Systematic investigation into the clonal variability of the non-conventional yeast *Pichia pastoris*

Dissertation

in fulfillment of the requirements for the degree

Doctor of Natural Sciences (Dr. rer. nat.)

Chair of Fermentation Engineering

Faculty of Technology

Bielefeld University

Submitted by

Jan-Philipp Schwarzhans

Bielefeld, September 2017

Chair of Fermentation Engineering

Faculty of Technology

Bielefeld University

Dissertation

in fulfillment of the requirements for the degree

Doctor of Natural Sciences (Dr. rer. nat.)

Systematic investigation into the clonal variability of the non-conventional yeast *Pichia pastoris*

Submitted by

Jan-Philipp Schwarzhans

Born in Hamburg on the 28th of June, 1987.

Bielefeld, September 2017

"*We are all agreed that your theory is crazy. The question that divides us is whether it is crazy enough to have a chance of being correct.*"

– Niels Bohr

The practical work of this thesis has been performed at the Chair of Fermentation Engineering at the Faculty of Technology and in the Research Group Microbial Genomics and Biotechnology of the Faculty of Biology at the Bielefeld University between 2013 and 2017 under the supervision of Prof. Dr. Karl Friehs and Prof. Dr. Jörn Kalinowski. The project was financially supported by the Cluster Industrial Biotechnology – Graduate Cluster (CLIB-GC) of the Federal State of North Rhine-Westphalia.

Reviewers:

Main Reviewer:	*Prof. Dr. Karl Friehs*
	Chair of Fermentation Engineering, Bielefeld University
Secondary Reviewers:	*Prof. Dr. Jörn Kalinowski*
	Research Group Microbial Genomics and Biotechnology at the Center for Biotechnology (CeBiTec), Bielefeld University
	Prof. Dr. Lars M. Blank
	Chair of Applied Microbiology, RWTH Aachen University

Printed on non-ageing paper according to ISO 9706

Parts of this thesis have been published in:

Schwarzhans, J.-P., Luttermann, T., Geier, M., Kalinowski, J., Friehs, K. **(2017)**. Towards systems-based metabolic engineering in *Pichia pastoris*. Biotechnology Advances 35, 681-710

Schwarzhans, J.-P., 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. Frontiers in Microbiology 8, 780. doi: 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. Scientific Reports 6, 38952. doi: 10.1038/srep38952

Schwarzhans, J.-P., Wibberg, D., Winkler, A., Luttermann, T., Kalinowski, J., Friehs, K. **(2016)**. Integration event induced changes in recombinant protein productivity in *Pichia pastoris* discovered by whole genome sequencing and derived vector optimization. Microbial Cell Factories 15, 84. doi: 10.1186/s12934-016-0486-7

Bachelor and Master theses supervised during the PhD project

No results from these theses are presented in this doctoral dissertation.

Liedtke, Maik **(2015)**. Rekombinante Expression einer α-Amylase aus *Bacillus licheniformis* und des grün fluoreszierenden Proteins (GFP) in der methylotrophen Hefe *Pichia pastoris*. Bachelor thesis

Kharatyan, Tigran **(2015)**. Produktion und Sekretion einer heterologen Xylanase in *Pichia pastoris*. Bachelor thesis

Topczewska, Patrycja **(2015)**. Heterologe Produktion und Sekretion eines Lysozyms in der methylotrophen Hefe *Pichia pastoris*. Bachelor thesis.

Hamidi, Jassin **(2016)**. Produktion und Sekretion einer heterologen Laccase (lcc2) in *Pichia pastoris*. Bachelor thesis

Wenzel, Carina **(2016)**. Charakterisierung und Identifizierung einer EDTA-abbauenden Spezies der *Pezizomycotina*. Bachelor thesis.

Widman, Christiane **(2016)**. Identifizierung von Stationärphasenpromotoren in *Pichia pastoris* mittels Transkriptomik. Master thesis

Luttermann, Tobias **(2016)**. Genomsequenzierung und Gesamtgenomtranskriptomik der biotechnologisch relevanten, nicht-konventionellen Hefe *Yarrowia lipolytica* DSM-3286. Master thesis

Yüzer, Ertug **(2017)**. EDTA-Abbau mit Hilfe eines neuartigen Pilzes. Bachelor thesis

Oral presentations:

Schwarzhans, J.-P., Luttermann, T., Wibberg, D., Winkler, A., Hübner, W., Huser, T., Kalinowski, J., Friehs, K. **(2017)**. 9th Conference on Recombinant Protein Production (RPP9), Dubrovnik, Croatia. Title "A Mitochondrial Autonomously Replicating Sequence from *Pichia pastoris* for Uniform High Level Recombinant Protein Production"

Schwarzhans, J.-P., Wibberg, D., Winkler, A., Luttermann, T., Kalinowski, J., Friehs, K. **(2016)**. ProcessNet-Jahrestagung und 32. DECHEMA-Jahrestagung der Biotechnologen 2016, Aachen, Germany. Title "Charting the clonal variability landscape of *Pichia pastoris*: Novel insights into integration events and their effects discovered by genome sequencing"

Schwarzhans, J.-P., Wibberg, D., Winkler, A., Luttermann, T., Kalinowski, J., Friehs, K. **(2016)** DECHEMA Himmelfahrtstagung: „New Frontiers for Biotech-Processes", Koblenz, Germany. Title "Connections between integration event and productivity: New insights into the protein production host *Pichia pastoris*"

Schwarzhans, J.-P., Wibberg, D., Winkler, A., Luttermann, T., Kalinowski, J., Friehs, K. **(2016)**. Pichia 2016, Antalya, Turkey. Title "Charting the clonal variability landscape of *Pichia pastoris*: Novel insights into integration events and their effects discovered by genome sequencing"

Poster presentations:

Schwarzhans, J.-P., Wibberg, D., Winkler, A., Luttermann, T., Kalinowski, J., Friehs, K. **(2016)**. 5th CeBiTec Scientific Advisory Board Meeting, Bielefeld, Germany. Title „Charting the clonal variability landscape in *Pichia pastoris* – New insights into protein productivity and off-target events"

Schwarzhans, J.-P., Wibberg, D., Winkler, A., Luttermann, T., Kalinowski, J., Friehs, K. **(2016)**. 6th International Bielefeld-CeBiTec Research Conference, Bielefeld, Germany. Title "Unraveling the clonal variability in *Pichia pastoris*: Effects of (non-canonical) integration events on productivity and morphology"

Schwarzhans, J.-P., Luttermann, T., Wibberg, D., Winkler, A., , Kalinowski, J., Friehs, K. **(2016)**. SAGS & SASBI conference „Bridging the gap: Decoding genetics for integration and application", Durban, South Africa. Title "Connections between integration event and productivity: New insights into the protein production host *Pichia pastoris*" (*presented by Luttermann, T.*)

Schwarzhans, J.-P., Wibberg, D., Winkler, A., Kalinowski, J., Friehs, K. **(2015)** DECHEMA Himmelfahrtstagung: „Scale-up and scale-down of bioprocesses", Hamburg, Germany. Title "Multiple and directed integration of recombinant protein expression cassettes into the genome of *Pichia pastoris*"

Schwarzhans, J.-P., Wibberg, D., Winkler, A., Kalinowski, J., Friehs, K. **(2015)** 10th CeBiTec Symposium: "Bioinformatics for Biotechnology and Biomedicine", Bielefeld, Germany. Title "Multiple and directed integration of recombinant protein expression cassettes into the genome of *Pichia pastoris*"

Table of contents

List of tables

Main body of text

Manuscript „Towards systems-based metabolic engineering in *Pichia pastoris*"

Manuscript "Integration event induced changes in recombinant protein productivity in *Pichia pastoris* discovered by whole genome sequencing and derived vector optimization"

List of figures

Main body of text

Manuscript „Towards systems-based metabolic engineering in *Pichia pastoris*"

Manuscript "Integration event induced changes in recombinant protein productivity in *Pichia pastoris* discovered by whole genome sequencing and derived vector optimization"

Manuscript "Non-canonical integration events in *Pichia pastoris* encountered during standard transformation analysed with genome sequencing"

Manuscript "A Mitochondrial Autonomously Replicating Sequence from *Pichia pastoris* for Uniform High Level Recombinant Protein Production"

Abstract

The non-conventional methylotrophic yeast *Pichia pastoris* has become a firmly established host for recombinant protein production in both the industry and academia. High product titers, an efficient secretory machinery and the ability to express complex proteins from bacterial to human origin have given *P. pastoris* an advantage over many other host systems. In recent years, its aptitude for foreign gene expression has also been applied in a rising number of metabolic engineering studies. However, scientists trying to create the *P. pastoris* strain optimal for their application are faced with a challenge. The high clonal variability results in clones from one transformation exhibiting wildly different expression levels, no detectable expression at all or altered growth behaviors. In consequence, a laborious screening process has to be applied to identify the desired strain from among hundreds or thousands of clones. Surprisingly, only few studies tried to analyze clonal variability in *P. pastoris* so far. Although the connections between gene dosage and product titers have been investigated thoroughly, the underlying causes and mechanisms of clonal variability remained unknown.

In this project, we present the first systematic investigation into the clonal variability of *P. pastoris*, the discovered genetic events and their impact on both recombinant protein production and growth behavior. By applying well-established standard methods for *P. pastoris* experiments, we aimed to provide relevant results and insights for other scientists working with this yeast. A library of 845 strains, transformed with an easy to detect reporter protein, was characterized for classic properties including colony morphology, gene dosage and productivity. Thereby, we analyzed a significantly larger clone library than previous *P. pastoris* publications, exceeding their size ca. 20 to 100 fold. Based on the characterization data, 31 strains with very peculiar features were selected for whole genome sequencing. Enabled by a combination of characterization and genome sequencing data, we discovered novel connections between integration event and strain properties.

A clear correlation between cassette-to-cassette orientation and productivity was found. Additionally, a surprising ratio between the different orientation forms suggested the existence of two competing integration mechanisms that excluded each other. We also observed a rather high occurrence of false-positive clones containing the same

integration event. Our combinatorial approach enabled us to identify a surplus homologous sequence inside the expression cassette as the likely cause for this secondary integration event. The theory was validated by optimization of the expression cassette and subsequent elimination of the undesired integration event.

Besides productivity related effects, we also analyzed strains that displayed a marked change in their colony morphology. Multiple new non-canonical integration events were discovered in them. Off-target gene disruptions could be correlated with the change in colony morphology. Particularly, the relocation of the knock-out target to a different chromosome and the subsequent gene disruption provided important insights for genetic engineering studies. In a number of clones we found *E. coli* DNA from the plasmid host, which had co-integrated in fusion with the expression cassette. Moreover, qRT-PCR experiments confirmed the transcriptional activity of the *E. coli* genes in *P. pastoris*.

Strikingly, the clonal variability also resulted in the creation of a novel genetic tool for recombinant protein production in *P. pastoris*. In one strain with exceptionally good productivity features, the creation of a circular plasmid consisting of the expression cassette and mitochondrial DNA was found. We could validate its replicative capabilities and successfully applied it for transformation of both *P. pastoris* and *Saccharomyces cerevisiae*. In *P. pastoris*, newly created pMito clones exhibited a highly uniform expression level that significantly exceeded a reference strain with a single copy of the expression cassette in its genome by up to fourfold.

Taken together, our project provides scientists working with *P. pastoris* with important references for studies both focused on recombinant protein production as well as genetic or metabolic engineering. Thereby, we aim to promote further development of this yeast and aid in the implementation of more complex genetic engineering strategies. Ways to reduce the frequency of low-producer strains enable streamlined screening procedures for high producer strains. Simultaneously, the documentation of off-target integration events helps to devise strategies that prevent their occurrence or highlight events that should be assayed for in constructed strains. Lastly, the novel episomal vector we discovered displayed great potential, especially for protein engineering studies in which a great number of different target variants need to be assayed.

1 Introduction

1.1 Yeasts and their biotechnological application

Biotechnology is the use of biological systems for the generation of products for human or animal use. Presently, such products can be found everywhere in daily life, the industry and medicine. Lipases and other enzymes enhance household detergents, versatile chemicals like citric acid are produced in large-scale operations with filamentous fungi and allogenic or even autogenic stem cells are applied in precision medicine. However, the history of biotechnology began much earlier.

It is suspected that in the Mesolithic time period (ca. 20,000 to 5,000 BC), when humans still lived a nomadic lifestyle, the first biotechnological products were discovered by accident [1]. Consumption of rotting fruits resulted in the first use of biotechnologically created foodstuffs. A conglomerate of various yeasts, filamentous fungi and bacteria can be found on the surface of most fruits. They likely facilitated the synthesis of ethanol from fruit sugars. Over the millennia, the process of brewing beer and wine from various sugar-rich substrates was explored. Aided by humans settling down and the rise of agriculture, various civilizations developed targeted brewing processes, e.g. in ancient Egypt, Rome and Greece. Unbeknownst to the brewers of the time, yeast was the key component necessary for successful fermentation and ethanol formation. This missing insight is exemplified by the original German "*Reinheitsgebot*" (English: purity law) from 1516, which stipulates that beer shall only be brewed using barley, hops and water but not yeast [2].

Yeasts are unicellular, lower eukaryotes. They belong to the kingdom fungi and are divided into the phyla ascomycetes and basidiomycetes. Ascomycetes propagate via sac shaped (Greek: *ascus*) spores that are generated endogenously, while basidiomycetes produce exogenous, club like (Greek: *basidium*) spores. Both are capable of asexual and sexual reproduction, but asexual reproduction plays only a minor role in basidiomycetes. The clear majority of known and applied yeast species, and the kingdom fungi in general, are ascomycetes [3]. Another common feature of most yeast is the so-called "budding". Budding yeasts (also called "true yeasts") form a small daughter cell (bud) on the mother cell during asexual reproduction. DNA, organelles and other important cellular components are duplicated and transferred from the mother cell to the bud. The bud

continues to grow on the parent cell and only separates when all essential components have been transferred [4]. On the other hand, fission yeasts don't form buds but rather divide into two daughter cells of equal size during asexual reproduction [5].

In nature, yeasts can be found in various environments ranging from foodstuffs to soils and animal guts [6, 7]. Common to these habitats is a high-abundance of sugar-rich materials that serve as their primary carbon source. Interestingly, although the number of known yeast species is steadily increasing (Figure 1) their exact ecology is not yet fully elucidated [7]. Newly discovered species often exhibit interesting properties suitable for biotechnological application [8, 9] or stem from far less sampled areas, e.g. the deep sea or atopic diseases in humans [10, 11]. The discovery of yeast species in the guts of mushroom-feeding insects led to a dramatic increase in biodiversity, owed to the low global diversity of previous sampling sites [12, 13]. It is estimated that, recollecting more yeast species from insect guts of the original site could lead to a ca. 50 % increase in total described yeast species [13].

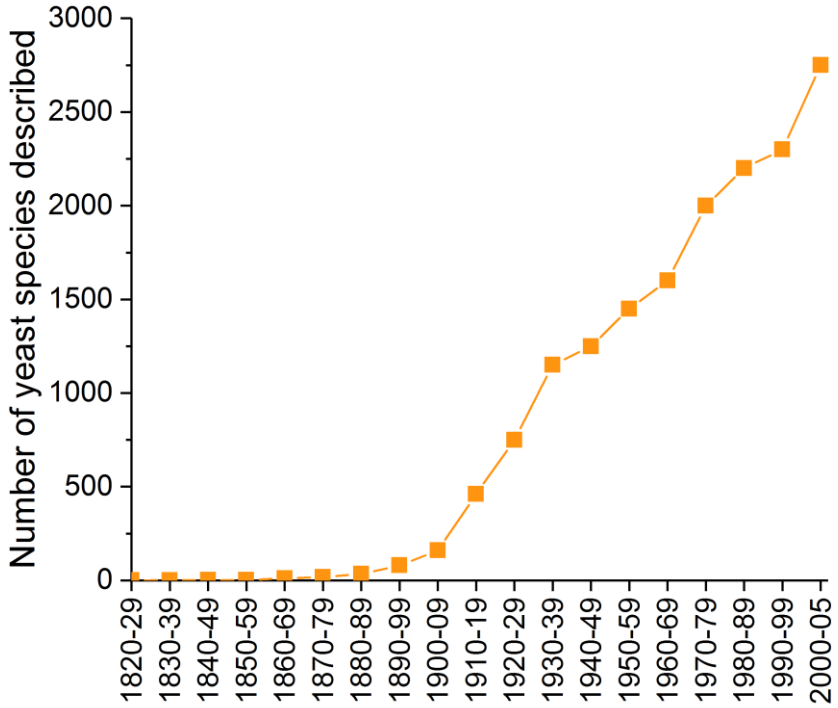

Figure 1: Illustration of the steady increase in described yeast species, based on Boekhout, 2005 [7]. After a stagnation between 1980 and 1999, the discovery of highly diverse yeasts in the gut of mushroom-feeding insect led to a drastic rise in the 2000s.

Especially the ascomycete, budding yeast *Saccharomyces cerevisiae* was unknowingly used for thousands of years. Originally found on grape skins, it was applied in wine and beer fermentation and later in bread making. During the 19th century the involvement of the microorganism in these processes was first discovered and soon after began the selection and cultivation of strains better suited for specific applications. Methods for isolating pure strains were published and various companies began the production of specialized yeast cultivation vessels [14]. This early popularity led to *S. cerevisiae* becoming one of the first model microorganisms. Ease of cultivation and genetic manipulation, the possibility for both sexual and asexual reproduction, combined with an already established economic interest made *S. cerevisiae* an ideal candidate for scientific studies. According to the confederation of European yeast producers (COFALEC, www.cofalec.com), about one million tons of yeast are produced annually in the EU. In contrast to bacterial systems like *Escherichia coli*, yeasts contain the same organelles and regulatory capabilities found in most higher eukaryotes. Therefore, many insights obtained in yeasts are also applicable to mammalian or plant systems. The short generation time and high robustness of yeast cells made them a highly viable alternative to the slow growing and sensitive animal or plant cell cultures.

Over the years, a large scientific community elected *S. cerevisiae* as their host-system of choice. Basic science studies conducted with this yeast analyzed chromosome and centromere organization [15, 16], the regulation and function of various genes and proteins [17–20] as well as the reproductive cycles [21, 22], amongst other topics. An international collaboration resulted in the full genome sequence of the reference strain S228C in 1995 [23], making it the first published complete genome of eukaryotic origin. But also from a biotechnological standpoint, *S. cerevisiae* has been applied for a variety of purposes.

Strains suitable for beer and wine fermentation have been well characterized and new ones created, e.g. via cross breeding existing strains [24, 25]. Furthermore, yeast cells can be disrupted and their cell walls removed to create yeast extract. Yeast extract is a glutamate- and protein-rich complex substrate suitable for multiple purposes ranging from use as a culture media supplement to whole-cell extract enabling *in vitro* enzymatic reactions [26, 27]. However, its main use is the addition to food for enhanced flavor and

texture [28]. The natural aptitude of *S. cerevisiae* for ethanol production, e.g. under high glucose and aerobic conditions (Crabtree effect), has been exploited for bio-ethanol production, a promising renewable energy source [29].

For food applications, mutagenesis methods are preferred due to law restrictions [30]. Besides exploiting its native capabilities, *S. cerevisiae* has also been applied in the production of heterologous targets. Targeted genetic manipulation was explored early and multiple techniques are currently available, including genomic integration cassettes and episomal vectors [31–33]. Modern technologies like CRISPR/Cas9 have been proven and refined for *S. cerevisiae* [34, 35] and *in vivo* assembly techniques enable rapid construction of vectors [32, 36]. In fact, *S. cerevisiae* is typically the first eukaryotic organism, if not even the first organism overall, where novel genetic manipulation techniques are being established [34, 37, 38].

These genetic manipulation techniques have been used to insert foreign genes, delete native ones or alter their expression levels. Thereby, strains were created for the production of recombinant proteins [39], synthesis of valuable metabolites [40] or the consumption of previously inaccessible substrates [41]. Particularly, metabolic engineering of *S. cerevisiae* was actively pursued. The ease of genetic manipulation and well-developed post-translational modification apparatus paired with simple and efficient cultivation techniques made it a suitable host for transferring heterologous pathways. In consequence, substances ranging from plant terpenoids [42, 43], to fungal polyketides [44] and platform chemicals [45] have been produced in *S. cerevisiae*. The broad scope of biofuels, bulk and fine chemicals that have been produced in this yeast has been reviewed by Hong & Nielsen (2012) [46]. On the other hand, recombinant protein production studies with *S. cerevisiae* are surprisingly sparse. In the last years, only a single-digit percentage of studies employed this yeast in favor of e.g. *E. coli* or mammalian cell-lines [47]. Problems regarding inefficient protein secretion and hyper-mannosylation of glycoproteins complicate the expression of heterologous targets [48, 49]. In consequence, scientists started exploring other yeast species for their biotechnological potential.

1.2 Non-conventional yeasts

Due to its prevalence in science, industry and daily life, *S. cerevisiae* has become almost synonymous with the term yeast. However, as mentioned previously, a multitude of other yeast species exists. Since *S. cerevisiae* was the first and remains the most used yeast by mankind, it has become a "conventional yeast". The model fission yeast *Schizosaccharomyces pombe* is the only other species classified as conventional, while all other yeast species are classified as "non-conventional". No genetic, morphologic or other phylogenetic criteria were used for this differentiation. It simply divides the two first discovered and applied yeast species from all others that followed. As shown in figure 1, close to 3,000 yeast species have been described. From this diverse group, many candidates for biotechnological application have emerged [50]. Figure 2 gives an overview of some of the most important conventional and non-conventional yeast species and their phylogenetic relationship to each other.

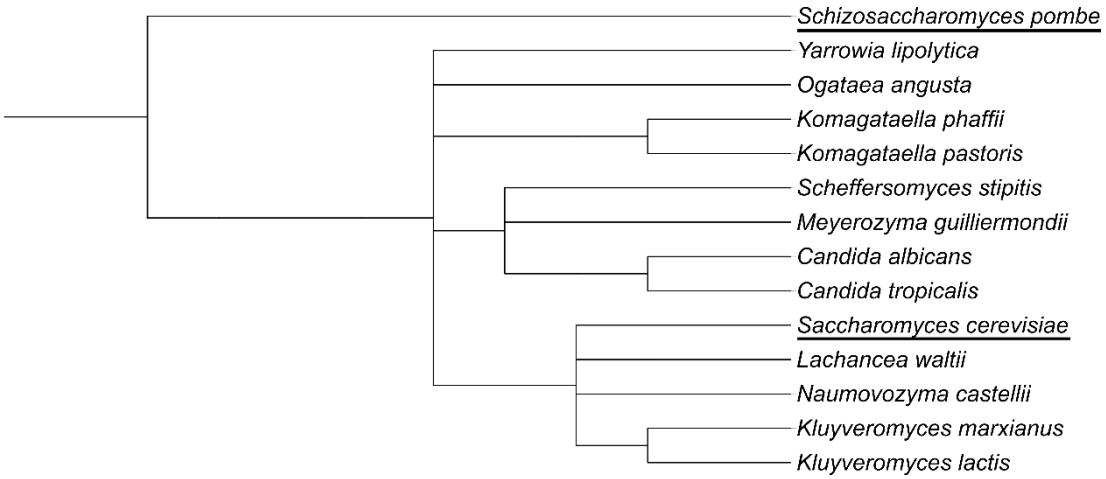

Figure 2: Phylogenetic tree of selected conventional and non-conventional yeast species. The tree was generated via the online tool phyloT (phylot.biobyte.de, version 2017.1) and visualized in iTOL (Interactive Tree of Life, version 3.5.2) [51]. Organization of the tree is based on information from the NCBI taxonomy database, which applies sequence data from GenBank for curated classification and nomenclature of organisms. The conventional yeasts *S. cerevisiae* and *S. pombe* have been underlined.

In recent years, an increased diversification of yeast biotechnology beyond *S. cerevisiae* and *S. pombe* has been postulated [52, 53]. While both systems are well understood and offer very good genetic tractability, some non-conventional yeasts outperform them in certain aspects. The wide range of described non-conventional yeasts contains members

adapted to different environments, showcasing distinct morphologies and having abilities beyond those of the conventional yeasts. In the following paragraphs, a few selected non-conventional yeasts with biotechnological relevance are presented. It should be mentioned that often the classification of these yeasts changed over time. Early classification relied primarily on morphological evidence. The supplementation with DNA sequencing information revealed a much wider diversity of yeast species, and made it necessary to divide previously related species into separate genera [54, 55]. Therefore, some species will be named with their current, corrected classification and the traditional one, which often is still the prevalent one in scientific communications.

Within the clade of budding yeasts, *Kluyveromyces lactis* is one of the closest relatives to *S. cerevisiae* that is biotechnologically applied. Transfer of molecular and cultivation methods between these two yeasts is rather straightforward and thus *K. lactis* gained early popularity, e.g. for recombinant protein expression [56]. However, concerns regarding low product titers and genetic integrity of production strains led to a decrease in scientific interest during the 2000s [57]. Its sibling *K. marxianus* shows promise for biotechnological application, in particular for bio-ethanol production. The high temperature resistance, lower tendency for fermentative growth and capability to utilize lignocellulosic carbon sources that are inaccessible to e.g. *S. cerevisiae* make it a suitable candidate for this purpose [58, 59].

Many *Candida* species are pathogenic and associated with human diseases, which are known as Candidiasis fungal infections [60]. Therefore, research on these yeasts is of great interest from a medical perspective. *C. albicans* is the most prominent Candida species and has been thoroughly studied. For example, the effect of gene deletions on cell wall integrity and virulence, as well as the interplay with other infections has been analyzed [61, 62]. Until recently, genetic manipulation of *C. albicans* was complicated by the absence of functional episomal vectors, a diploid genome and lack of meiotic phase. The recent establishment of CRISPR/Cas9 techniques will greatly help to better understand this pathogenic yeast on a basic level, which could help in the search of drug targets for treating Candidiasis patients [63]. However, due to its human pathogenicity *C. albicans* and other *Candida* species are typically not used for biotechnological purposes.

Yarrowia lipolytica first gained popularity due to its excellent lipid degradation and accumulation capabilities [64]. Furthermore, its dimorphic phenotype (yeast-like and (pseudo-)hyphae) was studied in detail [65]. In contrast to the dimorphic *C. albicans*, *Y. lipolytica* is not pathogenic, likely owed to a strict aerobic lifestyle and the inability to grow at temperatures >34 °C. Multiple processes involving *Y. lipolytica* have been classified as GRAS (generally regarded as safe) by the FDA (Food and Drug Administration, USA) [66]. Interestingly, *Y. lipolytica* is the only known member of its genus. Unlike most other yeasts, it exhibits a high GC content (ca. 49 %), high intron frequency and low similarity to homologous genes from other yeasts. Consequently, an early evolutionary branching off from *S. cerevisiae* and related yeasts is suspected [67]. It can grow on a range of (crude) carbon sources to high cell densities and exhibits an efficient carbon metabolism. In consequence, the popularity of *Y. lipolytica* for metabolic engineering projects steadily rose in recent years [68]. Often, the target compound can be derived from the β-oxidation pathway of fatty acids [69, 70]. Recently, CRISPR/Cas9 methods have been established and should significantly simplify further metabolic engineering studies [71].

In the 1970s *Ogataea angusta* (also known as *Hansenula polymorpha* or *Pichia angusta*) was discovered and soon applied for recombinant protein production. Compared to *S. cerevisiae,* far higher protein titers were possible [67]. *O. angusta* is a methylotrophic yeast, meaning it can utilize methanol as sole carbon source. The application of suitable promoters enabled methanol inducible foreign gene expression. Consequently, it also served as a model organism for studying methanol utilization and the peroxisome [72, 73]. For recombinant protein production, the target gene is typically integrated into the genome. Specialized techniques are available, enabling high dosage and thereby likely higher productivity [74, 75]. But issues regarding genetic accessibility and stability led to a stagnation in *O. angusta* scientific output. Instead, a different species of methylotrophic non-conventional yeast gained more traction as a host for recombinant protein production.

1.3 *Pichia pastoris*

Although originally isolated in France, currently used strains of the methylotrophic yeast *Pichia pastoris*, stem from California and were isolated by Herman Phaff and coworkers in the 1950s and 60s [76]. Owing to the scientific limitations of the time, many newly discovered yeast species were first erroneously assigned to the *Pichia* genus. Along with *P. pastoris*, other often applied species had to be moved to a new genus, including *Scheffersomyces stipites* (previously *P. stipites*) and *Meyerozyma guilliermondii* (previously *P. guilliermondii*) [54]. In the case of *P. pastoris,* sequencing work resulted in the reclassification of commonly applied *P. pastoris* strains into the closely related genera *Komagataella phaffii* and *K. pastoris* [77, 78]. However, over the years the original classification had engrained itself in the scientific community. Therefore, it remained common practice to use the term *P. pastoris* instead of *K. phaffii* or *K. pastoris* in scientific publications. The remainder of this work will adhere to this principle.

First reports of biotechnological application of *P. pastoris* can be found in the 1970s, when the Philipps Petroleum company used the ability of this yeast for generation of single cell protein on the cheap substrate methanol [79]. The yeast exhibits robust cultivation properties and can be grown to very high dry cell weight concentrations (>100 g/L) using uncomplicated fed-batch procedures. While growth on substrates like glucose and glycerol is possible, Philipps Petroleum harnessed its supply of inexpensive methanol as primary carbon source to increase profitability. Unfortunately, in the October of 1973 the Organization of Arab Petroleum Exporting Countries (OPEC) declared an oil embargo, leading to an international oil crisis and approximately fourfold increase in global oil prices. The oil crisis also caused the methanol price to dramatically increase, since methanol is predominantly synthesized from natural gas (methane), making *P. pastoris* cultivations on methanol media economically unfeasible. By the end of the embargo in the spring of 1974, Philipps Petroleum had stopped production of *P. pastoris* single cell proteins. Yet, the excellent cultivation properties of *P. pastoris* urged the scientists working at Philipps Petroleum to investigate other possible applications.

For its growth on methanol *P. pastoris* requires an effective methanol utilization apparatus. The first step in the methanol metabolism is the oxidation to formaldehyde

by an alcohol oxidase (*AOX*). *P. pastoris* contains two isoforms *AOX1* and *AOX2*, of which *AOX1* is expressed at significantly higher levels. The high growth rate of this yeast on methanol was accredited to high expression levels of its *AOX* genes [80]. This information was used as basis for establishing *P. pastoris* for recombinant protein production. The reasoning being that an organism adept at expressing high levels of an endogenous protein should also be capable of generating high product titers for foreign targets. Led by the "founding father" of *P. pastoris* biotechnology, James M. Cregg, transformation techniques were established and genes involved in methanol metabolism applied for the expression of heterologous genes [81–83]. Especially the promoter of *AOX1* (*pAOX1*) showcased favorable characteristics. Under repressed conditions, e.g. via glucose or glycerol, transcription is close to the lower detection limit and does not increase markedly even if the repressor has been exhausted (derepression). Induction is facilitated by methanol addition and leads to a drastic rise in transcription activity, with *aox1* accounting for up to 30 % of total soluble protein under methanol induced conditions [80]. Interestingly, a major reason for the high expression levels of *aox1* in *P. pastoris* is the low affinity of the protein to O_2, which is required for cofactor regeneration. It appears that the strategy of *P. pastoris* to compensate for this deficiency is to simply express more *aox1* [84]. In consequence, the organism has evolved its cellular machinery to allow for efficient production of high protein levels. Different vectors and strains were developed to allow use of *pAOX1* driven foreign gene expression [85].

Transformation of *P. pastoris* is eased by the stable vegetative haploid state, circumventing problems of different chromosomal alleles in diploid yeasts like *S. cerevisiae* and *O. angusta*. Only under specific stress situations does *P. pastoris* initiate sexual propagation, leading to a diploid phenotype [86]. Ease of transformation and efficient expression of foreign genes from bacteria, fungi, yeasts, plants and animals promoted the rise in popularity of *P. pastoris* as a recombinant protein production platform [84, 85]. This trend was further supported by the acquisition of a license for commercial distribution of *P. pastoris* strains and vectors by Invitrogen in 1993. Various recombinant proteins were produced in a g/L scale, often in secreted form. The emergence of the first GRAS sanctioned *P. pastoris* process and FDA approved biopharmaceuticals, widened the potential from industrial to food and pharmaceutical applications [87, 88]. Currently, *P. pastoris* ranks among the most applied non-

conventional yeasts. As per the website www.pichia.com, over 5000 different heterologous proteins have been produced to date. The popularity is exemplified by the rising yearly publication and citation rate (Figure 3). *P. pastoris* even outperforms the conventional yeast *S. pombe*, for which interest appears to have been declining since 2000. In contrast, *O. angusta* and *K. lactis* display stagnating frequencies. *Y. lipolytica* shows a promising upwards trend, fueled by multiple studies demonstrating its suitability to produce different value-added metabolites. It has to be noted, that the non-conventional yeast *C. albicans* averages at 520 publications per year over the last ten years. As mentioned before, scientific interest in *C. albicans* is chiefly justified by its human pathogenicity. It was therefore excluded from this comparison, which focuses on biotechnological application. Furthermore, *S. cerevisiae* remains by far the most popular yeast in science. On average, 900 new studies are published per year. The real number of yearly publications is likely higher, since many studies simply use "yeast" in their title when describing experiments with *S. cerevisiae*. While *P. pastoris* and other non-conventional yeast are gaining in popularity, it is unlikely that they'll overtake *S. cerevisiae* in the foreseeable future.

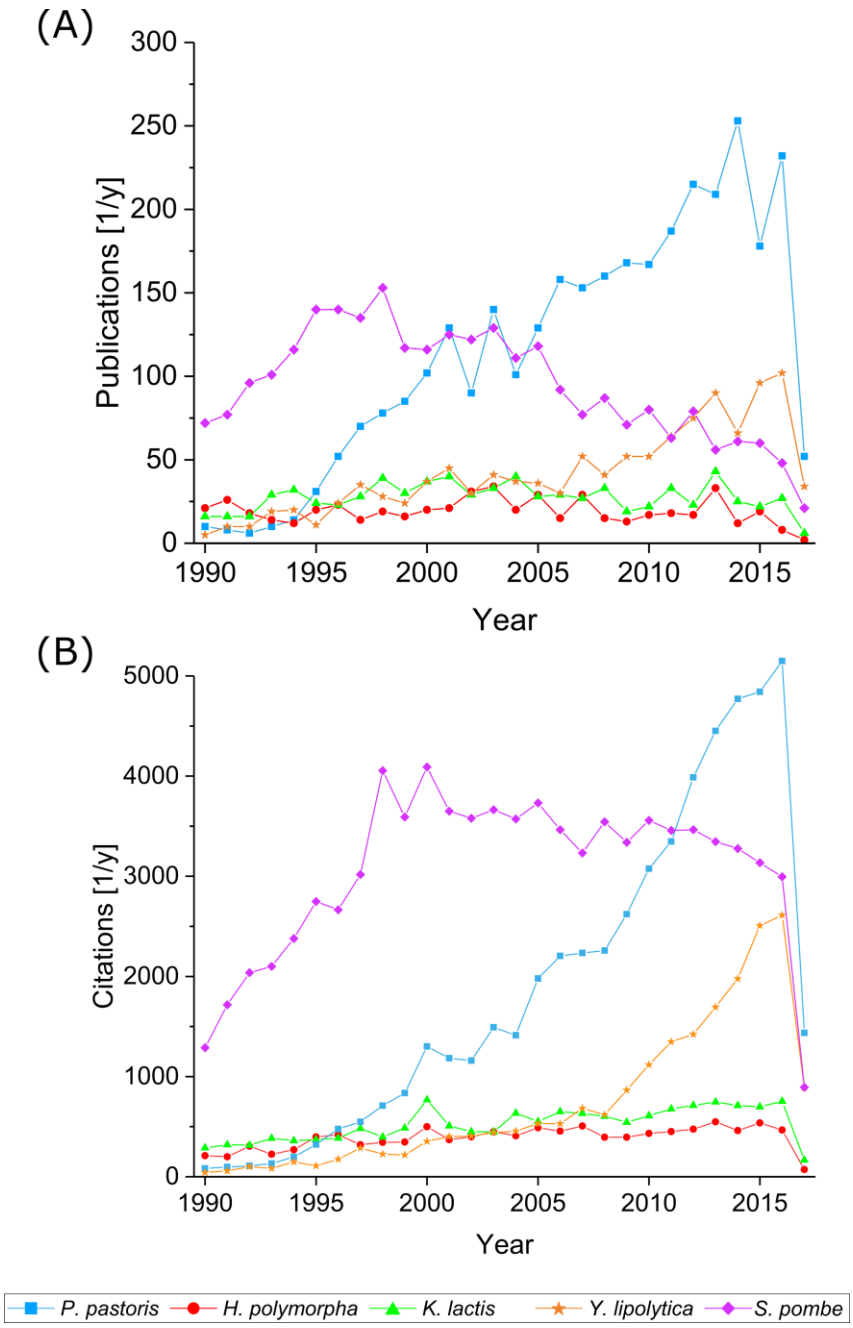

Figure 3: Publications **(A)** and citations **(B)** per year of selected popular (non-) conventional yeast species from 1990 till May 2017. The following yeasts are shown: *Komagataella phaffii* (*Pichia pastoris*), *Ogataea angusta* (*Hansenula polymorpha*), *Kluyveromyces lactis*, *Yarrowia lipolytica* and *Schizosaccharomyces pombe*. The data was collected via Web of Science™ (Version 5.24, Thomson Reuters, www.webofknowledge.com) on the 09.05.2017, searching for publications with the respective organism name in their title and citations of these articles. For *O. angusta* and *K. phaffii* the more commonly used terms (in parenthesis) were used as search query.

1.3.1 Recombinant protein production

There are two main reasons for the popularity of *P. pastoris*. Its methylotrophic lifestyle enabled its establishment as a model organism for studying the peroxisomal machinery. Under inducing conditions (e.g. methanol or oleate), the peroxisome can grow to several times its normal size, filling up to 80 % of the total cell volume [73]. Via knock-out studies, genes and their functions in peroxisome biogenesis, as matrix proteins and in trafficking processes across the peroxisomal membrane have been elucidated [89–91]. This topic and its implications for metabolic engineering of *P. pastoris* are discussed in more detail in chapter 1.3.3. However, *P. pastoris* suitability for recombinant protein production was the primary motivation for many scientists to apply this non-conventional yeast. The high success rate for heterologous protein synthesis led to the recommendation to include *P. pastoris* as standard tool in labs interested in protein studies [47]. Table 1 summarizes the key features of *P. pastoris* for recombinant protein production and compares them to other common expression platforms.

Table 1: Comparison of key characteristics of commonly applied hosts for recombinant protein expression. Information was gathered from Hoeffler (1999) [92] and Brondyk (2009) [93]. Abbreviations: CHO = Chinese hamster ovary cells; LPS = Lipopolysaccharides.

	E. coli	***S. cerevisiae***	***P. pastoris***	**Insect cells**	**CHO**
Viral infections	Bacteriophage	No	No	Virus	Virus
Endotoxins	LPS	No	No	No	No
Secretion capabilities	Low	Medium	Medium to high	High	High
Growth rate	High	Medium	Medium	Low	Low
Medium complexity	Low	Low	Low	High	High
Product titer	Medium to high	Low to medium	Medium to high	Low to high	Low to medium
Process complexity	Low	Low	Low to medium	High	High
Post-translational modifications					
Protein folding	Low	Medium	Medium to high	High	High
Glycosylation	No	Yes (high mannose)	Yes (high mannose)	Yes (no sialylation)	Yes
Phosphorylation	No	Yes	Yes	Yes	Yes
Acetylation	No	Yes	Yes	Yes	Yes
Acylation	No	Yes	Yes	Yes	Yes
γ-Carboxylation	No	No	No	No	Yes

Much effort has been put into enhancing positive traits and rectify shortcomings of heterologous protein production in *P. pastoris*. These efforts, particularly the genetic engineering of fully humanized N-glycosylation and improved secretion rates, are detailed in chapter 1.3.3. Here, the more general features regarding recombinant protein production are presented. *P. pastoris* combines some of the strengths of both traditional bacterial and higher eukaryotic systems. Media complexity, and therefore its cost, is comparable to that used for *E. coli* and far less complex than for mammalian or insect cell cultures. In contrast to many bacterial systems, *P. pastoris* is capable to secrete recombinant proteins directly into the medium, at a g/L-scale. Higher eukaryotic systems are vulnerable to viral infections, compromising applicability for

pharmaceutical processes. *E. coli* can be infected by bacteriophages and contains lipopolysaccharides which elicit strong immunogenic reactions in humans. No such issues exist in yeasts. Furthermore, protein folding is more efficient in *P. pastoris* than in *E. coli*, displaying a higher aptitude to functionally express proteins of bacterial to human origin. Most post-translational modifications can be performed by *P. pastoris*, excluding γ-carboxylation. In a direct comparison to *S. cerevisiae*, *P. pastoris* routinely delivers higher product titers and secretion rates, as well as a higher success rate for difficult to express targets, e.g. membrane proteins [85, 94, 95]. Consequently, the ratio of studies utilizing *P. pastoris* for recombinant gene expression has been steadily rising since from 5 % in 1995 to 11 % in 2014 [47].

Nevertheless, *P. pastoris* also has its shortcomings. While the doubling time of ca. 2 h is shorter than for cell cultures (ca. 24 h), *E. coli* doubling time of 20 – 30 min enables a better turnaround time. Cultivation procedures for *P. pastoris* can pose more challenges than for *E. coli* or *S. cerevisiae*, especially when facilitating protein production via methanol-induced *pAOX1* expression. During the methanol induction phase the O_2-demand of the culture rises drastically and high amounts of energy in the form of heat are emitted. This is particularly true for the commonly applied high cell-density fed-batch procedures. In consequence, more care during the planning and execution stage is needed, to ensure optimal productivity [96]. For heterologous proteins of bacterial origin, *E. coli* is often better suited, as the use of *P. pastoris* does not necessarily provide a benefit for enzymes with no or low level of post-translational modifications. Although, glycosylation is possible, the high mannose type is not suitable for pharmaceutical applications. *P. pastoris* glycosylation was engineered to fully resemble that of humans, including terminal sialylation, but also results in a markedly reduced strain fitness [97, 98]. Lastly, it has been shown that secretion rates in *P. pastoris* are not on-par with those of CHO, but in certain cases the higher product titers can compensate for this drawback [99].

Two other major challenges exist regarding recombinant protein production in *P. pastoris*. The unfolded protein response (UPR) pathway can be triggered by high levels of unfolded or misfolded recombinant protein inside the endoplasmic reticulum (ER). Activation of the UPR pathway initiates increased expression of chaperones to aid

in protein folding, but also proteases and proteins of the ER-associated degradation (ERAD) pathway are upregulated. Thereby, the cell aims to restore homeostasis and relieve secretion stress. This mechanism is not unique to *P. pastoris*, but can be found in all eukaryotic organisms [100]. Increased chaperone expression is beneficial for recombinant protein production, but its positive influence is outweighed by the negative effects associated with UPR activation. Especially the ERAD pathway compromises recombinant protein productivity. It facilitates translocation of misfolded proteins from the ER lumen to the cytosol, poly-ubiquitination and transport to proteasomes [101]. While the exact mechanism that triggers the UPR response has not yet been fully elucidated, the key transcription factor *HAC1* has been identified and characterized in *P. pastoris* [102]. The transcription factor is involved in the regulation of many hundred genes. Constitutive overexpression of *HAC1* leads to constant UPR activation. This strategy has proven useful for secretion of some recombinant proteins in *P. pastoris*, however it is not always beneficial [103]. Another transcription factor with suspected involvement in UPR regulation has recently been identified. Its overexpression has been shown to improve the production of a heterologous protein more than twofold [104].

Currently, a comprehensive solution for the UPR problem is not available. The optimal conditions have to be determined on a protein-to-protein basis. In theory, constant UPR activation prevents the cells from reaching intracellular protein levels that would result in an “overreaction” and subsequent degradation of more protein than is necessary for restoration of homeostasis. However, this strategy might also result in recombinant cells that are unable to reach the maximal productivity, if the native UPR would only be triggered at higher product titers. A thorough review of the UPR pathway in *P. pastoris*, its effects and different strategies to enable optimal secretion was published by Delic *et al.* (2013) [105].

Besides the UPR pathway, clonal variability is a significant factor complicating experiments with *P. pastoris*. The principle of clonal variability, its origin, effects and ways to control clonal variability in *P. pastoris* are presented in more detail in chapter 1.3.3 on page 28, as well as the introduction section of both section 2.1.2 and 2.2.2. Here, a brief summary is coupled with an overview of procedures used for screening large numbers of transformants.

For the creation of producer strains, the target gene is usually integrated into the genome of *P. pastoris* [85]. A specific locus is targeted via homologous sequences and homologous recombination is intended to facilitate correct integration of the expression cassette and, if desired, excision of the target locus. Yet, after transformation a diverse landscape of clones with different productivities and sometimes altered morphology or growth behavior is encountered [96, 106]. Consequently, time-intense screening procedures have to be performed in order to find the strain with desired characteristics. In many cases, unintended integration events are the cause for clonal variability. The non-homologous end joining (NHEJ) pathway mediates integration of DNA without requiring any sequence homologies and is involved in the double-strand break (DSB) repair mechanism [31]. Integration of the expression cassette via this pathway often results in off-target insertion and potential disruption of genes or regulatory elements. In *P. pastoris* this pathway is more pronounced than homologous recombination. Most other yeasts, and other eukaryotic systems, tend to exhibit even lower targeting frequencies than *P. pastoris* [107, 108]. However, *S. cerevisiae* is atypical in this regard, displaying extraordinarily high homologous recombination frequencies. Often, homologous sequences as short as 50 bp are sufficient for high targeting efficiencies [109, 110], enabling rapid PCR-based integration and alteration of targeting sequences in plasmids. In *P. pastoris* ca. 1000 bp long sequences are needed for efficient targeting, depending on the target locus [111, 112], making vector construction more cumbersome. Therefore, one goal of the *P. pastoris* community is to improve homologous recombination and reduce the prevalence of the NHEJ pathway, in order to raise the genetic tractability to that of *S. cerevisiae*.

To date, no study has been published that systematically investigated the different integration events encountered after transformation of *P. pastoris*. Multiple reports on the observed clonal variability exist, with productivities of clones from one experiment ranging from extremely high to below the detection limit [113, 114]. Näätsaari *et al.* (2012) [111] determined the off-target integration locus for 14 strains via combined nested PCR and genome walking. Hits of different coding and non-coding regions were determined. But the authors abstained from drawing further conclusions from these findings since they were concerned by the small sample size. In addition, the analyzed strains were generated using intentionally short homologous sequences for integration. Thereby, the

recorded events might not be applicable for standard operations in *P. pastoris*. Also, strains were selected based on retaining the wild type phenotype, with curing of an auxotrophy being the goal of the transformation. No recombinant protein production was assayed. It would be of interest, from a biotechnological standpoint, to correlate integration event and productivity of a clone and thereby derive which integration events are advantageous and which disadvantageous.

In order to facilitate the efficient selection of desired clones from a large number of transformants with unknown characteristics, many screening procedures have been developed. For high throughput cultivation of multiple clones, a method for *P. pastoris* cultivation in 96 deep-well plates was reported [115]. In contrast to previously published methods [116], the cultivation parameters were optimized to reduce apoptosis rates and the transferability to bioreactor conditions was demonstrated. The approach has been proven to be suitable for characterizing a library of promoters for their expression strength [117], as well as screening clones for protein or metabolite productivity [118, 119]. Specialized variants have been developed for screening the expression of membrane proteins [120],monoclonal antibodies [121] or enzymes with easy to detect activities [122, 123]. Although handling time was significantly reduced compared to e.g. shake flask cultivations, a drawback of this approach is the comparatively high demand for manual labor. Furthermore, the use of inexpensive 96 deep-well plates does not allow for the online control of cultivation parameters or application of fed-batch strategies. To this end, Hemmerich *et al.* (2014) [124] presented a fully automated microfermentation platform. It is capable of carrying out different fed-batch strategies and online monitoring of important cultivation parameters (e.g. optical density, fluorescence, pO_2) in a 48 well format. Besides directly screening for clones with desirable product titer or activity, their fitness is of interest, if industrial application is desired. Activation of the UPR pathway is a good indicator of high levels of stress in the cell. Assaying the transcript levels of associated genes is possible, but work-intense. Recently, the detection of metabolites correlating with induction of the UPR pathway was proposed to streamline screening of clones for their fitness level [125].

It was the goal of this thesis to better understand the underlying causes of clonal variability on the genetic level using standard *P. pastoris* genetic manipulation

techniques and to utilize these insights to better control the variability. Furthermore, the interplay between integration event and productivity was of interest, as to make the results applicable in the field of recombinant protein production with *P. pastoris*. A large library size of transformed clones should ensure that the frequency of discovered events can be estimated and rare events are not easily overlooked. Traditional characterization data was to be combined with whole-genome sequencing data for added insights. This approach was designed to enable correlation of integration event on the genome with the recombinant protein productivity of the clone. Due to the large sample size, results were expected to give information on different events. In turn, this information could enable future experiments to be optimized, reduce clonal variability and obtain more of the desired strains. On a bigger scale, reducing clonal variability, e.g. by preventing unintended integration events from occurring, could further the *P. pastoris* platform as a whole and open up the path towards genetic engineering projects of higher complexity.

1.3.2 Genetic engineering

Besides recombinant protein production, the heterologous production of metabolites and the creation of chassis strains optimized for specialized applications are ongoing projects in *P. pastoris* biotechnology. Central to these topics is the application of genetic engineering tools to alter the cell and make it more suitable for the targeted application. Genes for heterologous pathways need to be introduced, native genes deleted or down-regulated. These endeavors also heavily benefit from a deeper understanding of clonal variability and ways of controlling it. The less clones that need to be screened to find the right one, the more time that is available for application experiments. The lower the burden of clonal variability, the higher the inclination of scientists to conduct their studies in *P. pastoris* and to perform genetic engineering projects of higher complexity.

Future aspects of genetic engineering in *P. pastoris* in the context of systems-biology, its history as well as recent developments are discussed in the following section. This information is presented in the form of the manuscript “Towards systems-based metabolic engineering in *Pichia pastoris*”. The manuscript was submitted to the journal *Biotechnology Advances* in May, 2017. In contrast to the recent review by Kang *et al.* (2017) [126], a broader scope of topics is presented. Aspects of systems-biology, physiology, metabolite production and clonal variability are discussed and

connected to the overarching aim of establishing systems-based metabolic engineering in *P. pastoris*. On the other hand, Kang *et al*. solely focuses on tools for strain engineering developed in the last five years. Another recent review was published, from which the presented manuscript had to distinguish itself. Wagner & Alper (2016) [127] reviewed the latest developments in genetic engineering of the non-conventional yeasts *O. angusta*, *K. lactis*, *P. pastoris* and *Y. lipolytica*. Detailed information on new or established regulatory elements and techniques for targeted recombination in all four yeasts is presented. However, owing to the inclusion of four different yeast species and the publication date (14.12.2015), some key publications for *P. pastoris* were not discussed. For instance, the establishment of CRISPR/Cas9 methods for *P. pastoris* and the use of cell cycle synchronization for improved targeting efficiency are not mentioned [128, 129]. Like Kang *et al.*, the review focuses on genetic tools and does not include the aforementioned additional topics, contained in the following manuscript.

1.3.3 Manuscript "Towards systems-based metabolic engineering in *Pichia pastoris*"

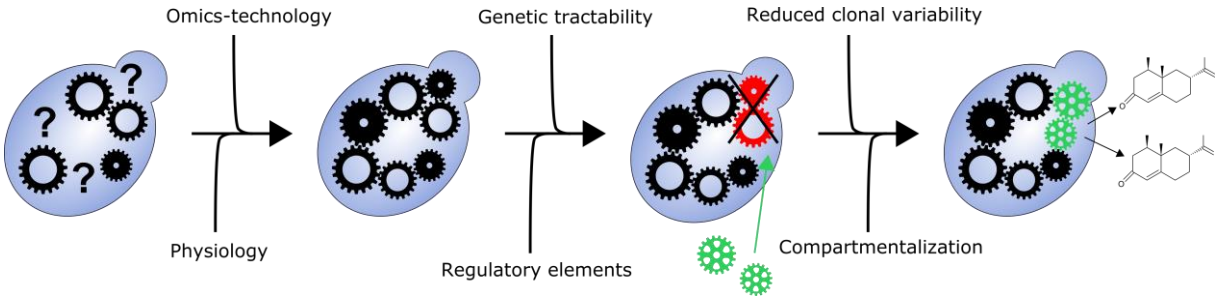

Figure 4: Graphical abstract for the manuscript "Towards systems-based metabolic engineering in *Pichia pastoris*". The illustration shows how new insights from physiology and omics-technology reveal the inner workings of *P. pastoris*. Subsequently, new regulatory elements and improved genetic tractability enable the targeted engineering of (heterologous) pathways. A reduced clonal variability eases strain selection and promotes projects of higher complexity, while the compartmentalization of pathways to organelles opens new possibilities for value-added metabolite production.

+ 39

Towards systems-based metabolic engineering in *Pichia pastoris*

Jan-Philipp Schwarzhans[1,2], Tobias Luttermann[1,2], Martina Geier[3], Jörn Kalinowski[2], Karl Friehs[1,*]

[1] Fermentation Engineering, Bielefeld University, Universitätsstr. 25, Bielefeld, 33615, Germany

[2] Microbial Genomics and Biotechnology, Center for Biotechnology (CeBiTec), Bielefeld University, Universitätsstr. 27, Bielefeld, 33615, Germany

[3] Austrian Center of Industrial Biotechnology (acib), Petersgasse 14, Graz, 8010, Austria

[*]Corresponding author:

Karl Friehs

Universitätsstrasse 25

33615, Bielefeld, Germany

Tel. : +49 (0)5 21 - 106-5301

Fax : +49 (0)5 21 - 106-6475

Email: karl.friehs@uni-bielefeld.de

Abstract

The methylotrophic yeast *Pichia pastoris* is firmly established as a host for the production of recombinant proteins, frequently outperforming other heterologous hosts. Already, a sizeable amount of systems biology knowledge has been acquired for this non-conventional yeast. By applying various omics-technologies, productivity features have been thoroughly analyzed and optimized via genetic engineering. However, challenging clonal variability, limited vector repertoire and insufficient genome annotation have hampered further developments. Yet, in the last few years a reinvigorated effort to establish *P. pastoris* as a host for both protein and metabolite production is visible. A variety of compounds from terpenoids to polyketides have been synthesized, often exceeding the productivity of other microbial systems. The clonal variability was systematically investigated and strategies formulated to circumvent untargeted events, thereby streamlining the screening procedure. Promoters with novel regulatory properties were discovered or engineered from existing ones. The genetic tractability was increased via the transfer of popular manipulation and assembly techniques, as well as the creation of new ones. A second generation of sequencing projects culminated in the creation of the second best functionally annotated yeast genome. In combination with landmark physiological insights and increased output of omics-data, a good basis for the creation of refined genome-scale metabolic models was created. The first application of model-based metabolic engineering in *P. pastoris* showcased the potential of this approach. Recent efforts to establish yeast peroxisomes for compartmentalized metabolite synthesis appear to fit ideally with the well-studied high capacity peroxisomal machinery of *P. pastoris*. Here, these recent developments are collected and reviewed with the aim of supporting the establishment of systems-based metabolic engineering in *P. pastoris*.

Keywords: *Pichia pastoris*; *Komagataella phaffii*; Non-conventional yeasts; Genetic Engineering; Metabolic Engineering; Recombinant protein production; Promoters; Systems-biology; Physiology

Abbreviations

ADH = Alcohol dehydrogenase

AOX1/2 = Alcohol oxidase 1/2

ARS = Autonomously replicating sequence

ATF = Artificial transcription factor

CDW = Cell dry weight

CHiP = Chromatin immunoprecipitation

COBRA = Constraint-based reconstruction and analysis

CPR = Cytochrome P450 reductase

CYP = Cytochrome P450 monooxygenase

DAS1/2 = Dihydroxyacetone synthase isoform 1/2

DDS = Dammarenediol-II synthase

DMAPP = Dimethylallyl pyrophosphate

DSB = Double-strand breaks

GAP = Glyceraldehyde-3-phosphate

GCN = Gene copy number

GEM = Genome-scale metabolic model

GlcNAc = N-acetylglucosamine

gRNA = guide RNA

HR = Homologous recombination

IPP = Isopentenyl pyrophosphate

IR = Inverted repeat

MIG1/2 = Multicopy inhibitor of GAL gene expression 1/2

MIT1 = Methanol-induced transcription factor 1

MUT = Methanol utilization

MVA = Mevalonate

MXR1 = Methanol expression regulator 1

NGS = Next generation sequencing

NHEJ = Non-homologous end joining

NRG1 = Negative regulator of glucose-repressed genes 1

OCH1 = α-1,6-mannosyltransferase

ORF = Open reading frame

PRM1 = Positive regulator of methanol 1

PTS = Peroxisomal targeting sequence

RNA-Seq = RNA-sequencing

ROS = reactive oxygen species

S1,7BP = Sedoheptulose-1,7-bisphosphate

UPR = Unfolded protein response

VPS = Vacuolar protein sorting

WSC = Cell wall integrity and stress response component

XYL5P = Xylulose-5-phosphate

ZF = Zinc finger

1. Introduction

In principle, systems-based metabolic engineering aims to bring the predictability and model based approach of the engineering world to biological systems. For the application of systems-based metabolic engineering an organism thoroughly studied via omics-technologies, the availability of computational systems biology tools and the capability for targeted genetic engineering, including synthetic biology, is required (Keasling, 2010; Lee *et al.*, 2012; Lee and Kim, 2015). So far, the biotechnological bacterial workhorse *Escherichia coli*, the gram-positive model bacterium *Bacillus subtilis* and the model yeast *Saccharomyces cerevisiae* are the only microorganisms for which systems-based metabolic engineering is firmly established (Kelwick *et al.,* 2014). Many other organisms show promising features, but not all requirements are fulfilled yet or systems-based metabolic engineering is currently being established. The very recent progress in *Pichia pastoris* research points towards a growing interest and effort to enable the system-based approach in this yeast. By giving a brief overview of *P. pastoris* history, summarizing the newest findings in detail and describing potential future applications, this review intents to aid this line of research.

The methylotrophic, non-conventional budding yeast *P. pastoris* has been established as a wide-spread recombinant protein expression platform in both academia and the industry. According to the web platform www.pichia.com, over 5000 different proteins have been produced in this yeast. The popularity stems from the availability of simple and robust high-cell density cultivation procedures, tightly regulated and extraordinarily strong promoters, good post-translational modification and secretion capabilities, as well as ease of genetic manipulation (Ahmad *et al.*, 2014; Macauley-Patrick *et al.*, 2005). While early ventures focused on technical enzymes (Cereghino and Cregg, 2000), the acquisition of the FDA GRAS (generally regarded as safe) status (Ciofalo *et al.*, 2006) promoted the development of biopharmaceuticals, e.g. the kallikrein inhibitor Kalbitor® or the aglycosylated protease Jetrea® (Corchero *et al.*, 2013; Meehl and Stadheim, 2014). *P. pastoris* has demonstrated its suitability for the expression of targets that proved problematic in other host systems, e.g. membrane bound proteins (Byrne, 2015; Vogl *et al.*, 2014) or glycoproteins (Laukens *et al.*, 2015).

Therefore, it has been recommended to consider *P. pastoris* as a standard tool for labs interested in the production of recombinant proteins (Bill, 2014).

Typically, the expression cassette harboring the target gene is integrated into a chromosomal locus via homologous recombination (HR), ensuring a high genetic stability (Cereghino and Cregg, 2000). Alternatively, episomal vectors using the native autonomously replicating sequences (ARS) *PARS1* and *PARS2* are available, but are merely used in a few applications (Cregg *et al.*, 1985; Lee *et al.*, 2005). While most protein expression studies only require the knock-in of a single target gene, the secretion of fully humanized and terminally sialylated glycoproteins was the largest genetic engineering project in *P. pastoris* to date (Hamilton *et al.*, 2006). This project required deletion of four genes and the integration of 14 foreign genes, including the transfer of the complete human CMP-N-actelyneuraminic acid biosynthesis pathway. Other genetic engineering ventures include the introduction of biotin-prototrophy (Gasser *et al.*, 2010), modification of the methanol utilization (MUT) pathway (Krainer *et al.*, 2012) and improvement of protein folding and secretion features (Guerfal *et al.*, 2010). Compared to the wealth of publications focused on protein production and its optimization, relatively few studies concern the biosynthesis of metabolites. The use of non-conventional yeasts in metabolic engineering projects in order to further the yeast platform as a whole was postulated. Only one example from *P. pastoris* was cited (Liu *et al.*, 2013). Nevertheless, chemically different metabolites from riboflavin (Marx *et al.*, 2008) and poly-3-hydroxybutyrat (Poirier *et al.*, 2002) to different carotenoids (Araya-Garay *et al.*, 2012a, 2012b) have been successfully produced in the past.

Although *P. pastoris* shares many properties with the conventional yeast *S. cerevisiae*, it also has its own distinguishing features that provide opportunities or present challenges. Genome integrations are stable but often a high clonal variability of clones from one transformation is encountered, displaying various productivity characteristics or changes in their physiology (Cregg *et al.*, 1985; Schwarzhans *et al.*, 2016a, 2016b). A time-consuming screening process has to be employed in order to find the clone with the optimal features for the desired application (Looser *et al.*, 2015). One of the causes for the clonal variability is the non-homologous end joining (NHEJ) pathway, which mediates integration of foreign DNA at untargeted locations. The clonal variability is

named as a key factor holding back the further development of *P. pastoris* as a platform for producing value-added chemicals (Kelwick *et al.*, 2014). On the other hand, *P. pastoris* inherently good protein production features, along with a Crabtree-negative phenotype and a well-developed and well-studied peroxisomal machinery, should make it a desirable host for metabolic engineering projects, surpassing *S. cerevisiae* in certain applications.

The methanol inducible alcohol oxidase 1 (*AOX1*) promoter *pAOX1* and the constitutive glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) promoter *pGAP* are the most popular choices for facilitating foreign gene expression (Vogl and Glieder, 2013). *pAOX1* offers tight regulation with near-zero transcriptional activity under repressed or derepressed conditions, and exceptionally high activity when induced with methanol (Cereghino and Cregg, 2000). However, the toxicity and flammability of methanol can be of concern as well as the increased oxygen consumption and heat generation of induced high-cell density cultures, necessitating adapted cultivation procedures (Looser *et al.*, 2015). *pGAP* enables constitutive expression at very similar levels to *pAOX1*, reducing process time and handling complexity (Waterham *et al.*, 1997a). However, *pGAP* is not suitable for the expression of host-toxic products, since production and growth phase cannot be decoupled. Variants of both *pAOX1* and *pGAP* with adjusted transcriptional activity have been developed to enable fine-tuned expression experiments (Hartner *et al.*, 2008; Qin *et al.*, 2011). In addition, many new promoters with different regulatory properties and expression strengths have been discovered and applied. They include alternative, methanol inducible promoters of the MUT pathway (Shen *et al.*, 1998; Tschopp *et al.*, 1987; Vogl *et al.*, 2016) as well as other constitutive or repressible promoters (Moreira de Almeida *et al.*, 2005; Stadlmayr *et al.*, 2010).

P. pastoris exhibits a high genetic accessibility. Since its first discovery, scientists have applied various approaches to shape *P. pastoris* towards their needs. While random mutagenesis and subsequent screening procedures were used in the beginning (Liu *et al.*, 1992), selectable markers (Lin Cereghino *et al.*, 2001) and more sophisticated genetic engineering tools like the Cre-Lox recombinase system (Pan *et al.*, 2011) were established over the years. Applying these techniques, genetic and metabolic engineering projects of different scopes have been realized. However, the aforementioned challenges

regarding clonal variability and NHEJ off-target integration events complicated genetic engineering projects of higher complexity. Therefore, improvements to the genetic tractability are an ongoing project. In tandem, the repertoire of integrative and episomal vectors requires expansion to facilitate further development of synthetic biology methods in *P. pastoris* (Kelwick *et al.*, 2014).

The commonly used *P. pastoris* strains were genome sequenced between 2009 and 2011 (De Schutter *et al.*, 2009; Küberl *et al.*, 2011; Mattanovich *et al.*, 2009). In the following years, multiple whole genome transcriptomics (Dragosits *et al.*, 2010; Hesketh *et al.*, 2013; Liang *et al.*, 2012), proteomics (Baumann *et al.*, 2010; Dragosits *et al.*, 2009) and metabolomics (Carnicer *et al.*, 2012; Heyland *et al.*, 2011) studies were carried out. In combination with the detailed characterization of the peroxisome (Wriessnegger *et al.*, 2007) and cellular physiology under protein production conditions (Puxbaum *et al.*, 2015), a good basis for *P. pastoris* systems biology was created. Building on these insights, the first generation of genome-scale metabolic models (GEM) was created (Caspeta *et al.*, 2012; Chung *et al.*, 2010; Sohn *et al.*, 2010). However, the incomplete nature of the available genome data and its annotation was cited as a factor holding back systems biology of *P. pastoris* (Dikicioglu *et al.*, 2014).

In this review, we highlight recent advancements to establish *P. pastoris* not only as a recombinant protein production host but as an entire microbial cell factory. Its capability to synthesize various value-added metabolites and to be genetically engineered in a model-based approach have been successfully demonstrated in recent years. The ease of expressing heterologous proteins of bacterial to human origin in *P. pastoris* ought to make it ideal for transferring metabolic pathways from organisms that produce interesting metabolites, but have other disadvantages like difficulties to be cultivated or genetic inaccessibility. Shortcomings like the clonal variability and insufficient genome annotation have been better understood and improved upon. Multiple studies were lately published regarding metabolic engineering of *P. pastoris* towards producing terpenes, polyketides and aromatic compounds, amongst others. Simultaneously, the genetic tractability of *P. pastoris* has been considerably improved and principles of synthetic biology have been applied successfully. Popular methods like CRISPR/Cas9 have been transferred to *Pichia*, as well as novel systems towards

optimized pathway engineering created. New promoters with unique expression profiles were discovered via transcriptome sequencing (RNA-Seq) and existing regulatory elements have been comprehensively characterized. Furthermore, genetic engineering led to the modification of classic promoters like *pAOX1*, creating methanol-independent variants. Advanced curation of the Pichia genome annotation resulted in the next best functionally annotated yeast genome behind that of *S. cerevisiae*. Omics-based studies elucidated e.g. the exact organization of key metabolic pathways and allowed creation of improved GEM. Combined with physiological findings like the discovery of methanol-sensing receptors, a deeper understanding of *P. pastoris* systems biology was obtained.

2 Properties of *Pichia pastoris*

Although it has been revealed that biotechnologically used *P. pastoris* strains need to be reclassified as *Komagataella phaffii* or *K. pastoris* (Kurtzman, 2009), the original name remains the common designator. From the perspective of metabolic engineering, *P. pastoris* exhibits the same advantages as *S. cerevisiae* when compared to bacterial systems. More complex enzymes can be functionally expressed, viral or bacteriophage infections are not problematic, secretion is more efficient, intra cellular compartmentalization is possible and genome integrations are stable. In addition, *P. pastoris* exhibits certain advantages over *S. cerevisiae*. *P. pastoris* suitability for the expression of various heterologous proteins from bacteria, fungi, algae, plants, animals and humans is well-documented and beyond doubt. Especially recombinant membrane protein expression shows markedly higher success rates in *P. pastoris* than in *S. cerevisiae* (Bill, 2014; Öberg *et al.*, 2011). The hyper-mannosylation is less pronounced, making glycoproteins expressed in *P. pastoris* more likely to be active and less likely to induce hyper-antigenic reactions (Hamilton and Gerngross, 2007). In contrast to *S. cerevisiae*, *P. pastoris* is Crabtree negative, making substrate utilization more efficient under aerobic conditions. The peroxisome of *P. pastoris* serves as model for studying varying effects and processes (Waterham *et al.*, 1997b; Wriessnegger *et al.*, 2007). In combination with the recently suggested compartmentalization of metabolic pathways to yeast peroxisomes (DeLoache *et al.*, 2016; Shabbir Hussain *et al.*, 2016), *P. pastoris* distinguishes itself as a very promising candidate for advanced metabolic engineering projects.

Nevertheless, the system also has its drawbacks and challenges. The *S. cerevisiae* research community is one of the largest in the world, enabling considerable scientific output. Due to its smaller size, the *P. pastoris* community exhibits a lower publication rate, although its dedication to better understanding and improving the system by applying the newest technological advances has to be emphasized. Furthermore, the repertoire of genetic tools is not as comprehensively developed for *P. pastoris*. For instance, genome editing and engineering via the CRISPR/Cas9 system was first pioneered in *S. cerevisiae* in 2013 (DiCarlo *et al.*, 2013) and expanded upon in the following years (Generoso *et al.*, 2016; Jakočiūnas *et al.*, 2015; Mans *et al.*, 2015). In the case of *P. pastoris* the method was first established in 2016 (Weninger *et al.*, 2016) and not yet further refined. But the major challenge holding back more wide-spread application of this yeast is the high clonal variability associated with it (Kelwick *et al.*, 2014).

2.1 Clonal variability

As shown in Fig. 1, clones created during one transformation event can display widely different productivity characteristics or changes in their growth behavior and physiology. This phenomenon has been encountered in early studies with *P. pastoris*, e.g. strains producing the tetanus toxin fragment C exhibited an up to 30 fold difference in product concentration (Clare *et al.*, 1991) and is still prevalent (Cámara *et al.*, 2016; Krainer *et al.*, 2016; Schwarzhans *et al.*, 2016a). However, clonal variability can be beneficial for obtaining clones with high recombinant protein production levels. High producers are usually multicopy strains, an integration event which likely results from *in vivo* multimerization and subsequent integration of several expression cassettes at once (Aw and Polizzi, 2013; Clare *et al.*, 1991). On the other hand, clonal variability poses a challenge during metabolic and genetic engineering of strains with clearly defined features. Gasser *et al.* (2010) observed significantly different growth rates of biotin-prototrophy engineered clones from one transformation. It was suspected that variations in the gene copy number (GCN) were responsible. However, in other cases no clear correlation between gene dosage and productivity can be found (Wang *et al.*, 2015). Two strains exhibiting a ca. two fold difference in product titer did not differ in growth rate or mRNA level of the target gene. In consequence, extensive screening procedures

are needed to isolate the transformant with the desired features. Many established screening procedures are available (Looser *et al.*, 2015), with new metabolite-based and automated procedures having been reported lately (Hemmerich *et al.*, 2014; Tredwell *et al.*, 2017). A main culprit for the clonal variability is assumed to be the untargeted integration of expression cassettes at random sites of the genome by the NHEJ pathway (Näätsaari *et al.*, 2012). It mediates the repair of double-strand breaks (DSB) via integration of DNA without requiring homologous sequences, thereby circumventing the targeted HR pathway. This can result in disrupting an untargeted gene and compromising the genetic integrity of the clone. In contrast to *S. cerevisiae*, NHEJ is dominant over HR in *P. pastoris* (Daley *et al.*, 2005). The resulting lower targeting efficiency can be improved, to an extent, by using comparatively long homology sequences of ca. 1000 bp or more (Näätsaari *et al.*, 2012; Nett *et al.*, 2005). It has to be noted, that *S. cerevisiae* has an exceptionally strong HR mechanism. Most other biotechnologically applied microorganisms either exhibit a more predominant NHEJ pathway than *P. pastoris* (Guirouilh-Barbat *et al.*, 2004; Meyer *et al.*, 2007; Verbeke *et al.*, 2013) or have such a weak HR pathway that it is essentially not useful for directing foreign gene integration (Belhaj *et al.*, 2015; Rasala and Mayfield, 2015).

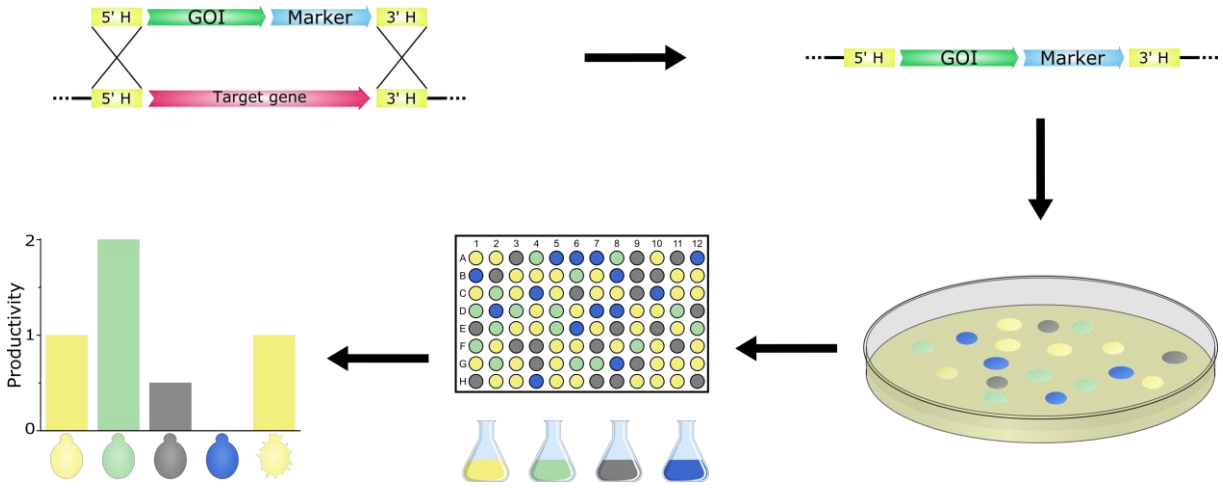

Figure 1: Common way of integrating the gene of interest in *P. pastoris* and the associated clonal variability. A wide range of productivity as well as physiological changes are encountered in clones from one transformation, making a time consuming screening procedures necessary. 5' H/ 3'H = 5' and 3' homologous sequence; GOI = Gene of interest

2.2 Ways of controlling clonal variability

If the gene dosage is the key challenge (Fig. 2 (A)), the GCN can be determined using proven qPCR or digital droplet PCR methods (Abad *et al.*, 2010a; Cámara *et al.*, 2016). The majority of clones after transformation will contain only a single copy of the gene. Typically, multicopy clones are encountered with a ca. 5-10 % frequency and "jackpot" strains (GCN > 10) with about 1 % frequency (Aw and Polizzi, 2013; Schwarzhans *et al.*, 2016a). Multiple approaches are possible in order to increase the GCN (reviewed in Aw and Polizzi (2013)). A recent publication presents a post-transformational vector amplification method, aiming to ease the process and make it more cost-efficient by using liquid media procedures (Aw and Polizzi, 2016). Although, fluctuations of productivity between strains can often be traced back to a different gene dosage, there are more possible causes. In addition, a high GCN values can also negatively impact productivity, e.g. by triggering the UPR pathway due to elevated stress levels in the cell.

P. pastoris multi copy strains typically contain all copies of the expression cassette adjacent to one another in a tandem array (Clare *et al.*, 1991). Fig. 2 (B) shows how this organization can lead to genetic instability and loss of cassettes due to so called "loop out" events via inter-cassette homologous sequences (Zhu *et al.*, 2009). Using different loci for insertion of multiple copies can mediate the problem but is often not applicable or to work-intense. Instead, optimizing cultivation conditions is advisable in order to find the most stable copy number. Loss of cassettes is often associated with high copy numbers and stress conditions, meaning cells with lower copy numbers have growth benefits. With this in mind, a system has been developed in which only multiple copies of an adenine auxotrophy selection marker with truncated promoter enable survival of the transformant, ensuring constantly high copy numbers (Du *et al.*, 2012). Due to the accumulation of an adenine intermediate in cells with insufficient complementation their colonies appear pink, facilitating direct selection of high copy clones via their white color. Accordingly, the commercially available system has been termed "PichiaPink™".

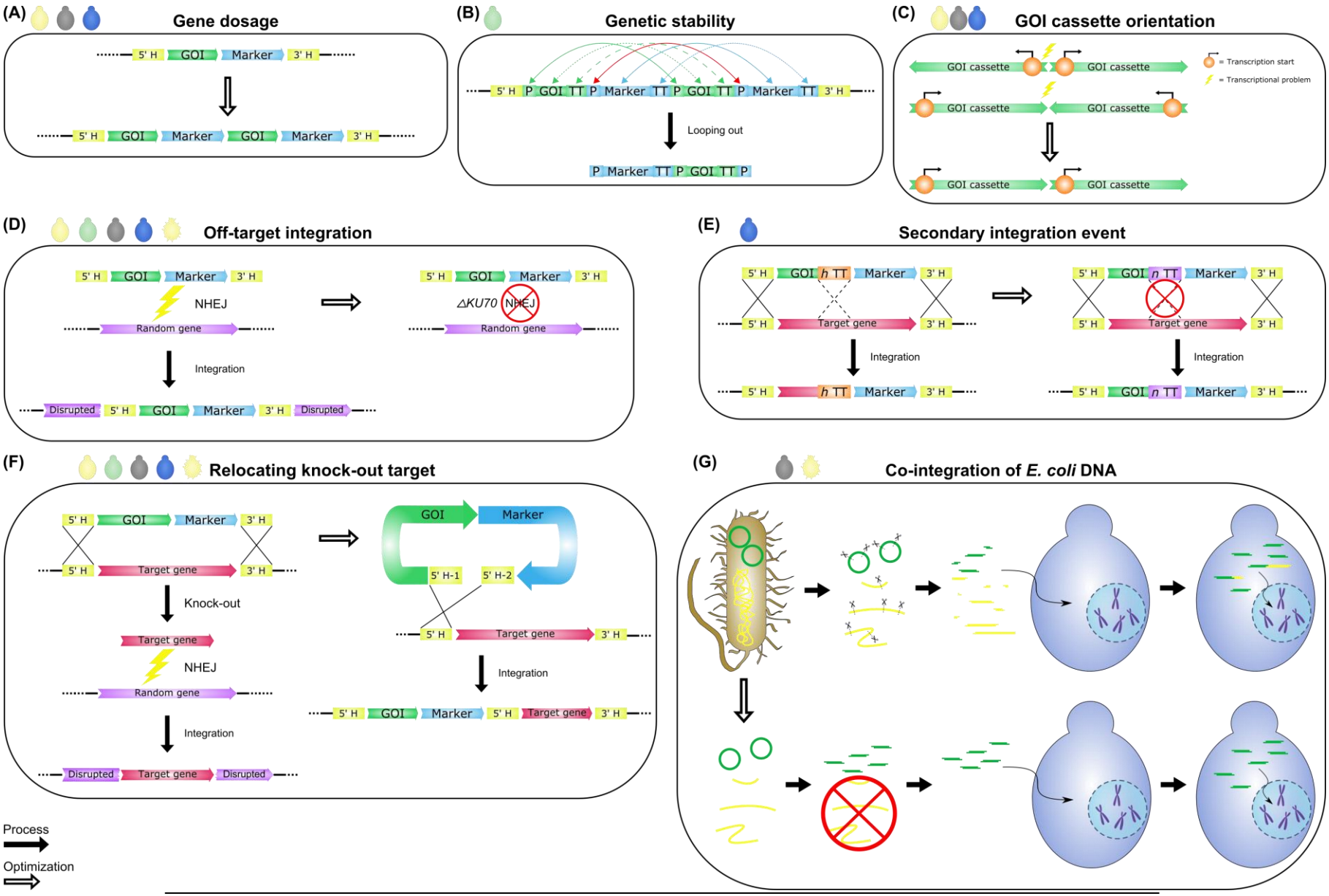
(A)
Gene dosage
5' H
GOI
Marker
3' H
(B)
Genetic stability
P
TT
Looping out
(C)
GOI cassette
= Transcription start
= Transcriptional problem
(D)
Off-target integration
NHEJ
Random gene
Integration
Disrupted
ΔKU70
(E)
Secondary integration event
h TT
n TT
Target gene
(F)
Relocating knock-out target
Knock-out
5' H-1
5' H-2
(G)
Co-integration of E. coli DNA
Process
Optimization

Figure 2 (previous page): Possible solutions on the genetic level to solve clonal variability issues. **(A)** Low gene dosage. Multiple strategies to increase the copy number of the target gene, and thus expression level, are available. **(B)** Loss of expression cassettes due to homologous sequences between adjacent cassettes. Fermentation procedures have to be optimized in order to minimize loss of cassettes and find a stable copy number. **(C)** Head-to-head and tail-to-tail cassette organization can be detrimental to productivity due to problems on transcriptional level. It is recommended to target head-to-tail assembly for optimal productivity. **(D)** Off-target integrations due to NHEJ can lead to unforeseen consequences. Multiple solutions are available, e.g. using a *KU70* deficient strain that is incapable of NHEJ. **(E)** Integration of only parts of the expression cassette due to internal homologous sequences, e.g. the terminator (h TT). This can be solved by replacing them with non-homologous sequence, e.g. a different non-homologous terminator (n TT). **(F)** Knock-out target relocates to different locus via NHEJ and causes same problems as described in (D). By using only an ends-in approach no gene is knocked-out. This approach is not suitable for knock-out studies. Here, screening of transformants must be performed accordingly. **(G)** DNA from *E. coli* plasmid propagation strain can co-integrate and be actively transcribed in *P. pastoris*. For example, by amplifying the expression cassette via PCR the issue can be circumvented.

Due to the location of expression cassettes in a tandem array, their orientation to one another can also have an impact on productivity (Fig. 2 (C)). Early it was found, that the head-to-tail organization form is predominant over the head-to-head and tail-to-tail variants (Clare *et al.*, 1991). Recently, a clear correlation between orientation and productivity was discovered via genome sequencing of clones from a library of over 800 strains (Schwarzhans *et al.*, 2016a). While high producer strains contained almost exclusively head-to-tail arrays, mostly head-to-head and tail-to-tail was found in low producing clones. It is suspected, that transcriptional problems cause expression levels markedly lower than the GCN would suggest. In a tail-to-tail organization RNA polymerases from neighboring cassettes run on a converging path, potentially leading to transcriptional arrest due to head-on collision events (Crampton *et al.*, 2006). On the other hand, in the head-to-head organization promoters for the target gene are directly next to one another on adjacent cassettes. This could lead to two RNA polymerases hindering each other from properly starting transcription due to steric interferences. Directed assembly of multiple cassettes into a head-to-tail array prior to integration can circumvent these problems (Vassileva *et al.*, 2001). Interestingly, analyzed strains contained either only head-to-tail or a 50-50 mixture of tail-to-tail and head-to-head

arrays. This hints at the existence of two competing integration mechanisms that exclude one another.

As mentioned earlier, NHEJ events can lead to integration of the expression cassette at a random locus on the chromosome (Fig. 2 (D)). Depending on the site, genes themselves or 5' and 3' regulatory intergenic regions can be disrupted. A preliminary investigation into the distribution of off-target integration events could find no pattern (Näätsaari *et al.*, 2012), but for a conclusive answer a genome wide approach as used for *Kluyveromyces lactis* would be necessary (Kegel *et al.*, 2006). Furthermore, it has been reported for eukaryotic systems that the integration site can influence expression strength of the GOI due to epigenetic factors, chromatin structure and other factors (Day *et al.*, 2000). Using cassettes targeted for different loci, this theory was also investigated in *P. pastoris*. No effect of the integration site on productivity was found (Love *et al.*, 2012; Perez-Pinera *et al.*, 2016). It has to be noted, that loci were chosen deliberately and random NHEJ mediated integration could lead to the disruption of genes of important metabolic pathways that would not be considered in a rationally designed experiment. However, a change of physiological properties due to off-target integration was reported (Schwarzhans *et al.*, 2016b). A gene involved in the oxidative stress response was disrupted, likely causing the change in growth behavior. The strain exhibited a changed colony morphology when grown on agar plates, indicating growth defects. In order to prevent off-target integrations the *P. pastoris KU70* homologue was deleted, thereby eliminating the NHEJ pathway (Näätsaari *et al.*, 2012). Affected cells displayed markedly increased targeting efficiencies and enable the use of shorter homologous sequences. The generated strain provides a good chassis for further metabolic engineering projects and has been successfully applied to that end (Krainer *et al.*, 2013; Wriessnegger *et al.*, 2014). Nevertheless, the removed NHEJ pathway also lead the cells being more susceptible to DNA damage and ca. 20 % lower growth rates, limiting their applicability for industrial purposes. For knock-out studies, a proven alternative presents itself with the split-marker method (Heiss *et al.*, 2013). Here, two fragments of the selection marker are split onto the distal ends of two linear DNA fragments that are transformed into *P. pastoris*. Only simultaneous homologous recombination between both fragments and the target locus leads to integration of the complete selection marker. While transformation frequencies are lower, the approach

does not require a specific host strain and is therefore more easily applicable to both academic and industrial research. The method has been applied in various genetic engineering studies (Nocon *et al.*, 2014; Prielhofer *et al.*, 2013; Ruth *et al.*, 2014).

The presence of an additional homologous sequence with the expression cassette can lead to a secondary integration event (Fig. 2 (E)). In the analyzed strains, the event was caused by the *AOX1* terminator (*AOX1 TT*) (Schwarzhans *et al.*, 2016a). The chromosomal *AOX1* locus was targeted by two distal homologous sequences (*pAOX1* and 3' UTR *AOX1* region) on the expression cassette. However, the homologous *AOX1 TT* on the expression cassette mediated integration of the selection marker region between *AOX1 TT* and 3' UTR *AOX1*. The resulting clones pass the selection process but cannot produce any target protein, since they are missing the GOI. They are false-positive. Theoretically, an integration of the GOI region between *pAOX1* and *AOX1 TT* without the selection marker is also possible, but clones would not survive the selection process. It was determined that 8 % of all >800 analyzed clones were the result of this secondary integration event (Schwarzhans *et al.*, 2016a). The relatively high frequency of this event means that they pose a burden for the screening process and should be avoided. In order to validate the assumed mechanism, *AOX1 TT* was replaced with *CYC1 TT* from *S. cerevisiae*. With the non-homologous terminator the secondary integration event was eliminated. However, average productivity was lower likely because *CYC1 TT* was too weak for *pAOX1* (Curran *et al.*, 2013). Many commercial and non-commercial vectors routinely used for *P. pastoris* contain the setup of *pAOX1*, GOI, *AOX1 TT* and 3' UTR *AOX1* (Ahmad *et al.*, 2014; Invitrogen, 2010). Terminators not homologous to the targeted region should replace the *AOX1 TT* in such plasmids. To this end, Vogl *et al.* (2016) assayed various native *P. pastoris* terminators for their effectiveness and highlighted potential replacement candidates. Regulatory elements like *DAS1 TT*, displayed effectiveness on the same level as *AOX1 TT* and should therefore be suitable for high level production applications.

Another kind of gene disruption was recently discovered. If two homologous sequences are used for targeting the expression cassette (e.g. for knock-out studies), integration results in the excision of the targeted region (Fig. 2 (F)). Typically, the excised DNA fragment will be degraded afterwards. In one particular case however, re-integration of

the previously deleted *AOX1* locus was observed (Schwarzhans *et al.*, 2016b). Likely mediated by the NHEJ pathway, the knock-out target moved from chromosome 4 to chromosome 2, where it disrupted an untargeted gene during the re-integration event. Only because the disrupted gene belonged to a family of genes involved in signal transduction, membrane trafficking and cytoskeletal organization (Nguyen *et al.*, 2005) the reintegration was discovered. While the relocated knock-out locus was fully functional at its new site, the untargeted gene disruption led to abnormal growth of the affected strain. Therefore, it is hard to gauge how often similar re-integration events occurred that did not result in physiological changes. Again, a genome scale analysis of a large number of clones would be necessary (Kegel *et al.*, 2006). During knock-out studies this re-integration event has to be taken into account, since it cannot fully be avoided. Simply checking strains with PCR assays targeting the knock-out locus might miss its relocation to a new site. In consequence, the effect of gene deletions might be misinterpreted. PCR assays checking for the absence of the knock-out target should complement the established methods. If a knock-out is not essential to the experiment, using only one homologous sequence for integration in an ends-in vector eliminates the possibility of relocating knock-out targets.

Last but not least, the co-integration of *E. coli* DNA from the plasmid propagation strain was detected (Fig. 2 (G)) (Schwarzhans *et al.*, 2016b). Common *P. pastoris* techniques use shuttle-vectors for amplifying the plasmid in *E. coli*, extract and digest it, and transform the linear DNA into the yeast cell. During this process contaminating gDNA, and depending on the *E. coli* strain F plasmid DNA, can also be digested and transformed into *P. pastoris*. In vivo ligation capabilities of yeasts in general (Suzuki *et al.*, 1983) and *P. pastoris* in particular (Clare *et al.*, 1991) are well-documented. Based on these abilities, and the integration events discovered during genome sequencing, *P. pastoris* ligated expression cassettes with *E. coli* DNA fragments and integrated the hybrids into its chromosomes. Interestingly, in one strain the integrated *E. coli* DNA was modified after the transformation. Two short palindromic sequences are missing compared to the *E. coli* genome, probably caused by yeast inherent problems with palindromic sequences during DSB repair (Nag *et al.*, 1989). The integrations were discovered in clones with aberrant colony morphology, which was assumed to be due to the *E. coli* DNA fragments containing genes coding for membrane associated proteins. Using qRT-PCR methods,

the presence of mRNA transcripts of the *E. coli* genes could be detected in the *P. pastoris* mutants (Schwarzhans *et al.*, 2016b). However, no proteomic study was carried out, leaving it unclear whether transcripts were indeed translated into proteins. Nevertheless, the co-integration and potential activity of *E. coli* DNA in *P. pastoris* is highly undesirable, especially from an industrial standpoint. The commonly applied gel purification of the expression cassette prior to transformation could mediate some of the observed co-integrations. Yet, integrated *E. coli* DNA fragments ranged in size between 1.5 and 9.3 kb, and could therefore by-pass such a step. Applying PCR for amplification of the expression cassette or the use of rare and blunt-end cutters for digestion should effectively prevent the co-transformation and subsequent co-integration of DNA from the *E. coli* plasmid host.

2.3 Compartmentalization to the peroxisome

Typically, metabolic engineering of microorganisms introduces or modifies metabolic pathways that occur in the cytoplasm. A multitude of native pathways occur simultaneously in the cytoplasm, leading to potential negative crosstalk between targeted and untargeted reactions. This problem severely complicates metabolic engineering projects and can limit achievable product yields. Compartmentalization of engineered pathways to the organelles of eukaryotes presents itself as a good solution to this challenge (Zecchin *et al.*, 2015). Deletion of genes involved in side reaction pathways to reduce crosstalk is a commonly applied tool, but negative impacts on cell growth or the inability to delete essential genes limit this method. In *S. cerevisiae*, localization to the mitochondrion for isobutanol production (Avalos *et al.*, 2013) and vacuolar localization of methyl iodide synthesis have been reported (Bayer *et al.*, 2009). The peroxisome offers distinct advantages over other organelles. Translocation of proteins to the peroxisome can be efficiently realized using peroxisomal targeting sequences (PTS) (Purdue and Lazarow, 2001) and transport of metabolites across the organelle membrane is relatively well understood (Antonenkov and Hiltunen, 2012). Additionally, peroxisome biogenesis can be decoupled from cell growth on certain carbon sources, expanding its use for processes independent of the native cellular machinery (Purdue and Lazarow, 2001). With the ß-oxidation of fatty acids, a pathway capable of supplying

the building blocks for a variety of value added chemicals is present (Fig. 3) (Poirier *et al.*, 2006).

Recently, three independent studies proposed using the yeast peroxisome for metabolite synthesis (DeLoache *et al.*, 2016; Sheng *et al.*, 2016; Zhou *et al.*, 2016). The product yields for fatty acid derivatives like fatty alcohols and alkenes in *S. cerevisiae* were drastically improved via peroxisomal localization, improving titers more than fourfold (Zhou *et al.*, 2016). A similar approach was used by Sheng *et al.* (2016) to produce medium chain fatty alcohols, starting from the ß-oxidation of fatty acids in the peroxisomes. Both studies clearly show the potential of producing ß-oxidation derived metabolites in the yeast peroxisome. DeLoache *et al.* (2016) work provides a better understanding of peroxisomal membrane transport processes. They improved the common PTS1 tag to lessen the effect of the protein properties on its transport rate. The enhanced PTS1 was used to develop an assay for recombinant protein translocation rates to the peroxisome. Furthermore, the permeability of the peroxisomal membrane was characterized and allowed the construction of a pathway with a permeable substrate but impermeable intermediate. Byproduct formation was significantly reduced and product yields increased.

The peroxisome of *P. pastoris* has been well studied and used as model organelle (Johnson *et al.*, 1999; Liu *et al.*, 1992; Waterham *et al.*, 1997b), making it an ideal target for the proposed compartmentalization of pathways. Specialized isolation protocols have been developed for targeted purification (Wriessnegger *et al.*, 2007). Under inducing conditions, the peroxisome can constitute up to 80 % of the total cell volume, indicating a large potential for protein intake (Gleeson and Sudbery, 1988). Peroxisome biogenesis can be induced with methanol or oleate containing medium (Waterham and Cregg, 1997). Interestingly, it was found that while methanol induces peroxisome genesis, down-regulation of ß-oxidation pathway genes was observed (Prielhofer *et al.*, 2015; Rußmayer *et al.*, 2015a). On the other hand, glucose limitation led to increased ß-oxidation activity. Therefore, fermentation strategies combining glucose growth phase with oleate induced peroxisome biogenesis seem advisable, if high ß-oxidation activity is desired. To this end, the characterization of genes involved in oleate-induced peroxisome biogenesis could provide targets for engineering tailored peroxisomes with improved properties for metabolite synthesis (Yan *et al.*, 2008). This approach would

also enable a methanol-free cultivation, eliminating negative traits associated with methanol-based processes (Looser *et al.*, 2015). Native (Bhataya *et al.*, 2009; Waterham *et al.*, 1997a) and heterologous (Poirier *et al.*, 2002) PTS have been successfully applied.

Already, the peroxisome has been used for metabolite production in *P. pastoris*. Poirier *et al.* (2002) demonstrated the peroxisomal synthesis of polyhydroxyalkanoate. Polyhydroxyalkanoate is a 3-hydroxyacyl-CoA polymer, which is an intermediate of the ß-oxidation pathway (Fig. 3). Using a whole cell catalyst approach, the enzymatic conversion of cephalosporin C by a D-amino acid oxidase localized at the peroxisome was realized (Abad *et al.*, 2010b). The ß-oxidation pathway also feeds acetyl-CoA into the MVA pathway, thereby providing IPP for terpenoid synthesis. This circumstance was exploited by Bhataya *et al.* (2009) for construction of a lycopene biosynthesis pathway in the peroxisome (Fig. 3). Compared to attempts in which pathway enzymes were localized to the cytosol, fourfold higher lycopene yields were achieved (Araya-Garay *et al.*, 2012b). Along with these proven metabolites, Fig. 3 displays other targets that have been produced in *P. pastoris* and whose biosynthetic pathways could be located to the peroxisome, due to their connection to the ß-oxidation of fatty acids. Besides metabolite production, peroxisomal targeting has also been used to express peptides that proved to be toxic for cytosolic expression (Xiao *et al.*, 2016).

By combining recent advances towards establishing yeast peroxisomes as ideal compartments for engineered metabolic pathways and *P. pastoris* naturally well-developed peroxisomal apparatus, metabolic engineering projects in this non-conventional yeast might be considerably boosted. More advanced techniques could enable selective repression of peroxisome biogenesis to reduce undesired metabolic activities in the growth phase. For instance, repression of the peroxisome biogenesis gene *PAS5* via the repressible *pTHI11* promoter should stop peroxisome formation without affecting growth on glucose (Spong and Subramani, 1993).

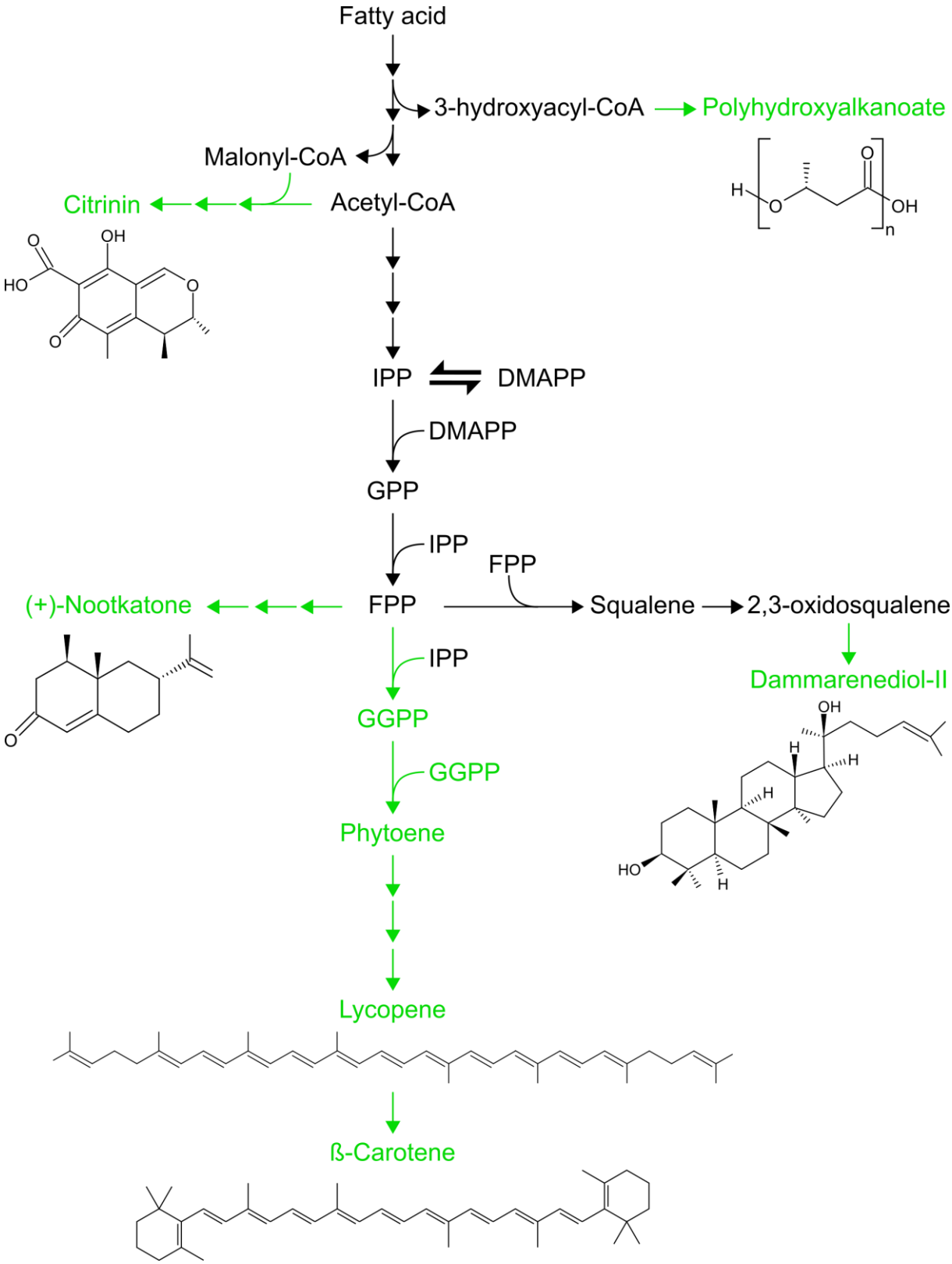

Figure 3: Examples of heterologous metabolites produced in *Pichia* and their biosynthetic pathway, starting from the ß-oxidation of fatty acids. Green arrows/names highlight heterologous pathways and their intermediates or end products. For ease of representation some pathways are shown condensed, indicated by multiple adjacent arrows. DMAPP = Dimethylallyl pyrophosphate; FPP = Farnesyl pyrophosphate; GPP = Geranyl pyrophosphate; IPP = Isopentenyl pyrophosphate

3 Regulatory elements

To efficiently produce fine chemicals or pharmaceuticals in heterologous expression systems, such as *P. pastoris*, strategies are required to engineer enzyme levels and avoid pathway bottlenecks or resource misdirection (Roquet and Lu, 2014). In eukaryotes, this is often achieved by fine-tuning of transcript levels via employment of different promoters. For metabolic engineering, the ideal promoter is tightly regulated by induction. This allows to decouple product formation from cell growth which is a suitable strategy particularly to produce toxic compounds that hamper cell growth. A comprehensive list of established *P. pastoris* promoters and their regulation can be found in the review by Vogl and Glieder (2013). Recently, novel derepressible or inducible promoters were discovered and researchers engineered methanol-free *pAOX1*-based expression systems (Table 1).

Table 1: New promotors and expression systems in *P. pastoris*. The 'modification' column describes genetic changes to the promoter itself or the strain. Approximate levels of protein or gene expression relative to *pAOX1* or *pGAP* are given, where available. Upwards arrows indicate that its corresponding gene is being overexpressed. Abbreviations: mi = methanol-independent; GFP = green fluorescent protein; HRP = horseradish peroxidase; CALB = lipase B from *Candida antarctica*; HAS = human serum albumin; LACB = ß-galactosidase; SCP = synthetic core promoter; URS= upstream regulatory sequence.

Name	Regulation	Modification	Expression level	Reference
ADH3	Ethanol induction	Native	Strong (similar to *pAOX1*)	Karaoglan *et al*. (2016)
AOX1	Methanol induction	Native	Strong (naturally ca. 30% of total protein)	Yurimoto *et al*. (2011)
$AOX1_{mi1}$	Glucose derepression; no Glycerol repression	Δ*nrg1*, Δ*mig1*, Δ*mig2*, *mit1*$^{\uparrow}$	77% of *pAOX1* (GFP)	Wang *et al*. (2017)
$AOX1_{mi2}$	Glucose derepression; no Glycerol repression	Δ*gut1*, Hp*gcy1*$^{\uparrow}$	25% of *pAOX1* (GFP)	Shen *et al*. (2016)
$AOX1_{mi3}$	Dihydroxyacetone induction	Δ*dak1*	50 - 60% of *pAOX1*	Shen *et al*. (2016)
$AOX1_{syn}$ (d6)	Glucose/Glycerol derepression	Synthetic	400 % of *pAOX1* (GFP)	Hartner *et al*. (2008), Looser *et al*. (2017)
CAT1	Glucose derepression, methanol/oleate induction	Native	100 – 180% of p*AOX1* (HRP and CALB)	Vogl *et al*. (2016)
DAS2	Methanol induction	Native	Strong (similar to *pAOX1*)	Vogl *et al*. (2016)
FDH1	Methanol induction	Native	Strong (similar to *pAOX1*)	Vogl *et al*. (2016)
FLD1	Methanol induction	Native	Strong (similar to *pAOX1*)	Shen *et al*. (1998), Vogl *et al*. (2016)
G1 (*GTH1*)	Glucose derepression	Native	230% of *pGAP* (HSA)	Prielhofer *et al*. (2013)
GAP	Constitutive	Native	Strong (similar to *pAOX1*)	Waterham *et al*. (1997)
LRA3	Rhamnose induction	Native	80% of *pGAP* (LACB)	Liu *et al*. (2016)
pCore11	Methanol induction	SCP with *AOX1* URS	10% of *pAOX1* (GFP)	Vogl *et al*. (2014)
pCoreAOX1	Methanol induction	SCP *pAOX1* hybrid variants with *AOX1* URS	35 – 117% of *pAOX1* (GFP)	Vogl *et al*. (2014)
Synthetic	Methanol induction/Constitutive	*De novo* SCP variants of *pAOX1*, *pCAT1*, *pDAS1* and *pGAP*	27 - 122 % of *pAOX1* (GFP)	(Portela *et al*., 2017)
Synthetic	ß-estradiol induction	Synthetic promoter and transcription factor	Stronger than *pAOX1*	Perez-Pinera *et al*. (2016)
THI11	Thiamine derepression	Native	Strong (similar to *pGAP*)	Stadlmayr *et al*. (2010), Landes *et al*. (2016)

Although several promising promoters with interesting features have been discovered in the recent years, there might be still a variety of novel, differently regulated genes hidden in *P. pastoris*. Typically, DNA microarray analysis has been carried out to identify genes that offer the desired regulatory features (Prielhofer *et al.*, 2013; Vogl *et al.*, 2016). However, since this technique can only give information about relative changes in expression levels, data derived from whole-transcriptome RNA sequencing (RNA-seq) should improve promoter identification. Love *et al.* (2016) investigated transcriptomic profiles of *P. pastoris* during cultivation on different carbon sources and provide a comprehensive dataset on differently regulated genes for each carbon source. Furthermore, RNA-seq in combination with ribosome profiling offers the possibility to find Kozak sequences that are characterized by high translation efficiencies. Ribosome profiling allows to analyze the translatome by only sequencing the mRNA that has been recovered from translating ribosomes (Ingolia *et al.*, 2009). With this strategy, it is possible to identify genes whose expression is translationally induced under specific conditions, for instance at distinct growth phases (Jeong *et al.*, 2016). Combination of results from both, RNA-seq and ribosome profiling might enable the construction of synthetic promoters in *P. pastoris* that exhibit high transcription as well as translation efficiencies under desired conditions. In a comparable approach, ribosome occupancy was used to fraction mRNA prior to microarray analysis and distinguish between highly- and lowest-translated transcripts (Prielhofer *et al.*, 2015).

3.1 *pAOX1* based systems

The classic methanol inducible *pAOX1* system remains the most popular one in *P. pastoris*, and it has recently been further developed for methanol-free expression. To this end, it is useful to understand how the MUT pathway and thus aox1 activity is regulated. When *P. pastoris* cells are grown on repressing carbon sources, such as glucose or glycerol, no aox1 activity can be observed (Couderc and Baratti, 1980). This tight regulation is achieved by a variety of activating and repressing transcription factors.

Several trans-acting elements have recently been identified that play a crucial role in the regulatory states of repression, derepression, and induction (Fig. 4 (A)). During induction, three transcriptional activators (prm1, mit1, and mxr1) independently bind

pAOX1 at different sites and activate the promotor through a cascade (Wang *et al.*, 2016). Of these, mxr1 (methanol expression regulator 1) was the first one to be described (Lin-Cereghino *et al.*, 2006). It codes for a protein with a zinc finger binding domain that has high similarity to adr1p from *S. cerevisiae*. When cells are grown on methanol or other gluconeogenic substrates, mxr1 is localized to the nucleus where it binds to sequences upstream of *AOX1*. Cells deficient of mxr1 are unable to utilize methanol, induce transcription of *AOX1* and to form normal-appearing peroxisomes (Lin-Cereghino *et al.*, 2006). In contrast, methanol-induced transcription factor 1 (mit1) does not participate in peroxisome proliferation in response to methanol, but regulates the expression of many genes involved in the MUT pathway (Wang *et al.*, 2016). mit1, together with prm1 (positive regulator of methanol 1), is localized to the nucleus independently of the carbon source. In contrast, mxr1 is only present in the nucleus in the absence of glucose, indicating that mxr1 is involved in derepression of *pAOX1* whereas mit1 and prm1 respond to methanol. The constitutively expressed prm1 binds upstream of *MIT1* and thus induces mit1 expression, if cells are exposed to methanol (Takagi *et al.*, 2012; Wang *et al.*, 2016). All three transcriptional activators are necessary for *pAOX1* activation. However, it remains still unclear how the methanol induction signal is transmitted to prm1 and how it activates prm1 to induce mit1 expression.

In the presence of glucose or glycerol, three trans-acting elements (nrg1, mig1 and mig2) have been recently described to repress *pAOX1* (Wang *et al.*, 2017, 2016). nrg1 (negative regulator of glucose-repressed genes 1) was shown to bind directly to five sites of *pAOX1*. These include two binding sites for the transcriptional activator mxr1. nrg1 deficient strains can express aox1 at low glucose concentrations, indicating that nrg1 competes with mxr1 for *pAOX1* binding sites (Wang *et al.*, 2016). Via live cell imaging of GFP tagged mig1 (multicopy inhibitor of *GAL* gene expression 1) and mig2, Wang *et al.* (2017) observed that both repressors are localized to the nucleus when cells were grown on glucose or glycerol. In cells grown on methanol they were predominantly transferred to the cytoplasm. Deletion of these genes led to cells that can express aox1 in the presence of 10 g/L glycerol and the absence of methanol. However, expression of the *AOX1* gene was still repressed if cells were grown on 10 g/L glucose as the sole carbon source. To further elucidate the role of these repressors individually, different deletion strains, including single and double knock-out strains of *MIG1* and *MIG2* were created. A

colorimetric aox1 enzyme assay and western blots revealed that only the double knock-out mutant showed significant aox1 expression, if cells were grown on 10 g/L glycerol. Δ*MIG1* and Δ*MIG2* double knock-out clones showed significantly higher *MIT1* expression levels than the Δ*MIG1* single knock-out strain. Expression levels of prm1 and mxr1 did not change in any of the investigated mutants. This indicates that both mig1 and nrg1 regulate mit1 expression by either binding its promoter or by interacting with the activating transcription factor itself in the presence of glycerol, and that mig2 might increase the repressing effect of mig1 on mit1. Accordingly, overexpression of mit1 from *pGAP* in wild-type cells grown on glycerol leads to minor aox1 activity demonstrating that mig1 and nrg1 rather bind to *pMIT1* than interact with the protein to prevent aox1 activity (Wang *et al*., 2017). However, no data on mRNA levels of mit1, prm1, and mxr1 if cultivated on 10 g/L glucose was presented. The complete interplay of all known activating and repressing trans-acting elements is not fully understood yet, but the characterization of the aforementioned transcription factors shed light on the regulation of *pAOX1* (Fig. 4 (A)).

Based on their characterization results of mig1, mig2 and nrg1, Wang *et al*. (2017) developed a methanol-independent *pAOX1* expression system via combinatorial engineering of trans-acting elements. Deletion of all three genes resulted in a strain that showed detectable aox1 activity (17% of wild-type cells grown in methanol), if grown on 10 g/L glycerol but was unable to express aox1 in the presence of glucose. Introduction of an additional copy of *MIT1*, constitutively expressed under the control of *pGAP*, let to a further increase of aox1 activity (up to 36 % of the wild-type). The corresponding strain will be referred to as MF1 (Fig. 4 (B)). To evaluate the potential of this expression system, GFP was expressed under the control of *pAOX1*. Relative GFP expression in MF1 reached 77 to 130 % of the wild-type if grown on 10 g/L glycerol or 0.5 % (w/w) methanol, respectively. Differences in aox1 expression and *pAOX1* driven GFP expression in MF1 can be explained by the lower content of peroxisomes in *P. pastoris* cells cultured in glycerol containing media, influencing aox1 content but not recombinant GFP expression. With this system, activation of *pAOX1* transcription can be achieved by derepression from glucose, eliminating the need for methanol induction. Glycerol can be used in the fed-batch phase as the sole carbon source. This allows for high cell density fermentation strategies with auto-induced expression of a recombinant protein via

glucose-glycerol shift. However, this engineered expression system also showed detectable GFP expression at low glucose concentrations, indicating that the tight repression of *pAOX1* by glucose was disturbed. This phenomenon might be ascribed to the deletion of *NRG1*, leading to insufficient prevention of *pAOX1* activation by mxr1. Full repression of *pAOX1* in MF1 is only achievable at glucose concentrations of 40 g/L, while low GFP expression could already be observed in the presence of 20 g/L, making this system not suitable for the production of toxic compounds. Nevertheless, this methanol-independent system represents a good alternative to the classic *pAOX1* expression system for regulated gene expression.

A different approach to construct a methanol-free expression system based on *pAOX1* was reported by Shen *et al.* (2016b). Instead of engineering trans-acting elements to achieve methanol independence of *pAOX1*, they focused on the deletion of kinases to alter or disrupt alternative carbon source metabolism. Their study is built upon the aforementioned kinase single knock-out library (Shen *et al.*, 2016a). The regulation of *pAOX1* depends not only on directly acting, trans-acting elements but also on transporters and sensors that import and recognize carbon sources and thereby initiate complex signaling pathways to either induce or to repress *AOX1* expression. For example, deletion of both the hexose transporter *HXT1* and the hexose sensor *GSS1* led to derepressed expression of *AOX1* in response to glucose depletion (Polupanov *et al.*, 2012; Zhang *et al.*, 2010). However, the exact mechanism of signal transduction from carbon source molecules to the transcription factors is still largely unknown. Shen *et al.* (2016b) identified the kinases *GUT1* and *DAK*, whose deletion led to methanol-independent *pAOX1* activity. Both kinases are involved in the glycerol metabolism. Their deletion enabled glycerol inducible expression at up to 25 % the level of methanol induced *pAOX1*. Despite having a reduced growth rate on glycerol, this system can still present a promising alternative for auto-inducible protein expression when applying a process strategy that comprises a glucose batch phase for biomass generation and a glycerol limited fed batch phase for methanol-free product formation. Induction is facilitated by the glycerol metabolism intermediate dihydroxyacetone (*DHA*). Supplementation of a Δ*DAK* strain with *DHA* led to 80 – 90 % of the expression level of the wild-type cells grown on methanol. The approach was validated with different technical enzymes, reaching up to 60 % of the native *pAOX1* expression yields.

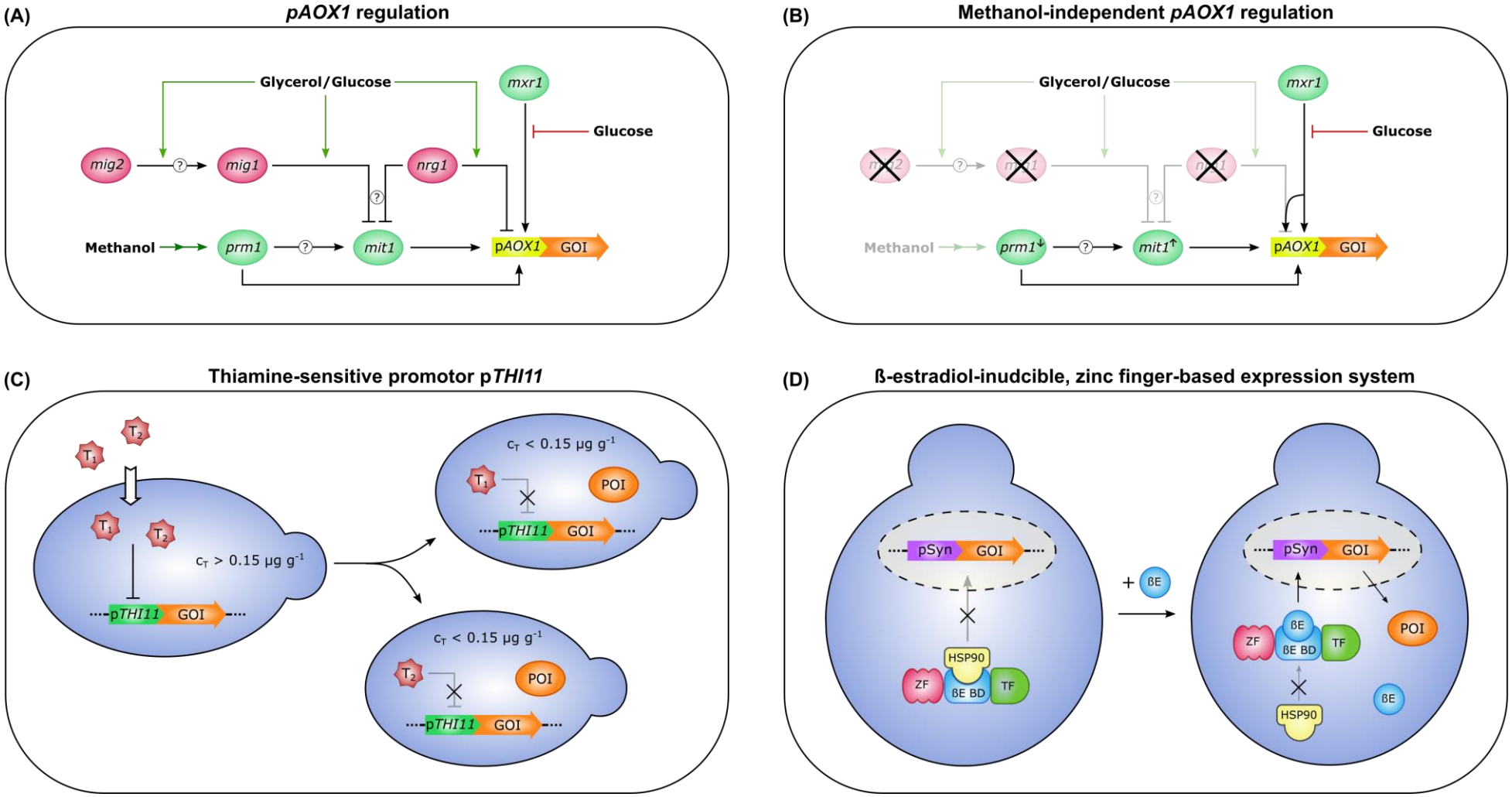
(A)
pAOX1 regulation
Glycerol/Glucose
mxr1
Glucose
mig2
mig1
nrg1
Methanol
prm1
mit1
pAOX1
GOI
(B)
Methanol-independent pAOX1 regulation
Glycerol/Glucose
mxr1
Glucose
Methanol
prm1↓
mit1↑
pAOX1
GOI
(C)
Thiamine-sensitive promotor pTHI11
T1
T2
cT > 0.15 µg g-1
pTHI11
GOI
cT < 0.15 µg g-1
POI
(D)
ß-estradiol-inudcible, zinc finger-based expression system
pSyn
GOI
HSP90
ZF
ßE BD
TF
+ ßE
ßE

Figure 4 (previous page): Regulatory mechanisms of classic and novel *P. pastoris* promoters. **(A)** The native *pAOX1*. *pAOX1* is induced by methanol and repressed with glycerol or glucose, via a complex and not yet fully understood cascade of transcription factors. **(B)** Newly engineered methanol-independent *pAOX1*. Via the deletion of three genes and overexpression of one gene the promoter is not activated by methanol anymore. Rather, a glucose repressible expression was achieved. Induction can be realized by shifting the cells from glucose to glycerol medium. **(C)** Thiamin repressible *pTHI11* promoter. Cells import thiamin but cannot utilize it. The intracellular content decreases during cultivation due to cell division. Once a threshold of 0.15 µg g-1 is surpassed, *pTHI11* transcription is initiated. **(D)** Synthetic expression system. An ATF is first inactivated via *HSP90* and localized to the cytoplasm. Addition of ß-estradiol leads to replacement of *HSP90* and translocation of the ATF to the nucleus. There, the ZF of the ATF binds to ZF recognition sites and activates the downstream minimal promoter, leading to transcription of the target gene.

3.2 Other systems

Many controllable promoters in *P. pastoris* are derived from genes that are involved in peroxisome biogenesis or the MUT pathway and therefore often are induced by methanol (Vogl and Glieder, 2013). Alternatively regulated promoters are rare. To overcome this limitation, Stadlmayr *et al.* (2010) identified 24 novel potential regulatory sequences. Of those, the promoter of the thiamine biosynthesis gene *THI11* seemed to be the most promising. The regulation of *pTHI11* is quite interesting since thiamine cannot be utilized by the cells, the biosynthesis of it is very energy-costly, and the import of exogenous thiamine is irreversible (Iwashima *et al.*, 1973; Landes *et al.*, 2016). Landes *et al.* (2016) were the first to characterize the thiamine-sensitive *pTHI11* and to investigate its suitability for recombinant protein production. In the presence of extracellular thiamine, *P. pastoris* cells rapidly imported the vitamin up to a thiamin level of 1.12 mg/g CDW. During cell growth, the intracellular thiamine level dropped to 0.15 µg/g, indicating that the basal level was reached and maintained by *in vivo* synthesis. No expression of native and heterologous genes under the control of *pTHI11* was observed, if the intracellular thiamine concentration is above a threshold content of 15 µg/g CDW. After the maximal import of exogenous thiamine, intracellular thiamine levels decrease only via dilution caused through cell division as thiamine cannot be utilized by the cells (Praekelt *et al.*, 1994). After a specific number of cell cycles the thiamine content drops below the basal level, allowing the expression of genes under control of *pTHI11* (Fig. 4(C)). Under these non-repressing conditions, *pTHI11*-driven

protein expression was shown, via chemostat cultivations, to be proportional to cell growth and inversely correlated to intracellular thiamine concentration. Transcript levels of *pTHI1* controlled genes remained constant at different dilution rates, indicating a constitutive behavior of *pTHI1*. Based on these results, Landes *et al.* (2016) developed a tailor-made fed batch process strategy to fully exploit the regulatory potential of *pTHI1*. Based on the inability of *P. pastoris* to utilize thiamine and the threshold content needed for repression, they were able to calculate the necessary thiamine concentration in the medium for auto-induction at a desired cell density. After the cells have reached the specific density, a glucose-limited fed batch can be applied to enable a constant growth rate optimal for *pTHI1*-driven expression. The strategy was evaluated against a *pGAP* driven process and similar space time yields were achieved. *pTHI1* regulatory properties represent a promising alternative to *pAOX1*, since it enables programmable, auto-induced protein production. Additionally, *pTHI1* can be used if expression of a recombinant protein needs to be repressed since *P. pastoris* cells take up exogenous thiamine immediately and *pTHI1* responses to intracellular thiamine contents instantaneously (Landes *et al.*, 2016). This strategy was employed by Liu *et al.* (2015), as discussed in chapter 6.1.

To guarantee orthogonality of the promoter, Perez-Pinera *et al.* (2016) developed a synthetic promoter that is induced by ß-estradiol and can exceed *pAOX1* in terms of promoter strength. This promoter consists of the constitutively expressed zinc finger (ZF) DNA binding domain ZF43-8 (Khalil *et al.*, 2012), coupled with the ß-estradiol binding domain of the human estrogen receptor. The receptor is expressed in fusion with the VP64 transcriptional activation domain, creating an artificial transcription factor (ATF) (McIsaac *et al.*, 2013). In the absence of ß-estradiol, the estrogen receptor interacts with HSP90 which leads to ATF translocation to the cytoplasm (Fliss *et al.*, 2000) and thus prevents the ATF form activating gene expression. If ß-estradiol is added to the medium, it is imported by the cells and displaces HSP90 from the estrogen receptor binding pocket. Thereby, the ATF translocates to the nucleus and activates gene expression regulated by a minimal promoter located downstream of multiple ZF binding sites (Fig. 4 (D)). To test their ATF, Perez-Pinera *et al.* (2016) expressed GFP under the control of a minimal *pGAP* and found that expression could be detected with only 0.01 µM ß-estradiol. Expression was highest if nine binding sites for ZF43-8 were

preceding the promoter. One of the advantages of this system is its highly flexible architecture that enables fine tuning of promoter strength at different levels. These include: promotor driving expression of the ZF, affinity of the ZF DNA binding domain, number of binding sides for the ZF, strength of the transcriptional activation domain and minimal promoter driving expression of the gene of interest. It was found that certain combinations of the ATF regulated, synthetic promoter surpassed *pAOX1* expression levels. Often, these combinations also had greater background expression levels in the absence of an inducer. To characterize their expression system for its suitability to produce multiple proteins induced by different signals, they engineered three different *P. pastoris* strains harboring the ATF to express two different proteins. One target was always expressed under the control of *pAOX1* and the other, as well as the ATF, was regulated by varying promoters. All strains were cultivated in glycerol-containing minimal media for 48 h and could express both proteins selectively at high levels. The induction took place after 48 h of outgrowth to guarantee that *pAOX1* was not repressed by glycerol and could be activated by methanol. Overall, a synthetic ß-estradiol-inducible promoter that offers high expression levels and allows for multiplexed expression in *P. pastoris* was developed.

Vogl *et al*. (2016) carried out a microarray experiment to identify genes involved in the MUT, PPP and ROS (reactive oxygen species) pathways to compare the transcriptional response of *P. pastoris* under glucose-repressed, derepressed, and methanol-induced conditions. They cloned promoters from the aforementioned pathways upstream of GFP to characterize the promoter strength. Many of the identified promoters from MUT genes showed strong methanol-induced activity which was at least half of the *pAOX1* activity. Interestingly, a promoter associated with the MUT pathway displaying high activity (29% of *pGAP*) under derepressed conditions was discovered. The promoter regulates expression of a protein involved in ROS defense and was referred to as *pCAT1*. cat1 likely has functions related to H_2O_2 detoxification arising from ROS stress, which in turn is assumed to activate *pCAT1*. Furthermore, *pCAT1* exhibited tight repression on glucose and could be further induced by methanol, leading to expression levels on par with *pAOX1*. It is also the only promoter of MUT or other related pathways that can be induced by oleate, resulting in comparable expression levels to methanol induction. This feature could proof useful for the aforementioned strategy to compartmentalize

metabolic pathways to peroxisomes and use oleate-induced peroxisome biogenesis (chapter 2.3). In total, 15 methanol-regulated promoters were identified that offer good options for combinatorial pathway fine-tuning. As a proof of concept, they used a carotenoid biosynthesis pathway for the synthesis of ß-carotene, consisting of four enzymes. ß-carotene has been produced via the use of a single promoter before (Araya-Garay *et al.*, 2012b). Application of multiple promoters allows for fine-tuning of every single reactions step, reducing the risk of creating bottlenecks in the synthesis of ß-carotene. As a result, the ß-carotene yield was increased to 5 mg/g CDW, compared to 0.3 mg/g CDW in the original publication. This amount is comparable to optimized *S. cerevisiae* strains (Verwaal *et al.*, 2007). Since different terminator sequences are also beneficial for assembling recombinant pathways, they analyzed a set of 20 endogenous terminators qualified for multiplexed gene expression in *P. pastoris*. All tested terminators showed similar capabilities to *AOX1 TT*, but did not surpass it (Vogl *et al.*, 2016).

Another strategy for coordinated multigene expression involves the usage of bidirectional promoters. Bidirectional promoters facilitate the concurrent transcription of both 5‘ and 3‘ open reading frames (ORF). They can be found in many eukaryotic organisms, including humans (Trinklein *et al.*, 2004), and are suspected to play an important role in regulatory signal transcription in yeast (Xu *et al.*, 2009). The *GAL1-GAL10* promoter of *S. cerevisiae* is the best characterized native bidirectional promoter in yeast (Johnston and Davis, 1984). It is used in popular vectors as a strong galactose-inducible promoter to drive the transcription of two GOI simultaneously (Partov *et al.*, 2010). Vogl *et al.* (2015b) published a patent detailing the application of *P. pastoris* bidirectional promoters. The *pHTX1* promoter exhibits a strong constitutive expression profile and consists of a head-to-head fusion of the histone *H2A* and *H2B* promoters of *P. pastoris*. Similarly, fusions of *pGAP* and *pTEF* (translation elongation factor 1-α) as well as *pDAS1* and *pDAS2* were used for the creation of constitutive and methanol inducible bidirectional promoters, respectively (Vogl *et al.*, 2015b). These bidirectional promoters were successfully applied for simultaneous transcription of CAS9 and a gRNA (Weninger *et al.*, 2016), as well as for expression of a target gene and an UPR pathway transcription factor (Krainer *et al.*, 2016). Their use simplifies vector construction,

reduces the risk of loop-out events, and makes the controlled co-expression of targets easier and more predictable.

4 Genetic tractability

P. pastoris exhibits a high genetic accessibility with well-established methods for transformation of various commercial and non-commercial vectors bearing different auxotrophic and antibiotic resistance markers (reviewed in Ahmad *et al.* (2014). More specialized applications like the aforementioned use of the split marker system (Heiss *et al.*, 2013) and *KU70* deletion strains (Näätsaari *et al.*, 2012), as well as the Cre-Lox recombinase system (Pan *et al.*, 2011) are also available. However, the discussed challenges of clonal variability, NHEJ pathway and lack of episomal vectors complicate genetic engineering projects of higher complexity. To this end, much progress has been made in the last few years to expand the genetic toolbox for *P. pastoris* and streamline the generation of metabolically engineered strains.

4.1 Pathway assembly

In yeast biotechnology genes are typically transcribed in a monocistronic manner, meaning each target gene requires its own promoter and terminator, complicating the construction of larger metabolic pathways. On the other hand, many bacterial operons like the *E. coli* lac operon transcribe a single mRNA for multiple proteins. Such polycistronic operons simplify the necessary vector construction and can enable easier pathway assembly, expression and modification. Using internal ribosome entry sequences, the separate translation of the encoded proteins is realized in bacteria. However, their use is limited by the comparatively high length (~ 500 bp) and the tendency of downstream genes to be expressed only at low levels (Douin *et al.*, 2004; Nielsen *et al.*, 2009). An emerging alternative are 2A sequences derived from the mRNA translation mechanism of RNA viruses. They are ca. 50 bp short and result in a ribosomal skip during translation, facilitating self-cleavage of the 2A sequence and synthesis of individual proteins from polycistronic mRNA (De Felipe *et al.*, 2006). This co-translational self-processing of multiple proteins under control of one promoter has already been applied in biotechnology (Szymczak *et al.*, 2004). Geier *et al.* (2015b) transferred and expanded the use of the 2A system in *P. pastoris*. Assaying the suitability

of a proven 2A sequence (Shah *et al.*, 2015; Wang *et al.*, 2007) and three variants novel to *P. pastoris*, they were able to express up to nine genes under control of a single promoter. Furthermore, they successfully tested the possibility of controlling expression strength by placing the genes in different orders, enabling simplified fine-tuning of an entire metabolic pathway. By combining the 2A-mediated polycistronic expression with a bidirectional promoter the simultaneous expression of two different heterologous pathways (ß-carotene and violacein synthesis) was realized (Fig. 5 (A)). This approach facilitates the efficient and space-saving construction of pathways in *P. pastoris*. It requires only a single transformation step, contains less internal homologous sequences for looping out events, enables easy tuning of expressional strength and simplifies the regulatory control of expressing multiple pathway genes. Nevertheless, remaining challenges of the 2A system, e.g. incomplete self-processing leading to fusion proteins, have to be addressed in future studies (Geier *et al.*, 2015b).

An essential part to facilitate efficient metabolic engineering of a microorganism, is the possibility to quickly and modularly assemble vectors. This way it is possible to generate large constructs encoding entire pathways by transforming cells with easily synthesizable and transferable small DNA fragments. Assembly can be realized *in vitro* or *in vivo* and many techniques like Gibson Assembly, Golden Gate, BioBricks and more have been developed in the last years (reviewed in Casini *et al.* (2015)). *In vivo* DNA assembly mediated by complementary overlaps has been widely applied in *S. cerevisiae*, due to the extraordinary HR capability of this yeast (Essani *et al.*, 2015). In contrast, only few reports of its use in *P. pastoris* exist (Yu *et al.*, 2012). Recently, the concept was further developed and characterized. Camattari *et al.* (2016) demonstrated that overlaps of 20 to 50 bp length were sufficient for efficient assembly of a three-part vector *in vivo* (Fig. 5 (B)). Molar ratios between the different fragments were optimized, resulting in 90 % of transformants containing the correctly assembled vector. The split-marker method (Heiss *et al.*, 2013) was used to increase the frequency of correctly assembled vectors after transformation. Transformation efficiencies were lowered compared to transforming the entire vector at once. Nevertheless, the *in vivo* assembly enables easy construction of new vector combinations simply by PCR amplification and transformation of the desired components, by-passing the typical *E. coli* cloning step.

On a similar note, an important synthetic biology toolkit from *S. cerevisiae* was transferred to *P. pastoris* very recently. Based on the Golden Gate assembly technique and the MoClo (modular cloning) system of standardized DNA building blocks, a large library of well characterized parts was published for *S. cerevisiae* (Lee *et al.*, 2015). 96 parts are split into categories like promoter, terminator, selection marker and various coding sequences for reporter proteins. All parts were thoroughly characterized for promoter and terminator strength, efficiency of protein degradation tags and targeting efficiency of chromosomal integration, amongst others. To facilitate straight-forward assembly of parts, they contain predefined overhangs compatible with the *Bsa*I and *BsmB*I restriction enzymes used for Golden Gate assembly. Using this approach, parts can easily be combined into a single-gene vector, which in-turn can be expanded to a multi-gene vector. The modular nature of the parts, the large number of them and their well characterized features make this tool box ideal for combinatorial metabolic engineering in yeast. Building upon this toolkit, Obst *et al.* (2017) added *P. pastoris* specific parts (e.g. *pAOX1* and *pGAP*) and characterized them together with selected *S. cerevisiae* parts in *P. pastoris* (Fig. 5 (C)). In total, 26 parts can be combined into 264 different vectors. Addition of the remaining *S. cerevisiae* parts would increase this number to over 4000. While the *P. pastoris* toolkit focuses on (secreted) protein expression, it can easily be extended and applied for metabolite production projects. The publication marks the first *P. pastoris* library of characterized, sequenced and standardized DNA parts for modular vector construction. In order to facilitate efficient distribution to the scientific community, the plasmids bearing the standardized parts are available at the Addgene repository. On a related note, a family of 40 plasmids for use with type IIS restriction enzymes like *Bsa*I was reported (Vogl *et al.*, 2015a).

4.2. New integrative and episomal techniques

The CRISPR/Cas9 technology has been discovered and subsequently developed for the application in various organisms over the past years (reviewed in Komor *et al.* (2016)). Based on a bacterial defense mechanism against foreign DNA invasion, the system was successfully applied to delete or mutate endogenous genes, as well as integrate foreign DNA at a targeted locus. Targeting is facilitated by a ca. 20 bp long gRNA (guide RNA) complementary to the target locus, which directs the cas9 nuclease to the target locus

and introduce a strand break (Fig. 5 (D)). Breaks can either be repaired by the error-prone NHEJ pathway, resulting in deletions and mutations, or via the HR pathway by providing a homologous DNA repair fragment. Using the HR pathway, foreign DNA can be integrated and due to the DSB break HR efficiency is markedly increased (Storici *et al.*, 2003). In *S. cerevisiae* CRISPR/Cas9 is well-established (DiCarlo *et al.*, 2013) and has been refined for multiplexed-genome editing and transcriptional silencing (Jensen *et al.*, 2017; Mans *et al.*, 2015). Recently, the system was transferred to *P. pastoris* (Weninger *et al.*, 2016). The study is based on the discovery and characterization of multiple nuclear localization sequences necessary for cas9 transfer to the nucleus (Weninger *et al.*, 2015) and bidirectional promoters for co-expression of cas9 and the gRNA (Vogl *et al.*, 2015b). Strikingly, 95 different combinations of cas9 variants (e.g. codon optimized), cas9 and gRNA promoters with and without ribozymes as well as different gRNA sequences needed to be tested to find a functioning setup. Out of these 95 combinations eleven were functional, but only six exhibited targeting efficienes >75 % with the best one having a 94 % efficiency. Cytotoxic effects due to cas9 hyperactivity are suspected to cause the inefficiency of most combinations. Nevertheless, Weninger *et al.* (2016) demonstrated the highly efficient, targeted and multiplexed genome editing capabilities of their system, combined with an HR integration efficiency on-par with conventional systems. Hence, the currently available CRISPR/Cas9 method should help e.g. in knock-out studies by allowing for time-saving procedures, while further improvements are needed for its application for foreign DNA integration. Potentially, newly developed Cas9 "successors" like the vertebrate based activation-induced cytidine deaminase (AID) tool could help overcome the limitations of the current system. AID was successfully applied in *S. cerevisiae*, where it displayed markedly reduced cytotoxicity (Nishida *et al.*, 2016).

In addition to the established Cre-Lox method (Pan *et al.*, 2011), a new recombinase based system was developed. Three recombinases (BxbI, R4 and Tp901-1) can be used to facilitate the integration of the vector (Perez-Pinera *et al.*, 2016). For targeted integration, a designated "landing pad" of heterologous DNA was previously integrated into the genome of the *P. pastoris* strain. The system offered high targeting efficiencies and via the landing pad an added security measure against off-target activities. To this end it was also applied in the previously mentioned creation of a modular vector toolkit

(Obst *et al.*, 2017). It has to be noted, that the use of a landing pad necessitates previous strain engineering work and restricts the potential for future applications. However, an interesting aspect of the presented approach is its capability to be expanded for synthetic gene circuit design. Gene circuits consist of genes flanked by recombinase recognition sites that enable inverting the gene to turn its expression on or off (Roquet *et al.*, 2016). This technology is still in its infancy, but could prove useful for biotechnological applications in the future, as has been demonstrated by creating expression systems that automatically start induction upon reaching a predefined cell-density (Soma and Hanai, 2015).

A different approach for enhanced targeting efficiency was proposed by Tsakraklides *et al.* (2015). Cells were reversibly arrested in the S-phase of mitosis, in which DNA is duplicated, via hydroxyurea treatment. It has been shown for *S. cerevisiae* that arrest of cell in this phase led to increased HR activity (Galli and Schiestl, 1996). Similarly, this approach facilitated increased targeting efficiencies in the assayed non-conventional yeasts (Tsakraklides *et al.*, 2015). Especially for *Yarrowia lipolytica* and *Arxula adeninivorans* significant improvements were observed. On the other hand, *P. pastoris* targeting efficiency was only increased from 1.6 % to 5.4 %. However, very short homologous sequences (50 bp) were used that are not representative of the long sequences of ca. 1000 bp, typically employed in *P. pastoris*. It would be of interest to investigate whether cell cycle synchronization also benefits targeted integration in *P. pastoris* with commonly employed homologous sequences. The system might enable the use of shorter homologues sequences, while enhancing targeting efficiency and not requiring specifically engineered strains.

Besides the clonal variability, the lack of (episomal) plasmids has been cited as a drawback of the *P. pastoris* system (Kelwick *et al.*, 2014). In yeast ARS are needed for replication of episomal vectors. A genome wide search for ARS consensus sequences resulted in the discovery a large number of potential ARS (Liachko *et al.*, 2014). Some of these were characterized and it was found that, unlike most other yeasts, *P. pastoris* contains mostly GC-rich ARS. In-tandem, an ARS from *K. lactis* (panARS) was shown to exhibit a very broad host-range of budding yeasts, including *P. pastoris* (Liachko and Dunham, 2014). These insights were combined to assay the suitability of ARS based

vectors for recombinant protein production in *P. pastoris* (Camattari *et al.*, 2016). panARS facilitated higher expression levels than the two assayed chromosomal *P. pastoris* ARS and a strain with an integrated expression cassette. Furthermore, the panARS vector did not only outperform the productivity of the integrative system, it also provided a much higher clonal homogeneity. Similar observations have been made with an ARS vector system based on a mitochondrial ARS from *P. pastoris* (Schwarzhans *et al.*, 2017). Here, an accidental integration event resulted in the creation of a circular plasmid containing the mitochondrial ARS. Further studies revealed its capability for high level recombinant protein production on-par with panARS, lowered clonal variability and suitability for transformation into *S. cerevisiae*. A preliminary scan of the mitochondrial genome revealed the presence of over 500 putative ARS, making them promising candidates for construction of novel *P. pastoris* ARS vectors. Both panARS and the mitochondrial ARS significantly expand the episomal plasmid repertoire of *P. pastoris*. Simultaneously, they present an option to tackle clonal variability issue by providing uniform expression levels and causing no genetic perturbance. In consequence, they appear to be best suited for screening experiments, e.g. protein engineering studies. The homogenous productivity would emphasize effects resulting from protein features over those from different expression levels. Both newly reported vectors displayed high stability (close to 100%) under selective conditions, but these values dropped dramatically if the selective pressure was removed (Camattari *et al.*, 2016; Schwarzhans *et al.*, 2017). For auxotrophic markers, a decreasing selective pressure during cultivation, due to release of the relevant metabolite into the medium by prototrophic strains, has to be considered (Pronk, 2002). Suitable selection markers and adapted processes are necessary to ensure stable productivity. Interestingly, for both ARS vectors it was found that productivity was not directly correlated with the GCN (Camattari *et al.*, 2016; Schwarzhans *et al.*, 2017). Strains with markedly different GCN exhibited highly similar protein expression levels. Post-translational and epigenetic factors are suspected to cause this phenomenon (Love *et al.*, 2010). Possibly, *P. pastoris* is capable of higher expression levels with ARS based vectors but an unknown bottleneck is stopping it from reaching the theoretical potential.

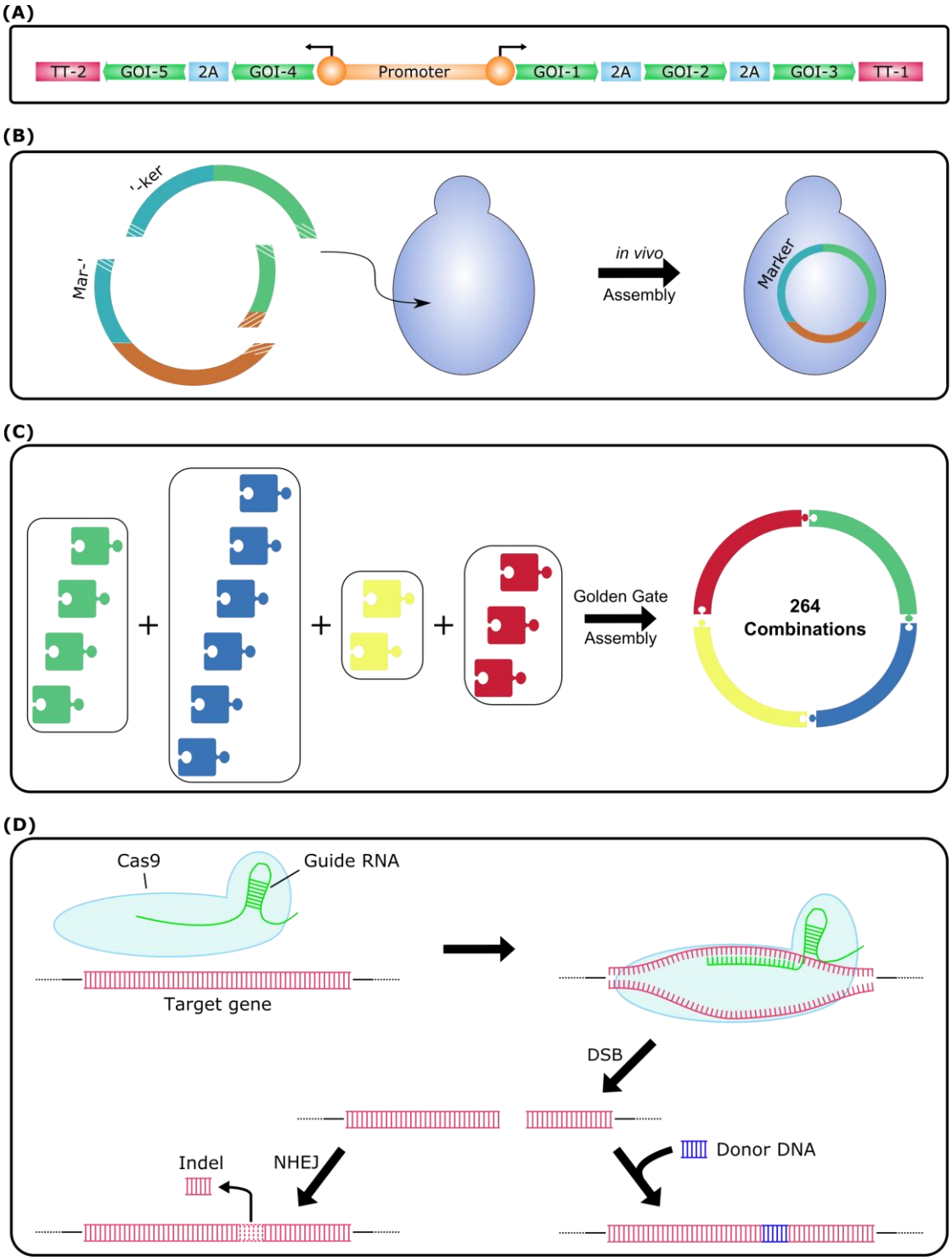

Figure 5: New tools for genetic manipulation in *P. pastoris*. **(A)** Combination of bi-directional promoter and 2A-mediated polycistronic expression of a multi-gene pathway. **(B)** *In vivo* assembly of an ARS vector via overlapping sequences on the different fragments. Using the split-marker system, the frequency of correctly assembled plasmids is increased. **(C)** Transfer and adaption of a modular multipart assembly toolkit to *P. pastoris*. Via Golden Gate assembly a total of 264 plasmid variants can be constructed from standardized and characterized parts. **(D)** CRISPR/Cas9 cartoon. A DBS is generated and either repaired by NHEJ, leading to indels, or with donor DNA for gene insertion.

5. Genetic engineering of platform strains

P. pastoris has been extensively engineered to improve protein production characteristics, study gene or organelle functions and to produce metabolites. To emphasize the increased complexity of these pursuits, the numbers of modified (knocked-in, knocked-out, upregulated and/or downregulated) genes are shown in the following chapters. A steady increase in engineering complexity can be noticed, with more studies focusing on simultaneous knock-out and overexpression of genes, instead of constructing separate strains for each modification. First, an overview of key glycoengineering, recombinant protein production and other studies focusing on non-metabolite targets is given. Second, metabolic engineering publications are detailed, showcasing the increased publication frequency in this field since 2014.

5.1 Glycoengineering

The production of glycoproteins received much attention in the first decade of this century (Table 2). *P. pastoris* tends to hyper-mannosylate N-glycans with typically around 10-20 mannose residues, displaying a far lower hyper-mannosylation extent than *S. cerevisiae*, in which up to 200 mannose residues can be found (Dean, 1999; Vervecken *et al.*, 2004). This circumstance eased the humanization of the N-glycosylation in *P. pastoris*. In 2006, scientists from the company GlycoFi (Lebanon, NH), later acquired by Merck & Co (Boston, MA), succeeded in fully humanizing the N-glycosylation (Hamilton *et al.*, 2006; Li *et al.*, 2006). Additionally, they introduced the full biosynthetic pathway for the human CMP-N-actelyneuraminic acid biosynthesis pathway and necessary genes for transfer of the synthesized sialic acid to the mature N-glycan. This marked the genetic engineering project of highest complexity in *P. pastoris* to date with 14 knock-ins and 4 knock-outs (Hamilton *et al.*, 2006). A detailed review on the latest developments of glycoengineering of *P. pastoris* can be found in Laukens *et al.* (2015).

However, after this boom period relatively few publications on the subject were made. One main concern regarding industrial applicability of the engineered strains was the lowered strain fitness. This was in a large part due to the deletion of *OCH1* (α-1,6-mannosyltransferase), resulting in increased flocculation, cell lysis and temperature

sensitivity (Davidson *et al.*, 2004; Krainer *et al.*, 2013). Jiang *et al.* (2015) discovered that deletion of *ATT1*, a homolog of the *S. cerevisiae GAL4* transcriptional activator, improved strain fitness. On the one hand, this result benefits efforts to create glycoengineered *P. pastoris* strains with increased robustness, capable of withstanding the harsh conditions of high-cell density fermentations in large-scale applications. On the other hand, it also furthers our understanding of *P. pastoris* physiological properties. In *S. cerevisiae GAL4* is part of the galactose utilization pathway, but *P. pastoris* is incapable of metabolizing galactose, due to the evolutionary loss of the involved metabolism pathway genes. The conservation of *ATT1* indicates its involvement in the transcriptional regulation of genes unrelated to galactose metabolism. Besides the increased strain fitness, this theory is further corroborated by experiments showing the translocation of att1 to the nucleus and the presence of *ATT1*-associated binding motifs in over 400 promoters of the *P. pastoris* genome (Jiang *et al.*, 2015).

A quite different glycoengineering strategy has recently emerged. Its goal is to completely remove N-glycans from proteins that do not require them for their activity and thereby improve product homogeneity. Termed “GlycoDelete”, the system uses a fungal endoglycosidase endoT which removes the entire N-glycan (except a single GlcNAc residue) from the glycoprotein when it passes through the golgi (Claes *et al.*, 2016). This approach is especially helpful for the production of membrane proteins, providing a uniform product better suited for pharmaceutical application. Many (viral) membrane proteins can be used as basis for vaccine development, as they often play an important part in receptor-based cell recognition (Fogg *et al.*, 2004). This approach should also have a lower impact on strain fitness than the extensive genetic engineering procedures necessary for humanizing the N-glycosylation. The GlycoDelete system might therefore prove to be of interest for expanding the capabilities of *P. pastoris* for functional expression of glycoproteins whose activity is changed by an incorrect glycosylation, often associated with expression in a heterologous eukaryotic system.

Table 2: Overview of key genetic engineering projects in *P. pastoris* not targeting the production of metabolites. The number of knocked-out, knocked-in, downregulated and upregulated genes is shown. In case multiple modifications were done separately from one another the number is given as "n x z", with n as the amount of modification targets and z as the number of simultaneous modifications in a strain. Items are ordered chronologically.

Target	Knock-in	Knock-out	Upregulated	Downregulated	Reference
Glycoengineering					
GlcNAcMan5GlcNAc2 N-glycans	3	1	0	0	(Choi *et al*., 2003)
Man5GlcNAc2 N-glycans	3	1	0	0	(Vervecken *et al*., 2004)
GlcNAc2Man3GlcNAc2 N-glycans	5	2	0	0	(Bobrowicz *et al*., 2004)
Fully humanized N-glycans (terminally sialylated)	14	4	0	0	(Hamilton *et al*., 2006)
Fully humanized N-glycans (nonsialylated)	8	4	0	0	(Li *et al*., 2006)
Gal2GlcNAc2Man3GlcNAc2 N-glycans	8	1	0	0	(Jacobs *et al*., 2009)
Strain fitness	0	1	0	0	(Jiang *et al*., 2015)
N-glycan trimming	3	0	0	0	(Claes *et al*., 2016)
Recombinant protein production					
Proteolysis	0	2	0	0	(Werten and De Wolf, 2005)
Protein secretion and folding	2	0	0	0	(Gasser *et al*., 2006)
Protein secretion and folding	0	0	1	0	(Guerfal *et al*., 2010)
Protein secretion	0	0	8 x 1	0	(Baumann *et al*., 2011)
Engineered cell cycle phase	0	0	8 x 1	0	(Buchetics *et al*., 2011)
Methanol utilization	0	0	3	0	(Krainer *et al*., 2012)
Glycoprotein homogeneity	0	1	0	0	(Krainer *et al*., 2013)
Central carbon metabolism	0	4 x 1	7 x 1	0	(Nocon *et al*., 2014)
Heme cofactor availability	0	0	8 x 1	0	(Krainer *et al*., 2015)
Pentose pathway flux	0	0	4	0	(Nocon *et al*., 2016)
Protein secretion and proteolysis	0	3	2 x 1	0	(Marsalek *et al*., 2017)
Other					
Biotin-prototrophy	4	0	0	0	(Gasser *et al*., 2010)
Peroxisome biogenesis	0	3 x 1	0	0	(Polupanov *et al*., 2011)
UPR pathway study	1	0	3 x 1	0	(Delic *et al*., 2012)
Deletion of native secreted protein	0	1	0	0	(Heiss *et al*., 2013)
NADH regeneration	0	2	0	0	(Geier *et al*., 2015a)
Fatty acid composition stress response	0	3 x 1	0	0	(Zhang *et al*., 2015)
Methanol-free AOX1 induction	0	2 x 1	0	0	(Shen *et al*., 2016b)
Rhamnose utilization pathway	0	1	0	0	(B. Liu *et al*., 2016)
Kinase involvement in cell growth	0	92 x 1	0	0	(Shen *et al*., 2016a)
Methanol-free AOX1 induction	0	3	1	0	(Wang *et al*., 2017)

5.2 Recombinant protein production

Besides glycoproteins in particular, the improvement of expression and secretion of heterologous proteins in general has been in the focus of genetic engineering research. Early it was discovered that the co-expression of chaperones and foldases, both heterologous (*S. cerevisiae*) or native, can improve folding and secretion of proteins (Gasser *et al.*, 2006; Guerfal *et al.*, 2010). Endogenous protease genes like *KEX2* and *YPS1* were deleted, thereby reducing proteolysis (Werten and De Wolf, 2005). Often, a specific combination beneficial for one heterologous protein does not show the same effect for another target. This led to the development of several approaches aimed at knocking-in helper proteins and removing proteases to tailor the expression strain to the desired application (reviewed in Puxbaum *et al.* (2015)). Complex influences of membrane properties (Baumann *et al.*, 2011), the cell cycle (Buchetics *et al.*, 2011) and the MUT pathway (Krainer *et al.*, 2012) on recombinant protein production yields have been assayed in genetic engineering studies by overexpressing selected genes. Typically, the native promoter was replaced with the strong constitutive *pGAP* or the methanol-inducible *pAOX1* promoter.

Historically, genes designated for modification were chosen based on textbook knowledge about metabolic pathways, e.g. genes involved in the ergosterol biosynthesis or methanol metabolism. Nocon *et al.* (2014) expanded upon this principle by selecting target genes using a model-based approach. Both overexpression and knock-out targets were identified using a GEM by Sohn *et al.* (2010). Overexpression targets were selected using the Flux Scanning based on Enforced Objective Function (FSEOF) (Choi *et al.*, 2010) and knock-out targets via Minimization Of Metabolic Adjustment (MOMA) (Segrè *et al.*, 2002), respectively. Five out of nine predicted targets resulted in significantly improved protein production. Although many candidate genes were assayed, genetic manipulations were conducted separately with no strain containing more than one modification. It would be of interest to see whether combinations of multiple knock-outs or upregulations result in even better yields and how well these combinatorial effects agree with the model predictions. One of the targets predicted by the models, the pentose phosphate pathway (PPP), was indeed further analyzed and four pathway genes simultaneously overexpressed (Nocon *et al.*, 2016). Some combinations

enabled added protein productivity, but others led to unpredicted imbalances in the PPP and lowered product titers.

To increase availability of the heme co-factor for recombinant proteins, the effect of separate overexpression of eight different pathway genes was assayed (Krainer *et al.*, 2015). Interestingly, no increase in heme-requiring protein production was observed. Instead, supplementation of the medium with a heme precursor delivered favourable results. Six more genes involved in the heme biosynthesis pathway were not targeted and no combination of multiple overexpression targets was performed. It is therefore possible, that future endeavors of combinatorial overexpression, potentially based on model predictions, could yield different results.

Besides glycoengineering, a new application for *OCH1* deletions in *P. pastoris* was reported by Weinhandl *et al.* (2016). Secretion of a recombinant branched chain aminotransferase was improved approximately threefold by using a *OCH1* deletion strain, compared to the wildtype. Two factors are suspected to facilitate this effect, (i) reduced hyper-mannosylation easing correct exocytosis (ii) increased permeability of the cell wall due to lowered integrity. However, large amounts of native proteins involved in cell wall assembly were secreted concomitantly, complicating purification of the target protein. It was suspected that inefficient cell wall assembly in growing cells was the root cause. To investigate this thesis, two deletions of genes involved in cell wall assembly were separately introduced in the *OCH1* deficient strain. While secretion rates of the target protein were slightly reduced, the secretion of contaminating proteins was almost completely eliminated in one case. Furthermore, the double knock-out strain displayed reduced cell clumping and a specific growth rate similar to the wild type, in contrast to the ca. 30 % decreased growth rate of the *OCH1* deletion strain. These results might also prove valuable for glycoengineering studies, providing more robust and faster growing host strains, similar to the previously mentioned results by Jiang *et al.* (2015).

Recently, the effect of vacuolar protein sorting (VPS) on protein secretion was investigated by generating VPS mutant strains (Marsalek *et al.*, 2017). The goal was to increase secretion rates of a recombinant protein, tackling a decreased protein yield due to intracellular missorting. While knock-outs of two genes involved in endosomal sorting had a drastic negative effect on secretion rates, a combination with deletion of a

cellular protease led to extracellular recombinant protein yields up to 80 % higher than in the original strain. Another insight from this study is that VPS3 is an essential gene in *P. pastoris*, whereas its deletion mutants are viable in *S. cerevisiae* (Peplowska *et al.*, 2007). Together with the diametric opposite effect of deleting VPS genes on protein secretion, a different mechanism of VPS in *P. pastoris* compared to *S. cerevisiae* can be suspected.

5.3 Other projects

Other genetic engineering studies in *P. pastoris* include the construction of a biotin-prototrophic strain (Gasser *et al.*, 2010), studies on peroxisome biogenesis (Polupanov *et al.*, 2011) or the unfolded protein response (UPR) pathway (Delic *et al.*, 2012), as well as removal of the major secreted native protein (Heiss *et al.*, 2013). While some of these studies are closely tied to protein production intents, they also give insights on a more fundamental research level. For example, the biotin-prototrophic strain was presented as being advantageous for protein production processes due to e.g. simpler and cheaper media composition. Nevertheless, the co-expression of four foreign genes in order to construct the biotin synthesis pathway was realized in a more sophisticated way than simply putting all genes under the control of *pGAP*. Integrating results from Stadlmayr *et al.* (2010) regarding novel native promoters, Gasser *et al.* (2010) expressed the four heterologous genes under the control of two different promoters. This approach offered the ability to fine-tune expression levels of the pathway genes by using differently strong promoters and also reduced the risk of genetic instabilities due to internal sequence homologies.

In 2015, a chassis strain with markedly improved NADH generation was presented (Geier *et al.*, 2015a). Two genes from the methanol assimilation pathway were deleted, encoding dihydroxyacetone synthase isoform 1 (*DAS1*) and isoform 2 (*DAS2*). Without the gene products of *DAS1* and *DAS2*, the methanol assimilatory pathway is disrupted and methanol cannot be converted into biomass. Instead, methanol is funneled into the dissimilatory pathway, resulting in the generation of two molecules of NADH per molecule of methanol. Geier *et al.* (2015a) could demonstrate that cells with disrupted methanol assimilation pathway were better suited for a NADH dependent enzymatic whole cell catalysis reaction, if cells were grown in methanol. The constructed strain

outperformed the wildtype with an approximately 5 fold higher conversion rate. Due to the wide applicability of NADH as cofactor, this approach could benefit many enzymatic conversion projects. The method does seem more suited for whole cell catalysis studies, since the disrupted methanol assimilation pathway results in very low growth rates on methanol. However, by using non-repressing carbon sources in a combined feed the *DAS1 DAS2* double knock-out could potentially also be used in a living cell system (Inan and Meagher, 2001). Thereby, NADH regeneration would be enhanced by the metabolized methanol and e.g. sorbitol used for biomass formation.

In an impressive endeavor, 92 of the 152 annotated protein kinases on the *P. pastoris* genome were separately knocked-out (Shen *et al.*, 2016a). It is suspected, that the remaining 60 kinase genes could not be deleted due to their potential essential functions. The important role kinases play in activating and deactivating proteins via phosphorylation and thereby affecting the cellular machinery on multiple levels is well known. For all knock-out variants the growth behavior on the carbon sources methanol, glucose and glycerol as well as the *pAOX1* activity on the respective media was assayed. Using this approach, 27 strains with impaired growth on specific carbon sources or changes in the regulation of *pAOX1* expression were discovered. This project constitutes the largest single-gene knock-out library in *P. pastoris* to date. Single-gene knock-out libraries, like the Keio collection of *E. coli* with ca. 4000 mutants (Baba *et al.*, 2006), provide an important tool for scientists. In themselves they offer valuable information about the roles of gene products in an organism and serve to predict potential essential genes. Furthermore, they enable researchers to quickly obtain knock-out strains for their experiments and more easily generate multi-gene knock-out strains. Although the library of Shen *et al.* (2016a) only contains knock-outs for genes of a specific class, it could serve as a starting point for a similar collection of *P. pastoris* single-gene knock-out strains. Deletions were performed in the popular GS115 strain and the basic growth characteristics of all mutants have been characterized and documented, providing a good basis for further expansion.

6 Metabolite production

Over the years a variety of metabolites belonging to different substance classes have been produced in *P. pastoris* (Table 3). However, the metabolite portfolio is by far not as well developed as for recombinant proteins. In the early days of establishing biotechnological applications of *P. pastoris*, metabolite and protein production were simultaneously developed. Researchers from DuPont succeeded in constructing strains producing pyruvic acid (Eisenberg *et al.*, 1997), glyoxylate (Payne *et al.*, 1997), or glyphosate (Gavagan *et al.*, 1997; Payne *et al.*, 1995) and strains capable of stereoselective nitrile hydrolysis (Wu *et al.*, 1999). These achievements are even more impressive due to the lack of knowledge about genomic, transcriptomic and metabolomics properties of *P. pastoris* at that time. Sadly, little progress was made in the following years. Research on establishing *P. pastoris* as protein production host intensified and interest in other projects apparently declined. Nevertheless, in the late 2000s interest seemed to slowly be reignited. Marx *et al.* (2008) upregulated six genes involved in the biosynthesis of the vitamin riboflavin by replacing their native promoters with *pGAP*. The synthesis of the carotenoids lycopene, ß-carotene and astaxanthin was realized (Araya-Garay *et al.*, 2012a, 2012b; Bhataya *et al.*, 2009). Additionally, the hydroxylation of bufuralol to 1'-hydroxybufuralol by a membrane-bound human cytochrome P450 monooxygenase (CYP) in a whole cell catalysis approach demonstrated the possible application of *P. pastoris* as biocatalyst of drug metabolites (Geier *et al.*, 2012). The exceptionally high expression levels of the employed CYP, thanks to *P. pastoris* good recombinant protein production capabilities, enabled the strain to significantly outperform other assayed microbial hosts during whole cell catalysis. In 2013 the first polyketide of heterologous origin was synthesized in *P. pastoris* (Gao *et al.*, 2013). Requiring the introduction of only two genes, up to 2.2 g/L of the polyketide 6-methylsalicylic acid was produced.

Potentially fueled by the available genome sequences of popular strains and rising amounts of transcriptomics, proteomics and metabolomics data, an increased number of published metabolic engineering studies was visible in recent years (Table 3). In 2014 to 2017, *P. pastoris* was used for the production of compounds belonging to various chemical classes. In the following sections highlights of the recent advancements regarding terpenoids, isoflavonoids, polyketide and other metabolites are discussed.

Table 3: Comprehensive list of genetic engineering projects in *P. pastoris* aimed at producing various metabolites and the number of introduced genetic modifications. Multiple alterations that were done separately from one another are shown as "n x z", with n as the amount of modification targets and z as the number of simultaneous changes in a strain. Entries are ordered chronologically.

Target	Substance class	Knock-in	Knock-out	Upregulated	Downregulated	Reference
Pyruvic acid[1]	Organic acid	1	0	0	0	(Eisenberg *et al*., 1997)
Glyoxylate	Organic acid	2	0	0	0	(Payne *et al*., 1997)
Nitrile hydratase activity[1]	Amide	3	0	0	0	(Wu *et al*., 1999)
Polyhydroxyalkanoate[2]	Polyester	1	0	0	0	(Poirier *et al*., 2002)
Riboflavin	Vitamin	0	0	6	0	(Marx *et al*., 2008)
Lycopene[2]	Terpenoid	3	0	0	0	(Bhataya *et al*., 2009)
Conversion of cephalosporin C[1,2]	Polyketide	1	0	0	0	(Abad *et al*., 2010b)
Lycopene, ß-carotene & astaxanthin	Terpenoid	6	0	0	0	(Araya-Garay *et al*., 2012a, 2012b)
1'-Hydroxybufuralol[1]	Beta blocker	2	0	0	0	(Geier *et al*., 2012)
6-Methylsalicylic acid	Polyketide	2	0	0	0	(Gao *et al*., 2013)
Xylitol	Saccharide	2	0	0	0	(Cheng *et al*., 2014)
Biodiesel[1]	Fatty acid derivative	1	0	0	0	(Yan *et al*., 2014)
S-adenosyl-L-methionine	Amino acid	3	0	0	0	(Kant *et al*., 2014)
Long-chain fatty acids	Fatty acid	2	0	0	0	(Jiang *et al*., 2014)
Very long-chain fatty acids	Fatty acid	3	0	0	0	(Kim *et al*., 2014)
Hyaluronic acid	Polysaccharide	2	0	3	0	(Jeong *et al*., 2014)
Ergosterol	Steroid	1	0	0	0	(Y. Liu *et al*., 2014)
(+)-nootkatone	Terpenoid	4	0	33 x 1	0	(Wriessnegger *et al*., 2016, 2014)
Cytochrome P450 reductase stabilization[1]	Terpenoid	2	0	1	0	(Emmerstorfer *et al*., 2015)
Oosperin	Polyketide	3	0	0	0	(Feng *et al*., 2015)
Violacein & ß-carotene	Polyketide & Terpenoid	9	0	0	0	(Geier *et al*., 2015b)
Bioethanol	Alcohol	3	0	0	0	(Shin *et al*., 2015)
Dammarenediol-II	Terpenoid	1	0	1	1	(Liu *et al*., 2015)
Ricinoleic acid	Fatty acid derivative	2	1	0	0	(Meesapyodsuk *et al*., 2015)
Fatty acid branched-chain esters	Fatty acid	3	0	0	0	(Tao *et al*., 2015)
Medium-chain volatile flavour esters[1]	Fatty acid derivative	2	0	0	0	(Zhuang *et al*., 2015)
Δ9-tetrahydrocannabinolic acid[1]	Cannabinoid precursor	1	0	0	0	(Zirpel *et al*., 2015)
6-Hydroxydaidzein & 3'-Hydroxygenistein[1]	Isoflavonoid	3	0	0	0	(Chiang *et al*., 2016; Wang *et al*., 2015)
Stylopine	Alkaloid	3	0	0	0	(Hori *et al*., 2016)
Glucaric acid	Organic acid	2	0	0	0	(Y. Liu *et al*., 2016)
L-lactic acid	Organic acid	2	0	0	0	(de Lima *et al*., 2016)
Testosterone[1]	Steroid	2	0	0	0	(Shao *et al*., 2016)

Dammarenediol-II	Terpenoid	2	0	0	0	(Zhao *et al*., 2016)
Citrinin	Polyketide	7	0	0	0	(Xue *et al*., 2017)
Bioethanol[3]	Alcohol	2	0	0	0	(Zhang *et al*., 2017)

6.1. Terpenoids

Terpenoids are organic compounds that consist of multiple units of isopentenyl pyrophosphate (IPP) and its isomer dimethylallyl pyrophosphate (DMAPP). Many thousand terpenoid variants are known and a variety of them have proven or potential applications in medicine, cosmetics, food, animal feed, or as platform chemicals (Withers and Keasling, 2007). In *P. pastoris* IPP is generated from Acetyl-CoA via the mevalonate (MVA) pathway (Fig. 3). Unlike bacteria, the MVA pathway is the only IPP synthesis pathway in yeasts and exhibits high efficiencies, resulting in IPP making up 5 % of the dry weight under ideal conditions (Lamacka and Sajbidor, 1997). Therefore, they are the ideal host for heterologous production of industrially relevant terpenoids.

The synthesis of (+)-nootkatone was recently developed and improved upon in subsequent publications (Fig. 3). Nootkatone derives its name from its original discovery in the cypress tree *Callitropsis nootkatensis* from the Nootkat region in British Columbia, Canada. It can also be found in citrus fruits like pomelo and grapefruits. A pleasant grapefruit aroma makes nootkatone popular for flavoring, but it also exhibits insect repellent and pharmacological activities such as inhibition of human CYPs. Extraction from fruits is hampered by low yields, whereas chemical synthesis mostly results in the (-)-nootkatone enantiomer with less favourable characteristics (Fraatz *et al.*, 2009). Therefore biotechnological production of (+)-nootkatone appears to be a promising alternative. Wriessnegger *et al.* (2014) realized this goal using two different strategies. *P. pastoris* expressing a plant CYP and a cytochrome P450 reductase (CPR) from *Hysocyamus muticus* and *Arabidopsis thaliana*, respectively, was employed for whole-cell catalysis of supplemented (+)-valencene to (+)-nootkatone. A yield of 80 mg/L (+)-nootkatone was achieved. However, problems with the volatility of (+)-valencene complicated this strategy. Therefore, it was decided to enable *in situ* (+)-valencene synthesis in *P. pastoris* via introduction of a valencene synthase from *C. nootkatensis*. Additionally, in the course of the whole cell catalysis experiments the involvement of a native alcohol dehydrogenase (*ADH*) in (+)-nootkatone synthesis was discovered. Subsequently, the *ADH* was overexpressed. A strain combining heterologous CPR, CYP and valencene synthase expression with overexpression of the endogenous *ADH* was capable of producing ca. 100 mg/L (+)-nootkatone. Lastly, the metabolic flux

through the MVA pathway was enhanced by expression of a hydroxy-methylglutaryl-CoA reductase from *S. cerevisiae*. This final step allowed for more than a doubling of the product titer to ca. 208 mg/L. Positive effects of increased flux through the MVA pathway on terpenoid production in *S. cerevisiae* are well documented (Asadollahi *et al.*, 2010). Wriessnegger *et al.* (2014) demonstrated the applicability of this strategy to *P. pastoris* for the first time. Yields were significantly higher than in *S. cerevisiae*, in which 4 mg/L (+)-nootkatone with added (+)-valencene (Cankar *et al.*, 2014) or 0.14 mg/L with *in situ* valencene synthesis (Gavira *et al.*, 2013) were achieved. Furthermore, unlike *S. cerevisiae* no cytotoxicity of nootkatone at concentrations > 100 mg/L was observed (Gavira *et al.*, 2013). Building upon these promising results, the *P. pastoris* (+)-nootkatone production system was further optimized in two subsequent studies. Overexpression of either the endogenous ice2 or heterologous *S. cerevisae* ice2 membrane protein was used to improve CPR stability (Emmerstorfer *et al.*, 2015). While CPR stability was improved, no beneficial effect on (+)-nootkatone production was observed. Based on transcriptomic findings, 33 genes were separately overexpressed to assay whether they benefited (+)-valencene hydroxylation by CYP (Wriessnegger *et al.*, 2016). Interestingly, it was found that overexpression of the gene *RAD52* resulted in two fold increased productivity. *RAD52* is involved in DNA repair and recombination events in *P. pastoris*, an assumption based on its similarity to the *S. cerevisiae* homologue (Symington, 2002). It is suspected that *RAD52* plays a role in relieving oxidative stress caused by overproduction of terpenoids. Although a direct effect of *RAD52* on CYP and CPR conversion rates could be disproved, no conclusive explanation was found for the increased terpenoid productivity.

The plant terpenoid dammarenediol-II from *Panax ginseng* was produced in a study by Liu *et al.* (2015). Dammarenediol-II is a cyclized form of 2,3-oxidosqualene, which in turn is formed from the assembly of two squalene molecules (Fig. 3). It can be applied for the synthesis of ginsenosides that are widely used in pharmaceutical applications (Kim *et al.*, 2015). While 2,3-oxidosqualene is a naturally occurring intermediate of the ergosterol pathway in *P. pastoris*, a dammarenediol-II synthase gene (DDS) from *P. ginseng* needed to be introduced to enable product formation. This approach resulted in only minute dammarenediol-II yields, compared to high concentrations of ergosterol. It was suspected that diverting the metabolic flux away from ergosterol towards

dammarenediol-II should increase productivity. Due to the essential role of ergosterol for cell membrane integrity, a simple knock-out of pathway genes was not feasible. Instead, the integration of additional native *ERG1* genes under control of *pAOX1* for increased 2,3-oxidosqualene synthesis was realized. The erg1 catalyzed production of 2,3-oxidosqualene is thought to be the rate-limiting step in ergosterol biosynthesis (Leber *et al.*, 2001). In addition to boosting the precursor pool, flux towards ergosterol was reduced by replacing the native promoter of *ERG7* (catalyzing the first transformation step from 2,3-oxidosqualene towards ergosterol) with the repressible promoter *pTHI11* (Delic *et al.*, 2013; Landes *et al.*, 2016). Applying *pTHI11*, ergosterol synthesis was unaffected during normal growth but could be repressed with the addition of thiamine once cells reached a desired density. To the best of our knowledge, this marks the first published use of downregulated gene expression for metabolic engineering purposes in *P. pastoris*. Compared to the initial attempt with only the heterologous DDS, the dammarenediol-II yield was increased from < 0.1 to 0.7 mg/g DCW. Further improvements by upregulating key genes of the MVA pathway are to be expected, as was demonstrated in *S. cerevisiae* (Dai *et al.*, 2013). Dammarenediol-II yields are slightly higher in *P. pastoris* than in *S. cerevisiae* with an unaltered MVA pathway. This suggests that enhancing flux through this pathway in *P. pastoris*, as carried out by Wriessnegger *et al.* (2014), could result in strains outperforming current *S. cerevisiae* top producers. Another study assayed the impact of co-localizing DDS and ERG1 as fusion proteins or via protein-protein interaction, in order to direct the flux of squalene towards dammarenediol-II (Zhao *et al.*, 2016). Co-localization was realized successfully. However, dammarenediol-II yields of 0.078 mg/g DCW were markedly below the ones achieved by Liu *et al.* (2015), highlighting the positive impact additional metabolic engineering steps had on productivity.

6.2 Isoflavonoid

Isoflavonoids are a subgroup of flavonoids that can be found in plants and fungi. Over 10,000 different compounds are presumed to exist in nature, fulfilling important roles in plant physiology and antimicrobial activity. Industrial use of flavonoids focuses on dietary supplements, since they have been attributed cognitive preservation effects. Furthermore, their application as part of a personalized medicine (precision medicine)

approach to prevent neurodegenerative diseases like Alzheimer's is currently being evaluated (Dixon and Pasinetti, 2010). The soybean isoflavonoids genistein and daidzein have been attributed anti-cancer properties, amongst other potential pharmaceutical features. It is suspected, that the high intake of soy-based food products in Asian countries contributes to the lower occurrence of hormone-related breast cancer, compared to western countries (Barnes, 2010).

Both soybean isoflavonoids have been produced in their hydroxylated form in *P. pastoris*. Regioselective hydroxylation is assumed to enhance bioactivity of the compounds, but these forms cannot be found in nature (Roh *et al.*, 2009). A membrane-bound fusion protein consisting of a *S. cerevisiae* CPR, Streptomyces avermitilis CYP and the transmembrane region of an *Aspergillus oryzae* CYP, was used for regioselective conversion of genistein and daidzein to 3'-hydroxygenistein and 6-hydroxydaidzein, respectively (Chiang *et al.*, 2016). This study expands on previous work in which the separate synthesis of both compounds was first described (Chang *et al.*, 2013; Wang *et al.*, 2015). The triple fusion protein facilitated efficient hydroxylation of genistein and daidzein, with a maximal 3'-hydroxygenistein titer of 15.0 mg/L and a 6-hydroxydaidzein titer of 7.5 mg/L. *P. pastoris* clearly outperformed other recombinant host systems, including *E. coli* and *S. cerevisiae*, in both maximal product titer and volumetric productivity. Again, the high aptitude of *P. pastoris* for expression of heterologous membrane bound proteins like CYP and CPR was an important factor towards its application as a host for metabolic engineering. It has to be noted, that production required the supplementation of the corresponding isoflavonoid and a heme precursor, and was realized in the form of whole cell catalysis. CYP utilizes heme as cofactor and therefore, alleviated heme levels benefit productivity. *In situ* isoflavonoid biosynthesis might enhance yields and could make the process more economical, since addition of external precursors would be avoided. First steps towards reconstructing the necessary biosynthetic pathway in yeast have been made, resulting in *S. cerevisiae* capable of synthesizing various soybean (iso)flavonoids, including genistein, from phenylalanine (Ralston *et al.*, 2005; Trantas *et al.*, 2009). For a fully integrated approach, the overproduction of heme via an augmented metabolic engineering strategy expanding on the results of Krainer *et al.* (2015), would be advisable.

6.3 Polyketides

Polyketides are secondary metabolites mostly produced by fungi, plants and bacteria in order to impair growth of competing organisms. Many of them are used as antibiotic agents (e.g. tetracycline and erythromycin) or as anti-cancer drug (e.g. candidaspongiolide) (Donsbach and Rück-Braun, 2005; Staunton and Weissman, 2001). Their biosynthesis starts from malonyl-CoA that can be provided e.g. by the ß-oxidation of fatty acids. Naturally polyketide producing microorganisms often exhibit low growth rates, product titers and robustness, complicating process control. This led to the exploration of heterologous hosts for producing polyketides, with the one of the all-time success stories being the synthesis of the antimalarial drug precursor artemisinic acid in *S. cerevisiae* (Ro *et al.*, 2006). Since polyketides can be of very high complexity, multiple enzymes are involved in their synthesis. Therefore, a host proven at expressing various recombinant proteins is advantageous.

Citrinin is a naturally occurring polyketide that can be found in several *Penicillium* and *Aspergillus* species. It acts as antibiotic by inhibiting sterol synthesis but also displays carcinogenic and mutational effects and is classified as mycotoxin (Ostry *et al.*, 2013). Therefore, it was chosen for production in *P. pastoris* solely as well-studied model polyketide compound with a relatively complex structure (Fig. 3). Aided by the discovery and characterization of genes involved in citrinin biosynthesis in a native producer (He and Cox, 2016), Xue *et al.* (2017) set out to engineer citrinin producing *P. pastoris* cells. Based on their earlier study focused on the synthesis of the polyketide 6-methylsalicylic acid (Gao *et al.*, 2013), a citrinin polyketide synthase and an activating phosphopantetheinyl transferase were introduced into *P. pastoris*. Fermentation of the recombinant strain led to the formation of a citrinin intermediate but not to the final product. In a stepwise engineering strategy, five heterologous genes from the native producer coding for the (proposed) citrinin biosynthetic pathway were added. In order to minimize the enzyme expression burden on *P. pastoris*, only genes essential to citrinin synthesis were integrated and e.g. associated transporter proteins omitted. This strategy enabled the production of 0.6 mg/L citrinin. Furthermore, a six step synthesis pathway starting from acetyl-CoA and malonyl-CoA condensation was proposed, based on the intermediates discovered during strain engineering steps. All genes were expressed

under control of *pAOX1*. Considering the length of the pathway and potential differences in reaction speeds, it is encouraging that citrinin synthesis was successful. Optimizing the expressional strength of selected pathway genes to avoid bottle-necks could improve citrinin yields in future experiments.

The synthesis of the bacterial pigment violacein, a polyketide with potential as both antibiotic and anticancer drug (Choi *et al.*, 2015), was recently reported (Geier *et al.*, 2015b). Five genes from the natural producer *Chromobacterium violaceum* were co-expressed and led to the formation of violacein. Although no product titers were given, the publication underlines the suitability of *P. pastoris* for polyketide synthesis. Starting with the simple polyketide 6-methylsalicylic acid ($C_8H_8O_3$), the complexity was increased to citrinin ($C_{13}H_{14}O_5$) and violacein ($C_{20}H_{13}N_3O_3$). The ease of establishing intricate heterologous pathways, even when expression levels are not optimized, suggests that further development of *P. pastoris* for polyketide synthesis is a promising prospect.

6.4 Other metabolites

Testosterone is a steroid and human sex hormone. It is widely used in medical applications and serves as platform chemical for the synthesis of other pharmacological active steroids (Fernandes *et al.*, 2003). Its biotechnological production has been hampered by side product formation and low conversion rates. Recently, the conversion of a testosterone precursor to testosterone via recombinant *P. pastoris* whole-cell catalysis was reported (Shao *et al.*, 2016). The expression of a single protein from human testis efficiently catalyzed the reaction and no byproducts were detected. However, productivity was further improved by co-expression of the *S. cerevisiae* glucose 6-phosphate dehydrogenase to enhance NADPH regeneration. This step led to a ca. 70 % increase in product titer. The beneficial effects of increased NADPH regeneration could be useful for metabolic engineering experiments, akin to the improved NADH regeneration approach discussed earlier (Geier *et al.*, 2015a).

An interesting approach for the production of bio-ethanol with *P. pastoris* was recently realized. Bio-ethanol production from lignocellulosic biomass represents a largely untapped energy source, hindered by the difficulty of fermentation with cellulose, hemicellulose or lignin as carbon source. Typically, complex pretreatment procedures

are necessary to convert these polymers into monosaccharides that can be metabolized by e.g. *S. cerevisiae* (Sánchez and Cardona, 2008). In order to reduce the need for pretreatment, a co-cultivation system of recombinant *P. pastoris* and *S. cerevisiae* was developed (Zhang *et al.*, 2017). Here, a *P. pastoris* strain expressing two types of recombinant xylanases mediated hemicellulose to xylose conversion. On the other hand, *S. cerevisiae* expressed an endo- and exoglucanase as well as a ß-glucosidase in order to hydrolyze cellulose to glucose. The adaption of *P. pastoris* to fermentation with lignocellulosic biomass via xylanase and cellobiohydrolase expression was demonstrated earlier (Mellitzer *et al.*, 2012; Shin *et al.*, 2015). However, the strategy by Zhang *et al.* (2017) combined the advantageous ethanol productivity of *S. cerevisiae* with the efficient xylose fermentation of *P. pastoris*. Using mildly pretreated (0.875 % (w/w) H_2O_2, pH 11.5, 35 °C, 1 h) wheat straw as sole carbon source, the ethanol yield was increased to ca. 33 g/L in co-cultivation. In comparison, 25 g/L and 12 g/L ethanol were produced if cultivating *S. cerevisiae* and *P. pastoris* separately, respectively. Complementation of *S. cerevisiae* high ethanol productivity but low xylose utilization with *P. pastoris* high xylose utilization and low ethanol productivity via integrated saccharification and co-fermentation highlighted the possibilities of combining yeasts for metabolite production. The strategy markedly surpassed the yields of similar attempts in which *S. cerevisiae* or *P. pastoris* had been used for simultaneous cellulose and xylose saccharification but were not co-cultivated (Shin *et al.*, 2015; Sun *et al.*, 2012).

7 Systems biology of *P. pastoris*

The goal of systems biology is to fundamentally understand a biological system in a holistic approach and apply the knowledge to formulate derived mathematical models (Bruggeman and Westerhoff, 2007; Palsson, 2006). By applying these models, desired engineering steps and their outcome can be predicted. Experiments are carried out, the data compared with the predictions and models adjusted accordingly. The systems-based approach aims to bring the computability and standardized procedures of the mechanical engineering world to biological systems. These do not only benefit basic but also applied research, cutting-down on time necessary for generating strains with desired production characteristics, providing ways to prevent negative engineering effects and enabling deeper insights into hosts for discovery of novel production

approaches. An important cornerstone of systems biology are the "omics" technologies, which enable systems-level analysis of the innards of a cell. Genomics, transcriptomics, proteomics and metabolomics are at the forefront when it comes to understanding the composition of an organism, cell or tissue and its reaction to different environmental stimuli. Understanding the physiological properties of the system and connecting them with the molecular level is another essential part of a systems biology approach (Bruggeman and Westerhoff, 2007).

7.1 Omics-based studies

Beginning in 2009, the three commonly applied *P. pastoris* strains GS115 (ATCC 20864) (De Schutter *et al.*, 2009), CBS7435 (ATCC 76273, NRRL Y-11430) (Küberl *et al.*, 2011) and DSMZ 70382 (ATCC 28485, NRRL Y-1603) (Mattanovich *et al.*, 2009) were subject to genome sequencing. These projects primarily relied on 454 next-generation sequencing technology, aided by first generation Illumina methods for the more difficult to resolve sequence parts. Although previous studies applying transcriptomic (Gasser *et al.*, 2007; Graf *et al.*, 2008), proteomic (Dragosits *et al.*, 2009) and metabolomic (Solà *et al.*, 2007) assays had been performed, the availability of genome data allowed for a much deeper understanding of cellular processes. Pre-genomic transcriptomic projects often relied either on heterologous microarrays with *S. cerevisiae* probes or on early designs for *P. pastoris* based on limited insights into the genome. Thanks to the genome sequence data, subsequent endeavors could more easily connect observations between different omics-levels. In the following years, studies involving omics-technologies were published at an increasing rate (Fig. 6). Milestones include the first application of genome-scale RNA-Seq, focused on the transcriptional response to growth on different carbon sources and improved annotation (Liang *et al.*, 2012). Methods for metabolic sampling were developed (Tredwell *et al.*, 2011) and metabolic flux analysis revealed the interplay of the central carbon metabolism (Jordà *et al.*, 2013; Srivastava *et al.*, 2012). Especially recombinant protein production associated characteristics, like the UPR pathway (Dragosits *et al.*, 2010; Hesketh *et al.*, 2013) along with production and secretion capabilities (Baumann *et al.*, 2011, 2010, Pfeffer *et al.*, 2012, 2011) were in the focus of systems level experiments. However, recent advancements might enable not

only a deeper understanding of *P. pastoris*, but also provide a fundament upon which to build a truly holistic model of *P. pastoris* cellular processes.

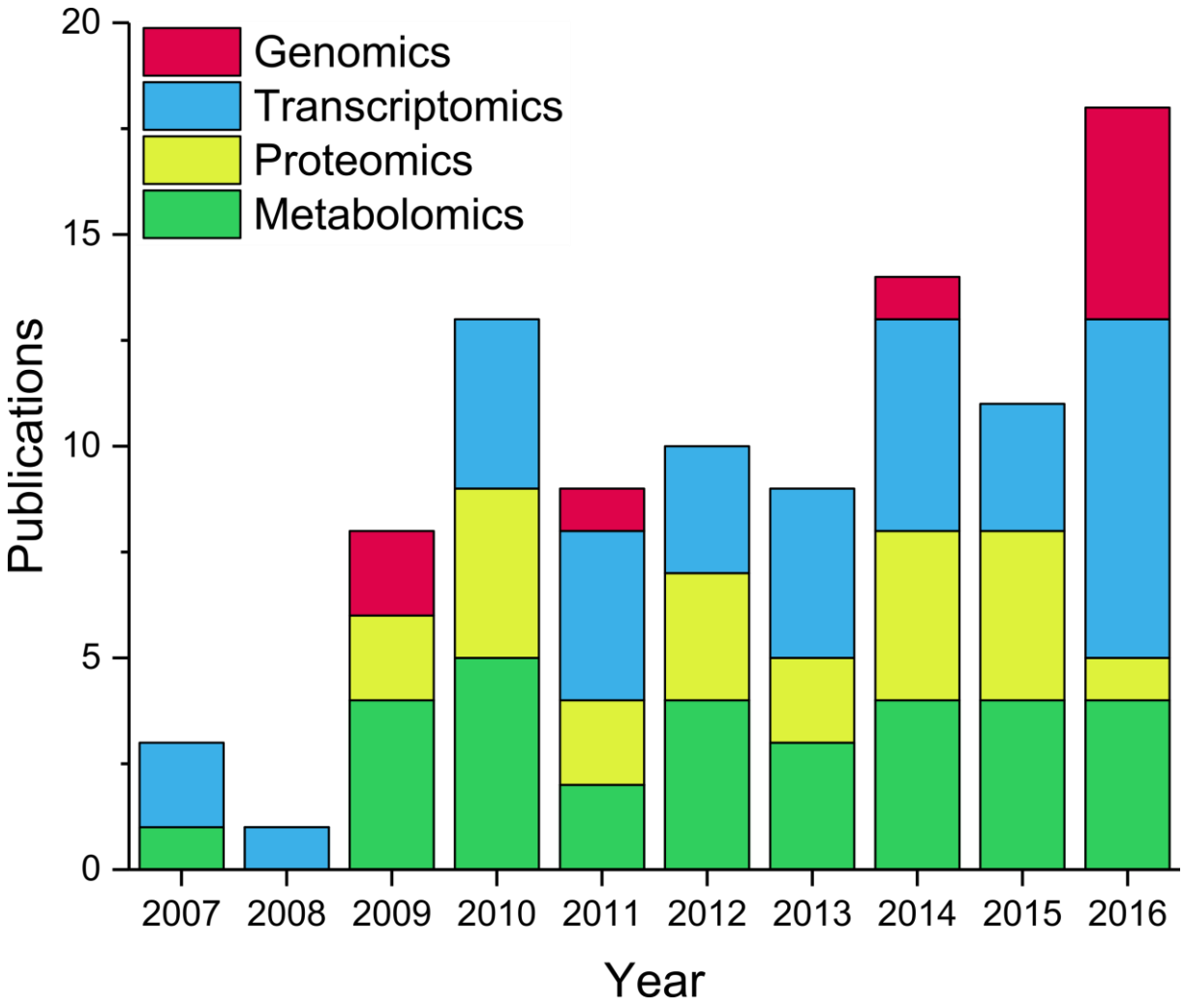

Figure 6: Omic studies in *P. pastoris* between 2007 and 2016. A first spike at 2009-2010 is visible, when the first genome sequences were published. The Second, larger, spike in 2016 is likely due to a combination of multiple studies focused on improving genome data and utilizing transcriptomic techniques.

Owed to the limits of the technology available at the time, the first *P. pastoris* genome sequencing projects were not able to resolve the entire genome. Multiple gaps (> 10) of up to 6 kb were left open. The last years have seen the rise of "next generation sequencing" (NGS) technology, facilitating higher coverage, sequencing depths and resolution of previously difficult to sequence regions like homopolymers, while costs dropped dramatically (Van Dijk *et al.*, 2014). Recently, multiple studies focused on improving the available genome data by applying the latest NGS technology (Table 4). Love *et al.* (2016) simultaneously assayed GS115, CBS7435 and DMSZ 70382 via PacBio

SMRT and Illumina HiSeq2000 systems. In this comparative approach, they were able to close all remaining gaps. The experiment was complemented by RNA-Seq analysis of cells grown in glucose, glycerol and methanol medium with four time points during a 48 h cultivation. Differences in transcriptional regulation between strains and sampling points were elucidated, leading to the discovery of multiple genes with interesting transcription patterns and strengths. They were proposed as candidates from which novel *P. pastoris* promoters can be derived. The accumulated data was also used to update the nomenclature and determine > 150 new alternative splicing variants. However, the work was focused on the comparative aspect of the three analyzed strains and the implications of the discovered differences on both genome and transcriptome level. Two additional studies published that year made it their prime objective to improve the *P. pastoris* CBS7435 genome data and provide a highly resolved, characterized and annotated basis for other scientists to use as basis for their systems level experiments. While Sturmberger *et al.* (2016) applied PacBio SMRT and Illumina HiSeq2500 NGS technology, Valli *et al.* (2016) analyzed RNA-Seq results of 37 different conditions. Both studies were complemented with available proteomic data (Renuse *et al.*, 2014; Rußmayer *et al.*, 2015a). Among their key genomic results are the closing of all remaining gaps, identification of two linear plasmids as well as resolving the telomeric and ribosomal repeats. The transcriptomic data was used to manually curate nearly all ORF (5111 of 5256), detect almost 500 new ORF (ca. 150 confirmed via proteomics data), identify ca. 5000 putative UTR and detect a large number of alternative translation start sites (ca. 700) and upstream ORF (ca. 900). Important elements involved in post-transcriptional (UTR, upstream ORF) and translational (alternative translation start sites) regulatory elements were characterized. Furthermore, the *P. pastoris* gene nomenclature was updated genome-wide using data for *S. cerevisiae* homologs and available *P. pastoris* experimental data. Genes without homologs or with uncharacterized homologs were annotated with their hypothetical function based on e.g. BLASTp results. In addition to the updated annotation, gene names were standardized according to the established three letter plus number code system. Via this systematic annotation and renaming process the second best functionally annotated yeast genome was generated.

Table 4: Recent studies concerning the advancement of *P. pastoris* genome data.

Project	Strains	Genomics	Transcriptomics	Proteomics	Key findings
(Love *et al*., 2016)	GS115, CBS7435, DSM70382	PacBio SMRT & Illumina HiSeq2000	10 conditions	-	-All gaps closed -Linear plasmid -Nomenclature updated -196 (putative) alternative splicing variants -Comparative transcriptomics
(Sturmberger *et al*., 2016)	CBS7435	PacBio SMRT & Illumina HiSeq2500	2 conditions	(Renuse *et al*., 2014)	-All gaps closed -Two linear plasmids -Telomeric region and ribosomal repeats resolved -5111 ORF manually curated -Putative chromosomal centromere
(Valli *et al*., 2016)	CBS7435	-	37 conditions	(Rußmayer *et al*., 2015a)	-Nomenclature updated -492 new ORF; 152 confirmed via proteome -4916 putative UTR -341 ORF predictions corrected -659 / 885 genes with alternative start site / uORF

In order to provide an easy access to the second generation *P. pastoris* genome data, the website www.pichiagenome.org was updated with the new genome, annotation and nomenclature data (Sturmberger *et al*., 2016; Valli *et al*., 2016). By modeling the nomenclature and annotation data on *S. cerevisiae*, the transferability of methods and results between these two yeasts will be significantly improved. Researchers can rely on a central database with comprehensive as well as exact information on genome sequence, annotation of ORFs, predicted protein domains and localizations as well as available *S. cerevisiae* homolog data. In comparison to its "bigger brother", the *S. cerevisiae* database at www.yeastgenome.org, additional features like metabolic pathway information are currently missing. But the current development of www.pichiagenome.org signals an increasing interest of its community to create a similar service.

Many studies in the last years applied different omics-technologies to interrogate *P. pastoris*. Transcriptomic analysis has been employed to elucidate the effect of growth rate on protein production and secretion, stress response, mating and maintenance energy demands (Rebnegger *et al*., 2016, 2014). Furthermore, a transcription factor with

potential benefits for recombinant protein production was characterized (Ruth *et al.*, 2014) and the tendency of *P. pastoris* to regulate its expression on the transcriptional rather than the translational level revealed (Prielhofer *et al.*, 2015). Metabolomic studies focused on the ^{13}C method, in order to analyze fluxes through the central carbon metabolism under different conditions. To this end, Jordà *et al.* (2014, 2013) applied instationary ^{13}C-metabolic flux to improve upon previous NMR data. Valuable insights into how *P. pastoris* adjusts important metabolic pathways (including glycolysis, PPP, TCA and methanol oxidation) under different conditions like recombinant protein production or changing carbon sources were obtained. A combination of metabolomic and transcriptomic methods led to the discovery of stress induced arrest of growth and recombinant protein production following methanol induction (Edwards-Jones *et al.*, 2015). The observed effect was reversible, leading to a delayed production of the target protein and temporary growth reduction. Interestingly, the results indicated no clear correlation of the phenomena to the UPR pathway, suggesting involvement of other factors like the redox state of the cell. An optimized metabolome sampling technique was developed, reducing the loss of low molecular weight compounds during cell separation (Rußmayer *et al.*, 2015b). The spectrum of measurable metabolites in ^{13}C-flux analysis was considerably expanded by the establishment of GC-Quadrupole Time-of-Flight Mass Spectrometry (GC-QTOFMS) procedures for *P. pastoris* experiments (Mairinger *et al.*, 2015).

The multi-omics study of the greatest scope so far for *P. pastoris* analyzed the methanol metabolism on transcriptomic, proteomic, and on metabolomic level (Rußmayer *et al.*, 2015a). Genome-wide transcription analysis was combined with the measurement of 575 proteins, 141 metabolite pools and 39 fluxes involved in MUT under different carbon source regiments. Supporting previous findings (Prielhofer *et al.*, 2015), a strong positive correlation between transcript abundance and protein concentration was found. A key finding of the study was the presence of all enzymes of the MUT pathway in the peroxisome. Previously, it was assumed that parts of the pathway take advantage of the cytosolic PPP for recycling of xylulose-5-phosphate (XYL5P). However, Rußmayer *et al.* (2015a) discovered isoforms of PPP enzymes containing PTS1 signal sequences that are involved in XYL5P recycling. The prevalence of peroxisomal localization of methanol induced XYL5P recycling was further corroborated by increased transcription for PTS1

containing isoforms in the presence of methanol, while those without were downregulated. Isolation of peroxisome organelles and subsequent proteome analysis validated that PTS1 isoforms were present in methanol induced cells and absent in non-induced cells. In addition to its peroxisomal localization, a novel rearrangement in the XYL5P recycling pathway was discovered. The XYL5P intermediate sedoheptulose-1,7-bisphosphate (S1,7BP), along with the associated transcript and enzyme, were found in methanol grown cells. S1,7BP is an important part of the Calvin cycle in chloroplasts (Raines *et al.*, 1999), but its presence in *S. cerevisiae* has also been shown (Clasquin *et al.*, 2011). Via S1,7BP and the aforementioned PPP enzyme isoforms, the in-organelle generation and regeneration of XYL5P from methanol is enabled. This eliminates the need for XYL5P import from the cytosol, as proposed by previous models of MUT in methylotrophic yeasts (van der Klei *et al.*, 2006). A model was postulated, equating peroxisomal compartmentalization of MUT in methylotrophic yeast to the carbon assimilation pathways of phototrophic organisms to their chloroplasts. The evidence presented supports this thesis and additional experiments with strains deficient in key peroxisomal isoforms could further cement its validity. Due to the high expression levels of proteins for MUT, the synthesis of related cofactors was also increased. Particularly the observed regulatory effects on genes involved in heme biosynthesis could serve as basis for optimizing previous attempts at increasing heme production in *P. pastoris* via metabolic engineering (Krainer *et al.*, 2015).

7.2 Genome-scale modeling

Systems biology relies upon the use of data gained from high throughput omics-technologies for the creation of genome-scale metabolic models (GEM) (Bruggeman and Westerhoff, 2007). In particular for biotechnological applications, the predictability that comes with having modeled central cellular processes is beneficial (Otero and Nielsen, 2010). Model-based approaches have been successfully used in several industrial microorganisms to purposefully predict genetic engineering targets for enhanced product yields and improved cultivation characteristics. Constraint-based reconstruction and analysis (COBRA) methods have proven their suitability for such applications (King *et al.*, 2015; Schellenberger *et al.*, 2011). Shortly after the first genome projects were published, GEM for *P. pastoris* followed. For *P. pastoris* DSMZ 70382 the

model *PpaMBEL1254* (Sohn *et al.*, 2010) and for strain GS115 the model *iPP668* (Chung *et al.*, 2010) were published. They are of similar size with ca. 1300 metabolic reactions involving ca. 1200 metabolites, segregated into 8 compartments. Two years later, a refined GS115 model (*iLC915*) introduced simultaneous computation of recombinant protein production and MUT, as well as optimized predictability of growth behavior on methanol based media (Caspeta *et al.*, 2012). However, *iLC915* omitted extracellular and nuclear reactions of *iPP668*. Notably, the gene coverage was increased from 10 % (*PpaMBEL1254*) to 17 % (*iLC915*). Compared to current *S. cerevisiae* GEM, less reactions but a higher percentage of genes are covered (Österlund *et al.*, 2013; Sánchez and Nielsen, 2015). Recently, the established *P. pastoris* GEM were further developed and also applied in metabolic engineering projects.

iLC915 was refined to include both native *P. pastoris* and humanized N-glycosylation of glycoproteins (Irani *et al.*, 2016). Integration of N-glycosylation required reconstruction of the dolichol pathway for native glycosylation. Hyper-mannosylation was set to 9 mannose residues at the lower end of the observed range (Vervecken *et al.*, 2004). The necessary pathways for synthesis of fully humanized and terminally sialylated N-glycosylation were taken from literature (Hamilton *et al.*, 2006). Subsequently, the GEM was named *ihGlycopastoris*. For validation, the GEM was evaluated against published *P. pastoris* data. Both growth rate and recombinant protein productivity were assayed under various cultivation conditions and with multiple target proteins. *ihGlycopastoris* displayed excellent consensus (R^2 = 0.92) for the growth rate, but failed at predicting recombinant protein production rates. Protein production rates were often considerably overestimated by the model. It is suspected that stress factors like UPR, secretion capacity and others (Puxbaum *et al.*, 2015) that are not considered by *ihGlycopastoris* lead to the overestimation of production rates. Despite the aforementioned challenges, the model is to be regarded as the first step towards a fully realized N-glycosylation GEM. Experimental data on proteome-wide native glycosylation in *P. pastoris*, as was done for *S. cerevisiae* (Chen *et al.*, 2014), and implementation of the effects of various stress factors on recombinant protein production rates are necessary to further develop *ihGlycopastoris*. Nevertheless, the results are a clear indication of the burden on *P. pastoris* metabolism due to glycoprotein expression. The model highlights the need

for expanding current GEM to increase their applicability and make them suitable for a wider range of tasks.

In a different approach, previously published GEM were combined and expanded upon into a new "meta"-GEM named *iMT1026* (Tomàs-Gamisans *et al.*, 2016). *iPP668*, *PpaMBEL1254* and *iLC915* were used for assembly of *iMT1026*. *ihGlycopastoris* was published later and could therefore not be considered. In addition to combining the three previous GEM, existing pathways in the models were reevaluated based on published data. For example, carbon source utilization pathways that were originally transferred from *S. cerevisiae* models, but for which the non-utilization in *P. pastoris* was proven, were removed. On the other hand, the fatty acid metabolism and oxidative phosphorylation required the addition of new reactions or condensation of existing ones. In the end, 185 new reactions were added, based on literature data. The resulting *iMT1026* exhibits the highest gene coverage (ca. 19 %) as well as number of metabolites and reactions (1689 and 2035, respectively) of any *P. pastoris* GEM published so far. The validity of its prediction capability was confirmed via comparison to published data. *iMT1026* growth rate, CO_2 production and O_2 consumption simulations were in very good agreement with experimental data (< 10 % deviation), outperforming *iPP668*, *PpaMBEL1254*, and *iLC915*. While strains producing a recombinant protein were included in the simulations, production rates or yields were not evaluated. Nonetheless, the approach of Tomàs-Gamisans *et al.* (2016) resulted in the most comprehensive *P. pastoris* GEM yet. Similar to the ongoing *S. cerevisiae* work on a collaborative GEM refinement (Sánchez and Nielsen, 2015), *iMT1026* and the methodology that led to its creation should encourage the *P. pastoris* community to cooperate on improving the GEM for its chosen microorganism.

To date only one practical example of genome engineering based on COBRA assisted prediction from a *P. pastoris* GEM has been published. As mentioned previously, Nocon *et al.* (2014) used *PpaMBEL1254* by Sohn *et al.* (2010) as GEM and applied MOMA and FSEOF algorithms for prediction of suitable knock-out or overexpression targets, respectively. A total of nine predicted targets were assayed separately and five of them led to increased titers of the target recombinant protein. Interestingly, the employed GEM is based on DMSZ 70382 while the experiments were conducted with X-33, a

reconstituted variant of the histidine-auxotroph GS115 (Ahmad *et al.*, 2014). With DSMZ 70382 having been reclassified as *K. pastoris* and GS115 as *K. phaffii*, a certain evolutionary distance can be inferred (Kurtzman, 2009). Distinct differences on both the genomic and transcriptomic levels were recently detailed (Love *et al.*, 2016). Potentially, this circumstance led to a lower accuracy of the GEM in this particular application. Nevertheless, the first implementation of GEM-based genetic engineering in *P. pastoris* was a success with product yields having been increased up to 40 % (Nocon *et al.*, 2014). It showcases the potential COBRA methods have for strain engineering in *P. pastoris*. The newly developed *iMT1026*, or future iterations of it, will enable more accurate target predictions and the combination of both *K. phaffii* and *K. pastoris* data should make it suitable for a wider range of applications.

7.3 Physiological insights

A key component to the improvement of a GEM is the inclusion of physiological data (Liu *et al.*, 2010). Understanding the physiological implications and limitations of the host organism is essential for network construction in a COBRA approach (Schellenberger *et al.*, 2011). For example, the capacity of the cellular membrane for anchoring membrane proteins has been incorporated into an *E. coli* GEM in order to increase its applicability for modeling expression of membrane bound proteins (J. K. Liu *et al.*, 2014). Similarly, physiological insights have been used to refine *P. pastoris* GEM (Tomàs-Gamisans *et al.*, 2016). Various genes involved in peroxisome biogenesis, membrane transport and central peroxisomal metabolic pathways were studied (Polupanov *et al.*, 2011; Spong and Subramani, 1993; Yan *et al.*, 2008) and its composition analyzed (Wriessnegger *et al.*, 2007). Soon after *P. pastoris* suitability for recombinant protein production was established, related cellular physiological properties were investigated. Often, the goal was to understand factors limiting protein expression, folding and secretion (reviewed in Puxbaum *et al.* (2015)). As a result, a sizeable body of knowledge about peroxisomal and protein related physiological properties of *P. pastoris* is available. Here, we highlight some of the most recent achievements that further deepen these subjects or provide insights into as of yet less-studied physiological aspects.

Much progress has been made in understanding the transcriptional regulation of *AOX1* and other MUT genes (see chapter 3.1). However, the exact mechanism of how the cell senses methanol remained unclear. Recently, two proteins homolog to the WSC (cell Wall integrity and Stress response Component) membrane protein family of *S. cerevisiae* (Verna *et al.*, 1997) were found to mediate extracellular methanol sensing in *P. pastoris* (Ohsawa *et al.*, 2017), as shown in Fig. 7 (A). A total of three WSC genes were found on the *P. pastoris* genome, *PpWSC1*, *PpWSC2* and *PpWSC3*. Different deletion strains were created and their growth behavior as well as the transcriptional activity of methanol associated genes at different concentrations analyzed. In combination with complementation experiments, it was elucidated that *PpWSC1* and *PpWSC3*, but not *PpWSC2* are involved in methanol sensing. They showed different responses to changing methanol concentration, suggesting that *PpWSC1* is responsible for methanol sensing at low concentration (0.01 to 0.05 % (w/w)) and *PpWSC3* for higher levels (0.1 to 0.5 % (w/w)). Compared to the high range ethanol sensing mechanism of *S. cerevisiae* (Stanley *et al.*, 2010) such a sensitive and two-tiered system is necessary due to the toxicity of intermediates of the methanol metabolism. *P. pastoris* must avoid the accumulation of toxic formaldehyde by fine-tuning the *AOX1* expression based on the methanol concentration. In the next step, knowledge about WSC involvement in cell wall integrity pathways (Levin, 2011) was used to identify a homolog of the signaling gene *ROM2* in *P. pastoris*. The homolog *PpROM2* was found and via a combination of knock-out and complementation studies, its involvement in methanol induced transcription and its binding to *PpWSC1*/*PpWSC2* was discovered. In consequence, a model was proposed by which *PpWCS1* and *PpWSC2* are transmembrane proteins that sense extracellular methanol and upon detection release *PpROM2* from their cytosolic tails. *PpROM2* translocates to the nucleus where it induces transcription of MUT genes. More than 30 mutated strains were constructed and protein-protein interactions assayed via protein engineering (Ohsawa *et al.*, 2017). Thereby, important insights into the mechanism of methanol sensing were revealed. These insights open up a plethora of new genetic engineering opportunities. For instance, methanol sensing sensitivity could be altered to reduce the needed methanol concentration or the mechanism could be employed to express other non-methanol related pathways.

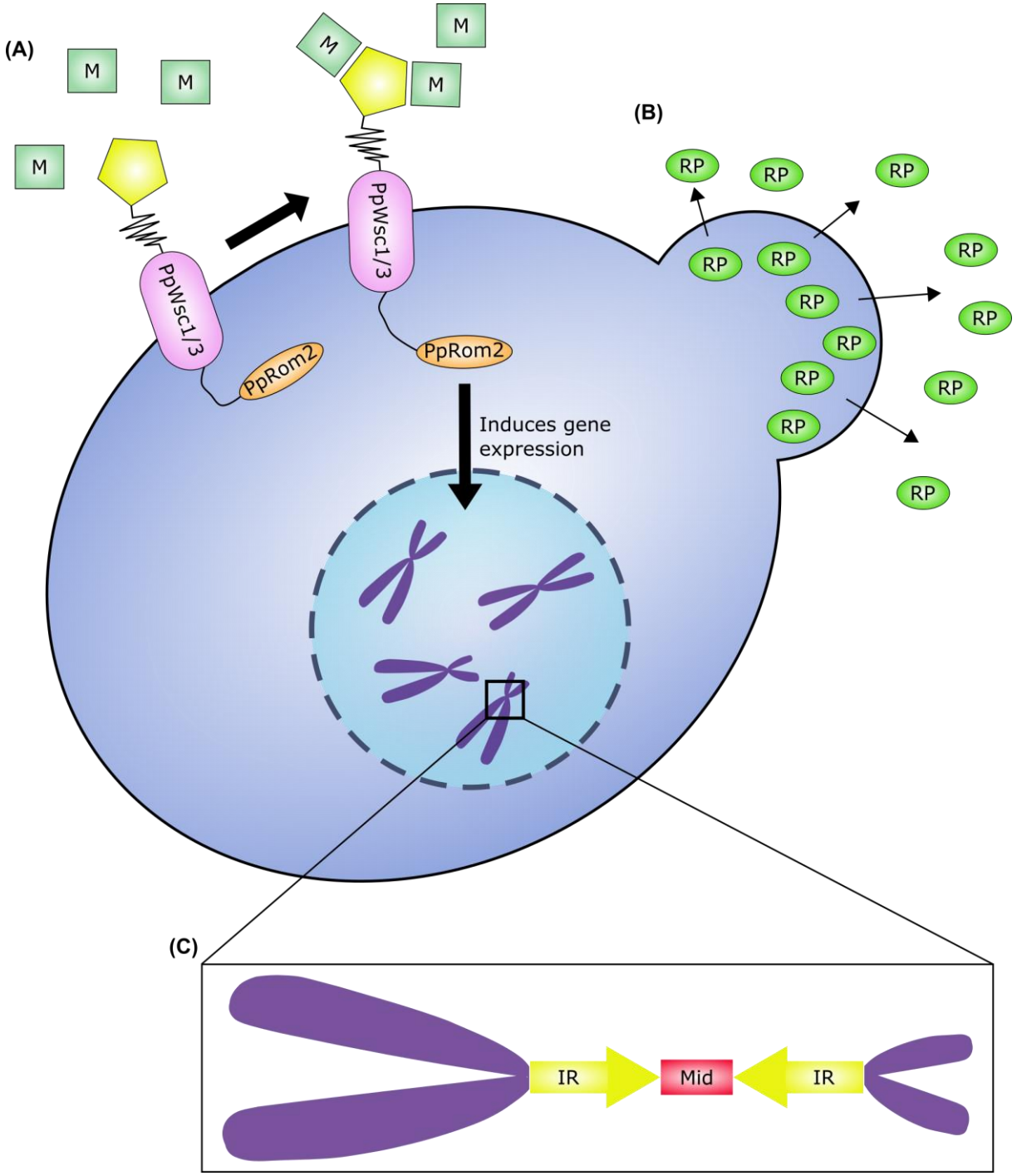

Figure 7: Cartoon of three recent physiological studies. **(A)** Methanol sensing complex. *PpWsc1* and *PpWsc3* sense extracellular methanol. The attached *PpRom2* is suspected to translocate to the nucleus and activate methanol inducible gene expression. **(B)** Recombinant proteins (RP) are almost exclusively secreted at the bud tip. **(C)** The centromere structure was fully elucidated. It consists of a middle section (Mid) flanked by two inverted repeats (IR).

Although *P. pastoris* has an efficient native secretion apparatus, further optimization is possible. Puxbaum *et al.* (2016) investigated the secretion behavior of different

recombinant proteins in regard to their cellular localization. It was found, that proteins transition from the cytosol to the medium predominately at the bud surface (Fig. 7 (B)). Increased activity of the cell wall biogenesis and transport of involved proteins to the growing bud in *S. cerevisiae* has been reported before (Levin, 2011). However, so far no data was available regarding the cellular localization of exocytosis in budding yeasts. The secretory pathway of recombinant fluorescent fusion proteins was monitored in live cells via 4D fluorescence microscopy. During cultivation, cells were placed every few minutes on a confocal microscope and a z stack covering the entire cell was recorded. This approach allowed for the spatially and temporally resolved intracellular localization of the fusion proteins. As cultivation time increased, an increasing transition of fusion protein from the ER of the mother cell to the bud tip was visible. Interestingly, as soon as the ER in the bud was fully inherited from the mother cell, the fusion proteins were not transported to the bud surface anymore. While useful, the constant loss of signal due to secretion of fusion protein into the medium complicated this method. In the next step, recombinant and endogenously secreted proteins were fused to a membrane anchor domain to hold it in place. Target proteins were visualized via immunostaining. Again, 4D fluorescence microscopy was applied. The results clearly show, that both heterologous and native proteins were first secreted at the bud tip and later at the entire bud surface. It cannot be excluded that the procedures necessary for 4D fluorescence microscopy had an impact on cellular physiology, but the evidently high protein trafficking activity at the yeast bud (Levin, 2011) supports the findings of Puxbaum *et al.* (2016). Similarly, it has been found that in filamentous fungi exocytosis occurs predominantly at the growing hyphae tip (Shoji *et al.*, 2014). Nevertheless, the present study elucidates the physiological properties of protein secretion in *P. pastoris*, and potentially all budding yeasts. It should therefore prove valuable for devising strategies to increase the secretion capacity by increasing the ratio of budding cells. These findings would also explain the improved secretion features of *P. pastoris* cells genetically engineered to exhibit prolonged G2 and M phases of cell division (Buchetics *et al.*, 2011).

A three-step signal peptide was recently discovered (Heiss *et al.*, 2015). Pursuing earlier studies on epx1 (Heiss *et al.*, 2013), the endogenous protein of highest abundance in the supernatant, its secretion signal was characterized. It was found that it contains a bipartite prosequence, requiring two endoprotease steps before the final signal sequence

is generated. Interestingly, the cleavage site inside the bipartite prosequence could not be found in *EPX1* homologous of other yeasts. Modifications of the site lead to misstargeting of epx1 in some cases. Besides its physiological relevance, truncated versions of the epx1 secretion signal were also shown to facilitate efficient secretion of recombinant proteins. No detailed secretion rates were published so far, but initial results were promising and could lead to the development of an alternative to the *S. cerevisiae* α-mating factor secretion signal. In contrast, computational analysis and *in vivo* validation of various endogenous secretion signals yielded no improved secretion in comparison to the α-mating factor secretion signal (Massahi and Çalik, 2016). However, the epx1 secretion signal was not included in these studies.

Centromeres play an important role during cell division as the anchor point between cytoskeleton and chromosome. In particular the centromere of *S. cerevisiae* has been well characterized and used as model to better understand centromeres of higher eukaryotes like plants and humans (Blackburn and Szostak, 1984). Until recently, the centromere structure and position of *P. pastoris* was not fully elucidated. Applying chromatin immunoprecipitation (ChIP) sequencing, Coughlan *et al*. (2016) revealed the centromere organization (Fig. 7 (C)). First, non-transcribed regions of the chromosomes were identified in available transcriptome data (Liang *et al*., 2012), reasoning that they should indicate the location of the centromeres. On each chromosome a 5 - 7 kb long non-transcribed region consisting of two ca. 2 kb inverted repeats (IR) separated by a ca. 1 kb non-repetitive middle section were found. Via a ChIP-Seq targeting the yeast centromeric histone 4, the proposed regions were confirmed to be the centromeres of *P. pastoris*. Their position on the chromosomes was in agreement with replication start data (Liachko *et al*., 2014). This setup differs markedly from the small point centromere of *S. cerevisiae* and is more similar to the organization found in *Candida tropicalis* and *Schizosaccharomyces pombe* (Coughlan *et al*., 2016). *P. pastoris* centromeres displayed only low similarity between chromosomes, making them more accessible to genetic manipulation. Centromeres can be applied for plasmid propagation in yeast and due to their low copy number are often used in conjunction with toxic targets (Rose *et al*., 1987). Engineering of synthetic chromosomes in *P. pastoris* will require further research, but the elucidation of the centromere organization is an important first step towards this application.

Compared to *S. cerevisiae*, exact information on biosynthetic pathways is less developed in non-conventional yeasts. To fill this knowledge gap, it is often assumed that *S. cerevisiae* pathways are conserved in other yeasts. In a computational approach, Förster *et al*. (2014) investigated the amino acid biosynthesis pathways of eight non-conventional yeasts, including *P. pastoris*. All analyzed yeasts had a more or less different enzymatic apparatus for amino acid biosynthesis, compared to *S. cerevisiae*. Isoenzymes were missing, different amount of gene copies were present and for some enzymes different intracellular localizations were predicted. The *in silico* approach was complemented by an experimental assessment of differences between *S. cerevisiae* and *P. pastoris* L-leucine biosynthesis pathways. Thereby, two steps that occur predominantly in the mitochondrion of *S. cerevisiae* (Kohlhaw, 2003) could be located exclusively to the cytosol of *P. pastoris*. The results highlight the necessity to use inferred conservation of pathways from *S. cerevisiae* cautiously. Future *P. pastoris* GEM should implement these updated biosynthetic pathways.

8. Conclusion and future perspective

The key drawbacks of *P. pastoris*, clonal variability, lack of vectors and incomplete genome data, were significantly improved upon (Dikicioglu *et al*., 2014; Kelwick *et al*., 2014). Simultaneously, existing strengths like protein production were refined and as of yet untapped potential for metabolite synthesis revealed. The concerted scientific output since 2014 considerably accelerated the process of establishing *P. pastoris* as a microorganism suitable for systems-based metabolic engineering. In multiple studies the successful production of value-added metabolites from different classes was demonstrated. Concurrently, novel genetic engineering projects enabled new approaches for expression of glycoproteins, improved protein production characteristics and engineered traditional promoters to exhibit drastically new regulatory properties. The ability to fine-tune expression induction and strength has been distinctly augmented. Methanol-free *pAOX1* variants, repressible, fully synthetic and bidirectional promoters were introduced, exhibiting a wider range of expressional strengths. Clonal variability was systematically investigated and many causes for aberrant productivity and morphology were discovered. The elucidation of the underlying events and mechanisms in play will allow researchers to device strategies that prevent detrimental

integration events from occurring. Thereby, the tedious screening procedures associated with *P. pastoris* might be significantly shortened. Popular genetic manipulation techniques like CRISPR and Golden Gate Assembly were successfully transferred, enabling the application of the newest synthetic biology methods. In addition, the novel concept of expressing pathways in a polycistronic way via 2A sequences could open up new avenues of metabolic engineering. The repertoire of available plasmids was markedly expanded with both integrative and episomal vectors. ARS based vectors displayed favourable productivity characteristics and a markedly reduced clonal variability. Applying NGS technology, the second generation of *P. pastoris* genome sequences was reported. Their high resolution, comprehensive functional annotation and unified nomenclature make them the ideal basis for future systems biology research. In particular transcriptomic and metabolomic analysis was used to gain further insights into cellular behavior under various environmental conditions. Also physiological discoveries were made, e.g. the first description of the methanol sensing apparatus. These physiological and biochemical results can be combined to form the next generation of *P. pastoris* GEM. Already, a renewed interest of refining existing GEM is discernible. Unfortunately, these efforts predated the publication of the second generation genome data.

Many promising optimization targets and potential products are available, but the compartmentalization of metabolite synthesis pathways to the peroxisome appears to be especially auspicious. The *P. pastoris* peroxisome has served as model peroxisome since many years and is therefore very well-characterized. In concert with the recent trend in yeast biotechnology to utilize peroxisomes for pathway compartmentalization (DeLoache *et al.*, 2016; Shabbir Hussain *et al.*, 2016), *P. pastoris* appears to be the ideal candidate to apply this strategy. As shown in Fig. 3, valuable metabolites like terpenoids and polyketides can be derived from the ß-oxidation pathway in the peroxisome, and in some cases already have been. Especially terpenoids have gained much interest, due to the wide spectrum of applications they can be used for. By implementing tried model-based strategies from *S. cerevisiae*, terpenoid productivity in *P. pastoris* might be increased (Gruchattka *et al.*, 2013; Gruchattka and Kayser, 2015). Already, it has been shown that for certain metabolites *P. pastoris* outperforms *S. cerevisiae* (Wriessnegger *et al.*, 2014).

On the other hand, the well-established and characterized suitability of *P. pastoris* for expression of recombinant proteins can serve as the corner stone, upon which heterologous metabolic pathways can be implemented. Membrane proteins and glycoproteins that often are problematic in other heterologous host have been proven to be suited for high level active expression in *P. pastoris* (Byrne, 2015; Öberg *et al.*, 2011). The large classes of CYP and CPR enzymes are often membrane anchored and are involved in a multitude of reactions that can be used for the production of valuable compounds. Published results indicate this to be a key factor in enabling higher yields in *P. pastoris* metabolic engineering for specific targets, compared to *E. coli* and *S. cerevisiae* (Chiang *et al.*, 2016; Geier *et al.*, 2012).

To fully realize systems-based metabolic engineering then, these discoveries need to be combined and expanded upon. For example, the new genomic, transcriptomic, metabolomic and physiological data should form the fundament upon which to create the newest generation of *P. pastoris* GEM. It has been shown that great potential lies in expanding GEM, facilitating more accurate predictions and the use of next-generation COBRA methods (King *et al.*, 2015; Lee *et al.*, 2012). The first reported model-based genetic engineering project in *P. pastoris* proved the potential this technology holds (Nocon *et al.*, 2014). And while the updated version of www.pichiagenome.org has to be commended, there is still room for further developments. However, the inclusion of the new genome data will facilitate a stream lined transfer of knowledge between *P. pastoris* and *S. cerevisiae*. The *P. pastoris* community can now more easily benefit from the vast knowledge gathered on all systems biology levels of *S. cerevisiae*. This progress will need to go hand-in-hand with integrating new and improved genetic manipulation and synthetic biology techniques. Recent publications have provided the basis for rapid, modular and *in vivo* vector assembly, as well as shown more accurate ways of manipulating the genome. Via genome-sequencing or genotyping a large library of clones, further insights into random integrations might be obtained (Kegel *et al.*, 2006; Zhang *et al.*, 2014).

Groundbreaking studies, like point-of-care drug synthesis in a microfluidic reactor or detection of single-cell protein efflux, demonstrate the growing interest of scientists in this non-conventional yeast (Landry *et al.*, 2017; Perez-Pinera *et al.*, 2016). It therefore

appears that the future of *P. pastoris* is bright and full of possibilities, but it will require the concerted effort of its scientific community to realize this potential.

Acknowledgments

Grants from the Federal State of North Rhine-Westphalia for the CLIB-Graduate Cluster Industrial Biotechnology are gratefully acknowledged.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 685778.

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 Business Agency Vienna through the COMET-Funding Program managed by the Austrian Research Promotion Agency FFG.

References

Abad, S., Kitz, K., Hörmann, A., Schreiner, U., Hartner, F.S., Glieder, A., 2010a. Real-time PCR-based determination of gene copy numbers in *Pichia pastoris* . Biotechnol. J. 5, 413–420. doi:10.1002/biot.200900233

Abad, S., Nahalka, J., Bergler, G., Arnold, S.A., Speight, R., Fotheringham, I., Nidetzky, B., Glieder, A., 2010b. Stepwise engineering of a *Pichia pastoris* D-amino acid oxidase whole cell catalyst. Microb. Cell Fact. 9, 24. doi:10.1186/1475-2859-9-24

Ahmad, M., Hirz, M., Pichler, H., Schwab, H., 2014. Protein expression in *Pichia pastoris* : recent achievements and perspectives for heterologous protein production. Appl. Microbiol. Biotechnol. 98, 5301–5317. doi:10.1007/s00253-014-5732-5

Antonenkov, V.D., Hiltunen, J.K., 2012. Transfer of metabolites across the peroxisomal membrane. Biochim. Biophys. Acta - Mol. Basis Dis. 1822, 1374–1386. doi:10.1016/j.bbadis.2011.12.011

Araya-Garay, J.M., Ageitos, J.M., Vallejo, J. a, Veiga-Crespo, P., Sánchez-Pérez, A., Villa, T.G., 2012a. Construction of a novel *Pichia pastoris* strain for production of xanthophylls. AMB Express 2, 24. doi:10.1186/2191-0855-2-24

Araya-Garay, J.M., Feijoo-Siota, L., Rosa-Dos-Santos, F., Veiga-Crespo, P., Villa, T.G., 2012b. Construction of new *Pichia pastoris* X-33 strains for production of lycopene and β-carotene. Appl. Microbiol. Biotechnol. 93, 2483–2492. doi:10.1007/s00253-011-3764-7

Asadollahi, M.A., Maury, J., Schalk, M., Clark, A., Nielsen, J., 2010. Enhancement of farnesyl diphosphate pool as direct precursor of sesquiterpenes through metabolic engineering of the mevalonate pathway in *Saccharomyces cerevisiae*. Biotechnol. Bioeng. 106, 86–96. doi:10.1002/bit.22668

Avalos, J.L., Fink, G.R., Stephanopoulos, G., 2013. Compartmentalization of metabolic pathways in yeast mitochondria improves the production of branched-chain alcohols. Nat. Biotechnol. 31, 335–341. doi:10.1038/nbt.2509

Aw, R., Polizzi, K.M., 2013. Can too many copies spoil the broth? Microb. Cell Fact. 12, 128. doi:10.1186/1475-2859-12-128

Aw, R., Polizzi, K.M., 2016. Liquid PTVA: a faster and cheaper alternative for generating multi-copy clones in *Pichia pastoris*. Microb. Cell Fact. 15, 29. doi:10.1186/s12934-016-0432-8

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., Mori, H., 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol. 2, 2006.0008. doi:10.1038/msb4100050

Barnes, S., 2010. The biochemistry, chemistry and physiology of the isoflavones in soybeans and their food products. Lymphat. Res. Biol. 8, 89–98. doi:10.1089/lrb.2009.0030

Baumann, K., Adelantado, N., Lang, C., Mattanovich, D., Ferrer, P., 2011. Protein trafficking, ergosterol biosynthesis and membrane physics impact recombinant protein secretion in *Pichia pastoris*. Microb. Cell Fact. 10, 93. doi:10.1186/1475-2859-10-93

Baumann, K., Carnicer, M., Dragosits, M., Graf, A.B., Stadlmann, J., Jouhten, P., Maaheimo, H., Gasser, B., Albiol, J., Mattanovich, D., Ferrer, P., 2010. A multi-level study of recombinant *Pichia pastoris* in different oxygen conditions. BMC Syst. Biol. 4, 141. doi:10.1186/1752-0509-4-141

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. doi:10.1021/ja809461u

Belhaj, K., Chaparro-Garcia, A., Kamoun, S., Patron, N.J., Nekrasov, V., 2015. Editing plant genomes with CRISPR/Cas9. Curr. Opin. Biotechnol. 32, 76–84. doi:10.1016/j.copbio.2014.11.007

Bhataya, A., Schmidt-Dannert, C., Lee, P.C., 2009. Metabolic engineering of *Pichia pastoris* X-33 for lycopene production. Process Biochem. 44, 1095–1102. doi:10.1016/j.procbio.2009.05.012

Bill, R.M., 2014. Playing catch-up with *Escherichia coli*: using yeast to increase success rates in recombinant protein production experiments. Front. Microbiol. 5, 85. doi:10.3389/fmicb.2014.00085

Blackburn, E.H., Szostak, J.W., 1984. The molecular structure of centromeres and telomeres. Annu. Rev. Biochem. 53, 163–194. doi:10.1146/annurev.bi.53.070184.001115

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, 757–766. doi:10.1093/glycob/cwh104

Bruggeman, F.J., Westerhoff, H. V., 2007. The nature of systems biology. Trends Microbiol. 15, 45–50. doi:10.1016/j.tim.2006.11.003

Buchetics, M., Dragosits, M., Maurer, M., Rebnegger, C., Porro, D., Sauer, M., Gasser, B., Mattanovich, D., 2011. Reverse engineering of protein secretion by uncoupling of cell cycle phases from growth. Biotechnol. Bioeng. 108, 2403–2412. doi:10.1002/bit.23198

Byrne, B., 2015. *Pichia pastoris* as an expression host for membrane protein structural biology. Curr. Opin. Struct. Biol. 32, 9–17. doi:10.1016/j.sbi.2015.01.005

Cámara, E., Albiol, J., Ferrer, P., 2016. Droplet digital PCR-aided screening and characterization of *Pichia pastoris* multiple gene copy strains. Biotechnol. Bioeng. 113, 1542–1551. doi:10.1002/bit.25916

Camattari, A., Goh, A., Yip, L.Y., Tan, A.H.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 Fact. 15, 139. doi:10.1186/s12934-016-0540-5

Cankar, K., Van Houwelingen, A., Goedbloed, M., Renirie, R., De Jong, R.M., Bouwmeester, H., Bosch, D., Sonke, T., Beekwilder, J., 2014. Valencene oxidase CYP706M1 from Alaska cedar (*Callitropsis nootkatensis*). FEBS Lett. 588, 1001–1007. doi:10.1016/j.febslet.2014.01.061

Carnicer, M., Ten Pierick, A., Van Dam, J., Heijnen, J.J., Albiol, J., Van Gulik, W., Ferrer, P., 2012. Quantitative metabolomics analysis of amino acid metabolism in recombinant *Pichia pastoris* under different oxygen availability conditions. Microb. Cell Fact. 11, 1. doi:10.1186/1475-2859-11-83

Casini, A., Storch, M., Baldwin, G.S., Ellis, T., 2015. Bricks and blueprints: methods and standards for DNA assembly. Nat. Rev. Mol. Cell Biol. 16, 568–576. doi:10.1038/nrm4014

Caspeta, L., Shoaie, S., Agren, R., Nookaew, I., Nielsen, J., 2012. Genome-scale metabolic reconstructions of *Pichia stipitis* and *Pichia pastoris* and *in silico* evaluation of their potentials. BMC Syst. Biol. 6, 24. doi:10.1186/1752-0509-6-24

Cereghino, J.L., Cregg, J.M., 2000. Heterologous protein expression in the methylotrophic yeast *Pichia pastoris*. FEMS Microbiol. Rev. 24, 45–66. doi:http://dx.doi.org/10.1111/j.1574-6976.2000.tb00532.x

Chang, T.S., Chao, S.Y., Chen, Y.C., 2013. Production of ortho-hydroxydaidzein derivatives by a recombinant strain of *Pichia pastoris* harboring a cytochrome P450 fusion gene. Process Biochem. 48, 426–429. doi:10.1016/j.procbio.2013.02.014

Chen, W., Smeekens, J.M., Wu, R., 2014. Comprehensive Analysis of Protein N-Glycosylation Sites by Combining Chemical Deglycosylation with LC–MS. J. Proteome Res. 13, 1466–1473. doi:10.1021/pr401000c

Cheng, H., Lv, J., Wang, H., Wang, B., Li, Z., Deng, Z., 2014. Genetically engineered *Pichia pastoris* yeast for conversion of glucose to xylitol by a single-fermentation process. Appl. Microbiol. Biotechnol. 98, 3539–3552. doi:10.1007/s00253-013-5501-x

Chiang, C.M., Ding, H.Y., Lu, J.Y., Chang, T.S., 2016. Biotransformation of isoflavones daidzein and genistein by recombinant *Pichia pastoris* expressing membrane-anchoring and reductase fusion chimeric CYP105D7. J. Taiwan Inst. Chem. Eng. 60, 26–31. doi:10.1016/j.jtice.2015.10.015

Choi, B., Bobrowicz, P., Davidson, R.C., Hamilton, S.R., Kung, D.H., Li, H., Miele, R.G., Nett, J.H., Wildt, S., Tillman, U., 2003. Use of Combinatorial Genetic Libraries to Humanize N-Linked Glycosylation in the Yeast *Pichia pastoris*. Proc. Natl. Acad. Sci. U. S. A. 100, 5022–5027. doi:10.1073/pnas.0931263100

Choi, H.S., Lee, S.Y., Kim, T.Y., Woo, H.M., 2010. *In silico* identification of gene amplification targets for improvement of lycopene production. Appl. Environ. Microbiol. 76, 3097–3105. doi:10.1128/AEM.00115-10

Choi, S.Y., Yoon, K.H., Lee, J. Il, Mitchell, R.J., 2015. Violacein: Properties and production of a versatile bacterial pigment. Biomed Res. Int. 2015. doi:10.1155/2015/465056

Chung, B.K.S., Selvarasu, S., Camattari, A., Ryu, J., Lee, H., Ahn, J., Lee, H., Lee, D., 2010. Genome-scale metabolic reconstruction and *in silico* analysis of methylotrophic yeast *Pichia pastoris* for strain improvement. Microb. Cell Fact. 9.

Ciofalo, V., Barton, N., Kreps, J., Coats, I., Shanahan, D., 2006. Safety evaluation of a lipase enzyme preparation, expressed in *Pichia pastoris* , intended for use in the degumming of edible vegetable oil. Regul. Toxicol. Pharmacol. 45, 1–8. doi:10.1016/j.yrtph.2006.02.001

Claes, K., Vandewalle, K., Laukens, B., Laeremans, T., Vosters, O., Langer, I., Parmentier, M., Steyaert, J., Callewaert, N., 2016. Modular Integrated Secretory System Engineering in *Pichia pastoris* to Enhance G-Protein Coupled Receptor Expression. ACS Synth. Biol. 5, 1070–1075. doi:10.1021/acssynbio.6b00032

Clare, J., Rayment, F., Sreekrishna, K., Romanos, M.A., 1991. High-level expression of tetanus toxin fragment C in *Pichia pastoris* strains containing multiple tandem integrations of the gene. Nat. Biotechnol. 9, 455–460. doi:10.1038/nbt0591-455

Clasquin, M.F., Melamud, E., Singer, A., Gooding, J.R., Xu, X., Dong, A., Cui, H., Campagna, S.R., Savchenko, A., Yakunin, A.F., Rabinowitz, J.D., Caudy, A.A., 2011. Riboneogenesis in yeast. Cell 145, 969–980. doi:10.1016/j.cell.2011.05.022

Corchero, J.L., Gasser, B., Resina, D., Smith, W., Parrilli, E., Vázquez, F., Abasolo, I., Giuliani, M., Jäntti, J., Ferrer, P., Saloheimo, M., Mattanovich, D., Schwartz, S., Tutino, M.L., Villaverde, A., 2013. Unconventional microbial systems for the cost-efficient production of high-quality protein therapeutics. Biotechnol. Adv. 31, 140–153. doi:10.1016/j.biotechadv.2012.09.001

Couderc, R., Baratti, J., 1980. Oxidation of methanol by the yeast, *Pichia pastoris* : Purification and properties of the alcohol oxidase. Agric. Biol. Chem. 44, 2279–2289.

Coughlan, A.Y., Hanson, S.J., Byrne, K.P., Wolfe, K.H., 2016. Centromeres of the Yeast *Komagataella phaffii* (*Pichia pastoris*) Have a Simple Inverted-Repeat Structure. Genome Biol. Evol. 8, 2482–2492. doi:10.1093/gbe/evw178

Crampton, N., Bonass, W.A., Kirkham, J., Rivetti, C., Thomson, N.H., 2006. Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy. Nucleic Acids Res. 34, 5416–5425. doi:10.1093/nar/gkl668

Cregg, J., Barringer, K., Hessler, A., Madden, K., 1985. *Pichia pastoris* as a host system for transformations. Mol. Cell. Biol. 5, 3376–3385. doi:10.1128/MCB.5.12.3376

Curran, K.A., Karim, A.S., Gupta, A., Alper, H.S., 2013. Use of expression-enhancing terminators in *Saccharomyces cerevisiae* to increase mRNA half-life and improve gene expression control for metabolic engineering applications. Metab. Eng. 19, 88–97. doi:10.1016/j.ymben.2013.07.001

Dai, Z., Liu, Y., Zhang, X., Shi, M., Wang, B., Wang, D., Huang, L., Zhang, X., 2013. Metabolic engineering of *Saccharomyces cerevisiae* for production of ginsenosides. Metab. Eng. 20, 146–156. doi:10.1016/j.ymben.2013.10.004

Daley, J.M., Palmbos, P.L., Wu, D., Wilson, T.E., 2005. Nonhomologous end joining in yeast. Annu. Rev. Genet. 39, 431–451. doi:10.1146/annurev.genet.39.073003.113340

Davidson, R.C., Nett, J.H., Renfer, E., Li, H., Stadheim, T.A., Miller, B.J., Miele, R.G., Hamilton, S.R., Choi, B.K., Mitchell, T.I., Wildt, S., 2004. Functional analysis of the *ALG3* gene encoding the Dol-P-Man: Man5GlcNAc2-PP-Dol mannosyltransferase enzyme of *P. pastoris*. Glycobiology 14, 399–407. doi:10.1093/glycob/cwh023

Day, C.D., Lee, E., Kobayashi, J., Holappa, L.D., Albert, H., Ow, D.W., 2000. Transgene integration into the same chromosome location can produce alleles that express at a predictable level, or alleles that are differentially silenced. Genes Dev. 14, 2869–2880. doi:10.1101/gad.849600

De Felipe, P., Luke, G.A., Hughes, L.E., Gani, D., Halpin, C., Ryan, M.D., 2006. *E unum pluribus*: Multiple proteins from a self-processing polyprotein. Trends Biotechnol. 24, 68–75. doi:10.1016/j.tibtech.2005.12.006

de Lima, P.B.A., Mulder, K.C.L., Melo, N.T.M., Carvalho, L.S., Menino, G.S., Mulinari, E., de Castro, V.H., Dos Reis, T.F., Goldman, G.H., Magalhães, B.S., Parachin, N.S., 2016. Novel homologous lactate transporter improves L-lactic acid production from glycerol

in recombinant strains of *Pichia pastoris*. Microb. Cell Fact. 15, 158. doi:10.1186/s12934-016-0557-9

Dean, N., 1999. Asparagine-linked glycosylation in the yeast Golgi. Biochim. Biophys. Acta - Gen. Subj. 1426, 309–322. doi:10.1016/S0304-4165(98)00132-9

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

Delic, M., Mattanovich, D., Gasser, B., 2013. Repressible promoters - a novel tool to generate conditional mutants in *Pichia pastoris*. Microb. Cell Fact. 12, 6. doi:10.1186/1475-2859-12-6

Delic, M., Rebnegger, C., Wanka, F., Puxbaum, V., Haberhauer-Troyer, C., Hann, S., Köllensperger, G., Mattanovich, D., Gasser, B., 2012. Oxidative protein folding and unfolded protein response elicit differing redox regulation in endoplasmic reticulum and cytosol of yeast. Free Radic. Biol. Med. 52, 2000–2012. doi:10.1016/j.freeradbiomed.2012.02.048

DeLoache, W.C., Russ, Z.N., Dueber, J.E., 2016. Towards repurposing the yeast peroxisome for compartmentalizing heterologous metabolic pathways. Nat. Commun. 7, 11152. doi:10.1038/ncomms11152

DiCarlo, J.E., Norville, J.E., Mali, P., Rios, X., Aach, J., Church, G.M., 2013. Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. Nucleic Acids Res. 41, 4336–4343. doi:10.1093/nar/gkt135

Dikicioglu, D., Wood, V., Rutherford, K.M., McDowall, M.D., Oliver, S.G., 2014. Improving functional annotation for industrial microbes: a case study with *Pichia pastoris* . Trends Biotechnol. 1–4. doi:10.1016/j.tibtech.2014.05.003

Dixon, R.A., Pasinetti, G.M., 2010. Flavonoids and Isoflavonoids: From Plant Biology to Agriculture and Neuroscience. Plant Physiol. 154, 453–457. doi:10.1104/pp.110.161430

Donsbach, K., Rück-Braun, K., 2005. Combinatorial biosynthesis of polyketides. Nat. Rev. Mircobiology 3, 925–936. doi:10.1002/9783527619979.ch46

Douin, V., Bornes, S., Creancier, L., Rochaix, P., Favre, G., Prats, A.-C., Couderc, B., 2004. Use and comparison of different internal ribosomal entry sites (IRES) in tricistronic retroviral vectors. BMC Biotechnol. 4, 16. doi:10.1186/1472-6750-4-16

Dragosits, M., Stadlmann, J., Albiol, J., Baumann, K., Maurer, M., Gasser, B., Sauer, M., Altmann, F., Ferrer, P., Mattanovich, D., Cerdanyola, B., 2009. The Effect of Temperature on the Proteome of Recombinant *Pichia pastoris*. J. Proteome Res. 8, 1380–1392.

Dragosits, M., Stadlmann, J., Graf, A., Gasser, B., Maurer, M., Sauer, M., Kreil, D.P., Altmann, F., Mattanovich, D., 2010. The response to unfolded protein is involved in osmotolerance of *Pichia pastoris* 1–16.

Du, M., Battles, M.B., Nett, J.H., 2012. A color-based stable multi-copy integrant selection system for *Pichia pastoris* using the attenuated *ADE1* and *ADE2* genes as auxotrophic markers. Bioeng. Bugs 3, 32–37. doi:10.4161/bbug.3.1.17936

Edwards-Jones, B., Aw, R., Barton, G.R., Tredwell, G.D., Bundy, J.G., Leak, D.J., 2015. Translational Arrest Due to Cytoplasmic Redox Stress Delays Adaptation to Growth on Methanol and Heterologous Protein Expression in a Typical Fed-Batch Culture of *Pichia pastoris*. PLoS One 10, e0119637. doi:10.1371/journal.pone.0119637

Eisenberg, A., Seip, J.E., Gavagan, J.E., Payne, M.S., Anton, D.L., Dicosimo, R., 1997. Pyruvic acid production using methylotrophic as catalyst yeast transformants. J. Mol. Catal. 2, 223–232. doi:http://dx.doi.org/10.1016/S1381-1177(96)00021-5

Emmerstorfer, A., Wimmer-Teubenbacher, M., Wriessnegger, T., Leitner, E., Müller, M., Kaluzna, I., Schürmann, M., Mink, D., Zellnig, G., Schwab, H., Pichler, H., 2015. Over-expression of *ICE2* stabilizes cytochrome P450 reductase in *Saccharomyces cerevisiae* and *Pichia pastoris*. Biotechnol. J. 10, 623–635. doi:10.1002/biot.201400780

Essani, K., Glieder, A., Geier, M., 2015. Combinatorial pathway assembly in yeast. AIMS Bioeng. 2, 423–436. doi:10.3934/bioeng.2015.4.423

Feng, P., Shang, Y., Cen, K., Wang, C., 2015. Fungal biosynthesis of the bibenzoquinone oosporein to evade insect immunity. Proc. Natl. Acad. Sci. U. S. A. 112, 11365–11370. doi:10.1073/pnas.1503200112

Fernandes, P., Cruz, A., Angelova, B., Pinheiro, H.M., Cabral, J.M.S., 2003. Microbial conversion of steroid compounds: Recent developments. Enzyme Microb. Technol. 32, 688–705. doi:10.1016/S0141-0229(03)00029-2

Fliss, A.E., Benzeno, S., Rao, J., Caplan, A.J., 2000. Control of estrogen receptor ligand binding by Hsp90. J. Steroid Biochem. Mol. Biol. 72, 223–230. doi:10.1016/S0960-0760(00)00037-6

Fogg, C., Lustig, S., Whitbeck, J.C., Eisenberg, R.J., Cohen, G.H., Moss, B., 2004. Protective Immunity to Vaccinia Virus Induced by Vaccination with Multiple Recombinant Outer Membrane Proteins of Intracellular and Extracellular Virions. J. Virol. 78, 10230–10237. doi:10.1128/JVI.78.19.10230

Förster, J., Halbfeld, C., Zimmermann, M., Blank, L.M., 2014. A blueprint of the amino acid biosynthesis network of hemiascomycetes. FEMS Yeast Res. 14, 1090–1100. doi:10.1111/1567-1364.12205

Fraatz, M.A., Berger, R.G., Zorn, H., 2009. Nootkatone-a biotechnological challenge. Appl. Microbiol. Biotechnol. 83, 35–41. doi:10.1007/s00253-009-1968-x

Galli, A., Schiestl, R.H., 1996. Hydroxyurea induces recombination in dividing but not in G1 or G2 cell cycle arrested yeast cells. Mutat. Res. 354, 69–75. doi:10.1016/0027-5107(96)00037-1

Gao, L., Cai, M., Shen, W., Xiao, S., Zhou, X., Zhang, Y., 2013. Engineered fungal polyketide biosynthesis in *Pichia pastoris*: a potential excellent host for polyketide production. Microb. Cell Fact. 12, 77. doi:10.1186/1475-2859-12-77

Gasser, B., Dragosits, M., Mattanovich, D., 2010. Engineering of biotin-prototrophy in *Pichia pastoris* for robust production processes. Metab. Eng. 12, 573–80. doi:10.1016/j.ymben.2010.07.002

Gasser, B., Maurer, M., Gach, J., Kunert, R., Mattanovich, D., 2006. Engineering of *Pichia pastoris* for improved production of antibody fragments. Biotechnol. Bioeng. 94, 353–361. doi:10.1002/bit

Gasser, B., Maurer, M., Rautio, J., Sauer, M., Bhattacharyya, A., Saloheimo, M., Penttilä, M., Mattanovich, D., 2007. Monitoring of transcriptional regulation in *Pichia pastoris* under protein production conditions. BMC Genomics 8, 179. doi:10.1186/1471-2164-8-179

Gavagan, J.E., Fager, S.K., Seip, J.E., Clark, D.S., Payne, M.S., Anton, D.L., Dicosimo, R., 1997. Chemoenzymic Synthesis of N-(Phosphonomethyl) glycine. J. Org. Chem. 62, 5419–5427. doi:10.1021/jo970455f

Gavira, C., Höfer, R., Lesot, A., Lambert, F., Zucca, J., Werck-Reichhart, D., 2013. Challenges and pitfalls of P450-dependent (+)-valencene bioconversion by *Saccharomyces cerevisiae*. Metab. Eng. 18, 25–35. doi:10.1016/j.ymben.2013.02.003

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., Braun, A., Emmerstorfer, A., Pichler, H., Glieder, A., 2012. Production of human cytochrome P450 2D6 drug metabolites with recombinant microbes - a comparative study. Biotechnol. J. 7, 1346–1358. doi:10.1002/biot.201200187

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

Generoso, W.C., Gottardi, M., Oreb, M., Boles, E., 2016. Simplified CRISPR-Cas genome editing for *Saccharomyces cerevisiae*. J. Microbiol. Methods 127, 203–205. doi:10.1016/j.mimet.2016.06.020

Gleeson, M.A., Sudbery, P.E., 1988. The methylotrophic yeasts. Yeast 4, 1–15. doi:10.1002/yea.320040102

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. BMC Genomics 9, 390. doi:10.1186/1471-2164-9-390

Gruchattka, E., Hädicke, O., Klamt, S., Schütz, V., Kayser, O., 2013. *In silico* profiling of *Escherichia coli* and *Saccharomyces cerevisiae* as terpenoid factories. Microb. Cell Fact. 12, 84. doi:10.1186/1475-2859-12-84

Gruchattka, E., Kayser, O., 2015. *In vivo* validation of *in silico* predicted metabolic engineering strategies in yeast: Disruption of alpha-ketoglutarate dehydrogenase and expression of ATP-citrate lyase for terpenoid production. PLoS One 10, 1–25. doi:10.1371/journal.pone.0144981

Guerfal, M., Ryckaert, S., Jacobs, P.P., Ameloot, P., Van Craenenbroeck, K., Derycke, R., Callewaert, N., 2010. The *HAC1* gene from *Pichia pastoris*: characterization and effect of its overexpression on the production of secreted, surface displayed and membrane proteins. Microb. Cell Fact. 9, 49. doi:10.1186/1475-2859-9-49

Guirouilh-Barbat, J., Huck, S., Bertrand, P., Pirzio, L., Desmaze, C., Sabatier, L., Lopez, B.S., 2004. Impact of the *KU80* pathway on NHEJ-induced genome rearrangements in mammalian cells. Mol. Cell 14, 611–623. doi:10.1016/j.molcel.2004.05.008

Hamilton, S.R., Davidson, R.C., Sethuraman, N., Nett, J.H., Jiang, Y., Rios, S., Bobrowicz, P., Stadheim, T. a, Li, H., Choi, B., Hopkins, D., Wischnewski, H., Roser, J., Mitchell, T., Strawbridge, R.R., Hoopes, J., Wildt, S., Gerngross, T.U., 2006. Humanization of Yeast to Produce Complex Terminally Sialylated Glycoproteins. Science 313, 1441–1443. doi:10.1126/science.1130256

Hamilton, S.R., Gerngross, T.U., 2007. Glycosylation engineering in yeast: the advent of fully humanized yeast. Curr. Opin. Biotechnol. 18, 387–392. doi:10.1016/j.copbio.2007.09.001

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. doi:10.1093/nar/gkn369

He, Y., Cox, R.J., 2016. The molecular steps of citrinin biosynthesis in fungi. Chem. Sci. 7, 2119–2127. doi:10.1039/C5SC04027B

Heiss, S., Maurer, M., Hahn, R., Mattanovich, D., Gasser, B., 2013. Identification and deletion of the major secreted protein of *Pichia pastoris* . Appl. Microbiol. Biotechnol. 97, 1241–9. doi:10.1007/s00253-012-4260-4

Heiss, S., Puxbaum, V., Gruber, C., Altmann, F., Mattanovich, D., Gasser, B., 2015. Multistep processing of the secretion leader of the extracellular protein Epx1 in *Pichia pastoris* and implications for protein localization. Microbiology 161, 1356–1368. doi:10.1099/mic.0.000105

Hemmerich, J., Adelantado, N., Barrigón, J.M., Ponte, X., Hörmann, A., Ferrer, P., Kensy, F., Valero, F., 2014. Comprehensive clone screening and evaluation of fed-batch strategies in a microbioreactor and lab scale stirred tank bioreactor system: application on *Pichia pastoris* producing *Rhizopus oryzae* lipase. Microb. Cell Fact. 13, 36. doi:10.1186/1475-2859-13-36

Hesketh, A.R., Castrillo, J.I., Sawyer, T., Archer, D.B., Oliver, S.G., 2013. Investigating the physiological response of *Pichia* (*Komagataella*) *pastoris* GS115 to the heterologous expression of misfolded proteins using chemostat cultures. Appl. Microbiol. Biotechnol. 97, 9747–9762. doi:10.1007/s00253-013-5186-1

Heyland, J., Fu, J., Blank, L.M., Schmid, A., 2011. Carbon metabolism limits recombinant protein production in *Pichia pastoris*. Biotechnol. Bioeng. 108, 1942–1953. doi:10.1002/bit.23114

Hori, K., Okano, S., Sato, F., 2016. Efficient microbial production of stylopine using a *Pichia pastoris* expression system. Sci. Rep. 6, 22201. doi:10.1038/srep22201

Inan, M., Meagher, M.M., 2001. Non-repressing carbon sources for alcohol oxidase (*AOX1*) promoter of *Pichia pastoris* . J. Biosci. Bioeng. 92, 585–9.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., Weissman, J.S., 2009. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. Science 324, 218–223. doi:10.1126/science.1168978

Invitrogen, 2010. *Pichia* Fermentation Process Guidelines.

Irani, Z.A., Kerkhoven, E., Shojaosadati, S.A., Nielsen, J., 2016. Genome-scale metabolic model of *Pichia pastoris* with native and humanized glycosylation of recombinant proteins. Biotechnol. Bioeng. 113, 961–969. doi:10.1002/bit.25863

Iwashima, A., Nishino, H., Nose, Y., 1973. Carrier-mediated transport of thiamine in baker's yeast. Biochim. Biophys. Acta 330, 222–234. doi:10.1016/0005-2736(73)90227-7

Jacobs, P.P., Geysens, S., Vervecken, W., Contreras, R., Callewaert, N., 2009. Engineering complex-type N-glycosylation in *Pichia pastoris* using GlycoSwitch technology. Nat. Protoc. 4, 58–70. doi:10.1038/nprot.2008.213

Jakočiūnas, T., Bonde, I., Herrgård, M., Harrison, S.J., Kristensen, M., Pedersen, L.E., Jensen, M.K., Keasling, J.D., 2015. Multiplex metabolic pathway engineering using CRISPR/Cas9 in *Saccharomyces cerevisiae*. Metab. Eng. 28, 213–222. doi:10.1016/j.ymben.2015.01.008

Jensen, E.D., Ferreira, R., Jakočiūnas, T., Arsovska, D., Zhang, J., Ding, L., Smith, J.D., David, F., Nielsen, J., Jensen, M.K., Keasling, J.D., 2017. Transcriptional reprogramming in yeast using dCas9 and combinatorial gRNA strategies. Microb. Cell Fact. 1–16. doi:10.1186/s12934-017-0664-2

Jeong, E., Shim, W.Y., Kim, J.H., 2014. Metabolic engineering of *Pichia pastoris* for production of hyaluronic acid with high molecular weight. J. Biotechnol. 185, 28–36. doi:10.1016/j.jbiotec.2014.05.018

Jeong, Y., Kim, J.-N., Kim, M.W., Bucca, G., Cho, S., Yoon, Y.J., Kim, B.-G., Roe, J.-H., Kim, S.C., Smith, C.P., Cho, B.-K., 2016. The dynamic transcriptional and translational landscape of the model antibiotic producer *Streptomyces coelicolor* A3(2). Nat. Commun. 7, 11605. doi:10.1038/ncomms11605

Jiang, B., Argyros, R., Bukowski, J., Nelson, S., Sharkey, N., Kim, S., Copeland, V., Davidson, R.C., Chen, R., Zhuang, J., Sethuraman, N., Stadheim, T.A., 2015. Inactivation of a *GAL4*-like transcription factor improves cell fitness and product yield in glycoengineered *Pichia pastoris* strains. Appl. Environ. Microbiol. 81, 260–271. doi:10.1128/AEM.02619-14

Jiang, M., Guo, B., Wan, X., Gong, Y., Zhang, Y., Hu, C., 2014. Isolation and characterization of the diatom *Phaeodactylum* delta5-elongase gene for transgenic LC-PUFA production in *Pichia pastoris*. Mar. Drugs 12, 1317–1334. doi:10.3390/md12031317

Johnson, M.A., Waterham, H.R., Ksheminska, G.P., Fayura, L.R., Cereghino, J.L., Stasyk, O. V, Veenhuis, M., Kulachkovsky, a R., Sibirny, a a, Cregg, J.M., 1999. Positive selection of novel peroxisome biogenesis-defective mutants of the yeast *Pichia pastoris*. Genetics 151, 1379–1391.

Johnston, M., Davis, R.W., 1984. Sequences that regulate the divergent *GAL1-GAL10* promoter in *Saccharomyces cerevisiae*. Mol. Cell. Biol. 4, 1440–1448. doi:10.1128/MCB.4.8.1440

Jordà, J., Rojas, H.C., Carnicer, M., Wahl, A., Ferrer, P., Albiol, J., 2014. Quantitative Metabolomics and Instationary ^{13}C-Metabolic Flux Analysis Reveals Impact of Recombinant Protein Production on Trehalose and Energy Metabolism in *Pichia pastoris*. Metabolites 4, 281–299. doi:10.3390/metabo4020281

Jordà, J., Suarez, C., Carnicer, M., ten Pierick, A., Heijnen, J.J., van Gulik, W., Ferrer, P., Albiol, J., Wahl, A., 2013. Glucose-methanol co-utilization in *Pichia pastoris* studied by metabolomics and instationary ^{13}C flux analysis. BMC Syst. Biol. 7, 17. doi:10.1186/1752-0509-7-17

Kant, H.R., Balamurali, M., Meenakshisundaram, S., 2014. Enhancing precursors availability in *Pichia pastoris* for the overproduction of S-adenosyl-L-methionine employing molecular strategies with process tuning. J. Biotechnol. 188, 112–121. doi:10.1016/j.jbiotec.2014.08.017

Keasling, J.D., 2010. Manufacturing Molecules Through Metabolic Engineering. Science 330, 1355–1358. doi:10.1126/science.1193990

Kegel, A., Martinez, P., Carter, S.D., Åström, S.U., 2006. Genome wide distribution of illegitimate recombination events in *Kluyveromyces lactis*. Nucleic Acids Res. 34, 1633–1645. doi:10.1093/nar/gkl064

Kelwick, R., MacDonald, J.T., Webb, A.J., Freemont, P., 2014. Developments in the tools and methodologies of synthetic biology. Front. Bioeng. Biotechnol. 2, 60. doi:10.3389/fbioe.2014.00060

Khalil, A.S., Lu, T.K., Bashor, C.J., Ramirez, C.L., Pyenson, N.C., Joung, J.K., Collins, J.J., 2012. A synthetic biology framework for programming eukaryotic transcription functions. Cell 150, 647–658. doi:10.1016/j.cell.2012.05.045

Kim, S.H., Roh, K.H., Kim, K.-S., Kim, H.U., Lee, K.-R., Kang, H.-C., Kim, J.-B., 2014. Coexpression of multiple genes reconstitutes two pathways of very long-chain

polyunsaturated fatty acid biosynthesis in *Pichia pastoris*. Biotechnol. Lett. 36, 1843–51. doi:10.1007/s10529-014-1550-1

Kim, Y.J., Zhang, D., Yang, D.C., 2015. Biosynthesis and biotechnological production of ginsenosides. Biotechnol. Adv. 33, 717–735. doi:10.1016/j.biotechadv.2015.03.001

King, Z.A., Lloyd, C.J., Feist, A.M., Palsson, B.O., 2015. Next-generation genome-scale models for metabolic engineering. Curr. Opin. Biotechnol. 35, 23–29. doi:10.1016/j.copbio.2014.12.016

Kohlhaw, G.B., 2003. Leucine Biosynthesis in Fungi : Entering Metabolism through the Back Door. Microbiol. Mol. Biol. Rev. 67, 1–15. doi:10.1128/MMBR.67.1.1

Komor, A.C., Badran, A.H., Liu, D.R., 2016. CRISPR-Based Technologies for the Manipulation of Eukaryotic Genomes. Cell 168, 1–17. doi:10.1016/j.cell.2016.10.044

Krainer, F.W., Capone, S., Jäger, M., Vogl, T., Gerstmann, M., Glieder, A., Herwig, C., Spadiut, O., 2015. Optimizing cofactor availability for the production of recombinant heme peroxidase in *Pichia pastoris*. Microb. Cell Fact. 14, 4. doi:10.1186/s12934-014-0187-z

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 Fact. 11, 22. doi:10.1186/1475-2859-11-22

Krainer, F.W., Gerstmann, M.A., Darnhofer, B., Birner-Gruenberger, R., Glieder, A., 2016. Biotechnological advances towards an enhanced peroxidase production in *Pichia pastoris*. J. Biotechnol. 233, 181–189. doi:10.1016/j.jbiotec.2016.07.012

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, 3279. doi: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. J. Biotechnol. 154, 312–20. 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

Lamacka, M., Sajbidor, J., 1997. Ergosterol determination in *Saccharomyces cerevisiae*. Comparison of different methods. Biotechnol. Tech. 11, 723–725. doi:10.1023/A:1018484015806

Landes, N., Gasser, B., Vorauer-Uhl, K., Lhota, G., Mattanovich, D., Maurer, M., 2016. The vitamin-sensitive promoter *P THI11* enables pre-defined autonomous induction of recombinant protein production in *Pichia pastoris*. Biotechnol. Bioeng. 113, 2633–2643. doi:10.1002/bit.26041

Landry, M.P., Ando, H., Chen, A.Y., Cao, J., Kottadiel, V.I., Chio, L., Yang, D., Dong, J., Lu, T.K., Strano, M.S., 2017. Single-molecule detection of protein efflux from microorganisms using fluorescent single-walled carbon nanotube sensor arrays. Nat. Nanotechnol. 102, 26–35. doi:10.1038/nnano.2016.284

Laukens, B., De Visscher, C., Callewaert, N., 2015. Engineering yeast for producing human glycoproteins: where are we now? Future Microbiol. 10, 21–34. doi:10.2217/fmb.14.104

Leber, R., Zenz, R., Schröttner, K., Fuchsbichler, S., Pühringer, B., Turnowsky, F., 2001. A novel sequence element is involved in the transcriptional regulation of expression of the *ERG1* (squalene epoxidase) gene in *Saccharomyces cerevisiae*. Eur. J. Biochem. 268, 914–924. doi:10.1046/j.1432-1327.2001.01940.x

Lee, C.C., Williams, T.G., Wong, D.W.S., Robertson, G.H., 2005. An episomal expression vector for screening mutant gene libraries in *Pichia pastoris*. Plasmid 54, 80–85. doi:10.1016/j.plasmid.2004.12.001

Lee, J.W., Na, D., Park, J.M., Lee, J., Choi, S., Lee, S.Y., 2012. Systems metabolic engineering of microorganisms for natural and non-natural chemicals. Nat. Chem. Biol. 8, 536–546. doi:10.1038/nchembio.970

Lee, M.E., DeLoache, W.C., Cervantes, B., Dueber, J.E., 2015. A Highly Characterized Yeast Toolkit for Modular, Multipart Assembly. ACS Synth. Biol. 4, 975–986. doi:10.1021/sb500366v

Lee, S.Y., Kim, H.U., 2015. Systems strategies for developing industrial microbial strains. Nat. Biotechnol. 33, 1061–1072. doi:10.1038/nbt.3365

Levin, D.E., 2011. Regulation of cell wall biogenesis in *Saccharomyces cerevisiae*: The cell wall integrity signaling pathway. Genetics 189, 1145–1175. doi:10.1534/genetics.111.128264

Li, H., Sethuraman, N., Stadheim, T. a, Zha, D., Prinz, B., Ballew, N., Bobrowicz, P., Choi, B.-K., Cook, W.J., Cukan, M., Houston-Cummings, N.R., Davidson, R., Gong, B., Hamilton, S.R., Hoopes, J.P., Jiang, Y., Kim, N., Mansfield, R., Nett, J.H., Rios, S., Strawbridge, R., Wildt, S., Gerngross, T.U., 2006. Optimization of humanized IgGs in glycoengineered *Pichia pastoris* . Nat. Biotechnol. 24, 210–215. doi:10.1038/nbt1178

Liachko, I., Dunham, M.J., 2014. An autonomously replicating sequence for use in a wide range of budding yeasts. FEMS Yeast Res. 14, 364–7. doi:10.1111/1567-1364.12123

Liachko, I., Youngblood, R.A., Tsui, K., Bubb, K.L., Queitsch, C., Raghuraman, M.K., Nislow, C., Brewer, B.J., Dunham, M.J., 2014. GC-rich DNA elements enable replication origin activity in the methylotrophic yeast *Pichia pastoris*. PLoS Genet. 10, e1004169. doi:10.1371/journal.pgen.1004169

Liang, S., Wang, B., Pan, L., Ye, Y., He, M., Han, S., Zheng, S., Wang, X., Lin, Y., 2012. Comprehensive structural annotation of *Pichia pastoris* transcriptome and the response to various carbon sources using deep paired-end RNA sequencing. BMC Genomics 13, 738. doi:10.1186/1471-2164-13-738

Lin-Cereghino, G.P., Godfrey, L., de la Cruz, B.J., 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*. Mol. Cell. Biol. 26, 883–897. doi:10.1128/MCB.26.3.883

Lin Cereghino, G.P., Lin Cereghino, J., Sunga, A.J., Johnson, M.A., Lim, M., Gleeson, M.A., Cregg, J.M., 2001. New selectable marker/auxotrophic host strain combinations

for molecular genetic manipulation of *Pichia pastoris*. Gene 263, 159–169. doi:10.1016/S0378-1119(00)00576-X

Liu, B., Zhang, Y., Zhang, X., Yan, C., Zhang, Y., Xu, X., Zhang, W., 2016. Discovery of a rhamnose utilization pathway and rhamnose-inducible promoters in *Pichia pastoris*. Sci. Rep. 6, 27352. doi:10.1038/srep27352

Liu, H., Tan, X., Veenhius, M., McCollum, D., Cregg, J.M., 1992. An Efficient Screen for Peroxisome-Deficient Mutants of *Pichia pastoris*. J. Bacteriol. 174, 4943–4951.

Liu, J.K., O'Brien, E.J., Lerman, J. a, Zengler, K., Palsson, B.O., Feist, A.M., 2014. Reconstruction and modeling protein translocation and compartmentalization in *Escherichia coli* at the genome-scale. BMC Syst. Biol. 8, 110. doi:10.1186/s12918-014-0110-6

Liu, L., Agren, R., Bordel, S., Nielsen, J., 2010. Use of genome-scale metabolic models for understanding microbial physiology. FEBS Lett. 584, 2556–2564. doi:10.1016/j.febslet.2010.04.052

Liu, L., Redden, H., Alper, H.S., 2013. Frontiers of yeast metabolic engineering: Diversifying beyond ethanol and *Saccharomyces*. Curr. Opin. Biotechnol. 24, 1023–1030. doi:10.1016/j.copbio.2013.03.005

Liu, Y., Gong, X., Wang, C., Du, G., Chen, J., Kang, Z., 2016. Production of glucaric acid from myo-inositol in engineered *Pichia pastoris*. Enzyme Microb. Technol. 91, 8–16. doi:10.1016/j.enzmictec.2016.05.009

Liu, Y., Zhu, X., Li, W., Wen, H., Gao, Y., Liu, Y., Liu, C., 2014. Enhancing production of ergosterol in *Pichia pastoris* GS115 by over-expression of 3-hydroxy-3-methylglutaryl CoA reductase from Glycyrrhiza uralensis. Acta Pharm. Sin. B 4, 161–166. doi:10.1016/j.apsb.2014.02.007

Liu, X. Bin, Liu, M., Tao, X.Y., Zhang, Z.X., Wang, F.Q., Wei, D.Z., 2015. Metabolic engineering of *Pichia pastoris* for the production of dammarenediol-II. J. Biotechnol. 216, 47–55. doi:10.1016/j.jbiotec.2015.10.005

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. doi:10.1016/j.biotechadv.2015.05.008

Love, K.R., Panagiotou, V., Jiang, B., Stadheim, T.A., Love, J.C., 2010. Integrated single-cell analysis shows *Pichia pastoris* secretes protein stochastically. Biotechnol. Bioeng. 106, 319–325. doi:10.1002/bit.22688

Love, K.R., Politano, T.J., Panagiotou, V., Jiang, B., Stadheim, T.A., Love, J.C., 2012. Systematic single-cell analysis of *Pichia pastoris* reveals secretory capacity limits productivity. PLoS One 7, 1–11. doi:10.1371/journal.pone.0037915

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, 550. doi:10.1186/s12864-016-2876-y

Macauley-Patrick, S., Fazenda, M.L., McNeil, B., Harvey, L.M., 2005. Heterologous protein production using the *Pichia pastoris* expression system. Yeast 22, 249–170. doi:10.1002/yea.1208

Mairinger, T., Steiger, M., Nocon, J., Mattanovich, D., Koellensperger, G., Hann, S., 2015. Gas Chromatography-Quadrupole Time-of-Flight Mass Spectrometry-Based Determination of Isotopologue and Tandem Mass Isotopomer Fractions of Primary Metabolites for ^{13}C-Metabolic Flux Analysis. Anal. Chem. 87, 11792–11802. doi:10.1021/acs.analchem.5b03173

Mans, R., van Rossum, H.M., Wijsman, M., Backx, A., Kuijpers, N.G.A., van den Broek, M., Daran-Lapujade, P., Pronk, J.T., van Maris, A.J.A., Daran, J.-M.G., 2015. CRISPR/Cas9: a molecular Swiss army knife for simultaneous introduction of multiple genetic modifications in *Saccharomyces cerevisiae*. FEMS Yeast Res. 1–15. doi:10.1093/femsyr/fov004

Marsalek, L., Gruber, C., Altmann, F., Aleschko, M., Mattanovich, D., Gasser, B., Puxbaum, V., 2017. Disruption of genes involved in CORVET complex leads to enhanced secretion of heterologous carboxylesterase only in protease deficient *Pichia pastoris*. Biotechnol. J. 1600584. doi:10.1002/biot.201600584

Marx, H., Mattanovich, D., Sauer, M., 2008. Overexpression of the riboflavin biosynthetic pathway in *Pichia pastoris*. Microb. Cell Fact. 7, 23. doi:10.1186/1475-2859-7-23

Massahi, A., Çalik, P., 2016. Endogenous signal peptides in recombinant protein production by *Pichia pastoris*: From *in-silico* analysis to fermentation. J. Theor. Biol. 408, 22–33. doi:10.1016/j.jtbi.2016.07.039

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* . Microb. Cell Fact. 8, 29. doi:10.1186/1475-2859-8-29

McIsaac, R.S., Oakes, B.L., Wang, X., Dummit, K.A., Botstein, D., Noyes, M.B., 2013. Synthetic gene expression perturbation systems with rapid, tunable, single-gene specificity in yeast. Nucleic Acids Res. 41, 1–10. doi:10.1093/nar/gks1313

Meehl, M.A., Stadheim, T.A., 2014. Biopharmaceutical discovery and production in yeast. Curr. Opin. Biotechnol. 30, 120–127. doi:10.1016/j.copbio.2014.06.007

Meesapyodsuk, D., Chen, Y., Ng, S.H., Chen, J., Qiu, X., 2015. Metabolic engineering of *Pichia pastoris* to produce ricinoleic acid, a hydroxy fatty acid of industrial importance. J. Lipid Res. 56, 2102–2109. doi:10.1128/AEM.02827-12

Mellitzer, A., Weis, R., Glieder, A., Flicker, K., 2012. Expression of lignocellulolytic enzymes in *Pichia pastoris*. Microb. Cell Fact. 11, 61. doi:10.1186/1475-2859-11-61

Meyer, V., Arentshorst, M., El-Ghezal, A., Drews, A.-C., Kooistra, R., van den Hondel, C.A.M.J.J., Ram, A.F.J., 2007. Highly efficient gene targeting in the *Aspergillus niger* kusA mutant. J. Biotechnol. 128, 770–775. doi:10.1016/j.jbiotec.2006.12.021

Moreira de Almeida, J.R., Pepe de Moraes, L.M., Gonçalves Torres, F.A., 2005. Molecular characterization of the 3-phosphoglycerate kinase gene (PGK1) from the methylotrophic yeast *Pichia pastoris*. Yeast 22, 725–737. doi:10.1002/yea.1243

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

Nag, D.K., White, M.A., Petes, T.D., 1989. Palindromic sequences in heteroduplex DNA inhibit mismatch repair in yeast. Nature 340, 318–320. doi:10.1038/340318a0

Nett, J.H., Hodel, N., Rausch, S., Wildt, S., 2005. Cloning and disruption of the *Pichia pastoris ARG1*, *ARG2*, *ARG3*, *HIS1*, *HIS2*, *HIS5*, *HIS6* genes and their use as auxotrophic markers. Yeast 22, 295–304. doi:10.1002/yea.1202

Nguyen, P.H., Hasek, J., Kohlwein, S.D., Romero, C., Choi, J.H., Vancura, A., 2005. Interaction of Pik1p and Sjl proteins in membrane trafficking. FEMS Yeast Res. 5, 363–371. doi:10.1016/j.femsyr.2004.09.007

Nielsen, D.R., Leonard, E., Yoon, S.H., Tseng, H.C., Yuan, C., Prather, K.L.J., 2009. Engineering alternative butanol production platforms in heterologous bacteria. Metab. Eng. 11, 262–273. doi:10.1016/j.ymben.2009.05.003

Nishida, K., Arazoe, T., Yachie, N., Banno, S., Kakimoto, M., Tabata, M., Mochizuki, M., Miyabe, A., Araki, M., Hara, K.Y., Shimatani, Z., Kondo, A., 2016. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. Science 102, 553–563. doi:10.1126/science.aaf8729

Nocon, J., Steiger, M., Mairinger, T., Hohlweg, J., Rußmayer, H., Hann, S., Gasser, B., Mattanovich, D., 2016. Increasing pentose phosphate pathway flux enhances recombinant protein production in *Pichia pastoris*. Appl. Microbiol. Biotechnol. 100, 5955–5963. doi:10.1007/s00253-016-7363-5

Nocon, J., Steiger, M.G., Pfeffer, M., Sohn, S.B., Kim, T.Y., Maurer, M., Rußmayer, H., Pflügl, S., Ask, M., Haberhauer-Troyer, C., Ortmayr, K., Hann, S., Koellensperger, G., Gasser, B., Lee, S.Y., Mattanovich, D., 2014. Model based engineering of *Pichia pastoris* central metabolism enhances recombinant protein production. Metab. Eng. 24, 129–38. doi:10.1016/j.ymben.2014.05.011

Öberg, F., Sjöhamn, J., Conner, M.T., Bill, R.M., Hedfalk, K., 2011. Improving recombinant eukaryotic membrane protein yields in *Pichia pastoris* : the importance of

codon optimization and clone selection. Mol. Membr. Biol. 28, 398–411. doi:10.3109/09687688.2011.602219

Obst, U., Lu, T.K., Sieber, V., 2017. A modular toolkit for generating *Pichia pastoris* secretion libraries. ACS Synth. Biol. acssynbio.6b00337. doi:10.1021/acssynbio.6b00337

Ohsawa, S., Yurimoto, H., Sakai, Y., 2017. Novel function of Wsc proteins as a methanol-sensing machinery in the yeast *Pichia pastoris*. Mol. Microbiol. 1–47. doi:10.1111/mmi.13631

Österlund, T., Nookaew, I., Bordel, S., Nielsen, J., 2013. Mapping condition-dependent regulation of metabolism in yeast through genome-scale modeling. BMC Syst. Biol. 7, 36. doi:10.1186/1752-0509-7-36

Ostry, V., Malir, F., Ruprich, J., 2013. Producers and important dietary sources of ochratoxin A and citrinin. Toxins (Basel). 5, 1574–1586. doi:10.3390/toxins5091574

Otero, J.M., Nielsen, J., 2010. Industrial systems biology. Biotechnol. Bioeng. 105, 439–460. doi:10.1002/bit.22592

Palsson, B.O., 2006. Systems biology - Properties of Reconstructed Networks. Cambridge University Press, Cambridge, UK.

Pan, R., Zhang, J., Shen, W.L., Tao, Z.Q., Li, S.P., Yan, X., 2011. Sequential deletion of *Pichia pastoris* genes by a self-excisable cassette. FEMS Yeast Res. 11, 292–298. doi:10.1111/j.1567-1364.2011.00716.x

Partov, S., Siewers, V., Bjorn, S., Nielsen, J., Maury, J., 2010. Characterization of different promoters for designing a new expression vector in *Saccharomyces cerevisiae*. Yeast 27, 955–964. doi:10.1002/yea.1806

Payne, M.S., Petrillo, K.L., Gavagan, J.E., Dicosimo, R., Wagner, L.W., Anton, D.L., 1997. Engineering *Pichia pastoris* for biocatalysis: Co-production of two active enzymes. Gene 194, 179–182. doi:10.1016/S0378-1119(97)00120-0

Payne, M.S., Petrillo, K.L., Gavagan, J.E., Wagner, L.W., DiCosimo, R., Anton, D.L., 1995. High-level production of spinach glycolate oxidase in the methylotrophic yeast *Pichia pastoris*: engineering a biocatalyst. Gene 167, 215–219. doi:10.1016/0378-1119(95)00661-3

Peplowska, K., Markgraf, D.F., Ostrowicz, C.W., Bange, G., Ungermann, C., 2007. The CORVET Tethering Complex Interacts with the Yeast Rab5 Homolog Vps21 and Is Involved in Endo-Lysosomal Biogenesis. Dev. Cell 12, 739–750. doi:10.1016/j.devcel.2007.03.006

Perez-Pinera, P., Han, N., Cleto, S., Cao, J., Purcell, O., Shah, K.A., Lee, K., Ram, R., Lu, T.K., 2016. Synthetic biology and microbioreactor platforms for programmable production of biologics at the point-of-care. Nat. Commun. 7, 12211. doi:10.1038/ncomms12211

Pfeffer, M., Maurer, M., Köllensperger, G., Hann, S., Graf, A.B., Mattanovich, D., 2011. Modeling and measuring intracellular fluxes of secreted recombinant protein in *Pichia pastoris* with a novel 34S labeling procedure. Microb. Cell Fact. 10, 47. doi:10.1186/1475-2859-10-47

Pfeffer, M., Maurer, M., Stadlmann, J., Grass, J., Delic, M., Altmann, F., Mattanovich, D., 2012. Intracellular interactome of secreted antibody Fab fragment in *Pichia pastoris* reveals its routes of secretion and degradation. Appl. Microbiol. Biotechnol. 93, 2503–12. doi:10.1007/s00253-012-3933-3

Poirier, Y., Antonenkov, V.D., Glumoff, T., Hiltunen, J.K., 2006. Peroxisomal ß-oxidation-A metabolic pathway with multiple functions. Biochim. Biophys. Acta - Mol. Cell Res. 1763, 1413–1426. doi:10.1016/j.bbamcr.2006.08.034

Poirier, Y., Erard, N., Petétot, J.M.C., 2002. Synthesis of polyhydroxyalkanoate in the peroxisome of *Pichia pastoris*. FEMS Microbiol. Lett. 207, 97–102. doi:10.1016/S0378-1097(01)00567-5

Polupanov, A.S., Nazarko, V.Y., Sibirny, A.A., 2011. *CCZ1*, *MON1* and *YPT7* genes are involved in pexophagy, the Cvt pathway and non-specific macroautophagy in the methylotrophic yeast *Pichia pastoris*. Cell Biol. Int. 35, 311–9. doi:10.1042/CBI20100547

Polupanov, A.S., Nazarko, V.Y., Sibirny, A.A., 2012. Gss1 protein of the methylotrophic yeast *Pichia pastoris* is involved in glucose sensing, pexophagy and catabolite repression. Int. J. Biochem. Cell Biol. 44, 1906–1918. doi:10.1016/j.biocel.2012.07.017

Portela, R.M.C., Vogl, T., Kniely, C., Fischer, J.E., Oliveira, R., Glieder, A., 2017. Synthetic Core Promoters as Universal Parts for Fine-Tuning Expression in Different Yeast Species. ACS Synth. Biol. 6, 471–484. doi:10.1021/acssynbio.6b00178

Praekelt, U.M., Byrne, K.L., Meacock, P.A., 1994. Regulation of THI4(MOL1), a thiamine-biosynthetic gene of *Saccharomyces cerevisiae*. Yeast 10, 481–490. doi:10.1002/yea.320100407

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, 167. doi:10.1186/s12864-015-1393-8

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

Pronk, J.T., 2002. Auxotrophic Yeast Strains in Fundamental and Applied Research. Appl. Environ. Microbiol. 68, 2095–2100. doi:10.1128/AEM.68.5.2095

Purdue, P.E., Lazarow, P.B., 2001. Peroxisome Biogenesis. Annu. Rev. Cell Biol. 17, 701–752. doi:10.1146/annurev.cellbio.17.1.701

Puxbaum, V., Gasser, B., Mattanovich, D., 2016. The bud tip is the cellular hot spot of protein secretion in yeasts. Appl. Microbiol. Biotechnol. 100, 8159–8168. doi:10.1007/s00253-016-7674-6

Puxbaum, V., Mattanovich, D., Gasser, B., 2015. Quo vadis? The challenges of recombinant protein folding and secretion in *Pichia pastoris*. Appl. Microbiol. Biotechnol. doi:10.1007/s00253-015-6470-z

Qin, X., Qian, J., Yao, G., Zhuang, Y., Zhang, S., Chu, J., 2011. *GAP* promoter library for fine-tuning of gene expression in *Pichia pastoris* . Appl. Environ. Microbiol. 77, 3600–8. doi:10.1128/AEM.02843-10

Raines, C.A., Lloyd, J.C., Dyer, T.A., 1999. New insights into the structure and function of but neglected Calvin cycle enzyme. J. Exp. Bot. 50, 1–8. doi:https://doi.org/10.1093/jxb/50.330.1

Ralston, L., Subramanian, S., Matsuno, M., Yu, O., 2005. Partial Reconstruction of Flavonoid and Isoflavonoid Biosynthesis in Yeast Using Soybean Type I and Type II Chalcone Isomerases. Plant Physiol. 137, 1375–1388. doi:10.1104/pp.104.054502.1

Rasala, B.A., Mayfield, S.P., 2015. Photosynthetic biomanufacturing in green algae; production of recombinant proteins for industrial, nutritional, and medical uses. Photosynth. Res. 123, 227–239. doi:10.1007/s11120-014-9994-7

Rebnegger, C., Graf, A.B., Valli, M., Steiger, M.G., Gasser, B., Maurer, M., Mattanovich, D., 2014. In *Pichia pastoris* , growth rate regulates protein synthesis and secretion, mating and stress response. Biotechnol. J. 9, 511–525. doi:10.1002/biot.201300334

Rebnegger, C., Vos, T., Graf, A.B., Valli, M., Pronk, J.T., Daran-Lapujade, P., Mattanovicha, D., 2016. *Pichia Pastoris* exhibits high viability and a low maintenance energy requirement at near-zero specific growth rates. Appl. Environ. Microbiol. 82, 4570–4583. doi:10.1128/AEM.00638-16

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 14, 2769–2779. doi:10.1002/pmic.201400267

Ro, D.-K., Paradise, E.M., Ouellet, M., Fisher, K.J., Newman, K.L., Ndungu, J.M., Ho, K.A., Eachus, R.A., Ham, T.S., Kirby, J., Chang, M.C.Y., Withers, S.T., Shiba, Y., Sarpong, R., Keasling, J.D., 2006. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature 440, 3–6. doi:10.1038/nature04640

Roh, C., Seo, S.H., Choi, K.Y., Cha, M., Pandey, B.P., Kim, J.H., Park, J.S., Kim, D.H., Chang, I.S., Kim, B.G., 2009. Regioselective hydroxylation of isoflavones by *Streptomyces avermitilis* MA-4680. J. Biosci. Bioeng. 108, 41–46. doi:10.1016/j.jbiosc.2009.02.021

Roquet, N., Lu, T.K., 2014. Digital and analog gene circuits for biotechnology. Biotechnol. J. 9, 597–608. doi:10.1002/biot.201300258

Roquet, N., Soleimany, A.P., Ferris, A.C., Aaronson, S., Lu, T.K., 2016. Synthetic recombinase-based state machines in living cells. Science 353, aad8559. doi:10.1126/science.aad8559

Rose, M.D., Novick, P., Thomas, J.H., Botstein, D., Fink, G.R., 1987. A *Saccharomyces cerevisiae* genomic plasmid bank based on a centromere-containing shuttle vector. Gene 60, 237–243. doi:10.1016/0378-1119(87)90232-0

Rußmayer, H., Buchetics, M., Gruber, C., Valli, M., Grillitsch, K., Modarres, G., Guerrasio, R., Klavins, K., Neubauer, S., Drexler, H., Steiger, M., Troyer, C., Al Chalabi, A., Krebiehl, G., Sonntag, D., Zellnig, G., Daum, G., Graf, A.B., Altmann, F., Koellensperger, G., Hann, S., Sauer, M., Mattanovich, D., Gasser, B., 2015a. Systems-level organization of yeast methylotrophic lifestyle. BMC Biol. 13, 80. doi:10.1186/s12915-015-0186-5

Rußmayer, H., Troyer, C., Neubauer, S., Steiger, M.G., Gasser, B., Hann, S., Koellensperger, G., Sauer, M., Mattanovich, D., 2015b. Metabolomics sampling of *Pichia pastoris* revisited: rapid filtration prevents metabolite loss during quenching. FEMS Yeast Res. 15, fov049. doi:https://doi.org/10.1093/femsyr/fov049

Ruth, C., Buchetics, M., Vidimce, V., Kotz, D., Naschberger, S., Mattanovich, D., Pichler, H., Gasser, B., 2014. *Pichia pastoris* Aft1-a novel transcription factor, enhancing recombinant protein secretion. Microb. Cell Fact. 13, 120. doi:10.1186/s12934-014-0120-5

Sánchez, B.J., Nielsen, J., 2015. Genome scale models of yeast: Towards standardized evaluation and consistent omic integration. Integr. Biol. 7, 846–858. doi:10.1039/C5IB00083A

Sánchez, Ó.J., Cardona, C.A., 2008. Trends in biotechnological production of fuel ethanol from different feedstocks. Bioresour. Technol. 99, 5270–5295. doi:10.1016/j.biortech.2007.11.013

Schellenberger, J., Que, R., Fleming, R.M.T., Thiele, I., Orth, J.D., Feist, A.M., Zielinski, D.C., Bordbar, A., Lewis, N.E., Rahmanian, S., Kang, J., Hyduke, D.R., Palsson, B.Ø., 2011. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat. Protoc. 6, 1290–1307. doi:10.1038/nprot.2011.308

Schwarzhans, J.-P., 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. Front. Microbiol. 8, 780. doi:10.3389/fmicb.2017.00780

Schwarzhans, J.-P., Wibberg, D., Winkler, A., Luttermann, T., Kalinowski, J., Friehs, K., 2016a. Integration event induced changes in recombinant protein productivity in *Pichia pastoris* discovered by whole genome sequencing and derived vector optimization. Microb. Cell Fact. 15, 84. doi:10.1186/s12934-016-0486-7

Schwarzhans, J.-P., Wibberg, D., Winkler, A., Luttermann, T., Kalinowski, J., Friehs, K., 2016b. Non-canonical integration events in *Pichia pastoris* encountered during standard transformation analysed with genome sequencing. Sci. Rep. 6, 38952. doi:10.1038/srep38952

Segrè, D., Vitkup, D., Church, G.M., 2002. Analysis of optimality in natural and perturbed metabolic networks. Proc. Natl. Acad. Sci. U. S. A. 99, 15112–15117. doi:10.1073/pnas.232349399

Shabbir Hussain, M., M Rodriguez, G., Gao, D., Spagnuolo, M., Gambill, L., Blenner, M., 2016. Recent advances in bioengineering of the oleaginous yeast *Yarrowia lipolytica*. AIMS Bioeng. 3, 493–514. doi:10.3934/bioeng.2016.4.493

Shah, K.A., Clark, J.J., Goods, B.A., Politano, T.J., Mozdzierz, N.J., Zimnisky, R.M., Leeson, R.L., Love, J.C., Love, K.R., 2015. Automated pipeline for rapid production and screening of HIV-specific monoclonal antibodies using *Pichia pastoris*. Biotechnol. Bioeng. 112, 2624–2629. doi:10.1002/bit.25663

Shao, M., Zhang, X., Rao, Z., Xu, M., Yang, T., Li, H., Xu, Z., Yang, S., 2016. Efficient testosterone production by engineered *Pichia pastoris* co-expressing human 17β-hydroxysteroid dehydrogenase type 3 and *Saccharomyces cerevisiae* glucose 6-phosphate dehydrogenase with NADPH regeneration. Green Chem. 18, 1774–1784. doi:10.1039/C5GC02353J

Shen, S., Sulter, G., Jeffries, T.W., Cregg, J.M., 1998. A strong nitrogen source-regulated promoter for controlled expression of foreign genes in the yeast *Pichia pastoris*. Gene 216, 93–102. doi:10.1016/S0378-1119(98)00315-1

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 (*PAOX1*) Regulation. PLoS One 11, e0167766. doi: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 Fact. 15, 178. doi:10.1186/s12934-016-0578-4

Sheng, J., Stevens, J., Feng, X., 2016. Pathway Compartmentalization in Peroxisome of *Saccharomyces cerevisiae* to Produce Versatile Medium Chain Fatty Alcohols. Sci. Rep. 6, 26884. doi:10.1038/srep26884

Shin, S.K., Hyeon, J.E., Kim, Y.I., Kang, D.H., Kim, S.W., Park, C., Han, S.O., 2015. Enhanced hydrolysis of lignocellulosic biomass: Bi-functional enzyme complexes expressed in *Pichia pastoris* improve bioethanol production from Miscanthus sinensis. Biotechnol. J. 10, 1912–1919. doi:10.1002/biot.201500081

Shoji, J., Kikuma, T., Kitamoto, K., 2014. Vesicle trafficking, organelle functions, and unconventional secretion in fungal physiology and pathogenicity. Curr. Opin. Microbiol. 20, 1–9. doi:10.1016/j.mib.2014.03.002

Sohn, S.B., Graf, A.B., Kim, T.Y., Gasser, B., Maurer, M., Ferrer, P., Mattanovich, D., Lee, S.Y., 2010. Genome-scale metabolic model of methylotrophic yeast *Pichia pastoris* and its use for *in silico* analysis of heterologous protein production. Biotechnol. J. 5, 705–15. doi:10.1002/biot.201000078

Solà, A., Jouhten, P., Maaheimo, H., Sánchez-Ferrando, F., Szyperski, T., Ferrer, P., 2007. Metabolic flux profiling of *Pichia pastoris* grown on glycerol/methanol mixtures in chemostat cultures at low and high dilution rates. Microbiology 153, 281–290. doi:10.1099/mic.0.29263-0

Soma, Y., Hanai, T., 2015. Self-induced metabolic state switching by a tunable cell density sensor for microbial isopropanol production. Metab. Eng. 30, 7–15. doi:10.1016/j.ymben.2015.04.005

Spong, A.P., Subramani, S., 1993. Cloning and Characterization of Pas5 - a Gene Required for Peroxisome Biogenesis in the Methylotrophic Yeast Pichia Pastoris. J. Cell Biol. 123, 535–548. doi:10.1083/jcb.123.3.535

Srivastava, A., Somvanshi, P., Mishra, B.N., 2012. Reconstruction and visualization of carbohydrate, N-glycosylation pathways in *Pichia pastoris* CBS7435 using computational and system biology approaches. Syst. Synth. Biol. 7, 7–22. doi:10.1007/s11693-012-9102-2

Stadlmayr, G., Mecklenbräuker, A., Rothmüller, M., Maurer, M., Sauer, M., Mattanovich, D., Gasser, B., 2010. Identification and characterisation of novel *Pichia pastoris* promoters for heterologous protein production. J. Biotechnol. 150, 519–529. doi:10.1016/j.jbiotec.2010.09.957

Stanley, D., Bandara, A., Fraser, S., Chambers, P.J., Stanley, G.A., 2010. The ethanol stress response and ethanol tolerance of *Saccharomyces cerevisiae*. J. Appl. Microbiol. 109, 13–24. doi:10.1111/j.1365-2672.2009.04657.x

Staunton, J., Weissman, K.J., 2001. Polyketide biosynthesis: a millennium review. Nat. Prod. Rep. 18, 380–416. doi:10.1039/a909079g

Storici, F., Durham, C.L., Gordenin, D. a, Resnick, M. a, 2003. Chromosomal site-specific double-strand breaks are efficiently targeted for repair by oligonucleotides in yeast. Proc. Natl. Acad. Sci. U. S. A. 100, 14994–14999. doi:10.1073/pnas.2036296100

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. 235, 121–131. doi:10.1016/j.jbiotec.2016.04.023

Sun, J., Wen, F., Si, T., Xu, J.H., Zhao, H., 2012. Direct conversion of xylan to ethanol by recombinant *Saccharomyces cerevisiae* strains displaying an engineered minihemicellulosome. Appl. Environ. Microbiol. 78, 3837–3845. doi:10.1128/AEM.07679-11

Suzuki, K., Imai, Y., Yamashita, I., 1983. In vivo ligation of linear DNA molecules to circular forms in the yeast *Saccharomyces cerevisiae*. J. Bacteriol. 155, 747–754.

Symington, L.S., 2002. Role of *RAD52* epistasis group genes in homologous recombination and double-strand break repair. Microbiol. Mol. Biol. Rev. 66, 630–670. doi:10.1128/MMBR.66.4.630

Szymczak, A.L., Workman, C.J., Wang, Y., Vignali, K.M., Dilioglou, S., Vanin, E.F., Vignali, D.A.A., 2004. Correction of multi-gene deficiency *in vivo* using a single "self-cleaving" 2A peptide-based retroviral vector. Nat. Biotechnol. 22, 589–594. doi:10.1038/nbt957

Takagi, S., Tsutsumi, N., Terui, Y., Kong, X.Y., 2012. Method for methanol independent induction from methanol inducible promoters in Pichia. US8236528 B2.

Tao, H., Guo, D., Zhang, Y., Deng, Z., Liu, T., 2015. Metabolic engineering of microbes for branched-chain biodiesel production with low-temperature property. Biotechnol. Biofuels 8, 92. doi:10.1186/s13068-015-0270-7

Tomàs-Gamisans, M., Ferrer, P., Albiol, J., 2016. Integration and validation of the genome-scale metabolic models of *Pichia pastoris*: A comprehensive update of protein glycosylation pathways, lipid and energy metabolism. PLoS One 11, 1–24. doi:10.1371/journal.pone.0148031

Trantas, E., Panopoulos, N., Ververidis, F., 2009. Metabolic engineering of the complete pathway leading to heterologous biosynthesis of various flavonoids and stilbenoids in *Saccharomyces cerevisiae*. Metab. Eng. 11, 355–366. doi:10.1016/j.ymben.2009.07.004

Tredwell, G.D., Aw, R., Edwards-Jones, B., Leak, D.J., Bundy, J.G., 2017. Rapid screening of cellular stress responses in recombinant *Pichia pastoris* strains using metabolite profiling. J. Ind. Microbiol. Biotechnol. 44, 413–417. doi:10.1007/s10295-017-1904-5

Tredwell, G.D., Edwards-Jones, B., Leak, D.J., Bundy, J.G., 2011. The development of metabolomic sampling procedures for *Pichia pastoris*, and baseline metabolome data. PLoS One 6. doi:10.1371/journal.pone.0016286

Trinklein, N.T., Force Aldred, S., Hartman, S.J., Schroeder, D.I., Otillar, R.P., Myers, R.M., 2004. An abundance of bidirectional promoters in the human genome. Genome Res. 14, 62–66. doi:10.1101/gr.1982804

Tsakraklides, V., Brevnova, E., Stephanopoulos, G., Shaw, A.J., 2015. Improved gene targeting through cell cycle synchronization. PLoS One 10, 1–11. doi:10.1371/journal.pone.0133434

Tschopp, J.F., Brust, P.F., Cregg, J.M., Stillman, C.A., Gingeras, T.R., 1987. Expression of the *lacZ* gene from two methanol-regulated promoters in *Pichia pastoris*. Nucleic Acids Res. 15, 3859–3876.

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. 1–28. doi:10.1093/femsyr/fow051

van der Klei, I.J., Yurimoto, H., Sakai, Y., Veenhuis, M., 2006. The significance of peroxisomes in methanol metabolism in methylotrophic yeast. Biochim. Biophys. Acta - Mol. Cell Res. 1763, 1453–1462. doi:10.1016/j.bbamcr.2006.07.016

Van Dijk, E.L., Auger, H., Jaszczyszyn, Y., Thermes, C., 2014. Ten years of next-generation sequencing technology. Trends Genet. 30, 418–426. doi:10.1016/j.tig.2014.07.001

Vassileva, A., Chugh, D.A., Swaminathan, S., Khanna, N., 2001. Effect of copy number on the expression levels of hepatitis B surface antigen in the methylotrophic yeast *Pichia pastoris* . Protein Expr. Purif. 21, 71–80. doi:10.1006/prep.2000.1335

Verbeke, J., Beopoulos, A., Nicaud, J.M., 2013. Efficient homologous recombination with short length flanking fragments in Ku70 deficient *Yarrowia lipolytica* strains. Biotechnol. Lett. 35, 571–576. doi:10.1007/s10529-012-1107-0

Verna, J., Lodder, A., Lee, K., Vagts, A., Ballester, R., 1997. A family of genes required for maintenance of cell wall integrity and for the stress response in *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. U. S. A. 94, 13804–13809. doi:10.1073/pnas.94.25.13804

Vervecken, W., Kaigorodov, V., Callewaert, N., Geysens, S., De Vusser, K., Contreras, R., 2004. *In vivo* synthesis of mammalian-like, hybrid-type N-glycans in *Pichia pastoris*. Appl. Environ. Microbiol. 70, 2639–2646. doi:10.1128/AEM.70.5.2639-2646.2004

Verwaal, R., Wang, J., Meijnen, J.P., Visser, H., Sandmann, G., Van Den Berg, J.A., Van Ooyen, A.J.J., 2007. High-level production of beta-carotene in *Saccharomyces cerevisiae* by successive transformation with carotenogenic genes from *Xanthophyllomyces dendrorhous*. Appl. Environ. Microbiol. 73, 4342–4350. doi:10.1128/AEM.02759-06

Vogl, T., Ahmad, M., Krainer, F.W., Schwab, H., Glieder, A., 2015a. Restriction site free cloning (RSFC) plasmid family for seamless, sequence independent cloning in *Pichia pastoris*. Microb. Cell Fact. 14, 103. doi:10.1186/s12934-015-0293-6

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

Vogl, T., Kickenweiz, T., Sturmberger, L., Glieder, A., 2015b. Bidirectional promoter. EP2862933A3.

Vogl, T., Ruth, C., Pitzer, J., Kickenweiz, T., Glieder, A., 2014. Synthetic core promoters for *Pichia pastoris*. ACS Synth. Biol. 3, 188–191. doi:10.1021/sb400091p

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. 5, 172–186. doi: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, 41850. doi:10.1038/srep41850

Wang, S., Yao, Q., Tao, J., Qiao, Y., Zhang, Z., 2007. Co-ordinate expression of glycine betaine synthesis genes linked by the FMDV 2A region in a single open reading frame in *Pichia pastoris*. Appl. Microbiol. Biotechnol. 77, 891–899. doi:10.1007/s00253-007-1222-3

Wang, T.Y., Tsai, Y.H., Yu, I.Z., Chang, T.S., 2015. Improving 3'-hydroxygenistein production in recombinant *Pichia pastoris* using periodic hydrogen peroxide-shocking strategy. J. Microbiol. Biotechnol. 26, 498–502. doi:10.4014/jmb.1509.09013

Wang, X., Wang, Q., Wang, J., Bai, P., Shi, L., Shen, W., Zhou, M., Zhou, X., Zhang, Y., Cai, M., 2016. Mit1 Transcription Factor Mediates Methanol Signaling and Regulates Alcohol Oxidase 1 Promoter in *Pichia pastoris*. J. Biol. Chem. 291, 6245–6261. doi:10.1074/jbc.M115.692053

Waterham, H.R., Cregg, J.M., 1997. Peroxisome biogenesis. BioEssays 19, 57–66. doi:10.1002/bies.950190110

Waterham, H.R., Digan, M.E., Koutz, P.J., Lair, S. V, Cregg, J.M., 1997a. Isolation of the *Pichia pastoris* glyceraldehyde-3-phosphate dehydrogenase gene and regulation and use of its promoter. Gene 186, 37–44.

Waterham, H.R., Russell, K.A., De Vries, Y., Cregg, J.M., 1997b. Peroxisomal targeting, import, and assembly of alcohol oxidase in *Pichia pastoris*. J. Cell Biol. 139, 1419–1431. doi:10.1083/jcb.139.6.1419

Weinhandl, K., Ballach, M., Winkler, M., Ahmad, M., Glieder, A., Birner-Gruenberger, R., Fotheringham, I., Escalettes, F., Camattari, A., 2016. *Pichia pastoris* mutants as host strains for efficient secretion of recombinant branched chain aminotransferase (BCAT). J. Biotechnol. 235, 84–91. doi:10.1016/j.jbiotec.2016.06.004

Weninger, A., Glieder, A., Vogl, T., 2015. A toolbox of endogenous and heterologous nuclear localization sequences for the methylotrophic yeast *Pichia pastoris*. FEMS Yeast Res. 15, fov082. doi:10.1093/femsyr/fov082

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. doi:10.1016/j.jbiotec.2016.03.027

Werten, M.W.T., De Wolf, F.A., 2005. Reduced proteolysis of secreted gelatin and Yps1-mediated alpha-factor leader processing in a *Pichia pastoris* kex2 disruptant. Appl. Environ. Microbiol. 71, 2310–2317. doi:10.1128/AEM.71.5.2310-2317.2005

Withers, S.T., Keasling, J.D., 2007. Biosynthesis and engineering of isoprenoid small molecules. Appl. Microbiol. Biotechnol. 73, 980–990. doi:10.1007/s00253-006-0593-1

Wriessnegger, T., Augustin, P., Engleder, M., Leitner, E., Müller, M., Kaluzna, I., Schürmann, M., Mink, D., Zellnig, G., Schwab, H., Pichler, H., 2014. Production of the sesquiterpenoid (+)-nootkatone by metabolic engineering of *Pichia pastoris*. Metab. Eng. 24, 18–29. doi:10.1016/j.ymben.2014.04.001

Wriessnegger, T., Gübitz, G., Leitner, E., Ingolic, E., Cregg, J., de la Cruz, B.J., Daum, G., 2007. Lipid composition of peroxisomes from the yeast *Pichia pastoris* grown on different carbon sources. Biochim. Biophys. Acta - Mol. Cell Biol. Lipids 1771, 455–461. doi:10.1016/j.bbalip.2007.01.004

Wriessnegger, T., Moser, S., Emmerstorfer-Augustin, A., Leitner, E., Müller, M., Kaluzna, I., Schürmann, M., Mink, D., Pichler, H., 2016. Enhancing cytochrome P450-mediated conversions in *P. pastoris* through *RAD52* over-expression and optimizing the cultivation conditions. Fungal Genet. Biol. 89, 114–125. doi:10.1016/j.fgb.2016.02.004

Wu, S., Fallon, R.D., Payne, M.S., 1999. Engineering *Pichia pastoris* for stereoselective nitrile hydrolysis by co-producing three heterologous proteins. Appl. Microbiol. Biotechnol. 52, 186–190. doi:10.1007/s002530051507

Xiao, S., Gao, Y., Wang, X., Shen, W., Wang, J., Zhou, X., Cai, M., Zhang, Y., 2016. Peroxisomes targeted and tandem repeat multimer expressions of human antimicrobial peptide LL37 in *Pichia pastoris*. Prep. Biochem. Biotechnol. 1–7. doi:http://dx.doi.org/10.1080/10826068.2016.1201684

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Münster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., Steinmetz, L.M., 2009. Bidirectional promoters generate pervasive transcription in yeast. Nature 457, 1033–7. doi:10.1038/nature07728

Xue, Y., Kong, C., Shen, W., Bai, C., Ren, Y., Zhou, X., Zhang, Y., Cai, M., 2017. Methylotrophic yeast *Pichia pastoris* as a chassis organism for polyketide synthesis via the full citrinin biosynthetic pathway. J. Biotechnol. 242, 64–72. doi:10.1016/j.jbiotec.2016.11.031

Yan, J., Zheng, X., Du, L., Li, S., 2014. Integrated lipase production and *in situ* biodiesel synthesis in a recombinant *Pichia pastoris* yeast: an efficient dual biocatalytic system composed of cell free enzymes and whole cell catalysts. Biotechnol Biofuels 7, 55. doi:10.1186/1754-6834-7-55

Yan, M., Rachubinski, D.A., Joshi, S., Rachubinski, R.A., Subramani, S., 2008. Dysferlin Domain-containing Proteins, Pex30p and Pex31p, Localized to Two Compartments, Control the Number and Size of Oleate-Induced Peroxisomes in *Pichia pastoris*. Mol. Biol. Cell 19, 885–898. doi:10.1091/mbc.E07-10-1042

Yu, X.-W., Wang, R., Zhang, M., Xu, Y., Xiao, R., 2012. Enhanced thermostability of a *Rhizopus chinensis* lipase by *in vivo* recombination in *Pichia pastoris*. Microb. Cell Fact. 11, 102–112. doi:10.1186/1475-2859-11-102

Zecchin, A., Stapor, P.C., Goveia, J., Carmeliet, P., 2015. Metabolic pathway compartmentalization: An underappreciated opportunity? Curr. Opin. Biotechnol. 34, 73–81. doi:10.1016/j.copbio.2014.11.022

Zhang, M., Liu, Z., Yu, Q., Mao, J., Zhang, B., Xing, L., Li, M., 2015. Deletion of genes encoding fatty acid desaturases leads to alterations in stress sensitivity in *Pichia pastoris*. FEMS Yeast Res. 15, fov020. doi:10.1093/femsyr/fov020

Zhang, P., Zhang, W., Zhou, X., Bai, P., Cregg, J.M., Zhang, Y., 2010. Catabolite Repression of Aox in *Pichia pastoris* is dependent on hexose transporter *PpHxt1* and pexophagy. Appl. Environ. Microbiol. 76, 6108–6118. doi:10.1128/AEM.00607-10

Zhang, R., Patena, W., Armbruster, U., Gang, S.S., Blum, S.R., Jonikas, M.C., 2014. High-Throughput Genotyping of Green Algal Mutants Reveals Random Distribution of Mutagenic Insertion Sites and Endonucleolytic Cleavage of Transforming DNA. Plant Cell 26, 1398–1409. doi:10.1105/tpc.114.124099

Zhang, Y., Wang, C., Wang, L., Yang, R., Hou, P., 2017. Direct bioethanol production from wheat straw using xylose/glucose co-fermentation by co-culture of two recombinant yeasts. J. Ind. Microbiol. Biotechnol. 44, 453–464. doi:10.1007/s10295-016-1893-9

Zhao, C., Gao, X., Liu, X., Wang, Y., Yang, S., Wang, F., Ren, Y., 2016. Enhancing biosynthesis of a ginsenoside precursor by self-assembly of two key enzymes in *Pichia pastoris*. J. Agric. Food Chem. 64, 3380–3385. doi:10.1021/acs.jafc.6b00650

Zhou, Y.J., Buijs, N.A., Zhu, Z., Gómez, D.O., Boonsombuti, A., Siewers, V., Nielsen, J., 2016. Harnessing Yeast Peroxisomes for Biosynthesis of Fatty-Acid-Derived Biofuels and

Chemicals with Relieved Side-Pathway Competition. J. Am. Chem. Soc. 138, 15368–15377. doi:10.1021/jacs.6b07394

Zhu, T., Guo, M., Sun, C., Qian, J., Zhuang, Y., Chu, J., Zhang, S., 2009. A systematical investigation on the genetic stability of multi-copy *Pichia pastoris* strains. Biotechnol. Lett. 31, 679–84. doi:10.1007/s10529-009-9917-4

Zhuang, S., Fu, J., Powell, C., Huang, J., Xia, Y., Yan, R., 2015. Production of medium-chain volatile flavour esters in *Pichia pastoris* whole-cell biocatalysts with extracellular expression of *Saccharomyces cerevisiae* acyl-CoA:ethanol O-acyltransferase Eht1 or Eeb1 Background. Springerplus 4. doi:10.1186/s40064-015-1195-0

Zirpel, B., Stehle, F., Kayser, O., 2015. Production of Delta9-tetrahydrocannabinolic acid from cannabigerolic acid by whole cells of *Pichia* (*Komagataella*) *pastoris* expressing Delta9-tetrahydrocannabinolic acid synthase from *Cannabis sativa* L. Biotechnol. Lett. 37, 1869–1875. doi:10.1007/s10529-015-1853-x

2 Results and discussion

2.1 Clonal variability and protein productivity

2.1.1 Motivation

As described above, finding the strain with the optimal productivity characteristics after transformation is markedly complicated by the high clonal variability. Consequently, a work- and time-intense screening procedure has to be performed. In this thesis, this phenomenon was to be investigated in order to better understand it and potentially provide clues on obtaining more strains with the desired characteristics. To this end, it was crucial that the experimental setup closely followed established guidelines for recombinant protein expression studies in *P. pastoris*.

Using firmly established methods from literature for transformation, cultivation and expression, a simple reporter protein was intended to quickly assay protein productivity. *P. pastoris* CBS7435 Δ*HIS4* was selected as host strain because of the popularity of CBS7435 in literature and the complete deletion of *HIS4* [111]. The deletion prevents cases of spontaneous histidine prototrophy as encountered with strain GS115, in which histidine auxotrophy is mediated by a sole single nucleotide polymorphism (SNP) [130]. The well characterized integrative vector pAHBgl from the "*Pichia* Pool" of the Graz University of Technology, Austria, was specifically developed for *P. pastoris* and used as chassis plasmid [85]. Due to the demonstrated suitability of GFPuv (Cycle-3-GFP) for estimating protein productivity in *P. pastoris* [131, 132], it was chosen as reporter. Intracellular expression was chosen over secretion, to avoid effects of the UPR pathway on productivity, which could have complicated result interpretation. The method of Wu & Letchworth (2004) [133] for electroporation was chosen, because of favorable results in preliminary experiments, compared to the protocol by Cereghino *et al.* (2005) [134]. Furthermore, the protocol forgoes a regeneration phase between transformation and plating. Thereby, the influences of cell doubling on observed integration event frequency could be excluded. For cultivation, the deep-well plate technique published by Weis *et al.* (2004) [115] was selected. In contrast to other deep-well plate methods for *P. pastoris*, it was thoroughly tested for reproducibility and scalability. Furthermore, the method was already successfully used in conjunction with GFPuv, to characterize a promoter library in *P. pastoris* [117]. For determination of the gene copy number (GCN)

an established qPCR method for *P. pastoris* was chosen [135]. It has been shown, that digital droplet PCR (ddPCR) can outperform qPCR in regards to high-throughput applications and is less susceptible to pipetting errors during serial dilution. An adapted *P. pastoris* protocol was recently published [113]. However, ddPCR requires specialized equipment, which was not available at the time and could therefore not be applied. A total of seven methods from literature and kits from different suppliers were tested for the quantity and quality of genomic DNA (gDNA) extracted from *P. pastoris* cells. For routine GCN determination and genome sequencing, the gDNA should be of sufficient concentration and of high molecular weight with little degradation. From all analyzed variants, only the kit "MasterPure™ Yeast DNA Purification" from epicentre fulfilled these criteria. Protocols from literature, e.g. "Bust n' Grab" [136], displayed a high degree of degradation, while other commercial kits extracted only low quantities of gDNA. For genome sequencing, the comparatively small size of the *P. pastoris* genome (9.4 Mb) enabled the simultaneous processing of multiple strains in one run. The Illumina MiSeq platform can generate an output of up to ca. 15 Gb per run (reagent kit v3), depending on the read length. In addition, the expertise of the collaborator Dr. Daniel Wibberg (Genome Research of Industrial Microorganisms, CeBiTec, Bielefeld University), who was involved in the *de-novo* genome assembly of *P. pastoris* CBS7435 [137], was of great aid. He performed the assemblies for sequenced clones, using the original genome as reference. More detailed descriptions of the different methods can be found in the methods section of the following publication on pages 151 to 153.

Applying these procedures, a typical *P. pastoris* recombinant protein production experiment was resembled. By choosing methods also used by other scientists working with *P. pastoris*, the hope was that results better reflect events encountered in their experiments. Additionally, it was important to correctly choose mutants for further analysis. From a biotechnological perspective, clones that produce high amounts of the target protein are of high interest for obvious reason. But also clones that produce only small or no amounts of the target protein are relevant, as elucidating their origin might provide pointers to preventing their occurrence in the future. Thereby, more desirable clones can be found after transformation, easing the discovery of high producer strains. If a clone displays a productivity that is significantly different from one theoretically possible based on its GCN, understanding the cause might also be helpful for

biotechnological purposes. For instance, if the cells produce less than what they theoretically are able to, revealing the bottleneck can in the future aid engineer cells to overcome this hurdle.

The publication "Integration event induced changes in recombinant protein productivity in *Pichia pastoris* discovered by whole genome sequencing and derived vector optimization" (Schwarzhans *et al.*, (2016), Microbial Cell Factories, [138]) in the following chapter provides an overview of the experimental setup used for generating the strains and explains the rationale for choosing certain strains for genome sequencing. Consequently, productivity and genome data are combined to find connections between integration event and the features of the strain. Lastly, these insights are used to modify the original vector to avoid a certain detrimental integration event. The supplementary information of the publication can be found in chapter 2.1.3. To the best of our knowledge, no previous study applied this systematic approach to better understand clonal variability on a genetic level and its influence on productivity in *P. pastoris*.

2.1.2 Manuscript

Schwarzhans *et al. Microb Cell Fact (2016) 15:84*
DOI 10.1186/s12934-016-0486-7

Microbial Cell Factories

RESEARCH Open Access

Integration event induced changes in recombinant protein productivity in *Pichia pastoris* discovered by whole genome sequencing and derived vector optimization

Jan-Philipp Schwarzhans[1,2], Daniel Wibberg[3], Anika Winkler[2], Tobias Luttermann[1], Jörn Kalinowski[2,4] and Karl Friehs[1,4*]

Abstract

Background: The classic *AOX1* replacement approach is still one of the most often used techniques for expression of recombinant proteins in the methylotrophic yeast *Pichia pastoris*. Although this approach is largely successful, it frequently delivers clones with unpredicted production characteristics and a work-intense screening process is required to find the strain with desired productivity.

Results: In this project 845 *P. pastoris* clones, transformed with a GFP expression cassette, were analyzed for their methanol-utilization (Mut)-phenotypes, GFP gene expression levels and gene copy numbers. Several groups of strains with irregular features were identified. Such features include GFP expression that is markedly higher or lower than expected based on gene copy number as well as strains that grew under selective conditions but where the GFP gene cassette and its expression could not be detected. From these classes of strains 31 characteristic clones were selected and their genomes sequenced. By correlating the assembled genome data with the experimental phenotypes novel insights were obtained. These comprise a clear connection between productivity and cassette-to-cassette orientation in the genome, the occurrence of false-positive clones due to a secondary recombination event, and lower total productivity due to the presence of untransformed cells within the isolates were discovered. To cope with some of these problems, the original vector was optimized by replacing the *AOX1* terminator, preventing the occurrence of false-positive clones due to the secondary recombination event.

Conclusions: Standard methods for transformation of *P. pastoris* led to a multitude of unintended and sometimes detrimental integration events, lowering total productivity. By documenting the connections between productivity and integration event we obtained a deeper understanding of the genetics of mutation in *P. pastoris*. These findings and the derived improved mutagenesis and transformation procedures and tools will help other scientists working on recombinant protein production in *P. pastoris* and similar non-conventional yeasts.

Keywords: *Pichia pastoris*, Recombinant protein production, *AOX1* promoter, Genome sequencing, Insertion locus, Non-conventional yeast, Expression cassette orientation, False-positive

Background

Pichia (*Komagatella*) *pastoris* is a non-conventional methylotrophic yeast that is widely used as a host for recombinant protein production [1, 2]. Its capability to perform post-translational modifications such as disulphide isomerization or glycosylation, an efficient secretion apparatus and the relative ease of reaching high dry cell weights >100 g/L during bioreactor fermentation make this eukaryote a popular choice for protein expression in industry as well as in research [3–5]. Over 500

*Correspondence: karl.friehs@uni-bielefeld.de
[1] Fermentation Engineering, Bielefeld University, Universitätsstr. 25, 33615 Bielefeld, Germany
Full list of author information is available at the end of the article

proteins, from industrial enzymes to biopharmaceuticals, have been expressed in *P. pastoris* [6]. A growing number of commercial products have reached the market in recent years [7]. Among them are the FDA-approved drugs Kalbitor® and Jetrea®, a kallikrein inhibitor and an aglycosylated protease, respectively [8].

The most common approach for heterologous protein expression in *P. pastoris* is the insertion of the target gene into the genome under the control of the *AOX1* (alcohol oxidase 1) promoter (p*AOX1*). This approach offers tight regulation and a very strong, methanol-inducible expression [6]. Two different modes of homologous recombination-mediated insertion are typically used (i) ends–in insertion leads to additive insertion of the target gene and (ii) ends–out insertion facilitates the replacement of a genomic region, most commonly the native *AOX1* gene [9]. Knock-out of *AOX1* leads to the Mut^S-phenotype (methanol utilization slow), since only the lesser transcribed *AOX2* gene remains. Clones with an additive insertion retain the native phenotype Mut^+ (methanol utilization plus, full growth on methanol). The optimal phenotype for a given application can differ, with Mut^S-strains exhibiting higher productivity than Mut^+-clones in some cases [10, 11]. Recently much progress has been made in understanding the regulation of p*AOX1* as well as creating novel synthetic variants with improved characteristics, underlining that the importance of the promoter still holds [12–17].

A frequently encountered problem during generation of *P. pastoris* clones via homology-mediated integration of the expression cassette is the low targeting efficiency, being as low as <1 % in certain cases like the mannosyltransferase *OCH1* [18]. In addition, an off-target insertion can lead to the disruption of a gene and potentially affect production characteristics. Different techniques have been proposed to improve the targeting efficiency, e.g. preventing random insertion due to non-homologous end-joining (NHEJ) via deletion of a *KU70* homologue or increasing the genetic redundancy [18, 19]. While these methods help to reduce the number of untargeted insertions, they are better suited for genetic engineering studies rather than the generation of a production strain. Scientists working with *P. pastoris* are faced with the task of identifying the optimal producer from a diverse group of clones with varying production characteristics. Similar problems have been reported for other non-conventional yeasts that are often used for recombinant protein expression like *Hansenula polymorpha*, *Yarrowia lipolytica* and *Kluyveromyces lactis* [20–22].

The publicly available genome sequences for the most commonly used *P. pastoris* strains CBS 7435 [23] and GS115 [24] gave rise to multiple genome-scale experiments, most of which focused on better understanding metabolic pathways in order to improve yields in recombinant protein production [25–28]. However, to present no study has been published that investigates the effects of random insertions on the productivity in *P. pastoris* as well as the integration events on the genome scale. In essence a sort of "black box" is present during transformation of *P. pastoris* and it is uncertain if the clone with the desired characteristics will be generated. Unknown events during integration of the expression cassette can lead to drastically different production characteristics of clones from one transformation experiment.

For researchers working with *P. pastoris*, or similar non-conventional yeasts, it would be of great value to gain insights into what might cause unexpected expression levels. By correlating an insertion event seen on the genome with the production characteristics their interaction can be determined. Once these events are known steps can be taken to e.g. optimize vectors to prevent particular integration events.

Using methods previously described and established specifically for *P. pastoris*, a library of 845 clones was characterized for their expression levels and gene copy numbers (GCN) of GFP*uv* (cycle-3-GFP) [29] as well as their Mut-phenotypes [30–33]. Based on these characteristics the clones were grouped and the 31 most outstanding ones selected for genome sequencing. By correlating experimental and genome data novel insights into the integration event and its effect on productivity were discovered and the original vector optimized.

Results and discussion

Characterization and grouping of pAHBgl-GFP *P. pastoris* clones

In total, 845 *P. pastoris* clones transformed with the GFP expression cassette were characterized for their Mut-phenotypes, GFP gene expressions and GCN. The intent of the transformation strategy used in our study was to replace the native *AOX1* gene with a single copy of the GFP expression cassette. Therefore a "regular" clone should have the Mut^S phenotype as well as GCN and GFP expression level of around 1. Overall, 347 out of all 845 clones fall into this category, accounting for approximately 41 % of all clones. This targeting efficiency for *AOX1* is above previously reported values of around 25 % [34, 35]. It has to be considered that in the present study a different histidine auxotroph strain, CBS7435 (Δ*HIS4*) with a fully deleted *HIS4* gene was used [19]. Thereby the background of spontaneous histidine prototrophy conversion clones found in GS115, in which histidine-auxotrophy is mediated by a single nucleotide polymorphism (c.1669C > T resulting in p.557Arg > Cys), is eliminated and the proportion of Mut^S strains increased. Strains with the Mut^+ phenotype can exhibit negative traits due

to illegitimate recombination of the expression cassette into the genome and require more methanol for continuous induction. Nevertheless, they might present suitable hosts for protein expression if adjusting process parameters accordingly [11]. Hence, Mut^+ strains that otherwise displayed the same features as regular clones can also be considered suitable for most applications. To this end, they were added to the "regular" clones in this study.

All strains not falling into either of these categories were designated as "irregular". They displayed certain properties that were not expected based on the transformation modus. Table 1 shows the distribution of strains based on these criteria and their Mut-phenotypes. While about a quarter of all clones exhibited irregular features only five of these were Mut^S strains, underlining the higher genetic variance of Mut^+ clones. A total of 45 multi-copy clones (GCN $\geq$1.5) were found, accounting for ca. 5 % of all clones. Among them seven "jackpot" strains with a GCN >10 are present. All our subsequent analysis concentrated on the irregular clones, in order to elucidate the genetic cause of their aberrant properties.

For a better insight into the diversity of the irregular clones, the relation between GFP expression level and GCN has to be looked at. *P. pastoris* is an industrially important host for recombinant protein expression, therefore these characteristics are at the forefront when it comes to determine whether a clone can meet the requirements of a production process. Interestingly, no clear correlation between GCN and GFP expression level could be seen evaluating all cones (Fig. 1). A wide distribution of clones is visible with no clear pattern. This is in contrast to previously reported results for the relation between GCN and expression level in *P. pastoris* for intracellular protein expression [36]. In other studies, a good linear correlation for intracellular expression was found [37, 38], while secretory expression showcased more complex correlations due to mechanisms like the UPR pathway (unfolded protein response) [34, 39, 40].

Table 1 Distribution of all 845 *P. pastoris* clones, transformed with the GFP expression cassette, based on "regular" and "irregular" properties as well as the Mut-phenotype

Group	No. of clones/ %
Regular clones	
Mut^S	347/41
Mut^+	308/36
Combined	655/77
Irregular clones	
Mut^S	5/1
Mut^+	185/22
Combined	190/23
Total	845/100

To facilitate a clearer understanding of the different kinds of irregular features, the clones were grouped. As shown in Fig. 1 irregular clones could be identified and grouped based on their GCN and GFP expression values. Strains that displayed high GFP expression levels that is paired with a high GCN, two very desirable features for a production strain, were categorized as "high producers". On the other hand, clones that showed a distinct discrepancy between their GCN and GFP expression were designated "over-"or "underachievers". Characteristic for these strains is an expression level markedly exceeding (overachiever) or falling below (underachiever) the expression level expected based on the GCN. Additionally, strains with an expression level or GCN notably below 1 or even at 0 were considered "low-/non-producers" making up the group least desirable as a production strain. Table 2 lists all groups the 190 irregular clones were divided into and the basic characteristics of each group. In Additional file 1: Table S1 the concrete criteria used for dividing clones into these groups can be found. By dividing the strains into groups with similar features, a strategy to identify clones of interest for genome sequencing could be conceived.

Selection of clones for sequencing and general sequencing results

Using a scoring system (Additional file 1: Table S1), the clones with the most outstanding characteristics from the various groups of irregular clones were selected for genome sequencing. In summary, the scoring system emphasized high deviations from the expected GCN or expression level of 1 and an expression level that did not correlate with the GCN. Based on the scoring system, 31 clones were submitted to genome sequencing. Of these 29 had the Mut^+ and 2 the Mut^S phenotype.

The sequencing runs (2 × 300 bp) on the Illumina MiSeq platform resulted in 80,608,298 reads comprising 24.2 Gb of sequence information. De novo assemblies for each sample generated an average of 37 scaffolds, 76 scaffolded contigs and a size of 9.35 Mb. This represents an average sequence coverage of 83-fold. The average GC content (41.1 %) is in accordance with the findings of Küberl et al. [23], who reported the first genome of *P. pastoris* CBS 7435. Detailed sequencing statistics for each strain can be found in Additional file 2: Table S2.

After the sequencing and assembly phase, a "contig-length vs. read-count" plot analysis was performed gaining deeper insights into the composition of the 31 samples. Figure 2 shows an example for one of the Low-producer genomes. In general, assembled contigs can be classified in three different groups. Group I contigs (lower 0.5×) represent low amounts of additional

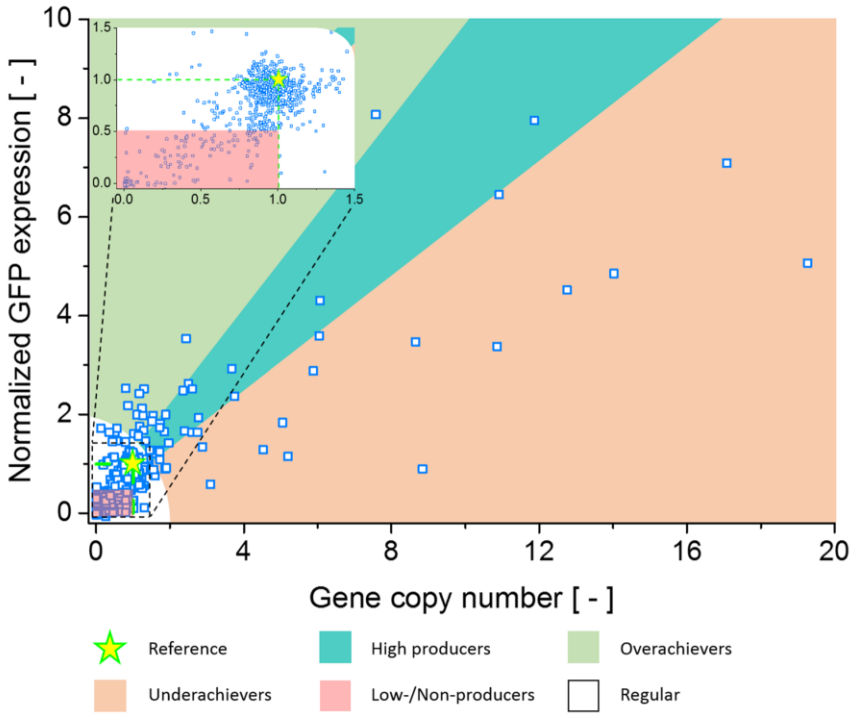

Fig. 1 GCN and normalized GFP expression of all 845 characterized *P. pastoris* clones. The reference clone JPS066 with GCN and GFP expression = 1 is shown with a *yellow star*. Areas for clones categorized as high producers, overachievers, underachievers or low-/non-producers (shown in the *cut-out*) are highlighted. Regular clones are found in the *white area*

Table 2 Grouping of all 190 irregular clones based on shared properties

Group	No. of clones	Characteristics
High producers	29	High GCN and high expression
Overachievers	21	Low GCN and high expression
Underachievers	40	High GCN and low expression
Low-/non-producers	100	Low/no GCN and expression

P. pastoris isolates in the sample, with small changes in comparison to the main isolate identified by a BLAST approach. The subpopulations of group I, essentially represent mixed-cultures. These sometimes contained low amounts of untransformed *P. pastoris* CBS 7435 (Δ*HIS4*) cells that presumably were supplied with L-histidine from the transformed cells. Contigs of group II (0.5 × to 2×) represent the almost complete chromosomal genome. The contigs of groups III (above 2×) were mostly allocated to the more abundant mitochondrial DNA and the most abundant DNA encoding ribosomal RNAs (rRNA) or other repetitive elements.

While the majority of sequenced clones contained no subpopulation, the occurrence of mixed-cultures with e.g. untransformed cells could have been reduced or completely eliminated by performing dilution plating experiments with transformed strains. If the mixed-cultures were the result of two or more cells adhering to each other after transformation they could form a single colony on the plate, containing both cell types. Dilution plating on suitable plates should separate such mixtures, whereby untransformed cells would be removed. The sequencing results, and especially the "contig-length vs. read-count" analysis, emphasize the necessity of this procedure for *P. pastoris* experiments. Using the high accuracy and sequencing depth of next generation sequencing (NGS) even small contaminations (<5 %) can be identified and culture heterogeneity better understood.

Based on a BLAST approach, the expression cassette was identified in 26 strains, but five strains lacked the *gfp* gene. The amount of contigs for each GFP cassette varies between five and seven contigs, whereas JPS495 stands out with 10 contigs for the cassette. Via in silico finishing, all gaps between cassette contigs were closed.

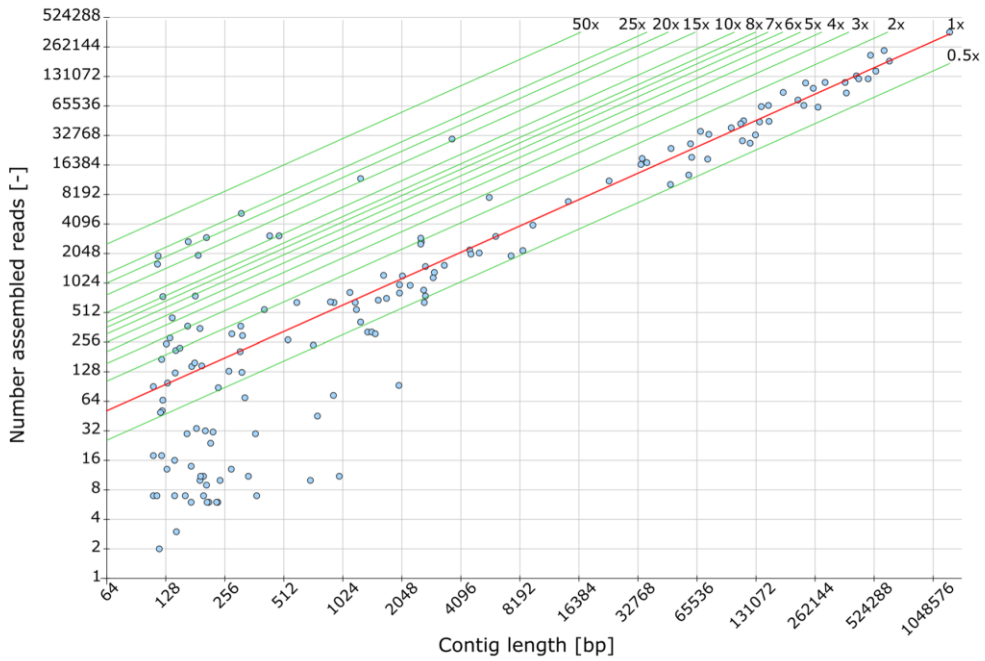

Fig. 2 "Contig length vs. read-count" plot for the Low-producer strain JPS379 (EMBL FBTF01000000) contigs in $\log_2$ scale. *Dots* represent the length of a contig plotted against the number of reads assembled into that contig. The *lines* denote the different predicted coverage ranges within the genome. Underrepresented contigs (group I) are found at a coverage of 0.5 or lower and indicate a secondary isolate in the sequenced sample. Unique contigs are represented by the *red line*, indicating a onefold coverage. Together with the twofold coverage line, they account for group II contigs which make up the chromosome of the main isolate. Above the twofold coverage line contigs belonging to group III can be found. They represent more abundant DNA elements, e.g. mitochondrial DNA

To this end, reads protruding contig ends were used to identify contigs that flank a certain source contig and were used to close the gaps between these contigs by applying CONSED [41]. Based on these results, further analysis was performed. By applying the "contig-length vs. read-count" plot analysis, the different amounts of inserted vector cassettes were calculated. Proportions between one inserted copy and about 20 copies were determined, as described in the methods section. Applying in silico finishing, four linear plasmids were closed and analyzed. Linear plasmids were identified by their left and right end. No additional reads were found that overlap at that position, whereas the coverage of the end contigs were often overrepresented in comparison to the chromosome. No connection to the chromosome and no chromosomal insertion side could be identified, therefore this sequences should represent linear plasmids. Possibly the linear plasmids are the result of expression cassettes being "looped out" from tandem arrays, resulting in instable GCN values for high copy *P. pastoris* clones [42], as illustrated in Additional file 3: Figure S2. Mainly such plasmids contain the vector cassette flanked by parts originating from the *E. coli* backbone of pAHBgl-GFP. The identified plasmids have a size of about 7–9 kb and include only 1–2 more genes on the non-vector cassette parts. On the other hand, it remains unclear how these linear plasmids would withstand degradation or segregational loss. Therefore more experiments are necessary to confirm their existence. In the following sections the various discovered integration events in sequenced *P. pastoris* strains are further discussed and the influence they might have on productivity analyzed.

Correlation between sequencing and experimental results

Using the data obtained from the sequenced *P. pastoris* clones multiple integration events were discovered (Fig. 3). In addition to the expected replacement of *AOX1* with the GFP cassette (Fig. 3a) the most common event were additional insertions of the expression cassette up- or downstream of the *AOX1* locus (Fig. 3b1). Often not only the cassette, but the complete pAHBgl–GFP vector, was found integrated into the genome hinting at either

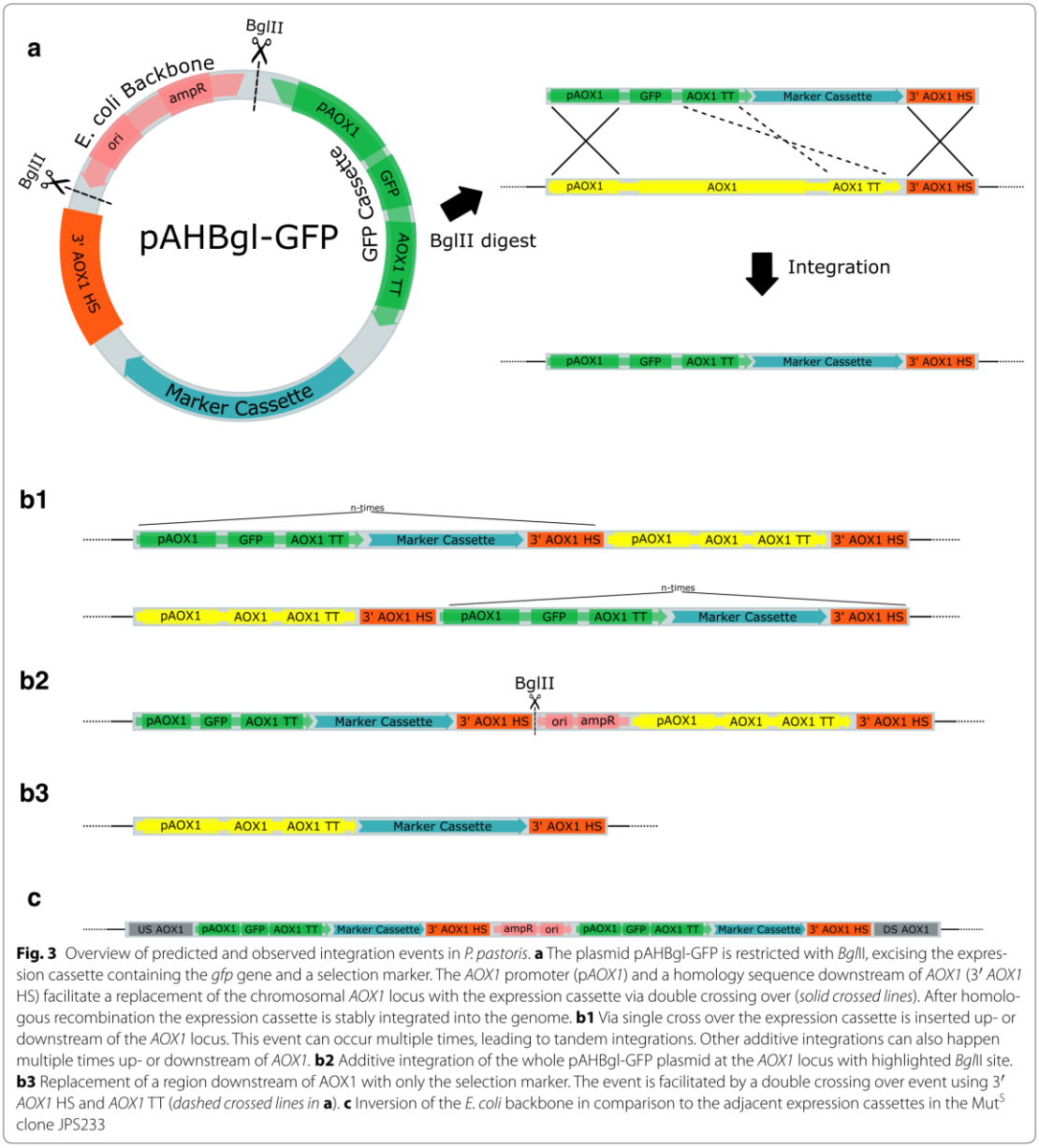

Fig. 3 Overview of predicted and observed integration events in *P. pastoris*. **a** The plasmid pAHBgl-GFP is restricted with *Bgl*II, excising the expression cassette containing the *gfp* gene and a selection marker. The *AOX1* promoter (p*AOX1*) and a homology sequence downstream of *AOX1* (3′ *AOX1* HS) facilitate a replacement of the chromosomal *AOX1* locus with the expression cassette via double crossing over (*solid crossed lines*). After homologous recombination the expression cassette is stably integrated into the genome. **b1** Via single cross over the expression cassette is inserted up- or downstream of the *AOX1* locus. This event can occur multiple times, leading to tandem integrations. Other additive integrations can also happen multiple times up- or downstream of *AOX1*. **b2** Additive integration of the whole pAHBgl-GFP plasmid at the *AOX1* locus with highlighted *Bgl*II site. **b3** Replacement of a region downstream of AOX1 with only the selection marker. The event is facilitated by a double crossing over event using 3′ *AOX1* HS and *AOX1* TT (*dashed crossed lines in* **a**). **c** Inversion of the *E. coli* backbone in comparison to the adjacent expression cassettes in the Mut^S clone JPS233

incomplete digestion or in vivo re-annealing of previously separated restriction fragments (Fig. 3b2). The in vivo re-annealing of fragments after digestion and prior to integration into the chromosome is supported by the orientation of the *E. coli* backbone found in e.g. the Mut^S clone JPS233 (Fig. 3c); *EMBL FBTV01000000*). Here, the elements of the *E. coli* backbone are inverted in comparison to the original vector (Fig. 3a), while the adjacent expression cassettes have the same orientation as located on the plasmid. The observed organization cannot be the result of an incompletely digested vector, but rather was caused by *P. pastoris* ligating an inversed *E. coli* backbone

to an expression cassette prior to integration. A gel purification step after plasmid digestion and prior to transformation would prevent the in vivo relegation between expression cassette and *E. coli* backbone.

Multiple tandem-integrations of the cassette or the complete vector were found. This is the first genome-sequence based evidence for the prevalence of tandem-integrations as the genetic organization in multi-copy clones of *P. pastoris*, first postulated by Clare et al. [43]. No dispersed distribution of cassettes in multiple different loci on the chromosomes was discovered in the analyzed multi-copy strains. In the majority of sequenced clones with multiple tandem integrations, the head-to-tail order of cassettes was the only organization found. Tail-to-tail or head-to-head sequences were only present in five of the sequenced strains, and in all but one they occurred in equal quantities. The different cassette orders are shown in Fig. 4. Collectively, in 86 cases head-to-tail was encountered, tail-to-tail 14 times and head-to-head 15 times. The exclusivity of head-to-tail in comparison to the inclusivity of head-to-head with tail-to-tail hints at an "either/or" relationship between two main integration mechanisms in *P. pastoris*. Head-to-tail tandem insertions likely integrated via the mechanism described by Clare et al. [43]. A different path of integration was probably responsible for head-to-head and tail-to-tail insertions. Multiple adjacent head-to-head and tail-to-tail integrations would culminate in the observed equilibrium between both arrangements. They could be the result of consecutive integrations of singular expression cassettes, each one using the previous one has homology sequence for additive integration. Only in strain JPS300 (*EMBL FBTO01000000*), head-to-tail as well as head-to-head was found. How these paths differ and why they are seemingly exclusive to one another remains unclear. However, in the following sections the observed correlation between organization of cassettes and their GFP expression are discussed. Furthermore, other findings of the genome sequencing data are correlated with the experimental results for the different groups of irregular clones.

High producers

Strains in the high producer group displayed a markedly higher GCN and expression level than the reference clone. Multi-copy clones with up to 12 (±1.4) copies of the *gfp* gene and expression levels up to eightfold (±1.2) higher than the reference clone are present in this group. In addition, a linear correlation between GCN and GFP expression was apparent. In Fig. 5, a good agreement with $R^2 = 0.83$ between experimental data is shown. The correlation found here is in accordance with previous reports for the relation between GCN and intracellular protein expression in *P. pastoris* [37, 38]. Notably, the

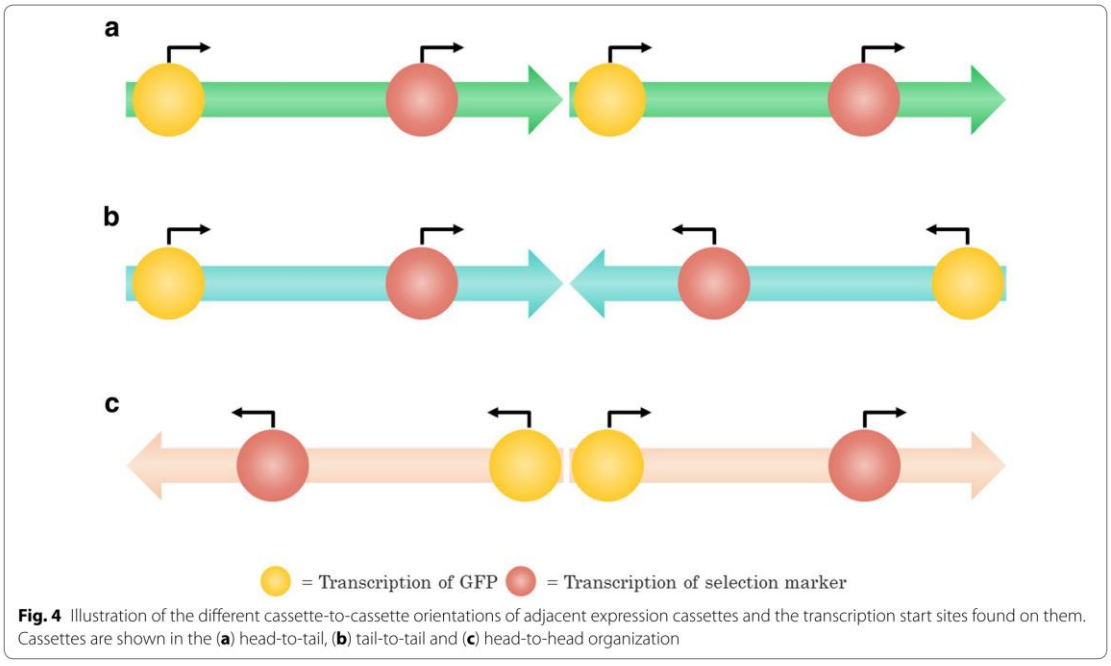

Fig. 4 Illustration of the different cassette-to-cassette orientations of adjacent expression cassettes and the transcription start sites found on them. Cassettes are shown in the (**a**) head-to-tail, (**b**) tail-to-tail and (**c**) head-to-head organization

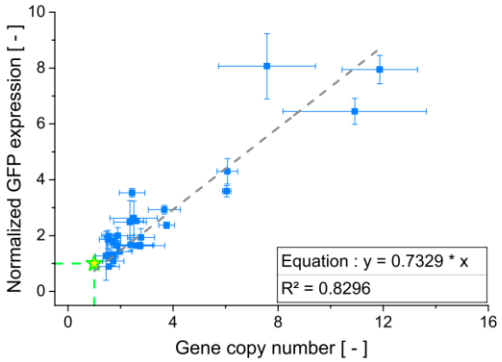

Fig. 5 Experimental values for GCN and GFP expression as well as linear regression for clones belonging to the high producers. The reference clone JPS066 is shown as a *green star*. *Error bars* display the standard deviation, with n = 3 × 3

slope of the regression line (0.73) is below 1. At a slope of 1 two copies would produce twice the amount of GFP. Especially clones with a high copy number are likely the cause of this lower incline. High copy clones put more stress on the protein synthesis apparatus of the cell, thereby diminishing the productivity per GFP cassette and lowering the slope of the linear regression curve.

In total, six clones from this group were selected for genome sequencing. As was expected based on the qPCR experiments multiple copies of the cassette could be detected. Based on the read frequency, up to 15 copies of the *gfp* gene were present. In addition to the GFP cassette, in some cases the *E. coli* backbone of pAHBgl–GFP was additionally inserted in-between two cassettes or directly adjacent to another one. All integration events occurred at the *AOX1* locus, even in Mut$^+$ strains. While it can not be excluded that unsequenced high producer clones contain integrations at other loci, it appears no off-target integrations occurred in this group. Since high producers are most likely to be selected for production purposes, the lack of random gene disruptions increases their suitability for industrial applications.

Except clone JPS535 (*EMBL FBTY01000000*), only head-to-tail tandem integrations were found in high producer clones, suggesting that this organization might be beneficial for productivity. For gaining further insight, the differences seen in comparison to over- and underachiever clones in the next section have to be included.

Over- and underachievers

Individual clones assigned to one of these groups displayed either a markedly higher or lower GFP expression as expected based on the GCN. In total, eight over- and nine underachiever clones were selected for genome sequencing. The discrepancy between experimental and theoretical expression level (calculated using the GCN and the equation shown in Fig. 5) are clearly visible in Fig. 6.

For certain overachiever clones the observed GFP expression was more than twofold increased in comparison to the theoretical value. On the other hand, some underachiever strains showcased expression levels far lower than half the expected GFP expression level. Notably, in three sequenced overachiever clones the abundances of the *gfp* gene calculated from sequence coverage indicate the presence of two copies. These clones would therefore have an expression level in line with their GCN and belong to the group of high producer strains. Multiple explanations are possible as to why the GCN of the affected clones was consistently lower in the qPCR experiments. Potentially, the PCR efficiency in these clones differed too much from the reference, thereby producing inaccurate qPCR predictions. Alternatively, the physical organization of the *gfp* locus in the isolated genomic DNA (gDNA) hindered the annealing of primers during qPCR. Nevertheless, in five sequenced overachiever strains only a single copy of *gfp* was found.

For an in-depth evaluation of the cause for deviant production characteristics the orientation of the cassettes to one another was analyzed. In strains of the overachiever group only head-to-tail tandem integrations were found. On the other hand underachiever clones showed the highest proportion of head-to-head and tail-to-tail integrations with 30 % of all tandem integrations. In some

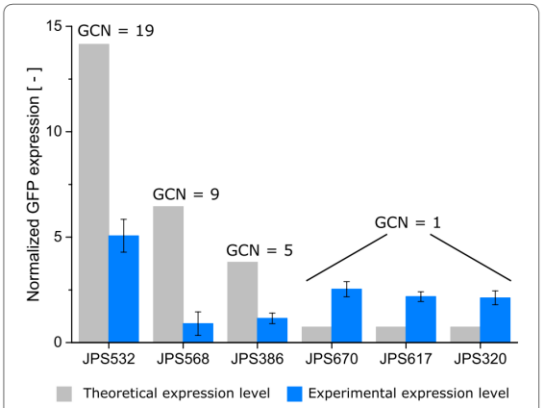

Fig. 6 Comparison between theoretical and experimental GFP expression for selected over—(JPS532, JPS568 and JPS386) and underachiever (JPS670, JPS617, JPS320) clones. The theoretical values were calculated using the GCN and the equation shown in Fig. 5. *Error bars* represent the standard deviation, with n = 3 × 3

underachiever clones a GCN >10 was found. It is possible that in these clones the decreased productivity was due to the high gene dosage triggering cytosolic proteases. The distribution of different cassette orientations among the groups of high producers, as well as under- and overachievers is shown in Fig. 7.

In combination with the findings described for the high producer group a trend is visible. Good production characteristics correlate with a head-to-tail orientation of adjacent cassettes. Head-to-tail organization ensures that all cassettes are read in the same direction. The lower GFP expression of strains with head-to-head and tail-to-tail integrations potentially was due to physical obstructions between RNA polymerases on adjacent cassettes during transcription. Crampton et al. [44] demonstrated via atomic force microscopy the premature stop of RNA polymerases on DNA with convergent promoters due to collision events. For the tail-to-tail orientation a "Head-on collision" of RNA polymerases seems less relevant. In this constellation only the marker cassettes are directly convergent with two 3′ *AOX1* HS (homology sequence downstream of *AOX1*, each 0.7 kb) separating them (Fig. 4b). However, in a head-to-head orientation the p*AOX1* of neighboring cassettes are directly adjacent to one another (Fig. 4c). This could culminate in mutual obstruction of RNA polymerases binding during the initiation of transcription, resulting in lower expression of the target protein.

In consequence, the construction of vectors that contain multiple expression cassettes in the head-to-tail orientation should lead to clones with an on average higher productivity than alternative approaches for the generation of multi-copy clones. Vassileva et al. [38] demonstrated the successful application of this strategy for the production of hepatitis B surface antigen, displaying a good correlation between GCN and product titer. If the transformation technique allows for other tandem integrations of cassettes, especially the head-to-head orientation, clones with a lower productivity are to be expected. This could be an explanation for some of the discrepancies found in other studies between GCN and expression level, if the orientation of expression cassettes was not determined [36].

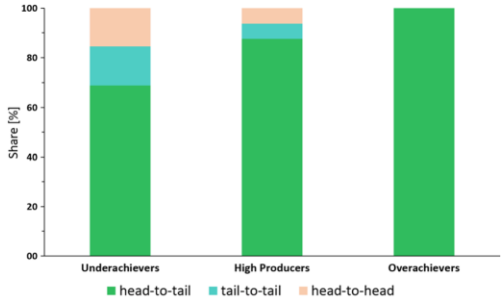

Fig. 7 Distribution of different cassette orientations among the sequenced multi-copy clones of the under—and overachievers as well as high producers

Low-/non-producers

Figure 1 displays a large number of clones producing only low amounts or no detectable GFP, the so called low- or non-producers. Six clones belonging to these groups were analyzed via genome sequencing. Many of these strains gave GCN values below 0.5 in qPCR experiments, indicating that a copy of the GFP cassette was present but the copy frequency was lower than that of the calibrator gene. Either the PCR efficiency for the *gfp* target was affected by e.g. the physical configuration of the GFP cassette or the mixed-culture phenomena caused a greater abundance of *ARG4* genes compared to *gfp* genes. With the mixed-culture phenomena the lower production, illustrated with the Low-producer strain JPS379 (*EMBL FBTF01000000*) in Fig. 2, can also be explained. A higher ratio of untransformed cells would imply that more cells in the culture media consume methanol without producing GFP. Thus, while GCN values <1 during qPCR experiments still indicate the presence of a full length expression cassette, they might also be used to identify clones with a lower productivity than is to be expected with GCN = 1. The findings of the sequencing experiments of four low-producer clones support this theory. Cultures of these strains always contained likely untransformed cells, as discovered by "contig-length vs. read-count"-plots discussed earlier.

In contrast to the low-producer clones, non-producer clones exhibited no detectable GFP expression and also no detectable GCN. Since they also grew normally in minimal medium without histidine, they were not untransformed cells. Genome sequencing of the non-producer strains JPS056 and JPS060 (*EMBL FBTH01000000* and *FBVQ01000000*) revealed that they shared the same genotype. Only the marker cassette region of pAHBgl-GFP had integrated in the *AOX1* locus (Fig. 3b3). This insertion was most likely facilitated by a double crossover homologous recombination between the *AOX1* terminator *(AOX1* TT) and the 3′ *AOX1* HS regions on the expression cassette and the *AOX1* chromosomal locus. As shown in Fig. 3a with dashed lines, *AOX1* TT presents a third homology sequence in addition to the two used for replacement of *AOX1*. In total, 68 clones exhibited the same phenotype (GCN and GFP expression level = 0) as JPS056 and JPS060. Therefore, this secondary homologous recombination event appeared to occur quite frequently. As a consequence of the insertion via *AOX1* TT

and 3′ *AOX1* HS, an uncharacterized gene downstream of *AOX1*, the 810 bp long *YDR514C*, is disrupted. At the 5′ end of Y*DR514C,* 165 bp are homologous to the *AOX1* TT, while at the 3′ end 509 bp are homologous to 3′ *AOX1* HS. Thus, base 510 to 645 of *YDR514C* are replaced with the marker cassette during this secondary homologous recombination event, which can be seen in the genome sequences of clone JPS056 and JPS060. Using this information, other clones with similar phenotype were analyzed via PCR for the integrity of *YDR514C*. A total of 64 strains with the same genotype as discovered in JPS056 and JPS060 could be identified. Thus, such false-positive clones accounted for approximately 8 % of all clones. They bypass the selection process and unnecessarily increase the workload e.g. in studies in which every clone is of interest. The irregular integration event seemed to depend on the close proximity of *AOX1* TT and 3′ *AOX1* HS on the expression cassette allowing insertion of the marker cassette without the target gene via homologous recombination. Resulting clones can grow in selective media and show the Mut^+ phenotype, but cannot produce the target gene. For further analysis of this phenomenon a novel vector was constructed and transformed into *P. pastoris*.

An optimized vector to prevent false-positive clones

Based on the findings for non-producer clones a variant of pAHBgl-GFP was constructed, aiming to prevent the creation of false-positive clones due to the integration of only the marker cassette. Since the erroneous integration event was assumed to be facilitated by a homologous recombination between the *AOX1* TT and 3′ *AOX1* HS regions, they were the key optimization targets. By replacing the terminator and leaving the 3′ *AOX1* HS unchanged, the core functionality of the vector would remain unaltered, thus this strategy was chosen. The *CYC1* (cytochrome c iso-1) terminator (*CYC1* TT) from *S. cerevisiae* was selected as replacement for multiple reasons: It shows no sequence similarities to the genome of *P. pastoris* CBS 7435 and is a well-studied, widely used terminator in *S. cerevisiae* [45, 46]. Additionally, *CYC1* TT has been used before as part of *P. pastoris* vectors, albeit not in combination with p*AOX1* [47, 48]. The resulting vector was named pAHBgl–GFP-CYC.

After transformation with pAHBgl-GFP-CYC, 120 clones were picked and characterized for their GFP expression using the same methods as described above. Their expression level was normalized to the same reference clone JPS066. In all 120 clones GFP expression was detectable. Using the Pearson's Chi squared test, it was determined that the lack of false-positive clones was significant (Table 3).

Table 3 Occurrence of false-positive clones when using pAHBgl-GFP compared to pAHBgl-GFP-CYC

Vector	Total clones	False-positive clones (no./ %)	X^2 value	p value
pAHBgl-GFP	845	64/8	–	–
pAHBgl-GFP-CYC	120	0/0	10.91	0.0001–0.001

The absence of false-positive clones strongly suggests that the integration of the marker cassette found in JPS056 and JPS060 was indeed mediated by a double crossing over event using the *AOX1* TT and 3′ *AOX1* HS regions. Multiple commercial and non-commercial vectors for *P. pastoris* are targeted for *AOX1* replacement, while putting the gene of interest under the control of p*AOX1* and *AOX1* TT [49, 50]. Integration of only the marker cassette by the mechanism described here can potentially also occur using these plasmids. Therefore it would seem advisable to switch to a different terminator for the gene of interest in order to prevent an increased workload for finding the producer strains, due to false-positive clones.

It has to be noted, however, that pAHBgl-GFP-CYC strains produced markedly lower amounts of GFP than pAHBgl-GFP clones on average (Additional file 3: Figure S3). Likely *CYC1* TT is not strong enough as a terminator for the exceptionally strong p*AOX1*. Thereby, faulty transcription termination and inactive gene products occur. For high-level production, a different terminator ought to be used. In a recent study multiple terminators (and other regulatory elements) from *P. pastoris* were characterized, providing a good starting point for finding a more effective replacement terminator [51].

Conclusion

Multiple unexpected integration events were discovered during genome sequencing and correlated with the production characteristics of the clones. By analyzing the connection between genome sequence and classic characterization experiments, many novel insights were obtained. The findings demonstrate that the combination of both methods enables deeper understanding than using them separately. Previously postulated theories regarding the generation of multi-copy tandem head-to-tail integrations and in vivo ligation events prior to integration could be verified [43]. It was found that the head-to-tail modus is the dominant insertion pathway, markedly outweighing head-to-head and tail-to-tail integrations. Both pathways seem to be exclusive to one another. The data also suggests that head-to-head and tail-to-tail integrations have a negative impact

on productivity. A likely cause is the close proximity of p*AOX1* of neighboring head-to-head cassettes. As a result RNA polymerases obstruct each other during transcription. Therefore it seems advisable to use methods specifically generating head-to-tail multi-copy clones, if aiming to increase the product titer via the gene dosage.

In some sequenced strains the presence of multiple genotypes in the form of a mixed-culture was observed, sometimes containing untransformed cells likely provided with L-histidine by transformed cells. Using dilution plating procedures after transformation should eliminate most, if not all, of the mixed-cultures containing untransformed cells. Employing antibiotics like Zeocin for selection ought to reduce the risk of such contaminations as well. The discovery of these subpopulations via genome sequencing supports the validity of the dilution plating procedure, often used in experiments involving yeast.

A secondary double crossing over event using *AOX1* TT and the 3′ *AOX1* HS led to the integration of only the marker cassette and the creation of false positive clones in about 8 % of all clones. Such clones result in an increased workload when assaying transformed cells for their productivity. By replacing *AOX1* TT with the non-homologous *CYC1* TT, we could show that no more false-positive clones occurred after transformation. Thereby underlining the validity of theories derived from correlating experimental and genome sequencing results. However, productivity was markedly lower, likely caused by inefficient transcription termination, suggesting that a more suitable terminator needs to be implemented.

Notably, the expression cassette was always found at the *AOX1* locus in the analyzed clones, for both Mut^S and Mut^+ strains. Especially for high producers, the best suited strains for industrial applications, the apparent absence of integrations at other loci is desirable. Random integrations as a result of NHEJ have been described for *P. pastoris*. In the present study a vector system was used, designed to prevent off-target integration [5]. Additionally, the selection process for sequencing emphasized productivity characteristics. It is therefore possible that random integrations at other loci were overlooked as they had no or only a small impact on the expression of GFP. The majority of clones had a GCN and expression level of ca. 1. Potentially, many of these clones harbor an expression cassette integrated at a locus other than *AOX1*, which did not affect productivity.

Methods

Microorganisms and cultivation conditions

Escherichia coli KRX (Promega, USA) was used for plasmid construction and propagation work. KRX was cultivated in LB (Lysogeny Broth) medium supplemented with 100 µg/mL ampicillin. For experiments involving *P. pastoris* CBS 7435 (Δ*HIS4*), obtained from ACIB (Austrian Center of Industrial Biotechnology, Austria) as well as the wild type CBS 7435 (CECT 11047 at Spanish Type Culture Collection, Spain), were used. *Saccharomyces cerevisiae* wild type strain LBG H620 was provided by the Institute for Agricultural Bacteriology and Fermentation Biology, ETH Zurich, Switzerland. Yeast shake flask cultivations were carried out in BMD (Buffered Minimal Dextrose) [49] or YPD (Yeast Peptone Dextrose) medium, supplemented with 4 mg/L L-histidine when necessary. Experiments in 96-deep-well plates with 2.4 mL total volume (Eppendorf, Germany) used BMD, BMM2 (Buffered Minimal Methanol) and BMM10 as previously described by Weis et al. [30] and Hartner et al. [31]. In brief, BMD is used for the growth phase while BMM2 and BMM10 induce expression of the target gene by maintaining a 0.5 % (*v*/*v*) methanol content in the culture medium. The 96-deep-well plates contained up to 500 µL of culture media, were sealed with sterile Breathseal film (Greiner, Germany) and were shaken at 340 rpm at 28 °C.

Plasmid construction and transformation

Primers were designed using SnapGene (GSL Biotech, USA). The sequences of all primers used in this study can be found in Additional file 4: Table S3. In order to construct a vector for intracellular expression of GFP*uv* in *P. pastoris*, the plasmid pAHBgl from ACIB, Austria was used as the basis vector. pAHBgl allows for selection based on complementation of the histidine auxotrophy, lacks a secretion signal and can be used for ends-out insertion via linearization with *Bgl*II prior to transformation [5]. Using the Gibson assembly technique [52], the *gfpuv* gene was amplified via PCR from the plasmid pBAD-GFP*uv* [29] and inserted into linearized pAHBgl, resulting in pAHBgl–GFP. Similarly the *AOX1* TT of pAHBgl–GFP was replaced by means of Gibson assembly with the *CYC1* TT amplified from gDNA of *S. cerevisiae* LBG H620, creating pAHBgl-GFP-CYC.

pAHBgl-GFP and pAHBgl-GFP-CYC were amplified in *E. coli* KRX. For transformation into *P. pastoris* CBS 7435 (Δ*HIS4*) the plasmids were extracted with the Wizard® Plus SV Minipreps DNA Purification System (Promega, USA). Purified plasmids were digested with *Bgl*II to facilitate ends-out insertion of the GFP expression cassette targeted for replacement of the native *AOX1*. *P. pastoris* CBS 7435 (Δ*HIS4*) was transformed according to Wu and Letchworth [53], with 2–3 µg of digested plasmid DNA per transformation. Cells were spread immediately after transformation onto MD (Minimal Dextrose) plates [49] and incubated at 28 °C for 3–4 days before picking clones. The transformants were both used for the following characterization experiments and stored at −80 °C in 12.5 % (*w*/*v*) glycerol to serve as a master strain bank for later analysis.

Characterization of *P. pastoris* clones

For determination of the Mut-phenotype the plating test as described in the EasySelect™ *Pichia* Expression Kit Manual (Invitrogen, USA) was used, employing MD and MM (minimal methanol) plates [49].

GFP expression was assayed in a 96-deep-well plate format using established protocols for *P. pastoris* [30, 31]. Clones were always cultivated in triplicate per plate, with clones belonging to high interest groups being cultivated on two additional deep-well plates. In order to normalize the GFP expression values of each clone, independent of the plate and experiment batch a reference clone had to be selected. The reference clone (JPS066) is a Mut^S, single-copy clone chosen from among the first 100 clones in a preliminary experiment. JPS066 exhibited a GFP fluorescence level closest to the mean of all single-copy clones of that test. Therefore the GFP fluorescence per OD_{600} of the reference clone was set as 1 and all other clones normalized to it. In the following experiments the reference clone as well as the untransformed CBS 7435 (Δ*HIS4*) were always cultivated on each deep-well plate for normalization. Eq. (1) was used for calculating the GFP expression level of each strain. Both GFP fluorescence (excitation 390 nm, emission 510 nm) and OD_{600} were measured using a SPECTRAFluor Plus microplate reader (Tecan, Switzerland). The value for the GFP expression level in the results and discussion section always represents the normalized expression level 60 h after the start of the methanol induction.

$$GFP_X = \frac{\frac{RFU_X - RFU_B}{OD_X - OD_B}}{\frac{RFU_R - RFU_B}{OD_R - OD_B}} \quad (1)$$

where GFP_X is the normalized GFP expression level of clone X, $RFU_{X/R/B}$ the relative fluorescence units of GFP for clone X, the reference clone R or the blank [CBS 7435 (Δ*HIS4*)] and $OD_{X/R/B}$ is the OD_{600} value of clone X, the reference clone R or the blank (medium), respectively.

gDNA was extracted from *P. pastoris* using the MasterPure™ Yeast DNA Purification Kit (Epicentre, USA). The GCN was determined in qPCR experiments in technical duplicates according to previously reported methods [32] using the Rotor-Gene SYBR® Green PCR Kit (Qiagen, Germany) and a LightCycler® 96 system (Roche, Switzerland). Irregular clones were assayed in two additional biological replicates, each with technical duplicates. In brief, the GCN of the *gfp* gene was assayed using relative quantification based on the $2^{-\Delta\Delta Ct}$ method [54]. The single copy gene *ARG4* was chosen as calibrator gene. The primers designed for qPCR exhibited similar T_M-values (59 ± 1 °C) and amplicon sizes (100 ± 1 bp). Their nucleotide sequences can be found in Additional file 4: Table S3. Using gDNA from the reference clone JPS066 in quadruple determination over 3.5 logs of copy quantity, a calibration curve for the *ARG4* and *gfp* target was created to determine the PCR-efficiencies for each target (Additional file 3: Figure S1). These PCR-Efficiencies were used for the following assays. Approximately 1 ng gDNA was used per qPCR reaction.

Sequencing of 31 selected *P. pastoris* strains and genome assembly

gDNA of 31 selected *P. pastoris* strains was isolated for high throughput sequencing as described above. The quality of the DNA was assessed by gel-electrophoresis and the quantity was estimated by a fluorescence-based method using the Quant-iT PicoGreen dsDNA kit (Invitrogen, USA) and the Tecan Infinite 200 Microplate Reader (Tecan, Switzerland).

For sequencing of the *P. pastoris* strains, paired-end sequencing libraries (TruSeq sample preparation kit; Illumina, USA) were constructed according to the manufacturer's protocol. The genome sequences of the *P. pastoris* strains were established on the Illumina MiSeq system by two paired-end sequencing runs (2 × 300 bp) with a distance range of about 500 bp. Upon sequencing and processing of the raw data, *de novo* assemblies were performed using the GS *De Novo* Assembler, software release version 2.8. (Roche, Switzerland) with default settings. The assembled draft sequences for the *P. pastoris* genomes were deposited in the EMBL-EBI database under the study id *PRJEB12220*.

Bioinformatic analyses of the 31 Pichia pastoris strains

First insights into the quality of the assemblies were provided by a "Contig-length vs. read-count plot" analysis [55–57]. In the plot the length of each contig (x-axis) is assessed against the number of assembled reads from the corresponding contig (y-axis). The values of lines representing the contig coverage were calculated from the contig length and the number of assembled reads of contigs longer than 10,000 bases, because the probability for these is higher than that for shorter contigs. The contig length distribution covers several orders of magnitude, so a double logarithmic scale was chosen for the axes. Both axes are logarithmic to the base two. This way even very long contigs can be presented clearly. For the calculation of the coverage lines also a double-logarithmic form was used.

Using the readcount and length values the relative abundance of contigs (e.g. the expression cassette) was calculated. With Eq. (2) and the median raw coverage (C) of all contigs longer than 10,000 bp the normalized coverage of a contig [$cov(c_i)$] can be determined. Contigs longer than 10,000 bp are expected to have a single fold coverage.

$$cov(c_i) = \frac{readcount(c_i)}{length(c_i) \cdot C} \quad (2)$$

$cov(c_i)$ values between 0.5 and 1.5 indicate that the corresponding contigs are represented once in the genome; ratios lower than 0.5 indicate underrepresented contigs and ratios higher 1.5 indicate overrepresented contigs (e.g. a multi-copy clone).

For more insight, genomic contigs resulting after the assembly were analyzed for large local similarities applying the BLASTn algorithm [58]. Each contig was compared to a local database including the pAHBgl-GFP vector sequence. Hits with an e-value $>1 \times 10^{-20}$ and a sequence identity of 100 % were analyzed in detail. In the first step, an in silico based finishing approach was used to close the gaps of vector sequence [23, 59, 60]. In the next step, the raw sequence coverage of the vector sequence was calculated and normalized based on "contig-length vs. read-count plot" analysis to detect the amount of inserted units. In addition, using the in silico based finishing approach the insertion site of each vector into the *P. pastoris* genome was identified. Annotation of relevant contigs from sequenced clones was done using SnapGene (GSL Biotech, USA).

Additional files

Additional file 1. Criteria used for sorting *P. pastoris* strains into different groups of "regular" and "irregular" clones, as well as scoring system for selection of clones for sequencing.

Additional file 2. Detailed genome sequencing statistics for all sequenced clones.

Additional file 3. Additional Figure 1 (qPCR calibration curve), Figure 2 (looping-out events) and Figure 3 (Comparison of GFP expression between pAHBgl-GFP and pAHBgl-GFP-CYC clones).

Additional file 4. Primer sequences.

Abbreviations

AOX1/2: alcohol oxidase 1/2; p*AOX1*: alcohol oxidase 1 promoter; *AOX1* TT: alcohol oxidase 1 transcription terminator; 3′ *AOX1* HS: homology sequence downstream of *AOX1*; BMD: buffered minimal dextrose; BMM: buffered minimal methanol; *CYC1* TT: cytochrome c iso-1 transcription terminator; GCN: gene copy number; gDNA: genomic DNA; LB: lysogeny broth; MD: minimal dextrose; MM: minimal methanol; $Mut^{S/+}$: methanol utilization slow/plus; NHEJ: non-homologous end-joining; YPD: yeast peptone dextrose.

Authors' contributions

JPS, JK and KF designed, analyzed and interpreted wet lab experiments. JPS and TL performed wet lab experiments. AW performed genome sequencing work. DW analyzed and interpreted sequencing data. JPS and DW wrote the manuscript. JK and KF revised the manuscript. JPS, JK and KF conceived the study. JK and KF supervised the research. All authors read and approved the final manuscript.

Author details

[1] Fermentation Engineering, Bielefeld University, Universitätsstr. 25, 33615 Bielefeld, Germany. [2] Microbial Genomics and Biotechnology, Center for Biotechnology (CeBiTec), Bielefeld University, Universitätsstr. 27, 33615 Bielefeld, Germany. [3] Genome Research of Industrial Microorganisms, CeBiTec, Bielefeld University, Universitätsstr. 27, 33615 Bielefeld, Germany. [4] CeBiTec, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany.

Acknowledgements

Grants from the Federal State of North Rhine-Westphalia for the CLIB-Graduate Cluster Industrial Biotechnology are gratefully acknowledged.

The bioinformatics support of the BMBF-funded project "Bielefeld-Gießen Center for Microbial Bioinformatics—BiGi (Grant number 031A533)" within the German Network for Bioinformatics Infrastructure (de.NBI) is gratefully acknowledged.

We acknowledge support for the Article Processing Charge by the Deutsche Forschungsgemeinschaft and the Open Access Publication Fund of Bielefeld University.

Competing interests

The authors declare that they have no competing interests.

Received: 9 February 2016 Accepted: 11 May 2016
Published online: 20 May 2016

References

1. Gasser B, Prielhofer R, Marx H, Maurer M, Nocon J, Steiger M, Puxbaum V, Sauer M, Mattanovich D. *Pichia pastoris*: protein production host and model organism for biomedical research. Future Microbiol. 2013;8:191–208.
2. Cregg J, Cereghino J, Shi J, Higgins D. Recombinant protein expression in *Pichia pastoris*. Mol Biotechnol. 2000;16:23–52.
3. Jahic M, Rotticci-Mulder J, Martinelle M, Hult K, Enfors SO. Modeling of growth and energy metabolism of *Pichia pastoris* producing a fusion protein. Bioprocess Biosyst Eng. 2002;24:385–93.
4. Bill RM. Playing catch-up with *Escherichia coli*: using yeast to increase success rates in recombinant protein production experiments. Front Microbiol. 2014;5(March):85.
5. Ahmad M, Hirz M, Pichler H, Schwab H. Protein expression in *Pichia pastoris*: recent achievements and perspectives for heterologous protein production. Appl Microbiol Biotechnol. 2014;98:5301–17.
6. Macauley-Patrick S, Fazenda ML, McNeil B, Harvey LM. Heterologous protein production using the *Pichia pastoris* expression system. Yeast. 2005;22:249–70.
7. Bollok M, Resina D, Valero F, Ferrer P. Recent patents on the *Pichia pastoris* expression system: expanding the toolbox for recombinant protein production. Recent Pat Biotechnol. 2009;3:192–201.
8. Meehl MA, Stadheim TA. Biopharmaceutical discovery and production in yeast. Curr Opin Biotechnol. 2014;30:120–7.
9. Klinner U, Schäfer B. Genetic aspects of targeted insertion mutagenesis in yeasts. FEMS Microbiol Rev. 2004;28:201–23.
10. Krainer FW, Dietzsch C, Hajek T, Herwig C, Spadiut O, Glieder A. Recombinant protein expression in *Pichia pastoris* strains with an engineered methanol utilization pathway. Microb Cell Fact. 2012;11:22.
11. Pla IA, Damasceno LM, Vannelli T, Ritter G, Batt CA, Shuler ML. Evaluation of $Mut^{+}$and Mut^{S} *Pichia pastoris* phenotypes for high level extracellular scFv expression under feedback control of the methanol concentration. Biotechnol Prog. 2006;22:881–8.
12. Vogl T, Ruth C, Pitzer J, Kickenweiz T, Glieder A. Synthetic core promoters for *Pichia pastoris*. ACS Synth Biol. 2014;3:188–91.
13. Vogl T, Glieder A. Regulation of *Pichia pastoris* promoters and its consequences for protein production. N Biotechnol. 2013;30:385–404.
14. Lin-Cereghino GP, Godfrey L, de la Cruz BJ, Johnson S, Khuongsathiene S, Tolstorukov I, Yan M, Lin-Cereghino J, Veenhuis M, Subramani S, Cregg JM. Mxr1p, a key regulator of the methanol utilization pathway and peroxisomal genes in *Pichia pastoris*. Mol Cell Biol. 2006;26:883–97.
15. Parua PK, Ryan PM, Trang K, Young ET. *Pichia pastoris* 14-3-3 regulates transcriptional activity of the methanol inducible transcription factor Mxr1 by direct interaction. Mol Microbiol. 2012;85:282–98.
16. Kumar NV, Rangarajan PN. The zinc finger proteins Mxr1p and repressor of phosphoenolpyruvate carboxykinase (ROP) have the same DNA binding specificity but regulate methanol metabolism antagonistically in *Pichia pastoris*. J Biol Chem. 2012;287:34465–73.

17. Capone S, Horvat J, Herwig C, Spadiut O. Development of a mixed feed strategy for a recombinant *Pichia pastoris* strain producing with a de-repression promoter. Microb Cell Fact. 2015;14:101.
18. Chen Z, Sun H, Li P, He N, Zhu T, Li Y. Enhancement of the gene targeting efficiency of non-conventional yeasts by increasing genetic redundancy. PLoS One. 2013;8:e57952.
19. Näätsaari L, Mistlberger B, Ruth C, Hajek T, Hartner FS, Glieder A. Deletion of the *Pichia pastoris* KU70 homologue facilitates platform strain generation for gene expression and synthetic biology. PLoS One. 2012;7:e39720.
20. Saraya R, Krikken AM, Kiel JA, Baerends RJ, Veenhuis M, van der Klei IJ. Novel genetic tools for *Hansenula polymorpha*. FEMS Yeast Res. 2012;12:271–8.
21. Kretzschmar A, Otto C, Holz M, Werner S, Hübner L, Barth G. Increased homologous integration frequency in Yarrowia lipolytica strains defective in non-homologous end-joining. Curr Genet. 2013;59:63–72.
22. Kegel A, Martinez P, Carter SD, Åström SU. Genome wide distribution of illegitimate recombination events in *Kluyveromyces lactis*. Nucleic Acids Res. 2006;34:1633–45.
23. Küberl A, Schneider J, Thallinger GG, Anderl I, Wibberg D, Hajek T, Jaenicke S, Brinkrolf K, Goesmann A, Szczepanowski R, Pühler A, Schwab H, Glieder A, Pichler H. High-quality genome sequence of *Pichia pastoris* CBS7435. J Biotechnol. 2011;154:312–20.
24. De Schutter K, Lin YC, Tiels P, Van Hecke A, Glinka S, Weber-Lehmann J, Rouzé P, Van de Peer Y, Callewaert N. Genome sequence of the recombinant protein production host *Pichia pastoris*. Nat Biotechnol. 2009;27:561–6.
25. Stadlmayr G, Benakovitsch K, Gasser B, Mattanovich D, Sauer M. Genome-scale analysis of library sorting (GALibSo): isolation of secretion enhancing factors for recombinant protein production in *Pichia pastoris*. Biotechnol Bioeng. 2010;105:543–55.
26. Chung BKS, Selvarasu S, Camattari A, Ryu J, Lee H, Ahn J, Lee H, Lee D. Genome-scale metabolic reconstruction and in silico analysis of methylotrophic yeast *Pichia pastoris* for strain improvement. Microb Cell Fact. 2010;9:50.
27. Caspeta L, Shoaie S, Agren R, Nookaew I, Nielsen J. Genome-scale metabolic reconstructions of *Pichia stipitis* and *Pichia pastoris* and in silico evaluation of their potentials. BMC Syst Biol. 2012;6:24.
28. Sohn SB, Graf AB, Kim TY, Gasser B, Maurer M, Ferrer P, Mattanovich D, Lee SY. Genome-scale metabolic model of methylotrophic yeast *Pichia pastoris* and its use for in silico analysis of heterologous protein production. Biotechnol J. 2010;5:705–15.
29. Crameri A, Whitehorn E, Tate E, Stemmer WPC. Improved green fluorescent protein by molecular evolution using DNA Shuffling. Nat Biotechnol. 1996;14:315–9.
30. Weis R, Luiten R, Skranc W, Schwab H, Wubbolts M, Glieder A. Reliable high-throughput screening with *Pichia pastoris* by limiting yeast cell death phenomena. FEMS Yeast Res. 2004;5:179–89.
31. Hartner FS, Ruth C, Langenegger D, Johnson SN, Hyka P, Lin-Cereghino GP, Lin-Cereghino J, Kovar K, Cregg JM, Glieder A. Promoter library designed for fine-tuned gene expression in *Pichia pastoris*. Nucleic Acids Res. 2008;36:e76.
32. Abad S, Kitz K, Hörmann A, Schreiner U, Hartner FS, Glieder A. Real-time PCR-based determination of gene copy numbers in *Pichia pastoris*. Biotechnol J. 2010;5:413–20.
33. Cregg JM, Madden KR, Barringer KJ, Thill GP, Stillman CA. Functional characterization of the two alcohol oxidase genes from the yeast *Pichia pastoris*. Mol Cell Biol. 1989;9:1316–23.
34. Clare JJ, Romanos MA, Rayment FB, Rowedder JE, Smith MA, Payne MM, Sreekrishna K, Henwood CA. Production of mouse epidermal growth factor in yeast: high-level secretion using *Pichia pastoris* strains containing multiple gene copies. Gene. 1991;105:205–12.
35. Daly R, Hearn MTW. Expression of heterologous proteins in *Pichia pastoris*: a useful experimental tool in protein engineenring and production. J Mol Recognit. 2005;18:119–38.
36. Aw R, Polizzi KM. Can too many copies spoil the broth? Microb Cell Fact. 2013;12:128.
37. Marx H, Mecklenbräuker A, Gasser B, Sauer M, Mattanovich D. Directed gene copy number amplification in *Pichia pastoris* by vector integration into the ribosomal DNA locus. FEMS Yeast Res. 2009;9:1260–70.
38. Vassileva A, Chugh DA, Swaminathan S, Khanna N. Effect of copy number on the expression levels of hepatitis B surface antigen in the methylotrophic yeast *Pichia pastoris*. Protein Expr Purif. 2001;21:71–80.
39. Inan M, Aryasomayajula D, Sinha J, Meagher MM. Enhancement of protein secretion in *Pichia pastoris* by overexpression of protein disulfide isomerase. Biotechnol Bioeng. 2006;93:771–8.
40. Zhu T, Guo M, Tang Z, Zhang M, Zhuang Y, Chu J, Zhang S. Efficient generation of multi-copy strains for optimizing secretory expression of porcine insulin precursor in yeast *Pichia pastoris*. J Appl Microbiol. 2009;107:954–63.
41. Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. Genome Res. 1998;8:195–202.
42. Zhu T, Guo M, Sun C, Qian J, Zhuang Y, Chu J, Zhang S. A systematical investigation on the genetic stability of multi-copy *Pichia pastoris* strains. Biotechnol Lett. 2009;31:679–84.
43. Clare J, Rayment F, Sreekrishna K, Romanos MA. High-level expression of tetanus toxin fragment C in *Pichia pastoris* strains containing multiple tandem integrations of the gene. Nat Biotechnol. 1991;9:455–60.
44. Crampton N, Bonass WA, Kirkham J, Rivetti C, Thomson NH. Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy. Nucleic Acids Res. 2006;34:5416–25.
45. Mumberg D, Müller R, Funk M. Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. Genetics. 1995;156:119–22.
46. Zaret K, Sherman F. Mutationally altered 3′ ends of yeast CYC1 mRNA affect transcript stability and translational efficiency. J Mol Biol. 1984;176:107–35.
47. Stadlmayr G, Mecklenbräuker A, Rothmüller M, Maurer M, Sauer M, Mattanovich D, Gasser B. Identification and characterisation of novel *Pichia pastoris* promoters for heterologous protein production. J Biotechnol. 2010;150:519–29.
48. Prielhofer R, Maurer M, Klein J, Wenger J, Kiziak C, Gasser B, Mattanovich D. Induction without methanol: novel regulated promoters enable high-level expression in *Pichia pastoris*. Microb Cell Fact. 2013;12:5.
49. Invitrogen: EasySelect pichia expression kit. 2010:95.
50. Vogl T, Ahmad M, Krainer FW, Schwab H, Glieder A. Restriction site free cloning (RSFC) plasmid family for seamless, sequence independent cloning in *Pichia pastoris*. Microb Cell Fact. 2015;14:103.
51. Vogl T, Sturmberger L, Kickenweiz T, Wasmayer R, Schmid C, Hatzl AM, Gerstmann MA, Pitzer J, Wagner M, Thallinger GG, Geier M, Glieder A. A toolbox of diverse promoters related to methanol utilization—functionally verified parts for heterologous pathway expression in *Pichia pastoris*. ACS Synth Biol. 2016;5:172–86.
52. Gibson DG, Young L, Chuang R, Venter JC, Iii CAH, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods. 2009;6:343–5.
53. Wu S, Letchworth GJ. High efficiency transformation by electroporation of *Pichia pastoris* pretreated with lithium acetate and dithiothreitol. Biotechniques. 2004;36:152–4.
54. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods. 2001;25:402–8.
55. Heinl S, Wibberg D, Eikmeyer F, Szczepanowski R, Blom J, Linke B, Goesmann A, Grabherr R, Schwab H, Pühler A, Schlüter A. Insights into the completely annotated genome of *Lactobacillus buchneri* CD034, a strain isolated from stable grass silage. J Biotechnol. 2012;161:153–66.
56. Schwientek P, Szczepanowski R, Rückert C, Kalinowski J, Klein A, Selber K, Wehmeier UF, Stoye J, Pühler A. The complete genome sequence of the acarbose producer *Actinoplanes* sp. SE50/110. BMC Genomics. 2012;13:112.
57. Wibberg D, Jelonek L, Rupp O, Hennig M, Eikmeyer F, Goesmann A, Hartmann A, Borriss R, Grosch R, Pühler A, Schlüter A. Establishment and interpretation of the genome sequence of the phytopathogenic fungus *Rhizoctonia solani* AG1-IB isolate 7/3/14. J Biotechnol. 2013;167:142–55.
58. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
59. Maus I, Stantscheff R, Wibberg D, Stolze Y, Winkler A, Pühler A, König H, Schlüter A. Complete genome sequence of the methanogenic neotype strain *Methanobacterium formicicum* MFT. J Biotechnol. 2014;192:40–1.

60. Wibberg D, Blom J, Jaenicke S, Kollin F, Rupp O, Scharf B, Schneiker-Bekel S, Sczcepanowski R, Goesmann A, Setubal JC, Schmitt R, Pühler A, Schlüter A. Complete genome sequencing of *Agrobacterium* sp. H13-3, the former *Rhizobium lupini* H13-3, reveals a tripartite genome consisting of a circular and a linear chromosome and an accessory plasmid but lacking a tumor-inducing Ti-plasmid. J Biotechnol. 2011;155:50–62.

2.1.3 Supplemental data

Table S1 Supplemental Data - Schwarzhans et al. (2016) Microbial Cell Factories

Group	Criteria
High producer	GCN ≥ 1.5 AND ExpLvl/GCN ≥ 0.5
Overachiever	ExpLvl ≥ 1.0 AND GCN ≤ 1.5 AND ExpLvLl/GCN ≥ 1.5
Underachiever	GCN ≥ 1.5 AND ExpLvl/GCN ≤ 0.5 OR GCN ≥ 0.5 AND ExpLvl ≤ 0.5 AND ExpLvl/GCN ≤ 0.5
Low-/Non-producer	ExpLvl > 0.1 AND GCN ≤ 0.5 AND 0.5 ≤ ExpLvl/GCN ≤ 1.5 OR ExpLvl <0.1 AND GCN <0.1 AND Growth under selective conditions

Criteria/Points	**2**	**0**	**1**	**2**	**3**	**4**	**5**
GCN	< 0.5	0.5-1.4	1.5-3.0	3.1-4.4	4.5-6.0	6.1-7.4	≥ 7.5
ExpLvl	< 0.5	0.5-1.4	1.5-2.4	2.5-3.4	3.5-4.4	4.5-5.4	≥ 5.5
ExpLvl/GCN	< 0.5	0.5-1.4	1.5-2.0	2.1-3.0	3.1-4.0	4.1-5.0	≥5.1
Mut-Phenotype	-	Mut+	-	-	-	-	MutS

Comments
Clones that already were confirmed false-positive got a malus of -6
The range for the different criteria is based on the range of the corresponding value found for analyzed strains

Abbreviations

GCN	Gene Copy Number
ExpLvl	Normalized GFP expression level

Table S2 Supplemental Data - Schwarzhans et al. (2016) Microbial Cell Factories

	JPS014	JPS019	JPS056	JPS060	JPS072	JPS086	JPS087
Aligned Reads (All/PE)	1,490,918/742,073	3,777,116/1,869,958	2,942,397/1,469,649	3,031,122/1,513,722	2,206,477/1,094,219	1,841,379/915,965	3,507,907/1,752,435
Assembled Bases:	431783646	1085242059	731820926	813557554	636755661	537302104	840670703
PE-Size(s):	803 +/- 243	678 +/- 232	672 +/- 253	693 +/- 252	756 +/- 245	753 +/- 259	655 +/- 250
Scaffolds (All/True):	35/35	37/37	45/45	35/35	37/37	33/33	38/38
Contigs (S/L/A):	80/89/145	70/81/166	79/87/129	68/74/125	83/94/177	77/85/165	69/74/135
Bases in Scaffolds:	9354958	9357288	9357362	9358081	9357047	9362065	9356638
Coverage:	46.16	115.98	78.21	86.94	68.05	57.39	89.85
GC content (%):	41.06	41.06	41.07	41.07	41.07	41.06	41.06
Avg. Scaffold:	267284	252899	207941	267373	252893	283698	246227
N50 Scaffold:	878042	1384376	700434	1384662	877751	1404188	751087
Largest Scaffold:	1437898	2030893	1383988	2030991	2032349	2032133	1432184
Avg. Scaf. Contig:	116559	133420	118176	137306	112406	121211	135321
Avg. Contig:	104919	115453	107427	126303	99411	109919	126304
N50 Contig:	417447	546675	389918	437961	437906	501543	475534
Largest Contig:	1276501	1276507	1276506	1063037	1275592	1276502	937051

Table S2 Supplemental Data - Schwarzhans et al. (2016) Microbial Cell Factories

JPS094	JPS095	JPS194	JPS230	JSP233	JPS289	JPS300	JPS379
2,002,140/997,133	1,914,151/956,1173	524,019/1,745,014	1,797,090/892,900	1,798,640/891,247	2,738,184/1,358,725	4,114,506/2,037,693	3,166,844/1,571,427
583164108	502017127	1010121966	519714112	523571475	786590958	1193119978	909911866
739 +/- 261	720 +/- 259	674 +/- 231	779 +/- 244	737 +/- 257	719 +/- 244	706 +/- 238	705 +/- 242
31/31	45/45	38/38	33/33	31/31	32/32	38/38	33/33
67/84/159	82/98/151	72/86/164	72/84/127	69/91/598	82/89/154	75/80/155	71/84/151
9366612	9350451	9344511	9353532	9353799	9364780	9359558	9349790
62.26	53.69	108.1	55.56	55.97	83.99	127.48	97.32
41.06	41.06	41.08	41.07	41.06	41.07	41.06	41.07
302148	207787	245908	283440	301735	292649	246304	283326
1402906	529362	877252	878172	1347668	1384951	1389575	919282
2033202	1019891	2031675	2032154	1431089	2036364	2031083	2032518
139451	113728	129530	129541	135210	113840	124524	131373
111421	95351	108655	111218	102693	105000	116826	111250
453759	317582	504044	504138	503369	406130	475863	426132
1276503	1019891	1142259	1142301	1276502	1253611	1253611	1276506

Table S2 Supplemental Data - Schwarzhans et al. (2016) Microbial Cell Factories

JPS391	JPS394	JPS478	JPS495	JPS496	JPS500	JPS535	JPS536
1,922,145/952,493	2,263,322/1,124,271	1,787,459/887,604	3,980,241/1,976,350	2,961,104/1,467,834	2,525,225/1,254,673	2,817,080/1,395,088	1,976,046/977,790
553239744	654296034	519026174	1145075983	845403064	730467405	802657705	570488846
756 +/- 246	763 +/- 245	790 +/- 245	694 +/- 237	705 +/- 239	755 +/- 245	692 +/- 236	771 +/- 245
36/36	35/35	34/34	49/49	39/39	52/52	35/35	33/33
74/83/131	71/81/143	75/84/134	89/105/195	79/90/170	88/92/162	80/89/172	77/84/142
9355429	9357898	9360199	9349423	9358173	9347773	9353273	9359661
59.14	69.92	55.45	122.48	90.34	78.14	85.82	60.95
41.06	41.08	41.08	41.06	41.06	41.06	41.06	41.06
259873	267368	275299	190804	239953	179764	267236	283626
1384696	1384578	1384192	598460	1387215	477329	1385263	1384494
2172106	2172063	2031406	1403570	2032718	1194180	2031732	2032101
126089	131469	124419	104789	118146	105974	116594	121207
112578	115417	111215	89023	103879	101443	104950	111200
535908	591225	504031	291273	437957	289199	406130	475509
1276492	1276501	1282582	704226	1276502	678038	1276505	1276437

Table S2 Supplemental Data - Schwarzhans et al. (2016) Microbial Cell Factories

JPS568	JPS604	JPS664	JPS670	JPS682	JPS733	JPS749	JPS782
2,100,950/1,040,588	2,480,747/1,229,773	3,872,357/1,917,506	4,007,049/1,981,457	5,121,780/2,537,408	1,651,934/818,627	4,238,528/2,105,376	1,766,186/875,947
609920992	705197405	1116960618	1148503436	1480416492	478478744	1219211427	497801188
773 +/- 246	696 +/- 237	684 +/- 234	684 +/- 231	669 +/- 228	792 +/- 245	669 +/- 230	701 +/- 240
37/37	34/34	34/34	35/35	36/36	31/31	33/33	39/39
75/82/142	79/93/165	75/90/223	72/91/196	71/97/207	71/83/143	74/90/180	91/104/170
9355180	9350177	9343272	9330933	9330482	9354243	9345632	9350874
65.2	75.42	119.55	123.09	158.66	51.15	130.46	53.24
41.07	41.07	41.08	41.11	41.1	41.07	41.07	41.1
252842	275005	274802	266598	259180	301749	283200	239766
1384487	915390	1390064	915423	878121	1403183	1384235	584100
2031672	1448892	2031914	1448578	1447415	2032693	2174735	1448326
124398	118009	124293	129308	131147	131358	125995	102448
113896	100451	103785	102566	96264	112535	103794	89828
475889	373200	453821	425138	437961	503382	443745	273823
1275604	766927	1142284	1276456	1275597	1142342	1276506	624293

Supplemental Figures - Schwarzhans et al. (2016) Microbial Cell Factories

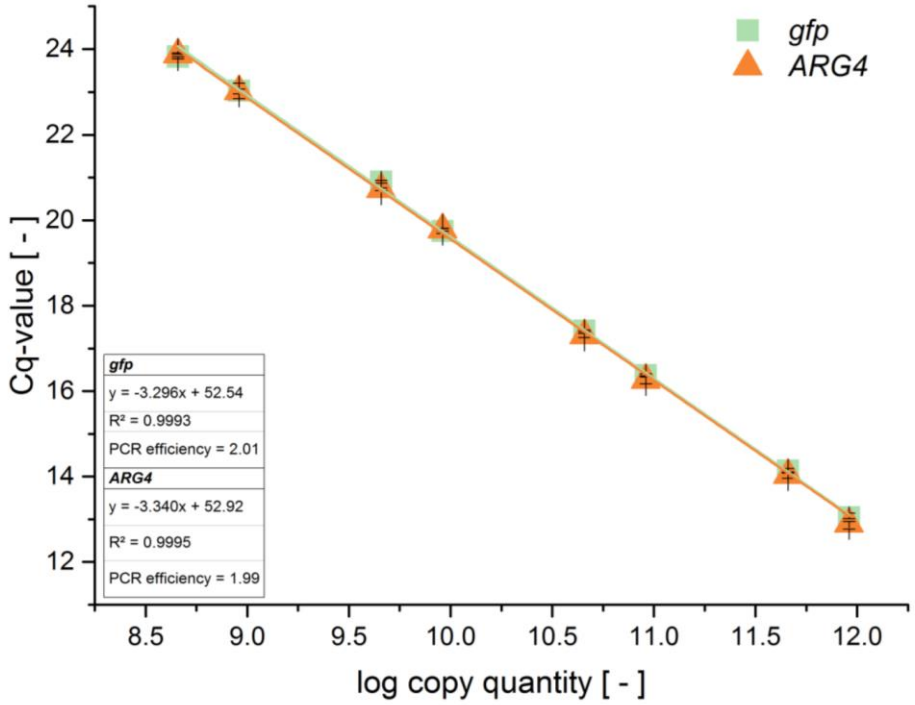

Fig. S1: qPCR calibration curve for the target (*gfp*) and the calibrator gene (*ARG4*). gDNA from the reference strain JPS066 was used as a template. The equation for the linear fit, R^2- and PCR efficiency values for both *gfp* and *ARG4* are shown in the integrated table. Error bars represent the standard deviation with n = 4

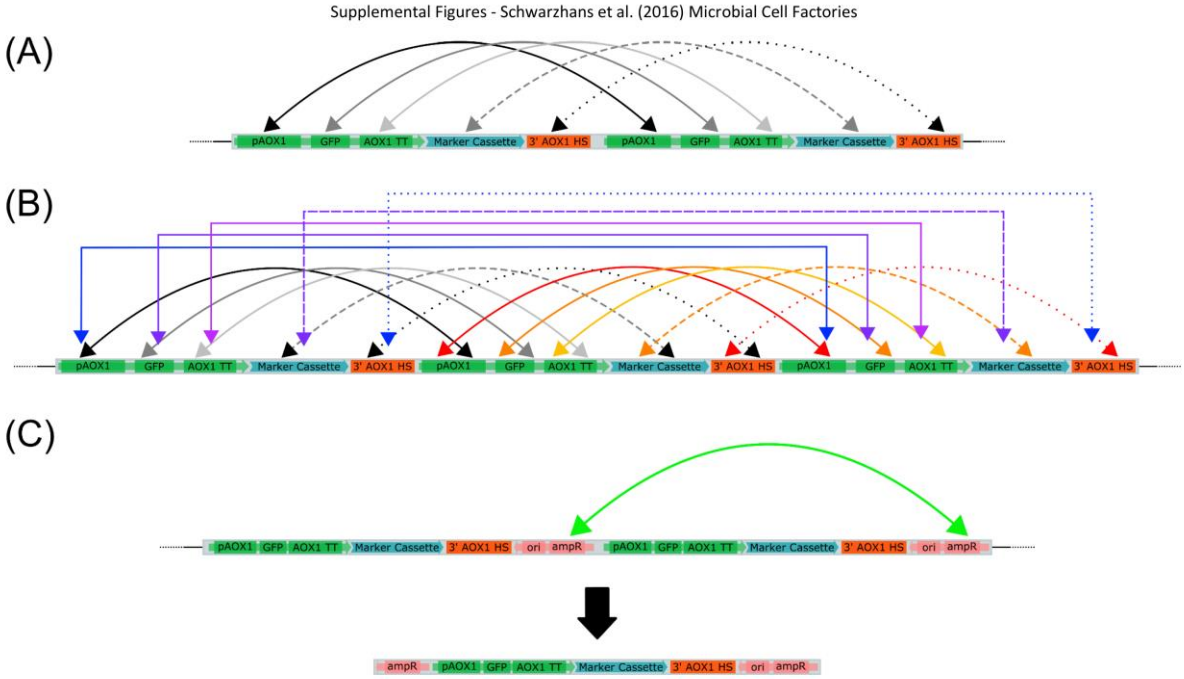

Fig. S2: Illustration of possible looping out recombination events between adjacent expression cassettes based on the vector design and a head-to-tail integration pattern. **(A)** Potential recombination events between two adjacent expression cassettes. **(B)** With three neighboring cassettes the amount of possible recombination events is increased threefold, compared to two cassettes. **(C)** Example loop out recombination and the resulting linear plasmid. The depicted reaction is based on the linear plasmid found during genome sequencing of strain JPS094 (**EMBL FBTE01000000**).

Supplemental Figures - Schwarzhans et al. (2016) Microbial Cell Factories

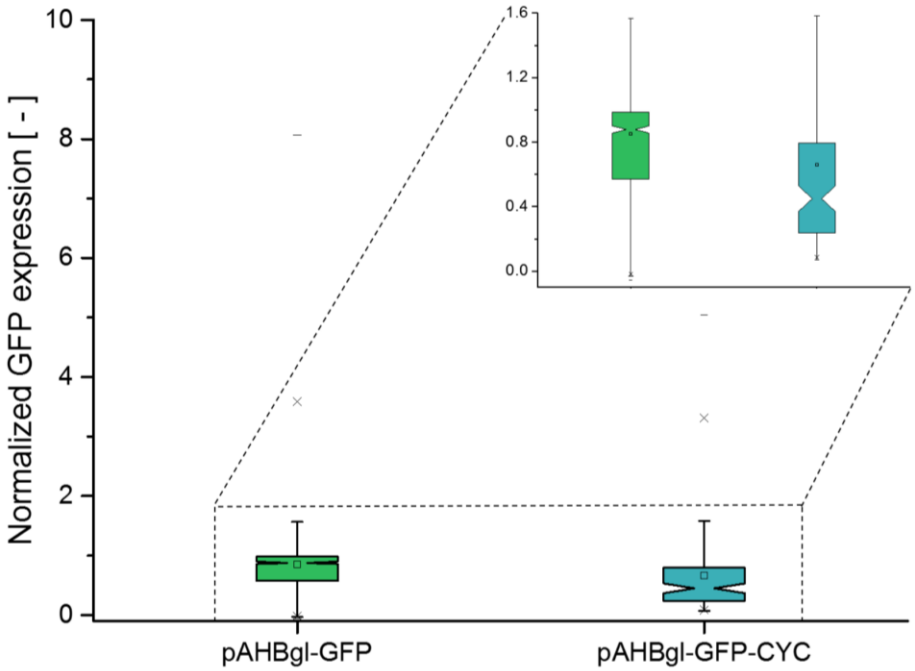

Fig. S3: Comparison of the normalized GFP expression of *P. pastoris* clones transformed either with pAHBgl-GFP or pAHBgl-GFP-CYC. 845 paHBgl-GFP and 120 pAHBgl-GFP-CYC strains were analyzed. The cut-out highlights the on-average lower expression level for clones transformed with pAHBgl-GFP-CYC in comparison to the original vector.

Table S3 Supplemental Data - Schwarzhans et al. (2016) Microbial Cell Factories

Primer	Sequence	Purpose
pAHBgl-Gib-FW	AGGCTTGGCGGCCGCAAACGGTTTTAG	Linearization of pAHBgl for Gibson Assembly
pAHBgl-Gib-RV	CGTTTCGAAGAATTCGTTGTTTTTTG	Linearization of pAHBgl for Gibson Assembly
GFP-Gib-FW	ACGAATTCTTCGAAACGATGGCTAGCAAAGGAGAAGAAC	Amplification of GFPuv with overhangs for Gibson Assembly
GFP-Gib-RV	GCGGCCGCCAAGCCTTTATTTGTAGAGCTCATCCATG	Amplification of GFPuv with overhangs for Gibson Assembly
pAHBgl-GFP-Gib-FW	GGAACGGAACGTATCTTAGCATGGTTGTGCGACAG	Linearization of pAHBgl-GFP for Gibson Assembly
pAHBgl-GFP-Gib-RV	CGTTTGCGGCCGCCAAGCCT	Linearization of pAHBgl-GFP for Gibson Assembly
CYC1-Gib-FW	AGGCTTGGCGGCCGCAAACGCCCCTTTTCCTTTGTCGAT	Amplification of CYC1TT with overhangs for Gibson Assembly
CYC1-Gib-RV	GCTAAGATACGTTCCGTTCCAGCTTGCAAATTAAAGCCTTC	Amplification of CYC1TT with overhangs for Gibson Assembly
GFP-qPCR-FW	GCGTGCTGAAGTCAAGTTTGAAGG	qPCR targeting GFPuv
GFP-qPCR-RV	CGAGTTTGTGTCCGAGAATGTTTCC	qPCR targeting GFPuv
ARG4-qPCR-FW	CTGACATACAACCGCATATCAGTGG	qPCR targeting ARG4
ARG4-qPCR-RV	TCGCTTGGGTGAGTTGATTGG	qPCR targeting ARG4
YDR514C-FW	GGTATTTCCCACTCCTCTTC	Amplification of YDR514C
YDR514C-RV	GAATGCGGGTATCACCATTC	Amplification of YDR514C

2.2 Non-canonical integration events

2.2.1 Motivation

During the evaluation of the plating assay, used for determining the Mut (Methanol utilization) phenotype, aberrant colony morphologies of certain strains became apparent in the first study. No direct correlation to the strains' productivity was observed, therefore the event was not discussed in the first publication. Nevertheless, the change in colony morphology captured our interest.

Initial investigations into *P. pastoris* literature suggested a connection of the aberrant colony morphology to the deletion of *OCH1* (α-1,6-mannosyltransferase). Knock-out of *OCH1* reduces hyper-mannosylation significantly in *P. pastoris*, but also leads to a crenulated colony morphology, increased temperature sensitivity, lower growth rates and increased flocculation in liquid culture [61, 98, 118]. Due to its involvement in hyper-mannosylation, *OCH1* became a key deletion target for glycoengineering *P. pastoris* [97]. This approach is complicated by the notoriously low targeting efficiencies for *OCH1* deletions, which typically are < 1% [139]. Because of the low frequency of correct integrations, in some cases the change in colony morphology was used for selection of correct clones after transformation [118]. Our hope was to find integration events that were related to *OCH1* or genes of similar function, thereby providing insights valuable for glycoengineering projects. Of course, other explanations for the observed alterations were also possible. Disruption of genes participating in central metabolic pathways or cell wall integrity can lead to decreased growth rates and a different colony appearance [140]. The potential scientific and industrial relevance urged us to further analyze clones with abnormal colony morphology.

However, from the perspective of genetic engineering, the exact event is less relevant than its occurrence in the first place. If a study aims to engineer *P. pastoris* strains with precise gene deletions or knock-ins, each off-target event is undesirable and a burden on the screening procedure. While the complexity of genetic engineering projects in *P. pastoris* has increased lately (see table 2 on page 70), the comparatively high clonal variability has hampered further advances. By discovering the integration event via genome sequencing, its mechanism of origin might be understood, and measures taken to prevent similar events from occurring in future experiments. Parallel knock-in of

foreign genes, coupled with knock-out, upregulation or downregulation of native genes has become a routine method in *S. cerevisiae*. Many techniques are available for multiplexed genetic engineering [45]. Optimized CRISPR/Cas9 methods were used for the simultaneous deletion of five genes with 100 % efficiency [141]. In a similar experiment, CRISPR/Cas9 was used for facilitating six genetic modifications (including foreign gene integration) in a single step with 50 - 100 % efficiency [142]. Contrary to this highly efficient and multiplexed strain construction in *S. cerevisiae*, genetic engineering of *P. pastoris* primarily relies on consecutive steps, introducing a single modification per step. An overview of established and newly developed strain engineering methods in *P. pastoris* is given in chapter 4 on page 61. The incremental engineering approach also raises the probability of aberrant integration events occurring. Therefore, off-target events as described here are especially relevant for such experiments.

In summary, the second part of the analysis of the strain library presented in chapter 2.1 concentrated on integration events relevant for genetic engineering studies, but less pertinent to the creation of high producer strains. Much of the ground work necessary for investigation had already been accomplished within the frame of the first publication. Further experiments concentrated on elucidating the novel and often complex integration events, and validating the effects of the ones with important implications for strain engineering. A detailed description of the experimental procedures, results and their implications is presented in the next chapter in the form of the publication “Non-canonical integration events in *Pichia pastoris* encountered during standard transformation analyzed with genome sequencing” (Schwarzhans *et al.*, (2016), Scientific Reports, [143]). The supplementary information of this publication is shown in chapter 2.2.3.

www.nature.com/scientificreports

OPEN

Received: 13 July 2016
Accepted: 15 November 2016
Published: 13 December 2016

Non-canonical integration events in *Pichia pastoris* encountered during standard transformation analysed with genome sequencing

Jan-Philipp Schwarzhans[1,2], Daniel Wibberg[3], Anika Winkler[2], Tobias Luttermann[1], Jörn Kalinowski[2] & Karl Friehs[1]

The non-conventional yeast *Pichia pastoris* is a popular host for recombinant protein production in scientific research and industry. Typically, the expression cassette is integrated into the genome *via* homologous recombination. Due to unknown integration events, a large clonal variability is often encountered consisting of clones with different productivities as well as aberrant morphological or growth characteristics. In this study, we analysed several clones with abnormal colony morphology and discovered unpredicted integration events *via* whole genome sequencing. These include (i) the relocation of the locus targeted for replacement to another chromosome (ii) co-integration of DNA from the *E. coli* plasmid host and (iii) the disruption of untargeted genes affecting colony morphology. Most of these events have not been reported so far in literature and present challenges for genetic engineering approaches in this yeast. Especially, the presence and independent activity of *E. coli* DNA elements in *P. pastoris* is of concern. In our study, we provide a deeper insight into these events and their potential origins. Steps preventing or reducing the risk for these phenomena are proposed and will help scientists working on genetic engineering of *P. pastoris* or similar non-conventional yeast to better understand and control clonal variability.

The non-conventional yeast *Pichia pastoris* is a popular host for recombinant protein production, due to a highly efficient secretion mechanism and the possibility of reaching high product titres for simpler enzymes such as phytase and complex proteins containing multiple post-translational modifications, e.g. monoclonal antibodies[1–5]. Research by Kurtzman *et al.*[6,7] resulted in the reclassification of the *P. pastoris* genus as *Komagatella*, including the subspecies *K. phaffii* and *K. pastoris*. However, they are still commonly referred to as *P. pastoris*. In recent years, the genetic toolbox for *P. pastoris* has been markedly expanded with several newly discovered native promoters, synthetic promoters and other regulatory elements[8–11]. The construction of optimized strains and vectors enabled new applications, e.g. the production of metabolites or expression of proteins lacking yeast specific hypermannosylation patterns[12,13]. Additionally, in a very recent publication Weninger *et al.*[14] reported the first CRISPR/Cas9 system for *P. pastoris* opening up new possibilities for genetic engineering approaches.

Nevertheless, the most frequently used approach for introducing the target gene in *P. pastoris* is still the integration of an expression cassette into the genome *via* homologous recombination. The most popular target for integration is the *AOX1* (alcohol oxidase 1) locus that represents the stronger expressed of the two alcohol oxidases in *P. pastoris*. This approach usually involves the utilization of the *AOX1* promoter (p*AOX1*) as homologous sequence and as promoter of the target gene, because it offers very high expression levels and tight regulation[15]. After a successful integration, the gene expression can be induced with methanol. However, a clone with an intact *AOX1* can metabolize methanol at a higher rate, designated as "methanol utilization plus" (Mut$^+$), complicating the maintenance of a constant induction[16,17]. Therefore, the application of expression cassettes with two homologous sequences targeted for mediating the replacement of *AOX1* is one possible technique. A knock-out mutation

[1]Fermentation Engineering, Bielefeld University, Universitätsstr. 25, Bielefeld, 33615, Germany. [2]Microbial Genomics and Biotechnology, Center for Biotechnology (CeBiTec), Bielefeld University, Universitätsstr. 27, Bielefeld, 33615, Germany. [3]Genome Research of Industrial Microorganisms, CeBiTec, Bielefeld University, Universitätsstr. 27, Bielefeld, 33615, Germany. Correspondence and requests for materials should be addressed to K.F. (email: karl.friehs@uni-bielefeld.de)

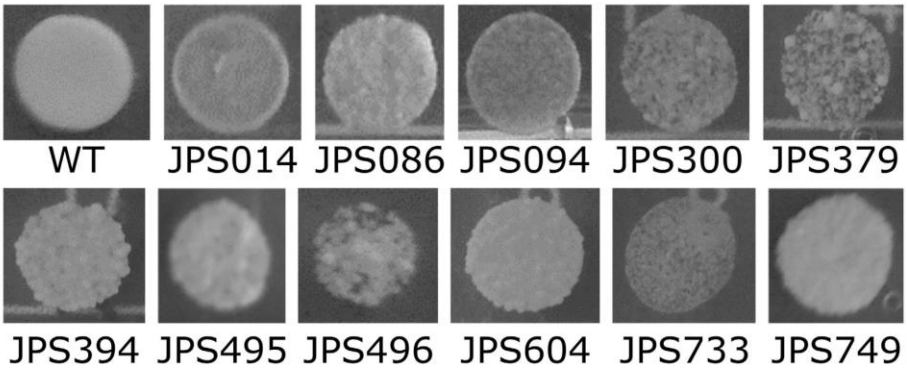

Figure 1. Colony morphology of wildtype *P. pastoris* CBS 7435 (WT) and irregular clones with crenulated morphology, grown on MD plates for 3 days at 28 °C. All mutant strains shown here were selected for genome sequencing.

of *AOX1* leads to the phenotype "methanol utilization slow" (Mut^S) easing process control and allowing to select for correct integration based on the phenotype.

Despite using comparatively long homologous sequences (ca. 1000 bp), a high variance in targeting efficiency is observed, indicating the prevalence of the non-homologous end joining (NHEJ) pathway in *P. pastoris*[18–20]. As a result, a high clonal variability is found after transformation, necessitating a time and labour-intensive screening process for the clone with the desired characteristics[21,22]. Concerning productivity characteristics, the different expression levels of clones typically originate from varying gene copy numbers[23,24]. The clonal variability is regarded as an inherent property of *P. pastoris* that is a by-product of the available and established transformation techniques in combination with the strong NHEJ pathway in this yeast species. Many other yeasts, filamentous fungi and higher eukaryotes that are used in biotechnological applications display similar or more pronounced clonal variabilities due to a predominant NHEJ pathway[25–27], while in the model yeast *Saccharomyces cerevisiae* homologous recombination is dominant over the NHEJ pathway[28]. Different techniques to reduce the clonal variability in *P. pastoris* have been proposed[14,18,20,29]. However, disadvantages like high complexity, lower strain fitness or not yet fully realized methods for donor cassette integration have kept these techniques from replacing the established ones for the efficient construction of producer strains. Therefore, successive genetic manipulation steps, e.g. for the construction of biosynthetic pathways, have been comparatively challenging in *P. pastoris* and only a few applications have been reported so far[30–34]. To date, the humanisation of the N-glycosylation pathway in *P. pastoris* via multiple consecutive cloning steps has been the most sophisticated and successful genetic engineering endeavour[13,35]. Similar observations have been made for other non-conventional yeast like *Hansenula polymorpha*, *Kluyveromyces lactis* and *Yarrowia lipolytica*[25,36,37].

In our previous study[24], we concentrated on analysing the clonal variability found in 845 *P. pastoris* strains transformed with a GFP expression cassette targeted for *AOX1* replacement. By screening all clones for their Mut-phenotype, GFP productivity and gene copy number 31 clones with interesting features were selected for genome sequencing. The combination of experimental and genome data allowed the discovery of novel insights into the correlation between productivity and integration event. In the majority of cases variations in the productivity could be traced back to gene copy number related effects. No off-target integrations were found in clones with a high productivity, suggesting that the impact of such events on productivity was low.

During the Mut-phenotype assay, strains were discovered that displayed an abnormal colony morphology on plate. Some of these clones were also selected for genome sequencing. Here, we discuss the results for these strains, elucidating a variety of as yet unreported off-target integration events. Besides the disruption of random genes, the co-integration of DNA elements from the plasmid host *E. coli* KRX as well as the relocation of the *AOX1* locus to another chromosome were discovered. The detrimental effect of these events on the genetic integrity and strain morphology presents an additional burden for the screening process. These observations are especially relevant for metabolic engineering or knock-out experiments in *P. pastoris*, while the integration events detailed in our previous study[24] are more pertinent to experiments targeting the creation of high producer strains.

In recent years, metabolic engineering of non-conventional yeast in general and *P. pastoris* in particular has gained more interest[11,38]. Therefore, the discussed integration events and the proposed theories explaining their origin will enable scientists working in this field to modify existing transformation and mutagenesis protocols in order to avoid these events. Based on this strategy, clonal diversity can be reduced and genetic engineering processes of higher complexity can be realized.

Results

Identification of clones with abnormal colony morphology. During the plating assay for determination of the Mut phenotype multiple clones with a crenulated colony morphology were found (Fig. 1). The change in colony morphology indicated growth deficiencies, similar to the ones reported for *OCH1* knock-out strains in *P. pastoris* and other non-conventional yeast[20,39,40]. *OCH1* codes for a mannosyltransferase and its loss leads to a

Strain	Mut-Phenotype	GCN	ExpLvl
JPS014	Mut^+	0.7	0.3
JPS086	Mut^+	0.4	1.1
JPS094	Mut^+	2.5	2.6
JPS165	Mut^+	0.9	1.0
JPS247	Mut^+	0.9	0.6
JPS277	Mut^S	0.9	0.9
JPS280	Mut^+	2.4	3.5
JPS300	Mut^+	0.5	0.5
JPS315	Mut^S	0.9	0.8
JPS336	Mut^+	1.0	0.1
JPS347	Mut^+	0.3	1.1
JPS379	Mut^+	0.1	0.0
JPS386	Mut^+	5.2	1.2
JPS394	Mut^+	0.6	0.8
JPS495	Mut^+	0.1	1.7
JPS496	Mut^+	0.5	0.5
JPS581	Mut^+	0.8	1.4
JPS603	Mut^+	3.1	0.6
JPS604	Mut^+	0.2	0.0
JPS626	Mut^+	0.6	0.7
JPS636	Mut^+	0.8	1.0
JPS733	Mut^+	0.6	0.3
JPS749	Mut^+	0.7	0.4
JPS768	Mut^+	0.6	0.5
JPS804	Mut^+	0.7	0.6
JPS835	Mut^+	0.8	0.8

Table 1. Characteristics of P. pastoris strains displaying an abnormal colony morphology, determined as detailed in Schwarzhans *et al.*[24]. A gene copy number (GCN) <1 can result from e.g. a reduced PCR efficiency during qPCR, but is still indicative of a single copy of the GFP cassette. The GFP expression level (ExpLvl) was normalized against a reference Mut^S strain with one copy of the cassette. Underlined strains were selected for genome sequencing.

lower cell wall integrity, negatively affecting growth and temperature sensitivity as well as an increased flocculation[13]. It is also a popular knock-out target for humanizing the glycosylation pattern in *P. pastoris*[40]. Therefore, it was of interest to discover a potential genetic cause affecting the growth behaviour of clones with abnormal colony morphology. In total, 26 (=3%) of 845 analysed *P. pastoris* strains displayed a crenulated colony morphology and were divided into the group "Abnormal morphology". Most of the clones of this group also belonged to other (productivity related) groups, mentioned in our previous report[24]. Table 1 summarizes the characteristics for all 26 clones with abnormal morphology. Except for the two strains JPS277 and JPS315, all clones had the Mut^+-phenotype. Furthermore, only JPS094 and JPS280 displayed clearly elevated GFP expression levels of 2.6 and 3.5, respectively. Since most clones with an abnormal colony morphology had a gene copy number and GFP expression level of approximately 1, they would likely not be selected if looking for high producer strains. However, from a strain engineering perspective the integration events that lead to their change in morphology could be relevant and the strains were therefore further investigated. Eleven strains belonging to this group were sequenced and a variety of different integration events were discovered.

Gene disruptions. In some cases, a gene disruption is the likely cause of the morphology change. For example, JPS496 (EMBL FBTT01000000) contained a disruption in the *MXR1* gene on chromosome 4, which codes for a methionine sulfoxide reductase (not to be confused with the similarly abbreviated methanol expression regulator 1 (*MXR1*)[41]). The gene product of *MXR1* in *S. cerevisiae* is part of the oxidative stress response and involved in the repair of damaged proteins by reducing methionine sulfoxide to methionine. Upon inactivation of *MXR1*, a change in colony morphology was noted in *S. cerevisiae*[42]. A disrupted *MXR1* in clone JPS496 could lead to the accumulation of damaged proteins in the cell and disrupt metabolic pathways, potentially explaining the observed morphology change. Interpretation of sequencing results for this clone were complicated by the fact that downstream of *MXR1* the paralogous gene *yahK* (alcohol dehydrogenase GroES domain protein) is located and also present on chromosome 3. Furthermore, *MXR1* itself has a paralogous copy on chromosome 3, although this copy is a fragment and only 567 bp long (*MXR1* on chromosome 4 is 669 bp long) and contains multiple mutations (>90 mismatches). In consequence, multiple possibilities existed to assemble the contigs related to the expression cassette and the *MXR1* locus in JPS496. To obtain a clearer picture, a PCR targeting the *MXR1* locus on chromosome 4 was performed, the primers are described in Table 2. The PCR assay indicated the integration of an > 6 kb fragment in the *MXR1* locus on chromosome 4. Figure 2(A) shows the proposed integration event in JPS496.

Primer name	Sequence	Purpose
MXR1-FW	GAGAAGGATAGGATCACGGTGGCC	*MXR1* locus amplification
MXR1-RV	GCGCCAGAGACATTCTACAGGGAG	
ADH-FW	CCCTTTATCACCAACTTGAACTCTCACATTCCCCC	Amplification of Adhesin *AidA* precursor *ychA* found in JPS094
ADH-RV	GGCCTGGTGCAACAGGATGGATTAATATTTTTAATGG	
FIM-FW	CTCAAATATCTTTAATAGCACCAATAACCAGCCAGGG	Amplification of Adhesin fragment *fimH* and Fimbrial protein found in JPS086
FIM-RV	CATCTGCCTAAATCTACCGTCTTATCAATATCCGC	
ACT1-FW	GGTGTGGTGCCAGATCTTTT	qRT-PCR; Housekeeping gene *ACT1* in *P. pastoris*; Amplicon = 181 bp
ACT1-RV	AGTGTTCCCATCGGTCGTAG	
sfmF-FW	TACTGAACGCTGGCGATACC	qRT-PCR; *sfmF* in JPS086 and *E. coli* KRX, Amplicon = 163 bp
sfmF-RV	AATTCGATGGCGACGGTTTG	
ychA-FW	GCAGCGTGCCATTATTACCG	qRT-PCR; *ychA* in JPS094 and *E. coli* KRX, Amplicon = 189 bp
ychA-RV	TTGGGCAATCTGATGCACCT	
ybaA-FW	TCGGAACGGAACGGATAAGAC	qRT-PCR; *ybaA* in JPS300 and *E. coli* KRX, Amplicon = 160 bp
ybaA-RV	TGCAAGTGTGAAAGCTCAGCA	
finO-FW	AATGTCACCACGCCACCAAA	qRT-PCR; *finO* in JPS496 and *E. coli* KRX, Amplicon = 159 bp
finO-RV	CTTCAGGGTGTTCACGGCAT	

Table 2. Primers used for PCR assays and qRT-PCR.

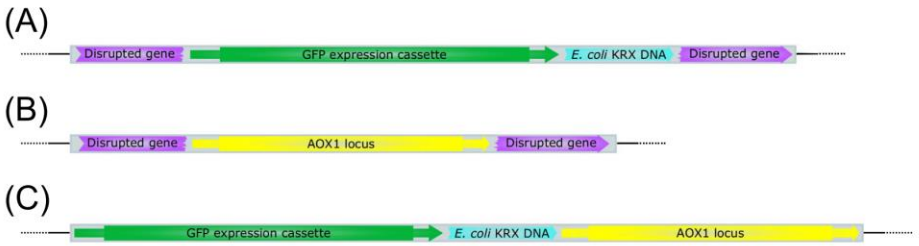

Figure 2. The different kinds of integration events discovered in sequenced *P. pastoris* clones with abnormal colony morphology. (**A**) Disruption of an untargeted gene by a GFP expression cassette fused to *E. coli* KRX DNA (**B**) Disruption of an untargeted gene by a re-integrated *AOX1* locus. (**C**) Co-integration of *E. coli* KRX DNA in fusion with an expression cassette at the *AOX1* locus.

Accordingly, the integration occurred between the base pairs 40 and 41 of the *MXR1* gene. In frame transcription of the *MXR1* gene would thus continue in the newly integrated p*AOX1* and end after adding 10 amino acids (AA) at the stop codon TGA. Additionally, due to the integration between base pairs 40 and 41, a p. Lys14Arg mutation occurs. The resulting 23 AA gene product of the *MXR1*-p*AOX1*-hybrid is of different composition and misses the catalytic domain compared to the full length *MXR1* gene product (222 AA) (Fig. S1). It is therefore probably inactive. No significant homologies between the integrated DNA and *MXR1* exist, indicating that the insertion was likely the result of the NHEJ pathway. The effect of a *MXR1* deletion on the morphology of *P. pastoris* has not been reported yet.

A different kind of gene disruption was found in JPS014 (EMBL FBTG01000000). In this strain the *AOX1* locus relocated from its native site on chromosome 4 to chromosome 2 between base 2,062,817 and 2,062,818. As a result, gene PP7435_Chr2-1130 was disrupted (Fig. 2(B)). PP7435_Chr2-1130 is annotated as a hypothetical protein in *P. pastoris* CBS 7435[43], but shows 100% identity to PAS_chr2-1_0178 in *P. pastoris* GS115[44], which codes for a polyphosphatidylinositol (PtdIns) phosphatase. In JPS014 the in frame transcription of PP7435_Chr2-1130 would continue after the 781st AA in the newly inserted p*AOX1*, adding 7 AA and end at a TGA stop codon. Furthermore, the insertion results in a c.G2346A silent mutation (p.Lys782Lys). In comparison to the full-length PP7435_Chr2-1130 protein (1069 AA), the gene product in JPS014 (788AA) is distinctly shorter and 79 AA of the predicted catalytic centre of the PtdIns phosphatase are mutated or missing (Fig. S2.). It is therefore likely, that the mutant protein in JPS014 is impaired in its activity. Due to the absence of sequence similarities between the disrupted gene and *AOX1*, the integration was likely caused by the NHEJ pathway. No knock-out study of this gene in *P. pastoris* was available at the time of writing the manuscript. However, it was shown that PtdIns phosphatases in *S. cerevisiae* are membrane bound proteins involved in cytoskeletal organization, signal transduction and membrane trafficking[45]. Stolz *et al.*[46] could also observe morphological changes, namely increased cell wall thickness, of *S. cerevisiae* clones by performing knock-out experiments with different PtdIns phosphatases. Therefore, it is likely that the observed growth deficiency of strain JPS014 was caused by the disrupted PP7435_Chr2-1130 gene. PP7435_Chr2-1130 was disrupted by the *AOX1* region targeted for replacement by the GFP cassette, implying that *P. pastoris* used the excised *AOX1* locus (including the *AOX1* promoter and

terminator) as a substrate for DNA repair. JPS014 displayed the Mut^{+}-phenotype indicating the full functionality of the re-integrated *AOX1* locus.

Based on the assembled genome data, it was deduced, which chain of events could have led to the abnormal colony morphology observed for JPS014 (Fig. 3). Firstly, the native *AOX1* locus on chromosome 4 was replaced with the GFP expression cassette *via* the homologous double crossing over event, as intended. Secondly, the excised linear *AOX1* locus was not degraded, but instead managed to reach chromosome 2. Potentially, a double-strand break (DSB) in PP7435_Chr2-1130 induced the DSB repair pathway. The NHEJ pathway accepted the linear *AOX1* locus as substrate for the DSB repair, disrupting PP7435_Chr2-1130 in the process. In consequence, a *P. pastoris* clone was created with a single copy of the expression cassette at the desired locus, a functioning *AOX1* locus on chromosome 2 and a disrupted PP7435_Chr2-1130, likely causing the observed morphological changes.

It seems that even the successful replacement of the *AOX1* locus with an expression cassette can result in unforeseen and detrimental off-target effects. It is difficult to estimate, how frequently the excised *AOX1* locus was re-integrated, because in case of JPS014 it was only discovered due to morphological changes. Multiple other re-insertions of the *AOX1* locus with no or other gene disruptions could have occurred, but bypassed the selection process. If screening for deletion mutants in *P. pastoris*, it seems possible to miss successful recombination events due to the re-integration of the knock-out target at a different locus. Depending on the aim of the experiment, clones with interesting properties might be overlooked this way. It seems advisable to check for reintegration of the knock-out target at a different locus in knock-out experiments, in addition to confirming the removal of the target from its original site. Additionally, transformation strategies with only one homology sequence for integration should prevent the described event and ought to reduce the clonal variation caused by relocating knock-out targets.

Co-integration of *E. coli* DNA. In four clones with irregular colony morphology (JPS086, JPS094, JPS300 and JPS496; EMBL FBTK01000000, FBTE01000000, FBTO01000000 and FBTT01000000, respectively) genome sequencing revealed the integration of *E. coli* KRX DNA elements directly adjacent to the GFP cassette (Fig. 2(C)). *E. coli* KRX was used as the plasmid host in all experiments. For a better understanding of the exact origin of these elements, we sequenced the *E. coli* KRX genome (unpublished). Using the genome data, the KRX DNA origin could be traced back to the F plasmid and the chromosome of its original host. In one *P. pastoris* clone a fusion of chromosomal and F plasmid DNA was discovered. Elements of *E. coli* KRX DNA present in strain JPS086, JPS094 and JPS300 contained multiple genes coding for proteins that are membrane associated in *E. coli*, identified *via* BLASTn. If not mentioned otherwise, KRX DNA elements in *P. pastoris* showed 100% identity to the chromosomal or F plasmid sequence of *E. coli* KRX. The ca. 4.6 kb of chromosomal *E. coli* KRX DNA in JPS086 contained genes coding for fimbria (*sfmF*) or that are involved in their expression and assembly (*sfmZ* and *smfH*). Adhesin associated genes (e.g. *ychA*) were encoded on the approx. 9.3 kb of F plasmid DNA integrated in JPS094. KRX F plasmid DNA (ca. 4.4 kb) found in clone JPS300 contained multiple membrane proteins or hypothetical proteins with assumed membrane association (*ybdA*, *yuaD*, *ybbA* and *ybaA*). In JPS496, a fusion of 553 bp chromosomal DNA and 947 bp F plasmid DNA from KRX was found. On the chromosomal portion, two sequences associated with transposases and on the F plasmid portion *finO* (fertility Inhibition) as well as a fragment of *traX* (acetylation of F-pilus) were identified. No genes coding for KRX proteins with (hypothetical) membrane association were detected in JPS496.

Interestingly, the putative F plasmid DNA in clone JPS094 contained two 12 bp long deletions in comparison to the reference (Fig. 4). Both missing 12 bp sequences are palindromic, which indicates a common cause for their deletion. Potentially, the DNA repair mechanism that facilitated the integration of the pAHBgl-GFP and F plasmid hybrid in *P. pastoris* caused the deletions. It cannot be excluded that the excision already occurred in *E. coli* KRX. However, the presence of both 12 bp sequences in the KRX genome data contraindicates this theory. DNA repair mechanisms in yeast have problems with (long) palindromic sequences. In some cases, this phenomenon can lead to DSBs or inhibition of mismatch repair[47,48]. This fact suggests that the deletion occurred post-transformational in *P. pastoris* during or after integration of the foreign DNA. The deleted palindromic sequences lead to the excision of two stop-codons (of *yuaO* and KRX_F_plasmid_20) and the creation of a ca. 157 kDa large fusion protein. Even more surprisingly, the putative ORF ends 4 bp outside the integrated F plasmid DNA. In the adjacent *E. coli* backbone of pAHBgl-GFP, a TGA base-triplet serves as a presumed stop-codon. The DNA sequence of the resulting fusion protein shows similarity with *ycbB* (98% identity, BLASTn) of the *E. coli* ECC-1470 plasmid pECC-1470_100 (GenBank: CP010345) and the encoded protein has a high similarity (99% identity, BLASTp) to an outer membrane auto transporter barrel domain protein from *E. coli* OK1357 (GenBank: EFZ69416). In comparison to the auto transporter from *E. coli* OK1357, the putative fusion protein in JPS094 is 131 AA shorter. Using the NCBI Conserved Domain Database, two conserved domains associated with membrane anchoring and transport function (GenBank: PRK14849 and TIGR01414; BLASTx e-values 5.85×10^{-33} and 6.49×10^{-23}, respectively) were found in the fusion protein[49].

With the exception of strain JPS496, a high abundance of genes coding for proteins that interact with the cell membrane were found on the integrated *E. coli* KRX DNA. This suggests that the observed morphological change have been the result of these proteins compromising the cellular integrity of *P. pastoris*. Since the irregular morphology was observed on both MD and MM plates, the expression of the *E. coli* proteins seemed to be independent of the methanol induction. In order to confirm the transcription of the *E. coli* fragments in the *P. pastoris* mutant strains, four representative targets were chosen for a qRT-PCR experiment. The targets were *sfmF* in JPS086, *ychA* in JPS094, *ybaA* in JPS300 and *finO* in JPS496. All four transcripts could be detected during the exponential growth phase in MD medium in their respective strains (Fig. 5). *sfmF* exhibited the highest and *ychA* the lowest relative transcript level compared to the endogenous control *ACT1*, with 0.38 and 0.16, respectively. At the same time, no expression of all four targets could be detected in *P. pastoris* CBS 7435 wild type.

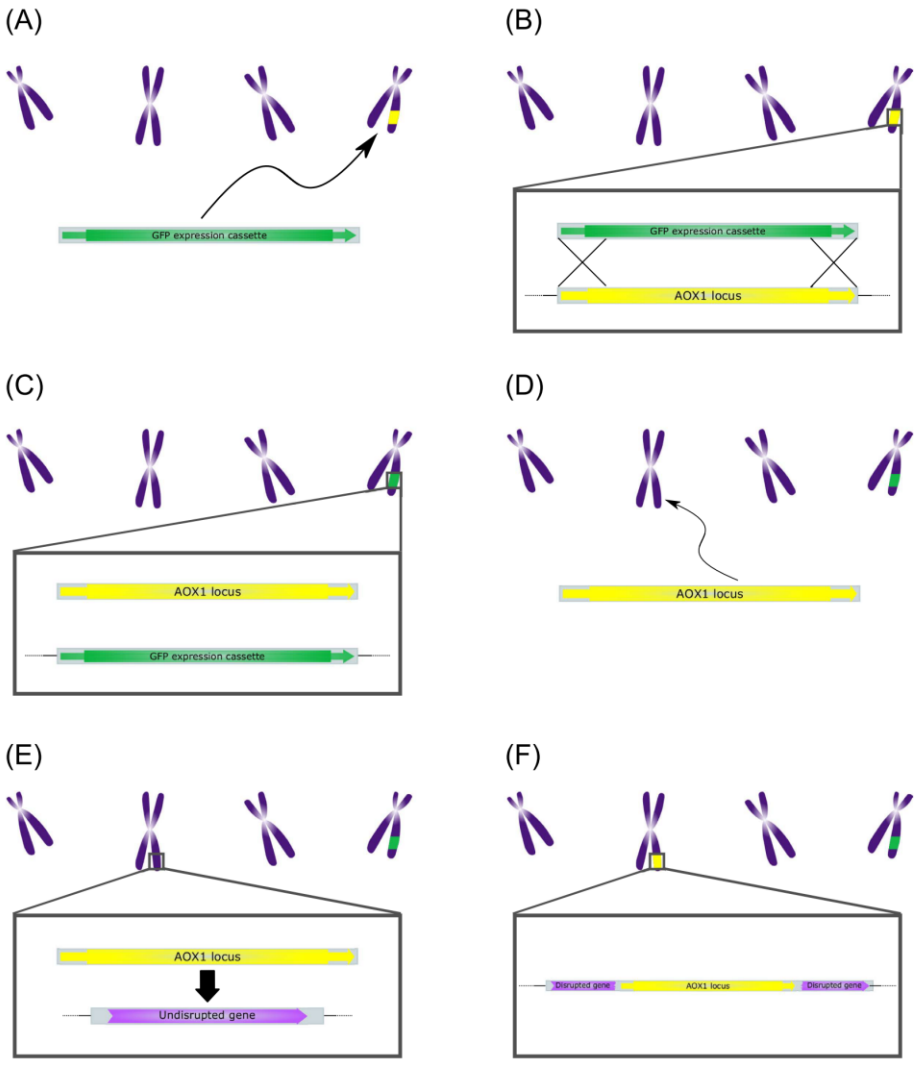

Figure 3. Proposed mechanism facilitating the relocation of the *AOX1* locus as found in clone JPS014. (**A**) After transformation the linear GFP expression cassette from pAHBgl-GFP moved to chromosome 4, where the *AOX1* locus is situated. (**B**) Two distal homologous sequences of the cassette with the chromosomal *AOX1* locus mediate a double cross-over recombination event (**C**) The cassette has replaced the *AOX1* locus on chromosome 4 (**D**) The excised linear *AOX1* locus moved to chromosome 2 (**E**) Potentially due to a DSB initiating the NHEJ pathway, the *AOX1* locus is integrated into a random gene on chromosome 2 (**F**) The untargeted gene PP7435_Chr2-1130 was disrupted. While the *AOX1* locus is still functional, the disruption of the untargeted gene likely led to the observed abnormal colony morphology in the case of JPS014.

Interestingly, for most targets the Cq-value suggests that the transcript level was similar or higher in the mutant strains than in *E. coli* KRX (Fig. S3). It remains unclear how the transcription of the *E. coli* fragments is regulated in *P.pastoris*, but the results strongly suggest their activity independent of the methanol induction. Although the aberrant phenotypes indicate that the transcripts are also translated, further analysis *via* e.g. LC-MS would be needed to confirm the presence of the translated *E. coli* proteins in the mutant strains.

The morphological change of JPS496 was likely due to the disruption of *MXR1*, mentioned above. Other clones with irregular morphology were tested *via* PCR for the presence of the adhesin and fimbria genes seen in JPS094 and JPS086 *ychA* and *fimH*, respectively. It was found that three of them contained adhesin genes and six were positive for fimbria synthesis genes. Considering that these particular clones were only found to contain *E. coli* KRX DNA due to their abnormal colony morphology, the yet undiscovered presence of different elements from KRX in other strains appears possible. Taken together with the activity of these elements without methanol

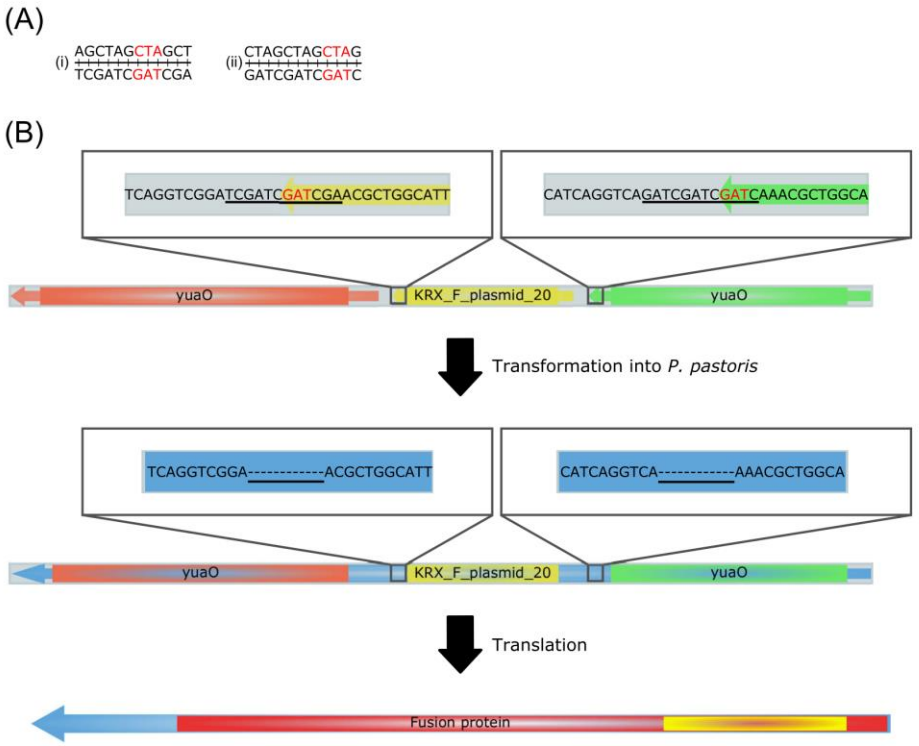

Figure 4. Observed modification of *E. coli* KRX DNA in *P. pastoris* clone JPS094. (**A**) Sequence of the two 12 bp long palindromic repeats found at the ends of (i) KRX_F_plasmid_20 and (ii) *yuaO*, respectively. The stop codons have been highlighted in red. (**B**) In its native form on the F plasmid of *E. coli* KRX both palindromic repeats (underlined) are present and contain a stop codon. After transformation into *P. pastoris* a deletion of both palindromic repeats was found in the genome assembly of JPS094. The loss of the stop codons indicated the formation of a ca. 1500 AA long fusion protein (blue), containing two conserved domains (red and yellow) after translation. The larger domain is associated with lipoproteins and autotransporters and the smaller one with outer membrane autotransporters.

induction this could present multiple challenges for the biotechnological application of *P. pastoris*. Integration *via* the NHEJ pathway could lead to the disruption of untargeted genes. Gene products that interfere with the metabolism of *P. pastoris* could lead to decreased product yields or negatively affect growth behaviour. Further analysis on the activity of *E. coli* operons in *P. pastoris* is needed to clarify this theory. It has to be noted that no separate integration of KRX DNA was observed, it only occurred in fusion with an expression cassette. Therefore it seems, that *P. pastoris* efficiently degraded most foreign DNA and only the combination of KRX DNA and expression cassette facilitated an integration of *E. coli* DNA into the genome of *P. pastoris*.

Integrated *E. coli* KRX DNA was always directly adjacent to at least one *Bgl*II site on the chromosome or F plasmid of its original host. *Bgl*II was used for linearization of the expression cassette from pAHBgl-GFP, prior to transformation. Based on these results a theory regarding the mechanism underlying the observed phenomenon was formulated (Fig. 6). Presumably F plasmids and fragmented chromosomal DNA were co-extracted with the pAHBgl-GFP vector from *E. coli* KRX during plasmid preparation. *Via Bgl*II digestion the DNA was linearized, compatible sticky-ends were introduced and enabled the *in vivo* ligation of various fragment combinations in *P. pastoris* after transformation. *In vivo* ligation of linear DNA fragments in yeast has been reported before[50,51]. For *P. pastoris* it has been suspected that this mechanism plays a role in the generation of multi-copy clones[21]. The different combinations were then integrated into the genome of *P. pastoris* leading to the observed morphological changes. Combinations of ligated foreign DNA without homology sequence were likely degraded rather than integrated. While integration of such elements via the NHEJ pathway is possible, no integration events of that kind were found in any of the sequenced clones. Since the unwanted integration and activity of *E. coli* host DNA is not considered in commonly used *P. pastoris* transformation protocols, prevention mechanisms should be employed. *E. coli* strains containing F plasmids (like JM109, TOP10F', XL10-Gold or DH5alphaF') are often used for plasmid propagation in studies with *P. pastoris*[52–55], but should be avoided. In order to prevent insertion of chromosomal DNA, either gel purification or PCR amplification of the expression cassette seems advisable. It has to be noted that in some cases the integrated *E. coli* host DNA fragments were of similar size as the expression

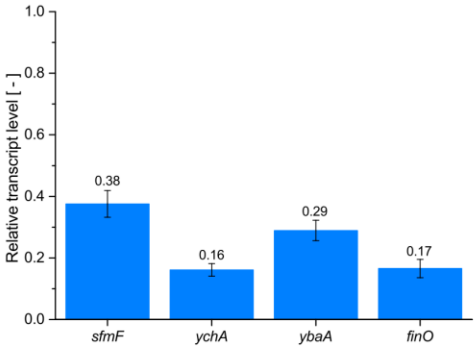

Figure 5. Relative transcriptional levels of *E. coli* KRX genes found in *P. pastoris* strains. *Sfmf* (fimbrial protein) in JPS086, *ychA* (adhesin *AidA* precursor) in JPS094, *ybaA* (signalling protein) in JPS300 and *finO* (fertility inhibitor) in JPS496. Samples for qRT-PCR experiments were taken in the exponential growth phase from cells grown in MD-medium. *ACT1* was used as housekeeping gene and values were normalized relative to the expression level of the *ACT1* gene in each sample. Error bars represent the standard deviation with n = 3.

cassette, reducing the potential effectiveness of gel purification. Furthermore, the use of blunt-end or rare cutters for expression cassette linearization ought to reduce the likelihood of hybrid formation.

Discussion

Several non-canonical integration events were discovered in *P. pastoris* clones displaying a change in colony morphology. The disruption of genes analogous to *MXR1* and a PtdIns phosphatse from *S. cerevisiae* was shown to be the likely cause for growth deficiencies in two of the sequenced strains. In the case of the PtdIns phosphatase disruption, a relocation of the *AOX1* locus from chromosome 4 to chromosome 2 was observed. This event raises the question of how efficiently *P. pastoris* can re-integrate knock-out targets. The full functionality of the relocated *AOX1* locus urges scientists interested in efficient knock-out studies in *P. pastoris* (and other non-conventional yeast) to take similar events into consideration for the design of their experiments. Successful knock-outs are complicated, if the locus simply moves to a new site and clonal diversity is increased in case untargeted genes are disrupted by the relocated knock-out target. Novel techniques like the recently presented CRISPR/Cas9 system could help in knock-out studies[14].

In multiple other strains morphological changes were possibly caused by the insertion and activity of DNA elements coding for membrane associated proteins from the plasmid propagation strain *E. coli* KRX. The methanol induction independent transcription of these *E. coli* elements in *P. pastoris* could be confirmed in a qRT-PCR assay. These findings strongly suggest that more care is needed, if transforming *P. pastoris*. Potentially the integration of other *E. coli* sequences remained undetected. Interestingly, in one case a presumed post-transformational modification of the KRX DNA was observed resulting in the excision of two stop codons and the creation of a triple-fusion protein with two membrane domains. The unwanted insertion of *E. coli* elements could cause multiple problems, e.g. in the generation of producer strains. Therefore, *E. coli* F^- strains should be used for studies involving *P. pastoris* transformation. Amplifying the expression cassette *via* PCR instead of using plasmid isolations should remediate the co-integration problem more effectively than gel purification. While gel purification of the expression cassette is a commonly used technique, we observed the co-integration of *E. coli* DNA elements of various sizes ranging from 1.5 to 9.3 kb. In some cases, these elements have a similar length in comparison to the expression cassette and thereby could by-pass a gel purification step. Nevertheless, PCR amplification could prove unsuitable in certain cases. Using blunt-end or rare cutters for plasmid linearization should also impede inadvertent co-integrations in these situations. Newly developed plasmids were recently published that offer sites for blunt-end linearization of an expression cassette targeting the *AOX1* locus for replacement[56]. While evolutionary horizontal gene transfer from bacteria to yeast (e.g. the acquisition of *URA1*[57]) was described previously, no co-integration of *E. coli* host DNA during genetic manipulation of *P. pastoris* has been reported so far.

Some of the sequenced strains (JPS379, JPS394, JPS495, JPS604 and JPS733) contained integration events that could not be related to the phenotype. They include the integration of a truncated expression cassette and mutations in *AOX1*. However, no gene disruptions were observed. While these events in some cases explained the low productivity, no clear correlation to the colony morphology could be deduced. In all analysed clones, no SNPs (single-nucleotide polymorphism) were detected that would explain the change in colony morphology.

In summary, the results elucidate the reason behind some of the causes for the clonal variability encountered in transformation experiments with *P. pastoris*. By focusing on strains with abnormal colony morphology, a group of clones undesirable for biotechnological applications can be better understood. Their change in morphology or growth behaviour is detrimental for production processes and in similar cases like Δ*OCH1* strains, special process strategies had to be developed to accommodate the changed physiology[58]. Additionally, the documented integration events and the proposed methods of preventing them enable scientists working with *P. pastoris* to devise optimized strategies for genetic engineering and mutagenesis of strains.

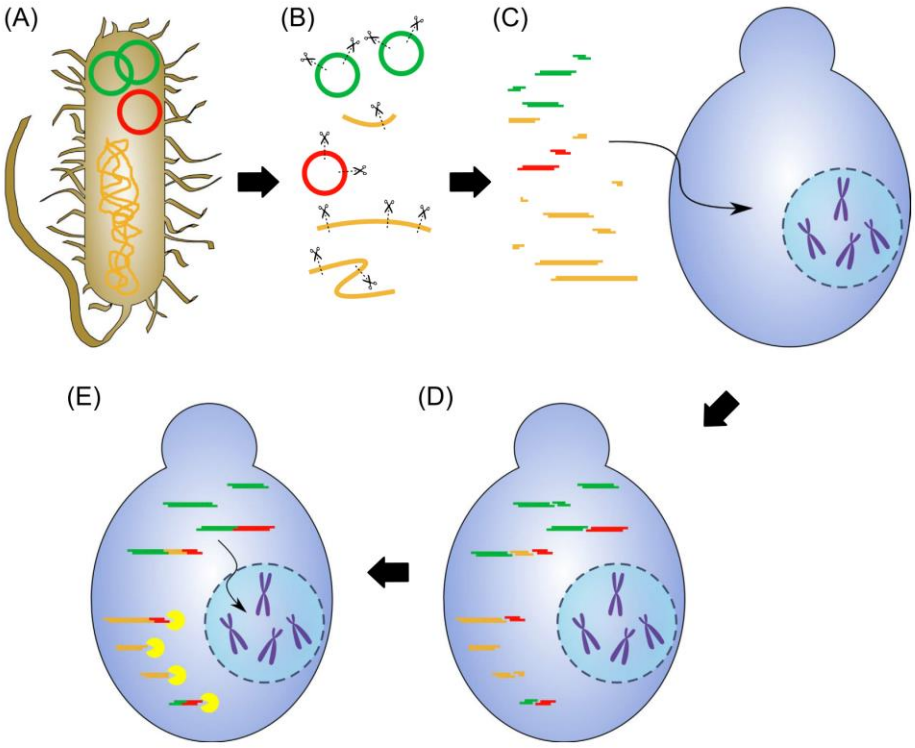

Figure 6. Schematic depiction of the proposed mechanism facilitating integration of *E. coli* host DNA in fusion with the expression cassette into the genome of *P. pastoris*. (**A**) *E. coli* cells containing chromosomal (orange) and F plasmid (red) DNA are used to produce the expression vector (green) (**B**) During plasmid isolation F plasmids and fragmented chromosomal DNA are co-extracted with the expression vector. Subsequent enzymatic digestion creates compatible sticky ends (**C**) Restricted DNA is transferred into *P. pastoris* cells via e.g. electroporation (**D**) *in vivo* various fragment combinations are ligated and form new hybrids (**E**) Expression cassettes and hybrids containing expression cassettes with homologies to chromosomal loci of *P. pastoris* are integrated into the genome, while non-homologous DNA is degraded.

Methods

Strains and vector. Plasmids were constructed and propagated in *E. coli* KRX (Promega, Madison, WI; Genotype: [F′, traD36, ΔompP, proA + B+, lacIq, Δ(lacZ)M15] ΔompT, endA1, recA1, gyrA96 (Nalr), thi-1, hsdR17 (rk–, mk+), e14– (McrA–), relA1, supE44, Δ(lac-proAB), Δ(rhaBAD)::T7 RNA polymerase). Yeast experiments involved *P. pastoris* CBS 7435 (Δ*HIS4*) (Austrian Center of Industrial Biotechnology, Graz, Austria) and the wild type CBS 7435 (identical to NRRL Y-11430 and ATCC76273), which was obtained from the Spanish Type Culture Collection (Valencia, Spain) under the strain number CECT 11047. *P. pastoris* CBS 7435 (Δ*HIS4*) was transformed with *Bgl*II linearized expression cassette from pAHBgl-GFP according to Wu and Letchworth[59]. The composition and construction of pAHBgl-GFP was described previously[24]. In brief, a linearized GFP expression cassette (purified with Wizard® Plus SV Minipreps DNA Purification System, Promega, Madison, WI) was transformed into *P. pastoris* and targeted the *AOX1* locus on chromosome 4 for replacement. Before transformation the linearized DNA was purified with the Wizard® SV Gel and PCR Clean-Up System (Promega, Madison, WI).

Cultivation conditions. *E. coli* KRX was cultivated in shake flasks at 37 °C with 120 rpm using Lysogeny Broth (LB) medium supplemented with 100 μg/mL Ampicillin.

P. pastoris was cultivated either in shake flasks, in 96-deep-well plates (Eppendorf, Germany) or on agar plates. Shake flask cultivations were performed with (Buffered) Minimal Dextrose ((B)MD, Invitrogen[60]), or Yeast Peptone Dextrose (YPD) medium and supplemented with 4 mg/L L-histidine when necessary. Deep-well plate experiments were performed according to Weis *et al.*[61] and Hartner *et al.*[9], with BMD and Buffered Minimal Methanol medium with 1 or 5 g/L methanol (BMM2 and BMM10, respectively). Agar plates contained Minimal Dextrose (MD) or Minimal Methanol (MM) medium[60]. All cultivations were carried out at 28 °C, with shake flasks being agitated at 120 rpm and deep-well plates at 340 rpm.

Identification of Mut-phenotype and colony morphology. The Mut-phenotype was assayed based on the procedure described in the EasySelect™ Pichia Expression Kit[60]. Briefly, cells were grown in BMD medium for 60 hours, washed twice and then 5 μL aliquots of the resuspended cells were plated onto both MD and MM plates. After 2–3 days of incubation at 28 °C, the Mut-phenotype could be determined based on the growth behaviour. Additionally, the plating test allowed the visual identification of clones with irregular colony morphology.

Isolation of genomic DNA and PCR assays. For the extraction of genomic DNA (gDNA) from *P. pastoris* the MasterPure™ Yeast DNA Purification Kit from Epicentre (Madison, WI) was used according to the manufacturer's protocol.

gDNA from relevant clones was used for different PCR assays. The sequence of the primers used in these assays can be found in Table 2.

Genome sequencing and bioinformatic analysis of *P. pastoris* strains. The performed procedures for *P. pastoris* clones selected for genome sequencing were recently reported[24]. In brief, the gDNA quantity and quality was assayed *via* the Quant-iT PicoGreen dsDNA kit (Invitrogen, Waltham, MA) and gel-electrophoresis, respectively. gDNA of sufficient quality was sequenced on an Illumina MiSeq system using a paired-end library prepared with the TruSeq sample preparation kit (Illumina, San Diego, CA). The raw data was assembled *de novo* in the GS *De Novo* Assembler (Version 2.8, Roche, Basel, Switzerland) with default settings. All assembled genomes can be found under the study id **PRJEB12220** in the EBI database.

Bioinformatic analysis involved similarity analysis using the BLASTn algorithm[62] and a local database including the pAHBgl-GFP vector sequence. Hits with an e-value $>1 \times 10^{-20}$ and a sequence identity of 100% were analysed in more detail. Gaps in the vector sequence were closed in an *in silico* approach applying CONSED[43,63,64]. This approach also allowed the identification of insertion sites of the expression cassette into the genome of *P. pastoris*.

Genome sequencing, annotation and sequence comparison of the *E. coli* KRX genome. DNA was extracted using the Wizard® Genomic DNA Purification KIT (Promega, Madison, WI). *E. coli* KRX genome sequencing and bioinformatic analysis was performed analogous to the method described for *P. pastoris*, resulting in the draft sequence of the chromosome and the F plasmid (unpublished). *E. coli* KRX contigs were compared to the *P. pastoris* draft genomes sequences using BLASTn[62]. Hits with an e-value $>1 \times 10^{-20}$ and a sequence identity of 95% were analysed in detail to detect horizontal gene transfer between the *P. pastoris* strains and *E. coli* KRX chromosomal or F plasmid DNA.

qRT-PCR experiments. *E. coli* KRX, *P. pastoris* CBS 7435 and the mutant strains JPS086, JPS094, JPS300 and JPS496 were grown as described above in shake flasks containing LB-Medium or MD-Medium, respectively. Samples taken in the exponential growth phase were employed for RNA isolation using the RNeasy Mini Kit (Qiagen, Hilden, Germany), with both an on- and off-column gDNA digestion step. RNA quantity and the absence of gDNA and other contaminants was assayed with the DropSense16 system (Trinean, Ghent, Belgium). Four different primer sets targeting the various *E. coli* fragments found in the sequenced *P. pastoris* mutant strains were designed using the NCBI Primer-BLAST tool, making sure no off-target activity on either the *E. coli* or *P. pastoris* genome was found. Additionally, the primer pair for quantification of the housekeeping gene *ACT1* (β-actin) in *P. pastoris* was taken from literature[65]. All primers exhibited similar T_M-values (59 ± 2 °C) and amplicon sizes (170 ± 19 bp). The sequences and amplicon sizes of all employed primer pairs can be found in Table 2. Per reaction 50 ng RNA free of gDNA was used as template. Employing the SensiFAST™ SYBR® No-ROX One-Step Kit by Bioline (London, UK), samples were measured on the LightCycler® 96 system (Roche, Basel, Switzerland) in biological triplicates with technical duplicates. For *P. pastoris* samples the transcript level of *E. coli* genes was normalized against the *ACT1* level of the same strain via the $2^{-\Delta\Delta Ct}$ method[66]. After 40 cycles the specificity of the amplicons was confirmed via a melting curve.

References

1. Rodriguez, E., Mullaney, E. J. & Lei, X. G. Expression of the *Aspergillus fumigatus* phytase gene in *Pichia pastoris* and characterization of the recombinant enzyme. *Biochem. Biophys. Res. Commun.* **268,** 373–378 (2000).
2. Kim, S. *et al.* Real-time monitoring of glycerol and methanol to enhance antibody production in industrial *Pichia pastoris* bioprocesses. *Biochem. Eng. J.* **94,** 115–124 (2015).
3. Gasser, B. *et al. Pichia pastoris* : protein production host and model organism for biomedical research. *Future Microbiol.* **8,** 191–208 (2013).
4. Ahmad, M., Hirz, M., Pichler, H. & Schwab, H. Protein expression in *Pichia pastoris* : recent achievements and perspectives for heterologous protein production. *Appl. Microbiol. Biotechnol.* **98,** 5301–5317 (2014).
5. Cregg, J. M. *et al.* Expression in the yeast *Pichia pastoris*. *Methods in enzymology* 463, (Elsevier Inc., 2009).
6. Kurtzman, C. P. 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 (2005).
7. Kurtzman, C. P. Biotechnological strains of *Komagataella (Pichia) pastoris* are *Komagataella phaffii* as determined from multigene sequence analysis. *J. Ind. Microbiol. Biotechnol.* **36,** 1435–1438 (2009).
8. Qin, X. *et al.* GAP promoter library for fine-tuning of gene expression in Pichia pastoris. *Appl. Environ. Microbiol.* **77,** 3600–8 (2011).
9. Hartner, F. S. *et al.* Promoter library designed for fine-tuned gene expression in *Pichia pastoris*. *Nucleic Acids Res.* **36,** e76 (2008).
10. Prielhofer, R. *et al.* Induction without methanol: novel regulated promoters enable high-level expression in *Pichia pastoris*. *Microb. Cell Fact.* **12,** 5 (2013).
11. Vogl, T. *et al.* A toolbox of diverse promoters related to methanol utilization – functionally verified parts for heterologous pathway expression in *Pichia pastoris*. *ACS Synth. Biol.* **5,** 172–186 (2016).
12. Geier, M., Fauland, P., Vogl, T. & Glieder, A. Compact multi-enzyme pathways in *P. pastoris*. *Chem. Commun.* **51,** 1643–1646 (2015).

13. Choi, B. *et al.* Use of Combinatorial Genetic Libraries to Humanize N-Linked Glycosylation in the Yeast *Pichia pastoris*. *Proc. Natl. Acad. Sci. USA*. **100,** 5022–5027 (2014).
14. Weninger, A., Hatzl, A.-M., Schmid, C., Vogl, T. & Glieder, A. Combinatorial optimization of CRISPR/Cas9 expression enables precision genome engineering in the methylotrophic yeast *Pichia pastoris*. *J. Biotechnol.*, doi: 10.1016/j.jbiotec.2016.03.027 (2016).
15. Lin Cereghino, G. P. *et al.* New selectable marker/auxotrophic host strain combinations for molecular genetic manipulation of *Pichia pastoris*. *Gene* **263,** 159–169 (2001).
16. Romanos, M. A., Scorer, C. A. & Clare, J. J. Foreign gene expression in yeast: a review. *Yeast* **8,** 423–488 (1992).
17. Krainer, F. W. *et al.* Recombinant protein expression in *Pichia pastoris* strains with an engineered methanol utilization pathway. *Microb. Cell Fact.* **11,** 22 (2012).
18. Näätsaari, L. *et al.* Deletion of the *Pichia pastoris* KU70 homologue facilitates platform strain generation for gene expression and synthetic biology. *PLoS One* **7,** e39720 (2012).
19. Nett, J. H., Hodel, N., Rausch, S. & Wildt, S. Cloning and disruption of the *Pichia pastoris* ARG1, ARG2, ARG3, HIS1, HIS2, HIS5, HIS6 genes and their use as auxotrophic markers. *Yeast* **22,** 295–304 (2005).
20. Chen, Z. *et al.* Enhancement of the gene targeting efficiency of non-conventional yeasts by increasing genetic redundancy. *PLoS One* **8,** e57952 (2013).
21. Clare, J., Rayment, F., Sreekrishna, K. & Romanos, M. A. High-level expression of tetanus toxin fragment C in *Pichia pastoris* strains containing multiple tandem integrations of the gene. *Nat. Biotechnol.* **9,** 455–460 (1991).
22. Looser, V. *et al.* Cultivation strategies to enhance productivity of *Pichia pastoris*: A review. *Biotechnol. Adv.* **33,** 1177–1193 (2015).
23. Aw, R. & Polizzi, K. M. Can too many copies spoil the broth? *Microb. Cell Fact.* **12,** 128 (2013).
24. Schwarzhans, J.-P. *et al.* Integration event induced changes in recombinant protein productivity in *Pichia pastoris* discovered by whole genome sequencing and derived vector optimization. *Microb. Cell Fact.* **15,** 84 (2016).
25. Verbeke, J., Beopoulos, A. & Nicaud, J. M. Efficient homologous recombination with short length flanking fragments in Ku70 deficient *Yarrowia lipolytica* strains. *Biotechnol. Lett.* **35,** 571–576 (2013).
26. Meyer, V. *et al.* Highly efficient gene targeting in the Aspergillus niger kusA mutant. *J. Biotechnol.* **128,** 770–775 (2007).
27. Guirouilh-Barbat, J. *et al.* Impact of the KU80 pathway on NHEJ-induced genome rearrangements in mammalian cells. *Mol. Cell* **14,** 611–623 (2004).
28. Daley, J. M., Palmbos, P. L., Wu, D. & Wilson, T. E. Nonhomologous end joining in yeast. *Annu. Rev. Genet.* **39,** 431–451 (2005).
29. Heiss, S., Maurer, M., Hahn, R., Mattanovich, D. & Gasser, B. Identification and deletion of the major secreted protein of *Pichia pastoris*. *Appl. Microbiol. Biotechnol.* **97,** 1241–9 (2013).
30. Marx, H., Mattanovich, D. & Sauer, M. Overexpression of the riboflavin biosynthetic pathway in *Pichia pastoris*. *Microb. Cell Fact.* **7,** 23 (2008).
31. Araya-Garay, J. M., Feijoo-Siota, L., Rosa-Dos-Santos, F., Veiga-Crespo, P. & Villa, T. G. Construction of new *Pichia pastoris* X-33 strains for production of lycopene and β-carotene. *Appl. Microbiol. Biotechnol.* **93,** 2483–2492 (2012).
32. Wriessnegger, T. *et al.* Production of the sesquiterpenoid (+)-nootkatone by metabolic engineering of *Pichia pastoris*. *Metab. Eng.* **24,** 18–29 (2014).
33. Liu, X. B. *et al.* Metabolic engineering of *Pichia pastoris* for the production of dammarenediol-II. *J. Biotechnol.* **216,** 47–55 (2015).
34. Nocon, J. *et al.* Model based engineering of *Pichia pastoris* central metabolism enhances recombinant protein production. *Metab. Eng.* **24,** 129–38 (2014).
35. Hamilton, S. R. *et al.* Humanization of Yeast to Produce Complex Terminally Sialylated Glycoproteins. *Science* (80−). **313,** 1441–1443 (2006).
36. Kegel, A., Martinez, P., Carter, S. D. & Åström, S. U. Genome wide distribution of illegitimate recombination events in *Kluyveromyces lactis*. *Nucleic Acids Res.* **34,** 1633–1645 (2006).
37. Saraya, R. *et al.* Novel genetic tools for *Hansenula polymorpha*. *FEMS Yeast Res.* **12,** 271–278 (2012).
38. Liu, L., Redden, H. & Alper, H. S. Frontiers of yeast metabolic engineering: Diversifying beyond ethanol and *Saccharomyces*. *Curr. Opin. Biotechnol.* **24,** 1023–1030 (2013).
39. Nakayama, K., Nagasu, T., Shimma, Y., Kuromitsu, J. & Jigami, Y. *OCH1* encodes a novel membrane bound mannosyltransferase: outer chain elongation of asparagine-linked oligosaccharides. *EMBO J.* **11,** 2511–2519 (1992).
40. Krainer, F. W. *et al.* Knockout of an endogenous mannosyltransferase increases the homogeneity of glycoproteins produced in *Pichia pastoris*. *Sci. Rep.* **3,** 3279 (2013).
41. Lin-Cereghino, G. P. *et al.* Mxr1p, a Key Regulator of the Methanol Utilization Pathway and Peroxisomal Genes in *Pichia pastoris*. *Mol. Cell. Biol.* **26,** 883–897 (2006).
42. Hansen, J. Inactivation of MXR1 abolishes formation of dimethyl sulfide from dimethyl sulfoxide in *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.* **65,** 3915–3919 (1999).
43. Küberl, A. *et al.* High-quality genome sequence of *Pichia pastoris* CBS7435. *J. Biotechnol.* **154,** 312–20 (2011).
44. De Schutter, K. *et al.* Genome sequence of the recombinant protein production host *Pichia pastoris*. *Nat. Biotechnol.* **27,** 561–566 (2009).
45. Nguyen, P. H. *et al.* Interaction of Pik1p and Sjl proteins in membrane trafficking. *FEMS Yeast Res.* **5,** 363–371 (2005).
46. Stolz, L. E., Huynh, C. V., Thorner, J. & York, J. D. Identification and characterization of an essential family of inositol polyphosphate 5-phosphatases (INP51, INP52 and INP53 gene products) in the yeast *Saccharomyces cerevisiae*. *Genetics* **148,** 1715–1729 (1998).
47. Nag, D. K., White, M. A. & Petes, T. D. Palindromic sequences in heteroduplex DNA inhibit mismatch repair in yeast. *Nature* **340,** 318–320 (1989).
48. Nasar, F., Jankowski, C. & Nag, D. K. Long palindromic sequences induce double-strand breaks during meiosis in yeast. *Mol. Cell. Biol.* **20,** 3449–3458 (2000).
49. Marchler-Bauer, A. *et al.* CDD: A Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39,** 225–229 (2011).
50. Kunes, S., Botstein, D. & Fox, M. S. Transformation of yeast with linearized plasmid DNA. Formation of inverted dimers and recombinant plasmid products. *J. Mol. Biol.* **184,** 375–387 (1985).
51. Suzuki, K., Imai, Y. & Yamashita, I. *In vivo* ligation of linear DNA molecules to circular forms in the yeast *Saccharomyces cerevisiae*. *J. Bacteriol.* **155,** 747–754 (1983).
52. Hong, Y., Zhou, H., Tu, X., Li, J. & Xiao, Y. Cloning of a laccase gene from a novel basidiomycete Trametes sp. 420 and its heterologous expression in *Pichia pastoris*. *Curr. Microbiol.* **54,** 260–5 (2007).
53. Liu, M. Q. & Liu, G. F. Expression of recombinant *Bacillus licheniformis* xylanase A in *Pichia pastoris* and xylooligosaccharides released from xylans by it. *Protein Expr. Purif.* **57,** 101–107 (2008).
54. Hu, F. *et al.* A visual method for direct selection of high-producing *Pichia pastoris* clones. *BMC Biotechnol.* **11,** 23 (2011).
55. Rumjantsev, A., Bondareva, O., Padkina, M. & Sambuk, E. Effect of nitrogen source and inorganic phosphate concentration on methanol utilization and PEX genes expression in *Pichia pastoris*. *Sci. World J.* **2014,** 9 (2014).
56. Vogl, T., Ahmad, M., Krainer, F. W., Schwab, H. & Glieder, A. Restriction site free cloning (RSFC) plasmid family for seamless, sequence independent cloning in *Pichia pastoris*. *Microb. Cell Fact.* **14,** 103 (2015).
57. Hall, C., Brachat, S. & Dietrich, F. S. Contribution of Horizontal Gene Transfer to the Evolution of *Saccharomyces cerevisiae*. *Eukaryot. Cell* **4,** 1102–1115 (2005).

58. Gmeiner, C. *et al.* Development of a fed-batch process for a recombinant *Pichia pastoris Δoch1* strain expressing a plant peroxidase. *Microb. Cell Fact.* **14,** 1 (2015).
59. Wu, S. & Letchworth, G. J. High efficiency transformation by electroporation of *Pichia pastoris* pretreated with lithium acetate and dithiothreitol. *Biotechniques* **36,** 152–154 (2004).
60. Invitrogen. Pichia Expression Kit. **31,** 79 (2010).
61. Weis, R. *et al.* Reliable high-throughput screening with *Pichia pastoris* by limiting yeast cell death phenomena. *FEMS Yeast Res.* **5,** 179–189 (2004).
62. Altschul, S. F. *et al.* Gapped BLAST and PSI- BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
63. Gordon, D., Abajian, C. & Green, P. Consed : A Graphical Tool for Sequence Finishing. *Genome Res.* **8,** 195–202 (1998).
64. Wibberg, D. *et al.* Complete genome sequencing of *Agrobacterium* sp. H13-3, the former *Rhizobium lupini* H13-3, reveals a tripartite genome consisting of a circular and a linear chromosome and an accessory plasmid but lacking a tumor-inducing Ti-plasmid. *J. Biotechnol.* **155,** 50–62 (2011).
65. Zhu, T., Guo, M., Zhuang, Y., Chu, J. & Zhang, S. Understanding the effect of foreign gene dosage on the physiology of *Pichia pastoris* by transcriptional analysis of key genes. *Appl. Microbiol. Biotechnol.* **89,** 1127–1135 (2011).
66. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25,** 402–408 (2001).

Acknowledgements

We acknowledge support for the Article Processing Charge by the Deutsche Forschungsgemeinschaft and the Open Access Publication Fund of Bielefeld University. Grants from the Federal State of North Rhine-Westphalia for the CLIB-Graduate Cluster Industrial Biotechnology are gratefully acknowledged. The bioinformatics support of the BMBF-funded project "Bielefeld-Gießen Center for Microbial Bioinformatics – BiGi (Grant number 031A533)" within the German Network for Bioinformatics Infrastructure (de. NBI) is gratefully acknowledged. We would like to thank Timo Wolf for support with the qRT-PCR experiments.

Author Contributions

J.P.S., J.K. and K.F. designed, analysed and interpreted wet lab experiments. J.P.S. and T.L. performed wet lab experiments. A.W. performed genome sequencing work. D.W. analysed and interpreted sequencing data. J.P.S. and D.W. wrote the manuscript. J.K. and K.F. revised the manuscript. J.P.S., J.K. and K.F. conceived the study. J.K. and K.F. supervised the research. All authors reviewed the final manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Schwarzhans, J.-P. *et al.* Non-canonical integration events in *Pichia pastoris* encountered during standard transformation analysed with genome sequencing. *Sci. Rep.* **6**, 38952; doi: 10.1038/srep38952 (2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

2.2.3 Supplemental data

Supplementary information

Non-canonical integration events in *Pichia pastoris* encountered during standard transformation analysed with genome sequencing

Jan-Philipp Schwarzhans[1,2], Daniel Wibberg[3], Anika Winkler[2], Tobias Luttermann[1], Jörn Kalinowski[2], Karl Friehs[1,*]

[1] Fermentation Engineering, Bielefeld University, Universitätsstr. 25, Bielefeld, 33615, Germany

[2] Microbial Genomics and Biotechnology, Center for Biotechnology (CeBiTec), Bielefeld University, Universitätsstr. 27, Bielefeld, 33615, Germany

[3] Genome Research of Industrial Microorganisms, CeBiTec, Bielefeld University, Universitätsstr. 27, Bielefeld, 33615, Germany

*Corresponding author at: karl.friehs@uni-bielefeld.de

```
MXR1             1     MKGYIFLLAPSSSKASGFAFMGSFRNFFKRSSSHMPVNSSLVSSNIRKSPQDKVVTVAGGCFWGLEHIYKMHFKDRIVDT   80
JPS496 Mutant    1     MKGYIFLLAPSSSRSNIQRRKVE*-------------------------------------------------------   24
MXR1             81    QVGFANGNVANPTYTQICLGMTHHAEVLQISYNPDVTSFNELIDFFFLVHDPTQKDRQGNDIGSQYRSAVFYLDEKEKEI   160
JPS496 Mutant          --------------------------------------------------------------------------------
MXR1             161   TEQSLAEAQKKWFPHHKIVTQVEKLTSYLDADDYHQDYLMKNPNGYHCPTHVLRKEPKSMSI*                   223
JPS496 Mutant          ------------------------------------------------------------
```

Figure S1: Alignment of the AA sequence of the native *mxr1* protein (GenBank: CCA41165.1) from *P. pastoris* CBS 7435 and the mutant found in strain JPS496, based on the coding sequence found during genome sequencing. Mutations are highlighted in red. The alignment was performed using the Constraint-based Multiple Alignment tool (Cobalt) of the NCBI. Via the NCBI Conserved Domain Database a peptide methionine sulfoxide reductase domain (underlined) was detected in the region AA 56-203 of the native protein.

```
PP7435_CHR2-1130  1    MRVLIHEKTRLIALVSSTHCLLIKYIKDKNQCAVDFVEKESFSPKHMKPLSRNKAHGFLGLINVNEDVFLCVITGKTEVA  80
JPS014 Mutant     1    MRVLIHEKTRLIALVSSTHCLLIKYIKDKNQCAVDFVEKESFSPKHMKPLSRNKAHGFLGLINVNEDVFLCVITGKTEVA  80
PP7435_CHR2-1130  81   TPTDGETVNKIYNVEFHCLNRDAWDFLELDGNGYPITTSDVNEEALRSRGSTEASDNHLSYEQNPTYELRRLLSNGSFYY  160
JPS014 Mutant     81   TPTDGETVNKIYNVEFHCLNRDAWDFLELDGNGYPITTSDVNEEALRSRGSTEASDNHLSYEQNPTYELRRLLSNGSFYY  160
PP7435_CHR2-1130  161  STNFDLTSTLQSRDVNSDSLSFDSFHLDYMWNSYMMKEVVNFRDRLPTDSKKILDRNGFLTTVIRGFAETFRTRIGHQKC  240
JPS014 Mutant     161  STNFDLTSTLQSRDVNSDSLSFDSFHLDYMWNSYMMKEVVNFRDRLPTDSKKILDRNGFLTTVIRGFAETFRTRIGHQKC  240
PP7435_CHR2-1130  241  NATIISKQSWKRAGTRYNARGIDDEGYVANFVETELILHSKDFIYAYTEVRGSVPIFWEQDTALVNPKVTITRSLEATEP  320
JPS014 Mutant     241  NATIISKQSWKRAGTRYNARGIDDEGYVANFVETELILHSKDFIYAYTEVRGSVPIFWEQDTALVNPKVTITRSLEATEP  320
PP7435_CHR2-1130  321  VFEKHFAALNGKYGPVHIVNLLSTKPSEIGLSNTYRKHFEIVNKKGSPQAYLTEFDFHKETGKNYALATKVIPFLEESIY  400
JPS014 Mutant     321  VFEKHFAALNGKYGPVHIVNLLSTKPSEIGLSNTYRKHFEIVNKKGSPQAYLTEFDFHKETGKNYALATKVIPFLEESIY  400
PP7435_CHR2-1130  401  DFDYFSYDVKNQKVLTLQKGVFRTNCLDCLDRTNVVQQVISNATLNMFLQRHNLNTNYDSDLFNKHNTLWADHGDAISQI  480
JPS014 Mutant     401  DFDYFSYDVKNQKVLTLQKGVFRTNCLDCLDRTNVVQQVISNATLNMFLQRHNLNTNYDSDLFNKHNTLWADHGDAISQI  480
PP7435_CHR2-1130  481  YTGTNALKSSFTRSGKMGIAGALSDATKSISRMYINNFQDKAKQVTIDTLLGKMSNQVEVRIFDPVSDHVDNELSKLKSQ  560
JPS014 Mutant     481  YTGTNALKSSFTRSGKMGIAGALSDATKSISRMYINNFQDKAKQVTIDTLLGKMSNQVEVRIFDPVSDHVDNELSKLKSQ  560
PP7435_CHR2-1130  561  FSAEDDIRIFTGSYNLGGTAYADDFTDWLFPKENGGIEGAPDVVILGFQEVVELTASNILNSDSSRSHYWSEEIKTQLNK  640
JPS014 Mutant     561  FSAEDDIRIFTGSYNLGGTAYADDFTDWLFPKENGGIEGAPDVVILGFQEVVELTASNILNSDSSRSHYWSEEIKTQLNK  640
PP7435_CHR2-1130  641  ISSSKYILLRSEQMTSLLLLFFIKEDKMPKVTQVEGCSKKTGLGGITANKGAVALRFSFGSTTFCLLNSHLAAGLNSVVE  720
JPS014 Mutant     641  ISSSKYILLRSEQMTSLLLLFFIKEDKMPKVTQVEGCSKKTGLGGITANKGAVALRFSFGSTTFCLLNSHLAAGLNSVVE  720
PP7435_CHR2-1130  721  RNNDFTTISQGIRFSRNKTIYDHDCVIWLGDLNYRVPLPNELVRSSALNGVYDELLAEDQLKTEMVHKGAFADFYEMKIN  800
JPS014 Mutant     721  RNNDFTTISQGIRFSRNKTIYDHDCVIWLGDLNYRVPLPNELVRSSALNGVYDELLAEDQLKHPKTKG*-----------  789
PP7435_CHR2-1130  801  FLPTYKYDKGTSVFDTSEKQRVPSWTDRILYRGKRLQQVNYNSVQSITISDHKPIYGTFKAHVTYVDERTKLTLMRKIYD  880
JPS014 Mutant          ------------------------------------------------------------------------------
```

```
PP7435_CHR2-1130  881   DYRRDPKSEASSASSNGDKFGDLIDFNDASSSSNTSIYTETEGKQVSQPVPGHFPPPPPPPPRNRVTPPQPSTPETPPPPG   960
JPS014 Mutant           ---------------------------------------------------------------------------------
PP7435_CHR2-1130  961   YALNMAPLIPSRANSGNNTRPSSPFEVSTDRAKAEISTKPVRPMVPDKPSSLKAMSPENTIQTVPPVQKAISTDSLNTAS   1040
JPS014 Mutant           ---------------------------------------------------------------------------------
PP7435_CHR2-1130  1041  AAPPPPPPRKQESINSSMSAFPVLMPKKK*                                                    1070
JPS014 Mutant           -----------------------------
```

Figure S2: Alignment of the AA sequence of the native PP7435_Chr2-1130 protein (GenBank CCA38807.1) from *P. pastoris* CBS 7435 and the mutant found in strain JPS014, based on the coding sequence found during genome sequencing. Mutations are highlighted in red, with the silent mutation c.G2346A in green. The alignment was performed via Cobalt of the NCBI. Two domains (underlined) were predicted in the native protein, using the NCBI Conserved Domain Database: a *Sac*I homology domain (AA 56-381) and a catalytic inositol polyphosphate 5-phosphatase (INPP5c) domain (AA 567-861).

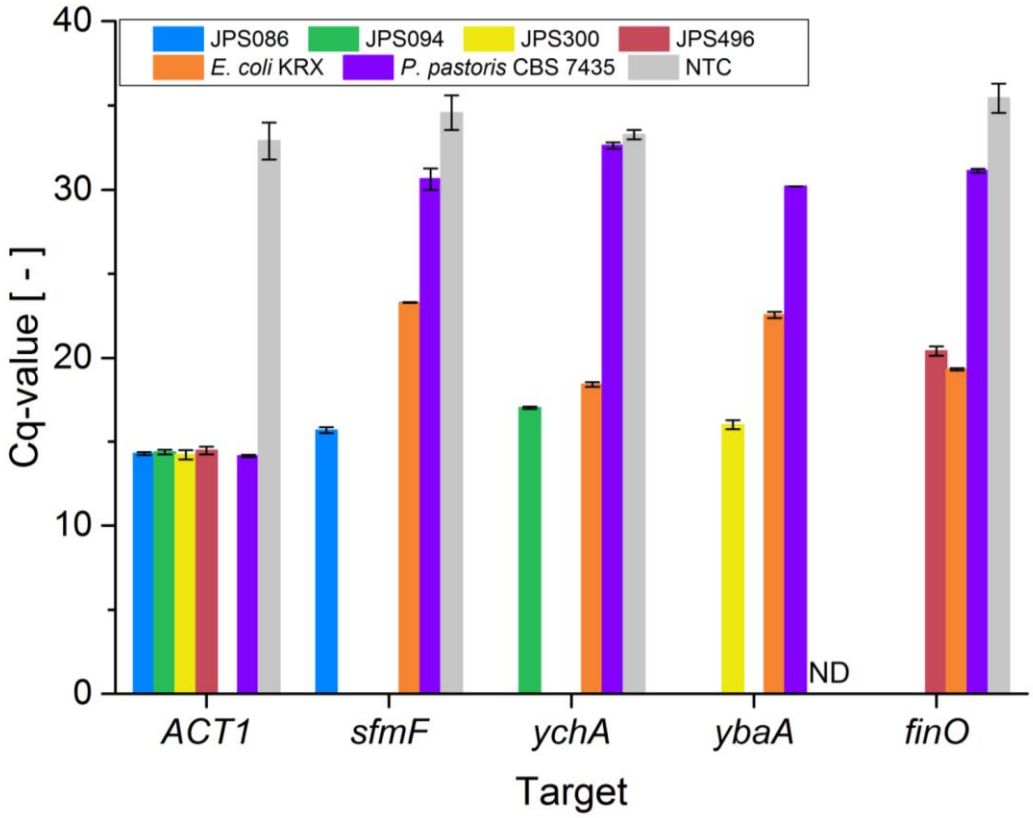

Figure S3: Cq-values measured during qRT-PCR experiments for the different target amplicons in *E. coli* KRX, *P. pastoris* CBS 7435 and the mutant strains JPS086, JPS094, JPS300 and JPS496. Additionally, the no template control (NTC) is shown. The targets are the housekeeping gene *ACT1* of *P. pastoris* as well as the *E. coli* genes *sfmF* (fimbrial protein), *ychA* (adhesin *AidA* precursor), *ybaA* (signaling protein) and *finO* (fertility inhibition). For the target *ybaA* no detectable (ND) signal was found in the NTC after 40 cycles. The melting curve analysis after 40 cycles revealed that in the *P. pastoris* CBS 7435 samples secondary amplicons were the cause for Cq-values < 33.3. Error bars indicate the standard deviation with n = 3.

2.3 Discovery and application of a novel episomal vector

2.3.1 Motivation

The previous two publications focused on the description of discovered integration events and their implications for genetic engineering and recombinant protein production. In many cases, recorded events had a negative impact on the productivity, genetic integrity or growth behavior of the strain. However, the high clonal variability did not only present challenges, it also brought forth a promising solution.

As mentioned in chapter 2.1, certain clones displayed a normalized GFP expression level markedly higher than what was expected based on their GCN. However, analysis of integration events in these strains did not reveal favourable integration loci, genome modifications or similar factors. Only a correlation between cassette-to-cassette orientation and productivity could be inferred, as discussed in chapter 2.1.2 on page 148. Yet, during genome sequencing a very unique event was detected in strain JPS664. The assembly data suggested that the expression cassette was fused to a 1.4 kb fragment of mitochondrial DNA and thereby formed a circular plasmid.

Initially, we were skeptical regarding the validity of this finding. We had found linear plasmids in some sequenced strains previously (see page 145). However, they were likely caused by loop out events from tandem arrays of multiple expression cassettes, integrated in the genome. Therefore, they did not represent replicating plasmids, but rather were a symptom of genetic instability associated with high copy clones in *P. pastoris* [144]. In JPS664 however, the plasmid was circular, making a similar origin unlikely. Additionally, the presence of mitochondrial DNA on the putative plasmid was highly surprising. Taken together with the favourable productivity characteristics of the affected strain, four key questions were sought to be clarified. (i) Is it truly a circular, replicating plasmid? (ii) What part of the plasmid mediates replication? (iii) Is the plasmid responsible for the exceptionally high productivity of the strain? (iv) Can the productivity be replicated by transforming a new strain with the plasmid?

Plasmids in yeasts have many forms and have been applied for various purposes. Native “killer plasmids” (providing growth advantages over competing yeast species) and the 2 µm plasmid (named after its monomeric contour length) of *S. cerevisiae* are the most prominent examples [145]. Autonomously replicating sequences (ARS) or centromeric

regions facilitate their replication, analogous to the ori regions of *E. coli* plasmids. In *S. cerevisiae*, engineered episomal vectors are routinely used for recombinant protein production or as part of genetic engineering approaches like CRISPR/cas9 [33, 142, 146]. On the other hand, availability and application of episomal vectors in *P. pastoris* is underdeveloped and was cited as a factor hampering further development of this non-conventional yeast as a biotechnological platform [147]. From this perspective, we were further encouraged to investigate the possible circular plasmid in strain JPS664. It might not only aid in recombinant protein production studies, but also serve as a starting point to expand the plasmid portfolio of *P. pastoris*.

The experimental setup, results and their implications for the study on this circular plasmid are detailed in the following chapter in the form of the publication "A Mitochondrial Autonomously Replicating Sequence from *Pichia pastoris* for Uniform High Level Recombinant Protein Production" (Schwarzhans *et al.*, (2017), Frontiers in Microbiology, [148]). Supplementary data of the publication is presented in the subsequent chapter 2.3.3.

2.3.2 Manuscript

frontiers in Microbiology

ORIGINAL RESEARCH
published: 02 May 2017
doi: 10.3389/fmicb.2017.00780

A Mitochondrial Autonomously Replicating Sequence from *Pichia pastoris* for Uniform High Level Recombinant Protein Production

*Jan-Philipp Schwarzhans[1,2], Tobias Luttermann[1,2], Daniel Wibberg[3], Anika Winkler[2], Wolfgang Hübner[4], Thomas Huser[4], Jörn Kalinowski[2] and Karl Friehs[1]**

[1] Fermentation Engineering, Faculty of Technology, Bielefeld University, Bielefeld, Germany, [2] Microbial Genomics and Biotechnology, Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany, [3] Genome Research of Industrial Microorganisms, CeBiTec, Bielefeld University, Bielefeld, Germany, [4] Biomolecular Photonics, Faculty of Physics, Bielefeld University, Bielefeld, Germany

Pichia pastoris is a non-conventional methylotrophic yeast that is widely used for recombinant protein production, typically by stably integrating the target gene into the genome as part of an expression cassette. However, the comparatively high clonal variability associated with this approach usually necessitates a time intense screening step in order to find strains with the desired productivity. Some of the factors causing this clonal variability can be overcome using episomal vectors containing an autonomously replicating sequence (ARS). Here, we report on the discovery, characterization, and application of a fragment of mitochondrial DNA from *P. pastoris* for use as an ARS. First encountered as an off-target event in an experiment aiming for genomic integration, the newly created circular plasmid named "pMito" consists of the expression cassette and a fragment of mitochondrial DNA. Multiple matches to known ARS consensus sequence motifs, but no exact match to known chromosomal ARS from *P. pastoris* were detected on the fragment, indicating the presence of a novel ARS element. Different variants of pMito were successfully used for transformation and their productivity characteristics were assayed. All analyzed clones displayed a highly uniform expression level, exceeding by up to fourfold that of a reference with a single copy integrated in its genome. Expressed GFP could be localized exclusively to the cytoplasm via super-resolution fluorescence microscopy, indicating that pMito is present in the nucleus. While expression levels were homogenous among pMito clones, an apparent upper limit of expression was visible that could not be explained based on the gene dosage. Further investigation is necessary to fully understand the bottle-neck hindering this and other ARS vectors in *P. pastoris* from reaching their full capability. Lastly, we could demonstrate that the mitochondrial ARS from *P. pastoris* is also suitable for episomal vector transformation in *Saccharomyces cerevisiae*, widening the potential for biotechnological application. pMito displayed strong potential to reduce clonal variability in experiments targeting recombinant protein production. These findings also showcase the as of yet largely untapped potential of mitochondrial ARS from different yeasts for biotechnological applications.

Keywords: episomal vectors, mitochondrial DNA migration, non-homologous end joining, recombinant protein production, autonomously replicating sequence consensus sequence, *Pichia pastoris*, *Saccharomyces cerevisiae*, *Komagatella phaffii*

OPEN ACCESS

Edited by:
Qiang Wang,
Institute of Hydrobiology (CAS), China

Reviewed by:
Feng Ge,
Institute of Hydrobiology (CAS), China
Cesare Indiveri,
University of Calabria, Italy

****Correspondence:***
Karl Friehs
karl.friehs@uni-bielefeld.de

Specialty section:
This article was submitted to Microbiotechnology, Ecotoxicology and Bioremediation, a section of the journal Frontiers in Microbiology

Received: *14 February 2017*
Accepted: *18 April 2017*
Published: *02 May 2017*

Citation:
Schwarzhans J-P, Luttermann T, Wibberg D, Winkler A, Hübner W, Huser T, Kalinowski J and Friehs K (2017) A Mitochondrial Autonomously Replicating Sequence from Pichia pastoris for Uniform High Level Recombinant Protein Production. Front. Microbiol. 8:780. doi: 10.3389/fmicb.2017.00780

INTRODUCTION

Since its discovery in the 1970s and the development of first molecular genetic tools in the 1980s, the non-conventional yeast *Pichia pastoris* has become a widely used host for recombinant protein production (Cregg et al., 1985; Ellis et al., 1985). Although recent research resulted in the reclassification of the most commonly used *P. pastoris* strains into *Komagatella phaffii* or *K. pastoris* (Kurtzman, 2005, 2009), the old name remains the popular choice for describing these organisms. The capability for high level protein production and secretion, post-translational modifications and ease of cultivation allowed the successful expression of a multitude of proteins, ranging from technical enzymes like phytase to biopharmaceuticals like the kallikrein inhibitor Kalbitor® (Ahmad et al., 2014; Bill, 2014). Consequently, much effort has been put into better understanding the genomic (Sturmberger et al., 2016; Valli et al., 2016), transcriptomic (Love et al., 2016) and metabolic (Rußmayer et al., 2015; Irani et al., 2016) properties of this host organism, in order to improve recombinant protein yields. In the last years, many studies provided novel regulatory elements, especially promoters for recombinant protein production in *P. pastoris* (Qin et al., 2011; Prielhofer et al., 2013; Vogl et al., 2016). The best studied and most commonly applied promoter in *P. pastoris* is the alcohol oxidase 1 (*AOX1*) promoter (Vogl and Glieder, 2013). It offers tight regulation, exceptionally high expression levels and can be induced with methanol.

Typically, the expression cassette containing the gene of interest is integrated into the chromosome via homologous recombination, enabling high genetic stability and, if desired, a simultaneous knock-out at the targeted locus (Klinner and Schäfer, 2004; Ahmad et al., 2014). Multicopy clones can be generated with different strategies to increase gene dosage and productivity (Marx et al., 2009; Aw and Polizzi, 2013, 2016). However, high copy numbers can also lead to genetic instability and the loss of expression cassettes during cultivation (Zhu et al., 2009). Furthermore, off-target integrations due to non-homologous end joining (NHEJ) events, increased cell stress caused by high gene dosage, and other as of yet not fully understood factors can lead to a heterogeneous productivity landscape in strains transformed with an integrative expression cassette (Clare et al., 1991; Hohenblum et al., 2004; Mattanovich et al., 2004; Cámara et al., 2016; Schwarzhans et al., 2016a). Our previous study revealed non-canonical NHEJ mediated integration events including reintegration of the knock-out target on a different chromosome and co-integration of *Escherichia coli* plasmid host DNA (Schwarzhans et al., 2016b). In contrast to *Saccharomyces cerevisiae*, the NHEJ pathway dominates over homologous recombination in *P. pastoris*, similar to many other yeasts and higher eukaryotes (Guirouilh-Barbat et al., 2004; Daley et al., 2005; Meyer et al., 2007; Näätsaari et al., 2012).

The classic way of overcoming some of the disadvantages associated with integrative expression cassettes, in particular genetic perturbance, is the use of episomal vectors. Since the discovery of the 2 μm plasmid in *S. cerevisiae*, several plasmids containing autonomously replicating sequences (ARS) for plasmid propagation have been developed (Futcher, 1988; Christianson et al., 1992; Newlon and Theis, 1993). Further research led to the detection of ARS in other yeasts like *Kluyveromyces lactis*, *Schizosaccharomyces pombe,* and *P. pastoris* (Peng et al., 2015). In some cases, episomal ARS vectors and integration of the expression cassette are combined into one strategy. For example, the classic *P. pastoris* ARS *PARS1* has been used both for episomal circular vectors as well as for *in vivo* amplification of linear plasmids prior to genomic integration (Lee et al., 2005; Madsen et al., 2016). A genome-wide study of *P. pastoris* GS115 led to the discovery of a multitude of (putative) ARS elements on the chromosomal DNA (Liachko et al., 2014). However, this analysis excluded the mitochondrial genome. Recently, a novel ARS originating from *K. lactis* and capable of plasmid propagation in a wide range of (non-) conventional yeasts was shown to be a promising candidate for episomal recombinant protein expression in *P. pastoris* (Liachko and Dunham, 2014; Camattari et al., 2016).

In many eukaryotes, ranging from yeasts to higher plants, animals and humans, the occurrence of mitochondrial DNA (mtDNA) on chromosomal DNA has been observed (Blanchard and Schmidt, 1996; Ricchetti et al., 2004; Hazkani-Covo et al., 2010). While the exact mechanism of mtDNA migration from the mitochondrion to the nucleus is not yet fully understood, the data suggests that the number of mtDNA integrations correlates with the genome size (Hazkani-Covo et al., 2010). It could also be shown that the integration of mtDNA into chromosomal DNA relies on the NHEJ repair of double-strand breaks (DSBs) (Ricchetti et al., 1999). In extreme cases, mtDNA integration can lead to genetic diseases (Turner et al., 2003), but most integrations have been localized to intergenic, intron or telomeric regions (Bernatzky et al., 1989; Louis and Haber, 1991; Noutsos et al., 2007). Nuclear mtDNA elements have been well-studied in *S. cerevisiae*, revealing their localizations, frequencies and properties (Sacerdot et al., 2008; Chatre and Ricchetti, 2011; Dujon, 2012). Some mtDNA elements from *S. cerevisiae* exhibit ARS activity (Gunge, 1983; Hyman et al., 1983; Delouya and Nobrega, 1991). Furthermore, in a study by Schiestl et al. (1993) aimed to induce non-homologous integrations in *S. cerevisiae*, an *in vivo* ligation of transformed DNA and mtDNA leading to the creation of a replicating plasmid was detected. So far, no data has been published on mtDNA migration, on ARS elements of mtDNA or the application of such elements from a biotechnological perspective in *P. pastoris*.

Here, we report on the discovery of a novel mtDNA ARS in *P. pastoris* and its application for episomal plasmid propagation. The ARS was first found by using genome sequencing in an experiment employing an integrative expression cassette. After validation of the presence of a circular plasmid in the affected strain, the ARS vector was assessed for its productivity characteristics. The characterization experiments indicate a uniform and high level recombinant protein production and favorable cellular localization of the product, confirmed by super-resolution fluorescence microscopy. Lastly, we could demonstrate that the mtDNA ARS can also be used for episomal transformation in *S. cerevisiae*.

RESULTS

Discovery of pMito

In our previous study, we analyzed a library of 845 *P. pastoris* clones transformed with an integrative GFP expression cassette for their productivity characteristics (Schwarzhans et al., 2016a). Based on the assayed features, interesting clones were selected for genome sequencing. Some strains were selected, because they displayed GFP expression levels that far exceeded the one predicted based on their gene copy number (GCN). One of these was strain JPS664 (EMBL FBUC01000000). With a GCN = 1 a normalized GFP expression level of ca. 1 was to be expected. However, a normalized expression level of 2.4 ± 0.3 was found. Consequently, JPS664 was selected for genome sequencing. After sequencing, it was revealed that JPS664 did not contain an expression cassette in its chromosome. Rather, a ligation of the GFP cassette to a 1.4 kb fragment of mitochondrial DNA was found. As a result of the fusion, a circular 7.3 kb plasmid was formed that was named "pMito" (**Figure 1**, EMBL LT724168). Due to the plasmid character of pMito, an ARS was suspected to be encoded on the mtDNA fragment. The mitochondrial DNA shows 100% identity to the bases 27,552–28,993 of the mitochondrial genome of *P. pastoris* CBS7435 containing a fragment of the *COX1* (cytochrome c oxidase I) gene. The last 73 bp of the second *COX1* exon as well as 1369 bp of the second *COX1* intron were detected on the segment. The combination of a novel mtDNA *in vivo* ligation to an expression cassette in *P. pastoris*, potential ARS activity, and the apparently high suitability of the plasmid for recombinant protein production prompted us to conduct further experiments on pMito.

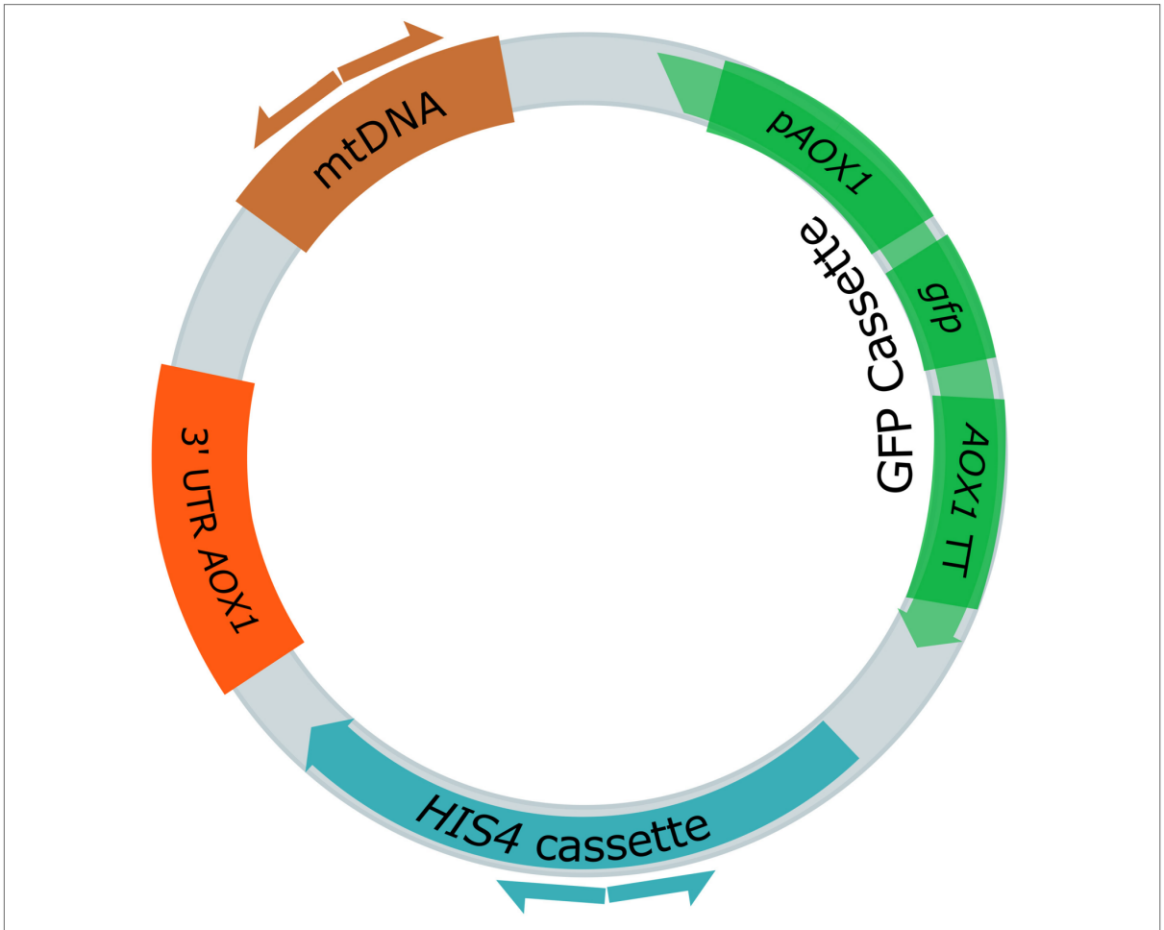

FIGURE 1 | Plasmid map of the 7.3 kb pMito, as discovered in *Pichia pastoris* JPS664. The displayed features include the GFP cassette, consisting of the *AOX1* promoter (p*AOX1*), *gfp* gene and *AOX1* terminator (*AOX1* TT), the *HIS4* cassette as selection marker, the 3′ UTR *AOX1* (in original vector for homologous integration purposes) and a fragment of mtDNA. Direction and binding sites of primers pairs used for PCR assays are shown as differently colored half arrows: pMito_Circ-mtDNA-FW/RV (brown) and pMito_Circ-HIS4-FW/RV (blue).

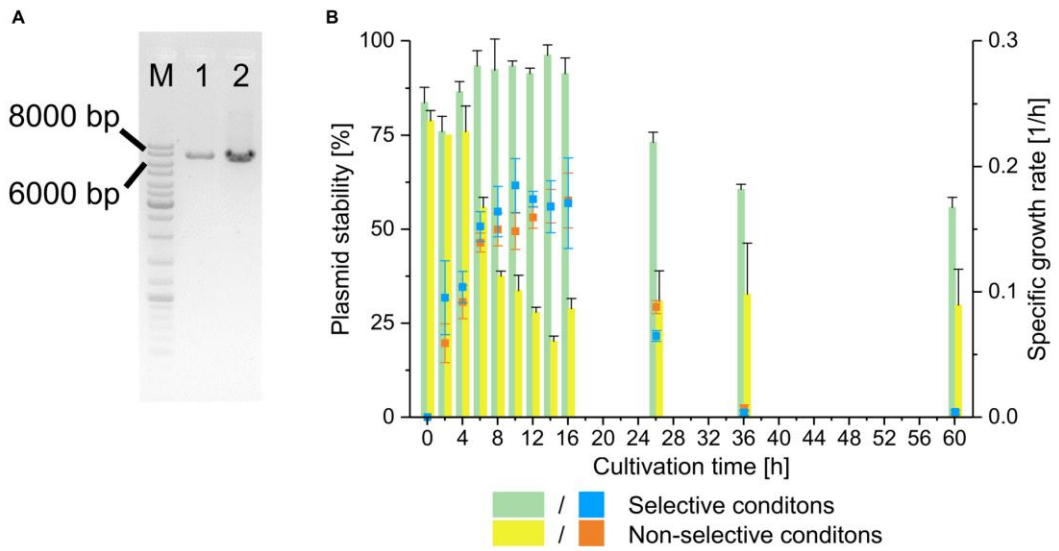

FIGURE 2 | Validation of pMito circular plasmid characteristics. (A) PCR assay for presence of circular pMito in JPS664. Lanes: (M) Marker (1) Primer pair pMito_Circ-HIS4-FW/RV (2) Primer pair pMito_Circ-mtDNA-FW/RV. Relevant band sizes have been highlighted. **(B)** Plasmid stability (columns) and specific growth rate (symbols) of pMito in JPS664, grown under selective and non-selective conditions. Error bars represent the standard deviation, with $n = 2$.

Validation of Plasmid Character of pMito

First, the plasmid character of pMito had to be validated in order to ensure that it was not an artifact based on genome sequencing and assembly. To this end, a PCR assay with two sets of directly adjacent, diverging primer pairs was designed. This way, a PCR product of the same size as the predicted pMito would validate its circular structure. One primer pair binds in the *HIS4* region of pMito and the other in the mtDNA fragment. In **Figure 2A** the PCR assay visibly indicates the presence of a full length circular plasmid as shown in **Figure 1**. In addition, the plasmid stability of pMito was investigated (**Figure 2B**). Under selective conditions pMito is well-maintained at up to 96.2 ± 2.7%, while under non-selective conditions the plasmid stability was lowest at 20.2 ± 1.4%. In combination with the specific growth rate, a growth associated loss and increase of pMito content is apparent, respectively. On the one hand, pMito content rose, while cells grew in selective media and stagnated or fell when they reached the stationary phase, indicating that cells containing pMito lost their growth advantage over plasmid-free cells. On the other hand, cells grown in non-selective media quickly lost pMito during the exponential growth phase and exhibited constant plasmid contents in the stationary phase, indicating a growth advantage of plasmid-free cells. Taken together, the plasmid stability and PCR assay clearly validated the circular plasmid character of pMito.

In Silico Analysis of pMito

It was therefore highly likely, that the mtDNA fragment in pMito conferred ARS activity. Using the Find Individual Motif Occurrences (FIMO) tool (Grant et al., 2011), the mtDNA fragment was scanned for known ARS consensus sequence (ACS) motifs. Both AT- and GC-rich ACS motifs from *P. pastoris* as well as the 11 and the 17 bp ACS motifs of *S. cerevisiae* were included (Newlon and Theis, 1993; Theis and Newlon, 1997; Liachko et al., 2014). Despite the low GC-content (22%) of the mtDNA fragment, one putative GC-ACS was found in addition to several AT-ACS and *S. cerevisiae* ACS with a total of 33 matches (**Table 1**). A more detailed description of the FIMO matches can be found in Supplementary Table S1. By mapping the FIMO matches to the sequence of the mtDNA fragment, multiple clusters of predicted ARS become visible (**Figure 3**). All putative ARS were found in the *COX1* intron on the mtDNA. Especially in the regions at 1256–1300 bp and 701–753 bp matching sequences were found, with up to 6 and 4 cumulative matches, respectively. Here, similar sequences to the *P. pastoris* AT-ACS, *S. cerevisiae* 11 bp ACS and *S. cerevisiae* 17 bp ACS were detected. The accumulation of different types of ACS indicates a high likelihood of actual ARS functionality in these regions. Additionally, between 2 and 3 cumulative matches can

TABLE 1 | Number of ACS motif matches found on the mtDNA fragment in pMito using FIMO.

Motif	Number of matches
P. pastoris AT-ACS	15
P. pastoris GC-ACS	1
S. cerevisiae 11 bp ACS	11
S. cerevisiae 17 bp ACS	6

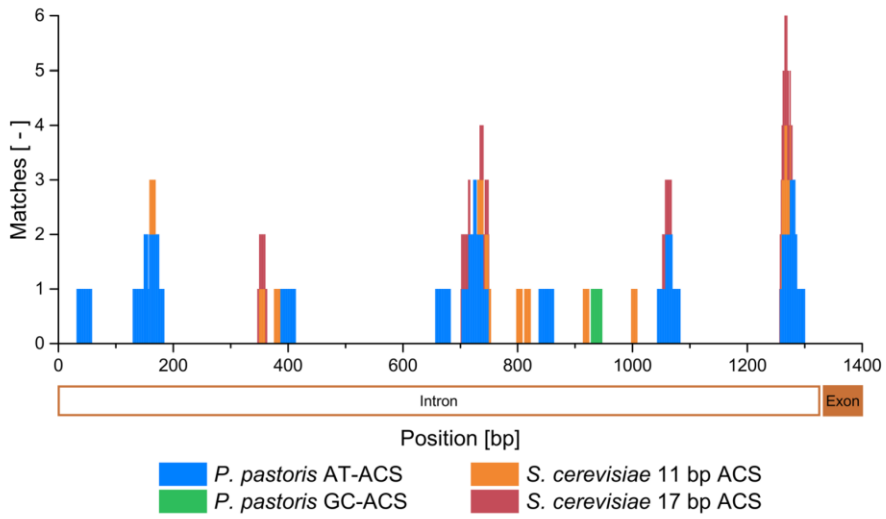

FIGURE 3 | Cumulative matches of the FIMO scan for *P. pastoris* AT- and GC-ACS, as well as *S. cerevisiae* 11 and 17 bp ACS motifs mapped to the mtDNA fragment of pMito. The corresponding second exon and second intron region of the *COX1* gene are shown below the x-axis in an orientation from right to left.

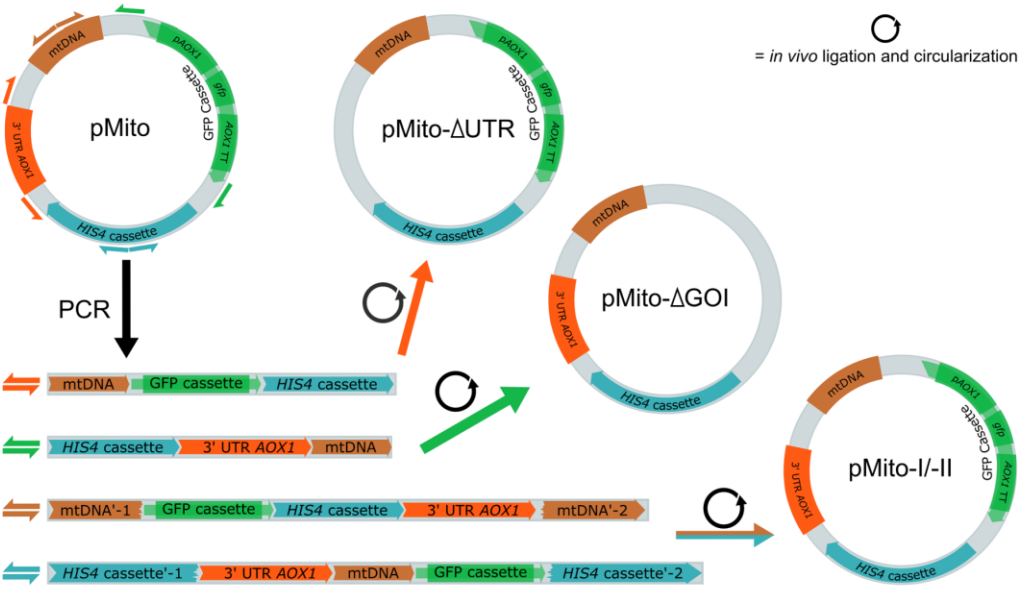

FIGURE 4 | Schematic overview of the construction and *in vivo* assembly of four different pMito variants used for transformation. Direction and primer binding sites of primers pairs used for construction are shown as differently colored half arrows: pMito_Circ-mtDNA-FW/RV (brown), pMito_Circ-HIS4-FW/RV (blue), pMito_No_GOI-FW/RV (green) and pMito_No_UTR-FW/RV (orange). In the first step, the variants were amplified as linear DNA via PCR. After successful transformation into *P. pastoris*, linear DNA was *in vivo* ligated into circular plasmids.

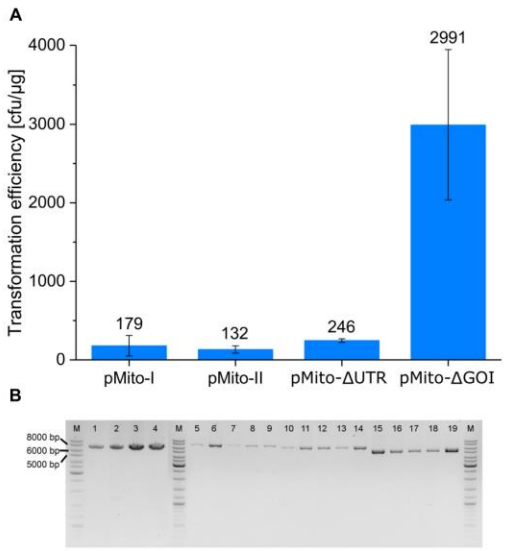

FIGURE 5 | (A) *P. pastoris* transformation efficiencies of the four pMito variants. Error bars represent the standard deviation, with $n = 3$. **(B)** PCR assay for circular plasmids in different pMito clones, using the primer pair pMito_Circ-mtDNA-FW/RV. Lanes: (M) Marker (1–4) pMito-I strains (5–9) pMito-II strains (10–14) pMito-ΔUTR strains (15–19) pMito-ΔGOI strains. Relevant band sizes have been highlighted.

be found in the regions of 130–184, 347–363, and 1043–1083 bp. In these regions as well, ARS activity could be present. Although *P. pastoris* ACS motif matches were found, no region of the mtDNA fragment could be directly aligned via BLASTn to the library of (putative) ARS sequences on the *P. pastoris* GS115 chromosomes (Liachko et al., 2014). This indicates that the ARS on pMito, while similar to the ARS found on the chromosomes, has its own distinguishable sequence. A preliminary FIMO scan of the mitochondrial genome of *P. pastoris* CBS7435 (GenBank: FR839632) identified 20 GC-ACS and over 500 AT-ACS motif occurrences (data not shown).

The expression cassette portion of pMito, excluding the mtDNA fragment, was also scanned for ACS motif occurrences, as described above. In spite of being approximately four times as long as the mtDNA fragment, only five matches were found. They are detailed in Supplementary Table S2 and the mapping result is shown in Supplementary Figure S1. Interestingly, three of these matches accumulate within a 43 bp region of the *AOX1* terminator and two occur in the 3′ UTR *AOX1* element. Therefore, we conducted transformation experiments with different variants of the plasmid.

Transformation of pMito Variants into *P. pastoris*

Using the original pMito from JPS664 as template, four different variants were created via PCR (**Figure 4**). They differed in two aspects. Firstly, the constructs pMito-I and pMito-II encompass the complete plasmid, while pMito-ΔUTR (untranslated region) omits the 3′ UTR *AOX1* region and pMito-ΔGOI (gene of interest) does not contain the *AOX1* promoter, *gfp* gene and *AOX1* terminator. Secondly, different loci were used for linearization. pMito-ΔUTR and pMito-ΔGOI were linearized by removing the aforementioned segments. pMito-I was linearized inside the mtDNA fragment and pMito-II inside the *HIS4* gene. In accordance with our previous recommendations, *P. pastoris* was transformed with linear, PCR amplified DNA (Schwarzhans et al., 2016b). This method also offered the advantage to assay *P. pastoris* capability for *in vivo* circularization of the transformed DNA. In the case of pMito-II, only a successfully circularized plasmid could bestow histidine prototrophy, since the *HIS4* gene was split onto the distal ends of the transformed, linear DNA.

All four constructs produced transformants (**Figure 5A**). Using the PCR assay described in the previous chapter, the presence of full-length circular plasmids could be confirmed in strains of all four constructs (**Figure 5B**). The successful transformation of pMito-II confirmed that *P. pastoris* was capable of *in vivo* ligating the linear DNA into circular plasmids. Depending on the aim of an experiment, this ability can be used, e.g., for DNA assembly purposes. Transformation of pMito-I, pMito-II, and pMito-ΔUTR resulted in comparable efficiencies, with an average of about 190 colony forming units (cfu)/µg. In contrast, an efficiency more than 10-fold higher at ca. 3000 cfu/µg was encountered using pMito-ΔGOI.

Productivity Characterization of pMito-I Strains

Since the original strain JPS664 in which pMito was found exhibited favorable productivity characteristics, it was of interest to investigate if these properties could be replicated. To this end, 16 clones transformed with pMito-I were randomly selected and submitted to further analysis. The productivity of the selected clones was assayed in relation to the original JPS664 and the reference strain JPS066, which contained a single expression cassette integrated into the *AOX1* locus on the genome (Schwarzhans et al., 2016a). All selected pMito-I clones produced high amounts of GFP, markedly exceeding the reference strain at least two- and up to fourfold (**Figure 6A**). Applying the Student's *t*-test (two-sided), it was determined that all pMito clones, and the original strain JPS664, produced GFP at significantly higher levels than JPS066, with the vast majority of pMito-I strains scoring a p-value < 0.01. A highly uniform expression level of pMito-I strains is visible, with all but one clone (pMito-I 2) exhibiting normalized expression levels in the range of 2.5–3.9. It seemed, that from the normalized expression levels an upper limit of GFP productivity can be deduced. To investigate whether the gene dosage could explain this behavior all clones were subjected to GCN analysis. As can be seen in **Figure 6B**, expression level and GCN did not correlate. While GCN values between 0.6 and 5.5 were encountered, the gene dosage of a strain had apparently no impact on the expression level.

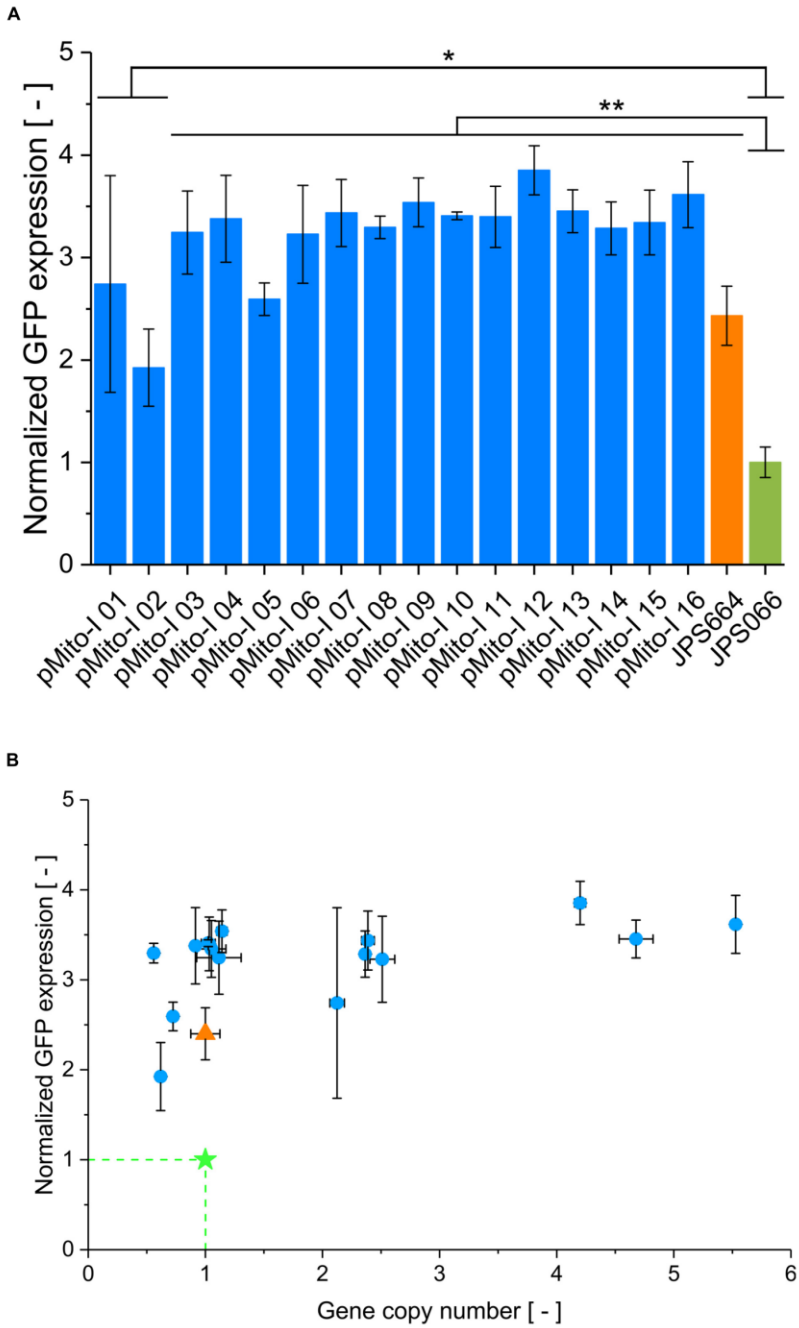

FIGURE 6 | Productivity characterization of 16 pMito-I strains (A) Normalized expression level of 16 pMito-I strains, JPS664 and JPS066. Applying the Student's *t*-test (two-sided), the significance of divergence from JPS066 was assayed. *p-value < 0.05 and **p-value < 0.01. Error bars represent the standard deviation, with $n = 3$. **(B)** Normalized expression level and GCN of the 16 pMito-I strains (blue dots). JPS066 is shown as green star and JPS664 as orange triangle. Error bars indicate the standard deviation, with $n = 3$.

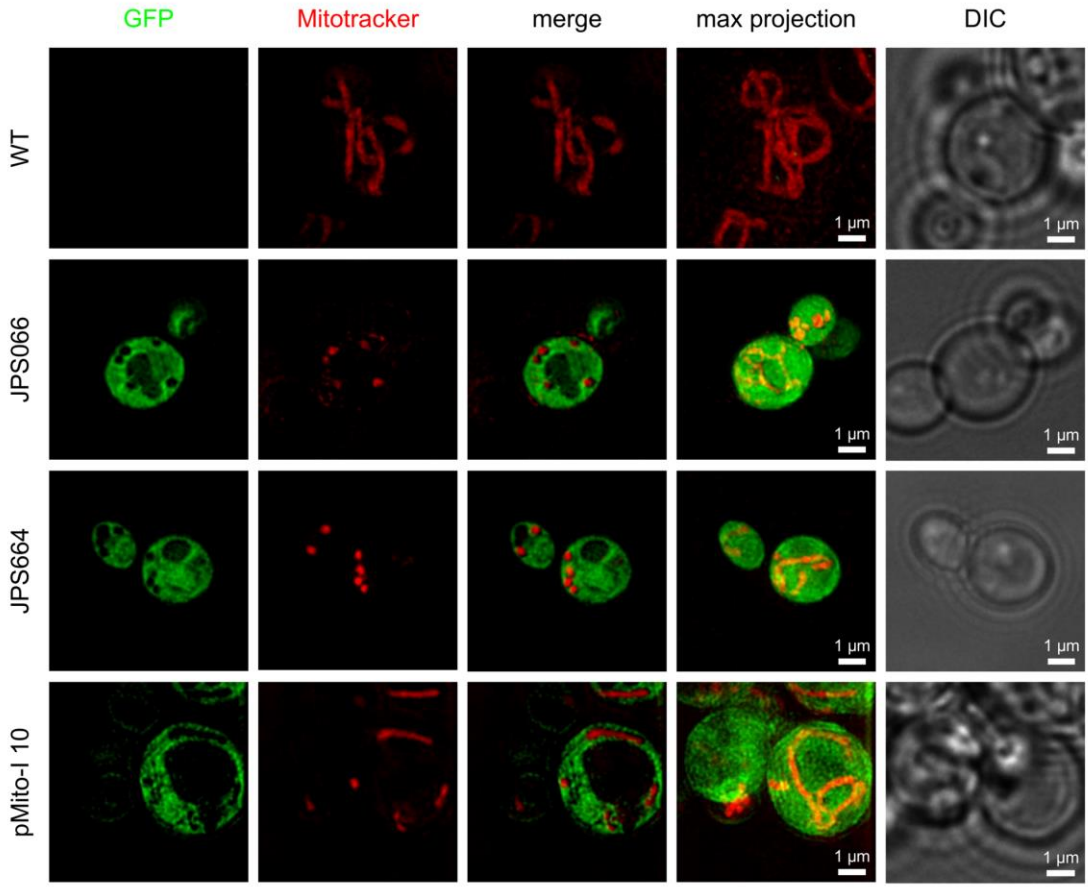

FIGURE 7 | Super-resolution fluorescence microscopy images of living *P. pastoris* cells expressing GFP and stained with MitoTracker® Red CM-H_2XRos. Images were acquired in the 3D SR-SIM mode. The *P. pastoris* CBS7435 wild type (WT), strains JPS066 and JPS664, as well as clone pMito-I 10 were analyzed. For all strains, a single z plane image of the GFP and MitoTracker signal are shown separately, merged and then as maximum intensity projection of the recorded entire cell z-stack (4–8 µm depth). In the last column a corresponding DIC image is displayed.

Intracellular Localization of Expressed GFP

Not only the expression level but also the localization of the product is of interest. For correct post-translational modification and secretion of the recombinant protein, it must first be present in the cytoplasm, so it can translocate to the endoplasmic reticulum and Golgi apparatus. A post-translational translocation of GFP to the mitochondria was unlikely due to the absence of a corresponding signal sequence. However, it was possible that copies of pMito were present in the mitochondria, especially for JPS664. Here, the exact origin of pMito was uncertain and could result in GFP accumulation in the mitochondria, complicating product capture. Living cells of the *P. pastoris* CBS7435 wild type (WT), JPS664, JPS066 and strain pMito-I 10 were analyzed via super-resolution fluorescence microscopy. In all GFP expressing strains the recombinant protein was present in the entire cytoplasm (**Figure 7**). Two major compartments absent of GFP fluorescence are distinguishable. They are likely the peroxisome and the mitochondria. This and more findings are analyzed in more detail in Section "Discussion".

mtDNA ARS from *P. pastoris* in *S. cerevisiae*

As previously detailed, ACS motifs from *S. cerevisiae* could be matched to the mtDNA segment of pMito. Therefore, we were interested to see if pMito could also be used as an episomal vector in *S. cerevisiae*. A variant of the plasmid pYES2 was constructed, replacing the 2 µm origin with the mtDNA fragment of pMito, resulting in pYES2-Mito. For the purpose of better comparison with the *Pichia* experiments, *S. cerevisiae* was transformed with linear, PCR amplified constructs. Both constructs facilitated creation of uracil prototroph *S. cerevisiae*

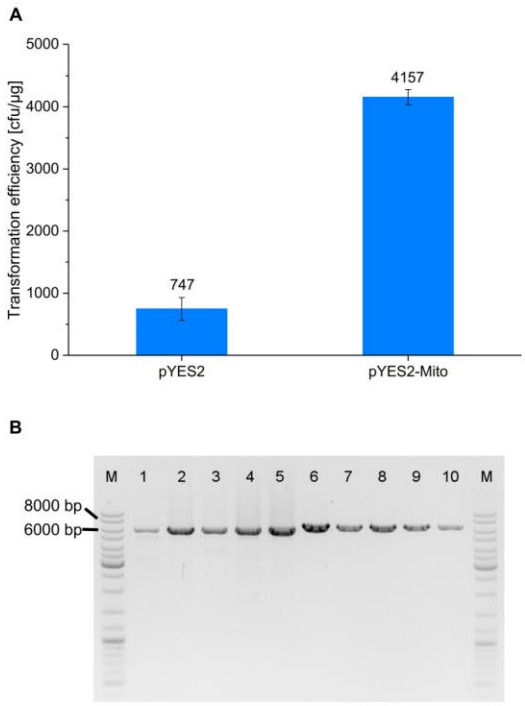

FIGURE 8 | (A) *S. cerevisiae* transformation efficiencies of pYES2 (5.9 kb) and pYES2-Mito (6.4 kb). Error bars represent the standard deviation, with $n = 3$. **(B)** PCR assay for circular plasmids in pYES2 and pYES2-Mito clones. Lanes: (M) Marker (1–5) PCR product of five pYES2 strains, using the primer pair pYES2_Circ-FW/RV (6–10) PCR product of five pYES2-Mito strains, applying the primer pair pMito_Circ-mtDNA-FW/RV. The size of relevant bands has been highlighted.

clones (**Figure 8A**) and the presence of circular plasmids was confirmed via PCR (**Figure 8B**). Surprisingly, the transformation efficiency of pYES2-Mito was almost sixfold higher as for pYES2.

DISCUSSION

During an experiment targeting the genomic integration of a GFP expression cassette in *P. pastoris* the random creation of the circular plasmid pMito, consisting of the expression cassette and a 1.4 kb fragment of mtDNA, was discovered. Due to the good productivity characteristics of the affected strain the plasmid was further investigated.

It is likely that pMito is the result of NHEJ mediated DSB repair, consistent with previous reports on the involvement of mtDNA in DSB repair (Ricchetti et al., 1999). So far no involvement of mtDNA in DSB repair or the formation of replicating plasmids due to its involvement have been reported for *P. pastoris*. Potentially, *P. pastoris* ligated the transformed, linear expression cassette with the fragment of mtDNA in order to "repair" the expression cassette. This untargeted incident is reminiscent of other non-canonical NHEJ mediated integration events in *P. pastoris* we have reported on previously, e.g., the *in vivo* ligation of *E. coli* DNA fragments to the expression cassette (Schwarzhans et al., 2016b). The presence of mtDNA in the nucleus has been reported before, although the exact mechanism of the mtDNA migration is unclear (Hazkani-Covo et al., 2010). However, similar events have been observed in *S. cerevisiae*. For this model yeast it is known that its mtDNA is rich in ARS and can migrate to the chromosomes (Hyman et al., 1983; Sacerdot et al., 2008). Furthermore, it has been shown that under conditions that favor NHEJ, *in vivo* ligation of transformed DNA to mtDNA can occur, forming replicating plasmids (Schiestl et al., 1993).

Applying two sets of diverging primer pairs the circular structure of pMito was confirmed via PCR. During the plasmid stability assay it is possible that L-histidine was released into the medium by prototrophic strains, making it easier for plasmid-free cells to propagate, even under selective conditions. Similar observations have been made in experiments with other auxotrophic yeast strains (Pronk, 2002). It seems likely that if nutrient limitation could be avoided and cells had continued to grow, the plasmid content would have dropped further under non-selective conditions.

Applying the FIMO tool and known ACS motifs from *P. pastoris* and *S. cerevisiae*, multiple matches for potential ARS sites were found on the mtDNA fragment. Besides the mtDNA segment, five additional ACS motif matches were found on the *AOX1* terminator associated regions of pMito. Some *S. cerevisiae* terminator regions are associated with ARS activity (Chen et al., 1996). However, Vogl et al. (2016) tested various *P. pastoris* terminators for ARS functionality and determined that the *AOX1* terminator has no ARS activity. Especially in the regions 701–753 and 1256–1300 bp of the mtDNA segment, a clustering of putative ARS sequences was observed, making them promising candidates for further experiments. For instance, they could serve as starting point to reduce the size of the mtDNA fragment needed for ARS activity. Interestingly, none of the predicted ARS matched the ones found on the chromosomes of *P. pastoris* (Liachko et al., 2014), indicating the presence of a novel ARS sequence on pMito. Over 500 additional ACS matches were found in a preliminary FIMO scan of the mitochondrial genome of *P. pastoris* CBS7435, suggesting it is tightly packed with potential ARS elements, similar to *S. cerevisiae* (Hyman et al., 1983).

We noticed a more than 10-fold increase in transformation efficiency, if using the pMito-ΔGOI construct compared to pMito-I, -II, and pMito-ΔUTR. Two main factors might have contributed to this increase in transformation efficiency. Firstly, pMito-ΔGOI is the shortest of the four constructs with 5.3 kb (pMito-I and -II: 7.3 kb, pMito-ΔUTR: 6.6 kb). The smaller size could have eased entry of the foreign DNA into the cell or nucleus and made circularization of the linear DNA via *in vivo* ligation more efficient, e.g., by containing less secondary structures than longer variants. To this end, the above mentioned clustering of putative ARS on the mtDNA segment could serve as starting point for constructing smaller variants with higher transformation efficiency. Secondly, by omitting the GOI region three ACS motif matches in the *AOX1* terminator are eliminated. Although, it has been shown that the *AOX1* terminator possesses

no ARS activity, the matching sequences could still facilitate interaction with the origin recognition complex (ORC), needed for DNA replication (Bell and Stillman, 1992). Therefore, in constructs containing the *AOX1* terminator potentially less ORC was free to initiate DNA replication at one of the ARS sites of the mtDNA of pMito. Especially, directly after transformation, when cell survival is most dependent on the histidine prototrophy conferred by the plasmid encoded *HIS4*, a delayed or inhibited replication of pMito could lead to cell death. This interference would also cause lower plasmid stability, meaning a lower frequency of daughter cells containing pMito. In consequence, even a successful transformation might have resulted in no cell growth or a markedly delayed cell growth. A similar, albeit much less pronounced, trend is visible when comparing the slightly increased transformation efficiency of pMito-ΔUTR to pMito-I and -II. Here, two ACS matches contained in the 3′ *AOX1* UTR are removed and the vector size is slightly reduced compared to the full length pMito.

In comparison to recently published results for ARS based protein production in *P. pastoris*, pMito surpasses the two tested chromosomal ARS vectors from *P. pastoris* in productivity and is approximately on par with the ARS vector from *K. lactis* (Camattari et al., 2016). Normalized expression levels were up to fourfold higher than the reference strain JPS066 and uniform among 16 pMito-I clones and the original JPS664 strain. Typically, transformation of *P. pastoris* results in strains with a wide range of productivities. For example, Clare et al. (1991) observed up to 30-fold differences in product titer, while Cámara et al. (2016) and our previous study (Schwarzhans et al., 2016a) reported on clonal variabilities ranging from non-producing to exceedingly high producing strains in transformants from one experiment. The high uniformity of pMito strains could help streamline screening procedures, which typically are time and work intense steps necessary in order to find strains with the desired productivity (Looser et al., 2015).

However, a seemingly gene dosage independent upper limit of GFP productivity was apparent. A similar phenomenon has been encountered by Camattari et al. (2016) in their characterization of different *P. pastoris* strains expressing blue fluorescent protein (BFP) from an ARS-based episomal vector. As in their case, recombinant protein expression might have been affected by post-translational or epigenetic factors (Love et al., 2010). It has to be noted that in our case GFP was expressed in the cytoplasm. However, the occurrence of highly similar phenomena in two studies working with episomal vectors in *P. pastoris* could point to a common origin. Especially, the potential connection to ARS based vectors would need additional investigation for clarifying the root cause, e.g., whether *Pichia* can distinguish between episomal and chromosomal expression and direct its resources accordingly. Further research is needed to fully understand the bottle-neck that might be preventing ARS based vectors from reaching their theoretical potential in *P. pastoris*.

It is known, that yeast mitochondria are capable of facilitating protein synthesis within the organelle (Herrmann et al., 2013). We wanted to ensure, that recombinant protein produced with pMito localized to the cytoplasm, enabling biotechnological application. To this end, we applied super-resolution fluorescence microscopy, which allows for the intracellular localization of fluorescent targets (Huang et al., 2009). This study marks the first reported implementation of this technique for analysis of *P. pastoris*. We could localize expressed GFP to the cytoplasm and identified two major compartments devoid of GFP. Firstly, what is assumed to be peroxisomes which have increased to multiple times their normal size due to the methanol induction (Johnson et al., 1999). Secondly, via a combination of the GFP and MitoTracker signal it becomes apparent, that the mitochondria contained no GFP. Best visible in the maximum intensity projection images, a tubular organization of the mitochondria can be clearly seen. This projection uses the entire recorded z-stack of the cell and thereby gives augmented information on the 3D spatial organization of fluorescence tagged targets. A similar structural organization was reported for *S. cerevisiae* mitochondria (Egner et al., 2002). Super-resolution fluorescence microscopy allowed for a more detailed insight into the spatial organization of mitochondria inside the living cell. This technique could therefore be of benefit for experiments studying intracellular localization of fluorescence-tagged targets in *P. pastoris*, where so far confocal microscopy has been used primarily (Heiss et al., 2015; Rueda et al., 2016). Cells were not fixed to avoid loss of the MitoTracker signal, therefore the "rings" visible in the DIC images are suspected to be an artifact caused by the living cells.

Lastly, the ARS activity in *S. cerevisiae* of the mtDNA segment from pMito was demonstrated via the plasmid pYES2-Mito. Higher transformation efficiencies were recorded for pYES2-Mito than for the original pYES2. This suggests that the ARS of pYES2-Mito is not only active in *S. cerevisiae*, but also easier to transform than the 2 μm based vector. However, it does not necessarily mean that pYES2-Mito outperforms pYES2 and other 2 μm based vectors in regards to recombinant protein productivity. 2 μm plasmids are known for their high copy number, enabling an increased gene dosage of the target gene (Christianson et al., 1992). It has to be noted, that the recombinant protein productivity in *S. cerevisiae* was not yet assayed but characterization of the applicative potential of pMito-derived vectors in this organism is subject to further studies. As it stands, the results clearly suggest the inter-genus capability of the ARS encoded on the mtDNA fragment of pMito.

In summary, the mtDNA fragment of pMito represents a promising candidate for ARS based recombinant protein production in *P. pastoris*, reducing clonal variability while providing increased expression levels. To fully gauge its production capabilities, the expression of additional (secreted) heterologous proteins via pMito is advisable. It could present itself as an alternative to integrative expression cassettes, especially when screening procedures are the limiting step. Combined with the high density of putative ARS on the whole mitochondrial genome of *P. pastoris*, a wealth of so far unused mtDNA ARS could benefit biotechnological and basic science approaches. Considering the relatively wide taxonomic distance of *P. pastoris* (order: Saccharomycetales; family: *Phaffomycetaceae*) and *S. cerevisiae* (order: Saccharomycetales; family: *Saccharomycetaceae*), the discovered ARS seems to be

capable of a wider host range activity, akin to the *K. lactis* ARS "panARS," discovered and optimized by Liachko and Dunham (2014). For panARS a very broad host range of 10 budding yeasts was shown, prompting us to explore a wider host range for pMito in further studies.

MATERIALS AND METHODS

Microorganisms and Cultivation Conditions

All plasmids were constructed and propagated in *E. coli* Top 10 [Invitrogen, USA; genotype: F- *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) Φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *ara*D139 Δ(*araleu*)7697 *gal*U *gal*K *rps*L (StrR) *end*A1 *nup*G]. *E. coli* cultivations were carried out in LB (Lysogeny Broth) medium supplemented with 100 μg/mL ampicillin. Experiments involving *P. pastoris* employed the WT strain CBS7435 (identical to NRRL Y-11430 and ATCC76273), obtained from the Spanish Type Culture Collection (Valencia, Spain) under the designator CECT11047, as well as the histidine-auxotrophic strain *P. pastoris* CBS7435 (Δ*HIS4*) from the Austrian Center of Industrial Biotechnology (Graz, Austria). *S. cerevisiae* INVSc1 (genotype: *MATa HIS3*D1 *LEU2 TRP1*-289 *URA3*-52 *MAT HIS3*D1 *LEU2 TRP1*-289 *URA3*-52) from Invitrogen (Waltham, MA, USA) was used for *Saccharomyces* experiments. Shake flask cultivations of both *P. pastoris* and *S. cerevisiae* were performed using YPD (Yeast Peptone Dextrose) medium, as well as BMD (Buffered Minimal Dextrose) medium (Invitrogen, 2010) for *P. pastoris* and SC minimal medium (Invitrogen, 2008) for *S. cerevisiae*. If necessary, BMD medium was supplemented with 4 mg/L L-histidine, and SC medium was supplemented with 0.05 g/L L-histidine, 0.1 g/L L-leucine, and 0.1 g/L L-tryptophan. *P. pastoris* expression levels were assayed in 96-deep-well plates with 2.4 mL total volume (Eppendorf, Hamburg, Germany) according to Weis et al. (2004) and Hartner et al. (2008), using BMD, BMM2 (Buffered Minimal Methanol) and BMM10 medium. By applying this method, the cells are first grown in BMD and expression is induced with BMM2 and BMM10 by maintaining a 0.5% (*v/v*) methanol concentration in the culture medium. All yeast cultivations were carried out at 28°C, with shake flasks (baffled) being agitated at 120 min^{-1} and deep-well plates at 340 min^{-1}.

Genome Sequencing and Bioinformatic Analysis

Genomic DNA (gDNA) was isolated from yeast cultures using the MasterPure™ Yeast DNA Purification kit (Epicentre, Madison, WI, USA). The method used for genome sequencing of relevant *P. pastoris* strains was recently described (Schwarzhans et al., 2016a). In short, gDNA quality was assayed via gel-electrophoresis and gDNA of sufficient quality was quantified using the Quant-iT PicoGreen dsDNA kit by Invitrogen (Waltham, MA, USA). From samples of high quality and quantity, paired-end libraries were prepared by applying the TruSeq sample preparation kit (Illumina, San Diego, CA, USA). The libraries were sequenced on an Illumina MiSeq system. Raw data was *de novo* assembled using the GS *De Novo* Assembler (Version 2.8, Roche, Basel, Switzerland) with default settings. The assembled draft genome of *P. pastoris* JPS664 can be found under FBUC01000000, and the finalized sequence of pMito under LT724168 in the EBI database.

For bioinformatic analysis the BLASTn algorithm (Altschul et al., 1997) and a local database including the pAHBgl-GFP vector sequence were used for database comparison. Only hits with a sequence identity of 100% and an e-value $> 1 \times 10^{-20}$ were further analyzed in more detail. If necessary, gaps in the vector were closed via an *in silico* approach with CONSED (Gordon et al., 1998; Küberl et al., 2011; Wibberg et al., 2011). This approach allowed to determine the exact locus for the expression cassette in the sequenced *P. pastoris* genome.

Construction and Transformation of Vectors

Primers were designed in SnapGene (GSL Biotech, Chicago, IL, USA) and their sequences can be found in Supplementary Table S3. PCR procedures employed the Phusion® High-Fidelity DNA Polymerase (New England Biolabs, Ipswich, MA, USA). For agarose gel electrophoresis analysis, the GeneRuler DNA Ladder Mix by Thermo Scientific (Waltham, MA, USA) was used as size marker. The *P. pastoris* strain JPS664 was created as described in Schwarzhans et al. (2016a) with a *Bgl*II digested pAHBgl-GFP plasmid. pMito and its variants were amplified via PCR using the original pMito as template. Potentially due to secondary structures, the variants pMito-ΔUTR and pMito-ΔGOI consistently resisted PCR amplification. As a compromise, primers were designed that bind slightly inside the region targeted for omission. In consequence, pMito-ΔUTR and pMito-ΔGOI still contain 63 and 65 bp of the targeted region, respectively. pYES2 was obtained from Invitrogen (Waltham, MA, USA). It can be used for the transformation of *S. cerevisiae* with an episomal vector containing a 2 μm sequence for replication and an *URA*3 selection marker (Invitrogen, 2008). For the construction of pYES2-Mito the mtDNA fragment on pMito was PCR amplified and combined via Gibson Assembly (Gibson et al., 2009) with the linearized pYES2 without 2 μm sequence.

Pichia pastoris CBS7435 (Δ*HIS4*) was transformed according to Wu and Letchworth (2004) and *S. cerevisiae* INVSc1 according to Thompson et al. (1998) using PCR amplified, linear DNA. Per transformation approximately 500 ng of purified (Wizard® Plus SV Minipreps DNA Purification System, Promega, Madison, WI, USA) DNA was used. After transformation the *Pichia* cells were immediately spread onto MD (Minimal Dextrose) plates (Invitrogen, 2010), and *Saccharomyces* cells onto SC plates without uracil (Invitrogen, 2008) in 200 μL aliquots and incubated for 3 days at 28°C. Following the incubation, the total number of transformants was counted and randomly selected clones picked for dilution plating. Single colonies from dilution plating were used for following experiments, including PCR assays for the correct construct, expression screenings and GCN determination. Experiments for assaying the transformation efficiency were carried out in biological triplicates.

In Silico Analysis of pMito ARS

The mtDNA fragment on pMito was scanned for ACS motifs via the FIMO tool (Grant et al., 2011) of the MEME suite (Bailey et al., 2009). Both the 11 and 17 bp ACS motifs of *S. cerevisiae* (Newlon and Theis, 1993; Theis and Newlon, 1997), as well as the GC-ACS and AT-ACS motifs of *P. pastoris* (Liachko et al., 2014) were used as references.

Plasmid Stability Assay

Precultures of *P. pastoris* JPS664 were grown overnight under selective conditions in MD medium without L-histidine. On the following day the main cultures were inoculated to OD 0.2. For selective conditions MD medium without L-histidine and for non-selective conditions YPD medium were used, respectively. Samples for OD measurement and the plasmid stability assay were taken every 2 h until 16 h of cultivation and again after 26, 36, and 60 h. OD values were used to calculate the specific growth rate at each time point. Before plating onto YPD plates, the samples for the plasmid stability assay were diluted based on the OD so that about 100–1000 colonies were to be expected per plate (Invitrogen, 2010). After 2 days of incubation, 52 colonies were picked per sampling point and individually washed twice with 200 µL of 9 g/L NaCl. Washed cells were resuspended in 20 µL of 9 g/L NaCl and pipetted onto MD plates without L-histidine. Following 2–3 days of incubation the colonies were counted in order to determine the plasmid stability. The plasmid stability assay was carried out in biological duplicates with technical triplicates each.

Determination of GFP Expression Level and Gene Copy Number

The procedures employed for assaying the GFP expression level and GCN were described previously in more detail (Schwarzhans et al., 2016a). In brief, a reference strain containing a single copy of the GFP expression cassette in the *AOX1* locus was used for normalization of the GFP/OD expression level. Strains were grown in 96 deep-well plates as described above and values 60 h after the start of the methanol induction are presented in this study. All strains were cultivated in biological triplicates with technical triplicates each. The GCN was determined based on the protocol by Abad et al. (2010) via the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001), with *ARG4* as the calibrator gene. Biological triplicates with technical duplicates each were used in these procedures.

Super-Resolution Fluorescence Microscopy

Mitochondria were stained with MitoTracker® Red CM-H_2XRos (Thermo Scientific, Waltham, MA, USA). The staining procedure was performed as described by Farre et al. (2007), using *P. pastoris* cells that had been induced with 0.5% (*v/v*) methanol for 3 days. Per experiment 3 µL of cells were mounted between a microscope slide and a high precision #1.5 coverglass (Marienfeld-Superior, Germany). Images were acquired on a DeltaVision OMX V4 system from GE Healthcare (United Kingdom) with a 60x 1.42 NA oil immersion PlanApoN objective (Olympus, Japan) and sCMOS camera. This setup applies the principal of three dimensional super-resolved structured illumination microscopy (3D SR-SIM), gaining a twofold resolution increase compared to conventional fluorescence microscopy. GFPuv was excited at 488 nm and the emission recorded at 504–552 nm. For MitoTracker® Red CM-H_2-XRos an excitation wavelength of 568 nm and emission wavelength band of 590–627 nm were employed. Multiple z planes encompassing the entire cell from top to bottom were recorded at a distance of 125 nm. For reference, differential interference contrast (DIC) images were recorded. Super-resolved fluorescent images were reconstructed with the corresponding recorded optical transfer function (OTF) in the softWoRx 6.5.2 software (GE Healthcare, United Kingdom) at a Wiener filter setting of 0.006. The GFP background was adjusted by subtracting the value of the WT, non-expressing strain. Maximum intensities were individually adjusted for optimal representation. The raw data, OTFs and calibration settings can be provided upon request.

AUTHOR CONTRIBUTIONS

J-PS, JK, and KF designed, analyzed and interpreted wet lab experiments. J-PS and TL performed wet lab experiments. AW performed genome sequencing work. DW analyzed and interpreted sequencing data. WH carried out fluorescence microscopy experiments. J-PS, DW, and WH wrote the manuscript. TH, JK, and KF revised the manuscript. J-PS, JK, and KF conceived the study. TH, JK, and KF supervised the research. All authors read an approved the final manuscript.

ACKNOWLEDGMENTS

Grants from the Federal State of North Rhine-Westphalia for the CLIB-Graduate Cluster Industrial Biotechnology are gratefully acknowledged. The bioinformatics support of the BMBF-funded project "Bielefeld-Gießen Center for Microbial Bioinformatics (BiGi) (Grant number 031A533)" within the German Network for Bioinformatics Infrastructure (de.NBI) is gratefully acknowledged. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 685778. We acknowledge support for the Article Processing Charge by the Deutsche Forschungsgemeinschaft and the Open Access Publication Fund of Bielefeld University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2017.00780/full#supplementary-material

REFERENCES

Abad, S., Kitz, K., Hörmann, A., Schreiner, U., Hartner, F. S., and Glieder, A. (2010). Real-time PCR-based determination of gene copy numbers in *Pichia pastoris*. *Biotechnol. J.* 5, 413–420. doi: 10.1002/biot.200900233

Ahmad, M., Hirz, M., Pichler, H., and Schwab, H. (2014). Protein expression in *Pichia pastoris* : recent achievements and perspectives for heterologous protein production. *Appl. Microbiol. Biotechnol.* 98, 5301–5317. doi: 10.1007/s00253-014-5732-5

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI- BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Aw, R., and Polizzi, K. M. (2013). Can too many copies spoil the broth? *Microb. Cell Fact.* 12:128. doi: 10.1186/1475-2859-12-128

Aw, R., and Polizzi, K. M. (2016). Liquid PTVA: a faster and cheaper alternative for generating multi-copy clones in *Pichia pastoris*. *Microb. Cell Fact.* 15, 29. doi: 10.1186/s12934-016-0432-8

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME suite: tools for motif discovery and searching. *Nucleic Acids Res.* 37, 202–208. doi: 10.1093/nar/gkp335

Bell, S. P., and Stillman, B. (1992). ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. *Nature* 357, 128–134. doi: 10.1038/357128a0

Bernatzky, R., Mau, S. L., and Clarke, A. E. (1989). A nuclear sequence associated with self-incompatibility in *Nicotiana alata* has homology with mitochondrial DNA. *Theor. Appl. Genet.* 77, 320–324. doi: 10.1007/BF00305822

Bill, R. M. (2014). Playing catch-up with *Escherichia coli*: using yeast to increase success rates in recombinant protein production experiments. *Front. Microbiol.* 5:85. doi: 10.3389/fmicb.2014.00085

Blanchard, J. L., and Schmidt, G. W. (1996). Mitochondrial DNA migration events in yeast and humans: integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution patterns. *Mol. Biol. Evol.* 13, 537–548. doi: 10.1093/oxfordjournals.molbev.a025614

Cámara, E., Albiol, J., and Ferrer, P. (2016). Droplet digital PCR-aided screening and characterization of *Pichia pastoris* multiple gene copy strains. *Biotechnol. Bioeng.* 113, 1542–1551. doi: 10.1002/bit.25916

Camattari, A., Goh, A., Yip, L. Y., Tan, A. H. M., Ng, S. W., Tran, A., et al. (2016). Characterization of a panARS-based episomal vector in the methylotrophic yeast *Pichia pastoris* for recombinant protein production and synthetic biology applications. *Microb. Cell Fact.* 15, 139. doi: 10.1186/s12934-016-0540-5

Chatre, L., and Ricchetti, M. (2011). Nuclear mitochondrial DNA activates replication in *Saccharomyces cerevisiae*. *PLoS ONE* 6:e17235. doi: 10.1371/journal.pone.0017235

Chen, S., Reger, R., Miller, C., and Hyman, L. E. (1996). Transcriptional terminators of RNA polymerase II are associated with yeast replication origins. *Nucleic Acids Res.* 24, 2885–2893. doi: 10.1093/nar/24.15.2885

Christianson, T. W., Sikorski, R. S., Dante, M., Shero, J. H., and Hieter, P. (1992). Multifunctional yeast high-copy-number shuttle vectors. *Gene* 110, 119–122. doi: 10.1016/0378-1119(92)90454-W

Clare, J., Rayment, F., Sreekrishna, K., and Romanos, M. A. (1991). High-level expression of tetanus toxin fragment C in *Pichia pastoris* strains containing multiple tandem integrations of the gene. *Nat. Biotechnol.* 9, 455–460. doi: 10.1038/nbt0591-455

Cregg, J., Barringer, K., Hessler, A., and Madden, K. (1985). *Pichia pastoris* as a host system for transformations. *Mol. Cell. Biol.* 5, 3376–3385. doi: 10.1128/MCB.5.12.3376

Daley, J. M., Palmbos, P. L., Wu, D., and Wilson, T. E. (2005). Nonhomologous end joining in yeast. *Annu. Rev. Genet.* 39, 431–451. doi: 10.1146/annurev.genet.39.073003.113340

Delouya, D., and Nobrega, F. G. (1991). Mapping of the ARS-Like activity and transcription initiation sites in the non-canonical yeast mitochondrial ori 6 region. *Yeast* 7, 51–60. doi: 10.1002/yea.320070106

Dujon, B. (2012). Yeast evolutionary genomics. *Nat. Rev. Genet.* 11, 512–524. doi: 10.1002/9783527659180.ch16

Egner, A., Jakobs, S., and Hell, S. W. (2002). Fast 100-nm resolution three-dimensional microscope reveals structural plasticity of mitochondria in live yeast. *Proc. Natl. Acad. Sci. U.S.A.* 99, 3370–3375.

Ellis, S. B., Brust, P. F., Koutz, P. J., Waters, A. F., Harpold, M. M., and Gingeras, T. R. (1985). Isolation of alcohol oxidase and two other methanol regulatable genes from the yeast *Pichia pastoris*. *Mol. Cell. Biol.* 5, 1111–1121. doi: 10.1128/MCB.5.5.1111.Updated

Farre, J. C., Shirahama-Noda, K., Zhang, L., Booher, K., and Subramani, S. (2007). "Localization of proteins and organelles using fluorescence microscopy," in *Methods in Molecular Biology: Pichia Protocols*, 2nd Edn, Vol. 389, ed. J. M. Cregg (Totowa, NJ: Humana Press), 239–250.

Futcher, A. B. (1988). The 2 micron circle plasmid of *Saccharomyces cerevisiae*. *Yeast* 4, 27–40. doi: 10.1002/yea.320040104

Gibson, D. G., Young, L., Chuang, R., Venter, J. C., Hutchison, C. A. III., and Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* 6, 343–345. doi: 10.1038/NMETH.1318

Gordon, D., Abajian, C., and Green, P. (1998). Consed : a graphical tool for sequence finishing. *Genome Res.* 8, 195–202. doi: 10.1101/gr.8.3.195

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27, 1017–1018. doi: 10.1093/bioinformatics/btr064

Guirouilh-Barbat, J., Huck, S., Bertrand, P., Pirzio, L., Desmaze, C., Sabatier, L., et al. (2004). Impact of the KU80 pathway on NHEJ-induced genome rearrangements in mammalian cells. *Mol. Cell* 14, 611–623. doi: 10.1016/j.molcel.2004.05.008

Gunge, N. (1983). Yeast DNA plasmids. *Annu. Rev. Microbiol.* 37, 253–276. doi: 10.1146/annurev.mi.37.100183.001345

Hartner, F. S., Ruth, C., Langenegger, D., Johnson, S. N., Hyka, P., Lin-Cereghino, G. P., et al. (2008). Promoter library designed for fine-tuned gene expression in *Pichia pastoris*. *Nucleic Acids Res.* 36, e76. doi: 10.1093/nar/gkn369

Hazkani-Covo, E., Zeller, R. M., and Martin, W. (2010). Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet.* 6:e1000834. doi: 10.1371/journal.pgen.1000834

Heiss, S., Puxbaum, V., Gruber, C., Altmann, F., Mattanovich, D., and Gasser, B. (2015). Multistep processing of the secretion leader of the extracellular protein Epx1 in *Pichia pastoris* and implications for protein localization. *Microbiology* 161, 1356–1368. doi: 10.1099/mic.0.000105

Herrmann, J. M., Woellhaf, M. W., and Bonnefoy, N. (2013). Control of protein synthesis in yeast mitochondria: the concept of translational activators. *Biochim. Biophys. Acta* 1833, 286–294. doi: 10.1016/j.bbamcr.2012.03.007

Hohenblum, H., Gasser, B., Maurer, M., Borth, N., and Mattanovich, D. (2004). Effects of gene dosage, promoters, and substrates on unfolded protein stress of recombinant *Pichia pastoris*. *Biotechnol. Bioeng.* 85, 367–375. doi: 10.1002/bit.10904

Huang, B., Bates, M., and Zhuang, X. (2009). Super-resolution fluorescence microscopy. *Annu. Rev. Biochem.* 78, 993–1016. doi: 10.1146/annurev.biochem.77.061906.092014

Hyman, B. C., Cramer, J. H., and Rownd, R. H. (1983). The mitochondrial genome of *Saccharomyces cerevisiae* contains numerous, densely spaced autonomously replicating sequences. *Gene* 26, 223–230. doi: 10.1016/0378-1119(83)90192-0

Invitrogen (2008). *pYES2 Manual*. Carlsbad CA: Invitrogen.

Invitrogen (2010). *EasySelect Pichia Expression Kit*. Carlsbad CA: Invitrogen.

Irani, Z. A., Kerkhoven, E., Shojaosadati, S. A., and Nielsen, J. (2016). Genome-scale metabolic model of *Pichia pastoris* with native and humanized glycosylation of recombinant proteins. *Biotechnol. Bioeng.* 113, 961–969. doi: 10.1002/bit.25863

Johnson, M. A., Waterham, H. R., Ksheminska, G. P., Fayura, L. R., Cereghino, J. L., Stasyk, O. V., et al. (1999). Positive selection of novel peroxisome biogenesis-defective mutants of the yeast *Pichia pastoris*. *Genetics* 151, 1379–1391.

Klinner, U., and Schäfer, B. (2004). Genetic aspects of targeted insertion mutagenesis in yeasts. *FEMS Microbiol. Rev.* 28, 201–223. doi: 10.1016/j.femsre.2003.10.002

Küberl, A., Schneider, J., Thallinger, G. G., Anderl, I., Wibberg, D., Hajek, T., et al. (2011). High-quality genome sequence of *Pichia pastoris* CBS7435. *J. Biotechnol.* 154, 312–320. doi: 10.1016/j.jbiotec.2011.04.014

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

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

Lee, C. C., Williams, T. G., Wong, D. W. S., and Robertson, G. H. (2005). An episomal expression vector for screening mutant gene libraries in *Pichia pastoris*. *Plasmid* 54, 80–85. doi: 10.1016/j.plasmid.2004.12.001

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

Liachko, I., Youngblood, R. A., Tsui, K., Bubb, K. L., Queitsch, C., Raghuraman, M. K., et al. (2014). GC-rich DNA elements enable replication origin activity in the methylotrophic yeast *Pichia pastoris*. *PLoS Genet.* 10:e1004169. doi: 10.1371/journal.pgen.1004169

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Looser, V., Bruhlmann, B., Bumbak, F., Stenger, C., Costa, M., Camattari, A., et al. (2015). Cultivation strategies to enhance productivity of *Pichia pastoris*: a review. *Biotechnol. Adv.* 33, 1177–1193. doi: 10.1016/j.biotechadv.2015.05.008

Louis, E. J., and Haber, J. E. (1991). Evolutionarily recent transfer of a group I mitochondrial intron to telomere regions in *Saccharomyces cerevisiae*. *Curr. Genet.* 20, 411–415. doi: 10.1007/BF00317070

Love, K. R., Panagiotou, V., Jiang, B., Stadheim, T. A., and Love, J. C. (2010). Integrated single-cell analysis shows *Pichia pastoris* secretes protein stochastically. *Biotechnol. Bioeng.* 106, 319–325. doi: 10.1002/bit.22688

Love, K. R., Shah, K. A., Whittaker, C. A., Wu, J., Bartlett, M. C., Ma, D., et al. (2016). Comparative genomics and transcriptomics of *Pichia pastoris*. *BMC Genomics* 17:550. doi: 10.1186/s12864-016-2876-y

Madsen, C. K., Vismans, G., and Brinch-Pedersen, H. (2016). The PARS sequence increase the efficiency of stable *Pichia pastoris* transformation. *J. Microbiol. Methods* 129, 1–7. doi: 10.1016/j.mimet.2016.07.015

Marx, H., Mecklenbräuker, A., Gasser, B., Sauer, M., and 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. doi: 10.1111/j.1567-1364.2009.00561.x

Mattanovich, D., Gasser, B., Hohenblum, H., and Sauer, M. (2004). Stress in recombinant protein producing yeasts. *J. Biotechnol.* 113, 121–135. doi: 10.1016/j.jbiotec.2004.04.035

Meyer, V., Arentshorst, M., El-Ghezal, A., Drews, A.-C., Kooistra, R., van den Hondel, C. A. M. J. J., et al. (2007). Highly efficient gene targeting in the *Aspergillus niger kusA* mutant. *J. Biotechnol.* 128, 770–775. doi: 10.1016/j.jbiotec.2006.12.021

Näätsaari, L., Mistlberger, B., Ruth, C., Hajek, T., Hartner, F. S., and 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

Newlon, C. S., and Theis, J. F. (1993). The structure and function of yeast ARS elements. *Curr. Opin. Genet. Dev.* 3, 752–758. doi: 10.1146/annurev.ge.19.120185.000333

Noutsos, C., Kleine, T., Armbruster, U., DalCorso, G., and Leister, D. (2007). Nuclear insertions of organellar DNA can create novel patches of functional exon sequences. *Trends Genet.* 23, 597–601. doi: 10.1016/j.tig.2007.08.016

Peng, C., Luo, H., Zhang, X., and Gao, F. (2015). Recent advances in the genome-wide study of DNA replication origins in yeast. *Front. Microbiol.* 6:117. doi: 10.3389/fmicb.2015.00117

Prielhofer, R., Maurer, M., Klein, J., Wenger, J., Kiziak, C., Gasser, B., et al. (2013). Induction without methanol: novel regulated promoters enable high-level expression in *Pichia pastoris*. *Microb. Cell Fact.* 12:5. doi: 10.1186/1475-2859-12-5

Pronk, J. T. (2002). Auxotrophic yeast strains in fundamental and applied research. *Appl. Environ. Microbiol.* 68, 2095–2100. doi: 10.1128/AEM.68.5.2095

Qin, X., Qian, J., Yao, G., Zhuang, Y., Zhang, S., and Chu, J. (2011). GAP promoter library for fine-tuning of gene expression in *Pichia pastoris*. *Appl. Environ. Microbiol.* 77, 3600–3608. doi: 10.1128/AEM.02843-10

Ricchetti, M., Fairhead, C., and Dujon, B. (1999). Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature* 402, 96–100. doi: 10.1038/47076

Ricchetti, M., Tekaia, F., and Dujon, B. (2004). Continued colonization of the human genome by mitochondrial DNA. *PLoS Biol.* 2:e273. doi: 10.1371/journal.pbio.0020273

Rueda, F., Gasser, B., Sánchez-Chardi, A., Roldán, M., Villegas, S., Puxbaum, V., et al. (2016). Functional inclusion bodies produced in the yeast *Pichia pastoris*. *Microb. Cell Fact.* 15, 166. doi: 10.1186/s12934-016-0565-9

Rußmayer, H., Buchetics, M., Gruber, C., Valli, M., Grillitsch, K., Modarres, G., et al. (2015). Systems-level organization of yeast methylotrophic lifestyle. *BMC Biol.* 13:80. doi: 10.1186/s12915-015-0186-5

Sacerdot, C., Casaregola, S., Lafontaine, I., Tekaia, F., Dujon, B., and Ozier-Kalogeropoulos, O. (2008). Promiscuous DNA in the nuclear genomes of hemiascomycetous yeasts. *FEMS Yeast Res.* 8, 846–857. doi: 10.1111/j.1567-1364.2008.00409.x

Schiestl, R. H., Dominska, M., and Petes, T. D. (1993). Transformation of *Saccharomyces cerevisiae* with nonhomologous DNA: illegitimate integration of transforming DNA into yeast chromosomes and in vivo ligation of transforming DNA to mitochondrial DNA sequences. *Mol. Cell. Biol.* 13, 2697–2705. doi: 10.1128/MCB.13.5.2697

Schwarzhans, J.-P., Wibberg, D., Winkler, A., Luttermann, T., Kalinowski, J., and Friehs, K. (2016a). Integration event induced changes in recombinant protein productivity in *Pichia pastoris* discovered by whole genome sequencing and derived vector optimization. *Microb. Cell Fact.* 15, 84. doi: 10.1186/s12934-016-0486-7

Schwarzhans, J.-P., Wibberg, D., Winkler, A., Luttermann, T., Kalinowski, J., and Friehs, K. (2016b). Non-canonical integration events in *Pichia pastoris* encountered during standard transformation analysed with genome sequencing. *Sci. Rep.* 6:38952. doi: 10.1038/srep38952

Sturmberger, L., Chappell, T., Geier, M., Krainer, F., Day, K. J., Vide, U., et al. (2016). Refined *Pichia pastoris* reference genome sequence. *J. Biotechnol.* 235, 121–131. doi: 10.1016/j.jbiotec.2016.04.023

Theis, J. F., and Newlon, C. S. (1997). The ARS309 chromosomal replicator of *Saccharomyces cerevisiae* depends on an exceptional ARS consensus sequence. *Proc. Natl. Acad. Sci. U.S.A.* 94, 10786–10791. doi: 10.1073/pnas.94.20.10786

Thompson, J. R., Register, E., Curotto, J., Kurtz, M., and Kelly, R. (1998). An improved protocol for the preparation of yeast cells for transformation by electroporation. *Yeast* 14, 565–571.

Turner, C., Killoran, C., Thomas, N. S., Rosenberg, M., Chuzhanova, N. A., Johnston, J., et al. (2003). Human genetic disease caused by *de novo* mitochondrial-nuclear DNA transfer. *Hum. Genet.* 112, 303–309. doi: 10.1007/s00439-002-0892-2

Valli, M., Tatto, N. E., Peymann, A., Gruber, C., Landes, N., Ekker, H., et al. (2016). Curation of the genome annotation of *Pichia pastoris* (*Komagataella phaffii*) CBS7435 from gene level to protein function. *FEMS Yeast Res.* 16:fow051. doi: 10.1093/femsyr/fow051

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

Vogl, T., Sturmberger, L., Kickenweiz, T., Wasmayer, R., Schmid, C., Hatzl, A.-M., et al. (2016). A toolbox of diverse promoters related to methanol utilization–functionally verified parts for heterologous pathway expression in *Pichia pastoris*. *ACS Synth. Biol.* 5, 172–186. doi: 10.1021/acssynbio.5b00199

Weis, R., Luiten, R., Skranc, W., Schwab, H., Wubbolts, M., and Glieder, A. (2004). Reliable high-throughput screening with *Pichia pastoris* by limiting yeast cell

death phenomena. *FEMS Yeast Res.* 5, 179–189. doi: 10.1016/j.femsyr.2004.06.016

Wibberg, D., Blom, J., Jaenicke, S., Kollin, F., Rupp, O., Scharf, B., et al. (2011). Complete genome sequencing of *Agrobacterium* sp. H13-3, the former *Rhizobium lupini* H13-3, reveals a tripartite genome consisting of a circular and a linear chromosome and an accessory plasmid but lacking a tumor-inducing Ti-plasmid. *J. Biotechnol.* 155, 50–62. doi: 10.1016/j.jbiotec.2011.01.010

Wu, S., and Letchworth, G. J. (2004). High efficiency transformation by electroporation of *Pichia pastoris* pretreated with lithium acetate and dithiothreitol. *Biotechniques* 36, 152–154. doi: 10.2144/3601A0152

Zhu, T., Guo, M., Sun, C., Qian, J., Zhuang, Y., Chu, J., et al. (2009). A systematical investigation on the genetic stability of multi-copy *Pichia pastoris* strains. *Biotechnol. Lett.* 31, 679–684. doi: 10.1007/s10529-009-9917-4

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer FG and handling Editor declared their shared affiliation, and the handling Editor states that the process nevertheless met the standards of a fair and objective review.

2.3.3 Supplemental data

frontiers

Supplementary Material

A Mitochondrial Autonomously Replicating Sequence from *Pichia pastoris* for Uniform High Level Recombinant Protein Production

Jan-Philipp Schwarzhans, Daniel Wibberg, Anika Winkler, Tobias Luttermann, Wolfgang Hübner, Thomas Huser, Jörn Kalinowski, Karl Friehs*

*** Correspondence:** Karl Friehs: karl.friehs@uni-bielefeld.de

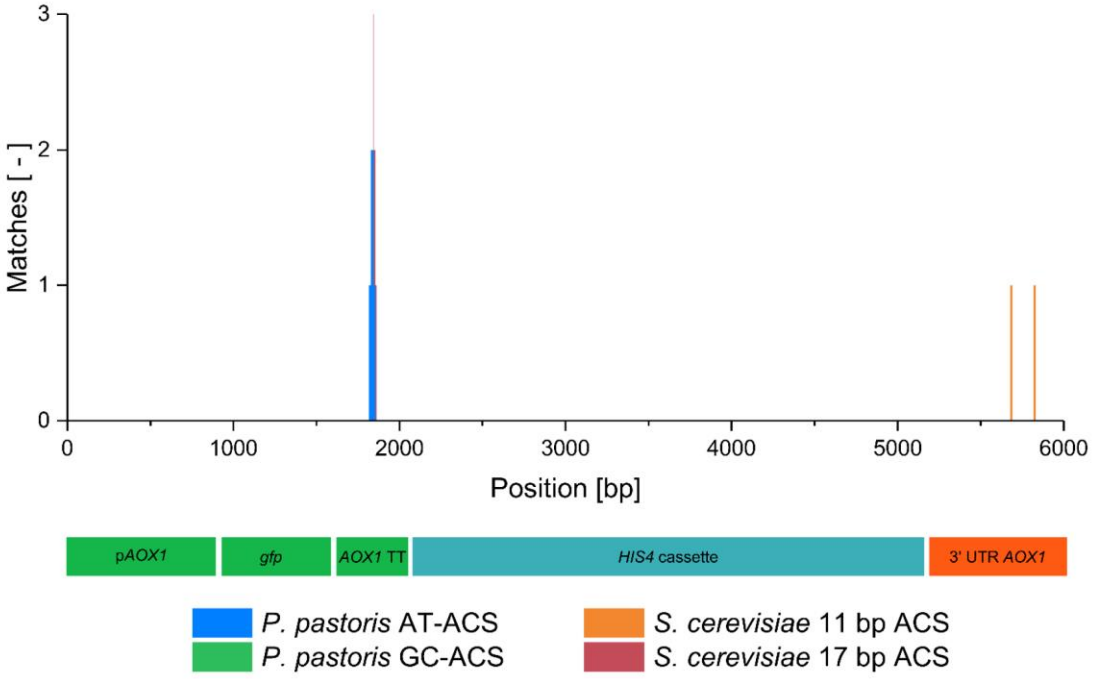

Figure S1: Cumulative matches of the FIMO scan for *P. pastoris* AT- and GC-ACS, as well as *S. cerevisiae* 11 and 17 bp ACS motifs mapped to the expression cassette segment of pMito, excluding the mtDNA fragment. The regions of the different features on the expression cassette are shown below the x-axis.

Supplementary Material

Table S1: FIMO matches of the AT-ACS and GT-ACS from *P. pastoris*, as well as the 11 and 17 bp ACS from *S. cerevisiae* to the 1442 bp mtDNA fragment in pMito.

Motif	Strand	Start	End	p-value	q-value	Matched Sequence
***P. pastoris* AT-ACS**	+	1260	1286	1.51e-06	0.000891	AAATATAAACTTAAAAATATAGTTTCC
	+	702	728	3.38e-06	0.000891	TTATGTATTTTTATTTCTATACTTTTT
	+	32	58	3.61e-06	0.000891	TGATACAAATAATACTTTATATCATTT
	+	387	413	4.52e-06	0.000891	AATTCTAAATTACATTCTATACGATTT
	-	1257	1283	7.67e-06	0.00121	AACTATATTTTTAAGTTTATATTTACT
	+	657	683	1.1e-05	0.00144	ATATGTAATTTATCTTGTATACGAGTT
	+	1043	1069	2.27e-05	0.00256	TAACATAACAAGTTTTATATTGTTTTG
	+	1057	1083	2.75e-05	0.00271	TTATATTGTTTTGACTATTAATTTTAT
	+	149	175	5.9e-05	0.00435	TTCTATAAATTTTAATATTTTTGATAT
	-	158	184	6.42e-05	0.00435	TTTTATAGAATATCAAAAATATTAAAA
	+	1274	1300	6.45e-05	0.00435	AAATATAGTTTCCAGCATATAGTCAAA
	-	130	156	6.63e-05	0.00435	TTATAGAAATACGGAAATATGCCAATT
	+	723	749	7.33e-05	0.00436	CTTTTTAATTTATTTTTTCCATTTATT
	+	714	740	7.74e-05	0.00436	ATTTCTATACTTTTTAATTTATTTTTT
	+	837	863	9.6e-05	0.00505	TTTAATAATTCTATTTTTGTATTTCTC
***P. pastoris* GC-ACS**	+	928	947	1.96e-06	0.00558	CGGGGATCGAACCTAGTCCG
***S. cerevisiae* 11 bp ACS**	+	159	169	1.01e-05	0.00461	TTTAATATTTT
	-	1259	1269	1.47e-05	0.00461	GTTTATATTTA
	+	998	1008	1.61e-05	0.00461	TTTTAAGTTTT

	+	730	740	2.66e-05	0.00461	ATTTATTTTTT
	+	743	753	2.66e-05	0.00461	ATTTATTTTTT
	+	376	386	2.82e-05	0.00461	TATTATATTTT
	-	1265	1275	4.01e-05	0.00522	TTTTAAGTTTA
	+	914	924	4.25e-05	0.00522	ATCTATATTTT
	+	350	360	7.79e-05	0.00768	ATTTAAATTTA
	+	812	822	8.39e-05	0.00768	ATCTATATTTA
	+	798	808	8.6e-05	0.00768	TTTAAAATTTT
	+	347	363	5.2e-06	0.00433	ATTATTTAAATTTAATT
	+	734	750	1.31e-05	0.00543	ATTTTTTCCATTTATTT
***S. cerevisiae* 17 bp ACS**	-	1262	1278	6.04e-05	0.0106	TATTTTTAAGTTTATAT
	-	1256	1272	7.21e-05	0.0106	TAAGTTTATATTTACTA
	+	1052	1068	7.34e-05	0.0106	AAGTTTTATATTGTTTT
	+	701	717	7.66e-05	0.0106	ATTATGTATTTTTATTT

3

Supplementary Material

Table S2: FIMO matches of the AT-ACS and GT-ACS from *P. pastoris*, as well as the 11 and 17 bp ACS from *S. cerevisiae* to the 5884 bp expression cassette region of pMito, excluding the mtDNA fragment.

Motif	Strand	Start	End	p-value	q-value	Matched Sequence
***P. pastoris* AT-ACS**	+	1818	1844	9.25e-05	0.487	ATTTGTAACCTATATAGTATAGGATTT
	-	1829	1855	9.33e-05	0.487	AATGACAAAAAAAATCCTATACTATAT
***S. cerevisiae* 11 bp ACS**	-	5820	5830	4.14e-05	0.276	TTTCACGTTTA
	-	5680	5690	5.03e-05	0.276	TTTAACGTTTA
***S. cerevisiae* 17 bp ACS**	+	1844	1860	9.85e-06	0.106	TTTTTTGTCATTTTGTT

4

Table S3: Names, nucleotide sequences and purposes of all primers used in this study.

Name	**Sequence**	**Purpose**
pMito_Circ_mtDNA-FW	GTCCTTCGTACTATTTAAAGAGAAC	Amplification of pMito and pYES2-Mito, linearized inside mtDNA fragment
pMito_Circ_mtDNA-RV	GAATTAAGTTCGACCTATCGGAG	
pMito_Circ_HIS4-FW	GAATTGGATTGTGATGGAGACTGC	Amplification of pMito, linearized inside *HIS4*
pMito_Circ_HIS4-RV	GATACCCAGCAACTTTTGAGTTGC	
pMito_No_GOI-FW	GGAACGGAACGTATCTTAGCATGGT	Amplification of pMito-ΔGOI from pMito
pMito_No_GOI-RV	TGGACCTGTGGATGTCGGATGG	
pMito_No_UTR-FW	TGCGGAGATGCGATATCGAG	Amplification of pMito-ΔUTR from pMito
pMito_No_UTR-RV	AATGCGAGGATGCTGCTGG	
pYES2_Circ-FW	CCTAATATCCGACAAACTGTTTTACAG	Amplification of pYES2, linearized ca. 300 bp upstream of 2 µm region
pYES2_Circ-RV	CTGTATCTCCTCAAAGCGTATTCG	
mtDNA_Insert-FW	TCAATAGCATATCTTTGTTATATAGTGGACTATCTCTTCAATCTTAATTTTG	Amplification of mtDNA fragment from pMito for insertion via Gibson assembly into pYES2. Overlapping bases to pYES2 are underlined.
mtDNA_Insert-RV	ACCGGGGTGGAGCTTCCCAGGATTATCAGGGGTAGTATTATCAAATG	
pYES2_No_2um-FW	GGGAAGCTCCACCCCGGTTG	Linearization of pYES2 without 2 µm region for Gibson assembly of pYES2-Mito
pYES2_No_2um-RV	AACAAAGATATGCTATTGAAGTGC	

5

3. Conclusion

3.1 Exploration of clonal variability in *Pichia pastoris*

From the outset of the project, it was clear that a broad range of integration events would likely be encountered. Albeit, the exact nature of these was uncertain. A large scope of the study was required, to cover as many potential events as possible. Concomitantly, the methods employed for strain generation and characterization needed to be well-established in the *P. pastoris* community, to ensure transferability and scientific significance of the result for other researchers working with this yeast. Despite the early discovery of clonal variation in 1989 [149], relatively little research was conducted to better understand it comprehensively. Along with the publications from this project, other results from literature are shown in table 2. Here, we compare the experimental design and goal of our study with those found in literature to help illustrate its scientific significance. Unique characteristics, shared properties and missing aspects are detailed.

Table 2: Comparison of studies investigating effects of clonal variability on different strain properties in *P. pastoris*. In the case of parameters for which a marked progress in analytical techniques occurred over the last years, the applied method is given in parenthesis. Entries are ordered chronologically, beginning with the most recent publication. Abbreviations: AFLP = Amplified-fragment length polymorphism, IE = Integration event, RPP = Recombinant protein production, TRG = Transcription level of recombinant gene

Main subject	**Analyzed parameters**	**No. of strains**	**Reference**
TRG and RPP	-RPP -GCN (qPCR) -TRG (qRT-PCR) -IE (PCR) -Transcriptome (microarray) -Cell viability (flow cytometry)	9	[150]
TRG and RPP	-RPP -GCN (ddPCR) -TRG (qRT-PCR) -Substrate consumption -Transcriptome (microarray)	6	[151]

Main subject	Analyzed parameters	No. of strains	Reference
IE and RPP/colony morphology	-RPP -GCN (qPCR) -IE (genome sequencing) -Colony morphology -*E. coli* gene activity (qRT-PCR)	845 (31 genome-sequenced)	This study [138, 143]
GCN and RPP	-RPP -GCN (ddPCR) -Substrate consumption	37	[113]
IE	-IE (genome walking)	14	[111]
GCN and RPP	-RPP -GCN (qPCR) -TRG (qRT-PCR)	9	[152]
GCN and RPP	-RPP -GCN (qPCR)	6	[135]
GCN and RPP	-RPP -GCN (qPCR) -UPR activity (qRT-PCR)	30	[153]
GCN and RPP	-RPP -GCN (qPCR)	24	[154]
IE and RPP	-RPP -IE (AFLP)	14	[155]
GCN and RPP	-RPP -GCN (Southern blot) -TRG (Northern blot)	15	[156]
GCN and RPP	-RPP -GCN (Southern blot)	8	[157]
GCN and RPP	-RPP -GCN (Southern blot)	3	[158]
GCN and RPP	-RPP -GCN (Southern blot) -TRG (Northern blot)	6	[159]
GCN and RPP	-RPP -GCN (Southern blot)	12	[114]

A clear focus on investigating the correlation between GCN and production of the recombinant protein is apparent (reviewed in [160]). Some studies also included determination of the mRNA expression level of the target gene, elucidating how well it correlated with measured protein concentrations [151, 152, 156, 159]. On the one hand, it was conclusively shown that for intracellular protein production GCN and productivity are directly correlated over a wide range [114, 154, 159]. This is not true for secreted protein production. Here, factors like the UPR pathway and correct targeting of the protein complicate the effect of the gene dosage on productivity [152, 154, 156]. In particular the study by Marx *et al.* (2009) [154] highlights this challenge. Multiple strains with different GCN values, either secreting human serum albumin (HSA) into the medium or accumulating human superoxide dismutase in the cytosol, were analyzed. For intracellular production a good linear correlation between GCN and expression level was found, in a range of 1 to 30 copies of the target gene. In contrast, secreted production of HSA increased linearly only from 1 to 6 copies and dropped for strains with a higher gene dosage. More recent studies have improved upon this knowledge by applying newly developed analysis techniques like ddPCR and measuring so far neglected parameters. ddPCR and qPCR allow for a more accurate determination of the GCN, compared to previously applied Southern blot protocols that could result in a standard deviation of up to ten copies [161]. Specially for secreted proteins, determining the mRNA level of the target gene or UPR pathway genes can help to identify bottlenecks in productivity due to activation of the UPR pathway [151–153]. Using a subset of six clones with different GCN, selected from 37 previously screened strains, substrate consumption in a chemostat cultivation was analyzed [113]. The analysis revealed an increased glycerol consumption of multicopy strains compared to the single copy and wild type strain, while methanol consumption decreased. Interestingly, no linear correlation was found. Rather, the strain with the highest productivity and two copies of the target gene displayed by far the lowest methanol and highest glycerol consumption rate, respectively. Potentially, the observed behavior was due to the cells adapting to the increased metabolic burden caused by foreign gene expression.

Two previous projects investigated integration events as part of the clonal variability. Viader-Salvadó *et al.* (2006) [155] utilized different lengths in band size after PCR to classify 14 clones and three controls into separate categories of integration events, but did not determine the exact insertion locus. *P. pastoris* gDNA was digested with *XhoI*, ligated with an adapter and primers specific for the adapter and the expression cassette were used for PCR amplification. The resulting bands were analyzed for their size via gel electrophoresis and clones sorted into related groups based on the observed patterns. This approach is also known as amplified-length polymorphism (AFLP). Five major clusters were determined for the analyzed recombinant strains. Interestingly, the cluster with the highest similarity to the control strains (wild type and strain transformed with empty vector) contained the transformants with the highest productivity levels. A correlation between lower genetic perturbance and high protein productivity can be assumed. Sadly, the applied analysis did not allow for further interpretations.

Näätsaari *et al.* (2012) [111] identified the exact integration site for 14 strains, originating from an experiment aiming for creation of auxotrophic strains. The affected strains did not display the desired phenotype after transformation and were therefore suspected to contain off-target integrations. The integration loci were determined via genome walking and nested PCR, applying different adapters for the first and second round of PCR. In 8 of the 14 transformants, off-target integration resulted in gene disruptions. However, the effects of these disruptions were not analyzed. Interestingly, while the majority of off-target integrations appeared to be at random sites of the genome, two strains contained an identical insertion locus. Yet, the authors point out that after DNA transformation a regeneration phase (2 h in YPD) was applied prior to plating and selection. It is difficult to say whether the identical integration site was simply the result of the affected cell doubling during the regeneration phase, thereby producing two genetically identical colonies on the plate. Furthermore, the authors submit that the sample size of 14 strains is not large enough to deduce general theses for random integrations in *P. pastoris*. Due to the focus on creating auxotrophic strains and preventing off-target integrations, no recombinant protein production was assayed. The effect of targeted integration at different loci was investigated in two previous studies, but no significant impact on protein productivity was found [162, 163]. Also, strictly

speaking in those studies no clonal variability was examined, since only correct integrations at the predefined sites were considered for further analysis.

Very recently (22.05.2017) a new study focused on understanding clonal variability in *P. pastoris* was published by Aw *et al.* (2017) [150]. Secreted expression of HSA in nine single-copy strains with the expression cassette in the *AOX1* locus (confirmed via PCR) was analyzed on multiple levels. Product titers were used to separate the clones into groups of high-, mid- and low-producers. Subsequently, transcription levels of HSA and key UPR-associated genes were determined via qRT-PCR. The genome-wide transcriptome was assessed in a microarray approach. Flow cytometry was used for investigating the cell viability during cultivation. No clear correlation between HSA transcript level and product titer was found, with one of the low-producer strains exhibiting the second highest HSA transcriptional activity. However, this particular clone displayed various idiosyncrasies and was therefore considered an outlier. Inclusion of qRT-PCR and microarray data on UPR activity did not produce a significant correlation to the HSA productivity. Interestingly, an upregulation of genes involved in ribosome biosynthesis was observed in all analyzed strains, compared to the wild type. The upregulation was inferred to be caused by a "demand-led" mechanism, meaning the cells actively upregulated ribosome synthesis to cope with the increased burden on protein synthesis from recombinant protein production. Cell viability analysis revealed that all high-producers were characterized by a lower live/dead ratio, while low-producers featured overall larger cells. A presumed connection of decreased viability with increased ROS (reactive oxygen species) accumulation or UPR activity could not be confirmed, based on the transcriptome data.

Although no specific pattern differentiating the different producer strains was found in the transcriptome, the upregulated proteasomal activity appeared to be a potential indicator for identifying high producer strains. Whether high secretion rates caused increased proteolytic activity or vice versa remained unclear, but is planned to be analyzed in future experiments. For the above mentioned outlier strain, a defect in the amino acid recycling pathway could be inferred from the microarray analysis. Potentially, a constrained amino acid pool induced the discrepancy between high amounts of HSA mRNA paired with a low product titer. Since the correct integration at

the *AOX1* locus was solely confirmed via PCR, the exact origin of altered transcriptional activity in the analyzed strains could not be determined. Expanded analysis via genome sequencing might reveal e.g. SNPs in genes with altered expression levels.

Nevertheless, the study provides relevant insights into clonal variability. In particular, the inconsistency between transcriptome and productivity data highlights the complicated nature of clonal variability in *P. pastoris*. This is even more striking, considering the pre-selection of only single-copy strains with cassettes in the right locus. It is possible that the applied time interval for sampling (24 h after start of methanol induction) was too early to observe all relevant effects. Recently published comparative RNA-Seq data demonstrated significant differences between 24 and 48 h sampling points [164]. A combination of both our approach and the one by Aw *et al.* (2017) [150], e.g. inclusion of transcriptome and genome data for strains with interesting properties, might produce even further insights. On a side note, the publication repeatedly suggests that our first study [138] referenced different integration sites as a leading cause for different product titers. However, our study did not describe any off-target insertions. Rather, we emphasized the favourable finding of high producer strains only containing integrations at the *AOX1* locus and speculated that the effect of random integrations on productivity was too low to pass our selection criteria.

In this context, the first two publications of this thesis focused on providing new insights into clonal variability [138, 143]. Aspects of gene dosage dependent recombinant product production and (off-target) integration events, that previous *P. pastoris* publications analyzed separately, were unified in one study. The to date largest library of *P. pastoris* clones transformed with the same expression cassette, surpassing earlier collections 20 to 100 fold, was systematically probed. Simultaneously, genome sequencing of 31 selected strains resulted in the most extensive determination of integration events reported so far. With the exception of the analysis of a mutagenesis derived library [165], genome sequencing was not previously applied for *P. pastoris* strain characterization on a larger scale. Multiple events (e.g. relocation of the *AOX1* locus, correlation between cassette orientation and productivity, integration of *E. coli* DNA) would not have been noticed without sequencing the entire genome. Genome data alone however would have only allowed for a limited insight into clonal variability, less relevant for

biotechnological applications. Only by combining these results with the ones derived from classical strain characterization (Mut-phenotype, colony morphology, productivity, GCN), could events and effects be correlated. Thus, statements relevant for both basic and applied science could be made. This augmented level of analysis and interpretation is a key difference to the previous clonal variance studies listed in table 2.

The experimental setup we chose ensured optimal comparability by exclusively employing standard *P. pastoris* methods, established specifically for this yeast and routinely used by other scientists working in this field. Yet, due to the scope of the investigated clone library, analysis of some interesting parameters was unfeasible. Cultivation in deep-well plates prohibited intricate process strategies (e.g. fed-batch) and monitoring of relevant culture parameters like pO_2, pH or substrate concentration. Most industrial applications of *P. pastoris* employ fed-batch processes, or to a lesser degree the chemostat mode [96].

The target protein was not secreted to the medium, which is a common technique in industrial processes to simplify downstream purification. We settled on intracellular protein expression to exclude an additional dimension complicating result interpretation. Extracellular protein production would have meant that UPR activation, failed protein targeting during secretion and extracellular protein degradation would all have played a part in affecting the measured expression level. Deriving connections between integration event and productivity was expected to be significantly more difficult. Therefore, we chose to pass over the added information from secreted protein production in favor of more reliable results and increased confidence in derived statements.

No measurement of the transcriptional level of the target gene via qRT-PCR was performed. GCN determination via qPCR required gDNA. gDNA is relatively stable, DNases in the samples can easily be removed or inactivated and contaminating RNA is efficiently degraded by the addition of RNase A. On the other hand, mRNA is unstable, small amounts of contaminating gDNA can lead to unreliable results and RNases are ubiquitous as well as difficult to be inactivated. Small scale preparation of mRNA suitable for qRT-PCR is generally unproblematic. In our case, technical complications were expected for the isolation of mRNA from 845 clones with biological triplicates,

necessary to generate meaningful results. However, it was shown that for intracellular protein expression both the transcript and protein levels display a good linear correlation in *P. pastoris* [159]. This is not true for secreted proteins, in which case qRT-PCR experiments would have been more advisable [152, 156].

Many publications centered around the effects of multicopy strains on productivity, employed specialized transformation techniques to ensure specific GCN values [113, 154, 159]. While these methods allow for directed creation of desired strains and are beneficial for biotechnological purposes, they make the results less comparable with multicopy strains arising from random integration events. Our and previous studies clearly implicate *in vivo* multimerization of transformed expression cassettes prior to integration for generating multicopy clones during standard transformation [114, 138]. By using one of the most common transformation strategies, replacement of *AOX1* with the expression cassette, we aimed to cover more and representative integration events. Furthermore, the electroporation protocol we used does not require a regeneration phase between transformation and plating [133]. Thereby, the above mentioned uncertainty for the frequency of integration events is prevented [111].

It should be noted that the clonal variability studies discussed above typically performed a screening step after transformation and prior to the main experiments. In consequence, a large part of the clonal variability was removed before the main analysis began. For example, only strains with a certain GCN, protein expression level or phenotype were selected for further analysis [111, 150, 152, 159]. We applied a scoring matrix to rank all 845 clones based on their properties (Mut-phenotype, GCN and expression level). Depending on the rank, 31 strains with particularly interesting properties were selected for genome sequencing. In this respect, we performed a selection process for further analysis. This decision was made due to technical limitations. Genome sequencing all 845 strains was expected to markedly exceed the scope of this project, regarding both time and cost. Nevertheless, we tried to ensure that selected strains were representative for larger groups of clones with relevant properties. Furthermore, insights from genome sequencing were used to investigate related clones via PCR assays. Thereby, the ranking and subsequent selection process did not necessarily exclude the remaining strains from further analysis. Our approach of

characterizing all clones obtained after transformation was a more "holistic" experimental design, compared to previous studies. It was the first attempt to chart the entirety of clonal variance that a scientist applying standard methods might encounter in a *P. pastoris* experiment.

The following three sections summarize the results and compare them to insights from literature. For enhanced readability, they are divided into implications for recombinant protein production or genetic engineering, as well as development of new genetic tools.

3.2 Implications for recombinant protein production

Four key insights into the relationship between integration event and recombinant productivity were discovered. Three of these had a negative impact on productivity, with only one correlating with increased product titers. Details for all events can be found in table 3.

Table 3: Discovered connections between integration event and protein productivity [138]. The suspected relationship and the recommendation to prevent or target the respective integration event are given. Abbreviations: EC = Expression cassette, IE = Integration event, RPP = Recombinant protein production, TT = Transcription terminator

Event	**RPP**	**Inferred relation**	**Recommendation**
Head-to-head and tail-to-tail orientation of EC	Reduced	Transcriptional issues between neighboring EC	Avoid these orientations
Head-to-tail orientation of EC	Increased	No transcriptional issues	Aim for this orientation
Mixed cultures	Reduced	Wild type cells consume methanol	Dilution plating; antibiotic selection marker
Secondary IE	Reduced	Additional homologous sequence in EC	Replace *AOX1 TT* on EC

For a visualization of the different cassette orientation forms and their distribution among genome sequenced clones please refer to figure 4 and 7 on pages 147 and 149, respectively. It was found that head-to-head and tail-to-tail orientations are predominantly present in strains with an expression level markedly lower than the one expected based on their GCN. In contrast, none of the genome sequenced multicopy strains with an expression level markedly exceeding the predicted one contained these orientation forms. This clearly indicated a negative impact on productivity by these

orientations. Two different mechanisms are suspected to facilitate this effect. For cassettes in a tail-to-tail orientation, a "head-on collision" of RNA polymerases II (RNAPII) during transcription is possible. Such collisions, and the subsequent transcriptional arrest, have been shown before for converging transcription events. Atomic force microscopy could visualize the event *in vitro* [166]. For S. *cerevisiae, in vivo* collision events have been proven [167]. Surprisingly, the collided RNAPII did not quickly dissociate from the DNA strands, but rather remained attached for a while and required ubiquitylation-associated proteolysis for removal [167]. Regarding the discovered clones this could mean that productivity was not as expected, since transcription of neighboring cassettes led to lasting transcriptional arrest via collided RNAPII molecules. The high transcriptional strength of *pAOX1* would result in large of amounts of RNAPII being recruited for transcription. Thereby, more RNAPII were supplied to the DNA strand, collided and were not dissociated efficiently. It has to be considered that in this orientation form, the *pAOX1* of neighboring cassettes were distal to one another. Combined with the high termination efficiency of *AOX1 TT* [168], the negative effect was potentially mitigated to an extent. In the case of cassettes in the head-to-head orientation, the *pAOX1* of adjacent cassettes were directly next to one another. With ca. 550 kDa, yeast RNAPII is a large enzyme [169]. Directly adjacent promoters could lead to these molecules colliding or otherwise sterically blocking each other from binding to the DNA or properly starting transcription. Deletion analysis of *pAOX1* has shown that the binding site for the transcription factor *MXR1* (Methanol expression regulator 1) exists very close to its 5' end, which is suspected to facilitate attraction of RNAPII [170]. Again, the high transcriptional activity of *pAOX1* might have compounded this issue by recruiting large amounts of RNAPII molecules. It is difficult to estimate which of these orientations had a greater effect on productivity, because genome sequenced clones always contained a 50:50 mixture of both variants. Targeted assembly of one orientation form in a vector, subsequent transformation and analysis of transcription level of the target genes might answer this question.

In contrast to the above mentioned cassette organization forms, head-to-tail orientation correlated well with improved protein productivity. While it was the predominant orientation form in all analyzed multicopy strains, its share increased markedly in the high producer and overachiever groups. The head-to-tail orientation form circumvents

the above mentioned challenges with converging or diverging promoter orientations and RNAPII molecules. Consequently, strains could fully express each copy of the target gene without negatively affecting neighboring cassettes. Besides the avoidance of negative effects, head-to-tail orientation might have also had a promoting effect on transcriptional activity. In *S. cerevisiae* it has been shown that RNAPII molecules remain attached after inefficient termination and continue transcription in the downstream gene [171]. This results in bicistronic mRNA and potentially increased expression levels. Admittedly, the highly efficient termination of *AOX1 TT* might prevent such events in *P. pastoris* [168]. Interestingly, clones with head-to-tail cassettes did not contain the alternative orientation forms (head-to-head and tail-to-tail). This circumstance suggests the existence of two competing integration mechanisms, either resulting in all cassettes being head-to-tail or a 50:50 mixture of the alternative orientation forms. However, the exact nature of these mechanisms remains unclear and requires further studies to be elucidated.

Taken together, for creation of high producer strains, achieving the head-to-tail orientation and preventing alternative orientations seems advisable. One possibility to achieve this is to construct vectors with multiple expression cassettes in the correct orientation [113, 159]. However, the potential GCN is limited by the vector size. Large vectors are problematic for proliferation in *E. coli* and transformation into *P. pastoris*. Directed *in vivo* assembly of single expression cassettes into an oligomer of defined orientation might prove more efficient. A recently published protocol promises high *in vivo* assembly efficiencies in *P. pastoris* [172]. Due to the predominant NHEJ pathway, irregular integration events might arise during *in vivo* assembly and subsequent integration. Ensuring that only the head-to-tail orientation produces viable transformants might be facilitated by implementing the split-marker system [173].

Some genome sequenced clones with low expression levels were mixed cultures, often containing untransformed cells. In theory, these cells could have taken up methanol during induction but did not produce any target protein. Thereby, protein production per OD dropped. From a biotechnological perspective, they represented an additional burden on the screening process. Untransformed cells were potentially supplied with histidine from prototrophic cells, as has been reported for similar auxotrophic yeast

markers [174]. The problem arose, since dilution plating after transformation was not performed. Application of this standard technique, or the use of an antibiotic selection marker like Zeocin, should solve the issue.

In multiple strains, no productivity and no target gene was detected. Via a combination of genome sequencing and PCR Assay, it was discovered that a secondary integration event was the culprit and that it had occurred in ca. 8 % of all strains. The high frequency of the event meant that it markedly complicated the screening process and is of concern for protein production studies. A particular setup of the expression cassette mediated the secondary integration event. The *AOX1 TT* of the expression cassette was the third homologous sequence to the chromosomal *AOX1* locus, besides *pAOX1* and a fragment of *3'* UTR (untranslated region). While the latter two were designed to mediate exchange of the native *AOX1* with the expression cassette, *AOX1 TT* was simply used because of its efficient termination capabilities. However, its homology to the chromosome led to integration of the selection marker without the target gene (see figure 3 on page 146). Essentially, false positive clones were created. Exchange of *AOX1 TT* with the *CYC1 TT* from *S. cerevisiae* with no homologies to the *AOX1* locus eliminated the secondary integration event. Thereby, insights from genome sequencing were applied for vector optimization. No reports of similar secondary integration events could be found in literature. Unfortunately, expression levels with this new terminator were markedly lower (ca. 40 % lower on average). Insufficient termination capabilities of *CYC1 TT* are the suspected cause [175]. Several native terminators from *P. pastoris,* with strength on-par with *AOX1 TT,* have recently been characterized and should be more suitable replacements [168].

Despite the scope of the library and the applied analytics, an explanation for their peculiar characteristics could not be determined for all strains. Interestingly, all strains with high product levels contained no off-target integrations. All expression cassettes were found at the *AOX1* locus. This is reassuring for scientists interested in high producer strains, suggesting that their genetic integrity is not affected. Nevertheless, it means that the effect of random integration on productivity could not be assayed in our study. The majority of the 845 clones had a normalized expression level and GCN of ca. 1. Potentially, many of these contained an expression cassette integrated at a random

site. But as previous research suggests [162, 163], the effect of the insertion locus on productivity was not significant or at least not significant enough to qualify for the ranking criteria we applied. As mentioned in the previous chapter, we did not measure certain parameters (e.g. the transcription level). It is possible that the behavior of certain clones would have been elucidated by applying these techniques. In this respect, another study required the application of transmission electron microscopy to reveal the presence of two nuclei in a high producer *P. pastoris* clone as likely cause for its favourable features [176]. The clone was isolated from a library previously screened via methods analogous to ours, and additionally characterized with transcriptomic and proteomic methods. Despite this thorough approach, the underlying mutation causing the binuclear phenotype and thereby high productivity has not been conclusively found. Perhaps, genome sequencing could aid in this case.

3.3 Implications for genetic engineering

Off-target integration of the expression cassette is a well-known issue during transformation of *P. pastoris*. Multiple strains with a crenulated colony morphology were detected during plating assays. Various off-target integration events were discovered in them via genome sequencing. They all negatively impacted genetic integrity to different extents. Table 4 lists the recorded events; the following paragraphs provide more detailed information on them and discuss their implications for genetic engineering in *P. pastoris*.

Table 4: Description of the untargeted integration events, discovered during genome sequencing of *P. pastoris* strains with aberrant colony morphology [143]. The suspected cause for the event and recommendations to avoid its occurrence are listed. Abbreviation: EC = Expression cassette

Event	**Inferred relation**	**Recommendation**
Off-target EC integration	NHEJ pathway	NHEJ deficient strain
Relocation of knock-out target	Two homologous sequences in EC	Single homologous sequence in EC
Co-integration of *E. coli* DNA	Contamination of transformed EC	PCR amplification or Gel purification of EC
Post-transformational modification of EC before integration	Problems of DSB repair in yeast	Avoid tandem repeats

The disruption of untargeted genes was recorded in two strains (see figure 2 on page 170). Integration was likely facilitated by the NHEJ pathway, due to the absence of homologous regions between the genes and the inserted DNA. Considering the annotated function of the disrupted genes, or their homologue in *S. cerevisiae*, their disruption appeared to be the most probable cause for the changed colony morphology. Both genes were involved in processes related to cell wall integrity and previous studies in *S. cerevisiae* had demonstrated that their inactivation caused morphological changes. From a genetic engineering standpoint however, it is more relevant to reflect upon ways of preventing similar events to occur. Random integrations complicate the creation of defined strains for gene function studies, metabolic engineering or the creation of platform strains. As the observed disruptions were likely mediated by the NHEJ pathway, it appears to be most prudent to try and circumvent this pathway. A *P. pastoris* strain deficient in NHEJ activity is available and was shown to have significantly increased targeting efficiencies [111]. As a side-effect, homologous sequences of shorter length can be used without affecting the targeting efficiency, simplifying vector construction. Although the strain was already successfully applied in genetic and metabolic engineering studies [118, 119], its applicability for industrial purposes is impaired. Inactivation of the NHEJ pathway also results in lowered specific growth rate and genetic stability, since an essential part of the DSB repair mechanism is missing. Another strategy is the split-marker system, wherein the marker is separated onto two DNA fragments with overlaps and only their simultaneous insertion at the targeted locus facilitates expression of the functional selection marker [173]. No specific strain is required, but so far the system has only been demonstrated for gene deletions in *P. pastoris* and not knock-ins [104, 177]. Other systems are available, but have not been applied outside of their proof-of-concept establishment [128, 129]. They are discussed in more detail in the review (section 4), starting on page 61.

An interesting subclass of NHEJ mediated off-target integrations was found in the form of a relocated *AOX1* locus (figure 3 on page 172). Likely, the native *AOX1* locus was knocked-out during the double crossing over event used for insertion of the expression cassette, and subsequently was reintegrated via NHEJ at a random locus of another chromosome. Since the entire *AOX1* locus including *pAOX1* and *AOX1TT* was relocated, the strain retained full *AOX1* activity and grew normal on methanol. Besides aspects of

preventing random integrations as discussed above, the excision of the *AOX1* locus is the main point of possible intervention. The observed event could have been prevented by only using one homologous sequence for directing integration (ends-in) instead of two (ends-out) [31]. Of course this strategy is only applicable, if no replacement but simple addition is desired. For recombinant protein production studies a trend towards using strains with predefined Mut^S phenotype in conjuncture with ends-in integration is visible [178, 179]. This approach removes the screening step for correct Mut-phenotype, offers higher transformation efficiencies and is more likely to result in high copy strains. On the downside, targeting efficiency is reduced [31, 180]. If gene knock-out is an integral part of the experimental setup, the relocation of the deletion target cannot be fully prevented. Besides removal at the original locus, reintegration at a new one has to be assessed in transformants. Otherwise, results might be misinterpreted since the "deleted" gene is still active on a different part of the genome. Genome sequencing will likely not be suitable for routine applications, but methods like genome walking could prove useful and more cost-efficient [111]. Unfortunately, we could not determine the frequency of the relocation event. It was found solely due to the gene disruption it caused.

Four of the sequenced strains contained *E. coli* DNA from the plasmid host, integrated in fusion with the expression cassette. Fragments of both genomic and F plasmid DNA were found. Their size ranged from 1.5 to 9.3 kb. The DNA was possibly co-extracted during plasmid isolation, transformed into *P. pastoris* where *in vivo* ligation to expression cassettes and subsequent integration resulted in the detected integration events (see figure 6 on page 175). A variety of genes coding for (hypothetical) proteins with membrane association in *E. coli* were present on these fragments. The theory that the expression of these proteins in *P. pastoris* led to the observed change in colony morphology, was corroborated by the detection of the respective transcripts via qRT-PCR in affected clones. From a strain engineering perspective, the presence and transcriptional activity of *E. coli* DNA is highly undesirable. In our case, the strain physiology was negatively affected. Reduced productivity, reduced strain fitness and expression of contaminating proteins or metabolites are other possible impairments. It has to be emphasized, that the production of substances compromising the safety for human use is highly unlikely. *E. coli* strains employed for plasmid propagation are not

pathogenic and expression of lipopolysaccharides, the major endotoxin from *E. coli*, requires several genes for synthesis and assembly [181]. Nevertheless, taking measures to prevent *E. coli* DNA co-integration are advisable. *E. coli* strains with an F plasmid should be avoided in *P. pastoris* studies. Purification of the excised expression cassette via gel purification should remove most contaminating DNA, but we encountered *E. coli* DNA fragments of similar size to the expression cassette, which would by-pass this measure. Rather, PCR amplification of the expression cassette should yield DNA free of contaminations. For excision of the expression cassette from the plasmid, we used a restriction enzyme that created sticky ends (*BglII*). The cohesive ends likely promoted the *in vivo* ligation of the expression cassette to the *E. coli* DNA fragments, which were also (unintentionally) digested with the same enzyme. Switching to blunt-end or rare-cutters would reduce the likelihood of *in vivo* fusion or the digestion of untargeted DNA, respectively. No previous reports about contaminating *E. coli* DNA co-integration (or related events) into the *P. pastoris* genome during standard transformation procedures were available at the time of writing.

Lastly, in one of the co-integration events of *E. coli* DNA, the removal of two 12 bp long palindromic sequences was discovered (figure 4 on page 173). It is possible that yeast-inherent problems with palindromic sequences during DSB repair resulted in their excision in the course of the integration event [182, 183]. The deletions did not occur at the distal ends of the DNA, excluding exonuclease activity as possible explanation. In our case, an undesired gene was affected. Nevertheless, it appears feasible that similar events can occur in experiments aiming for foreign gene expression or metabolic engineering. Potentially, the DNA sequence of the target gene is mutated during insertion, resulting in altered activity or even inactive gene products. Premature stop of transcription due to newly created stop codons is also possible. Regulatory elements of the expression cassette like the promoter and the terminator could also be affected, hampering correct expression. Small deletions in the DNA might not be noticed during routine PCR assays, and the encoded protein would appear of very similar size in a SDS-PAGE test. Only during e.g. an enzyme activity experiment, the unfavorable features would be revealed. In such cases, it might make sense to completely sequence the integrated expression cassette to ascertain its genetic integrity. Removal of

palindromic sequences in the target gene via codon alteration would be a probable solution.

3.4 Newly developed genetic tools

Two tools for genetic manipulation were developed in the context of this thesis. An optimized integrative vector for reduced occurrence of false-positive clones and a novel ARS based episomal vector. The identification and elimination of the secondary integration event mediated by *AOX1 TT* in the expression cassette has been described in chapter 3.2 in the context of discovered integration events. Here, the results and potential of pMito and its mitochondrial ARS are discussed.

Originally discovered in a strain (JPS664) with unexpectedly high expression levels, pMito is a fusion product of the expression cassette and 1.4 kb of mitochondrial DNA (mDNA) [148]. The event occurred during creation of the 845 clone library used in the prior studies. Previous reports regarding ligation of mDNA to transformed heterologous DNA and its involvement in DSB repair in yeast, suggest that the NHEJ pathway facilitated the creation of pMito [184–186]. Its circular form and replication capabilities were experimentally validated. In combination with *in silico* ARS analysis, the presence of a novel ARS on pMito was highly likely. Different variants of pMito were successfully transformed into *P. pastoris*, resulting in strains with favourable productivity characteristics, on-par with JPS664 (figure 6 on page 192). The expression level of a reference strain with a single copy of the expression cassette in the *AOX1* locus was significantly surpassed by all assayed pMito-strains. On average, pMito clones showed a three- to fourfold increase, relative to the reference strain. In addition, these strains displayed a far more homogenous clonal variability than the integrative strains.

The homogenous and high recombinant protein expression level among pMito strains makes the system ideal for screening procedures. Experiments in which a high number of different target proteins need to be expressed could benefit from this technology. By providing strains with highly similar expression levels the clone selection is simplified and results from enzyme activity assays can more easily be related to the enzyme itself, rather than to varying expression levels of different strains. The uniformly high expression level should ensure that sufficient amounts of the desired protein are synthesized for analytical purposes. In consequence, such screening experiments could

be considerably accelerated. Projects focused on optimizing production of a small number of target proteins, could benefit from the lowered clonal variability. Rather than screening several hundreds, if not thousands, of strains with different integration events [96, 122], all transformants would exhibit comparable expression levels.

Nevertheless, recombinant protein production via pMito also has its pitfalls. Although expression levels were highly similar among clones, the GCN was not. No apparent correlation between gene dosage and expression level could be deduced. This phenomenon and the apparent upper limit of expression (fourfold of single-copy integrative strain) was identical to the one described for the episomal vector "panARS" [172]. For both cases it was speculated, that post-translational or epigenetic factors limit the achievable expression level in *P. pastoris*. The recent discovery of epigenetic changes following gene deletion and their effect on the transcription of various untargeted genes, suggests the existence of so far unknown epigenetic regulation patterns in *P. pastoris* [187]. It would be of interest to see whether *P. pastoris* can differentiate episomal from chromosomal expression and direct its resources accordingly. Potentially, identification of this bottleneck could lead to a break-through, enabling increased episomal productivity. As it stands, pMito should not be viewed as a complete replacement for integrative vectors. Despite the decreased clonal variability, figure 5 demonstrates that integrative clones had the potential for higher productivity.

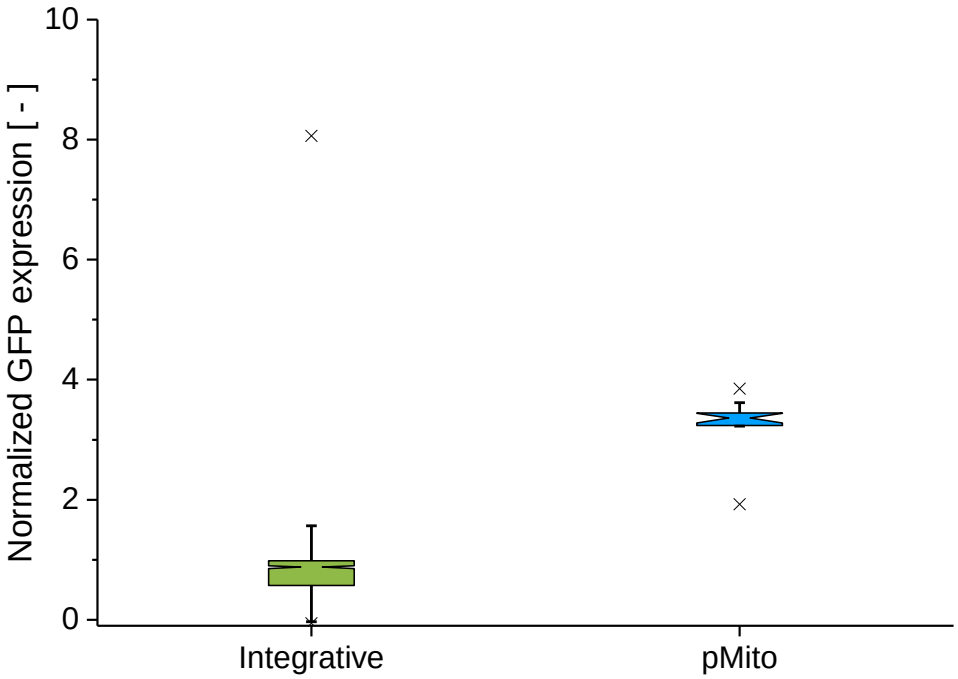

Figure 5: Box plot of the normalized GFP expression level for the integrative and pMito *P. pastoris* strains, characterized in the course of this thesis. The 845 integrative clones stem from the first publication [138] and the 17 pMito clones from the third one [148]. Notches of the box indicate the median, the ends of the whiskers the highest and lowest value within a 1.5 interquartile range of the upper and lower quartile, respectively. Crosses display the absolute maximum and minimum of the respective dataset.

The high genetic stability of integrative clones is a distinct advantage over the available episomal vectors in *P. pastoris*. pMito displayed high plasmid stability (ca. 100%) under selective conditions, but this value quickly dropped to ca. 25% if no selective pressure was present (figure 2(B) on page 189). A declining plasmid stability over time, even under selective conditions, was visible. It was suspected that release of L-histidine into the medium by prototrophic cells caused this phenomenon [174]. Often, the growth advantage of plasmid-free cells over plasmid-containing cells is the key factor causing dropping plasmid stability rates, once the selective pressure decreases [188]. In consequence, the system would be unsuitable for industrial applications, in which stable productivity is paramount. Switching to an antibiotic selection marker akin to the panARS system would ensure constantly high stability, but would require steady addition of the respective antibiotic. Again, this would be highly undesirable for industrial processes. High costs during cultivation, possible problems during purification and potential mutagenic effects exclude this method. Instead, a positive

selection marker system could be applied in which plasmid-bearing cells have a growth advantage over plasmid-free cells.

Transformation of differently truncated and linearized pMito variants highlighted the high *in vivo* ligation capabilities of *P. pastoris*. As demonstrated by Camattari et al. (2016) [172], this property can be exploited for efficiently directing *in vivo* assembly. Interestingly, the smallest pMito variant we used for transformation resulted in dramatically higher (ca. tenfold) transformation efficiencies (figure 5(A) on page 191). On these truncated variants, putative ARS sites on the *AOX1 TT* (associated) part of the expression cassette were removed. However, their inactivity has been shown previously [168]. Rather, the reduced vector size was suspected to be the main explanation, facilitating improved transformation and *in vivo* ligation. If this suspicion holds true, it would be an additional property promoting the application of *in vivo* assembly in *P. pastoris*. Use of smaller fragments makes their generation via PCR and subsequent transformation more efficient.

By replacing the 2 µm ARS of a *S. cerevisiae* vector with the mDNA of pMito, its ARS-activity in *S. cerevisiae* was demonstrated. Transformation efficiencies markedly exceeded the original vector (see figure 8(A) on page 194). The results suggest the applicability of the mDNA encoded ARS in a wide range of budding yeasts, considering the phylogenetic distance between *P. pastoris* and *S. cerevisiae* (see figure 2 on page 7). Similar observations were made for panARS [189]. Ten different budding yeast species were successfully transformed. In the case of pMito, further research is needed to clarify the host range, and whether protein productivity is also high and uniform in other yeasts.

3.5 Summary

The first systematic investigation into the clonal variability of *P. pastoris*, analyzing integration events and their connection to productivity related and unrelated features, yielded several insights into so far undocumented incidents and their effects. A strain library, markedly exceeding previous ones in size, was generated and characterized using established *P. pastoris* methods. Classic productivity features were combined with insights from genome sequencing, augmenting the validity of derived results. This expansion should help scientists working with *P. pastoris*, and similar non-conventional

yeasts, to get a better understanding of what they might encounter during their experiments. More importantly, the suggested measures for preventing negative and promoting positive events will aid in genetic or metabolic engineering projects, as well as studies focused on recombinant protein production. Only by understanding the causes for clonal variability, it can be controlled. The overarching goal being to streamline screening procedures and reduce the requirement for assessing several hundred to thousands of clones to find the right one.

Although much of this project focused on describing undesirable events, its key message is not one of despair. Many of the encountered integration events were rare and would likely not have survived the classic selection process, employed for finding high producer strains. Furthermore, the high producer strains that were analyzed in more detail via genome sequencing were of the expected genotype. Only integrations at the *AOX1* locus were detected and cassettes were orientated in the head-to-tail form. For recombinant protein production studies clonal variability is even beneficial, with high copy strains essentially also being a result of an unforseen incident. No significant effect of the integration site on the productivity was observed. Hence, one major concern for scientists working with *P. pastoris* could be to an extent relieved. The results pertinent to genetic and metabolic engineering studies should enable easier realization of projects with higher complexity. It was the stated goal of the review in chapter 1.3.3 to support the establishment of more intricate engineering approaches in *P. pastoris*. Streamlining the construction process will lower the inhibition level of scientists to start using this yeast in their experiments or expand *P. pastoris* projects they are already working on.

Strikingly, one promising solution for many issues related to clonal variability was itself the product of a very curious integration event. The plasmid pMito and the ARS encoded on its mDNA are good candidates for a variety of applications, in particular those focused on screening a large number of different variants of one target. More research needs to be conducted to understand the inferred bottleneck of productivity and to ensure high plasmid stability during industrial processes. Nevertheless, its features and origin suggest a wealth of other potential ARS on the mDNA of *P. pastoris*. An expansion of the episomal vector repertoire is essential to diversify the applicable genetic techniques and tap new fields of application.

Lastly, one has to see clonal variability of *P. pastoris* in the context of other organisms routinely used in science in general, and recombinant protein expression in particular. Many systems cannot compare to the highly efficient homologous recombination apparatus of *S. cerevisiae* and its suitability for utilizing complex genetic manipulation techniques. Despite its shortcomings however, *P. pastoris* has a distinctively higher genetic accessibility and tractability than many of its competitors. Other non-conventional yeasts like *Y. lipolytica* and *O. angusta* or plant and CHO cell lines typically display worse targeting efficienes and higher clonal variabilities. For comparison, in their review about genetic engineering in the microalgae *Chlamydomonas reinhaardtii* Sizova *et al.* (2013) [190] wrote "In *Chlamydomonas*, the ratio of HR over non-homologous integration of the delivered DNA was $<10^{-4}$, making isolation of homologous recombinants almost impossible". Additionally, the relatively small genome of *P. pastoris* markedly simplifies genome sequencing requirements. For production strains that are designated for large-scale application, implementation of routine genome sequencing appears advisable.

4 Outlook

Despite the large clone library and thorough characterization using conventional and more intricate techniques, not all phenomena could be explained. As discussed in chapter 3.1, parameters like transcript level, growth rate or substrate consumption were not measured. Camara *et al.* (2017) [151] demonstrated the validity of assessing strain features in a chemostat environment, while Aw *et al.* (2017) [150] highlighted the complex relationship between transcriptome data and productivity. A more detailed description of high interest strains might be obtained by implementing these techniques into our approach. To this end, inclusion of ddPCR methods would increase the accuracy of GCN values [113]. The integration of new techniques might be eased by reducing the work-load required for cultivation and productivity screening of the clone library. Thereby, more time would be available for other experiments. Automated systems and appropriate protocols are available in literature, but require specialized equipment [124].

Not all (off-target) integration events could be detected, due to limited genome sequencing capabilities. Unfortunately, sequencing of a pooled sample containing equal amounts of gDNA of all 845 strains did not produce new insights. In yeast, plasmid rescue combined with inverted PCR has been used to characterize off-target integrations of 54 *S. cerevisiae* [191] and 157 *K. lactis* [57] mutants. However, a method initially developed for characterizing gut bacteria and refined for analyzing random integrations in *C. reinhardtii* appears more promising [192]. Called “insertion sequencing”, a library of over 40.000 strains was genotyped for the integration locus of an expression cassette. The method relies on type II restriction enzymes, suitable adapters and high-throughput sequencing on a next-generation sequencing platform. Its application could elucidate whether “hot-spots” of illegitimate insertion exist in *P. pastoris*, potentially coinciding with sites of high DSB repair activity.

Weinhandl *et al.* (2016) [178] demonstrated the suitability of *P. pastoris* strains with impaired cell wall assembly capability for secreted protein production. Similarly, the mutants we found with crenulated colony morphology might be interesting for this purpose. More generally, secreted protein expression represents an important refinement that could be applied to our experimental design. We deliberately chose to focus on intracellular production, to enhance the validity of integration event deduced

statements. Nevertheless, improving on the presented results by switching to secreted expression would likely result in discovery of novel cause-and-effect relationships. Additionally, the transferability of the results to industrial processes would be enhanced.

No scale-up experiments were performed. It would be of interest to see, if specific integration events show the same beneficial or detrimental effect in bioreactor cultivations. While the transfer of results from shake-flask to bioreactor scale was shown to be difficult [193], the good transferability of the deep-well plate method we employed has been demonstrated before [113, 115]. Replication of results from small scale in the big scale can be further improved by employing automated screening methods that allow monitoring of process parameters and running of fed-batch procedures [124].

For pMito, the above mentioned (secreted) expression of other recombinant proteins could serve to further characterize its production capabilities. Similarly, additional (non-conventional) yeast species should be assayed for the ability to be transformed with pMito. Reducing the size of the mDNA element necessary for ARS-activity based on the *in silico* results, would have two benefits. Our findings indicate that transformation efficiency might be increased by the reduced plasmid size. Furthermore, *in vivo* assembly and general vector construction would be aided by removing superfluous parts of the mDNA.

The exact nature of the bottleneck responsible for the (suspected) upper limit of expression of pMito strains remains unclear. Epigenetic factors, like the ones described for a *P. pastoris* strain with altered flocculation [187], are possible but require intricate techniques for analysis. Ribosome profiling, as described in section 3 (page 52) of the review, could give an impression on the presence of such influences. Interestingly, in their analysis of panARS strains, Camattari *et al.* (2016) [172] discovered the presence of two distinct subpopulations via flow cytometry. One high expression and one low expression population were reported, in contrast to integrative clones displaying a broad spectrum of expression levels within one culture. The reduced population heterogeneity was suspected to be a key factor for the high and uniform expression level of panARS strains. Applying microfluidic techniques, pMito and its suitability to reduce population heterogeneity might be assessed in more detail [194].

References

1. Hornsey IS: *A History of Beer and Brewing*. 34th edition. London, UK: Royal Society of Chemistry; 2003.

2. Wilhelm IV, Herzog: **Bayerisches Reinheitsgebot**. 1516, http://www.reinheitsgebot.de/startseite/reinheitsgebot/entstehung/wortlaut/

3. Charlie MJ, Watkinson SC, Gooday GW: *The Fungi*. San Diego, CA: Academic Press; 2001.

4. Yeong FM: **Severing all ties between mother and daughter: Cell separation in budding yeast**. *Mol Microbiol* 2005, **55**:1325–1331.

5. Balasubramanian MK, Bi E, Glotzer M: **Comparative analysis of cytokinesis in budding yeast, fission yeast and animal cells**. *Curr Biol* 2004, **14**:806–818.

6. Martini A: **Biodiversity and conservation of yeasts**. *Biodivers Conserv* 1992, **1**:324–333.

7. Boekhout T: **Biodiversity: gut feeling for yeasts.** *Nature* 2005, **434**:449–451.

8. Liu Y, Wang Y, Chen X, Wu Q, Wang M, Zhu D, Ma Y: **Regio- and stereoselective reduction of 17-oxosteroids to 17β-hydroxysteroids by a yeast strain *Zygowilliopsis* sp. WY7905.** *Steroids* 2017, **118**:17–24.

9. Molnar O, Schatzmayr G, Fuchs E, Prillinger H: ***Trichosporon mycotoxinivorans* sp. nov., a new yeast species useful in biological detoxification of various mycotoxins.** *Syst Appl Microbiol* 2004, **27**:661–671.

10. Sugita T, Takashima M, Kodama M, Tsuboi R, Nishikawa A: **Description of a New Yeast Species, *Malassezia japonica*, and Its Detection in Patients with Atopic Dermatitis and Healthy Subjects**. *J Clin Microbiol* 2003, **41**:4695–4699.

11. Nagahama T, Hamamoto M, Nakase T, Horikoshi K: ***Kluyveromyces nonfermentans* sp. nov., a new yeast species isolated from the deep sea**. *Int J Syst Bacteriol* 1999, **49**:1899–1905.

12. Suh SO, McHugh J V, Pollock DD, Blackwell M: **The beetle gut: a hyperdiverse source of novel yeasts.** *Mycol Res* 2005, **109**:261–265.

13. Blackwell M: **The fungi: 1, 2, 3 ... 5.1 million species?** *Am J Bot* 2011, **98**:426–438.

14. Marx JL: *A Revolution in Biotechnology*. 1st edition. Cambridge, UK: Cambridge University Press; 1989.

15. Clarke L, Carbon J: **Isolation of a yeast centromere and construction of functional small circular chromosomes**. *Nature* 1980, **287**:504–509.

16. Blackburn EH, Szostak JW: **The molecular structure of centromeres and telomeres**. *Annu Rev Biochem* 1984, **53**:163–194.

17. Iwashima A, Nishino H, Nose Y: **Carrier-mediated transport of thiamine in baker's yeast**. *Biochim Biophys Acta* 1973, **330**:222–234.

18. Johnston M, Davis RW: **Sequences that regulate the divergent GAL1-GAL10 promoter in *Saccharomyces cerevisiae*.** *Mol Cell Biol* 1984, **4**:1440–1448.

19. Praekelt UM, Byrne KL, Meacock PA: **Regulation of THI4(MOL1), a thiamine-biosynthetic gene of *Saccharomyces cerevisiae***. *Yeast* 1994, **10**:481–490.

20. Verna J, Lodder A, Lee K, Vagts A, Ballester R: **A family of genes required for maintenance of cell wall integrity and for the stress response in *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci U S A* 1997, **94**:13804–13809.

21. Herskowitz I: **Life cycle of the budding yeast *Saccharomyces cerevisiae*.** *Microbiol Rev* 1988, **52**:536–553.

22. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization**. *Mol Biol Cell* 1998, **9**:3273–3297.

23. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: **Life with 6000 Genes**. *Science* 1995, **274**:546–567.

24. Pretorius IS: **Tailoring wine yeast for the new millennium: Novel approaches to the ancient art of winemaking**. *Yeast* 2000, **16**:675–729.

25. Lodolo EJ, Kock JLF, Axcell BC, Brooks M: **The yeast *Saccharomyces cerevisiae* - The main character in beer brewing**. *FEMS Yeast Res* 2008, **8**:1018–1036.

26. Zhang Z, Wippo CJ, Wal M, Ward E, Korber P, Pugh BF: **A Packing Mechanism for Nucleosome Organization Reconstituted Across a Eukaryotic Genome**. *Science* 2011, **332**:977–980.

27. Smith JS, Hillier AJ, Lees GJ, Jago GR: **The nature of the stimulation of the growth of *Streptococcus lactis* by yeast extract**. *J Dairy Res* 1975, **42**:123.

28. Chae HJ, Joo H, In MJ: **Utilization of brewer's yeast cells for the production of food-grade yeast extract. Part 1: Effects of different enzymatic treatments on solid and protein recovery and flavor characteristics**. *Bioresour Technol* 2001, **76**:253–258.

29. Hahn-Hägerdal B, Galbe M, Gorwa-Grauslund MF, Lidén G, Zacchi G: **Bio-ethanol - the fuel of tomorrow from the residues of today**. *Trends Biotechnol* 2006, **24**:549–556.

30. Lawrence CW: **Mutagenesis in *Saccharomyces cerevisiae***. In *Advances in Genetics*. *Volume 21*. 1st edition. Edited by Caspari EW. New York: Academic Press; 1982:173–254.

31. Klinner U, Schäfer B: **Genetic aspects of targeted insertion mutagenesis in yeasts.** *FEMS Microbiol Rev* 2004, **28**:201–23.

32. Lee ME, DeLoache WC, Cervantes B, Dueber JE: **A Highly Characterized Yeast Toolkit for Modular, Multipart Assembly**. *ACS Synth Biol* 2015, **4**:975–986.

33. Christianson TW, Sikorski RS, Dante M, Shero JH, Hieter P: **Multifunctional yeast high-copy-number shuttle vectors**. *Gene* 1992, **110**:119–122.

34. DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM: **Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems.** *Nucleic Acids Res* 2013, **41**:4336–4343.

35. Generoso WC, Gottardi M, Oreb M, Boles E: **Simplified CRISPR-Cas genome editing for *Saccharomyces cerevisiae***. *J Microbiol Methods* 2016, **127**:203–205.

36. Essani K, Glieder A, Geier M: **Combinatorial pathway assembly in yeast**. *AIMS Bioeng* 2015, **2**:423–436.

37. Nishida K, Arazoe T, Yachie N, Banno S, Kakimoto M, Tabata M, Mochizuki M, Miyabe A, Araki M, Hara KY, Shimatani Z, Kondo A: **Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems.** *Science* 2016, **102**:553–563.

38. Christian M, Cermak T, Doyle EL, Schmidt C, Zhang F, Hummel A, Bogdanove AJ, Voytas DF: **Targeting DNA double-strand breaks with TAL effector nucleases**. *Genetics* 2010, **186**:756–761.

39. Liu Z, Tyo KEJ, Martínez JL, Petranovic D, Nielsen J: **Different expression systems for production of recombinant proteins in *Saccharomyces cerevisiae***. *Biotechnol Bioeng* 2012, **109**:1259–1268.

40. Krivoruchko A, Nielsen J: **Production of natural products through metabolic engineering of *Saccharomyces cerevisiae***. *Curr Opin Biotechnol* 2015, **35**:7–15.

41. Liu Z, Ho S-H, Sasaki K, den Haan R, Inokuma K, Ogino C, van Zyl WH, Hasunuma T, Kondo A: **Engineering of a novel cellulose-adherent cellulolytic *Saccharomyces cerevisiae* for cellulosic biofuel production.** *Sci Rep* 2016, **6**(April):24550.

42. Gruchattka E, Kayser O: **In vivo validation of in silico predicted metabolic engineering strategies in yeast: Disruption of alpha-ketoglutarate dehydrogenase and expression of ATP-citrate lyase for terpenoid production**. *PLoS One* 2015, **10**:1–25.

43. Farhi M, Marhevka E, Masci T, Marcos E, Eyal Y, Ovadis M, Abeliovich H, Vainstein A: **Harnessing yeast subcellular compartments for the production of plant terpenoids**. *Metab Eng* 2011, **13**:474–481.

44. Rugbjerg P, Naesby M, Mortensen UH, Frandsen RJ: **Reconstruction of the biosynthetic pathway for the core fungal polyketide scaffold rubrofusarin in *Saccharomyces cerevisiae*.** *Microb Cell Fact* 2013, **12**:31.

45. Borodina I, Nielsen J: **Advances in metabolic engineering of yeast *Saccharomyces cerevisiae* for production of chemicals**. *Biotechnol J* 2014, **9**:609–620.

46. Hong KK, Nielsen J: **Metabolic engineering of *Saccharomyces cerevisiae*: A key cell factory platform for future biorefineries**. *Cell Mol Life Sci* 2012, **69**:2671–2690.

47. Bill RM: **Playing catch-up with *Escherichia coli*: using yeast to increase success rates in recombinant protein production experiments.** *Front Microbiol* 2014, **5**(March):85.

48. Smith R, Duncan M, Moir D: **Heterologous protein secretion from yeast**. *Science* 1985, **229**:1219–1224.

49. Dean N: **Asparagine-linked glycosylation in the yeast Golgi**. *Biochim Biophys Acta - Gen Subj* 1999, **1426**:309–322.

50. Wolf K, Breuning KD, Barth G: *Non-Conventional Yeasts in Genetics, Biochemistry and Biotechnology*. 1st edition. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003.

51. Letunic I, Bork P: **Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation**. *Bioinformatics* 2007, **23**:127–128.

52. Liu L, Redden H, Alper HS: **Frontiers of yeast metabolic engineering: Diversifying beyond ethanol and *Saccharomyces***. *Curr Opin Biotechnol* 2013, **24**:1023–1030.

53. Radecka D, Mukherjee V, Mateo RQ, Stojiljkovic M, Foulquié-Moreno MR, Thevelein JM: **Looking beyond *Saccharomyces*: The potential of non-conventional yeast species for desirable traits in bioethanol fermentation**. *FEMS Yeast Res* 2015, **15**:1–13.

54. Johnson EA: **Biotechnology of non-*Saccharomyces* yeasts--the ascomycetes.** *Appl Microbiol Biotechnol* 2013, **97**:503–517.

55. Toju H, Tanabe AS, Yamamoto S, Sato H: **High-coverage ITS primers for the DNA-based**

identification of ascomycetes and basidiomycetes in environmental samples. *PLoS One* 2012, **7**.

56. van Ooyen AJJ, Dekker P, Huang M, Olsthoorn MMA, Jacobs DI, Colussi PA, Taron CH: **Heterologous protein production in the yeast *Kluyveromyces lactis***. *FEMS Yeast Res* 2006, **6**:381–392.

57. Kegel A, Martinez P, Carter SD, Åström SU: **Genome wide distribution of illegitimate recombination events in *Kluyveromyces lactis***. *Nucleic Acids Res* 2006, **34**:1633–1645.

58. Fonseca GG, Heinzle E, Wittmann C, Gombert AK: **The yeast *Kluyveromyces marxianus* and its biotechnological potential.** *Appl Microbiol Biotechnol* 2008, **79**:339–54.

59. Limtong S, Sringiew C, Yongmanitchai W: **Production of fuel ethanol at high temperature from sugar cane juice by a newly isolated *Kluyveromyces marxianus*.** *Bioresour Technol* 2007, **98**:3367–3374.

60. Odds FC: ***Candida* Infections: An Overview**. *CRC Crit Rev Microbiol* 1987, **15**:1–5.

61. Bates S, Hughes HB, Munro C a, Thomas WPH, MacCallum DM, Bertram G, Atrih A, Ferguson M a J, Brown AJP, Odds FC, Gow N a R: **Outer chain N-glycans are required for cell wall integrity and virulence of *Candida albicans*.** *J Biol Chem* 2006, **281**:90–98.

62. Falsetta ML, Klein MI, Colonne PM, Scott-Anne K, Gregoire S, Pai C-H, Gonzalez-Begne M, Watson G, Krysan DJ, Bowen WH, Koo H: **Symbiotic Relationship between *Streptococcus mutans* and *Candida albicans* Synergizes Virulence of Plaque Biofilms *In Vivo***. *Infect Immun* 2014, **82**:1968–1981.

63. Vyas VK, Barrasa MI, Fink GR: **A *Candida albicans* CRISPR system permits genetic engineering of essential genes and gene families**. *Sci Adv* 2015, **1**:e1500248–e1500248.

64. Pomraning KR, Kim Y-M, Nicora CD, Chu RK, Bredeweg EL, Purvine SO, Hu D, Metz TO, Baker SE: **Multi-omics analysis reveals regulators of the response to nitrogen limitation in *Yarrowia lipolytica***. *BMC Genomics* 2016, **17**:138.

65. Guevara-Olvera L, Calvo-Mendez C, Ruiz-Herrera J: **The role of polyamine metabolism in dimorphism of *Yarrowia lipolytica*.** *J Gen Microbiol* 1993, **139**:485–93.

66. Madzak C, Beckerich J-M: **Heterologous Protein Expression and Secretion in *Yarrowia lipolytica***. In *Yarrowia Lipolytica*. *Volume 25*. Edited by Barth G. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013:1–76. [*Microbiology Monographs*]

67. Gellissen G, Kunze G, Gaillardin C, Cregg JM, Berardi E, Veenhuis M, van der Klei I: **New yeast expression platforms based on methylotrophic *Hansenula polymorpha* and *Pichia pastoris* and on dimorphic *Arxula adeninivorans* and *Yarrowia lipolytica* - a comparison.** *FEMS Yeast Res* 2005, **5**:1079–96.

68. Shabbir Hussain M, M Rodriguez G, Gao D, Spagnuolo M, Gambill L, Blenner M: **Recent advances in bioengineering of the oleaginous yeast *Yarrowia lipolytica***. *AIMS Bioeng* 2016, **3**:493–514.

69. Xue Z, Sharpe PL, Hong S-P, Yadav NS, Xie D, Short DR, Damude HG, Rupert R a, Seip JE, Wang J, Pollak DW, Bostick MW, Bosak MD, Macool DJ, Hollerbach DH, Zhang H, Arcilla DM, Bledsoe S a, Croker K, McCord EF, Tyreus BD, Jackson EN, Zhu Q: **Production of omega-3 eicosapentaenoic acid by metabolic engineering of *Yarrowia lipolytica*.** *Nat Biotechnol* 2013, **31**:734–740.

70. Beopoulos A, Verbeke J, Bordes F, Guicherd M, Bressy M, Marty A, Nicaud J-M: **Metabolic**

engineering for ricinoleic acid production in the oleaginous yeast *Yarrowia lipolytica*. *Appl Microbiol Biotechnol* 2014, **98**:251–262.

71. Schwartz CM, Hussain MS, Blenner M, Wheeldon I: **Synthetic RNA polymerase III promoters facilitate high efficiency CRISPR-Cas9 mediated genome editing in *Yarrowia lipolytica***. *ACS Synth Biol* 2015, **5**:356–359.

72. Yurimoto H, Sakai Y, Kato N: **Methanol metabolism**. In *Hansenula polymorpha: Biology and applications*; 2002:61–75.

73. Gleeson MA, Sudbery PE: **The methylotrophic yeasts**. *Yeast* 1988, **4**:1–15.

74. Saraya R, Krikken AM, Kiel J a KW, Baerends RJS, Veenhuis M, van der Klei IJ: **Novel genetic tools for *Hansenula polymorpha***. *FEMS Yeast Res* 2012, **12**:271–278.

75. Gatzke R, Weydemann U, Janowicz Z a., Hollenberg CP: **Stable multicopy integration of vector sequences in *Hansenula polymorpha***. *Appl Microbiol Biotechnol* 1995, **43**:844–849.

76. Phaff HJ: **My Life With Yeasts**. *Annu Rev Microbiol* 1986, **40**:1–53.

77. Kurtzman CP: **Biotechnological strains of *Komagataella (Pichia) pastoris* are *Komagataella phaffii* as determined from multigene sequence analysis**. *J Ind Microbiol Biotechnol* 2009, **36**:1435–1438.

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

79. Wegner GH: **Emerging applications of the methylotrophic yeasts**. *FEMS Microbiol Lett* 1990, **87**:279–284.

80. Couderc R, Baratti J: **Oxidation of methanol by the yeast, *Pichia pastoris* : Purification and properties of the alcohol oxidase**. *Agric Biol Chem* 1980, **44**:2279–2289.

81. Cregg J, Barringer K, Hessler A, Madden K: ***Pichia pastoris* as a host system for transformations.** *Mol Cell Biol* 1985, **5**:3376–3385.

82. Ellis SB, Brust PF, Koutz PJ, Waters AF, Harpold MM, Gingeras TR: **Isolation of alcohol oxidase and two other methanol regulatable genes from the yeast *Pichia pastoris* .** *Mol Cell Biol* 1985, **5**:1111–1121.

83. Siegel R, Brierley R: **Methylotrophic Yeast *Pichia pastoris* Produced in High Cell Density Fermentations with High Cell Yields as Vehicle for Recombinant Protein Production**. *Biotechnol Bioeng* 1989, **34**:403–404.

84. Cregg J, Cereghino J, Shi J, Higgins D: **Recombinant protein expression in *Pichia pastoris***. *Mol Biotechnol* 2000, **16**:23–52.

85. Ahmad M, Hirz M, Pichler H, Schwab H: **Protein expression in *Pichia pastoris* : recent achievements and perspectives for heterologous protein production.** *Appl Microbiol Biotechnol* 2014, **98**:5301–5317.

86. Chen M-T, Lin S, Shandil I, Andrews D, Stadheim T a, Choi B-K: **Generation of diploid *Pichia pastoris* strains by mating and their application for recombinant protein production**. *Microb Cell Fact* 2012, **11**:91.

87. Ciofalo V, Barton N, Kreps J, Coats I, Shanahan D: **Safety evaluation of a lipase enzyme preparation, expressed in *Pichia pastoris* , intended for use in the degumming of edible vegetable oil.** *Regul Toxicol Pharmacol* 2006, **45**:1–8.

88. Meehl MA, Stadheim TA: **Biopharmaceutical discovery and production in yeast**. *Curr Opin Biotechnol* 2014, **30**:120–127.

89. Polupanov AS, Nazarko VY, Sibirny AA: **CCZ1, MON1 and YPT7 genes are involved in pexophagy, the Cvt pathway and non-specific macroautophagy in the methylotrophic yeast *Pichia pastoris*.** *Cell Biol Int* 2011, **35**:311–9.

90. Waterham HR, Cregg JM: **Peroxisome biogenesis**. *BioEssays* 1997, **19**:57–66.

91. Lüers GH, Advani R, Wenzel T, Subramani S: **The *Pichia pastoris* dihydroxyacetone kinase is a PTS1-containing, but cytosolic, protein that is essential for growth on methanol**. *Yeast* 1998, **14**:759–771.

92. Hoeffler JP: **So many possibilities: How to choose a system to achieve your specific goal**. In *Gene Expression Systems*. 1st edition. Edited by Fernandez J, Hoeffler JP. New York: Academic Press; 1999:1–5.

93. Brondyk WH: **Selecting an appropriate method for expressing a recombinant protein.** *Methods Enzymol* 2009, **463**(C):131–147.

94. Tran A-M, Nguyen T-T, Nguyen C-T, Huynh-Thi X-M, Nguyen C-T, Trinh M-T, Tran L-T, Cartwright SP, Bill RM, Tran-Van H: ***Pichia pastoris* versus *Saccharomyces cerevisiae*: a case study on the recombinant production of human granulocyte-macrophage colony-stimulating factor.** *BMC Res Notes* 2017, **10**:148.

95. Byrne B: ***Pichia pastoris* as an expression host for membrane protein structural biology**. *Curr Opin Struct Biol* 2015, **32**:9–17.

96. Looser V, Bruhlmann B, Bumbak F, Stenger C, Costa M, Camattari A, Fotiadis D, Kovar K: **Cultivation strategies to enhance productivity of *Pichia pastoris*: A review**. *Biotechnol Adv* 2015, **33**:1177–1193.

97. Laukens B, De Visscher C, Callewaert N: **Engineering yeast for producing human glycoproteins: where are we now?** *Future Microbiol* 2015, **10**:21–34.

98. Gmeiner C, Saadati A, Maresch D, Krasteva S, Frank M, Altmann F, Herwig C, Spadiut O: **Development of a fed-batch process for a recombinant *Pichia pastoris Δoch1* strain expressing a plant peroxidase.** *Microb Cell Fact* 2015, **14**:1.

99. Maccani A, Landes N, Stadlmayr G, Maresch D, Leitner C, Maurer M, Gasser B, Ernst W, Kunert R, Mattanovich D: ***Pichia pastoris* secretes recombinant proteins less efficiently than Chinese hamster ovary cells but allows higher space-time yields for less complex proteins**. *Biotechnol J* 2014, **9**:526–537.

100. Schröder M, Kaufman RJ: **The mammalian unfolded protein response.** *Annu Rev Biochem* 2005, **74**:739–789.

101. Smith MH, Ploegh HL, Weissman JS: **Road to ruin: targeting proteins for degradation in the endoplasmic reticulum.** *Science* 2011, **334**:1086–1090.

102. Graf A, Gasser B, Dragosits M, Sauer M, Leparc GG, Tüchler T, Kreil DP, Mattanovich D: **Novel insights into the unfolded protein response using *Pichia pastoris* specific DNA microarrays**. *BMC Genomics* 2008, **9**:390.

103. Delic M, Göngrich R, Mattanovich D, Gasser B: **Engineering of protein folding and secretion-strategies to overcome bottlenecks for efficient production of recombinant proteins.** *Antioxid Redox Signal* 2014, **21**:414–37.

104. Ruth C, Buchetics M, Vidimce V, Kotz D, Naschberger S, Mattanovich D, Pichler H, Gasser B: ***Pichia pastoris* Aft1-a novel transcription factor, enhancing recombinant protein secretion**. *Microb Cell Fact* 2014, **13**:120.

105. Delic M, Valli M, Graf AB, Pfeffer M, Mattanovich D, Gasser B: **The secretory pathway: Exploring yeast diversity**. *FEMS Microbiol Rev* 2013, **37**:872–914.

106. Cregg JM, Tolstorukov I, Kusari A, Sunga J, Madden K, Chappell T: Expression in the Yeast *Pichia Pastoris* . 1st edition. *Volume 463*. Elsevier Inc.; 2009.

107. Guirouilh-Barbat J, Huck S, Bertrand P, Pirzio L, Desmaze C, Sabatier L, Lopez BS: **Impact of the KU80 pathway on NHEJ-induced genome rearrangements in mammalian cells**. *Mol Cell* 2004, **14**:611–623.

108. Daley JM, Palmbos PL, Wu D, Wilson TE: **Nonhomologous end joining in yeast.** *Annu Rev Genet* 2005, **39**:431–451.

109. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, Connelly C, Davis K, Dietrich F, Dow SW, El Bakkoury M, Foury F, Friend SH, Gentalen E, Giaever G, Hegemann JH, Jones T, Laub M, Liao H, Liebundguth N, Lockhart DJ, Lucau-Danila A, Lussier M, M'Rabet N, Menard P, et al.: **Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis.** *Science* 1999, **285**:901–906.

110. Brachmann CB, Davies A, Cost GJ, Caputo E, Li J, Hieter P, Boeke JD: **Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications.** *Yeast* 1998, **14**:115–132.

111. Näätsaari L, Mistlberger B, Ruth C, Hajek T, Hartner FS, Glieder A: **Deletion of the *Pichia pastoris* KU70 homologue facilitates platform strain generation for gene expression and synthetic biology.** *PLoS One* 2012, **7**:e39720.

112. Nett JH, Hodel N, Rausch S, Wildt S: **Cloning and disruption of the *Pichia pastoris* ARG1, ARG2, ARG3, HIS1, HIS2, HIS5, HIS6 genes and their use as auxotrophic markers.** *Yeast* 2005, **22**:295–304.

113. Cámara E, Albiol J, Ferrer P: **Droplet digital PCR-aided screening and characterization of *Pichia pastoris* multiple gene copy strains**. *Biotechnol Bioeng* 2016, **113**:1542–1551.

114. Clare J, Rayment F, Sreekrishna K, Romanos MA: **High-level expression of tetanus toxin fragment C in *Pichia pastoris* strains containing multiple tandem integrations of the gene**. *Nat Biotechnol* 1991, **9**:455–460.

115. Weis R, Luiten R, Skranc W, Schwab H, Wubbolts M, Glieder A: **Reliable high-throughput screening with *Pichia pastoris* by limiting yeast cell death phenomena.** *FEMS Yeast Res* 2004, **5**:179–189.

116. Boettner M, Prinz B, Holz C, Stahl U, Lang C: **High-throughput screening for expression of heterologous proteins in the yeast *Pichia pastoris* .** *J Biotechnol* 2002, **99**:51–62.

117. Hartner FS, Ruth C, Langenegger D, Johnson SN, Hyka P, Lin-Cereghino GP, Lin-Cereghino J, Kovar K, Cregg JM, Glieder A: **Promoter library designed for fine-tuned gene expression in *Pichia pastoris* .** *Nucleic Acids Res* 2008, **36**:e76.

118. Krainer FW, Gmeiner C, Neutsch L, Windwarder M, Pletzenauer R, Herwig C, Altmann F, Glieder A, Spadiut O: **Knockout of an endogenous mannosyltransferase increases the**

homogeneity of glycoproteins produced in *Pichia pastoris* . *Sci Rep* 2013, **3**:3279.

119. Wriessnegger T, Augustin P, Engleder M, Leitner E, Müller M, Kaluzna I, Schürmann M, Mink D, Zellnig G, Schwab H, Pichler H: **Production of the sesquiterpenoid (+)-nootkatone by metabolic engineering of *Pichia pastoris***. *Metab Eng* 2014, **24**:18–29.

120. Brooks CL, Morrison M, Joanne Lemieux M: **Rapid expression screening of eukaryotic membrane proteins in *Pichia pastoris* .** *Protein Sci* 2013, **22**:425–33.

121. Shah KA, Clark JJ, Goods BA, Politano TJ, Mozdzierz NJ, Zimnisky RM, Leeson RL, Love JC, Love KR: **Automated pipeline for rapid production and screening of HIV-specific monoclonal antibodies using *Pichia pastoris***. *Biotechnol Bioeng* 2015, **112**:2624–2629.

122. Zheng J, Guo N, Lin F, Wu L, Zhou H: **Screening of multi-copy mannanase recombinants of *Pichia pastoris* based on colony size.** *World J Microbiol Biotechnol* 2014, **30**:579–584.

123. Hu F, Li X, Lü J, Mao PH, Jin X, Rao B, Zheng P, Zhou YL, Liu SY, Ke T, Ma XD, Ma LX: **A visual method for direct selection of high-producing *Pichia pastoris* clones.** *BMC Biotechnol* 2011, **11**:23.

124. Hemmerich J, Adelantado N, Barrigón JM, Ponte X, Hörmann A, Ferrer P, Kensy F, Valero F: **Comprehensive clone screening and evaluation of fed-batch strategies in a microbioreactor and lab scale stirred tank bioreactor system: application on *Pichia pastoris* producing *Rhizopus oryzae* lipase.** *Microb Cell Fact* 2014, **13**:36.

125. Tredwell GD, Aw R, Edwards-Jones B, Leak DJ, Bundy JG: **Rapid screening of cellular stress responses in recombinant *Pichia pastoris* strains using metabolite profiling**. *J Ind Microbiol Biotechnol* 2017, **44**:413–417.

126. Kang Z, Huang H, Zhang Y, Du G, Chen J: **Recent advances of molecular toolbox construction expand *Pichia pastoris* in synthetic biology applications**. *World J Microbiol Biotechnol* 2017, **33**:19.

127. Wagner JM, Alper HS: **Synthetic biology and molecular genetics in non-conventional yeasts: Current tools and future advances**. *Fungal Genet Biol* 2016, **89**:126–136.

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

129. Tsakraklides V, Brevnova E, Stephanopoulos G, Shaw AJ: **Improved gene targeting through cell cycle synchronization**. *PLoS One* 2015, **10**:1–11.

130. Daly R, Hearn MTW: **Expression of heterologous proteins in *Pichia pastoris*: A useful experimental tool in protein engineenring and production**. *J Mol Recognit* 2005, **18**:119–138.

131. Qin X, Qian J, Yao G, Zhuang Y, Zhang S, Chu J: **GAP promoter library for fine-tuning of gene expression in *Pichia pastoris*.** *Appl Environ Microbiol* 2011, **77**:3600–8.

132. Vogl T, Ruth C, Pitzer J, Kickenweiz T, Glieder A: **Synthetic core promoters for *Pichia pastoris***. *ACS Synth Biol* 2014, **3**:188–191.

133. Wu S, Letchworth GJ: **High efficiency transformation by electroporation of *Pichia pastoris* pretreated with lithium acetate and dithiothreitol.** *Biotechniques* 2004, **36**:152–154.

134. Lin-Cereghino J, Wong WW, Xiong S, Giang W, Luong LT, Vu J, Johnson SD, Lin-Cereghino GP: **Condensed protocol for competent cell preparation and transformation of the methylotrophic yeast *Pichia pastoris* .** *Biotechniques* 2005, **38**:44, 46, 48.

135. Abad S, Kitz K, Hörmann A, Schreiner U, Hartner FS, Glieder A: **Real-time PCR-based determination of gene copy numbers in *Pichia pastoris* .** *Biotechnol J* 2010, **5**:413–420.

136. Harju S, Fedosyuk H, Peterson KR: **Rapid isolation of yeast genomic DNA: Bust n' Grab.** *BMC Biotechnol* 2004, **4**:8.

137. Küberl A, Schneider J, Thallinger GG, Anderl I, Wibberg D, Hajek T, Jaenicke S, Brinkrolf K, Goesmann A, Szczepanowski R, Pühler A, Schwab H, Glieder A, Pichler H: **High-quality genome sequence of *Pichia pastoris* CBS7435.** *J Biotechnol* 2011, **154**:312–20.

138. Schwarzhans J-P, Wibberg D, Winkler A, Luttermann T, Kalinowski J, Friehs K: **Integration event induced changes in recombinant protein productivity in *Pichia pastoris* discovered by whole genome sequencing and derived vector optimization**. *Microb Cell Fact* 2016, **15**:84.

139. Vervecken W, Kaigorodov V, Callewaert N, Geysens S, De Vusser K, Contreras R: **In vivo synthesis of mammalian-like, hybrid-type N-glycans in *Pichia pastoris***. *Appl Environ Microbiol* 2004, **70**:2639–2646.

140. Shen W, Kong C, Xue Y, Liu Y, Cai M, Zhang Y, Jiang T, Zhou X, Zhou M: **Kinase Screening in *Pichia pastoris* Identified Promising Targets Involved in Cell Growth and Alcohol Oxidase 1 Promoter (PAOX1) Regulation.** *PLoS One* 2016, **11**:e0167766.

141. Jakočiūnas T, Bonde I, Herrgård M, Harrison SJ, Kristensen M, Pedersen LE, Jensen MK, Keasling JD: **Multiplex metabolic pathway engineering using CRISPR/Cas9 in *Saccharomyces cerevisiae***. *Metab Eng* 2015, **28**:213–222.

142. Mans R, van Rossum HM, Wijsman M, Backx A, Kuijpers NGA, van den Broek M, Daran-Lapujade P, Pronk JT, van Maris AJA, Daran J-MG: **CRISPR/Cas9: a molecular Swiss army knife for simultaneous introduction of multiple genetic modifications in *Saccharomyces cerevisiae***. *FEMS Yeast Res* 2015(January):1–15.

143. Schwarzhans J-P, Wibberg D, Winkler A, Luttermann T, Kalinowski J, Friehs K: **Non-canonical integration events in *Pichia pastoris* encountered during standard transformation analysed with genome sequencing**. *Sci Rep* 2016, **6**:38952.

144. Zhu T, Guo M, Sun C, Qian J, Zhuang Y, Chu J, Zhang S: **A systematical investigation on the genetic stability of multi-copy *Pichia pastoris* strains.** *Biotechnol Lett* 2009, **31**:679–84.

145. Gunge N: **Yeast DNA plasmids**. *Annu Rev Microbiol* 1983, **37**:253–276.

146. Rose MD, Novick P, Thomas JH, Botstein D, Fink GR: **A *Saccharomyces cerevisiae* genomic plasmid bank based on a centromere-containing shuttle vector**. *Gene* 1987, **60**:237–243.

147. Kelwick R, MacDonald JT, Webb AJ, Freemont P: **Developments in the tools and methodologies of synthetic biology.** *Front Bioeng Biotechnol* 2014, **2**:60.

148. Schwarzhans J-P, Luttermann T, Wibberg D, Winkler A, Hübner W, Huser T, Kalinowski J, Friehs K: **A Mitochondrial Autonomously Replicating Sequence from *Pichia pastoris* for Uniform High Level Recombinant Protein Production**. *Front Microbiol* 2017, **8**:780.

149. Cregg JM, Madden KR, Barringer KJ, Thill GP, Stillman CA: **Functional characterization of the two alcohol oxidase genes from the yeast *Pichia pastoris* .** *Mol Cell Biol* 1989,

9:1316–1323.

150. Aw R, Barton GR, Leak DJ: **Insights into the prevalence and underlying causes of clonal variation through transcriptomic analysis in *Pichia pastoris***. *Appl Microbiol Biotechnol* 2017.

151. Cámara E, Landes N, Albiol J, Gasser B, Mattanovich D, Ferrer P: **Increased dosage of AOX1 promoter-regulated expression cassettes leads to transcription attenuation of the methanol metabolism in *Pichia pastoris***. *Sci Rep* 2017, **7**:44302.

152. Shen Q, Wu M, Wu M, Wang H-B, Naranmandura H, Chen S-Q: **The effect of gene copy number and co-expression of chaperone on production of albumin fusion proteins in *Pichia pastoris* .** *Appl Microbiol Biotechnol* 2012, **96**:763–772.

153. Zhu T, Guo M, Tang Z, Zhang M, Zhuang Y, Chu J, Zhang S: **Efficient generation of multi-copy strains for optimizing secretory expression of porcine insulin precursor in yeast *Pichia pastoris***. *J Appl Microbiol* 2009, **107**:954–963.

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

155. Viader-Salvadó JM, Cab-Barrera EL, Galán-Wong LJ, Guerrero-Olazarán M: **Genotyping of recombinant Pichia pastoris strains.** *Cell Mol Biol Lett* 2006, **11**:348–359.

156. Inan M, Aryasomayajula D, Sinha J, Meagher MM: **Enhancement of protein secretion in *Pichia pastoris* by overexpression of protein disulfide isomerase.** *Biotechnol Bioeng* 2006, **93**:771–8.

157. Mansur M, Cabello C, Hernández L, País J, Varas L, Valdés J, Terrero Y, Hidalgo A, Plana L, Besada V, García L, Lamazares E, Castellanos L, Martínez E: **Multiple gene copy number enhances insulin precursor secretion in the yeast *Pichia pastoris* .** *Biotechnol Lett* 2005, **27**:339–345.

158. Hohenblum H, Gasser B, Maurer M, Borth N, Mattanovich D: **Effects of gene dosage, promoters, and substrates on unfolded protein stress of recombinant *Pichia pastoris* .** *Biotechnol Bioeng* 2004, **85**:367–375.

159. Vassileva A, Chugh DA, Swaminathan S, Khanna N: **Effect of copy number on the expression levels of hepatitis B surface antigen in the methylotrophic yeast *Pichia pastoris* .** *Protein Expr Purif* 2001, **21**:71–80.

160. Aw R, Polizzi KM: **Can too many copies spoil the broth?** *Microb Cell Fact* 2013, **12**:128.

161. Sunga AJ, Tolstorukov I, Cregg JM: **Posttransformational vector amplification in the yeast *Pichia pastoris***. *FEMS Yeast Res* 2008, **8**:870–876.

162. Love KR, Politano TJ, Panagiotou V, Jiang B, Stadheim TA, Love JC: **Systematic single-cell analysis of *Pichia pastoris* reveals secretory capacity limits productivity**. *PLoS One* 2012, **7**:1–11.

163. Perez-Pinera P, Han N, Cleto S, Cao J, Purcell O, Shah KA, Lee K, Ram R, Lu TK: **Synthetic biology and microbioreactor platforms for programmable production of biologics at the point-of-care**. *Nat Commun* 2016, **7**:12211.

164. Love KR, Shah KA, Whittaker CA, Wu J, Bartlett MC, Ma D, Leeson RL, Priest M, Borowsky J, Young SK, Love JC: **Comparative genomics and transcriptomics of *Pichia pastoris***. *BMC Genomics* 2016, **17**:550.

165. Jiang B, Argyros R, Bukowski J, Nelson S, Sharkey N, Kim S, Copeland V, Davidson RC, Chen R, Zhuang J, Sethuraman N, Stadheim TA: **Inactivation of a GAL4-like transcription factor improves cell fitness and product yield in glycoengineered *Pichia pastoris* strains**. *Appl Environ Microbiol* 2015, **81**:260–271.

166. Crampton N, Bonass WA, Kirkham J, Rivetti C, Thomson NH: **Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy**. *Nucleic Acids Res* 2006, **34**:5416–5425.

167. Hobson DJ, Wei W, Steinmetz LM, Svejstrup JQ: **RNA polymerase II collision interrupts convergent transcription.** *Mol Cell* 2012, **48**:365–374.

168. Vogl T, Sturmberger L, Kickenweiz T, Wasmayer R, Schmid C, Hatzl A-M, Gerstmann MA, Pitzer J, Wagner M, Thallinger GG, Geier M, Glieder A: **A toolbox of diverse promoters related to methanol utilization – functionally verified parts for heterologous pathway expression in *Pichia pastoris***. *ACS Synth Biol* 2016, **5**:172–186.

169. Li B, Howe L, Anderson S, Yates JR, Workmann JL: **The Set2 Histone Methyltransferase Functions through the Phosphorylated Carboxyl-terminal Domain of RNA Polymerase II**. *J Biol Chem* 2003, **278**:8897–8903.

170. Vogl T, Glieder A: **Regulation of *Pichia pastoris* promoters and its consequences for protein production.** *N Biotechnol* 2013, **30**:385–404.

171. Gullerova M, Proudfoot NJ: **Transcriptional interference and gene orientation in yeast: noncoding RNA connections.** *Cold Spring Harb Symp Quant Biol* 2010, **75**:299–311.

172. Camattari A, Goh A, Yip LY, Tan AHM, Ng SW, Tran A, Liu G, Liachko I, Dunham MJ, Rancati G: **Characterization of a panARS-based episomal vector in the methylotrophic yeast *Pichia pastoris* for recombinant protein production and synthetic biology applications**. *Microb Cell Fact* 2016, **15**:139.

173. Heiss S, Maurer M, Hahn R, Mattanovich D, Gasser B: **Identification and deletion of the major secreted protein of *Pichia pastoris* .** *Appl Microbiol Biotechnol* 2013, **97**:1241–9.

174. Pronk JT: **Auxotrophic Yeast Strains in Fundamental and Applied Research**. *Appl Environ Microbiol* 2002, **68**:2095–2100.

175. Curran KA, Karim AS, Gupta A, Alper HS: **Use of expression-enhancing terminators in *Saccharomyces cerevisiae* to increase mRNA half-life and improve gene expression control for metabolic engineering applications**. *Metab Eng* 2013, **19**:88–97.

176. Buchetics M, Baumann K, Valli M, Grillitsch K, Daum G, Zellnig G, Mattanovich D, Gasser B: **A binuclear mutantt of *Pichia pastoris* was selected as an eficient protein producer**. *RPP 9 Conf Poster* 2017, **1**:1.

177. Nocon J, Steiger MG, Pfeffer M, Sohn SB, Kim TY, Maurer M, Rußmayer H, Pflügl S, Ask M, Haberhauer-Troyer C, Ortmayr K, Hann S, Koellensperger G, Gasser B, Lee SY, Mattanovich D: **Model based engineering of *Pichia pastoris* central metabolism enhances recombinant protein production.** *Metab Eng* 2014, **24**:129–38.

178. Weinhandl K, Ballach M, Winkler M, Ahmad M, Glieder A, Birner-Gruenberger R, Fotheringham I, Escalettes F, Camattari A: ***Pichia pastoris* mutants as host strains for efficient secretion of recombinant branched chain aminotransferase (BCAT).** *J Biotechnol* 2016, **235**:84–91.

179. Krainer FW, Gerstmann MA, Darnhofer B, Birner-Gruenberger R, Glieder A: **Biotechnological advances towards an enhanced peroxidase production in *Pichia***

pastoris. *J Biotechnol* 2016, **233**:181–189.

180. Du M, Battles MB, Nett JH: **A color-based stable multi-copy integrant selection system for *Pichia pastoris* using the attenuated ADE1 and ADE2 genes as auxotrophic markers.** *Bioeng Bugs* 2012, **3**:32–37.

181. Petsch D, Anspach FB: **Endotoxin removal from protein solutions**. *J Biotechnol* 2000, **76**:97–119.

182. Nasar F, Jankowski C, Nag DK: **Long palindromic sequences induce double-strand breaks during meiosis in yeast.** *Mol Cell Biol* 2000, **20**:3449–3458.

183. Nag DK, White MA, Petes TD: **Palindromic sequences in heteroduplex DNA inhibit mismatch repair in yeast**. *Nature* 1989, **340**:318–320.

184. Manivasakam P, Schiestl RH: **Nonhomologous end joining during restriction enzyme-mediated DNA integration in *Saccharomyces cerevisiae*.** *Mol Cell Biol* 1998, **18**:1736–1745.

185. Hazkani-Covo E, Zeller RM, Martin W: **Molecular poltergeists: Mitochondrial DNA copies (numts) in sequenced nuclear genomes**. *PLoS Genet* 2010, **6**.

186. Ricchetti M, Fairhead C, Dujon B: **Mitochondrial DNA repairs double-strand breaks in yeast chromosomes.** *Nature* 1999, **402**:96–100.

187. De S, Rebnegger C, Mattanovich D, Gasser B: **Role of epigenetic regulation in flocculation and pseudo-hyphal growth of the protein production host *Pichia pastoris***. *RPP 9 Conf Poster* 2017, **1**:1.

188. Friehs K: **Plasmid Copy Number and Plasmid Stability**. In *New Trends and Developments in Biochemical Engineering*. Edited by Scheper T. Berlin: Springer Berlin Heidelberg; 2004:47–82.

189. Liachko I, Dunham MJ: **An autonomously replicating sequence for use in a wide range of budding yeasts.** *FEMS Yeast Res* 2014, **14**:364–7.

190. Sizova I, Greiner A, Awasthi M, Kateriya S, Hegemann P: **Nuclear gene targeting in *Chlamydomonas* using engineered zinc-finger nucleases**. *Plant J* 2013, **73**:873–882.

191. Zhu J, Schiestl RH: **Topoisomerase I involvement in illegitimate recombination in *Saccharomyces cerevisiae*.** *Mol Cell Biol* 1996, **16**:1805–1812.

192. Zhang R, Patena W, Armbruster U, Gang SS, Blum SR, Jonikas MC: **High-Throughput Genotyping of Green Algal Mutants Reveals Random Distribution of Mutagenic Insertion Sites and Endonucleolytic Cleavage of Transforming DNA.** *Plant Cell* 2014, **26**:1398–1409.

193. Clare JJ, Romanos MA, Rayment FB, Rowedder JE, Smith MA, Payne MM, Sreekrishna K, Henwood CA: **Production of mouse epidermal growth factor in yeast: high-level secretion using *Pichia pastoris* strains containing multiple gene copies.** *Gene* 1991, **105**:205–212.

194. Grünberger A, Wiechert W, Kohlheyer D: **Single-cell microfluidics: Opportunity for bioprocess development**. *Curr Opin Biotechnol* 2014, **29**:15–23.

Acknowledgments/Danksagung

Meinem Doktorvater Herrn **Prof. Dr. Karl Friehs**, Leiter des Lehrstuhls für Fermentationstechnik, gilt mein äußerst großer Dank dafür, dass ich auf diesem Projekt arbeiten konnte. Er legte den Grundstein für das Projekt und sorgte durch seine Betreuung dafür, dass meine Begeisterung für *P. pastoris* entfacht wurde. Seine fachlichen Ratschlägen und die Gelegenheit erste Indizien weiter verfolgen zu dürfen, ermöglichten es die Dissertation in der hier präsentierten Form zu vollbringen. Oftmals kam ich auf Ihn mit scheinbar absurden Ergebnissen zu, aber er ließ sich nicht aus der Ruhe bringen und machte mir Mut an das „Unmögliche" zu glauben. Auch wenn ich den Wald vor lauter Bäumen nicht sehen konnte, wusste er den passenden Hinweis, um Ordnung in das Chaos zu bringen. Ich bin sehr dankbar für die Unterstützung auf wissenschaftlicher und besonders persönlicher Ebene, und auch für die kollegiale und freundliche Arbeitsatmosphäre auf die der „**Einmalige**" viel Wert legt. Auch bin ich dankbar, dass mir die Möglichkeit gegeben wurde an diversen Konferenzen teilzunehmen und das kleine bisschen extra Zeit, um das Projekt abzuschließen.

Ebenso bin ich sehr dankbar für die Betreuung durch **Prof. Dr. Jörn Kalinowski**, Leiter der AG „Microbial Genomics and Biotechnology". Er übernahm eine sehr aktive Rolle als Zweit-Betreuer der Dissertation und ermöglichte dadurch eine erfolgreiche Kooperation zwischen beiden AGs. Auch Ihn überraschte ich ab und zu mit den seltsamen Vorkommnissen in *P. pastoris*, aber er hatte stets einen guten Hinweis für weitere Analysen parat. In der Diskussion über Ergebnisse fielen Ihm Lösungsansätze ein, an die ich nie gedacht hätte. Er zeigte mir somit wiederholt den hohen Wert davon sich eine zweite Meinung einzuholen, um sich nicht zu sehr in der eigenen Denkweise einzufahren. Zusätzlich ermöglichte es die moderne Ausstattung seiner AG die Genomsequenzierungsarbeiten in diesem Umfang zu erledigen. Die Bereitschaft und das große Interesse an kooperativer Arbeit war einer der Aspekte, den ich im Rahmen dieser Doktorarbeit sehr zu schätzen lernte.

Herrn **Prof. Dr. Lars M. Blank**, Leiter des Institut für Angewandte Mikrobiologie der RWTH Aachen, danke ich vielmals für die Bereitschaft diese Arbeit zu begutachten. Er kennt sich mit den Eigenheiten von *P. pastoris* nur zu gut aus, und ich hoffe die hier präsentierten Ergebnisse erweisen sich für seine Forschung als hilfreich.

Ohne die Hilfe mehrerer Kooperationspartner hätte die Arbeit nicht in dieser Form abgelegt werden können. Allen voran ist **Dr. Daniel Wibberg** von der AG „Genome Research of Industrial Microorganisms" zu nennen. Seine Expertise mit Genomdaten generell und *P. pastoris* speziell, war essentiell für das Verständnis der verschiedenen Integrationsevents. Ich denke auch er war oftmals verblüfft von den Ergebnissen, aber er fand immer einen Weg alles aus den

Rohdaten rauszuholen, damit wir mehr über *P. pastoris* lernen konnten. Es hat mich sehr gefreut meinen Beitrag zur mittlerweile erweiterten Liste der „Daniel Wibberg Library of Science“ zu leisten. Dem Einsatz von **Dr. Wolfgang Hübner** aus der AG „Biomolecular Photonics“ sind die schönen Fluoreszenzmikroskopiebilder zu verdanken. Er hat sich voller Begeisterung für das Projekt eingesetzt und mir nebenbei einiges über die Möglichkeiten und Techniken der Fluoreszenzmikroskopie beigebracht. Frau **Dr. Martina Geier** vom „Austrian Center of Industrial Biotechnology“ in Graz, Österreich danke ich vielmals für die Mitarbeit beim hier präsentierten Review. Ihr wertvolles *feedback* sorgte für eine bessere Strukturierung und größeren Fokus auf Themen mit hoher Relevanz für die Pichia-Community. Auch möchte ich **Anika Winkler** aus der AG von Jörn Kalinowski für die Unterstützung bei der praktischen Durchführung der Sequenzierarbeiten danken.

Meinen (Ex-)Mit-Doktoranden **Dr. Dominik Cholewa, Dr. Ram Shankar, Gabriele Kleiner-Grote, Dr. Ying Fei Shi, Dr. Jakob Müller** und **Dr. Philipp Grimm** ist ein guter Teil der großartigen Arbeitsatmosphäre auf „D5“ zu verdanken. Lustige Unterhaltungen in der Mensa, im Labor, im Büro oder auf dem Flur und das ein oder andere Späßchen ließen einen vergessen, dass man eigentlich noch ein Experiment am Laufen hatte. **Ram** und **Dominik** sei hier nochmal besonders für das Korrekturlesen dieser Arbeit gedankt!

Ich hatte die Gelegenheit mehrere Bacheloranden und Masteranden im Rahmen ihrer Abschlussarbeiten zu betreuen: **Christiane Widmann, Patrycja Topczewska, Tigran Karatyan, Jassin Hamidi, Maik Liedtke, Tobias Luttermann, Carina Wenzel** und **Ertuğ Yüzer.** Ihnen gilt mein Dank für die gemeinsame Arbeit, bei der auch ich einiges dazugelernt habe. Besonders hervorzuheben sind hierbei **Christiane Widmann** und **Tobias Luttermann**, die verrückt genug waren Bachelor- und Masterarbeit unter meiner Betreuung zu machen. Ich wünsche ihnen viel Erfolg für die Zukunft und **Tobi** zusätzlich viel Freude mit den Eigenheiten von *P. pastoris* und pMito!

Der gesamten AG Fermentationstechnik möchte ich für die schöne Zeit während der Doktorarbeit und auch schon davor danken. Aufgrund der kollegialen, hilfsbereiten und einfach freundlichen Mitarbeiter wie **Joe Max Risse, Thomas Schäffer, Eberhard Wünsch, Galina Beck** und **Kirsten Leiwig** ging ich immer gerne zur „Arbeit“ und fühlte mich dort auch sehr zuhause.

Auch der AG „Microbial Genomics and Biotechnology“ bin ich sehr dankbar für die fruchtbare Kooperation und große Hilfsbereitschaft, insbesondere von **Manuel Wittchen** und **Timo Wolf.**

Ohne die Finanzierung durch das Bundesland Nordrhein-Westfalen und den **"Cluster Industrial Biotechnology – Graduate Cluster" (CLIB-GC)** wäre das Projekt nicht zustande gekommen. Neben der finanziellen Unterstützung waren die verschiedenen Seminare, Workshops und Retreats eine gute Gelegenheit sich mit anderen Doktoranden auszutauschen und den Blick über den Tellerrand zu wagen. Ein großer Teil dieser Veranstaltungen ist dem organisatorischen Talent von **Dr. Iris Brune** zu verdanken.

Abseits der Arbeit in Labor und Büro sorgten **Freunde** und **Familie** dafür, die schönen Zeiten schöner und die schweren Zeiten leichter zu machen. Ich kann nicht genug für den Rückhalt und die gemeinsamen Aktivitäten Danken, die dafür sorgten, dass ich die Studien- und Doktorzeit nicht vergessen werde. Ein ganz besonderer Dank gilt meiner Freundin **Shoa Naqvi**. Ihre Stärke und Unterstützung hat mir ebenso geholfen, wie die Zeit miteinander, von der wir in Zukunft hoffentlich noch mehr gemeinsam verbringen werden können.

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbstständig und nur unter der Benutzung der angegebenen Quellen angefertigt zu haben. Alle aus der Literatur wörtlich oder sinngemäß entnommenen Zitate und Abbildungen habe ich als solche kenntlich gemacht. Weiterhin erkläre ich, dass diese Arbeit weder vollständig noch teilweise einer anderen Prüfungsbehörde mit dem Ziel vorgelegt wurde, einen akademischen Titel zu erwerben.

Bielefeld, 21.09.2017 _____

Jan-Philipp Schwarzhans