nature biotechnology

Genome sequence of the recombinant protein production host *Pichia pastoris*

Kristof De Schutter[1,2,7], Yao-Cheng Lin[3,4,7], Petra Tiels[1,5,7], Annelies Van Hecke[1,5], Sascha Glinka[6], Jacqueline Weber-Lehmann[6], Pierre Rouzé[3,4], Yves Van de Peer[3,4] & Nico Callewaert[1,5]

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.

The methylotrophic yeast *Pichia pastoris* is by far the most commonly used yeast species in the production of recombinant proteins[1] and is employed in laboratories around the world to produce proteins for basic research and medical applications. It is also an important model organism for the investigation of peroxisomal proliferation and methanol assimilation. The *P. pastoris* expression technology has been commercially available for many years. *P. pastoris* grows to high cell density, provides tightly controlled methanol-inducible transgene expression and efficiently secretes heterologous proteins in defined media. Several *P. pastoris*–produced biopharmaceuticals that are either not glycosylated (such as human serum albumin[2]) or for which glycosylation is needed only for proper folding (such as several vaccines[3]) are already on the market. An important recent breakthrough has been the development of *P. pastoris* strains with human-type N-glycosylation[4–6]. Humanized glycosylation will further increase the importance of *P. pastoris* for biopharmaceutical production; indeed, proteins produced with this system are moving into clinical development[7]. Moreover, monoclonal antibodies can be made at gram-per-liter scale in the humanized glycosylation-homogenous strains[8].

For further strain engineering, a better understanding of all aspects of the yeast's protein production machinery is needed, and a number of studies relating to *P. pastoris*'s secretory system and engineered promoters have been forthcoming[9,10]. To facilitate the investigation of *P. pastoris* and other methylotrophic yeasts, we present the 9.43 Mbp genomic sequence of the GS115 strain of *P. pastoris*.

RESULTS

Genome sequencing and assembly

Very little is known about the genomic features of *P. pastoris*. The *P. pastoris* genome has been shown to be organized in four chromosomes with a total estimated size of 9.7 Mbp by pulsed-field gel electrophoresis[11]. In addition they assigned 13 *P. pastoris* genes to the different chromosomes. The absence of a genetic map makes chromosome assembly a challenging task, which we completed according to the strategy outlined in **Figure 1a**. We made use of 454/Roche sequencing[12] (GS-FLX version) to highly oversample the genome (20× coverage) and generated 70,500 paired-end sequence tags, to enable the assembly of all but seven contigs into nine 'supercontigs' (plus the mitochondrial genome) using automated shotgun assembly and BLASTN-based contig end-joining (Online Methods and **Supplementary Fig. 1** online). Upon assigning these (super)contigs to the four chromosomes (Online Methods and **Supplementary Fig. 2** online), the order of the supercontigs was determined through PCR and Sanger sequencing of the amplification products. These finishing experiments allowed the reconstruction of the four chromosomal sequences (**Fig. 1b** and **Table 1**), with only two gaps remaining (one each on chromosomes 1 and 4). A ribosomal DNA (rDNA) repeat sequence was present in the assembly as a separate contig of 7,450 bp, with exceptionally high coverage (328.8-fold). Given that sequence coverage all over our assembly very closely approximates 20×, we interpret that there are ~16 copies of the rDNA repeat region, thus accounting for about 119 kbp in sequence. We detected these rDNA loci on all chromosomes (Online Methods, **Fig. 1b** and **Supplementary Fig. 2**). The rDNA locus contains the 18S, 5.8S and 26S rRNA coding sequences. Unlike the *Saccharomyces cerevisiae* 5S rRNA gene, which is localized to the repeated rDNA locus, the 21 copies of the *P. pastoris* 5S rRNA are spread across the entire length of all chromosomes. Based on pulsed-field gel electrophoresis (PFGE), the chromosomes of *P. pastoris* GS115 were estimated to be 2.9, 2.6, 2.3 and 1.9 Mbp[11], whereas we obtained 2.88 (2.8 + 0.08), 2.39, 2.24

[1]Unit for Molecular Glycobiology, Department for Molecular Biomedical Research, VIB, Ghent-Zwijnaarde, Belgium. [2]Department for Biomedical Molecular Biology, Ghent University, Ghent- Zwijnaarde, Belgium. [3]Department of Plant Systems Biology, VIB, Ghent-Zwijnaarde, Belgium. [4]Department of Plant Biotechnology and Genetics, Ghent University, Ghent, Belgium. [5]Unit for Molecular Glycobiology, L-ProBE, Department of Biochemistry and Microbiology, Ghent University, Ghent-Zwijnaarde, Belgium. [6]Eurofins MWG Operon, Ebersberg, Germany. [7]These authors contributed equally to this work. Correspondence should be addressed to N.C. (Nico.Callewaert@dmbr.vib-UGent.be).

Received 1 April; accepted 6 May; published online 24 May 2009; doi:10.1038/nbt.1544

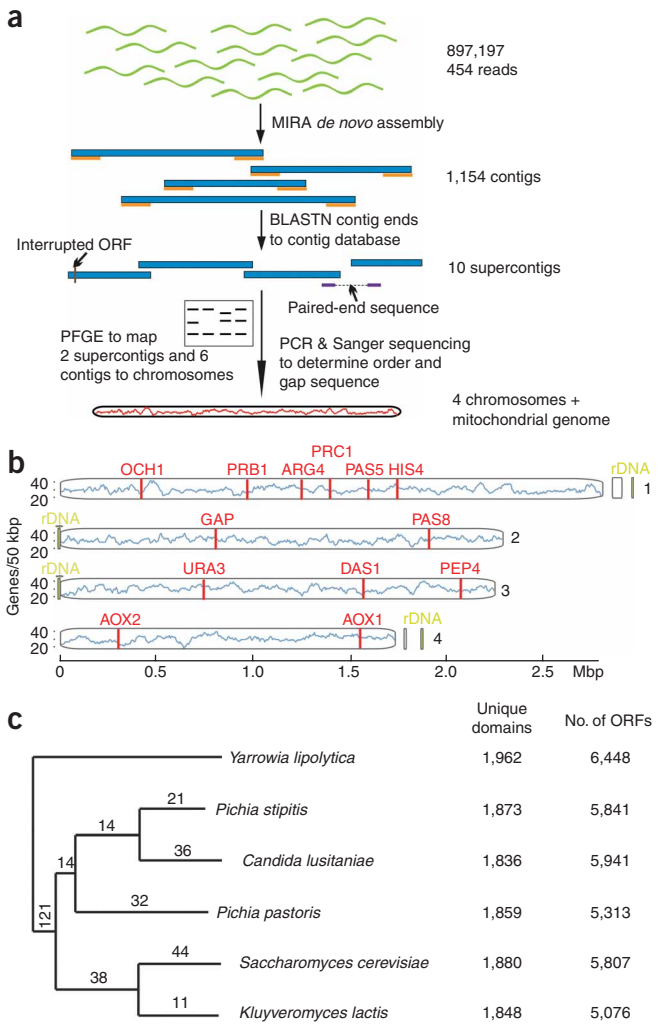

Figure 1 *Pichia pastoris* genome sequencing and overview. (**a**) Genome sequencing and assembly strategy. (**b**) *P. pastoris* gene density and known markers position. Gene density is plotted as a histogram, showing a uniform distribution of genes across each chromosome. The gene density is calculated 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. (**c**) Phylogenetic tree. The phylogenetic tree was built on the concatenated sequence of 200 single-copy orthologous genes in all of the six species. Numbers next to each branch correspond to the number of Pfam domains uniquely present in the corresponding lineage.

and 1.8 (1.78 + 0.017) Mbp after assembly (assembled chromosome + assigned contig). Including the estimated 0.12 Mbp of rRNA repeats, we calculate a genome size of 9.43 Mbp.

Genome sequence accuracy estimation

A concern with genome sequences largely generated through 454 sequencing is the potential for 'indel errors' at homopolymeric sequences[13]. An analysis of the occurrence of such sequences in the *P. pastoris* genome is provided in **Supplementary Figure 3** online. Two approaches were followed to estimate the accuracy of our genome sequence. First, we retrieved 39 peer-reviewed Genbank coding sequences of *P. pastoris* strain GS115 (**Supplementary Table 1** online; total sequence length 70,295 bp). These sequences were compared to our genome sequence, and 84 differences were encountered. To establish which sequence was correct, we amplified these genes by PCR and Sanger-sequenced the PCR products. In all but two cases, the Sanger sequences confirmed our genome sequence, and we thus estimate the error rate to be 1 in 35,147 bp. In an alternative approach, we analyzed all open reading frames (ORFs) encoding proteins with at least one clear homolog in the databases. Where we found an interrupted ORF with clear homology to the 5′ part of the homologs, immediately followed by a coding sequence with clear homology to the 3′ part, the most logical interpretation was that there was a frameshift error mutation in our genome sequence (that is, both coding sequences are extremely likely to be linked into one open reading frame (ORF)). We found such frameshift errors in 2.7% (108) of the 3,997 genes for which such analysis could be made, totaling 6.11 Mbp of coding sequence. Conservatively estimating that we would only have detected such error if it occurred in the first two-thirds of the ORF, we then calculated a frameshift error rate in the coding sequences of 1 in 37,716 bp. Both estimates show that high-coverage 454 sequencing can indeed yield highly accurate genome sequences.

Pichia pastoris phylogenetic position

Phylogenetic analysis (**Fig. 1c**; Online Methods) shows that *P. pastoris* diverged before the formation of the CTG clade (yeasts which translate the CUG codon into serine instead of leucine[14]).

Genome sequence annotation: protein-coding genes

Protein-coding genes were automatically predicted using EuGène[15] (Online Methods and **Supplementary Fig. 4** online). The gene models were manually curated for functional annotation, accurate translational start-and-stop assignment, and intron location. This resulted in a 5,313 protein-coding gene set of which 3,997 (75.2%) have at least one homolog in the National Center for Biotechnology Information protein database (BLASTP e-value 1e-5, sequence length $\leq$20% difference and sequence similarity $\geq$50%). The protein-coding genes occupy 80% of the genome sequence. According to recently proposed measures for genome completeness, we searched the genome for highly conserved single (or low) copy gene sets: core eukaryotic genes (CEGs) with 248 genes across six model organisms[16] and FUNYBASE[17] with 246 genes with orthologs in 21 fungi. All genes from both gene sets were present in our proteome with full domain coverage.

We assigned 1,285 genes to the Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathways, and 4,262 of the genes were annotated with Gene Ontology (GO) terms[18]. The GO slim categories of *P. pastoris* are presented in **Supplementary Figure 5** online. A secretion signal peptide was predicted in 9% of the genes[19], and 4,274 of proteins contain InterPro domains. These include 2,320 distinct Pfam domains. In comparing the presence and absence of protein domains with five other yeast proteomes, 32 domains in 32 genes are identified as specific to *P. pastoris* (**Supplementary Table 2** online). The two fungi in the CTG clade whose genomes have been sequenced (*P. stipitis* and *C. lusitaniae*) share 71 gene families that are absent in *P. pastoris* (**Supplementary Table 2**).

Codon (pair) optimization of transgenes to the expression host organism often yields substantial improvements in recombinant protein yield[20]. *P. pastoris*'s codon usage is shown in **Figure 2a**, which will guide synthetic gene design for protein production in this organism. Overall, the codon usage is similar to the one for *S. cerevisiae*. Some synonymous codon pairs are also more or less frequently used than expected (the codon pair bias)[21]. As reported for

Table 1 Genome sequencing and assembly statistics and contents overview

a. Genome sequencing and assembly statistics

454 Sequencing						
Sequenced reads		Sequenced length (bp)		Paired-end reads		
897,197		218,602,026		11,538		
MIRA assembly						
Assembled reads	Assembled contigs	Contigs (>500 bp)	Length (bp)	N50	L50	Average coverage
885,659	1,154	230	9,658,092	40	77	20
Contig joining			**Chromosomes**			
Joined contigs	Supercontigs	Length (Mbp)				
203	10	9.3	4			

b. Genome contents overview

General information	Coding genes	RNA genes	Mitochondrial genome
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	Coding (%): 79.6	5S rRNA genes: 21	Genome GC content (%): 22
Assembled chromosomes: 4	Coding GC (%): 41.6		Coding genes: 16
	Mean gene length (bp): 1,442		tRNA genes: 31
	Single exon genes: 4,680		

N50, number of contigs that collectively cover at least 50% of the assembly. L50, length of the shortest contig among those that collectively cover 50% of the assembly.

S. cerevisiae[22], under-represented and over-represented codon pair clusters were observed (**Fig. 2b**). It remains untested in *P. pastoris* whether optimizing genes to this codon pair bias results in higher protein expression levels.

Genome sequence annotation: tRNA genes

tRNA coding genes were automatically predicted and manually confirmed by BLASTN with *S. cerevisiae* homologs, which identified 123 nuclear tRNA genes (**Supplementary Table 3** online), compared

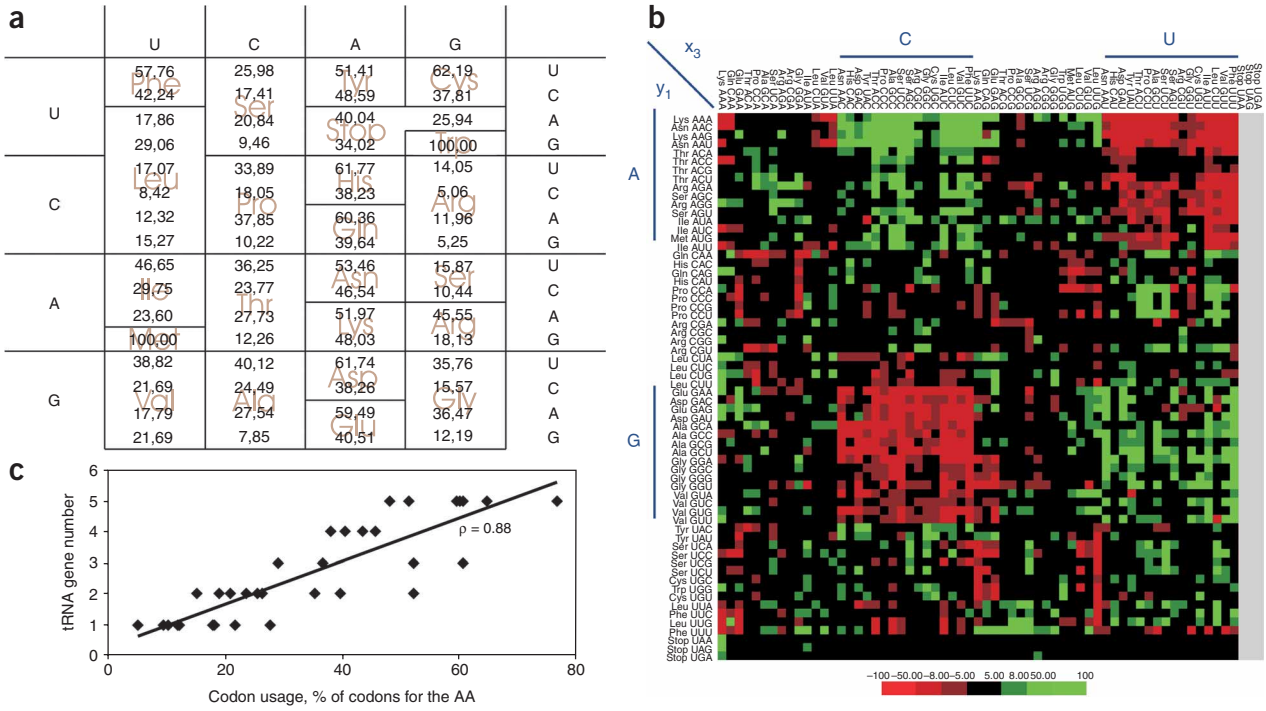

Figure 2 *Pichia pastoris* codon usage. (**a**) Codon usage. Codon usage in the *P. pastoris* ORFeome. The relative abundance of a codon is represented as a percentage of the total codon usage for the amino acid. (**b**) Codon pair usage. Codon pair residual values for *P. pastoris*. The horizontal and vertical axis show, respectively, the 5′ P-site and 3′ A-site codon. Each pixel represents a codon pair residual value. Favored codon pairs are represented in green, under-represented pairs in red. Grouping codon pairs by the x_3 and y_1 nucleotides in the $x_1x_2x_3$ and $y_1y_2y_3$ codon pair reveals over- and under-represented clusters. (**c**) Correlation of tRNA genes and codon usage. Graph shows correlation between the codon usage in relation to the number of genes coding for tRNAs recognizing this codon (Spearman $\rho = 0.88$, $P < 0.0001$).

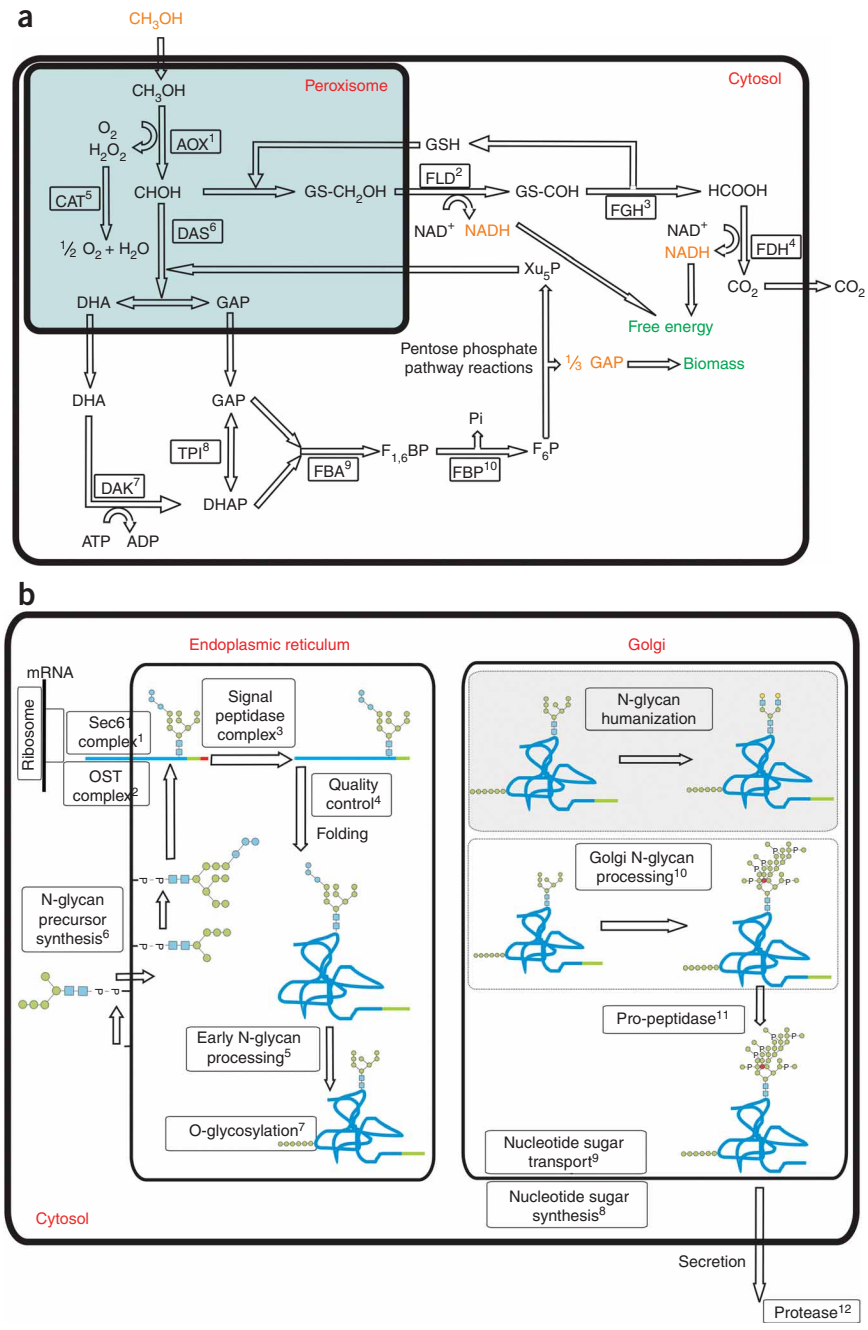

Figure 3 *Pichia pastoris* pathways. (**a**) Methanol utilization pathway in *Pichia pastoris*. A detailed table with the genes coding for the respective enzymes is shown in **Supplementary Table 4a**. [1]AOX, alcohol oxidase; [2]FLD, formaldehyde dehydrogenase; [3]FGH, S-formylglutathione hydrolase; [4]FDH, formate dehydrogenase; [5]CAT, catalase; [6]DAS, dihydroxyacetone synthase; [7]DAK, dihydroxyacetone kinase; [8]TPI, triosephosphate isomerase; [9]FBA, fructose-1,6-bisphosphate aldolase; [10]FBP, fructose-1,6-bisphosphatase; DHA, dihydroxyacetone; GAP, glyceraldehyde-3-phosphate; DHAP, dihydroxyacetone phosphate; $F_{1,6}BP$, fructose-1,6-bisphosphate; F_6P, fructose-6-phosphate; P_i, phosphate; Xu_5P, xylulose-5-phosphate; GSH, glutathione. (**b**) Protein secretion pathway. Schematic representation of the secretion pathway in *P. pastoris*. A detailed table with the genes coding for the components involved in the represented complexes or processes is shown in **Supplementary Table 4b**. The nascent protein is translocated to the ER by the Sec61 complex, and N-glycosylation sites are glycosylated with the dolichol-linked $Glc_3Man_9GlcNAc_2$ oligosaccharide precursor by the OST complex. After processing of the signal peptide, the protein is folded with the aid of chaperones. ER N-glycan processing results in $Man_8GlcNAc_2$ type glycan. O-glycosylation is also initiated in the ER by the protein-*O*-mannosyltransferases. After transport to the Golgi apparatus, the N-glycans are further processed to the yeast-typical hypermannosyl-type glycans. In strains with humanized glycosylation pathways,[4–6] the hypermannosylation is abolished and the glycans are processed to $Gal_2GlcNAc_2Man_3GlcNAc_2$. After processing of the pro-domain, the protein is secreted in the growth medium, where it may be a substrate for yeast proteases.

to 274 in the *S. cerevisiae* genome[23]. *P. pastoris* has three tRNA families not present in *S. cerevisiae* (tR(UCG), tL(CAG) and tP(CGG)), but also lacks one tRNA family (tL(GAG)).

Notably, a positive correlation was found between the number of tRNA genes for a given codon and the frequency of use of this codon (Spearman $\rho = 0.88$; $P < 0.0001$, **Fig. 2c**).

DISCUSSION

The genomic sequence of *P. pastoris* presented here will facilitate the development of improved strains with customized properties for high-yield protein production with defined post-translational modifications. Promising targets for genetic engineering include inducible promoters for transgene expression, chaperones that assist protein folding, proteins involved in the secretory pathway and enzymes catalyzing protein glycosylation, proteolytic processing and other post-translational modifications.

The commonly used methanol-inducible promoters in *P. pastoris*—the alcohol oxidase I promoter[10,24] and the formaldehyde dehydrogenase promoter[25]—drive the production of enzymes needed for methanol assimilation and therefore produce extremely high levels of these transcripts upon switching the carbon source to methanol. The genome sequence has allowed identification of all genes coding for enzymes involved in methanol assimilation and their promoters (**Fig. 3a** and **Supplementary Table 4a** online), which can now be studied for their suitability for transgene expression in *P. pastoris*. A first comparative analysis of these promoters did not reveal obvious commonalities in sequence motifs or promoter organization (data not shown).

Secretion of heterologous proteins rather than cytoplasmic accumulation is most often the preferred option in *Pichia*-based production processes. The yeast secretory system (overview in **Fig. 3b**; **Supplementary Table 4b** summarizes the genes discussed in the remainder of the text) is thus an important engineering target to obtain optimized strains that are capable of folding and processing a large flux of recombinant protein. However, many aspects of the secretory pathway are insufficiently characterized. For example, the

knowledge on the *Pichia* chaperones is incomplete, and we here provide the complete catalog of orthologs to the *S. cerevisiae* endoplasmic reticulum (ER) folding machinery, which should enable more efficacious folding-system engineering in the future[26].

The heterologous preprot signal sequence of the *S. cerevisiae* alpha-mating factor is most often used to induce Sec61p-mediated translocation of the protein into the endoplasmic reticulum of *P. pastoris* (http://faculty.kgi.edu/cregg/). This signal sequence works in most cases, although there have been almost no studies to compare it to other signal sequences. Moreover, the Kex2p/Ste13p-mediated processing of the propeptide in this *S. cerevisiae* sequence is often problematic in *Pichia*[27], resulting in nonnative amino acids at the N-terminus of the heterologous protein. The genome sequence now reveals a multitude of endogenous signal sequences (**Supplementary Fig. 6** online shows a subset of such signal sequences, derived from homologs of functionally annotated secreted *S. cerevisiae* proteins). This database of secretion signals will allow screening for the optimal signal-ORF combination, which may result in augmented protein expression levels. Multiple sequence alignment also allowed derivation of a consensus signal sequence (**Supplementary Fig. 6**), which may be suited for mediating heterologous protein secretion.

The secretory system is also the site of post-translational modification (especially glycosylation), and yeasts differ substantially from higher eukaryotes in this respect. In terms of N-glycosylation, yeasts such as *P. pastoris* modify proteins with a range of heterogenous high-mannose glycans[28], which introduce a large amount of heterogeneity in the protein (reducing downstream processing efficiency and complicating product characterization) and induce fast clearance from the bloodstream. The highly immunogenic terminal alpha-1, 3-mannosyl glycotopes that are abundantly produced by *S. cerevisiae* are not detected on *Pichia*-produced glycoproteins[29]. Indeed, we did not find an ortholog of the *S. cerevisiae* gene *MNN1* (encoding the alpha-1,3-mannosyltransferase) in the *Pichia* genome. However, *Pichia* glycoproteins can in some cases be modified with β-1,2-mannose residues[30], reminiscent of antigenic epitopes on the *Candida albicans* cell wall[31]. We find the patented *P. pastoris* AMR2 β-mannosyltransferase in the genome, and three homologs, thus providing the basis for reducing the levels of this undesired glycan modification.

To overcome the difficulties with *Pichia*'s glycosylation, strains have been developed with an entirely re-engineered glycosylation pathway to produce human IgG–type N-glycans (N-glycosylation humanization technology; **Fig. 3b**)[4–6]. The heterologous glycosyltransferases needed for this use the sugar-nucleotides UDP-GlcNAc and UDP-Gal as monosaccharide donors. Although UDP-GlcNAc is synthesized in yeasts for the synthesis of cell wall chitin (we have identified a UDP-GlcNAc transporter in the genome), no galactosylated glycoconjugates in *P. pastoris* have been described. We have shown previously that the mere overexpression of a *Pichia* Golgi-targeted version of human β-1,4-galactosyltransferase I is sufficient to achieve galactosylation of secreted glycoproteins, indicating that *Pichia* produces UDP-Gal and transports it into the Golgi apparatus[32]. Indeed, we now find an endogenous cytoplasmic UDP-Glc-4-epimerase and clear homologs of Golgi UDP-Galactose transporters in the *P. pastoris* genome (**Supplementary Table 4b**). These findings are relevant to glycan engineering in this yeast as researchers have previously overexpressed a heterologous UDP-Glc-4-epimerase in fusion to the galactosyltransferase to achieve higher levels of UDP-Gal in the yeast Golgi apparatus[6,33].

Yeasts also O-glycosylate secreted proteins with oligomannosyl-glycans that differ from the mucin-type O-glycosylation in humans[34]. No robust engineering approach has yet been developed to overcome this issue. The identification of the *Pichia* protein-O-mannosyltransferases that initiate this modification in the ER in the genome will help toward this goal.

Finally, an often-observed problem is degradation of the product by endogenous proteases. If the heterologous protein is toxic to the cell, much of this proteolytic activity can be of vacuolar origin (released in the growth medium upon cell lysis), but *Pichia* also expresses secreted proteases. It would be of great interest to have a panel of *P. pastoris* strains in which the most active proteases had been disrupted. Only few such strains are currently available because knowledge on the protease gene sequences was unavailable. We here provide a catalog of the *Pichia* vacuolar and secreted proteases (**Supplementary Table 4b**), which will speed up the development of protease-deficient strains.

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. With a large academic and industrial user base, human-type N-glycosylation already in place, gram-per liter monoclonal antibody production recently reported[8] and the genome now publicly available, the stage is set for *Pichia pastoris* to become an even more important expression system for biopharmaceutical proteins.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession numbers. The *P. pastoris* genomic sequence has been deposited in the EMBL Nucleotide Sequence Database (Accession numbers FN392319–FN392325).

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This research was supported by a Marie Curie Excellence Grant to N.C. (EU-FP6), IUAP P6/25 (BioMaGNet) and by the Fund for Scientific Research-Flanders (FWO). Y.-C.L. is supported by the EVOLTREE (EU-FP6) fellowship. Contributions towards the sequencing cost were received from VIB and from Research Corporation Technologies. We thank Lieven Sterck and Kenny Billiau for setup and maintenance of the BOGAS annotation portal and Cindy Martens for the Dollo parsimony analysis. We thank Mark Veugelers and Jo Bury for continuous support of this project.

AUTHOR CONTRIBUTIONS

K.D.S. and P.T. assembled and finished the genome sequence, manually curated the computer-generated annotation, analyzed the annotation and wrote parts of the manuscript. Y.-C.L. performed all post-shotgun assembly bio-informatics aspects of the study under guidance of Y.V.d.P. and P.R. and wrote parts of the manuscript. A.V.H. assisted in gap closure and in determining sequence accuracy. S.G. performed the 454/Roche sequencing and J.W.-L. processed the raw data and performed shotgun assembly and contig scaffolding. Both provided the corresponding methods sections of the manuscript. N.C. designed and coordinated the study, initiated the BLAST-based contig joining approach and wrote parts of the manuscript.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Cregg, J.M., Cereghino, J.L., Shi, J. & Higgins, D.R. Recombinant protein expression in *Pichia pastoris*. *Mol. Biotechnol.* **16**, 23–52 (2000).
2. Watanabe, H. *et al.* In vitro and in vivo properties of recombinant human serum albumin from *Pichia pastoris* purified by a method of short processing time. *Pharm. Res.* **18**, 1775–1781 (2001).
3. Hardy, E. *et al.* Large-scale production of recombinant hepatitis B surface antigen from *Pichia pastoris*. *J. Biotechnol.* **77**, 157–167 (2000).

4. Hamilton, S.R. *et al.* Humanization of Yeast to Produce Complex Terminally Sialylated Glycoproteins. *Science* **313**, 1441–1443 (2006).
5. Hamilton, S.R. & Gerngross, T.U. Glycosylation engineering in yeast: the advent of fully humanized yeast. *Curr. Opin. Biotechnol.* **18**, 387–392 (2007).
6. Jacobs, P.P., Geysens, S., Vervecken, W., Contreras, R. & Callewaert, N. Engineering complex-type N-glycosylation in *Pichia pastoris* using GlycoSwitch technology. *Nat. Protoc.* **4**, 58–70 (2009).
7. Ratner, M. Pharma swept up in biogenerics gold rush. *Nat. Biotechnol.* **27**, 299–301 (2009).
8. Potgieter, T.I. *et al.* Production of monoclonal antibodies by glycoengineered *Pichia pastoris*. *J. Biotechnol.* **139**, 318–325 (2009).
9. Mogelsvang, S., Gomez-Ospina, N., Soderholm, J., Glick, B.S. & Stachelin, L.A. Tomographic evidence for continuous turnover of Golgi cisternae in *Pichia pastoris*. *Mol. Biol. Cell* **14**, 2277–2291 (2003).
10. Hartner, F.S. *et al.* Promoter library designed for fine-tuned gene expression in *Pichia pastoris*. *Nucleic Acids Res.* **36**, e76 (2008).
11. Ohi, H., Okazaki, N., Uno, S., Miura, M. & Hiramatsu, R. Chromosomal DNA patterns and gene stability of *Pichia pastoris*. *Yeast* **14**, 895–903 (1998).
12. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
13. Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L. & Welch, D.M. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* **8**, R143 (2007).
14. Fitzpatrick, D.A., Logue, M.E., Stajich, J.E. & Butler, G. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol. Biol.* **6**, 99 (2006).
15. Foissac, S. *et al.* Genome annotation in plants and fungi: EuGene as a model platform. *Current Bioinformatics* **3**, 89–97 (2008).
16. Parra, G., Bradnam, K., Ning, Z., Keane, T. & Korf, I. Assessing the gene space in draft genomes. *Nucleic Acids Res.* **37**, 289–297 (2009).
17. Marthey, S. *et al.* FUNYBASE: a FUNgal phYlogenomic dataBASE. *BMC Bioinformatics* **9**, 456 (2008).
18. Schmid, R. & Blaxter, M. annot8r: GO, EC and KEGG annotation of EST datasets. *BMC Bioinformatics* **9**, 180 (2008).
19. Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* **2**, 953–971 (2007).
20. Hu, S. *et al.* Codon optimization, expression, and characterization of an internalizing anti-ErbB2 single-chain antibody in *Pichia pastoris*. *Protein Expr. Purif.* **47**, 249–257 (2006).
21. Gutman, G.A. & Hatfield, G.W. Nonrandom utilization of codon pairs in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **86**, 3699–3703 (1989).
22. Friberg, M., von Rohr, P. & Gonnet, G. Limitations of codon adaptation index and other coding DNA-based features for prediction of protein expression in *Saccharomyces cerevisiae*. *Yeast* **21**, 1083–1093 (2004).
23. Hani, J. & Feldmann, H. tRNA genes and retroelements in the yeast genome. *Nucleic Acids Res.* **26**, 689–696 (1998).
24. Tschopp, J.F., Brust, P.F., Cregg, J.M., Stillman, C.A. & Gingeras, T.R. Expression of the lacZ gene from two methanol-regulated promoters in *Pichia pastoris*. *Nucleic Acids Res.* **15**, 3859–3876 (1987).
25. Shen, S., Sulter, G., Jeffries, T.W. & Cregg, J.M. A strong nitrogen source-regulated promoter for controlled expression of foreign genes in the yeast *Pichia pastoris*. *Gene* **216**, 93–102 (1998).
26. Gasser, B., Sauer, M., Maurer, M., Stadlmayr, G. & Mattanovich, D. Transcriptomics-based identification of novel factors enhancing heterologous protein secretion in yeasts. *Appl. Environ. Microbiol.* **73**, 6499–6507 (2007).
27. Prabha, L., Govindappa, N., Adhikary, L., Melarkode, R. & Sastry, K. Identification of the dipeptidyl aminopeptidase responsible for N-terminal clipping of recombinant Exendin-4 precursor expressed in *Pichia pastoris*. *Protein Expr. Purif.* **64**, 155–161 (2009).
28. Grinna, L.S. & Tschopp, J.F. Size distribution and general structural features of N-linked oligosaccharides from the methylotrophic yeast, *Pichia pastoris*. *Yeast* **5**, 107–115 (1989).
29. Bretthauer, R.K. & Castellino, F.J. Glycosylation of *Pichia pastoris*-derived proteins. *Biotechnol. Appl. Biochem.* **30**, 193–200 (1999).
30. Mille, C. *et al.* Identification of a new family of genes involved in beta-1,2-mannosylation of glycans in *Pichia pastoris* and *Candida albicans*. *J. Biol. Chem.* **283**, 9724–9736 (2008).
31. Dalle, F. *et al.* Beta-1,2- and alpha-1,2-linked oligomannosides mediate adherence of *Candida albicans* blastospores to human enterocytes in vitro. *Infect. Immun.* **71**, 7061–7068 (2003).
32. Vervecken, W. *et al.* *In vivo* synthesis of mammalian-like, hybrid-type N-glycans in *Pichia pastoris*. *Appl. Environ. Microbiol.* **70**, 2639–2646 (2004).
33. Bobrowicz, P. Engineering of an artificial glycosylation pathway blocked in core oligosaccharide assembly in the yeast *Pichia pastoris*: production of complex humanized glycoproteins with terminal galactose. *Glycobiology* **14**, 757–766 (2004).
34. Trimble, R.B. *et al.* Characterization of N- and O-linked glycosylation of recombinant human bile salt-stimulated lipase secreted by *Pichia pastoris*. *Glycobiology* **14**, 265–274 (2004).

ONLINE METHODS

DNA preparation. The *P. pastoris* GS115 strain (Invitrogen) is derived from the wild-type strain NRRL-Y 11430 (Northern Regional Research Laboratories). It has a mutation in the histinol dehydrogenase gene (HIS4) and was generated by nitrosoguanidine mutagenesis at Phillips Petroleum Co[35]. It is the most frequently used *Pichia* strain for heterologous protein production.

P. pastoris genomic DNA was prepared according to a published protocol[36] with minor modifications. Instead of vortexing, the samples were shaken in a Mixer Mill (Retsch) for 2 min.

Sample preparation and sequencing with Roche/454 Genome Sequencer FLX. The shotgun library of *P. pastoris* for sequencing on the Genome Sequencer FLX (GS FLX) was prepared from 5 μg of intact genomic DNA. Based on random cleavage of the genomic DNA[12] with subsequent removal of small fragments with AMPure SPRI beads (Agencourt), the resulting single-stranded (ss) DNA library showed a fragment distribution between 300 and 900 bp with a maximum of 574 bp. The optimal amount of ssDNA library input for the emulsion PCR[12] (emPCR) was determined empirically through two small-scale titrations leading to 1.5 molecules per bead used for the large-scale approach. A total of 64 individual emulsion PCRs were performed to generate 3,974,400 DNA-carrying beads for two two-region-sized 70 × 75 PicoTiterPlates (PTP) and each region was loaded with 850,000 DNA-carrying beads. Each of the two sequencing runs was performed for a total of 100 cycles of nucleotide flows[12] (flow order TACG), and the 454 Life Sciences/Roche Diagnostics software Version 1.1.03 was used to perform the image and signal processing. The information about read flowgram (trace) data, basecalls and quality scores of all high-quality shotgun library reads was stored in a Standard Flowgram Format (SFF) file which is used by the subsequent computational analysis (see below).

Within this sequencing project, a paired end library of *P. pastoris* (strain GS115) was prepared for subsequent ordering and orienting of contigs (see computational analysis below). Six micrograms of intact genomic DNA was sheared hydrodynamically (Hydroshear, Genomic Solutions) and purified with AMPure™ SPRI beads into DNA fragments ~3 kbp in length. After methylation of EcoRI restriction sites, a biotinylated hairpin adaptor was ligated to the ends of the *P. pastoris* DNA fragments, followed by EcoRI digestion with a subsequent circularization[37]. The restriction of the circularized DNA fragments with MmeI, the subsequent ligation of paired-end adaptors and the amplification of the remaining DNA fragments resulted in a double-stranded paired-end library 130 bp in length. For the following eight individual emPCRs of the paired-end library, 1.5 molecules per bead were used to generate 339,480 DNA-carrying beads of which 280,000 were loaded onto a region of a four-region sized 70 × 75 PTP. The subsequent sequencing run with the GS FLX was performed for a total of 42 cycles of nucleotide flow (see above), and the 454 Life Sciences/Roche Diagnostics software Version 1.1.03 was used to perform the image and signal processing. The information about read flowgram (trace) data, basecalls and quality scores of all high-quality shotgun library reads was also stored in an standard flowgram format file, which is used by the subsequent computational analysis.

Computational analysis of GS FLX shotgun and paired-end reads. An automatic assembly pipeline (in-house software, Eurofins MWG Operon) was used to assemble *de novo* the generated shotgun and paired-end reads.

For *de novo* assembly of the *P. pastoris* genome sequence, a total of 897,197 good quality base-called, clipped shotgun reads with an average read length of 243 bp and a total of 70,500 good quality base-called, clipped 20 bp paired-end tag reads were used.

Within this pipeline, the information about all sequences and their quality was extracted from the SFF-file into a FASTA-file and subsequently converted into CAF format, the input format of choice of the used assembler mira (version 2.9 26×3; http://www.chevreux.org/projects_mira.html) for contig creation. The provided mate and size information (that is, forward and reverse read and the 3 kbp of length) of the paired end reads was used to scaffold the resulting contigs from the *de novo* assembly[38].

Assembly (**Fig. 1a** and **Supplementary Fig. 2**). The initial assembly contained 1,154 contigs with 9.6 Mbp sequence and 20× sequencing depth. The contig N/L50 was 40/77 kbp. Assembly of the contigs was performed manually, based on homology between the contig ends. 13 contigs were assigned to chromosomes by identification of the chromosomal markers previously described[11] (Chromosome 1: HIS4, ARG4, OCH1, PAS5, PRB1, PRC1; Chromosome 2: PAS8, GAP; Chromosome 3: DAS1, URA3, PEP4; Chromosome 4: AOX1, AOX2). Starting from these contigs, contigs with homologous contig ends were identified by BLASTN search with 500–1,000 bp of the contig ends to a database with the contig sequences. Contigs sharing homology with a *P*-value < e-20 are assumed to be linked. Pools of potentially linked contigs were assembled to supercontigs by the SeqMan assembly software (DNASTAR). The resulting contig junctions were curated by removing the low-coverage ends of either joined contig. In the cases where the BLASTN P-value was >e-50, the junction was PCR-amplified and Sanger-sequenced (primer sequences: **Supplementary Table 5** online). This resulted in ten supercontigs, with 9.1 Mbp of sequence and a remaining seven unassembled contigs. The supercontig N/L 50 was 3/1.544 Mbp.

The mitochondrial genome was also assembled and had extremely high coverage (859.9-fold), indicating the presence of ~43 mitochondrial genomes per cell in *P. pastoris* when grown on glucose as a carbon source.

Gap joining and finishing. Supercontigs were linked by mapping contigs to paired-end scaffolds ($n = 1$), and automated prediction of protein-coding sequences revealed a partial ORF at the end of a supercontig, homologous to a WD40 domain protein in other yeasts (including, *Pichia guillermondii* homolog PGUG 04385). Finding the other part of this ORF on one of the unassembled contigs allowed joining of this supercontig to one of the as-yet unassembled contigs. This was confirmed by PCR and Sanger sequencing.

Seven of the nine thus-generated supercontigs could be assigned to a specific chromosome when they contained one or more of the 13 genes for which chromosomal location had been previously established[11] (**Fig. 1b** and **Supplementary Fig. 1c**). For those two supercontigs and the six unassembled contigs where this was not the case, Southern blot analysis of pulsed-field gel electrophoresis-separated *Pichia pastoris* chromosomes (see below) was used for the assignment (**Supplementary Fig. 2**). After assignment to the chromosomes, orientation of the supercontigs and contigs on the chromosomes was determined by PCR analysis with primers on the contig ends (**Supplementary Table 5**). Gaps were PCR-amplified using primers flanking these regions (**Supplementary Table 5**) and sequenced by Sanger sequencing for finishing.

We detected rDNA repeat regions by Southern blot analysis on all four PFGE-separated chromosomes (**Supplementary Fig. 2**). The Southern signal on chromosomes 1 and 4 was as strong as those on chromosomes 2 and 3 combined. Subtelomeric location of rDNA loci is frequent in yeast genomes[39]. Because of their direct repeat character, these loci resist assembly by the current methods[40]. Through PCR, we determined the location and orientation of the rDNA locus at one end of chromosomes 2 and 3 (**Fig. 1b**). Our attempts at verification of the rDNA locus position on chromosomes 1 and 4 (still containing one gap) have so far been inconclusive.

Pulsed-field gel electrophoresis. A BioRad contour-clamped homogenous electric field CHEF DRIII system was used for PFGE. Chromosomal DNA was prepared in agarose plugs with the CHEF Genomic DNA Plug kit (BioRad) following the instructions of the manufacturer. A 0.8% agarose gel in 1× modified TBE (0.1 M Tris, 0.1 M Boric Acid, 0.2 mM EDTA) was used to separate the chromosomes. The gel was electrophoresed with a 106° angle at 14 °C at 3 V/cm for 32 h, with a switch interval of 300 s, followed by 32 h with a switch interval of 600 s and 24 h with a switch interval of 900 s (ref. 11). After separation, the chromosomes were visualized with ethidium bromide, and the different contigs were mapped onto the chromosomes by Southern blot analysis. Therefore, the gel was incubated in 0.25 M HCl for 30 min, followed by capillary alkali transfer of the DNA onto a Hybond N+ membrane (Amersham). The probes were prepared by PCR on an open reading frame. For chromosome specific probes[11], a part of the coding sequence of HIS4 (chromosome 1), GAP (chromosome 2), URA3 (chromosome 3) and AOX1 (chromosome 4) was used. The probes were random labeled with α ^{32}P dCTP, using the High Prime kit (Roche).

Automatic gene structure prediction & functional annotation. Protein-coding genes were predicted by the integrative gene prediction platform

EuGene[15] (**Supplementary Fig. 4**). A specific EuGene version was trained based on 108 manually checked *P. pastoris* genes. Documented genes from *P. stipitis* and *S. cerevisiae* were used to build *P. pastoris* orthologous gene models allowing the training of *P. pastoris*-specific Interpolated Markov Models for coding sequences and introns. Splice sites were predicted by NetAspGene[41] and gene prediction from GeneMarkHMM-ES[42] trained for *P. pastoris* and AUGUSTUS[43] (*Pichia stipitis* model) were used to provide alternative gene models for EuGene prediction. The UniProt and the fungi RefSeq protein database were searched against the supercontig sequence by BLASTX to identify the coding area. We used DeCypher-TBLASTX to search the conserved sequence area between the *P. pastoris, P. stipitis* and *Candida guilliermondii* genomes.

All predicted protein-coding genes were searched against the yeast protein database, UniProt and RefSeq fungi protein database by BLASTP. Protein domains were detected by InterProScan with various databases (BlastProDom, FPrintScan, PIR, Pfam, Smart, HMMTigr, SuperFamily, Panther and Gene3D) through the European Bioinformatics Institute Web Services SOAP-based web tools. Signal peptide and transmembrane helices were predicted by SignalP and TMHMM respectively (http://www.cbs.dtu.dk/services/). GO (Gene Ontology) terms were derived from the InterProScan result and the KEGG (Kyoto Encyclopedia for Genes and Genomes) pathway and EC (Enzyme Commission) numbers were annotated by the annot8r pipeline[18].

Expert gene structure/functional annotation. The gene structure prediction and the database search results from various databases were formatted and stored in a MySQL relational database. A multiple alignment of each protein-coding gene with the top ten best hits against the UniProt, RefSeq fungi and yeast protein database was built by MUSCLE[44]. A BOGAS (Bioinformatics Online Genome Annotation System) *P. pastoris* annotation website was setup as the workspace for expert annotators. The initial aim of BOGAS is to provide a workspace for gene structure and functional annotation. The editing of gene structure or gene function assignment is directly updated to the MySQL relational database through the web interface. All of the modification from expert annotators is traceable and reversible by the database system. Once the expert annotator modifies the gene structure and changes the translated protein product, the system will automatically trigger the update function to check the protein domain and protein database. BOGAS also provides a search function where users can search for genes by sequence similarity (BLAST), gene id, gene name or InterPro domain. Each predicted *Pichia* gene's structure and the similarity search result was visually inspected through an embedded strip-down version of Artemis[45]. The splice sites of each gene were carefully checked and compared with *S. cerevisiae* and *P. stipitis* loci. A functional description of each gene was added to the gene annotation when a closely related homologous gene was available. The result of the annotation effort is available at http://bioinformatics.psb.ugent.be/webtools/bogas/.

Estimate of the gene space completeness. Parra *et al.*[16] proposed a set of core eukaryotic genes (CEGs) to estimate the completeness of genome sequencing and assembly programs. The CEGs contains 248 genes across six model organisms (*Homo sapiens, Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana, S. cerevisiae and Saccharomyces pombe*) of which ~90% are single copy in *D. melanogaster, C. elegans, S. cerevisiae* and *S. pombe*. We checked our protein-coding genes with the HMM profile from the CEGs data set by the HMMER package. All of the 248 CEGs were present in our curated gene set with full HMM domain coverage. On the other hand, FUNYBASE (FUNgal phYlogenomic dataBASE)[17] provides 246 single-copy ortholog clusters in 21 sequenced fungal genomes. We extracted these single-copy protein sequences from the FUNYBASE website and built the HMM model for each cluster. The corrected *P. pastoris* protein sequences were searched with the FUNYBASE HMM database. All of the FUNYBASE models were presented in our gene catalog with complete domain coverage.

Detection of rRNA and tRNA loci. Ribosomal RNAs were detected automatically by INFERNAL 1.0 (INFERence of RNA ALignment) against the Rfam[46] database and manually confirmed by BLASTN search with *S. cerevisiae* homologs to the *P. pastoris* genome sequence. Localization of the rDNA locus was assayed by PFGE and PCR.

Transfer RNAs were automatically predicted by tRNA Scan-s.e.m. 1.21 (ref. 47) and manually confirmed by BLASTN search with the *S. cerevisiae* homologs to the *P. pastoris* genome sequence.

Codon usage. Nucleotide sequences of the predicted *P. pastoris* ORFeome were analyzed with ANACONDA 1.5 (ref. 48). In addition to calculation of the codon use, the analysis by ANACONDA generates a codon-pair context map for the ORFeome. This map shows one colored square for each codon-pair, the first codon corresponds to rows and the second corresponds to columns in the map. Favored codon pairs are shown in green, underrepresented ones are shown in red.

Phylogenetic tree reconstruction of fungal genomes. The phylogenetic tree was based on 200 single-copy genes which were present in 12 sequenced fungal genomes. A multiple sequence alignment was constructed using the MUSCLE program and gap removal by in-house script based on the BLOSUM62 scoring matrix. The maximum likelihood tree reconstruction program TREE-PUZZLE[49] (quartet puzzling, WAG model, estimated gama distribution rate with 1000 puzzling step) was used for phylogenetic tree reconstruction. The tree was well supported by 1,000 bootstraps in each node.

Comparative analysis of gene family and protein domain. The predicted proteomes used in this study were those of six hemiascomycetes (*P. pastoris, S. cerevisiae, K. lactis, P. stipitis, C. lustianiae and Y. lipolytica*)[50,51]. In order to obtain the gene families, a similarity search of all protein sequences from the six fungi (all-against-all BLASTP, e-value 1e-10) was performed. Gene families were constructed by Markov clustering[52] based on the BLASTP result. All predicted protein sequences from the six genomes were searched against the Pfam[53] database to obtain the protein domain occurrence in each species. The protein domain loss and acquisition was counted based on the Dollo parsimony principle by the DOLLOP program from the PHYLIP package[54].

Gene annotation. Available at http://bioinformatics.psb.ugent.be/webtools/bogas/.

35. Cregg, J.M., Barringer, K.J., Hessler, A.Y. & Madden, K.R. *Pichia pastoris* as a host system for transformations. *Mol. Cell. Biol.* **5**, 3376–3385 (1985).
36. Weiss, H.M., Haase, W. & Reilander, H. Expression of an integral membrane protein, the 5HT5A receptor. *Methods Mol. Biol.* **103**, 227–239 (1998).
37. Korbel, J.O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
38. Pop, M., Kosack, D.S. & Salzberg, S.L. Hierarchical scaffolding with Bambus. *Genome Res.* **14**, 149–159 (2004).
39. Venema, J. & Tollervey, D. Ribosome synthesis in *Saccharomyces cerevisiae*. *Annu. Rev. Genet.* **33**, 261–311 (1999).
40. James, S.A. *et al.* Repetitive sequence variation and dynamics in the ribosomal DNA array of *Saccharomyces cerevisiae* as revealed by whole-genome resequencing. *Genome Res.* **19**, 625–635 (2009).
41. Wang, K., Ussery, D.W. & Brunak, S. Analysis and prediction of gene splice sites in four *Aspergillus* genomes. *Fungal Genet. Biol.* **46** Suppl 1, S14–S18 (2009).
42. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. & Borodovsky, M. Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990 (2008).
43. Stanke, M. *et al.* Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006).
44. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
45. Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944–945 (2000).
46. Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124 (2005).
47. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
48. Pinheiro, M. *et al.* Statistical, computational and visualization methodologies to unveil gene primary structure features. *Methods Inf. Med.* **45**, 163–168 (2006).
49. Schmidt, H.A., Strimmer, K., Vingron, M. & von Haeseler, A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502–504 (2002).
50. Rossignol, T. *et al.* CandidaDB: a multi-genome database for *Candida* species and related Saccharomycotina. *Nucleic Acids Res.* **36**, D557–D561 (2007).

doi:10.1038/nbt.1544

51. Jeffries, T. *et al.* Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. *Nat. Biotechnol.* **25**, 319–326 (2007).
52. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575 (2002).
53. Finn, R. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36**, D281 (2008).
54. Felsenstein, J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* **266**, 418–427 (1996).