Tracking the origin of Baker's yeast: Seven new long-read genomes of highly divergent, wild *Saccharomyces cerevisiae* strains from East Asia

Devin P Bendixsen*, Noah Gettle, Ciaran Gilchrist, Zebin Zhang, and Rike Stelkens

Division of Population Genetics, Department of Zoology, Stockholm University, Svante Arrheniusväg 18 B, 106 91 Stockholm, Sweden

* Corresponding author: Devin P Bendixsen, Department of Zoology Stockholm University;
email: devin.bendixsen@zoologi.su.se

Running title Tracking the origin of Baker's yeast

Keywords *Saccharomyces cerevisiae, yeast, long-read, genome assembly, structural variation*

Abstract

Comparative genome analyses have suggested East Asia to be the cradle of the domesticated microbe Baker's yeast (*Saccharomyces cerevisiae*), used in the food and biotechnology industry worldwide. Here, we provide seven new, high quality long read genomes of non-domesticated yeast strains isolated from primeval forests and other natural environments in China and Taiwan. In a comprehensive analysis of our new genome assemblies, along with all long read *Saccharomycetes* genomes available to date, we show that the newly sequenced East Asian strains are among the closest living relatives of the ancestors of the global diversity of Baker's yeast, confirming predictions made from short read genomic data. Three of these strains (termed the East Asian Clade IX Complex here) share a recent ancestry and evolutionary history suggesting an early divergence from other *S. cerevisiae* strains prior to the larger radiation of the species. Our genomic analyses reveal that the wild East Asian strains contain elevated levels of structural variations (balanced and unbalanced), including large inversions and non-reciprocal chromosomal translocations. Two strains have severely reduced genomes, due to the loss of exon content rather than coding genes. The new genomic resources provided here contribute to our understanding of the natural diversity of *S. cerevisiae*, expand the intraspecific genetic variation found in this this heavily domesticated microbe, and provide a foundation for understanding its origin and global colonization history.

Introduction

The history of the yeast *S. cerevisiae* is deeply interwoven with that of human migration, agriculture, and food technology. The sequencing of hundreds of *Saccharomyces* genomes has provided insight about the domestication of the species, its population genetic structure, and eukaryotic genome evolution (Scannell et al. 2011; Wang et al. 2012; Duan et al. 2018; Peter et al. 2018). The vast majority of these genomes have been analyzed using short-read sequencing, resulting in a focus on single nucleotide variants (SNVs). However, larger structural variations (SVs), such as inversions, deletions and gene duplications, are difficult to resolve using only short-read data (Goodwin et al. 2016). SVs are vital for evolutionary adaptation in yeast, especially to stressful conditions (Yona et al. 2012; Zheng et al. 2014; Jeffares et al. 2017; Zhang et al. 2020), and an increasing number of studies in other taxa support the role of inversions in adaptation and speciation (Hoffmann and Rieseberg 2008; Hooper and Price 2017) and the evolution of disease (Feder et al. 2014; Merker et al. 2018; Sanchis-Juan et al. 2018; Wellenreuther et al. 2019). To be able to dependably identify and resolve SVs, we need high quality long-read genome assemblies, spanning complex or repetitive genomic regions with a single continuous read. This approach has recently been applied to 15 *Saccharomyces* strains (Hebly et al. 2015; Yue et al. 2017; Naseeb et al. 2018; Macías et al. 2019) which, together with the annotation of individual genes and proteins, offers valuable insights into the functional genomics of the natural yeast diversity.

Among the known *Saccharomyces cerevisiae* strains are several highly divergent wild strains originating in East Asia (Wang et al. 2012; Duan et al. 2018; Peter et al. 2018). These wild lineages (termed CHN I-IX and Taiwanese) represent the most divergent *S. cerevisiae* strains described to date, with up to 1.1% genome-wide divergence, and have been suggested to represent the geographic origin of the species. Here, we applied long-read sequencing to seven of the highly divergent East Asian strains and one common laboratory strain (Y55, **Table 1**) and generated eight new high-quality genomes. Using these new genomes, we generated new phylogenetic reconstructions that place these East Asian strains at the

base of *S. cerevisiae* and provide further evidence for an out-of-China colonization history of this species. Furthermore, we find notable intraspecific sequence divergence among these newly sequenced East Asian strains and vast amounts of chromosomal SVs, including small and large inversions, translocations, chromosome fusions and aneuploidies, shedding light on the genomic history and dynamics of *S. cerevisiae.*

Table 1: Descriptions of the *Saccharomyces cerevisiae* strains sequenced in this study.

Lineage	Strain	Species	Source	Geographic Location
CHN I	HN1	*S. cerevisiae*	rotten wood, primeval forest	Diaoluo Mountain, Hainan, China
CHN II	SX2	*S. cerevisiae*	bark of a Fagaceae tree, primeval forest	Qinling Mountain, Shaanxi, China
CHN IV	HLJ1	*S. cerevisiae*	bark of Quercus mongokica, secondary forest	Jingbo lake, Heilongjiang, China
CHN VI	BJ4	*S. cerevisiae*	intestine of a butterfly, park	Haidian, Beijing, China
CHN IX	JXXY16.1	*S. cerevisiae*	bark, primeval forest	Xiangxiyuan, Hubei Province, China
CHN IX	XXYS1.4	*S. cerevisiae*	bark, primeval forest	Xiangxiyuan, Hubei Province, China
Taiwanese	EN14S01-3B	*S. cerevisiae*	soil	Sinyi, Nantou, Taiwan
West African+	Y55	*S. cerevisiae*	wine grapes	France

Results

Genome sequencing and assembly

We applied long-read PacBio sequencing to seven highly divergent and one common lab strain of *S. cerevisiae*, with an average per base genomic coverage of 235 (median 203). The nuclear and mitochondrial assemblies were highly complete with all chromosomes assembled into single contigs. Assessment of the completeness of the genome assembly and annotation using BUSCO found that six of our eight genomes had similarly high BUSCO scores (C > 97.2%). The remaining two genomes (BJ4 and HLJ1) had marginally lower scores (84.0% and 89.8%, respectively). Correspondingly, we found that BJ4 and HLJ1 had reduced genome sizes (11.42 and 11.37 Mbp compared to on average 11.93 Mbp in the other genomes sequenced here).

Phylogenomics

Our newly constructed consensus species tree, based on 5847 genes, placed six of the newly assembled East Asian strains in a basal position within the *S. cerevisiae* radiation (**Fig. 1**). Interestingly, three of these strains (EN14S01-3B, XXYS1.4 and JXXY16.1), which were previously reported as being the most divergent (CHN IX and Taiwanese clades), show early divergence from all other *S. cerevisiae* strains. The common lab strain, Y55, clustered with two other domesticated strains (DBVPG6044 and SK1) within the West African+ clade. Construction of an Alignment and Assembly-Free (AAF) phylogeny shows similar clustering among the Far East strains and the common lab strain (**Fig. S1**) as well as a large amount of divergence of the CHNIX and Taiwanese strains (hereon referred to as the Clade IX Complex) from the rest of *S. cerevisiae.* Relative to the constructed consensus phylogeny, the AAF analysis places the seventh East Asian strain (BJ4) closer to the other East Asian strains. AAF analysis comparing the long-read sequencing data generated in this study and previous short-read data found a high-level of similarity (**Fig. S2**).

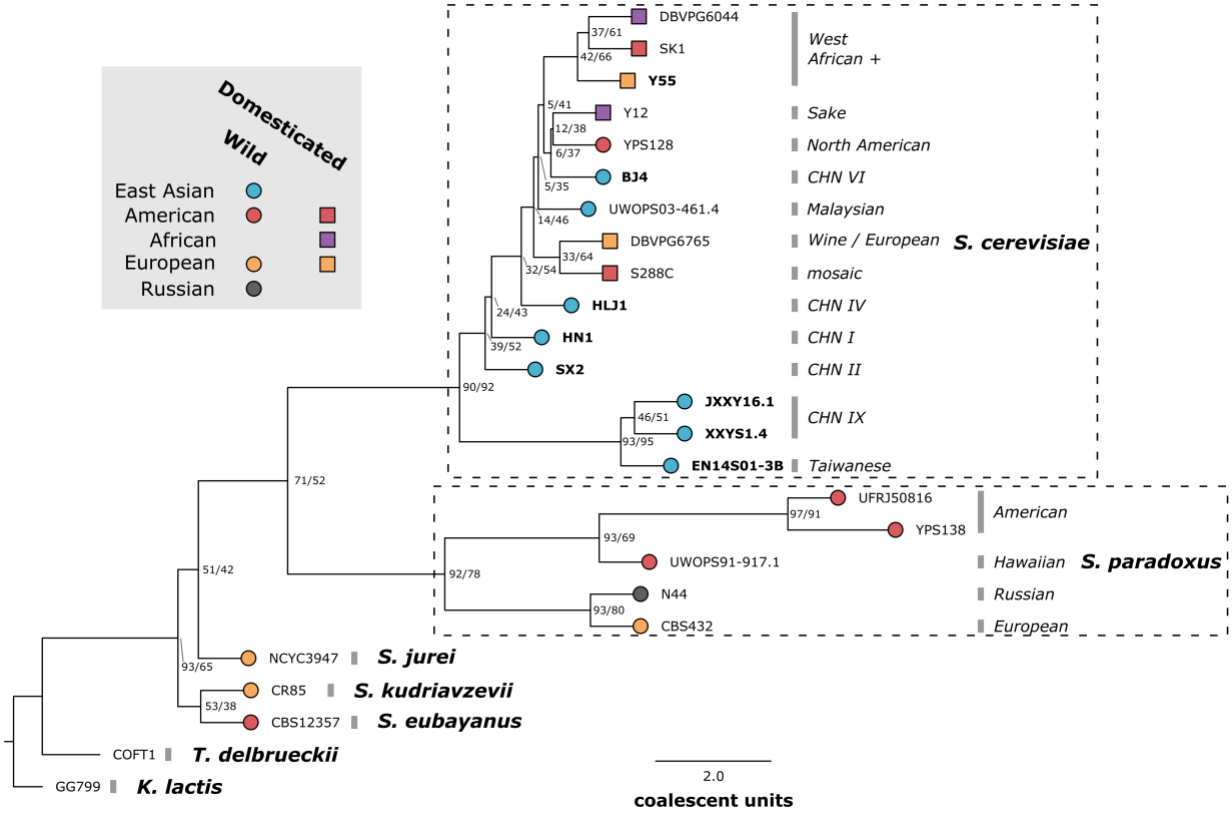

Figure 1: Maximum-likelihood phylogenetic tree of yeast long-read genomes. The tree was built from the alignment of all orthologous genes found in at least 5 strains (5847 genes). A maximum-likelihood single-tree inference was made for each locus and the consensus species tree is reported. Internal branches are labeled with gene concordance factor (gCF)/site concordance factor (sCF). gCF indicates the percent of gene trees supporting a given branch. sCF indicates the percentage of total aligned sites supporting a given branch. All internal branches were also supported with quartet scores equal to 1. For species with more than a single long-read genome assembly (*S. cerevisiae and S. paradoxus*), species clades are indicated in italics. *Saccharomyces* strains are colored according to their location of origin and branch tip shape indicates whether it is a domesticated (square) or wild (circle) strain. New long-read genome assemblies presented in this study are indicated in bold.

Comparative nuclear genomics

A comparison of our eight assembled *S. cerevisiae* genomes to each other and to previously published long-read genomes reveal a high-level of collinearity with the *S. cerevisiae* reference genome (S288C), with some notable deviations (**Fig. 2**, **Fig. S3-S4**). In particular, we found several large rearrangements within the Clade IX Complex. In all three member strains, we found a large terminal translocation from chromosome XI to chromosome XII (**Fig. 2 inset**). This ~80kb translocation is found conserved in these three strains but is not found in any other *Saccharomyces* strain. Another ~75kb inversion, was found in chromosome IV of the Taiwanese isolate EN14S01-3B but not in any other strain. We also found an apparent chromosomal fusion between chromosome VIII and chromosome XIII in strain JXXY16.1. We did not, however, find a loss or clear disruption of centromeric region from either chromosome. Other notable rearrangements are a large inversion in chromosome X of BJ4, three large translocations in chromosomes V, VII and X of HN1, and a large translocation and inversion in chromosome X of SX2 (**Fig. S3-S4**). The common lab strain, Y55, showed a high level of collinearity, with only minor deviations from homology.

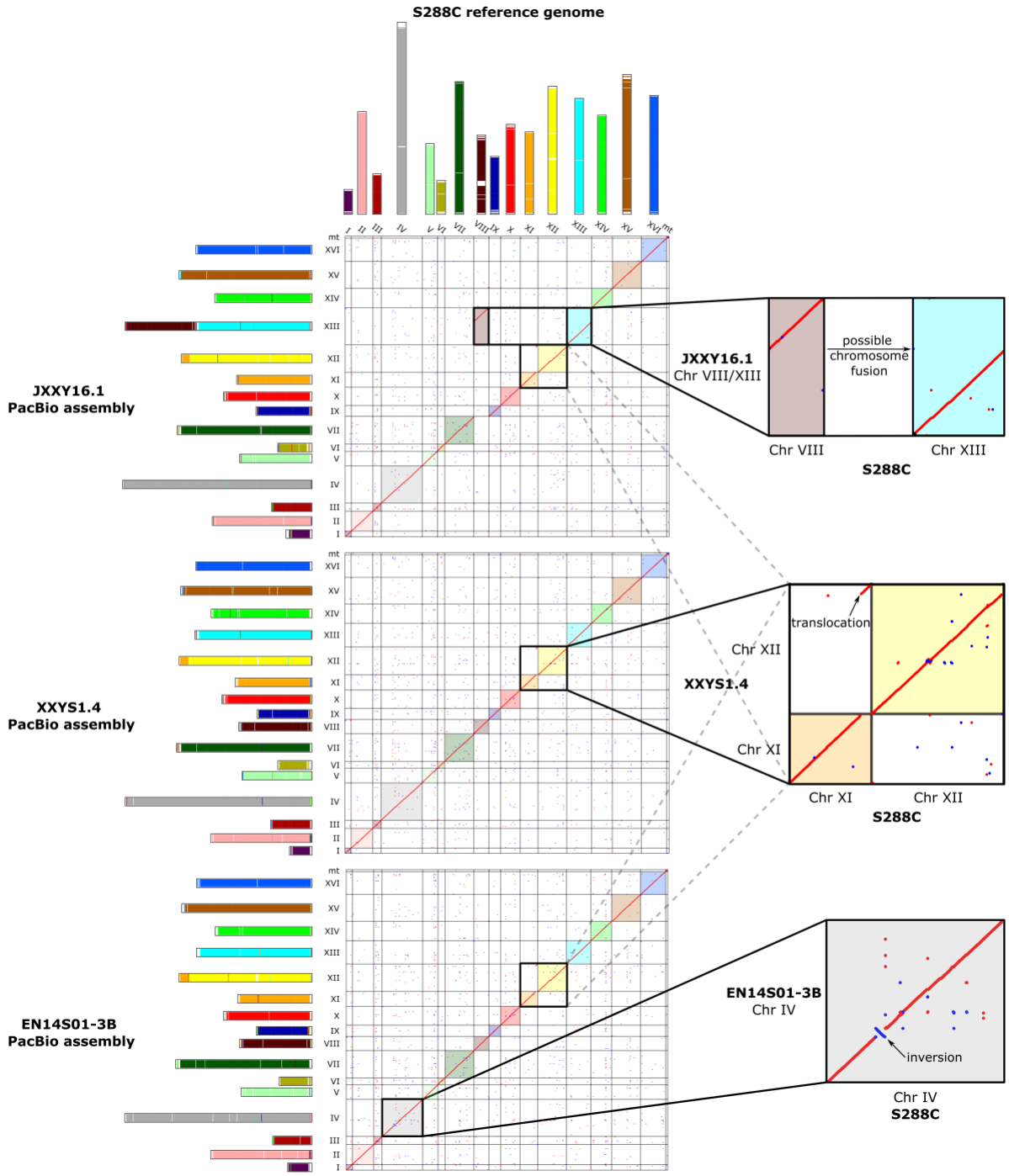

Figure 2: *Saccharomyces cerevisiae* long-read PacBio genome assemblies Genome comparison of the reference strain, S288C, and three members of the East Asian Clade IX Complex: JXXY16.1, XXYS1.4 and EN14S01-3B. Sequence homology within the dot plots is indicated by red dots for forward matches and blue dots for reverse matches. Chromosomal paintings and chromosome intersections between genome assemblies are colored according to the S288C reference genome. Insets depict examples of deviations from homology: 1) a possible chromosome fusion between chromosome VIII and XIII in JXXY16.1, 2) a large translocation between XI and chromosome XII found conserved in JXXY16.1, XXYS1.4 and 3) EN14SO1-3B and a large inversion in chromosome IV of EN14S01-3B.

Nuclear genome content

With regards to genome content, our newly sequenced East Asian genomes largely resemble those of Y55 and other assembled *S. cerevisiae* with a slightly lower range in gene counts (6095 – 6257 compared to 6289 – 6573) and 1.08 exons per gene. On average, 95.2% of these are directly orthologous to genes identified in the reference strain, S288C. Similarly, numbers of ncRNA and rRNA regions resemble those in other strains. To determine the source of the reduced genome sizes in strains BJ4 and HLJ1 we also examined genomic feature content in terms of overall length. While these strains do not show reduced numbers of genes, much of the differences in overall size are accounted for in loss of exon content with regard to their total length rather than reductions in unannotated genome content that is presumably non-coding (**Fig. S5-S7**). In addition to being outliers in exon content, BJ4 and HLJ1 also have considerably more intronic content than other *S. cerevisiae* strains (**Fig. 3A**). Composite lengths in base pairs of genomic features relative to overall feature mean for *S. cerevisiae* for total genomic length ($\bar{x}$ = 11.92 x 10^6; s = 2.26 x 10^5), exon regions ($\bar{x}$ = 8.46 x 10^6; s = 2.54 x 10^5), intron regions ($\bar{x}$ = 9.50 x 10^4; s = 1.47 x 10^4), and transposable elements (TEs) ($\bar{x}$ = 2.36 x 10^5; s = 9.68 x 10^4).

The amount of each genome composed of transposable elements (TEs) varied considerably among strains, but does not appear to follow a clear geographical or phylogenetic pattern nor is it correlated with genome size (**Fig. 3B**). In terms of TE make-up, all seven East Asian genomes, including the Clade IX Complex, have fewer functional TY1 elements than other *S. cerevisiae* strains with relative expansions in TY3 and TY5 elements, but clear patterns matching phylogenetic history are lacking.

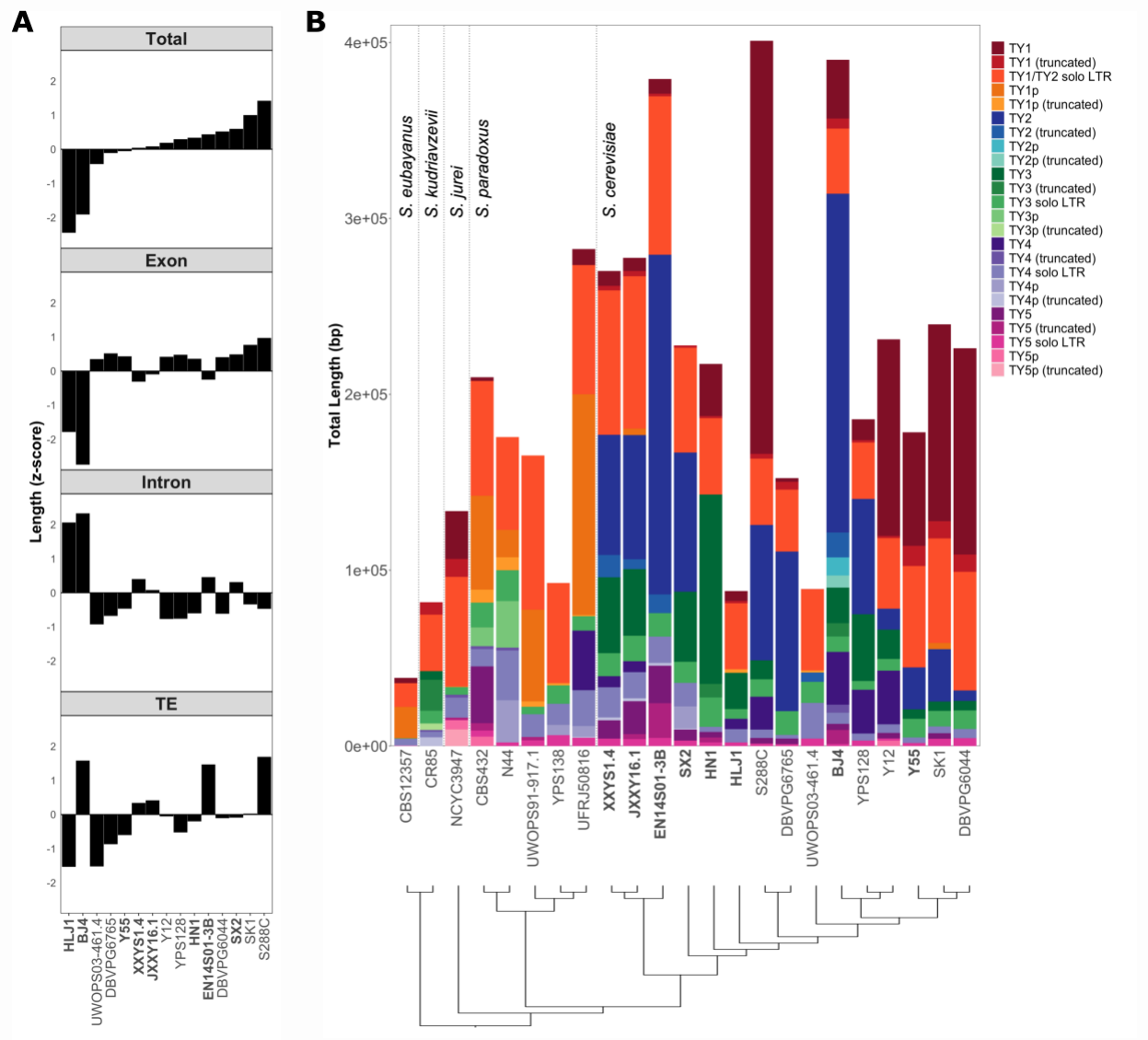

Figure 3: Genomic feature composition of *Saccharomyces* (A) Composite lengths in base pairs of genomic features relative to overall feature mean for *S. cerevisiae* for total genomic length, exon regions, intron regions, and transposable elements (TEs). Strains are ordered according to the total genomic length. **(B)** Transposable element composition in total length subdivided by TY classification for *Saccharomyces sensu stricto* strains. Phylogram is taken from species tree described in **Fig. 1**. solo LTR = long terminal repeat components of TY elements without replicative machinery.

Unbalanced and balanced structural variations

In addition to the large rearrangements mentioned above, which are readily identified in a collinearity analysis, we performed a comprehensive analysis to quantify the amount of structural variations in our genomes. We assessed both unbalanced and balanced variations using pairwise comparisons between

the 15 *S. cerevisiae* strains with long-read assemblies. Unbalanced variations are those that change the overall genomic content, such as insertions and deletions. Balanced variations are those that maintain genomic content such as inversions and translocations. Unbalanced variations were largely composed of repeat expansions/contractions (**Fig. S8**), whereas most of the balanced variations were translocations (**Fig. S9**). Pairwise comparisons of all *S. cerevisiae* strains revealed that the three Clade IX Complex strains (JXXY16.1, XXYS1.4, EN14S01-3B), showed significantly more balanced and unbalanced variation relative to each other and the rest of the analyzed *S. cerevisiae* strains (**Fig. 4**). On average these strains had ~38 unbalanced and ~36 balanced variations, whereas the average for all other comparisons was ~23 for each unbalanced and balanced. Interestingly, we also found large amounts of variation among the three divergent strains. However, the largest number of structural rearrangements, especially unbalanced variation, was found in the genomes of BJ4 and HLJ1, corresponding to their overall reduced genome size.

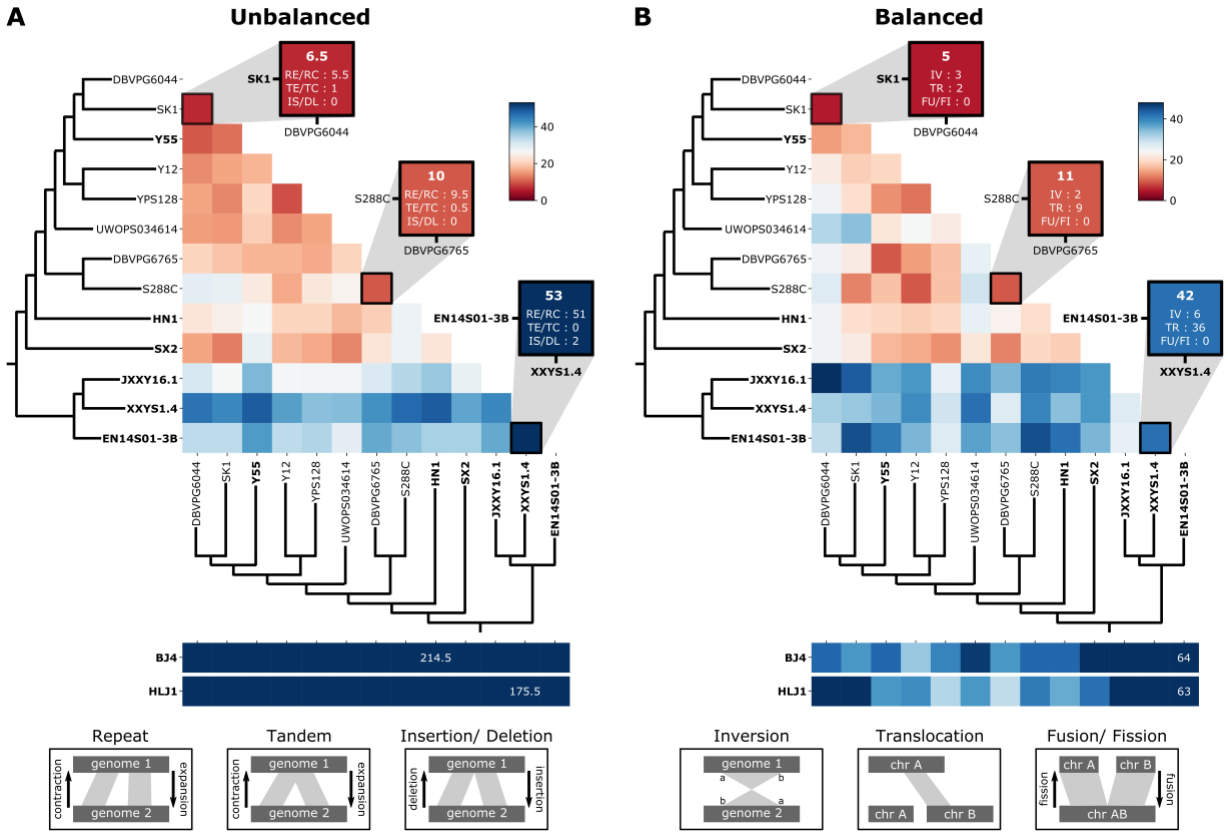

Figure 4: Structural variations within *Saccharomyces cerevisiae*. Pairwise comparisons among all *Saccharomyces cerevisiae* genome assemblies depicting the total **(A)** unbalanced and **(B)** balanced variations. Cladograms are adapted from the maximum-likelihood tree (**Fig. 1**). New long-read genome assemblies presented in this study are indicated in bold. For unbalanced variations, repeat expansions (RE) and contractions (RC), tandem expansions (TE) and contractions (TC), insertions (IS) and deletions (DL) were assessed. For balanced variations, inversions (IV), translocations (TR), fusions (FU) and fissions (FI) were assessed. Cartoons illustrate the different types of variation. Insets depict examples of pairwise comparisons. BJ4 and HLJ1 are plotted separately due to outlying counts of unbalanced SVs. The color scale is retained from the main heatmap with all values above the maximum of the main heatmap coloured dark blue. The maximum number of variants for either BJ4 or HLJ1 is shown for reference.

Comparative mitochondrial genomics

Overall, the mitochondrial genomes of the *S. cerevisiae* strains showed high-levels of collinearity (**Fig. S3-S4**). We did not, however, find a direct correspondence between our consensus gene phylogeny and that based on mitochondrial content (**Fig. 5**). Rather, we found a large division within *S. cerevisiae* strains with no clear geographical or evolutionary pattern. Five of the six *S. cerevisiae* strains that are domesticated were found on one side of the division. The nine wild strains, including the seven divergent East Asian strains along with a single domesticated strain (Y12) clustered on the other side of the division. Among the seven East Asian strains presented in this study, we found several interesting rearrangements including a duplication/inversion of COX1, ATP 6/8 in SX2, a duplication/inversion of 21S rRNA in JXXY16.1 and two duplications and an inversion of RPM1 in HN1. Most surprisingly, the 15S rRNA was not found in the mitochondrial assembly of BJ4 and a BLAST search of the raw sequencing data was not able to detect the gene. We also found that RPM1 was missing in XXYS1.4 and there was a severe reduction in size of VAR1.

Previous analyses have suggested that hybridization events can generate discordance between species and mitochondrial phylogenies (Peris et al. 2017; De Chiara et al. 2020). To investigate this, we also included other *Saccharomyces* species with available long read data in our mitochondrial phylogenetic analyses. For the most part, our mitochondrial phylogenies matched our species-level phylogeny with the notable exception of a strain of *S. jurei* (NCYC3947), a recently described European

species (Naseeb et al. 2018) that appears to share mitochondrial ancestry with a subgroup of European strains of *S. paradoxus*. The mitochondrial genomes from this subgroup also contain large structural variations (previously described in Yue et al. 2017) relative to other strains of *S. paradoxus* and *S. cerevisae* further supporting their shared ancestry (**Fig. 5**).

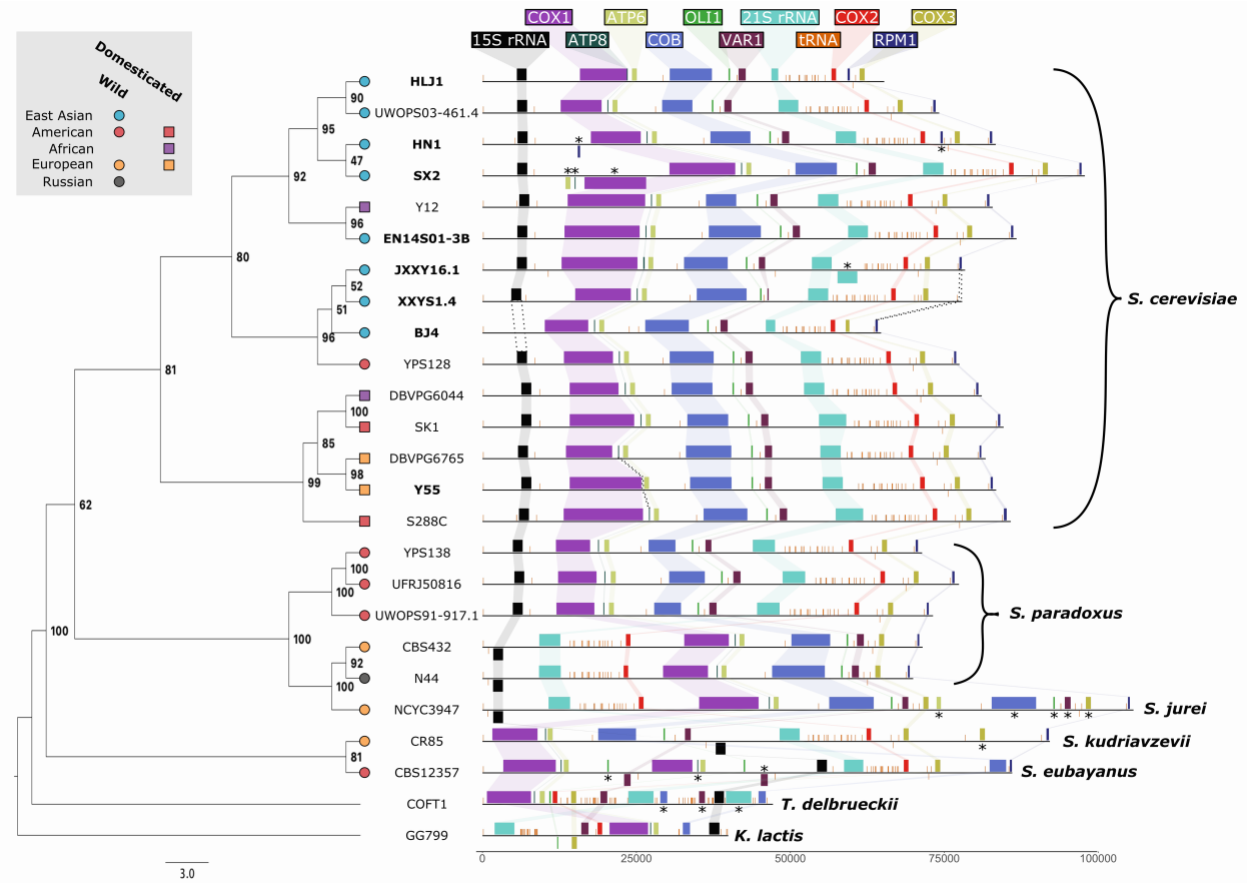

Figure 5: Mitochondrial phylogenetics and genomic arrangements. Phylogenetic tree based on mitochondrial genomic content. Internal branches are labeled with bootstrap support. *Saccharomyces* strains are colored according to their location of origin and branch tip shape indicates whether it is a domesticated (square) or wild (circle) strain. New long-read genome assemblies presented in this study are indicated in bold. Major genomic elements found on yeast mitochondria are shown and colored according to guide elements at the top. Inverted elements appear on the underside of the line. Missing elements are indicated by dashed lines. Duplicated elements are indicated with asterisk.

Discussion

Comparative genomic analyses have provided clues about the origin of Baker's yeast (*S. cerevisiae*) and have suggested an out-of-China origin (Peter et al. 2018). Here, we provide seven new, high quality long

220 read genomes of highly divergent wild strains recently isolated in East Asia. Phylogenomic analyses of the
221 long-read assemblies agree with previous findings that the wild East Asian strains (CHN, Taiwanese) are
222 basal relative to other *S. cerevisiae* strains (Duan et al. 2018; Peter et al. 2018) and, in the case of the CHN
223 IX and Taiwanese clades, show considerable divergence (**Fig. 1**). However, this is the first report where
224 representatives from these two clades are directly assessed and compared. We found that the two strains
225 from the CHN IX clade (JXXY16.1 and XXYS1.4) and the one strain from the Taiwanese clade (EN14S01-3B)
226 (collectively referred to as the Clade IX Complex) were highly divergent from other strains but share a
227 recent ancestry and evolutionary history. The placement of this clade relative to both the other East Asian
228 strains we describe, as well as other domesticated and lab strains from Africa and Europe, also indicates
229 that it diverged prior to the larger radiation of *S. cerevisiae* into other parts of the world.

230 In addition to providing insight into the evolutionary history of *S. cerevisiae*, our results in the
231 context of the seven previously assembled long-read genomes help reveal the extent of evolutionary
232 diversity in both genome structure and content. For example, BJ4 and HLJ1, belonging to CHNIV and CHVI
233 clades respectively, have genome sizes considerably smaller than not only all the *S. cerevisiae* strains
234 considered in our study (**Fig. 3A**) but also all *S. paradoxus*, *S. jurei*, *S. eubayanus*, and *S. jurei strains* (**Fig.**
235 **S5**). Contrary to our expectations that such size reductions would correspond to losses in non-coding
236 regions, we actually observed reduced amounts of genomic space taken up by coding regions (although
237 not in total number of genes) and an increase in intronic material in these shorter genomes (**Fig. 3A**).

238 The inclusion of the standard reference strain S288C in our analyses of genome structure and
239 content also adds to a growing consensus that, despite not being an outlier like BJ4 and HLJ1, S288C does
240 not reflect the average *S. cerevisiae* genome very well. One area where the use of the S288C genome as
241 representative appears problematic is the study of yeast transposable element evolution. Based on the
242 collective long read assemblies of divergent *S. cerevisiae* strains we analyzed here, we were able to
243 generate a broad scale map of *Ty* elements throughout this species (**Fig. 3B**). Previous conclusions based

244 largely on the S288C reference assembly have suggested that *Ty2* elements are a recent addition to *S.*
245 *cerevisiae*, arriving in the genome through horizontal gene transfer (HGT) from *S. mikatae* (Carr et al.
246 2012). While the lack of this element type in other closely related *Saccharomyces* species does provide
247 support for an HGT event, we found this element existing to varying degrees in even the most divergent
248 *S. cerevisiae* strains. This suggests an early origin of the *Ty2* element in the evolution of *S. cerevisiae.*
249 Without further genomic data available from diverse strains of *S. mikatae* it is difficult to delineate the
250 direction of this HGT. Previous findings have suggested that S288C may be a phenotypic extreme among
251 *S. cerevisiae* strains (Warringer et al. 2011). Overall, S288C may not be as representative of the species as
252 previously thought.

253 A similar level of diversity in TY composition among strains of *S. paradoxus*, *S. jurei*, *S. eubayanus*,
254 and *S. kudriavzevii* suggests a complex evolutionary history of TEs in *Saccharomyces*. TY2 and TY2p
255 elements were present in all *S. cerevisiae* strains but absent in other species, indicating their origin and
256 evolution only occurred after the divergence of *S. cerevisiae* from other Saccharomycotina.

257 Our comprehensive analyses of structural variations (SVs) further elucidates the evolutionary
258 history and intraspecific diversity of *S. cerevisiae*. Balanced and unbalanced SVs were identified for each
259 strain pair revealing patterns of genomic divergence. As expected, strains with reduced genome size (BJ4
260 and HLJ1) showed a large number of SVs dominated by unbalanced variations (**Fig. 4**). There is also a
261 noteworthy pattern of higher amounts of SVs in wild East Asian strains, especially the three strains within
262 the Clade IX Complex. As a species, *S. cerevisiae* has been shown to accumulate balanced variations at a
263 slower rate compared to *S. paradoxus* (Yue et al. 2017). This is likely due to the different selection histories
264 of these species. Many *S. cerevisiae* strains have long been associated with human activities where
265 domestication, cross-breeding and admixture have resulted in largely mosaic genomes (Liti et al. 2009;
266 Hyma and Fay 2013; Liti 2015), whereas *S. paradoxus* strains are recently isolated wild strains.
267 Interestingly, we found that wild East Asian strains accumulated both balanced and unbalanced SVs at a

high rate, more similar to rates normally seen in *S. paradoxus* (**Fig. 4**). It has been suggested that the geographical isolation of some *S. paradoxus* subpopulations may have favored quick fixation of structural rearrangements (Leducq et al. 2016). We may be witnessing similar patterns in the wild East Asian *S. cerevisiae* strains here.

Overall, this study generates new, valuable genomic resources and expands our understanding of the genetic variation and evolutionary history of this important microbial model system. This set of high-quality genomes, encompassing both domesticated and wild populations from different ecological backgrounds, provides a great testing field for open questions in evolutionary biology and fodder for future research into the dynamics that govern eukaryotic genome evolution.

Methods

Yeast strain origins

We selected eight *Saccharomyces cerevisiae* strains for long-read sequencing and genome assembly (**Table 1**). Seven of these strains are highly divergent strains originating from East Asia. Six strains were isolated in China (Wang et al. 2012; Duan et al. 2018) from a variety of ecological niches and one in Taiwan (Peter et al. 2018). The six Chinese strains cover many of the lineages (CHN I, II, IV, VI, and IX) previously shown to be highly divergent from other *S. cerevisiae* strains based on short-read sequencing. The final strain (Y55) is a common laboratory strain with a known mosaic genomic background originating in West Africa, but first isolated in France. To place our analyses in context, we also utilized all currently available *Saccharomyces* sensu stricto long read genome assemblies as well as assemblies from *Torulaspora delbrueckii* and *Kluyveromyce lactis* (**Table S1**).

290 *DNA preparation and long-read sequencing*

291 Before sequencing, strains were sporulated and tetrads were dissected to allow for autodiploidization,
292 making strains homozygous across all loci. Strains were incubated at 30°C in 5ml YEPD (1% yeast extract,
293 2% peptone, 2% dextrose) in a shaking incubator for 24h before we harvested cells by centrifugation. We
294 extracted genomic DNA using NucleoSpin Microbial DNA extraction kit according to the manufacturer's
295 instructions (Macherey-Nagel). Genomic DNA for strain Y55 was extracted independently using the
296 QIAGEN Blood & Culture DNA Midi Kit. Samples were sequenced on PacBio Sequel and Sequel II platforms
297 at the NGI/Uppsala Genome Center (Science for Life Laboratory, Sweden) and the University of Minnesota
298 Sequencing Center (USA). In addition to these PacBio data, we also used publicly available paired-end
299 Illumina sequence data previously generated for each strain (**Table S1**).

300

301 *Genome assembly and annotation*

302 Contigs were assembled using raw PacBio reads with Flye (default settings) (Kolmogorov et al. 2019). We
303 used short read sequences for each strain to error-correct the long reads using FMLRC (Wang et al. 2018).
304 These corrected long reads and the short reads were subsequently used to polish the Flye assemblies
305 using Racon (Vaser et al. 2017) and POLCA (Zimin and Salzberg 2019) respectively. We further scaffolded
306 the contigs based on the reference S288C genome (GCA_000146045.2) using RaGOO (Alonge et al. 2019)
307 and filled any gaps this generated using multiple iterations of LR Gapcloser (Xu et al. 2019) and Gapcloser
308 (Luo et al. 2012). To account for any errors introduced by using long reads to fill gaps, we polished each
309 assembly once more using POLCA. We used REAPR (Hunt et al. 2013) to check for any misassemblies.
310 Assembly processing was conducted independently for each nuclear and mitochondrial genome. We
311 further adjusted the mitochondrial genomes by replacing large deletions of >500bp (relative to S288C)
312 with proportionate lengths of N sequence filler and repeated previous gap filling and polishing steps.

313 Completeness of the final genome assemblies was assessed using BUSCO v4.0.5 (Simão et al. 2015;
314 Waterhouse et al. 2018).

315 We annotated nuclear genes, mitochondrial genes, centromeres, transposable elements, core X
316 elements and Y-prime elements using pipelines within the LRSDAY package (Yue and Liti 2018). In addition
317 to our 8 newly assembled genomes, we also used the same method to annotate the other long read
318 assemblies publicly available for S. cerevisiae strains as well as related yeast species (**Table S2**). Nuclear
319 genes orthologous to annotated genes in the S. cerevisiae S288C reference genome were identified using
320 Proteinortho (Lechner et al. 2011). Genes for which no orthologous protein was found in the reference
321 were clustered based on orthology to each other.

322

323 *Phylogenomic analysis*

324 To place our 8 assembled genomes within the context of other *Saccharomyces* strains, we employed both
325 a consensus gene tree and Assembly and Alignment-Free (AAF) approach to phylogenetic tree
326 construction. For the consensus gene tree approach, we aligned all orthologous genes found in at least 5
327 strains (5847 genes) using MUSCLE (Edgar 2004) and performed maximum-likelihood single-tree inference
328 for each locus using RAxML-NG (Kozlov et al. 2019) with a discrete GAMMA model of rate heterogeneity.
329 We used Astral-III (Zhang et al. 2018) with these gene trees to generate a consensus species tree (**Fig. 1**).
330 Alternatively, AAF (Fan et al. 2015) was used with a k-mer size of 40 nucleotides and a threshold frequency
331 of 3 for each k-mer to be included in the analysis. First, AAF was used on the long-read sequencing data
332 generated in this study and compared to previously published Illumina short-read sequencing data for the
333 same strains (**Fig. S2, Table S1**). We then used a non-parametric bootstrap (n=100) to generate a
334 consensus tree using the N44 genome as an outgroup of the analysis resulting in 100% bootstrap support
335 across the tree. We then used AAF with the same parameters to build a phylogenetic network of the 8 *S.*

336 *cerevisiae* strains presented in this study along with 13 other *Saccharomyces* genomes that have
337 previously been sequenced using long-read sequencing (**Fig. S1, Table S1**).
338 To generate the mitochondrial phylogeny, we reoriented the start of each assembly based on the
339 position of the tRNA gene, trnP(ugg), then aligned these assemblies to each other using Mugsy (Angiuoli
340 and Salzberg 2011). This multiple sequence alignment was then used to create a maximum likelihood tree
341 using IQ-TREE (v2.0.5) (options: -m TPM2u+F+R3 -B 1000 -bnni -alrt 1000) (Nguyen et al. 2015; Hoang et
342 al. 2018). The model was determined using the ModelFinder component of IQ-Tree (Kalyaanamoorthy et
343 al. 2017).
344
345 *Unbalanced and balanced structural variation*
346 To identify the structural variations between strains within *Saccharomyces cerevisiae*, we performed
347 exhaustive pairwise comparisons between the 15 strains with long-read assemblies (210 comparisons).
348 To identify unbalanced variation, we used nucmer (Kurtz et al. 2004) to generate a whole-genome
349 alignment for each strain pair (options: -maxmatch -c 500). We then used Assemblytics (Nattestad and
350 Schatz 2016) to identify repeat expansions (RE) and contractions (RC), tandem expansions (TE) and
351 contractions (TC), insertions (IS) and deletions (DL) (options: unique sequence length required = 10000,
352 minimum variant size = 50, maximum variant size = 10000). We then filtered these variants to retain only
353 those that encompassed at least a single protein-coding gene (CDS) by intersecting our variants with our
354 genome annotations using bedtools intersect (Quinlan and Hall 2010). This method of identifying
355 structural variations results in two sets of identified variations within a pairwise comparison, one with
356 each genome acting as reference sequence and the other acting as query sequence. These sets are often
357 identical (albeit reversed), however some variants might be found in only one direction and missed by the
358 algorithm in the other. In these instances, the average of the two sets of variations is reported (sometimes
359 resulting in a half variation being reported).

In order to identify balanced variations, we first identified conserved synteny blocks for each pairwise comparison using SynChro from the CHROnicle package (Drillon et al. 2013; Drillon et al. 2014). The input files were prepared for SynChro by separating genome fasta files and GFF annotation files by chromosome and converting them to EMBL format using EMBLmyGFF3 (Norling et al. 2018). We then used ReChro from the CHROnicle package (Vakirlis et al. 2016), with synteny block stringency parameter (delta) set to 0, to identify inversions (IV), translocations (TR), fusions (FU) and fissions (FI).

Data Access

All raw sequencing reads and complete genome assemblies from this study have been submitted to the European Nucleotide Archive under accession number **XXXXXXXXX**.

Acknowledgments

We would like to thank Feng-Yan Bai and Gianni Liti for donating strains, Jia-Xing Yue for advice on the LRSDAY pipeline, and Gianni Liti and Chris Wheat for discussion. We acknowledge the support of the National Genomics Infrastructure (NGI)/Uppsala Genome Center and UPPMAX for providing assistance in massive parallel sequencing and computational infrastructure. Work performed at NGI/Uppsala Genome Center (project SNIC 2019/8-23) has been funded by RFI/VR and Science for Life Laboratory, Sweden. This work was supported by grants from Stockholm University (Science for Life Laboratory sequencing grant SU FV-2.2.2-1843-17 to RS), the Swedish Research Council (2017-04963 to RS), the Knut and Alice Wallenberg Foundation (2017.0163 to RS), the Carl Trygger foundation (to ZZ), the Wenner-Gren Foundations (UPD2018-0196, UPD2019-0110 to DPB), and The University of Minnesota Department of Ecology, Evolution, and Behavior.

382 ***References***

383 Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC.
384 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome*
385 *biology* **20**: 1-17.
386 Angiuoli SV, Salzberg SL. 2011. Mugsy: fast multiple alignment of closely related whole
387 genomes. *Bioinformatics* **27**: 334-342.
388 Carr M, Bensasson D, Bergman CM. 2012. Evolutionary genomics of transposable elements in
389 Saccharomyces cerevisiae. *PloS one* **7**.
390 De Chiara M, Friedrich A, Barré B, Breitenbach M, Schacherer J, Liti G. 2020. Discordant
391 evolution of mitochondrial and nuclear yeast genomes at population level. *BMC biology*
392 **18**: 1-15.
393 Drillon G, Carbone A, Fischer G. 2013. Combinatorics of chromosomal rearrangements based on
394 synteny blocks and synteny packs. *Journal of Logic and Computation* **23**: 815-838.
395 Drillon G, Carbone A, Fischer G. 2014. SynChro: a fast and easy tool to reconstruct and visualize
396 synteny blocks along eukaryotic chromosomes. *PLoS One* **9**.
397 Duan S-F, Han P-J, Wang Q-M, Liu W-Q, Shi J-Y, Li K, Zhang X-L, Bai F-Y. 2018. The origin and
398 adaptive evolution of domesticated populations of yeast from Far East Asia. *Nature*
399 *communications* **9**: 1-13.
400 Fan H, Ives AR, Surget-Groba Y, Cannon CH. 2015. An assembly and alignment-free method of
401 phylogeny reconstruction from next-generation sequencing data. *BMC genomics* **16**:
402 522.
403 Feder JL, Nosil P, Flaxman SM. 2014. Assessing when chromosomal rearrangements affect the
404 dynamics of speciation: implications from computer simulations. *Frontiers in genetics* **5**:
405 295.
406 Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation
407 sequencing technologies. *Nature Reviews Genetics* **17**: 333.
408 Hebly M, Brickwedde A, Bolat I, Driessen MR, de Hulster EA, van den Broek M, Pronk JT,
409 Geertman J-M, Daran J-M, Daran-Lapujade P. 2015. S. cerevisiae× S. eubayanus
410 interspecific hybrid, the best of both worlds and beyond. *FEMS Yeast Research* **15**.
411 Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the
412 ultrafast bootstrap approximation. *Molecular biology and evolution* **35**: 518-522.
413 Hoffmann AA, Rieseberg LH. 2008. Revisiting the impact of inversions in evolution: from
414 population genetic markers to drivers of adaptive shifts and speciation? *Annual review*
415 *of ecology, evolution, and systematics* **39**: 21-42.
416 Hooper DM, Price TD. 2017. Chromosomal inversion differences correlate with range overlap in
417 passerine birds. *Nature ecology & evolution* **1**: 1526-1534.
418 Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. 2013. REAPR: a universal tool
419 for genome assembly evaluation. *Genome biology* **14**: R47.
420 Hyma KE, Fay JC. 2013. Mixing of vineyard and oak-tree ecotypes of Saccharomyces cerevisiae
421 in North American vineyards. *Molecular ecology* **22**: 2917-2930.
422 Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck
423 FJ. 2017. Transient structural variations have strong effects on quantitative traits and
424 reproductive isolation in fission yeast. *Nature communications* **8**: 1-11.

Kalyaanamoorthy S, Minh BQ, Wong TK, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods* **14**: 587.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology* **37**: 540-546.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome biology* **5**: R12.

Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ. 2011. Proteinortho: detection of (co-) orthologs in large-scale analysis. *BMC bioinformatics* **12**: 124.

Leducq JB, Nielly-Thibault L, Charron G, Eberlein C, Verta JP, Samani P, Sylvester K, Hittinger CT, Bell G, Landry CR. 2016. Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nat Microbiol* **1**: 15003.

Liti G. 2015. The fascinating and secret wild life of the budding yeast S. cerevisiae. *eLife* **4:e0585**.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V. 2009. Population genomics of domestic and wild yeasts. *Nature* **458**: 337-341.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**: 2047-2217X-2041-2018.

Macías LG, Morard M, Toft C, Barrio E. 2019. Comparative genomics between Saccharomyces kudriavzevii and S. cerevisiae applied to identify mechanisms involved in adaptation. *Frontiers in Genetics* **10**: 187.

Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, Waggott D, Utiramerur S, Hou Y, Smith KS. 2018. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genetics in Medicine* **20**: 159-163.

Naseeb S, Alsammar H, Burgis T, Donaldson I, Knyazev N, Knight C, Delneri D. 2018. Whole genome sequencing, de novo assembly and phenotypic profiling for the new budding yeast species Saccharomyces jurei. *G3: Genes, Genomes, Genetics* **8**: 2967-2977.

Nattestad M, Schatz MC. 2016. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**: 3021-3023.

Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution* **32**: 268-274.

Norling M, Jareborg N, Dainat J. 2018. EMBLmyGFF3: a converter facilitating genome annotation submission to European Nucleotide Archive. *BMC research notes* **11**: 1-5.

Peris D, Arias A, Orlić S, Belloch C, Perez-Traves L, Querol A, Barrio E. 2017. Mitochondrial introgression suggests extensive ancestral hybridization events among Saccharomyces species. *Molecular phylogenetics and evolution* **108**: 49-60.

Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, Sigwalt A, Barre B, Freel K, Llored A. 2018. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* **556**: 339.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

Sanchis-Juan A, Stephens J, French CE, Gleadall N, Mégy K, Penkett C, Shamardina O, Stirrups K, Delon I, Dewhurst E. 2018. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short-and long-read genome sequencing. *Genome medicine* **10**: 95.

Scannell DR, Zill OA, Rokas A, Payen C, Dunham MJ, Eisen MB, Rine J, Johnston M, Hittinger CT. 2011. The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. *G3 (Bethesda)* **1**: 11-25.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210-3212.

Vakirlis N, Sarilar V, Drillon G, Fleiss A, Agier N, Meyniel J-P, Blanpain L, Carbone A, Devillers H, Dubois K. 2016. Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. *Genome research* **26**: 918-932.

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research* **27**: 737-746.

Wang JR, Holt J, McMillan L, Jones CD. 2018. FMLRC: Hybrid long read error correction using an FM-index. *BMC bioinformatics* **19**: 50.

Wang QM, Liu WQ, Liti G, Wang SA, Bai FY. 2012. Surprisingly diverged populations of Saccharomyces cerevisiae in natural environments remote from human activity. *Molecular ecology* **21**: 5404-5417.

Warringer J, Zörgö E, Cubillos FA, Zia A, Gjuvsland A, Simpson JT, Forsmark A, Durbin R, Omholt SW, Louis EJ. 2011. Trait variation in yeast is defined by population history. *PLoS genetics* **7**.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular biology and evolution* **35**: 543-548.

Wellenreuther M, Mérot C, Berdan E, Bernatchez L. 2019. Going beyond SNPs: the role of structural genomic variants in adaptive evolution and species diversification. *Molecular ecology* **28**: 1203-1209.

Xu G-C, Xu T-J, Zhu R, Zhang Y, Li S-Q, Wang H-W, Li J-T. 2019. LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *GigaScience* **8**: giy157.

Yona AH, Manor YS, Herbst RH, Romano GH, Mitchell A, Kupiec M, Pilpel Y, Dahan O. 2012. Chromosomal duplication is a transient evolutionary solution to stress. *Proceedings of the National Academy of Sciences* **109**: 21010-21015.

Yue J-X, Li J, Aigrain L, Hallin J, Persson K, Oliver K, Bergström A, Coupland P, Warringer J, Lagomarsino MC. 2017. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nature genetics* **49**: 913.

Yue J-X, Liti G. 2018. Long-read sequencing data analysis for yeasts. *Nature protocols* **13**: 1213.

Zhang Z, Bendixsen DP, Janzen T, Nolte AW, Greig D, Stelkens R. 2020. Recombining your way out of trouble: The genetic architecture of hybrid fitness under environmental stress. *Molecular biology and evolution* **37**: 167-182.

Zheng D-Q, Chen J, Zhang K, Gao K-H, Li O, Wang P-M, Zhang X-Y, Du F-G, Sun P-Y, Qu A-M. 2014. Genomic structural variations contribute to trait improvement during whole-genome shuffling of yeast. *Applied microbiology and biotechnology* **98**: 3059-3070.

Zimin AV, Salzberg SL. 2019. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *bioRxiv*.