Journal of Genetics (2020)99:49
https://doi.org/10.1007/s12041-020-01201-1

RESEARCH ARTICLE

Genomic data mining reveals the transaminase repertoire of *Komagataella phaffii* (*Pichia pastoris*) strain GS115 and supports a systematic nomenclature

MARIA-JOSÉ HENRÍQUEZ, ROSALÍA PAULA CARDÓS-ELENA and DARREN NICHOLAS NESBETH*

Department of Biochemical Engineering, University College London, Bernard Katz Building, London WC1E 6BT, UK
*For correspondence. E-mail: d.nesbeth@ucl.ac.uk.

Received 22 July 2019; revised 20 January 2020; accepted 3 March 2020

Abstract. Transaminases are an industrially important class of enzyme, due to their ability to catalyse amination reactions for production of chiral amines, and are key building blocks of small molecule pharmaceuticals. We analysed the genome of strain GS115 of the methylotrophic yeast *Komagataella phaffii*, formerly known as *Pichia pastoris*, to identify the transaminase genes and propose a systematic nomenclature based on both phylogeny and structuro-functional features. *K. phaffii* is an increasingly attractive industrial host cell due to its ability to grow to high biomass, up to 60% wet cell weight by volume, using methanol as carbon source and inducer of transgene expression. Thirty-nine UniProt database hits were reduced to 19 on the basis of sequence similarity and hidden Markov model. Of the 19 genes, the open-reading frames of three (KpTam I-II.1b, KpTam I-II.7 and KpTam V.2) had strong homology with no characterized protein and four (KpTam III.1a, KpTam III.1b, KpTam III.2a and KpTam III.2b) had relatively high sequence similarity to ω-type transaminases, a subtype that typically accepts the broadest range of substrates. Comparison with *Saccharomyces cerevisiae* S288C suggested functions for KpTam I-II.1b and KpTam I-II.7. *K. phaffii* GS115 was originally generated by mutagenesis of *K. phaffii* CBS7435 and comparison revealed that one transaminase gene may have been deleted during this mutagenesis. These insights can advance fundamental understanding of yeast biology and can inform industrial screening and engineering of yeast transaminases.

Keywords. transaminase; aminotransferase; bioinformatics; genome; phylogeny; nomenclature.

Introduction

Transaminases, also referred to as aminotransferases, are homodimeric enzymes that use pyridoxal-5′-phosphate (PLP) as a cofactor in catalysing the transfer of an amino group from a donor molecule, typically an amino acid, to the oxo- or carbonyl group of a keto-acid. In cellular metabolism, transaminases play an important role in nitrogen utilization and in synthesis of amino acids, vitamins, bacterial cell walls and antibiotics. As such, transaminases are ubiquitous in nature and the genomes of most microorganisms encode several of them (Ward and Wohlgemuth 2010). The Enzyme Commission (EC) (Bairoch 2000) has defined 112 transaminase types, assigning EC numbers 2.6.1.X to each type based on the reaction catalysed.

Mehta *et al*. (1993) analysed the homology and hydropathy patterns of 51 transaminase amino acid sequences to map their structural relatedness. From this analysis they suggested four overall classes of transaminase existed, based on substrates preference, and concluded that all the transaminases had originated from the same ancestor. The growth in sequence data in the post-genomic era greatly increased the number of putative transaminase sequences available and Hwang *et al*. (2005) used protein sequences from the Pfam database (Finn *et al*. 2016) to suggest classification of transaminases into six classes (I–VI), within five subgroups (table 1).

Transaminases can also be classified according to the position of an acceptor keto- or aldehyde group, relative to the carboxylic acid or other major group of the amino donor molecule. Class I, II, IV and V transaminases are α-transaminases, as they transfer the amino group from the α position. Class III contains β-transaminases, γ-transaminases, and ω-transaminases. Class III transaminases are especially sought after as they have the widest substrate spectrum. Class III, ω-type transaminases are the only

Published online: 03 June 2020

Table 1. The subgroups of transaminases set out by Hwang *et al.* (2005).

Subgroup	Notes
Class I+II	Classes I and II were first proposed by Mehta *et al.* (1993) and comprise, respectively, the aspartate and aromatic transaminases. These are the most studied, and in general they use L-aspartate and L-tyrosine as amino donor, respectively, and α-ketoglutarate as an amino acceptor. The difference between an aspartate and aromatic transaminase is the hydrophobicity of the active site binding pocket (Hwang *et al.* 2005).
Class III	Have a wide variety of substrate acceptors and can transfer an amino group to aldehydes and ketones of different types. Include β-transaminases, ϒ-aminobutyrate transaminases, and ω-transaminases. Have no requirement for 2-ketocarboxylate (Ward and Wohlgemuth 2010).
Class IV	Differ structurally from the other types in that the positions of the large and small binding pockets are reversed.
Class V	Act on structurally and biosynthetically related substrates (Mehta *et al.* 1993).
Class VI	Comprise sugar aminotransferases, the majority of which use L-glutamate as the amino donor. Are derived from antibiotic operons (Ward and Wohlgemuth 2010).

enzyme known to perform stereoselective amination of ketones. As a result, class III ω-type transaminases have been investigated extensively for industrial production of amino acids, chiral amines and amino alcohols, all of which are valuable key intermediates for chemical synthesis of chiral, small-molecule therapeutics (Malik *et al.* 2012).

Class III ω-type transaminases are highly active for a given substrate and also stable to a broad range of pH, substrate concentration, temperature and product concentration. Class III ω-type transaminases with industrial promise have been identified by Shin *et al.* (2003) in *Vibrio fluvialis* JS17 by Yonaha *et al.* (1992) in *Pseudomonas putida* and by Kaulmann *et al.* (2007) in *Chromobacterium violaceum.*

These transaminases have also been coexpressed with other exogenous enzymes in a common host cells, such as *Escherichia coli*, to provide *de novo* multistep, whole cell biosynthetic pathways (Kaulmann *et al.* 2007). Cho *et al.* (2003) coupled two transaminase reactions (α and ω) for whole cell kinetic resolution of chiral amines. Ingram *et al.* (2007) achieved whole cell asymmetric synthesis of a chiral amino alcohol using co-overexpression of transketolase and transaminase in *E. coli*.

In addition to experimental screening, bioinformatic data mining has emerged as a successful route to discover the novel transaminases and prediction of their activity and substrate specificities (Hohne *et al.* 2010; Valli *et al.* 2016). Statistical techniques, such as hidden Markov modelling (HMM) can be used to predict whether a given transaminase belongs to a particular classification using a temporal pattern recognition which enables the creation of protein structure profiles that may discriminate whether an input sequence belongs to a protein family or subtype, despite apparent nonsignificant sequence homology (Krogh *et al.* 1994). Using HMMER software (Finn *et al.* 2015), it is possible to search profile databases for sequence homologs employing HMM.

The methylotrophic yeast *P. pastoris* was reassigned to the genus *Komagataella* following phylogenetic analysis (Kurtzman 2005), and the major strains split into three species: *K. pastoris*, *K. phaffii* GS115 (formerly *P. pastoris* strain GS115) and *K. pseudopastoris*. The *K. phaffii* GS115 is a reliable and robust expression system (Invitrogen 2010) that has become widely used in research and industrial settings for the production of recombinant protein (Templar *et al.* 2016) with well-established tools (Bollok *et al.* 2009) for genetic manipulation, strong native promoters to direct overexpression of transgenes and the ability to grow high cell densities rapidly by volume up to 60% wet cell weight (wcw) on a chemically defined culture media (Wei *et al.* 2018). Further, *K. phaffii* GS115 is thermos tolerant, able to grow at 47°C (Van der Klei *et al.* 2006), and tolerant to pH 3–7 (Macauley-Patrick *et al.* 2005).

The availability of the complete genome sequence (De Schutter *et al.* 2009; Mattanovich *et al.* 2009) has made bioinformatic data mining possible for transaminases in *K. phaffii* GS115. Here we attempt to predict the function of transaminase using sequence analyses and assign each identified transaminases to a Hwang subclasses (Hwang *et al.* 2005) using HMM analysis. We also propose a rational nomenclature for *K. phaffii* GS115 transaminases. This information will assist future investigators, who wish to exploit or rationally design transaminases for enhanced stability, substrate specificity, PLP binding and other properties.

Material and methods

Identifying transaminase genes in yeast genome sequences

The UniProt database (The Uniprot Consortium 2015) was first searched using the terms ‘pichia pastoris’, ‘gs115’ and ‘GO:0008483’ (gene ontology term for transaminase/aminotransferase activity); and then, replacing the GO terms by the keywords ‘transaminase’ and ‘aminotransferase’ to generate a wider pool of initial sequences. JalView v.2.8.2 (Waterhouse *et al.* 2009) was then used to remove duplicate results by examining identity percentages and pairwise alignments. The first round of UniProt hits was further refined using the terms ‘pyridoxal binding site’, ‘aminotransferase’ and ‘transaminase’. The nucleotide sequence of each transaminase identified in UniProt was obtained from the corresponding NCBI entry. The methods above were also used to identify transaminase genes in the published genome sequences of *K. phaffii* CBS7435 (Küberl *et al.* 2011) and *S. cerevisiae* S288C (Goffeau *et al.* 1996).

Table 2. Transaminases identified in *K. phaffii* (*P. pastoris*) GS115 genome, by protein accession number, gene location name, Hwang subclass, predicted by HMMER HMMscan, and proposed systematic transaminase nomenclature.

Accession number	Gene name	Predicted function	Class	Systematic nomenclature
C4QWE4	PAS_chr1-1_0200	Aspartate aminotransferase (EC 2.6.1.1)	I-II	KpTam I-II.1a
C4R862	PAS_chr4_0974	Uncharacterized protein	I-II	KpTam I-II.1b
C4R1J1	PAS_chr2-1_0716	5-Aminolevulinate synthase, delta-aminolevulinate synthase (EC 2.3.1.37)	I-II	KpTam I-II.2a
C4QZ99	PAS_FragB_0040	Serine palmitoyltransferase 1	I-II	KpTam I-II.2b
C4R5E6	PAS_chr3_0733	Component of serine palmitoyltransferase	I-II	KpTam I-II.3
C4R6Z7	PAS_chr4_0147	Aromatic aminotransferase II	I-II	KpTam I-II.4a
C4QYZ3	PAS_chr1-4_0608	Aromatic aminotransferase I, expression is regulated by general control of amino acid biosynthesis	I-II	KpTam I-II.4b
C4R366	PAS_chr3_1132	Kynurenine aminotransferase, catalyzes formation of kynurenic acid from kynurenine	I-II	KpTam I-II.5
C4R4P4	PAS_chr3_0482	Putative alanine transaminase (glutamic pyruvic transaminase)	I-II	KpTam I-II.6a
C4R1F7	PAS_chr2-1_0684	Histidinol-phosphate aminotransferase, catalyzes the seventh step in histidine biosynthesis	I-II	KpTam I-II.6b
C4QYW2	PAS_chr1-4_0579	Uncharacterized protein	I-II	KpTam I-II.7
C4QZN6	PAS_chr2-1_0107	Gamma-aminobutyrate (GABA) transaminase (4-aminobutyrate aminotransferase)	III	KpTam III.1a
C4R8L1	PAS_chr4_0677	GABA transaminase (4-aminobutyrate aminotransferase)	III	KpTam III.1b
C4R4H3	PAS_chr3_0410	L-ornithine transaminase (OTAse)	III	KpTam III.2a
C4R8H9	PAS_chr4_0645	Acetylornithine aminotransferase	III	KpTam III.2b
C4R7A4	PAS_chr4_0248	Branched-chain-amino-acid aminotransferase (EC 2.6.1.42)	IV	KpTam IV.1
C4R7U0	PAS_chr4_0416	Alanine:glyoxylate aminotransferase (AGT), catalyzes the synthesis of glycine from glyoxylate	V	KpTam V.1a
C4R4X8	PAS_chr3_0566	3-Phosphoserine aminotransferase	V	KpTam V.1b
C4R6Z6	PAS_chr4_0146	Uncharacterized protein	V	KpTam V.2

Assembly of transaminase dendrograms and phylograms

Protein sequences were aligned using ClustalX v.2.1 (Larkin *et al*. 2007) and a phylogenetic tree (Qian and Goldstein 2003) assembled using the neighbour-joining clustering algorithm and 1000 bootstrap replicates. The alignments were used to generate a tree diagram using TreeView v.1.6.6 (Page 1996) and dendrograms plotted using Dendroscope v.3.3.2 (Huson and Scornavacca 2012). Dendrogram images were edited for graphical brevity, to indicate bootstrap values and Hwang subgroup.

Assigning transaminases to a Hwang subgroup

Hits from the above UniProt search were then used to query the HMM database HMMER (Finn *et al*. 2015), provided by the European Molecular Biology Organisation - Biology European Bioinformatics Institute (EMBO-EBI) website (http://www.ebi.ac.uk/Tools/hmmer/search/hmmscan), by accession number. Predictions were deemed acceptable if the score was positive and expectation values (*E*-values) lower than 0.001. All protein families available on the search tool (Pfam, TIGRFAM, Gene3D, Superfamily and PIRSF) were selected for the search and those transaminases scoring highest for a HMM model pertaining to a particular subgroup (Hwang *et al*. 2005) were assigned to that subgroup.

Alignment of ω-transaminases

Vibrio fluvialis JS17 and *C. violaceum* DSM30191 ω-transaminase sequences were aligned with the four putative *K. phaffii* GS115 class III transaminases using ClustalW. The alignment file was inserted in JalView v.2.8.2 and residues colour-coded according to their BLOSUM62 score (Henikoff and Henikoff 1992). Secondary structure elements were generated in ESPript 3.0 (Robert and Gouet 2014). Pairwise alignments were performed in JalView v.2.8.2 to determine protein-to-protein identities. Analyses of *K. phaffii* CBS7435, *S. cerevisiae* S288C and *K. phaffii* GS115 class III transaminases with respect to similarity were performed using protein BLAST.

Results and discussion

Analysis and classification of K. phaffii GS115 transaminases

A total of 39 potential transaminases/aminotransferase genes were identified in an initial search of the *K. phaffii* GS115 genome using the UniProt webtool (table 2). A second UniProt search was performed within the 39 genes to rule out duplicates arising from the tautological terms ‘aminotransferase’ and ‘transaminase’. This identified 17 duplicate entries, which were then removed.

HMM screening with the HMMER tool was used to identify the sequences for which a HMM model exists that is associated with a given transaminase subgroup, as set out by Hwang *et al*. (2005). This procedure revealed three proteins: C4R864, C4R277 and C4R194, which did not match with any transaminase class by HMM profiling. UniProt entries for these three proteins also did not contain the search term 'pyridoxal-5′-phosphate'. Future investigation of the structure of these proteins may resolve their status as transaminases.

A standard nomenclature for K. phaffii GS115 transaminases

The search methodology described above extracted 19 putative transaminases from the *K. phaffii* GS115 genome (table 2). We used ClustalX v.2.1 to align the sequences and TreeView v.1.6.6 to make a tree diagram to illustrate the relatedness of the sequences and plotted a dendrogram (figure 1). We proposed a systematic nomenclature for the 19 transaminases abbreviating *Komagataella phaffii* to 'Kp' and transaminase to 'Tam' in 'KpTam', followed by modifiers that encompass structural and functional predictions of the Hwang *et al*. (2005) subgroupings (table 2) and phylogenetic relationships (figure 1). The modifier features, firstly the Hwang subgroup based on our HMMER analysis (table 2), secondly phylogenetic branching within a Hwang subgroup, and thirdly, pairings of proteins of high similarity within a branch where they exist (figure 1). For example, in 'KpTam III.2a', 'III' refers to the Hwang subclass III, '2' indicates that the protein sequence falls within the second of at least two phylogenetic branches of proteins within Hwang subgroup III and the 'a' indicates the protein is one of a pair (a and b) of closely related proteins within type 'III.2'. In most cases, closely related sequences appear also to be related with respect to function, where characterization data has been reported (table 2).

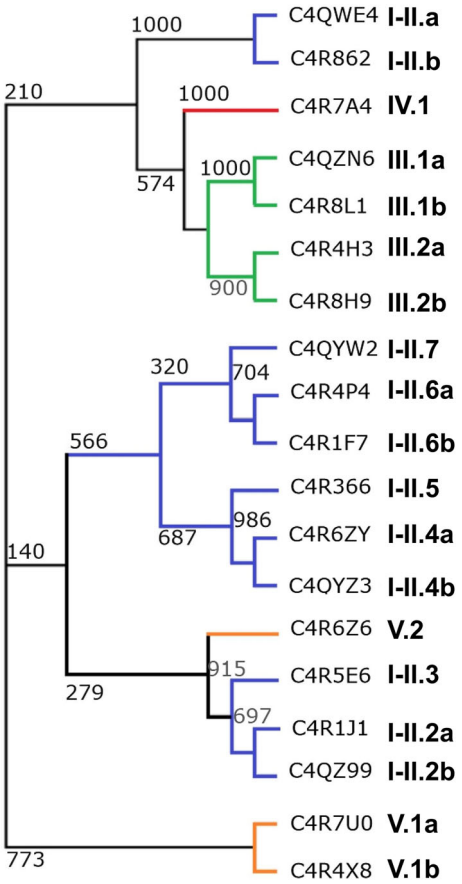

Figure 1. Dendrogram of the 19 *K. phaffii* (*P. pastoris*) GS115 transaminases identified *in silico*. The numbers in nodes are the bootstrap values. Transaminase class assigned by HMMER analysis. Branches in blue indicate class I–II; green, class III; red, class IV; orange, class V. There are no transaminases belonging to class VI. Accession numbers are indicated at the end of each branch followed by systematic nomenclature group in bold.

Putative K. phaffii GS115 class III ω-transaminases

Currently HMMER (Finn *et al*. 2015) does not encompass sufficient mechanistic data to meaningfully predict the topology and chemistry of the active site of a given transaminase. As such, although HMMER analysis can be used to predict the Hwang subgroup of a given transaminase, it cannot predict mechanistic information such as whether a given transaminase is of the β-, γ- or ω- type. Transaminases of the ω-type mechanism from *V. fluvialis* JS17 (Yonaha *et al.* 1992) and *C. violaceum* DSM30191 (Kaulmann *et al.* 2007) have been investigated mechanistically and shown to have industrially favourable substrate ranges. We identified four putative class III transaminases in *K. phaffii* GS115: KpTam III.1a, KpTam III.1b, KpTam III.2a and KpTam III.2b. We then determined the sequence similarity between these four genes and the ω-transaminases of *V. fluvialis* JS17 and *C. violaceum* DSM30191 by a multiple amino acid sequence alignment as a rudimentary *in silico* measure of the likelihood they act on substrates via an ω-type mechanism. Table 3 shows a level of similarity between the four putative *K. phaffii* GS115 class III transaminases and the two known ω-transaminases of 21–28%. Although this is inconclusive, sequence similarity between the two proven ω-transaminases is only 38%, the analysis shown in table 3, thus, at least recommends the four *K. phaffii* GS115 class III transaminases for further investigation to establish if they are in fact ω-transaminases.

We next performed a substitution matrix to align the most highly conserved residues between the four *K. phaffii* class III transaminases, KpTam III.1a, KpTam III.1b, KpTam III.2a and KpTam III.2b, the *C. violaceum* DSM30191-transaminase, F2XBU9, and the *V. fluvialis* JS17 transaminase, Q7NWG4 (figure 2). Previous alignment studies suggested a small number of residues are common to many transaminases, including a glutamic acid of unknown function which is only conserved in ω-transaminases (Mehta

Table 3. Identity percentages obtained in pairwise alignments between the class III transaminases from *K. phaffii* (*P. pastoris*) GS115, and the reference ω-transaminases from *V. fluvialis* JS17 and *C. violaceum* DSM30191.

Systematic nomenclature	*V. fluvialis* JS17 (%)	*C. violaceum* DSM30191 (%)
KpTam III.1a	22.46	20.82
KpTam III.1b	22.68	22.45
KpTam III.2a	25.76	28.14
KpTam III.2b	22.78	26.91

et al. 1993). Shen *et al*. (1998) also identified a conserved threonine residue, understood to form part of the cofactor phosphate-binding site. Three conserved residues are most commonly reported across transaminases alignments studies: aspartic acid, lysine and arginine (Yonaha *et al.* 1992; Mehta *et al*. 1993; Shen *et al.* 1998; Hwang *et al.* 2005; Kaulmann *et al.* 2007). The invariant lysine is understood to participate in Schiff base formation with the 4'-aldehyde group of PLP. The invariant arginine participates in a hydrogen bond / salt bridge with the α-carboxylate group within substrates. The invariant aspartic acid generates a hydrogen bond/salt bridge to protonate the pyridine of the PLP cofactor.

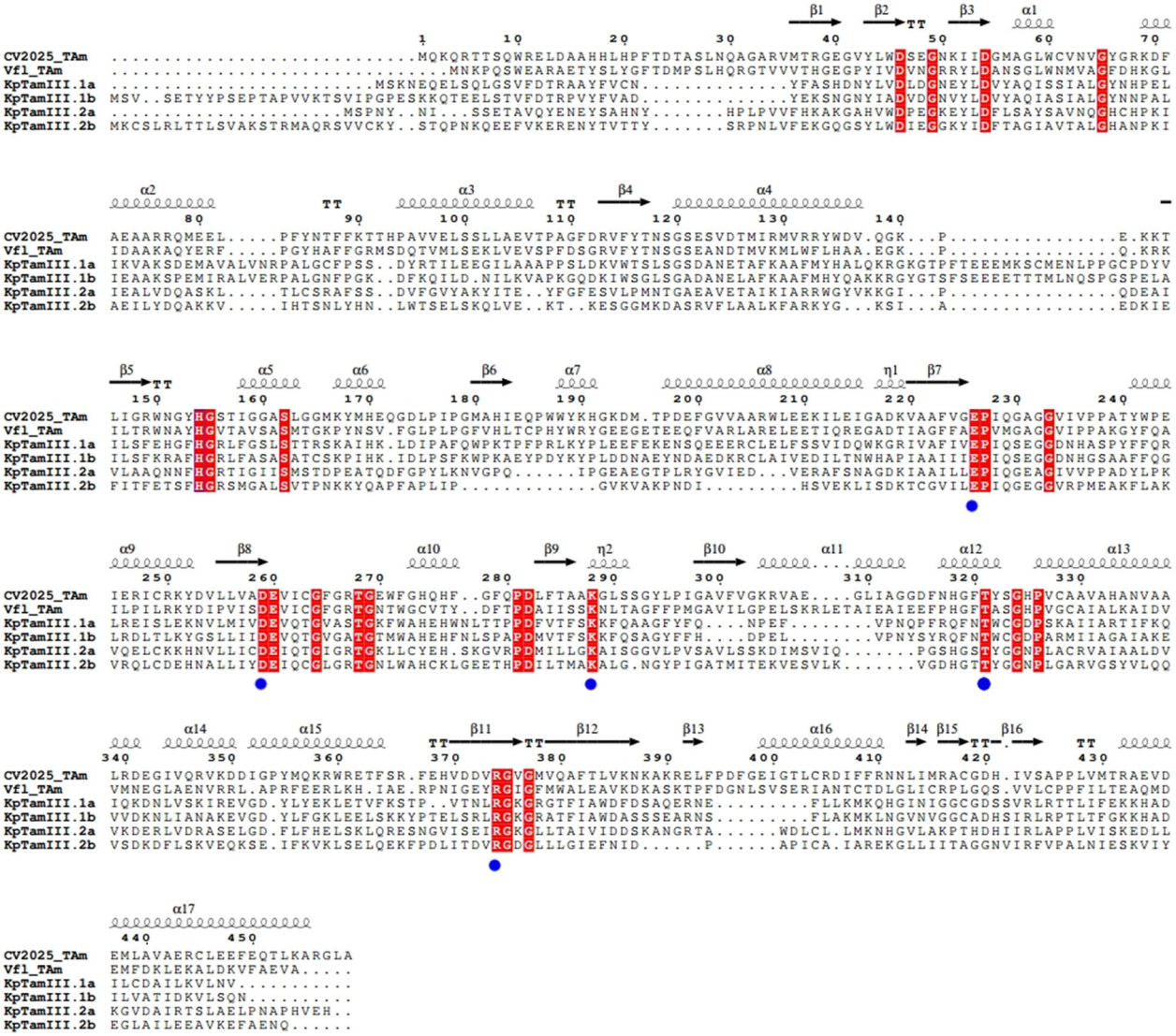

Figure 2. Amino-acid sequence alignment of four putatively class-III transminases from *K. phaffii* (*P. pastoris*) GS115, transaminase F2XBU9 from *V. fluvialis* JS17 and transaminase Q7NWG4 from *C. violaceum* DSM30192. Red colour indicates residues of highest homology by BLOSUM62 scores. Residues conserved in ω-TAms are marked by blue circles. Secondary structure elements in the spatial structure of CV2025 TAm (Sayer *et al.* 2007) are indicated at the top of each block: α-helices are displayed as squiggles, β-strands as arrows, strict β-turns as TT letters and 310-helix as η symbol.

Table 4. Topographical alignment of active site residues between the class III transaminases from *K. phaffii* (*P. pastoris*) GS115 and the reference ω-transaminases from *C. violaceum* DSM30191 and *V. fluvialis* JS17.

C. violaceum DSM30191	*V. fluvialis* JS17	KpTam III.1a	KpTam III.1b	KpTam III.2a	KpTam III.2b
E226	E223	E240	E262	E211	E218
D259	D256	D273	D295	D244	D251
K288	K285	K304	K326	K274	K282
T321	T322	T329	T351	T304	T311
R374	R373	R380	R404	R357	R364

Our alignment (figure 2) identified 24 residues as being highly conserved among the six transaminases, including lysine, aspartic acid, arginine, threonine and the glutamic acid characteristic of ω-transaminases. Table 4 lists the topographically equivalent active site residues in the four putative *K. phaffii* GS115 class III transaminases and the *C. violaceum* and *V. fluvialis* ω-transaminases.

Comparison with K. phaffii CBS7435 and S. cerevisiae S288C

We next compared the transaminase repertoire of *K. phaffii* GS115 to those of two other budding yeasts, *K. phaffii* CBS7435 and *S. cerevisiae* S288C (GenBank assembly accession: GCA_000146045.2), by performing the same data-mining procedures described previously for transaminase identification.

Strain GS115 of *K. phaffii* was originally developed by mutagenesis (Valli *et al.* 2016) of its parental strain *K. phaffii* CBS7435 (Küberl *et al.* 2011). Data mining revealed a repertoire of 20 putative transaminases for *K. phaffii* CBS7435 (table 5). All 19 putative transaminases of *K. phaffii* GS115 had orthologues in the *K. phaffii* CBS7435 genome, with 93.2–100% similarity. One *K. phaffii* CBS7435 transaminase (accession number: F2QVZ3) had a zero similarity with any putative *K. phaffii* GS115 transaminase, suggesting that this gene may have been lost during the mutagenesis procedure used to generate *K. phaffii* GS115.

S. cerevisiae S288C is a highly-characterised and widely-utilized strain whose genome was used as the reference sequence (Engel *et al.* 2014) for the Saccharomyces Genome Database (SGD). The genome of *S. cerevisiae* S288C is also the basis of efforts to synthetically refactor the entire genome of *S. cerevisiae* to improve its industrial utility (Richardson *et al.* 2017). Our data-mining procedure identified 20 putative transaminases within the *S. cerevisiae* S288C genome (table 5), all representing some similarity with those of *K. phaffii* GS115 (27–73% identity). Comparing *S. cerevisiae* S288C and *K. phaffii* GS115, transaminases were illuminating as it revealed KpTam

Table 5. Comparison of transaminases identified *in silico* in *K. phaffii* strains GS115 and CBS7435 and *S. cerevisiae* strain S288C. Hwang class assigned by HMMER HMMscan. Percentage identity shows sequence homology between strain pair.

Systematic nomenclature GS115 protein	Accession number CBS7435	Accession number S288C	Class (HMMER)	% Identity GS115 with CBS7435	% Identity GS115 with S288C
KpTam I-II.1a	F2QML6	Q01802	I–II	100%	40%
KpTam I-II.1b	F2QYY5	P23542	I–II	93.20%	57%
KpTamI-II.2a	F2QS71	P09950	I–II	100%	68%
KpTamI-II.2b	F2QQF8	P25045	I–II	100%	42%
KpTam I-II.3	F2QVJ5	P40970	I–II	100%	62%
KpTam I-II.4a	F2R043	P38840	I–II	100%	34%
KpTam I-II.4b	F2QQ55	P53090	I–II	100%	52%
KpTam I-II.5	F2QUC6	P47039	I–II	100%	56%
KpTam I-II.6a	F2QWA4	P52892	I–II	100%	58%
		P52893			59%
KpTam I-II.6b	F2QSA5	P07172	I–II	100%	53%
				100%	
KpTam I-II.7	F2QQ23	P10356	I–II	100%	40%
KpTam III.1a	F2QTZ9	–	III	100%	–
KpTam III.1b	F2QYI5	P17649	III	100%	68%
KpTam III.2a	F2QWH8	P50277	III	100%	27%
		P07991			70%
KpTam III.2b	F2QYL6	P18544	III	98%	51%
KpTam IV.1	F2QZT3	P38891	IV	100%	73%
		P47176			69%
KpTam V.1a	F2QZA3	P43567	V	100%	48%
KpTam V.1b	F2QW19	P33330	V	100%	59%
KpTam V.2	F2R044	–	V	100%	–
–	F2QVZ3	–	V	–	–

I-II.1b to have 57% identity with an aspartate aminotransferase (accession number P23542) and KpTam I-II.7 to have 40% identity with a 2-aminoadipate transaminase (accession number: P10356).

In conclusion, the information presented here will be useful for those seeking to engineer *K. phaffii* GS115 transaminases to build *de novo* biocatalytic pathways, modification of cellular metabolism and efforts to improve biological understanding of the organism. The data mining results provide a first overview of the transaminases of a methylotrophic yeast species. Our analysis suggested that in the order of 20 transaminases are encoded by the genomes of both *K. phaffii* GS115 and *S. cerevisiae* S288C. Phylogenetically divergent *K. phaffii* GS115 transaminase genes were often predicted to share the functional properties of Hwang *et al*. (2005) subgroup and reactive mechanism.

Acknowledgements

This work was supported by the Government of Chile and its Program CONICYT-Becas Chile (72120390), who supported the doctoral studies of M.-J.H., and the ERA-IB IPCRES Consortium, funded in the UK by the BBSRC (BB/M004880/1), who supported all other authors.

References

Bairoch A. 2000 The ENZYME database in 2000. *Nucleic Acids Res.* **28**, 304–305.

Bollok M., Resina D., Valero F. and Ferrer P. 2009 Recent patents on the *Pichia pastoris* expression system: expanding the toolbox for recombinant protein production. *Recent Pat. Biotechnol.* **3**,192–201

Cho B. K., Cho H. J., Park S. H., Yun H. and Kim B. G. 2003 Simultaneous synthesis of enantiomerically pure (S)-amino acids and (R)-amines using coupled transaminase reactions. *Biotechnol. Bioeng.* **81**, 783–789.

De Schutter K., Lin Y. C., Tiels P., Van Hecke A., Glinka S., Weber-Lehmann J. *et al*. 2009 Genome sequence of the recombinant protein production host *Pichia pastoris*. *Nat. Biotechnol.* **27**, 561–566.

Engel S. R., Dietrich F. S., Fisk D. G., Binkley G., Balakrishnan R., Costanzo M. C. *et al*. 2014 The reference genome sequence of *Saccharomyces cerevisiae*: then and now. *G3 (Bethesda)* **4**, 389–398.

Finn R. D., Clements J., Arndt W., Miller B. L., Wheeler T. J., Schreiber F. *et. al*. 2015 HMMER web server: 2015 update. *Nucleic Acids Res*. **43**, W30–W38.

Finn R. D., Coggill P., Eberhardt R. Y., Eddy S. R., Mistry J., Mitchell A. L. *et al*. 2016 The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*. **44**, D279–D285.

Goffeau A., Barrell B. G., Bussey H., Davis R. W., Dujon B., Feldmann H. *et al.* 1996 Life with 6000 genes. *Science* **274**, 563–567.

Henikoff S. and Henikoff J. G. 1992 Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919.

Hohne M., Schatzle S., Jochens H., Robins K. and Bornscheuer U. T. 2010 Rational assignment of key motifs for function guides in silico enzyme identification. *Nat. Chem. Biol.* **6**, 807–813.

Huson D. H. and Scornavacca C. 2012 Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol*. **61**, 1061–1067.

Hwang B-Y., Cho B-K., Yun H., Koteshwar K. and Kim B.-G. 2005 Revisit of aminotransferase in the genomic era and its application to biocatalysis. *J. Mol. Cat. B: Enzym*. **37**, 47–55.

Ingram C. U., Bommer M., Smith M. E., Dalby P. A., Ward J. M., Hailes H. *et al.* 2007 One-pot synthesis of amino-alcohols using a de-novo transketolase and beta-alanine: pyruvate transaminase pathway in *Escherichia coli*. *Biotechnol. Bioeng*. **96**, 559–569.

Invitrogen 2010 Multi-Copy *Pichia* Expression kit. For the isolation and expression of recombinant proteins from *Pichia pastoris* strains containing multiple copies of a particular gene. User Manual part no. 25–0170.

Kaulmann U., Smithies K., Smith M. E. B., Hailes H. C. and Ward J. M. 2007 Substrate spectrum of ω-transaminase from *Chromobacterium violaceum* DSM30191 and its potential for biocatalysis. *Enzyme Microb. Technol.* **41**, 628–637.

Krogh A., Brown M., Mian I. S., Sjolander K. and Haussler D 1994 Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol*. **235**, 1501–1531.

Küberl A., Schneider J., Thallinger G. G., Anderl I., Wibberg D., Hajek T. *et al.* 2011 High-quality genome sequence of *Pichia pastoris* CBS7435. *J. Biotechnol*. **154**, 312–320.

Kurtzman C. P. 2005 Description of *Komagataella phaffii* sp. nov. and the transfer of *Pichia pseudopastoris* to the methylotrophic yeast genus Komagataella. *Int. J. Syst. Evol. Microbiol*. **55**, 973–976.

Larkin M. A, Blackshields G, Brown N. P., Chenna R., McGettigan P. A. *et al*. 2007 Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948.

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

Malik M. S., Park E. S. and Shin J. S. 2012 Features and technical applications of omega-transaminases. *Appl. Microbiol. Biotechnol.* **94**, 1163–1171.

Mattanovich D., Callewaert N., Rouze P., Lin Y. C., Graf A., Redl A. *et. al* 2009 Open access to sequence: browsing the *Pichia pastoris* genome. *Microb. Cell Factories* **8**, 53.

Mehta P. K., Hale T. I. and Christen P. 1993 Aminotransferases: demonstration of homology and division into evolutionary subgroups. *Eur. J. Biochem*. **214**, 549–561.

Qian B. and Goldstein R. A. 2003 Detecting distant homologs using phylogenetic tree-based HMMs. *Proteins* **52**, 446–453.

Richardson S. M., Mitchell L. A., Stracquadanio G., Yang K., Dymond J. S., DiCarlo J. E. *et al.* 2017 Design of a synthetic yeast genome. *Science* **355**, 1040–1044.

Robert X. and Gouet P. 2014 Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res*. **42**, W320–W324.

Sayer C., Isupov M. N. and Littlechild J. A. 2007 Crystallization and preliminary X-ray diffraction analysis of omega-amino acid: puruvate transaminase from Chromobacterium violaceum. *Acta Crystallogr. Sect. F Strct. Biol. Cryst. Commun*. **63**, 117–119.

Shen B. W., Hennig M., Hohenester E., Jansonius J. N. and Schirmer T. 1998 Crystal structure of human recombinant ornithine aminotransferase. *J. Mol. Biol.* **277**, 81–102.

Shin J. S., Yun H., Jang J. W, Park I. and Kim B. G. 2003 Purification, characterization, and molecular cloning of a novel amine:pyruvate transaminase from *Vibrio fluvialis* JS17. *Appl. Microbiol. Biotechnol.* **61**, 463–471.

Templar A., Woodhouse S., Keshavarz-Moore E. and Nesbeth D. N. 2016 Influence of *Pichia pastoris* cellular material on polymerase

chain reaction performance as a synthetic biology standard for genome monitoring. *J. Microbiol. Methods* **127**, 111–122.

Valli M., Tatto N. E., Peymann A., Gruber C., Landes N., Ekker H. *et al*. 2016 Curation of the genome annotation of *Pichia pastoris* (*Komagataella phaffii*) CBS7435 from gene level to protein function. *FEMS Yeast Res.* **16**, fow051.

Van der Klei I. J., Yurimoto H., Sakai Y. and Veenhuis M. 2006 The significance of peroxisomes in methanol metabolism in methylotrophic yeast. *Biochim. Biophys. Acta* **1763**, 1453–1462.

Ward J. and Wohlgemuth R. 2010 High-yield biocatalytic amination reactions in organic synthesis. *Curr. Org. Chem*. **14**, 1914–1927.

Waterhouse A. M., Procter J. B., Martin D. M., Clamp M. and Barton G. J. 2009 Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191.

Wei Y. C., Braun-Galleani S., Henriquez M. J., Bandara S. and Nesbeth D. N. 2018 Biotransformation of beta-Hydroxypyruvate and Glycolaldehyde to L-Erythrulose by *Pichia pastoris* strain GS115 overexpressing native Transketolase. *Biotechnol. Prog*. **34**, 99–106.

Yonaha K., Nishie M. and Aibara S. 1992 The primary structure of omega-amino acid:pyruvate aminotransferase. *J. Biol. Chem.* **267**, 12506–12510.

Corresponding editor: H. A. Ranganath