

The *Candida* Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data

Marek S. Skrzypek, Jonathan Binkley, Gail Binkley, Stuart R. Miyasato, Matt Simison and Gavin Sherlock*

Department of Genetics, Stanford University Medical School, Stanford, CA 94305-5120, USA

Received September 14, 2016; Accepted October 05, 2016

ABSTRACT

The *Candida* Genome Database (CGD, <http://www.candidagenome.org/>) is a freely available online resource that provides gene, protein and sequence information for multiple *Candida* species, along with web-based tools for accessing, analyzing and exploring these data. The mission of CGD is to facilitate and accelerate research into *Candida* pathogenesis and biology, by curating the scientific literature in real time, and connecting literature-derived annotations to the latest version of the genomic sequence and its annotations. Here, we report the incorporation into CGD of Assembly 22, the first chromosome-level, phased diploid assembly of the *C. albicans* genome, coupled with improvements that we have made to the assembly using additional available sequence data. We also report the creation of systematic identifiers for *C. albicans* genes and sequence features using a system similar to that adopted by the yeast community over two decades ago. Finally, we describe the incorporation of JBrowse into CGD, which allows online browsing of mapped high throughput sequencing data, and its implementation for several RNA-Seq data sets, as well as the whole genome sequencing data that was used in the construction of Assembly 22.

INTRODUCTION

The *Candida* Genome Database (CGD, <http://www.candidagenome.org/>) is a freely available online resource, based on the *Saccharomyces* Genome Database (SGD, www.yeastgenome.org; (1)), which collects, organizes, and makes available *Candida* gene, protein and sequence information to the fungal research community. CGD also provides web-based tools for the visualization and analysis of biological data.

Candida albicans is the most thoroughly studied of the human fungal pathogens and also serves as a model organism for the study of other more experimentally challenging fungal pathogens (2). The frequency of fungal infections has risen dramatically over the past few decades, with the current annual incidence of invasive *Candida* infections recently estimated to range between 72 and 228 infections per million (3,4). These infections have high mortality rates (>35%) and, despite the availability of antifungal drugs, exceed those associated with bacterial infections among intensive care unit patients (5). As a result of medical progress, particularly with respect to chemotherapy, organ transplantation, intensive care and neonatal care, the population of immunocompromised patients susceptible to fungal infection, is expanding (6–8), and resistance to antifungal drugs is increasing (3,9,10). Furthermore, *C. albicans* is not the only disease-causing species in the genus; of great concern is an emerging clinical prevalence of non-*albicans* *Candida* species (11,12). Among these, *C. tropicalis* is common, virulent, and increasingly resistant to antifungal therapy (13), *C. parapsilosis* is particularly problematic, causing severe infections in neonates (14), and *C. glabrata* exhibits a special ability to evade the immune system and survive after cellular engulfment, and can resist antifungal treatment (15–17).

In fungal taxonomy, *C. albicans* is placed in the order Endomycetales, to which the family Saccharomycetaceae also belongs (18,19), though *Saccharomyces cerevisiae* and *C. albicans* are separated by 140–850 million years of evolution (20). Despite sharing many similarities, including 3453 pairs of orthologous genes, the two fungi inhabit very different environmental niches, with *S. cerevisiae* existing as a saprophyte, and *C. albicans* living in close association with its mammalian hosts. Consequently, the two fungi show significant divergence in their genome structure and function. Extensive ‘transcriptional rewiring’ has occurred in their regulatory networks, such that evolved divergence in the activity of orthologous transcription factors, or divergence in *cis*-regulatory elements, has created distinct regulatory programs in each class (e.g. galactose metabolism

*To whom correspondence should be addressed. Tel: +1 650 498 6012; Email: gsherloc@stanford.edu

(21), mating type (22,23) and ribosomal proteins (24)). Unlike *S. cerevisiae*, *C. albicans* exists primarily in a diploid state, with no conventional sexual reproduction that leads to meiosis. Instead, *C. albicans* undergoes a parasexual cycle that involves mating between two diploid cells of the opposite mating type and formation of a tetraploid cell that subsequently returns to diploidy by chromosome loss (25). Alternatively, chromosome loss may occur prior to mating yielding transient haploids that mate and return to a diploid state (26). Meiosis has not been observed in either case, and the homologous chromosomes remain mostly intact making the haplotypes much more stable than in other organisms. Thus, in order to understand and combat *Candida* pathogenesis, it is important to study *Candida* biology directly, and having access to the phased genome assembly opens up multiple new research opportunities.

Genome history, and the incorporation of *C. albicans* assembly 22 into CGD

Since its inception in 2004, CGD has been tasked with maintenance of the primary genome sequence and annotation for the *C. albicans* reference strain SC5314. CGD was originally based upon Assembly 19, a diploid, contig-level assembly built from the genomic sequence that had just been completed (27), and which included gene models, sequence corrections, and functional annotations provided by a group of volunteer researchers known as the Annotation Working Group (28). CGD then took over the task of curating all the *Candida*-related information based on published experimental data, comparative genomics predictions and improved sequence information (29). In 2006, Assembly 20 was incorporated into CGD (30); Assembly 20 was the first chromosome level assembly for *C. albicans*, though it consisted of what are referred to as reftigs, wherein each of the chromosomes in the assembly consist of a mosaic of the two haplotypes (31). Assembly 20 was superseded by Assembly 21, because a sequence from a different strain, WO-1, had been inadvertently incorporated into Assembly 20. A major update to Assembly 21, based on additional sequence data and comparative analyses (32) was subsequently also incorporated into CGD. The most recent major update to the sequence is Assembly 22 (33), which was generated using Illumina-based sequencing of SC5314 as well as a collection of congenic strains homozygous for specific chromosomes, which allowed the construction of a phased diploid assembly of the genome. Assembly 22 was incorporated into CGD and became the default genome sequence in 2014. All the previous assemblies, however, remain available at the CGD website.

Incorporation of Assembly 22 into CGD entailed careful reconciliation of sequence ambiguities, potential assembly errors and differences from previous assemblies. We have established a pipeline that takes advantage of all available data, including the Illumina-based sequences from Muzzey et al. (33), Roche 454 sequences generated at CGD, as well as independently published, gene-specific sequence data derived from our curation of the *Candida* literature. We have used this pipeline to correct the sequences for more than 300 features, eliminating many reading frame errors and resolving multiple sequence ambiguities (Table 1). We have

Table 1. Number of Assembly 22 features corrected, by error type (in either or both haplotypes)

Error Type	Features Corrected
Boundary/annotation	455
Ambiguous sequence	268
Nonsense codons	48
Missing stop codons	46
Misc. coding sequence	8
Misc. non-coding sequence	5
Missing start codons	2

also re-analyzed the haplotype assignments for the entire Chromosome 3, which led to an exchange of 845 kb between chromosome 3A and chromosome 3B. Improving the reference sequence is an ongoing project and we are currently focusing our efforts on intergenic regions, repetitive sequences and segmental duplications.

Sequence versioning

Since 2010, CGD has used a versioning system to track genome sequence releases, and their associated annotations. The version designation appears in the name of each of the relevant sequence and feature annotation files that are available at CGD, so the exact source of the sequence data is always clear. Version designations appear in the following format: sXX-mYY-rZZ where XX, YY, and ZZ are zero-padded integers. XX is incremented when there is any change to the underlying genomic (i.e., chromosome) sequence. YY is incremented when there is any change to the coordinates of any feature annotated in the genome (e.g. any change in location or boundary, or addition or removal of a feature from the annotation). YY is reset to '01' when XX is incremented (when a sequence change is made). ZZ is incremented in response to curatorial changes that affect information that appears in the GFF file, specifically gene names, gene aliases, gene IDs, gene descriptions, feature types (e.g. gene or pseudogene), and ORF classifications or qualifiers (e.g. Verified, Uncharacterized, Deleted, Merged). Files are checked on a weekly basis, as well as any time that a GFF file is regenerated manually, to determine if changes have occurred that warrant a change in the ZZ number. ZZ is reset to '01' when either XX or YY is incremented (when a sequence change is made, or when the coordinates of any feature are updated). All versions are archived on the CGD download site.

As a hypothetical example, say that we start with s05-m01-r01 as the current version. When the next weekly file check is performed, and the new file is noted to contain curatorial updates to gene names in the database, but no new changes to the structural annotation or to the sequence itself, the new version designation becomes s05-m01-r02. Subsequently, the chromosomal coordinates of a gene are changed, based on curation of a paper that provides evidence for updating the gene model. Consequently, the new version designation becomes s05-m02-r01. Later, a change to correct a sequencing error is made, and the new version designation becomes s06-m01-r01. CGD recommends that authors note in their materials and methods sections of published papers what version of the genome they were working

with when they performed an analysis, so that the analysis is reproducible.

Systematic ORF nomenclature

With Assembly 22, it became possible to implement a new position based systematic nomenclature for chromosomal features. The previous systematic names, dating back to Assembly 19, consisted of the ‘orf19’ prefix followed by a unique but somewhat arbitrary number. The new systematic name is based on the known chromosomal location and haplotype, and it consists of the chromosome (C1-C7 and CR for the eight nuclear chromosomes, CM for the mitochondrial chromosome), a unique number indicating the order of features along chromosomes, the strand (W for Watson or C for Crick) and the haplotype (A or B). For example, C4_03570W_A denotes a feature located on chromosome 4, Watson strand and haplotype A. Feature numbers start at the left end of the chromosome and increase by 10 to allow for adding new features in the intervening spaces as they are discovered. Since systematic identifiers from all the previous assemblies remain in the literature, and some, especially ‘orf19’ names, continue to be used by researchers, CGD includes all the previous identifiers as searchable aliases to allow seamless and unambiguous transition between various nomenclature systems. Assembly 19/21 identifiers, ‘orf19 names’, are also prominently displayed on each Locus Summary page.

Locus Summary pages (LSP) exist for every chromosomal feature in CGD and provide a primary way to access gene-specific information. In addition to locus-specific data, such as description, functional annotations (Gene Ontology terms, mutant phenotypes), or orthologous genes in other species, LSPs now also provide access to allele-specific data: DNA and protein sequences for both alleles, as well as a listing of allelic variations if they exist. LSPs also show graphical representations of the chromosomal context for each allele as thumbnails that lead to genome browser windows and allow farther exploration of a chosen haplotype.

Assembly 22 allows powerful, allele- and haplotype-specific analyses of the overall genome structure, function and evolution. To facilitate genome-wide research, CGD provides access to all the data in downloadable files that can be used by researchers’ own bioinformatics tools. All DNA sequences for the entire chromosomes, coding and non-coding regions, as well as translation products, are available for download at <http://www.candidagenome.org/download/>.

Incorporation of JBrowse

An important part of CGD’s mission is to incorporate and annotate large-scale datasets from published experiments on *Candida* species, and to make them easily available for exploration and analysis by our users. For several years we have provided GeneXplorer (34) to display and analyze microarray datasets. Recently, to provide access to the growing number of datasets from experiments that use high-throughput sequencing technologies, we have deployed the JBrowse genome browser (35). JBrowse is a mature and widely used application that is fast, intuitive, and compatible with most web browsers (36). JBrowse allows users to

quickly view large-scale sequence data in a genomic context, at multiple zoom-levels of resolution, from base pairs in individual sequence reads to read-density summaries across large genomic regions. The display includes parallel tracks of annotated sequence features, allowing seamless navigation between JBrowse and the Locus Summary pages for each feature (Figure 1). Quantitative tracks graphically display comparative information, such as relative expression level or sequence conservation. JBrowse is highly flexible and customizable: users may easily load their own sequence datasets and analysis tracks, for display in the context of genomic features, or for comparison with datasets and tracks provided by CGD.

CGD currently offers a number of *C. albicans* datasets for viewing in JBrowse. We provide the high-throughput DNA sequence data (33) that was the basis of Assembly 22, described above. Optional tracks highlight the sequence variation between the two Assembly 22 haplotypes, as well as the variation between the common strains SC5314 and WO-1 (32). We also make available RNA-Seq datasets from a number of gene expression studies in *C. albicans*, including comparisons of different stress conditions (37), hyphae-inducing conditions (37), biofilm vs. planktonic growth (38), white-opaque switching (39), and allele-specific expression differences (40). In addition, we provide chromatin occupancy data (ChIP-Seq) for the Wor1p transcription factor during white-opaque switching (41). We also have RNA-Seq datasets for two other *Candida* species: gene expression under pH and nitrosative stress in *C. glabrata* (42), and during biofilm vs. planktonic growth in *C. parapsilosis* (43). We will add new datasets to JBrowse as they become available, and in response to user requests.

Future directions

The reference sequence for an organism is not static—while sequencing technologies continue to advance (both in accuracy and read lengths), there is always the possibility that updates will improve the reference sequence. For example, either PacBio or Oxford Nanopore sequencing (very long reads, though with high error rates), coupled with existing Illumina data (short reads, but low error rates) may better resolve telomeric and other repeat sequences, which are hard to resolve even with the original Sanger reads. Another question is whether the sequence of a single instance of a strain is a reasonable representation of the reference sequence – indeed, given *Candida albicans*’ propensity to undergo rearrangement and loss of heterozygosity under stress, different lab isolates of ostensibly the same strain might have different sequences. It is likely in the near future that many distinct strains of *Candida* species will be sequenced, and CGD will endeavor to incorporate these into the database as well. They will provide insight into the genomic variation that exists in each of the *Candida* species.

Additional high throughput sequencing datasets may also allow refinement of the genome annotations, such as novel additional transcripts, or the addition of 5’ and 3’ UTRs to each of the genes. They may also allow the annotation of functional elements within non-transcribed regions, such as transcription factor binding sites from ChIP-Seq studies. CGD will strive to incorporate such refinements

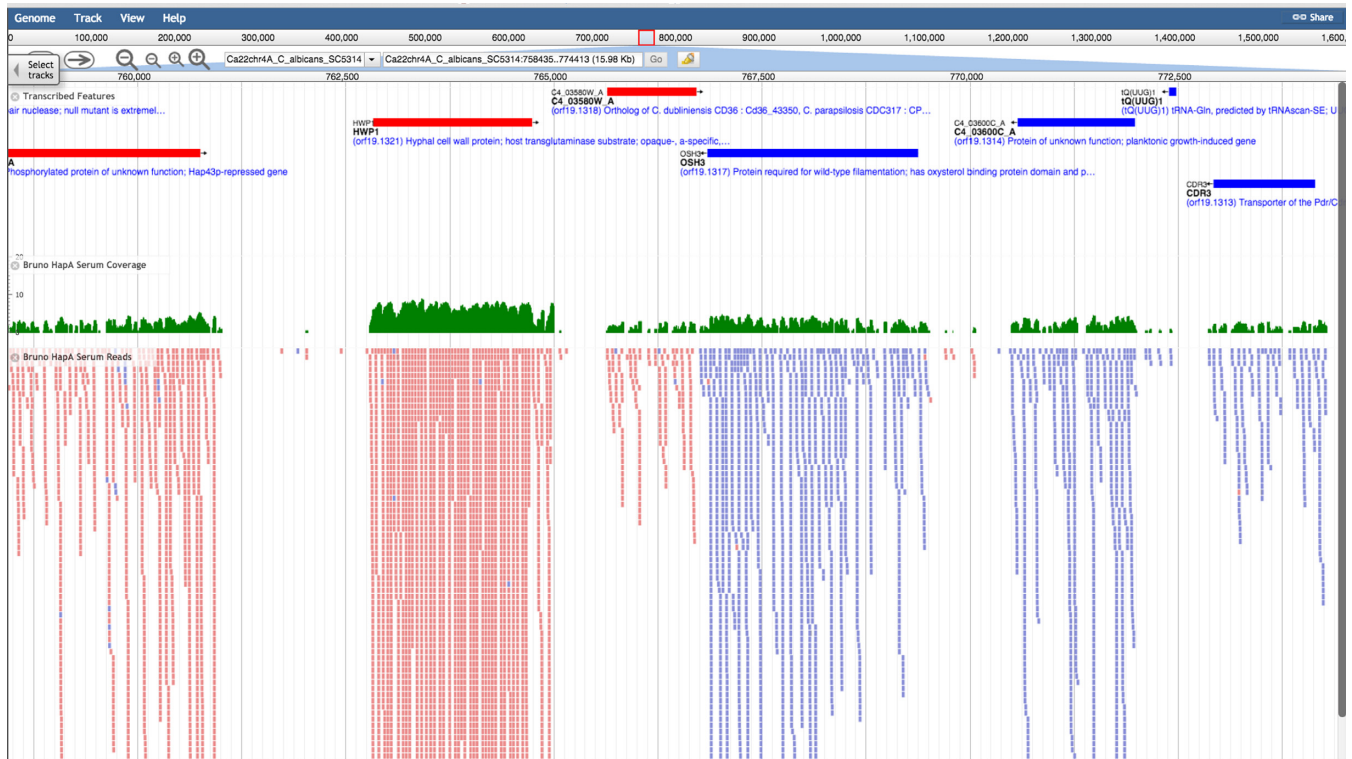


Figure 1. JBrowse Visualization of RNA-Seq data at CGD. JBrowse display of the region around the *C. albicans* serum-inducible gene *HWP1*, showing aligned RNA-Seq reads from serum-treated cells (37). The red and blue bars in the top track of the main display window show genes annotated at CGD: red for genes encoded on the 'W' strand (+), blue for genes on the 'C' strand (-). *HWP1* is the second gene from the left. Clicking on a bar brings up an information window for that gene, and includes a link to its CGD Locus Summary Page. The green bar graph below the gene track shows the density of aligned RNA-Seq reads along to the chromosome, plotted on a log scale. The bottom track shows all the aligned RNA-Seq reads along the chromosome: each short bar in the bottom track represents a unique read. In this example the sequence reads are strand-specific: pink bars indicate reads transcribed from the 'W' strand, and light-blue bars indicate reads transcribed from the 'C' strand. Clicking on a bar brings up information about the read, including the sequence and quality score for each base. Menus and controls at the top of the browser provide navigation, zoom and search functionalities, and allow users to load their own data.

as soon as they become available to maintain the reference sequence and annotation current.

ACKNOWLEDGEMENTS

CGD thanks the *Candida* research community for their advice, support and enthusiasm.

FUNDING

National Institute of Dental and Craniofacial Research at the US National Institutes of Health [R01 DE015873]. Funding for open access charge: NIH.

Conflict of interest statement. None declared.

REFERENCES

- Sheppard, T.K., Hitz, B.C., Engel, S.R., Song, G., Balakrishnan, R., Binkley, G., Costanzo, M.C., Dalusag, K.S., Demeter, J., Hellerstedt, S.T. *et al.* (2016) The *Saccharomyces* genome database variant viewer. *Nucleic Acids Res.*, **44**, D698–D702.
- Kabir, M.A., Hussain, M.A. and Ahmad, Z. (2012) *Candida albicans*: a model organism for studying fungal pathogens. *ISRN Microbiol.*, **2012**, 1–12.
- Pfaller, M.A. and Diekema, D.J. (2007) Epidemiology of invasive candidiasis: a persistent public health problem. *Clin. Microbiol. Rev.*, **20**, 133–163.
- Wilson, L.S., Reyes, C.M., Stolpman, M., Speckman, J., Allen, K. and Beney, J. (2002) The direct cost and incidence of systemic fungal infections. *Value Health*, **5**, 26–34.
- Kett, D.H., Azoulay, E., Echeverria, P.M., Vincent, J.L. and Extended Prevalence of Infection in, I.C.U.S.G.o.I. (2011) *Candida* bloodstream infections in intensive care units: analysis of the extended prevalence of infection in intensive care unit study. *Crit. Care Med.*, **39**, 665–670.
- Fisher-Hoch, S.P. and Hutwagner, L. (1995) Opportunistic candidiasis: an epidemic of the 1980s. *Clin. Infect. Dis.*, **21**, 897–904.
- Groll, A.H., De Lucca, A.J. and Walsh, T.J. (1998) Emerging targets for the development of novel antifungal therapeutics. *Trends Microbiol.*, **6**, 117–124.
- Mayer, F.L., Wilson, D. and Hube, B. (2013) *Candida albicans* pathogenicity mechanisms. *Virulence*, **4**, 119–128.
- Giri, S. and Kindo, A.J. (2012) A review of *Candida* species causing blood stream infection. *Indian J. Med. Microbiol.*, **30**, 270–278.
- Schelenz, S. (2008) Management of candidiasis in the intensive care unit. *J. Antimicrob. Chemother.*, **61**(Suppl. 1), i31–i34.
- Mikulska, M., Del Bono, V., Ratto, S. and Viscoli, C. (2012) Occurrence, presentation and treatment of candidemia. *Expert. Rev. Clin. Immunol.*, **8**, 755–765.
- Sipsas, N.V. and Kontoyiannis, D.P. (2012) Invasive fungal infections in patients with cancer in the Intensive Care Unit. *Int. J. Antimicrob. Agents*, **39**, 464–471.
- Negri, M., Silva, S., Henriques, M. and Oliveira, R. (2012) Insights into *Candida tropicalis* nosocomial infections and virulence factors. *Eur. J. Clin. Microbiol. Infect. Dis.*, **31**, 1399–1412.
- Chow, B.D., Linden, J.R. and Bliss, J.M. (2012) *Candida parapsilosis* and the neonate: epidemiology, virulence and host defense in a unique patient setting. *Expert Rev. Anti Infect. Ther.*, **10**, 935–946.

15. Brunke,S. and Hube,B. (2013) Two unlike cousins: *Candidaalbicans* and *C. glabrata* infection strategies. *Cell Microbiol.*, **15**, 701–708.
16. Lewis,R.E., Viale,P. and Kontoyiannis,D.P. (2012) The potential impact of antifungal drug resistance mechanisms on the host immune response to *Candida*. *Virulence*, **3**, 368–376.
17. Pfaller,M.A. (2012) Antifungal drug resistance: mechanisms, epidemiology, and consequences for treatment. *Am. J. Med.*, **125**, S3–S13.
18. Chen,M.W., Anne,J., Volckaert,G., Huysmans,E., Vandenberghe,A. and De Wachter,R. (1984) The nucleotide sequences of the 5 S rRNAs of seven molds and a yeast and their use in studying ascomycete phylogeny. *Nucleic Acids Res.*, **12**, 4881–4892.
19. Hendriks,L., Goris,A., Van de Peer,Y., Neefs,J.M., Vancanneyt,M., Kersters,K., Hennebert,G.L. and De Wachter,R. (1991) Phylogenetic analysis of five medically important *Candida* species as deduced on the basis of small ribosomal subunit RNA sequences. *J. Gen. Microbiol.*, **137**, 1223–1230.
20. Berman,J. and Sudbery,P.E. (2002) *Candida albicans*: a molecular revolution built on lessons from budding yeast. *Nat. Rev.*, **3**, 918–930.
21. Martchenko,M., Levitin,A., Hogues,H., Nantel,A. and Whiteway,M. (2007) Transcriptional rewiring of fungal galactose-metabolism circuitry. *Curr. Biol.*, **17**, 1007–1013.
22. Baker,C.R., Tuch,B.B. and Johnson,A.D. (2011) Extensive DNA-binding specificity divergence of a conserved transcription regulator. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 7493–7498.
23. Tsong,A.E., Tuch,B.B., Li,H. and Johnson,A.D. (2006) Evolution of alternative transcriptional circuits with identical logic. *Nature*, **443**, 415–420.
24. Ihmels,J., Bergmann,S., Gerami-Nejad,M., Yanai,I., McClellan,M., Berman,J. and Barkai,N. (2005) Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science (New York, N. Y.)*, **309**, 938–940.
25. Bennett,R.J. and Johnson,A.D. (2003) Completion of a parasexual cycle in *Candida albicans* by induced chromosome loss in tetraploid strains. *EMBO J.*, **22**, 2505–2515.
26. Hickman,M.A., Zeng,G., Forche,A., Hirakawa,M.P., Abbey,D., Harrison,B.D., Wang,Y.M., Su,C.H., Bennett,R.J., Wang,Y. *et al.* (2013) The 'obligate diploid' *Candida albicans* forms mating-competent haploids. *Nature*, **494**, 55–59.
27. Jones,T., Federspiel,N.A., Chibana,H., Dungan,J., Kalman,S., Magee,B.B., Newport,G., Thorstenson,Y.R., Agabian,N., Magee,P.T. *et al.* (2004) The diploid genome sequence of *Candida albicans*. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 7329–7334.
28. Braun,B.R., van Het Hoog,M., d'Enfert,C., Martchenko,M., Dungan,J., Kuo,A., Inglis,D.O., Uhl,M.A., Hogues,H., Berriman,M. *et al.* (2005) A human-curated annotation of the *Candida albicans* genome. *PLoS Genet.*, **1**, 36–57.
29. Arnaud,M.B., Costanzo,M.C., Skrzypek,M.S., Binkley,G., Lane,C., Miyasato,S.R. and Sherlock,G. (2005) The *Candida* Genome Database (CGD), a community resource for *Candida albicans* gene and protein information. *Nucleic Acids Res.*, **33**, D358–D363.
30. Arnaud,M.B., Costanzo,M.C., Skrzypek,M.S., Shah,P., Binkley,G., Lane,C., Miyasato,S.R. and Sherlock,G. (2007) Sequence resources at the *Candida* Genome Database. *Nucleic Acids Res.*, **35**, D452–D456.
31. van het Hoog,M., Rast,T.J., Martchenko,M., Grindle,S., Dignard,D., Hogues,H., Cuomo,C., Berriman,M., Scherer,S., Magee,B.B. *et al.* (2007) Assembly of the *Candida albicans* genome into sixteen supercontigs aligned on the eight chromosomes. *Genome Biol.*, **8**, R52.
32. Butler,G., Rasmussen,M.D., Lin,M.F., Santos,M.A., Sakthikumar,S., Munro,C.A., Rheinbay,E., Grabherr,M., Forche,A., Reedy,J.L. *et al.* (2009) Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature*, **459**, 657–662.
33. Muzzey,D., Schwartz,K., Weissman,J.S. and Sherlock,G. (2013) Assembly of a phased diploid *Candida albicans* genome facilitates allele-specific measurements and provides a simple model for repeat and indel structure. *Genome Biol.*, **14**, R97.
34. Rees,C.A., Demeter,J., Matese,J.C., Botstein,D. and Sherlock,G. (2004) GeneXplorer: an interactive web application for microarray data visualization and analysis. *BMC Bioinformatics*, **5**, 141.
35. Skinner,M.E., Uzilov,A.V., Stein,L.D., Mungall,C.J. and Holmes,I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.
36. Buels,R., Yao,E., Diesh,C.M., Hayes,R.D., Munoz-Torres,M., Helt,G., Goodstein,D.M., Elisk,C.G., Lewis,S.E., Stein,L. *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66.
37. Bruno,V.M., Wang,Z., Marjani,S.L., Euskirchen,G.M., Martin,J., Sherlock,G. and Snyder,M. (2010) Comprehensive annotation of the transcriptome of the human fungal pathogen *Candida albicans* using RNA-seq. *Genome Res.*, **20**, 1451–1458.
38. Desai,J.V., Bruno,V.M., Ganguly,S., Stamper,R.J., Mitchell,K.F., Solis,N., Hill,E.M., Xu,W., Filler,S.G., Andes,D.R. *et al.* (2013) Regulatory role of glycerol in *Candida albicans* biofilm formation. *MBio*, **4**, e00637–00612.
39. Xie,J., Tao,L., Nobile,C.J., Tong,Y., Guan,G., Sun,Y., Cao,C., Hernday,A.D., Johnson,A.D., Zhang,L. *et al.* (2013) White-opaque switching in natural MTL α /alpha isolates of *Candida albicans*: evolutionary implications for roles in host adaptation, pathogenesis, and sex. *PLoS Biol.*, **11**, e1001525.
40. Muzzey,D., Sherlock,G. and Weissman,J.S. (2014) Extensive and coordinated control of allele-specific expression by both transcription and translation in *Candida albicans*. *Genome Res.*, **24**, 963–973.
41. Lohse,M.B. and Johnson,A.D. (2016) Identification and Characterization of Wor4, a New Transcriptional Regulator of White-Opaque Switching. *G3 (Bethesda)*, **6**, 721–729.
42. Linde,J., Duggan,S., Weber,M., Horn,F., Sieber,P., Hellwig,D., Riege,K., Marz,M., Martin,R., Guthke,R. *et al.* (2015) Defining the transcriptomic landscape of *Candida glabrata* by RNA-Seq. *Nucleic Acids Res.*, **43**, 1392–1406.
43. Holland,L.M., Schroder,M.S., Turner,S.A., Taff,H., Andes,D., Grozer,Z., Gacser,A., Ames,L., Haynes,K., Higgins,D.G. *et al.* (2014) Comparative phenotypic analysis of the major fungal pathogens *Candida parapsilosis* and *Candida albicans*. *PLoS Pathog.*, **10**, e1004365.