

Genome sequence of the recombinant protein production host Pichia pastoris

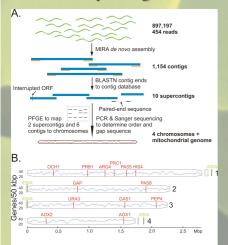
Kristof De Schutter^{1*}, <u>Yao-Cheng Lin^{2*}</u>, Petra Tiels^{1*}, Annelies Van Hecke¹, Lieven Sterck², Thomas Abeel², Sascha Glinka³, Jacqueline Weber-Lehmann³, Pierre Rouzé², Yves Van de Peer² & Nico Callewaert¹ e-mail: yalin@psb.vib-ugent.be

1 Unit for Molecular Glycobiology, VIB Department for Molecular Biomedical Research, UGhent, Belgium 2 Division of Bioinformatics and Systems Biology, VIB Department of Plant Systems Biology, UGhent, Belgium. 3 Eurofins MWG Operon, Ebersberg, Germany. (* These authors contributed equally to this work)

Introduction

The methylotrophic yeast Pichia pastoris is widely used for the production of proteins and as a model organism for studying peroxisomal biogenesis and methanol assimilation. P. pastoris strains capable of human-type N-glycosylation are now available, which increases the utility of this organism for biopharmaceutical production. Despite its biotechnological importance, relatively few genetic tools or engineered strains have been generated for P. pastoris. To facilitate progress in these areas, we present the 9.43 Mbp genomic sequence of the GS115 strain of P. pastoris. We also provide manually curated annotation for its 5,313 protein-coding genes. (Genome portal http://bioinformatics.psb.ugent.be/webtools/bogas/)

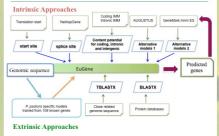
Genome Sequencing and Assembly



Pichia pastoris genome sequencing and overview. (a) Genome sequencing and assembly strategy. (b) P. pastoris gene density and known markers position. Gene density is plotted as a histogram, showing a uniform distribution of genes across each chromosome. The gene density is calculated in a window size of 50 kbp with 5 kbp sliding window. Genes that had been previously mapped to the chromosomes through PFGE are indicated in red, and rDNA repeats in green.

Automatic Gene Prediction

EuGène: integrated gene prediction platform



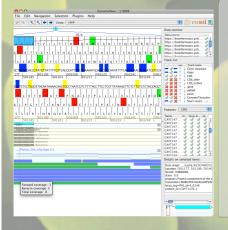
A schematic representation of the EuGéne integrated gene prediction platform. The species specific prediction parameters could be trained based on the known gene structure of 108 P. pastoris genes to improve the gene prediction accuracy.

bioinformatics

Gent

Genome Portal and Next-Gen Tool

Pichia pastoris		Protein Domains \varTheta	
0.03	Showing region from base position 497561 to 503909 (6.4 Kts) ID D		PR023445
Locus ID	chr4 0246	External ID	IPR013845
Functional Description	Protein component of the small (40S) ribosomal subunit	From Database	InterPro
ene Type	mRNA	Description	Ribosomal protein S4e, central
ontig	chr4	External ID	IPR005824
ist Modified On	06 January 2009 17h23 Select a date	From Database	InterPro
istory w	Select a date	Description	KOW
1	select a date	External ID	IPR002942
	Editing is disabled	From Database	InterPro
	Europis osabed	Description	RNA-binding S4
otator \varTheta		External ID	IPR013843
uno. 0		From Database	InterPro
пе	Kristof De Schutter	Description	Ribosomal protein S4e, N-terminal
ail .	kristofds@dmbr.ugent.be	External ID	IPR013844
ib atus	DMBR	From Database	InterPro
tus	active	Description	Ribosomal protein S4e, N-terminal and RNA-binding
e Function O		External ID	IPR000876
		From Database	InterPro
ort Name	n/a	Description	Ribosomal protein S4e
ernative Names	n/a		
initional Functional	Protein component of the small (40S) ribosomal subunit		
scription	YHR203C: RPS4B SGDID:S000001246. Chr VIII from		
	505530-505517.505247-504476, reverse complement, Verified ORF.		
	"Protein component of the small (40S) ribosomal subunit; identical to		
	Rps4Ap and has similarity to rat S4 ribosomal protein*		



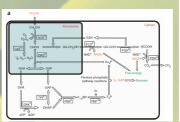
The BioinformaticsGent Online Genome Annotation System (BOGAS) provides a workspace of gene structure and functional annotation for expert annotators. This system is the backbone for tracking the manual annotation process and is an entry point to access the genome information. The GenomeView with genome browsing and annotation functionality enables the single-base resolution of short read data.

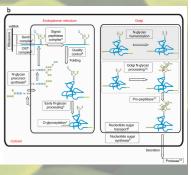
Genome Contents Overview

Size (Mbp): 9.3 (not including rDNA loci, estimated at 0.12 Mbp)	Coding genes: 5,313	tRNA genes: 123	Size (bp): 36, 119
Genome GC content (%): 41.1 Assembled chromosomes: 4	Coding (%): 79.6 Coding GC (%): 41.6 Mean gene length (bp): 1,442 Single exon genes: 4,680	5s rRNA genes: 21	Genome GC content (%): 22 Coding genes: 16 tRNA genes: 31

Pichia pastoris Pathways

Methanol utilization pathway (a) and Protein secretion pathway (b) in Pichia pastoris





Conclusion & Perspectives

• The pyrosequencing technique and a proper assembly strategy can deliver a reliable eukaryotic genome assembly.

• The development of the assembly, annotation and genome visualization tools will improve the usefulness of new sequencing data.

• The wealth of information provided by a full genome sequence will enable a more rapid development of P. pastoris as a protein expression host, building on its exceptional natural capacity for heterologous protein production.

