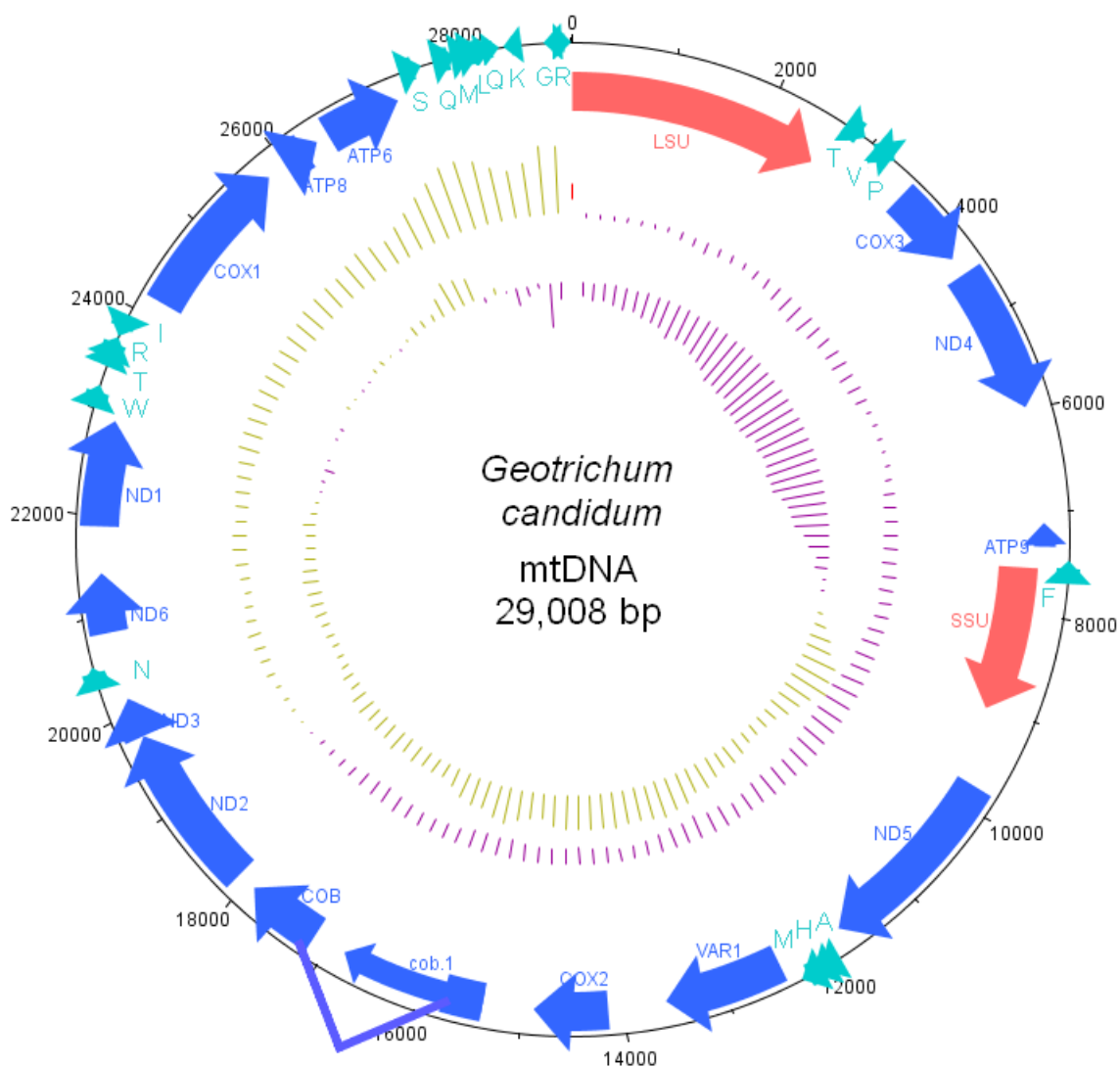


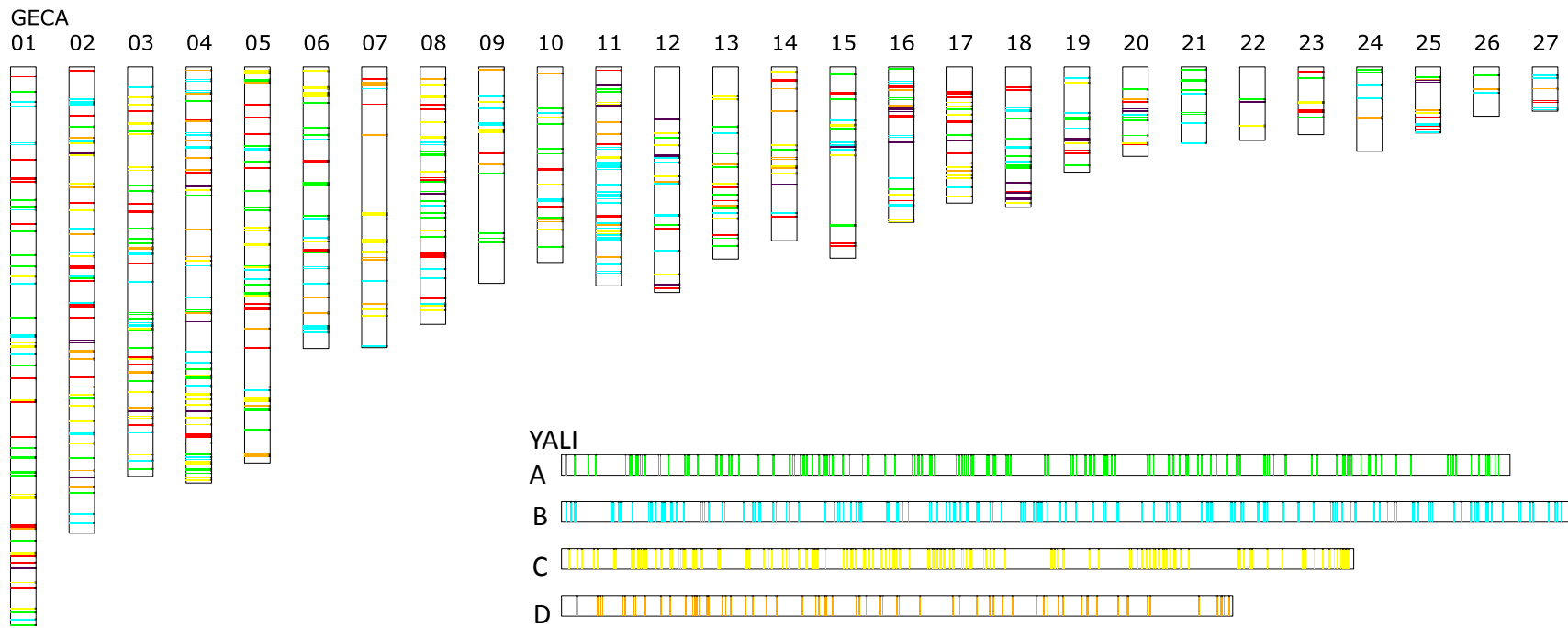
Differential gene retention as an evolutionary mechanism to generate biodiversity and adaptation in yeasts

Guillaume Morel, Lieven Sterck, Dominique Swennen, Marina Marcet-Houben, Djamilia Onesime, Anthony Levasseur, Noémie Jacques, Sandrine Mallet, Arnaud Couloux., Karine Labadie, Joëlle Amselem, Jean-Marie Beckerich, Bernard Henrissat, Yves Van de Peer, Patrick Wincker, Jean-Luc Souciet, Toni Gabaldón, Colin R. Tinsley & Serge Casaregola

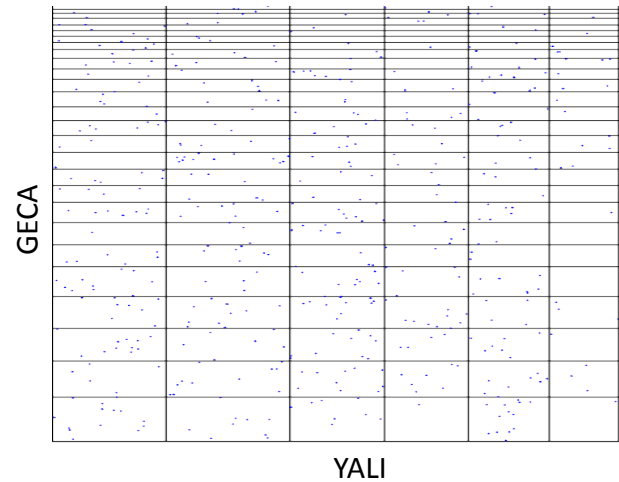
Supplementary figures

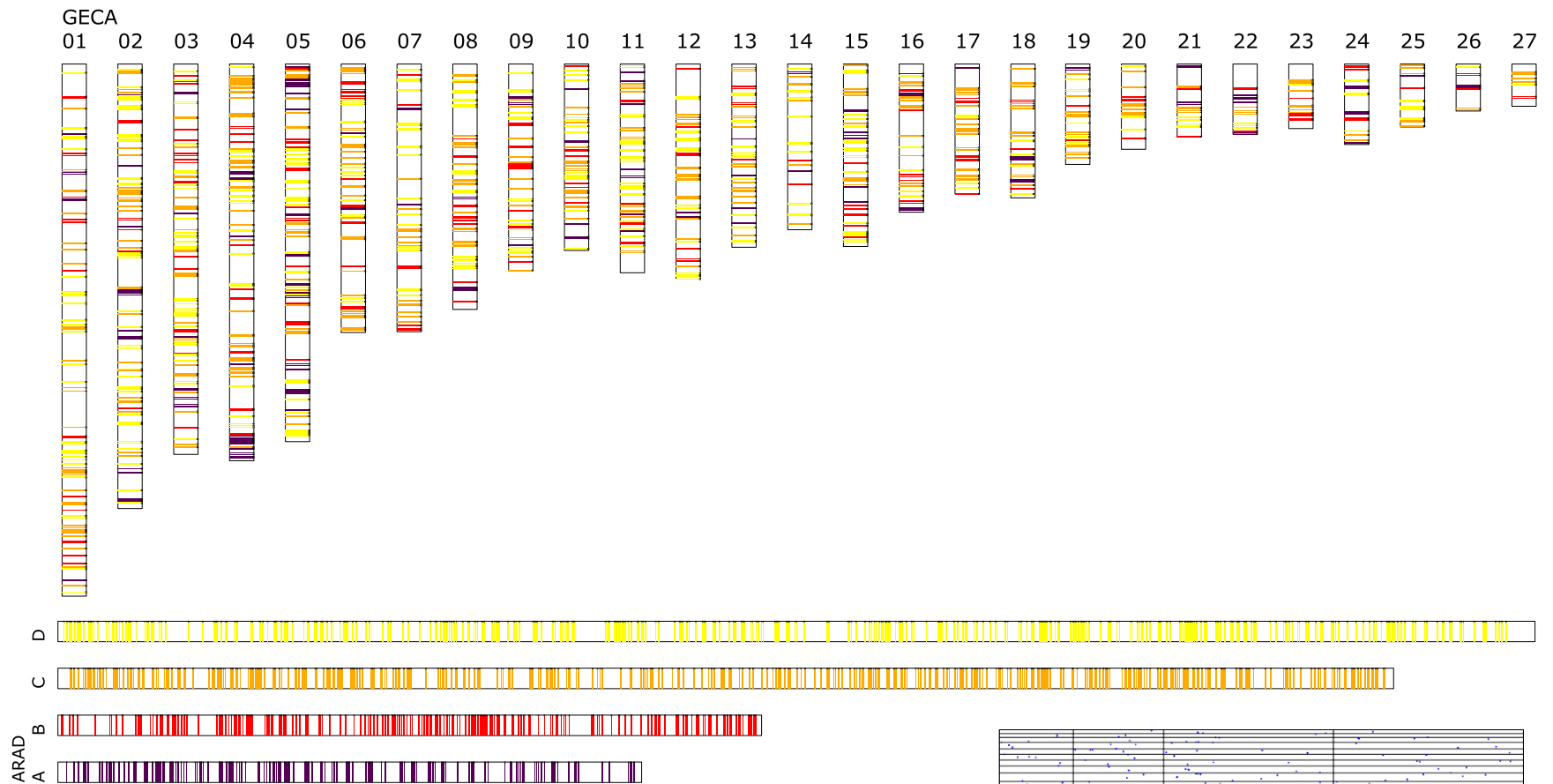


Supplementary Figure S1. **Circular map of *Geotrichum candidum* mtDNA.** Map was drawn using dnaplotter (<https://www.sanger.ac.uk/.../dnaplotter/>). Protein coding gene and ribosomal gene models were detected using blastX against the available Saccharomycotina mtDNAs. tRNA genes were detected using tRNAscan-SE. Protein coding genes are indicated by large blue arrows, rDNA genes by red arrows, tRNAs by green arrows with charged amino acid indicated in one letter code. The intronic ORFcob.1 is indicated by a thin blue arrow and the link between the two COB exons is shown in blue. GC% variations (outer bars) and GC skew (inner bars) are indicated in mauve (below average) and green (above average).

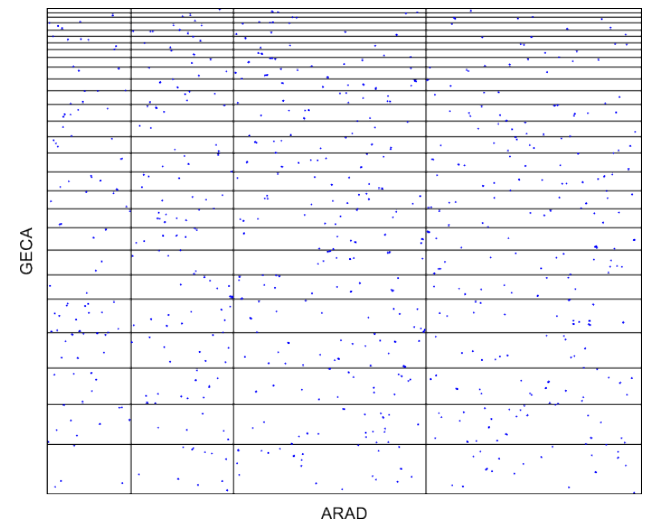


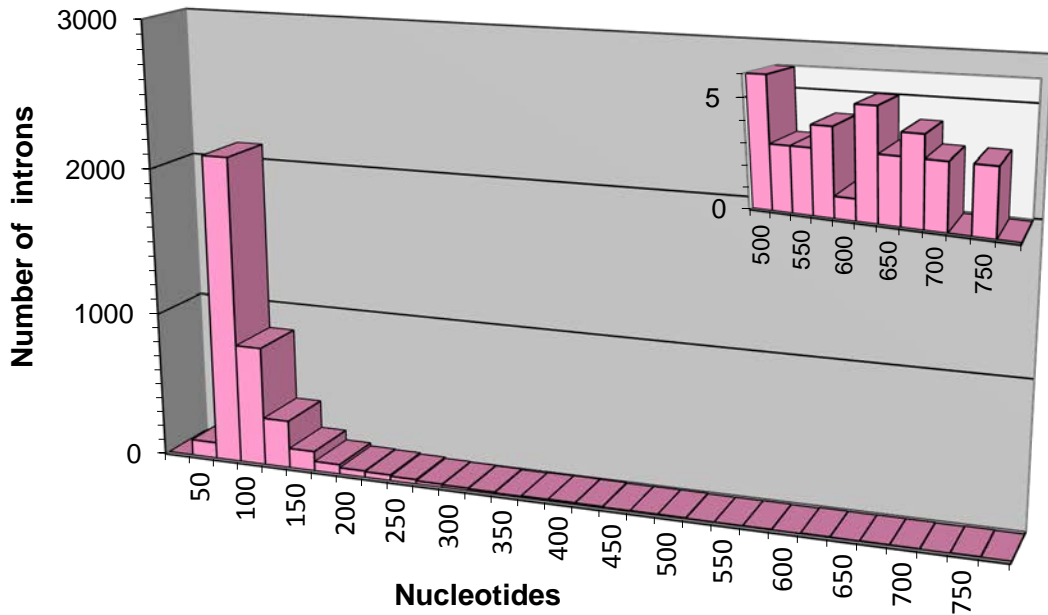
Supplementary Figure S2a. **Synteny between *G. candidum* (GECA) and *Y. lipolytica* (YALI)**. Top: distribution in the GECA scaffolds of YALI orthologs defined on the basis of bidirectional hits in blastp comparison. Each color is attributed to the genes of one YALI chromosome. Right: dotplot representation.



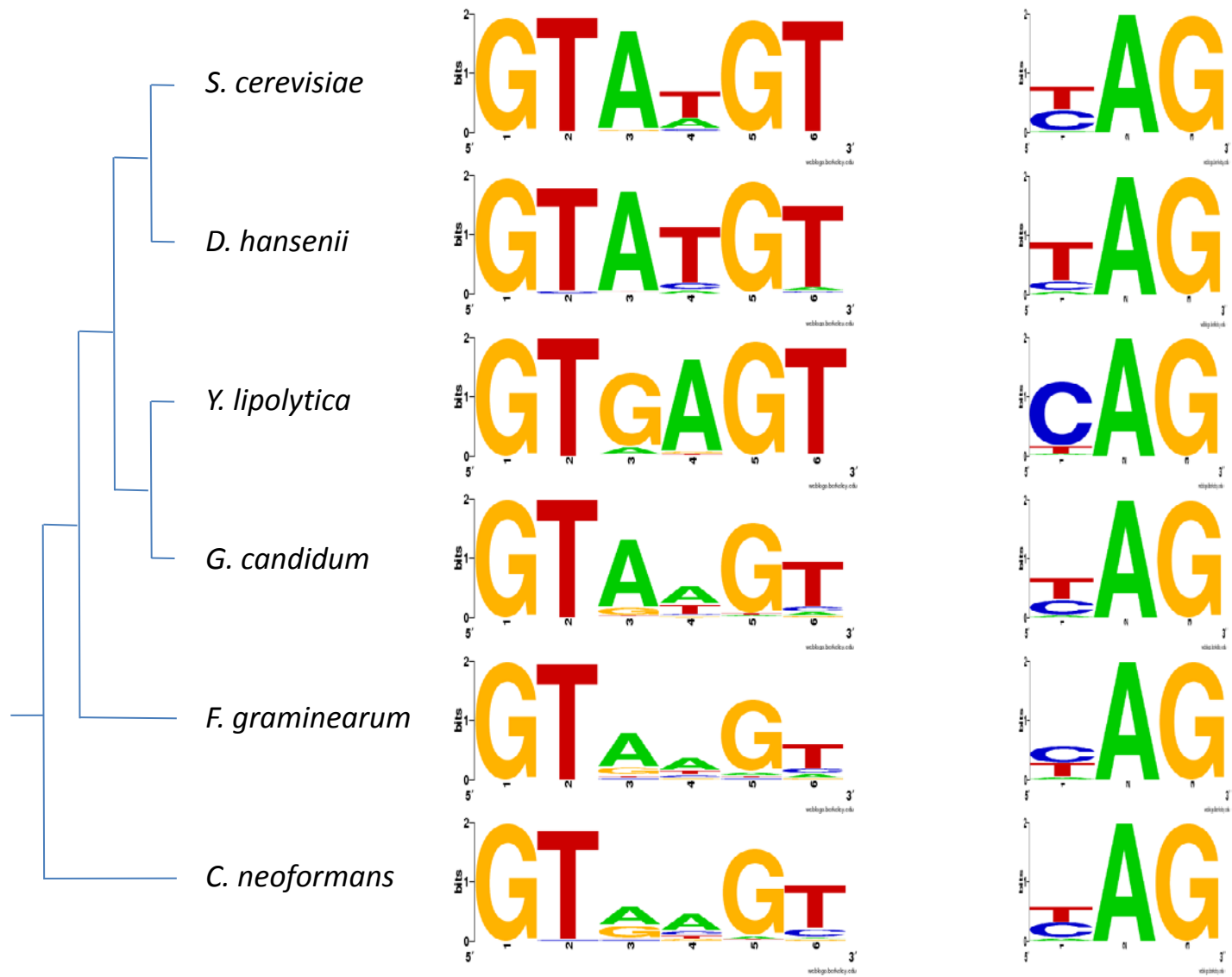


Supplementary Figure S2b. **Synteny between *G. candidum* (GECA) and *B. adenivorans* (ARAD)**. Top: distribution in the GECA scaffolds of ARAD orthologs defined on the basis of bidirectional hits in blastp comparison. Each color is attributed to the genes of one ARAD chromosome. Right: dotplot representation.

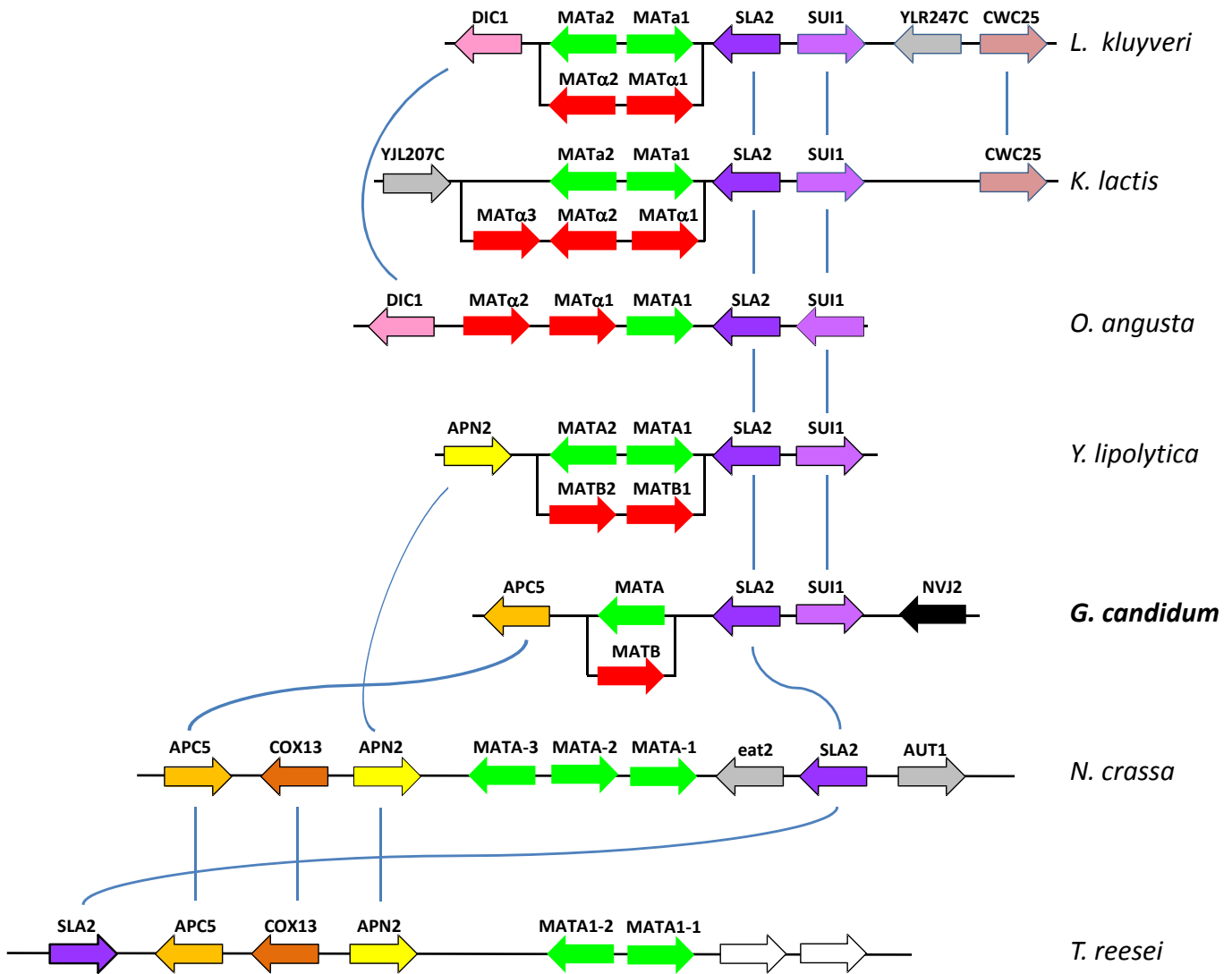




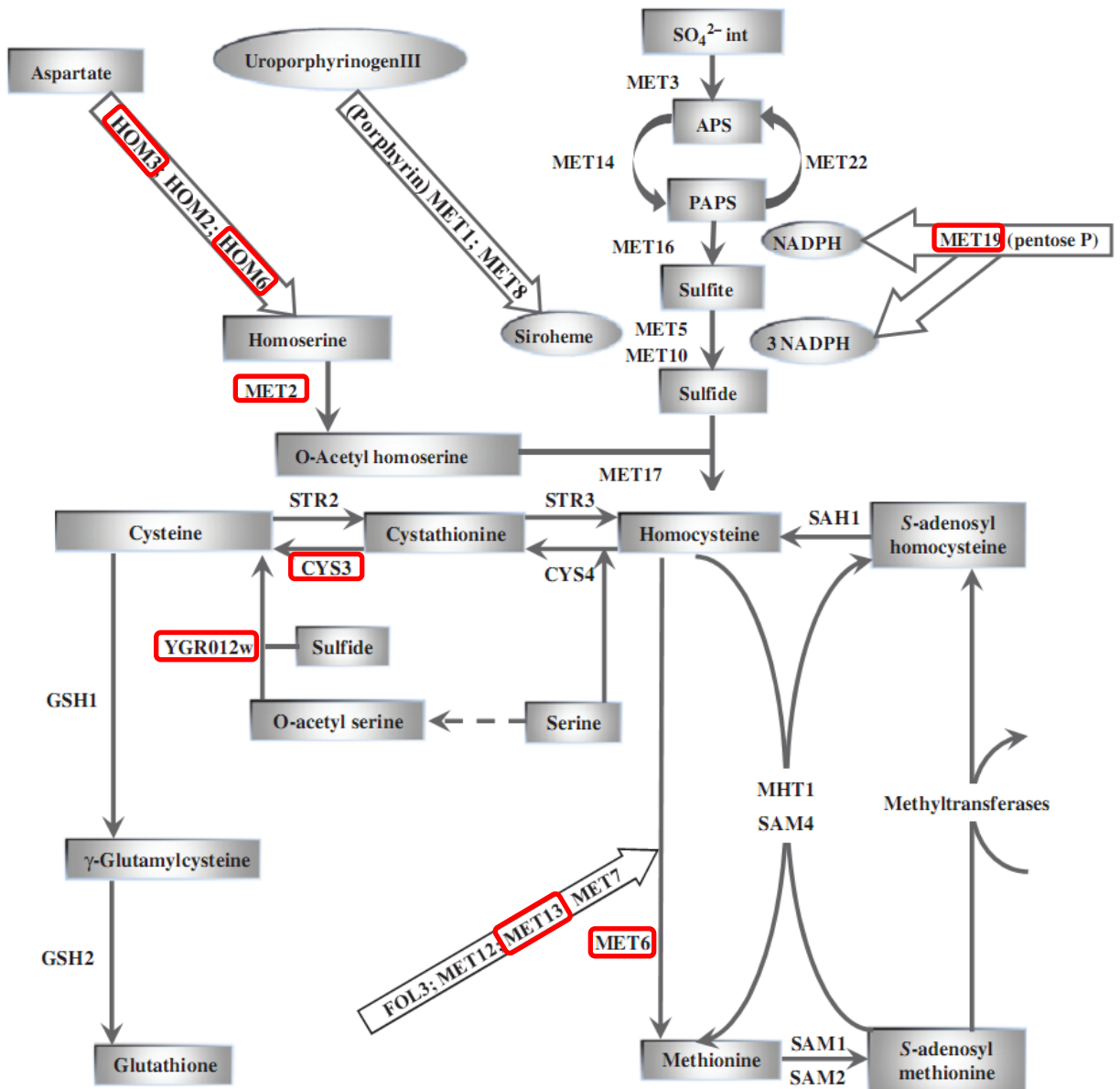
Supplementary Figure S3. **Distribution of intron length in *G. candidum*.** The number of introns is plotted against their length in nucleotides. The inset is an expansion of the higher end of the intron length distribution, demonstrating a continuation of the distribution by a small number of unusually long introns.



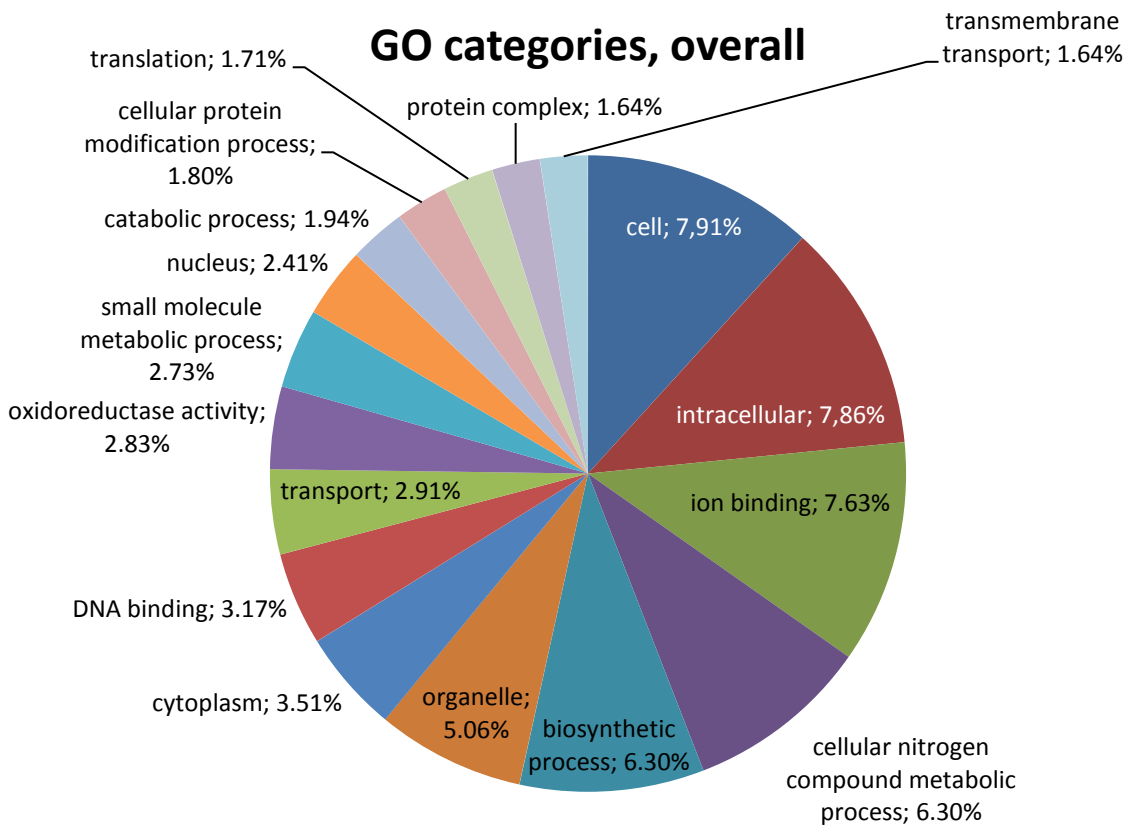
Supplementary Figure S4. **Conservation of 5' (left) and 3' (right) splice sites in *G. candidum* and various yeasts and fungi.** The consensus sequences were obtained using weblogo (<http://weblogo.berkeley.edu/>) version 2.8.2 . The sequences from *S. cerevisiae* were extracted from the Saccharomyces Genome Database (<http://www.yeastgenome.org/>) and those from *Y. lipolytica*, *D. hansenii*, *F. graminearum* and *C. neoformans* from NCBI (<http://www.ncbi.nlm.nih.gov/genome/>).



Supplementary Figure S5. **Organization of the mating type loci and their surroundings in *G. candidum* and in six fungal species.** Gene names are indicated above the genes. Orthologs are represented with the same color. MATA(a) genes are represented in green, MATB(α) genes are represented in red. Orientation of genes on chromosomes (black lines) are indicated. Data are from Butler *et al.* (2004) and Seidl *et al.* (2009).

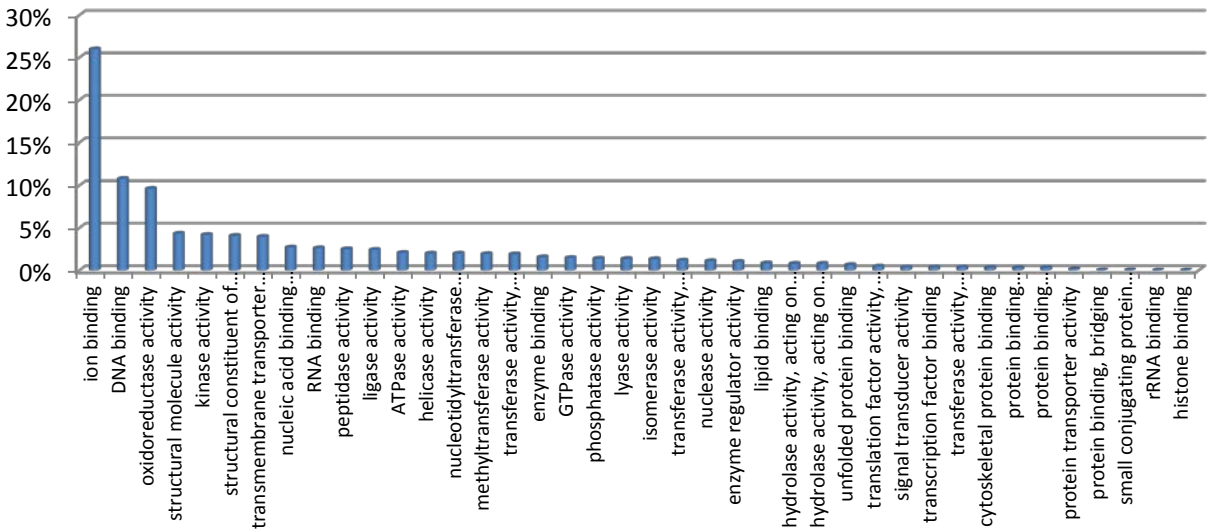


Supplementary Figure S6. **Duplicated genes of the sulfur metabolism pathway in *G. candidum*.** The scheme was taken from Hebert *et al.* (2011). Duplicated genes in *G. candidum* are boxed in red.



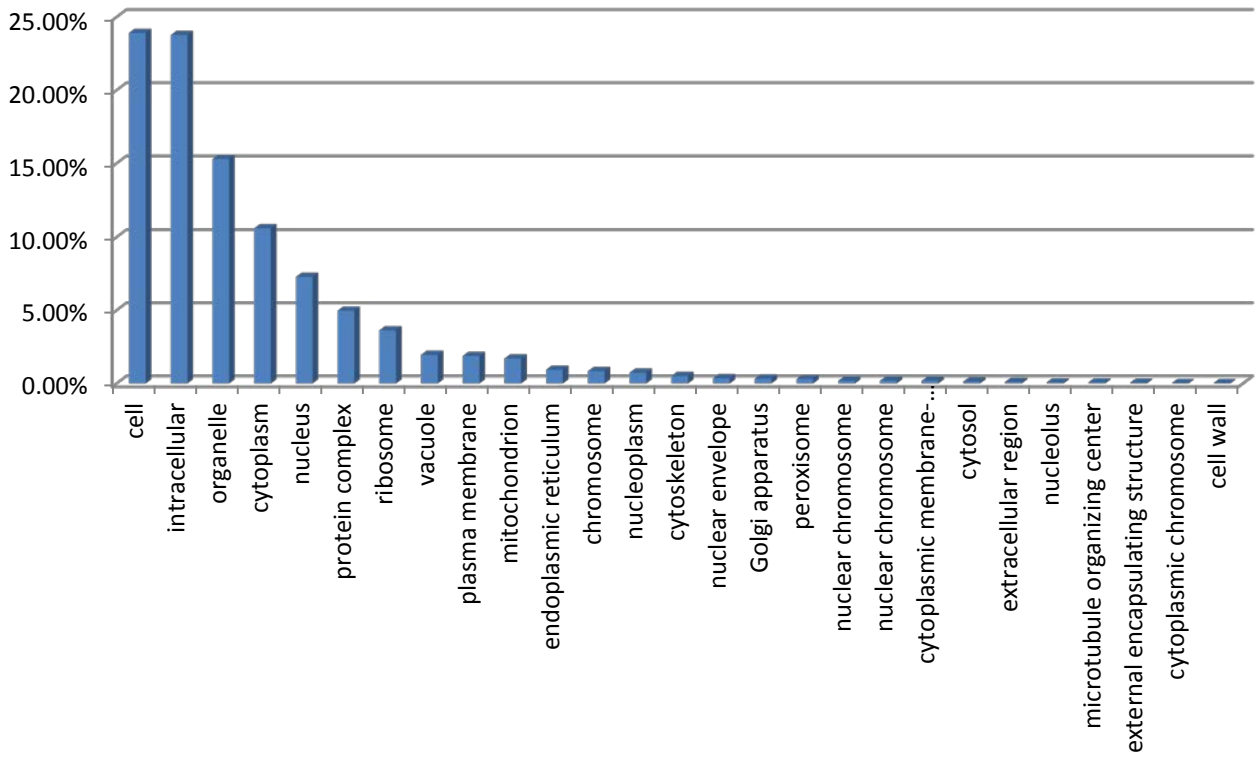
Supplementary Figure S7a. Overall distribution of predicted functions in the genome of *G. candidum* in Gene ontology categories.

Molecular Function



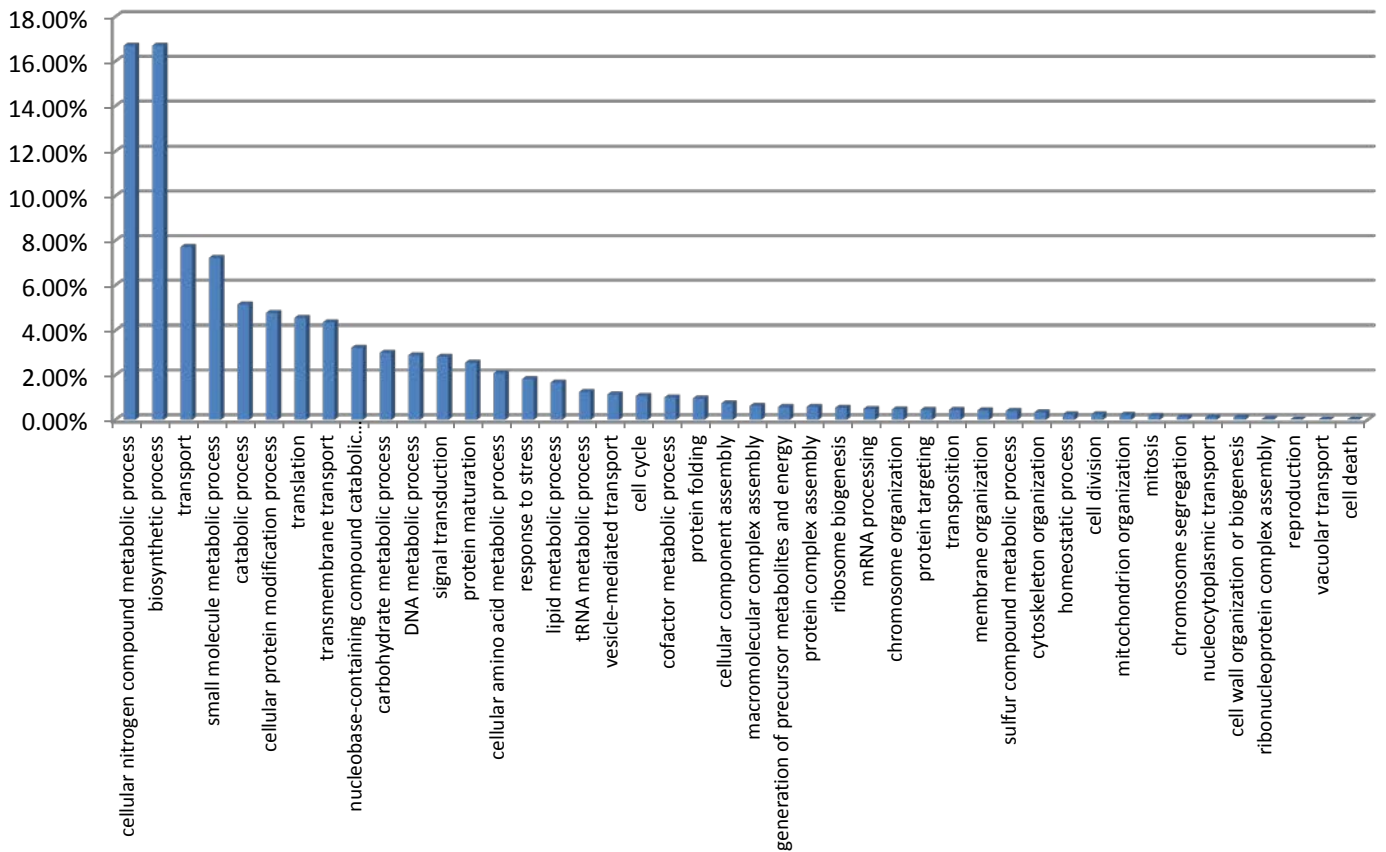
Supplementary Figure S7b. Distribution of predicted molecular functions in the genome of *G. candidum* in Gene ontology categories.

Cellular Component



Supplementary Figure S7c. **Distribution of predicted functions in the genome of *G. candidum* in Gene ontology Cellular component categories.**

Biological Process



Supplementary Figure S7d. **Distribution of predicted functions in the genome of *G. candidum* in Gene ontology Biological process categories.**

Supplementary Fig S8. **Phylogenetic trees from possible cases of HGT extracted from the *G. candidum* phylum.**

Fig S8a

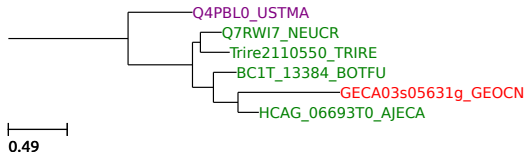


Fig S8b

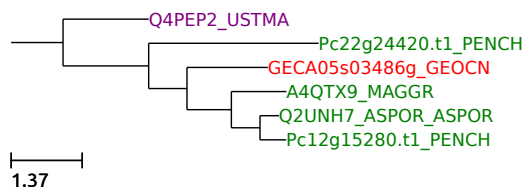


Fig S8c

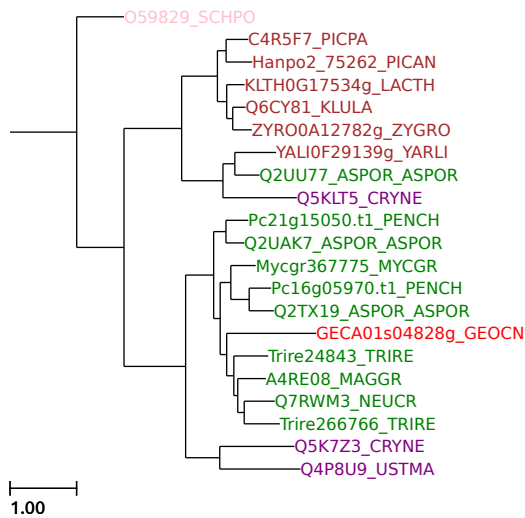


Fig S8d

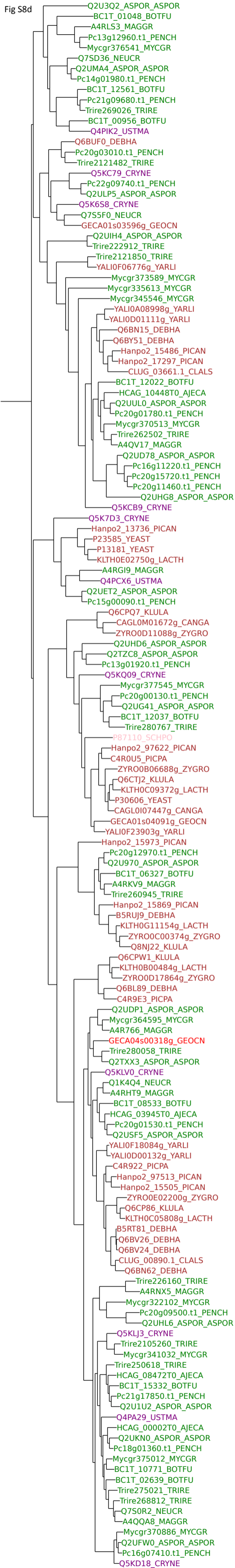
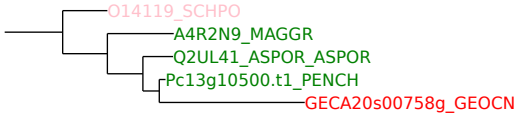
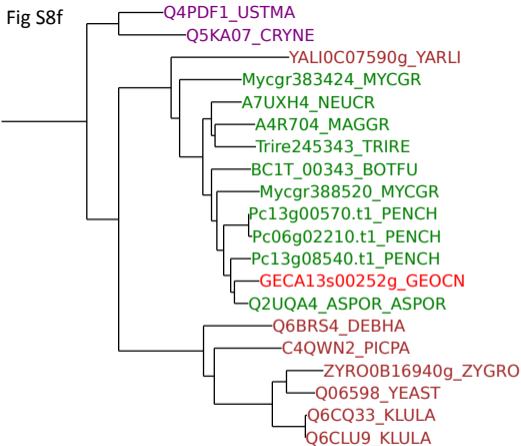


Fig S8e



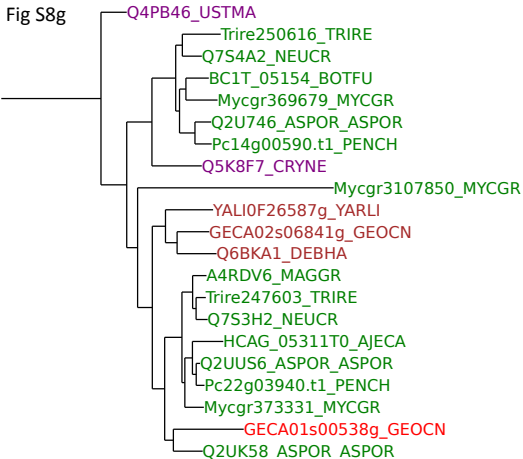
1.16

Fig S8f



0.77

Fig S8g



0.65

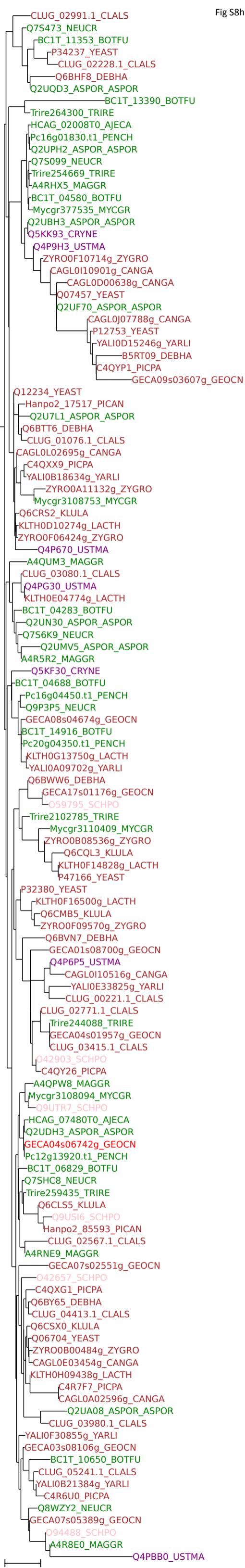
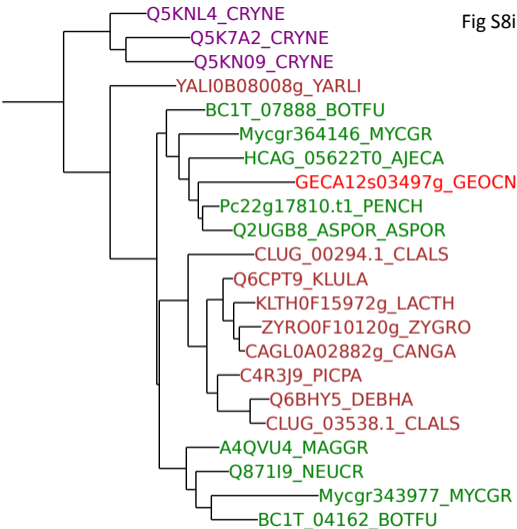
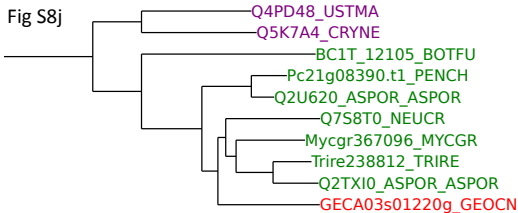


Fig S8i



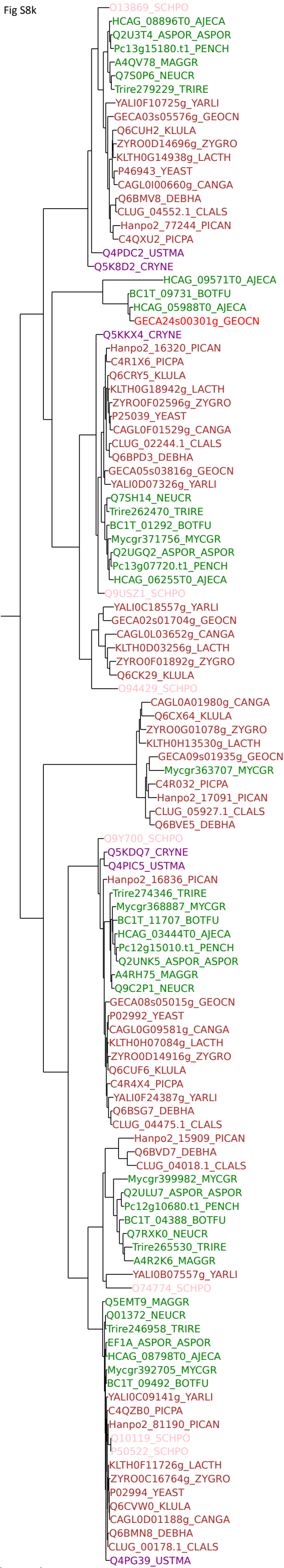
1.16

Fig S8j



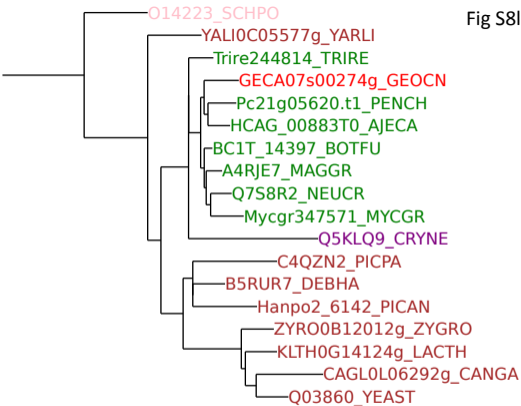
0.71

Fig S8k



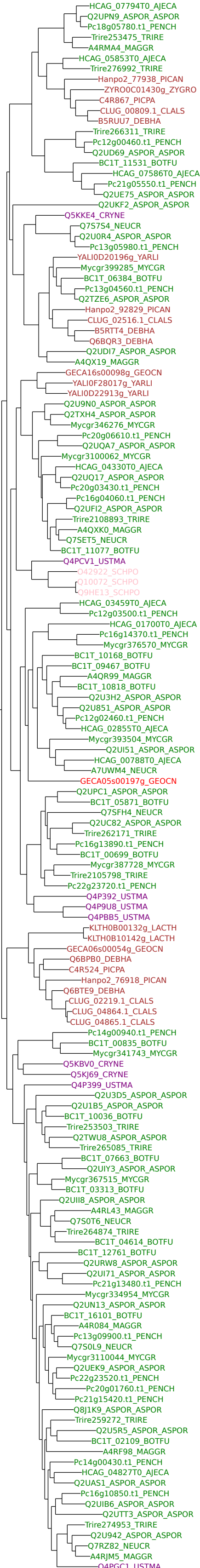
2.07

Fig S8l



0.81

Fig S8m



0.77

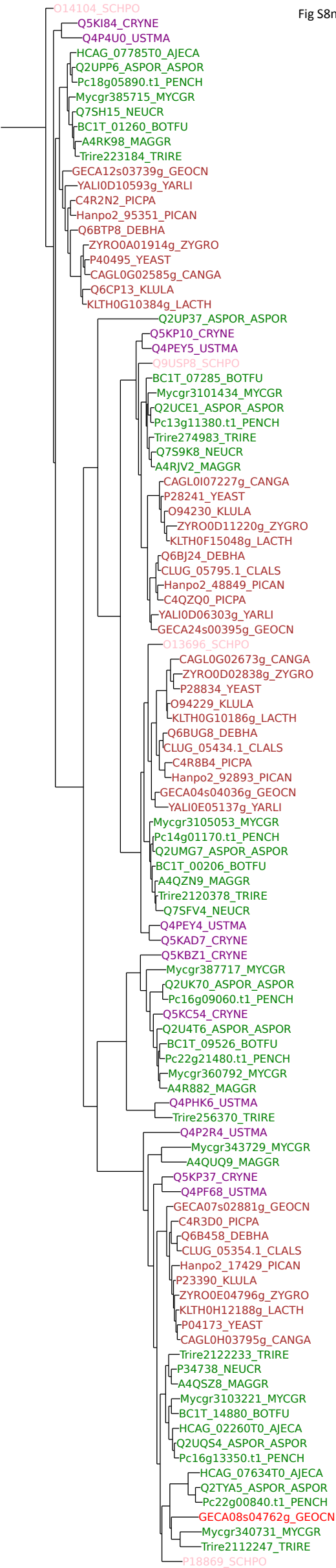
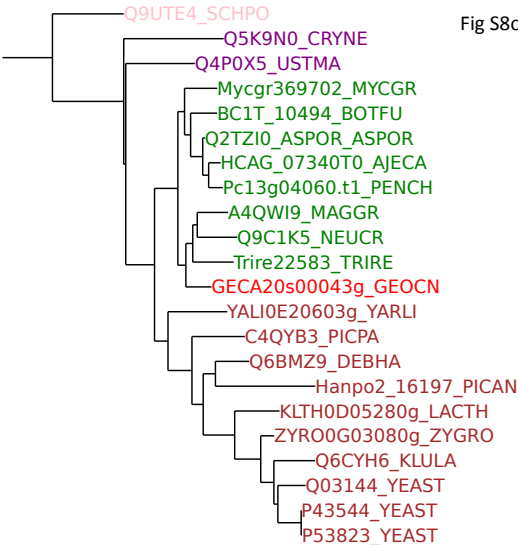
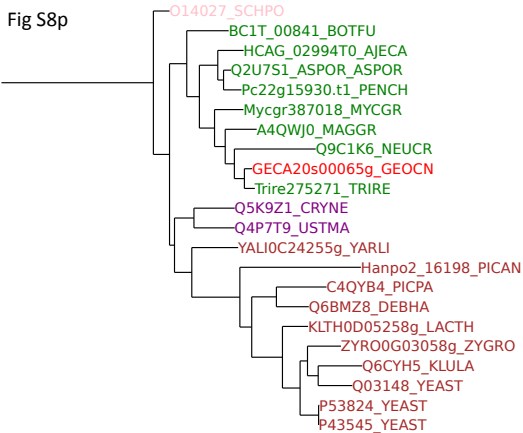


Fig S8o

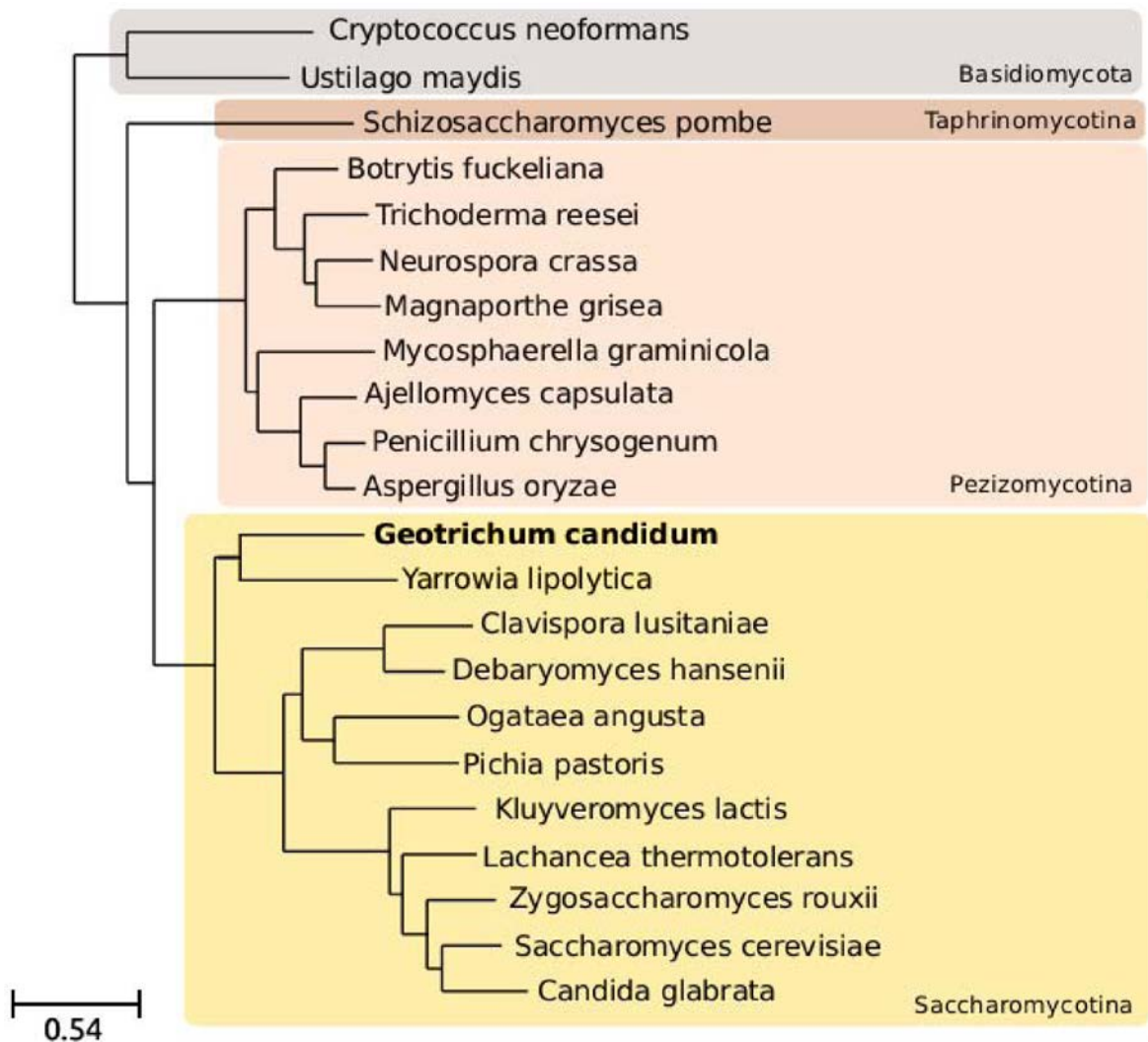


1.46

Fig S8p



0.38



Supplementary Figure S9. **Species tree reconstruction from the *G. candidum* phylome.** The tree species was reconstructed using the RAxML program as described in Materials and Methods. Proteins with a one-to-one orthology relationship to all the considered species were selected from the *G. candidum* phylome. A total of 302 protein alignments were concatenated into a multiple sequence alignment. The final alignment contained 170,787 amino acids. The bar indicates the scale of branch length.

Differential gene retention as an evolutionary mechanism to generate biodiversity in yeasts

Guillaume Morel, Lieven Sterck, Dominique Swennen, Marina Marcet-Houben, Djamila Onesime, Anthony Levasseur, Noémie Jacques, Sandrine Mallet, Arnaud Couloux., Karine Labadie, Joëlle Amselem, Jean-Marie Beckerich, Bernard Henrissat, Yves Van de Peer, Patrick Wincker, Jean-Luc Souciet, Toni Gabaldón, Colin R. Tinsley & Serge Casaregola

Supplementary tables

Supplementary Table 1. **Genome overview**

Scaffold	Size (bp)	Number of gaps	Total gap size (bp)	Number of protein coding genes (*)	Number of pseudogenes	Number of tRNA genes
scaffold 01	2,581,591	134	209,110 (8.10%)	668	13	36
scaffold 02	2,106,585	124	178,429 (8.47%)	558	6	20
scaffold 03	1,796,091	87	143,169 (7.97%)	492	4	33
scaffold 04	1,686,461	74	110,146 (6.53%)	498	5	18
scaffold 05	1,653,926	83	92,238 (5.57%)	474	1	16
scaffold 06	1,264,337	60	108,269 (8.56%)	338	0	15
scaffold 07	1,211,666	64	104,510 (8.62%)	336	2	22
scaffold 08	1,159,651	53	91,539 (7.89%)	309	5	15
scaffold 09	1,063,388	50	73,588 (6.92%)	259	4	11
scaffold 10	918,757	61	121,292 (13.20%)	234	2	22
scaffold 11	909,630	50	76,909 (8.45%)	262	2	9
scaffold 12	890,415	43	56,095 (6.29%)	270	2	5
scaffold 13	871,522	38	59,148 (6.78%)	230	5	5
scaffold 14	795,198	46	74,252 (9.33%)	208	2	12
scaffold 15	711,411	28	49,387 (6.94%)	229	2	13
scaffold 16	665,493	32	49,807 (7.48%)	186	3	12
scaffold 17	608,101	36	57,666 (9.48%)	163	2	7
scaffold 18	562,025	20	41,863 (7.44%)	168	3	5
scaffold 19	507,124	22	38,528 (7.59%)	126	2	12
scaffold 20	368,852	14	19,981 (5.41%)	107	1	4
scaffold 21	313,304	13	17,857 (5.69%)	91	2	5
scaffold 22	313,014	13	17,044 (5.44%)	88	4	1
scaffold 23	307,402	13	22,984 (7.47%)	81	1	6
scaffold 24	271,541	13	16,510 (6.08%)	101	2	8
scaffold 25	237,622	17	22,898 (9.63%)	79	0	3
scaffold 26	220,840	10	15,595 (7.06%)	59	1	2
scaffold 27	219,481	12	13,837 (6.30%)	53	2	7
scaffold32	620,629	206	61,001 (9.82%)	137	2	25
Total	24,836,057	1416	1,929,829 (7.77%)	6804	80	349

(*) excluding pseudogenes and transposons

Supplementary Table 2. List of *G. candidum* genes encoding NAD dehydrogenase complex I proteins and their corresponding orthologs in *Y. lipolytica*

Complex I proteins	<i>Geotrichum candidum</i>	<i>Yarrowia lipolytica</i>
NUAM	GECA17s02276g/GECA08s04113g	YALIOD05467g
NUBM	GECA04s04729g	YALIOB20372g
NUCM	GECA09s04641g	YALIOF17248g
NUGM	GECA02s07666g	YALIOF02123g
NUHM	GECA02s06929g	YALIOD00737g
NUIM	GECA10s03684g	YALIOF00924g
NUKM	GECA01s00098g	YALIOF06050g
NUEM	GECA12s01440g	YALIOD24585g
ST1	GECA04s07567g	YALIOF23551g
NESM	GECA04s01517g	YALIOE29095g
NUJM	GECA01s05532g	YALIOE11891g
NUZM	GECA18s02386g	YALIOA02651g
NUPM	GECA05s05884g	YALIOA20680g
NUXM	GECA10s02782g	YALIOE28424g
N7BM	GECA02s04916g	YALIOB00792g
NUYM	GECA07s00098g	YALIOB14861g
NUFM	GECA03s07314g	YALIOE23089g
NIAM	GECA03s00571g	YALIOD04939g
NB4M	GECA13s02562g	YALIOA01419g
NUMM	GECA04s05059g	YALIOD19030g
NB6M	GECA01s02859g	YALIOE28930g
NUNM	No hit	YALIOF14003g
NI2M	GECA09s00604g	YALIOD07216g
NB8M	GECA11s01715g	YALIOE31766g
NIDM	GECA05s00164g	YALIOA17946g
NB5M	GECA07s00719g	YALIOF06061g
ACPM1	GECA03s01275g	YALIOD24629g
ACPM2	GECA07s04509g	YALIOD14850g
NIPM	GECA20s01044g	YALIOF18359g
NIMM	GECA08s03189g	YALIOC21789g
NI8M	GECA02s07908g/GECA19s00890g	YALIOC03201g
NI9M	GECA06s05081g	YALIOE23749g
NB2M	GECA11s01638g	YALIOD10274g

The *Y. lipolytica* data are from Morgner *et al*, 2008 and Dröse *et al*, 2011.

Supplementary Table 3. **Syntenic blocks**

Nb of genes per syntenic block	Number of syntenic blocks		
	ARAD/YALI	ARAD/GECA	YALI/GECA
2	424	390	328
3	199	241	134
4	104	79	31
5	38	44	12
6	20	16	6
7	8	3	0
8	1	4	0
9	2	1	0
10	0	0	0
11	1	0	0
Total nb of syntenic blocks	797	778	511
Nb of genes in syntenic blocks	2264	2197	1278

Supplementary Table 4. **Distribution of spliceosomal introns per *G. candidum* protein coding gene.**

Genome content (excluding transposons)	Nb introns/gene	Nb genes
Number of protein coding genes		6804
Total number of spliceosomal introns		3830
Number of genes with	0	2414
	1	1526
	2	549
	3	227
	4	68
	5	27
	6	8
	7	5
	8	2
	9	1
	10	1

Supplementary Table 5. Chitin metabolism genes in *G. candidum*.

Gene name in <i>S. cerevisiae</i>	Function	Copy number in <i>S. cerevisiae</i>	Copy number in <i>G. candidum</i>	Copy number in <i>N. crassa</i>	Copy number in <i>A. oryzae</i>
<i>CTS1</i>	endochitinase	1	4	2	2
<i>CTS2</i>	chitinase	1	2	3	2
<i>CHS1</i>	chitin synthase I	1	2	1	1
<i>CHS2</i>	chitin synthase II	1	2	1	1
<i>CHS3</i>	chitin synthase III	1	6	4	3
<i>CHS7</i>	chitin biosynthesis through the regulation of chitin synthase III	1	2	1	1
np	filamentous fungal-type chitin synthase	np	2*	1	3
<i>CYK3</i>	chitin synthase activator	1	2	1	1
<i>SKT5</i>	chitin synthase III activator	1	5 (+ 1 pseudogene)	1	2
<i>CDA1</i>	chitin deacetylase	1	2	1	0
<i>CRH1</i>	chitin transglycosylase	1	2	2	2
<i>GFA1</i>	glutamine-fructose-6-phosphate amidotransferases	1	2	1	2
<i>UTR2</i>	chitin transglycosylase	1	2	1	1

np: not present

*GECA04s06797g and GECA14s02386g

Supplementary Table 6. Lipases and β -oxydation genes in *G. candidum*

<i>S. cerevisiae</i> gene name	<i>G. candidum</i> genes with orthologs in Saccharomycotina	<i>G. candidum</i> genes with orthologs in Pezizomycotina	Annotation
Lipase genes			
-		GECA03s04685g*	Lipase: extracellular triglyceride lipase
-		GECA03s04652g	//
-		GECA03s04718g	//
-		GECA03s04212g	//
B-oxidation genes			
<i>POX1</i>	GECA14s03310g		Fatty-acyl coenzyme A oxidase
<i>CTA1</i>	GECA27s00395g	GECA01s02936g	Catalase A
<i>POT1</i>	GECA17s00318g		3-ketoacyl-CoA thiolase with broad chain length specificity
<i>FOX2</i>	GECA16s01759g		3-hydroxyacyl-CoA dehydrogenase and enoyl-CoA hydratase
<i>FOX3</i>	GECA02s07259g		3-ketoacyl-CoA thiolase with broad chain length specificity
//	GECA11s01451g		//
<i>ECI1</i>	GECA07s05048g		Peroxisomal delta3,delta2-enoyl-CoA isomerase
<i>SPS19</i>	GECA07s03772g		Peroxisomal 2,4-dienoyl-CoA reductase

*Similar lipases were found in *Candida rugosa*, *Candida cylindracea* and an undescribed species *Candida* sp. ACITM

Supplementary Table 7. List of putative *G. candidum* genes encoding proteins involved in intron splicing.

	Saccharomycotina*	Pezizomycotina**	Pezizomycotina and <i>Y. lipolytica</i> ***
snRNP proteins			
Core Sm proteins			
<i>SMB1</i>	GECA10s03244g	GECA02s09393g	
<i>SMD1</i>	GECA01s11065g		
<i>SMD2</i>	GECA08s04608g		
<i>SMD3</i>	GECA19s01594g		
<i>SME1</i>	GECA03s06555g		
<i>SMX2</i>	GECA03s04850g		
<i>SMX3</i>	GECA08s03530g		
Lsm proteins			
<i>LSM1</i>	GECA13s00054g		
<i>LSM2</i>	GECA02s09470g		
<i>LSM3</i>	GECA08s03838g		
<i>LSM4</i>	GECA11s03486g		
<i>LSM5</i>	-	-	GECA03s02815g
<i>LSM6</i>	GECA01s06225g		
<i>LSM7</i>	GECA05s02749g		
<i>LSM8</i>	GECA05s00901g		
U1			
<i>SNP1</i>	GECA16s02694g		
<i>MUD1</i>	GECA17s02001g		
<i>YHC1</i>	GECA15s02848g		
<i>LUC7</i>	GECA08s03079g		
<i>NAM8</i>	GECA15s02716g		
<i>PRP39</i>	GECA06s05158g		
<i>PRP40</i>	GECA05s06247g		
<i>PRP42</i>	GECA26s00285g		
<i>SNU56</i>	No hit		
<i>SNU71</i>	GECA02s02903g		
U2			
<i>LEA1</i>	GECA03s04113g		
<i>MSL1</i>	GECA17s02001g		
//	GECA02s04234g		
<i>CUS2</i>	GECA05s00626g		
SF3a			
<i>PRP9</i>	GECA10s03420g		
<i>PRP11</i>	GECA06s00087g		
<i>PRP21</i>	GECA06s02518g		
SF3b			
<i>CUS1</i>	GECA01s06764g		
<i>RSE1</i>	GECA14s00109g		

HSH49 GECA03s03684g
HSH155 GECA07s00175g
RDS3 GECA02s01990g
YSF3 GECA18s02463g

RES

IST3 GECA15s01913g
BUD13 GECA05s07358g
PML1 GECA11s03662g

U4

PRP3 GECA23s00538g
PRP4 GECA08s03464g
SNU13 GECA14s01429g

U5

DIB1 GECA02s09492g
PRP8 GECA07s01220g
PRP28 GECA05s02265g
BRR2 GECA23s00197g
LIN1 GECA02s08832g
SNU114 GECA07s02419g
PRP6 GECA03s04201g
AAR2 GECA08s05169g

U6

PRP24 GECA01s06819g

U4/U6

PRP3 GECA23s00538g
PRP4 GECA08s03464g
SNU13 GECA14s01429g
PRP31 GECA23s00802g

U4/U6/U5

PRP3 GECA23s00538g
PRP4 GECA08s03464g
SNU13 GECA14s01429g
PRP31 GECA23s00802g
DIB1 GECA02s09492g
PRP8 GECA07s01220g
PRP28 GECA05s02265g
BRR2 GECA23s00197g
SNU114 GECA07s02419g
PRP6 GECA03s04201g
PRP38 No hit
SNU23 GECA12s01682g
SNU66 GECA01s06786g
SPP381 No hit
SAD1 GECA05s02463g

Non-snRNP proteins

Assembly

MSL5 GECA07s00318g
MUD2 GECA01s01143g

<i>SUB2</i>	GECA19s00604g		
<i>PRP5</i>	GECA06s04410g		
Activation			
<i>PRP19</i>	GECA01s10471g		
<i>SNT309</i>	No hit		
<i>CEF1</i>	GECA10s03651g		
<i>SYF1</i>	GECA02s02914g		
<i>CLF1</i>	GECA20s01055g		
<i>SYF2</i>	GECA05s03035g		
<i>ISY1</i>	GECA02s06742g		
<i>NTC20</i>	No hit		
<i>CWC2</i>	GECA05s02254g		
<i>PRP45</i>	GECA11s00032g		
<i>PRP46</i>	GECA16s01440g		
<i>ECM2</i>	GECA18s00736g		
<i>CWC21</i>	GECA03s00285g		
First reaction			
<i>CWC22</i>	GECA09s01143g		
<i>SPP2</i>	No hit		
<i>PRP2</i>	GECA17s02375g		
	GECA32s02001g		
<i>CWC25</i>	GECA05s02705g		
<i>YJU2</i>	GECA08s04630g	GECA15s01110g	
Second reaction			
<i>PRP17</i>	GECA07s00538g		
<i>PRP16</i>	No hit		
<i>SLU7</i>	GECA05s02892g		
<i>PRP18</i>	GECA05s02848g		
<i>PRP22</i>	GECA18s00395g		
Disassembly			
<i>SPP382/NTR1</i>	GECA17s00054g		
<i>NTR2</i>	GECA03s07039g		
<i>PRP43</i>	GECA20s01264g	GECA16s02188g	
Debranching enzyme			
<i>DBR1</i>	GECA04s02529g		
Unknown			
<i>BUD31</i>	GECA07s03739g		
<i>CWC15</i>	-	-	GECA04s02056g
<i>CWC24</i>	GECA01s05400g		
<i>CWC27</i>	GECA26s00703g		
<i>URN1</i>	GECA05s06346g		

* genes which can also be found in other Saccharomycotina

**genes with Pezizomycotina orthologs only

***genes with Pezizomycotina and *Y. lipolytica* orthologs

Supplementary Table 8. **Comparison of the distribution of the homologous genes involved in hyphal growth in *Neurospora crassa*, *G. candidum* and *Y. lipolytica***

<i>Neurospora crassa</i> functions and nomenclature	<i>Neurospora crassa</i>	<i>Geotrichum candidum</i>	<i>Yarrowia lipolytica</i>
Cell wall			
Glucan synthase regulator gs-1	NCU04189	GECA01s04894g	YALIOB20570g
Chitin synthase chs-1, chs-2, chs-3, chs-4, chs-5, chs-6, chs-7	NCU03611	GECA04s06797g	YALIOD03179g
	NCU05239	GECA14s02386g	YALIOB16324g
	NCU04251	GECA09s01715g	YALIOD25938g
	NCU09324	GECA05s00582g	YALIOC24354g
	NCU04352	GECA06s02375g	YALIOD17006g
	NCU05268	GECA01s05081g	YALIOE10417g
	NCU05350	GECA01s06676g	YALIOE22198g
		GECA01s06687g	YALIOF28655g
		GECA05s05191g	YALIOE16170g
		GECA06s01830g	
		GECA09s02287g	
		GECA20s01385g	
		GECA11s03596g	
	GECA13s00978g		
	GECA32s00472g		
Cytoskeleton genes			
Beta-tubulin bml	NCU04054	GECA32s02265g GECA04s07127g	YALIOE00726g YALIOE26961g
Fimbrin, an actin-binding protein fim	NCU003992	GECA02s02650g	YALIOB17622g
Tropomyosin, an actin-binding protein tpm-1	NCU001204	GECA18s02606g GECA01s02771g	YALIOF27049g
Subunit of the Arp2/3 complex arp-3	NCU001756	GECA12s01121g	YALIOE34170g
Nuclei			
NDR protein kinase dbf-2	NCU09071	GECA05s06973g	YALIOB14201g
Histone hh-1	NCU06863	GECA04s00890g	YALIOB16280g
Heterochromatin protein HP1 Hpo	NCU04018		
Nucleoporin son-1	NCU04288	GECA16s01198g	YALIOE23265g
Endoplasmic reticulum			
ER associated HSP grp-78	NCU03982	GECA07s00527g	YALIOE13706g
Dolichol-phosphate mannosyltransferase dpm	NCU07965	GECA14s02067g	YALIOD06281g
Vacuole			
Subunit A of vacuolar ATPase vma-1	NCU01207	GECA02s07754g	YALIOA09900g
Vacuole-associated SNARE protein vma-3	NCU06777	GECA06s02727g	YALIOE11825g
		GECA01s05279g	YALIOC22275g
Subunit C of vacuolar ATPase vma-5	NCU09897	GECA12s01572g	YALIOA11143g

Mitochondria

Mitochondrial acetylornithine-glutamate transacetylase. Arginine biosynthesis arg-4 NCU10468 GECA02s00604g YALIOE13057g

Golgi

Component of Golgi body-associated retrograde protein complex vps-52 NCU05273 GECA01s04157g YALIOF07381g

Calcium transporters

Ca²⁺/H⁺-ATPase nca-1, nca-2, nca-3 NCU03305 GECA06s05246g YALIOE09471g
NCU04736 GECA09s01066g YALIOD04873g
NCU05154 GECA03s01000g

Ca²⁺/H⁺ exchange protein cax NCU07075 GECA06s00263g YALIOB09273g
GECA12s00120g

Exocyst

Exocyst complex sec-3, sec-5, sec-6, sec-8, sec-15, exo-70 NCU09869 GECA17s01242g YALIOF21681g
NCU07698 GECA06s03123g YALIOE33759g
NCU03341 GECA25s00681g YALIOA19052g
NCU04190 GECA02s05917g YALIOF12969g
NCU00117 GECA16s00263g YALIOC01595g
NCU08012 GECA01s07512g YALIOD08492g
NCU06631 GECA06s00604g YALIOF11143g
GECA02s04586g YALIOC11946g

Polarity

MAP kinase activator bem-1 NCU06593 GECA04s04883g YALIOF27643g
GECA21s00131g

Involved in actin organization bni-1 NCU01431 GECA12s01407g YALIOD10879g
GECA04s04894g

Putative involved in septum formation cla-4 NCU00406 GECA11s02155g YALIOC22770g
GECA04s06214g YALIOF00572g

NDR kinase, essential for polar cell extension cot-1 NCU07296 GECA02s00043g YALIOB04268g
GECA10s01297g

RHO-1-specific GAP. Involved in coordinating apical tip growth lrg-1 NCU02689 GECA06s02760g YALIOF13211g
GECA03s04938g

MAP-kinase2, required for cell fusion mak-2 NCU02393 GECA01s10548g YALIOE23496g

NDR kinase pod-6 NCU02537 GECA13s03233g YALIOE34001g

Subunit of the polarisome complex spa-2 NCU03115 GECA09s03684g YALIOF16665g
GECA05s03904g

Septation

Involved in septum formation bud-3 NCU06579 GECA18s01121g YALIOC16346g
GECA23s00054g

Involved in septum formation and putative landmark protein bud-4 NCU00152 GECA03s02749g YALIOD11880g
GECA13s02529g

Putatively involved in septum formation <i>cdc-12</i>	NCU03795	GECA22s01297g GECA03s00736g	YALIOD27148g YALIOF26873g
Involved in septum formation and putative landmark protein <i>rgf-3</i>	NCU02131	GECA15s00659g	YALIOE12155g
Rho GTPase <i>rho-4</i>	NCU03407	GECA03s01451g	YALIOE23001g YALIOD01045g
Contributes to septal plugging <i>so</i>	NCU02794	GECA07s04663g	YALIOE10549g
Hyphal fusion			
Pheromone-regulated membrane protein 1-like, involved in cell fusion events <i>prm-1</i>	NCU09337	GECA09s00516g	YALIOE08580g

Supplementary Table 9. List of genomes used in this study

Species	Strain	Sequence source
<i>Eremothecium gossypi</i>	ATCC 10895 ^T	Ashbya Genome Database
<i>Aspergillus fumigatus</i>	Af293	NCBI
<i>Aspergillus nidulans</i>	FGSC A4	Broad
<i>Candida glabrata</i>	CBS 138 ^T	Genolevures
<i>Candida lusitanae</i>	ATCC 42720	Broad
<i>Coccidioides immitis</i>	RS	Broad
<i>Cryptococcus neoformans</i>	JEC21	NCBI
<i>Debaryomyces hansenii</i>	CBS 767 ^T	Genolevures
<i>Fusarium graminearum</i>	PH-1	Broad
<i>Geotrichum candium</i>	CLIB 918 (ATCC 204307)	ORCAE
<i>Kluyveromyces lactis</i>	CLIB 210	Genolevures
<i>Komagataella pastoris</i>	CBS 7435	ORCAE
<i>Lachancea kluyveri</i>	CBS 3082 ^T	Genolevures
<i>Lachancea thermotolerans</i>	CBS 6340 ^T	Genolevures
<i>Magnaporthe grisea</i>	70-15	Broad
<i>Millerozyma (sorbitophila) farinosa</i>	CBS 7064	Genolevures
<i>Neurospora crassa</i>	OR74A	Broad
<i>Ogataea parapolyomorpha</i>	CBS 4732 ^T	JGI
<i>Penicillium chrysogenum</i>	Wisconsin 54-1255	JGI
<i>Phanerochaete chrysosporium</i>	RP-78	JGI
<i>Saccharomyces (uvarum) bayanus</i>	623-6C	Broad
<i>Saccharomyces cerevisiae</i>	S288c	SGD
<i>Saccharomyces paradoxus</i>	NRRL Y-17217T	Broad
<i>Scheffersomyces stipitis</i>	CBS 6054 ^T	JGI
<i>Schizosaccharomyces pombe</i>	972h	NCBI
<i>Sclerotinia sclerotiorum</i>	1980 UF-70	Broad
<i>Trichoderma reesei</i>	QM6a	JGI
<i>Ustilago maydis</i>	521	Broad
<i>Yarrowia lipolytica</i>	CLIB 122	Genolevures
<i>Zygosaccharomyces rouxii</i>	CBS 732 ^T	Genolevures

The addresses of the databases are: Ashbya Genome Database, <http://agd.vital-it.ch>, Broad, <http://www.broadinstitute.org/annotation/fungi> ; Genolevures, <http://www.genolevures.org>; JGI, <http://genome.jgi-psf.org/programs/fungi>; NCBI, <http://www.ncbi.nlm.nih.gov/genome> ; ORCAE, <http://bioinformatics.psb.ugent.be/orcae>; SGD, <http://yeastgenome.org>.

Differential gene retention as an evolutionary mechanism to generate biodiversity in yeasts

Guillaume Morel, Lieven Sterck, Dominique Swennen, Marina Marcet-Houben, Djamila Onesime, Anthony Levasseur, Noémie Jacques, Sandrine Mallet, Arnaud Couloux., Karine Labadie, Joëlle Amselem, Jean-Marie Beckerich, Bernard Henrissat, Yves Van de Peer, Patrick Wincker, Jean-Luc Souciet, Toni Gabaldón, Colin R. Tinsley & Serge Casaregola

Supplementary note

Transposable elements

Long terminal repeat (LTR) retrotransposons and LINEs were found as remnants, in the assembly. At least five families of DNA transposons closely related to Tc elements; two families of Mutator-like elements that comprise four entire copies, one partial copy and three pseudogenes were detected. The Mutator-like elements were found to be closely related to Pezizomycotina elements, but not to the *Yarrowia lipolytica* Mutator-like Mutyl element ¹. In addition, analysis of reads using the REPET program ² was performed; it suggested the presence of the first MITE elements (Miniature Inverted-repeat Transposable Element) to be found in a Saccharomycotina species. This 452 bp element is represented in the assembly by five full-length copies and several degenerate copies. Like other known MITEs, the *G. candidum* element is able to form a stable hairpin-like secondary structure (not shown), but it lacks the conserved terminal inverted repeats.

A solo LTR of 382 bp was also found in 66 copies in 12 of the scaffolds. The detectable copies of this repeated element were all at the ends of the scaffolds, except for internal clusters found in scaffold 3 and in scaffold 6 (11 and 8 copies, respectively). In these two scaffolds, the solo LTRs are located in regions covering over 20 kb and containing large gaps, suggesting that these regions also contain entire transposons. The presence of these clusters in the ends of 10 scaffolds very likely prevented the assembly into larger scaffolds. The presence of a number of transposable element hot-spots is reminiscent of the situation described in two yeasts, *Scheffersomyces stipitis*³ and *D. hansenii* (Genolevures consortium, unpublished; ⁴), where the Tps5 and Tdh5 LTR-retrotransposons, respectively, of each species were all found clustered in one locus per chromosome. These regions were proposed to contain the centromeres ⁴, as already seen in *N. crassa* ^{5,6}. A similar observation was recently made for the yeast *Hansenula (Ogataea parapolymorpha) polymorpha* ⁷. Although pulsed-field gel electrophoresis failed to separate the chromosomes of *G. candidum* to estimate the chromosome number, a cytological analysis by ⁸ found eight chromosomes in this species. The number of LTR clusters that we identified (two internally and 10 at the ends of other scaffolds) fits the presence of eight chromosomes in this strain.

Various features of the *G. candidum* genome

Sequence homology using fungal sequences in the Rfam database (<http://rfam.sanger.ac.uk/>) detected TPP riboswitch candidates in the 5' sequence of the *THI4* and the *DUR3* homologs.

An intein of 1,237 bp was detected in GECA09s00186g encoding a subunit of the translation factor eIF2. It is a full-length intein presenting around 40% sequence similarity with VMA1-derived endonucleases from Saccharomycotina over 485 aa. In addition to the intein in the vacuolar ATPase gene VMA, inteins have previously been found in the glutamate synthase gene (GLT1) and in the threonyl-tRNA synthetase gene (ThrRS) in various Saccharomycotina species ⁹. Various inteins have been found in filamentous fungi ⁹. This is the first to be found in a fungal gene encoding a translation factor.

The mitochondrial DNA sequence is 28,008 bp long and has a GC content of 28%, placing it midway between those of *S. cerevisiae* (20%) and *D. hansenii* (38%) (Supplementary Figure 1). It carries 14 protein-coding genes: *COB*, *COXI*, *COXII*, *COXIII*, *ATP6*, *ATP8*, *ATP9*, *VAR1*, plus six ubiquinone oxidoreductase complex I protein genes *NADH1*, *NADH2*, *NDH3*, *NADH4*, *NADH5* and *NADH6*. A total of 23 tRNA genes and SSU and LSU rRNA genes were found (Supplementary Fig. 1). The *ATP9* gene, the downstream tRNA-Phe gene and a tRNA-Arg gene are oriented counterclockwise. Interestingly, there is only one intron, (carrying an endonuclease, in the *COB* gene), in the mtDNA, rendering *G. candidum* the Saccharomycotina species with the lowest number of introns. *Candida phangngensis* carries two introns splitting the *COB* gene¹⁰.

Spliceosomal introns and spliceosome

A total of 14 introns departed from the consensus GT in 5' with 12 introns starting with GC, one with GA and one with GG. In addition only one intron did not end with AG. These results were verified using RNA sequencing. A maximum of 10 introns was found in GECA13s02177g, a gene coding for an *MRE11* homolog which has functions in DNA double-strand break repair and telomere stability. An intron in GECA18s01957g encoding a mitochondrial isocitrate dehydrogenase reaches the unusual size for *G. candidum* of 1428 nt, whereas the rest of the introns have a size between 44 nt and 750 nt, the median being 71 nt. Intron size distribution is overwhelmingly centered between 60 nt and 80 nt (Supplementary Figure 3). Finally, the most striking feature of the spliceosomal introns in *G. candidum* is the poor conservation of the 5' splice site and the branch point. In previously-sequenced yeast genomes, this sequence is relatively well conserved. The consensus is GTATGT in *S. cerevisiae* and the large majority of the Saccharomycotina yeasts, whereas *Y. lipolytica* stands out with a GTGAGT consensus. However, *G. candidum* presents a mix of these two patterns. Outside the first two bases¹¹, the other bases are less conserved than in other yeasts (Supplementary Figure 4). Indeed, this pattern is closer to that observed in the Pezizomycotina/Basidiomycota than to other Saccharomycotina. Examples of 5' splice site patterns for a Pezizomycotina species, *Fusarium graminearum*, and a Basidiomycota species, *Cryptococcus neoformans*, are shown in Supplementary Figure 4.

Similarly, the branch point sequence is also less conserved than is the case in the other Saccharomycotina yeasts. The most common sequence in *G. candidum* was found to be NNCTAAC (72% of the total), followed by NNCTAAT (12%), NNTTAAC (7%) and NNCTGAC (5%). By comparison, the consensus sequence NNCTAAC is found in 95% of *Y. lipolytica*, 92% of *D. hansenii* and 93% of *S. cerevisiae* introns. The median S2 distance separating the branch point sequence from the 3' splice site (AG) was 12 nucleotides.

The spliceosome is a large ribonucleoprotein complex that catalyzes intron removal from precursor mRNAs via two trans-esterification steps. The spliceosome comprises five small nuclear RNAs (snRNAs) organized in small nuclear ribonucleoprotein particles (snRNPs) and numerous proteins¹². *G. candidum* homologous snRNP and non-snRNP protein coding genes are listed in Supplementary Table 7. Among these genes, three (GECA02s09393g, GECA15s01110g and GECA16s02188g) do not have counterparts in other Saccharomycotina and have orthologs only in filamentous fungi. Two other *G. candidum* genes have orthologs only in *Y. lipolytica*

and in filamentous fungi (GECA03s02815g and GECA04s02056g). Interestingly, *G. candidum* has two paralogs coding for Yju2p and for Prp43p with, in each family, one paralog of the filamentous fungal type (GECA15s01110g for Yju2p and GECA16s02188g for Prp43p). In *S. cerevisiae*, Yju2p functions in the first step of trans-esterification and has a possible role in positioning the branch point during the reaction¹³. Prp43p codes for a DExD/H-box RNA helicase involved in spliceosome disassembly together with the two cofactors Ntr1p and Ntr2p¹⁴. *G. candidum* has only one of these two cofactors; it displays low sequence similarity with the Saccharomycotina counterparts but it is well conserved compared to the filamentous fungal orthologs.

Mating type

The *GcMATA* coding sequence presented similarity with the previously-described HMG-box proteins in yeasts and is located between the *APC5* and *SLA2* orthologs in *G. candidum* (Supplementary Figure 6). To confirm that this gene corresponded to a mating type gene, we ran a search for the other idiomorphs by PCR amplification of different parts of this region in a set of *G. candidum* stains of various origins. First, primers were chosen to amplify a fragment of 279 bp within the *MATA* gene and were subsequently used in PCR on 62 *G. candidum* strains preserved at the CIRM-Levures (<http://www6.inra.fr/cirm/Levures>). PCR products of the expected size were obtained for 37 strains. No PCR product was observed for 25 strains, suggesting that these strains may not carry the *MATA* gene but the opposite mating type. Hence, primers located in the regions flanking the *MATA* gene were designed and used to amplify the corresponding region in three strains that failed to give a positive signal with the *MATA*-specific primers. A 1.5 kb region was successfully amplified from the three strains including the *G. candidum* type strain CBS 615.84^{NT}, and was entirely sequenced: it contained a CDS of 339 aa, different from the *GcMATA* gene, in an otherwise identical environment. The product of this gene presented sequence similarity with various MAT α proteins in the region of the alpha box (not shown). This gene was therefore called *MATB*. We thus confirmed that *G. candidum* is heterothallic.

Comparison of the structure of the mating type loci in several yeast and fungal species (Supplementary Figure 6) indicated that the *G. candidum* locus contains only two genes—*GcMATA* and *GcMATB*—and thus resembles some filamentous fungi such as *Aspergillus* species¹⁵, but not Saccharomycotina yeasts. A roughly 2 kb region separates the *MATA* gene from the gene upstream of *MATA*, but we did not identify a valid CDS in this region. The *SLA2* and the *SUI1* genes are the immediate neighbors to the right of the sexual locus in the examples of Saccharomycotina species shown in Supplementary Figure 6. However, the neighboring genes on the left-end side of the mating type loci showed more variability: an ortholog of *DIC1* is found in *Lachancea kluyveri* and *Ogatea angusta* and an ortholog of *APN2* in *Y. lipolytica*, *Neurospora crassa* and *Trichoderma reesei*. Interestingly, the left flanking region of the *G. candidum* sexual locus is not conserved and has very likely been rearranged. Indeed, the neighboring gene of the *G. candidum* sexual locus is *APC5*, which is found near the sexual locus of some Pezizomycotina species but separated from this locus by two genes, *COX13* and *APN2*. A search for *G. candidum* *COX13* and *APN2* indicated that they are localized on different scaffolds. Genomic rearrangements located at the border of the sexual locus have previously been seen in yeast species that are

able to switch mating type; this switching was proposed as responsible for an erosion of this locus ¹⁶. Our description of the organization of the *G. candidum* sexual locus indicates that rearrangement at the border of the locus may be more widespread, potentially affecting Dipodascaceae as well as other Saccharomycetes family yeasts in which mating type switching takes place.

Cazymes in the *G. candidum* genome

Among the 133 glycoside hydrolase (GH) families integrated in the CAZy database (<http://www.cazy.org>; ¹⁷), fungal endoglucanases are so far classified into nine GH families. Endo- β -1,4-glucanases (EC 3.2.1.4) act in synergy with other enzymatic activities (including hydrolytic and oxidative enzymes) and help degrade the main plant cell wall component, i.e. cellulose ¹⁸. The endoglucanases of one of the GH families, GH45, randomly cleave glycosidic bonds on cellulose polymers, releasing cello-oligosaccharides as end-products ¹⁹. One GH45 family, comprising four members (not found in Saccharomycotina except one gene belonging to one of these families in *K. pastoris*) was detected in *G. candidum* (Supplementary data 3).

Surprisingly, four members of family AA9 of lytic polysaccharide monooxygenases (LPMOs) were also identified. LPMOs participate in cellulose targeting but via oxidative mechanisms, contrary to the hydrolytic mechanism of the GH counterparts. The presence of four LPMOs is unexpected in *G. candidum* as AA9 members are only found in the wood-decaying fungi, prevalently in white-rot basidiomycete fungi. Comparison with other yeasts (Supplementary Data "CAZy annotation") showed that AA9 members were exclusively identified in *G. candidum*. Moreover, *G. candidum* is the only Saccharomycotina yeast to possess genes encoding enzymes containing domains of the carbohydrate-binding module family 1 (CBM1), which specifically binds crystalline cellulose ²⁰. CBM1 are primarily found in the fungal kingdom ²¹ and are usually restricted to the genome of wood-rot fungi. Previous works showed that carbohydrate binding modules enable increase of the enzyme concentration in the vicinity of the substrate ²². Moreover, CBMs may also be involved in the destructuration of polysaccharides on the substrate fibrils ²³. Remarkably, *G. candidum* carries eight CBM1 members, and each CBM1 is linked to all the GH45 endoglucanases. Finally, *G. candidum* has two AA1_2-family ferroxidases like other yeasts but only *G. candidum* has one AA1 multicopper oxidase close to laccases. To our knowledge, *G. candidum* is the only yeast to retain this broad lignocellulolytic repertoire with representatives of typical filamentous fungi-associated families (AA1, AA9, CBM1).

HGT from Basidiomycota

Polyamines are involved in numerous processes and are essential for growth ²⁴. In *S. cerevisiae*, polyamine synthesis is initiated by two reactions: decarboxylation of L-ornithine by the *SPE1* gene results in putrescine, and decarboxylation of S-adenosyl-L-methionine by the *SPE2* gene results in S-adenosyl-methionamine. Transfer of an aminopropyl group from S-adenosyl-methionamine to putrescine by spermidine synthase (encoded by the *SPE3* gene) results in spermidine. A second aminopropyl group is then incorporated into

spermidine by spermine synthase (*SPE4*) to yield spermine. While *SPE3* is essential for *S. cerevisiae* growth, *SPE4* is not²⁵. Most filamentous fungi do not contain spermine, and these organisms contain a spermidine synthase encoded by an ortholog of *SPE3* as well as a second gene encoding a spermidine synthase, phylogenetically unrelated to either *SPE3* or *SPE4*.

G. candidum possesses a gene, GECA15s02364g, which shows a high degree of conservation with *SPE3* and *SPE4*. Interestingly, *G. candidum* also carries another gene, GECA13s02485g, that is similar to the second spermidine synthase of filamentous fungi, and indeed groups with the Basidiomycota sequences in phylogenetic analysis (Figure 4). This indicates that *G. candidum* has a very unusual complement of spermidine synthases, an *SPE3*-like spermidine synthase and a second spermidine synthase, derived from that of the filamentous fungi. This could imply that the *SPE4* gene, very likely derived from a duplication of the *SPE3* gene²⁵, was lost in *G. candidum* and that a spermidine synthase has been acquired through HGT from a basidiomycete. Therefore, the polyamine synthase gene complement of *G. candidum* is consistent with an involvement of polyamines in hyphal and pseudo-hyphal growth by a mechanism similar to that acting in filamentous fungi²⁶.

References of supplementary material

1. Neueglise, C., Chalvet, F., Wincker, P., Gaillardin, C. & Casaregola, S. Mutator-like element in the yeast *Yarrowia lipolytica* displays multiple alternative splicings. *Eukaryot Cell* **4**, 615-24 (2005).
2. Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in de novo annotation approaches. *PLoS One* **6**, e16526 (2011).
3. Jeffries, T.W. et al. Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. *Nat Biotechnol* **25**, 319-26 (2007).
4. Lynch, D.B., Logue, M.E., Butler, G. & Wolfe, K.H. Chromosomal G + C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres. *Genome Biol Evol* **2**, 572-83 (2010).
5. Borkovich, K.A. et al. Lessons from the genome sequence of *Neurospora crassa*: tracing the path from genomic blueprint to multicellular organism. *Microbiol Mol Biol Rev* **68**, 1-108 (2004).
6. Cambareri, E.B., Aisner, R. & Carbon, J. Structure of the chromosome VII centromere region in *Neurospora crassa*: degenerate transposons and simple repeats. *Mol Cell Biol* **18**, 5465-77 (1998).
7. Ravin, N.V. et al. Genome sequence and analysis of methylotrophic yeast *Hansenula polymorpha* DL1. *BMC Genomics* **14**, 837 (2013).
8. Gente, S. et al. Intra-species chromosome-length polymorphism in *Geotrichum candidum* revealed by pulsed field gel electrophoresis. *Int J Food Microbiol* **76**, 127-34 (2002).
9. Poulter, R.T., Goodwin, T.J. & Butler, M.I. The nuclear-encoded inteins of fungi. *Fungal Genet Biol* **44**, 153-79 (2007).
10. Gaillardin, C., Neueglise, C., Kerscher, S. & Nicaud, J.M. Mitochondrial genomes of yeasts of the *Yarrowia* clade. *FEMS Yeast Res* **12**, 317-31 (2012).
11. Goodwin, S.B. et al. Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. *PLoS Genet* **7**, e1002070 (2011).
12. Chen, H.C. & Cheng, S.C. Functional roles of protein splicing factors. *Biosci Rep* **32**, 345-59 (2012).
13. Chiang, T.W. & Cheng, S.C. A weak spliceosome-binding domain of Yju2 functions in the first step and bypasses Prp16 in the second step of splicing. *Mol Cell Biol* **33**, 1746-55 (2013).
14. Tsai, R.T. et al. Spliceosome disassembly catalyzed by Prp43 and its associated components Ntr1 and Ntr2. *Genes Dev* **19**, 2991-3003 (2005).
15. Dyer, P.S. & O'Gorman, C.M. Sexual development and cryptic sexuality in fungi: insights from *Aspergillus* species. *FEMS Microbiol Rev* **36**, 165-92 (2012).
16. Gordon, J.L. et al. Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents. *Proc Natl Acad Sci U S A* **108**, 20024-9 (2011).
17. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* **42**, D490-5 (2014).
18. Davies, G.J. et al. Structure and function of endoglucanase V. *Nature* **365**, 362-4 (1993).
19. Karlsson, J., Siika-aho, M., Tenkanen, M. & Tjerneld, F. Enzymatic properties of the low molecular mass endoglucanases Cel12A (EG III) and Cel45A (EG V) of *Trichoderma reesei*. *J Biotechnol* **99**, 63-78 (2002).
20. Boraston, A.B., Bolam, D.N., Gilbert, H.J. & Davies, G.J. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J* **382**, 769-81 (2004).
21. Palomares-Rius, J.E. et al. Distribution and evolution of glycoside hydrolase family 45 cellulases in nematodes and fungi. *BMC Evol Biol* **14**, 69 (2014).
22. Bolam, D.N. et al. *Pseudomonas* cellulose-binding domains mediate their effects by increasing enzyme substrate proximity. *Biochem J* **331** (Pt 3), 775-81 (1998).

23. Wang, L., Zhang, Y. & Gao, P. A novel function for the cellulose binding module of cellobiohydrolase I. *Sci China C Life Sci* **51**, 620-9 (2008).
24. Wallace, I.M., O'Sullivan, O., Higgins, D.G. & Notredame, C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res* **34**, 1692-9 (2006).
25. Hamasaki-Katagiri, N., Katagiri, Y., Tabor, C.W. & Tabor, H. Spermine is not essential for growth of *Saccharomyces cerevisiae*: identification of the SPE4 gene (spermine synthase) and characterization of a *spe4* deletion mutant. *Gene* **210**, 195-201 (1998).
26. Cheng, Y. et al. Polyamines stimulate hyphal branching and infection in the early stage of *Glomus etunicatum* colonization. *World J Microbiol Biotechnol* **28**, 1615-1621 (2012).