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

jia Chang
Nankai University
Jinlong Bei
Guangdong Academy of Agricultural Sciences
Hemu Wang
LTDS: Laboratoire de tribologie et dynamique des systemes
Jun Yang
LTDS: Laboratoire de tribologie et dynamique des systemes
Xin Li
Nankai University
Tung On Yau
Nottingham Trent University https://orcid.org/0000-0002-3283-0370
Wenjun Bu
Nankai University
Jishou Ruan
Nankai University
Guangyou Duan
Qilu Normal University
Shan Gao (gao_shan@mail.nankai.edu.cn)
Nankai University https://orcid.org/0000-0002-8919-1338

Research article

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

Posted Date: May 18th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-530556/v1

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 subtelomeric regions. Three novel findings are: (1) O. polymorpha NCYC495 is so phylogenetically close to CBS4732/HU-11 that the syntenic regions covers nearly 100% of their genomes with a nucleotide identity of 99.5%, while NCYC495 is significantly distinct from DL-1; (2) large segment duplication in subtelomeric 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 into yeast genomes.

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 HU-11 and CBS4732. 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, NCYC495, and their derivative strains.

Full Text

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed. However, the manuscript can be downloaded and accessed as a PDF.

Figures

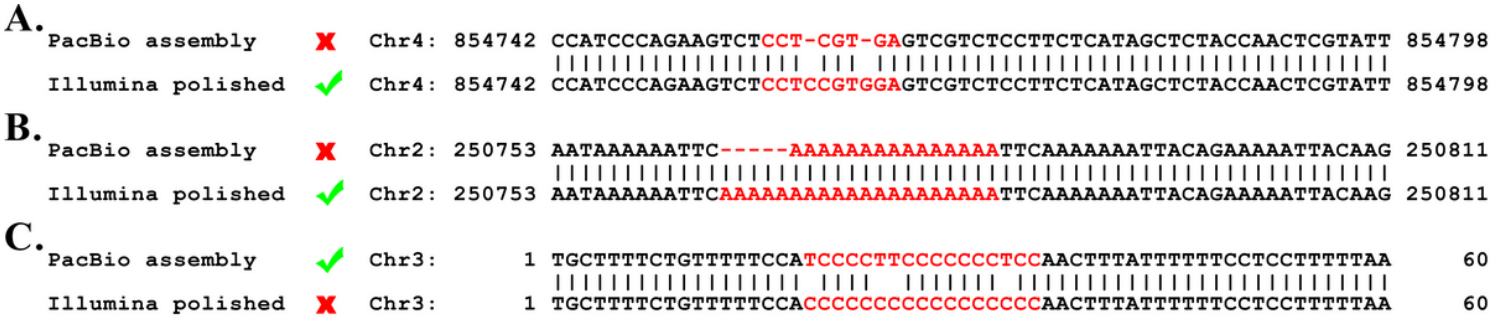

Figure 1

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.

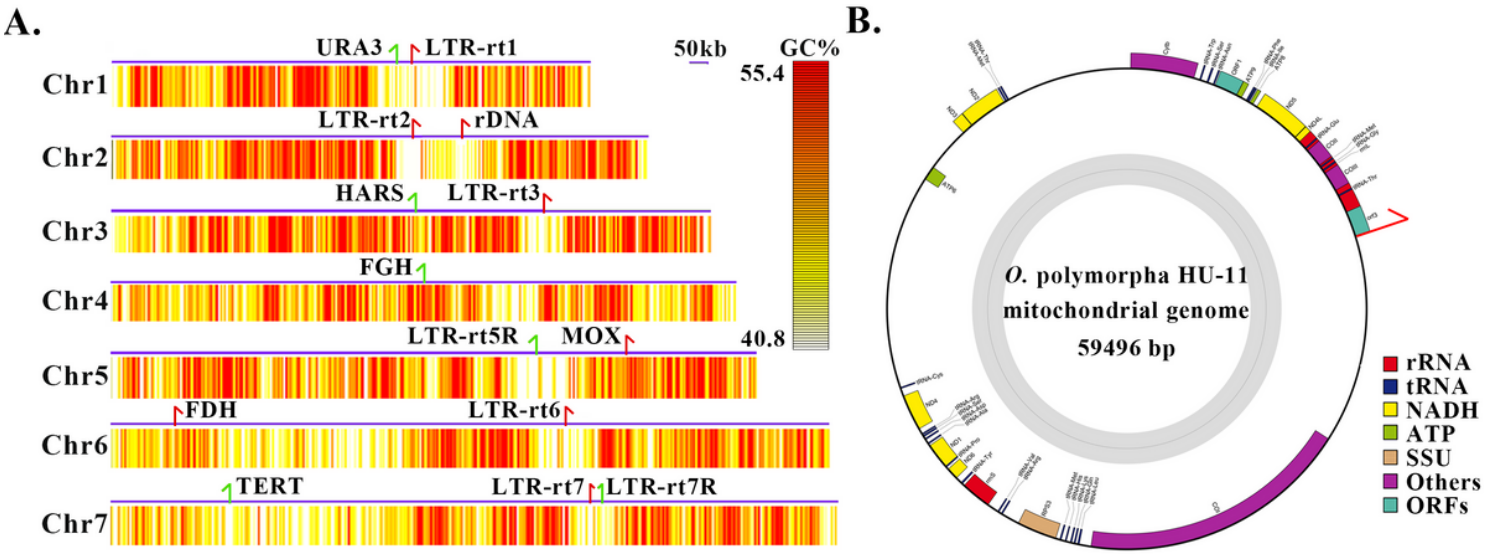

Figure 2

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, Q90, Q75, Q50, Q25, Q10 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 Q10 and Q90 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).

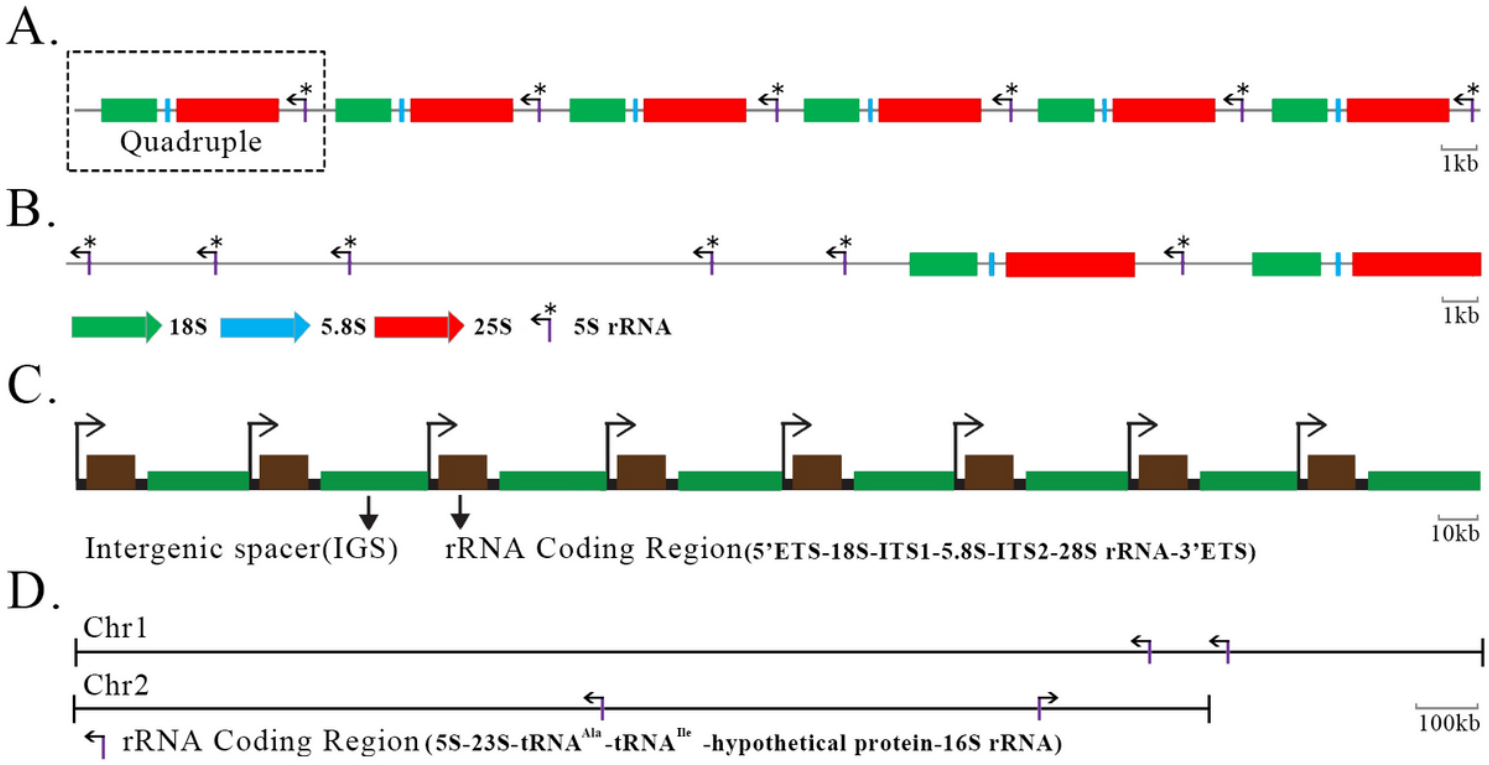

Figure 3

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

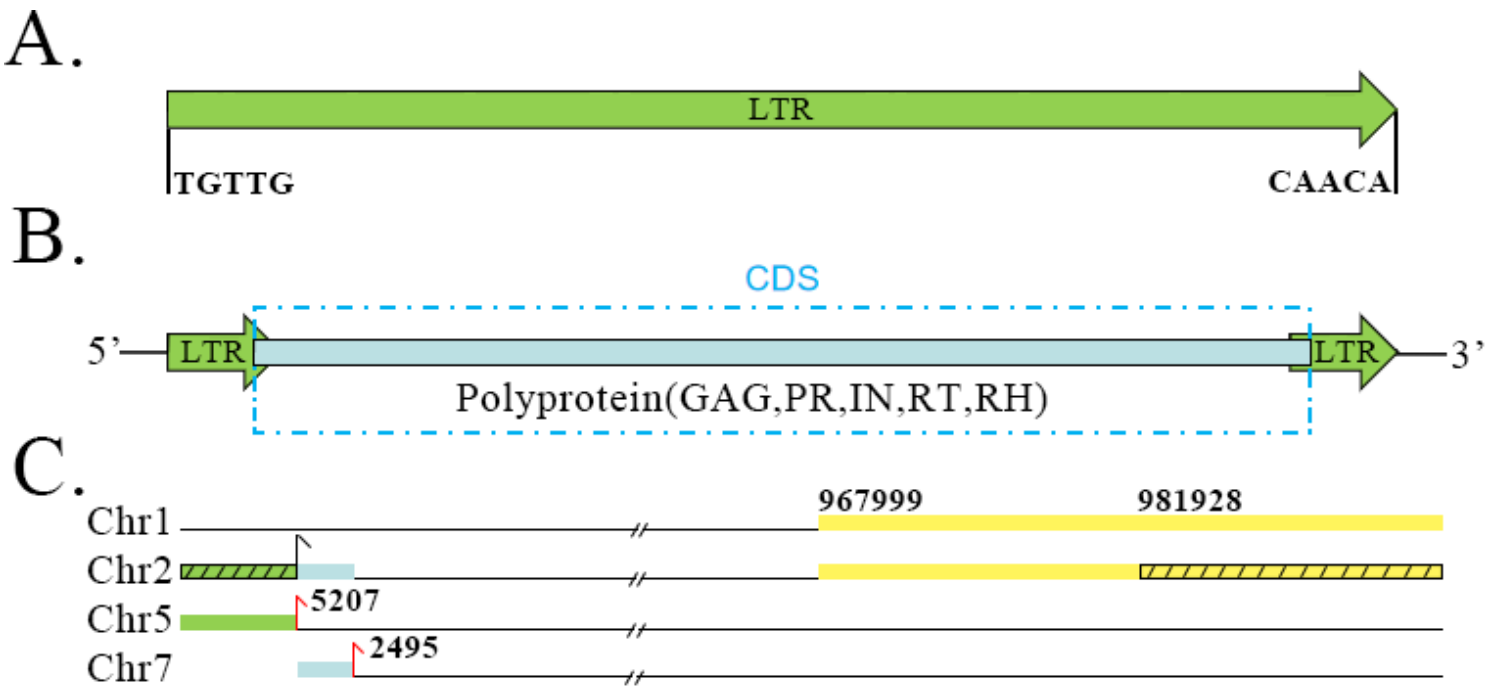

Figure 4

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). O. polymorpha NCYC495 is so phylogenetically close to HU-11 and CBS4732 that the syntenic regions covers nearly 100% of their genomes. All the large deletion or 428 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.