OPENPichia: building a free-to-operate *Komagataella phaffii* protein expression toolkit

Dries Van Herpe[1,2,3*], Robin Vanluchene[1,2*], Kristof Vandewalle[1,2], Sandrine Vanmarcke[1,2], Elise Wyseure[1,2], Berre Van Moer[1,2], Hannah Eeckhaut[1,2], Daria Fijalkowska[1,2], Hendrik Grootaert[1,2], Chiara Lonigro[1,2], Leander Meuris[1,2], Gitte Michielsen[1,2], Justine Naessens[1,2], Charlotte Roels[1,2], Loes van Schie[1,2], Riet De Rycke[4,5], Michiel De Bruyne[4,5], Peter Borghgraef[5], Katrien Claes[1,2**] and Nico Callewaert[1,2**]

[1]Unit for Medical Biotechnology, Center for Medical Biotechnology, VIB, Technologiepark 75, 9052 Ghent, Belgium.

[2]Department of Biochemistry and Microbiology, Ghent University, K.L.-Ledeganckstraat 35, 9000 Ghent, Belgium.

[3]Inbiose NV, Ghent, Belgium.

[4]Department of Biomedical Molecular Biology, Ghent University, Ghent, Belgium.

[5]BioImaging Core, VIB, Technologiepark 71, 9052 Ghent, Belgium.

[*] These authors contributed equally to this manuscript.

[**]Correspondence should be addressed to N.C. (Nico.Callewaert@vib-ugent.be) and K.C (Katrien.Claes@vib-ugent.be).

Abstract

In the standard toolkit for recombinant protein expression, the yeast known in biotechnology as *Pichia pastoris* (formally: *Komagataella phaffii*) takes up the position between *E. coli* and HEK293 or CHO mammalian cells, and is used by thousands of laboratories both in academia and industry. The organism is eukaryotic yet microbial, and grows to extremely high cell densities while secreting proteins into its fully defined growth medium, using very well established strong inducible or constitutive promoters. Many products made in *Pichia* are in the clinic and in industrial markets. *Pichia* is also a favoured host for the rapidly emerging area of 'precision fermentation' for the manufacturing of food proteins. However, the earliest steps in the development of the industrial strain (NRRL Y-11430/CBS 7435) that is used throughout the world were performed prior to 1985 in industry (Phillips Petroleum Company) and are not in the public domain. Moreover, despite the long expiry of associated patents, the patent deposit NRRL Y-11430/CBS 7435 that is the parent to all commonly used industrial strains, is not or no longer made freely available through the resp. culture collections. This situation is far from ideal for what is a major chassis for synthetic biology, as it generates concern that novel applications of the system are still encumbered by licensing requirements of the very basic strains. In the spirit of open science and freedom to operate for what is a key component of biotechnology, we set out to resolve this by using genome sequencing of type strains, reverse engineering where necessary, and comparative protein expression and strain characterisation studies. We find that the industrial strains derive from the *K. phaffii* type strain lineage deposited as 54-11.239 in the UC Davis Phaff Yeast Strain collection by Herman Phaff in 1954. This type strain has valid equivalent deposits that are replicated/derived from it in other yeast strain collections, incl. in ARS-NRRL NRRL YB-4290 (deposit also made by Herman Phaff) and NRRL Y-7556, CBS 2612 and NCYC 2543. We furthermore discovered that NRRL Y-11430 and its derivatives carry an ORF-truncating mutation in the *HOC1* cell wall synthesis gene, and that reverse engineering of a similar mutation in the NCYC 2543 type strain imparts the high transformability that is characteristic of the industrial strains. Uniquely, the NCYC 2543 type strain, which we propose to call 'OPENPichia' henceforth, is freely available from the NCYC culture collection, incl. resale and commercial production licenses at nominal annual licensing fees[1]. Furthermore, our not-for-profit research institute VIB has also acquired a resale/distribution license from NCYC, which we presently use to openly provide to end-users our genome-sequenced OPENPichia subclone strain and its derivatives, i.e., currently the highly transformable *hoc1*^{tr} and the *his4* auxotrophic mutants. To complement the OPENPichia platform, a fully synthetic modular gene expression vector building toolkit was developed, which is also openly distributed, for any purpose. We invite other researchers to contribute to our open science resource-building effort to establish a new unencumbered standard chassis for *Pichia* synthetic biology.

Introduction

Presently, a wide variety of microbial hosts is available for the production of recombinant proteins. However, for general laboratory use as well as biopharmaceutical protein manufacturing, a strong consolidation has taken place over the past years. *E. coli* remains the most-frequently used prokaryotic system for proteins of prokaryotic origin and for relatively simple stably folding proteins of eukaryotic origin, especially those with no or few disulphide bonds. On the other end, production in HEK293 or CHO cells is used for production of proteins of higher eukaryotic origin, affording *i.a.* complex-type N-glycosylation. The methylotrophic yeast known in biotechnology as *Pichia pastoris* (and formally classified as *Komagataella phaffii*) takes up the intermediate-complexity position in most protein expression lab's toolkit, as it combines the easy cultivation and fast growth of a micro-organism with the presence of a eukaryotic secretory system and the ensuing ability to perform complex post-translational modifications such as N-glycosylation and strong capacity for the formation and isomerization of disulphide bonds[2–4]. Hundreds of papers are published every year reporting on the use of *Pichia* for protein production (and increasingly also for engineered metabolite production, incl. in the area of sustainable chemical building block manufacturing from methanol and even CO_2[5]). Quite often, *Pichia*-produced proteins are developed by academic and industrial scientists alike with an eventual applied/commercial use in mind, being it as a therapeutic compound, an industrial biocatalyst[6,7] or more recently, as a food ingredient.

Historically, in 1954 Herman Phaff deposited a methylotrophic yeast strain from a black oak tree (*Quercus*) in the Yosemite region[8]. The isolate was stored in the culture collection of the University of California at Davis as UCD-FST K-239, with formally equivalent type strain deposits in other culture collections as NRRL YB-4290, NRRL Y-7556, CBS 2612 and NCYC 2543. At this time, these isolated strains of Phaff could not be distinguished from similar strains isolated in 1919 by A. Guilliermond, and Herman Phaff categorized both together as a new species: *Pichia pastoris* (the genus *Pichia* was established half a century before, in 1904, by E. Hanssen[9]) (Figure 1). In 1995, *Pichia pastoris* was re-classified into the new genus of *Komagataella*, named after the Japanese scientist Kazuo Komagata as a tribute to his contributions to yeast systematics, in particular the methanol-assimilating yeasts[10]. In 2005, the two distinctly evolved isolates from Phaff and Guilliermond were divided into two separate species and renamed *Komagataella phaffii* and *Komagataella pastoris* by C. Kurtzman[11], based on the sequencing of 26S rDNA. Consequently, the strain of Phaff (UCD-FST K-239, NRRL YB-4290, NRRL Y- 7556, CBS 2612, and NCYC 2543) was now considered the type strain of the species *Komagataella phaffii,* while the strain of Guillermond (CBS 704 or NRRL Y-1603) was regarded as the type strain of the species *Komagataella pastoris*.

90   By the 1970's, these yeast species that can utilize methanol as the sole carbon source[12–14]

91   sparked the interest of Phillips Petroleum Company, since they had a vast supply of cheap

92   methane gas (a by-product of their oil refinement process), which can be easily oxidized to

93   methanol by chemical oxidation. Hence, they explored the available methylotrophic yeast species

94   through procedures that are not in the public domain, and selected a *Pichia pastoris* to grow on

95   the synthesized methanol and produce a single cell protein source for animal feed using

96   fermentation. The application was patented in 1980 (with the earliest priority application on April

97   12th, 1979)[15], which required the strain to be deposited once more in public culture collections and

98   it became known as NRRL Y-11430 and CBS 7435. Which exact strain was deposited, is not

99   described in the patent, nor available in the public domain. It was already assumed that it could

100   be either the isolate from Guillermond (NRRL Y-1603) or that of H. Phaff (NRRL Y-4290)

101   (GRN 737). However, recent efforts of genomic sequencing show that the genetic differences

102   between both lineages are large enough to conclude that the latter was used[16,17].

103   In the early 80's, Phillips Petroleum Company contracted with the Salk Institute

104   Biotechnology/Industrial Associates (SIBIA) to develop the organism for recombinant protein

105   production, based primarily on the extremely strong and tightly regulated Alcohol Oxidase 1

106   promoter (pAOX1). In this context, NRRL Y-11430 derived strains were generated with

107   auxotrophies, such as the GS115 strain, which is a *his4* auxotrophic mutant obtained through

108   nitrosoguanidine mutagenesis of NRRL-Y11430[18], and the X33 strain that is a *HIS4*

109   complemented strain deriving from GS115, likely via an intermediate[18–20]. In 1993, Phillips

110   Petroleum sold its patent position in the *Pichia* system to Research Corporation Technologies

111   (RCT, Tucson, AZ)[21], apparently including the strain patent deposits associated to this, and their

112   derived strains. Unfortunately, NRRL Y-11430 is not distributed anymore, likely because the

113   patentee is no longer under an obligation to provide the strain, given that the associated patent(s)

114   have expired almost 20 years ago. Indeed, the NRRL Y-11430 in the Agricultural Research

115   Service culture collection (ARS-NRRL) is no longer available, and the same is the case for the

116   CBS 7435 strain in the collection of the Westerdijk Fungal Biodiversity Institute (CBS). Also, it

117   appears that the strain and its popular derivatives, incl. their distribution, are still controlled by the

118   original patent holder. This occurs through the enforcement of Material Transfer Agreements

119   (MTAs) that were associated to obtaining the strain from the NRRL culture collection during the

120   time it was maintained as a patent deposit, or conditions of sale associated to obtaining derivative

121   strains through the commercial distribution by Invitrogen (currently a brand of Thermo Fisher

122   Scientific) as part of *Pichia* expression system kits. To our knowledge, the NRRL Y-11430

123   parental industrial strain can currently only be obtained at the American Type Culture Collection

124   (ATCC 76273) under a similarly restrictive MTA. Because of the RCT-mandated distribution of

125   *Pichia* expression technology kits by Invitrogen, for more than two decades, thousands of protein

126 expression labs have extensive experience with the GS115 family of strains that were included in

127 these kits[22,23] and the currently approved biopharmaceuticals are manufactured by use of the

128 NRRL Y-11430 industrialized strain lineage. Many researchers are not fully aware of the

129 restrictions to strain distribution and potential applied use downstream of discoveries and

130 inventions made with them. Without prejudicing the legal aspects of some of this, which is outside

131 of our field of competence, as scientists we find it entirely unsatisfactory that the very basis of a

132 critical mainstay of academic and industrial biotechnology remains not simply openly accessible

133 exceedingly long after the original patents have expired. Given that *Pichia* takes such an ever

134 more prominent place in the toolkit for diverse sectors of synthetic biology and biotechnology, we

135 believe that it is long overdue for an open-access alternative strain platform to be set up by the

136 community. To continue to take benefit of the extensive regulatory agency acquaintance with

137 *Pichia*, it is important to achieve this by use of the same parental strain lineage as used in the

138 strains that were industrialized so far.

139 To achieve this goal, we and others have recently turned to the *Komagataella phaffii* type strains

140 that are present in culture collections throughout the world, armed with now highly affordable

141 genome resequencing, to try and identify the original isolate from nature that the Phillips

142 Petroleum researchers used in their derivation of NRRL Y-11430. In an exploration in the largest

143 US yeast strain collection (ARS-NRRL), the Love lab at MIT found that two deposits indicated by

144 the culture collection as equivalent *Komagataella phaffii* type strains (NRRL YB-4290, directly

145 deposited by Herman Phaff, and NRRL Y-7556), are genetically identical. Furthermore, in

146 comparison to this type strain, in their analysis they found only 1 point mutation leading to a

147 truncation of the Rsf2p open reading frame in NRRL Y-11430[24], establishing the type strain

148 lineage of this industrial strain. In this study, the NRRL type strains were found to have problems

149 with easily generating high copy number transformants and the authors recommended to keep

150 using the industrial NRRL Y-11430 strain background for recombinant protein production,

151 especially for expressions under methanol control. The NRRL YB-4290 strain is also present in

152 the UC Davis Phaff Yeast Culture Collection as its deposit made by Herman Phaff in 1954, and

153 numbered 54-11.239 (referred to as UCD-FST K-239 in the NRRL entries). It is documented in

154 the collections as having been isolated at Mather, Central Sierra Nevada, California, USA, from

155 the exudate flux of a black oak (*Quercus kelloggii*). Given its origin, no Access and Benefit Sharing

156 restrictions apply to its use for any purpose (Nagoya Protocol or Convention on Biological

157 Diversity). This type strain was also deposited in European culture collections: at the Westerdijk

158 Fungal Biodiversity Institute (CBS, Delft, The Netherlands, deposit CBS 2612) and at the National

159 Collection of Yeast Cultures (Norwich, UK, deposit NCYC 2543). In our present study, we started

160 out by analysing the genome of the four type strains. We then focused on the NCYC 2543 strain,

161 as the NCYC collection uniquely provides standard very affordable distribution and commercial

162   use licenses for its strains and derivatives thereof, fulfilling our requirements for a community

163   open-access *Pichia* platform strain. Different from all other culture collections where the *K. phaffii*

164   type strain deposits are found, technology developers can acquire a distribution license to end-

165   users for their newly generated materials, or have the choice of also depositing these in the NCYC

166   collection for full open access. Hence, we propose to nickname this NCYC 2543 type strain

167   deposit as 'OPENPichia'. In an effort to establish confidence in the OPENPichia strain as a new

168   standard open-access *Pichia* chassis for use by the research and biotechnological industry

169   community, we set out on a comprehensive characterization of its genome, growth characteristics,

170   transformation efficiency and recombinant protein expression, in comparison to the NRRL

171   Y-11430 industrial strain. We discovered a key frameshift mutation in the *HOC1* gene of the

172   industrial strain that enhances conduciveness to transformation, and introduced this mutation in

173   OPENPichia using open-source genome engineering technology, to overcome this important

174   limitation. Further characterization of this OPENPichia-*hoc1^{tr}* strain in comparison to NRRL

175   Y-11430 in terms of growth rate, cell wall density and cell sensitivity to cell envelope-binding

176   chemicals confirmed it to be indistinguishable phenotypically from the industrial strain. As set forth

177   below, we complement this OPENPichia strain set with an OPENPichia modular protein

178   expression vector toolkit, entirely made up of synthetic DNA to avoid any third-party MTAs, and

179   following the Golden Gate cloning standard for compatibility with complementary toolkits from

180   other *Pichia* developer labs[25]. The basic NCYC 2543 OPENPichia strain is available from NCYC,

181   incl. straightforward and low-cost distribution licenses for labs that develop novel strains. Our own

182   derived OPENPichia strains described in this paper are openly available for end-users (incl. for

183   royalty-free *Pichia*-made commercial product manufacturing) under such distribution license that

184   our not-for-profit research institution VIB obtained from NCYC. OPENPichia vector cloning

185   materials are openly distributed for any utilization purpose in association with the public plasmid

186   collection of the Belgian Coordinated Collection of Microorganisms.

187

188   **Results**

189   **The genomes of the *K. phaffii* type strains and commercial strains are virtually identical**

190   To evaluate alternative *K. phaffii* strains at the genomic level, we deeply (avg. 180x genome

191   coverage) resequenced the genome of several type strains available at culture collections (i.e.,

192   NRRL YB-4290 / NRRL Y-7556 / CBS 2612 / NCYC 2543) in comparison to the NRRL Y-11430

193   industrial strain (**Error! Reference source not found.**) (Supplementary Table 1 and 2). The reads

194   were mapped against the published reference genome of the CBS 7435 strain incl. the

195   mitochondrial genome sequence and that of the two *K. phaffii* linear killer plasmids[26]. CBS 7435

196   is a deposit of the NRRL Y-11430 industrial strain in the CBS culture collection (Delft, The

197   Netherlands). To this end, the Breseq[27] software package was used in consensus mode. There
198   was some variation in overall GC content, which can be attributed to the varying amount of
199   mitochondrial DNA and the presence of the two killer plasmids in some of the strains, as these
200   have an average GC content of 24%, 28% and 29%, respectively (unlike the nuclear
201   chromosomes, which have an average GC content of 41%). Indeed, the read alignment showed
202   that the proportion of reads that mapped to the mitochondrial genome varied between 10 and
203   47%.

204   The proportion of reads originating from the two killer plasmids, varied between 0% and 9%
205   (Supplementary Table 2). The *K. phaffii* killer plasmids are linear autonomously replicating DNA
206   fragments of 9.5 kb and 13.1 kb[19] that encode exotoxins that can kill other yeast cells[17,19]. These
207   toxins may also be toxic to cells from the same species, in case these have lost resistance to the
208   toxin, and this phenomenon is therefore undesired in large scale culturing of yeast as it may
209   reduce the overall viability of the culture. To verify the presence of such killer plasmids in the
210   tested strains, their sequences, as reported by Sturmberger *et. al.*, were included in the reference
211   for the alignment of the reads[26]. Surprisingly, the killer plasmids could not be detected in both the
212   CBS 2612 and NCYC 2543 strain, while such reads were abundant for the analysed NRRL strains
213   (NRRL YB-4290, Y-7556 and Y-11430) (Table 2). According to the strain history, the NRRL
214   YB-4290 and CBS 2612 strains were deposited by Phaff, while the NRRL Y-7556 strain was a
215   redeposit of CBS 2612 by D. Yarrow (CBS). Since the daughter strain (NRRL Y-7556), still has
216   these killer plasmids, the mother strain (CBS 2612) should have had them as well, leading to the
217   conclusion that these linear plasmids are easily lost while propagating *K. phaffii* strains. In our lab
218   as well, several NRRL Y-11430 descendants were identified that have lost these killer plasmids
219   upon the standard microbial practice of single colony streaking (unpublished data). Also, the
220   commercial strain GS115, does not have any killer plasmids[18].

221   Using a Maximum Likelihood method and the Hasegawa-Kishino-Yano model, a phylogenetic
222   tree based on the nuclear genome was computed to visualize the genetic distances between the
223   sequenced strains and a variety of other *K. phaffii* strains whose genomes were published
224   previously[18,24,28]. All *K. phaffii* type strains are strongly clustered together with the NRRL Y-11430
225   and CBS 7435 strains, as well as other close relatives (Figure 2). Thus, these data support
226   previous literature[24,29], where it was hypothesized that all these strains are derived from the same
227   isolate, originally isolated by Phaff[28].

228   Based on the resequencing data, we identified single nucleotide polymorphisms (SNPs) and short
229   insertion-deletions (indels) (Table 2). First comparing the mutation calling of strain NRRL Y-11430
230   with the reference genome of the equivalent-deposit CBS 7435 strain reference revealed only
231   one protein-coding difference in one hypothetical protein and about 20 intergenic/intronic/silent
232   exonic differences. All these alterations were also identified in the type strains, indicating that

7

233   these are likely to be considered as the wild type *K. phaffii* genotype and that the CBS 7435

234   reference genome was most likely miscalled at these sites. The same holds for two insertions that

235   were observed in each of the sequenced strains in a region annotated to encode for a hypothetical

236   papain-like cysteine protease. This region appears to have been difficult to sequence in previous

237   experiments, as also in the datasets associated with a previous type strain genome sequencing

238   study[24], we found only a few reads mapping to this area, which were accompanied by many

239   mutations, low quality bases (Phred scores of <28) and low overall mapping quality score (<20).

240   Note that the type strain deposits of the different culture collections also differ from one another

241   each at one other coding sequence-altering genomic position and a few non-coding ones, likely

242   reflecting drift due to background mutational rate during strain propagation (Table 2 and 3).

243   We then focused on the very few protein-coding alterations that consistently distinguish the

244   industrial strain NRRL Y-11430 from the equivalent type strain deposits NRRL YB-4290, NRRL

245   Y-7556, CBS 2612, and NCYC 2543. Three coding sequence altering mutations (in *SEF1*, *RSF2*,

246   and *HOC1*) were shared by all type strains but were absent in the NRRL Y-11430 strain. Our re-

247   analysis of the raw sequencing reads from a previous genome analysis of deposits NRRL

248   YB-4290 and NRRL Y-7556 confirmed the presence of these mutations and their absence in the

249   NRRL Y-11430 strain also based on these datasets[24]. Data quality in the area of the *SEF1* and

250   *HOC1* mutation is rather poor in these earlier datasets, and the authors only detected the mutation

251   in *RSF2*[24]. Since all three mutations are shared by the type strains, it must be concluded that they

252   represent the original state of *K. phaffii*, and that NRRL Y-11430 is actually the mutant strain at

253   these three loci.

254

Three mutations in protein-coding regions in the NRRL Y-11430/CBS 7435 industrial strain

255

lineage vs. the type strain deposits

256

257   *SEF1* encodes for a putative transcription factor (UniProt ID F2QV09) and the observed SNP

258   causes a S315C mutation in NRRL Y-11430 as compared to the type strains. *RSF2* encodes for

259   a transcription factor which is involved in methanol- and biotin-starvation (UniProt ID F2QW29).

260   The observed mutation is a SNP which causes the introduction of a stop codon (W748*) in NRRL

261   Y-11430, resulting in the deletion of 183 amino acids from the C-terminus of the protein. The non-

262   truncated Rsf2p which is found in the *K. phaffii* type strains, is very similar to its homolog in *S.*

263   *cerevisiae*[30], additionally indicating that this was the original genomic state, as previously

264   reported[24]. *HOC1* (homolog of *OCH1*) encodes for an α-1,6-mannosyltransferase (UniProt ID

265   F2QVW2) involved in the synthesis of cell wall mannan and is part of the M-PolII complex[31]. Here,

266   the industrial strain NRRL Y-11430 has a single base pair deletion in a poly-A stretch (at bp 755

267   of the 1191 bp long coding sequence), as compared to the type strains. This causes a frameshift

268   and premature stop codon, resulting in a C-terminally truncated protein (274 versus 398 amino

8

269   acids), with the last 22 amino acids up to the new stop codon being different from the type strain

270   sequence. The indel in the homopolymer was confirmed by Sanger sequencing (data not shown).

271   In parallel to our work, the same mutation was identified by the lab of Kenneth Wolfe (UC Dublin),

272   as the phenotype-causative mutation for a quantitative trait locus (QTL) that yielded 2-3 fold

273   higher secretion of a reported beta-glucosidase (personal communication, publication in press[29]).

274   With this knowledge on which mutations could be involved in phenotypic differences between the

275   type strains and the industrial strains, we set out for a comparison of characteristics important to

276   the use of *Pichia* for recombinant protein production. Given that the NCYC 2543 deposit of the

277   type strain was the only one for which the resp. culture collection openly provides both commercial

278   and re-sale/strain distribution licenses as part of its standard business practice[1], we further mainly

279   focused on this deposit's characteristics, and we called it OPENPichia.

280

281   **Strain comparison for growth rate, protein production and transformation efficiency**

282   NCYC 2543/OPENPichia was compared to NRRL Y-11430 and GS115, both in terms of growth

283   rate and their capacity for expressing recombinant proteins. GS115 is a *his4* auxotrophic mutant

284   derived from NRRL Y-11430 by nitrosoguanidine mutagenesis and, depending on the analysis,

285   its genome contains about 69[32], 74[24] or 71 (our unpublished data) mutations vs. that of its parent.

286   To be able to evaluated the impact of *his4*-mediated histidine auxotrophy on strain characteristics,

287   we generated an OPENPichia *his4* strain using the split-marker method[33,34]. No significant

288   difference in growth rate between the type strains and NRRL Y-11430 could be observed

289   (**Error! Reference source not found.**), but the GS115 strain grew significantly more slowly (one-

290   way ANOVA, p=0.0034, Tukey test), as reported[24]. Interestingly, our histidine auxotrophic

291   OPENPichia does not show the slower growth phenotype, demonstrating that this GS115

292   phenotype is not due to histidine auxotrophy, but rather must be due to one or more of the other

293   nitrosoguanidine-induced mutations in its genome. As *his4* complementation is an often-used

294   antibiotic-free selection marker, this unaffected growth rate in the OPENPichia *his4* strain is a

295   useful feature.

296

297   To compare the protein production capacity of the strains, a selection of model proteins was

298   produced in NRRL Y-11430 and OPENPichia (

299   Table). Four proteins of very different types in use in biotechnology were chosen: a cytokine (GM-

300   CSF), a redox enzyme (GaOx), a VHH-hFcα fusion (Cdiff-VHH-IgA), and a VHH-hFcγ fusion

301   (CovidVHH-IgG). To enable recombinant expression, the two most commonly used off-patent *K.*

302   *phaffii* promoters were tested: pGAP for strong constitutive and pAOX1 for strong methanol-

303   inducible expression, respectively. Protein expression in *K. phaffii* is prone to clonal variations

304   that can interfere with the comparison of expression capabilities between strains. Most of the

305   variation is due to the integration site and the copy number of the construct[35]. To this end, a single-

306   copy was targeted to the respective promoter regions in the genome, the copy number and

307   integration site were confirmed by qPCR and integration-site specific PCR, and two independent

308   clones of each setup were grown in triplicate.

309   For both the pGAP-driven and pAOX1-driven expressions, there is in general no major difference

310   in protein production titres between the two hosts. However, there is a clear difference in cell

311   density at harvest of the pGAP cultures, with OPENPichia growing to higher densities than NRRL

312   Y-11430 (Supplementary Figure 1), whereas this is not the case on methanol. Additionally, in all

313   cases, NRRL Y-11430 shows slightly more host cell proteins (HCPs) in the medium of the pGAP

314   expressions (on limiting glucose), as visible on the SDS-PAGE gels (Figure 4), but not when

315   grown on methanol. Both observations point towards a minimally increased cell lysis of the NRRL

316   Y-11430 strain vs. the OPENPichia strain when grown on glucose.

317

318   **_HOC1_-truncation restores the transformation efficiency in NCYC 2543**

319   Upon generating expression clones, a strongly reduced transformation efficiency was observed

320   for OPENPichia as compared to NRRL Y-11430 (Figure 5C), as was also observed by others[24].

321   Out of the three consistent protein-coding differences between the type strains and NRRL

322   Y-11430, we reasoned that the one in the _HOC1_ open reading frame could be causative for the

323   low transformation efficiency of the type strains, as the _S. cerevisiae_ Hoc1p ortholog is an

324   α-1,6-mannosyltransferase that is part of the mannan polymerase II complex in the yeast Golgi

325   apparatus. The hypermannosyl-N-glycans, of which the backbone is synthesized by this mannan

326   polymerase complex, form the outermost layer of the ascomycete cell wall, and carry most of their

327   charge under the form of mannosylphosphate modifications on the side chains. _HOC1_ deletion in

328   _S. cerevisiae_ is viable though the phenotyping and genetic interaction data available in the

329   _Saccharomyces_ Genome Database clearly indicate a cell wall stress response. As negatively

330   charged plasmid DNA during transformation has to traverse the cell wall, a cell wall that presents

331   less of a diffusional/charge barrier due to less mannan/mannosylphosphate density could explain

332   the more highly transformable phenotype of NRRL Y-11430. Using the split-marker method, we

333   introduced this single base pair deletion in the genome of OPENPichia, hence reverse

334   engineering the genetic makeup of the NRRL Y-11430 strain (Figure 5A) in this locus. Two mutant

335   versions were generated: NCYC 2543 _hoc1^{tr}_-1, where only the single base pair was deleted,

336   resulting in the truncated ORF and a Lox72-scar downstream of the novel stop codon; and NCYC

337   2543 _hoc1^{tr}_-2, where additionally 115 bp downstream of the novel stop codon were removed

338   (Figure 5B). The transformation efficiency was compared between the two wild type strains and

339   the _hoc1^{tr}_ mutants, using a plasmid encoding for a VHH under control of the _GAP_ or _AOX1_

340   promoter. The _HOC1_-truncation drastically increased the transformation efficiency of the

10

341   OPENPichia strain, and even showed a 2-3-fold improvement as compared to the NRRL Y-11430

342   in this experiment (Figure 5C). In general, it was also observed that the transformation efficiency

343   for pGAP-based plasmids is lower as compared to similar plasmids that are pAOX1-based, which

344   we speculate is due to the metabolic burden imposed by constitutive GBP VHH production during

345   recovery of the cells after transformation.

346

Characterization of the NRRL Y-11430, NCYC 2543 and NCYC 2543 _hoc1tr_ cell wall

348   These strains were further characterized in terms of their cell wall composition. First, the cell wall

349   mannoprotein N-glycans were profiled by capillary electrophoresis[36] after the cells were grown on

350   glucose or glycerol (Supplementary Figure 2). This method mainly detects the lower-degree of

351   polymerization N-glycans, and the profiles were very similar for all four strains, indicating that the

352   pathway of synthesis of the mannan core was intact in all strains. This capillary electrophoresis

353   method is however unsuited to the detailed profiling of the higher-polymerized mannan N-glycans,

354   as there are so many isomeric structures formed that are all in part or completely resolved, such

355   that they form one long trail of overlapping low-abundance peaks that is impossible to interpret.

356   Most of the mannosylphosphate negative charge-imparting moieties are added to the mannan

357   side branches of these long chains. They bind to cationic dyes such as Alcian Blue and hence

358   the staining density of yeast cells with such dye reflects the density of these negative charges on

359   the outermost layer of the cell wall. Indeed, as compared to the type strain NCYC

360   2543/OPENPichia, both NRRL Y-11430 and the two OPENPichia _hoc1tr_ mutants showed a

361   reduced Alcian Blue staining intensity (Figure 6A). This is consistent with the reduced Alcian Blue

362   staining of _S. cerevisiae hoc1_ mutants[37,38].

363   To further investigate the cell wall integrity, resistance of the strains towards Congo red and

364   Calcofluor white was analysed (Figure 6B) using previously reported methods[24,29]. The

365   OPENPichia type strain is much more resistant than NRRL Y-11430 towards both dyes, a

366   difference which is lost in the _HOC1_-truncated OPENPichia mutants. This indicates the

367   importance of Hoc1p in cell wall integrity. We also performed transmission electron microscopy

368   (TEM) using a freeze substitution technique that is optimized to pull in the osmium tetroxide

369   membrane-staining contrast reagent as well as fixatives through the cell wall during the slow

370   dehydration of the cells, which is then reversed in subsequent sample preparation. We observe

371   a much stronger electron scattering at the outermost cell wall layer of the wild type OPENPichia

372   type strain than for the other, _HOC1_-truncated strains (NRRL Y-11430 and OPENPichia _hoc1tr_)

373   (Figure 6C). We interpret that this is caused by OsO_4-accumulation at the mannan layer of the

374   cell wall due to a stronger barrier to diffusion of the reagent in the wild type strain during freeze

375   substitution. Scanning electron microscopy looked very similar for the four strains (Figure 6C),

376   indicating no gross malformations. In summary, the data are consistent with the _hoc1tr_ mutation

377   resulting in a relatively mild cell wall integrity deficiency, resulting in increased passage of plasmid

378   DNA during transformation and, depending on the target protein, in some cases also increased

379   production or cell wall passage of secreted recombinant proteins[29].

380

381   **Strain comparison for growth rate and protein production of OPENPichia *hoc1^tr***

382   The growth rate, as well as protein production capacity of the newly generated OPENPichia

383   *HOC1*-truncated strains were compared to those of NRRL Y-11430 and wild type OPENPichia.

384   No significant difference in growth rate is observed between the 4 tested strains (Figure 7A). As

385   we earlier found most phenotypic difference between NRRL Y-11430 and wild type OPENPichia

386   in terms of growth and HCP levels when the strains were grown on glucose, we tested pGAP-

387   based protein expression for the different strains (Table 4), using an anti-GFP VHH (GBP) as test

388   protein. To this end, 24 clones of each strain were screened, except for wild type OPENPichia,

389   where only 12 transformants were obtained (Figure 7B and C). For pGAP-based GBP expression,

390   the NCYC 2543 *hoc1^tr* strains outperform the NCYC 2543 strain, and NCYC 2543 *hoc1^tr*-1 even

391   outperforms NRRL Y-11430, although the differences are small and clonal distributions overlap.

392   The result is in line with the observations made by the Wolfe lab (personal communication,

393   publication in press[29]), where they observed that the truncated *HOC1* genotype in NRRL Y-11430

394   resulted in doubling of the secretion level of a beta-glucosidase under control of the pGAP

395   promoter, vs. rather divergent *K. phaffii* NRRL strains.

396   The presented data together with the growth and expression analysis, show that the NCYC 2543

397   background is at least as good as an expression host as the industrial NRRL Y-11430 strain (or

398   its derivatives). With the discovery of the *HOC1* truncation as the basis for the higher

399   transformability of NRRL Y-11430, and the finding that it can easily be accomplished using free-

400   to-operate recombination-based genome engineering, also this last remaining handicap of the

401   parental NCYC 2543 was removed. Indeed, as reported recently by Brady *et al.*[24], the low

402   transformability, which makes it laborious to find clones with multicopy integration of the

403   expression vector, was the key reason for them to decide on continued use of NRRL Y-11430-

404   based strains over more wild type strains. This problem is now solved.

405   As no difference between the two NCYC 2543 *HOC1*-mutant versions could be observed and

406   given their excellent performance, we decided to continue all future *Pichia* work with the version

407   *hoc1^tr-1*, which has the smallest genomic scar downstream of the truncated *HOC1* coding

408   sequence (i.e., insertion of a remaining Lox72 site).

409

410   **OPENPichia modular protein expression vector construction toolkit**

411   Many scientists have used or still use commercial *K. phaffii* expression kits, which conveniently

412   match expression strains and vectors. Indeed, constraints in vector choice are sometimes

413 imposed by the properties of the chosen *K. phaffii* strain (e.g., selection markers, auxotrophies,
414 methanol utilization phenotypes, etc). Although such kits are attractive to beginner users and are
415 hence widespread within the scientific community, their conditions of sale are legally restrictive
416 and forbid further distribution and reutilization of vector parts, let alone use in commercial
417 production, which requires commercial licensing from the provider. Fortunately, issues with non-
418 FTO DNA constructs can nowadays be avoided by *de novo* synthesis, combined with novel and
419 fast cloning techniques which both emerged rapidly in the last decade[39]. However, establishing a
420 new properly documented FTO genetic toolkit is still a relatively expensive and time-consuming
421 occupation for most labs.

422 To enable scientists to express their genes of interest, a well-equipped genetic toolkit and
423 corresponding cloning framework is provided to the community (Figure 8), and deposited at the
424 BCCM plasmid collection. The cloning system that was chosen is modular, to enable maximum
425 flexibility and based on Golden Gate cloning, similar to other toolkits[25,40–48]. These cloning systems
426 have gained a lot of popularity as they are user-friendly and easily expandable, while also boasting
427 high versatility in construct design. In essence, the strength of Golden Gate assembly is based
428 on the use of Type IIS restriction endonucleases that cut outside their recognition sites, which
429 allows users to flank DNA fragments of interest with customizable 4 nt overhangs. As such, a
430 4 nt overhang of one fragment can be made complementary to a 4 nt overhang of another
431 fragment, and such compatible overhangs are ligated much more efficiently, enabling directional,
432 multi-insert cloning in a single reaction. The MoClo system takes this concept a step further, as it
433 standardizes Golden Gate assembly by designating *a priori* all DNA elements of a desired vector,
434 which are typically referred to as 'parts', to a particular 'part type' (e.g., promoter, CDS, etc) and
435 flanking each part type by unique 4 nt overhangs and Type IIS restriction sites[48]. In practice, parts
436 are derived from PCR fragments or synthetic constructs, which are first subcloned in entry vectors,
437 also known as 'Level 0' vectors (Figure 8). As such, a collection of sequence-verified Level 0
438 vectors is established and vectors of interest can then be assembled into expression vectors of
439 interest, which are termed 'Level 1' vectors. By providing proper connector sequences with
440 additional Type IIS restriction sites, the resulting expression vectors or Level 1 vectors can then
441 be assembled again to obtain multigene or Level 2 vectors, which is the top level in the system's
442 hierarchy. In the current toolkit, all 4 nt overhangs were adopted to ensure a high degree of
443 compatibility with existing yeast toolkits[25,41,45] and to ensure a near 100% predicted ligation
444 fidelity[49]. Since this toolkit is essentially derived from the *S. cerevisiae* MoClo system, it shares
445 the restriction enzymes (BsmBI and BsaI), most of the 4 nt overhangs, and the number and design
446 of the individual part types[41]. As such, the system is comprised of eight part types, of which
447 Part 3 (Coding Sequence) and Part 4 (Terminator) can still be split up to allow additional
448 modularity, for example to incorporate N- and C-terminal fusion partners for the protein of interest

13

449   (Figure 8). An overview of the part types and the parts that are provided in this OPENPichia toolkit

450   is included (Supplementary Figure 3). Part sequences are available in Supplementary Information

451   and materials can be obtained from the BCCM GeneCorner plasmid collection[50]. The Material

452   Transfer Agreement associated with it was custom-designed in collaboration with GeneCorner to

453   allow for any use of resulting plasmids, including royalty-free commercial manufacturing.

454

Discussion

456   *Pichia pastoris (*formally known as *K. phaffii*) is an important protein production host in both

457   academia and industry, but the most common industrially developed strains are currently

458   distributed with restrictive MTAs, or not any longer. To facilitate academic and commercial host

459   strain development for recombinant protein expression and easy distribution throughout the

460   biotechnology community, alternative *K. phaffii* strains were investigated. We put forward an

461   open-access wild type strain, i.e., NCYC 2543 or OPENPichia, as well as two derivatives: a

462   histidine-auxotrophic strain and a *HOC1*-truncated strain with an improved transformation

463   efficiency. Moreover, a compatible genetic engineering toolkit is made available, which contains

464   all the necessary components for the expression of recombinant proteins and which can be easily

465   expanded with more genetic parts to fit the researcher's needs. This toolkit can be obtained with

466   an open-usage MTA from the GeneCorner plasmid collection, but it can also be assembled from

467   scratch, based on the sequences provided in the Supplementary Files 2.

468   It was shown that the proposed alternative OPENPichia strain (NCYC 2543) and its derivatives

469   are almost identical to the common NRRL Y-11430 strain. Only a handful of mutations could be

470   identified in our direct comparative genome analysis, of which only 4 are protein coding-altering

471   (SNPs and indels). Additionally, OPENPichia does not contain the undesired killer plasmids and

472   the strain shows the same maximum growth rate under the tested conditions. With respect to the

473   protein production capacity, the data presented here demonstrate that small differences can occur

474   between the *K. phaffii* type strain NCYC 2543/OPENPichia and NRRL Y-11430, but that there is

475   no consistently better performing strain, considering the variety of proteins tested in this study.

476   Previously, Brady et al[24], performed a similar experiment where NRRL Y-11430 showed to have

477   the highest protein expression level as compared to a series of other *K. phaffii* strains. However,

478   none of the type strains from which NRRL Y-11430 directly derives were included in this study;

479   instead, Y-12729 and Y-48124 (amongst others) were included, which are all members of Cluster

480   1 of their transcriptomics experiment. Y-7556 and YB-4290, which are type strains like NCYC

481   2543, are members of the transcriptomics Cluster 2, and would have allowed a better comparison.

482   Due to the increased cell wall robustness and reduced transformation efficiencies of the type

483   strains, they were excluded from the protein expression comparison in that study. Indeed, we also

14

484   observed that the transformation efficiency of the NCYC 2543 strain is dramatically lower as
485   compared to NRRL Y-11430. However, we could completely overcome this after our discovery
486   by more in-depth genome resequencing that NRRL Y-11430 has a truncating frameshift mutation
487   in the *HOC1* cell wall synthesis gene. We introduced the same frameshift mutations in *HOC1* of
488   OPENPichia, resulting in an even improved transformation efficiency as compared to NRRL
489   Y-11430. Hoc1p is part of one of two Golgi mannan polymerase complexes that in *S. cerevisiae*
490   also contains Anp1p, Mnn9p, Mnn10p and Mnn11p[51], and which mediates elongation of the
491   α-1,6-mannan backbone that is initiated by the activity of Och1p. In *S. cerevisiae*, *HOC1* has
492   synthetic positive genetic interactions with the *PKC1* pathway that mediates responses to cell wall
493   stress, as well as genetic interactions with a multitude of genes involved in cell wall integrity,
494   protein secretion, vesicular transport, all key pathways in the yeast cell wall maintenance[52].
495   Indeed, we observed a lower cell wall mannoprotein N-glycan mannosylphosphorylation density,
496   which is a sensitive hallmark of cell wall stress. The truncated Hoc1p that is produced in NRRL
497   Y-11430 vs. the type strain still has the alpha-helicoidal N-terminal luminal protein regions that
498   type II Golgi proteins typically use to space their catalytic domains away from the membrane, and
499   to interact with other Golgi proteins of similar topology. As Hoc1p forms part of such multi-protein
500   complex, we hypothesize that the truncated *Hoc1* allele maintains assembly of this complex but
501   lacks its own catalytic activity (see AlphaFold 2 model of the protein in Supplementary Figure 4).
502   In this way, likely a milder phenotype is obtained than with full *HOC1* deletion, making the *hoc1^tr^*
503   strains grow equally well as the wild type.

504   Interestingly, under pGAP expression conditions, NRRL Y-11430 has somewhat more HCPs in
505   its culture supernatant and grows to a lower cell culture density as compared to OPENPichia. We
506   hypothesize that both observations are related and due to slightly increased cell lysis in NRRL
507   Y-11430, which can have an impact on the need for additional purification steps. Whether the
508   *hoc1^tr^* mutation in OPENPichia has the same effect remains to be determined.

509   The *K. phaffii* strain that is proposed here as OPENPichia is one deposit of the type strain which
510   is widely available in many culture collections in different countries (see Global Catalogue of
511   Microorganisms)[53]. For instance, strain CBS 2612 also does not have killer plasmids and is
512   identical but for a few drift mutations. If a *Pichia*-user is an end-user and wishes to merely
513   manufacture a product in these type strains, multiple culture collections provide cost-effective
514   licenses. However, for labs who are developing improved *Pichia* technology and implement these
515   novel inventions in the strains, strains have to be distributable to other laboratories, and this is
516   most often not allowed by the MTAs even of the public strain collections. NCYC should be
517   commended for uniquely transparently providing cost-effective resale/redistribution as well as

518 commercial manufacturing use licenses as part of standard culture collection practice, fulfilling an
519 essential need for yeast-based technology developers.

520 More broadly, our study illustrates the need to build 'generic' biotechnological platforms after
521 patents on these foundational inventions of our field expire, much in the same way as
522 'generic/biosimilar' medicines need to be developed to increase access to more affordable
523 medicines. We have previously also accomplished this for the HEK293 cell lineage that is used
524 for viral vector and vaccine manufacturing and hope that others will join us in such open science
525 endeavours for other synthetic biology 'chassis' systems[54]. For now, we invite all *Pichia*
526 researchers and users to contribute to this OPENPichia resource and make best use of it.

527

528 **Materials & Methods**

529 **Strains and Media**

530 The wild type *K. phaffii* strains NRRL YB-4290, NRRL Y-7556 and NRRL Y-11430 were obtained
531 from the Agricultural Research Service (ARS, USA), CBS 2612 was obtained from Westerdijk
532 Institute (Netherlands) and NCYC 2543 was obtained from the National Collection of Yeast
533 Cultures (NCYC, UK). All mentioned strains were cultured and maintained on YPD or YPD agar.

534 All entry vectors and expression vectors were propagated and are available in the *E. coli* DH5α
535 strain. MC1061 and MC1061λ strains were used successfully as well and generally showed
536 higher transformation efficiency and easier green-white or red-white screening than was the case
537 for DH5α. All *E. coli* strains were cultured and maintained on LB agar.

538 Antibiotics were used in the following concentrations for selection in *E. coli*: Zeocin 50 µg/ml,
539 Nourseothricin 50 µg/ml, Hygromycin 50 µg/ml, Kanamycin 50 µg/ml, Chloramphenicol 50 µg/ml
540 and Carbenicillin 50 µg/ml. Antibiotics were used in the following concentrations for selection in
541 *Pichia*: Zeocin 100 µg/ml, Nourseothricin 100 µg/ml, Hygromycin 100 µg/ml, Geneticin 100 µg/ml,
542 Blasticidin 100 µg/ml.

543 Several media were used: LB (1% Tryptone, 0.5% Yeast Extract, 0.5% NaCl), YPD (1% Yeast
544 Extract, 2% Peptone, 2% D-glucose), YPG (1% Yeast Extract, 2% Peptone, 1% Glycerol), BMY
545 (1% Yeast Extract, 2% Peptone, 1,34% YNB without amino acids, 100 mM Potassium phosphate
546 buffer pH6), BMGY (BMY with 1% Glycerol), BMDY (BMY with 2% D-glucose), BMMY (BMY with
547 1% Methanol), and limiting glucose (1% Yeast Extract, 2% Peptone, 100mM Potassium
548 phosphate buffer pH6, 50g/l Enpresso EnPump substrate, 5ml/l Enpresso EnPump enzyme
549 solution (Enpresso GmbH, Germany)). For plates, 1.5% agar was added for LB media and 2%
550 for YPD media; when Zeocin selection was used, media were set to pH7.5.

551   All oligonucleotides and synthetic DNA fragments were ordered at Integrated DNA Technologies

552   (IDT), Leuven, Belgium. All synthetic DNA fragments (gBlocks® and Genes®) were designed and

553   adapted for synthesis using the Codon Optimization Tool and the gBlocks Gene Fragments Entry

554   Tool available at the website of IDT Europe.

555   **Illumina sequencing**

556   The strains were cultured overnight in YPD medium and the genomic DNA was extracted using

557   the Epicentre MasterPure™ Yeast DNA Purification Kit. Sample preparation (DNA fragmentation,

558   adapter ligation, size selection and amplification) and next generation sequencing (5M 150bp

559   paired end reads) was done by Eurofins, using Illumina technology. The reads were checked for

560   quality using fastqc[55], from which the %GC and number of reads was obtained. From the number

561   of   reads,   the   average   overall   coverage   was   calculated   with   the   formula

562    $\frac{\#reads \times read\ lengt\ \ (in\ bp)}{leng\ \ of\ genomic\ DNA\ + mitochondrial\ DNA\ (in\ bp)}$.

563   *NGS analysis*

564   The reads were trimmed using Trimmomatic[56] to remove adapter, leading and trailing low quality

565   bases (cut off quality 3), low quality reads (4-base sliding window quality below 15) and reads

566   below 100 bp. Next, the reads were aligned to a reference and the mutations were identified using

567   Breseq[27] in consensus mode. As reference, the genome sequence published by Sturmberger *et*.

568   *al*.[26] was used. The reference sequences for killer plasmids and the mitochondrial DNA were

569   obtained from Sturmberger *et*. *al*.[26] and Brady *et*. *al*.[17], respectively. The reported coverage depth

570   was calculated by the Breseq algorithm. This is done by fitting a negative binomial distribution to

571   the read coverage depth observed at unique reference positions. The mean of this binomial fit is

572   used as the coverage depth. The killer plasmid copy number was estimated by comparing their

573   coverage depth with the average of the four chromosomes. The coverage depth for each molecule

574   was calculated as the mean of a binomial fit for the coverage depth for each reference position.

575   *Phylogenetic tree*

576   In order to make a phylogenetic tree, the sequencing data from this study was combined with the

577   raw reads that were published before[24] and also aligned as described above. From the predicted

578   mutations of both datasets, a whole genome alignment was constructed from which a

579   phylogenetic tree was calculated using the Mega X[57] software package. A maximum likelihood

580   algorithm was used with an HKY substitution matrix.

581   **Creation of the NCYC2543 *his4* strain**

582   The NCYC2543 *his4* strain was generated using the split-marker method that was described

583   previously by Heiss *et al.*[58]. The homology arms of the *HIS4* gene were selected from Näätsaari

17

584   *et al.*[59], and the reference genome of the CBS7435 strain. First, a construct containing the two
585   homology arms with a floxed Nourseothricin acetyltransferase marker was created. Next two
586   overlapping fragments containing one of the homologies and a part of the antibiotic marker were
587   generated by PCR using Taq polymerase (Promega), which overlap for a length of 594 bp. These
588   fragments were purified by a phenol chloroform precipitation. In brief, after adding an equal
589   volume of phenol:chloroform:isoamyl alcohol (25:24:1), the solution was mixed, centrifuged
590   (5 min at 12,000 g) and the liquid phase was isolated by decanting. To this, 1/10th volume of 3 M
591   sodium acetate pH 5.5 and 2 volumes of 100% ethanol was added and the sample was mixed
592   and centrifuged (15 min at 12,000 g). Afterwards, the pellet, containing the amplified DNA, was
593   isolated, washed with 70% ethanol, air-dried, and resuspended in water.

594   Both purified fragments were transformed into NCYC 2543 competent cells by electroporation
595   and transformants were streaked to single clone onto YPD plates containing Nourseothricin and
596   grown for 2 days at room temperature. The resulting clones were replica plated onto CSM-his
597   plates for growth screening and grown for 2 days at room temperature. Strict non-growers were
598   checked by colony PCR for replacement of the *HIS*4 gene with the antibiotic marker cassette.

599   The Nourseothricin acetyltransferase marker was finally removed by transient expression of a
600   Cre-recombinase. This gene was cloned into a plasmid with an ARS[60] and a Zeocin resistance
601   cassette, which was then transformed in the *his4* strain. Transformants were incubated overnight
602   on a YPD plate containing Zeocin and colonies were transferred to YPD plates without antibiotics.
603   The removal of the antibiotic cassettes of the plasmid and *HIS4* knock-out was verified with replica
604   plating on YPD containing the respective antibiotics and double-checked via colony PCR.

605   **Creation of the NCYC2543 *hoc1^tr* strains**

606   The NCYC2543 *hoc1^tr* strains were generated using the split-marker method as described above.
607   The left homology arm of the *HOC1* gene was chosen such as to contain about 1 kb upstream of
608   the premature stop codon. To introduce the single nucleotide deletion, genomic DNA of NRRL
609   Y-11430 instead of NCYC 2543 was used as the PCR template. The right homology arm was
610   chosen as to contain about 1 kb downstream of the premature stop codon, also for this PCR,
611   genomic DNA of NRRL Y-11430 was used, although NCYC 2543 would have also worked. The
612   left and right homology arms were respectively fused by PCR to the first and last two thirds of the
613   floxed Nourseothricin acetyltransferase marker. The PCR fragments were purified over gel and
614   the DNA was recovered using the Wizard SV Gel and PCR Clean-Up System (Promega),
615   according to the manufacturer's instructions. Both purified fragments were transformed into NCYC
616   2543 competent cells by electroporation and transformants were streaked to single clone onto
617   YPD plates containing Nourseothricin and grown for 2 days at room temperature. The resulting
618   clones were screened using colony PCR using a forward primer annealing upstream of the left

619 homology arm and a reverse primer annealing to the Nourseothricin selection marker. The
620 Nourseothricin acetyltransferase marker was removed by transient expression of a Cre-
621 recombinase as described above. The engineered *HOC1* locus was confirmed for both strategies
622 by colony PCR and Sanger sequencing. The sequences for the PCR primers and split-marker
623 cassettes can be found in Supplementary Tables 3 and 4.

624 **Growth Analysis**

625 The different *Pichia* strains were cultured on YPD agar for 2 days, inoculated in triplicate into a
626 5 ml preculture of BMDY and grown overnight at 28 °C, shaking at 225 rpm. The optical density
627 at 600 nm (OD_{600}) of each culture was measured and 250 ml of BMDY was inoculated at a starting
628 OD_{600} of 0.05. Samples of 1 ml were immediately isolated from each culture to measure the OD_{600}
629 again. Then, samples of 1 ml were isolated every 2 h for 22 hours and again after 26 hours and
630 29 hours. All samples were diluted accordingly and measured within an OD_{600} range of 0.05 – 1.

631 **Recombinant protein expressions**

632 The expression vectors were made using the MoClo toolkit, based on Golden Gate cloning as
633 described in this paper. Briefly, the protein coding sequences were ordered synthetically with Part
634 3b type BsaI overhangs (NEB R3733) and cloned into the entry vector with BsmBI (NEB R0739).
635 Next, expression vectors were made by assembly of the Level 0 parts.

636 The cloning procedure was as follows: 1 µl of T4 DNA Ligase (400 U; NEB M0202), 2 µl of T4
637 DNA Ligase Buffer (NEB M0202), 1 µl of Restriction Enzyme (20 U) were added to 20 fmoles of
638 backbone (pPTK081 for entry vectors; any P8 backbone for destination vectors). An excess of
639 insert (>1000 fmoles of PCR amplicon or synthetic gene; 10 pmoles of annealed oligonucleotides)
640 was added for a BsmBI assembly, while equimolar amounts (20 fmoles) of each entry vector were
641 added for a BsaI assembly. BsmBI assembly mixtures were incubated according to the following
642 protocol: >25 cycles of 42 °C for 2 min (digest) and 16 °C for 5 min (ligation), followed by 60 °C
643 for 10 min (final digest) and 80 °C for 10 min (heat inactivation step). BsaI assembly mixtures
644 were incubated similarly, except that the digestion steps were performed at 37 °C.

645 *Pichia* electrocompetent cells were generated, using the lithium acetate method as described by
646 Wu *et al.*[61]. In brief, precultures were inoculated in 5 ml YPD and grown overnight in an incubator
647 at 28 °C and 250 rpm. The precultures were diluted and grown to an OD_{600} of approximately 1.5.
648 50 ml of the culture was isolated and the cells were harvested by centrifugation (1,519 g for 5 min
649 at 4 °C), resuspended in 200 ml of a lithium acetate (LiAc)/dithiothreitol (DTT) solution (100 mM
650 LiAc, 10 mM DTT, 0.6 M sorbitol, 10 mM Tris-HCl pH 7.5) and incubated for 30 minutes at 28 °C
651 and 100 rpm. Next, the cells were collected by centrifugation (1,519 g for 5 min at 4 °C), washed
652 two times with 1 M ice-cold sorbitol and finally resuspended in 1.875 ml of 1 M ice-cold sorbitol.

19

653   0.5 to 1 µg of DNA was added to aliquots of 80 µl and electroshocked (1.5kV, 200Ω, 25µF).

654   Immediately, 1 ml of 1 M sorbitol was added and the suspension was incubated at 28 °C for

655   2-5 h. Next, the cells were plated on YPD agar containing the appropriate antibiotic, and colonies

656   were isolated after 2 days of incubation at 30 °C.

657   To be able to compare expression, only colonies with single copy integration of the construct were

658   selected. The copy number was determined by quantitative PCR on a Lightcycler 480 (Roche)

659   using primers that bind pAOX1 and pGAP. The genes *OCH1* and *ALG9* were used as reference.

660   Genomic DNA (gDNA) of NCYC 2543 was included as a single copy positive control. A single

661   copy plasmid integration will yield one additional copy and more than two copies would be the

662   result of multiple plasmid integrations. Amplification efficiencies were determined using serial

663   dilutions of gDNA samples. Reactions were set up in 10 µl with final concentrations of 300 nM

664   forward primer, 300 nM reverse primer, 1x SensiFast SYBR no-rox mastermix (Bioline), 10 ng

665   gDNA and the following cycling conditions: 3 min at 95 °C, followed by 45 cycles of 95 °C for 3 s,

666   60 °C for 30 s at ramp rate 2.5 °C/s, 72 °C for 1 s, ending with 0.11 °C/s from 65 °C to 95 °C for

667   melting curve determination (5 acquisitions/s). Copy numbers were calculated using the

668   ΔΔCt method[62].

669   The different strains expressing the recombinant proteins, were cultured on YPD agar plates for

670   2 days, inoculated in triplicate into a 5 ml preculture of BMDY and grown overnight at 28 °C,

671   shaking at 225 rpm. Next, the cultures for pAOX1-driven expression, were inoculated in BMDY,

672   grown for 24 h, subsequently transferred to BMMY and incubated for 48 h. After 24 h in BMMY,

673   an extra 1% of methanol was added. The cultures for pGAP-driven expression, were instead

674   inoculated in limiting glucose medium and incubated for 48 h. Then, optical density at 600nm was

675   measured for all cultures and the supernatant was collected by centrifuging (2,500 g for 5 min).

676   The samples were incubated with EndoH (produced in-house) to remove N-glycans and analysed

677   by SDS-PAGE.

678   **ELISA-based quantification of GBP**

679   Nunc MaxiSorp™ 96-well plates were coated with 75 ng/well of penta-His antibody in PBS

680   solution (Qiagen, 34660) and incubated overnight at 4 °C. Plates were washed three times with

681   200 µl/well of wash buffer (PBS + 0.05% Tween-20) and any residual liquid was removed. Plates

682   were blocked with 100 µl/well Reagent Diluent (1% Probumin (Millipore, 82-045-1) in PBS pH 7.2)

683   for 2 h. Plates were washed three times with 200 µl/well of wash buffer and residual liquid was

684   removed. Dilutions of the yeast supernatant were prepared in 96-deepwell plates and 100 µl of a

685   100,000-fold dilution was applied to the plates and incubated for 1 hour while shaking gently on

686   a table top plate shaker. Plates were washed three times with 200 µl/well of wash buffer and

687   residual liquid was removed. Plates were provided with 100 µl of a 250 ng/ml MonoRab™ Rabbit

20

688 Anti-Camelid VHH Antibody coupled to HRP in Reagent Diluent and incubated for 1 hour while
689 shaking gently on a table top plate shaker. Plates were washed three times with 200 µl/well of
690 wash buffer and residual liquid was removed. TMB substrate was prepared according to the
691 manufacturer's instructions (BD OptEIA™) and 100 µl/well was applied to the plate before
692 incubating for 10 min. Next, 50 µl of stop solution (2N H_2SO_4) was added to each well and the
693 plate was read at 450 nm by a plate reader. Absorbance units were background corrected. All
694 strains were compared in a Kruskal-Wallis omnibus test, followed by pairwise comparison
695 corrected with Dunn's multiple comparison procedure.

Transformation efficiency testing

697 Competent cells were prepared using the lithium acetate method as described above. 200 ng of
698 linearized plasmid was transformed to each strain and several dilutions of the transformation mix
699 were plated on either non-selective YPD agar or YPD agar containing 100 µg/ml Zeocin. For each
700 transformation, colonies were counted from the plates where clear individual colonies could be
701 observed after 2 days at 30 °C incubation. Both the selective and non-selective plates were
702 counted to correct for a potential difference in the number of competent cells per transformation.

703 A linear model (estimated using ordinary least squares) in the statistical software R was fitted[63].
704 As the outcome variable, the log-transformed, normalized transformation efficiency (natural log of
705 the number of transformants per million clones) and as predictor variables the strain and promotor
706 type, including an interaction effect were used. The model explains a statistically significant and
707 substantial proportion of variance ($R^2=0.94$, $F(7,38)=81.33$, $p<.001$, $adj.R^2=0.93$). Model-
708 predicted group means with 95% confidence intervals were obtained using the ggeffects package
709 with heteroscedasticity-consistent variance estimators from the sandwich package (vcovHC, type
710 HC0)[64,65]. Further, we defined specific contrasts using the multcomp package[66], again with
711 heteroscedasticity-consistent variance estimators to obtain multiple-comparison corrected
712 estimates for the ratios of transformation efficiencies between the different strains and using
713 different plasmids.

DSA-FACE-based glycan analysis of the cell wall mannoproteins

715 Strains were inoculated in YPD or YPG, from their respective precultures, at an OD_{600} of 0.05 and
716 grown overnight at 28 °C and 200 rpm. The next day, 500 OD units per strain were pelleted
717 (10 min at 1,500 g) and the mannoproteins were isolated, as follows. The pellets were washed
718 three times with Milli-Q, after which 20 mM of citrate buffer pH6.6 was added at 1 ml per 150 µg
719 of wet cell weight. The resuspended cells were autoclaved for 1.5 hours at 120 °C in cryovials
720 and then centrifuged for 10 min at 16,000 g. To the supernatant fractions, 3 volumes of ice-cold
721 methanol were added and the vials were incubated for 15 min at 20 °C. The mannoproteins were

21

722   spun down for 10 min at 16,000 g and the pellets were left to dry until transparent. The pellets

723   were resuspended in 50 μl RCM buffer (8 M Urea, 360 mM Tris-HCl pH 8.6, 3.2 mM EDTA) and

724   stored at 4 °C until further analysis.

725   The N-linked oligosaccharides were prepared from the purified mannoproteins upon blotting to

726   PVDF membrane in the wells of 96-well plate membrane plates, and were analysed by capillary

727   electrophoresis with laser-induced fluorescence detection (CE-LIF) using an ABI 3130 capillary

728   DNA sequencer as described previously[36].

Alcian blue assay

730   The assay was performed as described previously[37], with adaptations. Briefly, Alcian blue was

731   prepared in 0.02 N HCl at a concentration of 63 μg/ml and the solution was centrifuged to remove

732   insoluble precipitates. An overnight culture of each strain was grown in YPD at 28 °C and

733   200 rpm. The next day, the cells were pelleted and the supernatant was removed. The cells were

734   washed with 0.02 N HCl and the pellet was resuspended again in 0.02 N HCl to an OD_{600} of

735   10 OD/ml. 100 μl (1 OD_{600}) of cells was transferred to a 96-V-bottom plate to which 100 μl of the

736   Alcian blue solution was added. After 15 min of incubation at room temperature, the plate was

737   centrifuged for 15 min at 3,220 g, after which the pellets were visually checked.

Congo red and Calcofluor white test

739   The test was performed as described elsewhere[67], with slight adaptations. Briefly, the strains were

740   grown overnight in BMGY. The next day, dilutions were made in order to obtain 10E5 to 10E1

741   cells in 5 μl BMGY. 5 μl drops were spotted on the different plates and the plates were incubated

742   for 3 days at 30 °C. Congo red (Sigma, C6767) and Calcofluor white (Fluka, 18909) were present

743   at final concentrations of 75 μg/ml and 10 μg/ml, respectively.

Electron microscopy

745   *Transmission electron microscopy*

746   The strains were cultivated in BMGY at 28 °C and 200 rpm, overnight. High Pressure Freezing,

747   as described previously[68], was carried out in a high-pressure freezer (Leica EM ICE; Leica

748   Microsystems, Vienna, Austria). Cells were pelleted and frozen as a paste in 150 μm cupper

749   carriers. HPF was followed by Quick Freeze Substitution as described previously[69]. Briefly,

750   carriers were placed on top of the frozen FS solution inside a cryovial containing 1% ddH_2O,

751   1% OsO_4 and 0.5% glutaraldehyde in dried acetone. After reaching 4 °C for 30 min, samples were

752   infiltrated stepwise over three days at 0-4 °C in Spurr's resin and embedded in capsules. The

753   polymerization was performed at 70 °C for 16 h. Ultrathin sections of a gold interference colour

22

754   were cut using an ultra-microtome (Leica EM UC6), followed by a post-staining in a Leica EM

755   AC20 for 40 min in uranyl acetate at 20 °C and for 10 min in lead stain at 20 °C.

756   Sections were collected on formvar-coated copper slot grids. Grids were viewed with a JEM-

757   1400Plus transmission electron microscope (JEOL, Tokyo, Japan) operating at 60 kV.

758   *Scanning electron microscopy*

759   The strains were cultivated in BMGY at 28 °C and 200 rpm, overnight. Cells were fixed overnight

760   in 1.5% Paraformaldehyde, 3% Glutaraldehyde in 0.05 M Na-Cacodylate buffer pH7.4. The fixed

761   cells were centrifuged for 2 min at 1,000 g between each following step. First the cells were

762   washed 3 times with 0.1 M Na-Cacodylate buffer pH7.4 and then incubated for 30 min in 2% OsO_4

763   in 0.1 M Na-Cacodylate pH7.4. Osmicated samples were washed 3 times with Milli-Q, prior to a

764   stepwise ethanol dehydration (50%, 70%, 90%, 2 x 100%). Samples were incubated twice in

765   hexamethyldisilazane (HMDS) solution (Sigma-Aldrich), as a final dehydration step, after which

766   they were spotted on silicon grids (Ted Pella) and air-dried overnight at room temperature.

767   Samples were next coated with 5 nm Platinum (Pt) in a Quorum Q 150T ES sputter coater

768   (Quorum Technologies) and placed in a Gemini 2 Cross beam 540 microscope from Zeiss for

769   SEM imaging at 1.50 kV using a SE2 detector.

770

771   **Associated content**

772   •   All plasmids from the modular cloning kit are available at the Belgian Co-ordinated

773   Collections of Micro-organisms (BCCM)/GeneCorner Plasmid Collection

774   (http://bccm.belspo.be/about-us/bccm-genecorner).

775   •   All raw reads of the genomes sequenced in this study have been submitted to NCBI and

776   can be found under the following accession numbers: NRRL Y-11430 (SAMN32067769),

777   NRRL YB-4290 (SAMN32067770), NRRL Y-7556 (SAMN32067773), NCYC 2543

778   (SAMN32067771), CBS 2612 (SAMN32067772).

779   •   Supplementary Figure 1: Summary of the end-ODs of the pGAP- and pAOX1-based

780   cultivations at harvest.

781   •   Supplementary Figure 2: DSA-FACE profiles of the cell wall mannoproteins of the different

782   strains grown on YPD or YPG.

783   •   Supplementary Figure 3: Overview of the available elements for the different parts of the

784   MoClo toolbox.

785   •   Supplementary Figure 4: AlphaFold 2 models of the type strain Hoc1p and of the NRRL

786   Y-11430 lineage derived strain.

787 • Supplementary Table 1: Overview of the NGS results, including the number of reads, GC% and average overall coverage.

789 • Supplementary Table 2: Proportion of NGS reads mapping to different molecules of the reference genome, mitochondrial DNA and killer plasmids.

791 • Supplementary Table 3: List of oligonucleotides that were used as primers for PCR, cPCR or sequencing.

793 • Supplementary Table 4. Sequences of the split-marker fragments used to generate the two *HOC1* mutants.

795 • Supplementary File 1: GenBank files of the expression constructs used in the study.

796 • Supplementary File 2: FASTA files of the available MoClo Parts.

797

Author Information

804

Acknowledgements

825 **References**

826 1. Licences | National Collection of Yeast Cultures. https://www.ncyc.co.uk/licences.
827 2. Karbalaei, M., Rezaee, S. A. & Farsiani, H. *Pichia pastoris* : A highly successful expression
828 system for optimal synthesis of heterologous proteins. *J Cell Physiol* jcp.29583 (2020)
829 doi:10.1002/jcp.29583.
830 3. Adivitiya, Dagar, V. K. & Khasa, Y. P. Yeast Expression Systems: Current Status and Future
831 Prospects. in *Yeast Diversity in Human Welfare* (eds. Satyanarayana, T. & Kunze, G.) 215–
832 250 (Springer Singapore, 2017). doi:10.1007/978-981-10-2621-8_9.
833 4. Yang, Z. & Zhang, Z. Engineering strategies for enhanced production of protein and bio-
834 products in Pichia pastoris: A review. *Biotechnology Advances* **36**, 182–195 (2018).
835 5. Baumschabl, M. *et al.* Conversion of CO2 into organic acids by engineered autotrophic yeast.
836 *Proc Natl Acad Sci U S A* **119**, e2211827119 (2022).
837 6. Liu, L. *et al.* How to achieve high-level expression of microbial enzymes: strategies and
838 perspectives. *Bioengineered* **4**, 212–223 (2013).
839 7. Gasser, B. *et al.* Pichia pastoris: protein production host and model organism for biomedical
840 research. *Future Microbiology* **8**, 191–208 (2013).
841 8. Phaff, H. J., Miller, M. W. & Shifrine, M. The taxonomy of yeasts isolated fromDrosophila in
842 the Yosemite region of California. *Antonie van Leeuwenhoek* **22**, 145–161 (1956).
843 9. Phaff, H. J. A proposal for amendment of the diagnosis of the genusPichia hansen. *Antonie
844 van Leeuwenhoek* **22**, 113–116 (1956).
845 10. Yamada, Y., Matsuda, M., Maeda, K. & Mikata, K. The Phylogenetic Relationships of
846 Methanol-assimilating Yeasts Based on the Partial Sequences of 18S and 26S Ribosomal
847 RNAs: The Proposal of *Komagataella* Gen. Nov. (Saccharomycetaceae). *Bioscience,
848 Biotechnology, and Biochemistry* **59**, 439–444 (1995).
849 11. Kurtzman, C. P. Description of Komagataella phaffii sp. nov. and the transfer of Pichia
850 pseudopastoris to the methylotrophic yeast genus Komagataella. *INTERNATIONAL
851 JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY* **55**, 973–976 (2005).
852 12. Ogata, K., Nishikawa, H. & Ohsugi, M. A yeast capable of utilizing methanol. *Agricultural and
853 biological chemistry* **33**, 1519–1520 (1969).
854 13. Tani, Y., Miya, T., Nishikawa, H. & Ogata, K. The Microbial Metabolism of Methanol: Part I.
855 Formation and Crystallization of Methanol-oxidizing Enzyme in a Methanol-utilizing Yeast,
856 Kloeckera sp. No. 2201 Part II. Properties of Crystalline Alcohol Oxidase from Kloeckera sp.
857 No. 2201. *Agricultural and Biological Chemistry* **36**, 68–83 (1972).
858 14. TANI, Y., MIYA, T. & OGATA, K. The microbial metabolism of methanol Part II. *Agricultural
859 and Biological Chemistry* **36**, 76–83 (1972).
860 15. Wegner Eugene Herman. A Process For Producing Single Cell Protein Material And Culture.
861 (1980).
862 16. Kurtzman, C. P. Biotechnological strains of Komagataella (Pichia) pastoris are Komagataella
863 phaffii as determined from multigene sequence analysis. *J Ind Microbiol Biotechnol* **36**, 1435–
864 1438 (2009).
865 17. Love, K. R. *et al.* Comparative genomics and transcriptomics of Pichia pastoris. *BMC
866 genomics* **17**, 550 (2016).

867   18. De Schutter, K. *et al.* Genome sequence of the recombinant protein production host Pichia
868   pastoris. *Nature biotechnology* **27**, 561 (2009).

869   19. Sturmberger, L. *et al.* Refined Pichia pastoris reference genome sequence. *Journal of*
870   *biotechnology* **235**, 121–131 (2016).

871   20. Mattanovich, D. *et al.* Open access to sequence: browsing the Pichia pastoris genome.
872   *Microbial cell factories* **8**, 53 (2009).

873   21. Pichia Technology from RCT. https://pichia.com/.

874   22. Gasser, B. *et al. Pichia pastoris* : protein production host and model organism for biomedical
875   research. *Future Microbiology* **8**, 191–208 (2013).

876   23. Fischer, J. E. & Glieder, A. Current advances in engineering tools for Pichia pastoris. *Current*
877   *Opinion in Biotechnology* **59**, 175–181 (2019).

878   24. Brady, J. R. *et al.* Comparative genome-scale analysis of *Pichia pastoris* variants informs
879   selection of an optimal base strain. *Biotechnology and Bioengineering* **117**, 543–555 (2020).

880   25. Prielhofer, R. *et al.* GoldenPiCS: a Golden Gate-derived modular cloning system for applied
881   synthetic biology in the yeast Pichia pastoris. *BMC Syst Biol* **11**, 123 (2017).

882   26. Sturmberger, L. *et al.* Refined Pichia pastoris reference genome sequence. *Journal of*
883   *Biotechnology* **235**, 121–131 (2016).

884   27. Deatherage, D. E. & Barrick, J. E. Identification of Mutations in Laboratory-Evolved Microbes
885   from Next-Generation Sequencing Data Using breseq. in *Engineering and Analyzing*
886   *Multicellular Systems* (eds. Sun, L. & Shou, W.) vol. 1151 165–188 (Springer New York,
887   2014).

888   28. Braun-Galleani, S. *et al. Tetrad analysis without tetrad dissection: Meiotic recombination and*
889   *genomic diversity in the yeast* Komagataella phaffii (Pichia pastoris).
890   http://biorxiv.org/lookup/doi/10.1101/704627 (2019) doi:10.1101/704627.

891   29. Offei, B. *et al.* Identification of genetic variants of the industrial yeast Komagataella phaffii
892   (Pichia pastoris) that contribute to increased yields of secreted heterologous proteins. *PLoS*
893   *Bio* **20**, (2022).

894   30. Lu, L., Roberts, G. G., Oszust, C. & Hudson, A. P. The YJR127C/ZMS1 gene product is
895   involved in glycerol-based respiratory growth of the yeast Saccharomyces cerevisiae. *Current*
896   *genetics* **48**, 235–246 (2005).

897   31. Jungmann, J. & Munro, S. Multi-protein complexes in the cis Golgi of Saccharomyces
898   cerevisiae with alpha-1,6-mannosyltransferase activity. *EMBO J* **17**, 423–434 (1998).

899   32. Braun-Galleani, S. *et al.* Genomic diversity and meiotic recombination among isolates of the
900   biotech yeast Komagataella phaffii (Pichia pastoris). *Microb Cell Fact* **18**, 211 (2019).

901   33. Fairhead, C., Llorente, B., Denis, F., Soler, M. & Dujon, B. New vectors for combinatorial
902   deletions in yeast chromosomes and for gap-repair cloning using 'split-marker'recombination.
903   *Yeast* **12**, 1439–1457 (1996).

904   34. Heiss, S., Maurer, M., Hahn, R., Mattanovich, D. & Gasser, B. Identification and deletion of
905   the major secreted protein of Pichia pastoris. *Appl Microbiol Biotechnol* **97**, 1241–1249
906   (2013).

907   35. Vogl, T., Gebbie, L., Palfreyman, R. W. & Speight, R. Effect of Plasmid Design and Type of
908   Integration Event on Recombinant Protein Expression in *Pichia pastoris. Appl Environ*
909   *Microbiol* **84**, e02712-17, /aem/84/6/e02712-17.atom (2018).

910   36. Laroy, W., Contreras, R. & Callewaert, N. Glycome mapping on DNA sequencing equipment.
911   *Nat Protoc* **1**, 397–405 (2006).

912   37. Conde, R., Pablo, G., Cueva, R. & Larriba, G. Screening for new yeast mutants affected in
913   mannosylphosphorylation of cell wall mannoproteins. *Yeast* **20**, 1189–1211 (2003).

914   38. Friis, J. & Ottolenghi, P. The genetically determined binding of alcian blue by a minor fraction
915   of yeast cell walls. *C R Trav Lab Carlsberg* **37**, 327–341 (1970).

916 39. Casini, A., Storch, M., Baldwin, G. S. & Ellis, T. Bricks and blueprints: methods and standards
917   for DNA assembly. *Nat Rev Mol Cell Biol* **16**, 568–576 (2015).
918 40. Moore, S. J. *et al.* EcoFlex: A Multifunctional MoClo Kit for *E. coli* Synthetic Biology. *ACS*
919   *Synth. Biol.* **5**, 1059–1069 (2016).
920 41. Lee, M. E., DeLoache, W. C., Cervantes, B. & Dueber, J. E. A Highly Characterized Yeast
921   Toolkit for Modular, Multipart Assembly. *ACS Synth. Biol.* **4**, 975–986 (2015).
922 42. van Dolleweerd, C. J. *et al.* MIDAS: A Modular DNA Assembly System for Synthetic Biology.
923   *ACS Synth. Biol.* **7**, 1018–1029 (2018).
924 43. Hernanz-Koers, M. *et al.* FungalBraid: A GoldenBraid-based modular cloning platform for the
925   assembly and exchange of DNA elements tailored to fungal synthetic biology. *Fungal*
926   *Genetics and Biology* **116**, 51–61 (2018).
927 44. Sarrion-Perdigones, A. *et al.* GoldenBraid: An Iterative Cloning System for Standardized
928   Assembly of Reusable Genetic Modules. *PLoS ONE* **6**, e21622 (2011).
929 45. Obst, U., Lu, T. K. & Sieber, V. A Modular Toolkit for Generating *Pichia pastoris* Secretion
930   Libraries. *ACS Synth. Biol.* **6**, 1016–1025 (2017).
931 46. Andreou, A. I. & Nakayama, N. Mobius Assembly: A versatile Golden-Gate framework
932   towards universal DNA assembly. *PLoS ONE* **13**, e0189892 (2018).
933 47. Engler, C. *et al.* A Golden Gate Modular Cloning Toolbox for Plants. *ACS Synth. Biol.* **3**, 839–
934   843 (2014).
935 48. Weber, E., Engler, C., Gruetzner, R., Werner, S. & Marillonnet, S. A Modular Cloning System
936   for Standardized Assembly of Multigene Constructs. *PLoS ONE* **6**, e16765 (2011).
937 49. Potapov, V. *et al.* Comprehensive Profiling of Four Base Overhang Ligation Fidelity by T4
938   DNA Ligase and Application to DNA Assembly. *ACS Synth Biol* **7**, 2665–2674 (2018).
939 50. OPENPichia plasmid set | BCCM Belgian Coordinated Collections of Microorganisms.
940   https://bccm.belspo.be/catalogues/plasmid-sets/openpichia.
941 51. Jungmann, J., Rayner, J. C. & Munro, S. The Saccharomyces cerevisiae protein
942   Mnn10p/Bed1p is a subunit of a Golgi mannosyltransferase complex. *J Biol Chem* **274**, 6579–
943   6585 (1999).
944 52. HOC1 Interactions | SGD. https://www.yeastgenome.org/locus/S000003836/interaction.
945 53. Wu, L. *et al.* Global catalogue of microorganisms (gcm): a comprehensive database and
946   information retrieval, analysis, and visualization system for microbial resources. *BMC*
947   *Genomics* **14**, 933 (2013).
948 54. Lin, Y.-C. *et al.* Genome dynamics of the human embryonic kidney 293 lineage in response
949   to cell biology manipulations. *Nat Commun* **5**, 4767 (2014).
950 55. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data. Available*
951   *online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.* (2010).
952 56. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence
953   data. *Bioinformatics* **30**, 2114–2120 (2014).
954 57. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary
955   Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution* **35**, 1547–
956   1549 (2018).
957 58. Heiss, S., Maurer, M., Hahn, R., Mattanovich, D. & Gasser, B. Identification and deletion of
958   the major secreted protein of Pichia pastoris. *Applied microbiology and biotechnology* **97**,
959   1241–1249 (2013).
960 59. Näätsaari, L. *et al.* Deletion of the Pichia pastoris KU70 Homologue Facilitates Platform Strain
961   Generation for Gene Expression and Synthetic Biology. *PLoS ONE* **7**, e39720 (2012).
962 60. Weninger, A., Hatzl, A.-M., Schmid, C., Vogl, T. & Glieder, A. Combinatorial optimization of
963   CRISPR/Cas9 expression enables precision genome engineering in the methylotrophic yeast
964   Pichia pastoris. *Journal of Biotechnology* **235**, 139–149 (2016).

61. Wu, S. & Letchworth, G. J. High efficiency transformation by electroporation of *Pichia pastoris* pretreated with lithium acetate and dithiothreitol. *BioTechniques* **36**, 152–154 (2004).

62. 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).

63. R: The R Project for Statistical Computing. https://www.r-project.org/.

64. Lüdecke, D. ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. *The Journal of Open Source Software* **3**, (2018).

65. Zeileis, A., Köll, S. & Graham, N. Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R. *Journal of Statistical Software* **95**, 1–36 (2020).

66. Hothorn, T., Bretz, F. & Westfall, P. Simultaneous inference in general parametric models. *Biom J* **50**, 346–363 (2008).

67. Ram, A. F. J. & Klis, F. M. Identification of fungal cell wall mutants using susceptibility assays based on Calcofluor white and Congo red. *Nat Protoc* **1**, 2253–2256 (2006).

68. Arendt, P. *et al.* An endoplasmic reticulum-engineered yeast platform for overproduction of triterpenoids. *Metab Eng* **40**, 165–175 (2017).

69. McDonald, K. L. & Webb, R. I. Freeze substitution in 3 hours or less. *J Microsc* **243**, 227–233 (2011).

70. Walter, M. R. *et al.* Three-dimensional structure of recombinant human granulocyte-macrophage colony-stimulating factor. *Journal of Molecular Biology* **224**, 1075–1085 (1992).

71. Wrapp, D. *et al.* Structural Basis for Potent Neutralization of Betacoronaviruses by Single-Domain Camelid Antibodies. *Cell* **181**, 1004-1015.e15 (2020).

72. Yang, Z. *et al.* A Novel Multivalent, Single-Domain Antibody Targeting TcdA and TcdB Prevents Fulminant Clostridium difficile Infection in Mice. *The Journal of Infectious Diseases* **210**, 964–972 (2014).

73. McPherson, M. J. *et al.* Galactose oxidase of Dactylium dendroides. Gene cloning and sequence analysis. *Journal of Biological Chemistry* **267**, 8146–8152 (1992).

74. Kubala, M. H., Kovtun, O., Alexandrov, K. & Collins, B. M. Structural and thermodynamic analysis of the GFP:GFP-nanobody complex. *Protein Science* **19**, 2389–2401 (2010).

Tables

Table 1. Strains used in this publication. * Type strain; Abbreviations: UCD: University of California, Davis, USA; CBS: Centraalbureau voor Schimmelculturen, currently known as Westerdijk Fungal Biodiversity Institute, The Netherlands; NCYC: National Collection of Yeast Cultures, UK; NRRL: Northern Regional Research Laboratory, currently known as Agricultural Research Service (ARS), USA; NTG: nitrosoguanidine.

Strain ID	Origin of strain isolate	Source	Original depositor
CBS 2612 *	*Quercus kelloggii* (California, USA)	CBS culture collection	HJ Phaff, UCD
NCYC 2543 *	*Quercus kelloggii* (California, USA)	NCYC culture collection	CBS
NRRL YB-4290 *	*Quercus kelloggii* (California, USA)	NRRL culture collection	HJ Phaff, UCD
NRRL Y-7556 *	*Quercus kelloggii* (California, USA)	NRRL culture collection	D. Yarrow, CBS
UCD FST K-239 *	*Quercus kelloggii* (California, USA)	UCD culture collection	HJ Phaff, UCD
NRRL Y-11430	Most likely a subclone of the type strain	NRRL culture collection	Patent deposit Phillips Petroleum Company (USA)
GS115	NTG-mutagenized derivative of NRRL Y-11430	Internal VIB culture collection	
NCYC 2543 *his4*	Mutagenesis	This publication	
NCYC 2543 *hoc1^{tr}-1*	Mutagenesis	This publication	
NCYC 2543 *hoc1^{tr}-2*	Mutagenesis	This publication	

Table 2. Overview of the functional mutations of the analysed K. phaffii strains compared to the CBS 7435 (eq. strain deposit of NRRL- Y-11430) reference genome, non-functional SNPs and indels are between brackets. Functional mutations are those resulting in a SNP or indel in a gene.

Strain Designation	SNP	Indel	Total	Mutations/Mbp	Killer plasmids copy number (KP1/KP2)
NRRL Y-11430	0 (2)	2 (18)	2 (20)	2.3	21/15
NRRL YB-4290	2 (2)	3 (20)	5 (22)	2.9	140/82
NCYC 2543	3 (5)	3 (18)	6 (23)	3.1	none detected
CBS 2612	3 (4)	3 (19)	6 (23)	3.1	none detected
NRRL Y-7556	3 (3)	3 (21)	6 (24)	3.2	138/86

Table 3. Summary of the coding mutations in the Philips Petroleum strain and the type strains, compared to the CBS 7435 reference genome. *The exact mutations are mentioned in the second column, with the CBS 7435 amino acid as reference amino acid; although the original genetic makeup is as present in the type strains, and it is the CBS 7435/NRRL Y-11430 that mutated. The mutations indicated with an asterisk were also reported by Brady et. al.[24]. All mutations found in these strains are concentrated in eight locations. The mutations in SEF1, ROP100/RSF2, and HOC1 are shared by all type strains. Two mutations in a gene encoding for a hypothetical protein were found in all sequenced genomes. The other mutations (in SRB7, RAD18 and PRP46) are present in one of the type strains, NCYC 2543, CBS 2612 and NRRL Y-7556, respectively.*

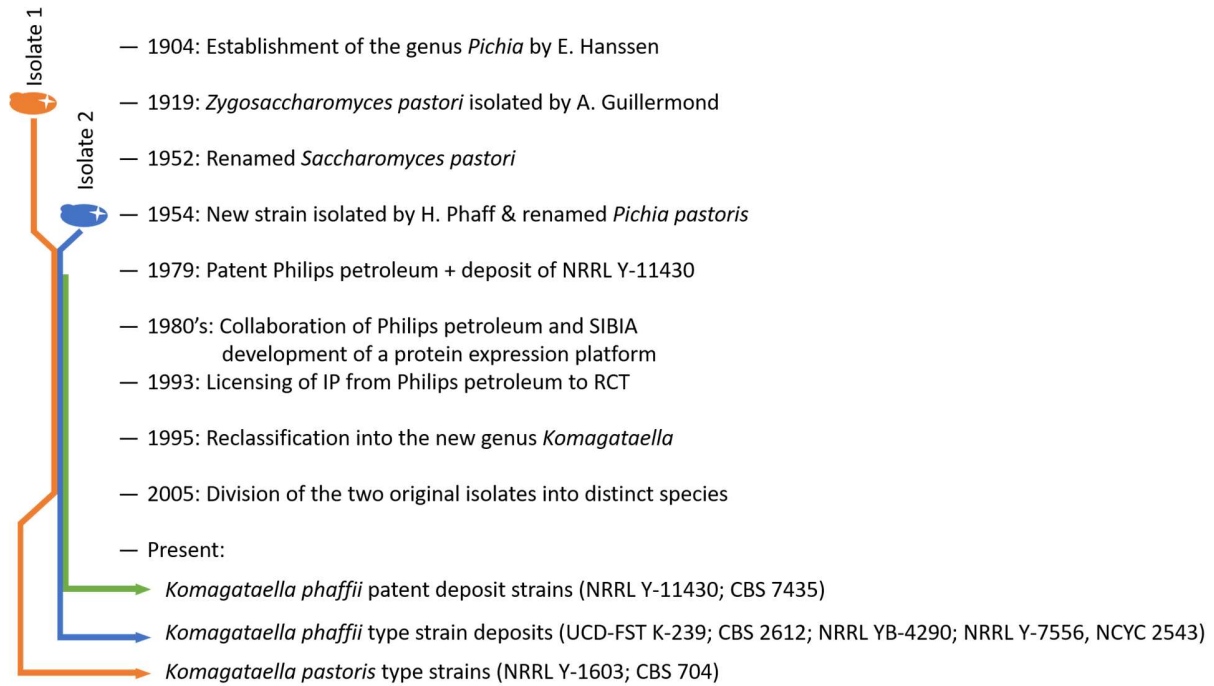| Gene	Mutation	NRRL Y-11430	NRRL YB-4290	NCYC 2543	CBS 2612	NRRL Y-7556	Gene function
SEF1	SNP: C315S		x	x	x	x	Putative transcription factor
ROP100/RSF2	SNP: *748W		x*	x	x	x*	Methanol- and biotin-starvation-inducible zinc finger protein
HOC1	Indel: (T)$_{5\rightarrow6}$ Premature *		x	x	x	x	Alpha1,6mannosyltransferase
Hypothetical	Indel 1: +G Indel 2: +GT GNYD821GFPND	x	x	x	x	x	Papain-like cysteine protease
SRB7	SNP: E131Q			x			RNA polymerase II mediator complex subunit
RAD18	SNP: K210N				x		E3 ubiquitin ligase
PRP46	SNP: Q320L					x	NineTeen Complex (NTC) component
Total mutations in protein-coding sequences		**2**	**5**	**6**	**6**	**6**	

Table 4. Details of the selected proteins used for the protein expression comparisons.

Abbreviation	Name	Type	Molecular mass (kDa)	Secretion leader	Ref.
GM-CSF	Granulocyte-macrophage colony-stimulating factor	Cytokine	15.7	αMF	[70]
GaOx	Galactose oxidase from *Fusarium graminearum*	Enzyme	69.9	αMF	[73]
Cdiff-VHH-IgA	Anti-*C.difficile* toxin VHH IgA fusion protein	VHH-IgA	41.2	Ost1	[72]
CovidVHH-IgG	SARS-CoV-2 neutralizing VHH hIgG1 fusion protein	VHH-IgG	40.6	Ost1	[71]
GBP	*GFP-binding protein*	VHH	14.1	MF	[74]

1017 **Figures**

1018

Isolate 1

Isolate 2

— 1904: Establishment of the genus *Pichia* by E. Hanssen

— 1919: *Zygosaccharomyces pastori* isolated by A. Guillermond

— 1952: Renamed *Saccharomyces pastori*

— 1954: New strain isolated by H. Phaff & renamed *Pichia pastoris*

— 1979: Patent Philips petroleum + deposit of NRRL Y-11430

— 1980's: Collaboration of Philips petroleum and SIBIA
       development of a protein expression platform
— 1993: Licensing of IP from Philips petroleum to RCT

— 1995: Reclassification into the new genus *Komagataella*

— 2005: Division of the two original isolates into distinct species

— Present:

Komagataella phaffii patent deposit strains (NRRL Y-11430; CBS 7435)

Komagataella phaffii type strain deposits (UCD-FST K-239; CBS 2612; NRRL YB-4290; NRRL Y-7556, NCYC 2543)

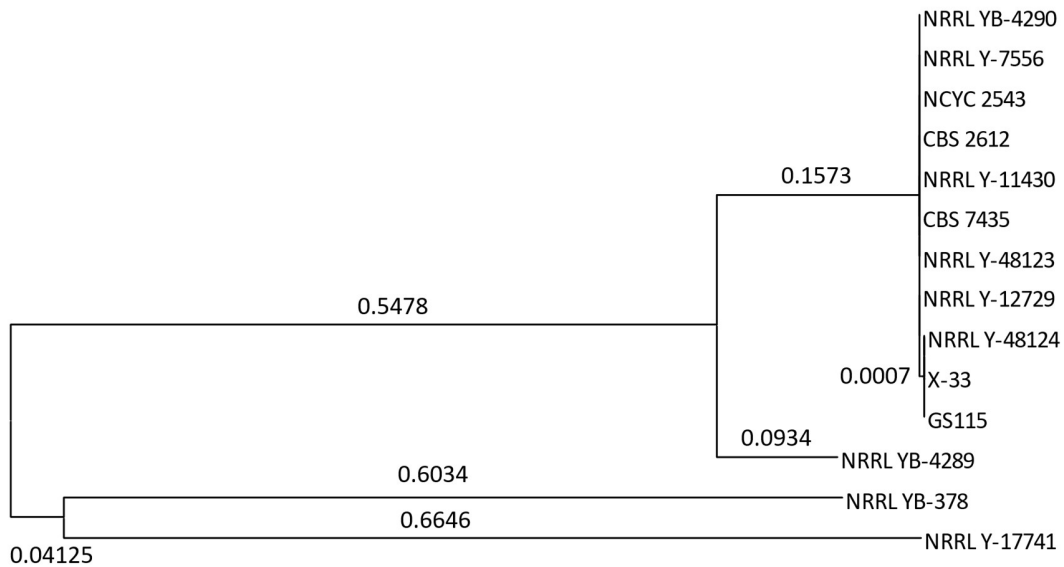*Komagataella pastoris* type strains (NRRL Y-1603; CBS 704)

1019
1020 *Figure 1. Schematic timeline of the history of the Komagataella species and the available type strain deposits*
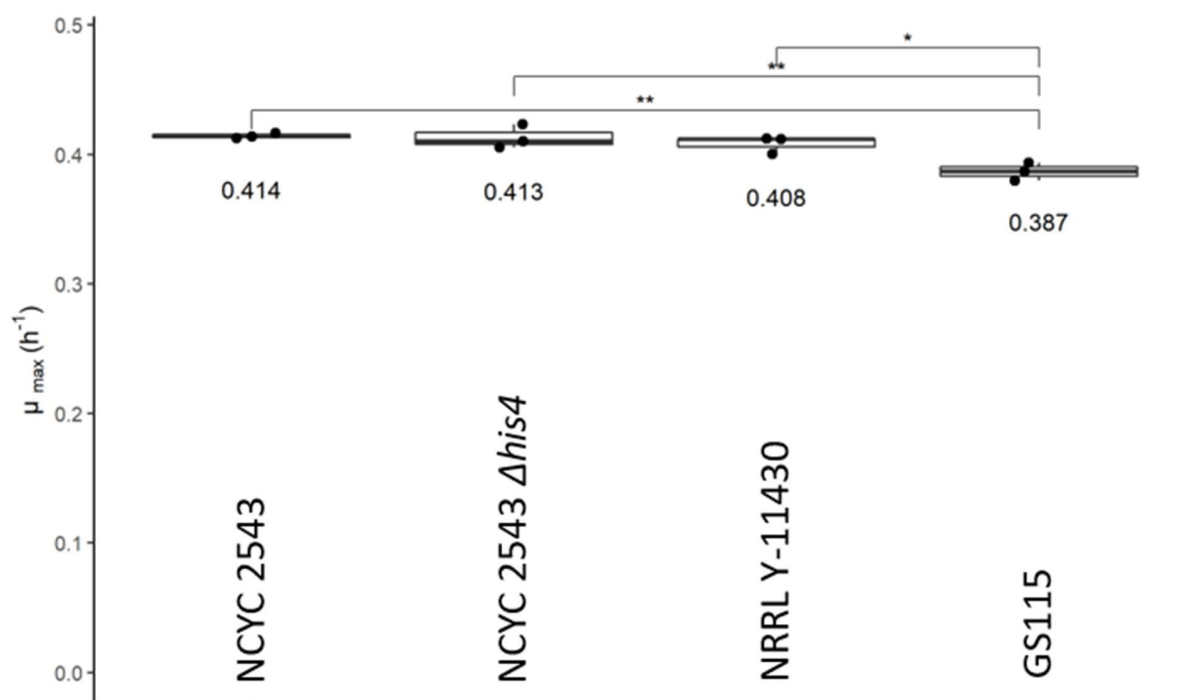1021 *and patent strain deposits.*

1022

1023
1024 *Figure 1: Phylogenetic tree of the K. phaffii strains with node lengths. The tree was constructed using a Maximum*
1025 *Likelihood method and Hasegawa-Kishino-Yano model. Node lengths of less than 0.0001 were neglected.*
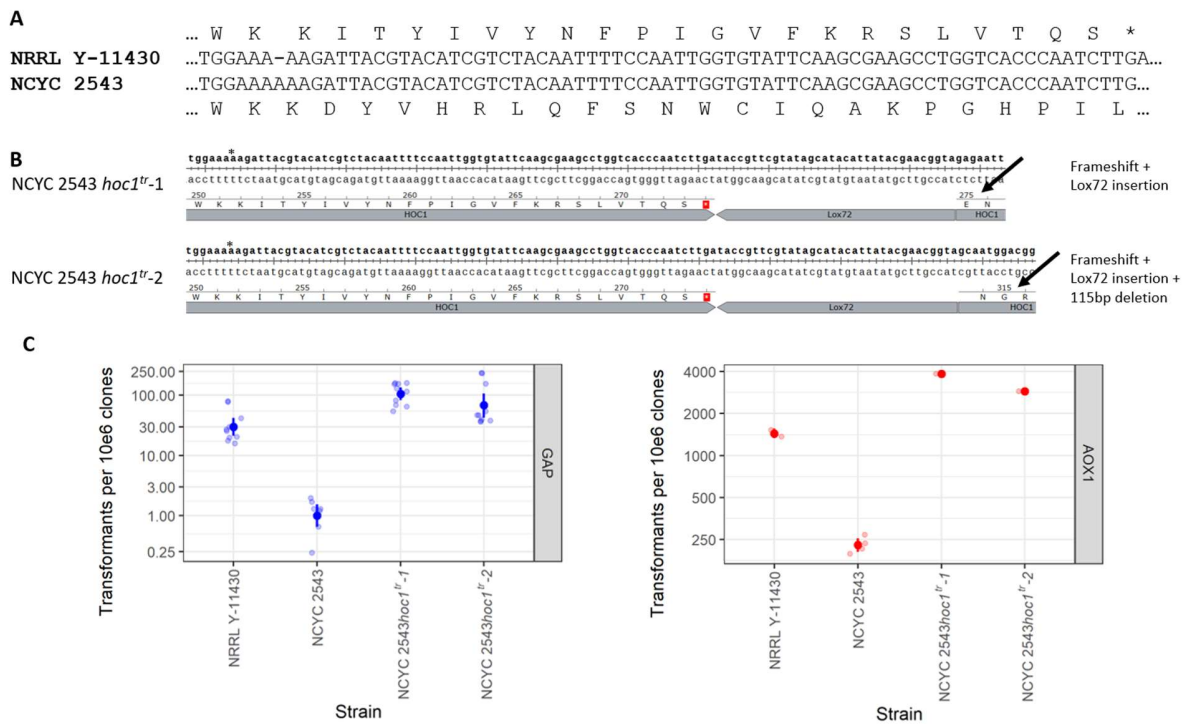
1026

Figure 3. Maximum growth rate of the different Pichia strains, plotted individually as dots, as well as a boxplot with median value. *A one-way ANOVA test showed significant influence of the strain on the growth rate (p = 0.0034). A post-hoc Tukey test shows that GS115 grows significantly slower than the other strains. The asterisks indicate the p-value of the Tukey test: **, 0.01>p>0.001; *, 0.05>p>0.01.*
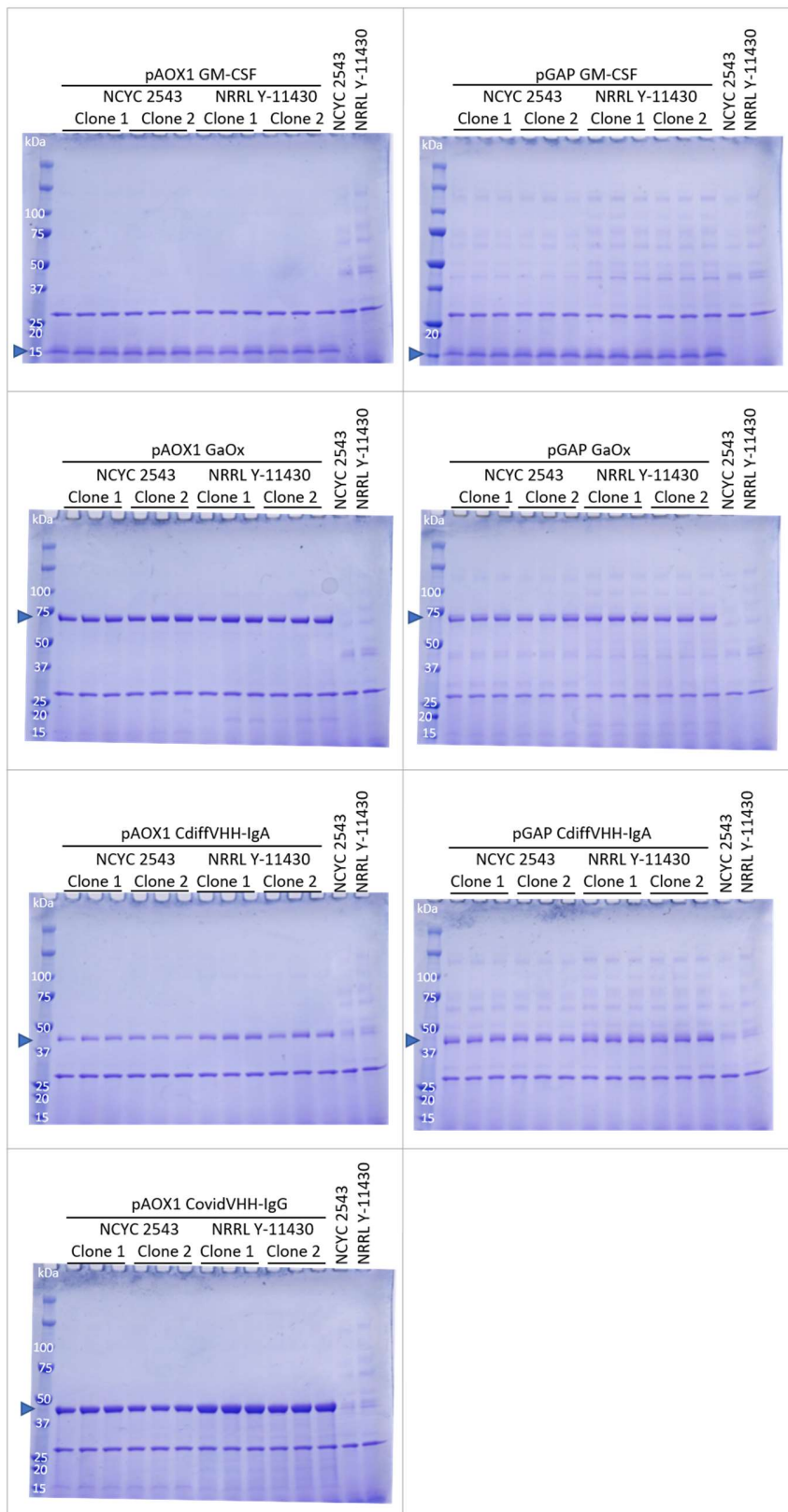
Figure 4. Expression comparison between NCYC 2543 and NRRL Y-11430. *The proteins were expressed using the GAP or AOX1 promoter. As controls, both wild type strains were grown and analysed as well. Supernatant samples were treated with EndoH to remove N-glycans and samples were analysed on SDS-PAGE. EndoH is also visible on the gels at around 30 kDa.*

33

A

```
         … W   K   K   I   T   Y   I   V   Y   N   F   P   I   G   V   F   K   R   S   L   V   T   Q   S   *
NRRL Y-11430  …TGGAAA-AAGATTACGTACATCGTCTACAATTTTCCAATTGGTGTATTCAAGCGAAGCCTGGTCACCCAATCTTGA…
NCYC 2543     …TGGAAAAAAGATTACGTACATCGTCTACAATTTTCCAATTGGTGTATTCAAGCGAAGCCTGGTCACCCAATCTTG…
         … W   K   K   D   Y   V   H   R   L   Q   F   S   N   W   C   I   Q   A   K   P   G   H   P   I   L …
```

B

NCYC 2543 *hoc1*^{tr}-1   Frameshift + Lox72 insertion

NCYC 2543 *hoc1*^{tr}-2   Frameshift + Lox72 insertion + 115bp deletion

C

Figure 5. Overview of the HOC1 genome engineering strategy and the effect on transformation efficiency of the resulting strains. *A. Alignment of a part of the HOC1 gene as present in NRRL Y-11430 vs. NCYC 2543, showing the frameshift resulting in a premature stop codon in the NRRL Y-11430. B. Resulting genomic HOC1 sequence upon split-marker-based gene editing. Two strategies were followed, where either the single base pair deletion (indicated with *) resulting in the Hoc1p truncation and a Lox72 scar is introduced downstream of the stop codon; or where an additional 115 bp deletion downstream of the resulting stop codon and Lox72 scar is introduced. C. Transformation efficiency in the two wild type strains and two HOC1-engineered strains, either using a pGAP-based plasmid (left) or a pAOX1-based plasmid (right). The analysis was performed as described in the Materials & Methods section.*
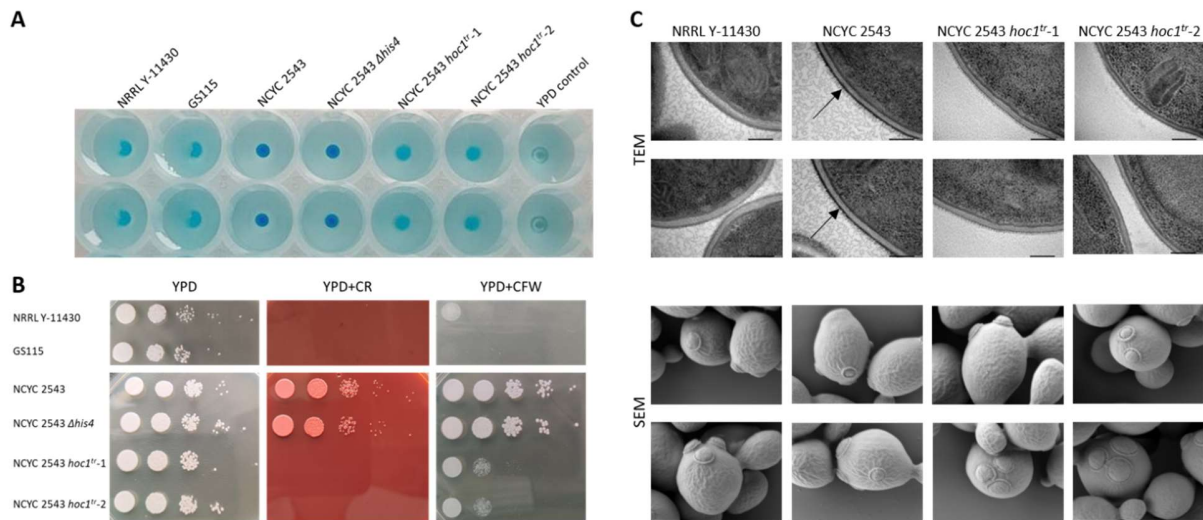
**Figure 6. Characterization of the cell walls of NRRL Y-11430, NCYC 2543 and the two NCYC 2543 hoc1^{tr} mutants.**
A. Alcian blue staining of the strains to determine the density of negative charges at the yeast cell wall. Alcian blue is a cationic dye that binds negative charges at the cell wall. The more intense the blue staining of the cells, the more negative charge, i.e., mannosylphosphate moieties, are present on the glycan trees of the cell wall mannoproteins. Duplicate wells are shown per strain (vertical). B. Sensitivity of the strains towards Congo red and Calcofluor white, compared to growth on YPD agar, as an indicator for cell wall integrity. The plates were incubated for 3 days at 30 °C. C. Transmission electron microscopy (TEM) and scanning electron microscopy (SEM) images of the four strains. The increased electron density of the outermost layer, i.e., the cell wall is indicated with an arrow in the TEM images of the NCYC 2543 strain. Only two individual images per strain are shown.
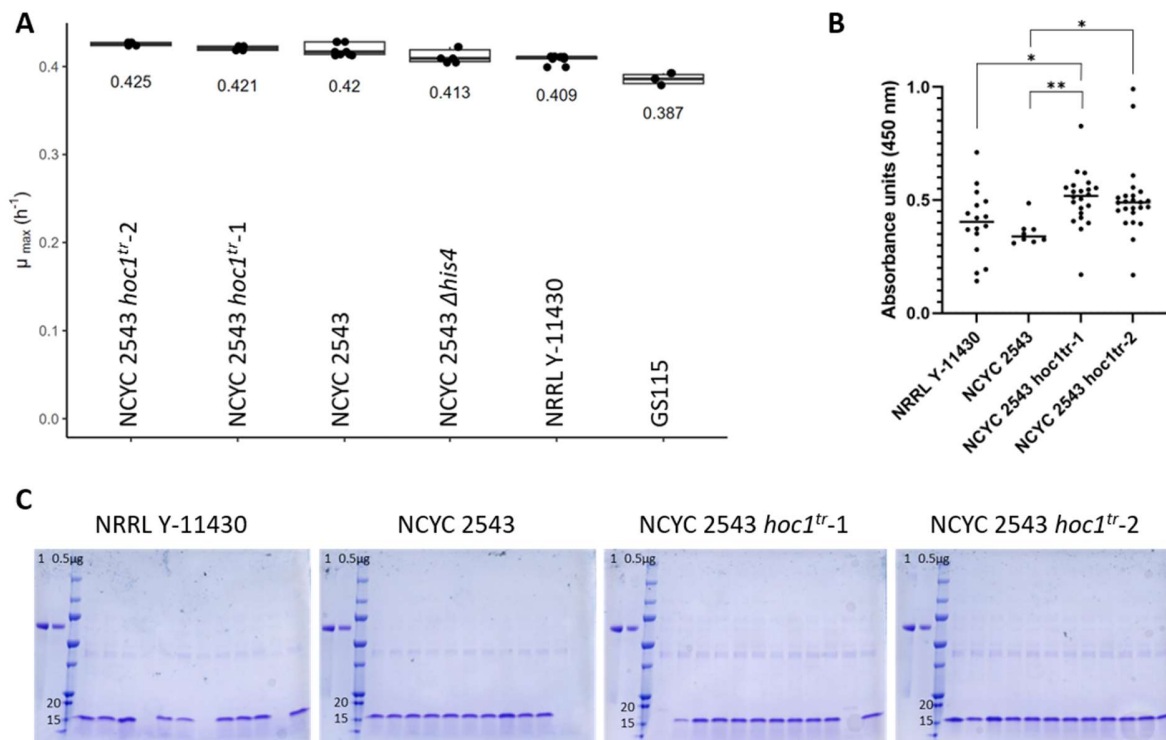
Figure 7. Overview of the strain performance of NRRL Y-11430, NCYC 2543 and the two NCYC 2543/OPENPichia hoc1^{tr} mutants. *A. Comparison of the maximal growth rate of the four main strains, accompanied by GS115 and the NCYC 2543/OPENPichia his4. B. ELISA results of 24 clones of pGAP-based GBP expression in the different strains (no selection for single copies was done, wells were excluded when no expression was observed on SDS-PAGE, assuming these clones contain no expression cassette, or due to a technical issue during the ELISA procedure). Absorbance units were background corrected. All strains were compared in a Kruskal-Wallis omnibus test, followed by pairwise comparison corrected with Dunn's multiple comparison procedure. Significance scores are annotated as follows, *: $0.002 < p < 0.03$; **: $0.001 < p < 0.002$; non-significant differences are not shown. C. SDS-PAGE analysis of the first 12 clones per transformation. Equal volumes of supernatant were loaded.*

Figure 8. The modular cloning or MoClo principle. *In a first phase, source DNA, such as PCR fragments, synthetic genes or annealed oligonucleotides are flanked with the proper Type IIS restriction sites and 4 nt overhangs, which are then accommodated in a Level 0 entry vector through BsmBI digest and T4 DNA ligation. Then, selected Level 0 vectors are assembled into a Level 1 expression vector by means of a BsaI digest and T4 DNA ligation. Finally, the system allows the assembly of multiple transcription units (promoter, CDS, terminator) from the individual Level 1 vectors, into a higher order Level 2 vector, in case the assembly connector sequences were properly selected during the assembly of the Level 1 vectors. Assembled assembly connectors are depicted with two horizontal lines and different shades of blue and purple. Note that the Part 3 Coding Sequence can be split up in a Part 3a and Part 3b Coding Sequence, to allow additional modularity. Likewise, the Part 4 Terminator can be split up in a Part 4a Coding Sequence and a Part 4b Terminator. Figure adapted from Lee et al[41].*