PNAS RESEARCH ARTICLE | EVOLUTION

Convergent expansions of keystone gene families drive metabolic innovation in Saccharomycotina yeasts

Kyle T. David[a,b], Joshua G. Schraiber[c,d], Johnathan G. Crandall[e], Abigail L. Labella[a,b,f], Dana A. Opulente[e,g], Marie-Claire Harrison[a,b], John F. Wolters[e], Xiaofan Zhou[h], Xing-Xing Shen[i], Marizeth Groenewald[j], Chris Todd Hittinger[e,1], Matt Pennell[c,d,k,1], and Antonis Rokas[a,b,1]

Affiliations are included on p. 5.

Edited by David Hillis, The University of Texas at Austin, Austin, TX; received January 6, 2025; accepted May 2, 2025

Many remarkable phenotypes have repeatedly occurred across vast evolutionary distances. When convergent traits emerge on the tree of life, they are sometimes driven by the same underlying gene families, while other times, many different gene families are involved. Conversely, a gene family may be repeatedly recruited for a single trait or many different traits. To understand the general rules governing convergence at both genomic and phenotypic levels, we systematically tested associations between 56 binary metabolic traits and gene count in 14,785 gene families from 993 Saccharomycotina yeasts. Using a recently developed phylogenetic approach that reduces spurious correlations, we found that gene family expansion and contraction were significantly linked to trait gain and loss in 45/56 (80%) traits. While 595/739 (81%) significant gene families were associated with only one trait, we also identified several "keystone" gene families that were significantly associated with up to 13/56 (23%) of all traits. Strikingly, most of these families are known to encode metabolic enzymes and transporters, including all members of the industrially relevant *MAL* tose fermentation loci in the baker's yeast *Saccharomyces cerevisiae*. These results indicate that convergent evolution on the gene family level may be more widespread across deeper timescales than previously believed.

deep homology | convergent evolution | gene duplication | innovation | yeasts

Significance

Convergent evolution occurs when the same trait arises independently on the tree of life. How often such traits stem from changes in the same genetic elements remains poorly understood, particularly above the species level. Studying the convergent evolution of dozens of metabolic traits across a species-rich and ancient yeast lineage, we found that gene gain and loss in specific gene families reliably predicted the convergent evolution of most traits. We also found that certain gene families predicted the convergent evolution of multiple traits, suggesting that they encode general functions that are repeatedly utilized for diverse biochemical processes. Our study shows that select gene families have been repeatedly recruited in many convergently evolved metabolic traits, even across vast evolutionary timescales.

The repeated emergence of convergent traits has long been used to provide strong evidence for the power of natural selection and predictability of evolution (1). However, the extent to which convergent traits are achieved via the same genetic elements remains a longstanding question (2–5) (Fig. 1). For example, image-forming eyes are a classic example of convergent evolution, having emerged several times across Metazoa (6). Due to the vast evolutionary distances and morphological variation found across eyes, they were long assumed to have evolved via an equally diverse number of genetic pathways (7). However, in a remarkable case of deep homology, it is now established that the development of all animal eyes is globally controlled by an ancient regulatory system governed by a *PAX6* homolog (8, 9). In contrast, antifreeze proteins, another classic case of convergent evolution, have repeatedly evolved from diverse genomic origins. Even proteins with virtually identical structures and sequences have evolved from at least three distinct gene families (10).

Image-forming eyes and antifreeze proteins are highly specialized traits associated with specific functions. More general traits that are essential to the basic maintenance of an organism are expected to be more conserved and experience stronger purifying selection, such that convergence may be limited (5, 7, 11). However, even for fundamental housekeeping processes, convergent solutions occur. For example, oxygen transport has evolved many times through both homologous and nonhomologous means (7): The protein hemerythrin has been recruited for oxygen transport at least four times across three animal phyla (12) and is itself analogous to other convergently evolving O_2-binding protein families like the hemocyanins and hemoglobins (13).

Examples of nonhomologous and (especially) homologous convergent evolution across deeper evolutionary distances, such as those described above, are few and far between. Much of the previous work studying convergence at genomic and phenotypic levels has focused on relatively recent changes, identifying convergent patterns in single genes or even nucleotides across populations (14). For example, the repeated fixation of ectodysplasin alleles has been shown to reduce armor plates in freshwater populations of stickleback fish (15). However, such cases are likely dependent on preexisting shared genetic variation (11, 15, 16). By contrast, the rules governing the convergence of truly independent events, such as those that occur across gene families, are still not well understood.

Author contributions: K.T.D., C.T.H., M.P., and A.R. designed research; K.T.D., J.G.S., and J.G.C. performed research; K.T.D., J.G.S., J.G.C., A.L.L., D.A.O., M.-C.H., J.F.W., X.Z., X.-X.S., M.G., C.T.H., M.P., and A.R. contributed new reagents/analytic tools; K.T.D., J.G.S., J.G.C., C.T.H., M.P., and A.R. analyzed data; and K.T.D., M.P., and A.R. wrote the paper.

Competing interest statement: A.R. is a scientific consultant for LifeMine Therapeutics, Inc. The other authors declare no other competing interests.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: cthittinger@wisc.edu, mpennell@cornell.edu, or antonis.rokas@vanderbilt.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2500165122/-/DCSupplemental.

Published June 3, 2025.

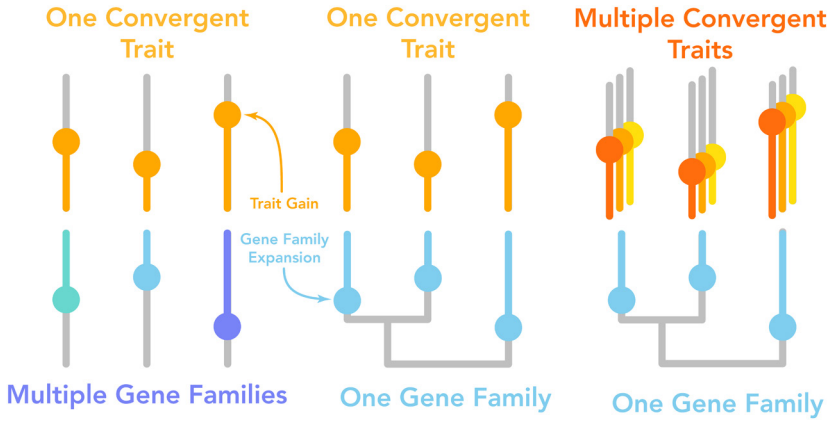

Fig. 1. Models of convergent evolution. Gains of novel traits (*Top*) and genes (*Bottom*) are represented by colored dots which appear across lineages (lines), with unique traits/genes represented by separate colors. *Top* row: convergence occurs when the same trait appears multiple times across separate lineages on the tree of life. *Bottom* row: convergent traits can originate in a homologous fashion by changes within a single gene family (*Center*), as is the case with image-forming eyes, or in an analogous fashion by changes across multiple different gene families (*Left*), as with freezing resistance. Likewise, gene family evolution may cause convergence in one trait (*Center*) or many different traits (*Right*). Our study reports that homologous events in gene family evolution explain convergence in 80% of analyzed traits (*Center*). While the majority of these families are associated with a single trait (*Center*), we also find evidence of select few "keystone" gene families that exhibit pleiotropic effects across many different traits (*Right*).

Gene family evolutionary events such as gene duplication (17–20), gene loss (21, 22), and horizontal gene transfer (23, 24) are often invoked in trait evolution and innovation. However, support for causal relationships is often weak, as traits often evolve only once (25), leaving no degrees of freedom to statistically test gene family–trait evolution associations (26). To overcome these issues, we tested phylogenetic correlations between gene family size and growth across 56 carbon and nitrogen substrates for 993 Saccharomycotina yeasts ("yeasts" hereafter). These 56 metabolic traits (27, 28) have been gained dozens to hundreds of times across more than 400 My of evolution, providing unprecedented statistical power for investigating convergent evolutionary events.

Yeasts are able to metabolize alcohols, ketones, organic acids, and more, which has enabled them to colonize virtually every continent and biome on the planet (29, 30). The spectacular diversity of yeast metabolism has not gone unnoticed by humans. In addition to the genus *Saccharomyces*, whose metabolisms underwrite the baking, brewing, and winemaking industries, many yeasts, such as *Yarrowia lipolytica, Lipomyces starkeyi,* and *Komagataella (Pichia) pastoris*, have unique metabolisms that are exploited for technological and industrial applications (31–33). By sampling genomes from ~80% of described Saccharomycotina species and traits from across the metabolic spectrum, we were able to systematically quantify the extent of gene family convergence within and across traits throughout a wide swath of genetic and phenotypic diversity.

Convergent Gene Family Expansion Undergirds Metabolic Diversity in Yeasts. Metabolic variation in Saccharomycotina is extensive, from specialist species able to metabolize one or two compounds to extreme generalists found growing on 47/56 (84%) tested substrates (*SI Appendix*, Fig. S1). Genetic variation is also broad; a family of RNA-directed DNA polymerases (K00986) is absent or represented by a single gene in certain species but can have as many as 290 genes in others (*SI Appendix*, Fig. S2). To test associations between gene family size and metabolic breadth, we ran a modified phylogenetic logistic regression model for each combination of traits and gene families (N = 817,591 tests). We identified 739/14,785 (5%) families with significant (FDR < 10^{-6}) associations with one or more traits. These families were significantly (FDR < 0.05) enriched in five KEGG pathways, all of which belong to the metabolism category, including metabolism in diverse environments (*SI Appendix*, Table S1). 45/56 (80%) traits were significantly associated with at least one (on average 12) gene families, wherein the size of the family repeatedly predicted metabolic ability across lineages (Fig. 2). This relationship was positive in 684/976 (70%) of cases, strongly implicating gene family expansion as an engine of convergent metabolic innovation in yeasts. To investigate this engine, we estimated the evolutionary history of the strongest association between a trait (raffinose) and gene family (*SUC*) in our analyses. In *Saccharomyces cerevisiae*, *SUC2* encodes an enzyme that cleaves the glycosidic link in raffinose, producing fructose and melibiose. As expected, we found that gene duplication and horizontal transfer events within the family map closely with trait gains across the yeast phylogeny (Fig. 2).

The extent of convergence on the gene family level also reflects the biochemical properties of the metabolic substrates examined. For example, 8/11 (73%) traits without evidence for convergence concerned carboxylic acids or nitrogenous substrates. By contrast, all alcohols and glucosides had strong evidence of convergence. To further compare patterns of gene family usage across traits, we performed hierarchical clustering of z-scores for each regression, meaning traits that clustered closer together shared more associations with the same gene families (Fig. 2). We found that closely related compounds (e.g., nitrate and nitrite, isomers butane-2,3-diol and propane-1,2-diol, xylose, and its alcohol derivative xylitol) were often reciprocal nearest neighbors. Shared usage of gene families between similar chemical substrates implies that certain families may be repeatedly recruited for the convergent evolution of multiple traits.

Keystone Gene Families Drive Convergent Evolution of Multiple Traits. Our results raise the hypothesis that the same gene families may be repeatedly co-opted in the evolution of multiple convergent traits. To formally test this hypothesis and measure the extent of shared usage, we examined how many traits were significantly associated with each gene family (Fig. 3). As expected, the majority (595/739) of significant gene families were only associated with one specific trait. However, we also identified 144/739 (19%) gene families associated with multiple traits. The repeated recruitment of the same gene family for multiple traits suggests that these families provide fundamental functions

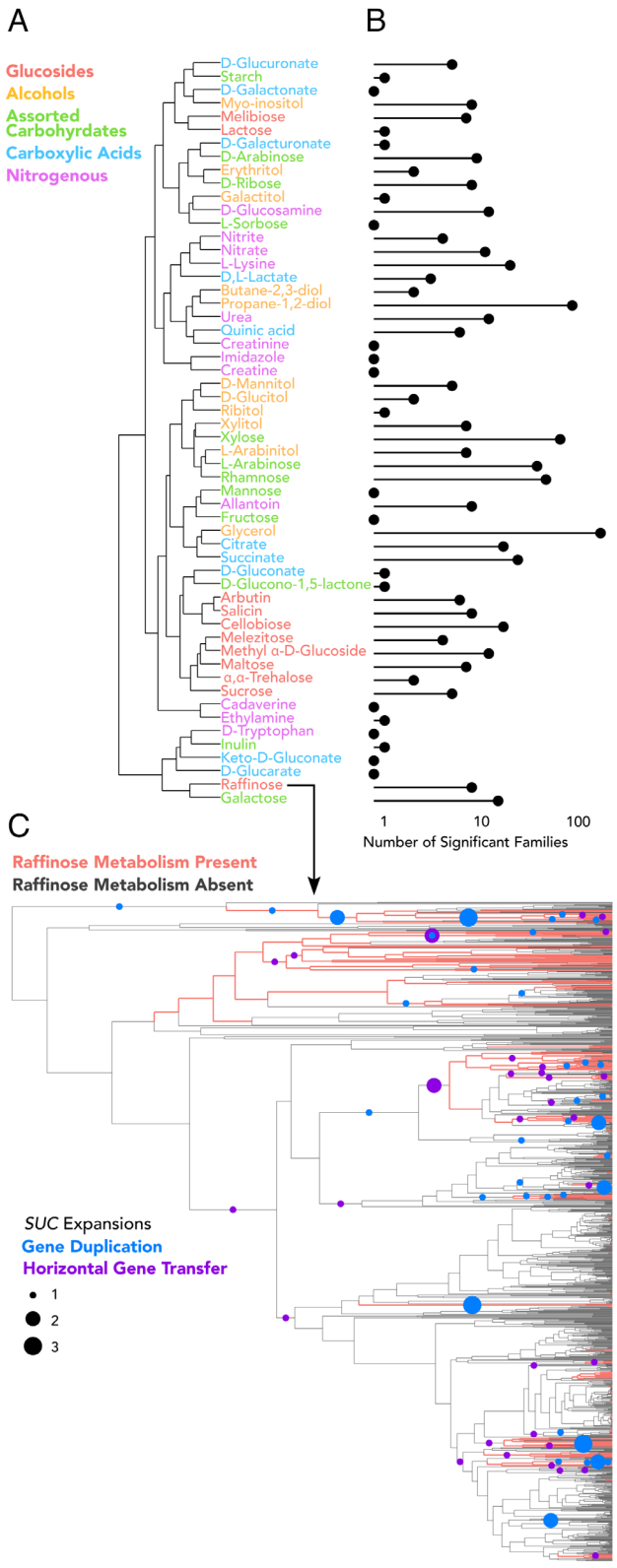

Fig. 2. Gene family expansion and contraction was significantly linked to trait gain and loss in 80% of metabolic traits. (*A*) All 56 metabolic traits examined by this study, clustered by gene family affinity. Traits are clustered by z-scores across families; closer traits share more similar correlations with the same gene families. (*B*) Number of gene families that are significantly associated with each metabolic trait. (*C*) Example of a convergent trait (raffinose metabolism), driven by convergent expansions within the same gene family (*SUC*). The presence of raffinose metabolism is denoted by the red branches and its absence by the black branches. Note that the presence of raffinose metabolism (in red) coincides with gene duplication (blue dots) and horizontal gene transfer (purple dots) events of the *SUC* gene family across the yeast species phylogeny.

intersecting many different metabolic pathways. Just as keystone species demonstrate an outsized impact on their ecological communities (34), these keystone families (defined as those that exhibited significant associations with the convergent evolution of at least 4 different traits) appear to have been instrumental in the evolution of metabolism across a wide variety of chemical substrates.

Of particular note was a gene family encoding α-glucoside transporters (*AGT*), typified by *S. cerevisiae MAL11*, which was significantly associated with the convergent evolution of 13 different traits, almost twice as many as any other family. We found strong evidence that this family has been repeatedly recruited to transport all known Mal11 substrates across the evolution of yeasts, in addition to many more. The size of the *AGT* gene family significantly ($P < 0.025$) predicted metabolic breadth generally across all 56 tested substrates. Other "keystone" families include the oligo-1,6-glucosidase enzymes (typified by *MAL12*) and MAL-activator transcription factors (typified by *MAL13*). These families, along with *AGT*, contain each member of the *MAL* loci responsible for maltose fermentation in *S. cerevisiae*. The observed pattern of gene families accumulating functional affinity as they increase in size is strongly suggestive of neofunctionalization or subfunctionalization (19, 35). In particular, the "Escape from Adaptive Conflict" (EAC) model of subfunctionalization has empirical support in each of these three gene families. EAC describes a scenario wherein all functions of a multifunctional ancestral protein cannot be optimized simultaneously, until duplication frees additional copies to optimize the individual ancestral functions (36–39). Both *MAL11* and *MAL12* homologs are thought to have specialized from promiscuous ancestors with broad but weak affinity for a variety of α-glucoside sugars after several rounds of gene duplication (40, 41). Likewise, the family of associated transcription factors containing *MAL13* has been shown to have specialized its regulatory targets following duplication from a promiscuous ancestor (42). Our results substantiate these scenarios and further indicate EAC as a common pathway to innovation in Saccharomycotina metabolism, acting across many more gene families, lineages, and substrates than previously believed. This view is supported by the fact that all 20 keystone gene families encode transporters, enzymes, or transcription factors (Table 1 and *SI Appendix*, Tables S2 and S3). We hypothesize that these families are characterized by ancestral promiscuity, rendering them amenable to repeated bouts of specialization for additional primary functions following rounds of duplication (43, 44).

Conclusion

Using genomes and metabolic trait data for 56 substrates from 993 Saccharomycotina yeasts, we found that the convergent evolution of metabolic traits was associated with the same gene families in 45/56 (80%) cases, showing widespread evidence of convergence on the gene family level. The incidence of homology in convergent evolution is expected to decrease with divergence time (14, 45). Beyond the species level, examples of deep homology in convergent evolution are restricted to just a few examples, such as the role of *PAX6* in the development of image-forming eyes (2). However, rather than being the exception, we find extensive evidence of deep homology across 400 My of metabolic evolution in Saccharomycotina, which possess roughly as much genetic diversity as the plant and animal kingdoms (22). Gene families are repeatedly recruited to serve parallel functions in the majority of metabolic traits, even across vast evolutionary distances.

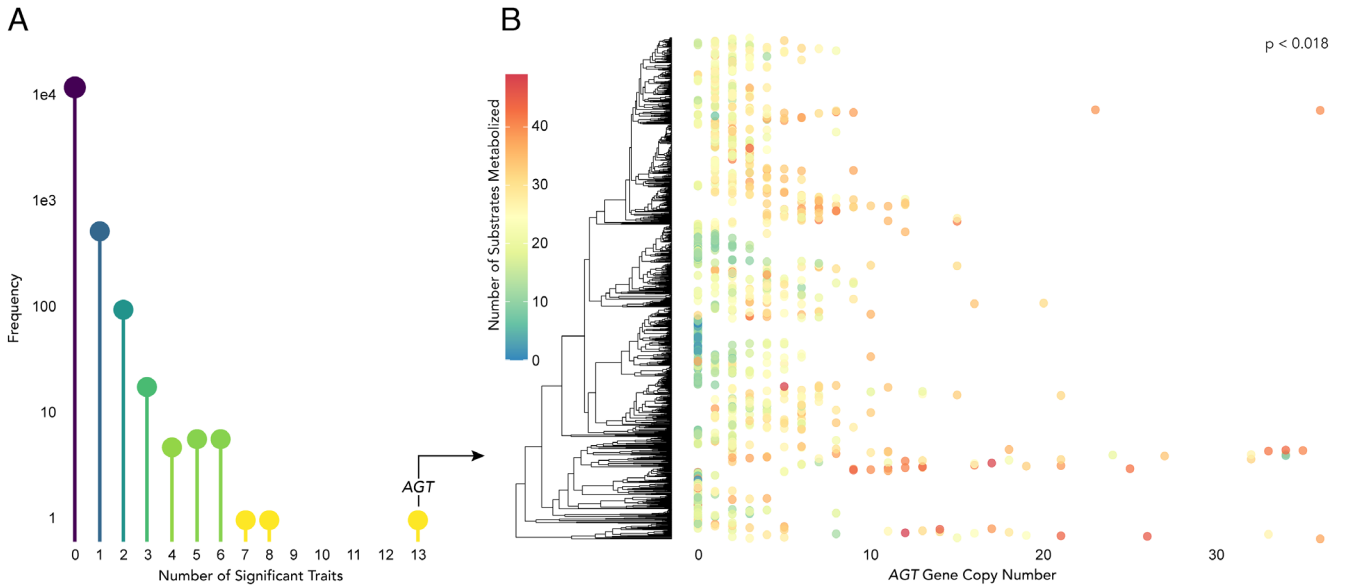

Fig. 3. Keystone gene families influence innovation of multiple metabolic traits. (*A*) Histogram of the number of traits significantly associated with each of 14,785 gene families. Of the 739 families with a significant relationship, most (81%) are associated with just one trait, with the remaining 19% associated with two or more traits. Furthermore, a minority of keystone families show significant effects on several (≥4) traits. (*B*) An extreme example of a keystone gene family encoding *AGT*. Size of the *AGT* gene family mapped onto a species phylogeny, colored by the total number of substrates metabolized by each species. On average, the *AGT* gene family is significantly ($P < 0.018$) larger in generalist species than in specialist species.

Perhaps more remarkably, we also found that gene family evolution sometimes coincided with convergent gains of *multiple* traits. In the most extreme example, we found that over a dozen diverse metabolic traits are seemingly contingent on gene gains and losses within a single family of transporters. We report several additional keystone gene families associated with pleiotropic effects across four or more substrates. All annotated keystone families encode transporters, enzymes, or associated transcription factors, several of which are known to have acquired additional primary functions following duplication events. These results support an EAC model of evolution where multiple traits can be gained through specialization following gene family expansion, particularly for promiscuous proteins. The repeated association between trait gain and expansion within the same families across independent lineages may also suggest a more deterministic view of evolutionary innovation, wherein novel traits are primarily

Table 1. The number of significant traits associated with each keystone gene family

Name	Class	*S. cerevisiae* homolog	Significant traits
General alpha glucoside:H+ symporter	Sugar transporter	*MAL11/MAL31/MPH2*	13
D-xylulose reductase	Oxidoreductase	*SOR1/SOR2/XYL2*	8
Endoglucanase	Hydrolase		7
Sugar:H+ symporter	Sugar transporter	*STL1*	6
Glucose/mannose:H+ symporter	Sugar transporter	*BSC6*	6
Oligo-1,6-glucosidase	Hydrolase	*MAL12/MAL32/IMA1*	6
L-fuconate dehydratase	Lyase		6
Sorbose reductase	Oxidoreductase	*OAR1*	6
Putative transcription factor	Transcription factor		6
Putative transcription factor	Transcription factor		5
MAL-activator	Transcription factor	*MAL13/MAL31/ZNF1*	5
MATE family	Solute carrier	*ERC1*	5
Beta-glucosidase	Hydrolase		5
Beta-mannosidase	Hydrolase		5
2-keto-3-deoxy-L-rhamnonate aldolase	Lyase		5
L-rhamnonate dehydratase	Lyase		4
L-rhamnono-1,4-lactonase	Hydrolase		4
Sarcosine oxidase	Oxidoreductase		4
Cysteine synthase	Transferase		4
Alpha-D-xyloside xylohydrolase	Hydrolase		4

See *SI Appendix*, Table S2 for more details and *SI Appendix*, Table S3 for all *S. cerevisiae* and *C. albicans* keystone genes.

acquired through a narrow set of possible genetic elements and mechanisms (2, 5, 46).

Finally, we identify hundreds of gene families that are significantly linked to novel metabolic ability across Saccharomycotina, offering putative targets for genetic engineering. Identifying shared mechanisms of metabolic innovation is of particular interest for this clade, whose metabolisms are harnessed by humans for a slew of purposes across medical, scientific, and industrial fields (29).

Methods

Genomic Dataset. Genomes and the species phylogeny were obtained from Opulente et al. (27). Briefly, 1,154 yeast taxa were paired-end sequenced on an Illumina HiSeq 2500, and genes were predicted with the BRAKER (47) v2.1.6 pipeline. After aligning, a species phylogeny was inferred with IQ-TREE (48) v2.0.7 using the general amino acid substitution matrix (49) with four gamma discretized rate categories. The phylogeny was then time-calibrated using the RelTime method implemented in MEGA7 (50) using calibration points from Shen et al. (22). Additional information, as well as all genomes, alignments, and phylogenies can be found at the original publication (27). OrthoFinder (51) v3.0 was run under default parameters on the 1,154 genomes, resulting in 72,381 gene families (18) which were then filtered to 14,785 to include only gene families with at least 10 taxa represented. Of these, 338 (2%) were conserved across all genomes with the average gene family occurring in 356 (30%) of taxa. The average size of a gene family spanned 4 orders of magnitude from 0.0087 to 10.2 homologs/taxon. On average, a gene family had 0.38 homologs per taxon (*SI Appendix*, Fig. S2). Following Opulente et al. (27), gene models were annotated with KEGG orthologs (52) using KofamScan (53). Keystone gene families were annotated from the most common KO term found in each family, assuming that term was represented in at least 5% of genes. Of the remaining three families that did not meet these criteria, one contained known transcription factors regulating α-glucoside metabolism in *S. cerevisiae* and *Candida albicans*. For the final two families, we used InterProScan (54) v5.72 to annotate protein domains of each member, from which we derived a consensus prediction of protein function. If no KO terms were found, annotations were instead taken from *S. cerevisiae* homologs if available. Enrichment analysis was performed using the enrichKEGG() function in the R package clusterProfiler (55) v4.10. Five significant metabolic pathways were identified, including galactose and fructose + mannose. These pathways are very well characterized in yeasts and represent some of the only substrates with their own annotated KEGG pathways. We interpret their significance to be indicative of the broad convergence observed across most sugars, rather than any unique quality of these particular pathways.

Trait Dataset. Metabolic traits for each species were sourced from the Westerdijk Fungal Biodiversity Institute as reported in Harrison et al. (28). These data were supplemented with experimental assays from Opulente et al. (27) wherein growth rates were obtained for 853 yeast species across 24 carbon and nitrogen substrates. The quantitative results of Opulente et al. (27) were binarized for this study. Metabolic traits were considered present for a given species if any growth was observed on 96-well plates containing a given carbon or nitrogen source in minimal media, and absent if it was not. If there was disagreement between the two data sources, preference was given to Opulente et al. (27). As none of the 56 metabolic traits exhibited identical presence/absence patterns across all taxa, each was considered a unique trait and analyzed separately. Traits and genomes were merged based on the most recently available taxonomy (56) hosted on the MycoBank database (57). Then, 161 of the 1,154 taxa analyzed by Opulente et al. (27) had no metabolic data associated with them and were excluded from our analysis. Traits were retained if they had no more than 50% missing data across the remaining 993 taxa. Two additional metabolic traits were further removed: growth on glucose, which was present in every species, and growth in the absence of carbon, which is not directly comparable with positive substrate-specific metabolic traits. The final trait matrix had 19% missing data across 46 carbon and 10 nitrogen substrates (*SI Appendix*, Fig. S1). Evolution of each trait was modeled using a stochastic character mapping approach implemented in RevBayes (58) v1.2.2 using two hidden states (59). Models were run for 1,000 generations across two chains.

Phylogenetic Analyses. To assess the effect of gene family size on trait evolution, we used a phylogenetic logistic regression as implemented in the phylolm (60) v2.6.2 R (61) v4.3.2 package using the maximum penalized likelihood estimation method. A square root transformation was applied to gene family size to reflect the expected diminishing contribution of individual genes at high dosage in large families. This package was also used to test the relationship between metabolic breadth and *AGT* family size (Fig. 3*B*) using a phylogenetic linear regression.

When studying patterns of convergence, care must be taken to distinguish truly independent events from synapomorphies shared by common descent (7). Conventional phylogenetic comparative methods have been shown to have trouble distinguishing between these scenarios, attributing significance for spurious correlations even for single unreplicated events (26, 62). It was recently shown that many approaches in phylogenetic comparative methods and statistical genetics represent special cases of the same general model (63). Leveraging this discovery, we adapt a common strategy in genome-wide association studies shown to reduce these types of spurious correlations (26, 63) by including eigenvectors of the phylogenetic variance–covariance matrix in our model. We selected the first two leading eigenvectors, which together contain over 50% of shared variance across the phylogeny. As expected, these eigenvectors explain the greatest variance in traits with the fewest number of transitions, supporting the idea that they reduce false positives resulting from shared ancestral events (*SI Appendix*, Fig. S3). These eigenvectors were included as fixed effects in the traditional phylogenetic logistic regression model. To further reduce the possibility of false positives and correct for multiple tests, we calculated the false discovery rate (64) adopting a conservative alpha value of 10^{-6}. Evolution of the *SUC* family was estimated with GeneRax (65) v2.0.4, a species–gene tree reconciliation program, using a maximum subtree prune and regraft distance of 3. The starting gene tree was inferred using IQ-TREE (48) v2.2.2 and both initial and reconciled gene trees used 10 FreeRate (66, 67) categories while also allowing for invariant sites, the model parameters with the highest Bayesian information criterion according to ModelFinder (68).

Data, Materials, and Software Availability. Data have been deposited in Figshare and GitHub. The data are available at [Figshare (https://doi.org/10.6084/m9.figshare.26440963) (69) and Github (https://github.com/KyleTDavid/YeastConvergence2025)(70)][data (https://doi.org/10.6084/m9.figshare.26440963) and code (https://github.com/KyleTDavid/YeastConvergence2024)].

ACKNOWLEDGMENTS. This work was performed using resources contained within the Advanced Computing Center for Research and Education at Vanderbilt University in Nashville, TN. This work was supported by the NSF (grants DBI-2305612 to K.T.D., DGE-1747503 to J.G.C. DEB-2110403 to C.T.H., and DEB-2110404 to A.R.) and the NIH (grant T32GM007133 to J.G.C. and R35GM151348 to M.P.). X.-X.S. was supported by the NSF for Distinguished Young Scholars of Zhejiang Province (LR23C140001), the Fundamental Research Funds for the Central Universities (226-2023-00021), and the key research project of Zhejiang Lab (2021PE0AC04). Research in the Hittinger Lab is also supported by the United States Department of Agriculture National Institute of Food and Agriculture (Hatch Projects 1020204 and 7005101), in part by the Department of Energy (DOE) Great Lakes Bioenergy Research Center (DOE Biological and Environmental Research Office of Science DE–SC0018409), and an H.I. Romnes Faculty Fellowship (Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation). Research in the Rokas Lab is also supported by the NIH/National Institute of Allergy and Infectious Diseases (R01 AI153356) and the Burroughs Wellcome Fund.

Author affiliations: [a]Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235; [b]Evolutionary Studies Initiative, Vanderbilt University, Nashville, TN 37235; [c]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089; [d]Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089; [e]Laboratory of Genetics, James Franklin Crow Institute for the Study of Evolution, Center for Genomic Science Innovation, Department of Energy Great Lakes Bioenergy Research Center, Wisconsin Energy Institute, University of Wisconsin-Madison, Madison, WI 53726; [f]Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223; [g]Department of Biology, Villanova University, Villanova, PA 19085; [h]Guangdong Laboratory for Lingnan Modern Agriculture, Guangdong Province Key Laboratory of Microbial Signals and Disease Control, Integrative Microbiology Research Center, South China Agricultural University, Guangzhou 510642, China; [i]Key Laboratory of Biology of Crop Pathogens and Insects of Zhejiang Province, Institute of Insect Sciences, Zhejiang University, Hangzhou 310058, China; [j]Westerdijk Fungal Biodiversity Institute, Utrecht 3584, The Netherlands; and [k]Department of Computational Biology, Cornell University, Ithaca, NY 14853

1. S. D. Smith, M. W. Pennell, C. W. Dunn, S. V. Edwards, Phylogenetics is the new genetics (for most of biodiversity). *Trends Ecol. Evol.* **35**, 415–425 (2020).
2. T. Oakley, Building, maintaining, and (re-) deploying genetic toolkits during convergent evolution. *Integr. Comp. Biol.* **64**, 1505–1512 (2024).
3. T. B. Sackton *et al.*, Convergent regulatory evolution and loss of flight in paleognathous birds. *Science* **364**, 74–78 (2019).
4. J. B. Losos, Convergence, adaptation, and constraint. *Evolution* **65**, 1827–1840 (2011).
5. D. B. Wake, M. H. Wake, C. D. Specht, Homoplasy: From detecting pattern to determining process and mechanism of evolution. *Science* **331**, 1032–1035 (2011).
6. M. F. Land, R. D. Fernald, The evolution of eyes. *Annu. Rev. Neurosci.* **15**, 1–29 (1992).
7. G. R. McGhee, *Convergent Evolution: Limited Forms Most Beautiful* (MIT press, 2011). https://books.google.com/books?hl=en&lr=&id=QwDSr1qdqXUC&oi=fnd&pg=PR7&dq=Convergent+Evolution+:+Limited+Forms+Most+Beautiful&ots=El89QCLmtu&sig=CsJauXowwo55WhmiZktyi-iGr8k.
8. S. I. Tomarev *et al.*, Squid Pax-6 and eye development. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 2421–2426 (1997).
9. G. Halder, P. Callaerts, W. J. Gehring, Induction of Ectopic Eyes by Targeted Expression of the *eyeless* Gene in *Drosophila*. *Science* **267**, 1788–1792 (1995).
10. N. Rives, V. Lamba, C.-H.C. Cheng, X. Zhuang, Diverse origins of near-identical antifreeze proteins in unrelated fish lineages provide insights into evolutionary mechanisms of new gene birth and protein sequence convergence. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.03.12.584730 (Accessed 1 August 2024).
11. D. L. Stern, The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764 (2013).
12. X. Bailly, S. Vanin, C. Chabasse, K. Mizuguchi, S. N. Vinogradov, A phylogenomic profile of hemerythrins, the nonheme diiron binding respiratory proteins. *BMC Evol. Biol.* **8**, 244 (2008).
13. C. J. Coates, H. Decker, Immunological properties of oxygen-transport proteins: Hemoglobin, hemocyanin and hemerythrin. *Cell. Mol. Life Sci.* **74**, 293–317 (2017).
14. M. Bohutínská, C. L. Peichel, Divergence time shapes gene reuse during repeated adaptation. *Trends Ecol. Evol.* **39**, 396–407 (2024).
15. P. F. Colosimo *et al.*, Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science* **307**, 1928–1933 (2005).
16. V. Soria-Carrasco *et al.*, Stick insect genomes reveal natural selection's role in parallel speciation. *Science* **344**, 738–742 (2014).
17. S. Ohno, *Evolution by Gene Duplication* (Springer Science & Business Media, 1970).
18. B. Feng *et al.*, Unique trajectory of gene family evolution from genomic analysis of nearly all known species in an ancient yeast lineage. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.06.05.597512 (Accessed 1 August 2024).
19. E. Kuzmin, J. S. Taylor, C. Boone, Retention of duplicated genes in evolution. *Trends Genet.* **38**, 59–72 (2022).
20. S. D. Copley, Evolution of new enzymes by gene duplication and divergence. *FEBS J.* **287**, 1262–1283 (2020).
21. R. Albalat, C. Cañestro, Evolution by gene loss. *Nat. Rev. Genet.* **17**, 379–391 (2016).
22. X.-X. Shen *et al.*, Tempo and mode of genome evolution in the budding yeast subphylum. *Cell* **175**, 1533–1545.e20 (2018).
23. S. M. Soucy, J. Huang, J. P. Gogarten, Horizontal gene transfer: Building the web of life. *Nat. Rev. Genet.* **16**, 472–482 (2015).
24. J. V. Etten, D. Bhattacharya, Horizontal gene transfer in eukaryotes: Not if, but how much? *Trends Genet.* **36**, 915–925 (2020).
25. G. J. Vermeij, Historical contingency and the purported uniqueness of evolutionary innovations. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 1804–1809 (2006).
26. J. C. Uyeda, R. Zenil-Ferguson, M. W. Pennell, Rethinking phylogenetic comparative methods. *Syst. Biol.* **67**, 1091–1109 (2018).
27. D. A. Opulente *et al.*, Genomic factors shape carbon and nitrogen metabolic niche breadth across Saccharomycotina yeasts. *Science* **384**, eadj4503 (2024).
28. M.-C. Harrison *et al.*, Machine learning enables identification of an alternative yeast galactose utilization pathway. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2315314121 (2024).
29. C. P. Kurtzman, J. W. Fell, T. Boekhout, *The Yeasts: A Taxonomic Study* (Elsevier, 2011).
30. K. T. David *et al.*, Saccharomycotina yeasts defy long-standing macroecological patterns. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2316031121 (2024).
31. M. Spagnuolo, A. Yaguchi, M. Blenner, Oleaginous yeast for biofuel and oleochemical production. *Curr. Opin. Biotechnol.* **57**, 73–81 (2019).
32. T. Gassler *et al.*, The industrial yeast Pichia pastoris is converted from a heterotroph into an autotroph capable of growth on CO_2. *Nat. Biotechnol.* **38**, 210–216 (2020).
33. P. Srinivasan, C. D. Smolke, Biosynthesis of medicinal tropane alkaloids in yeast. *Nature* **585**, 614–619 (2020).
34. R. T. Paine, A note on trophic complexity and community stability. *Am. Nat.* **103**, 91–93 (1969).
35. M. W. Hahn, Distinguishing among evolutionary models for the maintenance of gene duplicates. *J. Hered.* **100**, 605–617 (2009).
36. M. Lynch, V. Katju, The altered evolutionary trajectories of gene duplicates. *Trends Genet.* **20**, 544–549 (2004).
37. X. He, J. Zhang, Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**, 1157–1164 (2005).
38. C. T. Hittinger, S. B. Carroll, Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* **449**, 677–681 (2007).
39. D. L. Des Marais, M. D. Rausher, Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* **454**, 762–765 (2008).
40. J. G. Crandall, X. Zhou, A. Rokas, C. T. Hittinger, Specialization restricts the evolutionary paths available to yeast sugar transporters. *Mol. Biol. Evol.* **41**, msae228 (2024).
41. K. Voordeckers *et al.*, Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biol.* **10**, e1001446 (2012).
42. K. Pougach *et al.*, Duplication of a promiscuous transcription factor drives the emergence of a new regulatory network. *Nat. Commun.* **5**, 4868 (2014).
43. P. J. O'Brien, D. Herschlag, Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.* **6**, R91–R105 (1999).
44. R. A. Jensen, Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409–425 (1976).
45. S. Yeaman *et al.*, Convergent local adaptation to climate in distantly related conifers. *Science* **353**, 1431–1433 (2016).
46. J. B. Losos, T. R. Jackman, A. Larson, K. de Queiroz, L. Rodriguez-Schettino, Contingency and determinism in replicated adaptive radiations of island lizards. *Science* **279**, 2115–2118 (1998).
47. T. Brůna, K. J. Hoff, A. Lomsadze, M. Stanke, M. Borodovsky, BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinf.* **3**, lqaa108 (2021).
48. L.-T. Nguyen, H. A. Schmidt, A. Von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
49. S. Q. Le, O. Gascuel, An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
50. S. Kumar, G. Stecher, K. Tamura, MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
51. D. M. Emms, S. Kelly, OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14 (2019).
52. M. Kanehisa, S. Goto, KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
53. T. Aramaki *et al.*, KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
54. P. Jones *et al.*, InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
55. T. Wu *et al.*, ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* **2** (2021).
56. M. Groenewald *et al.*, A genome-informed higher rank classification of the biotechnologically important fungal subphylum Saccharomycotina. *Stud. Mycol.* **105**, 1–22 (2023).
57. P. W. Crous, W. Gams, J. A. Stalpers, V. Robert, G. Stegehuis, MycoBank: An online initiative to launch mycology into the 21st century. *Stud. Mycol.* **50**, 19–22 (2004).
58. S. Höhna *et al.*, RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol.* **65**, 726–736 (2016).
59. J. M. Beaulieu, B. C. O'Meara, M. J. Donoghue, Identifying hidden rate changes in the evolution of a binary morphological character: The evolution of plant habit in campanulid angiosperms. *Syst. Biol.* **62**, 725–737 (2013).
60. L. S. T. Ho *et al.*, Package 'phylolm'. See http://cran.r-project.org/web/packages/phylolm/index.html (Accessed February 2018) (2016). https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=51bf6041cf997bbcfe1db63d6bcc40a33c01fee6.
61. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2017), https://www.R-project.org/.
62. W. P. Maddison, R. G. FitzJohn, The unsolved challenge to phylogenetic correlation tests for categorical characters. *Syst. Biol.* **64**, 127–136 (2015).
63. J. G. Schraiber, M. D. Edge, M. Pennell, Unifying approaches from statistical genetics and phylogenetics for mapping phenotypes in structured populations. *PLoS biology* **22,** e3002847 (2024).
64. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.: Series B (Methodol.)* **57**, 289–300 (1995).
65. B. Morel, A. M. Kozlov, A. Stamatakis, G. J. SzöllHosi, GeneRax: A tool for species-tree-aware maximum likelihood-based gene family tree inference under gene duplication, transfer, and loss. *Mol. Biol. Evol.* **37**, 2763–2774 (2020).
66. Z. Yang, A space-time process model for the evolution of DNA sequences. *Genetics* **139**, 993–1005 (1995).
67. J. Soubrier *et al.*, The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Mol. Biol. Evol.* **29**, 3345–3358 (2012).
68. S. Kalyaanamoorthy, B. Q. Minh, T. K. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587 (2017).
69. K. T. David, Data from "Convergent expansions of keystone *gene* families drive metabolic innovation in Saccharomycotina yeasts." Figshare. doi.org/10.6084/m9.figshare.26440963. Deposited 17 April 2025.
70. K. T. David, YeastConvergence2025. GitHub. https://github.com/KyleTDavid/YeastConvergence2025. Deposited 1 August 2024.