The diploid genome sequence of *Candida albicans*

Ted Jones*, Nancy A. Federspiel*†, Hiroji Chibana‡, Jan Dungan§, Sue Kalman*, B. B. Magee¶, George Newport§, Yvonne R. Thorstenson*, Nina Agabian§‖, P. T. Magee¶, Ronald W. Davis*††, and Stewart Scherer‡‡**

*Stanford Genome Technology Center, Palo Alto, CA 94304; ‡Research Center for Pathogenic Fungi and Microbial Toxicoses, Chiba University, Chiba 260-8673, Japan; Departments of §Stomatology, ‖Microbiology and Immunology, and **Pharmaceutical Chemistry, University of California, San Francisco, CA 94143; ¶Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, MN 55455; and ‡‡3938 Paseo Grande, Moraga, CA 94556

Contributed by Ronald W. Davis, March 8, 2004

We present the diploid genome sequence of the fungal pathogen *Candida albicans*. Because *C. albicans* has no known haploid or homozygous form, sequencing was performed as a whole-genome shotgun of the heterozygous diploid genome in strain SC5314, a clinical isolate that is the parent of strains widely used for molecular analysis. We developed computational methods to assemble a diploid genome sequence in good agreement with available physical mapping data. We provide a whole-genome description of heterozygosity in the organism. Comparative genomic analyses provide important clues about the evolution of the species and its mechanisms of pathogenesis.

Candida albicans, one of the first eukaryotic pathogens selected for genome sequencing, is the most commonly encountered human fungal pathogen, causing skin and mucosal infections in generally healthy individuals and life-threatening infections in persons with severely compromised immune function. The many clinical isolates of *C. albicans* used for laboratory study and genetic analysis are generally diploid and exhibit considerable natural heterozygosity, and some have translocations in their genomes (1–3). Although mating governed by a mating-type-like locus can occur, a meiotic phase for the sexual cycle remains obscure and, unlike most species chosen for sequencing, a haploid or homozygous form for *C. albicans* is not available (4–6). Strain SC5314 (7) was chosen for large-scale sequencing because of its widespread and increasing use in molecular analyses, virulence in animal models, and apparent standard diploid electrophoretic karyotype.

Whole-genome shotgun (WGS) sequencing has been successfully applied to very large genomes; however, standard assembly software does not allow for the possibility of two homologs with varying degrees of similarity and does not assemble such sequences correctly unless the sequence is nearly homozygous throughout the genome. To assemble the *C. albicans* diploid genome sequence, we began with PHRAP, the widely used assembly program (www.phrap.org). Application of PHRAP resulted in an assembly (Assembly 6) in which the sum of the contigs significantly exceeded the haploid genome size. Here we describe the conversion of this standard PHRAP assembly into a diploid assembly that is in good agreement with available physical mapping data. The diploid sequence assembly reveals the nature and extent of heterozygosity in strain SC5314. Together with the gene set inferred from the sequence, these results provide significant insights into *C. albicans* evolution and pathogenesis.

Methods

Assembly. The *C. albicans* WGS sequence was initially assembled with PHRAP (www.phrap.org), at 10.9× (haploid trimmed coverage; see *Supporting Text*, which is published as supporting information on the PNAS web site), equivalent to 7.1× PHRED20 coverage, using special methods to handle mitochondrial and rDNA sequence, which had very high sequence coverage. Because the PHRAP assembler assumes single-copy sequence, we expected problems in its application to a heterozygous genome. PHRAP assembly resulted in 2,519 contigs. Even after discarding short low-coverage contigs typical in large assemblies, the number was far larger than expected given the coverage (8). The sum of the contig lengths exceeded the genome size by ≈20%. Genes believed to be single copy were often found on two contigs, suggesting that homologous sequences were sometimes assembled into separate contigs. Standard finishing experiments designed to close gaps, normally undertaken after completing an assembly, were inappropriate if most apparent gaps were caused by separate assembly of heterozygous sequence, not lack of data.

We call regions of the assembly where homologs assembled together "nearly homozygous" and regions of separated assembly "heterozygous." Although separated homologs usually had similar sequence, similarity alone was insufficient to identify them amid the many duplicated sequences in the genome. Sequence alignments between separated homologs, however, do have distinctive properties that are created as a byproduct of assembly. To reconstruct the diploid genome sequence from the PHRAP assembly, it was necessary to identify heterozygous regions of the assembly, align separated homologs, and appropriately join them. As shown in Fig. 1 *b* and *c*, the logic of single-copy assembly, applied to diploid sequence, dictates that separated homologs must give rise to what we call "terminal alignments."

We began the diploid assembly with an all-against-all pairwise BLASTN comparison of the PHRAP contigs. Each BLASTN alignment was examined to determine whether it was terminal (either type 1*b* or 1*c*). Alignments containing repeat sequences located at contig ends were treated as nonterminal (see *Supporting Text*). When terminal alignments attributable to separately assembled homologs were identified, the process described in Fig. 1 was reversed, reconstructing the original diploid sequence as shown in Fig. 1*a*. The process was continued in both directions as far as possible, producing two homologous supercontigs assembled from multiple PHRAP contigs. Fig. 2 shows how the process created one typical supercontig pair from five PHRAP contigs.

Diploid assembly does not necessarily reconstruct haplotypes. Across homozygous genomic regions larger than the read length, the WGS does not provide data from which to determine phase. PHRAP does not examine phase information; however, the diploid assembler places each PHRAP contig entirely into one or the other homologous supercontig, preserving phase.

In practice, identification of terminal alignments was not always straightforward. In general, sequence at the ends of PHRAP contigs came from a single read and therefore was often of low quality and sometimes chimeric. This prevented most homologous terminal alignments from reaching the very end of the contig. We used PHRAP's base quality scores to perform a

Abbreviations: WGS, whole-genome shotgun; MRS, major repeat sequence; TR, tandem repeat; VNTR, variable number of TR.

Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. AACQ00000000). The version described in this paper is the first version, AACQ01000000.

†Present address: Department of Anesthesia, Stanford University, Stanford, CA 94305.

††To whom correspondence should be addressed. E-mail: dbowe@stanford.edu.

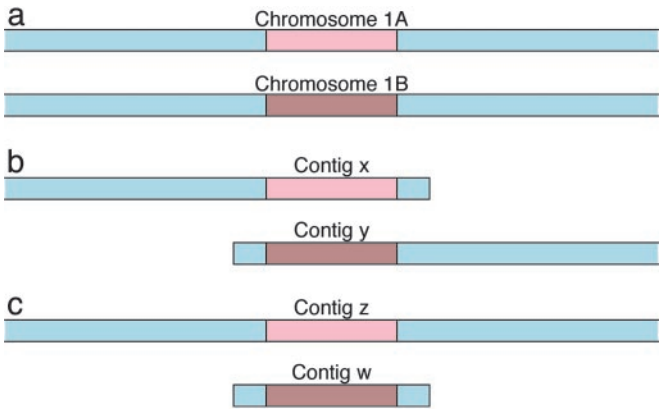

Fig. 1. Assembly strategy. Effects of separate assembly of diverged homologs by a single-copy assembler such as PHRAP. (*a*) Hypothetical configuration of genomic sequence. Two diverged homologous regions are shown in pink and brown, flanked by nearly homozygous sequence shown in blue. Reads containing pink sequence look different from brown reads and must not assemble into the same contig. In the blue regions, reads from either homolog look alike and be assembled together. (*b* and *c*) The two possible ways in which these conditions can be met by the assembler. In both cases, two contigs are produced, one containing pink reads and the other, brown. In *b*, the two blue flanking regions assemble into different contigs. The first contig contains a small amount of blue sequence on the right because of reads that are mostly pink but extend into the blue region. The second similarly contains a small amount of blue sequence (on the left). In *c*, both blue flanking regions are assembled into the contig containing the pink homolog. The second contig consists only of the brown homolog plus a small amount of blue sequence, as described for *b*. In both cases, the PHRAP contig numbers x, y, z, and w are arbitrary, and the separated homologs must be located by sequence alignment. In *b*, it is predicted that the alignment will extend to the right end of contig x and the left end of y. In *c*, the alignment will include both ends of contig w, running the entire length of the contig. We call such alignments terminal.

statistical test for assessing terminality, and suspected chimeric contig ends were identified and trimmed in the diploid assembly (see *Supporting Text*).

Special methods were devised to assemble across transposon insertions and large substitutions such as the mating-type-like locus. In a process analogous to finishing in single-copy WGS, manual assembler directives based on physical mapping data, paired plasmid clone sequences, and known GenBank *C. albicans* sequences were used to guide the assembly. A more detailed description of the assembly is presented in *Supporting Text*, Table 6, and Fig. 6, which are published as supporting information on the PNAS web site.

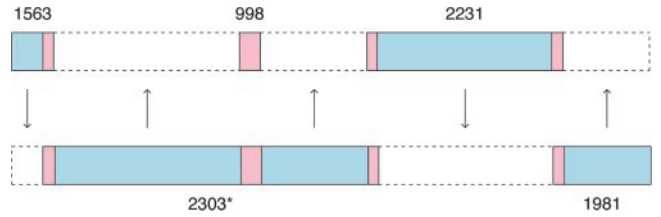

Fig. 2. Diploid assembly of a pair of homologous supercontigs. Shown is a pair of homologous supercontigs (10065 and 20065) built from PHRAP contigs 1563, 2303*, 998, 2231, and 1981, where * denotes sequence complementation. There are terminal alignments indicating separately assembled homologous regions occurring between 1563:2303*, 998:2303*, 2303*:2231, and 2231:1981. For simplicity, both diverged homologs are shown in pink. In nearly homozygous regions of the PHRAP assembly, where a single sequence represents both homologs, sequence is copied in the direction shown by the arrows to fill in the dotted regions in the opposite homologous supercontig, reversing the process described in Fig. 1. In the heterozygous regions, low-quality bases at the ends of PHRAP contigs corresponding to the small blue regions in Fig. 1 *b* and *c* are also replaced with sequence from the other homolog. Not shown is a small region of internal trimming in contig 998 (see *Supporting Text*).

Identification of Heterozygosity in Strain SC5314. In nearly homozygous regions of the PHRAP assembly, where highly similar homologs assembled together, polymorphisms were identified by scanning PHRAP contigs for positions having a pattern of high-quality disagreements between individual reads. Similar methods have been used to find polymorphisms in the human (9) and *Anopheles* (10) whole-genome assemblies. By aligning homologous supercontig pairs, we were able to identify many additional polymorphisms between homologs that PHRAP had assembled separately. Both methods of polymorphism discovery use base quality scores to distinguish true polymorphisms from sequencing errors (see *Supporting Text*).

Results

Global Genome Characteristics. The final diploid sequence (released as Assembly 19) is distributed over 412 supercontigs: 146 homologous pairs, 119 PHRAP contigs lacking usable joins with others, and a single supercontig formed from two PHRAP contigs joined on the basis of GenBank sequence. A reference haploid genome consisting of 266 supercontigs was created by arbitrarily discarding one from each pair of the homologous supercontigs. The reference haploid genome contains 7,677 ORFs of 100 amino acids or greater, including incomplete ORFs at the ends of supercontigs. A "reduced" set of 6,419 ORFs was derived by eliminating the smaller of a pair of ORFs that overlap by >50%. Even ORF overlaps of <50% were expected to be rare, but we preferred to err on the side of completeness. In most cases, assignment of the ORF encoding the second allele of a pair is relatively straightforward.

The genome size and physical map of *C. albicans* has been examined primarily in strain CBS5736 and its derivatives (3). No significant differences were found between the electrophoretic karyotype of the sequencing strain SC5314 and CBS5736. Size estimates of the SC5314 chromosomes are presented in Table 1. Given the assumptions made in determining genome size, the assembled haploid genome sequence of *C. albicans* is in remarkably good agreement with estimates of genome size derived from physical criteria. Supercontigs with sequenced map markers were assigned to the chromosomes from which the markers derive. The varying levels of coverage of individual chromosomes, lowest on chromosome R, relate to the number and distribution of markers on the physical map.

Two PHRAP contigs that appear to conflict with available physical mapping data are retained in the diploid sequence (Table 1). These known discrepancies between the final diploid assembly and the physical map involve <1% of the genome. In some cases, PHRAP contigs span the major repeat sequence (MRS) of *C. albicans*. Assemblies across large repeats such as the MRS are probably not reliable.

Our assembled rDNA sequence (see additional data at http://genome-www.stanford.edu/candida-pnas2004-supplement) gives a repeating unit of 12,756 base pairs and indicates that the haploid genome encodes ≈55 copies of the shorter, intronless class of rDNA (see Table 1). The arrangement of the rRNA genes in strain SC5314 is similar to that in *Saccharomyces cerevisiae* with the addition of a low-complexity region of ≈2 kb. This region varies among strains and is used in various DNA typing schemes. Analysis of traces that contain partial rDNA sequences suggest that the repeat is located between supercontigs 10247 and 2511. Physical mapping data had previously placed the rDNA near markers on 2511.

As in *S. cerevisiae*, a relatively small fraction of *C. albicans* genes contain introns. Unlike some other fungal species, *C. albicans* does not appear to have extensively spliced genes. The

Table 1. Assignment of supercontigs by chromosome

Chromosome		Assigned supercontigs			Polymorphisms	
Name	Size, kb	Sum, kb	Count	Coverage, %	Count	Per kilobase covered
chr R	2,530	1,743	22	68.9	6,198	3.56
chr 1	3,165	3,034	26	95.9	12,969	4.27
chr 2	2,300	1,989	18	86.5	8,681	4.36
chr 3	1,820	1,666	10	91.6	2,962	1.78
chr 4	1,700	1,485	15	87.4	6,854	4.61
chr 5	1,230	1,081	10	87.9	10,254	9.48
chr 6	1,090	902	12	82.7	6,720	7.45
chr 7	1,020	950	8	93.1	1,640	1.73
Subtotal	14,855	12,851	121	86.5	56,278	4.38
Conflicts		120	2		298	2.48
Unmapped		1,879	143		5,958	3.17
Total	14,855	14,851	266	100.0	62,534	4.21

The haploid supercontig set totals 14,855 kb, which is very close to the estimated haploid genome size derived from physical criteria. Chromosome sizes were calculated from the summation of SfiI fragments assigned to each. The actual size of chr R is ≈ 3.2 Mb; in the table, 700 kb was deducted for its rDNA cluster (measured as *Bam*HI and *Xho*I fragments). The repeat is present only once in the assembly. The only significant difference from published maps relates to the SfiI fragments of chromosome 1 (24), where fragment J2 appears to be replaced by three small fragments. Also, some SfiI sites that are heterozygous in the mapping strain are absent or present on both homologs in SC5314. ''Conflicts'' gives statistics for the two PHRAP contigs that contained map markers from more than one chromosome; these contigs were passed unchanged into the diploid assembly. ''Unmapped'' refers to supercontigs not containing any map markers. Values in the table are rounded.

C. albicans intron structure is generally similar to that of *Saccharomyces*. *C. albicans* and its close relatives translate the codon CUG as serine rather than the usual leucine in nuclear genes (11). Approximately two-thirds of the ORFs make use of this unusual codon.

Heterozygosity. The diploid assembly highlights the extent of natural heterozygosity in *C. albicans*. The analysis described in *Methods* yielded a total of 62,534 high-confidence polymorphisms for the entire genome. Single base substitutions made up >89% of the high-confidence polymorphism set, with a 2:1 ratio of transitions to transversions (Table 7, which is published as supporting information on the PNAS web site). Homologs assembled separately by PHRAP account for 19% of the genome but contain 65% of the polymorphisms. The overall average frequency of polymorphism is one in 237 bases, considerably higher than observed in human or *Anopheles* sequence, probably in part due to our detection of separately assembled regions as homologs. The significance of the extensive allelic differences in *C. albicans* is unknown but may function to increase genetic diversity (12, 13) and contribute to the evolution of drug resistance (14).

The polymorphisms in the *C. albicans* genome, like those in human and *Anopheles*, are distributed quite unevenly across its genome. Table 1 lists the overall frequencies by chromosome. The excess polymorphisms on chromosomes 5 and 6 are explained by just a few highly diverged regions described below. The low overall polymorphism on chromosomes 3 and 7 results from very large nearly homozygous regions; Fig. 3 shows the distribution of polymorphisms along chromosome 7. The large nearly homozygous regions are near the telomeres, likely the result of mitotic recombination. Although the general location of the centromere is known from translocation data (B.B.M., unpublished data), the more polymorphic regions do not point to a more specific location.

We identified 11 highly polymorphic regions in the genome (Table 2). The largest of these is MTL, the mating-type-like locus. A second large polymorphic region is located ≈50 kb from MTL on the same supercontig. At this latter site, examination of the homologous supercontig indicates that it contains an inversion of sequence with otherwise low levels of polymorphism. The inversion is itself flanked by inverted repeats and could have occurred *in vivo* or as an outcome of the PHRAP assembly step. Localized inversions are a major feature of fungal genome evolution (15). Among the other highly polymorphic sites are a second inversion and known gene families containing low-complexity sequences.

Comparisons between the nucleotide sequence of homologous supercontigs reveal the presence of 82 large insertion or deletion (indel) polymorphisms (111–6,901 bp); these are generally located within intergenic regions. PHRAP contigs from our sequence data had previously been examined for dispersed repeated sequences (16). A number of these transposons and LTRs are heterozygous at certain loci. Fifteen of the indels contain ORFs encoding peptides >100 aa, with five related to previously described *C. albicans* transposable elements or the retrotransposons of *Drosophila*. Most ORFs encoded within indels are <200 amino acids and have no counterpart in current databases.

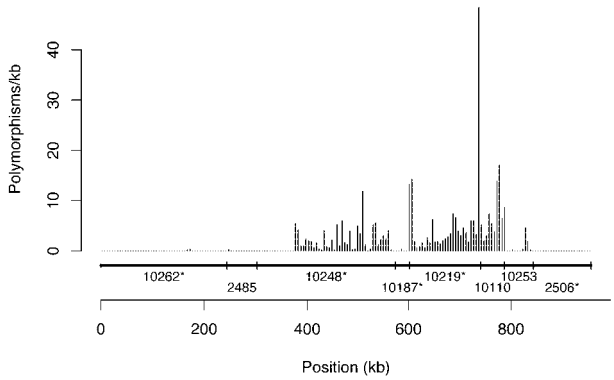

Fig. 3. Polymorphism distribution on chromosome 7. Shown are eight supercontigs accounting for 93% of the sequence of chromosome 7, ordered and oriented by physical map data. The orientation of supercontigs 10110 and 10253 is uncertain. The position of the polymorphism is shown on the *x* axis. The polymorphism position is assigned much like the base position except that insertion polymorphisms are given a position, and multibase deletions are collapsed to a single position. Bar heights show polymorphism frequency per kilobase in 5,000-position windows across the concatenated supercontigs.

GENETICS

Table 2. Highly polymorphic regions

Supercontig	Coordinates	Chromosome	Description	Gene
10035	37975–40874	6	Partial linked inverted duplication	MET22
10045	27175–29031	2	Low-complexity protein	
10150	35418–36351	1	Low-complexity protein	
10170	494–3725	5	Inversion	
10170	69722–78471	5	Substitution	MTL
10173	96750–97115	1	Substitution	
10196	9101–11664	2	Substitution	
10202	20748–21788	5	Partial linked duplication	
10219	362–1825	7	Partial unlinked duplication	HYR related
10233	79127–81042	6	Low-complexity protein	ALS related
10247	29766–30670	R	Inversion	

The polymorphic regions above were obtained by locating all regions of length at least 100 whose polymorphism score could not be increased either by expanding or shrinking the region. Polymorphic positions were scored +12 and nonpolymorhpic positions, −1.

Exceptions include a predicted protein related to oleate-inducible proteins of *Candida maltosa* and *Candida tropicalis* and another having sequence similarity to *Mycoplasma hominis* Lmp1p, a cell surface protein containing variable internal repeats. ORFs with large indels generally contain long low-complexity regions or have multiple internal repeats.

We found 3,579 ORFs containing high-confidence polymorphisms. In 2,792 of these, the polymorphisms alter protein translation. Among the protein differences, for 94 there was no ORF (100 amino acids or greater) on the homologous supercontig obviously encoding an allele, and for 57 others the ORF was fragmented into more than one ORF on the homologous supercontig. The effects of heterozygosity in *C. albicans* coding regions have not yet been extensively explored; however, significant phenotypic differences between parent strains and heterozygous mutants have been reported (17).

Among the 6,699 indels, there is a general decline in frequency with increasing length except at multiples of three bases. The excess of indels with length a multiple of three is concentrated almost completely in the coding fraction of the genome as defined by the reduced ORF set (Fig. 4). Three-base indels are not surprising in low-complexity regions of proteins such as homopolymer tracts.

It has been reported that a difference in the number of pentamer repeats in the upstream region of the *SAP2* gene function in regulating its expression (18). To understand the extent and distribution of tandem repeat (TR) variation in *C. albicans*, we scanned the genome for TRs of short sequences. The results for repeat unit sizes between 2 and 5 are summarized in Table 3. About half of the TRs are trinucleotides, and more than half of these are found within coding regions. For TRs of other sizes, the majority are found in noncoding regions. Except for dinucleotide repeats, once there are five TRs of any size, 9% or more of the loci have a different number of repeats on the homologous supercontig. Only a handful of the variable number of TR (VNTR) polymorphisms lead to frame-shift mutations in coding regions. The TR loci are also the sites of a significant number of non-VNTR polymorphisms.

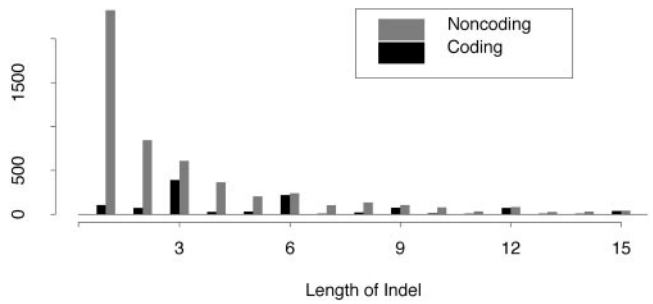

Fig. 4. Size distribution of indel polymorphisms up to 15 bp in coding and noncoding sequence. The coding fraction is determined from the reduced ORF set. Indel frequency in coding sequence decreases with increasing length, but multiples of three are present at higher frequency than other lengths. Overrepresentation of multiples of three nearly disappears in noncoding sequence.

Whole-Genome Comparisons. We used each ORF in the reduced haploid set to search *S. cerevisiae*, *Schizosaccharomyces pombe*, and human protein databases. Almost half of the ORFs, 3,027, found matches in all three genomes, with an additional 939 having matches in both other yeast genomes. With three genomes for comparison, only 22% of the ORFs lacked matches. It is noteworthy that nearly as many *C. albicans* genes matched only human genes as matched only *S. pombe* genes (83 vs. 91) (Fig. 5). Sixty-four percent of the ORFs have their best match in *Saccharomyces*; the remaining 14% of ORFs found matches to other more distantly related species than *S. cerevisiae*. With additional comparative genomes, one would expect this latter class to increase. As described below, examination of *C. albicans* genes not found in *S. cerevisiae*, or more closely related to genes in more distantly related species, highlights functions of potential significance for *C. albicans* pathogenesis.

The mitochondrial genome of *C. albicans* provides another example of how *C. albicans* is similar to more distantly related species than *S. cerevisiae*. The mtDNA encodes several NADH dehydrogenase subunits, which are more typical of eukaryotes but not found in *S. cerevisiae*. Translation of the encoded proteins indicates that the mitochondrial genetic code is likely to be the one used by *S. pombe* and filamentous fungi rather than the one used by *Saccharomyces*. Table 4 shows BLASTP scores when *C. albicans* ATPase subunit 6 is translated with the two genetic codes. Even though the *S. cerevisiae* gene is the closest match, the suggested genetic code is different.

Although *C. albicans* does not have large gene families obviously connected to antigenic variation, as do other pathogens, it does have a number of large gene families related to pathogenesis. These include the ALS (19), iron transport (20), secreted aspartyl proteinase (21), and secreted lipase (22) genes. Search of the haploid gene set reveals several additional gene families that may play a role in infection including oligopeptide transport (eight or nine genes), eight genes related to an estrogen-binding protein, and four acid sphingomyelinases. Also present are 12 cytochrome P450s, many more than in *S. cerevisiae* or *S. pombe*.

The most striking differences between *S. cerevisiae* and *C.*

Table 3. Variability of tandem repeats in the reference haploid genome

Unit size	Units	TR count	VNTR	Coding	VNTR and coding
2	6	449	38	11	0
2	7	266	43	7	0
2	8+	436	72	15	4
3	5	1,554	142	920	81
3	6	804	100	466	56
3	7+	908	191	554	109
4	4	458	35	20	3
4	5	217	28	7	2
4	6+	151	30	8	1
5	3	689	26	48	2
5	4	155	11	5	0
5	5+	107	26	7	0

The table was generated by scanning the reference haploid set of supercontigs for TRs and then examining polymorphisms at those locations. The size distribution of repeats in the reference haploid set is biased toward the larger allele because in the nearly homozygous regions, PHRAP generally reports the larger allele. The genome has additional VNTR loci where the allele in the reference haploid set was too small for inclusion in the table.

albicans are found in oxidative metabolism. In addition to common components in their electron transport chains, *C. albicans* also encodes a typical complex I. *C. albicans* has both the mitochondrially and nuclear encoded subunits of this complex found in most eukaryotes but absent in *S. cerevisiae*. An increased role for respiration in *C. albicans* is suggested by numerous differences, including a pyruvate dehydrogenase kinase to regulate the flow from glycolysis into the trichloroacetic acid cycle, the lipase family mentioned above and other enzymes in fatty acid catabolism, and additional amino acid catabolic pathways.

Sulfur metabolism appears also to differ between these two yeasts. *C. albicans* has genes likely to encode a direct pathway to cysteine in addition to a transsulfuration pathway from homocysteine. Genes encoding cysteine catabolic enzymes may also be present. These additional cysteine pathways might reflect an increased significance for glutathione metabolism in *C. albicans*.

The increased filamentation responses found in *C. albicans* would be expected to require alterations in genes for structural proteins and for cell cycle regulation. Among the differences in structural proteins, a kinesin-like gene most closely resembles the type found in *Aspergillus*. In the cell cycle, a number of differences from *S. cerevisiae* appear in the subunits of the anaphase-promoting complex.

Finally, the genome sequence reveals a number of adaptations for environmental sensing and response. *C. albicans*' ability to pass through the digestive tract requires it to cope with widely varying pH environments. *C. albicans* has a number of genes related to the pH regulatory genes of *Aspergillus* (23) and encodes a small family of chloride channels with members resembling types expressed in a variety of mammalian tissues. Also of note are differences in genes in the calmodulin signaling pathway, including a protein kinase related to one implicated in sensing surface contact in a plant pathogen.

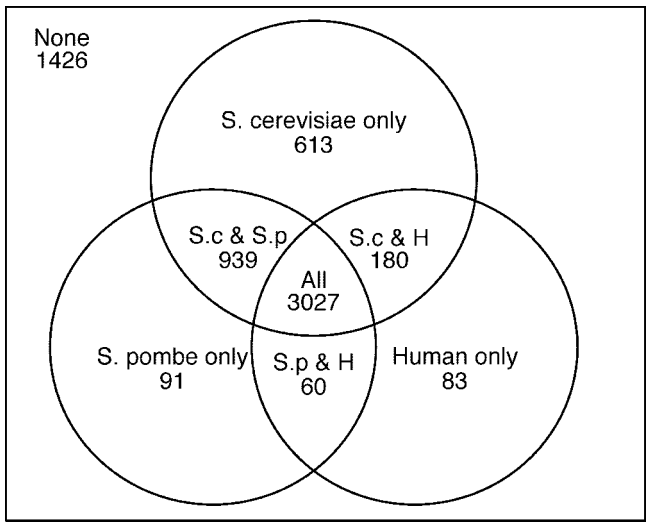

Fig. 5. Genome comparisons with other species. Matches of the 6,419 ORFs to human, *Saccharomyces*, and *Schizosaccharomyces* proteins. Shown are BLASTP hits (*E* value 10^{-8} or better) of *C. albicans* ORFs against *S. cerevisiae* (S.c), *S. pombe* (S.p), and human (H) protein sequences. The protein comparison sets are described in *Supporting Text*.

Discussion

Assembly of heterozygous diploid WGS sequence presents several challenges as compared with conventional genome assemblies, as well as opportunities for new types of analysis. These challenges and opportunities derive from the nature of the genome and are not readily avoided by taking a different approach to sequencing. For example, shotgun sequencing of bacterial artificial chromosome (BAC) clones eliminates diploid assembly in exchange for difficulties in constructing a tiling path in a diploid organism with diverged homologs. Both assembly and tiling path problems can be avoided by brute force, e.g., covering the genome redundantly (e.g., 7×) by BAC clones and then sequencing each clone to 10× shotgun coverage, but the cost of such an approach is very high. The benefits of directly addressing the assembly of diploid WGS sequence may extend to

Table 4. ATPase subunit 6 BLAST scores

Matching species	Code 3 (*Saccharomyces*)	Code 4 (others)
S. cerevisiae	612	681
Emericella nidulans	485	545
Neurospora crassa	468	535
S. pombe	465	522

The *C. albicans* protein was used to search proteins in other species assuming either the *S. cerevisiae* mitochondrial genetic code or the mitochondrial genetic code used in the other fungi. Highest raw scores are reported. Even though the *S. cerevisiae* protein is the most closely related, higher scores are obtained when the other code is used to translate the *C. albicans* gene. Similar results are obtained for COX3.

Table 5. Assembly as a function of sequence coverage

Assembly	19 (haploid)	6	5	4
Traces	—	313,165	211,629	170,725
Depth	—	10.9	7.3	5.7
Largest contig	421 kb	282 kb	151 kb	104 kb
Contigs >2 kb	239	1,213	1,680	1,631
Total size of contigs >2 kb, Mb	14.8	17.4	16.2	14.9
Contigs <2 kb, Mb	< 0.05	0.8	1.4	1.2

Trends in the assembly are given for increasing sequence coverage. The diploid assembly (19) is derived from the PHRAP assembly (6) with the additional steps described in the text so the trace/depth figures do not apply.

future sequencing efforts in organisms that are both diploid and heterozygous.

It is possible that additional sequence data might have closed some of the remaining true coverage gaps. From the eight chromosomes and the assembly gaps due to the copies of the MRS, one has ≈20 contigs as a lower bound. The remaining gaps have diverse origins in other repeat sequences, true gaps in the coverage, regions that may not be readily cloned in *Escherichia coli*, and overlaps too short for the conservative approach to joins we have used. We periodically assembled available sequence, and the results are summarized in Table 5. Increasing coverage yielded contigs whose sum clearly exceeded the haploid genome size; however, with assembly 6, there was a precipitous drop in the number of large contigs and in the total sequence contained within small contigs. The superassembly process continued these trends while delivering a product very close to independently derived estimates of the genome size.

Although PHRAP does not examine phase information, the diploid assembler places each PHRAP contig entirely into one or the other homologous supercontig. Our strategy was designed to preserve phase to the extent that it is preserved by PHRAP. Although additional coverage, increased read length, and double-end sequencing of clones would all identify more polymorphisms and assemble them with their correct phase, the presence of long homozygous regions, large repeat structures, and statistical limits derived from the sampling of the two homologs in the shotgun suggests diminishing returns from much higher levels of sequencing.

C. albicans biology also suggests that the limitations of the superassembly will not have severe practical effects. The only highly suspect areas of the sequence relate to the largest repeated sequences, particularly the ALS genes and the MRS (3, 19). Both of these sequence families have allelic and strain variation. Naturally occurring translocations via the MRS have been observed. Because much of the interest in *C. albicans* derives from the diversity of clinical isolates, the disproportionate effort required to assemble these sequences in one strain would have limited value.

Our diploid genome sequence catalogs polymorphisms in both protein encoding genes and potential regulatory sequences. This should greatly facilitate the search for additional loci where allelic differences are significant for pathogenesis. In addition to providing likely sites for regulation, variable numbers of TRs are useful markers for both population genetics and epidemiology.

The release of the *C. albicans* genome sequence to the public domain at various stages of completion has already accelerated research in the biology and disease processes of this important pathogen. The availability of a diploid genome sequence will now take these studies to a new level.

We thank Anja Forche, Suzanne Grindle, Alan Kuo, Paul Lephart, Curtis J. Palm, Audrey Southwick, and Lars Steinmetz for valuable discussions, technical contributions, and comments on the manuscript. Sequencing of *C. albicans* was supported by National Institute of Dental and Craniofacial Research Grant DE12302-02S2 and by the Burroughs Wellcome Fund. Additional work reported here was supported by National Institutes of Health Grants RO1AI16567, RO1AI46351, and NO1AI05406 (to P.T.M.) and R01DE12940 and P01DE07946 (to N.A.). We thank Bristol-Myers Squibb for making the SC5314 strain available to us without restrictions.

1. Poulter, R. T. (1987) *Crit. Rev. Microbiol.* **15,** 97–101.
2. Chibana, H., Magee, B. B., Grindle, S., Ran, Y., Scherer, S. & Magee, P. T. (1998) *Genetics* **149,** 1739–1752.
3. Chibana, H., Beckerman, J. L. & Magee, P. T. (2000) *Genome Res.* **10,** 1865–1877.
4. Hull, C. M., Raisner, R. M. & Johnson, A. D. (2000) *Science* **289,** 307–310.
5. Magee, B. B. & Magee, P. T. (2000) *Science* **289,** 310–313.
6. Tzung, K. W., Williams, R. M., Scherer, S., Federspiel, N., Jones, T., Hansen, N., Bivolarevic, V., Huizar, L., Komp, C., Surzycki, R., *et al*. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 3249–3253.
7. Fonzi, W. A. & Irwin, M. Y. (1993) *Genetics* **134,** 717–728.
8. Lander, E. S. & Waterman, M. S. (1988) *Genomics* **2,** 231–239.
9. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al*. (2001) *Science* **291,** 1304–1351.
10. Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., Wincker, P., Clark, A. G., Ribeiro, J. M., Wides, R., *et al*. (2002) *Science* **298,** 129–149.
11. Santos, M. A., Keith, G. & Tuite, M. F. (1993) *EMBO J.* **12,** 607–616.
12. Tsang, P. W., Cao, B., Siu, P. Y. & Wang, J. (1999) *Microbiology* **145,** 1623–1629.
13. Yesland, K. & Fonzi, W. A. (2000) *Microbiology* **146,** 2097–2104.
14. Cowen, L. E., Anderson, J. B. & Kohn, L. M. (2002) *Annu. Rev. Microbiol.* **56,** 139–165.
15. Seoighe, C., Federspiel, N., Jones, T., Hansen, N., Bivolarovic, V., Surzycki, R., Tamse, R., Komp, C., Huizar, L., Davis, R. W., *et al*. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 14433–14437.
16. Goodwin, T. J. & Poulter, R. T. (2000) *Genome. Res.* **10,** 174–191.
17. Kohler, J. R. & Fink, G. R. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 13223–13228.
18. Staib, P., Kretschmar, M., Nichterlein, T., Hof, H. & Morschhauser, J. (2002) *Mol. Microbiol.* **44,** 1351–1366.
19. Hoyer, L. L. (2001) *Trends Microbiol.* **9,** 176–180.
20. Ramanan, N. & Wang, Y. (2000) *Science* **288,** 1062–1064.
21. Hube, B. & Naglik, J. (2001) *Microbiology* **147,** 1997–2005.
22. Hube, B., Stehr, F., Bossenz, M., Mazur, A., Kretschmar, M. & Schafer, W. (2000) *Arch. Microbiol.* **174,** 362–374.
23. Davis, D., Wilson, R. B. & Mitchell, A. P. (2000) *Mol. Cell. Biol.* **20,** 971–978.
24. Chu, W. S., Magee, B. B. & Magee, P. T. (1993) *J. Bacteriol.* **175,** 6637–6651.