8th International Conference of the Brazilian Association for Bioinformatics and Computational Biology
OCTOBER 14th to 17th
X-meeting 2012
Unicamp Convention Center | Campinas - São Paulo

Draft genome sequencing of the Yeast Spathaspora arborariae sp.

Edmar de Melo Santos, Gloria Regina Franco, Boris Juan Carlos Ugarte Stambuk, Carlos Augusto Rosa, Francisco Pereira Lobo
UNICAMP, UFMG, UFSC, UFMG, Embrapa Informática Agropecuária

In 2009, a new yeast was isolated from rotting wood samples collected in the Atlantic Rain Forest and the Cerrado ecosystems in Brazil. This yeast was found to be a new species from the Spathaspora genus and was denominated Spathaspora arborariae. It has gained interest due to its ability to utilize D-xylose for producing high ethanol yields. Due to its potential to convert cellulosic biomass, an underused source of biofuel, this yeast is a potential source of genes, enzymes and knowledge for metabolic xylose fermentation. This work provides the description of a draft genome of S. arborariae. The following steps are necessary for the assembly of a genome: sequencing, assembly and gene prediction and annotation. Genome sequencing was performed by a whole-genome shotgun strategy using the 454 platform. The final assembly was produced using Newbler. Pulsed Field Gel Eletrophoresis (PFGE) was performed to estimate the number and approximate size of the chromosomes. Ab initio Gene prediction was done using the GeneMark-hmm software, trained with Saccharomyces cerevisiae gene models. Gene annotation was done using the BLAST+ software and the nr database. Minimum cutoffs for automatic annotation were an identity greater than 50% in more thant 80% of the query sequence, with an e-value smaller than 0.00001. Prediction of rRNA genes was done using the RNAmmer software, and prediction of tRNA genes was done using tRNAscan-SE with parameters to achieve maximum accuracy. The raw sequence data comprised a total of 915.700 reads with 657.682 mate-pairs totalizing 291.670.584 nucleotides. The PFGE estimated the S. arborariae genome to be 12 Mb, generating and average sequencing coverage of 23X. The raw sequence data was assembled in 439 contigs and 41 scaffolds with a total length of 12.708.019 bp after excluding the 162.563 non-determined nucleotides. The assembled genome has an N50 of ~679 kb (6 scaffolds) and an N90 of ~ 202 kb (18 scaffolds), with an average GC content of 31,7%. We found 6595 gene models greater than 100 nucleotides, of which 5569 (84,4%) were found to possess similar sequences in the nr database. The rRNA gene location was predicted to be at the scaffold 9, and 185 distinct tRNA genes were found scattered across the scaffolds. To validate our assembly we used the four sequences of S. arborariae available in NCBI. These sequences comprised of distinct portions of the rRNA gene structure, and when aligned to the prediced rRNA, we observed 100% matches with no gaps, therefore demonstrating that our assembly is coherent with S. arborariae genomic sequences from other sources. This draft opens perspectives for studies that increase the understanding of S. arborariae metabolism, and future steps will include the prediction metabolic pathway, comparative genomics and gene prospecting.

Keywords: Draft genome sequencing of the Yeast Spathaspora arborariae sp.
Concentration area: Genomics Evolution

⬅