Supplementary Materials for

Mitochondrial introgression suggests extensive ancestral hybridization events among *Saccharomyces* species

Authors: David Peris, Armando Arias, Sandi Orlić, Carmela Belloch, Laura Pérez-Través, Amparo Querol, Eladio Barrio

Correspondence to: david.perisnavarro@wisc.edu

This PDF file includes:

Supplementary Text

Table of Contents

Supplementary Text

COX2 nucleotide statistics

A total of five hundred and seventeen *COX2* sequences of worldwide distributed wild *Saccharomyces* strains and forty-nine sequences of domesticated natural hybrids were sequenced or retrieved from public databases (Table S1, Figure 1). The *COX2* alignment is 585 bp, with 106 variable positions (18.11%) (Figure S2) where 43.4% of polymorphisms are found in the last 89 nucleotides. 80 of the variable positions are phylogenetically informative. 27 variable sites correspond to 0-fold degenerated, but only 10 of them are informative. 20 variable sites correspond to 2-fold degenerated positions (19 informative and 1 singleton), being 5 of them non-synonymous substitutions. Finally, 42 variable sites are 4-fold degenerated and 38 of them are informative.

COX2 genetic diversity statistics are summarized in Table S4. The most diverse sequences were found in three species, *S. cerevisiae, S. paradoxus* and *S. mikatae*. In the case of *S. cerevisiae* nucleotide diversity values of 2.1% are so much higher than values inferred for the nuclear genome, around 0.9% (Liti et al. 2009; Wang et al. 2012), suggesting that recombination is increasing the values. For *S. eubayanus* and *S. uvarum*, the *COX2* nucleotide diversity values, 0.5 and 0.6% respectively, are lower than those found for the nuclear genome, around 0.8% (Patagonia A-Patagonia B *S. eubayanus*) and 1% (Holarctic-South America B *S. uvarum*) (Almeida et al. 2014; Peris et al. 2014) in agreement with the low mutation rate of yeast mitochondrial genomes (Clark-Walker 1991). *COX2* values for *S. kudriavzevii* strains are similar to those inferred for the nuclear genome, 1.11% (Hittinger et al. 2010).

Maximum-Likelihood phylogenetic tree of *COX2* 3'end region is incongruent with the species tree due to a recombination hotspot

COX2 alignment was split in two segments based on the most common recombination point. We reconstructed the Maximum-Likelihood tree with `PhyML v3.0` (Guindon and Gascuel 2003) using the best fitted substitution model inferred by `jModeltest` (Posada 2008). Phylogenetic trees comparison with the *Saccharomyces* species tree was performed in `Tree Puzzle v5.2` (Schmidt et al. 2002) implemented with the conservative Shimodaira-Hasegawa (Shimodaira and Hasegawa 1999) and Expected

Likelihood weights (ELW) (Strimmer and Rambaut 2002) tests. Both *COX2* 5'end segment and the species tree agree each other; however, the *COX2* 3'end was best explained by the inferred ML using that sequence region, and significantly rejected the topology of the species tree (p-value < $1*10^{-5}$), suggesting incongruence between the inferred ML *COX2* 3'end phylogenetic tree and the species tree.

A set of *COX2* representative sequences of each haplotype was the input of `RDPv3.44` package (Martin et al. 2010). RDP includes six methods to detect recombination: RDP (Martin and Rybicki 2000), Bootscanning (Salminen et al. 1995), MaxChi (Smith 1992), Chimaera (Smith 1992), GeneConv (Padidam et al. 1999) and Sis-scan (Gibbs et al. 2000). Default settings were set up with a statistical significance threshold of p-value < 0.05, with Bonferroni correction for multiple comparisons.

Codons (nucleotide sequences 526-528 and 535-537) in the *COX2* 3' end show several convergent nucleotide substitutions among *S. cerevisiae* and *S. paradoxus* strains. This finding might be the result of a "patchy-tachy" effect due to different substitution rate in those positions (Sun et al. 2011). Homoplasies can drive to recombinant false positives. One confounding effect can be observed in L1528 and CECT11757 strains which is difficult to know if they were truly introgressed or they suffered two different recombination events. A recent study shown transfers from European *S. paradoxus* GC48 cluster to L1528. Also the presence of type II *ORF1* segments supports the introgression scenario and a double recombination scenario.

***ORF1* gene structure**

52 different haplotypes were found in our representative strains (Table S1). *ORF1* start codon is nineteen nucleotides inside the 3' end of COX2 gene. The average GC-composition of the ORF1 is extremely low, 18%. Eleven strains have a GTG as a start codon, which is uncommon in yeast mitochondria: haplotypes M2-M7 and M9-M12. The translation of the *ORF1* gene predicts fourteen strains having a premature stop codon: haplotypes M2, M3, M7-M11, M17-M20, M23 and M46. *ORF1* sequence length range from 1363bp (ZA17 strain) to 1516bp (VRB strain), translated to 454 and 505, respectively. Note that we sequenced a partial *ORF1* gene, 45 nucleotides before the last base pair annotated for the S288c *ORF1* gene.

ORF1 secondary structure and domains (LAGLIDADG and NUMOD) were annotated in Jalview according to Dalgaard *et al.* (1997) and using the Conserved Domains tool in NCBI (http://www.ncbi.nlm.nih.gov/Structure/cdd/ wrpsb.cgi) (Marchler-Bauer et al. 2009). WebLogo profiles of LAGLIDADG and NUMOD1 domains were done in WebLogo 2.8.2 tool (http://weblogo.berkeley.edu/). Three different domains were found in all *ORF1* sequences: two LADGLIDADG (P1 and P2) and one NUMOD1 (Figure 4, S10 and S11). Two non-*Saccharomyces ORF1* were detected by PSI-BLAST (Altschul et al. 1997) showing some similarities with the *Saccharomyces ORF1* sequences. We generated a new alignment in Jalview 4.0.b2 (Waterhouse et al. 2009) including the *Williopsis saturnus* var. *suaveolans* (*ORF1* and *ORF3*), and one from *Kazachstania servazii* (SasefMp08) sequences (Figure S11). The sequence from *K. servazii* was found to be closely related to the *Saccharomyces ORF1* sequences (Figure S5). Although, *W. saturnus* is phylogenetically fairly distant from *Saccharomyces* genus (James SA et al. 1998), they have been diverging since 235mya, we observed some conservation among LAGLIDADG and NUMOD aminoacids when compared with *Saccharomyces* and *Kazachstania ORF1* sequences (Figure S10 and S11).

Differences in size among *ORF1* gene sequences were due to the presence of GC insertions (Figure S5) and AT repeats (Table S3). Seven different GC clusters insertion points were found along the *ORF1* alignment. GC cluster 1 is observed in VRB; three types of GC cluster 2 were found in CBS435, CECT11757 and 120MX (Figure S5 and S12); GC cluster 3 is shown in CBS10644; GC cluster 4 is shown in VRB; GC cluster 5 is inserted in *ORF1* from CBS435, and was structurally similar to GC cluster 3 of CBS10644; GC cluster 6 and 7 were the most frequent among *Saccharomyces* strains (Figure S9). 40 *S. cerevisiae* and one hybrid *S. cerevisiae* x *S. kudriavzevii* (AMH) have the GC cluster 6, and 36 *S. cerevisiae*, 2 hybrids *S. cerevisiae* x *S. kudriavzevii,* 2 Far Eastern and 2 American *S. paradoxus* shown the GC cluster 7 in their *ORF1* sequences (Figure S9). Some GC clusters sequences were found inverted (Figure S5). GC cluster 7 of the two American *S. paradoxus* (120MX and CBS5313) and three *S. cerevisiae*, two from Japan and one from USA (CBS435, Y9 and YPS606), showed identical GC cluster sequence (Figure S9).

ORF1 GC clusters were classified according to Zamaroczy and Bernardi (1986), except for GC cluster 2. GC cluster 1 and 4 are similar to a1 family, and GC cluster 3 and 5 were similar to a4 family. In the case of GC cluster 1, 2, 4, and 5 are on the opposite strand. As Séraphin *et al.* (1987) and Weiller *et al.* (1989) described, the GC clusters of the *ORF1* gene were flanked by TAG and AGGAG, or CTA and CTCCT when cluster was inserted in the opposite strand (Figure S12). These conserved nucleotides were flanked by A+T rich sequences. Flanking sequences TAG and AG (CTA and CT) are conserved in most of the sequences with and without GC clusters. All GC clusters in *ORF1* belong to group M1 (Weiller et al. 1989). Conserved flanking regions might help to the transposition of GC clusters.

Differences in size were also due to the presence of AT repeats. Tandem repeat sequences for *COX2* and *ORF1* genes were detected and described using `Tandem Repeat Finder` software (Benson 1999). Twenty-one different A+T rich sequences that repeated at least twice were found in the *ORF1* alignment (Table S3). The length of A+T rich tandem repeats ranged from three nucleotides to twenty five nucleotides. The most repeated sequence (AAT) was repeated ten times in haplotypes M1, M12, M22, M39, M40, and M50. The A+T rich tandem repeats were located near to GC clusters (Figure 4). In *COX2* gene, we found three A+T rich sequences repeated twice (Table S3).

ORF1 detection of selection

The single likelihood ancestor counting (SLAC), fixed effects likelihood (FEL), and random effects likelihood (REL) methods (Pond and Frost 2005), implemented in the `HYPHY` web based version, `Datamonkey` (Delport et al. 2010), were used to detect the signatures of selection operating on *ORF1* protein gene. NJ phylogenetic trees were reconstructed under the REV (GTR) substitution model. Two different approaches were performed to describe selection signatures. In the first approach, a complete *ORF1* alignment, without GC Insertions and indels, partitioned by recombinant sections detected by GARD, was used. In a second study, the two LADGLIDADG and the NUMOD domains were analyzed independently.

Aminoacid positions were described as being under positive or purifying selection when significant values were generated by two of the three tested methods. Codon-

specific selection pressure along the sequences (i.e. site specific dN-dS) was measured and *p*-values were estimated at each site. We analyzed 417 *ORF1* codons of the total 458. A 61 codons, corresponding to 15% of the total were found to be under purifying selection, where 43 of the 61 codons were found in LAGLIDADGs and NUMOD1 domains (Figure S10). A functional characterization might help to understand how important these aminoacids are for the activity of this homing endonuclease.

References

Almeida P, Gonçalves C, Teixeira S *et al*. 2014. A Gondwanan imprint on global diversity and domestication of wine and cider yeast *Saccharomyces uvarum*. Nat Commun. 5:

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res. 25:3389-3402.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucl Acids Res. 27:573-580.

Clark-Walker GD. 1991. Contrasting mutation rates in mitochondrial and nuclear genes of yeasts versus mammals. Curr Genet. 20:195-198.

Dalgaard JZ, Klar AJ, Moser MJ, Holley WR, Chatterjee A, Mian IS. 1997. Statistical modeling and analysis of the LAGLIDADG family of site- specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family. Nucl Acids Res. 25:4626-4638.

de Zamaroczy M, Bernardi G. 1986. The GC clusters of the mitochondrial genome of yeast and their evolutionary origin. Gene. 41:1-22.

Delport W, Poon AFY, Frost SDW, Kosakovsky Pond SL. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics. 26:2455-2457.

Gibbs MJ, Armstrong JS, Gibbs AJ. 2000. Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics. 16:573-582.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by Maximum Likelihood. Syst Biol. 52:696-704.

Hittinger CT, Gonçalves P, Sampaio JP, Dover J, Johnston M, Rokas A. 2010. Remarkably ancient balanced polymorphisms in a multi-locus gene network. Nature. 464:54-58.

James SA, Roberts IN, Collins MD. 1998. Phylogenetic heterogeneity of the genus *Williopsis* as revealed by 18S rRNA gene sequences. Int J Syst Bacteriol. 48:591-596.

Leducq JB, Charron G, Samani P *et al*. 2014. Local climatic adaptation in a widespread microorganism. Proc R Soc Lond B Biol Sci. 281:

Liti G, Carter DM, Moses AM *et al*. 2009. Population genomics of domestic and wild yeasts. Nature. 458:337-341.

Livingstone C, Barton GJ. 1993. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput Appl Biosci. 9:745-756.

Marchler-Bauer A, Anderson JB, Chitsaz F *et al*. 2009. CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res. 37:D205-D210.

Martin D, Rybicki E. 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics. 16:562-563.

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics. 26:2462-2463.

Padidam M, Sawyer S, Fauquet CM. 1999. Possible emergence of new Geminiviruses by frequent recombination. Virology. 265:218-225.

Peris D, Sylvester K, Libkind D, Gonçalves P, Sampaio JP, Alexander WG, Hittinger CT. 2014. Population structure and reticulate evolution of *Saccharomyces eubayanus* and its lager-brewing hybrids. Mol Ecol. 23:2031-2045.

Pond SLK, Frost SDW. 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics. 21:2531-2533.

Posada D. 2008. jModelTest: phylogenetic model averaging. Mol Biol Evol. 25:1253-1256.

Salminen MO, Carr JK, Burke DS, McCutchan FE, Garcia-Martinez J. 1995. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS Res Hum Retroviruses. 11:1423-1425.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 18:502-504.

Séraphin B, Simon M, Faye G. 1987. The mitochondrial reading frame RF3 is a functional gene in *Saccharomyces uvarum*. J Biol Chem. 262:10146-10153.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 16:1114-1116.

Smith JM. 1992. Analyzing the mosaic structure of genes. J Mol Evol. 34:126-129.

Strimmer K, Rambaut A. 2002. Inferring confidence sets of possibly misspecified gene trees. Proceedings Biological sciences / The Royal Society. 269:137-142.

Sun S, Evans BJ, Golding GB. 2011. "Patchy-Tachy" leads to false positives for recombination. Mol Biol Evol. 28:2549-2559.

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

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics. 25:1189-1191.

Weiller G, Schueller CME, Schweyen RJ. 1989. Putative target sites for mobile G+C rich clusters in yeast mitochondrial DNA: Single elements and tandem arrays. Mol Gen Genet. 218:272-283.

Supplementary Tables and Figures

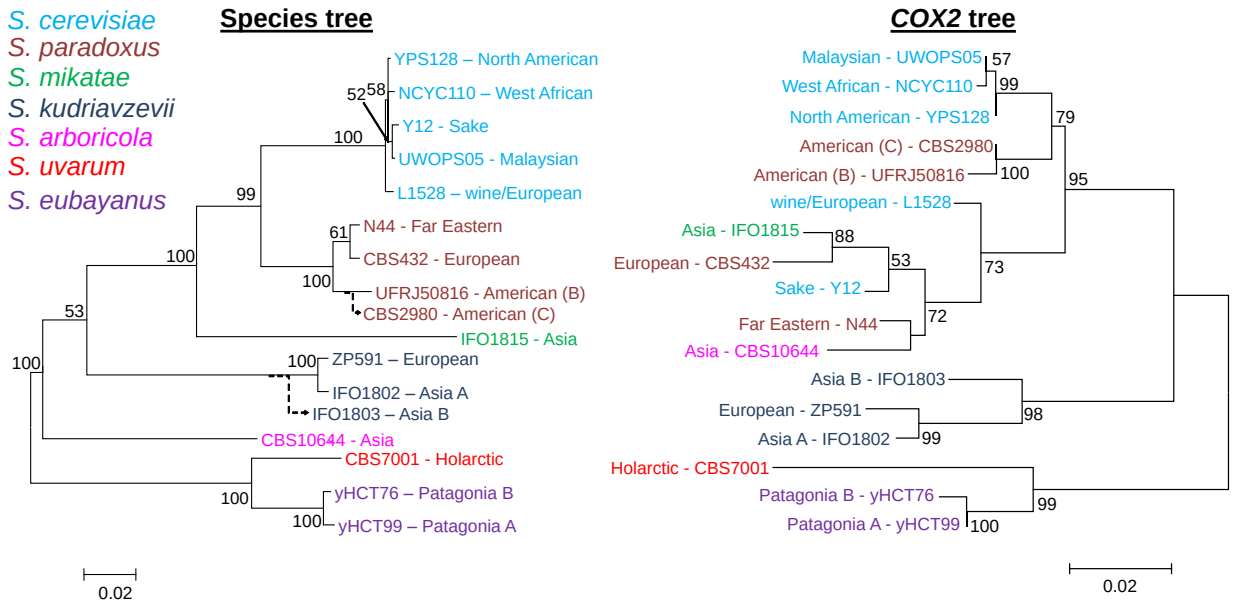
Table S1. Strain information.

Table S2. Primer pairs used in this study.

Table S3. Distribution of AT tandem repeats.

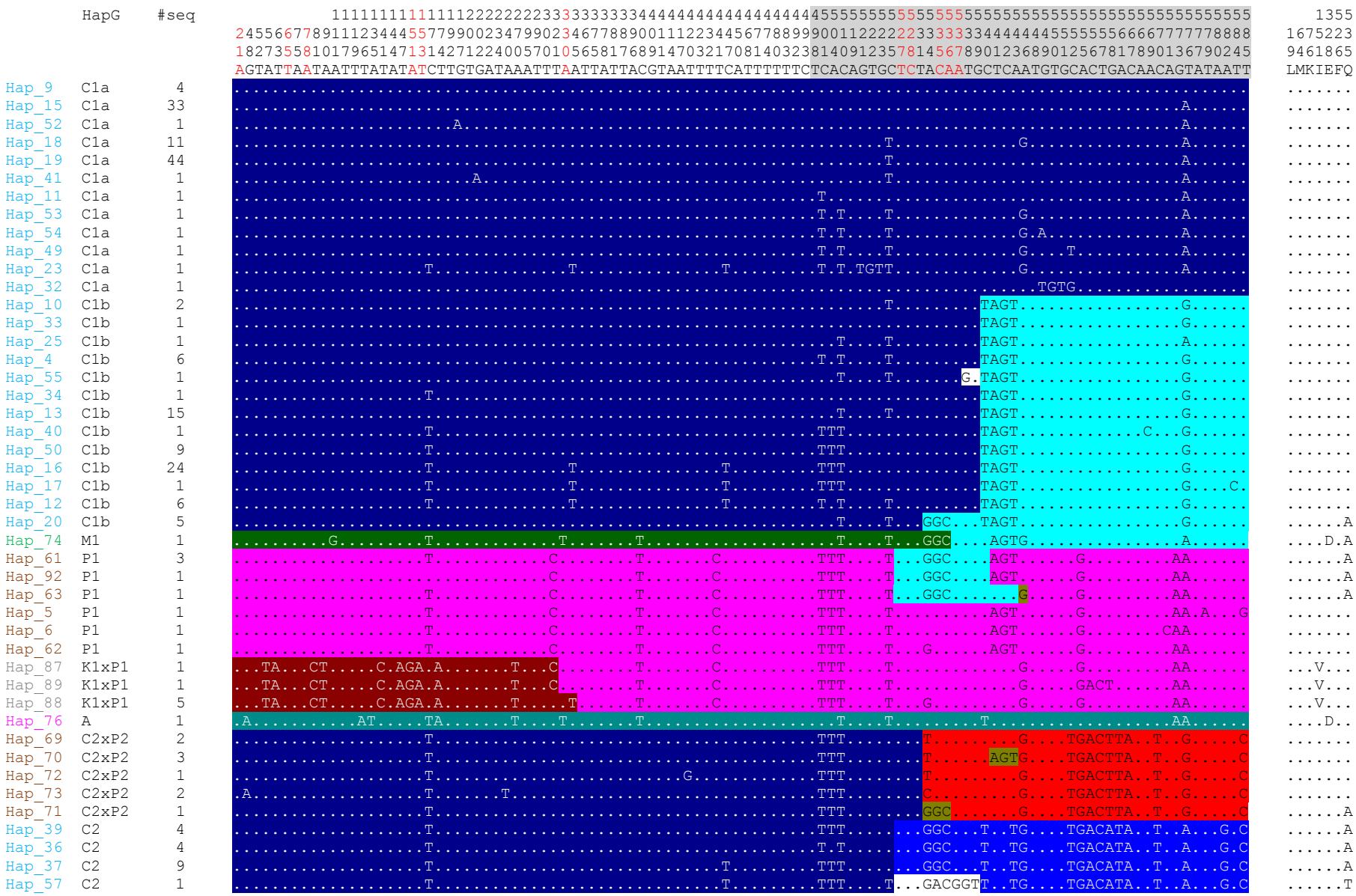
Table S4. Summary statistics for all *Saccharomyces* and each species using *COX2*.

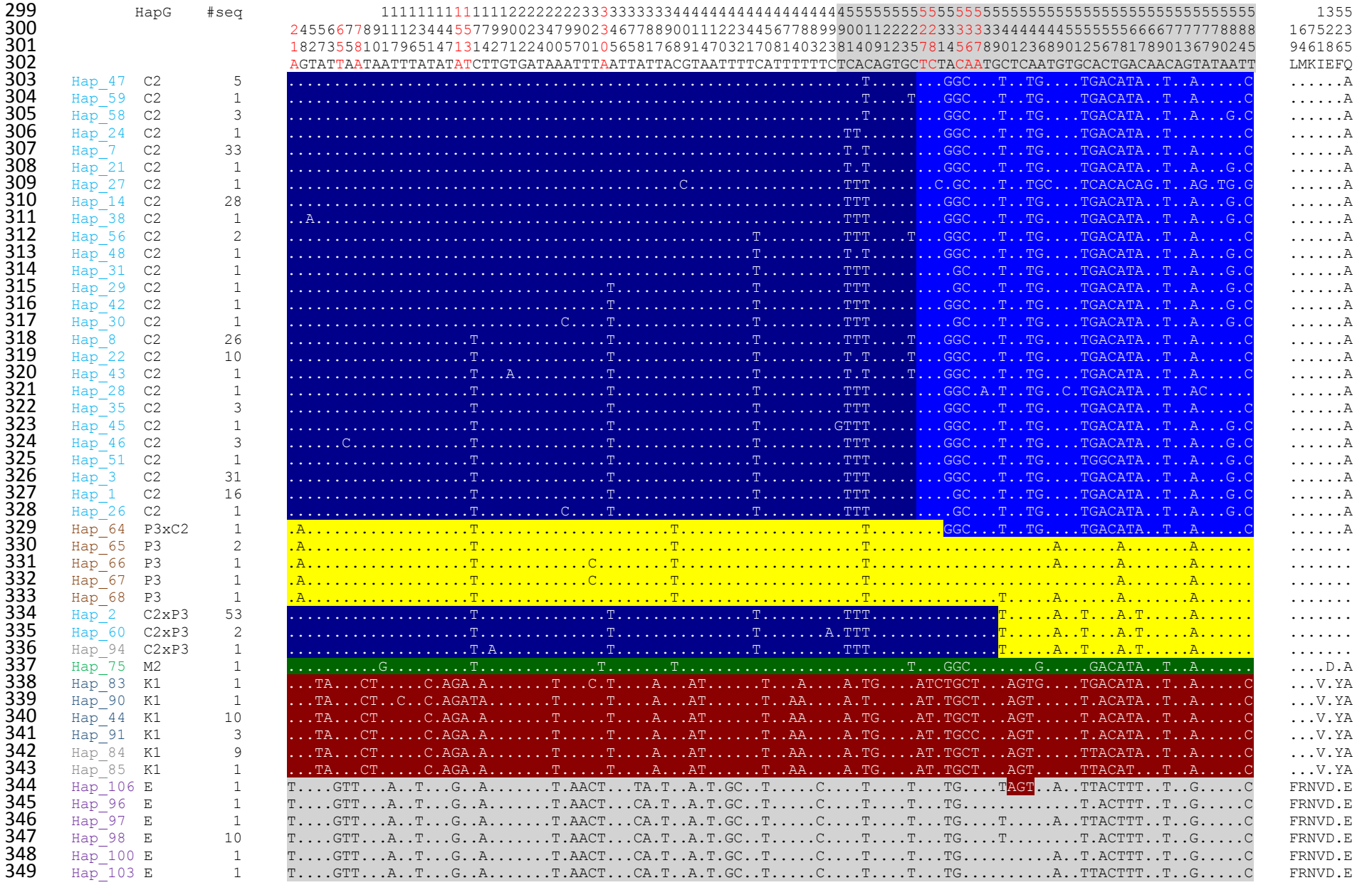
Figure S1. Maximum-Likelihood phylogenetic trees.

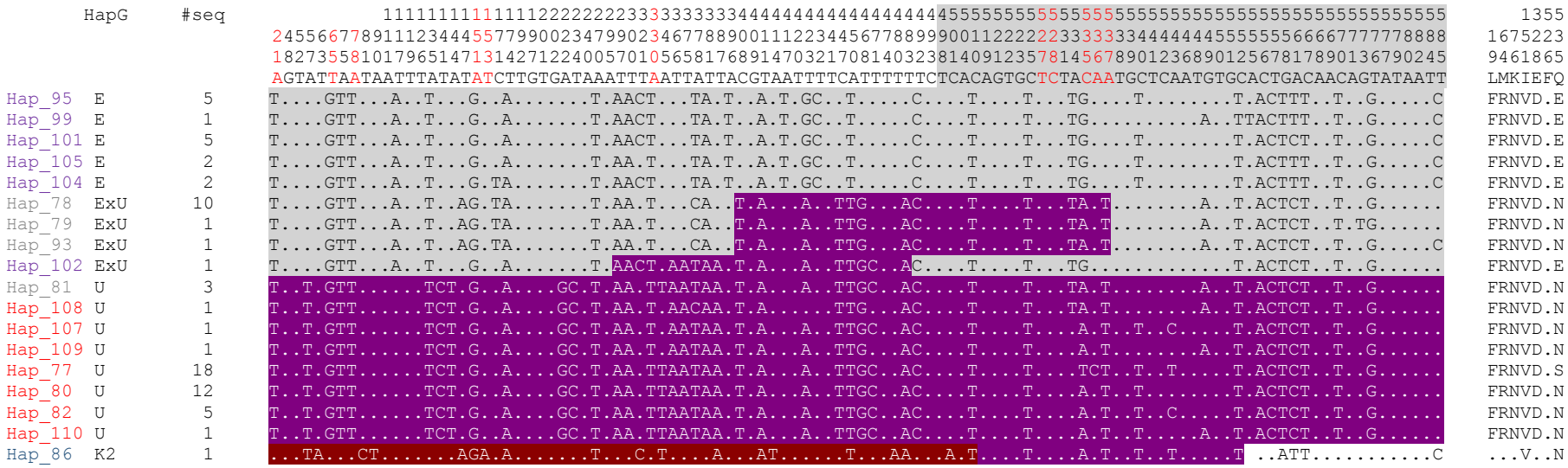


ML phylogenetic tree using a concatenated alignment of *RIP1, MET2* and *FUN14* (left), and *COX2* (right) are represented. Strains were colored according to the species designation. CBS2980 and IFO1803 strain positions in the phylogenetic species tree were inferred according to Leducq *et al.* (2014) and Hittinger *et al.* (2010). Bootstrap values above 50 are given for each branch. Scales are given in nucleotide substitution per site.

Figure S2. *COX2* polymorphic sites.

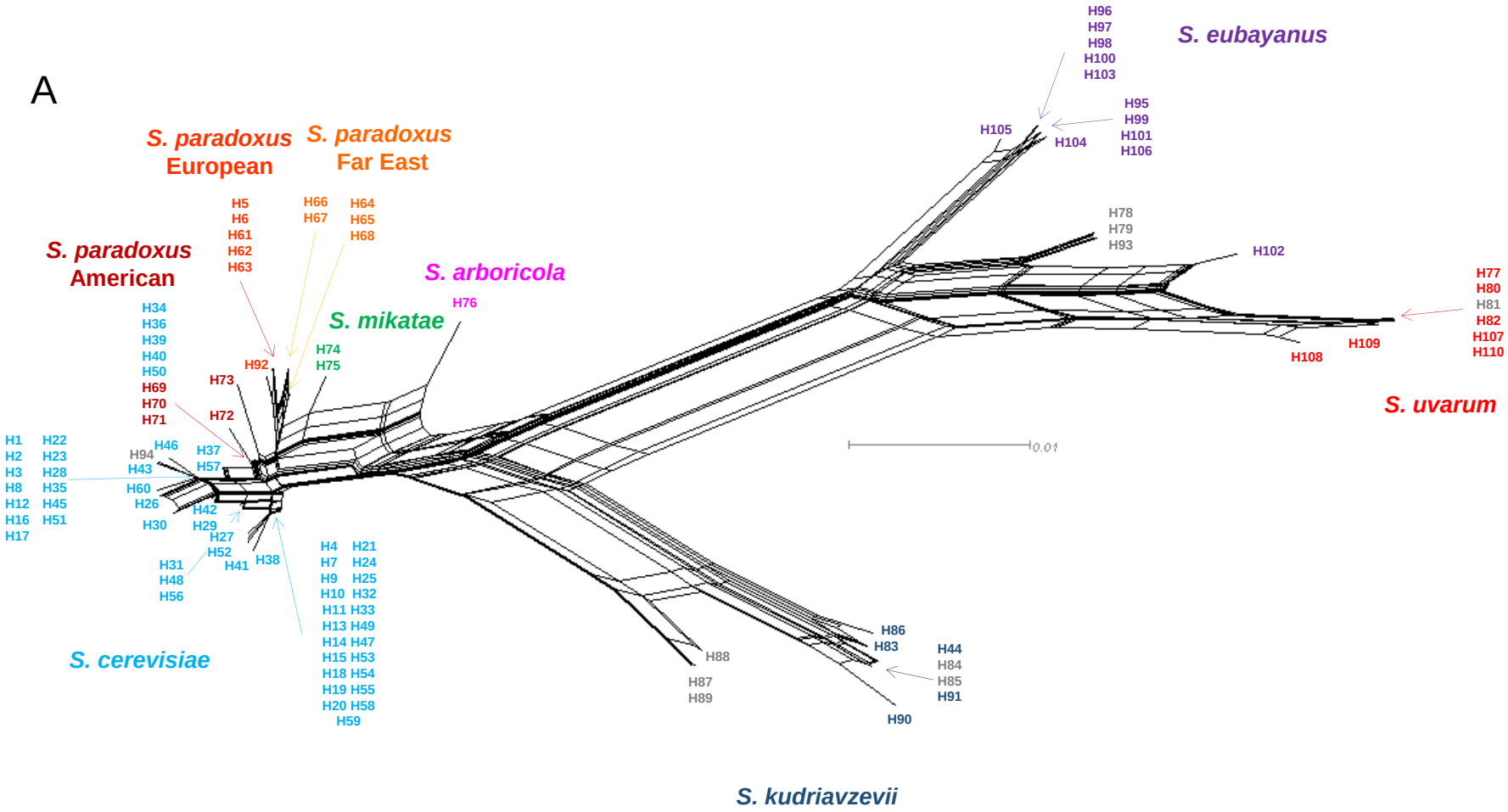




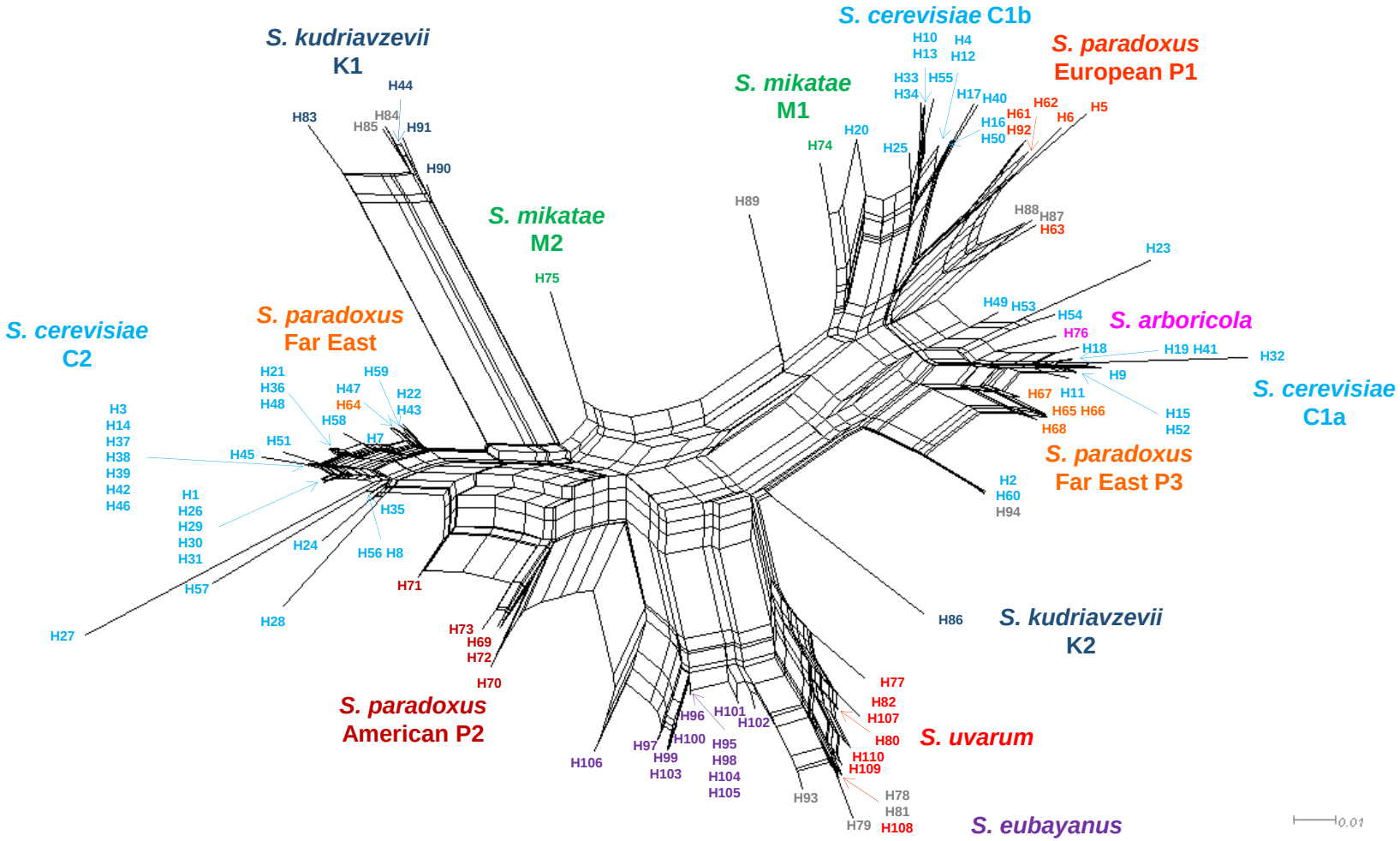


Variable *COX2* nucleotide positions and informative aminoacid changes among haplotypes are shown. Haplotype number are colored according to the species designation, as in Figure 2. Numbers above polymorphic sites represent their position in the alignment. Grey shadow highlights the *COX2* 3'end region. Haplogroup (HapG) designation and the number of sequences (#seq) for each haplotype are shown. Red nucleotide positions correspond to codon nucleotides, encoding for the aminoacid shown in the right hand, which position of the first nucleotide of the codon is shown above the aminoacid sequence. Sequences were colored according to their similarity as inferred from Figure S3B.

Figure S3. *COX2* Neighbor-Net phylogenetic network for each *COX2* segment.

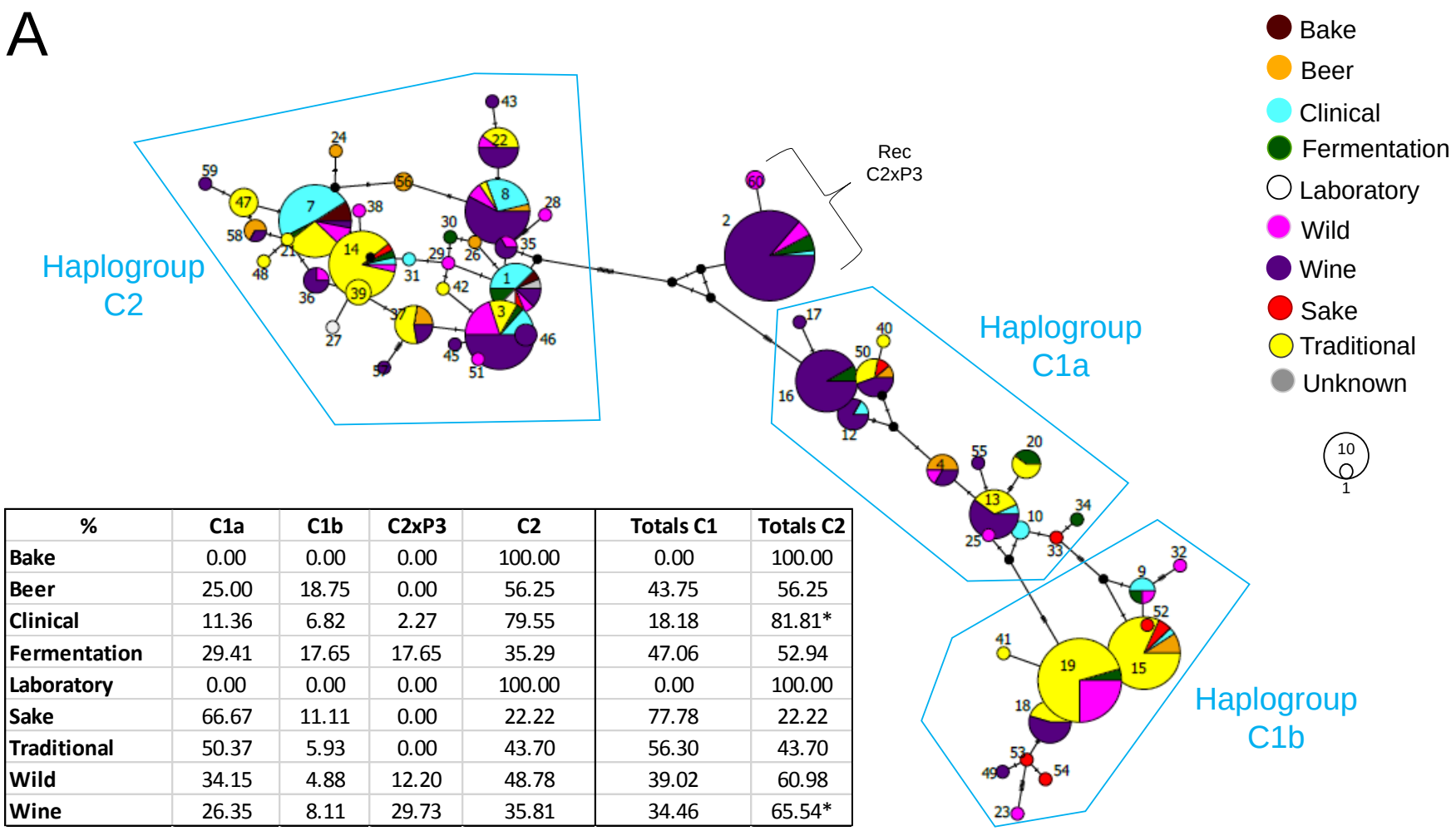


B

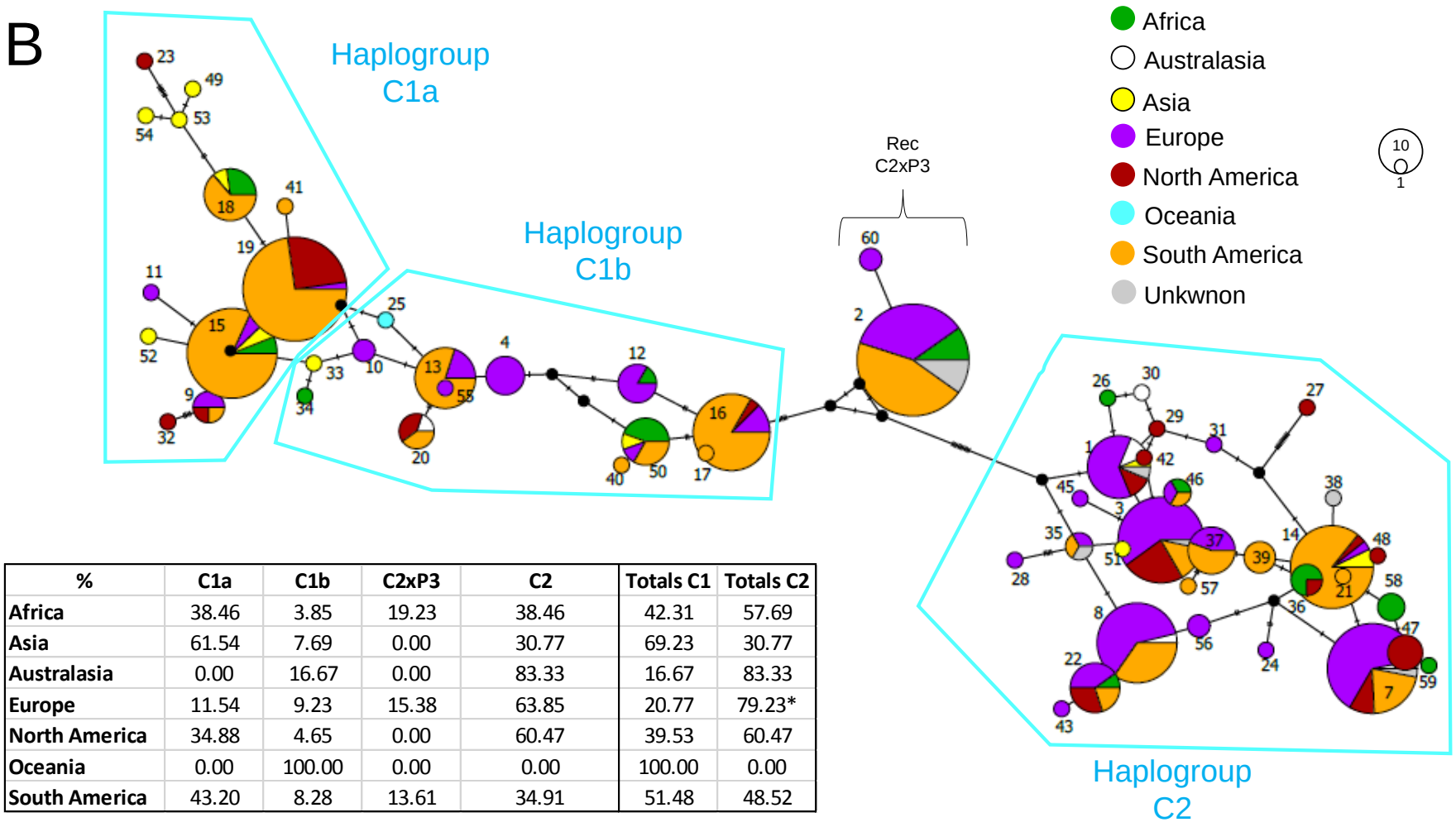


Neighbor-net phylogenetic networks for *COX2* 5' end and 3' end segments are represented in A) and B), respectively. Haplotypes were colored according to each species designation. Haplotypes exclusively from hybrids are colored in grey. Scale is given in nucleotide substitution per site.

Figure S4. *S. cerevisiae COX2* MJ networks.



%	C1a	C1b	C2xP3	C2	Totals C1	Totals C2
Bake	0.00	0.00	0.00	100.00	0.00	100.00
Beer	25.00	18.75	0.00	56.25	43.75	56.25
Clinical	11.36	6.82	2.27	79.55	18.18	81.81*
Fermentation	29.41	17.65	17.65	35.29	47.06	52.94
Laboratory	0.00	0.00	0.00	100.00	0.00	100.00
Sake	66.67	11.11	0.00	22.22	77.78	22.22
Traditional	50.37	5.93	0.00	43.70	56.30	43.70
Wild	34.15	4.88	12.20	48.78	39.02	60.98
Wine	26.35	8.11	29.73	35.81	34.46	65.54*

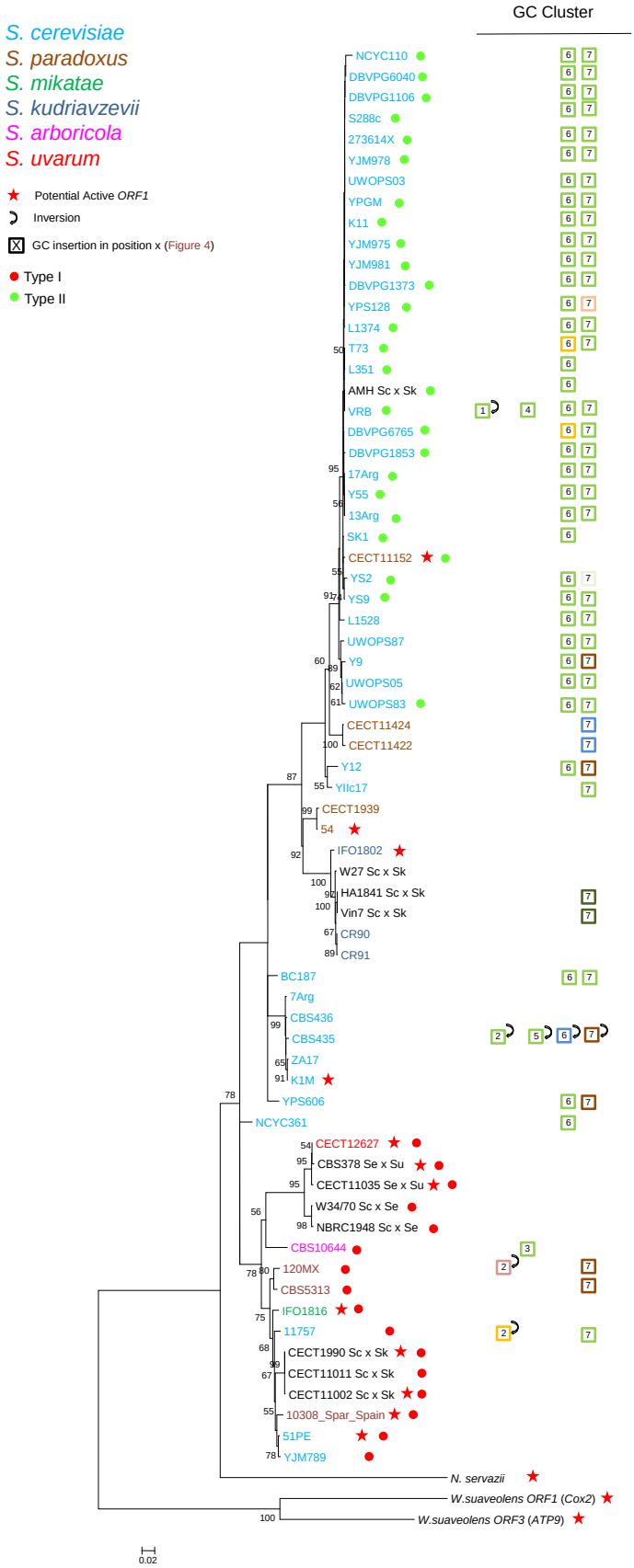


%	C1a	C1b	C2xP3	C2	Totals C1	Totals C2
Africa	38.46	3.85	19.23	38.46	42.31	57.69
Asia	61.54	7.69	0.00	30.77	69.23	30.77
Australasia	0.00	16.67	0.00	83.33	16.67	83.33
Europe	11.54	9.23	15.38	63.85	20.77	79.23*
North America	34.88	4.65	0.00	60.47	39.53	60.47
Oceania	0.00	100.00	0.00	0.00	100.00	0.00
South America	43.20	8.28	13.61	34.91	51.48	48.52

S. cerevisiae COX2 MJ networks were represented in A) and B) where haplotype pie charts were colored according to the isolation source and continent of isolation, respectively. Circles sizes represents the number of sequences in a haplotype and is scaled according to the legend. Number of mutations from one haplotype to another are indicated by lines in the edges connecting

391 the haplotypes. A table showing the distribution in each haplogroup or recombinant group by isolation source or continent is also
392 displayed in A) and B), respectively.

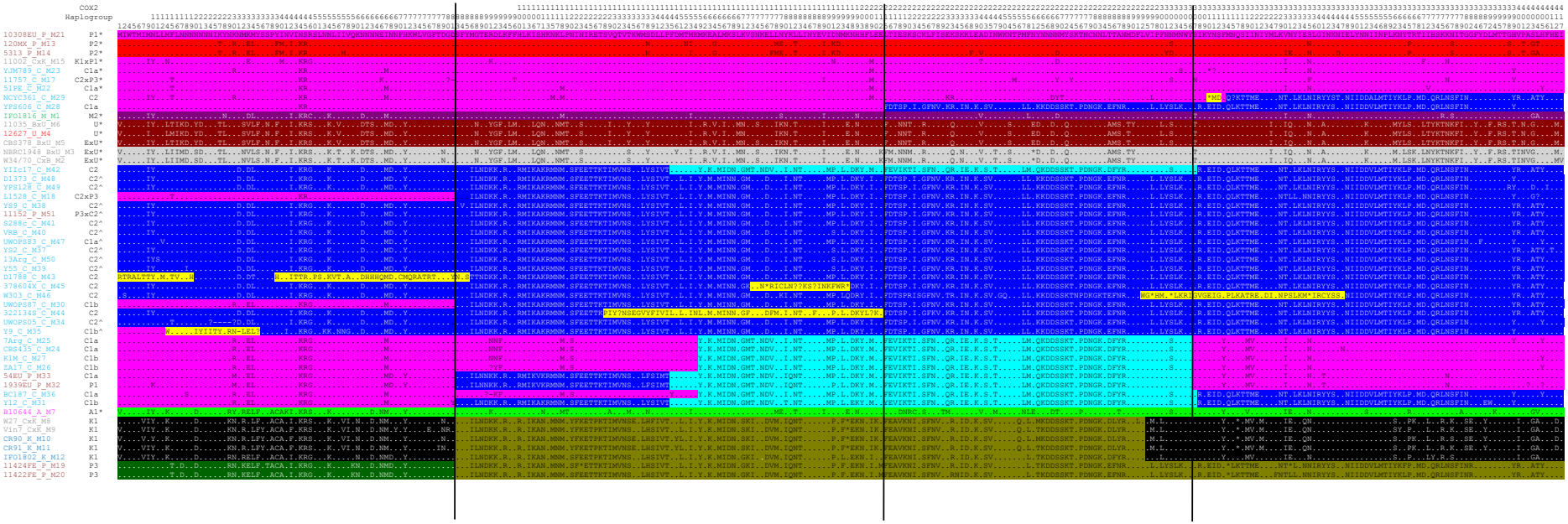
393 **Figure S5. *ORF1* Neighbor-Joining phylogenetic tree.**



394

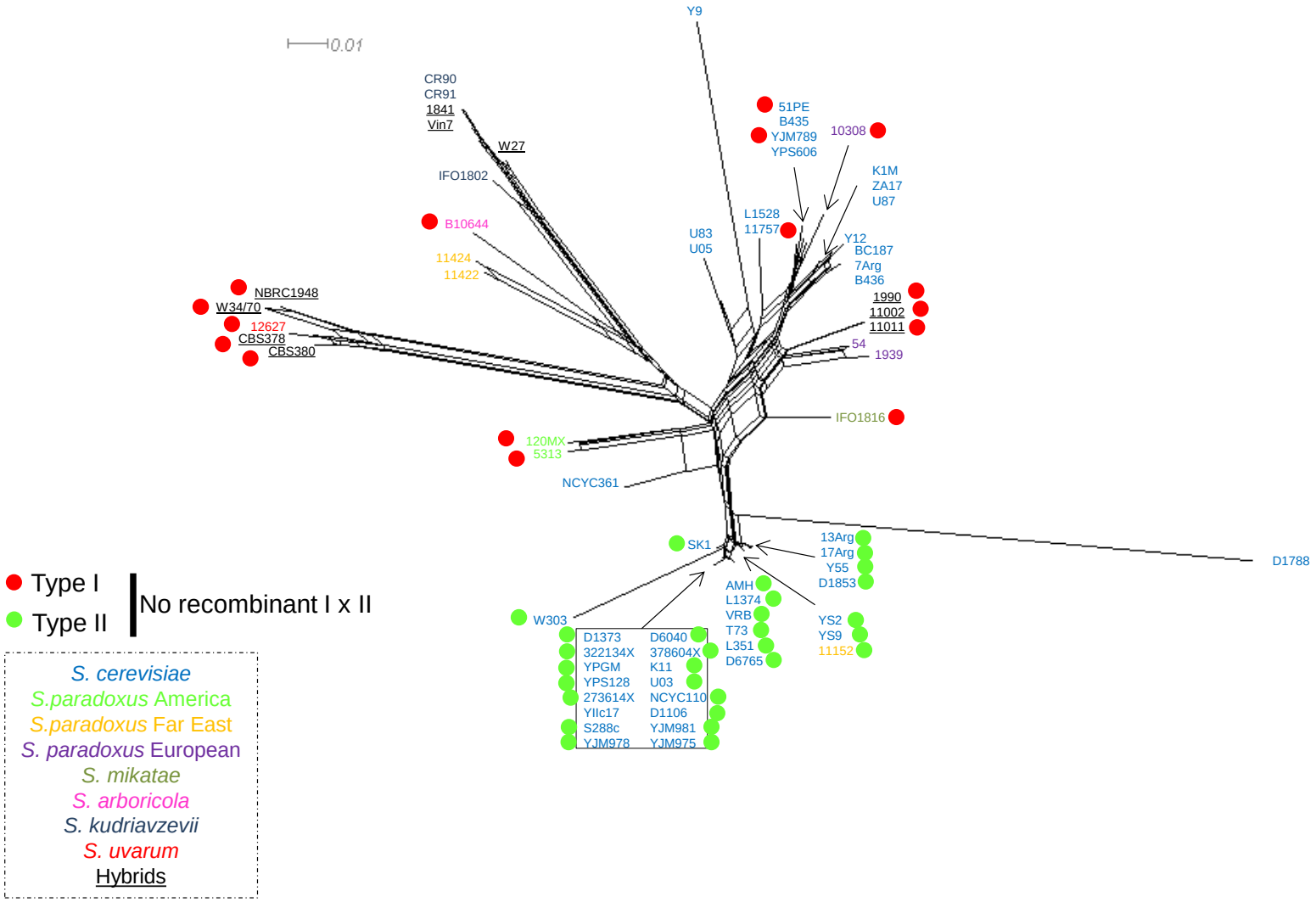
A NJ phylogenetic tree was reconstructed for the *ORF1* sequence alignment. Strains were colored according to the species designation. Bootstrap values above 50 are given for each branch. A red star represents a potential active *ORF1* based on the absence of GC clusters and premature stop codons which disrupt the coding sequence. Type I and Type II *ORF1* sequences, detected as non-recombinant, were colored in red and green, respectively. GC clusters in the *ORF1* sequence were represented by squares and the number represents the position in the *ORF1* alignment (see Figure 4). An arrow indicates that a particular GC cluster was inverted to infer the GC cluster family. Square colors represents GC cluster similarity according to the NJ trees from Figure S9 and S12.

Figure S6. *ORF1* aminoacid polymorphic sites.



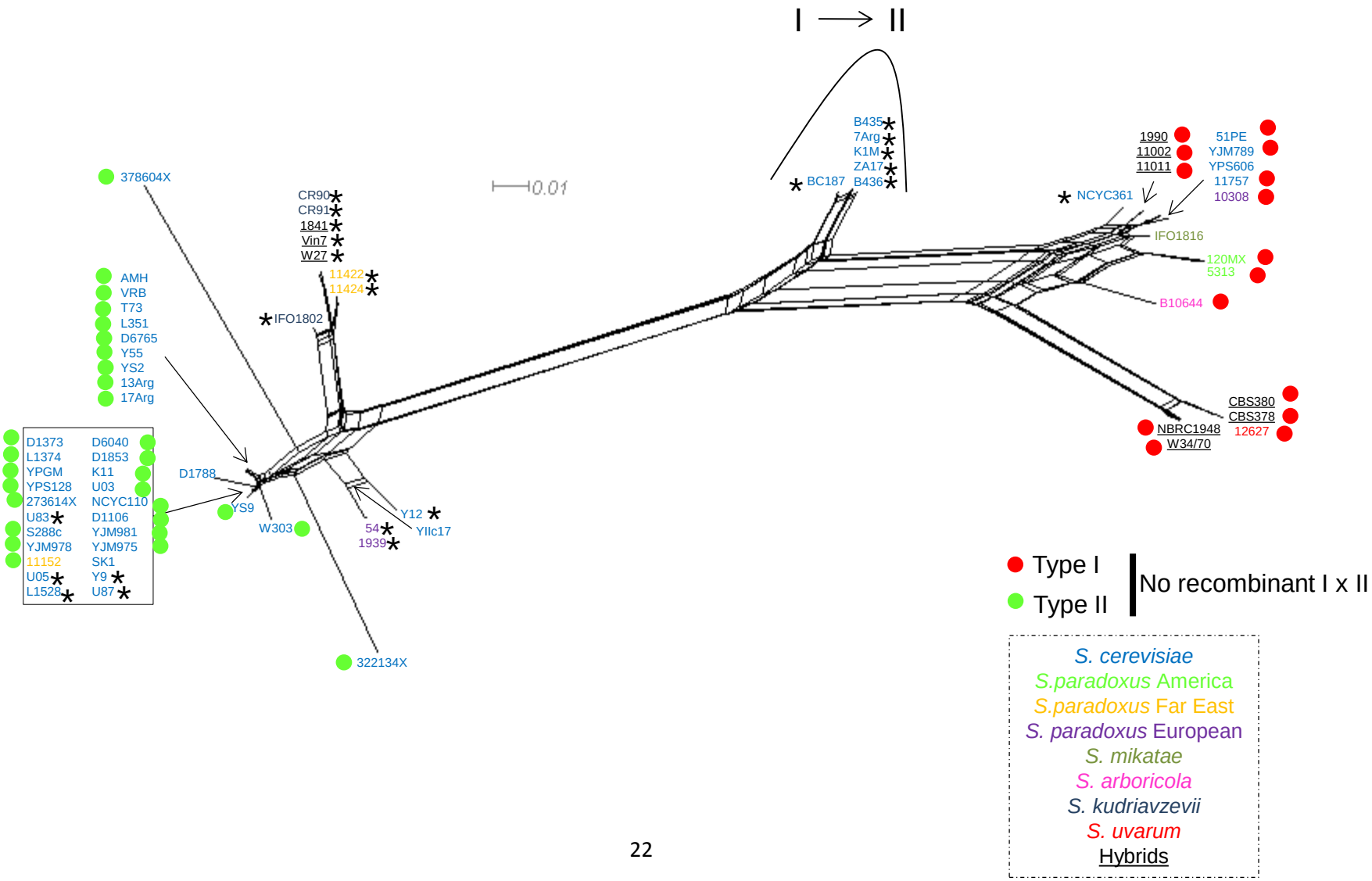
Variable *ORF1* aminoacid positions among haplotypes. Haplotype number are colored according to the species designation, as in Figure 2. Alignment aminoacid positions for each polymorphic site are also shown. The *COX2* Haplogroup designation is shown. Sequences were colored according to their similarity as inferred from Figure S7. Symbols * and ^ represent type I and type II *ORF1* sequences, respectively. Regions colored in yellow are sequences from unknown source. Lines indicate the sites corresponding to each alignment partition to reconstruct the *ORF1* phylogenetic networks by segments (Figure S7).

Figure S7. *ORF1* NN phylogenetic networks by segments.



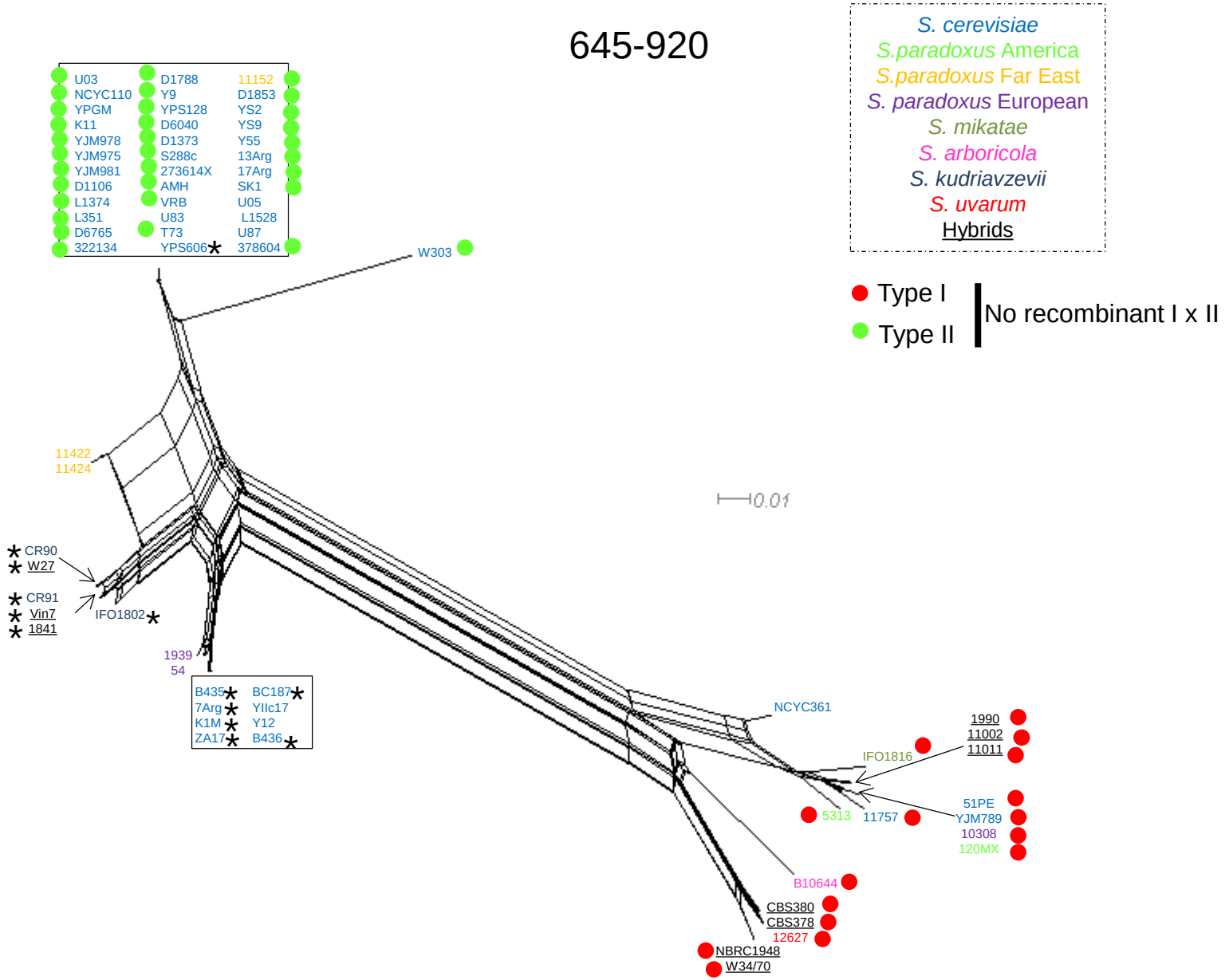
B

247-644


I → II
0.01
378604X
AMH
VRB
T73
L351
D6765
Y55
YS2
13Arg
17Arg
D1373
L1374
YPGM
YPS128
273614X
U83
S288c
YJM978
11152
U05
L1528
D6040
D1853
K11
U03
NCYC110
D1106
YJM981
YJM975
SK1
Y9
U87
CR90
CR91
1841
Vin7
W27
11422
11424
IFO1802
D1788
YS9
W303
Y12
YIIc17
54
1939
322134X
B435
7Arg
K1M
ZA17
B436
BC187
NCYC361
1990
11002
11011
51PE
YJM789
YPS606
11757
10308
IFO1816
120MX
5313
B10644
CBS380
CBS378
12627
NBRC1948
W34/70
Type I
Type II
No recombinant I x II
S. cerevisiae
S.paradoxus America
S.paradoxus Far East
S. paradoxus European
S. mikatae
S. arboricola
S. kudriavzevii
S. uvarum
Hybrids

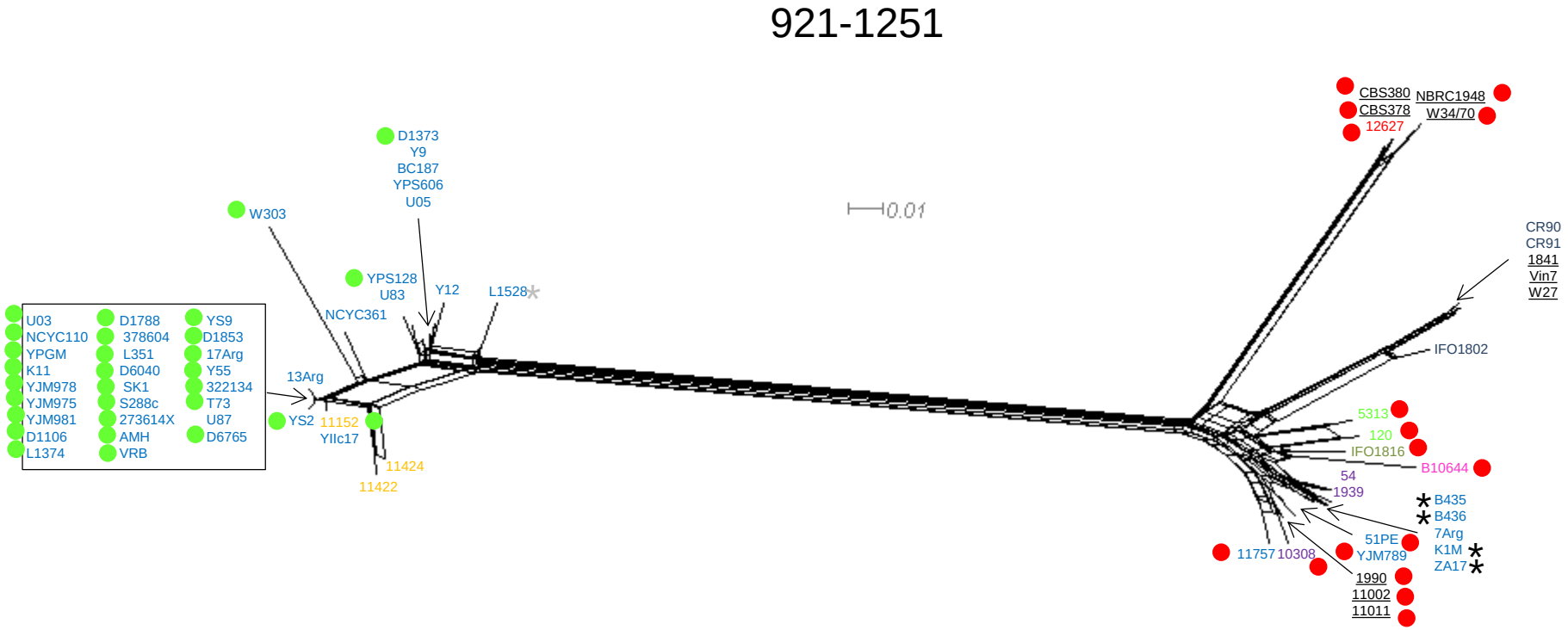
412

C


645-920
S. cerevisiae
S.paradoxus America
S.paradoxus Far East
S. paradoxus European
S. mikatae
S. arboricola
S. kudriavzevii
S. uvarum
Hybrids
Type I
Type II
No recombinant I x II
U03
NCYC110
YPGM
K11
YJM978
YJM975
YJM981
D1106
L1374
L351
D6765
322134
D1788
Y9
YPS128
D6040
D1373
S288c
273614X
AMH
VRB
U83
T73
YPS606
11152
D1853
YS2
YS9
Y55
13Arg
17Arg
SK1
U05
L1528
U87
378604
W303
11422
11424
0.01
CR90
W27
CR91
Vin7
1841
IFO1802
1939
54
B435
7Arg
K1M
ZA17
BC187
YIIc17
Y12
B436
NCYC361
IFO1816
1990
11002
11011
5313
11757
51PE
YJM789
10308
120MX
B10644
CBS380
CBS378
12627
NBRC1948
W34/70

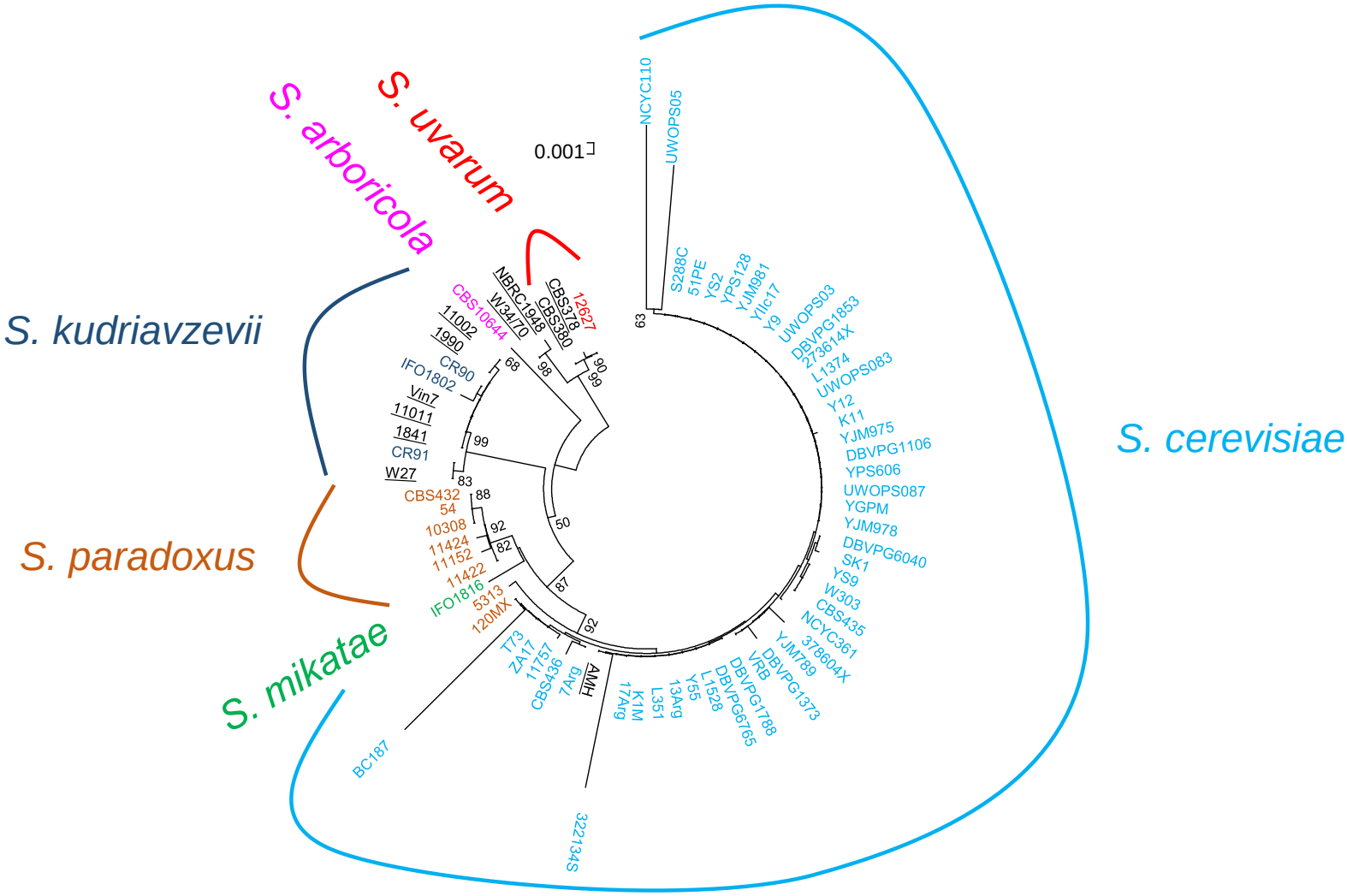
413

D


921-1251
0.01
CBS380
NBRC1948
CBS378
W34/70
12627
D1373
Y9
BC187
YPS606
U05
W303
YPS128
U83
Y12
L1528
NCYC361
13Arg
YS2
11152
YIIc17
11424
11422
U03
NCYC110
YPGM
K11
YJM978
YJM975
YJM981
D1106
L1374
D1788
378604
L351
D6040
SK1
S288c
273614X
AMH
VRB
YS9
D1853
17Arg
Y55
322134
T73
U87
D6765
CR90
CR91
1841
Vin7
W27
IFO1802
5313
120
IFO1816
B10644
54
1939
B435
B436
7Arg
K1M
ZA17
51PE
YJM789
11757
10308
1990
11002
11011
Type I
Type II
No recombinant I x II
S. cerevisiae
S.paradoxus America
S.paradoxus Far East
S. paradoxus European
S. mikatae
S. arboricola
S. kudriavzevii
S. uvarum
Hybrids

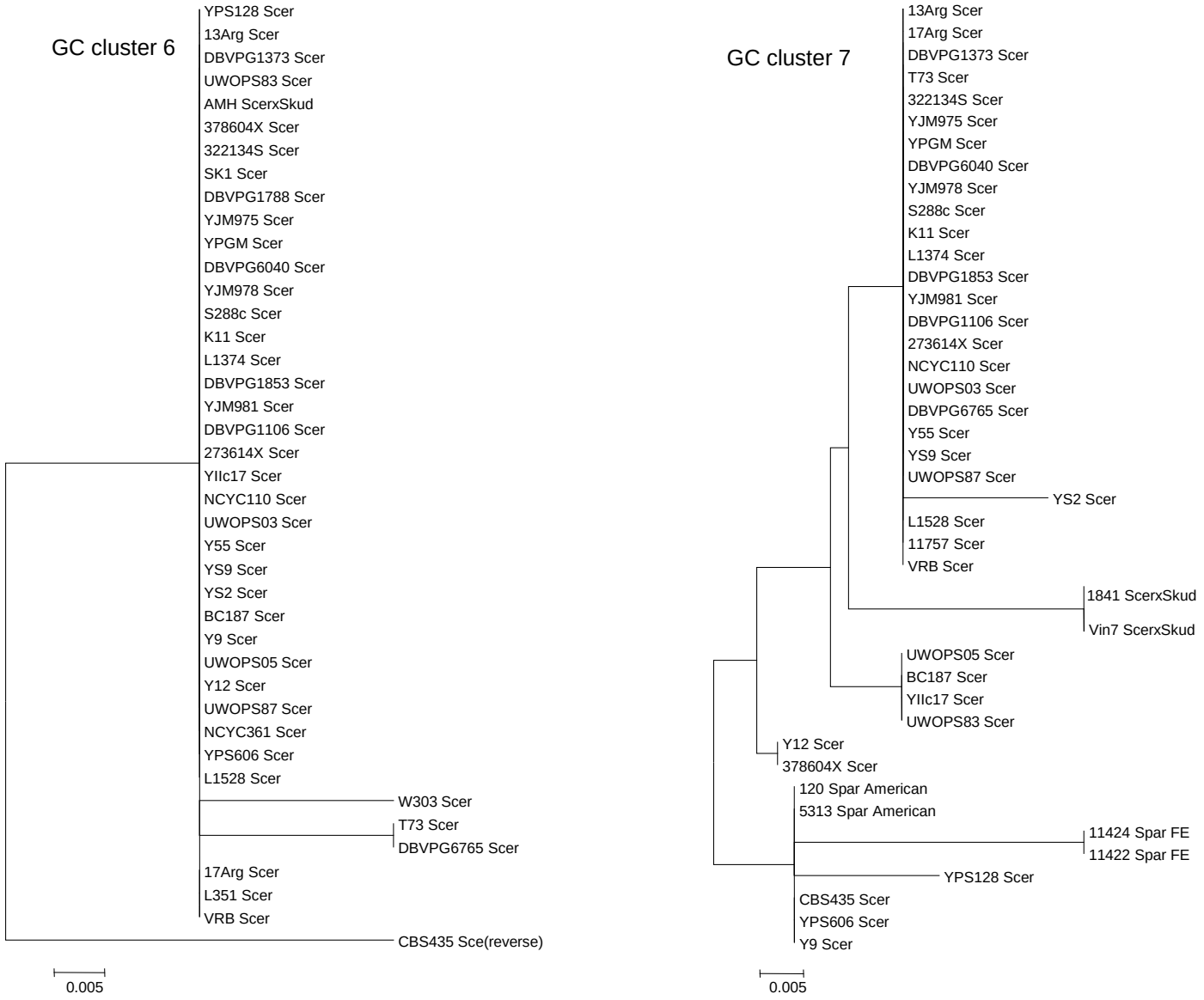
NN phylogenetic networks. for each *ORF1* alignment partitions, inferred by GARD and RDP, are shown in A), B), C) and D). A) shows the NN phylogenetic network corresponding to the *COX2* 497 nucleotide position until the position 246 of the *ORF1* alignment. Nucleotide positions from *ORF1* alignment are shown in figures B-D. Scale bar are given in nucleotide substitution per site. Type I and Type II *ORF1* sequences, detected as non-recombinant, are represented by red and green circles, respectively. Sequences are colored to each species designation. In figure B) some *S. cerevisiae* sequences are grouped in a I -> II group, indicating that the region used for inferring that NN phylonetwork contains a recombination point for those particular strains driving to the ambiguous position. Asterisks highlight sequences which clustering changed from one *ORF1* type to another due to its recombinant character.

Figure S8. *COX3* NJ phylogenetic tree.



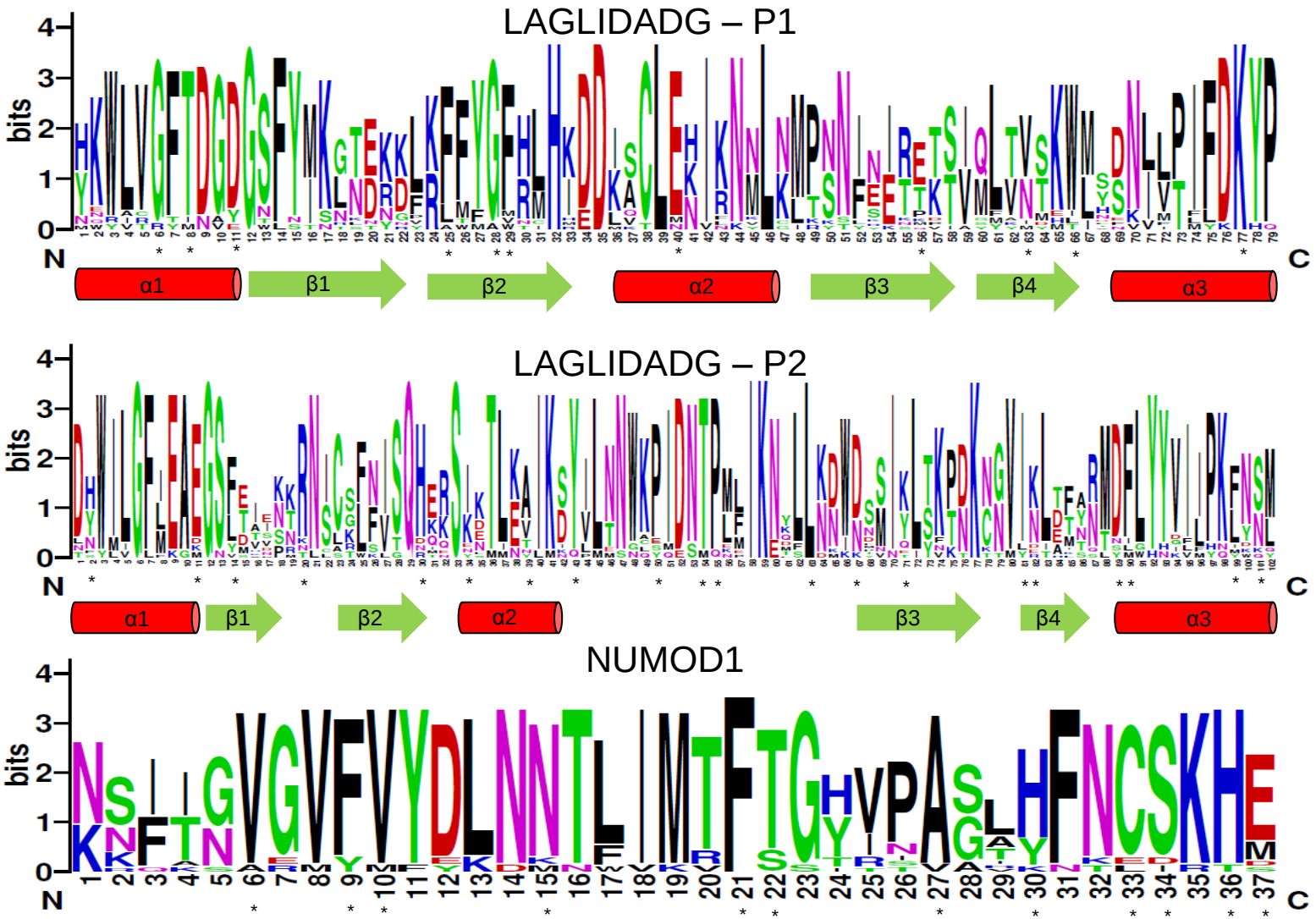
A *COX3* Neighbor-Joining phylogenetic tree is shown. Strains are colored according to the species designation. Bootstrap values above 50 are given for each branch. Scale is given in nucleotide substitution per site.

Figure S9. GC cluster 6 and 7 Neighbor-Joining phylogenetic trees.



To classify the GC cluster according to sequence similarity we reconstructed the NJ phylogenetic tree of GC cluster 6 and 7. This classification is shown by numbers in Figure S5. Scale bars represent number of substitutions per site.

Figure S10. Weblogo representation of LAGLIDADG and NUMOD domains.

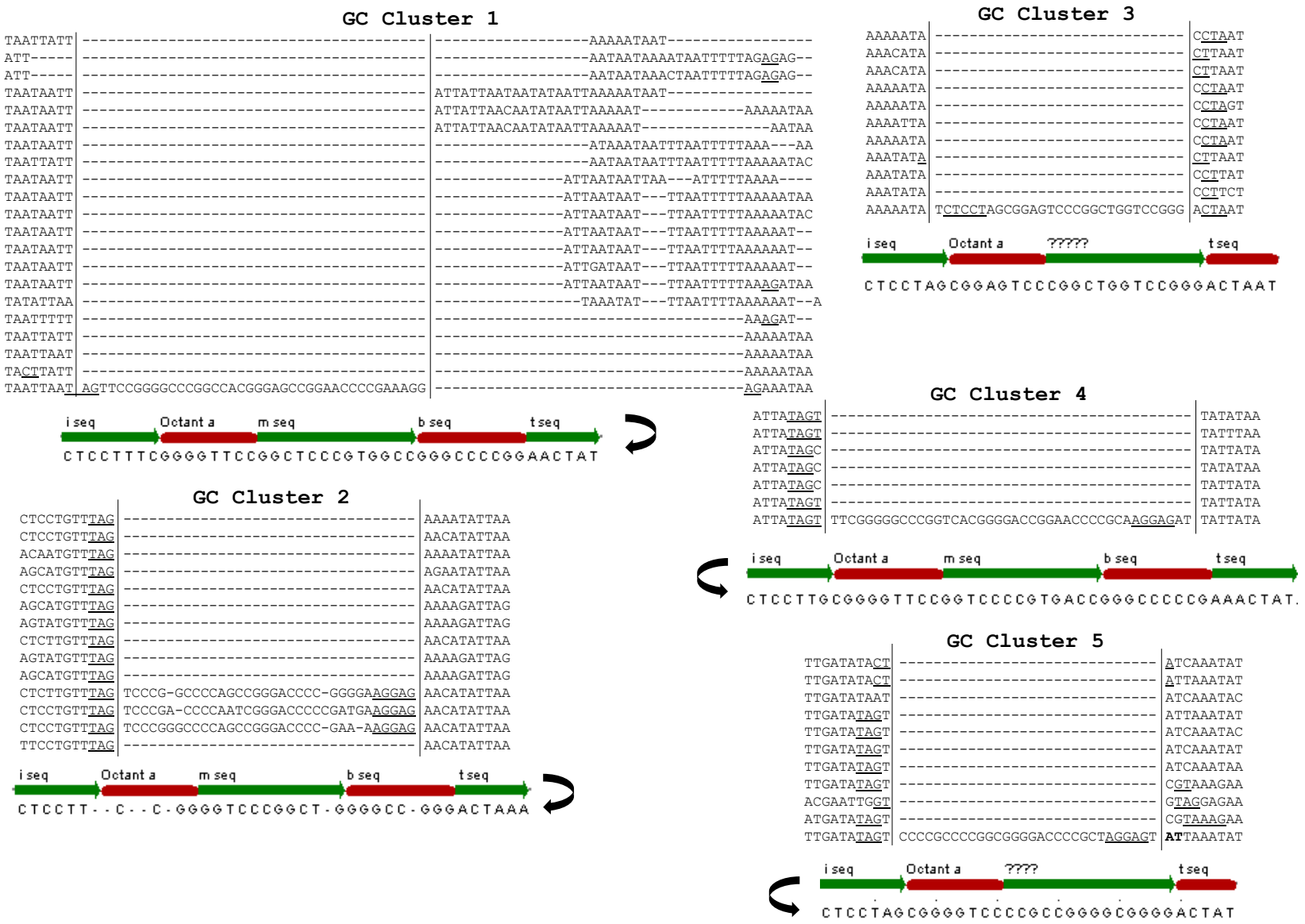


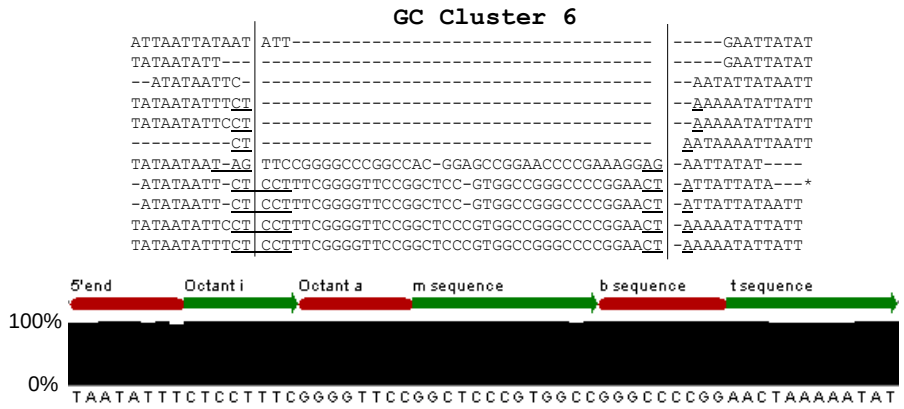
Weblogo (http://weblogo.berkeley.edu/) of the three homing endonuclease domains for *Saccharomyces ORF1* genes plus *SasefMp08* from *Kazchastania servazii*, and *ORF1* and ORF3 from *Williopsis saturnus* var. *suaveolens* are represented. Cylinders and arrows represent α-helix and β-sheet, as previously described (Dalgaard et al. 1997). Asterisk symbol indicates the codons for those aminoacids under purifying selection detected by `DataMonkey`.

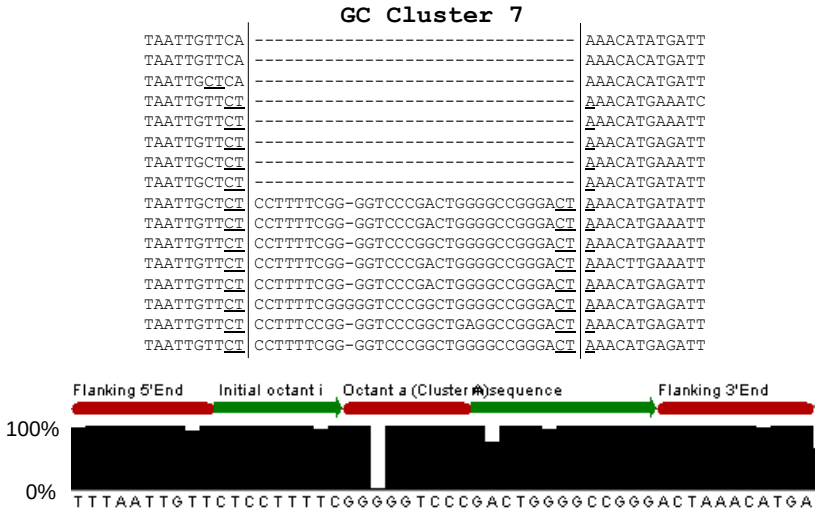
Figure S11. *COX2* and *ORF1* aminoacid alignment (Supplementary File S11).

The complete *COX2* and *ORF1* aminoacid alignments are shown. Protein secondary structure and domains are indicated for *ORF1.* Jalview also display the alignment quality (based on BLOSUM 62), physicochemical conservation calculated according to Livingstone and Barton (1993), and consensus sequence.

Figure S12. GC clusters from M1 family insertion regions.







443

444 The Flanking regions were the seven GC clusters were found are shown. In the case of GC cluster 6 and 7 a consensus sequence is

445 shown and bars show the percentage of sequences showing a particular nucleotide substitution. An arrow indicates that the GC cluster

446 sequence was inverted to annotate its structure. Question marks indicate an unknown GC cluster structure.

447