Supplemental Figures, Tables, and Codes

Supplemental Figures

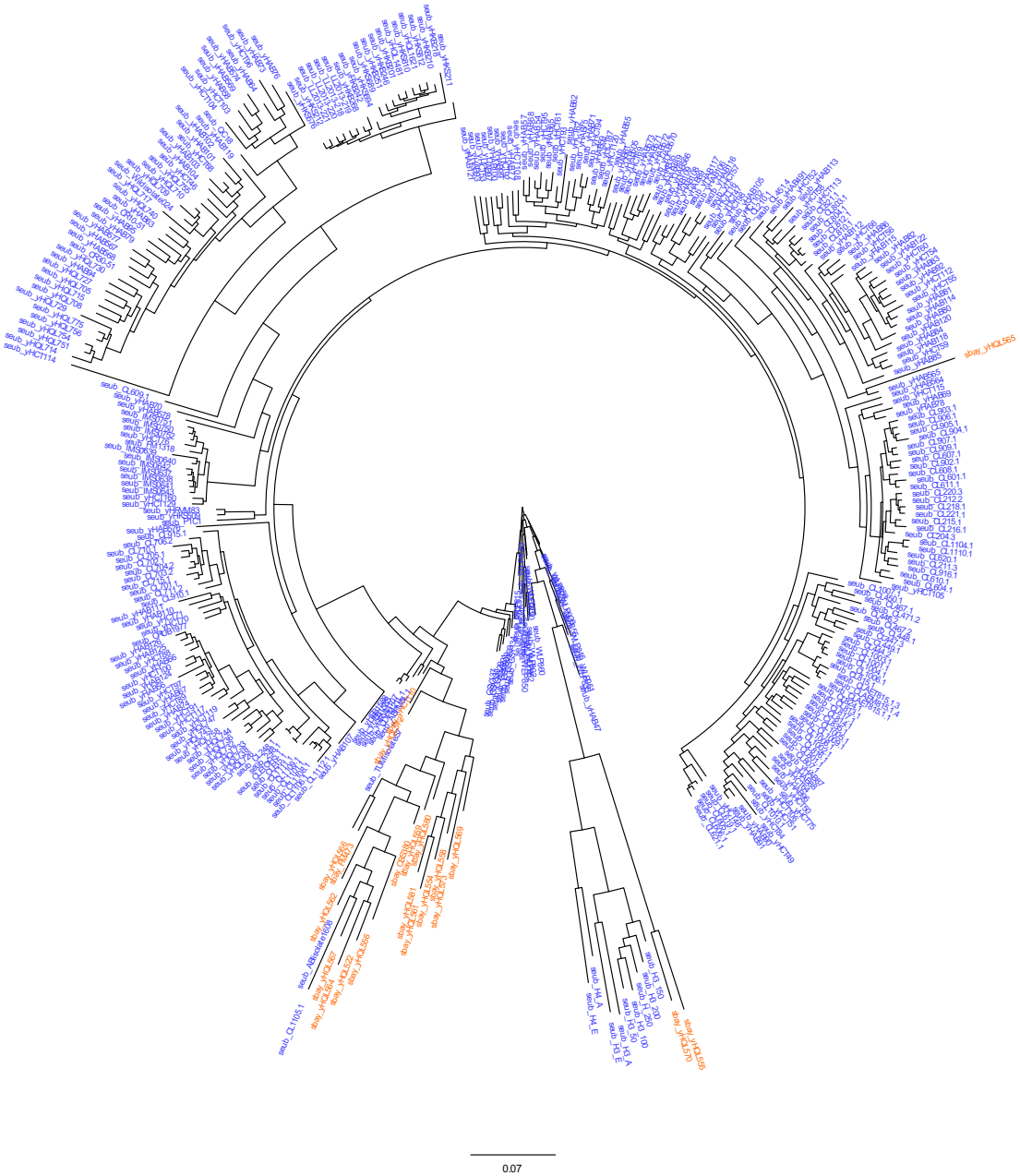
- Supplemental Figure S1: Phylogenetic tree showing the evolutionary relationships among *S. bayanus* and *S. eubayanus* strains.
- Supplemental Figure S2: Flow cytometry analysis supports a diploid genome of *S. bayanus* CBS380.
- Supplemental Figure S3: Overview of *S. bayanus* CBS 380 genome sequencing results using Oxford Nanopore's MinION.
- Supplemental Figure S4: Percentages of Nanopore reads mapped to different reference genomes.
- Supplemental Figure S5: Percentages of MinION reads mapped to different mitochondrial genomes.
- Supplemental Figure S6: Method for inferring parental origin of hybrid sequences using Local BLAST Analysis.
- Supplemental Figure S7: Inference of genomic origin based on re-mapping of reads assigned to *S. uvarum* and *S. eubayanus*.
- Supplemental Figure S8: Distribution of LTR and breakpoints of major recombination events.
- Supplemental Figure S9: Identification of chromosome aneuploidy based on normalized read depth.
- Supplemental Figure S10: Chromosomal collinearity analysis for Chr VI and Chr X in different *S. uvarum* strains.
- Supplemental Figure S11: Histogram showing distribution of sequence divergence levels between *S. bayanus* genes and their orthologous genes in six *Saccharomyces* species.
- Supplemental Figure S12: Boxplot showing the distribution of genomic content originated from *S. eubayanus*.
- Supplemental Figure S13: Phylogenetic trees of the *MALS* (A) and *MALT* (B) gene families.
- Supplemental Figure S14: A chromosome map showing locations of members of the three *MAL* gene families in the *S. bayanus* genome.

Supplemental Tables

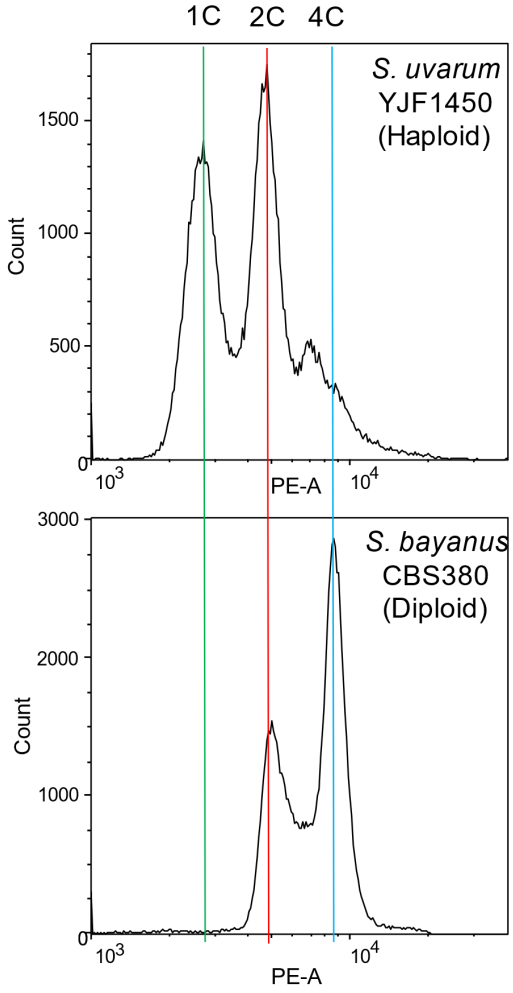
- Supplemental Table S1: Assembly statistics of the *S. bayanus* CBS 380 genome generated by different bioinformatic tools.
- Supplemental Table S2: A list of assembled sequence lengths for both haplotypes in *S. bayanus*.
- Supplemental Table S3: A list of telomere repeat sequences identified in *S. bayanus* chromosomes.
- Supplemental Table S4: A list of proportions of parental genome content in each *S. bayanus* chromosome estimated by two different methods.
- Supplemental Table S5: Breakpoints of major recombination events are supported by sequencing reads.
- Supplemental Table S6: A list of enriched GO terms "orphan genes" in *S. bayanus* CBS380.

Supplemental Codes

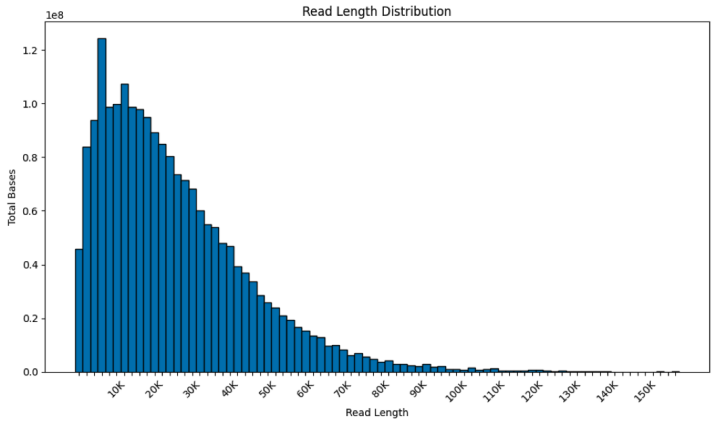
- Supplemental Code C1: blast_comparison.py
- Supplemental Code C2: pairwise_p_distance_calc.py



Supplemental Figure S1: **Phylogenetic tree showing the evolutionary relationships among *S. bayanus* and *S. eubayanus* strains.** The phylogenetic tree was constructed based on SNP variants in 21 *S. bayanus* strains and 347 *S. eubayanus* strains. SNPs were identified by mapping raw sequencing reads to Chr VI and Chr X. Sample names in blue are *S. eubayanus* strains and in orange are *S. bayanus* strains. *S. bayanus* strains form polyphyletic groups suggesting they were created by hybridization from different *S. eubayanus* strains.

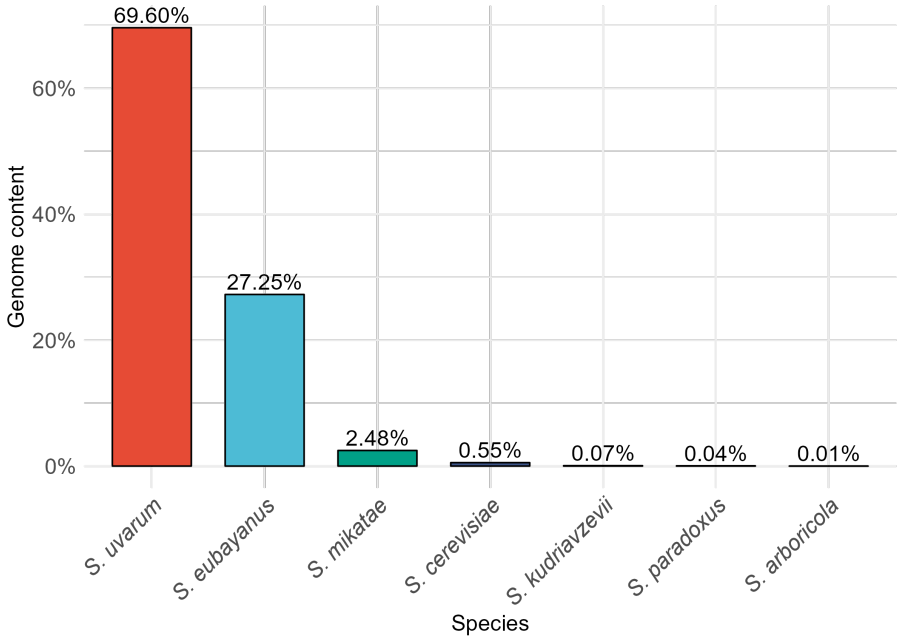


Supplemental Figure S2: Flow cytometry analysis supports a diploid genome of *S. bayanus* CBS380. The top histogram shows cell count of *S. uvarum* YJF1450 and the bottom histogram is CBS380. The x-axis indicates the amount of DNA that is stained by propidium iodide. The green line shows the DNA amount of the G1 phase of the YJF1450 cells (one copy of haploid genome, 1C). The red line shows the G2 phase of the YJF1450 (two copies of haploid genome, 2C) and the G1 phase of the CBS380 (one copy of diploid genome). The blue line indicates the G2 phase of CBS380 (two copies of diploid genomes, 4C).

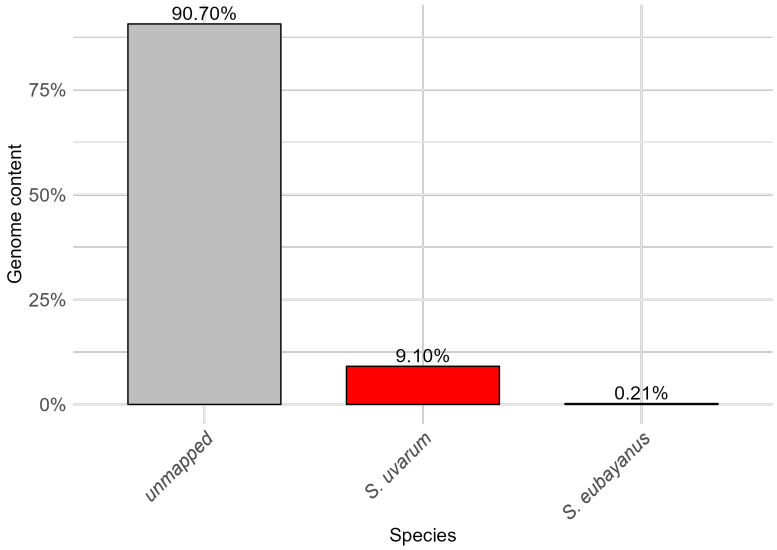


	Passing basecaller QC	After Porechop
Sequences	192,999	192,837
Bases	2,052,301,802 bp	2,036,617,616 bp
Mean length	10,634 bp	10,561 bp
Median length	6,016 bp	5,949 bp
N50	21,807 bp	21,844 bp
Longest sequence	158,341 bp	158,255 bp

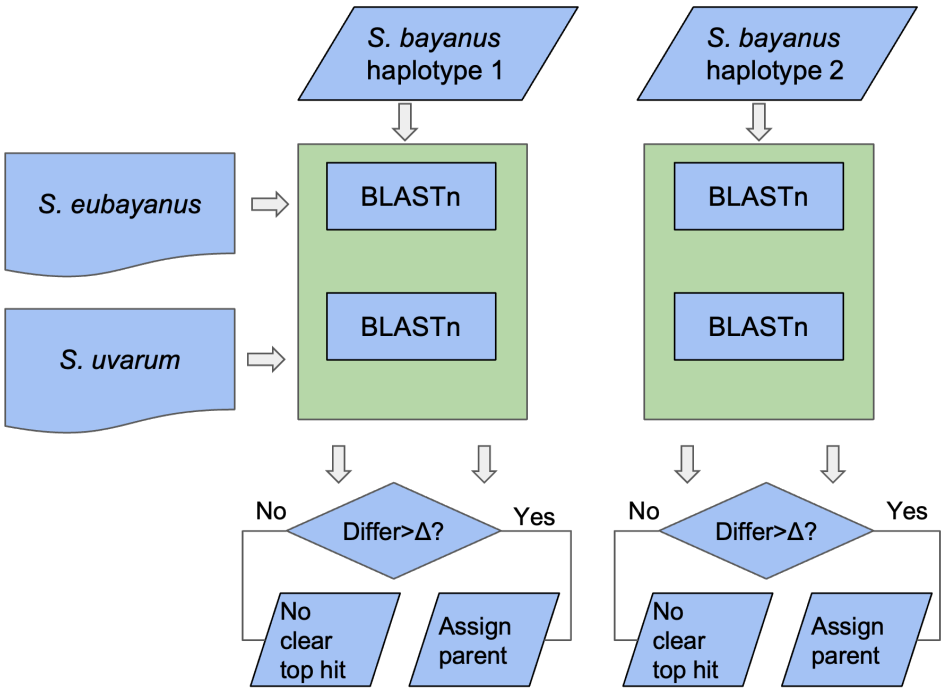
Supplemental Figure S3: **Overview of *S. bayanus* CBS 380 genome sequencing results using Oxford Nanopore's MinION.** The sequencing effort generated 2.2 Gb of data, achieving approximately 170-fold coverage. After quality control, 2.04 Gb of data were retained. Notably, the sequencing run produced 100 reads surpassing 100 Kb, with the longest read measuring 158,255 bp.



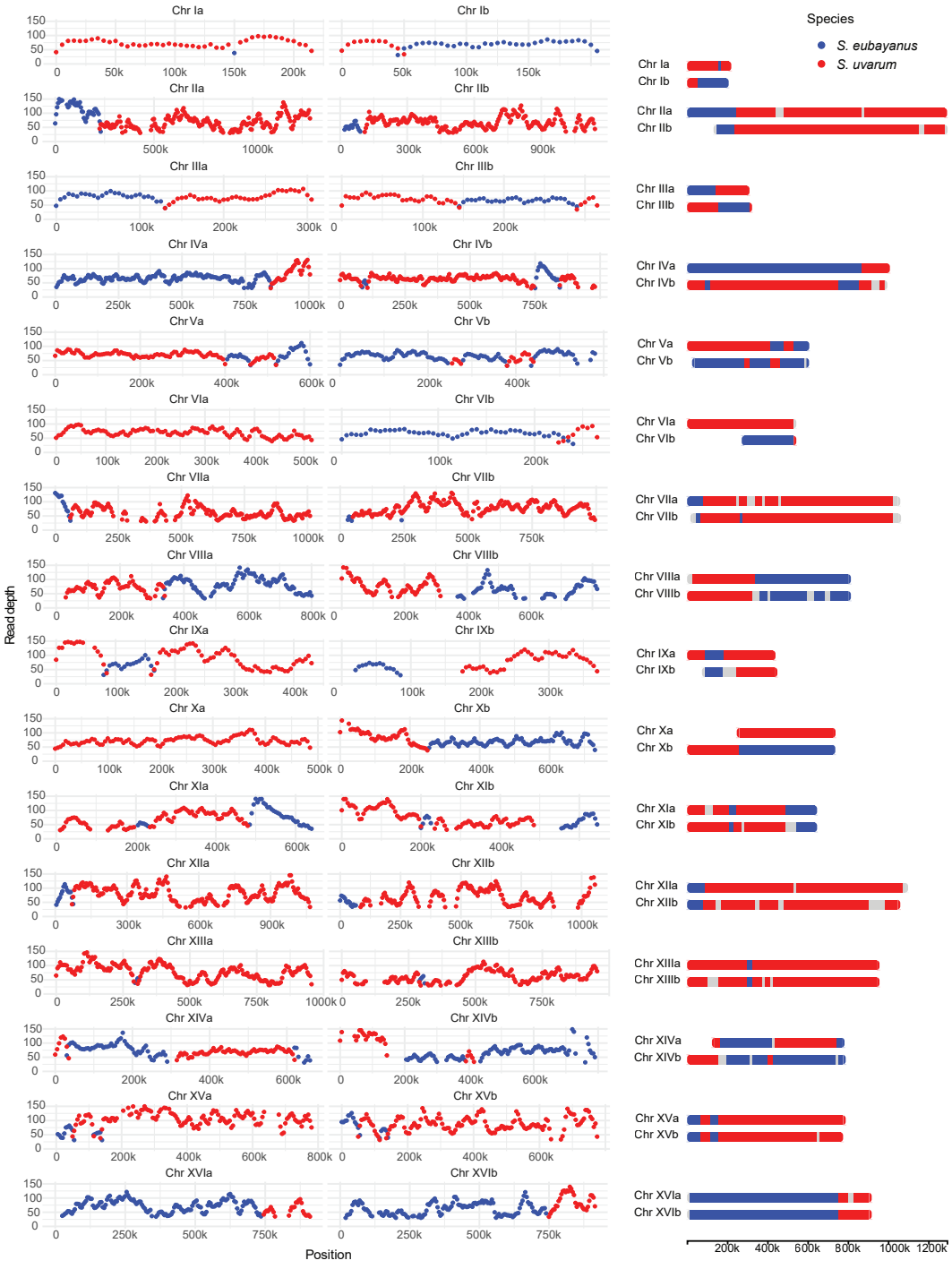
Supplemental Figure S4: Percentages of Nanopore reads mapped to different reference genomes. Nanopore reads were mapped to combination genomes of seven *Saccharomyces* species**.** The percentages of genomic contribution of each potential parental species were inferred by sppIDer.



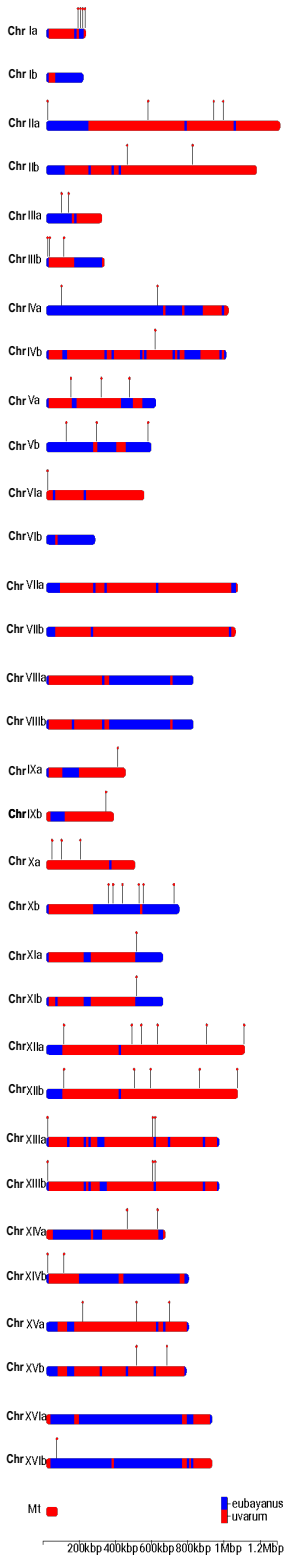
Supplemental Figure S5: Percentages of MinION reads mapped to different mitochondrial genomes. A total of 9.31% of all raw MinION reads were mapped to combined mitochondrial genomes of *S. uvarum* and *S. eubayanus*. Among them, 9.10% of reads were mapped to *S. uvarum* mitochondrial genome and only 0.21% of reads were mapped to *S. eubayanus* mitochondrial genome, suggesting that the mitochondrial genome of CBS380 was inherited from *S. uvarum*.



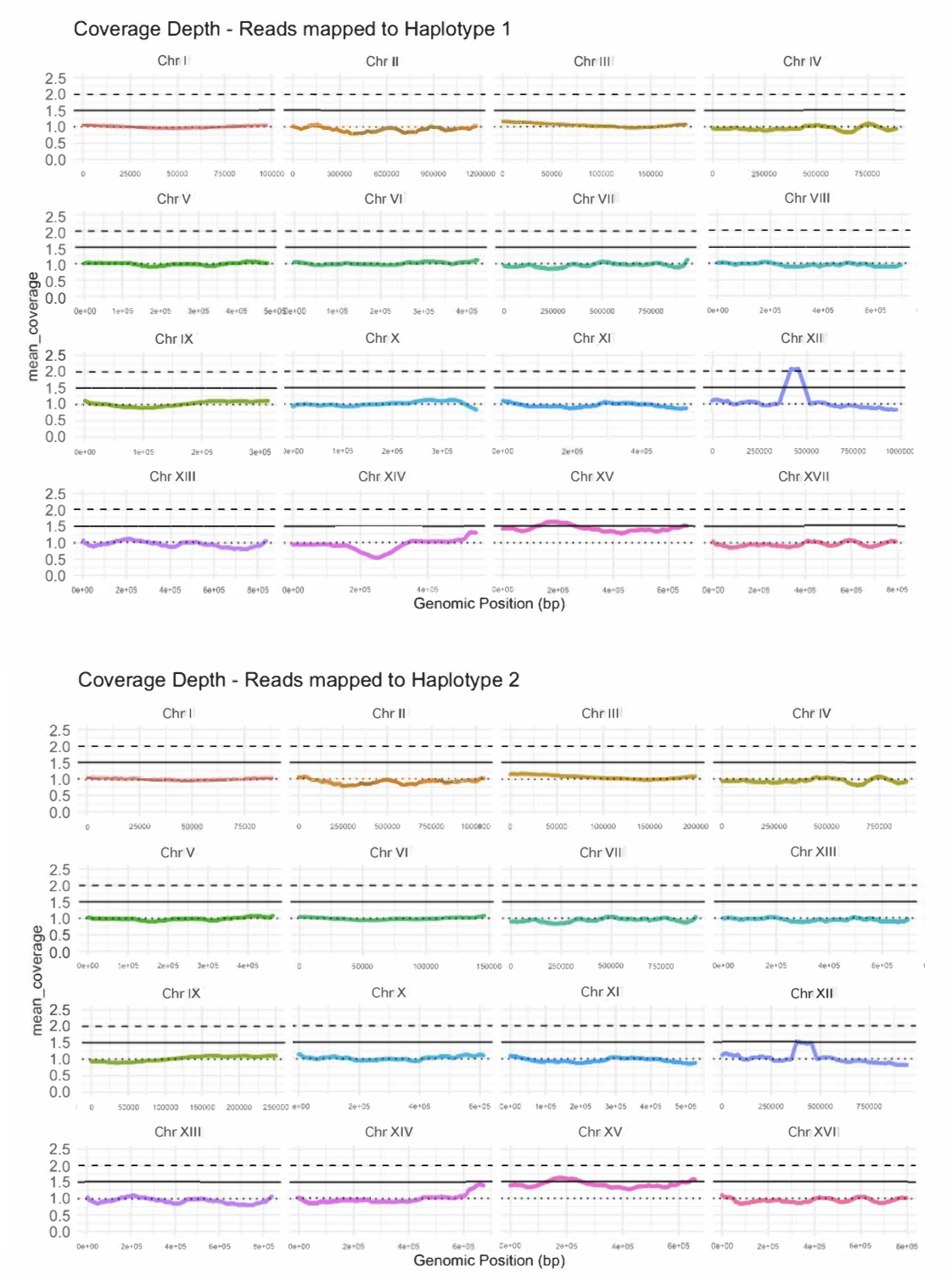
Supplemental Figure S6: Method for inferring parental origin of hybrid sequences using Local BLAST Analysis. Hybrid assembly sequences from *S. bayanus* haplotypes a and b were processed in 5k non-overlapping windows and subjected to BLASTn searches against the parental genomes of *S. eubayanus* and *S. uvarum*. Parental assignments were made based on the highest bitscore.



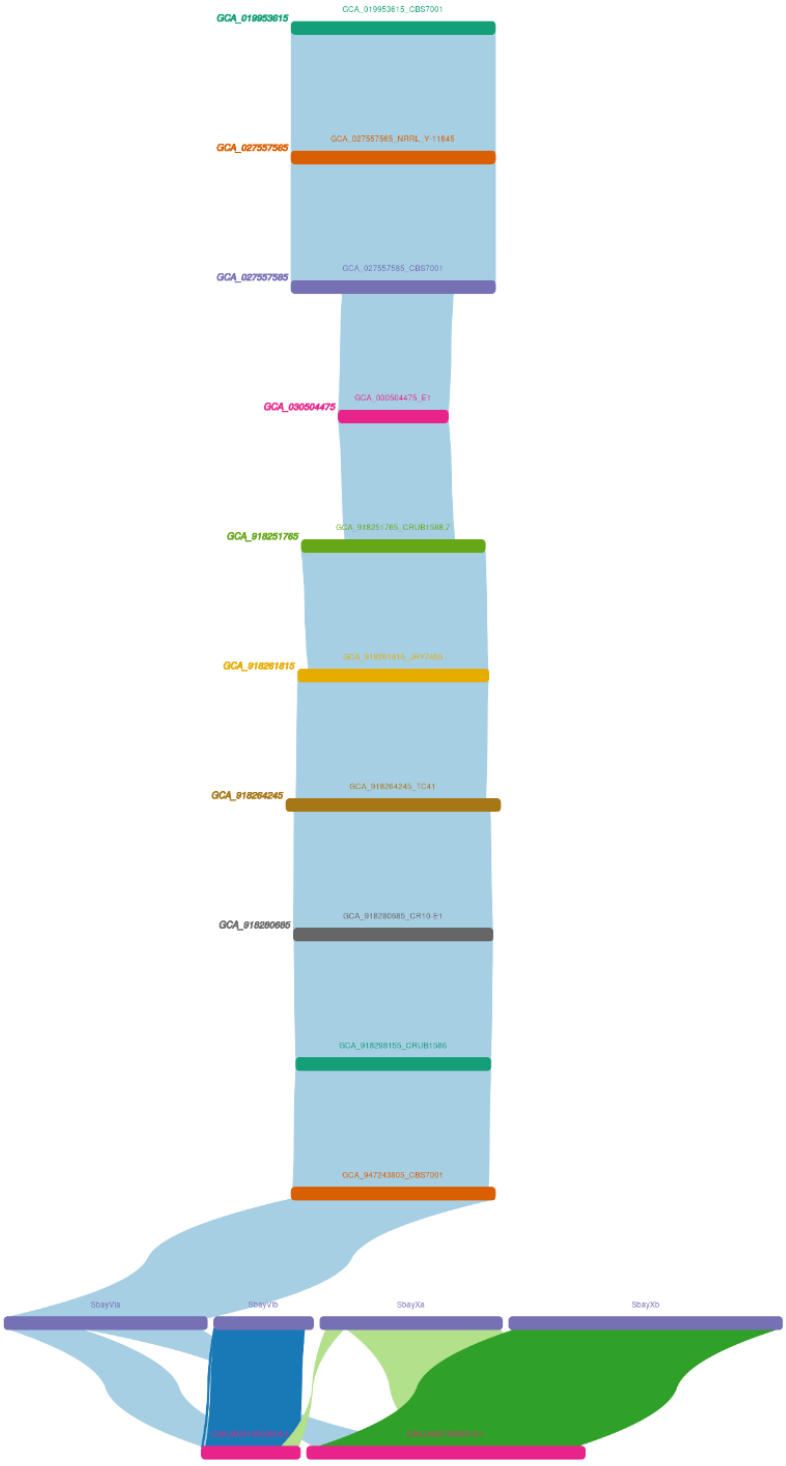
Supplemental Figure S7: Inference of genomic origin based on re-mapping of reads assigned to *S. uvarum* and *S. eubayanus*. The left panels show the distributions of read depth on each chromosome. Each row shows a pair of homologous chromosomes. A dot represents a non-overlapping 5 kb window. Dots in red indicate regions supported by reads assigned to *S. uvarum* and blue dots represent regions supported by reads assigned to *S. eubayanus*. The right panel is a schematic illustration showing the origin of genomic regions in each chromosome based on results shown in the left panel.



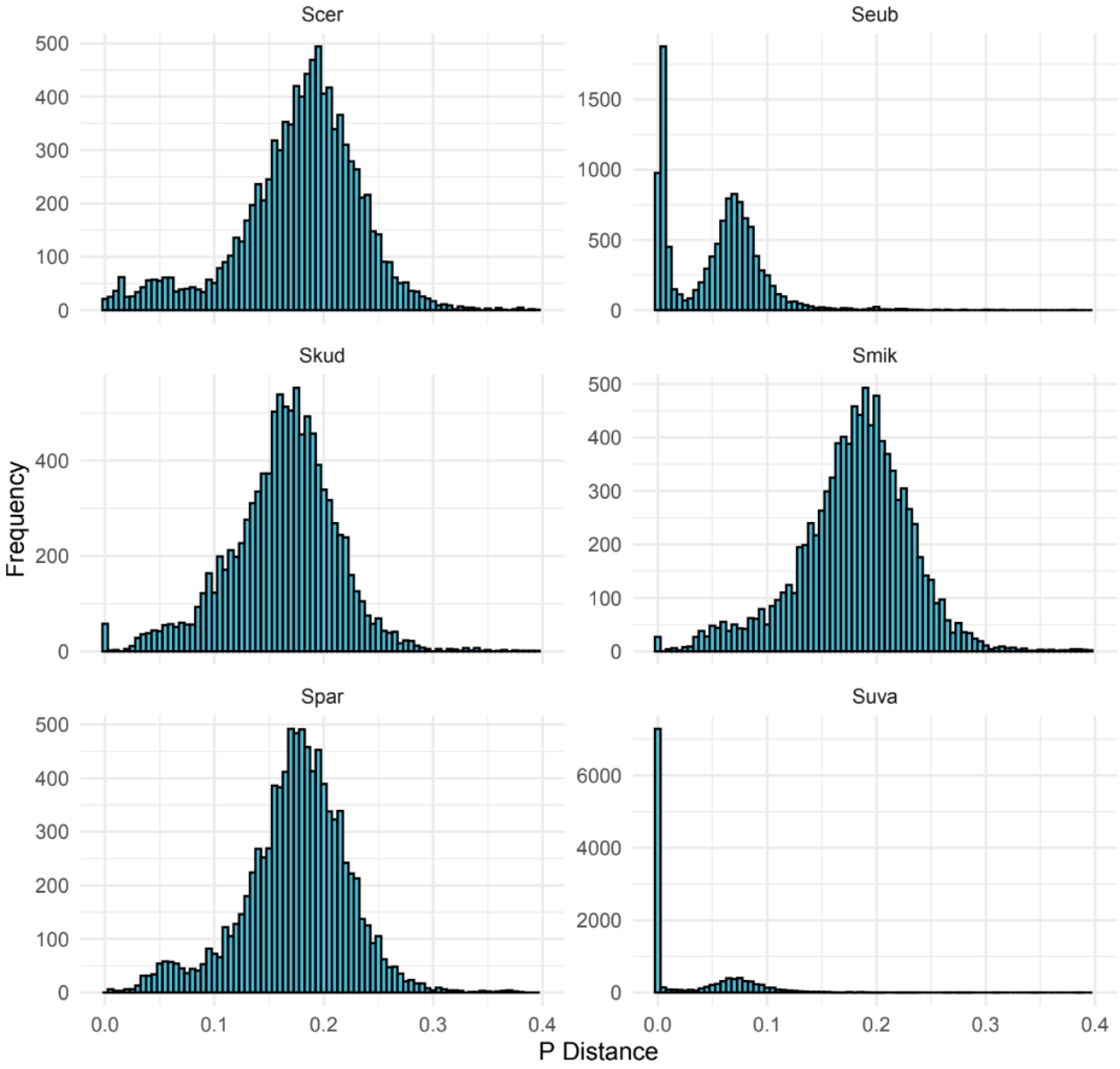
Supplemental Figure S8: **Distribution of LTR and breakpoints of major recombination events.** LTR retrotransposons are significantly enriched within 5 kb regions flanking the breakpoints of recombination in both subgenomes (Fisher's Exact Test: p-value < 2.2e-16 for Subgenome a and p = 7.5e-05 for Subgenome b), suggesting LTR may have facilitated the recombination between parental subgenomes.



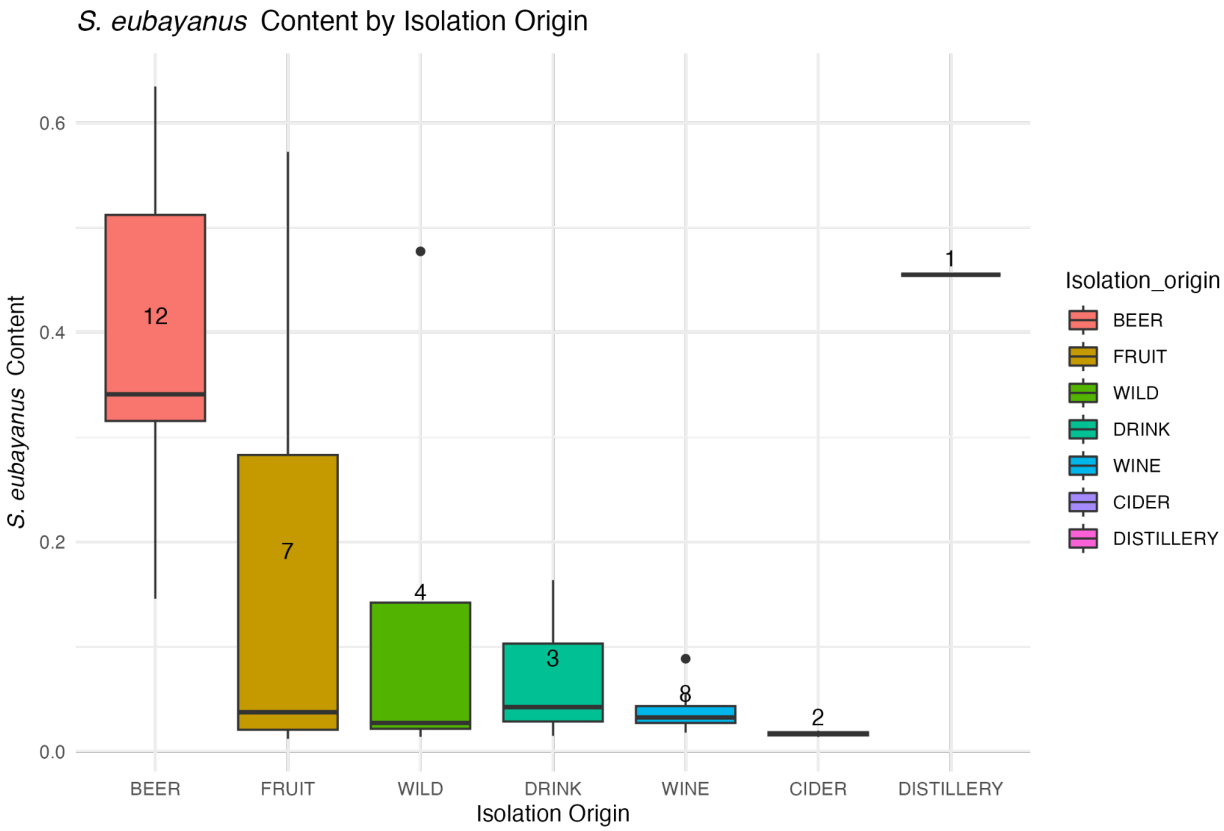
Supplemental Figure S9: Identification of chromosome aneuploidy based on normalized read depth. Y axis indicates the normalized sequencing depth by genome average for each chromosome. Our analysis shows that, while most of the genome maintains a diploid state, Chr XV shows a coverage pattern consistent with triploidy. The increased coverage observed in a region of Chr XII is consistent and expected due to the presence of ribosomal DNA.



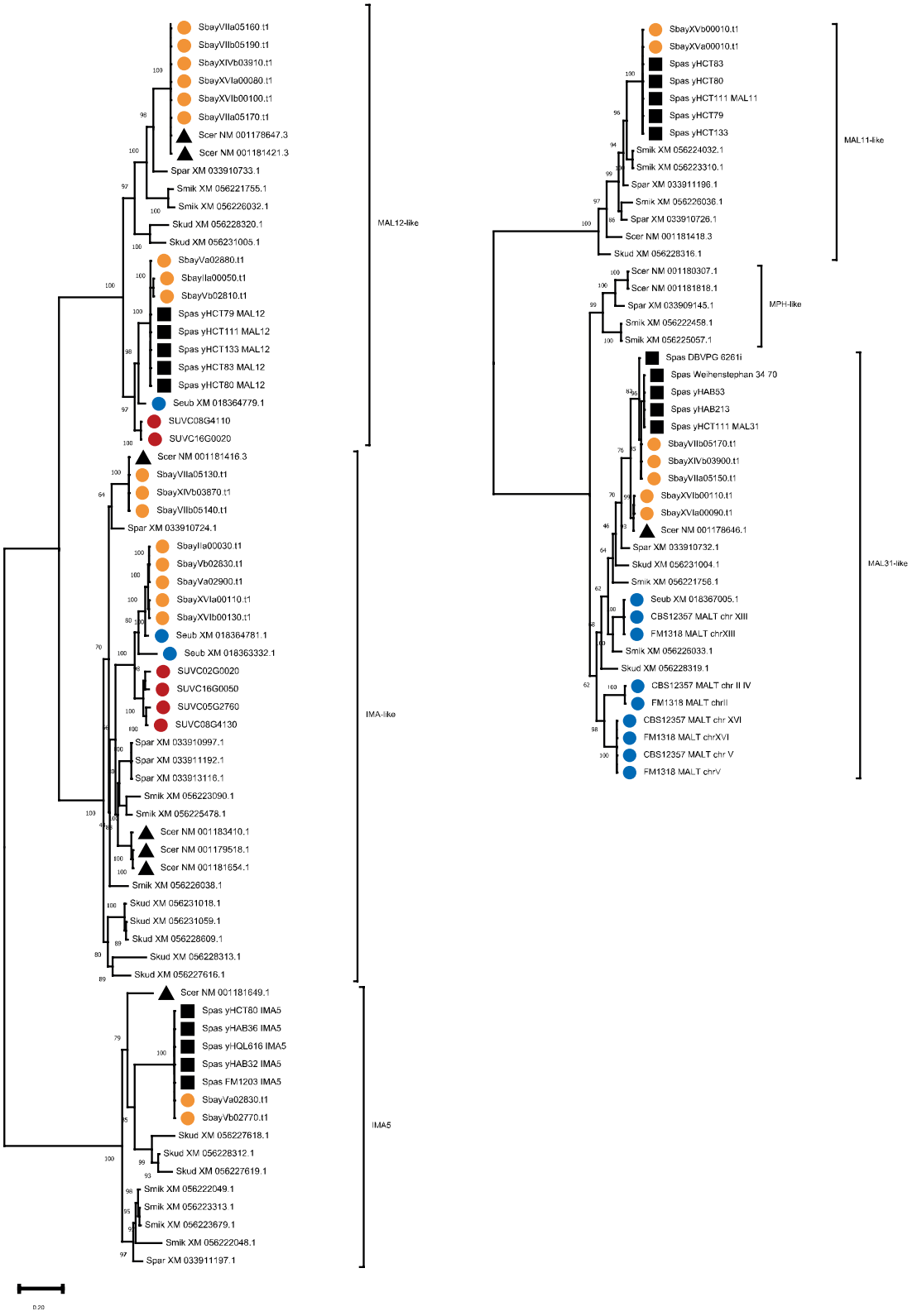
Supplemental Figure S10: **Chromosomal collinearity analysis for Chr VI and Chr X in different *S. uvarum* strains.**



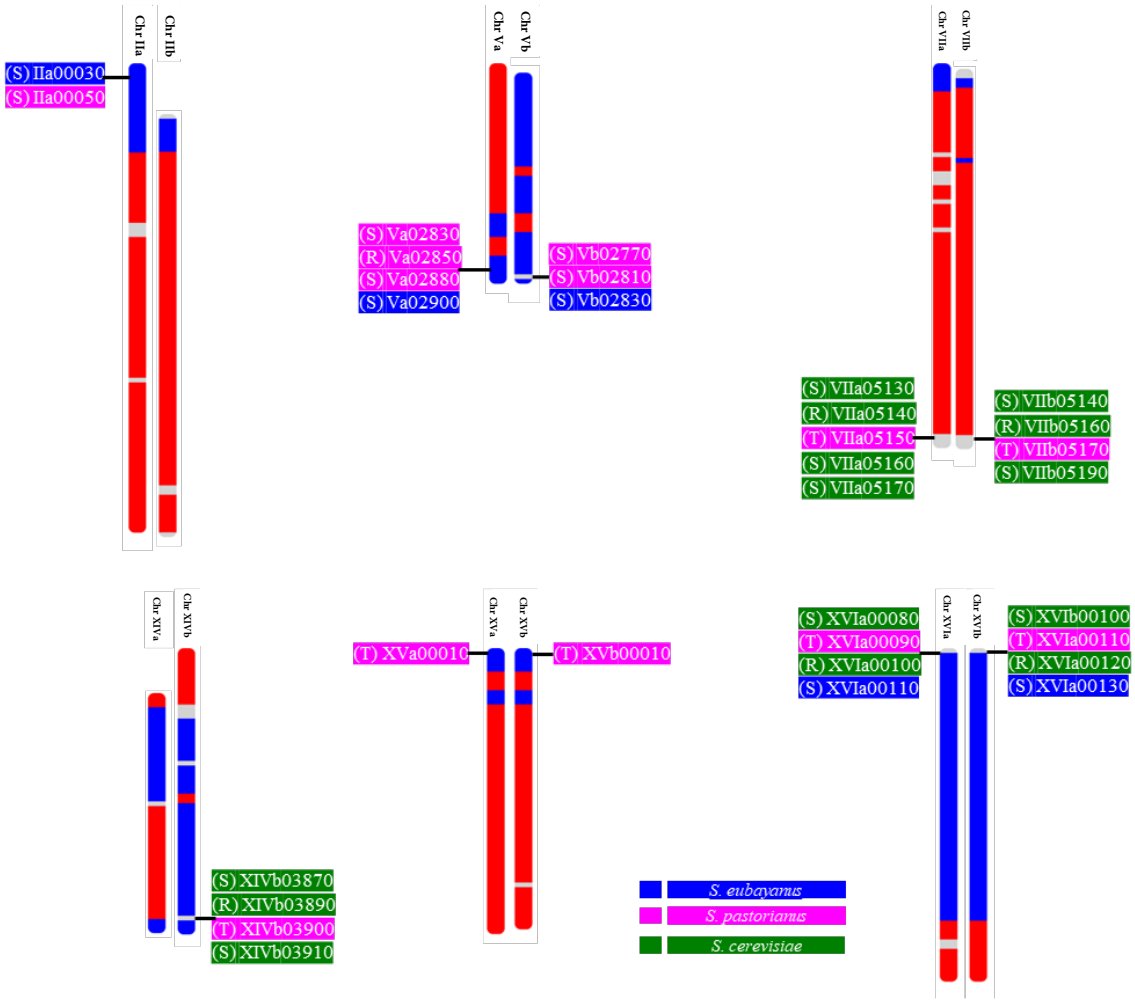
Supplemental Figure S11: **Histogram showing distribution of sequence divergence levels between *S. bayanus* genes and their orthologous genes in six *Saccharomyces* species.** Sequence divergence levels were calculated as proportion of different nucleotides between two sequences in coding region.



Supplemental Figure S12: **Boxplot showing the distribution of genomic content originated from *S. eubayanus*.** The genomic content data was obtained from Langdon et al (2019). Strains with unknown isolation origins were excluded. It was observed that strains directly related to beer have a higher contribution to *S. eubayanus'* nuclear genome content compared to other environments. The number in each boxplot indicates the number of *S. bayanus* genomes examined in each group.



Supplemental Figure S13: Phylogenetic trees of the *MALS* (A) and *MALT* (B) gene families. Both trees were inferred using the maximum likelihood method based on nucleotide sequences with 1000 bootstrap and reported as percentages on each node. The trees are drawn to scale, with branch lengths denoting the genetic distance.



Supplemental Figure S14: A chromosome map showing locations of members of the three *MAL* gene families in the *S. bayanus* genome. The origin of each gene based on phylogenetic inferences was indicated by different colors.

Supplemental Table S1: Assembly statistics of the *S. bayanus* CBS 380 genome generated by different bioinformatic tools.

Assembler	Flye	NextDeNovo	WTDBG	Ra	NECAT	SmartDeNovo	Canu	SmartDeNovo*	Flye*
Length (bp)	14,635,841	11,870,454	11,811,864	12,017,687	15,428,960	11,578,645	16,115,554	12,391,933	13,538,319
Total seq. #	38	17	21	24	32	16	49	21	62
Max seq.	1,286,589	1,293,072	2,164,800	1,101,538	1,293,344	1,293,262	1,292,830	1,101,177	1,285,351
Min seq.	2,346	122,189	17,265	8,314	59,794	61,853	10,281	59,343	2,074
Med seq len.	301,095	773,390	462,238	528,214	442,886	780,948	213,365	610,360	84,583
N50 (length)	647,014	816,895	813,059	785,083	646,646	919,172	612,488	787,374	785,325
N50 (ctg #)	9	6	5	7	9	6	9	7	7
BUSCO	C:96.6%[S:74.5%,D:22.1%],F:2.3%,M:1.1%,n:2137	C:93.0%[S:90.5%,D:2.5%],F:2.9%,M:4.1%,n:2137	C:90.0%[S:88.3%,D:1.7%],F:6.4%,M:3.6%,n:2137	C:96.5%[S:94.3%,D:2.2%],F:2.2%,M:1.3%,n:2137	C:97.7%[S:74.9%,D:22.8%],F:1.7%,M:0.6%,n:2137	C:91.6%[S:89.1%,D:2.5%],F:2.5%,M:5.9%,n:2137	C:96.6%[S:73.8%,D:22.8%],F:2.3%,M:1.1%,n:2137	C:93.4%[S:89.2%,D:4.2%],F:4.4%,M:2.2%,n:2137	C:96.1%[S:86.3%,D:9.8%],F:2.7%,M:1.2%,n:2137

Notes: This table summarizes the performance of eight different assembly algorithms, including Flye, NextDeNovo, WTDBG, Ra, NECAT, SmartDeNovo, Canu, and their optimized versions (denoted by an asterisk), with an emphasis on addressing diploidy complexities.

Supplemental Table S2: A list of assembled sequence lengths for both haplotypes in *S. bayanus*.

Chromosome	Haplotype-a	Haplotype-b
Chr I	217,795	208,383
Chr II	1,292,201	1,163,801
Chr III	306,526	319,184
Chr IV	1,007,622	1,000,830
Chr V	601,693	581,353
Chr VI	544,124	266,821
Chr VII	1,055,917	1,042,311
Chr VIII	817,384	817,171
Chr IX	433,831	370,629
Chr X	488,404	732,591
Chr XI	646,725	646,334
Chr XII	1,096,630	1,057,074
Chr XIII	959,152	958,029
Chr XIV	664,155	791,919
Chr XV	784,543	778,848
Chr XVI	912,922	919,249
Mt	64,655	

Supplemental Table S3: A list of telomere repeat sequences identified in *S. bayanus* chromosomes.

Chr. Name	side	type	start	end	Length (bp)	note
Chr Ia	Left	term	1	176	176	
Chr Ia	Right	term	217757	217795	39	
Chr Ib	Left	term	1	100	100	
Chr Ib	Right	term	208276	208383	108	
Chr IIa	Left	term	1	6	6	motif search
Chr IIa	Right	term	1292098	1292201	104	
Chr IIb	Left	term	1163744	1163801	58	motif search
Chr IIb	Right	term	1163741	1163801	61	
Chr IIIa	Left	term	1	37	37	
Chr IIIa	Right	term	306374	306526	153	
Chr IIIb	Left	term	1	11	11	motif search
Chr IIIb	Right	term	319135	319184	50	
Chr IVa	Left	term	1	22	22	motif search
Chr IVa	Right	term	1007512	1007622	111	
Chr IVb	Left	term	1	106	106	
Chr IVb	Right	term	1000737	1000830	94	
Chr Va	Left	term	1	128	128	
Chr Va	Right	term				
Chr Vb	Left	term	1	38	38	
Chr Vb	Right	term	581264	581353	90	
Chr VIa	Left	term				
Chr VIa	Right	term				
Chr VIb	Left	term	1	100	100	
Chr VIb	Right	term				
Chr VIIa	Left	term				
Chr VIIa	Right	term	1055895	1055917	23	
Chr VIIb	Left	term				
Chr VIIb	Right	term				
Chr VIIIa	Left	term	1	71	71	
Chr VIIIa	Right	term	817320	817384	65	
Chr VIIIb	Left	term	1	148	148	
Chr VIIIb	Right	term				
Chr IXa	Left	term	1	303	303	
Chr IXa	Right	term	433664	433831	168	
Chr IXa	Left	term				

Chr IXa	Right	term	370577	370629	53	
Chr Xa	Left	term	1	82	82	
Chr Xa	Right	term	488284	488404	121	
Chr Xb	Left	term	1	42	42	
Chr Xb	Right	term	732554	732591	38	
Chr XIa	Left	term				
Chr XIa	Right	intern	646705	646724	20	
Chr XIb	Left	term	1	124	124	
Chr XIb	Right	term	646254	646334	81	
Chr XIIa	Left	term	1	115	115	
Chr XIIa	Right	term	1096471	1096630	160	
Chr XIIb	Left	term	1	158	158	
Chr XIIb	Right	term	1056943	1057074	132	
Chr XIIIa	Left	term	1	46	46	
Chr XIIIa	Right	term	959061	959152	92	
Chr XIIIb	Left	term	1	95	95	
Chr XIIIb	Right	term	957913	958029	117	
Chr XIVa	Left	term				
Chr XIVa	Right	intern	659515	659534	20	
Chr XIVb	Left	term				
Chr XIVb	Right	term	791893	791919	27	
Chr XVa	Left	intern	8274	8293	20	
Chr XVa	Right	term	784500	784543	44	
Chr XVb	Left	term	1	143	143	
Chr XVb	Right	term				
Chr XVIa	Left	term	1	23	23	
Chr XVIa	Right	term	912857	912922	66	
Chr XVIb	Left	term				
Chr XVIb	Right	term	918807	919249	443	

Supplemental Table S4: A list of proportions of parental genome content in each *S. bayanus* chromosome estimated by two different methods.

	BLAST-based		SppIDer_based	
Chromosome	***S. uvarum* (%)**	***S. eubayanus* (%)**	***S. uvarum* (%)**	***S. eubayanus* (%)**
Chr Ia	86	13	97.78	2.22
Chr Ib	23	76	25	75
Chr IIa	79	20	80.42	19.58
Chr IIb	86	13	92.86	7.14
Chr IIIa	47	52	57.14	42.86
Chr IIIb	47	52	54.55	45.45
Chr IVa	16	83	16	84
Chr IVb	77	22	87.1	12.9
Chr Va	68	31	76.42	23.58
Chr Vb	14	85	15.79	84.21
Chr VIa	95	4	100	0
Chr VIb	3	96	15.52	84.48
Chr VIIa	89	10	92.61	7.39
Chr VIIb	92	7	97.41	2.59
Chr VIIIa	40	59	40.13	59.87
Chr VIIIb	39	60	48.41	51.59
Chr IXa	72	27	80.22	19.78
Chr IXb	73	26	74.07	25.93
Chr Xa	92	7	100	0
Chr Xb	33	66	34.69	65.31
Chr XIa	68	31	67.8	32.2
Chr XIb	66	33	77.88	22.12
Chr XIIa	89	10	92.65	7.35
Chr XIIb	87	12	91.52	8.48
Chr XIIIa	88	11	97.94	2.06
Chr XIIIb	87	12	97.02	2.98
Chr XIVa	53	46	53.85	46.15
Chr XIVb	27	72	26.67	73.33
Chr XVa	83	16	89.44	10.56
Chr XVb	83	16	88.46	11.54
Chr XVIa	17	82	15.06	84.94
Chr XVIb	19	80	19.88	80.12

Supplemental Table S5: Breakpoints of major recombination events are supported by sequencing reads.

Chromosome	Position	# of Supporting Reads	Total Mapped Reads	Ratio of supporting reads	Classification
Chr Ib	67484	81	81	1.0	High support
Chr IIa	256497	58	58	1.0	High support
Chr IIb	497484	18	18	1.0	High support
Chr IIb	1158355	15	15	1.0	High support
Chr IIb	1162798	34	34	1.0	High support
Chr IIb	1163226	34	34	1.0	High support
Chr IIb	1163523	35	36	0.972	High support
Chr IVa	480694	56	56	1.0	High support
Chr Va	455951	52	56	0.929	High support
Chr Va	455705	52	52	1.0	High support
Chr Va	451600	55	55	1.0	High support
Chr Va	454049	56	56	1.0	High support
Chr Va	451426	56	56	1.0	High support
Chr Va	454466	53	53	1.0	High support
Chr Vb	579009	31	32	0.969	High support
Chr Vb	1355	25	25	1.0	High support
Chr Vb	1987	26	26	1.0	High support
Chr Vb	7249	51	51	1.0	High support
Chr Vb	434146	94	95	0.989	High support
Chr Vb	435276	88	88	1.0	High support
Chr Vb	435639	86	86	1.0	High support
Chr Vb	561796	56	64	0.875	Moderate support
Chr Vb	564281	63	63	1.0	High support
Chr Vb	580014	31	32	0.969	High support
Chr Vb	433784	94	94	1.0	High support
Chr Vb	431733	87	88	0.989	High support
Chr Vb	580796	31	31	1.0	High support
Chr VIa	538007	13	14	0.929	High support
Chr VIa	538672	14	15	0.933	High support
Chr VIIa	1	5	5	1.0	High support
Chr Xa	134000	66	68	0.971	High support
Chr Xa	423601	81	81	1.0	High support
Chr XIIa	1091018	14	14	1.0	High support
Chr XIIa	526512	51	51	1.0	High support
Chr XIIa	1095257	14	14	1.0	High support
Chr XIIa	1091578	14	14	1.0	High support

Chr XIIa	1092207	14	14	1.0	High support
Chr XIIa	1093765	14	14	1.0	High support
Chr XIIa	1094267	13	13	1.0	High support
Chr XIIa	1094668	13	13	1.0	High support
Chr XIIa	1096175	13	13	1.0	High support
Chr XIIa	528623	34	34	1.0	High support
Chr XIIb	492179	58	145	0.4	Low support
Chr XIVa	636829	62	62	1.0	High support
Chr XIVa	638390	66	66	1.0	High support
Chr XIVa	638028	65	65	1.0	High support
Chr XIVa	637204	63	63	1.0	High support
Chr XIVa	636567	70	74	0.946	High support
Chr XIVa	636567	70	74	0.946	High support
Chr XIVa	618729	84	85	0.988	High support
Chr XIVa	213985	45	45	1.0	High support
Chr XIVb	715097	68	68	1.0	High support
Chr XIVb	714348	69	69	1.0	High support
Chr XVb	157665	73	73	1.0	High support
Chr XVIa	841998	26	26	1.0	High support

Supplemental Table S6: A list of enriched GO terms "orphan genes" in *S. bayanus* CBS380.

Gene Ontology term	Cluster frequency	Genome frequency	Corrected P-value	Genes annotated to the term
ATP hydrolysis activity	13 of 53 genes, 24.5%	260 of 7166 genes, 3.6%	0.00000357	YIL013C, YER172C, YLR419W, YER013W, YGR281W, YML061C, YMR128W, YGL150C, YLL048C, YLR106C, YNL250W, YKR054C, YGR271W
ribonucleoside triphosphate phosphatase activity	14 of 53 genes, 26.4%	349 of 7166 genes, 4.9%	0.0000153	YIL013C, YER172C, YLR419W, YER013W, YGR281W, YML061C, YGL150C, YMR128W, YBR179C, YGR271W , YLR106C, YLL048C, YKR054C, YNL250W
ATP-dependent activity	14 of 53 genes, 26.4%	356 of 7166 genes, 5.0%	0.0000196	YML061C, YLR424W, YGR281W, YLR419W, YER013W , YIL013C, YER172C, YGR271W, YKR054C, YNL250W, YLL048C, YLR106C, YMR128W, YGL150C
pyrophosphatase activity	14 of 53 genes, 26.4%	385 of 7166 genes, 5.4%	0.0000514	YGL150C, YMR128W, YBR179C, YKR054C, YNL250W, YLL048C, YLR106C, YGR271W, YER013W, YLR419W, YER172C, YIL013C, YML061C, YGR281W
hydrolase activity, acting on acid anhydrides	14 of 53 genes, 26.4%	387 of 7166 genes, 5.4%	0.0000548	YIL013C, YER172C, YLR419W, YER013W, YGR281W, YML061C, YMR128W, YBR179C, YGL150C, YGR271W , YLL048C, YLR106C, YNL250W, YKR054C
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	14 of 53 genes, 26.4%	387 of 7166 genes, 5.4%	0.0000548	YGR281W, YML061C, YIL013C, YER172C, YLR419W, YER013W, YGR271W, YLR106C, YLL048C, YKR054C, YNL250W, YMR128W, YGL150C, YBR179C
RNA helicase activity	6 of 53 genes, 11.3%	46 of 7166 genes, 0.6%	0.0001	YER172C, YMR128W, YER013W, YLR419W, YLR424W , YGR271W
ATP-dependent activity, acting on RNA	6 of 53 genes, 11.3%	49 of 7166 genes, 0.7%	0.00014	YMR128W, YER172C, YLR419W, YER013W, YLR424W , YGR271W
helicase activity	8 of 53 genes, 15.1%	113 of 7166 genes, 1.6%	0.00016	YER013W, YLR419W, YER172C, YML061C, YLR424W, YMR128W, YGL150C, YGR271W
nucleotide binding	20 of 53 genes, 37.7%	944 of 7166 genes, 13.2%	0.00062	YML061C, YBL098W, YGR005C, YOL100W, YKR054C, YLR096W, YGL150C, YBR179C, YGR281W, YER172C, YIL013C, YLR419W, YKL203C, YER013W, YLR305C, YLR106C, YLL048C, YNL250W, YGR271W, YMR128W

nucleoside phosphate binding	20 of 53 genes, 37.7%	944 of 7166 genes, 13.2%	0.00062	YER172C, YIL013C, YER013W, YLR419W, YKL203C, YGR281W, YMR128W, YLL048C, YLR305C, YLR106C, YNL250W, YGR271W, YOL100W, YML061C, YGR005C, YBL098W, YLR096W, YGL150C, YBR179C, YKR054C
ATP binding	17 of 53 genes, 32.1%	719 of 7166 genes, 10.0%	0.00097	YGL150C, YLR096W, YKR054C, YOL100W, YML061C, YMR128W, YGR271W, YNL250W, YLR305C, YLR106C, YLL048C, YLR419W, YER013W, YKL203C, YIL013C, YER172C, YGR281W
adenyl ribonucleotide binding	17 of 53 genes, 32.1%	723 of 7166 genes, 10.1%	0.00105	YGR281W, YIL013C, YER172C, YLR419W, YER013W, YKL203C, YLL048C, YLR305C, YLR106C, YNL250W, YGR271W, YMR128W, YML061C, YOL100W, YKR054C, YLR096W, YGL150C
heterocyclic compound binding	20 of 53 genes, 37.7%	994 of 7166 genes, 13.9%	0.00137	YKR054C, YBR179C, YGL150C, YLR096W, YGR005C, YBL098W, YML061C, YOL100W, YGR271W, YNL250W, YLR305C, YLL048C, YLR106C, YMR128W, YGR281W, YLR419W, YER013W, YKL203C, YIL013C, YER172C
purine ribonucleoside triphosphate binding	18 of 53 genes, 34.0%	822 of 7166 genes, 11.5%	0.00141	YML061C, YOL100W, YKR054C, YBR179C, YGL150C, YLR096W, YGR281W, YLR419W, YER013W, YKL203C, YIL013C, YER172C, YGR271W, YNL250W, YLL048C, YLR305C, YLR106C, YMR128W
purine ribonucleotide binding	18 of 53 genes, 34.0%	826 of 7166 genes, 11.5%	0.00151	YIL013C, YER172C, YER013W, YLR419W, YKL203C, YGR281W, YMR128W, YLR305C, YLL048C, YLR106C, YNL250W, YGR271W, YOL100W, YML061C, YLR096W, YBR179C, YGL150C, YKR054C
ribonucleotide binding	18 of 53 genes, 34.0%	846 of 7166 genes, 11.8%	0.00211	YOL100W, YML061C, YLR096W, YBR179C, YGL150C, YKR054C, YIL013C, YER172C, YKL203C, YLR419W, YER013W, YGR281W, YMR128W, YLL048C, YLR305C, YLR106C, YNL250W, YGR271W
adenyl nucleotide binding	17 of 53 genes, 32.1%	765 of 7166 genes, 10.7%	0.00223	YER172C, YIL013C, YLR419W, YKL203C, YER013W, YGR281W, YMR128W, YLR305C, YLL048C, YLR106C, YNL250W, YGR271W, YOL100W, YML061C, YLR096W, YGL150C, YKR054C
carbohydrate derivative binding	18 of 53 genes, 34.0%	860 of 7166 genes, 12.0%	0.00265	YKR054C, YGL150C, YBR179C, YLR096W, YML061C, YOL100W, YGR271W, YNL250W, YLL048C, YLR305C, YLR106C, YMR128W, YGR281W, YLR419W, YER013W, YKL203C, YIL013C, YER172C
purine nucleotide binding	18 of 53 genes, 34.0%	868 of 7166 genes, 12.1%	0.003	YLR305C, YLR106C, YLL048C, YNL250W, YGR271W, YMR128W, YGR281W, YIL013C, YER172C, YKL203C, YLR419W, YER013W, YKR054C, YLR096W, YGL150C, YBR179C, YML061C, YOL100W

anion binding	19 of 53 genes, 35.8%	974 of 7166 genes, 13.6%	0.00391	YMR128W, YGR271W, YNL250W, YLR305C, YLL048C , YLR106C, YLR419W, YER013W, YKL203C, YIL013C, YER172C, YGR281W, YBR179C, YGL150C, YLR096W, YKR054C, YOL100W, YBL098W, YML061C
voltage-gated monoatomic cation channel activity	2 of 53 genes, 3.8%	2 of 7166 genes, 0.0%	0.00595	YJL093C, YGR217W

Gene Ontology term	Cluster frequency	Genome frequency	Corrected P-value	Genes annotated to the term
ATP hydrolysis activity	13 of 53 genes, 24.5%	260 of 7166 genes, 3.6%	0.00000357	YIL013C, YER172C, YLR419W, YER013W, YGR281W, YML0 61C, YMR128W, YGL150C, YLL048C, YLR106C, YNL250W, YKR054C, YGR271W
ribonucleoside triphosphate phosphatase activity	14 of 53 genes, 26.4%	349 of 7166 genes, 4.9%	0.0000153	YIL013C, YER172C, YLR419W, YER013W, YGR281W, YML0 61C, YGL150C, YMR128W, YBR179C, YGR271W, YLR106C, YLL048C, YKR054C, YNL250W
ATP-dependent activity	14 of 53 genes, 26.4%	356 of 7166 genes, 5.0%	0.0000196	YML061C, YLR424W, YGR281W, YLR419W, YER013W, YIL0 13C, YER172C, YGR271W, YKR054C, YNL250W, YLL048C, YLR106C, YMR128W, YGL150C
pyrophosphatas e activity	14 of 53 genes, 26.4%	385 of 7166 genes, 5.4%	0.0000514	YGL150C, YMR128W, YBR179C, YKR054C, YNL250W, YLL0 48C, YLR106C, YGR271W, YER013W, YLR419W, YER172C, YIL013C, YML061C, YGR281W
hydrolase activity, acting on acid anhydrides	14 of 53 genes, 26.4%	387 of 7166 genes, 5.4%	0.0000548	YIL013C, YER172C, YLR419W, YER013W, YGR281W, YML0 61C, YMR128W, YBR179C, YGL150C, YGR271W, YLL048C, YLR106C, YNL250W, YKR054C

hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	14 of 53 genes, 26.4%	387 of 7166 genes, 5.4%	0.0000548	YGR281W, YML061C, YIL013C, YER172C, YLR419W, YER013W, YGR271W, YLR106C, YLL048C, YKR054C, YNL250W, YMR128W, YGL150C, YBR179C
RNA helicase activity	6 of 53 genes, 11.3%	46 of 7166 genes, 0.6%	0.0001	YER172C, YMR128W, YER013W, YLR419W, YLR424W, YGR271W
ATP-dependent activity, acting on RNA	6 of 53 genes, 11.3%	49 of 7166 genes, 0.7%	0.00014	YMR128W, YER172C, YLR419W, YER013W, YLR424W, YGR271W
helicase activity	8 of 53 genes, 15.1%	113 of 7166 genes, 1.6%	0.00016	YER013W, YLR419W, YER172C, YML061C, YLR424W, YMR128W, YGL150C, YGR271W
nucleotide binding	20 of 53 genes, 37.7%	944 of 7166 genes, 13.2%	0.00062	YML061C, YBL098W, YGR005C, YOL100W, YKR054C, YLR096W, YGL150C, YBR179C, YGR281W, YER172C, YIL013C, YLR419W, YKL203C, YER013W, YLR305C, YLR106C, YLL048C, YNL250W, YGR271W, YMR128W
nucleoside phosphate binding	20 of 53 genes, 37.7%	944 of 7166 genes, 13.2%	0.00062	YER172C, YIL013C, YER013W, YLR419W, YKL203C, YGR281W, YMR128W, YLL048C, YLR305C, YLR106C, YNL250W, YGR271W, YOL100W, YML061C, YGR005C, YBL098W, YLR096W, YGL150C, YBR179C, YKR054C
ATP binding	17 of 53 genes, 32.1%	719 of 7166 genes, 10.0%	0.00097	YGL150C, YLR096W, YKR054C, YOL100W, YML061C, YMR128W, YGR271W, YNL250W, YLR305C, YLR106C, YLL048C, YLR419W, YER013W, YKL203C, YIL013C, YER172C, YGR281W

adenyl ribonucleotide binding	17 of 53 genes, 32.1%	723 of 7166 genes, 10.1%	0.00105	YGR281W, YIL013C, YER172C, YLR419W, YER013W, YKL203C, YLL048C, YLR305C, YLR106C, YNL250W, YGR271W, YMR128W, YML061C, YOL100W, YKR054C, YLR096W, YGL150C
heterocyclic compound binding	20 of 53 genes, 37.7%	994 of 7166 genes, 13.9%	0.00137	YKR054C, YBR179C, YGL150C, YLR096W, YGR005C, YBL098W, YML061C, YOL100W, YGR271W, YNL250W, YLR305C, YLL048C, YLR106C, YMR128W, YGR281W, YLR419W, YER013W, YKL203C, YIL013C, YER172C
purine ribonucleoside triphosphate binding	18 of 53 genes, 34.0%	822 of 7166 genes, 11.5%	0.00141	YML061C, YOL100W, YKR054C, YBR179C, YGL150C, YLR096W, YGR281W, YLR419W, YER013W, YKL203C, YIL013C, YER172C, YGR271W, YNL250W, YLL048C, YLR305C, YLR106C, YMR128W
purine ribonucleotide binding	18 of 53 genes, 34.0%	826 of 7166 genes, 11.5%	0.00151	YIL013C, YER172C, YER013W, YLR419W, YKL203C, YGR281W, YMR128W, YLR305C, YLL048C, YLR106C, YNL250W, YGR271W, YOL100W, YML061C, YLR096W, YBR179C, YGL150C, YKR054C
ribonucleotide binding	18 of 53 genes, 34.0%	846 of 7166 genes, 11.8%	0.00211	YOL100W, YML061C, YLR096W, YBR179C, YGL150C, YKR054C, YIL013C, YER172C, YKL203C, YLR419W, YER013W, YGR281W, YMR128W, YLL048C, YLR305C, YLR106C, YNL250W, YGR271W

adenyl nucleotide binding	17 of 53 genes, 32.1%	765 of 7166 genes, 10.7%	0.00223	YER172C, YIL013C, YLR419W, YKL203C, YER013W, YGR281W, YMR128W, YLR305C, YLL048C, YLR106C, YNL250W, YGR271W, YOL100W, YML061C, YLR096W, YGL150C, YKR054C
carbohydrate derivative binding	18 of 53 genes, 34.0%	860 of 7166 genes, 12.0%	0.00265	YKR054C, YGL150C, YBR179C, YLR096W, YML061C, YOL100W, YGR271W, YNL250W, YLL048C, YLR305C, YLR106C, YMR128W, YGR281W, YLR419W, YER013W, YKL203C, YIL013C, YER172C
purine nucleotide binding	18 of 53 genes, 34.0%	868 of 7166 genes, 12.1%	0.003	YLR305C, YLR106C, YLL048C, YNL250W, YGR271W, YMR128W, YGR281W, YIL013C, YER172C, YKL203C, YLR419W, YER013W, YKR054C, YLR096W, YGL150C, YBR179C, YML061C, YOL100W
anion binding	19 of 53 genes, 35.8%	974 of 7166 genes, 13.6%	0.00391	YMR128W, YGR271W, YNL250W, YLR305C, YLL048C, YLR106C, YLR419W, YER013W, YKL203C, YIL013C, YER172C, YGR281W, YBR179C, YGL150C, YLR096W, YKR054C, YOL100W, YBL098W, YML061C
voltage-gated monoatomic cation channel activity	2 of 53 genes, 3.8%	2 of 7166 genes, 0.0%	0.00595	YJL093C, YGR217W

Supplemental Code C1: blast_comparison.py

```
import os
import subprocess
import tempfile
import argparse
from Bio import SeqIO
from Bio.SeqRecord import SeqRecord
from Bio.Seq import Seq

def local_blast_analysis(hybrid_assembly_path, parent_genomes_paths, output_path, chunk_size,
score_type, score_threshold):
    # header
    with open(output_path, 'w') as output_file:
        headers = ["qseqid", "sseqid", "pident", "length", "mismatch", "gapopen", "qstart", "qend",
"sstart", "send", "evalue", "bitscore"]
        output_file.write("\t".join(headers) + "\n")

    with open(hybrid_assembly_path, 'r') as hybrid_assembly_file:
        for record in SeqIO.parse(hybrid_assembly_file, "fasta"):
            sequence = str(record.seq)
            seq_id = record.id
            current_chunk_sequence = ''
            chunk_start = 1

            for i in range(0, len(sequence), chunk_size):
                chunk_end = chunk_start + chunk_size - 1
                chunk_sequence = sequence[i:i+chunk_size]
                chunk_id = f"{seq_id}_{chunk_start}-{chunk_end}"

                print("\nChunk: ", chunk_start, " : ", chunk_end)
                perform_local_blast(chunk_sequence, chunk_id, parent_genomes_paths, output_path,
score_threshold, score_type)

                chunk_start = chunk_end + 1

    # end (partial) chunk
    if current_chunk_sequence:
        chunk_end = chunk_start + len(current_chunk_sequence) - 1
        perform_local_blast(chunk_to_process, parent_genomes_paths, output_path, score_threshold,
args.score_type)

def perform_local_blast(chunk_sequence, chunk_id, parent_genomes_paths, output_path,
score_threshold, score_type):
    chunk_record = SeqRecord(Seq(chunk_sequence), id=chunk_id, description="")
    best_scores = []
    best_hits = []
    # Can't pass directly -- throws error due to length
```

```
    with tempfile.NamedTemporaryFile(mode='w+', delete=False) as temp_seq_file:
        SeqIO.write(chunk_record, temp_seq_file, "fasta")
        temp_seq_file_path = temp_seq_file.name

    for parent_genome_path in parent_genomes_paths:
        try:
            blastn_output = subprocess.check_output(['blastn', '-query', temp_seq_file_path, '-db',
parent_genome_path, '-outfmt', '6'])
            decoded_blastn_output = blastn_output.decode('utf-8')
            hits = blastn_output.decode('utf-8').strip().split('\n')
            print("BLAST output for: ", parent_genome_path)
            print(hits[0])
            if hits and hits != ['']:
                best_hit = hits[0].split('\t')
                #if scoretype == 'bitscore':
                score = float(best_hit[11])  # add escore for comparison type (?)
                #if scoretype == 'escore':
                #score=flat(best_hit[10]):
                best_scores.append(score)
                best_hits.append(best_hit)
            else:
                print("no hits")
                best_scores.append(-1)  # no hits (shouldn't happen)
                best_hits.append(None)

        except subprocess.CalledProcessError as e:
            print("Error: ", e.output)
            best_scores.append(-1)
            best_hits.append(None)

    os.remove(temp_seq_file_path)

    for hit in best_hits:
        if hit:
            #if scoretype=='bitscore':
            score = float(hit[11])  # implement escore (column 10)
            #if scoretype=='escore':
            #score=float(hit[10])
            best_scores.append(score)
        else:
            best_scores.append(-1)

    if best_scores:
        sorted_scores_with_parents = sorted(zip(best_scores, parent_genomes_paths, best_hits),
reverse=True)

        if len(sorted_scores_with_parents) > 1 and sorted_scores_with_parents[0][0] -
sorted_scores_with_parents[1][0] >= score_threshold:
```

```
            inferred_parent = sorted_scores_with_parents[0][1]
            best_hit = sorted_scores_with_parents[0][2]
        else:
            inferred_parent = "Undetermined"
            best_hit = None

        # output
        print("Inferred parent: ",inferred_parent)
        with open(output_path, 'a') as output_file:
            output_line = '\t'.join(best_hit) if best_hit else "No clear hit"
            output_file.write(inferred_parent + "\t" + output_line + "\n")
    else:
        print("No hits found for either parent.")

if __name__ == '__main__':
    parser = argparse.ArgumentParser(description="Run BLAST, output inferred parent based")

    parser.add_argument("query_sequence", help="query sequence")
    parser.add_argument("target_sequences", nargs='+', help="target (parental) sequences")
    parser.add_argument("output_path", help="output file")
    parser.add_argument("--chunk_size", type=int, default=5000, help="Chunk size, default is 5k")
    parser.add_argument("--score_type", choices=['bitscore', 'escore'], type=str, default='bitscore',
help="'bitscore' or 'escore' (only bitscore currently), bitscore is default")
    parser.add_argument("--score_threshold", type=float, default=100, help="threshold for
bitscore/escore (default 100)")

    args = parser.parse_args()

    local_blast_analysis(args.query_sequence, args.target_sequences, args.output_path,
args.chunk_size, args.score_type, args.score_threshold)
```

Supplemental Code C2: pairwise_p_distance_calc.py

```
import os
from Bio import SeqIO

def calculate_p_distance(seq1, seq2):
    differences = 0
    valid_positions = 0
    for c1, c2 in zip(seq1, seq2):
        if c1 != '-' and c2 != '-':
            valid_positions += 1
            if c1 != c2:
                differences += 1
    return differences / valid_positions if valid_positions > 0 else 0
```

```
# Path settings
input_dir = './mafft_all'
output_path = './output_p_distance.csv'

# Prepare the output file
with open(output_path, 'w') as output_file:
    output_file.write("Filename,Sbay_Gene_Pair,P_Distance\n")

    # Iterate over all files in the mafft_all folder
    for filename in os.listdir(input_dir):
        if filename.endswith("mafft.out"):
            file_path = os.path.join(input_dir, filename)
            sbay_seqs = []
            non_sbay_seqs = []

            # Read sequence files
            for record in SeqIO.parse(file_path, "fasta"):
                if record.id.startswith("Sbay"):
                    sbay_seqs.append(record)
                else:
                    non_sbay_seqs.append(record)

            # Compare each pair of Sbay and non-Sbay genes
            for sbay in sbay_seqs:
                for non_sbay in non_sbay_seqs:
                    p_distance = calculate_p_distance(sbay.seq, non_sbay.seq)
                    pair_name = f"{sbay.id}_vs_{non_sbay.id}"
                    output_file.write(f"{filename},{pair_name},{p_distance:.4f}\n")
                    output_file.flush()
            print(f"Processed file: {filename}")  # Print processing progress

print("Calculation complete, results saved in: output_p_distance.csv")
```