Comparative genomics of biotechnologically important yeasts

Robert Riley[a], Sajeet Haridas[a], Kenneth H. Wolfe[b], Mariana R. Lopes[c,d], Chris Todd Hittinger[c,e], Markus Göker[f], Asaf A. Salamov[a], Jennifer H. Wisecaver[g], Tanya M. Long[h,i], Christopher H. Calvey[j], Andrea L. Aerts[a], Kerrie W. Barry[a], Cindy Choi[a], Alicia Clum[a], Aisling Y. Coughlan[b], Shweta Deshpande[a], Alexander P. Douglass[b], Sara J. Hanson[b], Hans-Peter Klenk[f,k], Kurt M. LaButti[a], Alla Lapidus[a,1], Erika A. Lindquist[a], Anna M. Lipzen[a], Jan P. Meier-Kolthoff[f], Robin A. Ohm[a,2], Robert P. Otillar[a], Jasmyn L. Pangilinan[a], Yi Peng[a], Antonis Rokas[g], Carlos A. Rosa[d], Carmen Scheuner[f], Andriy A. Sibirny[l,m], Jason C. Slot[n], J. Benjamin Stielow[f,o], Hui Sun[a], Cletus P. Kurtzman[p], Meredith Blackwell[q,r], Igor V. Grigoriev[a,3], and Thomas W. Jeffries[h,3]

[a]Department of Energy Joint Genome Institute, Walnut Creek, CA 94598; [b]University College Dublin Conway Institute, School of Medicine, University College Dublin, Dublin 4, Ireland; [c]Laboratory of Genetics, Genetics/Biotechnology Center, University of Wisconsin–Madison, Madison, WI 53706; [d]Departamento de Microbiologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, 31270-901, Brazil; [e]Department of Energy Great Lakes Bioenergy Research Center, University of Wisconsin–Madison, Madison, WI 53726; [f]Deutsche Sammlung von Mikroorganismen und Zellkulturen German Collection of Microorganisms and Cell Cultures, Leibniz Institute, 38124 Braunschweig, Germany; [g]Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235; [h]Department of Bacteriology, University of Wisconsin–Madison, Madison, WI 53706; [i]US Department of Agriculture Forest Products Laboratory, Madison, WI 53726; [j]Xylome Corporation, Madison, WI 53719; [k]School of Biology, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom; [l]Department of Molecular Genetics and Biotechnology, Institute of Cell Biology, National Academy of Sciences of Ukraine, Lviv 79005, Ukraine; [m]Department of Biotechnology and Microbiology, University of Rzeszow, Rzeszow 35-601, Poland; [n]Department of Plant Pathology, Ohio State University, Columbus, OH 43210; [o]Centraalbureau voor Schimmelcultures Fungal Biodiversity Centre, Royal Netherlands Academy of Arts and Sciences, 3508 AD, Utrecht, The Netherlands; [p]Agricultural Research Service, National Center for Agricultural Utilization Research, US Department of Agriculture, Peoria, IL 61604; [q]Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803; and [r]Department of Biological Sciences, University of South Carolina, Columbia, SC 29208

Edited by Chris R. Somerville, University of California, Berkeley, CA, and approved July 11, 2016 (received for review March 10, 2016)

Ascomycete yeasts are metabolically diverse, with great potential for biotechnology. Here, we report the comparative genome analysis of 29 taxonomically and biotechnologically important yeasts, including 16 newly sequenced. We identify a genetic code change, CUG-Ala, in *Pachysolen tannophilus* in the clade sister to the known CUG-Ser clade. Our well-resolved yeast phylogeny shows that some traits, such as methylotrophy, are restricted to single clades, whereas others, such as L-rhamnose utilization, have patchy phylogenetic distributions. Gene clusters, with variable organization and distribution, encode many pathways of interest. Genomics can predict some biochemical traits precisely, but the genomic basis of others, such as xylose utilization, remains unresolved. Our data also provide insight into early evolution of ascomycetes. We document the loss of H3K9me2/3 heterochromatin, the origin of ascomycete mating-type switching, and panascomycete synteny at the *MAT* locus. These data and analyses will facilitate the engineering of efficient biosynthetic and degradative pathways and gateways for genomic manipulation.

genomics | bioenergy | biotechnological yeasts | genetic code | microbiology

Yeasts are fungi that reproduce asexually by budding or fission and sexually without multicellular fruiting bodies (1, 2). Their unicellular, largely free-living lifestyle has evolved several times (3). Despite morphological similarities, yeasts constitute over 1,500 known species that inhabit many specialized environmental niches and associations, including virtually all varieties of fruits and flowers, plant surfaces and exudates, insects and other invertebrates, birds, mammals, and highly diverse soils (4). Biochemical and genomic studies of the model yeast *Saccharomyces cerevisiae*—essential for making bread, beer, and wine—have established much of our understanding of eukaryotic biology. However, in many ways, *S. cerevisiae* is an oddity among the yeasts, and many important biotechnological applications and highly divergent physiological capabilities of lesser-known yeast species have not been fully exploited (5). Various species can grow on methanol or *n*-alkanes as sole carbon and energy sources, overproduce vitamins and lipids, thrive under acidic conditions, and ferment unconventional carbon sources. Many features of yeasts make them ideal platforms for biotechnological processes. Their thick cell walls help them survive osmotic shock, and in contrast to bacteria, they are resistant to viruses. Their unicellular form is easy to cultivate, scale up, and harvest. The objective of this study was, therefore, to put yeasts with diverse biotechnological applications in a phylogenomic context and relate their physiologies to genomic

Significance

The highly diverse Ascomycete yeasts have enormous biotechnological potential. Collectively, these yeasts convert a broad range of substrates into useful compounds, such as ethanol, lipids, and vitamins, and can grow in extremes of temperature, salinity, and pH. We compared 29 yeast genomes with the goal of correlating genetics to useful traits. In one rare species, we discovered a genetic code that translates CUG codons to alanine rather than canonical leucine. Genome comparison enabled correlation of genes to useful metabolic properties and showed the synteny of the mating-type locus to be conserved over a billion years of evolution. Our study provides a roadmap for future biotechnological exploitations.

Author contributions: A.L., C.P.K., M.B., I.V.G., and T.W.J. designed research; T.M.L., C.H.C., C.C., A.Y.C., S.D., S.J.H., H.-P.K., Y.P., A.A. Sibirny, J.B.S., C.P.K., and T.W.J. performed research; C.H.C. contributed new reagents/analytic tools; R.R., S.H., K.H.W., M.R.L., C.T.H., M.G., A.A. Salamov, J.H.W., A.L.A., K.W.B., A.C., A.P.D., K.M.L., A.L., E.A.L., A.M.L., J.P.M.-K., R.A.O., R.P.O., J.L.P., A.R., C.A.R., C.S., J.C.S., H.S., C.P.K., I.V.G., and T.W.J. analyzed data; R.R., S.H., K.H.W., C.T.H., M.G., C.P.K., M.B., I.V.G., and T.W.J. wrote the paper; and K.W.B., C.P.K., M.B., I.V.G., and T.W.J. coordinated the project.

Conflict of interest statement: C.H.C. and T.W.J. are employees of Xylome Corporation, which is developing nonconventional yeasts for biotechnological applications.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. LWKO00000000, LYME00000000, LXTC00000000, LYBQ00000000, LYBR00000000, LWUO00000000, LSKT00000000, LTAD00000000, LXPE00000000, AECK00000000, LSGR00000000, LXPB00000000, LZCH00000000, AEHA00000000, AEUO00000000, and LWUN00000000).

[1]Present address: Center for Algorithmic Biotechnology, St. Petersburg State University, St. Petersburg 199004, Russia.

[2]Present address: Microbiology, Department of Biology, Utrecht University, 3508, Utrecht, The Netherlands.

[3]To whom correspondence may be addressed. Email: IVGrigoriev@lbl.gov or twjeffri@wisc.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1603941113/-/DCSupplemental.

features, so that their useful properties may be developed through genetic techniques. Backgrounds on 16 yeasts are given in *SI Appendix*.

Results

Organism Phylogeny. Genomes were sequenced and assembled as described in *Materials and Methods* (*SI Appendix*, Table S1). Using the predicted proteomes (*SI Appendix*, Fig. S1 and Table S2) of these plus an additional 13 ascomycete yeasts and 9 fungal outgroups, we generated three phylogenomic data matrices: "full" (7,297 genes with four or more sequences), "MARE" (1,559 genes from the full set filtered for informative quality), and "core" (418 genes present in all organisms). The MARE-filtered supermatrix tree (6, 7) is shown (Fig. 1). The full supermatrix and core gene trees differed regarding the positions of *Ascoidea rubescens*, *Debaryomyces hansenii*, and *Metschnikowia bicuspidata*, but conflicting branches were not supported. The maximum parsimony MARE-filtered supermatrix and core genes matrix trees were topologically identical to Fig. 1 (*SI Appendix*, Figs. S5 and S6). Overall, our data show a phylogenetic tree with three major Saccharomycotina clades (CUG-Ser, Methylotrophs, and Saccharomycetaceae) along with significant differences among the early diverging members, such as *Lipomyces starkeyi*, *Tortispora caseinolytica*, *Yarrowia lipolytica*, and *Nadsonia fulvescens*.

Alternative Genetic Codes: CUG Coding for Ser and Ala. Yeasts in the CUG-Ser clade, including *Candida albicans*, use an altered genetic code, in which CUG codons are translated as Ser rather than the canonical Leu (8, 9) because of alterations in the $tRNA_{CAG}$ that decodes CUG. To investigate the origins of this change, we inspected predicted $tRNA_{CAG}$s for the presence of three sequence features indicative of Ser translation (9): a G33 residue 5′ to the anticodon, which may lower rates of leucylation (10), an Ser identity element in the variable loop, and a G discriminator base. Most CUG-Ser clade species contained all three serylation features in their predicted $tRNA_{CAG}$s, indicating translation of CUG to Ser (*SI Appendix*, Fig. S7*A*). However, in the most basal taxa of this clade, not all of the features are present: *M. bicuspidata* lacks the Ser identity element, and *Babjeviella inositovora* lacks the discriminator base. This finding may reflect stepwise accumulation of $tRNA_{CAG}^{Ser}$ features in the evolution of alternative CUG translation. Species branching deeper in the tree do not show any of the three features.

To investigate CUG translation beyond CUG-Ser, we analyzed multiple alignments of 700 orthologous groups of proteins (*SI Appendix*). For each yeast, we identified its CUG-encoded positions in the alignments and tabulated the frequencies of the amino acids in other species to which CUG sites aligned, restricting the analysis to conserved regions of proteins. In the CUG-Ser clade, CUG codons most frequently aligned with Ser rather than Leu (Fig. 2).

In the eight genomes from this clade, CUG codons aligned with Ser in 26–42% of aligned positions. For most of the other yeasts, CUG aligned predominantly with Leu (58–75%). However, two yeasts outside the CUG-Ser clade, *Pachysolen tannophilus* and *A. rubescens*, show unusual CUG alignment patterns. They were previously proposed to translate CUG as Ser based on interspecies alignments (11) (*SI Appendix*, Table S3), but their $tRNA_{CAG}$ genes lack the serylation features (*SI Appendix*, Fig. S7*A*). In *P. tannophilus*, CUG unexpectedly aligned mostly with Ala (25%) (Fig. 2), more than with Leu (7%) or Ser (9%). In *A. rubescens*, CUG codons are remarkably rare and showed no strong preference for any amino acid, suggesting the possibility that its genome could be edited to use CUG to code for unconventional amino acids.

To determine the genetic code of *P. tannophilus* directly, we sequenced tryptic peptides de novo by liquid chromatography

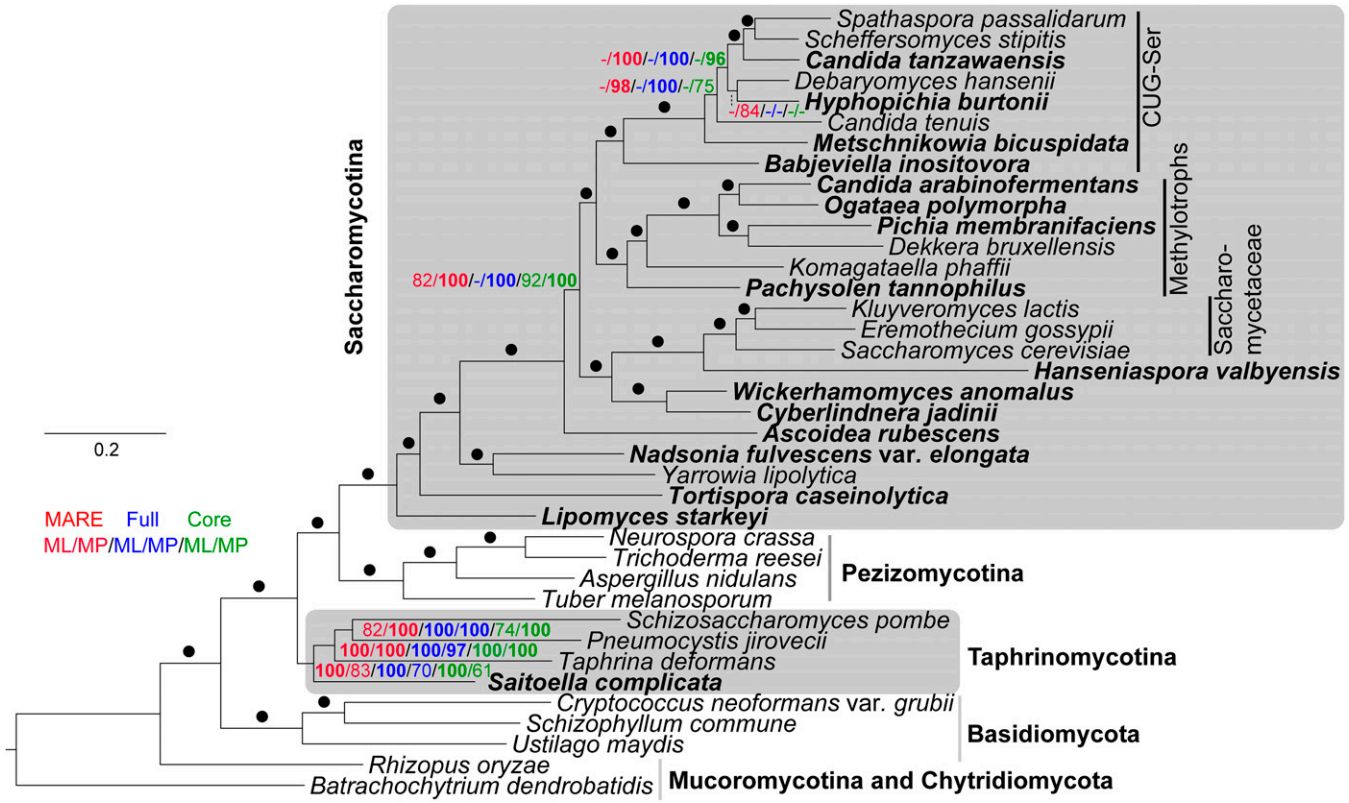

Fig. 1. Phylogenetic tree inferred from the MARE-filtered supermatrix (364,126 aligned amino acid residues) using maximum likelihood (ML) and rooted with *Batrachochytrium*. Organisms sequenced in this study are shown in bold. Numbers on the branches indicate ML and maximum parsimony (MP) bootstrap support values for the MARE-filtered (red), full (blue), and core genes (green) supermatrices. Values less than 60% are shown as dashes; dots indicate branches with maximum support under all settings.

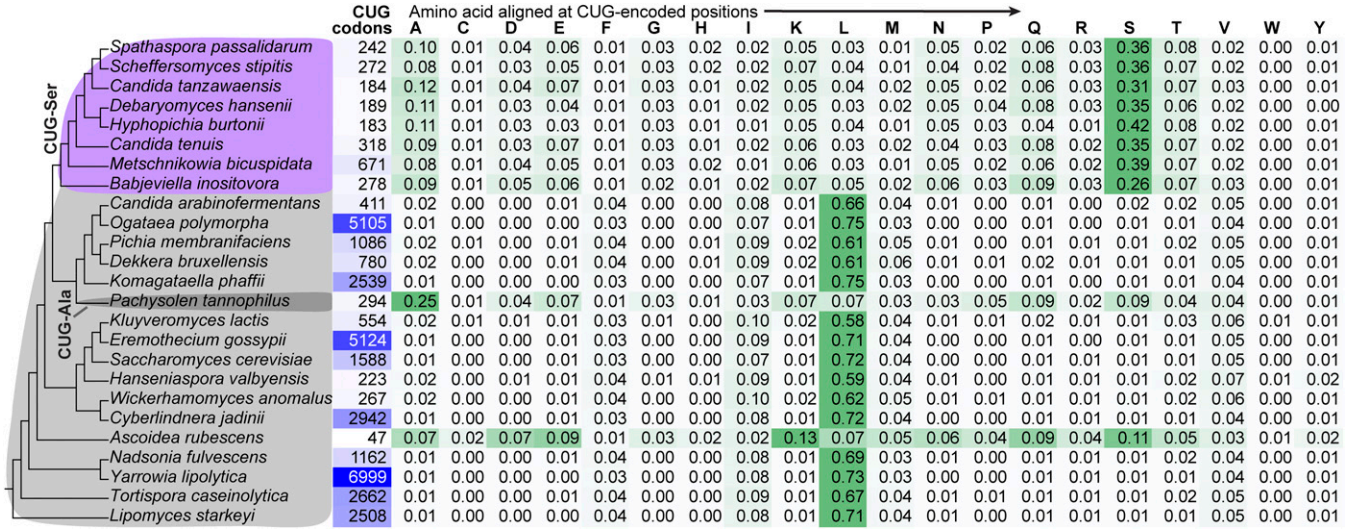

Fig. 2. CUG-Ala genetic code inferred from amino acids aligned to yeast CUG codons based on 700 orthologous groups of orthologous proteins. Blue-shaded boxes indicate the number of CUG-encoded amino acid positions in the ortholog set. Green-shaded boxes indicate the fraction of each organism's CUG-encoded positions aligned to each amino acid.

(LC)-MS/MS and compared them with the genome sequence. Among 6,836 peptides that mapped to unique sites in the genome, 178 span a CUG codon site (in 170 different genes). Of these CUG codon sites, 160 (90%) align with Ala in the sequenced peptide, 16 align with Leu in the sequenced peptide, and 2 align with other amino acids in the sequenced peptide (Dataset S1). The possibility of mRNA editing can be excluded, because no editing was seen at these sites in expressed sequence tag (EST) data from *P. tannophilus* for all 166 sites that were covered by EST reads. To confirm these findings, we transformed *P. tannophilus* with a selectable marker for hygromycin resistance, in which all CUG-Leu codons were changed to other Leu codons. Several thousand resistant transformants were obtained with the altered selectable marker, but none were obtained with the vector containing the native *hyg*^{r} gene (*SI Appendix*, Fig. S7 *B* and *C*). We conclude that *P. tannophilus* translates most CUG codons in mRNA as Ala.

Correlation of Genomically Encoded Enzymes to Metabolic Traits. We correlated several metabolic capabilities (2) with genome content (Fig. 3 and *SI Appendix*, Figs. S8–S10). Genes for D-xylose assimilation appear to be widespread in the ascomycete yeasts, including unexpectedly, several yeasts lacking the ability to grow on D-xylose in classical assays (2). Candidate genes (generally in multiple paralogs) for xylose reductase (*XYL1*), xylitol dehydrogenase (*XYL2*), and D-xylulokinase (*XYL3*) exist in nearly all

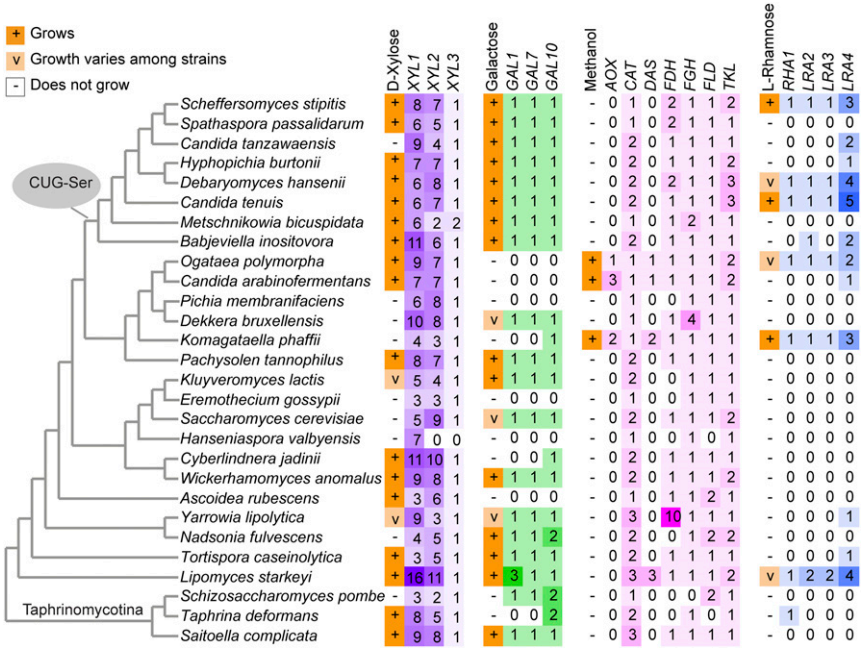

Fig. 3. Distribution of metabolic traits and their genes. For D-xylose, galactose, methanol, and L-rhamnose, numbered boxes indicate the numbers of predicted pathway genes. Genes for D-xylose, galactose, methanol, and L-rhamnose metabolism were identified by homology to characterized sequences (*Materials and Methods*).

of the ascomycete yeasts (the sole exception, *Hanseniaspora valbyensis*, lacks *XYL2* and *XYL3*). Unexpectedly, we found that *Candida tanzawaensis*, a member of the CUG-Ser clade in which xylose metabolism is well-characterized (12, 13), did not grow on xylose, despite possessing homologous genes for the full pathway. Similarly, despite the near ubiquity of the *XYL1*/*XLY2*/*XYL3* pathway in the yeasts, several species beyond the CUG-Ser clade either lack or have strain-specific ability to metabolize D-xylose (2, 14, 15). Variation in this trait may, therefore, involve factors other than gene presence or absence (e.g., regulation of gene expression or enzyme specificity) (16).

Galactose utilization. In contrast to D-xylose, galactose utilization correlates well with characterized pathways (17). The ability to use galactose varies widely across yeasts (2), and previous research has shown a handful of gene losses in the galactose utilization pathway (*GAL*) and at least one reacquisition by horizontal gene transfer (17, 18). The dense sampling of yeast genomes in this study provides an enriched picture of dynamic pathway evolution. Parsimony suggests that key *GAL* genes and the ability to use galactose have been lost at least seven times among the examined taxa for a total of at least 11 losses among the subphyla Saccharomycotina and Taphrinomycotina (Fig. 3). Because *GAL* homologs are conserved across all domains of life (19), our analyses suggest that the dominant mode of evolution for yeast galactose consumption is one of repeated and independent loss of the trait along with its required genes from an ancestral yeast that could consume galactose.

Methylotrophy. Relatively few yeasts can use methanol as a sole carbon source (20). All three methylotrophic yeasts in this study—*Ogataea polymorpha*, *Candida arabinofermentans*, and *Komagataella phaffii*—possess a full complement of genes for methanol utilization pathway enzymes (Fig. 3) and belong to the same multigenus clade (Fig. 1). In particular, this rare trait correlates perfectly with the presence of genes encoding alcohol oxidases (*AOX*s) and dihydroxyacetone synthases (*DAS*s) but not with other members of the pathway. Methylotrophy seems to have been lost because of loss of *AOX* and *DAS* in *Pichia membranifaciens* and *Dekkera bruxellensis*, which likely had a methylotrophic ancestor.

L-Rhamnose utilization. L-Rhamnose metabolism can proceed by phosphorylated (isomerase) and nonphosphorylated (oxidative) pathways, with the latter occurring in fungi (21). Its catabolism requires several enzymatic steps, all of which are necessary to enter central metabolism. In *Scheffersomyces stipitis*, the four genes of the oxidative L-rhamnose pathway are situated side by side in a cluster, which is repeated in various conformations among six of the yeasts in this study (*Candida tenuis*, *D. hansenii*, *O. polymorpha*, *L. starkeyi*, and *K. phaffii* in addition to *S. stipitis*), and the occurrence of the conserved cluster across broad phylogenetic distances correlates with a yeast's capacity to oxidize L-rhamnose. However, L-rhamnose assimilation is sparsely distributed throughout Saccharomycotina (Fig. 3), with losses of the pathway apparent in several clades.

Complex I (NADH Dehydrogenase)/Dihydroorotate Dehydrogenase (URA9/URA1). Our broad phylogenetic study elaborated prior findings (22) that loss of respiration complex I (RC1) preceded the gain of bacterial dihydroorotate dehydrogenase (*URA1*) in *S. cerevisiae* and closely related genera. These two features along with other evolutionary changes, such as expansion of genes for facilitated sugar uptake, contributed to the capacity for anaerobic growth by these yeasts. Two other fermentative species (*Nadsonia* and *Schizosaccharomyces*) showed diminished RC1 components but did not acquire *URA1* (*SI Appendix*, Fig. S9).

Metabolic Gene Clusters. Many genes for degradative and biosynthetic traits were proximal on chromosomes across wide phylogenetic ranges (*SI Appendix* and Datasets S2 and S3). Our analysis identified previously recognized gene clusters for urea, allantoin, galactose and *N*-acetyl glucosamine catabolism, starch, cellulose, and nitrate utilization and extended these associations across multiple genomes. Genes for lipid synthesis and amino acid metabolism (Gly, Ser, Phe, Tyr, and Trp) likewise were found in clusters. Three successive enzymatic steps in the biotin synthesis pathway were found in pairwise clusters that differed with phylogeny. Genes for the first two of these steps are frequently clustered in Saccharomycotina (23), whereas clusters with genes for the second and third steps were prevalent in the other clades.

Mating-Type Locus Organization, Mating-Type Switching, and H3K9me Heterochromatin. Mating-type locus (*MAT*) structures of the sequenced species (*SI Appendix*, Fig. S11) suggest a genetic mechanism to explain biological data about homothallism or heterothallism of the species. Comparison of *MAT* loci revealed evidence of conservation of synteny at this locus among all three subphyla of Ascomycota, showing that cell type has been controlled during 1 billion years (24) of ascomycete evolution by a single orthologous locus (*MAT*), despite gross changes of its gene content. The stability of *MAT* contrasts with the frequent turnover of sex determination loci in animals and plants. We also found that *P. tannophilus* and *A. rubescens* have mating-type switching systems that operate by inversion of a region of chromosome, similar to *O. polymorpha* and *K. phaffii* (25, 26) (*SI Appendix*, Figs. S11 and S12). Most Saccharomycotina species have lost the ancestral form of eukaryotic heterochromatin, in which lysine 9 of histone H3 is methylated (27); *L. starkeyi* is the only Saccharomycotina species with orthologs of *Schizosaccharomyces pombe* Clr4 (H3K9 methyltransferase), Epe1 (H3K9 demethylase), and Swi6 (H3K9me2/3 binding chromodomain protein). We infer that mating-type switching, which requires a mechanism to silence the nonexpressed copies of *MAT* genes, evolved in the Saccharomycotina lineage relatively soon after the ancestral H3K9me2/3 form of heterochromatin was lost (*SI Appendix*, Fig. S11).

Discussion

By filling in key taxonomic gaps in yeast phylogeny, we have been able to identify major evolutionary transitions in yeast metabolism. For example, pentose and cellobiose fermenting yeasts are often associated with wood-ingesting beetles and mainly found in the CUG-Ser clade. *P. tannophilus*, which also ferments xylose to ethanol, is in a newly recognized sister CUG-Ala clade. The ethanologenic yeasts, which include *S. cerevisiae*, *Kluyveromyces lactis*, and several other genera, gained the capacity for anaerobic growth by acquiring genes that enable anaerobic uracil synthesis, while losing NADH dehydrogenase. Loss of RC1 occurred more than once (*SI Appendix*, Fig. S9), and diminished respiration correlates with increased fermentative activity. The methylotrophic yeasts occur in a single clade that has retained high oxidative pentose phosphate pathway activity, while gaining tolerance to salt and low pH. The most lipogenic yeasts are closest to the base of divergence from filamentous ascomycetes.

The tree topology presented here (Fig. 1) generally agrees with previous work (1, 3, 28), but almost all branches receive stronger support. The clade including *S. cerevisiae* and *K. lactis* remains strongly supported as the sister of the clade comprising *Wickerhamomyces* and *Cyberlindnera*. Furthermore, there is now strong support for placement of *Dekkera*/*Brettanomyces* in the *Pichia* clade and a sister relationship between methylotrophs and the CUG-Ser clade. Inclusion of *Saitoella complicata* in the analysis greatly improved support for a monophyletic grouping of some members of the Taphrinomycotina. Our analysis provides strong support for a number of previous phylogenetic conclusions (1, 3, 28), but more data are needed to resolve the remaining poorly sampled clades. In addition to providing an understanding of biochemical pathway evolution, a strongly supported tree with more inclusive sampling will give a more stable higher-level taxonomy of the yeasts.

Common morphological features seldom resolve phylogenetic classification of yeasts. For example, ascospore shape and presence or absence of pseudohyphae and true hyphae are seldom exclusive to one clade. Although *Eremothecium* species have unique elongated ascospores with a tail-like extension, other morphologies, such as hat-shaped ascospores, are found in many clades, including members of Pezizomycotina. Bipolar budding is known for four genera, *Hanseniaspora*, *Saccharomycodes*, *Nadsonia*, and *Wickerhamia*, but *Hanseniaspora* and *Nadsonia* are only distantly related (Fig. 1). Similarly, many metabolic traits are shared too broadly to

have phylogenetic value, such as the fermentation of glucose. However, metabolism of some compounds, such as methanol, is monophyletic. Species that assimilate methanol belong to the multigenus "methylotrophs" clade (Fig. 1) that includes *Ogataea*, *Kuraishia*, *Komagataella*, *Pichia*, and *Dekkera*, but the latter two genera do not metabolize methanol, which suggests loss of this physiological trait in these two lineages.

Certain enzymes are found in some clades to a much greater extent than in others. For example, genes for cellulose utilization are found largely in the CUG-Ser clade, where β-glucosidases, endoglucanases, and transporters occur in functional clusters. *S. stipitis*, a member of the CUG-Ser clade, has six β-glucosidases, three endoglucanases, and multiple cellobiose transporters along with some xylanase activities. Most of the β- and endoglucosidases are induced in response to growth on cellobiose, where they confer the ability for rapid assimilation and fermentation of cellobiose. Enzymes for L-rhamnose metabolism are also found in functional clusters in the CUG-Ser yeasts (21). Based on clusters of the enzymes for L-rhamnose metabolism, it is possible to predict the capacity for L-rhamnose metabolism in other yeasts.

By sampling early branching lineages of Saccharomycotina, our study elucidates many aspects of genome evolution in this subphylum. Synteny of the *MAT* locus has been conserved across Ascomycota, and the capacity for sexual recombination has surely contributed to yeasts' metabolic and physiological versatility. Contrastingly, rDNA organization is highly dynamic (*SI Appendix*, Fig. S4). Gene clusters seem to be common across yeast taxa, with the most conspicuous examples attributable to the utilization of substrates or biosynthetic pathways that require several metabolic steps. Reassignment of the CUG codon was more complex than previously thought, affecting a broader phylogenetic range of species, and involved at least two distinct events.

The genetic code change in *P. tannophilus* has practical implications for the use of this species in biotechnology, because heterologous genes containing CUG codons may not produce functional proteins. Indeed, in addition to the results with the CUG-less and CUG-containing hygromycin resistance markers described above, we and others (29) have been unable to transform *P. tannophilus* with the kanamycin resistance marker gene (Kan^R), which contains four CUG codons. In contrast, the *S. cerevisiae HXK2* gene, which has no CUG codons, was used successfully to complement a *P. tannophilus* mutant (30). Our discovery [in agreement with a report published (31) while the present study was in review] that CUG codes for Ala in *P. tannophilus* nuclear genes is only the second known example of a naturally occurring sense codon reassignment in any organism, the other being the CUG to Ser reassignment (9). All other genetic code changes involve the capture or reassignment of stop codons. The CUG codon in *A. rubescens* also has biotechnological potential, because the very low frequency of its use suggests the possibility of assigning nonstandard amino acids to this codon through genome editing (32).

The use of "nonconventional" yeast species in biotechnology is still in its infancy, and the industry uses only a tiny fraction of the thousands of species that are potentially exploitable (5). The species in widespread use have generally been chosen based on classical assays of their enzymatic or physiological properties in laboratory conditions, without regard to the possible full potential of their genomes. *S. pombe* provides an informative illustration. Traditional tests indicate that *S. pombe* cannot grow on galactose (2), but its genome contains a set of *GAL* pathway genes predicted to be functional. Indeed, mutants of *S. pombe* have been isolated that grow on galactose and constitutively express their *GAL* pathways (33), suggesting that *S. pombe* may respond to a different induction signal. Similarly, *S. cerevisiae*, although not classified as a xylose-using yeast (2), has been shown to grow on xylose when endogenous genes are overexpressed (34). The ability of *Y. lipolytica* to grow on D-xylose varies among strains (2, 14, 15), and it may be that the D-xylose pathways of *C. tanzawensis* and other yeasts, which possess *XYL* gene homologs but do not grow on D-xylose in classical assays (2), are likewise rewired. If this type of latent metabolic capability is commonplace, traditional biochemical assays may have overlooked a significant portion of genomic potential. As sequencing costs decrease, mining the genomes of the thousands of currently unsequenced yeast species offers an efficient route toward discovering the next generation of workhorse yeasts for biotechnology.

Materials and Methods

Genome Sequencing and Assembly. Genomes were sequenced using the Illumina, 454 (Roche), and PacificBiosciences platforms and assembled with AllPathsLG (35), Velvet (36), gapResolution (37), and PBJelly (38) as platform-appropriate (*SI Appendix*).

Genome Annotation. Genomes were annotated using the JGI Annotation pipeline and made available through JGI fungal genome portal MycoCosm (39) (jgi.doe.gov/fungi). The data are also deposited at GenBank under the following accession numbers (LWKO00000000, LYME00000000, LXTC00000000, LYBQ00000000, LYBR00000000, LWUO00000000, LSKT00000000, LTAD00000000, LXPE00000000, AECK00000000, LSGR00000000, LXPB00000000, LZCH00000000, AEHA00000000, AEUO00000000, and LWUN00000000). Annotation statistics are summarized in *SI Appendix*, Table S2.

Organism Phylogeny. Thirty-eight genome sequences were phylogenetically analyzed using the DSMZ phylogenomics pipeline as previously described (40–42) (*SI Appendix*).

Alternative Genetic Code. tRNAs were predicted with tRNAscan-SE (43). Those predicted tRNAs with CAG anticodons (complement of CUG) were chosen and designated as $tRNA_{CAG}$. Secondary structure-based multiple alignment of predicted $tRNA_{CAG}$ sequences was then performed with R-Coffee (44). Resulting alignments were manually inspected to identify the sequence features previously described (9) as important for translation of CUG to Ser (*SI Appendix*, Fig. S7). To further investigate the role of CUG codons beyond the CUG clade, we performed best bidirectional hits analysis using BLAST to identify orthologous groups and selected 700 such groups, such that each group had one protein from each of the 25 Saccharomycotina yeasts plus 3 filamentous fungi outgroups: *Neurospora crassa*, *Aspergillus nidulans*, and *Schizophyllum commune*. All protein sequences were downloaded from the JGI MycoCosm site (jgi.doe.gov/fungi). Mapping the proteins back to the genome sequences from which they were predicted, we replaced all predicted amino acids coded by a CUG to an X in all yeasts. From 700 orthologous groups, we then performed multiple alignment using MAFFT and extracted conserved regions using Gblocks. We then looked for resulting alignment positions that had at least one X, giving 28,640 such positions, and counted statistics on what non-Xs (i.e., coded by a codon other than CUG) were present in these positions. We reasoned that, particularly in the case of highly conserved positions, the amino acids present ought to suggest what amino acid was coded by a yeast's CUG codon.

Genetic Code of *P. tannophilus*. A culture of *P. tannophilus* grown in yeast extract peptone dextrose was analyzed by LC-MS/MS. Triplicate samples were run on a Thermo Scientific Q Exactive Mass Spectrometer connected to a Dionex Ultimate 3000 (RSLCnano) Chromatography System. Tryptic peptides were resuspended in 0.1% formic acid. Each sample was loaded onto a fused silica emitter (75 μm i.d.; pulled using a laser puller; Sutter Instruments P2000), packed with 1.8 μm 120 Å UChrom C18 packing material (NanoLCMS Solutions), and separated by an increasing acetonitrile gradient over 60 min at a flow rate of 250 nL/min. The mass spectrometer was operated in positive ion mode, with a capillary temperature of 320 °C and with potential of 2,300 V applied to the frit. All data were acquired with the mass spectrometer operating in automatic data-dependent switching mode. A high-resolution (70,000) MS scan (300–1600 *m/z*) was performed using the Q Exactive to select the 12 most intense ions before MS/MS analysis using HCD.

De novo peptide sequences were determined from the LC-MS/MS data using PEAKS Studio 7 (Bioinformatics Solutions) set to a false discovery rate of 1%. The 34,116 nonredundant peptide sequences obtained by combining three replicate samples were compared with a protein database of all annotated *P. tannophilus* genes translated using the universal genetic code. Only 6,836 peptides that mapped to unique sites in this database with zero or one amino acid mismatch per peptide were retained for analysis. They mapped to 1,982 genes. All 71,455 de novo sequenced residues in these peptides were then compared with the corresponding codons in their genes (Dataset S1).

Assigning Genes to Enzyme Functions. Predicted protein sequences were assigned enzyme function using a combination of TBLASTN searches with query protein sequences from the characterized pathways in model organisms vs.

genome assemblies, BLASTP searches of the same query proteins against the full protein catalog of each organism, and PRIAM (45) profile searches (*SI Appendix*).

Gene Cluster Analysis. Gene families were defined by clustering all protein sequences of the full study set with MCL (46). Fungal genomes were screened for metabolic gene clusters as previously described (47). Briefly, two EC genes were considered clustered if they were separated by no more than six intervening genes according to published annotation and their EC numbers were nearest neighbors in one or more Kyoto Encyclopedia of Genes and Genomes pathways. Gene clusters were inferred by joining overlapping metabolic gene pair ranges that were separated by no more than six intervening genes.

ACKNOWLEDGMENTS. We thank Marco A. Soares for computational advice. K.H.W. thanks G. Cagney, E. Dillon, and K. Wynne (University College Dublin Conway Institute Proteomics Core Facility) for help with MS. M.B. thanks Drs. S. O. Suh, H. Urbina, and N. H. Nguyen and numerous Louisiana State University undergraduates for their assistance. The work conducted by the US Department of Energy (DOE) Joint Genome Institute, a DOE Office of Science User Facility, is supported by Office of Science of the US DOE Contract DE-AC02-05CH11231. This material is based on work supported by National Science Foundation Grant DEB-1442148 (to C.T.H. and C.P.K.) and supported in part by DOE Great Lakes Bioenergy Research Center, DOE Office of Science Grant BER DE-FC02-07ER64494, and US Department of Agriculture (USDA) National Institute of Food and Agriculture Hatch Project 1003258. K.H.W. acknowledges European Research Council Grant 268893, Science Foundation Ireland Grant 13/IA/1910, and the Wellcome Trust. M.R.L. acknowledges a fellowship from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (process no. 7371/13-6). C.T.H. is a Pew Scholar in the Biomedical Sciences and an Alfred Toepfer Faculty Fellow, which are supported by the Pew Charitable Trusts and the Alexander von Humboldt Foundation, respectively. C.A.R. acknowledges support from the Conselho Nacional de Desenvolvimento Científico e Tecnológico-CNPq. Funding from National Science Foundation Grants DEB-0072741 (to M.B.) and 0417180 (to M.B.) supported discovery and study of many new yeast strains that contributed to this study. T.W.J. acknowledges DOE Great Lakes Bioenergy Research Center DOE Office of Science Grant BER DE-FC02-07ER64494 and the USDA, Forest Products Laboratory for financial support.

1. Dujon B (2010) Yeast evolutionary genomics. *Nat Rev Genet* 11(7):512–524.
2. Kurtzman CP, Fell JW, Boekhout T, eds (2011) *The Yeasts, a Taxonomic Study* (Elsevier, Amsterdam).
3. Nagy LG, et al. (2014) Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. *Nat Commun* 5:4471.
4. Sylvester K, et al. (2015) Temperature and host preferences drive the diversification of *Saccharomyces* and other yeasts: A survey and the discovery of eight new yeast species. *FEMS Yeast Res* 15(3):fov002.
5. Hittinger CT, et al. (2015) Genomics and the making of yeast biodiversity. *Curr Opin Genet Dev* 35:100–109.
6. de Queiroz A, Gatesy J (2007) The supermatrix approach to systematics. *Trends Ecol Evol* 22(1):34–41.
7. DeGiorgio M, Degnan JH (2010) Fast and consistent estimation of species trees using supermatrix rooted triples. *Mol Biol Evol* 27(3):552–569.
8. Kawaguchi Y, Honda H, Taniguchi-Morimura J, Iwasaki S (1989) The codon CUG is read as serine in an asporogenic yeast *Candida cylindracea*. *Nature* 341(6238):164–166.
9. Santos MA, Gomes AC, Santos MC, Carreto LC, Moura GR (2011) The genetic code of the fungal CTG clade. *C R Biol* 334(8-9):607–611.
10. Santos MA, Ueda T, Watanabe K, Tuite MF (1997) The non-standard genetic code of *Candida* spp.: An evolving genetic code or a novel mechanism for adaptation? *Mol Microbiol* 26(3):423–431.
11. Mühlhausen S, Kollmar M (2014) Molecular phylogeny of sequenced *Saccharomycetes* reveals polyphyly of the alternative yeast codon usage. *Genome Biol Evol* 6(12): 3222–3237.
12. Jeffries TW, et al. (2007) Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. *Nat Biotechnol* 25(3):319–326.
13. Wohlbach DJ, et al. (2011) Comparative genomics of xylose-fermenting fungi for enhanced biofuel production. *Proc Natl Acad Sci USA* 108(32):13212–13217.
14. Ryu S, Hipp J, Trinh CT (2015) Activating and elucidating metabolism of complex sugars in *Yarrowia lipolytica*. *Appl Environ Microbiol* 82(4):1334–1345.
15. Tsigie YA, Wang CY, Truong CT, Ju YH (2011) Lipid production from *Yarrowia lipolytica* Po1g grown in sugarcane bagasse hydrolysate. *Bioresour Technol* 102(19): 9216–9222.
16. Urbina H, Blackwell M (2012) Multilocus phylogenetic study of the *Scheffersomyces* yeast clade and characterization of the N-terminal region of xylose reductase gene. *PLoS One* 7(6):e39128.
17. Hittinger CT, Rokas A, Carroll SB (2004) Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. *Proc Natl Acad Sci USA* 101(39): 14144–14149.
18. Wolfe KH, et al. (2015) Clade- and species-specific features of genome evolution in the Saccharomycetaceae. *FEMS Yeast Res* 15(5):fov035.
19. Johnston M (1987) A model fungal gene regulatory mechanism: The GAL genes of *Saccharomyces cerevisiae*. *Microbiol Rev* 51(4):458–476.
20. Wegner GH (1990) Emerging applications of the methylotrophic yeasts. *FEMS Microbiol Rev* 7(3-4):279–283.
21. Koivistoinen OM, et al. (2012) Characterisation of the gene cluster for l-rhamnose catabolism in the yeast *Scheffersomyces* (*Pichia*) *stipitis*. *Gene* 492(1):177–185.
22. Gojković Z, et al. (2004) Horizontal gene transfer promoted evolution of the ability to propagate under anaerobic conditions in yeasts. *Mol Genet Genomics* 271(4): 387–393.
23. Hall C, Dietrich FS (2007) The reacquisition of biotin prototrophy in *Saccharomyces cerevisiae* involved horizontal gene transfer, gene duplication and gene clustering. *Genetics* 177(4):2293–2307.
24. Hedges SB, Blair JE, Venturi ML, Shoe JL (2004) A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol Biol* 4:2.
25. Hanson SJ, Byrne KP, Wolfe KH (2014) Mating-type switching by chromosomal inversion in methylotrophic yeasts suggests an origin for the three-locus *Saccharomyces cerevisiae* system. *Proc Natl Acad Sci USA* 111(45):E4851–E4858.
26. Maekawa H, Kaneko Y (2014) Inversion of the chromosomal region between two mating type loci switches the mating type in *Hansenula polymorpha*. *PLoS Genet* 10(11):e1004796.
27. Hickman MA, Froyd CA, Rusche LN (2011) Reinventing heterochromatin in budding yeasts: Sir2 and the origin recognition complex take center stage. *Eukaryot Cell* 10(9): 1183–1192.
28. Kurtzman CP, Robnett CJ (2013) Relationships among genera of the Saccharomycotina (Ascomycota) from multigene phylogenetic analysis of type species. *FEMS Yeast Res* 13(1):23–33.
29. Liu X (2012) Conversion of the biodiesel by-product glycerol by the non-conventional yeast *Pachysolen tannophilus*. PhD thesis (Technical University of Denmark, Kongens Lyngby, Denmark).
30. Wedlock DN, Thornton RJ (1989) Transformation of a glucose negative mutant of *Pachysolen tannophilus* with a plasmid carrying the cloned hexokinase-PII gene from *Saccharomyces cerevisiae*. *Biotechnol Lett* 11(9):601–604.
31. Mühlhausen S, Findeisen P, Plessmann U, Urlaub H, Kollmar M (2016) A novel nuclear genetic code alteration in yeasts and the evolution of codon reassignment in eukaryotes. *Genome Res* 26(7):945–955.
32. Bain JD, Switzer C, Chamberlin AR, Benner SA (1992) Ribosome-mediated incorporation of a non-standard amino acid into a peptide through expansion of the genetic code. *Nature* 356(6369):537–539.
33. Matsuzawa T, et al. (2011) New insights into galactose metabolism by *Schizosaccharomyces pombe*: Isolation and characterization of a galactose-assimilating mutant. *J Biosci Bioeng* 111(2):158–166.
34. Toivari MH, Salusjärvi L, Ruohonen L, Penttilä M (2004) Endogenous xylose pathway in *Saccharomyces cerevisiae*. *Appl Environ Microbiol* 70(6):3681–3686.
35. Gnerre S, et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* 108(4):1513–1518.
36. Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18(5):821–829.
37. LaButti K, Foster B, Han C, Brettin T, Lapidus A (2009) *Gap Resolution: A Software Package for Improving Newbler Genome Assemblies*, Rep No. LBNL-1899E Abs. Available at https://publications.lbl.gov/islandora/object/ir%3A152968. Accessed May 27, 2009.
38. English AC, et al. (2012) Mind the gap: Upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* 7(11):e47768.
39. Grigoriev IV, et al. (2014) MycoCosm portal: Gearing up for 1000 fungal genomes. *Nucleic Acids Res* 42(Database issue):D699–D704.
40. Göker M, Scheuner C, Klenk HP, Stielow JB, Menzel W (2011) Codivergence of mycoviruses with their hosts. *PLoS One* 6(7):e22252.
41. Scheuner C, et al. (2014) Complete genome sequence of *Planctomyces brasiliensis* type strain (DSM 5305(T)), phylogenomic analysis and reclassification of *Planctomycetes* including the descriptions of Gimesia gen. nov., Planctopirus gen. nov. and Rubinisphaera gen. nov. and emended descriptions of the order Planctomycetales and the family Planctomycetaceae. *Stand Genomic Sci* 9:10.
42. Spring S, et al. (2010) The genome sequence of *Methanohalophilus mahii* SLP(T) reveals differences in the energy metabolism among members of the Methanosarcinaceae inhabiting freshwater and saline environments. *Archaea* 2010:690737.
43. Lowe TM, Eddy SR (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25(5):955–964.
44. Di Tommaso P, et al. (2011) T-Coffee: A web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res* 39(Web server issue):W13–W17.
45. Claudel-Renard C, Chevalet C, Faraut T, Kahn D (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res* 31(22):6633–6639.
46. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30(7):1575–1584.
47. Wisecaver JH, Slot JC, Rokas A (2014) The evolution of fungal metabolic pathways. *PLoS Genet* 10(12):e1004816.