A human-curated annotation of the Candida albicans genome

Burkhard Braun, Marco van Het Hoog, Christophe D'Enfert, Mikhail Martchenko, Jan Dungan, Alan Kuo, Diane Inglis, M. Andrew Uhl, Hervé Hogues, Matthew Berriman, et al.

▶ To cite this version:

Burkhard Braun, Marco van Het Hoog, Christophe D'Enfert, Mikhail Martchenko, Jan Dungan, et al.. A human-curated annotation of the Candida albicans genome. PLoS Genetics, 2005, 1 (1), pp.36-57. 10.1371/journal.pgen.0010001 . hal-02682460

HAL Id: hal-02682460

https://hal.inrae.fr/hal-02682460

Submitted on 1 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

A Human-Curated Annotation of the *Candida albicans* Genome

Burkhard R. Braun[1], Marco van het Hoog[2], Christophe d'Enfert[3], Mikhail Martchenko[2], Jan Dungan[4], Alan Kuo[4], Diane O. Inglis[1], M. Andrew Uhl[1], Hervé Hogues[2], Matthew Berriman[5], Michael Lorenz[6], Anastasia Levitin[2], Ursula Oberholzer[2], Catherine Bachewich[2], Doreen Harcus[2], Anne Marcil[2], Daniel Dignard[2], Tatiana Iouk[2], Rosa Zito[2], Lionel Frangeul[7], Fredj Tekaia[8], Kim Rutherford[5], Edwin Wang[2], Carol A. Munro[9], Steve Bates[9], Neil A. Gow[9], Lois L. Hoyer[10], Gerwald Köhler[4], Joachim Morschhäuser[11], George Newport[4], Sadri Znaidi[12], Martine Raymond[12], Bernard Turcotte[13], Gavin Sherlock[14], Maria Costanzo[14], Jan Ihmels[15], Judith Berman[16], Dominique Sanglard[17], Nina Agabian[4], Aaron P. Mitchell[18], Alexander D. Johnson[1], Malcolm Whiteway[2], André Nantel[2*]

1 Department of Microbiology and Immunology, University of California, San Francisco, California, United States of America, **2** Biotechnology Research Institute, National Research Council Canada, Montreal, Quebec, Canada, **3** Unité Postulante Biologie et Pathogénicité Fongiques, INRA USC 2019, Institut Pasteur, Paris, France, **4** Department of Stomatology, University of California, San Francisco, California, United States of America, **5** The Sanger Centre, Cambridge, United Kingdom, **6** Department of Microbiology and Molecular Genetics, Utah-Houston Medical School, Houston, Texas, United States of America, **7** Plate-Forme Intégration et Analyse Génomique, Institut Pasteur, Paris, France, **8** Unité de Génétique Moléculaire des Levures, Institut Pasteur, Paris, France, **9** School of Medical Sciences, University of Aberdeen, Institute of Medical Sciences, Foresterhill, Aberdeen, United Kingdom, **10** Department of Veterinary Pathobiology, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **11** Institut für Molekulare Infektionsbiologie, Universität Wurzburg, Wurzburg, Germany, **12** Institut de Recherches Cliniques de Montreal, Montreal, Quebec, Canada, **13** Department of Medicine, Royal Victoria Hospital, McGill University, Montreal, Quebec, Canada, **14** Department of Genetics, Stanford University School of Medicine, Palo Alto, California, United States of America, **15** Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel, **16** Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, Minnesota, United States of America, **17** Institute of Microbiology, University Hospital Lausanne, Lausanne, Switzerland, **18** Department of Microbiology and Institute of Cancer Research, Columbia University, New York, New York, United States of America

Recent sequencing and assembly of the genome for the fungal pathogen *Candida albicans* used simple automated procedures for the identification of putative genes. We have reviewed the entire assembly, both by hand and with additional bioinformatic resources, to accurately map and describe 6,354 genes and to identify 246 genes whose original database entries contained sequencing errors (or possibly mutations) that affect their reading frame. Comparison with other fungal genomes permitted the identification of numerous fungus-specific genes that might be targeted for antifungal therapy. We also observed that, compared to other fungi, the protein-coding sequences in the *C. albicans* genome are especially rich in short sequence repeats. Finally, our improved annotation permitted a detailed analysis of several multigene families, and comparative genomic studies showed that *C. albicans* has a far greater catabolic range, encoding respiratory Complex 1, several novel oxidoreductases and ketone body degrading enzymes, malonyl-CoA and enoyl-CoA carriers, several novel amino acid degrading enzymes, a variety of secreted catabolic lipases and proteases, and numerous transporters to assimilate the resulting nutrients. The results of these efforts will ensure that the *Candida* research community has uniform and comprehensive genomic information for medical research as well as for future diagnostic and therapeutic applications.

Citation: Braun BR, van het Hoog M, d'Enfert C, Martchenko M, Dungan J, et al. (2005) A human-curated annotation of the *Candida albicans* genome. PLoS Genet 1(1): e1.

Introduction

Candida albicans is a commonly encountered fungal pathogen responsible for infections generally classed as either superficial (thrush and vaginitis) or systemic (such as life-threatening blood-borne candidiasis) [1,2]. Its life cycle has fascinating aspects that have generated great excitement over the last decade, with an influx of workers and new molecular techniques brought to bear on long-standing problems [3]. Topics of particular interest are the organism's capacity to shift into several different phenotypic states, some with distinct roles in infection, and its recently discovered capacity to mate, providing at least part of a sexual cycle, although population genetic studies indicate that it is still largely a clonal diploid population. Other special adaptations for infection include a battery of externally displayed proteins and secreted digestive enzymes; complex interactions with the host immune system normally keep *C. albicans* at bay as a minor part of the mucosal flora [1,4,5].

Here, we report a detailed annotation of the genome sequence of this organism, bringing the previously available raw sequence to a new level of stability and usability. The genome of *C. albicans* has previously been shotgun sequenced to a level of 10.9-fold coverage [6]. However the assembly of

Received January 10, 2005; Accepted March 14, 2005; Published June 17, 2005
DOI: 10.1371/journal.pgen.0010001

Copyright: © 2005 Braun et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abbreviations: ABC, ATP-binding cassette; CGD, *Candida* Genome Database; *e*-value, expect value; EC, Enzyme Commission; GO, Gene Ontology; IPF, individual protein file; NR, non-redundant; ORF, open reading frame; SGD, *Saccharomyces* Genome Database; SGTC, Stanford Genome Technology Center; STR, short tandem repeat; TMS, transmembrane segment

Editor: Michael Snyder, Yale University, United States of America

*To whom correspondence should be addressed. E-mail: andre.nantel@nrc-cnrc.gc.ca

Synopsis

Candida albicans is a commonly encountered fungal pathogen usually responsible for superficial infections (thrush and vaginitis). However, an estimated 30% of severe fungal infections, most due to *Candida,* result in death. Those who are most at risk include individuals taking immune-suppressive drugs following organ transplantation, people with HIV infection, premature infants, and cancer patients undergoing chemotherapy. Current therapies for this pathogen are made more difficult by the significant secondary effects of anti-fungal drugs that target proteins that are also found in the human host.

Recent sequencing and assembly of the genome for the fungal pathogen *C. albicans* used simple automated procedures for the identification of putative genes. Here, we report a detailed annotation of the 6,354 genes that are present in the genome sequence of this organism, essentially writing the dictionary of the *C. albicans* genome.

Comparison with other fungal genomes permitted the identification of numerous fungus-specific genes that are absent from the human genome and whose products might be targeted for antifungal therapy. The results of these efforts will thus ensure that the *Candida* research community has uniform and comprehensive genomic information for medical research, for the development of functional genomic tools as well as for future diagnostic and therapeutic applications.

this sequence faced special difficulties because the organism is diploid but with little or no gene exchange in the wild. Thus homologous chromosomes show substantial divergence, and many genes are present as two distinctive alleles. This required that the assembly process be aware of the diploid status and be prepared to segregate reads into two alleles for any section of the genome. At the same time, the genome is rich in recently diverged gene families that are easily confused with alleles. This task was further complicated by the absence of a complete physical map of the *C. albicans* genome. Nevertheless, this arduous assembly process resulted in a dataset (assembly 19, with 266 primary contigs over eight chromosomes) that has already yielded a number of significant advances including the production of DNA microarrays [7], libraries of systematic gene knockouts [8], large-scale transposon mutagenesis [9], and the ability of many individual researchers to identify novel genes using bioinformatic tools [10]. Unfortunately, due to the mostly computational methods used in its development, the current genome assembly still contains a significant number of predicted genes that are fragmented, overlapping, or otherwise erroneous. As a consequence, different groups have been using different methods for the identification and classification of *C. albicans* genes, which has hindered communication and complicated comparisons between large-scale datasets.

Following the publication of these early functional genomics studies, it was realized that the needs of the *C. albicans* research community would be better served by a unified gene nomenclature. The results of this community-based effort were initially based on the version 19 computational assembly and preliminary annotation produced independently by various research groups. We used visual inspection of 11,615 putative coding sequences and various bioinformatic tools to refine the quality and description of each open reading frame (ORF).

In all, we provide unique identifiers, coordinates, names, and descriptions for 6,354 genes. With the exception of certain large gene families, we have not annotated the portion of the assembly 19 DNA that was set aside as secondary alleles, instead concentrating on the primary sequence that forms one haploid genome equivalent. Investigation of the identity and relative divergence of all alleles will be an important further project for the *C. albicans* genome, as will finishing and linking the small number of gaps that remain in the primary sequence. In addition, we describe a variety of gene families and we discuss insights into virulence. Finally, we use comparative genomics to point out a variety of additional insights that are illuminated by the high-quality annotation provided here. This project serves as a model for community-based annotation that could be applied by other research communities that wish to improve on automated sequencing pipeline output that may be available for their organisms of interest.

Results/Discussion

The Annotation Process

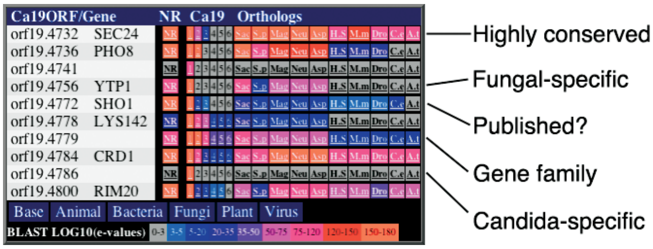
Compilation of *Candida* annotation data. As detailed in Materials and Methods, we used assembly version 19 of the *C. albicans* genome [6] to identify 11,615 putative ORFs. These included genes encoding proteins greater than 150 aa as well as genes encoding smaller proteins of 50–149 aa that have a coding function greater than 0.5 as determined with a GeneMark matrix [11]. These ORFs were then compared to the set of 7,680 *C. albicans* ORFs defined by the Stanford Genome Technology Center (SGTC), thus permitting their classification using the same systematic identifiers of the format orf19.*n* [6]. The 3,936 novel ORFs without an orf19.*n* counterpart were assigned a new reference number of the format orf19.*n.i* where orf19.*n* is the five-prime closest (contig-wise) ORF defined by the SGTC and *i* is an integer that varies between one and the number of novel ORFs found in the orf19.*n* to orf19.($n+1$) interval. To simplify correlation with previously published data that use the orf6.*n* or earlier nomenclatures, we have produced a Web-accessible translation tool (http://candida.bri.nrc.ca).

Positional information for each ORF was merged with data from a variety of different sources, including the SGTC (http://www-sequence.stanford.edu/group/candida/index.html), CandidaDB (http://genolist.pasteur.fr/CandidaDB; [12]), the Agabian laboratory (http://agabian.ucsf.edu/canoDB/anno.php), and the Johnson/Fink laboratories [13], whose annotation data had been updated with a Magpie annotation [7]. This large dataset was then reformatted into EMBL-style files, thus allowing for input in the Artemis annotation software [14]. Volunteer annotators accessed a custom-made database to reserve and download EMBL files containing sequence and annotation data for each of the 266 DNA sequence contigs. To help in validating various and sometimes conflicting sources of information, translated protein sequences from putative *C. albicans* ORFs were compared to putative protein sequences extracted from five fungal genomes—*Saccharomyces cerevisiae* [15,16], *Schizosaccharomyces pombe* [17], *Neurospora crassa* [18], *Aspergillus nidulans* (*Aspergillus nidulans* Database, http://www-genome.wi.mit.edu/annotation/

fungi/aspergillus/), and *Magnaporthe grisea* (*Magnaporthe grisea* Database, http://www-genome.wi.mit.edu/annotation/fungi/magnaporthe/)—as well as to the genomes of five other eukaryotes—*Arabidopsis thaliana* [19], *Drosophila melanogaster* [20], *Caenorhadbitis elegans* [21], *Mus musculus* [22], and *Homo sapiens* [23]—and the GenBank non-redundant (NR) protein database. Comparisons against the translated *C. albicans* genome were also performed to help identify overlapping genes and putative gene families.

To help interpret such a large number of sequence comparisons, we organized sequence similarity data in a Web-accessible database using a novel visualization concept whereby we used a colorimetric display to indicate BLAST similarity, which was easy and rapid to scan visually (Figure 1). The annotators could thus rapidly determine which genes are potentially unique to *C. albicans* (e.g., orf19.4741 and orf19.4786), those that are members of gene families (e.g., orf19.4736 and orf19.4779), genes that only have homologs in fungal genomes (e.g., orf19.4756 and orf19.4778), or those with homologs in all eukaryotic genomes (e.g., orf19.4732 and orf19.4784). Finally, a strong hit against the complete NR database, but not in the other genomes (orf19.4772 and orf19.4800), allowed us to identify *C. albicans* genes that had already been described and submitted to the sequence databases prior to the publication of assembly orf19. Clicking on the relevant boxes opened an additional window containing the precompiled sequence alignments, thus permitting the validation of interesting observations. These visualization tools and the results of sequence comparisons are available at http://candida.bri.nrc.ca/candida/index.cfm?page=blast.

The coordinates and annotations for all 11,615 putative ORFs were thus verified, corrected, and (if necessary) rewritten by the annotators. We removed ORFs smaller than 300 bp with no significant sequence similarity to other genes, either within the *C. albicans* genome or in the sequence databases. In cases where two ORFs overlapped by more than 50%, the smallest gene was removed unless it showed even a slight sequence similarity to another gene in the sequence databases. In other cases, we encountered two, or more, contiguous ORFs that obviously were part of the same gene. These interruptions were usually due to unidentified introns or presumed sequencing errors. In these cases, we decided to merge the relevant gene fragments into a single entry. A total of 5,262 ORF entries were thus removed from the database, or merged with neighboring ORFs, leaving 6,354 confirmed genes. Sequence and/or annotation data can be obtained in Dataset S1 or at http://candida.bri.nrc.ca.



Ca19ORF/Gene		NR	Ca19						Orthologs									
orf19.4732	SEC24	NR	1	2	3	4	5	6	Sac	S.p	Mag	Neu	Asp	H.S	M.m	Dro	C.e	A.t
orf19.4736	PHO8	NR	1	2	3	4	5	6	Sac	S.p	Mag	Neu	Asp	H.S	M.m	Dro	C.e	A.t
orf19.4741		NR	1	2	3	4	5	6	Sac	S.p	Mag	Neu	Asp	H.S	M.m	Dro	C.e	A.t
orf19.4756	YTP1	NR	1	2	3	4	5	6	Sac	S.p	Mag	Neu	Asp	H.S	M.m	Dro	C.e	A.t
orf19.4772	SHO1	NR	1	2	3	4	5	6	Sac	S.p	Mag	Neu	Asp	H.S	M.m	Dro	C.e	A.t
orf19.4778	LYS142	NR	1	2	3	4	5	6	Sac	S.p	Mag	Neu	Asp	H.S	M.m	Dro	C.e	A.t
orf19.4779		NR	1	2	3	4	5	6	Sac	S.p	Mag	Neu	Asp	H.S	M.m	Dro	C.e	A.t
orf19.4784	CRD1	NR	1	2	3	4	5	6	Sac	S.p	Mag	Neu	Asp	H.S	M.m	Dro	C.e	A.t
orf19.4786		NR	1	2	3	4	5	6	Sac	S.p	Mag	Neu	Asp	H.S	M.m	Dro	C.e	A.t
orf19.4800	RIM20	NR	1	2	3	4	5	6	Sac	S.p	Mag	Neu	Asp	H.S	M.m	Dro	C.e	A.t

Base Animal Bacteria Fungi Plant Virus

BLAST LOG10(e-values)	0-3	3-5	5-20	20-35	35-50	50-75	75-120	120-150	150-180

Figure 1. Visualization of Protein Sequence Similarities

Sample from a Web page used by annotators of the *C. albicans* genome to visualize the significance of the best hit from whole-proteome BLASTP searches. Each putative ORF was compared to the NR database, the *Candida* ORF list itself (Ca19; showing results from the four top hits), and amino acid sequences from the proteomes of *S. cerevisiae* (Sac), *S. pombe* (S.p), *M. grisea* (Mag), *N. crassa* (Neu), *H. sapiens* (H.S), *M. musculus* (M.m), *D. melanogaster* (Dro), *C. elegans* (C.e), and *A. thaliana* (A.t). The BLASTP *e*-value from the top hit was converted to a color scale as indicated. Examples of *C. albicans* genes with interesting similarity patterns are indicated.

DOI: 10.1371/journal.pgen.0010001.g001

A nomenclature for *C. albicans* genes. Following consultations with the *C. albicans* research community during the fifth and sixth American Society for Microbiology Conferences on *Candida* and Candidiasis, it was agreed that *C. albicans* gene names should follow the format established for *S. cerevisiae* [24]. Gene names consist of three letters (the gene symbol) followed by an integer (e.g., *ADE12*); the gene symbol should be an acronym for, or relate to, the gene function, gene product, or mutant phenotype. It is preferable that a given gene symbol have only one meaning, so that all genes using that symbol are related in some way, for instance, by sharing a function, participating in a shared pathway, or belonging to the same gene family. In addition, gene symbols that are used in *S. cerevisiae* gene names should retain the same meaning when used for *C. albicans* genes. The prefix 'Ca' has sometimes been used on gene names to denote that a gene is derived from *C. albicans;* however, while the use of prefixes adds clarity to discussions of genes from different species that share a name (e.g., comparing Ca*URA3* to Sc*URA3*), the prefix is not considered part of the gene name proper. Finally, allele designations and deletion symbols should come after the gene name (*ICG1–8* and *icg1Δ* for example). For more details on genetic nomenclature, see the *Candida* Genome Database (CGD; [25]) Web page on this topic (http://www.candidagenome.org/Nomenclature.html).

Wherever possible, genes that are orthologous between *C. albicans* and *S. cerevisiae* should share the same name. We have provided 3,409 suggested names (in the SuggGene field of the EMBL files) for many *C. albicans* ORFs based on their orthology to *S. cerevisiae* genes; these are not yet considered the standard *C. albicans* gene names, but rather provide guidance for investigators wishing to name these genes. CGD assigns standard names to *C. albicans* genes for which there are published data (the PubGene field). The annotation contains 355 such entries. Generally, CGD considers the first published name in the correct format to be the standard name; common usage and uniqueness are also considered. All names that have been used for a gene are collected in CGD, regardless of their format, so that information from the literature can be traced to the correct gene. In the current annotation, additional published gene names have been placed in the Synonym field.

Public access to the data. The complete annotation dataset, results of BLAST sequence similarity searches, and the identification of conserved protein domains can be obtained from our Web page (http://candida.bri.nrc.ca/). Furthermore, CGD (http://www.candidagenome.org), funded by the National Institute for Dental and Craniofacial Research of the National Institutes of Health, will curate the scientific literature and provide tools for accessing and analyzing the *C. albicans* genome sequence. In addition, CGD will act as a central repository for gene names and modifications, as approved by the *C. albicans* research community at the American Society for Microbiology *Candida* and Candidiasis meeting in Austin, Texas, in March 2004. CGD itself will not name *C. albicans* genes, but instead will act as a clearinghouse for the standard gene names and aliases, as the *Saccharomyces* Genome Database (SGD) does for the *S. cerevisiae* community. CGD hopes that researchers will follow CGD's gene nomenclature guidelines (see above) and keep CGD informed of any new gene names. Prior to publication, researchers may reserve a gene name, which will then become the standard name upon publication. Finally, the CandidaDB database (http://genolist.pasteur.fr/CandidaDB) [12], which has provided an annotation of the *C. albicans* genome sequence since January 2001, will be updated to take into account the complete annotation dataset and will continue to provide tools for accessing and analyzing the *C. albicans* genome sequence complementary to those available at the CGD and the Biotechnology Research Institute.

Content and General Statistics

As detailed in Tables 1 and 2, we identified 6,354 genes in version 19 of the *C. albicans* genome assembly. This number is certain to change slightly with time as more data come to light. For instance, 80 of these genes are probably duplicates, having almost identical counterparts near the extremities of sequence contigs. Novel genes may also lie in unsequenced/unassembled gaps between the DNA sequence contigs. We identified 246 genes containing mutations or sequencing errors that result in a frameshift, or the insertion of a stop codon, that will have to be confirmed through resequencing. In the meantime, these elements have been joined as a single ORF entry and tagged with the entry "sequencing error?" inside their Note field. We have also identified 190 genes truncated at the ends of contigs, only 35 of which have an identical counterpart on a potentially overlapping contig. New information will be continuously integrated into the community data as it is submitted.

The mean protein coding length of 1,439 bp (480 aa) is almost identical to what has been observed in *S. cerevisiae* and *S. pombe,* while the gene density stands at one gene per 2,342 bp. Short descriptions for all gene products were provided by annotators, usually based on sequence similarity. A total of 1,218 (19.2%) genes encode unique proteins with no significant homologs in the sequence databases, a percentage almost identical to that observed in the current version of the *S. cerevisiae* annotation [16]. An additional 819 (12.9%) gene products exhibited significant similarities to other proteins of unknown function. Furthermore, we have provided Enzyme Commission (EC) numbers and Gene Ontology (GO) terms for 1,334 and 3,586 gene products, respectively.

Intron analysis. There are 215 ORFs containing at least one intron, four of which have two introns, one gene (encoding the Hxt4p transporter) has three, and the *SIN3* gene has four. A total of 43 (20.2%) of these genes encode ribosomal proteins, 63 (29.6%) encode products with enzymatic activity, and 26 (12.2%) encode *trans*-membrane proteins involved in small molecule transport. We measured the relative position of introns in their host ORFs and observed that a significant proportion of them are located in the 5′ end of ORFs, with 32% of introns being located within the first 10% of the coding sequences. A survey of the distribution of introns in 18 eukaryotic genomes, including *S. cerevisiae* and *H. sapiens,* also indicated a similar bias in intron-poor genomes. It has been argued that this 5′ bias is an indication that introns are particularly difficult to remove by cDNA recombination, because of the high activity of these genes and paucity of full-length cDNA, and that this finding lends some support to the idea that introns are being lost more frequently than they are being gained in these lineages [26], although a more recent study of four fungal genomes suggests the presence of additional mechanisms [27].

We surveyed the intron phase distribution and found that *C. albicans* has 50.5%, 20.4%, and 29.1% of phase zero, one, and two introns, respectively. A similar result was observed in fungal, plant, and animal genomes [27,28], suggesting that a similar intron phase distribution may be present in ancient introns and that the intron loss has no preference selection on intron phases. Seventy out of 215 intron-containing ORFs have reciprocal best matches with *S. cerevisiae* genes that also contain introns. Among these 70 ORFs, 25 introns (35.7%) share the same position and the same phase. This suggests that these commonly positioned introns descended from a common ancestor, as suggested previously [29].

Analysis of protein domains. Table 3 shows the most abundant protein domains that were identified in the *C. albicans* proteome. As a comparison, we also performed this analysis on the same ten eukaryotic proteomes that were used in the BLASTP sequence comparisons. Compared to the *S.*

Table 1. Features of Completed Fungal Genomes

Species	Length (Mb)	Number of Genes	Mean Coding Length (bp)	Gene Density[a]	Coding Percent	Introns	Unique Proteins[b]	Reference
C. albicans	14.88	6,354	1,439	2,342	61.5%	224	1,218 (19.2%)	
S. cerevisiae	12.16	5,726	1,485	2,124	69.9%	272	1,104 (19.1%)	16
S. pombe	12.46	4,929	1,426	2,528	57.5%	2,034	681 (14%)	17
N. crassa	38.64	10,082	1,673	3,832	43.6%	17,139	4,140 (41%)	18
C. glabrata	12.28	5,283	1,479	2,324	65.0%	nd	nd	128
Kluyveromyces lactis	10.63	5,329	1,383	1,995	71.6%	nd	nd	128
Debaryomyces hansenii	12.22	6,906	1,167	1,769	79.2%	nd	nd	128
Yarrowia lipolytica	20.50	6,703	1,428	3,058	46.3%	nd	nd	128
Cryptococcus neoformans	19.05	6,572	1,909	2,925	65.8%	34,909	35%	129

[a] Number of base pairs in genome divided by number of genes.

[b] Number and proportion of proteins with no significant similarity to known proteins.

nd, not determined.

DOI: 10.1371/journal.pgen.0010001.t001

Table 2. Statistics of the *C. albicans* Annotation

Parameter	Value
Number of genes	6,354
Published/reserved gene names	355
Suggested gene names	3,409
Genes with synonyms/multiple names	85
Genes in putative contig overlaps	80
Truncated genes at end of contigs	190
Sequencing errors/frameshift mutations	246
Gene product descriptions	4,317
Conserved hypothetical proteins	819
Hypothetical proteins	1,218
Products with EC numbers	1,334
Total number of EC numbers	1,463
Products with GO Terms[a]	3,438
Number of GO terms[a]	13,835

[a] Excluding "unknown."

DOI: 10.1371/journal.pgen.0010001.t002

pombe and *S. cerevisiae* proteomes, the *C. albicans* proteome shows a slight increase in the abundance of leucine-rich repeats (IPR001611), some zinc finger transcription factors (IPR001138), esterases/lipases (IPR000379), and *trans*-membrane transporters for polyamines (IPR002293) and for amino acids (IPR004841). If the analysis is expanded to the other fungal proteomes, only the increased abundance in leucine-rich repeats appears to be unique to *C. albicans*.

Genome-Based Identification of Antifungal Targets

One of the main arguments supporting large-scale sequencing projects for fungal pathogens is the hope of finding novel antifungal targets, particularly those that are absent from the genome of their host. Table 4 shows a list of 228 *C. albicans* genes that have a very strong sequence homolog (based on a top hit BLASTP expect value (*e*-value) $< 1e^{-45}$) in all five fungal genomes but no significant sequence similarity (best BLASTP *e*-value $> 1e^{-10}$) to genes in the genomes of either humans or mice. For example, this list includes *FKS1,* which encodes a 1,3-beta-glucan synthase that is the target for the cell wall agents called echinocandins [30]. The list includes 46 gene products that are assumed to be located on the plasma membrane, 71 that are predicted to be involved in the transport of small molecules, and 21 that appear to be involved, directly or indirectly, with cell wall synthesis. Furthermore, 41 gene products have been associated with an EC number, indicating an enzymatic activity, with phospholipases being the most abundant. The roles and sites of action of these gene products suggest that they would be both accessible and theoretically amenable to inhibition by small molecules.

Short Tandem Repeats

Short tandem repeats (STRs), also called short sequence repeats or microsatellite DNA, play an important role in evolution and have been used to characterize population variability. Although they can arise through DNA polymerase slippage and unequal recombination, whole-genome analysis has suggested that additional mechanisms for the control of STR production/correction remain to be identified [31–33]. Jones et al. [6] scanned the *C. albicans* genome for STRs of unit sizes between two and five and identified 1,940 trinucleotide repeats in their ORF sequences. To confirm that this high STR frequency is indeed a hallmark of the *C. albicans* genome, we used a statistical approach to measure repeat frequencies in four completed fungal genomes with an emphasis on STRs that affect protein sequences. We used randomized genome sequences to calculate the probability that each potential STR (including mutations that may arise following the amplification event) is nonrandom, and used only those with greater than 95% probability.

As can be seen in Datasets S2–S5 and Table 5, the STR frequencies in *C. albicans* and *N. crassa* are significantly greater than the frequencies observed in *S. cerevisiae* and *S. pombe*. Repeats that occur inside coding sequences are further characterized in Table 5. As would be expected, repeats with a modulo of three are more common in coding sequences, although we note that species with the greatest STR frequency have the smallest proportion of repeats that would break a reading frame. While coding sequence STRs in *C. albicans* and the other fungi most commonly encode for repeats of glutamine, asparagine, glutamic acid, and aspartic acid, we note that some of the repeats that are prevalent in *C. albicans* genes are distinct. Repeats of the ACT (threonine) and TCA (serine) codons are known to be especially rare in most taxa [31,33]. Correlating STR distribution with Gene Ontology annotations shows that a significant proportion of the *C. albicans* genes whose products are classified as DNA-binding proteins or cytoskeletal elements also contain STRs. Several gene products have been shown to play a role in the generation/correction of novel STRs in eukaryotes [34]. A comparison of the aa sequences of Rad51p, Rad52p, Mre1p, Hpr5p, and Pob3p from *C. albicans, S. cerevisiae, S. pombe,* and *N. crassa* did not reveal any significant correlation that could be associated with changes in the STR distribution. The high proportion of STRs in *C. albicans* genes argues that this organism would make a better model than *S. cerevisiae* for studying the creation and elongation of these elements that cause a variety of neuromuscular pathologies in humans. Our observations further indicate that future studies on STR frequency in eukaryotic genomes should include a broader spectrum of fungal genomes. The *S. cerevisiae* genome has been used as the fungal representative in comparative studies published to date [31–33].

Identification of Spurious Genes

Some of the 6,354 predicted ORFs are likely to be spurious. We used data from *S. cerevisiae* to model an approach that combines gene length, gene homology, and gene expression data to search for spurious gene candidates. Theoretically, genes with no sequence similarity and with expression profiles that do not correlate with other known genes are much more likely to be spurious. In an earlier study, spurious genes in *S. cerevisiae* were identified by sequence comparison between four closely related yeast species [16]. Most did not have orthologs with other eukaryotes, were of short length, and had expression profiles that were not significantly correlated with those of other genes in the genome (Figure 2A and 2B). Combining both the criteria of sequence homology and expression correlation produced a list of *S. cerevisiae* candidate genes that was highly enriched for ORFs that were considered to be spurious based on the separate sequence comparison between the closely related species. We repeated this homology/expression/length analysis on genes

Table 3. Number, Abundance Ranking, and Proportion of Gene Products Containing the Indicated Interpro Protein Domain in *C. albicans* and Other Eukaryotes

Interpro	Description	*C. albicans*	*S. cerevisiae*	*S. pombe*	*A. niger*	*M. grisea*	*N. crassa*	*A. thaliana*	*C. elegans*	*D. melanogaster*	*M. musculus*	*H. sapiens*
IPR011009	Protein kinase-like	124 (1) 1.95%	136 (1) 2.19%	129 (2) 2.59%	140 (4) 1.47%	146 (3) 1.31%	142 (1) 1.41%	1,146 (1) 3.98%	1,015 (2) 3.05%	472 (1) 2.64%	592 (6) 2.33%	700 (4) 2.76%
IPR008938	ARM repeat fold	112 (2) 1.76%	109 (6) 1.76%	131 (1) 2.63%	116 (12) 1.22%	108 (11) 0.97%	120 (5) 1.19%	390 (12) 1.35%	384 (16) 1.15%	270 (7) 1.51%	364 (11) 1.43%	388 (13) 1.53%
IPR011046	WD40-like	111 (3) 1.74%	120 (3) 1.93%	121 (3) 2.43%	125 (9) 1.31%	105 (12) 0.95%	120 (4) 1.19%	267 (23) 0.93%	310 (18) 0.93%	245 (11) 1.37%	278 (22) 1.10%	326 (20) 1.29%
IPR000719	Protein kinase	105 (4) 1.65%	123 (2) 1.98%	109 (5) 2.19%	117 (11) 1.23%	117 (7) 1.05%	126 (2) 1.25%	1,122 (2) 3.90%	908 (3) 2.73%	401 (3) 2.24%	572 (7) 2.25%	678 (5) 2.68%
IPR001680	G protein beta WD40 repeat	100 (5) 1.57%	101 (8) 1.63%	119 (4) 2.39%	126 (8) 1.32%	114 (9) 1.03%	121 (3) 1.20%	297 (18) 1.03%	283 (22) 0.85%	238 (12) 1.33%	286 (19) 1.13%	333 (19) 1.32%
IPR002290	Serine/threonine protein kinase	96 (6) 1.51%	114 (4) 1.84%	105 (6) 2.11%	102 (17) 1.07%	99 (14) 0.89%	99 (7) 0.98%	1,068 (3) 3.71%	829 (4) 2.49%	359 (4) 2.01%	509 (8) 2.01%	635 (6) 2.51%
IPR008271	Serine/threonine protein kinase, active site	92 (7) 1.44%	111 (5) 1.79%	94 (8) 1.89%	86 (19) 0.90%	78 (27) 0.70%	80 (13) 0.79%	848 (5) 2.95%	453 (12) 1.36%	253 (9) 1.42%	322 (14) 1.27%	415 (10) 1.64%
IPR001138	Fungal transcriptional regulatory protein, N-terminal	76 (9) 1.19%	53 (16) 0.85%	31 (24) 0.62%	217 (2) 2.27%	122 (6) 1.10%	93 (9) 0.92%	1 (3125) 0%	—	—	—	1 (3209) 0%
IPR003593	AAA ATPase	75 (10) 1.18%	82 (9) 1.32%	69 (11) 1.39%	107 (15) 1.12%	92 (16) 0.83%	93 (10) 0.92%	322 (16) 1.12%	185 (41) 0.56%	160 (24) 0.89%	136 (52) 0.54%	172 (43) 0.68%
IPR007114	Major facilitator superfamily	71 (11) 1.11%	62 (14) 1%	52 (14) 1.05%	293 (1) 3.07%	175 (1) 1.58%	115 (6) 1.14%	107 (66) 0.37%	211 (34) 0.63%	139 (30) 0.78%	88 (88) 0.35%	97 (85) 0.38%
IPR008941	TPR-like	71 (12) 1.11%	63 (13) 1.02%	70 (10) 1.41%	112 (13) 1.17%	86 (19) 0.77%	82 (12) 0.81%	623 (7) 2.16%	218 (32) 0.65%	184 (18) 1.03%	225 (30) 0.89%	245 (31) 0.97%
IPR001410	DEAD/DEAH box helicase	65 (13) 1.02%	78 (10) 1.26%	68 (12) 1.37%	71 (29) 0.74%	71 (33) 0.64%	74 (16) 0.73%	156 (41) 0.54%	178 (43) 0.53%	99 (49) 0.55%	117 (66) 0.46%	126 (63) 0.50%
IPR001650	Helicase, C-terminal	63 (14) 0.99%	76 (11) 1.23%	67 (13) 1.35%	70 (30) 0.73%	73 (31) 0.66%	70 (18) 0.69%	153 (43) 0.53%	165 (52) 0.50%	101 (44) 0.56%	110 (68) 0.43%	120 (68) 0.47%
IPR001611	Leucine-rich repeat	62 (15) 0.97%	14 (98) 0.23%	13 (95) 0.26%	16 (142) 0.17%	14 (157) 0.13%	16 (104) 0.16%	556 (8) 1.93%	158 (55) 0.47%	174 (19) 0.97%	261 (24) 1.03%	288 (25) 1.14%
IPR000504	RNA-binding region RNP-1 (RNA recognition motif)	60 (16) 0.94%	65 (12) 1.05%	86 (9) 1.73%	81 (24) 0.85%	81 (24) 0.73%	76 (15) 0.75%	346 (14) 1.20%	266 (23) 0.80%	296 (6) 1.66%	305 (17) 1.20%	333 (18) 1.32%
IPR000379	Esterase/lipase/ thioesterase	54 (17) 0.85%	37 (22) 0.60%	26 (33) 0.52%	103 (16) 1.08%	115 (8) 1.04%	76 (14) 0.75%	237 (25) 0.82%	223 (28) 0.67%	165 (22) 0.92%	118 (65) 0.47%	112 (73) 0.44%
IPR007087	Zinc finger, C_2H_2 type	53 (18) 0.83%	53 (15) 0.85%	34 (20) 0.68%	84 (20) 0.88%	99 (13) 0.89%	97 (8) 0.96%	190 (29) 0.66%	389 (15) 1.17%	433 (2) 2.42%	712 (4) 2.81%	879 (1) 3.47%
IPR000345	Cytochrome c heme-binding site	43 (19) 0.68%	39 (21) 0.63%	37 (17) 0.74%	61 (35) 0.64%	80 (25) 0.72%	51 (26) 0.51%	281 (19) 0.98%	238 (27) 0.71%	202 (16) 1.13%	262 (23) 1.03%	303 (24) 1.20%
IPR001841	Zinc finger, RING	41 (20) 0.64%	35 (26) 0.56%	46 (16) 0.92%	38 (54) 0.40%	48 (41) 0.43%	52 (23) 0.52%	489 (10) 1.70%	261 (24) 0.78%	159 (25) 0.89%	293 (18) 1.15%	339 (17) 1.34%
IPR009057	Homeodomain-like	40 (21) 0.63%	33 (31) 0.53%	27 (31) 0.54%	67 (31) 0.70%	128 (5) 1.15%	42 (31) 0.42%	490 (9) 1.70%	404 (14) 1.21%	210 (15) 1.17%	362 (12) 1.43%	397 (11) 1.57%
IPR005828	General substrate transporter	38 (22) 0.60%	39 (20) 0.63%	29 (27) 0.58%	128 (7) 1.34%	75 (30) 0.68%	52 (24) 0.52%	88 (87) 0.31%	101 (85) 0.30%	95 (51) 0.53%	38 (230) 0.15%	45 (202) 0.18%
IPR008994	Nucleic-acid-binding OB-fold	38 (23) 0.60%	41 (18) 0.66%	51 (15) 1.02%	44 (50) 0.46%	45 (43) 0.41%	45 (30) 0.45%	268 (22) 0.93%	92 (91) 0.28%	70 (61) 0.39%	112 (67) 0.44%	94 (86) 0.37%
IPR002048	Calcium-binding EF-hand	38 (24) 0.60%	26 (45) 0.42%	33 (21) 0.66%	41 (51) 0.43%	37 (57) 0.33%	46 (28) 0.46%	269 (2%1) 0.93	219 (31) 0.66%	172 (20) 0.96%	286 (20) 1.13%	321 (21) 1.27%
IPR002293	Amino acid/polyamine transporter I	37 (25) 0.58%	25 (49) 0.40%	21 (47) 0.42%	58 (36) 0.61%	26 (85) 0.23%	18 (90) 0.18%	18 (459) 0.06%	32 (283) 0.10%	27 (211) 0.15%	26 (337) 0.10%	23 (411) 0.09%
IPR005225	Small GTP-binding protein domain	37 (26) 0.58%	43 (17) 0.69%	35 (19) 0.70%	35 (61) 0.37%	38 (54) 0.34%	38 (36) 0.38%	128 (53) 0.44%	118 (72) 0.35%	117 (34) 0.65%	163 (39) 0.64%	180 (39) 0.71%
IPR005829	Sugar transporter superfamily	35 (27) 0.55%	40 (19) 0.64%	23 (40) 0.46%	107 (14) 1.12%	61 (34) 0.55%	47 (27) 0.47%	109 (64) 0.38%	70 (130) 0.21%	76 (60) 0.43%	49 (168) 0.19%	58 (142) 0.23%
IPR004841	Amino acid permease-associated region	34 (28) 0.53%	24 (52) 0.39%	21 (48) 0.42%	48 (47) 0.50%	25 (89) 0.23%	18 (91) 0.18%	18 (470) 0.06%	30 (295) 0.09%	27 (203) 0.15%	25 (367) 0.10%	22 (418) 0.09%
IPR001440	TPR repeat	34 (29) 0.53%	33 (29) 0.53%	37 (18) 0.74%	53 (43) 0.56%	42 (47) 0.38%	38 (35) 0.38%	152 (44) 0.53%	125 (69) 0.38%	108 (36) 0.60%	151 (46) 0.60%	160 (47) 0.63%
IPR001993	Mitochondrial substrate carrier	33 (30) 0.52%	34 (28) 0.55%	23 (38) 0.46%	36 (60) 0.38%	36 (60) 0.32%	35 (40) 0.35%	61 (140) 0.21%	59 (160) 0.18%	69 (63) 0.39%	58 (135) 0.23%	61 (137) 0.24%

Numbers represent how many gene products have the given domain. Ordered ranking of each domain is given in parentheses. Percentages represent the proportion of gene products that contain at least one of the domains.

DOI: 10.1371/journal.pgen.0010001.t003

Table 4. Genes from *C. albicans* with a Strong Homolog in the *S. cerevisiae*, *S. pombe, A. niger, M. gri*sea, and *N. crassa* genomes but Absent from the *H. sapiens* and *M. musculus* Genomes

Systematic ID	Name	Product
orf19.2929	*GSL2*	1,3-Beta-D-glucan synthase subunit
orf19.2495	*FKS1*	1,3-Beta-glucan synthase; target for echinocandin antifungal drugs
orf19.1517		2-Dehydro-3-deoxy-phosphoheptonate aldolase
orf19.6086	*LEU4*	2-Isopropylmalalate synthase
orf19.3106	*MET16*	3′-Phosphoadenylylsulfate reductase
orf19.99	*MET222*	3′(2′)5′-Bisphosphate nucleotidase, possibly involved in salt tolerance and methionine synthesis
orf19.4060	*ARO4*	3-Deoxy-D-arabino-heptulosonate 7-phosphate (DAHP) synthase isoenzyme
orf19.2269		3-Phosphoserine phosphatase
orf19.5113	*ADH2*	Alcohol dehydrogenase
orf19.309	*DAL5*	Allantoate permease
orf19.3208	*DAL6*	Allantoate permease
orf19.5023	*DAL7*	Allantoate permease
orf19.5859	*DAL8*	Allantoate permease
orf19.6522		Allantoate permease
orf19.6956	*DAL9*	Allantoate permease
orf19.313	*DAL4*	Allantoin permease
orf19.1663	*KRE2*	Alpha-1,2-mannosyltransferase
orf19.1665	*KTR1*	Alpha-1,2-mannosyltransferase involved in N- and O-linked glycosylation
orf19.1375	*LEU6*	Alpha-isopropylmalate synthase
orf19.2810	*AAP1*	Amino acid permease
orf19.3795	*AGP3*	Amino acid permease
orf19.4679	*AGP2*	Amino acid permease
orf19.5672	*MEP2*	Ammonia permease
orf19.7428	*APN1*	Apurinic/apyrimidinic endonuclease/3′-repair diesterase
orf19.111	*CAN1*	Arginine permease
orf19.97	*CAN2*	Arginine permease
orf19.1235	*HOM3*	Aspartate kinase (L-aspartate 4-P-transferase)
orf19.6959	*HOM32*	Aspartate kinase (L-aspartate 4-P-transferase); L-aspartate 4-P-transferase
orf19.1559	*HOM2*	Aspartate-semialdehyde dehydrogenase; threonine and methionine pathway
orf19.4026	*HIS1*	ATP phosphoribosyltransferase
orf19.5970	*HPR5*	ATP-dependent DNA helicase involved in DNA repair
orf19.7213		ATP-dependent RNA helicase
orf19.5604	*MDR1*	Benomyl/methotrexate resistance protein
orf19.7670		Ca^{2+}/H^{+} antiporter conserved in fungi
orf19.5796	*SHE9*	Causes growth arrest when overexpressed
orf19.5531	*CDC37*	Cell division control protein
orf19.807	*CHS5*	Chitin biosynthesis protein
orf19.5188	*CHS1*	Chitin synthase
orf19.7298	*CHS2*	Chitin synthase 2
orf19.5384	*CHS8*	Chitin synthase 8
orf19.2946	*HNM4*	Choline permease
orf19.1170		Chorismate mutase
orf19.1986	*ARO2*	Chorismate synthase
orf19.3489	*ARO22*	Chorismate synthase
orf19.5932		Conserved hypothetical membrane protein
orf19.1240		Conserved hypothetical protein
orf19.246		Conserved hypothetical protein
orf19.2703		Conserved hypothetical protein
orf19.3288		Conserved hypothetical protein
orf19.4907		Conserved hypothetical protein
orf19.5342		Conserved hypothetical protein
orf19.5541		Conserved hypothetical protein
orf19.5605		Conserved hypothetical protein
orf19.5667	*MNR2*	Conserved hypothetical protein; putative ion transporter
orf19.1427		Conserved hypothetical transporter
orf19.988		Conserved membrane protein
orf19.778	*PIL1*	Conserved protein
orf19.1989	*DCW1*	Defective cell wall
orf19.2445	*DIP5*	Dicarboxylic amino acid permease
orf19.579	*FOL1*	Dihydroneopterin aldolase, dihydro-6-hydroxymethylpterin pyrophosphokinase
orf19.4040	*ILV3*	Dihydroxyacid dehydratase
orf19.843		DNA repair exonuclease
orf19.3417	*ACF2*	Endo-1,3-beta-glucanase; involved in actin polymerization
orf19.3066	*ACF3*	Endo-1,3-beta-glucanase
orf19.979	*FAS1*	Fatty-acyl-CoA synthase, beta chain
orf19.3203	*RCY1*	F-box protein: endocytic membrane traffic, recycling

Table 4. Continued

Systematic ID	Name	Product
orf19.2075	*DFG5*	Filamentous growth, cell polarity, and elongation
orf19.5285	*PST3*	Flavodoxin
orf19.5286	*YCP4*	Flavodoxin
orf19.4618	*FBA1*	Fructose-bisphosphate aldolase
orf19.6882	*OSM1*	Fumarate reductase flavoprotein subunit
orf19.5658	*MNN10*	Mannosyltransferase
orf19.1232	*VRG4*	GDP-mannose transporter into the lumen of the Golgi
orf19.1799	*GAP6*	General amino acid permease
orf19.3195	*GAP3*	General amino acid permease
orf19.4304	*GAP1*	General amino acid permease
orf19.6659	*GAP5*	General amino acid permease
orf19.2990	*XOG1*	Glucan 1,3-beta-glucosidase
orf19.1719	*SGA1*	Glucoamylase
orf19.5815	*SCT2*	Glycerol-3-phosphate acyltransferase
orf19.1289	*SCT1*	Glycerol-3-phosphate O-acyltransferase
orf19.1978	*GIT2*	Glycerophosphoinositol permease
orf19.1979	*GIT3*	Glycerophosphoinositol permease
orf19.1980	*GIT4*	Glycerophosphoinositol permease
orf19.377	*PHR3*	Glycophospholipid
orf19.4035	*GAS1*	Glycosylphosphatidylinositol anchored surface protein
orf19.2862	*RIB1*	GTP cyclohydrolase II, first step in riboflavin biosynthesis
orf19.2626	*RGD2*	GTPase activating protein
orf19.730	*RGD3*	GTPase activating protein (GAP) for Rho
orf19.1422	*FZO1*	GTPase required for biogenesis of mitochondria
orf19.6387	*HSP104*	Heat shock protein 104
orf19.1193	*GNP2*	High-affinity glutamine permease
orf19.7565	*GNP3*	High-affinity glutamine permease
orf19.7566	*GNP1*	High-affinity glutamine permease
orf19.2337	*ALP1*	High-affinity permease for basic amino acids
orf19.3222		Highly conserved hypothetical protein
orf19.3395		Highly conserved hypothetical protein
orf19.2350		Highly conserved hypothetical protein, MFS family transporter
orf19.4940	*HIP1*	Histidine permease
orf19.5639	*HIS4*	Histidinol dehydrogenase
orf19.4506	*LYS22*	Homocitrate synthase
orf19.772	*LYS21*	Homocitrate synthase
orf19.923	*THR1*	Homoserine kinase
orf19.2618	*MET2*	Homoserine O-acetyltransferase
orf19.2987		Hypothetical membrane protein
orf19.5505	*HIS7*	Imidazole glycerol phosphate synthase; histidine biosynthesis
orf19.183	*HIS3*	Imidazoleglycerol-phosphate dehydratase
orf19.3355	*ISN1*	Inosine 5′-monophosphate 5′-nucleotidase
orf19.4379	*PRP13*	Integral membrane mitochondrial protein
orf19.1112	*BUD7*	Involved in bud-site selection
orf19.6068	*SVF1*	Involved in diauxic shift
orf19.5986	*THI4*	Involved in thiamine biosynthesis pathway and DNA repair
orf19.2179	*ARN1*	Iron-siderophore transporter
orf19.6844	*ICL1*	Isocitrate lyase
orf19.3412	*ATG15*	Lipase involved in autophagy
orf19.5839	*PDR17*	Lipid biosynthesis and multidrug resistance
orf19.3149	*LSP1*	Long-chain base; stimulates phosphorylation
orf19.1614	*MEP1*	Low-affinity high-capacity ammonium permease
orf19.600	*TRK1*	Low-affinity potassium transporter
orf19.3663	*PHO91*	Low-affinity phosphate transporter
orf19.3622	*ANP1*	Mannan 8; Golgi mannosyltransferase required for protein glycosylation
orf19.3171	*ACH1*	Mannose-containing glycoprotein that binds concanavalin A; acetyl-CoA hydrolase
orf19.4494	*KTR2*	Mannosyltransferase
orf19.1010	*KTR3*	Mannosyltransferase involved in O- and N-linked glycosylation
orf19.7158		Member of allantoate permease family
orf19.2160	*NAG4*	Membrane transporter
orf19.2170		Membrane transporter
orf19.4805	*RSN1*	Membrane transporter
orf19.4737	*DHA12*	Membrane transporter of the MFS-MDR family
orf19.7391	*OCH1*	Membrane-bound alpha-1,6-mannosyltransferase
orf19.5811	*MET1*	Methionine metabolism; siroheme synthase; uroporphyrin-3 C-methyltransferase
orf19.2551	*MET6*	Methionine-synthesizing 5-methyltetrahydropteroyltriglutamate-homocysteine methyltransferase
orf19.339	*NDE1*	Mitochondria-directed NADH dehydrogenase

Table 4. Continued

Systematic ID	Name	Product
orf19.2956	*MGM101*	Mitochondrial genome maintenance protein
orf19.4351	*PRP12*	Mitochondrial inner membrane protein
orf19.88	*ILV5*	Mitochondrial ketol-acid reductoisomerase
orf19.2597	*MRS2*	Mitochondrial magnesium ion transporter, essential for splicing of group II introns
orf19.3422	*FMP27*	Mitochondrial protein
orf19.2115		Molybdopterin-converting factor
orf19.943	*FET33*	Multicopper ferro-O_2-oxidoreductase
orf19.6577	*TPO1*	Multidrug resistance protein
orf19.7148	*TPO2*	Multidrug resistance proteins; polyamine transport protein
orf19.4779		Multidrug resistance transporter
orf19.304		Multidrug-resistance-type transporter
orf19.367	*CNH1*	Na^+/H^+ antiporter
orf19.5713	*NDE1*	NADH dehydrogenase
orf19.125	*EBP1*	NADH: flavin oxidoreductase (old yellow enzyme)
orf19.3612	*PST2*	NADH: quinone oxidoreductase; 1,4-benzoquinone reductase
orf19.2192	*GDH2*	NAD-specific glutamate dehydrogenase
orf19.3443	*OYE2*	NAPDH dehydrogenase (old yellow enzyme)
orf19.3433	*OYE23*	NAPDH dehydrogenase (old yellow enzyme), isoform 2
orf19.7176	*NPT1*	Nnicotinate phosphoribosyltransferase
orf19.2055	*NPL6*	Nuclear protein localization factor
orf19.176		Oligopeptide transporter
orf19.2292	*OPT4*	Oligopeptide transporter protein
orf19.3746	*OPT2*	Oligopeptide transporter protein
orf19.3749	*OPT3*	Oligopeptide transporter protein
orf19.4655	*OPT6*	Oligopeptide transporter protein
orf19.5121	*OPT5*	Oligopeptide transporter protein
orf19.5673	*OPT7*	Oligopeptide transporter protein
orf19.2602	*OPT1*	Oligopeptide transporter specific for tetra- and pentapeptides
orf19.6500	*ECM40*	Ornithine acetyltransferase
orf19.1291	*ABZ1*	Para-aminobenzoate synthase (PABA)
orf19.4704	*ARO1*	Pentafunctional arom polypeptide
orf19.3718		Peptide transporter
orf19.34	*GIT1*	Permease involved in the uptake of glycerophosphoinositol (GroPIns)
orf19.6081	*PHR2*	pH-regulated cell wall protein
orf19.3829	*PHR1*	pH-regulated GPI-anchored membrane protein that is required for morphogenesis
orf19.169	*CHO2*	Phosphatidyl-ethanolamine N-methyltransferase
orf19.1027	*PDR16*	Phosphatidylinositol transfer protein; drug resistance
orf19.677	*CHO1*	Phosphatidylserine synthase
orf19.5102	*PLB5*	Phospholipase B/lysophospholipase
orf19.6594	*PLB4*	Phospholipase B/lysophospholipase
orf19.689	*PLB1*	Phospholipase B/lysophospholipase
orf19.690	*PLB2*	Phospholipase B/lysophospholipase
orf19.7484	*ADE1*	Phosphoribosyl-amidoimidazole-succinocarboxamide synthetase
orf19.5906	*ADE2*	Phosphoribosylamino-imidazole-carboxylase; purine biosynthesis
orf19.4381	*VTC2*	Polyphosphate synthetase
orf19.3363	*VTC4*	Polyphosphate synthetase
orf19.1504		Potential patatin-like phospholipase
orf19.5426		Predicted esterase of the alpha-beta hydrolase superfamily
orf19.4605	*TYR1*	Prephenate dehydrogenase; tyrosine biosynthesis
orf19.2945	*PUT4*	Proline permease
orf19.7577	*MSS51*	Protein involved in maturation of COX1 and COB mRNA
orf19.6520		Putative allantoate permease
orf19.5995	*MCA1*	Putative cysteine protease
orf19.1607	*ALR1*	Putative divalent cation transporter
orf19.2798		Putative helicase
orf19.4475	*MNT4*	Putative mannosyltransferase
orf19.5029	*MODF*	Putative membrane protein
orf19.685	*YHM1*	Putative mitochondrial carrier protein
orf19.4446		Putative permease
orf19.5031	*SSK1*	Putative reponse regulator two-component phosphorelay gene
orf19.2151	*SEY1*	Putative stress-related vesicular transport protein
orf19.3232		Putative transporter
orf19.4608	*PDC12*	Pyruvate decarboxylase I
orf19.4650	*ILV6*	Regulatory subunit of acetolacetate synthase
orf19.1311	*SPO75*	Related to yeast sporulation protein
orf19.7383	*MNN9*	Required for complex glycosylation
orf19.4024	*RIB5*	Riboflavin synthase

Table 4. Continued

Systematic ID	Name	Product
orf19.6727	*RIT1*	Ribosyltransferase of initiator tRNA methionine
orf19.5071	*NRP1*	RNA-binding Ran zinc finger protein
orf19.1789.1	*LYS1*	Saccharopine dehydrogenase
orf19.1631	*ERG6*	S-adenosyl-methionine delta-24-sterol-C-methyltransferase
orf19.1474	*SLA1*	SH3 domain protein involved in assembly of cortical actin cytoskeleton
orf19.3693	*GAS12*	Similar to GPI-anchored surface protein GAS1
orf19.4151	*SPO1*	Similar to phospholipase B
orf19.473	*TPO4*	Sperimidine transporter
orf19.341	*TPO3*	Spermidine exporter, MDR-type pump
orf19.5827	*BUB2*	Spindle body component required for cell cycle arrest in response to loss of microtubule function
orf19.2947	*SNZ1*	Stationary phase protein
orf19.1203	*SRO77*	Suppressor of defect in the small GTPase Rho3p
orf19.277	*THI6*	Thiamin-phosphate pyrophosphorylase and hydroxyethylthiazole kinase
orf19.889	*THI20*	Thiamine biosynthesis; phosphomethylpyrimidine kinase
orf19.4290	*TRR1*	Thioredoxin reductase
orf19.3038	*TPS2*	Threalose-6-phosphate phosphatase
orf19.814	*SSY1.5*	Transcriptional regulator of multiple amino acid permeases
orf19.4335	*TNA1*	Transporter of nicotinic acid
orf19.6640	*TPS1*	Trehalose-6-phosphate synthase
orf19.6511	*TRL1*	tRNA ligase
orf19.7205	*DUR7*	Urea active transport protein
orf19.781	*DUR3*	Urea active transport protein
orf19.5677	*DUR4*	Urea permease
orf19.405	*VCX1*	Vacuolar H^+/Ca^{2+} exchanger
orf19.3344	*VPS17*	Vacuolar sorting protein
orf19.6324	*VID27*	Vacuole import and degradation
orf19.6738	*VAN1*	Vanadate resistance protein
orf19.4621		Weak similarity to pig tubulin-tyrosine ligase

DOI: 10.1371/journal.pgen.0010001.t004

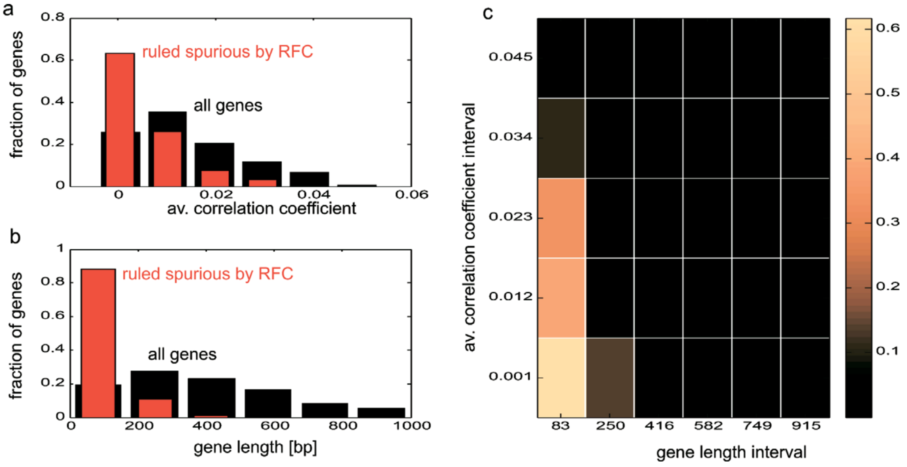


Figure 2. Identification of Spurious Genes

Assessing criteria that identify candidate spurious genes in *S. cerevisiae,* using a reference set of known spurious genes [16].

(A) For every gene in *S. cerevisiae,* the average Pearson correlation coefficient with all other genes was calculated. Shown are histograms of the correlations associated with genes characterized as spurious in the reading frame conservation test ([16]; red) and all genes in the genome (black).

(B) The distribution of gene lengths is shown for genes characterized as spurious (red) and for all genes of the genome (black).

(C) Assessing the likelihood of being spurious as a function of gene length and correlation score. Shown is the proportion of spurious genes out of all genes whose length and correlation score fall into each of the intervals. The proportion is color-coded according to the color bar shown. *S. cerevisiae* genes with an ortholog in *C. albicans* were excluded from the analysis.

DOI: 10.1371/journal.pgen.0010001.g002

of the *C. albicans* genome. *C. albicans* genes with an ortholog in other eukaryotes are assumed to be real and were excluded as candidates (510 of 513 *S. cerevisiae* genes ruled spurious by the reading frame conservation test [16] had no ortholog in *C. albicans*). In the above analysis, approximately 1,000 gene expression experiments were analyzed for *S. cerevisiae* [35], while approximately 200 currently available experiments were analyzed for *C. albicans* (see Materials and Methods). Table S1 includes a ranked list of the 349 *C. albicans* genes that are the most likely to be spurious.

Table 5. Frequency and Characteristics of Short Tandem Repeats in the Coding Sequences of Fungal Genomes

Characteristic	*S. pombe*	*S. cerevisiae*	*C. albicans*	*N. crassa*
Number of STR95s[a] in coding sequence	225	922	2,640	4,721
Number of genes	4,984	5,888	6,354	10,082
Percent of genes with one or more STR95s	4.5%	15.7%	41.5%	46.8%
Number of STR of the indicated periodicity in coding sequences				
1	32	179	457	514
2	2	8	18	15
3	64	445	3,697	6,876
4	6	3	35	74
5	7	1	27	31
6	78	304	1,316	2,140
7	2	3	8	16
8	5	1	7	23
9	52	163	469	924
10	0	8	6	17
11	0	7	1	14
12	6	127	288	376
13	0	0	2	4
14	0	0	2	0
15	0	44	45	122
16	0	0	0	1
17	0	0	0	0
18	0	14	47	51
19	0	0	0	0
20	0	0	0	0
Proportion of Modulo3 repeats	78.7%	83.9%	91.4%	93.7%
Distribution of encoded amino acids sequences in trinucleotide repeats (rank, amino acid, number)	5 A 122	6 A 562	9 A 1,942	4 A 8,198
	C —	19 C 17	19 C 68	19 C 103
	4 D 153	4 D 1,334	5 D 4,599	6 D 6,141
	1 E 276	3 E 1,351	3 E 5,553	2 E 8,824
	16 F 10	18 F 21	17 F 225	17 F 445
	9 G 77	10 G 250	8 G 2,493	3 G 8,205
	12 H 20	11 H 208	11 H 1,109	12 H 1,900
	15 I 15	15 I 71	14 I 589	15 I 626
	3 K 158	7 K 472	10 K 1,837	9 K 4,282
	13 L 18	14 L 87	12 L 771	14 L 1,395
	14 M 16	17 M 53	18 M 181	16 M 523
	8 N 88	2 N 1,369	2 N 6,448	10 N 3,717
	6 P 109	9 P 304	7 P 2,504	7 P 5,085
	7 Q 97	1 Q 1,942	1 Q 7,805	1 Q 9,389
	11 R 34	12 R 113	15 R 486	11 R 2,155
	2 S 207	5 S 1,099	4 S 4,706	5 S 7,881
	10 T 43	8 T 394	6 T 4,518	8 T 5,072
	12 V 20	13 V 89	13 V 615	13 V 1,539
	W —	W —	20 W 33	20 W 82
	17 Y 8	16 Y 60	16 Y 399	18 Y 384
Codon (and encoded amino acid) frequency in trinucleotide repeats >13 bp (number [percentage])				
AAA (K)	2 (1.0)	6 (1.0)	21 (0.7)	3 (0.1)
AAC (N)	2 (1.0)	39 (6.5)	242 (7.8)	194 (5.5)
AAG (K)	12 (6.1)	23 (3.9)	30 (1.0)	136 (3.8)
AAT (N)	6 (3.0)	81 (13.6)	252 (8.1)	4 (0.1)
ACA (T)	2 (1.0)	0 (0.0)	132 (4.3)	61 (1.7)
ACC (T)	0 (0.0)	0 (0.0)	98 (3.2)	159 (4.5)
ACG (T)	0 (0.0)	0 (0.0)	1 (0.0)	38 (1.1)
ACT (T)	4 (2.0)	5 (0.8)	137 (4.4)	19 (0.5)
AGA (R)	1 (0.5)	7 (1.2)	15 (0.5)	6 (0.2)
AGC (S)	1 (0.5)	5 (0.8)	6 (0.2)	86 (2.4)
AGG (R)	0 (0.0)	1 (0.2)	2 (0.1)	20 (0.6)
AGT (S)	1 (0.5)	1 (0.2)	37 (1.2)	14 (0.4)
ATA (I)	0 (0.0)	2 (0.3)	5 (0.2)	1 (0.0)
ATC (I)	0 (0.0)	1 (0.2)	3 (0.1)	8 (0.2)
ATG (M)	0 (0.0)	1 (0.2)	14 (0.5)	8 (0.2)
ATT (I)	1 (0.5)	1 (0.2)	6 (0.2)	0 (0.0)
CAA (Q)	7 (3.6)	71 (11.9)	623 (20.1)	270 (7.6)
CAC (H)	0 (0.0)	2 (0.3)	22 (0.7)	51 (1.4)
CAG (Q)	1 (0.5)	48 (8.1)	54 (1.7)	326 (9.2)
CAT (H)	3 (1.5)	4 (0.7)	37 (1.2)	21 (0.6)
CCA (P)	5 (2.5)	7 (1.2)	109 (3.5)	89 (2.5)
CCC (P)	0 (0.0)	0 (0.0)	2 (0.1)	0 (0.0)
CCG (P)	0 (0.0)	0 (0.0)	0 (0.0)	68 (1.9)
CCT (P)	11 (5.6)	8 (1.3)	14 (0.5)	79 (2.2)
CGA (R)	0 (0.0)	0 (0.0)	3 (0.1)	3 (0.1)
CGC (R)	0 (0.0)	0 (0.0)	0 (0.0)	8 (0.2)
CGG (R)	0 (0.0)	0 (0.0)	0 (0.0)	2 (0.1)
CGT (R)	3 (1.5)	0 (0.0)	2 (0.1)	4 (0.1)
CTA (L)	0 (0.0)	0 (0.0)	5 (0.2)	1 (0.0)
CTC (L)	0 (0.0)	0 (0.0)	0 (0.0)	21 (0.6)
CTG (L/S)	0 (0.0)	0 (0.0)	7 (0.2)	13 (0.4)
CTT (L)	0 (0.0)	0 (0.0)	4 (0.1)	0 (0.0)
GAA (E)	48 (24.4)	97 (16.3)	371 (12.0)	125 (3.5)
GAC (D)	1 (0.5)	18 (3.0)	30 (1.0)	113 (3.2)
GAG (E)	4 (2.0)	8 (1.3)	29 (0.9)	257 (7.2)
GAT (D)	25 (12.7)	67 (11.2)	235 (7.6)	103 (2.9)
GCA (A)	4 (2.0)	7 (1.2)	25 (0.8)	62 (1.7)
GCC (A)	0 (0.0)	0 (0.0)	4 (0.1)	119 (3.4)
GCG (A)	0 (0.0)	0 (0.0)	1 (0.0)	57 (1.6)
GCT (A)	19 (9.6)	14 (2.3)	70 (2.3)	167 (4.7)
GGA (G)	0 (0.0)	0 (0.0)	18 (0.6)	131 (3.7)
GGC (G)	0 (0.0)	0 (0.0)	1 (0.0)	110 (3.1)
GGG (G)	0 (0.0)	0 (0.0)	2 (0.1)	2 (0.1)
GGT (G)	3 (1.5)	14 (2.3)	124 (4.0)	196 (5.5)
GTA (V)	0 (0.0)	0 (0.0)	2 (0.1)	0 (0.0)
GTC (V)	0 (0.0)	0 (0.0)	0 (0.0)	7 (0.2)
GTG (V)	0 (0.0)	0 (0.0)	5 (0.2)	17 (0.5)
GTT (V)	1 (0.5)	5 (0.8)	13 (0.4)	4 (0.1)
TAA (*)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
TAC (Y)	0 (0.0)	1 (0.2)	3 (0.1)	1 (0.0)
TAG (*)	0 (0.0)	0 (0.0)	1 (0.0)	0 (0.0)
TAT (Y)	2 (1.0)	0 (0.0)	7 (0.2)	0 (0.0)
TCA (S)	1 (0.5)	12 (2.0)	122 (3.9)	56 (1.6)
TCC (S)	4 (2.0)	8 (1.3)	10 (0.3)	152 (4.3)
TCG (S)	2 (1.0)	2 (0.3)	8 (0.3)	27 (0.8)
TCT (S)	16 (8.1)	19 (3.2)	77 (2.5)	110 (3.1)
TGA (*)	0 (0.0)	0 (0.0)	2 (0.1)	0 (0.0)
TGC (C)	0 (0.0)	0 (0.0)	2 (0.1)	0 (0.0)
TGG (C)	0 (0.0)	0 (0.0)	4 (0.1)	3 (0.1)
TGT (C)	1 (0.5)	1 (0.2)	0 (0.0)	0 (0.0)
TTA (L)	0 (0.0)	5 (0.8)	16 (0.5)	0 (0.0)
TTC (F)	0 (0.0)	1 (0.2)	3 (0.1)	4 (0.1)
TTG (L)	0 (0.0)	0 (0.0)	9 (0.3)	6 (0.2)
TTT (F)	4 (2.0)	4 (0.7)	26 (0.8)	5 (0.1)

[a] STRs with a less than 5% chance of being random

DOI: 10.1371/journal.pgen.0010001.t005

Multigene Families

Many putative and demonstrated virulence factors of *C. albicans* are members of large multigene families. Well-known examples of such families encode secreted aspartyl proteinases [36,37]), agglutinins [38], secreted lipases [39], high-affinity iron transporters [40], and ferric reductases [41]. Members of each of these families are differentially expressed as a function of the yeast–hyphae transition, phenotypic switching, or timing during experimental infection. Also, each of these families is large relative to the corresponding homolog or family of homologs in *S. cerevisiae,* leading to the concept that expansion of many *C. albicans* gene families may be an adaptation to a commensal lifestyle and may be, in part, responsible for *C. albicans*'s unusual ability to occupy a variety of host niches.

The sequencing of the genome provides an opportunity to survey the global occurrence and extent of multigene families as a first step in assessing their contribution to colonization and disease. We devised a purely computational method to define a comprehensive list of multigene families using NCBI-BLAST and custom Perl scripts. Each translated ORF in the annotated ORF set was compared to every other ORF in the set; if an ORF pair's BLAST alignment had an expectation value less than $1e^{-30}$ and a length greater than 60% of the length of the longer of the two ORFs, then the two ORFs were considered to be members of the same family. A transitive closure rule was applied to ensure that each ORF had membership in one and only one family. In all, 23% of the ORFs were members of families, a percentage comparable to that seen in other eukaryotes [18]. The approach yielded 451 families, with an average of 3.27 members each; 13 of the families have ten or more members, while the largest family has 39 members, consisting of proteins with possible leucine-rich repeat domains.

A striking difference between *C. albicans* and *S. cerevisiae* is the manner in which they acquire nutrients from the environment. In addition to the well-described secreted aspartyl proteinases, lipases, and high-affinity iron transporters, *C. albicans* possesses expanded families of acid sphingomyelinases (with four genes per haploid genome), phospholipases B (six genes), oligopeptide transporters (seven genes), and amino acid permeases (23–24 genes). Another striking difference is the emphasis by *C. albicans* on respiratory catabolism, as reflected in expanded families of peroxisomal enzymes. These include families of acyl-CoA oxidases (three genes), 3-ketoacyl-CoA thiolases (four genes), acyl-CoA thioesterases (three or four genes), fatty acid–CoA synthases (five genes), and glutathione peroxidases (four genes).

Additional families that may pertain to colonization or pathogenesis include those encoding the estrogen-binding protein OYE1 (seven genes), the fluconazole-resistance transporter FLU1 (13 genes), and the vacuolar protein PEP3/VPS16 (four genes), whose *Aspergillus* homolog is required for nuclear migration and polarized growth.

The ATP-binding cassette transporter superfamily. The ATP-binding cassette (ABC) protein superfamily represents one of the largest protein families known to date among available genome sequences. These proteins share similar molecular architecture with the presence of at least one conserved ABC domain and the presence of membrane-spanning segments (transmembrane segments [TMSs]). The ABC domain typically contains Walker A and Walker B motifs and an ABC signature motif. The ABC domain and TMSs can be arranged in a duplicated forward $(TMS_6\text{-}ABC)_2$ or reverse $(ABC\text{-}TMS_6)_2$ topology, however "half size" ABC proteins also exist. As indicated in Table 6, the *C. albicans* genome contains at least 27 genes with ABC domains that include these topologies. These genes have been categorized, according to a classification established in *S. cerevisiae,* into six subfamilies (the MDR, PDR, MRP/CFTR, ALD, YEF3, and RLI subfamilies) [42]. The MDR, PDR, MRP/CFTR, and ALD subfamilies likely all encode transporter proteins, while the other subfamilies, YEF3 and RLI, generally lack TMSs and are considered as non-transporter ABC proteins. The *C. albicans* ABC proteins fall neatly into the categories developed for *S. cerevisiae,* and they are also present in approximately the same numbers (with the exception of the MRP/CFTR subfamily; see below). The predicted topology of each protein detailed in Table 6 is also largely comparable between the two yeast species. Among the 27 ABC proteins so far identified in *C. albicans,* the functions of only nine have been previously characterized. The largest group of known ABC transporters belongs to the *CDR* gene family, among which are *CDR1* and *CDR2,* two genes upregulated in azole-resistant clinical isolates that function in multidrug resistance [43–45]. *CDR3* and *CDR4* have been shown to function as phospholipid flippases and their expression is controlled by the white–opaque switching system [46,47]. Four MRP/CFTR-like transporters are present in *C. albicans,* and among them three show the NH_2-terminal extension with additional transmembrane segments that is typical for many MRP-like transporters (see Table 6). For unknown reasons, homologs of additional members of this family, such as the *S. cerevisiae* genes *ScYBT1, ScNFT1,* and *ScVMR1,* are lacking in *C. albicans* [42,48]. Interestingly, the vacuolar MRP-like transporter encoded by *MLT1* has been implicated in virulence [49]. Since MRP/CFTR transporters are often involved in detoxification of heavy metals or xenobiotics, the presence or absence of discontinuous alleles of some ABC transporter genes (e.g., orf19.6383) may indicate strain differences in ABC transporter function and resulting susceptibility to environmental stresses. Most of the ABC transporter genes listed in Table 6 were given names through their closest homologs in *S. cerevisiae;* however, the functional assignments of these genes awaits further investigation.

The ALS family. The *ALS* genes encode large cell-surface glycoproteins that function in host–pathogen interactions [50,51]. The *ALS* genes are composed of three domains: a 5′ domain that is approximately 1,300 bp in length and relatively conserved in sequence across the family, a central domain composed entirely of tandemly repeated copies of a 108-bp sequence, and a 3′ domain of variable length and sequence that encodes a serine/threonine-rich portion of the protein [50]. Efforts to characterize the *ALS* genes started independently of the *C. albicans* genome project [38,52–54]) and were aided greatly by information that emerged as the genome sequencing effort progressed [55–57]. Table 7 lists the current ORFs that correspond to genes in the *ALS* family. The *ALS* family includes eight different genes [55], each with an extensive degree of allelic variability, sometimes within a given strain (Table 7) or across the wider population of *C. albicans* isolates [58–60]. Because of sequence assembly difficulties, mainly attributable to the length and repetitive nature of sequences within the *ALS* central domain, only three of *ALS* ORFs in this project are in agreement with *ALS* gene sequences derived independently of the genome project and reported in the literature (Table 7). The annotation effort described here did not edit the underlying assembly 19 sequence. However, gap sequencing that is presently being

Table 6. Genes Encoding Members of the ABC Transporter Family

ORF	Subfamily[a]	Topology	Contig	Length (Base Pairs)	Suggested or Published Name[b]	Product/Note	References
orf19.6000	PDR	(ABC-TMS$_6$)$_2$	Contig19–10236	1,502	*CDR1*	ABC transporter, multidrug resistance protein	43
orf19.5958	PDR	(ABC-TMS$_6$)$_2$	Contig19–10236	1,500	*CDR2*	ABC transporter, multidrug resistance protein	45
orf19.1313	PDR	(ABC-TMS$_6$)$_2$	Contig19–10109	1,501	*CDR3*	ABC transporter, opaque-specific, merged with orf19.1312	41,130
orf19.5079	PDR	(ABC-TMS$_6$)$_2$	Contig19–10218	1,491	*CDR4*	ABC transporter, white-specific	41,131
orf19.918	PDR	(ABC-TMS$_6$)$_2$	Contig19–10079	1,513	*CDR5*	ABC transporter, merged with orf19.919	—
orf19.5759	PDR	(ABC-TMS$_6$)$_2$	Contig19–10233	1,496	*SNQ2*	ABC transporter	—
orf19.3120	PDR	ABC-TMS$_6$	Contig19–10166	580	orf19.3120	Possible half-size ABC transporter, best similarity with YOL075c	—
orf19.4531	PDR	TMS$_7$-ABC-TMS$_6$-ABC	Contig19–10209	1,275	orf19.4531	ABC transporter, best similarity withYOL075c	—
orf19.459	PDR	TMS$_2$-ABC-TMS$_7$	Contig19–10052	1,039	*ADP1*	ABC transporter	—
orf19.1077	MDR	TMS$_6$-ABC	Contig19–10090	751	*ATM1*	ABC transporter, putative half-size mitochondrial transporter	—
orf19.2615	MDR	TMS$_6$-ABC	Contig19–10151	685	*MDL1*	ABC transporter, putative half-size mitochondrial transporter	—
orf19.13043	MDR	TMS$_6$-ABC	Contig19–20230	783	*MDL2*	ABC transporter, putative half-size mitochondrial transporter, fragmented allele orf19.5599–orf19.5600	—
orf19.7440	MDR	(TMS$_6$-ABC)$_2$	Contig19–2514	1,324	*HST6*	ABC transporter, putative pheromone transporter	132
orf19.1783	MRP/CFTR	(TMS$_6$-ABC)$_2$	Contig19–10125	1,487	*YOR1*	ABC transporter, closely related to *S. cerevisiae* YOR1 transporter, merged with orf19.1784, continuous allele sequence not confirmed	133
orf19.5100	MRP/CFTR	TMS$_5$-(TMS$_6$-ABC)$_2$	Contig19–10218	1,606	*MLT1*	Vacuolar ABC transporter, best similarity with ScBPT1	49
orf19.6383	MRP/CFTR	TMS$_5$-(TMS$_6$-ABC)$_2$	Contig19–10247	1,490	orf19.6383	ABC transporter, merged with orf19.6382, continuous allele sequence confirmed in some strains, best similarity with ScBPT1	G. Köhler, unpublished
orf19.6478	MRP/CFTR	TMS$_5$-(TMS$_6$-ABC)$_2$	Contig19–10248	1,581	*YCF1*	ABC transporter, closely related to *S. cerevisiae* YCF1 transporter	—
orf19.7500	ALD	TMS$_6$-ABC	Contig19–2516	769	*PXA1*	ABC transporter, putative half-size peroxisomal transporter	—
orf19.5255	ALD	TMS$_6$-ABC	Contig19–10223	668	*PXA2*	ABC transporter, putative half-size peroxisomal transporter	—
orf19.2183	YEF3	ABC$_2$	Contig19–10141	610	*KRE30*	Non-transporter ABC protein, best similarity with YER036c	—
orf19.4152	YEF3	ABC$_2$	Contig19–10198	1,051	*CEF3*	Non-transporter ABC protein, translation elongation factor 3	134
orf19.6060	YEF3	ABC$_2$	Contig19–10237	752	*GCN20*	Non-transporter ABC protein, best similarity with YFR009w	—
orf19.7332	YEF3	ABC$_2$	Contig19–2511	1,196	*ELF1*	Non-transporter ABC protein, elongation-like factor	135
orf19.3034	RLI	ABC$_2$	Contig19–10163	623	*RLI1*	Non-transporter ABC protein, best similarity with YDR091c	—
orf19.388	Others	ABC	Contig19–10051	321	*CAF16*	Non-transporter ABC protein, best similarity with YFL028c	—
orf19.5029	Others	ABC	Contig19–10216	546	orf19.5029	Non-transporter ABC protein, best similarity with YDR061w	—

[a] Subfamily nomenclature as proposed by Bauer et al. [42].
[b] Published names are underlined.
DOI: 10.1371/journal.pgen.0010001.t006

carried out and the production of a final genome assembly will correct these errors. Published *ALS* gene sequences can be found on the CGD Web site.

Assembly of the *C. albicans* genome sequence revealed the contiguous positions of *ALS5, ALS1,* and *ALS9* on Chromosome 6, which was verified by independent studies [57]. Additional testing revealed that, in SC5314, the large alleles of *ALS5, ALS1,* and *ALS9* occupy the same chromosome while the small alleles of each gene are found on the homologous chromosome [57]. But allelic variability and arrangement on homologous chromosomes will vary for each *C. albicans* strain. Allelic variation can be extreme for *ALS* genes, and is most

Table 7. Assembly 19 ORFs That Correspond to ALS Genes

Systematic ID	Gene Name	Number of Tandem Repeat Copies per Allele	Method for Determining Repeat Copy Number	Reference	Comments
orf19.5741	*ALS1*	8, 20	Southern blot	52	This ORF has about ten copies of the tandem repeat sequence. This number is too large for the small *ALS1* allele in strain SC5314 and far too small for the large *ALS1* allele. Some sequences from within the tandem repeat domain must have been assembled incorrectly.
orf19.2355	*ALS2*	31, 36	Southern blot	L. L. Hoyer, unpublished	This ORF encodes about 24 copies of the tandem repeat sequence, making it too short to be either allele from strain SC5314. However, the remainder of the coding region is intact and correctly assigned as *ALS2*.
orf19.1816	*ALS3*	9, 12	DNA sequencing	57	This ORF is missing the 3′ end of the coding region. The ORF has almost 12 copies of the tandem repeat sequence and corresponds to the large *ALS3* allele in strain SC5314. Its name should be *ALS3–1*. DNA sequences of both *ALS3* alleles were derived outside of the genome sequencing effort and are available in GenBank (*ALS3–1* accession number AY223552; *ALS3–2* accession number AY223551).
orf19.1097	*ALS4*	18, 36	Southern blot	L. L. Hoyer, unpublished	This sequence contains about 48 copies of the tandem repeat sequence. *ALS4* allele sizes in strain SC5314 are estimated by Southern blot to have 18 or 36 copies of the repeated sequence. The ORF is missing the 3′ end of the gene. It is possible that this ORF has the 5′ end of *ALS4* fused to the 3′ end of *ALS2*.
orf19.2121	*ALS4*	18, 36	Southern blot	L. L. Hoyer, unpublished	This ORF starts at amino acid 252 of the *ALS4* sequence and contains about 20 copies of the tandem repeat sequence. Since the Southern blotting method, by which *ALS4* alleles sizes were judged, has some error, it is possible that this sequence is the smaller SC5314 allele, *ALS4–2*. There is a stop codon in the middle of the sequence due to a frameshift that reads HHL* and then resumes with the correct APSTET sequence. A DNA sequence for *ALS4–2* is available in GenBank (accession number AF272027) and includes 20 tandem repeat copies.
orf19.4555	*ALS4*	18, 36	Southern blot	L. L. Hoyer, unpublished	This ORF encodes about 36 copies of the tandem repeat sequence, which is the correct number for the larger *ALS4* allele *(ALS4–1)* from strain SC5314. The ORF has a frame shift within the tandem repeat domain that prematurely truncates a repeat copy and adds ETSKLHGYHN*. The reading frame then resumes with another repeat copy, but in the middle of the consensus sequence.
orf19.5736	*ALS5*	4, 5	PCR/acrylamide gel; DNA sequencing of both alleles	L. L. Hoyer, unpublished	This ORF contains about four copies of the tandem repeat sequence, and represents the small *ALS5* allele *(ALS5–2)* from strain SC5314. *ALS5–2* is found on the same chromosome as the short allele of *ALS1 (ALS1–2)* and the short allele of *ALS9 (ALS9–2)*. The sequence of *ALS5–2* was derived independently of the genome project and is deposited in GenBank (accession number AY227439). The sequence of the large *ALS5* allele (*ALS5–1*) has GenBank accession number AY227440.
orf19.7414	*ALS6*	4, 4	PCR/acrylamide gel; DNA sequence of one allele	L. L. Hoyer, unpublished	The corresponding sequence from strain SC5314 is GenBank accession number AY225310, which was derived independently of the genome sequencing project. Both *ALS6* alleles in strain SC5314 have the same number of tandem repeat copies.
orf19.7400	*ALS7*	15, 15	PCR and genome sequence	59	This ORF has about 15 copies of the tandem repeat sequence, which is correct for both alleles from strain SC5314.
orf19.5742	*ALS9–1*	14, 17	DNA sequencing	57	This ORF should be *ALS9–1* as defined by Zhao et al. [57] but has the wrong 5′ end matched with the correct *ALS9–1* 3′ end. The correct sequence for *ALS9–1* from strain SC5314 is GenBank accession number AY269423. *ALS9–1* should be on the same chromosome copy as *ALS1–1* and *ALS5–1*. Separate alleles of *ALS1* and *ALS5* were not maintained in the genome assembly process.
orf19.45	*ALS9–2*	14, 17	DNA sequencing	57	The closest match to this ORF is *ALS9–2,* but this ORF is only a partial sequence. It is likely that the rest of *ALS9–2* was collapsed into *ALS9–1* and this sequence is left since it does not directly match *ALS9–1*. The *ALS9–2* sequence should be on the same chromosome copy as *ALS1–2* (closest match is orf19–5741) and *ALS5–2* (orf19.5736). The correct sequence for *ALS9–2* in SC5314 is GenBank accession number AY269422.
orf19.79	Unknown			50	This ORF is composed entirely of tandem repeat sequences that come from *ALS1, ALS2, ALS3,* or *ALS4.* These four genes have tandem repeat sequences that cross-hybridize by Southern blotting and comprise one subfamily of the ALS genes.

DOI: 10.1371/journal.pgen.0010001.t007

commonly associated with the tandem repeat domain, although it is also present within other domains of the coding region [56,57,59]. Presenting the sequence of a single *ALS* allele, as done in these annotation data, loses the sense of allelic diversity that can have a significant effect on evaluation of ALS protein function. For example, testing the two *ALS3* alleles from strain SC5314 in a common adhesion assay format showed that the allele with more tandem repeat copies produced a protein with greater adhesive capability than the smaller allele [60]. Table 7 notes GenBank entries for *ALS* alleles from strain SC5314 that aid understanding of allelic diversity for the various *ALS* genes.

The *MEP* family. Members of the *MEP* gene family encode ammonium permeases and, along with the *OPT* family described below, feature prominently in our list of fungal-specific genes. They thus represent potentially interesting targets for the development of antifungal drugs. Experimental evidence suggests that *MEP1* and *MEP2* encode the only specific ammonium permeases in *C. albicans*, since Δ*mep1* Δ*mep2* double mutants exhibited no detectable ammonium uptake and were unable to grow at ammonium concentrations below 5 mM [61], a phenotype that is similar to that of *S. cerevisiae* mutants deleted for all three ammonium permeases [62,63]. The third *C. albicans* gene, represented by orf19.4446, encodes a protein with much lower similarity to the other ammonium permeases of *C. albicans* and *S. cerevisiae* (approximately 44% to all proteins) but might encode an ammonium permease that is not expressed under the growth conditions used in these assays.

In addition to its role in ammonium transport, Mep2p also controls nitrogen-starvation-induced filamentous growth of *C. albicans*. Mutants in which only the *MEP2* gene was deleted grew as well as the wild-type strain at low ammonium concentrations but failed to filament under these conditions. This role of *MEP2* in filamentous growth of *C. albicans* at low ammonium concentrations is similar to the function of its counterpart *ScMEP2* in pseudohyphal growth of *S. cerevisiae* under limiting ammonium conditions [63]. However, in contrast to the latter, *MEP2* seems to have a much broader role in filamentous growth of *C. albicans* since Δ*mep2* mutants also had a filamentation defect when amino acids or urea instead of ammonium served as the limiting nitrogen source (J. Morschhäuser, personal communication).

The *OPT* family. Oligopeptide transporters represent another group of fungal-specific surface proteins that transport peptides of four or five amino acids in length into the cell and together with the di- and tripetide transporters allow growth when peptides are the only available nitrogen source. This is presumably the position of *C. albicans* cells when they have invaded host tissues and are secreting their battery of peptidases and other catabolic enzymes. The founding member of the oligopeptide transporter gene family was *OPT1* from *C. albicans* [64]. Analysis of the *C. albicans* genome sequence as well as cloning of the corresponding genes demonstrated that *C. albicans* in fact possesses a large gene family encoding putative oligopeptide transporters. The *OPT* genes were annotated according to their decreasing similarity to *OPT1*. The *OPT2, OPT3,* and *OPT4* genes are highly similar to each other. The similarity of the remaining members of the family then drops considerably, but we have detected genes now named *OPT6, OPT7,* and *OPT8*. Deletion of the *OPT1* alleles in the *C. albicans* wild-type strain SC5314 resulted in increased resistance of the mutants to a toxic tetrapeptide, providing experimental evidence that Opt1p indeed functions as an oligopeptide transporter in *C. albicans* [65]. Preliminary observations indicate that at least the *OPT2* to *OPT5* genes also encode functional oligopeptide transporters (O. Reuß and J. Morschhäuser, unpublished data).

Zinc cluster transcription factors. Proteins of the zinc finger superfamily represent one of the largest classes of DNA-binding proteins in eukaryotes. Several different classes of zinc finger domains exist that differ in the arrangement of their zinc-binding residues [66]. One of these domains, which appears to be restricted to fungi, consists of the $Zn(II)_2Cys_6$ binuclear cluster motif in which six cysteines coordinate two zinc atoms [67,68]. *S. cerevisiae* possesses 54 zinc cluster factors defined by the presence of the zinc cluster signature motif $CX_2CX_6CX_{5-16}CX_2CX_{6-8}C$, which is generally located at the N-terminus of the protein. These proteins function as transcriptional regulators involved in various cellular processes including primary and secondary metabolism (e.g., Gal4p, Ppr1p, Hap1p, Cha4p, Leu3p, Lys14p, and Cat8p), pleiotropic drug resistance (e.g., Pdr1p, Pdr3p, and Yrr1p), and meiosis (Ume6p) [68,69]. Quite often, they bind as homo- or heterodimers to two CGG triplets organized as direct, indirect, or inverted repeats and separated by sequences of variable length [68,70]. A large proportion of these factors (50%) also contain a middle homology region (Fungal_*trans* in the Pfam Protein Families Database) located in the central portion of the protein that has been proposed to participate in DNA binding and to assist in DNA target discrimination [67].

Analysis of the *C. albicans* proteome using a combination of sequence analyses tools (SMART, Pfam, and PHI-BLAST) allowed us to identify 77 binuclear cluster proteins. These factors are characterized by the presence of the zinc cluster signature motif $CX_2CX_6CX_{5-24}CX_2CX_{6-9}C$ generally located at the N-terminus of the protein (72 out of 77) and with a spacing between cysteines 3–4 and 5–6 slightly different from the *S. cerevisiae* motif. As observed in *S. cerevisiae,* a large proportion of the *C. albicans* factors also contain a middle homology region (29 out of 77). To our knowledge, only six of the *C. albicans* zinc cluster genes have been characterized in detail, including *SUC1,* involved in sucrose utilization [71], *FCR1,* implicated in pleiotropic drug resistance [72], *CWT1,* required for cell wall integrity [73], and *CZF1, FGR17,* and *FGR27,* involved in filamentous growth [9,74]. The functions of many uncharacterized *C. albicans* zinc cluster factors (approximately 20%) can be inferred from the fact that they display high levels of sequence similarity (top BLASTP *e*-value $\leq 1e^{-20}$) with the products of *S. cerevisiae* genes with a known function. In the case of *GAL4,* however, the *C. albicans* homologous ORF identified (orf19.5338) encodes a significantly smaller protein (261 aa) than *S. cerevisiae* Gal4p (881 aa), lacking the C-terminal two-thirds of the protein that contains one of two transcriptional activating domains, and must therefore have a somewhat different function. Approximately half of the *C. albicans* zinc cluster genes do not appear to have homologs in *S. cerevisiae* (using a BLAST cutoff of $< 1e^{-20}$) and are therefore likely to participate in processes specific to *C. albicans*. Finally, it is noteworthy that many of the zinc cluster factors known to be involved in pleiotropic drug resistance in *S. cerevisiae,* such as Pdr1p, Pdr3p, Yrr1p, Yrm1p, Rds1p, and Rdr1p, do not appear to possess close structural homologs in *C. albicans*. Since pleiotropic drug resistance is frequently observed in *C. albicans,* it is likely that this organism possesses functional homologs of these genes or other novel processes that remain to be identified.

Lipid and Amino Acid Metabolism

Some of the *C. albicans* ORFs that do not have clear homologs in *S. cerevisiae* but do have homologs in other fungi, bacteria, and/or vertebrates encode catabolic enzymes, oxidoreductases, and proteins involved in environmental sensing pathways. The list of genes that *C. albicans* does not share with *S. cerevisiae* is skewed towards enzymes involved in the catabolism of fatty acids and ketone bodies in the peroxisome. There are also numerous oxidoreductases, some of which may be involved in activating hydrophobic organic compounds as a prelude to their oxidative degradation. This metabolic arrangement may reflect, in part, the state of the common ancestor with *S. cerevisiae,* as also reflected in *Yarrowia lipolytica, C. antartica, C. rugosa, C. tropicalis, C. maltosa,* and *C. deformans,* which are model organisms in the study of lipases and alkane oxidation for industrial purposes. It is worth mentioning, however, that the genus *Candida* arose originally to identify fungi that were unclassifiable, asexual, and ascomycetous—properties that appear to correlate with parasitism and the presence of catabolic gene families, such as lipases and alkane-assimilating cytochrome P-450 enzymes. Beta-oxidation in fungi is predominantly peroxisomal, and the number of enzymes participating in the process is greater in *C. albicans* than in *S. cerevisiae*. *C. albicans* also encodes a related ethanolamine kinase (orf19.6912), a malonyl-CoA acyl carrier protein acyltransferase (MCT1), and an enoyl-CoA hydratase (orf19.6830) not found in *S. cerevisiae*. Further supplying substrates for oxidation are several enzymes encoded by *C. albicans* that participate in the degradation of asparagine (asparaginase; orf19.3791), cysteine (cysteine dioxygenase [CDG1] and cysteine sulfinate decarboxylase [orf19.5393]), valine (3-hydroxyisobutyrate dehydrogenase [orf19.5565]), and arginine (orf19.3498). Other catabolic enzymes come as a surprise in that they may relate to the scavenging of unsuspected carbon sources. *C. albicans* encodes three D-amino acid oxidases (IFG3, DAO1, and DAO2) whose substrates might be derived from bacterial cell walls, various oxidoreductases whose substrates are likely to be aromatic and aliphatic compounds not used by the host, a pathway consistent with omega oxidation of fatty acids (which would convert alkanes into alpha-omega diols, fatty acids, and dicarboxylic acids), and a benzene desulfurase (orf19.3901).

Acetyl-CoA generated in the peroxisome is transferred to the mitochondrion, where the most notable difference from *S. cerevisiae* is the presence of a respiratory Complex I, which can now largely be reconstructed based on sequence similarity to components found in other organisms. The importance of Complex I in the biology of *C. albicans* is inferred from the observation that deletion of one of its subunits results in a defect in filamentation [75] and the observation that subunit 49 is essential for vegetative growth [8]. An additional difference is the presence of two alternative oxidases that may be involved in protection against oxidative stress [76]. Thus, it is not yet clear whether the omnivorous catabolic capacity of *C. albicans* reflects its heritage and role as a fungal saprophyte aiding organic decomposition, or whether these capacities have been elaborated and tuned in response to the specific problem of consuming mammalian host cells.

Phospholipases. Depending on the site of attack, phospholipases are classified as phospholipase A, B, C, or D. Phospholipase A enzymes hydrolyze the 1-acyl ester (PLA_1) or the 2-acyl ester (PLA_2) of phospholipids. In fungi, phospholipase B enzymes hydrolyze both acyl groups and often also have lysophospholipase activity, removing the remaining acyl moiety on lysophospholipids [77]. Phospholipase C and phospholipase D enzymes are phosphodiesterases that cleave the glycerophosphate bond and remove the base group of phospholipids, respectively. While a major role of phospholipase function is membrane homeostasis, additional functions comprise nutrient digestion and generation of signaling molecules. Some phospholipases are toxins or components of venoms. Bacterial phospholipases have been shown to be involved in pathogenesis by promoting hemolysis, cytolysis, and tissue destruction, as well as interfering with host signal transduction [78].

As indicated in Table 8, the largest and best-characterized group of phospholipases in *C. albicans* is the five-member phospholipase B gene family. A related gene family is present in *S. cerevisiae,* albeit with three members, again reflecting the general increase in gene numbers for enzymes involved in lipid metabolism in *C. albicans*. All PLB proteins harbor NH_2-terminal signal peptides for secretion; Plb3p, Plb4p, and Plb5p additionally contain hydrophobic COOH termini with putative GPI anchor attachment sites for localization to the plasma membrane or further processing for tethering to the cell wall [79,80]. To date, *PLB1* and *PLB2* are the best-characterized members of the gene family [81–84]. Inactivation of *PLB1* [82,83] and *PLB5* (S. Theiss, G. Ishdorj, M. Kretschman, C. Y. Lan, T. Nichterlein, et al., unpublished data) reduced virulence in animal models.

Putative PLC and PLD phosphodiesterases are also represented in the *C. albicans* genome. Orf19.6629 is a likely homolog to *S. cerevisiae ScISC1,* which encodes a PLC with neutral sphingomyelinase activity. Besides the recently published *PLC1* gene [85], two almost identical genes encode phosphatidylinositol phospholipase C proteins (PI-PLC). The latter lack homologs in *S. cerevisiae,* but are similar to bacterial PI-PLCs. *PLD1* was shown to be involved in the morphological transition from yeast to hyphae and required for full virulence in animal models [86]. Interestingly, the *PLD1* gene product and another phospholipase-like protein (encoded by orf19.4151) show significant sequence similarity to *S. cerevisiae* proteins that are involved in meiosis and sporulation (ScSpo1p, ScSpo14p, and ScSpo22p). As already shown for *PLD1,* the *ScSPO1* homolog, the functional roles of these proteins are likely to differ from their counterparts in *S. cerevisiae* since *C. albicans* has not been shown to undergo meiosis [10].

Another intriguing group of phospholipase genes in *C. albicans* are patatin-like phospholipases encoded by orf19.1504, orf19.5426, and orf19.6396. These proteins might account for phospholipase A activities in *C. albicans* that could be involved in intracellular storage or mobilization of lipids.

Sphingolipid metabolism. *C. albicans* also displays differences from *S. cerevisiae* with respect to sphingolipid metabolism. Pathways leading to and from fungal-type sphingomyelins have been studied extensively in *S. cerevisiae,* where by-products mediate many important structural and signaling functions that affect cell proliferation, the definition of cell membrane domains and polarity, apoptosis, and stress responses [87,88]. Many of the associated enzymes are essential and are targets of fungal toxins, and thus are candidates for anti-fungal drug development [88]. *C. albicans* shares the same fundamental pathways in sphingolipid biosynthesis/degradation plus four additional enzymes. Two of these, a glucosyl transferase (CGT1) and a delta-4

Table 8. Phospholipases in *C. albicans*

ORF	Domain (Pfam)	Contig	Length (Base Pairs)	Suggested Name/ Published Name[a]	Product/Note	References
orf19.689	Lysophospholipase catalytic domain	Contig19_10064	605	*PLB1*	Phospholipase B, gene in tandem with *PLB2*	81–83
orf19.690	Lysophospholipase catalytic domain	Contig19_10064	609	*PLB2*	Phospholipase B, gene in tandem with *PLB1*	84
orf19.1442	Lysophospholipase catalytic domain	Contig19_10119	702	*PLB3*	Phospholipase B, merged with orf19.1443, continuous allele sequence confirmed	G. Köhler, unpublished
orf19.6594	Lysophospholipase catalytic domain	Contig19_2449	632	*PLB4*	Phospholipase B	
orf19.5102	Lysophospholipase catalytic domain	Contig19_10218	754	*PLB5*	Phospholipase B	G. Köhler, unpublished
orf19.4151	Lysophospholipase catalytic domain	Contig19_10198	684	*SPO1*	Similar to *S. cerevisiae SPO1* product, a meiosis-specific protein with similarity to phospholipase B	
orf19.5506	Phosphatidylinositol-specific phospholipase C, X and Y domain	Contig19_2335	1,100	*PLC1*	Phospholipase C	85
orf19.5797	Phosphatidylinositol-specific phospholipase C, X domain	Contig19_10234	295	*PLC2/PI-PLC*	Phosphatidylinositol phospholipase C, closely related to bacterial PI-PLCs, highly similar to orf19.1586 protein	136
orf19.1586	Phosphatidylinositol-specific phospholipase C, X domain	Contig19_10119	295	*PLC3/PI-PLC*	Phosphatidylinositol phospholipase C, closely related to bacterial PI-PLCs, highly similar to orf19.5797 protein	
orf19.6629	Endonuclease/exonuclease/ phosphatase family	Contig19_10251	438	*ISC1*	Similar to *S. cerevisiae* YER019w product, an inositol phosphosphingolipid phospholipase C	
orf19.1161	Phospholipase D	Contig19_10097	1,710	*PLD1*	Phospholipase D, similar to *S. cerevisiae* SPO14 product which is required for meiosis and spore formation	86,137
orf19.1504	Patatin-like phospholipase	Contig19_10119	853	orf19.1504	Potential patatin-like phospholipase similar to *Sc. pombe* SPAC31G5.20c and to *S. cerevisiae* YOR081c	
orf19.5426	Patatin-like phospholipase	Contig19_10227	949	orf19.5426	Potential patatin-like phospholipase similar to *S. cerevisiae* YKR089c	
orf19.6396	Patatin-like phospholipase	Contig19_10247	1,386	orf19.6396	Potential patatin-like phospholipase with cyclic nucleotide binding domain, similar to *S. cerevisiae* YML059c	
orf19.13603		Contig19_20241	835	*SPO22*	Similar to *S. cerevisiae SPO22* product, a meiosis-specific phospholipase A2 homolog	

[a]Published names are underlined.

DOI: 10.1371/journal.pgen.0010001.t008

sphingolipid desaturase (DES1), have been previously studied. The presence of glycosyl ceramides in *C. albicans* has been known for some time [89,90], and the gene responsible for their synthesis has been cloned and expressed in *Pichia* [91]. The molecules play a common role in differentiation in dimorphic fungi [92]. Homologs of the delta-4 sphingolipid desaturase enzyme include the mouse, human, and *Drosophila* degenerative spermatocyte proteins, which play a role in meiosis [93]; its function in *C. albicans* may relate to membrane structure, or the production of signaling molecules, as is the case in plants. An interesting component of sphingolipid metabolism in *C. albicans* is a sphingomyelin transfer protein (Het1p) similar to the *Podospora anserina* HET-C2 protein. The *P. anserina* protein is involved in self/non-self discrimination, a fungal version of the vertebrate major histocompatibility locus [94,95]. It is possible that the protein is involved in regulating the sphingomyelin composition of *C. albicans* membranes, a factor that may relate to acquisition of resistance to amphotericin B and azoles [96]. Finally, *C. albicans* encodes four acid sphingomyelinases, two of which may be secreted, that have not been studied in fungi. Based on the actions of metazoan secreted acid sphingomyelinases, these enzymes may be involved in regulation of membrane raft formation and generation of ceramide, a second messenger that is known to regulate apoptosis in higher eukaryotes. Secreted sphingomyelinases of pathogenic bacteria, which are enzymatically similar but structurally unrelated to those of *C. albicans,* have been shown to lyse phagosomal membranes [97], facilitate entry into both phagocytic and nonphagocytic cells [98,99], act as hemolysins that abet piracy of iron from the host [100,101], and induce host cell apoptosis [102,103].

Signal Transduction

Differences in signal transduction and regulatory pathways between *C. albicans* and *S. cerevisiae* are numerous. Many of these *C. albicans*–specific genes encode proteins that are responsive to changes in the environment. They may thus be responsive to colonization of a new anatomical site (e.g., passage through the stomach), fluctuations in the availability of nutrients, or the appearance of host inflammatory reactions. Gene products falling into this category include (1) a homolog (TIP120) of a TBP-interacting protein in humans and rats, which acts as global regulator of class I, II, and III genes in response to abrupt changes in ambient conditions [104], (2) a relative (orf 19.1798) of tuberin, a negative regulator of cell growth in response to low cellular energy levels in mammals [105], (3) a conserved group of stomatin-like proteins (orf 19.7296 and SLP2) that may play a role in mechanoreception, (4) a family of pirin homologs that obviously arose from a recent duplication event *(PRN1, PRN2, PRN3,* and *PRN4)*—these are nuclear factors whose homologs interact with the human oncogene Bcl-3 product and with an *A. thaliana* G protein alpha-subunit involved in regulating seed germination and early seedling development [106])—and (5) a rhomboid protein (orf 19.5234), probably located on the plasma membrane, whose homologs in eukaryotes and bacteria mediate the proteolytic release of signaling peptides from a larger precursor [107]. In addition to differences traceable to novel genes, other pathways that share components have doubtless been altered in their role and regulation, such as the mating pathway [108].

Two of the most important enzyme families that are involved in signal transduction pathways are the kinases and small GTPases. The *C. albicans* annotation identifies 96 protein kinases, most of which have strong orthologs in *S. cerevisiae.* The *C. albicans* genome contains two genes encoding GTPases of the heterotrimeric G protein alpha-subunit family—*GPA1* and *GPA2*. In addition, it contains 29 small GTPases of the p21 superfamily. These include a single Ras protein (Ras1p), various members of the Rho and Rab families, the Ran1 homolog Gsp1p, and several members of the ADP ribosylation subfamily. Most of these proteins have clear *S. cerevisiae* orthologs. However, *S. cerevisiae* does not have a Rac homolog, while orf19.6237 appears to encode a *C. albicans* Rac protein and has thus been named *RAC1*. As well, orf19.5902 appears to be distantly related to Ras but lacks any strong equivalent in any organism, and has been designated Rlp1p, for Ras-like protein, while orf19.2975 is a YPT/RAB family member that has been named *RAB7* because it has no clear *S. cerevisiae* YPT ortholog.

Conclusions

We have coordinated a community-wide effort to manually confirm, edit, and annotate 6,354 genes from assembly 19 of the *C. albicans* genome. This annotation includes 214 intron-containing genes, 246 genes with either missense mutations or sequencing errors, and 190 truncated genes that terminate at the ends of the sequence contigs. *C. albicans* genes were found to be exceptionally rich in short sequence repeats, especially compared to the genomes of *S. pombe* and *S. cerevisiae.* Correlation with transcriptional profiling data was used to identify potentially spurious genes. This improved dataset allowed the identification fungal-specific genes and permitted a detailed analysis of several large multigene families. Comparative genomic studies indicate that *C. albicans* is much more versatile in its production of secreted lipid- and amino-acid-degrading enzymes and in its ability to import the resulting nutrients.

Materials and Methods

Identification of *C. albicans* ORFs and merging of preliminary annotations. Nucleotide sequence data for assembly 19 were retrieved from the SGTC Web site (http://www-sequence.stanford.edu//group/candida/). Assembly 19 is composed of a haploid supercontig set (contigs 19–831 to 19–10262), here referred to as the haploid set, and a allelic supercontig set (contigs 19–20001 to 19–20161), here referred to as the allelic set [6].

The CAAT-Box software package [109] was used to identify annotation-relevant ORFs in assembly orf19. A set including ORFs longer than 300 codons and a set with all intergenic regions obtained after subtraction of ORFs larger than 80 codons were created. These sets were used to build a GeneMark matrix [11] that was subsequently used to evaluate the coding probability of all ORFs in assembly 19. ORFs longer than 150 codons, and ORFs longer than 40 codons and with a GeneMark coding function greater than 0.5 [11] over their whole length, were selected and assigned a reference number of the format IPF*n.i* where IPF stands for individual protein file, *n* is an integer specific to the IPF, and *i* corresponds to the number of times the IPF has been modified between assembly 5, 6, and 19 of the *C. albicans* genome sequence. In total, 11,025 and 9,089 IPFs were selected in the haploid and allelic sets, respectively. IPFs shorter than 150 codons in the haploid set were further inspected for (1) overlaps with larger IPFs on a different frame and (2) homology to proteins in the NR database of non-redundant proteins from GenBank. IPFs that overlapped with a larger IPF or did not show a significant homolog (BLASTP *e*-value $< 1e^{-3}$) [110] were designated FALSORF. Of the 11,025 IPFs identified in the haploid set, 3,505 were FALSORFs.

All IPFs identified in the haploid set were compared through reciprocal BLASTP to the set of 7,680 *C. albicans* ORFs defined at the SGTC that uses the systematic designation orf19.*n*. BLASTP results were parsed using Readblast [111]. IPFs without an orf19.*n* counterpart were assigned a new reference number of the format orf19.*n.i,* where orf19.*n* is the closest upstream (using SGTC contig coordinates) ORF defined by the SGTC and *i* is an integer that varies between 1 and the number of IPFs located between orf19.*n* and orf19.(n + 1). For instance, if three ORFs were found between orf19.1234 and orf19.1235, these would be referred to as orf19.1234.1, orf19.1234.2, and orf19.1234.3. Taken together, 11,616 orf19 ORFs were identified in the haploid set, of which 3,936 were not present in the SGTC orf19 set.

A similar procedure was applied to the allelic set of sequences, and a total of 9,552 orf19 ORFs were identified, of which 3,012 were not present in the SGTC orf19 set. The haploid and allelic sets of orf19 ORFs were compared by reciprocal BLASTP in order to define allelic and unique sequences in the allelic set (see below).

The 11,615 orf19 ORFs identified in the haploid set were compared by reciprocal BLASTP to the 9,168 ORFs identified by the SGTC using assembly 6 of the *C. albicans* genome sequence (designated orf6.*n*). A similar reciprocal comparison was run using the set of 6,165 *C. albicans* proteins available in the CandidaDB database that have been defined by applying a procedure similar to that outlined above on assembly 6 and through a manual curation aiming to reach a non-redundant protein set (http://genolist.pasteur.fr/CandidaDB; [12]). Furthermore, orf19 ORFs were reciprocally compared to the *S. cerevisiae* proteome using data available at the SGD [112]. All data were parsed using Readblast [111], and a matrix was generated that correlated ORFs from each dataset.

We used the genome annotation tool Artemis [14], which provides very detailed annotation capability, visually mapping desired features onto the target sequence. In preparation, we loaded our heterogeneous data into the required EMBL-style files: (1) the orf19 reference (field: Assembly__ID); (2) the orf6 reference (field: old__Assembly__ID); (3) the CandidaDB entry number (field: db__xref); (4) the entry number for the Comprehensive Yeast Genome Database, which provides a detailed analysis of protein features of all entries available in CandidaDB [113] (field: db__xref); (5) the GenBank entry number for *C. albicans* proteins previously characterized and annotated (field: db__xref); (6) the IPF reference (field: db__xref); (7) annotation data available from CandidaDB (proposed gene name and proposed function; field: Annotator); (8) annotation data of the Agabian's laboratory based on the orf6 protein set (http://agabian.ucsf.edu/canoDB/anno.php) (field: Annotator); (9) annotation data of the Fink's and Johnson's laboratories based on the orf6 protein set (unpublished) (field: note__AJ); (10) annotation comments available from CandidaDB (field: note__GF); (11) Pfam matches [114] obtained using the orf19 protein set (field: pfam__match); (12) Clusters of Orthologous Groups matches [115] obtained from the Comprehensive Yeast Genome Database using the CandidaDB protein set (field: COGs__MIPS); (13) EC number matches obtained from the Comprehensive Yeast Genome Database using the CandidaDB protein set (field: EC__number__MIPS); (14) *S. cerevisiae* closest protein including *e*-value, putative orthology, protein function, gene name, and alternate gene names, genome reference number obtained from SGD (field: Note); (15) GO annotation of the *S. cerevisiae* protein obtained from SGD (field: GO); (16) the reference of the allelic orf19 in the allelic set of sequences including supercontig number and location (field: Allele); and (17) chromosome assignment data available from the Magee's and Whiteway's laboratories (http://206.167.190.233/candida/index.cfm?page=CaChrom) (field: Chromosome).

Furthermore all ORFs were assigned a color code in order to facilitate annotation using the annotation tool Artemis [14]. All orf19 ORFs corresponding to an IPF classified as FALSORF and orf19 ORFs identified at the SGTC and not found among IPFs were color-coded in grey. orf19 ORFs with an unambiguous allele (90% identical amino acids over the whole length of the longest ORF) were color-coded in red (>150 codons) or pink (<150 codons). orf19 ORFs with a questionable allele (90% identical amino acids over the whole length of the shortest ORF) were color-coded in green (>150 codons) or pale green (<150 codons). orf19 ORFs without a clear allele (less than 90% identical amino acids over the whole length of the shortest ORF or no reciprocal match) were color-coded in blue (>150 codons) or light blue (<150 codons).

From this "master" file of sequences and their corresponding preliminary annotation, groups of contigs were selected and saved as partially annotated subsequences that were reserved and retrieved by members of the annotation consortium, and once fully annotated, were returned to a central Web site. A version of Artemis was distributed to the consortium that included a modified "options" file [116] allowing project-specific qualifiers to be used, and also featuring the *C. albicans*–specific translation Table 12 [117].

Whole genome BLAST searches and visualization of sequence homologies. ORF sequences were translated to proteins using the translation table for *C. albicans* [117], and compared using the BLASTP algorithm [118] with the NR database, the *C. albicans* proteome itself, the putative proteomes of five fungi *(S. cerevisiae, S. pombe, N. crassa, A. nidulans,* and *M. grisea),* and the proteomes of five other eukaryotes *(A. thaliana, D. melanogaster, C. elegans, M. musculus,* and *H. sapiens).* Sequence data were obtained from the EMBL-EBI Integr8 Browser (http://www.ebi.ac.uk/integr8/EBI-Integr8-HomePage.do) and the Broad Institute for Genome Research (http://www.broad.mit.edu/annotation/). For the most similar proteins by BLAST, negative exponents of the *e*-values were parsed from the output files, collected in a relational database, and visualized as a color range associated with each reference protein (see Figure 1). For this purpose, a Web-accessible visualization tool was constructed using Maromedia Cold Fusion and an Apache2 Web server. These results can be consulted at http://candida.bri.nrc.ca/candida/index.cfm?page=blast.

Bioinformatic identification of putative introns. Intron locations were predicted by constructing regular expressions based on consensus data drawn from several known introns in *C. albicans* (SOD1, EFB1, CMD1, CSK1, and others) as well as the extensive knowledge of the splice site consensus in *S. cerevisiae* [119] and matching of those expressions to the genomic DNA. Two regular expressions were used: /(.{15,}?TG[AT]A[CT]G)/ and /(.{700})([ATC]TACTAAC.{4,24}?[ATC][CTA]AG)(.{90})/. These incorporate several conserved aspects of mRNA splice sites including (1) the splice donor site of G(T/C)A(T/A)GT, where the initial guanine is the first residue within the intron, (2) a gap of between 15 and 700 nucleotides between the donor site and the branch point, (3) the branch point consensus sequence (A/T/C)TACTAAC, (4) a gap of between four and 24 nucleotides between the branch point and the splice acceptor site, (5) the splice acceptor site (T/A/C)(T/A/C)AG, where the final guanine is the final residue within the intron. This procedure provided landmarks that annotators could use to decide whether a gene contained introns, a decision also based on the position of nearby ORFs, on the ability of the intron/exon assembly to extend the size of the reading frame, and, most important, on the ability of the putative intron/exon assembly to improve similarity to sequences in other species. It predicted 1,297 introns in the *C. albicans* genome, many of which were not near ORFs or were otherwise incorrect. Nevertheless, this procedure was useful in providing markers for confirmation by the annotators.

Identification and classification of STRs. A STR was defined as a short nucleotide element (1–20 nt) repeated a number of times with a periodicity P, a length L, and a tolerated mutation window W. Since allelic indels are rare in *C. albicans,* we considered only mutations in STR sequences and not insertion or deletion from the consensus periodic pattern. W indicates the minimun distance between two mutations within a STR. For example, the STR "CTACAACAACAGCAAC" has $P = 3$, $L = 16$, and $W \leq 10$. When two STRs overlap they are merged together defining a single STR domain. Since short STRs can randomly occur, depending on the G/C content of the genome, we established a significant threshold length L_{MIN} for every periodicity P and mutation window W to represent the STR length for which the number of STR domains found in a randomized genome is less than 5% of the number found in the real genome. We call STR95 the set of all STR domains that are longer than the threshold value. In the STR95 set, every STR domain has a less than 5% chance of being a random event. With this method of setting the minimal STR length, no significant STR can be found in any randomized genome or in genomes that do not contain STRs in sufficient number to beat the odds 20 to one. More information and data can be found in Dataset S6.

Identification of spurious genes. For the calculation of gene expression correlation, we used a set of approximately 1,000 *S. cerevisiae* microarray experiments [127], and 216 genome-wide *C. albicans* expression profiles [7,13,108,120–126]. The iterative signature algorithm [35,127] was applied to the *C. albicans* expression data as described. We analyzed ORFs present on the arrays for which an orf19 number could be determined, including some that were subsequently removed from the final set of annotated genes. Pairwise Pearson correlation coefficients were calculated for each ORF with respect to all other ORFs across all of the experiments contained in the dataset. Random subsets of the *S. cerevisiae* data were generated by randomly selecting 200 experiments from the complete set of approximately 1,000 profiles. ORFs whose correlation coefficient exceeded the threshold value of 3σ with at least one other ORF in the dataset were recorded and excluded from the list of spurious gene candidates. The standard deviation σ of the background correlation of random gene pairs was measured to be $\sigma = 0.16$ for *S. cerevisiae* and $\sigma = 0.21$ for *C. albicans*. Similarly, genes possessing an ortholog in *S. cerevisiae* were excluded from the list of *C. albicans* candidates, and vice versa. All remaining ORFs were subsequently ordered by their length, and the 50 shortest ORFs were excluded (many of the shortest 50 genes correspond to real genes in *S. cerevisiae*).

Supporting Information

Dataset S1. Coordinates and All of the Annotation Fields for the 6,354 Confirmed *C. albicans* Genes, Based on the Version 19 Genome Assembly

Please note that Microsoft Excel may convert some of the gene names to dates and fail to import some of the largest fields.

Found at DOI: 10.1371/journal.pgen.0010001.sd001 (2 MB TXT).

Dataset S2. Sequence and Position of All Statistically Significant STRs in *C. albicans* Coding Sequences

Found at DOI: 10.1371/journal.pgen.0010001.sd002 (291 KB TXT).

Dataset S3. Sequence and Position of All Statistically Significant STRs in *S. pombe* Coding Sequences

Found at DOI: 10.1371/journal.pgen.0010001.sd003 (11 KB TXT).

Dataset S4. Sequence and Position of All Statistically Significant STRs in *S. cerevisiae* Coding Sequences

Found at DOI: 10.1371/journal.pgen.0010001.sd004 (66 KB TXT).

Dataset S5. Sequence and Position of All Statistically Significant STRs in *N. crassa* Coding Sequences

Found at DOI: 10.1371/journal.pgen.0010001.sd005 (488 KB TXT).

Dataset S6. Detailed Description of Our STR Identification Algorithm

Found at DOI: 10.1371/journal.pgen.0010001.sd006 (4 KB TXT).

Table S1. List of Potentially Spurious Genes

Found at DOI: 10.1371/journal.pgen.0010001.st001 (34 KB XLS).

Accession Numbers

The GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) accession numbers for ORFs discussed in this paper are *ALS3–1* (AY223552), *ALS3–2* (AY223551), *ALS4–2* (AF272027), *ALS5–1* (AY227440), *ALS5–2* (AY227439), strain SC5314 sequence corresponding to *ALS6* (AY225310), strain SC5314 *ALS9–1* (AY269423), and strain SC5314 *ALS9–2* (AY269422).

Acknowledgments

This project, and many others that applied functional genomic techniques to the study of *C. albicans,* would not have been possible without the public release of genomic sequence data by Ted Jones, Stewart Scherer, and their colleagues at the SGTC. We also wish to thank Keith James for Perl-scripting, Susumu Goto for corrections to the EC numbers, and numerous members of the *Candida* research community for their encouragements. We thank the *Magnaporthe* and *Aspergillus* Sequencing Projects of North Carolina State University and the Broad Institute of the Massachusetts Institute of Technology and Harvard (http://www.broad.mit.edu) for access to preliminary genome sequence data. Many of the annotators were supported by National Institutes of Health grant # RO1 AI49187 (to ADJ), European Comission grants # QLK2–2000–00795 and MCRTN-CT-2003–504148 (to CD), National Institute for Dental and Craniofacial Research grant # P01 DE07946 (to NA), and Canadian Institutes of Health Research grant # MOP-42516 (to MW). Finally, we gratefully acknowledge direct support for the community effort from the Pharmacia Corporation, the Burroughs Wellcome Fund, the Wellcome Trust, and the National Research Council Canada's Genome and Health Initiative. This is National Research Council Canada publication #46227.

Competing interests. BRB is an employee of both Incyte and University of California at San Francisco (UCSF). Incyte played no role in this work, and the resources used were all from UCSF.

Author contributions. BRB, CD, JD, DOI, MB, ML, APM, ADJ, MW, and AN conceived and designed the experiments. BRB, MM, JD, DOI, MAV, AL, VO, CB, DH, AM, DD, TL, RZ, MW, and AN performed the experiments. BRB, MV, CD, AK, HH, MB, CAM, SB, NAG, LLH, GK, JM, GN, SZ, MR, BT, JHI, JB, DS, NA, MW, and AN analyzed the data. BRB, MV, CD, JD, AK, DOI, MAV, HH, MB, ML, LF, FT, KR, EW, GS, MC, NA, APM, ADJ, and AN contributed reagents/materials/analysis tools. BRB, CD, AK, MB, EW, NAG, LLH, GK, JM, GN, SZ, MR, BT, GS, JB, DS, ADJ, MW, and AN wrote the paper. ■

References

1. Odds FC (1994) Pathogenesis of *Candida* infections. J Am Acad Dermatol 31: S2–S5.
2. Fradin C, Hube B (2003) Tissue infection and site-specific gene expression in *Candida albicans*. Adv Appl Microbiol 53: 271–290.
3. Berman J, Sudbery PE (2002) *Candida albicans:* A molecular revolution built on lessons from budding yeast. Nat Rev Genet 3: 918–930.
4. Fidel PL Jr (2002) Distinct protective host defenses against oral and vaginal candidiasis. Med Mycol 40: 359–375.
5. Naglik JR, Challacombe SJ, Hube B (2003) *Candida albicans* secreted aspartyl proteinases in virulence and pathogenesis. Microbiol Mol Biol Rev 67: 400–428.
6. Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, et al. (2004) The diploid genome sequence of *Candida albicans*. Proc Natl Acad Sci U S A 101: 7329–7334.
7. Nantel A, Dignard D, Bachewich C, Harcus D, Marcil A, et al. (2002) Transcription profiling of *Candida albicans* cells undergoing the yeast to hyphal transition. Mol Biol Cell 13: 3452–3465.
8. Roemer T, Jiang B, Davison J, Ketela T, Veillette K, et al. (2003) Large-scale essential gene identification in *Candida albicans* and applications to antifungal drug discovery. Mol Microbiol 50: 167–181.
9. Uhl MA, Biery M, Craig N, Johnson AD (2003) Haploinsufficiency-based large-scale forward genetic analysis of filamentous growth in the diploid human fungal pathogen *C. albicans*. EMBO J 22: 2668–2678.
10. Tzung KW, Williams RM, Scherer S, Federspiel N, Jones T, et al. (2001) Genomic evidence for a complete sexual cycle in *Candida albicans*. Proc Natl Acad Sci U S A 98: 3249–3253.
11. Lukashin AV, Borodovsky M (1998) GeneMark.hmm: New solutions for gene finding. Nucleic Acids Res 26: 1107–1115.
12. d'Enfert C, Goyard S, Rodriguez-Arnaveilhe S, Frangeul L, Jones L, et al. (2005) CandidaDB: A genome database for *Candida albicans* pathogenomics. Nucleic Acids Res 33: D353–D357.
13. Bennett RJ, Uhl MA, Miller MG, Johnson AD (2003) Identification and characterization of a *Candida albicans* mating pheromone. Mol Cell Biol 23: 8189–8201.
14. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, et al. (2000) Artemis: Sequence visualization and annotation. Bioinformatics 16: 944–945.
15. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, et al. (1996) Life with 6000 genes. Science 274: 546, 563–547.
16. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423: 241–254.
17. Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, et al. (2002) The genome sequence of *Schizosaccharomyces pombe*. Nature 415: 871–880.
18. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, et al. (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422: 859–868.
19. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815.
20. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, et al. (2000) The genome sequence of *Drosophila melanogaster*. Science 287: 2185–2195.
21. C. elegans Sequencing Consortium (1998) Genome sequence of the nematode C. elegans: A platform for investigating biology. Science 282: 2012–2018.
22. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.
23. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
24. Cherry JM (1995) Genetic nomenclature guide. *Saccharomyces cerevisiae.* Trends Genet 1995: 11–12.
25. Arnaud MB, Constanzo MC, Skrzypek MS, Binkley G, Lane C, et al. (2005) The *Candida* Genome Database (CGD), a community resource for *Candida albicans* gene and protein information. Nucleic Acid Res 33: D358–D363.
26. Mourier T, Jeffares DC (2003) Eukaryotic intron loss. Science 300: 1393.
27. Nielsen CB, Friedman B, Birren B, Burge CB, Galagan JE (2004) Patterns of intron gain and loss in fungi. PLoS Biol 2: DOI: 10.1371/journal.pbio.0020422
28. Fedorov A, Merican AF, Gilbert W (2002) Large-scale comparison of intron positions among animal, plant, and fungal genes. Proc Natl Acad Sci U S A 99: 16128–16133.
29. Kersanach R, Brinkmann H, Liaud MF, Zhang DX, Martin W, et al. (1994) Five identical intron positions in ancient duplicated genes of eubacterial origin. Nature 367: 387–389.
30. Boucher HW, Groll AH, Chiou CC, Walsh TJ (2004) Newer systemic antifungal agents: Pharmacokinetics, safety and efficacy. Drugs 64: 1997–2020.
31. Toth G, Gaspari Z, Jurkat J (2000) Microsatellites in different eukaryotic genomes: Survey and analysis. Genome Res 10: 967–981.
32. Katti MV, Ranjekar PK, Gupta VS (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 18: 1161–1167.
33. Astolfi P, Bellizzi D, Sgaramella V (2003) Frequency and coverage of trinucleotide repeats in eukaryotes. Gene 317: 117–125.
34. Nag DK, Suri M, Stenson EK (2004) Both CAG repeats and inverted DNA repeats stimulate spontaneous unequal sister-chromatid exchange in *Saccharomyces cerevisiae*. Nucleic Acids Res 32: 5677–5684.
35. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, et al. (2002) Revealing modular organization in the yeast transcriptional network. Nat Genet 31: 370–377.
36. Magee BB, Hube B, Wright RJ, Sullivan PJ, Magee PT (1993) The genes encoding the secreted aspartyl proteinases of *Candida albicans* constitute a family with at least three members. Infect Immun 61: 3240–3243.
37. White TC, Miyasaki SH, Agabian N (1993) Three distinct secreted aspartyl proteinases in *Candida albicans*. J Bacteriol 175: 6126–6133.
38. Hoyer LL, Payne TL, Bell M, Myers AM, Scherer S (1998) *Candida albicans* ALS3 and insights into the nature of the ALS gene family. Curr Genet 33: 451–459.
39. Hube B, Stehr F, Bossenz M, Mazur A, Kretschmar M, et al. (2000) Secreted lipases of *Candida albicans:* Cloning, characterisation and expression analysis of a new gene family with at least ten members. Arch Microbiol 174: 362–374.
40. Ramanan N, Wang Y (2000) A high-affinity iron permease essential for *Candida albicans* virulence. Science 288: 1062–1064.
41. Lan CY, Rodarte G, Murillo LA, Jones T, Davis RW, et al. (2004) Regulatory networks affected by iron availability in *Candida albicans*. Mol Microbiol 53: 1451–1469.
42. Bauer BE, Wolfger H, Kuchler K (1999) Inventory and function of yeast ABC proteins: About sex, stress, pleiotropic drug and heavy metal resistance. Biochim Biophys Acta 1461: 217–236.
43. Prasad R, De Wergifosse P, Goffeau A, Balzi E (1995) Molecular cloning and characterization of a novel gene of *Candida albicans*, CDR1, conferring multiple resistance to drugs and antifungals. Curr Genet 27: 320–329.
44. Sanglard D, Kuchler K, Ischer F, Pagani JL, Monod M, et al. (1995) Mechanisms of resistance to azole antifungal agents in *Candida albicans* isolates from AIDS patients involve specific multidrug transporters. Antimicrob Agents Chemother 39: 2378–2386.
45. Sanglard D, Ischer F, Monod M, Bille J (1997) Cloning of *Candida albicans* genes conferring resistance to azole antifungal agents: Characterization of CDR2, a new multidrug ABC-transporter gene. Microbiology 143: 405–416.
46. Lan CY, Newport G, Murillo LA, Jones T, Scherer S, et al. (2002) Metabolic specialization associated with phenotypic switching in *Candida albicans*. Proc Natl Acad Sci U S A 99: 14907–14912.
47. Smriti Krishnamurthy S, Dixit BL, Gupta CM, Milewski S, et al. (2002) ABC transporters Cdr1p, Cdr2p and Cdr3p of a human pathogen *Candida albicans* are general phospholipid translocators. Yeast 19: 303–318.
48. Mason DL, Mallampalli MP, Huyer G, Michaelis S (2003) A region within a lumenal loop of *Saccharomyces cerevisiae* Ycf1p directs proteolytic processing and substrate specificity. Eukaryot Cell 2: 588–598.
49. Theiss S, Kretschmar M, Nichterlein T, Hof H, Agabian N, et al. (2002)

Functional analysis of a vacuolar ABC transporter in wild-type *Candida albicans* reveals its involvement in virulence. Mol Microbiol 43: 571–584.

50. Hoyer LL (2001) The ALS gene family of *Candida albicans*. Trends Microbiol 9: 176–180.
51. Sheppard DC, Yeaman MR, Welch WH, Phan QT, Fu Y, et al. (2004) Functional and structural diversity in the Als protein family of *Candida albicans*. J Biol Chem 279: 30480–30489.
52. Hoyer LL, Scherer S, Shatzman AR, Livi GP (1995) *Candida albicans* ALS1: Domains related to a *Saccharomyces cerevisiae* sexual agglutinin separated by a repeating motif. Mol Microbiol 15: 39–54.
53. Hoyer LL, Payne TL, Hecht JE (1998). Identification of *Candida albicans* ALS2 and ALS4 and localization of Als proteins to the fungal cell surface. J Bacteriol 180: 5334–5343.
54. Gaur NK, Klotz SA (1997) Expression, cloning, and characterization of a *Candida albicans* gene, ALA1, that confers adherence properties upon *Saccharomyces cerevisiae* for extracellular matrix proteins. Infect Immun 65: 5289–5294.
55. Hoyer LL, Hecht JE (2000) The ALS6 and ALS7 genes of *Candida albicans*. Yeast 16: 847–855.
56. Hoyer LL, Hecht JE (2001) The ALS5 gene of *Candida albicans* and analysis of the Als5p N-terminal domain. Yeast 18: 49–60.
57. Zhao X, Pujol C, Soll DR, Hoyer LL (2003) Allelic variation in the contiguous loci encoding *Candida albicans* ALS5, ALS1 and ALS9. Microbiology 149: 2947–2960.
58. Lott TJ, Holloway BP, Logan DA, Fundyga R, Arnold J (1999) Towards understanding the evolution of the human commensal yeast *Candida albicans*. Microbiology 145: 1137–1143.
59. Zhang N, Harrex AL, Holland BR, Fenton LE, Cannon RD, et al. (2003) Sixty alleles of the ALS7 open reading frame in *Candida albicans:* ALS7 is a hypermutable contingency locus. Genome Res 13: 2005–2017.
60. Oh SH, Cheng G, Nuessen JA, Jajko R, Yeater KM, et al. (2004) Functional specificity of *Candida albicans* Als3p proteins and clade specificity of ALS3 alleles discriminated by the number of copies of the tandem repeat sequence in the central domain. Microbiology 151: 673–681.
61. Biswas K, Morschhäuser J (2005) The Mep2p ammonium permease controls nitrogen starvation-induced filamentous growth in *Candida albicans*. Mol Microbiol 56: 649–669.
62. Marini AM, Soussi-Boudekou S, Vissers S, Andre B (1997) A family of ammonium transporters in *Saccharomyces cerevisiae*. Mol Cell Biol 17: 4282–4293.
63. Lorenz MC, Heitman J (1998) The MEP2 ammonium permease regulates pseudohyphal differentiation in *Saccharomyces cerevisiae*. EMBO J 17: 1236–1247.
64. Lubkowitz MA, Hauser L, Breslav M, Naider F, Becker JM (1997) An oligopeptide transport gene from *Candida albicans*. Microbiology 143: 387–396.
65. Reuß O, Vik Å, Kolter R, Morschhäuser J (2004) The SAT1 flipper, an optimized tool for gene disruption in *Candida albicans*. Gene 341: 119–127.
66. Matthews JM, Sunde M (2002) Zinc fingers—Folds for many occasions. Life 54: 351–355.
67. Schjerling P, Holmberg S (1996) Comparative amino acid sequence analysis of the C6 zinc cluster family of transcriptional regulators. Nucleic Acids Res 24: 4599–4607.
68. Todd RB, Andrianopoulos A (1997) Evolution of a fungal regulatory gene family: The Zn(II)2Cys6 binuclear cluster DNA binding motif. Fungal Genet Biol 21: 388–405.
69. Akache B, Wu K, Turcotte B (2001) Phenotypic analysis of genes encoding yeast zinc cluster proteins. Nucleic Acids Res 29: 2181–2190.
70. Hellauer K, Rochon MH, Turcotte B (1996) A novel DNA binding motif for yeast zinc cluster proteins: The Leu3p and Pdr3p transcriptional activators recognize everted repeats. Mol Cell Biol 16: 6096–6102.
71. Kelly R, Kwon-Chung KJ (1992) A zinc finger protein from *Candida albicans* is involved in sucrose utilization. J Bacteriol 174: 222–232.
72. Talibi D, Raymond M (1999) Isolation of a putative *Candida albicans* transcriptional regulator involved in pleiotropic drug resistance by functional complementation of a pdr1 pdr3 mutation in *Saccharomyces cerevisiae*. J Bacteriol 181: 231–240.
73. Moreno I, Pedreno Y, Maicas S, Sentandreu R, Herrero E, et al. (2003) Characterization of a *Candida albicans* gene encoding a putative transcriptional factor required for cell wall integrity. FEMS Microbiol Lett 226: 159–167.
74. Brown DH Jr, Giusani AD, Chen X, Kumamoto CA (1999) Filamentous growth of *Candida albicans* in response to physical environmental cues and its regulation by the unique CZF1 gene. Mol Microbiol 34: 651–662.
75. McDonough JA, Bhattacherjee V, Sadlon T, Hostetter MK (2002) Involvement of *Candida albicans* NADH dehydrogenase complex I in filamentation. Fungal Genet Biol 36: 117–127.
76. Huh WK, Kang SO (2001) Characterization of the gene family encoding alternative oxidase from *Candida albicans*. Biochem J 356: 595–604.
77. Ghannoum MA (2000) Potential role of phospholipases in virulence and fungal pathogenesis. Clin Microbiol Rev 13: 122–143.
78. Schmiel DH, Miller VL (1999) Bacterial phospholipases and pathogenesis. Microbes Infect 1: 1103–1112.
79. Eisenhaber B, Schneider G, Wildpaner M, Eisenhaber F (2004) A sensitive predictor for potential GPI lipid modification sites in fungal protein sequences and its application to genome-wide studies for *Aspergillus nidulans*, *Candida albicans*, *Neurospora crassa*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. J Mol Biol 337: 243–253.
80. Lee SA, Wormsley S, Kamoun S, Lee AF, Joiner K, et al. (2003) An analysis of the *Candida albicans* genome database for soluble secreted proteins using computer-based prediction algorithms. Yeast 20: 595–610.
81. Hoover CI, Jantapour MJ, Newport G, Agabian N, Fisher SJ (1998) Cloning and regulated expression of the *Candida albicans* phospholipase B (PLB1) gene. FEMS Microbiol Lett 167: 163–169.
82. Leidich SD, Ibrahim AS, Fu Y, Koul A, Jessup C, et al. (1998) Cloning and disruption of caPLB1, a phospholipase B gene involved in the pathogenicity of *Candida albicans*. J Biol Chem 273: 26078–26086.
83. Mukherjee PK, Seshan KR, Leidich SD, Chandra J, Cole GT, et al. (2001) Reintroduction of the PLB1 gene into *Candida albicans* restores virulence in vivo. Microbiology 147: 2585–2597.
84. Sugiyama Y, Nakashima S, Mirbod F, Kanoh H, Kitajima Y, et al. (1999) Molecular cloning of a second phospholipase B gene, caPLB2 from *Candida albicans*. Med Mycol 37: 61–67.
85. Bennett DE, McCreary CE, Coleman DC (1998) Genetic characterization of a phospholipase C gene from *Candida albicans:* Presence of homologous sequences in *Candida* species other than *Candida albicans*. Microbiology 144: 55–72.
86. Hube B, Hess D, Baker CA, Schaller M, Schafer W, et al. (2001) The role and relevance of phospholipase D1 during growth and dimorphism of *Candida albicans*. Microbiology 147: 879–889.
87. Dickson RC, Lester RL (2002) Sphingolipid functions in *Saccharomyces cerevisiae*. Biochim Biophys Acta 1583: 13–25.
88. Obeid LM, Okamoto Y, Mao C (2002) Yeast sphingolipids: Metabolism and biology. Biochim Biophys Acta 1585: 163–171.
89. Matsubara T, Hayashi A, Banno Y, Morita T, Nozawa Y (1987) Cerebroside of the dimorphic human pathogen, *Candida albicans*. Chem Phys Lipids 43: 1–12.
90. Ghannoum MA, Janini G, Khamis L, Radwan SS (1986) Dimorphism-associated variations in the lipid composition of *Candida albicans*. J Gen Microbiol 132: 2367–2375.
91. Leipelt M, Warnecke D, Zahringer U, Ott C, Muller F, et al. (2001) Glucosylceramide synthases, a gene family responsible for the biosynthesis of glucosphingolipids in animals, plants, and fungi. J Biol Chem 276: 33621–33629.
92. Barreto-Bergter E, Pinto MR, Rodrigues ML (2004) Structure and biological functions of fungal cerebrosides. An Acad Bras Cienc 76: 67–84.
93. Ternes P, Franke S, Zahringer U, Sperling P, Heinz E (2002) Identification and characterization of a sphingolipid delta 4-desaturase family. J Biol Chem 277: 25512–25518.
94. Saupe S, Descamps C, Turcq B, Begueret J (1994) Inactivation of the *Podospora anserina* vegetative incompatibility locus het-c, whose product resembles a glycolipid transfer protein, drastically impairs ascospore production. Proc Natl Acad Sci U S A 91: 5927–5931.
95. Mattjus P, Turcq B, Pike HM, Molotkovsky JG, Brown RE (2003) Glycolipid intermembrane transfer is accelerated by HET-C2, a filamentous fungus gene product involved in the cell-cell incompatibility response. Biochemistry 42: 535–542.
96. Mukhopadhyay K, Prasad T, Saini P, Pucadyil TJ, Chattopadhyay A, et al. (2004) Membrane sphingolipid-ergosterol interactions are important determinants of multidrug resistance in *Candida albicans*. Antimicrob Agents Chemother 48: 1778–1787.
97. Gonzalez-Zorn B, Dominguez-Bernal G, Suarez M, Ripio MT, Vega Y, et al. (1999) The smcL gene of *Listeria ivanovii* encodes a sphingomyelinase C that mediates bacterial escape from the phagocytic vacuole. Mol Microbiol 33: 510–523.
98. Grassme H, Gulbins E, Brenner B, Ferlinz K, Sandhoff K, et al. (1997) Acidic sphingomyelinase mediates entry of *N. gonorrhoeae* into non-phagocytic cells. Cell 91: 605–615.
99. Hauck CR, Grassme H, Bock J, Jendrossek V, Ferlinz K, et al. (2000) Acid sphingomyelinase is involved in CEACAM receptor-mediated phagocytosis of *Neisseria gonorrhoeae*. FEBS Lett 478: 260–266.
100. Projan SJ, Kornblum J, Kreiswirth B, Moghazeh SL, Eisner W, et al. (1989) Nucleotide sequence: The beta-hemolysin gene of *Staphylococcus aureus*. Nucleic Acids Res 17: 3305.
101. Marshall MJ, Bohach GA, Boehm DF (2000) Characterization of *Staphylococcus aureus* beta-toxin induced leukotoxicity. J Nat Toxins 9: 125–138.
102. Esen M, Schreiner B, Jendrossek V, Lang F, Fassbender K, et al. (2001) Mechanisms of *Staphylococcus aureus* induced apoptosis of human endothelial cells. Apoptosis 6: 431–439.
103. Tseng HJ, Chan CC, Chan EC (2004) Sphingomyelinase of *Helicobacter pylori*-induced cytotoxicity in AGS gastric epithelial cells via activation of JNK kinase. Biochem Biophys Res Commun 314: 513–518.
104. Makino Y, Yogosawa S, Kayukawa K, Coin F, Egly JM, et al. (1999) TATA-binding protein-interacting protein 120, TIP120, stimulates three classes of eukaryotic transcription via a unique mechanism. Mol Cell Biol 19: 7951–7960.
105. Inoki K, Zhu T, Guan KL (2003) TSC2 mediates cellular energy response to control cell growth and survival. Cell 115: 577–590.
106. Lapik YR, Kaufman LS (2003) The Arabidopsis cupin domain protein

AtPirin1 interacts with the G protein alpha-subunit GPA1 and regulates seed germination and early seedling development. Plant Cell 15: 1578–1590.

107. Gallio M, Sturgill G, Rather P, Kylsten P (2002) A conserved mechanism for extracellular signaling in eukaryotes and prokaryotes. Proc Natl Acad Sci U S A 99: 12208–12213.
108. Tsong AE, Miller MG, Raisner RM, Johnson AD (2003) Evolution of a combinatorial transcriptional circuit: A case study in yeasts. Cell 115: 389–399.
109. Frangeul L, Glaser P, Rusniok C, Buchrieser C, Duchaud E, et al. (2004) CAAT-Box, contigs-assembly and annotation tool-box for genome sequencing projects. Bioinformatics 20: 790–797.
110. Altschul SF, Gish W, Miller W, Myers EW, Lipman D (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
111. Tekaia F, Blandin G, Malpertuy A, Llorente B, Durrens P, et al. (2000) Genomic exploration of the hemiascomycetous yeasts: 3. Methods and strategies used for sequence analysis and annotation. FEBS Lett 487: 17–30.
112. Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, et al. (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. Nucleic Acids Res 32: D311–D314.
113. Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, et al. (2004) MIPS: Analysis and annotation of proteins from whole genomes. Nucleic Acids Res 32: D41–D44.
114. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. (2004) The Pfam protein families database. Nucleic Acids Res 32: D138–D141.
115. Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, et al. (2004) Database resources of the National Center for Biotechnology Information: Update. Nucleic Acids Res 32: D35–D40.
116. Berriman M, Rutherford K (2003) Viewing and annotating sequence data with Artemis. Brief Bioinform 4: 124–132.
117. Ohama T, Suzuki T, Mori M, Osawa S, Ueda T, et al. (1993) Non-universal decoding of the leucine codon CUG in several *Candida* species. Nucleic Acids Res 21: 4039–4045.
118. Altschul SF (1997) Evaluating the statistical significance of multiple distinct local alignments. In: Suhai S, editor. Theoretical and computational methods in genome research. New York: Plenum. pp. 1–14
119. Cheng SC (1994) Formation of the yeast splicing complex A1 and association of the splicing factor PRP19 with the pre-mRNA are independent of the 3′ region of the intron. Nucleic Acids Res 22: 1548–1554.
120. Garcia-Sanchez S, Aubert S, Iraqui I, Janbon G, Ghigo JM, et al. (2004) *Candida albicans* biofilms: A developmental state associated with specific and stable gene expression patterns. Eukaryot Cell 3: 536–545.
121. Bensen ES, Martin SJ, Li M, Berman J, Davis DA (2004) Transcriptional profiling in *C. albicans* reveals new adaptive responses to extracellular pH and functions for Rim101p. Mol Microbiol 54: 1335–1351.
122. Cowen LE, Nantel A, Tessier D, Whiteway M, Thomas DY, et al. (2002) Population genomics of drug resistance in experimental populations of *Candida albicans*. Proc Natl Acad Sci U S A 99: 9284–9289.
123. Enjalbert B, Nantel A, Whiteway M (2003) Stress induced gene expression in *Candida albicans:* Absence of a general stress response. Mol Biol Cell 14: 1460–1467.
124. Lee CM, Nantel A, Jiang L, Whiteway M, Shen SH (2004) The serine/threonine protein phosphatase SIT4 modulates yeast-to-hypha morphogenesis and virulence in *Candida albicans*. Mol Microbiol 51: 691–709.
125. Karababa M, Coste AT, Rognon B, Bille J, Sanglard D (2004) Comparison of gene expression profiles of *Candida albicans* azole-resistant clinical isolates and laboratory strains exposed to drugs inducing multidrug transporters. Antimicrob Agents Chemother 48: 3064–3079.
126. Rogers PD, Barker KS (2003) Genome-wide expression profile analysis reveals coordinately regulated genes associated with stepwise acquisition of azole resistance in *Candida albicans* clinical isolates. Antimicrob Agents Chemother 47: 1220–1227.
127. Ihmels J, Bergmann S, Barkai N (2004) Defining transcription modules using large-scale gene expression data. Bioinformatics 20: 1993–2003.
128. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola I, et al. (2004) Genome evolution in yeast. Nature 430: 35–44.
129. Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, et al. (2005) The genome of the basidiomycetous yeast and human pathogen *Cryptococcus Neoformans*. Science 307: 1321–1324.
130. Balan I, Alarco AM, Raymond M (1997) The *Candida albicans* CDR3 gene codes for an opaque-phase ABC transporter. J Bacteriol 179: 7210–7218.
131. Franz R, Michel S, Morschhauser J (1998) A fourth gene from the *Candida albicans* CDR family of ABC transporters. Gene 220: 91–98.
132. Raymond M, Dignard D, Alarco AM, Mainville N, Magee BB, et al. (1998) A Ste6p/P-glycoprotein homolog from the asexual yeast *Candida albicans* transports the a-factor mating pheromone in *Saccharomyces cerevisiae*. Mol Microbiol 27: 587–598.
133. Ogawa A, Hashida-Okado T, Endo M, Yoshioka H, Tsuruo T, et al. (1998) Role of ABC transporters in aureobasidin A resistance. Antimicrob Agents Chemother 42: 755–761.
134. Di Domenico BJ, Lupisella J, Sandbaken M, Chakraburtty K (1992) Isolation and sequence analysis of the gene encoding translation elongation factor 3 from *Candida albicans*. Yeast 8: 337–352.
135. Sturtevant J, Cihlar R, Calderone R (1998) Disruption studies of a *Candida albicans* gene, ELF1: A member of the ATP-binding cassette family. Microbiology 144: 2311–2321.
136. Andaluz E, Coque JJ, Cueva R, Larriba G (2001) Sequencing of a 4.3 kbp region of chromosome 2 of *Candida albicans* reveals the presence of homologs of SHE9 from *Saccharomyces cerevisiae* and of bacterial phosphatidylinositol-phospholipase C. Yeast 18: 711–721.
137. Kanoh H, Nakashima S, Zhao Y, Sugiyama Y, Kitajima Y, et al. (1998) Molecular cloning of a gene encoding phospholipase D from the pathogenic and dimorphic fungus, *Candida albicans*. Biochim Biophys Acta 9: 359–364.