Journal of Biotechnology 154 (2011) 312–320

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

High-quality genome sequence of *Pichia pastoris* CBS7435

Andreas Küberl[a,b,1], Jessica Schneider[c,d,1], Gerhard G. Thallinger[b,e,1], Ingund Anderl[a], Daniel Wibberg[c], Tanja Hajek[a,b], Sebastian Jaenicke[d], Karina Brinkrolf[c], Alexander Goesmann[d], Rafael Szczepanowski[c], Alfred Pühler[c], Helmut Schwab[a,b], Anton Glieder[a,b], Harald Pichler[a,b,*]

[a] *Institute of Molecular Biotechnology, Graz University of Technology, Petersgasse 14, A-8010 Graz, Austria*
[b] *Austrian Centre of Industrial Biotechnology (ACIB), Petersgasse 14, A-8010 Graz, Austria*
[c] *Institut für Genomforschung und Systembiologie, Centrum für Biotechnologie, Universität Bielefeld, Universitätsstraße 27, D-33615 Bielefeld, Germany*
[d] *Bioinformatics Resource Facility, Centrum für Biotechnologie, Universität Bielefeld, Universitätsstraße 25, D-33615 Bielefeld, Germany*
[e] *Institute for Genomics and Bioinformatics, Graz University of Technology, Petersgasse 14, A-8010 Graz, Austria*

ARTICLE INFO

Article history:
Received 8 February 2011
Received in revised form 21 April 2011
Accepted 26 April 2011
Available online 6 May 2011

Keywords:
Pichia pastoris
Genome sequence
Genome annotation
Mitochondrial genome
Alpha mating factor

ABSTRACT

The methylotrophic yeast *Pichia pastoris* (*Komagataella phaffii*) CBS7435 is the parental strain of commonly used *P. pastoris* recombinant protein production hosts making it well suited for improving the understanding of associated genomic features. Here, we present a 9.35 Mbp high-quality genome sequence of *P. pastoris* CBS7435 established by a combination of 454 and Illumina sequencing. An automatic annotation of the genome sequence yielded 5007 protein-coding genes, 124 tRNAs and 29 rRNAs. Moreover, we report the complete DNA sequence of the first mitochondrial genome of a methylotrophic yeast. Fifteen genes encoding proteins, 2 rRNA and 25 tRNA loci were identified on the 35.7 kbp circular, mitochondrial DNA. Furthermore, the architecture of the putative alpha mating factor protein of *P. pastoris* CBS7435 turned out to be more complex than the corresponding protein of *Saccharomyces cerevisiae*.

1. Introduction

The methylotrophic yeast *Pichia pastoris* has become one of the most important hosts for heterologous protein expression (Cregg et al., 2009). The name *P. pastoris* is commonly used for a number of distantly related strain types. One of these strains, *P. pastoris* NRRL Y-11430, which has been deposited as CBS7435 (CBS, Centraalbureau voor Schimmelcultures) and has been classified as *Komagataella phaffii* (Kurtzman, 2009), is the parental strain for the two most frequently used protein expression host strains GS115 (US Patent 4,879,231, Phillips Petroleum, 1989) and the prototrophic strain X-33 (Cregg, 2007). Especially the use of the strong, methanol-inducible *P. pastoris* alcohol oxidase 1 promoter (P_{AOX1}) and its variants provides efficient tools for heterologous protein production by *P. pastoris* (Hartner et al., 2008). Numerous publications report on heterologous protein expression in *P. pastoris* based on methanol induction employing P_{AOX1} (Hartner et al., 2008; Cereghino and Cregg, 2000).

Oxidizing methanol to formaldehyde, alcohol oxidases Aox1p and Aox2p are responsible for the initial step of the methanol utilization (MUT) pathway. This pathway is the main metabolic route in *P. pastoris* when protein expression is induced by growth on methanol as a carbon source (Hartner and Glieder, 2006). Formaldehyde derived from methanol oxidation can be directed into two different routes. It is either further oxidized to CO_2 by formaldehyde dehydrogenase, S-formylglutathione hydrolase and formate dehydrogenase (dissimilative pathway) or condensed with xylulose-5-phosphate and subsequently converted into dihydroxyacetone and glyceraldehyde-3-phosphate (GAP) by dihydroxyacetone synthase (assimilative pathway). The dissimilative pathway yields two equivalents of NADH with formaldehyde dehydrogenase (Fld1p) being the rate limiting enzyme (Schroer et al., 2010). The assimilative pathway is used to produce biomass *via* glycolysis (Hartner and Glieder, 2006).

While regulatory patterns have been investigated by transcriptome analysis upon methanol induction for the methylotrophic yeasts *Hansenula polymorpha* (van Zutphen et al., 2010) and *P. pastoris* (Sauer et al., 2004), the genes themselves have not been entirely analyzed yet. A list of all genes that were expected to encode for enzymes involved in the MUT pathway and their loci in the *P. pastoris* GS115 genome has been published (De Schutter et al., 2009). Thereby, it was recognized that the *P. pastoris* MUT

* Corresponding author at: Institute of Molecular Biotechnology, Graz University of Technology, Petersgasse 14/2, A-8010 Graz, Austria. Tel.: +43 316 873 4089; fax: +43 316 873 4071.
E-mail address: harald.pichler@tugraz.at (H. Pichler).
[1] Authors contributed equally.

0168-1656/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbiotec.2011.04.014

pathway differs from other methylotrophic yeasts by providing two dihydroxyacetone synthase (*DAS*) genes.

An important asset of the *P. pastoris* protein expression system besides the strong, methanol inducible promoters is the option to secrete recombinant proteins into the culture medium which is largely devoid of yeast proteins (Mattanovich et al., 2009). Most often, the alpha mating factor pre–pro signal sequence derived from *Saccharomyces cerevisiae* is employed to drive protein secretion in *P. pastoris* (Ghosalkar et al., 2008). The corresponding gene of *P. pastoris* has not yet been identified from the previously published genome sequences.

In 2009, the first draft genome sequence of the most frequently used *P. pastoris* strain GS115, auxotrophic for histidine, was determined (De Schutter et al., 2009). This strain had been created applying nitrosoguanidine mutagenesis to the parental strain NRRL Y-11430 (Cregg et al., 1985). The genome of a second *P. pastoris* strain, CBS704 (=DSMZ70382, Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH), which has been assigned as the European *P. pastoris* type strain also has been sequenced (Mattanovich et al., 2009).

Here, we present the annotated genome of *P. pastoris* CBS7435 generated by using the pyrosequencing (454 Life Sciences, Roche) and the sequencing-by-synthesis (Illumina) methods. The high quality of the genome sequence allowed a detailed description of the MUT pathway. We identified the locus of the *P. pastoris* alpha mating factor and describe its architecture for the first time in detail. Furthermore, this study includes the first assembly and annotation of the mitochondrial genome sequence of a methylotrophic yeast.

2. Materials and methods

2.1. Sequencing of the P. pastoris CBS7435 genome

Genomic DNA of *P. pastoris* CBS7435 was isolated using the Easy-DNA™ Kit (Invitrogen) according to the manufacturer's instructions. DNA identity and integrity were checked by amplification and subsequent sequencing of an 18S rDNA specific region on an ABI 3730xl DNA Analyzer (Life Technologies) using the following primer sequences: f439, 5′-GTTCGATTCCGGAGAGGGA, and r1487, 5′-GGGCATCACAGACCTG. Quality of genomic DNA was assessed by agarose gel electrophoresis and DNA quantity was determined with a fluorescence-based method using the Quant-iT PicoGreen dsDNA kit (Invitrogen) and the Tecan Infinite 200 Microplate Reader (Tecan Deutschland GmbH). A whole-genome-shotgun library was generated using 5 μg of genomic DNA and the GS FLX Titanium General Library Preparation Kit (Roche) according to the protocol provided by the manufacturer. The DNA library was amplified by emulsion PCR and sequenced using the Titanium sequencing chemistry on one half of a PicoTiterPlate (454 Life Sciences, Roche). In order to scaffold the resulting contigs, a mate-pair library with an insert size of 3 kbp was sequenced on two lanes of a flow cell on Illumina's Genome Analyzer IIx system at Ambry Genetics (Aliso Viejo, CA, USA), with a read length of 54 bp. The library was generated from 200 ng of genomic DNA with the Illumina Mate-Pair Kit V1 according to the manufacturer's instructions (Part # 1005363 Rev. B, February 2009).

2.2. Genome assembly and contig scaffolding

Genome assembly was divided into two steps. Creation of a draft genome sequence and improvement of the sequence to get the final consensus sequence. Sequence assembly of the 454 data was performed with default settings using the GS De Novo Assembler software (version 2.3, 454 Life Sciences, Roche Diagnostics, Penzberg, Germany). The resulting contigs were used as a reference for a mapping of the Illumina mate-pair reads with CLC Genomics Workbench (Version 4.0.2, CLC, Aarhus, Denmark). To scaffold the contigs, the CLC mapping was analyzed applying BAMBUS (Pop et al., 2004). As BAMBUS includes repetitive contigs in the scaffolds, its result file was further analyzed with a custom R script (R Development Core Team, Vienna). The position of mate-pair reads connecting non-repetitive contigs was extracted and the gap size was estimated based on the nominal insert size of 3 kbp. Sequence gaps were either closed *in silico* based on the contig graph information created by the GS De Novo Assembler or by Sanger sequencing of specific PCR fragments directly or subsequently to cloning them into pJet1.2blunt (Fermentas). PCR amplifications were performed as conventional or nested reactions using NEB LongAmp Taq DNA Polymerase (Ipswich, MA, USA), Finnzymes Phusion Polymerase (Espoo, Finland), HotStarTaq Polymerase (Qiagen, Hilden, Germany) or Pfu Ultra High Fidelity Polymerase (Stratagene, La Jolla, CA, USA) according to the manufacturers' instructions. Primer sequences used in gap-closure are given in the Supplemental Table. The draft genome sequence of *P. pastoris* CBS7435 was created with a custom R script based on the obtained scaffold information, which was extended with the Sanger contigs from the gap closure.

In the second step, the draft genome was improved to get the final consensus sequence. All available sequences, i.e. 454, Illumina and Sanger, were mapped to the draft sequence with CLC Genomics Workbench to produce an improved consensus sequence. Conflicts were solved by counting instances of each nucleotide and then letting the majority decide the nucleotide in the consensus sequence. Differences between the draft and the consensus sequence were manually compared with a mapping of the Illumina reads with SARUMAN (Blom et al., 2011) and with an improved *de novo* assembly (454 and Sanger reads) in CONSED (Gordon et al., 1998). Differences not supported by the manual inspection were corrected, resulting in the final high-quality genome sequence.

2.3. Gene prediction and sequence annotation

Regional and functional annotation of the final genome sequence of *P. pastoris* CBS7435 was performed using the yeast annotation platform RAPYD (Schneider et al., 2010) that is based on the automated genome annotation program GenDB (Meyer et al., 2003). The prediction of coding sequences was done using the *ab initio* gene prediction program AUGUSTUS version 2.3.1 (Stanke and Waack, 2003) applying the pre-computed training data set of *Schizosaccharomyces pombe* supplied with the tool. In addition, tRNA and rRNA predictions were performed by ARAGORN (Laslett and Canback, 2004) and RNAmmer (Lagesen et al., 2007), respectively. The Metanor-Euk pipeline within RAPYD was used for the automatic functional annotation of predicted coding sequences. This pipeline includes the computation of BLAST runs vs. the NCBI (nt and nr), Swissprot (Boutet et al., 2007), KEGG (Ogata et al., 1999), COG (Tatusov et al., 2003) and the CDD databases (Marchler-Bauer et al., 2009), as well as searches against the InterPro database (Hunter et al., 2009). Furthermore, Hidden Markov model based sequence analysis was performed against the PFAM (Finn et al., 2010) and the TIGRFAMs database (Haft et al., 2003). Signal peptides were predicted using SignalP (Bendtsen et al., 2004) and transmembrane helices by TMHMM (Chaudhuri et al., 2008).

Gene prediction of *P. pastoris* CBS7435 mitochondrial coding sequences was performed using the open reading frame (ORF) prediction tool of SeqBuilder (DNASTAR Lasergene®,