JHU THESIS TEMPLATE TITLE

by

Yunfan Fan

A dissertation submitted to Johns Hopkins University
in conformity with the requirements for the degree of
Doctor of Philosophy

Baltimore, Maryland
August, 2022

Abstract

While next generation sequencing (NGS) has enabled massively parallel DNA sequencing for lower and lower cost, the development of third generation nanopore sequencing offers several key advantages over older sequencing methods. Nanopore sequencers are pocket-sized, making them orders of magnitude cheaper than the next most affordable alternative and the ideal option for wide deployment. They are capable of providing data in real-time, saving valuable hours before data analysis can begin. Additionally, they are able to sequence reads several thousand basepairs long, as opposed to the hundreds of basepairs NGS platforms are capable of, and they embed base modification data without the need for specific treatment beforehand. Given these advantages, in this thesis I examine the application of nanopore sequencing to the study of human pathogens.

First, we use nanopore sequencing to characterize anti-microbial resistance (AMR) in forty clinical isolates. We analyzed real-time data to quickly identify AMR genes, assembled genomes to identify chromosomal mutations, and used short-read sequencing data to correct the errors in the assemblies. With sequencing data, we found that time to effective antibiotic therapy could be shortened by as much as 20 hours compared to standard antimicrobial

susceptibility testing (AST).

Second, we leverage the long reads of nanopore sequencing to assemble the genome of a pathogenic yeast, /textitCandida nivariensis. Previous efforts to assemble this yeast genome relied solely on NGS data, resulting in a highly fragmented genome. Using nanopore data, we achieve a much higher contiguity, capture previously missing portions of the genome. Furthermore, we demonstrate that our more contiguous genome can be used to better study long and repetative genes, such as those involved in pathogenticity to humans.

Third, we use the base modification information embedded in nanopore sequencing data to call methylation in metagenomic assemblies. These calls enable the binning of metagenomic contigs according to methylation signature without the need to collect additional data. We demonstrate the efficacy of this method on a synthetic community sample, a simple two-bacteria system, and a clinical sample with matched proximity ligation binning data.

These applications of nanopore sequencing demonstrate its potential and its utility for all fronts of pathogen genomics research.

Thesis Committee

Primary Readers

Dr. Winston Timp (Primary Advisor)
Associate Professor
Department of Biomedical Engineering
Department of Molecular Biology and Genetics
Johns Hopkins University School of Medicine

Dr. Patricia Simner
Associate Professor
Department of Pathology
Johns Hopkins University School of Medicine

Dr. Steven Salzberg
Bloomberg Distinguished Professor
Department of Computer Science
Johns Hopkins University Whiting School of Engineering
Department of Biomedical Engineering
Johns Hopkins University School of Medicine
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Alternate Readers

First Lastname
Professor
Department of ChangeMe
Johns Hopkins Bloomberg School of Public Health

First Lastname
Assistant Professor
Department of ChangeMe
Johns Hopkins Bloomberg School of Public Health

Acknowledgments

I have tremendous gratitude
to those people,
numerous and uncountable,
who have contributed,
directly or in subtler ways,
to this work.

Some of them are listed here.

To my advisor, Winston: I remember writing to you as a sophomore in college many years ago, asking to do research in your brand new lab, which at the time was but a few months old. Back then, I hardly knew what research was and had no relevant skills or credentials to offer, only my time and my interest to learn. Over these years I've learned so much from you, and will always be grateful to you for building the place where I was able to grow.

To my thesis committee, Trish and Steven: Thank you for your patient guidance, encouragement, advice, and for helping me to keep an eye on the bigger picture.

To the @yfan arc of the #core channel - @isac, @brochael, @shao, @gilfunk,

@narley, @broham, @gmoney, @Brittany, @sherbear, @Sam Sholes, @Paul Hook, @amymeltzer39, and @alice: Thank you for those times when you patiently watched over me as I learned new lab techniques, answered my dumb questions, and generally saved me from my own buffoonery. Thank you even more for commiserating with me as we struggled together through the singular challenges of research, and celebrating the equally singular triumphs.

To the crew that moved me into Boonique, and Charles, and Charlotte, and Sven, and Manolo: Thanks for being there.

To mom and dad, and family further away: It was your labor that first cultivated my growth. Accomplishments in my name are as much yours as they are mine. I flourish for you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Introduce your thesis (Aardvark, 1900)

References

Aardvark, A. A. (1900). “Article title”. In: *Journal One* 1.1, pp. 1–8.

Chapter 2

Genome assembly of *Candida nivariensis*

Portions of this chapter originally appeared in:

Fan Y, Gale AN, Bailey A, Barnes K, Colotti K, Mass M, et al. Genome and transcriptome of a pathogenic yeast, *Candida nivariensis*. G3 Genes | Genomes | Genetics. 2021;11. doi:10.1093/g3journal/jkab137

2.1 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 11.8Mb 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*.

2.2 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, Zala, and McDonald, 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, Naumova, and Louis, 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, Murray, and Verstrepen, 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, Murray, and Verstrepen, 2010).

One of the gene families of great interest to the pathogenic yeast field are the GPI-anchored cell wall proteins. 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 248Kb (Gabaldón et al., 2013), meaning that at least half of the total genome length is contained in contigs 248Kb 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 has been shown to produce much more contiguous assemblies, and have been crucial

in sequencing through large repetitive regions, as well as assessing structural variants. However, read accuracy on the ONT platform in particular ranges from 86% for early basecaller versions (Wick, Judd, and Holt, 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, Judd, and Holt, 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, Judd, and Holt, 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 this data with liftover of annotations from evolutionary "cousins" of nivariensis, we have generated a new and annotated reference genome for the community.

2.3 Results

2.3.1 Genome statistics

Using our nanopore and Illumina sequencing data, we generated a new assembly of *Candida nivariensis*, JHU_Cniv_v1 (Methods). Our assembly consists of 11.8 Mb of sequence in 16 contigs with an N50 of 886 Kb (**Figure 2.2**, **Table 2.1**). Compared to the reference genome we have 275kb of additional sequence, 218kb 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 multi-mapping short reads (>10% in gap regions vs 7% average across the genome).

	Contigs	N50	Longest Contig	Shortest Contig	Total Length
Reference	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

Table 2.1: Assembly Statistics. Assembly statistics of JHU_Cniv_v1 and the reference genome for *C. nivariensis*

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 CTGGGTGCTGTGGGGT, the telomere sequence of *Candida glabrata*(McEachern and Blackburn 1994). The other 4

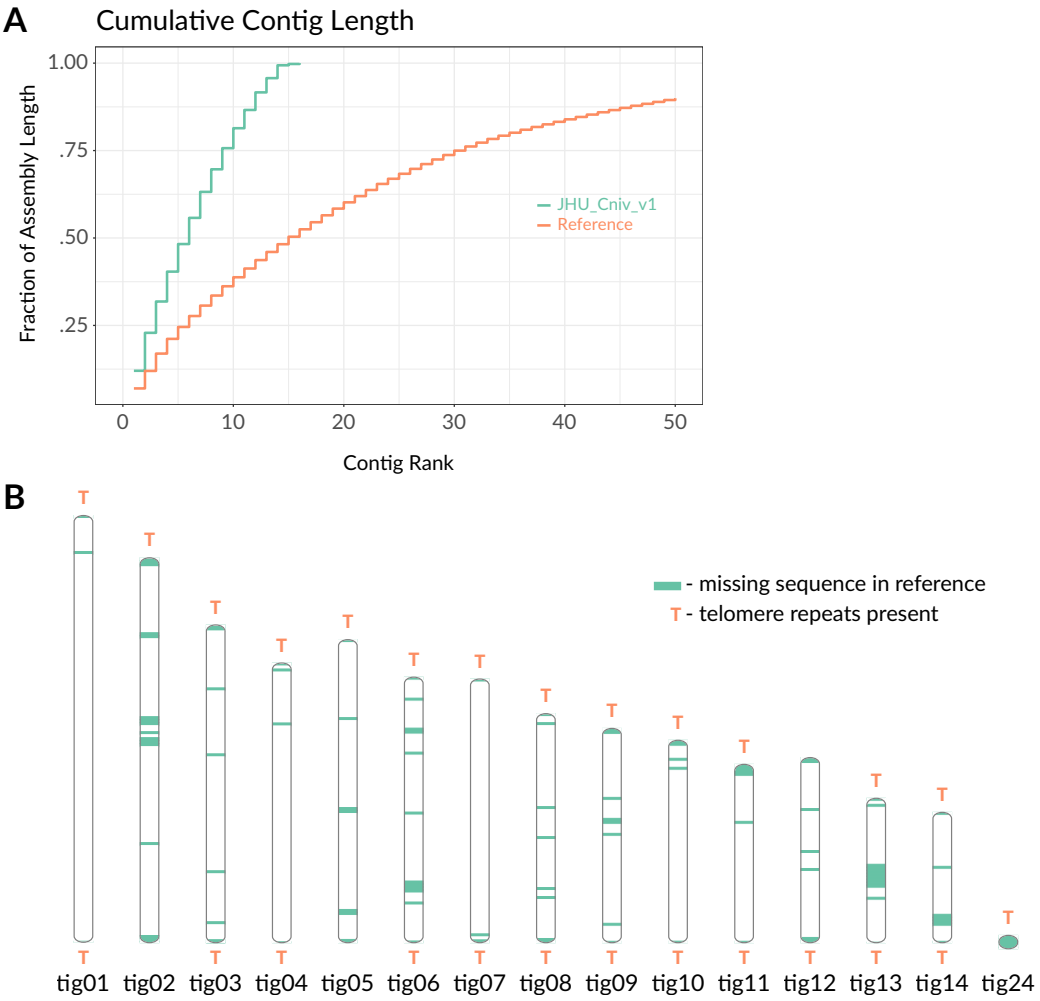

Figure 2.1: Characteristics of the JHU_Cniv_v1 assembly. (A) Cumulative lengths of the 50 longest sequences in our assembly and previous reference genome. **(B)** Ideogram of assembly. Sequence that is missing in the reference genome is shown along each non-mitochondrial contig, and the positions of telomere repeats are marked.

non-mitochondrial sequences terminate only at one end in this telomeric repeat (Figure 1b, Supp. Table 1), 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 (Supp. Figure 3). Upon aligning the *C. glabrata* genome to JHU_Cniv_v1 using Mummer, we found only sporadic shared segments of negligible length (Supp. Figure 4), as opposed to a nearly perfect 1:1 alignment between JHU_Cniv_v1 and the current *C. nivariensis* reference genome (Supp. Figure 5). 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.

2.3.2 Genome completeness

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/

2.3.3 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 ribosomal genes appear once in the reference, and

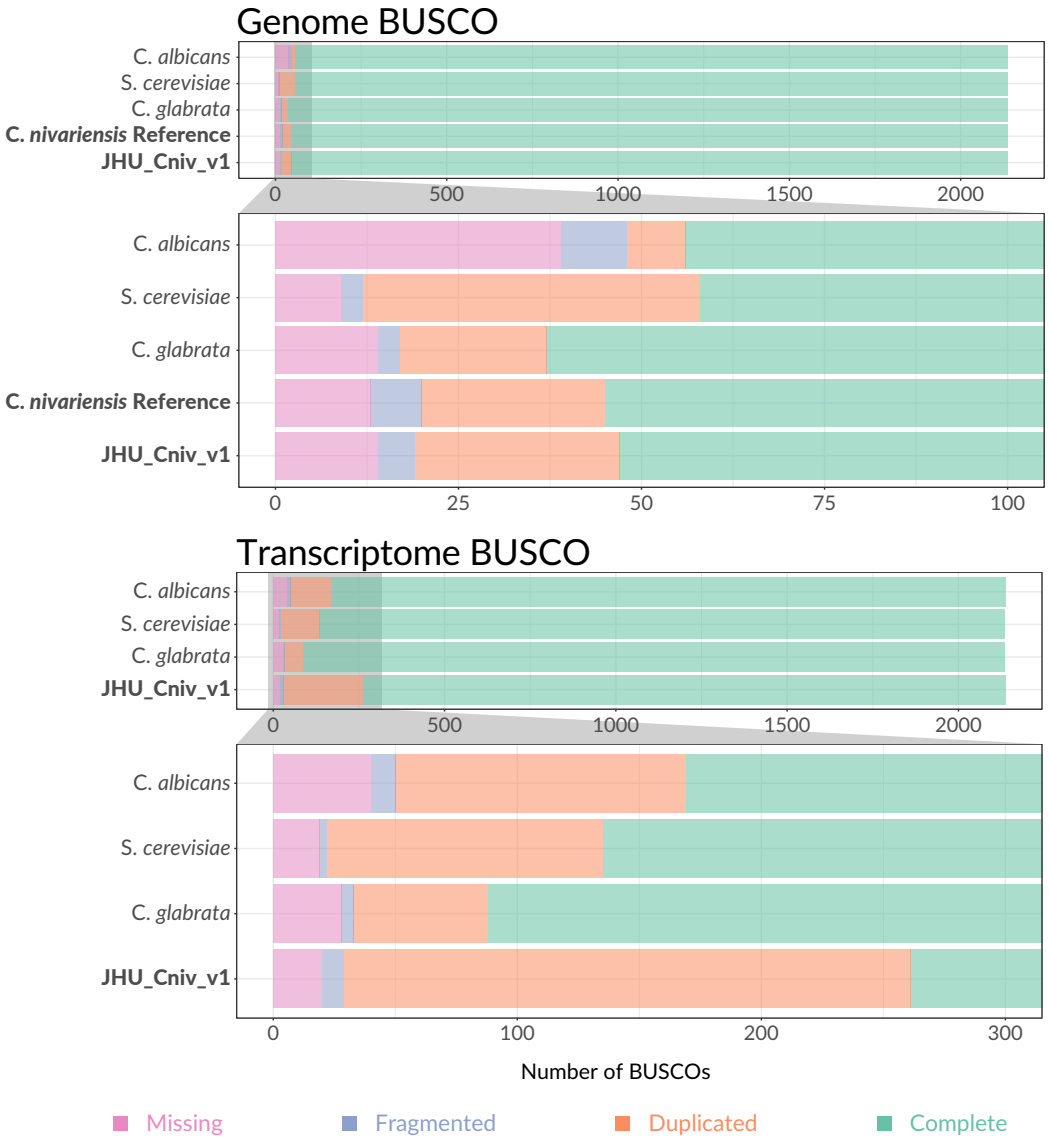

Figure 2.2: Completeness of the JHU_Cniv_v1 assembly. Genome and transcriptome completeness Bar charts comparing BUSCOs detected in JHU_Cniv_v1 and accompanying transcriptome to those of the current C. albicans, extitS. cerevisiae, extitC. glabrata, and extitC. nivariensis reference genomes. No reference transcriptome is currently available for extitC. nivariensis.

either four or six times in JHU_Cniv_v1 (Figure 3A).

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 3.5kb, and 27 of these adhesin genes are not found continuously, suggesting the previous reference either truncated or did not continuously assemble these important pathogenicity genes.

2.4 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.*

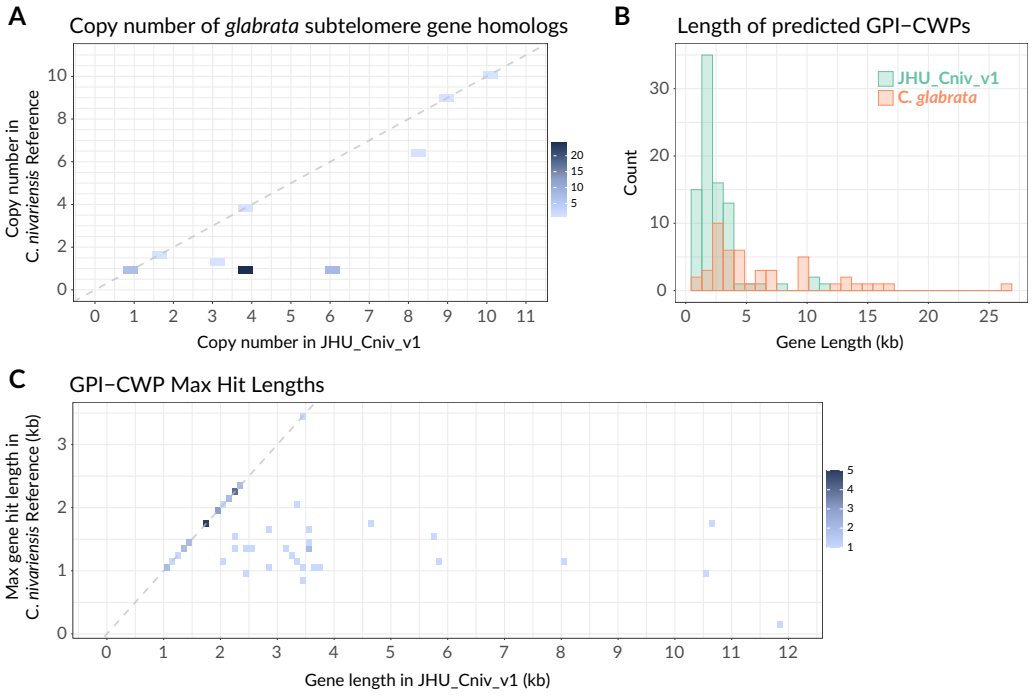

Figure 2.3: GPI genes. (A) Scatterplot showing the number of times each glabrata subtelomere gene homolog appears in the extitC. nivariensis reference genome and in JHU_Cniv_v1. Overlapping points are shown on the color scale, and the y=x line is shown in dashed gray. **(B)** Histogram of adhesion protein 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 extitC. nivariensis reference genome. Overlapping points are shown on the color scale, and the y=x line is shown in dashed gray.

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).

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.

2.5 Methods

2.5.1 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.

2.5.2 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 400ng of extracted DNA. Both Illumina libraries were then sequenced on a single iSeq 100 run.

2.5.3 RNA isolation and 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 200ng 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 400ng of cDNA. Both library replicates were then sequenced on a single iSeq 100 run, generating 2×150 paired-end reads.

2.5.4 Genome assembly

Nanopore data were basecalled using Guppy v3.2.4 on default settings. Reads greater than 3kb 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 11m. 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 3662bp 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 (Xu et al. 2020): `match = 2, mismatch = 7, delta = 7, pm = 80, pi = 10, minscore = 50, maxperiod = 600`. Multimapping short reads were identified using bwa mem (Li 2013) on default settings.

2.5.5 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 (Pertea and Pertea 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.

2.5.6 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 here 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/timplab/nivar.

Contig	Length (bp)	Forward Telomeres	Reverse Telomeres
tig01	1423475	35	38
tig02	1283968	0	39
tig03	1060011	35	39
tig04	933062	36	26
tig05	1010854	0	36
tig06	885783	35	38
tig07	879540	39	35
tig08	763992	34	33
tig09	714796	35	47
tig10	675194	36	36
tig11	594828	32	26
tig12	617546	36	0
tig13	481613	38	41
tig14	434809	33	33
tig24	44616	0	39
JHU_Cniv_v1_mito	28512	0	0

Table 2.2: Contig and telomere lengths. Contig lengths and the number of times the forward and reverse telomere sequence appears in each

	Total	Gene	Exon
Augustus (BRAKER)	23,497	5,028	6,109
Genemark.hmm (BRAKER)	36	6	12
Liftoff glabrata	263	130	2
Liftoff cerevisiae	42	21	0
Liftoff albicans	0	0	0
StringTie	2,141	824	1,175

Table 2.3: Contributions from each annotation software. Number of genes and exons added by each software

	Total Exons	Total Genes
JHU_Cniv_v1	7,298	5,859
C. *glabrata*	5,629	5,448
S. *cerevisiae*	6,760	6,420
C. *albicans*	6,732	6,263

Table 2.4: Gene and exon counts of JHU_Cniv_v1 and related yeasts. Gene and exon counts of our annotation and currently available reference annotations

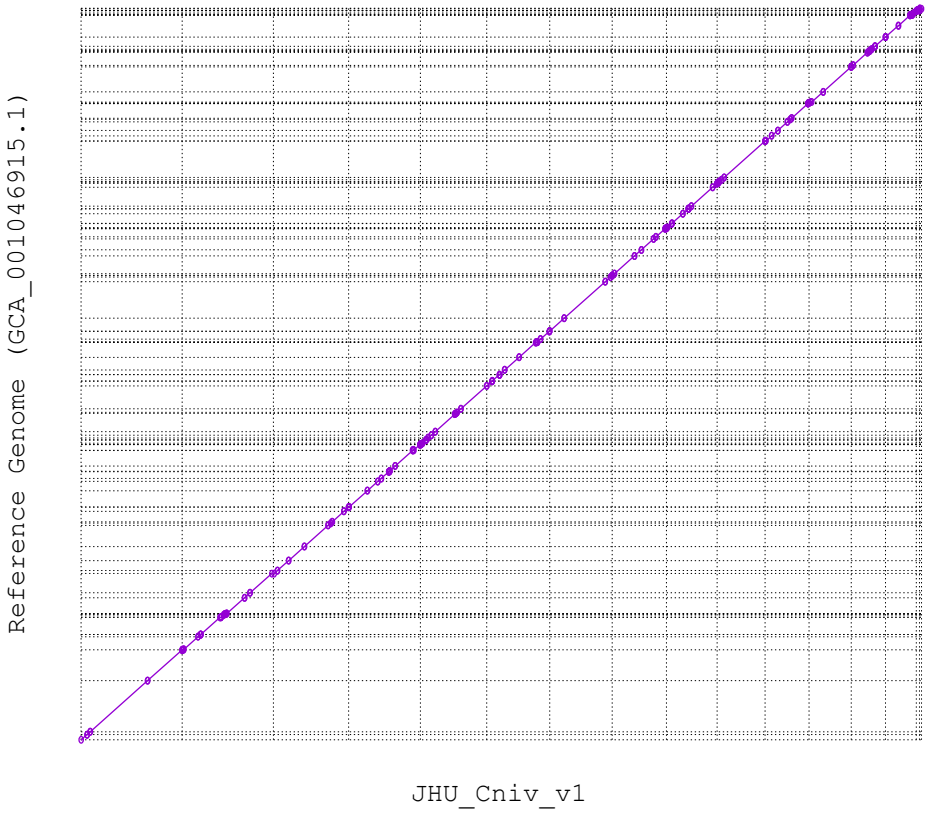

Figure 2.4: Whole genome alignment of JHU_Cniv_v1 and the extitC. nivariensis reference genome. Whole genome alignment of the current reference genome (y axis) compared to our new assembly (x axis). Alignments match with no notable structural variants, and very little missing or duplicated sequence.

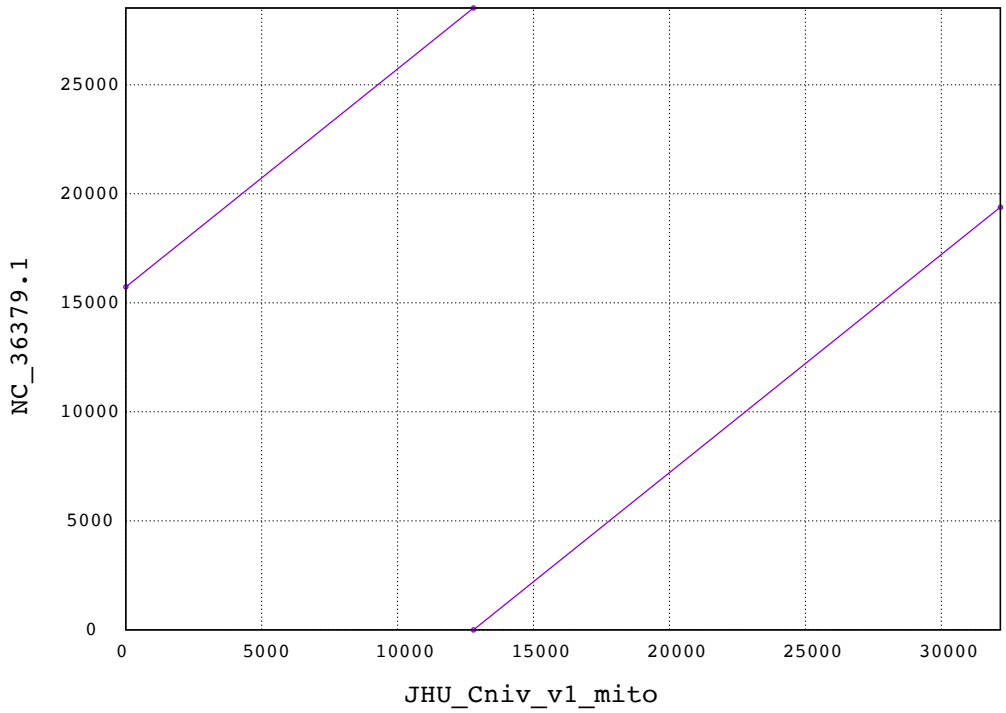

Figure 2.5: Alignment of JHU_Cniv_v1 mitochrondrial contig and the extitC. nivariensis mitochondrial genome. Alignment of our 32Kb circular contig (x axis) with the completed mitochondrial genome of the extitC. nivariensis reference genome (y axis). The final 3662bp of this contig appears twice in the reference genome.

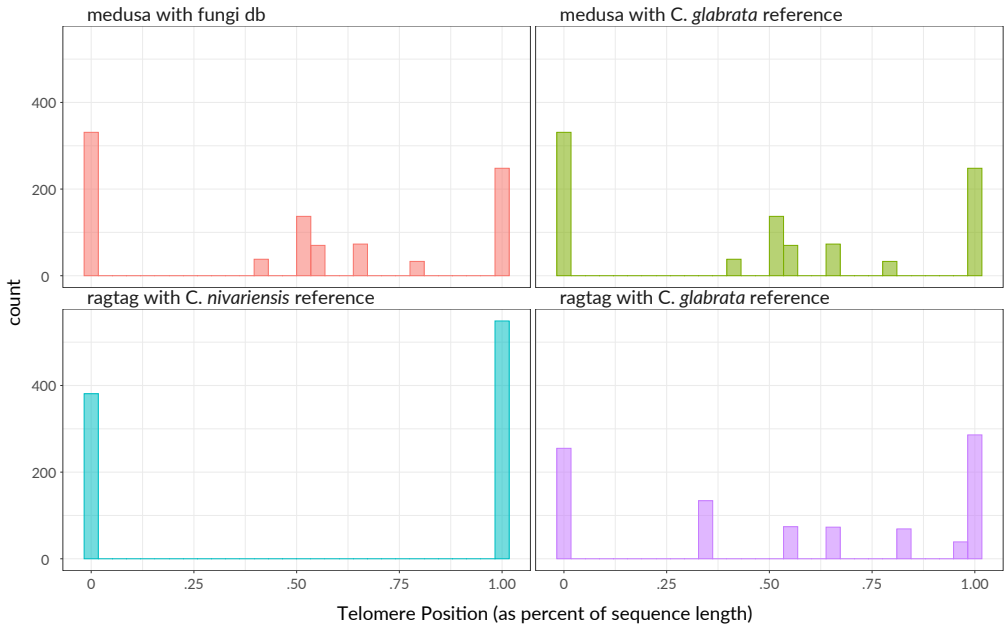

Figure 2.6: Telomere positions reference based scaffolds. Histogram of telomere repeat positions in our assembly, and in scaffolds produced by RagTag and MeDuSa. When MeDuSa is used with a database including the reference genomes of extitC. nivariensis, extitC. glabrata, C. bracarensis, and N. delphensis, telomeres are placed in the middle of contigs. The same result is produced when only the extitC. glabrata genome is used for scaffolding with MeDuSa, and MeDuSa fails to run when only the extitC. nivariensis reference is used. When the extitC. nivariensis reference genome is used for scaffolding with RagTag, no changes are made. When the more contiguous extitC. glabrata genome is used with RagTag, telomere sequences are again placed in the middle of sequences, suggesting a scaffolding error.

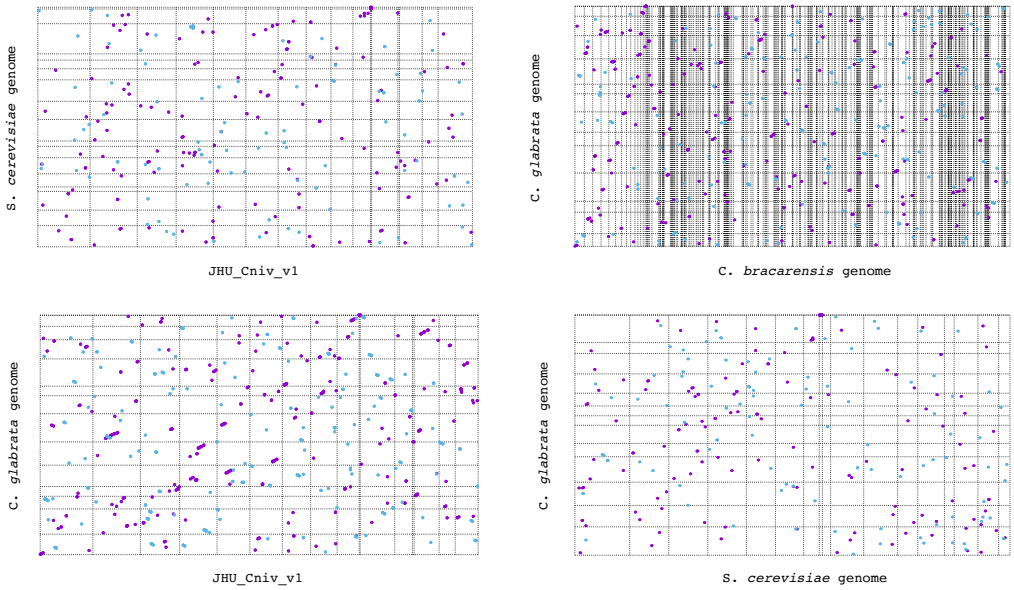

Figure 2.7: Whole genome alignments between related yeasts. Whole genome alignment of our new assembly against the extitS. cerevisiae (top left), and extitC. glabrata (bottom left) reference genomes. For both, there are no long alignments, suggesting that there is little similarity in genome structure between these species and extitC. nivariensis. C. bracarensis, a close relative to both extitC. glabrata and extitC. nivariensis, also shares little genome similarity to extitC. glabrata (top right), suggesting that yeast genomes within the glabrata clade are not generally similar enough to support inter-species reference based scaffolding. We also compared extitC. glabrata to the highly contiguous and complete extitS. cerevisiae genome (bottom right) to check that genome contiguity alone did not bias the genome similarity detected.

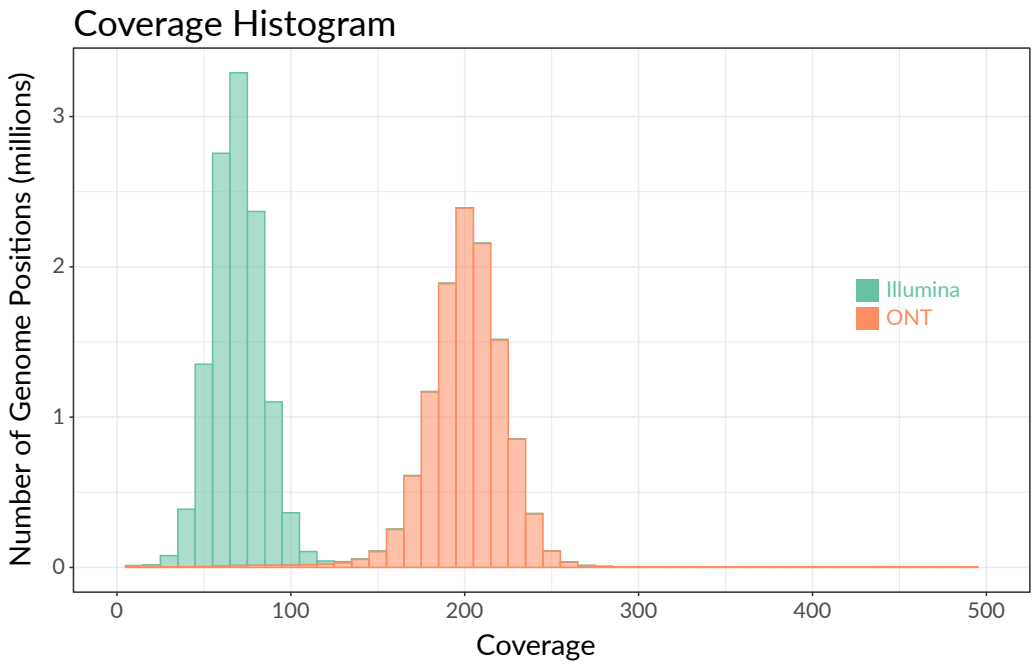

Figure 2.8: Coverage histograms. Histogram of coverage per base in our assembly by filtered (>3kb) ONT reads and trimmed Illumina reads.

References

Borman, Andrew M, Rebecca Petch, Christopher J Linton, Michael D Palmer, Paul D Bridge, and Elizabeth M Johnson (2008). "Candida nivariensis, an emerging pathogenic fungus with multidrug resistance to antifungal agents". en. In: *J. Clin. Microbiol.* 46.3, pp. 933–938.

Aznar-Marin, Pilar, Fátima Galan-Sanchez, Pilar Marin-Casanova, Pedro García-Martos, and Manuel Rodríguez-Iglesias (2016). "Candida nivariensis as a New Emergent Agent of Vulvovaginal Candidiasis: Description of Cases and Review of Published Studies". en. In: *Mycopathologia* 181.5-6, pp. 445–449.

Gabaldón, Toni, Tiphaine Martin, Marina Marcet-Houben, Pascal Durrens, Monique Bolotin-Fukuhara, Olivier Lespinet, Sylvie Arnaise, Stéphanie Boisnard, Gabriela Aguileta, Ralitsa Atanasova, Christiane Bouchier, Arnaud Couloux, Sophie Creno, Jose Almeida Cruz, Hugo Devillers, Adela Enache-Angoulvant, Juliette Guitard, Laure Jaouen, Laurence Ma, Christian Marck, Cécile Neuvéglise, Eric Pelletier, Amélie Pinard, Julie Poulain, Julien Recoquillay, Eric Westhof, Patrick Wincker, Bernard Dujon, Christophe Hennequin, and Cécile Fairhead (2013). "Comparative genomics of emerging pathogens in the Candida glabrata clade". en. In: *BMC Genomics* 14, p. 623.

Croll, Daniel, Marcello Zala, and Bruce A McDonald (2013). "Breakage-fusion-bridge cycles and large insertions contribute to the rapid evolution of accessory chromosomes in a fungal pathogen". en. In: *PLoS Genet.* 9.6, e1003567.

Ford, Christopher B, Jason M Funt, Darren Abbey, Luca Issi, Candace Guiducci, Diego A Martinez, Toni Delorey, Bi Yu Li, Theodore C White, Christina Cuomo, Reeta P Rao, Judith Berman, Dawn A Thompson, and Aviv Regev (2015). "The evolution of drug resistance in clinical isolates of Candida albicans". en. In: *Elife* 4, e00662.

López-Fuentes, Eunice, Guadalupe Gutiérrez-Escobedo, Bea Timmermans, Patrick Van Dijck, Alejandro De Las Peñas, and Irene Castaño (2018). "Candida glabrata's Genome Plasticity Confers a Unique Pattern of Expressed Cell Wall Proteins". en. In: *J Fungi (Basel)* 4.2.

Carreté, Laia, Ewa Ksiezopolska, Emilia Gómez-Molero, Adela Angoulvant, Oliver Bader, Cécile Fairhead, and Toni Gabaldón (2019). "Genome Comparisons of Candida glabrata Serial Clinical Isolates Reveal Patterns of Genetic Variation in Infecting Clonal Populations". en. In: *Front. Microbiol.* 10, p. 112.

Todd, Robert T, Tyler D Wikoff, Anja Forche, and Anna Selmecki (2019). "Genome plasticity in Candida albicans is driven by long repeat sequences". en. In: *Elife* 8.

Barry, J D, M L Ginger, P Burton, and R McCulloch (2003). "Why are parasite contingency genes often associated with telomeres?" en. In: *Int. J. Parasitol.* 33.1, pp. 29–45.

De Las Peñas, Alejandro, Shih-Jung Pan, Irene Castaño, Jonathan Alder, Robert Cregg, and Brendan P Cormack (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". en. In: *Genes Dev.* 17.18, pp. 2245–2258.

Naumov, G I, E S Naumova, and E J Louis (1995). "Genetic mapping of the alpha-galactosidase MEL gene family on right and left telomeres of Saccharomyces cerevisiae". en. In: *Yeast* 11.5, pp. 481–483.

Iraqui, Ismail, Susana Garcia-Sanchez, Sylvie Aubert, Françoise Dromer, Jean-Marc Ghigo, Christophe d'Enfert, and Guilhem Janbon (2005). "The Yak1p kinase controls expression of adhesins and biofilm formation in Candida glabrata in a Sir4p-dependent pathway". en. In: *Mol. Microbiol.* 55.4, pp. 1259–1271.

Carreto, Laura, Maria F Eiriz, Ana C Gomes, Patrícia M Pereira, Dorit Schuller, and Manuel A S Santos (2008). "Comparative genomics of wild type yeast strains unveils important genome diversity". en. In: *BMC Genomics* 9, p. 524.

Brown, Chris A, Andrew W Murray, and Kevin J Verstrepen (2010). "Rapid expansion and functional divergence of subtelomeric gene families in yeasts". en. In: *Curr. Biol.* 20.10, pp. 895–903.

Anderson, Matthew Z, Lauren J Wigen, Laura S Burrack, and Judith Berman (2015). "Real-Time Evolution of a Subtelomeric Gene Family in Candida albicans". en. In: *Genetics* 200.3, pp. 907–919.

Timmermans, Bea, Alejandro De Las Peñas, Irene Castaño, and Patrick Van Dijck (2018). "Adhesins in Candida glabrata". en. In: *J Fungi (Basel)* 4.2.

McCall, Andrew D, Ruvini U Pathirana, Aditi Prabhakar, Paul J Cullen, and Mira Edgerton (2019). "Candida albicans biofilm development is governed by cooperative attachment and adhesion maintenance proteins". en. In: *NPJ Biofilms Microbiomes* 5.1, p. 21.

Carreté, Laia, Ewa Ksiezopolska, Cinta Pegueroles, Emilia Gómez-Molero, Ester Saus, Susana Iraola-Guzmán, Damian Loska, Oliver Bader, Cecile Fairhead, and Toni Gabaldón (2018). "Patterns of Genomic Variation in the Opportunistic Pathogen Candida glabrata Suggest the Existence of Mating and a Secondary Association with Humans". en. In: *Curr. Biol.* 28.1, 15–27.e7.

Wick, Ryan R, Louise M Judd, and Kathryn E Holt (2019). "Performance of neural network basecalling tools for Oxford Nanopore sequencing". en. In: *Genome Biol.* 20.1, p. 129.

Fox, Edward J, Kate S Reid-Bayliss, Mary J Emond, and Lawrence A Loeb (2014). "Accuracy of Next Generation Sequencing Platforms". en. In: *Next Gener Seq Appl* 1.

Watson, Mick and Amanda Warr (2019). "Errors in long-read assemblies can critically affect protein prediction". en. In: *Nat. Biotechnol.* 37.2, pp. 124–126.

Garrison, Erik and Gabor Marth (2012). "Haplotype-based variant detection from short-read sequencing". In: arXiv: 1207.3907 [q-bio.GN].

Walker, Bruce J, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K Young, and Ashlee M Earl (2014). "Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement". en. In: *PLoS One* 9.11, e112963.

Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić (2017). "Fast and accurate de novo genome assembly from long uncorrected reads". en. In: *Genome Res.* 27.5, pp. 737–746.

Salzberg, Steven L (2019). "Next-generation genome annotation: we still struggle to get it right". en. In: *Genome Biol.* 20.1, p. 92.

Chapter 3

Discussion and Conclusion

Discuss and conclude your thesis (Abramson, Barbie, and Rider, 1900)

References

Abramson, A. A., B. B. Barbie, and C. C. Rider (1900). "Article title". In: *Journal Three* 1.1, pp. 192–244.

John Doe

Resumé title

Some quote

Education

year–year **Degree**, *Institution*, City, *Grade*.
Description

year–year **Degree**, *Institution*, City, *Grade*.
Description

Master thesis

title *Title*

supervisors Supervisors

description Short thesis abstract

Experience

Vocational

year–year **Job title**, *Employer*, City.
General description no longer than 1–2 lines.
Detailed achievements:
- Achievement 1;
- Achievement 2, with sub-achievements:
  - Sub-achievement (a);
  - Sub-achievement (b), with sub-sub-achievements (don't do this!);
    - Sub-sub-achievement i;
    - Sub-sub-achievement ii;
    - Sub-sub-achievement iii;
  - Sub-achievement (c);
- Achievement 3.

year–year **Job title**, *Employer*, City.
Description line 1
Description line 2

Miscellaneous

street and number – postcode city – country
+1 (234) 567 890 • *+2 (345) 678 901* • *+3 (456) 789 012*
john@doe.org • *www.johndoe.com* • *john.doe* • *jdoe*
jdoe • *additional information*

year–year **Job title**, *Employer*, City.
Description

Languages

Language 1 Skill level *Comment*
Language 2 Skill level *Comment*
Language 3 Skill level *Comment*

Computer skills

category 1 XXX, YYY, ZZZ
category 2 XXX, YYY, ZZZ
category 3 XXX, YYY, ZZZ
category 4 XXX, YYY, ZZZ
category 5 XXX, YYY, ZZZ
category 6 XXX, YYY, ZZZ

Interests

hobby 1 Description
hobby 2 Description
hobby 3 Description

Extra 1

- Item 1
- Item 2
- Item 3. This item is particularly long and therefore normally spans over several lines. Did you notice the indentation when the line wraps?

Extra 2

- Item 1
- Item 2
- Item 3
- Item 4
- Item 5[3]
- Item 6. Like item 3 in the single column list before, this item is particularly long to wrap over several lines.

References

Category 1

- Person 1
- Person 2
- Person 3

Category 2

Amongst others:

- Person 1, and
- Person 2

(more upon request)

All the rest & some more

That person, and **those** also (all available upon request).

Publications

[1] John Doe. Title, year.

street and number – postcode city – country
+1 (234) 567 890 • *+2 (345) 678 901* • *+3 (456) 789 012*
john@doe.org • *www.johndoe.com* • *john.doe* • *jdoe*
jdoe • *additional information*

[2] John Doe. Title, year.

[3] John Doe and Author 1. *Title*. Publisher, edition edition, year.

[4] John Doe and Author 2. *Title*. Publisher, edition edition, year.

[5] John Doe and Author 3. Title, year.

Company Recruitment team January 01, 1984
Company, Inc.
123 somestreet
some city

Dear Sir or Madam,

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis ullamcorper neque sit amet lectus facilisis sed luctus nisl iaculis. Vivamus at neque arcu, sed tempor quam. Curabitur pharetra tincidunt tincidunt. Morbi volutpat feugiat mauris, quis tempor neque vehicula volutpat. Duis tristique justo vel massa fermentum accumsan. Mauris ante elit, feugiat vestibulum tempor eget, eleifend ac ipsum. Donec scelerisque lobortis ipsum eu vestibulum. Pellentesque vel massa at felis accumsan rhoncus.

Suspendisse commodo, massa eu congue tincidunt, elit mauris pellentesque orci, cursus tempor odio nisl euismod augue. Aliquam adipiscing nibh ut odio sodales et pulvinar tortor laoreet. Mauris a accumsan ligula. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Suspendisse vulputate sem vehicula ipsum varius nec tempus dui dapibus. Phasellus et est urna, ut auctor erat. Sed tincidunt odio id odio aliquam mattis. Donec sapien nulla, feugiat eget adipiscing sit amet, lacinia ut dolor. Phasellus tincidunt, leo a fringilla consectetur, felis diam aliquam urna, vitae aliquet lectus orci nec velit. Vivamus dapibus varius blandit.

Duis sit amet magna ante, at sodales diam. Aenean consectetur porta risus et sagittis. Ut interdum, enim varius pellentesque tincidunt, magna libero sodales tortor, ut fermentum nunc metus a ante. Vivamus odio leo, tincidunt eu luctus ut, sollicitudin sit amet metus. Nunc sed orci lectus. Ut sodales magna sed velit volutpat sit amet pulvinar diam venenatis.

Albert Einstein discovered that $e = mc^2$ in 1905.

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n$$

Yours faithfully,

John Doe

Attached: curriculum vitæ

John Doe
street and number – postcode city – country
+1 (234) 567 890 • *+2 (345) 678 901* • *+3 (456) 789 012*
john@doe.org • *www.johndoe.com* • *john.doe* • *jdoe*
jdoe • *additional information*