Utilization of Hybrid Assembly Approach to Determine the Genome of an Opportunistic Pathogenic Fungus, *Candida albicans* TIMM 1768

Suresh Panthee[1], Hiroshi Hamamoto[1], Sanae A. Ishijima[1], Atmika Paudel[1], and Kazuhisa Sekimizu[1,2,*]

[1]Institute of Medical Mycology, Teikyo University, Hachioji, Tokyo, Japan

[2]Genome Pharmaceuticals Institute Co., Ltd, Bunkyo, Tokyo, Japan

*Corresponding author: E-mail: sekimizu@main.teikyo-u.ac.jp.

Accepted: July 28, 2018

Data deposition: This whole-genome project has been deposited at NCBI BioProject under the accession PRJNA471195.

Abstract

Candida albicans TIMM1768 is a highly virulent strain utilized as a model organism for the study of gastrointestinal and oral candidiasis. Despite being a model strain, identification of its genetic determinants of pathogenesis is hindered by the unavailability of its genome sequence. In this study, we determined the genome sequence of *C. albicans* TIMM1768 using reads obtained from portable MinION and benchtop Ion PGM sequencers. Genome annotation and a comparative analysis with published genomes revealed that the TIMM1768 strain was close to *Candida albicans* CHN1, and we identified a significant number of genes encoding for pathogenesis. The availability of the *C. albicans* TIMM1768 genome will facilitate comparative genomic analysis of Candida species, including studies of its virulence mechanisms and the development of treatment strategies for severe candidiasis.

Key words: *Candida albicans*, de novo assembly, genomics, pathogenesis, MinION sequencing.

Introduction

Candida albicans, a fungus usually present in the gastrointestinal tract, mouth, and genital tract of human beings as a part of the human microbiome, is capable of causing both superficial and systemic infections, especially in patients with compromised immune function. Recently, there has been a tremendous rise in the fungal infections caused by candida (Odds 2010). Enhancing our understanding of the biological and pathogenic features of the fungus will significantly advance the discovery of novel approaches to counteract the infections caused by this organism. In the pregenomic era, *C. albicans* virulence studies were performed by the functional elucidation of a few candidate genes. Global studies were recently facilitated by the availability of the *C. albicans* genome sequence and functional genomic approaches (Anderson and Bennett 2016).

Candida albicans TIMM1768 is a highly virulent strain isolated from the feces of a candidiasis patient. Hyphal growth is associated with its virulence and N-acetylglucosamine (GlcNAc), an inducer of the hyphal growth, leads to severe colonization in the oral and gastrointestinal models of TIMM1768 candidiasis (Mizutani et al. 2000; Hisajima et al. 2008; Ishijima et al. 2012; Ishijima and Abe 2015). Despite noteworthy progress in elucidating TIMM1768 pathogenicity, the genetic element responsible for its high pathogenicity has yet to be identified. One approach to elucidate this phenotype is to mine the genome sequence for the genes with a potential pathogenic role, such as virulence factors and proteases. The last decade has seen significant development in next-generation sequencers that produce high-throughput short reads. Although this has greatly advanced genomic and transcriptomic studies, fungal genome assembly remains challenging, due in part to the high repeat content.

The Oxford Nanopore Technologies third-generation sequencer MinION is a new sequencing technology that offers ultra-long reads and can be particularly beneficial for resolving repeat-rich regions. Further, MinION has a low capital cost, produces data in real-time, and is portable. Because the availability of the genome is essential for genome mining, here, we sequenced the genome of *C. albicans* TIMM1768 by combining Oxford Nanopore long reads and Ion PGM short reads. The 14.4-Mb TIMM1768 genome was assembled into eight

chromosomes and one mitochondrion. Annotation and comparative analysis revealed 424 genes with a potential role in pathogenesis.

Materials and Methods

Strain and Minimum Inhibitory Concentration Assay

The strain used in this study was *C. albicans* TIMM1768. The minimum inhibitory concentration of antifungal agents was determined based on CLSI M27-A3 using a commercial kit (cat. 9FEF27, Eiken Chemical Co., Ltd., Japan).

DNA Extraction

Candida albicans TIMM1768 was cultured on YPD medium, and the genomic DNA from 300 µl of culture broth was isolated with a Qiagen DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) using zymolyase for cell lysis.

Library Preparation and Oxford Nanopore Sequencing

The library preparation for Oxford Nanopore sequencing was performed using the manufacturer's recommended protocol—1D Genomic DNA by ligation for the SQK-LSK108 kit (Oxford Nanopore Technologies, Oxford, UK). Briefly, 1 µg genomic DNA was end-prepped using the NEBNext End repair/dA-tailing Module (New England Biolabs, Inc., Ipswich, MA, USA) and the DNA was cleaned up using Agencourt AMPure XP (Beckman Coulter Inc., Brea, CA, USA). The DNA was then ligated to the adapter using NEB Blunt/TA Ligase Master Mix (NEB). The adapted library was purified using Agencourt AMPure XP (Beckman Coulter Inc.) and applied to a primed FLO-MIN106 R9.4 SpotON Flow Cell attached to MinION (Oxford Nanopore Technologies). The sequence reads were obtained with MinKNOW software using the 48-h protocol and live base-calling.

Library Preparation and Ion PGM Sequencing

The library for the Ion semiconductor sequencing was prepared as described previously (Panthee et al. 2017a, 2017b). Briefly, the barcoded library of 400 base-reads was prepared after fragmentation of 100 ng of the DNA using the Ion Xpress™ Plus Fragment Library Kit (ThermoFisher Scientific, Waltham, MA, USA). The libraries were enriched in an Ion 318™ Chip v2 using Ion Chef (ThermoFisher Scientific), and subsequent sequencing was performed in the Ion PGM System (ThermoFisher Scientific).

Read Correction and Genome Assembly

We obtained 10.8 M reads (mean length 216 bp) from Ion PGM, and 129 K reads from MinION (mean length 8.3 kb), accounting for approximately 165-fold and 75-fold genome coverage, respectively. Genome assembly was performed using Flye 2.3.3 assembler (Kolmogorov et al. 2018). Briefly, the short reads from Ion PGM were corrected using SPADES 3.11 (Bankevich et al. 2012), and the high quality MinION long reads were filtered using Filtlong followed by self-correction and trimming using canu 1.7 (Koren et al. 2017). The short reads were then used for error correction of the long trimmed reads using LoRDEC (Salmela and Rivals 2014). The final assembly of the corrected reads was then performed with the Flye genome assembler. Further polishing of the assembly was performed by mapping the short reads to the assembly followed by consensus generation. Scaffolding of the contigs was performed on the MeDuSa webserver (Bosi et al. 2015) using *Candida albicans* SC5314 genome as reference.

Comparative Genomic Analysis

The concatenated sequence of seven housekeeping genes (Bougnoux et al. 2003) used for multi-locus sequence typing was utilized to perform the comparative analysis. The TIMM1768 genome and assemblies available in the NCBI database were blasted against the trimmed sequence of the seven genes, namely *AAT1a, ACC1*, *ADP1*, *MPIb*, *SYA1*, *VPS13*, and *ZWF1b*. The extracted sequences were then concatenated in order. The concatenated sequences of the 3,483 sequence types were downloaded from PubMLST (https://pubmlst.org/calbicans/; last accessed August 6, 2018). Finally, the sequences were aligned, and the phylogenetic tree was constructed using CLC Genomics Workbench ver 11.0 (CLC bio, Aarhus, Denmark). For comparative analysis of the mitochondrial genome, six complete mitochondrial genomes were downloaded from NCBI and analyzed for genome rearrangement using ProgressiveMauve (Darling et al. 2010). A phylogenetic tree was constructed after aligning the sequences in CLC Genomics Workbench ver 11.0 (CLC bio).

To determine the extent of sequence divergence among orthologs, the genome sequences of *Candida dubliniensis* CD36 (GCA_000026945.1) and *Saccharomyces cerevisiae* S288C (GCA_000146045.2) were downloaded from NCBI. The protein sequences of TIMM1768, CD36 and S228C were then submitted to OrthoFinder 2.2.6 (Emms and Kelly 2015) to identify orthologous groups. The orthologous sequences were aligned using ParaAT 1.0 (Zhang et al. 2012) and sequence divergence was calculated using KaKs_calculator (Zhang et al. 2006).

Data Availability

This whole-genome project has been deposited at NCBI BioProject under the accession PRJNA471195.

Results and Discussion

The Antibiogram of TIMM1768

The antibiotic resistance profile of the clinical isolate *Candida albicans* TIMM1768 was tested against the most commonly

used antifungal agents. The strain exhibited sensitivity to all of the tested agents (supplementary table 1, Supplementary Material online).

Sequencing and de novo Genome Assembly

This study aimed to achieve the de novo assembly of *C. albicans* TIMM1768. The sequencing and assembly workflow is explained in the supplementary scheme 1, Supplementary Material online. To examine the efficiency of assembly, we mapped short reads to the assembly followed by variant detection, where the final assembly had 444 variants, a value comparable to 302 variants of the short reads only assembly (supplementary table 2, Supplementary Material online). Assembly of the short reads alone resulted in a total of 3,398 contigs with an L_{50} of 119 and the nanopore-only reads resulted in 113,662 variants indicating the higher error rates in the ONT MinION reads. The assembly of long reads after correction and consensus generation drastically improved the assembly and successfully circularized the 40-kb mitochondrial genome. Scaffolding by using *C. albicans* SC5314 chromosome as reference resulted in seven linear chromosomes (# of runs of Ns 2) and a circular mitochondrion (supplementary fig. 1, Supplementary Material online). The key assembly parameters, that is, the largest contig, N_{50}, and L_{50}, were 3.1, 1.7, and 3 Mb, respectively (supplementary table 2, Supplementary Material online).

Analysis of the TIMM1768 Genome

The assembled chromosome was analyzed using Yeast Genome Annotation Pipeline (Proux-Wéra et al. 2012) to reveal 5,649 coding sequences (CDSs). Functional annotation of the CDSs was performed in Blast2GO ver 5.1.13 (BioBam Bioinformatics, Valencia, Spain). Of the 5,626 (99.6%) CDSs matched to known protein sequences in the fungal Nr database, Gene Ontology (GO) terms were assigned to 4,886 (78%) of the CDSs based on the match database. Of the total 26,404 assigned GO terms, 11,841 (45%), 7,410 (28%), and 7,153 (27%) were assigned for the biological process, molecular function, and cellular component category, respectively (fig. 1*a*–*c*). Among these categories, most of the genes mapped for oxidation-reduction process, ATP binding, and integral component of the membrane, respectively. Statistical analysis of the blasted sequences showed that nearly 80% of the sequences had an *E*-value of $<e^{-100}$ (fig. 1*d*), most of the sequences had an alignment identity of >90% (fig. 1*e*), and nearly 70% sequences matched to different strains of *C. albicans* (fig. 1*f*).

Comparative Genomic Analysis

To assess the preliminary genomic comparison of the TIMM1768 strain with *C. albicans* SC5314, we mapped the Ion PGM reads from TIMM1768 to the SC5314 assembly to find the mapping of >99% of the reads. This suggested a conserved genomic architecture between these two strains. Further, based on the phylogenetic tree created from the aligned concatenated genes used for multi-locus sequence typing, we found that TIMM1768 claded with *C. albicans* CHN1 (fig. 2*a*). These findings suggested that although SC5314 and TIMM1768 shared a conserved genome architecture, these two strains underwent a high level of divergence during evolution. This was further evident from the differences in the preferred codons among these species (supplementary table 3, Supplementary Material online). As the whole genome sequence of multiple strains in the TIMM1768 clade is not publicly available, this is one of the few strains to be sequenced in this group. Furthermore, we performed a comparative analysis of the mitochondrial chromosome. The mitochondrial DNA encodes a handful of proteins involved in the mitochondrial electron transport chain and ATP synthesis (Gray 2012). The TIMM1768 mitochondrion harbored eight CDSs encoding for ATP synthesis (ATP synthase and NADH dehydrogenase), protein maturation (cytochrome C oxidase), and LAGLIDADG endonuclease. A comparative analysis using mauve (Darling et al. 2004) showed a high degree of similarity with the previously published mitochondrial DNA sequences of *C. albicans* (supplementary fig. 2*a* and *b*, Supplementary Material online).

To elucidate the rate of protein sequence evolution, we first identified the orthogroups between the TIMM1768, CD36, and S288C strains and then determined the extent of synonymous (Ks) and nonsynonymous (Ka) sequence divergence among the orthologous gene pairs of TIMM1768 and CD36 (fig. 2*b*). Orthofinder analysis revealed 3,174 single copy orthogroups present in these three strains and 1170 orthogroups were present only in TIMM1768 and CD36. We were interested to examine the pattern of Ka/Ks ratio between the former and the latter gene sets. Although we did not find the genes with strong positive selection (Ka/Ks >1.0), larger proportion of genes present in the latter set had a higher Ka/Ks ratio (fig. 2*c*). Given that a higher value of Ka/Ks ratio is considered to indicate the genes likely to have experienced or be experiencing positive selection, we selected 43 genes with Ka/Ks > 0.3 and checked the assigned GO terms to find that majority of the genes were categorized under pathogenesis, integral component of membrane and DNA binding (supplementary fig. 3*a*–*c*, Supplementary Material online).

Genes Involved in Pathogenesis

Several factors that contribute to the pathogenicity of *C. albicans* have been identified (Mayer et al. 2013). Most of these factors either enable adhesion and invasion of the host cell or modify the pathogenic form to avoid host defense mechanisms. Further, it has been shown that mutants that are unable to form hyphae have attenuated virulence

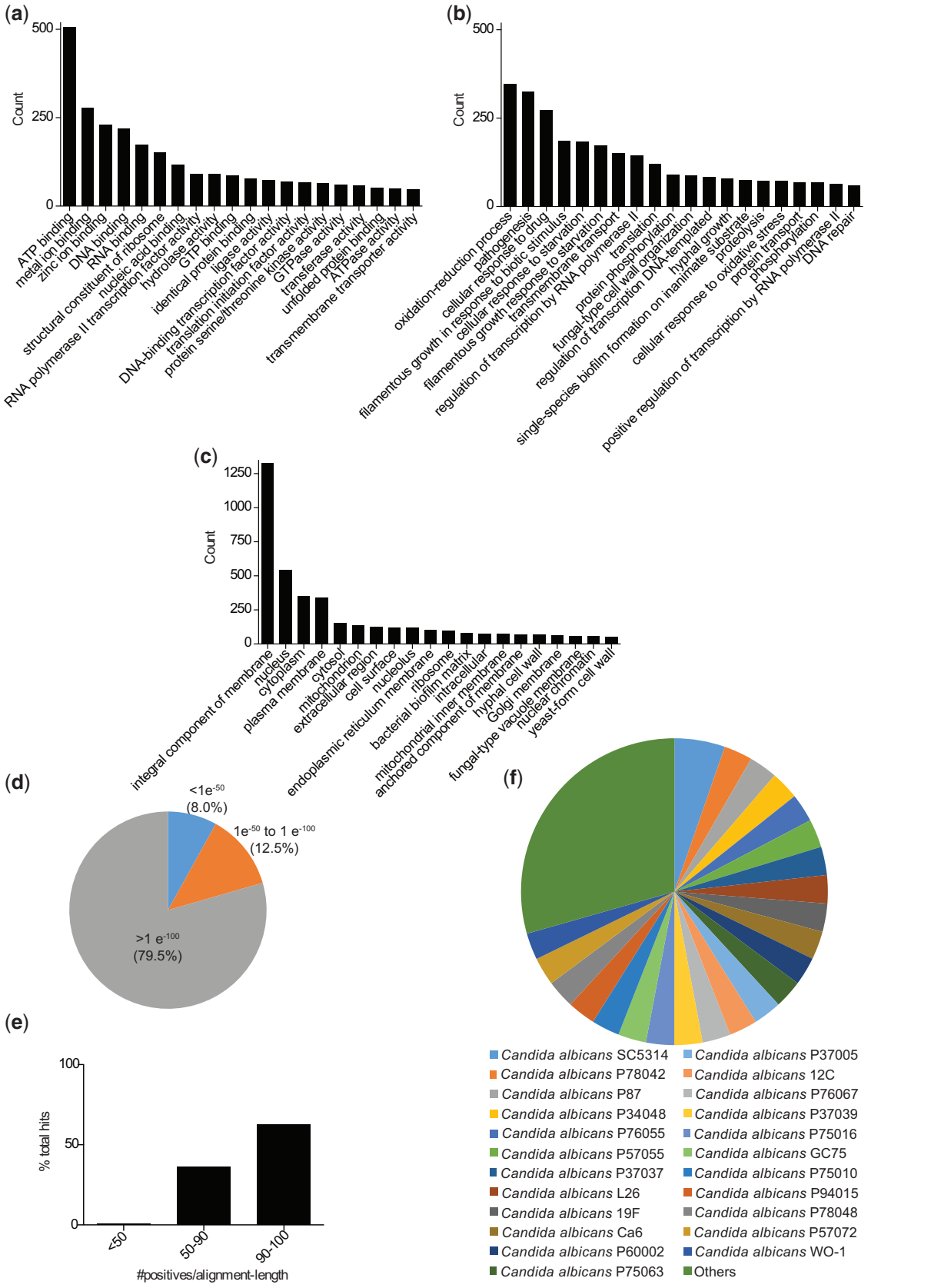

Fig. 1.—Genome annotation and analysis of *C. albicans* TIMM1768 genome. Top Gene Ontology (GO) term counts for—(*a*) molecular functions, (*b*) biological processes, and (c) cellular component. Summary of BLAST results of the proteins identified—(*d*) *E*-value distribution, (*e*) Similarity distribution, and (*f*) Origin of species distribution.

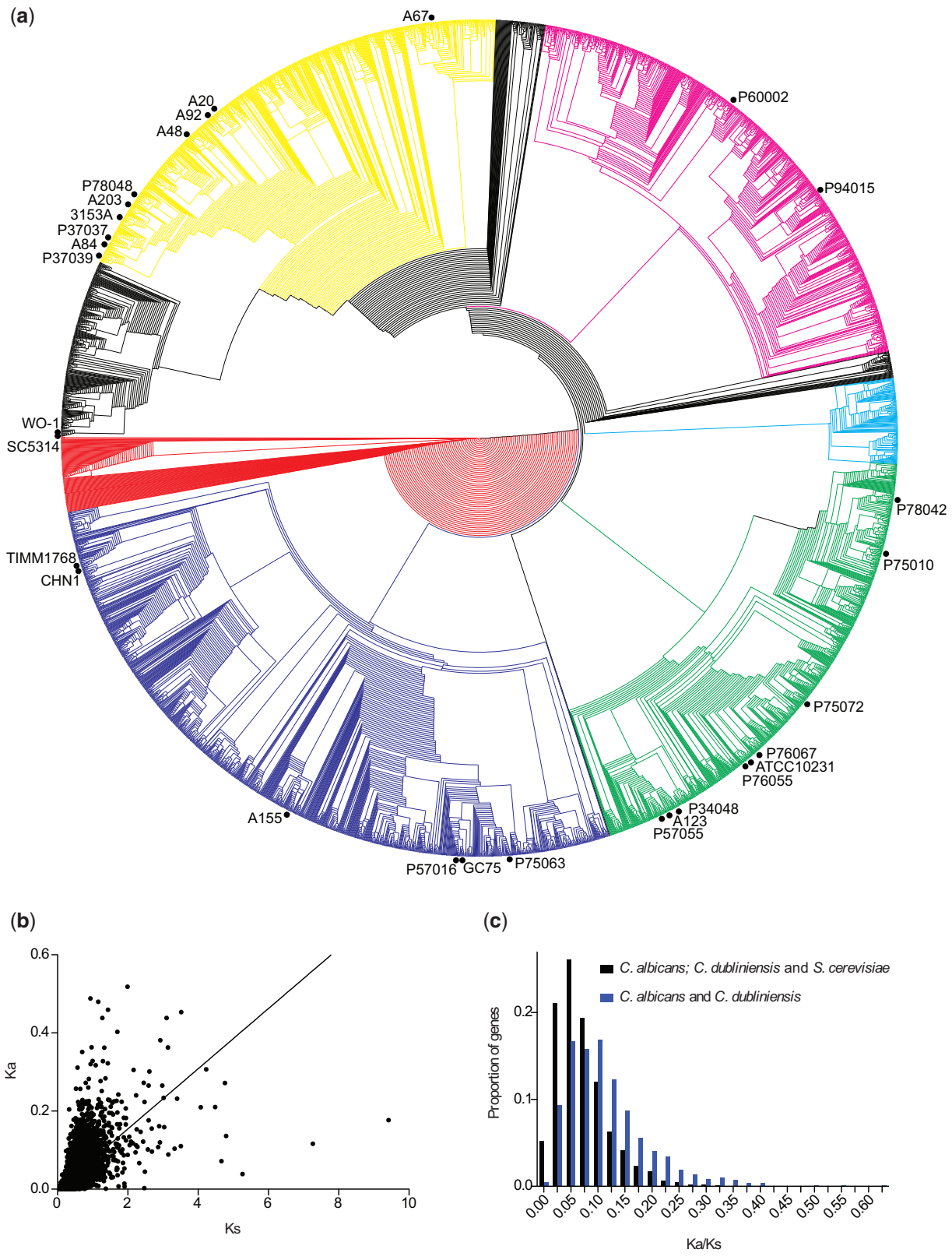

Fig. 2.—Comparative analysis of TIMM1768. (*a*) Phylogenetic tree was constructed based on the concatenated sequence of the genes used for multilocus sequence typing. The positions of TIMM1768 and other sequences downloaded from NCBI are indicated. (*b*) Scatter plot showing the nonsynonymous and synonymous substitution levels in 4,344 ortholog pairs between *C. albicans* TIMM1768 and *C. dubliniensis* CD36. (*c*) Histogram of distributions of Ka/Ks ratio among the single copy orthologs present in three and two species as determined by KaKs_calculator.

(Lo et al. 1997). We found 324 genes possibly involved in pathogenesis in the TIMM1768 genome. We attempted to analyze the function of these genes and found that the majority of the genes were categorized as cellular response to starvation, cellular response to drug, and genes involved in the filamentous growth. Further, we found the presence of various virulence factors, such as Als (agglutinin-like sequence), Hwp (hyphal wall protein), Iff (IPF family F), Sap (secreted aspartic protease), Phr1 (β-1,3 glucanosyl transferase), GPI-anchored proteins (Ecm, Pga, Rhd, and Utr), Hsp (heat-shock protein), Tec1 (TEA/ATTS transcription factor), other transcriptional factors (NDT8 and Brg1), members of the RIM101 pathway, Plb (phospholipase); Lip (lipase), catalase, superoxide dismutase, and kinases (CEK1 and Hog1).

In summary, we performed a de novo assembly of the TIMM1768 genome using the Oxford nanopore MinION long reads and Ion PGM short reads. Comparative analysis with the published Candida genomes indicated that this strain claded with the strains for whom most of the genome is not sequenced. Further, comparison with *C. dubliniensis* CD36 indicated that the rate of synonymous sequence divergence was much higher compared with rate of nonsynonymous divergence. The availability of the TIMM1768 genome sequence will further facilitate the Candida research, especially investigations of the genetic requirements for commensalism, clinical candida virulence, site-specific colonization, and the development of antivirulence drugs.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported in part by a Grant-in-Aid for scientific research (S) (JP15H05783) to K.S.; Grant-in-Aid for scientific research (C) (JP16K08790) to S.A.I; and Grant-in-Aid for JSPS Fellows (JP17F17421) to S.P. and K.S.

Literature Cited

Anderson MZ, Bennett RJ. 2016. Budding off: bringing functional genomics to *Candida albicans*. Brief Funct Genomics 15(2):85–94.

Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19(5):455–477.

Bosi E, et al. 2015. MeDuSa: a multi-draft based scaffolder. Bioinformatics 31(15):2443–2451.

Bougnoux M-E, et al. 2003. Collaborative consensus for optimized multilocus sequence typing of *Candida albicans*. J Clin Microbiol. 41(11):5265–5266.

Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14(7):1394–1403.

Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5(6):e11147.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16(1):157.

Gray MW. 2012. Mitochondrial evolution. Cold Spring Harbor Perspect Biol. 4(9):a011403.

Hisajima T, et al. 2008. Invasion process of *Candida albicans* to tongue surface in early stages of experimental murine oral candidiasis. Med Mycol. 46(7):697–704.

Ishijima SA, Abe S. 2015. A novel murine candidiasis model with severe colonization in the stomach induced by N-acetylglucosamine-treatment and its scoring system based on local characteristic stomach symptoms. Med Mycol J. 56(4):E31–E39.

Ishijima SA, et al. 2012. N-acetylglucosamine increases symptoms and fungal burden in a murine model of oral candidiasis. Med Mycol. 50(3):252–258.

Kolmogorov M, Yuan J, Lin Y, Pevzner P. 2018. Assembly of long error-prone reads using repeat graphs. *bioRxiv*: 247148.

Koren S, et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27(5):722–736.

Lo H-J, et al. 1997. Nonfilamentous *C. albicans* mutants are avirulent. Cell 90(5):939–949.

Mayer FL, Wilson D, Hube B. 2013. *Candida albicans* pathogenicity mechanisms. Virulence 4(2):119–128.

Mizutani S, et al. 2000. CD4+-T-cell-mediated resistance to systemic murine candidiasis induced by a membrane fraction of *Candida albicans*. Antimicrob Agents Chemother. 44(10):2653–2658.

Odds FC. 2010. Molecular phylogenetics and epidemiology of *Candida albicans*. Future Microbiol. 5(1):67–79.

Panthee S, Hamamoto H, Paudel A, Sekimizu K. 2017. Genomic analysis of vancomycin resistant *Staphylococcus aureus* VRS3b and its comparison with other VRSA isolates. Drug Discov Ther. 11(2):78–83.

Panthee S, Paudel A, Hamamoto H, Sekimizu K. 2017. Draft genome sequence of the vancomycin-resistant clinical isolate *Staphylococcus aureus* VRS3b. Genome Announc. 5(22): e00452-17.

Proux-Wéra E, Armisén D, Byrne KP, Wolfe KH. 2012. A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach. BMC Bioinformatics 13(1):237.

Salmela L, Rivals E. 2014. LoRDEC: accurate and efficient long read error correction. Bioinformatics 30(24):3506–3514.

Zhang Z, et al. 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics 4(4):259–263.

Zhang Z, et al. 2012. ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. Biochem Biophys Res Commun. 419(4):779–781.

Associate editor: Howard Ochman