De novo whole-genome assembly of a wild type yeast isolate using nanopore sequencing [version 1; referees: 4 approved with reservations]

Hans J. Jansen¹, Ron P. Dirks¹, Michael Liem², Christiaan V. Henkel², G. Paul H. van Heusden², Richard J.L.F. Lemmers³, Trifa Omer⁴, Shuai Shao², Peter J. Punt²,⁴, Herman P. Spaink²

¹Future Genomics Technologies B.V., Leiden, 2333 CH, The Netherlands
²Institute of Biology, Leiden University, Leiden, 2300 RA, The Netherlands
³Department of Human Genetics, Leiden University Medical Center, Leiden, 333 ZA, The Netherlands
⁴Dutch DNA Biotech B.V., Zeist, 3700 AJ, The Netherlands

Abstract

Background: The introduction of the MinION™ sequencing device by Oxford Nanopore Technologies may greatly accelerate whole genome sequencing. It has been shown that the nanopore sequence data, in combination with other sequencing technologies, is highly useful for accurate annotation of all genes in the genome. However, it also offers great potential for de novo assembly of complex genomes without using other technologies. In this manuscript we used nanopore sequencing as a tool to classify yeast strains.

Methods: We compared various technical and software developments for the nanopore sequencing protocol, showing that the R9 chemistry is, as predicted, higher in quality than R7.3 chemistry. The R9 chemistry is an essential improvement for assembly of the extremely AT-rich mitochondrial genome.

Results: In this study, we used this new technology to sequence and de novo assemble the genome of a recently isolated ethanologenic yeast strain, and compared the results with those obtained by classical Illumina short read sequencing. This strain was originally named (Candida vartiovaarae Torulopsis) based on ribosomal RNA sequencing. We show that the assembly using nanopore data is much more contiguous than the assembly using short read data.

Conclusions: The mitochondrial and chromosomal genome sequences showed that our strain is clearly distinct from other yeast taxons and most closely related to published Cyberlindnera species. In conclusion, MinION-mediated long read sequencing can be used for high quality de novo assembly of new eukaryotic microbial genomes.

Keywords

Nanopore sequencing, de novo genome assembly, wild type yeasts, ethanologenic, Candida, Cyberlindnera
This article is included in the Nanopore Analysis gateway.

Corresponding author: Michael Liem (m.liem@biology.leidenuniv.nl)

Competing interests: HJJ and CVH are members of the Nanopore Community, and have previously received flow cells free of charge, as well as travel expense reimbursements from Oxford Nanopore Technologies.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2017 Jansen HJ et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

How to cite this article: Jansen HJ, Dirks RP, Liem M et al. De novo whole-genome assembly of a wild type yeast isolate using nanopore sequencing [version 1; referees: 4 approved with reservations] F1000Research 2017, 6:618 (doi: 10.12688/f1000research.11146.1)

First published: 03 May 2017, 6:618 (doi: 10.12688/f1000research.11146.1)
Introduction

With the development of robust second generation bioethanol processes, next to the use of highly engineered Saccharomyces cerevisiae strains\(^1\), non-classical ethanologenic yeasts are also being considered as production organisms\(^2\). In particular, aspects concerning the ability to use both C6 and C5 sources and feedstock derived inhibitor resistance have been identified as important for the industrial applicability of different production hosts\(^3\). In our previous studies we have identified a novel ethanologenic yeast, Wickerhamomyces anomalala, as a potential candidate\(^1\). Based on this research, a further screen for alternative yeast species was initiated (Punt and Omer, unpublished study) Here we describe the isolation and genomic characterization of one of these new isolates, which was typed as Candida variovaarae based on ribosomal RNA analysis.

With the arrival of next generation sequencing and the assemblers that can use this type of sequencing data, whole genome shotgun sequencing of completely novel organisms has become affordable and accessible. As a result, a wealth of genomic information has become available to the scientific community leading to many important discoveries. While generating whole draft genomes has become accessible, these genomes are often fragmented due to the nature of these short read technologies\(^4\). Assembling the short read data into large contigs proved to be difficult because the short reads do not contain the information to span repeated structures in the genome. Approaches to sequence the ends of larger fragments partially mitigated this problem\(^5\).

The new long read platforms from Pacific Biosciences and Oxford Nanopore Technologies made it possible to obtain reads that span many kilobases\(^6\). Assemblies using this type of data are often more contiguous than assemblies based on short read data\(^6\).

We have employed the Oxford Nanopore Technologies MinION\(\text{TM}\) device to sequence genomic DNA from the isolated Candida variovaarae strain. The same DNA was also used to prepare a paired end library for sequencing on the Illumina HiSeq2500. The sequence data were used in various assemblers to obtain the best assemblies.

Materials and methods

Strain selection and cultivation conditions

In our previous research\(^1\), a screening approach was developed to select for potential ethanologens using selective growth on industrial feedstock hydrolysates. Based on this approach, a previously identified microflora from grass silage was screened for growth on different hydrolysates from both woody and cereal residues. From this microflora, a strain was isolated (DDNA#1) after selection on a growth medium consisting of 10% acid-pretreated corn stover hydrolysate, which was shown to be most restrictive in growth due to the presence of relatively high amounts of furanic inhibitors.

DNA purification

Cells were grown at 30°C on plates with YNB (without amino acids) medium supplemented with 0.5% glucose. Cells were scraped from plates and resuspended in 5 ml TE. High MW chromosomal DNA was isolated from yeast isolate DNA#1 and Saccharomyces cerevisiae S288C using a Genomic-tip 100/G column, according to the manufacturer’s instructions (Qiagen).

Pulsed field gel electrophoresis

To isolate intact chromosomal DNA from DDNA#1, a BioRad CHEF Genomic DNA Plug Kit was used. Briefly, yeast cells were treated with lyticase and the resulting spheroplasts were embedded in low melting point agarose. After incubation with RNase A and Proteinase K, the agarose plugs were thoroughly washed in TE. The DNA in the agarose plugs was separated on a 0.88% agarose gel in 1xTAE buffer on a Bio-Rad CHEF DRII system. The DNA was separated in four subsequent 12 hour runs at 3V/cm; run one and two used a constant switching time of 500 seconds, and in run three and four the switching time increased from 60 seconds to 120 seconds. The gel was afterwards stained with ethidium bromide and imaged.

Illumina library preparation and sequencing

High molecular weight DNA from both DDNA#1 and Saccharomyces cerevisiae S288C was sheared using a nebulizer (Life Technologies). The sheared DNA was used to make genomic DNA libraries using the TrueSeq\(\text{TM}\) DNA sample preparation kit, according to the manufacturer’s instructions (Illumina Inc.). In the size selection step, a band of 330–350 bp was cut out of the gel to obtain an insert length of ~270 bp. From the resulting libraries, 4.5 million fragments were sequenced in paired end reads with a read length of 150 nt on an Illumina HiSeq2500, according to the manufacturer’s instructions. The HiSeq control software (HCS) and real time analysis (RTA) software, versions were 2.2.38 and 1.18.61, respectively, were used.

MiniON library preparation and sequencing

The genomic DNA was sequenced using nanopore sequencing technology. First the DNA was sequenced on R7.3 Flow Cells. Subsequently, multiple R9 and R9.4 Flow Cells were used to sequence the DNA. For R7.3 sequencing runs, we prepared the library using the SQK-MAP006 kit from Oxford Nanopore Technologies. In short, high molecular weight DNA was sheared with a g-TUBE (Covaris) to an average fragment length of 20 kbp. The sheared DNA was repaired using the FFPE Repair Mix, according to the manufacturer’s instructions (New England Biolabs). After cleaning the DNA with using an extraction process, using a ratio of 0.4:1 Ampure XP beads (Beckman Coulter) to DNA, the DNA ends were polished and an A overhang was added with the NEBNext DNA End Prep Module (New England Biolabs). Then, prior to ligation, the DNA was again cleaned with an extraction using a ratio of 1:1 Ampure XP beads to DNA. The adaptor and hairpin adapter were ligated using Blunt/TA Ligase Master Mix (New England Biolabs).
The final library was prepared by cleaning the ligation mix using MyOne C1 beads (Invitrogen).

To prepare 2D libraries for R9 sequencing runs, we used the SQK-NSK007 kit from Oxford Nanopore Technologies. The procedure to prepare a library with this kit is largely the same as with the SQK-MAP006 kit. 1D library preparation was done with the SQK-RAD001 kit from Oxford Nanopore Technologies. In short, high molecular weight DNA was tagmented with a transposase. The final library was prepared by ligation of the sequencing adapters to the tagmented fragments using the Blunt/TA Ligase Master Mix (New England Biolabs).

The prepared libraries were loaded on the MinION flow cell, which was docked on the MinION device. The MinKNOW software (version 0.50.2.15 for SQK-MAP006 libraries and version 1.0.5 for SQK-NSK007 and SQK-RAD001 libraries) was used to control the sequencing process and the read files were uploaded to the cloud based Metrichor EPI2ME platform for base calling. Base called reads were downloaded for further processing and assembly.

**Genome assembly**

The sequence data from the Illumina platform was assembled using the Spades assembler (version 3.6.0), either alone or in combination with the nanopore data. From the base called read files produced by the Metrichor EPI2ME platform, a sequence file in fasta format was extracted using the R-package poRe v0.17. For the assembly of the nanopore data, Canu v1.3 was used. After assembly, the resulting contigs were polished with the short read data using PILON v1.18. The sequencing data has been submitted to the European Nucleotide Archive and can be accessed at http://www.ebi.ac.uk/ena/data/view/PRJEB19912.

**Gene size estimation and heterozygosity**

A k-mer count analysis was done using Jellyfish (version 2.2.6) on the Illumina data. From the paired end reads, only the first read was truncated to 100 bp to avoid the lower quality part of the read. The second read was omitted from this analysis to avoid counting overlapping k-mers. Different k-mer sizes were used ranging from k=17 to 23. After converting the k-mer counts into a histogram format, this file was analyzed using the Genomescope tool, available at http://qb.cshl.edu/genomescope/ and https://github.com/schatzlab/genomescope.

**Full genome comparison**

From 26S ribosomal RNA sequences available in the nucleotide database, Chen et al. have constructed a phylogenetic tree. The closest relative for which whole genome sequences are available is *Cyberlindnera jadinii*. To compare our draft genome assembly to this yeast species, we retrieved assemblies of two *Cyberlindnera jadinii* strains, namely NBRC 0988 (GenBank accession number, DG000077.1) and CBS1600 (GenBank accession number, CDQK00000000.1). We also used *Saccharomyces cerevisiae* S288C (GenBank accession number, GCA_000146045.2) in this comparison. We aligned those assemblies to the corrected draft assembly of our strain using MUMmer’s alignment generator NUCmer (version 3.1). NUCmer’s output was filtered with delta-filter, and the filtered results parsed to MUMmerplot, generating full-genome visualization between the pairs of different yeast species.

**Read mapping**

Reads generated on the Illumina platform were aligned to the published *Candida vartiovaarae* mitochondrial genome (Genbank accession number, KC993190.1) using Bowtie2 (version 2.2.5). Reads generated on the MinION platform were aligned using BWA-mem (version 0.7.15) with -x ont2d settings. Resulting bam files were sorted and viewed in IGV viewer (version 2.3).

**Results and discussion**

**Pure cultures of candidate ethanologenic yeasts**

From a screen on 10% acid-pretreated corn stover hydrolysate, about 70 individual clones were obtained, only five of which were able to grow well on purely synthetic YNB-based medium. To determine the taxonomic status of these clones, chromosomal DNA was isolated and used for PCR amplification of the ribosomal ITS sequence using ITS specific primers (ITS1 and ITS4).

BLAST analysis of these ITS sequences of all 5 isolates revealed a 100% identity to *Candida vartiovaarae* (*Torulopsis vartiovaarae*; NCBI accession number KY102493).

All five isolates were grown on different C-sources and showed growth on glucose, mannose, cellobiose, xylose and glycerol, while growth on L-arabinose was variable. No significant growth was found on galactose and rhamnose. Good growth (on glucose) occurred between 20–30°C, at pH3-7 (optimum 25°C, pH4-5). Based on the results, we concluded that all five isolates originated from a single source in the grass silage sample. Subsequent experiments were therefore carried out with a single isolate now named DDNA#1.

**Illumina and MinION de novo genome assembly**

We took three approaches to assemble the genome of DDNA#1. The first approach used only short reads produced by the Illumina platform. After merging the paired end reads we obtained 1.08 Gbp of ~240 bp reads. The genome sequence that we obtained using the Spades assembler showed a very fragmented assembly that consisted of 14,764 contigs. The N50 of this assembly was only 124 kbp. In the second approach, we used the Spades assembler alone, either with 1D or 2D reads. Here we obtained 1.08 Gbp with an average read length of 240 bp. The genome sequence that we obtained using Spades assembler showed a fragmented assembly that consisted of 617 contigs with an average read length of 7.9 kbp. This hybrid assembly consisted of 1904 contigs with an average read length of 7.5 kbp, 389 Mbp was left after correction by Canu. This corrected dataset had an average read length of 7.9 kbp. This hybrid assembly consisted of 1904 contigs with an N50 of 255 kbp. As a third approach, we only used the long read data set and let the Canu assembler correct the longest reads with the shorter reads and then attempt an assembly. In this assembly
we obtained 61 contigs with a N50 of 455 kbp (Table 1). It is clear from these results that using the long read data set alone produced the most contiguous assembly, as has been shown previously.\(^8,9\)

We also used the nanopore datasets made with the R7.3 and R9 chemistry separately in the Canu assembler. The most notable difference between these assemblies is found in the mitochondrial genome. Only 16 kbp of this 33 kbp genome could be assembled with the R7.3 data, whereas the R9 assembly contained the complete mitochondrial genome (NCBI reference sequence, NC_022164.1). The mitochondrial genome has a very low GC content (21%) and in the extragenic regions more A and T homopolymers are found. Very few R7.3 reads mapped to this region, but in the R9 dataset there are many more reads that represent this region (Figure 1). It has been shown that the R7.3 data especially has a bias against A and T homopolymers. This bias is reduced in R9, but not completely absent.\(^8,9\) Even after correction of the long reads and assembly in Canu the contig sequences still contain errors.\(^11\) We have used PILON\(^12\) and the complementary Illumina data from this strain to correct the assembled contigs. This led to a minor increase in size of the assembly.

### Table 1. Canu assembly parameters and results.

<table>
<thead>
<tr>
<th>Property</th>
<th>R7.3 and R9 assembly</th>
<th>R9 assembly</th>
<th>R7.3 assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contigs</td>
<td>61</td>
<td>96</td>
<td>134</td>
</tr>
<tr>
<td>Assembly length</td>
<td>13027450</td>
<td>12823090</td>
<td>14433920</td>
</tr>
<tr>
<td>N25</td>
<td>943924</td>
<td>910563</td>
<td>1341435</td>
</tr>
<tr>
<td>N50</td>
<td>445592</td>
<td>421627</td>
<td>676845</td>
</tr>
<tr>
<td>N75</td>
<td>152971</td>
<td>252826</td>
<td>168187</td>
</tr>
<tr>
<td>Max length</td>
<td>1259066</td>
<td>1114421</td>
<td>2458927</td>
</tr>
<tr>
<td>Mean length</td>
<td>213565</td>
<td>133574</td>
<td>107716</td>
</tr>
<tr>
<td>Min length</td>
<td>23844</td>
<td>10334</td>
<td>1370</td>
</tr>
<tr>
<td>Est. genome size (Mbp)</td>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
</tr>
<tr>
<td>Error rate setting in Canu</td>
<td>0.025</td>
<td>0.025</td>
<td>0.035</td>
</tr>
<tr>
<td>Data used in Canu</td>
<td>R7.3 2D pass, R9 1D pass</td>
<td>R9 2D pass</td>
<td>R7.3 2D pass</td>
</tr>
</tbody>
</table>

AT rich mitochondrial DNA (21% GC, 79% AT)
The mtDNA assembly was incomplete when only R7.3 data was used.

---

Figure 1. Coverage plot of the *Candida vartiovaarae* DDNA#1 mitochondrial genome. Reads from both the Illumina, and the nanopore platform were aligned to the *Candida vartiovaarae* mitochondrial genome (Genbank accession number, KC993190.1) to show the difference in coverage between the different platforms and chemistry versions.
Figure 2. Genome size estimation generated by Genomescope, providing a k-mer analysis ($k = 19$, from Jellyfish) to estimate haploid genome size, fraction of heterozygosity and coverage. Genomescope attempts to find k-mer count peaks, low and high coverage peaks indicating hetero- and homozygosity. (A) We find ~13× and ~28× coverage for hetero- and homozygous fractions in our dataset. Exact peak positions are determined with a log transformation. Evaluating the slope between coverage points reveals the peak positions indicating hetero-and homozygosity, for lower and higher coverage, respectively. (B) Table showing the most important metrics from this k-mer analysis.

Pulsed field gel electrophoresis
As a further means to validate our assembled contigs and determine if they match the actual chromosome length, we have separated the chromosomes on an agarose gel using pulsed field gel electrophoresis. The gel image in Figure 3 shows five bands that represent the chromosomes of this yeast strain. The smallest band has a length that corresponds to the length of the mitochondrial genome (33 kbp). Additional fragments of 450, 1200, and 1500 kbp are also found. The intensity of the band that runs above the 2200 kbp marker band suggests that it actually contains more than one distinct fragment. To make the genome size fit to the estimate derived from the assembly and k-mer analysis (~12.5 mbp), three ~3 Mbp chromosomes should be postulated. The uncertainty in chromosome size estimate based on pulsed field electrophoresis gels is high because of the large chromosome size and the fact that it is difficult to determine if more than one fragment is...
present in the gel at a given position. Our conclusion that the top band represents three or more chromosomes is in agreement with the genome sequences of two related *C. jadinii* strains, namely CBS1600 and NBRC 0988.

**Genome comparison**

We have compared the assembled contigs of our *C. vartiovaarae* isolate DDNA#1 strain to yeast genome sequences that are already deposited in the nucleotide database. Comparison of our yeast strain with the well characterized *S. cerevisiae* assembly showed negligible genomic similarity (Figure 4A). From 26S ribosomal RNA sequences available in the nucleotide database, Chen *et al.* have constructed a phylogenetic tree. The closest relatives for which whole genome sequences are available are *C. jadinii* strains CBS1600 and NBRC 0988. An initial comparison between CBS1600 and NBRC 0988 revealed that these two strains show high homology (Figure 4B). The genomic similarity between our strain and *C. jadinii* strains CBS1600 and NBRC 0988 is much lower (Figures 4C and D). In conclusion, these data show that wild type yeast strains are very heterogeneous, despite a high similarity based on ribosomal RNA ITS sequences. Therefore, the data suggest that nanopore sequencing is an essential new tool to classify yeast strains. Of course, the nanopore sequence data in combination with other sequencing technologies is highly useful for accurate annotation of all genes in the genome.

![Figure 3. Pulsed field gel electrophoresis of Candida varitiovaarae DDNA#1 chromosomes.](image)

In lane 1, the chromosomes of *Saccharomyces cerevisiae* were loaded as a marker. Sizes of the chromosomes in the marker lane are indicated. In lane 2, the chromosomes of *Candida varitiovaarae* DDNA#1 were loaded.
Figure 4. Full genome comparisons between different yeast species. Dashed lines indicate contigs (start and stop positions) and the area between dashed lines indicates the contig size. Blue and orange dots are hits in reverse and forward orientation, respectively. Diagonal lines indicate sequence and synteny conservation across species: (A) Comparison between *Saccharomyces cerevisiae* S288c (horizontal axis) and *Candida variovaarae* isolate DDNA#1 (vertical axis). (B) Comparison between *Cyberlindnera jadinii* strains CBS1600 (horizontal axis) and NBRC 0988 (vertical axis). (C) Comparison between *Candida variovaarae* isolate DDNA#1 (vertical axis) and *Cyberlindnera jadinii* strain CBS1600 (horizontal axis). (D) Comparison between *Candida variovaarae* isolate DDNA#1 (vertical axis) and *Cyberlindnera jadinii* strain NBRC 0988 (horizontal axis).
Author contributions
HPS conceived the study, PJP, HPS, HJJ, and RPD designed the experiments. HJJ, RJLFL, PvH, TO, and SS performed the experiments. HJJ, ML, and CVH contributed to the data analysis. HJJ, RPD, and HPS prepared the first draft of the manuscript. All authors were involved in the revision of the draft manuscript and have agreed to the final content.

Competing interests
HIJJ and CVH are members of the Nanopore Community, and have previously received flow cells free of charge, as well as travel expense reimbursements from Oxford Nanopore Technologies.

Grant information
The author(s) declared that no grants were involved in supporting this work.

References
Open Peer Review

Current Referee Status: ?? ?? ?? ?

Version 1

Referee Report 27 July 2017

doi:10.5256/f1000research.12025.r23807

Hayan Lee
Department of Genetics, School of Medicine, Stanford University, California, CA, USA

Jansen et al. used Oxford Nanopore Technology with other short read sequencing technology, HiSeq 2500, to perform high-quality de novo genome assembly and classify yeast strain isolates, Candida variovaarae DDNA#1 from Saccharomyces cerevisiae S288C and Cyberlindrena jadinii CBS1600/NBRC 0988. They also exploited two versions of Nanopore flowcell chemistry and related software. Especially AT-rich mitochondria assembly using R7.3 and R9 comparison is very interesting.

Using similar short read data, N50 of DDNA#1 is 2.2kbp and that of S277C was 124Kbp. Probably authors want to perform repeat analysis for both strains to further study what makes such a performance gap.

For assembly approach two and three, authors used Canu to correct Nanopore reads with short reads. So basically all three approaches adopted short reads for correction or assembly purpose. Since Canu can perform self-correction with only long reads, it would be very interesting to compare self-corrected Nanopore reads assembly contiguity vs. short reads corrected Nanopore reads assembly contiguity.

Authors used two error correction methods; Canu and PILON, It would be helpful to consistently compare the correction performance of two software.

Although C. jadinii stains are proposed to be the closest strain, given Figure 4, S288C looks much closer to DDNA#1. Probably authors want to take a close look at this.

All sequencing data should be available online for reproducibility.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes
If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions drawn adequately supported by the results?
Partly

**Competing Interests:** No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

---

**Author Response 04 Jul 2018**

**Michael Liem, Leiden University, The Netherlands**

Using similar short read data, N50 of DDNA#1 is 2.2kbp and that of S277C was 124Kbp. Probably authors want to perform repeat analysis for both strains to further study what makes such a performance gap.

- **We hope to have shown that the performance gap can be overcome using long reads.** Using long reads in either hybrid or with data exclusively from nanopore decreased fragmentation and increased contiguity. Suggesting genomic complexity caused initial difficulties during assembly of this strain.

For assembly approach two and three, authors used Canu to correct Nanopore reads with short reads. So basically all three approaches adopted short reads for correction or assembly purpose. Since Canu can perform self-correction with only long reads, it would be very interesting to compare self-corrected Nanopore reads assembly contiguity vs. short reads corrected Nanopore reads assembly contiguity.

- **We have compared Canu (self-corrected) results to assemblies made with Miniasm, TULIP and Smartdenovo corrected with Racon.** It appears the assembly strategy is a crucial difference to contiguity and fragmentation as opposed to self- or post-assembly correction.

Authors used two error correction methods; Canu and PILON, It would be helpful to consistently compare the correction performance of two software.

- **The comparison between different assemblers and correction procedures should be more consistent now that we have separated the two task more prominently.**

Although *C. jadinii* stains are proposed to be the closest strain, given Figure 4, S288C looks much closer to DDNA#1. Probably authors want to take a close look at this.

- **Although alignment hits between *C. jadinii* and S288C are more targeted towards the diagonal in this figure the alignment length is very short and the number of alignment hits is significantly lower compared to the other two strains.** This underlines the poor synteny conservation between *C. jadinii* and S288C as compared to CBS1600 and NBRC 0988. *C. jadinii* compared to these two strains show many more alignment hits hence these strains are taken to be more similar.

All sequencing data should be available online for reproducibility.

- **Data has status in process, should be publicly accessible very soon**
This report by Jansen et al. describes comparison of de novo assemblies generated using Illumina or Oxford Nanopore sequence for the yeast Candida varitavaarae. The sequenced isolate was collected from a screen for new ethanologenic yeast species. Genomic DNA was sequenced using both platforms and de novo assemblies compared for overall metrics and representation of the mitochondrial genome. The final assembly was compared to those of other related yeast species to view conservation of synteny.

Overall this is an interesting study in showing the advantage of utilizing long Oxford nanopore reads for assembly of a genome that was difficult to assemble using Illumina data. This description would be more compelling if the authors could address a few issues with the presentation of this data.

1. In addition to genome size, the major factors that can influence the outcome of a de novo assembly are the repetitive sequence content, GC content, and level of heterozygosity. The authors suggest that repetitive sequence could explain large number of contigs; this could be directly addressed by identifying repetitive sequences in the assembly and evaluating contig ends. However there is also the suggestion in the text of some level of heterozygosity, which could better account for the low contig N50 they report in the Illumina assemblies. Whether or not the species is diploid and if so the level of heterozygosity is important to address in evaluating the performance of the two sequencing approaches and documenting the genomes for which long reads are most useful. This could be addressed for example using the Illumina data to identify heterozygous variants across the assembly.

2. The authors use Pilon to correct the assembled contigs with Illumina data and note that this led to a minor increase in size of the assembly, suggesting there were some misassembled regions in the original Canu assembly. As the other genomes compared using Nucmer are distantly related, with many rearrangements, this could not be used to validate the Canu assembly. It would be helpful if the authors could more fully describe the errors identified and fixed by Pilon.

3. Along the same lines, which statistics are for the final, best version of the assembly? Table 1 compares different combinations of Oxford Chemistry, however the authors also describe an additional step of Pilon polishing. It would be useful to contrast metrics, including sequence coverage levels and GC content, to those from the 2 Spades assemblies, as well as note which assembly is the final version.

4. In Figure 1, the top scale is too small to read. Plotting the GC as a separate track would be helpful to compare to the R7 coverage level.

5. For the PFG in Figure 3, a longer run may help separate the bright high MW band into separate chromosomes.
6. The data does not appear to be submitted to a public repository; both the raw sequence and the final best assembly should be submitted to NCBI or the ENA.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions drawn adequately supported by the results?
Partly

**Competing Interests:** No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

---

**Author Response 04 Jul 2018**

Michael Liem, Leiden University, The Netherlands

1. In addition to genome size, the major factors that can influence the outcome of a *de novo* assembly are the repetitive sequence content, GC content, and level of heterozygosity. The authors suggest that repetitive sequence could explain large number of contigs; this could be directly addressed by identifying repetitive sequences in the assembly and evaluating contig ends. However there is also the suggestion in the text of some level of heterozygosity, which could better account for the low contig N50 they report in the Illumina assemblies. Whether or not the species is diploid and if so the level of heterozygosity is important to address in evaluating the performance of the two sequencing approaches and documenting the genomes for which long reads are most useful. This could be addressed for example using the Illumina data to identify heterozygous variants across the assembly.

   - **The estimated genome size comparison between nanopore mediated assemblies and hybrid Spades assembly is a first indication of the polyploid genome of our strain.** Together with the abundant double gene copy BUSCO gene identification analysis and Spades hybrid contigs alignment to TULIP contigs we hope to have shown the diploid characteristics of DDNA#1, at least to partial extend.

2. The authors use Pilon to correct the assembled contigs with Illumina data and note that this led
to a minor increase in size of the assembly, suggesting there were some misassembled regions in the original Canu assembly. As the other genomes compared using Nucmer are distantly related, with many rearrangements, this could not be used to validate the Canu assembly. It would be helpful if the authors could more fully describe the errors identified and fixed by Pilon.

- **Increased assembly length after PILON correction is mainly due to corrected homopolymer stretches that are often underrepresented due to sequencing complexities of low complexity regions. This explanation has been added to the manuscript under results and discussion – Illumina and MinION *de novo* genome assembly.**

3. Along the same lines, which statistics are for the final, best version of the assembly? Table 1 compares different combinations of Oxford Chemistry, however the authors also describe an additional step of Pilon polishing. It would be useful to contrast metrics, including sequence coverage levels and GC content, to those from the 2 Spades assemblies, as well as note which assembly is the final version.

- **The final assembly is now described under results and discussion – Illumina and MinION *de novo* genome assembly. We have added sequence data statistics such as coverage and total amount of data. And aimed to highlight the error correction effect using BUSCO gene identification analysis.**

4. In Figure 1, the top scale is too small to read. Plotting the GC as a separate track would be helpful to compare to the R7 coverage level.

- **GC-content is now added to this figure and numbers and text have been made more clear.**

5. For the PFG in Figure 3, a longer run may help separate the bright high MW band into separate chromosomes.

- **We have tried many different run conditions and failed to properly resolve the largest bands. This may be different on a different system but we do not have access to such a system.**

6. The data does not appear to be submitted to a public repository; both the raw sequence and the final best assembly should be submitted to NCBI or the ENA.

- **Data has status in process, should be publicly accessible very soon**

*Competing Interests:* No competing interests were disclosed.
assemble a yeast strain using different methodologies: a short read strategy with Illumina reads alone and two hybrid approaches, the first one combining both short and long reads for the assembly and the second using long reads for the assembly and short reads for the correction of the consensus. In general, we think that this is a well put together study that reflects the current standard approaches for assembling genomes with both short and long reads. However, we have some questions/remarks that we would like the authors to answer.

1. It seems that the high level of polymorphism complicate the de novo assembly. If some regions are heterozygous, it should lead to a higher than expected assembly size. We think the authors should describe in more details the Illumina-only assembly especially the cumulative size (add a column in Table 1). As the error rate is low, with a high level of SNPs, both (Is the DDNA#1 isolate is a diploid yeast?) haplotypes should be segregated. On the contrary, the assembly length of the nanopore-only assemblies seems to be near the expected size (12Mb), does it mean that the error rate prevent to distinguish haplotypes? We think the authors should discuss in more details how haplotypes are resolved in their different assemblies.

2. The whole dataset (reads + final assembly) should be submitted in public repository to ensure full reproducibility.

3. Paragraph Illumina and MinION de novo genome assembly, line 38. Contigs were polished using the Pilon tool but line 7 of the same paragraph, authors indicate that the Spades assembly that was generated from Illumina reads alone was highly fragmented possibly due to a high level of SNPs in the DDNA#1 isolate. I think that to verify if the Pilon correction didn’t do more harm than good, authors could run the Busco tool (http://busco.ezlab.org/) on the assemblies, or annotate genes, before and after correction to verify if it didn’t introduce errors in the consensus due to heterogeneous input reads.

4. Paragraph Illumina and MinION de novo genome assembly, lines 14-15 it is said that the cumulative size of reads that was given as input to Canu was 2.05 Gb and that the corrected reads cumulative output size was equal to 389 Mb. I think that by default Canu only corrects 30X of the input read set (controlled by the corOutCoverage parameter) and since it is relatively close to 30-fold coverage of a yeast genome, I was wondering if authors leaved this parameter as default or if they moved up the limit and it could only correct around 30X of coverage. If this parameter was changed, I think it would be a good idea to indicate it.

5. Authors should add a table that contains standard metrics about the sequencing data (nanopore and illumina): number of reads, cumulative size, coverage, average read length...

6. Paragraph Full genome comparison, lines 12-15 it is said that the Nucmer’s ouput was filtered with the delta-filter software; please add the parameters used to filter out alignments. Moreover, if the yeast genomes used for the comparison are highly variable the nucmer software is not the best suited; maybe lastz (https://github.com/lastz/lastz) should better perform.

7. The smartdenovo assembler has been successfully applied to yeast genomes (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5466710/), it would be interesting to compare their results with a smartdenovo assembly.

Is the work clearly and accurately presented and does it cite the current literature?

Yes
Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions drawn adequately supported by the results?
Yes

**Competing Interests:** We declare that we have no competing interests; however we should mention that we are part of the MinION® Access Programme (MAP) and JMA received travel and accommodation expenses to speak at Oxford Nanopore Technologies conferences.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

**Author Response 04 Jul 2018**

Michael Liem, Leiden University, The Netherlands

1. It seems that the high level of polymorphism complicate the *de novo* assembly. If some regions are heterozygous, it should lead to a higher than expected assembly size. We think the authors should describe in more details the Illumina-only assembly especially the cumulative size (add a column in Table 1). As the error rate is low, with a high level of SNPs, both (Is the DDNA#1 isolate is a diploid yeast?) haplotypes should be segregated. On the contrary, the assembly length of the nanopore-only assemblies seems to be near the expected size (12Mb), does it mean that the error rate prevent to distinguish haplotypes? We think the authors should discuss in more details how haplotypes are resolved in their different assemblies.
   - Statistical information on the Illumina derived assembly is now added to Table 1. Indeed the majority of assemblies based exclusively on nanopore data are haploid genomes, this comes together with the notion that most of these assembler are designed to reconstruct bacterial genomes. However Canu should be able to differentiate diploid haplotypes, that is for high coverage datasets. It appears 17x coverage high quality long length read data is insufficient to resolve the (partial) diploid genome of DDNA#1.

2. The whole dataset (reads + final assembly) should be submitted in public repository to ensure full reproducibility.
   - These should be publicly available now

3. Paragraph *Illumina and MinION de novo genome assembly, line 38*. Contigs were polished
using the Pilon tool but line 7 of the same paragraph, authors indicate that the Spades assembly that was generated from Illumina reads alone was highly fragmented possibly due to a high level of SNPs in the DDNA#1 isolate. I think that to verify if the Pilon correction didn’t do more harm than good, authors could run the Busco tool (http://busco.ezlab.org/) on the assemblies, or annotate genes, before and after correction to verify if it didn’t introduce errors in the consensus due to heterogeneous input reads.

- **Thank you for your suggestion this is now incorporated into the manuscript under methods/results and discussion – genome assembly assessment based on gene expectation using BUSCO**

4. Paragraph **Illumina and MinION de novo genome assembly, lines 14-15** it is said that the cumulative size of reads that was given as input to Canu was 2.05 Gb and that the corrected reads cumulative output size was equal to 389 Mb. I think that by default Canu only corrects 30X of the input read set (controlled by the corOutCoverage parameter) and since it is relatively close to 30-fold coverage of a yeast genome, I was wondering if authors leaved this parameter as default or if they moved up the limit and it could only correct around 30X of coverage. If this parameter was changed, I think it would be a good idea to indicate it.

- **In our Canu version corOutCoverage is set to 40x coverage by default and has not been changed.**

5. Authors should add a table that contains standard metrics about the sequencing data (nanopore and illumina): number of reads, cumulative size, coverage, average read length…

- **Table is now added.**

6. Paragraph **Full genome comparison, lines 12-15** it is said that the Nucmer’s ouput was filtered with the delta-filter software; please add the parameters used to filter out alignments. Moreover, if the yeast genomes used for the comparison are highly variable the nucmer software is not the best suited; maybe lastz (https://github.com/lastz/lastz) should better perform.

- **Thank you for your suggestion, we have performed similar whole genome alignments with Lastz and mummer, however we didn’t observe a noticeable difference based on whole genome comparison alone. It appears the assembly algorithm and input data characteristics are the major factors that influenced the contiguity and fragmentation of our assemblies.**

7. The smartdenovo assembler has been successfully applied to yeast genomes (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5466710/), it would be interesting to compare their results with a smartdenovo assembly.

- **Thank you for your suggestion, Smartdenovo has now been added to the set of assemblers and results are denoted in our manuscript. Indeed Smartdenovo is an assembler that performs relatively well on the dataset of our yeast strain.**

**Competing Interests:** No competing interests were disclosed.
Mile Šikić
Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia

The authors presented de novo whole-genome assembly of a wild type yeast isolate using nanopore sequencing. They tried three different approaches to assemble the genome: using Illumina reads only, using both Illumina and nanopore reads in a hybrid approach, and using the only nanopore reads for assembling and Illumina reads for polishing. The third approach resulted in the most contiguous assembly. In they work they use nanopore datasets made with R7.3, R9 and R9.4 chemistries.

Although they used a correct procedure for genome assembly it would be interesting to compare their results with the following methods in the third approach:

- Using minimap+ miniasm assembler in combination with Racon consensus tool and PILON
- Using Canu + racon + PILON
- Try to polish nanopore assembly using Nanopolish

In addition, it would be valuable if they make their data publicly available to enable others to reproduce their results.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 04 Jul 2018
Michael Liem, Leiden University, The Netherlands

Using minimap+ miniasm assembler in combination with Racon consensus tool and PILON
• Thank you for your suggestion, this strategy is now included in our study.

Using Canu + racon + PILON
• Since Canu contains an integrated self-correction procedure prior to assembly we have not corrected the Canu contigs with Racon, however the combination Canu – PILON correction is part of our study, thank you.

Try to polish nanopore assembly using Nanopolish
• Thank you for your suggestion, however, since we have combined different data sets from different chemistries and different laboratories, at different times, including filtering of these data, it’s relatively complicated to polish such datasets with Nanopolish. To balance out the effort-result ratio we have performed a double iteration PILON correction which shows to be sufficient to identify the majority of genes stored in the Fungi 0db9 database used by BUSCO.

In addition, it would be valuable if they make their data publicly available to enable others to reproduce their results.
• Data has status in process, should be publicly accessible very soon

Competing Interests: No competing interests were disclosed.