Ancient balancing selection maintains incompatible versions of the galactose pathway in yeast

Boocock, James

Publication Date
2021

UNIVERSITY OF CALIFORNIA

Los Angeles

Ancient balancing selection maintains incompatible versions of the galactose pathway in yeast

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Human Genetics

by

James Boocock

2021

ABSTRACT OF THE DISSERTATION

Ancient balancing selection maintains incompatible versions of the galactose pathway in yeast

by

James Boocock
Doctor of Philosophy in Human Genetics
University of California, Los Angeles, 2021
Professor Leonid Kruglyak

Variation in nutrient availability between environments has led to the evolution of diverse metabolic pathways. These pathways differ across species, but are expected to be similar within a species. To identify large genetic differences in pathways within a single species, We performed a genome-wide scan for higher-order genetic interactions in segregants from 16 highly diverse S. cerevisiae crosses grown in 38 different conditions. We observed a large effect genetic interaction for growth in galactose among three loci in crosses involving the soil strain CBS2888. We used precisely engineered alleles to show that this genetic interaction arises from variation in the genes *GAL2*, *GAL1/10/7*, and *PGM1* from the galactose metabolic pathway. The CBS2888 alleles of these galactose genes were highly diverged from the reference. We hereafter refer to the divergent galactose alleles found in CBS2888 as the alternative alleles, and the alleles found in other strains as the reference alleles. Strains with alternative alleles are found primarily in galactose-rich dairy environments, and they grow faster in galactose, but slower in glucose, revealing a tradeoff, on which population genetics analyses suggest that balancing selection could have acted on. Our results show that balancing selection can preserve, functionally distinct states of a multi-locus genetic network, providing a general mechanism for maintenance of complex, interacting genetic variation at co-adapted alleles.

The dissertation of James Boocock is approved.

Dan Geschwind

Jake Lusis

Bogdan Pasanuic

Leonid Kruglyak, Committee Chair

University of California, Los Angeles

2021

This work is dedicated to my Mother, Father, Sisters, and my partner Elsie.

Table of contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

There are many people who made possible my journey across the ocean to pursue a PhD. Firstly, I would like to thank my parents and sisters for supporting me throughout my PhD, during the good times and a global pandemic. I would like to thank my partner Elsie for her constant support from New Zealand, and more recently in the US. I have learned loads from the Kruglyak lab members past and present, including Olga Schubert, Oliver Brandenburg, Eyal Ben-David, Longhua Guo, Meru Sadhu, Daniel Leighton, Alejandro Burga, Giancarlo Bruni, Tzitziki Vergara, Stefan Zdraljevic, Matthieu Delcourt, and Liangke Gou. I am grateful to my American friends Alex Chubick, Colin Farrell, and Arun Durvasala, for making me feel welcome. To all the Student Researchers and UAW staff I have organized with during the last two years of my PhD, I appreciate you, together we can make academia a better place to work. I would like to thank my commitee members, Bogdan Pasaniuc, Dan Geschwind, and Jake Lusis for their encouragement and advice. My Masters supervisor Tony Merriman introduced me to the facisnating puzzle of complex traits. My other Masters supervisor Mik Black encouraged me to study in the US, without him I would not be here today. I extend my thanks to Joshua Bloom a project scientist in the lab, for his guidance and friendship throughout my PhD. I hope we can continue to work together in the future. Finally, to my PhD advisor Leonid Kruglyak, thank you for your mentorship, guidance, and the awesome research culture you have created in your lab.

Permissions: Chapter 1 is a version of Boocock J, Sadhu MJ, Durvasula A, Bloom JS, Kruglyak L (2021) Ancient balancing selection maintains incompatible versions of the galactose pathway in yeast. Science 371.6527 (2021), 415-419.

The reprint is used here with permission from the authors and under the terms of the distribution license.

Funding: This work was supported by funding from the Howard Hughes Medical Institute and an NIH grant (2RO1GM102308-06 to L.K.). A.D. was supported by the NSF Graduate Research Fellowship (DGE-1650604).

VITA

EDUCATION

Masters of Science, The University of Otago, Dunedin, NZ	2016
Diploma for Graduates, The University of Otago, Dunedin, NZ	2013
BS Computer Science, The University of Otago, Dunedin, NZ	2012

PUBLICATIONS

Boocock J, Sadhu MJ, Durvasula A, Bloom JS, Kruglyak L. Ancient balancing selection maintains incompatible versions of the galactose pathway in yeast (2021). *Science.*

Bloom JS, **Boocock J**, Treusch S, Sadhu MJ, Day L, Oates-Barker H, Kruglyak L. Rare variants contribute disproportionately to quantitative trait variation in yeast (2021). *Elife.*

Ben-David E, **Boocock J**, Guo L, Zdraljevic S, Bloom JS, and Kruglyak L. Whole-organism mapping of the genetics of gene expression at cellular resolution (2019). *Elife.*.

Burga A, Ben-David E, Vergara TL, **Boocock J**, Kruglyak L. Fast genetic mapping of complex traits in C. elegans using millions of individuals in bulk (2019). *Nature Communications.*

Introduction

To grow and reproduce, organisms must extract chemicals from their environments and convert them into energy and cellular building blocks. Variation in nutrient avaliability between environments has led to the evolution of diverse and complex interconnecting metabolic enzymatic pathways. In humans, mutations in these pathways give rise to diseases known as inborn errors of metabolism [1]. The budding yeast *Saccharyomyces cerevisiae* has been extensively used as a model system for unravelling the genetic and biochemical basis of eukaryotic metabolism. Although S cerevisiae prefers glucose as a carbon source, it can utilize a wide variety of other sugars, and work over many decades has identified and characterized the regulatory and enzymatic components of pathays that process various carbon sources [2].

One exceptionally well-studied example of alternative carbon utilization is the Leloir pathway of galactose metabolism [3]. This pathway consists of a galactose transporter, encoded by the gene *GAL2*, enzymes that work together to convert galactose into glucose-1-phosphate, and regulatory components that control their expression. The enzymatic components are encoded by *GAL1*, *GAL10*, and *GAL7*, together known as the structural galactose genes. Phosphoglucomutase, encoded by *PGM1* and *PGM2*, then converts glucose-1-phosphate to glucose-6-phosphate—the substrate for glycolysis. The GAL genes are strongly repressed when either glucose or glycerol is present but are induced up to 1000-fold when only galactose is available [4]. Induction of the galactose pathway is controlled by the *GAL4* transcription factor and its regulators *GAL80* and *GAL3* [3].

Although the enzymes involved in galactose metabolism are highly conserved from yeast to humans, the regulation of galactose metabolism varies substantially [5, 6]. Within the yeast family *Saccharomycetaceae*, species show radically different responses to galactose [7]. For example, in *S. uvarum*, a species that is separated from *S. cerevisiae* by approximately 10-20 million years, GAL genes are not repressed by glucose [8, 9]. Some strains of S. kudriavzevii can metabolize galactose, while others have lost this ability through pseudogenization of multiple genes in the pathway, and it has been proposed that the two versions of the pathway have been maintained by multi-locus balancing selection [10]. Recent population surveys of S. cerevisae identified substantial sequence diversity in genes involved in galactose metabolism and uncovered some strains that lack glucose repression, a feature thought to be characteristic of galactose regulation in this species [11, 12, 13, 14].

In the course of mapping the genetic basis of variation in growth on multiple carbon sources, we discovered a three-way genetic interaction for growth in galactose [15]. Here, we fine-mapped these three loci to genes in the galactose pathway and identified specific allelic combinations of these genes that are incompatible for growth in galactose. We characterized the global distribution of these alleles in over 1000 yeast strains [11, 12] and found that most strains contain one of two functionally distinct versions of the galactose metabolic pathway that have been maintained by ancient balancing selection. We experimentally identified a fitness trade-off between these versions which provides a possible explanation for the maintenance of these alleles. These findings are explored in detail in Chapter 1.

References

[1] Jens Nielsen. Systems biology of metabolism. *Annual review of biochemistry*, 86:245–275, 2017.

[2] Jure Piškur and Concetta Compagno. *Molecular mechanisms in yeast carbon metabolism*. Springer, 2014.

[3] Christopher A Sellick, Robert N Campbell, and Richard J Reece. 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, 2008.

[4] D Lohr, P Venkov, and J Zlatanova. Transcriptional regulation in the yeast gal gene family: a complex genetic network. *The FASEB Journal*, 9(9):777–787, 1995.

[5] Chiraj K Dalal, Ignacio A Zuleta, Kaitlin F Mitchell, David R Andes, Hana El-Samad, and Alexander D Johnson. Transcriptional rewiring over evolutionary timescales changes quantitative and qualitative properties of gene expression. *Elife*, 5:e18981, 2016.

[6] Robert M Cohn and Stanton Segal. Galactose metabolism and its regulation. *Metabolism*, 22(4):627–642, 1973.

[7] Meihua Christina Kuang, Jacek Kominek, William G Alexander, Jan-Fang Cheng, Russell L Wrobel, and Chris Todd Hittinger. Repeated cis-regulatory tuning of a metabolic bottleneck gene during evolution. *Molecular biology and evolution*, 35(8):1968–1981, 2018.

[8] Jeremy I Roop, Kyu Chul Chang, and Rachel B Brem. Polygenic evolution of a sugar specialization trade-off in yeast. *Nature*, 530(7590):336–339, 2016.

[9] Meihua Christina Kuang, Paul D Hutchins, Jason D Russell, Joshua J Coon, and Chris Todd Hittinger. Ongoing resolution of duplicate gene functions shapes the diversification of a metabolic network. *Elife*, 5:e19027, 2016.

[10] Chris Todd Hittinger, Paula Gonçalves, José Paulo Sampaio, Jim Dover, Mark Johnston, and Antonis Rokas. Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature*, 464(7285):54–58, 2010.

[11] Jackson Peter, Matteo De Chiara, Anne Friedrich, Jia-Xing Yue, David Pflieger, Anders Bergström, Anastasie Sigwalt, Benjamin Barre, Kelle Freel, Agnès Llored, et al. Genome evolution across 1,011 saccharomyces cerevisiae isolates. *Nature*, 556(7701):339–344, 2018.

[12] Shou-Fu Duan, Pei-Jie Han, Qi-Ming Wang, Wan-Qiu Liu, Jun-Yan Shi, Kuan Li, Xiao-Ling Zhang, and Feng-Yan Bai. The origin and adaptive evolution of domesticated populations of yeast from far east asia. *Nature communications*, 9(1):1–13, 2018.

[13] Jean-Luc Legras, Virginie Galeote, Frédéric Bigey, Carole Camarasa, Souhir Marsit, Thibault Nidelet, Isabelle Sanchez, Arnaud Couloux, Julie Guy, Ricardo Franco-Duarte, et al. 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, 2018.

[14] Shou-Fu Duan, Jun-Yan Shi, Qi Yin, Ri-Peng Zhang, Pei-Jie Han, Qi-Ming Wang, and Feng-Yan Bai. Reverse evolution of a classic gene network in yeast offers a competitive advantage. *Current Biology*, 29(7):1126–1136, 2019.

[15] Joshua S Bloom, James Boocock, Sebastian Treusch, Meru J Sadhu, Laura Day, Holly Oates-Barker, and Leonid Kruglyak. Rare variants contribute disproportionately to quantitative trait variation in yeast. *Elife*, 8:e49212, 2019.

Chapter 1: Ancient balancing selection maintains incompatible versions of the galactose pathway in yeast

METABOLIC EVOLUTION

Ancient balancing selection maintains incompatible versions of the galactose pathway in yeast

James Boocock[1,2,3,4], Meru J. Sadhu[1,2,3,4]*, Arun Durvasula[1], Joshua S. Bloom[1,2,3,4]†, Leonid Kruglyak[1,2,3,4]†

Metabolic pathways differ across species but are expected to be similar within a species. We discovered two functional, incompatible versions of the galactose pathway in *Saccharomyces cerevisiae*. We identified a three-locus genetic interaction for growth in galactose, and used precisely engineered alleles to show that it arises from variation in the galactose utilization genes *GAL2*, *GAL1/10/7*, and phosphoglucomutase (*PGM1*), and that the reference allele of *PGM1* is incompatible with the alternative alleles of the other genes. Multiloci balancing selection has maintained the two incompatible versions of the pathway for millions of years. Strains with alternative alleles are found primarily in galactose-rich dairy environments, and they grow faster in galactose but slower in glucose, revealing a trade-off on which balancing selection may have acted.

Variation in nutrient availability between environments has led to the evolution of diverse metabolic pathways. In humans, mutations in these pathways give rise to diseases known as inborn errors of metabolism (*1*). The budding yeast *Saccharomyces cerevisiae* is commonly used for studying eukaryotic metabolism (*2*). A classic well-studied pathway for galactose metabolism includes a galactose transporter, encoded by the gene *GAL2*, and the enzymes encoded by *GAL1*, *GAL10*, and *GAL7*, which convert galactose to glucose-1-phosphate (*3*). Phosphoglucomutase, encoded by *PGM1* and *PGM2*, then converts glucose-1-phosphate to glucose-6-phosphate—the substrate for glycolysis.

Within the same genus, some strains of *Saccharomyces kudriavzevii* can metabolize galactose, whereas others have lost this ability through pseudogenization of multiple genes in the pathway, and it was proposed that the two versions of the pathway have been maintained through multiloci balancing selection (*4*). Balancing selection maintains genetic diversity against the forces of genetic drift and has typically been observed to act on single loci (*5*). Multiloci balancing selection is expected to be extremely rare because it has to overcome the independent segregation of alleles at the different loci.

We studied growth in galactose in a large set of crosses in *S. cerevisiae* (*6*) and observed a genetic interaction among three loci in crosses involving the soil strain CBS2888 (three-way effect size 0.19 SD units, chi-square test, $P < 10^{-15}$) (Fig. 1A, figs. S1 and S2, and tables S1 to S3). The nonadditive nature of the effects of the three loci is best illustrated by the phenotype of segregants that inherit the CBS2888 allele at the loci on chromosome II (ChrII) and ChrXII and the non-CBS2888 allele at the locus on ChrXI; these segregants grow much slower in galactose than those with any other combination of alleles (Fig. 1A). The three loci contain genes that encode components of galactose

[1]Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA, USA. [2]Department of Biological Chemistry, University of California, Los Angeles, Los Angeles, CA, USA. [3]Howard Hughes Medical Institute, University of California, Los Angeles, Los Angeles, CA, USA. [4]Institute for Quantitative and Computational Biology, University of California, Los Angeles, Los Angeles, CA, USA.
*Present address: Genetic Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA.
†Corresponding author. Email: jbloom@mednet.ucla.edu (J.S.B.); lkruglyak@mednet.ucla.edu (L.K.)

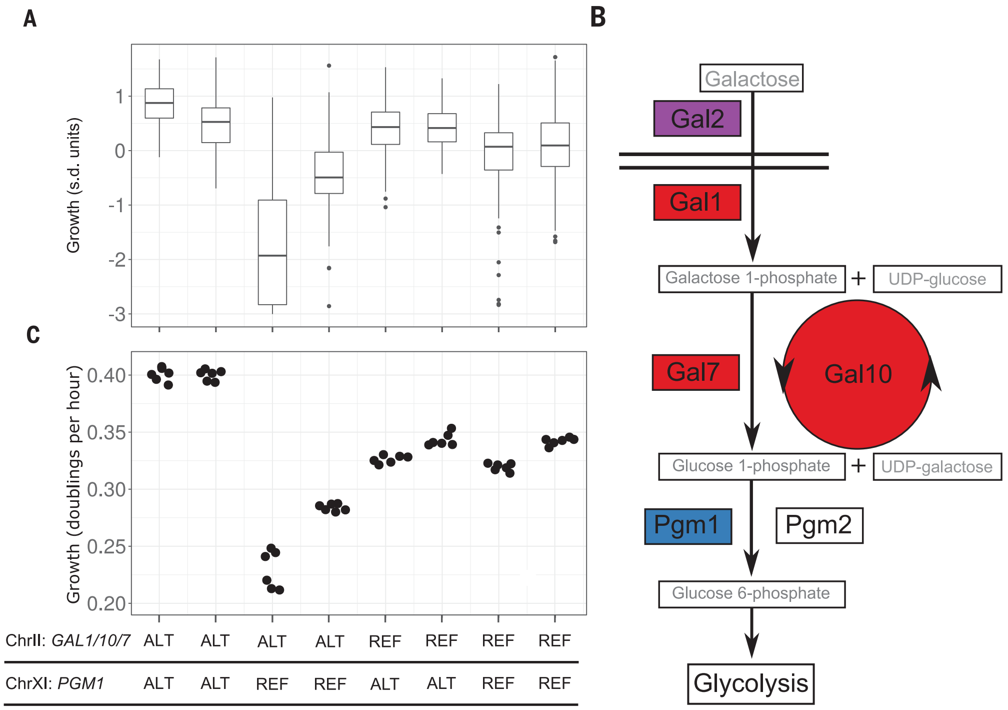


Fig. 1. Three-locus genetic interaction for growth in galactose. (**A**) Boxplots show growth in galactose of yeast segregants (*n* = 867) derived from a cross between CBS2888 and YJM981. Each boxplot corresponds to segregants with one of eight distinct combinations of alleles at the three loci (ChrII, *GAL1/10/7*; ChrXI, *PGM1*; and ChrXII, *GAL2*). (**B**) Galactose metabolic pathway. Components of the pathway corresponding to the three loci are shown in different colors. UDP, uridine 5′-diphosphate. (**C**) Growth of allele replacement strains in galactose. BY alleles [reference (REF)] and CBS2888 alleles [alternative (ALT)].

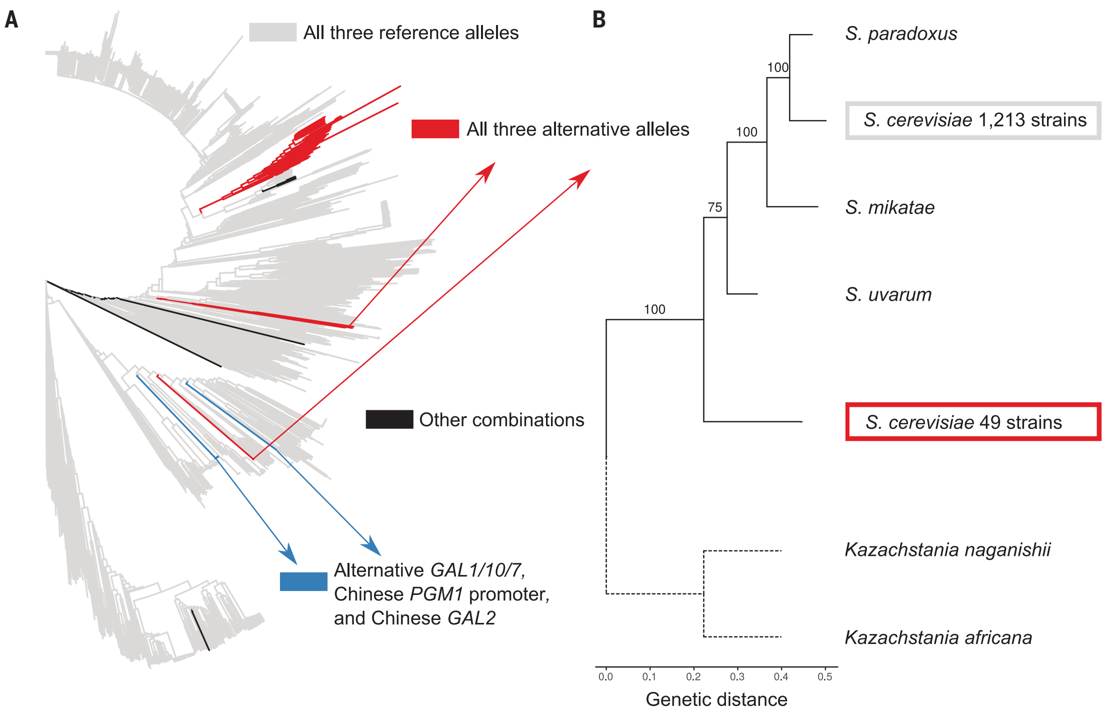


Fig. 2. The alternative galactose alleles are broadly distributed and fall outside the *Saccharomyces* genus. (**A**) Genome-wide neighbor-joining tree of 1276 sequenced yeast isolates. (**B**) Bootstrapped maximum likelihood phylogenetic tree of the *GAL1/10/7* alleles from CBS2888 (alternative), BY (reference), other species in the *Saccharomyces* genus, and two outgroup species. The outgroup branches (dotted lines) were rescaled to the average branch length.

metabolism: *GAL1*, *GAL10*, and *GAL7* on ChrII; *PGM1* on ChrXI; and *GAL2* on ChrXII (Fig. 1B and fig. S1) (*3*). The CBS2888 alleles of these genes were highly diverged from the reference (fig. S3 and table S4). We hereafter refer to the divergent galactose alleles found in CBS2888 as the alternative alleles and alleles observed in the other strains as the reference alleles.

We used CRISPR-Cas9 to engineer strains with all eight possible combinations of the three alternative and three reference galactose alleles in a common genetic background (fig. S4 and tables S5 to S7) (*7*, *8*). We then measured the growth rates of the eight engineered strains in galactose and recapitulated the mapping results (Fig. 1C, fig. S5, and tables S1 and S2), demonstrating that variants in the coding and intergenic regions of *GAL1/10/7* and *GAL2* and in the promoter region of *PGM1* are responsible for the observed genetic interaction. In particular, the strain with the reference *PGM1* promoter allele and the alternative *GAL1/10/7* and *GAL2* alleles exhibited a severe growth defect in galactose, confirming that the components of the reference and alternative pathways are incompatible.

To better understand cis-acting regulatory differences between the alternative and reference hybrid strain (CBS2888xBY) in glucose, transferred it to galactose medium, and sequenced RNA from samples collected throughout a growth time course (*7*). In glucose, the expression of the CBS2888 allele of *PGM1* was slightly lower than that of the reference allele (fig. S6). By contrast, 1 hour after the switch to galactose, the expression of the CBS2888 allele of *PGM1* was 15.5 times greater than that of the reference allele (binomial test, $P < 10^{-100}$), and this difference persisted for the rest of the time course (fig. S6 and table S8).

The alternative *PGM1* promoter allele contains a *GAL4* upstream activating sequence (UAS) (*10*), whereas the reference allele does not (fig. S4). We engineered a point mutation disrupting the UAS in a strain with all three alternative galactose alleles (*7*). This single mutation recapitulated the growth defect in galactose observed in a strain with a combination of the reference allele of the *PGM1* promoter and alternative alleles of the other *GAL* genes (fig. S7). We conclude that the induction of *PGM1* in galactose, mediated through a *GAL4* UAS, is critical for the proper functioning of the alternative galactose pathway.

We searched for the alternative and reference of sequenced *S. cerevisiae* isolates comprising 1276 strains (*11*, *12*) and found three common combinations: only reference alleles (1213 strains), only alternative alleles (49 strains), and 8 strains from China with the alternative *GAL1/10/7* allele and alleles of *GAL2* and the *PGM1* promoter that differ from both the reference and the alternative alleles (Fig. 2A, fig. S8, and table S9) (*7*). No strains carried the reference *PGM1* promoter allele and the alternative *GAL1/10/7* and *GAL2* alleles, suggesting that this combination causes a fitness disadvantage in natural environments and has been purged by selection. This hypothesis is further supported by a high linkage disequilibrium index (ε = 0.59) for the three loci (fig. S9) (*13*).

The alternative galactose alleles are fixed in two lineages of strains found in dairy products, including Camembert cheese from France, kefir grains from Japan, and fermented yak and goat milk from China (table S9). These environments are rich in lactose, a disaccharide of glucose and galactose. *S. cerevisiae* relies on the activity of other fungi and bacteria to break down lactose into glucose and galactose, which it then metabolizes (*14*). This observation suggests that the alternatives alleles are maintained by natural selection in dairy

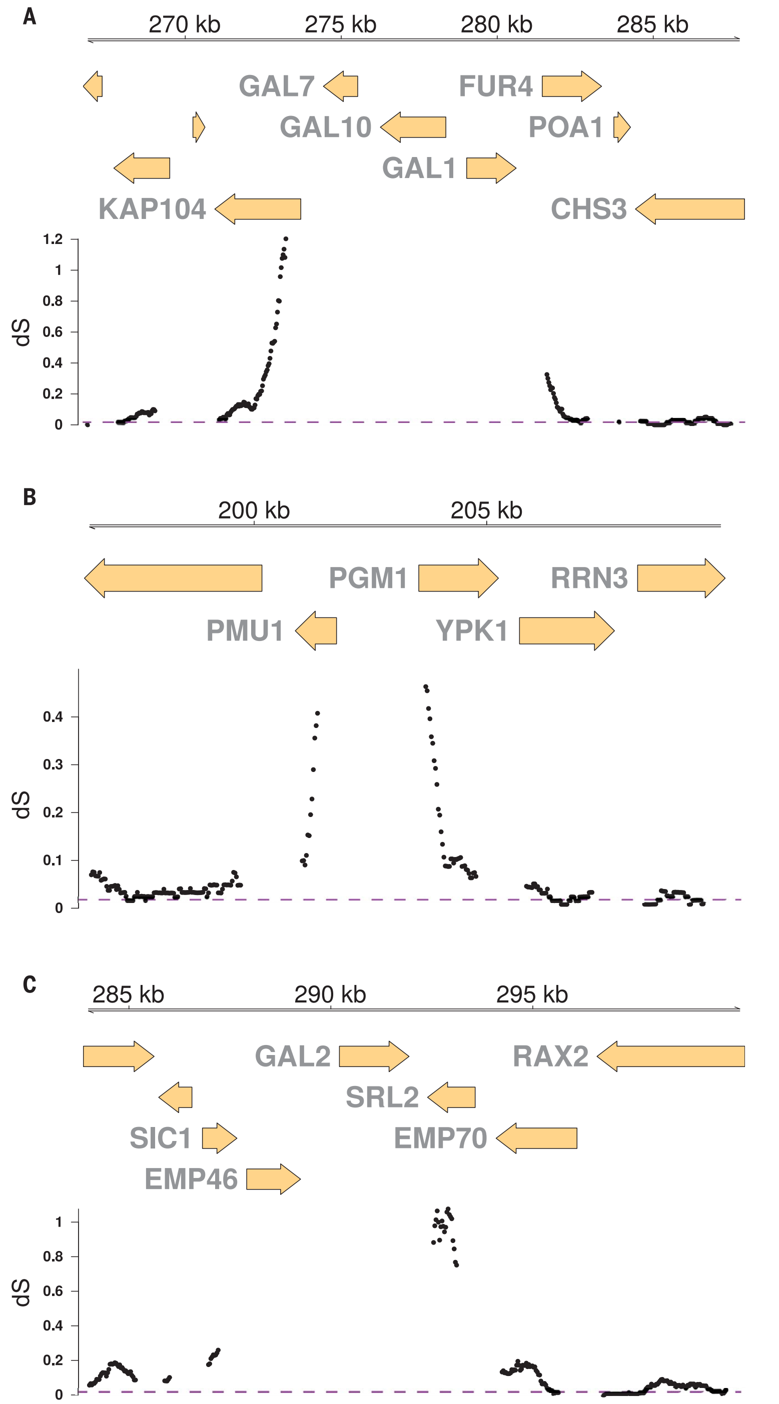


Fig. 3. A signature of ancient balancing selection. Estimated rate of synonymous substitutions per site (dS) in 200-codon windows stepped every 10 base pairs between CBS2888 (alternative) and BY (reference) genes surrounding the galactose loci. (**A** to **C**) Genes adjacent to *GAL1/10/7* (A), genes adjacent to the *PGM1* promoter (B), and genes adjacent to *GAL2* (C). The purple dashed line shows the genome-wide average dS of 0.014. The

We dated the split between the alternative and reference galactose alleles to ~3.2 billion generations ago (95% confidence interval = 2.5 to 4.5 billion generations), which predates the most recent common ancestor of the *Saccharomyces* genus (figs. S10 and S11 and tables S4 and S10) (*7*, *15*). Phylogenetic clustering placed the alternative galactose alleles outside the *Saccharomyces* genus and supports an ancient origin of the alternative alleles (Fig. 2B and figs. S12 and S13) (*7*, *16*).

One force that can maintain highly diverged alleles within a species is balancing selection. This process is expected to generate a signature of elevated sequence divergence at linked neutral sites that decays with genetic distance from the selected variant (*5*). We examined the rate of synonymous substitutions per site (dS) across the CBS2888 genome relative to the reference and observed a strong signature of ancient balancing selection at all three galactose loci (Fig. 3 and figs. S14 to S18) (*7*).

The strains with the alternative or Chinese alleles contain *GAL2* genes duplicated in tandem, and *GAL2* is also duplicated in two other yeast species: *Saccharomyces uvarum* and *Saccharomyces eubayanus*. We aligned all the *GAL2* paralogs and observed that the N-terminal cytosolic regions (amino acids 1 to 67) were highly dissimilar within species and phylogenetically clustered across species (fig. S19). These results suggest that the N-terminal regions of the *GAL2* paralogs are functionally distinct and maintained by selection, and they also provide evidence that the alternative alleles have an ancient origin in *Saccharomyces* (fig. S20).

It has been proposed that the alternative galactose alleles arose through introgression around the time humans domesticated milk-producing animals, but no species that could have donated the alleles has been identified (*7*, *17*). A relatively recent introgression would generate a sharp boundary between dS at the *GAL* genes and the rest of the genome. Instead, our data suggest that the variation at these loci has accumulated within *S. cerevisiae* over time.

We performed forward genetic simulations to distinguish between scenarios that could have given rise to the observed signatures of balancing selection (figs. S21 to S23) (*7*). Models of recent introgression (<50 million generations ago), with or without balancing selection, were not well-supported when compared with a model of ancient balancing selection (figs. S24 to S27 and table S11) (*7*).

Balancing selection can act on fitness tradeoffs, in which alleles with higher fitness in one environment have lower fitness in another (*5*). Although all strains grow faster in glucose than in galactose [*t* statistic (*T*) = 7.80, *t* test, $P < 10^{-5}$], the strains with the alternative alleles grow

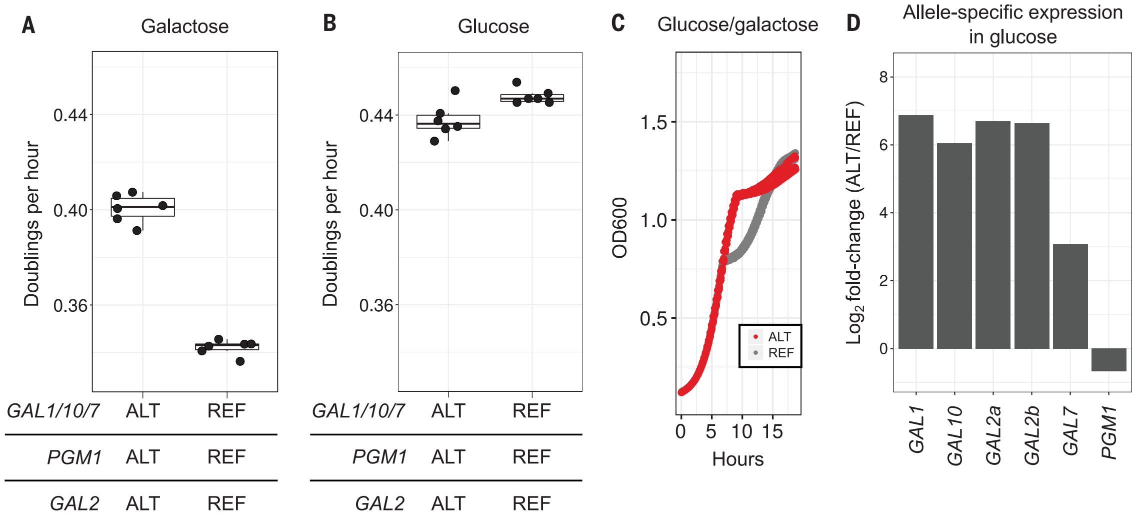


Fig. 4. Trade-offs between the alternative and reference alleles in the galactose pathway. (**A**) Growth rates of allele replacement strains (n = 6) with all three reference (right) or all three alternative (left) alleles in galactose as a sole carbon source. (**B**) As in (A), but for cells grown in glucose. (**C**) Growth curves of allele replacement strains with all three alternative (red) or reference (gray) galactose alleles in mixed glucose and galactose medium. OD600, optical density at 600 nm. (**D**) Allele-specific expression of the galactose genes from a diploid hybrid (CBS2888xBY) strain grown in glucose.

reference alleles (Figs. 1C and 4A). *S. cerevisiae* encounters and metabolizes a wide variety of sugars (*18*) but prefers glucose (*19*). In glucose, the strains with the reference alleles grow 2% faster than strains with the alternative alleles (T = –3.12, t test, P = 0.017) (Fig. 4B and fig. S28). This faster growth provides an explanation for the maintenance of the reference alleles in the strains that do not frequently encounter galactose.

In the strains with reference alleles, the *GAL* genes are robustly repressed by glucose and induced by galactose (*3*). This leads to a pause in growth known as the diauxic shift, when yeast switch from metabolizing glucose to metabolizing galactose. Strains with the three alternative galactose alleles do not undergo a diauxic shift (Fig. 4C and fig. S29). RNA sequencing showed that in glucose, the reference alleles are repressed, whereas the alternative *GAL* alleles are constitutively expressed (fold change = 40.6, binomial test, $P < 10^{-16}$) (Fig. 4D, fig. S30, and table S8) (*7*). The constitutive expression of the *GAL* genes eliminates the diauxic shift (*20*), providing a fitness benefit when galactose is encountered. However, gene expression can be costly (*21*), and this could explain why the alternative galactose pathway leads to a growth disadvantage in glucose.

The incompatible allele combinations we identified may provide a model for classical galactosemia, an inborn error of metabolism caused by recessive mutations in *GALT*, the human homolog of *GAL7* (*22*), that can lead to life-threatening symptoms if galactose is not eliminated from diet. The precise molecular basis of galactosemia is not fully understood (*23*), but yeast models of galactose toxicity suggest that the incompatibility observed in this work arises from the same metabolic defect that underlies galactosemia. Finally, our results go beyond previous findings (*4*) in showing that balancing selection can preserve two alternate, functionally distinct states of a multiloci genetic network, providing a general mechanism for the maintenance of complex, interacting genetic variation at coadapted alleles.

REFERENCES AND NOTES

1. J. M. Saudubray, À. Garcia-Cazorla, *Pediatr. Clin. North Am.* **65**, 179–208 (2018).
2. J. Nielsen, *Annu. Rev. Biochem.* **86**, 245–275 (2017).
3. C. A. Sellick, R. N. Campbell, R. J. Reece, *Int. Rev. Cell Mol. Biol.* **269**, 111–150 (2008).
4. C. T. Hittinger *et al.*, *Nature* **464**, 54–58 (2010).
5. D. Charlesworth, *PLOS Genet.* **2**, e64 (2006).
6. J. S. Bloom *et al.*, *eLife* **8**, e49212 (2019).
7. Materials and methods and supplementary text are available as supplementary materials.
8. M. J. Sadhu *et al.*, *Nat. Genet.* **50**, 510–514 (2018).
9. F. W. Albert, L. Kruglyak, *Nat. Rev. Genet.* **16**, 197–212 (2015).
10. A. Traven, B. Jelicic, M. Sopta, *EMBO Rep.* **7**, 496–499 (2006).
11. S.-F. Duan *et al.*, *Nat. Commun.* **9**, 2690 (2018).
12. J. Peter *et al.*, *Nature* **556**, 339–344 (2018).
13. Y. Okada, *Hum. Genome Var.* **5**, 29 (2018).
14. J. L. Legras *et al.*, *Mol. Biol. Evol.* **35**, 1712–1727 (2018).
15. M. Kellis, N. Patterson, M. Endrizzi, B. Birren, E. S. Lander, *Nature* **423**, 241–254 (2003).
16. D. R. Scannell *et al.*, *Genes Genomes Genet.* **1**, 11–25 (2011).
17. S. F. Duan *et al.*, *Curr. Biol.* **29**, 1126–1136.e5 (2019).
18. B. Turcotte, X. B. Liang, F. Robert, N. Soontorngun, *FEMS Yeast Res.* **10**, 2–13 (2010).
19. J. H. Kim, A. Roy, D. Jouandot II, K. H. Cho, *Biochim. Biophys. Acta* **1830**, 5204–5210 (2013).
20. J. I. Roop, K. C. Chang, R. B. Brem, *Nature* **530**, 336–339 (2016).
21. G. I. Lang, A. W. Murray, D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 5755–5760 (2009).
22. A. I. Coelho, M. E. Rubio-Gozalbo, J. B. Vicente, I. Rivera,
23. K. Lai, L. J. Elsas, K. J. Wierenga, *IUBMB Life* **61**, 1063–1074 (2009).
24. J. Boocock, M. J. Sadhu, A. Durvasula, J. S. Bloom, L. Kruglyak, Ancient balancing selection maintains incompatible versions of the galactose pathway in yeast (Figure creation), Dryad (2020); https://doi.org/10.5068/D14370.
25. J. Boocock, theboocock/ancient_bal_scripts: submission, Zenodo (2020); https://doi.org/10.5281/zenodo.4132713.
26. J. Boocock, theboocock/popgen_utilities: submission, Zenodo (2020); https://doi.org/10.5281/zenodo.4131787.
27. J. Boocock, theboocock/gal_bal: submission, Zenodo (2020); https://doi.org/10.5281/zenodo.4107954.

ACKNOWLEDGMENTS

We thank O. Schubert, E. Ben-David, L. Guo, and S. Zdraljevic for helpful manuscript feedback and edits. **Funding:** This work was supported by funding from the Howard Hughes Medical Institute and an NIH grant (2R01GM102308-06 to L.K.). A.D. is supported by the NSF Graduate Research Fellowship (DGE-1650604). **Author contributions:** J.B. performed experiments with assistance from M.J.S. and J.S.B.; J.B. analyzed data with assistance from J.S.B. and A.D.; L.K. and J.S.B. supervised the project. J.B., A.D., J.S.B., and L.K. wrote the manuscript. All authors discussed and agreed on the final version of the manuscript. **Competing interests:** The authors declare no competing financial interests. **Data and materials availability:** Sequencing data are available under NCBI BioProject PRJNA575066. The processed datasets are available on Dryad (*24*). The following code is available on Zenodo: custom scripts used to analyze processed data and create figures (*25*), a custom Python package used with our population genetic analysis (*26*), and the code used to perform and analyze the forward genetic simulations (*27*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/371/6527/415/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S30
References (*28*–*58*)
Tables S1 to S11
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

1 November 2019; resubmitted 8 May 2020
Accepted 17 December 2020

Science

Ancient balancing selection maintains incompatible versions of the galactose pathway in yeast

James Boocock, Meru J. Sadhu, Arun Durvasula, Joshua S. Bloom and Leonid Kruglyak

Science **371** (6527), 415-419.
DOI: 10.1126/science.aba0542

Yeast switches for glucose and galactose

Some organisms can switch metabolic pathways depending on their environment. One such example is yeast, which can transition between the sugars glucose and galactose as carbon sources. Boocock *et al.* show that this ability has undergone selection, resulting in the maintenance of two incompatible metabolic pathways in a select set of yeast strains within a single species. A phylogenetic analysis supports that these different pathways are mediated by three genes that differ between strains within and among yeast species and likely have been maintained over 10 million to 20 million years.

Science, this issue p. 415

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.

science.sciencemag.org/content/371/6527/415/suppl/DC1

Supplementary Material for

Ancient balancing selection maintains incompatible versions of the galactose pathway in yeast

James Boocock, Meru J. Sadhu, Arun Durvasula, Joshua S. Bloom*, Leonid Kruglyak*

*Corresponding author. Email: jbloom@mednet.ucla.edu (J.S.B.); lkruglyak@mednet.ucla.edu (L.K.)

Published 22 January 2021, *Science* **371**, 415 (2021)
DOI: 10.1126/science.aba0542

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S30
References

Other Supplementary Material for this manuscript includes the following:
(available at science.sciencemag.org/content/371/6527/415/suppl/DC1)

Tables S1 to S11
MDAR Reproducibility Checklist

Materials and Methods

Quantitative trait mapping

We obtained the additive QTL for each of the 38 phenotypes that were measured in ~14,000 progeny from 16 parental crosses in Bloom et al (*6*). We tested all triplets of these QTL to see whether they were involved in any three-way QTL interactions. Specifically, for unique triplet combination of QTLs for each trait and cross we built a linear model with all additive QTLs, two-way QTLs, and the three-way QTL interaction terms. We tested whether this model fit significantly better than a nested model that included only additive QTLs and two-way interaction terms using a likelihood ratio test. We considered any three-way interaction with a q-value less 0.1 to be significant. The coefficient of variation, R^2, was used to quantify how much variance in the phenotype was explained by the different QTL models. Boxplots of the normalized growth rate were made using ggplot2 (v3.2.0)(*28*). A chi-square test was used to evaluate whether the three galactose loci segregated independently. Unless otherwise specified, all computational analyses were performed in R (v3.6.1)(*29*).

Yeast strains

Strains, plasmids, and primers used in this study are listed in Tables S5 to S7. To generate allele replacement strains, we used a two-guide RNA CRISPR system to introduce double-strand breaks flanking our regions of interest and provided linear repair templates of the desired replacement allele. The precise details for the construction of each of the strains are described below.

We engineered a lab yeast strain derived from BY4741 (YLK3221: Mata met15Δ his3Δ1 leu2Δ0 ura3Δ0 nej1Δ::KanMX) with all eight combinations of the CBS2888 and BY4741 galactose alleles (*8*). To engineer strains with the CBS2888 *PGM1* promoter we used a plasmid that contained galactose inducible *CAS9* (PLK77, p415-GalL-Cas9-CYC1t)(*30*). For the CBS2888 *GAL2* and *GAL1/10/7* allele, we used a plasmid that contained a constitutively expressed *CAS9* (PLK91, pRS414-TEF1p-Cas9-CYC1t-NatMx).

For the *PGM1* locus, we replaced the BY4741 *PGM1* promoter sequence with the CBS2888 *PGM1* promoter allele. We generated a repair template that contained the 2.9kb CBS2888 promoter sequence flanked by homology arms that were identical to the ends of the nearest flanking genes, *PMU1* and *PGM1*. For the *GAL2* locus, we replaced the BY4741 *GAL2* sequence and the surrounding non-coding region with the CBS2888 *GAL2* allele. We generated a repair template that contained the 5.7kb CBS28888 GAL2 sequence flanked by homology arms that were identical to the ends of the nearest flanking genes *EMP46* and *SRL2*. For the *GAL1/10/7* locus, we replaced the BY4741 *GAL1/10/7* sequence and the surrounding non-coding region with the CBS2888 *GAL1/10/7* allele. We generated a repair template that contained the 7.3kb CBS2888 *GAL1/10/7* sequence flanked by homology arms that were identical to the ends of the nearest flanking genes *KAP104* and. We co-transformed these repair templates with selectable plasmids expressing two guide-RNAs that exclusively cut near the 3' and 5' ends of the BY4741 region. We picked colonies and confirmed they had the exact intended replacement allele sequences using multiple sanger sequencing reactions (YLK3267, YLK3268, YLK3269).

We mated strains with the CBS2888 *GAL2* and *PGM1* promoter to obtain a heterozygous diploid (YLK3270), which we sporulated to get a haploid strain with both CBS2888 alleles. This strain was then mated to a strain with the CBS2888 *GAL1/10/7* allele to obtain a heterozygous diploid (YLK3271). We also took care to ensure that auxotrophies and the drug marker were homozygous in this strain. We sporulated this strain and PCR genotyped the progeny. From this cross, we obtained two isolates of each of the eight possible combinations of the CBS2888 and BY4741 galactose alleles.

We made a point mutation in the GAL4-binding site of the CBS2888 promoter in strains with the CBS2888 *GAL1/10/7*, *PGM1* promoter, and *GAL2*. We used a plasmid that expressed a single guide RNA, which was co-transformed with a repair template. We confirmed that strains had the desired mutation using sanger sequencing (YLK3288-YLK3291).

Growth measurements.

All growth experiments were performed at 30 degrees in YP media (2% bacto-peptone, 1% yeast extract) supplemented with 2% glucose, 2% galactose, or 1% glucose/1% galactose. Strains were always incubated with fast shaking in a BioTek Synergy™ 2 plate reader. Before each experiment, strains were grown to saturation in our plate reader in 96-well plates (Corning, Flat Bottom with Lid, #3370) in 2 % glucose. Strains were then diluted 1:100 into new 96-well plates and transferred to our plate reader, which automatically took optical density measurements (OD600) measurements every 15 minutes.

Growth rate calculations

Growth rate was quantified as the geometric mean rate of growth (GMR). Our procedure for calculating the GMR follows that described in Brem et al (*20*). Briefly, a spline was fit in R using the splinefun function, and the time spent (t) between OD 0.2 and 0.8 was calculated. The GMR was then estimated as the $\log(0.8/0.2)\ /t$. We then converted this GMR into doublings per hour. In each experiment, we measured the growth of three biological replicates of each strain. To determine whether the diauxic shift differed between our allele replacement strains, we grew these strains in 1% glucose/1% galactose medium and calculated the GMR between 0.8 and 1.1. This range of OD captures the range over which the reference strain pauses and restarts growth in 1% glucose / 1 % galactose medium.

To determine whether our allele replacement experiments recapitulated our QTL results, we fit a linear model with additive, two-way, and three-way interactions terms. We performed a likelihood ratio test to determine whether this model fit significantly better than the two-way interaction model. The coefficient of variation R^2 was used to quantify the variance explained by the different QTL models. In the model with all additive and interaction terms we used a t-test to determine whether the three-way interaction term was significant.

In all of our growth experiments, we performed Welch's t-tests to determine whether certain allele replacement strains had significantly different growth rates from each other.

To align the growth curves across a 96-well experiment for visualization purposes, for each well we identified the last time point that was less than OD 0.2, and the first time point that was greater than or equal to OD 0.2. We linearly interpolated these time-points to get an estimate for t at OD = 0.2 and calculated an adjusted time for all wells.

Allele-specific expression of a hybrid diploid strain (CBS2888xBY)
throughout a galactose-induction time-course

For our allele-specific RNA-sequencing experiments we mated a prototrophic BY strain (YLK1881) to CBS2888, to obtain a diploid hybrid. We performed an RNA-seq experiment for this diploid hybrid strain over a time course growth experiment where the strain was transferred from YP +2% glucose media to YP+ 2% galactose media. In more detail, we collected yeast from mid-log (OD ~ 0.5) in glucose media. We then spun down the culture and resuspended it in galactose media (OD ~ 0.1), we then collected yeast at 30 minutes, 1 hour, 2 hours, 4 hours, and 5.5 hours. These samples were placed in the -80 freezer for further processing. We extracted RNA from each time point using the Quick-RNA Fungal/Bacterial Kit from Zymo research. We constructed RNA-sequencing libraries using the KAPA mRNA hyperprep kits. These libraries were then sequenced on an Illumina MiSeq.

To quantify allele-specific expression (ASE) in our time course experiment, we used the WASP software to generate allele counts for each SNP site within every gene (v0.2.2)(*31*). We quantified the significance of ASE using a binomial test. For the galactose genes (*GAL1/10/7* and *GAL2*), we could not obtain ASE estimates because the reads from the CBS2888 alleles do not align to the reference. For these genes we quantified expression using Kallisto (v.0.44.0) with a reference transcriptome that contained the coding sequences of the CBS2888 galactose genes (*32*). The fold-change was calculated as the $\log_2$ ratio of the estimated counts of the reference and CBS2888 galactose genes.

<u>Curation of sequencing data used for the population genetic and phylogenetic analysis of the galactose alleles of *S. cerevisiae*</u>

We obtained the genome assemblies from two large collections of sequenced yeast isolates, comprising 1,277 total isolates (*11*, *12*). One of these isolates had poor sequencing coverage (YCL), and we removed it from all downstream analyses. We also obtained the genome assemblies and gene annotations for the *Saccharomyces* genes: *S. mikatae (S. mik)*, *S. uvarum (S. uva),* and two outgroup species *Kazaschstania africana* (*K. afr*) and *Kazaschstania naganishii* (*K. nag*) from the yeast gene order browser (YGOB)(*33*). We obtained the genome assembly and gene annotations for *S. paradoxus* (*S. par,* CBS432) from the Yeast Population Reference panel (*34*). We obtained the genome assembly and gene annotations for *S. eubayanus* (*S. eub,* FM1318) from Ensembl Fungi (*35*, *36*). We used the moleculo long-read assembly of CBS2888 provided by Bloom et al. as a representative strain for the alternative alleles, and the reference genome (SacCer3) as a representative strain for the reference alleles (*37*). We obtained all 332 publicly available budding yeast genomes, this list of strains and the location of the data on public databases was provided in Shen et al (*38*).

<u>Annotation of the galactose alleles from a global collection of 1,276 sequenced yeast strains</u>

We used BLAST (v2.6.0) to align the alternative (CBS2888) and reference *GAL1*, *GAL10*, *GAL7*, *GAL2*, and *PGM1* promoter alleles to each of the 1,276 genome assemblies (*39*). We removed any alignments that did not cover greater than 80% of the length of the query sequence and retained the best alignment. We classified the galactose genes in each strain as alternative if the alignment had greater than 90% identity to the alternative allele and less than 90% to the reference allele. Equivalent criteria were used to classify strains as having the reference allele. The reference PGM1 promoter contained a Ty1 transposon, which fragmented most of the assemblies with the reference allele. For this gene, we only required that the alignment covered greater than 40% of the query sequence. Even with this relaxed criterion for some genes in some strains we were still not able to make a classification, usually due to additional assembly fragmentation. For 70 strains we were not able to assign the *PGM1* promoter, for 62 strains we were not able to assign *GAL2*, and for one strain we were not able to assign *GAL7*. In these cases, we determined which allele was present by manually combining multiple partial alignments. During this process, we identified 8 strains collected in China that had neither a reference or alternative *GAL2* or *PGM1* promoter allele. On closer inspection, all of these strains had distinct alleles (<90% sequence identity to both the reference and alternative alleles) at both of these loci.

Analysis of the linkage disequilibrium between galactose alleles

To analyze the patterns of linkage disequilibrium (LD) between the galactose alleles, we could not use a standard measure of LD, R^2, because it cannot be calculated between more than two loci. Instead, we used an entropy-based method (eLD) which generates a LD index (ϵ), which can be applied to arbitrary numbers of alleles (*13*). We calculated ϵ using the genotypes of the galactose alleles that we inferred for the 1,276 strains. Strains with the Chinese alleles were conservatively assigned as the reference. The yeast population is highly structured, and as such the null expectation for ϵ will be inflated relative to an unstructured population. We calculated a null distribution by randomly selecting 10,000 triplets of SNPs from different chromosomes and calculating ϵ.

For this analysis, we used SNPs with frequencies of between 4-5%, which is close to the observed frequency of the alternative alleles (4.4%). These SNPs were obtained by merging a VCF file provided by Peter et al. (*12*), and a VCF file we generated from data provided by Duan et al. (*11*). For the data from Peter et al. we removed sites with more than 5% missing data, indels, and sites that were not biallelic. The study by Duan et al. contained an additional 266 strains but did not provide a VCF file. We therefore downloaded the reads from the short-read archive (SRA) and generated a VCF file using the standard Genome Analysis Toolkit (GATK, v4.1.3.0) variant calling pipeline (*40*). We removed sites that had overall coverage less than 30, QUAL scores less than 100, and more than 5% missing data. We also removed sites that were indels and were not biallelic using vcftools (v0.1.15). We merged these VCF files together and set any missing sites to the reference using the "—missing-to-ref" option in bcftools (v1.9)(*41*). This merged VCF file contained a total of 1,803,186 SNPs and had 16,451 SNPs with a frequency of between 4-5%.

Phylogenetic analysis of the alternative and reference galactose alleles

To visualize the phylogenetic relationships between all 1,276 strains, we created a neighbor joining tree following similar methods as described by Peter et al. (*12*). In more detail, we extracted all variants from a merged VCF file generated using bcftools with default parameters. We filtered this merged VCF to remove sites with greater than 5% missing data and sites with a minor allele frequency less than 5%. Next we generated a dissimilarity matrix using SNPRelate (v1.18.1)(*42*). This matrix was used to build a neighbor joining tree with ape (v5.3)(*43*). We visualized this tree using the ggtree (v1.16.4) package (*44*). We colored the branches based on which combination of the galactose alleles each strain had.

To visualize the phylogeny of the alternative galactose alleles, we performed phylogenetic clustering of the *GAL1, GAL10, GAL7* coding sequences from CBS2888 (alternative), SacCer3 (reference), *S. par*, *S. mik*, *S. uva*, *K. afr*, and *K. nag* (*45*, *46*). We did not use *GAL2* in this analysis because this gene is only found in the *Saccharomyces* genus. We aligned each gene using Clustal Omega (v1.2.4) and concatenated the sequences. We used RAxML (v 8.1.21) with a GTR + GAMMA model for tree generation (*47*). We set the outgroups to be *K. afr* and *K. nag*. We performed 100 rapid bootstraps to assess the confidence of the branches.

To identify divergent regions of the duplicated *GAL2* genes, we calculated the pairwise sequence identity of *GAL2a* and *GAL2b* for the alternative, Chinese, *S. uva*, and *S. eubayanus* (*S. eub*) sequences. This pairwise sequence identity was calculated in 50 base-pair windows with a 10 base-pair step. To investigate the phylogenetic relationships between the alternative, Chinese, and reference *GAL2* genes, we performed phylogenetic clustering of the *GAL2* coding sequence from CBS2888 (alternative), BAM (Chinese), SacCer3 (reference), *S. par, S. mik*, *S. uva,* and *S. eub*. *GAL2* is only found in the *Saccharomyces* genus*,* so we used the K. afr and K. nag *HXT7* genes as the outgroups. We aligned the protein sequences of all of these genes using DECIPHER. We annotated the cytosolic, extracellular, and trans-membrane domains of the S. cer *GAL2* gene using TMHMM (*48*). We performed phylogenetic sub-clustering analysis using the codon-aligned protein and DNA sequences of the N-terminal domain comprising the first 67 amino acids of the reference sequence. We also performed phylogenetic sub-clustering analysis of the trans-membrane and C-terminal cytosolic domain using the codon-aligned protein and DNA sequences comprising amino acids 68 through 575 of the reference sequences.

Population genetic analysis of the alternative and reference galactose alleles

We estimated the synonymous substitutions per site (dS) between the reference and alternative galactose genes using the codonseq package of Biopython (v1.70) with the NG86 method (*49*, *50*). We note that this method utilizes the Jukes-Cantor correction that explicitly incorporates back mutations and converts the raw proportion of observed synonymous changes into a time linear distance between the sequences. If there are many differences between the sequences this can lead to estimates of dS that are greater than 1. To calculate a 95% confidence interval on all our dS estimates, we performed 200 bootstraps in which we resampled codons with replacement and then recalculated the dS. When the number of synonymous substitutions is too large the NG86 method can return an error, for these bootstrap samples dS was set to 3.23. To obtain a background distribution of synonymous differences, we performed a global alignment using ssearch36 (v36.3.8f) to identify the reference genome open-reading frames in the CBS2888 genome (*51*). We only considered alignments that contained a complete open reading frame. We calculated the dS for each aligned gene using codonseq. We aligned the alternative and reference *PGM1* promoters and calculated the distance between the sequences using the DistanceCalculator from Biopython with the ‘identity’ method. For every pair of neighboring genes in the reference genome, if such pairs both align to the same CBS2888 contig we extracted the sequences between them for both the reference and CBS2888 assemblies and calculated percent identity using the method described above for the *PGM1* promoters.

Analysis of synonymous substitutions per site in CBS2888

Using the aligned genes between CBS2888 and the reference genome, we calculated the dS in 200 amino acid overlapping windows genome-wide using codonseq. We used a step of 10 amino-acids for the windows. We repeated our analysis of loci linked to the alternative galactose alleles using non-overlapping windows of 50, 75, 100, 150 and 200 base-pairs.

Analysis of sequence dissimilarity in *EMP46*

The CBS2888 *EMP46* contains a premature stop codon, in addition to many insertions and deletion relative to the reference gene. We calculated sequence identity of the alternative EMP46 allele relative to the reference using Biopython with the ‘identity’ method. We performed this calculation in 300 base-pair windows with a 100 base-pair step. We calculated these sequence identities using insertions and deletions as both missing and mismatches.

Population genetic simulations

Forward simulations were performed with SLiM (v3.4)(*52*).

Parameter choices

The population genetics of yeast has not been extensively characterized and several important evolutionary parameters remain to be estimated. However, we were able to obtain a number of realistic fixed parameters from the literature (Table S11). For all of our simulations, we used an effective population size (Ne) of 10,000,000 (*12*), a mutation rate (μ) of 3.8e-10 (*53*), and a meiotic recombination rate (r) of 3.133483e-06 (*6*). We set the total number of generations in all our simulations to 3.2 billion, which is our estimate for how many generations have passed since the alternative and galactose alleles diverged. These parameters were fixed in every simulation regardless of the model. Yeast are extremely inbred and often divide clonally, and when undergoing meiosis mostly self. To model this behavior, we sampled the cloning rate (c) from a uniform distribution from between 0.9 and 0.9995. Yeast outcross approximately once every 50,000 generations (*54*), so we set the selfing rate (s) to be whatever was necessary to ensure that the outcrossing events happened at this rate. We implemented selfing indirectly by setting the meiotic recombination rate to r_{adj} =r(1-s) (*55*). SliM does not implement mitotic recombination, which leads to loss of heterozygosity events when yeast divide clonally. To model these loss of heterozygosity events in SLiM, we set a portion of the clonal evolution to be performed by the selfing module in SLiM. The mitotic recombination rate was drawn from a uniform distribution from between 0 and 0.1. All simulations were run for a neutral burn-in of 8Ne with the same population set-up used for the rest of the simulation and a migration rate between all populations of 50%.

Simulated models

1) Ancient balancing selection

We modeled the multi-locus ancient balancing selection as a two-population model where the populations were assigned to either a galactose-rich or a glucose-rich environment with migration occurring between the populations. The two subpopulations were created with sizes of Ne/2. We created six point mutations that are used to represent the reference and alternative galactose alleles of *GAL1/10/7, PGM1* promoter, and *GAL2*. After the burn-in, we placed these reference mutations at positions 5kb, 15kb, and 25kb into all the genomes from the population in the glucose-rich environment, and vice-versa for the alternative mutations into the galactose-rich environment.

We used fitness callbacks to implement the multi-locus balancing selection so that in the galactose-rich environment individuals with all three alleles of the alternative pathway had a selective advantage, and in the glucose-rich environment all three alleles of the reference pathway had a selective advantage. For the alternative pathway the dominance was set to 0.98, and for the reference pathway we set the dominance to 0.0001. These dominance coefficients were chosen based on our experimental results, where the cost of the alternative pathway in the reference environment was due to the constitutive expression, and the benefit of the alternative pathway in the alternative environment was partly due to a switching advantage, which is a dominant trait. The unknown parameters in this simulation were the relative advantage of the reference and alternative pathways in their environments. These selection coefficients were drawn from a uniform distribution between 0.0002 and 0.3. We found that when using SLiMs default settings, different selection coefficients and migration rates between the environments influenced the observed outcrossing rate between individuals from the different environments. To prevent this, we fixed the migration rate between the environments to 50% and chose the first parent based on fitness and the second parent randomly from within each population. This approach gave realistic outcrossing rates while not affecting the other dynamics of the simulation.

2) Recent neutral introgression

We modelled introgression as a three-population model with two populations of size Ne/2 assigned to either a galactose-rich or a glucose-rich environment with migration occurring between the populations. In this simulation, we included a third population of size Ne completely isolated from the other populations. Introgressions events were modelled as a single generation pulse from this third population into the galactose-rich environment at a specific frequency. After the burn-in, the reference galactose alleles were fixed in both the galactose-rich and glucose-rich environments, and the alternative alleles were fixed in the isolated population. If introgression occurs at a frequency of ~0.5, then, in the absence of balancing selection, fixation or loss is expected to occur on the order of 2.77Ne generations (*56*). Therefore, we only simulated neutral introgression occurring between 1 and 50,000,000 generations ago. For each simulation, we drew the generation from a uniform distribution. The pulse fraction was drawn from a uniform distribution of between 0.5 and 1.

3) Introgression followed by the maintenance of the two pathways by balancing selection.

We modelled introgression followed by balancing selection using the same setup for the populations as the recent neutral introgression model until the generation the alternative galactose alleles were introgressed. After this generation, we switch to using the same set-up as the multi-locus ancient balancing selection model. We simulated three versions of this model to capture different ranges of generations ago for when the introgression could have occurred, these were: 1-25 million; 25-50 million, and 50 million-1 billion. For each of these models, the number of generations in each simulation was drawn from a uniform distribution.

Computational speed-up

Forward simulations using the estimated population sizes are intractable with existing computational resources and software. To proceed, we rescaled the parameters of our model by dividing the population sizes by a factor and increased the mutation, recombination rate, and selection coefficients by the same factor. To check whether this scaling affected the dynamics of our simulations, we fixed a set of parameters and simulated from our ancient balancing selection model with different factors (1,000, 5000, 10000, 50000, and 100,000) and investigated whether the site frequency spectra of the resulting simulations differed between scaling factors (Fig. S22).

Summary statistic computation

We calculated the two summary statistics from the simulated data that support ancient balancing selection. These were the entropy linkage disequilibrium index (eLD) and a vector representing the synonymous rate of substitution (dS) at each base pair. For our observed data, we used the dS between the CBS2888 and S288C (SacCer3) galactose alleles shown in Figure 3. To obtain an equivalent statistic from the observed data, we calculated the pairwise diversity between haplotypes with the reference and galactose alleles in 600 base-pair windows with a 30 base-pair step. This window size and step is identical to the 200-amino acid windows and 10-amino acid windows used to calculate the dS from the observed data.

To enable comparison of the observed and simulated dS signals, we removed the *GAL1/10/7*, *PGM1* promoter, and *GAL2* regions from the observed data. This approximates our simulations where we modelled these selected regions as point mutations. We also had to convert the dS signal to the coordinate system of the simulations. To achieve this, we first assigned each locus to one of three 10kb chunks on a 30kb chromosome, making sure to center the observed data in each chunk around the selected locus. Secondly, the windows in the observed data are spaced according to the genes nearby the galactose alleles. We aligned each window in the observed data to the window in the simulated data with which it had the largest overlap.

Approximate Bayesian Computation

We performed Approximate Bayesian Computation (ABC) with the R package abc (v2.1)(*57*) and the rejection method to construct the posterior distribution of the random parameters and the simulated and observed summary statistics described above. We used the postpr function to calculate Bayes factors for the five evolutionary scenarios we simulated: ancient balancing selection, neutral introgression, and introgression followed by balancing selection in three time periods. We calculated these Bayes factors for a range of tolerances (0.01, 0.05, 0.1,0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50). The approximations used to generate the Bayes Factors hold when the different models are all *a priori* equally likely, and the same number of simulations are from each model are used in the comparison (*58*). We calculated Bayes factors using 20,000 simulations from each of the five models.

Supplementary Text

Global distribution of the galactose alleles

We identified three most common combinations of galactose alleles: only reference alleles (1213 strains), only alternative alleles (49 strains), and 8 strains from China with the alternative *GAL1/10/7* allele and alleles of *GAL2* and the *PGM1* promoter that differ from both the reference and the alternative alleles (Table S9). In addition, one strain had a deletion of the entire region containing *GAL1/10/7*, three strains contained reference alleles at all loci except *GAL7*, and two strains were heterozygous at the *PGM1* promoter and homozygous for the reference alleles at the other loci, indicating that strains with alternative and reference galactose alleles have not been completely reproductively isolated from each other. This is also demonstrated by the phylogenetic distribution of the isolates with alternative *GAL* alleles (Figs. 2A and S8).

Dating the split between the alternative and reference galactose alleles

We combined the estimated number of synonymous substitutions per site (dS) with the measured mutation rate μ=3.8x10^{-10} per site per generation in yeast (*53*). Under neutral theory, the dS value of 2.4 for the galactose alleles corresponds to a split between these alleles approximately 3.2 billion generations ago (95% C.I.=2.5-4.5 billion generations), which predates the most recent common ancestor of the *Saccharomyces* genus (*15*). This date is over 100 times older than the divergence between CBS2888 and the reference strain based on the average number of synonymous substitutions per site (0.014) for all genes (Fig. S10, Table S10)

We calculated the synonymous substitutions per site between the reference genome and the *GAL1*, *GAL10*, and *GAL7* genes of the 1,276 sequenced strains (Fig. S11). In agreement with our comparison between CBS2888 and reference galactose alleles, the number of synonymous substitutions per site was high in strains classified as having the alternative alleles (*GAL7* dS=2.50, range=0.94-2.88; *GAL10* dS=2.64, range=2.53-2.85; *GAL1* dS=2.13, range=2.05-2.52). The number of synonymous substitutions per site was low when comparing the alleles of strains that we annotated as having the reference alleles (*GAL7* dS=0.01, range=0-0.26, *GAL10* dS=0.01, range=0-0.016, *GAL1*=0.01, range=0-0.041).

The alternative and Chinese galactose alleles fall outside the *Saccharomyces* genus

We extracted *GAL1, GAL10,* and *GAL7* alleles from the genomes of 1,234 of the 1,276 strains with complete assembled ORFs at these sites. We removed the 3 strains with reference alleles at all loci except *GAL7*. We performed phylogenetic clustering of the genes from the remaining 1,231 strains and verified that all of the alternative *GAL1/10/7* alleles fall outside the *Saccharomyces* genus (Fig. S12).

The Chinese and alternative *GAL2* alleles both contain a *GAL2* duplication. We compared the centromere-proximal alternative and Chinese *GAL2* alleles and found that these two alleles are more similar to each other (dS=1.37) than they are to the reference (Chinese *GAL2* compared to reference dS=1.91, alternative GAL2 compared to reference dS=2.38). Phylogenetic clustering revealed that the transmembrane and C-terminal regions of the Chinese and alternative *GAL2* genes cluster with each other (Fig. S19C). This result provides evidence that the Chinese, alternative, and reference galactose alleles have an ancient origin in *Saccharomyces*.

The Chinese and alternative *PGM1* promoter alleles are more similar to each other (70.6% sequence identity) than either is to the reference (Chinese *PGM1* promoter compared to reference, sequence identity=47.0%; alternative *PGM1* promoter compared to reference, sequence identity=46.7%). The low sequence conservation between the *PGM1* promoters in outgroup species prevented us from performing phylogenetic clustering, however, we did note that the alternative and Chinese *PGM1* promoter alleles are missing a lysine tRNA that is present in every species of the *Saccharomyces* genus (Fig. S13). This places the Chinese and alternative *PGM1* promoter alleles outside the *Saccharomyces* genus

A signature of ancient balancing selection

We examined the rate of synonymous substitutions per site (dS) across the CBS2888 genome relative to the reference and observed a strong signature of ancient balancing selection at all three galactose loci (Figs. 3 and S15). A signature that persists when using non-overlapping windows for calculating dS (Fig. S14). No comparable signatures were seen at other genomic loci in the CBS2888 genome (Fig. S16). The diversity of alternative alleles of the strains that carry them falls within the distribution of diversity among other regions of the genome, providing evidence that the alternative galactose pathway was not recently introgressed (Fig. S17).

In agreement with our analysis of the CBS2888 genome, we observed a strong signature of balancing selection at the galactose genes of the 1,276 sequenced strains classified as having the alternative or Chinese galactose alleles (Fig. S18). We did not observe this signature when comparing strains classified as having the reference galactose alleles.

Searching for the alternative galactose alleles in other budding yeast species

We searched for the alternative galactose alleles in the genome sequences of 332 budding yeast species (*38*). We did not find any matches of greater than 90% identity to the alternative of Chinese galactose alleles, indicating that these alleles were not recently introgressed from any of these species.

Forward genetic simulations

To distinguish between evolutionary scenarios that could have given rise to the signatures found in the observed data, we performed forward simulations under five possible models using SLiM (*52*) and used Approximate Bayesian Computation (ABC) to fit parameters and perform model selection (*57*). The models we simulated were: 1) ancient multi-locus balancing selection that has been acting since the galactose alleles diverged; 2) a recent neutral introgression of the alternative galactose alleles occurring between 50 million and 1 generation ago; and 3) a introgression of the alternative galactose alleles followed by maintenance of the alleles by multi-locus balancing selection in three different time periods (M1, M2, and M3)(Fig. S21). The large effective population size (Ne=10^6) and extremely large number of generations ($3.2x10^9$) made these models computationally intractable. To overcome this challenge, we rescaled our simulations by a factor of 50,000, which although large, preserved the behavior of our simulations (Fig. S22). The two summary statistics we calculated were entropy linkage disequilibrium index (eLD) and a vector representing the synonymous rate of substitution (dS). We converted the dS signal from the observed data to the coordinate system of the simulations (Fig. S23).

We first investigated whether the simple model of neutral introgression occurring between 0 and 50 million generations ago was a better fit to the observed data than a model of ancient balancing selection. We performed 24,679 simulations of the neutral introgression model and found that in only 5,180 (21%) did all six alleles survive until the end of the simulation. This is expected, as without balancing selection these alleles will be lost or become fixed over time. The timing of the introgression was more recent in simulations where all six alleles were retained, with the introgression occurring on average ~7 million generations ago (95% CI=250,000-23 million)(Fig. S24). We compared 20,000 neutral introgression simulations to 20,000 ancient balancing selection simulations, and found decisive evidence that the neutral introgression model was a worse model (Bayes Factor=298.6 for the ancient balancing selection model compared to the neutral introgression model at a tolerance of 0.05) (Fig. S25A). To understand more about the resulting summary statistics from the neutral introgression model, we examined the two accepted simulations with the lowest distance to the observed data (i.e. the two best simulations). We found that these simulations did not display a clear signature of elevated sequence divergence at linked neutral sites (Fig. S25B) and had eLD values (ϵ) of 0.24 and 0.44, which are much lower than our estimate for the observed data (ϵ=0.59). These results provide strong evidence that recent introgression without selection is not the evolutionary process that gave rise to the reference and alternative galactose alleles.

Another evolutionary scenario that could explain our observations is that the alternative galactose alleles were introgressed at some point in the past and have been maintained by multi-locus balancing selection ever since. To investigate this scenario, we performed simulations from three models representing ranges of possible dates when the introgression could have occurred. While pinning down the exact timing of such an introgression is challenging in this case because of uncertainty around many important parameter values, we wanted to see if there was support for a recent (M1), moderately old (M2), or ancient (M3) introgression followed by balancing selection. We specified the time periods in terms of generations as follows: 1 and 25,000,000 (M1); 25 million and 50 million (M2); and 50 million and 1 billion generations ago (M3)(Fig. S21). We compared these three models (n=20,000 per model) to the ancient balancing selection model (n=20,000) and found evidence that the M1 and M2 models were not well supported across a range of tolerances (Fig. S26A). We compared the ancient introgression model with balancing selection to the model of ancient balancing selection (M3; n=20,000) and found that neither model was favored. For all of these models, the two simulations with the lowest distance match the observed data well (Fig. S26B). The analysis further showed that if the alternative alleles were introgressed, the best estimate for the time of introgression is ~108 million generations ago (Fig. S27). These results suggest that we can reject the recent and moderately old introgression models in favor of the ancient introgression with balancing selection or ancient balancing selection model. Overall, these results reinforce our hypothesis that ancient balancing selection has maintained the alternative and reference galactose alleles.

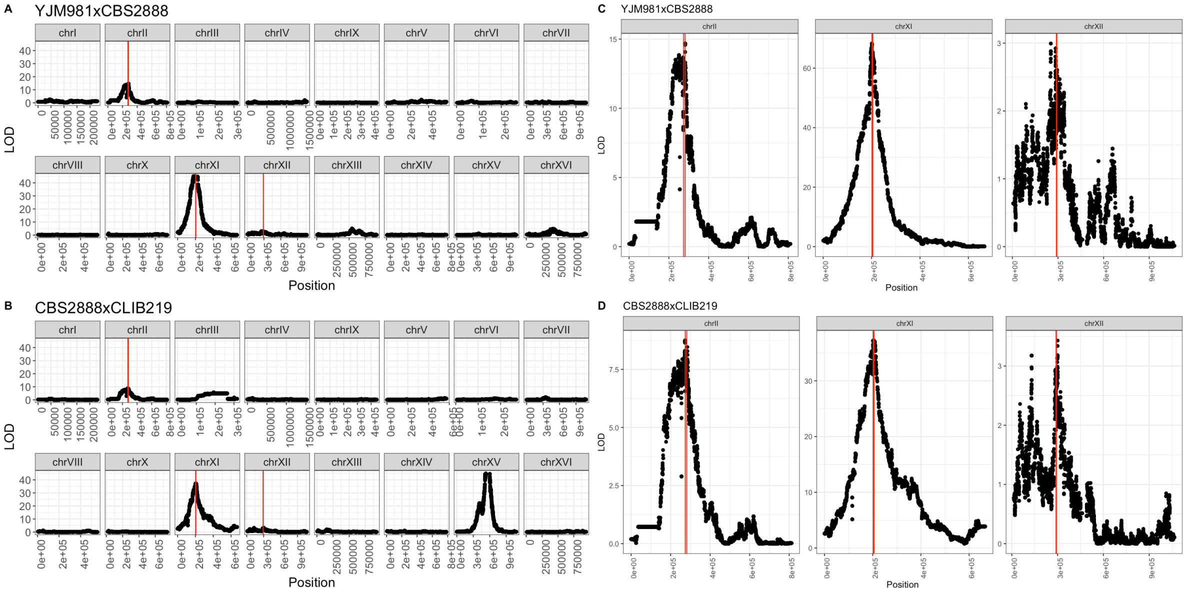


Fig. S1.
Additive QTL for the three galactose regions in two independent crosses. A) logarithm of the odds (LOD) traces for 943 segregants derived from a cross between CBS2888 and YJM981. B) LOD traces for 867 segregants derived from a cross between CBS2888 and YJM981. The y-axis was truncated at 45. Note the large-effect QTL on chromosome XV coming from the CBS2888xCLIB219 cross. CLIB219 was found to have a *de novo* mutation in ADE2 on this chromosome that affects all traits. C) LOD traces for chromosome II, XI, and XII of segregants from a cross between CBS2888 and YJM981. D) LOD traces for chromosome II, XI, and XII of segregants derived from a cross between CBS2888 and CLIB219. On all panels the red vertical lines show the regional boundaries of the galactose genes. On chromosome II, the red bars highlight the *GAL1/10/7* region. On chromosome XI, the red bars highlight the *PGM1* promoter. On chromosome XII, the red bars highlight the *GAL2* region.

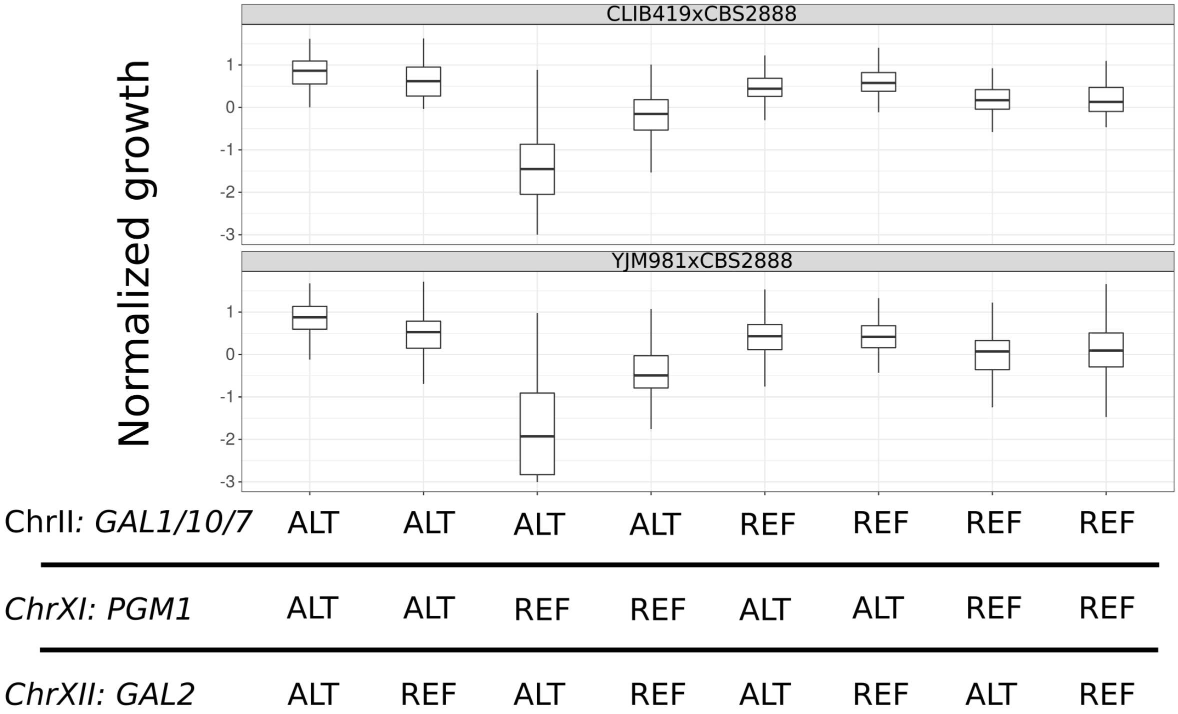


Fig. S2.
Replication of the three-way genetic interaction another panel of segregants. top) Boxplots showing the growth of 943 segregants on 2% galactose agar plates. These segregants were derived from a cross between CBS2888 and CLIB219. bottom) Boxplots showing the growth (s.d units) of 867 segregants on 2% galactose agar plates. These segregants were derived from a cross between CBS2888 and YJM981. Segregants are partitioned on the x-axis based on their genotypes at eight combinations of the three galactose loci (ChrII: *GAL1/10/7*, ChrXI*: PGM1*, and ChrXII: *GAL2*). Alleles at the QTL loci from CBS2888 are designated as ALT, and alleles from CLIB219 and YJM981 are designated as REF.

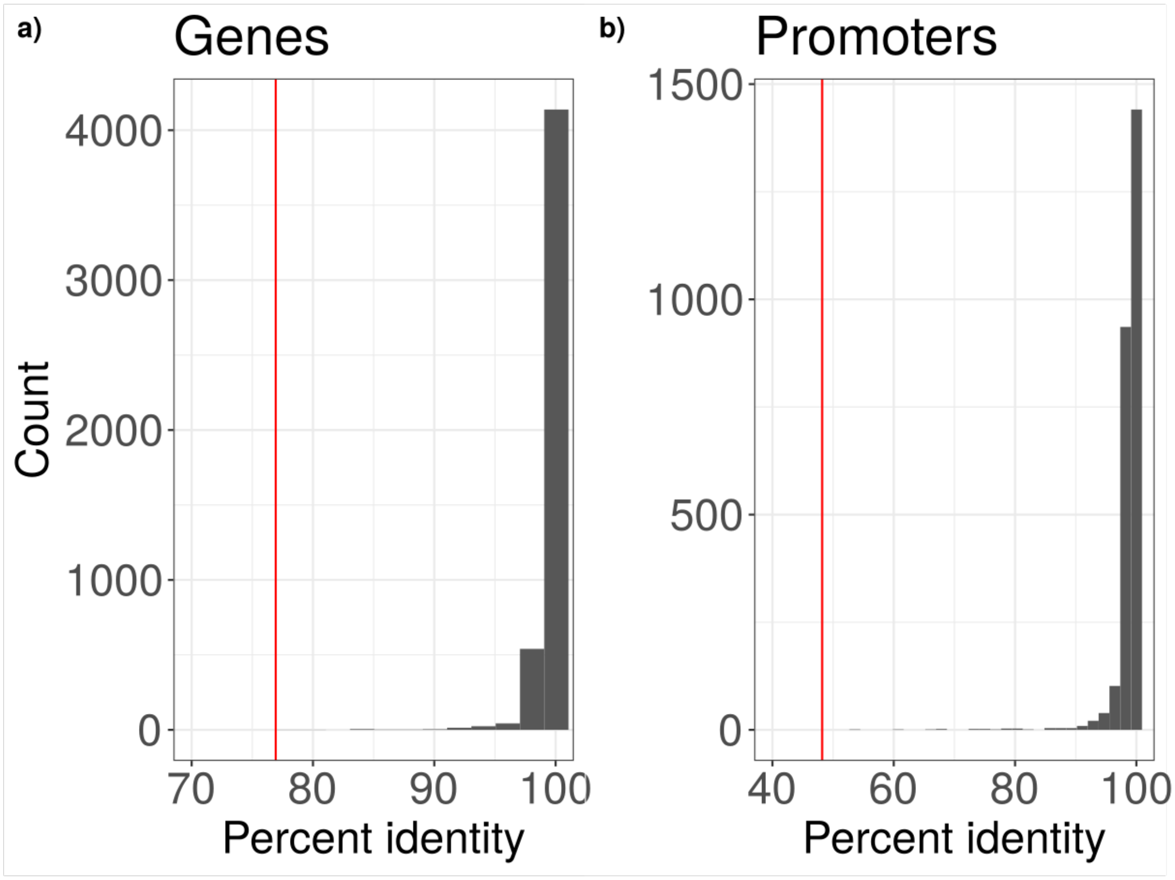


Fig. S3.
Sequence identity between the genes and promoters of the reference and CBS2888.
a) Histogram of the percent identity of 4,780 genes in the CBS2888 genome when these genes were aligned to the reference genes. The average sequence identity of the diverged *GAL1*, *GAL10*, *GAL7*, and *GAL2* (77%) when aligned to the reference galactose alleles is shown in red.
b) Histogram of the percent identity of 2,583 promoters of CBS2888 when aligned to the reference. The sequence identity of the diverged PGM1 promoter (48%) when aligned to the reference promoter is shown in red.

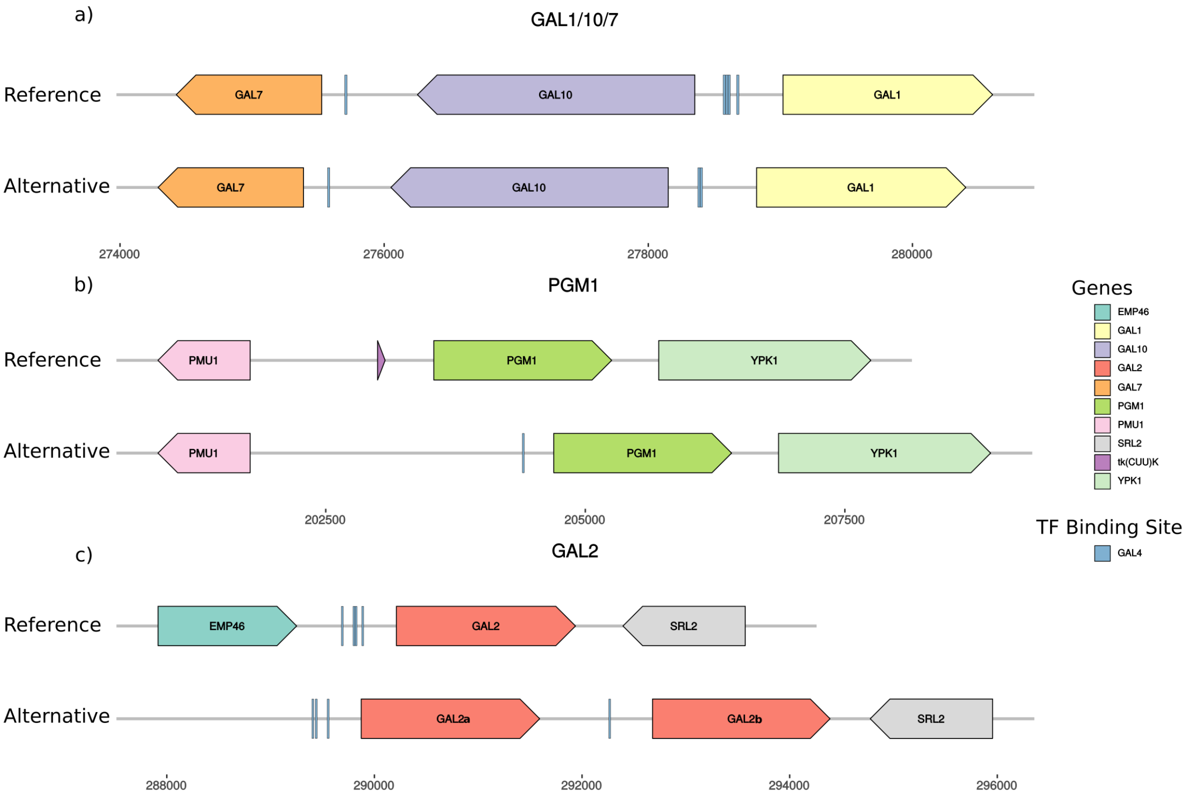


Fig. S4.
Genomic organization of the alternative and reference galactose alleles. Regional genome plots showing the relative lengths and locations of the *GAL* loci and surrounding regions. We annotated any *GAL4* upstream activating sequence (UAS) that we found at each of the loci. a) *GAL1/10/7*, b) *PGM1*, c) *GAL2*. The centromere-proximal copy of *GAL2* in CBS2888 (alternative) is denoted as *GAL2a*, and the other copy as *GAL2b*.

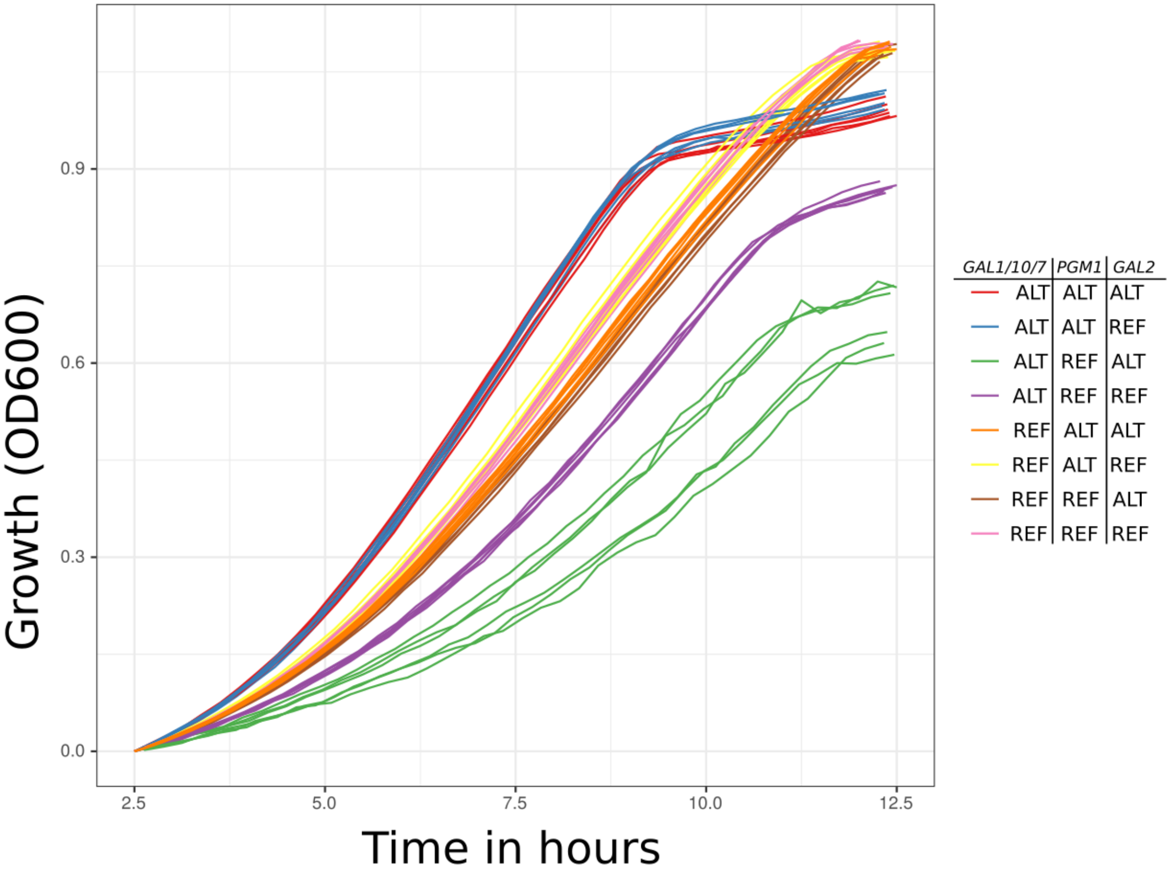


Fig. S5.

Growth curves for allele replacement strains that contain all eight combinations of the alternative and reference galactose alleles. Growth curves for 6 replicates of each allele replacement strain grown in 2% glucose. Each growth curve is colored according to the genotype of the corresponding strain.

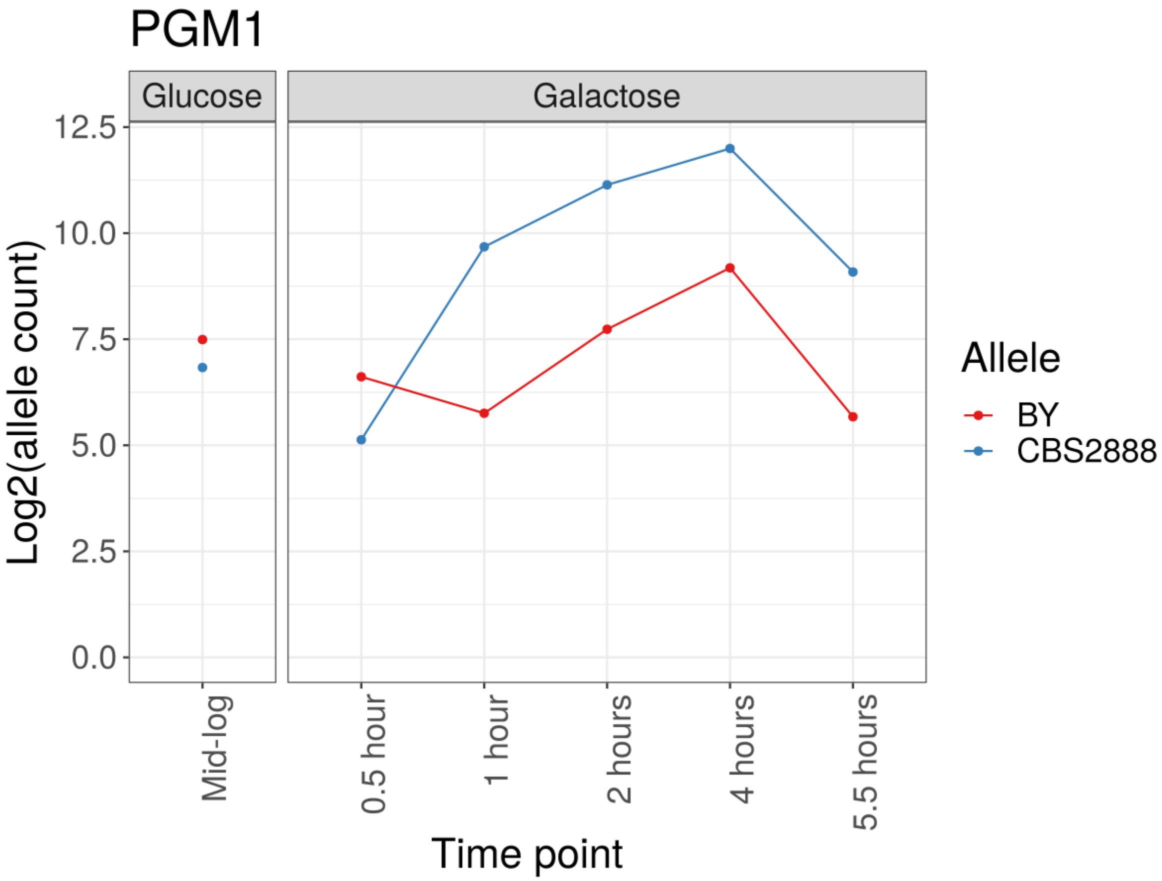
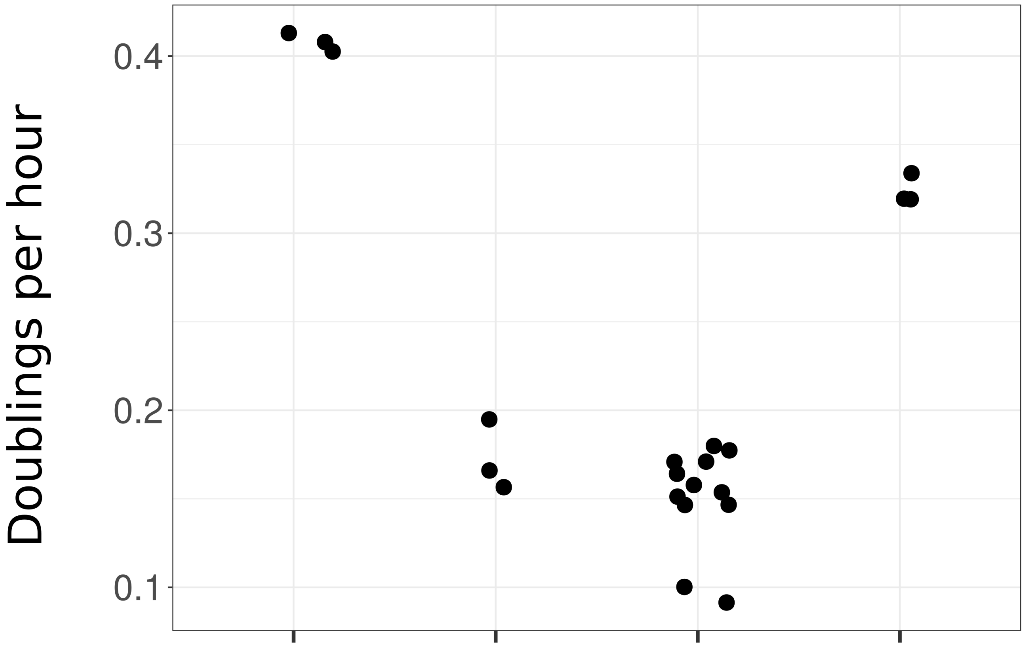


Fig. S6.
Allele-specific expression of PGM1 throughout a galactose induction time course. Allele-specific expression of a hybrid (CBS2888xBY) that is heterozygous for the alternative and reference alleles when grown in 2% glucose medium and transferred to 2% galactose medium. The line graph shows the $\log_2$ allele counts for the CBS2888 (alternative) alleles and BY (reference) alleles.



ChrII:GAL1/10/7	ALT	ALT	ALT	REF
ChrXI: PGM1	ALT	REF	ALT	REF
GAL4-UAS	WT	N.A.	ΔGAL4-UAS	N.A.
ChrXII: GAL2	ALT	ALT	ALT	REF

Fig. S7.

Galactose responsive *PGM1* is necessary for the proper functioning of the alternative galactose pathway. Growth rates of allele replacement strains when grown in 2% galactose medium. The strains are from left to right: A strain with all three alternative alleles. A strain with the reference promoter and the alternative *GAL1/10/7* and *GAL2*. A strain with all three alternative alleles and a non-functional GAL4-UAS (CGGN$_{11}$CCG-> CGGN$_{11}$ACG). A strain with all three reference alleles. Each dot shows a biological replicate.

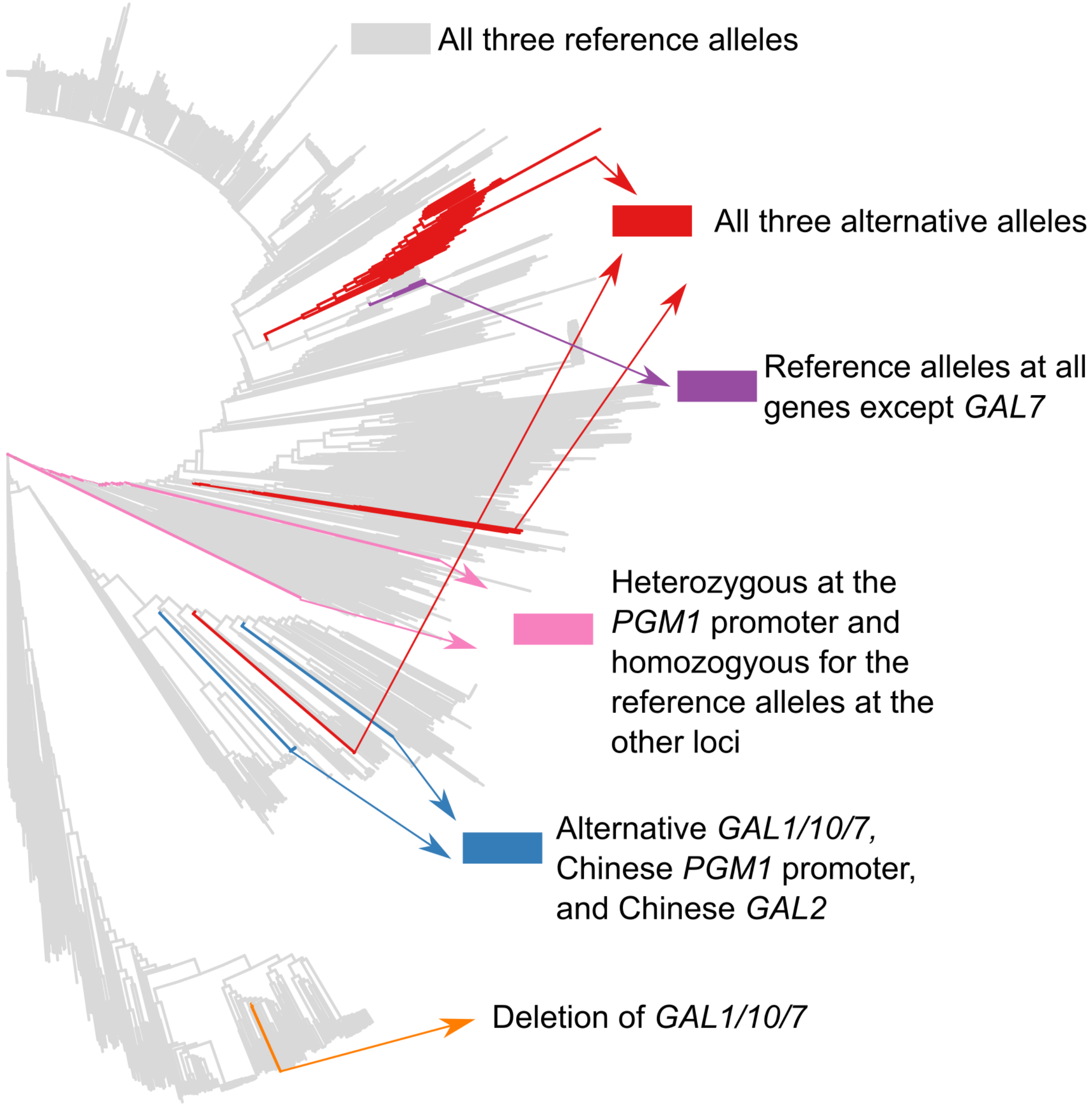

All three reference alleles
All three alternative alleles
Reference alleles at all genes except *GAL7*
Heterozygous at the *PGM1* promoter and homozogyous for the reference alleles at the other loci
Alternative *GAL1/10/7*, Chinese *PGM1* promoter, and Chinese *GAL2*
Deletion of *GAL1/10/7*

Fig. S8.
Genome-wide neighbor-joining tree of 1,276 sequenced yeast isolates. Clusters of branches with the same genotypes at the three galactose loci are colored as follows: all three reference alleles (grey), all three alternative alleles (red), alternative *GAL1/10/7* allele and Chinese *PGM1* and *GAL2* alleles (blue), reference alleles of all genes except *GAL7* (purple), heterozygous at the *PGM1* promoter and homozygous for the reference alleles at the other loci (pink), and a deletion of the entire region containing *GAL1/10/7* and reference alleles at the other loci (orange).

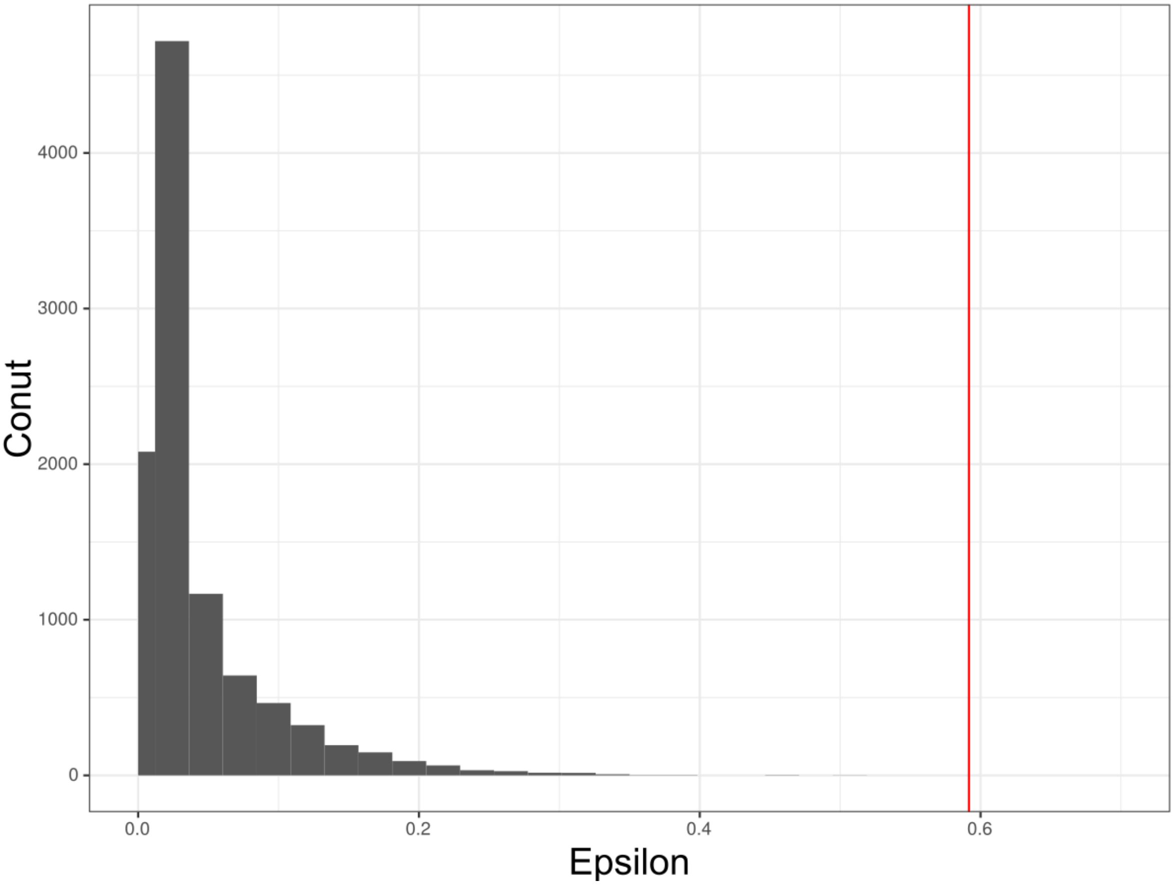


Fig. S9.
Linkage disequilibrium between 10,000 random sets of three SNPs.
Distribution of the entropy-based linkage disequilibrium index (ϵ) of 10,000 random sets of three SNPs with a similar frequency (4-5%) to the diverged galactose alleles. In red is the value of ϵ of the galactose alleles (0.592). The largest value we observed in our 10,000 permutations was 0.50. A larger ϵ indicates that the sites are in high linkage disequilibrium.

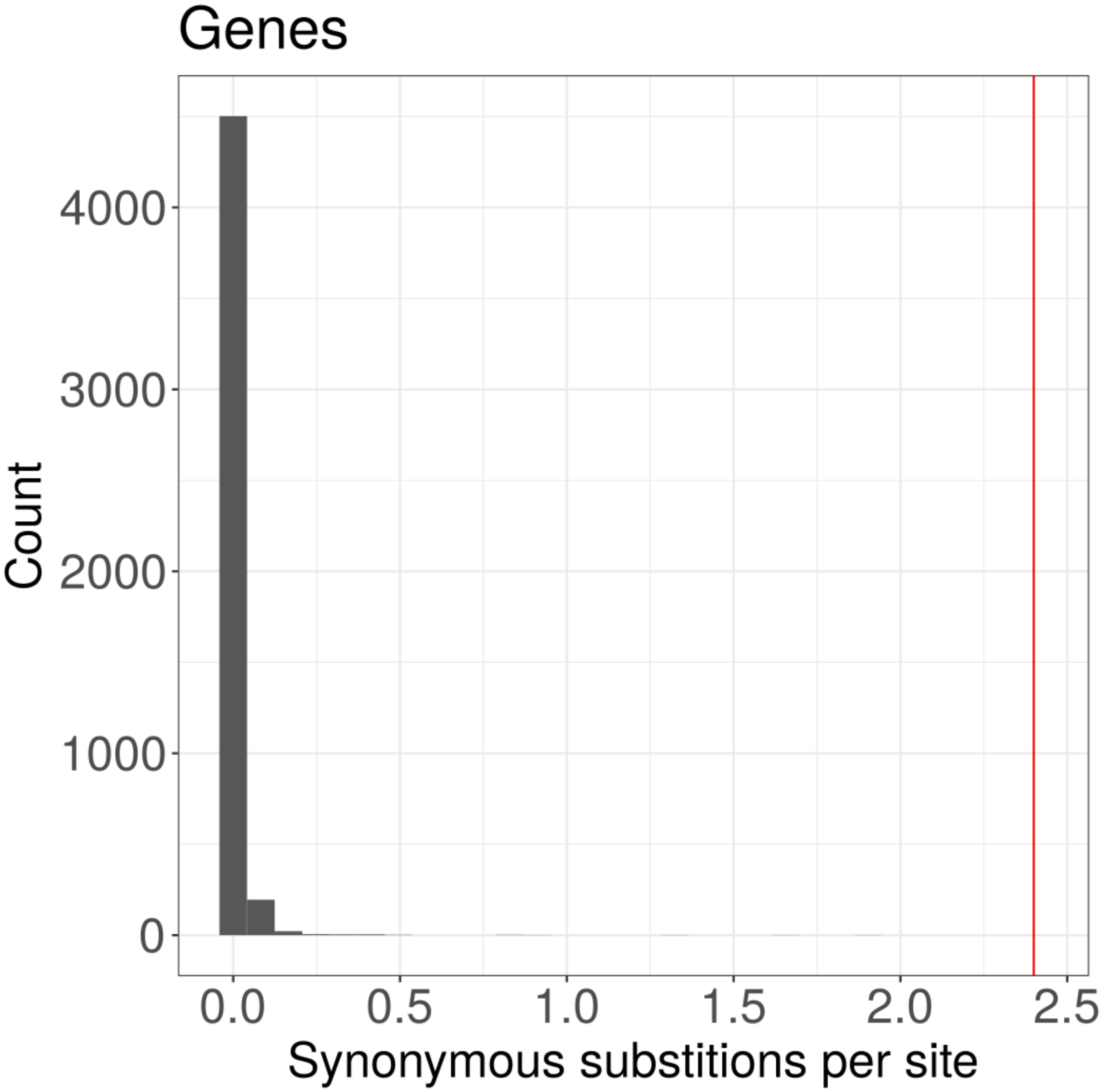


Fig. S10.
Synonymous substitutions per site (dS) when comparing the reference and CBS2888 genes. Histogram of the dS of the CBS2888 genes when compared to the reference genes. In red is shown the average dS of the alternative *GAL1*, *GAL7*, *GAL10*, and *GAL2* galactose alleles when compared to the reference genes.

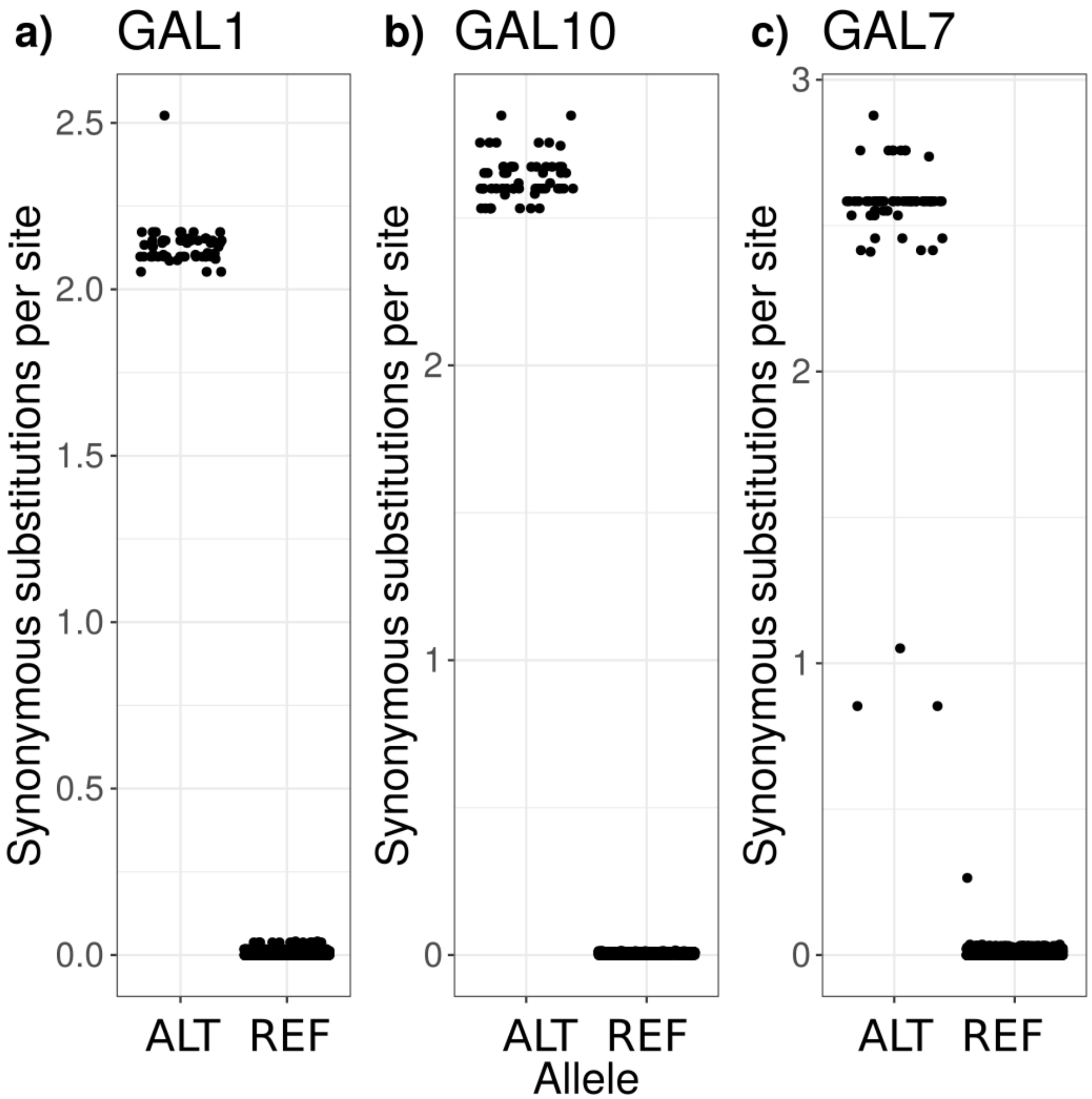


Fig. S11.
Synonymous substitutions per site (dS) for the alleles of *GAL1*, *GAL10*, and *GAL7* found in 1,276 sequenced yeast strains. Boxplots of the dS of *GAL1*, *GAL10*, and *GAL7* extracted where possible from the genomes of the 1,276 yeast strains when aligned to the reference genes. Strains are partitioned on the x-axis based on whether they were assigned as having the alternative (ALT) or reference (REF) allele for each gene.

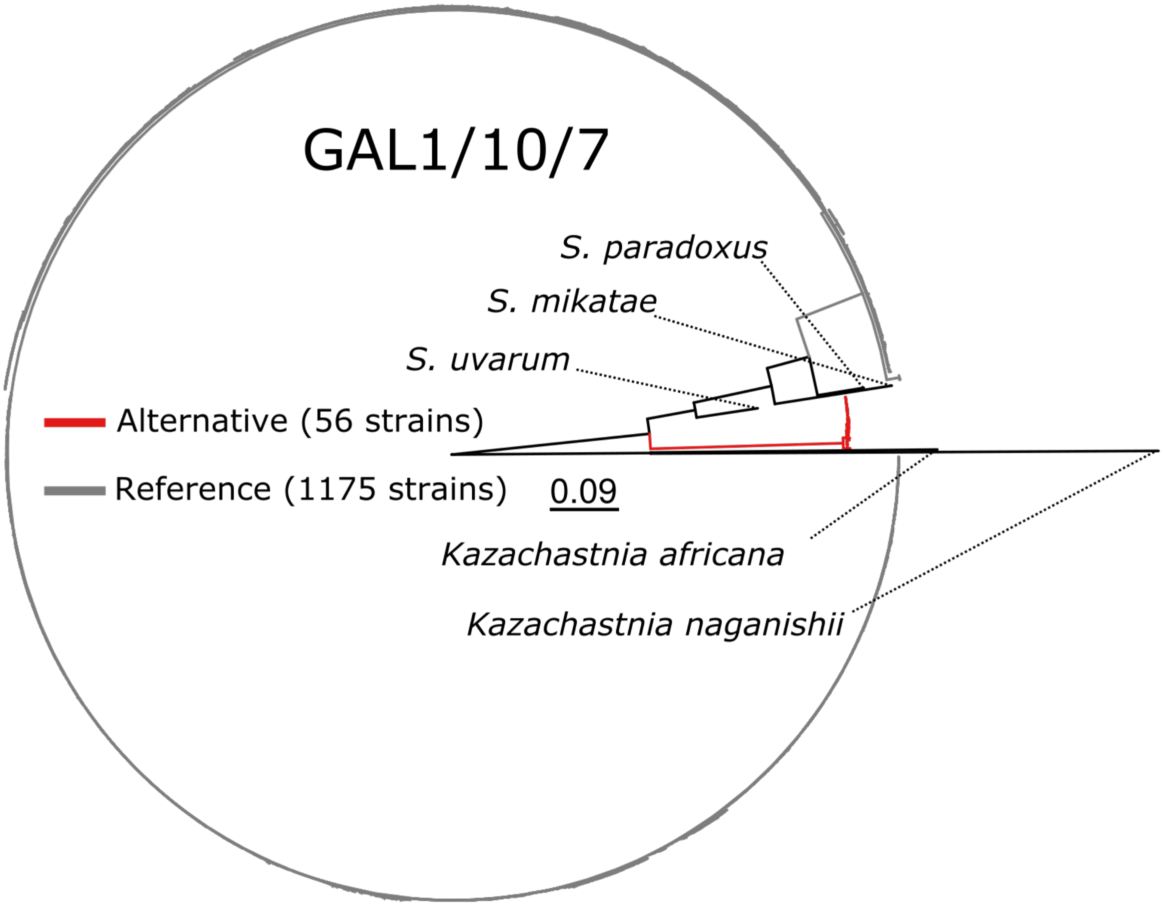


Fig. S12.
Phylogenetic clustering of the *GAL1/10/7* alleles in a global sample of 1,276 sequenced yeast strains. Maximum likelihood phylogenetic tree of the *GAL1/10/7* alleles.
Gene sequences were successfully extracted from the 1,231 strains with high-quality assemblies in this region. Gene sequences were also extracted from members of the *Saccharomyces* genus (*S. uvarum, S. mitakae, S. paradoxus*), and the outgroup species (*Kazachstania africana* and *Kazachstania naganashii*). Branches are colored according to the genotype of each strain at *GAL1/10/7*; alternative alleles (red), reference alleles (grey), and other species (black). Scale bar shows the estimated number of substitutions per site.

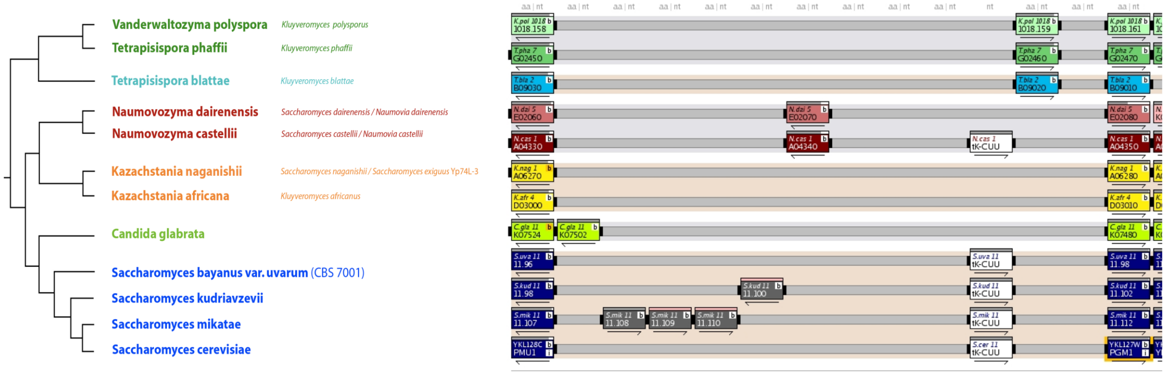


Fig. S13.
The missing lysine tRNA in the alternative PGM1 promoter is conserved within the *Saccharomyces* genus. Screenshot of the PGM1 promoter from the yeast gene order browser (*33*). Species from the *Saccharomyces* genus (*S. uvarum*, *S. kudriavzevii*, *S. mikatae*, and *S. cerevisiae*) all have a lysine tRNA (tk-CUU) in their *PGM1* promoters whereas the alternative *PGM1* promoter does not.

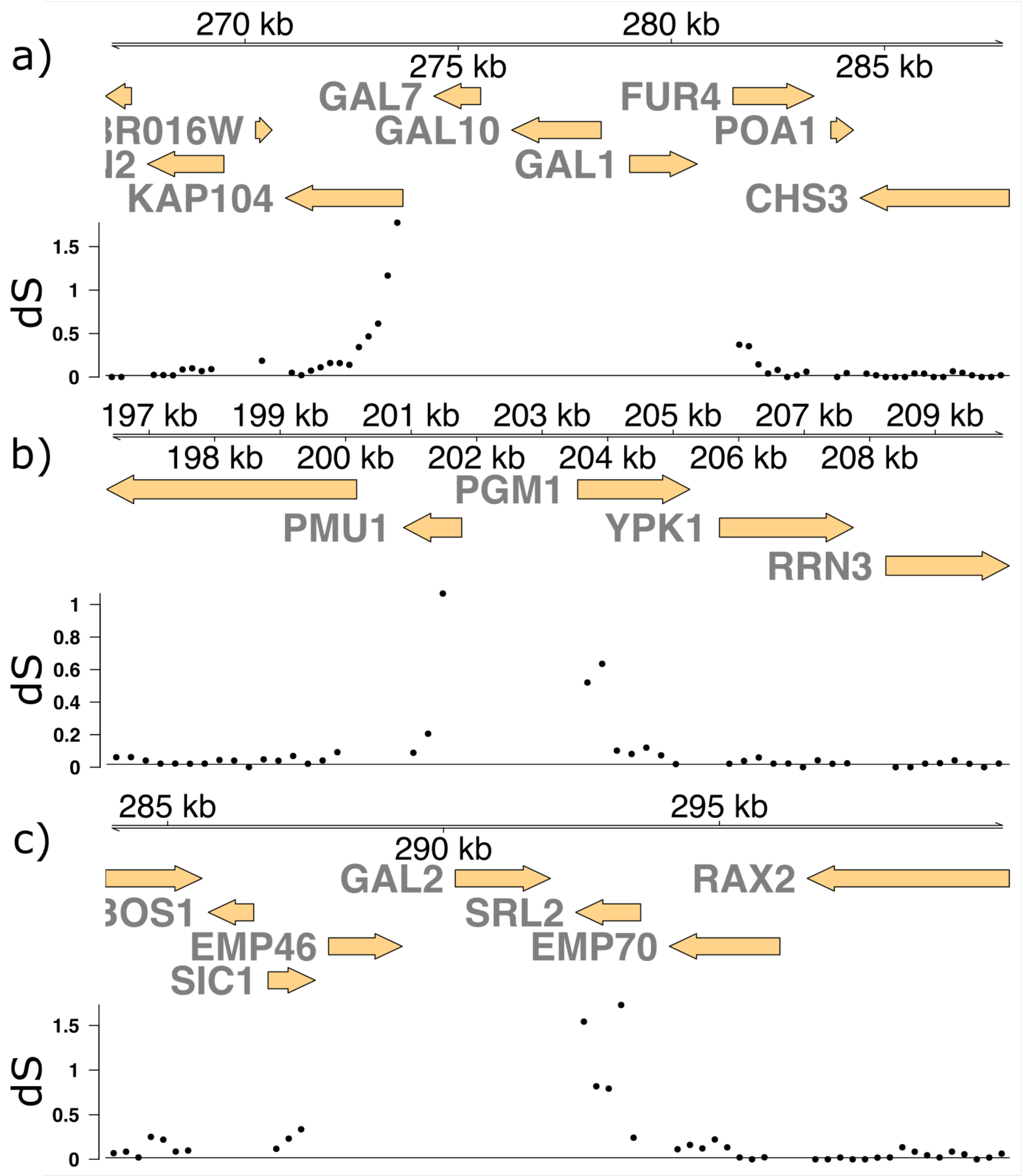


Fig. S14.
A signature of ancient balancing selection. Estimated rate of synonymous substitutions per site (dS) between CBS2888 (alternative) and BY (reference) genes shown for regions surrounding the galactose loci. Estimates of dS are plotted as dots for 75-codon windows stepped every 75 codons. a) genes adjacent to *GAL1/10/7*, b) genes adjacent to the *PGM1* promoter, c) genes adjacent to *GAL2*. dS was not estimated for *EMP46* due to the presence of an early stop codon interrupting this gene in CBS2888.

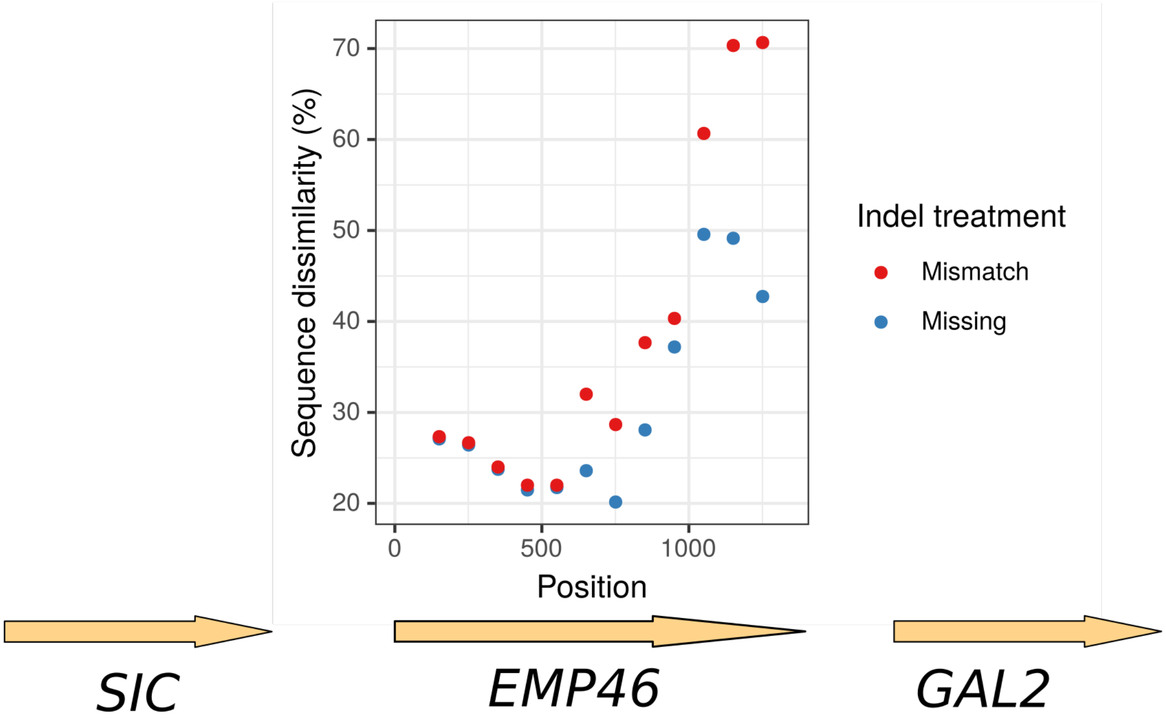


Fig. S15.
A signature of ancient balancing selection in the *EMP46* gene. Sequence dissimilarity (%) between CBS2888 (alternative) and BY reference) *EMP46* gene. The sequence dissimilarity (%) is plotted in 300 base-pair windows with a 100 base-pair step. We calculated the sequence dissimilarity by treating the indels as mismatches (red) or removing those positions (blue) from the alignment. The relative position of the nearby genes (*SIC1*, *GAL2*) are shown below.

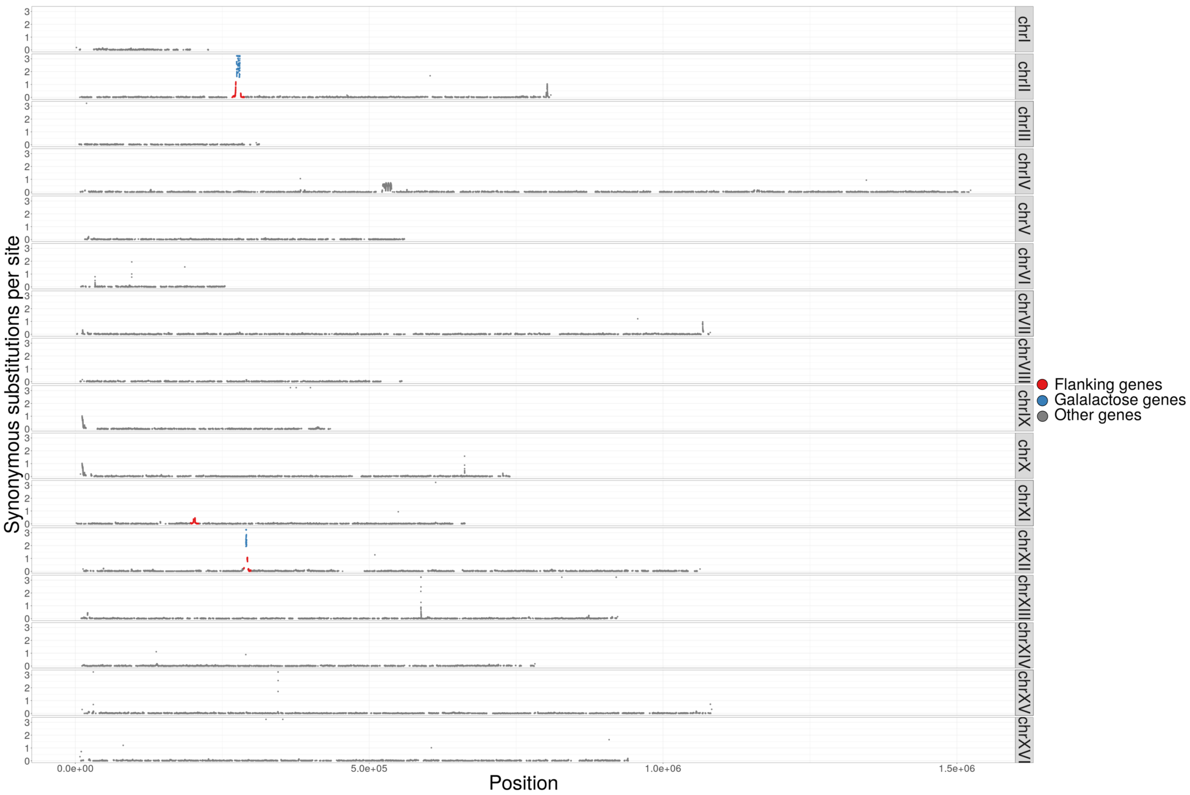


Fig. S16.
Genome-wide distribution of the synonymous substitutions per site (dS) 200-amino acid sliding windows between the CBS2888 and reference genes. The distribution of dS in 200-amino acid sliding windows with a step of 10-amino acids when the CBS2888 genes are aligned to the reference genes. We have highlighted in red genes near the galactose loci, in blue are the galactose genes (*GAL1, GAL10, GAL7,* and *GAL2*), and in grey are all other genes. *GAL1/10/7* are on ChrII, *PGM1* is on ChrXI, and *GAL2* is on ChrXII.

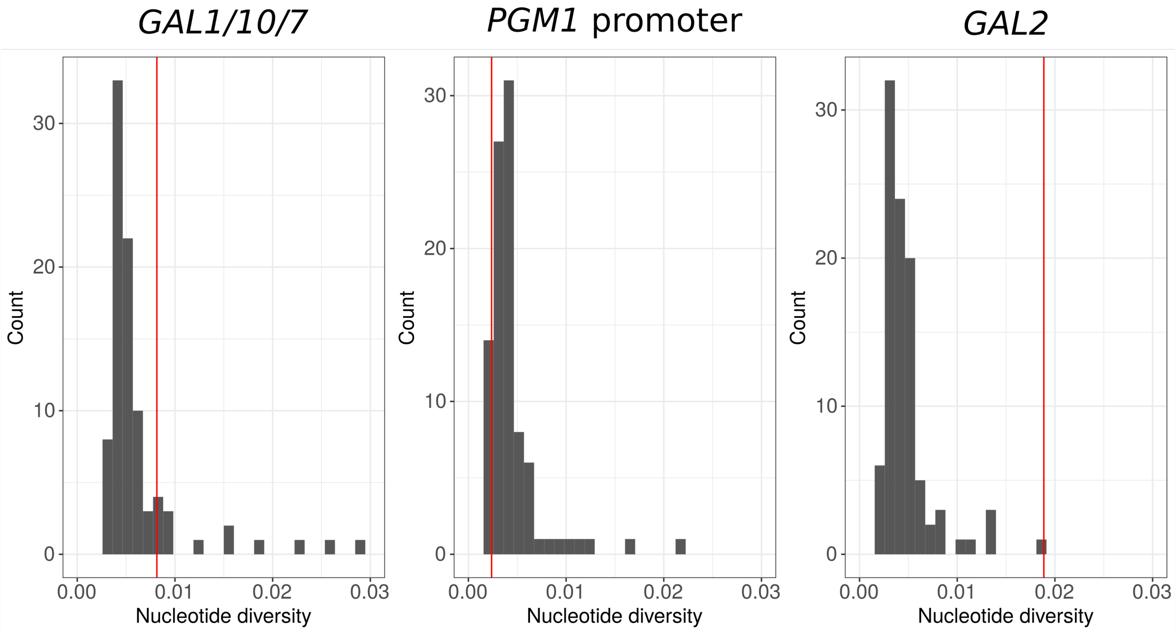


Fig. S17.
Intra-allelic diversity of the alternative alleles compared to other regions of the genome. Intra-allelic diversity was calculated between the alleles of strains with the a) alternative *GAL1/10/7* (n=57)*,* b) alternative *PGM1* promoter (n=49), and c) alternative *GAL2* (n=49). The observed sequence diversity for each of the loci is shown in red. The histograms show the intra-allelic diversity of 100 randomly selected background regions of the same length as each of the alternative regions.

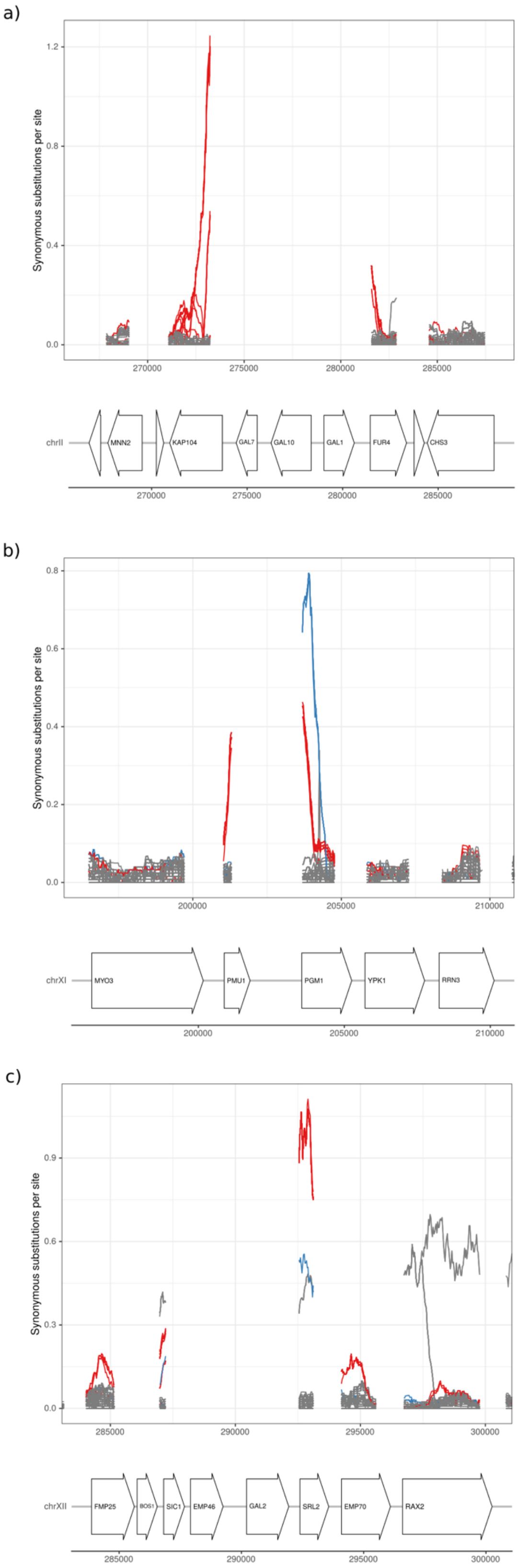


Fig. S18.
A signature of balancing selection in the global yeast population. Synonymous substitutions per site (dS) in 200-amino acid sliding windows between the 1,276 sequenced yeast strains and the reference in regions surrounding the galactose loci. The dS for strains with only alternative alleles are shown in red. The dS for strains with only reference alleles are shown in grey. The dS for strains from China with the alternative *GAL1/10/7* allele and alleles of *GAL2* and the *PGM1* promoter that differ from both the reference and the alternative alleles are shown in blue. The dS was not estimated for *EMP46* due to the presence of an early stop codon interrupting this gene in strains with the alternative alleles.

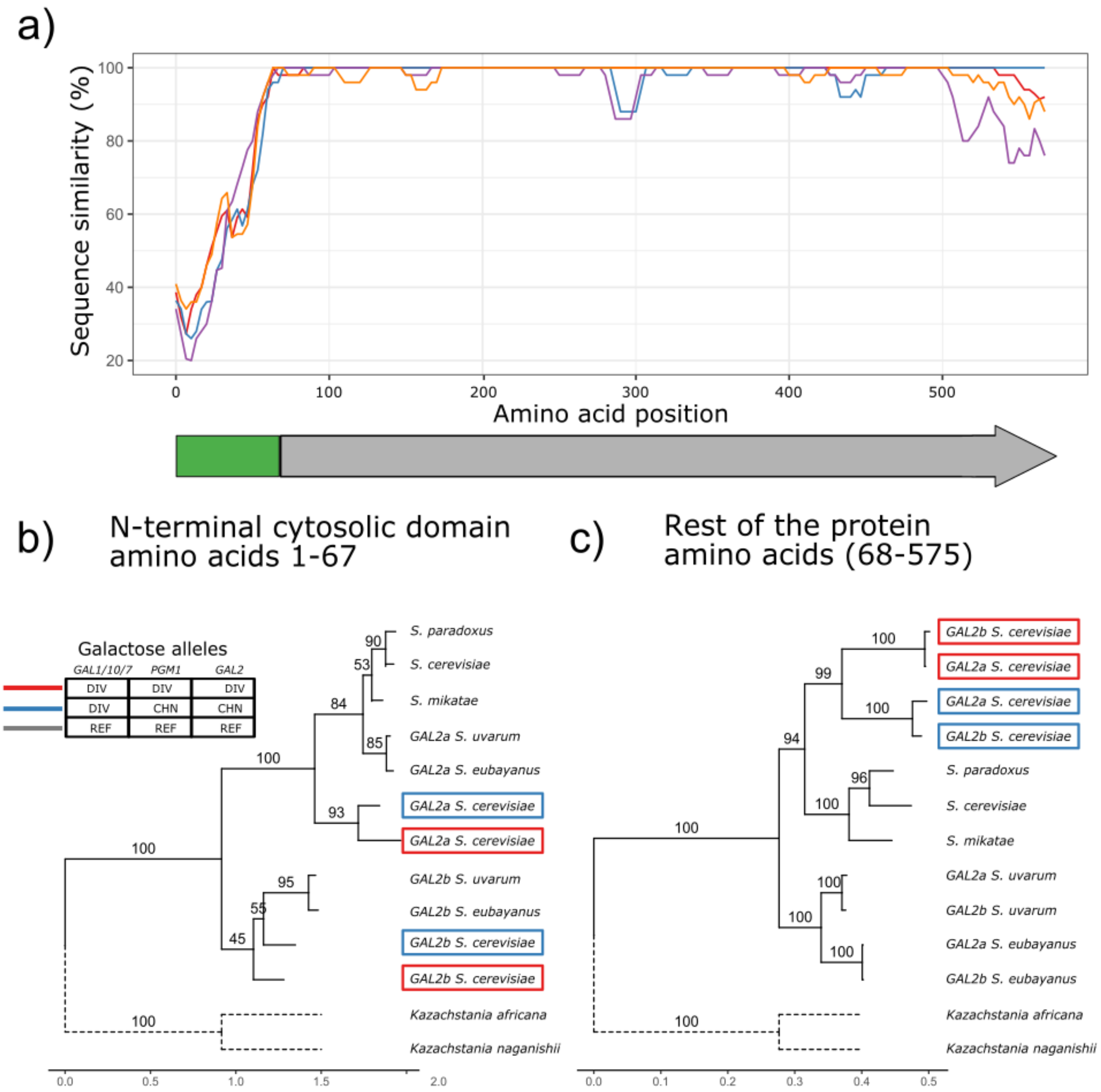


Fig. S19.
Phylogenetic analysis of *GAL2* across *Saccharomyces* species. a) Pairwise sequence similarity (%) between *GAL2a* and *GAL2b* calculated in 50 base-pair windows with a 10-base-pair step. Alternative (red) and Chinese (blue) alleles; *S. uvarum* (purple) and *S. eubayanus* (orange). In all cases, we denoted the centromere proximal *GAL2* as *GAL2a*, and the other as *GAL2b*. Annotated under the graph is the *GAL2* gene, with the N-terminal cytosolic domain highlighted in green. b) Bootstrapped maximum likelihood phylogenetic clustering of the N-terminal cytosolic region of *GAL2* (amino acids 1-67 reference aligned) from members of the *Saccharomyces* genus. Gene sequences were extracted from CBS2888 (alternative), BAM (Chinese), other members of the genus (*S. uvarum, S. eubayanus, S. paradoxus, S. mikatae*), and the outgroup species (*Kazachstania naganashii*, and *Kazachstania africana*). The outgroup branches (dotted lines)

were rescaled to the average branch length. The *S. cerevisiae* alleles are colored according to their classifications, alternative (red), Chinese (blue), and reference (grey). Scale bar shows the estimated number of substitutions per site. c) Bootstrapped maximum likelihood phylogenetic clustering of the remaining portion of the *GAL2* gene (amino acids 68-575 reference aligned).

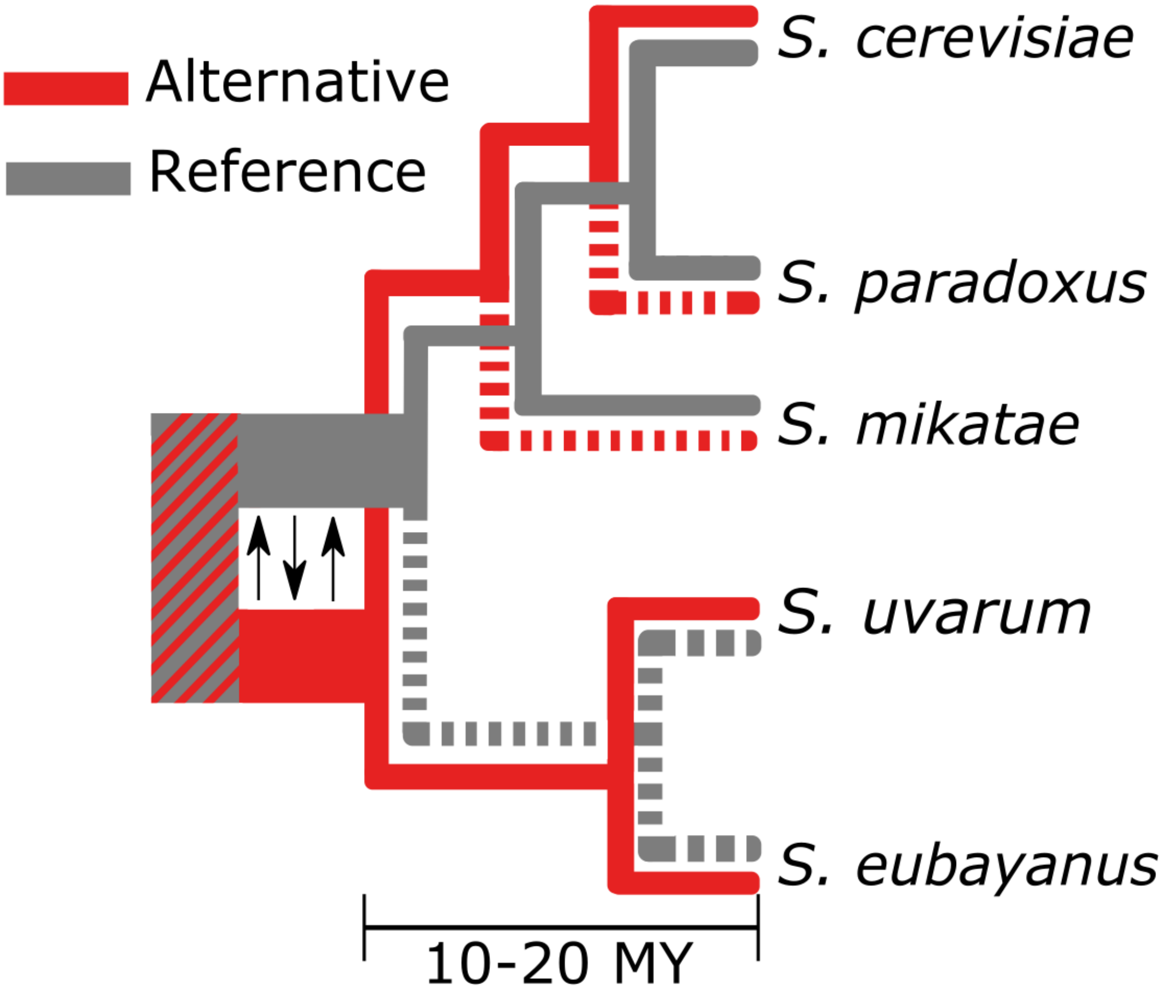


Fig. S20.
Proposed evolutionary history of the alternative and reference galactose pathways. The ancestral pathway of both the alternative and reference galactose pathways arose before the birth of the *Saccharomyces* genus 10-20 MYA. This is represented on the phylogenetic tree by the striped red and grey region. This ancestral pathway diverged into the alternative and reference pathway and balancing selection maintained both of these pathways until present day in *S. cerevisiae*. The evolutionary history of a pathway that is known to be present is drawn with a solid line. The missing pathways (dotted lines) may have been lost or remain unsampled. Gene flow has been occurring across the rest of the genome within all species since these pathways diverged millions of years ago.

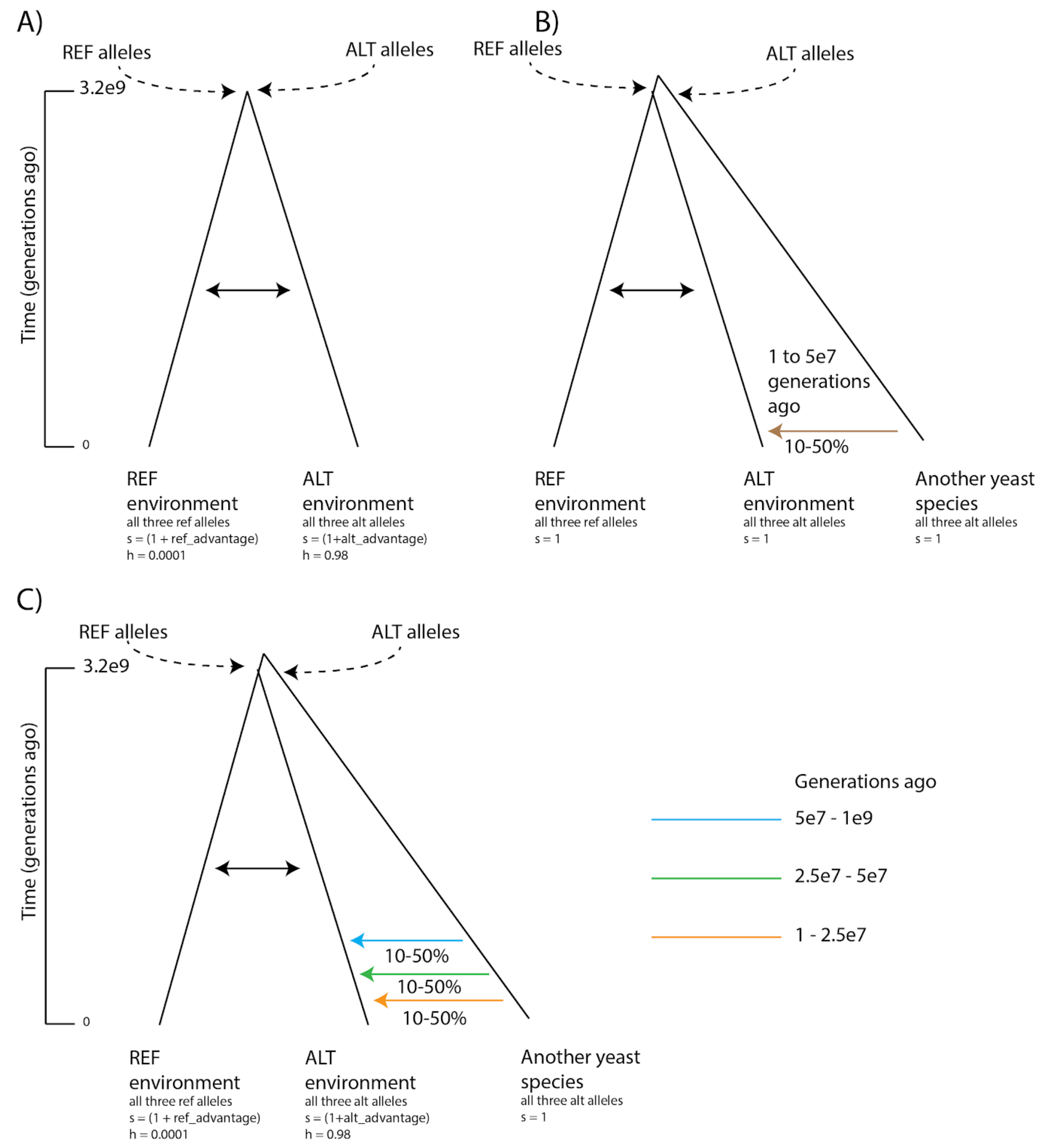


Fig. S21.
Diagrams of models explored using forward simulations. A) Ancient balancing selection B) Neutral introgression C) Introgression in three different time periods followed by maintenance of the alternative and reference galactose alleles with multi-locus balancing selection. The black arrow depicts migration occurring at 50% throughout the entire history of the two populations. The colored arrows in B and C depict the introgression pulse of 10-50% and the ranges of generations ago we simulated the introgression to occur. For all models, the simulation was run for $3.2x10^9$ generations with a burn-in of 8Ne.

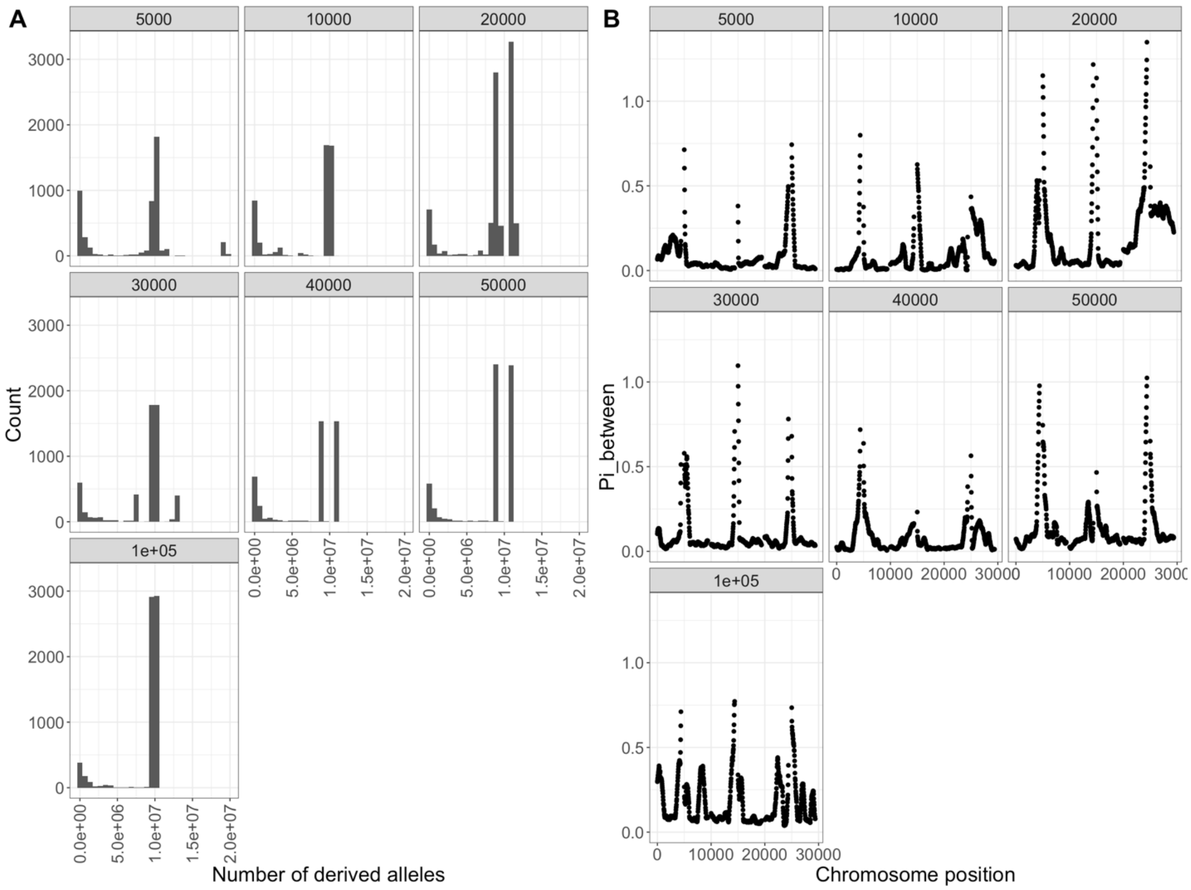


Fig. S22.
Parameter rescaling does not change the behavior of simulations using our ancient balancing selection model. A) The site frequency spectrum for simulations run using seven different scaling factors 5000, 10000, 20000, 30000, 40000, 50000, and 100000. We rescaled the population size of our models by this factor and increased the mutation rate, recombination rate, and selection coefficients by the same factor. B) The pairwise diversity between haplotypes with the reference and alternative alleles, summarized in 600 base-pair windows with a 30 base-pair step using these same seven scaling factors. For this simulation, we used a fixed set of parameters and only modified the scaling factor. These parameters were: a mitotic recombination rate of 0.1, a selfing rate of 0.998, a cloning rate of 0.99, a relative fitness for the alternative pathway in the alternative environment of 0.1, and a relative fitness for the reference pathway in the reference environment of 0.1.

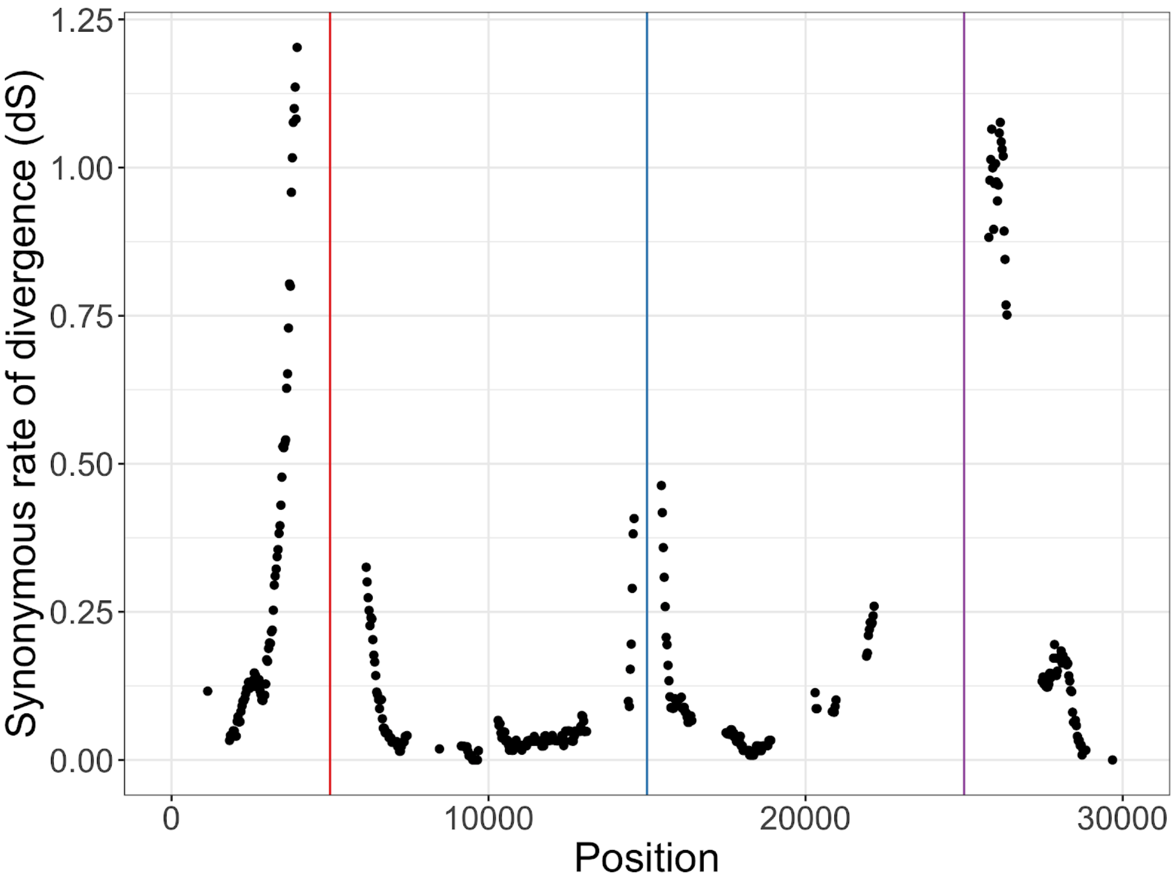


Fig. S23.
Synonymous rate of divergence (dS) presented in Figure 3, converted to the coordinate system of the simulations.

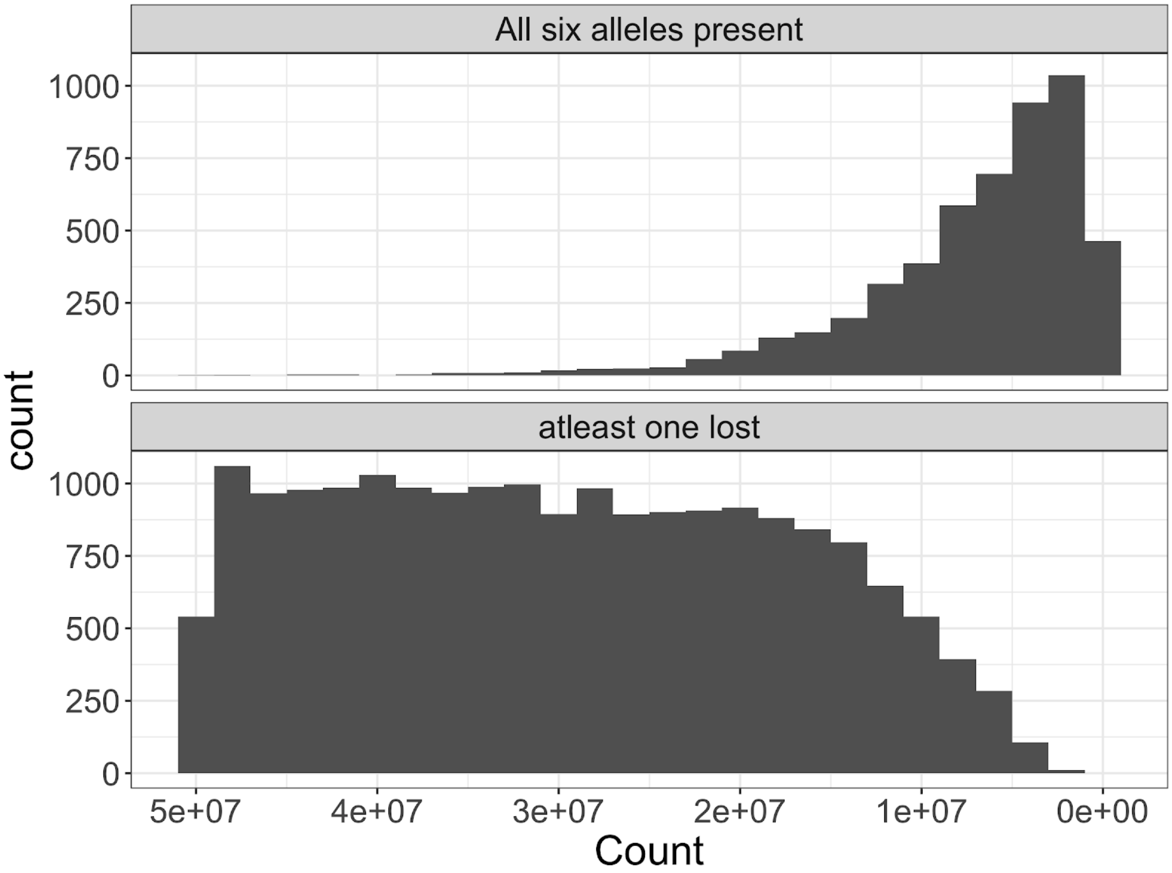


Fig. S24.
Loss of alleles in neutral introgression simulations. Categorization of whether all six galactose alleles survived to the end of each simulation (top), or at least one of the alleles got lost (bottom). For these simulations, we used a pulse introgression model, where the introgression was randomly chosen to occur between 50 million and 1 generation ago.

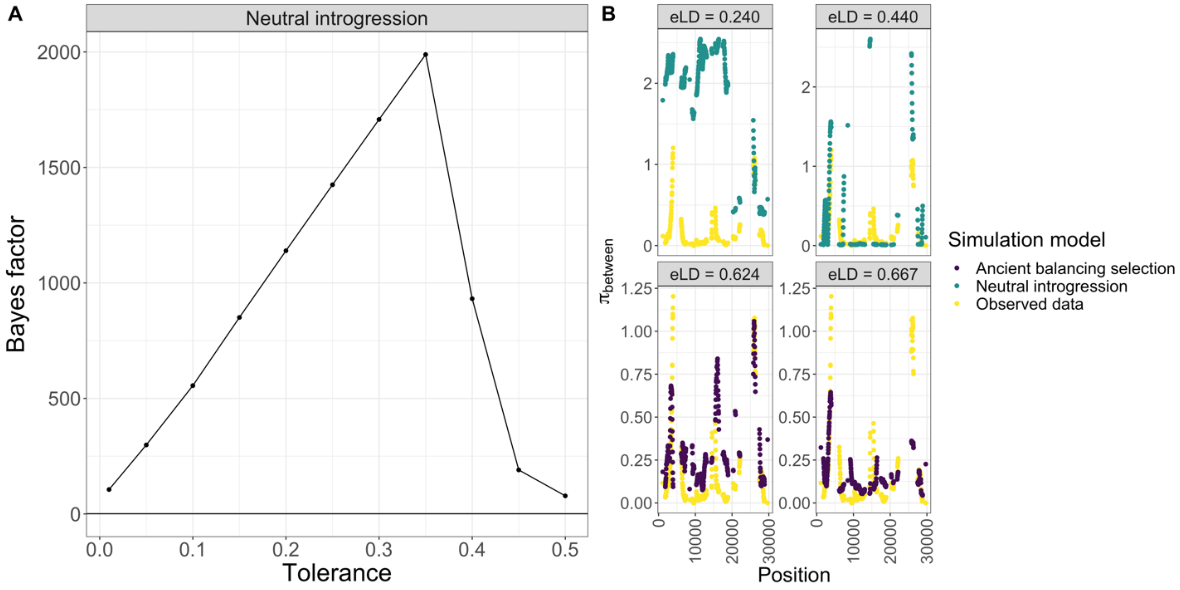


Fig. S25.
Comparisons of the ancient balancing selection to the neural recent introgression model. A) Bayes factor P(D | ancient balancing selection)/ P(D| neutral introgression) for the support of the ancient balancing selection over the neutral recent introgression model. B) The two simulations most similar to the observed data from the neutral recent introgression model (top), and the ancient balancing selection model (bottom). The dS values from the observed data are plotted in yellow. The eLD (ϵ) for each of these simulations is written at the top of each subplot. For reference the observed ϵ was 0.59.

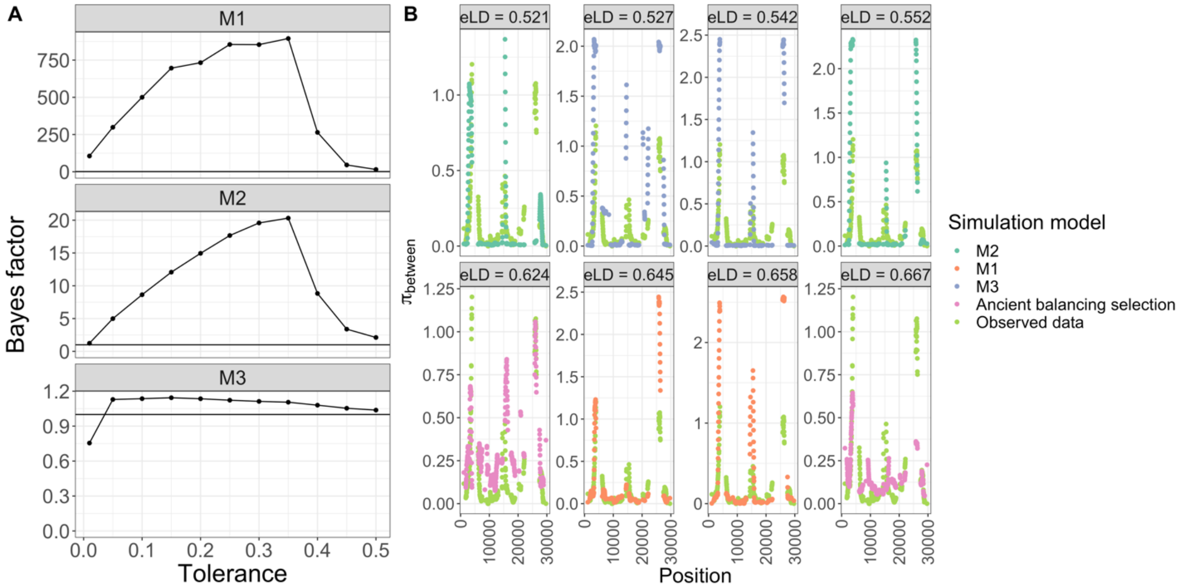


Fig. S26.
Comparisons of the ancient balancing selection to three models of introgression followed by maintenance of the alleles by balancing selection. A) Bayes factor P(D | Ancient balancing selection)/ P(D| M) for the support of the ancient balancing selection over a model of introgression followed by balancing selection over three time periods. The three time periods for the introgression followed by balancing selection model was between: 1 and 25,000,000 (M1); 25 million and 50 million (M2); and 50 million and 1 billion generations ago (M3). B) The two simulations with the lowest distance to the observed data from all four models (M1, M2, M3, and Ancient balancing selection). The observed dS signature is plotted in light green. The eLD (ϵ) for each of these simulations is shown at the top of each subplot.

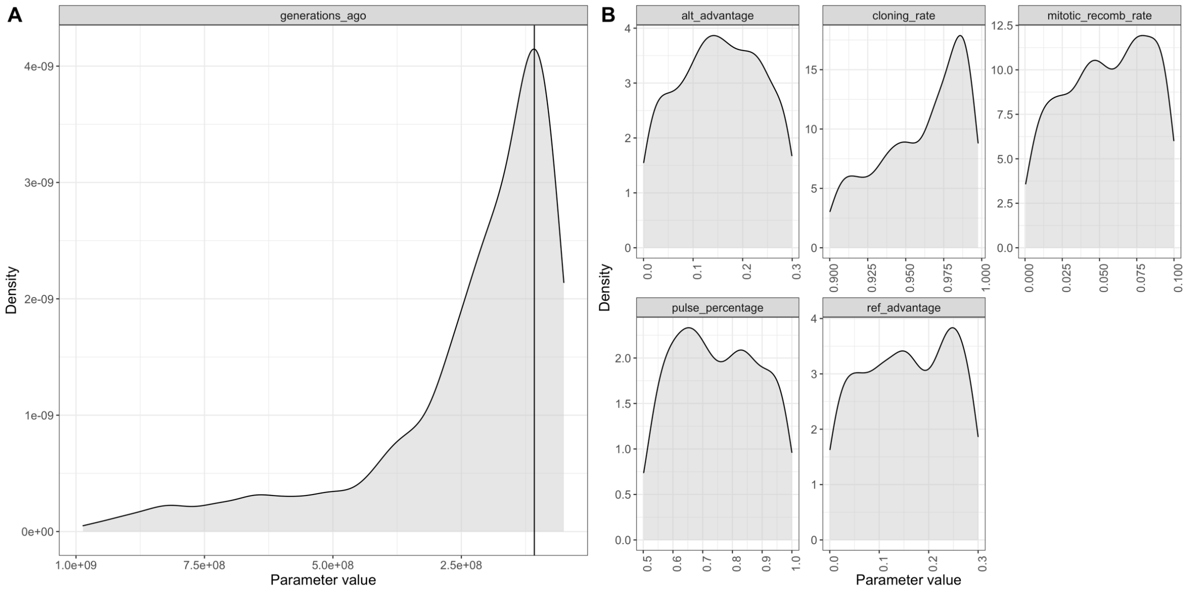
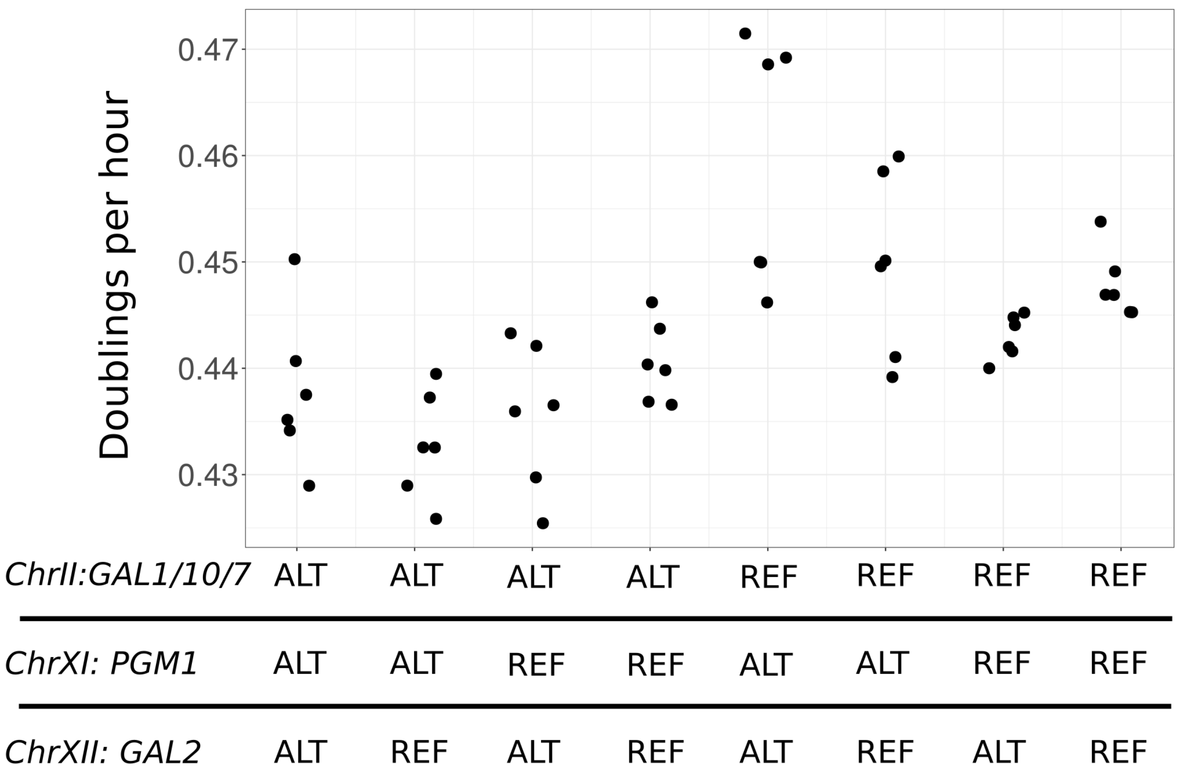


Fig. S27.
Posterior distributions of parameters for the introgression followed by balancing selection model, in which the introgression occurred between 1 billion and 50 million generations ago. A) Shows the posterior distributions of the generations ago each simulation was performed. A vertical black line was drawn at the posterior mode of 108 million generations ago. B) Posterior distributions of other parameters randomly chosen in each simulation. Simulations were accepted in the ABC framework using the reject method with the tolerance threshold set to 0.05. The prior distribution for each parameter was uniform.



ChrII:GAL1/10/7	ALT	ALT	ALT	ALT	REF	REF	REF	REF
ChrXI: PGM1	ALT	ALT	REF	REF	ALT	ALT	REF	REF
ChrXII: GAL2	ALT	REF	ALT	REF	ALT	REF	ALT	REF

Fig. S28.
Growth of allele replacement strains in 2% glucose medium. The doublings per hour in 2% glucose medium for allele replacement strains. Alleles at the QTL loci from CBS2888 are designated as ALT, and alleles from BY are designated as REF. Each dot shows a biological replicate.

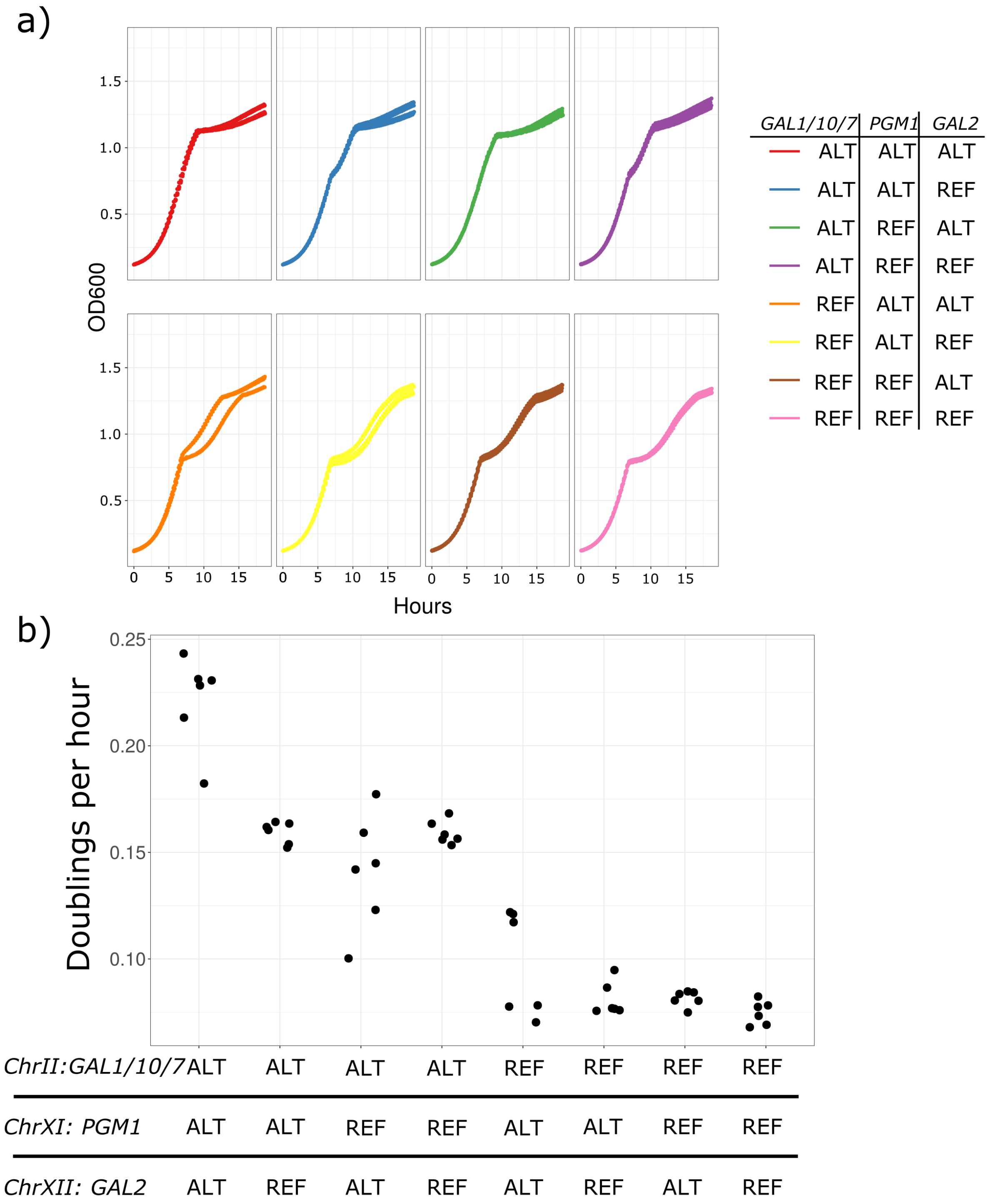

a)
1.5
1.0
0.5
OD600
0
5
10
15
Hours
GAL1/10/7 PGM1 GAL2
ALT ALT ALT
ALT ALT REF
ALT REF ALT
ALT REF REF
REF ALT ALT
REF ALT REF
REF REF ALT
REF REF REF
b)
Doublings per hour
0.25
0.20
0.15
0.10
ChrII:GAL1/10/7 ALT ALT ALT ALT REF REF REF REF
ChrXI: PGM1 ALT ALT REF REF ALT ALT REF REF
ChrXII: GAL2 ALT REF ALT REF ALT REF ALT REF

Fig. S29.
Strains with the diverged galactose alleles do not experience a diauxic shift when they switch from glucose to galactose medium. a) Growth curves for all eight allele replacement strains when grown in 1% glucose/1% galactose medium. Each growth curve is a summary of three biological replicates from each of the two strains for all eight possible combinations of alleles. b) The growth rate between 0.8 and 1.1 OD600 units for all eight allele replacement strains. Each dot represents a biological replicate. Quantifying this range of the growth curve captures the degree of diauxic shift.

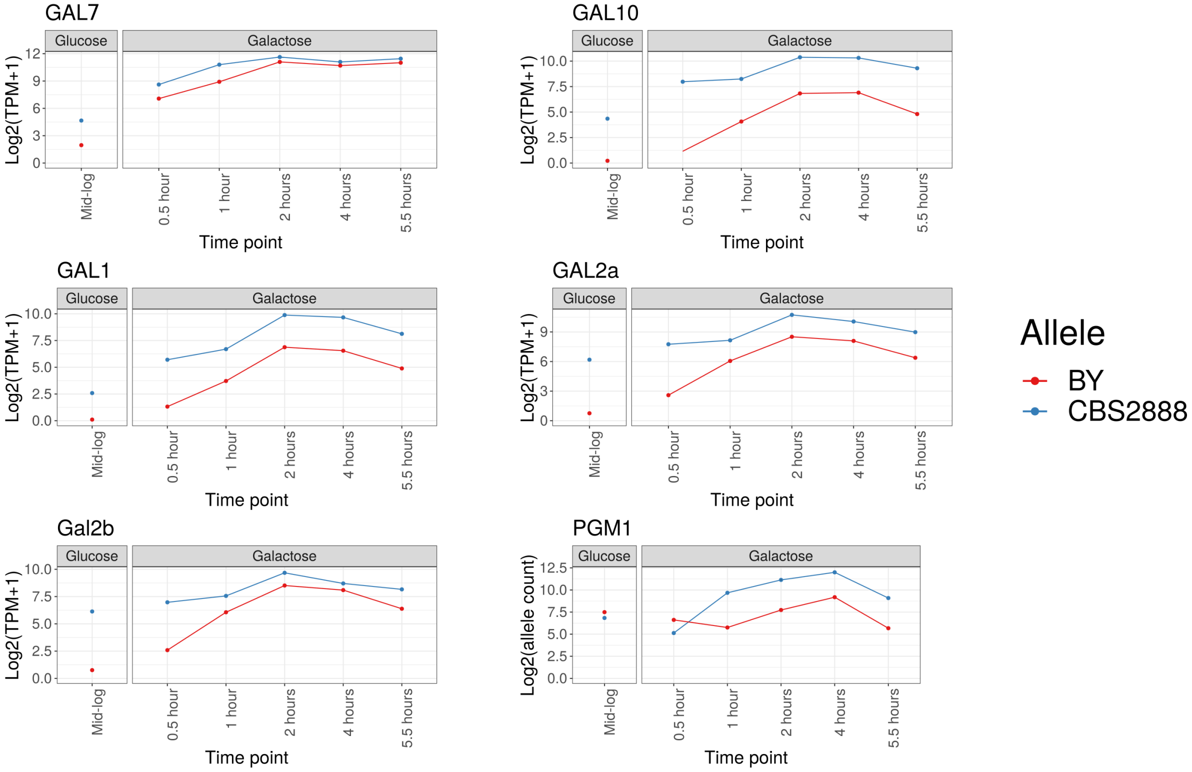


Fig. S30.

The alternative (CBS2888) galactose alleles are constitutively expressed in glucose and are more highly expressed in galactose. Allele-specific expression of a hybrid (CBS2888xYJM981) that is heterozygous for the alternative and reference alleles when grown in 2% glucose medium and transferred to 2% galactose medium. The line graph shows the $\log_2$ allele counts or transcripts per millions (TPM) for the CBS2888 (alternative) alleles and BY (reference) alleles. We denote the centromere-proximal copy of *GAL2* in CBS2888 as *GAL2a*, and the other one as *GAL2b*.

Table S1.
Three-way linear model coefficients and fits for segregants and allele replacement strains when grown in galactose. The model with the segregants included the intercept term, a term for each of the QTL markers, the three two-way QTL interaction terms, and the three-way QTL interaction term. The model with the allele replacement strains included the intercept term, a term for the identity of the three galactose loci, the three two-way interaction terms, and the three-way interaction term. The normalized colony size was used for the segregant model. The doublings per hour were used as the phenotype for the model with the allele replacement strains. For the QTL mapping the CBS2888 allele was coded positive, and the YJM981 allele was coded negative. For the allele replacement strains, the CBS2888 allele was coded as 1, and the reference allele was coded as 0.

Table S2.
Variance explained from a additive, two-way, and three-way linear model for segregants and allele replacement strains grown in galactose. Coefficient of determination (R^2) for the different models.

Table S3.
Significant three-way QTL interactions. The three-locus interaction term comes from a joint linear model, that includes all additive, and two-way interaction effects between the three loci. The p-value is derived from a likelihood ratio test in which the full model is compared to a model where only the additive and two-way interactions terms were fit.

Table S4.
Summary statistics for alignments between the alternative and reference galactose alleles. Synonymous substitutions per site (dS) could not be calculated for the PGM1 promoter. Confidence intervals were calculated by bootstrapping codons 200 times.

Table S5.
Strains used in this study.

Table S6.
Primer used in this study.

Table S7.
Plasmids used in this study.

Table S8.
Allele-specific expression of a hybrid (CBS2888xYJM981) that is heterozygous for the alternative and reference alleles when grown in 2% glucose medium and transferred to 2% galactose medium. a) All genes with at least 5 reference and 5 alternate counts, a false-discovery rate adjusted P-value of less than 5% from a binomial test, and a $\log_2$ fold-change of greater than 2 are reported in the table. b) Allele-specific expression for the galactose genes (*GAL1*, *GAL10*,

GAL7, *GAL2*). The $\log_2$ fold-change is calculated using the estimated counts of the reference and CBS2888 galactose genes.

Table S9.
Classification of galactose loci in the global collection of 1,276 yeast strains. For each strain we report the geographical location, isolation, and the assignment of each allele to reference or alternative version. The alternative *GAL1/10/7* allele was further subdivided into each gene since we found that some strains did not exclusively have the alternative or reference version of the entire region.

Table S10.
Estimates of the synonymous substitutions per site (dS) between the genes of the reference and CBS2888.

Table S11.
Parameters used for the population genetic forward simulations.

References and Notes

1. J. M. Saudubray, À. Garcia-Cazorla, Inborn Errors of Metabolism Overview: Pathophysiology, Manifestations, Evaluation, and Management. *Pediatr. Clin. North Am.* **65**, 179–208 (2018).

2. J. Nielsen, Systems Biology of Metabolism. *Annu. Rev. Biochem.* **86**, 245–275 (2017).

3. C. A. Sellick, R. N. Campbell, R. J. Reece, Galactose metabolism in yeast-structure and regulation of the leloir pathway enzymes and the genes encoding them. *Int. Rev. Cell Mol. Biol.* **269**, 111–150 (2008).

4. C. T. Hittinger, P. Gonçalves, J. P. Sampaio, J. Dover, M. Johnston, A. Rokas, Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature* **464**, 54–58 (2010).

5. D. Charlesworth, Balancing selection and its effects on sequences in nearby genome regions. *PLOS Genet.* **2**, e64 (2006).

6. J. S. Bloom, J. Boocock, S. Treusch, M. J. Sadhu, L. Day, H. Oates-Barker, L. Kruglyak, Rare variants contribute disproportionately to quantitative trait variation in yeast. *eLife* **8**, e49212 (2019).

7. Materials and methods and supplementary text are available as supplementary materials.

8. M. J. Sadhu, J. S. Bloom, L. Day, J. J. Siegel, S. Kosuri, L. Kruglyak, Highly parallel genome variant engineering with CRISPR-Cas9. *Nat. Genet.* **50**, 510–514 (2018).

9. F. W. Albert, L. Kruglyak, The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).

10. A. Traven, B. Jelicic, M. Sopta, Yeast Gal4: A transcriptional paradigm revisited. *EMBO Rep.* **7**, 496–499 (2006).

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

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

13. Y. Okada, eLD: Entropy-based linkage disequilibrium index between multiallelic sites. *Hum. Genome Var.* **5**, 29 (2018).

14. J. L. Legras, V. Galeote, F. Bigey, C. Camarasa, S. Marsit, T. Nidelet, I. Sanchez, A. Couloux, J. Guy, R. Franco-Duarte, M. Marcet-Houben, T. Gabaldon, D. Schuller, J. P. Sampaio, S. Dequin, Adaptation of s. Cerevisiae to fermented food environments reveals remarkable genome plasticity and the footprints of domestication. *Mol. Biol. Evol.* **35**, 1712–1727 (2018).

15. M. Kellis, N. Patterson, M. Endrizzi, B. Birren, E. S. Lander, Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).

16. D. R. Scannell, O. A. Zill, A. Rokas, C. Payen, M. J. Dunham, M. B. Eisen, J. Rine, M. Johnston, C. T. Hittinger, The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. *Genes Genomes Genet.* **1**, 11–25 (2011).

17. S. F. Duan, J.-Y. Shi, Q. Yin, R.-P. Zhang, P.-J. Han, Q.-M. Wang, F.-Y. Bai, Reverse Evolution of a Classic Gene Network in Yeast Offers a Competitive Advantage. *Curr. Biol.* **29**, 1126–1136.e5 (2019).

18. B. Turcotte, X. B. Liang, F. Robert, N. Soontorngun, Transcriptional regulation of nonfermentable carbon utilization in budding yeast. *FEMS Yeast Res.* **10**, 2–13 (2010).

19. J. H. Kim, A. Roy, D. Jouandot II, K. H. Cho, The glucose signaling network in yeast. *Biochim. Biophys. Acta* **1830**, 5204–5210 (2013).

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

21. G. I. Lang, A. W. Murray, D. Botstein, The cost of gene expression underlies a fitness trade-off in yeast. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 5755–5760 (2009).

22. A. I. Coelho, M. E. Rubio-Gozalbo, J. B. Vicente, I. Rivera, Sweet and sour: An update on classic galactosemia. *J. Inherit. Metab. Dis.* **40**, 325–342 (2017).

23. K. Lai, L. J. Elsas, K. J. Wierenga, Galactose toxicity in animals. *IUBMB Life* **61**, 1063–1074 (2009).

24. J. Boocock, M. J. Sadhu, A. Durvasula, J. S. Bloom, L. Kruglyak, Ancient balancing selection maintains incompatible versions of the galactose pathway in yeast (Figure creation), Dryad (2020); https://doi.org/10.5068/D14370.

25. J. Boocock, theboocock/ancient_bal_scripts: submission, Zenodo (2020); https://doi.org/10.5281/zenodo.4132713.

26. J. Boocock, theboocock/popgen_utilities: submission, Zenodo (2020); https://doi.org/10.5281/zenodo.4131787.

27. J. Boocock, theboocock/gal_bal: submission, Zenodo (2020); https://doi.org/10.5281/zenodo.4107954.

28. H. Wickham, ggplot2: Elegant Graphics for Data Analysis (Springer, 2016).

29. R Core Team, R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, 2018); www.r-project.org.

30. J. E. DiCarlo, J. E. Norville, P. Mali, X. Rios, J. Aach, G. M. Church, Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Res.* **41**, 4336–4343 (2013).

31. B. van de Geijn, G. McVicker, Y. Gilad, J. K. Pritchard, WASP: Allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).

32. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

33. K. P. Byrne, K. H. Wolfe, The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* **15**, 1456–1461 (2005).

34. J. X. Yue, J. Li, L. Aigrain, J. Hallin, K. Persson, K. Oliver, A. Bergström, P. Coupland, J. Warringer, M. C. Lagomarsino, G. Fischer, R. Durbin, G. Liti, Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat. Genet.* **49**, 913–924 (2017).

35. E. Baker, B. Wang, N. Bellora, D. Peris, A. B. Hulfachor, J. A. Koshalek, M. Adams, D. Libkind, C. T. Hittinger, The genome sequence of Saccharomyces eubayanus and the domestication of lager-brewing yeasts. *Mol. Biol. Evol.* **32**, 2818–2831 (2015).

36. K. L. Howe, B. Contreras-Moreira, N. De Silva, G. Maslen, W. Akanni, J. Allen, J. Alvarez-Jarreta, M. Barba, D. M. Bolser, L. Cambell, M. Carbajo, M. Chakiachvili, M. Christensen, C. Cummins, A. Cuzick, P. Davis, S. Fexova, A. Gall, N. George, L. Gil, P. Gupta, K. E. Hammond-Kosack, E. Haskell, S. E. Hunt, P. Jaiswal, S. H. Janacek, P. J. Kersey, N. Langridge, U. Maheswari, T. Maurel, M. D. McDowall, B. Moore, M. Muffato, G. Naamati, S. Naithani, A. Olson, I. Papatheodorou, M. Patricio, M. Paulini, H. Pedro, E. Perry, J. Preece, M. Rosello, M. Russell, V. Sitnik, D. M. Staines, J. Stein, M. K. Tello-Ruiz, S. J. Trevanion, M. Urban, S. Wei, D. Ware, G. Williams, A. D. Yates, P. Flicek, Ensembl Genomes 2020-enabling non-vertebrate genomic research. *Nucleic Acids Res.* **48**, D689–D695 (2020).

37. J. M. Cherry, E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, E. T. Chan, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, B. C. Hitz, K. Karra, C. J. Krieger, S. R. Miyasato, R. S. Nash, J. Park, M. S. Skrzypek, M. Simison, S. Weng, E. D. Wong, Saccharomyces Genome Database: The genomics resource of budding yeast. *Nucleic Acids Res.* **40**, D700–D705 (2012).

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

39. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

40. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

41. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

42. X. Zheng, D. Levine, J. Shen, S. M. Gogarten, C. Laurie, B. S. Weir, A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).

43. E. Paradis, K. Schliep, ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

44. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T. Y. Lam, Ggtree: An R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. *Methods Ecol. Evol.* **8**, 28–36 (2017).

45. P. K. Strope, D. A. Skelly, S. G. Kozmin, G. Mahadevan, E. A. Stone, P. M. Magwene, F. S. Dietrich, J. H. McCusker, The 100-genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Res.* **25**, 762–774 (2015).

46. B. Istace, A. Friedrich, L. d'Agata, S. Faye, E. Payen, O. Beluche, C. Caradec, S. Davidas, C. Cruaud, G. Liti, A. Lemainque, S. Engelen, P. Wincker, J. Schacherer, J.-M. Aury, de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *Gigascience* **6**, giw018 (2017).

47. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

48. A. Krogh, B. Larsson, G. von Heijne, E. L. L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).

49. P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, M. J. L. de Hoon, Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

50. M. Nei, T. Gojobori, Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).

51. W. R. Pearson, Finding protein and nucleotide similarities with FASTA. *Curr. Protoc. Bioinformatics* **53**, 3.9.1–3.9.25 (2016).

52. B. C. Haller, P. W. Messer, SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. *Mol. Biol. Evol.* **36**, 632–637 (2019).

53. G. I. Lang, A. W. Murray, Estimating the per-base-pair mutation rate in the yeast Saccharomyces cerevisiae. *Genetics* **178**, 67–82 (2008).

54. D. M. Ruderfer, S. C. Pratt, H. S. Seidel, L. Kruglyak, Population genomic analysis of outcrossing and recombination in yeast. *Nat. Genet.* **38**, 1077–1081 (2006).

55. M. Nordborg, Linkage disequilibrium, gene trees and selfing: An ancestral recombination graph with partial self-fertilization. *Genetics* **154**, 923–929 (2000).

56. M. Kimura, T. Ohta, The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population. *Genetics* **61**, 763–771 (1969).

57. K. Csilléry, O. François, M. G. B. Blum, Abc: An R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* **3**, 475–479 (2012).

58. C. P. Robert, J.-M. Cornuet, J.-M. Marin, N. S. Pillai, Lack of confidence in approximate Bayesian computation model choice. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 15112–15117 (2011).