Published online 14 October 2021

https://doi.org/10.1093/nargab/lqab085

Establishment of a near-contiguous genome sequence of the citric acid producing yeast *Yarrowia lipolytica* DSM 3286 with resolution of rDNA clusters and telomeres

Tobias Luttermann[1,2], Christian Rückert[1], Daniel Wibberg[3,4], Tobias Busche[1], Jan-Philipp Schwarzhans[2], Karl Friehs[2] and Jörn Kalinowski[1,*]

[1]Microbial Genomics and Biotechnology, Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, NRW 33615, Germany, [2]Fermentation Engineering, Bielefeld University, Bielefeld, NRW 33615, Germany, [3]Genome Research of Industrial Microorganisms, Bielefeld University, Bielefeld, NRW 33615, Germany and [4]German Network for Bioinformatics Infrastructure (de.NBI), Bielefeld, NRW 33615, Germany

Received December 22, 2020; Revised September 01, 2021; Editorial Decision September 07, 2021; Accepted October 13, 2021

ABSTRACT

Yarrowia lipolytica **is an oleaginous yeast that is particularly suitable for the sustainable production of secondary metabolites. The genome of this yeast is characterized by its relatively large size and its high number of different rDNA clusters located in its telomeric regions. However, due to the presence of long repetitive elements in the sub-telomeric regions, rDNA clusters and telomeres are missing in current genome assemblies of *Y. lipolytica*. Here, we present the near-contiguous genome sequence of the biotechnologically relevant strain DSM 3286. We employed a hybrid assembly strategy combining Illumina and nanopore sequencing reads to integrate all six rDNA clusters as well as telomeric repeats into the genome sequence. By fine-tuning of DNA isolation and library preparation protocols, we were able to create ultra-long reads that not only contained multiples of mitochondrial genomes but also shed light on the inter- and intra-chromosomal diversity of rDNA cluster types. We show that there are ten different rDNA units present in this strain that additionally appear in a predefined order in a cluster. Based on single reads, we also demonstrate that the number of rDNA repeats in a specific cluster varies from cell to cell within a population.**

INTRODUCTION

Yarrowia lipolytica represents one of the most important non-conventional yeasts and serves as a model organism in various research fields based on its specific morphological and physiological characteristics (1). As a dimorphic organism, this yeast commands the ability, depending on the cultivation conditions, to either grow as yeast cells or to form hyphae and pseudohyphae (2). Mainly isolated not only from dairy products and sausages but also from waste waters or hypersaline habitats, such as marine waters, *Y. lipolytica* naturally produces and secrets high amounts of lipases and proteases (1–2 g l^{-1}) (1). These features make this yeast promising for recombinant protein expression and *Y. lipolytica* constitutes an attractive alternative to the well-studied *Pichia pastoris* (*Komagataella phaffii*) as a yeast-based expression system (3). With the oleic acid-inducible promoter p*POX2*, *Y. lipolytica* even has a more ecological alternative to the methanol-inducible *AOX1* promoter in *P. pastoris* (4,5). Additionally, *Y. lipolytica* is capable to effectively utilize hydrophobic substrates as a sole carbon source and to secrete large amounts of organic acids (up to 240 g l^{-1}), such as citric acid (CA), which created a high industrial interest in this yeast already in the mid-1960s. A comparative study of various *Yarrowia lipolytica* strains revealed *Y. lipolytica* DSM 3286 as one of the best CA-producing strains (6,7).

The current reference genome for *Y. lipolytica* is that of strain CLIB 122 (8) which was obtained by large insert clone-based Sanger sequencing and consists of six large scaffolds. While the estimated total size of the six chromosomes was 21 Mb (9), the CLIB 122 assembly corresponded to only 20.5 Mb as the assembly contained 13 gaps. Additionally, the telomeric ends as well as rDNA repeats could not be integrated into this genome assembly. Owing to the repetitive character of rDNA clusters, telomeres and sub-telomeric regions in general, chromosome ends are usually difficult to integrate into the complete genomic sequence using short-read sequencing technologies. This is especially true for *Y. lipolytica* strains that contain between

[*]To whom correspondence should be addressed. Tel: +49 521 106 8756; Fax: +49 521 106 89041; Email: joern@cebitec.uni-bielefeld.de

five and seven rDNA clusters, ranging from 190 to 620 kb in size (8,9). The rDNA clusters are composed of tandem repeat units with each of them containing the 18S, 5.8S and 28S rRNA gene, that together encode the 35–45S precursor RNA, as well as an intergenic region (IGR). It was also shown that *Y. lipolytica* has two major size classes of rDNA units that have a length of 7.7 and 8.7 kb. The different size classes vary in the length of their IGRs. Specifically, they differ in their amount of PvuII restriction sites as well as in their number of short direct repeats (10). Via segregation studies and Southern hybridization, it was proven that the 7.7 kb size class encompasses multiple sub-types (10,11). In the model yeast *Saccharomyces cerevisiae*, the 2.3 kb long IGR is divided by the 5S rRNA gene into a 0.9 kb non-transcribed spacer 1 (NTS1) containing transcription terminators, an enhancer element, and a replication fork barrier (RFB), and the 1.2 kb NTS2 that harbors a cohesin-associated region, important for sister-chromatid separation, the upstream promoter element, the core promoter encompassing a ∼50 bp region upstream of the transcription start site (TSS), and the rDNA autonomous replication sequence (rARS) (12,13). Replication of yeast rDNA starts bi-directionally at the rARSs that are located next to actively transcribed rRNA genes (14). However, replication that occurs in the direction of the terminators is stopped at the RFB to prevent the collision of the transcription and the replication machinery (15). In contrast to *S. cerevisiae* but similar to *P. pastoris*, the IGR of *Y. lipolytica* does not carry the 5S rRNA genes which are instead dispersed throughout the genome (10,16). In a recent study applying Illumina short-read together with PacBio long-read sequencing (17), it was shown via Iris long-range genome mapping that rDNA clusters in *Y. lipolytica* are located at the ends of the chromosomes in proximity to the telomeres. However, the authors (17) were not able to assemble the rDNA clusters and the telomeric sequences into the complete genomic sequence due to the length of these repetitive elements. Therefore, new sequencing approaches must be developed to overcome this challenge.

Here, we present for the first time a *Yarrowia lipolytica* hybrid genome sequencing and assembly approach by using very long reads obtained by nanopore sequencing, combined with short Illumina reads to create the *Y. lipolytica* DSM 3286 genome on a novel level of quality and contiguity. Additionally, the individual very long reads gave insights into the architecture of the mitochondrial genome as wells as into the heterogeneity of rDNA IGRs and cluster lengths within a population. The data presented here provide a strategy to address other genomes of eukaryotic origin and an annotated genome for future biotechnological application.

MATERIALS AND METHODS

DNA isolation and next-generation sequencing

For the isolation of genomic DNA, 25 ml YPD medium (2% yeast extract, 1% peptone, 2% dextrose) were inoculated with a single colony of *Y. lipolytica* DSM 3286. This strain was obtained from the DSMZ (Leibniz Institute German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany). The cells were cultivated in a baffled shake flask overnight at 28°C and 120 min^{-1}. In case of Illumina sequencing, genomic DNA was then isolated with the MasterPure™ Yeast DNA Purification Kit (Epicentre). Illumina sequencing libraries were prepared using either the TruSeq DNA PCR-Free library prep kit (for paired-end sequencing) or the Nextera Mate Pair Library Prep Kit, both from Illumina Inc. Both libraries were sequenced on a MiSeq sequencer using the 600 cycle v3 chemistry in 2×300 bp sequencing runs (Supplementary Table S1).

For the generation of high molecular weight DNA and ultra-long nanopore sequencing reads, we at first generated spheroplasts based on the EasySelect™ *Pichia* Expression kit (Invitrogen) manual. Therefore, an overnight culture (see above) was used to inoculate 10 ml YPD medium with an initial OD_{600} of 0.1. The cells were then grown to an OD_{600} of 5–10 under the same conditions as described above. Afterwards, the cells were centrifuged at 1500 × *g* and room temperature for 5 min followed by two washing steps with 10 ml sterile water. The cells were then resuspended in 2 ml freshly prepared SCED (1 M sorbitol, 1 mM EDTA, 10 mM sodium citrate, pH 5.8, 10 mM DTT) and transferred to a tube. After that, 75 U of Zymolyase E1004 (Zymo Research) were added, mixed briefly, and the cell suspension was incubated on a rotating shaker for 50 min at 37°C. Eventually, the resulting spheroplasts were used as the input material for the enzymatic lysis protocol of the NucleoBond® HMW DNA kit (Macherey-Nagel) starting with step 2. Different from the protocol, the DNA was resuspended overnight in 16 μl nuclease-free water. From this, 2 μl were used for quality control and concentration measurements. To begin the DNA sequencing library preparation for Rapid Sequencing (SQK-RAD004, Oxford Nanopore Technologies, ONT), the fragmentation mix FRA was added directly to the DNA droplet in the 50 ml tube from the DNA isolation. However, only 1 μl FRA was added to the DNA template to create less but longer fragments. Contents were mixed by carefully flicking the tube. Due to the thickness of the 50 ml tube, the subsequent incubation times were increased to 2 and 5 min for incubation at 30 and then 80°C, respectively, and incubation was carried out in a water bath. Consequently, 15 μl *tag*mented DNA was taken into the next step. One microliter of the rapid adapter RAP was added to the droplet and contents were mixed carefully. The reaction mixture was incubated for 5 min at room temperature. Flow cell priming and the remaining steps of the library preparation were performed according to the protocol. The prepared library was loaded into the flow cell using a wide-bore pipette tip. For sequencing of the *Y. lipolytica* DSM 3286 genome, we used an ONT MinION device and R9.4 as well as R9.5 flow cells (Supplementary Table S1; for more details see GenBank deposition under accession number PRJNA641784).

Hybrid genome assembly

The genome sequence of *Y. lipolytica* DSM-3286 was established in an approach combining short but highly accurate Illumina reads with significantly longer but more error-prone Oxford nanopore reads to assemble a complete high-quality genome using a hybrid assembly pipeline. First, long 1D reads produced by Oxford Nanopore's MinION using the R9.4 as well as the R9.5 flow cells were assembled into contigs using the canu long read assembler v.1.6 (18). To improve consensus accuracy, the assembled contigs were then polished using Nanopolish (19). Therefore, the reads were mapped to the contigs assembled with canu using minimap2 (20) and, after sorting and indexing with samtools (21), the resulting bam file was fed to Nanopolish. For final polishing, pilon (22) was applied using Illumina MiSeq sequencing reads to correct substitutions and smaller indels. In total, 19 rounds of pilon polishing were performed, the first 6 using bwa mem (23) as for mapping of the mate Pair and paired-end reads, followed by 13 rounds of mapping using bowtie2 (24) until no further changes were introduced by pilon.

Gene prediction and genome annotation

Gene prediction and genome annotation for *Y. lipolytica* DSM 3286 was performed essentially as recently described (25,26) with smaller modifications. In brief, gene prediction was performed by applying Augustus version 3.3 (27) using default settings and a parameter set designed for *Y. lipolytica*. Afterward, predicted genes were functionally annotated using a modified version of the genome annotation platform GenDB 2.0 for eukaryotic genomes as described (28). Complete genomes were compared using Mauve (29).

RNA isolation, RNA-Seq and transcription start site (TSS) detection

The following procedures were used to generate data that will be published in a transcriptomic and proteomic study of *Y. lipolytica* DSM 3286 elsewhere. However, the generated data were used here to give new insight into the structure of the rDNA IGRs.

In short, two shake flask cultivations were carried out to generate a diverse RNA pool. Therefore, 25 ml of YPD medium and 25 ml of buffered YNB minimal medium (1.34% YNB, yeast nitrogen base, 0.00001% Biotin, 2% glucose, 100 mM potassium phosphate, pH 6.0) were inoculated and grown overnight as described above. Subsequently, the overnight cultures were used to inoculate 100 ml YPD and 100 ml buffered YNB with an initial OD_{600} of 0.1. The cells were cultivated in 1 l baffled shake flasks as described above and three 1 ml samples were taken from each cultivation at an OD_{600} of 4 as well as once the cells reached the stationary growth phase. After that, the cells were immediately centrifuged for 30 s at 3,000 × *g*. The total RNA was isolated with the RNeasy Mini Kit (Qiagen), pooled and sent to Vertis Biotechnology AG where a 5′-end-specific cDNA library was constructed. The library was sequenced on an Illumina MiSeq.

For the automated TSS detection with ReadXplorer (30), only uniquely mapping reads were used. Different from the default setting for the TSS analysis parameters, the maximum distance to features (e.g. coding sequences) to be assigned to an TSS was increased from 300 to 2000 nt to allow for the detection of long 5′ UTRs. Additionally, the sequence window in which all detected TSS will be associated with the most significant TSS was increased from 3 to 10 nt.

Statistical analysis of the rDNA units

The analysis of the rDNA unit length variations and the IGR characteristics of the different chromosomes of *Y. lipolytica* DSM 3286, respectively, was performed based on single nanopore reads and on the assembly. Therefore, a nucleotide BLAST (31) was performed with a single sequence of the 18S rRNA gene of *Y. lipolytica* DSM 3286 on all nanopore reads from the libraries RAD1 and RAD_HMW (Supplementary Table S1) to identify the reads that contain rDNA units. After that, the length of an rDNA unit was determined by calculating the distance from the 5′ end of an 18S rRNA gene to the 5′ end of the next gene. A second nucleotide BLAST was carried out with unique sequences from all chromosome ends to allocate the identified rDNA-containing reads to the specific chromosome ends. In case of the detailed IGR analysis, the intergenic sequences were analyzed in regards of the features analyzed by van Heerikhuizen *et al.* (10). PvuII recognition sites as well as other features were annotated with Geneious Prime 2020.1.1 and repeats were identified with Tandem repeat finder (32). For identification of motif sites within the repetitive regions and to scan for the discovered motifs, the MEME Suite was used (33). Comparative analysis of the IGRs was performed using MUSCLE (34) and visualized via Geneious Prime 2020.1.1.

Copy number analysis of rDNA via Droplet Digital™ PCR

The Droplet Digital™ PCR (ddPCR, BIO-RAD) was used to calculate the amount of rDNA unit copies within the genome of *Y. lipolytica* DSM 3286. The target for the detection of rDNA unit copies was selected to be the 28S rRNA gene whereas the *ENA2* gene of chromosome A and the *SPT6* gene of chromosome B would function as single-copy reference markers. Primers (Supplementary Table S2) were designed to allow for an amplification of a 157–186 nt product. Genomic DNA was isolated as described above for the generation of high molecular weight DNA. Prior to the sample preparation for the ddPCR, we digested the genomic DNA of *Y. lipolytica* DSM 3286 with HindIII (Thermo Fisher Scientific) that cuts once per rDNA unit within the rARS in *Y. lipolytica* and therefore divides every rDNA cluster into its single units. This facilitates that every DNA fragment would carry only one PCR target at maximum. Based on the reported 100 rDNA unit copies present in the genome of *Y. lipolytica* (10,35), we prepared two QX200™ ddPCR™ EvaGreen® Supermix (BIO-RAD) reaction mixtures per target and reference gene from the digested DNA with 20 and 200 genome copies per 20 μl reaction corresponding to 2000 and 20 000 rDNA unit copies, respectively. After the PCR, the droplet generation and all subsequent steps of the ddPCR were performed according to the Droplet Digital™ PCR application guide (BIO-RAD).

RESULTS

Assembly of a near-complete genome sequence of *Yarrowia lipolytica* DSM 3286

The first assembly of the *Y. lipolytica* DSM 3286 genome was created using paired-end and shotgun sequencing data obtained from a HiSeq 1500 sequencing platform (Illumina Inc.) with Newbler v.2.8 (36). A total of 1.53 billion bases was assembled, yielding 26 scaffolds representing all chromosomes and the mitochondrial genome. The largest contig was 427 680 bp long (Table 1) and the total length of the six chromosomes was 20.4 Mb. Telomeric repeats as well as rDNA clusters were absent in this Illumina-only genome assembly, due to their repetitive nature.

In order to reconstruct the complete genome sequence including the telomeric and sub-telomeric regions, we adapted and combined the protocols for DNA isolation and the preparation of a nanopore rapid sequencing library (see Materials and Methods). With this protocol we were able to reduce the number of pipetting steps to one while also using a wide-bore pipette tip to further minimize the shear forces. The library (RAD_HMW, Supplementary Table S1) was sequenced on a MinION (ONT) R9.5 flow cell resulting in 1.42 billion base pairs that were assembled into nine high-quality contigs of which six represent the chromosomes of *Y. lipolytica* DSM 3286 with a total size of 22.4 Mb and an average GC content of 49.0%, two the mitochondrion, 153 and 179 kb in length with a GC content of 22.5%, and the remaining one a rDNA cluster that is 504 kb long and has no unique anchor points to any chromosome end. The two mitochondrial contigs contain multiple copies of the mitochondrial scaffold from the Illumina-only assembly. Compared to the latter, the total length of the chromosomes in the nanopore-based assembly was increased by 1.0 Mb indicating an extension of the chromosome ends with sub-telomeric and telomeric regions. Both assemblies had a similar coverage with 75.1- and 67.5-fold for the Illumina-based and the nanopore-based assembly, respectively, while the coverage of the two mitochondrial contigs from the ONT assembly corresponds to a 1183-fold coverage of a single circular *Y. lipolytica* mitochondrion. For the final assembly, we used nine nanopore datasets followed by a polishing step with the pylon tool applying four Illumina datasets (Supplementary Table S1). Compared to the assembly that was obtained from a single ONT run, the two mitochondrial contigs were merged into one with the same size as the mitochondrial reference genome of the *Y. lipolytica* strain CBS 599 (GenBank: KC993177.1).

Automated annotation of the genome

The automated annotation resulted in 6439 protein coding sequences (CDS) present in the genome sequence. Additionally, 511 tRNA genes as well as 334 rRNA genes were predicted. In total, six rDNA clusters could be identified that are located at the end of the chromosomes. In detail, the rDNA clusters were situated at both ends of chromosome B, the right end of chromosome C, as well as the left ends of chromosomes D, E and F. For *Y. lipolytica* DSM 3286, 114 copies of the 5S rRNA gene were found, spread in the genome of which 40 were present as hybrid dicistronic tRNA-5S rRNA fusions. Additionally, we were also able to annotate 173 telomeric repeats with the sequence 5′-GGGTTAGTCA-3′ that are located at both ends of chromosomes A and D as well as the right end of chromosomes E and F. Out of all rDNA cluster-containing chromosome ends, only the cluster D ended in telomeric repeats.

Insights into mitochondrial replication and architecture

By applying our improved protocol to generate ultra-long reads, we were able to increase the maximum read length from 244 to 493 kb and to overall raise the ratio of longer reads. In comparison to the normal transposon-based library preparation protocol for rapid nanopore sequencing that is usually characterized by an N50 read length of 20–25 kb (our unpublished data), the adapted protocol led to an increase of the N50 read length to 64 kb. However, the total sequencing yield was decreased from 6.16 to 1.42 billion sequenced bases (Supplementary Figure S1A). The lower yield can be linked to too many unligated adapters that reduce the life of a pore and therefore limit its sequencing capability. The higher abundance of ultra-long reads led to a significant longer assembly of three chromosome ends (Supplementary Figure S1B). A closer investigation of the read length distribution also reveals a sharp peak at a sequence length of approximately 50 kb. Most of those reads have a low GC content of 22.5% and can be mapped onto the mitochondrial reference genome indicating that they represent full-length mitochondrial genomes. Additionally, the apparent high coverage is in accordance with the high copy number of the mitochondrion.

When taking a closer look at the reads to investigate the large mitochondrial contigs originating from the single run ONT assembly, we were also able to identify 380 reads that are exceeding the length of a circular mitochondrial genome. When sorting all reads that can be fully aligned to the mitochondrion by length from smallest to longest, they can be divided into three groups with the first set of reads approximating a limit that is of the size of the single circular reference mitochondrial genome (Figure 1). Interestingly, the second group also approximates a limit that appears to be 94.0 kb, roughly the size of two genomic copies. The remaining 50 reads show a sharp increase with the longest read being 159.7 kb long that would be equivalent to more than three copies of the mitochondrial genome. Annotation of the mitochondrial reads based on the reference genome with Geneious Prime 2020.1.1 did not reveal any telomeric features which excludes the possibility that these species constitute linear-mapping (37) variants. However, it could be observed for most of the reads at the thresholds that the features of one end transitioned into the other end thereby depicting a circular characteristic.

Ultra-long reads reveal the structure and heterogeneity of rDNA clusters in *Y. lipolytica* DSM 3286

Although our assembly pipeline already led to chromosome ends containing rDNA clusters as well as telomeric repeats, the terminal DNA sequences had to be refined manually by using individual reads. This refinement had two reasons, the first being the heterogeneity of rDNA copy numbers (see below) in individual chromosomes that disturbed the assembly, and the second being the likeliness of a wrong assembly due to the fact that the nanopore reads have a significant error rate.

To prepare for the manual assembly refinement step, we first wanted to verify the presence of the two major types for our strain *in silico*. For this reason, we analyzed all

Table 1. *Y. lipolytica* DSM 3286 sequencing read and assembly statistics

Assembly	Illumina HiSeq 1500	ONT MinION[1]	ONT MinION[2]
Total length	20 432 711	22 439 472	23 647 606
Number of reads	6 968 540	36 515	1 308 298
Number of bases	1 533 837 762	1 417 586 013	12 189 829 440
Scaffolds (Illumina) / Contigs (ONT, after polishing)	26	n.d.	8
Long contigs (Illumina) / contigs (ONT)	477	11	379
Coverage	75.07	67.50	580.47
Median read length / bp	-	27 507	4618
Mean read length / bp	-	38 822	9317
N50 read / bp	-	64 822	18 398
N50 contig (Iillumina)	86,187	-	-
Maximum read / bp	300	493 808	493 808
Maximum contig (Illumina) / bp	427 680	-	-

n.d. = not determined
[1] = assembly based on one run (RAD_HMW)
[2] = assembly based on all nine runs

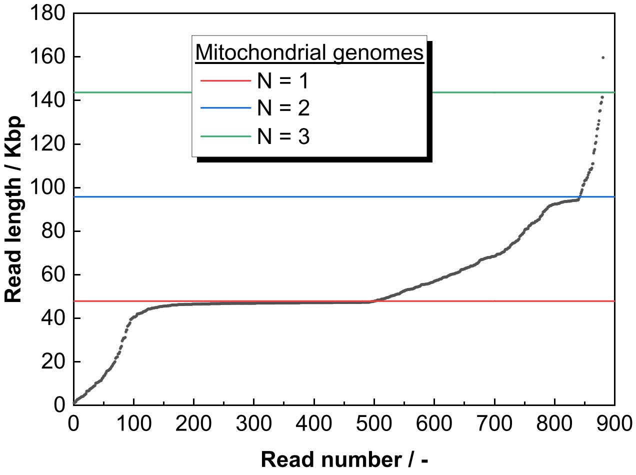

Figure 1. Length characteristic of 881 nanopore sequencing reads that can be mapped to the mitochondrial genome of *Y. lipolytica* DSM 3286. Reads were ordered by length. Colored lines indicate multiples of the mitochondrial genome size.

nanopore reads that contained multiple rDNA units and calculated the length of each rDNA unit in every read. In total, 148 sequencing reads with unique anchor points to specific chromosome ends from two libraries (RAD1 and RAD_HMW, Supplementary Table S1) with overall 519 rDNA units were used for this analysis. The sample sizes *n* (number of units) for the determination of the unit length distribution for each cluster are as followed: 98 (chromosome B, right end; B_R), 123 (B_L), 87 (C_R), 51 (D_L), 76 (E_L) and 84 (F_L), respectively.

The two identified major size classes had a length of 7.8 and 8.4 kb, respectively. However, the variance of the rDNA unit size in the 8.4 kb class was wider than for the smaller class and analysis of the rDNA unit length specific for chromosome ends revealed that there are indeed not two but three major size classes with the third being 8.6 kb in size (Figure 2A). Interestingly, all rDNA clusters except those of chromosome B seem to have rDNA units of the 8.4 kb class while chromosome B, which is the only chromosome that has two rDNA clusters, represents the other two major classes. While the unit length distribution for the ends of chromosomes C, D, E and F appears to be differing from each other in the direction of the minimum unit length, the third quartile is approximately at 8450 bp in all cases (Figure 2B). Reads with an average rDNA unit size of the 7.8 kb can be mapped solely to the left end of chromosome B whereas the third class can only be observed at its right end. Nevertheless, the broad distribution of unit lengths indicated the presence of IGR sub-types.

Now, to verify the automated assembly regarding correct rDNA cluster lengths, we compared each chromosome end of the final assembly with the longest read that could be mapped to a unique part of the specific chromosome end. It turned out that out of the rDNA cluster containing ends only the left end of chromosome D was assembled correctly in the automatic assembly process. The right end of chromosome C and the left end of E showed a significant difference between the analyzed reads and the assembly in terms of numbers of rDNA units. While the rDNA cluster of chromosome C only contained seven rDNA units in the assembly, the longest reads had eight additional copies and furthermore finished in telomeric repeats, indicating a correct chromosome end. In the manually refined assembly, this chromosome end was therefore extended by 63.4 kb. Interestingly, two reads were found for this cluster that end in telomeric repeats. However, the reads differ in their amount of rDNA units with one containing 15 units while the other only has twelve. Upon further inspection, we found that also the reads mapped to rDNA cluster D showed a variance regarding their number of encoded units. Out of the nine reads, that contained telomeric repeats, one showed six rDNA units, two had seven, another one eight, and five reads had nine copies indicating a heterogeneity in rDNA cluster length within a population. Additionally, out of the reads with no telomeric repeats, none had more than nine rDNA units. In contrast, the rDNA cluster of chromosome E was 166 kb longer in the assembly than the longest read suggested. After closer examination of this assembled rDNA cluster, we noticed that the cluster appeared to have two parts. One, that was matched by the longest read, only seemed to contain mostly rDNA units of the 8.7 kb type whereas the second part, that was only present in the assembly, showed many units of the 7.8 kb type.

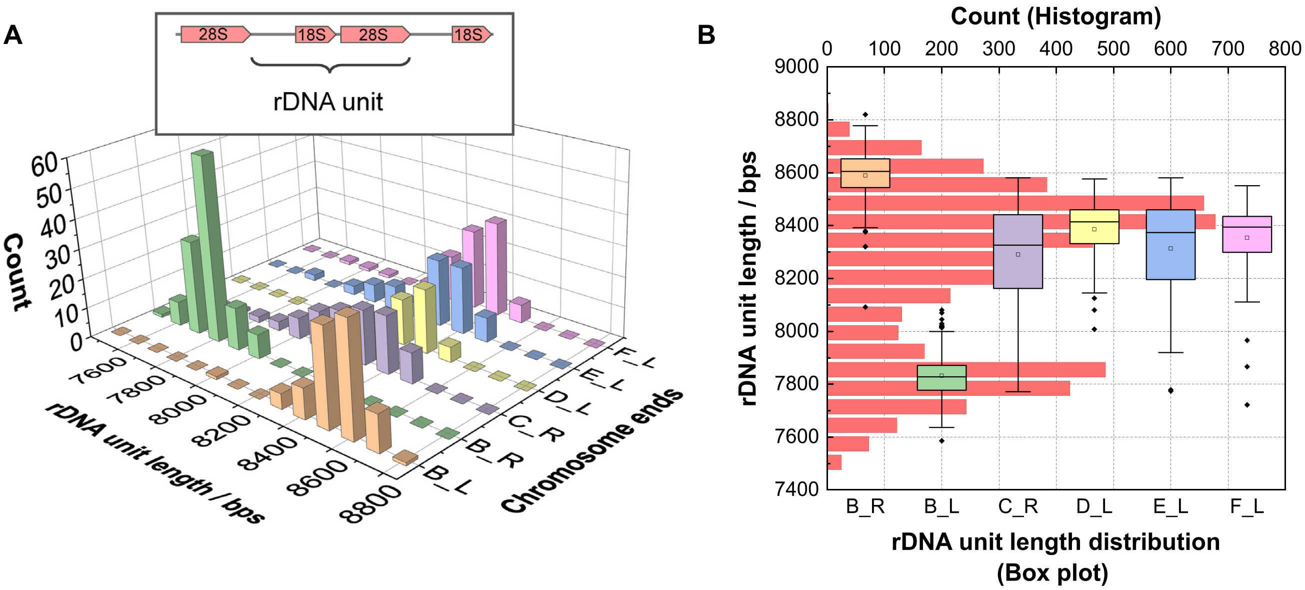

Figure 2. Analysis of the rDNA unit length distribution per rDNA cluster in *Y. lipolytica* DSM 3286. The rDNA unit length spans a sequence containing the 18S and 28S gene unit and the subsequent IGR. (**A**) For each nanopore sequencing read that could be allocated to a specific rDNA cluster, the distance from the 5′end of an 18S gene to the 5′ end of the next downstream 18S gene was calculated for every rDNA unit present on the read. Each histogram represents the unit length distribution for the specific rDNA cluster. In total, 519 rDNA units were identified which can be allocated as the following: 98, 123, 87, 51, 76 and 84 units belong to the rDNA clusters B_R, B_L, C_R, D_L, E_L and F_L, respectively. Indices L (left) and R (right) indicate the chromosome end. (**B**) Histogram of the length of all rDNA units independent of the cluster affiliation ($n = 5305$; this also includes all reads that only contain rDNA units and could not be mapped to a specific chromosome end) depicted on the *y*-axis and a cluster-specific box plot of the unit lengths depicted on the *x*-axis.

Although our improved protocol yielded ultra-long reads, we could not assemble telomeres for every rDNA cluster. While the overall longest read could be mapped to a central part of chromosome E, the read containing the most rDNA units was with 283 kb only 57% the size of the top read and did not have any telomeric repeats. However, to get an idea of the total amount rDNA copies, we performed ddPCR and the number of rDNA units was estimated to be 126 (Supplementary Figure S2).

Assessing diversity of rDNA clusters and their intergenic regions (IGR)

To investigate rDNA unit subtypes in *Y. lipolytica* DSM 3286, we first looked for short direct repeats present in the IGR of the rDNA clusters in the assembly and annotated PvuII restriction sites. Two prominent repeats were identified with one being 12 nt long and the other 87 nt. Independent from the chromosome, IGRs of all different rDNA units showed prevalence of both repeats whereas the 12 nt repeats clustered at the beginning of the IGR (repeat region 1, RPR1) and the 87 nt repeats could be found solely in the middle (repeat region 2, RPR2). Interestingly, while the short 12 nt repeat only had one sequence variant and is organized primarily as two sets of tandem repeats, the 87 nt repeat showed significant sequence variations and this region contained a significant number of partial repeats that together with the complete repeats combine for a tandem array. The PvuII sites are mostly scattered in RPR2 but can also be found between the repeat regions.

When looking at individual clusters, it became apparent that there are two distinct IGR sub-types present at three chromosome ends (Figure 3A). In the following, clusters, rDNA units and IGRs will be named after the associated chromosome where the indices L, R, 1 and 2 describe the cluster location and the subunit type (L = left; R = right; 1 = subtype 1; 2 = subtype 2). In general, the length differences of the IGRs can be explained as size variations of the two repeat regions, especially of RPR2, and the space in between (Supplementary Figure S3A). However, analysis of the different RPR2s emerged to be challenging since several initially detected repeats were partial and the truncation did not appear to be systematic. Therefore, we analyzed the IGRs with the sequence analysis tools MEME and FIMO (33). The most prominent discovered motif had a width of 41 nt and was found 188 times across all different IGR types whereas all motif copies were in RPR2. However, analysis of the sequence logo revealed a shorter but more conserved motif (motif 1) that was 24 nt in length (Supplementary Figure S3C) and a FIMO scan identified 166 occurrences of the motif corresponding to the amount of the previously discovered 87 nt repeats. Interestingly, a second MEME search with changed parameter settings revealed two more motifs (motifs 2 and 3) present in RPR2 that surrounded some of the main motifs which reflects the presence of complete and truncated repeats (Supplementary Figure S3B). This also explains why the distance between the main motifs varies between IGR types as well within the RPR2 itself. For example, both IGR subtypes of the cluster B_L have twelve occurrences of motif 1 but the RPR2 of IGR of B_{L1} is 138 bp longer and contains three additional copies of motif 3. The number of motifs and thereby 87 nt repeats ranged from twelve in the IGR of B_{L2} to 19 in the IGR of B_R.

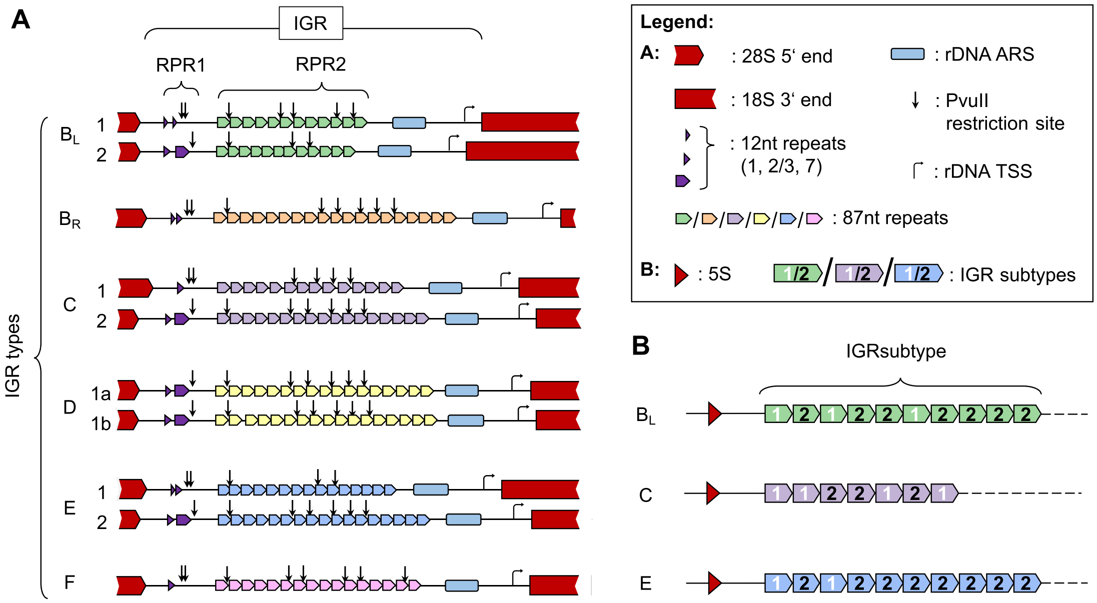

Figure 3. Sequence analysis of the different IGR types and subtypes of *Y. lipolytica* DSM 3286. A: Graphical representation of all identified IGRs. Originally, the IGR was identified as the region between the 28S rRNA gene and the subsequent downstream 18S rRNA gene. But it also carries a lot of functional features necessary for rDNA silencing (13) or replication, and therefore can be found upstream of every rDNA unit. In addition, it exhibits features that can be easily identified on a nucleotide level. Every IGR contains two repeat regions and a PvuII restriction site pattern. The combination of these sequence features that can be highly divergent is specific for each rDNA cluster and allows to distinguish between different sub-types within a cluster. The rARS, as well as its up- and downstream region are highly conserved among all sub-types. Note that the position of the repeats in RPR2 does not correlate with the actual position in the sequence but was simplified for the sake of clarity. B: Graphical representation of the order of IGR subtypes within the rDNA clusters B_L, E and C. The order was determined based on the final assembly.

The differences regarding the number of 12nt repeats and PvuII sites between the repeat regions are abundant and could be observed for all ends with two distinct rDNA unit sub types and for all types in general as every unit's IGR either starts with five or less 12 nt repeats followed by two adjacent PvuII sites (type A) or begins with ten 12 nt repeats and only one PvuII site (type B). But also based on the PvuII sites present in RPR2, one could divide the subtypes. It is noticeable that all subtypes of chromosome D as well as the subtypes 2 of chromosome C and E not only share the same RPR1 but also have the same PvuII site pattern in RPR2. A local sequence alignment via MUSCLE (Supplementary Figure S4) (34) revealed that these four subtypes are highly similar except for the region between the TSS and the rARS (38) that is different for each chromosome, and that the IGR of B_{L2} is a truncated version of this type.

While both chromosome D subtypes also share this TSS upstream region, subtype 2 has a slightly longer RPR2 than all other subtypes of the whole group. Interestingly, this marginally longer subunit is predominant in the rDNA cluster of chromosome D. Based on this similarity, we analyzed all remaining IGR subtypes (Supplementary Table S3) and identified a high pairwise sequence similarity, where the difference can be attributed mostly to gaps in the consensus sequence, between B_{L1} and F as well as B_R and E_1 (Supplementary Figures S5 and S6). The rARS and its upstream region are highly conserved among all sub-types. A 100 bp window between RPR1 and RPR2 exhibited a 100% pairwise identity in all subunits. Interestingly, alignment of a 100 bp long region downstream of the TSS of all units, that most likely contains the core promoter, only showed a pairwise identity of 82%. However, only six of them were unique: the core promoter region of units B_{L1}, D_1 and D_2, the units B_{L2} and B_R, as well as the units C_1 and C_2 were identical, respectively. Noteworthy, the group of C_1 and C_2 had the lowest pairwise identity to the consensus sequence. By determining the subtype of an rDNA unit within a cluster, we could also specify the order in which they appear in the cluster (Figure 3B). Remarkably, the clusters of all chromosomes that have two distinct subtypes start with the one that is characterized by a type A RPR1. While the rDNA cluster of chromosome C seems to have an equal distribution of IGR types, the chromosome B_L and E clusters appear to contain more subtypes with a type B RPR1.

Since the assignment of an rDNA unit in a cluster to one of the discovered subtypes was not always clear, especially for telomere proximal units, we wanted to confirm the subtypes and their order within the cluster on a single-read level. Based on RPR1, most of the reads that could be mapped onto the end of a chromosome allowed us to deduce the same order of units as given by the assembly. For some reads though, we observed deviations in the pattern. Particularly, cluster C showed a high variance at position seven and eight with 34% and 43% of the reads having a different subtype, respectively (Figure 4). Among the three clusters, E has the most consistent pattern while B_L was characterized by the smallest number of evaluable reads.

Analysis of telomeric and sub-telomeric regions

In contrast, the rDNA cluster-free chromosome ends could all be assembled completely. A comprehensive analysis of

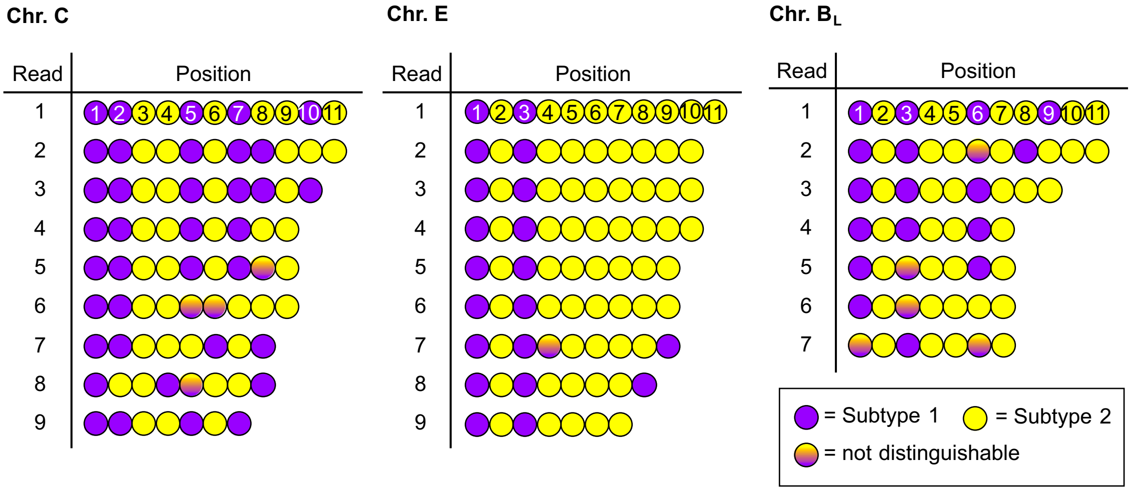

Figure 4. Graphical representation of the order of IGR subtypes in the rDNA clusters C, E and B_L of *Y. lipolytica* DSM 3286. For the clusters C and E, nine reads had a base quality sufficient for identification of subtypes while only seven could be used for cluster B_L. The colors of the circles indicate the subtype. The position of an rDNA unit within a cluster is stated for the first read where 1 corresponds to the first rDNA unit in the respective cluster.

the sub-telomeric regions of these chromosome ends revealed that the six regions share many similarities and although no feature itself is unique for a region, the different sub-telomeric regions when compared to each other still are (Figure 5 and Supplementary Figure S7). As for instance, sub-telomeric region A_L and F_R are almost identical feature-wise except for a 5S rRNA gene copy that can only be observed for A_L. Interestingly, all sub-telomeric regions include an approximately 8 kb long sequence that is also present in all sub-rDNA cluster regions. Additionally, IGR-like sequences can also be found in these regions. They harbour the same core features as the intergenic regions between the rDNA genes but cannot be allocated to any of the aforementioned types. Specifically, A_L, D_R and E_R are characterized by the longest RPR2 of all IGRs whereas A_R and C_L have 3′-truncated versions indicated by the missing rARS (not shown in the figure). Although F_R has similarities to both A_L as well as the rDNA IGR of chromosome F, it is still significantly different. Besides their unique characteristic, the sub-telomeric regions can be roughly divided into two groups depending on whether they encode for RecQ-like helicases (39). The IGR, independent if truncated or not, is always accompanied by the above mentioned 8 kb long sequence. The 8 kb region constitutes a genomic feature that can not only be observed for the chromosome ends described in this paragraph but also for those that comprise the rDNA clusters as already implied above and this feature connects the sub-telomeric regions with the unique genome part of the respective chromosome.

Comparative genomics for *Yarrowia lipolytica*

In accordance with the literature (8,17), the DSM 3286 genome was named YALI2 and a comparison with CLIB122 (YALI0) and CLIB89 (YALI1) revealed that they were similar in total genome length but differed in individual chromosome length (Supplementary Table S4). The structural variations between the genomes of YALI0, YALI1, and YALI2 were analyzed by aligning all three assemblies with Mauve (Supplementary Figure S8) (29). The alignment shows 14 locally co-linear blocks (LCB) and two additional regions which are, however, only a few hundred base pairs long (Supplementary Table S5). In the assembly of YALI2, most of these regions without a sequence conservation represent the rDNA clusters and telomeres. Interestingly, YALI1 exhibits a relatively large region in chromosome C from base 2 692 172 to 2 753 172 that cannot be aligned with the other genomes. However, the gap is part of an approximately 60 kb direct repeat making an alignment difficult. Except for a small 70 kb inversion at the beginning of chromosome B and the mentioned direct repeat, YALI0 and YALI1 share otherwise a structurally identical genome. In contrast, *Y. lipolytica* DSM 3286 has many differences. While chromosome A seems to be similar among the three strains, chromosome B of YALI2 appears to consist of the first 684 kb of chromosome B from YALI0 and YALI1 followed by the last 1.728 Mb of chromosome C. Consequently, the end of the YALI2 chromosome C corresponds to the end of chromosome B from the other two strains. A similar exchange can be observed for chromosomes E and F where the last 1.365 Mb of the YALI2 chromosome E align with the respective last bases of chromosome F of the other two strains. Finally, a 206 kb inversion can be observed in the middle of chromosome F of YALI2 compared to YALI0 and YALI1. To further investigate the chromosomal rearrangements between YALI2 and YALI0 as well as YALI1, the genomes were compared to the recently published draft genome of the *Y. lipolytica* strain H222 (40). The alignment (Supplementary Figures S9 and S10) confirmed the two reciprocal translocations and the large inversion described by the authors. Interestingly, we found indications that one of the two large reciprocal translocations (involving chromosomes C/B and E/F) observed between YALI2 and YALI0/YALI1 also at least partly happened for H222. While YALI2C conforms to H222s03, YALI2B corresponds to H222s06, s07, s10 and s12. However, it seems that the second translocation event did not occur in H222 since there are two more translo-

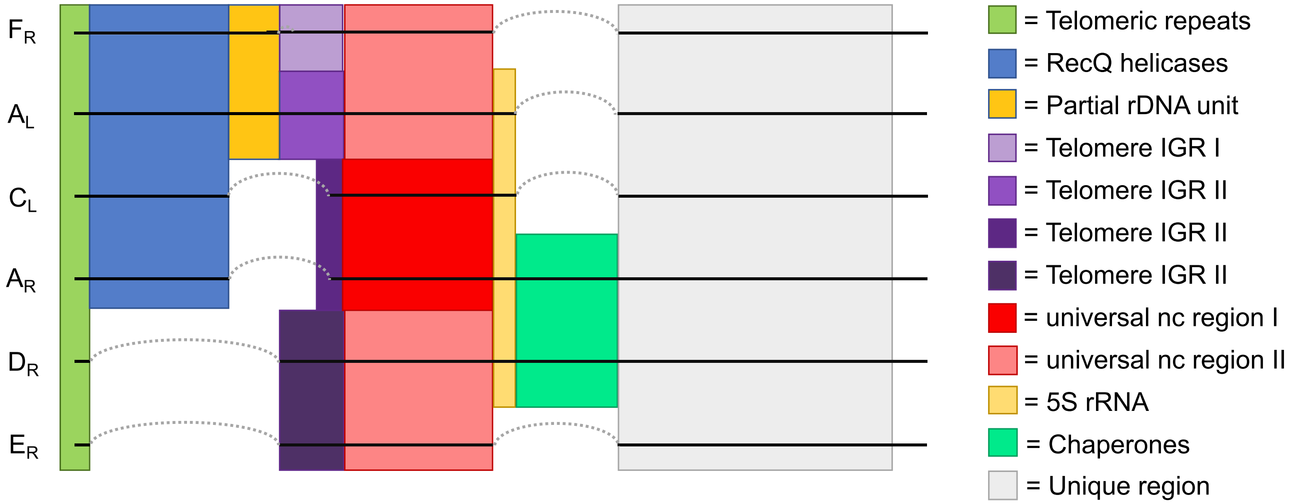

Figure 5. Graphical alignment of sub-telomeric features of *Y. lipolytica* DSM 3286. Each colored block represents a distinct sequence feature that can be observed for multiple sub-telomeric regions. Each region belongs to specific chromosome ends as indicated left (R = right; L = left) where the straight black line represents the DNA sequence of the respective end. Dotted lines appear when there is a break in the alignment meaning that the respective end is missing a feature or only has a truncated version. The sequence features were identified based on multiple iterative sequence alignments with MUSCLE (34) and mean sequence regions of homology. The telomeric repeats are identical for all ends and only differ in their number of annotated 10 nt repeats. RecQ helicases are highly conserved in prokaryotes and eukaryotes and are involved in maintaining the genomic and especially telomeric structure (39). The partial rDNA unit consists of the complete 18S and a 5′-fragment of the 28S gene. While all the aforementioned features have a high degree of pairwise identity across the various chromosomes (99.8–99.9%), the different colour shades in case of the telomere IGR and the universal nc region features describe different levels of similarity. As for the universal non-coding (nc) regions, C_L and A_R have a 99.4% pairwise identity while the remaining have a pairwise identity to each other in range from 89.2 to 94.3%. The differences are even higher for the telomeric IGR feature where D_R and E_R as well as C_L and A_R have a 100% pairwise identity while F_R and A_L are only 70.1% pairwise identical.

cations between YALI2E/YALI2F and H222s01/H222s02 while, except for the inversion, H222s01 and s02 correspond to chromosomes E and F of YALI1 and YALI2, respectively.

To compare YALI0, YALI1 and YALI2 also on a functional level, we uploaded the three genomes to the software platform EDGAR for comparative genomics (41). The core genome of the three *Yarrowia* strains consists of 6045 CDS with YALI0 as a reference genome. A phylogenetic tree (Supplementary Figure S11A) confirmed that YALI0 and YALI1 are closer related to each other than to YALI2. The more distant relationship of the DSM 3286 strain that had already been observed after comparison of the genome structures is therefore also reflected in the genome content. Accordingly, YALI0 and YALI1 share more genes with each other (204) than YALI2 with YALI0 (80) or YALI1 (132) (Supplementary Figure S11B). In addition to the core genes, EDGAR also provided a list of singletons. Altogether, 60 singletons for YALI0, 1377 singletons for YALI1, and 86 singletons for YALI2 have been calculated. While the list of singletons for YALI0 mainly contained hypothetical proteins and transposable elements, two singletons from YALI0 were annotated as 60S ribosomal proteins. In contrast, the singleton list for YALI2 showed a larger variety comprising genes encoding an acidic extracellular protease, an argininosuccinate synthase, among others. Most interestingly, features such as five pif-like ATP-dependent DNA helicase genes are not found in the assemblies of the two other strains. In the YALI2 assembly, these genes are located in the sub-telomeric regions. Compared to the previous assembly of DSM 3286, the genome sequence resulting from the hybrid approach contained 11 more CDS than the Illumina-only assembly emphasizing the advantage of long sequencing reads combined with short reads.

DISCUSSION

Ultra-long nanopore reads enable near contiguous assemblies of complex yeast genomes and constitute a new gold standard

In this paper, we present a nearly contiguous genome assembly of the yeast *Yarrowia lipolytica* DSM 3286 that for the first time includes the assembly of the highly repetitive rDNA clusters and telomeric regions. By applying our improved protocol for long read nanopore sequencing that merges the DNA isolation and sequencing library preparation protocols, we could generate ultra-long sequencing reads and thereby increase the size of the genome assembly of *Y. lipolytica* DSM 3286 by 1.0 Mb compared to an assembly that is solely based on short reads. The additional assembled base pairs were identified as extensions of the chromosome ends and constitute the rDNA clusters, telomeres and sub-telomeric regions that were missing from previous assemblies (8,17,42). While it was reported that *Y. lipolytica* has four chromosomes that contain terminal rDNA cluster (9,10,17,43) and shown for YALI1 that each cluster belongs to a separate chromosome (17), we could demonstrate that YALI2 has six clusters connected to five chromosomes with one carrying two clusters. Besides this structural genome variation, the MAUVE (29) alignments revealed 14 large locally co-linear blocks (LCB) between YALI2 and YALI0/YALI1, including two large reciprocal translocations. Furthermore, we showed that strain H222 shares one of the large translocations. These chromosome rearrangements could have been generated by crossing-over events mediated by the various classes of transposable elements present in *Y. lipolytica* (17) and indicate that *Y. lipolytica* H222 could be a connecting link between the evolutionary separation of YALI2 from YALI0 and YALI1. However, more analyses need to be conducted to prove these findings.

Although we could assemble all six rDNA clusters into the genomic sequence, only two of them are complete and finish in telomeric repeats (the clusters of chromosome C and D). In order to completely assemble an rDNA cluster with sequencing data obtained from the current nanopore technology, a read is required that starts with a unique sequence which would allow to allocate the read to a specific chromosome end, terminate in telomeric repeats, and thereby span the whole cluster. Considering that the longest reported cluster in YALI0 is 620 kb long (9), it is highly possible that the longest cluster of YALI2 is of similar size. Consequently, a maximum read length of 493 kb obtained in this study would not be enough to enable a complete assembly of all rDNA clusters. Since integration of all clusters into the genomic sequence is hence depending on the maximum read length, getting a successful assembly is not only a technical matter but also a stochastic one. Additionally, the observed heterogeneity in terms of the number of rDNA units within a cluster requires a manual refinement since heterogenic reads cannot be resolved with an automated assembly. However, we believe that with coming improvements of the nanopore technology, that will lead to longer reads but especially to a higher base calling accuracy, these obstacles can be overcome.

Looking beyond the contig: analyzing genomic regions on a single nanopore read level delivers single cell sequencing information

Ultra-long sequencing reads do not only improve the continuity and completeness of a genome assembly but also reveal new information on a single-read level that can describe the status of a single cell or a single molecule. While the principle of single sequencing read analyses is already used to study population structures via single nucleotide variants (44) or to assess the adaptive immune receptor repertoire (AIRR) where reads are used to identify unique antibody gene segments of individual B cells (45), analyses are limited to the length of the read in context of the provided single cell information. However, with the help of ultra-long reads, it is possible to extend this information content.

For instance, due to the high copy number of mitochondria per cell, the size of a single copy of the mitochondrial genome can be deduced from the read length distribution (Supplementary Figure S1A). In our case, the very narrow peak at 48 kb confirmed the size of the reference mitochondrial genome. Additionally, we could identify many apparently circular multimers of the mitochondrial genome with some of them being longer than the equivalent of three copies of the genome. In general, the landscape of mitochondrial genome architecture in yeast is quite complex and can differ even between closely related organisms (37). Mitochondrial DNA (mtDNA) in yeast is mostly present as linear molecules and only to small amounts as circular DNA (46). However, the linear variants can either be found as linear-mapping monomers, that terminate in specific telomeric sequences, or as polydisperse molecules that are circular-mapping. Additionally, multipartite linear forms have also been reported (37). Consequently, it is quite challenging to fully enlighten the mechanism of mitochondrial DNA replication in yeast. In a recent study, it was suggested that mtDNA in yeast is replicated by the interplay of recombination, rolling circle replication and template switching (47). If in this case the template-switching would fail, the synthesis of concatemers would take place, leading to the reads observed in our sequencing data. However, the proposed replication mechanism also means that an observed concatemer can result from sequencing of either circular multimers or linear polydisperse molecules. On closer examination of the read lengths (Figure 1), it is apparent that there are two size thresholds at multiples of the size of the reference genome. This indicates the presence of two distinct forms. In addition, it is very likely that each molecule was cleaved by just one transposome complex and that the two forms therefore depict the monomeric as well as the dimeric circular form of the YALI2 mitochondria. This is additionally supported by our finding that the reads at the size thresholds strongly converge to multiples of full-length mtDNA genomes. In contrast to circular DNA, fragmentation of polydisperse linear DNA by a single transposome complex would lead to two fragments instead of one and a size threshold would probably not be observable at multiples of the monomer. Interestingly, the amount of reads larger than the monomer was unexpectedly high. More than 43% of reads were longer than the monomer indicating a low efficiency of the template switching during replication. Naturally, a detailed analysis of our data also revealed that none of the reads contained telomeric or mobile elements and all sequences could be mapped onto a circular monomer of the YALI2 mtDNA thereby confirming its circular-mapping nature (48).

Ultra-long reads provide insights into rDNA cluster replication and heterogeneity

Besides the insights given into the architecture of the mitochondrial genome in YALI2, the ultra-long reads also shed light on the spectrum of intergenic regions in rDNA clusters of this yeast. With the aid of the long sequencing reads that contained a multitude of rDNA units, we were able to show the presence of rDNA unit subtypes in specific clusters. For different strains of *Y. lipolytica*, it was reported that they contain up to five types of rDNA units and it was suggested that each type groups into one or more distinct clusters (11). Here, we showed that YALI2 has ten different IGR types in total, divided among six cluster of which two have only one type while the other four clusters exhibit two types each. One possible explanation for the appearance of two different IGRs in an rDNA cluster could be that YALI2 is diploid and that the observed IGR subtypes are a form of heterozygosity. However, an analysis of genes and DNA sequences relevant for mating (data not shown) did not reveal any idiomorphs indicating that our strain is haploid. Compared to other yeast species and to the previous findings for *Y. lipoly-*

tica, the number of different IGR types and thereby rDNA types is significantly higher in YALI2 (35). However, while the reported number of rDNA types usually does not exceed five, it remains unclear if the actual number of rDNA types in the reported yeast species is higher and it would be interesting to see in future studies which new insights could be obtained if our analysis strategy would be applied to other yeast species.

The biggest length variability of the IGRs of YALI2 was observed in a repetitive region that contains multiple sites of a specific 24 nt sequence motif (Supplementary Figure S2B). Naturally, variations in specific sequence segments of the IGR have a chance to affect one of the many regulatory elements that are important for transcription or replication. However, no repetitive elements have been found in the well-studied IGR of *S. cerevisiae* that would explain the biological function of the significant variations in the IGRs of *Y. lipolytica* (13). However, in contrast to the Baker's yeast, metazoans have much larger IGRs and additionally, contain repetitive, regulatory elements. Typically, enhancers in higher eukaryotes are repetitive elements that share sequence or functional similarities with promoter domains and increase the amount of stable promoter complexes by binding activating transcription factors and thereby enhancing its activity (49). In addition, metazoan IGRs also contain a spacer promoter which is absent in *S. cerevisiae* (12), and phylogenetic analysis suggested that enhancers evolved from the spacer promoter by repeating duplication and truncation events. In the frog *Xenopus laevis*, spacer promoter duplications of two sizes (60/81 bp) were found that each share a 42 bp region with high sequence similarity to an upstream domain of the promoter and it was shown that these elements enhance rDNA transcription (49). We found that in YALI2 the sequences that separate the 24 nt motifs could be divided into four size classes (∼61, ∼81, ∼109 and ∼130 bp; data not shown). However, no sequence similarity to the core promoter or a sequence that could represent a spacer promoter could be identified in YALI2. Nevertheless, we will not exclude the possibility that the identified 24 nt motif could constitute a binding site for a transcription factor and hence function as an enhancer for the transcription of rRNA genes in *Y. lipolytica*, especially since enhancers in mouse also lack sequence similarity with their promoters (50). Still, the question remains why *Y. lipolytica* has such a high variance in the number of motifs in RPR2 and in the sequence upstream of the TSS (data not shown) but it is likely that the heterogenous IGRs influence rDNA expression differently (35).

Another advantage of the ultra-long sequencing reads for assembling the rDNA clusters was that they allowed to determine the order of the IGR subtypes within a cluster even on a single-read level where each read represents a different cell. While all rDNA clusters exhibit different IGRs, they also share common features regarding their subunit order. For instance, all clusters, that showed significantly different subunits, start with rDNA units that are characterized by a type A RPR1 while type B RPR1 units are the most frequently occurring subunits in a cluster (Figure 4). However, we observed that not all sequencing reads of a cluster have the same rDNA unit order or even contain the same number of units. Besides the differences regarding the extend of possible transcription factor binding sites in RPR2, this adds another level of heterogeneity to the rDNA clusters of YALI2. However, this heterogeneity is not surprising since rDNA clusters are relatively unstable and a subject for recombination (13). In *S. cerevisiae*, a key factor for homologous recombination is the fork-blocking protein Fob1 that activates the RFB (15). A stalled replication at an RFB is susceptible for double-strand breaks that, in case of an unequal sister chromatid exchange (USCE), lead to an extension or contraction of the rDNA cluster (51). Variations in the number of units in a cluster from molecule to molecule can thus be explained by the USCE. While we could observe variations in the copy number of the rDNA repeats per chromosome on a single-read level, the ultra-long reads do not provide information on the total count per cell and single cell data are limited to individual chromosomes. It is possible that in *Y. lipolytica* RPR1 contains the RFBs and that the efficiency for replication fork stalling is different between type A and type B RPR1s. Since RPR1 is also present in ever sub-telomeric region, homologous recombination could also provide an explanation for the subtelomeric heterogeneity. Unfortunately, we could not yet identify an ortholog of Fob1 in YALI2 to further investigate possible RFB sites. This will be subject for future studies.

The data that we presented emphasize the importance of ultra-long sequencing reads and we encourage to look behind the contig in genome assemblies more often since single reads can provide information regarding the status of single cells or single molecules. In the case of *Y. lipolytica* DSM 3286 we were able to demonstrate with single ultra-long reads an interesting mitochondrial architecture as well as to show the vast heterogeneity of the rDNA clusters, once again highlighting the closeness of this yeast to metazoans.

DATA AVAILABILITY

The annotated genome sequence as well as the raw sequence reads have been deposited with GenBank and the Sequence Read Archive under accession number PRJNA641784. Command lines that were used for the hybrid genome assembly and genome analyses have been uploaded to GitHub (https://github.com/tluttermann/YALI2_NARGAB_2020).

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

FUNDING

European Regional Development Fund (EFRE) [34.EFRE-0300095/1703FI04 to T.B., J.K.]. Funding for open access charge: Open Access Publication Fund of Bielefeld University.

Conflict of interest statement. None declared.

REFERENCES

1. Nicaud,J.-M. (2012) Yarrowia lipolytica. *Yeast*, **29**, 409–418.
2. Ruiz-Herrera,J. and Sentandreu,R. (2002) Different effectors of dimorphism in yarrowia lipolytica. *Arch. Microbiol.*, **178**, 477–483.
3. Boonvitthya,N., Bozonnet,S., Burapatana,V., O'Donohue,M.J. and Chulalaksananukul,W. (2013) Comparison of the heterologous

expression of trichoderma reesei endoglucanase II and cellobiohydrolase II in the yeasts pichia pastoris and yarrowia lipolytica. *Mol. Biotechnol.*, **54**, 158–169.

4. Gasmi,N., Fudalej,F., Kallel,H. and Nicaud,J.-M. (2011) A molecular approach to optimize hIFN α2b expression and secretion in yarrowia lipolytica. *Appl. Microbiol. Biotechnol.*, **89**, 109–119.
5. Schwarzhans,J.-P., Luttermann,T., Geier,M., Kalinowski,J. and Friehs,K. (2017) Towards systems metabolic engineering in pichia pastoris. *Biotechnol. Adv.*, **35**, 681–710.
6. Levinson,W.E., Kurtzman,C.P. and Kuo,T.M. (2007) Characterization of yarrowia lipolytica and related species for citric acid production from glycerol. *Enzyme Microb. Technol.*, **41**, 292–295.
7. Anastassiadis,S., Aivasidis,A. and Wandrey,C. (2002) Citric acid production by candida strains under intracellular nitrogen limitation. *Appl. Microbiol. Biotechnol.*, **60**, 81–87.
8. Dujon,B., Sherman,D., Fischer,G., Durrens,P., Casaregola,S., Lafontaine,I., Montigny,J. de, Marck,C., Neuvéglise,C., Talla,E. *et al.* (2004) Genome evolution in yeasts. *Nature*, **430**, 35–44.
9. Casarégola,S., Feynerol,C., Diez,M., Fournier,P. and Gaillardin,C. (1997) Genomic organization of the yeast yarrowia lipolytica. *Chromosoma*, **106**, 380–390.
10. van Heerikhuizen,H., Ykema,A., Klootwijk,J., Gaillardin,C., Ballas,C. and Fournier,P. (1985) Heterogeneity in the ribosomal RNA genes of the yeast yarrowia lipolytica; cloning and analysis of two size classes of repeats. *Gene*, **39**, 213–222.
11. Fournier,P., Gaillardin,C., Persuy,M.-A., Klootwijk,J. and van Heerikhuizen,H. (1986) Heterogeneity in the ribosomal family of the yeast yarrowia lipolytica: genomic organization and segregation studies. *Gene*, **42**, 273–282.
12. Paule,M.R. and White,R.J. (2000) Survey and summary: transcription by RNA polymerases i and III. *Nucleic Acids Res.*, **28**, 1283–1298.
13. Srivastava,R., Srivastava,R. and Ahn,S.H. (2016) The epigenetic pathways to ribosomal DNA silencing. *Microbiol. Mol. Biol. Rev.*, **80**, 545–563.
14. Muller,M., Lucchini,R. and Sogo,J.M. (2000) Replication of yeast rDNA initiates downstream of transcriptionally active genes. *Mol. Cell*, **5**, 767–777.
15. Kobayashi,T. (2003) The replication fork barrier site forms a unique structure with fob1p and inhibits the replication fork. *Mol. Cell. Biol.*, **23**, 9178–9188.
16. Schutter,K., Lin,Y.-C., Tiels,P., van Hecke,A., Glinka,S., Weber-Lehmann,J., Rouzé,P., van de Peer,Y. and Callewaert,N. (2009) Genome sequence of the recombinant protein production host pichia pastoris. *Nat. Biotechnol.*, **27**, 561–566.
17. Magnan,C., Yu,J., Chang,I., Jahn,E., Kanomata,Y., Wu,J., Zeller,M., Oakes,M., Baldi,P. and Sandmeyer,S. (2016) Sequence assembly of yarrowia lipolytica strain W29/CLIB89 shows transposable element diversity. *PLoS One*, **11**, e0162363.
18. Koren,S., Walenz,B.P., Berlin,K., Miller,J.R., Bergman,N.H. and Phillippy,A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, **27**, 722–736.
19. Loman,N.J., Quick,J. and Simpson,J.T. (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods*, **12**, 733–735.
20. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
21. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
22. Walker,B.J., Abeel,T., Shea,T., Priest,M., Abouelliel,A., Sakthikumar,S., Cuomo,C.A., Zeng,Q., Wortman,J., Young,S.K. *et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9**, e112963.
23. Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
24. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with bowtie 2. *Nat. Methods*, **9**, 357–359.
25. Wibberg,D., Rupp,O., Blom,J., Jelonek,L., Kröber,M., Verwaaijen,B., Goesmann,A., Albaum,S., Grosch,R., Pühler,A. *et al.* (2015) Development of a rhizoctonia solani AG1-IB specific gene model enables comparative genome analyses between phytopathogenic r. solani AG1-IA, AG1-IB, AG3 and AG8 isolates. *PLoS One*, **10**, e0144769.
26. Wibberg,D., Andersson,L., Tzelepis,G., Rupp,O., Blom,J., Jelonek,L., Pühler,A., Fogelqvist,J., Varrelmann,M., Schlüter,A. *et al.* (2016) Genome analysis of the sugar beet pathogen rhizoctonia solani AG2-2IIIB revealed high numbers in secreted proteins and cell wall degrading enzymes. *BMC Genomics*, **17**, 245.
27. Stanke,M., Steinkamp,R., Waack,S. and Morgenstern,B. (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.*, **32**, W309–W312.
28. Rupp,O., Becker,J., Brinkrolf,K., Timmermann,C., Borth,N., Pühler,A., Noll,T. and Goesmann,A. (2014) Construction of a public CHO cell line transcript database using versatile bioinformatics analysis pipelines. *PLoS One*, **9**, e85568.
29. Darling,A.C.E., Mau,B., Blattner,F.R. and Perna,N.T. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.*, **14**, 1394–1403.
30. Hilker,R., Stadermann,K.B., Schwengers,O., Anisiforov,E., Jaenicke,S., Weisshaar,B., Zimmermann,T. and Goesmann,A. (2016) ReadXplorer 2-detailed read mapping analysis and visualization from one single source. *Bioinformatics*, **32**, 3702–3708.
31. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
32. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
33. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
34. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
35. Torres-Machorro,A.L., Hernández,R., Cevallos,A.M. and López-Villaseñor,I. (2010) Ribosomal RNA genes in eukaryotic microorganisms: witnesses of phylogeny? *FEMS Microbiol. Rev.*, **34**, 59–86.
36. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.-J., Chen,Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
37. Valach,M., Farkas,Z., Fricova,D., Kovac,J., Brejova,B., Vinar,T., Pfeiffer,I., Kucsera,J., Tomaska,L., Lang,B.F. *et al.* (2011) Evolution of linear chromosomes and multipartite genomes in yeast mitochondria. *Nucleic Acids Res.*, **39**, 4202–4219.
38. Vernis,L., Chasles,M., Pasero,P., Lepingle,A., Gaillardin,C. and Fournier,P. (1999) Short DNA fragments without sequence similarity are initiation sites for replication in the chromosome of the yeast yarrowia lipolytica. *Mol. Biol. Cell*, **10**, 757–769.
39. Bernstein,K.A., Gangloff,S. and Rothstein,R. (2010) The RecQ DNA helicases in DNA repair. *Annu. Rev. Genet.*, **44**, 393–417.
40. Devillers,H. and Neuvéglise,C. (2019) Genome sequence of the oleaginous yeast yarrowia lipolytica H222. *Microbiol. Res. Announ.*, **8**, 1–2.
41. Blom,J., Kreis,J., Spänig,S., Juhre,T., Bertelli,C., Ernst,C. and Goesmann,A. (2016) EDGAR 2.0: an enhanced software platform for comparative gene content analyses. *Nucleic Acids Res.*, **44**, W22–W28.
42. Walker,C., Ryu,S., Haridas,S., Na,H., Zane,M., LaButti,K., Barry,K., Grigoriev,I.V. and Trinh,C.T. (2020) Draft genome assemblies of ionic liquid-resistant yarrowia lipolytica PO1f and its superior evolved strain, ylcw001. *Microbiol. Res. Announ.*, **9**, e01356–19.
43. Gaillardin,C., Mekouar,M. and Neuvéglise,C. (2013) Comparative genomics of yarrowia lipolytica. In: Barth,G. (ed.) *Yarrowia Lipolytica*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–30.
44. Roth,A., Khattra,J., Yap,D., Wan,A., Laks,E., Biele,J., Ha,G., Aparicio,S., Bouchard-Côté,A. and Shah,S.P. (2014) PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*, **11**, 396–398.
45. Mroczek,E.S., Ippolito,G.C., Rogosch,T., Hoi,K.H., Hwangpo,T.A., Brand,M.G., Zhuang,Y., Liu,C.R., Schneider,D.A., Zemlin,M. *et al.* (2014) Differences in the composition of the human antibody repertoire by b cell subsets in the blood. *Front. Immunol.*, **5**, 96.

46. Malina,C., Larsson,C. and Nielsen,J. (2018) Yeast mitochondria: an overview of mitochondrial biology and the potential of mitochondrial systems biology. *FEMS Yeast Res.*, **18**, 1–17.
47. Chen,X.J. and Clark-Walker,G.D. (2018) Unveiling the mystery of mitochondrial DNA replication in yeasts. *Mitochondrion*, **38**, 17–22.
48. Gaillardin,C., Neuvéglise,C., Kerscher,S. and Nicaud,J.-M. (2012) Mitochondrial genomes of yeasts of the yarrowia clade. *FEMS Yeast Res.*, **12**, 317–331.
49. Pikaard,C.S. (1994) Ribosomal gene promoter domains can function as artificial enhancers of RNA polymerase i transcription, supporting a promoter origin for natural enhancers in xenopus. *Proc. Natl. Acad. Sci. USA*, **91**, 464–468.
50. Grozdanov,P., Georgiev,O. and Karagyozov,L. (2003) Complete sequence of the 45-kb mouse ribosomal DNA repeat: analysis of the intergenic spacer☆. *Genomics*, **82**, 637–643.
51. Johzuka,K. and Horiuchi,T. (2002) Replication fork block protein, fob1, acts as an rDNA region specific recombinator in s. cerevisiae. *Genes to Cells*, **7**, 99–113.