Full-length Genome of the *Ogataea polymorpha* strain HU-11 reveals large duplicated segments in subtelomeic regions

Jia Chang[1$], Jinlong Bei[23$], Hemu Wang[4], Jun Yang[4], Xin Li[1]

Tung On Yau[5], Wenjun Bu[1], Jishou Ruan[6], Guangyou Duan[7*], Shan Gao[1*]

1. College of Life Sciences, Nankai University, Tianjin, Tianjin 300071, P.R.China.
2. Guangdong Provincial Key Laboratory for Crop Germplasm Resources Preservation and Utilization, Agro-Biological Gene Research Center, Guangdong Academy of Agricultural Sciences, Guangzhou, Guangdong 510640, P. R. China.
3. Guangdong Laboratory for Lingnan Modern Agriculture, Guangzhou, Guangdong 510642, P. R. China.
4. Tianjin Hemu Health Biotechnological Co., Ltd, Tianjin, Tianjin 300384, P.R.China
5. John Van Geest Cancer Research Centre, School of Science and Technology, Nottingham Trent University, Nottingham, NG11 8NS, United Kingdom;
6. School of Mathematical Sciences, Nankai University, Tianjin, Tianjin 300071, P.R.China
7. School of Life Sciences, Qilu Normal University, Jinan, Shandong 250200, P.R.China.

[$] These authors contributed equally to this paper.

* The corresponding authors.

SG：gao_shan@mail.nankai.edu.cn

GD：guangyou.duan@qlnu.edu.cn

Abstract

Background: Currently, methylotrophic yeasts (e.g., *Pichia pastoris*, *Hansenula polymorpha*, and *Candida boindii*) are subjects of intense genomics studies in basic research and industrial applications. In the genus *Ogataea*, most research is focused on three basic *O. polymorpha* strains—CBS4732, NCYC495, and DL-1. However, these three strains are of independent origin and unclear relationship. As a high-yield engineered *O. polymorpha* strain, HU-11 can be regarded as identical to CBS4732, because the only difference between them is a 5-bp insertion.

Results: In the present study, we have assembled the full-length genome of *O. polymorpha* HU-11 using high-depth PacBio and Illumina data. Long terminal repeat (LTR) retrotransposons, rDNA, 5' and 3' telomeric, subtelomeric, low complexity and other repeat regions were curated to improve the genome quality. We took advantage of the full-length HU-11 genome sequence for the genome annotation and comparison. Particularly, we determined the exact location of the rDNA genes and LTR retrotransposons in seven chromosomes and detected large duplicated segments in the subtelomeic regions. **Three novel findings are**: (1) the *O.* polymorpha NCYC495 is so phylogenetically similar to HU11 that a nearly 100% of their genomes is covered by their syntenic regions, while NCYC495 is significantly distinct from DL-1; (2) large segment duplication in subtelomeic regions is the main reason for genome expansion in yeasts; and (3) the duplicated segments in subtelomeric regions may be integrated at telomeric tandem repeats (TRs) through a molecular mechanism, which can be used to develop a simple and highly efficient genome editing system to integrate or cleave large segments at telomeric TRs.

Conclusions: Our findings provide new opportunities for in-depth understanding of genome evolution in methylotrophic yeasts and lay the foundations for the industrial applications of *O. polymorpha* CBS4732 and HU11. The full-length genome of the *O. polymorpha* strain HU-11 should be included into the NCBI RefSeq database for future studies of *O. polymorpha* CBS4732 and its derivatives LR9, RB11 and HU-11.

Keywords: Methylotrophic yeast; Ogataea; CBS4732; rDNA quadruple; retrotransposon

Introduction

Currently, methylotrophic yeasts (e.g., *Pichia pastoris*, *Hansenula polymorpha*, and *Candida boindii*) are subjects of intense genomics studies in basic research and industrial applications. However, genomic research on *Ogataea (Hansenula) polymorpha* trails behind that on *P. pastoris* [1], given that they both are popular and widely used species of methylotrophic yeasts. In the genus *Ogataea*, most research is focused on three basic *O. polymorpha* strains—CBS4732 (synonymous to NRRL-Y-5445 or ATCC34438), NCYC495 (synonymous to NRRL-Y-1798, ATCC14754, or CBS1976), and DL-1 (synonymous to NRRL-Y-7560 or ATCC26012). These three strains are of independent origin: CBS4732 was originally isolated from soil irrigated with waste water from a distillery in Pernambuco, Brazil in 1959 [2]; NCYC495 is identical to a strain first isolated from spoiled concentrated orange juice in Florida and initially designated as *Hansenula angusta* by Wickerham in 1951 [3]; and DL-1 was isolated from soil by Levine and Cooney in 1973 [4]. Although it is known that CBS4732 and NCYC495 can be mated while DL-1 cannot be mated with the either one of them, the relationship of all three strains is remains unclear [5]. CBS4732 and its derivatives—LR9, RB11, and HU-11—have been developed as genetically engineered strains to produce many heterologous proteins, including enzymes (e.g., feed additive phytase), anticoagulants (e.g., hirudin and saratin), and an efficient vaccine against hepatitis B infection [5]. As a nutritionally deficient mutant derived from CBS4732, the *O. polymorpha* strain HU-11 [6] is being used for high-yield production of several important protein or peptides, particularly including recombinant hepatitis B surface antigen (HBsAg) vaccine [7] and hirudin [8]. CBS4732 and HU-11 can be regarded as identical strains, as the only difference between them is a 5-bp insertion caused by frame-shift mutation of its *URA3* gene, which encodes orotidine 5'-phosphate decarboxylase.

To facilitate genomic research of yeasts, genome sequences have been increasingly submitted to the Genome- NCBI datasets. Among the genomes of 34 species in the *Ogataea* or *Candida* genus (**Supplementary file 1**), those of NCYC495 and DL-1 have been assembled at the chromosome level. However, the other genomes have been assembled at the contig or scaffold level. Furthermore, the genome sequence of CBS4732 was not available in the Genome- NCBI datasets until this manuscript was drafted. Among the genomes of 33 *Komagataella (Pichia)* spp., the genome of the *P. pastoris* strain GS115 is the only genome assembled at the chromosome level. The main problem of these *Ogataea*, *Candida*, or *Pichia* genomes is their incomplete sequences and poor annotations. For example, the rDNA sequence (GenBank: FN392325) of *P. pastoris* GS115 cannot be well aligned to its genome (Genbank assembly: GCA_001708105). Most genome sequences do not contain complete subtelomeric regions and, as a result, subtelomeres are often overlooked in comparative genomics [9]. For example, the genome of DL-1 has been analyzed for better understanding the phylogenetics and molecular basis of *O. polymorpha* [1]; however, it does not contain complete subtelomeric regions due to assembly using short sequences. Another problem of current yeast genome data is that the complete sequences of mitochondrial genomes is not simultaneously

released with those of nuclear genomes. The only complete mitochondrial genome in the NCBI GenBank database is the *O. polymorpha* DL-1 mitochondrial genome (RefSeq: NC_014805). More high-quality complete genome sequences of *Ogataea* spp. need to be sequenced to bridge the gap in *Ogataea* basic research and industrial applications.

In the present study, we have assembled the full-length genome of *O. polymorpha* HU-11 using high-depth PacBio and Illumina data and conducted the annotation and analysis to achieve the following research goals: (1) to determine the relationship between CBS4732/HU-11, NCYC495, and DL-1; (2) to provide a high-quality and well-curated reference genome for future studies of *O. polymorpha* CBS4732 and its derivatives- LR9, RB11, and HU-11; and (3) to discover important genomic features (e.g., high yield) of *Ogataea* spp. for basic research (e.g., synthetic biology) and industrial applications.

Results and Discussion

Genome sequencing, assembly and annotation

One 500 bp and one 10 Kbp DNA library each were prepared using fresh cells of *O. polymorpha* HU-11 and sequenced on the Illumina HiSeq X Ten and PacBio Sequel platforms, respectively, to obtain the high-quality genome sequences. Firstly, 18,319,084,791 bp cleaned PacBio DNA-seq data were used to assembled the complete genome, except for the rDNA region (**analyzed in further detail in subsequent sections**), with an extremely high depth of ~1800X. However, the assembled genome using high-depth PacBio data still contained two types of error in the low complexity (**Figure 1A**) and the short tandem repeat (STR) regions (**Figure 1B**). Then, 6,628,480,424 bp cleaned Illumina DNA-seq data were used to polish the complete genome of HU-11 to remove the two types of error. However, Illumina DNA-seq data contained errors in the long (>10 copy numbers) poly(GC) regions. Following this, the poly(GC) regions, polished using Illumina DNA-seq data, were curated using PacBio subreads (**Figure 1C**). Finally, Long Terminal Repeat (LTR) retrotransposons, rDNA (**analysed in more details in following sections**), 5' and 3' telomeric, subtelomeric, low complexity, and other repeat regions were curated to obtain the full-length genome using 103,345 long (> 20 Kbp) PacBio subreads (**Supplementary file 1**).

Ogataea polymorpha HU-11 has a nuclear genome (**Figure 2A**) with a summed sequence length of 9.1 Mbp and a mitochondrial (mt) genome (**Figure 2B**) with a sequence length of 59,496 bp (**Table 1**). For the data submission to the GenBank database, the sequence of circular mt genome was linearized, starting at the first codon of the ORF3 coding sequence (CDS). Analysis of long PacBio subreads revealed that the telomeric regions at 5' and 3' ends of each chromosome consist of tendam repeats (TRs) $[ACCCCGCC]_n$ and $[GGCGGGGT]_n$ (n is the copy number) with average lengths of 166 bp and 168 bp (~20 copy numbers), respectively. As these TRs vary in lengths, the 5' and 3' telomeric regions were not included into the seven linear chromosomes of HU-11, which were named as 1 to 7 from the smallest to the largest, respectively (**Table 1**). The full-length *O. polymorpha* HU-11 genome includes the complete sequences of all seven

chromosomes, while the 5' and 3' ends of NCYC495 or DL-1 chromosomes have many errors (**Supplementary file 1**). Therefore, we recommend the inclusion of our genome sequences into the NCBI RefSeq database to facilitate future studies on *O. polymorpha* CBS4732 and its derivatives- LR9, RB11, and HU-11.

The 9.1 Mbp length of the HU-11 genome is close to the estimated length of the *O. polymorpha* DL-1 genome [1], while the published NCYC495 and DL-1 genomes (**Table 1**) have shorter lengths of 8.97 and 8.87 Mbp, respectively (**Table 1**). The GC contents of the HU-11, NCYC495, and DL-1 genomes are comparable (~48%). Taking advantage of the full-length HU-11 genome sequence for the genome annotation and comparison, we determined the exact location of the rDNA genes and LTR retrotransposons in seven chromosomes (**Figure 2A**) and detected large duplicated segments in the subtelomeic regions (**described in more detail in succeeding sections**). Syntenic comparison (**Methods and Materials**) revealed that NCYC495 is so phylogenetically similar to HU11 that a nearly 100% of their genomes is covered by their syntenic regions; whereas NCYC495 is significantly distinct from DL-1. Using the full-length HU-11 genome, we corrected the NCYC495 genome sequence (**described in more detail further**) and annotations (**Methods and Materials**). Finally, we estimated that both HU-11 and NCYC495 encode 5,145 protein-coding genes (**Table 1**). In addition, HU-11, NCYC495, and DL-1 encode 80 identical tRNA genes.

Organization of rDNA genes

An rDNA TR of HU-11, NCYC495, or DL-1 encodes 5S, 18S, 5.8S, and 25S rRNAs (named as quadruple in the present study), with a length of 8,145 bp (**Supplementary file 1**). The copy number of rDNA TRs was estimated as 20 in the HU-11 genome (**Figure 3A**), while that was estimated as 6 and 25 in NCYC495 and DL-1, respectively [1]. TRs of HU-11 and NCYC495 rDNAs share a very high nucleotide (nt) sequence identity of 99.5% (8,115/8,152), while those of HU-11 and DL-1 rDNAs share a comparatively low nt sequence identity of 97% (7,530/7,765). As the largest TR region (~162 Kbp) in the HU-11 genome, the only rDNA locus is located in chromosome 2 and the organization of rDNA genes with different copy numbers may be conserved in the *Ogataea* genus. A TR of rDNAs in *Saccharomyces cerevisiae* also contains 5S, 18S, 5.8S, and 25S rDNAs as a quadruple, repeating two times on the chromosome 7 of its genome (**Figure 3B**). Four other 5S rDNAs are located separately away from the rDNA quadruples in *S. cerevisiae*. Compared to *O. polymorpha* or *S. cerevisiae* with only one rDNA locus, *Pichia pastoris* GS115 carries several rDNA loci, which are interspersed in three of its four chromosomes. Since the genome of *P. pastoris* GS115 (Genbank assembly: GCA_001708105) is incomplete and poorly annotated, we only estimated the copy number of its rDNAs as three. In eukaryotes, rDNAs encoding 18S, 5.8S, and 28S rRNAs that are transcribed into a single RNA precursor by RNA polymerase I are also organized in TRs. For example, there are approximately 200–600 rDNA copies (**Figure 3C**) distributed in

short arms of the five acrocentric chromosomes (chromosomes 13, 14, 15, 21, and 22) of human. [10]. In prokaryotic cells, 5S, 23S, and 16S rRNA genes are typically organized as a co-transcribed operon. There may be one or more copies of the operon dispersed in the genome and the copy numbers typically range from 1 to 15 in bacteria. For example, there are four copies at two rDNA loci in the chromosome 1 (GenBank: CP022603) and 2 (GenBank: CP022604) of *Ochrobactrum quorumnocens* (**Figure 3D**). Compared to those of human, bacteria, and *P. pastoris* rDNAs, 20 copies of *O. polymorpha* rDNA quadruples are very closely organized, suggesting that their transcription is regulated with high efficiency. This genomic feature may contribute to the high yield characteristics of *O. polymorpha*.

Besides the high similarity of genomic arrangement, the rDNAs of *S. cerevisiae* and *O. polymorpha* HU-11 share high nt sequence identities of 95.3% (1720/1805), 96.2% (152/158), 92% (3,111/3,381), and 96.7% (117/121) for 18S, 5.8S, 25S, and 5S rDNAs, respectively. However, the rDNAs (Genbank: FN392325) of *P. pastoris* GS115 and *O. polymorpha* HU-11 have nt sequence identities of 87.3% (1477/1691), 80% (84/105), and 80.5% (2,073/2,576) for 18S, 5.8S, and 25S rDNAs, respectively. This finding contradicts the results of a previous study [1] in which phylogenetic analysis using 153 protein-coding genes had shown that *Pichia pastoris* GS115 and *O. polymorpha* are members of a clade that is distinct from the one that *S. cerevisiae* belongs to. However, our studies revealed that HU-11 is phylogenetically closest to NCYC495, followed by DL-1, *S. cerevisiae*, and *P. pastoris* GS115. In addition, our studies have shown that the rDNA genes are more conservative than that are protein-coding genes in yeasts, and rDNA is an important feature of yeasts for their detection, identification, classification and phylogenetic analysis.

Long terminal repeat retrotransposons

LTR elements with lengths of 322 bp were discovered in all seven chromosomes of HU-11. These LTR elements with low GC content of 29% (94/322) are flanked by TCTTG and CAACA at their 5' and 3' ends (**Figure 4A**). All the LTR elements in HU-11 were identified as components of Tpa5 LTR-retrotransposons (GenBank: AJ439553) from *Pichia angusta* CBS4732 (a former name of *O. polymorpha* CBS4732) in a previous study. An LTR retrotransposon (**Figure 4B**) consists of 5' LTR, 3' LTR, and a single open reading frame (ORF) encoding a putative polyprotein. This polyprotein, if translated, can be processed into truncated Gag (GAG), protease (PR), integrase (IN), reverse transcriptase (RT), and RNase H (RH). Based on the gene order (PR, IN, RT, and RH), the LTR retrotransposons of HU-11 were classified into the Ty5 type of the Ty1/copia group (Ty1, 2, 4, and 5 types) [11].

With the length corrected from 4,883 bp to 4,882 bp, AJ439558 was used as reference of Tpa5 LTR-retrotransposons to search for homologs. The results confirmed that HU-11 is phylogenetically closest to NCYC495 and they share identical 322 bp LTR elements. However, the LTR elements of HU-11 and NCYC495 are quite distinct from 282 bp LTR elements of DL-1, which were reported as 290-bp solo LTR

elements in the previous study [1]. In addition, the amino acid (aa) sequences of the polyprotein with the length of 1417-aa in HU-11 and NCYC495 LTR retrotransposons are distinct from those in DL-1. Based on the records in the UniProt Knowledgebase (UniProtKB), *O. polymorpha* strains DL-1, ATCC26012, BCRC20466, JCM22074, and NRRL-Y-7560 have nearly the same aa sequences (UniProt: W1QI12) of the polyprotein. These results reveal that the LTR retrotransposon is another important feature of yeasts useful for their detection, identification, classification, and phylogenetic analysis.

In the previous study, 50,000 fragments of 13 *Hemiascomycetes* species were used to identify LTR retrotransposons. However, the analysis was probably biased as it was based on only random sequences of approximately 1 kb on an average and not the complete genome sequences [11]. In the present study, seven copies of intact LTR retrotransposons (**Supplementary file 1**) were discovered and accurately positioned in the HU-11 genome (**Figure 2A**); five of them were located on the sense strands of chromosome 1, 2, 3, 6, and 7 (named LTR-rt1, 2, 3, 6, and 7), while the other two were located on the antisense strands of chromosome 5 and 7 (named LTR-rt5R and 7R). LTR-rt1, 3, and 6 share very high nt identities of 99.9% with each other. LTR-rt1 or 3 contains a single ORF encoding a polyprotein with the same aa sequence, while LTR-rt6 contains a single ORF with a 42-bp insertion (encoding RSSLFDVPCSPTVD), compared to LTR-rt1 and 3. LTR-rt2, 5R, 7, and 7R contain several single nt substitutions and small Insertion/Deletions (InDels), which break the single ORFs into several ORFs. The homologs of LTR-rt2, 3, and 5R in HU-11 are present in the NCYC495 genome with very high nt identities of 99.9%. The homologs of LTR-rt1, 7, and 7R, however, were not detected in the NCYC495 genome. Further analysis suggested that their absence in the NCYC495 genome was resultant from assembly errors (**described in more detail subsequently**).

Structural variation and large segment duplication

Sequence comparison between the NCYC495 and HU-11 genomes showed that they share a nt identity of 99.5% through the whole genomes, including the rDNA region and LTR retrotransposons. However, the DL-1 and HU-11 genomes share a comparatively low nt identity (< 95%) through the whole genomes. Syntenic comparison revealed that NCYC495 is so phylogenetically similar to HU11 that a nearly 100% of their genomes is covered by their syntenic regions, whereas NCYC495 is significantly distinct from DL-1 (**As shown in preceding sections**). Subsequently, the detection of structural variations (SVs) was performed between the NCYC495 and HU11 genomes. Further analysis revealed that all the large deletion or insertions (two types of SVs) are errors in the assembly of NCYC495 genome (**Figure 4C**): (1) LTR-rt1, 7, and 7R (lost in NCYC495) should have been included in the NCYC495 genome; (2) two large deletions (lost in NCYC495) are large duplicated segments at 5' and 3' ends of chromosome 2; and (3) a large insertion (acquired by NCYC495) is a large duplicated segment at 3' end of chromosome 6 (NW_017264698:1509870-1541475), the upstream and downstream parts of which are identical to segments from chromosome 2 and 3, respectively. In addition, telomeric TRs $[GGCGGGGT]_n$

(NW_017264698:1509840-1509869) were discovered at the 5' ends of this large insertion, confirming that it resulted from assembly errors.

All large duplicated segments were correctly assembled in the subtelomeric regions of the HU-11 genome. Human curation was performed to verify their locations using long (> 30 Kb) PacBio subreads. A 27,850 bp duplicated segment is present at 3' ends of chromosome 1 and 2 in the HU-11 genome. There are only four mismatches and one 1-bp gap between these two sequences. The homolog of this 27,850 bp duplicated segment is present at the 3' end of chromosome 1 in the NCYC495 genome. The downstream part of it, however, is absent at 3' end of chromosome 2 in the NCYC495 genome (**Figure 4C**), which corresponds to one 14,090 bp large deletion of NCYC495. An approximately 5,100 bp duplicated segment is present at 5' ends of both chromosomes 2 and 5 in the HU-11 genome. The homolog of this ~5,100 bp duplicated segment is present at 5' end of chromosome 5, but absent at 5' end of chromosome 2 in the NCYC495 genome (**Figure 4C**), which corresponds to the other large deletion of NCYC495. An approximately 2,500 bp duplicated segment is present at 5' ends of chromosome 7 and downstream of the ~5,100 bp duplicated segment of chromosome 2 in the HU-11 genome. Different from homologs of the 27,850-bp and ~5,100 bp duplicated segments, the homolog of this ~2,500 bp duplicated segment is correctly assembled in the NCYC495 genome (**Figure 4C**). As an important finding, telomeric TRs $[ACCCCGCC]_n$ (n > 2) were discovered at 3' end of the ~5,100 bp duplicated segment located in both chromosomes 2 and 5, and at 3' end of the ~2,500 bp duplicated segment located in chromosome 7; this finding indicated that these duplicated segments were integrated at telomeric TRs.

Conclusions

HU-11 is a nutritionally deficient mutant derived from CBS4732 by a 5 bp insertion of “GAAGT” into the 31st position of the *URA3* CDS; this insertion causes a frame-shift mutation of the *URA3* CDS, resulting in the loss of the *URA3* functions. Since the difference between the genomes of CBS4732 and HU-11 is only five nts, CBS4732 and HU-11 can be regarded as identical strains. As the syntenic regions cover nearly 100% of their genomes, NCYC495 is considered to be phylogenetically very close to HU11. This fact clearly reveals the origin and relationship of NCYC495 and CBS4732/HU11. Therefore, the full-length genome of the *O. polymorpha* strain HU-11 can be used as a reference for future studies of CBS4732, NCYC495, and their derivative strains. For example, we corrected assembly errors in the NCYC495 genome using the full-length genome of HU-11, which facilitated in obtaining the full-length genome of NCYC495.

Using the high-quality full-length HU11 genome, large duplicated segments in subtelomeric regions were first comprehensively discovered within a methylotrophic yeast, which was overlooked in the previous studies due to the incomplete genomes. A computational study showed that subtelomeric families are evolving and expanding much faster than that are the families that do not contain subtelomeric genes in the

yeasts. These findings thus, indicate that the extraordinary instability of eukaryotic subtelomeres supports rapid adaptation to novel niches by promoting gene recombination and duplication followed by functional divergence of the alleles [9]. Our results suggest that large segment duplication in subtelomeric regions is the main reason for genome expansion in yeasts, also suggesting that subtelomeric families are evolving and expanding much faster. The discovery of telomeric TRs in these segments indicate that the duplicated segments in subtelomeric regions were integrated at telomeric TRs. However, the underlying molecular mechanism is still unknown. The duplicated segments in different regions are nearly identical; suggesting that this level of divergence was acquired shortly after the duplication. Such unknown molecular mechanism may still function in *O. polymorpha* and the exogenous genes may get integrated into the subtelomeric regions, which is extremely important for the industrial applications of *O. polymorpha*. Based on the molecular mechanism, a simple and highly efficient genome editing system can be developed to integrate or cleave large segments at the telomeric TRs.

Methods and Materials

The *Ogataea polymorpha* strain HU-11 was obtained from Tianjin Hemu Health Biotechnological Co., Ltd. DNA extraction and quality control were performed as described in our previous study [12]. A 500 bp DNA library was constructed as described in our previous study [12] and sequenced on the Illumina HiSeq X Ten platform. A 10 Kb DNA library was constructed and sequenced on the PacBio Sequel platforms, according to the manufacturer's instruction. The software SMRTlink v5.0 (--minLength=50, --minReadScore=0.8) was used for PacBio data cleaning and quality control, while the software Fastq_clean v2.0 [13] was used for Illumina data cleaning and quality control. The software MECAT v1.2 was used to assemble the HU-11 draft genome using PacBio data. To polish the HU-11 genome, Illumina data was aligned to the HU-11 draft genome using the software BWA. Then, samtools was used to obtain the BAM and pileup files from the alignment results. A Perl script was used to extract the consensus sequence from the pileup file. This procedure was repeatedly performed to obtain the final genome sequence. The curation of genome and genes was performed using the software IGV. Statistical computation and plotting were performed using the software R v2.15.3 with the Bioconductor packages [14].

Gene annotation and SV detection were performed using the software blast v2.9.0. Syntenic comparison were performed using the CoGe website (https://genomevolution.org/CoGe). 5,155 protein-coding genes of NCYC495 were annotated in the previous study. Among them, two genes were removed as they are ITS sequences. 17 genes in the large insertion in the chromosome 6 of NCYC495 were removed as they are assembly errors. Three and six genes in two large deletions (lost in NCYC495) were added, as the losses of them in chromosome 2 resulted from assembly errors. Finally, the corrected 5,145 genes of NCYC495 were aligned to the HU-11 genome using the software blast v2.9.0 to annotate protein-coding genes of HU-11.

Supplementary information

Additional file 1: Figure S1. A workflow to generate full-length genome sequence of an RNA virus. Table S1. Collection of ticks. Table S2. Primers for PCR amplification coupled with Sanger sequencing. Table S3. PCR reagent for each sample. Table S4. 17 complete genomes for further analysis.

Abbreviations

TR: tandem repeat; STR: short tandem repeat; LTR: long terminal repeat; mt: mitochondrial; nt: nucleotide; aa: amino acid; ORF: Open Reading Frame; CDS: Coding Sequence; SV: Structural Variation; InDel: Insertion/Deletion.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent to publish

Not applicable.

Availability of data and materials

The complete genome sequence of *Hansenula Polymorpha* HU-11 is available at the NCBI GenBank database under the accession numbers CP073033-40, in the project PRJNA687834.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Natural Science Foundation of Guangdong Province of China (2021A1515011072) to Jinlong Bei, Tianjin Key Research and Development Program of China (19YFZCSY00500) to Shan Gao, and Shandong Province Natural Science Foundation (ZR2020QC071 and ZR2020MC027) to Guangyou Duan. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

SG conceived the project. SG and GD supervised the present study. SG drafted the manuscript. JC assembled the HU-11 genome. JB and JY executed the experiments. SG, GD, TY and XL analyzed the data. GD prepared the figures and tables. SG, HW, WB and JR revised the manuscript. All authors have read and approved the manuscript.

Acknowledgments

We appreciate the help equally from the people listed below. They are Wenjun Bu, Huaijun Xue, Dawei Huang, Yanqiang Liu, Bingjun He, Qiang Zhao, Zhen Ye and Xiufeng Jin from College of Life Sciences, Nankai University.

REFERENCES

1. N.V. Ravin, M.A. Eldarov, V.V. Kadnikov, A.V. Beletsky and K.G. Skryabin, *Genome sequence and analysis of methylotrophic yeast Hansenula polymorpha DL1*. BMC Genomics, 2013. **14**(1): p. 1-20.
2. J.O.F. Morais and M.H.D. Maia, *Estudos de microorganismos encontrados em leitos de despejos de caldas de destilarias de Pernambuco. II. Uma nova especie de Hansenula: H. polymorpha.* An Esc Super Quim Univ Recife, 1959. **1**: p. 15-20.
3. L.J. Wickerham, *Taxonomy of yeasts.* Technical Bulletin United States Department of Agriculture, 1951. **6**(4715): p. 781-782.
4. D.W. Levine and C.L. Cooney, *Isolation and Characterization of a Thermotolerant Methanol-Utilizing Yeast.* Appl. Environ. Microbiol., 1973.
5. R.R. Massoud, C.P. Hollenberg, L. Juergen, W. Holger, G. Eike, W. Christian, A. Kaj, H. Jean, P. Michael and D. Ulrike, *The Hansenula polymorpha (strain CBS4732) genome sequencing and analysis.* FEMS Yeast Research, 2003. **4**(2): p. 207-215.
6. H. Wang, X. He and B. Zhang, *A method to construct Ogataea polymorpha strains and its application*.
7. H. Wang, C. Wang and J. Yang, *High-dose recombinant B hepatitis vaccine in Ogataea polymorpha*.
8. H. Wang, C. Wang, Y. Wang and J. Yang, *A recombinant hirudin gene and its application*.
9. C.A. Brown, *Rapid expansion and functional divergence of subtelomeric gene families in yeast*. 2010, Harvard University.;.
10. S. Agrawal and A. Ganley, *The conservation landscape of the human ribosomal RNA gene repeats.* PloS one, 2018. **13**(12): p. e0207531.
11. Neuveglise and C., *Genomic Evolution of the Long Terminal Repeat Retrotransposons in Hemiascomycetous Yeasts.* Genome Research, 2002. **12**(6): p. 930-943.
12. Y. Wang, Z. Wang, X. Chen, H. Zhang, F. Guo, K. Zhang, H. Feng, W. Gu, C. Wu and L. Ma, *The Complete Genome of Brucella Suis 019 Provides Insights on Cross-Species Infection.* Genes, 2016. **7**(2): p. 1-12.
13. M. Zhang, F. Zhan, H. Sun, X. Gong, Z. Fei and S. Gao. *Fastq_clean: An optimized pipeline to clean the Illumina sequencing data with quality control*. in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*. 2014. IEEE.
14. S. Gao, J. Ou and K. Xiao, *R language and Bioconductor in bioinformatics applications(Chinese Edition)*. 2014, Tianjin: Tianjin Science and Technology Translation Publishing Ltd.

Figure legends

Figure 1 Errors in PacBio data and Illumina data

The errors in the low complexity and short tandem repeat (STR) regions can be corrected during the genome polishment using Illumina data, while the error in the long (>10 copy numbers) poly(GC) regions need be curated after the genome polishment. **A.** A example to show that the assembled genomes using high-depth PacBio data still contain errors in the low complexity regions. **B.** A example to show that the assembled genomes using high-depth PacBio data still contain errors in the STR regions. **C.** A example to show that the genome polishment using Illumina data causes error in the long poly(GC) regions.

Figure 2 Full-length genome of the *Ogataea polymorpha* strain HU-11

A. The full-length *O. polymorpha* HU-11 genome includes the complete sequences of seven linear chromosomes, which were named as 1 to 7 from the smallest to the largest. The 5' and 3' telomeric regions were not included. The minimum, Q_{90}, Q_{75}, Q_{50}, Q_{25}, Q_{10} and maximum of GC contents (%) are 0.08, 0.408 0.436, 0.472, 0.514, 0.554 and 0.732. The GC contents (%) were calculated by 500-bp sliding windows and then trimmed between Q_{10} and Q_{90} for plotting the heatmaps. Long terminal repeat retrotransposons are indicated by LTR-rt in the chromosomes. Marker: URA3 (encoding orotidine 5'-phosphate decarboxylase), HARS(Hansenula autonomously replicating sequence), FGH (S-formylglutathione hydrolase), MOX(methanol oxidase), FDH (Formate dehydrogenase) and TERT (telomerase reverse transcriptase).**B.** For the data submission to the GenBank database, the genome sequence of circular mitochondrion was linearized, starting at the first codon (indicated by a red arrow) of the ORF3 coding sequence (CDS).

Figure 3 Organization of rDNA genes in yeast, human and bacteria

A. The only rDNA locus is located in chromosome 2 of the *Ogataea polymorpha* strain HU-11, containing 20 times of TRs. Here only six TRs are shown. **B.** A TR of rDNAs in *Saccharomyces cerevisiae* also contains 5S, 18S, 5.8S and 25S rDNAs as a quadruple, repeating 2 times on the chromosome 7 of its genome. Four other 5S rDNAs are located separately away from the rDNA quadruples in *S.* cerevisiae. **C.** Each human rDNA unit has an rRNA coding region and an intergenic spacer (IGS). Here only eight units are shown. ITS: internal transcribed spacer; ETS: external transcribed spacers. **D.** There are four copies at two rDNA loci in the chromosome 1 (GenBank: CP022603) and 2 (GenBank: CP022604) of the *Ochrobactrum quorumnocens* genome

Figure 4 LTR retrotransposons and structural variations in NCYC495

A. NCYC495 and HU-11 share identical 322-bp LTR elements, which are flanked by TCTTG and CAACA at their 5' and 3' ends. Three of seven LTR retrotransposons of HU-11 does not have homologs in NCYC495B due to assembly errors. **B.** A LTR retrotransposon consists of 5' LTR, 3' LTR and a single open reading frame (ORF) encoding a putative polyprotein. This polyprotein, if translated, can be processed into trunculated gag (GAG), protease (PR), integrase (IN), reverse transcriptase (RT) and RNase H (RH). **D.** Chr1, 2, 5 and 7 represent the sequences (GenBank: NW_017264703, 704, 700 and 699) of NCYC495. The numbers indicate the starting genomic position of the large duplicated segment (in yellow color) in chromosome 1 and the ending genomic position of the large duplicated segments in chromosome 5 (in green color) and 7 (in blue color). NCYC495 is so phylogenetically close to HU11 that the syntenic regions cover

almost 100% of their genomes. All the large deletion or insertions are errors in the assembly of NCYC495 genome. Two large deletions (indicated by black slash lines) in chromosome 2 should have been included in the NCYC495 genome. rrnS: small subunit ribosomal RNA; rps3: ribosomal protein S3.

Tables

Table 1. Genomes of three basic *O. polymorpha* strains.

Chromosome	**HU-11**	**NCYC495**	**DL-1**	**HU-11 Size(bp)**	**Marker**
Chr1	CP073033	NW_017264703	NC_027865	1,001,202	URA3
Chr2	CP073034	NW_017264704	NC_027866	1,011,645	rDNA
Chr3	CP073035	NW_017264702	NC_027864	1,265,386	HARS
Chr4	CP073036	NW_017264701	NC_027863	1,316,701	FGH
Chr5	CP073037	NW_017264700	NC_027862	1,357,511	MOX
Chr6	CP073038	NW_017264698	NC_027860	1,513,389	FDH
Chr7	CP073039	NW_017264699	NC_027861	1,525,948	TERT
ChrM	CP073040	NA	NC_014805	59,473	COIII
Total(Mbp)	9.1	8.97	8.87		
GC%	47.76	47.86	47.83		
mRNA#	5145	5145	5325*		
tRNA#	80	80	80		
rRNA#	4×20	4×6	4×25		

*the expression of 5238 out of 5325 genes were detected. # The numbers of mitochondrial genes and CDSs in the LTR retrotransposons were not counted. As these 5' and 3' telomeric TRs are varied in their lengths, they were not included into the seven linear chromosomes of HU-11, which were named as 1 to 7 from the smallest to the largest. The accession numbers of NCYC495 and DL-1 were mapped to the chromosome numbers of HU-11. Marker genes were used to identify seven chromosomes. CBS4732 and HU-11 can be used as identical strains for the investigation, as the only difference between them is a 5-bp insertion. Marker: URA3 (encoding orotidine 5'-phosphate decarboxylase), HARS(*Hansenula* autonomously replicating sequence), FGH (S-formylglutathione hydrolase), MOX(methanol oxidase), FDH (Formate dehydrogenase), TERT (telomerase reverse transcriptase) and COIII.

A.

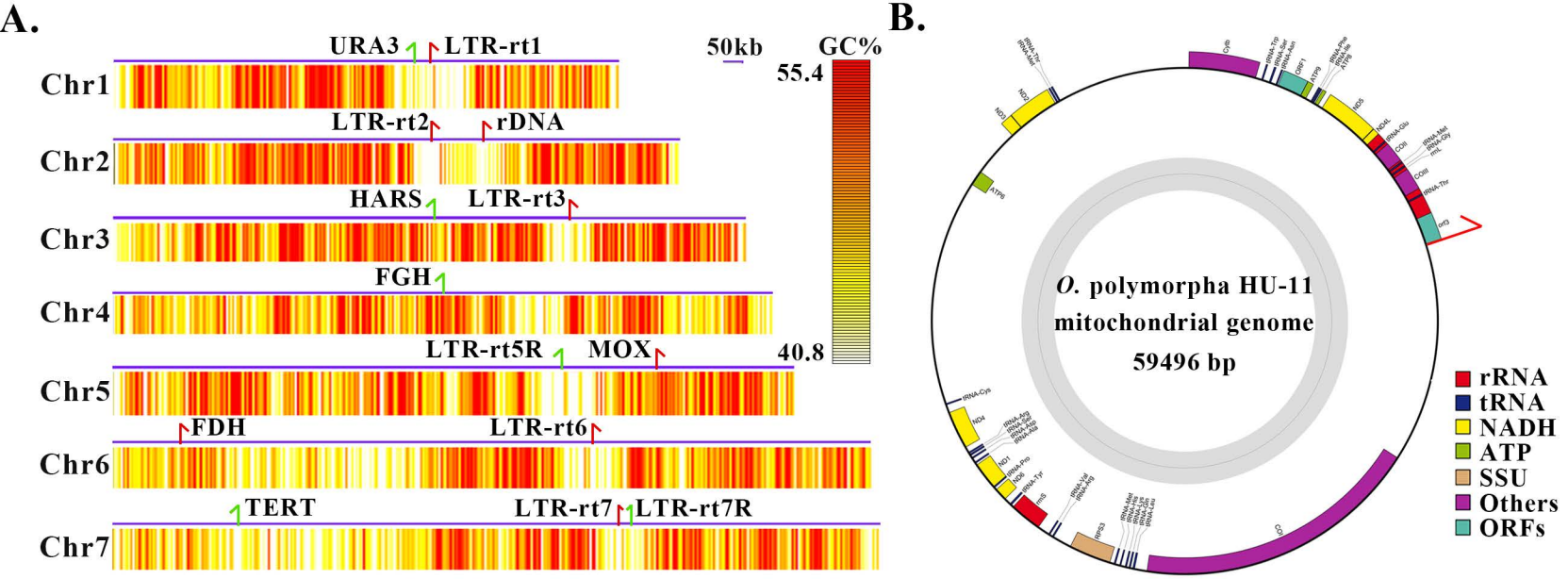
```
PacBio assembly     ✗  Chr4: 854742 CCATCCCAGAAGTCTCCT-CGT-GAGTCGTCTCCTTCTCATAGCTCTACCAACTCGTATT 854798
                                    |||||||||||||||||| ||| |||||||||||||||||||||||||||||||||||||
Illumina polished   ✓  Chr4: 854742 CCATCCCAGAAGTCTCCTCCGTGGAGTCGTCTCCTTCTCATAGCTCTACCAACTCGTATT 854798
```

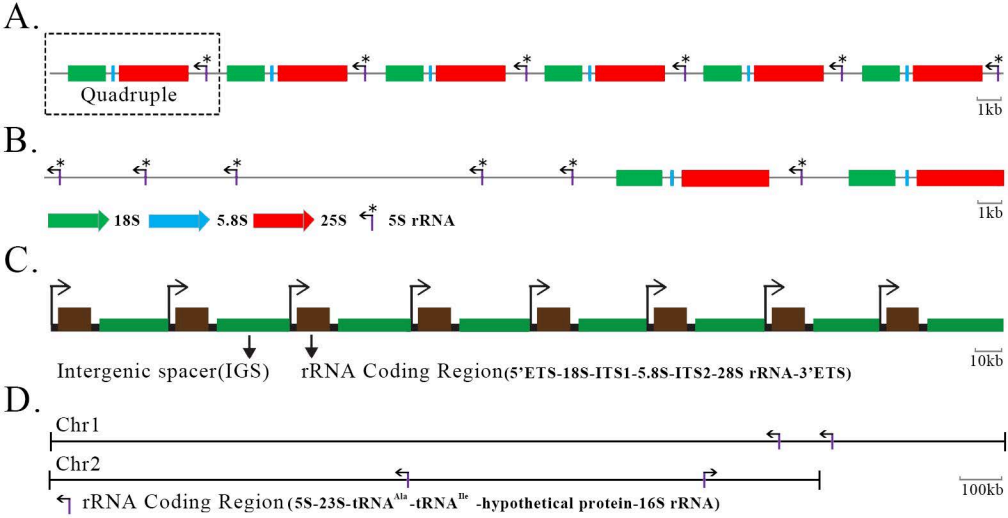
B.

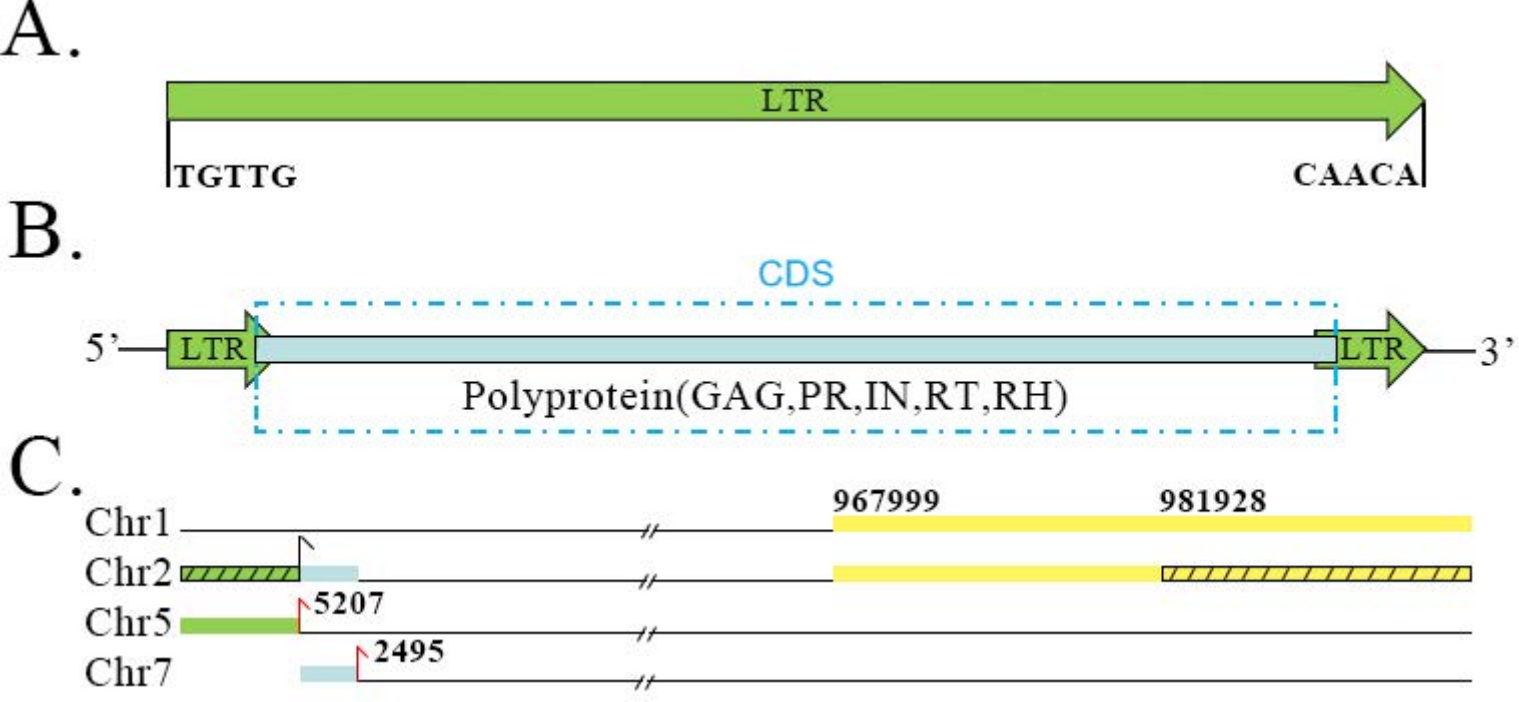
```
PacBio assembly     ✗  Chr2: 250753 AATAAAAAATTC-----AAAAAAAAAAAAAAAATTCAAAAAAATTACAGAAAAATTACAAG 250811
                                    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Illumina polished   ✓  Chr2: 250753 AATAAAAAATTCAAAAAAAAAAAAAAAAAAAAATTCAAAAAAATTACAGAAAAATTACAAG 250811
```

C.

```
PacBio assembly     ✓  Chr3:      1 TGCTTTTCTGTTTTTCCATCCCCTTCCCCCCCTCCAACTTTATTTTTTCCTCCTTTTTAA     60
                                    |||||||||||||||||| ||||  ||||||| |||||||||||||||||||||||||||||
Illumina polished   ✗  Chr3:      1 TGCTTTTCTGTTTTTCCACCCCCCCCCCCCCCCCCAACTTTATTTTTTCCTCCTTTTTAA     60
```


A.
URA3
LTR-rt1
50kb
GC%
55.4
Chr1
LTR-rt2
rDNA
Chr2
HARS
LTR-rt3
Chr3
FGH
Chr4
LTR-rt5R
MOX
40.8
Chr5
FDH
LTR-rt6
Chr6
TERT
LTR-rt7
LTR-rt7R
Chr7
B.
O. polymorpha HU-11
mitochondrial genome
59496 bp
rRNA
tRNA
NADH
ATP
SSU
Others
ORFs


A.
Quadruple
1kb
B.
18S
5.8S
25S
5S rRNA
1kb
C.
Intergenic spacer(IGS)
rRNA Coding Region(5'ETS-18S-ITS1-5.8S-ITS2-28S rRNA-3'ETS)
10kb
D.
Chr1
Chr2
100kb
rRNA Coding Region (5S-23S-tRNA^{Ala}-tRNA^{Ile} -hypothetical protein-16S rRNA)


A.
LTR
TGTTG
CAACA
B.
CDS
5'
LTR
LTR
3'
Polyprotein(GAG,PR,IN,RT,RH)
C.
Chr1
Chr2
Chr5
Chr7
967999
981928
5207
2495