

1 **Repeated horizontal gene transfer of *GAL*actose metabolism genes violates**

2 **Dollo's law of irreversible loss**

3

4 Max A. B. Haase^{1,2&}, Jacek Kominek^{1&}, Dana A. Opolente¹, Xing-Xing Shen^{3,4}, Abigail L.

5 LaBella³, Xiaofan Zhou^{3,5}, Jeremy DeVirgilio⁶, Amanda Beth Hulfachor¹, Cletus P.

6 Kurtzman^{6,7}, Antonis Rokas^{3*}, Chris Todd Hittinger^{1*}

7

8 ¹Laboratory of Genetics, Wisconsin Energy Institute, DOE Great Lakes Bioenergy

9 Research Center, Center for Genomic Science Innovation, J. F. Crow Institute for the

10 Study of Evolution, University of Wisconsin-Madison, Madison, Wisconsin, USA

11 ²Sackler Institute of Graduate Biomedical Sciences and Institute for Systems Genetics,

12 NYU Langone Health, New York, NY, USA

13 ³Department of Biological Sciences, Vanderbilt University, Nashville, TN, USA

14 ⁴State Key Laboratory of Rice Biology and Ministry of Agriculture Key Lab of Molecular

15 Biology of Crop Pathogens and Insects, Institute of Insect Sciences, Zhejiang

16 University, Hangzhou 310058, China

17 ⁵Guangdong Province Key Laboratory of Microbial Signals and Disease Control,

18 Integrative Microbiology Research Centre, South China Agricultural University, 510642

19 Guangzhou, China

20 ⁶Mycotoxin Prevention and Applied Microbiology Research Unit, National Center for

21 Agricultural Utilization Research, Agricultural Research Service, U.S. Department of

22 Agriculture, Peoria, IL 61604, USA

23 ⁷Deceased

24

25 &Equal authorship

26

27 *To whom correspondence should be addressed: cthittinger@wisc.edu and

28 antonis.rokas@vanderbilt.edu

29

30 Keywords: gene loss, evolution, *GAL* cluster, Dollo's law, horizontal gene transfer,

31 yeasts, lateral gene transfer

32 **Abstract**

33 Dollo's law posits that evolutionary losses are irreversible, thereby narrowing the
34 potential paths of evolutionary change. While phenotypic reversals to ancestral states
35 have been observed, little is known about their underlying genetic causes. The
36 genomes of budding yeasts have been shaped by extensive reductive evolution, such
37 as reduced genome sizes and the losses of metabolic capabilities. However, the extent
38 and mechanisms of trait reacquisition after gene loss in yeasts have not been
39 thoroughly studied. Here, through phylogenomic analyses, we reconstructed the
40 evolutionary history of the yeast galactose utilization pathway and observed widespread
41 and repeated losses of the ability to utilize galactose, which occurred concurrently with
42 the losses of *GAL*actose (*GAL*) utilization genes. Unexpectedly, we detected three
43 galactose-utilizing lineages that were deeply embedded within clades that underwent
44 ancient losses of galactose utilization. We show that at least two, and possibly three,
45 lineages reacquired the *GAL* pathway via yeast-to-yeast horizontal gene transfer. Our
46 results show how trait reacquisition can occur tens of millions of years after an initial
47 loss via horizontal gene transfer from distant relatives. These findings demonstrate that
48 the losses of complex traits and even whole pathways are not always evolutionary
49 dead-ends, highlighting how reversals to ancestral states can occur.

50

51 **Introduction**

52 Understanding the interactions between a species' phenotype, genotype, and
53 environment is a central goal of evolutionary biology. Of particular interest are the
54 mechanisms by which the environment selects for changes in phenotype and

55 subsequently genome content. Due to their remarkable physiological diversity, budding
56 yeasts are present in an extraordinary range of environments¹. Alongside robustly
57 characterized physiologies² and the availability of an unrivaled set of genome
58 sequences^{1,3,4}, budding yeasts provide a unique subphylum-level eukaryotic model for
59 studying the interplay between the genome, phenotype, and the environment.

60 Trait reversal is an intriguing phenomenon whereby the character state of a
61 particular evolutionary lineage returns to its ancestral state. For more than a century,
62 trait reversal after a loss event has been thought to be highly unlikely; Dollo's law of
63 irreversibility states that, once a trait is lost, it is unlikely for the same trait to be found in
64 a descendant lineage, thereby excluding certain evolutionary paths^{5,6}. Despite this
65 purist interpretation, many examples of apparent violations to Dollo's law have been
66 documented⁷⁻¹⁵, and it is clear that evolutionary processes sometimes break Dollo's
67 law¹⁶⁻¹⁸. Nonetheless, the molecular and genetic mechanisms leading to trait reversal
68 have only been determined in a few cases^{17,18}. For example, it was recently shown that
69 flower color reversal in a *Petunia* species was facilitated by the resurrection of a
70 pseudogene¹⁸. In this case, the reversal was temporally rapid, which is in agreement
71 with the hypothesis that traits flicker on and off during speciation¹⁶. These results
72 underscore that complex traits do indeed undergo reversal and help identify one
73 possible genetic mechanism for doing so. In other cases, traits have been reversed long
74 after the speciation process and long after pseudogenes are undetectable^{7,19}, raising
75 the question of how trait reversal can occur millions of years after the initial loss.

76 The Leloir pathway of galactose utilization in the model budding yeast
77 *Saccharomyces cerevisiae* (subphylum Saccharomycotina) is one of the most intensely

78 studied and well-understood genetic, regulatory, and metabolic pathways of any
79 eukaryote^{20–29}. Although its regulatory genes are unlinked, the *GAL* genes encoding the
80 three key catabolic enzymes (*GAL1*, *GAL7*, and *GAL10*) are present in a localized gene
81 cluster²⁵. A critical consequence of clustering genes in fungi is a marked increase in the
82 rate of gene loss^{22,25,30–32} and a striking increase in the incidence of horizontal gene
83 transfer (HGT) of those genes^{32,33}. The principal mode of evolution for the *GAL* gene
84 cluster has been differential gene loss from an ancestral species that possessed the
85 *GAL* genes in a cluster^{4,22,25,34}. In one case, the budding yeast *GAL* enzymatic gene
86 cluster was horizontally transferred into the fission yeast *Schizosaccharomyces pombe*
87 (subphylum Taphrinomycotina)²⁵. Nonetheless, this transferred cluster is not functional
88 in typical growth assays, suggesting *Sc. pombe* *GAL* cluster may not be deployed
89 catabolically or may respond to induction signals other than galactose³⁵. Dairy and
90 some other strains of *Saccharomyces cerevisiae* may have horizontally acquired a more
91 active, transcriptionally rewired *GAL* pathway from an unknown outgroup of the genus
92 *Saccharomyces*^{36,37}, or they may have preserved these two versions of the pathway
93 through extreme balancing selection³⁸, but trait reversal is highly unlikely under either
94 interpretation. Collectively, these prior observations suggest that both cis-regulatory
95 features and unlinked regulators play crucial roles in determining the function of
96 horizontally transferred genes. Due to the widespread loss of *GAL* genes and the
97 apparent ability for the *GAL* enzymatic gene cluster to be horizontally transferred intact,
98 we hypothesized that budding yeast *GAL* clusters might break Dollo's law under some
99 conditions.

100 To address this hypothesis, we explored the genetic content and phenotypic
101 capabilities of a diverse set of budding yeast genomes. Despite being deeply embedded
102 within clades that underwent ancient losses of galactose metabolism, the genera
103 *Brettanomyces* and *Wickerhamomyces* both contained representatives that could utilize
104 galactose. Analyses of their genome sequences revealed *GAL* gene clusters that
105 exhibited an unusually high degree of synteny with gene clusters in distantly related
106 species. Further analysis of the genome of *Nadsonia fulvescens* showed that it also
107 contains a *GAL* gene cluster that is remarkably similar to a distantly related species.
108 Through rigorous phylogenetic hypothesis testing, we found strong evidence for the
109 complete losses of the genes encoding the enzymes necessary for galactose
110 catabolism, followed by their reacquisitions via independent yeast-to-yeast HGT events
111 in at least two, and possibly three, cases. Genes lost in budding yeasts have been
112 regained via HGT from bacterial donors in several cases^{39–45}, but here we demonstrate
113 an exceptionally clear example of a complex trait and its corresponding genes being lost
114 and then regained to its ancestral eukaryotic form. We conclude that multiple distantly
115 related lineages of yeasts have circumvented evolutionary irreversibility, both at the
116 molecular and phenotypic level, via eukaryotic HGT and that evolutionary paths are not
117 absolutely constrained after trait loss.

118

119 **Results**

120 Genome selection and sequencing

121 To reconstruct the evolution of galactose metabolism in the budding yeast
122 subphylum Saccharomycotina, we first selected a set of genomes to analyze that

123 spanned the backbone of the subphylum^{3,4}. Next, we sequenced the genomes of five
124 additional species at strategically positioned branches: *Brettanomyces naardenensis*; a
125 yet-to-be described *Wickerhamomyces* species, *Wickerhamomyces* sp. UFMG-CM-
126 Y6624; *Candida chilensis*; *Candida cylindracea*; and *Candida silvatica*. All strains used
127 in this study can be found in Supplemental Table 1. Finally, we reconstructed a species-
128 level phylogeny, analyzing the genome sequences of 96 Saccharomycotina and 10
129 outgroup species (Supplemental Figures 1 and 2).

130

131 Recurrent loss of yeast *GAL* clusters

132 This dataset suggests that the *GAL* enzymatic gene cluster (hereafter *GAL*
133 cluster) of budding yeasts formed prior to the last common ancestor of the CUG-Ser1,
134 CUG-Ser2, CUG-Ala, Phaffomycetaceae, Saccharomycodaceae, and
135 Saccharomycetaceae major clades (Figure 1 and Supplemental Figure 3)²⁵. This
136 inference is supported by the presence of the fused bifunctional *GAL10* gene in these
137 lineages and the absence of the fused protein in species outside these lineages (Figure
138 1 and Supplemental Figure 3)²⁵. Since galactose metabolism has been repeatedly lost
139 over the course of budding yeast evolution and the enzymatic genes are present in a
140 gene cluster, we next asked whether the trait of galactose utilization had undergone trait
141 reversal. We reasoned that species or lineages who utilize galactose, but who are
142 deeply embedded in clades that predominantly cannot utilize galactose, would
143 represent prime candidates for possible trait reversal events. When we mapped both
144 *GAL* gene presence and galactose utilization onto our phylogeny (Figure 1 and
145 Supplemental Figure 3), we inferred repeated loss of the *GAL* gene clusters (Figure 1

146 and Supplemental Figure 3) and a strong association between genotype and phenotype
147 (Supplemental Table 2). However, we identified two genera, *Brettanomyces* and
148 *Wickerhamomyces*, as containing candidates for trait reversal (Figure 1). This unusual
149 trait distribution led us to consider the possibility that the *GAL* clusters of these two
150 lineages were not inherited vertically.

151

152 Unusual synteny patterns of *GAL* clusters

153 If the observed distribution of galactose metabolism were to be explained by only
154 vertical reductive evolution, then *GAL* cluster losses have occurred even more
155 frequently than currently appreciated. Interestingly, we noted that the structures of
156 *Brettanomyces* and *Wickerhamomyces* *GAL* clusters are strikingly syntenic to the *GAL*
157 clusters belonging to distantly related yeasts, specifically those belonging to the CUG-
158 Ser1 clade, which includes *Candida albicans* (Figure 2 and Supplemental Figure 4).
159 Since the CUG-Ser1 clade *GAL* cluster structure is evolutionarily derived²⁵, it is highly
160 unlikely that these two additional lineages would independently evolve such similar
161 structures. Instead, one might expect *Brettanomyces* to share a structure with
162 *Pacchysolen tannophilus*, its closest relative containing a *GAL* cluster. These
163 observations suggest, a model wherein the *Brettanomyces* and *Wickerhamomyces* *GAL*
164 clusters share ancestry with *GAL* clusters from the CUG-Ser1 clade, rather than with
165 those from their much closer organismal relatives.

166 Unexpectedly, we observed distinct *GAL* clusters in *Lipomyces starkeyi* and
167 *Nadsonia fulvescens* (Figure 2 and Supplemental Figure 3), two species that diverged
168 from the rest of the Saccharomycotina prior to the formation of the canonical *GAL*

169 cluster. *L. starkeyi*, a species belonging to a lineage that is sister to the rest of the
170 budding yeasts, contains a large gene cluster consisting of two copies of *GAL1*, a single
171 copy of *GAL7*, *GALE* (predicted to only encode the epimerase domain, instead of the
172 fused *GAL10* gene, which additionally encodes the mutarotase domain), and a gene
173 encoding a zinc-finger domain (Supplemental Figure 3). The novel content and
174 configuration of this cluster suggests that the *L. starkeyi* *GAL* gene cluster formed
175 independently of the canonical budding yeast *GAL* cluster.

176 Remarkably, the structure of the *GAL* cluster of *N. fulvescens* is nearly identical
177 to that of the CUG-Ser1 species *Cephaloascus albidus* (Figure 2 and Supplemental
178 Figures 3 and 4), despite the fact that these two lineages are separated by hundreds of
179 millions of years of evolution⁴. This synteny suggests that the *GAL* cluster of *N.*
180 *fulvescens* was either horizontally acquired or that it independently evolved the
181 bifunctional *GAL10* gene (fusion of galactose mutarotase (*GALM*) and UDP-galactose
182 4-epimerase (*GALE*) domains) and a *GAL* cluster with the same gene arrangement.
183 Interestingly, *N. fulvescens* var. *elongata* has a pseudogenized *GAL10* gene (indicated
184 by multiple inactivating mutations along the gene; Supplemental Figure 5), while *N.*
185 *fulvescens* var. *fulvescens* has an intact *GAL10* gene, and the varieties' phenotypes
186 were consistent with their inferred *GAL10* functionality (Supplemental Figure 1 and
187 Supplemental Table 3). Both varieties also contain a linked *GALE* gene, which resides
188 ~20 kb downstream of *GAL7*, suggesting the ongoing replacement of an ancestral
189 *GALE*-containing *GAL* cluster by a CUG-Ser1-like *GAL* cluster containing *GAL10*.
190 Notably, *GALE* or *GAL10* genes are present in some budding yeast species that do not
191 utilize galactose³⁴, and *N. fulvescens* var. *fulvescens* has only CUG-Ser1-like copies of

192 the *GAL7* and *GAL1* genes required for galactose utilization. While parsimony suggests
193 that the last common ancestor of *N. fulvescens* and its relative *Yarrowia lipolytica* was
194 able to utilize galactose, *N. fulvescens* rests on an unusually long branch with no other
195 known closely related species. Thus, in this case, we cannot infer whether partial cluster
196 loss and trait loss (i.e. to the state of possessing only *GALE* and not utilizing galactose)
197 preceded acquisition of the new functional cluster.

198

199 Allowing reacquisition is more parsimonious than enforcing loss

200 These synteny observations suggest three independent reacquisitions of the
201 *GAL* cluster and at least two independent reacquisitions of the galactose utilization trait.
202 To test the hypothesis of trait reversal, we next investigated whether, in some cases,
203 reacquisition of the *GAL* cluster offered a more parsimonious explanation than reductive
204 evolution. To reconcile the observed topologies of the gene and species phylogenies,
205 we reconstructed the evolutionary events using a parsimony framework, either
206 assuming Dollo's law of irreversibility to be true (only gene loss was possible) or false
207 (both gene loss and reacquisition were possible). When there was variation segregating
208 below the species level (e.g. *N. fulvescens* and *S. kudriavzevii*⁴⁶), we treated the
209 species as positive for galactose utilization. When Dollo's law was enforced, we inferred
210 15 distinct loss events for galactose metabolism (Figure 3A). When we allowed for the
211 violation of Dollo's law, we replaced a portion of the loss events with two reacquisition
212 events, arriving at a more parsimonious inference of 11 distinct events: 9 losses and 2
213 reacquisitions (Figure 3A). The most parsimonious scenario did not infer trait loss for
214 *Nadsonia*, but even adding one loss and one gain of galactose metabolism, instead of

215 the cluster replacement scenario, still yielded a more parsimonious solution of 13
216 distinct events.

217

218 Yeast *GAL* gene clusters have been horizontally transferred multiple times

219 From these synteny and trait reconstructions, we hypothesized that the *GAL*
220 clusters of *Brettanomyces*, *Wickerhamomyces*, and *Nadsonia* were horizontally
221 transferred from the CUG-Ser1 clade. This hypothesis predicts that the coding
222 sequences of their *GAL* genes should be more similar to species in the CUG-Ser1 clade
223 than to their closest relative possessing *GAL* genes. Thus, we calculated the percent
224 identities of Gal1, Gal7, and Gal10 proteins between four groups of species; (A)
225 between species in the candidate HGT recipient clade, (B) between the candidate HGT
226 recipient clade and their closest relative with *GAL* genes, (C) between the candidate
227 HGT recipient clade and the candidate donor clade, and (D) between the candidate
228 HGT recipient clade and an outgroup lineage (Figure 3B, C). If the genes were vertically
229 acquired, one would expect the percent identities to be highest in group A and then
230 decrease in the order of group B to C to D. If the genes were acquired horizontally, then
231 the percent identities would be higher in group C than in group B. Indeed, we found that
232 the percent identities of the Gal proteins of group C were significantly greater than
233 group B (Figure 3C, p -value = $1.79e-4$). These results suggest that the *GAL* clusters of
234 *Brettanomyces*, *Wickerhamomyces*, and *Nadsonia* were acquired horizontally from the
235 CUG-Ser1 clade.

236 To further explore whether HGT occurred in these taxa, we reconstructed
237 maximum-likelihood (ML) phylogenies for each of the *GAL* genes, as well as for the

238 concatenation of all three (Supplemental Figures 6-9). Interestingly, we observed a
239 consistent pattern of phylogenetic placement of *Brettanomyces*, *Wickerhamomyces*,
240 and *Nadsonia* *GAL* genes, which grouped to different lineages than would be expected
241 based on their species taxonomy or phylogeny (Supplemental Figure 1). The
242 *Wickerhamomyces* *GAL* genes clustered with *Hyphopichia*; the *Brettanomyces* *GAL*
243 genes clustered with several genera from the families Debaryomycetaceae and
244 Metschnikowiaceae; and the *Nadsonia* *GAL* genes clustered with those from the family
245 Cephaloascaceae. These observations are consistent with three independent horizontal
246 gene transfers of *GAL* clusters into these lineages from the CUG-Ser1 clade.

247 To formally test the hypothesis of *GAL* HGT, we used Approximately Unbiased
248 (AU) tests (Figure 4A). Specifically, we generated multiple maximum likelihood
249 phylogenetic trees using alignments of *GAL* genes with constraints on the placements
250 of various taxa: (i) fully constrained to follow the species tree, (ii) unconstrained in the
251 *Brettanomyces* lineage, (iii) unconstrained in the *Wickerhamomyces* lineage, (iv)
252 unconstrained in the *Nadsonia* lineage, and (v) unconstrained in all three candidate
253 HGT lineages (*Brettanomyces*, *Wickerhamomyces*, and *Nadsonia*). By comparing the
254 partially constrained trees to the fully constrained tree with AU tests, we found that each
255 of the proposed horizontal transfer events was statistically supported (Figure 4B). These
256 results were consistent across individual alignments of the *GAL* genes and when all
257 three lineages were examined together (Figure 4B). From these results, we conclude
258 that the *GAL* clusters of the *Brettanomyces*, *Wickerhamomyces*, and *Nadsonia* lineages
259 were likely acquired via HGT from ancient CUG-Ser1 yeasts.

260

261 Regulatory mode correlates with the horizontal gene transfers

262 Gal4 is the key transcriptional activator of the *GAL* cluster in *S. cerevisiae* and
263 responds to galactose through the co-activator Gal3 and co-repressor Gal80. This mode
264 of regulation is thought to be restricted to the family Saccharomycetaceae and is absent
265 in other yeasts and fungi⁴⁷. In other budding yeasts (including *C. albicans*, the most
266 thoroughly studied CUG-Ser1 species, as well as *Y. lipolytica*, an outgroup to *S.*
267 *cerevisiae* and *C. albicans*), regulation of the *GAL* cluster is thought to be under the
268 control of the activators Rtg1 and Rtg3⁴⁸. These two regulatory mechanisms respond to
269 different signals and have dramatically different dynamic ranges. In Gal4-regulated
270 species, the *GAL* cluster is nearly transcriptionally silent in the presence of glucose and
271 is rapidly induced to high transcriptional activity when only galactose is present. In
272 contrast, Rtg1/Rtg3-regulated species have high basal levels of transcription and are
273 weakly induced in the presence of galactose⁴⁸.

274 Intriguingly, all putative donor lineages of the *GAL* genes were from the CUG-
275 Ser1 clade of yeasts, and no transfers occurred from or into the family
276 Saccharomycetaceae. To examine whether the relaxed Rtg1/Rtg3 regulatory regimen of
277 the CUG-Ser1 yeasts might have facilitated their role as an HGT donor, as opposed to
278 the Gal4-mediated regulation of the Saccharomycetaceae, we identified sequence
279 motifs that were enriched 800 bp upstream from the coding regions of the *GAL1*, *GAL7*,
280 and *GAL10* genes (Supplemental Table 4). Then, based on the existing experimental
281 evidence on the regulation of the *GAL* genes^{23,24,48}, we divided the yeast species into
282 Saccharomycetaceae and non-Saccharomycetaceae species. We then ran a selective
283 motif enrichment analysis to determine if any regulatory motifs were enriched in one

284 group, but not the other. We found that the top enriched motifs corresponded to the
285 known Gal4-binding site in the Saccharomycetaceae²⁰ and the known Rtg1-binding site
286 in the non-Saccharomycetaceae species⁴⁸ (Figure 5A and B, Supplemental Table 4),
287 consistent with the previously documented regulatory rewiring of the *GAL* genes that
288 occurred at the base of the family Saccharomycetaceae⁴⁸. In general, the enrichment of
289 Rgt1-binding sites was patchier and did not include the HGT recipient lineages, the
290 previously characterized Rtg1-regulated *GAL* cluster of *Y. lipolytica*⁴⁸, or several CUG-
291 Ser1 clade species (e.g. *Ce. albidus*).

292 Taken together, our new results suggest that the switch to the Gal4-mode of
293 regulation, which is tighter and involves multiple unlinked and dedicated regulatory
294 genes, reduced the likelihood of horizontal transfer into naïve genomes or genomes that
295 had lost their *GAL* pathways. Specifically, any *GAL* cluster regulated by Gal4 would not
296 be able to be transcribed or properly regulated if it were horizontally transferred into a
297 species lacking *GAL4* and other regulatory genes. In contrast, Rtg1 and Rtg3 are more
298 broadly conserved, and any horizontally transferred *GAL* cluster regulated by them
299 would likely be sufficiently transcriptionally active, providing an initial benefit to the
300 organism.

301

302 Discussion

303 Budding yeasts have diversified from their metabolically complex most recent
304 common ancestor over the last 400 million years^{2,4}. While they have evolved
305 specialized metabolic capabilities, their evolutionary trajectories have been prominently
306 shaped by reductive evolution^{2,4,49,50}. Here, we present evidence that losses of the *GAL*

307 genes and galactose metabolism in some lineages were offset, tens of millions of years
308 after their initial losses, by eukaryote-to-eukaryote horizontal gene transfer (Figure 6).
309 While reacquired ancestral traits have been documented in several eukaryotic lineages,
310 our observation of galactose metabolism reacquisition differs in a few regards. First, the
311 majority of reported events did not identify the molecular mechanism or the genes
312 involved in the reacquired traits. Second, few studies have comprehensively sampled
313 taxa and constructed robust genome-scale phylogenies onto which the examined traits
314 were mapped, a requirement for robustly inferring trait evolution. Remarkably, we
315 observed trait reversal in at least two independent lineages, with a third possible
316 lineage, suggesting that the recovery of lost eukaryotic metabolic genes may be an
317 important and underappreciated driver in trait evolution in budding yeasts, and perhaps
318 more generally in fungi and other eukaryotes. In line with our study, budding yeasts also
319 have reacquired lost metabolic traits from bacteria, supporting the hypothesis that
320 regains via HGT offset reductive evolution⁴⁴.

321 The dearth of HGT from Saccharomycetaceae into other major clades provides
322 clues into the potential limits on ancestral trait reacquisition via HGT. We propose the
323 transcriptional rewiring to Gal4-mediated regulation imposed a restriction on the
324 potential for benefit of transferred *GAL* clusters. Since Gal4-mediated gene activation is
325 tightly coordinated and the off-state is less leaky⁴⁷, any transferred *GAL* cluster lacking
326 Gal4-binding sites into a species with exclusively Gal4-mediated activation in response
327 to galactose would not be able to activate the transferred genes. Similarly, transfer of a
328 Gal4-regulated gene cluster into a species lacking *GAL4* and other upstream regulators
329 would have limited potential for activation. For the case of transfer between two species

330 whose regulation does not rely on Gal4, the transferred *GAL* cluster would be
331 transcriptionally active because the broadly conserved transcription factors Rtg1 and
332 Rtg3 could further enhance moderate basal transcriptional activity⁴⁸. Thus, even leaky
333 levels of transcription would provide a benefit in the presence of galactose that could
334 further be refined, possibly to become regulated by lineage-specific networks. Under
335 this model, the likelihood of HGT is partly determined by the potential activity of the
336 transferred genes and by the recipient's ancestral regulatory mode.

337 More generally, our findings demonstrate that reductive evolution is not always a
338 dead end, and gene loss can be circumvented by HGT from distantly related taxa.
339 However, the scope of genes that can be regained in this fashion is likely limited. In
340 particular, the *GAL* genes of the CUG-Ser1 clade of budding yeasts represent
341 something of a best-case scenario. First, all enzymatic genes needed for phenotypic
342 output are encoded in a cluster, facilitating the likelihood that all necessary genes for
343 function are transferred together^{32,51}. Second, the regulatory mode of these *GAL* genes
344 is conducive to function in the recipient species, as they are loosely regulated by
345 conserved factors with moderate basal activity. Third, the genes would provide a clear
346 competitive advantage in environments with galactose.

347 The modern interpretation of Dollo's law is that species cannot return to a
348 previous character state after loss. Alongside recently reported character state reversals
349 in petunias after pseudogene reactivation⁵², our results of reacquisition of galactose
350 metabolism and *GAL* genes by HGT can be considered a case of character state
351 reversal. However, the previous example fits into the model that, for groups undergoing
352 adaptive radiations, lost traits seem to "flicker" on and off, resulting in an unusual

353 distribution of character states on the phylogeny. Here, and in the recently described
354 reacquisition of alcoholic fermentation genes from bacteria in fructophilic yeasts⁴⁴, the
355 ancestral genes were completely lost from the genome, and they were restored far later
356 than could be explained by the flickering of traits during adaptive radiations. The
357 reacquisition of galactose metabolism in budding yeasts represents a striking example
358 of gene and trait reversal by eukaryote-to-eukaryote horizontal gene transfer and
359 provides insight into the mechanisms by which Dollo's law can be broken.

360

361 **Methods**

362 GAL gene identification

363 We analyzed 96 publicly available genome sequences used in a recent study of
364 the Saccharomycotina phylogeny³ (86 Saccharomycotina, 10 outgroups), as well ten
365 additional species belonging to clades where we identified potentially deep losses of the
366 *GAL* gene cluster. Of the latter ten species, five genome sequences, including *Nadsonia*
367 *fulvescens* var. *fulvescens*, were published recently⁴, while genome sequences for five
368 new species are published here. Due to their importance to this study and since
369 previously published genome sequences may have been from different strains that were
370 unavailable for phenotyping, eight additional genome sequences were generated for
371 taxonomic type strains. In total, 104 genome sequences were analyzed. All genome
372 sequences generated after a backbone phylogeny was compiled from data published
373 before 2016³ are denoted Y1000+ in Supplemental Figures 6-9. The presence of *GAL*
374 genes in the genome assemblies was inferred with TBLASTN⁵³ v2.7.1 using the *C.*
375 *albicans* Gal1, Gal7, and Gal10 sequences as queries, followed by extraction of the

376 open reading frame centered on the location of the best hit. The structure and synteny
377 of the clusters were manually curated and documented. For *S. kudriavzevii*, where
378 balanced variation is segregating for the *GAL* pathway⁴⁶, phylogenetic analyses were
379 performed with the taxonomic type strain (cannot grow on galactose), whereas
380 summary figures (Supplemental Figure 3 and Figures 1, 3, and 6) show a reference
381 strain (ZP591) that can grow on galactose.

382

383 Sequencing and assembly of genomes

384 For the five new genomes sequenced here, genomic DNA was sonicated and
385 ligated to Illumina sequencing adaptors as previously described²⁶. The paired-end
386 library was sequenced on an Illumina HiSeq 2500 instrument, conducting a rapid 2x250
387 run. To generate whole genome assemblies, paired-end Illumina reads were used as
388 input to a meta-assembler pipeline iWGS⁵⁴. The quality of the assemblies was assessed
389 using QUAST⁵⁵ v3.1, and the best assembly for the newly described species was
390 chosen based on N50 statistics.

391

392 *GAL* gene similarity analysis

393 To calculate the percent identities between Gal proteins, we first aligned the
394 protein sequences for each species (see Supplemental Table 1 for species used) of
395 Gal1, Gal7, and Gal10 and generated percent identity matrices using Clustal Omega⁵⁶.
396 These results were then subdivided into four groups: (1) the percent identities between
397 species within the potential HGT recipient clade, (2) the percent identities between
398 species of the recipient clade and their closest relative with *GAL* genes, (3) the percent

399 identities between species of the recipient clade and species in the donor lineage, and
400 (4) the percent identities between species of the recipient clade and an outgroup
401 lineage (i.e. *S. cerevisiae*). Next, a similarity score was calculated by normalizing the
402 percent identity values of each group to the average value of the fourth group:

$$403 \quad \textit{Similarity Score} = \textit{Log}2\left(\frac{x_i}{\textit{ave}X_4}\right)$$

404

405 Phylogenetic analyses

406 Sequence alignments were conducted using MAFFT⁵⁷ v 7.409 run in the "--auto"
407 mode. Alignments were subjected to maximum-likelihood phylogenetic reconstruction
408 using RAxML⁵⁸ v8.1.0 with 100 rapid bootstrap replicates. Constrained phylogenetic
409 trees were generated with RAxML using the "-g" option, with the constraint tree identical
410 to the species tree, except for the species/lineage of interest, whose position on the tree
411 was allowed to be optimized by the ML algorithm. Statistical support for the HGT events
412 involving *GAL* genes was determined using the Approximately Unbiased (AU) test, by
413 comparing the various partially constrained ML phylogenies and the fully constrained
414 phylogeny. The AU test was performed with IQ-TREE⁵⁹ v1.6.8 (-au option), which was
415 run with the General Time Reversible model, substitution rate heterogeneity
416 approximated with the gamma distribution (-m GTR+G), and with 10,000 replicates (-zb
417 10000)

418

419 Regulatory motif enrichment

420 Sequences of 800 bp upstream of the start codon of all identified *GAL* genes
421 were extracted and subjected to a regulatory motif identification analysis using MEME⁶⁰

422 v5.0.2, with the following constraints: maximum number of motifs = 20 (-nmotifs 20),
423 maximum length of motif = 25 bases (-maxw 25), any number of motif repetitions (-anr),
424 active search of reverse complement of the used sequence (-revcomp), and the log-
425 likelihood ratio method (-use_llr). Selective enrichment of motifs was determined by
426 splitting the sequences into Saccharomycetaceae and non-Saccharomycetaceae
427 groups and running AME⁶¹ v5.0.2, with each group being the control group in one
428 analysis and the test group in a second analysis.

429

430 Species tree reconstruction

431 Our data matrix was composed of 104 budding yeasts and 10 outgroups,
432 comprising of 1,219 BUSCO genes (601,996 amino acid sites); each gene had a
433 minimum sequence occupancy ≥ 57 taxa and sequence length ≥ 167 amino acid
434 residues. For the concatenation-based analysis, we used RAxML version 8.2.3 and IQ-
435 TREE⁵⁹ version 1.5.1 to perform maximum likelihood (ML) estimations under an
436 unpartitioned scheme (a LG+GAMMA model) and a gene-based partition scheme
437 (1,219 partitions; each gene has its own model), respectively. As a result, four ML trees
438 produced by two different phylogenetic programs and two different partition strategies
439 were topologically identical. Branch support for each internode was evaluated with 100
440 rapid bootstrap replicates using RAxML⁶². For the coalescence-based analysis, we first
441 estimated individual gene trees with their best-fitting amino acid models, which were
442 determined by IQ-TREE⁵⁹ (the “-m TESTONLY” option); we then used those individual
443 gene trees to infer the species tree implemented in the ASTRAL program⁶³, v4.10.2.
444 The reliability for each internode was evaluated using the local posterior probability

445 measure⁶⁴. Finally, internode certainty (IC) was used to quantify the incongruence by
446 considering the most prevalent conflicting bipartitions for each individual internode
447 among individual gene trees^{65,66,67} implemented in RAxML⁵⁸ v8.2.3. The relative
448 divergence times were estimated using the RelTime⁶⁸ in MEGA7⁶⁹. The ML topology
449 was used as the input tree.

450

451 Growth assays

452 We previously published galactose growth data for the majority of
453 species^{4,70}. Growth experiments were performed for an additional nine species
454 separately (Supplemental Table 3). All species were struck onto yeast extract peptone
455 dextrose (YPD) plates from freezer stocks and grown for single colonies. Single
456 colonies were struck onto three types of plates minimal media base (5g/L ammonium
457 sulfate, 1.71g/L Yeast Nitrogen Base (w/o amino acids, ammonium sulfate, or carbon),
458 20g/L agar) treatments with either: 2% galactose, 1% galactose, or 2% glucose (to test
459 for auxotrophies). We also re-struck the specific colony onto YPD plates as a positive
460 control. All growth experiments were performed at room temperature. After initial growth
461 on treatment plates, growth was recorded for the first round, and we struck colonies
462 from each treatment plate onto a second round of the respected treatment to ensure
463 there was no nutrient carryover from the YPD plate. For example, a single colony from
464 2% galactose minimal media plate was struck for a second round of growth on a 2%
465 galactose minimal media plate. We inspected plates every three days for growth for up
466 to a month. Yeasts were recorded as having no growth on galactose if they did not grow
467 on either the first or second round of growth on galactose.

468 **References**

- 469 1. Hittinger, C. T. *et al.* Genomics and the making of yeast biodiversity. *Current Opinion in*
470 *Genetics and Development* vol. 35 100–109 (2015).
- 471 2. Kurtzman, C. P., Fell, J. W. & Boekhout, T. *The Yeasts: A Taxonomic Study, 5th Edition.*
472 *The Yeasts* vols 1–3 (2011).
- 473 3. Shen, X.-X. *et al.* Reconstructing the Backbone of the Saccharomycotina Yeast Phylogeny
474 Using Genome-Scale Data. *G3: Genes, Genomes, Genetics* **6**, 3927–3939 (2016).
- 475 4. Shen, X.-X. *et al.* The tempo and mode of genome evolution in the budding yeast
476 subphylum. *Cell* **175**, 1–13 (2018).
- 477 5. Dollo, L. Les lois de l'évolution. *Bulletin de la Société Belge de Géologie* **VII**, 164–166
478 (1893).
- 479 6. Simpson, G. G. *The Major Features of Evolution.* (Columbia University Press, 1953).
- 480 7. Collin, R. & Cipriani, R. Dollo's law and the re-evolution of shell coiling. *Proceedings of the*
481 *Royal Society B: Biological Sciences* **270**, 2551–2555 (2003).
- 482 8. Whiting, M. F., Bradler, S. & Maxwell, T. Loss and recovery of wings in stick insects.
483 *Nature* **421**, 264–267 (2003).
- 484 9. Kohlsdorf, T. & Wagner, G. P. Evidence for the reversibility of digit loss: a phylogenetic
485 study of limb evolution in *Bachia* (Gymnophthalmidae: Squamata). *Evolution* **60**, 1896–1912
486 (2006).
- 487 10. Brandley, M. C., Huelsenbeck, J. P. & Wiens, J. J. Rates and patterns in the evolution of
488 snake-like body form in squamate reptiles: Evidence for repeated re-evolution of lost digits
489 and long-term persistence of intermediate body forms. *Evolution* **62**, 2042–2064 (2008).

- 490 11. Kohlsdorf, T., Lynch, V. J., Rodrigues, M. T., Brandley, M. C. & Wagner, G. P. Data and
491 data interpretation in the study of limb evolution: A reply to Galis et al. On the reevolution of
492 digits in the lizard genus *Bachia*. *Evolution* **64**, 2477–2485 (2010).
- 493 12. Lynch, V. J. & Wagner, G. P. Did egg-laying boas break Dollo’s law? Phylogenetic evidence
494 for reversal to oviparity in sand boas (*Eryx*: Boidae). *Evolution* **64**, 207–216 (2010).
- 495 13. Wiens, J. J. Re-evolution of lost mandibular teeth in frogs after more than 200 million years,
496 and re-evaluating Dollo’s law. *Evolution* **65**, 1283–1296 (2011).
- 497 14. Xu, F. *et al.* On the reversibility of parasitism: adaptation to a free-living lifestyle via gene
498 acquisitions in the diplomonad *Trepomonas* sp. PC1. *BMC Biology* **14**, 62 (2016).
- 499 15. Recknagel, H., Kamenos, N. A. & Elmer, K. R. Common lizards break Dollo’s law of
500 irreversibility: Genome-wide phylogenomics support a single origin of viviparity and re-
501 evolution of oviparity. *Molecular Phylogenetics and Evolution* (2018)
502 doi:10.1016/j.ympev.2018.05.029.
- 503 16. Collin, R. & Miglietta, M. P. Reversing opinions on Dollo’s Law. *Trends in Ecology and*
504 *Evolution* vol. 23 602–609 (2008).
- 505 17. Seher, T. D. *et al.* Genetic basis of a violation of Dollo’s law: Re-evolution of rotating sex
506 combs in *Drosophila bipectinata*. *Genetics* **192**, 1465–1475 (2012).
- 507 18. Esfeld, K. *et al.* Pseudogenization and Resurrection of a Speciation Gene. *Current Biology*
508 (2018) doi:10.1016/j.cub.2018.10.019.
- 509 19. Chippindale, P. T. & Ronald M. Bonett c Andrew S. Baldwin, a and John J. Wiensd, e, A.
510 Phylogenetic evidence for a major reversal of life-history evolution in plethodontid
511 salamanders. *Evolution* **58**, 2809–2822 (2004).

- 512 20. Johnston, M. A model fungal gene regulatory mechanism: the GAL genes of *Saccharomyces*
513 *cerevisiae*. *Microbiological reviews* **51**, 458–476 (1987).
- 514 21. Jayadeva Bhat, P. & Murthy, T. V. S. Transcriptional control of the GAL/MEL regulon of
515 yeast *Saccharomyces cerevisiae*: Mechanism of galactose-mediated signal transduction.
516 *Molecular Microbiology* **40**, 1059–1066 (2001).
- 517 22. Hittinger, C. T., Rokas, A. & Carroll, S. B. Parallel inactivation of multiple GAL pathway
518 genes and ecological diversification in yeasts. *Proceedings of the National Academy of*
519 *Sciences of the United States of America* **101**, 14144–9 (2004).
- 520 23. Hittinger, C. T. & Carroll, S. B. Gene duplication and the adaptive evolution of a classic
521 genetic switch. *Nature* **449**, 677–681 (2007).
- 522 24. Martchenko, M., Levitin, A., Hogues, H., Nantel, A. & Whiteway, M. Transcriptional
523 Rewiring of Fungal Galactose-Metabolism Circuitry. *Current Biology* **17**, 1007–1013
524 (2007).
- 525 25. Slot, J. C. & Rokas, A. Multiple GAL pathway gene clusters evolved independently and by
526 different mechanisms in fungi. *Proceedings of the National Academy of Sciences of the*
527 *United States of America* **107**, 10136–10141 (2010).
- 528 26. Hittinger, C. T. *et al.* Remarkably ancient balanced polymorphisms in a multi-locus gene
529 network. *Nature* **464**, 54–58 (2010).
- 530 27. Wolfe, K. H. *et al.* Clade- and species-specific features of genome evolution in the
531 *saccharomycetaceae*. *FEMS Yeast Research* vol. 15 (2015).
- 532 28. Kuang, M. C., Hutchins, P. D., Russell, J. D., Coon, J. J. & Hittinger, C. T. Ongoing
533 resolution of duplicate gene functions shapes the diversification of a metabolic network.
534 *eLife* **5**, (2016).

- 535 29. Kuang, M. C. *et al.* Repeated Cis-Regulatory Tuning of a Metabolic Bottleneck Gene during
536 Evolution. *Mol Biol Evol* **35**, 1968–1981 (2018).
- 537 30. Campbell, M. A., Staats, M., van Kan, J. A. L., Rokas, A. & Slot, J. C. Repeated loss of an
538 anciently horizontally transferred gene cluster in *Botrytis*. *Mycologia* **105**, 1126–1134
539 (2013).
- 540 31. Wisecaver, J. H., Slot, J. C. & Rokas, A. The Evolution of Fungal Metabolic Pathways. *PLoS*
541 *Genetics* **10**, (2014).
- 542 32. Wisecaver, J. H. & Rokas, A. Fungal metabolic gene clusters-caravans traveling across
543 genomes and environments. *Frontiers in Microbiology* vol. 6 (2015).
- 544 33. Slot, J. C. Fungal Gene Cluster Diversity and Evolution. *Advances in Genetics* (2017)
545 doi:10.1016/bs.adgen.2017.09.005.
- 546 34. Riley, R. *et al.* Comparative genomics of biotechnologically important yeasts. *Proceedings*
547 *of the National Academy of Sciences* **113**, 9882–9887 (2016).
- 548 35. Matsuzawa, T. *et al.* New insights into galactose metabolism by *Schizosaccharomyces*
549 *pombe*: Isolation and characterization of a galactose-assimilating mutant. *Journal of*
550 *Bioscience and Bioengineering* (2011) doi:10.1016/j.jbiosc.2010.10.007.
- 551 36. Duan, S.-F. *et al.* Reverse Evolution of a Classic Gene Network in Yeast Offers a
552 Competitive Advantage. *Curr. Biol.* **29**, 1126-1136.e5 (2019).
- 553 37. Legras, J.-L. *et al.* Adaptation of *S. cerevisiae* to Fermented Food Environments Reveals
554 Remarkable Genome Plasticity and the Footprints of Domestication. *Mol. Biol. Evol.* **35**,
555 1712–1727 (2018).

- 556 38. Boocock, J., Sadhu, M. J., Bloom, J. S. & Kruglyak, L. Ancient balancing selection
557 maintains incompatible versions of a conserved metabolic pathway in yeast. *bioRxiv* 829325
558 (2019) doi:10.1101/829325.
- 559 39. Keeling, P. J. & Palmer, J. D. Horizontal gene transfer in eukaryotic evolution. *Nature*
560 *Reviews Genetics* **9**, 605–618 (2008).
- 561 40. Fitzpatrick, D. A. Horizontal gene transfer in fungi. *FEMS Microbiology Letters* vol. 329 1–
562 8 (2012).
- 563 41. Marcet-Houben, M. & Gabaldón, T. Acquisition of prokaryotic genes by fungal genomes.
564 *Trends in Genetics* vol. 26 5–8 (2010).
- 565 42. Hall, C. & Dietrich, F. S. The reacquisition of biotin prototrophy in *Saccharomyces*
566 *cerevisiae* involved horizontal gene transfer, gene duplication and gene clustering. *Genetics*
567 **177**, 2293–2307 (2007).
- 568 43. Alexander, W. G., Wisecaver, J. H., Rokas, A. & Hittinger, C. T. Horizontally acquired
569 genes in early-diverging pathogenic fungi enable the use of host nucleosides and nucleotides.
570 *Proceedings of the National Academy of Sciences* **113**, 4116–4121 (2016).
- 571 44. Gonçalves, C. *et al.* Evidence for loss and adaptive reacquisition of alcoholic fermentation in
572 an early-derived fructophilic yeast lineage. *eLife* (2018) doi:10.7554/eLife.33034.
- 573 45. Kominek, J. *et al.* Eukaryotic Acquisition of a Bacterial Operon. *Cell* **176**, 1356-1366.e10
574 (2019).
- 575 46. Hittinger, C. T. *et al.* Remarkably ancient balanced polymorphisms in a multi-locus gene
576 network. *Nature* **464**, 54–58 (2010).

- 577 47. Choudhury, B. I. & Whiteway, M. Evolutionary Transition of GAL Regulatory Circuit from
578 Generalist to Specialist Function in Ascomycetes. *Trends in Microbiology* (2018)
579 doi:10.1016/j.tim.2017.12.008.
- 580 48. Dalal, C. K. *et al.* Transcriptional rewiring over evolutionary timescales changes quantitative
581 and qualitative properties of gene expression. *eLife* **5**, (2016).
- 582 49. Dujon, B. *et al.* Genome evolution in yeasts. *Nature* **430**, 35 (2004).
- 583 50. Steenwyk, J. L. *et al.* Extensive loss of cell-cycle and DNA repair genes in an ancient lineage
584 of bipolar budding yeasts. *PLOS Biology* **17**, e3000255 (2019).
- 585 51. Rokas, A., Wisecaver, J. H. & Lind, A. L. The birth, evolution and death of metabolic gene
586 clusters in fungi. *Nat. Rev. Microbiol.* **16**, 731–744 (2018).
- 587 52. Esfeld, K. *et al.* Pseudogenization and Resurrection of a Speciation Gene. *Current Biology*
588 (2018) doi:10.1016/j.cub.2018.10.019.
- 589 53. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
590 search tool. *Journal of Molecular Biology* (1990) doi:10.1016/S0022-2836(05)80360-2.
- 591 54. Zhou, X. *et al.* in silico Whole Genome Sequencer & Analyzer (iWGS): A
592 Computational Pipeline to Guide the Design and Analysis of de novo Genome Sequencing
593 Studies. *G3 (Bethesda, Md.)* **6**, 3655–3662 (2016).
- 594 55. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUASt: Quality assessment tool for
595 genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
- 596 56. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence
597 alignments using Clustal Omega. *Mol Syst Biol* **7**, 539 (2011).
- 598 57. Katoh, K. & Standley, D. M. MAFFT: Iterative refinement and additional methods. *Methods*
599 *in Molecular Biology* **1079**, 131–146 (2014).

- 600 58. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large
601 phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
- 602 59. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and
603 effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular*
604 *Biology and Evolution* **32**, 268–274 (2015).
- 605 60. Bailey, T. L. *et al.* MEME Suite: Tools for motif discovery and searching. *Nucleic Acids*
606 *Research* (2009) doi:10.1093/nar/gkp335.
- 607 61. McLeay, R. C. & Bailey, T. L. Motif Enrichment Analysis: A unified framework and an
608 evaluation on ChIP data. *BMC Bioinformatics* (2010) doi:10.1186/1471-2105-11-165.
- 609 62. Stamatakis, A., Hoover, P., Rougemont, J. & Renner, S. A Rapid Bootstrap Algorithm for
610 the RAxML Web Servers. *Systematic Biology* **57**, 758–771 (2008).
- 611 63. Mirarab, S. & Warnow, T. ASTRAL-II: Coalescent-based species tree estimation with many
612 hundreds of taxa and thousands of genes. in *Bioinformatics* vol. 31 i44–i52 (2015).
- 613 64. Sayyari, E. & Mirarab, S. Fast Coalescent-Based Computation of Local Branch Support from
614 Quartet Frequencies. *Molecular biology and evolution* **33**, 1654–1668 (2016).
- 615 65. Salichos, L. & Rokas, A. Inferring ancient divergences requires genes with strong
616 phylogenetic signals. *Nature* **497**, 327–331 (2013).
- 617 66. Salichos, L., Stamatakis, A. & Rokas, A. Novel information theory-based measures for
618 quantifying incongruence among phylogenetic trees. *Molecular Biology and Evolution* **31**,
619 1261–1271 (2014).
- 620 67. Kobert, K., Salichos, L., Rokas, A. & Stamatakis, A. Computing the Internode Certainty and
621 Related Measures from Partial Gene Trees. *Molecular Biology and Evolution* **33**, 1606–1617
622 (2016).

- 623 68. Tamura, K. *et al.* Estimating divergence times in large molecular phylogenies. *Proceedings*
624 *of the National Academy of Sciences of the United States of America* **109**, 19333–8 (2012).
- 625 69. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis
626 Version 7.0 for Bigger Datasets. *Molecular biology and evolution* **33**, 1870–1874 (2016).
- 627 70. Opulente, D. A. *et al.* Factors driving metabolic diversity in the budding yeast subphylum.
628 *BMC Biology* **16**, 26 (2018).
- 629
- 630

631 **Acknowledgments**

632 We are grateful to Carlos A. Rosa for providing the strain *Wickerhamomyces* sp.
633 UFMG-CM-Y6624. We thank the Rokas and Hittinger labs for comments and
634 discussions and the University of Wisconsin Biotechnology Center DNA Sequencing
635 Facility for providing Illumina sequencing facilities and services. This material is based
636 upon work supported by the National Science Foundation under Grant Nos. DEB-
637 1442113 (to A.R.) and DEB-1442148 (to C.T.H. and C.P.K.), in part by the DOE Great
638 Lakes Bioenergy Research Center (DOE BER Office of Science DE-SC0018409 and
639 DE-FC02-07ER64494 to Timothy J. Donohue), USDA National Institute of Food and
640 Agriculture (Hatch Project 1020204), and National Institutes of Health (NIAID AI105619
641 to A.R.), and a Guggenheim fellowship (to A.R). C.T.H. is a Pew Scholar in the
642 Biomedical Sciences, a Vilas Early Career Investigator, and a H. I. Romnes Faculty
643 Fellow, supported by the Pew Charitable Trusts, Vilas Trust Estate, and Office of the
644 Vice Chancellor for Research and Graduate Education with funding from the Wisconsin
645 Alumni Research Foundation, respectively.

646

647 **Data deposition**

648 Raw DNA sequencing data were deposited in GenBank under Bioproject ID
649 PRJNA647756. Whole Genome Shotgun assemblies have been deposited at
650 DDBJ/ENA/GenBank under the accessions XXX-XXXX (pending processing). The
651 versions described in this paper are versions XXX-XXXX (pending processing).

652

653 **Author contributions**

654 M.A.B.H. (study design, preliminary phylogenetic analyses, sequence analyses, cluster
655 analyses, text); J. K. (study design, genome assemblies, phylogenetic analyses, motif
656 enrichment analyses, text); D.A.O. (genomic DNA isolation, library preparation, yeast
657 growth assays); X.S. (phylogenomic analyses); A.L.L. (cluster analyses); X.Z.
658 (preliminary genome annotations and analyses); J.DeV., A.B.H. (genomic DNA
659 isolation, library preparation); C.P.K. (support and supervision, study design); and A.R.,
660 and C.T.H. (support and supervision, study design, text).

661

662 **Competing interests**

663 The authors declare no competing interests

664 **Figure legends:**

665 Figure 1. Evolutionary history of galactose metabolism in budding yeasts.

666 Species-level presence or absence of galactose utilization is mapped onto the relative

667 divergence timetree (Supplemental Figures 2 and 3) with some clades collapsed.

668 Branch color denotes the ability to metabolize galactose; blue (+) and red(-). The black

669 bars mark the branches of three key events in the evolution of the *GAL* cluster (I-cluster

670 formation, II-translocation of *ORF-Y* into the cluster, and III-translocation of *ORF-X* into

671 the cluster). Numbered groups indicate the three clades with unexpected *GAL* clusters.

672 The dashed branch of the *Nadsonia* lineage indicates the ambiguity of the ancestral

673 character state due to its extremely long branch (Supplemental Figures 2 and 3).

674

675 Figure 2. Surprisingly syntenic *GAL* clusters between distantly related groups of yeasts.

676 The *GAL* clusters of five representative species are shown. Numbers correspond to

677 positions in each scaffold or contig. Further details and examples are provided in

678 Supplemental Figures 3 and 4.

679

680 Figure 3. Comparison of Dollo's law versus reacquisition of the *GAL* genes from the

681 CUG-Ser1 clade.

682 **(A)** Evolutionary trait reconstruction, based on a parsimony framework either assuming

683 that traits cannot be regained (left) or that traits can be regained (right).

684 **(B)** Similarity score of the Gal1, Gal7, and Gal10 proteins as calculated by protein

685 sequence similarity and the comparisons shown in the upper right; means with standard

686 deviations are depicted. Raw percent identity values are shown in Supplemental Figure

687 10. Comparisons used to calculate similarity scores: A, between species within the
688 clade with potentially transferred *GAL* genes (recipient clade); B, between the recipient
689 clade and their closest relative with *GAL* genes; C, between the recipient clade and the
690 potential donor lineage (CUG-Ser1 clade); and D, between the recipient clade and an
691 outgroup lineage.

692 **(C)** Student's t-test of the mean difference between groups. Negative values violate the
693 assumptions of vertical inheritance, and the critical comparison between B and C is
694 bolded in blue.

695

696 Figure 4. The *GAL* clusters of three lineages were acquired by HGT.

697 **(A)** Diagrammatic representation of AU tests performed by either constraining the tree
698 in selected lineages (as indicated in red) or not.

699 **(B)** p-values of the AU tests are shown. All tests significantly reject their null
700 hypotheses, indicating that the unconstrained topologies better explain the observed
701 distribution of *GAL* genes, which is consistent with HGT as the mechanism of
702 reacquisition.

703

704 Figure 5. Enrichment of transcription factor-binding sites in the promoters of *GAL*
705 enzymatic genes.

706 **(A)** Maximum likelihood phylogeny of Saccharomycotina. Colors indicate highlighted
707 clades: light blue – *Nadsonia*, red – *Brettanomyces*, yellow – CUG-Ser1 clade, green –
708 *Wickerhamomyces*, and blue – Saccharomycetaceae.

709 **(B)** Heatmap of enrichment for either Rtg1- or Gal4-binding sites in the promoters of the
710 *GAL* genes (*GAL1*, *GAL10*, *GAL7*). White-shaded boxes indicate lineages lacking the
711 *GAL* gene cluster.

712

713 Figure 6. The CUG-Ser1 clade serves as a common donor of the *GAL* gene cluster to
714 other yeasts.

715 Cladogram of the ML phylogeny is presented with the leaf labels removed for simplicity.

716 The colored boxes represent the species' ability to utilize galactose (blue =

717 positive/variable, red = negative), gray circles indicate the presence of a full set of *GAL*

718 enzymatic genes, and gray stars indicate that those *GAL* genes are clustered. Five

719 lineages on the cladogram are colored: pink - *Schizosaccharomyces pombe* (a member

720 of the subphylum Taphrinomycotina with a transferred *GAL* cluster that does not confer

721 utilization), light blue – *Nadsonia*, red – *Brettanomyces*, yellow – CUG-Ser1 clade, and

722 green – *Wickerhamomyces*. Numbered boxes and arrows depict the four horizontal

723 transfer events of the *GAL* cluster. The colored arcs encompassing the cladogram

724 represent the predicted regulatory mode of the *GAL* genes: orange – Rtg1/Rtg3 (non-

725 Gal4) and purple – Gal4.

726 **Supplemental Figures and Tables**

727 Supplemental Table 1. Strains used in this study.

728

729 Supplemental Table 2. Chi-squared (χ^2) test of genotype-to-phenotype associations of
 730 species presented in Supplemental Figure 3. We used our phenotypes in cases where
 731 our data disagreed with *The Yeasts* book².

Observed	GAL genes present	GAL genes absent	Total
Gal ⁺	63	1	64
Gal ⁻	3	28	31
Total	66	29	95
χ^2	77.5816 (p-value < 0.00001)		

732

733

734 Supplemental Table 3. Galactose growth phenotyping of key species. NT, not tested.

Species	Controls		1st streak		2nd streak	
	YPD	2% Glu	2% Gal	1% Gal	2% Gal	1%Gal
<i>Brettanomyces anomalus</i>	+	+	-	-	NT	NT
<i>Brettanomyces naardensis</i>	+	+	+	+	+	+
<i>Kluyveromyces marxianus</i>	+	+	NT	+	NT	+
<i>Metschnikowia bicuspidata</i> var. <i>bicuspidata</i>	+	+	NT	+	NT	+
<i>Nadsonia fulvescens</i> var. <i>fulvescens</i>	+	+	+	-	+	NT
<i>Ogataea parapolyomorpha</i>	+	+	-	-	-	-
<i>Zygosaccharomyces bailii</i>	+	+	-	-	NT	NT
<i>Starmerella bombicola</i>	+	+	+	+	+	+
<i>Wickerhamomyces anomalus</i>	+	+	+	+	+	+

735

736

737 Supplemental Table 4. Per-species p-values for the presence of Gal4- and Rtg1-binding
 738 site motifs in individual GAL genes.

739

740 Supplemental Figure 1. Genome-scale maximum likelihood phylogeny.

741

742 Supplemental Figure 2. Genome-scale internode certainty cladogram.

743

744 Supplemental Figure 3. Distribution of the structure of *GAL* gene clusters.

745 Both cluster structure and growth characteristics are mapped onto the relative
746 divergence timetree. Growth on galactose is indicated by the colored squares next to
747 each species (green=blue, yellow=variable, red=negative). Asterisks next to certain
748 species names indicated either a new genome sequence published here (**) or an
749 additional genome sequence from a recent study (*)⁴, including *Nadsonia fulvescens*
750 var. *fulvescens*. To ensure phenotyping could be performed on sequenced strains, we
751 also sequenced the genomes of the taxonomic type strains for eight species and report
752 those *GAL* clusters here (^). The syntenic structure of the *GAL* genes are displayed to
753 the right of the growth characteristics for each species. The structure of the *Nadsonia*
754 *fulvescens* var. *elongata* cluster is shown in Supplemental Figure 4.

755

756 Supplemental Figure 4. Surprisingly syntenic *GAL* clusters between diverse lineages.

757

758 Supplemental Figure 5. Alignment of the *GAL10* genes of *N. fulvescens* var. *fulvescens*
759 and *N. fulvescens* var. *elongata*. Genes were aligned using MAFFT v 7.409 using --
760 auto. Likely inactivating mutations are shown in various colors: mutation of the start
761 codon in orange, frameshift mutations in blue, in-frame nonsense mutations in red, and
762 insertions in green. One in-frame deletion is shown in purple.

763

764 Supplemental Figure 6. Gene tree of *GAL1* genes.

765 Supplemental Figure 7. Gene tree of *GAL7* genes.

766 Supplemental Figure 8. Gene tree of *GAL10* genes.

767 Supplemental Figure 9. Concatenated gene tree of the *GAL*actose enzymatic gene

768 cluster.

769 Supplemental Figure 10. Percent identities of *GAL* genes as calculated by the

770 comparisons shown in Figure 3.

771

Figure 1

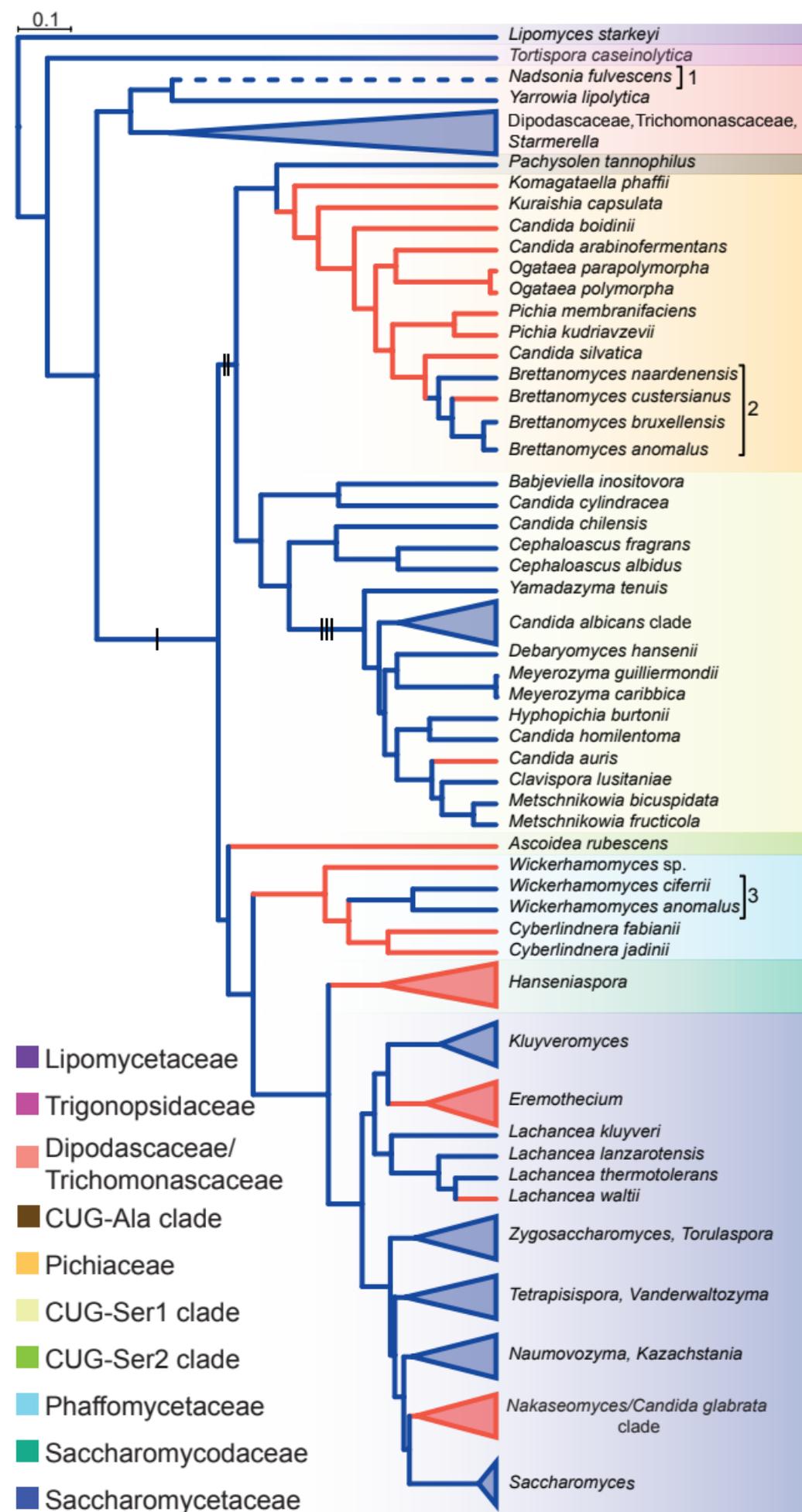


Figure 2

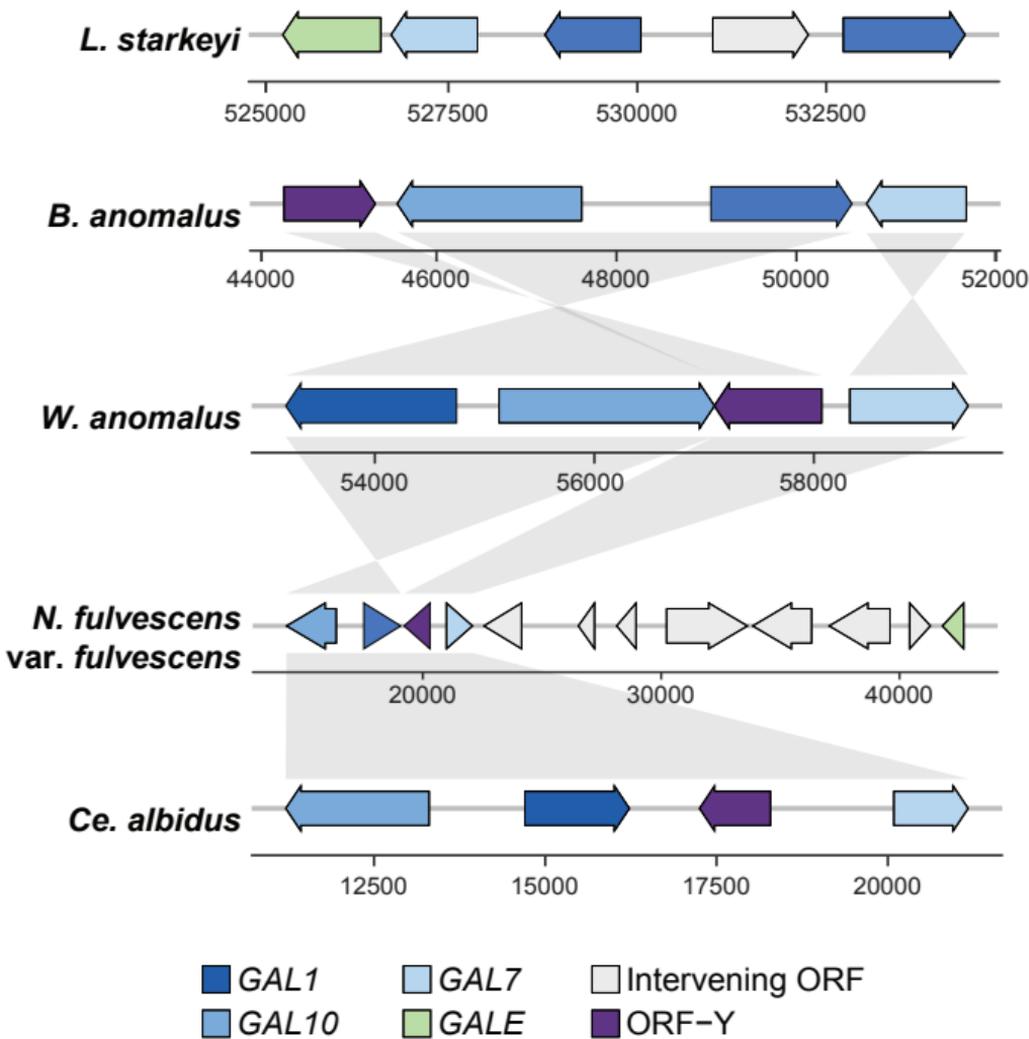
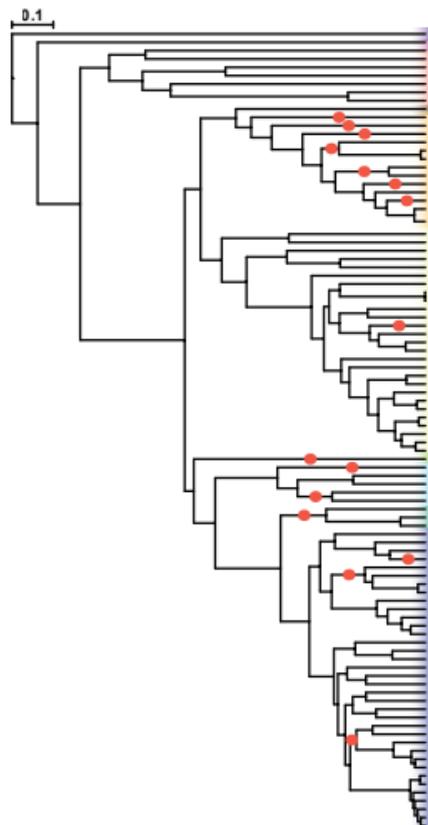


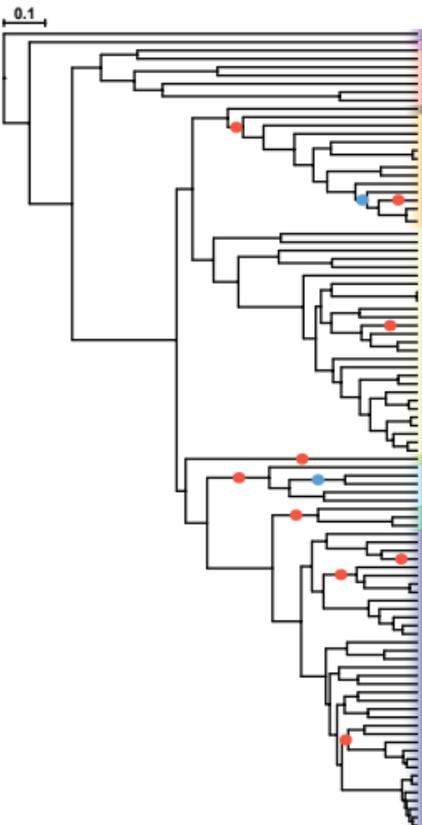
Figure 3**A****Dollo's law**

● Number of losses = 15

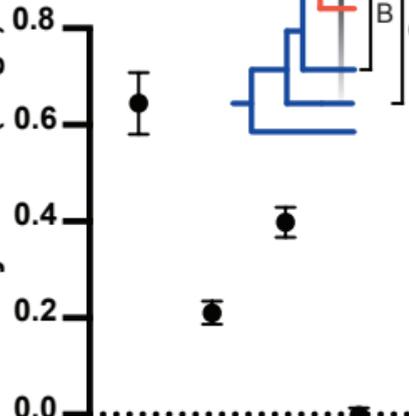
**No Dollo's law**

● Number of losses = 9

● Number of reacquisitions = 2

**B**

Similarity Score (Log2)



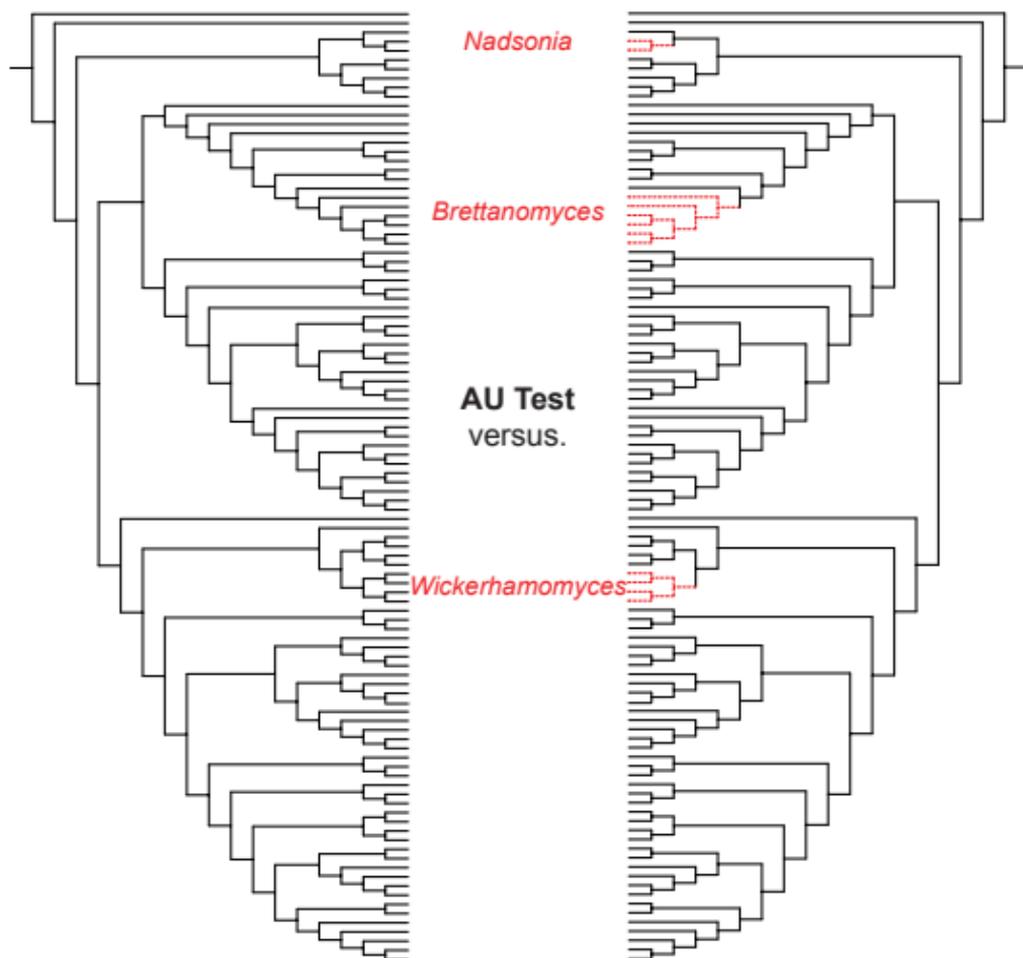
A B C D

C

Test	Mean Diff.	Significant?	Adj. P Value
A vs B	0.4339	Yes	<1e-10
A vs C	0.2467	Yes	2.73e-5
A vs D	0.6464	Yes	<1e-10
B vs C	-0.1872	Yes	5.78e-4
B vs D	0.2125	Yes	7.87e-5
C vs D	0.3997	Yes	<1e-10

Figure 4**A** Species tree constrained topology

Partially unconstrained topology

**B**

Unconstrained clade	<i>GAL1</i>	<i>GAL7</i>	<i>GAL10</i>	Merged
<i>Brettanomyces</i>	3.44e-04	9.52e-03	5.97e-03	1.15e-90
<i>Wickerhamomyces</i>	3.29e-55	3.36e-09	7.08e-07	5.17e-60
<i>Nadsonia</i>	1.05e-73	8.41e-05	5.32e-44	8.19e-06
All	5.19e-12	7.90e-79	3.22e-06	3.91e-29

Figure 5

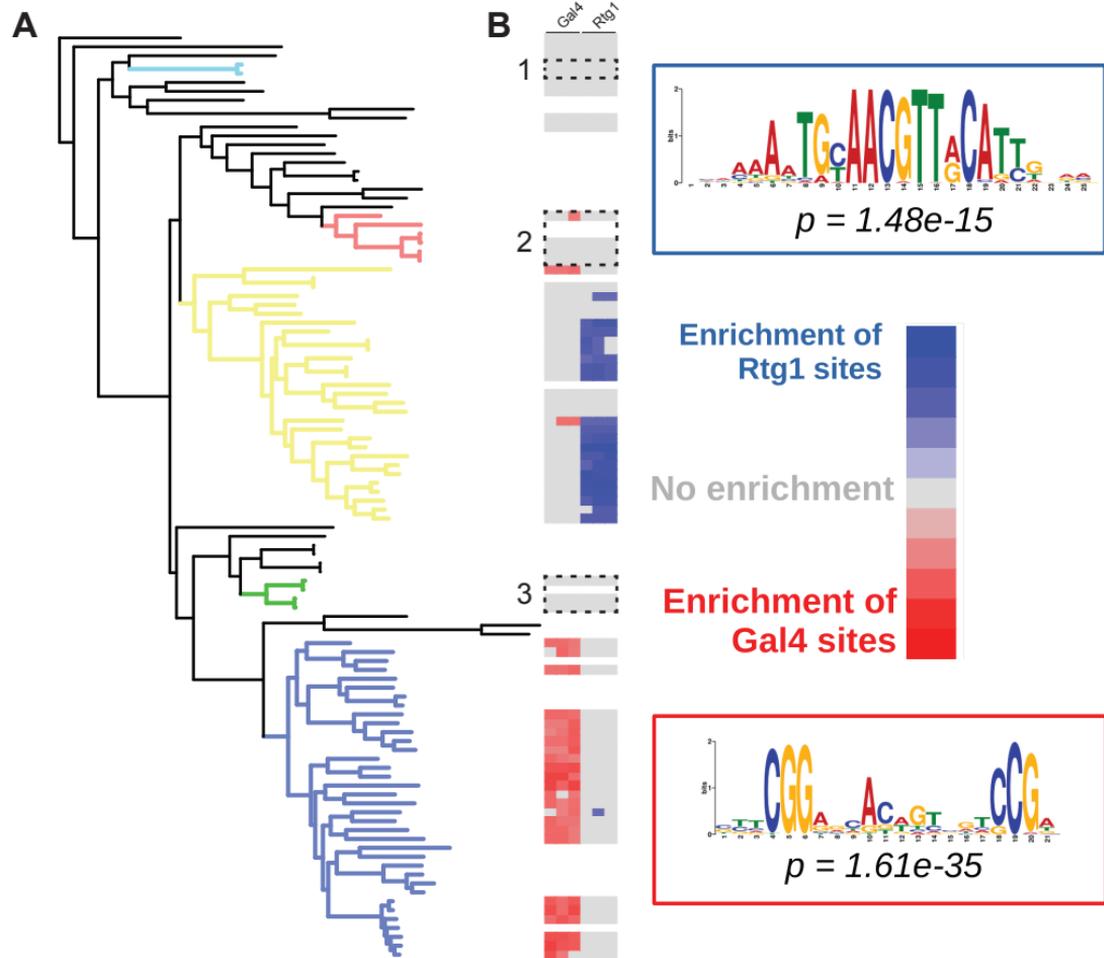


Figure 6

