reads[1]. Moreover, we checked several known cases of copy number variants (CNVs) (for example, Y′ elements[25], the *CUP1* locus[6] and *ARR*[6] gene clusters) and SVs (for example, those in the Malaysian *S. cerevisiae* UWOPS03-461.4)[26] and they were all correctly recaptured in our assemblies.

The final assembly sizes of these 12 strains ranged from 11.73 to 12.14 Mb for the nuclear genome (excluding rDNA gaps) and from 69.95 to 85.79 kb for the mitochondrial genome (**Fig. 1c**,**d** and **Supplementary Tables 8** and **9**). The abundance of Ty and Y′ elements substantially contributed to the nuclear genome size differences (**Fig. 1c** and **Supplementary Table 8**). For example, we observed strain-specific enrichment of full-length Ty1 in *S. cerevisiae* S288C, Ty4 in *S. paradoxus* UFRJ50816 and Ty5 in *S. paradoxus* CBS432, whereas no full-length Ty was found in *S. cerevisiae* UWOPS03-461.4 (**Supplementary Table 6**). Similarly, >30 copies of the Y′ element were found in *S. cerevisiae* SK1 but none in *S. paradoxus* N44 (**Supplementary Table 5**). Mitochondrial genome size variation is heavily shaped by the presence or absence of group I and group II introns in *COB1*, *COX1* and 21S rRNA *(rnl)* (**Fig. 1d** and **Supplementary Tables 9** and **10**). Despite large-scale interchromosomal rearrangements in a few strains (*S. cerevisiae* UWOPS03-461.4, *S. paradoxus* UFRJ50816 and *S. paradoxus* UWOPS91-917.1), all 12 strains maintained 16 nuclear chromosomes.

Molecular evolutionary rate and diversification timescale

To gauge structural dynamics in a well-defined evolutionary context, we performed phylogenetic analysis for the 12 strains and 6

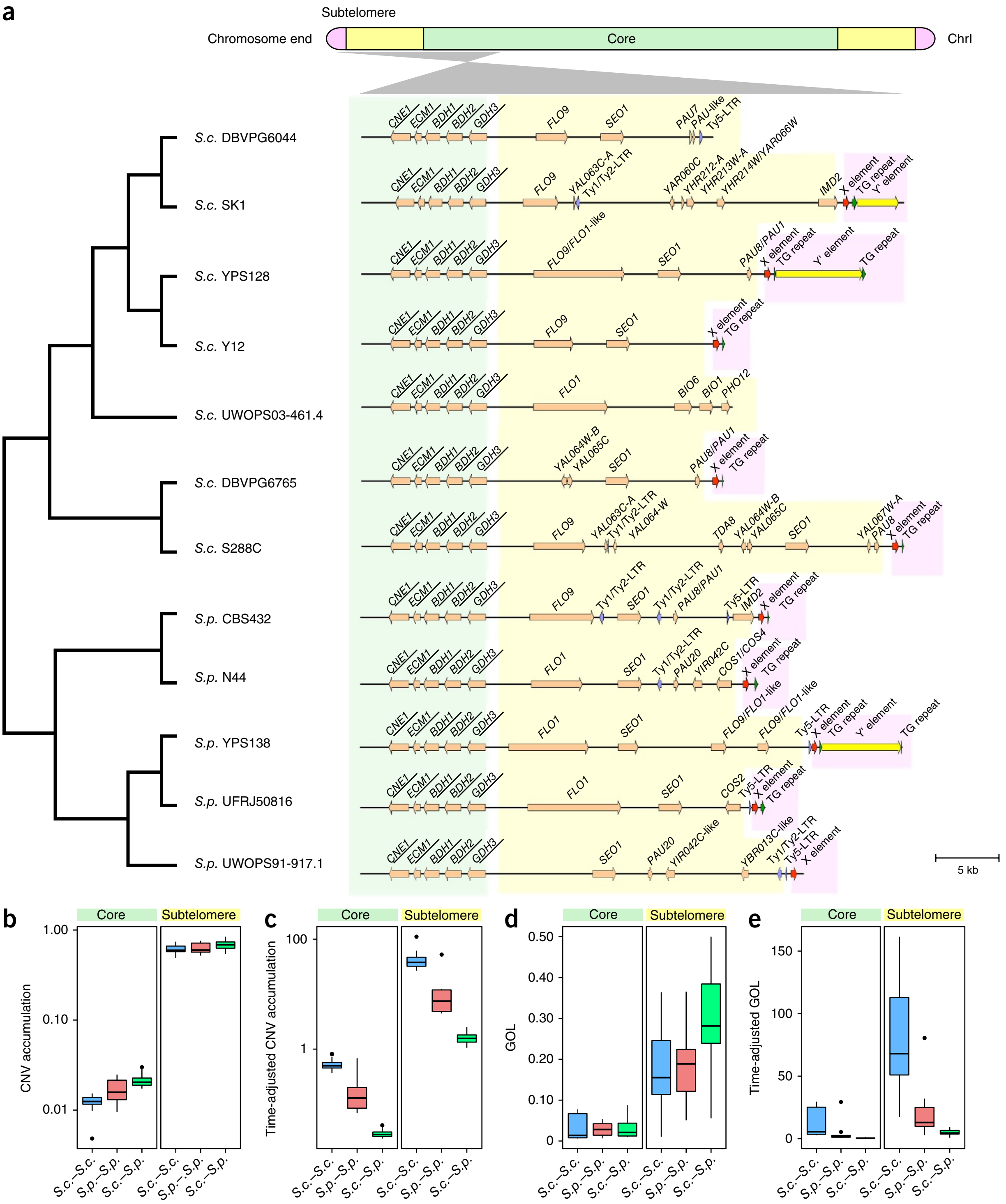

Figure 2 Explicit nuclear chromosome partitioning. (**a**) Partitioning of the left arm of chromosome I into the core (green), subtelomere (yellow) and chromosome end (pink) based on synteny conservation and the yeast telomere-associated core X and Y′ elements. Cladogram (left) shows the phylogenetic relationships of the 12 strains; gene arrangement map (right) illustrates the syntenic conservation profile in both the core and subtelomeric regions. The names of genes within the syntenic block are underlined. (**b**,**c**) CNV accumulation (**b**) and CNV accumulation adjusted by diversification time (**c**) of strain pairs within *S. cerevisiae (S.c.–S.p.)*, within *S. paradoxus* (*S.c.–S.p.*) and between the two species (*S.c.–S.p.*) ($\log_{10}$ scale). (**d**,**e**) GOL (**d**) and GOL adjusted by diversification time (**e**) of strain pairs. Center lines, median; boxes, interquartile range (IQR); whiskers, 1.5× IQR. Data points beyond the whiskers are outliers.

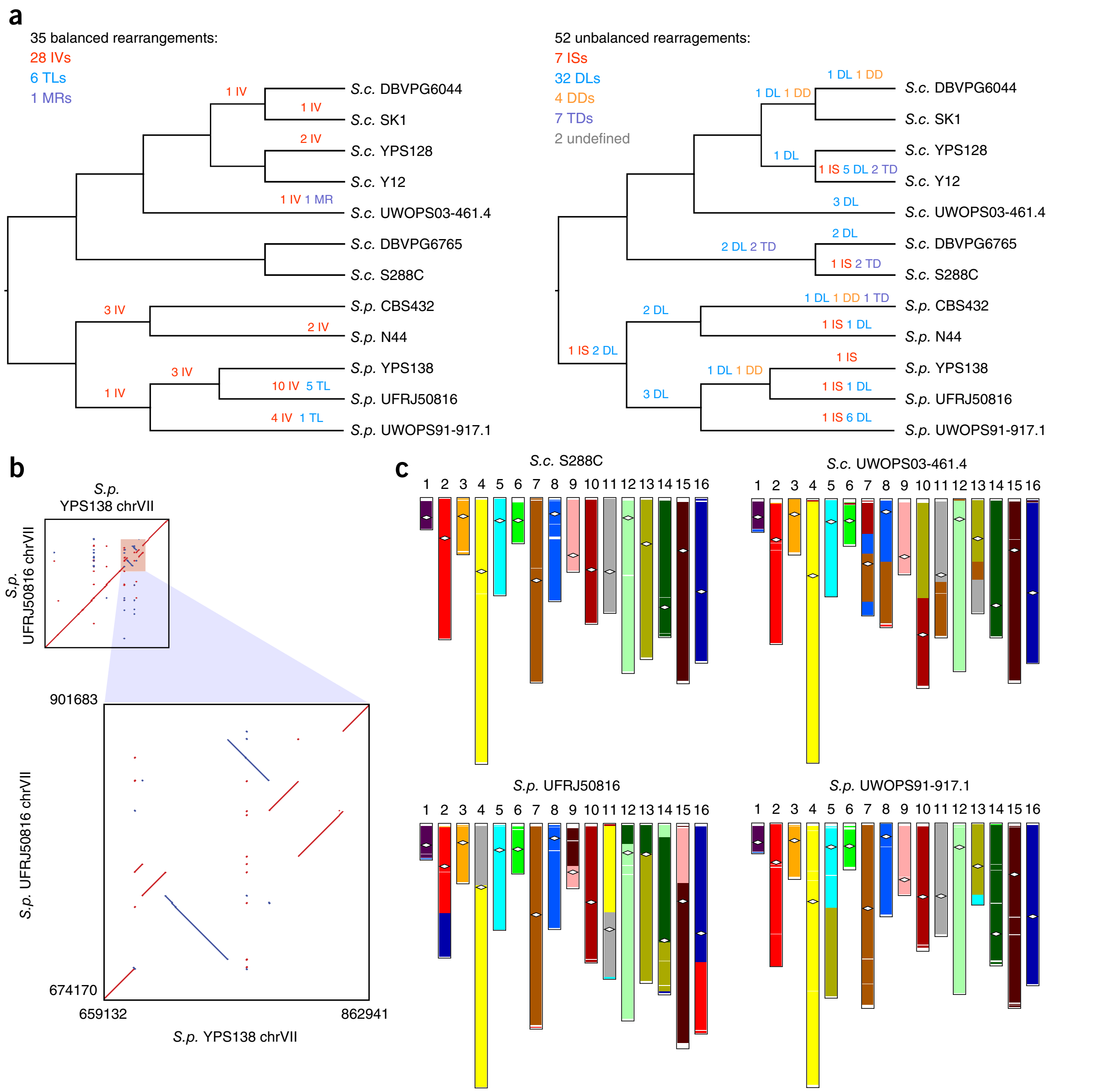

Figure 3 Structural rearrangements in the nuclear chromosomal cores. (**a**) Balanced (left) and unbalanced (right) structural rearrangements occurred along the evolutionary history of the 12 strains. IV, inversion; TL, translocation; MR, massive rearrangement; IS, insertion; DL, deletion; DD, dispersed duplication; TD, tandem duplication. (**b**) The six clustered inversions on chromosome VII (chrVII) of the *S. paradoxus* strain UFRJ50816; highlighted region (top) is shown in zoomed-in plot (bottom). (**c**) Genome organization of UWOPS03-461.4, UFRJ50816 and UWOPS91-917.1 relative to that of S288C, which is free from large-scale interchromosomal rearrangements. White diamonds indicate positions of centromeres. Different colors are used to differentiate gene contents in different ancestral *S. cerevisiae* chromosomes.

Saccharomyces sensu stricto outgroups based on 4,717 one-to-one orthologs of nuclear protein-coding genes (**Supplementary Data Set 1**). The resulting phylogeny is consistent with our prior knowledge about these strains (**Fig. 1e**). Analyzing this phylogenetic tree, we found the entire *S. cerevisiae* lineage to have evolved faster than the *S. paradoxus* lineage, as indicated by the overall longer branch from the common ancestor of the two species to each tip of the tree (**Fig. 1e**). We confirmed such rate differences by Tajima's relative rate test[27] for all *S. cerevisiae*–*S. paradoxus* strain pairs, using *Saccharomyces mikatae* as the outgroup ($P < 1 \times 10^{-5}$ for all pairwise comparisons). In contrast, molecular dating analysis shows that the cumulative diversification time for the five *S. paradoxus* strains was 3.87-fold that for the seven *S. cerevisiae* strains, suggesting a much longer time span for accumulating species-specific genetic changes in the former (**Supplementary Fig. 2a**). This timescale difference was further supported by the synonymous substitution rate (dS) (**Supplementary Fig. 2b**).

Core–subtelomere chromosome partitioning

Conceptually, linear nuclear chromosomes can be partitioned into internal chromosomal cores, interstitial subtelomeres and terminal chromosome ends. However, their precise boundaries are challenging to demarcate without a rigid subtelomere definition. Here we propose an explicit way to pinpoint yeast subtelomeres on the basis of multi-genome comparison, which can be further applied to other eukaryotic organisms. For each subtelomere, we located its proximal boundary on the basis of the sudden loss of synteny conservation and

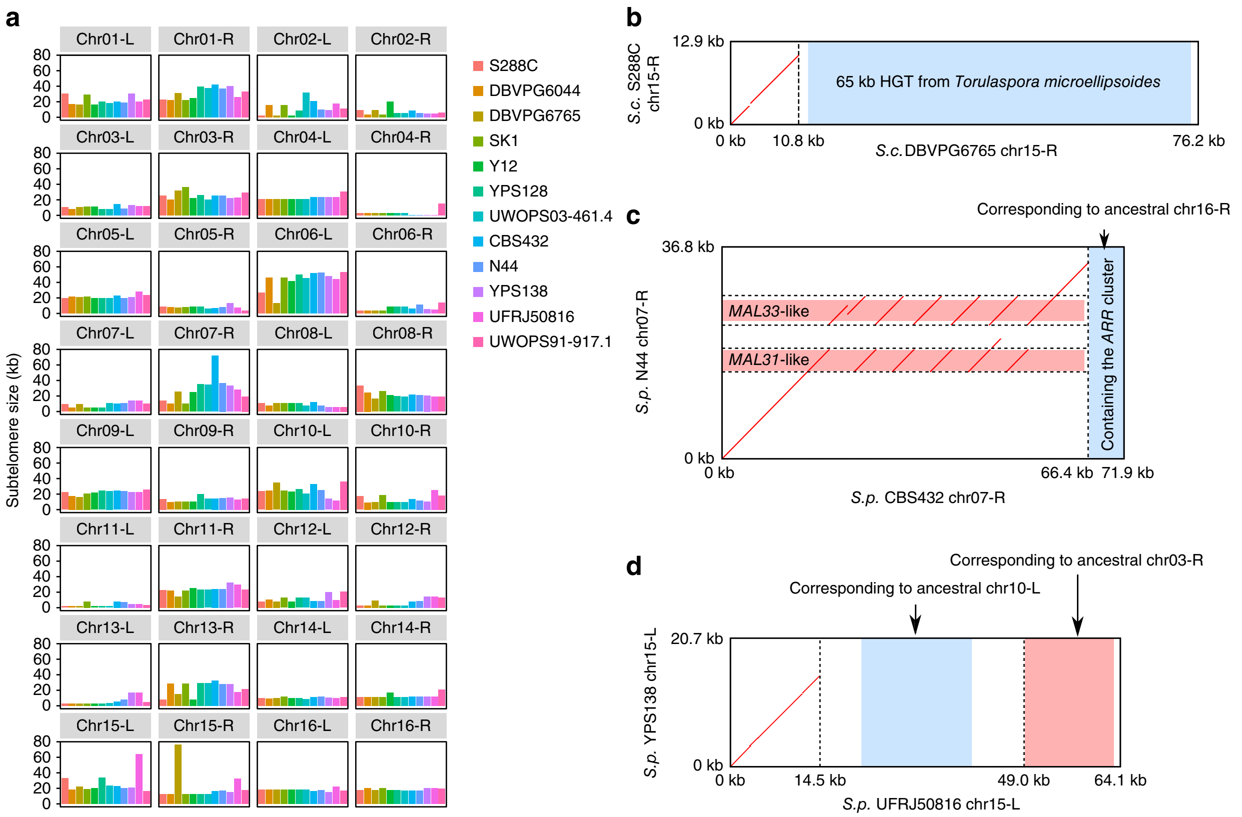

Figure 4 Subtelomere size plasticity and structural rearrangements. (**a**) Size variation of the 32 orthologous subtelomeres across the 12 strains. (**b**) Chromosome 15-R (chr15-R) subtelomere comparison between *S. cerevisiae* DBVPG6765 and S288C. The extended DBVPG6765 chr15-R subtelomere is explained by a eukaryote-to-eukaryote HGT event[39]. (**c**) Chromosome 07-R (chr07-R) subtelomere comparison between *S. paradoxus* CBS432 and N44. The chr07-R subtelomere expansion in CBS432 is explained by a series of tandem duplications of the *MAL31*-like and *MAL33*-like genes and an addition of the *ARR*-containing segment from the ancestral chromosome 16-R subtelomere. (**d**) Chromosome 15-L subtelomere comparison between *S. paradoxus* UFRJ50816 and YPS138. The expanded chromosome 15-L subtelomere in UFRJ50816 is explained by the relocated subtelomeric segments from the ancestral chromosome 10-L and chromosome 03-R subtelomeres. Region coordinates in **b**–**d** are based on the defined subtelomeres rather than the full chromosomes.

demarcated its distal boundary by the telomere-associated core X and Y′ elements (Online Methods and **Supplementary Fig. 3**). The partitioning for the left arm of chromosome I is illustrated in **Figure 2a**. The strict gene synteny conservation is lost after *GDH3*, thus marking the boundary between the core and the subtelomere for this chromosome arm (**Fig. 2a**). All chromosomal cores and subtelomeres and 358 out of 384 chromosome ends across the 12 strains could be defined in this way (**Supplementary Tables 11–13** and **Supplementary Data Sets 2** and **3**). For the remaining 26 chromosome ends, X and Y′ elements and telomeric repeats (TG_{1-3}) were missing. We assigned the orthology of subtelomeres from different strains on the basis of the ancestral chromosomal identity of their flanking chromosomal cores (Online Methods). Here we use Arabic numbers to denote such ancestral chromosomal identities and the associated subtelomeres, taking into account the large-scale interchromosomal rearrangements that have occurred in some strains (**Supplementary Fig. 4** and **Supplementary Table 12**). Such accurately assigned subtelomere orthology, together with explicit chromosome partitioning, allows an in-depth examination of subtelomeric evolutionary dynamics.

Our analysis captures distinct properties of chromosomal cores and subtelomeres. All previously defined essential genes in *S. cerevisiae* S288C[28] fell into the chromosomal cores, whereas all previously described subtelomeric duplication blocks in S288C (http://www2.le.ac.uk/colleges/medbiopsych/research/gact/images/clusters-fixed-large.jpg) were fully enclosed in our defined S288C subtelomeres. Furthermore, the genes from our defined subtelomeres showed 36.6-fold higher CNV accumulation than those from the cores (one-sided Mann–Whitney *U* test, $P < 2.2 \times 10^{-16}$) (**Fig. 2b,c**). When considering only one-to-one orthologs, the subtelomeric genes showed 8.4-fold higher gene order loss (GOL)[29–31] than their core counterparts (one-sided Mann–Whitney *U* test, $P < 2.2 \times 10^{-16}$) (**Fig. 2d,e**). Additionally, subtelomeric one-to-one orthologs also showed significantly higher nonsynonymous-to-synonymous substitution rate ratio (dN/dS) than those from the cores in the *S. cerevisiae*–*S. cerevisiae* and *S. cerevisiae*–*S. paradoxus* comparisons (one-sided Mann–Whitney *U* test, $P < 2.2 \times 10^{-16}$), although no clear trend was found in the *S. paradoxus*–*S. paradoxus* comparison (one-sided Mann–Whitney *U*-test, $P = 0.936$). These observations fit well with known properties of cores and subtelomeres and provide the first quantitative assessment of the core–subtelomere contrasts in genome dynamics. Notably, aside from such core–subtelomere contrasts, we also observed clear interspecific differences in all three measurements. *S. cerevisiae* strains showed faster CNV accumulation (one-sided Mann–Whitney *U*-test; $P = 6.7 \times 10^{-5}$ for cores, $P = 5.1 \times 10^{-5}$ for subtelomeres) and more rapid GOL (one-sided Mann–Whitney *U*-test, $P = 5.5 \times 10^{-5}$ for cores and $P = 2.6 \times 10^{-5}$ for subtelomeres) than *S. paradoxus* strains in both core and subtelomeres, respectively (**Fig. 2c,e**). Similarly, *S. cerevisiae* subtelomeric genes also showed higher dN/dS than their *S. paradoxus* counterparts (one-sided Mann–Whitney *U*-test, $P = 4.3 \times 10^{-4}$), although their core genes appear to have similar dN/dS (one-sided Mann–Whitney *U*-test, $P = 1.000$). These observations collectively suggest accelerated evolution in *S. cerevisiae* relative to *S. paradoxus*, especially in subtelomeres.

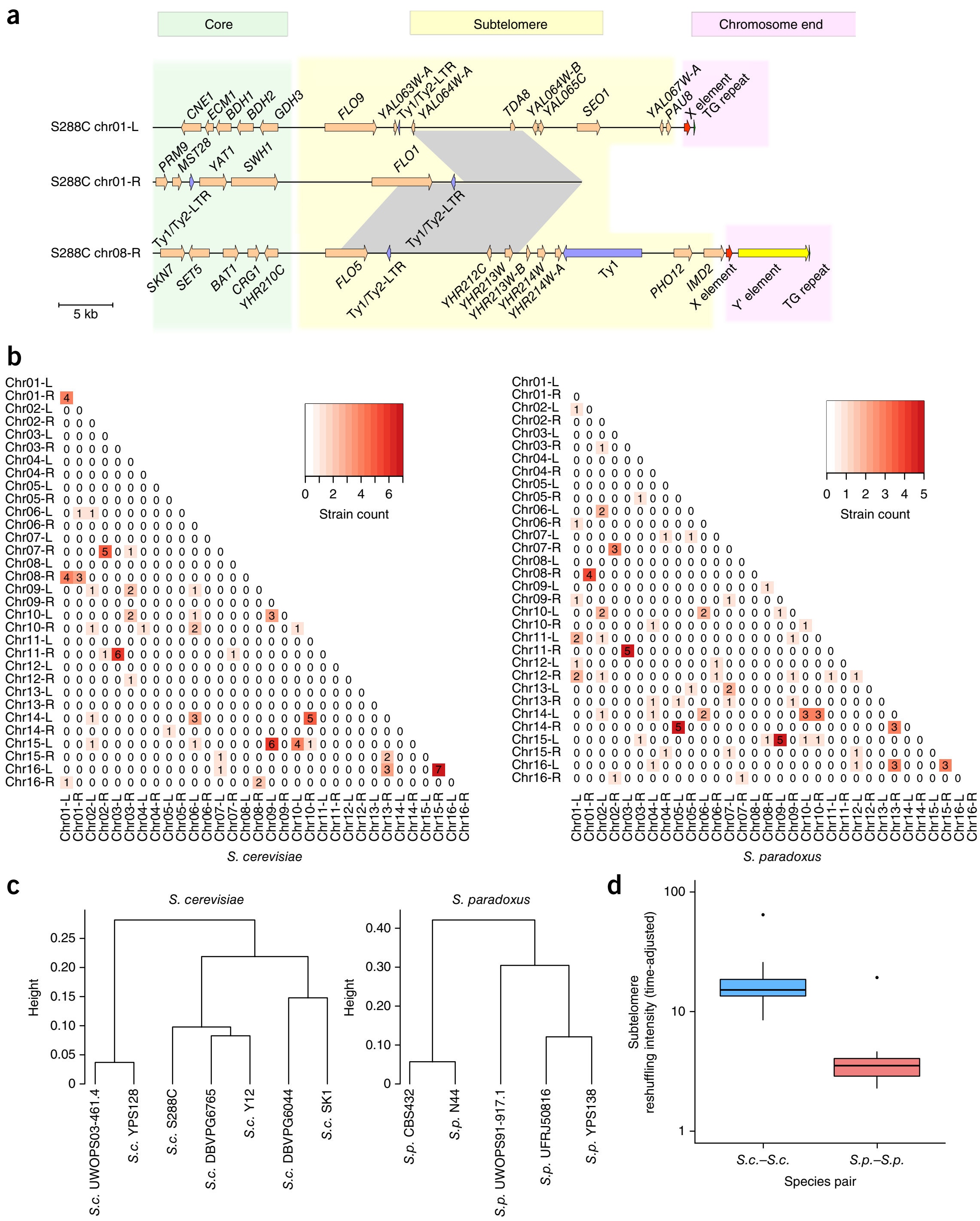

Figure 5 Evolutionary dynamics of subtelomeric duplications. (**a**) An example of subtelomeric duplication blocks shared among the chromosome 01-L (chr01-L), chr01-R and chr08-R subtelomeres in *S. cerevisiae* S288C. Gray shading indicates shared homologous regions with ≥90% sequence identity. (**b**) Subtelomeric duplication signals shared across strains. For each subtelomere pair, the number of strains showing strong sequence homology (BLAT score ≥5,000 and identity ≥90%) is indicated in the heat map. (**c**) Hierarchical clustering based on the proportion of conserved orthologous subtelomeres in cross-strain comparisons within *S. cerevisiae* and *S. paradoxus*. (**d**) Subtelomere reshuffling intensities ($\log_{10}$ scale) within *S. cerevisiae* (*S.c.*–*S.c.*) and within *S. paradoxus* (*S.p.*–*S.p.*), adjusted by the diversification time of the compared strain pair. Center lines, median; boxes, interquartile range (IQR); whiskers, 1.5× IQR. Data points beyond the whiskers are outliers.

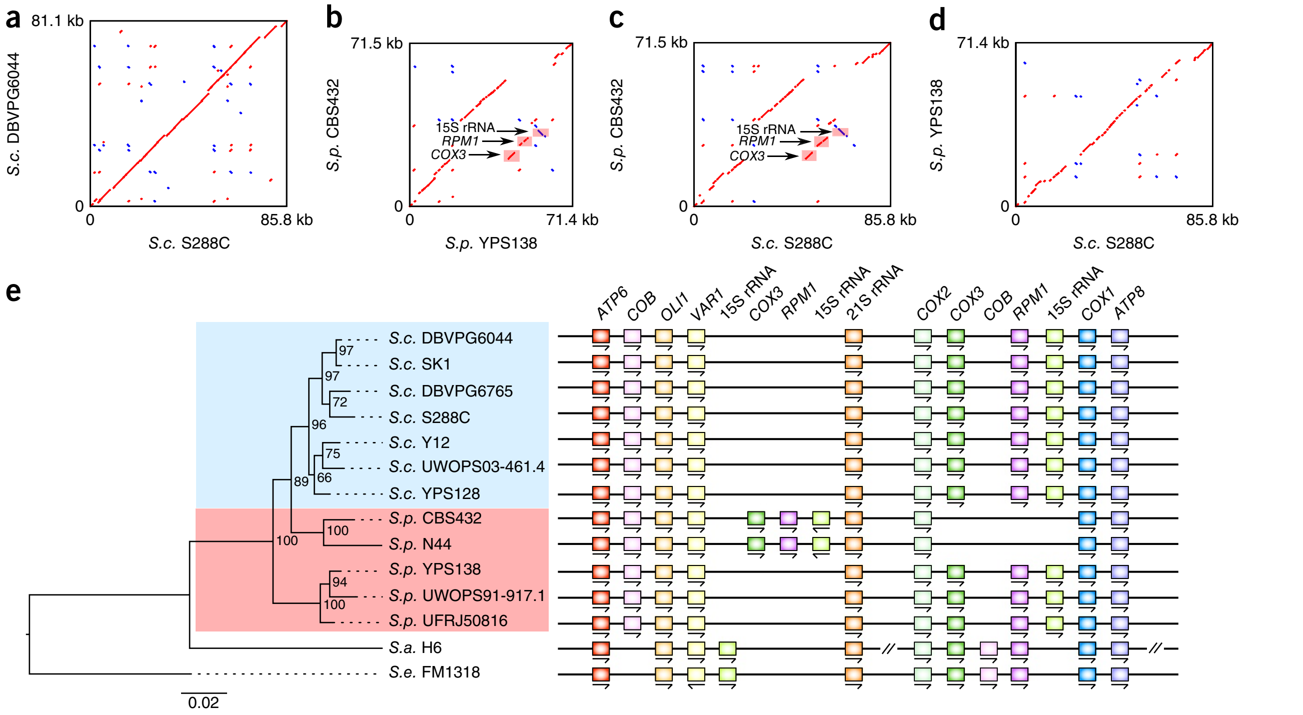

Figure 6 Comparative mitochondrial genomics. (**a**–**d**) Pairwise comparisons for the mitochondrial genomes of S288C and DBVPG6044 from *S. cerevisiae* (**a**), CBS432 and YPS138 from *S. paradoxus* (**b**), *S. cerevisiae* S288C and *S. paradoxus* CBS432 (**c**) and *S. cerevisiae* S288C and *S. paradoxus* YPS138 (**d**). (**e**) Genomic arrangement of the mitochondrial protein-coding genes and RNA genes across the 12 sampled strains. Left, phylogenetic tree constructed on the basis of mitochondrial protein-coding genes, with the number at each internal node showing rapid bootstrap support. The detailed gene arrangement map is shown on the right. There is a large inversion in *S. arboricolus* that encompasses the entire *COX2*–*ATP8* (according to its original mitochondrial genome assembly), which we inverted back this segment for better visualization.

Structural rearrangements in chromosomal cores

Structural rearrangements can be balanced (as with inversions, reciprocal translocations and transpositions) or unbalanced (as with large-scale novel insertions, deletions and duplications) depending on whether the copy number of genetic material is affected[10]. We identified 35 balanced rearrangements in total, including 28 inversions, 6 reciprocal translocations and 1 massive rearrangement (**Fig. 3a**, **Supplementary Fig. 5a**–**c** and **Supplementary Data Set 4**). All events occurred during the species-specific diversification of the two species, with 29 events occurring in *S. paradoxus* and only 6 in *S. cerevisiae*. Factoring in the cumulative evolutionary time difference, *S. paradoxus* still showed 1.25-fold faster accumulation of balanced rearrangements than *S. cerevisiae*. Six inversions were tightly packed into a ~200-kb region on chromosome VII of South American *S. paradoxus* UFRJ50816, indicating a strain-specific inversion hot spot (**Fig. 3b**). With regard to interchromosomal rearrangements, six were reciprocal translocations that occurred in two *S. paradoxus* strains (**Fig. 3c** and **Supplementary Fig. 5a,b**). The remaining one, in the Malaysian *S. cerevisiae* UWOPS03-461.4, was particularly notable: chromosomes VII, VIII, X, XI and XIII were heavily reshuffled, confirming recent chromosomal contact data[26] (**Fig. 3c** and **Supplementary Fig. 5c**). We describe this as a massive rearrangement because it cannot be explained by typical independent reciprocal translocations but is more likely to result from a single catastrophic event resembling the chromothripsis observed in tumor cells[32,33]. This massive rearrangement in the Malaysian *S. cerevisiae* and the rapid accumulation of inversions and translocations in the South American *S. paradoxus* resulted in extensively altered genome configurations, explaining the reproductive isolation of these two lineages[34,35]. As previously observed in yeasts on larger divergence scales[36,37], the breakpoints of those balanced rearrangements are associated with tRNAs and Tys, highlighting the roles of these elements in triggering genome instability and suggesting nonallelic homologous recombination as the mutational mechanism.

Considering unbalanced structural rearrangements in chromosomal cores, we identified 7 novel insertions, 32 deletions, 4 dispersed duplications and at least 7 tandem duplications (**Fig. 3a** and **Supplementary Data Set 5**). There were two additional cases of which the evolutionary history could not be confidently determined owing to multiple potential independent origins or secondary deletions (**Supplementary Data Set 5**). Although this is a conservative estimate, our identified unbalanced structural rearrangements clearly outnumbered the balanced ones, as recently reported in *Lachancea* yeasts[38]. We found that *S. cerevisiae* accumulated as many unbalanced rearrangements as *S. paradoxus* despite its much shorter cumulative diversification time. We noticed that the breakpoints of these unbalanced rearrangements (except for tandem duplications) were also frequently associated with Tys and tRNAs, mirroring our observation for balanced rearrangements. Finally, we found genes involved in unbalanced rearrangements to be significantly enriched for Gene Ontology (GO) terms related to the binding, transporting and detoxification of metal ions (for example, Na^+, K^+, Cd^{2+} and Cu^{2+}) (**Supplementary Table 14**), hinting that these events are probably adaptive.

Structural evolutionary dynamics of subtelomeres

The complete assemblies and well-defined subtelomere boundaries enabled us to examine subtelomeric regions with unprecedented resolution. We found both the size and gene content of the subtelomere to be highly variable across different strains and chromosome arms (**Fig. 4a** and **Supplementary Data Set 3**). The subtelomere size ranged from

0.13 to 76 kb (median = 15.6 kb), the number of genes enclosed in each subtelomere varied between 0 and 19 (median = 4), and the total number of subtelomeric genes varied between 134 and 169 (median = 146) per strain. Whereas the very short subtelomeres (for example, chromosome 04-R and chromosome 11-L) can be explained by an unexpected high degree of synteny conservation extending all the way to the end, some exceptionally long subtelomeres are the products of multiple mechanisms. For example, the chromosome 15-R subtelomere

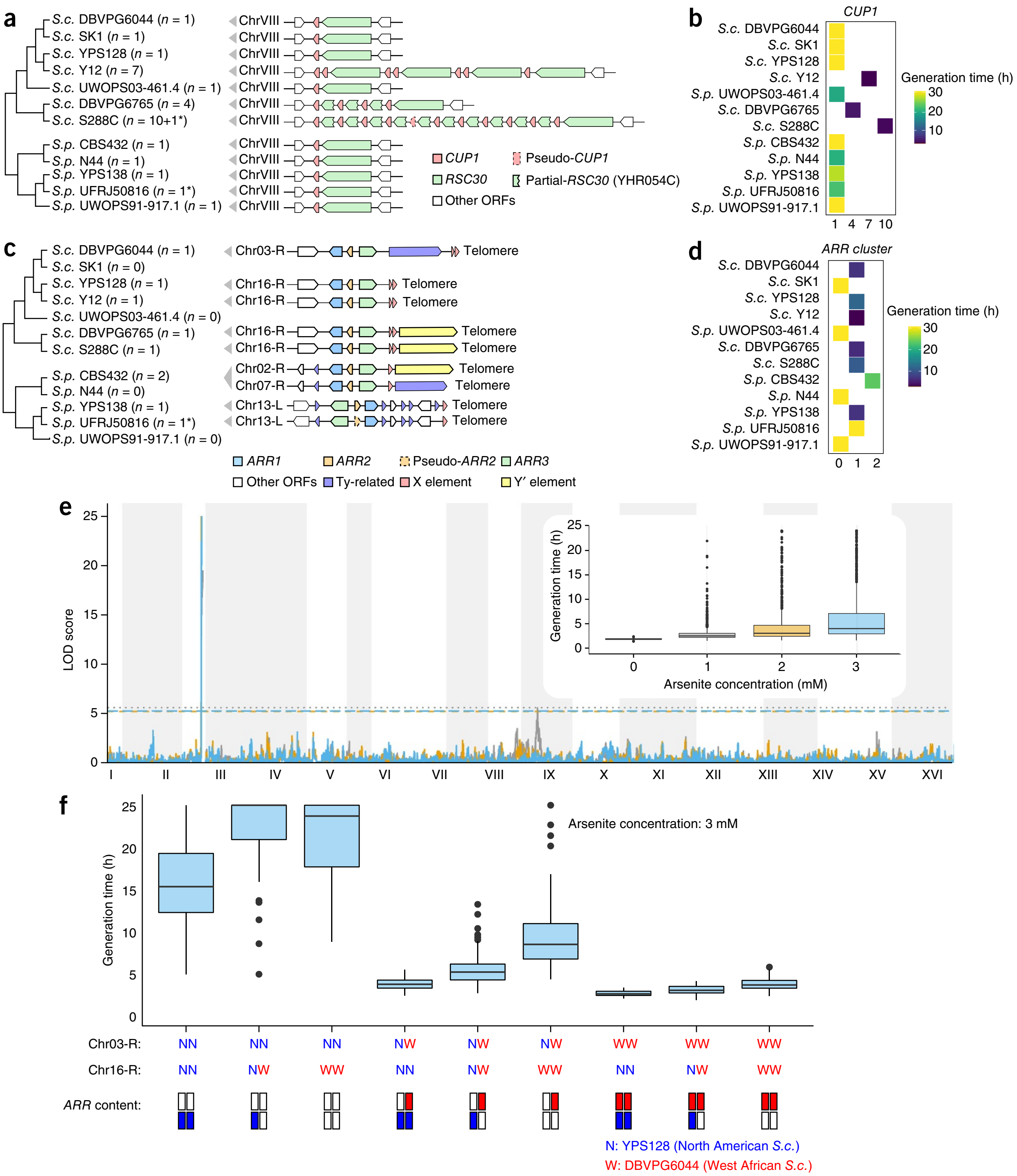

Figure 7 Structural rearrangements illuminate complex phenotypic variation. (**a**–**d**) Copy number and gene arrangement of the *CUP1* locus (**a**) and the *ARR* cluster (**c**) across the 12 strains (asterisks denote involvement of pseudogenes), and generation time of the 12 strains in high-copper (**b**) and high-arsenic conditions (**d**). (**e**) The rearrangement that relocates the *ARR* cluster to the chromosome 03-R (chr03-R) subtelomere in the West African *S. cerevisiae* DBVPG6044 is consistent with the linkage mapping analysis using phased outbred lines (POLs) derived from North American (YPS128) and West African (DBVPG6044) *S. cerevisiae*. (**f**) Phenotypic distribution of the 826 POLs for generation time in arsenic condition partitioned for genotype positions at the chr03-R and chr16-R subtelomeres and inferred copies of *ARR* clusters (bottom). Center lines, median; boxes, interquartile range (IQR); whiskers, 1.5× IQR. Data points beyond the whiskers are outliers.

of *S. cerevisiae* DBVPG6765 has been drastically elongated by a 65-kb horizontal gene transfer (HGT)[39] (**Fig. 4b** and **Supplementary Fig. 6a**). The chromosome 07-R subtelomere of *S. paradoxus* CBS432 was extended by a series of tandem duplications of *MAL31*-like and *MAL33*-like genes, as well as the addition of the *ARR* cluster (**Fig. 4c** and **Supplementary Fig. 6b**). The chromosome 15-L subtelomere of *S. paradoxus* UFRJ50816 increased size by duplications of subtelomeric segments from two other chromosomes (**Fig. 4d** and **Supplementary Fig. 6c**). Inversions have also occurred in subtelomeres, including one affecting the *HMRA1*–*HMRA2* locus in UFRJ50816 and another affecting a *MAL11*-like gene in CBS432 (**Supplementary Fig. 7**).

The enrichment of segmental duplication blocks occurring via ectopic sequence reshuffling is a common feature of eukaryotic subtelomeres; however, incomplete genome assemblies have prevented population-level quantitative analysis of this phenomenon. Here we identified subtelomeric duplication blocks based on pairwise comparisons of different subtelomeres within the same strain (**Fig. 5a** and **Supplementary Data Set 6**). In total, we identified 173 pairs of subtelomeric duplication blocks across the 12 strains, with 8–26 pairs for each strain (**Supplementary Table 15**). Among the 16 pairs of subtelomeric duplication blocks previously identified in S288C (mentioned above), all the 12 larger pairs passed our filtering criteria. Notably, the Hawaiian *S. paradoxus* UWOPS91-917.1 had the most subtelomeric duplication blocks, and half of these were strain-specific, suggesting unique subtelomere evolution in this strain. The duplicated segments always maintained the same centromere–telomere orientation, supporting a mutational mechanism of double-strand break (DSB) repair like those previously suggested in other species[40,41]. We further summarized those 173 pairs of duplication blocks according to the orthologous subtelomeres involved. This led to 75 unique duplicated subtelomere pairs, 59 (78.7%) of which have not been described before (**Supplementary Data Set 7**). We found 31 (41.3%) of these unique pairs to be shared between strains or even between species with highly dynamic strain-sharing patterns (**Fig. 5b** and **Supplementary Fig. 8a**). Most (87.1%) of this sharing pattern could not be explained by the strain phylogeny (**Supplementary Data Set 7**). This suggests a constant gain-and-loss process of subtelomeric duplications throughout evolutionary history.

Given the rampant subtelomere reshuffling, we investigated to what extent the similarity in orthologous subtelomere composition reflects the intra-species phylogenies. We measured the proportion of conserved orthologous subtelomeres in all strain pairs within the same species and performed hierarchical clustering accordingly (**Fig. 5c**). The clustering in *S. paradoxus* correctly recapitulated the true phylogeny, whereas the clustering in *S. cerevisiae* showed a different topology, and only the relationship of the most recently diversified strain pair (DBVPG6044 versus SK1) was correctly recovered. Notably, the distantly related Wine/European (DBVPG6765) and Sake (Y12) *S. cerevisiae* strains were clustered together, suggesting possible convergent subtelomere evolution during their respective domestication for alcoholic beverage production. The proportion of conserved orthologous subtelomeres among *S. cerevisiae* strains (56.3–81.3%) is comparable to that among *S. paradoxus* strains (50.0–81.3%), despite the much smaller diversification timescales of *S. cerevisiae*. This translates into a 3.8-fold difference in subtelomeric reshuffling intensity between the two species during their respective diversifications (one-sided Mann–Whitney *U*-test, $P = 2.93 \times 10^{-8}$) (**Fig. 5d**). The frequent reshuffling of subtelomeric sequences often has drastic impacts on gene content, both qualitatively and quantitatively. For example, four genes (*PAU3*, *ADH7*, *RDS1* and *AAD3*) were lost in *S. cerevisiae* Y12 owing to a single subtelomeric duplication event (chromosome 08-L to chromosome 03-R) (**Supplementary Fig. 8b**). Therefore, the accelerated subtelomere reshuffling in *S. cerevisiae* is likely to have important functional implications.

Native noncanonical chromosome end structures

S. cerevisiae chromosome ends are characterized by two telomere-associated sequences: the core X and the Y′ element[42]. The core X element is present in nearly all chromosome ends, whereas the number of Y′ elements varies across chromosome ends and strains. The two previously described chromosome end structures have (i) a single core X element or (ii) a single core X element followed by 1–4 distal Y′ elements[42]. *S. paradoxus* chromosome ends also contain core X and Y′ elements[43], but their detailed structures and genome-wide distributions have not been systematically characterized. Across our 12 strains, most (~85%) chromosome ends had one of the two structures described above, but we also discovered novel chromosome ends (**Supplementary Table 13**). We found several examples of tandem duplications of the core X element in both species. In most cases, including the ones in the *S. cerevisiae* reference genome (chromosome VIII-L and chromosome XVI-R), the proximal duplicated core X elements had degenerated, but we found two examples where intact duplicated copies were retained: chromosome XII-R in *S. cerevisiae* Y12 and chromosome III-L in *S. paradoxus* CBS432. The latter was especially notable, with six core X elements (including three complete copies) arranged in tandem. We discovered five chromosome ends consisting of only Y′ elements (one or more copies) but no core X elements. This was unexpected given the importance of core X elements in maintaining genome stability[44,45]. The discovery of these noncanonical chromosome end structures offers a new paradigm to investigate the functional role of core X elements.

Mitochondrial genome evolution

Despite being highly repetitive and AT-rich, the mitochondrial genomes of the *S. cerevisiae* strains showed high degrees of collinearity (**Fig. 6a**). In contrast, *S. paradoxus* mitochondrial genomes showed lineage-specific structural rearrangements. The two Eurasian strains (CBS432 and N44) share a transposition of the entire *COX3*–*RPM1* (rnpB)–15s rRNA (rns) segment, in which 15s rRNA was further inverted (**Fig. 6b–d**). In addition, given the gene order in two outgroups, the *COB* gene was relocated in the common ancestor of *S. cerevisiae* and *S. paradoxus* (**Fig. 6e**). The phylogenetic tree inferred from mitochondrial protein-coding genes showed clear deviation from the nuclear tree (**Fig. 6e**). In particular, the Eurasian *S. paradoxus* lineage (CBS432 and N44) clustered with the seven *S. cerevisiae* strains before joining with the other *S. paradoxus* strains, which supports the idea of mitochondrial introgression from *S. cerevisiae*[46] (**Fig. 6e**). We found low topology consensus (normalized quartet score = 0.59, versus 0.92 for the nuclear gene tree) across different mitochondrial gene loci, suggesting heterogeneous phylogenetic histories. Together with the drastically dynamic presence and absence patterns of mitochondrial group I and group II introns (**Supplementary Table 10**), this reinforces the argument for extensive cross-strain recombination in yeast mitochondrial evolution[47]. In addition, the *COX3* gene in *S. paradoxus* UFRJ50816 and UWOPS91-917.1 started with GTG rather than the typical ATG start codon, which was further supported by Illumina reads. This suggests either an adoption of an alternative ATG start codon nearby (for example, 45 bp downstream) or a rare case of a near-cognate start codon[48–50].

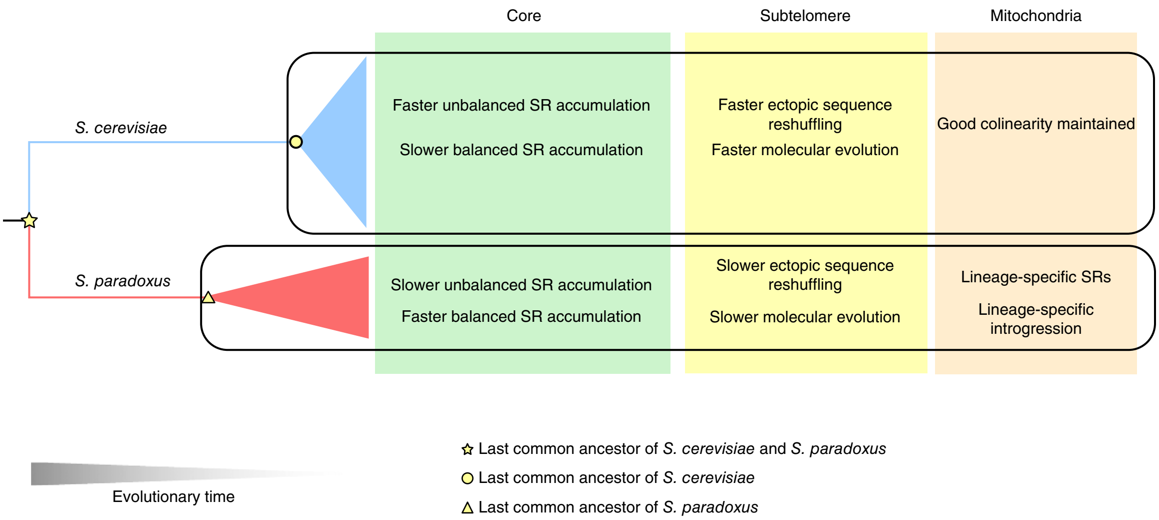
Core
Subtelomere
Mitochondria
S. cerevisiae
Faster unbalanced SR accumulation
Slower balanced SR accumulation
Faster ectopic sequence reshuffling
Faster molecular evolution
Good colinearity maintained
S. paradoxus
Slower unbalanced SR accumulation
Faster balanced SR accumulation
Slower ectopic sequence reshuffling
Slower molecular evolution
Lineage-specific SRs
Lineage-specific introgression
Evolutionary time
☆ Last common ancestor of S. cerevisiae and S. paradoxus
○ Last common ancestor of S. cerevisiae
△ Last common ancestor of S. paradoxus

Figure 8 Contrasting evolutionary dynamics across the genomic landscape between *S. cerevisiae* and *S. paradoxus*. The interspecific contrasts in nuclear chromosomal cores, subtelomeres and mitochondrial genomes are summarized. SR, structural rearrangement.

Fully resolved SVs illuminate complex phenotypic traits

SVs are expected to account for a substantial fraction of phenotypic variation; fully resolved SVs can therefore be crucial in understanding complex phenotypic traits. We used the copper tolerance–related *CUP1* locus and the arsenic tolerance–related *ARR* cluster as examples of associations between fully characterized genomic compositions (i.e., copy numbers and genotypes) and conditional growth rates. The PacBio assemblies precisely resolved these complex loci, and phenotype associations were consistent with previous findings based on copy number analysis[6,21,51] (**Fig. 7a**–**d** and **Supplementary Note**). We further illustrated their phenotypic contributions via linkage mapping using 826 phased outbred lines (POLs) derived from crossing the North American (YPS128) and West African (DBVPG6044) *S. cerevisiae*[52] (Online Methods). The linkage analysis accurately mapped a large-effect quantitative trait locus (QTL) at the chromosome 03-R subtelomere (the location of the *ARR* genes in DBVPG6044), but showed no arsenic resistance association with the YPS128 *ARR* locus on the chromosome 16-R subtelomere (**Fig. 7e**). This profile is consistent with the relocation of an active *ARR* cluster to the chromosome 03-R subtelomere in DBVPG6044 and the presence of deleterious mutations predicted to inactivate the *ARR* cluster in YPS128 (refs. 6,35). Thus, a full understanding of the relationship between genome sequence and arsenic resistance phenotype is not provided by the knowledge of copy number alone but rather requires the combined knowledge of genotype, genomic location and copy number as provided by our end-to-end assemblies (**Fig. 7f**).

DISCUSSION

The landscape of genetic variation is shaped by multiple evolutionary processes, including mutation, drift, recombination, gene flow, natural selection and demographic history. The combined effect of these factors can vary considerably both across the genome and between species, resulting in different patterns of evolutionary dynamics. The complete genome assemblies that we generated for multiple strains from both domesticated and wild yeasts provide a unique data set for exploring such patterns with unprecedented resolution.

Considering the evolutionary dynamics across the genome, eukaryotic subtelomeres are exceptionally variable compared to chromosomal cores[40,53,54], with accelerated evolution manifest in extensive CNV accumulation, rampant ectopic reshuffling and rapid functional divergence[6,41,55–57]. Our study provides a quantitative comparison of subtelomeres and cores in structural genome evolution and a high-resolution view of the extreme evolutionary plasticity of subtelomeres. This rapid evolution of subtelomeres can substantially alter the gene repertoire and generate novel recombinants with adaptive potential[57]. Given that subtelomeric genes are highly enriched in functions mediating interactions with external environments (for example, stress response, nutrient uptake and ion transport)[6,55,58], it is tempting to speculate that the accelerated subtelomeric evolution reflects selection for evolvability, i.e., the ability to respond and adapt to changing environments[59].

With regard to the genome dynamics between species, external factors such as selection and demographic history have important roles. The ecological niches and recent evolutionary history of *S. cerevisiae* have been intimately associated with human activities, with many strains isolated from human-associated environments such as breweries, bakeries and even clinical patients[60]. Consequently, this wide spectrum of selection schemes could significantly shape the genome evolution of *S. cerevisiae*. In addition, human activities also promoted admixture and cross-breeding of *S. cerevisiae* strains from different geographical locations and ecological niches[61], resulting in many mosaic strains with mixed genetic backgrounds[1]. In contrast, the wild-living *S. paradoxus* occupies very limited ecological niches, with most strains isolated from trees in the *Quercus* genus[62]. *S. paradoxus* strains from different geographical subpopulations are genetically well differentiated with partial reproductive isolations[34,63]. Such interspecific differences in their history could result in distinct evolutionary genome dynamics, which is captured in our study (**Fig. 8**). In chromosomal cores, *S. cerevisiae* strains show slower accumulation of balanced structural rearrangements compared with *S. paradoxus* strains. This pattern might be explained by the admixture between different *S. cerevisiae* subpopulations during their recent association with human activities, which would considerably impede the fixation of balanced structural rearrangements. In contrast, geographical isolation of different *S. paradoxus* subpopulations would favor relatively fast fixation of balanced structural rearrangements[64]. We observed an opposite pattern for unbalanced rearrangements in chromosomal cores. The *S. cerevisiae* strains accumulate such changes more rapidly than their *S. paradoxus* counterparts, which is probably driven by selection, considering the biological functions of those affected genes. Likewise, the more rapid subtelomeric reshuffling and higher dN/dS of subtelomeric genes in *S. cerevisiae* than in *S. paradoxus* are probably also driven by selection. As a consequence of such unbalanced rearrangements and subtelomeric reshuffling, *S. cerevisiae* strains show more rapid CNV accumulation and GOL, which reinforces this argument. In addition, the mitochondrial genomes of *S. cerevisiae* strains maintained high degrees of collinearity, whereas those of *S. paradoxus* strains showed lineage-specific structural rearrangements and introgression, suggesting distinct modes of mitochondrial evolution. Taken together, many of these observed differences between *S. cerevisiae* and *S. paradoxus* probably reflect the influence of human activities on structural genome evolution, which sheds new light on why *S. cerevisiae*, but not its wild relative, is one of our most biotechnologically important organisms.

URLs. Previously identified subtelomeric duplication blocks in *S. cerevisiae* S288C, http://www2.le.ac.uk/colleges/medbiopsych/research/gact/images/clusters-fixed-large.jpg; RepeatMasker, http://www.repeatmasker.org; FastQC, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/; Picard tools, http://broadinstitute.github.io/picard/; vcflib, https://github.com/vcflib/vcflib; MFannot, http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl; The *Saccharomyces* Genome Database (SGD), http://www.yeastgenome.org; FigTree, http://tree.bio.ed.ac.uk/software/figtree/.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank G. Drillon for the help with using the program CHROnicle. We thank O. Croce and R. Marangoni for help with maintaining the computing server and various bioinformatics tools. We thank A. Llored for technical help with yeast strains and DNA samples. This work was supported by ATIP-Avenir (CNRS/INSERM), Fondation ARC pour la Recherche sur le Cancer (PJA20151203273), Marie Curie Career Integration Grants (322035), Agence Nationale de la Recherche (ANR-16-CE12-0019, ANR-13-BSV6-0006-01 and ANR-11-LABX-0028-01), Cancéropôle PACA (AAP émergence 2015) and a DuPont Young Professor Award to G.L.; the Wellcome Trust (WT098051 to R.D.) and Vetenskapsrådet (The Swedish Research Council, 325-2014-4605 to J.W.). J.-X.Y. is supported by a postdoctoral fellowship from Fondation ARC pour la Recherche sur le Cancer (PDF20150602803). J.L. is supported by a postdoctoral fellowship from Fondation ARC pour la Recherche sur le Cancer (PDF20140601375). J.H. is supported by the Labex SIGNALIFE program from Agence Nationale de la Recherche (ANR-11-LABX-0028-01).

AUTHOR CONTRIBUTIONS

J.-X.Y. conceived, designed and performed bioinformatics analysis and wrote the manuscript; J.L. prepared DNA samples for sequencing, performed the experiment on verifying structural rearrangements and contributed to the manuscript; L.A. performed PacBio sequencing and helped with diagnosing the assembly pipeline; J.H. performed experiments and data analysis for phenotyping and contributed to the manuscript; K.P. performed experiments and data analysis for phenotyping and contributed to the manuscript; K.O. performed the PacBio sequencing and ran the standard assembly pipeline; A.B. helped with discussion on data analysis and manuscript preparation; P.C. performed the PacBio sequencing for the pilot phase project; J.W. designed the phenotyping experiment and helped with data interpretation; M.C.L. helped with the analysis on measuring the sequence homology for subtelomeres; G.F. helped with study design, results discussion and manuscript writing; R.D. conceived and designed the study; G.L. conceived, designed and guided the study and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Liti, G. *et al.* Population genomics of domestic and wild yeasts. *Nature* **458**, 337–341 (2009).
2. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
3. Cao, J. *et al.* Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43**, 956–963 (2011).
4. Mackay, T.F.C. *et al.* The *Drosophila melanogaster* genetic reference panel. *Nature* **482**, 173–178 (2012).
5. Huang, W. *et al.* Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome Res.* **24**, 1193–1208 (2014).
6. Bergström, A. *et al.* A high-definition view of functional genetic variation from natural yeast genomes. *Mol. Biol. Evol.* **31**, 872–888 (2014).
7. Strope, P.K. *et al.* The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Res.* **25**, 762–774 (2015).
8. 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsisthaliana*. *Cell* **166**, 481–491 (2016).
9. Gallone, B. *et al.* Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. *Cell* **166**, 1397–1410 (2016).
10. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
11. Rieseberg, L.H. Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* **16**, 351–358 (2001).
12. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J.O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
13. Goodwin, S., McPherson, J.D. & McCombie, W.R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
14. Chaisson, M.J.P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
15. Gordon, D. *et al.* Long-read sequence assembly of the gorilla genome. *Science* **352**, aae0344 (2016).
16. Pryde, F.E., Gorham, H.C. & Louis, E.J. Chromosome ends: all the same under their caps. *Curr. Opin. Genet. Dev.* **7**, 822–828 (1997).
17. Mefford, H.C. & Trask, B.J. The complex structure and dynamic evolution of human subtelomeres. *Nat. Rev. Genet.* **3**, 91–102 (2002).
18. Eichler, E.E. & Sankoff, D. Structural dynamics of eukaryotic chromosome evolution. *Science* **301**, 793–797 (2003).
19. Dujon, B. Yeast evolutionary genomics. *Nat. Rev. Genet.* **11**, 512–524 (2010).
20. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546–567, 563–567 (1996).
21. Warringer, J. *et al.* Trait variation in yeast is defined by population history. *PLoS Genet.* **7**, e1002111 (2011).
22. Wheelan, S.J., Scheifele, L.Z., Martínez-Murillo, F., Irizarry, R.A. & Boeke, J.D. Transposon insertion site profiling chip (TIP-chip). *Proc. Natl. Acad. Sci. USA* **103**, 17632–17637 (2006).
23. Shibata, Y., Malhotra, A., Bekiranov, S. & Dutta, A. Yeast genome analysis identifies chromosomal translocation, gene conversion events and several sites of Ty element insertion. *Nucleic Acids Res.* **37**, 6454–6465 (2009).
24. Hoang, M.L. *et al.* Competitive repair by naturally dispersed repetitive DNA during non-allelic homologous recombination. *PLoS Genet.* **6**, e1001228 (2010).
25. Liti, G., Peruffo, A., James, S.A., Roberts, I.N. & Louis, E.J. Inferences of evolutionary relationships from a population survey of LTR-retrotransposons and telomeric-associated sequences in the *Saccharomycessensu stricto* complex. *Yeast* **22**, 177–192 (2005).
26. Marie-Nelly, H. *et al.* High-quality genome (re)assembly using chromosomal contact data. *Nat. Commun.* **5**, 5695 (2014).
27. Tajima, F. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135**, 599–607 (1993).
28. Winzeler, E.A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
29. Rocha, E.P.C. DNA repeats lead to the accelerated loss of gene order in bacteria. *Trends Genet.* **19**, 600–603 (2003).
30. Rocha, E.P.C. Inference and analysis of the relative stability of bacterial chromosomes. *Mol. Biol. Evol.* **23**, 513–522 (2006).
31. Fischer, G., Rocha, E.P.C., Brunet, F., Vergassola, M. & Dujon, B. Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. *PLoS Genet.* **2**, e32 (2006).

32. Stephens, P.J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
33. Zhang, C.-Z., Leibowitz, M.L. & Pellman, D. Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. *Genes Dev.* **27**, 2513–2530 (2013).
34. Liti, G., Barton, D.B.H. & Louis, E.J. Sequence diversity, reproductive isolation and species concepts in *Saccharomyces*. *Genetics* **174**, 839–850 (2006).
35. Cubillos, F.A. *et al.* Assessing the complex architecture of polygenic traits in diverged yeast populations. *Mol. Ecol.* **20**, 1401–1413 (2011).
36. Fischer, G., James, S.A., Roberts, I.N., Oliver, S.G. & Louis, E.J. Chromosomal evolution in *Saccharomyces*. *Nature* **405**, 451–454 (2000).
37. 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).
38. Vakirlis, N. *et al.* Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. *Genome Res.* **26**, 918–932 (2016).
39. Marsit, S. *et al.* Evolutionary advantage conferred by an eukaryote-to-eukaryote gene transfer event in wine yeasts. *Mol. Biol. Evol.* **32**, 1695–1707 (2015).
40. Linardopoulou, E.V. *et al.* Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* **437**, 94–100 (2005).
41. Fairhead, C. & Dujon, B. Structure of *Kluyveromyces lactis* subtelomeres: duplications and gene content. *FEMS Yeast Res.* **6**, 428–441 (2006).
42. Louis, E.J. The chromosome ends of *Saccharomyces cerevisiae*. *Yeast* **11**, 1553–1573 (1995).
43. Liti, G. *et al.* Segregating *YKU80* and *TLC1* alleles underlying natural variation in telomere properties in wild yeast. *PLoS Genet.* **5**, e1000659 (2009).
44. Marvin, M.E. *et al.* The association of yKu with subtelomeric core X sequences prevents recombination involving telomeric sequences. *Genetics* **183**, 453–467 (2009).
45. Marvin, M.E., Griffin, C.D., Eyre, D.E., Barton, D.B.H. & Louis, E.J. In *Saccharomyces cerevisiae*, yKu and subtelomeric core X sequences repress homologous recombination near telomeres as part of the same pathway. *Genetics* **183**, 441–451 (2009).
46. Wu, B. & Hao, W. A dynamic mobile DNA family in the yeast mitochondrial genome. *G3 (Bethesda)* **5**, 1273–1282 (2015).
47. Wu, B., Buljic, A. & Hao, W. Extensive horizontal transfer and homologous recombination generate highly chimeric mitochondrial genomes in yeast. *Mol. Biol. Evol.* **32**, 2559–2570 (2015).
48. Blattner, F.R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462 (1997).
49. Cole, S.T. *et al.* Deciphering the biology of *Mycobacteriumtuberculosis* from the complete genome sequence. *Nature* **393**, 537–544 (1998).
50. Abramczyk, D., Tchórzewski, M. & Grankowski, N. Non-AUG translation initiation of mRNA encoding acidic ribosomal P2A protein in *Candida albicans*. *Yeast* **20**, 1045–1052 (2003).
51. Zhao, Y. *et al.* Structures of naturally evolved *CUP1* tandem arrays in yeast indicate that these arrays are generated by unequal nonhomologous recombination. *G3 (Bethesda)* **4**, 2259–2269 (2014).
52. Hallin, J. *et al.* Powerful decomposition of complex traits in a diploid model. *Nat. Commun.* **7**, 13311 (2016).
53. Anderson, J.A., Song, Y.S. & Langley, C.H. Molecular population genetics of *Drosophila* subtelomeric DNA. *Genetics* **178**, 477–487 (2008).
54. Kuo, H.-F., Olsen, K.M. & Richards, E.J. Natural variation in a subtelomeric region of *Arabidopsis*: implications for the genomic dynamics of a chromosome end. *Genetics* **173**, 401–417 (2006).
55. Brown, C.A., Murray, A.W. & Verstrepen, K.J. Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Curr. Biol.* **20**, 895–903 (2010).
56. Louis, E.J. & Haber, J.E. Mitotic recombination among subtelomeric Y′ repeats in *Saccharomyces cerevisiae*. *Genetics* **124**, 547–559 (1990).
57. Anderson, M.Z., Wigen, L.J., Burrack, L.S. & Berman, J. Real-time evolution of a subtelomeric gene family in *Candida albicans*. *Genetics* **200**, 907–919 (2015).
58. Ames, R.M. *et al.* Gene duplication and environmental adaptation within yeast populations. *Genome Biol. Evol.* **2**, 591–601 (2010).
59. Kirschner, M. & Gerhart, J. Evolvability. *Proc. Natl. Acad. Sci. USA* **95**, 8420–8427 (1998).
60. Liti, G. The fascinating and secret wild life of the budding yeast *S. cerevisiae*. *eLife* **4**, 1–9 (2015).
61. Hyma, K.E. & Fay, J.C. Mixing of vineyard and oak-tree ecotypes of *Saccharomyces cerevisiae* in North American vineyards. *Mol. Ecol.* **22**, 2917–2930 (2013).
62. Borneman, A.R. & Pretorius, I.S. Genomic insights into the *Saccharomycessensu stricto* complex. *Genetics* **199**, 281–291 (2015).
63. Sniegowski, P.D., Dombrowski, P.G. & Fingerman, E. *Saccharomyces cerevisiae* and *Saccharomycesparadoxus* coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. *FEMS Yeast Res.* **1**, 299–306 (2002).
64. Leducq, J.-B. *et al.* Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nat. Microbiol.* **1**, 15003 (2016).

ONLINE METHODS

Strain sampling, preparation and DNA extraction. On the basis of previous population genomics surveys[1], we sampled seven *S. cerevisiae* and five *S. paradoxus* strains (all in the haploid or homozygous diploid forms) to represent major evolutionary lineages of the two species (**Supplementary Table 1**). The reference strains for *S. cerevisiae* (S288C) and *S. paradoxus* (CBS432) were included for quality control. All strains were taken from our strain collection stored at −80 °C and cultured on yeast extract–peptone–dextrose (YPD) plates. A single colony for each strain was picked and cultured in 5 mL YPD liquid at 30 °C 220 r.p.m. overnight. DNA extraction was carried out using the MasterPure Yeast DNA Purification Kit (Epicentre).

PacBio sequencing and raw assembly. The sequencing center at the Wellcome Trust Sanger Institute performed library preparation and sequencing using the PacBio Single Molecule, Real-Time (SMRT) DNA sequencing technology (platform: PacBio RS II; chemistry: P4-C2 for the pilot phase and P6-C4 for the main phase). The raw reads were processed using the standard SMRT analysis pipeline (v2.3.0). The *de novo* assembly was carried out following the hierarchical genome-assembly process (HGAP) assembly protocol with Quiver polishing[65].

Assembly evaluation and manual refinement. We retrieved the reference genomes (**Supplementary Note**) for both species to assess the quality of our PacBio assemblies. For each polished PacBio assembly, we first used RepeatMasker (v4.0.5) (URLs) to soft-mask repetitive regions (option: -species fungi -xsmall -gff). The soft-masked assemblies were subsequently aligned to the reference genomes using the nucmer program from MUMmer (v3.23)[66] for chromosome assignment. For most chromosomes, we have single contigs covering the entire chromosomes. For the cases where internal assembly gaps occurred, we performed manual gap closing by consulting the assemblies generated in the pilot phase of this project. The only gap we were unable to close is the highly repetitive rDNA array (usually consisting 100–200 copies of a 9.1-kb unit) on chromosome XII. The *S. cerevisiae* reference genome used a 17,357-bp sequence of two tandemly arranged rDNA copies to represent this complex region. For our assemblies, we trimmed off the partially assembled rDNAs around this gap and re-linked the two contigs with 17,357-bp Ns to keep consistency. The mitochondrial genomes of the 12 strains were recovered by single contigs in the raw HGAP assemblies. We further circularized them and reset their starting position as the *ATP6* gene using Circlator (v1.1.4)[67]. The circularized mitochondrial genome assemblies were further checked by consulting the raw PacBio reads and manual adjustment was applied when necessary.

Illumina sequencing, read mapping and error correction. In addition to the PacBio sequencing, we also performed Illumina 151-bp paired-end sequencing for each strain at Institut Curie. We examined the raw Illumina reads via FastQC (v0.11.3) (URLs) and performed adaptor-removing and quality-based trimming by trimmomatic (v0.33)[68] (options: ILLUMINACLIP: adapters.fa:2:30:10 SLIDINGWINDOW:5:20 MINLEN:36). For each strain, the trimmed reads were mapped to the corresponding PacBio assemblies by BWA (v0.7.12)[69]. The resulting read alignments were subsequently processed by SAMTools (v1.2)[70], Picard tools (v1.131) (URLs) and GATK (v3.5-0)[71]. On the basis of Illumina read alignments, we further performed error correction with Pilon (v1.12)[72] to generate final assemblies for downstream analysis.

Base-level error rate estimation for the final PacBio assemblies. Eight of our twelve strains were previously sequenced using Illumina technology with moderate to high depths[6]. We retrieved those raw reads and mapped them to our PacBio assemblies (both before and after Pilon correction) following the protocol described above. SNPs and indels were called by FreeBayes (v1.0.1-2)[73] (option: -p 1) to assess the performance of the Pilon correction and estimate the remaining base-level error rate in our final assemblies. The raw SNP and indel calls were filtered by the vcffilter tool from vcflib (URLs) with the filter expression: QUAL > 30 & QUAL / AO > 10 & SAF > 0 & SAR > 0 & RPR > 1 & RPL > 1.

Assembly completeness evaluation. We compared our S288C PacBio assembly with three published *S. cerevisiae* assemblies generated by different sequencing technologies (PacBio, Oxford Nanopore and Illumina)[74,75]. We aligned these three assemblies as well as our S288C PacBio assembly to the *S. cerevisiae* reference genome using nucmer from MUMmer (v3.23)[66]. The nucmer alignments were filtered by delta-filter (from the same package) (option: -1). We converted the output file to BED format and used bedtools (v2.15.0)[76] to calculate the intersection between our genome alignment and various annotation features (such as chromosomes, genes, retrotransposable elements, telomeres) of the *S. cerevisiae* nuclear reference genome. The percentage coverage of these annotation features by different assemblies were summarized accordingly.

Annotation of the protein-coding genes, tRNA genes and other genomic features. For nuclear genomes, we assembled an integrative pipeline that combines three existing annotation tools to form an evidence-leveraged protein-coding gene annotation. First, we used the RATT package[77] for directly transferring the nondubious *S. cerevisiae* reference gene annotations to our PacBio assemblies on the basis of whole genome alignments. Furthermore, we used the Yeast Genome Annotation Pipeline (YGAP)[78] to annotate our PacBio assemblies (default options without scaffolds reordering) based on gene sequence homology and synteny conservation. A custom Perl script (available on request) was used to remove redundant, truncated, or frameshifted genes annotated by YGAP. Finally, we used the Maker pipeline (v2.31.8)[79] to perform *de novo* gene discovery with EST–protein alignment support (**Supplementary Note**). As a by-product, tRNA genes were also annotated via the tRNAscan-SE (v1.3.1)[80] module of the Maker pipeline. Gene annotations produced by RATT, YGAP and Maker together with the EST–protein alignment evidences generated by Maker were further leveraged by EVidenceModeler (EVM)[81] to form an integrative annotation. Manual curation was carried out for selected cases (for example, the *CUP1* and *ARR* clusters) and pseudogenes were manually labeled when verified. The same pipeline was used for upgrading the protein-coding gene annotation of *S. arboricolus*, for which the originally annotated coding sequences (CDSs) and protein sequences was used for initial EST–protein alignment. In addition, for the 12 strains, we systematically annotated other genomic features encoded in their nuclear genomes, such as centromeres, Ty retrotransposable elements and telomere-associated core X and Y′ elements (**Supplementary Note**). Protein-coding genes that overlap with truncated or full-length Tys, core X or Y′ elements were removed from our final annotation.

As for mitochondrial genomes, the protein-coding genes, tRNA genes and other mitochondrial RNA genes such as *RPM1* (RNase P RNA), 15S rRNA (small) and 21S rRNA (large) subunit rRNA were annotated by MFannot (URLs). The exon–intron boundaries of annotated mitochondrial genes were manually curated based on BLAST and the 12-way mitochondrial genome alignment generated by mVISTA[82].

Orthology group identification. For nuclear protein-coding genes, we used Proteinortho (v5.15)[83] to identify gene orthology across the 12 strains and six other *sensu stricto* outgroups: *Saccharomyces mikatae* (strain IFO1815), *Saccharomyces kudriavzevii* (strain IFO1802), *Saccharomyces kudriavzevii* (strain ZP591), *Saccharomyces arboricolus* (strain H6), *Saccharomyces eubayanus* (strain FM1318) and *Saccharomyces bayanus* var. *uvarum* (strain CBS7001). The orthology identification took into account both sequence homology and synteny conservation (the PoFF feature[84] of Proteinortho). For each annotated strain, the systematic names of nondubious genes in the *Saccharomyces* Genome Database (SGD) (URLs) were mapped to our annotated genes based on the orthology groups identified above.

Phylogenetic reconstruction. For nuclear genes, we performed the phylogenetic analysis on the basis of one-to-one orthologs that are shared across all 18 strains (seven *S. cerevisiae* + five *S. paradoxus* + six outgroups) using two complementary approaches: the concatenated tree approach and the consensus tree approach. For each one-to-one ortholog, we used MUSCLE (v3.8.1551)[85] to align protein sequences and PAL2NAL (v14)[86] to align codons accordingly. For the concatenated tree approach, we generated a concatenated codon alignment across all orthology groups and fed it into RAxML (v8.2.6)[87] for maximum likelihood (ML) tree building. Alignment partition was configured by the first, second, and third codon positions. The GTRGAMMA model was used for phylogenetic inference. The rapid bootstrapping method built in RAxML was used to assess the stability of internal nodes (option: -# 100). The final ML tree was visualized in FigTree (v1.4.2) (URLs). For the consensus tree approach, we built individual gene trees with RAxML using the same method described above, which were further summarized into a coalescent-based consensus species tree by ASTRAL (v4.7.12)[88]. The normalized quartet score was calculated to assess the reliability of the final species tree given individual gene trees. For mitochondrial genes, we performed the same phylogenetic analysis based on the eight mitochondrial protein-coding genes.

Relative rate test. To test the rate heterogeneity between *S. cerevisiae* and *S. paradoxus* in molecular evolution, we constructed three-way sequence alignments by sampling one strain for each species together with *S. mikatae* as the outgroup. The sequences were drawn from the concatenated nuclear CDS alignment described above. The extracted sequences were fed into MEGA (v7.0.16)[89] for Tajima's relative rate test[27]. We conducted this test for all possible *S. cerevisiae*–*S. paradoxus* strain pairs.

Molecular dating. As no yeast fossil record can be used for reliable calibration, we performed molecular dating analysis using a relative time scale. We used the phylogenetic tree constructed from the nuclear one-to-one orthologs as the input and performed least-squares-based fast dating with LSD[90] (options: -c -v -s). We specified *S. bayanus* var. *uvarum* CBS7001 and *S. eubayanus* FM1318 as outgroups for this analysis.

Conserved synteny block identification. We used SynChro from the CHROnicle package (January 2015 version)[91,92] to identify conserved synteny blocks. We prepared the input files for SynChro with custom Perl scripts (available on request) to provide the genomic coordinates of all annotated features together with the genome assembly and proteome sequences. SynChro subsequently performed exhaustive pairwise comparisons to identify synteny blocks shared in the given strain pair.

Subtelomere definition and chromosome partitioning. An often-used yeast subtelomere definition is 20–30 kb from the chromosome ends. However, this definition is arbitrary in the sense that it treats all subtelomeres indiscriminately. In this study, we defined yeast subtelomeres on the basis of gene synteny conservation profiles across the 12 strains. For each chromosome arm, we examined all syntenic blocks shared across the 12 strains and used the most distal one to define the distal boundary for the chromosomal core (**Supplementary Table 11**). Meanwhile, we defined the proximal boundary of the chromosome end for this chromosome arm according to the first occurrence of core X or Y′ elements. The region between these two boundaries was defined as the subtelomere for this chromosome arm, with 400–bp interstitial transition zones on both sides (**Supplementary Fig. 3**).

Given that some strains (i.e., UWOPS03-461.4, UFRJ50816 and UWOPS91-917.1) are involved in large-scale interchromosomal rearrangements, the current chromosomal identities (determined by centromeres) might not necessarily agree with the ancestral chromosomal identities (determined by gene contents). Therefore, we used Roman and Arabic numbers, respectively, to denote these two identities for all 12 strains and avoid potential confusion about those interchromosomal rearrangements (**Supplementary Fig. 4** and **Supplementary Table 12**). Each defined subtelomere was named according to the ancestral chromosomal identity of its flanking chromosomal core and denoted also using Arabic numbers (**Supplementary Data Sets 2** and **3**).

Identification of balanced and unbalanced structural rearrangements in chromosomal cores. To identify balanced rearrangements, we first used ReChro from CHROnicle (January 2015 version)[91,92]. We set the synteny block stringency parameter "delta=1" for the main analysis. A complementary run was performed with "delta=0" to identify single gene inversions. Alternatively, we started with the one-to-one ortholog gene pairs (identified by our orthology group identification) in chromosomal cores between any given strain pair and examined their relative orientation and chromosomal locations. If the two one-to-one orthologous genes are located on the same chromosome but have opposite orientations, an inversion should be involved. If they reside on different chromosomes, a translocation or transposition should be involved.

As for unbalanced rearrangements, we first generated whole-genome alignment for every strain pair by nucmer[66] (options: -maxmatch -c 500) and used Assemblytics[93] to identify potential insertions, deletions and duplications or contractions. All candidates were further intersected with our gene annotations by bedtools intersect[76] to only keep those encompassing at least one protein-coding gene. Alternatively, we started with all the genes enclosed in chromosomal cores of any given strain pair and filtered out those completely covered by unique genome alignment between this strain pair. All the remaining genes were classified as candidates potentially involved in unbalanced rearrangements.

All identified candidate cases were manually examined by dot plots using Gepard (v1.30)[94]. All verified rearrangements in chromosomal cores were further mapped to the phylogeny of the 12 strains to reconstruct their evolutionary histories based on the maximum parsimony principle. The corresponding genomic regions in those six outgroups were also checked by dot plots to provide further support for our evolutionary history inferences.

Gene Ontology analysis. The CDSs of the *S. cerevisiae* nondubious reference genes were BLASTed against the NCBI nonredundant (nr) database using blastx (E-value = 1×10^{-3}) and further annotated by BLAST2GO (v.3.2)[95,96] to generate Gene Ontology (GO) mapping for each gene. We performed Fisher's exact test[97] to detect significantly enriched GO terms of our test gene set relative to the genome-wide background. False discovery rate (FDR) (cutoff 0.05)[98] was used for multiple correction. Significantly enriched GO terms were further processed by the 'Reduce to most specific terms' function implemented in BLAST2GO to keep only child terms.

Molecular evolutionary rates, CNV accumulation and GOL estimation. For the one-to-one orthologs in each strain pair, we calculate synonymous substitution rate (dS), nonsynonymous substitution rate (dN) and nonsynonymous-to-synonymous substitution rate ratio (dN/dS) using the yn00 program from PAML (v4.8a)[99] based on Yang and Nielsen[100]. We also measured the proportion of genes involved in CNVs (i.e., those are not one-to-one orthologs) in any strain pair. We denoted this measurement as P_{CNVs}, a quantity analogous to the P-distance in sequence comparison. To correct for multiple changes at the same gene loci, the Poisson distance D_{CNVs} can be given by $-\ln(1 - P_{\mathrm{CNVs}})$. This value can be further adjusted with evolutionary time by dividing $2T$, where T is the diversification time of the two compared strains obtained from our molecular dating analysis. To further capture evolutionary dynamics in terms of gene order changes, we further measured GOL for those one-to-one orthologs using the method proposed by previous studies without allowing for intervening genes[29–31]. For GOL, we performed similar Poisson correction and evolutionary time adjustment as for CNV accumulation. The calculation values for dN/dS, CNV accumulation and GOL were further summarized by 'core genes' and 'subtelomeric genes' on the basis of genome partitioning described above.

Subtelomeric homology search. For each defined subtelomeric region, we hard-masked all the enclosed Ty-related features (i.e., full-length Ty, truncated Ty and Ty solo-LTRs) and then searched against all the other subtelomeric regions for shared sequence homology. The search was performed by BLAT[101] (options: -noHead -stepSize = 5 -repMatch = 2253 -minIdentity = 80 -*t* = dna -q = dna -mask = lower -qMask = lower). We used pslCDnaFilter (options: -minId = 0.9 -minAlnSize = 1000 -bestOverlap -filterWeirdOverlapped) to filter out trivial signals and pslScore to calculate sequence alignment scores for those filtered BLAT matches. As the two reciprocal scores obtained from the same subtelomere pair are not symmetrical (depending on which sequence was used as the query), we took their arithmetic mean in our analysis. Such subtelomeric homology search was carried out for both within-strain and cross-strain comparisons, and subtelomere pairs with strong sequence homology (BLAT alignment score ≥5000 and sequence identity ≥90%) were recorded.

Hierarchical clustering analysis and reshuffling rate calculation for orthologous subtelomeres. For all strains within the same species, we performed pairwise comparisons of their subtelomeric regions to identify conserved orthologous subtelomeres in any given strain pairs on the basis of homology search described above. For each strain pair, the proportion of conserved orthologous subtelomeres was calculated as a measurement of the overall subtelomere conservation between the two strains. Such measurements were converted into a distance matrix by the dist() function in R (v3.1)[102], based on which the hclust() function was further used for hierarchical clustering. We gauged the reshuffling intensity of orthologous subtelomeres similarly to how we measured CNV accumulation and GOL. For any given strain pair, we first calculated the proportion of the nonconserved orthologous subtelomeres in this strain pair as $P_{\text{reshuffling}}$ and then applied the Poisson correction and evolutionary time adjustment by $-\ln(1 - P_{\text{reshuffling}})/2T$, in which T is the diversification time of the two compared strains.

Phenotyping the growth rates of yeast strains in copper- and arsenite-rich medium. The homozygous diploid versions of the 12 strains were pre-cultured in synthetic complete (SC) medium overnight to saturation. To examine their conditional growth rates in copper- and arsenite-rich environment, we mixed 350 μl conditional medium ($CuCl_2$ (0.38 mM) and arsenite (As(III), 3 mM) for the two environment respectively) with 10 μl saturated culture to the wells of honeycomb plates. Oxygen-permeable films were placed on top of the plates to enable uniform oxygen distribution throughout the plate. The automatic screening was done with Bioscreen Analyser C (Thermo Labsystems Oy) at 30 °C for 72 h, measuring in 20-min intervals using a wide-band filter at 420–580 nm (ref. 103). Growth data pre-processing and phenotypic trait extraction were performed by PRECOG[104].

Linkage analysis in diploid *S. cerevisiae* hybrids. A total of 826 phased outbred lines (POLs) were constructed and phenotyped as previously described[52]. Briefly, advanced intercrossed lines (AILs) were generated by successive rounds of mating and sporulation from the YPS128 and DBVPG6044 strains[105]. The resulting haploid AILs were sequenced[106] and crossed in different combinations to yield the 826 POLs used for the analysis. The POL diploid genotypes can be accurately inferred from the haploid AILs. Effectively, these 826 POLs constitute a subset of the larger set of POLs in Hallin *et al.*[52] but were constructed and phenotyped independently. Phenotyping of the POLs, each with four replicates, was performed using Scan-o-Matic[107] on solid agar plates (0.14% yeast nitrogen base, 0.5% ammonium sulfate, 2% (w/v) glucose and pH buffered to 5.8 with 1% (w/v) succinic acid, 0.077% complete supplement mixture (CSM, Formedium), 2% agar) supplemented with varying arsenite concentrations (0, 1, 2, and 3 mM). Using the deviations between the POL phenotype and the estimated parental mean phenotype in the mapping to combat population structure issues[52], quantitative trait loci (QTLs) were mapped using the scanone() function in R/qtl[108] with the marker regression method.

Statistics. Tajima's relative rate test[27] was performed in MEGA (v7.0.16)[89]. Fisher's exact test[97] with FDR correction[98] was performed in BLAST2GO (v.3.2)[95,96]. The Mann–Whitney U-test was performed in R (v3.1)[102] using the wilcox.test() function, with one.sided alternative hypothesis. $P < 0.05$ was considered statistically significant in all statistical tests.

Data availability. All genome sequencing, assembly and annotation data that support the findings of this study have been deposited in public repositories. The PacBio sequencing reads for this project has been deposed in the European Nucleotide Archive (ENA) under accession code PRJEB7245. Illumina sequencing reads have been deposed in Short Reads Archive (SRA) under accession code PRJNA340312. The genome assemblies and annotations generated by this study are available at https://yjx1217.github.io/Yeast_PacBio_2016/data/ and in GenBank under accession code PRJEB7245.

65. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
66. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
67. Hunt, M. *et al.* Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.* **16**, 294 (2015).
68. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
69. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
70. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
71. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
72. Walker, B.J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
73. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. https://arxiv.org/abs/1207.3907 (2012).
74. Kim, K.E. *et al.* Long-read, whole-genome shotgun sequence data for five model organisms. *Sci. Data* **1**, 140045 (2014).
75. Goodwin, S. *et al.* Oxford Nanopore sequencing, hybrid error correction, and *denovo* assembly of a eukaryotic genome. *Genome Res.* **25**, 1750–1756 (2015).
76. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
77. Otto, T.D., Dillon, G.P., Degrave, W.S. & Berriman, M. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res.* **39**, e57 (2011).
78. Proux-Wéra, E., Armisén, D., Byrne, K.P. & Wolfe, K.H. A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach. *BMC Bioinformatics* **13**, 237 (2012).
79. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
80. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
81. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
82. Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* **32**, W273–W279 (2004).
83. Lechner, M. *et al.* Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* **12**, 124 (2011).
84. Lechner, M. *et al.* Orthology detection combining clustering and synteny for very large datasets. *PLoS One* **9**, e105015 (2014).
85. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
86. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).
87. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
88. Mirarab, S. *et al.* ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541–i548 (2014).
89. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
90. To, T.H., Jung, M., Lycett, S. & Gascuel, O. Fast dating using least-squares criteria and algorithms. *Syst. Biol.* **65**, 82–97 (2016).
91. Drillon, G., Carbone, A. & Fischer, G. Combinatorics of chromosomal rearrangements based on synteny blocks and synteny packs. *J. Log. Comput.* **23**, 815–838 (2013).
92. Drillon, G., Carbone, A. & Fischer, G. SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PLoS One* **9**, e92621 (2014).
93. Nattestad, M. & Schatz, M.C. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021–3023 (2016).
94. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).
95. Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
96. Götz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435 (2008).
97. Fisher, R. On the interpretation of χ^2 from contingency tables, and the calculation of *P*. *J.R. Stat. Soc.* **85**, 87–94 (1922).
98. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
99. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
100. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43 (2000).
101. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
102. R Developement Core Team. R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, 2015).
103. Warringer, J. & Blomberg, A. Automated screening in environmental arrays allows analysis of quantitative phenotypic profiles in *Saccharomyces cerevisiae*. *Yeast* **20**, 53–67 (2003).
104. Fernandez-Ricaud, L., Kourtchenko, O., Zackrisson, M., Warringer, J. & Blomberg, A. PRECOG: a tool for automated extraction and visualization of fitness components in microbial growth phenomics. *BMC Bioinformatics* **17**, 249 (2016).
105. Parts, L. *et al.* Revealing the genetic structure of a trait by sequencing a population under selection. *Genome Res.* **21**, 1131–1138 (2011).
106. Illingworth, C.J.R., Parts, L., Bergström, A., Liti, G. & Mustonen, V. Inferring genome-wide recombination landscapes from advanced intercross lines: application to yeast crosses. *PLoS One* **8**, e62266 (2013).
107. Zackrisson, M. *et al.* Scan-o-matic: high-resolution microbial phenomics at a massive scale. *G3 (Bethesda)* **6**, 3003–3014 (2016).
108. Broman, K.W., Wu, H., Sen, S. & Churchill, G.A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889–890 (2003).

doi:10.1038/ng.3847