RESEARCH ARTICLE

# Proteomic analysis and genome annotation of *Pichia pastoris*, a recombinant protein expression host

*Santosh Renuse[1,2], Anil K. Madugundu[1,3], Praveen Kumar[1]*, Bipin G. Nair[2], Harsha Gowda[1], T. S. Keshava Prasad[1,2,3] and Akhilesh Pandey[4]*

[1] Institute of Bioinformatics, International Technology Park, Bangalore, India
[2] Amrita School of Biotechnology, Amrita Vishwa Vidyapeetham, Kollam, India
[3] Centre of Excellence in Bioinformatics, School of Life Sciences, Pondicherry University, Puducherry, India
[4] McKusick-Nathans Institute of Genetic Medicine and Departments of Biological Chemistry, Oncology and Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, USA

*Pichia pastoris* is a widely used eukaryotic host for production of recombinant proteins. We performed a proteogenomic analysis using high resolution Fourier transform MS to characterize the proteome of the GS115 strain. Our analysis resulted in identification of 46 889 unique peptides mapping to 3914 unique protein groups, which corresponds to ~80% of the predicted genes. In addition, our proteogenomic analysis led to the discovery of 64 novel genes and correction of 11 predicted gene models. The strategy used here demonstrates the utility of high resolution MS-derived peptide sequence data to cover near complete proteomes of organisms. Given the popularity of *P. pastoris* as a protein expression host, this proteome map should provide a list of contaminants derived from the host to assist in optimization of heterologous protein production. All MS data have been deposited in the ProteomeXchange with identifier PXD000627 (http://proteomecentral.proteomexchange.org/dataset/PXD000627).

Received: June 11, 2014
Revised: October 1, 2014
Accepted: October 20, 2014

**Keywords:**
Expression host / Heterologous protein production / Microbiology / Model system / Peroxisome biosynthesis

Additional supporting information may be found in the online version of this article at the publisher's web-site

## 1 Introduction

The methylotrophic yeast, *Pichia pastoris,* recently designated as *Komagataella pastoris,* has been widely used as a host for expression of recombinant proteins [1–3]. Expression of recombinant proteins in *P. pastoris* is under the control of a strong inducible promoter of alcohol oxidase gene, which is inducible by methanol [4]. The recombinant protein is secreted into the medium, which minimizes contamination from cytosolic proteins. Over the last several years, a number of strains of *Pichia* have been developed for use in specific applications. More than 500 recombinant proteins have been synthesized in *P. pastoris* in the recent past including various PTM proteins, monoclonal antibodies, and therapeutic proteins [5]. There is an extensive literature describing the advantages of *P. pastoris* over other expression systems [4]. For instance, the GS115 strain of *P. pastoris* has been used in the production of many recombinant proteins including hepatitis B surface antigen [6].

Surprisingly, although *P. pastoris* is frequently used as a recombinant protein expression host, the annotation of its genome has not been carried out in detail. In addition, only a limited number of proteomic studies have been carried out to verify the proteins encoded by its genome. Shevencheko

**Correspondence**: Dr. Akhilesh Pandey, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
**E-mail**: pandey@jhmi.edu
**Fax**: +1-410-502-7544

**Abbreviations: FDR**, false discovery rate; **NSAF**, normalized spectral abundance factor; **SCX**, strong-cation exchange chromatography; **YEPD**, yeast extract peptone dextrose

*Current Address: Praveen Kumar, Department of Biomedical Informatics and Computational Biology, University of Minnesota, Minneapolis, MN 55455, USA

**Colour Online:** See the article online to view Figs. 2–5 in colour.

 

et al. carried out a proteomic analysis of *Pichia* using a de novo sequencing approach in 2001 because its genome was still unsequenced at that time [7]. Mattonowich et al. identified 20 secreted proteins from the DSMZ 70382 strain of *P. pastoris* [8]. Dragosits et al. [9] used 2D gels to identify 49 proteins that are differentially regulated between low and high temperature conditions in two *P. pastoris* strains (X-33 and its variant 3H6 Fab). Effect of low oxygen conditions have also been studied in *P. pastoris* at transcriptomic, proteomic, and metabolic levels during heterologous protein production [10]. More recently, a SILAC compatible strain of *P. pastoris* has been developed, which allows the production of isotopically labeled heavy protein standards [11].

The GS115 strain of *P. pastoris* is a His auxotroph and has a methanol-utilization-plus phenotype ($Mut^+His^-$). This strain belongs to a group of *P. pastoris* strains that utilize methanol faster. As the expression of recombinant proteins is under the control of alcohol oxidase gene promoter, which is induced by methanol, rapid methanol utilization results in enhanced protein expression. The genome of GS115 strain of *P. pastoris* was sequenced in 2009 and reported to contain 5313 genes including 5040 protein-coding genes based on EuGene prediction algorithm [12]. Most of these protein-coding genes are annotated as hypothetical or putative proteins with unknown functions. Although, gene predictions are required initially to catalog the functional genome, it is essential to validate the annotated genes by detecting expressed transcripts/proteins. Expressed proteins can be detected and identified using various approaches. Recent advances in MS instrumentation have enabled in-depth proteome identification. Our group has systematically carried out genome annotation of several organisms of biomedical importance including *Anopheles gambiae* [13], *Mycobacterium tuberculosis* [14], *Candida glabrata* [15], *Leishmania donovani* [16], and *Homo sapiens* [17]. Here, we describe the proteome of GS115 strain of *P. pastoris* using high resolution Fourier transform MS. Using proteogenomics, we validated the expression of ~80% of predicted proteins and identified 64 novel protein-coding genes in *P. pastoris*. We also corrected 11 cases of erroneous protein-coding gene models predicted by computational means. To our knowledge, this is the first high-throughput proteomic study to describe the proteome of *P. pastoris*. We anticipate that the *P. pastoris* GS115 proteome described in this study will help the scientific community working in the field of recombinant protein expression using *P. pastoris* by providing a list of potential contaminants that can be copurified with the recombinant proteins.

## 2 Materials and methods

### 2.1 *Pichia pastoris* culture

GS115 strain of *P. pastoris* was obtained from American Type Cell Culture (ATCC number 20864). It was grown on yeast extract peptone dextrose (YEPD) agar plates and a single colony was further grown in YEPD medium supplemented with adenine sulphate for 16–18 h at 25°C. The cells were harvested at a final $O.D._{600}$ between 0.8–1.0. Cells were centrifuged for 10 min at 2000 × *g* for 10 min at 4°C, washed ten times with ice-cold PBS and the cell pellets stored at −80°C until further analysis.

### 2.2 In-gel trypsin digestion

Approximately $5 \times 10^8$ cells were resuspended in 200 μL alkaline lysis buffer (0.1 M NaOH, 0.05 M EDTA, 2% SDS, 2% β-mercaptoethanol) as described by Von der Haar [18], incubated at 90°C for 10 min. The sample was acidified with 5 μL of 4 M acetic acid and vortexed for 30 s and incubated further at 90°C for 10 min. A total of ~50 μL loading buffer (0.25 M Tris-HCl pH 6.8, 50% glycerol, 0.05% bromophenol blue) was added and lysate was cleared by centrifugation. A total of 200 μL supernatant was loaded on 10% 15 × 15 cm SDS-PAGE. The proteins were stained with colloidal Coomassie blue. Thirty-three gel bands were excised and subjected to in-gel trypsin digestion, as described previously [19]. Briefly, proteins were reduced using 5 mM DTT at 60°C for 45 min and alkylated by 20 mM iodoacetamide at room temperature for 10 min in dark. Trypsin (modified sequencing grade; Promega, Madison, WI, US) was added and digested at 37°C for 16 h. The reaction was quenched by adding 1 μL formic acid to each vial. Peptides were extracted, vacuum dried, reconstituted in 60 μL 0.1% formic acid, desalted using $C_{18}$ STAGE [20] tips and stored at −80°C until LC-MS/MS analysis.

### 2.3 In-solution trypsin digestion

Approximately $1 \times 10^9$ cells were resuspended in urea lysis buffer (9 M urea, 20 mM HEPES pH 8.0). Proteins were extracted by bead beating followed by ultrasonication. Protein estimation was carried out using bichinoic acid assay. Approximately, 2 mg protein was reduced for 20 min at 60°C in 5 mM DTT and then alkylated for 15 min by 10 mM iodoacetamide at room temperature in the dark. It was then diluted three times to make final urea concentration below 2 M. TPCK-treated trypsin (Worthington Biochemical Corporation, Lakewood, NJ, USA) was added at 1:20 w/w and digested overnight at 37°C. Trypsin activity was quenched by acidification with formic acid. Peptides were desalted using Sep-Pak $C_{18}$ cartridge (WAT051910, Waters Corporation, Milford, MA, USA) as per manufacturer's instructions. The eluted peptides were divided into two fractions, lyophilized and used for strong-cation exchange chromatography (SCX) and basic pH RPLC.

### 2.4 Strong-cation exchange chromatography

Peptides were separated on a PolySulfoethyl A column (200 × 4.6 mm, 5 μm, 200 Å, PolyLC Inc., Columbia, MD, USA)

 

using an Agilent 1200 series HPLC system containing a quaternary pump, manual injector, variant wavelength detector, and a fraction collector. The peptides were resuspended in 1 mL solvent A (10 mM $KH_2PO_4$, 20% ACN pH 2.8) and loaded on PolySulfoethyl A column using manual injector. Peptides were fractionated by a linear gradient of 0–100% solvent B (10 mM $KH_2PO_4$, 350 mM KCl, 20% acetonitrile, pH 2.8) over 50 min at a flow rate of 200 μL/min. A total of 96 fractions were collected for 50 min. The fractions were completely dried, reconstituted in 40 μL of 0.1% FA, pooled in to 32 fractions based on chromatography profile, desalted using $C_{18}$ STAGE [20] tips and stored at −80°C until LC-MS/MS analysis.

### 2.5 Basic pH RPLC

Basic pH RPLC was carried out at pH 9.5 on a XBridge $C_{18}$ column (250 × 4.6 mm, 5 μm, 200 Å, Waters Corporation, Milford, MA, USA). Peptides were reconstituted in bRPLC solvent A (10 mM triethylammonium bicarbonate, pH 9.5) and loaded on $C_{18}$ column using an Agilent 1200 series HPLC system. Peptides were fractionated by a 40 min gradient of 10–35% solvent B (10 mM trimethyl ammonium bicarbonate, 90% acetonitrile, pH 9.5) at a flow rate of 1.5 mL/min. The peptide fractions were collected in 96-well plate with preadded 5 μL of 20% TFA and vacuum dried. Peptides were reconstituted in 40% acetonitrile, concatenated into 36 fractions, vacuum dried, and stored at −20°C until LC-MS/MS analysis.

### 2.6 MS analysis

Peptides from in-gel digested protein bands were analyzed on a nanoflow HPLC system (1200 series, Agilent Technologies) while peptide fractions from SCX and bRPLC fractionation were analyzed on an Easy nano-LC (Thermo Scientific) connected to a hybrid LTQ–Orbitrap Velos ETD (Thermo Scientific) equipped with a nanoelectrospray ion source. In total, 101 fractions from all the methods were analyzed by MS. The peptide fractions were loaded on a 2 cm trap column (75 μm id) packed with $C_{18}$ material (Magic $C_{18}$AQ, 5 μm, 100 Å, Michrom Bioresources Inc.) using 0.1% formic acid with a flow rate 4 μL/min. The peptides were separated on a 15 cm analytical column (75 μm id) packed with $C_{18}$ material (Magic $C_{18}$AQ, 3 μm, 100 Å, Michrom Bioresources Inc.) with a 70 min gradient from 7 to 30% acetonitrile in 0.1% formic acid with a flow rate of 400 nL/min. The spray voltage was set to 2 kV while capillary temperature was set to 250°C. The MS instrument was operated in data-dependent acquisition mode. A survey full scan MS (from $m/z$ 350–1800) was acquired in the Orbitrap with resolution 60 000 at $m/z$ 400 with a maximum AGC target value of 1 000 000 ions in the linear ion trap. The twenty most intense peptide ions with charge states ≥2 were sequentially isolated to a target value of 50 000 ions and fragmented in the higher-energy collisional dissociation cell using 39% normalized collision energy. The maximum ion injection time for MS and MS/MS were set to 100 and 200 ms, respectively. Fragment ion spectra were detected in Orbitrap mass analyzer with a resolution $R = 15\,000$ at $m/z$ 400. For all measurements with the Orbitrap detector, a lock-mass ion from ambient air ($m/z$ 445.120025) was used for internal calibration as described [21].

### 2.7 Data analysis

The MS/MS searches were carried out using MASCOT, SEQUEST, and MS Amanda search algorithms through Proteome Discoverer (Version 1.4) software (Thermo Scientific) against GS115 strain of *P. pastoris* NCBI RefSeq protein database (Release 52; 5,040 sequences) and customized six-frame translated genome. The search workflow consisted of spectrum selector, MASCOT (Version 2.2.0), SEQUEST and MS Amanda search nodes, peptide validator and annotation (Supporting Information Fig. 1). Oxidation of methionine and deamidation of asparagine and glutamine were set as variable modifications and carbamidomethylation of cysteine was set as a fixed modification for all three search nodes. Protein N-terminal acetylation was set as a variable modification for MASCOT search node. MS and MS/MS mass tolerances were set to 20 ppm and 0.1 Da, respectively. Peptide identification was based on 1% false discovery rate (FDR) calculated using target-decoy database searches. MASCOT significant threshold based on peptide score, "Xcorr versus charge state" for SEQUEST node and Amanda score for MS Amanda search node, were used as filter settings for FDR calculation. The peptides passing 1% FDR criteria from all three search algorithms were then grouped into proteins based on maximum parsimony principles using Proteome Discoverer software suite.

### 2.8 Genome annotation pipeline

Six-frame translated database of *P. pastoris* genome was made using in-house scripts. The raw MS data were searched against the six-frame translated database using MASCOT, SEQUEST, and MS Amanda search algorithms. The peptides uniquely identified from six-frame translated genome database search (genome search specific peptides; GSSPs) were filtered and categorized based on their location with respect to known gene models such as intergenic, gene_exonic, and gene_intronic, as described previously [13] The quality of MS/MS spectra of these GSSPs were manually verified. The peptides from known proteins, GSSPs, gene prediction models predicted by Augustus, FGenesh and publicly available RNA-Seq data on GS115 strain of *P. pastoris* [22] were overlaid locally onto the gene models present in Ensembl Genome Browser. We also overlaid gene models from CBS7435 strain of *P. pastoris*. GSSPs were analyzed in the respective genomic

 

regions by overlaying the experimental data onto predicted gene models as well as publicly available RNA-Seq data. Novel and/or corrected gene model refinements were attributed based on supporting evidence from at least one prediction algorithm with additional gene model evidence from CBS7435 strain of *P. pastoris* and *P. pastoris* GS115 RNA-seq data. In cases, where two different gene models were predicted by these two programs, we used one where we also have evidence from CBS7435 strain of *P. pastoris* and/or RNA-Seq evidence from GS115 strain of *P. pastoris*.

### 2.9 Functional annotation

GO annotation was done using ProteinCenter (Version 1.0, Thermo Scientific) through Proteome Discoverer software suite to gene ontology domains—molecular function, biological process, and cellular component.

### 2.10 Domain structure prediction

Domain structure prediction of identified novel gene models and gene refinements were carried out using SMART prediction algorithm [23]. We used SMART in the ''Genomic Mode'' which includes protein sequences only from completely sequenced genomes. Briefly, SMART incorporates a library of Hidden Markov models, which utilizes a statistical model of amino acid preferences and insertion/deletion probabilities at each position in a sequence alignment.

### 2.11 Label-free quantitation

Protein abundance analysis was carried out using spectral counting based label-free quantitation as described by Paoletti et al. [24]. Briefly, the normalized spectral abundance factor (NSAF) was calculated for each identified protein. NSAF for a protein k was calculated by dividing the total number of peptide spectrum matches (S) identified for protein k by protein length (L) and then divided by the sum of S/L ratio for all proteins. The protein with highest NSAF value was termed as most abundant while the one with lowest was termed as least abundant.

## 3 Results and discussion

The goal of this study was to carry out deep proteomic profiling to generate a detailed proteome map of recombinant protein expression host *P. pastoris*. We undertook a multipronged approach including SDS-PAGE, SCX, and bRPLC based fractionation methods followed by MS analysis to obtain comprehensive proteome coverage of *P. pastoris*. The overall workflow used for proteogenomic analysis is shown in Fig. 1.

### 3.1 Validation of computationally predicted proteins and start sites

#### 3.1.1 In-depth proteomic analysis of *P. pastoris*

The current build of *P. pastoris* GS115 genome lists 5040 protein-coding gene annotations that were generated by the EuGene prediction program [12] with very little or no evidence in the form of transcripts or ESTs. MS-derived peptide data are direct evidence validating these predicted proteins. Using multiple prefractionation methods, we generated 101 fractions that were then analyzed on an LTQ-Orbitrap Velos mass spectrometer in a high- resolution mode, this resulted in acquisition of 647 093 MS/MS spectra. The database searches against GS115 strain of *P. pastoris* RefSeq protein database using MASCOT, SEQUEST, and MS Amanda resulted in identification of 187 938 peptide-spectrum matches. The median mass error was found to be 400 parts per billion. Supporting Information Fig. 2A shows the mass error distribution for all identified peptides. A total of 46 889 unique peptides were identified at 1% FDR, which mapped to 3914 protein groups accounting for ~80% of the *P. pastoris* proteome. To our knowledge, this study describes the most in-depth proteomic analysis for GS115 strain of *P. pastoris*. The identified proteins and peptides are listed in Supporting Information Tables 1 and 2, respectively.

A total of 931 proteins (23%) were identified with >50% coverage and the large majority (87%) of the identified proteins were identified based on two or more unique peptides (Supporting Information Fig. 2B). On an average, ten unique peptides were identified for each of the identified proteins and the average sequence coverage per protein was 30%. Figure 2A shows confirmation of a 154 amino acid protein, cytosolic superoxide dismutase (PAS_chr4_0786–1) in which seventeen unique peptides (indicated as red rectangles) were mapped to this protein covering 100% of the protein sequence.

#### 3.1.2 Confirmation of translational start sites

In eukaryotes, a large majority of proteins are known to be N-terminally acetylated during protein synthesis. This process is governed by aminopeptidases and N-acetyl transferases [25]. In proteogenomics, we employ this phenomenon to delineate and confirm translational start sites of proteins [26, 27]. Identification of N-terminally acetylated peptides using MS approach has been utilized to confirm and/or correct translational start sites [13, 15, 17]. In this study, translational start sites of 680 (17%) proteins out of 3914 total identified proteins were confirmed by N-terminally acetylated peptides (Supporting Information Table 3). Figure 2B shows an N-terminally acetylated peptide, aDGVFQGAIGIDLGTTYScVATYDSAVEIIANEQGNR, derived from cytoplasmic ATPase (PAS_chr3_0731–1). The distribution of amino acids

 

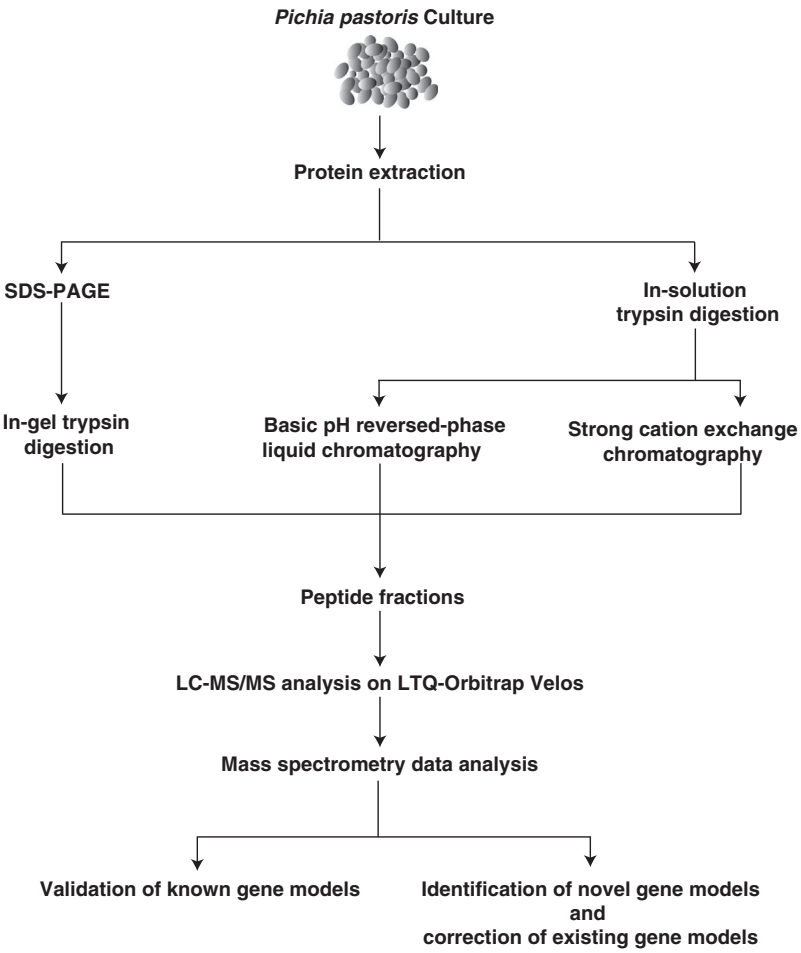


**Figure 1.** Schematic workflow depicting the steps in proteogenomic analysis. *Pichia pastoris* (GS115 strain) was cultured in yeast extract peptone dextrose medium. Protein lysate was subjected to in-gel and in-solution trypsin digestion. In-solution trypsin digested peptides were separated using strong-cation exchange chromatography and basic pH RPLC. Peptide fractions were analyzed using high resolution Fourier transform MS. The MS data were searched against protein database of *P. pastoris* GS115 to confirm the known/predicted proteome and against the six-frame translation of genome database to identify novel gene models and gene refinements using proteogenomics approaches.

present at P2 position was in accordance with that described in yeast [28]. In the majority (69%) of identified acetylated peptides, serine at position P2 was found to be acetylated while in 18% of cases alanine was acetylated (Fig. 3A and B). It has been shown that the presence of serine at P2 position is a common occurrence which is widely present in most types of yeast [29]. Peptides acetylated at methionine (P1 position) contain amino acids other than Ser, Ala, Pro, Thr, Gly, and Val at P2 position. This clearly shows the acetylation of methionine occurs only when none of seven amino acids mentioned earlier are present at P2 position and methionine plays the role of stabilization of N-terminus (Fig. 3C).

### 3.2 Identification of novel protein-coding genes

MS data were searched against six-frame translated genome of the GS115 strain. A list of peptides identified exclusively from genome database search that uniquely mapped to a single location in the genome (GSSPs) was generated. A total of 1345 GSSPs were identified that resulted in 326 clusters with two or more GSSPs based on grouping of GSSPs according to their chromosome and genome coordinates. Each GSSP was analyzed to evaluate the gene-coding potential of the corresponding genomic regions. Gene prediction from FGenesh and Augustus served as base gene models while CBS7435 strain of *P. pastoris* gene models were used as orthologous evidence. A total of 64 novel gene models were identified based on $\geq$2 GSSPs. Almost all of these (63 out of 64) were observed in *P. pastoris* CBS7435. Supporting Information Table 4 provides a list of novel gene models identified in this study. Figure 4A shows an example of a novel gene model (Supporting Information Table 4, IOB_PPAS_024) where 31 GSSPs (indicated by green blocks) were mapped to an unannotated genomic region on chromosome 4 of *P. pastoris* GS115. Gene prediction by Augustus and FGenesh depict the presence of a coding gene in this region with two exons; in-frame with all identified GSSPs. Annotation of IMP dehydrogenase from the CBS7435 strain of *P. pastoris* supports this identified novel gene. Gene model built using RNA-Seq derived transcript data also supports the novel gene model. A known snoRNA gene is also located in this region, residing in the intron of this novel gene, IOB_PPAS_024. Presence of noncoding RNA genes such as snoRNA gene in the intron of another gene is a known phenomenon in eukaryotes which controls and fine tunes gene expression [30]. snoRNA

 

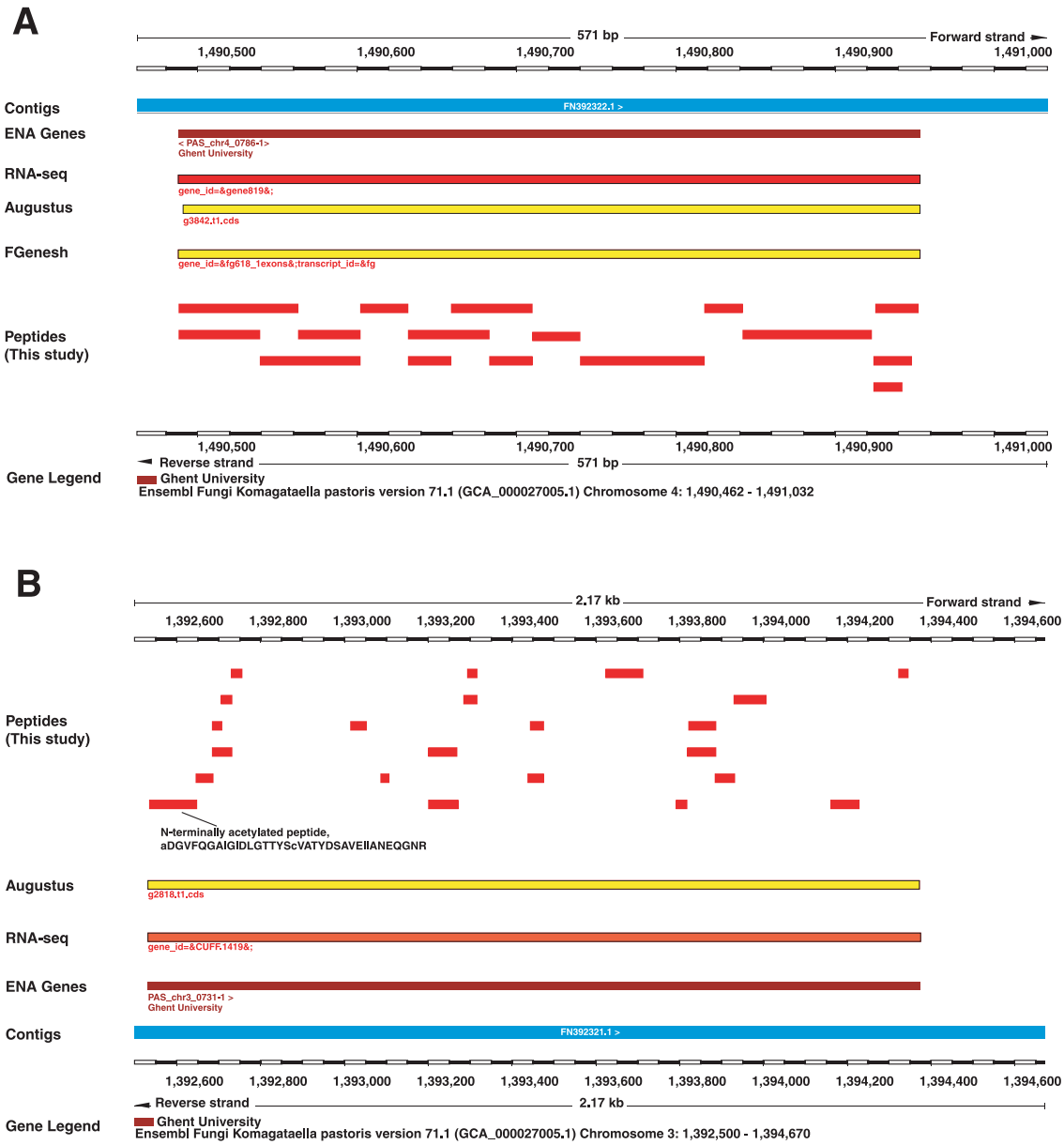


**Figure 2.** Confirmation of known gene models (A). Cytosolic superoxide dismutase (PAS_chr4_0786–1) was identified on the basis of 17 unique peptides (shown as red rectangles) covering 100% of protein sequence. Ensembl gene models as provided by Ghent University are labeled as ENA genes. Other tracks include gene models predicted by Augustus and FGenesh algorithms and gene model from RNA-Seq data (B). An N-terminally acetylated peptide, aDGVFQGAIGIDLGTTYScVATYDSAVEIIANEQGNR, was identified confirming the translation start site of cytoplasmic ATPase protein (PAS_chr3_0731–1, shown by ENA gene track). Other tracks show the gene models by Augustus and RNA-Seq derived transcript.

present in introns generally help in the biosynthesis and efficient splicing of the host gene [31].

### 3.3 Refinement of computationally predicted gene models

We identified a number of GSSPs that mapped to exon–intron, exon-intergene, and intergenic regions adjacent to existing genes in *P. pastoris*. Genes present in these regions could thus be refined based on such peptide identifications. We identified 11 instances of existing gene model refinements, which included two cases of gene extension, eight examples of novel exons, and one example of joining of genes.

#### 3.3.1 Identification of novel exons

Ornithine transporter of the mitochondrial inner membrane (PAS_chr3_100–1) was identified in this study

 

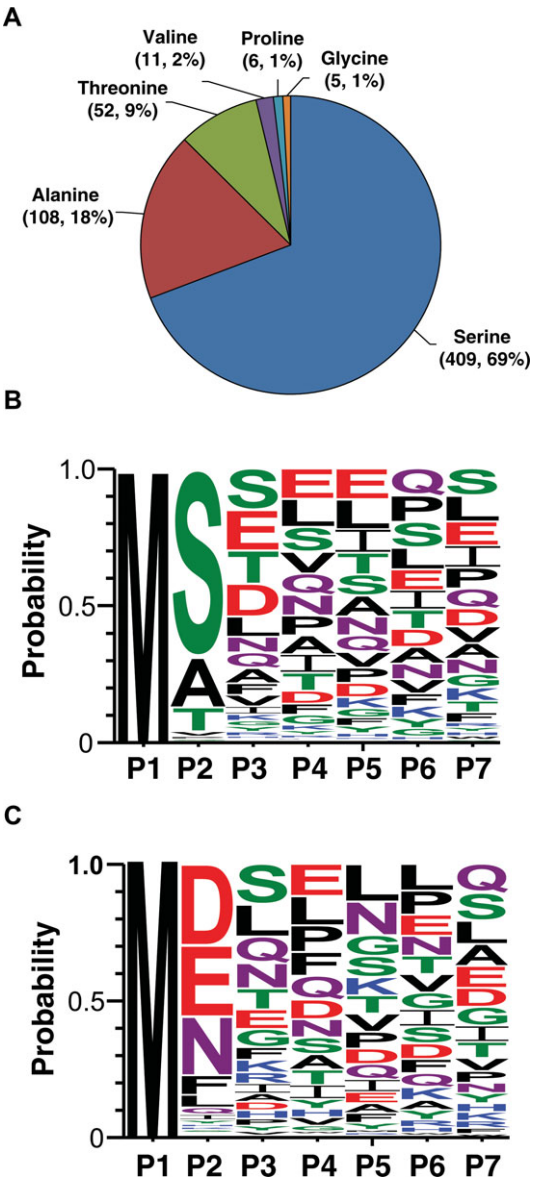


**Figure 3.** Protein N-terminal acetylation in *Pichia pastoris* GS115 (A). Distribution of N-terminally acetylated peptides based on presence of specific amino acid at P2 position (B). Logo showing probability of presence of an amino acid at P2 position in cases where the amino acid at P2 position was acetylated (C). Logo showing probability of an amino acid at P2 position in cases where the initiator methionine was acetylated.

based on four peptides. We identified four GSSPs YLAAYLLLQQGGVQTPSAEQISDLLQTVGSEVDESR, LSV LLKELEGK, ELEGKDISELIAEGSSK, and DISELIAEGSSK in the intergenic region downstream of PPAS_chr3_100–1 on chromosome 3 (shown as green rectangles Fig. 4B). Gene prediction by FGenesh identified a gene model in this region with two exons, one mapping to the known gene PPAS_chr3_100–1 and other overlapping with the four GSSPs. Thus, an additional novel exon was proposed downstream of the existing gene annotation. Figure 4B shows the known gene, PPAS_chr3_100–1, with peptides mapping to existing gene as well as the proposed novel exon. The protein sequence coded by this novel exon contains 60s acidic ribosomal protein domain. In total, eight novel exons were identified in eight different gene models. Supporting Information Table 4 lists the novel exons identified in this study.

### 3.3.2 Joining of two protein-coding genes

PAS_chr4_973–1 and PAS_chr4_974–1 are annotated as hypothetical proteins. These proteins were identified in this study based on a single peptide matching PAS_chr4_973–1 and 13 peptides matching PAS_chr4_974–1 (Fig. 5A). Two GSSPs, TAIAHEGKFPIIICQSFAK and FPIIICQSFAK (green rectangles) were mapped in the intronic region of PAS_chr4_974–1. Gene prediction algorithms FGenesh and Augustus predicted a protein-coding gene model exactly spanning these two genes. This is also supported by the presence of a gene in CBS7435 strain of *P. pastoris* and RNA-Seq evidence from publicly available RNA-Seq data of GS115 strain of *P. pastoris*.

### 3.3.3 Extension of a gene

Protein PAS_chr4_0455–1 was identified with four unique peptides (IALLK, IELLENTTK, AEPSNESNETEKLNPNAIK, and ELSPLRIELLENTTK) covering 41% of protein sequence (shown as red rectangles in Fig. 5B). We identified four GSSPs (SKNNIPVAVTVK, ENVSVVHLFKR, ENVSVVHLFK, and SQTISAVKK) that mapped to N-terminal intergenic region of PAS_chr4_0455–1 and one GSSP (QLCEMLNSTGGLSNDADANIEGNEIQIPK) that mapped to intergene-gene boundary (Shown as green rectangles in Fig. 5B). Predicted gene model by FGenesh and Augustus in this region is in-frame with existing gene and GSSPs indicating the extension of N-terminus of the gene PAS_chr4_0455–1. This gene encodes for a hypothetical protein with no known function. A protein from *P. pastoris* CBS7435, PP7435_chr4–0525 shows similar length as per the predicted N-terminus as shown in Fig. 5B. We identified one example each of N-terminal and C-terminal protein extension (Supporting Information Table 4).

## 3.4 Domain structure analysis of novel gene models

The functional activity of a protein is governed primarily by its structure and the presence of particular domains and motifs. Thus, it is important to know the domains present in newly identified and/or predicted proteins to ascertain their

 

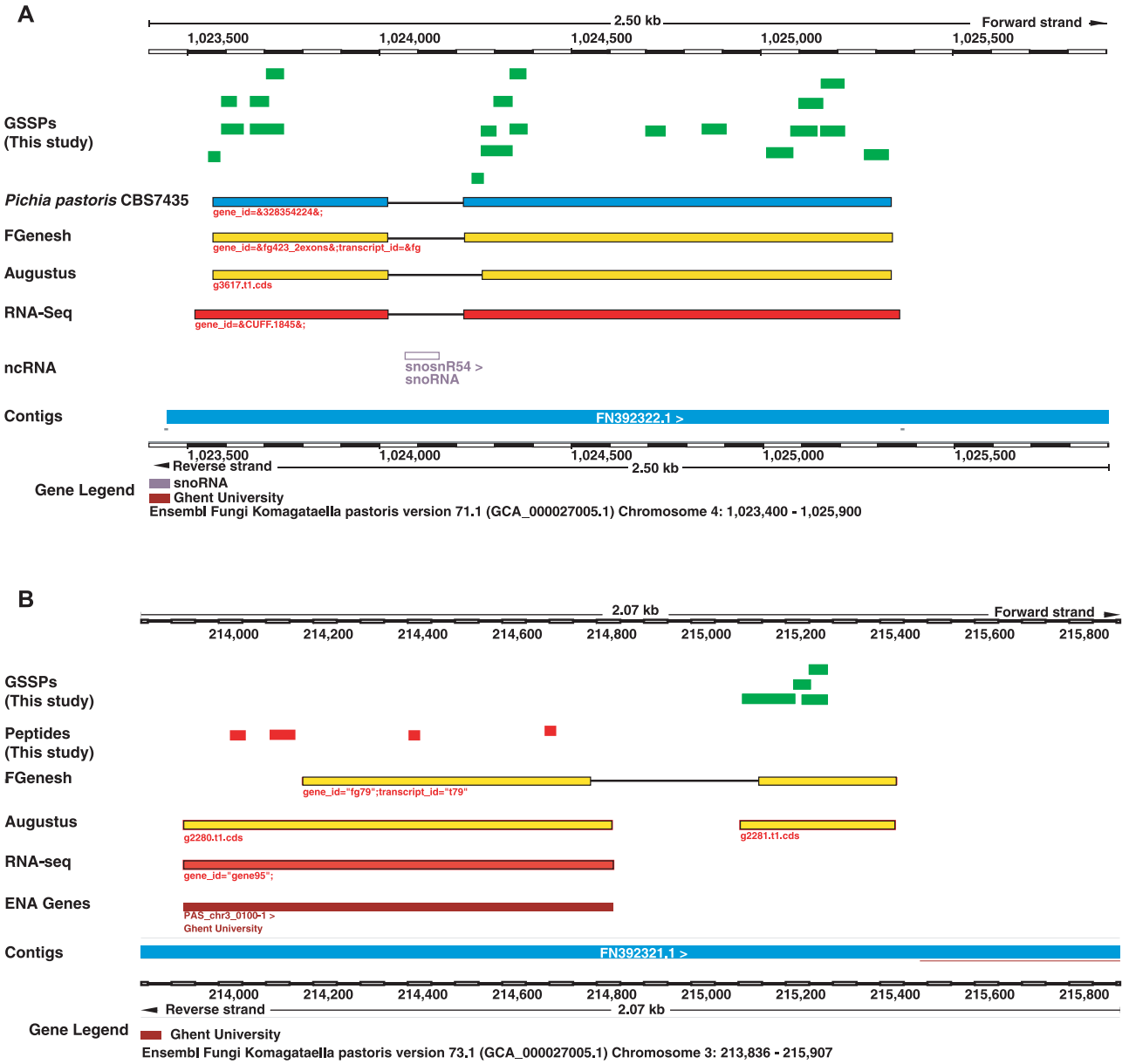


**Figure 4.** Identification of novel genes and exons (A). An example of a novel gene identification using 19 -genome-search specific peptides (GSSPs) (green rectangles) mapped to an unannotated region. Ensembl gene models as provided by Ghent University are labeled as ENA genes. Gene prediction algorithms Augustus and FGenesh predicted a coding gene in this region with two exons; in-frame with all identified GSSPs. A gene from a different strain of *Pichia pastoris* (CBS7435) contains a gene model similar to that predicted by Augustus and FGenesh. Gene model built using RNA-Seq data also supports the prediction model (B). An example of novel exon identification using four GSSPs (green rectangles) that were mapped in the intergenic region. Gene prediction algorithms FGenesh predicted a gene model in this region with two exons, one mapping to known gene, Ornithine transporter of the mitochondrial inner membrane (PAS_chr3_100–1) and other mapping to identified GSSPs.

potential function. Computational methods have been used for prediction of domains of the proteins. Presence of a domain in newly identified and/or predicted protein can provide additional confirmation for hypothetical/putative proteins. Identified novel gene models as well as gene model refinements were subjected to domain structure analysis using SMART algorithm [23]. We found that 40 out of 64 (80%) novel gene models contain well-characterized domains. Although, the remaining novel gene models do not have any known domain structures, protein evidence by MS supports the presence of these proteins and warrants further investigation to characterize their function. Supporting

 

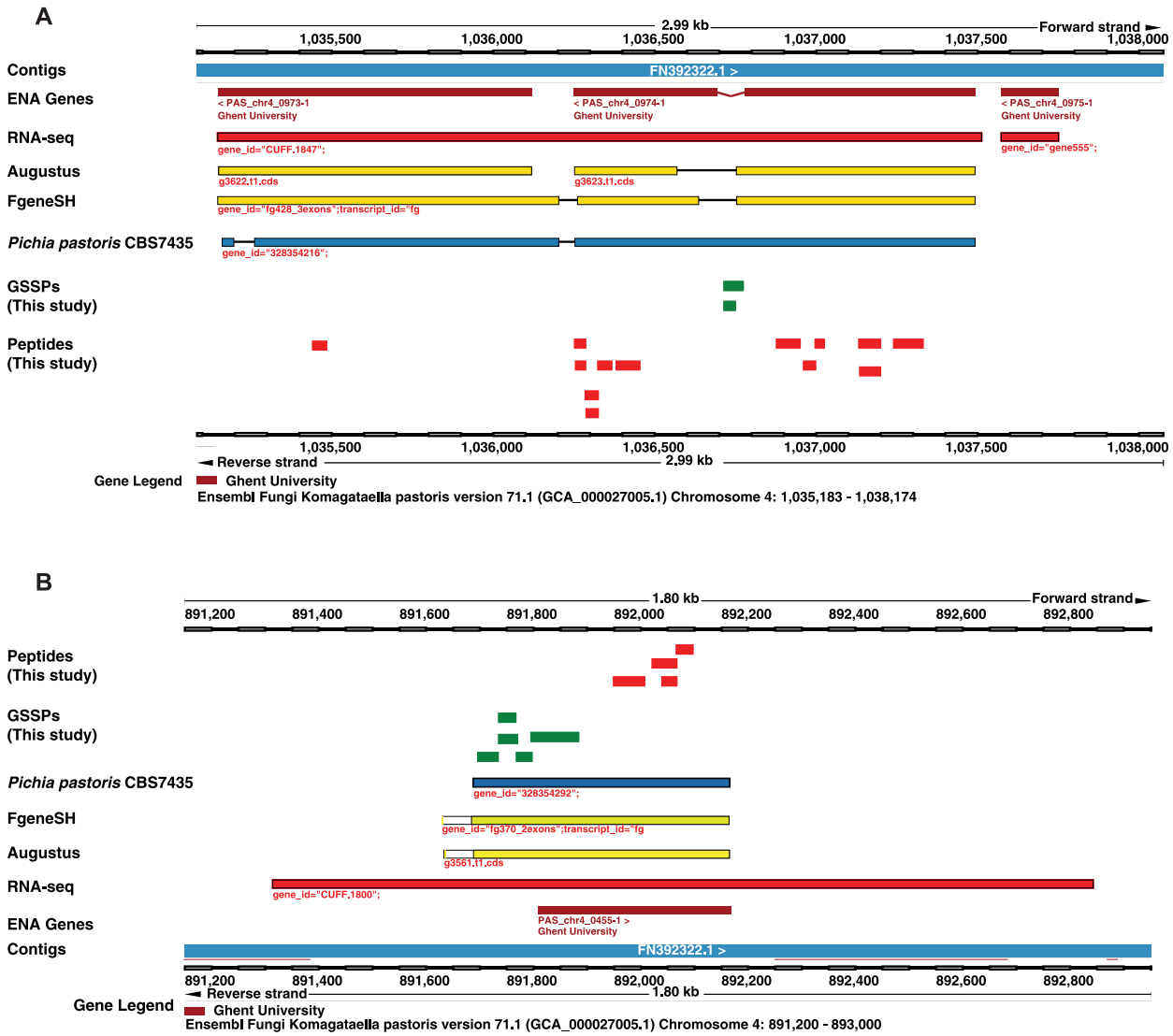


**Figure 5.** Gene model refinement using genome search-specific peptides (A). An example of joining of genes PAS_chr_973–1 and PAS_chr_974–1 annotated as hypothetical proteins. A total of 14 peptides (red rectangles) were mapped to these two proteins. Two GSSPs, TAIAHEGKFPIIICQSFAK and FPIIICQSFAK (green rectangles) were mapped to the intronic region of the gene PAS_chr4_974–1. Gene prediction algorithms, FGenesh and Augustus, predicted a single gene model in this region which is also corroborated by presence of an orthologous gene in *Pichia pastoris* CBS7435 and the gene model based on RNA-Seq derived transcript (C). An example showing N-terminal extension of a protein PAS_chr4_0455–1. Protein PAS_Chr4_0455–1 was identified with four unique peptides (IALLK, IELLENTTK, AEPSNESNETEKLNPNAIK, and ELSPLRIELLENTTK, shown as red rectangles). Four GSSPs (SKNNIPVAVTVK, ENVSVVHLFKR, ENVSVVHLFK, and SQTISAVKK, shown as green rectangles) were mapped to N-terminal intergenic region of PAS_chr4_0455–1 and one GSSP (QLCEMLNSTGGLSNDADANIEGNEIQIPK) mapping to intergene-gene boundary. FGenesh and Augustus predicted a gene in this region in-frame with existing gene and GSSPs indicating the extension of N-terminus of the gene PAS_chr4_0455-1.

Information Table 4 lists all the identified domains for novel gene models as well as gene model refinements.

### 3.5 Functional annotation of identified proteins

The proteins identified in this study were classified based on GO (Supporting Information Fig. 3). The annotations based on molecular function (Supporting Information Fig. 3A) revealed that majority of the proteins were involved in catalytic activity (28%). Around 30% of identified proteins were involved in binding such as nucleotide binding (13%), protein binding (10%), and metal ion binding (7%). About 26% of identified proteins had no annotations available. Based on cellular component (Supporting Information Fig. 3B), the identified proteins were localized to membrane (38%),

 

cytoplasm (12%), and nucleus (14%). Proteins localized to ribosome (3%) and mitochondria (2%) were also identified. According to biological process annotated in GO (Supporting Information Fig. 3C), the majority of the identified proteins were involved in metabolic processes (39%) whereas others were involved in transport (8%), regulation of biological process (6%), and cell organization and biogenesis (3%). Biological processes for around 42% of identified proteins were not annotated. Overall, 25–40% of identified proteins have not been annotated for any of the GO categories.

### 3.6 Protein abundance analysis using spectral counting

Label-free approaches have been used for quantitation of abundance in a number of studies. Spectral counting is considered as a standard estimation for protein abundance in MS-based proteomic experiments where no isotope labeling techniques are used [32]. Protein abundance of all identified proteins was calculated using spectral counting label-free quantitation approach as per the NSAF [24]. The NSAF values for each identified protein ranges from 0 to 1. Values close to 1 denote the most abundant protein while that close to 0 denote less abundant. Glyceraldehyde-3-phosphate dehydrogenase, one of the enzymes in the glycolytic pathway, was found to be most abundant protein in the current dataset with several other metabolic enzymes. Interestingly, enzymes involved in the glycolysis and TCA cycle pathway were found to be in top 500 most abundant proteins (Supporting Information Table 5).

### 3.7 Data availability

The MS proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository [33] with the dataset identifier PXD000627.

## 4 Concluding remarks

High resolution MS-based approaches are becoming popular to characterize the proteome of various organisms. Here, we describe the proteome of a methylotrophic recombinant protein expression host, *P. pastoris*. Given that *P. pastoris* is perhaps the most widely used protein expression host, a reference proteome map should be helpful in protein expression studies. The N-terminal acetylation data provided evidence for the presence of specific conserved amino acids at P2 position. In addition to this, using proteogenomics approaches, we proposed novel protein-coding genes and gene refinement of existing protein-coding genes. These proposed novel protein-coding gene models contain a number of well-characterized domains. Overall, a reference proteome map of *P. pastoris* was described which will facilitate future heterologous protein expression studies using *Pichia* as a host system.

*We thank Department of Biotechnology, Government of India for research support to Institute of Bioinformatics. Santosh Renuse is a recipient of a Senior Research Fellowship from the University Grants Commission (UGC), Government of India. Anil K. Madugundu is recipient of a Senior Research Fellowship from Bioinformatics National Certification Examination (BINC), Department of Biotechnology (DBT), Government of India. Harsha Gowda is a Wellcome Trust/DBT India Alliance Early Career Fellow. T. S. Keshava Prasad is supported by a research grant on "Development of Infrastructure and a Computational Framework for Analysis of Proteomic Data" from Department of Biotechnology, Government of India.*

*The authors have declared no conflict of interest.*

## 5 References

[1] Cregg, J. M., Vedvick, T. S., Raschke, W. C., Recent advances in the expression of foreign genes in Pichia pastoris. *Biotechnology (N Y)* 1993, *11*, 905–910.

[2] Cereghino, J. L., Cregg, J. M., Heterologous protein expression in the methylotrophic yeast Pichia pastoris. *FEMS Microbiol Rev* 2000, *24*, 45–66.

[3] Weidner, M., Taupp, M., Hallam, S. J., Expression of recombinant proteins in the methylotrophic yeast Pichia pastoris. *J. Vis. Exp* 2010, *36*, 1–5.

[4] Cregg, J. M., Tolstorukov, I., Kusari, A., Sunga, J. et al., Expression in the yeast Pichia pastoris. *Meth. Enzymol.* 2009, *463*, 169–189.

[5] Macauley-Patrick, S., Fazenda, M. L., McNeil, B., Harvey, L. M., Heterologous protein production using the Pichia pastoris expression system. *Yeast* 2005, *22*, 249–270.

[6] Hardy, E., Martinez, E., Diago, D., Diaz, R. et al., Large-scale production of recombinant hepatitis B surface antigen from Pichia pastoris. *J. Biotechnol.* 2000, *77*, 157–167.

[7] Shevchenko, A., Sunyaev, S., Loboda, A., Shevchenko, A. et al., Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.* 2001, *73*, 1917–1926.

[8] Mattanovich, D., Graf, A., Stadlmann, J., Dragosits, M. et al., Genome, secretome and glucose transport highlight unique features of the protein production host Pichia pastoris. *Microb. Cell Fact.* 2009, *8*, 29.

[9] Dragosits, M., Stadlmann, J., Albiol, J., Baumann, K. et al., The effect of temperature on the proteome of recombinant Pichia pastoris. *J. Proteome Res.* 2009, *8*, 1380–1392.

[10] Baumann, K., Carnicer, M., Dragosits, M., Graf, A. B. et al., A multi-level study of recombinant Pichia pastoris in different oxygen conditions. *BMC Syst. Biol.* 2010, *4*, 141.

[11] Austin, R. J., Kuestner, R. E., Chang, D. K., Madden, K. R., Martin, D. B., SILAC compatible strain of Pichia pastoris for

 

expression of isotopically labeled protein standards and quantitative proteomics. *J. Proteome Res.* 2011, *10*, 5251–5259.

[12] De Schutter, K., Lin, Y. C., Tiels, P., Van Hecke, A. et al., Genome sequence of the recombinant protein production host Pichia pastoris. *Nat. Biotechnol.* 2009, *27*, 561–566.

[13] Chaerkady, R., Kelkar, D. S., Muthusamy, B., Kandasamy, K. et al., A proteogenomic analysis of Anopheles gambiae using high-resolution Fourier transform mass spectrometry. *Genome Res.* 2011, *21*, 1872–1881.

[14] Kelkar, D. S., Kumar, D., Kumar, P., Balakrishnan, L. et al., Proteogenomic analysis of Mycobacterium tuberculosis by high resolution mass spectrometry. *Mol. Cell Proteomics* 2011, *10*, M111 011627.

[15] Prasad, T. S., Harsha, H. C., Keerthikumar, S., Sekhar, N. R. et al., Proteogenomic analysis of Candida glabrata using high resolution mass spectrometry. *J. Proteome Res.* 2012, *11*, 247–260.

[16] Pawar, H., Sahasrabuddhe, N. A., Renuse, S., Keerthikumar, S. et al., A proteogenomic approach to map the proteome of an unsequenced pathogen—Leishmania donovani. *Proteomics* 2012, *12*, 832–844.

[17] Kim, M. S., Pinto, S. M., Getnet, D., Nirujogi, R. S. et al., A draft map of the human proteome. *Nature* 2014, *509*, 575–581.

[18] von der Haar, T., Optimized protein extraction for quantitative proteomics of yeasts. *PLoS One* 2007, *2*, e1078.

[19] Nagarajha Selvan, L. D., Kaviyil, J. E., Nirujogi, R. S., Muthusamy, B. et al., Proteogenomic analysis of pathogenic yeast Cryptococcus neoformans using high resolution mass spectrometry. *Clin. Proteomics* 2014, *11*, 5.

[20] Rappsilber, J., Ishihama, Y., Mann, M., Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* 2003, *75*, 663–670.

[21] Olsen, J. V., de Godoy, L. M., Li, G., Macek, B. et al., Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* 2005, *4*, 2010–2021.

[22] Liang, S., Wang, B., Pan, L., Ye, Y. et al., Comprehensive structural annotation of Pichia pastoris transcriptome and the response to various carbon sources using deep paired-end RNA sequencing. *BMC Genomics* 2012, *13*, 738.

[23] Letunic, I., Doerks, T., Bork, P., SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* 2012, *40*, D302–D305.

[24] Paoletti, A. C., Parmely, T. J., Tomomori-Sato, C., Sato, S. et al., Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. *Proc. Natl. Acad. Sci. USA* 2006, *103*, 18928–18933.

[25] Molina, H., Bunkenborg, J., Reddy, G. H., Muthusamy, B. et al., A proteomic analysis of human hemodialysis fluid. *Mol. Cell. Proteomics* 2005, *4*, 637–650.

[26] Giglione, C., Boularot, A., Meinnel, T., Protein N-terminal methionine excision. *Cell. Mol. Life Sci.* 2004, *61*, 1455–1474.

[27] Starheim, K. K., Gevaert, K., Arnesen, T., Protein N-terminal acetyltransferases: when the start matters. *Trends Biochem. Sci.* 2012, *37*, 152–161.

[28] Huang, S., Elliott, R. C., Liu, P. S., Koduri, R. K. et al., Specificity of cotranslational amino-terminal processing of proteins in yeast. *Biochemistry* 1987, *26*, 8242–8246.

[29] Bonissone, S., Gupta, N., Romine, M., Bradshaw, R. A., Pevzner, P. A., N-terminal protein processing: a comparative proteogenomic analysis. *Mol. Cell Proteomics* 2013, *12*, 14–28.

[30] Rearick, D., Prakash, A., McSweeny, A., Shepard, S. S. et al., Critical association of ncRNA with introns. *Nucleic Acids Res.* 2011, *39*, 2357–2366.

[31] Vincenti, S., De Chiara, V., Bozzoni, I., Presutti, C., The position of yeast snoRNA-coding regions within host introns is essential for their biosynthesis and for efficient splicing of the host pre-mRNA. *RNA* 2007, *13*, 138–150.

[32] Liu, H., Sadygov, R. G., Yates, J. R., 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 2004, *76*, 4193–4201.

[33] Vizcaino, J. A., Cote, R. G., Csordas, A., Dianes, J. A. et al., The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* 2013, *41*, D1063–D1069.

 