




Genome and transcriptome of a pathogenic yeast, *Candida nivariensis*

Yunfan Fan ¹, Andrew N. Gale ², Anna Bailey,¹ Kali Barnes,¹ Kiersten Colotti,¹ Michal Mass,¹ Luke B. Morina,¹ Bailey Robertson,¹ Remy Schwab,¹ Niki Tselepidakis,¹ and Winston Timp ^{1,3,4,*}

¹Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA

²Department of Biology, Johns Hopkins University, Baltimore, MD 21218, USA

³Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

⁴Division of Infectious Disease, Department of Medicine, Johns Hopkins University, Baltimore, MD 21205, USA

*Corresponding author: Email: wtimp@jhu.edu

Abstract

We present a highly contiguous genome and transcriptome of the pathogenic yeast, *Candida nivariensis*. We sequenced both the DNA and RNA of this species using both the Oxford Nanopore Technologies and Illumina platforms. We assembled the genome into an 11.8 Mb draft composed of 16 contigs with an N50 of 886 Kb, including a circular mitochondrial sequence of 28 Kb. Using direct RNA nanopore sequencing and Illumina cDNA sequencing, we constructed an annotation of our new assembly, supplemented by lifting over genes from *Saccharomyces cerevisiae* and *Candida glabrata*.

Keywords: *Candida nivariensis*; genome assembly; pathogenic yeast; nanopore sequencing; transcriptome

Introduction

For immunocompromised hosts, opportunistic infections caused by drug-resistant fungi of the *Candida* genus are a major source of morbidity and mortality (Borman et al. 2008). In particular, *Candida nivariensis*, a close relative to *Candida glabrata*, has emerged in recent years as especially resistant to antifungal therapies (Borman et al. 2008). However, due to its phenotypic similarities to *C. glabrata*, *C. nivariensis* is generally underidentified and easily misdiagnosed, and currently, only molecular approaches can distinguish the two (Aznar-Marin et al. 2016), spurring whole-genome sequencing studies on the clade (Gabaldón et al. 2013).

Accurate assembly of repetitive genomic regions is crucial for understanding genetic diversity and virulence in pathogenic species. Fungal pathogens have long been known to exhibit a high degree of genome plasticity to enhance fitness in various environments (Croll et al. 2013; Ford et al. 2015; López-Fuentes et al. 2018; Carreté et al. 2019; Todd et al. 2019). Repetitive subtelomeric regions in particular play a crucial role in virulence for many pathogenic organisms (Barry et al. 2003; De Las Peñas et al. 2003). Many yeasts' subtelomeric regions contain and regulate the expression of genes crucial for biofilm formation, carbohydrate utilization, and cellular adhesion (Naumov et al. 1995; De Las Peñas et al. 2003; Iraqui et al. 2005). These gene families often undergo rapid evolution through changes in copy number and sequence through either SNPs or indels (Carreto et al. 2008; Brown et al. 2010; Anderson et al. 2015). However, these subtelomeric regions remain one of the most difficult sections of the genome to accurately assemble due to their repetitive nature and high sequence

similarity between genes, making genetic analysis cumbersome (Brown et al. 2010).

One of the gene families of great interest to the pathogenic yeast field is the glycosylphosphatidylinositol-anchored cell wall proteins (GPI-CWPs). This protein family includes many genes that encode for adhesion proteins that are found in various members of the *Candida* genus, and play a key role in pathogenicity, being involved in regulation of biofilm formation, cell-to-cell contact, and host-pathogen interactions (Timmermans et al. 2018; McCall et al. 2019). With the many roles, these genes play in infection, the accurate identification and understanding of the genetic variation of these genes vital to combating fungal pathogens.

Unfortunately, like many eukaryotic pathogens, the current reference genome for *C. nivariensis* (GenBank: GCA_001046915.1) is highly fragmented. Constructed from sequencing of strain CBS9983, the reference genome consists of 123 contigs with an N50 of 248 Kb (Gabaldón et al. 2013), meaning that at least half of the total genome length is contained in contigs 248 Kb or longer. This is typical of genomes assembled from limited short-read sequencing data; though short reads are highly accurate, assembling them into contiguous genomes is challenging depending on the size and complexity of the genome. Such short-read assemblies have limited utility since large-scale variants, repetitive regions, and genome structure remain difficult to elucidate, though they are often involved in the genome plasticity of pathogenic yeasts (Carreté et al. 2018). In contrast, long-read sequencing data have been shown to produce much more contiguous assemblies and have been crucial in sequencing through large repetitive regions, as well as assessing structural variants.

Received: February 5, 2021. Accepted: April 11, 2021

© The Author(s) 2021. Published by Oxford University Press on behalf of Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

However, read accuracy on the Oxford Nanopore Technologies (ONT) platform in particular ranges from 86% for early basecaller versions (Wick et al. 2019) to 97% as currently reported by ONT. This is lower than the read accuracy of short-read Illumina sequencing, which achieves 99.9% accuracy (Fox et al. 2014). In consensus sequences, most random errors can be corrected by other reads covering the same genomic loci, resulting in >99% consensus accuracy (Wick et al. 2019). However, systematic errors occurring in most or all of the reads cannot be corrected this way. For ONT data, indels at homopolymers dominate systematic errors (Wick et al. 2019). These persistent errors can be problematic for gene prediction and annotation in downstream analysis (Watson and Warr 2019) and are typically corrected with more accurate short-read data in mappable regions (Garrison and Marth 2012; Walker et al. 2014; Vaser et al. 2017).

Having a genome alone is not enough; we need to annotate it with genes and other functional elements for the genome to be of greatest use. Knowledge of gene loci is critical to constructing phylogenetic relationships between organisms, and to studying the functional implications of variants, both common uses of reference genomes. While model-based, purely computational gene predictors can be highly accurate in bacteria, gene sparsity and intronic regions make this task more difficult in eukaryotes (Salzberg 2019). For improved annotations, some RNA-seq information is required (Salzberg 2019).

Here, as part of our newly developed Methods in Nucleic Acid Sequencing university course, we used a hybrid approach, applying long-read nanopore sequencing to assemble a highly contiguous genome of *C. nivariensis*, followed by short-read sequencing to polish or correct errors in our assembly. We followed this by a combination of nanopore direct RNA sequencing as well as short-read RNA-seq to annotate our assembly. By combining these data with liftover of annotations from evolutionary “cousins” of *nivariensis*, we have generated a new and annotated reference genome for the community.

Materials and methods

Media and growth conditions

For genomic extractions, a single colony of *C. nivariensis* CBS9983, originally isolated from a blood culture of a Spanish woman (Alcoba-Flórez et al. 2005), was inoculated into synthetic complete (SC) medium supplemented with 2% glucose and shaken overnight at 30°C in a glass culture tube. For RNA extractions, *C. nivariensis* CBS9983 was grown to log phase in SC medium supplemented with 2% glucose at 30°C in a glass culture tube.

DNA isolation and sequencing

DNA was extracted from liquid culture using the Zymo Fungal/Bacterial DNA MiniPrep Kit according to manufacturer specifications. Two ONT sequencing libraries were prepared from the extracted DNA using the ONT rapid barcoding sequencing kit (SQK-RBK004), and each was sequenced on a separate MinION flowcell (R9.4). Two Illumina libraries were prepared with the Nextera Flex Library Prep Kit, each using 400 ng of extracted DNA. Both Illumina libraries were then sequenced on a single iSeq 100 run.

RNA sequencing

RNA was extracted from liquid culture using the Zymo Fungal/Bacterial RNA MiniPrep Kit. Using the NEBNext Poly(A) mRNA Magnetic Isolation Module, polyA tailed mRNA was isolated from the total RNA. Two ONT direct RNA sequencing libraries were

prepared and sequenced on separate MinION flowcells, each using ~200 ng of polyA selected RNA and the SQK-RNA002 sequencing kit. With the NEBNext Ultra II RNA First-Strand Synthesis Module and the NEBNext Ultra II Non-Directional RNA Second Strand Synthesis Module, cDNA was prepared from the isolated mRNA. Two individual Illumina libraries were then prepared with the Nextera Flex Library Prep Kit, each using 400 ng of cDNA. Both library replicates were then sequenced on a single iSeq 100 run, generating 2×150 paired-end reads.

Genome assembly

Nanopore data were basecalled using Guppy v3.2.4 on default settings. Reads greater than 3 kb long with an average basecalling quality score greater than 7 were assembled into 21 contigs using Canu v2.1 (Koren et al. 2017) on default settings with the genome size set to 11 m. Illumina DNA reads were trimmed for adapters and quality using Trimmomatic v0.39 (Bolger et al. 2014) using settings LEADING:3 TRAILING:3 SLIDINGWINDOW:4:30 MINLEN:36. The trimmed reads were then used to iteratively correct draft assembly using FreeBayes v1.3.4-pre1 (Garrison and Marth 2012) with alignments made by bwa mem v0.7.17-r1198-dirty (Li 2013) using default settings. Changes were made at positions where both the alternative allele frequency was greater than 0.5 and the total number of alternate allele observations was greater than 5. We aligned and corrected the assembly iteratively for three rounds, after which further rounds of corrections made no changes.

Of our 21 corrected contigs, 5 were flagged as repeats by Canu and originally constructed from fewer than 180 nanopore reads. The remaining 16 contigs were constructed from over 1800 nanopore reads each. Because the five repetitive contigs were constructed from so few reads and were found to occur elsewhere in the assembly through Mummer v4.0.0rc1 (Marçais et al. 2018) and nanopore read alignment Minimap2 v2.17 (Li 2018), we excluded them from the final assembly. One 32-Kb contig was suggested to be circular by Canu, and therefore likely to be a mitochondrial sequence. To confirm, we aligned this contig to the complete mitochondrial genome of *C. nivariensis* (NCBI: NC_036379.1) using Mummer, and observed a 3662-bp sequence in the reference mitochondrial genome which appears at both ends of our 32-kb circular contig. Using the Mummer alignments (Supplementary Figure S1), we removed the extraneous 3662 bp from the end of our contig, resulting in a 28-kb mitochondrial genome, which we named “JHU_Cniv_v1_mito.” Lastly, we remapped the ONT and Illumina reads back to the assembly, and found no bases with zero coverage, indicating that none of our contigs need to be further broken (Supplementary Figure S2). Henceforth, we refer to this assembly as “JHU_Cniv_v1.”

Repeat regions were identified by Tandem Repeats Finder v4.09 (Benson 1999) with settings [match = 2, mismatch = 7, delta = 7, pm = 80, pi = 10, minscore = 50, maxperiod = 600] (Xu et al. 2020). Multimapping short reads were identified using bwa mem (Li 2013) on default settings.

Annotation

Illumina RNA-seq reads were trimmed using Trimmomatic v0.39 (Bolger et al. 2014) in order to check for any remaining adapter sequences and to filter out reads with low base quality. HISAT2 v2.1.0 was used on default settings to align the trimmed cDNA reads to the assembly. The BRAKER v2.1.5 (Hoff et al. 2019) pipeline was then used to make gene predictions using these alignments. Currently, ONT dRNA compatibility with BRAKER is in development, and that data was thus not used for prediction.

Instead, ONT dRNA reads were aligned to the genome assembly using Minimap2 on recommended settings for nanopore direct RNA reads (-ax splice -uf -k14). Transcripts were then assembled from the dRNA alignments using StringTie2 v2.1.5 (Kovaka et al. 2019) with the long read option (-L). Using Liftoff v1.5.0 (Shumate and Salzberg 2020), we lifted over the annotations from *C. glabrata* (NCBI: GCF_000002545.3), *Saccharomyces cerevisiae* (NCBI: GCF_000146045.2), *Candida albicans* (NCBI: GCF_000182965.3).

Starting with the BRAKER predictions, Gffcompare v0.12.1 (Perteau and Perteau 2020) was used to add nonoverlapping annotations lifted from *C. glabrata*, *S. cerevisiae*, and *C. albicans* in that order. Specifically, we add any annotation with class code “u” in the Gffcompare .tmap outputs when comparing our list of genes with a list of potential genes to add, since these refer to intergenic regions devoid of any overlap or proximity to previous annotations. Finally, we compared and added nonredundant transcripts assembled by stringtie2 to the annotation using gffcompare.

Data availability

All sequence data are available in the Sequence Read Archive, under BioProject PRJNA686979. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JAEVGP000000000. The version described in this paper is version JAEVGP010000000. The JHU_Cniv_v1 assembly and annotation are also available in Zenodo (<http://doi.org/10.5281/zenodo.4644506>). Code used for analysis is available at <https://github.com/timlplab/nivar>. Supplementary materials and data files are available on figshare: <https://doi.org/10.25387/g3.14381858>.

Results

Using our nanopore and Illumina sequencing data, we generated a new assembly of *C. nivariensis*, JHU_Cniv_v1 (Methods). Our assembly consists of 11.8 Mb of sequence in 16 contigs with an N50 of 886 Kb (Figure 1A, Table 1). Compared to the reference genome, we have 275 kb of additional sequence, 218 kb of which is accounted for by gaps in the reference, which are newly spanned by JHU_Cniv_v1. Of the 69 newly spanned gap sequences, 54 were identified as repeat regions. Another 13 gap regions were identified to contain a higher than average proportion of multimapping short reads (>10% in gap regions vs 7% average across the genome).

To determine whether JHU_Cniv_v1 contigs represent full chromosomes, we looked for telomere repeats in our assembly and attempted to use related yeast reference genomes to scaffold. In our assembly, 11 contigs terminate at both ends in repeats of CTGGGTGCTGTTGGGT, the telomere sequence of *C. glabrata* (McEachern and Blackburn 1994). The other four nonmitochondrial sequences terminate only at one end in this telomeric repeat (Figure 1B, Supplementary Table S1), suggesting they may scaffold to form two additional chromosomes. This suggests that, like *C. glabrata*, the *C. nivariensis* genome also contains 13 chromosomes.

We tried to further scaffold our assembly using the more contiguous and highly related *glabrata* genome as a reference, but we found that reference-based scaffolders such as Medusa v1.6 (Bosi et al. 2015) and RagTag v1.0.2 (Alonge et al. 2019) either placed telomeric sequences in the middle of scaffolds or made no improvement (Supplementary Figure S3). Upon aligning the *C. glabrata* genome to JHU_Cniv_v1 using Mummer, we found only sporadic shared segments of negligible length (Supplementary Figure S4), as opposed to a nearly perfect 1:1 alignment between

JHU_Cniv_v1 and the current *C. nivariensis* reference genome (Supplementary Figure S5). This indicated that the *C. glabrata* genome is not sufficiently similar to *C. nivariensis* to use as a reference for contig scaffolding. Using the *C. nivariensis* reference genome for scaffolding similarly results in erroneous placement of telomere repeats in the middle of scaffolds, or no change to our assembly. This is unsurprising, as the *C. nivariensis* reference genome is so highly fragmented.

To assess assembly completeness, fungal single-copy orthologs were checked using BUSCO v5.0.0 (Simão et al. 2015) and its available saccharomycetes_odb10 database. Out of 2137 BUSCOs searched, JHU_Cniv_v1 has only 14 missing, 13 of which are also missing in the current reference (Figure 2). This missing gene, RNA polymerase archaeal subunit P/eukaryotic subunit RPABC4 (buscoID 41996at4891), though present in the reference, has the second lowest combined match length and match score among all genes searched. From the reference, we extracted the nucleotide sequence of this match using the coordinates reported by BUSCO, and searched for it in JHU_Cniv_v1 using BLAST. We found a full-length match with 99.9% identity, suggesting that this BUSCO is not actually absent in JHU_Cniv_v1. Upon further examination of this alignment, we found that all seven non-matching nucleotides consist of small deletions associated with poly-A or poly-T homopolymers, known error-prone regions for nanopore sequencing data (Watson and Warr 2019).

We annotated our new assembly by lifting over genes from related yeasts and adding gene predictions based on long- and short-read RNA sequencing from the same strain (Methods). Our final annotation of JHU_Cniv_v1 comprises 25,979 features, 5859 of which are genes (Supplementary Table S2), the rest of which are more detailed features including individual exons, coding sequence, and start/stop codons. Current annotations of closely related yeasts report similar gene counts (Supplementary Table S3). In order to assess transcriptome completeness, BUSCO was used in transcriptome mode, again with its saccharomycetes_odb10 database. Because no annotation of the *C. nivariensis* reference genome currently exists, we compared our transcriptome to those of *C. glabrata*, *S. cerevisiae*, and *C. albicans*. Compared to these highly characterized yeast transcriptomes, ours contains slightly fewer complete and single-copy BUSCOs (1876 of 2137 searched) and roughly double the number of complete and duplicated BUSCOs (232 of 2137 searched). The numbers of missing and fragmented BUSCOs between the three are comparable (Figure 2).

Repetitive genes

As *C. glabrata* subtelomeric regions have been proven to be difficult to correctly assemble using short-read data (Xu et al. 2020), we compare the copy number of *C. glabrata* subtelomere gene homologs between the *C. nivariensis* reference genome and JHU_Cniv_v1. Using the assembly and re-annotation of *C. glabrata* from Xu et al. (2020), we extracted the sequences of the *C. glabrata* subtelomere genes and used BLAST (v2.6.0+) to find any matches in the *C. nivariensis* reference and JHU_Cniv_v1. We observed an identical set of 48 *C. glabrata* subtelomere genes in both *C. nivariensis* genomes but found that the copy number for several genes was greater in JHU_Cniv_v1 (Figure 3A). To account for genes truncated by short contigs in the reference genome, we calculate copy number by summing the alignment lengths of all the hits of a particular gene and dividing by gene length. Of the 48 *C. glabrata* genes with homology in *C. nivariensis*, 35 are ribosomal. With the exception of just three ribosomal genes, which occur a similar number of times in both *C. nivariensis* genomes, all homologous

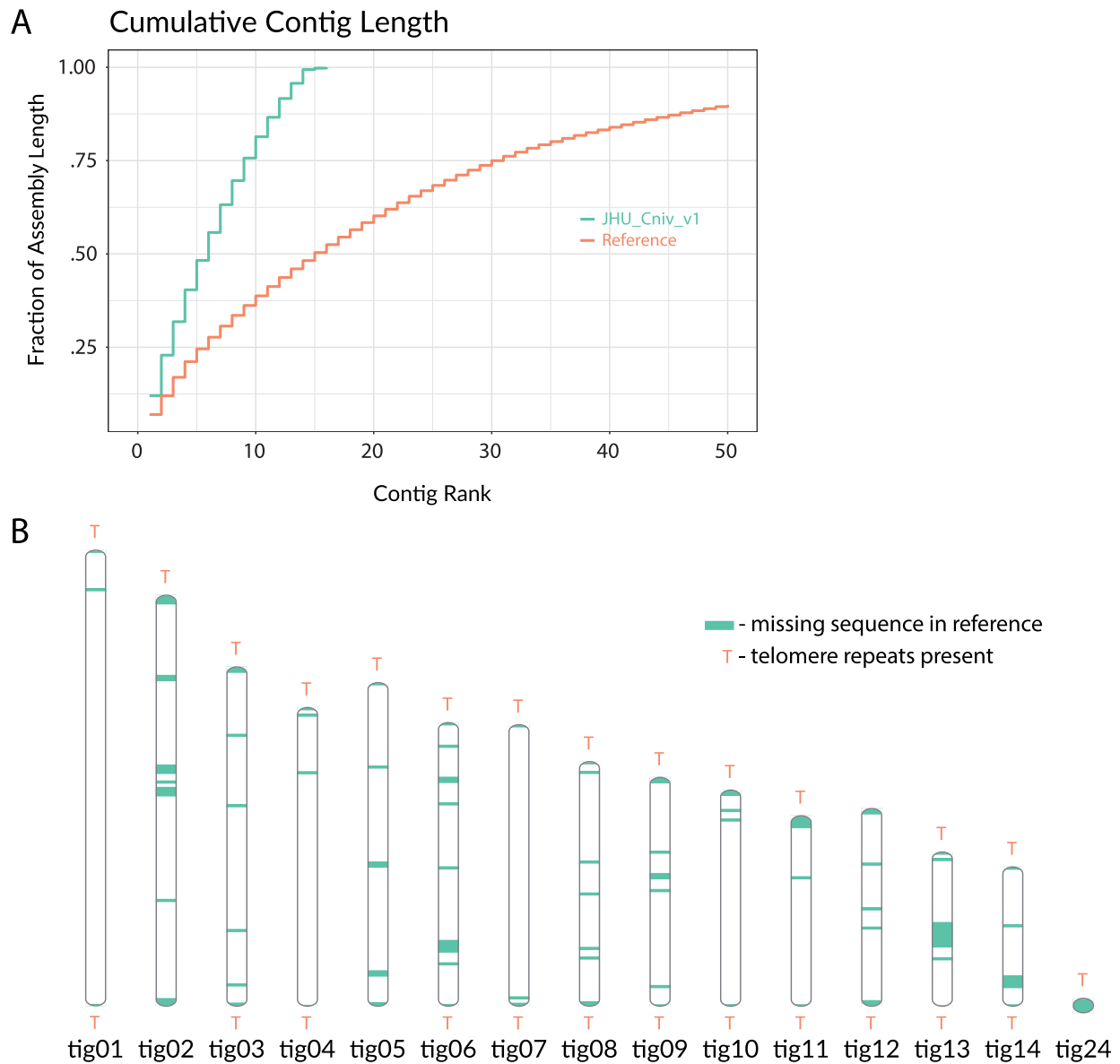


Figure 1 Characteristics of the JHU_Cniv_v1 (A) Cumulative lengths of the 50 longest sequences in our assembly and previous reference genome. (B) Ideogram of assembly. Sequence that is missing in the reference genome is shown along each nonmitochondrial contig, and the positions of telomere repeats are marked.

Table 1 Assembly statistics

	Contigs	N50	Longest contig	Shortest contig	Total length
Reference (GCA_001046915.1)	123	248 Kb	807 Kb	666 bp	11.56 Mb
JHU_Cniv_v1	16	886 Kb	1.42 Mb	28.5 Kb	11.83 Mb

ribosomal genes appear once in the reference, and either four or six times in JHU_Cniv_v1 (Figure 3A, Supplementary Data S1).

Using JHU_Cniv_v1, we identified GPI-anchored membrane proteins among annotated genes >1000-nt long. Using GffRead (Pertea and Pertea 2020), we constructed the amino acid sequences for these genes and excluded any with internal stop codons. We then used PredGPI (Pierleoni et al. 2008) to predict which of these encoded GPI proteins, using an FDR cutoff of <0.0005 (Xu et al. 2020) to find 86 total genes. As GPI-anchored fungal adhesins typically contain tandem repeats (Lipke 2018; Xu et al. 2020),

we further filtered for genes overlapping with tandem repeats as classified by Tandem Repeat Finder and identified 53 of the GPI genes as putative adhesins. As with *C. glabrata*, the putative adhesins typically spanned multiple kilobases (Figure 3B), though we do not find very long (>13 kb) genes in contrast to several *glabrata* GPI-CWPs. To find the corresponding adhesin genes in the *C. nivariensis* reference genome, we again used BLAST, and compared the longest hit of each adhesin gene to the true length of the gene as predicted in JHU_Cniv_v1 (Figure 3C). Notably, no hit in the reference genome exceeded 3.5 kb, and 27 of these adhesin

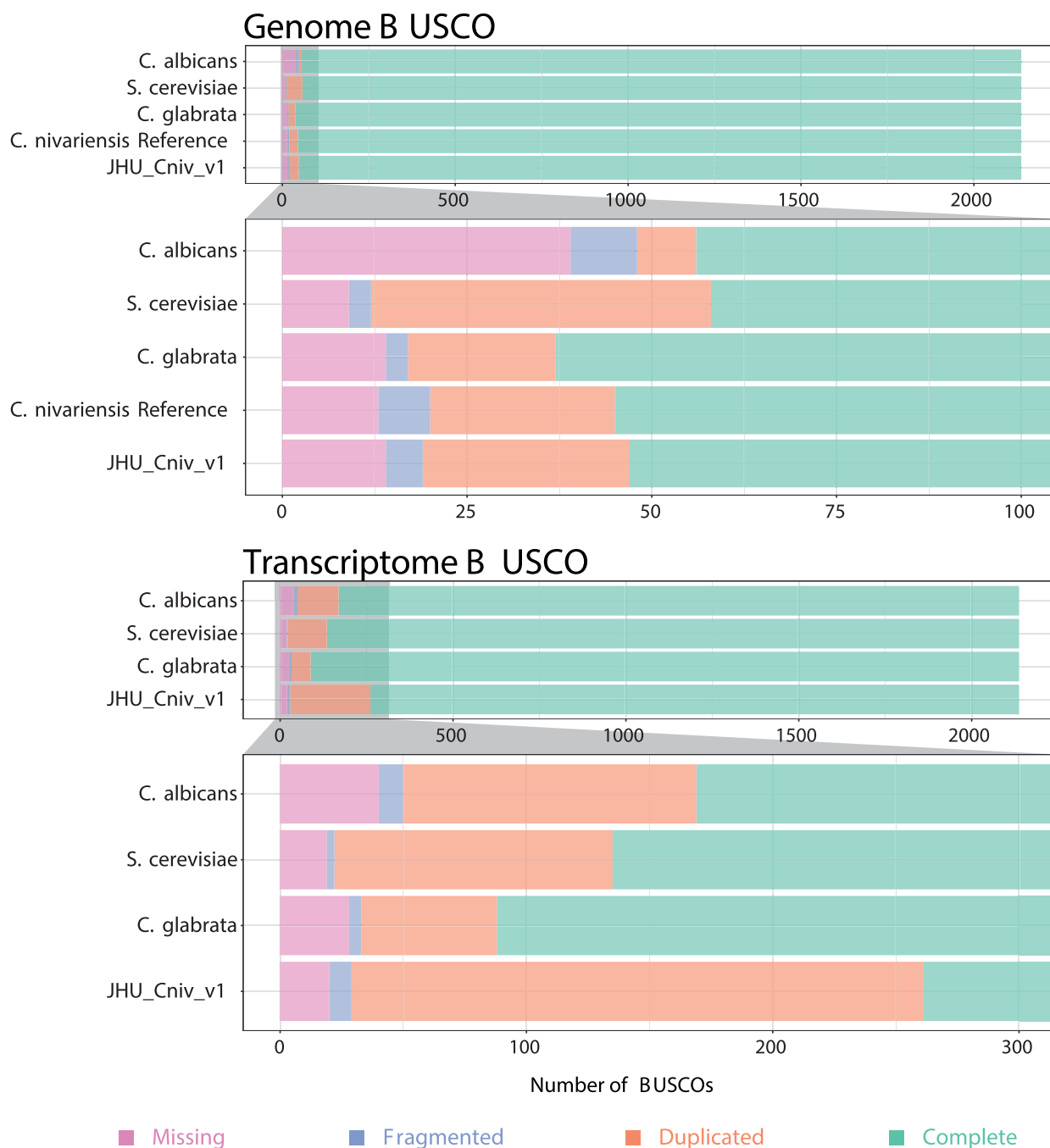


Figure 2 Genome and transcriptome completeness Bar charts comparing BUSCOs detected in JHU_Cniv_v1 and accompanying transcriptome to those of the current *C. albicans*, *S. cerevisiae*, *C. glabrata*, and *C. nivariensis* reference genomes. No reference transcriptome is currently available for *C. nivariensis*.

genes are not found continuously, suggesting the previous reference either truncated or did not continuously assemble these important pathogenicity genes.

Discussion

JHU_Cniv_v1 is a high quality, extremely contiguous assembly of *Candida nivariensis* constructed by long reads and polished by short reads. It spans large, repetitive gaps in the *nivariensis* genome that have fragmented short-read assemblies thus far, and includes a full mitochondrial chromosome, as well as telomere

repeats. These telomere repeats are identical to those in *C. glabrata* and have been found to be shared within the entire “glabrata group” (Gabaldón et al. 2013). The orientation of the telomeres suggests that *C. nivariensis* has 13 chromosomes, which is in agreement with previous PFGE data (Gabaldón et al. 2013). Furthermore, of the contigs missing telomere repeats on one end, we note that scaffolding tig05 with tig12 and tig02 with tig24 would result in 13 chromosomes that would all match PFGE length estimates to 8% error or less, which is within the expected range of PFGE error for very large DNA fragments (Cutting et al. 1988).

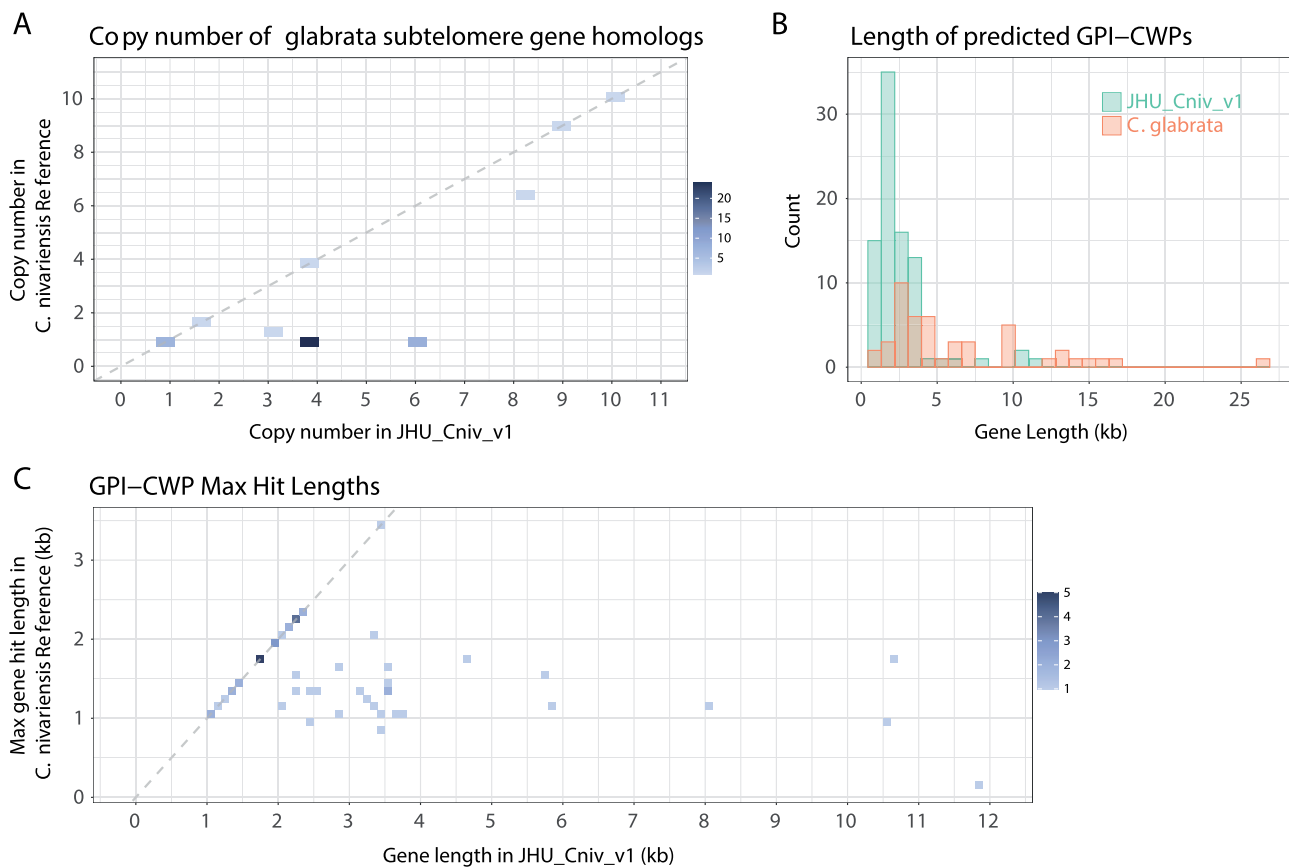


Figure 3 GPI genes (A) Scatterplot showing the number of times each *glabrata* subtelomere gene homolog appears in the *C. nivariensis* reference genome and in JHU_Cniv_v1. Overlapping points are shown on the color scale, and the $y = x$ line is shown in dashed gray. (B) Histogram of adhesion protein lengths in *glabrata* as annotated by Xu *et al.*, and the lengths of predicted adhesion proteins found in JHU_Cniv_v1. (C) Scatterplot showing the maximum BLAST alignment lengths for each predicted *nivariensis* GPI gene in JHU_Cniv_v1 and the *C. nivariensis* reference genome. Overlapping points are shown on the color scale, and the $y = x$ line is shown in dashed gray.

As assessed by BUSCO, genome completeness of the current *C. nivariensis* reference and JHU_Cniv_v1 are comparable to other related yeasts, with our genome slightly improved over the previous reference. However, while JHU_Cniv_v1 is a much more contiguous assembly than any *C. nivariensis* genome preceding it, the few remaining sequence errors still can pose a problem to downstream analyses, as evidenced by the seemingly absent BUSCO we manually identified.

Our accompanying RNA-seq data enabled us to annotate this genome, achieving a similar level of BUSCO completeness to some of the most highly studied model organisms. Our annotation has comparable or lower levels of missing and fragmented BUSCOs compared to the reference annotations, though more duplicated ones. While our annotation is largely comparable to those of similar yeasts, it has not been manually curated, and should thus be treated as preliminary. Of course, as these organisms were grown under only one condition before RNA extraction, it remains unlikely that this annotation is fully complete.

To demonstrate the utility of genome and annotation contiguity, we examine genes from a difficult to assemble region in *C. glabrata*. For each subtelomeric *C. glabrata* gene with homology in *C. nivariensis*, more copies were found in JHU_Cniv_v1, as its contiguity allows it to more easily capture repeated genome elements. We note that of subtelomeric *glabrata* genes found, the majority are ribosomal, and of these, only three do not show a four or six times increased copy number in JHU_Cniv_v1. Due to the repetitive nature of rDNA arrays, it can be difficult for short-

read genome assemblies to capture them in their full complexity. Conversely, our long-read assembly more easily spans these regions, potentially providing a clearer look at the biology in which they are involved.

In addition to genes arranged in complex and repetitive patterns, our more contiguous assembly enables analysis of large genes with internal repeats, such as GPI adhesins. Since these genes are so large, it can be difficult or impossible to predict them from fragmented assemblies which are unable to capture them in their full length. As adhesins are critical to understanding elements of pathogenicity in these yeasts, fragmented genome assemblies and missing gene annotations can be crippling to this dimension of research in these organisms.

Acknowledgments

Much of the data described in this paper were collected as part of the EN.580.454 Methods in Nucleic Acid Sequencing course at Johns Hopkins University in Spring of 2019. The authors thank Illumina Inc. and ONT for supporting educational initiatives and supplying reagents for collecting these data.

Funding

This work was funded in part by the Johns Hopkins University Department of Biomedical Engineering, National Institute of Health (NIH) grant 1R01HG009190 (W.T.), and Canadian

Institutes of Health Research (CIHR) Doctoral Foreign Study Award DFS-157831 (Y.F.).

Conflicts of interest

W.T. holds two patents (US 8,748,091 and US 8,394,584) licensed to ONT. Y.F. and W.T. have received travel funding from ONT.

Literature cited

- Alcoba-Flórez J, Méndez-Alvarez S, Cano J, Guarro J, Pérez-Roth E, et al. 2005. Phenotypic and molecular characterization of *Candida nivariensis* sp. nov., a possible new opportunistic fungus. *J Clin Microbiol.* 43:4107–4111.
- Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, et al. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* 20:224.
- Anderson MZ, Wigen LJ, Burrack LS, Berman J. 2015. Real-time evolution of a subtelomeric gene family in *Candida albicans*. *Genetics.* 200:907–919.
- Aznar-Marin P, Galan-Sanchez F, Marin-Casanova P, García-Martos P, Rodríguez-Iglesias M. 2016. *Candida nivariensis* as a new emergent agent of vulvovaginal candidiasis: description of cases and review of published studies. *Mycopathologia.* 181:445–449.
- Barry JD, Ginger ML, Burton P, McCulloch R. 2003. Why are parasite contingency genes often associated with telomeres? *Int J Parasitol.* 33:29–45.
- Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27:573–580. 10.1093/nar/27.2.573 9862982
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 30:2114–2120. 10.1093/bioinformatics/btu170 24695404
- Borman AM, Petch R, Linton CJ, Palmer MD, Bridge PD, et al. 2008. *Candida nivariensis*, an emerging pathogenic fungus with multi-drug resistance to antifungal agents. *J Clin Microbiol.* 46:933–938.
- Bosi E, Donati B, Galardini M, Brunetti S, Sagot M-F, et al. 2015. MeDuSa: a multi-draft based scaffolder. *Bioinformatics.* 31:2443–2451.
- Brown CA, Murray AW, Verstrepen KJ. 2010. Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Curr Biol.* 20:895–903.
- Carreté L, Ksiezopolska E, Gómez-Molero E, Angoulvant A, Bader O, et al. 2019. Genome comparisons of *Candida glabrata* serial clinical isolates reveal patterns of genetic variation in infecting clonal populations. *Front Microbiol.* 10:112.
- Carreté L, Ksiezopolska E, Pegueroles C, Gómez-Molero E, Saus E, et al. 2018. Patterns of genomic variation in the opportunistic pathogen *Candida glabrata* suggest the existence of mating and a secondary association with humans. *Curr Biol.* 28:15–27.e7.
- Carreto L, Eiriz MF, Gomes AC, Pereira PM, Schuller D, et al. 2008. Comparative genomics of wild type yeast strains unveils important genome diversity. *BMC Genomics.* 9:524.
- Croll D, Zala M, McDonald BA. 2013. Breakage-fusion-bridge cycles and large insertions contribute to the rapid evolution of accessory chromosomes in a fungal pathogen. *PLoS Genet.* 9:e1003567.
- Cutting GR, Antonarakis SE, Youssoufian H, Kazazian HH, Jr. 1988. Accuracy and limitations of pulsed field gel electrophoresis in sizing partial deletions of the factor VIII gene. *Mol Biol Med.* 5:173–184.
- De Las Peñas A, Pan S-J, Castaño I, Alder J, Cregg R, et al. 2003. Virulence-related surface glycoproteins in the yeast pathogen *Candida glabrata* are encoded in subtelomeric clusters and subject to RAP1- and SIR-dependent transcriptional silencing. *Genes Dev.* 17:2245–2258.
- Ford CB, Funt JM, Abbey D, Issi L, Guiducci C, et al. 2015. The evolution of drug resistance in clinical isolates of *Candida albicans*. *Elife.* 4:e00662.
- Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA. 2014. Accuracy of next generation sequencing platforms. *Next Gener Seq Appl.* 1:25699289
- Gabaldón T, Martin T, Marcet-Houben M, Durrens P, Bolotin-Fukuhara M, et al. 2013. Comparative genomics of emerging pathogens in the *Candida glabrata* clade. *BMC Genomics.* 14:623.
- Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]*.
- Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-Genome Annotation with BRAKER. *Methods Mol Biol.* 1962:65–95. 10.1007/978-1-4939-9173-0_5 31020555
- Iraqui I, Garcia-Sanchez S, Aubert S, Dromer F, Ghigo J-M, et al. 2005. The Yak1p kinase controls expression of adhesins and biofilm formation in *Candida glabrata* in a Sir4p-dependent pathway. *Mol Microbiol.* 55:1259–1271.
- Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27:722–736. 10.1101/gr.215087.116 28298431
- Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, et al. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* 20:10.1186/s13059-019-1910-1
- Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio. GN]*.
- Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 34:3094–3100. 10.1093/bioinformatics/bty191 29750242
- Lipke PN. 2018. What we do not know about fungal cell adhesion molecules. *J Fungi (Basel).* 4:29772751
- López-Fuentes E, Gutiérrez-Escobedo G, Timmermans B, Van Dijck P, De Las Peñas A, et al. 2018. *Candida glabrata*'s genome plasticity confers a unique pattern of expressed cell wall proteins. *J Fungi (Basel).* 4:29874814
- Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, et al. 2018. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol.* 14:e1005944 10.1371/journal.pcbi.1005944PMC: 29373581
- McCall AD, Pathirana RU, Prabhakar A, Cullen PJ, Edgerton M. 2019. *Candida albicans* biofilm development is governed by cooperative attachment and adhesion maintenance proteins. *NPJ Biofilms Microbiomes.* 5:21.
- McEachern MJ, Blackburn EH. 1994. A conserved sequence motif within the exceptionally diverse telomeric sequences of budding yeasts. *Proc Natl Acad Sci U S A.* 91:3453–3457.
- Naumov GI, Naumova ES, Louis EJ. 1995. Genetic mapping of the alpha-galactosidase MEL gene family on right and left telomeres of *Saccharomyces cerevisiae*. *Yeast.* 11:481–483.
- Pertea G, Pertea M. 2020. GFF utilities: GffRead and GffCompare. *F1000Res.* 9:304.
- Pierleoni A, Martelli PL, Casadio R. 2008. PredGPI: a GPI-anchor predictor. *BMC Bioinformatics.* 9:392.
- Salzberg SL. 2019. Next-generation genome annotation: we still struggle to get it right. *Genome Biol.* 20:92.
- Shumate A, Salzberg SL. 2020. Liftoff: accurate mapping of gene annotations. *Bioinformatics.* 10.1093/bioinformatics/btaa1016
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* 31:3210–3212. 10.1093/bioinformatics/btv351 26059717

- Timmermans B, De Las Peñas A, Castaño I, Van Dijck P. 2018. Adhesins in *Candida glabrata*. *J Fungi (Basel)*. 4:29783771
- Todd RT, Wikoff TD, Forche A, Selmecki A. 2019. Genome plasticity in *Candida albicans* is driven by long repeat sequences. *Elife*. 8: 31172944
- Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res*. 27:737–746.
- Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 9:e112963.
- Watson M, Warr A. 2019. Errors in long-read assemblies can critically affect protein prediction. *Nat Biotechnol*. 37:124–126.
- Wick RR, Judd LM, Holt KE. 2019. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol*. 20:129.
- Xu Z, Green B, Benoit N, Schatz M, Wheelan S, et al. 2020. De novo genome assembly of *Candida glabrata* reveals cell wall protein complement and structure of dispersed tandem repeat arrays. *Mol Microbiol*. 113:1209–1224.

Communicating editor: B. J. Andrews