Draft Genome Sequence of the Yeast Pachysolen tannophilus CBS 4044/NRRL Y-2460

Liu, Xiaoying ; Kaas, Rolf Sommer; Jensen, Peter Ruhdal; Workman, Mhairi

Published in:
Eukaryotic Cell (Online Edition)

Link to article, DOI:
10.1128/EC.00114-12

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
A draft genome sequence of the yeast *Pachysolen tannophilus* CBS 4044/NRRL Y-2460 is presented. The organism has the potential to be developed as a cell factory for biorefineries due to its ability to utilize waste feedstocks. The sequenced genome size was 12,238,196 bp, consisting of 34 scaffolds. A total of 4,463 genes from 5,346 predicted open reading frames were annotated with function.

**The yeast *Pachysolen tannophilus* was first isolated from wood extracts used in leather tanning (2) and has gained interest due to its ability to utilize D-xylose (1,7,8) for fuel ethanol production. It can ferment the common sugars (glucose, mannose, and galactose) occurring in hemicellulose hydrolysate mixtures (8) and can produce ethanol from glycerol (4). The whole-genome sequencing of *P. tannophilus* was performed to provide genetic information as a necessary step toward engineering of the strain and its potential development as an industrial cell factory.**

Sequencing was performed by a whole-genome shotgun strategy with an Illumina genome analyzer (Beijing Genomics Institute, Shenzhen, China). The raw data of short reads were assembled into 279 contigs which were ordered into 34 scaffolds (≥2 kb) with an N50 size of 1.1 Mb by using the SOAPdenovo package (3). Augustus software, version 2.5 (9), was utilized to identify protein-coding genes in the genome, and the functional annotation was accomplished by BLASTP analysis (E-value < 1 × 10^-5) of protein sequences in the databases (COG and KEGG), and the best hit was selected. Pulsed-field gel electrophoresis (PFGE) (5) was performed to predict the number and approximate sizes of chromosomes. The program for PFGE was 48 h at 3 V/cm with a 500-s switch time at an angle of 106° with 0.5 LMP (low melting point) agarose at 14°C.

The total length of the sequenced genome was 12,238,196 bp (without N), with a GC content of 29.82%. A total of 1970.8 Mb of raw data was sequenced, representing around 145-fold coverage of the genome. Five thousand three hundred forty-six protein-coding genes (coding sequences [CDSs]) were predicted, and 4,463 (83.5%) genes were annotated with function.

Four coding sequences of *P. tannophilus* were retrieved from GenBank to compare with the annotated CDSs. There were 2-bp differences among 4,803 bp of length. Furthermore, 13 full-length coding sequences were PCR amplified and resequenced to estimate accuracy. The total resequenced length was 21,111 bp, with 100% similarity obtained.

Based on PFGE results, six chromosomal bands were separated, with two of the bands probably migrating as doublets. The sizes of the chromosomal bands were estimated to be 2.9 ± 0.05 Mb (mean ± standard deviation), 2.1 ± 0.04 Mb, 1.9 ± 0.05 Mb, 1.6 ± 0.08 Mb (doublet), 1.3 ± 0.07 Mb (doublet), and 0.98 ± 0.02 Mb based on comparison with the yeast marker *Hansenula wingei*. The estimated genome size of *P. tannophilus* was approximately 13.6 ± 0.4 Mb with an estimated 8 chromosomes.

**Nucleotide sequence accession numbers.** The draft genome sequences of *P. tannophilus* were deposited in EMBL with contig accession numbers CAHV01000001 to CAHV01000267.

**ACKNOWLEDGMENTS**

This work was funded by the European Community’s 7th Framework Research Programme under grant agreement number 213506 (project GLYFINERY), providing financial support to X.L. and M.W.

**REFERENCES**