Giant *GAL* gene clusters for the melibiose-galactose pathway in *Torulaspora*

Anjan Venkatesh, Anthony L. Murray, Aisling Y. Coughlan, Kenneth H. Wolfe*

UCD Conway Institute and School of Medicine, University College Dublin, Dublin 4, Ireland

*To whom correspondence should be addressed. kenneth.wolfe@ucd.ie

Abstract

In many yeast species the three genes at the center of the galactose catabolism pathway, *GAL1*, *GAL10* and *GAL7*, are neighbors in the genome and form a metabolic gene cluster. We report here that some yeast strains in the genus *Torulaspora* have much larger *GAL* clusters that include genes for melibiase (*MEL1*), galactose permease (*GAL2*), glucose transporter (*HGT1*), phosphoglucomutase (*PGM1*), and the transcription factor *GAL4*, in addition to *GAL1*, *GAL10*, and *GAL7*. Together, these 8 genes encode almost all the steps in the pathway for catabolism of extracellular melibiose (a disaccharide of galactose and glucose). We show that a progenitor 5-gene cluster containing *GAL 7-1-10-4-2* was likely present in the common ancestor of *Torulaspora* and *Zygotorulaspora*. It added *PGM1* and *MEL1* in the ancestor of most *Torulaspora* species. It underwent further expansion in the *T. pretoriensis* clade, involving the fusion of three progenitor clusters in tandem and the gain of *HGT1*. These giant *GAL* clusters are highly polymorphic in structure, and subject to horizontal transfers, pseudogenization and gene losses. We identify recent horizontal transfers of complete *GAL* clusters from *T. franciscae* into one strain of *T. delbrueckii*, and from a relative of *T. maleeae* into one strain of *T. globosa.* The variability and dynamic evolution of *GAL* clusters in *Torulaspora* indicates that there is strong natural selection on the *GAL* pathway in this genus.

Introduction

Physical clusters of genes that function in the same process or metabolic pathway are relatively rare in yeasts (Riley et al., 2016; Rokas et al., 2018), but in budding yeasts (Saccharomycotina) the known examples include gene clusters for the pathways *NIT* (nitrate assimilation (Ávila et al., 2002)), *PUL* (pulcherrimin synthesis (Krause et al., 2018)), *NAG* (N-acetyl glucosamine catabolism (Yamada-Okabe et al., 2001)), *LAC* (lactose utilization (Varela et al., 2019)), *DAL* (allantoin degradation (Wong and Wolfe, 2005)), *MAL* (maltose utilization (Viigand et al., 2018)), and *GAL* (galactose utilization (Slot and Rokas, 2010)). The *GAL* pathway is one of the most intensively studied systems in yeast genetics. The canonical *GAL* gene cluster was first characterized in *Saccharomyces cerevisiae*, where it consists of three genes (*GAL1*, *GAL10* and *GAL7*) that code for the pathway to convert intracellular β-D-galactose to glucose-1-phosphate (Fig. 1) (Douglas and Hawthorne, 1964; St John and Davis, 1981). The same three genes are clustered in the same order in *Kluyveromyces lactis* (Webster and Dickson, 1988) and most other species in the family Saccharomycetaceae. A similar cluster of *GAL 1-10-7*, interspersed with two genes of unknown function, occurs in *Candida albicans* and other species in the CUG-Ser1 clade (Slot and Rokas, 2010). In more divergent yeasts the *GAL* genes are generally not clustered, except for four genera (*Schizosaccharomyces*, *Nadsonia*, *Brettanomyces* and *Wickerhamomyces*) that gained clusters by horizontal transfer from donors in the CUG-Ser1 clade, and two genera (*Cryptococcus* and *Lipomyces*) in which *GAL* clusters appear to have formed independently (Slot and Rokas, 2010; Haase et al., 2020).

It is widely thought that clustering of metabolic genes evolves as a mechanism for co-regulating the expression of genes, and that clustering can be selected for if an intermediate metabolite in the pathway is toxic – as is the case for galactose-1-phosphate in the *GAL* pathway – so that is it important to coordinate synthesis and removal of the toxin (McGary et al., 2013). The local order of genes within clusters often varies among species (Wong and Wolfe, 2005; Slot and Rokas, 2010; Naseeb and Delneri, 2012), and it is common to find that genes that are in a cluster in one species are completely absent from the genome in others (Hittinger et al., 2004; Wolfe et al., 2015). It is also common to find that the metabolic pathways encoded by clustered genes show presence/absence polymorphism within a

species: for example, the *GAL* genes (including the *GAL 1-10-7* cluster but also the unclustered genes *GAL4*, *GAL2* and *GAL80*) are intact in some populations of *S. kudriavzevii* but pseudogenes in others (Hittinger et al., 2010).

We previously reported that the genome sequence of the type strain of *Torulaspora delbrueckii* (CBS1146^{T}) contains a large cluster of *GAL* genes, occupying 22 kb near a telomere of chromosome 5 (Wolfe et al., 2015). As well as *GAL10* (2 copies), *GAL1* (2 copies) and *GAL7* (1 copy), the cluster also contained predicted genes *MEL1* (melibiase), *GAL2* (galactose permease), *PGM1* (phosphoglucomutase), *GAL4* (transcription factor) and *HGT1* (high-affinity glucose transporter, orthologous to *K. lactis HGT1* (Billard et al., 1996)). The genes in this cluster appeared to code for additional steps in the *GAL* pathway, both upstream and downstream of the steps encoded by the canonical *GAL1-10-7* cluster (Fig. 1). In the extended pathway, extracellular melibiose (a disaccharide) is hydrolyzed into its constituent monosaccharides β-D-galactose and D-glucose by secreted Mel1 enzyme (melibiase, an α(1,6)-galactosidase). The monosaccharides are then imported across the plasma membrane by Gal2 (for galactose) and Hgt1 (for glucose). The galactose is processed by the Gal10, Gal1 and Gal7 enzymes to yield glucose-1-phosphate, which is then converted to glucose-6-phosphate by Pgm1. A second molecule of glucose-6-phosphate is made by importing the glucose and phosphorylating it by hexokinase (Hxk1) or glucokinase (Glk1). The two molecules of glucose-6-phosphate then enter the glycolytic pathway. Thus, the *T. delbrueckii* gene cluster appeared to contain genes for all the steps needed to convert melibiose into two molecules of glucose-6-phosphate, except for hexokinase/glucokinase; there are *HXK1* and *GLK1* genes in the *T. delbrueckii* genome but they are not in the cluster. The *T. delbrueckii* cluster also contains an ortholog of *S. cerevisiae GAL4*, the transcription factor that positively regulates expression of the other *GAL* genes (Hittinger et al., 2004).

In this study, we used genome sequences from additional species and strains of *Torulaspora*, generated in other studies (Galeote et al., 2018; Shen et al., 2018; Coughlan et al., 2020), to investigate the origin and evolution of *GAL* clusters in *Torulaspora* and related genera. We find that the large *GAL* cluster in the type strain of *T. delbrueckii* is atypical of this species, because all 14 other *T. delbrueckii* strains that we examined have no cluster, and we show that the cluster in the type strain of *T. delbrueckii* was acquired from

96 *T. franciscae* recently by horizontal gene transfer. We also uncovered an extraordinary
97 diversity of allelic *GAL* gene cluster structures in *T. pretoriensis*, and a rich history of cluster
98 expansion, fusion, and degeneration.
99
100
101 **Results**
102
103 **Phylogeny and phenotypes**
104
105 We examined genome sequences from multiple strains of *T. delbrueckii*, *T. pretoriensis* and
106 *T. globosa*, and from single strains of other *Torulaspora* species, as well as *Zygotorulaspora*
107 *mrakii*, *Zygotorulaspora florentina*, *Zygosaccharomyces rouxii*, *Kluyveromyces lactis* and
108 *S. cerevisiae*. The phylogeny of the species, and a summary of the major events we infer to
109 have occurred during *GAL* cluster evolution in *Torulaspora*, is shown in Figure 2. One gene in
110 the well-known *GAL* system of *S. cerevisiae*, *GAL3*, is a paralog of *GAL1* that was formed by
111 the whole-genome duplication (WGD). *Torulaspora* and all the other genera considered
112 here diverged from *S. cerevisiae* before the WGD occurred, so their *GAL1* genes are
113 orthologous to both *GAL1* and *GAL3* in *S. cerevisiae*. Another gene, *GAL80*, coding for a
114 corepressor of *GAL* gene expression, is absent from most *Torulaspora* species (Fig. 2).
115
116 A *GAL* cluster is present in at least some strains of all the *Torulaspora* species we studied.
117 We tested the ability of several strains to grow on solid media containing galactose,
118 melibiose, or glucose as a sole carbon source (Fig. 3). We found that the ability to grow on
119 galactose correlates with the presence of intact copies of the genes *GAL1*, *GAL10* and *GAL7*
120 in the genome, and the ability to grow on melibiose correlates with the presence of an
121 intact *MEL1* gene (Fig. 3). The starting point for our study was the large *GAL* cluster on
122 chromosome 5 of *T. delbrueckii* strain CBS1146^{T} (Wolfe et al., 2015), and we found that this
123 strain can grow on galactose whereas *T. delbrueckii* strain L09, which lacks the cluster,
124 cannot (Fig. 3). However, we were surprised to find that *T. delbrueckii* CBS1146^{T} cannot
125 grow on melibiose despite apparently having a *MEL1* gene. We realized that the open
126 reading frame we originally annotated as *MEL1* (*TDEL0E00170*) is truncated at the 5' end
127 relative to other *MEL1* genes. Comparison to a functional *MEL1* gene previously

128 characterized by Oda and Fukunaga (1999) from *T. delbrueckii* strain IFO1255 shows that
129 CBS1146^{T} has a TGG (Trp) -> TGA (stop) mutation at codon 38 which removes the region
130 coding for the secretion signal, so the *MEL1* gene of CBS1146^{T} is a pseudogene. A second
131 discrepancy between genotypes and phenotypes occurs in *T. pretoriensis* CBS2187^{T}, which
132 grows poorly on galactose despite containing *GAL1*, *GAL10* and *GAL7* genes (Fig. 3). This
133 discrepancy is discussed later.
134
135
136 **Synteny relationships**
137
138 Synteny comparisons among the *Torulaspora* species and outgroups revealed a complex
139 pattern of relationships and gene relocations (Fig. 2). For some loci, we refer to the
140 Ancestral gene numbering system of Gordon et al. (2009), which numbers genes
141 sequentially along the 8 chromosomes inferred to have existed just prior to the WGD, for
142 example locus Anc_8.123 is the 123rd gene along Ancestral chromosome 8. This numbering
143 system is also used in our Yeast Gene Order Browser (ygob.ucd.ie) (Byrne and Wolfe, 2005).
144
145 In the outgroup species shown at the bottom of Figure 2 (*S. cerevisiae*, *K. lactis*, *Z. rouxii*),
146 the only genes in the *GAL* pathway that are clustered are *GAL1*, *GAL10* and *GAL7*, and they
147 occur in the order *GAL 1-10-7*. This arrangement is conserved in *T. microellipsoides*,
148 including the flanking genes *SNQ2* and *RPT2* (Anc_3.216 to Anc_3.220). This cluster is at an
149 internal chromosomal site in these species, i.e. it is not subtelomeric. In the outgroups, the
150 other genes in the pathway are at conserved, dispersed, places in the genome (*PGM1* =
151 Anc_2.445; *GAL4* = Anc_6.279; *HGT1* = Anc_1.432; *GAL80* = Anc_1.500), and *MEL1* is not
152 present at all.
153
154
155 **Formation of a large *GAL* cluster in the common ancestor of *Torulaspora* and**
156 ***Zygotorulaspora***
157
158 In *Zygotorulaspora mrakii*, the cluster has expanded to 6 genes: it contains *GAL 7-1-10-4-2*
159 and a *PGM1* gene (Fig. 2). *Z. mrakii* also has an unlinked *MEL1* gene, which was previously

shown to be functional by Oda and Fujisawa (2000). The 6-gene cluster has gained genes for the pathway steps upstream (*GAL2*) and downstream (*PGM1*) of the steps encoded by the 3-gene cluster, as well as gaining the transcription factor *GAL4*. It is interesting that the order of the 3 genes has also changed, from *GAL 1-10-7* in the outgroups to *GAL 7-1-10* in *Z. mrakii*. The *Z. mrakii* 6-gene cluster is located at an internal chromosomal site between *EST3* (Anc_7.128) and *URM1* (Anc_7.129). The cluster therefore appears to have become inserted between two genes that were ancestrally neighbors. In the genome assembly of a second *Zygotorulaspora* species, *Z. florentina* (accession number PPJY02000000), the same six genes are found on three small contigs: one containing only *PGM1-GAL7-GAL1*, one containing only *GAL10*, and one containing only *GAL4-GAL2*, so it is unclear whether *Z. florentina* has a *GAL* cluster organization identical to that in *Z. mrakii* or a more fragmented organization.

In *T. maleeae*, there is a 7-gene cluster with identical gene order to the 6-gene cluster of *Z. mrakii*, plus *MEL1* (Fig. 2). This cluster appears to be at a subtelomeric location, and the *EST3* and *URM1* genes (Anc_7.128/7.129) are adjacent in *T. maleeae*. Both *T. maleeae* and the two *Zygotorulaspora* species have two *PGM1* genes. The first, designated *PGM1_anc*, is at the ancestral *PGM1* location (Anc_2.445). It is syntenic with the *PGM1* genes of other yeasts, including the *PGM1/PGM2* gene pair of *S. cerevisiae*, which is a WGD pair. The second, designated *PGM1_dup*, is a duplicated copy of *PGM1* located in the *GAL* cluster.

The gene order *GAL 7-1-10-4-2*, as seen in *Z. mrakii* and *T. maleeae*, is a pattern that recurs throughout the *GAL* clusters of most *Torulaspora* species that will be described in the following sections. However, *T. microellipsoides* has an ancestral-type cluster (*GAL 1-10-7*) at the ancestral location (Anc_3.219), rather than the *GAL 7-1-10-4-2* pattern, even though phylogenomic analysis (Shen et al., 2018) has indicated that the genus *Torulaspora* is monophyletic and *Zygotorulaspora* is an outgroup to it. *T. microellipsoides* also has a *MEL1* gene at an unlinked, non-telomeric location (Fig. 2).

The organization of *GAL* genes in *T. microellipsoides* resembles the outgroup species more closely than it resembles other *Torulaspora* species, whereas the *Z. mrakii* organization resembles *Torulaspora* species (Fig. 2). In phylogenetic trees of individual *GAL* genes,

192 *T. microellipsoides* is often placed outside *Zygotorulaspora* (Fig. 4), in contrast to the
193 phylogenomic tree. Moreover, *GAL80* is present in *T. microellipsoides* but absent in the
194 other *Torulaspora* species and in *Zygotorulaspora* (Fig. 2). Together, these results suggest
195 that the phylogenomic tree might be incorrect regarding the branching order of
196 *Zygotorulaspora* and *T. microellipsoides*. Alternatively, there may have been horizontal
197 transfer of a *GAL* cluster between the *Zygotorulaspora* and *Torulaspora* branches in either
198 direction, after *T. microellipsoides* diverged from the rest of the genus *Torulaspora*, making
199 the *GAL* phylogeny different from the phylogeny of the rest of the genome.
200
201 In summary, the point of origin of the *GAL 7-1-10-4-2* cluster pattern is not fully clear, but it
202 appears to have been present in the common ancestor of the genera *Zygotorulaspora* and
203 *Torulaspora*. It is first seen with *PGM1* at one end, and later gained *MEL1* at the other end.
204
205
206 **Horizontal *GAL* cluster transfer into one strain of *T. globosa***
207
208 *T. globosa* is a sister species to *T. maleeae*. We sequenced the genomes of 12 strains of
209 *T. globosa* (Coughlan et al., 2020 and A.Y.C. and K.H.W., unpublished) and found that 11 of
210 them, including the type strain CBS764^T, have no *GAL* genes. However, one strain, *T. globosa*
211 NRRL YB-1481, has a *GAL* cluster, and the organization of this cluster is very similar to the
212 *T. maleeae* cluster (Fig. 2). Phylogenetic trees of *GAL 7*, *1*, *10*, *4*, *2* and *MEL1* all show that
213 the *T. globosa* NRRL YB-1481 genes group with the *T. maleeae* genes (Fig. 4). In plate tests,
214 *T. globosa* NRRL YB-1481 was able to grow on melibiose and galactose, whereas *T. globosa*
215 CBS764^T could not (Fig. 3).
216
217 Interestingly, the *GAL* cluster in *T. globosa* strain NRRL YB-1481 has formed at the ancestral
218 location of *PGM1* (Anc_2.445; Fig. 2). This strain has only one *PGM1* gene, in contrast to
219 *T. maleeae* and *Z. mrakii* which have two (*PGM1_anc* and *PGM1_dup*). Since most
220 *T. globosa* strains have no *GAL* genes, the most plausible scenario to explain the presence of
221 a cluster in NRRL YB-1481 is that it originated by horizontal transfer. In view of the relatively
222 low DNA sequence identity (74%) between the *T. globosa* NRRL YB-1481 and *T. maleeae*

223 clusters, the donor is more likely to have been an unidentified species related to
224 *T. maleeae*/*T. globosa*, rather than *T. maleeae* itself.
225
226 Although it is possible that recombination between the *PGM1* genes in the donor cluster
227 and the recipient *T. globosa* NRRL YB-1481 genome might have guided integration of the
228 cluster, this seems unlikely because the *T. maleeae* and *T. globosa PGM1* genes are
229 currently in opposite orientations relative to their neighbor *GAL7* (Fig. 2). Also, a
230 phylogenetic tree of *PGM1* sequences (Fig. 4) places the single, cluster-associated, *PGM1* of
231 *T. globosa* NRRL YB-1481 at the position expected for a *PGM1_anc* gene: it is in a clade with
232 the single *PGM1* gene of *T. globosa* CBS764^T and *T. maleeae PGM1_anc*, and far away from
233 *T. maleeae PGM1_dup* which lies in a clade with *PGM1_dup* genes from *Z. mrakii* and
234 *Z. florentina.*
235
236
237 **Horizontal *GAL* cluster transfer from *T. franciscae* into *T. delbrueckii***
238
239 *T. pretoriensis*, *T. franciscae* and *T. delbrueckii* form a clade of three species whose *GAL*
240 clusters, when present, are greatly expanded and contain numerous *GAL* pseudogenes as
241 well as functional genes. We analyzed data from multiple strains of *T. delbrueckii* and
242 *T. pretoriensis*, but we have only one genome sequence from *T. franciscae* (the type strain,
243 CBS2926^T).
244
245 In the set of 15 *T. delbrueckii* strains that we analyzed, none except CBS1146^T contains a
246 *GAL* cluster, which suggests that the cluster was gained by horizontal transfer. The CBS1146^T
247 cluster is identical in gene organization to a cluster in the type strain of *T. franciscae*, and
248 the two clusters have 97% DNA sequence identity over 22 kb. The similarity between these
249 two species is much higher than between either of them and *T. pretoriensis*, even though
250 *T. pretoriensis* is a sister species to *T. franciscae* (Fig. 2). Phylogenetic trees from individual
251 genes in the cluster consistently place *T. delbrueckii* CBS1146^T beside *T. franciscae* (Fig. 4).
252 We therefore infer that horizontal transfer occurred from *T. franciscae* to *T. delbrueckii*.
253 Curiously, although the cluster is near a telomere in both species, the two species have
254 opposite orientations of the cluster relative to the telomere (Fig. 2).

255

256 The *MEL1* genes in the clusters in the type strains of both *T. franciscae* and *T. delbrueckii* are
257 pseudogenes, and these strains are unable to grow on melibiose but able to grow on
258 galactose (Fig. 3). In a previous study by Oda and Tonomura (1996), 12 of 28 *T. delbrueckii*
259 strains examined, including the type strain, were found to be able to grow on galactose.
260 Only one of the *T. delbrueckii* strains (IFO 1255) could grow on melibiose as well as
261 galactose and was shown to have an intact *MEL1* gene (Oda and Tonomura, 1996; Oda and
262 Fukunaga, 1999).

263

264

265 **Extensive structural polymorphism of *T. pretoriensis GAL* clusters**

266

267 We analyzed genome sequences from nine strains of *T. pretoriensis*, of which five have large
268 and variable *GAL* clusters, and the other four have none. The four strains without clusters
269 (CBS11100, CBS11121, CBS11123, CBS11124) are closely related to each other, so only
270 CBS11100 is shown in Figure 2. Among the five strains with clusters, there is extensive
271 structural polymorphism, with only two strains (CBS2187^{T} and CBS9333) having similar
272 organization. All the *GAL* clusters in *T. pretoriensis* strains appear to be near telomeres.

273

274 The most complex *GAL* cluster in *T. pretoriensis* is in strain UWOPS 83-1046.2 (Fig. 2; we
275 refer to this strain hereafter as UWOPS). It spans 42 kb and contains 8 intact genes and 8
276 pseudogenes related to galactose metabolism. It also contains 2 unrelated genes and 1
277 unrelated pseudogene, which appear to be of subtelomeric origin. These unrelated genes
278 occupy a region of 15 kb inside the cluster and divide it into two parts, left and right. The
279 right part is almost identical in gene organization to the large *GAL* cluster that was
280 transferred between *T. franciscae* and *T. delbrueckii* CBS1146^{T}, the only differences being
281 some genes that are pseudogenes in *T. pretoriensis* UWOPS but intact in *T. franciscae* and
282 *T. delbrueckii* CBS1146^{T}, or vice versa (*HGT1*, *MEL1*, and one copy each of *GAL1* and *GAL10*;
283 Fig. 2). Phylogenetic analysis of the genes in this region (Fig. 4) shows that, in all cases,
284 *T. franciscae* and *T. delbrueckii* CBS1146^{T} form a clade with *T. pretoriensis* UWOPS outside,
285 which contradicts the expected species phylogeny (Fig. 2) and supports the hypothesis of
286 horizontal transfer between *T. franciscae* and *T. delbruckii*.

287

288 We tested the phenotypes of four *T. pretoriensis* strains (Fig. 3). As expected, only UWOPS
289 can grow on melibiose – it is the only strain with intact *MEL1*. On galactose, CBS11100
290 cannot grow (it has no *GAL* cluster), CBS5080 and UWOPS grow well, and the type strain
291 CBS2187^T grows more slowly. The poor growth of the type strain of *T. pretoriensis* on
292 galactose is consistent with previous studies. Oda and colleagues reported that
293 fermentation of galactose or melibiose by strain YK-1, which is a non-sedimenting derivative
294 of *T. pretoriensis* CBS2187^T (syn. IFO 10218), was undetectable after 2 days, whereas
295 *T. pretoriensis* CBS5080 (IFO 0022) and *T. franciscae* CBS2926^T (IFO 1360) fermented
296 galactose but not melibiose (Oda and Tonomura, 1993; Oda and Tonomura, 1996). Oda's
297 results are consistent with our results in Figure 3, except that we find that growth of
298 CBS2187^T on galactose is slow rather than absent. A possible reason for the poor growth is
299 that there is no *GAL2* galactose transporter gene anywhere in the *T. pretoriensis* CBS2187^T
300 genome; it is the only strain tested in Figure 3 that has the *GAL* enzyme genes without the
301 transporter gene.

302

303

304 **Cluster expansion by tandem triplication of progenitor *GAL 7-1-10-4-2* clusters**

305

306 Closer examination of the *T. pretoriensis GAL* clusters shows that they have an internal
307 structure that is based on tandem triplication of the *GAL 7-1-10-4-2* pattern mentioned
308 earlier. This structure is most clearly seen in *T. pretoriensis* UWOPS which has three copies
309 of the pattern: including pseudogenes, it has *GAL 7-1-10-4-2* in the left part of the cluster,
310 and *GAL 7-1-10-4* (without *GAL2*) followed by *GAL 7-1-10-2* (without *GAL4*) in the right part.
311 The other genes in the cluster (*HGT1*, *MEL1*, *PGM1*, and the unrelated genes between the
312 left and right parts) are located at the junctions between these three copies of the pattern.

313

314 This arrangement suggests that the large UWOPS cluster was formed by tandem fusion of
315 three smaller progenitor clusters that we designate L, R1 and R2, corresponding to the left
316 part and two sections of the right part of the current cluster (Fig. 2). We postulate that L
317 contained *GAL 7-1-10-4-2*, R1 originally contained *HGT1 – GAL 7-1-10-4-2*, and R2 originally
318 contained *MEL1 – GAL 7-1-10-4-2 – PGM1*. Subsequently, many of the triplicated *GAL* gene

319 copies became pseudogenes or relics (very short pseudogenes), and no trace remains of
320 *GAL2* in R1 or *GAL4* in R2. Notably, although there are many pseudogenes in the
321 *T. pretoriensis* clusters (of all strains), there are no pseudogenes that indicate that *HGT1*,
322 *MEL1*, or *PGM1* was ever duplicated within the clusters; all the duplications are of *GAL*
323 genes. Therefore we suggest that the triple-size cluster did not arise by triplicating a single
324 progenitor cluster, but instead arose by fusion of three progenitor clusters that were similar
325 (containing *GAL 7-1-10-4-2*) but already different regarding their content of *HGT1*, *MEL1* and
326 *PGM1*.
327
328 The clusters in the other *T. pretoriensis* strains are smaller than in UWOPS but still
329 consistent with the hypothesis of cluster expansion by tandem fusion of progenitors. Strain
330 CBS2785 has an overall organization similar to UWOPS, but it has lost *MEL1* and adjacent
331 parts of R1 and R2. It has also sustained an inversion of *GAL1-10-4* in the L part, probably in
332 conjunction with the formation of an extra relic of *GAL7* that is also in inverted orientation.
333 Strain CBS5080 has parts L and R2 but not R1, and it also has additional *HGT1* and *GAL1*
334 genes to the right of R2. Strains CBS2187^{T} and CBS9333 have only part L and an additional
335 *GAL1* gene; they lack *MEL1*, *HGT1* and *PGM1* in the cluster and have only one *PGM1* gene in
336 their genomes (at the ancestral locus Anc_2.445). The phylogenies of most genes and
337 pseudogenes in the *T. pretoriensis* clusters (Fig. 4) generally support the relationships shown
338 in Figure 2, which are based on synteny as well as phylogenetic considerations. It is
339 impossible to infer the complete history of the *T. pretoriensis* clusters, but we can conclude
340 that (i) at least three progenitor clusters fused in tandem to form them, and (ii) they are
341 undergoing extensive within-species structural rearrangement and turnover.
342
343
344 **Vestigial *GAL* clusters and extra unclustered *GAL10* and *HGT1* genes**
345
346 The large *GAL* cluster in *T. delbrueckii* originated by horizontal transfer from *T. franciscae*.
347 Among our sequenced strains, it is only present in CBS1146^{T} and is located near a telomere
348 of chromosome 5. However, in addition, all 15 *T. delbrueckii* strains (including CBS1146^{T})
349 also contain an intact *GAL10* gene near a telomere of chromosome 7 (Fig. 2). It is located
350 beside four pseudogenes in the arrangement *HGT1 – GAL 7-1-10-4*, where *GAL10* is the only

351 intact gene, so it appears to be a remnant of a primordial *GAL* cluster that has almost
352 disappeared. Its structure is the same as the R1 primordial cluster inferred in *T. pretoriensis*.
353
354 Similarly, most strains of *T. pretoriensis* have an extra copy of *GAL10*, located near *HGT1* and
355 a telomere (Fig. 2). This *GAL10* gene is present even in strains such as CBS11100 that cannot
356 utilize galactose. Therefore, many strains of both *T. delbrueckii* and *T. pretoriensis* contain
357 *GAL10* but no other *GAL* genes. This situation has also been seen in other yeasts (Haase et
358 al., 2020) but its physiological significance is unknown.
359
360 An extra vestigial telomeric *GAL* cluster is also seen in *T. maleeae*, containing an intact *HGT1*
361 gene and pseudogenes of *GAL7* and *GAL1* (Fig. 2). Thus, in both *T. maleeae* and
362 *T. pretoriensis*, high-affinity glucose transporter function is provided by an *HGT1* gene that is
363 neither located at the ancestral *HGT1* locus (Anc_1.432), nor in an active *GAL* cluster
364 containing intact *GAL1* and *GAL7*, but in a remnant of a degraded cluster at a telomeric
365 location that sometimes also includes *GAL10*. Notably, in the only *T. pretoriensis* strain that
366 includes an intact *HGT1* in its *GAL* cluster (CBS5080), there are no additional telomeric *HGT1*
367 or *GAL10* genes (Fig 2).
368
369
370 **Discussion**
371
372 The *GAL* clusters of *Torulaspora* species are remarkably large and heterogeneous. There are
373 polymorphisms both for presence/absence of the cluster, and for gene order within the
374 cluster. Formation of pseudogenes is common. As a result, *Torulaspora* strains and species
375 vary in their ability to grow using galactose or melibiose as the sole carbon source. It is
376 difficult to correlate these differences with the ecology of the yeasts, because relatively
377 little is known about their natural environments. *T. delbrueckii* and *T. microellipsoides* are
378 frequently isolated from high-sugar anthropic environments such as food spoilage and
379 fermented fruit juices, whereas most isolates of *T. franciscae*, *T. pretoriensis*, *T. globosa*, and
380 *T. maleeae* come from soil (Kurtzman, 2011). For the two strains that gained *GAL* clusters by
381 horizontal transfer, *T. globosa* NRRL YB-1481 was isolated from soil in Ghana, and the origin
382 of the type strain of *T. delbrueckii* CBS1146^{T} is uncertain.

383

384 The cluster first expanded from a canonical 3-gene *GAL 1-10-7* structure by adding *GAL2* and
385 *GAL4*, around the time of the common ancestor of *Torulaspora* and *Zygotorulaspora*. The
386 synteny relationships in Figure 2 suggest that a duplicate copy of *PGM1* was then recruited
387 into the *GAL 7-1-10-4-2* cluster, followed later by relocation of *MEL1* and then *HGT1*.
388 However, the phylogeny of *PGM1* sequences (Fig. 4) shows that there must have been
389 multiple separate incorporations of *PGM1* into the cluster, because the *PGM1_dup* genes in
390 the giant *GAL* clusters of the *T. pretoriensis/T. delbrueckii/T. franciscae* clade originated
391 independently of the *PGM1_dup* genes in the smaller clusters of *T. maleeae* and
392 *Z. mrakii/Z. florentina*. Including the integration of a *GAL* cluster beside *PGM1_anc* in
393 *T. globosa* NRRL Y-1481, there were three separate, parallel, events of incorporation of
394 *PGM1* into *Torulaspora GAL* clusters – pointing to strong selection to incorporate it. In two
395 *Lachancea* species a *GAL* cluster including *GAL1, GAL7* and *GAL2* has formed beside *PGM1* at
396 its ancestral location (Kuang et al., 2018), similar to what we observe in *T. globosa* NRRL Y-
397 1481. *PGM1* is a bottleneck gene, coding for an enzyme that integrates metabolic flux from
398 several pathways including glycogen synthesis, trehalose synthesis and the pentose
399 phosphate pathway as well as the *GAL* pathway, and in the genera *Saccharomyces* and
400 *Lachancea*, regulation of *PGM1* by *GAL4* has been gained and lost multiple times (Kuang et
401 al., 2018). We find that in the species with two *PGM1* genes (Fig. 4), the *PGM1_dup* genes in
402 the cluster contain multiple putative Gal4 binding sites (CGG-N_{11}-CCG) in their upstream
403 regions, whereas the *PGM1_anc* genes do not. In *T. globosa* NRRL YB-1481, *PGM1* is not
404 duplicated but has Gal4 sites in the upstream region that it shares with *GAL7* (Fig. 2). Thus,
405 in all the clusters in the *Torulaspora* clade, a *PGM1* gene has come under the regulation of
406 *GAL4.*

407

408 Unexpectedly, our results indicate that duplication and fusion of whole clusters, rather than
409 duplication of individual genes, was the major mechanism of evolution of *GAL* clusters. In
410 *T. pretoriensis*, three primordial clusters fused to form one giant cluster and many of the
411 genes later became pseudogenes. Tandem fusion of clusters may have provided an
412 opportunity to experiment with shuffling the gene order, by allowing different gene copies
413 to become pseudogenes. For example, in the *T. pretoriensis* clusters, the intact gene
414 upstream of *GAL1* can be *GAL10, GAL2*, *GAL4*, or *MEL1* (Fig. 2). Haase et al. (2020) recently

415 identified a similar fusion of two *GAL* clusters (one ancestral and one horizontally
416 transferred) in *Nadsonia fulvescens*.
417
418 The *Torulaspora GAL* clusters include up to eight different functional genes, comprising the
419 whole *MEL-GAL-PGM* pathway except for hexokinase/glucokinase (Fig. 1). Since the sugar
420 kinases also function in the pathway for catabolism of glucose monomers imported into the
421 cell by hexose transporters, the eight genes in the cluster constitute the complete set of
422 genes that need to be activated in the presence of melibiose or galactose, and repressed in
423 their absence. In *K. lactis*, *HGT1* was originally described as a high-affinity glucose
424 transporter, but it can also transport galactose and is induced by galactose (Baruffini et al.,
425 2006).
426
427 To build clusters with eight functional genes by random genomic rearrangements, natural
428 selection on the *GAL* metabolic pathway must be exceptionally strong in *Torulaspora*.
429 However, we have no explanation for why selection to form clusters is stronger in
430 *Torulaspora* than in other budding yeast genera. It seems likely that regulatory changes,
431 involving duplication of *PGM1*, loss of *GAL80*, and movement of *GAL4* into the cluster were
432 central to expansion of the cluster. Previous work has shown that Gal4 became the major
433 regulator of the *GAL* pathway relatively recently, displacing Rtg1/Rtg3 in an ancestor of the
434 family Saccharomycetaceae (Choudhury and Whiteway, 2018; Haase et al., 2020). In the
435 *Torulaspora/Zygotorulaspora* clade, the further step of moving the *GAL4* gene into the
436 cluster has occurred. Relocation of *GAL4* into the cluster would have enabled the Gal4
437 protein to evolve in concert with its binding sites in the promoters of the nearby *GAL* genes.
438 Moreover, in the *Torulaspora/Zygotorulaspora* species (except *T. microellipsoides*), Gal4 has
439 lost the C-terminal region for interaction with the co-repressor Gal80 (Choudhury and
440 Whiteway, 2018), and the *GAL80* gene is absent from their genomes (Fig. 2). In each cluster,
441 multiple putative Gal4 binding sites are present upstream of each intact *GAL* gene (except
442 *GAL4*) as well as *PGM1* and *HGT1*, but not *MEL1*. These regulatory changes may have made
443 the cluster almost independent of other loci in the genome, and hence made it more
444 amenable to transfer among species.
445
446

447 **Methods**

448

449 Yeast strains were obtained from the Westerdijk Fungal Biodiversity Institute (CBS strains),
450 the USDA Agricultural Research Service (NRRL strains), Lallemand Inc. (L09), and M.-A.
451 Lachance (UWOPS 83-1046.2).

452

453 For growth tests, yeast strains were streaked onto agar plates made with YPD (2% dextrose)
454 (Formedium, catalog CCM0110), YNB (yeast nitrogen base; Sigma-Aldrich, 51483) with 2%
455 D-(+)-galactose (Sigma-Aldrich, G0625), or YNB with 2% D-(+)-melibiose (Sigma-Aldrich,
456 63630). Plates were incubated at 30° C for 48 hours before photographing.

457

458 For sequencing *T. globosa* strain NRRL YB-1481, cultures were grown under standard rich-
459 medium conditions. DNA was harvested from stationary-phase cultures by homogenization
460 with glass beads followed by phenol-chloroform extraction and ethanol precipitation.
461 Purified DNA was concentrated with the Genomic DNA Clean and Concentrator-10 (Zymo
462 Research, catalog D4010). Sequencing was done by BGI Tech Solutions (Hong Kong) using
463 Illumina HiSeq 4000 (paired end, 2 x 150 bp reads), and assembled using SPAdes version
464 3.11.1 (Bankevich et al., 2012). Coverage was approximately 85x. All other genome
465 sequences are from sources cited in Coughlan et al. (2020).

466

467 *GAL* clusters were annotated manually. In the *T. franciscae* genome assembly, the large
468 cluster was initially split into three contigs due to high similarity between the two *GAL10*
469 genes. Its organization was inferred by manually merging scaffold 86, scaffold 87, and contig
470 C4393.

471

472 Genes were inferred to be located in subtelomeric regions if the gene is near the end of a
473 chromosome-sized scaffold, or if DNA sequences neighboring the gene are repeat
474 sequences that occur only near the ends of multiple very large scaffolds, or if several
475 neighbors of the gene are members of gene families that are often found in subtelomeric
476 regions (Brown et al., 2010) and do not have Ancestral gene numbers (Gordon et al., 2009).

477

478 Phylogenetic trees were constructed from MUSCLE alignments of amino acid sequences,
479 using PhyML as implemented in version 5.0 of SeaView (Gouy et al., 2010). Approximate
480 translations of pseudogenes were made by manual annotation.
481
482
483

484 **Figure Legends**

485

486 **Figure 1.** The yeast biochemical pathway for catabolism of extracellular melibiose (Holden
487 et al., 2003). Colored backgrounds indicate genes that are located in clusters in *Torulaspora*
488 species. Gal10 has two distinct functions, mutarotase and epimerase, performed by two
489 domains of the protein. Hgt1 has been reported to transport galactose as well as glucose in
490 *K. lactis* (Baruffini et al., 2006).

491

492 **Figure 2.** Synteny relationships among *GAL* genes and clusters in *Torulaspora* species and
493 outgroups. Genes are labeled with their *GAL* gene number (7, 1, 10, 4, 2, or 80), or M
494 (*MEL1*), P (*PGM1*), or H (*HGT1*). Dashed borders on gene symbols indicate pseudogenes.
495 Gray backgrounds highlight groups of adjacent genes with the progenitor cluster gene order
496 *GAL 7-1-10-4-2* or subsets thereof. Large gray boxes indicate groups of genes that are at
497 syntenic locations in different strains/species, and are indicated as being either telomeric or
498 internal to chromosomes. Ancestral gene locations refer to the numbering system of
499 Gordon et al. (2009) and are internal to chromosomes. Different P symbols are used to
500 distinguish between *PGM1* genes at the ancestral location (*PGM1_anc*, dark brown), and
501 duplicate *PGM1* genes in *GAL* clusters (*PGM1_dup*, light brown). Tel indicates a region
502 inferred to be close to a telomere (subtelomeric), and zigzag symbols in *T. pretoriensis*
503 indicate intervening regions of 10-15 kb with no genes related to *GAL* metabolism. The tree
504 topology is from the phylogenomic analysis of Shen et al. (2018) with *T. globosa* added as in
505 (Saluja et al., 2012; Kaewwichian et al., 2020).

506

507 **Figure 3.** Growth of *Torulaspora* strains on galactose, melibiose, and glucose (YPD) media.
508 Plates were incubated at 30° C for 48 hours before photographing. The lower panel
509 indicates the presence or absence of intact genes in each genome.

510

511 **Figure 4.** Phylogenetic trees of *GAL*, *PGM1*, *HGT1* and *MEL1* genes. Branches are colored by
512 species. Some groups of closely related sequences have been collapsed (triangles).
513 Green/red braces mark gene pairs showing horizontal transfer between *T. franciscae* (TFRA)
514 and *T. delbrueckii* (TDEL) strain CBS1146^{T}. In the *PGM1* tree, gray rectangles indicate genes
515 that are located in *GAL* clusters, and for genomes with two *PGM1* genes the copies are

516 labeled *PGM1_anc* and *PGM1_dup*; other genomes have only one gene. Approximate
517 likelihood ratio test (aLRT) branch support values are shown.
518
519
520
521
522 **Acknowledgments**
523 This work was supported by the European Research Council (789341) and Science
524 Foundation Ireland (13/IA/1910).
525
526

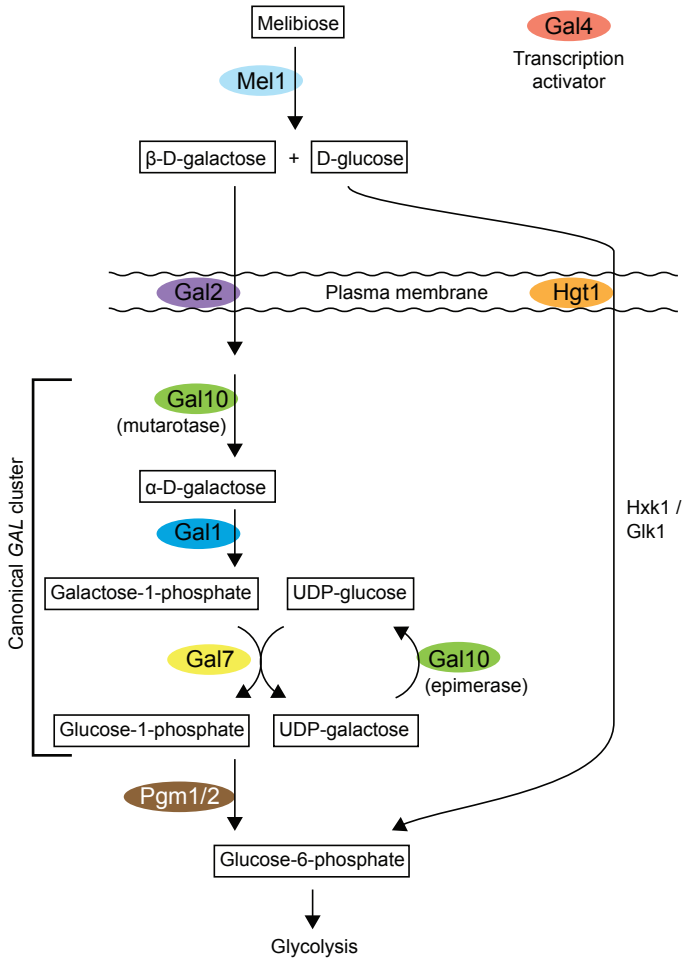
527 **References**

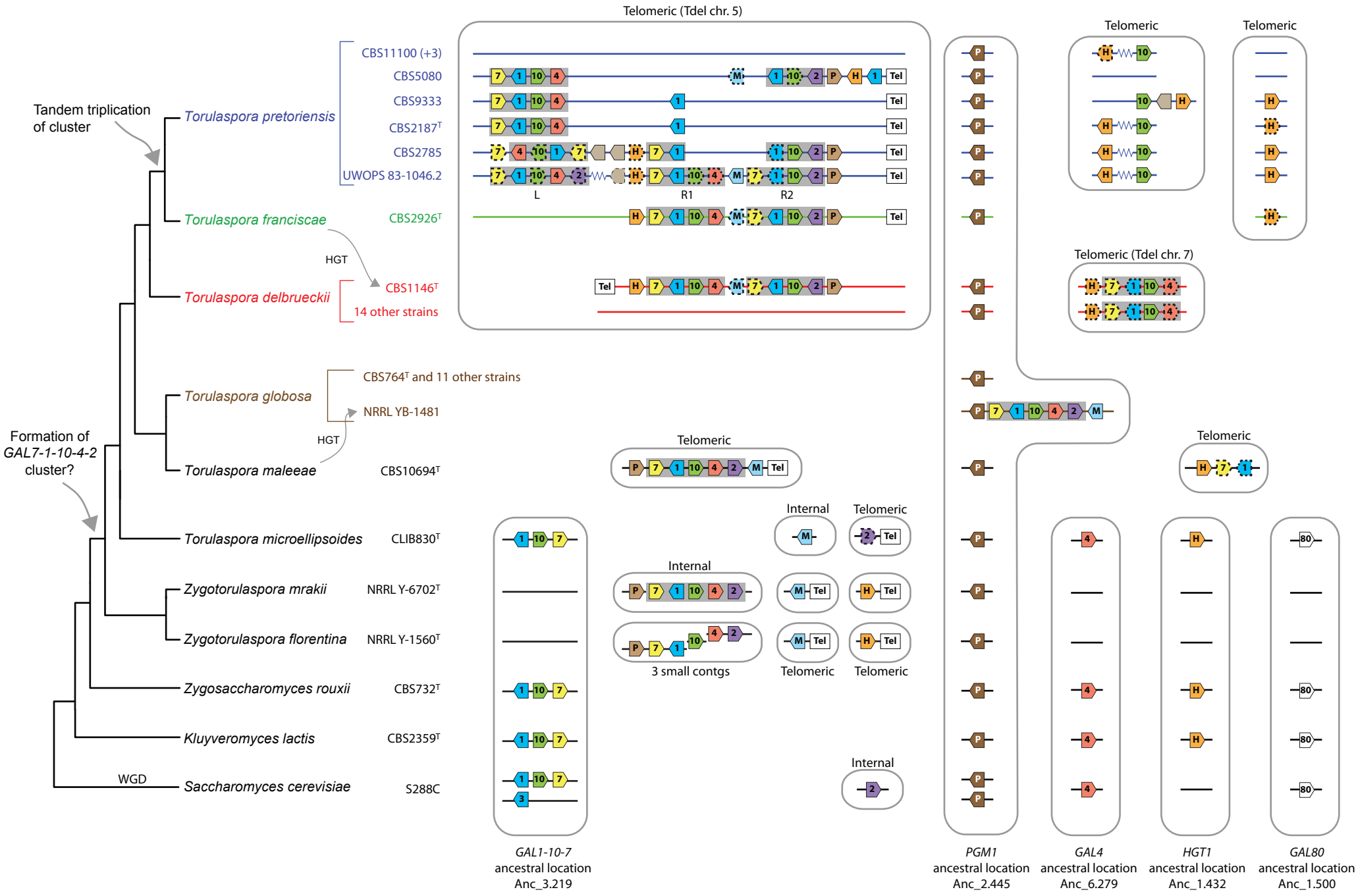
528

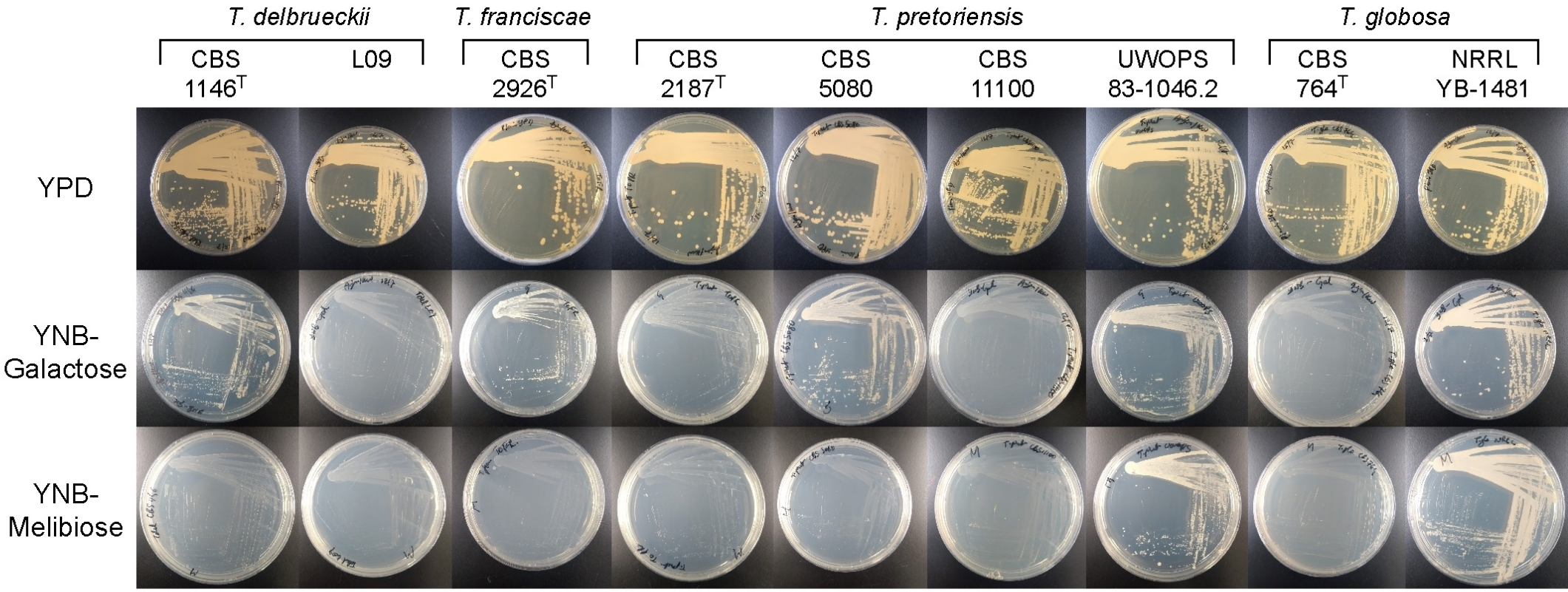
529 Ávila, J., González, C., Brito, N., Machín, M.F., Pérez, D., and Siverio, J.M. (2002). A second
530 Zn(II)(2)Cys(6) transcriptional factor encoded by the *YNA2* gene is indispensable for
531 the transcriptional activation of the genes involved in nitrate assimilation in the
532 yeast *Hansenula polymorpha*. Yeast *19*, 537-544.
533 Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M.,
534 Nikolenko, S.I., Pham, S., Prjibelski, A.D.*, et al.* (2012). SPAdes: a new genome
535 assembly algorithm and its applications to single-cell sequencing. J Comput Biol *19*,
536 455-477.
537 Baruffini, E., Goffrini, P., Donnini, C., and Lodi, T. (2006). Galactose transport in
538 *Kluyveromyces lactis*: major role of the glucose permease Hgt1. FEMS Yeast Res *6*,
539 1235-1242.
540 Billard, P., Menart, S., Blaisonneau, J., Bolotin-Fukuhara, M., Fukuhara, H., and Wesolowski-
541 Louvel, M. (1996). Glucose uptake in *Kluyveromyces lactis*: role of the *HGT1* gene in
542 glucose transport. J Bacteriol *178*, 5860-5866.
543 Brown, C.A., Murray, A.W., and Verstrepen, K.J. (2010). Rapid Expansion and Functional
544 Divergence of Subtelomeric Gene Families in Yeasts. Curr Biol *20*, 895-903.
545 Byrne, K.P., and Wolfe, K.H. (2005). The Yeast Gene Order Browser: combining curated
546 homology and syntenic context reveals gene fate in polyploid species. Genome Res
547 *15*, 1456-1461.
548 Choudhury, B.I., and Whiteway, M. (2018). Evolutionary Transition of *GAL* Regulatory Circuit
549 from Generalist to Specialist Function in Ascomycetes. Trends Microbiol *26*, 692-702.
550 Coughlan, A.Y., Lombardi, L., Braun-Galleani, S., Martos, A.A., Galeote, V., Bigey, F., Dequin,
551 S., Byrne, K.P., and Wolfe, K.H. (2020). The yeast mating-type switching
552 endonuclease HO is a domesticated member of an unorthodox homing genetic
553 element family. Elife *9*, e55336.
554 Douglas, H.C., and Hawthorne, D.C. (1964). Enzymatic Expression and Genetic Linkage of
555 Genes Controlling Galactose Utilization in Saccharomyces. Genetics *49*, 837-844.
556 Galeote, V., Bigey, F., Devillers, H., Ortiz-Merino, R.A., Dequin, S., Wolfe, K.H., and
557 Neuveglise, C. (2018). Genome Sequence of *Torulaspora microellipsoides* CLIB
558 830(T). Genome Announc *6*, e00615-00618.
559 Gordon, J.L., Byrne, K.P., and Wolfe, K.H. (2009). Additions, losses, and rearrangements on
560 the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces*
561 *cerevisiae* genome. PLoS Genet *5*, e1000485.
562 Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: A multiplatform graphical
563 user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol
564 *27*, 221-224.
565 Haase, M.A.B., Kominek, J., Opulente, D.A., Shen, X.-X., LaBella, A.L., Zhou, X., DeVirgilio, J.,
566 Hulfachor, A.B., Kurtzman, C.P., Rokas, A.*, et al.* (2020). Repeated horizontal gene
567 transfer of GALactose metabolism genes violates Dollo's law of irreversible loss.
568 BioRxiv.
569 Hittinger, C.T., Goncalves, P., Sampaio, J.P., Dover, J., Johnston, M., and Rokas, A. (2010).
570 Remarkably ancient balanced polymorphisms in a multi-locus gene network. Nature
571 *464*, 54-58.

572 Hittinger, C.T., Rokas, A., and Carroll, S.B. (2004). Parallel inactivation of multiple *GAL*
573 pathway genes and ecological diversification in yeasts. Proc Natl Acad Sci USA *101*,
574 14144-14149.
575 Holden, H.M., Rayment, I., and Thoden, J.B. (2003). Structure and function of enzymes of
576 the Leloir pathway for galactose metabolism. J Biol Chem *278*, 43885-43888.
577 Kaewwichian, R., Khunnamwong, P., Am-In, S., Jindamorakot, S., and Limtong, S. (2020).
578 *Torulaspora nypae sp. nov.*, a novel yeast species isolated from nipa (*Nypa fruticans*
579 Wurmb.) inflorescence sap in southern Thailand. Int J Syst Evol Microbiol *70*, 1112-
580 1116.
581 Krause, D.J., Kominek, J., Opulente, D.A., Shen, X.X., Zhou, X., Langdon, Q.K., DeVirgilio, J.,
582 Hulfachor, A.B., Kurtzman, C.P., Rokas, A.*, et al.* (2018). Functional and evolutionary
583 characterization of a secondary metabolite gene cluster in budding yeasts. Proc Natl
584 Acad Sci USA *115*, 11030-11035.
585 Kuang, M.C., Kominek, J., Alexander, W.G., Cheng, J.F., Wrobel, R.L., and Hittinger, C.T.
586 (2018). Repeated Cis-Regulatory Tuning of a Metabolic Bottleneck Gene during
587 Evolution. Mol Biol Evol *35*, 1968-1981.
588 Kurtzman, C.P. (2011). *Torulaspora* Lindner (1904). In The Yeasts, A Taxonomic Study, C.P.
589 Kurtzman, J.W. Fell, and T. Boekhout, eds. (Amsterdam: Elsevier), pp. 867-874.
590 McGary, K.L., Slot, J.C., and Rokas, A. (2013). Physical linkage of metabolic genes in fungi is
591 an adaptation against the accumulation of toxic intermediate compounds. Proc Natl
592 Acad Sci USA *110*, 11481-11486.
593 Naseeb, S., and Delneri, D. (2012). Impact of chromosomal inversions on the yeast *DAL*
594 cluster. PLoS One *7*, e42022.
595 Oda, Y., and Fujisawa, T. (2000). Nucleotide sequence of alpha-galactosidase *MEL* gene from
596 *Zygosaccharomyces mrakii*. Curr Microbiol *41*, 220-222.
597 Oda, Y., and Fukunaga, M. (1999). Isolation and characterization of *MELt* gene from
598 *Torulaspora delbrueckii* IFO 1255. Yeast *15*, 1797-1801.
599 Oda, Y., and Tonomura, K. (1993). Selection of a Novel Baking Strain from the *Torulaspora*
600 Yeasts. Biosci Biotech Bioch *57*, 1320-1322.
601 Oda, Y., and Tonomura, K. (1996). α-Galactosidase from the yeast *Torulaspora delbrueckii*
602 IFO 1255. Journal of Applied Bacteriology, *80*, 203–208.
603 Riley, R., Haridas, S., Wolfe, K.H., Lopes, M.R., Hittinger, C.T., Goker, M., Salamov, A.A.,
604 Wisecaver, J.H., Long, T.M., Calvey, C.H.*, et al.* (2016). Comparative genomics of
605 biotechnologically important yeasts. Proc Natl Acad Sci USA *113*, 9882-9887.
606 Rokas, A., Wisecaver, J.H., and Lind, A.L. (2018). The birth, evolution and death of metabolic
607 gene clusters in fungi. Nat Rev Microbiol *16*, 731-744.
608 Saluja, P., Yelchuri, R.K., Sohal, S.K., Bhagat, G., Paramjit, and Prasad, G.S. (2012).
609 *Torulaspora indica* a novel yeast species isolated from coal mine soils. Antonie Van
610 Leeuwenhoek *101*, 733-742.
611 Shen, X.X., Opulente, D.A., Kominek, J., Zhou, X., Steenwyk, J.L., Buh, K.V., Haase, M.A.B.,
612 Wisecaver, J.H., Wang, M., Doering, D.T.*, et al.* (2018). Tempo and mode of genome
613 evolution in the budding yeast subphylum. Cell *175*, 1533-1545 e1520.
614 Slot, J.C., and Rokas, A. (2010). Multiple *GAL* pathway gene clusters evolved independently
615 and by different mechanisms in fungi. Proc Natl Acad Sci USA *107*, 10136-10141.
616 St John, T.P., and Davis, R.W. (1981). The organization and transcription of the galactose
617 gene cluster of *Saccharomyces*. J Mol Biol *152*, 285-315.

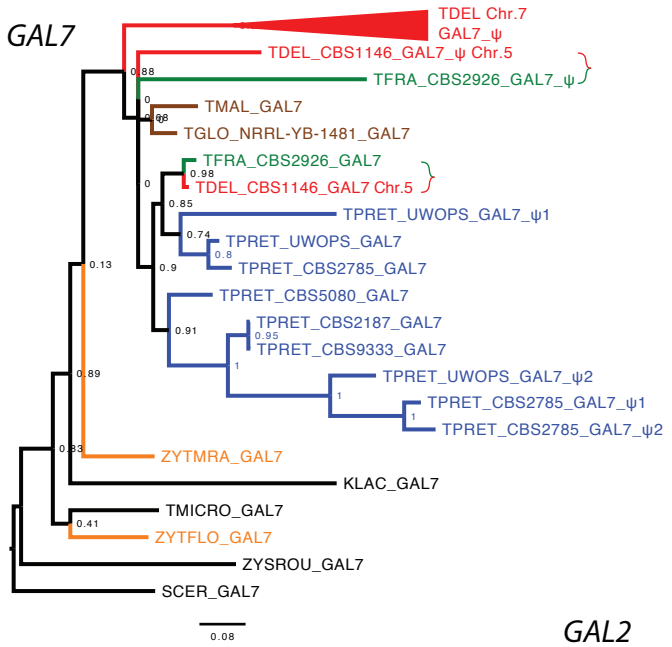
618 Varela, J.A., Puricelli, M., Ortiz-Merino, R.A., Giacomobono, R., Braun-Galleani, S., Wolfe,
619 K.H., and Morrissey, J.P. (2019). Origin of Lactose Fermentation in *Kluyveromyces*
620 *lactis* by Interspecies Transfer of a Neo-functionalized Gene Cluster during
621 Domestication. Curr Biol *29*, 4284-4290 e4282.
622 Viigand, K., Posnograjeva, K., Visnapuu, T., and Alamae, T. (2018). Genome Mining of Non-
623 Conventional Yeasts: Search and Analysis of MAL Clusters and Proteins. Genes
624 (Basel) *9*.
625 Webster, T.D., and Dickson, R.C. (1988). Nucleotide sequence of the galactose gene cluster
626 of *Kluyveromyces lactis*. Nucleic Acids Res *16*, 8192-8194.
627 Wolfe, K.H., Armisen, D., Proux-Wera, E., OhEigeartaigh, S.S., Azam, H., Gordon, J.L., and
628 Byrne, K.P. (2015). Clade- and species-specific features of genome evolution in the
629 Saccharomycetaceae. FEMS Yeast Res *15*, fov035.
630 Wong, S., and Wolfe, K.H. (2005). Birth of a metabolic gene cluster in yeast by adaptive gene
631 relocation. Nature Genet *37*, 777-782.
632 Yamada-Okabe, T., Sakamori, Y., Mio, T., and Yamada-Okabe, H. (2001). Identification and
633 characterization of the genes for N-acetylglucosamine kinase and N-
634 acetylglucosamine-phosphate deacetylase in the pathogenic fungus *Candida*
635 *albicans*. Eur J Biochem *268*, 2498-2505.
636

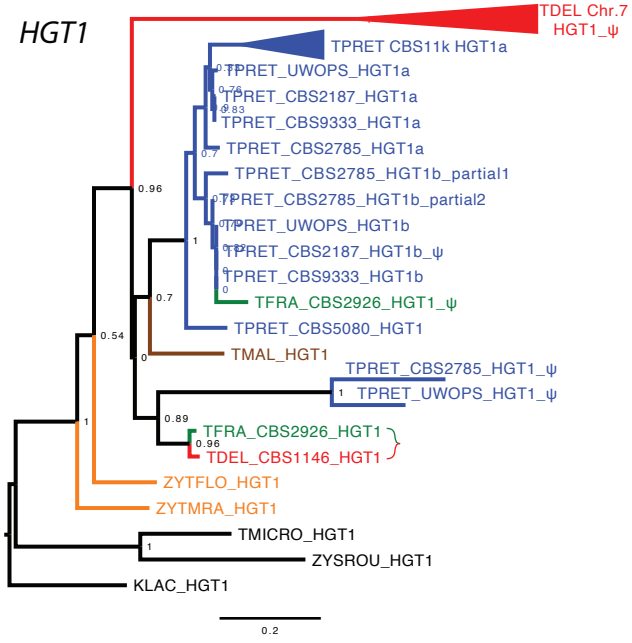

Melibiose
Gal4
Transcription activator
Mel1
β-D-galactose
+
D-glucose
Gal2
Plasma membrane
Hgt1
Gal10
(mutarotase)
α-D-galactose
Gal1
Galactose-1-phosphate
UDP-glucose
Gal7
Gal10
(epimerase)
Glucose-1-phosphate
UDP-galactose
Pgm1/2
Glucose-6-phosphate
Glycolysis
Canonical GAL cluster
Hxk1 /
Glk1

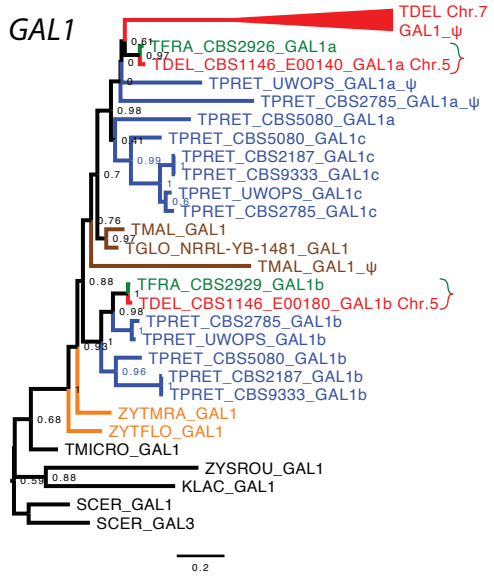

Tandem triplication of cluster
Formation of GAL7-1-10-4-2 cluster?
Torulaspora pretoriensis
CBS11100 (+3)
CBS5080
CBS9333
CBS2187T
CBS2785
UWOPS 83-1046.2
Torulaspora franciscae
CBS2926T
HGT
Torulaspora delbrueckii
CBS1146T
14 other strains
Torulaspora globosa
CBS764T and 11 other strains
NRRL YB-1481
HGT
Torulaspora maleeae
CBS10694T
Torulaspora microellipsoides
CLIB830T
Zygotorulaspora mrakii
NRRL Y-6702T
Zygotorulaspora florentina
NRRL Y-1560T
Zygosaccharomyces rouxii
CBS732T
Kluyveromyces lactis
CBS2359T
WGD
Saccharomyces cerevisiae
S288C
Telomeric (Tdel chr. 5)
L
R1
R2
Telomeric
Telomeric
Telomeric (Tdel chr. 7)
Telomeric
Telomeric
Internal
Telomeric
Internal
3 small contgs
Telomeric
Telomeric
Internal
GAL1-10-7 ancestral location Anc_3.219
PGM1 ancestral location Anc_2.445
GAL4 ancestral location Anc_6.279
HGT1 ancestral location Anc_1.432
GAL80 ancestral location Anc_1.500

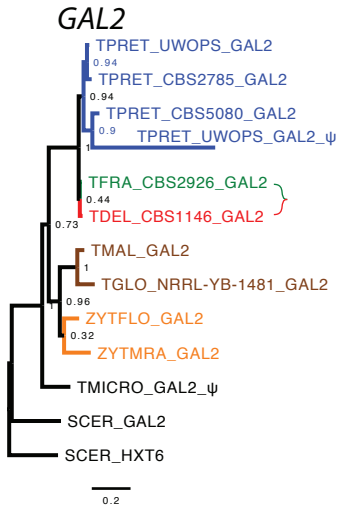


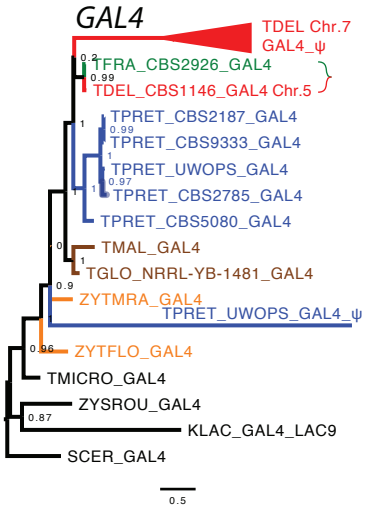
	CBS 1146T	L09	CBS 2926T	CBS 2187T	CBS 5080	CBS 11100	UWOPS 83-1046.2	CBS 764T	NRRL YB-1481
GAL1-10-7 genes	+	−	+	+	+	−	+	−	+
GAL2 gene	+	−	+	−	+	−	+	−	+
MEL1 gene	−	−	−	−	−	−	+	−	+

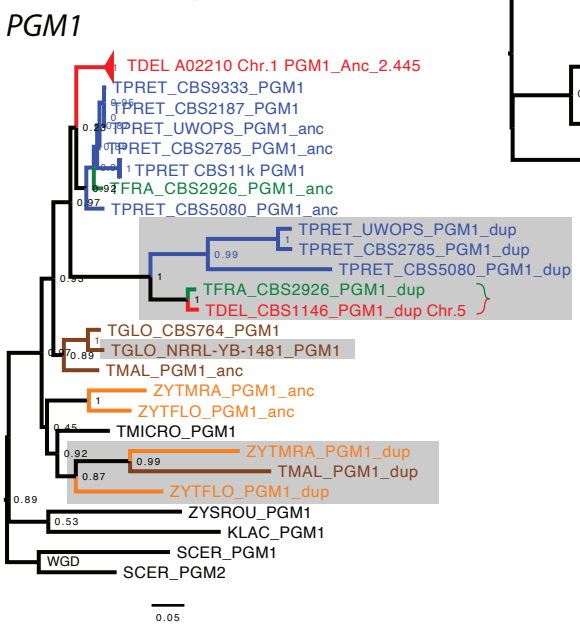

GAL7
TDEL Chr.7
GAL7_ψ
TDEL_CBS1146_GAL7_ψ Chr.5
TFRA_CBS2926_GAL7_ψ
TMAL_GAL7
TGLO_NRRL-YB-1481_GAL7
TFRA_CBS2926_GAL7
TDEL_CBS1146_GAL7 Chr.5
TPRET_UWOPS_GAL7_ψ1
TPRET_UWOPS_GAL7
TPRET_CBS2785_GAL7
TPRET_CBS5080_GAL7
TPRET_CBS2187_GAL7
TPRET_CBS9333_GAL7
TPRET_UWOPS_GAL7_ψ2
TPRET_CBS2785_GAL7_ψ1
TPRET_CBS2785_GAL7_ψ2
ZYTMRA_GAL7
KLAC_GAL7
TMICRO_GAL7
ZYTFLO_GAL7
ZYSROU_GAL7
SCER_GAL7
0.08

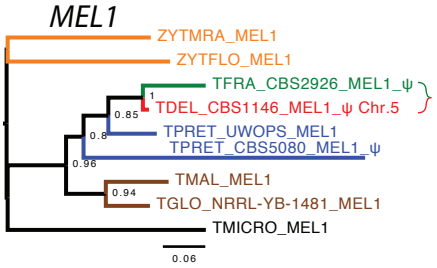

HGT1
TDEL Chr.7
HGT1_ψ
TPRET CBS11k HGT1a
TPRET_UWOPS_HGT1a
TPRET_CBS2187_HGT1a
TPRET_CBS9333_HGT1a
TPRET_CBS2785_HGT1a
TPRET_CBS2785_HGT1b_partial1
TPRET_CBS2785_HGT1b_partial2
TPRET_UWOPS_HGT1b
TPRET_CBS2187_HGT1b_ψ
TPRET_CBS9333_HGT1b
TFRA_CBS2926_HGT1_ψ
TPRET_CBS5080_HGT1
TMAL_HGT1
TPRET_CBS2785_HGT1_ψ
TPRET_UWOPS_HGT1_ψ
TFRA_CBS2926_HGT1
TDEL_CBS1146_HGT1
ZYTFLO_HGT1
ZYTMRA_HGT1
TMICRO_HGT1
ZYSROU_HGT1
KLAC_HGT1
0.2

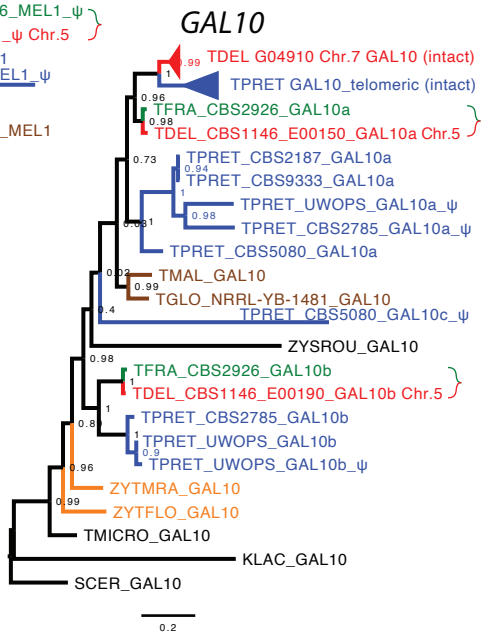

GAL1
TDEL Chr.7
GAL1_ψ
TFRA_CBS2926_GAL1a
TDEL_CBS1146_E00140_GAL1a Chr.5
TPRET_UWOPS_GAL1a_ψ
TPRET_CBS2785_GAL1a_ψ
TPRET_CBS5080_GAL1a
TPRET_CBS5080_GAL1c
TPRET_CBS2187_GAL1c
TPRET_CBS9333_GAL1c
TPRET_UWOPS_GAL1c
TPRET_CBS2785_GAL1c
TMAL_GAL1
TGLO_NRRL-YB-1481_GAL1
TMAL_GAL1_ψ
TFRA_CBS2929_GAL1b
TDEL_CBS1146_E00180_GAL1b Chr.5
TPRET_CBS2785_GAL1b
TPRET_UWOPS_GAL1b
TPRET_CBS5080_GAL1b
TPRET_CBS2187_GAL1b
TPRET_CBS9333_GAL1b
ZYTMRA_GAL1
ZYTFLO_GAL1
TMICRO_GAL1
ZYSROU_GAL1
KLAC_GAL1
SCER_GAL1
SCER_GAL3
0.2


GAL2
TPRET_UWOPS_GAL2
TPRET_CBS2785_GAL2
TPRET_CBS5080_GAL2
TPRET_UWOPS_GAL2_ψ
TFRA_CBS2926_GAL2
TDEL_CBS1146_GAL2
TMAL_GAL2
TGLO_NRRL-YB-1481_GAL2
ZYTFLO_GAL2
ZYTMRA_GAL2
TMICRO_GAL2_ψ
SCER_GAL2
SCER_HXT6
0.2


GAL4
TDEL Chr.7
GAL4_ψ
TFRA_CBS2926_GAL4
TDEL_CBS1146_GAL4 Chr.5
TPRET_CBS2187_GAL4
TPRET_CBS9333_GAL4
TPRET_UWOPS_GAL4
TPRET_CBS2785_GAL4
TPRET_CBS5080_GAL4
TMAL_GAL4
TGLO_NRRL-YB-1481_GAL4
ZYTMRA_GAL4
TPRET_UWOPS_GAL4_ψ
ZYTFLO_GAL4
TMICRO_GAL4
ZYSROU_GAL4
KLAC_GAL4_LAC9
SCER_GAL4
0.5


PGM1
TDEL A02210 Chr.1 PGM1_Anc_2.445
TPRET_CBS9333_PGM1
TPRET_CBS2187_PGM1
TPRET_UWOPS_PGM1_anc
TPRET_CBS2785_PGM1_anc
TPRET CBS11k PGM1
TFRA_CBS2926_PGM1_anc
TPRET_CBS5080_PGM1_anc
TPRET_UWOPS_PGM1_dup
TPRET_CBS2785_PGM1_dup
TPRET_CBS5080_PGM1_dup
TFRA_CBS2926_PGM1_dup
TDEL_CBS1146_PGM1_dup Chr.5
TGLO_CBS764_PGM1
TGLO_NRRL-YB-1481_PGM1
TMAL_PGM1_anc
ZYTMRA_PGM1_anc
ZYTFLO_PGM1_anc
TMICRO_PGM1
ZYTMRA_PGM1_dup
TMAL_PGM1_dup
ZYTFLO_PGM1_dup
ZYSROU_PGM1
KLAC_PGM1
SCER_PGM1
SCER_PGM2
WGD
0.05


MEL1
ZYTMRA_MEL1
ZYTFLO_MEL1
TFRA_CBS2926_MEL1_ψ
TDEL_CBS1146_MEL1_ψ Chr.5
TPRET_UWOPS_MEL1
TPRET_CBS5080_MEL1_ψ
TMAL_MEL1
TGLO_NRRL-YB-1481_MEL1
TMICRO_MEL1
0.06


GAL10
TDEL G04910 Chr.7 GAL10 (intact)
TPRET GAL10_telomeric (intact)
TFRA_CBS2926_GAL10a
TDEL_CBS1146_E00150_GAL10a Chr.5
TPRET_CBS2187_GAL10a
TPRET_CBS9333_GAL10a
TPRET_UWOPS_GAL10a_ψ
TPRET_CBS2785_GAL10a_ψ
TPRET_CBS5080_GAL10a
TMAL_GAL10
TGLO_NRRL-YB-1481_GAL10
TPRET_CBS5080_GAL10c_ψ
ZYSROU_GAL10
TFRA_CBS2926_GAL10b
TDEL_CBS1146_E00190_GAL10b Chr.5
TPRET_CBS2785_GAL10b
TPRET_UWOPS_GAL10b
TPRET_UWOPS_GAL10b_ψ
ZYTMRA_GAL10
ZYTFLO_GAL10
TMICRO_GAL10
KLAC_GAL10
SCER_GAL10
0.2