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.

International Journal of Food Microbiology

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

Corresponding author: J. Piskur, Department of Biology, Biologihuset A, Lund University, Soelvegatan 35, SE-22362 Lund, Sweden, e-mail: Jure.Piskur@cob.lu.se, phone: +46 46 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

There is an enormous diversity among yeast species, including those that play important roles in traditional food processes, often in mixed cultures in spontaneous fermentations. One such yeast is *Dekkera/Brettanomyces bruxellensis,* associated with lambic beer fermentation and wine production, especially as a contributor, in a positive or negative manner, to flavour development (Du Toit and Pretorius, 2000). This yeast can produce phenolic compounds, such as 4-ethylguaiacol and 4-ethylphenol, which could lead to wine spoilage if present in high enough concentration (Heresztyn, 1986; Vigentini et al., 2008). In fact, *D. bruxellensis* represents a serious problem in wine industry, causing enormous economic losses as a consequence of wine spoilage (Wedral et al., 2010). However, in spite of the economic impact of *D. bruxellensis,* this yeast remains poorly studied.

D. bruxellensis is apparently not a close relative of baker's yeast *Saccharomyces cerevisiae*, but the phylogenetic position of the *D. bruxellensis* group has so far been rather impossible to determine (Woolfit et al., 2007). Both yeasts share several “peculiar” and rather “unusual” traits important for food-related properties, such as production of high ethanol levels, high tolerance towards ethanol, and the ability to grow without oxygen and in acidic environments (Rozpędowska et al., 2011). Apparently, given the lack of relatedness, these traits evolved in parallel in both groups, but it is unclear if the molecular mechanisms behind these properties are similar or different (Rozpędowska et al., 2011). In other words, these two yeasts represent an ideal model to study molecular processes involved in convergent and parallel evolutionary routes.

Ethanol production and capability to survive without oxygen are highly relevant in food fermentations. In *S. cerevisiae*, but not in *D. bruxellensis*, the corresponding genetic factors that underlie these traits have been relatively well studied. For example, the whole genome duplication (WGD), duplicated gene profiles, the horizontal transfer of the *URA1* gene (coding for the DHODase, dihydroorotate dehydrogenase, catalysing the fourth pyrimidine *de novo* pathway step), and lineage-specific duplication of the *ADH* genes (encoding alcohol dehydrogenases), have been shown to be at least partially responsible for development of the *S. cerevisiae* high fermentation capacity and/or anaerobic properties (reviewed in Piskur and Langkjaer, 2004; Piskur et al., 2006). It is not known whether similar molecular strategies are responsible for the domination of the same environment by *D. bruxellensis*.

Recently, a partial genome sequence of one strain of *D. bruxellensis* has been reported, and the analysis estimated that this yeast has around 7.500 genes, of which many lack a homolog in the *S. cerevisiae* genome (Woolfit et al., 2007). Further analysis of the partial sequence has revealed that *D. bruxellensis* is not a simple haploid. Its genome contains approximately 1% polymorphic sites but the exact physical background for this heterozygocity is not known (Hellborg and Piskur, 2009).

Here we determined the whole genome sequence of the *D. bruxellensis* strain Y879 (CBS2499) and used it to deduce several “food-relevant” properties and evolution pathways of this yeast.

2. Materials and Methods

2.1. Genome sequencing and assembly

The genome of *D. bruxellensis* strain Y879 (CBS2499) was sequenced using a combination of 454 and Illumina sequencing platforms (GYBS 454 standard rapid, GYHO 454 standard rapid, GYHG 454 titanium 4kb, GYFW 454 titanium 4kb, GXXW Illumina 2x76 300bp, ICHI Illumna 2x150 270bp, and ICCY Illumina 2x100 4kb CLIP). All general aspects of library construction and sequencing can be found at the JGI website (http://www.jgi.doe.gov/). An initial assembly of GXXW was conducted for QC purposes using the Velvet assembler, version 0.7.55, with the following parameters: k 61 -min_contig_lgth 100 -exp_cov 81. A list of data to be excluded from the draft assembly was also created by identifying possible contaminant data in preliminary Newbler assemblies of the 454 data.   The resulting screened data was assembled along with shredded consensus from the initial Velvet assembly using the Newbler assembler, software release 2.5-internal-10Apr08-1, with the following parameters: -fe reads2remove.MPA -info -consed -finish -nrm –rip -sio a 50 -l 350 -g -ml 30 -mi 94 -e 87. The final draft assembly was assembled from the Illumina data, as well as 3kb and 15kb paired end data generated from the Newbler assembly using wgsim, with the AllpathsLG assembler software release R38445, to an estimated assembled coverage of 128x (Table 1A) with 84 scaffolds with an N50 of 1.7 Mb, and 880 contigs with an N50 of 30.5 Kb (Table 1B).

2.2. EST sequencing and assembly

Total RNA from two separate *D. bruxellensis* samples, “air” and “no air” were used to generate stranded RNASeq libraries. mRNA was purified from total RNA using the Absolutely mRNA™ purification kit (Stratagene,Santa Clara, CA). Subsequently, the mRNA samples were chemically fragmented to the size range 200-250 bp using 1x fragmentation solution for 5 minutes at 70 □ (RNA Fragmentation Reagents, AM8740 – Zn, Ambion, Carlsbad, CA). First strand cDNA was synthesized using Superscript II Reverse Transcriptase (Invitrogen, Carlsbad, CA) and random hexamers then the second strand was synthesized using *E. coli* RnaseH, DNA Ligase, and DNA polymerase I for nick translation. The dscDNA was then cleaned up using a double SPRI bead selection (Agencourt Ampure beads; Beckman Coulter, Brea CA). The TruSeq Sample Prep kit (Illumina Inc. San Diego, CA) was used for RNASeq library creation using the dscDNA and the manufacturer's instructions (Illumina). Briefly, dscDNA was end repaired, and ligated to Illumina adaptors. Then the second strand was removed by AmpErase UNG (Applied Biosystems, Carlsbad, CA) similar to the method described by (Parkhomchuk et al., 2009). Paired end 100 bp reads were generated by sequencing using the Illumina HiSeq instrument. 176,820,692 and 159,263,276 reads were generated for the "air" and "no air" samples respectively. Newbler assembled consensus EST sequence data was used to assess the completeness of the final genome assembly Fasta with alignment using 90% identity and 85% coverage thresholds. This resulted in 89.16% placement.

2.3. Genome Annotation

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

Genomic assembly scaffolds were masked using RepeatMasker (Smit et al., 2010) and the RepBase library of 234 fungal repeats (Jurka et al., 2005) and RepeatScout. Using the repeat-masked assembly, several gene prediction programs falling into three general categories were used: 1) *ab initio* - FGENESH (Salamov and Solovyev, 2000); GeneMark (Isono et al., 1994), 2) *homology-based* - FGENESH+; Genewise (Briney and Durbin, 2000) seeded by BLASTx alignments against GenBank's database of non-redundant proteins (NR: http://www.ncbi.nlm.nih.gov/BLAST/), and 3) *EST-based* - EST_map (http://www.softberry.com/) seeded by EST contigs. Genewise models were extended where possible using scaffold data to find start and stop codons. EST BLAT alignments (Kent, 2002) were used to extend, verify, and complete the predicted gene models. The resulting set of models was then filtered for the best models, based on EST and homology support, to produce a non-redundant representative set (see Table 1C). This representative set was subject to further analysis and manual curation. Measures of model quality include proportions of the models complete with start and stop codons (92%) consistent with ESTs (91%) supported by similarity with proteins from the NCBI NR database (87%) Quality metrics for gene models are summarized in Table 1D.

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

2.4. Phylome reconstruction

The *D. bruxellensis* CBS2499 predicted proteome described above, and those from a collection of 21 completely sequenced fungal genomes were downloaded from various sources (see Table 2, A&B). Using the phylomeDB pipeline (Huerta-Cepas et al., 2011) we reconstructed the complete collection of evolutionary histories of *D. bruxellensis* genes, i.e the phylome. In brief, the phylogenetic reconstruction pipeline involves Smith-Waterman searches for homologs (E-Value <1e-05, 50% sequence overlap) accross 21 related fungal species including: *Schizosaccharomyces pombe* and *Yarrowia lipolytica* as outgroups. These homologous groups are then aligned using 3 different programs, MUSCLE v3.7 (Edgar, 2004), MAFFT v6.712b (Katoh, 2008), and DIALIGN-TX (Subramanian, 2008), and in forward and reverse direction (i.e using the Head or Tail approach). The 6 resulting alignments were then combined with M-COFFEE (Wallace et al., 2006) and then trimmed with trimAl v1.3 (Capella-Gutiérrez et al., 2009) using consistency-score cutoff 0.1667 and gap-score cutoff 0.9. Multiple sequence alignments were then used to reconstruct maximum likelihood tree. For each gene, the best evolutionary model was chosen among seven competing models (JTT, LG, WAG, Blosum62, MtREV, VT and Dayhoff) reconstructing a NJ tree, using bioNJ (Gascuel, 1997) as implemented in PhyML (Guindon, 2009); The 2 best-fitting models, as determined by the AIC criterion (Akaike, 1973), were used to derive ML trees. The model used four rate categories and the fraction of invariant positions was inferred from the data. Branch support was computed using an aLRT (approximate likelihood ratio test) based on a chi-square distribution. Resulting trees and alignments are stored in phylomeDB, with the phylomeID 138 (www.phylomedb.org). Orthology and paralogy relationships for each gene in the phylome were obtained using the species-overlap algorithm implemented in ETE (Huerta-Cepas et al., 2010).

347 genes, which had a strict and phylogeny-based one-to-one orthology relationship in all species included in the phylome, were concatenated into a single alignment and then trimmed using trimAl (gap-score cutoff 0.5, conservation score 0.5). The species-tree was

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

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

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

2.6. Analysis of duplicated sequences

D. bruxellensis genome was split in non-overlapping regions spanning 2000 and 5000 nucleotides. Each sub-sequence was then used to do a local blast (Smith and Waterman, 1981) search (e-value < 1e-05, a continuous overlapping region longer than one-third of the query's total length) against the whole *D. bruxellensis* genome. The number of fragments with 2 or 3 hits, to exclude highly repetitive sequences (such as transposons), with similarity higher than 70%, 80% and 90% were recorded. For comparison, we applied the same method to *S. cerevisiae*, *K. lactis*, and *C. albicans*. In addition, we scanned *D. bruxellensis* phylome as well as the phylomes from *C. albicans* and *S. cerevisiae* deposited in phylomeDB (Huerta-Cepas et al., 2011) to detect and date lineage-specific duplications using a phylogeny-based dating methodology (Huerta-Cepas and Gabaldón, 2011). The relative number of duplication events per gene at each lineage of interest was estimated by dividing the number of duplication events detected at that stage by the number of trees rooted at a deeper branching point; for example, from a tree rooted on the sequences of *Y. lipolytica*, only duplications following the split between this species and *Saccharomyceteceae* were taken into account.

2.7. Anaerobic plate tests

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

3. Results and Discussion

3.1. General genome parameters

The 13.4 Megabase genome of *D. bruxellensis* CBS 2499 was sequenced using a combination of 454 and Illumina platforms, assembled with AllPaths assembler and annotated using JGI annotation pipeline to predict 5,600 genes (Table 1 A, B, C, D, E). The obtained genome size is significantly smaller from the one deduced from the previously determined partial sequence (Woolfit et al., 2007). The previous wrong prediction was likely due to the problems with ploidy because *D. bruxellensis* is not a simple haploid but rather contains several recently duplicated, and therefore more or less identical, genome segments (Hellborg and Piskur, 2009). Also the number of putative genes is smaller than the previously suggested 7,500. Approximately three quarters of the predicted genes were functionally annotated and over 90% were expressed (Table 1). The total number of scaffolds was 84 and the number of larger scaffolds (over 50 kb) was 21, which is higher than the estimated chromosome number, which varies between 4 and 9 among different strains of this species (Hellborg and Piskur, 2009).

3.2. Phylogenomics analyses

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

3.3. DHODase encoding genes and anaerobic properties

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

3.4. Duplicated genes

The WGD event, thought to have occurred app. 100 mya, was deemed important for the adaptations of *S. cerevisiae* to a fermentative life-style, for example, because the genes encoding the glycolytic pathway were duplicated (reviewed in Piskur et al., 2006). We thus investigated whether *D. bruxellensis* demonstrated any trace of recent larger gene duplication events. Analysis of duplicated regions in *D. bruxellensis, S. cerevisiae* and other species (Table 3) shows that *D. bruxellensis* displays a much lower number of duplicated regions as compared to *S. cerevisiae* and the deduced level of segment duplications is within the range of the non-WGD species *Candida albicans*. In addition, we scanned the *D. bruxellensis* phylome to measure the relative number of gene families duplicated specifically in the *D. bruxellensis* lineage, as compared to others (Figure 4). The results indicate a very small fraction of gene families exhibiting a *Dekkera*-specific duplication, this is much lower than those observed in the *S. cerevisiae* lineage and even lower to those observed in the non-WGD species *C. albicans* clade. Thus both results, from repeated genome segments and phylogenetic analysis of gene duplicates, suggest that a WGD-like event has not occurred in the lineage leading to *D. bruxellensis*. The apparent lower number of duplicated genes may be one of the reasons that *D. bruxellensis* has a lower fermentation capacity (Rozpędowska et al., 2011) than *S. cerevisiae.*

The *ADH* genes are crucial in yeast to promote the ability to ferment sugars into alcohol and to generate some aromatic compounds. In *S. cerevisiae,* there are seven *ADH* genes (*ADH1-7*), and five of them, *ADH1-5*, encode alcohol dehydrogenases involved in the catalysis of the reversible conversion of aldehydes to ethanol. Four of the corresponding enzymes, encoded by *ADH1*, *ADH3*, *ADH4*, and *ADH5*, reduce acetaldehyde to ethanol during glucose fermentation, while the *ADH2* encoded enzyme catalyzes the reverse reaction and oxidizes ethanol to acetaldehyde. The *ADH1* and *ADH2* represent a recent lineage-specific duplication, providing a very efficient regulation check-point for the ethanol accumulation and ethanol degradation metabolic activities (Thomson et al., 2005). When we analysed homologs of these genes in four species we found that in the *ADH1,2,3,5* group there is also a lineage-specific duplication in *D. bruxellensis* (Fig. 5). The three recently duplicated genes, which show a high degree of similarity, were not found in the closest relative *P. pastoris*, which is a Crabtree-negative yeast, could have in *D. bruxellensis* a similar function as the *ADH1* and *ADH2* genes in *S. cerevisiae* and these duplications represent a parallel evolutionary event. Regarding the group of *ADH6* and *ADH7*, which are in *S. cerevisiae* involved in the conversion of longer chain aldehydes and alcohols, one can again see several *D. bruxellensis* lineage-specific duplicates. *S. cerevisiae ADH6* and *ADH7* are involved in the synthesis of aromatic compounds (higher alcohols) and pre-cursors for aromatic esters. Our observation of the presence of several duplicated *ADH6-7*-like genes coincides with the previous observations that *D. bruxellensis* has a very intensive aromatic profile (Licker et al., 1999).

3.5. Conclusion: genome, evolution and food-related properties

The comparative analysis of the genome sequences of *S. cerevisiae* and *D. bruxellensis* revealed that the two lineages employed different or similar molecular mechanisms to evolve several similar traits. The WGD event and the lateral acquisition of genes needed for anaerobic growth such as *URA1* from other organisms were some of the events necessary for the establishment of a modern fermentative and anaerobic life style in the *Saccharomyces* lineage. *D. bruxellensis* has independently evolved into an organism able to grow under anaerobic conditions, producing large amounts of ethanol and tolerating high ethanol levels. Under oxygen limitation, the ethanol yield of *D. bruxellensis* is almost the same as in *S. cerevisiae* (Galafassi et al., 2011). However, under aerobic conditions *D. bruxellensis* produces less ethanol but has higher biomass than *S. cerevisiae* (Blomqvist et al., 2010) suggesting a less pronounced Crabtree effect. We show here that in contrast to *S. cerevisiae*, *D. bruxellensis* does not show traces of extensive gene duplications. On the other hand, both lineages used the same strategy with promoter rewiring in genes associated with the respiration (Ihmels et al., 2005; Rozpędowska et al., 2011), and likely with the *ADH* genes duplication, which promotes ultimate separation of the fermentation process from ethanol consumption.

D. bruxellensis also shows greater diversity among strains in chromosome number and ploidy than does *S. cerevisiae* (Hellborg and Piskur, 2009), suggesting that the increase in the gene dose/ploidy could be an important event in establishment of the yeasts in sugar-rich anaerobic food fermentation habitats. In the *S. cerevisiae* lineage this was achieved by WGD, for example duplication of the genes involved in glycolysis, but in the *D. bruxellensis* lineage it is apparently achieved through increased ploidy. The differences in production of some components that have a flavor impact by these two yeast species may also be due to the observed differences in the genome content, for example duplication of genes involved in generation of higher alcohols (Figure 5). The availability of the whole genome sequence now provides a tool to deduce the enzymatic background for production of off-flavor compounds. In conclusion, this work opens many opportunities to examine the genetic background for food-related properties as well as to understand the evolutionary processes behind evolution of the fermentative metabolism and the ability of these yeasts to establish themselves in anaerobic niches.

Acknowledgements

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

References

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

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

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

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

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

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

Du Toit, M., Pretorius, I. S., 2000. Microbial spoilage and preservation of wine: Using weapons from nature's own arsenal. South African Journal of Enology and Viticulture 21, 74-96.

Eddy, S. R., 2011. Accelerated Profile HMM Searches. PLoS Computational Biology 7, e1002195.

Edgar, R. C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32, 1792-1797.

Enright, A. J., Van Dongen, S., Ouzounis, C. A., 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 30, 1575-1584.

Felsenstein, J., 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Gabaldón, T., 2008. Large-scale assignment of orthology: back to phylogenetics? Genome Biology 9, 235.

Galafassi, S., Merico, A., Pizza, F., Hellborg, L., Molinari, F., Piskur, J., Compagno, C., 2011. *Dekkera/Brettanomyces* yeasts for ethanol production from renewable sources under oxygen-limited and low-pH conditions. Journal of Industrial Microbiology and Biotechnology 38, 1079-1088.

Gascuel, O., 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Molecular Biology and Evolution 14, 685-695.

Gojković, Z., 2004. Horizontal gene transfer promoted evolution of the ability to propagate under anaerobic conditions in yeasts. Molecular Genetics and Genomics 271, 387-393.

Guindon, S., Delsuc, F., Dufayard, J. F., Gascuel, O., 2009. Estimating maximum likelihood phylogenies with PhyML. Methods in Molecular Biology 537, 113-137.

Hellborg, L., Piskur, J., 2009. Complex nature of the genome in a wine spoilage yeast, *Dekkera bruxellensis*. Eukaryotic Cell 8, 1739-1749.

Heresztyn, T., 1986. Metabolism of volatile phenolic compounds from hydroxycinnamic acids by *Brettanomyces* strains. Archives of Microbiology 146, 96-98.

Huerta-Cepas, J., 2011. PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. Nucleic Acids Research 39, D556-560.

Huerta-Cepas, J., Dopazo, J., Gabaldón, T., 2010. ETE: a python Environment for Tree Exploration. BMC Bioinformatics 11, 24.

Huerta-Cepas, J., Gabaldón, T., 2011. Assigning duplication events to relative temporal scales in genome-wide studies. Bioinformatics 27, 38-45.

Ihmels, J., Bergmann, S., Gerami-Nejad, M., Yanai, I., McClellan, M., Berman, J., Barkai, N., 2005. Rewiring of the yeast transcriptional network through the evolution of motif usage. Science 309, 938-940.

Isono, K., McIninch, J. D., Borodovsky, M., 1994. Characteristic features of the nucleotide sequences of yeast mitochondrial ribosomal protein genes as analyzed by computer program GeneMark. DNA Research 1, 263-269.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenetics and Genome Research 110, 462-467.

Kanehisa, M., , 2008. KEGG for linking genomes to life and the environment. Nucleic Acids Research 36, D480-484.

Katoh, K., Toh, H., 2008. Recent developments in the MAFFT multiple sequence alignment program. Briefings in Bioinformatics 9, 286-298.

Kent, W. J., 2002. BLAT--the BLAST-like alignment tool. Genome Research 12, 656-664.

Koonin, E.V., 2004. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biology 5, R7.

Letunic, I., Bork, P., 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23, 127-128.

Licker, J. L., Acree, T. E., Henick-Kling, T., 1999. What is 'Brett' (*Brettanomyces*) flavour? A preliminary investigation. American Chemical Society Symposium Series 714, 96-115.

Melén, K., Krogh, A., von Heijne, G., 2003. Reliability measures for membrane protein topology prediction algorithms. Journal of Molecular Biology 327, 735-744.

Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G., 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Engineering 10, 1-6.

Parkhomchuk, D., 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Research 37, e123.

Piskur, J., Langkjaer, R. B., 2004. Yeast genome sequencing: the power of comparative genomics. Molecular Microbiology 53, 381-389.

Piskur, J., Rozpędowska, E., Polakova, S., Merico, A., Compagno, C., 2006. How did *Saccharomyces* evolve to become a good brewer? Trends in Genetics 22, 183-186.

Rozpędowska, E., Hellborg, L., Ishchuk, O. P., Orhan, F., Galafassi, S., Merico, A., Woolfit, M., Compagno, C., Piskur, J., 2011. Parallel evolution of the make-accumulate-consume strategy in *Saccharomyces* and *Dekkera* yeasts. Nature Communications 2, 302.

Salamov, A. A., Solovyev, V. V., 2000. Ab initio gene finding in Drosophila genomic DNA. Genome Research 10, 516-522.

Smit, A. F. A., Hubley, R., Green, P. *RepeatMasker Open-3.0*. 1996-2010.

Smith, T. F., Waterman, M. S., 1981. Identification of common molecular subsequences. Journal of Molecular Biology 147, 195-197.

Stamatakis, A., Ludwig, T., Meier, H., 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21, 456-463.

Subramanian, A. R., Kaufmann, M., Morgenstern, B., 2008. DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. Algorithms for Molecular Biology 3, 6.

The UniProt Consortium., 2011. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Research 39, D214-D219.

Thomson, J. M., Gaucher, E. A., Burgan, M. F., De Kee, D. W., Li, T., Aris, J. P., Benner, S. A., 2005. Resurrecting ancestral alcohol dehydrogenases from yeast. Nature Genetics 37, 630-635.

Vigentini, I., Romano, A., Compagno, C., Merico, A., Molinari, F., Tirelli, A., Foschino, R., Volonterio, G., 2008. Physiological and oenological traits of different *Dekkera/Brettanomyces bruxellensis* strains under wine-model conditions. FEMS Yeast Research 8, 1087-1096.

Wallace, I. M., O'Sullivan, O., Higgins, D. G., Notredame, C., 2006. M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Research 34, 1692-1699.

Wedral, D., Shewfelt, R., Frank, J., 2008. The challenge of *Brettanomyces* in wine. LWT-Food Science and Technology 43, 1474-1479.

Wehe, A., Bansal, M. S., Burleigh, J. G., Eulenstein, O., 2008. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics 24, 1540-1541.

Woolfit, M., Rozpędowska, E., Piskur, J., Wolfe, K. H., 2007. Genome survey sequencing of the wine spoilage yeast *Dekkera (Brettanomyces) bruxellensis*. Eukaryotic Cell 6, 721-733.

Zdobnov, E. M., Apweiler, R., 2001. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17, 847-848.

Figures

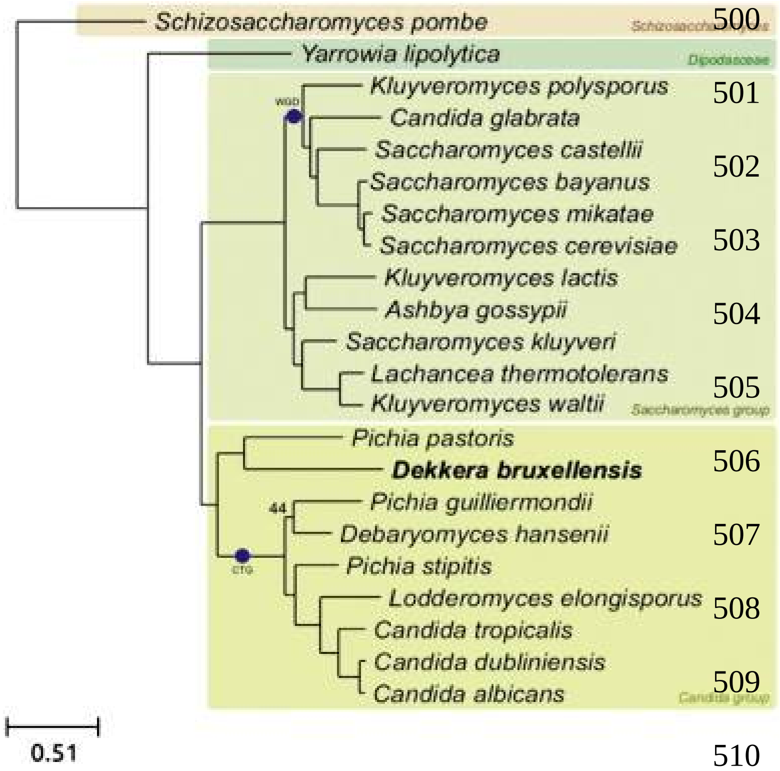


Figure 1. Phylogenetic relationships within the *Schizosaccharomyces Dipodasceael Candida Saccharomyces* group. Postion of *D. Bruxellensis* is in bold. Important evolutionaty events such as the WGD, or genetic code alteration in the *Candida* (CTG) clade are indicated. The tree is based in a Masximum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

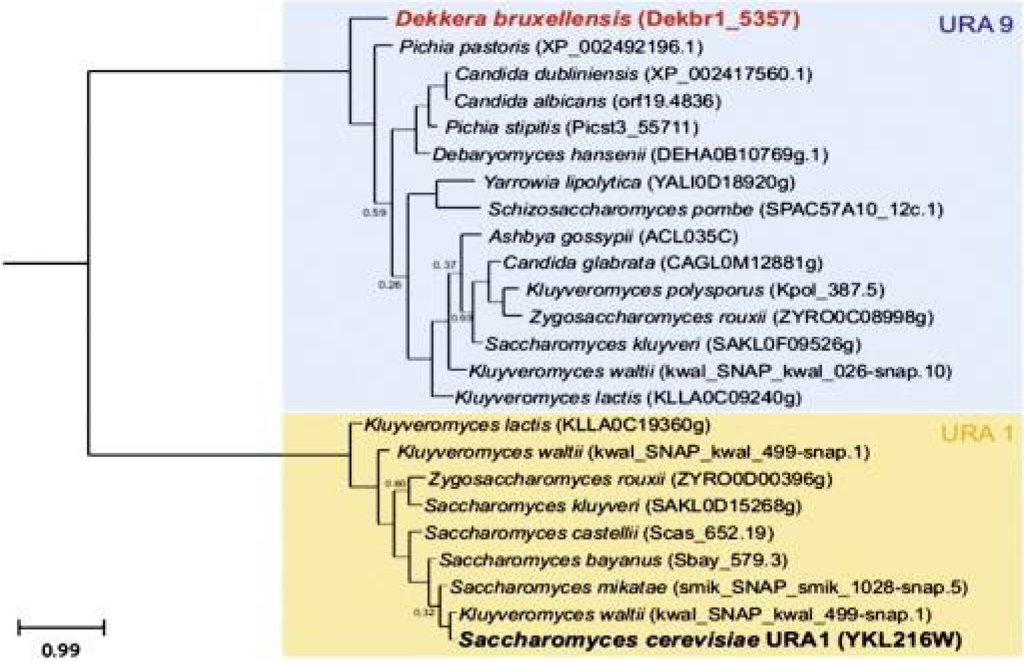

Dekkera bruxellensis (Dekbr1_5357)
URA 9
Pichia pastoris (XP_002492196.1)
Candida dubliniensis (XP_002417560.1)
Candida albicans (orf19.4836)
Pichia stipitis (Picst3_55711)
Debaryomyces hansenii (DEHA0B10769g.1)
Yarrowia lipolytica (YALI0D18920g)
0.59
Schizosaccharomyces pombe (SPAC57A10_12c.1)
Ashbya gossypii (ACL035C)
0.37
Candida glabrata (CAGL0M12881g)
0.26
Kluyveromyces polysporus (Kpol_387.5)
0.69
Zygosaccharomyces rouxii (ZYRO0C08998g)
Saccharomyces kluyveri (SAKL0F09526g)
Kluyveromyces waltii (kwal_SNAP_kwal_026-snap.10)
Kluyveromyces lactis (KLLA0C09240g)
Kluyveromyces lactis (KLLA0C19360g)
URA 1
Kluyveromyces waltii (kwal_SNAP_kwal_499-snap.1)
0.80
Zygosaccharomyces rouxii (ZYRO0D00396g)
Saccharomyces kluyveri (SAKL0D15268g)
Saccharomyces castellii (Scas_652.19)
Saccharomyces bayanus (Sbay_579.3)
Saccharomyces mikatae (smik_SNAP_smik_1028-snap.5)
0.32
Kluyveromyces waltii (kwal_SNAP_kwal_499-snap.1)
Saccharomyces cerevisiae URA1 (YKL216W)
0.99

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.

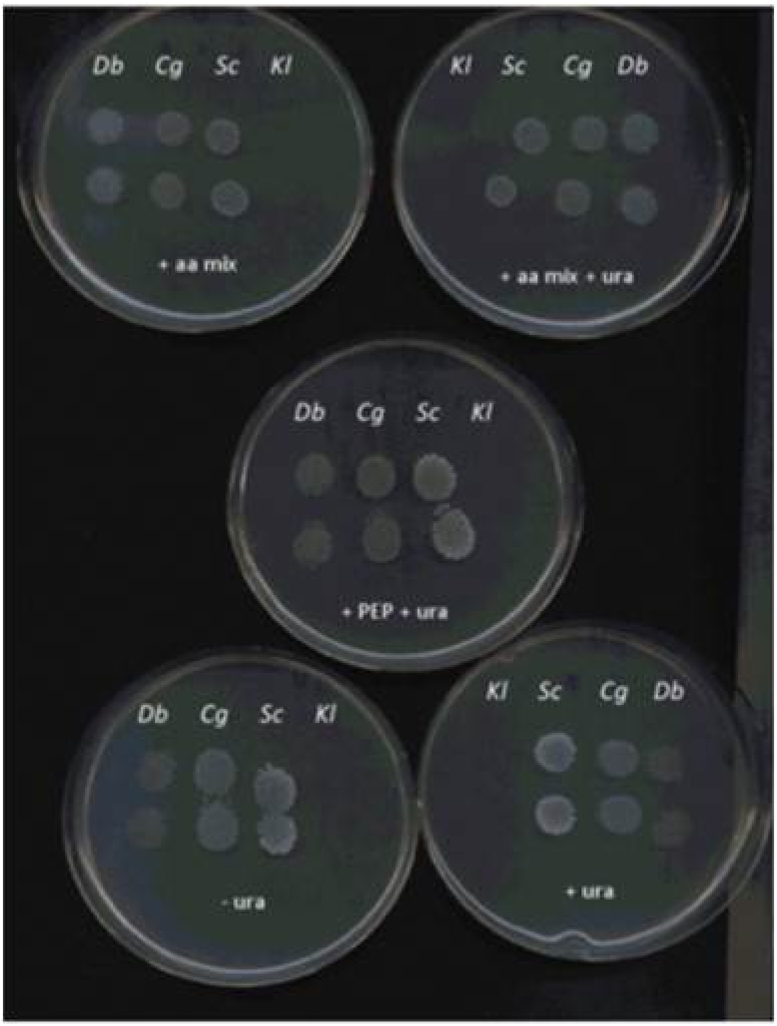


Figure 3. Anaerobic growth of *D. Bruxellensis* (CBS2499), *C. Glabrata* (CBS138), *S. Cerevisiae* (CEN.PK 113-7D) and *K. Lactis* (CBS2359) on a minimal medium (2% glucose) with different suplements, sucha as uracile, aminoacids mix and peptone.

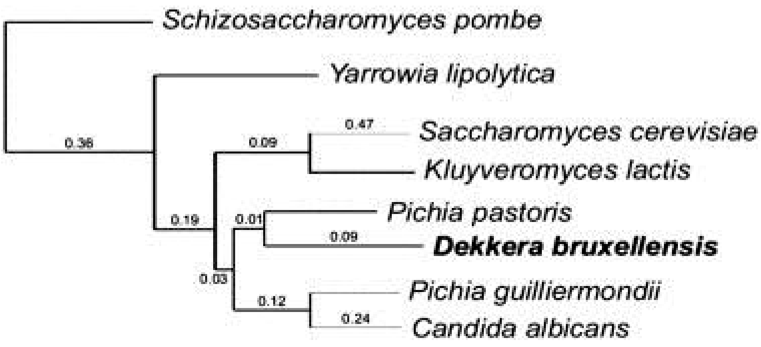


Figure 4. The relative number of duplication events per gene. The number on each branching point indicates the average duplication events per gene detected at each of the indicated lineages, as inferred from anlaysis of *S. Cerevisiae*, *C. Albicans* and *D. Bruxellensis* phylomes available at PhylomeDB.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

Bruxellensis lineage is among the samllest, despite the relatively long branch since its separation from *P. Pastoris.*

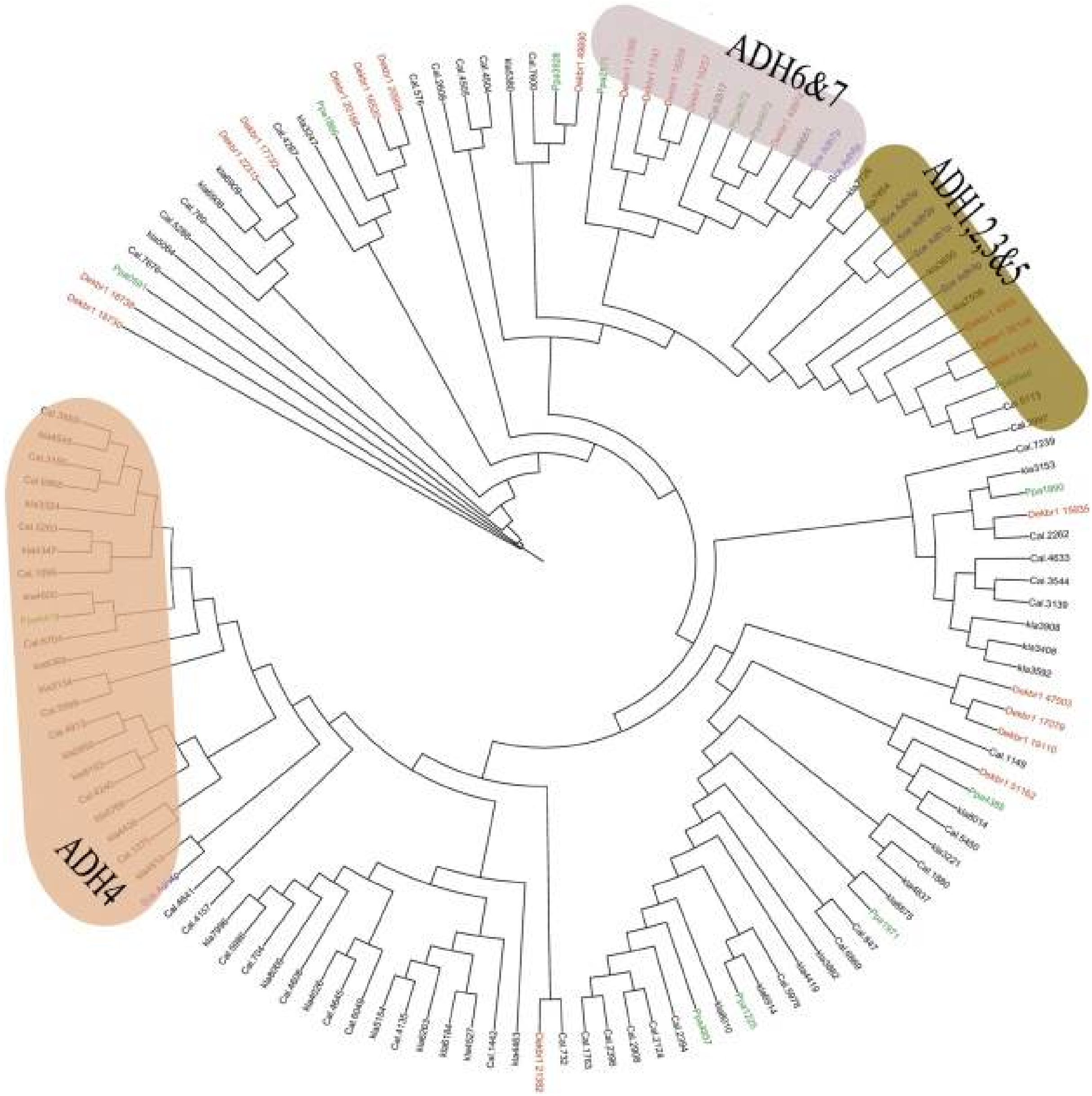
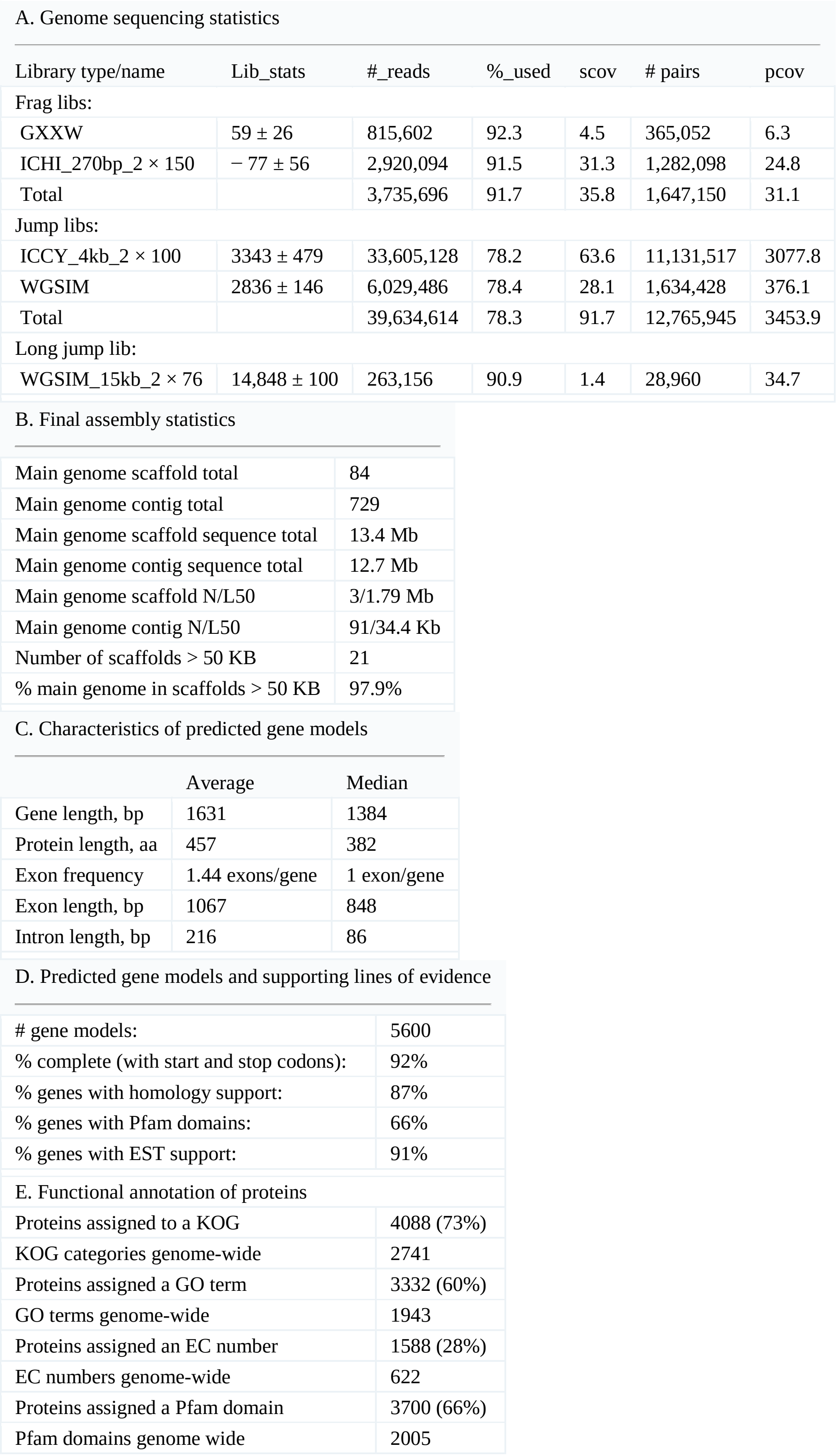


Figure 5. A phylogenetic tree of *ADH* homologs of *S. cerevisiae*, *D. Bruxellensis*, *C. Albicans* and *P. Pastoris*. The *ADM* genes (*ADH1-7*) of *S. Cerevisiae* are colored in blue while the the homologs of *D. Bruxellensis* are colored in red and *P. Pastoris* in green. The figure was produced with iTOL (Letunic and Bork, 2007).

Tables

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%

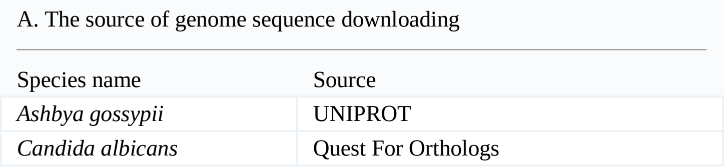
C. Characteristics of predicted gene models

	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

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/