Dipl.-Ing. Lukas Sturmberger

ENZYME DISCOVERY FOR BIOPOLYMER DEGRADATION AND MODIFICATION

DOCTORAL THESIS
to achieve the university degree of
Doktor der technischen Wissenschaften

submitted to
Graz University of Technology

Supervisor
Ao.Univ.-Prof. Mag.rer.nat. Dr.rer.nat. Anton Glieder
Institut für molekulare Biotechnologie

Graz, August 2020

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present doctoral thesis.

Date, Signature

Preamble

This thesis was initiated with the attempt to identify novel enzymes for biopolymer degradation and modification. I have set out to investigate complementary approaches and combine them to test their suitability and applicability for directed enzyme discovery. In doing so, I employed genomics-, transcriptomics- as well as proteomics-based techniques coupled with functional screenings in order to find novel proteins actively expressed in the eukaryotic host *Komagataella phaffii* (*Pichia pastoris*).

The most significant output of this thesis is the successful employment of synergistic approaches in enzyme discovery projects. Throughout this thesis I was able to show, that a course of action which incorporates several approaches in parallel leads to the successful identification of candidate proteins. The synergy between sequencing technologies such as RNAseq and the possibilities attained through bioinformatic homology searches coupled with functional screenings turned out to be a powerful tool in the overall process to discover surface-active hydrophobin proteins. In a different part of the thesis a synergy of approaches again proved to be very successful in identifying functionally expressed novel enzymes. By combining mRNA sequencing, proteomics and functional expression a novel thermostable beta-glycoside hydrolase was discovered. The technical challenges for novel enzyme discoveries from eukaryotic organisms throughout this thesis also demonstrated the need for a reliable and robust expression system capable of producing a multitude of different enzyme classes. In the yeast system *K. phaffii* (*P. pastoris*) we were able to find such a system, further improve its broad applicability and future potential by providing a completely new reference genome sequence and successfully employ it for the expression of enzyme candidates from the discovery stream.

Acknowledgements

First and foremost, I would like to thank the love of my life, my husband Markus, for his never-ending support and patience throughout all these years. I would not have been able to complete this journey and keep my sanity without you on my side – you are my bridge over troubled water, my best friend and my greatest confidant – I love you!

I would also like to thank my entire family, my parents Bettina and Kurt, my grandparents Ernst and Ida and my aunt Evelyn for shaping me into the person I am today. This thesis is as much your effort as it is mine. Thank you for your support throughout these many years of studies and for always encouraging me to strive for the best.

I also want to express my thanks to my supervisor Prof. Anton Glieder for the support of my PhD thesis. My sincere gratitude also goes to Prof. Ruth Birner-Grünberger, Prof. Gerhard Thallinger and all my co-authors and colleagues who ventured on this journey with me and contributed their knowledge, time and energy to this project.

At this point I also want to thank ACIB GmbH and Clariant Produkte Deutschland GmbH as well as all involved people for providing me with the funding and framework that allowed me to realize this project in the first place.

Abstract

The rise of biocatalytic applications in industry and academia has resulted in an increasing demand for enzymes capable of performing desired reactions under different conditions. The field of enzyme discovery deals with the search and identification of biocatalysts and their utilization for the synthesis, modification and degradation of specific compounds. This thesis explores complementary approaches for enzyme discovery and attempts to combine them to identify novel proteins. A combinatorial approach of mRNA sequencing, sequence-homology based searches and functional screening allowed the discovery of amphiphilic proteins in the transcriptome of the fern species *Pteridium aquilinum*. Five different surface-active hydrophobin sequences could be identified and were characterized for their surface-tension lowering activities as well as their suitability as a coating agent for in-frame fused horseradish peroxidase (HRP). In a different set of experiments a combinatorial approach of mRNA-sequencing and proteomics coupled with functional expression was postulated and successfully employed for the discovery of a novel thermostable beta-glycoside hydrolase in the transcriptome of the entomopathogenic fungus *Lecanicillium attenuatum*. In order to improve the widely used yeast expression system *Komagataella phaffii* (*Pichia pastoris*) we generated a new reference genome sequence with a special emphasis on intron prediction. The integration of Pacific Biosystems sequencing technologies as well as large RNAseq data sets allowed us to correct deletions, insertions, gaps, rearrangements and splicing variants in already published sequences and greatly enhanced our knowledge and capabilities for cell-engineering and protein expression within this thesis. The significance of this contribution lies in its focus on synergies between multiple streams of omics technologies and shows that a combinatorial approach leads to the successful identification of target proteins. Furthermore, it emphasizes the need for a robust expression system capable of producing significant amounts of biocatalysts.

Kurzzusammenfassung

Die zunehmende Anwendung von Biokatalysatoren in der Industrie und der wissenschaftlichen Forschung führt zu einer erhöhten Nachfrage an Enzymen, die in der Lage sind, spezifische Reaktionen unter unterschiedlichen Bedingungen zu katalysieren. Das Feld der Enzymentdeckung beschäftigt sich mit der Suche und Entdeckung geeigneter Biokatalysatoren und deren Anwendung für die Synthese, die Modifikation und den Abbau spezifischer Verbindungen. Im Rahmen dieser Arbeit wurden komplementäre Herangehensweisen analysiert und kombiniert, um neuartige Proteine zu identifizieren. Eine Kombination aus mRNA Sequenzierung, Sequenzhomologie-Suche und funktionalem Screening erlaubte die Entdeckung von amphiphilen Proteinen in den Transkriptomdaten des Farns *Pteridium aquilinum*. 5 oberflächenaktive Hydrophobine konnten entdeckt und charakterisiert werden. Es wurde untersucht, ob ein HRP Fusionsprotein eine spezifische Bindung an Glasoberflächen ermöglicht. In einem weiteren Teil wurde eine Kombination aus mRNA Sequenzierung, Proteomics und funktionaler Expression postuliert und erfolgreich für die Identifikation einer thermostabilen beta-Glykosidhydrolase im Transkriptom des Pilzes *Lecanicillium attenuatum* eingesetzt. Im Zuge der Verbesserung des häufig verwendeten Hefe Expressionssystems *Komagataella phaffii* (*Pichia pastoris*) konnte zusätzlich eine neue Referenzgenomsequenz erstellt werden. Die Verbindung der Pacific Biosystems Sequenziertechnologie und eines großen RNAseq Datensatzes erlaubte die Korrektur von Insertionen, Deletionen, Lücken und falscher Splicing Vorhersagen in bereits publizierten Sequenzen und ermöglichte es dieses System für die Proteinexpression in dieser Dissertation einzusetzen. Die Bedeutung dieser Thesis liegt vor allem in der Synergie zwischen verschiedenen Technologiesträngen und hat gezeigt, dass eine Herangehensweise, die verschiedene Methoden kombiniert, für die Entdeckung von Zielproteinen erfolgreich eingesetzt werden kann. Weiter konnte auch die Bedeutung eines robusten Expressionssystems, dass die Produktion von größeren Enzymmengen erlaubt, aufgezeigt werden.

CONTENTS

Introduction

Why enzyme discovery? - The need for novel biocatalysts

The concern for the reduction in carbon emissions, the depletion of finite natural resources as well as the use of environmentally problematic substances on an industrial scale (e.g. organic solvents) has driven the development of alternatives to the "classical" chemical synthesis of compounds (Anastas and Warner, 1998; Bornscheuer et al., 2012; Meyer, 2011; Sheldon, 2014). Biocatalysis is a key technology that allows the use of enzyme catalysts for organic synthesis. Lately, the use of enzymes in diverse fields is increasingly supplementing the classical organo- and metallo-catalysis (Reetz, 2013). While organo-catalysis still dominates major industrial operations, especially in the area of asymmetric synthesis, biocatalysts have proven themselves as viable alternatives to metal catalysts (Catalysis, 2018; Schulze and Wubbolts, 1999). The discovery of novel enzymes from environmental sources and the capability to tailor-make enzymes for specific applications by enzyme engineering opens up applications ranging from bulk commodity chemicals to highly specialized pharmaceutical intermediates. Enzyme catalysts are also employed in the food and fine chemical industry (Schoemaker, 2003; Truppo, 2017). Additionally, biocatalysts are not only used for synthesis but also for the degradation of chemical compounds. They assist in in the degradation of xenobiotics, plastics, biomass and unwanted components in wastewater in a very targeted and controlled manner (Cammarota and Freire, 2006; Eberl et al., 2008; Janssen et al., 2005; Kullman and Matsumura, 1996; Müller et al., 2005; Rabinovich et al., 2004; Ribitsch et al., 2012).

Several hundred industrial applications of biocatalysts have been implemented so far, resulting in a continuously growing number of published reactions (Clouthier and Pelletier, 2012; Erickson et al., 2012; Li et al., 2012a; Schmid et al., 2001; Woodley, 2013) with the majority of those being hydrolytic enzymes and oxidoreductases (Faber, 2011). Exact numbers on the status of the enzyme market are difficult to come by and only estimations are available. The global industrial enzyme market has increased steadily in consecutive years and is expected to rise to 6,3 Billion USD in 2022 at a compound annual growth rate of 5,8% from 2017 (https://www.marketsandmarkets.com/PressReleases/industrial-enzymes.asp).

Depending on the specialized application, enzyme characteristics differ quite drastically between highly engineered enzymes and naturally evolved ones. A prominent example found in the pharmaceutical industry is the halohydrin dehalogenase for the production of a key intermediate in the production of atorvastatin (cholesterol lowering drug) engineered via directed evolution. In combination with a KRED and a glucose dehydrogenase as cofactor regeneration system the drug Lipitor is produced (Ma et al., 2010; Patel, 2009). This application exemplifies the strengths of enzymes for high regio- and stereoselectivity during the synthesis of chiral intermediates. Concerning the production of bulk chemicals, prominent examples include protease and lipases used as laundry detergents (Banerjee et al., 1999; Jaouadi et al., 2008; Moreira et al., 2002; Sellami-Kamoun et al., 2008) or the isomerization of glucose to fructose using immobilized glucose isomerase (Bhosale et al., 1996; Takasaki, 1966).

With an increase in biocatalytic synthesis, also the need for novel enzymes rises. The natural biodiversity poses both opportunities but also some challenges. While our environment provides a huge abundance of enzyme sequences, a large number of currently employed biocatalysts stem from a limited number of organisms (Robertson and Steer, 2004). These limitations are most likely due to the inability of culturing the majority of naturally occurring microorganisms in a lab environment. According to estimates, less than 1% of microbes are accessible to cultivation due to the lack of essential nutrients from co-cultures in close proximity, slow growth behavior or simply unknown cultivation conditions (Amann et al., 1995; Ekkers et al., 2012). If these obstacles can be overcome, the seemingly endless natural diversity can be harnessed which would in turn allow the replacement of an even higher number of organo-chemical synthesis processes with biocatalytic driven reactions.

Genomics and transcriptomics-based discovery processes

Genomics and sequence homology-assisted identification of novel biocatalysts

So far, the majority of industrially employed enzymes derive from metagenomic and genomic library screenings (Ferrer et al., 2009; Uchiyama and Miyazaki, 2009). The ability to link genomic sequence information with enzymatic functions from activity screenings exemplifies the importance for genomics research. This synergy allowed the prediction of enzymatic functions solely based on the primary sequence - termed functional genomics (Lacerda and

Reardon, 2008). Based on the similarity to sequences deposited in databases, a putative function can be assigned (O'Leary et al., 2016). Additionally, the occurrence of sequence motifs and conserved amino acid residues can give valuable cues about the potential function of a gene sequence (Marchler-Bauer et al., 2015). However, certain open reading frames in genomic data sets bear no sequence homology to deposited proteins with known functions. Furthermore, even in case of high sequence identity, the enzymatic function of the closest relative sequence might not be annotated correctly. A completely novel enzyme fold or as yet unidentified sequence motif would most likely not appear in a search result based on the degree of relatedness to annotated sequences. It is these sequences, for which it is challenging to estimate protein function solely based on DNA or protein sequence information (Gray et al., 2015; O'Leary et al., 2016). Common to all *in silico* methods is that the result of the analysis remains a prediction of function which needs to be verified by cloning and expression of the respective gene sequence followed by enzyme function determination via enzymatic assays.

The basis for enzyme discovery with "omics" technologies is a gene or protein sequence database of a single or of multiple organisms. This can be achieved by benefiting from publicly deposited genome or proteome sequences. The National Center for Biotechnology Information (NCBI) is one of the major sources for sequence data besides the European Molecular Biology Laboratory (EMBL) and the DNA Data Bank of Japan (DDBJ). These databases are updated regularly and integrated through the International Nucleotide Sequence Database Collaboration (INSDC). For the generation of suitable databases also additional resources exist such as the Uniprot Knowledge database (UniProtKB) (Consortium, 2015) or the Joint Genome Institute's genome portal (JGI) (Nordberg et al., 2014). In case there exists no sequence information for the target organism in any public database, and also in case of complex environmental samples, custom databases specific to the sample can improve database quality (Sturmberger et al., 2016b). In case the generation of a sample-specific database is challenging, a collection of more or less related organisms can be used. The major drawback in doing so is however, the lack of knowledge of the specific sequence, which in turn prevents cloning, recombinant expression and a general study of the target (Bräutigam et al., 2008). In case the target organism or specific protein sequence is not covered by the underlying database, the discovery, especially of novel enzyme sequences, can be challenging and might not yield satisfactory results. To prevent this loss, a sample-specific transcriptome sequence and protein database can be prepared.

In the last couple of years, advances in the field of DNA and RNA sequencing have made it possible to sequence transcriptomes and entire genomes via *de novo* sequencing (Grabherr et al., 2011). The third generation of sequencers, most prominently the Pacific Biosciences sequencing platform have increased the read length of single reads to 15kb, in some cases even 20kbs (Ferrarini et al., 2013; Mccarthy, 2010). This extraordinary increase opens up the possibility to perform sequencing drastically decreasing the need for bioinformatic assembly. With older technologies several shorter reads had to assembled to a larger contig to yield a genome or transcriptome sequence and therefore the coding sequence for a candidate enzyme (Quail et al., 2012). Based on the genetic composition of the sample, the assembly process can be challenging, as gene isoforms, high AT sequence motifs and repeat structures pose difficulties to the current technologies. The misalignment of reads might therefore result in the loss of possible new enzyme candidates (McGettigan, 2013).

In particular, the ability to construct libraries and the developments in sequencing technologies resulted in a drop regarding price per sequenced base while the availability of sequencing services by several companies greatly enhanced the distribution of this technology. The basic workflow for sequencing projects is the extraction of DNA or RNA from either environmental samples or cultured organisms. Following extraction, any rRNA contamination is removed and normalization is performed to enrich for less abundant sequences. In case of RNA, reverse transcription to generate cDNA is performed, which is used for library preparation and sequencing. However, the raw data originating from sequencing still need to be aligned which creates several challenges. The misalignment of sequences with repetitive nature or slight variations (gene isoforms) termed shadowing (McGettigan, 2013) still poses problems in identifying the correct sequence. Although algorithms for de novo assembly exist (Hölzer and Marz, 2019), it still requires a specialized knowledge and access to the suitable computational facilities (Manfred G. Grabherr, Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W. and Friedman, 2013).

Obstacles associated with genomics and meta-genomics assisted discovery

In contrast to genome sequencing and assembly, the cloning and expression of entire genomes and metagenomes in suitable screening vectors circumvents these problems and enables the identification of biocatalyst with previously unknown functions and folds (Ferrer et al., 2009; Lanfranchi et al., 2017). A large number of studies have shown the applicability of functional screening approaches (Adrio and Demain, 2014; Davids et al., 2013; Hess, 2011; Kim et al., 2011; Lanfranchi et al., 2015; Rabausch et al., 2013; Taupp et al., 2011; Ufart et al., 2015; Warnecke and Hess, 2009). The largest caveat involving genomic and metagenomic libraries is the complexity and size of eukaryotic genomes. While their bacterial and archaeal counterparts display smaller overall sizes, eukaryotic genes tend to be more complex with several layers of regulatory elements. Above all the tendency to change the sequential arrangement of genes through splicing and shuffling (Deutsch and Long, 1999) makes the generation of genomic and meta-genomic sequences from eukaryotic origins challenging. Additionally, more factors need to be considered when constructing these gDNA based libraries. Recently, the borderline between homology- and activity-based screening was bridged. Steinkellner and colleagues showed that structure guided methods could supplement *in silico* methods by comparing a set of residues within the active site from proteins with known function, the catalophore, to unknown proteins thereby predicting enzyme function based on active site architecture (Steinkellner et al., 2014).

Transcriptome and meta-transcriptome driven discovery and its disadvantages

Applying transcriptomics and meta-transcriptomics circumvents some of the disadvantages associated with genome sequencing approaches (Warnecke and Hess, 2009). Since samples for library preparation are enriched for coding mRNA, the majority of non-coding transcripts and non-transcribed genes are removed thereby decreasing library size. However, this enrichment necessarily excludes genes that are not actively transcribed in the chosen culture conditions. It might therefore be necessary to include several samples grown under different environmental influences. This issue could on the other hand be used as an asset for cases where mRNA libraries enable the elucidation of rarely expressed genes, which might facilitate the discovery of biocatalysts directly or indirectly involved with a specific condition, e.g. presence of an inductor or certain metabolite.

The transcriptomics approach can be further complemented by analyzing the protein content derived from the mRNA information of the cell. The major advantage of proteomics in contrast to transcriptomics is the precise measurement of protein abundance, as mRNA abundance does not accurately reflect protein abundance (Gygi et al., 1999; Zhang et al., 2014). Furthermore, any posttranslational modifications crucial for protein function can be analyzed and inferred based on the proteomics data (Perchey et al., 2019). The issue of subcellular localization, especially in the case of eukaryotic species is difficult to answer based on the primary protein sequence. Although many different algorithms for the prediction of localization within the cells of different organisms exist (Almagro Armenteros et al., 2019; Frank and Sippl, 2008; Käll et al., 2007), it is proteomics which allows the direct analysis of subcellular compartments without the need of any predictive tools. The compartments in which proteins can be found are diverse and range from the cytosol, the secretory pathway (ER, Golgi Body and vesicles), peroxisomes, mitochondria, lysosomes or in the extracellular space (Tashiro, 1983). Each compartment of the cell represents a different condition in respect to pH, ionic strength, the presence of cofactors and protein partners and the respective oxidative environment (Go and Jones, 2014). The knowledge about the different media requirements can provide valuable cues about the circumstance under which the protein exhibits its optimal function and stability. Any insights gained in this manner can be incorporated into functional assays used for screening the respective enzymatic activities, in turn increasing the chances to find a protein with desired catalytic properties. However, I will focus both on the functional screening as well as on the use of proteomics for enzyme discovery in a later section.

Construction and functional screening of enzyme libraries

cDNA library generation as a basis for enzyme discovery

To generate transcriptomic libraries, the provision of proper RNA material is necessary. The ability to provide high quality genetic material is the basis for library construction, which differs between genomic and transcriptomic libraries. Depending on the transcriptional state, certain genes can be highly enriched in a sample whereas others might not be expressed at all and therefore lack from the library entirely. Furthermore, a high abundance of certain genes can displace lower abundant genes from the library, essentially eliminating them in favor of themselves. To capture also low abundant genes in a cDNA library, normalization of the library should be performed. The normalization procedure can eliminate a large part of the highly abundant transcripts from the library by several different mechanisms. The most common techniques used consist of the DSN-normalization (Bogdanova et al., 2008), normalization by hydroxyapatite chromatography (Vandernoot et al., 2012) or normalization by self-annealing (Patanjali et al., 1991). These treatments allow the capturing of rare transcripts, a necessary step considering the composition of a typical eukaryotic transcriptome. The majority of transcripts, more than 90%, are made up of a few highly transcribed genes, while the remaining 10% are divided into medium transcribed genes and a large part of extremely rare transcripts, which are only present with a few copies or in the worst case not at all (Soares et al., 1994). The transcriptional state of a certain gene can be influenced by the culturing conditions used to grow the cells. Therefore, the enrichment of rare transcripts can be expedited by choosing conditions conducive with the transcription of a certain gene. To increase the share that a polymer degrading enzyme has within a library for instance, the cells can be co-cultured with the polymer thereby inducing the transcription of genes necessary for degradation and metabolization of the substrate. The same stands for the enrichment of enzymes capable of degrading a toxic compound.

A further issue during cDNA library construction that needs to be addressed is the fragmentation of transcripts and the presence of truncated reverse transcribed transcripts. Premature termination of transcription on the one hand and incomplete reverse transcription on the other hand can both lead to the same phenomenon - a partial coding sequence. While premature termination of transcription can lead to in vivo degradation of the sequence due to a lack in poly-adenylation signals, the incomplete reverse transcription is more challenging to circumvent. The standard method to reverse transcribe protein coding RNA is the employment of poly-T primers binding to the poly-A tail of protein coding transcript as a priming site. Poly-T primers harbor between 18-25 Thymidine residues and specifically bind the poly-A tail. However, these primers can also bind the homopolymer stretches consisting of several adenosines occurring within a coding sequence and therefore produce N-terminal truncations of proteins. The likelihood of this occurrence can be minimized by using anchoring primer (Nam et al., 2002). Anchoring primer in cDNA generation consist of a stretch of Thymidin residues followed by two bases with an equal combinatorial distribution of A, T, G and C – essentially forming an anchor, hence the name anchor primer. This set-up improves the binding by minimizing the number of residues not taking part in hybridization during primer annealing through binding to the outermost 3’end of the poly-A tail rather than binding shorter internal stretches of adenosines within an mRNA transcript. Therefore, mis-priming and the generation of truncated cDNA transcripts can be minimized. With this approach however, broken, truncated or incomplete transcripts lacking a poly-adenylation signal will not be part of the created cDNA library pool (Nam et al., 2002).

To circumvent the potential loss a different approach can be used. The employment of primers with a randomized sequence, most often random hexamer primers, will also capture transcripts lacking any polyadenylation signal. While this approach will solve the problem of transcript loss due to lacking poly-adenylation signals, it creates a different challenge – random priming not only produces partial transcripts such as caused by primers binding within a transcript, but will however also amplify and translate non-mRNA into cDNA. The majority of the total RNA pool from eukaryotic species is made up of non-coding RNA such as ribosomal RNA (rRNA), long non-coding RNA (lcnRNA) or small interfering RNA (siRNA), while messenger RNA (mRNA) transcripts coding for proteins only make up a small proportion of the entire pool (David et al., 2006; Drinnenberg et al., 2009). The cDNA generated by these random primers will inadvertently also include the non-coding RNA molecules which have to be removed later on by normalization techniques to enrich for coding transcripts.

Once the majority of coding RNA transcripts has been converted to cDNA the cloning into suitable library expression vectors is possible. The structure of the reverse transcribed cDNA with its directionality will only allow the production of protein once the right orientation is chosen, with the core promoter regions and start-ATG at the 5‘-end of the transcript and the stop codon and polyadenylation site at the 3‘-end. Direction-less cloning such as blunt end cloning for instance will orient the transcript by chance and therefore would produce half the pool of cDNA orientated in a way uncapable of producing protein. It is therefore necessary to either guarantee directional cloning or to rely on a plasmid set-up that allows transcription in both orientations. Although such promoter set-ups exist and have been used for gDNA library generation in E. coli (Lämmle et al., 2007), the more common approach is to use a directional cloning technique via the introduction of adapter sequences. The binding of a poly-T primer to the poly-A tail of mRNA transcripts concurrently allows the introduction of adapter sequences used for later cloning steps. The incorporation of adapter sequences at the 5‘-end of transcripts however turns out to be more challenging, as the core promoter region lacks a common motive for primer design. Two different approaches can be used to resolve this issue. For one, the adapter sequences can be added by blunt-end ligation (Teeri et al., 1987). This process has a very low efficiency and additionally also risks adding identical adapter sequences at the undesirable 3‘end of the transcript (which already harbor adapter sequences). The more elegant solution is the employment of an enzymatic side activity of several reverse transcriptases, termed template switching. When a reverse transcriptase such as the Maloney mouse leukemia virus reverse transcriptase (MMLV-RT) scans along the mRNA (commencing from the polyT primer poly-A tail) and reaches the 5’end of the mRNA transcript it will eventually seize the reverse transcription and fall off or alternatively, if presented with RNA residues on a different molecule it will switch template and continue to read and reverse transcribe these parts onto the same cDNA template (Wellenreuther et al., 2004; Zhu et al., 2001). Furthermore, the same reverse transcriptase shows the tendency to add a triple C (CCC) residue at the 5’end of the transcript. These two occurrences can be used to incorporate adapter sequences in a directional manner to the 5’end of protein coding mRNA transcripts. The cDNA library now harbors unique sequences on both ends, which allow the directional cloning either via restriction digestions and ligation or restriction-free cloning techniques (Gibson et al., 2009; Quan and Tian, 2009).

Gene library expression and screening procedures

The capability to directly subclone desired genomic and metagenomic sequences from environmental sources circumvents issues with the cultivability of source organisms (Amann et al., 1995; Handelsman, 2004; Healy et al., 1995). This allows access to an extremely large pool of sequences only limited by the ability to process samples in a way that is amenable to downstream applications such as library preparation and cloning. Whole microbial communities and also key players within these communities harboring desirable enzymatic functions, which were impossible to access by classical enrichment and culturing techniques, can be discovered in this manner (Ekkers et al., 2012; Staley and Konopka, 1985). However, while function-based screenings are a viable alternative to unearth novel genes with desired functions, several issues need to be addressed to ensure optimal conditions for library preparation and screening.

Primarily, the quality of starting material from which either whole RNA or DNA is isolated determines the probability of obtaining nucleic acids with sufficient quality for further molecular biological manipulations (Friar, 2005; Nacke et al., 2016; Nordgård et al., 2005;

Tang et al., 2008; Valentin et al., 2005). Currently, a large number of commercial kits exist for the isolation of both RNA and DNA samples deriving from diverse sources such as water, soil, and plant material but also from challenging samples such as sludge waste and feces.

Secondly, the likelihood of identifying a certain gene of interest depends on its expressibility in the respective host. The choice of expression host in relation to the originating species has profound influence on issues like posttranslational modifications, codon usage and promoter recognition (Behrens et al., 2011). It is therefore a challenging endeavor to find a strategy with the highest probability of success in discovering novel enzymes. One of the most important factors is the interplay between the host and vector system. It is often challenging to functionally produce a desired protein when there is only insufficient expression from the fragment in the respective host. The presence of post-translational modifications such as glycosylation, phosphorylation or myristylation might be beneficial or even required for proper enzyme function (Audagnotto and Dal Peraro, 2017; Ryšlavá et al., 2013). Especially in libraries with a considerable content of eukaryotic sequences, the ability of the expression host to properly modify target enzymes plays an important role. The decision on an expression host closely related to the species diversity to be expected at the highest abundancy within the sample can increase the chances of successful expression (Li et al., 2005; Martinez et al., 2004).

Furthermore, the expression and secretion of a recombinant protein can be improved by co-expression of folding and secretion helper factors. Protein factors influencing successful expression are for instance folding helper resident in the cytosol, Golgi Body and ER that improve expression and secretion (Guan et al., 2016; Piva et al., 2017; Yu et al., 2017). As the target enzyme might harbor a secretion signal which directs the protein into the secretory pathway in its native host, the recognition of this signal will impact its subcellular localization also within the heterologous expression host (Hou et al., 2012; Idiris et al., 2010c). A countermeasure, albeit with considerable increase in both time and work effort, would be the expression and screening of libraries within more than one host.

Additionally, the absolute size and the abundance of the respective gene sequence within a complex sample play an important role in the probability of identifying certain enzymes. In case the fragments, either genomic or transcriptomic in nature, can be cloned successfully, the challenge of host interaction with the orthologous genetic elements remains. The genetic machinery of the host system and the interplay with the recombinant gene fragments can potentially cause issues when the host cell fails to recognize transcriptional and/or translational signals necessary for expression. The vector design cannot take into account the diverse regulatory elements present in a complex metagenomic sample and is rather relying on system promiscuity in decoding signals for proper transcription and translation. This exemplifies how important the relationship between the genetic source and the choice of host system is for successful expression. Especially in the case of complex metagenomic samples, the organismal origin of a specific sequence can be challenging to predict, as a multitude of different organism will be present concurrently. Therefore, the design of a vector system suitable for differently regulated orthologous genes from several species is an important factor. A possible way to circumvent compatibility issues is the generation of shuttle vector systems which enable transcription of cloned inserts in several different species and can therefore be used in more than one host. The existence of eukaryotic promoter systems derived from plant or yeast species which also show transcriptional activity in bacteria could potentially circumvent any gene-host incompatibility issues (Antonucci et al., 1989; Jacob et al., 2002; Lewin et al., 2004a). Other possible approaches to increase the chances of transcription include for example the design of vectors with dual-orientation promoters (Lämmle et al., 2007).

High-throughput assay development

Even if a library can be expressed successfully, a suitable high-throughput screening procedure needs to be available, as thousands of library clones need to be assayed for a desirable enzymatic function (Davids et al., 2013; Taupp et al., 2011). A crucial aspect of functional library screening is the availability of suitable enzyme assays. The basic benchmark characteristics of such an assay is the capability of allowing a high-throughput of clones while at the same time guaranteeing sufficient sensitivity and a low background-to-signal ratio (Ferrer et al., 2009; Steele et al., 2008; Taupp et al., 2011; Tuffin et al., 2009). The method of choice for enzyme activity screenings are agar plate assays supplemented with substrates which support the discovery of a desired enzymatic activity, as they offer the screening of a large number of clones. The signal can either be a clearing zone around the colony or a color reaction indicating enzymatic activity in the vicinity of or within a specific clone (Uchiyama and Miyazaki, 2009). The sensitivity and the ratio of background-to-signal intensity are strong contributing factors for the low hit rates typically observed in library screenings. Successfully produced enzymatic activity of a respective clone present in the library has to be above the activity of any background noise arising from the cell itself or from biophysical phenomena causing naturally occurring substrate degradation. Additionally, the time the assay allows to assess the read-out is crucial. The window of opportunity to differentiate the signal from the background activity needs to be wide enough to allow detection of clones with a comparably low enzymatic activity (Robertson and Steer, 2004; Steele et al., 2008).

However, these challenges can be alleviated by instead linking enzymatic activity with the survivability of the host organisms to the assay conditions (Mori et al., 2018; Simon et al., 2009). Survivability can either be linked to the degradation of toxic substance as in the case of antibiotic selection or to the metabolization of a substrate necessary for growth and survival. This could either be the degradation of a complex polymer releasing carbon substrates for growth or supply of nitrogen through the action of an enzyme with desired function. Assay set-ups such as this can be a powerful tool for the discovery of new enzymes but are limited to the possibility of linking desired function with degradation of available substrates and therefor might not apply to all enzymes. The degradation of cellulosic material for instance can be used for the screening of cellulases and hemi-cellulases. The presence of a highly active enzyme in such a library approach can act upon the polymer and release tri-, di- and monosaccharides used as growth substrates by the cell (Kickenweiz et al., 2018). The natural diffusion of both enzymes and released substrates occurring in agar plates limits the time it is possible to use the assay, as over time also neighboring cells will profit from the absence of a toxic compound or the presence of a growth limiting carbon or nitrogen substance. An example of the degradation of a toxic compound was presented by Bayer and co-workers who showed that they could successfully discover methyl-halide transferases by expressing and screening *E. coli* and *S. cerevisiae* cells (Bayer et al., 2009).

Enzyme identification via proteomics and functional proteomics

A complementary approach to discover new enzymes is proteomics. Several different angles are in use, however the most commonly employed one is bottom-up shotgun proteomics. A protein sample is enzymatically digested in order to break it down into peptides which are then separated and analyzed. The separation and analysis can either be done by a tandem mass spectrometry liquid chromatography system (ESI–LC–MS/MS) or by matrix-assisted laser desorption ionization and time of flight mass spectrometry (MALDI–TOF–MS). Based on the mass to charge ratios (m/z) of peptides and fragments thereof, proteins are compared to the in silico digested (m/z) values of theoretical proteins/peptides. Subsequently, search programs such as COMET (Eng et al., 2013), MaxQuant (Cox and Mann, 2008), MASCOT (Pappin et al., 1999), MSAmanda (Dorfer et al., 2014), X!Tandem (Craig and Beavis, 2004), MyriMatch (Tabb et al., 2008), SEQUEST (Eng et al., 1994), MS–GF+ (Kim and Pevzner, 2014) or OMSSA (Geer et al., 2004) are employed for peptide mapping.

These programs attempt to match at least one unique peptide to the protein found in the database. The underlying assumption however is that the protein is actually present in the database. De-novo protein sequencing is a possibility to identify proteins not present in any databases (Hughes et al., 2010). Although the possibility exists, it requires large amounts of pure protein and considerable effort in time as well as in resources. Additionally, metaproteomic and proteomic analyses have been conducted on environmental and cultured organisms in order to identify novel enzymes or more complex enzymatic functions requiring a pathway (Kirsch et al., 2012; Schneider et al., 2012; Weiß et al., 2013; Wilmes et al., 2008).

The synergy of proteomics and mRNA sequencing

While transcriptomics and proteomics as single -omics technologies can be used as powerful tools for enzyme discovery, the synergy of multiple -omics approaches can be even more potent to exploit nature's biodiversity. Several studies demonstrating the successful synergy of transcriptomics and proteomics have been shown recently. Employing transcriptomics derived sequences to construct protein databases allowed a more in-depth analysis resulting in the discovery of novel enzymes and entire pathways. In these cases, the availability of specifically generated protein databases drastically improved the likelihood of finding novel enzymes (Desgagné-Penix et al., 2012; Kirsch et al., 2012; Schilmiller et al., 2010; Schneider et al., 2012; Van Bogaert et al., 2013; Wilmes et al., 2008). Compared to other studies using only one –omics technology where an identification was not possible due to a lack of samples specific protein databases (Amore et al., 2015; Benson et al., 2013; Jacobs et al., 2005a; Kirsch et al., 2012; Tiwari et al., 2014).

The identification of target enzymes on the protein level allows to circumvent DNA and mRNA and focuses on the enzymatic activity measured in a sample directly without any reference to the issues of assembly, incomplete transcripts or the inability to express certain genes. On the other hand, the analysis is limited to a fixed set of proteins being actively expressed at the time of analysis. Therefore, it is limited to a subset of proteins in comparison to all potential ones coded in the genome. Additionally, the amounts and the specific activity of said enzymes plays a major role in the discovery process of protein-based identification.

Localizing specific enzymatic activities via functional proteomics

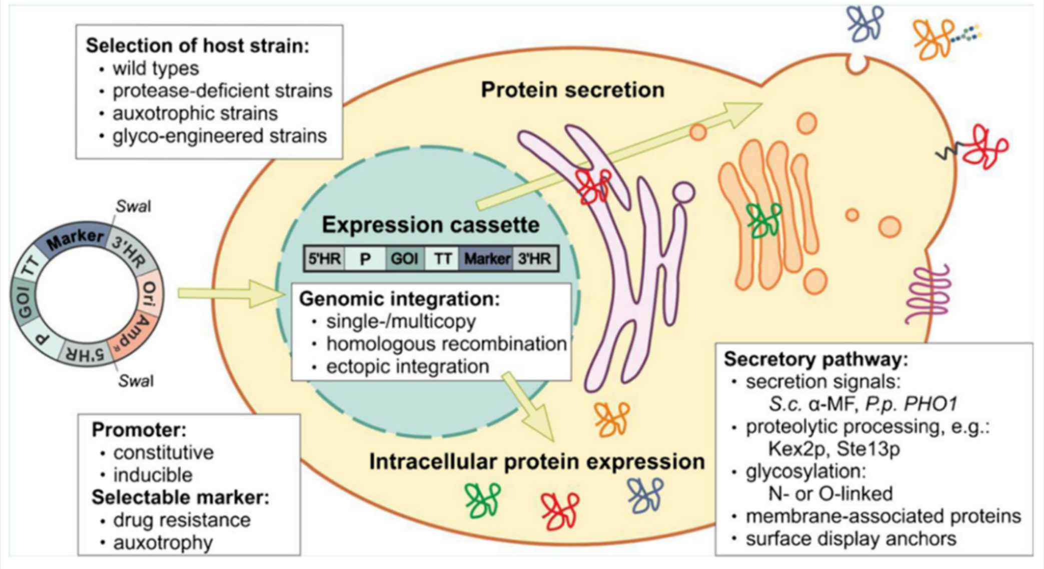
Adding even more layers of information to the analysis is the possibility to perform functional proteomics experiments. This technique relies on the linking of sequence and function on the protein level through the use of activity-specific probes. The probes are linked to a tag which can subsequently be used either for detection or purification of the desired proteins. The proteomics probe can be tailored to suit the substrate requirements of the target enzymes (Cravatt, 2014; Cravatt et al., 2008; Okerberg et al., 2005; Pohl, 2005). A broad array of probes exist covering diverse substrate classes and enzyme functions (Fonović and Bogyo, 2008; Salisbury and Cravatt, 2007). The most abundant enzyme class for which proteomic probes are available are hydrolases (Simon and Cravatt, 2010). However, also other enzyme classes are covered by probes and theoretically the substrate design for desired enzyme classes is one of the few application limits. The identification and retrieval of enzymes is simplified by a covalent bond generated between the enzyme and the probe. The probe can be designed in a way to only bind to enzymes with the correct three-dimensional architecture within the active center. The reacted enzymes can then be separated from the complex sample and further processed (Böttcher and Sieber, 2008). This exemplifies the strength of the functional proteomics approach, combining an -omics approach with enzymatic function determination thereby combining pieces of information which are usually extremely difficult to discover in one experiment – the relationship between the presence of an enzyme and its function.

Komagataella phaffii (*Pichia pastoris*) as an expression host for enzyme discovery

Yeast species are widely used in the production of recombinant proteins both on an industrial and laboratory scale (Buckholz, Richard; Gleeson, 1991; Cereghino and Cregg, 1999; Gellissen and Hollenberg, 1997; Porro et al., 2005). As eukaryotic organisms, they offer many advantages such as the capability of post-translational modifications (e.g. glycosylation, disulfide bridge formation) (Hamilton et al., 2003; Hamilton and Gerngross, 2007; Xiao et al., 2004) and secretory expression (Leonardo M Damasceno et al., 2012; Idiris et al., 2010a), while at the same time combining the ease in genetic manipulation (Cregg et al., 2000a), the growth on inexpensive synthetic media (Cereghino et al., 2002) as well as bioprocess relevant parameters such as high cell density fermentation (Heyland et al., 2010) typically associated with prokaryotic host expression systems. Yeast species employed for protein expression can roughly be categorized in non-methylotrophic and methylotrophic species, with the latter group playing a crucial role in industrial scale heterologous protein expression (Porro et al., 2005). In the group of non-methylotrophic yeast species are *Saccharomyces cerevisiae* (Ghaemmaghami et al., 2003), *Kluyveromyces lactis* (Fukuhara, 2006; Ooyen et al., 2006), *Yarrowia lipolytica* (Bordes et al., 2007; Madzak et al., 2004), *Pichia stiptidis* (Piotek et al., 1998a), and *Zygosaccharomyces bailii* (Branduardi et al., 2004). The species *Hansenula polymorpha* (Celik and Calık, 2012; Steinborn et al., 2006), *Pichia pastoris* (Cereghino et al., 2002) (Cereghino and Cregg, 2000a; Cregg et al., 2000b), *Candida boidinii* (Yurimoto, 2009) and *Pichia methanolica* (Yurimoto and Sakai, 2009) are the preferred host organisms for heterologous protein production using methanol for protein production purposes. Due to its prevalent position in industrial as well as laboratory scale expression the most prominent position among industrial yeasts has been taken by the methylotrophic yeast species *Pichia pastoris* (*Komagataella phaffii*).

The yeast species *Komagataella phaffii*, formerly known as *Pichia pastoris*, is one of the most widely used yeast species for protein expression both in academic as well as in industrial settings. Microbial isolates capable to utilize the single carbon alcohol methanol from the Yosemite National Park (California) were discovered and developed into the expression system we know today. *Komagataella phaffii* was chosen by Philips Petroleum for the methanol-based large-scale single cell protein production. The cheap and readily available petrochemical product methanol was used to grow these cells to densities up to 130 g/L (cell dry weight). However, the oil crisis starting in 1973 drastically increased the price for methane, from which methanol was derived, and made other sources of protein such as soybeans economically more competitive. The allure of this system was its capability to grow on simple defined media and multi cubic meter scale bioreactors to very high cell densities. In the 1980s, Phillips Petroleum Company together with the Salk Institute founded the Salk Institute Biotechnology/Industrial Associates Inc. (SIBIA, La Jolla, CA, USA) and further developed this system for protein expression (Cereghino and Cregg, 2000b; Cregg et al., 2000a). The species *Pichia pastoris*, which in 2009 was reclassified as *Komagataella phaffii,* includes the strains NRRL Y-11430 from the Agriculture Research Service culture collection (Peoria IL, USA), and NRRL Y-48124 (X-33, Invitrogen expression kit strains, Carlsbad CA, USA) (Kurtzman, 2009, 2005). Later the strain NRRL Y-11430 was also deposited in Utrecht (Netherlands) as CBS7435. Although the genome sequence of the type strain NRRL Y-7556 is not known, several other strains' genomes have been sequenced and published so far. The majority of employed *K. phaffii* are derived from the CBS7435 strain including the strain GS115 which was chemically mutagenized and selected for histidine auxotrophy and is the most widely used *K. phaffii* strain (De Schutter et al., 2009; Küberl et al., 2011; Sturmberger et al., 2016a). Over the last three decades this yeast species has developed into an important industrial expression system but also gained importance in basic research as a model eukaryote for protein secretion, peroxisome biogenesis and autophagic degradation. Recently, several attempts were made to improve on existing genome annotations (Love et al., 2016; Valli et al., 2016) and provide new reference genome sequence for *K. phaffii* CBS7435 (Love et al., 2016; Sturmberger et al., 2016a).

There exist many different cellular processes influencing the expression of heterologous proteins in *K. phaffii*. Vogl and coworkers, as well as Ahmad and colleagues attempted to give a short summary of the different steps involved in the production and/or secretion of recombinant proteins (Ahmad et al., 2014; Vogl et al., 2013a; Vogl and Glieder, 2013). These aspects include the selection of host strains (i.e. protease deficient strains, auxotrophic strains or glycoengineered strains), the promoter used for transcription of the recombinant gene (constitutive, de-repressible, inducible), the mode in which the expression cassette is supplied (genomic integration vs. episomal expression) and the possibility of secreting recombinant proteins (post-translational or co-translational protein translocation) (Ahmad et al., 2014).



Ahmad, M., Hirz, M., & Pichler, H. (2014). Protein expression in Pichia pastoris : recent achievements and perspectives for heterologous protein production, 5301–5317. https://doi.org/10.1007/s00253-014-5732-5

Figure 1: Cellular processes influencing expression and secretion of recombinant proteins. The cellular processes include amongst other factors, the selection of host strains, the choice of promoter, the mode of supply of genetic information and the choice of secretion signal for secretory expression.

Promoters and their regulatory profile employed for protein expression

Most likely due to the low affinity of the *AOX1* protein towards oxygen, the cells require large amounts of this enzyme. In cells fed with methanol the *AOX1* protein can sometimes make up 30% of soluble protein (Cereghino and Cregg, 2000b). The transcription and translation of the genes necessary for methanol metabolization are tightly regulated and repressed under the presence of glucose (Hartner and Glieder, 2006a). Through the addition of methanol, a substantial increase of *AOX1* transcript and protein levels can be observed. This regulatory profile makes the promoter a perfect candidate for recombinant protein expression (Vogl and Glieder, 2013). However, by now numerous other methanol inducible and constitutive promoters have been identified, such as the promoter of the *GAP* gene (Vassileva et al., 2001) commonly used for continuous protein expression or the promoter of the *DAS* gene (Tschopp et al., 1987) which shows an even higher induction profile compared to *AOX1* (Vogl et al., 2015). Recently, the application of de-repressible promoters has gained attention, allowing induced expression without the need for methanol induction (Fischer et al., 2019; Hartner et al., 2008; Hartner and Glieder, 2006a; Prielhofer et al., 2013; Vogl et al., 2015). These promoters are typically repressed on the carbon source used in the batch phase (e.g. glucose or glycerol) and after the carbon source is depleted, transcription and translation commence. One example of such a system was characterized by Vogl et al. using the promoter of the catalase 1 gene (Vogl et al., 2015) as well as the promoter of the formate dehydrogenase gene from *Ogataea angusta* (Vogl et al., 2020) in *K. phaffii*. Under carbon depleted as well as methanol induced conditions the promoter of the catalase 1 gene was superior compared to the standard P_{AOX1}. A similar strategy was pursued by Prielhofer and colleagues in developing a promoter system which is repressed under glycerol batch conditions and is induced by a shift to a glucose feed (Gasser et al., 2013; Prielhofer et al., 2018, 2015). The main advantage of inductive de-repressed systems lies in the decoupling of cell growth and protein production. This allows for the production of proteins otherwise interfering in the cellular metabolism in a negative manner, while still allowing a marked increase in transcript levels at a certain time-point (Vogl et al., 2016).

Different secretion signals and the respective pathway used for translocation of recombinant proteins

A further hallmark of *K. phaffii* is the possibility to secrete proteins into the medium without laborious and difficult cell disruption. This mode of expression has two advantages. Firstly, it simplifies down-stream processing since the number and amount of proteins contained in a typical *K. phaffii* secretome is drastically lower compared to a lysate prepared by cell disruption. Due to its very low complexity often the desired protein to be secreted is the primary protein found within the sample (Mattanovich et al., 2009). Secondly, the entire process of cell lysis is rendered redundant, which eliminates certain unit operations while saving both time and cost. However, the secretion of proteins requires the presence of a secretion signal which mediates translocation into the ER and the secretory pathway (Kurjan and Herskowitz, 1982a). Concerning the translocation itself, one can distinguish between post-translational and a co-translational translocation, each showing a different cellular mechanism (Leonardo M Damasceno et al., 2012). The most widely used signal sequence in *K. phaffii* is the *S. cerevisiae* derived pre-pro signal peptide from the alpha-factor mating protein. This signal sequence consists of a 19-amino acid signal (pre) sequence followed by a 66-residue (pro) sequence containing three consensus N-linked glycosylation sites and a dibasic Kex2 endopeptidase processing site. The first step in processing is the cleavage of the pre-signal by ER residing signal peptidases. Secondly, the endopeptidase Kex2 is hydrolyzing the peptide chain between the amino acid residues Arg and Lys of the pro-sequence, followed by cleavage in a Glu-Ala-Glu-Ala repeat by the carboxypeptidase Ste13. Proteins fused with this signal sequence are kept in an unfolded state prior to translocation through the ER membrane.

Within the ER the protein is finally folded with the help of foldases such as Kar2 and BiP (Kurjan and Herskowitz, 1982b). However, there also exists a different mode of protein secretion termed co-translational protein translocation, where the nascent protein chain is translocated through the ER membrane while simultaneously folding of the protein occurs (Fitzgerald and Glick, 2014). Depending on the biophysical and structural properties of the recombinant protein the choice of post- or co-translational translocation can influence the secretion efficiency accordingly. Recently, Barrero and co-workers were able to show an improved secretion of the challenging E2-Crimson red fluorescent protein by employing the *S. cerevisiae* OST1 pre and *K. phaffii* alpha factor pro region (Barrero et al., 2018). These results indicate the importance to investigate the type of secretion of an unknown protein rather than relying simply on one specific secretion signal.

The effect of host strains on the quality and quantity of recombinant proteins.

The most commonly used *K. phaffii* strain is the NRRL-Y 11430 (Northern Regional Research Laboratories, Peoria, IL) (Cregg et al., 2000a; Küberl et al., 2011). Certain strains harbor one or more mutations incorporated either through genetic manipulation or chemical/physical mutation creating auxotrophies for appropriate selectable markers. The outcome of such undertakings was for example the *K. phaffii* GS115 strain widely used in industrial protein expression (De Schutter et al., 2009). Other mutations can concern the *AOX1* gene severely slowing the pace of methanol metabolization and effectively generating a mutS phenotype (Krainer et al., 2012). To eliminate the occurrence of protease-mediated degradation of heterologous expressed proteins, protease deficient strains were constructed. The genes *his4*, *pep4*, *prb1* were knocked out and shown to improve heterologous protein titers, above all in fermenter cultures (Cereghino and Cregg, 2000b). In a different study, Krainer and colleagues showed that knocking out the *OCH1* gene in *K. phaffii* improves the homogeneity of recombinant horseradish peroxidase (HRP) expression (Florian W. Krainer et al., 2013).

Recently, Tillman Gerngross, Stephen Hamilton, Nico Callewaert and colleagues have generated *K. phaffii* strains capable of attaching human-like glycan structures (Hamilton et al., 2003; Hamilton and Gerngross, 2007; Jacobs et al., 2009). By knocking out mannosyltransferases (such as *OCH1*) residing in the ER or Golgi (Bobrowicz et al., 2004; Hamilton et al., 2003), supplying the cell with mannosidases and testing different oligosaccharyl-transferases (Choi et al., 2012) they deleted the hypermannosylation usually observed in *K. phaffii* glycoproteins (Jacobs and Callewaert, 2009) and replaced it with human-like glycan structures. Proteins produced in these strains are therefore suitable for use as human pharmaceuticals since the potentially allergenic properties of hypermannose structures are abolished.

Other than changing the glycosylation machinery, several other processes involved in protein expression can be optimized. The production and secretion of recombinant proteins exerts metabolic stress on cellular systems. The folding and post-translational modification machinery, to a large part situated in the ER and Golgi Body, induce the unfolded protein response (UPR) (Graf et al., 2008). The UPR induces the expression of genes related to protein folding but also ER associated degradation (ERAD), to alleviate the cell from ER related stress and potential blockage within the secretory pathway (Vanz et al., 2014; Zahrl et al., 2018). In a typical *K. phaffii* fermentation, high methanol concentrations and high osmolarity can lead to apoptosis phenomena and therefore loss of protein expressing cells. Furthermore, osmotic stress due to high osmolarity in fermentation media might also induce cell lysis and diffusion of extracellular proteases into the culture supernatants. This would further increase proteolytic activity and endanger the quantity and quality of recombinant proteins (Mattanovich et al., 2004). Furthermore, the medium pH and composition can influence the activity of proteases (Potvin et al., 2012). It is therefore essential to strike a balance between cellular and recombinant protein requirements for a certain pH and the optimization of pH to lower protease activity. Additionally, the use of complex media containing protein hydrolysates can act as a competing substrate and also decrease proteolytic degradation of the recombinant proteins (Mattanovich et al., 2004).

The incorporation of genetic material – genomic vs. episomal expression of recombinant proteins

The expression of heterologous genes in any organism requires the incorporation of foreign genetic material into the host cell. In principal, for *K. phaffii* there exist two different modes of gene presentation. On the one hand, foreign genetic material can be integrated into the genome either via homologous recombination or non-homologous-end-joining (Näätsaari et al., 2012a). For gene integration into the genome, a linear DNA fragment containing homologous sequences to the desired integration locus at both ends of the construct, is transformed into the host cell. Following the transformation, the cells will either use the homologous sequences present on the DNA fragment to direct it to the desired locus where it will be integrated into the genome (Näätsaari et al., 2012b). Alternatively, the cell can use the non-homologous-end-joining mechanism to achieve a quasi-random integration into the genome. However, during the process of integration, genes can be deleted wherever the linear DNA fragment is inserted. In case of a targeted integration this can be used to screen for a desired phenotype, which was shown to be very successful for integrating into the *AOX1* locus (Näätsaari et al., 2012a). By replica stamping transformation plates onto minimal media plates supplemented with methanol, mut- strains lacking the capability to metabolize methanol and therefore growth on this carbon source can be generated. Identifying the locus of randomly integrated constructs necessitates either the sequencing of a chosen transformant or other methods such as genome walking (Leoni et al., 2011; Tang et al., 2006). Compared to other yeast species such as *S. cerevisiae*, *K. phaffii* shows a high propensity for non-homologous-end-joining in contrast to homologous recombination (Miné-Hattab and Rothstein, 2013; Schwarzhans et al., 2016). This feature requires the screening of several transformants since the expression levels of a desired gene is highly dependent on the genetic locus amongst other factors. A possibility to circumvent this costly and time-intensive initial screening is the utilization of a knock-out strain deficient in non-homologous-end-joining, termed deltaKU70 (Näätsaari et al., 2012b). This strain lacks the key player of the NHEJ machinery, KU70, and shows a drastically reduced tendency for random integration. However, due to the abolished NHEJ the transformation efficiencies of these strains are lower than for wildtype cells.

The discovery and developments of the CRISPR-Cas9 technology opens up the possibility of targeted genome engineering in a whole host of different species and cell lines (Deltcheva et al., 2011; Jinek et al., 2012; Sapranauskas et al., 2011). Recently, this technology was developed for the yeast species *K. phaffii* (Gassler et al., 2019; Liu et al., 2019; Weninger et al., 2018, 2016). The generation of targeted knock-outs in *K. phaffii* is hampered by a tendency for non-homologous end joining over homologous recombination (Näätsaari et al., 2012a). CRISPR-Cas9 offers a possibility to facilitate genome engineering in *K. phaffii*. The efficiencies achieved during the integration of marker-less cassettes were close to 100% in the background of a *ku70* KO strain (Weninger et al., 2016), and about 25-fold improved in the wildtype strain by placing an autonomously replicating sequence on the donor cassette. Additionally, Weninger and colleagues were able to show that the generation of expression plasmids even allows the modification of existing production strains (Weninger et al., 2018). While any attempts to reduce off-target toxicity or improve donor cassette integration were not successful, they nevertheless could demonstrate the extreme value of this technology for facilitating homologous recombination and targeted genome engineering in *K. phaffii*.

While genomic integration is by far the most common method of foreign gene introduction, the availability of plasmid systems also allows episomal expression of heterologous proteins (patent application WO2017055436A1). Recently, several research groups have shown the possibility to use autonomously replicating sequence (ARS) containing plasmids for protein expression in *K. phaffii* (Camattari et al., 2016; Schwarzhans et al., 2017; Vogl et al., 2018a; Weninger et al., 2018). This latter method offers transformation efficiencies of up to 10^6 far exceeding the values attainable via genomic integration of linear constructs (WO2017055436A1). Additionally, the need to supply a linear fragment either generated via PCR or by linearization of a circular plasmid is also unnecessary. The combination of these two factors drastically simplifies the screening of gene libraries for protein engineering and enzyme discovery projects. However, drawbacks of using episomal plasmids for protein expression lie in the fact that a steady supply of selection pressure is necessary for plasmid retention. If the selection pressure is abolished and cell growth continues, the supplied plasmids will get lost in the replication process and daughter cells will therefore not contain any plasmid for heterologous gene expression (Clyne and Kelly, 1995; Liachko and Dunham, 2014).

An additional feature for fine-tuning the protein expression in *K. phaffii* is the gene dosage, i.e. the copy number of genes present at a given moment. Generally speaking, increasing the copy number of an expression cassette leads to an increase in the amount of expressed protein. However, increasing the gene copy number does not always benefit the protein titers as a knock-on effect of transcription and translation can overburden the cellular resources available. Especially the post-translational processes of folding, translocation and signal sequence processing can lead to the creation of bottlenecks during the production of recombinant proteins (Aw and Polizzi, 2013).

High-Cell-Density Fermentations and factors influencing protein yield and quality

The propensity of *K. phaffii* as a Crabtree negative yeast makes it more suitable for fermenter cultivations compared to *S. cerevisiae*, which accumulates high amounts of ethanol in high cell density cultivations (Baumann et al., 2011). The amount of ethanol can quickly reach toxic levels limiting growth as well as heterologous protein production. Without this preference for respiratory growth, *K. phaffii* can reach extremely high densities of up to 500 OD units. Combined with the possibility of secreting foreign proteins (secreted protein concentration is roughly proportional to cell concentration) this makes a compelling argument in favor of high-cell-density fermentations (Cregg et al., 2000a; Heyland et al., 2010). As the P_{AOX1} quickly became the promoter of choice for expression in *K. phaffii*, traditionally the bioprocesses involved some form of methanol addition for promoter induction (Cereghino and Cregg, 2000b). A typical high-cell-density fermentation of such a process starts with a batch phase on either glucose or glycerol and a transition phase at which the carbon source is fed in growth limiting amounts. Methanol induction is commenced upon depletion of the carbon source. In some cases, a mixture of glycerol and methanol can be beneficial for recombinant protein expression (Jahic et al., 2006; Zhang et al., 2003). The promoter engineering efforts of Hartner and colleagues however, proved that the P_{AOX1} could be engineered to allow de-repressed expression. These developments on the one hand allowed to separate the batch and production phase and induce gene expression at a specific time point and on the other hand eliminated the necessity for the addition of the highly volatile, toxic and flammable compound methanol (Franz and Hartner, 2007; Hartner and Glieder, 2006b). Several research works within the last years have used these engineered promoters and combined them with a limited glycerol feed to achieve promoter induction (Looser et al., 2017, 2015; Ruth et al., 2010). In a different approach, Prielhofer et al. combined promoters which are repressed on glycerol and show marked upregulation upon glucose addition for recombinant protein expression (Prielhofer et al., 2013). More recently, Fischer and colleagues were able to show markedly improved hGH secretion when employing the PDC under a limited glycerol feed (implemented via glycerol release feed disks) (Fischer et al., 2019). Alternatively, the transcriptional response of methanol addition can be mimicked by overexpression of key methanol utilization gene regulators (Lin-cereghino et al., 2006; WO2008/090211). This strategy was successfully implemented for the recombinant expression of industrially relevant enzymes in *K. phaffii* without the need for methanol induction (Shen et al., 2016b, 2016a; Takagi et al., 2009; Vogl et al., 2018b; Wang et al., 2017).

References

Adrio, J.L., Demain, A.L., 2014. Microbial enzymes: tools for biotechnological processes. Biomolecules 4, 117–139. https://doi.org/10.3390/biom4010117

Ahmad, M., Hirz, M., Pichler, H., 2014. Protein expression in Pichia pastoris : recent achievements and perspectives for heterologous protein production 5301–5317. https://doi.org/10.1007/s00253-014-5732-5

Almagro Armenteros, J.J., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., Nielsen, H., 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat. Biotechnol. 37, 420–423. https://doi.org/10.1038/s41587-019-0036-z

Amann, R.I., Ludwig, W., Schleifer, K.H., 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. 59, 143–169. https://doi.org/10.1016/j.jip.2007.09.009

Amore, A., Parameswaran, B., Kumar, R., Birolo, L., Vinciguerra, R., Marcolongo, L., Ionata, E., La Cara, F., Pandey, A., Faraco, V., 2015. Application of a new xylanase activity from Bacillus amyloliquefaciens XR44A in brewer's spent grain saccharification. J. Chem. Technol. Biotechnol. 90, 573–581. https://doi.org/10.1002/jctb.4589

Anastas, P.T., Warner, J.C., 1998. Green Chemistry: Theory and Practice. Oxford University Press.

Antonucci, T.K., Wen, P., Rutter, W.J., 1989. Eukaryotic promoters drive gene expression in Escherichia coli. J. Biol. Chem. 264, 17656–9.

Audagnotto, M., Dal Peraro, M., 2017. Protein post-translational modifications: In silico prediction tools and molecular modeling. Comput. Struct. Biotechnol. J. 15, 307–319. https://doi.org/10.1016/J.CSBJ.2017.03.004

Aw, R., Polizzi, K.M., 2013. Can too many copies spoil the broth ? 1–9.

Banerjee, U.C., Sani, R.K., Azmi, W., Soni, R., 1999. Thermostable alkaline protease from Bacillus brevis and its characterization as a laundry detergent additive. Process Biochem. 35, 213–219. https://doi.org/10.1016/S0032-9592(99)00053-9

Baumann, K., Dato, L., Graf, A.B., Frascotti, G., Dragosits, M., Porro, D., Mattanovich, D., Ferrer, P., Branduardi, P., 2011. The impact of oxygen on the transcriptome of recombinant S. cerevisiae and P. pastoris - a comparative analysis. BMC Genomics 12, 218. https://doi.org/10.1186/1471-2164-12-218

Bayer, T.S., Widmaier, D.M., Temme, K., Mirsky, E.A., Santi, D. V, Voigt, C.A., 2009. Synthesis of Methyl Halides from Biomass Using Engineered Microbes. J. Am. Chem. Soc. 131, 6508–6515. https://doi.org/10.1021/ja809461u

Behrens, G. a., Hummel, A., Padhi, S.K., Schätzle, S., Bornscheuer, U.T., 2011. Discovery and Protein Engineering of Biocatalysts for Organic Synthesis. Adv. Synth. Catal. 353, 2191–2215. https://doi.org/10.1002/adsc.201100446

Benson, D. a., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2013. GenBank. Nucleic Acids Res. 41, D36–D42. https://doi.org/10.1093/nar/gks1195

Bhosale, S.H., Rao, M.B., Deshpande, V. V, 1996. Molecular and industrial aspects of glucose isomerase. Microbiol. Rev. 60, 280–300.

Bobrowicz, P., Davidson, R.C., Li, H., Potgieter, T.I., Nett, J.H., Hamilton, S.R., Stadheim, T. a, Miele, R.G., Bobrowicz, B., Mitchell, T., Rausch, S., Renfer, E., Wildt, S., 2004. 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, 757–66. https://doi.org/10.1093/glycob/cwh104

Bogdanova, E. a, Shagin, D. a, Lukyanov, S. a, 2008. Normalization of full-length enriched cDNA. Mol. Biosyst. 4, 205–12. https://doi.org/10.1039/b715110c

Bordes, F., Fudalej, F., Dossat, V., Nicaud, J., Marty, A., 2007. A new recombinant protein expression system for high-throughput screening in the yeast Yarrowia lipolytica 70, 493–502. https://doi.org/10.1016/j.mimet.2007.06.008

Bornscheuer, U.T., Huisman, G.W., Kazlauskas, R.J., Lutz, S., Moore, J.C., Robins, K., 2012. Engineering the third wave of biocatalysis. Nature 485, 185–194. https://doi.org/10.1038/nature11117

Böttcher, T., Sieber, S.A., 2008. β-Lactones as Privileged Structures for the Active-Site Labeling of Versatile Bacterial Enzyme Classes. Angew. Chem. Int. Ed. 47, 4600–4603. https://doi.org/10.1002/anie.200705768

Branduardi, P., Valli, M., Brambilla, L., Sauer, M., Alberghina, L., Porro, D., 2004. The yeast Zygosaccharomyces bailii: A new host for heterologous protein production, secretion and for metabolic engineering applications. FEMS Yeast Res. 4, 493–504. https://doi.org/10.1016/S1567-1356(03)00200-9

Bräutigam, A., Shrestha, R.P., Whitten, D., Wilkerson, C.G., Carr, K.M., Froehlich, J.E., Weber, A.P.M., 2008. Comparison of the use of a species-specific database generated by pyrosequencing with databases from related species for proteome analysis of pea chloroplast envelopes. J Biotechnol 136, 44–53.

Buckholz, Richard; Gleeson, M., 1991. Nature Publishing Group http://www.nature.com/naturebiotechnology. Nat. Biotechnol. 9, 1067–1072.

Camattari, A., Goh, A., Yip, L.Y., Hee, A., Tan, M., Ng, S.W., Tran, A., Liu, G., Liachko, I., Dunham, M.J., Rancati, G., 2016. Characterization of a panARS - based episomal vector in the methylotrophic yeast Pichia pastoris for recombinant protein production and synthetic biology applications. Microb. Cell Factories 1–11. https://doi.org/10.1186/s12934-016-0540-5

Cammarota, M.C., Freire, D.M.G., 2006. A review on hydrolytic enzymes in the treatment of wastewater with high oil and grease content. Bioresour. Technol. 97, 2195–2210. https://doi.org/10.1016/j.biortech.2006.02.030

Catalysis, N., 2018. On advances and challenges in biocatalysis 1, 635–636. https://doi.org/10.1038/s41929-018-0157-7

Celik, E., Calık, P., 2012. Production of recombinant proteins by yeast cells. Biotechnol. Adv. 30, 1108–18. https://doi.org/10.1016/j.biotechadv.2011.09.011

Cereghino, G.P.L., Cereghino, J.L., Ilgen, C., Cregg, J.M., 2002. Production of recombinant proteins in fermenter cultures of the yeast Pichia pastoris 329–332. https://doi.org/10.1016/S0958166902003300

Cereghino, G.P.L., Cregg, J.M., 1999. Applications of yeast in biotechnology: Protein production and genetic analysis. Curr. Opin. Biotechnol. 10, 422–427. https://doi.org/10.1016/S0958-1669(99)00004-X

Cereghino, J.L., Cregg, J.M., 2000a. Heterologous protein expression in the methylotrophic yeast Pichia pastoris 24.

Cereghino, J.L., Cregg, J.M., 2000b. Heterologous protein expression in the methylotrophic yeast Pichia pastoris 24.

Choi, B., Warburton, S., Lin, H., 2012. Improvement of N -glycan site occupancy of therapeutic glycoproteins produced in Pichia pastoris 671–682. https://doi.org/10.1007/s00253-012-4067-3

Clouthier, C.M., Pelletier, J.N., 2012. Expanding the organic toolbox: A guide to integrating biocatalysis in synthesis. Chem. Soc. Rev. 41, 1585–1605. https://doi.org/10.1039/c2cs15286j

Clyne, R.K., Kelly, T.J., 1995. Genetic analysis of an ARS element from the fission yeast Schizosaccharomyces pombe. EMBO J. https://doi.org/10.1002/j.1460-2075.1995.tb00326.x

Consortium, T.U., 2015. UniProt: a hub for protein information. Nucleic Acids Res. 43, D204–D212. https://doi.org/10.1093/nar/gku989

Cox, J., Mann, M., 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372. https://doi.org/10.1038/nbt.1511

Craig, R., Beavis, R.C., 2004. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20, 1466–1467. https://doi.org/10.1093/bioinformatics/bth092

Cravatt, B., 2014. Activity-based proteomics &#151; applications for enzyme and inhibitor discovery (357.1). FASEB J. 28, 357.1. https://doi.org/10.1096/fasebj.28.1_supplement.357.1

Cravatt, B.F., Wright, A.T., Kozarich, J.W., 2008. Activity-Based Protein Profiling: From Enzyme Chemistry to Proteomic Chemistry. Annu. Rev. Biochem. 77, 383–414. https://doi.org/10.1146/annurev.biochem.75.101304.124125

Cregg, J.M., Cereghino, J.L., Shi, J., Higgins, D.R., 2000a. Recombinant Protein Expression in Pichia pastoris 16.

Damasceno, L.M., Huang, C., Batt, C.A., 2012. Protein secretion in Pichia pastoris and advances in protein production 31–39. https://doi.org/10.1007/s00253-011-3654-z

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., Steinmetz, L.M., 2006. A high-resolution map of transcription in the yeast genome. Proc. Natl. Acad. Sci. U. S. A. 103, 5320–5. https://doi.org/10.1073/pnas.0601091103

Davids, T., Schmidt, M., Böttcher, D., Bornscheuer, U.T., 2013. Strategies for the discovery and engineering of enzymes for biocatalysis. Curr. Opin. Chem. Biol. 17, 215–220. https://doi.org/10.1016/j.cbpa.2013.02.022

De Schutter, K., Lin, Y.-C., Tiels, P., Van Hecke, A., Glinka, S., Weber-Lehmann, J., Rouzé, P., Van de Peer, Y., Callewaert, N., 2009. Genome sequence of the recombinant protein production host Pichia pastoris. Nat. Biotechnol. 27, 561–566. https://doi.org/10.1038/nbt.1544

Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J., Charpentier, E., 2011. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature 471, 602–607. https://doi.org/10.1038/nature09886

Desgagné-Penix, I., Farrow, S.C., Cram, D., Nowak, J., Facchini, P.J., 2012. Integration of deep transcript and targeted metabolite profiles for eight cultivars of opium poppy. Plant Mol. Biol. 79, 295–313. https://doi.org/10.1007/s11103-012-9913-2

Deutsch, M., Long, M., 1999. Intron – exon structures of eukaryotic model organisms 27, 3219–3228.

Dorfer, V., Pichler, P., Stranzl, T., Stadlmann, J., Taus, T., Winkler, S., Mechtler, K., 2014. MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. J. Proteome Res. 13, 3679–84. https://doi.org/10.1021/pr500202e

Drinnenberg, I.A., Weinberg, D.E., Xie, K.T., Mower, J.P., Wolfe, K.H., Fink, G.R., Bartel, D.P., 2009. RNAi in budding yeast. Science 326, 544–550. https://doi.org/10.1126/science.1176945.RNAi

Eberl, a., Heumann, S., Kotek, R., Kaufmann, F., Mitsche, S., Cavaco-Paulo, a., Gübitz, G.M., 2008. Enzymatic hydrolysis of PTT polymers and oligomers. J. Biotechnol. 135, 45–51. https://doi.org/10.1016/j.jbiotec.2008.02.015

Ekkers, D.M., Cretoiu, M.S., Kielak, A.M., van Elsas, J.D., 2012. The great screen anomaly—a new frontier in product discovery through functional metagenomics. Appl. Microbiol. Biotechnol. 93, 1005–1020. https://doi.org/10.1007/s00253-011-3804-3

Eng, J.K., Jahan, T.A., Hoopmann, M.R., 2013. Comet: An open-source MS/MS sequence database search tool. PROTEOMICS 13, 22–24. https://doi.org/10.1002/pmic.201200439

Eng, J.K., McCormack, a L., Yates, J.R., 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–89. https://doi.org/10.1016/1044-0305(94)80016-2

Erickson, B., Nelson, Winters, P., 2012. Perspective on opportunities in industrial biotechnology in renewable chemicals. Biotechnol. J. 7, 176–85. https://doi.org/10.1002/biot.201100069

Faber, K., 2011. Biotransformations aid organic chemists, ChemInform. https://doi.org/10.1007/978-3-642-17393-6

Ferrarini, M., Moretto, M., Ward, J.A., Šurbanovski, N., Stevanović, V., Giongo, L., Viola, R., Cavalieri, D., Velasco, R., Cestaro, A., Sargent, D.J., 2013. An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. BMC Genomics 14, 670. https://doi.org/10.1186/1471-2164-14-670

Ferrer, M., Beloqui, A., Timmis, K.N., Golyshin, P.N., 2009. Metagenomics for Mining New Genetic Resources of Microbial Communities. J. Mol. Microbiol. Biotechnol. 16, 109–123. https://doi.org/10.1159/000142898

Fischer, J.E., Hatzl, A.-M., Weninger, A., Schmid, C., Glieder, A., 2019. Methanol Independent Expression by Pichia Pastoris Employing De-repression Technologies. J. Vis. Exp. e58589. https://doi.org/doi:10.3791/58589

Fitzgerald, I., Glick, B.S., 2014. Secretion of a foreign protein from budding yeasts is enhanced by cotranslational translocation and by suppression of vacuolar targeting. Microb. Cell Factories 13, 125. https://doi.org/10.1186/s12934-014-0125-0

Fonović, M., Bogyo, M., 2008. Activity-based probes as a tool for functional proteomic analysis of proteases. Expert Rev. Proteomics 5, 721–730. https://doi.org/10.1586/14789450.5.5.721

Frank, K., Sippl, M.J., 2008. High-performance signal peptide prediction based on sequence alignment techniques. Bioinformatics 24, 2172–2176. https://doi.org/10.1093/bioinformatics/btn422

Franz, D., Hartner, S., 2007. Engineering Pichia pastoris for whole-cell biotransformation Doctoral thesis.

Friar, E.A., 2005. Isolation of DNA from Plants with Large Amounts of Secondary Metabolites. Methods Enzymol. 395, 1–12. https://doi.org/10.1016/S0076-6879(05)95001-5

Fukuhara, H., 2006. Kluyveromyces lactis - A retrospective. FEMS Yeast Res. 6, 323–324. https://doi.org/10.1111/j.1567-1364.2005.00012.x

Gasser, B., Prielhofer, R., Marx, H., Maurer, M., Nocon, J., Steiger, M., Puxbaum, V., Sauer, M., Mattanovich, D., 2013. Pichia pastoris : 191–208.

Gassler, T., Heistinger, L., Mattanovich, D., Gasser, B., Prielhofer, R., 2019. CRISPR/Cas9-Mediated Homology-Directed Genome Editing in Pichia pastoris, in: Gasser, B., Mattanovich, D. (Eds.), Recombinant Protein Production in Yeast. Springer New York, New York, NY, pp. 211–225. https://doi.org/10.1007/978-1-4939-9024-5_9

Geer, L.Y., Markey, S.P., Kowalak, J. a, Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W., Bryant, S.H., 2004. Open mass spectrometry search algorithm. J Proteome Res 3, 958–964. https://doi.org/10.1021/pr0499491

Gellissen, G., Hollenberg, C.P., 1997. Application of yeasts in gene expression studies : a comparison of. Methods Enzymol. 190, 87–97.

Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., Weissman, J.S., 2003. Global analysis of protein expression in yeast. Nature 425, 737–41. https://doi.org/10.1038/nature02046

Gibson, D.G., Young, L., Chuang, R., Venter, J.C., Iii, C.A.H., Smith, H.O., America, N., 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases 6, 12–16. https://doi.org/10.1038/NMETH.1318

Go, Y.-M., Jones, D.P., 2014. Redox compartmentalization in eukaryotic cells. Nano Lett 6, 187401. https://doi.org/10.1016/j.bbagen.2008.01.011.Redox

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech 29, 644–652.

Graf, A., Gasser, B., Dragosits, M., Sauer, M., Leparc, G.G., Tüchler, T., Kreil, D.P., Mattanovich, D., 2008. Novel insights into the unfolded protein response using Pichia pastoris specific DNA microarrays 13, 1–13. https://doi.org/10.1186/1471-2164-9-390

Gray, A.N., Koo, B.-M., Shiver, A.L., Peters, J.M., Osadnik, H., Gross, C. a, 2015. High-throughput bacterial functional genomics in the sequencing era. Curr. Opin. Microbiol. 27, 86–95. https://doi.org/10.1016/j.mib.2015.07.012

Guan, B., Chen, F., Su, S., Duan, Z., Chen, Y., Li, H., Jin, J., 2016. Effects of co-overexpression of secretion helper factors on the secretion of a HSA fusion protein (IL2-HSA) in pichia pastoris. Yeast 33, 587–600. https://doi.org/10.1002/yea.3183

Gygi, S.P., Rochon, Y., Franza, B.R., Aebersold, R., 1999. Correlation between Protein and mRNA Abundance in Yeast 19, 1720–1730.

Hamilton, S.R., Bobrowicz, P., Bobrowicz, B., Davidson, R.C., Li, H., Mitchell, T., Nett, J.H., Rausch, S., Stadheim, T.A., Wischnewski, H., Wildt, S., Gerngross, T.U., 2003. Production of complex human glycoproteins in yeast. Science 301, 1244–6. https://doi.org/10.1126/science.1088166

Hamilton, S.R., Gerngross, T.U., 2007. Glycosylation engineering in yeast : the advent of fully humanized yeast 387–392. https://doi.org/10.1016/j.copbio.2007.09.001

Handelsman, J., 2004. Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiol. Mol. Biol. Rev. 68, 669–685. https://doi.org/10.1128/MBR.68.4.669

Hartner, F.S., Glieder, A., 2006a. Regulation of methanol utilisation pathway genes in yeasts 21, 1–21. https://doi.org/10.1186/1475-2859-5-39

Hartner, F.S., Glieder, A., 2006b. Regulation of methanol utilisation pathway genes in yeasts. Microb. Cell Factories 5, 39. https://doi.org/10.1186/1475-2859-5-39

Hartner, F.S., Ruth, C., Langenegger, D., Johnson, S.N., Hyka, P., Lin-Cereghino, G.P., Lin-Cereghino, J., Kovar, K., Cregg, J.M., Glieder, A., 2008. Promoter library designed for fine-tuned gene expression in Pichia pastoris. Nucleic Acids Res. 36, e76. https://doi.org/10.1093/nar/gkn369

Healy, F.G., Ray, R.M., Aldrich, H.C., Wilkie, A.C., Ingram, L.O., Shanmugam, K.T., 1995. Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose. Appl. Microbiol. Biotechnol. 43, 667–674. https://doi.org/10.1007/BF00164771

Hess, M., 2011. Metagenomic Discovery of. Science 463, 463–467. https://doi.org/10.1126/science.1200387

Heyland, J., Fu, J., Blank, L.M., Schmid, A., 2010. Quantitative physiology of Pichia pastoris during glucose-limited high-cell density fed-batch cultivation for recombinant protein production. Biotechnol. Bioeng. 107, 357–68. https://doi.org/10.1002/bit.22836

Hölzer, M., Marz, M., 2019. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. GigaScience 8. https://doi.org/10.1093/gigascience/giz039

Hou, J., Tyo, K.E.J., Liu, Z., Petranovic, D., Nielsen, J., 2012. Metabolic engineering of recombinant protein secretion by Saccharomyces cerevisiae 12, 491–510. https://doi.org/10.1111/j.1567-1364.2012.00810.x

Hughes, C., Ma, B., Lajoie, G.A., 2010. Proteome Bioinformatics 604. https://doi.org/10.1007/978-1-60761-444-9

Idiris, A., Tohda, H., Kumagai, H., 2010a. Engineering of protein secretion in yeast : strategies and impact on protein production 403–417. https://doi.org/10.1007/s00253-010-2447-0

Idiris, A., Tohda, H., Kumagai, H., Takegawa, K., 2010b. Engineering of protein secretion in yeast: Strategies and impact on protein production. Appl. Microbiol. Biotechnol. 86, 403–417. https://doi.org/10.1007/s00253-010-2447-0

Jacob, D., Lewin, A., Meister, B., Appel, B., 2002. Plant-specific promoter sequences carry elements that are recognised by the eubacterial transcription machinery. Transgenic Res. https://doi.org/10.1023/A:1015620016472

Jacobs, D.I., Gaspari, M., van der Greef, J., van der Heijden, R., Verpoorte, R., 2005. Proteome analysis of the medicinal plant Catharanthus roseus. Planta 221, 690–704. https://doi.org/10.1007/s00425-004-1474-4

Jacobs, P.P., Callewaert, N., 2009. N-glycosylation Engineering of Biopharmaceutical Expression Systems 774–800.

Jacobs, P.P., Geysens, S., Vervecken, W., Contreras, R., Callewaert, N., 2009. Engineering complex-type N-glycosylation in Pichia pastoris using GlycoSwitch technology 4. https://doi.org/10.1038/nprot.2008.213

Jahic, M., Veide, A., Charoenrat, T., Teeri, T., Enfors, S., 2006. Process Technology for Production and Recovery of Heterologous Proteins with 1465–1473.

Janssen, D.B., Dinkla, I.J.T., Poelarends, G.J., Terpstra, P., 2005. Bacterial degradation of xenobiotic compounds: evolution and distribution of novel enzyme activities. Environ. Microbiol. 7, 1868–1882. https://doi.org/10.1111/j.1462-2920.2005.00966.x

Jaouadi, B., Ellouz-Chaabouni, S., Rhimi, M., Bejar, S., 2008. Biochemical and molecular characterization of a detergent-stable serine alkaline protease from Bacillus pumilus CBS with high catalytic efficiency. Biochimie 90, 1291–1305. https://doi.org/10.1016/j.biochi.2008.03.004

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., Charpentier, E., 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337, 816–821. https://doi.org/10.1126/science.1225829

Käll, L., Krogh, A., Sonnhammer, E.L.L., 2007. Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. Nucleic Acids Res. 35, 429–432. https://doi.org/10.1093/nar/gkm256

Kickenweiz, T., Glieder, A., Wu, J., 2018. Construction of a cellulose-metabolizing Komagataella phaffii (Pichia pastoris) by co-expressing glucanases and β-glucosidase. Appl. Microbiol. Biotechnol. 102. https://doi.org/10.1007/s00253-017-8656-z

Kim, S., Pevzner, P. a, 2014. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun. 5, 5277. https://doi.org/10.1038/ncomms6277

Kim, T.W., Chokhawala, H.A., Hess, M., Dana, C.M., Baer, Z., Sczyrba, A., Rubin, E.M., Blanch, H.W., Clark, D.S., 2011. High-throughput in vitro glycoside hydrolase (HIGH) screening for enzyme discovery. Angew. Chem. - Int. Ed. 50, 11215–11218. https://doi.org/10.1002/anie.201104685

Kirsch, R., Wielsch, N., Vogel, H., Svatoš, A., Heckel, D.G., Pauchet, Y., 2012. Combining proteomics and transcriptome sequencing to identify active plant-cell-wall-degrading enzymes in a leaf beetle. BMC Genomics 13, 587. https://doi.org/10.1186/1471-2164-13-587

Krainer, F.W., Dietzsch, C., Hajek, T., Herwig, C., Spadiut, O., Glieder, A., 2012. Recombinant protein expression in Pichia pastoris strains with an engineered methanol utilization pathway. Microb. Cell Factories 11, 22. https://doi.org/10.1186/1475-2859-11-22

Krainer, F.W., Gmeiner, C., Neutsch, L., Windwarder, M., Pletzenauer, R., Herwig, C., Altmann, F., Glieder, A., Spadiut, O., 2013. Knockout of an endogenous mannosyltransferase increases the homogeneity of glycoproteins produced in Pichia pastoris. Sci. Rep. 3. https://doi.org/10.1038/srep03279

Küberl, A., Schneider, J., Thallinger, G.G., Anderl, I., Wibberg, D., Hajek, T., Jaenicke, S., Brinkrolf, K., Goesmann, A., Szczepanowski, R., Pühler, A., Schwab, H., Glieder, A., Pichler, H., 2011. High-quality genome sequence of Pichia pastoris CBS7435 154, 312–320. https://doi.org/10.1016/j.jbiotec.2011.04.014

Kullman, S.W., Matsumura, F., 1996. Metabolic pathways utilized by Phanerochaete chrysosporium for degradation of the cyclodiene pesticide endosulfan. Appl. Environ. Microbiol. 62, 593–600.

Kurjan, J., Herskowitz, I., 1982a. Structure of a yeast pheromone gene (MFα): A putative α-factor precursor contains four tandem copies of mature α-factor. Cell 30, 933–943. https://doi.org/10.1016/0092-8674(82)90298-7

Kurjan, J., Herskowitz, I., 1982b. Structure of a yeast pheromone gene (MFα): A putative α-factor precursor contains four tandem copies of mature α-factor. Cell 30, 933–943. https://doi.org/10.1016/0092-8674(82)90298-7

Lacerda, C.M.R., Reardon, K.F., 2008. Environmental proteomics: applications of proteome profiling in environmental microbiology and biotechnology. Brief. Funct. Genomic. Proteomic. 8, 75–87. https://doi.org/10.1093/bfgp/elp005

Lämmle, K., Zipper, H., Breuer, M., Hauer, B., Buta, C., Brunner, H., Rupp, S., 2007. Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. J. Biotechnol. 127, 575–592. https://doi.org/10.1016/J.JBIOTEC.2006.07.036

Leoni, C., Volpicella, M., De Leo, F., Gallerani, R., Ceci, L.R., 2011. Genome walking in eukaryotes. FEBS J. 278, 3953–3977. https://doi.org/10.1111/j.1742-4658.2011.08307.x

Lewin, A., Tran, T.T., Jacob, D., Mayer, M., Freytag, B., Appel, B., 2004. Yeast DNA sequences initiating gene expression in Escherichia coli. Microbiol. Res. https://doi.org/10.1016/j.micres.2004.01.006

Li, S., Yang, X., Yang, S., Zhu, M., Wang, X., 2012. Technology prospecting on enzymes: application, marketing and engineering. Comput. Struct. Biotechnol. J. 2, e201209017. https://doi.org/10.5936/csbj.201209017

Li, Y., Wexler, M., Richardson, D.J., Bond, P.L., Johnston, A.W.B., 2005. Screening a wide host-range, waste-water metagenomic library in tryptophan auxotrophs of Rhizobium leguminosarum and of Escherichia coli reveals different classes of cloned trp genes. Environ. Microbiol. 7, 1927–1936. https://doi.org/10.1111/j.1462-2920.2005.00853.x

Liachko, I., Dunham, M.J., 2014. An autonomously replicating sequence for use in a wide range of budding yeasts. FEMS Yeast Res. 14, 364–367. https://doi.org/10.1111/1567-1364.12123

Lin-cereghino, G.P., Godfrey, L., Cruz, B.J. De, Johnson, S., Khuongsathiene, S., Tolstorukov, I., Yan, M., Lin-cereghino, J., Veenhuis, M., Subramani, S., Cregg, J.M., 2006. Mxr1p , a Key Regulator of the Methanol Utilization Pathway and Peroxisomal Genes in Pichia pastoris 26, 883–897. https://doi.org/10.1128/MCB.26.3.883

Liu, Q., Shi, X., Song, L., Liu, H., Zhou, X., Wang, Q., Zhang, Y., Cai, M., 2019. CRISPR–Cas9-mediated genomic multiloci integration in Pichia pastoris. Microb. Cell Factories 18, 144. https://doi.org/10.1186/s12934-019-1194-x

Looser, V., Bruhlmann, B., Bumbak, F., Stenger, C., Costa, M., Camattari, A., Fotiadis, D., Kovar, K., 2015. Cultivation strategies to enhance productivity of Pichia pastoris: A review. Biotechnol. Adv. 33, 1177–1193. https://doi.org/10.1016/J.BIOTECHADV.2015.05.008

Looser, V., Lüthy, D., Straumann, M., Hecht, K., Melzoch, K., Kovar, K., 2017. Effects of glycerol supply and specific growth rate on methanol-free production of CALB by P. pastoris: functional characterisation of a novel promoter. Appl. Microbiol. Biotechnol. 101, 3163–3176. https://doi.org/10.1007/s00253-017-8123-x

Love, K.R., Shah, K.A., Whittaker, C.A., Wu, J., Bartlett, M.C., Ma, D., Leeson, R.L., Priest, M., Borowsky, J., Young, S.K., Love, J.C., 2016. Comparative genomics and transcriptomics of Pichia pastoris. BMC Genomics 17, 1–17. https://doi.org/10.1186/s12864-016-2876-y

Ma, S.K., Gruber, J., Davis, C., Newman, L., Gray, D., Wang, A., Grate, J., Huisman, G.W., Sheldon, R.A., 2010. A green-by-design biocatalytic process for atorvastatin intermediate. Green Chem. 12, 81–86. https://doi.org/10.1039/b919115c

Madzak, C., Gaillardin, C., Beckerich, J., 2004. Heterologous protein expression and secretion in the non-conventional yeast Yarrowia lipolytica : a review 109, 63–81. https://doi.org/10.1016/j.jbiotec.2003.10.027

Manfred G. Grabherr, Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W., N., Friedman, and A.R., 2013. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat. Biotechnol. 29, 644–652. https://doi.org/10.1038/nbt.1883.Trinity

Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Wang, Z., Yamashita, R.A., Zhang, D., Zheng, C., Bryant, S.H., 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res. 43, D222–D226. https://doi.org/10.1093/nar/gku1221

Martinez, A., Kolvek, S.J., Yip, C.L.T., Hopke, J., Brown, K.A., MacNeil, I.A., Osburne, M.S., 2004. Genetically Modified Bacterial Strains and Novel Bacterial Artificial Chromosome Shuttle Vectors for Constructing Environmental Libraries and Detecting Heterologous Natural Products in Multiple Expression Hosts. Appl. Environ. Microbiol. 70, 2452–2463. https://doi.org/10.1128/AEM.70.4.2452-2463.2004

Mattanovich, D., Gasser, B., Hohenblum, H., Sauer, M., 2004. Stress in recombinant protein producing yeasts 113, 121–135. https://doi.org/10.1016/j.jbiotec.2004.04.035

Mattanovich, D., Graf, A., Stadlmann, J., Dragosits, M., Redl, A., Maurer, M., Kleinheinz, M., Sauer, M., Altmann, F., Gasser, B., 2009. Genome , secretome and glucose transport highlight unique features of the protein production host Pichia pastoris 13, 1–13. https://doi.org/10.1186/1475-2859-8-29

Mccarthy, A., 2010. Third generation DNA sequencing: Pacific biosciences' single molecule real time technology. Chem. Biol. 17, 675–676. https://doi.org/10.1016/j.chembiol.2010.07.004

McGettigan, P. a, 2013. Transcriptomics in the RNA-seq era. Curr. Opin. Chem. Biol. 17, 4–11. https://doi.org/10.1016/j.cbpa.2012.12.008

Meyer, H.P., 2011. Sustainability and biotechnology. Org. Process Res. Dev. 15, 180–188. https://doi.org/10.1021/op100206p

Miné-Hattab, J., Rothstein, R., 2013. Gene Targeting and Homologous Recombination in Saccharomyces cerevisiae, in: Renault, S., Duchateau, P. (Eds.), Site-Directed Insertion of Transgenes. Springer Netherlands, Dordrecht, pp. 71–89. https://doi.org/10.1007/978-94-007-4531-5_3

Moreira, K.A., Albuquerque, B.F., Teixeira, M.F.S., Porto, A.L.F., Lima Filho, J.L., 2002. Application of protease from Nocardiopsis sp. as a laundry detergent additive. World J. Microbiol. Biotechnol. 18, 309–315. https://doi.org/10.1023/A:1015221327263

Mori, M., Tsuge, S., Fukasawa, W., Jeelani, G., Nakada-Tsukui, K., Nonaka, K., Matsumoto, A., Ōmura, S., Nozaki, T., Shiomi, K., 2018. Discovery of Antiamebic Compounds That Inhibit Cysteine Synthase From the Enteric Parasitic Protist Entamoeba histolytica by Screening of Microbial Secondary Metabolites. Front. Cell. Infect. Microbiol. 8, 409. https://doi.org/10.3389/fcimb.2018.00409

Müller, R.-J., Schrader, H., Profe, J., Dresler, K., Deckwer, W.-D., 2005. Enzymatic Degradation of Poly(ethylene terephthalate): Rapid Hydrolyse using a Hydrolase fromT. fusca. Macromol. Rapid Commun. 26, 1400–1405. https://doi.org/10.1002/marc.200500410

Näätsaari, L., Mistlberger, B., Ruth, C., Hajek, T., Hartner, F.S., Glieder, A., 2012a. Deletion of the Pichia pastoris KU70 homologue facilitates platform strain generation for gene expression and synthetic biology. PLoS One 7, e39720. https://doi.org/10.1371/journal.pone.0039720

Näätsaari, L., Mistlberger, B., Ruth, C., Hajek, T., Hartner, F.S., Glieder, A., 2012b. Deletion of the pichia pastoris ku70 homologue facilitates platform strain generation for gene expression and synthetic biology. PLoS ONE 7. https://doi.org/10.1371/journal.pone.0039720

Nacke, H., Kaiser, K., Marhan, S., Sikorski, J., Kandeler, E., Daniel, R., 2016. Estimates of Soil Bacterial Ribosome Content and Diversity Are Employed. Appl. Environ. Microbiol. 82, 2595–2607. https://doi.org/10.1128/AEM.00019-16.Editor

Nam, D.K., Lee, S., Zhou, G., Cao, X., Wang, C., Clark, T., Chen, J., Rowley, J.D., Wang, S.M., 2002. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription 1–5.

Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, a., Shabalov, I., Smirnova, T., Grigoriev, I. V., Dubchak, I., 2014. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res. 42, D26–D31. https://doi.org/10.1093/nar/gkt1069

Nordgård, L., Traavik, T., Nielsen, K.M., 2005. Nucleic Acid Isolation from Ecological Samples—Vertebrate Gut Flora. Methods Enzymol. 395, 38–48. https://doi.org/10.1016/S0076-6879(05)95003-9

Okerberg, E.S., Wu, J., Zhang, B., Samii, B., Blackford, K., Winn, D.T., Shreder, K.R., Burbaum, J.J., Patricelli, M.P., 2005. High-resolution functional proteomics by active-site peptide profiling. Proc. Natl. Acad. Sci. U. S. A. 102, 4996 LP – 5001. https://doi.org/10.1073/pnas.0501205102

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733–D745. https://doi.org/10.1093/nar/gkv1189

Ooyen, A.J.J. Van, Dekker, P., Huang, M., Olsthoorn, M.M.A., Jacobs, D.I., Colussi, P.A., Taron, C.H., 2006. Heterologous protein production in the yeast Kluyveromyces lactis 6, 381–392. https://doi.org/10.1111/j.1567-1364.2006.00049.x

Pappin, D.J.C., Creasy, D.M., Cottrell, J.S., 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data Proteomics and 2-DE. Electrophoresis 3551–3567.

Patanjali, S.R., Parimoo, S., Weissman, S.M., 1991. Construction of 88, 1943–1947.

Patel, J.M., 2009. Biocatalytic synthesis of atorvastatin intermediates. J. Mol. Catal. B Enzym. 61, 123–128. https://doi.org/10.1016/j.molcatb.2009.07.004

Perchey, R.T., Tonini, L., Tosolini, M., Fournié, J.J., Lopez, F., Besson, A., Pont, F., 2019. PTMselect: optimization of protein modifications discovery by mass spectrometry. Sci. Rep. 9, 5–11. https://doi.org/10.1038/s41598-019-40873-3

Piotek, M., Hagedorn, J., Hollenberg, C.P., Gellissen, G., Srasser, A.W.M., 1998. Two novel gene expression systems based on the yeasts _Schwanniomyces occidentalis_ and _Pichia stipitis_. Appl. Microbiol. Biotechnol. 50, 331–338.

Piva, L.C., Bentacur, M.O., Reis, V.C.B., Marco, J.L. De, de Moraes, L.M.P., Torres, F.A.G., 2017. Molecular strategies to increase the levels of heterologous transcripts in Komagataella phaffii for protein production. Bioengineered 8, 441–445. https://doi.org/10.1080/21655979.2017.1296613

Pohl, N.L., 2005. Functional proteomics for the discovery of carbohydrate-related enzyme activities. Curr. Opin. Chem. Biol. 9, 76–81. https://doi.org/10.1016/j.cbpa.2004.12.003

Porro, D., Sauer, M., Branduardi, P., Mattanovich, D., 2005. Recombinant Protein Production in Yeasts Recombinant Protein Production in Yeasts 31.

Potvin, G., Ahmad, A., Zhang, Z., 2012. Bioprocess engineering aspects of heterologous protein production in Pichia pastoris: A review. Biochem. Eng. J. 64, 91–105. https://doi.org/10.1016/j.bej.2010.07.017

Prielhofer, R., Cartwright, S.P., Graf, A.B., Valli, M., Bill, R.M., Mattanovich, D., Gasser, B., 2015. Pichia pastoris regulates its gene-specific response to different carbon sources at the transcriptional, rather than the translational, level. BMC Genomics 16, 1–17. https://doi.org/10.1186/s12864-015-1393-8

Prielhofer, R., Maurer, M., Klein, J., Wenger, J., Kiziak, C., Gasser, B., 2013. Induction without methanol : novel regulated promoters enable high-level expression in Pichia pastoris. Microb. Cell Factories 12, 1. https://doi.org/10.1186/1475-2859-12-5

Prielhofer, R., Reichinger, M., Wagner, N., Claes, K., Kiziak, C., Gasser, B., Mattanovich, D., 2018. Superior protein titers in half the fermentation time: Promoter and process engineering for the glucose-regulated GTH1 promoter of Pichia pastoris. Biotechnol. Bioeng. 115, 2479–2488. https://doi.org/10.1002/bit.26800

Quail, M., Smith, M.E., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., Gu, Y., 2012. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics 13, 1. https://doi.org/10.1186/1471-2164-13-341

Quan, J., Tian, J., 2009. Circular Polymerase Extension Cloning of Complex Gene Libraries and Pathways 4. https://doi.org/10.1371/journal.pone.0006441

Rabausch, U., Juergensen, J., Ilmberger, N., Böhnke, S., Fischer, S., Schubach, B., Schulte, M., Streit, W. V., 2013. Functional screening of metagenome and genome libraries for detection of novel flavonoid-modifying enzymes. Appl. Environ. Microbiol. 79, 4551–4563. https://doi.org/10.1128/AEM.01077-13

Rabinovich, M.L., Bolobova, A. V, Vasil'chenko, L.G., 2004. Fungal Decomposition of Natural Aromatic Structures and Xenobiotics: A Review. Appl. Biochem. Microbiol. 40, 1–17. https://doi.org/10.1023/B:ABIM.0000010343.73266.08

Reetz, M.T., 2013. Biocatalysis in organic chemistry and biotechnology: Past, present, and future. J. Am. Chem. Soc. 135, 12480–12496. https://doi.org/10.1021/ja405051f

Ribitsch, D., Acero, E.H., Greimel, K., Eiteljoerg, I., Trotscha, E., Freddi, G., Schwab, H., Guebitz, G.M., 2012. Characterization of a new cutinase from *Thermobifida alba* for PET-surface hydrolysis. Biocatal. Biotransformation 30, 2–9. https://doi.org/10.3109/10242422.2012.644435

Robertson, D.E., Steer, B. a., 2004. Recent progress in biocatalyst discovery and optimization. Curr. Opin. Chem. Biol. 8, 141–149. https://doi.org/10.1016/j.cbpa.2004.02.010

Ruth, C., Zuellig, T., Mellitzer, A., Weis, R., Looser, V., Kovar, K., Glieder, A., 2010. Variable production windows for porcine trypsinogen employing synthetic inducible promoter variants in Pichia pastoris 181–191. https://doi.org/10.1007/s11693-010-9057-0

Ryšlavá, H., Hýsková, V., Kavan, D., Vaněk, O., 2013. Effect of posttranslational modifications on enzyme function and assembly. J. Proteomics 92, 80–109. https://doi.org/10.1016/j.jprot.2013.03.025

Salisbury, C.M., Cravatt, B.F., 2007. Click Chemistry-Led Advances in High Content Functional Proteomics. QSAR Comb. Sci. 26, 1229–1238. https://doi.org/10.1002/qsar.200740090

Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., Siksnys, V., 2011. The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. Nucleic Acids Res. 39, 9275–9282. https://doi.org/10.1093/nar/gkr606

Schilmiller, a. L., Miner, D.P., Larson, M., McDowell, E., Gang, D.R., Wilkerson, C., Last, R.L., 2010. Studies of a Biochemical Factory: Tomato Trichome Deep Expressed Sequence Tag Sequencing and Proteomics. Plant Physiol. 153, 1212–1223. https://doi.org/10.1104/pp.110.157214

Schmid, A., Dordick, J.S., Hauer, B., Kiener, A., Wubbolt, M., Witholt, B., 2001. Industrial Biocatalysis and Tomorrow. Nature 409, 258–268.

Schneider, T., Keiblinger, K.M., Schmid, E., Sterflinger-Gleixner, K., Ellersdorfer, G., Roschitzki, B., Richter, A., Eberl, L., Zechmeister-Boltenstern, S., Riedel, K., 2012. Who is who in litter decomposition? Metaproteomics reveals major microbial players and their biogeochemical functions. ISME J. 6, 1749–1762. https://doi.org/10.1038/ismej.2012.11

Schoemaker, H.E., 2003. Dispelling the Myths--Biocatalysis in Industrial Synthesis. Science 299, 1694–1697. https://doi.org/10.1126/science.1079237

Schulze, B., Wubbolts, M.G., 1999. Biocatalysis for industrial production of fine chemicals. Curr. Opin. Biotechnol. 10, 609–615. https://doi.org/10.1016/S0958-1669(99)00042-7

Schwarzhans, J., Luttermann, T., Wibberg, D., Winkler, A., Hübner, W., Huser, T., Kalinowski, J., Friehs, K., 2017. A Mitochondrial Autonomously Replicating Sequence from Pichia pastoris for Uniform High Level Recombinant Protein Production 8, 1–15. https://doi.org/10.3389/fmicb.2017.00780

Schwarzhans, J.P., Wibberg, D., Winkler, A., Luttermann, T., Kalinowski, J., Friehs, K., 2016. Non-canonical integration events in Pichia pastoris encountered during standard transformation analysed with genome sequencing. Sci. Rep. 6, 1–12. https://doi.org/10.1038/srep38952

Sellami-Kamoun, A., Haddar, A., Ali, N.E.H., Ghorbel-Frikha, B., Kanoun, S., Nasri, M., 2008. Stability of thermostable alkaline protease from Bacillus licheniformis RP1 in commercial solid laundry detergent formulations. Microbiol. Res. 163, 299–306. https://doi.org/10.1016/j.micres.2006.06.001

Sheldon, R.A., 2014. Green and sustainable manufacture of chemicals from biomass: State of the art. Green Chem. 16, 950–963. https://doi.org/10.1039/c3gc41935e

Shen, W., Kong, C., Xue, Y., Liu, Y., Cai, M., Zhang, Y., Jiang, T., Zhou, X., Zhou, M., 2016a. Kinase screening in pichia pastoris identified promising targets involved in cell growth and alcohol oxidase 1 promoter (P AOX1) regulation. PLoS ONE 11. https://doi.org/10.1371/journal.pone.0167766

Shen, W., Xue, Y., Liu, Y., Kong, C., Wang, X., Huang, M., Cai, M., Zhou, X., Zhang, Y., Zhou, M., 2016b. A novel methanol-free Pichia pastoris system for recombinant protein expression. Microb. Cell Factories 15, 178. https://doi.org/10.1186/s12934-016-0578-4

Simon, C., Herath, J., Rockstroh, S., Daniel, R., 2009. Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice. Appl. Environ. Microbiol. 75, 2964–2968. https://doi.org/10.1128/AEM.02644-08

Simon, G.M., Cravatt, B.F., 2010. Activity-based proteomics of enzyme superfamilies: Serine hydrolases as a case study. J. Biol. Chem. 285, 11051–11055. https://doi.org/10.1074/jbc.R109.097600

Soares, M.B., Bonaldo, M.F., Jelene, P., Su, L., Lawton, L., Efstratiadis, a, 1994. Construction and characterization of a normalized cDNA library. Proc. Natl. Acad. Sci. U. S. A. 91, 9228–32.

Staley, J.T., Konopka, A., 1985. MEASUREMENT OF IN SITU ACTIVITIES OF NONPHOTOSYNTHETIC MICROORGANISMS IN AQUATIC AND TERRESTRIAL HABITATS. Annu. Rev. Microbiol. 39, 321–346. https://doi.org/10.1146/annurev.mi.39.100185.001541

Steele, H.L., Jaeger, K.E., Daniel, R., Streit, W.R., 2008. Advances in recovery of novel biocatalysts from metagenomes. J. Mol. Microbiol. Biotechnol. 16, 25–37. https://doi.org/10.1159/000142892

Steinborn, G., Böer, E., Scholz, A., Tag, K., Kunze, G., Gellissen, G., 2006. methylotrophic Hansenula polymorpha and other yeasts 13, 1–13. https://doi.org/10.1186/1475-2859-5-33

Steinkellner, G., Gruber, C.C., Pavkov-Keller, T., Binter, A., Steiner, K., Winkler, C., Łyskowski, A., Schwamberger, O., Oberer, M., Schwab, H., Faber, K., Macheroux, P., Gruber, K., 2014. Identification of promiscuous ene-reductase activity by mining structural databases using active site constellations. Nat. Commun. 5, 1–9. https://doi.org/10.1038/ncomms5150

Sturmberger, L., Chappell, T., Geier, M., Krainer, F., Day, K.J., Vide, U., Trstenjak, S., Schiefer, A., Richardson, T., Soriaga, L., Darnhofer, B., Birner-Gruenberger, R., Glick, B.S., Tolstorukov, I., Cregg, J., Madden, K., Glieder, A., 2016a. Refined Pichia pastoris reference genome sequence. J. Biotechnol. https://doi.org/10.1016/j.jbiotec.2016.04.023

Sturmberger, L., Wallace, P.W., Glieder, A., Birner-Gruenberger, R., 2016b. Synergism of proteomics and mRNA sequencing for enzyme discovery. J. Biotechnol. https://doi.org/10.1016/j.jbiotec.2015.12.015

Tabb, D.L., Fernando, C.G., Chambers, M.C., 2008. MyriMatch:highly accurate tandem mass spectral peptide identificaiton by multivariate hypergeometric analysis. J Proteome Res 6, 654–661. https://doi.org/10.1021/pr0604054.MyriMatch

Takagi, S., Tsutsumi, N., Terui, Y., Kong, X.Y., 2009. Method for methanol independent induction from methanol inducible promoters in Pichia PATENT APPLICATION PUBLICATION.

Takasaki, Y., 1966. Studies on sugar-isomerizing enzyme. Agric. Biol. Chem. 30, 1247–1253. https://doi.org/10.1080/00021369.1966.10858758

Tang, J., Zeng, Z., Wang, H., Yang, T., Zhang, P., Li, Y., Zhang, A., Fan, W., Zhang, Y., Yang, X., Zhao, S., Tian, G., Zou, L., 2008. An effective method for isolation of DNA from pig faeces and comparison of five different methods. J. Microbiol. Methods 75, 432–436. https://doi.org/10.1016/J.MIMET.2008.07.014

Tang, K., Utairungsee, T., Kanokratana, P., Sriprang, R., Champreda, V., Eurwilaichitr, L., Tanapongpipat, S., 2006. Characterization of a novel cyclomaltodextrinase expressed from environmental DNA isolated from Bor Khleung hot spring in Thailand. FEMS Microbiol. Lett. 260, 91–99. https://doi.org/10.1111/j.1574-6968.2006.00308.x

Tashiro, Y., 1983. Subcellular Compartments and Protein Topogenesis. Cell Struct. Funct. 8, 91–107. https://doi.org/10.1247/csf.8.91

Taupp, M., Mewis, K., Hallam, S.J., 2011. The art and design of functional metagenomic screens. Curr. Opin. Biotechnol. 22, 465–472. https://doi.org/10.1016/j.copbio.2011.02.010

Teeri, T.T., Kumar, V., Lehtovaara, P., Knowles, J., 1987. Construction of cDNA libraries by blunt-end ligation: High-frequency cloning of long cDNAs from filamentous fungi. Anal. Biochem. 164, 60–67. https://doi.org/10.1016/0003-2697(87)90367-8

Tiwari, R., Singh, S., Singh, N., Adak, A., Rana, S., Sharma, A., Arora, A., Nain, L., 2014. Unwrapping the hydrolytic system of the phytopathogenic fungus Phoma exigua by secretome analysis. Process Biochem. 49, 1630–1636. https://doi.org/10.1016/j.procbio.2014.06.023

Truppo, M.D., 2017. Biocatalysis in the Pharmaceutical Industry: The Need for Speed. ACS Med. Chem. Lett. 8, 476–480. https://doi.org/10.1021/acsmedchemlett.7b00114

Tschopp, J.F., Brust, P.F., Cregg, J.M., Stillman, C.A., Gingeras, T.R., 1987. Volume 15 Number 9 1987 Nucleic Acids Research 15, 3859–3876.

Tuffin, M., Anderson, D., Heath, C., Cowan, D.A., 2009. Metagenomic gene discovery: How far have we moved into novel sequence space? Biotechnol. J. 4, 1671–1683. https://doi.org/10.1002/biot.200900235

Uchiyama, T., Miyazaki, K., 2009. Functional metagenomics for enzyme discovery: challenges to efficient screening. Curr. Opin. Biotechnol. 20, 616–622. https://doi.org/10.1016/j.copbio.2009.09.010

Ufart, L., Potocki-Veronese, G., Laville, Ã., 2015. Discovery of new protein families and functions: new challenges in functional metagenomics for biotechnologies and microbial ecology. Front. Microbiol. 6, 1–10. https://doi.org/10.3389/fmicb.2015.00563

Valentin, K., John, U., Medlin, L., 2005. Nucleic Acid Isolation from Environmental Aqueous Samples. Methods Enzymol. 395, 15–37. https://doi.org/10.1016/S0076-6879(05)95002-7

Valli, M., Tatto, N.E., Peymann, A., Gruber, C., Landes, N., Ekker, H., Thallinger, G.G., Mattanovich, D., Gasser, B., Graf, A.B., 2016. Curation of the genome annotation of Pichia pastoris (Komagataella phaffii) CBS7435 from gene level to protein function. FEMS Yeast Res. 16. https://doi.org/10.1093/femsyr/fow051

Van Bogaert, I.N. a, Holvoet, K., Roelants, S.L.K.W., Li, B., Lin, Y.C., Van de Peer, Y., Soetaert, W., 2013. The biosynthetic gene cluster for sophorolipids: A biotechnological interesting biosurfactant produced by Starmerella bombicola. Mol. Microbiol. 88, 501–509. https://doi.org/10.1111/mmi.12200

Vandernoot, V. a, Langevin, S. a, Solberg, O.D., Lane, P.D., Curtis, D.J., Bent, Z.W., Williams, K.P., Patel, K.D., Schoeniger, J.S., Branda, S.S., Lane, T.W., 2012. cDNA normalization by hydroxyapatite chromatography to enrich transcriptome diversity in RNA-seq applications. BioTechniques 53, 373–80. https://doi.org/10.2144/000113937

Vanz, A.L., Nimtz, M., Rinas, U., 2014. Decrease of UPR- and ERAD-related proteins in Pichia pastoris during methanol-induced secretory insulin precursor production in controlled fed-batch cultures. Microb. Cell Factories 13, 1–10. https://doi.org/10.1186/1475-2859-13-23

Vassileva, A., Chugh, D.A., Swaminathan, S., Khanna, N., 2001. Expression of hepatitis B surface antigen in the methylotrophic yeast Pichia pastoris using the GAP promoter 88, 21–35.

Vogl, T., Glieder, A., 2013. Regulation of Pichia pastoris promoters and its consequences for protein production. New Biotechnol. 30, 385–404. https://doi.org/10.1016/j.nbt.2012.11.010

Vogl, T., Hartner, F.S., Glieder, A., 2013. New opportunities by synthetic biology for biopharmaceutical production in Pichia pastoris. Curr. Opin. Biotechnol. 1–8. https://doi.org/10.1016/j.copbio.2013.02.024

Vogl, T., Kickenweiz, T., Pitzer, J., Sturmberger, L., Weninger, A., Biggs, B.W., Köhler, E.M., Baumschlager, A., Fischer, J.E., Hyden, P., Wagner, M., Baumann, M., Borth, N., Geier, M., Ajikumar, P.K., Glieder, A., 2018a. Engineered bidirectional promoters enable rapid multi-gene co-expression optimization. Nat. Commun. 9. https://doi.org/10.1038/s41467-018-05915-w

Vogl, T., Sturmberger, L., Fauland, P.C., Hyden, P., Fischer, J.E., Schmid, C., Thallinger, G.G., Geier, M., Glieder, A., 2018b. Methanol independent induction in Pichia pastoris by simple derepressed overexpression of single transcription factors. Biotechnol. Bioeng. 115, 1037–1050. https://doi.org/10.1002/bit.26529

Vogl, T., Sturmberger, L., Kickenweiz, T., Wasmayer, R., Schmid, C., Hatzl, A.M., Gerstmann, M.A., Pitzer, J., Wagner, M., Thallinger, G.G., Geier, M., Glieder, A., 2016. A Toolbox of Diverse Promoters Related to Methanol Utilization: Functionally Verified Parts for Heterologous Pathway Expression in Pichia pastoris. ACS Synth. Biol. https://doi.org/10.1021/acssynbio.5b00199

Vogl, T., Sturmberger, L., Kickenweiz, T., Wasmayer, R., Schmid, C., Hatzl, A.-M., Gerstmann, M.A., Pitzer, J., Wagner, M., Thallinger, G.G., Geier, M., Glieder, A., 2015. A toolbox of diverse promoters related to methanol utilization – functionally verified parts for heterologous pathway expression in Pichia pastoris. ACS Synth. Biol. acssynbio.5b00199. https://doi.org/10.1021/acssynbio.5b00199

Wang, J., Wang, X., Shi, L., Qi, F., Zhang, P., Zhang, Y., Zhou, X., Song, Z., Cai, M., 2017. Methanol-Independent Protein Expression by AOX1 Promoter with trans-Acting Elements Engineering and Glucose-Glycerol-Shift Induction in Pichia pastoris. Sci. Rep. 7, 1–12. https://doi.org/10.1038/srep41850

Warnecke, F., Hess, M., 2009. A perspective: Metatranscriptomics as a tool for the discovery of novel biocatalysts. J. Biotechnol. 142, 91–95. https://doi.org/10.1016/j.jbiotec.2009.03.022

Weiß, S., Lebuhn, M., Andrade, D., Zankel, A., Cardinale, M., Birner-Gruenberger, R., Somitsch, W., Ueberbacher, B.J., Guebitz, G.M., 2013. Activated zeolite—suitable carriers for microorganisms in anaerobic digestion processes? Appl. Microbiol. Biotechnol. 97, 3225–3238. https://doi.org/10.1007/s00253-013-4691-6

Wellenreuther, R., Schupp, I., Poustka, A., Wiemann, S., Consortium, T.G. cDNA, 2004. SMART amplification combined with cDNA size fractionation in order to obtain large full-length clones. BMC Genomics 5, 36. https://doi.org/10.1186/1471-2164-5-36

Weninger, A., Fischer, J.E., Raschmanová, H., Kniely, C., Vogl, T., Glieder, A., 2018. Expanding the CRISPR/Cas9 toolkit for Pichia pastoris with efficient donor integration and alternative resistance markers. J. Cell. Biochem. 119, 3183–3198. https://doi.org/10.1002/jcb.26474

Weninger, A., Hatzl, A.-M., Schmid, C., Vogl, T., Glieder, A., 2016. Combinatorial optimization of CRISPR/Cas9 expression enables precision genome engineering in the methylotrophic yeast Pichia pastoris. J. Biotechnol. 235, 139–149. https://doi.org/10.1016/J.JBIOTEC.2016.03.027

Wilmes, P., Wexler, M., Bond, P.L., 2008. Metaproteomics Provides Functional Insight into Activated Sludge Wastewater Treatment. PLoS ONE 3, e1778. https://doi.org/10.1371/journal.pone.0001778

Wo, A., n.d. Mpp1 Japanese patent Yeast for transformation and process for producing protein.

Woodley, J.M., 2013. Protein engineering of enzymes for process applications. Curr. Opin. Chem. Biol. 17, 310–316. https://doi.org/10.1016/j.cbpa.2013.03.017

Xiao, R., Wilkinson, B., Solovyov, A., Winther, J.R., Holmgren, A., Lundström-Ljung, J., Gilbert, H.F., 2004. The contributions of protein bisulfide isomerase and its homologues to oxidative protein folding in the yeast endoplasmic reticulum. J. Biol. Chem. 279, 49780–49786. https://doi.org/10.1074/jbc.M409210200

Yu, X.W., Sun, W.H., Wang, Y.Z., Xu, Y., 2017. Identification of novel factors enhancing recombinant protein production in multi-copy Komagataella phaffii based on transcriptomic analysis of overexpression effects. Sci. Rep. 7, 1–12. https://doi.org/10.1038/s41598-017-16577-x

Yurimoto, H., 2009. Molecular Basis of Methanol-Inducible Gene Expression and Its Application in the Methylotrophic Yeast Candida boidinii. Biosci. Biotechnol. Biochem. 73, 793–800. https://doi.org/10.1271/bbb.80825

Yurimoto, H., Sakai, Y., 2009. Methanol-inducible gene expression and heterologous protein production in the methylotrophic yeast Candida boidinii. Biotechnol. Appl. Biochem. 53, 85–92. https://doi.org/10.1042/BA20090030

Zahrl, R.J., Mattanovich, D., Gasser, B., 2018. The impact of ERAD on recombinant protein secretion in pichia pastoris (Syn komagataella spp.). Microbiol. U. K. 164, 453–463. https://doi.org/10.1099/mic.0.000630

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S., Davies, S.R., Wang, S., Wang, P., Kinsinger, C.R., Rivers, R.C., Rodriguez, H., Townsend, R.R., Ellis, M.J.C., Carr, S. a, Tabb, D.L., Coffey, R.J., Slebos, R.J.C., Liebler, D.C., 2014. Proteogenomic characterization of human colon and rectal cancer. Nature 513, 382–7. https://doi.org/10.1038/nature13438

Zhang, W., Hywood Potter, K.J., Plantz, B.A., Schlegel, V.L., Smith, L.A., Meagher, M.M., 2003. Pichia pastoris fermentation with mixed-feeds of glycerol and methanol: growth kinetics and production improvement. J. Ind. Microbiol. Biotechnol. 30, 210–215. https://doi.org/10.1007/s10295-003-0035-3

Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R., Siebert, P.D., 2001. Reverse Transcriptase Template Switching: A SMART^{TM} Approach for Full-Length cDNA Library Construction. BioTechniques 30, 892–897. https://doi.org/10.2144/01304pf02

Research Article

The meta-transcriptome of the fern species *Pteridium aquilinum* reveals the presence of novel surface-active proteins

Lukas Sturmberger[a], Agnieszka Przylucka[d], Andrea Mellitzer[a], Anna Hatzl[a,b], Irina S. Druzhinina[d] and Anton Glieder[a,b,c]

[a] Austrian Center of Industrial Biotechnology (ACIB), Petersgasse 14, 8010 Graz, Austria
[b] Institute of Molecular Biotechnology, Graz University of Technology, Petersgasse 14, 8010 Graz, Austria
[c] bisy e.U., Wetzawinkel 20, 8200 Hofstaetten/Raab, Austria
[d] Institute of Chemical Engineering, Vienna University of Technology, Karlsplatz 13, 1040 Vienna, Austria

Abbreviations

RNA	ribonucleic acid
PET	polyethylene terephthalate
HFB	hydrophobin
HRP	horseradish peroxidase
TEF	transcription elongation factor
AOX	alcohol oxidase
EMBL	European molecular biology laboratory
BSA	bovine serum albumin
WCA	water contact angle measurement
BLAST	basic local alignment search tool
RPC	reverse phase column

Keywords

surface-active, hydrophobin, *Komagataella phaffii,* reversed-phase chromatography, HRP fusion, glass coating

Abstract

Hydrophobins are amphiphilic proteins found in many different fungal species. Members of this protein family have functions in reducing the surface tension of water, assist in adhesion and provide protective surface coatings. Here, we describe the identification of five hydrophobin genes from a fern meta-transcriptome (*Pteridium sp.*) by RNA sequencing and sequence similarity-based search methods. The identified sequences were codon optimized for protein expression in the methylotrophic yeast *Komagataella phaffii*. We succeeded in purifying the expressed target proteins by reversed-phase chromatography and characterized their surface activity through means of contact angle measurements. On glass matrices a surface tension lowering effect could be detected with four of the five tested hydrophobins, however not on the matrix polyethylene terephthalate (PET) matrix. Subsequently, fusion proteins expressing horseradish peroxidase in frame with the discovered hydrophobins were expressed in a secretory manner in *K. phaffii.* Their ability to specifically coat glass surfaces was tested and showed a specific binding on glass beads for one hydrophobin as measured by HRP activity.

Introduction

The surfaces of plants are colonized by different organisms such as bacteria and fungi. For most of this species the attachment to its surface is the first step of infecting the host organism (Łaźniewska et al., 2012; Nicholson and Epstein, 1991; Tucker and Talbot, 2001). As such, plant surfaces which have not been treated, washed or altered chemically will include surface-attached as well as endophytic microorganisms and their derived nucleic acids. During the isolation of genetic material from plant samples, inadvertently also bacterial and fungal genes are enriched. Subsequent analysis on the DNA and RNA level might therefore reveal sequences from different origins, depending on the degree of surface contamination. In addition to pathogenic and asymptomatically occurring species, endophytic organisms would also be unearthed in this manner. In this study a transcriptome sequence of the fern species *Pteridium aquilinum* (Lanfranchi et al., 2015) formed the basis for a protein discovery approach. As described above, the plant material used for RNA isolation and sequencing were not surface sterilized and therefore contained non-host RNA which was evident in the composition of the transcriptome. Interestingly, we were able to identify five hydrophobin genes from this fern's meta-transcriptome by RNA sequencing and sequence similarity-based search methods.

Hydrophobins are a class of small, secreted proteins which are unique to fungi. The sequences within the hydrophobin family share very little amino acid sequence similarity except for a characteristic eight cysteine residue pattern, forming four disulfide bonds. (Wösten et al., 1999b, 1999a). They are amphiphatic molecules and have a tendency to self-assemble at hydrophobic-hydrophilic interfaces. In doing so, they enable the formation of aerial hyphae. (Wösten, 2001; Wösten et al., 1994). The precise function of these proteins lies in reducing the surface tension of water, assisting in surface adhesion as well as providing protective surface coatings (Markus B Linder et al., 2005; Wessels, 2000; Wösten and de Vocht, 2000). The Wösten lab amongst others was instrumental in the discovery and characterization of this class of proteins and has initially shown the applicability of hydrophobins by reversing the wettability of Teflon and oil droplets (Wösten et al., 1994)(Lugones et al., 1996; Wösten et al., 1995). The application of hydrophobins to control the binding of cells to certain surfaces, as well as molecule attachment to surfaces that do not normally have a high affinity for, was also proposed. These implementations were suggested to take place in the field of nanotechnology, tissue engineering and drug delivery (Cox and Hooley, 2009; Scholtmeijer et al., 2001). Lipases were already non-covalently immobilized and activated through the use of hydrophobins, which were covalently applied as spots on hydrophilic surfaces (Palomo et al., 2003). In a similar manner, Espino-Rammer and colleagues were able to show two novel class II hydrophobins from *Trichoderma spp.* that stimulate enzymatic hydrolysis of poly-ethylene terephtalate when expressed as fusion protein with a *Humicola insolens* cutinase by an isotherm-like adsorption behavior (Espino-Rammer et al., 2013). Due to their small size and biophysical properties, hydrophobins have been investigated for the separation of biomolecules more specifically, as a purification tag. The hydrophobin HFBI of *Trichoderma reesei* was successfully expressed as an N-terminal fusion with chicken avidin in insect cells and enabled the one-step purification in an aqueous micellar two-phase system (Lahtinen et al., 2008). A similar two-phase system approach was later on used to aid in the recovery of a hydrophobin endoglucanase fusion protein (Collén et al., 2002). However, there also exist examples where immobilization was not successful. Linder and co-workers have shown that, while *Trichoderma reesei* HFBI and HFBII efficiently bind surfaces, these proteins did not cause immobilization as a fusion partner (Linder et al., 2002).

In this study we wanted to investigate the potential of the newly discovered hydrophobins for their ability to specifically attach to glass or PET surfaces (Espino-Rammer et al., 2013; Van der Vegt et al., 1996; Wösten et al., 1995). Due to their native secretory mode of production and the presence of several disulfide bridges we attempted to express the coding sequences in the methylotrophic yeast species *Komagataella phaffii*. Hydrophobin fusions with horseradish peroxidase C1A (HRP) proteins to test a potential biotechnological application were generated and the fusion proteins were screened for their ability to improve attachment to glass surfaces on glass beads via their hydrophobin protein partner in comparison to their native, hydrophobin-free counterparts.

Material and Methods

Strains, Plasmids, Chemicals, and Media. For all expression experiments outlined in this publication the *K. phaffii* CBS7435 mutS strain was employed (Näätsaari et al., 2012b; Sturmberger et al., 2016a). The hydrophobin expression vectors were cloned in the backbone of the pPpT4-S vector reported by Näätsaari et al. via Gibson assembly (Gibson Assembly® HiFi 1-Step Kit, SGI DNA, San Diego, CA, USA) (Näätsaari et al., 2012b). The promoter employed for expression for all constructs was the promoter of the *CTA1* gene of *K. phaffii* (Vogl et al., 2015). The pPpT4-S vector was engineered by removing the fragment encoding the *P_{AOX1}* and exchanged for the *P_{CAT1}* (500bps upstream of the ATG of the *CTA1* gene) initially described by Vogl and co-workers (Vogl et al., 2015). Downstream of the *P_{CAT1}* a Kozak sequence consisting of CGAAACG was placed followed by the ATG of the corresponding gene to be expressed. Additionally, an 819bp long fragment from *Schizosaccharomyces pombe* acting as an autonomously replicating sequence (ARS) was included into the plasmid sequence (Clyne and Kelly, 1995). The regulatory elements controlling the transcription of the zeocin selection cassette were exchanged for the *P_{TEF1}* and terminator sequence from *Ashbya gossypii* (S Steiner and Philippsen, 1994). This set-up allows the selection and expression in more than one yeast (Camattari et al., 2016; Liachko and Dunham, 2014).

The five hydrophobin genes were ordered as synthetic gene fragments (Integrated DNA Technologies, Leuven, Belgium) (supplementary File S1) and amplified using the primers outlined in table 1. Additional vectors to produce the respective hydrophobin gene in frame with the coding sequence of HRP isoenzyme C1A (Florian W Krainer et al., 2013) were designed by the primers in table 1. The HRP was produced in frame with the *Saccharomyces cerevisiae* alpha mating factor prepro-leader and the respective hydrophobin as a C-terminal fusion separated by a (GGGGS)$_3$ linker. The resulting plasmids were termed pPpT4_Alpha_S_CAT1_tig(1-5) as well as pPpT4_Alpha_S_CAT1_HRPC1A_tig(1-5). As an empty vector control the plasmid pPpT4_Alpha_S_CAT1_HRPC1A was used, which contains no C-terminal fusion partner. Additionally, we produced the hydrophobins in frame with the *Saccharomyces cerevisiae* alpha mating factor prepro-leader and the HRP as a C-terminal fusion separated by a (GGGGS)$_3$ linker. The resulting plasmids were termed pPpT4_Alpha_S_CAT1_tig(1-5)_HRPC1A.

As a selection agent zeocin (Invivogen, Toulouse, France) solubilized in water was used in concentrations of 25µg/mL for *E. coli* and 100µg/mL for *K. phaffii*. All cultivations in *E. coli* were performed in LB-Lennox (Carl Roth GmbH + Co. KG, Karlsruhe, Germany) at 37°C. *K. phaffii* was cultivated either in YPD2% (10g/L yeast extract (Carl Roth GmbH + Co. KG, Karlsruhe, Germany), 20g/L Bactopeptone (BD Sciences, Franklin Lakes, USA) 20g/L D-glucose (Carl Roth GmbH + Co. KG, Karlsruhe, Germany)) or BM media pH 6,0 (13,4g/L YNB BD Sciences, Franklin Lakes, USA, 6g/L K2HPO4, 24g/L KH2PO4, 0,0004g/L Biotin Carl Roth GmbH + Co. KG, Karlsruhe, Germany) either supplemented with 20g/L glucose or methanol (Carl Roth GmbH + Co. KG, Karlsruhe, Germany). 15g/L agar was added for plate media preparations (Carl Roth GmbH + Co. KG, Karlsruhe, Germany). Recombinant DNA manipulations were performed in *E. coli* TOP10F' (Invitrogen Corp., Carlsbad, CA, USA). Thermo Scientific Fisher fast digest restriction endonucleases were used for all restriction endonuclease digestions.

Electrocompetent *E. coli* TOP10 F' cells were cultivated according to the protocol of Seidman and co-workers (Seidman 2001) and cells stored at -80°C. For the electroporation, 3 µL Gibson assembly reaction mixture (no desalting) were mixed with 80µL frozen cells and thawed on ice. The transformation was performed with 2.5 kV/25 µF/200 Ω (Bio-Rad Gene Pulser System, Bio-Rad Laboratories Inc., Hercules, CA, United States). Immediately after shock supply, 1mL of SOC Medium (Carl Roth GmbH + Co. KG, Karlsruhe, Germany) was added and the reaction was regenerated at 37°C, 650rpm for 1 hour. Electrocompetent *K. phaffii* cells were prepared according to the condensed protocol (Lin-Cereghino et al., 2005). Before plating, cells were regenerated for two hours at 28°C with 500 µL 1M sorbitol and 500µL YPD2%. Prior to electroporation each plasmid was linearized by *SwaI* digestion and electroporation of 1 µg linear DNA was performed under following conditions: 1.5 kV, 25 µF and 200 Ω.

Table 1: Primer used within this study

Primer Designation	Primer Sequence 5'-3'	Resulting plasmids
tig1 SS fwd	5'-*AGAAGAGAGAGGCCGAAGCT***GTTCCTACCACTGCCACCGG**-3'	pPpT4_α_S_CAT1_ tig1-5
tig2 SS fwd	5'-*AGAAGAGAGAGGCCGAAGCT***CGTCCTCAGATCGCTGGTCC**-3'	
tig3 SS fwd	5'-*AGAAGAGAGAGGCCGAAGCT***ACTCCTGAGCCATcTCCaGAG**-3'	
tig4 SS fwd	5'-*AGAAGAGAGAGGCCGAAGCT***AACCCAGTTAGACTTGCCGC**-3'	
tig5 SS fwd	5'-*AGAAGAGAGAGGCCGAAGCT***GCACCAGATGCCAGACGTC**-3'	
tig1 SS rev	5´-*TGGCATTCTGACATCCTCT***TGAGCGGCCGCTTAgtgatggtgatggtgatg**-3'	
SSalpha-HRP fwd	5'-*GTCAAGACTTACAATTAAAACGAAACG***ATGAGATTCCCATCTATTTTCACCGCTG**-3'	pPpT4_α_S_CAT1_ HRPC1A_tig1-5
SSalpha-HRP rev	5'*CGATCCGCCACCGCCAGAGCCACCTCCGCCTGAACCGCCTCCACCTGA***GTTAGAGTTGACAACACGGCAGTT**-3'	
HRP tig1 fwd	5'-*GCTCTGGCGGTGGCGGATCG***GTTCCTACCACTGCCACCGG**-3'	
HRP tig2 fwd	5'-*GCTCTGGCGGTGGCGGATCG***CGTCCTCAGATCGCTGGTCC**-3'	
HRP tig3 fwd	5'-*GCTCTGGCGGTGGCGGATCG***ACTCCTGAGCCATcTCCaGA**-3'	
HRP tig4 fwd	5'-*GCTCTGGCGGTGGCGGATCG***AACCCAGTTAGACTTGCCGC**-3'	
HRP tig5 fwd	5'-*GCTCTGGCGGTGGCGGATCG***GCACCAGATGCCAGACGTCgTG**-3'	
HRP tig rev	5'-*GCATTCTGACATCCTCTTGA***TTAGTGATGGTGATGGTGATG**-3'	
tig1 HRP rev	5'-*CTCCGCCTGAACCGCCTCCACCTGA***CAAGATTGGAGCGCAGTTCTGGATG**-3'	pPpT4_α_S_CAT1_ tig1-5_HRPC1A
tig2 HRP rev	5'-*CTCCGCCTGAACCGCCTCCACCTGA***GAAGATACCGGTGGACTGAGATCC**-3'	
tig3 HRP rev	5'-*CTCCGCCTGAACCGCCTCCACCTGA***AGCATTGACGGCgtTGCaAGCCTGGAC**-3'	
tig4 HRP rev	5'-*CTCCGCCTGAACCGCCTCCACCTGA***AACCACTGGaGTGCAGGTTTCAAGTC**-3'	
tig5 HRP rev	5'-*CTCCGCCTGAACCGCCTCCACCTGA***AGATCCACCaACCTCGGtGCATcCCTC**-3'	
HRP (GS) fwd	5'*GGCGGTTCAGGCGGAGGTGGCTCTGGCGGTGGCGGATCG***CAACTTACTCCAACCTTCTAC**-3'	
HRP His rev	5'*ATTCTGACATCCTCTTGATTAGTGATGGTGATGGTGATG***TGAGTTAGAGTTGACAACACG**-3'	
pCAT1 fwd	*CCAGCTTAAGGCCGCCTCGGCCAGATCT***GAACTCCGAATGCGGTTCTCCTGTAACC**	pPpT4_α_S_CAT1_ HRPC1A
aMF rev	*GGAGTAAGTTGGTGATGGTGGTGATGGTG***AGCTTCGGCCTCTCTCTTCTCGAGAG**	
HRP ctrl fwd	*CGAGAAGAGAGAGGCCGAAGCT***CACCATCACCACCATCACCAACTTACTCC**	
HRP ctrl rev	*ATTCTGACATCCTC***TTGATTAGTGATGGTGATGGTGATGTGAGTTAGAGTTGACAACACG**	

The basis for all vector constructions is the pPpT4_Alpha_S_pCAT1_CalB vector described by Vogl and co-workers (Vogl et al., 2015). We adapted the vector to include an autonomously replicating sequence (ARS) from *Schizosaccharomyces pombe* (Clyne and Kelly, 1995) by digesting the vector with *BglII* and *SwaI* (Thermo Scientific Fisher, Waltham, MA, USA) and incorporating the PCR product derived from primer T4_SpARS_fwd and T4_SpARS_rev (*S. pombe* DSM 70576, L972 gDNA as template) via Gibson Assembly (SGI Inc., San Diego, CA, USA). In order to adapt the vector for use as a shuffle vector in several different yeast species the regulatory elements of the zeocin selection cassette were exchanged for the promoter and terminator of the *Ashbya gossypii* TEF1 gene (Sabine Steiner and Philippsen, 1994). This was achieved by digesting the vector with *BsaI* and *PstI* (Thermo Scientific Fisher, Waltham, MA, USA) and providing a synthetically synthesized gBlock (Integrated DNA Technologies, Leuven, Belgium) containing the *Ashbya gossypii* TEF1 promoter, the Sh ble gene product and the TEF1TT terminator with suitable vector overhangs in a Gibson Assembly reaction. The vector generated in this manner and isolated by plasmid minipreparation (Promega, Fitchburg, WI, USA) contained a zeocin selection cassette and an autonomously replicating sequence functional in at least in three tested yeast species – *Komagataella phaffii*, *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae* (Master Thesis Jasmin Elgin Fischer, Development of new Methods for reliable cDNA Library Generation and Expression, 2016). The promoter of the *K. phaffii CTA1* gene allows expression of the eGFP model protein under both derepressed and methanol induced conditions. The transcriptional termination is mediated by the terminator sequence of the *AOX1* gene and propagation in *E. coli* proceeded via the pUC origin of replication in LB-Lennox Medium (Carl Roth GmbH + Co. KG, Karlsruhe, Germany). All vector sequences were sequence verified by Sanger sequencing.

For the generation of hydrophobin expression vectors pPpT4_α_S_CAT1_tig1-5, the final plasmid as described above was digested with *EcoRI* and *NotI* (Thermo Scientific Fisher, Waltham, MA, USA) and gel-purified (Promega, Fitchburg, WI, USA). PCR products (Phusion DNA polymerase, Thermo Scientific Fisher, Waltham, MA, USA) and linearized vector were assembled with Gibson assembly and positive transformants selected via plating on LB zeocin plates. The plasmids pPpT4_α_S_CAT1_HRPC1A_tig1-5 were generated by entry vector digestion with *EcoRI* and *NotI* and PCR amplification of the HRP gene, and the respective hydrophobin genes containing a $(GGGGS)_3$ linker and a 6xHis tag. For N-terminal fusion of the hydrophin sequence to the HRPC1A protein as in plasmids pPpT4_α_S_CAT1_tig1-5_HRPC1A the *EcoRI/NotI* digested entry vector was used and assembled with PCR products as outlined in table 1 (Gibson Assembly® HiFi 1-Step Kit, SGI DNA, San Diego, CA, USA). As a control vector we generated pPpT4_α_S_CAT1_HRPC1A by digesting pPpT4_α_S_CAT1_tig1_HRPC1A with *BglII* and *NotI* and assembling PCR products generated by primer (table 1). All vector sequences were sequence verified by Sanger sequencing.

Transcriptome assembly and sequence discovery. The transcriptome sequence of the fern species *Pteridium aquilinum* utilized for sequence discovery was generated by Lanfranchi and co-workers and can be accessed via the EMBL-EBI European Nucleotide Archive under the study accession number PRJEB10897 (Lanfranchi et al., 2017). The discovery of the hydrophobin sequences was mediated by using the ClusterControl software package (Stocker et al., 2004) at reduced setting parameters with the conserved domains of Hydrophobin (pfam01185) and Hydrophobin_2 (pfam06766) as query sequences. The coding sequences of five positive hits discovered in this manner were subjected to a BLASTp search against the NCBI nr/nt database and the presence of conserved domains pfam01185 (fungal hydrophobin) was verified. All sequences were codon optimized for expression in *K. phaffii* (IDT codon optimization tool) with a C-terminal 6xhis tag added for detection and purification.

Production and purification via RPC. Clones were grown either in shake flaks at 28°C and 160rpm or 96 deep-well plates (Bel-Art Products, Wayne, NJ, United States) at 28°C and 320rpm. Clones were grown for 60h in 250µL BMD1%, followed by addition of 250µL BMM1% and 12 hour pulses of 50 µl BMM5% for an additional 72 hours. The supernatants were collected by centrifugation at 4000 rpm at 4°C for 5 min (Eppendorf 5810R, Hamburg, Deutschland). To account for plate-to-plate and cultivation condition variabilities, a subset of clones was subjected to a rescreen, also performed in 96 deep-well plates. At least four clones per construct were cultivated under the same conditions in octuplicates and the average and standard deviations were calculated from these values. After methanol induction culture supernatants were collected by centrifugation at 4000 rpm and 4°C (Eppendorf 5810R, Eppendorf, Hamburg, Germany) and ultra-filtered using a 0,22µM filter (GE Healthcare, Austria). Protein purification was performed on an Amersham Bioscience Äkta Purifier connected to a Resource RPC column (GE Healthcare, Austria). Protein binding was performed in 0.1% TFA solution (Buffer A). For the protein elution we used a 0.1% TFA in acetonitrile (Buffer B) gradient starting at 0% B with a target concentration of 100% B. The fractionation size was set to 0.2mL and the fractions collected in a chilled 96 deep well plate for subsequent lyophilization to dispose of the acetonitrile. Three different wavelengths were measured (280nm, 230nm and 215 nm). The acetonitrile gradient was set to spread over 15 column volumes.

Western Blot and Dot-Blot detection. Hydrophobins were detected via a C-terminal 6xhis tag incorporated in all hydrophobin expression vectors. The protein transfer was performed in a Bio-Rad chamber (Bio-Rad, Austria) using 0,22µM Amersham™ Protran™ nitrocellulose membrane and a Pierce™ 10X Western Blot Transfer Buffer. Penta·His Antibody (Qiagen, Austria) was diluted in TBS buffer in a ratio of 1:5000 (292,7g/L NaCl, 4,24g/L Tris, 26g/L Tris-HCl, pH 7,5 Carl Roth GmbH + Co. KG, Karlsruhe, Germany) and Goat anti-Mouse IgG Cross-Adsorbed HRP Secondary Antibody (Thermo Scientific Fisher) was diluted 1:20.000 in TBS. As a blocking agent a 3% BSA fraction V solution in TBS buffer was used (Carl Roth GmbH + Co.

KG, Karlsruhe, Germany). Western Blots and dot blots were washed with TBST Buffer (1ml/L Tween 20 Carl Roth GmbH + Co. KG, Karlsruhe, Germany). Blocking and antibody incubation were performed at room temperature for one hour followed by three consecutive washing steps for five minutes each. The detection was performed with Pierce™ ECL Western Blotting Substrate with recording times between 30 seconds and five minutes in a Syngene G:Box (Syngene, Cambridge, UK). Dot-Blot assays were performed with a Bio-Dot SF Microfiltration Apparatus (Bio-Rad, Austria) by applying 50µL of culture supernatant to each well, followed by an incubation for 15 minutes and two consecutive wash steps with 50µL TBS buffer. The apparatus was disassembled, and membranes subjected to regular western blot membrane treatment.

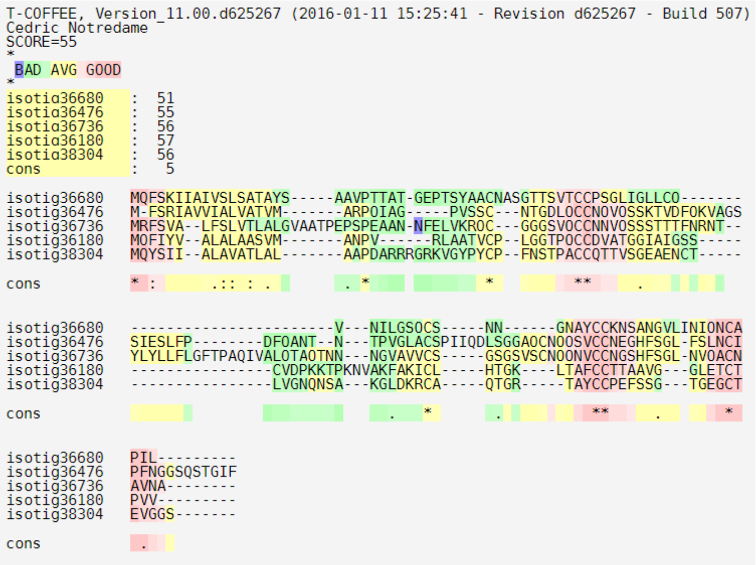
Water contact angle measurements (WCA). For all measurements, the samples were diluted with 0.1M potassium phosphate buffer pH 6.5. Water contact angle measurements were performed as described by Espino-Rammer et al. in a drop-shape analysis system (DSA 100; Kruss GmbH, Hamburg, Germany) (Espino-Rammer et al., 2013). Water droplets were applied to the PET and glass surfaces and the contact angle measured after 3 seconds.

Determination of HRP activity in buffer and coated beads. Protein concentrations were measured at 595nm by Bradford assay (Sigma-Aldrich Protein Assay Kit, Sigma-Aldrich, Austria). BSA standards ranging from 0.005mg-2mg/mL were used for a standard curve measurement. The activity of HRP and HRP-hydrophobin fusion proteins was determined by an ABTS assay (2,2'-azino-bis(3-ethylbenzthiazoline-6-sulfonic acid) diammonium salt, Sigma-Aldrich, Austria) and measured in a BioTek H1 plate reader (Biotek, Vermont, USA) (Florian W Krainer et al., 2013). 10µL of supernatants were mixed with 140µL of 1mM ABTS solution in a 50mM potassium phosphate buffer at pH 6.5. The reaction was pre-incubated at 37°C for 10 minutes and started by addition of 20µL 0.078% H2O2 (Carl Roth GmbH + Co. KG, Karlsruhe, Germany). Absorption changes at 415nm were recorded for 300 seconds and rates from these values calculated. From a commercial source of HRP (HRP 6A, Sigma-Aldrich, Austria) a standard curve from 0,05 U/mL to 1 U/mL was prepared.

Results and Discussion

By using the hydrophobin protein sequence (UniProtKB - D8QCG9) of the *HYD1* gene from *Schizophyllum commune* (strain H4-8 / FGSC 9210) and blasting it against the transcriptome of

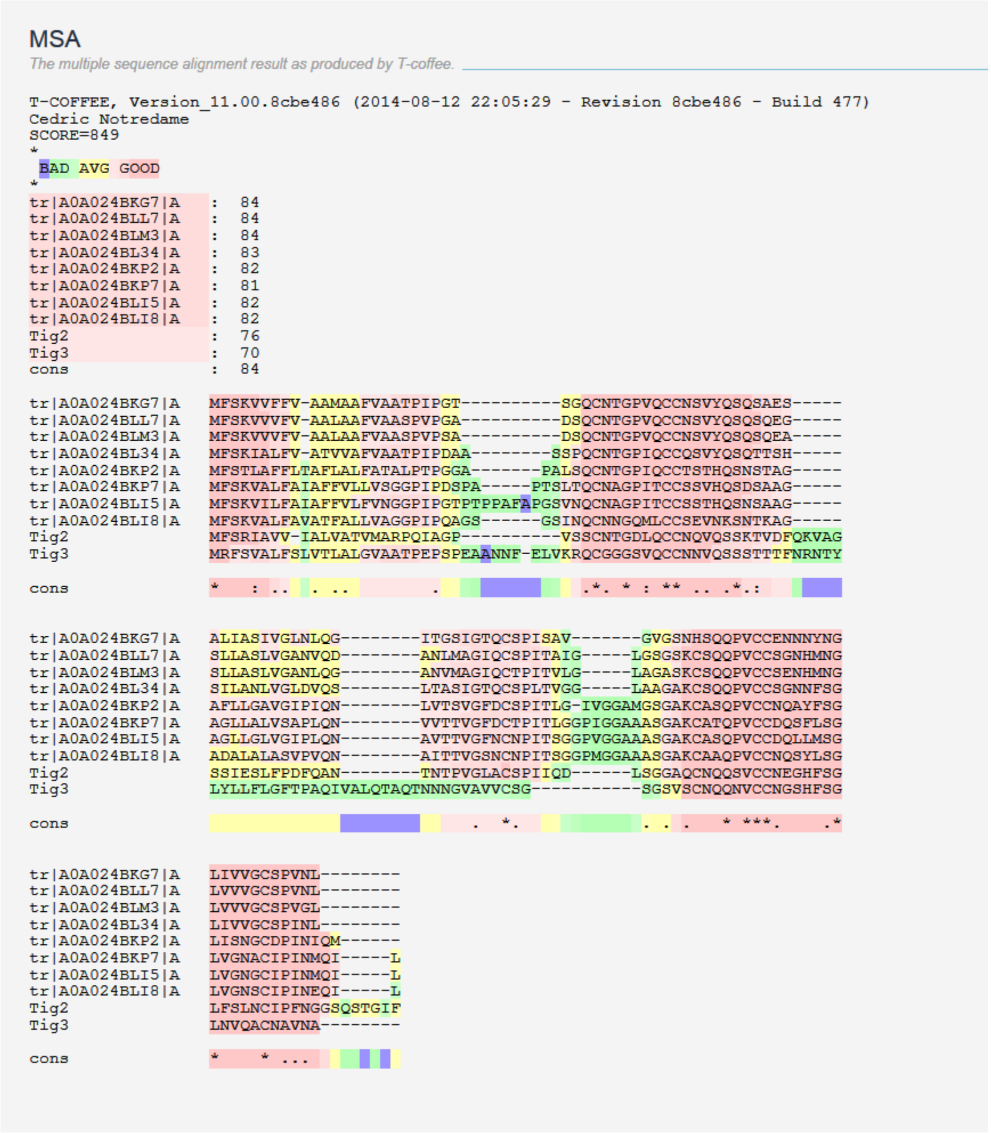
P. aquilinum (Lanfranchi et al., 2017) using the Cluster Control (Stocker et al., 2004) platform we were able to identify five potential hydrophobin sequences (figure 1).



```
T-COFFEE, Version_11.00.d625267 (2016-01-11 15:25:41 - Revision d625267 - Build 507)
Cedric Notredame
SCORE=55
*
 BAD AVG GOOD
*
isotia36680   :  51
isotia36476   :  55
isotia36736   :  56
isotia36180   :  57
isotia38304   :  56
cons          :   5

isotig36680   MQFSKIIAIVSLSATAYS-----AAVPTTAT-GEPTSYAACNASGTTSVTCCPSGLIGLLCO-------
isotig36476   M-FSRIAVVIALVATVM-------ARPOIAG-----PVSSC---NTGDLOCCNOVOSSKTVDFOKVAGS
isotig36736   MRFSVA--LFSLVTLALGVAATPEPSPEAAN-NFELVKROC---GGGSVOCCNNVOSSSTTTFNRNT--
isotig36180   MOFIYV--ALALAASVM-------ANPV------RLAATVCP--LGGTPOCCDVATGGIAIGSS-----
isotig38304   MQYSII--ALAVATLAL-------AAPDARRRGRKVGYPYCP--FNSTPACCQTTVSGEAENCT-----

cons          * :      .:: : .        . *             *         **     .

isotig36680   -----------------------V---NILGSOCS-----NN------GNAYCCKNSANGVLINIONCA
isotig36476   SIESLFP--------DFOANT--N---TPVGLACSPIIQDLSGGAOCNOOSVCCNEGHFSGL-FSLNCI
isotig36736   YLYLLFLGFTPAQIVALOTAOTNN---NGVAVVCS-----GSGSVSCNOONVCCNGSHFSGL-NVOACN
isotig36180   ----------------CVDPKKTPKNVAKFAKICL-----HTGK----LTAFCCTTAAVG---GLETCT
isotig38304   ----------------LVGNQNSA---KGLDKRCA-----QTGR-----TAYCCPEFSSG---TGEGCT

cons                                       .   *       .          **     .       *

isotig36680   PIL---------
isotig36476   PFNGGSQSTGIF
isotig36736   AVNA--------
isotig36180   PVV---------
isotig38304   EVGGS-------

cons           .
```

Figure 1. T-COFFEE multiple sequence alignment of hydrophobins tig1-tig5. All protein sequences were aligned using the algorithm published by Notredame et al. 2000 and executed under http://tcoffee.crg.cat/ (Expresso alignement). The conserved residues are marked with an asteriks and include a proline and 8 cysteine residues. The cysteine residues 2/3 and 6/7 are directly adjacent to each other, a characteristic of all hydrophobins (Scholtmeijer et al 2001). tig1 (isotig36680), tig2 (isotig36476), tig3 (isotig36736), tig4 (isotig36180), tig5 (isotig38304).

Aligning these sequences to each other (figure 1) and to *HYD1-HYD8* from *Tricholoma vaccinum* (figure 2) shows the presence of eight conserved cysteine residues. The location of the cysteine residues within the sequence as well as the order of these residues (the adjacency of residue 2/3 and 6/7) abides by the consensus of the hydrophobin family (Pfam pf01185). The overall length of the identified protein sequence shared the typical size of hydrophobins of about 100 aa (101 aa to 135 aa, well within the range of hydrophobins reported for *T. vaccinum* (*HYD1* 108 aa KJ507742– *HYD7* 140aa KJ507748) (Sammer et al., 2016). As expected, all hydrophobins showed N-terminal secretion signals (between 16aa and 20aa in length) and consensus sequences for signal sequence peptidases (SignalP www.cbs.dtu.dk/services/SignalP/).



MSA

The multiple sequence alignment result as produced by T-coffee.

```
T-COFFEE, Version_11.00.8cbe486 (2014-08-12 22:05:29 - Revision 8cbe486 - Build 477)
Cedric Notredame
SCORE=849
*
 BAD AVG GOOD
*
tr|A0A024BKG7|A   :  84
tr|A0A024BLL7|A   :  84
tr|A0A024BLM3|A   :  84
tr|A0A024BL34|A   :  83
tr|A0A024BKP2|A   :  82
tr|A0A024BKP7|A   :  81
tr|A0A024BLI5|A   :  82
tr|A0A024BLI8|A   :  82
Tig2              :  76
Tig3              :  70
cons              :  84

tr|A0A024BKG7|A   MFSKVVFFV-AAMAAFVAATPIPGT----------SGQCNTGPVQCCNSVYQSQSAES-----
tr|A0A024BLL7|A   MFSKVVVFV-AALAAFVAASPVPGA----------DSQCNTGPVQCCNSVYQSQSQEG-----
tr|A0A024BLM3|A   MFSKVVVFV-AALAAFVAASPVPSA----------DSQCNTGPVQCCNSVYQSQSQEA-----
tr|A0A024BL34|A   MFSKIALFV-ATVVAFVAATPIPDAA--------SSPQCNTGPIQCCQSVYQSQTTSH-----
tr|A0A024BKP2|A   MFSTLAFFLTAFLALFATALPTPGGA-------PALSQCNTGPIQCCTSTHQSNSTAG-----
tr|A0A024BKP7|A   MFSKVALFAIAFFVLLVSGGPIPDSPA-----PTSLTQCNAGPITCCSSVHQSDSAAG-----
tr|A0A024BLI5|A   MFSKVILFAIAFFVLFVNGGPIPGTPTPPAFAPGSVNQCNAGPITCCSSTHQSNSAAG-----
tr|A0A024BLI8|A   MFSKVALFAVATFALLVAGGPIPQAGS------GSINQCNNGQMLCCSEVNKSNTKAG-----
Tig2              MFSRIAVV-IALVATVMARPQIAGP----------VSSCNTGDLQCCNQVQSSKTVDFQKVAG
Tig3              MRFSVALFSLVTLALGVAATPEPSPEAANNF-ELVKRQCGGGSVQCCNNVQSSSTTTFNRNTY

cons              *   : .. . ..        .              .*. * : ** .. .*.:


tr|A0A024BKG7|A   ALIASIVGLNLQG--------ITGSIGTQCSPISAV-------GVGSNHSQQPVCCENNNYNG
tr|A0A024BLL7|A   SLLASLVGANVQD--------ANLMAGIQCSPITAIG-----LGSGSKCSQQPVCCSGNHMNG
tr|A0A024BLM3|A   SLLASLVGANLQG--------ANVMAGIQCTPITVLG-----LAGASKCSQQPVCCSENHMNG
tr|A0A024BL34|A   SILANLVGLDVQS--------LTASIGTQCSPLTVGG-----LAAGAKCSQQPVCCSGNNFSG
tr|A0A024BKP2|A   AFLLGAVGIPIQN--------LVTSVGFDCSPITLG-IVGGAMGSGAKCASQPVCCNQAYFSG
tr|A0A024BKP7|A   AGLLALVSAPLQN--------VVTTVGFDCTPITLGGPIGGAAASGAKCATQPVCCDQSFLSG
tr|A0A024BLI5|A   AGLLGLVGIPLQN--------AVTTVGFNCNPITSGGPVGGAAASGAKCASQPVCCDQLLMSG
tr|A0A024BLI8|A   ADALALASVPVQN--------AITTVGSNCNPITSGGPMGGAAASGAKCAAQPVCCNQSYLSG
Tig2              SSIESLFPDFQAN--------TNTPVGLACSPIIQD------LSGGAQCNQQSVCCNEGHFSG
Tig3              LYLLFLGFTPAQIVALQTAQTNNNGVAVVCSG-----------SGSVSCNQQNVCCNGSHFSG

cons                                        .  *.            . . .   * ***.    .*


tr|A0A024BKG7|A   LIVVGCSPVNL--------
tr|A0A024BLL7|A   LVVVGCSPVNL--------
tr|A0A024BLM3|A   LVVVGCSPVGL--------
tr|A0A024BL34|A   LIVVGCSPINL--------
tr|A0A024BKP2|A   LISNGCDPINIQM------
tr|A0A024BKP7|A   LVGNACIPINMQI-----L
tr|A0A024BLI5|A   LVGNGCIPINMQI-----L
tr|A0A024BLI8|A   LVGNSCIPINEQI-----L
Tig2              LFSLNCIPFNGGSQSTGIF
Tig3              LNVQACNAVNA--------

cons              *    * ...
```

Figure 2. T-COFFEE alignment between tig2 and tig3 and the cluster corresponding to Hyd1-8 from *Tricholoma vaccinum*. The protein sequences of the reported hydrophobins and hydrophobins found in Tricholoma vaccinum were aligned using the T-COFFEE alignement (Notredame et al. 2000). Conserved residues are marked with an asteriks. The cysteine residues present in the conserved domain structure of hydrophobins (pf01185) are present in all aligned sequences (http://pfam.xfam.org/family/pf01185).

The length of the hydrophobin gene sequences varied between 276bps (tig1) and 378bps (tig3) with an average GC content of 52,92%. As the underlying sequences were cDNA based, we could not identify any intron sequences. Based on the consensus sequence for N-linked glycosylation of Asn-Xaa-Ser/Thr (where Xaa is not Pro) we could not identify glycosylation sites in the five hydrophobin coding sequences (http://www.cbs.dtu.dk/services/NetNGlyc/). Within the phylogenetic tree tig1 showed clustering with the hyd3 protein from *Gibberella moniliformis* (Q6YF30) (Jurgenson et al., 2002) and with the *eas* protein from *Neurospora crassa*

(Q04561) (Winefield et al., 2007) while tig4 and tig 5 both located in close proximity to the hyd4 protein from *Gibberella moniliformis* (Q6YF29) (figure 3). Interestingly, both protein sequences of tig2 and tig3 occurred within the gene cluster of the hyd genes (hyd1-hyd9) of *Tricholoma vaccinum* (Wagner et al., 2015). Considering that tig1 as well as tig4 and tig5 cluster within hyd proteins of *Gibberella moniliformis*, while tig2 and tig3 occurred in close proximity to the hyd genes of *Tricholoma vaccinum* (figure 3) we speculate, that the origin of the five hydrophobin genes is of mixed organismal descendance, potentially from at least 2 species which co-occurred with the plant material. As an ascomycete *Gibberella moniliformis* shows distinct evolutionary distance to the basidiomycete *Tricholoma vaccinum* and therefore also proteins clustering with one or the other could indicate this evolutionary distance (Heilmann-Clausen et al., 2017; Leslie et al., 2004). This might indicate the presence of two organisms contributing their hydrophobin genes to the transcriptome sequence. *Gibberella moniliformis* is described as a facultative endophyte which can exist in biotrophic endophytic association with maize as well as saprophytically. On the other hand, *Tricholoma vaccinum* is a widely spread fungus which grows in a mycorrhizal association with pine trees. The presence of genes from other species than *P. aquilinum* could either have originated from a surface contamination or an endophytic organism associated with the fern. Due to the biology of *Tricholoma vaccinum* and its relatives, we speculate that tig2 and tig3 might therefore originate from a fungal species found on the plant surface at the time of collection. Petrini et al. have identified fungal endophytes of the bracken fern (*Pteridium aquilinum*) in dependence of the season, with harvest points in autumn and spring. In spring samples, they were able to show the presence of *Fusarium avenaceum,* a relative of *Gibberella moniliformis* and also a member of the *Nectriaceae* family. They were however not able to identify the presence of *F. avenaceum* in autumn samples of *P. aquilinum* (Fisher and Petrin, 1992). The sampling of fern specimen used for the generation of the transcriptome were also performed in spring. Sequence comparison of all five hydrophobin sequences via BLASTp did not result in any positive hits for *Fusarium avenaceum.* We therefore speculate, that tig1, tig4 and tig5 could be derived from a *Nectriaceae* family member. However, Blast searches are further complicated because except for the preserved cysteine pattern, the amino acid sequences are diverse, even among different hydrophobins from the same organism (Wessels, 2000). Interestingly, when comparing plant materials collected from spring samples of three consecutive years in close geographical proximity, we were only able to identify the gene of tig5 in gDNA samples (supporting information S1) of fern material of different organs. This repeated finding points into the direction of a certain degree of specificity in the interaction between the fungal species and the fern. However, the exact species of origin for the five hydrophobin sequences could at this point not be clarified.

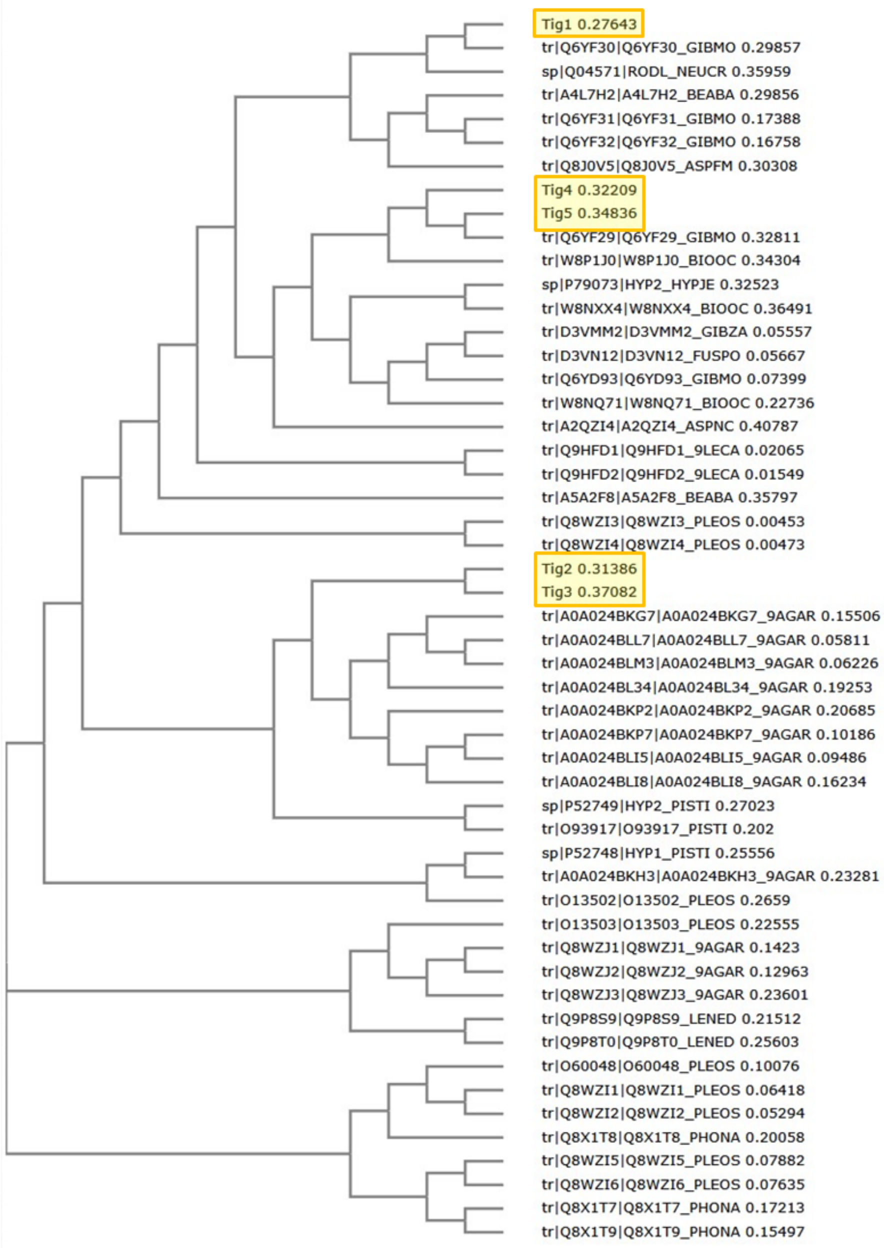

Tig1 0.27643
tr|Q6YF30|Q6YF30_GIBMO 0.29857
sp|Q04571|RODL_NEUCR 0.35959
tr|A4L7H2|A4L7H2_BEABA 0.29856
tr|Q6YF31|Q6YF31_GIBMO 0.17388
tr|Q6YF32|Q6YF32_GIBMO 0.16758
tr|Q8J0V5|Q8J0V5_ASPFM 0.30308
Tig4 0.32209
Tig5 0.34836
tr|Q6YF29|Q6YF29_GIBMO 0.32811
tr|W8P1J0|W8P1J0_BIOOC 0.34304
sp|P79073|HYP2_HYPJE 0.32523
tr|W8NXX4|W8NXX4_BIOOC 0.36491
tr|D3VMM2|D3VMM2_GIBZA 0.05557
tr|D3VN12|D3VN12_FUSPO 0.05667
tr|Q6YD93|Q6YD93_GIBMO 0.07399
tr|W8NQ71|W8NQ71_BIOOC 0.22736
tr|A2QZI4|A2QZI4_ASPNC 0.40787
tr|Q9HFD1|Q9HFD1_9LECA 0.02065
tr|Q9HFD2|Q9HFD2_9LECA 0.01549
tr|A5A2F8|A5A2F8_BEABA 0.35797
tr|Q8WZI3|Q8WZI3_PLEOS 0.00453
tr|Q8WZI4|Q8WZI4_PLEOS 0.00473
Tig2 0.31386
Tig3 0.37082
tr|A0A024BKG7|A0A024BKG7_9AGAR 0.15506
tr|A0A024BLL7|A0A024BLL7_9AGAR 0.05811
tr|A0A024BLM3|A0A024BLM3_9AGAR 0.06226
tr|A0A024BL34|A0A024BL34_9AGAR 0.19253
tr|A0A024BKP2|A0A024BKP2_9AGAR 0.20685
tr|A0A024BKP7|A0A024BKP7_9AGAR 0.10186
tr|A0A024BLI5|A0A024BLI5_9AGAR 0.09486
tr|A0A024BLI8|A0A024BLI8_9AGAR 0.16234
sp|P52749|HYP2_PISTI 0.27023
tr|O93917|O93917_PISTI 0.202
sp|P52748|HYP1_PISTI 0.25556
tr|A0A024BKH3|A0A024BKH3_9AGAR 0.23281
tr|O13502|O13502_PLEOS 0.2659
tr|O13503|O13503_PLEOS 0.22555
tr|Q8WZJ1|Q8WZJ1_9AGAR 0.1423
tr|Q8WZJ2|Q8WZJ2_9AGAR 0.12963
tr|Q8WZJ3|Q8WZJ3_9AGAR 0.23601
tr|Q9P8S9|Q9P8S9_LENED 0.21512
tr|Q9P8T0|Q9P8T0_LENED 0.25603
tr|O60048|O60048_PLEOS 0.10076
tr|Q8WZI1|Q8WZI1_PLEOS 0.06418
tr|Q8WZI2|Q8WZI2_PLEOS 0.05294
tr|Q8X1T8|Q8X1T8_PHONA 0.20058
tr|Q8WZI5|Q8WZI5_PLEOS 0.07882
tr|Q8WZI6|Q8WZI6_PLEOS 0.07635
tr|Q8X1T7|Q8X1T7_PHONA 0.17213
tr|Q8X1T9|Q8X1T9_PHONA 0.15497

Figure 3. Phylogenetic tree of hydrophobin protein sequences reported in this publication and Uniprot. Yellow boxes highlight the position of the identified proteins tig1-tig5. The phylogenetic tree was generated using the ClustalW2 algorithm at http://www.ebi.ac.uk/Tools/services/web/toolform.ebi (gap exclusion: off).

Based on their hydropathy profile, their solubility in solvents and the spacing between conserved cysteine residues, hydrophobins can conventionally be grouped into two classes -

class I and class II (Seidl-Seiboth et al., 2011). Generally, there exists more sequence variation and length in class I compared to class II hydrophobins. Looking at the overall sequence hydrophobicity, sequences after the first cysteine are higher in hydrophobicity in class Ib in comparison to class Ia and class II (Markus B. Linder et al., 2005). More specifically, the average of hydropathy is elevated in case of class I proteins and typically shows values of 0,5 and above. Based on the hydropathy plots in figure 5, we can observe a rather high overall hydrophobicity of tig1 which would indicate the presence of a class Ib hydrophobin (Basidiomycetes). The comparison with conserved residues reported by Linder et al. (Linder et al., 2005) shows the presence of amino acid residues conserved for Class Ia (Ascomycetes). When comparing the distance between the conserved cysteine residues (figure 4) to the consensus for class I and class II a mixed picture emerges. The hydrophobins tig2 and tig3 show a low overall hydrophobicity score (0,19 and 0,26, respectively) while possessing a cysteine distance pattern conserved for class I hydrophobins. In addition, the presence of conserved residues indicates the presence of a Class Ib (Basidiomycetes) hydrophobin. For tig4 the overall hydrophobicity after the first Cys is rather high (0,62) (figure 6) and would indicate a class I hydrophobin, however the conserved distance as well as the location of conserved residues would indicate the classification as a Class II hydrophobin. For tig5 both classification criteria (spacing and hydropathy) suggest the presence of a class II hydrophobin (figure 5, figure 6). Here, also the pattern of conserved amino acid residues clearly points into the direction of a Class II hydrophobin.

Class 1:	**C**	X_{5-8}	**CC**	X_{17-39}	**C**	X_{8-23}	**C**	X_{5-6}	**CC**	X_{6-18}	**C**	X_{2-13}
Class 2:	**C**	X_{9-10}	**CC**	X_{11}	**C**	X_{16}	**C**	X_{6-9}	**CC**	X_{10}	**C**	X_{3-7}
Tig1:		9		8		8		7		13		4
Tig2:		6		37		12		5		12		13
Tig3:		6		45		7		5		12		5
Tig4:		7		12		16		9		10		4
Tig5:		7		10		15		8		10		6

Figure 4. Distribution of cysteines present in hydrophobin tig1-tig5, compared with a commonly accepted consensus for class I and class II hydrophobins (Sammer et al. 2016). The distance between the conserved cysteine residues commonly found in class1 and class 2 is represented. Below, the distance values for hydrophobins tig1-tig5 is given. Based on this consensus sequence tig1 can either be classified as class 1 or class 2, while tig2 and tig3 can be grouped into class 1. The hydrophobin sequences termed tig4 and tig5 can be classified into class 2.

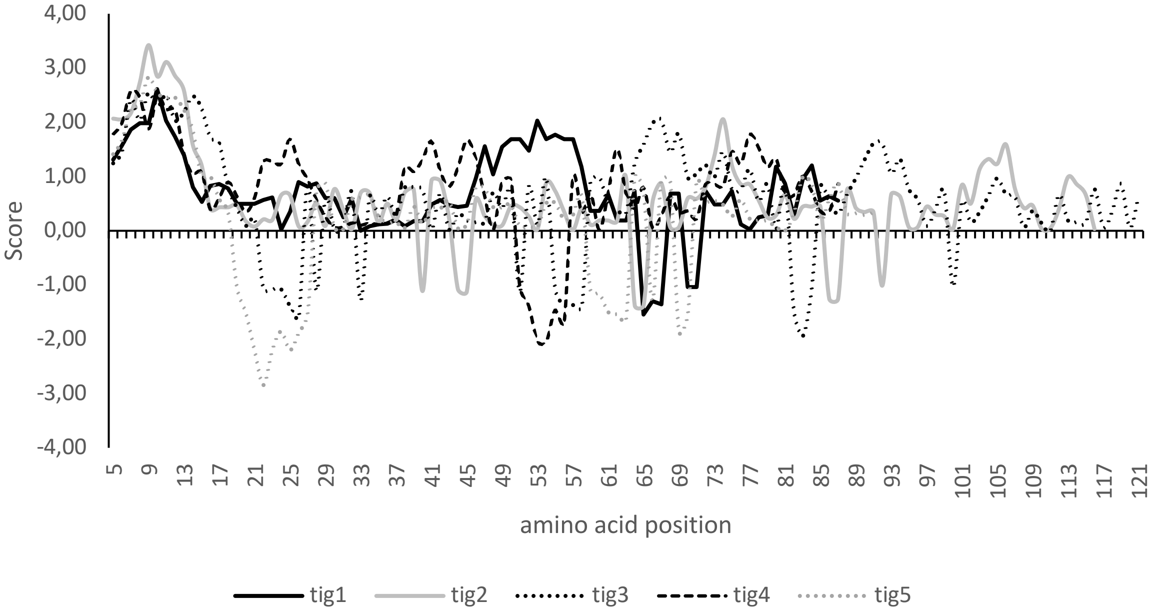
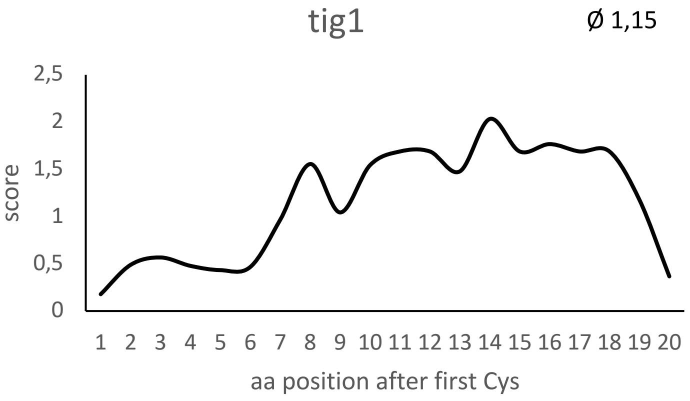
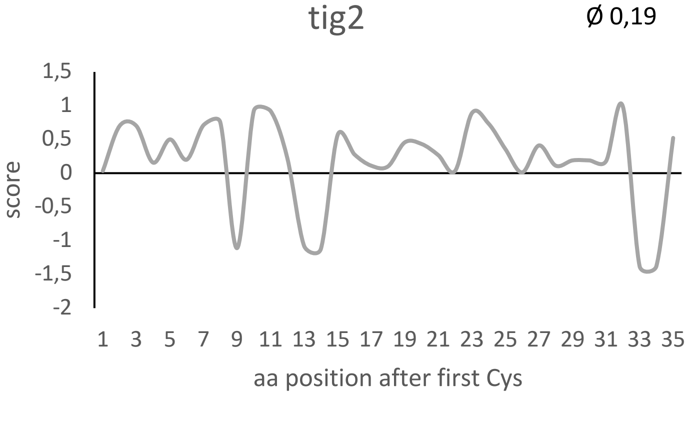


Figure 5. Hydropathy plot of entire coding sequence of hydrophobins tig1-tig5. The algorithm after Kyte & Doolittle was employed for the calcuation of hydrophobicity. The Kyte & Doolittle score of each protein is reported and ranged from 3,62 to -2,84. The x-axis shows the amino acid position of the respective hydrophobin. The tool was accessed under https://web.expasy.org/protscale/.





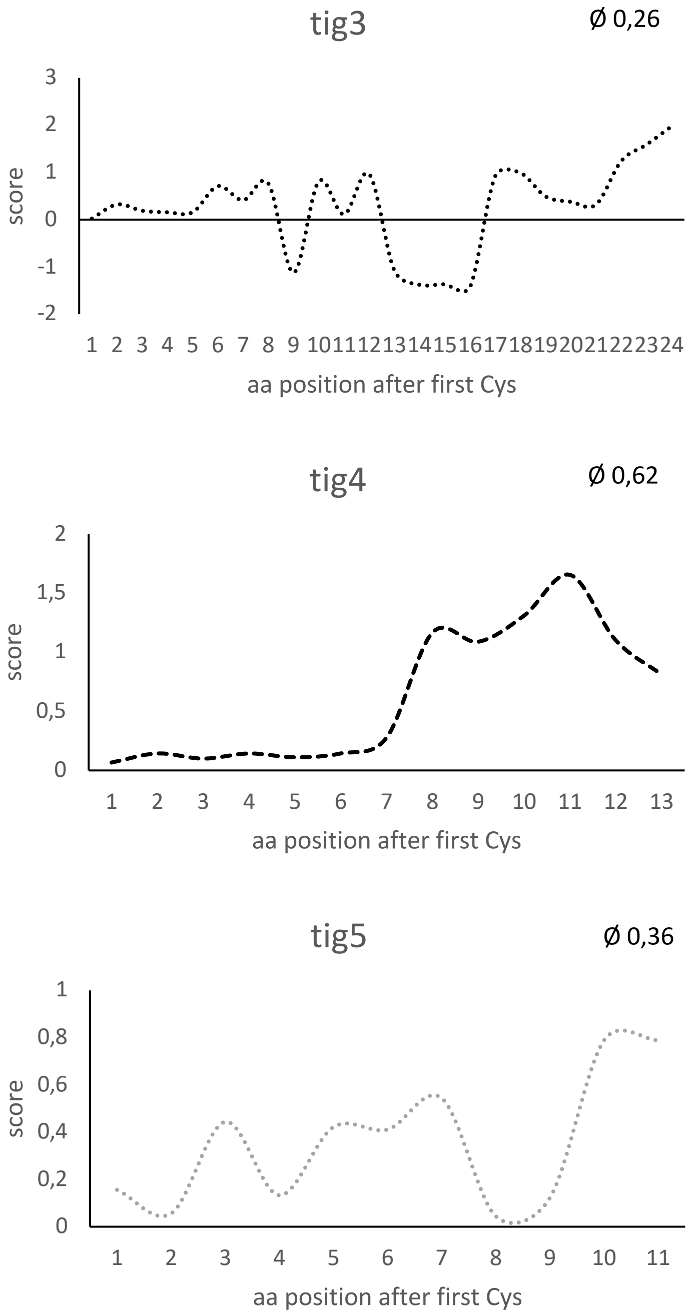


Figure 6. Hydropathy plot of amino acids between cys1 and cys2/3 of hydrophobins tig1-tig5. The algorithm after Kyte & Doolittle was employed for the calcuation of hydrophobicity. The Kyte & Doolittle score of each amino acid sequence is reported. The x-axis shows the amino acid position of the respective hydrophobin. The upper right corner shows the average value. The tool was accessed under https://web.expasy.org/protscale/.

However, a more definitive classification would need to be performed on the basis of rodlet formation and aggregation behavior in solvents (Wösten et al., 1994). To test the surface activity of the newly discovered hydrophobin sequences we generated expression vectors for production in the methylotrophic yeast species *Komagataella phaffii*. The five genes were codon optimized to adapt them to the *K. phaffii* codon usage and ordered as synthetic gene fragments. For the expression we chose a derepressed and methanol inducible promoter of *K. phaffii*, the promoter of the *CTA1* gene (Vogl et al., 2015). As a mode of production, we tested both the native hydrophobin secretion signal as well as the *S. cerevisiae* alpha mating factor pre-pro leader sequence cloned in frame with the respective hydrophobin sequence.

Interestingly, the *Ashbya gossypii TEF1* promoter driving the expression of the zeocin selection marker gene (*Sh ble* gene) showed functionable expression also in *E. coli* as no other bacterial promoter was necessary for selection of transformants under zeocin selection conditions. As early as 1989 Antonucci and co-workers showed expression of luciferase from different eukaryotic and viral promoters in *E. coli* (Antonucci et al., 1989). A similar result was obtained during the investigation of plant promoter sequences. In a study of plant promoters about 50% of all tested sequences showed activity also in *E. coli* (Jacob et al., 2002). Likewise, the activity of yeast promoter sequences in *E. coli* was shown by inserting random yeast sequences in a promoterless luciferase vector (Lewin et al., 2004b). It is therefore likely, that the *Ashbya gossypii TEF1* promoter also shows expression levels in *E. coli* conducive to successful selection under zeocin supplementation conditions.

The expression vectors for each hydrophobin gene were transformed into *K. phaffii* CBS7435 mutS (Sturmberger et al., 2016a), induced under standard minimal media and methanol conditions and cell supernatants harvested at the end of the cultivation. Dot-Blot analyses were conducted, and we could observe successful expression for all five proteins, however with different efficiencies (figure 7). Based on dot-blot results we have observed successful expression of both hydrophobins tig1 and tig2 albeit dependent on the employed secretion leader. While tig1 showed higher expression levels with the alpha mating factor pre-pro leader sequence in comparison to its native secretion leader, the expression results for hydrophobin tig2 resulted in opposite outcomes. The native secretion signal of tig2 resulted in higher secretion levels compared to the alpha mating factor pre-pro leader sequence (figure 7). For hydrophobins tig3, tig4 and tig5 we could only detect weak secretion into the culture supernatant however we were not able observe any differences between the protein’s native secretion signal and the alpha mating factor pre-pro leader. In comparison to tig1 and tig2, the expression levels are considerably lower with both secretion leaders.

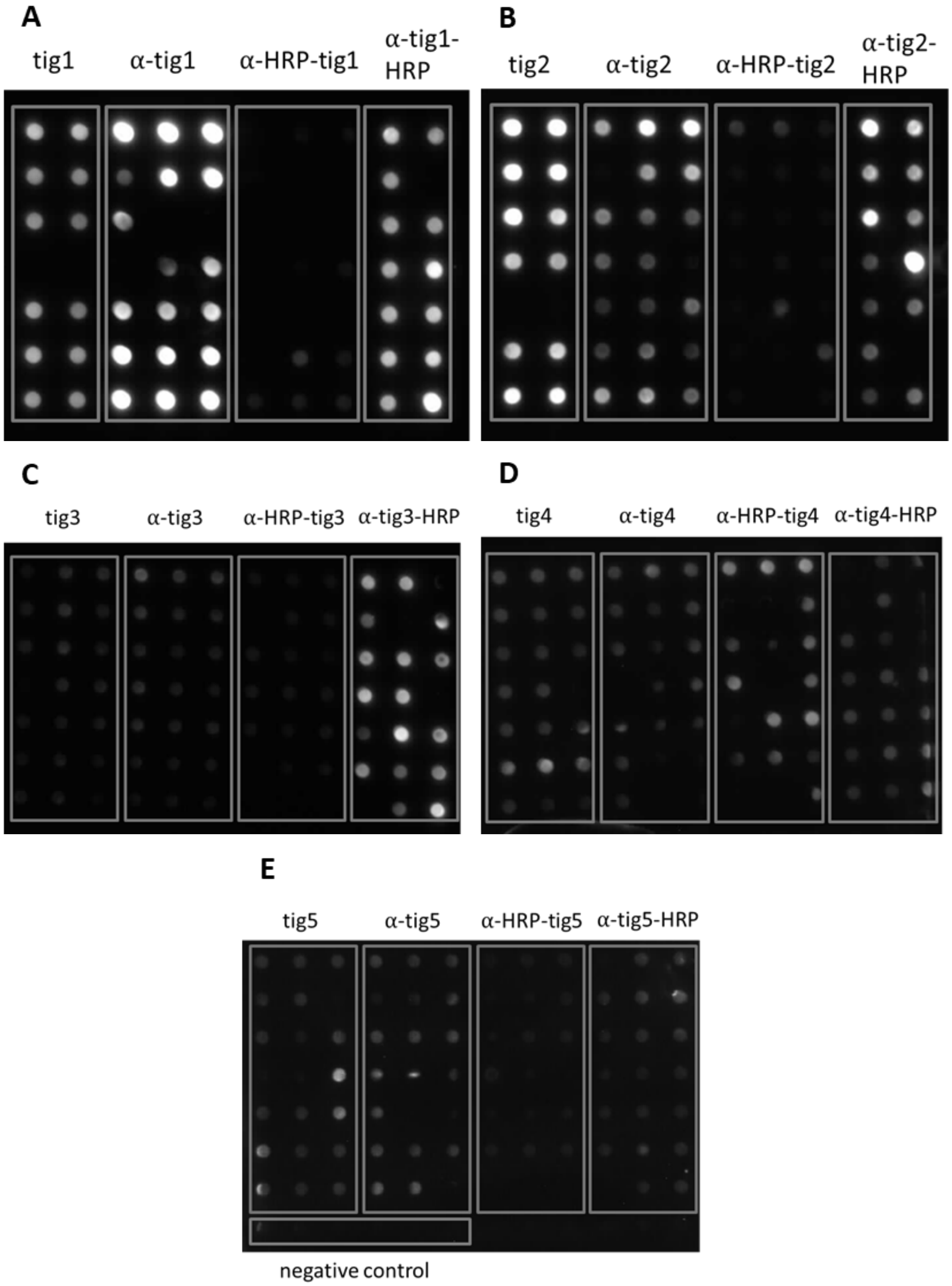


Figure 7. Dot Blot of hydrophobin expression strain supernatants. The hydrophobins were expressed under the regulation of the *K. phaffii* P_{CTA1} and methanol induction either with their native secretion signal, with an α mating factor pre-pro leader or as a fusion protein either with a N-terminally tagged HRP or a C-terminally tagged protein. The isoform of horse radish peroxidase chosen was the C1A (Krainer et al.) and linked by a (GGGGS)3 linker. A hexa-histidine tag was used for verification of expression in all experiments, in combination with a mouse anti-his and a goat anti-mouse Fc HRP antibody. As a negative control *K. phaffii* supernatant was used.

The collected culture supernatants were concentrated to an average of 50µg/mL and subjected to water contact angle measurements to assess the surface activity of hydrophobin candidates. Figure 8 shows the surface activity of hydrophobins on glass and polyethylenterephtalate (PET) surfaces. No shift in water contact angle could be observed for either of the five hydrophobins on polyethylen terephthalat. However, the angle shift and therefore the surface activity on glass was significantly increased for four hydrophobins. Hydrophobin tig1 showed the highest surface acitvity with almost 60 degrees (58°). Tig2, tig3 and tig4 all showed moderate angle shifts in comparison to tig1 of around 40 degress above the blank. For hydrophobin tig5 we were not able to detect any surface acitivtiy either on glass or on PET surfaces. It seems therefore, that four hydrophobins show a clear tendency for surface activity on glass. As we could prove a conserved hydrophobin domain with correctly positioned cysteine residues and successful expression of hydrophobin tig5 we speculate that the natural activity could be on a substrate with a different hydrophobicity pattern other than glass or PET. Potentially, the protein might not be active at all.

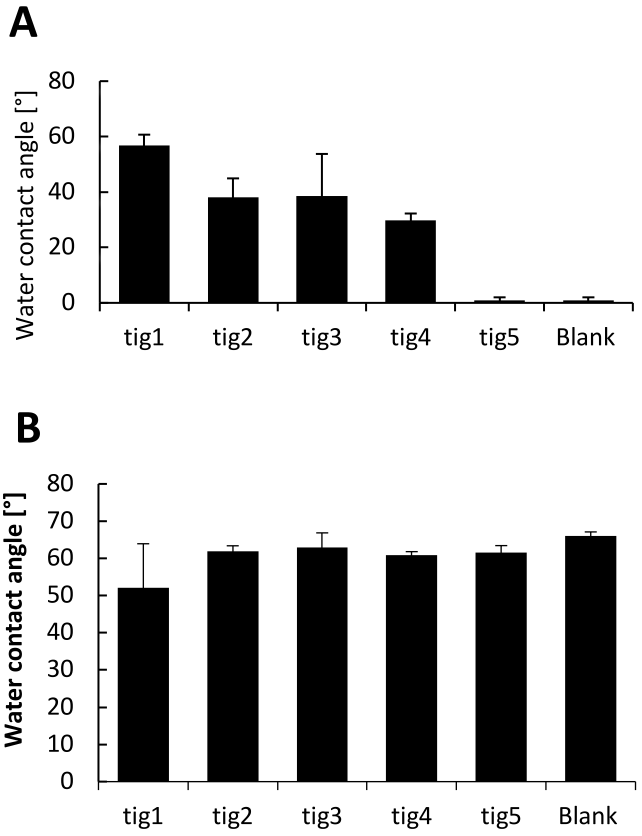
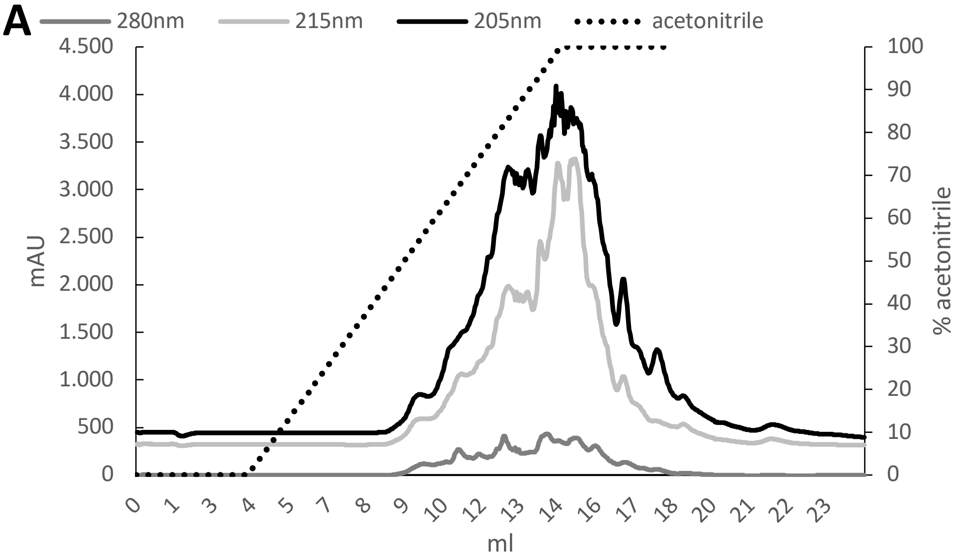
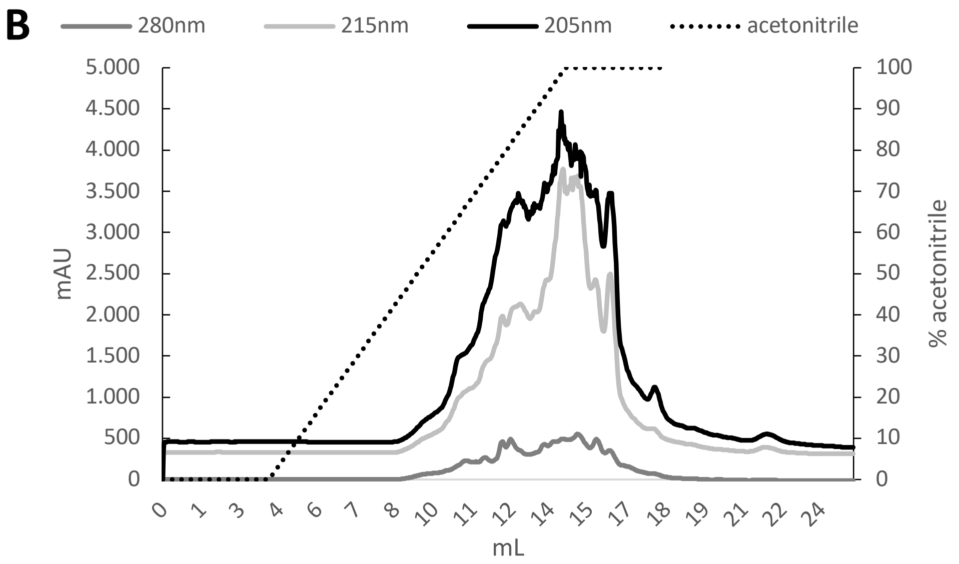


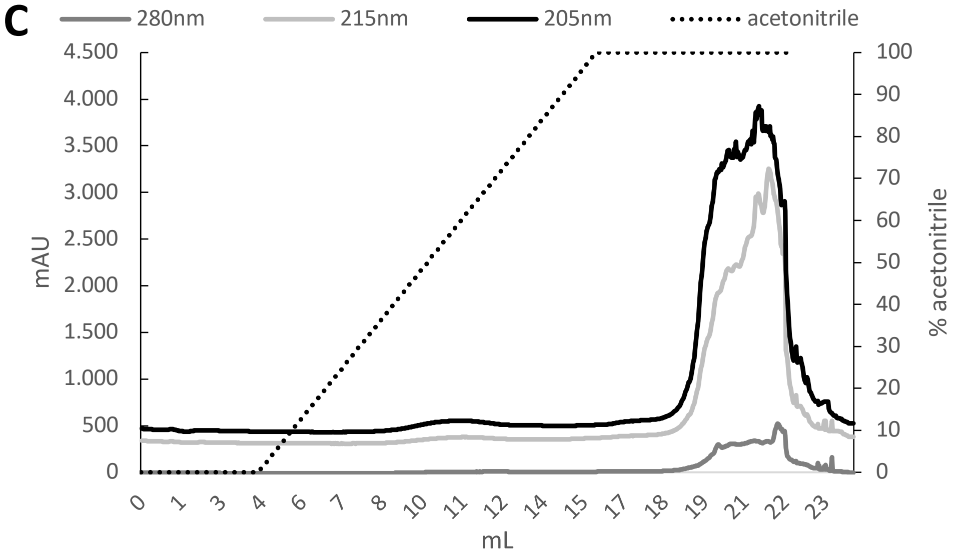
Fig. 8. WCA measurements of concentrated culture supernatants from hydrophobin expression strains. Surface activity of hydrophobins tig1, tig2, ig3, tig4 and tig5 expressed under the regulation of the *K. phaffii* P_{CTA1} and methanol induced with an α mating factor pre-pro leader was determined. For dilutions and blanks 0.1M potassium phosphate buffer was employed. The water contact angle of the blank was set to zero and the values represent the difference in angles measured. All measurements were performed on glass (A) and PET (B) surface.

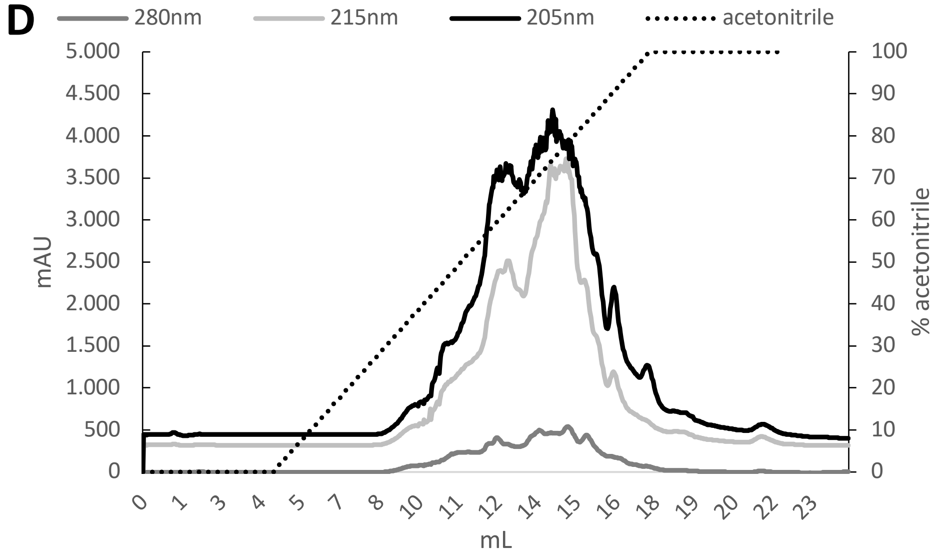
The possibility to purify the hydrophobins discovered in this study was tested as well. As IMAC purification via the expressed hexa-histidine tag did not yield any positive results (data not shown), the hydrophobic nature of this class of proteins provides a unique attribute for purification purposes. Reversed phase chromatography relies on the hydrophobic interaction between a protein and the column resin (Wang et al., 2010). By applying the filtered culture supernatant to an RPC column, the hydrophobin proteins are binding to the resin and were eluted by an acetonitrile gradient. Figure 9 shows the UV spectrophotometric curves at 280nm, 215nm and 205nm as well as the acetonitrile gradient for each hydrophobin purification. Fractions of 200µL were collected and analyzed again via dot-blot detection (figure 10). As can be seen for tig1, the protein was first eluted at an acetonitrile gradient of 80% up to 90%, with the majority between 84% and 89%. The protein containing fractions were subsequently pooled and buffer exchanged. We could therefore show the possibility to use reversed phase chromatography as means to purify the hydrophobins discovered in this study.

Since we were successful in showing hydrophobin surface activity on glass surfaces we speculated whether we might be able to use this activity to specifically coat glass surfaces. Although Linder and co-workers have shown that *Trichoderma reesei* HFBI and HFBII efficiently bind surfaces, these proteins did not cause immobilization as a fusion partner (Linder et al., 2002). In 2013 however, Espino-Rammer and colleagues were able to show two novel class II hydrophobins from *Trichoderma spp* that stimulate enzymatic hydrolysis of poly-ethylene terephthalate when expressed as fusion protein with a *Humicola insolens* cutinase by an isotherm-like adsorption behavior (Espino-Rammer et al., 2013). We initially started to design a fusion protein approach with the horse radish peroxidase (HRP) isoform C1A (Matsui, Krainer). Hydrophobin expression vectors with C-terminally and N-terminally labelled HRP were constructed and expressed in *K. phaffii*. Interestingly, the N-terminal fusion of HRP to the hydrophobins resulted in significantly lower secretion levels in comparison to the C-terminal fusion. In general, the fusion protein increased expression levels when benchmarked against fusionless hydrophobins (either with their native or alpha mating factor prepro leader sequence) (figure 7).




B
280nm
215nm
205nm
acetonitrile
mAU
5.000
4.500
4.000
3.500
3.000
2.500
2.000
1.500
1.000
500
0
% acetonitrile
100
90
80
70
60
50
40
30
20
10
0
0 1 3 4 6 7 8 10 11 12 14 15 17 18 19 21 22 24
mL


C
280nm
215nm
205nm
acetonitrile
mAU
4.500
4.000
3.500
3.000
2.500
2.000
1.500
1.000
500
0
% acetonitrile
100
90
80
70
60
50
40
30
20
10
0
0 1 3 4 6 7 8 10 11 12 14 15 17 18 19 21 22 23
mL


D
280nm
215nm
205nm
acetonitrile
mAU
5.000
4.500
4.000
3.500
3.000
2.500
2.000
1.500
1.000
500
0
% acetonitrile
100
90
80
70
60
50
40
30
20
10
0
0 1 3 4 5 7 8 10 11 12 14 15 16 18 19 20 22 23
mL

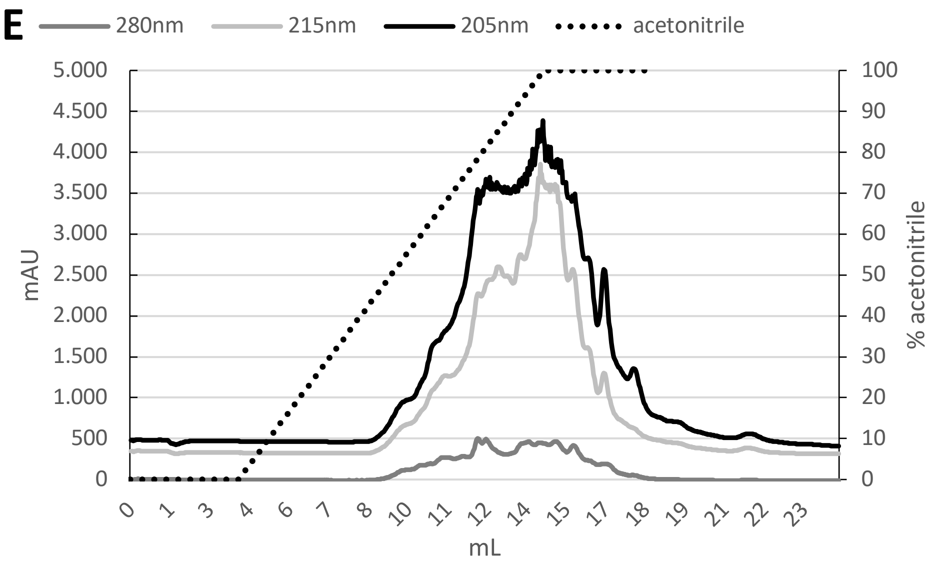


Figure 9. RP-chromatography profile of hydrophobin tig1 (A), tig2 (B), tig3 (C), tig4 (D) and tig5 (E) expressing *K. phaffii* culture supernatants. 50mL culture supernatant were loaded on an RPC column and an acetonitrile gradient ranging from 0% to 100% was applied. UV detection at three different wave lengths 280nm, 215nm and 205nm was performed. The acetonitrile gradient was set to spread over 15 column volumes.

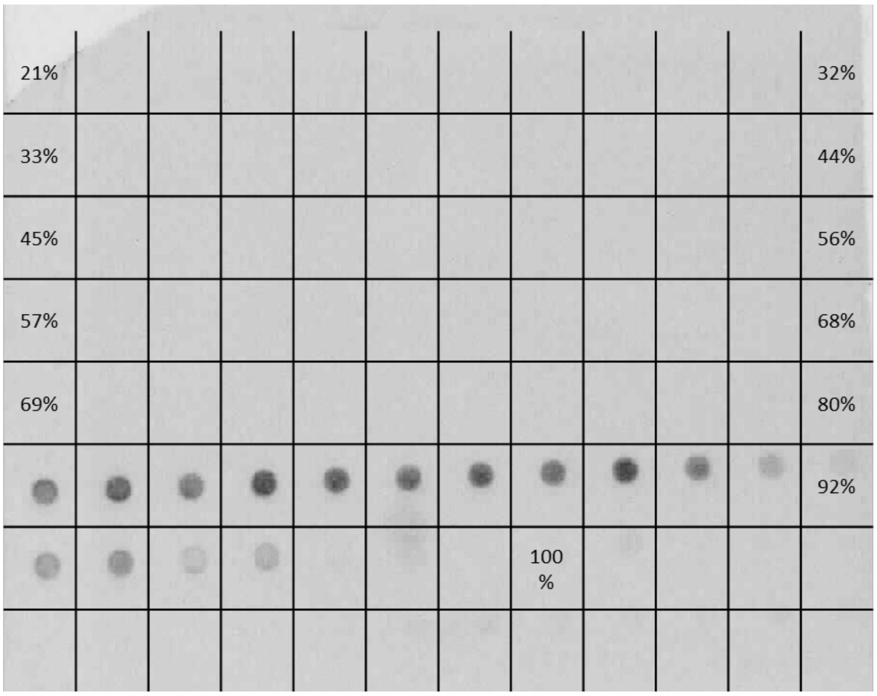


Figure 10. Dot blot analysis of eluted fractions from reversed chromatography of hydrophobin tig1 expression strain supernatants. The hydrophobin was eluted between an acetonitrile gradient of 80% and 90%, gradient can be seen in figure 6A. A hexa-histidine tag was used for verification of expression in all experiments, in combination with a mouse anti-his and a goat anti-mouse Fc HRP antibody.

As was also observed with prior experiments, tig1 and tig2 showed highest expression levels amongst all tested hydrophobins. Noteworthy, with tig3 we observed a considerable increase in secretion levels when expressed as a C-terminal fusion but not as an N-terminal one. The expression of tig4 and tig5 fusion proteins resulted in much lower yields in comparison to the rest of the proteins. Also, in these two cases the orientation of the HRP fusion partner did not influence secretion levels markedly. Based on the clonal variation occurring within expression cultures, a single clone with a comparably high expression level was chosen and re-cultivated at a larger scale (50mL culture). As a control we expressed a fusion and linker-less HRP C1A protein. A set amount of hydrophobin-HRP fusion and HRP protein (based on enzyme activity) was used and mixed with glass beads. The enzymes were allowed to bind to the glass surface and subsequently washed with buffer. The residual activity in the wash fraction, the bound fraction and the culture supernatant prior to binding were determined (figure 11). We were able to demonstrate a modest improvement in binding capacity of certain hydrophobin-HRP fusion proteins in comparison to a fusion-less HRP protein. The hydrophobin tig1 bound HRP showed a 38% improved binding capacity to glass surfaces compared to the HRP molecules alone. For the HRP fusion proteins with hydrophobins tig2-tig5 we could not detect any improvements in their glass-binding capacity. We were therefore able to show how a hydrophobin fusion protein can be employed to specifically coat glass surfaces. This technology could be used in a myriad of applications were a protein not naturally interacting with glass can be engineered to specifically bind this surface.

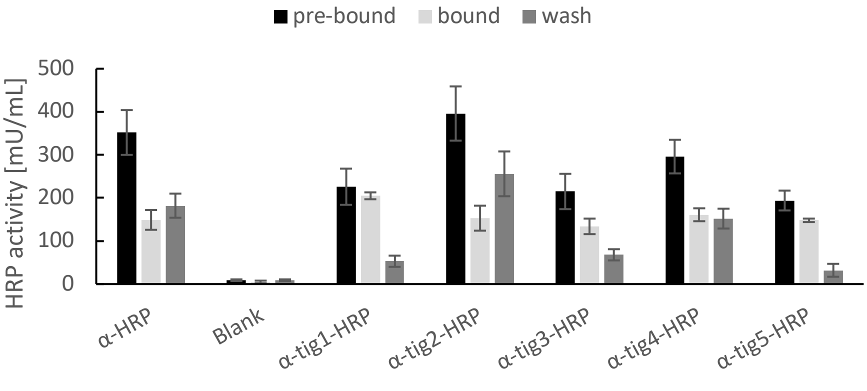
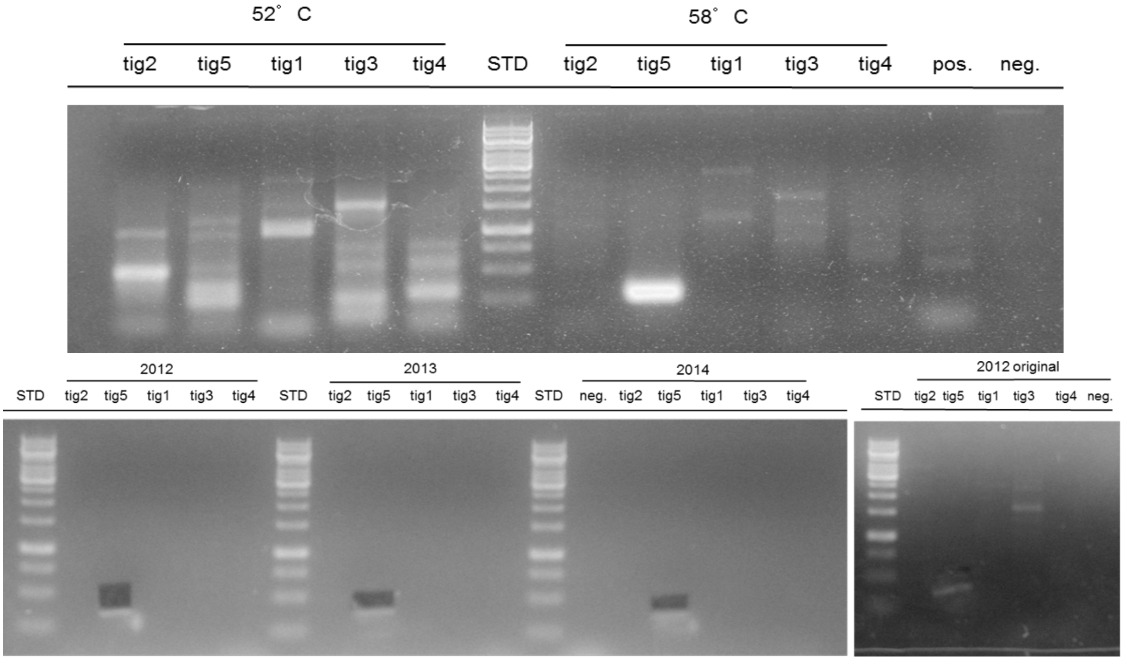


Figure 11. HRP activity of free (unbound) and glass bead bound hydrophobin-HRP fusion protein. 200-400 mU/mL of HRP activity of either HRPC1A itself or fusion protein with hydrophobin were mixed with 100mg of 0,1mm glas beads in HRP assay buffer. After a washing step with assay buffer, the wash fraction was stored, and beads allowed to settle. The HRP activity of the glass bead bound fraction and the wash fraction was determined. Hydrophobin mediated specific binding on glass could be shown for tig1-HRP fusion protein.


52° C
58° C
tig2 tig5 tig1 tig3 tig4 STD tig2 tig5 tig1 tig3 tig4 pos. neg.
2012
2013
2014
2012 original
STD tig2 tig5 tig1 tig3 tig4 STD tig2 tig5 tig1 tig3 tig4 STD neg. tig2 tig5 tig1 tig3 tig4
STD tig2 tig5 tig1 tig3 tig4 neg.

Supplementary file S1: PCR reaction of P. aquilinum gDNA from different sampling years. Fern plants from roughly the same geographical location were harvested around springtime in the years of 2012, 2013 and 2014. We isolated the gDNA and performed PCR reactions with primer binding in the original sequences discovered in the 2012 samples. Tig5 was the only gene sequence which yielded a positive PCR product in samples of all three sampling years.

Figures and Tables

Figure 1. T-COFFEE multiple sequence alignment of hydrophobins tig1-tig5.
Figure 2. Phylogenetic tree of hydrophobin protein sequences reported in this publication and Uniprot.
Figure 3. T-COFFEE alignment between tig2 and tig3 and the cluster corresponding to Hyd1-8 from *Tricholoma vaccinum*.
Figure 4. Distribution of cysteines present in hydrophobin tig1-tig5, compared with a commonly accepted consensus for class I and class II hydrophobins
Figure 5. Hydropathy plot of entire coding sequence of hydrophobins tig1-tig5.
Figure 6. Hydropathy plot of amino acids between cys1 and cys2/3 of hydrophobins tig1-tig5.
Figure 7. Dot Blot of hydrophobin expression strain supernatants.
Fig. 8. WCA measurements of concentrated culture supernatants from hydrophobin expression strains.
Figure 9. RP-chromatography profile of hydrophobin tig1 (A), tig2 (B), tig3 (C), tig4 (D) and tig5 (E) expressing *K. phaffii* culture supernatants.
Figure 10. Dot blot analysis of eluted fractions from reversed chromatography of hydrophobin tig1 expression strain supernatants.
Figure 11. HRP activity of free (unbound) and glass bead bound hydrophobin-HRP fusion protein.
Supplementary file S1. PCR reaction of P. aquilinum gDNA from different sampling years.

References

Antonucci, T.K., Wen, P., Rutter, W.J., 1989. Eukaryotic promoters drive gene expression in Escherichia coli. J. Biol. Chem. 264, 17656–9.

Camattari, A., Goh, A., Yip, L.Y., Hee, A., Tan, M., Ng, S.W., Tran, A., Liu, G., Liachko, I., Dunham, M.J., Rancati, G., 2016. Characterization of a panARS - based episomal vector in the methylotrophic yeast Pichia pastoris for recombinant protein production and synthetic biology applications. Microb. Cell Factories 1–11. https://doi.org/10.1186/s12934-016-0540-5

Clyne, R.K., Kelly, T.J., 1995. Genetic analysis of an ARS element from the fission yeast Schizosaccharomyces pombe. EMBO J. https://doi.org/10.1002/j.1460-2075.1995.tb00326.x

Collén, A., Persson, J., Linder, M., Nakari-Setälä, T., Penttilä, M., Tjerneld, F., Sivars, U., 2002. A novel two-step extraction method with detergent/polymer systems for primary recovery of the fusion protein endoglucanase I-hydrophobin I. Biochim. Biophys. Acta - Gen. Subj. 1569, 139–150. https://doi.org/10.1016/S0304-4165(01)00244-6

Cox, P.W., Hooley, P., 2009. Hydrophobins: New prospects for biotechnology. Fungal Biol. Rev. 23, 40–47. https://doi.org/10.1016/j.fbr.2009.09.001

Espino-Rammer, L., Ribitsch, D., Przylucka, A., Marold, A., Greimel, K.J., Acero, E.H., Guebitz, G.M., Kubicek, C.P., Druzhinina, I.S., 2013. Two novel class ii hydrophobins from Trichoderma spp. Stimulate enzymatic hydrolysis of poly(ethylene terephthalate) when expressed as fusion proteins. Appl. Environ. Microbiol. https://doi.org/10.1128/AEM.01132-13

Fisher, P., PetrinP, L., 1992. Fungal endophytes of bracken (Pteridium aquilinum), with some reflections on their use in biological control.

Heilmann-Clausen, J., Christensen, M., Frøslev, T.G., Kjøller, R., 2017. Taxonomy of Tricholoma in northern Europe based on ITS sequence data and morphological characters. Persoonia Mol. Phylogeny Evol. Fungi 38, 38–57. https://doi.org/10.3767/003158517X693174

Jacob, D., Lewin, A., Meister, B., Appel, B., 2002. Plant-specific promoter sequences carry elements that are recognised by the eubacterial transcription machinery. Transgenic Res. https://doi.org/10.1023/A:1015620016472

Jurgenson, J.E., Zeller, K.A., Leslie, J.F., 2002. Expanded genetic map of Gibberella moniliformis (Fusarium verticillioides). Appl. Environ. Microbiol. 68, 1972–1979. https://doi.org/10.1128/AEM.68.4.1972-1979.2002

Krainer, F.W., Pletzenauer, R., Rossetti, L., Herwig, C., Glieder, A., Spadiut, O., 2013. Purification and basic biochemical characterization of 19 recombinant plant peroxidase isoenzymes produced in Pichia pastoris q. PROTEIN Expr. Purif. https://doi.org/10.1016/j.pep.2013.12.003

Lahtinen, T., Linder, M.B., Nakari-Setälä, T., Oker-Blom, C., 2008. Hydrophobin (HFBI): A potential fusion partner for one-step purification of recombinant proteins from insect cells. Protein Expr. Purif. 59, 18–24. https://doi.org/10.1016/j.pep.2007.12.014

Lanfranchi, E., Köhler, E.-M., Darnhofer, B., Steiner, K., Birner-Gruenberger, R., Glieder, A., Winkler, M., 2015. Bioprospecting for Hydroxynitrile Lyases by Blue Native PAGE Coupled HCN Detection. Curr. Biotechnol. https://doi.org/10.2174/2211550104666150506225048

Lanfranchi, E., Pavkov-Keller, T., Koehler, E.-M., Diepold, M., Steiner, K., Darnhofer, B., Hartler, J., Van, T., Bergh, D., Joosten, H.-J., Gruber-Khadjawi, M., Thallinger, G.G., Birner-Gruenberger, R., Gruber, K., Winkler, M., Glieder, A., 2017. Enzyme discovery beyond homology: a unique hydroxynitrile lyase in the Bet v1 superfamily OPEN. https://doi.org/10.1038/srep46738

Łaźniewska, J., Macioszek, V.K., Kononowicz, A.K., 2012. Plant-fungus interface: The role of surface structures in plant resistance and susceptibility to pathogenic fungi. Physiol. Mol. Plant Pathol. https://doi.org/10.1016/j.pmpp.2012.01.004

Leslie, J.F., Zeller, K.A., Logrieco, A., Mulè, G., Moretti, A., Ritieni, A., 2004. Species Diversity of and Toxin Production by Gibberella fujikuroi Species Complex Strains Isolated from Native Prairie Grasses in Kansas. Appl. Environ. Microbiol. 70, 2254–2262. https://doi.org/10.1128/AEM.70.4.2254-2262.2004

Lewin, A., Tran, T.T., Jacob, D., Mayer, M., Freytag, B., Appel, B., 2004. Yeast DNA sequences initiating gene expression in Escherichia coli. Microbiol. Res. https://doi.org/10.1016/j.micres.2004.01.006

Liachko, I., Dunham, M.J., 2014. An autonomously replicating sequence for use in a wide range of budding yeasts. FEMS Yeast Res. 14, 364–367. https://doi.org/10.1111/1567-1364.12123

Lin-cereghino, J., Wong, W.W., Xiong, S., Giang, W., Linda, T., Vu, J., Johnson, S.D., Lin-cereghino, G.P., 2005. Condensed protocol for competent cell preparation and transformation of the methylotrophic yeast Pichia pastoris 38, 4–6.

Linder, M., Szilvay, G.R., Nakari-Setälä, T., Söderlund, H., Penttilä, M., 2002. Surface adhesion of fusion proteins containing the hydrophobins HFBI and HFBII from Trichoderma reesei. Protein Sci. Publ. Protein Soc. 11, 2257–2266. https://doi.org/10.1110/ps.0207902

Linder, Markus B, Szilvay, G.R., Nakari-Setälä, T., Penttilä, M.E., 2005. Hydrophobins: the protein-amphiphiles of filamentous fungi. FEMS Microbiol. Rev. 29, 877–96. https://doi.org/10.1016/j.femsre.2005.01.004

Linder, Markus B., Szilvay, G.R., Nakari-Setälä, T., Penttilä, M.E., 2005. Hydrophobins: The protein-amphiphiles of filamentous fungi. FEMS Microbiol. Rev. https://doi.org/10.1016/j.femsre.2005.01.004

Lugones, L.G., Bosscher, J. 5, Scholtmeyer, K., De Vries, O.M.H., Wessels, J.G.H., 1996. An abundant hydrophobin (ABHI) forms hydrophobic rodlet layers in Agarics bisporus fruiting bodies, Microbiology.

Näätsaari, L., Mistlberger, B., Ruth, C., Hajek, T., Hartner, F.S., Glieder, A., 2012. Deletion of the pichia pastoris ku70 homologue facilitates platform strain generation for gene expression and synthetic biology. PLoS ONE 7. https://doi.org/10.1371/journal.pone.0039720

Nicholson, R.L., Epstein, L., 1991. Adhesion of Fungi to the Plant Surface, in: The Fungal Spore and Disease Initiation in Plants and Animals. https://doi.org/10.1007/978-1-4899-2635-7_1

Palomo, J.M., Peñas, M.M., Fernández-Lorente, G., Mateo, C., Pisabarro, A.G., Fernández-Lafuente, R., Ramírez, L., Guisán, J.M., 2003. Solid-phase handling of hydrophobins: Immobilized hydrophobins as a new tool to study lipases. Biomacromolecules 4, 204–210. https://doi.org/10.1021/bm020071l

Sammer, D., Krause, K., Gube, M., Wagner, K., Kothe, E., 2016. Hydrophobins in the life cycle of the ectomycorrhizal basidiomycete Tricholoma vaccinum. PLoS ONE. https://doi.org/10.1371/journal.pone.0167773

Scholtmeijer, K., Wessels, J.G.H., Wösten, H. a. B., 2001. Fungal hydrophobins in medical and technical applications. Appl. Microbiol. Biotechnol. 56, 1–8. https://doi.org/10.1007/s002530100632

Seidl-Seiboth, V., Gruber, S., Sezerman, U., Schwecke, T., Albayrak, A., Neuhof, T., Von Döhren, H., Baker, S.E., Kubicek, C.P., 2011. Novel hydrophobins from trichoderma define a new hydrophobin subclass: Protein properties, evolution, regulation and processing. J. Mol. Evol. https://doi.org/10.1007/s00239-011-9438-3

Steiner, S, Philippsen, P., 1994. Sequence and promoter analysis of the highly expressed TEF gene of the filamentous fungus Ashbya gossypii. Mol. Gen. Genet. MGG 242, 263–71. https://doi.org/10.1007/bf00280415

Steiner, Sabine, Philippsen, P., 1994. Sequence and promoter analysis of the highly expressed TEF gene of the filamentous fungus Ashbya gossypii. MGG Mol. Gen. Genet. https://doi.org/10.1007/BF00280415

Stocker, G., Rieder, D., Trajanoski, Z., 2004. ClusterControl: A web interface for distributing and monitoring bioinformatics applications on a Linux cluster. Bioinformatics. https://doi.org/10.1093/bioinformatics/bth014

Sturmberger, L., Chappell, T., Geier, M., Krainer, F., Day, K.J., Vide, U., Trstenjak, S., Schiefer, A., Richardson, T., Soriaga, L., Darnhofer, B., Birner-Gruenberger, R., Glick, B.S., Tolstorukov, I., Cregg, J., Madden, K., Glieder, A., 2016. Refined Pichia pastoris reference genome sequence. J. Biotechnol. https://doi.org/10.1016/j.jbiotec.2016.04.023

Tucker, S.L., Talbot, N.J., 2001. S <scp>URFACE</scp> A <scp>TTACHMENT AND</scp> P <scp>RE</scp> -P <scp>ENETRATION</scp> S <scp>TAGE</scp> D <scp>EVELOPMENT BY</scp> P <scp>LANT</scp> P <scp>ATHOGENIC</scp> F <scp>UNGI</scp>. Annu. Rev. Phytopathol. https://doi.org/10.1146/annurev.phyto.39.1.385

Van der Vegt, W., Van der Mei, H.C., Wosten, H.A.B., Wessels, J.G.H., Busscher, H.J., 1996. A comparison of the surface activity of the fungal hydrophobin SC3p with those of other proteins. Biophys. Chem. 57, 253–260. https://doi.org/10.1016/0301-4622(95)00059-7

Vogl, T., Sturmberger, L., Kickenweiz, T., Wasmayer, R., Schmid, C., Hatzl, A.-M., Gerstmann, M.A., Pitzer, J., Wagner, M., Thallinger, G.G., Geier, M., Glieder, A., 2015. A toolbox of diverse promoters related to methanol utilization – functionally verified parts for heterologous pathway expression in Pichia pastoris. ACS Synth. Biol. acssynbio.5b00199. https://doi.org/10.1021/acssynbio.5b00199

Wagner, K., Org Linde, J.¨, Krause, K., Gube, M., Koestler, T., Sammer, D., Kniemeyer, O., Kothe, E., 2015. Tricholoma vaccinum host communication during ectomycorrhiza formation. FEMS Microbiol. Ecol. 91, 120. https://doi.org/10.1093/femsec/fiv120

Wang, Z., Feng, S., Huang, Y., Li, S., Xu, H., Zhang, X., Bai, Y., Qiao, M., 2010. Expression and characterization of a Grifola frondosa hydrophobin in Pichia pastoris. Protein Expr. Purif. 72, 19–25. https://doi.org/10.1016/j.pep.2010.03.017

Wessels, J.G.H., 2000. Hydrophobins, unique fungal proteins. Mycologist 14, 153–159. https://doi.org/10.1016/S0269-915X(00)80030-0

Winefield, R.D., Hilario, E., Beever, R.E., Haverkamp, R.G., Templeton, M.D., 2007. Hydrophobin genes and their expression in conidial and aconidial Neurospora species. Fungal Genet. Biol. FG B 44, 250–7. https://doi.org/10.1016/j.fgb.2006.11.008

Wösten, H. a, 2001. Hydrophobins: multipurpose proteins. Annu. Rev. Microbiol. 55, 625–646. https://doi.org/10.1146/annurev.micro.55.1.625

Wösten, H. a, de Vocht, M.L., 2000. Hydrophobins, the fungal coat unravelled. Biochim. Biophys. Acta 1469, 79–86.

Wösten, H.A., Schuren, F.H., Wessels, J.G., 1994. Interfacial self-assembly of a hydrophobin into an amphipathic protein membrane mediates fungal attachment to hydrophobic surfaces. EMBO J. 13, 5848–54.

Wösten, H.A.B., Richter, M., Willey, J.M., 1999a. Structural proteins involved in emergence of microbial aerial hyphae. Fungal Genet. Biol. https://doi.org/10.1006/fgbi.1999.1130

Wösten, H.A.B., Ruardy, T.G., van der Mei, H.C., Busscher, H.J., Wessels, J.G.H., 1995. Interfacial self-assembly of a Schizophyllum commune hydrophobin into an insoluble amphipathic protein membrane depends on surface hydrophobicity. Colloids Surf. B Biointerfaces 5, 189–195. https://doi.org/10.1016/0927-7765(94)01151-T

Wösten, H.A.B., Van Wetter, M.A., Lugones, L.G., Van der Mei, H.C., Busscher, H.J., Wessels, J.G.H., 1999b. How a fungus escapes the water to grow into the air. Curr. Biol. 9, 85–88. https://doi.org/10.1016/S0960-9822(99)80019-0

Review

Synergism of proteomics and mRNA sequencing for enzyme discovery

Lukas Sturmberger[a,1], Paal W. Wallace[a,b,c,1], Anton Glieder[a,d], Ruth Birner-Gruenberger[a,b,c]

[a]Austrian Center of Industrial Biotechnology (ACIB), Petersgasse 14, 8010 Graz, Austria
[b] Medical University of Graz, Institute of Pathology, Research Unit Functional Proteomics and Metabolic Pathways, Stiftingtalstrasse 24, 8010 Graz, Austria
[c] Omics Center Graz, BioTechMed-Graz, Stiftingtalstrasse 24, 8010 Graz, Austria
[d] TU Graz, Institute of Molecular Biotechnology, NAWI Graz, Petersgasse 14, 8010 Graz, Austria
[1] authors contributed equally

Abbreviations

mRNA	messenger ribonucleic acids
cDNA	complementary desoxyribonucleic acids
MALDI-TOF	matrix-assisted laser desorption/ionization followed by time of flight
LC–MS/MS	liquid chromatography tandem mass spectrometry
NCBI	National Center for Biotechnology Information
EMBL	European Molecular Biology Laboratory
DDBJ	DNA Data Bank of Japan
IINSDC	International Nucleotide Sequence Database Collaboration
JGI	Joint Genome Institute
rRNA	ribosomal ribonucleic acids
ESI	electrospray ionization
EST	expressed sequence tag
FDA	food and drug administration
SL	sophorolipid
SILAC	stable isotope labeling by amino acids in cell culture
TCA	tricarboxylic acid

Keywords

enzyme discovery, proteomics, transcriptomics, database, biotechnology

This manuscript was published as Sturmberger, L. *et al.* (2016) ‘Synergism of proteomics and mRNA sequencing for enzyme discovery’, *Journal of Biotechnology*. Elsevier B.V., 235, pp. 132–138. doi: 10.1016/j.jbiotec.2015.12.015.

Abstract

Enzyme catalyzed processes are increasingly complementing chemical manufacturing as new enzymes are being discovered. Although, many industrially applied biocatalysts have been identified by functional screenings technological advances in the omics fields have created a different path to access novelty. Here we describe how omics technologies, especially proteomics and transcriptomics, can complement each other in the aim of finding new enzymatic functions. Special emphasis is laid on how mRNA sequencing can improve proteomic experiments by allowing the generation of high-quality protein sequence databases, which subsequently facilitates protein identification.

Introduction

Enzymes have over the course of the last decade established themselves as useful alternatives to classical organo-and metallo-catalysis (Anastas and Warner, 1998; Bornscheuer et al., 2012; Illanes et al., 2012; Meyer, 2011). They are increasingly employed for the manufacturing of a diverse array of chemical products such as pharmaceuticals, agrochemicals, bulk chemicals or biofuels on preparative as well as industrial scale (Erickson et al., 2012). This focus also applies to biological degradation of products, such as xenobiotics, plastics, biomass and unwanted components in wastewater (Cammarota and Freire, 2006; Eberl et al., 2008; Janssen et al., 2005; Kullman and Matsumura, 1996; Müller et al., 2005; Rabinovich et al., 2004; Ribitsch et al., 2012). So far about 200 enzymes are industrially used (Li et al., 2012b), with the majority of those being hydrolytic enzymes and oxidoreductases (Faber, 2011). The importance of enzyme-catalyzed synthesis can also be visualized in economic terms. The global enzyme market in 2010 was worth 3.3 billion US dollars and is expected to reach 4.4 billion dollars by the end of 2015 (Li et al., 2012; Sarrouh et al., 2012). Due to the need for enzyme-catalyzed processes, the biotechnological sector is on a constant quest for new sources of enzymes. The majority of discovered and applied enzymes derive from a limited number of cultivable laboratory organisms, mainly of fungal or bacterial origin (Robertson and Steer, 2004). Estimates show that less than 1% of all microbes are readily cultivable due to unknown cultivation conditions, slow growth behavior or the absence of essential nutrients supplied by other organisms (Amann et al., 1995; Ekkers et al., 2012). This fact presents both a challenge and an opportunity for the discovery of novel enzymes and functions. Most industrially applied enzymes have been found by functional screenings of metagenomic and genomic libraries (Ferrer et al., 2009; Uchiyama and Miyazaki, 2009). The advantages of functional screens lie in the immediate identification of enzyme activity whereas methods based solely on sequence only give predictions about enzyme function derived from annotations. On the other hand, the complex makeup of eukaryotic genomes containing many layers of regulatory elements drastically increases library size, making functional expression and screening of these libraries very challenging. Even though functional screening of bacterial and to a lesser extent eukaryotic genomes and metagenomes have been most successful, mere in silico methods such as structure guided (Steinkellner et al., 2014) and sequence similarity-based approaches have also received attention in recent years. The latter approaches are limited to the discovery of new genes with high similarities to already deposited sequences, making it impossible to discover truly novel enzymatic functions. They nevertheless represent a powerful tool by eliminating cost-and labor-intensive wet-laboratory time (Behrens et al., 2011). Although genomic screenings have been successful in identifying new enzymes, recent technological advances in the field of transcriptomics and proteomics have made these attractive tools for further discoveries. A major disadvantage of metagenomic or genomic libraries is the high percentage of genomic DNA comprising of non-coding regions (Deutsch and Long, 1999), which need to be removed in order to produce functional proteins and which unnecessarily increase library size. In addition to problems related to incorrect positioning of the gene relative to its promoters, non-spliceable introns and different codon usage by different organisms, posttranslational modifications are challenging when expressing these genes in hosts other than the original organism (Behrens et al., 2011). By applying a transcriptomic or proteomic approach, some of these disadvantages are circumvented. Using transcriptomics instead of genomics the majority of non-coding DNA elements can be removed thereby reducing library size and avoiding non-functional genes arising due to incomplete splicing recognition. Furthermore, the positioning of cDNA transcript sequences in relation to core promoter regions allows for a higher likelihood of successful transcription and translation compared to the majority of sheared genomic DNA fragments. By using transcriptomics rather than genomics, it is also possible to investigate dynamic spatio-temporal gene expression patterns. Therefore, transcriptomics can capture differential gene expression arising due to changes in environmental factors such as the presence of certain chemicals or shifts in cultivation conditions. This dynamic can also be monitored using proteomics. Quantitative proteomics has an advantage over transcriptomics as mRNA abundance does not accurately reflect protein abundance (Gygi et al., 1999; Zhang et al., 2014) and only proteomics can capture posttranslational modifications.

Proteomic analysis moreover has the benefit of analysis directly performed on proteins expressed by the original organism. This circumvents problems related to transcription and translation of the gene and the potential subsequent posttranslational modification of the protein in different host organisms. Proteomics additionally allows the direct elucidation of subcellular localization whereas transcriptomics and genomics rely on predictions. Proteins can be localized to different sub-compartments, such as the cytosol, peroxisomes, mitochondria, endoplasmic reticulum, Golgi vesicles, lipid droplets, lysosomes, and membranes or to the extracellular space. This information provides valuable cues about the conditions under which the proteins are functional and stable. In turn, this allows for optimization of the conditions for functional screenings, which can increase the number of hits. One of the most direct ways to discover enzymes is through activity-based proteomics which relies on enzyme class specific probes for simultaneous identification of individual enzymatic activities sharing the same reaction mechanism (Cravatt et al., 2008; Schittmayer and Birner-Gruenberger, 2012). This does however require the development of appropriate probes.

Enzyme discovery workflow

The most commonly used proteomic approach for enzyme discovery is bottom-up shotgun proteomics. In this approach proteins are enzymatically digested into peptides which are analyzed either by tandem mass spectrometry after a separation on a liquid chromatography system and electrospray ionization (ESI–LC–MS/MS) or by matrix-assisted laser desorption/ionization followed by time of flight mass spectrometry (MALDI–TOF–MS) (Fig. 1). Identification of proteins is based on the comparison of the measured mass to charge ratios (m/z) of the peptides and their fragments to the respective m/z values for all theoretical peptides and fragments thereof present in the database, which are generated by in silico digestion of the database with the same enzyme. The comparisons are automatically performed by search programs such as MASCOT (Pappin et al., 1999), SEQUEST (Eng et al., 1994), MaxQuant (Cox and Mann, 2008), MSAmanda (Dorfer et al., 2014) OMSSA (Geer et al., 2004), X!Tandem (Craig and Beavis, 2004), COMET (Eng et al., 2013), MS–GF+ (Kim and Pevzner, 2014) and MyriMatch (Tabb et al., 2008). All these search engines infer the presence of a protein from identification of at least one peptide that is unique to this protein (within the used database). This however implies that for a protein to be correctly identified it needs to be present in the database. Although the possibility of performing de-novo sequencing of proteins exists (Hughes et al., 2010), and could circumvent the problem of a protein not being present in the database, it is time and labor intensive and requires large amounts of highly purified proteins. The most direct way to construct a high-quality database for peptide mapping is by building it from the same sample used for protein identification by either genome or transcriptome sequencing (Fig. 1). For cultivable and, especially, model organisms, genome sequences (and even curated protein databases like SwissProt) are publicly available and present an easy way to obtain a high-quality database. Among the most prominent suppliers of sequence databases are the National Center for Biotechnology Information (NCBI), the European Molecular Biology Laboratory (EMBL) and the DNA Data Bank of Japan (DDBJ), which are all integrated through the International Nucleotide Sequence Database Collaboration (INSDC). Other important public database resources are the Uniprot Knowledge database (UniProtKB) (Consortium, 2015) or the Joint Genome Institute's genome portal (JGI) (Nordberg et al., 2014). However, if public databases lack sequence information of the target organism, which is the case for most complex environmental samples, custom database creation is necessary. If nucleotide sequencing is unfeasible, a reasonable approach is to compile a database consisting of publicly available sequences from (more or less) related organisms. This has been shown to be less reliable than using databases based on the target organism alone (Bräutigam et al., 2008) and only allows prediction of protein function but not of the exact sequence, preventing cloning and expression of the protein. Also, by querying against a database not containing the target organism's sequences, potential novel enzymes, that lack any homology with other proteins, cannot be identified. In order to minimize the loss of information it is advisable to create a custom protein database translated from the sample's transcriptome. Advances within the field of RNA sequencing have allowed for the analysis of entire meta-transcriptomes and made them available for construction of protein databases. Particularly developments in cDNA generation protocols and increasing sequencing capacity (McGettigan, 2013) has made transcriptomics more accessible to a wider audience being offered as an affordable service by several companies. The preparation of cDNA involves three major steps: (1) the extraction of RNA from cultured organisms or environmental samples, (2) the removal of rRNA and normalization to enrich for rare transcripts and (3) the reverse transcription to generate cDNA, which subsequently is sequenced. Although considerable progress has been made in assembly of RNAseq data methods, analysis still poses some challenges. Among these is the misalignment of reads to closely related genes (such as sequences coding for isoenzymes), also termed transcript shadowing (McGettigan, 2013) and the fact that many de novo assembly algorithms are still highly memory intensive and therefore require access to advanced computing facilities (Grabherr et al., 2011). Proteomic and metaproteomic screenings have been performed on both cultured microorganisms as well as environmental samples with the purpose of discovering novel enzymatic functions or for determining optimal culture conditions leading to e.g. optimal degradation of biomass or phosphorus removal. Here we describe a few selected examples of enzyme or whole pathway discovery where either a protein database from a public repository or a genome or transcriptome sequence has been obtained prior to proteomic analyses. Moreover, we discuss the advantages offered by combining proteomics with protein sequence databases derived from mRNA sequencing for enzyme discovery in more depth.

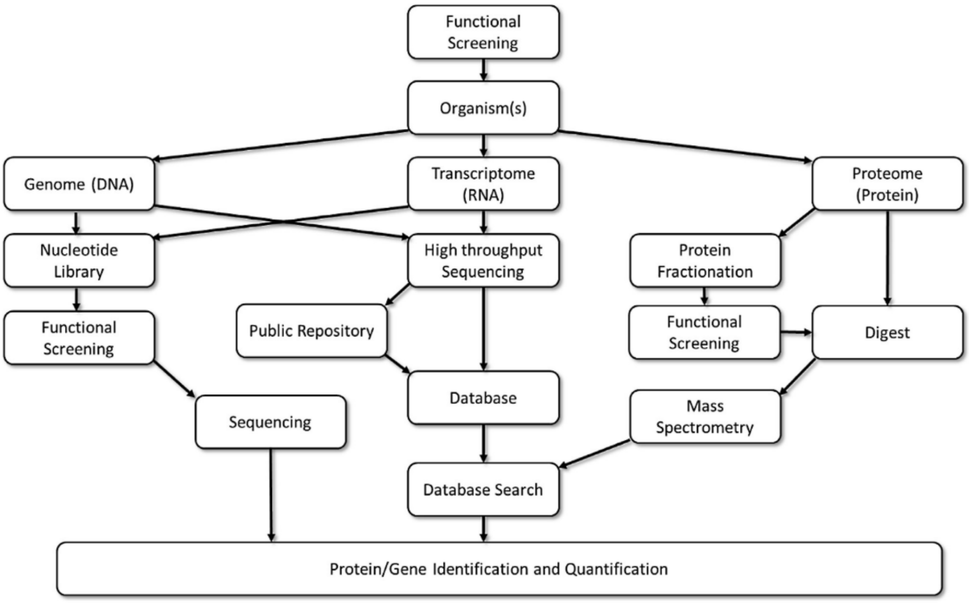


Fig. 1. Workflow of omics experiments for the discovery of novel enzymes. Functional screenings lead to the identification of organisms, tissues or other samples of interest. These can be analyzed omics technologies to identify the gene/protein of interest. Genome and transcriptome sequence data can be integrated into the proteomics workflow, by generating databases, which can improve the number of correct protein identifications from mass spectrometric data. This facilitates the access to novel enzymes by recombinant expression in microbial hosts.

Pathway discovery for production of secondary metabolites

As the chemical synthesis of many useful compounds, like alkaloids or biosurfactants, can be long and complicated, using organisms that naturally synthesize these secondary metabolites can be very beneficial. However, to allow the heterologous production, the knowledge of the enzymatic steps involved is essential. Here we present a few studies employing proteomic screenings using different types of protein sequence databases for this purpose and discuss the advantages and drawbacks of these databases.

Polyketide synthesis

One example, in which the availability of a high-quality protein database could have assisted protein identification in a more direct way, is the elucidation of a non-ribosomal peptide synthetase (NRPS) and a polyketide synthase gene cluster from an un-sequenced *Bacillus* sp. isolated from a soil sample. By exploiting the large size of these protein complexes, Evans et al. were able to identify peptides which served as the basis for PCR primer design (Evans et al., 2011). This approach allowed them to amplify the gene cluster containing several enzymes important for the synthesis of koranimine, a secondary metabolite, which falls into a family of compounds containing e.g. known mycotoxins and immunosuppressants. The gene cluster discovered was not highly related to any other NRPS cluster present in GenBank (the NCBI genetic sequence database) (Benson et al., 2013) at the time of analysis. Therefore, while publicly available databases would most likely not have yielded protein identification an improved protein sequence database derived from mRNA sequencing could have allowed the direct identification of the protein sequence in the gel band without the need to perform difficult PCR amplifications using degenerate primers derived from amino acid sequences.

Alkaloid production

Papaver somniferum (opium poppy) is the only commercial source of the medically important alkaloids noscapine, morphine, codeine, papaverine and sanguinarine (Berenyi et al., n.d.). The biosynthesis of these alkaloids have been studied extensively (Hagel and Facchini, 2013). Here we summarize the employed proteomic screening approaches and discuss how their outcome was improved by the use of protein sequence databases derived from Expressed Sequence Tag (EST, i.e. a short cDNA read) sequencing. The first proteomic approach to detect and localize enzymes responsible for alkaloid biosynthesis was performed on the latex of the opium poppy. Latex proteins were extracted and separated by 2D-gel electrophoresis prior to in gel digestion, and peptide sequencing of picked gel spots (Decker et al., 2000). The resultant MS/MS data were queried against all the present sequences in the NCBI non-redundant protein database, as no *P. somniferum* specific database was available at the time. Of the 75 excised gel spots, 69 contained peptides that could be assigned to homologous proteins with known functions based on sequence-similarity to other organisms. Among the identified proteins was codeine reductase, which is known to be involved in morphine synthesis, confirming the presence of this enzyme in latex. Proteomics was also used in combination with more conventional approaches for enzyme discovery: Peptide sequencing and homology searches in combination with sequence-similarity based PCR-primers lead to the identification of two methyl-transferases involved in alkaloid biosynthesis (Ounaroon et al., 2003). Similarly, Jacobs et al. (Jacobs et al., 2005) were challenged with a lack of sequence resources for database construction in their quest for novel genes for the biosynthesis of alkaloids. They performed MALDI–MS/MS analyses of samples from the medicinal plant Catharanthus roseus and due to poor genome and proteome sequence information (only 27 deposited protein sequences in SwissProt at the time of analysis) they had to rely on the NCBInr database restricted to *Viridiplantae*. 58 different proteins, among these several enzymes from the alkaloid biosynthesis pathways, were identified. Of these, only one was a *C. roseus* protein. Thus, without access to specific sequence databases for *C. roseus* the huge potential of the peptide sequence data set could not be fully explored. More recently, the availability of EST databases specific for opium poppies allowed for better identification of proteins involved in plant defense responses and synthesis of antimicrobial alkaloids in *P. somniferum* (Zulak et al., 2009). Latex proteins separated by 2D-gel electrophoresis were identified by LC–MS/MS and MALDI–TOF–MS and searching against both the NCBInr protein database restricted to *Viridiplantae* containing 381,872 sequences and a translated EST database containing 22,036 sequences. Of 219 identified proteins, 29% were found using the sequences contained only within the EST database while another 12% could be identified from sequences contained within both databases. In this study, transcriptomics was moreover included as a complementary method to quantify transcript levels under elicitor induced and non-induced conditions showing the induction of six known alkaloid biosynthetic enzymes. As deeper sequencing of the transcriptomes was performed the size of the EST databases used for proteomic screenings increased, and 1004 peptides and polypeptides could be identified from opium poppy cell lines (Desgagné-Penix et al., 2010). Again, the search results of the public NCBInr database restricted to *Viridiplantae* were compared to those obtained by using an EST database (427,369 sequences). While searching the NCBInr database yielded only 288 identified peptides or polypeptides, querying against the opium poppy specific EST database resulted in the identification of 1004 peptides and polypeptides. This example clearly demonstrates the advantage of complementing proteomic studies with the use of organism specific EST databases.

Isoprenoid production

In a similar example, enzymes responsible for isoprenoid biosynthesis in tomato trichrome were discovered by a combination of mRNA sequencing and proteomics of the same organism. Schilmiller et al. (Schilmiller et al., 2010) sequenced the transcriptome (mRNA) and subsequently translated it into a protein sequence database which was then used for proteomic screening. This synergistic approach enabled the identification of 1552 proteins including all eight proteins involved in the production of the precursors needed for isoprenoid biosynthesis. Moreover, it led to the discovery of a previously uncharacterized tomato trichome sesquiterpene synthase involved in the synthesis of the anti-inflammatory compounds caryophyllene (approved as a food additive by the FDA) and humulene.

Sophorolipid production

Another prominent example for the synergism of different omics technologies is the discovery of enzymes involved in sophorolipid (SL) production. The yeast *Starmerella bombicola* (formerly *Candida bombicola*) was found to produce SL (Spencer et al., 1970) which have been suggested as environmentally friendly biosurfactants for the use in fields ranging from cleaning agents to stabilization of nanoparticles (Basak et al., 2013). Therefore, people have focused on the discovery of the genes necessary for the SL biosynthesis pathway. Initially, genes were identified one by one using sequence homology searches combined with PCR and gene walking, as no genome sequence was known (Saerens and Soetaert, 2011; Saerens et al., 2011; Van Bogaert et al., 2007). In 2013 the genome of *S. bombicola* was sequenced (Ciesielska et al., 2013) allowing for more comprehensive studies of the organism. This quickly led to the identification of a gene cluster containing all proteins known to be involved in SL synthesis (Van Bogaert et al., 2013). The cluster also contained other genes, one of which was found to encode a necessary SL transporter. The first large scale proteomic study on *S. bombicola* identified and quantified 615 proteins. Quantitation was performed both by RNAseq and by a quantitative proteomic (stable isotope labeling by amino acids in cell culture (SILAC)) experiment, where the exponential and early stationary growth phases were compared. The authors reported a simultaneous production of all known proteins involved in SL biosynthesis (Ciesielska et al., 2013), which was followed up by an exoproteome analysis where the lactone esterase responsible for the final step of SL synthesis (lactonization) was identified (Ciesielska et al., 2014; Saerens et al., 2015). This example demonstrates the advantage of obtaining a genome sequence and combining RNAseq and quantitative proteomics for rapid identification of unknown enzymes, which is relatively easy and cheap for a cultivable organism.

Enzyme discovery

The importance of using a protein database containing the target protein or a very close homologue for the discovery of new enzymes, especially from uncultivable isolated organisms and environmental samples containing complex microbial communities, became similarly apparent in several studies published in recent years.

Isolated organisms

While searching for cellulases and xylanases Amore et al. isolated microorganisms from three different areas in the Western Ghat region of India (Amore et al., 2015). By analyzing eight different microorganisms obtained by cultivation, they were able to show xylanase activity in a *Bacillus amyloliquefaciens* XR44A strain. Subsequent HPLC–MS/MS analysis of peptides derived from a native gel band harboring xylanase activity and search of the LC–MS/MS data against the NCBInr database resulted in the identification of an endo-1,4-beta-xylanase. However, the protein identified in NCBInr was from *Paenibacillus macerans*, an organism not even belonging to the same family as the source organism (*Bacillaceae*). Thus, although the authors were able to identify an endo-1,4-beta-xylanase it is probably only a close homologue of the actual protein present in the sample, which could not be identified due to the absence of its protein sequence in the database. Since the analysis did not yield the actual sequence cloning and recombinant expression is restricted. The same problem was faced by (Tiwari et al., 2014). After identifying hydrolytic activities in the secretome of the phytopathogenic fungus *Phoma exigua*, LC–MS/MS analysis was performed to identify potential protein targets with glycosyl hydrolase activity. Since the *P. exigua* genome was not available, searches were conducted against the NCBInr database restricted to fungal sequences and resulted in the identification of 33 different homologous proteins. The majority of these proteins were annotated as glycosyl hydrolases, but none of them was actually derived from *P. exigua* and therefore the exact sequence of the full-length protein was not accessible by the employed approach, once again restricting the further use of the obtained sequences. (Kirsch et al., 2012) studied plant cell wall degradation in the leaf beetle gut. The degradation is proposed to be mediated by enzymes secreted by the beetle itself. Thus, screenings of the gut proteome and transcriptome were performed to identify the plant cell wall degrading enzymes. Gel electrophoresis combined with activity assays resulted in gel bands with activity towards cellulose, pectin or xylan. To create an appropriate protein database and to identify potential enzymes, a meta-transcriptomic analysis was performed, where several tissues were sampled during two developmental stages of the beetle and while being subjected to different environmental stress factors. The combined date resulted in an EST database with 644,940 entries, which was concatenated with the NCBInr database. The proteomic screening of the peptides from the active gel bands derived from the gut content identified 13 proteins from the beetle with a putative plant cell wall degrading enzymatic function, whereas the transcriptome of the same sample suggested 19 putative plant cell wall degrading enzymes. One may speculate that this difference might arise due to the higher specificity of the proteomic screen, low translation rates or protein stability of some candidates or wrong in silico functional assignment on the RNA level.

Microbial communities

With the aim of examining the suitability of activated zeolite as a carrier for microorganisms in anaerobic digestion processes, Weiß et al. performed LC–MS/MS of hydrolytically active protein bands after batch fermentation of grass silage (Weiß et al., 2013). Alongside this, they analyzed single strand conformation polymorphisms (SSCP) of the total bacterial community based on bacterial and archaeal 16S rRNA. By searching the LC–MS/MS data against the NCBInr database they were successful in identifying 36 biomass degradation associated enzymes, mainly from organisms in the *Paenibacillaceae* family. Meanwhile, predominantly species from the *Clostridium*, *Methanoculleus* and *Pseudomonas* families were shown to be the major organisms on activated zeolite by SSCP analysis. This can be interpreted in two ways: One possibility is that the specific proteomic analysis of enzymes highly complemented the 16S rRNA analysis of the total bacterial community. Since a small subset of all organisms present (*Paenibacillaceae* family) appeared to be responsible for the secretion of most of the hydrolytic enzymes, while not being among the most dominant organisms in terms of numbers. A second possible interpretation is that the protein identifications were assigned to the *Paenibacillaceae* family because the actual organisms that the proteins originate from were not present in the protein sequence database. In this case, the first interpretation seems to be more likely, since proteins from organisms more closely related to the organisms identified by SSCP than to the *Paenibacillaceae* family were present in the NCBInr database at the time.

The next example also shows that creation of custom protein databases by including relevant sequences into the analysis can improve the overall success in identification of novel enzymes. This approach of using focused databases was employed by Schneider et al. when querying their obtained peptide MS/MS spectra against a database compiled from both, a farm silage soil metagenome and the UniRef100 database (a clustered set of sequences from the UniProtKB) (Schneider et al., 2012). Using this metaproteomic approach they could show that on the one hand litter microbial communities differ between sampling sites and seasons and that on the other hand fungi are the main producers of litter-degrading enzymes. The majority of enzymes identified by this approach belonged to the group of cellulases, phosphatases, xylanases and lipases. An even higher degree of specificity in database creation was implemented by a metaproteomic study on wastewater sludge. Wilmes et al. used a compilation of metagenomes derived from three different wastewater sludge sites to elucidate the enzymatic functions related to biological phosphorus removal. The use of these specific databases allowed them to identify enzymes involved in fatty acid oxidation and polyhydroxyalkanoate synthesis, glycogen degradation, TCA cycle, phosphate bioenergetics and stress response (Wilmes et al., 2008). The metagenomes, however, were generated from wastewater sludge from other sites than the metaproteomic study. Thus, although these databases proved very useful for enzyme identification, the list of protein identifications may not be complete since the microbial composition of the sludge from the different sites may not be identical.

Conclusion

Proteomic approaches for gene/protein discovery depend on the usage of appropriate protein databases containing the sequences of the proteins present in the sample. Sequences of homologous proteins obtainable from public repositories have been used as databases instead, but the resulting data may be less comprehensive as demonstrated by the examples described above. In the case of the discovery of enzymes involved in SL biosynthesis in *S. bombicola* (Ciesielska et al., 2013), the acquisition of a genomic sequence improved protein identifications. Similarly, the generation of EST databases by transcriptomic analysis accelerated enzyme discovery significantly as demonstrated by the examples of alkaloid biosynthesis in *P. somniferum* (Desgagné-Penix et al., 2010; Zulak et al., 2009) and plant cell wall degradation in leaf beetles (Kirsch et al., 2012). Kirsch et al.'s study on leaf beetles also demonstrates the benefits of transcriptomic sequence data both on its own and in its use to improve the protein database by combining the generated EST database with the NCBInr database. This approach of integrating mRNA sequencing data into protein databases would also be beneficial to environmental samples where the microbial composition (and thus potential proteome) is not known. This will allow the identification of a higher number of full-length protein sequences and their true origin rather than relying on close homologues found in other databases. With the reduced cost and increasing sequencing capacities proteomic approaches benefit from the generation of improved protein sequence databases derived from mRNA sequencing experiments. Furthermore, one can expect that reanalysis of already acquired proteomic data with newly available improved protein sequence databases will yield valuable new information at very low cost.

Acknowledgements

This work was supported by the Austrian Science Fund (FWF) doctoral school “DK Metabolic and Cardiovascular Disease” (W1226), the Federal Ministry of Economy, Family and Youth (BMWFJ), the Federal Ministry of Traffic, Innovation and Technology (bmvit), the Styrian Business Promotion Agency SFG, the Standortagentur Tirol and ZIT-Technology Agency of the City of Vienna through the COMET-Funding Program managed by the Austrian Research Promotion Agency FFG.

Figures

Fig. 1. Workflow of omics experiments for the discovery of novel enzymes. Functional screenings lead to the identification of organisms, tissues or other samples of interest.

References

Amann, R.I., Ludwig, W., Schleifer, K.H., 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. 59, 143–169. https://doi.org/10.1016/j.jip.2007.09.009

Amore, A., Parameswaran, B., Kumar, R., Birolo, L., Vinciguerra, R., Marcolongo, L., Ionata, E., La Cara, F., Pandey, A., Faraco, V., 2015. Application of a new xylanase activity from Bacillus amyloliquefaciens XR44A in brewer’s spent grain saccharification. J. Chem. Technol. Biotechnol. 90, 573–581. https://doi.org/10.1002/jctb.4589

Anastas, P.T., Warner, J.C., 1998. Green Chemistry: Theory and Practice. Oxford University Press.

Basak, G., Das, D., Das, N., 2013. Dual role of acidic diacetate sophorolipid as biostabilizer for ZnO nanoparticle synthesis and biofunctionalizing agent against Salmonella enterica and Candida albicans. J. Microbiol. Biotechnol. 24, 87–96.

Behrens, G. a., Hummel, A., Padhi, S.K., Schätzle, S., Bornscheuer, U.T., 2011. Discovery and Protein Engineering of Biocatalysts for Organic Synthesis. Adv. Synth. Catal. 353, 2191–2215. https://doi.org/10.1002/adsc.201100446

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2013. GenBank. Nucleic Acids Res. 41, 36–42. https://doi.org/10.1093/nar/gks1195

Berenyi, S., Csutoras, C., Sipos, A., n.d. Recent Developments in the Chemistry of Thebaine and its Transformation Products as Pharmacological Targets. Curr. Med. Chem. 16, 3215–3242. https://doi.org/doi:10.2174/092986709788803295

Bornscheuer, U.T., Huisman, G.W., Kazlauskas, R.J., Lutz, S., Moore, J.C., Robins, K., 2012. Engineering the third wave of biocatalysis. Nature 485, 185–194. https://doi.org/10.1038/nature11117

Bräutigam, A., Shrestha, R.P., Whitten, D., Wilkerson, C.G., Carr, K.M., Froehlich, J.E., Weber, A.P.M., 2008. Comparison of the use of a species-specific database generated by pyrosequencing with databases from related species for proteome analysis of pea chloroplast envelopes. J Biotechnol 136, 44–53.

Cammarota, M.C., Freire, D.M.G., 2006. A review on hydrolytic enzymes in the treatment of wastewater with high oil and grease content. Bioresour. Technol. 97, 2195–2210. https://doi.org/10.1016/j.biortech.2006.02.030

Ciesielska, K., Li, B., Groeneboer, S., Van Bogaert, I., Lin, Y.C., Soetaert, W., Van De Peer, Y., Devreese, B., 2013. SILAC-based proteome analysis of starmerella bombicola sophorolipid production. J. Proteome Res. 12, 4376–4392. https://doi.org/10.1021/pr400392a

Ciesielska, K., Van Bogaert, I.N., Chevineau, S., Li, B., Groeneboer, S., Soetaert, W., Van de Peer, Y., Devreese, B., 2014. Exoproteome analysis of Starmerella bombicola results in the discovery of an esterase required for lactonization of sophorolipids. J. Proteomics 98, 159–174. https://doi.org/10.1016/j.jprot.2013.12.026

Consortium, T.U., 2015. UniProt: a hub for protein information. Nucleic Acids Res. 43, D204–D212. https://doi.org/10.1093/nar/gku989

Cox, J., Mann, M., 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372. https://doi.org/10.1038/nbt.1511

Craig, R., Beavis, R.C., 2004. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20, 1466–1467. https://doi.org/10.1093/bioinformatics/bth092

Cravatt, B.F., Wright, A.T., Kozarich, J.W., 2008. Activity-Based Protein Profiling: From Enzyme Chemistry to Proteomic Chemistry. Annu. Rev. Biochem. 77, 383–414. https://doi.org/10.1146/annurev.biochem.75.101304.124125

Decker, G., Wanner, G., Zenk, M.H., Lottspeich, F., 2000. Characterization of proteins in latex of the opium poppy (Papaver somniferum) using two-dimensional gel electrophoresis and microsequencing. Electrophoresis 21, 3500–3516. https://doi.org/10.1002/1522-2683(20001001)21:16<3500::AID-ELPS3500>3.0.CO;2-O [pii]\r10.1002/1522-2683(20001001)21:16<3500::AID-ELPS3500>3.0.CO;2-O

Desgagné-Penix, I., Khan, M.F., Schriemer, D.C., Cram, D., Nowak, J., Facchini, P.J., 2010. Integration of deep transcriptome and proteome analyses reveals the components of alkaloid metabolism in opium poppy cell cultures. BMC Plant Biol. 10, 252. https://doi.org/10.1186/1471-2229-10-252

Deutsch, M., Long, M., 1999. Intron – exon structures of eukaryotic model organisms 27, 3219–3228.

Dorfer, V., Pichler, P., Stranzl, T., Stadlmann, J., Taus, T., Winkler, S., Mechtler, K., 2014. MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. J. Proteome Res. 13, 3679–84. https://doi.org/10.1021/pr500202e

Eberl, a., Heumann, S., Kotek, R., Kaufmann, F., Mitsche, S., Cavaco-Paulo, a., Gübitz, G.M., 2008. Enzymatic hydrolysis of PTT polymers and oligomers. J. Biotechnol. 135, 45–51. https://doi.org/10.1016/j.jbiotec.2008.02.015

Ekkers, D.M., Cretoiu, M.S., Kielak, A.M., van Elsas, J.D., 2012. The great screen anomaly—a new frontier in product discovery through functional metagenomics. Appl. Microbiol. Biotechnol. 93, 1005–1020. https://doi.org/10.1007/s00253-011-3804-3

Eng, J.K., Jahan, T.A., Hoopmann, M.R., 2013. Comet: An open-source MS/MS sequence database search tool. PROTEOMICS 13, 22–24. https://doi.org/10.1002/pmic.201200439

Eng, J.K., McCormack, a L., Yates, J.R., 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–89. https://doi.org/10.1016/1044-0305(94)80016-2

Erickson, B., Nelson, Winters, P., 2012. Perspective on opportunities in industrial biotechnology in renewable chemicals. Biotechnol. J. 7, 176–85. https://doi.org/10.1002/biot.201100069

Evans, B.S., Ntai, I., Chen, Y., Robinson, S.J., Kelleher, N.L., 2011. Proteomics-based discovery of koranimine, a cyclic imine natural product. J. Am. Chem. Soc. 133, 7316–7319. https://doi.org/10.1021/ja2015795

Faber, K., 2011. Biotransformations aid organic chemists, ChemInform. https://doi.org/10.1007/978-3-642-17393-6

Ferrer, M., Beloqui, A., Timmis, K.N., Golyshin, P.N., 2009. Metagenomics for Mining New Genetic Resources of Microbial Communities. J. Mol. Microbiol. Biotechnol. 16, 109–123. https://doi.org/10.1159/000142898

Geer, L.Y., Markey, S.P., Kowalak, J. a, Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W., Bryant, S.H., 2004. Open mass spectrometry search algorithm. J Proteome Res 3, 958–964. https://doi.org/10.1021/pr0499491

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech 29, 644–652.

Gygi, S.P., Rochon, Y., Franza, B.R., Aebersold, R., 1999. Correlation between Protein and mRNA Abundance in Yeast 19, 1720–1730.

Hagel, J.M., Facchini, P.J., 2013. Benzylisoquinoline Alkaloid Metabolism: A Century of Discovery and a Brave New World. Plant Cell Physiol. 54, 647–672. https://doi.org/10.1093/pcp/pct020

Hughes, C., Ma, B., Lajoie, G.A., 2010. Proteome Bioinformatics 604. https://doi.org/10.1007/978-1-60761-444-9

Illanes, A., Cauerhff, A., Wilson, L., Castro, G.R., 2012. Recent trends in biocatalysis engineering. Bioresour. Technol. 115, 48–57. https://doi.org/10.1016/j.biortech.2011.12.050

Jacobs, D.I., Gaspari, M., van der Greef, J., van der Heijden, R., Verpoorte, R., 2005. Proteome analysis of the medicinal plant Catharanthus roseus. Planta 221, 690–704. https://doi.org/10.1007/s00425-004-1474-4

Janssen, D.B., Dinkla, I.J.T., Poelarends, G.J., Terpstra, P., 2005. Bacterial degradation of xenobiotic compounds: evolution and distribution of novel enzyme activities. Environ. Microbiol. 7, 1868–1882. https://doi.org/10.1111/j.1462-2920.2005.00966.x

Kim, S., Pevzner, P. a, 2014. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun. 5, 5277. https://doi.org/10.1038/ncomms6277

Kirsch, R., Wielsch, N., Vogel, H., Svatoš, A., Heckel, D.G., Pauchet, Y., 2012. Combining proteomics and transcriptome sequencing to identify active plant-cell-wall-degrading enzymes in a leaf beetle. BMC Genomics 13, 587. https://doi.org/10.1186/1471-2164-13-587

Kullman, S.W., Matsumura, F., 1996. Metabolic pathways utilized by Phanerochaete chrysosporium for degradation of the cyclodiene pesticide endosulfan. Appl. Environ. Microbiol. 62, 593–600.

Li, S., Yang, X., Yang, S., Zhu, M., Wang, X., 2012. Technology prospecting on enzymes: application, marketing and engineering. Comput. Struct. Biotechnol. J. 2, e201209017. https://doi.org/10.5936/csbj.201209017

McGettigan, P. a, 2013. Transcriptomics in the RNA-seq era. Curr. Opin. Chem. Biol. 17, 4–11. https://doi.org/10.1016/j.cbpa.2012.12.008

Meyer, H.P., 2011. Sustainability and biotechnology. Org. Process Res. Dev. 15, 180–188. https://doi.org/10.1021/op100206p

Müller, R.-J., Schrader, H., Profe, J., Dresler, K., Deckwer, W.-D., 2005. Enzymatic Degradation of Poly(ethylene terephthalate): Rapid Hydrolyse using a Hydrolase fromT. fusca. Macromol. Rapid Commun. 26, 1400–1405. https://doi.org/10.1002/marc.200500410

Ounaroon, A., Decker, G., Schmidt, J., Lottspeich, F., Kutchan, T.M., 2003. (R,S)-Reticuline 7- O -methyltransferase and (R,S)-norcoclaurine 6- O -methyltransferase of Papaver somniferum - cDNA cloning and characterization of methyl transfer enzymes of alkaloid biosynthesis in opium poppy. Plant J. 36, 808–819. https://doi.org/10.1046/j.1365-313X.2003.01928.x

Pappin, D.J.C., Creasy, D.M., Cottrell, J.S., 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data Proteomics and 2-DE. Electrophoresis 3551–3567.

Rabinovich, M.L., Bolobova, A. V, Vasil'chenko, L.G., 2004. Fungal Decomposition of Natural Aromatic Structures and Xenobiotics: A Review. Appl. Biochem. Microbiol. 40, 1–17. https://doi.org/10.1023/B:ABIM.0000010343.73266.08

Ribitsch, D., Acero, E.H., Greimel, K., Eiteljoerg, I., Trotscha, E., Freddi, G., Schwab, H., Guebitz, G.M., 2012. Characterization of a new cutinase from *Thermobifida alba* for PET-surface hydrolysis. Biocatal. Biotransformation 30, 2–9. https://doi.org/10.3109/10242422.2012.644435

Robertson, D.E., Steer, B.A., 2004. Recent progress in biocatalyst discovery and optimization. Curr. Opin. Chem. Biol. 8, 141–149. https://doi.org/10.1016/j.cbpa.2004.02.010

Saerens, K., Soetaert, W., 2011. Identification of key-genes from the sophorolipid biosynthetic pathway of Candida bombicola opens a new route to increased biosurfactant yields, in: COMMUNICATIONS IN AGRICULTURAL AND APPLIED BIOLOGICAL SCIENCES. pp. 65–68.

Saerens, K.M.J., Roelants, S.L.K.W., Van Bogaert, I.N. a., Soetaert, W., 2011. Identification of the UDP-glucosyltransferase gene UGTA1, responsible for the first glucosylation step in the sophorolipid biosynthetic pathway of Candida bombicola ATCC 22214. FEMS Yeast Res. 11, 123–132. https://doi.org/10.1111/j.1567-1364.2010.00695.x

Saerens, K.M.J., Van Bogaert, I.N.A., Soetaert, W., 2015. Characterization of sophorolipid biosynthetic enzymes from Starmerella bombicola. FEMS Yeast Res. 15.

Sarrouh, B., Santos, T.M., Miyoshi, A., Dias, R., Azevedo, V., 2012. Up-To-Date Insight on Industrial Enzymes Applications and Global Market. J. Bioprocess. Biotech. S4, 1–10. https://doi.org/10.4172/2155-9821.S4-002

Schilmiller, a. L., Miner, D.P., Larson, M., McDowell, E., Gang, D.R., Wilkerson, C., Last, R.L., 2010. Studies of a Biochemical Factory: Tomato Trichome Deep Expressed Sequence Tag Sequencing and Proteomics. Plant Physiol. 153, 1212–1223. https://doi.org/10.1104/pp.110.157214

Schittmayer, M., Birner-Gruenberger, R., 2012. Lipolytic proteomics. Mass Spectrom. Rev. 31, 570–582. https://doi.org/10.1002/mas.20355

Schneider, T., Keiblinger, K.M., Schmid, E., Sterflinger-Gleixner, K., Ellersdorfer, G., Roschitzki, B., Richter, A., Eberl, L., Zechmeister-Boltenstern, S., Riedel, K., 2012. Who is who in litter decomposition? Metaproteomics reveals major microbial players and their biogeochemical functions. ISME J. 6, 1749–1762. https://doi.org/10.1038/ismej.2012.11

Spencer, J.F.T., Gorin, P.A.J., Tulloch, A.P., 1970. Torulopsis bombicola sp. n. Antonie Van Leeuwenhoek 36, 129–133. https://doi.org/10.1007/BF02069014

Steinkellner, G., Gruber, C.C., Pavkov-Keller, T., Binter, A., Steiner, K., Winkler, C., Łyskowski, A., Schwamberger, O., Oberer, M., Schwab, H., Faber, K., Macheroux, P., Gruber, K., 2014. Identification of promiscuous ene-reductase activity by mining structural databases using active site constellations. Nat. Commun. 5, 1–9. https://doi.org/10.1038/ncomms5150

Tabb, D.L., Fernando, C.G., Chambers, M.C., 2008. MyriMatch:highly accurate tandem mass spectral peptide identificaiton by multivariate hypergeometric analysis. J Proteome Res 6, 654–661. https://doi.org/10.1021/pr0604054.MyriMatch

Tiwari, R., Singh, S., Singh, N., Adak, A., Rana, S., Sharma, A., Arora, A., Nain, L., 2014. Unwrapping the hydrolytic system of the phytopathogenic fungus Phoma exigua by secretome analysis. Process Biochem. 49, 1630–1636. https://doi.org/10.1016/j.procbio.2014.06.023

Uchiyama, T., Miyazaki, K., 2009. Functional metagenomics for enzyme discovery: challenges to efficient screening. Curr. Opin. Biotechnol. 20, 616–622. https://doi.org/10.1016/j.copbio.2009.09.010

Van Bogaert, I.N. a, Holvoet, K., Roelants, S.L.K.W., Li, B., Lin, Y.C., Van de Peer, Y., Soetaert, W., 2013. The biosynthetic gene cluster for sophorolipids: A biotechnological interesting biosurfactant produced by Starmerella bombicola. Mol. Microbiol. 88, 501–509. https://doi.org/10.1111/mmi.12200

Van Bogaert, I.N.A., Saerens, K., De Muynck, C., Develter, D., Soetaert, W., Vandamme, E.J., 2007. Microbial production and application of sophorolipids. Appl. Microbiol. Biotechnol. 76, 23–34. https://doi.org/10.1007/s00253-007-0988-7

Weiß, S., Lebuhn, M., Andrade, D., Zankel, A., Cardinale, M., Birner-Gruenberger, R., Somitsch, W., Ueberbacher, B.J., Guebitz, G.M., 2013. Activated zeolite—suitable carriers for microorganisms in anaerobic digestion processes? Appl. Microbiol. Biotechnol. 97, 3225–3238. https://doi.org/10.1007/s00253-013-4691-6

Wilmes, P., Wexler, M., Bond, P.L., 2008. Metaproteomics Provides Functional Insight into Activated Sludge Wastewater Treatment. PLoS ONE 3, e1778. https://doi.org/10.1371/journal.pone.0001778

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S., Davies, S.R., Wang, S., Wang, P., Kinsinger, C.R., Rivers, R.C., Rodriguez, H., Townsend, R.R., Ellis, M.J.C., Carr, S. a, Tabb, D.L., Coffey, R.J., Slebos, R.J.C., Liebler, D.C., 2014. Proteogenomic characterization of human colon and rectal cancer. Nature 513, 382–7. https://doi.org/10.1038/nature13438

Zulak, K.G., Khan, M.F., Alcantara, J., Schriemer, D.C., Facchini, P.J., 2009. Plant Defense Responses in Opium Poppy Cell Cultures Revealed by Liquid Chromatography-Tandem Mass Spectrometry Proteomics. Mol. Cell. Proteomics 8, 86–98. https://doi.org/10.1074/mcp.M800211-MCP200

Research Article

A combinatorial approach of transcriptomics and proteomics enables the discovery of a novel glycoside hydrolase from the fungus *Lecanicillium attenuatum*

Lukas Sturmberger[a], Barbara Darnhofer [a,d], Ruth Birner-Grünberger [d] and Anton Glieder[a,b,c]

[a] Austrian Center of Industrial Biotechnology (ACIB), Petersgasse 14, 8010 Graz, Austria
[b] Institute of Molecular Biotechnology, Graz University of Technology, Petersgasse 14, 8010 Graz, Austria
[c] bisy e.U., Wetzawinkel 20, 8200 Hofstaetten/Raab, Austria
[d] Research Unit Functional Proteomics and Metabolomic Pathways, Institute of Pathology, Medical University of Graz, Graz, Austria

Abbreviations

EST	expressed sequence tag
NCBI	National center for biotechnology information
LC-MS/MS	liquid chromatography mass spectrometry/mass spectrometry
CBS	Centraalbureau voor Schimmelcultures
ARS	autonomously replicating sequences
RP-HPLC	reversed-phase high-pressure liquid chromatography
HRP	horseradish peroxidase
BLAST	basic local alignemnet search tool

Keywords

proteomics, transcriptomics, enzyme discovery, *Komagataella phaffii,* glycoside hydrolase, synergy

Abstract

The rapid developments in the fields of omics technologies have drastically facilitated the discovery of novel enzymes. As enzymatic transformations are increasingly competing with classical organic synthesis counterparts, the need for enzymes with new functions is growing. Functional screenings as well as genomics-based methods have proven to be valuable tools in the quest for new biocatalysts. Here we describe a combinatorial approach of transcriptomics and proteomics of a fungal secretome, which allowed the discovery of a novel glycoside hydrolase from the entomopathogenic fungal parasite *Lecanicillium attenuatum*. The characterized enzyme showed activity against β-cellobioside, β-xylopyranoside and β-glucopyranoside substrates and a pH optimum of pH 5 as well as and temperature optimum of 70°C. Additionally, the pH and temperature stability of the enzyme expressed in *Komagataella phaffii* was determined.

Introduction

As the use of enzymes for the transformation of organic compounds increasingly becomes a useful complementation to classical organic catalysis, the need for suitable biocatalysts becomes apparent. Apart from engineering already existent enzyme backbones and adapt them to new applications, the discovery of new enzymes from nature is an important step in developing novel industrial applications (Bornscheuer et al., 2012; Erickson et al., 2012; Illanes et al., 2012). Since less than 1% of all microbial strains can be readily cultivated, an observation termed "the great plate anomaly" (Amann et al., 1995), the discovery of new biocatalysts mainly relies on a variety of culture independent methods such as functional screenings of nucleic acid libraries (Uchiyama and Miyazaki, 2009) and more recently the use of different -omics technologies (Davids et al., 2013; McGettigan, 2013). Amongst these are genomics- (DeAngelis et al., 2010; Lämmle et al., 2007; Lee et al., 2010; Uchiyama and Miyazaki, 2009), transcriptomics-(Baldrian and López-Mondéjar, 2014; Desgagné-Penix et al., 2010; Kirsch et al., 2012) and proteomics- (Adav et al., 2012; Cravatt, 2014; Lanfranchi et al., 2017; Pohl, 2005) based discovery approaches which have different advantages and disadvantages regarding the nature of the sample to be analyzed. A combinatorial approach of more than one technology allows to compensate for the shortcomings encountered with certain approaches. Amongst them, the synergy of transcriptomics and proteomics allows the identification of novel enzymes in complex environmental samples (Sturmberger et al., 2016c). For instance, this synergy allowed the elucidation of the exact nature of the alkaloid pathway of *Papaver somniferum*. Two studies attempted the identification of the entire pathway, however were only able to partially elucidate the complete sequence of enzymatic reactions (Decker et al., 2000; Jacobs et al., 2005b). The lack of a high-quality specific protein databases hampered a more thorough analysis in both cases. However, a study generating an EST database specific for opium poppy allowed for the discovery of 219 proteins, from which 29% were only discoverable through the employment of siad EST database (Zulak et al., 2009). When the size of the EST database was increased, 3 times as many peptides were identified compared to a strategy solely relying on NCBI nr for protein database generation, 1004 and 288 peptides, respectively (Desgagné-Penix et al., 2010). Similarly, in an effort to identify cell wall degrading enzymes in the gut of leaf beetles, Kirsch and colleagues prepared transcriptome sequences of gut samples to create a highly specific protein database. In combination with proteomic screenings of peptides they were able to identify 19 putative plant cell wall degrading enzymes (Kirsch et al., 2012). In another embodiment, the sophorolipid pathway from *Stamerella bombicola* was elucidated based on the synergy of proteomics and transcriptomics/genomics. Initially, this undertaking was limited to the discovery of single genes (Saerens and Soetaert, 2011; Saerens et al., 2011; Van Bogaert et al., 2007), however the use of specific protein databases based on genome and transcriptome data (Ciesielska et al., 2014, 2013) allowed the elucidation of all known proteins involved in SL biosynthesis. However, as the size of publicly available databases are increasing steadily, the likelihood of finding a specific sequence within them is increasing as well. This in turn indicates that at a certain point int the future, the generation of a specific protein database is rendered redundant.

In this study we have attempted to identify carbohydrate active enzymes in the secretome of the fungus *Lecanicillium attenuatum*. The generation of a transcriptome sequence and the attempt to predict the number of secreted enzymes allowed us to generate a highly specific protein database. LC-MS/MS peptide mass fingerprinting and querying against this database allowed the identification of several proteins with sequence homology to carbohydrate active enzymes. Subsequently, a subset of sequences was expressed in a secretory manner in the methylotrophic yeast species *Komagataella phaffii*. In doing so, we were successful in identifying a novel glycoside hydrolase with β-cellobioside, β-xylopyranoside and β-glucopyranoside substrates. The temperature and pH optima of this newly discovered enzyme were determined, as well as the pH and temperature stabilities.

Material and Methods

Strains, Plasmids, Chemicals, and Media. The *Lecanicillium attenuatum* strain CBS 170.76 was cultivated in buffered minimal media containing autoclaved insects (flies, bees and wasps) as sole carbon source. This mineral medium contained 0.17 g CaSO4^{2H2O}, 2.86 g K2SO4, 0.64 g KOH, 2.32 g MgSO4^{7H2O}, 0.22 g NaCl, 0.6 g EDTA disodium salt and 4.25 ml H3PO4 (85%). 4,35 mL PTM1 solution per Liter were added after sterilization. The PTM1 salt solution contained in 1 Liter following components: 65.0 g FeS- O4^{7H2O} , 0.2 g Na2MoO4^{2H2O}, 20.0 g ZnCl2, 3.0 g MnSO4, , 105 0.92 g CoCl2^{6H2O},3.84 g CuSO4, 0.02 g H3BO3, 0.08 g NaCI and 5ml H2SO4 (69%) (Carl Roth GmbH + Co. KG, Karlsruhe, Germany). As a buffer system 6g/L K2HPO4 and 24g/L KH2PO4 were used (Carl Roth GmbH + Co. KG, Karlsruhe, Germany). For the expression of genes in *K. phaffii* the *K. phaffii* CBS7435 mutS was employed (Näätsaari et al., 2012b; Sturmberger et al., 2016a) (CBS, Utrecht, The Netherlands). All cloning work performed in this study was done in *E. coli* TOP10 F‘ (Invitrogen, Carlsbard, CA, USA). Recombinant *E. coli* and *K. phaffii* cells were cultivated under Zeocin selection conditions (Invivogen, Tolouse, France) with 25µg/mL for *E. coli* and 100µg/mL for *K. phaffii*. The cell growth of *E. coli* was realized in LB-Lennox (Carl Roth GmbH + Co. KG, Karlsruhe, Germany) at 37°C. K. phaffii cells were cultivated in YPD2% (10g/L yeast extract (Carl Roth GmbH + Co. KG, Karlsruhe, Germany), 20g/L Bactopeptone (BD Sciences, Franklin Lakes, USA) 20g/L D-glucose (Carl Roth GmbH + Co. KG, Karlsruhe, Germany)) or BM media pH 6,0 (13,4g/L YNB BD Sciences, Franklin Lakes, USA, 6g/L K2HPO4, 24g/L KH2PO4, 0,0004g/L biotin Carl Roth GmbH + Co. KG, Karlsruhe, Germany) either supplemented with 20g/L glucose or methanol (Carl Roth GmbH + Co. KG, Karlsruhe, Germany). For plate media an agar concentration of 15g/L (Carl Roth GmbH + Co. KG, Karlsruhe, Germany) was used. The competency of *E. coli* cells was guaranteed as described in (Seidman and Struhl, 2001) and stored at -80°C. 3µL of assembly mixture without any prior desalting were mixed with 80µL frozen cells, thawed on ice and transformed with 2.5 kV/25 µF/200 Ω (Bio-Rad Gene Pulser System, Bio-Rad Laboratories Inc., Hercules, CA, United States). The regeneration of cells was mediated in 1 mL SOC medium (Carl Roth GmbH + Co. KG, Karlsruhe, Germany) at 37°C and 650rpm for one hour. In a similar manner, we attained competency of *K. phaffii* cells according to the protocol of (Lin-Cereghino et al., 2005). Linearized plasmids (*SwaI*) were mixed with cells and electroporated (1.5 kV, 25 µF and 200 Ω, Bio-Rad Gene Pulser). Subsequently cells were regenerated in 1mL of a 1:1 mixture of 1M D-sorbitol (Carl Roth GmbH + Co. KG, Karlsruhe, Germany) and YPD2% for 2 hours at 28°C without shaking.

Cloning. All genes were expressed in the backbone of the previously reported pPpT4-S vector (Näätsaari et al., 2012b) under the regulation of the *K. phaffii* P_{CTA1} (Vogl et al., 2016). In addition, an autonomously replicating sequence (ARS) from *Schizosaccharomyces pombe* was used (Clyne and Kelly, 1995). For expression and regulation of the zeocin selection marker (Sh ble) the P_{TEF1} and terminator sequence from *Ashbya gossypii* was used (S Steiner and Philippsen, 1994). The genes identified through the secretome screen were codon optimized for *K. phaffii* using the IDT codon optimizer tool (IDT, Leuven, Belgium) and ordered at IDT (Integrated DNA technologies) as gBlocks. Subsequently, the entry vector was linearized with *XhoI* and resuspended gBlocks were assembled in a Gibson assembly reaction. Selection of positive transformants was mediated via zeocine (Invivogen, Tolouse, France). With any mentioning of restriction endonucleases, Thermo Scientific Fisher fast digest enzymes (Thermo Scientific Fisher, Waltham, MA, USA) were employed. The cloning work was realized via Gibson Assembly (Gibson Assembly® HiFi 1-Step Kit, SGI DNA, San Diego, CA, USA) and all plasmids sequence verfied via Sanger sequencing at Microsynth (Vienna, Austria).

RNA isolation, transcriptome sequencing and assembly. The *Lecanicillium attenuatum* strain CBS 170.76 grown in the media described above in shake flasks at 28°C for 5 days. The cell material was harvested by centrifugation at 4000 rpm for 10 minutes (Eppendorf 5810R, Hamburger, Germany). The Kit InviTrap® Spin Plant RNA Mini Kit (50 purifications: 1064100300) was used for RNA isolation. Best results were obtained by using the Lysis Solution DCT at 22°C. Afterwards the isolated RNA was analyzed on a NanoDrop (Thermo Scientific Fisher, USA) and with a Agilent Bioanalyzer RNA chip (Agilent 2100 Bioanalyzer, Agilent Technologies). The samples were sent to LGC genomics (LGC, Berlin, Germany) for library preparation and transcriptome sequencing. A normalized transcriptome library was generated for the 454 FLX sequencing platform according to the default protocol by the manufacturer (Roche, Branford, CT, USA). Concatenated fragments of about 800bps in size were prepared by shearing them randomly through nebulization. The polished fragments were subsequently adaptor ligated and sequenced on one half of a picoliterplate on the GS FLX platform. The demultiplexing of the sample was performed using Illumina's CASAVA data analysis software (Clipping of Illumina TruSeq™ adapters in all reads). Any reads with final length < 20 bases were discarded and the filtering of rRNA sequences was done by using the RiboPicker 0.4.3 (http://ribopicker.sourceforge.net/) The clipping of sequencing adapters (Illumina TruSeq and Nextera mate pair linkers) and quality trimming of all reads was executed by removal of reads containing more than one N, removal of bases or complete reads with sequencing errors, which occur on Illumina reads mainly at the 3'-end, trimming of reads at 3'-end to get a minimum average Phred quality score of 10 over a window of ten bases. In addition, reads with final length < 20 bases were discarded (if one read in a pair has been discarded, the remaining mate read was written into a separate FASTQ file for single reads. The forward and reverse reads were combined by using FLASh 1.2.4 (http://ccb.jhu.edu/software/FLASH/) with a minimum overlap of 10 bases and a maximum mismatch rate of 25 %. The de novo assembly was performed with Newbler v 2.8 (http://www.454.com/products/analysis200/3.07-2014-04software/). Peptide identification was done with TransDecoder rel16JAN2014 (http://transdecoder.sourceforge.net/). The transcriptome sequence was provided in fasta file format.

Secretome analysis. The concentrated culture supernatant was tryptically digested and dissolved in a solution with 0.1% formic acid. The peptides were separated by HPLC and ionized using a nanospray source. Analysis took place in an Orbitrap mass spectrometer (Orbi CID top20, 120min). The generated data were analyzed by querying against the translated *L. attenuatum* transcriptome with contaminations included (e.g. keratin, trypsin) with the Proteome Discoverer (version 2.0.0.802) and MASCOT (Version 2.4.1). The positive hits were subsequently subjected to BLAST alignment against the NCBI non-redundant database.

Enzyme production. *K. phaffii* clones were grown in shake flasks at 28°C and 160rpm or in 96 deep-well plates (Bel-Art Products, Wayne, NJ, United States). Cells were cultivated in 250µL of BMD1%, prior to induction with 250µL BMM1% and 50µL pulses of BMM5% every 12 hours for an additional 72 hours. The harvested cell suspension was centrifuged at 4000rpm and 4°C for 5 min to collect the supernatant min (Eppendorf 5810R, Hamburg, Deutschland). All *K. phaffii* cells were used for a rescreen in 96-well plates with at least four clones per construct in octuplicates.

Enzyme activity assays. We performed enzyme assays with commercially available methylumbelliferyl fluorescence substrates (Carbosynth, UK). Following substrates were used in this study: a-xyl: methylumbelliferyl-a-xylopyranoside, a-gal: methylumbelliferyl-a-galactopyranoside, b-xyl: methylumbelliferyl-b-xylopyranoside, b-gal: methylumbelliferyl-b-galactopyranoside, b-glu: methylumbelliferyl-b-glucopyranoside, b-cel:

methylumbelliferyl-b-cellobioside, a-glu: methylumbelliferyl-a-glucopyranoside, b-man: methylumbelliferyl-b-manopyranoside, b-fuc: methylumbelliferyl-b-fucopyranoside. 100 µM of each substrate was dissolved in 50mM sodium acetate buffer at pH 4.8. The enzymatic activity of 20µL culture supernatant was evaluated at 45°C and at 60° by mixing the supernatant with substrate containing buffer and recording the fluorescence on a Synergy H1 fluorescence plate reader (Biotek). The excitation was performed at 365nm, and fluorescence measured at 450nm. The assay was incubated at 300rpm for 2 hours at 45°C or one hour at 60°C and subsequently quenched with 100µL of 1M sodium carbonate solution. Absolute fluorescence values for each reaction were determined. For control reactions we used the supernatant of a wildtype *K. phaffii* CBS7435 mutS strain. For the determination of pH optima we used a three reagent buffer system according to (Carmody, 1956) consisting of boric acid, citric acid and tertiary sodium phosphate (Carl Roth GmbH + Co. KG, Karlsruhe, Germany). The culture supernatant was incubated at the respective pH values for 10 minutes prior to the start of the assay. For pH stability assays the supernatant was incubated at the respective pH value for 1 hour prior to fluorescence assay measurements. For the determination of temperature optima, the supernatant was incubated for 10 minutes at the respective temperature and then the enzymatic activity measured as described above.

Dot-Blot detection. The enzymes were detected with a C-terminal hexa-histidine tag. Dot-Blot assays were performed with a Bio-Dot SF Microfiltration Apparatus (Bio-Rad, Austria) (0,22µM Amersham™ Protran™ nitrocellulose membrane) by applying 50µL of culture supernatant to each well, followed by an incubation for 15 minutes and two consecutive wash steps with 50µL TBS buffer. The apparatus was disassembled, and membranes subjected to regular western blot membrane treatment. The penta-his antibody (Qiagen, Austria) was diluted 1:5000 in TBS buffer (292,7g/L NaCl, 4,24g/L Tris, 26g/L Tris-HCl, pH 7,5 Carl Roth GmbH + Co. KG, Karlsruhe, Germany). The goat anti-mouse IgG cross adsorbed HRP secondary antibody (Thermo Scientific Fisher, Austria) was diluted in a ratio of 1:20000 in TBS buffer. Prior to antibody binding the membranes were blocked with 3% BSA fraction V solution in TBS buffer (Carl Roth GmbH + Co. KG, Karlsruhe, Germany). In-between antibody incubations the membranes were washed with TBST Buffer (1ml/L Tween 20 Carl Roth GmbH + Co. KG, Karlsruhe, Germany) consecutively three times for 5 min each. The detection was performed with Pierce™ ECL Western Blotting Substrate with recording times between 30 seconds and five minutes in a Syngene G:Box (Syngene, Cambridge, UK). Dot-Blot assays were performed with a Bio-Dot SF Microfiltration Apparatus (Bio-Rad, Austria) by applying 50µL of culture supernatant to each well, followed by an incubation for 15 minutes and two consecutive wash steps with 50µL TBS buffer. The apparatus was disassembled, and membranes subjected to regular western blot membrane treatment.

Results and Discussion

The assembly of transcript data from *L. attenuatum* resulted in the identification of 11.378 transcripts with 10.981 coding for a putative protein sequence. The average contig size was 1.396 bp with a GC content of 55% and the largest contig assembled being 13.837 bps (Table 1). The median length of contigs (large contigs) was counted at 909 bps and is depicted in Figure 1. These 10.981 coding sequences formed the basis for our gene discovery efforts by generating a protein database from said sequences. The separation of tryptically digested proteins from culture supernatants via HPLC-MS/MS and their subsequent alignment to the protein database allowed the identification of 11 contig hits with at least 15 unique peptides per hit. From these contigs three showed a complete open reading frame with a designated start and stop codon located within the full transcript. Two positive hit sequences are missing the C-terminal part of the coding sequences, while five coding sequences show truncations of their N-terminal coding sequence. One out of these eleven coding sequences is internally located, with the transcript lacking both the N- and C-terminal part of the protein. The largest protein translated from the identified contigs is contig01001 with a predicted molecular weight of 95,9 kDa and a pI of 5,52. The entire contig shows a length of 2736 bps with the protein coding sequence having a length of 1648 bps. The 5'UTR region of this mRNA spans 87 bps and the 3' UTR has a length of 1001 bps, with the coding sequence of contig01001 reaching from position 88 to position 1736. Based on the MS/MS data we were able to count 19 unique peptides for contig01001 (Table 2). The protein sequence of contig01001 shows the presence of several conserved domains, amongst them the pfam01915, a Glycosyl hydrolase family 3 C-terminal domain. Other conserved domains suggest the classification as a beta-xylosidase (PLN03080), a Glucoamylase from the GH15 family (COG3387) and a beta-D-glucoside glucohydrolase (PRK15098) (Marchler-Bauer et al., 2017). The protein sequence contains 9 cystein residues. Disulfide prediction tools do either predict no bond formation amongst them (Ceroni et al., 2006) or 4 bonds in total (DiANNA http://clavius.bc.edu/~clotelab/DiANNA/).

Table 1. Summary of the transcriptome sequencing and assembly results.

Transcriptome sequencing	
Raw total reads	7.878.310
Raw read pairs	3.939.155
Adapter clipped total reads	7.877.012
Adapter clipped read pairs	3.938.506
rRNA filtered total reads	6.626.842
rRNA filtered read pairs	3.313.421
Quality trimmed total reads	6.626.806
Quality trimmed read pairs	3.313.385
Quality trimmed unpaired reads	36
Combined reads	3.219.982
Un-combinable read pairs	93.403
Transcriptome assembly	
Assembler	Newbler v 2.8
All Contig count	19.046
All Contig base count	18.178.128
Contig GC content [%]	55
Large Contig count	11.378
Large Contig base count	15.885.540
Large Contig average Contig size	1.396
Large Contig N50 size	1.754
Large Contig largest contig size	13.837

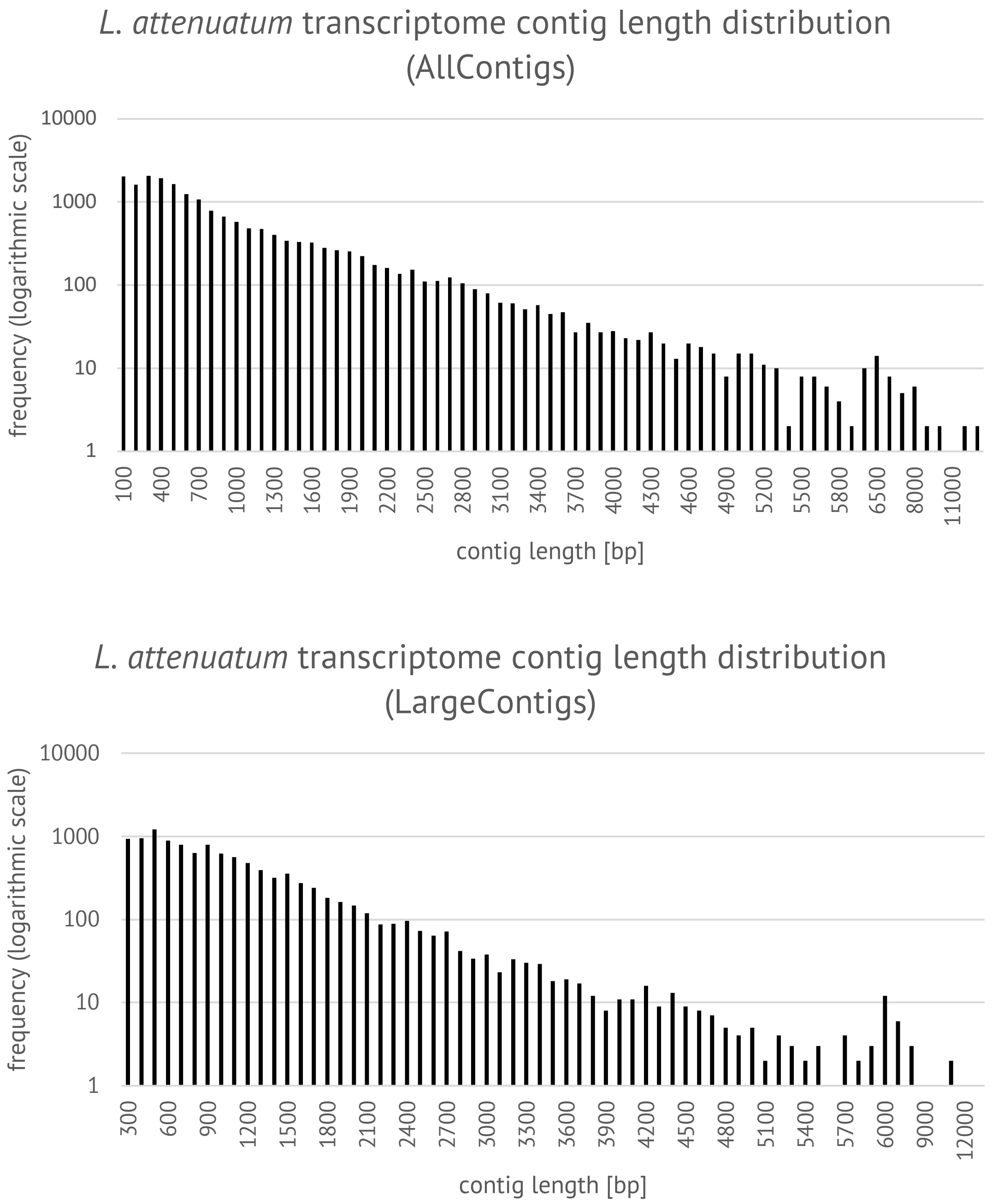


Figure 1. Contig length distribution of the L. Attenuatum transcriptome. (A) AllContigs: total number of reads: 19.046, average/median length 954/614 (B) LargeContigs: total number of reads: 11.378, average/median length 1126/909. All numeral in base pairs.

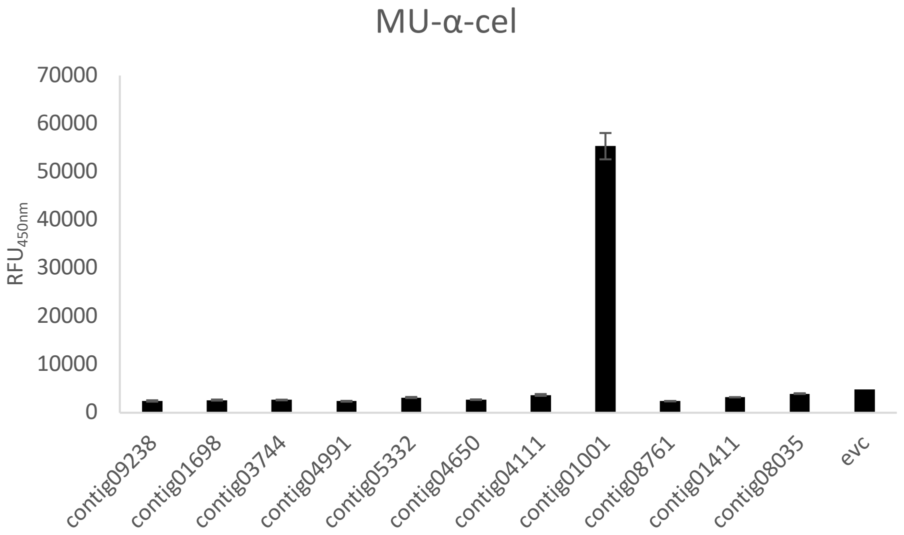
In order to be able to infer a putative enzymatic activity of the identified proteins in the secretome of *L. attenuatum* we performed BLASTp algorithm searches against the NCBI nr/nt database (Altschul et al., 1990). As can be seen in table 3 the majority of proteins showed high sequence similarity to glycoside hydrolases from different family origins as well as fungal cell wall proteins. A putative function based on sequence similarity could not be inferred for one protein, contig04650, which resulted in a hypothetical protein with no known function (Table

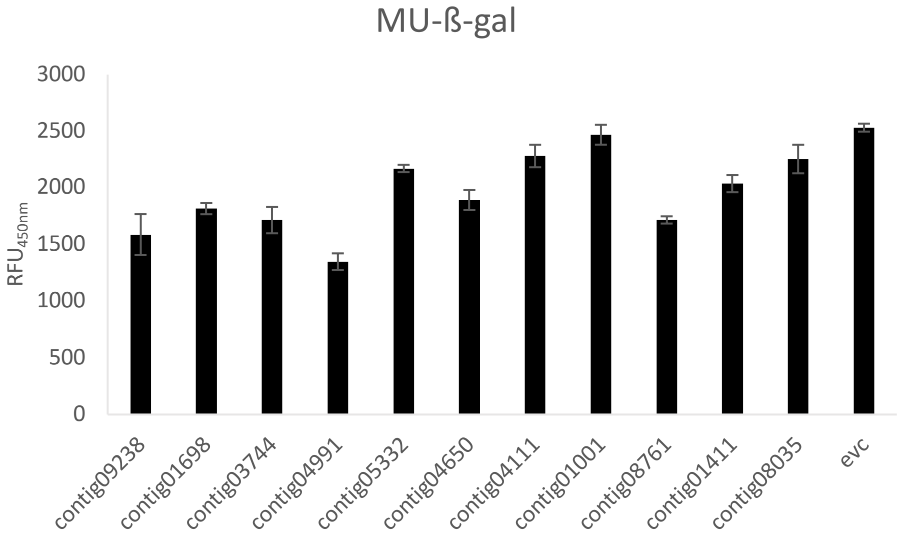
3). In general, sequence similarity showed the highest identity to proteins from the genus of Cordyceps with single proteins having high identity to proteins from different Cordyceps species (Sung et al., 2007). As this comparison is only based on *in silico* prediction of enzymatic function, we set out to express the sequences in the yeast species *K. phaffii* and test their putative enzymatic function with different model substrates. As the protein sample for analysis was derived from the supernatant of *L. attenuatum* and therefore constitutes the secretome, we performed signal peptide prediction wherever possible (complete and 3' partial open reading frames).

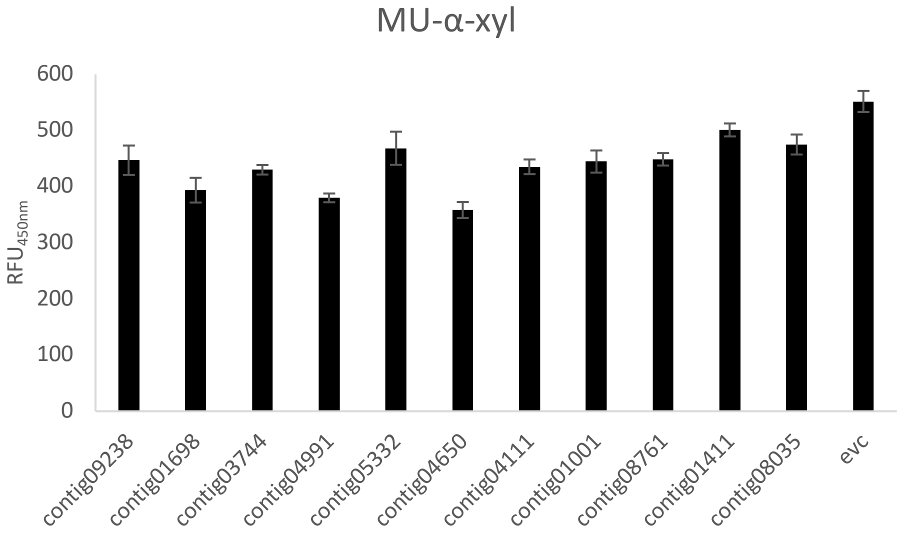
Except for contig04650, all sequences for which a full or 3' partial ORF was available showed the presence of a signal peptide for extracellular secretion (SignalP www.cbs.dtu.dk/services/SignalP/). Although contig04650 was also found in the cell culture supernatant it could have been found through cell lysis and would therefore constitute an intracellular enzyme. Indeed, the coding sequence could be partial in nature. Since the transcript is missing the 5' UTR we cannot rule out the possibility that the start ATG used for translation here is an internal methionine and the protein therefore lacks its native N-terminus. This missing part of the protein either might harbor a secretion signal or the enzyme uses a non-canonical form of secretion which cannot be detected by standard signal peptide prediction algorithms (Almagro Armenteros et al., 2019; Frank and Sippl, 2008; Horton et al., 2007; Käll et al., 2007).

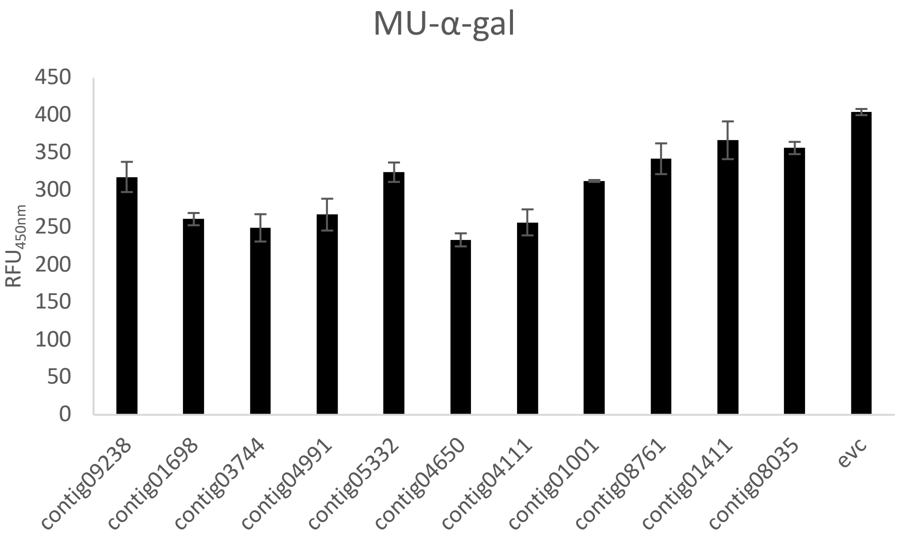
All coding sequences, complete CDSs as well as 5', 3' and internal sequences were cloned in the backbone of a pPpT4_S_alpha vector (Näätsaari et al., 2012b) under the control of the *K. phaffii CTA1* promoter. The genes were cloned downstream and in frame with the *S. cerevisiae* alpha mating protein pre-pro leader and are therefore targeted to the culture supernatant via the secretory pathway. Each target from table 3 was transformed into *K. phaffii* CBS7435 mutS (Sturmberger et al., 2016a), cultivated in minimal glucose media and methanol induced, and cell supernatants were harvested after an induction phase of 60 hours. The conducted dot-blot analyses resulted in a detection of expression for three coding sequences (contig01698, contig04650 and contig01001). For two more cases, we saw weak signals corresponding to low expression levels (contig03744, contig04991) (Figure 5). As can be seen from the dot-blot the remaining candidates could not be successfully expressed or secreted. Especially for the 5' and 3' partial sequences as well as for the internal transcript, the lack of a full open reading frame might hamper the protein's ability to fold correctly. In turn, if incorrectly folded protein accumulates within the lumen of the ER (targeted to the secretory pathway by the a-MF pre-pro leader) (Fitzgerald and Glick, 2014), the protein will most likely we recognized as misfolded and destined for proteasomal degradation. We therefore expected only full open reading frame or at least correctly folded proteins to be secreted to the culture supernatant (Raschmanová et al., 2019; Whyteside et al., 2011).

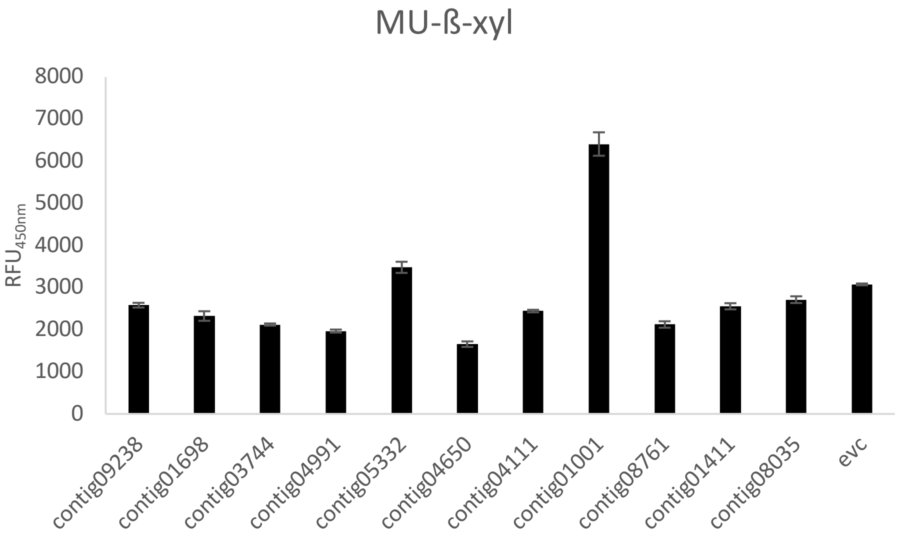
Subsequently, the cell supernatants were also employed for elucidating the enzymatic activity of the secreted proteins. As a model substrate we chose the florescence substrate group of methyl-umbelliferyls substituted with different sidechains, which give each substrate the specificity required. Through the choice of substrates, we attempted to identify the specificity for a certain sugar moiety as well as for its preferred conformation, either the alpha or beta form of the respective glycan. We tried to elucidate whether the expressed proteins show activity towards either galactopyranosides, glucopyranosides, xylopyranosides, cellobiosides, manopyranosides or fucopyranoside substrates (Bobey and Ederer, 1981), and if so, whether they prefer either the α or β form (Rye and Withers, 2000). Figure 2 shows the enzymatic activity of culture supernatants toward the respective substrates based on the fluorescence generated through cleavage of the methyl-umbelliferyl group. The analysis of enzymatic activity showed a high degree of conversion for the protein contig01001, however none of the other 10 sequences showed any enzymatic activity towards the tested substrates. Contig01001 resulted in the highest conversion with methylumbelliferyl-β-cellobioside (β-cel), as well as with methylumbelliferyl-β-glucopyranoside (β-glu). For both substrates we were not able to see any activity above the empty vector control background for the respective α conformations. The enzyme also did not show any activity towards methylumbelliferyl-β-fucopyranoside (β-fuc), methylumbelliferyl-β-manopyranoside (β-man) or methylumbelliferyl-β-xylopyranoside (β-xyl). Both conformations of the galactopyranoside substrates, methylumbelliferyl-α-galactopyranoside (α-gal) and methylumbelliferyl- β-galactopyranoside (β -gal) as well as methylumbelliferyl-α-xylopyranoside (α-xyl) showed no activity. We therefore conclude that contig01001 most likely is a hydrolase with a β-substrate specificity, with a clear preference for glucopyranosides as can be seen with MU-β-cel and MU-β-glu as preferred substrates. Contig01001 most likely is a beta-glucosidase. Any questions about a further restriction on substrate specificity, e.g. the preference for cellulosic polymers or shorter chain glucose polymers, is not possible on the basis of these data.

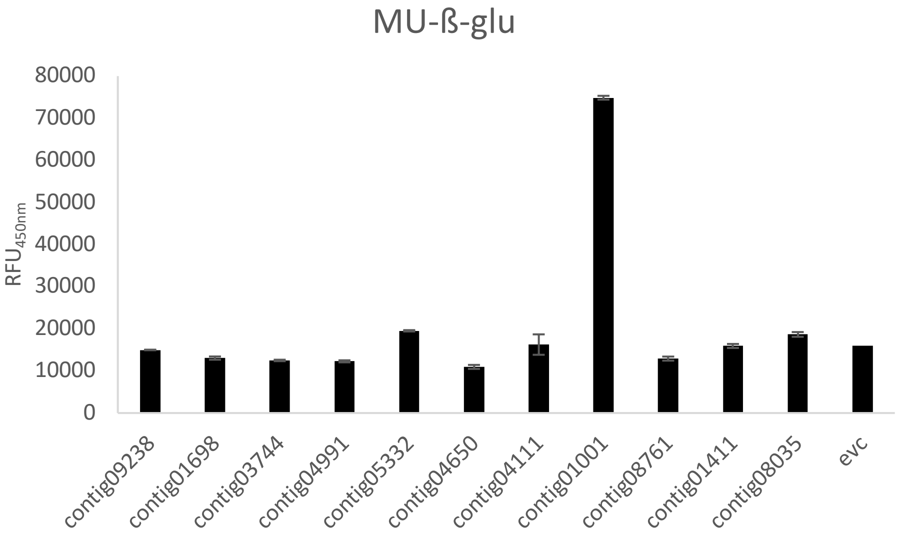

MU-α-cel
70000
60000
50000
40000
30000
20000
10000
0
RFU450nm
contig09238
contig01698
contig03744
contig04991
contig05332
contig04650
contig04111
contig01001
contig08761
contig01411
contig08035
evc

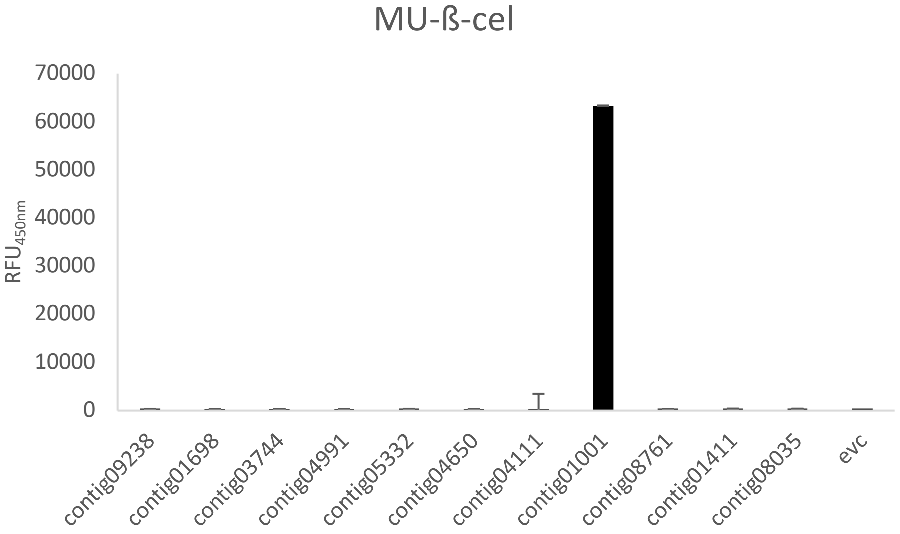

MU-ß-gal
3000
2500
2000
1500
1000
500
0
RFU450nm
contig09238
contig01698
contig03744
contig04991
contig05332
contig04650
contig04111
contig01001
contig08761
contig01411
contig08035
evc

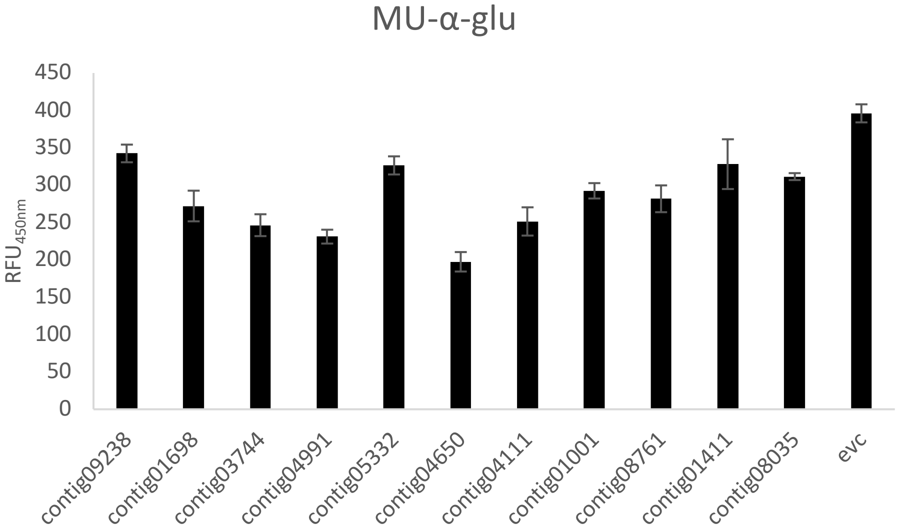

MU-α-xyl
600
500
400
300
200
100
0
RFU450nm
contig09238
contig01698
contig03744
contig04991
contig05332
contig04650
contig04111
contig01001
contig08761
contig01411
contig08035
evc

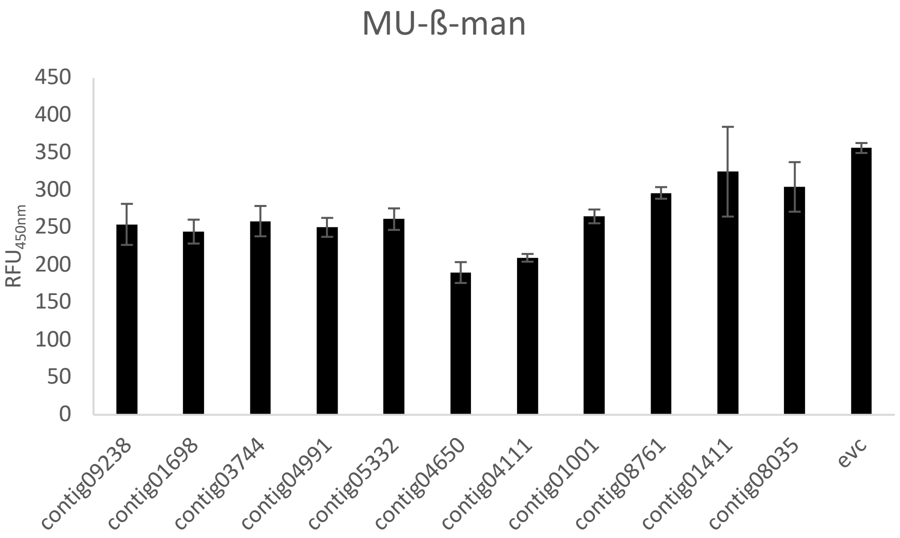

MU-α-gal
RFU$_{450nm}$
450
400
350
300
250
200
150
100
50
0
contig09238
contig01698
contig03744
contig04991
contig05332
contig04650
contig04111
contig01001
contig08761
contig01411
contig08035
evc


MU-ß-xyl
RFU$_{450nm}$
8000
7000
6000
5000
4000
3000
2000
1000
0
contig09238
contig01698
contig03744
contig04991
contig05332
contig04650
contig04111
contig01001
contig08761
contig01411
contig08035
evc


MU-ß-glu
RFU$_{450nm}$
80000
70000
60000
50000
40000
30000
20000
10000
0
contig09238
contig01698
contig03744
contig04991
contig05332
contig04650
contig04111
contig01001
contig08761
contig01411
contig08035
evc


MU-ß-cel
70000
60000
50000
40000
30000
20000
10000
0
RFU_{450nm}
contig09238
contig01698
contig03744
contig04991
contig05332
contig04650
contig04111
contig01001
contig08761
contig01411
contig08035
evc


MU-α-glu
450
400
350
300
250
200
150
100
50
0
RFU_{450nm}
contig09238
contig01698
contig03744
contig04991
contig05332
contig04650
contig04111
contig01001
contig08761
contig01411
contig08035
evc



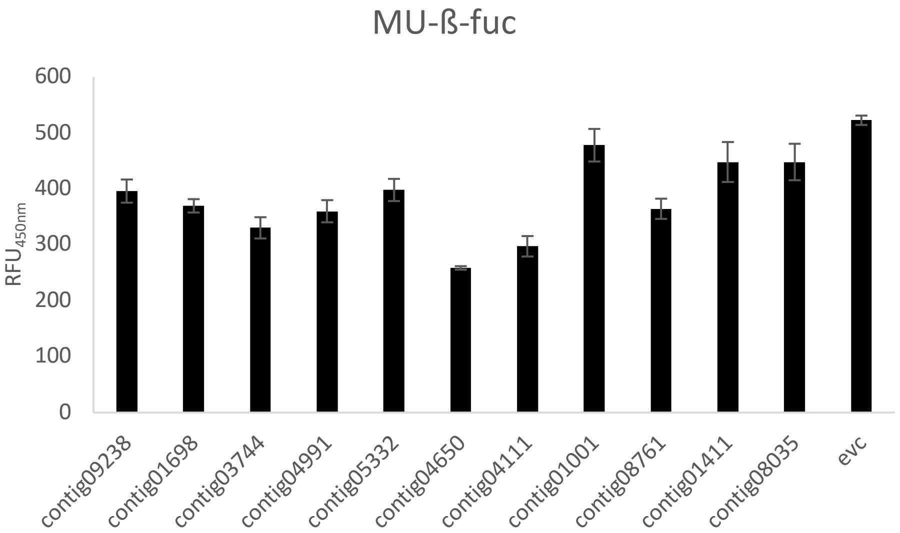


Figure 2. Enzymatic assay of strains expressing candidate enzymes listed in table 3 using various methylumbelliferyl substrates. The reactions were performed using culture supernatants in a sodium acetate buffer system with 100µM methylumbelliferyl substrate. For each reaction the supernatant of a strain expressing an empty plasmid was used as a negative control (=evc). β -gal: methylumbelliferyl-β-galactopyranoside, β-glu: methylumbelliferyl-β-glucopyranoside, β-xyl: methylumbelliferyl-β-xylopyranoside, α-xyl: methylumbelliferyl-α-xylopyranoside, α-gal: methylumbelliferyl-α-galactopyranoside, β-cel: methylumbelliferyl-β-cellobioside, α-glu: methylumbelliferyl-α-glucopyranoside, β-man: methylumbelliferyl-β-manopyranoside, β-fuc: methylumbelliferyl-β-fucopyranoside.

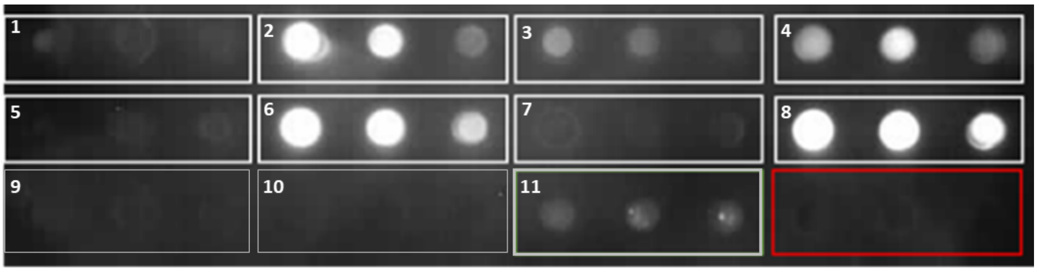


Figure 5. **Dot blot assay of culture supernatants from strains expressing secretome target proteins from table 3.** 20µL of cell supernatant from each was used for the dot-blot assay. An anti 6xHis antibody from mouse was used as primary antibody. The secondary antibody consisted of an HRP-anti mouse goat antibody fusion protein. The reaction was initiated by addition of chemiluminescent substrate (ECL) and pictures taken after 10, 30 and 60 seconds exposure. Shown here is the detection after 30 seconds. The negative control consisted of a culture supernatant from a *P. pastoris* wildtype cell (marked in red). 1 (contig09238), 2 (contig01698), 3 (contig03744), 4 (contig04991), 5 (contig05332), 6 (contig04650), 7 (contig04111), 8 (contig01001), 9 (contig08761), 10 (contig01411), 11 (contig08035).

To further characterize the identified beta-glucosidase we performed temperature and pH optima tests. Figure 3 reveals that the beta-glucosidase contig01001 optimally performs at a pH of around 5, however also shows considerable residual activity at pH 4 and pH 9 (~50% activity), sharply dropping outside of this range. Based on these results the enzyme could therefore be used for conversion between pH 4 and 9. Focusing on the temperature optimum we observed a very high value of 70°C, with 70,3% residual activity at 90°C. Interestingly, there exists a small local maximum at 40°C, although the highest conversion was attained at 70°C. The beta-glucosidase contig01001 is a thermostable hydrolase with a pH and temperature optimum at pH 5 and 70°C, respectively. Additionally, the pH and temperature stability of this beta-glucosidase was determined (Figure 4). The enzyme showed stability over a broad pH range. We were unable to measure a decline in activity between pH 4 to 9, however stability was rapidly declining outside these boundaries with no residual activity at pH 3 and 11. The stability of the enzyme was investigated at temperatures ranging between 40°C and 90°C by incubation at the respective temperature for one hour before activity determination. Contig01001 turned out to be stable up to 60°C without any loss in activity, however we were able to observe a considerable decline after this point with 48% residual activity at 80°C. No enzymatic activity could be determined at 90°C.

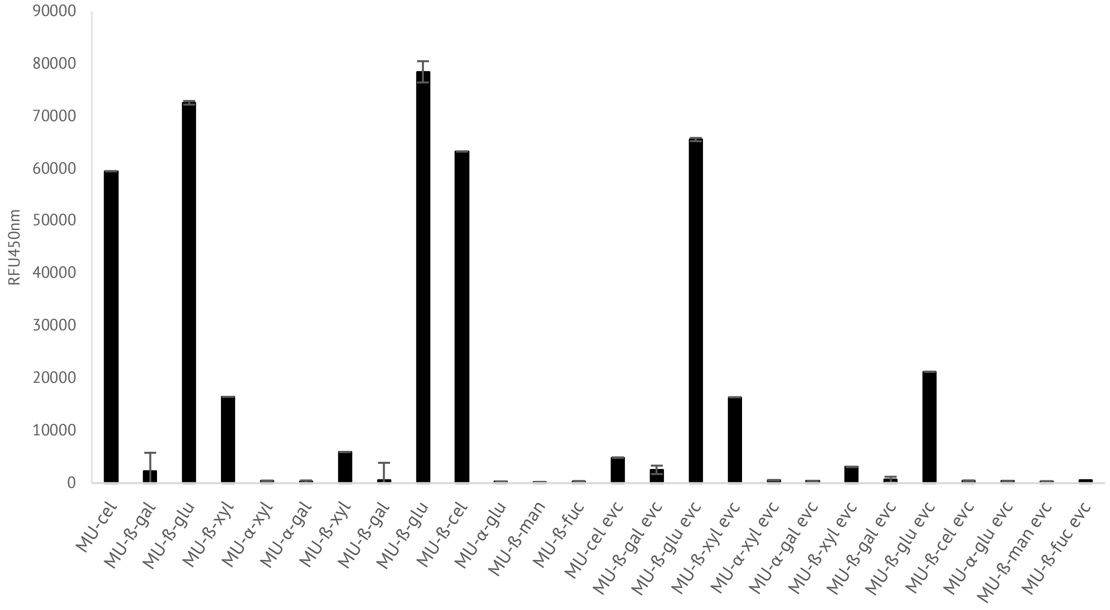


Figure 3. Enzymatic assay of contig01001 using various methylumbelliferyl substrates. The reactions were performed using culture supernatants in a sodium acetate buffer system with 100µM methylumbelliferyl substrate. For each reaction the supernatant of a strain expressing an empty plasmid was used as a negative control (=evc). a-xyl: methylumbelliferyl-a-xylopyranoside, a-gal: methylumbelliferyl-a-galactopyranoside, b-xyl: methylumbelliferyl-b-xylopyranoside, b-gal: methylumbelliferyl-b-galactopyranoside, b-glu: methylumbelliferyl-b-glucopyranoside, b-cel: methylumbelliferyl-b-cellobioside, a-glu: methylumbelliferyl-a-glucopyranoside, b-man: methylumbelliferyl-b-manopyranoside, b-fuc: methylumbelliferyl-b-fucopyranoside.

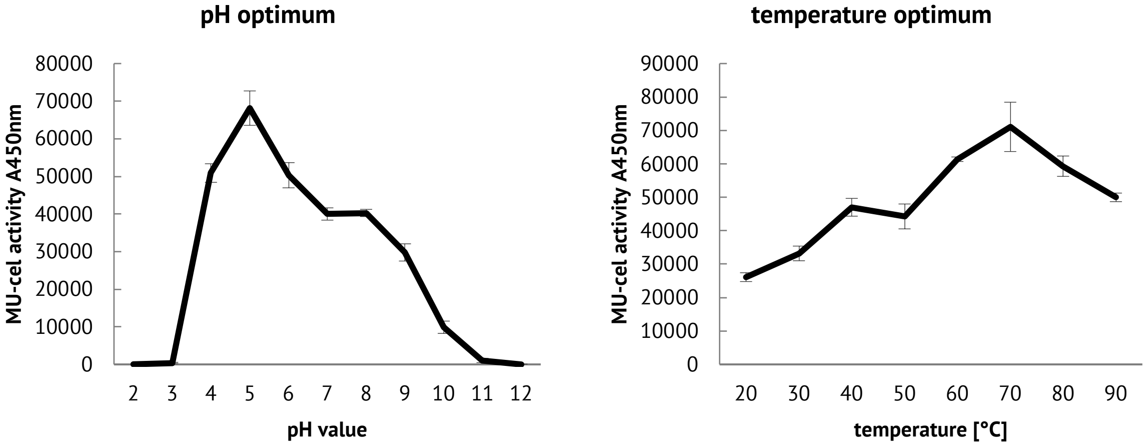


Figure 4a. Determination of pH and temperature optimum of glycoside hydrolase contig01001. The reactions were performed using culture supernatants in a three reagent buffer system according to (Carmody, 1956) consisting of boric acid, citric acid and tertiary sodium phosphate with 100µM methylumbelliferyl-β-glucopyranoside and tested at pH values between 2 and 12. For temperature optimum determination a sodium acetate buffer system with 100µM methylumbelliferyl-β-glucopyranoside was used at temperatures between 20°C and 90°C.

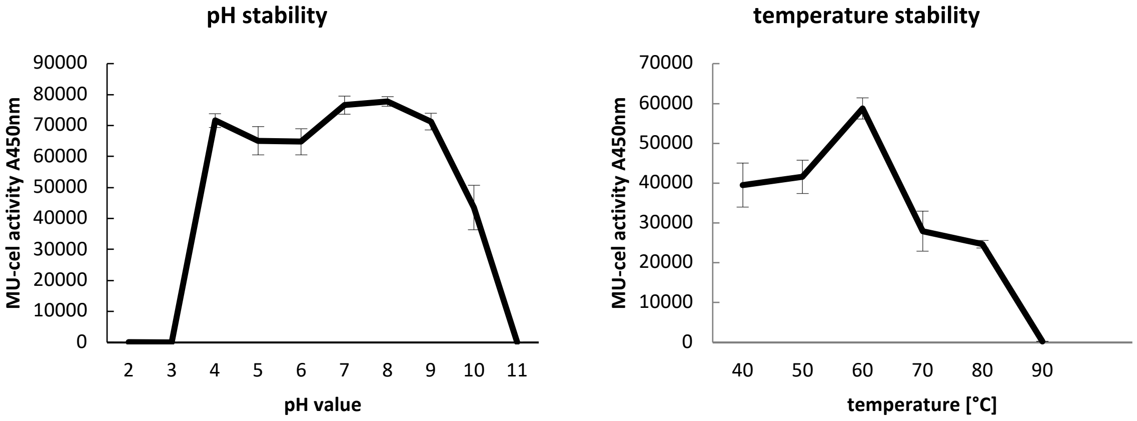


Figure 4b. Determination of pH and temperature stability. The pH stability was determined in a three reagent buffer system according to (Carmody, 1956) consisting of boric acid, citric acid and tertiary sodium phosphate with 100µM methylumbelliferyl-β-glucopyranoside. For temperature stability a sodium acetate buffer system with 100µM methylumbelliferyl-β-glucopyranoside was used at temperatures between 20°C and 90°C.

As we performed a transcriptomics and proteomics approach within this study, we were able reap the benefits derived from the synergy of both (Sturmberger et al., 2016b). The elucidation of transcriptome data on the same strain used for proteomics allowed the generation of a highly specific protein database. Standard peptide matching software uses publicly available protein databases such as NCBI nr protein database (Adav et al., 2012; Akeroyd et al., 2013; Pappin et al., 1999). This approach relies on the fact, that the protein sequence for identified peptides can be found in the database. In case the sequence is unique and not yet present in any public database the protein sequence cannot be determined precisely. Close homologues of the underlying sequence might be discovered in this way, however not the exact protein sequence in question (Amore et al., 2015; Tiwari et al., 2014). In this case, the presence of a sample specific protein database can drastically increase the likelihood of identifying the desired sequence. Table 3 exemplifies this situation, as BLASTp searches against the NCBI nr protein database resulted in hits with high sequence similarity, however none with an exact match. The resulting proteins derive from close relatives of *L. attenuatum*, but an exact match and therefore the knowledge of the precise protein sequence was only possible via peptide matching on our own created database. This exemplifies the importance of the synergy of at least two omics technologies (Ciesielska et al., 2014; Kirsch et al., 2012; Saerens et al., 2015; Schilmiller et al., 2010; Schneider et al., 2012). The availability of a genome sequence of *L. attenuatum* might have been even more promising for peptide matching as we could find only fragmented transcripts for eight out of eleven proteins. In order to employ a genome sequence for protein database generation, it can either be translated in all 6 open reading frames or a genome annotation (and CDS prediction) is available (Hernandez, Celine; Waridel, Patrice; Quadroni, 2014). In both cases, the occurrence of intron containing sequences complicates the analysis. Especially with genomes of eukaryotic origins, a genome annotation should include an intron-exon prediction step to identify spliced genes and predict their correct protein sequence (Deutsch and Long, 1999). The lack of such a prediction step might lower the quality of the underlying protein database as spliced genes could be missing from the protein sequence or could have been translated in a false frame also resulting in missing parts. The generation of a transcriptome sequence circumvents the exon-intro issue. However, as we have seen in our case, fragmented transcripts can also considerably stifle any enzyme discovery efforts. This study shows the importance of combining -omics technologies to improve proteomic experiments and enable the generation of high-quality protein sequence databases, which in turn facilitate enzyme discovery. The integration of data from several sources allow the identification of a higher number of full-length protein sequences, not relying on close homologues to be found in other databases. However, with the rising number of sequences deposited in public repositories, the likelihood of a specific sequence having already been sequenced is rising.

Table 2. Summary of peptides identified via mass spectrometric analysis. Reported here are the MASCOT results derived from the LC-MS/MS peptide spectra of the *L. attenuatum* secretome as well as the location of the corresponding coding sequence within each transcript (contig#). In addition, the molecular weight, and the pI of the coding sequences is reported.

Contig #	MW [kDa]	pI	transcript	transcript length [bp]	length cds [bp]	location within transcript [bp]	pI	# Alt. Proteins	Scores	Peptides				SC [%]	RMS90 [ppm]	Rank
										#peptides	M.expect	M.score	M.siglimit			
contig09238	16.81	5.3	5'-partial	634	449	185-634	16,80	5,30	1	261	0	261.4	13.0	7	71,81	0,89
contig01698	19.41	5.0	complete	1688	566	1122-1688	19,39	5,00	1	142	0	142.2	13.0	4	31,38	0,75
contig03744	51.58	5.84	5'-partial	1478	1397	81-1478	51,55	5,84	1	25	31.8	25.4	13.0	1	4,30	2,54
contig04991	31.12	4.77	5'-partial	942	942	1-942	31,10	4,77	1	17	221.1	17	13.0	1	3,19	0,58
contig05332	38.86	5.94	internal	1107	1107	1-1107	38,83	5,94	1	31	8.3	31.2	13.0	2	8,67	2,06
contig04650	44.02	5.78	3'-partial	1187	1184	3-1187	43,99	5,78	1	22	66.4	22.2	13.0	1	3,04	1,34
contig04111	50.01	5.74	5'-partial	1375	1319	56-1375	49,98	5,74	1	20	106	20.2	13.0	1	3,19	0,68
contig01001	95.87	5.52	complete	2736	1648	88-1736	95,81	5,52	1	19	134.5	19.1	13.0	1	1,70	0,78
contig08761	17.63	5.20	complete	607	512	95-607	17,62	5,20	1	17	238.1	16.6	13.0	1	5,88	0,48
contig01411	6.18	84.54	5'-partial	2464	2357	107-2464	84,48	6,18	1	16	287	15.8	13.0	1	1,66	0,84
contig08035	14.22	4.55	3'-partial	739	410	329-739	14,22	4,55	1	15	269.2	14.7	13.0	1	12,41	1,36

Table 3. Summary of BLASTp results from peptides/proteins identified via mass spectrometric analysis. Given here are the contig # containing the entry described in table 2. The most significant outcome of a BLASTp analysis for each protein is reported alongside with the hit description, max and total score, as well as the query coverage, sequence identity, E-value and the accession number of the respective database entry.

	BLASTp results						
contig #	**description**	**max score**	**total score**	**query cover [%]**	**E value**	**identity [%]**	**accession**
contig09238	Glycoside hydrolase, family 25 [Beauveria brongniartii RCEF 3172]	270	270	98%	1,00E-90	86.49%	OAA43861.1
contig01698	cell wall protein [Beauveria bassiana ARSEF 2860]	313	313	98%	1,00E-106	82.89%	XP_008595921.1
contig03744	beta-galactosidase [Cordyceps javanica]	855	855	99%	0.0	86.24%	TQV96700.1
contig04991	putative glucan endo-1,3-beta-glucosidase eglC [Beauveria bassiana]	394	394	94%	7,00E-133	71.06%	PMB70840.1
contig05332	glycoside hydrolase family 55 protein [Cordyceps javanica]	683	683	96%	0.0	92.39%	TQV99886.1
contig04650	hypothetical protein CCM_09573 [Cordyceps militaris CM01]	575	575	94%	0.0	75.00%	XP_006674769.1
contig04111	glycoside hydrolase family 5 [Cordyceps javanica]	845	845	97%	0.0	92.34%	TQW01025.1
contig01001	beta-glucosidase 1 precursor [Cordyceps militaris]	1680	1680	99%	0.0	91.05%	ATY61140.1
contig08761	mannose-6-phosphate isomerase, class I [Cordyceps fumosorosea ARSEF]	160	160	99%	4,00E-47	49.71%	XP_018701867.1
contig01411	glucan 1,3-beta-glucosidase [Cordyceps militaris]	1463	1463	98%	0.0	90.08%	ATY60133.1
contig08035	GPI-anchored cell wall beta-1,3-endoglucanase EglC [Cordyceps javanica]	244	244	100%	4,00E-77	89.86%	TQV96090.1

Figures and Tables

Table 1. Summary of the transcriptome sequencing and assembly results.
Table 2. Summary of peptides identified via mass spectrometric analysis.
Table 3. Summary of BLASTp results from peptides/proteins identified via mass spectrometric analysis.
Figure 1. Contig length distribution of the L. attenuatum transcriptome.
Figure 2. Enzymatic assay of strains expressing candidate enzymes listed in table 3 using various methylumbelliferyl substrates.
Figure 3. Determination of pH and temperature optimum of glycoside hydrolase contig01001.
Figure 4. Determination of pH and temperature stability.
Figure 5. Dot blot assay of culture supernatants from strains expressing secretome target proteins from table 3.

References

Adav, S.S., Ravindran, A., Cheow, E.S.H., Sze, S.K., 2012. Quantitative proteomic analysis of secretome of microbial consortium during saw dust utilization. J. Proteomics 75, 5590–5603. https://doi.org/10.1016/j.jprot.2012.08.011

Akeroyd, M., Olsthoorn, M., Gerritsma, J., Gutker-Vermaas, D., Ekkelkamp, L., van Rij, T., Klaassen, P., Plugge, W., Smit, E., Strupat, K., Wenzel, T., van Tilborg, M., van der Hoeven, R., 2013. Searching for microbial protein over-expression in a complex matrix using automated high throughput MS-based proteomics tools. J. Biotechnol. 164, 112–120. https://doi.org/10.1016/j.jbiotec.2012.11.015

Almagro Armenteros, J.J., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., Nielsen, H., 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat. Biotechnol. 37, 420–423. https://doi.org/10.1038/s41587-019-0036-z

Altschul, S.F., Gish, W., Pennsylvania, T., Park, U., 1990. Basic Local Alignment Search Tool 2Department of Computer Science 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amann, R.I., Ludwig, W., Schleifer, K.H., 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. 59, 143–169. https://doi.org/10.1016/j.jip.2007.09.009

Amore, A., Parameswaran, B., Kumar, R., Birolo, L., Vinciguerra, R., Marcolongo, L., Ionata, E., La Cara, F., Pandey, A., Faraco, V., 2015. Application of a new xylanase activity from Bacillus amyloliquefaciens XR44A in brewer’s spent grain saccharification. J. Chem. Technol. Biotechnol. 90, 573–581. https://doi.org/10.1002/jctb.4589

Baldrian, P., López-Mondéjar, R., 2014. Microbial genomics, transcriptomics and proteomics: new discoveries in decomposition research using complementary methods. Appl. Microbiol. Biotechnol. 98, 1531–1537. https://doi.org/10.1007/s00253-013-5457-x

Bobey, D.G., Ederer, G.M., 1981. Rapid detection of yeast enzymes by using 4-methylumbelliferyl substrates. J. Clin. Microbiol. 13, 393–394.

Bornscheuer, U.T., Huisman, G.W., Kazlauskas, R.J., Lutz, S., Moore, J.C., Robins, K., 2012. Engineering the third wave of biocatalysis. Nature 485, 185–194. https://doi.org/10.1038/nature11117

Carmody, W.R., 1956. An Easily Prepared Wide Range Buffer Series, J. Biol. Chem.

Ceroni, A., Passerini, A., Vullo, A., Frasconi, P., 2006. Disulfind: A disulfide bonding state and cysteine connectivity prediction server. Nucleic Acids Res. 34, 177–181. https://doi.org/10.1093/nar/gkl266

Ciesielska, K., Li, B., Groeneboer, S., Van Bogaert, I., Lin, Y.C., Soetaert, W., Van De Peer, Y., Devreese, B., 2013. SILAC-based proteome analysis of starmerella bombicola sophorolipid production. J. Proteome Res. 12, 4376–4392. https://doi.org/10.1021/pr400392a

Ciesielska, K., Van Bogaert, I.N., Chevineau, S., Li, B., Groeneboer, S., Soetaert, W., Van de Peer, Y., Devreese, B., 2014. Exoproteome analysis of Starmerella bombicola results in the discovery of an esterase required for lactonization of sophorolipids. J. Proteomics 98, 159–174. https://doi.org/10.1016/j.jprot.2013.12.026

Clyne, R.K., Kelly, T.J., 1995. Genetic analysis of an ARS element from the fission yeast Schizosaccharomyces pombe. EMBO J. https://doi.org/10.1002/j.1460-2075.1995.tb00326.x

Cravatt, B., 2014. Activity-based proteomics &#151; applications for enzyme and inhibitor discovery (357.1). FASEB J. 28, 357.1. https://doi.org/10.1096/fasebj.28.1_supplement.357.1

Davids, T., Schmidt, M., Böttcher, D., Bornscheuer, U.T., 2013. Strategies for the discovery and engineering of enzymes for biocatalysis. Curr. Opin. Chem. Biol. 17, 215–220. https://doi.org/10.1016/j.cbpa.2013.02.022

DeAngelis, K.M., Gladden, J.M., Allgaier, M., D'haeseleer, P., Fortney, J.L., Reddy, A., Hugenholtz, P., Singer, S.W., Vander Gheynst, J.S., Silver, W.L., Simmons, B. a., Hazen, T.C., 2010. Strategies for enhancing the effectiveness of metagenomic-based enzyme discovery in lignocellulolytic microbial communities. Bioenergy Res. 3, 146–158. https://doi.org/10.1007/s12155-010-9089-z

Decker, G., Wanner, G., Zenk, M.H., Lottspeich, F., 2000. Characterization of proteins in latex of the opium poppy (Papaver somniferum) using two-dimensional gel electrophoresis and microsequencing. Electrophoresis 21, 3500–3516. https://doi.org/10.1002/1522-2683(20001001)21:16<3500::AID-ELPS3500>3.0.CO;2-O [pii]\r10.1002/1522-2683(20001001)21:16<3500::AID-ELPS3500>3.0.CO;2-O

Desgagné-Penix, I., Khan, M.F., Schriemer, D.C., Cram, D., Nowak, J., Facchini, P.J., 2010. Integration of deep transcriptome and proteome analyses reveals the components of alkaloid metabolism in opium poppy cell cultures. BMC Plant Biol. 10, 252. https://doi.org/10.1186/1471-2229-10-252

Deutsch, M., Long, M., 1999. Intron – exon structures of eukaryotic model organisms 27, 3219–3228.

Erickson, B., Nelson, Winters, P., 2012. Perspective on opportunities in industrial biotechnology in renewable chemicals. Biotechnol. J. 7, 176–85. https://doi.org/10.1002/biot.201100069

Fitzgerald, I., Glick, B.S., 2014. Secretion of a foreign protein from budding yeasts is enhanced by cotranslational translocation and by suppression of vacuolar targeting. Microb. Cell Factories 13, 125. https://doi.org/10.1186/s12934-014-0125-0

Frank, K., Sippl, M.J., 2008. High-performance signal peptide prediction based on sequence alignment techniques. Bioinformatics 24, 2172–2176. https://doi.org/10.1093/bioinformatics/btn422

Hernandez, Celine; Waridel, Patrice; Quadroni, M., 2014. Database Construction and Peptide Identification Strategies for Proteogenomic Studies on Sequenced Genomes. Curr. Top. Med. Chem. 14.

Horton, P., Park, K.J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J., Nakai, K., 2007. WoLF PSORT: Protein localization predictor. Nucleic Acids Res. 35, 585–587. https://doi.org/10.1093/nar/gkm259

Illanes, A., Cauerhff, A., Wilson, L., Castro, G.R., 2012. Recent trends in biocatalysis engineering. Bioresour. Technol. 115, 48–57. https://doi.org/10.1016/j.biortech.2011.12.050

Jacobs, D.I., Gaspari, M., van der Greef, J., van der Heijden, R., Verpoorte, R., 2005. Proteome analysis of the medicinal plant Catharanthus roseus. Planta 221, 690–704. https://doi.org/10.1007/s00425-004-1474-4

Käll, L., Krogh, A., Sonnhammer, E.L.L., 2007. Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. Nucleic Acids Res. 35, 429–432. https://doi.org/10.1093/nar/gkm256

Kirsch, R., Wielsch, N., Vogel, H., Svatoš, A., Heckel, D.G., Pauchet, Y., 2012. Combining proteomics and transcriptome sequencing to identify active plant-cell-wall-degrading enzymes in a leaf beetle. BMC Genomics 13, 587. https://doi.org/10.1186/1471-2164-13-587

Lämmle, K., Zipper, H., Breuer, M., Hauer, B., Buta, C., Brunner, H., Rupp, S., 2007. Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. J. Biotechnol. 127, 575–592. https://doi.org/10.1016/J.JBIOTEC.2006.07.036

Lanfranchi, E., Pavkov-Keller, T., Koehler, E.-M., Diepold, M., Steiner, K., Darnhofer, B., Hartler, J., Van, T., Bergh, D., Joosten, H.-J., Gruber-Khadjawi, M., Thallinger, G.G., Birner-Gruenberger, R., Gruber, K., Winkler, M., Glieder, A., 2017. Enzyme discovery beyond homology: a unique hydroxynitrile lyase in the Bet v1 superfamily OPEN. https://doi.org/10.1038/srep46738

Lee, H.S., Kwon, K.K., Kang, S.G., Cha, S.S., Kim, S.J., Lee, J.H., 2010. Approaches for novel enzyme discovery from marine environments. Curr. Opin. Biotechnol. 21, 353–357. https://doi.org/10.1016/j.copbio.2010.01.015

Lin-Cereghino, J., Wong, W.W., Xiong, S., Giang, W., Luong, L.T., Vu, J., Johnson, S.D., Lin-Cereghino, G.P., 2005. Condensed protocol for competent cell preparation and transformation of the methylotrophic yeast Pichia pastoris. BioTechniques 38, 44–48. https://doi.org/10.2144/05381BM04

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C.J., Lu, S., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Wang, Z., Yamashita, R.A., Zhang, D., Zheng, C., Geer, L.Y., Bryant, S.H., 2017. CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 45, D200–D203. https://doi.org/10.1093/nar/gkw1129

McGettigan, P. a, 2013. Transcriptomics in the RNA-seq era. Curr. Opin. Chem. Biol. 17, 4–11. https://doi.org/10.1016/j.cbpa.2012.12.008

Näätsaari, L., Mistlberger, B., Ruth, C., Hajek, T., Hartner, F.S., Glieder, A., 2012. Deletion of the pichia pastoris ku70 homologue facilitates platform strain generation for gene expression and synthetic biology. PLoS ONE 7. https://doi.org/10.1371/journal.pone.0039720

Pappin, D.J.C., Creasy, D.M., Cottrell, J.S., 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data Proteomics and 2-DE. Electrophoresis 3551–3567.

Pohl, N.L., 2005. Functional proteomics for the discovery of carbohydrate-related enzyme activities. Curr. Opin. Chem. Biol. 9, 76–81. https://doi.org/10.1016/j.cbpa.2004.12.003

Raschmanová, H., Zamora, I., Borčinová, M., Meier, P., Weninger, A., Mächler, D., Glieder, A., Melzoch, K., Knejzlík, Z., Kovar, K., 2019. Single-cell approach to monitor the unfolded protein response during biotechnological processes with Pichia pastoris. Front. Microbiol. 10, 1–18. https://doi.org/10.3389/fmicb.2019.00335

Rye, C.S., Withers, S.G., 2000. Glycosidase mechanisms. Curr. Opin. Chem. Biol. 4, 573–580. https://doi.org/10.1016/S1367-5931(00)00135-6

Saerens, K., Soetaert, W., 2011. Identification of key-genes from the sophorolipid biosynthetic pathway of Candida bombicola opens a new route to increased biosurfactant yields, in: COMMUNICATIONS IN AGRICULTURAL AND APPLIED BIOLOGICAL SCIENCES. pp. 65–68.

Saerens, K.M.J., Roelants, S.L.K.W., Van Bogaert, I.N. a., Soetaert, W., 2011. Identification of the UDP-glucosyltransferase gene UGTA1, responsible for the first glucosylation step in the sophorolipid biosynthetic pathway of Candida bombicola ATCC 22214. FEMS Yeast Res. 11, 123–132. https://doi.org/10.1111/j.1567-1364.2010.00695.x

Saerens, K.M.J., Van Bogaert, I.N.A., Soetaert, W., 2015. Characterization of sophorolipid biosynthetic enzymes from Starmerella bombicola. FEMS Yeast Res. 15.

Schilmiller, a. L., Miner, D.P., Larson, M., McDowell, E., Gang, D.R., Wilkerson, C., Last, R.L., 2010. Studies of a Biochemical Factory: Tomato Trichome Deep Expressed Sequence Tag Sequencing and Proteomics. Plant Physiol. 153, 1212–1223. https://doi.org/10.1104/pp.110.157214

Schneider, T., Keiblinger, K.M., Schmid, E., Sterflinger-Gleixner, K., Ellersdorfer, G., Roschitzki, B., Richter, A., Eberl, L., Zechmeister-Boltenstern, S., Riedel, K., 2012. Who is who in litter decomposition? Metaproteomics reveals major microbial players and their biogeochemical functions. ISME J. 6, 1749–1762. https://doi.org/10.1038/ismej.2012.11

Seidman, C.E., Struhl, K., 2001. Introduction of plasmid DNA into cells. Curr. Protoc. Neurosci. Editor. Board Jacqueline N Crawley Al. https://doi.org/10.1002/0471142301.nsa01ls11

Steiner, S., Philippsen, P., 1994. Sequence and promoter analysis of the highly expressed TEF gene of the filamentous fungus Ashbya gossypii. Mol. Gen. Genet. MGG 242, 263–71. https://doi.org/10.1007/bf00280415

Sturmberger, L., Chappell, T., Geier, M., Krainer, F., Day, K.J., Vide, U., Trstenjak, S., Schiefer, A., Richardson, T., Soriaga, L., Darnhofer, B., Birner-Gruenberger, R., Glick, B.S., Tolstorukov, I., Cregg, J., Madden, K., Glieder, A., 2016a. Refined Pichia pastoris reference genome sequence. J. Biotechnol. https://doi.org/10.1016/j.jbiotec.2016.04.023

Sturmberger, L., Wallace, P.W., Glieder, A., Birner-Gruenberger, R., 2016b. Synergism of proteomics and mRNA sequencing for enzyme discovery. J. Biotechnol. 235, 132–138. https://doi.org/10.1016/j.jbiotec.2015.12.015

Sturmberger, L., Wallace, P.W., Glieder, A., Birner-Gruenberger, R., 2016c. Synergism of proteomics and mRNA sequencing for enzyme discovery. J. Biotechnol. https://doi.org/10.1016/j.jbiotec.2015.12.015

Sung, G.H., Hywel-Jones, N.L., Sung, J.M., Luangsa-ard, J.J., Shrestha, B., Spatafora, J.W., 2007. Phylogenetic classification of Cordyceps and the clavicipitaceous fungi. Stud. Mycol. 57, 5–59. https://doi.org/10.3114/sim.2007.57.01

Tiwari, R., Singh, S., Singh, N., Adak, A., Rana, S., Sharma, A., Arora, A., Nain, L., 2014. Unwrapping the hydrolytic system of the phytopathogenic fungus Phoma exigua by secretome analysis. Process Biochem. 49, 1630–1636. https://doi.org/10.1016/j.procbio.2014.06.023

Uchiyama, T., Miyazaki, K., 2009a. Functional metagenomics for enzyme discovery: challenges to efficient screening. Curr. Opin. Biotechnol. 20, 616–622. https://doi.org/10.1016/j.copbio.2009.09.010

Uchiyama, T., Miyazaki, K., 2009b. Functional metagenomics for enzyme discovery: challenges to efficient screening. Curr. Opin. Biotechnol. 20, 616–622. https://doi.org/10.1016/j.copbio.2009.09.010

Van Bogaert, I.N.A., Saerens, K., De Muynck, C., Develter, D., Soetaert, W., Vandamme, E.J., 2007. Microbial production and application of sophorolipids. Appl. Microbiol. Biotechnol. 76, 23–34. https://doi.org/10.1007/s00253-007-0988-7

Vogl, T., Sturmberger, L., Kickenweiz, T., Wasmayer, R., Schmid, C., Hatzl, A.M., Gerstmann, M.A., Pitzer, J., Wagner, M., Thallinger, G.G., Geier, M., Glieder, A., 2016. A Toolbox of Diverse Promoters Related to Methanol Utilization: Functionally Verified Parts for Heterologous Pathway Expression in Pichia pastoris. ACS Synth. Biol. https://doi.org/10.1021/acssynbio.5b00199

Whyteside, G., Mat, R., Alcocer, M.J.C., Archer, D.B., 2011. Activation of the unfolded protein response in Pichia pastoris requires splicing of a HAC1 mRNA intron and retention of the C-terminal tail of Hac1p. FEBS Lett. 585, 1037–1041. https://doi.org/10.1016/j.febslet.2011.02.036

Zulak, K.G., Khan, M.F., Alcantara, J., Schriemer, D.C., Facchini, P.J., 2009. Plant Defense Responses in Opium Poppy Cell Cultures Revealed by Liquid Chromatography-Tandem Mass Spectrometry Proteomics. Mol. Cell. Proteomics 8, 86–98. https://doi.org/10.1074/mcp.M800211-MCP200

Research Article

Refined *Pichia pastoris* reference genome sequence

Lukas Sturmberger[a,1], Thomas Chappell[e,1], Martina Geier[a], Florian Krainer[d], Kasey J. Day[f], Ursa Vide[d], Sara Trstenjak[d], Anja Schiefer[a], Toby Richardson[g], Leah Soriaga[g], Barbara Darnhofer[a,b,c], Ruth Birner-Gruenberger[a,b,c], Benjamin S. Glick[f], Ilya Tolstorukov[e,h], James Cregg[e,h], Knut Madden[e], Anton Glieder[a,d,i]

[1] authors contributed equally
[a] Austrian Center of Industrial Biotechnology (ACIB), Petersgasse 14, 8010 Graz, Austria
[b] Institute of Pathology, Research Unit Functional Proteomics and Metabolic Pathways, Medical University of Graz, Stiftingtalstrasse 24, 8010 Graz, Austria
[c] Omics Center Graz, BioTechMed-Graz, Stiftingtalstrasse 24, 8010 Graz, Austria
[d] Institute of Molecular Biotechnology, Graz University of Technology, Petersgasse 14, 8010 Graz, Austria
[e] BioGrammatics Inc., 2120 Las Palmas Drive, Carlsbad, CA 92011, United States
[f] Department of Molecular Genetics and Cell Biology, University of Chicago, 920 East 58th St., Chicago, IL 60637, United States
[g] Synthetic Genomics, Inc., 11149 North Torrey Pines Rd., La Jolla, CA 92037, United States
[h] Keck Graduate Institute, 535 Watson Drive, Claremont, CA 91711, United States
[i] bisy e.U., Wetzawinkel 20, 8200 Hofstaetten/Raab, Austria

Abbreviations

LC-MS/MS	liquid chromatography tandem mass spectrometry
ORF	open reading frame
SIBIA	Salk Institute Biotechnology/Industrial Assosciate
BLAST	basic local alignment search tool
NCBI	National Center for Biotechnology Information
HMM	Hidden Markov Model
SGI	Synthetic Genomics Inc.
LTR	long terminal repeat
PCR	polymerase chain reaction
CUT	cryptic unstable transcript
CBS	Centraalbureau voor Schimmelcultures, Utrecht

Keywords

P. pastoris, genome, NGS, killer-plasmid, centromere, K.phaffii

This manuscript was published as Sturmberger, L. *et al.* (2016) 'Refined Pichia pastoris reference genome sequence', *Journal of Biotechnology*. doi: 10.1016/j.jbiotec.2016.04.023.

Abstract

Strains of the species *Komagataella phaffii* are the most frequently used "*Pichia pastoris*" strains employed for recombinant protein production as well as studies on peroxisome biogenesis, autophagy and secretory pathway analyses. Genome sequencing of several different *P. pastoris* strains has provided the foundation for understanding these cellular functions in recent genomics, transcriptomics and proteomics experiments. This experimentation has identified mistakes, gaps and incorrectly annotated open reading frames in the previously published draft genome sequences. Here, a refined reference genome is presented, generated with genome and transcriptome sequencing data from multiple *P. pastoris* strains. Twelve major sequence gaps from 20 to 6000 base pairs were closed and 5111 out of 5256 putative open reading frames were manually curated and confirmed by RNA-seq and published LC–MS/MS data, including the addition of new open reading frames (ORFs) and a reduction in the number of spliced genes from 797 to 571. One chromosomal fragment of 76 kbp between two previous gaps on chromosome 1 and another 134 kbp fragment at the end of chromosome 4, as well as several shorter fragments needed re-orientation. In total more than 500 positions in the genome have been corrected. This reference genome is presented with new chromosomal numbering, positioning ribosomal repeats at the distal ends of the four chromosomes, and includes predicted chromosomal centromeres as well as the sequence of two linear cytoplasmic plasmids of 13.1 and 9.5 kbp found in some strains of *P. pastoris*.

Introduction

Methanol utilizing yeast isolates from the Yosemite region of California were used to establish the species *Pichia pastoris*, and selected by Phillips Petroleum for the large-scale production of single cell protein. Subsequently, *P. pastoris* clones capable of high cell density growth on simple defined medium in 100,000 l fermenters were deposited into the yeast culture collections by Phillips Petroleum for patent protection; and work was initiated in collaboration with scientists at the Salk Institute/Biotechnology Associate, SIBIA, to use these *P. pastoris* strains for the expression of recombinant proteins (Cregg et al., 2009). In 2009, a reclassification dictated that the *P. pastoris* strains most commonly used around the world for protein production now belong to the species *Komagataella phaffii*, and include the strains: NRRL Y-11430 from the Agriculture Research Service culture collection (Peoria IL, USA), and NRRL Y-48124 (X-33, Invitrogen expression kit strains, Carlsbad CA, USA) (Kurtzman, 2009, 2005). The same strain deposited in Peoria, IL, as NRRL Y-11430 was also deposited in Utrecht

(The Netherlands) as CBS7435. Although the genome sequence of the *K. phaffii* type strain NRRL Y-7556 (=CBS2612) is not yet known, the first draft *P. pastoris* genome (De Schutter et al., 2009), and subsequent CBS7435 genomic data (Küberl et al., 2011) have accelerated *P. pastoris* research. Both sequenced *P. pastoris* strains, like most gene expression studies, build on the NRRL Y-11430/CBS7435 strain or strains directly derived from those. For example, *P. pastoris* GS115 was derived from the NRLL Y11430 strain by chemical mutagenesis and selection for histidine auxotrophy (US Patent 4,879,231 A) and became one of the most frequently used *P. pastoris* strains. More recently we reported the construction of an alcohol oxidase (*AOX1*) gene knock out (mutS) variant of the CBS7435 strain by homologous recombination and marker recycling employing an FRT/flipper recombinase-based strategy (Näätsaari et al., 2012). However, initial sequence data from this CBS7435 strain (*aox1*-) did not match the published draft genome sequence as expected (i.e. outside of the AOX1 deletion). Although most data of RNA-seq experiments (Liang et al., 2012) and LC–MS/MS based proteomics (Renuse et al., 2014) mapped to predicted open reading frames(ORFs) of the published draft genomes, many were miscalled and additional new ORFs and alternative splice sites were recently identified. Here, state-of-the-art sequencing technologies including long read sequencing is used in resequencing the genome of *P.pastoris* CBS7435 mutS. With the advent of next generation sequencing technologies, Sanger-based shotgun sequencing was replaced by massively parallel, short read sequencing methods such as ABI’s SOLiD or Illumina’s Solexa platforms. While this development allowed for higher base coverage and cheaper per base sequencing costs, the assembly of short reads to generate full length genome sequences remained challenging. Current draft genomes (English et al., 2012) including the genomes of *P. pastoris* CBS7435 (Küberl et al., 2011) and *P. pastoris* GS115 (De Schutter et al., 2009) reflect these previous limitations which can be observed in gaps, insertions, deletions and rearrangements. Typically, repetitive genome features, skewed GC distributions and other genomic complexities limit the methods used for genome sequencing and assembly, giving rise to such errors (Quail et al., 2012; Roberts et al., 2013). Currently Pacific Biosciences (PacBio) single molecule, real-time sequencing technology, SMRT, provides an alternative. SMRT sequencing is a sequencing-by-synthesis approach based on the real-time imaging of fluorescently tagged nucleotides which are incorporated by a polymerase affixed at the bottom of a zero-mode waveguide (ZMW) well (Mccarthy, 2010). The advantages offered by this sequencing technology are two-fold. Generally, the average read length of the PacBio RS platform is 8 kb–15 kb. The availability of such long reads acting as anchoring sequences substantially improves eukaryotic genome assemblies and therefore the generation of high-quality full-length genome sequences. Compared to other techniques, PacBio is limited by modest per base throughput and a high error rate of approximately 13% observed in raw reads (Quail et al., 2012); these errors are however corrected for by the increased sequencing depth offered by shorter reads present in the sequencing reaction. Here new genomic sequence data from multiple closely related *P. pastoris* strains were combined to provide a new reference genome for this powerful eukaryotic expression system. For the first time, de novo sequencing of *P. pastoris* strains has been performed employing the Pacific Biosciences RSII platform (PacBio). In relation to the genome sequence published in 2011 (Küberl et al., 2011) deletions, insertions, repeats and larger inversions have been identified. By integrating the PacBio derived de novo genome sequence with new Illumina HiSeq data, as well as more traditional Sanger sequencing data, a first complete *P. pastoris* reference genome sequence has been generated.

Materials and methods

Strains. Strains used in this study are listed in Table 1.2.2. Strain cultivation and DNA extraction *P. pastoris* CBS7435 mutS (Näätsaari et al., 2012) and related strains overexpressing the beta-carotene biosynthesis pathway from *Pantoea ananatis* (Geier et al., 2015) were grown overnight in 50 ml YPD medium (20 g/L peptone, 10 g/L yeast extract, 20 g/L dextrose) at 28◦C and 120 rpm 1.5 × 10^9 cells were removed from the supernatant by centrifugation and washed once with TE Buffer before resuspension in 1 ml yeast lysis buffer (1 M sorbitol, 100 mM EDTA, 14 mM beta-mercaptoethanol). Spheroplasts were generated by addition of 100 µL of a zymolyase stock solution (1000 U/ml) and incubation of the suspension at 30°C for 30 min. Spheroplasts, pelleted by centrifugation at 3220g, 4°C for 10 min were resuspended in 2 ml digestion buffer (800 mM guanidine HCl, 30 mM Tris-HCl, pH 8, 30 mM EDTA, 5% Tween-20 and 0.5% Triton X-100) supplemented with RNase A. 45 µl of a Proteinase K stock solution (21.4 mg/ml) were then added to the suspension followed by incubation at 50°C for 30 min. Cellular debris were removed by centrifugation at 3220g (4°C for 10 min) and genomic DNA was then isolated from the obtained supernatant using Genomic tips 20/G (Qiagen, Hilden, Germany) according to the manufacturer´s instructions. The concentration and quality of the isolated gDNA was determined spectrophotometrically and via agarose gel electrophoresis. P. pastoris strain BG08 (BioGrammatics Inc., Carlsbad; CA, USA) is a single colony isolate from the Phillips Petroleum strain NRRLY-11430 obtained from the Agriculture Research Service culture collection. For genomic DNA isolation, zymolyase, Proteinase K and RNase A were used. P. pastoris BG10 (BioGrammatics Inc., Carlsbad, CA, USA) was derived from BG08 using Hoechst dye selection to remove cytoplasmic killer plasmids.

Table 1. *P. pastoris* strains used in this study.

strain	description	reference
P. pastoris CBS7435	wildtype strain received from CBS	(Küberl et al., 2011)
P. pastoris CBS7435 mutS	*AOX1* knockout derived from *P. pastoris* CBS7435	(Näätsaari et al., 2012)
P. pastoris CBS7435 Δ*das1/das2*	*das1*/das*2* double knockout derived from *P. pastoris* CBS7435	(Geier et al., 2015a)
P. pastoris CBS7436 *crtEBIY*	β-carotene producing strain derived from *P. pastoris* CBS7435	(Geier et al., 2014)
P. pastoris PPY12	*his4 arg4* auxotrophic strain	(Gouldi et al., 1992)
P. pastoris BG08	BioGrammatics Inc.	
P. pastoris BG10	BioGrammatics Inc.	Cat. No. PS001-01

Strain cultivation and RNA extraction. The wildtype *P. pastoris* strain CBS7435, as well as related strains with deletions of the dihydroxyacetone synthase (Δdas1, Δdas2, Δdas1/das2) (Geier et al., 2015), were cultivated for 24 hours in BMD2% (200 mM KPi, pH 6.0, 20 g/L dextrose, 13.4 g/L yeast nitrogen base and 0.4 mg/L biotin) at 28°C and 100 rpm. Cells were harvested by centrifugation (1000g, 5 min) and used to inoculate 200 ml of BMM (as BMD2% but supplemented with 0.5% methanol instead of dextrose) for growth to an OD600 of 8 at 28°C and 100 rpm. Samples for RNA-seq analysis (150–300 mg wet cell weight, wcw) were drawn after 24 h growth on glucose and 5 h growth on methanol; all cell samples were immediately frozen in liquid nitrogen after removing the supernatant. Total RNA samples were prepared in duplicate from 8 samples using a Fast RNA Yeast SPIN kit (MP Biomedicals, Santa Ana, CA, USA). Briefly, cell disruption was performed with 3 × 2 min bursts, using a BioSpec Products Mini–Beadbeater–96 (Bartlesville, OK, USA) and purified RNA samples were flash frozen in liquid nitrogen before storage at −80°C. cDNA libraries were constructed using Illumina TruSeq stranded mRNA library preparation kits and sequenced on an Illumina HiSeq2500 platform. TruSeq libraries of two additional *P. pastoris* strains expressing the beta-carotene biosynthetic pathway, either regulated by the constitutive *GAP* promoter or the inducible *AOX1* promoter, were also analyzed by RNA-seq; these samples were similarly drawn after 48 h growth on glucose and after 24 h of methanol induction, respectively. Additional libraries were prepared from BG10 strains expressing a variety of heterologous proteins, both intracellular and secreted. In total, 57 RNA-seq libraries were created and sequenced.

Sequencing and genome assembly. PacBio sequencing was performed on 10 µg of DNA by GATC Biotech (Konstanz, Germany). Preparation of a large insert library and subsequent sequencing was performed on a PacBio RS II instrument. No manual filtering of sequence reads was attempted and the assembly could be done by using the HGAP 3 based de-novo assembly protocol with standard settings, except for p assembleunitig.genomeSize = 9000000 andp assembleunitig. × Coverage = 15. The software was provided by GATC in the SMRT Portal Version 2.2.0. For *P. pastoris* BG08 and BG10 paired-end genomic DNA sequencing was performed by GeneWiz Inc. (New Jersey, USA) on an Illumina HiSeq 2500 platform. The de novo assembled reference genome sequence was evaluated by comparison with previously generated wildtype sequence data (Küberl et al., 2011). In order to obtain the full genome sequence, the deleted AOX1 ORF of the sequenced mutS strain was adjusted by in silico complementation into the PacBio genome assembly employing CLC Genomics Workbench software (Qiagen, Hilden, Germany).

Sequencing and transcriptome mapping. Library preparation was performed at the University of California, San Diego, Institute for Genomic Medicine. RNA samples were analyzed on an Agilent 2200 Tape Station to visualize intact ribosomal RNA prior to library preparation. Libraries were pre-pared using an Illumina TruSeq mRNA preparation kit and libraries were barcoded for multiplexing during sequencing. Subsequently, the libraries were size selected for >200 bp, with modes of approximately 300 bp and 50 cycles of single-end sequencing was performed on an Illumina HiSeq 2500 machine. Data was provided in standard FASTQ (Cock et al., 2009) format and TopHat2 (Trapnell et al., 2009) was used to map reads to the PacBio genome assembly.

Gene prediction and functional sequence annotation. A de novo transcriptome assembly was generated from the single-end strand-specific RNA-seq library using Trinity (Trinity version: trinityrnaseq r20140717) (Manfred G. Grabherr, Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W. and Friedman, 2013). Open reading frames (ORFs) were identified in the assembled transcripts using the Perl script, transcripts to best scoring ORFs pl (Trinity version: trinityrnaseq r2012-01-25p1). Consensus gene models were flagged in assembled transcripts using a combination of the predicted ORFs and homology-based annotation as evidence. Blast matches were made to the non-redundant database at NCBI, and Hidden Markov Models (HMM) matches were made using hmmer3 (Finn et al., 2011; Sonnhammer ELL, 1998) to models from PFAM (Finn et al., 2015) and TIGRfam (Haft et al., 2003). These consensus gene models, based on the de novo transcriptome assembly were in turn mapped to the *P. pastoris* genome assembly using gmap. These genome-mapped transcriptome-derived gene models were then used to improve the SGI/Archetype® eukaryotic gene prediction pipeline. The SGI/Archetype® eukaryotic gene prediction pipeline is divided into two primary components – the first trains a HMM for gene prediction and the second uses the trained HMM and any supporting gene evidence (e.g., genome-mapped transcriptome-derived gene models) to predict the final set of genes. A high-quality training set of *P. pastoris* gene models was obtained from the following series of steps: (1) blastx search of the input genomic sequences against eukaryotic protein sequences in UniProt (Consortium, 2015), (2) GeneWise (Birney Clamp, M., Durbin, R, 2004) generation of more precise genomic alignments, with splice sites as needed, from the blastx matches, and (3) filtering of the GeneWise output to ensure that predicted coding sequences each have a valid start codon, a valid stop codon, no inner stop codons, overlap with a transcriptome-derived gene model. The resulting filtered gene models were then used to train an HMM for the AUGUSTUS gene prediction algorithm (Grabherr et al., 2011; Stanke et al., 2008). AUGUSTUS was then used to predict genes using the trained HMM, the original input genomic sequences, and evidence provided in the form of transcriptome-derived gene models, protein alignments, and negative evidence in the form of internal data (which provides regions that are likely RNA non-coding and are therefore not good candidates for gene coding regions). Finally, any full-length transcriptome-derived gene models mapped to regions in the genomic sequence without any AUGUSTUS-predicted genes were added to the final catalog of predicted genes. Having generated the ORFs they were annotated through the Archetype® annotation pipeline (Robson et al., 2015).

Identification of *P. pastoris* killer plasmids. In addition to the assemblies described above, a Velvet *de novo* assembly (Zerbino and Birney, 2008) of paired-end Illumina data from the *P. pastoris* BG08 strain and PacBio sequence data of the *P. pastoris* CBS7435 mutS strain resulted in the discovery of two high coverage contigs with homology to *K. lactis* killer plasmid sequences (Schickel et al., 1996). Subsequently the ORFs on both plasmids were identified by manually checking for ORFs and gene by gene blastp. Both assembled contigs were flanked by inverted repeats.

Analysis of intron splicing. The TopHat 2.1.0 software tool was used to map individual RNA-seq data sets to the genome sequence presented here with intron length limited to 3000 bases. In this analysis, typically >98% of reads aligned to the genome sequence. A custom BioRuby (Goto et al., 2010) script was then employed to combine all junction files and identify introns with GT—AG ends. All TopHat alignments were then rerun, this time forcing the use of only GT—AG introns with the "–no-novel-juncs" option. In all cases, forcing alignments to predetermined splicing sites resulted in an ~0.1% increase in mapping. The resulting BAM files were filtered using SAMtools (Li et al., 2009) to regions spanning the predetermined splicing sites. The filtered BAM files for all experiments were combined and a mpileup output was generated. Subsequently a custom BioRuby script was used to analyze the mpileup output and to determine the splicing density at each predicted spliced nucleotide.

Identification of centromere regions. The *P. pastoris* strain used to visualize centromeres was a derivative of PPY12 (his4 arg4) (Gouldi et al., 1992). To label the endoplasmic reticulum, this strain was transformed with a HIS4 integrating vector encoding DsRed. T1-HDEL as previously described (Bevis et al., 2002). To label Cse4, PPY12 genomic DNA was used as a PCR template to amplify the CSE4 gene, including 589 bp of upstream sequence and 294 bp of downstream sequence, and this fragment was inserted into the polylinker of a pUC19 derivative containing the *Saccharomyces cerevisiae* ARG4 gene (Rossanese et al., 1999). In-Fusion cloning was then used to generate a chimeric gene encoding msGFP (Fitzgerald and Glick, 2014) fused to the C-terminus of Cse4 with an intervening GSS-GSSGSSGSS linker. This construct was linearized with *SpeI* for integration at the CSE4 locus, resulting in a tandem duplication of CSE4 in which one copy of the gene was fused to msGFP. Fluorescence microscopy was performed as previously described (Papanikou et al., 2015). In brief, cells grown to logarithmic phase in a non-fluorescent minimal medium were compressed beneath a coverslip, and a Z-stack of images in red, green, and transmitted light channels was collected using a Leica SP5 confocal micro-scope. This Z-stack was then deconvolved and average projected. Planning and simulation of cloning procedures, visualization of chromosome organization, and identification of inverted repeats were performed using SnapGene or SnapGene Viewer software (GSL Biotech, Chicago, IL).

Results and discussion

Sequencing and assembly of the *P. pastoris* CBS7435 genome

Pacific Bioscience's single, molecule real-time sequencing platform (SMRT) enabled the de novo sequencing and assembly of the methylotrophic yeast strain *P. pastoris* CBS7435 mutS genome. Sequencing of a PacBio RS II library with an insert size of 8–12 kbp in a 1 movie run mode resulted in 185,064 sequence reads with 948,348,000 sequenced bases. The *de novo* assembly of PacBio sequence reads resulted in the identification of 31 unitigs with four large unitigs of size 2.9 Mbp, 2.4 Mbp, 2.3 Mbp and 1.8 Mbp (Table 2). Focus is on these four large unitigs since most of the remaining unitigs correspond to fragmented mitochondrial and killer plasmid DNA generated during the fragmentation and size selection of PacBio libraries. By including the mitochondrial DNA sequence published in 2011, as well as, the two killer plasmid sequences, and remapping the PacBio raw reads onto this combinatorial data, 99.96% of all reads are correctly aligned. Two unitigs of13.1 kbp and 9.5 kbp bearing homology to *Kluveromyces lactis* killer plasmid sequences were also identified (Schickel et al., 1996). Based on a comparison with published genome data of different *P. pastoris* wildtype strains (De Schutter et al., 2009; Küberl et al., 2011), and by using blast algorithm to align all four unitigs to the *P. pastoris* CBS7435 genome, the four large unitigs correspond to the four chromosomes of *P. pastoris.* The entire length of the *P. pastoris* genome sequence presented here is 9.38 Mbp. For the first time, a reference genome with four un-gapped chromosomes, telomere sequences on each chromosome and ribosomal repeats is created. The chromosomes of different yeast species, including *P. pastoris*, show regions of rDNA tandem repeats, variable in number and located at the end of chromosomes (De Schutter et al., 2009; Küberl et al., 2011; Rustchenko et al., 1993). Here, the new genome sequence data orient the chromosomes so that these ribosomal repeats are at the distal "end" of each chromosome. This is done to stabilize the more proximal genome and annotation numberings employed for cataloging the genetic information. This results in "flipping" chromosome 1 relative to the first published draft genome (Fig. 1, (Küberl et al., 2011)). Moreover, chromosomal rearrangements ranging from as few as 1 kbp to 134 kbp were found. One chromosomal fragment of 76 kbp between two previous gaps on chromosome 1, 134.2 kbp fragment at the end of chromosome 4, as well as, several shorter fragments of 2–3 kbp were reoriented (De Schutter et al., 2009; Küberl et al., 2011). The genome sequence of *P. pastoris* CBS7435 published in 2011 (Küberl et al., 2011) shows 12 gaps ranging from 20 bp to 6 kbp insize.

Table 2. Summary of all major unitigs identified in the assembly of *P. pastoris* genomic DNA. The length, mean coverage and all protein coding sequences of the CBS 7435 mutS strain are presented here. Assembly metrics can be found in supplementary figure S1.

unitig	designation	length [bp]	mean coverage	protein-coding sequences
1	chromosome 1	2894792	66.42	1587
2	chromosome 2	2396129	64.72	1333
3	chromosome 3	2263199	66.21	1252
4	chromosome 4	1825687	66.70	1013

Genome assembly with short sequence reads cannot assemble longer, highly repetitive sequence elements. Here, the longer PacBio reads of up to 15 kbp allow these repetitive structure elements to be assembled and previously existing gaps to be closed. Within these regions nine ORFs with altered annotations relative to the previous genome sequences were identified and summarized in Table 3. An increase in size is also caused by additional bases within the open reading frame at the 3'end, and the correction of splicing events occurring in these genes as exemplified in Fig. 2. Interestingly, blastp database searches for protein homologues identified proteins involved in cell flocculation (agglutination) and cell surface recognition. In total we found more than 500 sequence differences relative to the 2011 published genome (Küberl et al., 2011). PCR amplification and Sanger sequencing of insertions and deletions observed in putative coding regions confirmed the validity of the new reference sequence. In 34 out of 35 regions tested, the *de novo* assembled Pacific Biosciences sequence was confirmed (data not shown). Additionally, the improvement with this new reference genome is confirmed by mapping the Roche 454 GS FLX Titanium reads to both the new reference genome and the 2011 genome (Küberl et al., 2011). More of the reads map to a combinatorial data set (*de novo* genome, killer plasmids and mitochondrial DNA sequence) of the new reference genome than the 2011 sequence (99.96% vs 99.63%, respectively). Furthermore, paired-end Illumina HiSeq reads from the strains BG08 and BG10 (BioGrammatics Inc., Carlsbad, USA) were aligned to the new reference genome sequence to determine the differences between closely related *P. pastoris strains*. 2 base differences are evident between BG08 and BG10 outside of the killer plasmids, and 24 differences were found between the BioGrammatics strains and the CBS7435 sequence presented here (supplementary Table S3). Except for one triple nucleotide insertion, all of the changes affected single nucleotide deletions and insertions; no inversions or larger rearrangements are found. Only small clonal variations occur between these closely related *P. pastoris* strains, even after storage at different sites, for many years indicating defined molecular manipulation can be precise with relatively little clonal drift. In this study, 5256 potential ORFs are identified, of which 5111 can be verified on the basis of either RNA-seq data or published peptide sequences (Renuse et al., 2014). In this manner more than 50 new reading frames were identified (supplementary table S4) relative to the 2011 draft genome sequence (Küberl et al., 2011). Under the growth conditions described in this manuscript, no consider-able transcript levels were found for 145 previously annotated open reading frames. Neither evidence by RNA-seq experiments nor data from published proteomics experiments (Renuse et al., 2014) find transcripts in these regions (supplementary Table S5). Additionally, 304 regions were identified to which transcript sequence reads map without the presence of either an ORF >50 aa or ORFs showing significant similarity to NCBInr protein database deposited entries (supplementary Table S6). These regions might contain non-coding RNA regulatory elements, such as described in *S. cerevisiae* (Sibthorp et al., 2013; Thompson and Parker, 2007) and *Schizosaccharomyces pombe* (Volpe et al., 2003; Wilhelm et al., 2008). One can speculate that similar cryptic unstable transcripts (CUTs), such as those involved in the regulation of meiosis (Lardenois et al., 2011), histone methylation (van Dijk et al., 2011) and telomere length (Luke et al., 2008) in *S. cerevisiae*, might exist *Pichia species* as well. The RNA-seq data also contains reads for overlapping ORFS, e.g. transcripts of two genes with opposite transcriptional orientation showing elevated read coverage. These overlaps of sense-antisense gene pairs might have a regulatory function in gene expression and silencing such as described in *S. cerevisiae* (David et al., 2006; Drinnenberg et al., 2009; Nagalakshmi et al., 2008). In order to create a reference sequence for the wildtype genome, the deleted *AOX1* gene of the sequenced mutS strain was complemented in silico to generate the new full reference sequence, which can be accessed from the NCBI web interface. Due to the large number of changes in this genome compared to the previously published genomes, we propose to use this new and completed whole genome sequence as a reference sequence for *P. pastoris*, which should facilitate future omics and systems biology studies, as well as, precise genome engineering approaches.

Table 3. Putative ORFs identified in the gaps of the *P. pastoris* CBS7435 genome sequence of 2011. Putative ORFs differing in length and were found on each of the four chromosomes. All 9 genes identified in these regions are characterized by the presence of highly repetitive sequence motifs. Amongst these, one previously undiscovered and new putative open reading frame (ACIB1EUKG55381) was identified. n.p. not present.

gene		chr	length [bp]		spliced		description
2011	2016		2011	2016	2011	2016	
PP7435_Chr1-1550	ACIB1EUKG51234	1	1020	4149	yes	yes	plaque matrix protein-like
PP7435_Chr2-0267	ACIB1EUKG53054	2	2394	3345	yes	no	cell surface glycoprotein
PP7435_Chr3-0722	ACIB1EUKG54633	3	1257	1407	yes	no	zinc metalloprotease zmpB
PP7435_Chr3-1225	ACIB1EUKG54124	3	1503	7308	yes	no	zonadhesin
PP7435_Chr3-1228	ACIB1EUKG54122	3	2628	3555	No	no	flocculation protein flo9
PP7435_Chr3-1236	ACIB1EUKG54115	3	1131	4407	yes	yes	agglutinin-like protein 3
PP7435_Chr4-0001	ACIB1EUKG55517	4	918	1230	no	no	cell surface glycoprotein
n.p.	ACIB1EUKG55381	4	-	6537	-	no	cell agglutination protein
PP7435_Chr4-1019	ACIB1EUKG55365	4	981	8088	yes	no	zonadhesin

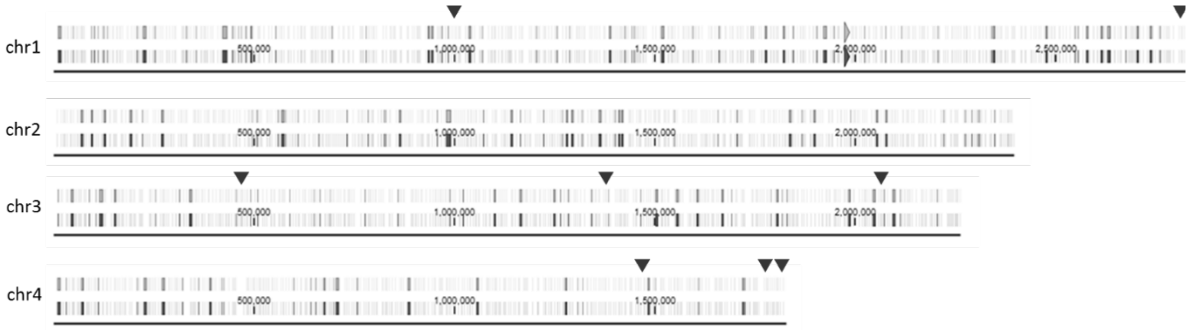


Figure 1: Open reading frames (ORFs) identified in the closed gaps of the *P. pastoris* chromosomes. Six small gaps of 60-200 bp as well as six larger gaps of 2-6 kbp present in the CBS 7435 genome sequence could be closed. The ORFs found in these regions are marked with dark triangles. In addition, the orientation of the four chromosomes was standardized to show the ribosomal clusters at the 3' ends. The vertical lines marked on all four chromosomes represent the annotation of ORFs.

Alternative splicing and RNA-seq data mapping

To further refine the automated annotation performed on the genome sequence of *P. pastoris* CBS7435, we performed an RNA-seq analysis using Illumina HiSeq sequencing data. An Illumina HiSeq mRNA library was run with an average read length of 50 bp reads from several different RNA samples of wildtype strains and strains harboring heterologous expression cassettes. These data were then mapped against the genome sequence which resulted in more than 98% of reads correctly aligned. Fig. 3 shows an example of three different events occurring in the analysis. In panel A, the RNA-seq data confirm the automated annotation as successful mapping depends on a gap opening in the reads. Alternatively, panel B and C depict two different examples in which the RNA-seq reads do not confirm previous predicted splicing events. Previous intron mis-calls in the 2011 draft genome sequence demonstrate: (1) in panel B, the mapped reads do not substantiate the presence of an intron in this position, and (2) in panel C, the RNA-seq mapping is forced to open a gap to correctly align to the genome sequence and therefore indicated the presence of an intron at this position. Based on these strategies, all four *P. pastoris* chromosomes and their corresponding open reading frames were manually corrected. The analysis resulted in the identification of 571 experimentally confirmed spliced genes in the genome sequence presented here compared with 797 reported in 2011 (based on computational predictions). In order to accurately map RNA-seq data onto the reference genome, it is important to allow mapping software to consider not only the introns annotated to create the protein encoding ORFs, but also splicing events that occur either outside these ORFs or as variants of the major splicing events. Forty-six variant splice acceptors and 33 variant splice donors were identified in the bed fileoutputs from TopHat. In addition, 11 exon "jumping" events were found where genes containing multiple introns were spliced from the donor site of one intron to the branch/acceptor site of a second intron. In general, alternative splicing occurs at low frequency and is the result of faulty selection of the proper donor or acceptor. For most cases that use an alternative splice acceptor, the same splice branch site is used and either the next upstream or downstream AG relative to the proper splice acceptor is used. In addition to alter-native splicing, a low-level of cryptic splicing events is observed in some highly transcribed mRNAs. Additionally, a number of mRNAs have introns in their 5'UTRs, which were previously not annotated; annotations for these probable locations have been included to help with the future identification of promoter-gene regions. Avery small number of genes appear to have 3'UTR introns. TopHat analysis of the complete set of RNA-seq data revealed 828 splicing events where the intron was flanked by GT—AG and there were at least 10 reads distributed across at least 4 experiments to support the existence of an intron. These 828 loci were hand curated to remove all locations that fell on the wrong strand of a highly transcribed mRNA or appeared to be the result of mis-mapping in genes with repetitive sequences. A final set of 771 junctions was used to generate mapping data for all RNA-seq experiments. A ".juncs" file (for use with TopHat –no-novel-juncs –raw-juncsoptions) containing the 771 locations is provided in the supplementary data. 771 possible splice sites can be confirmed by the current RNA-seq data. Wherever possible, the RNA-seq data set and proteomics data was used to confirm the predicted ORFs, to evaluate discrepancies in respect to translational starts and splice sites, and to re-evaluate or confirm sequence differences from the transcriptome comparisons (Renuse et al., 2014). Not surprisingly, for many ORFs the number of identified peptides was too low to clearly confirm deep sequencing RNA-seq data from the currently available proteomics data sets.

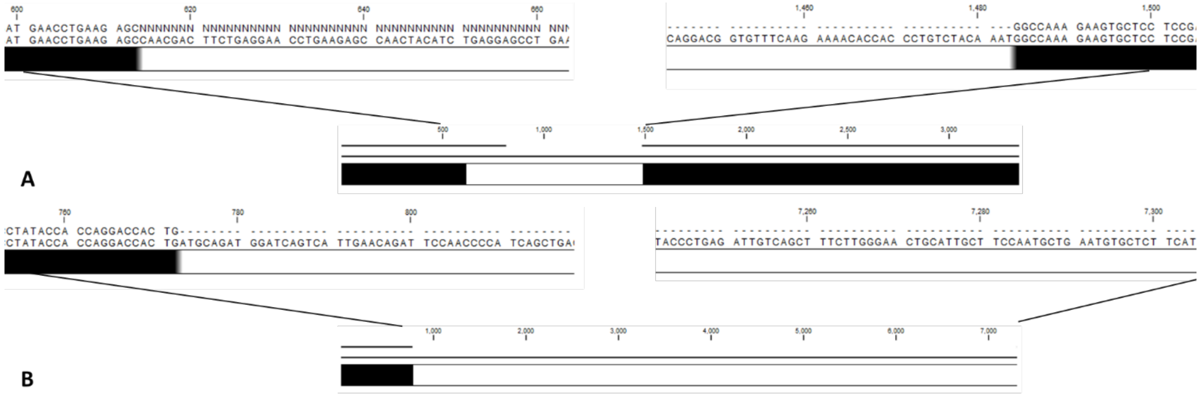

600
620
640
660
AT GAACCTGAAG AGCNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNN
AT GAACCTGAAG AGCCAACGAC TTCTGAGGAA CCTGAAGAGC CAACTACATC TGAGGAGCCT GA
1,460
1,480
1,500
---GGCCAAA GAAGTGCTCC TCCG
CAGGACG GTGTTTCAAG AAAACACCAC CCTGTCTACA AATGGCCAAA GAAGTGCTCC TCCG
500
1,000
1,500
2,000
2,500
3,000
A
760
780
800
CTATACCA CCAGGACCAC TG--
CTATACCA CCAGGACCAC TGATGCAGAT GGATCAGTCA TTGAACAGAT TCCAACCCCA TCAGCTGA
7,260
7,280
7,300
TACCCTGAG ATTGTCAGCT TTCTTGGGAA CTGCATTGCT TCCAATGCTG AATGTGCTCT TCAT
1,000
2,000
3,000
4,000
5,000
6,000
7,000
B

Figure 2. Multiple sequence alignment of P. pastoris CBS 7435 genome sequence (2011) and the P. pastoris CBS7435 reference genome sequence presented here. All alignments were performed using CLC Bio's proprietary alignment algorithm. The dark areas correspond to perfect nucleotide matches whereas white areas denote mismatches or missing bases. N denotes bases missing in the 2011 genome sequence. A. The genes PP7435_Chr2-0267 (top) and ACIB1EUKG53054 (bottom) were aligned against each other. B. The genes PP7435_Chr3-1225 (top) and ACIB1EUKG54124 (bottom) were aligned against each other.

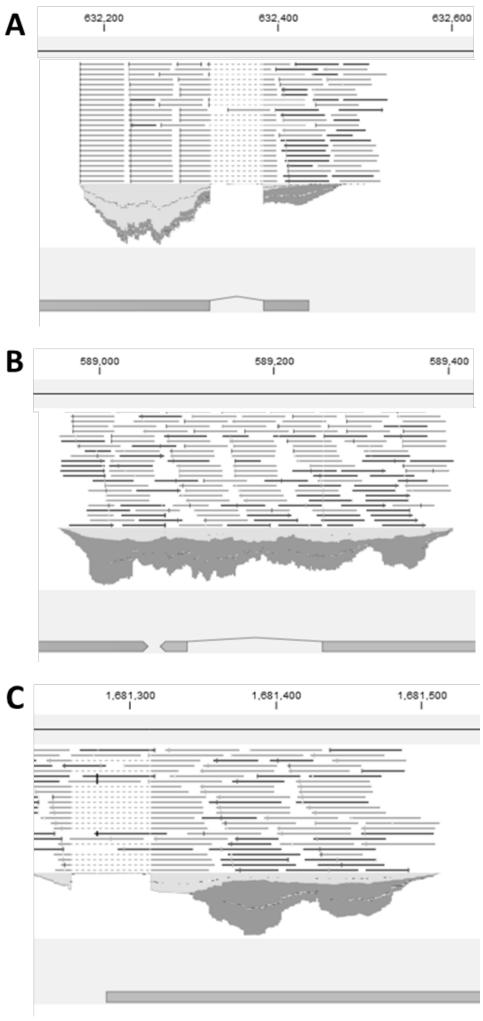


Figure 3. Exemplary intron splicing site prediction as identified by mapping RNA-seq reads to the *P. pastoris* reference sequence. **A** Gene ACIB1EUKG51520. Mapped reads verify the automated annotation. **B** Gene ACIB1EUKG56026. Due to the presence of mapped reads in the intron sequence the automated annotation had to be corrected. **C** Gene ACIB1EUKG55026. The intron identified in the RNA-seq data was incorporated into the automated annotation. The bottom part of each figure shows the gene as present in the genome sequence of *P. pastoris* CBS 7435 published in 2011. The middle part corresponds to the RNA-seq reads. CLC Genomics Workbench version 7 was used for visualizations and manual corrections.

Identification of putative *P. pastoris* centromeres

Budding yeast centromere sequences were originally identified in *S. cerevisiae*, where they are as short as 120 bp (Biggins, 2013; Clarke and Carbon, 1985). Those centromeres can be incorporated into plasmids to confer replicative stability (Clarke and Carbon, 1985), and can be used to generate yeast artificial chromosomes (Noskov et al., 2010). It would therefore be of value to identify centromere sequences for *P. pastoris*. However, centromere sequences diverge rapidly during evolution and are quite variable among yeast species (Malik and

Henikoff, 2009; Roy and Sanyal, 2011), so homology searches of the *P. pastoris* genome were unlikely to identify putative centromeres. As a first step toward identifying *P. pastoris* centromeres, we asked whether *P. pastoris* resembles other yeasts in possessing the centromere-specific histone H3 variant Cse4 (Biggins, 2013). Indeed, a clear CSE4 homolog on chromosome two (2011 genome location 1026080.1026598 or 2016 genome location1 026960.1027478) was present on Chr2. *P. pastoris* Cse4 was tagged with the monomeric superfolder GFP variant msGFP (Fitzgerald and Glick, 2014). For visualizing nuclei, the cells also expressed ER-targeted DsRed-HDEL, which fills the nuclear envelope with red fluorescence (Bevis et al., 2002). Fig. 6 shows that each cell contained a single green fluorescent spot in close proximity to the nuclear envelope. This pattern is typical for budding yeasts, in which the centromeres are clustered near the spindle pole bodies (Pearson et al., 2001). Therefore, it is inferred that *P. pastoris* likely has centromeres that are functionally similar to those found in other budding yeasts. A clue to the possible locations of *P. pastoris* centromeres came from visual inspection of the chromosome annotation patterns. Fig. 4 shows a portion of *P. pastoris* Chr1 with the predicted ORFs annotated in dark gray. The chromosome is densely packed with protein-coding genes, but a single gap of 9 kbp was observed. Upon closer inspection, this region was found to contain a perfect inverted repeat of 1991 bp. A similar inspection of the other three chromosomes revealed that each of them also has a single region of 9–11 kbp that is largely devoid of predicted ORFs, and that contains perfect or near perfect inverted repeats of 1991–2699 bp (Table 4).

Table 4. Locations of Hi-C centromere calls and RNA-seq mapping based prediction for *P. pastoris* CBS 7435 centromeres. Varoquaux et al. used the chromatin conformation capture assay, Hi-C, to predict the centromere regions of *P. pastoris* GS115 to within a 20 kbp region. Based on RNA-seq data mapping to the reference sequence presented here we were able observe a drastic drop in signal strength in the regions below, indicating a low transcriptional status in those regions. In those regions we identified near perfect inverted repeats. The reorientation of chr1, chr3 and chr4 described above resulted in a differing value when compared to GS115. The slightly differing value between GS115 and CBS7435 on chromosome 4 arises from the shorter length of GS115 chr4.

	GS115 2009		CBS7435 2016 predicted centromeres		ORF-free space [bp]	Inverted Repeats [bp]		
chr	Hi-C	predicted	before reorientation	after reorientation		individual repeats	total sequence spanned	Identity [%]
1	1408908 ± 20000	1400423..1409375	1487825..1493796	1401559..1407530	8,955	1991	8953	99
2	1556231 ± 20000	1542915..1551466	1545323..1551977	844482..851136	10,413	2699	8552	99
3	2226823 ± 20000	2202870..2211602	2222793..2228973	34486..40666	8,734	2649	8733	99
4	1719280 ± 20000	1701016..1712046	1762920..1769148	58794..65022	110586	2559	11031	99

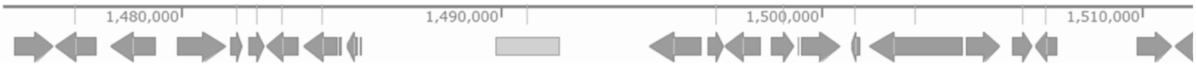


Figure 4. ***P. pastoris*** **chr1 with the predicted ORFs annotated.** The annotated ORFs are marked in dark grey. The putative centromere unique region is marked in bright gray.

Because centromeres tend to have few transcribed genes and sometimes contain inverted repeats (McFarlane and Humphrey, 2010), the ORF-free regions with inverted repeats are putative centromeres for *P. pastoris*. To confirm that the putative centromeres are largely devoid of transcribed genes, we examined RNA-seq data. As shown in Fig. 5, each of the putative centromere regions corresponds to a sharp and pronounced drop in the RNA-seq signal strength. When transcriptional profiles were generated for the full chromosomes using 4 kbp windows at 100 bp intervals, the predicted centromeres corresponded to the minimum values in the plots after telomere sequences were excluded (Fig. 5). Independent evidence that these regions are *P. pastoris* centromeres came from a recent analysis using a chromatin conformation capture assay called Hi-C (Varoquaux et al., 2015). That study took advantage of centromere clustering, of the type shown in Fig. 6 to map predicted centromere locations within approximately 20 kbp. The predicted centromere locations closely match the regions identified here (Table 4). Interestingly, the MATa1 and a2 locus on chr4 is found in between the *P. pastoris* centromere inverted repeats. Hanson et al. have determined that the flanking mating type inverted repeats are found in all four orientations (Hanson et al., 2014). While the centromere repeats are smaller, it is still possible that these can also undergo flipping and result in eight combinations at the end of chr4. If the centromere function is defined solely by the inverted repeats and their spacing, the resulting four possible arrangements of the centromeric core relative to immediate flanking regions and the chromosome as a whole would be functionally identical. The core and its relative orientation might potentially play a role during meiotic recombination and chromosome segregation during sporulation of a diploid cell. The combined data give high confidence that we have identified the four centromeres in *P. pastoris*. However, a rigorous demonstration will require further evidence, such as crosslinking of nucleosome-associated Cse4 to the putative centromeres (Meluh et al., 1998) or confirmation that the putative centromeres confer replicative stability to plasmids (Clarke and Carbon, 1985).

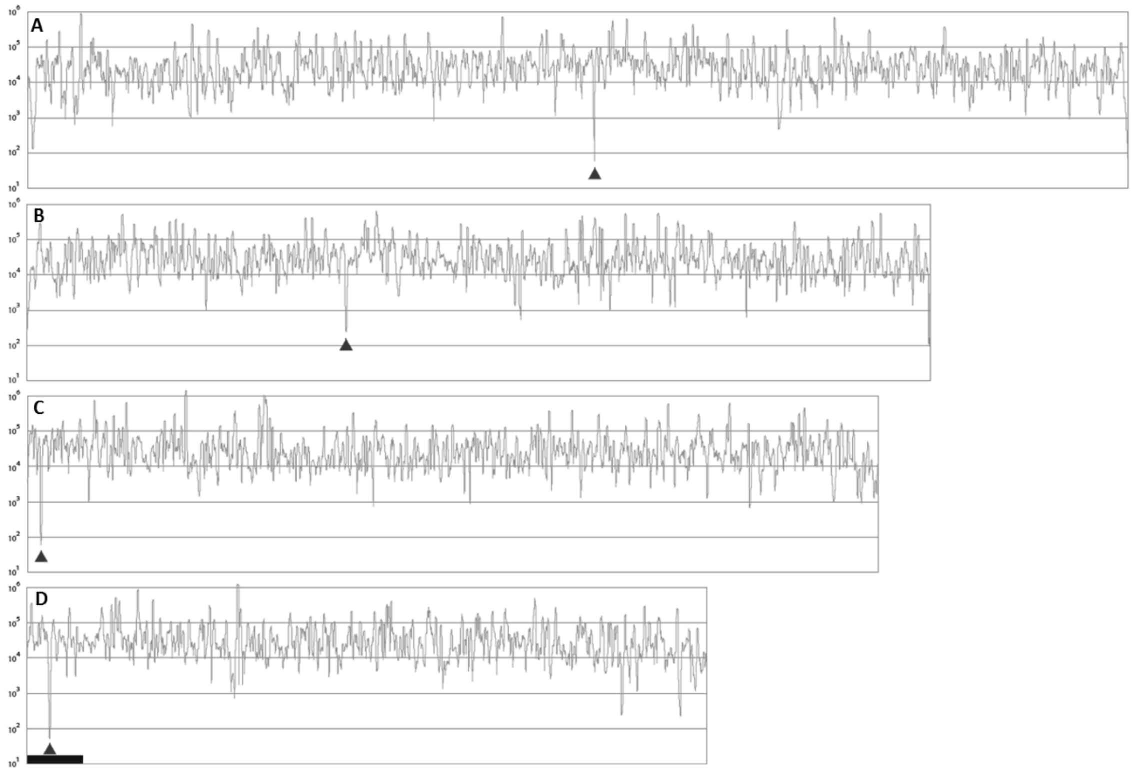


Figure 5. Putative location of *P. pastoris* centromeres indicated by RNA-seq reads mapped to this reference sequence. A-D corresponds to chromosomes 1-4. The putative centromere regions are largely devoid of transcribed genes as can be seen by the marked drop in the RNA-seq signal strength. The dark triangles correspond to the location of the putative centromere on each chromosome. The 138 kbp mating type chromosomal inversion region is indicated by the dark bar on chr4. The log scale plot shows the transcriptome density with 4 kbp windows at 100 bp intervals normalized to the maximum density window of 900,000 for chr1.

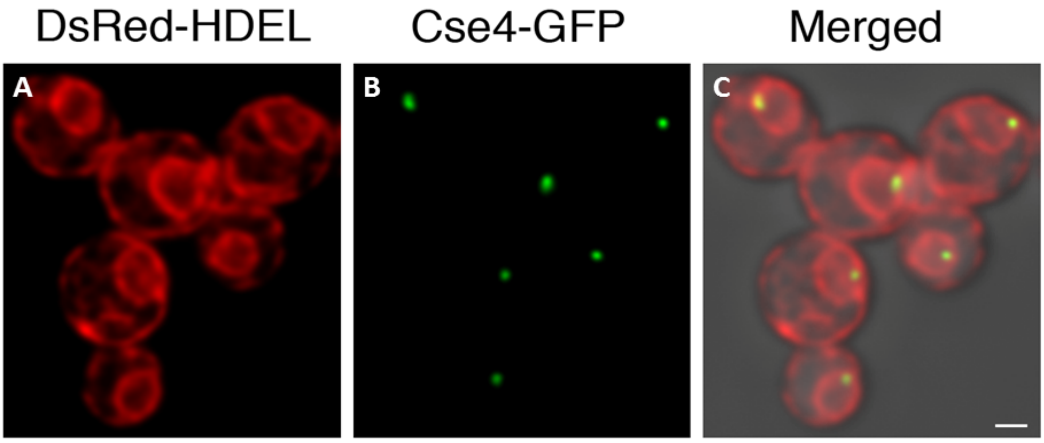


Figure 6. Visualization of clustered centromeres in *P. pastoris* by confocal microscopy. This strain expressed DsRed-HDEL to label the endoplasmic reticulum in red. The ring visible in each cell is the nuclear envelope. In addition, the strain expressed Cse4-GFP to label centromeres in green. The merged image shows the two fluorescence signals overlaid on a transmitted light image of the cells. A cluster of centromeres is visible at the nuclear periphery in each cell. Scale bar, 2 µm.

***P. pastoris* killer plasmids**

The dairy yeast *K. lactis* was one of the first yeast species proven to harbor a set of two different linear DNA fragments, which enabled the cells to kill other yeasts by secreting an exo-toxin (Gunge et al., 1981). These plasmids have been termed yeast killer plasmids and have so far been identified in several yeast genera such as *Botryascus, Pichia, Debaryomyces* and *Wingea* (Cong et al., 1994; Hayman and Bolen, 1991; Wickner, 1979; Wickner and Leibowitz, 1976; Worsham and Bolen, 1990). The genetic composition of these linear plasmids seems to be quite conserved among those species. The *K. lactis* pGKL1/pGKL2 system has been extensively studied and resulted in the elucidation of plasmid encoded gene functions (Butler et al., 1991; Gunge, 1986; Gunge and Kitada, 1988; Kikuchi et al., 1984; Sor et al., 1983; Stam et al., 1986; Tokunaga et al., 1987). Banerjee et al. (1998) were the first to describe the presence of RNase resistant double stranded DNA molecules sensitive towards DNase I digestion in the methylotrophic yeast *P. pastoris*. Using paired-end Illumina data from a *P. pastoris* wildtype strain (BG08) and a *P. pastoris* CBS7435 crtEBIY strain expressing the beta-carotene synthesis pathway (Geier et al., 2015) as well as PacBio sequence data of the *P. pastoris* CBS7435 mutS strain, two so far unpublished sequences of differently sized linear plasmids with intact LTR sequences were identified. These linear plasmids are13.1 and 9.5 kbp in size and show homologues of several annotated ORFs frequently found on killer plasmids from other yeast species such as *K. lactis, Pichia accaciae* and *S. cerevisiae*. Among the coding sequences found on the two plasmids are DNA polymerases, an RNA polymerase, a helicase, an mRNA capping enzyme and several homologues to *K. lactis* killer plasmid proteins. Due to the DNA size selection performed during PacBio library construction, the smaller killer plasmid was lost entirely, and the larger killer plasmid was significantly underrepresented in the PacBio library. In Illumina data from a library prepared with ~500 bp inserts, the copy number of the two killer plasmids was estimated at 80–100 in BG08 and *P. pastoris* CBS7435 crtEBIY. In sequencing other wild type *P. pastoris* strains, approximately 25% of sequencing data maps to killer plasmid sequences (supplementary Fig. S2), indicating about 3 Mbp of total killer plasmid DNA relative to 9.38 Mbp of genomic DNA. The 13.1 and 9.5 kbp large plasmids identified here show 8 and 6 ORFs, respectively (Fig. 7). These ORFs have high sequence identity to already known *K. lactis* killer plasmid proteins as seen in Table 5. Both *P. pastoris* killer plasmids showed very similar spatial organization compared to *K. lactis* and *S. cerevisiae* plasmids (Schickel et al., 1996) with regard to the order of the ORFs. Also the coding density is similarly high, with 90.9% and 91.8% of all nucleotides coding for proteins in the 13.1 kbp and 9.5 kbp *P. pastoris* killer plasmids, respectively. ORF1 on the 13.1 kbp large plasmid shows high sequence identity to a plasmid specific DNA polymerase. Together with ORF5 and ORF6, annotated as RNA polymerase and RNA polymerase subunits, respectively, these putative genes most likely allow for the replication and transcription of plasmid encoded genes (Jung et al., 1987; Wilson and Meacock, 1988). ORF2, a potential mRNA capping enzyme and ORF3, a helicase, could potentially stabilize the linear plasmid in the cytosolic space (Larsen et al., 1998). On the basis of *K. lactis* killer plasmid protein functions we could also identify potential DNA binding proteins (ORF4) and a terminal recognition factor (ORF8) which is potentially responsible for the protection of linear DNA present in the cytosol by binding to the LTR sequences found at the outer boundaries of the plasmids (Schaffrath and Meacock, 2001). As described, in *S. cerevisiae* the smaller 9.5 kbp plasmid contains a DNA polymerase (ORF1) and several different ORFs with sequence identities to killer plasmid toxins. The elements responsible for conferring toxicity are encoded on the smaller plasmid. In *K. lactis* and other yeast species the heterotrimeric killer toxin is made up of three different subunits, termed alpha, beta and gamma while a fourth protein coded on the same plasmid functions as an anti-toxin (Stark and Boyd, 1986). Based on sequence homology and localization on the plasmid, ORF4 might contain the alpha and beta subunit (Larsen et al., 1998). Neither ORF2, ORF3, ORF5 nor ORF6 could be identified as the killer toxin gamma subunit or anti-toxin protein. Although Banerjee and colleagues were able to isolate a DNase I digestible double stranded DNA fragment that was not susceptible to RNase degradation in *P. pastoris* (Banerjee et al., 1998), evaluation of 14 different *P. pastoris* strains for their killer activity showed no killer phenotype (Banerjee and Verma, 2000). In the sequencing data presented here major components of the killer plasmid system, such as the presence of DNA and RNA polymerases and putative open reading frames responsible for plasmid integrity, have been found. However, none of the remaining ORFs could be identified as homologues of the *K. lactis* toxin gamma subunit or the anti-toxin protein. The lack of these proteins could potentially cause a loss of the killer phenotype and therefore provide support for the finding of (Banerjee et al., 1998; Banerjee and Verma, 2000).

Table 5. Putative open reading frames identified on the 13.1 kbp and 9.5 kbp plasmids of *P. pastoris*. All protein sequences were analyzed using BLASTP (protein-protein blast) against the non-redundant protein database. The protein entries showing the highest sequence identity to the query are summarized. No homologues of the gamma toxin subunit or the antitoxin gene were identified on the two *P. pastoris* killer plasmids.

13.1 kbp plasmid	**Description**	**Organism and sequence identity**
ORF1	DNA-polymerase	*Debaryomyces hansenii* 67%
ORF2	mRNA capping enzyme	*Pichia etchellsii* 49%
ORF3	Helicase	*K. lactis* 59%
ORF4	DNA binding protein	*Millerozyma acacia* 54%
ORF5	RNA-polymerase	*K. lactis* 57%
ORF6	RNA-polymerase subunit	*M. acacia* 41%
ORF7	Killer toxin protein	*K. lactis* 50%
ORF8	terminal recognition factor	*M.-acacia* 58%
9.5 kbp plasmid		
ORF1	DNA-polymerase	*D. hansenii* 55%
ORF2	Hypothetical protein	no similarity
ORF3	Hypothetical protein	no similarity
ORF4	Killer toxin protein	*K. lactis* 46%
ORF5	Hypothetical protein	no similarity
ORF6	Hypothetical protein	no similarity

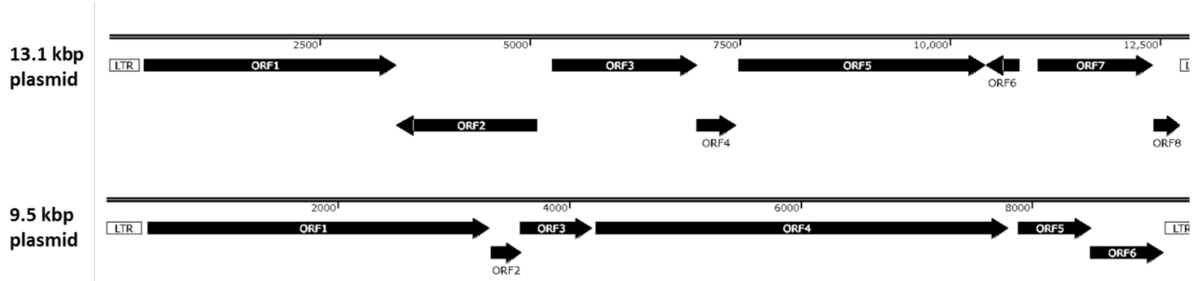


Figure 7. Genetic Organization of the two linear plasmids identified in *P. pastoris*. Based on the plasmid sequences we identified 8 open reading frames on the 13.1 kbp killer plasmid and 6 open reading frames on the 9.5 kbp large killer plasmid. Both plasmids are flanked by long terminal repeat sequences (LTR).

Author's contributions

LS, TC, MG, IT, JC, KM and AG planned and started the project collaboration. Wet laboratory work was carried out by LS, TC, MG, AS, KD and KM. In silico analysis was performed by LS, TC, MG, FK, UV, ST, TR and LSo. Experimental work and analysis of centromeres was done by LS, TC, KD and BG. Peptide mapping and analysis of proteome data was performed by TC, BD and RBG. LS, TC, BG, TR, IT, JC, KM, and AG wrote the manuscript. The research leading to these results has received funding from the Innovative Medicines Initiative Joint Undertaking project CHEM21 under grant agreement n°115360, resources of which are composed of financial contribution from the European Union's Seventh Framework Program (FP7/2007-2013) and EFPIA companies' in-kind contribution. In addition, this work has been supported by the Federal Ministry of Science, Research and Economy (BMWFW),the Federal Ministry of Traffic, Innovation and Technology (bmvit), the Styrian Business Promotion Agency SFG, the Standortagentur Tirol, the Government of Lower Austria and ZIT—Technology Agency of the City of Vienna through the COMET-Funding Program managed by the Austrian Research Promotion Agency FFG. BSG was supported by NIH grant R01 GM104010, and KJD was supported by NIH training grant T32 GM007183. AG gratefully acknowledges the support and fruitful discussion of the presented work with Prof. Suresh Subramani and group members during a sabbatical at UCSD. Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbiotec.2016.04.023.

Tables and Figures

Table 1. *P. pastoris* strains used in this study.

Table 2. Summary of all major unitigs identified in the assembly of *P. pastoris* genomic DNA.

Table 3. Putative ORFs identified in the gaps of the *P. pastoris* CBS7435 genome sequence of 2011.

Table 4. Locations of Hi-C centromere calls and RNA-seq mapping based prediction for *P. pastoris* CBS 7435 centromeres.

Table 5. Putative open reading frames identified on the 13.1 kbp and 9.5 kbp plasmids of *P. pastoris*.

Figure 1. Open reading frames (ORFs) identified in the closed gaps of the *P. pastoris* chromosomes.

Figure 2. Multiple sequence alignment of *P. pastoris* CBS 7435 genome sequence (2011) and the *P. pastoris* CBS7435 reference genome sequence presented here.

Figure 3. Exemplary intron splicing site prediction as identified by mapping RNA-seq reads to the *P. pastoris* reference sequence.

Figure 4. *P. pastoris* chr1 with the predicted ORFs annotated.

Figure 5. Putative location of *P. pastoris* centromeres indicated by RNA-seq reads mapped to this reference sequence.

Figure 6. Visualization of clustered centromeres in *P. pastoris* by confocal microscopy.

Figure 7. Genetic Organization of the two linear plasmids identified in *P. pastoris*.

References

Banerjee, H., Kopvak, C., Curley, D., 1998. Identification of Linear DNA Plasmids of the YeastPichia pastoris. Plasmid 40, 58–60. doi:http://dx.doi.org/10.1006/plas.1998.1341

Banerjee, H., Verma, M., 2000. Search for a novel killer toxin in yeast Pichia pastoris. Plasmid 43, 181–3. doi:10.1006/plas.1999.1452

Bevis, B.J., Hammond, A.T., Reinke, C.A., Glick, B.S., 2002. De novo formation of transitional ER sites and Golgi structures in Pichia pastoris. Nat Cell Biol 4, 750–756.

Biggins, S., 2013. The composition, functions, and regulation of the budding yeast kinetochore. Genetics 194, 817–846. doi:10.1534/genetics.112.145276

Butler, A.R., O'Donnell, R.W., Martin, V.J., Gooday, G.W., Stark, M.J., 1991. Kluyveromyces lactis toxin has an essential chitinase activity. Eur. J. Biochem. 199, 483–488.

Clarke, L., Carbon, J., 1985. The structure and function of yeast centromeres. Ann. Rev. Genet 19, 29–56.

Cong, Y., Yarrow, D., Li, Y., Fukuhara, H., 1994. Debaryomyces hansenii and Wingea robertsiae. Yeast 1327–1335.

Cregg, J.M., Tolstorukov, I., Kusari, A., Sunga, J., Madden, K., Chappell, T., 2009. Chapter 13 Expression in the Yeast Pichia pastoris, in: Enzymology, R.R.B. and M.P.D.B.T.-M. in (Ed.), Guide to Protein Purification, 2nd Edition. Academic Press, pp. 169–189. doi:http://dx.doi.org/10.1016/S0076-6879(09)63013-5

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., Steinmetz, L.M., 2006. A high-resolution map of transcription in the yeast genome. Proc. Natl. Acad. Sci. U. S. A. 103, 5320–5. doi:10.1073/pnas.0601091103

De Schutter, K., Lin, Y.-C., Tiels, P., Van Hecke, A., Glinka, S., Weber-Lehmann, J., Rouzé, P., Van de Peer, Y., Callewaert, N., 2009. Genome sequence of the recombinant protein production host Pichia pastoris. Nat. Biotechnol. 27, 561–566. doi:10.1038/nbt.1544

Drinnenberg, I.A., Weinberg, D.E., Xie, K.T., Mower, J.P., Wolfe, K.H., Fink, G.R., Bartel, D.P., 2009. RNAi in budding yeast. Science (80-.). 326, 544–550. doi:10.1126/science.1176945.RNAi

Fitzgerald, I., Glick, B.S., 2014. Secretion of a foreign protein from budding yeasts is enhanced by cotranslational translocation and by suppression of vacuolar targeting. Microb. Cell Fact. 13, 125. doi:10.1186/s12934-014-0125-0

Geier, M., Brandner, C., Strohmeier, G. a, Hall, M., Hartner, F.S., Glieder, A., 2015a. Engineering *Pichia pastoris* for improved NADH regeneration: A novel chassis strain for whole-cell catalysis. Beilstein J. Org. Chem. 11, 1741–1748. doi:10.3762/bjoc.11.190

Geier, M., Fauland, P., Vogl, T., Glieder, A., 2015b. Compact multi-enzyme pathways in P. pastoris. Chem. Commun. 51, 1643–1646. doi:10.1039/C4CC08502G

Geier, M., Fauland, P.C., Vogl, T., Glieder, A., 2014. Compact multi enzyme pathways in Pichia pastoris. Chem. Commun. doi:10.1039/C4CC08502G

Gouldi, S.J., Mccollum, D., Spong, A.P., Heyman, J.A., 1992. Development of the Yeast Pichiapastoris as a Model Organism for a Genetic and Molecular Analysis of Peroxisome Assembly 8.

Gunge, N., 1986. Linear DNA killer plasmids from the yeast Kluyveromyces. Yeast 2, 153–62. doi:10.1002/yea.320020303

Gunge, N., Kitada, K., n.d. Replication and maintenance of the Kluyveromyces linear pGKL plasmids. Eur. J. Epidemiol. 4, 409–414. doi:10.1007/BF00146390

Gunge, N., Tamaru, A., Ozawa, F., Sakaguchi, K., 1981. Isolation and Characterization of Linear Deoxyribonucleic Acid Plasmids from Kluyveromyces lactis and the Plasmid- Associated Killer Character. J. Bacteriol. 145, 382–390.

Hanson, S.J., Byrne, K.P., Wolfe, K.H., 2014. Mating-type switching by chromosomal inversion in methylotrophic yeasts suggests an origin for the three-locus Saccharomyces cerevisiae system. Proc. Natl. Acad. Sci. U. S. A. 1–8. doi:10.1073/pnas.1416014111

Harris, D., Goodnow, R., 1979. United States Patent . 119].

Hayman, G.T., Bolen, P.L., n.d. Linear DNA plasmids of Pichia inositovora are associated with a novel killer toxin activity. Curr. Genet. 19, 389–393. doi:10.1007/BF00309600

Jung, G.H., Leavitt, M.C., Ito, J., 1987. Yeast killer plasmid pGKL1 encodes a DNA polymerase belonging to the family B DNA polymerases. Nucleic Acids Res. 15, 9088.

Kikuchi, Y., Hirai, K., Hishinuma, F., 1984. The yeast linear DNA killer plasmids, pGKL1 and pGKL2, possess terminally attached proteins. Nucleic Acids Res. 12 , 5685–5692. doi:10.1093/nar/12.14.5685

Küberl, A., Schneider, J., Thallinger, G.G., Anderl, I., Wibberg, D., Hajek, T., Jaenicke, S., Brinkrolf, K., Goesmann, A., Szczepanowski, R., Pühler, A., Schwab, H., Glieder, A., Pichler, H., 2011. High-quality genome sequence of Pichia pastoris CBS7435 154, 312–320. doi:10.1016/j.jbiotec.2011.04.014

Kurtzman, C.P., 2009. Biotechnological strains of Komagataella (Pichia) pastoris are Komagataella phaffii as determined from multigene sequence analysis. J. Ind. Microbiol. Biotechnol. 36, 1435–1438. doi:10.1007/s10295-009-0638-4

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. doi:10.1099/ijs.0.63491-0

Lardenois, A., Liu, Y., Walther, T., Chalmel, F., Evrard, B., Granovskaia, M., Chu, A., Davis, R.W., Steinmetz, L.M., Primig, M., 2011. Execution of the meiotic noncoding RNA expression program and the onset of gametogenesis in yeast require the conserved exosome subunit Rrp6. Proc. Natl. Acad. Sci. U. S. A. 108, 1058–1063. doi:10.1073/pnas.1016459108

Larsen, M., Gunge, N., Meinhardt, F., 1998. Kluyveromyces lactis killer plasmid pGKL2: evidence for a viral-like capping enzyme encoded by ORF3. Plasmid 40, 243–246. doi:10.1006/plas.1998.1367

Luke, B., Panza, A., Redon, S., Iglesias, N., Li, Z., Lingner, J., 2008. The Rat1p 5??? to 3??? Exonuclease Degrades Telomeric Repeat-Containing RNA and Promotes Telomere Elongation in Saccharomyces cerevisiae. Mol. Cell 32, 465–477. doi:10.1016/j.molcel.2008.10.019

Malik, H.S., Henikoff, S., 2009. Major Evolutionary Transitions in Centromere Complexity. Cell 138, 1067–1082. doi:10.1016/j.cell.2009.08.036

Marx, H., Mecklenbr??uker, A., Gasser, B., Sauer, M., Mattanovich, D., 2009. Directed gene copy number amplification in Pichia pastoris by vector integration into the ribosomal DNA locus. FEMS Yeast Res. 9, 1260–1270.

McFarlane, R.J., Humphrey, T.C., 2010. A role for recombination in centromere function. Trends Genet. 26, 209–213. doi:10.1016/j.tig.2010.02.005

Meluh, P.B., Yang, P., Glowczewski, L., Koshland, D., Smith, M.M., 1998. Cse4p is a component of the core centromere of Saccharomyces cerevisiae. Cell 94, 607–613. doi:10.1016/S0092-8674(00)81602-5

Näätsaari, L., Mistlberger, B., Ruth, C., Hajek, T., Hartner, F.S., Glieder, A., 2012. Deletion of the Pichia pastoris KU70 homologue facilitates platform strain generation for gene expression and synthetic biology. PLoS One 7, e39720. doi:10.1371/journal.pone.0039720

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M., 2008. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science (80-.). 320, 1344–1349.

nath Banerjee, H., Verma, M., 2000. Search for a Novel Killer Toxin in Yeast Pichia pastoris. Plasmid 43, 181–183. doi:http://dx.doi.org/10.1006/plas.1999.1452

Noskov, V.N., Chuang, R.-Y., Gibson, D.G., Leem, S.-H., Larionov, V., Kouprina, N., 2010. Isolation of circular yeast artificial chromosomes for synthetic biology and functional genomics studies. Nat. Protoc. 6, 89–96.

Pearson, C.G., Maddox, P.S., Salmon, E.D.D., Bloom, K., 2001. Budding yeast chromosome structure and dynamics during mitosis. J. Cell Biol. 152, 1255–66. doi:10.1083/jcb.152.6.1255

Renuse, S., Madugundu, A.K., Kumar, P., Nair, B.G., Gowda, H., Prasad, T.S.K., Pandey, A., 2014. Proteomic analysis and genome annotation of Pichia pastoris, a recombinant protein expression host . PROTEOMICS . doi:10.1002/pmic.201400267

Robson, R.L., Jones, R., Robson, R.M., Schwartz, A., Richardson, T.H., 2015. Azotobacter Genomes: The Genome of Azotobacter chroococcum NCIMB 8003 (ATCC 4412). PLoS One 10, e0127997. doi:10.1371/journal.pone.0127997

Roy, B., Sanyal, K., 2011. Diversity in requirement of genetic and epigenetic factors for centromere function in fungi. Eukaryot. Cell 10, 1384–1395. doi:10.1128/EC.05165-11

Schaffrath, R., Meacock, P.A., 2001. An SSB encoded by and operating on linear killer plasmids from Kluyveromyces lactis. Yeast 18, 1239–1247. doi:10.1002/yea.773

Schickel, J., Helmig, C., Meinhardt, F., 1996. Kluyveromyces lactis killer system: Analysis of cytoplasmic promoters of the linear plasmids. Nucleic Acids Res. 24, 1879–1886. doi:10.1093/nar/24.10.1879

Sor, F., Wèsolowski, M., Fukuhara, H., 1983. Inverted terminal repetitions of the two linear DNA associated with the killer character of the yeast Kluyveromyces lactis. Nucleic Acids Res. 11 , 5037–5044. doi:10.1093/nar/11.15.5037

Stam, J.C., Kwakman, J., Meijer, M., Stuitje, A.R., 1986. Efficient isolation of the linear DNA killer plasmid of Kluyveromyces lactis: evidence for location and expression in the cytoplasm and characterization of their terminally bound proteins. Nucleic Acids Res. 14 , 6871–6884. doi:10.1093/nar/14.17.6871

Stark, M.J., Boyd, A., 1986. The killer toxin of Kluyveromyces lactis: characterization of the toxin subunits and identification of the genes which encode them. EMBO J. 5, 1995–2002.

Thompson, D.M., Parker, R., 2007. Cytoplasmic decay of intergenic transcripts in Saccharomyces cerevisiae. Mol. Cell. Biol. 27, 92–101. doi:10.1128/MCB.01023-06

Tokunaga, M., Wada, N., Hishinuma, F., 1987. Expression and identification of immunity determinants on linear DNA killer plasmids pGKL1 and pGKL2 in Kluyveromyces lactis. Nucleic Acids Res. 15 , 1031–1046. doi:10.1093/nar/15.3.1031

van Dijk, E.L., Chen, C.L., d'Aubenton-Carafa, Y., Gourvennec, S., Kwapisz, M., Roche, V., Bertrand, C., Silvain, M., Legoix-Né, P., Loeillet, S., Nicolas, A., Thermes, C., Morillon, A., 2011. XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. TL - 475. Nature 475 VN - , 114–117. doi:10.1038/nature10118

Varoquaux, N., Liachko, I., Ay, F., Burton, J.N., Shendure, J., Dunham, M.J., Vert, J.P., Noble, W.S., 2015. Accurate identification of centromere locations in yeast genomes using Hi-C. Nucleic Acids Res. 43, 5331–5339. doi:10.1093/nar/gkv424

Volpe, T., Schramke, V., Hamilton, G.L., White, S.A., Teng, G., Martienssen, R.A., Allshire, R.C., 2003. RNA interference is required for normal centromere function in fission yeast. Chromosom. Res. 11, 137–146. doi:10.1023/A:1022815931524

Wickner, R.B., 1979. The killer double-stranded RNA plasmids of yeast. Plasmid 2, 303–322. doi:http://dx.doi.org/10.1016/0147-619X(79)90015-5

Wickner, R.B., Leibowitz, M.J., 1976. Chromosomal genes essential for replication of a double-stranded RNA plasmid of Saccharomyces cerevisiae: The killer character of yeast. J. Mol. Biol. 105, 427–443. doi:http://dx.doi.org/10.1016/0022-2836(76)90102-9

Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., Bähler, J., 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453, 1239–1243. doi:10.1038/nature07002

Wilson, D.W., Meacock, P.A., 1988. Extranuclear gene expression in yeast: evidence for a plasmid-encoded RNA polymerase of unique structure. Nucleic Acids Res. 16, 8097–8112.

Worsham, P.L., Bolen, P.L., n.d. Killer toxin production in Pichia acaciae is associated with linear DNA plasmids. Curr. Genet. 18, 77–80. doi:10.1007/BF00321119

Wu, J., Delneri, D., Keefe, R.T.O., 2014. Europe PMC Funders Group Non-coding RNAs in Saccharomyces cerevisiae : What is the function ? 40, 907–911. doi:10.1042/BST20120042.Non-coding

Conclusion and future outlook

The increasing use of enzymatic catalysis in industry hinges on the discovery of novel enzymes. The discovery process can be designed to incorporate several different technologies depending on the source material and the aim of a specific project (Bornscheuer et al., 2012; Erickson et al., 2012; Illanes et al., 2012). In the past, the most successful approach was the functional screening of genomic or metagenomics libraries (Uchiyama and Miyazaki, 2009). These screenings rely on the recombinant expression of a constructed library and identifying the desired activity of clones in a high-throughput manner. The DNA of active clones can subsequently be sequenced, and activity inferred for a specific sequence. One of the biggest advantages with this approach is the low information content necessary for project initiation. The existence of a proper screening assay and a gDNA library is very often sufficient for initiating an experiment. In a similar manner, the functional screening of cDNA libraries employs a different source material, RNA, to construct libraries, which are subsequently screened for desired activity. Common to both techniques is the possibility to probe the enzyme sequence space in a manner not feasible with other technologies. Functional screens of genomic, transcriptomic as well as metagenomics/transcriptomic libraries represent a powerful tool for enzyme discovery.

With the advent of powerful sequencing technologies, the number of deposited sequences is rising continuously (English et al., 2012; Ferrarini et al., 2013; Zhang et al., 2012). The ability to link a gene sequence with a specific enzymatic function gave rise to the field of functional genomics and allowed prediction of enzymatic function solely based on in silico data (Gray et al., 2015). Based on the homology to enzymes with known function the overall sequence identity, but also the existence of conserved protein motifs and residues, is calculated and used for the assessment of enzymatic activity (Marchler-Bauer et al., 2015; O'Leary et al., 2016). This opens up the possibility to probe the entire sequence space of public and private databases for a sequence with a certain function. By using these homology-based search methods we were able to identify five hydrophobin sequences in the meta-transcriptome of a fern species, which were successfully expressed in *K. phaffii*. This exemplifies the validity of this approach. However, relying on sequence homology-based techniques alone will complicate the discovery of enzymes with completely novel folds or functions as no reference sequence, for which homology could be calculated, exists (Lanfranchi et al., 2017).

Proteomics offers the possibility to correlate enzymatic function with protein sequence via MS-based identification methods. The sequence determination of unique peptides, which are mapped to a protein database allows the discovery of proteins with new functions based on enzymatic activity. In a more specific embodiment, this process can be further streamlined through the use of probes specific for an enzyme class, which lead to the development of the field of functional proteomics (Cravatt, 2014; Cravatt et al., 2008; Schittmayer and Birner-Gruenberger, 2012). The basis for all proteomics discoveries is the existence of a high-quality protein database. This dependency exemplifies the synergies between different omics technologies for successful enzyme discovery. As protein databases are based on genomic or transcriptomic sequences, proteomics benefits from the availability of data sets, which are highly related to the specific sample used for peptide sequencing (Ciesielska et al., 2013; Desgagné-Penix et al., 2010; Kirsch et al., 2012; Sturmberger et al., 2016c). This was also the case during the discovery of a glycoside hydrolase from *L. attenuatum*. Since we performed mRNA sequencing experiments on the same samples used for secretomic analysis we were able to generate a highly specific protein database. This dataset allowed us to identify the specific protein, an achievement which would have been not possible via the NCBI nr protein database, as the sequence of interest was not part of the database.

Regardless of which discovery approach is applied, at a certain point recombinant protein expression is employed for the verification of enzymatic activity. Expression tools for a large number of recombinant expression hosts are available. Amongst the species suitable for library expression, yeasts (Bordes et al., 2007; Branduardi et al., 2004; Celik and Calık, 2012; Fukuhara, 2006; Ghaemmaghami et al., 2003; Nicaud et al., 2002; Ooyen et al., 2006; Piotek et al., 1998b; Steinborn et al., 2006; Yurimoto et al., 2000; Yurimoto and Sakai, 2009) offer both the availability of a large number of genetic information and synthetic biology tools as well as the simplicity and ease of manipulation associated with unicellular microbial cells, such as fast growth on defined media, the availability of cultivation protocols and the possibility for upscaling (Cereghino et al., 2002; Cregg et al., 2000a; Leonardo M. Damasceno et al., 2012; Hamilton et al., 2003; Hamilton and Gerngross, 2007; Heyland et al., 2010; Idiris et al., 2010b; Xiao et al., 2004). Amongst yeast species, *K. phaffii* was established as one of the most important expression hosts both in academia and in industry (Mattanovich et al., 2016; Vogl et al., 2013b). The applicability of this host for protein expression relies on the possibility to engineer the genetic make-up this yeast species. This requires the availability of a high-quality genome sequence. By exploiting state-of-the-art sequencing technologies, we could further improve on published genome sequences by closing gaps, correcting rearrangements and better predicting spliced genes. This improved reference genome sequence allowed us to better understand the transcriptomic landscape of *K. phaffii* (Sturmberger et al., 2016a). Furthermore, a high-quality genome offers the sequence information necessary for targeted knock-out and integration of linear expression cassettes, which is also needed in the process of library screenings for enzyme discovery.

The degree of automation in lab environments is increasing and the cost for sequencing and gene synthesis is steadily declining. This entails a rise in the number of deposited sequences in public databases as well as their functional annotation. In combination with the capability to screen a higher number of clones due to robotic, human-independent labor the size and coverage of libraries can be increased. This in turn increases the chances of finding novel enzymes. Furthermore, as more proteomics probes are generated to cover a broader range of enzymatic functions, the search for enzymes with specific functions will no doubt be facilitated as well.

References

Bordes, F., Fudalej, F., Dossat, V., Nicaud, J., Marty, A., 2007. A new recombinant protein expression system for high-throughput screening in the yeast Yarrowia lipolytica 70, 493–502. https://doi.org/10.1016/j.mimet.2007.06.008

Bornscheuer, U.T., Huisman, G.W., Kazlauskas, R.J., Lutz, S., Moore, J.C., Robins, K., 2012. Engineering the third wave of biocatalysis. Nature 485, 185–194. https://doi.org/10.1038/nature11117

Branduardi, P., Valli, M., Brambilla, L., Sauer, M., Alberghina, L., Porro, D., 2004. The yeast Zygosaccharomyces bailii: A new host for heterologous protein production, secretion and for metabolic engineering applications. FEMS Yeast Res. 4, 493–504. https://doi.org/10.1016/S1567-1356(03)00200-9

Celik, E., Calık, P., 2012. Production of recombinant proteins by yeast cells. Biotechnol. Adv. 30, 1108–18. https://doi.org/10.1016/j.biotechadv.2011.09.011

Cereghino, G.P.L., Cereghino, J.L., Ilgen, C., Cregg, J.M., 2002. Production of recombinant proteins in fermenter cultures of the yeast Pichia pastoris 329–332. https://doi.org/10.1016/S0958166902003300

Ciesielska, K., Li, B., Groeneboer, S., Van Bogaert, I., Lin, Y.C., Soetaert, W., Van De Peer, Y., Devreese, B., 2013. SILAC-based proteome analysis of starmerella bombicola sophorolipid production. J. Proteome Res. 12, 4376–4392. https://doi.org/10.1021/pr400392a

Cravatt, B., 2014. Activity-based proteomics &#151; applications for enzyme and inhibitor discovery (357.1). FASEB J. 28, 357.1. https://doi.org/10.1096/fasebj.28.1_supplement.357.1

Cravatt, B.F., Wright, A.T., Kozarich, J.W., 2008. Activity-Based Protein Profiling: From Enzyme Chemistry to Proteomic Chemistry. Annu. Rev. Biochem. 77, 383–414. https://doi.org/10.1146/annurev.biochem.75.101304.124125

Cregg, J.M., Cereghino, J.L., Shi, J., Higgins, D.R., 2000. Recombinant Protein Expression in Pichia pastoris 16.

Damasceno, L.M., Huang, C.J., Batt, C.A., 2012. Protein secretion in Pichia pastoris and advances in protein production. Appl. Microbiol. Biotechnol. 93, 31–39. https://doi.org/10.1007/s00253-011-3654-z

Desgagné-Penix, I., Khan, M.F., Schriemer, D.C., Cram, D., Nowak, J., Facchini, P.J., 2010. Integration of deep transcriptome and proteome analyses reveals the components of alkaloid metabolism in opium poppy cell cultures. BMC Plant Biol. 10, 252. https://doi.org/10.1186/1471-2229-10-252

English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C., Gibbs, R.A., 2012. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS ONE 7, 1–12. https://doi.org/10.1371/journal.pone.0047768

Erickson, B., Nelson, Winters, P., 2012. Perspective on opportunities in industrial biotechnology in renewable chemicals. Biotechnol. J. 7, 176–85. https://doi.org/10.1002/biot.201100069

Ferrarini, M., Moretto, M., Ward, J.A., Šurbanovski, N., Stevanović, V., Giongo, L., Viola, R., Cavalieri, D., Velasco, R., Cestaro, A., Sargent, D.J., 2013. An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. BMC Genomics 14, 670. https://doi.org/10.1186/1471-2164-14-670

Fukuhara, H., 2006. Kluyveromyces lactis - A retrospective. FEMS Yeast Res. 6, 323–324. https://doi.org/10.1111/j.1567-1364.2005.00012.x

Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., Weissman, J.S., 2003. Global analysis of protein expression in yeast. Nature 425, 737–41. https://doi.org/10.1038/nature02046

Gray, A.N., Koo, B.-M., Shiver, A.L., Peters, J.M., Osadnik, H., Gross, C. a, 2015. High-throughput bacterial functional genomics in the sequencing era. Curr. Opin. Microbiol. 27, 86–95. https://doi.org/10.1016/j.mib.2015.07.012

Hamilton, S.R., Bobrowicz, P., Bobrowicz, B., Davidson, R.C., Li, H., Mitchell, T., Nett, J.H., Rausch, S., Stadheim, T.A., Wischnewski, H., Wildt, S., Gerngross, T.U., 2003. Production of complex human glycoproteins in yeast. Science 301, 1244–6. https://doi.org/10.1126/science.1088166

Hamilton, S.R., Gerngross, T.U., 2007. Glycosylation engineering in yeast : the advent of fully humanized yeast 387–392. https://doi.org/10.1016/j.copbio.2007.09.001

Heyland, J., Fu, J., Blank, L.M., Schmid, A., 2010. Quantitative physiology of Pichia pastoris during glucose-limited high-cell density fed-batch cultivation for recombinant protein production. Biotechnol. Bioeng. 107, 357–68. https://doi.org/10.1002/bit.22836

Idiris, A., Tohda, H., Kumagai, H., 2010. Engineering of protein secretion in yeast : strategies and impact on protein production. https://doi.org/10.1007/s00253-010-2447-0

Illanes, A., Cauerhff, A., Wilson, L., Castro, G.R., 2012. Recent trends in biocatalysis engineering. Bioresour. Technol. 115, 48–57. https://doi.org/10.1016/j.biortech.2011.12.050

Kirsch, R., Wielsch, N., Vogel, H., Svatoš, A., Heckel, D.G., Pauchet, Y., 2012. Combining proteomics and transcriptome sequencing to identify active plant-cell-wall-degrading enzymes in a leaf beetle. BMC Genomics 13, 587. https://doi.org/10.1186/1471-2164-13-587

Lanfranchi, E., Pavkov-Keller, T., Koehler, E.-M., Diepold, M., Steiner, K., Darnhofer, B., Hartler, J., Van, T., Bergh, D., Joosten, H.-J., Gruber-Khadjawi, M., Thallinger, G.G., Birner-Gruenberger, R., Gruber, K., Winkler, M., Glieder, A., 2017. Enzyme discovery beyond homology: a unique hydroxynitrile lyase in the Bet v1 superfamily OPEN. https://doi.org/10.1038/srep46738

Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Wang, Z., Yamashita, R.A., Zhang, D., Zheng, C., Bryant, S.H., 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res. 43, D222–D226. https://doi.org/10.1093/nar/gku1221

Mattanovich, D., Sauer, M., Gasser, B., 2016. Industrial Microorganisms: Pichia pastoris, in: Industrial Biotechnology. John Wiley & Sons, Ltd, pp. 687–714. https://doi.org/10.1002/9783527807796.ch19

Nicaud, J.-M., Madzak, C., van den Broek, P., Gysler, C., Duboc, P., Niederberger, P., Gaillardin, C., Broek, P. Van Den, Gysler, C., Duboc, P., Niederberger, P., Gaillardin, C., 2002. Protein expression and secretion in the yeast Yarrowia lipolytica. FEMS Yeast Res. 2, 371–9. https://doi.org/10.1111/j.1567-1364.2002.tb00106.x

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733–D745. https://doi.org/10.1093/nar/gkv1189

Ooyen, A.J.J. Van, Dekker, P., Huang, M., Olsthoorn, M.M.A., Jacobs, D.I., Colussi, P.A., Taron, C.H., 2006. Heterologous protein production in the yeast Kluyveromyces lactis 6, 381–392. https://doi.org/10.1111/j.1567-1364.2006.00049.x

Piotek, M., Hagedorn, J., Hollenberg, C.P., Gellissen, G., Srasser, A.W.M., 1998. Two novel gene expression systems based on the yeasts _Schwanniomyces occidentalis_ and _Pichia stipitis_. Appl. Microbiol. Biotechnol. 50, 331–338.

Schittmayer, M., Birner-Gruenberger, R., 2012. Lipolytic proteomics. Mass Spectrom. Rev. 31, 570–582. https://doi.org/10.1002/mas.20355

Steinborn, G., Böer, E., Scholz, A., Tag, K., Kunze, G., Gellissen, G., 2006. methylotrophic Hansenula polymorpha and other yeasts 13, 1–13. https://doi.org/10.1186/1475-2859-5-33

Sturmberger, L., Chappell, T., Geier, M., Krainer, F., Day, K.J., Vide, U., Trstenjak, S., Schiefer, A., Richardson, T., Soriaga, L., Darnhofer, B., Birner-Gruenberger, R., Glick, B.S., Tolstorukov, I., Cregg, J., Madden, K., Glieder, A., 2016a. Refined Pichia pastoris reference genome sequence. J. Biotechnol. https://doi.org/10.1016/j.jbiotec.2016.04.023

Sturmberger, L., Wallace, P.W., Glieder, A., Birner-Gruenberger, R., 2016b. Synergism of proteomics and mRNA sequencing for enzyme discovery. J. Biotechnol. 235, 132–138. https://doi.org/10.1016/j.jbiotec.2015.12.015

Uchiyama, T., Miyazaki, K., 2009. Functional metagenomics for enzyme discovery: challenges to efficient screening. Curr. Opin. Biotechnol. 20, 616–622. https://doi.org/10.1016/j.copbio.2009.09.010

Vogl, T., Hartner, F.S., Glieder, A., 2013. New opportunities by synthetic biology for biopharmaceutical production in Pichia pastoris. Curr. Opin. Biotechnol. 24, 1094–1101. https://doi.org/10.1016/j.copbio.2013.02.024

Xiao, R., Wilkinson, B., Solovyov, A., Winther, J.R., Holmgren, A., Lundström-Ljung, J., Gilbert, H.F., 2004. The contributions of protein bisulfide isomerase and its homologues to oxidative protein folding in the yeast endoplasmic reticulum. J. Biol. Chem. 279, 49780–49786. https://doi.org/10.1074/jbc.M409210200

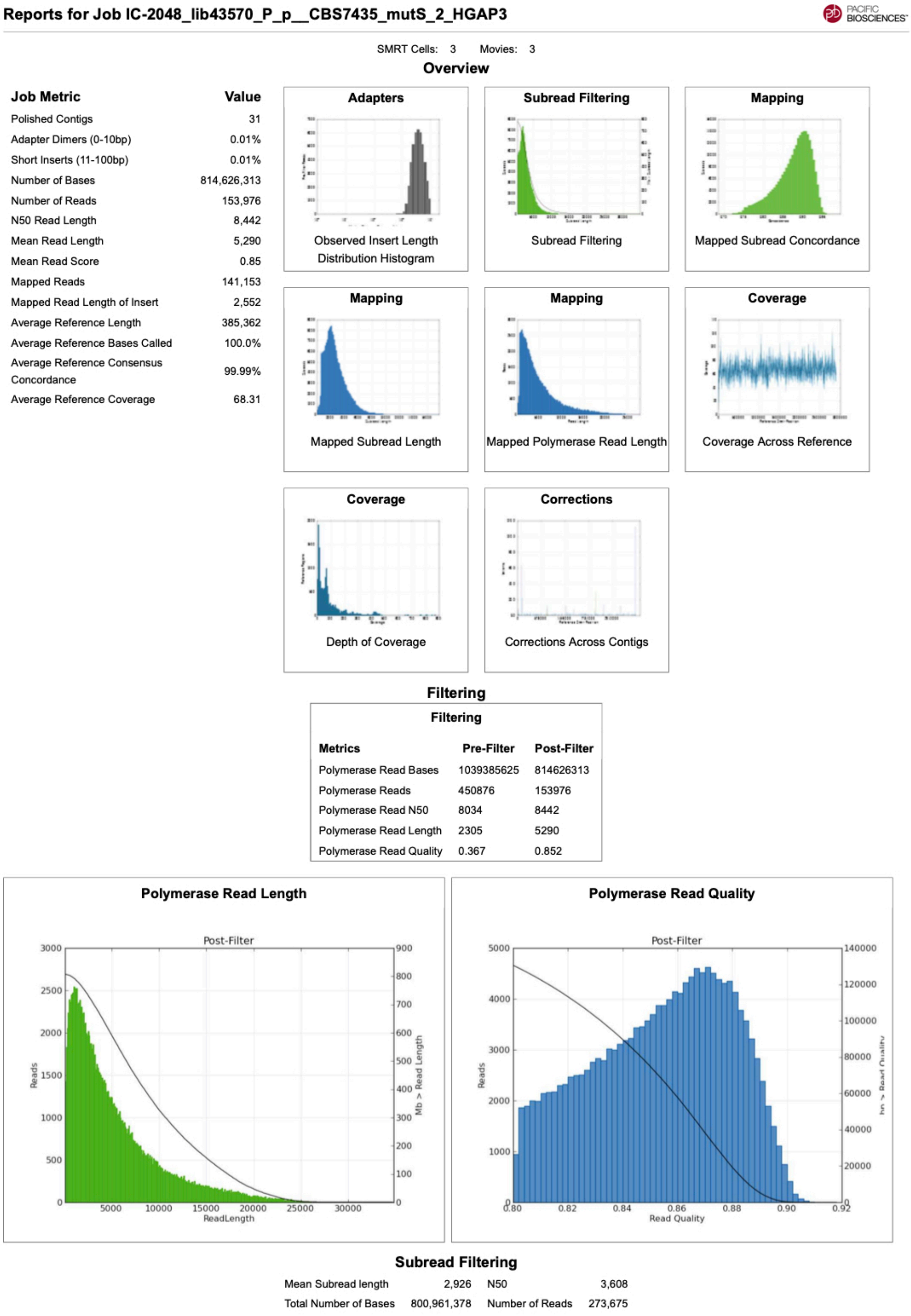
Yurimoto, H., Komeda, T., Lim, C.R., Nakagawa, T., Kondo, K., Kato, N., Sakai, Y., 2000. Regulation and evaluation of five methanol-inducible promoters in the methylotrophic yeast Candida boidinii. Biochim. Biophys. Acta 1493, 56–63.

Yurimoto, H., Sakai, Y., 2009. Methanol-inducible gene expression and heterologous protein production in the methylotrophic yeast Candida boidinii. Biotechnol. Appl. Biochem. 53, 85–92. https://doi.org/10.1042/BA20090030

Zhang, X., Davenport, K.W., Gu, W., Daligault, H.E., Christine Munk, A., Tashima, H., Reitenga, K., Green, L.D., Han, C.S., 2012. Improving genome assemblies by sequencing PCR products with PacBio. BioTechniques 53, 61–62. https://doi.org/10.2144/0000113891

Appendix

Assembly Report for K. phaffii CBS7435 mutS



Reports for Job IC-2048_lib43570_P_p__CBS7435_mutS_2_HGAP3

SMRT Cells: 3 Movies: 3

Overview

Job Metric	Value
Polished Contigs	31
Adapter Dimers (0-10bp)	0.01%
Short Inserts (11-100bp)	0.01%
Number of Bases	814,626,313
Number of Reads	153,976
N50 Read Length	8,442
Mean Read Length	5,290
Mean Read Score	0.85
Mapped Reads	141,153
Mapped Read Length of Insert	2,552
Average Reference Length	385,362
Average Reference Bases Called	100.0%
Average Reference Consensus Concordance	99.99%
Average Reference Coverage	68.31

PACIFIC BIOSCIENCES
Adapters
Observed Insert Length Distribution Histogram
Subread Filtering
Subread Filtering
Mapping
Mapped Subread Concordance
Mapping
Mapped Subread Length
Mapping
Mapped Polymerase Read Length
Coverage
Coverage Across Reference
Coverage
Depth of Coverage
Corrections
Corrections Across Contigs

Filtering

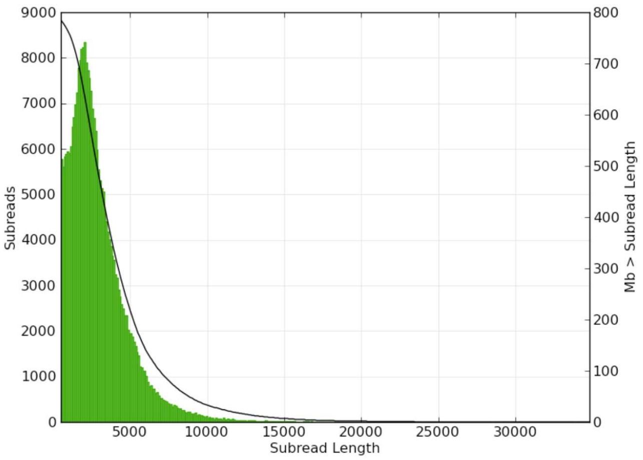
Filtering

Metrics	Pre-Filter	Post-Filter
Polymerase Read Bases	1039385625	814626313
Polymerase Reads	450876	153976
Polymerase Read N50	8034	8442
Polymerase Read Length	2305	5290
Polymerase Read Quality	0.367	0.852

Subread Filtering

Mean Subread length	2,926	N50	3,608
Total Number of Bases	800,961,378	Number of Reads	273,675

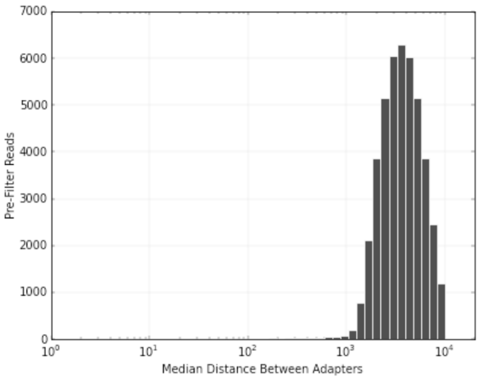
Subread Filtering



Adapters

Adapter Dimers (0-10bp)	0.01%
Short Inserts (11-100bp)	0.01%

Observed Insert Length Distribution



Loading

SMRT Cell ID	Productive ZMWs	ZMW Loading For Productivity 0	ZMW Loading For Productivity 1	ZMW Loading For Productivity 2
m140603_235014_42149_c100656362550000001823120810071490	150,292	44.65%	35.21%	20.14%
m140614_165525_42149_c100656232550000001823120810071455	150,292	39.68%	43.07%	17.26%
m140614_134246_42149_c100656232550000001823120810071454	150,292	36.33%	44.87%	18.79%

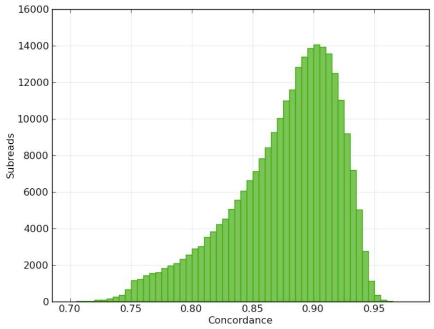
Mapping

Mapped Subread Length N50 (bp)	3,227	Mapped Polymerase Read Length 95% (bp)	14,760
Mapped Subread Length Mean (bp)	2,606	Mapped Polymerase Read Length Max (bp)	27,616

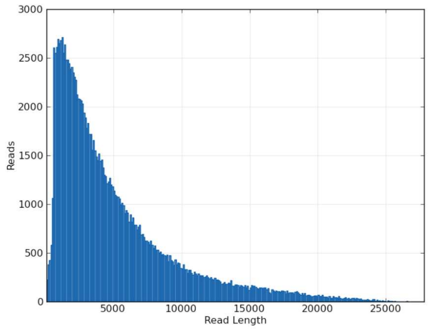
Mapping Stats Summary

Movie	Mapped Read	Mapped Polymerase Read Length	Mapped Polymerase Read Length N50	Mapped Subreads	Mapped Subread Bases	Mapped Subread Length	Mean Mapped Subread Concordance
All Movies	141,153	4,986	7729	261,372	681184688	2,606	0.871
m140603_235014_42149_c100656362550000001823120810071490_s1_p0	40,540	5,132	8189	77,461	201725749	2,604	0.865
m140614_134246_42149_c100656232550000001823120810071454_s1_p0	51,126	5,142	7963	95,826	253573439	2,646	0.873
m140614_165525_42149_c100656232550000001823120810071455_s1_p0	49,487	4,704	7148	88,085	225885500	2,564	0.874

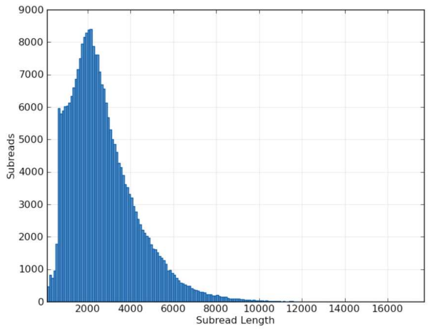
Mapped Subread Concordance



Mapped Polymerase Read Length



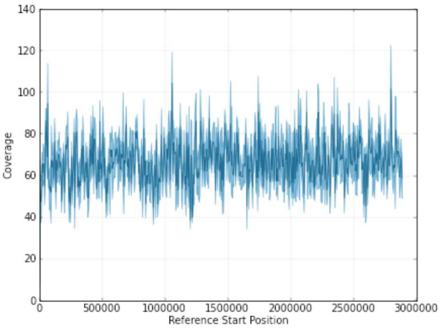
Mapped Subread Length



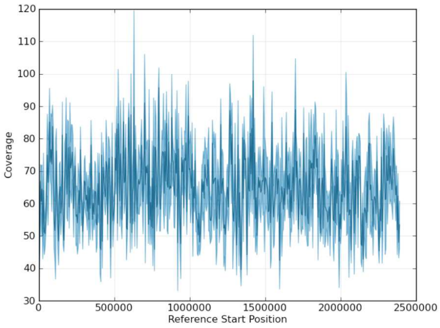
Coverage

Mean Coverage	68.31
Missing Bases (%)	0.0

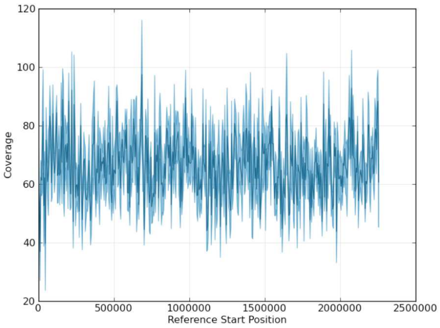
Coverage Across Reference



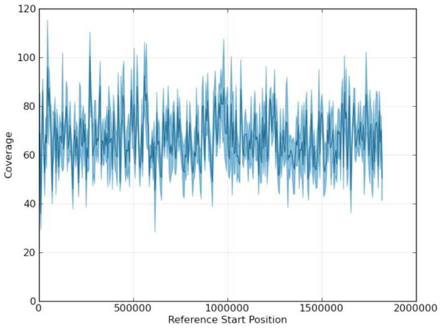
Observed depth of coverage across unitig_30 (window size = 5008bp).



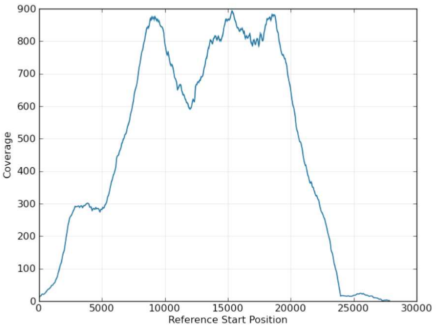
Observed depth of coverage across unitig_0 (window size = 5002bp).



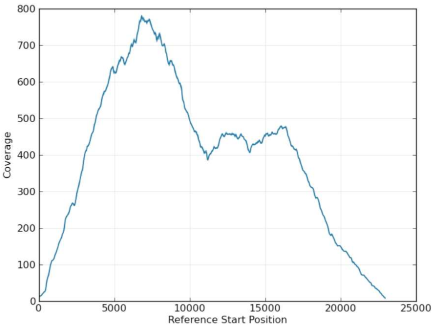
Observed depth of coverage across unitig_33 (window size = 5007bp).



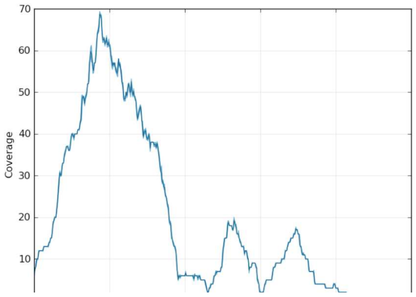
Observed depth of coverage across unitig_1 (window size = 5001bp).

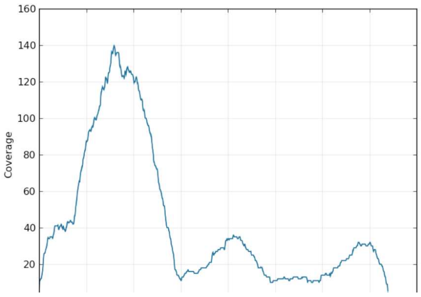


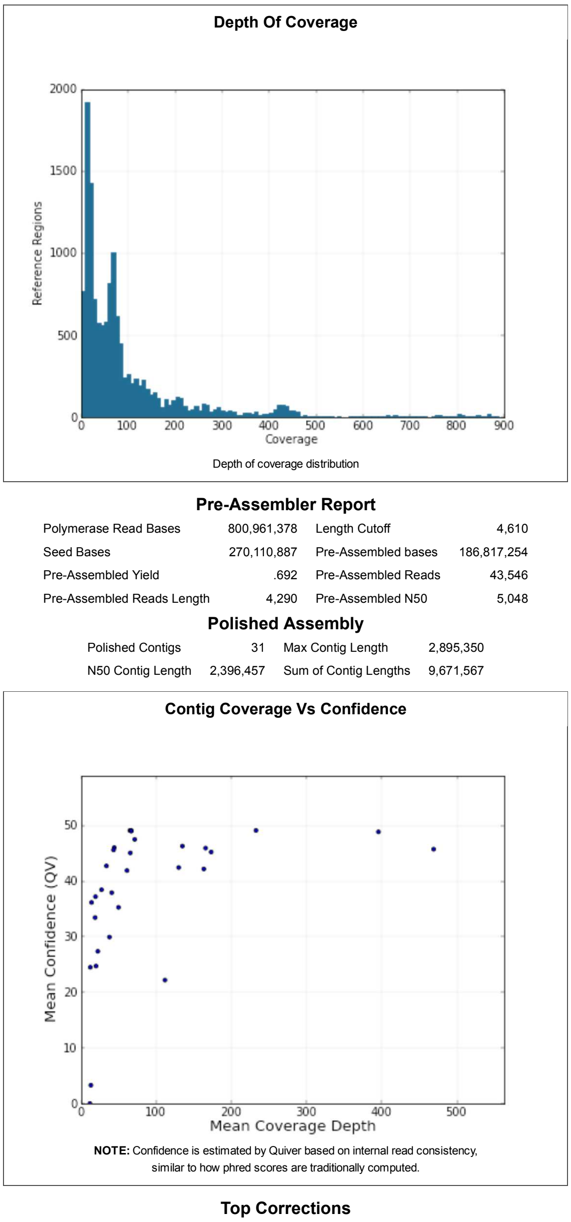
Observed depth of coverage across unitig_23 (window size = 50bp).



Observed depth of coverage across unitig_3 (window size = 50bp).







Depth Of Coverage

Depth of coverage distribution

Pre-Assembler Report

Polymerase Read Bases	800,961,378	Length Cutoff	4,610
Seed Bases	270,110,887	Pre-Assembled bases	186,817,254
Pre-Assembled Yield	.692	Pre-Assembled Reads	43,546
Pre-Assembled Reads Length	4,290	Pre-Assembled N50	5,048

Polished Assembly

Polished Contigs	31	Max Contig Length	2,895,350
N50 Contig Length	2,396,457	Sum of Contig Lengths	9,671,567

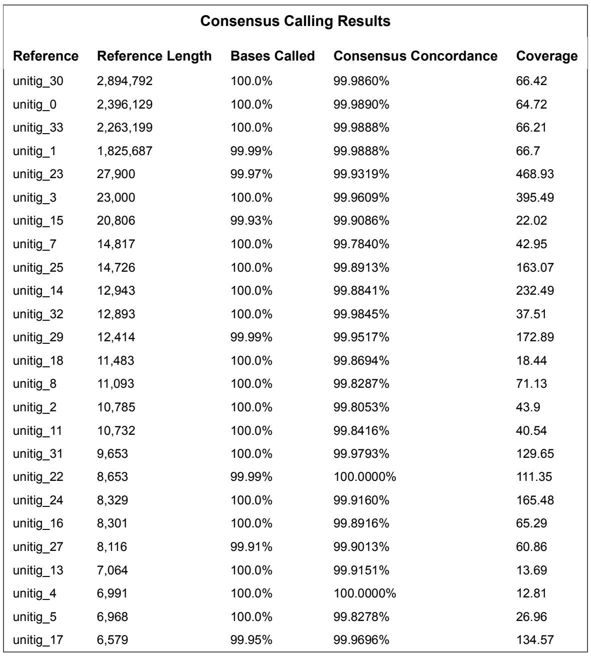
Contig Coverage Vs Confidence

NOTE: Confidence is estimated by Quiver based on internal read consistency, similar to how phred scores are traditionally computed.

Top Corrections

Sequence	Position	Correction	Type	Coverage	Confidence	Genotype
unitig_0	29,417	29417_29418insG	INS	49	53	haploid
unitig_33	1,519,582	1519582_1519583insA	INS	63	53	haploid
unitig_30	2,716,261	2716261_2716262insC	INS	49	53	haploid
unitig_1	1,525,638	1525638_1525639insG	INS	53	53	haploid
unitig_0	288,552	288552_288553insA	INS	67	52	haploid
unitig_0	747,500	747500_747501insC	INS	55	52	haploid
unitig_0	1,543,853	1543853_1543854insC	INS	65	52	haploid
unitig_23	24,971	24971_24972insA	INS	14	52	haploid
unitig_11	1,566	1566_1567insG	INS	11	52	haploid
unitig_33	10,135	10135_10136insG	INS	35	52	haploid
unitig_33	11,626	11626delGAC	DEL	27	52	haploid
unitig_33	248,758	248758_248759insG	INS	59	52	haploid
unitig_33	378,455	378455_378456insC	INS	66	52	haploid
unitig_33	468,465	468465_468466insG	INS	54	52	haploid
unitig_33	607,377	607377_607378insG	INS	55	52	haploid
unitig_33	862,983	862983_862984insC	INS	51	52	haploid
unitig_33	1,337,800	1337800_1337801insC	INS	63	52	haploid
unitig_14	188	188_189insT	INS	20	52	haploid
unitig_14	12,147	12147_12148insA	INS	40	52	haploid
unitig_30	547,093	547093_547094insG	INS	54	52	haploid
unitig_30	737,374	737374_737375insA	INS	52	52	haploid
unitig_30	1,245,070	1245070_1245071insG	INS	84	52	haploid
unitig_30	1,594,620	1594620_1594621insG	INS	54	52	haploid
unitig_1	92,484	92484_92485insC	INS	38	52	haploid
unitig_1	254,114	254114_254115insAC	INS	28	52	haploid
unitig_1	518,110	518110_518111insC	INS	65	52	haploid
unitig_1	756,027	756027_756028insC	INS	49	52	haploid
unitig_1	1,073,277	1073277_1073278insC	INS	70	52	haploid
unitig_1	1,286,623	1286623_1286624insG	INS	53	52	haploid
unitig_7	7,237	7237_7238insA	INS	20	52	haploid
unitig_0	1,054	1054_1055insA	INS	35	51	haploid
unitig_0	1,773	1773_1774insG	INS	41	51	haploid
unitig_0	138,820	138820_138821insT	INS	75	51	haploid
unitig_0	147,296	147296_147297insC	INS	60	51	haploid
unitig_0	288,568	288568_288569insGA	INS	66	51	haploid
unitig_0	297,309	297309_297310insA	INS	58	51	haploid
unitig_0	356,091	356091_356092insG	INS	45	51	haploid
unitig_0	459,005	459005_459006insG	INS	54	51	haploid
unitig_0	462,107	462107_462108insAC	INS	47	51	haploid
unitig_0	539,123	539123_539124insC	INS	55	51	haploid
unitig_0	550,942	550942_550943insG	INS	53	51	haploid
unitig_0	557,524	557524_557525insC	INS	64	51	haploid
unitig_0	629,885	629885_629886insT	INS	45	51	haploid
unitig_0	836,277	836277_836278insC	INS	82	51	haploid
unitig_0	894,874	894874_894875insG	INS	80	51	haploid
unitig_0	924,601	924601_924602insC	INS	56	51	haploid
unitig_0	953,756	953756_953757insC	INS	66	51	haploid
unitig_0	970,721	970721_970722insG	INS	87	51	haploid
unitig_0	1,057,746	1057746_1057747insG	INS	42	51	haploid
unitig_0	1,071,573	1071573_1071574insTT	INS	37	51	haploid
unitig_0	1,210,623	1210623_1210624insG	INS	74	51	haploid
unitig_0	1,236,414	1236414_1236415insG	INS	29	51	haploid
unitig_0	1,247,510	1247510_1247511insT	INS	69	51	haploid
unitig_0	1,274,453	1274453_1274454insC	INS	78	51	haploid
unitig_0	1,349,767	1349767_1349768insC	INS	71	51	haploid
unitig_0	1,520,805	1520805_1520806insG	INS	47	51	haploid
unitig_0	1,575,889	1575889_1575890insC	INS	61	51	haploid
unitig_0	1,758,425	1758425_1758426insC	INS	88	51	haploid
unitig_0	1,794,877	1794877_1794878insC	INS	43	51	haploid
unitig_0	1,809,564	1809564_1809565insC	INS	98	51	haploid
unitig_0	1,916,707	1916707_1916708insG	INS	45	51	haploid
unitig_0	1,954,042	1954042_1954043insT	INS	55	51	haploid
unitig_0	2,180,611	2180611_2180612insT	INS	61	51	haploid
unitig_0	2,280,649	2280649_2280650insT	INS	53	51	haploid
unitig_0	2,355,597	2355597_2355598insA	INS	67	51	haploid
unitig_0	2,386,097	2386097_2386098insA	INS	40	51	haploid

Corrections

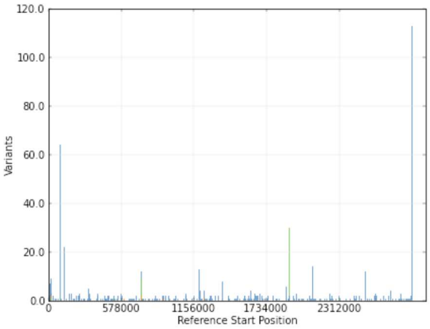


Consensus Calling Results

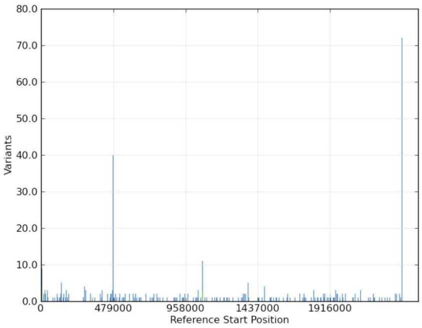
Reference	Reference Length	Bases Called	Consensus Concordance	Coverage
unitig_30	2,894,792	100.0%	99.9860%	66.42
unitig_0	2,396,129	100.0%	99.9890%	64.72
unitig_33	2,263,199	100.0%	99.9888%	66.21
unitig_1	1,825,687	99.99%	99.9888%	66.7
unitig_23	27,900	99.97%	99.9319%	468.93
unitig_3	23,000	100.0%	99.9609%	395.49
unitig_15	20,806	99.93%	99.9086%	22.02
unitig_7	14,817	100.0%	99.7840%	42.95
unitig_25	14,726	100.0%	99.8913%	163.07
unitig_14	12,943	100.0%	99.8841%	232.49
unitig_32	12,893	100.0%	99.9845%	37.51
unitig_29	12,414	99.99%	99.9517%	172.89
unitig_18	11,483	100.0%	99.8694%	18.44
unitig_8	11,093	100.0%	99.8287%	71.13
unitig_2	10,785	100.0%	99.8053%	43.9
unitig_11	10,732	100.0%	99.8416%	40.54
unitig_31	9,653	100.0%	99.9793%	129.65
unitig_22	8,653	99.99%	100.0000%	111.35
unitig_24	8,329	100.0%	99.9160%	165.48
unitig_16	8,301	100.0%	99.8916%	65.29
unitig_27	8,116	99.91%	99.9013%	60.86
unitig_13	7,064	100.0%	99.9151%	13.69
unitig_4	6,991	100.0%	100.0000%	12.81
unitig_5	6,968	100.0%	99.8278%	26.96
unitig_17	6,579	99.95%	99.9696%	134.57

Corrections Across Reference

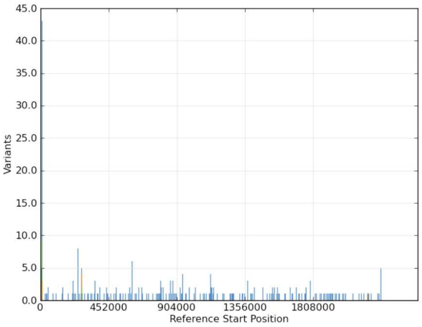
🟦 Insertions
🟩 Deletions
🟧 Substitutions



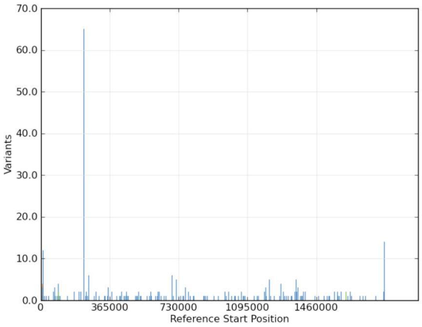
Observed variants across unitig_30.



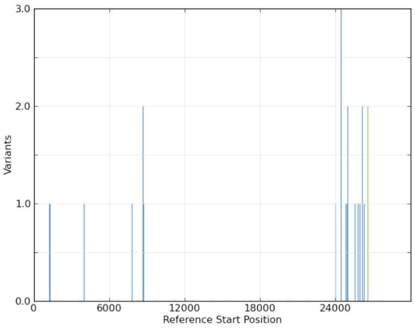
Observed variants across unitig_0.



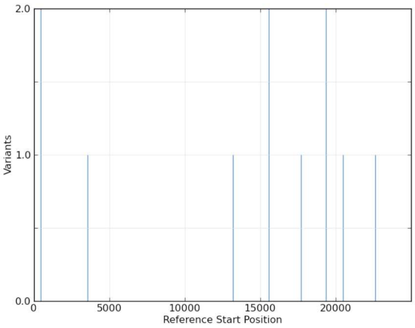
Observed variants across unitig_33.



Observed variants across unitig_1.



Observed variants across unitig_23.



Observed variants across unitig_3.

1. Mapping summary report

1.1 Summary statistics

	Count	Percentage of reads	Average length	Number of bases	Percentage of bases
References	2	-	11.270,00	22.540	-
Mapped reads	5.004.634	24,85%	101,00	505.468.034	24,85%
Not mapped reads	15.134.866	75,15%	101,00	1.528.621.466	75,15%
Total reads	20.139.500	100,00%	101,00	2.034.089.500	100,00%

1.2 Distribution of read length

Length	Count
101	20.139.500

1.3 Distribution of mapped read length

Length	Count
101	5.004.634

1.4 Distribution of un-mapped read length

Length	Count
101	15.134.866