Lawrence Berkeley National Laboratory
Lawrence Berkeley National Laboratory

Title

The genome of wine yeast Dekkera bruxellensis provides a tool to explore its food-related properties

Permalink

https://escholarship.org/uc/item/0vt635rv

Author

Piskur, Jure

Publication Date

2012-07-02

DOI

10.1016/j.ijfoodmicro.2012.05.008

eScholarship.org Powered by the California Digital Library
University of California

The genome of wine yeast *Dekkera bruxellensis* provides a tool to explore its food-related properties

Jure Piškur[1,2], Zhihao Ling[2] , Marina Marcet-Houben[3], Olena P. Ishchuk[2] , Andrea Aerts[4], Kurt LaButti[4], Alex Copeland[4], Erika Lindquist[4], Kerrie Barry[4], Concetta Compagno[5], Linda Bisson[6] , Igor V. Grigoriev[4], Toni Gabaldón[3] and Trevor Phister[7]

[1] Wine Research Centre, University of Nova Gorica, Slovenia

[2] Department of Biology, Lund University, Sweden

[3] Centre for Genomic Regulation (CRG) and UPF, Barcelona, Spain

[4]US Department of Energy Joint Genome Institute, 2800 Michell Dr, Walnut Creek, CA 94598; US

[5] Department of Biological Sciences and Biotechnology, University of Milan, Italy

[6] Department of Viticulture and Enology, University of California, Davis, US

[7] Division of Food Science, Brewing Science Program, University of Nottingham, UK

JULY 2012

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

1 International Journal of Food Microbiology

2

3 **The genome of wine yeast *Dekkera bruxellensis* provides a tool to explore its food-related**

4 **properties**

5

6 Jure Piškur[1,2], Zhihao Ling[2] , Marina Marcet-Houben[3], Olena P. Ishchuk[2] , Andrea Aerts[4],

7 Kurt LaButti[4], Alex Copeland[4], Erika Lindquist[4], Kerrie Barry[4], Concetta Compagno[5], Linda

8 Bisson[6] , Igor V. Grigoriev[4], Toni Gabaldón[3] and Trevor Phister[7]

9

10 [1] Wine Research Centre, University of Nova Gorica, Slovenia

11 [2] Department of Biology, Lund University, Sweden

12 [3] Centre for Genomic Regulation (CRG) and UPF, Barcelona, Spain

13 [4]US Department of Energy Joint Genome Institute, 2800 Michell Dr, Walnut Creek, CA

14 94598; US

15 [5] Department of Biological Sciences and Biotechnology, University of Milan, Italy

16 [6] Department of Viticulture and Enology, University of California, Davis, US

17 [7] Division of Food Science, Brewing Science Program, University of Nottingham, UK

18

19

20

21

22

23 Corresponding author: J. Piskur, Department of Biology, Biologihuset A, Lund University,

24 Soelvegatan 35, SE-22362 Lund, Sweden, e-mail: Jure.Piskur@cob.lu.se, phone: +46 46

25 2111982

Abstract

The yeast *Dekkera/Brettanomyces bruxellensis* can cause enormous economic losses in wine industry due to production of phenolic off-flavor compounds. *D. bruxellensis* is a distant relative of baker's yeast *Saccharomyces cerevisiae*. Nevertheless, these two yeasts are often found in the same habitats and share several food-related traits, such as production of high ethanol levels and ability to grow without oxygen. In some food products, like lambic beer, *D. bruxellensis* can importantly contribute to flavour development. We determined the 13.4 Mb genome sequence of the *D. bruxellensis* strain Y879 (CBS2499) and deduced the genetic background of several "food-relevant" properties and evolutionary history of this yeast. Surprisingly, we find that this yeast is phylogenetically distant to other food-related yeasts and most related to *Pichia (Komagataella) pastoris*, which is an aerobic poor ethanol producer. We further show that the *D. bruxellensis* genome does not contain an excess of lineage specific duplicated genes nor a horizontally transferred *URA1* gene, two crucial events that promoted the evolution of the food relevant traits in the *S. cerevisiae* lineage. However, *D. bruxellensis* has several independently duplicated *ADH* and *ADH*-like genes, which are likely responsible for metabolism of alcohols, including ethanol, and also a range of aromatic compounds.

Keywords: Comparative genomics; wine yeast; evolution; ethanol fermentations; aromatic compounds

1. Introduction

51 There is an enormous diversity among yeast species, including those that play important roles
52 in traditional food processes, often in mixed cultures in spontaneous fermentations. One such
53 yeast is *Dekkera/Brettanomyces bruxellensis*, associated with lambic beer fermentation and
54 wine production, especially as a contributor, in a positive or negative manner, to flavour
55 development (Du Toit and Pretorius, 2000). This yeast can produce phenolic compounds, such
56 as 4-ethylguaiacol and 4-ethylphenol, which could lead to wine spoilage if present in high
57 enough concentration (Heresztyn, 1986; Vigentini et al., 2008). In fact, *D. bruxellensis*
58 represents a serious problem in wine industry, causing enormous economic losses as a
59 consequence of wine spoilage (Wedral et al., 2010). However, in spite of the economic impact
60 of *D. bruxellensis*, this yeast remains poorly studied.
61
62 *D. bruxellensis* is apparently not a close relative of baker's yeast *Saccharomyces cerevisiae*,
63 but the phylogenetic position of the *D. bruxellensis* group has so far been rather impossible to
64 determine (Woolfit et al., 2007). Both yeasts share several “peculiar” and rather “unusual”
65 traits important for food-related properties, such as production of high ethanol levels, high
66 tolerance towards ethanol, and the ability to grow without oxygen and in acidic environments
67 (Rozpędowska et al., 2011). Apparently, given the lack of relatedness, these traits evolved in
68 parallel in both groups, but it is unclear if the molecular mechanisms behind these properties
69 are similar or different (Rozpędowska et al., 2011). In other words, these two yeasts represent
70 an ideal model to study molecular processes involved in convergent and parallel evolutionary
71 routes.
72
73 Ethanol production and capability to survive without oxygen are highly relevant in food
74 fermentations. In *S. cerevisiae*, but not in *D. bruxellensis*, the corresponding genetic factors
75 that underlie these traits have been relatively well studied. For example, the whole genome

76 duplication (WGD), duplicated gene profiles, the horizontal transfer of the *URA1* gene
77 (coding for the DHODase, dihydroorotate dehydrogenase, catalysing the fourth pyrimidine *de*
78 *novo* pathway step), and lineage-specific duplication of the *ADH* genes (encoding alcohol
79 dehydrogenases), have been shown to be at least partially responsible for development of the
80 *S. cerevisiae* high fermentation capacity and/or anaerobic properties (reviewed in Piskur and
81 Langkjaer, 2004; Piskur et al., 2006). It is not known whether similar molecular strategies are
82 responsible for the domination of the same environment by *D. bruxellensis*.
83
84 Recently, a partial genome sequence of one strain of *D. bruxellensis* has been reported, and
85 the analysis estimated that this yeast has around 7.500 genes, of which many lack a homolog
86 in the *S. cerevisiae* genome (Woolfit et al., 2007). Further analysis of the partial sequence has
87 revealed that *D. bruxellensis* is not a simple haploid. Its genome contains approximately 1%
88 polymorphic sites but the exact physical background for this heterozygocity is not known
89 (Hellborg and Piskur, 2009).
90
91 Here we determined the whole genome sequence of the *D. bruxellensis* strain Y879
92 (CBS2499) and used it to deduce several "food-relevant" properties and evolution pathways
93 of this yeast.
94
95
96 **2. Materials and Methods**
97
98 **2.1. Genome sequencing and assembly**
99 The genome of *D. bruxellensis* strain Y879 (CBS2499) was sequenced using a combination of
100 454 and Illumina sequencing platforms (GYBS 454 standard rapid, GYHO 454 standard

101 rapid, GYHG 454 titanium 4kb, GYFW 454 titanium 4kb, GXXW Illumina 2x76 300bp,
102 ICHI Illumna 2x150 270bp, and ICCY Illumina 2x100 4kb CLIP). All general aspects of
103 library construction and sequencing can be found at the JGI website
104 (http://www.jgi.doe.gov/). An initial assembly of GXXW was conducted for QC purposes
105 using the Velvet assembler, version 0.7.55, with the following parameters: k 61 -
106 min_contig_lgth 100 -exp_cov 81. A list of data to be excluded from the draft assembly was
107 also created by identifying possible contaminant data in preliminary Newbler assemblies of
108 the 454 data. The resulting screened data was assembled along with shredded consensus
109 from the initial Velvet assembly using the Newbler assembler, software release 2.5-internal-
110 10Apr08-1, with the following parameters: -fe reads2remove.MPA -info -consed -finish -nrm
111 –rip -sio a 50 -l 350 -g -ml 30 -mi 94 -e 87. The final draft assembly was assembled from the
112 Illumina data, as well as 3kb and 15kb paired end data generated from the Newbler assembly
113 using wgsim, with the AllpathsLG assembler software release R38445, to an estimated
114 assembled coverage of 128x (Table 1A) with 84 scaffolds with an N50 of 1.7 Mb, and 880
115 contigs with an N50 of 30.5 Kb (Table 1B).

116

117 **2.2. EST sequencing and assembly**

118 Total RNA from two separate *D. bruxellensis* samples, "air" and "no air" were used to
119 generate stranded RNASeq libraries. mRNA was purified from total RNA using the
120 Absolutely mRNA™ purification kit (Stratagene,Santa Clara, CA). Subsequently, the mRNA
121 samples were chemically fragmented to the size range 200-250 bp using 1x fragmentation
122 solution for 5 minutes at 70 ℃ (RNA Fragmentation Reagents, AM8740 – Zn, Ambion,
123 Carlsbad, CA). First strand cDNA was synthesized using Superscript II Reverse Transcriptase
124 (Invitrogen, Carlsbad, CA) and random hexamers then the second strand was synthesized

125 using *E. coli* RnaseH, DNA Ligase, and DNA polymerase I for nick translation. The dscDNA
126 was then cleaned up using a double SPRI bead selection (Agencourt Ampure beads; Beckman
127 Coulter, Brea CA). The TruSeq Sample Prep kit (Illumina Inc. San Diego, CA) was used for
128 RNASeq library creation using the dscDNA and the manufacturer's instructions (Illumina).
129 Briefly, dscDNA was end repaired, and ligated to Illumina adaptors. Then the second strand
130 was removed by AmpErase UNG (Applied Biosystems, Carlsbad, CA) similar to the method
131 described by (Parkhomchuk et al., 2009). Paired end 100 bp reads were generated by
132 sequencing using the Illumina HiSeq instrument. 176,820,692 and 159,263,276 reads were
133 generated for the "air" and "no air" samples respectively. Newbler assembled consensus EST
134 sequence data was used to assess the completeness of the final genome assembly Fasta with
135 alignment using 90% identity and 85% coverage thresholds. This resulted in 89.16%
136 placement.

137 **2.3. Genome Annotation**

138 The *D. bruxellensis* CBS 2499 genome was annotated using the JGI annotation pipeline,
139 which takes multiple inputs (scaffolds, ESTs, and known genes) and runs several analytical
140 tools for gene prediction and annotation, and deposits the results in the JGI fungal genome
141 portal MycoCosm (http://www.jgi.doe.gov/fungi) for further analysis and manual curation.

142 Genomic assembly scaffolds were masked using RepeatMasker (Smit et al., 2010) and the
143 RepBase library of 234 fungal repeats (Jurka et al., 2005) and RepeatScout. Using the repeat-
144 masked assembly, several gene prediction programs falling into three general categories were
145 used: 1) *ab initio* - FGENESH (Salamov and Solovyev, 2000); GeneMark (Isono et al.,
146 1994), 2) *homology-based* - FGENESH+; Genewise (Briney and Durbin, 2000) seeded by
147 BLASTx alignments against GenBank's database of non-redundant proteins (NR:
148 http://www.ncbi.nlm.nih.gov/BLAST/), and 3) *EST-based* - EST_map
149 (http://www.softberry.com/) seeded by EST contigs. Genewise models were extended where

150 possible using scaffold data to find start and stop codons. EST BLAT alignments (Kent, 2002)
151 were used to extend, verify, and complete the predicted gene models. The resulting set of
152 models was then filtered for the best models, based on EST and homology support, to produce
153 a non-redundant representative set (see Table 1C). This representative set was subject to
154 further analysis and manual curation. Measures of model quality include proportions of the
155 models complete with start and stop codons (92%) consistent with ESTs (91%) supported by
156 similarity with proteins from the NCBI NR database (87%) Quality metrics for gene models
157 are summarized in Table 1D.

158 All predicted gene models functionally annotated using SignalP (Nielsen et al., 1997),
159 TMHMM (Melen et al., 2003), InterProScan (Zdobnov and Apweiler, 2001), BLASTp
160 (Altschul et al., 1990) against nr, and hardware-accelerated double-affine Smith-Waterman
161 alignments (deCypherSW; http://www.timelogic.com/decypher_sw.html) against SwissProt
162 (http://www.expasy.org/sprot/), KEGG (Kanehisa et al., 2008), and KOG (Koonin et al.,
163 2004). KEGG hits were used to assign EC numbers (http://www.expasy.org/enzyme/), and
164 Interpro and SwissProt hits were used to map GO terms (http://www.geneontology.org/).
165 Multigene families were predicted with the Markov clustering algorithm (MCL (Enright et al.,
166 2002)) to cluster the proteins, using BLASTp alignment scores between proteins as a
167 similarity metric. Functional annotations are summarized in Table 1E. Manual curation of the
168 automated annotations was performed by using the web-based interactive editing tools of the
169 JGI Genome Portal to assess predicted gene structures, assign gene functions, and report
170 supporting evidence.

171

172 **2.4. Phylome reconstruction**

173 The *D. bruxellensis* CBS2499 predicted proteome described above, and those from a
174 collection of 21 completely sequenced fungal genomes were downloaded from various

175 sources (see Table 2, A&B). Using the phylomeDB pipeline (Huerta-Cepas et al., 2011) we
176 reconstructed the complete collection of evolutionary histories of *D. bruxellensis* genes, i.e
177 the phylome. In brief, the phylogenetic reconstruction pipeline involves Smith-Waterman
178 searches for homologs (E-Value <1e-05, 50% sequence overlap) accross 21 related fungal
179 species including: *Schizosaccharomyces pombe* and *Yarrowia lipolytica* as outgroups. These
180 homologous groups are then aligned using 3 different programs, MUSCLE v3.7 (Edgar,
181 2004), MAFFT v6.712b (Katoh, 2008), and DIALIGN-TX (Subramanian, 2008), and in
182 forward and reverse direction (i.e using the Head or Tail approach). The 6 resulting
183 alignments were then combined with M-COFFEE (Wallace et al., 2006) and then trimmed
184 with trimAl v1.3 (Capella-Gutiérrez et al., 2009) using consistency-score cutoff 0.1667 and
185 gap-score cutoff 0.9. Multiple sequence alignments were then used to reconstruct maximum
186 likelihood tree. For each gene, the best evolutionary model was chosen among seven
187 competing models (JTT, LG, WAG, Blosum62, MtREV, VT and Dayhoff) reconstructing a NJ
188 tree, using bioNJ (Gascuel, 1997) as implemented in PhyML (Guindon, 2009); The 2 best-
189 fitting models, as determined by the AIC criterion (Akaike, 1973), were used to derive ML
190 trees. The model used four rate categories and the fraction of invariant positions was inferred
191 from the data. Branch support was computed using an aLRT (approximate likelihood ratio
192 test) based on a chi-square distribution. Resulting trees and alignments are stored in
193 phylomeDB, with the phylomeID 138 (www.phylomedb.org). Orthology and paralogy
194 relationships for each gene in the phylome were obtained using the species-overlap algorithm
195 implemented in ETE (Huerta-Cepas et al., 2010).
196
197 347 genes, which had a strict and phylogeny-based one-to-one orthology relationship in all
198 species included in the phylome, were concatenated into a single alignment and then trimmed
199 using trimAl (gap-score cutoff 0.5, conservation score 0.5). The species-tree was

200 reconstructed using RaxML vesion:7.2.6 (Stamatakis, 2005), using a 4-categories *GAMMA*
201 distribution to account for rate heterogeneity and the LG model. Bootstrap support was
202 obtained by creating 100 random sequences with SeqBoot from the phylip package
203 (Felsenstein, 2005) and then reconstructing the tree for each sequence. A consensus tree is
204 finally inferred using phylip. In addition, we constructed the species-tree with a supertree
205 method implemented in DupTree using all the trees in the phylome (Wehe et al., 2008).
206

207 **2.5. Gene tree of the *URA1* and *ADH* genes**

208 Homologs of *URA1* from 10 yeast species were retrieved from UniProtKB (45). A HMMER
209 profile was then derived aligning the sequences with MUSCLE v3.7 and then using
210 HMMER3 (Eddy, 2011). The *URA1* profile was then used to search for homologs in the *D.*
211 *bruxellensis* proteome database and in the complete local fungi proteome database. The
212 homologs found in the above-mentioned 21 yeast species were used to reconstruct a
213 phylogenetic tree. These were aligned using MUSCLE v3.7 and then trimmed using trimAl
214 (gap-score cutoff 0.9, conservation score 0.33). A ML phylogenetic tree was obtained using
215 PhyML, the LG model and four rate categories was used. The fraction of invariant positions
216 was inferred from the data and branch support was computed using aLRT. A similar analysis
217 was performed for the *ADH* genes using homologs of *S. cerevisiae ADH1-7* and searching in
218 the genomes of *D. bruxellensis*, *Kluyveromyces lactis*, *P. pastoris* and *C. albicans*.
219

220 **2.6. Analysis of duplicated sequences**

221 *D. bruxellensis* genome was split in non-overlapping regions spanning 2000 and 5000
222 nucleotides. Each sub-sequence was then used to do a local blast (Smith and Waterman, 1981)
223 search (e-value < 1e-05, a continuous overlapping region longer than one-third of the query's
224 total length) against the whole *D. bruxellensis* genome. The number of fragments with 2 or 3

225 hits, to exclude highly repetitive sequences (such as transposons), with similarity higher than
226 70%, 80% and 90% were recorded. For comparison, we applied the same method to *S.*
227 *cerevisiae*, *K. lactis*, and *C. albicans*. In addition, we scanned *D. bruxellensis* phylome as well
228 as the phylomes from *C. albicans* and *S. cerevisiae* deposited in phylomeDB (Huerta-Cepas et
229 al., 2011) to detect and date lineage-specific duplications using a phylogeny-based dating
230 methodology (Huerta-Cepas and Gabaldón, 2011). The relative number of duplication events
231 per gene at each lineage of interest was estimated by dividing the number of duplication
232 events detected at that stage by the number of trees rooted at a deeper branching point; for
233 example, from a tree rooted on the sequences of *Y. lipolytica*, only duplications following the
234 split between this species and *Saccharomyceteceae* were taken into account.
235

236 **2.7. Anaerobic plate tests**

237 Anaerobic experiments were performed using Anaerocult A system on plates containing
238 YNB-based media (Rozpędowska et al., 2011) with and without supplements (aminoacids
239 mixture or peptone), an deither with or without uracil (50 mg/l). The environment contained
240 less than 1 p.p.m. of oxygen. Positive (*S. cerevisiae*) and negative (*K. lactis*) controls were
241 used.
242
243

244 **3. Results and Discussion**

245

246 **3.1. General genome parameters**

247 The 13.4 Megabase genome of *D. bruxellensis* CBS 2499 was sequenced using a combination
248 of 454 and Illumina platforms, assembled with AllPaths assembler and annotated using JGI
249 annotation pipeline to predict 5,600 genes (Table 1 A, B, C, D, E). The obtained genome size

250 is significantly smaller from the one deduced from the previously determined partial sequence
251 (Woolfit et al., 2007). The previous wrong prediction was likely due to the problems with
252 ploidy because *D. bruxellensis* is not a simple haploid but rather contains several recently
253 duplicated, and therefore more or less identical, genome segments (Hellborg and Piskur,
254 2009). Also the number of putative genes is smaller than the previously suggested 7,500.
255 Approximately three quarters of the predicted genes were functionally annotated and over
256 90% were expressed (Table 1). The total number of scaffolds was 84 and the number of larger
257 scaffolds (over 50 kb) was 21, which is higher than the estimated chromosome number, which
258 varies between 4 and 9 among different strains of this species (Hellborg and Piskur, 2009).
259

260 **3.2. Phylogenomics analyses**

261 In order to get an accurate view of the evolution of *D. bruxellensis,* we reconstructed the
262 complete collection of evolutionary histories of its genes in the context of 21 closely related
263 fungal species. This *phylome,* which is accessible through phylomeDB (Huerta-Cepas et al.,
264 2011, [http://phylomedb.org]) was used to predict orthology and paralogy relationships using
265 phylogenetic criteria (Gabaldón, 2008). A super-tree derived from the 3,930 individual gene
266 trees in the phylome using the Gene Tree Parsimony approach implemented in duptree (Wehe
267 et al., 2008) was constructed. In addition, 347 protein families with one-to-one orthology
268 relationships in all the species considered were used to reconstruct a Maximum Likelihood
269 species tree. Both approaches yielded an identical, highly-supported topology that
270 surprisingly places *D. bruxellensis* as a sister-group to *Pichia(Komagataella) pastoris* (Figure
271 1). The *Komagataella* genus and its closest relatives are known as aerobic poor ethanol
272 producer yeasts (reviewed in De Shutter et al., 2009), just opposite to *D. bruxellensis* and *S.*
273 *cerevisiae*.
274

275 **3.3. DHODase encoding genes and anaerobic properties**

276 The acquisition of *URA1* by *S. cerevisiae* promoted synthesis of pyrimidines in the absence of
277 oxygen and therefore provided one of the steps towards adaptation of this lineage to an
278 anaerobic life-style (Gojkovic et al., 2004). The horizontal gene transfer event took place at
279 the base of *Saccharomycetaceae*, and thus much later than the separation of the *S. cerevisiae*
280 and *D. bruxellensis* lineages (see Figure 1). As *D. bruxellensis* shares numerous traits with *S.*
281 *cerevisiae*, we searched for the presence of *URA1* in the newly sequenced genome. The *URA1*
282 phylogenetic tree was reconstructed from an alignment of homologs detected using a
283 HMMER profile based on yeast *URA1* homologs. As seen in Figure 2, the tree clearly shows
284 two groups. The first one belongs to the ancestral *URA9* gene, which can be found in most
285 eukaryotic species and encodes a mitochondrial respiratory chain associated DHODase. This
286 gene was lost in *S. cerevisiae* after the acquisition of the prokaryotic *URA1* gene (Gojkovic et
287 al., 2004). the second group in our analysis (Figure 2) contains orthologs of this gene. The
288 only homologous sequence found in *D. bruxellensis* clearly grouped with the *URA9* genes,
289 discarding the possibilities that (i) the transfer occurred earlier than predicted and (ii) that a
290 second gene transfer took place. However, *D. bruxellensis* can grow anaerobically on the
291 minimal medium without externally provided uracil (Figure 3). The anaerobic growth was
292 fully promoted if a defined mix of amino acids was added to the minimal medium, and the
293 ability to grow in the absence of uracil could be crucial to survival durning the anaerobic
294 phase of wine and beer fermentations since uracil levels are generally low in these
295 environments. Also *Candida glabrata* (a close relative of *S. cerevisiae,* see Figure 1)*,* which
296 only has an *URA9* ortholog (and has lost its *URA1*)*,* does not need uracil for anaerobic growth
297 (Figure 3). Apparently, in these two lineages different evolutionary mechanisms must have
298 operated to establish independence of the *de novo* pyrimidine biosynthesis from the presence

299 of oxygen. An alternative solution could be that the *URA9* gene encoded DHODase adopted a
300 novel acceptor of electrons, independent of the active respiratory chain.
301
302 **3.4. Duplicated genes**
303 The WGD event, thought to have occurred app. 100 mya, was deemed important for the
304 adaptations of *S. cerevisiae* to a fermentative life-style, for example, because the genes
305 encoding the glycolytic pathway were duplicated (reviewed in Piskur et al., 2006). We thus
306 investigated whether *D. bruxellensis* demonstrated any trace of recent larger gene duplication
307 events. Analysis of duplicated regions in *D. bruxellensis, S. cerevisiae* and other species
308 (Table 3) shows that *D. bruxellensis* displays a much lower number of duplicated regions as
309 compared to *S. cerevisiae* and the deduced level of segment duplications is within the range of
310 the non-WGD species *Candida albicans*. In addition, we scanned the *D. bruxellensis* phylome
311 to measure the relative number of gene families duplicated specifically in the *D. bruxellensis*
312 lineage, as compared to others (Figure 4). The results indicate a very small fraction of gene
313 families exhibiting a *Dekkera*-specific duplication, this is much lower than those observed in
314 the *S. cerevisiae* lineage and even lower to those observed in the non-WGD species *C.*
315 *albicans* clade. Thus both results, from repeated genome segments and phylogenetic analysis
316 of gene duplicates, suggest that a WGD-like event has not occurred in the lineage leading to
317 *D. bruxellensis*. The apparent lower number of duplicated genes may be one of the reasons
318 that *D. bruxellensis* has a lower fermentation capacity (Rozpędowska et al., 2011) than *S.*
319 *cerevisiae.*
320
321 The *ADH* genes are crucial in yeast to promote the ability to ferment sugars into alcohol and
322 to generate some aromatic compounds. In *S. cerevisiae*, there are seven *ADH* genes (*ADH1-*
323 *7*), and five of them, *ADH1-5*, encode alcohol dehydrogenases involved in the catalysis of the

324 reversible conversion of aldehydes to ethanol. Four of the corresponding enzymes, encoded
325 by *ADH1*, *ADH3*, *ADH4*, and *ADH5*, reduce acetaldehyde to ethanol during glucose
326 fermentation, while the *ADH2* encoded enzyme catalyzes the reverse reaction and oxidizes
327 ethanol to acetaldehyde. The *ADH1* and *ADH2* represent a recent lineage-specific duplication,
328 providing a very efficient regulation check-point for the ethanol accumulation and ethanol
329 degradation metabolic activities (Thomson et al., 2005). When we analysed homologs of these
330 genes in four species we found that in the *ADH1,2,3,5* group there is also a lineage-specific
331 duplication in *D. bruxellensis* (Fig. 5). The three recently duplicated genes, which show a
332 high degree of similarity, were not found in the closest relative *P. pastoris*, which is a
333 Crabtree-negative yeast, could have in *D. bruxellensis* a similar function as the *ADH1* and
334 *ADH2* genes in *S. cerevisiae* and these duplications represent a parallel evolutionary event.
335 Regarding the group of *ADH6* and *ADH7*, which are in *S. cerevisiae* involved in the
336 conversion of longer chain aldehydes and alcohols, one can again see several *D. bruxellensis*
337 lineage-specific duplicates. *S. cerevisiae ADH6* and *ADH7* are involved in the synthesis of
338 aromatic compounds (higher alcohols) and pre-cursors for aromatic esters. Our observation of
339 the presence of several duplicated *ADH6-7*-like genes coincides with the previous
340 observations that *D. bruxellensis* has a very intensive aromatic profile (Licker et al., 1999).
341
342 **3.5. Conclusion: genome, evolution and food-related properties**
343 The comparative analysis of the genome sequences of *S. cerevisiae* and *D. bruxellensis*
344 revealed that the two lineages employed different or similar molecular mechanisms to evolve
345 several similar traits. The WGD event and the lateral acquisition of genes needed for
346 anaerobic growth such as *URA1* from other organisms were some of the events necessary for
347 the establishment of a modern fermentative and anaerobic life style in the *Saccharomyces*
348 lineage. *D. bruxellensis* has independently evolved into an organism able to grow under

349 anaerobic conditions, producing large amounts of ethanol and tolerating high ethanol levels.
350 Under oxygen limitation, the ethanol yield of *D. bruxellensis* is almost the same as in *S.*
351 *cerevisiae* (Galafassi et al., 2011). However, under aerobic conditions *D. bruxellensis*
352 produces less ethanol but has higher biomass than *S. cerevisiae* (Blomqvist et al., 2010)
353 suggesting a less pronounced Crabtree effect. We show here that in contrast to *S. cerevisiae*,
354 *D. bruxellensis* does not show traces of extensive gene duplications. On the other hand, both
355 lineages used the same strategy with promoter rewiring in genes associated with the
356 respiration (Ihmels et al., 2005; Rozpędowska et al., 2011), and likely with the *ADH* genes
357 duplication, which promotes ultimate separation of the fermentation process from ethanol
358 consumption.
359
360 *D. bruxellensis* also shows greater diversity among strains in chromosome number and ploidy
361 than does *S. cerevisiae* (Hellborg and Piskur, 2009), suggesting that the increase in the gene
362 dose/ploidy could be an important event in establishment of the yeasts in sugar-rich anaerobic
363 food fermentation habitats. In the *S. cerevisiae* lineage this was achieved by WGD, for
364 example duplication of the genes involved in glycolysis, but in the *D. bruxellensis* lineage it is
365 apparently achieved through increased ploidy. The differences in production of some
366 components that have a flavor impact by these two yeast species may also be due to the
367 observed differences in the genome content, for example duplication of genes involved in
368 generation of higher alcohols (Figure 5). The availability of the whole genome sequence now
369 provides a tool to deduce the enzymatic background for production of off-flavor compounds.
370 In conclusion, this work opens many opportunities to examine the genetic background for
371 food-related properties as well as to understand the evolutionary processes behind evolution
372 of the fermentative metabolism and the ability of these yeasts to establish themselves in
373 anaerobic niches.

374

375

376 **Acknowledgements**

377

378 JP acknowledges a Slovenian ARRS grant and a Swedish Sörensen Foundation grant to
379 support his work on wine yeasts. TG and MMH acknowledge a grant from the Spanish
380 Ministry of Science (BFU09-09268). The sequencing and annotations were conducted by the
381 U.S. Department of Energy Joint Genome Institute and was supported by the Office of
382 Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

383

384 **References**

385

386 Akaike, M., 1973. Information theory and extension of the maximum likelihood principle.
387 Proceedings of the 2nd international symposium on information theory, pp. 267-281.

388 Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., 1990. Basic local alignment
389 search tool. Journal of Molecular Biology 215, 403-410.

390 Birney, E., Durbin, R., 2000. Using GeneWise in the Drosophila annotation experiment.
391 Genome Research 10, 547-548.

392 Blomqvist, J., Eberhard, T., Schnürer, J., Passoth, V., 2010. Fermentation characteristics of
393 *Dekkera bruxellensis* strains. Applied Microbiology and Biotechnology 87, 1487-1497.

394 Capella-Gutiérrez, S., Silla-Martínez, J. M., Gabaldón, T., 2009. trimAl: a tool for automated
395 alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972-1973.

396 De Schutter, K., 2009. Genome sequence of the recombinant protein production host *Pichia*
397 *pastoris*. Nature Biotechnology 27, 561-566.

398 Du Toit, M., Pretorius, I. S., 2000. Microbial spoilage and preservation of wine: Using

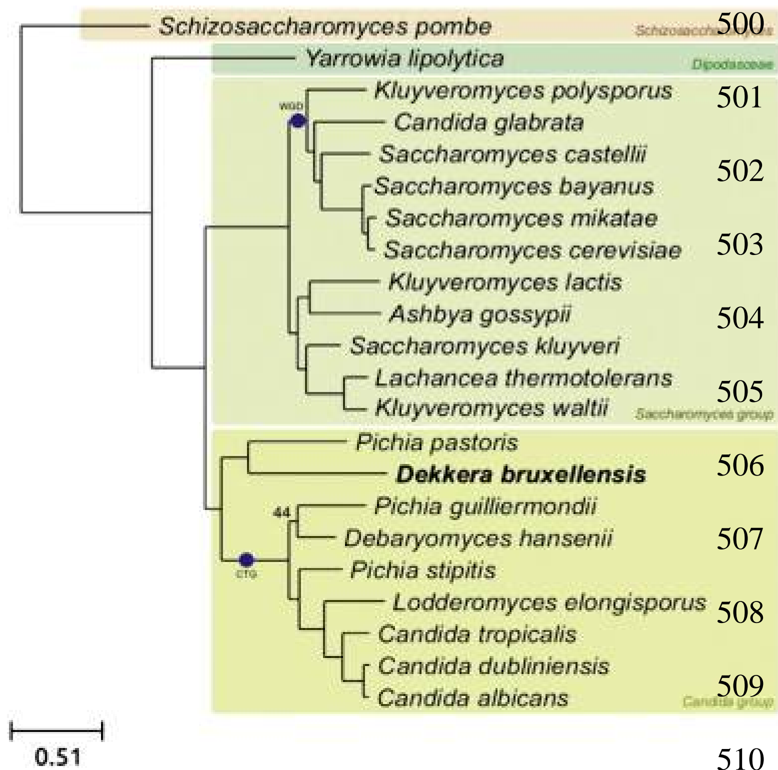
399 weapons from nature's own arsenal. South African Journal of Enology and Viticulture
400 21, 74-96.
401 Eddy, S. R., 2011. Accelerated Profile HMM Searches. PLoS Computational Biology 7,
402 e1002195.
403 Edgar, R. C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high
404 throughput. Nucleic Acids Research 32, 1792-1797.
405 Enright, A. J., Van Dongen, S., Ouzounis, C. A., 2002. An efficient algorithm for large-scale
406 detection of protein families. Nucleic Acids Research 30, 1575-1584.
407 Felsenstein, J., 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the
408 author. Department of Genome Sciences, University of Washington, Seattle.
409 Gabaldón, T., 2008. Large-scale assignment of orthology: back to phylogenetics? Genome
410 Biology 9, 235.
411 Galafassi, S., Merico, A., Pizza, F., Hellborg, L., Molinari, F., Piskur, J., Compagno, C., 2011.
412 *Dekkera/Brettanomyces* yeasts for ethanol production from renewable sources under
413 oxygen-limited and low-pH conditions. Journal of Industrial Microbiology and
414 Biotechnology 38, 1079-1088.
415 Gascuel, O., 1997. BIONJ: an improved version of the NJ algorithm based on a simple model
416 of sequence data. Molecular Biology and Evolution 14, 685-695.
417 Gojković, Z., 2004. Horizontal gene transfer promoted evolution of the ability to propagate
418 under anaerobic conditions in yeasts. Molecular Genetics and Genomics 271, 387-393.
419 Guindon, S., Delsuc, F., Dufayard, J. F., Gascuel, O., 2009. Estimating maximum likelihood
420 phylogenies with PhyML. Methods in Molecular Biology 537, 113-137.
421 Hellborg, L., Piskur, J., 2009. Complex nature of the genome in a wine spoilage yeast,
422 *Dekkera bruxellensis*. Eukaryotic Cell 8, 1739-1749.
423 Heresztyn, T., 1986. Metabolism of volatile phenolic compounds from hydroxycinnamic acids

424 by *Brettanomyces* strains. Archives of Microbiology 146, 96-98.
425 Huerta-Cepas, J., 2011. PhylomeDB v3.0: an expanding repository of genome-wide
426 collections of trees, alignments and phylogeny-based orthology and paralogy
427 predictions. Nucleic Acids Research 39, D556-560.
428 Huerta-Cepas, J., Dopazo, J., Gabaldón, T., 2010. ETE: a python Environment for Tree
429 Exploration. BMC Bioinformatics 11, 24.
430 Huerta-Cepas, J., Gabaldón, T., 2011. Assigning duplication events to relative temporal scales
431 in genome-wide studies. Bioinformatics 27, 38-45.
432 Ihmels, J., Bergmann, S., Gerami-Nejad, M., Yanai, I., McClellan, M., Berman, J., Barkai, N.,
433 2005. Rewiring of the yeast transcriptional network through the evolution of motif
434 usage. Science 309, 938-940.
435 Isono, K., McIninch, J. D., Borodovsky, M., 1994. Characteristic features of the nucleotide
436 sequences of yeast mitochondrial ribosomal protein genes as analyzed by computer
437 program GeneMark. DNA Research 1, 263-269.
438 Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005.
439 Repbase Update, a database of eukaryotic repetitive elements. Cytogenetics and
440 Genome Research 110, 462-467.
441 Kanehisa, M., , 2008. KEGG for linking genomes to life and the environment. Nucleic Acids
442 Research 36, D480-484.
443 Katoh, K., Toh, H., 2008. Recent developments in the MAFFT multiple sequence alignment
444 program. Briefings in Bioinformatics 9, 286-298.
445 Kent, W. J., 2002. BLAT--the BLAST-like alignment tool. Genome Research 12, 656-664.
446 Koonin, E.V., 2004. A comprehensive evolutionary classification of proteins encoded in
447 complete eukaryotic genomes. Genome Biology 5, R7.
448 Letunic, I., Bork, P., 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic

449 tree display and annotation. Bioinformatics 23, 127-128.
450 Licker, J. L., Acree, T. E., Henick-Kling, T., 1999. What is 'Brett' (*Brettanomyces*) flavour? A
451 preliminary investigation. American Chemical Society Symposium Series 714, 96-115.
452 Melén, K., Krogh, A., von Heijne, G., 2003. Reliability measures for membrane protein
453 topology prediction algorithms. Journal of Molecular Biology 327, 735-744.
454 Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G., 1997. Identification of prokaryotic
455 and eukaryotic signal peptides and prediction of their cleavage sites. Protein
456 Engineering 10, 1-6.
457 Parkhomchuk, D., 2009. Transcriptome analysis by strand-specific sequencing of
458 complementary DNA. Nucleic Acids Research 37, e123.
459 Piskur, J., Langkjaer, R. B., 2004. Yeast genome sequencing: the power of comparative
460 genomics. Molecular Microbiology 53, 381-389.
461 Piskur, J., Rozpędowska, E., Polakova, S., Merico, A., Compagno, C., 2006. How did
462 *Saccharomyces* evolve to become a good brewer? Trends in Genetics 22, 183-186.
463 Rozpędowska, E., Hellborg, L., Ishchuk, O. P., Orhan, F., Galafassi, S., Merico, A., Woolfit,
464 M., Compagno, C., Piskur, J., 2011. Parallel evolution of the make-accumulate-consume
465 strategy in *Saccharomyces* and *Dekkera* yeasts. Nature Communications 2, 302.
466 Salamov, A. A., Solovyev, V. V., 2000. Ab initio gene finding in Drosophila genomic DNA.
467 Genome Research 10, 516-522.
468 Smit, A. F. A., Hubley, R., Green, P. *RepeatMasker Open-3.0*. 1996-2010.
469 Smith, T. F., Waterman, M. S., 1981. Identification of common molecular subsequences.
470 Journal of Molecular Biology 147, 195-197.
471 Stamatakis, A., Ludwig, T., Meier, H., 2005. RAxML-III: a fast program for maximum
472 likelihood-based inference of large phylogenetic trees. Bioinformatics 21, 456-463.
473 Subramanian, A. R., Kaufmann, M., Morgenstern, B., 2008. DIALIGN-TX: greedy and

474 progressive approaches for segment-based multiple sequence alignment. Algorithms for
475 Molecular Biology 3, 6.
476 The UniProt Consortium., 2011. Ongoing and future developments at the Universal Protein
477 Resource. Nucleic Acids Research 39, D214-D219.
478 Thomson, J. M., Gaucher, E. A., Burgan, M. F., De Kee, D. W., Li, T., Aris, J. P., Benner, S. A.,
479 2005. Resurrecting ancestral alcohol dehydrogenases from yeast. Nature Genetics 37,
480 630-635.
481 Vigentini, I., Romano, A., Compagno, C., Merico, A., Molinari, F., Tirelli, A., Foschino, R.,
482 Volonterio, G., 2008. Physiological and oenological traits of different
483 *Dekkera/Brettanomyces bruxellensis* strains under wine-model conditions. FEMS Yeast
484 Research 8, 1087-1096.
485 Wallace, I. M., O'Sullivan, O., Higgins, D. G., Notredame, C., 2006. M-Coffee: combining
486 multiple sequence alignment methods with T-Coffee. Nucleic Acids Research 34, 1692-
487 1699.
488 Wedral, D., Shewfelt, R., Frank, J., 2008. The challenge of *Brettanomyces* in wine. LWT-Food
489 Science and Technology 43, 1474-1479.
490 Wehe, A., Bansal, M. S., Burleigh, J. G., Eulenstein, O., 2008. DupTree: a program for large-
491 scale phylogenetic analyses using gene tree parsimony. Bioinformatics 24, 1540-1541.
492 Woolfit, M., Rozpędowska, E., Piskur, J., Wolfe, K. H., 2007. Genome survey sequencing of
493 the wine spoilage yeast *Dekkera (Brettanomyces) bruxellensis*. Eukaryotic Cell 6, 721-
494 733.
495 Zdobnov, E. M., Apweiler, R., 2001. InterProScan--an integration platform for the signature-
496 recognition methods in InterPro. Bioinformatics 17, 847-848.
497
498

499 **Figures**



500 Figure 1. Phylogenetic
501 relationships within the
502 *Schizosaccharomyces*
503 *Dipodasceael Candida*
504 *Saccharomyces* group. Postion
505 of *D. Bruxellensis* is in bold.
506 Important evolutionaty events
507 such as the WGD, or genetic
508 code alteration in the *Candida*
509 (CTG) clade are indicated. The
510 tree is based in a Masximum
511 Likilihood analysis of a concatenated alignment of 347 proteins with one-to-one
512 orthologs in all species considered. All nodes received the highest support in terms of
513 approximate likelihood ratio test and of a boostrap analysis of 100 replicas. An
514 identical topology was obtained froma super-tree methods combining all trees in *D.*
515 *Bruxellensis* phylome.

516

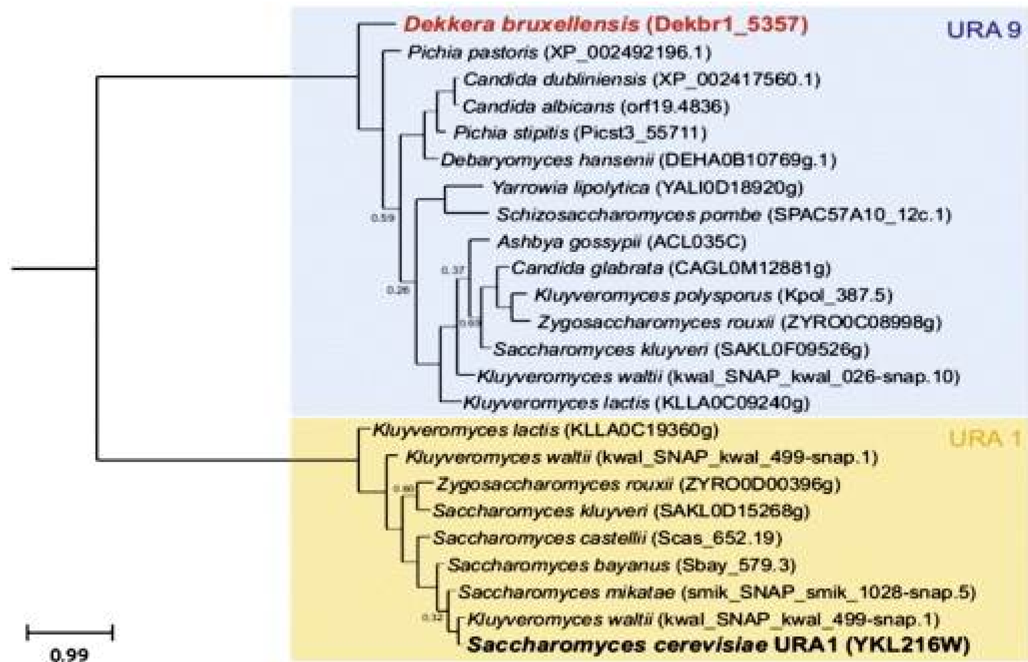
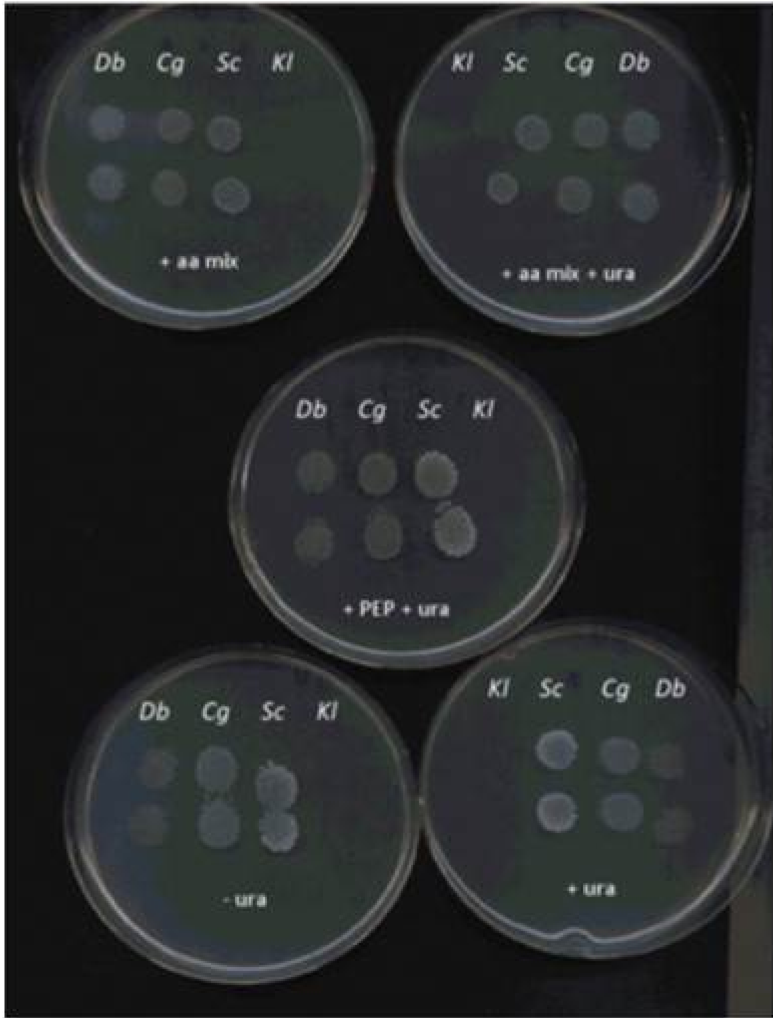


Figure 2. The gene tree of yeast *URA/URA9* honologs, the position of the *D. Bruxellensis* homolog is colored in red and belongs to the *URA9* group (shown with a blue background). The *URA1* group is colored in yellow. Nodes with an LRT value below 0.75 are shown in the tree.

517
518
519
520
521

522



523 Figure 3. Anaerobic growth of *D. Bruxellensis* (CBS2499), *C. Glabrata* (CBS138), *S.*
524 *Cerevisiae* (CEN.PK 113-7D) and *K. Lactis* (CBS2359) on a minimal medium (2%
525 glucose) with different suplements, sucha as uracile, aminoacids mix and peptone.

526

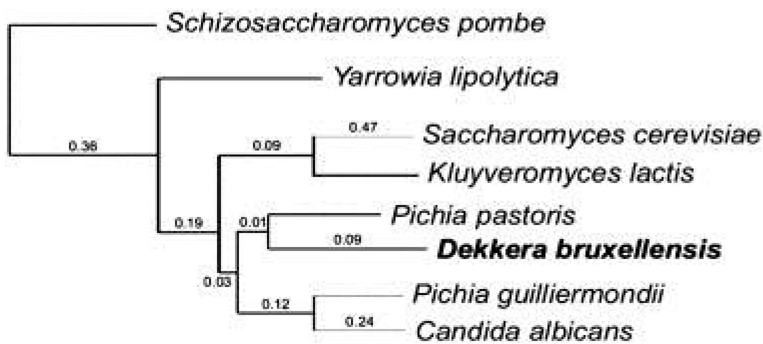
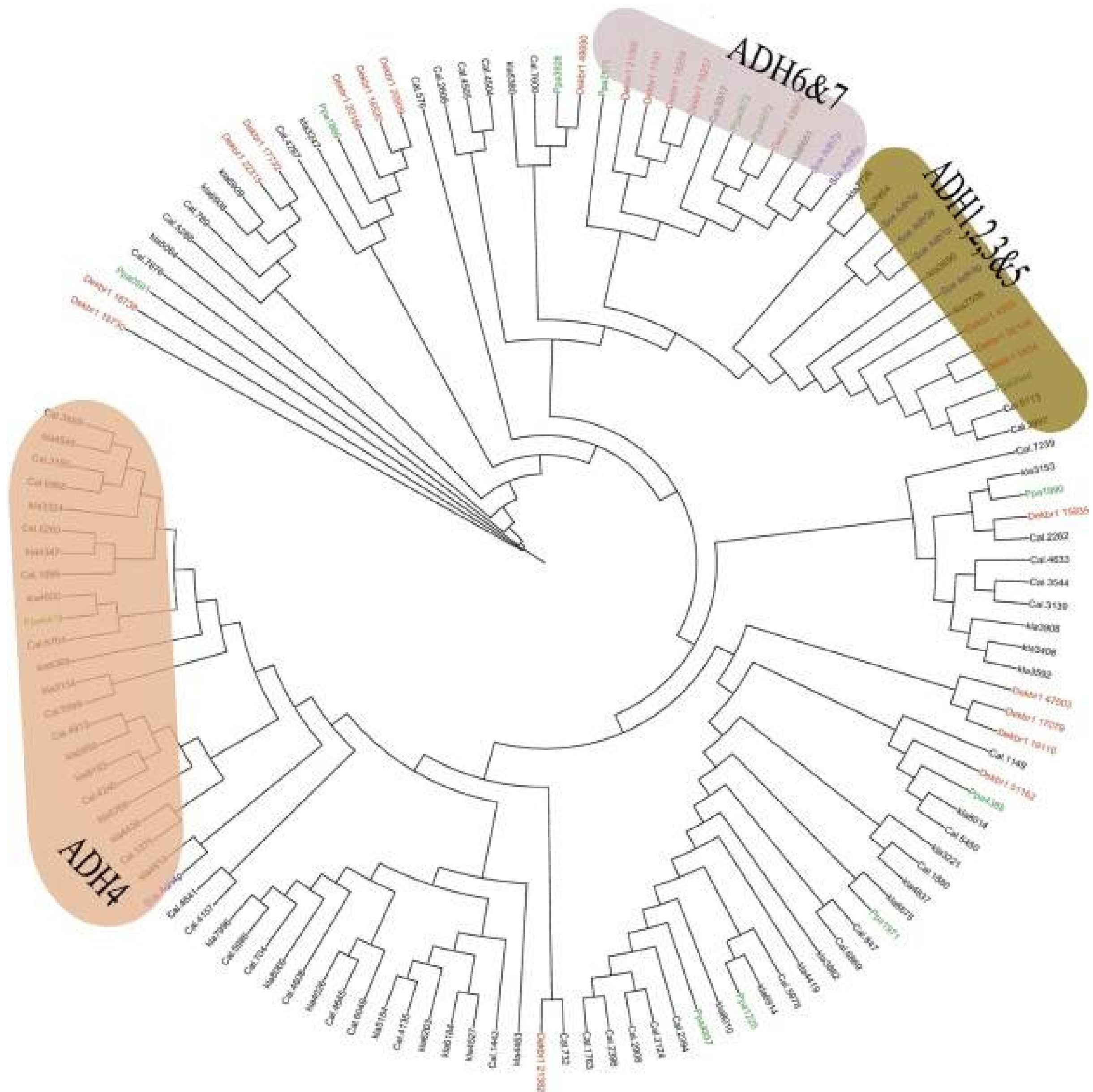


Figure 4. The relative number of duplication events per gene. The number on each branching point indicates the average duplication events per gene detected at each
532 of the indicated lineages, as inferred from anlaysis of *S. Cerevisiae*, *C. Albicans* and *D.*
533 *Bruxellensis* phylomes available at PhylomeDB. Notably, gene duplication rate in *D.*

534 *Bruxellensis* lineage is among the samllest, despite the relatively long branch since its

535 separation from *P. Pastoris.*

536



537

538 Figure 5. A phylogenetic tree of *ADH* homologs of *S. cerevisiae*, *D. Bruxellensis*, *C.*

539 *Albicans* and *P. Pastoris*. The *ADM* genes (*ADH1-7*) of *S. Cerevisiae* are colored in

540 blue while the the homologs of *D. Bruxellensis* are colored in red and *P. Pastoris* in

541 green. The figure was produced with iTOL (Letunic and Bork, 2007).

542 **Tables**

543 Table 1. Details on the genome sequence and annotation.

A. Genome sequencing statistics

Library type/name	Lib_stats	#_reads	%_used	scov	# pairs	pcov
Frag libs:						
GXXW	59 ± 26	815,602	92.3	4.5	365,052	6.3
ICHI_270bp_2 × 150	− 77 ± 56	2,920,094	91.5	31.3	1,282,098	24.8
Total		3,735,696	91.7	35.8	1,647,150	31.1
Jump libs:						
ICCY_4kb_2 × 100	3343 ± 479	33,605,128	78.2	63.6	11,131,517	3077.8
WGSIM	2836 ± 146	6,029,486	78.4	28.1	1,634,428	376.1
Total		39,634,614	78.3	91.7	12,765,945	3453.9
Long jump lib:						
WGSIM_15kb_2 × 76	14,848 ± 100	263,156	90.9	1.4	28,960	34.7

B. Final assembly statistics

Main genome scaffold total	84
Main genome contig total	729
Main genome scaffold sequence total	13.4 Mb
Main genome contig sequence total	12.7 Mb
Main genome scaffold N/L50	3/1.79 Mb
Main genome contig N/L50	91/34.4 Kb
Number of scaffolds > 50 KB	21
% main genome in scaffolds > 50 KB	97.9%

C. Characteristics of predicted gene models

	Average	Median
Gene length, bp	1631	1384
Protein length, aa	457	382
Exon frequency	1.44 exons/gene	1 exon/gene
Exon length, bp	1067	848
Intron length, bp	216	86

D. Predicted gene models and supporting lines of evidence

# gene models:	5600
% complete (with start and stop codons):	92%
% genes with homology support:	87%
% genes with Pfam domains:	66%
% genes with EST support:	91%
E. Functional annotation of proteins	
Proteins assigned to a KOG	4088 (73%)
KOG categories genome-wide	2741
Proteins assigned a GO term	3332 (60%)
GO terms genome-wide	1943
Proteins assigned an EC number	1588 (28%)
EC numbers genome-wide	622
Proteins assigned a Pfam domain	3700 (66%)
Pfam domains genome wide	2005

547

Table 2. The source of the genome sequences.

A. The source of genome sequence downloading

Species name	Source
Ashbya gossypii	UNIPROT
Candida albicans	Quest For Orthologs

A. The source of genome sequence downloading	
Species name	Source
Candida dubliniensis	Sanger
Candida glabrata	UNIPROT
Candida tropicalis	Broad_Institute
Debaryomyces hansenii	UNIPROT
Dekkera bruxellensis	JGI
Kluyveromyces lactis	UNIPROT
Kluyveromyces polysporus	YGOB
Kluyveromyces waltii	YGOB
Lachancea thermotolerans	Genolevures
Lodderomyces elongisporus	Broad_Institute
Pichia guilliermondii	Broad_Institute
Pichia pastoris	Ghent university
Pichia stipitis	integr8
Saccharomyces bayanus	YGOB
Saccharomyces castellii	YGOB
Saccharomyces cerevisiae	Quest For Orthologs
Saccharomyces kluyveri	Genolevures
Saccharomyces mikatae	SGD
Schizosaccharomyces pombe	Quest For Orthologs
Yarrowia lipolytica	Quest For Orthologs
B. The website of the source	
Sites	
Quest for Orthologs	http://www.ebi.ac.uk/reference_proteomes/
UNIPROT	http://www.uniprot.org/
Sanger	http://www.sanger.ac.uk/
Broad_Institute	http://www.broadinstitute.org/
JGI	http://www.jgi.doe.gov/
Genolevures	http://www.genolevures.org/
YGOB	http://wolfe.gen.tcd.ie/ygob/
integr8	http://www.ebi.ac.uk/integr8/

548