The Evolution of the *GAL*actose Utilization Pathway in Budding Yeasts

Marie-Claire Harrison[1], Abigail L. LaBella[1], Chris Todd Hittinger[2,*], & Antonis Rokas[1,*]

[1] *Department of Biological Sciences, Vanderbilt University, TN, USA*

[2] *Laboratory of Genetics, DOE Great Lakes Bioenergy Research Center, Wisconsin Energy Institute, Center for Genomic Science Innovation, J.F. Crow Institute for the Study of Evolution, University of Wisconsin-Madison, Madison, WI, USA*

*Authors for correspondence: cthittinger@wisc.edu and antonis.rokas@vanderbilt.edu

Running Title: The evolution of the *GAL* pathway

Keywords: galactose, horizontal gene transfer, regulatory rewiring, gene regulation, gene cluster, fungi, polymorphism

Abstract

The Leloir galactose utilization or *GAL* pathway of budding yeasts, including that of the baker's yeast *Saccharomyces cerevisiae* and the opportunistic human pathogen *Candida albicans*, breaks down the sugar galactose for energy and biomass production. The *GAL* pathway has long served as a model system for understanding how eukaryotic metabolic pathways, including their modes of regulation, evolve. More recently, the physical linkage of the structural genes *GAL1*, *GAL7*, and *GAL10* in diverse budding yeast genomes has been used as a model for understanding the evolution of gene clustering. In this opinion, we summarize exciting recent work on three different aspects of this iconic pathway's evolution: gene cluster organization, *GAL* gene regulation, and the population genetics of the *GAL* pathway.

Introduction

Galactose is a monosaccharide that is abundant in nature and is found in many forms. For example, galactose is a component of the disaccharide lactose, a main ingredient of dairy products, as well as the trisaccharide raffinose and disaccharide melibiose, which are both common in grains and other plants [1]. Galactose is also often found in the form of glycolipids and glycoproteins, which are produced by cells and are critical for a wide variety of cellular functions [2,3]. The assimilation of galactose has been extensively characterized both in humans [2] and in the baker's yeast *Saccharomyces cerevisiae* [3]. The enzymatic products of three genes, *GAL1*, *GAL7*, and *GAL10* in budding yeasts, or *GALK*, *GALT*, *GALM*, and *GALE* in humans, are responsible for converting galactose into glucose-1-phosphate (Figure 1) [3]. Glucose-1-phosphate is then isomerized to glucose-6-phosphate by Pgm1p or Pgm2p, which can then be used to generate energy for the cell via glycolysis.

Through multiple now-classic studies, the Leloir galactose utilization or *GAL* pathway of budding yeasts has been established as a model system for understanding the function and evolution of eukaryotic metabolic pathways and their regulation [4-8]. We will start by introducing the classic work on the evolution of the *GAL* pathway from comparative studies of *S. cerevisiae* and *Candida albicans*. Then, we will synthesize several discoveries in the last few years that have significantly enriched our understanding of how this iconic pathway evolved. These results include the discovery of repeated instances of wholesale pathway loss and of reacquisition via **horizontal gene transfer (HGT)** (see Glossary), extensive variation in sequence and function within species, and diversity in levels and modes of pathway regulation and genomic organization.

The classical view: evolution of the *GAL* pathway in *S. cerevisiae* and *C. albicans*

Comparison of the *GAL* genes, the *GAL* pathway's genomic organization, and regulation between *C. albicans* and *S. cerevisiae* shows that there is substantial variation between the two species. While the structural genes are functional orthologs of each other and have the same order and orientation in the two species, the *C. albicans* cluster also contains the genes *GAL102* and *ORF-X* (Figure 1). *GAL102* encodes a glucose-4,6-dehydratase and *ORF-X* encodes a transporter. Although transporters of galactose (encoded by *GAL2* and other *HXT* genes for *S. cerevisiae* and by *HXT* genes in *C. albicans*) and the *PGM1/PGM2* structural genes are important parts of both the *S. cerevisiae* and *C. albicans GAL* pathways, none of these genes are parts of their gene clusters [9,10].

The regulation of the *GAL* pathway also varies between the two organisms. In *S. cerevisiae*, the pathway is regulated by the Zn-binuclear cluster transcription factor Gal4p, which binds to a highly enriched regulatory motif whose consensus sequence is 5'-CGG-N_{11}-CCG-3'. Gal4p is repressed by Gal80p in the absence of galactose. In the presence of galactose, the repression of Gal80p is removed by Gal3p, and Gal4p induces transcription of the *GAL1*, *GAL7*, and *GAL10* genes. This leads to an "on-and-off switch" or bimodal mode of regulation, with strong suppression of the pathway when galactose is absent and a ~900-fold induction when galactose is detected [8]. In contrast, the *GAL* pathway of *C. albicans* is regulated by the heterodimeric **helix-loop-helix transcription factors** Rtg1p and Rtg3p. These transcription factors bind to a different binding motif, whose consensus sequence is 5'-TGYAACGTTRCA-3'. The basal rate of transcription of *GAL* genes in *C. albicans* is higher, and the induction is more graded, with a ~12-fold induction in the presence of galactose [8].

Budding yeasts are a model lineage for studying the evolution of metabolic pathways

There are ~1,200 known species of budding yeasts that belong to the subphylum Saccharomycotina, otherwise known as the budding yeast subphylum, one of the three subphyla in the phylum Ascomycota [11-14]. Advances in genome sequencing technologies have led to the sequencing of the genomes of hundreds of species across the subphylum, allowing for a greater understanding of how these organisms, their genes, and their metabolic traits evolved. The two biggest efforts to date have been the genome sequencing of 16 diverse species of biotechnologically important yeast species by a Joint Genome Institute-led effort [15]; and the genome sequencing of 220 species (including 24 from the RIKEN Institute in Japan) led by the Y1000+ Project[i], which placed emphasis on sequencing at least one representative species from each genus of budding yeasts [11]. These efforts have led to a robust higher-level phylogeny of budding yeasts and the identification of 12 major lineages or clades [11]; *S. cerevisiae* belongs to the family Saccharomycetaceae, while *C. albicans* belongs to the CUG-Ser1 clade (so called because the CUG codon encodes for serine rather than leucine in this clade [16]).

This richness of genomic and phylogenetic data is complemented by extensive aggregated metabolic and ecological trait data for a broad and representative set of budding yeast species. These include qualitative growth data on 44 substrates and environmental isolation data for up to 50 environments for 784 species, and quantitative growth rate data for galactose, mannose, and glucose for 258 species [14,17,18]. These genomes, metabolic and ecological data, and species phylogeny, coupled with the availability of strains from all known species, have provided an unprecedented resource that has allowed for deeper analysis of metabolic pathways, including the *GAL* pathway.

Evolution of genomic organization of the *GAL* pathway across the budding yeast subphylum

Approximately half of the ~350 species with available genomes studied to date contain the *GAL1, GAL7,* and *GAL10* genes and can grow on galactose [11,17]. Furthermore, out of the 174 species that grow on galactose, 127 species have the *GAL1, GAL7,* and *GAL10* genes clustered, and 23 additional species that do not grow on galactose also have the genes clustered, indicating they still may use the pathway or that it was recently inactivated [11,18]. The 47 species that grow on galactose without having the genes clustered are mostly in clades other than the CUG-Ser1 clade (which contains *C. albicans*) and Saccharomycetaceae (which contains *S. cerevisiae*), such as the Dipodascaceae/Trichomonascaceae clade, which contains the genera *Blastobotrys* and *Yarrowia*. However, a few species in Saccharomycetaceae that are descendants of an ancient whole genome duplication event [19,20], such as *Vanderwaltozyma polyspora* [7], also have a functional *GAL* pathway but lack a cluster; this genome duplication event was followed by extensive loss of duplicate genes, such that some *GAL* genes are now found in one ohnologous genomic region and the rest are found in the other.

The clustering of *GAL1, GAL7,* and *GAL10* has evolved multiple times in fungi. For example, *Cryptococcus* basidiomycetous yeasts also have a *GAL* cluster, but phylogenetic analysis suggests that this cluster evolved independently, and its organization differs from the budding yeast *GAL* clusters [7]. In budding yeasts, clustering of the *GAL* genes originated at least twice:

once in the common ancestor of *S. cerevisiae* and *C. albicans* and another time in *Lipomyces* and relatives (Figure 2) [21]. *Lipomyces* species have *GAL10* next to *GAL7*, instead of *GAL1*, and they often have two copies of *GAL1* with an uncharacterized transcription factor between them that is homologous to *ARA1*, the L-arabinose regulatory transcription factor in *Trichoderma reesei* [21,22]. The repeated origin of the clustering of *GAL1*, *GAL7,* and *GAL10* in budding yeasts and other fungi supports the hypothesis that this genomic organization may be selectively advantageous in certain conditions (Box 1).

The variety of budding yeast clusters shows that functional *GAL* pathways can evolve into several different organizations in the genome. One prominent example is the giant *GAL* clusters in *Torulaspora* species, where *GAL4, MEL1, GAL2*, *PGM1/2*, and *HGT1*, as well as multiple copies of *GAL1* and *GAL10* are parts of the same cluster (Figure 2) [23]. These additional components of the cluster are likely functionally significant in environments with high amounts of melibiose or galactose: Mel1p breaks down melibiose into galactose and glucose, Gal2p transports galactose into the cell, Gal4p can upregulate transcription of the pathway, Hgt1p can transport glucose (and possibly other sugars) generated by the pathway, and Pgm1/2p converts the glucose-1-phosphate generated by the Leloir pathway to glucose-6-phosphate, which can then go through glycolysis. Interestingly, the clustering of these genes with *GAL1, GAL7,* and *GAL10* is not observed in other budding yeasts, with the notable exception of *HGT1*, which is frequently found in *GAL* clusters in the CUG-Ser1 clade (e.g., in *Priceomyces medius* – see Figure 2).

The clustering of the *GAL* genes makes the *GAL* cluster genomic region a good candidate for HGT since acquisition of the *GAL* cluster would provide the minimal genetic information necessary for the utilization of the available galactose in the environment. Acquisition of the *GAL* cluster could occur in organisms with a functional *GAL* pathway as well as in organisms whose ancestors lost the pathway – there have been many instances of *GAL* pathway loss across the subphylum [5,7,15,19,21] – suggesting that pathway losses are potentially reversible. Support for this hypothesis comes from a recent molecular phylogenetic study that inferred HGT of a CUG-Ser1 type of *GAL* cluster independently occurred in the genera *Brettanomyces*, *Wickerhamomyces,* and *Nadsonia* after ancestral losses of the *GAL* pathway in at least two lineages [21]. This inference is supported by the observation that the *GAL* clusters of the recipient species share the same cluster organization as the donor species and by formal topology tests that show their phylogenetic placement to be closer to the CUG-Ser1 clade than to their known evolutionary relatives (Figure 2) [21].

CUG-Ser1 yeasts have repeatedly served as HGT donors of *GAL* clusters to organisms in other budding yeast major clades, but there are no known instances of *GAL* cluster HGTs from Saccharomycetaceae to lineages outside of the family. It has been hypothesized that this difference is due to the *RTG1/RTG3* mode of regulation used by the CUG-Ser1 clade. These transcription factors are more broadly conserved than *GAL4*, which is not known to regulate the *GAL* genes outside of the family Saccharomycetaceae [8,24]. In fact, even though *C. albicans* has a *GAL4* ortholog, it is much shorter in length and has been shown to regulate an entirely different set of genes [25]. Similarly, in the Pichiaceae clade, a *GAL4* ortholog has been found to regulate Crabtree-Warburg Effect in *Komagataella phaffii* instead of the *GAL* pathway [24].

GAL clusters in the CUG-Ser1 clade have low background levels of gene expression, whereas clusters in the family Saccharomycetaceae are typically actively repressed in the presence of glucose [21]. Thus, *GAL* clusters acquired from the CUG-Ser1 clade would be more likely to be basally expressed and less likely to require the evolution of a new mode of regulation in the recipient organisms. Interestingly, CUG-Ser1 yeasts can act as donors for HGT even to lineages outside of budding yeasts; for example, *Schizosaccharomyces* fission yeasts, an independently evolved lineage of yeasts in the subphylum Taphrinomycotina, also acquired their *GAL* cluster via HGT from CUG-Ser1 yeasts [7].

Evolution of Gene Regulation in the *GAL* Pathway

The *GAL4* regulatory system appears to be conserved throughout the family Saccharomycetaceae: Domain I encoded by *GAL4* (amino acid residues 1-76, encoding a DNA-binding domain homologous to dozens of transcription factors) is conserved throughout the budding yeast subphylum, but Domain V encoded by *GAL4* (residues 767-881, encoding the Gal80p-binding domain) is only conserved in Saccharomycetaceae (excluding some *Torulaspora* species), and its pattern of conservation mirrors that of *GAL80* (Figure 3) [30,31]. This suggests that the Gal4p-Gal80p mode of regulation is restricted to the Saccharomycetaceae clade. However, there are still significant differences in *GAL* gene induction and repression between species. For example, in 1% glucose medium supplement with galactose, *S. cerevisiae* waits until glucose is completely exhausted to start metabolizing galactose (a phenomenon known as "diauxic lag"), while the closely related species *Saccharomyces uvarum* does not [32]. The stronger repression and slower induction of the *GAL* genes in *S. cerevisiae* compared to *S. uvarum* gives *S. cerevisiae* a fitness advantage in environments where glucose is in excess but a disadvantage when switching from glucose to galactose [32,33]. When the promoter or the *GAL* coding regions found in *S. cerevisiae* were expressed in *S. uvarum*, this phenotype could be partially reconstructed, suggesting that both the coding and promoter regions contribute to this *S. cerevisiae* phenotype. Thus, even within the *GAL4* regulatory system, there can be significant differences in induction and repression of the *GAL* genes.

The growth delay or diauxic lag of *S. cerevisiae* when switching from metabolizing glucose to galactose is a well-characterized phenotype thought to be due to the repression of the *GAL* genes in the presence of glucose by Mig1p [34]. However, recent studies on wild strains of *S. cerevisiae* have found variations in this phenotype [34,35]. Further investigation revealed that the yeasts are responding to the extracellular ratio of glucose to galactose to begin expressing the *GAL* pathway as opposed to a threshold amount of galactose or glucose. Different strains induce the *GAL* genes at higher or lower ratios of galactose to glucose [35]. Inducing the *GAL* genes at a lower ratio of galactose to glucose leads to a fitness advantage in certain environments, likely due to a shorter diauxic lag [35]. More recent studies have described the regulation of the *GAL* pathway as that of a “dimmer-switch” where turning gene expression on or off is decoupled from the regulation of the level of expression [36]. The pathway is switched on or off by Gal3p (which removes Gal80p from Gal4p, activating the catalytic activity of Gal4p) due to the ratio of galactose to glucose, whereas the level of expression of the pathway is controlled by Mig1p based on the concentration of glucose by modulating expression of Gal4p [36]. This gives yeasts using the *GAL4-GAL80* mode of regulation the fitness advantage of the *GAL* pathway already being switched on prior to all glucose being depleted, lessening the diauxic lag, but limits the cost of expression of those genes through Mig1p reducing their expression according to glucose concentration.

Comparison of the *GAL* pathways of *S. cerevisiae*, *S. uvarum*, and related species has revealed several other differences. Regulatory differences were first noted between *S. uvarum* and *S. cerevisiae* when it was found that the *S. uvarum* genome has retained both *GAL80* and *GAL80B* (the two genes are ohnologs stemming from the ancient whole duplication event, and both were predicted by machine learning to be involved in galactose assimilation), while the *S. cerevisiae* genome has retained only *GAL80* [5,33]. Recent work demonstrated that the corepressor encoded by *GAL80B* in *S. uvarum* plays an important role in preventing metabolic overload when grown in galactose, especially when alternative sugars are also being metabolized [37]. Consequently, while a *GAL80* knockout in *S. cerevisiae* grew more quickly in galactose, a double knockout of *GAL80* and *GAL80B* caused a growth arrest of *S. uvarum* when grown in galactose. Finally, Gal4p-binding sites upstream of the metabolic bottleneck gene *PGM1* in *S. uvarum* and several other species of the family Saccharomycetaceae lead to significantly faster growth on galactose compared to species lacking these binding sites (either naturally or via knock out experiments) [38]. Interestingly, these more active *GAL* networks [32,33,37] tend to be found in the same yeast species that also have dual layers of repression [38]. Thus, it seems that species are continually dialing their regulatory systems to the availability of galactose and other sugars in their environments.

While the *GAL4* mode of regulation has been well-characterized in the clade Saccharomycetaceae, less is known about the regulation of the *GAL* pathway in other species of budding yeasts, especially those in other major clades. Recently, it was discovered that the transcription factors Rtg1p and Rtg3p regulate the *GAL* pathway of *C. albicans* [8]. This regulatory mode differs from the Gal4p mode because it has a higher basal level of expression and a more graded induction, rather than the bimodal, on-and-off mode of Gal4p regulation. Due to the conservation of *RTG1* across the budding yeast subphylum, as well as reduced induction of *GAL1* in the presence of galactose by a knockout of *RTG1* in the outgroup species *Yarrowia lipolytica,* it was suggested that Rtg1p-Rtg3p mode of regulation may be the ancestral one [8]. In support of this hypothesis, *RTG1* and *RTG3* are conserved through most of the budding yeast phylogeny (Figure 3). However, although Rtg1/3p-binding motifs are conserved in most galactose-growing species in the CUG-Ser1 clade, a lineage within the genus *Metschnikowia* and a few other species (all of which belong to the CUG-Ser1 clade) appear to lack these motifs (Figure 3). Furthermore, neither the Gal4p, nor the Rtg1/3p motifs, are highly enriched in the *GAL* cluster regions of budding yeast clades other than Saccharomycetaceae (which contains *S. cerevisiae*) and CUG-Ser1 (which contains *C. albicans*) (Figure 3). These data raise the hypothesis that there is a variety of modes of regulation of galactose metabolism in the budding yeast subphylum.

Population-level variation of *GAL* gene clusters

In recent years, examination of isolates from diverse environments has substantially enhanced our understanding of genomic and phenotypic variation in populations of budding yeast species [39-41]. While the genomic organization and regulation of *GAL* gene clusters vary between budding yeast species, numerous population genomic studies have shown that *GAL* function and regulation can also vary substantially within species. The first example of substantial population-level variation in the *GAL* gene cluster was described in *Saccharomyces kudriavzevii*, a close relative of *S. cerevisiae*. Whereas European isolates of *S. kudriavzevii* have a functional *GAL* pathway comprised of six genes (compared to *S. cerevisiae*, they only lack the optional co-inducer encoded by *GAL3*) and can grow on galactose, Eastern Asian isolates cannot grow on galactose and their *GAL* pathway is composed of pseudogenes that are syntenic with the functional alleles [6]. If a crossing occurred between these populations, meiotic progeny would be unlikely to harbor either completely functional or completely non-functional networks, and some of the partial networks are less fit than either parent. For example, *S. kudriavzevii* segregants that lacked the corepressor encoded by *GAL80,* but had other functional genes, would constitutively express them in environments that lack galactose, often at a substantial fitness cost [6]. This example shows that the *GAL* genes of budding yeast populations can adapt to different environments and regulatory systems, including maintaining completely non-functional versions as local adaptations and balanced polymorphisms.

Recent examinations of the genomes of more than one thousand *S. cerevisiae* isolates have revealed the existence of three different combinations of highly divergent alleles in their *GAL* gene networks [42-45], suggesting that substantial population-level variation in the *GAL* pathway may be more common than previously thought in yeast populations. The first combination, found in most isolates, is composed of the alleles found in the reference S288C strain of *S. cerevisiae.* The second combination, found in a small percentage of isolates from different environments, including dairy ones, is composed of highly diverged alleles in the *GAL2, GAL1/7/10*, and *PGM1* genes; these alleles' divergence from their reference counterparts predates both the origin of the species *S. cerevisiae*, as well as the origin of the genus *Saccharomyces* [42,43,45]. The third combination, found in a few Chinese isolates from soil or wood environments, is composed of the same highly diverged *GAL1/7/10* alleles as the second combination, but the alleles for the *GAL2* and *PGM1* loci differ substantially from those found in the other two combinations; these alleles' origin(s) also predates the origin of the genus *Saccharomyces* [42,43].

The common characteristic of these alternative combinations of *GAL* gene network variants is that they allow these isolates to grow faster on galactose and slower on glucose than the reference S288C strain (which contains the first combination). For example, the *PGM1* allele of some of these isolates has a Gal4p-regulated promoter, allowing for quick utilization of galactose for energy through glycolysis, while the *PGM1* allele of the reference strain does not; this allele is incompatible with the *GAL2* and *GAL1/7/10* alleles of isolates with the second combination [42]. Similarly, mutations in certain isolates have impaired or abolished Mig1p- and Gal80p-mediated repression of the *GAL* pathway [43]. Finally, these isolates show – in addition to the observed variation in their *GAL2* loci – extensive variation in their hexose transporters, with numerous examples of gene fusion events, gene truncations, and wholesale gene deletions [43,45].

The very deep origin of the observed variation in the *GAL* gene network of *S. cerevisiae* raises the question of the evolutionary processes involved in its making and maintenance. Two alternative hypotheses have been proposed: introgression or HGT from a yet-to-be-identified species closely related to the genus *Saccharomyces* [43,45] and ancient balancing selection [42]; a combination of these two evolutionary scenarios (e.g., **introgression** followed by balancing selection) is also a possibility. Extensive genetic simulations reject a scenario where the alternative variants stem from a recent introgression, perhaps around the beginnings of agriculture, favoring instead a scenario of ancient balancing selection or ancient introgression followed by balancing selection [42].

Recent studies have begun to identify similar types of adaptations in other budding yeast metabolic genes and pathways. For example, it was recently shown that *Kluyveromyces lactis* var. *lactis*, the dairy-based variety of *K. lactis,* acquired the ability to ferment the milk sugar lactose via a recent HGT of a *LAC4-LAC12* gene cluster from a dairy isolate of *Kluyveromyces marxianus* [46]. In this cluster, *LAC4* encodes the enzyme lactase that hydrolyzes lactose into galactose and glucose, and *LAC12* encodes the transporter lactose permease [46].

These population genomic examinations raise the hypothesis that the *GAL* pathway, and perhaps most other metabolic pathways, of budding yeasts are subject to varied selection for the utilization of the galactose present in different environments and conditions. Thus, their pathways are likely to harbor multiple, highly divergent gene network variants and exhibit strong signatures of local adaptation. If this turns out to be true, population genomic examinations of populations of the ~1,200 known species of budding yeasts could reveal a treasure-trove of novel *GAL* (and other metabolic) gene network variants that would not only enhance our understanding of eukaryotic pathway evolution, but also of yeast metabolic engineering.

Concluding Remarks

The *GAL* pathway of the budding yeast *S. cerevisiae* is a favorite textbook example of the regulation of gene expression in eukaryotes (e.g., [47]). Recent advances in genome sequencing of budding yeasts, coupled with evolutionary genomic and functional studies of diverse populations and species across the subphylum, have allowed for many significant insights on the evolution of the *GAL* pathway that extend far beyond the insights obtained from the classic work in *S. cerevisiae* (and more recently in *C. albicans*). Since this pathway has long served as a model for gene regulation and evolution in eukaryotes, these insights are crucial for broadening our understanding of how metabolic pathways, and their mode of regulation, change in response to different environments. Nevertheless, several questions remain (see Outstanding Questions). The addition of more genomes from diverse budding yeast species and populations found in different environments, coupled with the development of new genetic model systems, is likely to reveal new and interesting findings. Fleshing out the *GAL* pathway's evolutionary and ecological diversity will help bridge our understanding of how genotypic variation emerges as phenotypic variation, perpetuating the pathway's utility as a model.

Acknowledgements

We thank members of the Rokas and Hittinger labs, especially members of the Y1000+ Project team, for the helpful suggestions. M-C. H. was supported by the Biological Sciences Graduate Program at Vanderbilt University. Research in C.T.H's lab is supported by National Science Foundation (DEB-1442148 and DEB-2110403), in part by the DOE Great Lakes Bioenergy Research Center (DOE BER Office of Science DE-SC0018409), the USDA National Institute of Food and Agriculture (Hatch Project 1020204), a Pew Scholarship in the Biomedical Sciences (Pew Charitable Trusts), and an H. I. Romnes Faculty Fellowship (Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation). Research in A.R.'s lab is supported by the National Science Foundation (DEB-

1442113 and DEB-2110404), the Burroughs Wellcome Fund, and the National Institutes of Health/National Institute of Allergy and Infectious Diseases (R56AI146096).

Resources

i) http://y1000plus.org/

Glossary

Helix-loop-helix transcription factor: a class of dimeric transcription factors that are common in eukaryotic genomes and which contain two alpha-helices connected by a loop [48].

Horizontal gene transfer (HGT): the transfer of a gene or of a genomic region containing genes (e.g., a gene cluster) between organisms by means other than sexual reproduction [49].

Introgression: the incorporation of genetic material from one species into the genome of another as a result of their hybridization, followed by repeated backcrossing to one of the parental species [50].

References:

1. Acosta, P. B., & Gross, K. C. (1995). Hidden sources of galactose in the environment. *European Journal of Pediatrics*, *154*(7 Suppl 2), S87-92.
2. Coelho, A. I., Berry, G. T., & Rubio-Gozalbo, M. E. (2015). Galactose metabolism and health. *Current Opinion in Clinical Nutrition & Metabolic Care*, *18*(4), 422–427.
3. Sellick, C. A., Campbell, R. N., & Reece, R. J. (2008). Galactose metabolism in yeast-structure and regulation of the Leloir pathway enzymes and the genes encoding them. *International Review of Cell and Molecular Biology*, *269*, 111–150.
4. Johnston, M. (1987). A model fungal gene regulatory mechanism: The *GAL* genes of *Saccharomyces cerevisiae*. *Microbiology and Molecular Biology Reviews*, *51*(4), 458–476.
5. Hittinger, C. T., Rokas, A., & Carroll, S. B. (2004). Parallel inactivation of multiple *GAL* pathway genes and ecological diversification in yeasts. Proceedings of the National Academy of Sciences, 101(39), 14144–14149.
6. Hittinger, C. T., Gonçalves, P., Sampaio, J. P., Dover, J., Johnston, M., & Rokas, A. (2010). Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature*, *464*(7285), 54–58.
7. Slot, J. C., & Rokas, A. (2010). Multiple *GAL* pathway gene clusters evolved independently and by different mechanisms in fungi. *Proceedings of the National Academy of Sciences*, *107*(22), 10136–10141.
8. Dalal, C. K., Zuleta, I. A., Mitchell, K. F., Andes, D. R., El-Samad, H., & Johnson, A. D. (2016). Transcriptional rewiring over evolutionary timescales changes quantitative and qualitative properties of gene expression. *ELife*, *5*.

9. Van Ende, M., Wijnants, S., & Van Dijck, P. (2019). Sugar Sensing and Signaling in *Candida albicans* and *Candida glabrata*. *Frontiers in Microbiology*, *10*.

10. Brown, V., Sabina, J., & Johnston, M. (2009). Specialized Sugar Sensing in Diverse Fungi. *Current Biology*, *19*(5), 436–441.

11. Shen, X.-X., Opulente, D. A., Kominek, J., Zhou, X., Steenwyk, J. L., Buh, K. V., Haase, M. A. B., Wisecaver, J. H., Wang, M., Doering, D. T., Boudouris, J. T., Schneider, R. M., Langdon, Q. K., Ohkuma, M., Endoh, R., Takashima, M., Manabe, R., Čadež, N., Libkind, D., Rosa, C. A., DeVirgilio, J., Hulfachor, A. B., Groenewald, M., Kurtzman, C. P., Hittinger, C. T., Rokas, A. (2018). Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell*, *175*(6), 1533-1545.e20.

12. Shen, X.-X., Steenwyk, J. L., LaBella, A. L., Opulente, D. A., Zhou, X., Kominek, J., Li, Y., Groenewald, M., Hittinger, C. T., & Rokas, A. (2020). Genome-scale phylogeny and contrasting modes of genome evolution in the fungal phylum Ascomycota. *Science Advances*, *6*(45).

13. Li, Y., Steenwyk, J. L., Chang, Y., Wang, Y., James, T. Y., Stajich, J. E., Spatafora, J. W., Groenewald, M., Dunn, C. W., Hittinger, C. T., Shen, X.-X., & Rokas, A. (2021). A genome-scale phylogeny of the kingdom Fungi. *Current Biology: CB*, *31*(8), 1653-1665.e5.

14. Kurtzman, C., Fell, J. W., & Boekhout, T. (2011). *The Yeasts: A Taxonomic Study*. Elsevier.

15. Riley, R., Haridas, S., Wolfe, K. H., Lopes, M. R., Hittinger, C. T., Göker, M., Salamov, A. A., Wisecaver, J. H., Long, T. M., Calvey, C. H., Aerts, A. L., Barry, K. W., Choi, C., Clum, A., Coughlan, A. Y., Deshpande, S., Douglass, A. P., Hanson, S. J., Klenk, H.-P., … Jeffries, T. W. (2016). Comparative genomics of biotechnologically important yeasts. *Proceedings of the National Academy of Sciences*, *113*(35), 9882–9887.

16. Krassowski, T., Coughlan, A. Y., Shen, X.-X., Zhou, X., Kominek, J., Opulente, D. A., Riley, R., Grigoriev, I. V., Maheshwari, N., Shields, D. C., Kurtzman, C. P., Hittinger, C. T., Rokas, A., & Wolfe, K. H. (2018). Evolutionary instability of CUG-Leu in the genetic code of budding yeasts. *Nature Communications*, *9*(1), 1887.

17. Opulente, D. A., Rollinson, E. J., Bernick-Roehr, C., Hulfachor, A. B., Rokas, A., Kurtzman, C. P., & Hittinger, C. T. (2018). Factors driving metabolic diversity in the budding yeast subphylum. *BMC Biology*, *16*(1), 26.

18. LaBella, A. L., Opulente, D. A., Steenwyk, J. L., Hittinger, C. T., & Rokas, A. (2021). Signatures of optimal codon usage in metabolic genes inform budding yeast ecology. *PLOS Biology*, *19*(4), e3001185.

19. Wolfe, K. H., & Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, *387*(6634), 708–713.

20. Marcet-Houben, M., & Gabaldón, T. (2015). Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. *PLoS Biology*, *13*(8), e1002220.

21. Haase, M. A. B., Kominek, J., Opulente, D. A., Shen, X.-X., LaBella, A. L., Zhou, X., DeVirgilio, J., Hulfachor, A. B., Kurtzman, C. P., Rokas, A., & Hittinger, C. T. (2021). Repeated horizontal gene transfer of *GAL*actose metabolism genes violates Dollo's law of irreversible loss. *Genetics*, *217*(iyaa012).

22. Benocci, T., Aguilar-Pontes, M. V., Kun, R. S., Seiboth, B., Vries, R. P. de, & Daly, P. (2018). *ARA1* regulates not only l-arabinose but also D-galactose catabolism in *Trichoderma reesei*. *FEBS Letters*, *592*(1), 60–70.

23. Venkatesh, A., Murray, A. L., Coughlan, A. Y., & Wolfe, K. H. (2021). Giant *GAL* gene clusters for the melibiose-galactose pathway in *Torulaspora*. *Yeast*, *38*(1), 117–126.

24. Ata, Ö., Rebnegger, C., Tatto, N. E., Valli, M., Mairinger, T., Hann, S., Steiger, M. G., Çalık, P., & Mattanovich, D. (2018). A single Gal4-like transcription factor activates the Crabtree effect in *Komagataella phaffii*. *Nature Communications*, *9*(1), 4911.

25. Martchenko, M., Levitin, A., & Whiteway, M. (2007). Transcriptional activation domains of the *Candida albicans* Gcn4p and Gal4p homologs. *Eukaryotic Cell*, *6*(2), 291–301.

26. Lang, G. I., & Botstein, D. (2011). A test of the coordinated expression hypothesis for the origin and maintenance of the *GAL* cluster in yeast. *PloS One*, *6*(9), e25290.

27. McGary, K. L., Slot, J. C., & Rokas, A. (2013). Physical linkage of metabolic genes in fungi is an adaptation against the accumulation of toxic intermediate compounds. *Proceedings of the National Academy of Sciences*, *110*(28), 11481–11486.

28. Xu, H., Liu, J.-J., Liu, Z., Li, Y., Jin, Y.-S., & Zhang, J. (2019). Synchronization of stochastic expressions drives the clustering of functionally related genes. *Science Advances*, *5*(10), eaax6525.

29. Rokas, A., Wisecaver, J. H., & Lind, A. L. (2018). The birth, evolution and death of metabolic gene clusters in fungi. *Nature Reviews Microbiology*, *16*(12), 731–744.

30. Traven, A., Jelicic, B., & Sopta, M. (2006). Yeast Gal4: A transcriptional paradigm revisited. *EMBO Reports*, *7*(5), 496–499.

31. Pan, T., & Coleman, J. E. (1989). Structure and function of the Zn(II) binding site within the DNA-binding domain of the GAL4 transcription factor. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(9), 3145–3149.

32. Roop, J. I., Chang, K. C., & Brem, R. B. (2016). Polygenic evolution of a sugar specialization trade-off in yeast. *Nature*, *530*(7590), 336–339.

33. Caudy, A. A., Guan, Y., Jia, Y., Hansen, C., DeSevo, C., Hayes, A. P., Agee, J., Alvarez-Dominguez, J. R., Arellano, H., Barrett, D., Bauerle, C., Bisaria, N., Bradley, P. H., Breunig, J. S., Bush, E., Cappel, D., Capra, E., Chen, W., Clore, J., … Dunham, M. J. (2013). A New System for Comparative Functional Genomics of *Saccharomyces* Yeasts. *Genetics*, *195*(1), 275–287.

34. Wang, J., Atolia, E., Hua, B., Savir, Y., Escalante-Chong, R., & Springer, M. (2015). Natural Variation in Preparation for Nutrient Depletion Reveals a Cost–Benefit Tradeoff. *PLOS Biology*, *13*(1), e1002041.

35. Escalante-Chong, R., Savir, Y., Carroll, S. M., Ingraham, J. B., Wang, J., Marx, C. J., & Springer, M. (2015). Galactose metabolic genes in yeast respond to a ratio of galactose and glucose. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(5), 1636–1641.

36. Ricci-Tam, C., Ben-Zion, I., Wang, J., Palme, J., Li, A., Savir, Y., & Springer, M. (2021). Decoupling transcription factor expression and activity enables dimmer switch gene regulation. *Science*, *372*(6539), 292–295.

37. Kuang, M. C., Hutchins, P. D., Russell, J. D., Coon, J. J., & Hittinger, C. T. (2016). Ongoing resolution of duplicate gene functions shapes the diversification of a metabolic network. *ELife*, *5*.

38. Kuang, M. C., Kominek, J., Alexander, W. G., Cheng, J.-F., Wrobel, R. L., & Hittinger, C. T. (2018). Repeated Cis-Regulatory Tuning of a Metabolic Bottleneck Gene during Evolution. *Molecular Biology and Evolution*, *35*(8), 1968–1981.

39. Libkind, D., Peris, D., Cubillos, F. A., Steenwyk, J. L., Opulente, D. A., Langdon, Q. K., Rokas, A., & Hittinger, C. T. (2020). Into the wild: New yeast genomes from natural environments and new tools for their analysis. *FEMS Yeast Research*, *20*(2).

40. Hénault, M., Eberlein, C., Charron, G., Durand, É., Nielly-Thibault, L., Martin, H., & Landry, C. R. (2019). Yeast Population Genomics Goes Wild: The Case of Saccharomyces paradoxus. In M. F. Polz & O. P. Rajora (Eds.), *Population Genomics: Microorganisms* (pp. 207–230). Springer International Publishing.

41. Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., Cruaud, C., Labadie, K., Aury, J.-M., Istace, B., Lebrigand, K., Barbry, P., Engelen, S., Lemainque, A., Wincker, P., … Schacherer, J. (2018). Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature*, *556*(7701), 339–344.

42. Boocock, J., Sadhu, M. J., Durvasula, A., Bloom, J. S., & Kruglyak, L. (2021). Ancient balancing selection maintains incompatible versions of the galactose pathway in yeast. *Science*, *371*(6527), 415–419.

43. Duan, S.-F., Shi, J.-Y., Yin, Q., Zhang, R.-P., Han, P.-J., Wang, Q.-M., & Bai, F.-Y. (2019). Reverse Evolution of a Classic Gene Network in Yeast Offers a Competitive Advantage. *Current Biology: CB*, *29*(7), 1126-1136.e5.

44. Duan, S.-F., Han, P.-J., Wang, Q.-M., Liu, W.-Q., Shi, J.-Y., Li, K., Zhang, X.-L., & Bai, F.-Y. (2018). The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. *Nature Communications*, *9*(1), 2690.

45. Legras, J.-L., Galeote, V., Bigey, F., Camarasa, C., Marsit, S., Nidelet, T., Sanchez, I., Couloux, A., Guy, J., Franco-Duarte, R., Marcet-Houben, M., Gabaldon, T., Schuller, D., Sampaio, J. P., & Dequin, S. (2018). Adaptation of *S. cerevisiae* to Fermented Food

Environments Reveals Remarkable Genome Plasticity and the Footprints of Domestication. *Molecular Biology and Evolution*, *35*(7), 1712–1727.

46. Varela, J. A., Puricelli, M., Ortiz-Merino, R. A., Giacomobono, R., Braun-Galleani, S., Wolfe, K. H., & Morrissey, J. P. (2019). Origin of Lactose Fermentation in *Kluyveromyces lactis* by Interspecies Transfer of a Neo-functionalized Gene Cluster during Domestication. *Current Biology*, *29*(24), 4284-4290.e2.
47. Griffiths, A. J. F., Doebley, J., Peichel, C., & Wassarman, D. A. (2020). *An Introduction to Genetic Analysis*. Macmillan Learning.
48. Jones, S. (2004). An overview of the basic helix-loop-helix proteins. *Genome Biology*, *5*(6), 226.
49. Wisecaver, J. H., & Rokas, A. (2015). Fungal metabolic gene clusters—Caravans traveling across genomes and environments. *Frontiers in Microbiology*, *6*, 161.
50. Feurtey, A., & Stukenbrock, E. H. (2018). Interspecific Gene Exchange as a Driver of Adaptive Evolution in Fungi. *Annual Review of Microbiology*, *72*, 377–398.

Box 1. Testing the Evolutionary Advantage(s) of *GAL* Gene Clustering

Whether the clustering of *GAL* (and sometimes other functionally related) genes is evolutionary advantageous remains a major, outstanding question and several genetic (e.g., coordinated gene expression, genetic linkage) and phenotypic (e.g., avoidance of toxic intermediates) models have been proposed to explain its origin and maintenance [26-29]. For example, it has been previously observed that metabolic genes in pathways with toxic intermediates are more often clustered together than those in pathways lacking them [27]. In the context of the *GAL* pathway, galactose-1-phosphate is a toxic intermediate, which leads the occurrence of a disease called galactosemia in humans lacking the functional pathway due to their inability to metabolize this intermediate. Furthermore, the metabolic genes encoding the enzymes involved in the production and conversion of a toxic intermediate were most often divergently oriented (an arrangement typical of co-regulated genes), such as *GAL1* and *GAL10* in *C. albicans* and *S. cerevisiae* (Figure 1), suggesting that clustering may be associated with reducing the buildup of the toxic intermediate [27]. In support of this hypothesis, a recent paper by Xu et al. found that experimental unlinking of the *GAL1, GAL7,* and *GAL10* genes in *S. cerevisiae* leads to higher fluctuations of their expression ratios – and higher buildup of the toxic intermediate galactose-1-phosphate – compared to when the three genes are clustered [28].

Outstanding Questions

- What is the contribution of the observed extensive segregating variation on *GAL* genes, including the maintenance of two incompatible versions of the pathway in some budding yeast populations, to the macroevolution of the *GAL* pathway?
- What, if any, is the function of the *GAL102* and *ORF-X* genes in the pathway?
- How many times has similar clustering of galactose metabolism genes independently evolved in budding yeasts, and what evolutionary pressures could be driving this adaptation (what do these yeast species have in common)?
- How do clustering and regulation affect HGT and maintenance of polymorphisms in the canonical *GAL* genes?
- What regulates galactose metabolism in budding yeast lineages that lack known transcription factors and/or their binding sites but have *GAL* genes and grow on galactose (e.g., in yeasts of the genus *Metschnikowia* or *Lipomyces*)?
- Can we infer with any degree of confidence the genomic organization and regulation of the *GAL* pathway of the budding yeast common ancestor or at various key nodes?

Figure Legends

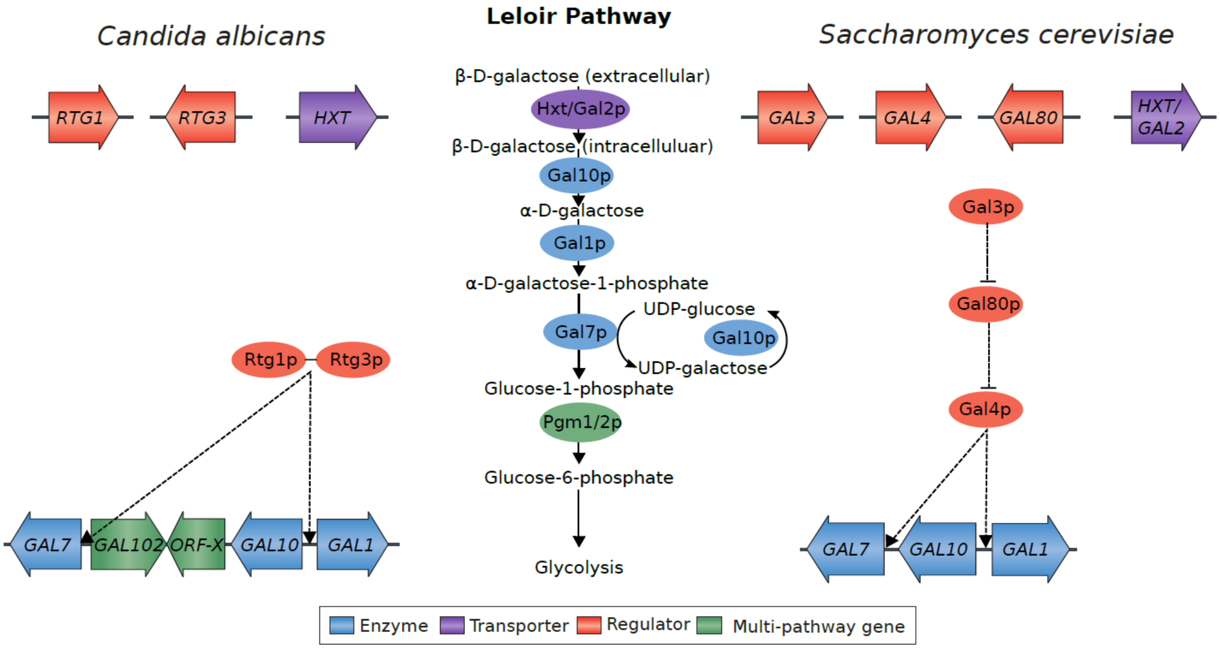

Figure 1. Comparison of genomic organization, function, and mode of regulation of the *GAL* pathway in the model organisms *C. albicans* (left panel) and *S. cerevisiae* (right panel). Although *GAL102* and *ORF-X* are nested within the *GAL* gene cluster in *C. albicans*, their functions are not known to be related to galactose assimilation. Information displayed in the figure based on: [3,8,21].

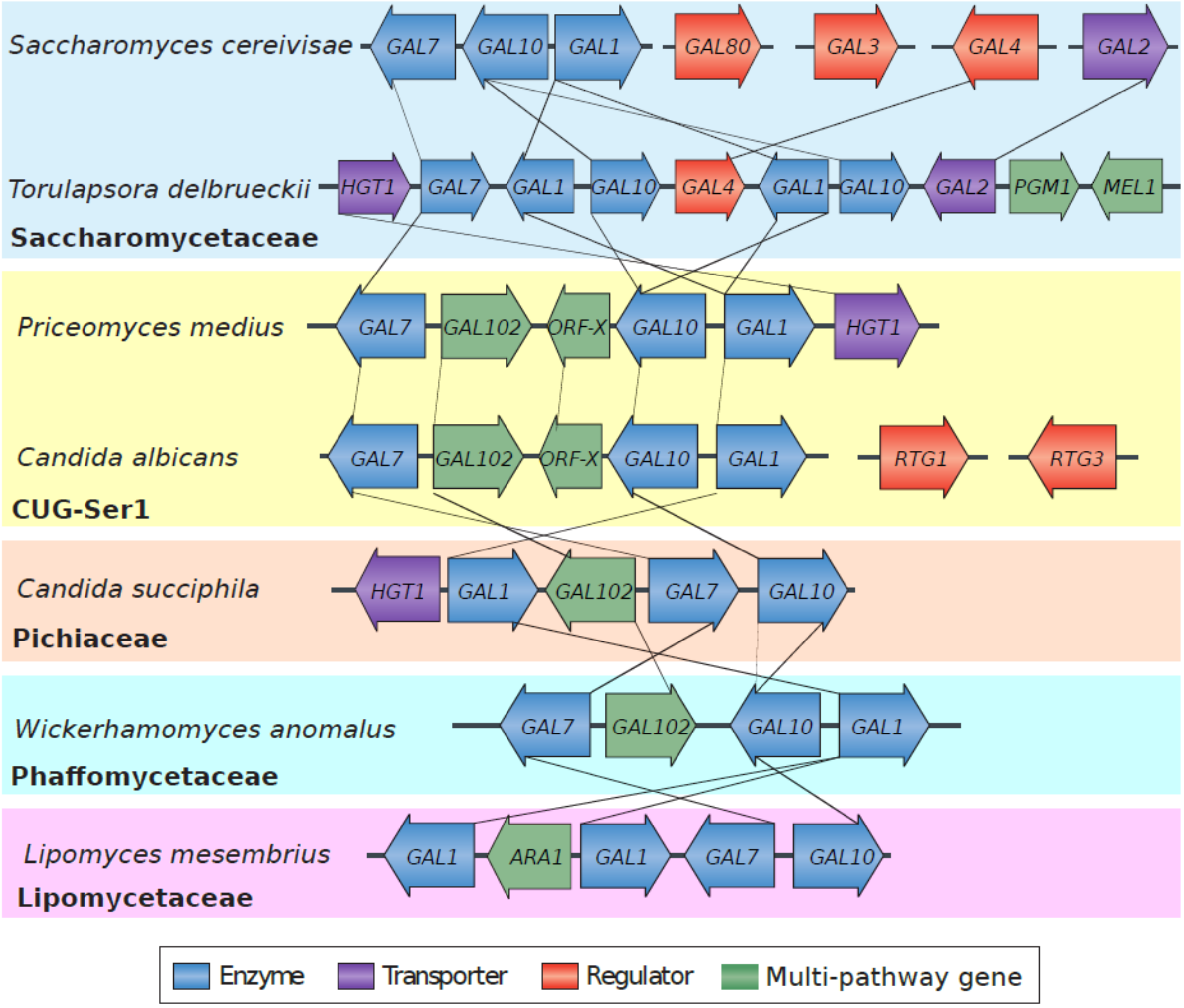

Figure 2. Genomic organization of *GAL* gene clusters in different budding yeast major clades. Note the differing patterns of presence / absence of *GAL* pathway genes in between major clades (indicated by the large colored rectangles). Lines correspond to homologs; gene box colors correspond to different functional categories of genes. Information displayed in the figure based on: [11,21,23].

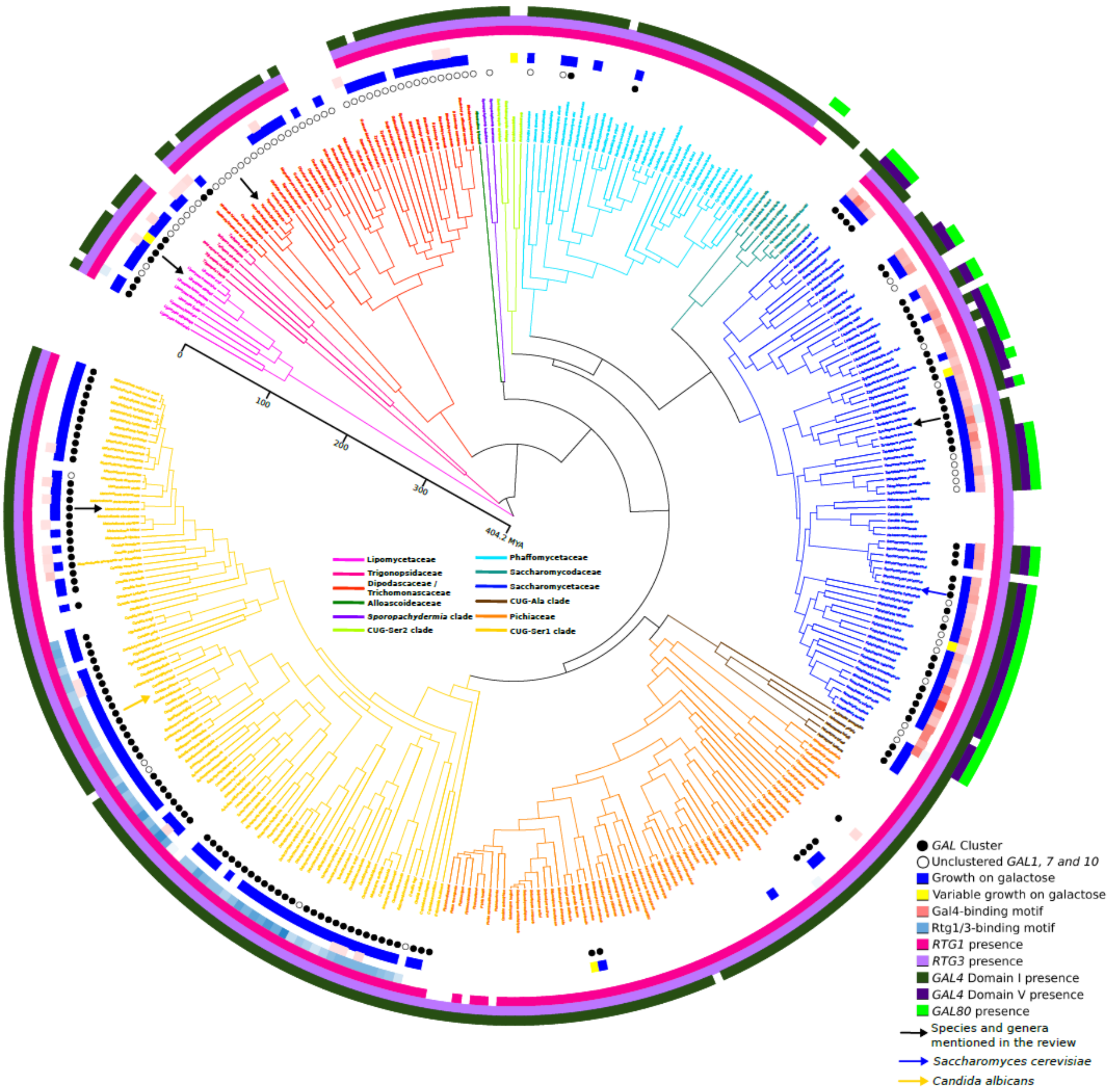

Figure 3. Gene presence and absence of key transcription factors involved in *GAL* pathway regulation, as well as variation in the presence and absence of their transcription factor-binding site sequence motifs across budding yeasts. Clustering is defined as *GAL1, GAL7*, and *GAL10* having 0-5 ORFs between them. Domain I of *GAL4* encodes amino acids 1-76 and Domain V amino acids 767-881 of the *S. cerevisiae* protein [30,31]. Information displayed in the figure based on [11,17,18].