

1 **Nanopore sequencing enables near-complete *de novo* assembly of**
2 ***Saccharomyces cerevisiae* reference strain CEN.PK113-7D**

3

4 Alex N. Salazar^{#,1,3}, Arthur R. Gorter de Vries^{#,2}, Marcel van den Broek², Melanie
5 Wijsman², Pilar de la Torre Cortés², Anja Brickwedde², Nick Brouwers², Jean-Marc
6 G. Daran² and Thomas Abeel^{*,1,3}

7 # These authors contributed equally to this publication and should be
8 considered co-first authors.

9 * Corresponding author

- 10 1. Delft Bioinformatics Lab, Delft University of Technology, Delft, The
11 Netherlands
- 12 2. Department of Biotechnology, Delft University of Technology, Delft, The
13 Netherlands
- 14 3. Broad Institute of MIT and Harvard, Boston, Massachusetts, USA

15

16 Alex N. Salazar	A.N.Salazar@tudelft.nl
17 Arthur R. Gorter de Vries	A.R.GorterdeVries@tudelft.nl
18 Marcel van den Broek	Marcel.vandenBroek@tudelft.nl
19 Melanie Wijsman	M.Wijsman@tudelft.nl
20 Pilar de la Torre Cortés	P.DeLaTorre@tudelft.nl
21 Anja Brickwedde	A.Brickwedde@tudelft.nl
22 Nick Brouwers	N.Brouwers-1@tudelft.nl
23 Jean-Marc G. Daran	J.G.Daran@tudelft.nl
24 Thomas Abeel	T.Abeel@tudelft.nl

25 Manuscript for publication in FEMS Yeast Research

26 Abstract

27 The haploid *Saccharomyces cerevisiae* strain CEN.PK113-7D is a popular model system for metabolic
28 engineering and systems biology research. Current genome assemblies are based on short-read
29 sequencing data scaffolded based on homology to strain S288C. However, these assemblies contain
30 large sequence gaps, particularly in subtelomeric regions, and the assumption of perfect homology to
31 S288C for scaffolding introduces bias.

32 In this study, we obtained a near-complete genome assembly of CEN.PK113-7D using only Oxford
33 Nanopore Technology's MinION sequencing platform. 15 of the 16 chromosomes, the mitochondrial
34 genome, and the 2-micron plasmid are assembled in single contigs and all but one chromosome starts or
35 ends in a telomere cap. This improved genome assembly contains 770 Kbp of added sequence
36 containing 248 gene annotations in comparison to the previous assembly of CEN.PK113-7D. Many of
37 these genes encode functions determining fitness in specific growth conditions and are therefore highly
38 relevant for various industrial applications. Furthermore, we discovered a translocation between
39 chromosomes III and VIII which caused misidentification of a *MAL* locus in the previous CEN.PK113-7D
40 assembly. This study demonstrates the power of long-read sequencing by providing a high-quality
41 reference assembly and annotation of CEN.PK113-7D and places a caveat on assumed genome stability
42 of microorganisms.

43

44

45

46 Keywords

47 *Saccharomyces cerevisiae*—Yeast—genome assembly—long read sequencing—Nanopore sequencing

48

49 Introduction

50 Whole Genome Sequencing (WGS) reveals important genetic information of an organism which can be
51 linked to specific phenotypes and enable genetic engineering approaches (Mardis 2008, Ng and Kirkness
52 2010). Short-read sequencing has become the standard method for WGS in the past years due to its low
53 cost, high sequencing accuracy and high output of sequence reads. In most cases, the obtained read
54 data is used to reassemble the sequenced genome either by *de novo* assembly or by mapping the reads
55 to a previously-assembled closely-related genome. However, the sequence reads obtained are relatively
56 short: between 35 and 1000 bp (van Dijk *et al.* 2014). This poses challenges as genomes have long
57 stretches of repetitive sequences of several thousand nucleotides in length and can only be
58 characterized if a read spans the repetitive region and has a unique fit to the flanking ends (Matheson *et*
59 *al.* 2017). As a result, *de novo* genome assembly based on short-read technologies “break” at repetitive
60 regions preventing reconstruction of whole chromosomes. The resulting assembly consists of dozens to
61 hundreds of sequence fragments, commonly referred to as *contigs*. These contigs are then either
62 analysed independently or ordered and joined together adjacently based on their alignment to a closely-
63 related reference genome. However, referenced based joining of contigs into so-called *scaffolds*, is
64 based on the assumption that the genetic structure of the sequenced strain is identical to that of the
65 reference genome—potentially concealing existing genetic variation.

66 Previous genome assemblies of the *Saccharomyces cerevisiae* strain CEN.PK113-7D have been based on
67 homology with the fully-assembled reference genome of *S. cerevisiae* strain S288C (Cherry *et al.* 2012,
68 Nijkamp *et al.* 2012). CEN.PK113-7D is a haploid strain used as a model organism in biotechnology-
69 related research and systems biology because of its convenient growth characteristics, its robustness
70 under industrially-relevant conditions, and its excellent genetic accessibility (Canelas *et al.* 2010,
71 González-Ramos *et al.* 2016, Nijkamp *et al.* 2012, Papapetridis *et al.* 2017). CEN.PK113-7D was

72 sequenced using a combination of 454 and Illumina short-read libraries and a draft genome was
73 assembled consisting of over 700 contigs (Nijkamp *et al.* 2012). After scaffolding using MAIA (Nijkamp *et*
74 *al.* 2010) and linking based on homology with the genome of S288C, it was possible to reconstruct all 16
75 chromosomes. However, there were large sequence gaps within chromosomes and the subtelomeric
76 regions were left unassembled, both of which could contain relevant open reading frames (ORFs)
77 (Nijkamp *et al.* 2012). Assuming homology to S288C, more than 90% of missing sequence was located in
78 repetitive regions corresponding mostly to subtelomeric regions and Ty-elements. These regions are
79 genetically unstable as repeated sequences promote recombination events (Pryde *et al.* 1995);
80 therefore the assumption of homology with S288C could be unjustified. Ty-elements are present across
81 the genome: repetitive sequences with varying length (on average ~6 Kbp) resulting from introgressions
82 of viral DNA (Kim *et al.* 1998). Subtelomeric regions are segments towards the end of chromosomes
83 consisting of highly repetitive elements making them notoriously challenging to reconstruct using only
84 short-read sequencing data (Bergström *et al.* 2014). While Ty-elements are likely to have limited impact
85 on gene expression, subtelomeric regions harbour various so-called subtelomeric genes. Several gene
86 families are present mostly in subtelomeric regions and typically have functions determining the cell's
87 interaction with its environment; such as nutrient uptake (Carlson *et al.* 1985, Naumov *et al.* 1995),
88 sugar utilisation (Teste *et al.* 2010), and inhibitor tolerance (Denayrolles *et al.* 1997). Many of these
89 subtelomeric gene families therefore contribute to the adaptation of industrial strains to the specific
90 environment they are used in. For example, the *RTM* and *SUC* gene families are relevant for bioethanol
91 production as they increase inhibitor-tolerance in molasses and utilization of extracellular sucrose,
92 respectively (Carlson *et al.* 1985, Denayrolles *et al.* 1997). Similarly, *MAL* genes enable utilization of
93 maltose and maltotriose and *FLO* genes enable calcium-dependent flocculation, both of which are
94 crucial for the beer brewing industry (Brown *et al.* 2010, Lodolo *et al.* 2008, Teunissen and Steensma

95 1995). As is the case for Ty-elements, subtelomeric regions are unstable due to repetitive sequences and
96 homology to various regions of the genome, which is likely to cause diversity across strains (Brown *et al.*
97 2010, Nijkamp *et al.* 2012, Pryde *et al.* 1995). Characterizing and accurately localizing subtelomeric gene
98 families is thus crucial for associating strain performance to specific genomic features and for targeted
99 engineering approaches for strain improvement (Bergström *et al.* 2014).

100 In contrast to short-read technologies, single-molecule sequencing technologies can output sequence
101 reads of several thousand nucleotides in length. Recent developments of long-read sequencing
102 technologies have decreased the cost and increased the accuracy and output, yielding near-complete
103 assemblies of diverse yeast strains (Giordano *et al.* 2017, McIlwain *et al.* 2016). For example, *de novo*
104 assembly of a biofuel production *S. cerevisiae* strain using PacBio reads produced a genome assembly
105 consisting of 25 chromosomal contigs scaffolded into 16 chromosomes. This assembly revealed 92 new
106 genes relative to S288C amongst which 28 previously uncharacterized and unnamed genes.
107 Interestingly, many of these genes had functions linked to stress tolerance and carbon metabolism
108 which are functions critical to the strains industrial application (McIlwain *et al.* 2016). In addition, rapid
109 technological advances in nanopore sequencing have matured as a competitive long-read sequencing
110 technology and the first yeast genomes assembled using nanopore reads are appearing (Giordano *et al.*
111 2017, Goodwin *et al.* 2015, Istace *et al.* 2017, Jansen *et al.* 2017, McIlwain *et al.* 2016). For example,
112 Istace *et al.* sequenced 21 wild *S. cerevisiae* isolates and their genome assemblies ranged between 18
113 and 105 contigs enabling the detection of 29 translocations and 4 inversions relative to the chromosome
114 structure of reference S288C. In addition, large variations were found in several difficult to sequence
115 subtelomeric genes such as *CUP1*, which was correlated to large differences in copper tolerance (Istace
116 *et al.* 2017). Nanopore sequencing has thus proven to be a potent technology for characterizing yeast.

117 In this study, we sequenced CEN.PK113-7D using Oxford Nanopore Technology's (ONT) MinION
118 sequencing platform. This nanopore *de novo* assembly was compared to the previous short-read
119 assembly of CEN.PK113-7D (Nijkamp *et al.* 2012) with particular attention for previously, poorly-
120 assembled subtelomeric regions and for structural variation potentially concealed due to the
121 assumption of homology to S288C.

122 [Materials and methods](#)

123 [Yeast strains](#)

124 The *Saccharomyces cerevisiae* strain "CEN.PK113-7D Frankfurt" (*MATa MAL2-8c*) was kindly provided by
125 Dr. P. Kötter in 2016 (Entian and Kötter 2007, Nijkamp *et al.* 2012). It was plated on solid YPD
126 (containing 10 g/l yeast extract, 20 g/l peptone and 20 g/l glucose) upon arrival and a single colony was
127 grown once until stationary phase in liquid YPD medium and 1 mL aliquots with 30% glycerol were
128 stored at -80°C since. The previously sequenced CEN.PK113-7D sample was renamed "CEN.PK113-7D
129 Delft" (Nijkamp *et al.* 2012). It was obtained from the same source in 2001 and 1 mL aliquots with 30%
130 glycerol were stored at -80°C with minimal propagation since (no more than three cultures on YPD as
131 described above).

132 [Yeast cultivation and genomic DNA extraction](#)

133 Yeast cultures were incubated in 500-ml shake-flasks containing 100 ml liquid YPD medium at 30°C on
134 an orbital shaker set at 200 rpm until the strains reached stationary phase with an OD₆₆₀ between 12
135 and 20. Genomic DNA of CEN.PK113-7D Delft and CEN.PK113-7D Frankfurt for whole genome
136 sequencing was isolated using the Qiagen 100/G kit (Qiagen, Hilden, Germany) according to the
137 manufacturer's instructions and quantified using a Qubit® Fluorometer 2.0 (ThermoFisher Scientific,
138 Waltham, MA).

139 Short-read Illumina sequencing

140 Genomic DNA of CEN.PK113-7D Frankfurt was sequenced on a HiSeq2500 sequencer (Illumina, San
141 Diego, CA) with 150 bp paired-end reads using PCR-free library preparation by Novogene Bioinformatics
142 Technology Co., Ltd (Yuen Long, Hong Kong). All Illumina sequencing data are available at NCBI
143 (<https://www.ncbi.nlm.nih.gov/>) under the bioproject accession number PRJNA393501.

144 MinION Sequencing

145 MinION genomic libraries were prepared using either nanopore Sequencing Kit SQK-MAP006 (2D-
146 ligation for R7.3 chemistry), SQK-RAD001 (Rapid library prep kit for R9 chemistries) or SQK-MAP007 (2D-
147 ligation for R9 chemistries) (Oxford Nanopore Technologies, Oxford, United Kingdom). Two separate
148 libraries of SQK-MAP006 and one library of SQK-RAD001 were used to sequence CEN.PK113-7D Delft.
149 Only one SQK-MAP007 library was used to sequence CEN.PK113-7D Frankfurt. With the exception of the
150 SQK-RAD001 library, all libraries used 2-3 µg of genomic DNA fragmented in a Covaris g-tube (Covaris)
151 with the “8-10 kbp fragments” settings according to manufacturer’s instructions. The SQK-RAD001
152 library used 200 ng of unsheared genomic DNA. Libraries for SQK-MAP006 and SQK-MAP007 were
153 constructed following manufacturer’s instructions with the exception of using 0.4x concentration of
154 AMPure XP Beads (Beckman Coulter Inc., Brea, CA) and 80% EtOH during the “End Repair/dA-tailing
155 module” step. The SQK-RAD001 library was constructed following manufacturer’s instructions. Prior to
156 sequencing, flow cell quality was assessed by running the MinKNOW platform QC (Oxford Nanopore
157 Technology). All flow cells were primed with priming buffer and the libraries were loaded following
158 manufacturer’s instructions. The mixture was then loaded into the flow cells for sequencing. The SQK-
159 MAP006 library of CEN.PK113-7D Delft was sequenced twice on a R7.3 chemistry flow cell (FLO-MIN103)
160 and the SQK-RAD001 library was sequenced on a R9 chemistry flow cell (FLO-MIN105)—all for 48 hours.
161 The SQK-MAP007 library for CEN.PK113-7D Frankfurt was sequenced for 48 hours on a R9 chemistry
162 flow cell (FLO-MIN104). Reads from all sequencing runs were uploaded and base-called using Metrichor

163 desktop agent (<https://metrichor.com/s/>). The error rate of nanopore reads in the CEN.PK113-7D
164 Frankfurt and Delft was determined by aligning them to the final CEN.PK113-7D assembly (see section
165 below) using Graphmap (Sović *et al.* 2016) and calculating mismatches based on the CIGAR strings of
166 reads with a mapping quality of at least 1 and no more than 500 nt of soft/hard clipping on each end of
167 the alignment to avoid erroneous read-alignments due to repetitive regions (i.e. paralogous genes,
168 genes with copy number variation). All nanopore sequencing data are available at NCBI under the
169 bioproject accession number PRJNA393501.

170 *De novo genome assembly*

171 FASTA and FASTQ files were extracted from base-called FAST5 files using Poretools (version 0.6.0)
172 (Loman and Quinlan 2014). Raw nanopore reads were filtered for lambda DNA by aligning to the
173 *Enterobacteria phage lambda* reference genome (RefSeq assembly accession: GCF_000840245.1) using
174 Graphmap (Sović *et al.* 2016) with *--no-end2end* parameter and retaining only unmapped reads using
175 Samtools (Li *et al.* 2009). All reads obtained from the Delft and the Frankfurt CEN.PK113-7D stock
176 cultures were assembled *de novo* using Canu (version 1.3) (Koren *et al.* 2017) with *--genomesize* set to
177 12 Mbp. The assemblies were aligned using the MUMmer tool package: Nucmer with the *--maxmatch*
178 parameter and filtered for the best one-to-one alignment using Delta-filter (Kurtz *et al.* 2004). The
179 genome assemblies were visualized using Mummerplot (Kurtz *et al.* 2004) with the *--fat* parameter.
180 Gene annotations were performed using MAKER2 annotation pipeline (version 2.31.9) using SNAP
181 (version 2013-11-29) and Augustus (version 3.2.3) as *ab initio* gene predictors (Holt and Yandell 2011).
182 S288C EST and protein sequences were obtained from SGD (*Saccharomyces* Genome Database,
183 <http://www.yeastgenome.org/>) and were aligned using BLASTX (BLAST version 2.2.28+) (Camacho *et al.*
184 2009). Translated protein sequence of the final gene model were aligned using BLASTP to S288C protein
185 Swiss-Prot database. Custom made Perl scripts were used to map systematic names to the annotated

186 gene names. Telomere cap sequences (TEL07R of size 7,306 bp and TEL07L of size 781 bp) from the
187 manually-curated and complete reference genome for *S. cerevisiae* S288C (version R64, Genbank ID:
188 285798) obtained from SGD were aligned to the assembly as a proxy to assess completeness of each
189 assembled chromosome. SGIDs for TEL07R and TEL07L are S000028960 and S000028887, respectively.
190 The Tablet genome browser (Milne *et al.* 2012) was used to visualize nanopore reads aligned to the
191 nanopore *de novo* assemblies. Short assembly errors in the Frankfurt assembly were corrected with
192 Nanopolish (version 0.5.0) using default parameters (Loman *et al.* 2015). Two contigs, corresponding to
193 chromosome XII, were manually scaffolded based on homology to S288C. To obtain the 2-micron native
194 plasmid in CEN.PK113-7D, we aligned S288C's native plasmid to the "unassembled" contigs file provided
195 by Canu (Koren *et al.* 2017) and obtained the best aligned contig in terms of size and sequence
196 similarity. Duplicated regions due to assembly difficulties in closing circular genomes were identified
197 with Nucmer and manually corrected. BWA (Li and Durbin 2010) was used to align Illumina reads to the
198 scaffolded Frankfurt assembly using default parameters. Pilon (Walker *et al.* 2014) was then used to
199 further correct assembly errors by aligning Illumina reads to the scaffolded Frankfurt assembly using
200 correction of only SNPs and short indels (*--fix bases* parameter) using only reads with a minimum
201 mapping quality of 20 (*--minmq 20* parameter). Polishing with structural variant correction in addition to
202 SNP and short indel correction was benchmarked, but not applied to the final assembly (Additional File
203 1).

204 [Analysis of added information in the CEN.PK113-7D nanopore assembly](#)

205 Gained and lost sequence information in the nanopore assembly of CEN.PK113-7D was determined by
206 comparing it to the previous short-read assembly (Nijkamp *et al.* 2012). Contigs of at least 1 Kbp of
207 short-read assembly were aligned to the nanopore CEN.PK113-7D Frankfurt assembly using the
208 MUMmer tool package (Kurtz *et al.* 2004) using *show-coords* to extract alignment coordinates. For

209 multi-mapped contigs, overlapping alignments of the same contig were collapsed and the largest
210 alignment length as determined by Nucmer was used. Unaligned coordinates in the nanopore assembly
211 were extracted and considered as added sequence. Added genes were retrieved by extracting the gene
212 annotations in these unaligned regions from the annotated nanopore genome; mitochondria and 2-
213 micron plasmid genes were excluded. For the lost sequence, unaligned sequences were obtained by
214 aligning the contigs of the nanopore assembly to the short-read contigs of at least 1 kb using the same
215 procedure as described above. Lost genes were retrieved by aligning the unaligned sequences to the
216 short-read CEN.PK113-7D assembly with BLASTN (version 2.2.31+) (Camacho *et al.* 2009) and retrieving
217 gene annotations. BLASTN was used to align DNA sequences of YHRCTy1-1, YDRCTy2-1, YILWTy3-1,
218 YHLWTy4-1, and YCLWTy5-1 (obtained from the *Saccharomyces Genome Database*; SGIDs: S000007006,
219 S000006862, S000007020, S000006991, and S000006831, respectively) as proxies for the location of
220 two known groups of Ty-elements in *Saccharomyces cerevisiae*, *Metaviridae* and *Pseudoviridae* (Kim *et*
221 *al.* 1998), in the CEN.PK113-7D Frankfurt assembly. Non-redundant locations with at least a 2 Kbp
222 alignment and an E-value of 0.0 as determined by BLASTN were then manually inspected.

223 [Comparison of the CEN.PK113-7D assembly to the S288C genome](#)

224 The nanopore assembly of CEN.PK113-7D and the reference genome of S2888C (Accession number
225 GCA_000146045.2) were annotated using the MAKER2 pipeline described in the “De novo genome
226 assembly” section. For each genome a list of gene names per chromosome was constructed and
227 compared strictly on their names to identify genes names absent in the corresponding chromosome in
228 the other genome. The ORFs of genes identified as absent in either genome were aligned using BLASTN
229 (version 2.2.31+) to the total set of ORFs of the other genome and matches with an alignment length of
230 half the query and with a sequence identity of at least 95% were listed. If one of the unique genes
231 aligned to an ORF on the same chromosome, it was manually inspected to check if it was truly absent in

232 the other genome. Merged ORFs and misannotations were not considered in further analysis. These
233 alignments were also used to identify copies and homologues of the genes identified as truly absent in
234 the other genome.

235 Gene ontology analysis was performed using the Gene Ontology term finder of SGD using the list of
236 unique genes as the query set and all annotated genes as the background set of genes for each genome
237 (Additional file 2A and 2C). The ORFs of genes identified as present in S288C but absent in CEN.PK113-7D
238 in previously made lists (Daran-Lapujade *et al.* 2003, Nijkamp *et al.* 2012) were obtained from SGD. The
239 ORFs were aligned both ways to ORFs from SGD identified as unique to S288C in this study using
240 BLASTN. Genes with alignments of at least half the query length and with a sequence identity of at least
241 95% were interpreted as confirmed by the other data set. In order to analyze the origin of genes
242 identified as unique to S288C, these ORFs were aligned using BLASTN to 481 genome assemblies of
243 various *S. cerevisiae* strains obtained from NCBI (Additional file 3) and alignments of at least 50% of the
244 query were considered. The top alignments were selected based on the highest sequence ID and only
245 one alignment per strain was counted per gene.

246 Chromosome translocation analysis

247 Reads supporting the original and translocated genomic architectures of chromosomes III and VIII were
248 identified via read alignment of raw nanopore reads. First, the translocation breakpoints coordinates
249 were calculated based on whole-genome alignment of CEN.PK113-7D Delft assembly to S288C with
250 MUMmer. A modified version of S288C was created containing the normal architectures of all 16
251 chromosomes and the mitochondrial genome plus the translocated architecture of chromosomes III-VIII
252 and VIII-III. The first nearest unique flanking genes at each breakpoint were determined using BLASTN
253 (version 2.2.31+) (English *et al.* 2012, Zhang *et al.* 2000) in reference to both S288C and the Delft
254 CEN.PK113-7D nanopore assembly. Raw nanopore reads from CEN.PK113-7D Delft and Frankfurt were

255 aligned to the modified version of S288C and nanopore reads that spanned the translocation
256 breakpoints as well as the unique flanking sequences were extracted. Supporting reads were validated
257 by re-aligning them to the modified version of S288C using BLASTN.

258 Results

259 Sequencing on a single nanopore flow cell enables near-complete genome assembly

260 To obtain a complete chromosome level *de novo* assembly of *Saccharomyces cerevisiae* CENPK113-7D,
261 we performed long read sequencing on the Oxford Nanopore Technology's (ONT) MinION platform. A
262 fresh sample of CEN.PK113-7D was obtained from the original distributor Dr. P. Kötter (further referred
263 to as "CEN.PK113-7D Frankfurt"), cultured in a single batch on YPD medium and genomic DNA was
264 extracted. CEN.PK113-7D Frankfurt was sequenced on a single R9 (FLO-MIN104) chemistry flow cell
265 using the 2D ligation kit for the DNA libraries producing more than 49x coverage of the genome with an
266 average read-length distribution of 10.0 Kbp (Supplementary Figure S1) and an estimated error rate of
267 10% (Supplementary Figure S2). We used Canu (Koren *et al.* 2017) to produce high-quality *de novo*
268 assemblies using only nanopore data. Before correcting for misassemblies, the assembly contained a
269 total of 21 contigs with an N50 of 756 Kbp (Supplementary Table S1). This represented a 19-fold
270 reduction in the number of contigs and a 15-fold increase of the N50 in comparison to the short-read-
271 only assembly of the first CEN.PK113-7D draft genome version (Nijkamp *et al.* 2012) (Table 1).

272 Most chromosomes of the nanopore *de novo* assembly are single contigs and are flanked by telomere
273 caps. Genome completeness was determined by alignment to the manually-curated reference genome
274 of the strain S288C (version R64, Genbank ID: 285798) (Supplementary Table S2). The two largest yeast
275 chromosomes, IV and XII, were each split in two separate contigs, and two additional contigs (31 and 38
276 Kbp in length) corresponded to unplaced subtelomeric fragments. In particular, the assembly for
277 chromosome XII was interrupted in the *RDN1* locus—a repetitive region consisting of gene encoding

278 ribosomal RNA estimated to be more than 1-Mbp long (Venema and Tollervey 1999). Since no reads
279 were long enough to span this region, the contigs were joined with a gap.

280 Manual curation resolved chromosome III, chromosome IV and the mitochondrial genome.
281 Chromosome IV was fragmented into two contigs at locus of 11.5 Kbp containing two Ty-elements in
282 S288C (coordinates 981171-992642). Interestingly, the end of the first contig and the start of the second
283 contig have 8.8 Kbp of overlap (corresponding to the two Ty-elements) and one read spans the
284 repetitive Ty-elements and aligns to unique genes on the left and right flanks (*EXG2* and *DIN7*,
285 respectively). We therefore joined the contigs without missing sequence resulting in a complete
286 assembly of chromosome IV. For chromosome III, the last ~27 Kbp contained multiple telomeric caps
287 next to each other. The last ~10 Kbp had little to no coverage when re-aligning raw nanopore reads to
288 the assembly (Supplementary Figure S3). The coordinates for the first telomeric cap were identified and
289 the remaining sequence downstream was removed resulting in a final contig of size of 347 Kbp. The
290 original contig corresponding to the mitochondrial genome had a size of 104 Kbp and contained a nearly
291 identical ~20 Kbp overlap corresponding to start of the *S. cerevisiae* mitochondrial genome (i.e. origin of
292 replication) (Supplementary Figure S4). This is a common artifact as assembly algorithms generally have
293 difficulties reconstructing and closing circular genomes (McIlwain *et al.* 2016, Venema and Tollervey
294 1999). The coordinates of the overlaps were determined with Nucmer (Kurtz *et al.* 2004) and manually
295 joined resulting to a final size of 86,616 bp.

296 Overall, the final CEN.PK113-7D Frankfurt assembly contained 15 chromosome contigs, 1 chromosome
297 scaffold, the complete mitochondrial contig, the complete 2-micron plasmid and two unplaced
298 telomeric fragments, adding up to a total of 12.1 Mbp (Table 1 and Supplementary Table S3). Of the 16
299 chromosomes, 11 were assembled up until both telomeric caps, four were missing one of the telomere
300 caps and only chromosome X was missing both telomere caps. Based on homology with S288C, the

301 missing sequence was estimated not to exceed 12 kbp for each missing (sub)telomeric region.
302 Furthermore, we found a total of 46 retrotransposons Ty-elements: 44 were from the *Pseudoviridae*
303 group (30 *Ty1*, 12 *Ty2*, 1 *Ty4*, and 1 *Ty5*) and 2 from *Metaviridae* group (*Ty3*). The annotated nanopore
304 assembly of CEN.PK113-7D Frankfurt is available at NCBI under the bioproject accession number
305 PRJNA393501.

306 [Comparison of the nanopore and short-read assemblies of CEN.PK113-7D](#)

307 We compared the nanopore assembly of CEN.PK113-7D to a previously published version to quantify
308 the improvements over the current state-of-the art (Nijkamp *et al.* 2012). Alignment of the contigs of
309 the short-read assembly to the nanopore assembly revealed 770 Kbp of previously unassembled
310 sequence, including the previously unassembled mitochondrial genome (Additional file 4A). This gained
311 sequence is relatively spread out over the genome (Figures 1A and 1B) and contained as much as 284
312 chromosomal gene annotations (Additional file 4B). Interestingly, 69 out of 284 genes had paralogs,
313 corresponding to a fraction almost twice as high as the 13% found in the whole genome of S288C (Wolfe
314 and Shields 1997). Gene ontology analysis revealed an enrichment in the biological process of cell
315 aggregation ($P=9.30 \times 10^{-4}$); in the molecular functions of mannose binding ($P=3.90 \times 10^{-4}$) and glucosidase
316 activity ($P=7.49 \times 10^{-3}$); and in the cellular components of the cell wall ($P=3.41 \times 10^{-7}$) and the cell periphery
317 component ($P=5.81 \times 10^{-5}$). Some newly-assembled genes are involved in central carbon metabolism,
318 such as *PDC5*. In addition, many of the added genes are known to be relevant in industrial applications
319 including hexose transporters such as *HXT* genes and sugar polymer hydrolases such as *IMA* and *MALx2*
320 genes; several genes relevant for cellular metal homeostasis, such as *CUP1-2* (linked to copper ion
321 tolerance) and *FIT1* (linked to iron ion retention); genes relevant for nitrogen metabolism in medium
322 rich or poor in specific amino acids, including amino acid transporters such as *VBA5*, amino acid
323 catabolism genes such as *ASP3-4* and *LEU2* and amino-acid limitation response genes such as many *PAU*

324 genes; several *FLO* genes which are responsible for calcium-dependent flocculation; and various genes
325 linked to different environmental stress responses, such as *HSP* genes increasing heat shock tolerance
326 and *RIM101* increasing tolerance to high pH.

327 To evaluate whether some previously assembled sequence was missing in the nanopore assembly, we
328 aligned the nanopore contigs to the short-read assembly (Nijkamp *et al.* 2012). Less than 6 Kbp of
329 sequence of the short-read assembly was not present in the nanopore assembly, distributed over 13
330 contigs (Additional file 4C). Only two ORFs were missing: the genes *BIO1* and *BIO6* (Additional file 4D).
331 Alignment of *BIO1* and *BIO6* sequences to the nanopore assembly showed that the right-end of the
332 chromosome I contig contains the first ~500 nt of *BIO1*. While *BIO1* and *BIO6* were present in the
333 nanopore sequences, they are absent in the final assembly likely due to the lack of long-enough reads to
334 resolve the repetitive nature of this subtelomeric region.

335 Overall an additional 770 Kbp sequence containing 284 genes was gained, while 6 Kbp containing two
336 genes was not captured compared to the previous assembly. In addition, the reduction from over 700 to
337 only 20 contigs clearly shows that the nanopore assembly is much less fragmented than the short-read
338 assembly (Table 1).

339 [Comparison of the Nanopore assembly of CEN.PK113-7D to S288C](#)

340 To identify unique and shared genes between CEN.PK113-7D and S288C, we compared annotations
341 made using the same method for both genomes (Additional Files 2A and 2C). We identified a total of 45
342 genes unique to CEN.PK113-7D and 44 genes unique to S288C (Additional Files 2B and 2D). Genes
343 located in regions that had no assembled counterpart in the other genome were excluded; 20 for S288C
344 and 27 for CEN.PK113-7D. Interestingly, the genes unique to either strain and genes present on different
345 chromosomes were found mostly in the outer 10% of the chromosomes, indicating that the

346 subtelomeric regions harbor most of the genetic differences between CEN.PK113-7D and S288C (Figure
347 1C).

348 In order to validate the genes identified as unique to S288C, we compared them to genes identified as
349 absent in CEN.PK113-7D in previous studies (Additional file 2D, Table 2). 25 genes of S288C were
350 identified as absent in CEN.PK113-7D by array comparative genomic hybridization (aCGH) analysis
351 (Daran-Lapujade *et al.* 2003) and 21 genes were identified as absent in CEN.PK113-7D based on short-
352 read WGS (Nijkamp *et al.* 2012). Of these genes, 19 and 10 respectively were identified as genes in
353 S288C by our annotation pipeline and could be compared to the genes we identified as unique to S288C.
354 While 19 of these 29 genes were also absent in the nanopore assembly, the remaining 10 genes were
355 fully assembled and annotated, indicating they were erroneously identified as missing (Table 2).

356 In order to determine if the genes unique to S288C have homologues elsewhere in the genome of
357 CEN.PK113-7D or if they are truly unique, we aligned the ORFs of the 44 genes identified as unique in
358 S288C to the ORFs in the nanopore CEN.PK113-7D assembly. 26 genes were completely absent in the
359 CEN.PK113-7D assembly, while the remaining 18 genes aligned to between 1 and 20 ORFs each in the
360 genome of CEN.PK113-7D with more than 95% sequence identity, indicating they may have close
361 homologues or additional copies in S288C (Additional file 2D). Gene ontology analysis revealed no
362 enrichment in biological process, molecular functions or cell components of the 26 genes without
363 homologues in CEN.PK113-7D. Five genes without homologues were labelled as putative. However,
364 there were many genes encoding proteins relevant for fitness under specific industrial conditions, such
365 as *PHO5* which is part of the response to phosphate scarcity, *COS3* linked to salt tolerance, *ADH7* linked
366 to acetaldehyde tolerance, *RDS1* linked to resistance to cycloheximide, *PDR18* linked to ethanol
367 tolerance and *HXT17* which is involved in hexose sugar uptake (Additional file 2D). In addition, we

368 confirmed the complete absence of *ENA2* and *ENA5* in CEN.PK113-7D which are responsible for lithium
369 sensitivity of CEN.PK113-7D (Daran-Lapujade *et al.* 2009).

370 Conversely, to determine if the genes unique to CEN.PK113-7D have homologues elsewhere in the
371 genome of S288C or if they are truly unique, we aligned the ORFs of the 45 genes identified as unique in
372 CEN.PK113-7D to the ORFs of S288C. A set of 16 genes were completely absent in S288C, while the
373 remaining 29 aligned to between one and 16 ORFs each in the genome of S288C with more than 95%
374 sequence (Additional File 2D). Gene ontology analysis revealed no enrichment in biological processes,
375 molecular functions or cell components of the 16 genes unique to CEN.PK113-7D without homologues.
376 However, among the genes without homologues a total of 13 were labelled as putative. The presence of
377 an additional copy of *IMA1*, *MAL31* and *MAL32* on chromosome III was in line with the presence of the
378 *MAL2* locus which was absent in S288C. Interestingly the sequence of *MAL13*, which belongs to this
379 locus, was divergent enough from other *MAL*-gene activators to not be identified as homologue.
380 Additionally, when performing the same analysis on the 27 genes on the two unplaced contigs of the
381 CEN.PK113-7D assembly, 7 of them did not align to any gene of S288C with more than 95% sequence
382 identity, indicating these unplaced telomeric regions are highly unique to CEN.PK113-7D.

383 Since the genome of CEN.PK113-7D contains 45 ORFs which are absent in S288C, we investigated their
384 origin by aligning them against all available *S. cerevisiae* nucleotide data at NCBI (Additional File 3). For
385 each ORF, we report the strains to which they align with the highest sequence identity and the sequence
386 identity relative to S288C in Additional File 2B. For most genes, several strains aligned equally well with
387 the same sequence identity. For 13 ORFs S288C is among the best matches, indicating these ORFs may
388 come from duplications in the S288C genome. However, S288C is not among the best matches for 32
389 ORFs. In these, laboratory strain “SK1” is among the best matches 9 times, the west African wine isolate
390 “DBVPG6044” appears 8 times, laboratory strain “W303” appears 7 times, the Belgian beer strain

391 “beer080” appears 3 times and the Brazilian bioethanol strain “bioethanol005” appears 3 times.
392 Interestingly, some grouped unique genes are most related to specific strains. For example, the unique
393 genes identified on the left subtelomeric regions of chromosome XVI (YBL109W, YHR216W and YOR392)
394 and of chromosome VIII (YJL225C and YOL161W) exhibited the highest similarity to DBVPG6044.
395 Similarly, the right end of the subtelomeric region of chromosome III (YPL283W-A and YPR202) and of
396 chromosome XI ((YPL283W-A and YLR466W) were most closely related to W303.

397 Interestingly, the nanopore assembly revealed a duplication of *LEU2*, a gene involved in synthesis of
398 leucine which can be used as an auxotrophy marker. In the complete reference genome of *S. cerevisiae*
399 S288C, both *LEU2* and *NFS1* are unique, neighboring genes located chromosome III. However, gene
400 annotations of the assemblies and raw nanopore reads support additional copies of *LEU2* and *NFS1* in
401 CEN.PK113-7D located on chromosome VII (Figure 2). The additional copy contained the complete *LEU2*
402 sequence but only ~0.5 kb of the 5’ end of *NFS1*. In CEN.PK113-7D and S288C, the *LEU2* and *NFS1* loci in
403 chromosome III were located adjacent to Ty-elements. Two such Ty-elements were also found flanking
404 the additional *LEU2* and *NFS1* loci in chromosome VII (Figure 2). It is likely that the duplication was the
405 result of a translocation based on homology of the Ty-elements which resulted in local copy number
406 increase during its strain development program (Entian and Kötter 2007).

407 [Long-read sequencing data reveals chromosome structure heterogeneity in](#) 408 [CEN.PK113-7D Delft](#)

409 CEN.PK113-7D has three confirmed *MAL* loci encoding genes for the uptake and hydrolysis of maltose:
410 *MAL1* on chromosome VIII, *MAL2* on chromosome III and *MAL3* on chromosome II (Additional file 2A). A
411 fourth *MAL* locus was identified in previous research on chromosome XI based on contour-clamped
412 homogeneous electric field electrophoresis (CHEF) and southern blotting with a probe for *MAL* loci
413 (Nijkamp *et al.* 2012). However, the nanopore assembly revealed no additional *MAL* locus despite the

414 complete assembly of Chromosome XI. The CEN.PK113-7D stock in which the fourth *MAL* locus was
415 obtained from Dr P. Kötter in 2001 and stored at -80°C since (further referred to as “CEN.PK113-7D
416 Delft”). In order to investigate the presence of the potential *MAL* locus, we sequenced CEN.PK113-7D
417 Delft using nanopore MinION sequencing. Two R7.3 flow cells (FLO-MIN103) produced 55x coverage
418 with an average read-length distribution of 8.5 Kbp and an R9 flow cell (FLO-MIN103) produced 47x
419 coverage with an average read-length distribution of 3.2 Kbp (Supplementary Figure S1). The error rate
420 was estimated to be 13% (Supplementary Figure S4) after aligning the raw nanopore reads to the
421 CEN.PK113-7D Frankfurt assembly. These reads were assembled into 24 contigs with an N50 of 736 Kbp
422 (Supplementary Table S1).

423 Alignment of the assembly of CEN.PK113-7D Delft to the Frankfurt assembly showed evidence of a
424 translocation between chromosomes III and VIII (Supplementary Figure S5). The assembly thus
425 suggested the presence of two new chromosomes: chromosomes III-VIII of size 680 Kbp and
426 chromosome VIII-III of size 217 Kbp (Figure 3). The translocation occurred between Ty-element
427 YCLW_{Ty2-1} on chromosome III and long terminal repeats YHRC_{delta5-7} on chromosome VIII. These
428 repetitive regions are flanked by unique genes *KCC4* and *NFS1* on chromosome III and *SPO13* and *MIP6*
429 on chromosome VIII (Figure 3). Nanopore reads spanning the whole translocated or non-translocated
430 sequence anchored in the unique genes flanking them were extracted for CEN.PK113-7D Delft and
431 Frankfurt. A total of eight reads from CEN.PK113-7D Delft supported the translocated
432 chromosome III-VIII architecture (largest read was 39 Kbp) and one 19 Kbp read supported the normal
433 chromosome III architecture. For CEN.PK113-7D Frankfurt, we found only one read of size 23 Kbp that
434 supported the normal chromosome III architecture but we found no reads that supported the
435 translocated architectures. This data suggested that CEN.PK113-7D Delft is in fact a heterogeneous
436 population containing cells with recombined chromosomes III and VIII and cells with original

437 chromosomes III and VIII. As a result, in addition to the *MAL2* locus on chromosome III, CEN.PK113-7D
438 Delft harboured a *MAL2* locus on recombined chromosome III-VIII. As the size of recombined
439 chromosome III-VIII was close to chromosome XI, the *MAL2* locus on chromosome III-VIII led to
440 misidentification of a *MAL4* locus on chromosome XI (Nijkamp *et al.* 2012). By repeating the CHEF gel
441 and southern blotting for MAL loci on several CEN.PK113-7D stocks, the *MAL2* on the translocated
442 chromosomes III-VIII was shown to be present only in CEN.PK113-7D Delft, demonstrating that there
443 was indeed chromosome structure heterogeneity (Additional File 5).

444 Discussion

445 In this study, we obtained a near-complete genome assembly of *S. cerevisiae* strain CEN.PK113-7D using
446 only a single R9 flow cell on ONT's MinION sequencing platform. 15 of the 16 chromosomes as well as
447 the mitochondrial genome and the 2-micron plasmid were assembled in single, mostly telomere-to-
448 telomere, contigs. This genome assembly is remarkably unfragmented, even when compared with other
449 *S. cerevisiae* assemblies made with several nanopore technology flow cells, in which 18 to 105
450 chromosomal contigs were obtained (Istace *et al.* 2017, McIlwain *et al.* 2016). Despite the long read
451 lengths obtained by Nanopore sequencing, the ribosomal DNA locus in chromosome XII could not be
452 completely resolved. In practice, this would require reads exceeding 1 Mb in length, which current
453 technology cannot yet deliver.

454 The obtained nanopore assembly is of vastly superior quality to the previous short-read-only assembly
455 of CEN.PK113-7D that was fragmented into over 700 contigs (Nijkamp *et al.* 2012). In addition to the
456 lesser fragmentation, the addition of 770 Kbp of previously unassembled sequence led to the
457 identification and accurate placement of 284 additional ORFs spread out over the genome. These newly
458 assembled genes showed overrepresentation for cell wall and cell periphery compartmentalization and
459 relate to functions such as sugar utilization, amino acid uptake, metal ion metabolism, flocculation and

460 tolerance to various stresses. While many of these genes are already present in the short-read assembly
461 of CEN.PK113-7D, copy number was shown to be an important factor determining the adaptation of
462 strains to specific growth conditions (Brown *et al.* 2010). The added genes may therefore be very
463 relevant for the specific physiology of CEN.PK113-7D under different industrial conditions (Brown *et al.*
464 2010). The ability of nanopore sequencing to distinguish genes with various similar copies is crucial in
465 *S. cerevisiae* as homologues are frequent particularly in subtelomeric regions, and paralogues are
466 widespread due to a whole genome duplication in its evolutionary history (Wolfe and Shields 1997).
467 Besides the added sequence, 6 Kbp of sequence of the short-read assembly was not present in the
468 nanopore assembly, mostly consisting of small unplaced contigs. Notably the absence of *BIO1* and *BIO6*
469 in the assembly was unexpected, as it constituted a marked difference between CEN.PK113-7D and
470 many other strains which enables biotin prototrophy (Bracher *et al.* 2017). Both genes were present in
471 the nanopore reads, but were unassembled likely due to the lack of reads long-enough to resolve this
472 subtelomeric region (a fragment of *BIO1* is located at the right-end of chromosome I). Targeted long-
473 read sequencing in known gaps of a draft assembly followed by manual curation could provide an
474 interesting tool to obtain complete genome assemblies (Loose *et al.* 2016). Alternatively, a more
475 complete assembly could be obtained by maximizing read length. The importance of read length is
476 illustrated by the higher fragmentation of the CEN.PK113-7D Delft assembly compared to the Frankfurt
477 one, which was based on reads with lower length distribution despite higher coverage and similar error
478 rate (Table 1, Supplementary Figures S1 and S5). Read-length distribution in nanopore sequencing is
479 highly influenced by the DNA extraction method and library preparation (Supplementary Figure S1). The
480 mitochondrial genome was completely assembled, which is not always possible with nanopore
481 sequencing (Giordano *et al.* 2017, Istace *et al.* 2017, McIlwain *et al.* 2016). Even with identical DNA
482 extraction and assembly methods, the mitochondrial genome cannot always be assembled, as illustrated

483 by its absence in the assembly of CEN.PK113-7D Delft. Overall, the gained sequence in the nanopore
484 assembly far outweighs the lost sequence relative to the previous assembly, and the reduction in
485 number of contigs presents an important advantage.

486 The use of long read sequencing enabled the discovery of a translocation between chromosomes III and
487 VIII, which led to the misidentification of a fourth MAL locus on chromosome XI of CEN.PK113-7D
488 (Nijkamp *et al.* 2012). Identification of this translocation required reads to span at least 12 Kbp due to
489 the large repetitive elements surrounding the translocation breakpoints, explaining why it was
490 previously undetected. While the translocation did not disrupt any coding sequence and is unlikely to
491 cause phenotypical changes (Naseeb *et al.* 2016), there may be decreased spore viability upon mating
492 with other CEN.PK strains. Our ability to detect structural heterogeneity within a culture shows that
493 nanopore sequencing could also be valuable in detecting structural variation within a genome between
494 different chromosome copies, which occurs frequently in aneuploid yeast genomes (Gorter de Vries *et*
495 *al.* 2017). These results highlight the importance of minimal propagation of laboratory microorganisms
496 to warrant genome stability and avoid heterogeneity which could at worst have an impact on phenotype
497 and interpretation of experimental results.

498 The nanopore assembly of CEN.PK113-7D constitutes a vast improvement of its reference genome
499 which should facilitate its use as a model organism. The elucidation of various homologue and paralogue
500 genes is particularly relevant as CEN.PK113-7D is commonly used as a model for industrial *S. cerevisiae*
501 applications for which gene copy number frequently plays an important role (Brown *et al.* 2010, Gorter
502 de Vries *et al.* 2017). Using the nanopore assembly as a reference for short-read sequencing of strains
503 derived from CEN.PK113-7D will yield more complete and more accurate lists of SNPs and other
504 mutations, facilitating the identification of causal mutations in laboratory evolution or mutagenesis
505 experiments. Therefore, the new assembly should accelerate elucidation of the genetic basis underlying

506 the fitness of *S. cerevisiae* in various environmental conditions, as well as the discovery of new strain
507 improvement strategies for industrial applications (Oud *et al.* 2012).

508 Acknowledgements

509 The authors would like to thank Dr. P. Kötter for sending CEN.PK113-7D Frankfurt, Dr. Kirsten Benjamin
510 for sending CEN.PK113-7D Amyris and Dr. Verena Siewers for sending CEN.PK113-7D Chalmers. We are
511 thankful to Prof. Jack T. Pronk (Delft University of Technology) and Dr. Niels Kuijpers (HEINEKEN
512 Supply Chain B.V.) for their critical reading of the manuscript.

513 This work was performed within the BE-Basic R&D Program (<http://www.be-basic.org/>), which was
514 granted an FES subsidy from the Dutch Ministry of Economic Affairs, Agriculture and Innovation (EL&I).
515 Anja Brickwedde was funded by the Seventh Framework Programme of the European Union in the
516 frame of the SP3 people support for training and career development of researchers (Marie Curie),
517 Networks for Initial Training (PITN-GA-2013 ITN-2013-606795) YeastCell.

518 **Author's contribution**

519 PdITC and ARGdV extracted high molecular weight DNA for Illumina and MinION sequencing. PdITC
520 performed Illumina sequencing. AS and MW constructed MinION sequencing libraries and performed
521 MinION genome sequencing. ARGdV and AB performed the CHEF and Southern-blot hybridization. AS,
522 ARGdV and MvdB performed the bioinformatics analysis. AS, ARGdV, MvdB, JMGD and TA were involved
523 in the data analysis and AS, ARGdV, JMGD and TA wrote the manuscript. JMGD and TA supervised the
524 study. All authors read and approved the final manuscript.

525 References

526

527 Bergström A , Simpson JT , Salinas F *et al.* (2014) A high-definition view of functional genetic variation
528 from natural yeast genomes. *Mol Biol Evol* **31**: 872-88.

529

530 Bracher JM , de Hulster E , Koster CC *et al.* (2017) Laboratory evolution of a biotin-requiring
531 *Saccharomyces cerevisiae* strain for full biotin prototrophy and identification of causal mutations. *Appl*
532 *Environ Microbiol* AEM. 00892-17.

533

534 Brown CA , Murray AW , Verstrepen KJ (2010) Rapid expansion and functional divergence of
535 subtelomeric gene families in yeasts. *Current biology : CB* **20**: 895-903.

536

537 Camacho C , Coulouris G , Avagyan V *et al.* (2009) BLAST+: architecture and applications. *BMC*
538 *bioinformatics* **10**: 421.

539

540 Canelas AB , Harrison N , Fazio A *et al.* (2010) Integrated multilaboratory systems biology reveals
541 differences in protein metabolism between two reference yeast strains. *Nat Commun* **1**: 145.

542

543 Carlson M , Celenza JL , Eng FJ (1985) Evolution of the dispersed *SUC* gene family of *Saccharomyces* by
544 rearrangements of chromosome telomeres. *Mol Cell Biol* **5**: 2894-902.

545

546 Cherry JM , Hong EL , Amundsen C *et al.* (2012) *Saccharomyces* Genome Database: the genomics
547 resource of budding yeast. *Nucleic Acids Res* **40**: D700-5.

548

549 Daran-Lapujade P , Daran J-MG , Kötter P *et al.* (2003) Comparative genotyping of the *Saccharomyces*
550 *cerevisiae* laboratory strains S288C and CEN. PK113-7D using oligonucleotide microarrays. *Fems Yeast*
551 *Res* **4**: 259-69.

552

553 Daran-Lapujade P , Daran J-MG , Luttik MAH *et al.* (2009) An atypical *PMR2* locus is responsible for
554 hypersensitivity to sodium and lithium cations in the laboratory strain *Saccharomyces cerevisiae* CEN.
555 PK113-7D. *Fems Yeast Res* **9**: 789-92.

556

557 Denayrolles M , de Villechenon EP , Lonvaud-Funel A *et al.* (1997) Incidence of *SUC-RTM* telomeric
558 repeated genes in brewing and wild wine strains of *Saccharomyces*. *Curr Genet* **31**: 457-61.

559

560 English AC , Richards S , Han Y *et al.* (2012) Mind the gap: upgrading genomes with Pacific Biosciences RS
561 long-read sequencing technology. *PLoS One* **7**: e47768.

562

563 Entian K-D , Kötter P (2007) 25 Yeast genetic strain and plasmid collections. *Method Microbiol* **36**: 629-
564 66.
565

566 Fischer G , James SA , Roberts IN *et al.* (2000) Chromosomal evolution in *Saccharomyces*. *Nature* **405**:
567 451-4.
568

569 Giordano F , Aigrain L , Quail MA *et al.* (2017) *De novo* yeast genome assemblies from MinION, PacBio
570 and MiSeq platforms. *Sci Rep* **7**: 3935.
571

572 González-Ramos D , Gorter de Vries AR , Grijseels SS *et al.* (2016) A new laboratory evolution approach
573 to select for constitutive acetic acid tolerance in *Saccharomyces cerevisiae* and identification of causal
574 mutations. *Biotechnol Biofuels* **9**: 173.
575

576 Goodwin S , Gurtowski J , Ethe-Sayers S *et al.* (2015) Oxford Nanopore sequencing, hybrid error
577 correction, and de novo assembly of a eukaryotic genome. *Genome Res* **25**: 1750-6.
578

579 Gorter de Vries AR , Pronk JT , Daran J-MG (2017) Industrial relevance of chromosomal copy number
580 variation in *Saccharomyces* yeasts. *Appl Environ Microbiol* **83**: e03206-16.
581

582 Holt C , Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for
583 second-generation genome projects. *BMC Bioinformatics* **12**: 491.
584

585 Istace B , Friedrich A , d'Agata L *et al.* (2017) *De novo* assembly and population genomic survey of natural
586 yeast isolates with the Oxford Nanopore MinION sequencer. *Gigascience* **6**: 1-13.
587

588 Jansen H , Dirks RP , Liem M *et al.* (2017) *De novo* whole-genome assembly of a wild type yeast isolate
589 using nanopore sequencing. *F1000Research* **6**.
590

591 Kim JM , Vanguri S , Boeke JD *et al.* (1998) Transposable elements and genome organization: a
592 comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome
593 sequence. *Genome Res* **8**: 464-78.
594

595 Koren S , Walenz BP , Berlin K *et al.* (2017) Canu: scalable and accurate long-read assembly via adaptive
596 k-mer weighting and repeat separation. *Genome Res* **27**: 722-36.
597

598 Kurtz S , Phillippy A , Delcher AL *et al.* (2004) Versatile and open software for comparing large genomes.
599 *Genome Biol* **5**: R12.
600

- 601 Li H , Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform.
602 *Bioinformatics* **26**: 589-95.
603
- 604 Li H , Handsaker B , Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools.
605 *Bioinformatics* **25**: 2078-9.
606
- 607 Lodolo EJ , Kock JLF , Axcell BC *et al.* (2008) The yeast *Saccharomyces cerevisiae* - the main character in
608 beer brewing. *Fems Yeast Res* **8**: 1018-36.
609
- 610 Loman NJ , Quick J , Simpson JT (2015) A complete bacterial genome assembled *de novo* using only
611 nanopore sequencing data. *Nat Methods* **12**: 733-U51.
612
- 613 Loman NJ , Quinlan AR (2014) Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics*
614 **30**: 3399-401.
615
- 616 Loose M , Malla S , Stout M (2016) Real-time selective sequencing using nanopore technology. *Nat*
617 *Methods* **13**: 751.
618
- 619 Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends in genetics* :
620 *TIG* **24**: 133-41.
621
- 622 Matheson K , Parsons L , Gammie A (2017) Whole-genome sequence and variant analysis of W303, a
623 widely-used strain of *Saccharomyces cerevisiae*. *G3*, DOI 10.1534/g3.117.040022g3. 117.040022.
624
- 625 McIlwain SJ , Peris D , Sardi M *et al.* (2016) Genome sequence and analysis of a stress-tolerant, wild-
626 derived strain of *Saccharomyces cerevisiae* used in biofuels research. *G3* **6**: 1757-66.
627
- 628 Milne I , Stephen G , Bayer M *et al.* (2012) Using Tablet for visual exploration of second-generation
629 sequencing data. *Brief Bioinform* bbs012.
630
- 631 Naseeb S , Carter Z , Minnis D *et al.* (2016) Widespread Impact of Chromosomal Inversions on Gene
632 Expression Uncovers Robustness via Phenotypic Buffering. *Mol Biol Evol* **33**: 1679-96.
633
- 634 Naumov GI , Naumova ES , Louis EJ (1995) Genetic mapping of the α -galactosidase *MEL* gene family on
635 right and left telomeres of *Saccharomyces cerevisiae*. *Yeast* **11**: 481-3.
636
- 637 Ng PC , Kirkness EF (2010) Whole genome sequencing. *Genetic variation*, p.^pp. 215-26. Springer.
638

- 639 Nijkamp J , Winterbach W , Van den Broek M *et al.* (2010) Integrating genome assemblies with MAIA.
640 *Bioinformatics* **26**: i433-i9.
641
- 642 Nijkamp JF , van den Broek M , Datema E *et al.* (2012) De novo sequencing, assembly and analysis of the
643 genome of the laboratory strain *Saccharomyces cerevisiae* CEN.PK113-7D, a model for modern industrial
644 biotechnology. *Microb Cell Fact* **11**: 36.
645
- 646 Oud B , Maris AJA , Daran JM *et al.* (2012) Genome-wide analytical approaches for reverse metabolic
647 engineering of industrially relevant phenotypes in yeast. *Fems Yeast Res* **12**: 183-96.
648
- 649 Papapetridis I , Dijk M , Maris AJA *et al.* (2017) Metabolic engineering strategies for optimizing acetate
650 reduction, ethanol yield and osmotolerance in *Saccharomyces cerevisiae*. *Biotechnol Biofuels* **10**: 107.
651
- 652 Pryde FE , Huckle TC , Louis EJ (1995) Sequence analysis of the right end of chromosome XV in
653 *Saccharomyces cerevisiae*: an insight into the structural and functional significance of sub-telomeric
654 repeat sequences. *Yeast* **11**: 371-82.
655
- 656 Sović I , Šikić M , Wilm A *et al.* (2016) Fast and sensitive mapping of nanopore sequencing reads with
657 GraphMap. *Nat Commun* **7**.
658
- 659 Teste M-A , François JM , Parrou J-L (2010) Characterization of a new multigene family encoding
660 isomaltases in the yeast *Saccharomyces cerevisiae*, the *IMA* family. *J Biol Chem* **285**: 26815-24.
661
- 662 Teunissen AW , Steensma HY (1995) Review: The dominant flocculation genes of *Saccharomyces*
663 *cerevisiae* constitute a new subtelomeric gene family. *Yeast* **11**: 1001-13.
664
- 665 van Dijk EL , Auger H , Jaszczyszyn Y *et al.* (2014) Ten years of next-generation sequencing technology.
666 *Trends in genetics : TIG* **30**: 418-26.
667
- 668 Venema J , Tollervey D (1999) Ribosome synthesis in *Saccharomyces cerevisiae*. *Annu Rev Genet* **33**: 261-
669 311.
670
- 671 Walker BJ , Abeel T , Shea T *et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant
672 detection and genome assembly improvement. *PLoS One* **9**: e112963.
673
- 674 Wolfe KH , Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome.
675 *Nature* **387**: 708.
676

677 Zhang Z , Schwartz S , Wagner L *et al.* (2000) A greedy algorithm for aligning DNA sequences. *Journal of*
678 *computational biology : a journal of computational molecular cell biology* **7**: 203-14.

679

680 **Tables and Figures**

681 **Table 1. Comparison of 454/Illumina and nanopore *de novo* assemblies of CEN.PK113-7D.** Summary of
682 *de novo* assembly metrics of CEN.PK113-7D Delft and CEN.PK113-7D Frankfurt. For the short-read
683 assembly, only contigs of at least 1 Kbp are shown (Nijkamp *et al.* 2012). The nanopore assembly of
684 CEN.PK113-7D Delft is uncorrected for misassemblies while CEN.PK113-7D Frankfurt was corrected for
685 misassemblies.

Data	CEN.PK113-7D Delft		CEN.PK113-7D Frankfurt
	Short read	Nanopore	Nanopore
Contigs (\geq 1 Kbp)	414	24	20
Largest contig	0.210 Mbp	1.08 Mbp	1.50 Mbp
Smallest contig	0.001 Mbp	0.013 Mbp	0.085 Mbp
N50	0.048 Mbp	0.736 Mbp	0.912 Mbp
Total assembly size	11.4 Mbp	11.9 Mbp	12.1 Mbp

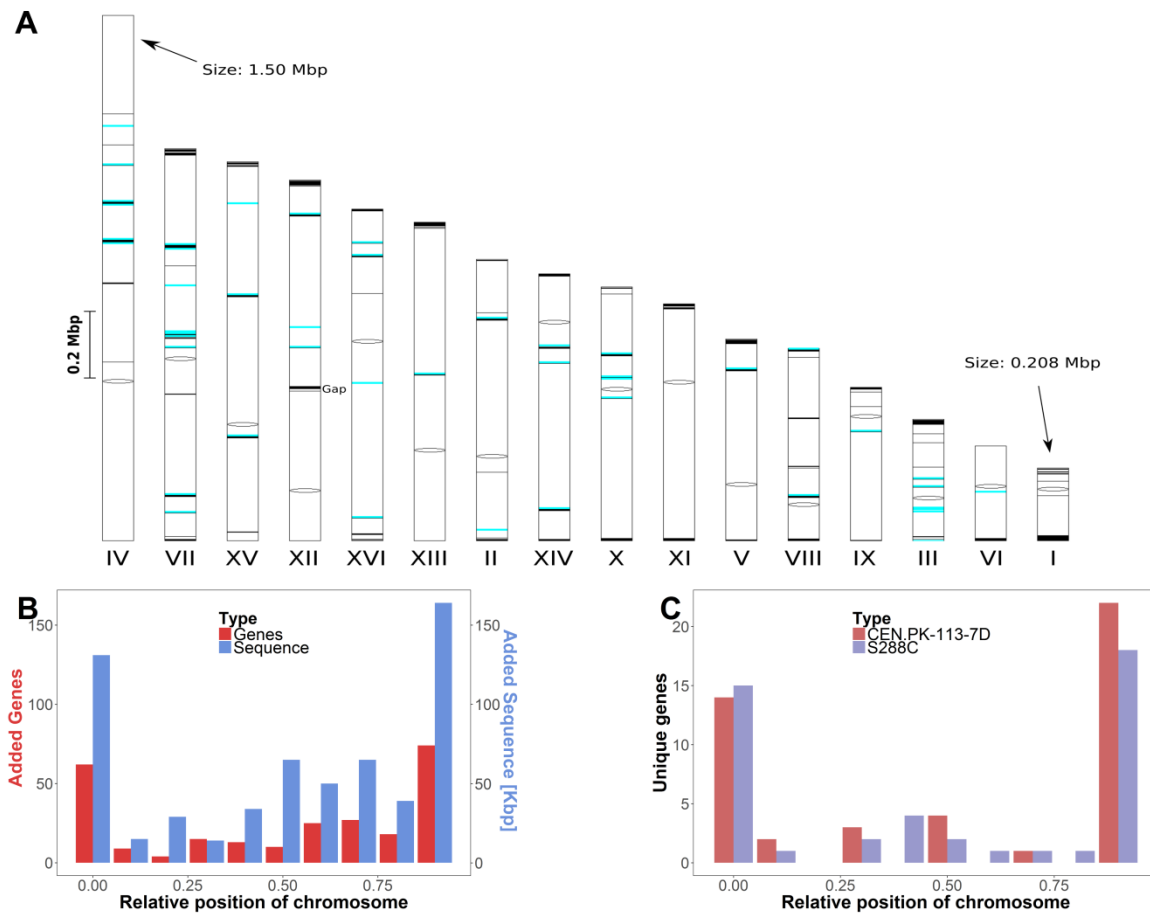
686

687 **Table 2: Presence in the nanopore assembly of genes identified as absent in CEN.PK113-7D in previous**
 688 **research.** For genes identified as absent in CEN.PK113-7D in two previous studies, the absence or
 689 presence in the nanopore assembly of CEN.PK113-7D is shown. 25 genes were identified previously by
 690 array comparative genome hybridisation (Daran-Lapujade *et al.* 2003) and 21 genes were identified by
 691 short-read genome assembly (Nijkamp *et al.* 2012). Genes which were not annotated by MAKER2 in
 692 S288C could not be analysed. Genes with an alignment to genes identified as missing in the nanopore
 693 assembly of at least 50% of the query length and 95% sequence identity were confirmed as being
 694 absent, while those without such an alignment were identified as present. The presence of these genes
 695 was verified manually, which revealed the misannotation of YPL277C as YOR389W.

	Not analysed	Absent in assembly	Present in assembly
Daran-Lapujade et al	YAL064C-A, YAL066W, YAR047C, YHL046W-A, YIL058W, YOLO13W-A	YAL065C, YAL067C, YBR093C, YCR018C, YCR105W, YCR106W, YDR038C, YDR039C, YHL047C, YHL048W, YNR070W, YNR071C and YNR074C	YAL069W, YDR036C, YDR037W, YJL165C, YNR004W, and YPL277C (misannotated as YOR389W)
Nijkamp et al	Q0140, YDR543C, YDR544C, YDR545W, YIL046W-A, YLR154C-H, YLR156C-A, YLR157C-C, YLR159C-A, YOR029W and YOR082C	YBR093C, YCR040W, YCR041W, YDR038C, YDR039C and YDR040C	YDR036C, YHL008C, YHR056C and YLR055C

696

697

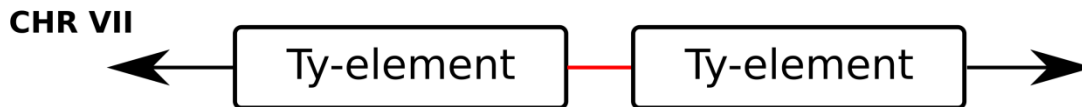
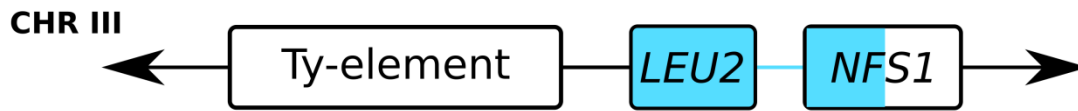


698 **Figure 1: Overview of gained and lost sequence and genes in the CEN.PK113-7D Frankfurt nanopore**
 699 **assembly relative to the short-read CEN.PK113-7D assembly and to the genome of S288C.** The two
 700 unplaced subtelomeric contigs and the mitochondrial DNA were not included in this figure. **(1A)**
 701 **Chromosomal location of sequence assembled in the nanopore assembly which was not assembled**
 702 **using short-read data.** The sixteen chromosome contigs of the nanopore assembly are shown.
 703 Chromosome XII has a gap at the *RDN1* locus, a region estimated to contain more than 1 Mbp worth of
 704 repetitive sequence (Venema and Tollervey 1999). Centromeres are indicated by black ovals, gained
 705 sequence relative to the short-read assembly is indicated by black marks and 46 identified
 706 retrotransposon Ty-elements are indicated by blue marks. The size of all chromosomes and marks is
 707 proportional to their corresponding sequence size. In total 611 Kbp of sequence was added within the

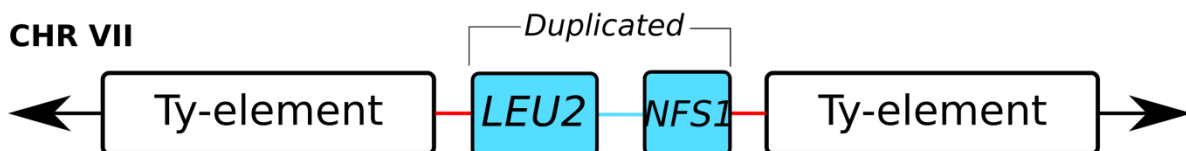
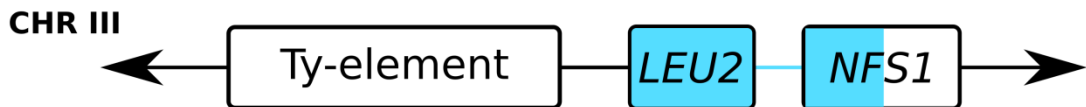
708 chromosomal contigs. **(1B) Relative chromosome position of sequences and genes assembled on**
709 **chromosome contigs of the nanopore assembly which were not assembled using short-read data.** The
710 positions of added sequence and genes were normalized to the total chromosome size. The number of
711 genes (red) and the amount of sequence (cyan) over all chromosomes are shown per tenth of the
712 relative chromosome size. **(1C) Relative chromosome position of gene presence differences between**
713 **S288C and CEN.PK113-7D.** The positions of the 45 genes identified as unique to CEN.PK113-7D and of
714 the 44 genes identified as unique to S288C were normalized to the total chromosome size. The number
715 of genes unique to CEN.PK113-7D (red) and to S288C (purple) are shown per tenth of the relative
716 chromosome position.

717

S288C:

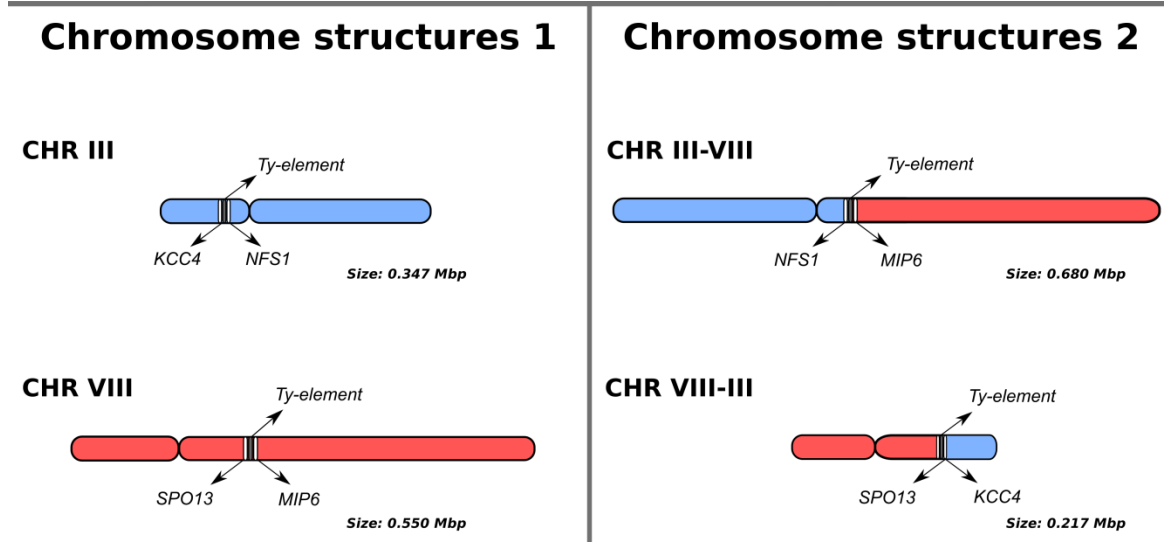


CEN.PK113-7D:



718

719 **Figure 2: LEU2 and NFS1 duplication in chromosome VII of CEN.PK113-7D.** The nanopore assembly
720 contains a duplication of *LEU2* and part of *NFS1* in CEN.PK113-7D. In S288C, the two genes are located in
721 chromosome III next to a Ty element. In CEN.PK113-7D, the two genes are present in chromosome III
722 and in chromosome VII. The duplication appears to be mediated by Ty-elements. Note that the
723 additional copy in chromosome VII is present in between two Ty-elements and contains only the first
724 ~500 bp of *NFS1*. The duplication is supported by long-read data that span across the *LEU2*, *NFS1*, the
725 two Ty-elements, and the neighboring flanking genes (not shown).



726

727 **Figure 3: Overview of chromosome structure heterogeneity in CEN.PK113-7D Delft for CHRIII and**

728 **CHRVIII which led to the misidentification of a fourth MAL locus in a previous short-read assembly**

729 **study of the genome of CEN.PK113-7D.** Nanopore reads support the presence of two chromosome

730 architectures: the normal chromosomes III and VIII (left panel) and translocated chromosomes III-VIII

731 and VIII-III (right panel). The translocation occurred in Ty-elements, large repetitive sequences known to

732 mediate chromosomal translocations in *Saccharomyces* species (Fischer *et al.* 2000). Long-reads are

733 required to diagnose the chromosome architecture via sequencing: the repetitive region between *KCC4*

734 to *NFS1* in chromosome III exceeds 15 Kbp, while the region between *SPO13* and *MIP6* in chromosome

735 VIII is only 1.4 Kbp long. For the translocated architecture, the region from *NFS1* to *MIP6* in chromosome

736 III-VIII exceeds 16 Kbp and the distance from *SPO13* to *KCC4* in chromosome VIII-III is nearly 10 Kbp.

737

738