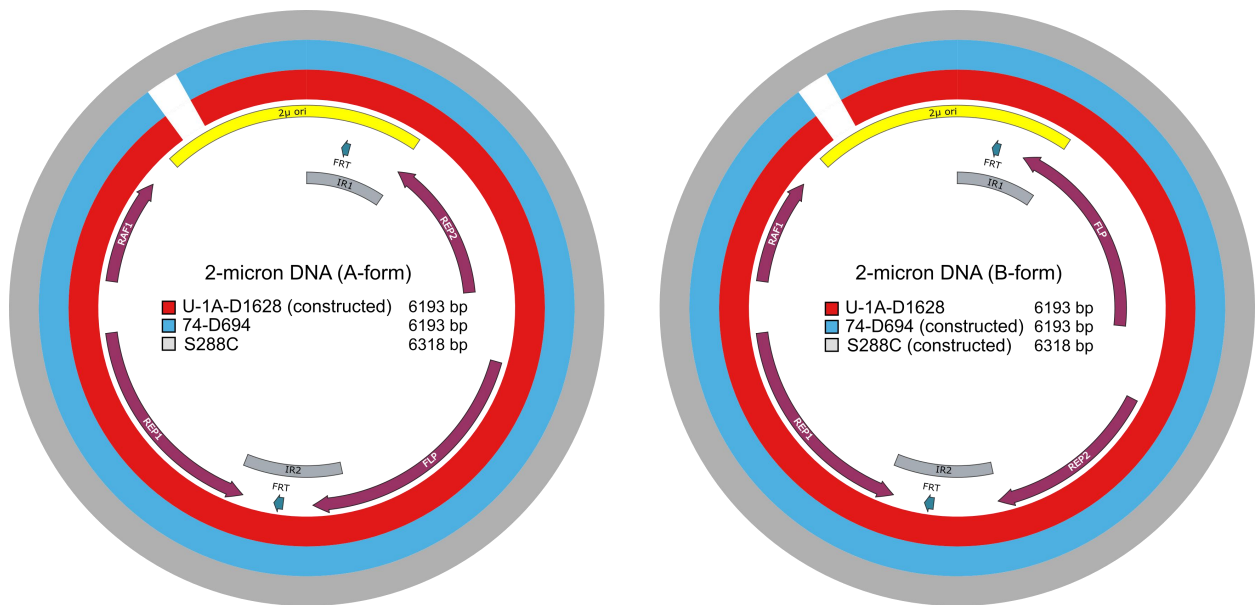


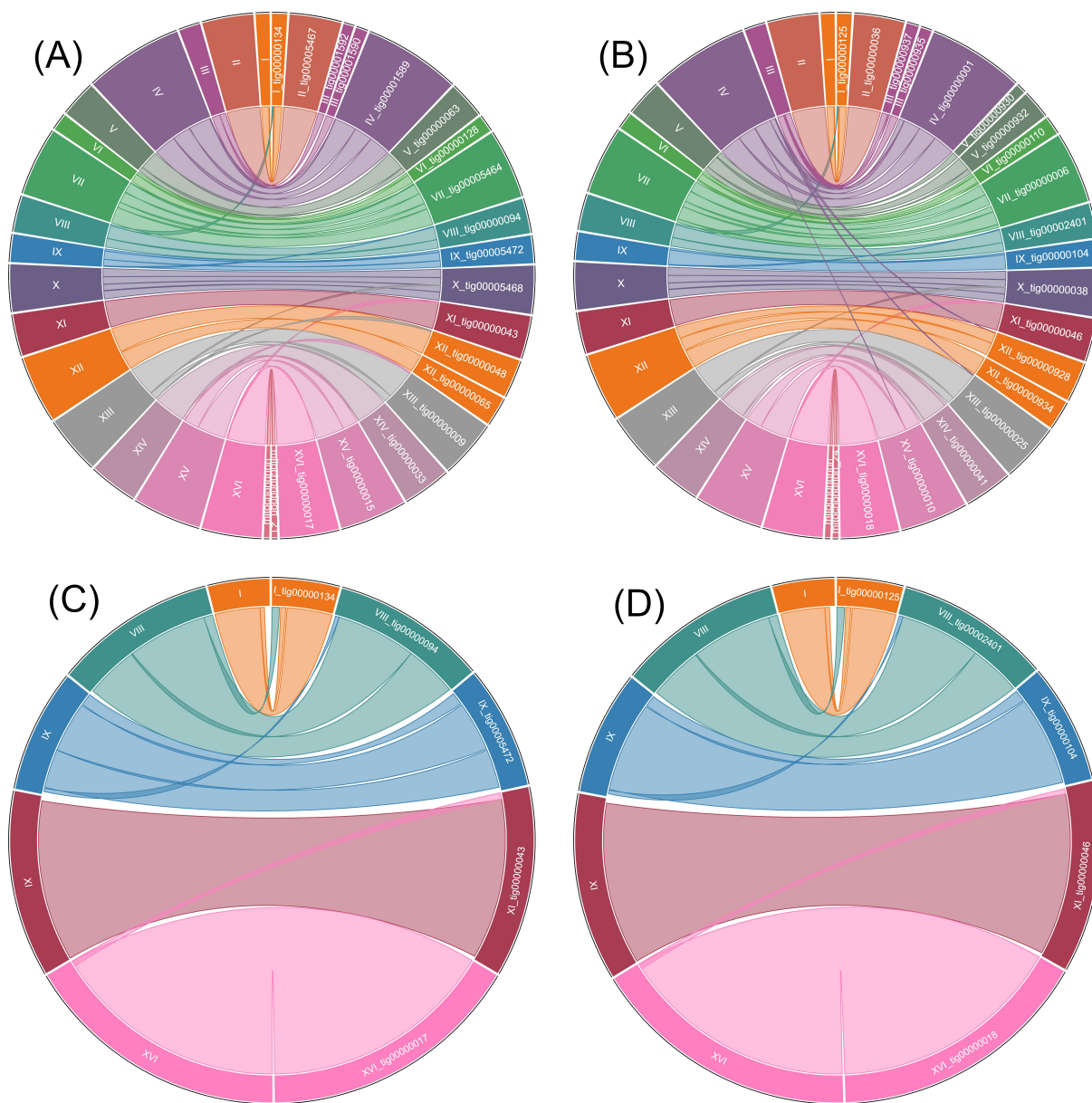
## Chromosome-level genome assembly and structural variant analysis of two laboratory yeast strains from the Peterhof Genetic Collection lineage

Yury A. Barbitoff, Andrew G. Matveenko, Anton B. Matiiv, Evgeniia M. Maksiutenko, Svetlana E. Moskalenko, Polina B. Drozdova, Dmitrii E. Polev, Alexandra Y. Beliavskaia, Lavrentii G. Danilov, Alexander V. Predeus, and Galina A. Zhouravleva

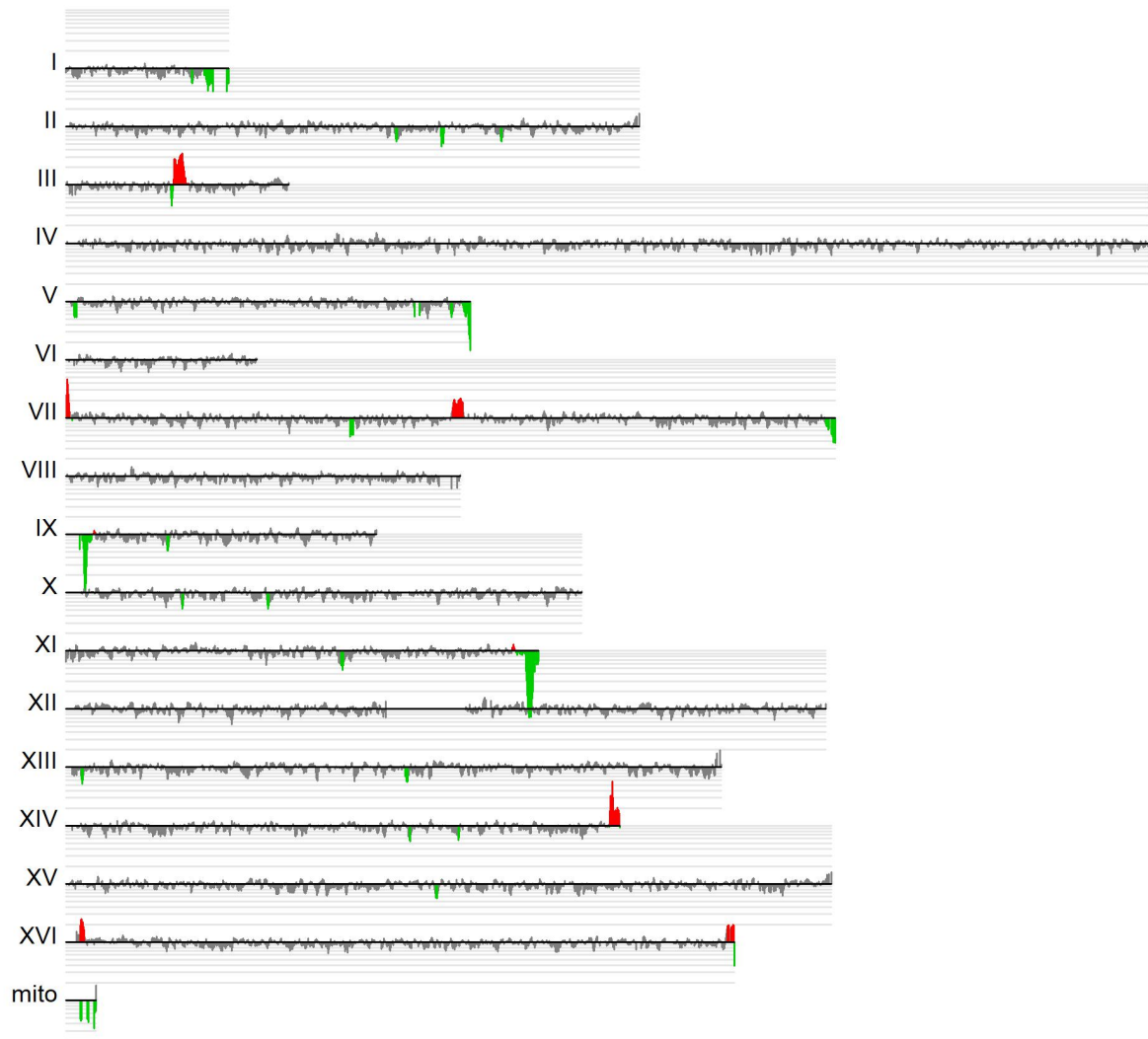
### Supplementary Figures and Tables



**Figure S1.** Alignment of the 2-micron DNA sequences of the respective strains. For PGC strains, 2-micron plasmid sequence was assembled using a hybrid assembly framework. For the U-1A-D1628 strain, the A-form of 2-micron DNA was constructed manually from the assembled B-form by flipping the sequence between two inverted repeats. Similarly, B-forms were constructed for 74-D694 and S288C. Sequences were aligned using ClustalW and the alignment was visualized using a custom set of scripts. Feature maps were created using SnapGene Viewer v. 5.2.3 ([www.snapgene.com/snapgene-viewer](http://www.snapgene.com/snapgene-viewer)); IR1 and IR2 - inverted repeats, FRT - Flp recognition target, 2 $\mu$  ori - origin of replication.

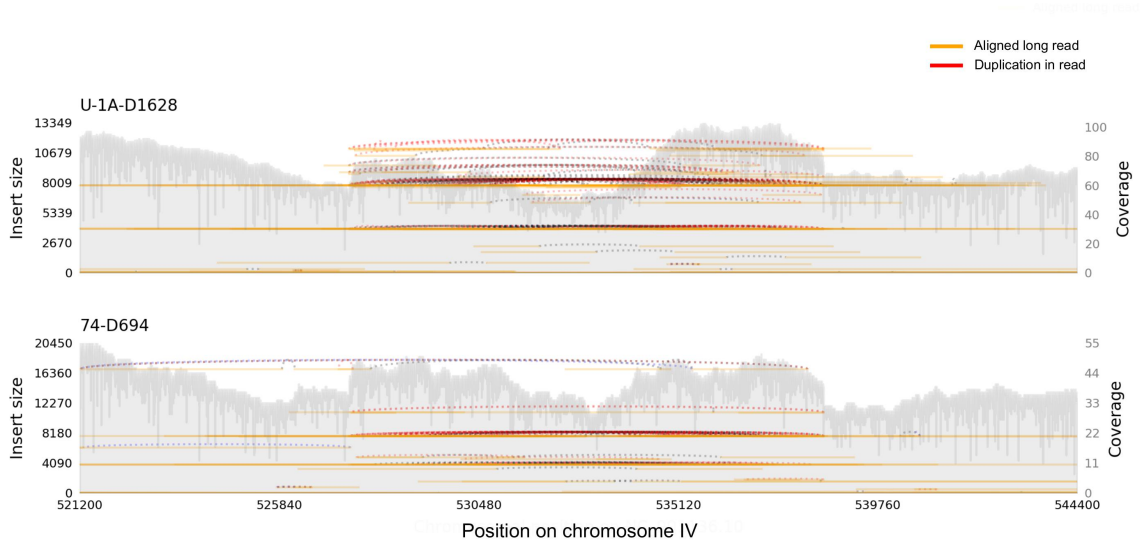


**Figure S2.** Alignment of genome assemblies of U-1A-D1628 and 74-D694 strains to the reference genome of S288C. Alignment was performed using MashMap and visualization was depicted as circo plots using Circa. A. Alignment of genome assembly of U-1A-D1628 (right semicircle) to the reference genome of S288C (left semicircle). B. Alignment of genome assembly of 74-D694 (right semicircle) to the reference genome of S288C (left semicircle). C. Alignment of the chromosomes I, VIII, IX, XI and XVI from assembly of U-1A-D1628 (right semicircle) to the corresponding chromosomes of the reference genome of S288C (left semicircle). D. Alignment of the chromosomes I, VIII, IX, XI and XVI from assembly of 74-D694 (right semicircle) to the corresponding chromosomes of the reference genome of S288C (left semicircle).

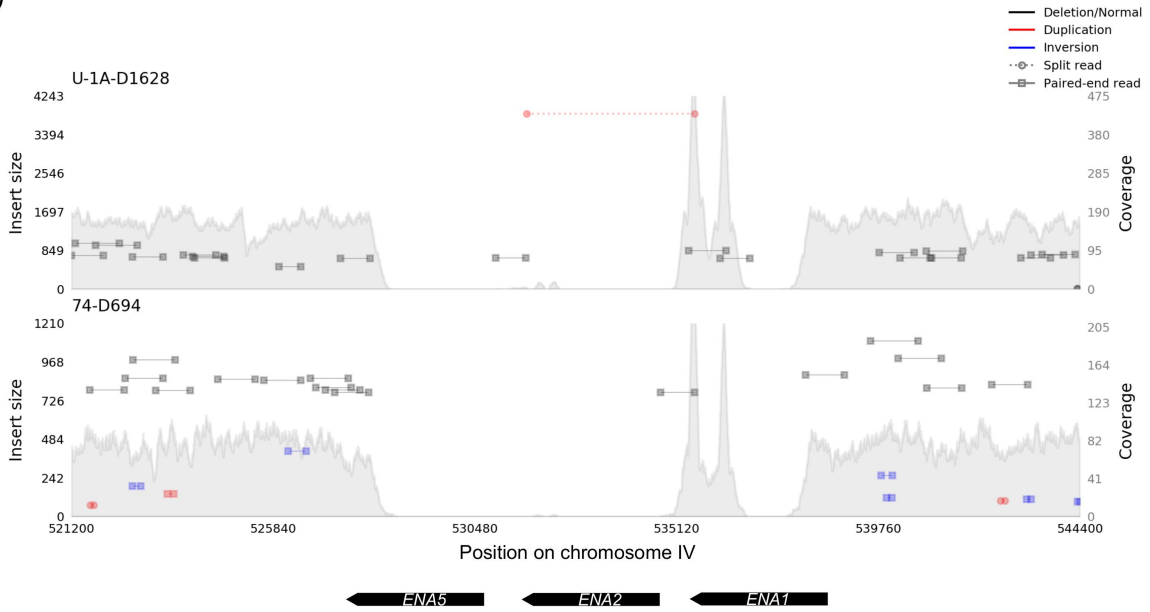


**Figure S3.** Representative CGH CLAC plot of U-1A-D1628. Regions of reference chromosomes with higher coverage compared to PSL2 strain (isogenic to W303-1A) are shown in red, while lower coverage shown in green (FDR=0.579). Total of three replicates were analysed. Raw CGH data are presented on Table S3.

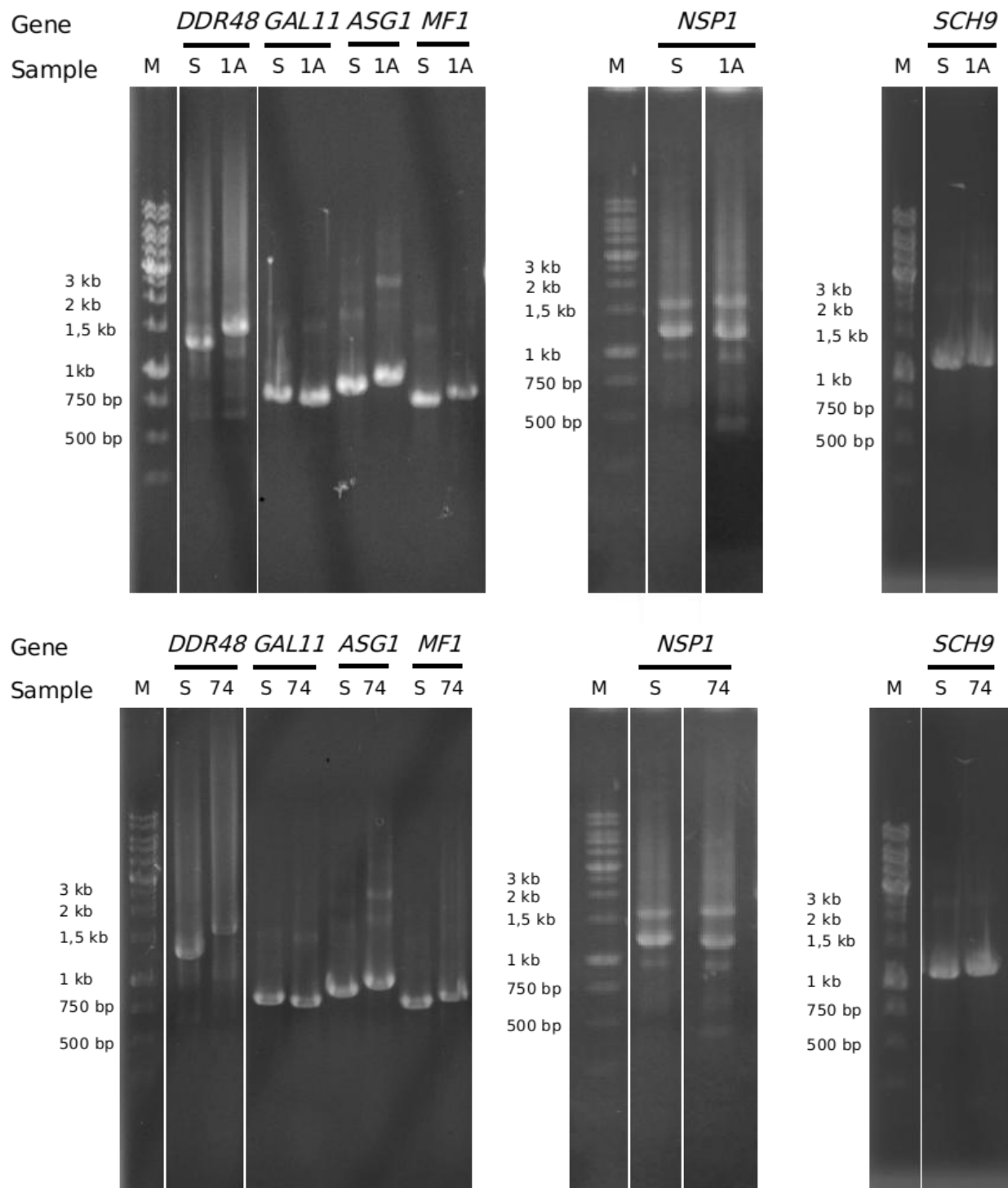
(A)



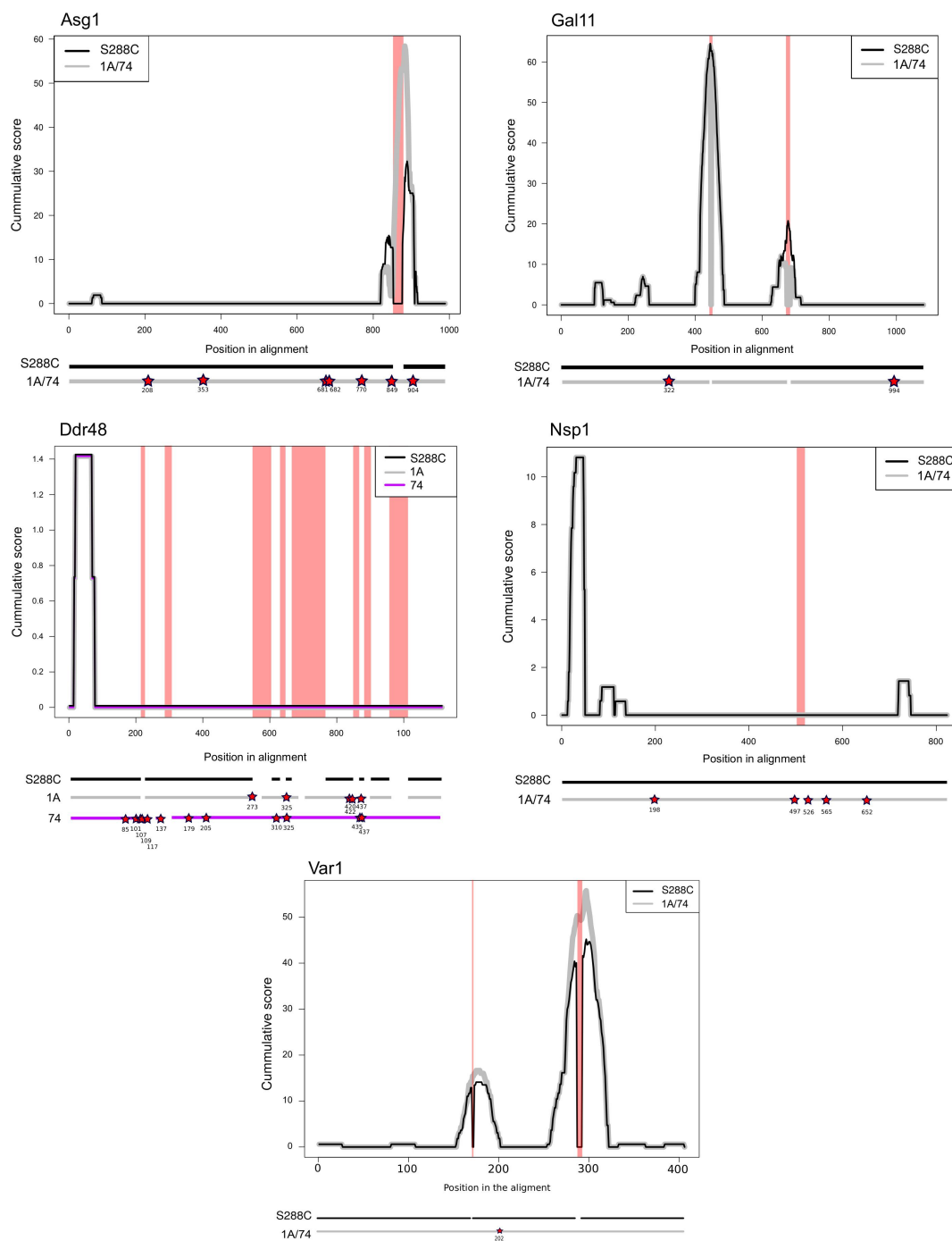
(B)



**Figure S4.** Visualization of read coverage profiles for U-1A-D1628 and 74-D694 stains in the *CUP1* locus on chromosome VIII (A) and *ENA* locus on chromosome IV. (A) - long (Oxford Nanopore) reads; (B) - short (Illumina HiSeq) reads. On (B), reads with MQ = 0 (ambiguously mapped) are excluded from visualization. Plots were generated using the samplot package. Locations of *ENA* genes are shown according to the RefSeq annotation.



**Figure S5.** PCR-based validation of structural variants in the U-1A-D1628 and 74-D694 strains. S, 74, and 1A correspond to S1 (S288C), 74-D694, and U-1A-D1628 genomic DNA, used as a template; M, DNA molecular weight marker (SibEnzyme, 1kb).



**Figure S6.** Amyloidogenicity prediction for the proteins harboring insertions or deletions in tandem repeat coding sequences. The stars indicate amino acid substitutions with respect to the S288C protein sequence. Coordinates of the substitutions represent positions in the alignment. For all proteins except Ddr48 the sequences derived from U-1A-D1628 and 74-D694 are identical. For Ddr48, amyloidogenicity profiles of sequences from PGC strains are identical. The plots are slightly shifted to avoid complete overlap of lines. On all plots, red areas indicate the positions of alignment gaps.

**Table S1.** Expected length of genomic DNA fragments used for PCR-based structural variant validation and estimation of *CUP1* gene copy number in each of the three strains.

Primer pairs	Sequences	Reference (S288C) (bp)	U-1A- D1628 (bp)	74- D694 (bp)
ASG1-2107-F-NcoI ASG1-R-SacII	GCAACCCATGGTATTCTTTCTTCCCAGGATACGAAG GATACCGCGGTCCTTCAGAGGGGTAATTTAAAGGTAGGTA	796	871	871
GAL11-1684-F GAL11-2431-R	CCACAAGTCTACATCATCACAAAG GAGAAGGATTTGTATTTGGGGTTG	748	718	718
DDR48-F-BamHI DDR48-R-SacII	AAAAGGATCCATGGGTTTATTTGATAAAGTGAAGCA ATCCCCGCGGTCCGTAATCGTCGTCACCACCG	1290	1488	1638
MF(ALPHA)1_pro-F MF(ALPHA)1_out-R-SalI	ACAACAGGTTTTGGATAAGTACAT CAAAGTCGACTTTGTGTACATCTACAC	693	756	756
Sch9-F-SpeI Sch9-R-402-BamHI	CGAATAACTAGTATGATGAATTTTTTTACATCAAAATCG GATTGGGATCCCGTGTCTGTTTGTAAGTCCATTG	1207	1255	1255
NSP1-963-F NSP1-2132-R	ATAAGACAACCTAACACAACCCC ACTGACTAATTTGTTACCTCC	1212	1155	1155
FLO5start-F (1)* FLO5-5B (2)	AAAAATGCCTGTGGCTGCTC GTTGACCGTTGGTACCGGT	1047	864	864
FLO5start-F (1) CSS1inner-R (4)	AAATGACAATTGCACACCACTG GATGCTGAAGAAGTAATGGAAGTCA	n.a.	1939	1939
CSS1start-F (3) CSS1inner-R (4)	ATGTTCAATCGTTTAAACAAATTCCAAG GATGCTGAAGAAGTAATGGAAGTCA	869	n.a.	n.a.
FLO10start-F (5) FLO10-547-R (6)	AAAAATGCCTGTGGCTGCTC CCAACCGATAAAATTGCTGAATC	573	572	572
FLO10-2663-F (7) YPR195Cout(-1467)-F (8)	CAAGGAAACCATGTCGTCTGA TACAAGTTGAGGGTGTAAGTGAAG	n.a.	434	434
CUP1-RT-FW CUP1-RT-RV	AAGGTCATGAGTGCCAATGC ATTTCCCAGAGCAGCATGAC	153, 2151	153, 2150	153, 2150
ACT1-F ACT1-R	TCGAACAAGAAATGCAAACCGCTGC GACTTGACCATCTGGAAGTTCGTAGG	74	74	74

\* Bold numbers in parentheses are the numbers of the primers on Figure 3A; n.a. indicates that the primer pair does not support amplification of any fragment.

**Table S2.** Amino acid substitutions in mitochondrial proteins in U-1A-D1628 and 74-D694 strains

Protein	Amino acid substitutions
Intron-encoded RNA maturase bI4	E433K, Q461K
Intron-encoded DNA endonuclease aI4	S224A
Cytochrome b mRNA maturase bI3	G269V, D350N, E412K, Q434M, S437N, T451N, Q452N, T466M, D467N, K468E, R463K, G471D, P515S
Cytochrome b mRNA maturase bI2	M139L
Intron-encoded DNA endonuclease aI5 beta	F137SPPASAPRAGRT, L331LVPGPVQGPEPRRG
ATP synthase subunit a	F177I, M231I, G241S, A245T, A255T, V256L
Cytochrome c oxidase subunit I	H241Q, P242T, E243V, V244A, Y245T, L247I, I248M, I249L, H378Y, N492A
Cytochrome b	H253Q, G260N, P262T, L263M, V264L, P266Y, A267M, S268N, I269C
Mitochondrial 37S ribosomal protein Var1	N170NNN, K200N, N284NNNNNNN

**Table S3.** Results of the structural variant analysis (available as a Supplementary File).

**Table S4.** Curated structural variations in protein coding genes (available as a Supplementary File).

**Table S5.** CGH experiments raw data and CGH-Miner output (available as a Supplementary File).



**Table S6.** Summary of *CUPI* copy number estimation

Strain	<i>CUPI</i> copy number		Cu <sup>2+</sup> inhibitory concentration, mM
	Sequencing data	Other data	
S288C	2 (SGD) 11 (PacBio, Yue et al., 2017) 1-34 (depending on the assembler and sequencing platform, Giordano et al., 2017)	14 (Southern blot hybridization, Zhao et al. 2014)	2 (this work and Zhao et al. 2014)
U-1A-D1628	8 (this work)	2 times less than S288C (qPCR, this work)	0.75 (this work)
74-D694	8 (this work)	2 times less than S288C (qPCR, this work)	0.75 (this work)