

The background of the entire page is a dark blue field densely populated with small, semi-transparent dots in various colors including purple, pink, white, and yellow. Faint, thin white lines are scattered across the field, some connecting clusters of dots, creating a network-like or starry pattern.

Species-wide genetic control of transcript and protein abundance across multiple environments in yeast

Gauthier Brach

HaploTeam · IGBMC UMR 7104 – UMR-S 1258
Université de Strasbourg · 2026



Abstract

Genetic variation shapes phenotypes through molecular layers, yet association studies often focus on common variants, gene-level RNA abundance and one environment. Using the natural diversity of the baker's yeast *Saccharomyces cerevisiae*, this thesis examines how structural variants (SVs), transcript structure and environmental context shape molecular phenotypes. A 26-parent half-diallel looked at the influence of low-frequency variants and SVs on gene expression variation. Oxford Nanopore direct RNA sequencing of six isolates revealed variation in transcript-end use and poly(A)-tail length beyond gene-level abundance. Matched transcriptomes and proteomes from more than 550 isolates across four environments showed genotype-by-environment interactions and largely condition-specific QTL. Transcript effects retained at protein level were concordant, but their transmission varied among environments. Together, this work tackles hidden dimensions of molecular variation and shows that understanding the genotype–phenotype relationship requires integrating variant classes, transcript structures, molecular layers and environmental context.

Résumé

La variation génétique façonne les phénotypes à travers différentes couches moléculaires, mais les études d'association se concentrent souvent sur les variants communs, l'abondance des ARN à l'échelle des gènes et un seul environnement. En exploitant la diversité naturelle de la levure de boulanger *Saccharomyces cerevisiae*, cette thèse examine comment les variants structurels (SV), la structure des transcrits et le contexte environnemental façonnent les phénotypes moléculaires. Un demi-diallèle de 26 parents a permis d'étudier l'influence des variants de faible fréquence et des SV sur la variation de l'expression génique. Le séquençage direct des ARN par Oxford Nanopore dans six isolats a révélé une variation des extrémités des transcrits et de la longueur des queues poly(A). Des transcriptomes et protéomes pour plus de 550 isolats dans quatre environnements ont mis en évidence des interactions génotype–environnement et des QTL en grande partie spécifiques aux conditions. Les effets observés sur les transcrits qui se maintenaient au niveau protéique étaient concordants, mais leur transmission variait selon l'environnement. Ensemble, ces travaux abordent des dimensions cachées de la variation moléculaire et montrent que comprendre la relation génotype–phénotype nécessite d'intégrer les classes de variants, la structure des transcrits, les couches moléculaires et le contexte environnemental.

Acknowledgments

This work was conducted primarily in the Molecular Genetics, Genomics, and Microbiology Unit (GMGM, UMR 7156 – CNRS) at the University of Strasbourg, and later at the Institute of Genetics and Molecular and Cellular Biology (IGBMC, UMR 7104 – UMR-S 1258 – CNRS, Inserm), which became the new home for the team in early 2026, under the supervision of Prof. Joseph Schacherer.

I would first like to thank the members of my thesis committee for agreeing to review my work and being among the very few who will ever actually read this manuscript: Prof. **Warren Albertin**, Prof. **Bernard Dujon**, Dr. **Susana Coelho** and Prof. **Carine Meignin**. Thank you for your time. I would also like to extend special thanks to Susana and Bernard, who also served on my mid-thesis committees, for their valuable insights and guidance, which helped steer this project along the way.

Joseph, I am really grateful to you for giving me the opportunity to join the lab six years ago. I rather haphazardly found my way here at a time when I was not really sure what I wanted to do, and I could easily have made the wrong choice and tried my hand at structural biology, but thankfully I was lured instead by the sweet sirens of genetics! It has clearly been a long and arduous journey, and the struggle has been intense over these years. I am still debating whether the low point was when I was coughing up my lungs after running the Ekiden or the despair of the dRNA-seq paper (which still gives me nightmares)... But this was balanced by so many wonderful moments across the years, and I am deeply thankful to have been able to learn so much from you.

I am grateful to everyone in the HaploTeam for your support over the years; navigating this journey would have been far more difficult without you. A special thanks to those who had trouble getting out of bed in time for the lab meetings, your sacrifices were appreciated. **Anne**, thank you for igniting a passion for sequencing and bioinformatics maaaaany years ago now, and for your endless patience with my confusing questions; your daily fuel of Bonne Maman cookies for the 228 kept me going (I am fairly certain we single-handedly kept their production line in business). **Jing**, your ability to show the artistry within science is something I deeply admire; I can only hope to one day approach your incredible intuition and depth of knowledge; you truly possess the ability to turn data into gold seemingly effortlessly, I hope you manage to keep running enough to get your VO2 max to the high tier. **Isabel**, thank you for our discussions and laughs; I promise to keep pretending I understand those papers with four-page equations on the role of variable dominance. Thank you for your constant help with awk and always pushing me to be better. **Claudia**, I really appreciated your constant support, your battles against the service financier to ensure we had everything on time, while carrying an unimaginably heavy burden on your side. **Claudine**, thank you for your advice and tips throughout the years.

To the Slackers who were part of the ride: **Arthur** for your encouragement right at the beginning, **Jan** for teaching me about sketchy port forwarding on clusters, and **Abhishek** for your deep knowledge of library troubleshooting. **Élodie**, your encouraging words when I was only starting meant a lot, and I'm glad to have been able to expand on your work. **Lucie**, I owe you so much for your help in general, always having to order even more stuff on a whim. Thank you for all the good moments, **Yasmine**. Thank you **Jessica** for feeding me yoghurts all these months. My gratitude also extends to those I had less time to interact with: **Emna**, **Sam**, **Théophile**, **Fabien**, **Gaëlle**, **Mathias**, **Coline**, **Ziyang**, **Mathieu**, **Julie**. Thanks **Elie** for the tips, the bioRxiv papers, and encouragement along the way. **Ana**, slacker fugaz, gracias por tu apoyo y por el mejor jamón que he probado en mi vida. Thank you **Axelle** for your cheerfulness and for your spoon every day. Thank you, **Marina**, for your smiles, the 00 flour and tomato sauce, and the 007 operation. Thanks to the unofficial slackers of the lab, the PIXL, ROTOR and epMotion96xl robots, for enduring my abuse and surgery sessions to fix you; you were of precious help, but I could have done without your tantrums during long weekend nights.

I also want to extend my gratitude to the other great people who had an influence during these years. Tack, **Emelie**, för att du har funnits där för mig genom åren, trots avståndet. **Enrico**, cheers to you for showing me just how loud someone can be, the beers (red please...), and the canards laqués. **Lucille** and **Carmen2**, our office discussions were always a joy, as well as the cookies exchanges. **Éléonore**, I am really grateful for your support. **Solène**, your moral support was precious, thank you. **Sandrine**, you're a wild horse, hard to tame but lovely once you open up. ¿Cuándo volvemos a hacer tacos? Thank you. **Marie**, I truly appreciated your kind words over the years. Finally, **Cindy**, your warm welcome at the IGBMC made a real difference. Te espero en casa a comer, que yo cocino la bandeja paisa (Sandrine también, btw). Thank you for everything.

Victor, I cannot overstate my appreciation for your patience in indulging my incessant questions and for showing me the kind of bioinformatician I should aspire to be. Thanks for taking the time to show me the ropes of GWAS. You're one of a kind.

Special thanks to **Andreas** for believing in me back in my first year of Master's, for teaching me basically everything, from the benchwork to the libraries, basic bioinformatics, R, for successfully converting me to the tidyverse, and for showing me the ropes of the robots, I couldn't have had a better teacher. The tetrad is now complete and the curse lifted!

Marion (Fig. 0D), your incredible help at the bench was a lifesaver; it would have taken so much time I didn't have if we had to replace your machine-like pipetting with my sloppy wetlab skills. You helped deliver this four-year-in-the-making baby, while managing to grow one in barely nine months! As I write these words, the name is still unknown, but I'm certain it's Cléa. 🧙 Last second edit... One letter off. Never doubt my powers. 🧙 Welcome, Cléo, eleven days of fight and light... ❤️

Pia (Fig. 0A), your constant support and your smiles were proof that sometimes an immovable optimist is exactly what an unstoppable pessimist needs. I still don't get how it is even possible to be so consistently positive. You kept the office bright on the days I made it grey. It has been so nice having you by my side for such a long time, I'm going to miss you (but I won't miss eating outside, swarmed by insects...)

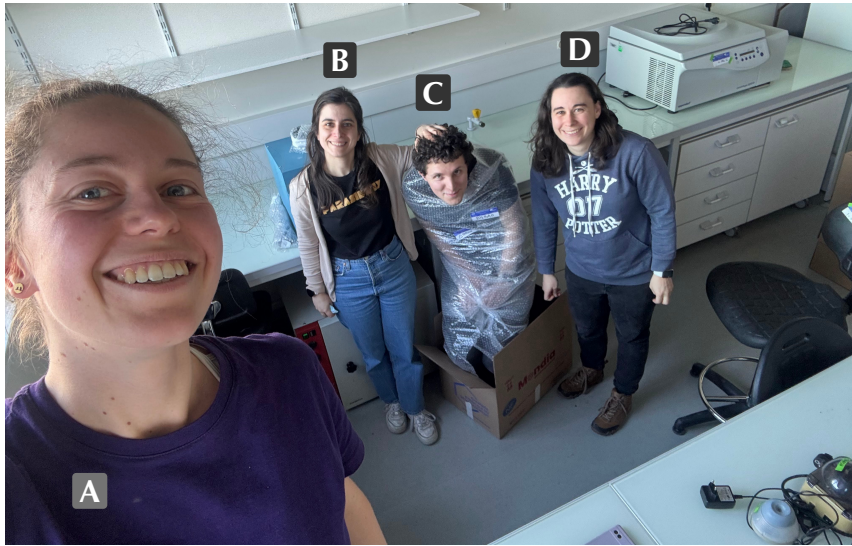


Figure 0 | The 228 Slackers are moving (being shipped) to the IGBMC!

(A) Pia, acquiring photographic evidence of the experiment. (B) Carmen, applying manual restraint to the (fragile) specimen. (C) The author, packed in bubble wrap and loaded into a standard cardboard container, following the packaging protocol of (Becerra-Rodríguez, Garin, and Thiele, 2026). (D) Marion, questioning the experimental robustness with $n = 1$.

Amélie, thank you so much for your support and friendship over the years. It is crazy how fast time has passed; it really went by in a blink. It feels like yesterday when we were thrown together in those genetics (poetically, almost) practicals, having no idea what we were doing, and here we are now. Thank you for sharing the stress and always having my back, from the beginning all the way to the very last page.

Carmen (Fig. 0B), words will truly fall short of expressing how important you were in helping me through this journey. From half-drunkenly getting yourself invited to my wedding, to being the best support and bike companion (+2 morale points when not getting run over) I could have hoped for in the trenches. There were countless moments when I felt overwhelmed and discouraged, but your unwavering support, pragmatic advice, and encouragements really helped. Brutally direct (too much?), but I needed that more than I am willing or able to admit; definitely my dear chef, but I still think it clearly would have needed more salt. I will forever be grateful to you for pushing me when I needed it (into an Uber, more than once), for always believing in me even when my belief faltered, and for not letting me quit. Multiple times, on multiple benches, both of us... You helped turn this from a disaster into a merely excruciating task, and for that, you have my everlasting appreciation. For better or for worse, this probably wouldn't exist without you. I'm glad to be able to call you my friend, and you know how difficult this can be for me. I (Fig. 0C) hope I was able to bring something to you in return. Ojalá estés orgullosa de mí. Gracias por estar siempre ahí.

Valentina, for you who had to endure all the worst parts of this PhD, even more than me if I'm being honest. You carried so much of the weight, mostly unknowingly. I could not have done this without you, nor would I have wanted to. From the early mornings to the long nights, what a journey it has been... and somehow, this is just the beginning.

Table of Contents

Introduction	1
Introduction	2
PART I. Building the yeast genomic reference	3
The first eukaryotic genome (1996-2000)	3
From sequence to function	7
PART II. Discovering that S288C is not the species	13
Early population genomics	13
Scaling to thousands of genomes	17
PART III. The pangenome era	25
Long reads and structural variation	25
Assembling the species pangenome	26
PART IV. From variation to mapping and function	34
Conceptual foundations of genotype–phenotype mapping	34
What are traits, really?	35
Measuring biological layers	40
Genetic mapping strategies	43
Toward a species-wide functional toolkit	57
Outlook	58
Overview of the project	83
1 Diallel panel reveals a significant impact of low-frequency genetic variants on gene expression variation in yeast	89
Introduction	90
Results	92
Diallel design and transcript abundance variation	92
Detection of the SNP-eQTL using the diallel panel	92
Impact of the low-frequency SNPs on gene expression variation	95

Characterization of the structural variants in the diallel panel	97
Impact of SVs on gene expression phenotypes	98
Discussion	100
Methods	101
Diallel panel generation and cell harvesting	101
cDNA library preparation	102
mRNA abundance quantification	103
Functional annotation of the SNP variants	103
Long-reads sequencing of the parental genomes	103
Genome assembly and structural variant detection	104
Generation of the genotype matrix	105
Genome-wide heritability estimation	105
Genome-wide association	105
Analysis of low frequency variants	106
Supplementary figures	106
Acknowledgements	110
Author contributions	110
Data availability	110

2 Natural variation in alternative polyadenylation across *Saccharomyces cerevisiae* isolates 117

Introduction	118
Results	120
Direct RNA sequencing generates a high-quality transcriptomic atlas of six natural isolates	120
Gene expression is broadly conserved but retains strain-specific signatures	121
Poly(A) tail length is gene-specific and largely conserved across strains .	121
A robust map of alternative polyadenylation across natural isolates . . .	123
APA architecture is only partially conserved across natural isolates . . .	126
APA generates substantial isoform diversity, including a limited set of novel transcripts	128
Poly(A) tail length and APA are associated in gene-specific and strain-associated patterns	131

Matched APA usage differs quantitatively among isolates	132
Discussion	136
Methods	138
Yeast strains, growth and RNA extraction	138
Direct RNA sequencing and basecalling	138
Read correction, alignment and gene expression	138
Depth equalization	139
Transcript-end detection and polyadenylation islands	139
Transcript-end validation	140
Isoform quantification and classification	140
Poly(A)-tail analyses	141
Cross-strain transcript-end comparisons	142
Differential isoform usage	142
Coupling between polyadenylation site and tail length	143
Strain relationships and statistical analysis	143
Data availability	143
Supplementary figures	144

3

Environmental context shapes the cross-omic consequences of natural genetic variation in yeast

155

Introduction	156
Results	158
A population-scale, cross-environment, matched multi-omic resource .	158
Environmental and population structure shape the multi-omic landscape	162
Environmental responses are coordinated but lineage dependent	166
Regulatory architecture is restricted by environment and molecular layer	170
Cross-omic transmission of genetic effects is incomplete and context dependent	175
A <i>GAL2</i> -linked allele marks a cross-omic vacuolar response	176
Discussion	181
Methods	184
Isolate panel and environmental conditions	184

Preculture and growth monitoring	184
Sample harvesting	184
RNA extraction, library preparation and sequencing	185
Sample identity verification and contamination screening	185
Pangenome-aware RNA quantification	185
Proteomics acquisition and processing	186
Growth phenotypes	186
Global expression maps and programme scores	187
Multi-omics factor analysis	187
Functional enrichment	188
Environmental responses and lineage differential expression	188
Global and clade-resolved GxE	189
Static eQTL and pQTL mapping	189
Cross-environment and cross-layer QTL comparisons	191
Direct cross-omic transmission analysis	191
<i>GAL2</i> locus and response analyses	192
Statistical analysis	193
Supplementary figures	194
Data and code availability	220
Author contributions	220
Competing interests	220

Discussion and perspectives **227**

From genotype to phenotype through molecular layers	228
Looking back at the three angles	229
1. Effects of structural and low-frequency variants on the transcriptome .	229
2. The transcript layer at full length	230
3. Matched transcriptomes and proteomes across four environments . .	231
Going further	232
A larger diallel: sequencing hybrids rather than parents	232
Looking at the transcriptome at single-cell resolution	234

Limitations of linear models for deciphering the genotype–phenotype relationship	238
--	-----

Concluding remarks	242
---------------------------	------------

Appendices	249
-------------------	------------

PART I. FaST-ER-LMM: a GPU FaST-LMM for yeast-scale GWAS	250
--	-----

Introduction	251
------------------------	-----

FaST-LMM	254
--------------------	-----

FaST-ER-LMM: making it faster on GPUs	255
---	-----

Why the reference implementation is too slow for our needs	255
--	-----

Porting the association calculations to GPUs	255
--	-----

Running a scan	256
--------------------------	-----

Sharing one eigendecomposition across phenotypes	257
--	-----

Using more than one GPU	258
-----------------------------------	-----

Results	259
-------------------	-----

Parity with the reference implementation	262
--	-----

Conclusion	264
----------------------	-----

PARTIE II. Version française de la revue	267
--	-----

Construire le cadre génomique de référence de la levure	268
---	-----

Le premier génome de référence eucaryote	268
--	-----

Établir un cadre évolutif	269
-------------------------------------	-----

De la séquence à la fonction	272
--	-----

Construire une boîte à outils expérimentale à l'échelle du génome	273
---	-----

Un cadre pour la génomique moderne	274
--	-----

Découvrir que la souche S288C n'est pas représentative de l'espèce	275
--	-----

S288C, une représentante peu fidèle de l'espèce	275
---	-----

Les premiers aperçus de la diversité des populations	276
--	-----

Le séquençage de génomes complets révèle l'étendue de la diversité génétique	277
--	-----

D'un génome de référence à une perspective d'espèce	278
---	-----

Passer de centaines à des milliers de génomes	278
---	-----

Élargir le recensement génomique de l'espèce	280
--	-----

Le projet 1002 Yeast Genomes	281
Reconstruire les histoires de domestication	281
L'Asie de l'Est et la recherche de l'origine de l'espèce	283
Vers un cadre à l'échelle de l'espèce	285
L'ère pangénomique : assemblages télomère-à-télomère et variation structurale	285
Les limites d'un génome de référence unique	285
Le séquençage en lectures longues change l'échelle de l'assemblage des génomes	286
La variation structurale s'impose comme une source majeure de diversité	287
Des assemblages télomère-à-télomère vers une vision pangénomique	287
Graphes de pangénome contre référence unique	288
Vers une représentation complète de la diversité de l'espèce . . .	288
De la diversité génétique aux relations génotype-phénotype	288
Les fondements conceptuels de la génétique quantitative	289
La transition vers les études d'association pangénomiques	290
Au-delà des SNP : variation structurale et génome accessoire . . .	291
Multi-omique et phénotypes intermédiaires	291
Conclusion et perspectives	293
PART III. Publications, communications and teaching	306
List of publications	307
Oral communications	308
Teaching	308



Introduction

This introduction served as the basis for a review marking the 30th anniversary of the publication of the *Saccharomyces cerevisiae* reference genome.

From one genome to thousands, and beyond

Gauthier Brach¹ and Joseph Schacherer^{1,2,3,4}

¹ Université de Strasbourg, CNRS, Inserm, IGBMC UMR 7104 – UMR-S 1258, 67400 Illkirch, France

² Institut Universitaire de France (IUF), 75231 Paris, France

³ CIC bioGUNE, Bizkaia Technological Park, Building 800, 48160 Derio, Spain

⁴ Ikerbasque, Basque Foundation for Science, Bilbao, Spain

FEMS Yeast Research, 2026

 [10.1093/femsyr/foag038](https://doi.org/10.1093/femsyr/foag038)

Introduction

I

In April 1996, an international consortium of more than 600 scientists from over 90 laboratories published the complete genome sequence of the baker's yeast *Saccharomyces cerevisiae* strain S288C (Goffeau et al., 1996; Dujon & Cooper, 2026). Spanning 12.07 megabases (Mb) across 16 nuclear chromosomes, it was the first eukaryotic genome to be sequenced and, at the time, the largest genome assembled (Mewes et al., 1997; Goffeau, 2000). Three decades later, that single reference has grown into a catalogue of more than 3,000 genomes divided into 39 clades, with population-scale transcriptomes and proteomes profiled across the species (Caudal et al., 2024; Loegler et al., 2024; Muenzner et al., 2024).

Each technological wave in yeast genomics expanded the biological scale at which *S. cerevisiae* could be studied: from a single genome, to genome-wide function, to population diversity, to pangenomes, and ultimately toward predictive genotype-to-phenotype biology. A recurring theme is reference bias: the progressive realization that limiting analyses to a single laboratory strain hides the diversity of the species, and the ongoing effort to replace the linear reference with a graph-pangenome. Here, we follow that trajectory and the conceptual reframing it brought over the years.

This introduction is organized into four parts. The first covers the building of the yeast genomic reference, from the 1996 sequence to the functional-genomic toolkit that enabled the systematic study of almost every gene. The second discusses the discovery that S288C is not representative of the whole species and the population-genomics era that resolved the species into structured populations. The third follows the pangenome era, in which long reads and structural-variant analyses finally made the full extent of genomic diversity across the species visible. Finally, the fourth turns from cataloguing variation to dissecting it, framing the genotype-to-phenotype problem and the mapping strategies designed to crack it.

S. cerevisiae is an ideal model for this endeavor. Its genome is compact enough to be assembled, annotated, and manipulated with relative ease, yet it retains eukaryotic complexity. This unique balance placed yeast at the center of discoveries recognized by multiple Nobel Prizes: from Leland Hartwell's identification of the universal regulators of the cell cycle and Randy Schekman's mapping of the secretory pathway, to Yoshinori Ohsumi's elucidation of the mechanisms of autophagy (Hartwell et al., 1974; Novick & Schekman, 1979; Tsukada & Ohsumi, 1993). The functional-genomic toolkit built around the S288C reference, including the deletion collection, the green fluorescent protein (GFP) and tandem affinity purification (TAP)-tag libraries, and the Synthetic Genetic Array, covers almost every gene and protein. This combination of a compact and well-annotated genome, a mature genetic toolkit, broad ecological and geographical diversity, and a community culture of open data sharing makes *S. cerevisiae* particularly well suited to bridge genotype and phenotype.

PART I. Building the yeast genomic reference

The first eukaryotic genome (1996-2000)

In 1989, a European consortium led by André Goffeau launched the *S. cerevisiae* genome project, convincing the European Commission to fund the sequencing of chromosome III as a pilot phase (Goffeau & Vassarotti, 1991; Dujon, 1992; Oliver et al., 1992). Rather than building dedicated sequencing centers, the program enlisted existing laboratories, each sequencing an assigned segment. This decentralized model was unprecedented in biology at the time (Dujon, 2015). Chromosome III was completed first, chosen because physical maps and clone libraries were already available. It became the first eukaryotic chromosome ever sequenced, through a collaboration of more than 140 scientists from 12 countries. At 315 kilobases (kb), it represented approximately 2.6% of the nuclear genome. On this chromosome, 182 open reading frames (ORFs) were identified, but only 37 corresponded to previously known genes. Interestingly, more than 80% of the newly discovered ORFs showed no significant homology to any previously sequenced gene (Oliver et al., 1992). The fact that such a large share of the genome of the most popular model eukaryote remained uncharacterized boosted and justified the completion of the genome project. Thus, this success gave the project its momentum, and successive European programs expanded the consortium to more than 80 laboratories, while efforts in North America and Japan tackled additional chromosomes.

The sequencing relied on Sanger's dideoxy chain-terminator method, which, since late 1970s, has been the DNA-sequencing method of choice (Sanger et al., 1977b). A DNA polymerase extends a primer along the template in the presence of a low concentration of chain-terminating 2',3'-dideoxynucleotides. Incorporation halts elongation at random positions, producing a set of fragments that are resolved by length on a gel to reveal the sequence one base at a time (Fig. 1a). The approach refined Sanger's earlier "plus and minus" method and had already delivered the first complete DNA genome of any organism, the 5,386-nucleotide bacteriophage ϕ X174 (Sanger & Coulson, 1975; Sanger et al., 1977a, 1978). A parallel approach based on chemical cleavage of DNA was rapidly discarded because of using toxic chemicals, especially a powerful carcinogen and rocket fuel, hydrazine (Maxam & Gilbert, 1977, 1980). By the mid-1990s, the radioactive labels were replaced with fluorescent dyes and slab gels were abandoned in favor of capillary array electrophoresis. These improvements helped scale the method to be able to sequence up to several kb per day (Smith et al., 1986; Luckey et al., 1990), a level of throughput that made a 12 Mb effort feasible.

Each group contributed finished Sanger sequences to what became the first eukaryotic genome assembled. In total, 633 authors from 92 laboratories distributed in 75 cities across 20 countries were involved in the project (Dujon & Cooper, 2026). In April 1996, the result was announced at a press conference in Brussels, only a few months after the first bacterial genome had been completed using a different, whole-genome shotgun strategy (Fleischmann et al., 1995). The collaboration culminated with a publication in *Science* titled "Life with 6000 genes", for the first eukaryotic genome ever assembled

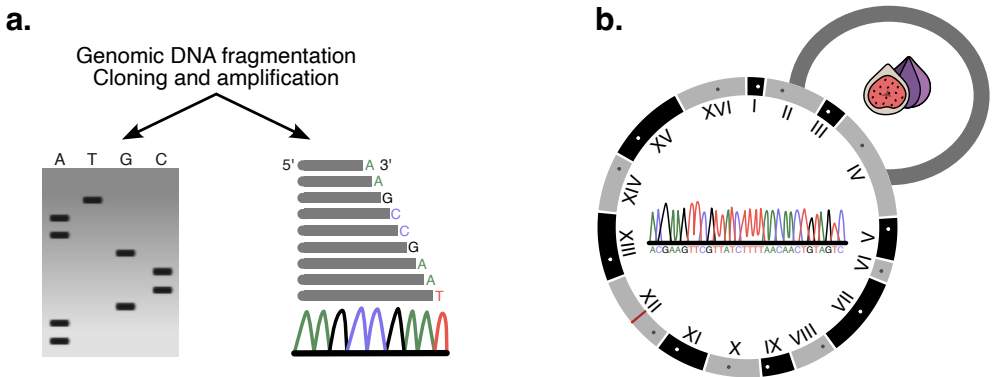


Figure 1 | Sequencing the first eukaryotic genome.

a. First-generation (Sanger) sequencing. Genomic DNA is fragmented, cloned and amplified, then copied in chain-termination reactions that randomly generate a set of fragments each ending in a labeled base. After electrophoresis (historically across four lanes, one per terminating base), the ladder is read from bottom to top to reconstruct the sequence (left); modern instruments display the same information as a fluorescent chromatogram (right).

b. *Saccharomyces cerevisiae*, strain S288C isolated from a rotten fig, became the first eukaryote with a fully sequenced genome: sixteen nuclear chromosomes (I–XVI; alternating bands with centromeres marked, the rDNA [ribosomal DNA] array on chromosome XII in red), assembled base-by-base from Sanger reads.

(Goffeau et al., 1996; Mewes et al., 1997) (Fig. 1b). In the 12.07 Mb sequence, 5,885 predicted protein-coding genes alongside a large number of non-coding elements were identified: approximately 140 rRNA (ribosomal RNA), 40 snRNA (small nuclear RNA), and 275 tRNA (transfer RNA) genes.

As expected, the genome was very compact: less than 5% of genes contained introns, and repetitive DNA was largely confined to transposable (Ty) elements and subtelomeric repeats. Intergenic regions averaged only a few hundred base pairs. The small genome size made yeast uniquely suited for reverse genetics. In principle, every ORF could be disrupted, tagged, or replaced by homologous recombination. Furthermore, only 43% of ORFs could be assigned a function, based either on direct experimental evidence or strong homology (Mewes et al., 1997). Around 30% of ORFs were complete orphans with no detectable homolog in any organism (Dujon, 1996). The fact that the best-studied eukaryote still harbored thousands of genes of unknown function was a major finding, and it would drive the next decade of systematic functional genomics.

The reference strain itself has a complex history. S288C was constructed by Carl Lindegren through a series of crosses and refined by Robert Mortimer for laboratory use. Nearly 90% of its genome derives from EM93, a strain isolated by Emil Mrak in 1938 from a rotting fig in California's Central Valley (Fig. 2). The rest comes from a handful of other strains, including a commercial baking yeast (Lindegren, 1949; Mortimer & Johnston, 1986; Engel et al., 2013). Mortimer shared S288C widely, and the derivatives used for sequencing each carried their own modifications. Isogenic strains carrying a repaired *GAL2* allele and non-reverting auxotrophic markers were created (Winston et al., 1995), two of which (FY23 and FY73) Bernard Dujon mated to produce the diploid

FY1679 (Dujon & Cooper, 2026). AB972, an ethidium-bromide-induced ρ^0 (lacking the entire mitochondrial genome) derivative of X2180-1B-trp1 (Olson et al., 1986), was used as the other major DNA source. The 1996 reference sequence was therefore not obtained from a single colony: AB972 was the source of most chromosomes, FY1679 provided chromosomes VII, X, XI, and XV (and contributed to IV and XIV), while chromosome II came from S288C itself (Dujon, 1996; Engel et al., 2013). The result was a mosaic assembly across three near-isogenic backgrounds, S288C, AB972, and FY1679, differing by only a few single nucleotide polymorphisms (SNPs) per 100 kb (Engel et al., 2013), but technically representing no single strain (Fig. 2).

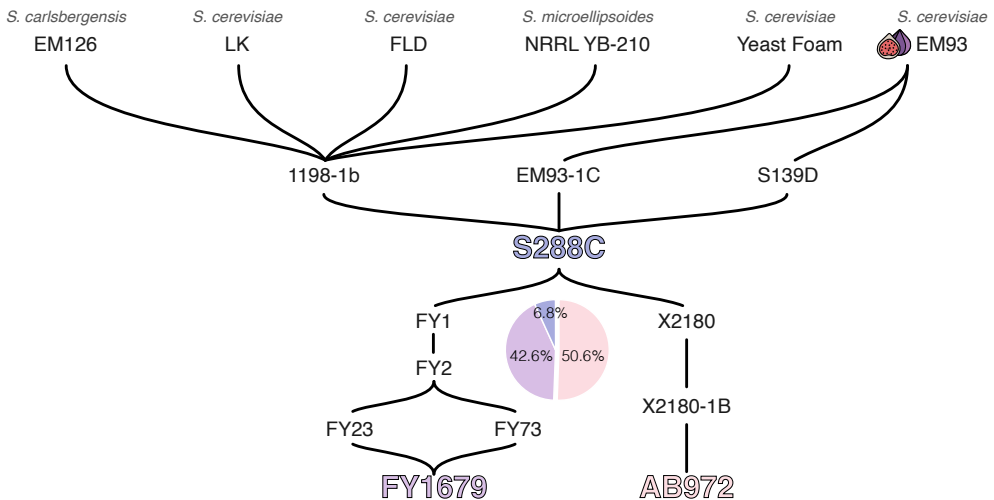


Figure 2 | Simplified genealogy of S288C and the strains used for sequencing.

S288C is not a wild isolate but a laboratory strain. The fig isolate EM93, recovered from a rotting fig in California in 1938, contributes 88% of the S288C (blue) genome. The two strains actually sequenced in the genome project descend directly from S288C and are near-isogenic to it: FY1679 (purple) and AB972 (pink). The pie chart gives the fraction sequenced from each strain. Because all three share the S288C background, this composite technically represents almost a single genome.

At the same time, the digital infrastructure required to browse the genome emerged. Through a series of meetings in the late 1980s and early 1990s, the sequencing consortium created a systematic nomenclature in which each ORF was labeled with "Y" for yeast, a letter for the chromosome, "L" or "R" for the chromosome arm, a sequential number counting outward from the centromere, as well as a "W" or "C" suffix for the Watson or Crick strands, promptly used by the community (Thierry et al., 1990; Jacquier et al., 1992; Engel et al., 2013). Meanwhile, the *Saccharomyces* Genome Database (SGD) was established in 1993 at Stanford by J. Michael Cherry and adopted these systematic names (YAL001C, YBR002W, ...). As it matured into its modern form, it also grew well beyond a sequence repository (Cherry et al., 1998, 2012). It progressively centralized genetic and physical maps, exhaustive protein catalogues (Chervitz et al., 1999; Weng et al., 2003), and Gene Ontology annotations (Dwight et al., 2002), integrating these layers with emerging gene-expression data to transform the static genome into a curated resource for the community (Engel et al., 2022).

As Engel and colleagues noted, the complete sequence transformed anecdotal characterizations into exhaustive genome-wide catalogues: from tRNA genes and retrotransposon insertions to ribosomal protein genes and subtelomeric repeats (Lowe & Eddy, 1997; Percudani et al., 1997; Pryde et al., 1997; Kim et al., 1998; Planta & Mager, 1998; Engel et al., 2013). At the turn of the millennium, the yeast community was already taking advantage of the S288C sequencing project, barely half a century after Lindegren's first hand-drawn chromosome maps (Lindegren, 1949).

Soon, comparative genomics placed S288C in its evolutionary context. Molecular evidence revealed that *S. cerevisiae* had undergone an ancient whole-genome duplication (WGD) event, explaining the ubiquitous duplicated synteny blocks visible across chromosomes (Wolfe & Shields, 1997; Seoighe & Wolfe, 1999). Ohnologs, which are paralogs originating from a whole-genome duplication event, were mapped by comparing the *S. cerevisiae* genome to the pre-WGD species *Ashbya gossypii* and *Kluyveromyces waltii* (Dietrich et al., 2004; Kellis et al., 2004). Today, 551 ohnolog pairs are found in the genome distributed across 16 chromosomes. These 16 descend from an eight-chromosome ancestor, and were reshuffled by ~144 post-WGD rearrangements (Wolfe, 2015). The duplication is now widely accepted not as a doubling within a single lineage (autopolyploidy) but as the outcome of an ancient hybridization between two already-diverged yeast lineages (allopolyploidy). Under this model, the two members of each ohnolog pair descend from distinct parental genomes and were therefore already divergent at the moment of duplication, which would account for much of the sequence divergence between them (Marcet-Houben & Gabaldón, 2015). When smaller-scale duplications are included, 40% of genes are in paralogous families (Dujon, 1996; Llorente et al., 2000). The majority of the duplicated genes were lost, which restored a single-copy gene dosage, while the remaining pairs underwent specific evolutionary trajectories such as subfunctionalization, where ancestral roles were partitioned between copies, or neofunctionalization, where one copy evolved entirely novel functions (Dujon et al., 2004; Kellis et al., 2004; Conant & Wolfe, 2008; Dujon, 2010). It turned out that this influx of genetic material provided raw material for adaptation to specialized niches. One example is the Crabtree fermentative lifestyle, in which a preferential retention of glycolytic enzyme duplicates contributed to increased glycolytic flux (De Deken, 1966; Piškur, 2001; Conant & Wolfe, 2007; Pfeiffer & Morley, 2014). The WGD also provided a structure for interpreting genetic interaction data: many synthetic lethal pairs turned out to be ohnologs whose functional redundancy has been preserved for over 100 million years (Guan et al., 2007; Kuzmin et al., 2020). For the genes that have a WGD partner, this redundancy acts as a functional buffer. Indeed, the effects of single-gene deletions are typically masked by the surviving paralog which preserves the ancestral function (Dean et al., 2008; DeLuna et al., 2008).

A comparative context was provided soon after by the Génolevures project. This project sequenced partial and then complete genomes of multiple Hemiascomycetes species spanning hundreds of millions of years of evolution (Souciet et al., 2000; Dujon et al., 2004; Sherman et al., 2004; Shen et al., 2018). By comparing gene content, synteny, and genome organization across the Saccharomycotina, Génolevures revealed that gene loss after the WGD had been massive and that horizontal gene acquisition from bacteria

was more common than once assumed. Together, these studies opened what was called “a rewarding route towards elucidating novel general principles of evolution” (Dujon, 2010, 2015).

The reference annotation also evolved over the years. Between 1996 and 2011, SGD issued a few major successive genome releases, incrementally correcting sequence errors, revising ORF boundaries, and deprecating dubious genes as new experimental evidence accumulated (Engel et al., 2013). One of the most significant updates came in 2011, after resequencing of a single colony of AB972 using Illumina technology. The resulting R64 assembly corrected sequence errors affecting 194 predicted protein sequences and 232 intergenic regions. Even a “finished” Sanger genome harbored non-trivial inaccuracies (Engel et al., 2013). Comparison of the resequenced AB972 with FY1679 showed only four single nucleotide polymorphisms per 100 kb, confirming that the two principal sequencing strains were virtually indistinguishable. The R64 coordinates remain the standard reference to this day, although the annotation continues to be refined with each SGD release, the latest at the time of writing being R64-5-1, released in 2024 (Engel et al., 2022, 2025).

Two and a half years after the nuclear genome was published, the ~86 kb mitochondrial genome completed the picture, sequenced from FY1679 (Foury et al., 1998). Its organization revealed mobile group I and group II introns encoding maturases and endonucleases, more than 150 short GC clusters of 10-80 bp scattered through otherwise quasi-pure AT intergenic spacers (83%), and all but one gene transcribed from a single strand (Foury et al., 1998). Mitochondrial genome variation would also later be recognized as an important dimension of population-level diversity, with for example, intron presence/absence polymorphisms and recombination patterns varying extensively across strains in the species (De Chiara et al., 2020).

The yeast community now had a complete genome, a curated database, and a continuously refined annotation. Yet more than half of the ~6,000 predicted genes still had no assigned function. A wave of tools was needed in order to study every gene one by one. They would be built in the S288C background, made in part possible by the compactness that had justified the sequencing project in the first place.

From sequence to function

The reference made genome-scale assays possible almost overnight. Within months of its publication in 1996, transcripts, proteins and chromatin were being studied across the entire genome for the first time. Less than 2 years after the genome publication, complementary DNA microarrays were used to monitor the expression of nearly every yeast gene during the metabolic transition from fermentation to respiration (diauxic shift) (DeRisi et al., 1997). The transcriptional programs of the cell-cycle (Spellman et al., 1998), sporulation (Chu et al., 1998), and the environmental stress response were then characterized (Gasch et al., 2000).

Furthermore, a compendium of expression profiles demonstrated that gene function could be inferred from co-expression patterns (Hughes et al., 2000), while chromatin

immunoprecipitation on microarrays (ChIP-chip) provided the first genome-wide maps of transcription factor binding (Lee et al., 2002; Harbison et al., 2004). At the same time, progress in proteomics helped researchers look beyond gene expression. Within a year of the genome publication, mass spectrometry was scaled up for large-scale protein identification, matching yeast proteins to their genes directly from two-dimensional gels (Shevchenko et al., 1996). A comparison of messenger RNA (mRNA) and protein abundances revealed a poor correlation, a result whose implications for genotype-to-phenotype prediction remain central to this day (Gygi et al., 1999; Teyssonnière et al., 2024b).

Finally, yeast was the testing ground for three-dimensional genome analysis. Chromosome conformation capture (3C), the foundational technology for subsequent chromatin interaction studies, was developed in yeast (Dekker et al., 2002). Nucleosome positions and histone modifications were mapped genome-wide, establishing that even a compact genome has complex chromatin architecture (Pokholok et al., 2005; Yuan et al., 2005).

These genome-scale experiments revealed what is expressed and when, but not what each gene does. Assigning function to the thousands of orphan ORFs then required systematic perturbation. Over the following decade, the community constructed a suite of strain collections and accompanying assays, all built in the BY4741/BY4743 genetic backgrounds, isogenic derivatives of S288C.

These resources tackled four related questions: what happens when each gene is removed or weakened? Where are the products located and how abundant are they? How do genes and proteins connect into networks? And how do cells respond to chemical and environmental perturbations?

(i) Perturbing genes

Removing each gene one at a time and studying what happens to the cell became the logical route in order to systematically assign functions to genes. One of the foundational tools for this endeavor was the yeast knockout (YKO) collection. Each targeted ORF was replaced by a kanamycin-resistance cassette flanked by two unique 20-mer oligonucleotide sequences. These barcoding sequences can be amplified from pooled genomic DNA and quantified by hybridization to high-density oligonucleotide arrays. Thanks to this collection, massively parallel competitive fitness assays became a possibility (Winzeler et al., 1999). Pooling 558 homozygous deletion strains and growing them for approximately 60 generations revealed fitness defects in nearly 40% of the strains and exposed a disconnect between transcriptional induction and fitness requirement. Indeed, genes upregulated under a given condition were rarely the ones whose deletion impaired growth in that condition.

The near-complete collection followed three years later. In total, 5,916 genes were deleted, representing 96% of the annotated ORFs. Up to four strain types were generated for each: MAT α and MAT α haploids, homozygous diploids, and heterozygous diploids (Giaever et al., 2002). Close to 19% of genes proved essential for viability on rich glucose medium, yet nearly half of these had never before been annotated as

essential. The collection was phenotyped under six growth conditions, revealing that approximately 15% of viable homozygous deletion strains grew slowly even in rich medium. In addition, it showed that the disconnect between expression and fitness was pervasive: less than 7% of genes induced by galactose were also required for optimal growth on that carbon source.

The deletion collection nonetheless exposed an issue. Approximately 19% of gene deletions were lethal and alternative strategies were therefore needed to study these essential genes. A titratable-promoter collection replaced the native promoters of more than 600 essential genes with a tetO7 cassette repressible by doxycycline. This enabled depletion of the encoded protein after transcriptional shutoff. This resource allowed machine-learning classifiers trained on the resulting multi-dimensional phenotypic profiles to assign functional categories to 75 previously uncharacterized essential genes (Mnaimneh et al., 2004). A few years later, a second approach exploited mRNA destabilization. In the DAMP (decreased abundance by mRNA perturbation) collection, a cassette inserted immediately after the stop codon of 842 essential genes is used to disrupt the 3' untranslated region (UTR). This reduces transcript levels two- to tenfold while leaving the encoded protein unmodified (Breslow et al., 2008). A third approach involved designing temperature-sensitive alleles for essential genes via the diploid-shuffle method, in which mutagenized polymerase chain reaction (PCR) products of each essential ORF are integrated at the native locus of a heterozygous diploid YKO strain. The resulting collection initially covered 250 genes (Ben-Aroya et al., 2008).

It was later expanded to nearly 500 essential genes (Li et al., 2011). All three collections were built in the same BY backgrounds and were compatible with the deletion set, enabling combinatorial screens. In combination with the YKO strains, they provided near-complete coverage of the *S. cerevisiae* gene repertoire through deletion, depletion, destabilization, or even conditional inactivation. When combined with growth assays, they helped reveal that the majority of the genome contributes to fitness even under a single laboratory condition (Breslow et al., 2008).

The most radical perturbation tool of this era came with the construction of an entirely synthetic genome. The Sc2.0 project set out to redesign all 16 chromosomes of *S. cerevisiae*. Among other modifications, all 1,416 TAG stop codons were recoded to TAA, 3,932 loxPsym palindromic recombination sites were inserted 3 bp downstream of all non-essential genes, introns and repetitive elements were discarded, and tRNA genes were relocated to a dedicated neochromosome (Richardson et al., 2017). The embedded loxPsym sites enable SCRaMBLE (Synthetic Chromosome Rearrangement and Modification by LoxP-mediated Evolution), in which Cre recombinases generate combinatorial genomic rearrangements. Individual synthetic chromosome arms were shown to be viable in 2011 (Dymond et al., 2011), and the first complete human-designed chromosome (synIII) soon followed (Annaluru et al., 2014).

The complete set of 16 synthetic chromosomes and a tRNA neochromosome were built and reported in 2023, although the most extensively consolidated strain, syn6.5, only carried seven synthetic chromosomes (Schindler et al., 2023; Zhao et al., 2023).

(ii) Tracking proteins

Another group of resources was used to determine where each protein localizes in the cell and how abundant it is, rather than looking at what happens when its gene is removed. A collection of more than 6,000 strains, each expressing a C-terminal GFP fusion from the endogenous chromosomal locus, provided the first proteome-wide subcellular localization atlas for any eukaryote. Of these, 4,156 showed GFP signal above background noise, covering 75% of the proteome. The detected proteins were assigned to 22 cellular compartments (Huh et al., 2003).

Proteins whose targeting depends on a C-terminal localization signal were, predictably, consistently mislocalized. This limitation was later addressed by the SWAp-Tag (SWAT) approach, in which an acceptor module inserted at the 5' end of each gene can be exchanged for any tag via an inducible double-strand break followed by homologous recombination (Yofe et al., 2016). Applied to 1,800 endomembrane proteins, N-terminal tagging revealed that a large proportion of them displayed a different localization from the one assigned by C-terminal tagging, showing that no single tagging strategy captures the full picture.

Quantifying protein abundance came next, when a genome-wide collection of C-terminal TAP fusions (Rigaut et al., 1999) was assembled. This collection enabled the quantification of protein levels across nearly the entire proteome. Abundances spanned more than four orders of magnitude, from fewer than 50 to more than one million molecules per cell. However, the correlation with mRNA levels was modest, implying that post-transcriptional regulation is pervasive (Ghaemmaghami et al., 2003). In total, 75% of them were present at fewer than 5,000 copies per cell, a range that mass spectrometry could barely reach at the time.

(iii) Mapping interactions

Connecting genes and proteins into the cellular network required a different kind of assay. The same TAP-tagged strains were then used for systematic affinity purification of protein complexes. Two large-scale efforts were published back-to-back, jointly defining the architecture of the yeast proteome. One study identified 491 complexes organized into stable cores and variable attachments. This combinatorial recycling of functional modules was defined as a general organizational principle (Gavin et al., 2006). The other study resolved 547 complexes. It identified 7,123 high-confidence interactions among 2,708 proteins, out of 4,087 proteins detected overall. Half of the complexes were absent from existing databases (Krogan et al., 2006). Both studies converged on a modular, hierarchically organized architecture of the proteome. But both relied on affinity purification, which captures stable complexes but misses transient or condition-dependent contacts.

To complement this view, genome-wide yeast two-hybrid (Y2H) screens were used to identify binary, often transient, protein-protein interactions, providing the first systematic physical pairwise interaction map. Approximately 6,000 ORFs were screened, identifying 957 putative interactions among 1,004 proteins. This was far from saturation but sufficient to reveal unexpected connections (Uetz et al., 2000). A later

study provided a complementary interaction dataset but the overlap between the two studies was modest. This would show how far from complete any single screening remained (Ito et al., 2001).

A further application of the deletion collection is the Synthetic Genetic Array (SGA). Briefly, this method automates the construction of double mutants by mating a query strain carrying a marked deletion to the full array of ~4,700 viable deletion strains. This is followed by sporulation of the resulting diploids and selection of haploid double-mutant progeny through a combination of mating-type-specific drug-resistance markers (Tong et al., 2001). Synthetic lethal and synthetic sick interactions (in which two individually viable deletions combine to produce lethality or fitness loss) identify pairs of genes that are individually dispensable but jointly essential: each can compensate for the loss of the other because they might operate in parallel pathways, and only when both are removed does the cell lose a critical function. One illustrative case is retained WGD pairs. Among 399 surveyable ohnolog pairs, 69 (17%) displayed a negative interaction, synthetic sickness or lethality, under standard laboratory conditions (Dean et al., 2008). These pairs retain enough functional overlap that each gene can buffer the loss of the other. The first large-scale application screened eight query genes and identified 291 genetic interactions. At the same time, it demonstrated that the approach could be used at a large scale (Tong et al., 2001). Less than 10 years later, a near-genome-wide map tested 5.4 million gene pairs and detected approximately 170,000 interactions. This revealed that genes cluster into discrete functional modules corresponding to core biological processes (Costanzo et al., 2010). Only six years later, the map was expanded to 23 million double mutants. Approximately 550,000 negative and 350,000 positive interactions (where the double mutant grows better than expected) were identified across close to 90% of all yeast genes. This included essential genes probed through temperature-sensitive alleles (Costanzo et al., 2016). This global genetic interaction map is the largest wiring diagram available for any organism, and it demonstrated that genetic interaction profiles predict gene function and complex membership more reliably than physical interactions alone.

(iv) Chemical genomics

Probing how each gene contributes to fitness when chemicals and environments are varied is a different axis that was explored by the community. Here, the deletion collection also enabled chemical genomics at scale. In haploinsufficiency profiling (HIP), the complete pool of approximately 6,000 heterozygous deletion strains is grown competitively in the presence of a bioactive compound. Reducing the copy number of a drug's target from two to one makes the strain more sensitive, causing it to be outcompeted (Giaever et al., 2004). Because the heterozygous collection includes deletions affecting 1,000 essential genes that are absent from the viable homozygous collection, HIP can identify targets that no haploid or homozygous screen could reach.

The complementary approach, homozygous deletion profiling (HOP), uses the homozygous diploid collection to identify genes that buffer the cell against a compound's effect. Combining HIP and HOP across approximately 1,144 chemical and environmental conditions revealed that 97% of gene deletions cause a fitness defect

I in at least one condition. Genes that showed no phenotype in rich medium weren't functionless but rather conditionally essential, with their effects becoming apparent only outside the usual laboratory conditions (Hillenmeyer et al., 2008). Genome-wide fitness profiling in the presence of bioactive compounds allowed drug targets and resistance pathways to be identified in a systematic way (Giaever et al., 2002, 2004; Hillenmeyer et al., 2008; Piotrowski et al., 2017). This made yeast a platform for both fundamental biology and pharmaceutical discovery.

Over this decade, the field moved from studying genes one at a time to systematically perturbing, tagging and measuring the entire cell. Functions could now be assigned at genome scale and networks could be drawn. The response to chemical or environmental perturbations could be measured across thousands of genotype-by-condition combinations.

Every resource described in this section was built in a single genetic background and optimized for a single reference. These tools, as powerful as they are, assume that S288C is representative of the species. A parallel line of research, starting in the mid-2000s, would show that it is not.

PART II. Discovering that S288C is not the species

Early population genomics

The reference strain itself turned out to be biologically atypical. S288C carries a defective Ty1 transposable element inserted in the 3' region of *HAP1*, disrupting the heme-dependent activator that couples respiratory gene expression to oxygen availability; in S288C-derived backgrounds this allele behaves as a null and disables a major branch of aerobic gene regulation (Gaisne et al., 1999). S288C also carries mutations in *FLO1* and *FLO8*, rendering it non-flocculent and abolishing the filamentous growth response that many strains rely on to colonize surfaces and form biofilms (Liu et al., 1996; Louis, 2016). These are not isolated examples: the reference carries null or loss-of-function alleles of *GAL2* (galactose permease), *MEL1* (alpha-galactosidase), *BIO1* and *BIO6* (biotin biosynthesis), among others. Several of these differences make S288C convenient at the bench: the loss of flocculation and filamentous growth keeps cells uniformly dispersed in liquid culture, allowing stable, reproducible optical-density measurements, while the auxotrophies engineered into its BY derivatives (*ura3*, *leu2*, *his3*, *met15*) provide the selectable markers needed for plasmid transformation and gene replacement. Others, such as the defective *HAP1* allele, are not advantages but are simply tolerated on rich medium. Taken together, these laboratory adaptations distinguish S288C from the diversity found in wild populations (Mortimer & Johnston, 1986; Louis, 2016).

The first attempts to quantify how much diversity lay outside the reference included microarray-based detection of single-feature polymorphisms (SNPs) in a small panel of 14 strains (Winzeler et al., 2003), genome-wide expression differences among Tuscany vineyard isolates (Townsend et al., 2003), and ecological surveys recovering wild *S. cerevisiae* from oak-associated soils alongside its sister species *S. paradoxus* (Sniegowski et al., 2002). Array-based comparative genome hybridization (aCGH) then extended this probe-based view of sequence variation to gene copy number. Four commercial wine strains profiled against S288C each carried amplifications or deletions of individual genes, many encoding transporters such as the HXT hexose permeases and the ENA sodium pumps (Dunn et al., 2005). A first population-level view emerged from a multilocus sequence survey of five unlinked loci across 81 strains isolated from diverse origins: vineyards, clinical samples, sake fermentations, traditional African beverages (Fay & Benavides, 2005). Conceptually, the results challenged the prevailing idea that *S. cerevisiae* was an entirely domesticated species and showed that the species consists of both wild and domesticated populations.

Subsequent studies tried to resolve finer structure within and between these groups. For example, oak-associated strains from a Pennsylvania forest were genetically uniform and cleanly separated from Italian vineyard isolates across all four sequenced loci (Aa et al., 2006), establishing ecological niche as a primary axis of differentiation. Selection acting on *SSU1*, a sulfite transporter whose substrate is used as a preservative in winemaking, provided the first evidence of adaptive evolution in a winemaking-relevant locus (Pérez-Ortín et al., 2002). This was direct evidence of

selection at a fermentation-specific locus in *S. cerevisiae*. At a larger scale, microsatellite genotyping in 651 strains showed that yeast diversity is shaped by human fermentation history: wine, beer, bread, and sake strains each form distinct clusters, with a Lebanese wine group near the root suggesting a migration out of Mesopotamia coincident with vine domestication (Legras et al., 2007). Human cultural history thus emerged, alongside geography and ecology, as a force structuring the species.

These multilocus and microsatellite surveys established the main axes of population-level structure, distinguishing wild from domesticated isolates and revealing ecological rather than purely geographic clustering. But these studies were limited, as they sampled only a handful of loci and could not resolve genome-wide patterns of variation and structural differences. The transition toward whole-genome sequencing (WGS) approaches began with the first complete genome of a non-laboratory strain, YJM789, a clinical isolate from an AIDS patient with pneumonia. Sanger shotgun sequencing revealed approximately 60,000 SNPs and 6,000 insertion–deletion variants (InDels) relative to S288C, a nucleotide divergence of 0.6%. Chromosome I carried a 12 kb region of *S. paradoxus* origin, providing early evidence that interspecific introgression shapes intraspecific variation (Wei et al., 2007). The following year, the wine strain AWRI1631 was sequenced and compared with both S288C and YJM789: the three genomes differed by an average of one SNP every 150 bp, the Ty elements differed in number and arrangement, and several AWRI1631-specific regions were predicted to encode proteins unique to that industrial strain (Borneman et al., 2008). A larger comparison of S288C with the vineyard strain M22 and the oak-tree isolate YPS163 catalogued close to 88,000 polymorphisms across the genomes (Doniger et al., 2008). They estimated that 12% of coding and 7% of non-coding SNPs were deleterious and demonstrated that predicted deleterious cis-regulatory variants correlated with allele-specific expression differences across nine environmental conditions. For the first time, the regulatory consequences of intraspecific variation could be measured directly. These first few genomes revealed tens of thousands of variants and regulatory differences between strains. The next step was to resolve how that variation was organized across the species at a larger scale.

In 2009, two studies published in the same issue of *Nature* provided the first genome-wide view of *S. cerevisiae* population structure, using different approaches (Fig. 3). The primarily Sanger-based sequencing Saccharomyces Genome Resequencing Project (SGRP) sequenced 36 *S. cerevisiae* strains at a low coverage, defining five lineages (Malaysian, Sake, North American, West African, and Wine/European) and a large group of mosaic strains in which different chromosomal segments could be traced to different lineages (Liti et al., 2009). In *S. paradoxus*, populations partitioned cleanly along geographic lines with little admixture, whereas in *S. cerevisiae* the same five lineages were heavily intermixed by mosaic genomes, likely the legacy of human-mediated dispersal and outcrossing between previously isolated populations (Liti et al., 2009). In parallel, tiling-array hybridization of 63 strains identified 101,343 segregating sites (30,097 SNPs per strain) as well as 3,985 deletion events larger than 200 base pairs (Schacherer et al., 2009). Together, these two studies provided the first species-wide picture, with five recognizable lineages superimposed on extensive

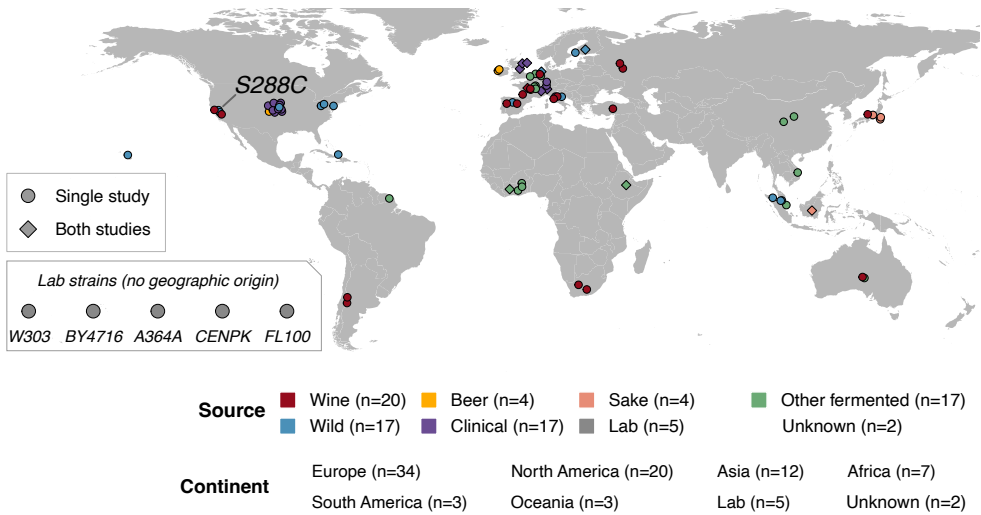


Figure 3 | The first global census of *Saccharomyces cerevisiae*, 2009.

Geographic origins of the 86 isolates pooled from the two foundational population surveys published that year, plotted at their collection sites and colored by ecological source. Circles mark strains characterized in a single study; diamonds mark the 13 strains analyzed in both. Five canonical laboratory strains have no natural provenance and appear in the inset. The species reaches far beyond the bench: isolates span vineyards, breweries, sake and ragi ferments, clinical samples, and a range of wild substrates across six continents. The sampling is biased toward human-associated niches and the global North, with Europe and North America accounting for the bulk of the panel.

mosaicism. Both studies converged on the conclusion that population structure reflects ecological niche as much as geography. A fundamental limitation of these approaches was their dependence on the S288C reference: any sequence absent from that reference (insertions, subtelomeric genes, repetitive regions, ...) was invisible.

That same year, two new genome assemblies demonstrated how much of the species' diversity was missing from the reference. Horizontal gene transfer (HGT) into *S. cerevisiae* was already a known phenomenon: a systematic screen of S288C had identified ten candidate HGT events, all from bacterial donors, the most striking being *URA1* (dihydroorotate dehydrogenase, acquired from *Lactococcus*), whose oxygen-independent activity uniquely enables pyrimidine biosynthesis under the anaerobic conditions of fermentation (Gojković et al., 2004; Hall et al., 2005). The genome of EC1118, a commercial wine strain, carried 34 genes and 5 pseudogenes absent from S288C clustered in three regions acquired by horizontal gene transfer from non-*Saccharomyces* species including *Zygosaccharomyces bailii* and *Z. rouxii*. These regions encoded proteins for carbon and nitrogen metabolism (functions directly relevant to wine fermentation). A survey of 53 isolates showed that the three regions are found almost exclusively among strains of winemaking origin (Novo et al., 2009). Further studies would show that the fungal oligopeptide transporters (FOTs) encoded in these regions confer a fitness advantage during wine fermentation, and that their diversity is greater than what was previously thought (Marsit & Dequin, 2015; Becerra-Rodríguez et al., 2020, 2021). JAY291, a haploid

derivative of the Brazilian bioethanol strain PE-2, showed massive structural variation, including reciprocal translocations, copy-number amplifications of stress-response gene families in subtelomeric regions, while the central portions of chromosomes remained conserved (Argueso et al., 2009). Conceptually, these two papers established two recurrent patterns. The structurally stable core surrounded by plastic subtelomeric regions emerged as a general rule, and the accessory pool of HGT-derived genes was identified as a recurrent niche-adaptation signature. These are zones of adaptive variation that reference-based approaches systematically miss. Together, these 2009 studies established that each strain is shaped by domestication bottlenecks, ecological adaptation, and horizontal acquisitions.

The geographic scope expanded rapidly. A survey of more than 2,000 samples from 11 provinces across China, including primeval forests remote from any human activity, yielded 99 isolates that harbored nearly twice as much genetic variation as all previously characterized lineages worldwide combined (Wang et al., 2012). Eight new lineages were defined. The two most divergent (CHN I and CHN II) were separated from all other lineages by genetic distances double those of any other pairwise comparison. Hainan, a tropical island in southern China, alone harbored three lineages (CHN I, III, and V) whose combined variation was equivalent to that of the entire species. These results relocated the species' ancestral reservoir to Far East Asia and challenged the assumption that European panels captured most of the diversity.

The next step was to characterize the phenotypic consequences of this genetic diversity. High-throughput phenotyping of 86 *Saccharomyces sensu stricto* isolates across 200 environments showed that trait variation within *S. cerevisiae* follows population history rather than ecological source (Warringer et al., 2011). A restriction site-associated DNA sequencing (RAD-seq) approach then extended sampling to 262 globally sourced strains, and showed that geography defined the primary structure, with ecological niche and fermentation history defining finer subpopulations (Cromie et al., 2013). This put a number on the reference-bias problem: laboratory strains miss approximately three quarters of the species' diversity. The phenotyping had already made the point in a single observation. Across all 200 environments, the most atypical strain was S288C itself, the reference strain: it excelled at maltose utilization and growth on rich medium, yet showed poor ethanol tolerance (Warringer et al., 2011). The field had built its functional-genomics toolkit around one of the least representative strains in the species, if not the single least representative one.

By 2013, depending on the study, five to nine populations had been defined. These populations were marked by extensive mosaicism and a structurally conserved chromosomal core flanked by plastic, HGT-prone subtelomeres. Most importantly, three-quarters of the species' allelic diversity was absent from any laboratory strain. The reference genome, in other words, was not the species, and any genotype-to-phenotype map built on S288C alone would capture only a fraction of it. Unfortunately, the field still lacked statistical resolution in order to look at the genotype–phenotype relationship. Indeed, linkage mapping in limited small panels of a few strains could only recover the largest-effect alleles. Resolving the genetic architecture of quantitative traits would require panels an order of magnitude larger (Peter & Schacherer, 2016).

Scaling to thousands of genomes

The transition from descriptive to statistical population genomics required two things: short-read sequencing cheap enough to apply at the scale of hundreds of isolates, and curated panels large enough to deliver fine lineage resolution. Despite having one of the smallest eukaryotic genomes, *S. cerevisiae* lagged noticeably behind other model organisms in the number of sequenced natural isolates. *Arabidopsis thaliana*, *Drosophila melanogaster*, and humans had each assembled panels spanning hundreds to thousands of sequenced genomes, while yeast population genomics remained limited to a few hundred strains, most of them haploid single-spore derivatives. The lag seems counterintuitive as yeast has been one of the leading model organisms. But, simply, the genotype–phenotype relationship in yeast had been dissected mainly through experimental crosses and reverse genetics (approaches requiring only two parents, in which a variant’s effect can be measured directly), so large natural-isolate panels were not the obvious tool. Its natural ecology, too, was recognized late: long treated as a “man-made” organism tied to fermentation and lacking a clear ecological niche, *S. cerevisiae* was only sampled systematically from nature in the 2000s. This hid the natural ploidy and heterozygosity of the isolates. Pushing sample sizes into the hundreds, and eventually thousands, would prove essential for fine-grained lineage resolution and for any realistic prospect of species-scale association mapping (Connelly & Akey, 2012; Peter & Schacherer, 2016).

The enabling technology was the second-generation sequencing wave that had been unfolding in parallel. The first commercial platform (pyrosequencing by 454 Life Sciences) sequenced DNA in millions of picoliter-scale wells and demonstrated the principle on a complete bacterial genome (Margulies et al., 2005). The chemistry itself is based on real-time detection of pyrophosphate released during nucleotide incorporation, a principle described seven years earlier (Ronaghi et al., 1998). The first individual human genome by next-generation technology, James D. Watson’s, was completed on the same platform in two months at around a one-hundredth of the Sanger cost (Wheeler et al., 2008). The Solexa, then Illumina, platform overtook the field, using a sequencing-by-synthesis approach, which adapted Sanger’s principle to a flowcell format (Bentley et al., 2008). In a nutshell, sheared DNA fragments are ligated to specific adapters (Fig. 4a) and hybridized to a lawn of complementary oligonucleotides on a glass surface (Fig. 4b). They are bridge-amplified *in situ* into clonal clusters. Deoxyribonucleoside triphosphates, each carrying a 3′-O-azidomethyl reversible terminator on the sugar and a removable fluorophore on the base, are incorporated one per cycle and the flowcell surface is imaged in four colors (Fig. 4c), though it is worth noting that one- and two-color chemistries do exist. The terminator and the dye are then cleaved and the next cycle can begin. The technique could produce hundreds of millions of reads of 35 nucleotides with greater than 99.7% per-base accuracy, all of that in parallel (Bentley et al., 2008; Goodwin et al., 2016). On the original Genome Analyzer machine, the Illumina technology delivered 35-base reads and sequenced a Yoruba diploid genome to 40x depth in eight weeks for \$250,000 in consumables (Bentley et al., 2008). Read lengths subsequently increased to 300 base pairs on later HiSeq and MiSeq platforms (Goodwin et al., 2016). Other

architectures popped up, but only Illumina stood the test of time. For example, Ion Torrent's proton-sensing chips (Rothberg et al., 2011) and Complete Genomics' DNA nanoballs (Drmanac et al., 2010) remained marginal technologies (Shendure & Ji, 2008; Metzker, 2010; Shendure et al., 2017). As such, per-genome cost fell by approximately 50,000-fold over the 15 years following the Human Genome Project (Goodwin et al., 2016). Population-genomic resources followed: more than 2,500 human genomes (Auton et al., 2015), 1,135 *A. thaliana* ecotypes (Alonso-Blanco et al., 2016), the *D. melanogaster* Genetic Reference Panel (Mackay et al., 2012) along with the 1,011 yeast isolates (Peter et al., 2018).

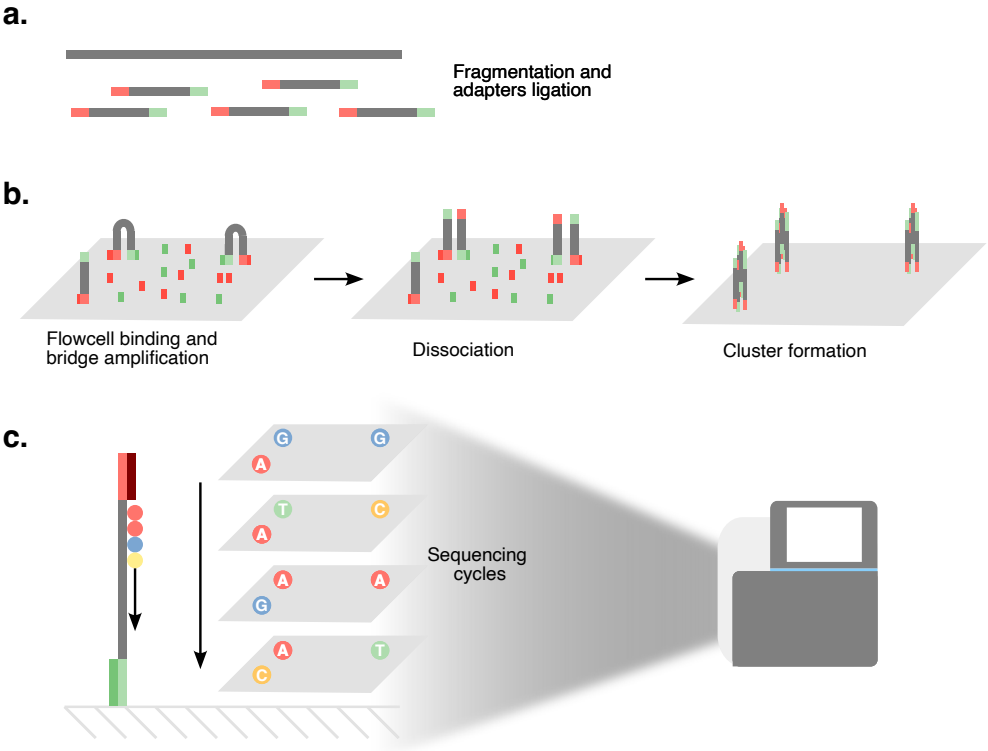


Figure 4 | Short-read sequencing by synthesis (Illumina).

a. DNA fragments, derived from genomic DNA or reverse-transcribed RNA, are flanked by two distinct adapters (red, green), yielding a sequencing library.

b. Library fragments hybridize to a lawn of complementary oligonucleotides on the flowcell surface and are clonally amplified by bridge amplification: each fragment folds over to prime synthesis off a neighboring oligo, and repeated cycles of synthesis and dissociation generate a spatially separated cluster of identical copies.

c. Each cluster is sequenced one base at a time. Per cycle, a single fluorescently labeled reversible-terminator nucleotide is incorporated, the flowcell is imaged to record the base identity at every cluster, and the terminator and dye are then cleaved before the next cycle. Successive images (stacked planes) are converted into per-cluster base calls, reconstructing the sequence of each fragment.

(i) Building the panels

The transition to thousands of genomes started with a series of studies that deepened sampling within particular ecological or geographic niches. At the same time, genome assemblies progressively improved. A precursor study was the 454 resequencing of six industrial strains, which identified novel non-S288C ORFs (Borneman et al., 2011). A study of 83 strains introduced the pangenome concept to yeast genomics (Dunn et al., 2012). The pangenome (or pan-genome) represents the full gene repertoire of a species, split into core genes shared by all individuals and accessory genes present in only some. Beyond gene content, the same study also catalogued copy-number variants (CNVs), subtelomeric rearrangements, introgressions, and Ty-element variation across industrial environments (Dunn et al., 2012). High-coverage Illumina sequencing and *de novo* assembly of 14 *S. cerevisiae* isolates spanning the known lineage diversity (alongside 13 *S. paradoxus* for cross-species comparison) revealed extensive CNVs within *S. cerevisiae*. This variation exceeded even that observed in the more nucleotide-divergent *S. paradoxus*. Loss-of-function variants were strongly enriched in subtelomeric regions and concentrated in genes mediating interaction with the environment, including sugar and metal transport or flocculation. This linked subtelomeric plasticity directly to environment-sensing functions (Bergström et al., 2014). The “100-genomes” resource (composed of 93 newly sequenced single-spore segregants along with 7 previously sequenced strains) was then assembled to investigate the potential opportunistic pathogenicity of *S. cerevisiae*, with 43 drawn from clinical origins (Strope et al., 2015). The project sequenced segregants rather than the parental isolates and therefore this involved trading ploidy and heterozygosity for easier assembly. Chromosomal rearrangements provided population-level insights; for example, the *ECM34-SSU1* reciprocal translocation (chrVIII-chrXVI) recurring in 12 strains (Pérez-Ortín et al., 2002; Strope et al., 2015). More region-specific surveys were carried out with for example the analysis of nine variable microsatellite loci in 172 strains across New Zealand (Goddard et al., 2010) or the characterization of *S. cerevisiae* isolates from cacao and coffee fermentation niches (Ludlow et al., 2016).

(ii) Domestication histories

These larger panels made it possible to explore the domestication histories of the species. If wine yeast was domesticated alongside the development of viticulture, the wild ancestor should be found around the Mediterranean basin. Integration of newly sequenced Mediterranean oak isolates identified this population as the closest wild relative of the Wine/European clade (Almeida et al., 2015). Demographic modeling under a population-growth scenario with asymmetric migration placed the wild-domesticated split between 1,300 and 10,300 years ago. It also inferred a one-way gene flow from oak into wine an order of magnitude stronger than the reverse, consistent with continued introgression from wild oak populations into the cultivated lineage (Almeida et al., 2015). The three horizontally transferred regions first discovered in EC1118 (Novo et al., 2009) (totaling 120 kb and 39 predicted genes acquired from other yeast species) were present in 31 of 40 wine isolates but in none of the 25 Mediterranean oak isolates surveyed, marking these acquisitions as a clear hallmark

of domestication. Among them is a tandem pair of *FOT* oligopeptide transporter genes from *Torulaspora microellipsoides*. Loss of this pair lowers biomass by 12% and leaves residual fructose at the end of fermentation (Marsit & Dequin, 2015; Becerra-Rodríguez et al., 2020) (Fig. 5a). Two further signatures recur across wine strains: a reciprocal translocation between chromosomes VIII and XVI that places *SSU1* under the *ECM34* promoter and confers sulfite resistance, selected through the use of sulfite as a wine preservative (Pérez-Ortín et al., 2002; Marsit & Dequin, 2015), and an amplification of *CUP1* whose copy number ranges from zero to fourteen across 132 wine strains (68% carrying more than one copy), and aligns with copper tolerance, a pattern attributed to the use of copper sulfate as a vineyard fungicide (Fogel et al., 1983; Steenwyk & Rokas, 2017).

In beer, multiple studies framed domestication. Sequencing 157 industrial strains spanning beer, wine, sake, and other origins partitioned them into five sublineages (Wine, Beer 1, Beer 2, Mixed, Asia) and dated the Beer 1 common ancestor to around the year 1600, predating the discovery of microbes by more than two centuries (Gallone et al., 2016). The Beer 1 clade also carried the most pronounced hallmarks of domestication, such as pervasive aneuploidy, numerous CNVs, amplification of the *MAL* locus, loss-of-function mutations in *PAD1/FDC1* (a subtelomeric cluster on chromosome IV whose loss abolishes the clove-like 4-vinylguaiacol phenolic off-flavor undesirable in most beer styles), and reduced sporulation and spore viability (Fig. 5d). This was consistent with prolonged propagation in nutrient-rich brewery wort. Furthermore, a parallel survey of 90 strains across top-fermenting beer, wine, bread, and sake confirmed that ale yeasts are polyphyletic and more than twice as diverse as the wine clade, with the same *PAD1/FDC1* and *MAL* signatures recurring as convergent hallmarks of beer adaptation across brewing lineages (Gonçalves et al., 2016). Norwegian kveik (farmhouse strains maintained by backslipping) form a tetraploid lineage related but genetically distinct from Beer 1. They carried the canonical *PAD1/FDC1* and *MAL* signatures alongside the ability to grow at high temperatures: pitched into wort at 28 to 40 °C, 19 of 25 kveik isolates grew at 40 °C, a temperature tolerated by only one of four control ale strains (Preiss et al., 2018). Later, a population-genetic analysis of 542 genomes traced the beer populations to a within-species admixture of a Wine/European-like lineage with an Asia/sake-like lineage (Fay et al., 2019). The same pattern was recovered in the *S. cerevisiae* sub-genome of allotetraploid lager strains, and polyploidy appears to have acted as a passive reproductive barrier that locked the ancestral admixture in place (Fay et al., 2019). Collectively, the beer studies established a view of domestication based on convergent signatures: the *PAD1/FDC1* loss and *MAL* amplification recurred across independently brewed lineages, identifying these as the molecular fingerprints of brewing-niche adaptation.

The picture was completed by the sequencing of 106 commercial wine starter strains from nine international suppliers, alongside 13 brewing strains and 93 wine isolates spanning 7 decades of industrial propagation. It revealed pervasive inbreeding within production lineages and near-identical commercial strains from independent manufacturers (Borneman et al., 2016).

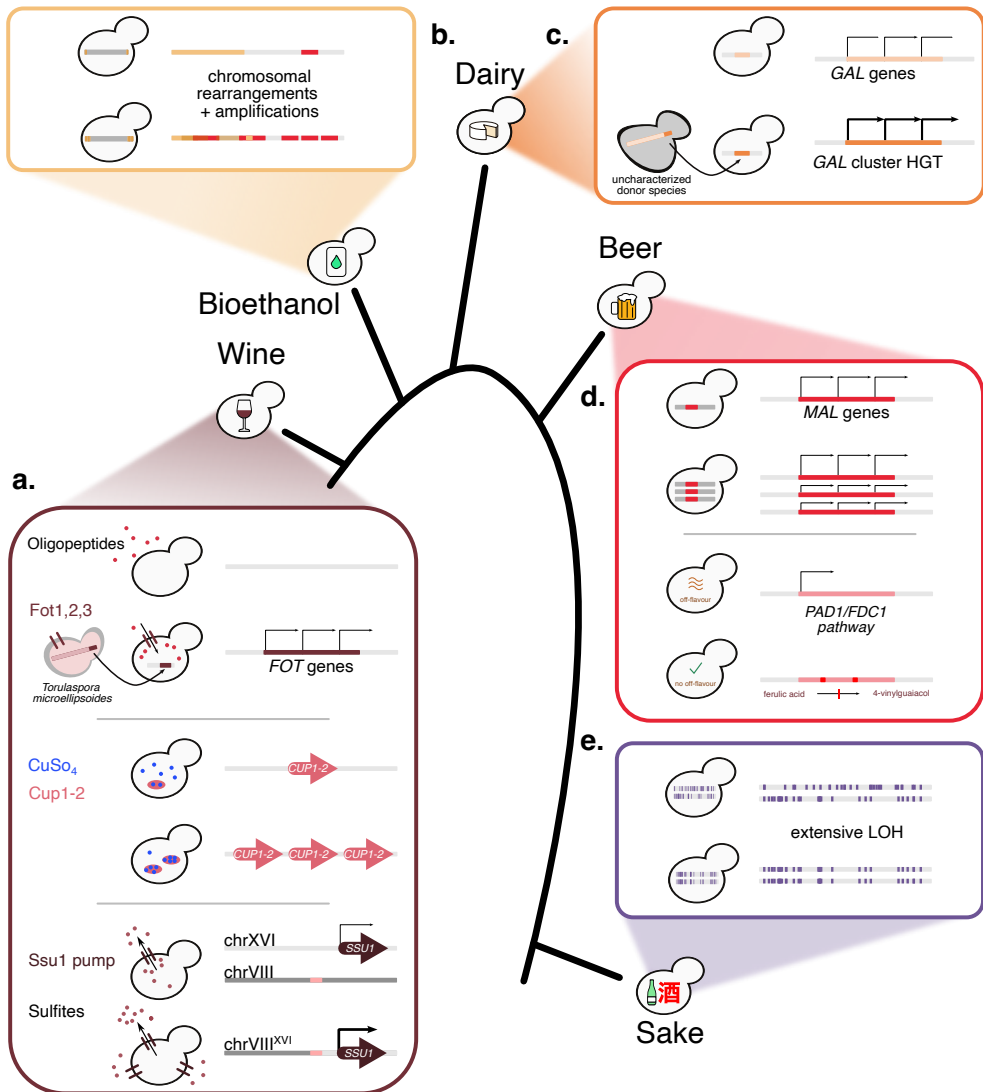


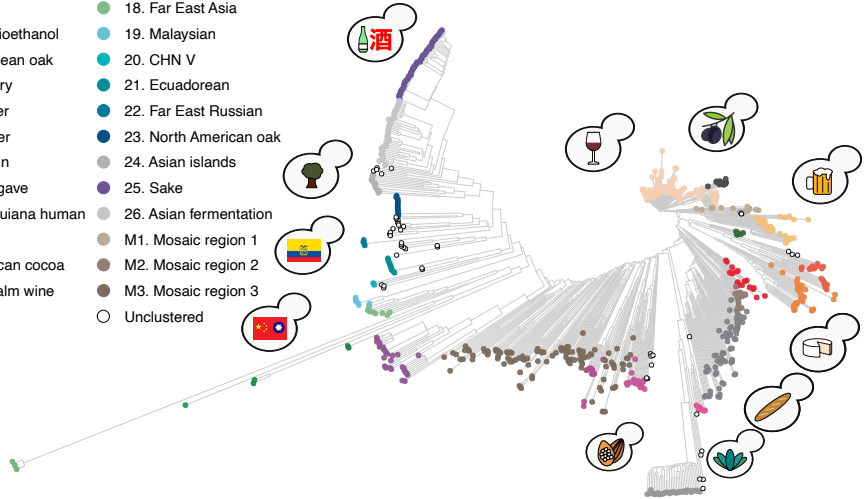
Figure 5 | Genomic signatures of domestication across *Saccharomyces cerevisiae* lineages.

Schematic phylogeny of major domesticated clades, labeled by environment (Wine, Bioethanol, Dairy, Beer, Sake), with characteristic adaptive features shown in the surrounding boxes. **a.** Wine: horizontally acquired *FOT1,2,3* oligopeptide transporters from *Torulaspora microellipsoides*, *CUP1-2* amplification (copper tolerance), and a chrVIII–chrXVI translocation driving *SSU1* overexpression (sulfite efflux). **b.** Bioethanol: chromosomal rearrangements and subtelomeric amplifications. **c.** Dairy: horizontal transfer (HGT) of a *GAL* gene cluster from an uncharacterized donor. **d.** Beer: *MAL* gene amplification, polyploidy, and *PAD1/FDC1* loss-of-function abolishing phenolic off-flavor. **e.** Sake: extensive loss of heterozygosity (LOH).

Cheese fermentation also revealed another niche-specific domestication trajectory. Sequencing of 82 strains placed dairy isolates in a distinct Mediterranean-oak-derived cluster dispersed across France, Morocco, and Kazakhstan, revealing a signature of niche specialization (Fig. 5c). Indeed, these strains are characterized by a loss of the native *GAL1-GAL7-GAL10-GAL2-HXT1-HXT4* locus and by the horizontal acquisition of a replacement *GAL7-GAL10-GAL1* cluster and a high-affinity *GAL2* from a basal *Saccharomyces* donor (Region T), the replacement orthologs being more closely related to those of *S. uvarum* and *S. eubayanus* than to any *S. cerevisiae* allele. This transfer let cheese-associated strains grow on galactose faster than any other surveyed lineage (Legras et al., 2018). HGT in *S. cerevisiae* is therefore not a wine-niche curiosity but a general mechanism of specialization, with an entire gene family (*GAL*) replaced. The origin of this alternative pathway is, however, contested: demographic modeling of the same alleles rejected a recent introgression, dating their divergence from the reference alleles to approximately three billion generations (before the most recent common ancestor of the genus) and attributing their persistence instead to multilocus balancing selection acting on a trade-off between faster growth on galactose and slower growth on glucose (Boocock et al., 2021).

(iii) Global population structure

In parallel, sampling went deeper into East Asia. A survey of 266 isolates collected across Chinese provinces (106 wild from primeval forests and other wild habitats; 160 from a wide range of fermentation sources) yielded two sharply distinct pools. The wild strains clustered into ten geographically structured lineages: nine CHN clades (CHN-I through CHN-X, with VI/VII as a merged lineage) and a Wine lineage that included four Chinese fruit and orchard isolates clustering with European ones (Duan et al., 2018). Moreover, the CHN-IX clade was reported as the most basal *S. cerevisiae* lineage. On the other hand, domesticated strains clustered into twelve Chinese lineages organized by fermentation mode: ten solid-state (Mantou 1-7, Baijiu, Huangjiu, Qingkejiu) and two liquid-state (Milk and Active Dry Yeast). The contrast between wild and domesticated strains was massive: wild isolates were almost entirely homozygous and retained ancestral life-history traits (sporulation ~60%, spore viability ~95%, nucleotide diversity $\pi \sim 8 \times 10^{-3}$), whereas domesticated isolates showed ~40-fold higher heterozygosity, a sharply reduced sporulation rate (~14%) and spore viability (~19%), as well as halved nucleotide diversity ($\pi \sim 4 \times 10^{-3}$). On top of that, domestication shaped CNVs, including an expansion of the *ARR* arsenite-resistance cluster and a *GAL7-GAL10-GAL1* introgression shared by the Milk lineage and the wild lineage CHN-III. Four Chinese isolates also clustered inside the European Wine lineage, and the wine-associated HGT cassettes were shared between European and Chinese strains. Furthermore, the diversity of domesticated populations within China exceeded that of any other region surveyed. On this basis, the authors proposed a Far East Asian origin for domesticated *S. cerevisiae*, arguing that the European Wine lineage was itself imported from Asia rather than domesticated from the Mediterranean oak population (Duan et al., 2018). The East Asian origin is also consistent with the archaeological record. Indeed, Neolithic Chinese potteries often carry chemical signatures of grain-based alcohol fermentation, predating even the earliest evidence of Mediterranean wine (Liu et al., 2019).



Neighbor-joining tree built from genome-wide biallelic SNPs. The collection resolves into 26 clades and three mosaic groups (M1-M3), named after the geographic origin or ecological niche of their defining isolates. Adapted from (Peter et al., 2018) and (Caudal et al., 2024).

23

dataset became one of the most comprehensive population resources available and remained the reference framework for yeast population genomics until its recent expansion.

(iv) *Structural complexity and introgressions*

Introgressions are pervasive across the species. The Alpechín clade, composed of strains isolated from olive-mill wastewater, carries the largest amount of *S. paradoxus*-derived sequence and was subsequently traced to an ancestral *S. cerevisiae* × *S. paradoxus* hybrid. Indeed, a clonal “living ancestor” that still retained the contiguous parental subgenomes revealed that the Alpechín introgressions arose through more than a hundred mitotic recombination events. These events converted heterozygous regions into homozygous blocks and restored meiotic recombination, rescuing fertility in the otherwise sterile interspecific hybrid (D’Angiolo et al., 2020). This gene flow into *S. cerevisiae* is not restricted to the sister species. Hybrids were identified among wine and cider production strains (Masneuf et al., 1998). Cool wine regions of France turned out to host *S. bayanus* var. *uvorum* at a high frequency (Naumov et al., 2000). Another study identified *S. cerevisiae* × *S. uvarum* hybridization events recurrent across successive years in a single Alsace cellar (Le Jeune et al., 2007). Many commercial beer and wine strains, used for fermentation at cool or cold temperatures, are *S. cerevisiae* × *S. kudriavzevii* hybrids that selectively retained *S. kudriavzevii* alleles associated with cold adaptation (González et al., 2008; Peris et al., 2012). The same pattern is present in lager brewing strains. *S. pastorianus*, an allotetraploid hybrid of *S. cerevisiae* and the cryotolerant *S. eubayanus* (a wild yeast that was only discovered in 2011 in forests from Patagonia), is widely used for cold-temperature brewing (Libkind et al., 2011). Across the population, *S. cerevisiae* genomes are often mosaics in which entire chromosomal segments were contributed by other species, traces of the complex evolutionary mechanisms acting on the individuals.

Increasing sample sizes to the thousands allowed demographic history, domestication, admixture between lineages, and adaptation to specific niches to be reconstructed from genetic data rather than inferred from prior ecological assumptions. The domestication histories of wine, beer, cheese, and Asian fermentation were resolved through distinct convergent signatures (*PAD1/FDC1* pathway defects, *MAL* amplification, *GAL* replacement, polyploidy, loss of heterozygosity (LOH)), and the ancestral reservoir of the species was relocated to East Asia. By the late 2010s, short-read platforms had expanded the sampled diversity from one genome to more than a thousand, but large parts of the genome remained refractory to assembly. Repetitive and subtelomeric regions and large structural variants (SVs) were difficult to detect with short reads. Resolving these features required long-read sequencing and, ultimately, telomere-to-telomere (T2T) assemblies.

PART III. The pangenome era

Long reads and structural variation

Short reads defined the SNP diversity across the species, but long reads would help reveal the full extent of the genome architecture. Long-read sequencing addressed the short-read blind spot through two distinct chemistries, with such impact that Nature Methods named long-read sequencing the Method of the Year for 2022 (Nature Methods, 2023). The first one, Pacific Biosciences' (PacBio) single-molecule real-time platform (SMRT) immobilized individual DNA polymerases at the bottom of zero-mode waveguides (Fig. 7a). These are wells on the nanometer scale, smaller than the wavelength of visible light, which restrict the optical excitation volume to the polymerase active site. When nucleotides are incorporated, a fluorescent flash is released and detected (Fig. 7b). The fluorophore, which is carried on the terminal phosphate, is lost upon catalysis, leaving a native DNA strand behind (Eid et al., 2009). Reads of 10 to 30 kb are routine with this technology, and the same molecule can be sequenced multiple times by circularizing it; the resulting consensus delivering accuracies above 99.8% for high-fidelity (HiFi) reads (Wenger et al., 2019).

Oxford Nanopore Technologies (ONT) took a different route, not relying on synthesis. Proposed by Deamer in the 1980s and first demonstrated by Kasianowicz in 1996, the principle is straightforward (Kasianowicz et al., 1996). The MinION Access Programme brought a portable, USB-powered sequencer to the research community in 2014 (Jain et al., 2016). A protein nanopore (α -hemolysin from *Staphylococcus aureus*, then MspA from *Mycobacterium smegmatis* and eventually engineered *E. coli* CsgG variants) is embedded in a synthetic membrane and a voltage difference is applied across the membrane. A single strand of DNA is ratcheted through by a motor enzyme (phi29 DNA polymerase in the founding implementation). The nucleotides passing through (usually 5 or 10 being sensed at a time, depending on the chemistry and pore type) modulate the ionic current in a sequence-specific manner (Jain et al., 2016, 2018) (Fig. 7c). These ionic variations are recorded 5,000 times a second on modern Nanopore flowcells and are then turned into base calls using a neural network, with the latest base-callers being Guppy and Dorado (Fig. 7d). Because the electrical signal is analyzed in real time, the voltage across individual pores can also be reversed to reject selected molecules, enabling computational enrichment or depletion through adaptive sampling (Payne et al., 2021). The first ONT MinION-based assembly of a human genome demonstrated the platform, with ultra-long reads (a read-length N50 above 100 kb, meaning that half of the sequenced bases were contained in reads longer than 100 kb, and individual reads up to 882 kb) (Jain et al., 2016, 2018). Furthermore, native DNA and native RNA can both be sequenced, base modifications are detectable, and read length is theoretically only limited by the input fragment, with multimegabase reads being routinely reported (Zhang et al., 2024). Nanopore reads were once noisy, with error rates of 5 to 10%. Advances in flowcell design and chemistry have reduced them below 1%, while deep-learning methods can improve accuracy 100-fold (Stanojević et al., 2026).

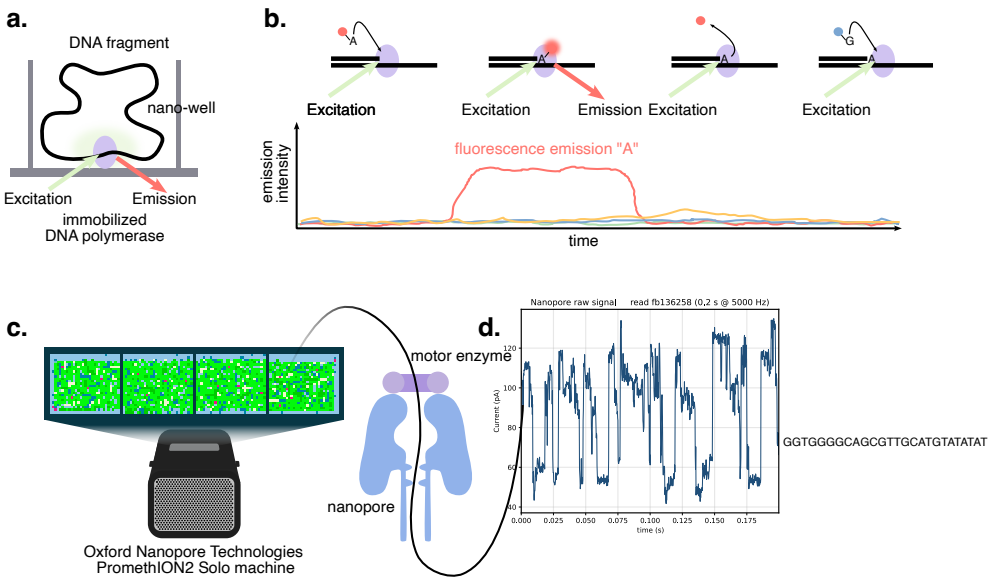


Figure 7 | Third-generation long-read technologies: PacBio and Nanopore.

a. PacBio single-molecule real-time sequencing. A single DNA polymerase is immobilized at the base of a zero-mode waveguide (ZMW), whose sub-wavelength diameter confines excitation light around the enzyme; as the polymerase copies the template, each incorporated nucleotide fluoresces and is recorded in real time.

b. Incorporation cycle: a fluorescently labeled nucleotide is held in the active site during base pairing and emits a pulse of light on excitation (here, "A"); the dye is then cleaved and diffuses away before the next nucleotide (G) is added. The trace plots emission intensity over time, and each base is called from the color and timing of pulses above the noise floor.

c. Oxford Nanopore sequencing. An applied voltage drives a single DNA strand through a protein nanopore set in a membrane, while a motor protein ratchets it through; thousands of pores are sampled in parallel across the flow cell by the sequencer.

d. Raw nanopore signal from a *S. cerevisiae* × *S. paradoxus* hybrid sequenced in the lab. The ionic current through one pore (picoamperes versus time) varies as successive bases occupy the pore constrictions; this trace is then base-called into a read.

Assembling the species pangenome

(i) *Telomere-to-telomere assemblies of yeast*

PacBio was used first, with a public SMRT dataset generated from a W303-derived strain. The study delivered a highly complete sequencing dataset, with more than 130× coverage and a read N50 of 8.8 kb. Such reads made *de novo* assembly easier than with short reads (Kim et al., 2014). Data were released for five model organisms, showing that reads beyond ten kilobases were sufficient for near-complete *de novo* assembly of small eukaryotic genomes.

Nanopore followed the next year with a *de novo* eukaryotic genome assembly. This assembly increased contig N50 more than tenfold and recovered individual rDNA repeat units, telomeres, and Ty elements that Illumina assemblies would collapse (Goodwin et al., 2015). A subsequent benchmark of ONT, PacBio, and Illumina MiSeq showed that both long-read platforms could reconstruct the sixteen nuclear chromosomes to near-completion (Giordano et al., 2017). The same year, an ONT MinION-only assembly of the industrial and laboratory strain CEN.PK113-7D recovered 15 chromosomes as single contigs and the remaining chromosome as a scaffold (Salazar et al., 2017). Furthermore, the complete mitochondrial contig, the complete 2µm plasmid, and two unplaced subtelomeric fragments were assembled. Together, they added close to 800 kb of sequence and 248 gene annotations absent from the prior short-read scaffold and uncovered a chromosome III–VIII translocation, showing that even classical laboratory strains concealed chromosome-scale rearrangements that short-read methods had failed to detect (Salazar et al., 2017).

In 2017, two studies used long-read sequencing to study population-wide panels rather than individual strains. Nanopore-only *de novo* assemblies of 21 natural isolates drawn from the 1002 Yeast Genomes Project achieved 14-fold higher contiguity than the corresponding Illumina assemblies. They produced a complete cartography of Ty1–Ty5 transposable-element insertions across a diverse panel and catalogued 29 translocations and inversions across 17 of the 21 strains (Istace et al., 2017). In parallel, PacBio assemblies of 7 *S. cerevisiae* and 5 *S. paradoxus* isolates delivered near-T2T sequences for 192 chromosomes. These assemblies allowed for the first systematic definition of the subtelomeric boundaries from synteny conservation and showed contrasting evolutionary dynamics between the two sister species (Yue et al., 2017). The same assemblies precisely localized the *CUP1* copper-resistance array (with one to ten tandem copies across strains) and the subtelomeric *ARR* arsenic-resistance cluster, further cementing subtelomeric structural variation as a direct (and now quantifiable) driver of phenotypic variation.

(ii) Structural variants at population scale

The move to T2T assemblies at the population scale came with the *Saccharomyces cerevisiae* Reference Assembly Panel (ScRAP), which used ONT sequencing to generate T2T assemblies for 142 strains, including phased haplotype assemblies for heterozygous diploid and polyploid isolates (O'Donnell et al., 2023). Across these genomes, approximately 4,800 non-redundant SVs were identified: deletions, contractions, insertions, duplications, inversions, and translocations. A median of 240 SVs per strain established structural variation as a pervasive component of yeast genomes. Moreover, telomere length varied fourfold across isolates, and horizontally transferred regions were found to form new telomeric caps, particularly in domesticated strains.

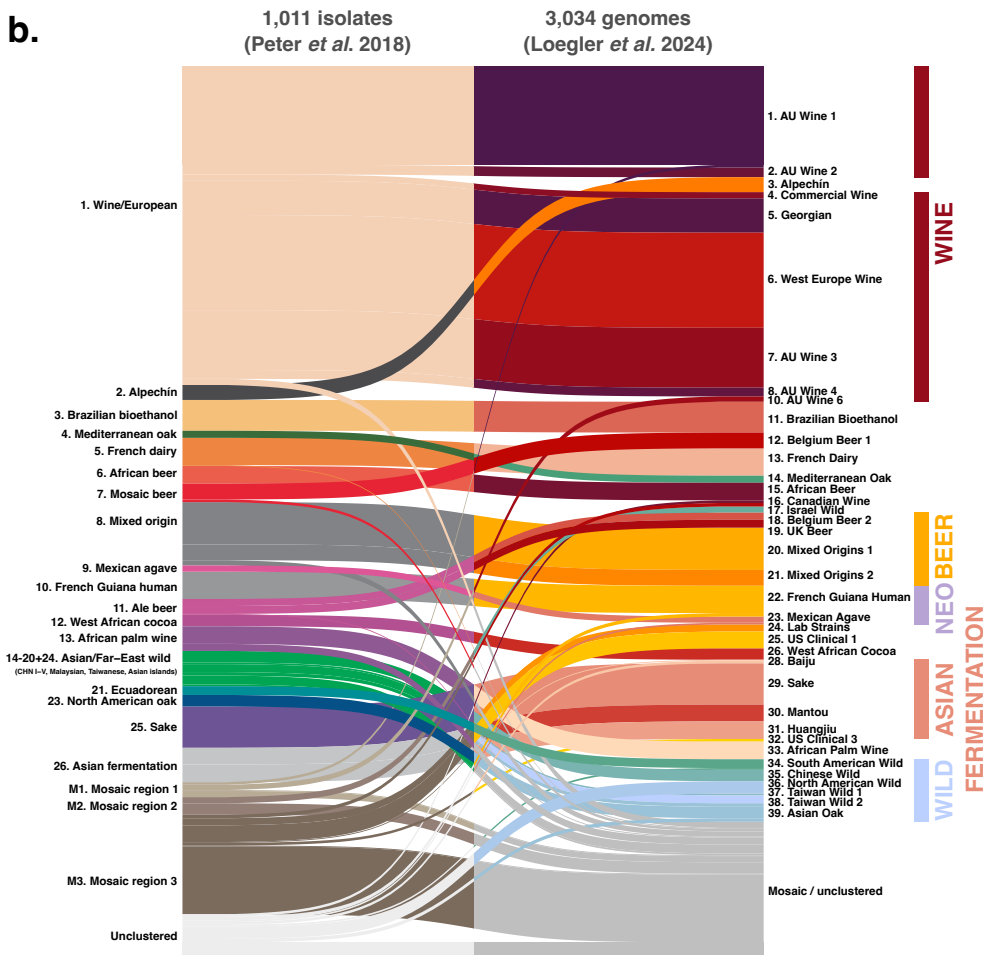
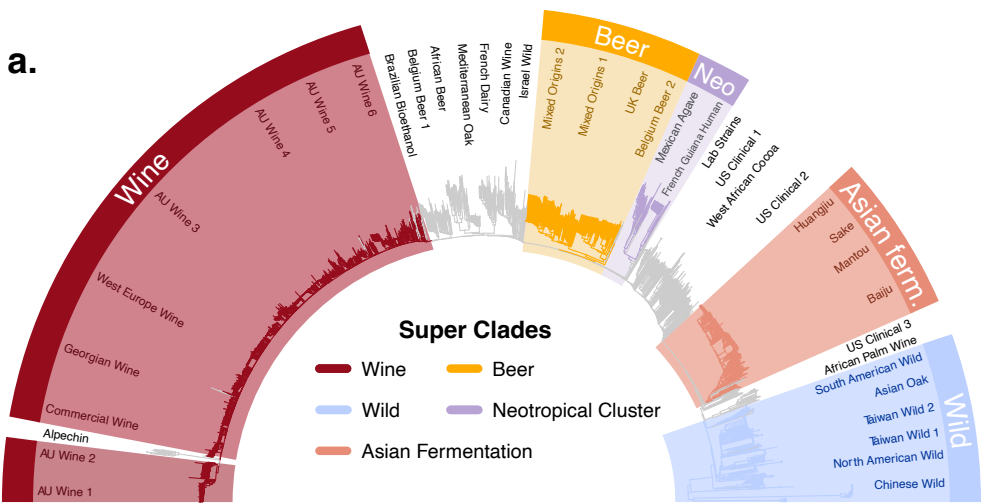


Figure 8 | A five-superclade view of *Saccharomyces cerevisiae* population structure.

a. Phylogeny of *S. cerevisiae* isolates, with the named subpopulations grouped into five superclades: Wine (dark red), Beer (orange), Asian Fermentation (coral), the Neotropical Cluster (purple), and Wild (light blue). The Neotropical cluster, absent from the 2024 survey, has been added to incorporate findings from (Avelar-Rivas et al., 2026). Tip labels give the individual clades within each superclade. Phylogenetic tree adapted from (Loegler et al., 2024).

b. Sankey diagram relating the classifications of the 1,011 isolates in (Peter et al., 2018) and (Loegler et al., 2024). Band width is proportional to the number of isolates.

(iii) A species-wide population census

While long-read assembly was proving its value, short-read data continued to accumulate. A compilation brought together 3,034 genomes from ~30 population-genomics surveys, incorporating data from numerous studies (Borneman et al., 2016; Han et al., 2021; Lee et al., 2022; Ward et al., 2024; Avelar-Rivas et al., 2026) (Fig. 8a). Strains were sampled across 94 countries and were used to define 39 clades, with some reassigned relative to the 2018 survey (Fig. 8b) (Peter et al., 2018). These 39 clades are organized into four superclades: Wine, Beer, Asian Fermentation, and Wild (Loegler et al., 2024). Rarefaction analysis indicated that this collection of more than 3,000 genomes captures more than 90% of the species' non-singleton variants, though the addition of any new deeply divergent population would substantially increase the variant count.

These four superclades reflect the species' main ecological divides. The Wine superclade, the largest, comprises the European wine lineage recognized since the SGRP (Liti et al., 2009) and closely related to Mediterranean-oak wild isolates (Almeida et al., 2015), along with many Australian Wine strains (Ward et al., 2024). The Beer superclade encompasses the ale lineages whose convergent *PAD1/FDC1* and *MAL* signatures were dissected in Part II (Gallone et al., 2016; Gonçalves et al., 2016), with the Mixed-Origins clades reflecting much of the beer–wine admixture (Fay et al., 2019). The Asian Fermentation superclade encompasses the East Asian grain ferments: the sake lineage recognized since 2009, alongside the solid-state baijiu, mantou, and huangjiu clades resolved by deep Chinese sampling (Duan et al., 2018). The Wild superclade contains the deepest-branching, least-domesticated populations and the species' ancestral reservoir, including the Chinese wild clades first identified in 2012 (Wang et al., 2012), the Taiwanese lineages that include the most divergent isolate known (Lee et al., 2022), and the Asian-oak and American wild groups.

Beyond the four superclades, a handful of clades carry signatures distinctive enough to be singled out, being recurrent outliers in the different surveys. Their defining features often include introgressions or structural rearrangements that a single linear reference cannot capture, as well as distinct transcriptomic and/or proteomic signatures (Caudal et al., 2024; Loegler et al., 2024; Teyssonnière et al., 2024b).

The Alpechín clade, already singled out for its hybrid origin, is the clearest case: with roughly 400 introgressed *S. paradoxus* ORFs per strain, the heaviest introgression load in the species (D'Angiolo et al., 2020).

In the Americas, the Brazilian bioethanol clade (Fig. 5b), isolated mainly from sugarcane mills and cachaça fermentations, carries introgressions from American *S. paradoxus* lineages and is characterized by specific amplifications of subtelomeric regions and stress-response genes (Argueso et al., 2009; Jacobus et al., 2021).

A survey sequenced 216 strains from traditional Mexican agave fermentations used to produce tequila and mezcal (Colón-González et al., 2025), revealing two new clades within the Neotropical cluster, a broader phylogenetic grouping that also encompasses lineages from French Guiana, Ecuador, and Brazil (Avelar-Rivas et al., 2026). This Neotropical cluster is shaped by at least three introgression waves from *S. paradoxus* lineages. Patterns of introgression conservation point to multiple hybridization events and suggest more permeable species barriers in this region (Avelar-Rivas et al., 2026).

The wild Ecuadorean clade (South American Wild) sits apart from this domesticated cluster within the deep wild lineages and shows no detectable subtelomeric Y' elements in the 1,011-genome copy-number data (Peter et al., 2018). A recent study reports the same pattern, with no Y' elements detected in the three Ecuadorean strains examined (Hénault et al., 2026).

The West African clades (Cocoa, Palm wine, and African beer) cluster basally within the domesticated groups, a placement consistent with a long, separate history of fermentation on the continent. The Ethiopian honey-wine (t'ej) strains are unusual for their very high heterozygosity (Han et al., 2021). The clinical clades, recovered from human infections, are not one pathogenic lineage but a genetically heterogeneous set whose most consistent feature is ploidy variation: diploids, triploids, and tetraploids, with frequent aneuploidy that correlates with drug-resistance phenotypes (Zhu et al., 2016). Several clinical strains also carry an introgressed copy of *PDR5*, which encodes a multidrug-efflux pump (Strope et al., 2015).

The Sake clade is a low-diversity lineage descended from the Kyokai brewing stock selected by the Japanese brewing societies in the mid-20th century. It is characterized by extensive LOH across its genomes. In 2011, the genome of the sake strain Kyokai no. 7 was resolved as a mosaic of homozygous segments, with repeated LOH in the ancestral brewing lineage covering more than 80% of the genome (Akao et al., 2011) (Fig. 5e).

The French dairy clade comprises cheese isolates that grow better on galactose. They are characterized by an unusually high abundance of Ty1c1 transposable elements (Bleykasten-Grosshans et al., 2021), together with horizontally acquired *GAL* genes (from a species whose nature is still debated, but related to the Saccharomycotina) that are constitutively expressed (Legras et al., 2018; Caudal et al., 2024).

Sampling continued over the following years. A survey of 121 Taiwanese isolates resolved nine lineages, three endemic to Taiwan, including TW1, which clusters with CHN-IX and represents the most divergent *S. cerevisiae* lineage identified to date. This deepened our understanding of the Far East Asian reservoir (Lee et al., 2022). Sequencing of 64 new African isolates showed that African strains form deep-branching

lineages within domesticated groups, implying a long domestication history of yeast in Africa (Han et al., 2021).

Entirely new lineages also keep surfacing. Sequencing 75 strains from British Columbian vineyards resolved four clades, including a newly defined Pacific West Coast Wine group of 34 strains with high nucleotide diversity. This group shares genomic characteristics with wild North American oak yeasts and shows gene flow from Wine/European and Ecuadorean lineages (Marr et al., 2023). The sequencing of 104 genomes from New Zealand traced the colonization of the country to ten independent European founder events (Higgins et al., 2021). Recently, sequencing of 411 Australian spontaneous-ferment isolates revealed an aneuploidy rate of 79% and three independent admixture events involving Oceanian lineages of putatively endemic origin (Ward et al., 2024). Sampling of the species' diversity therefore remains far from saturated, particularly in undersampled regions, highlighting the ever so important role of local sampling in uncovering new lineages and introgression events.

(iv) Graph pangenomes and the species reference

The T2T effort culminated in the generation of 1,482 near-T2T assemblies from 1,086 natural isolates, in which 97% of the chromosomes were resolved in single contigs (Loegler et al., 2025). This dataset revealed 6,587 unique structural variants: 4,755 presence-absence variants (PAVs), 1,207 copy-number variants, 231 inversions, and 394 translocations, with Ty elements spanning more than 50% of the sequence in 39% of PAVs (Fig. 9a). The gene-based pangenome comprised 8,541 gene families: 5,047 core genes shared by all isolates and 3,494 accessory genes. The accessory genes were further divided into 1,263 soft-core, 2,102 dispensable, and 129 private genes (Fig. 9b). The bulk of the accessory genome consisted of 1,233 introgressed genes, 516 fast-evolving genes, 358 acquired by horizontal transfer, and 92 putative *de novo* genes. A graph pangenome, constructed from 500 high-quality haplotypes, encoded this diversity in a framework no longer restricted to a single linear reference and captured over 97% of low-frequency and nearly 99% of common SNPs. Together, these near-T2T assemblies brought to near-completion the effort begun by the 2009 surveys, delivering a species-wide catalogue of genetic variation, from SNPs to large SVs, together with a gene-based pangenome estimated to capture 99.5% of the species' gene content (Loegler et al., 2025).

That pangenome is already being used for analyses. Combining 38 newly isolated North American sourdough strains with 2,908 previously available genomes resolved six superclades (Ruffieux et al., 2026). These superclades were organized around two primary domestication hubs: an Asian solid-state fermentation hub encompassing sake, baijiu, mantou, and sourdough, and a Mediterranean liquid-state fermentation hub spanning the Wine, Mediterranean Fermentation, and derived Baking-and-Brewing superclades (Ruffieux et al., 2026). Sourdough yeasts cluster with the Asian solid-state lineages but not with European-derived commercial baking strains, which group instead with European beer subclades within Baking-and-Brewing, consistent with earlier evidence for two distinct breadmaking trajectories (Bigey et al., 2021). Furthermore, a phylogeny based on pangenome gene presence/absence recovered the structure of

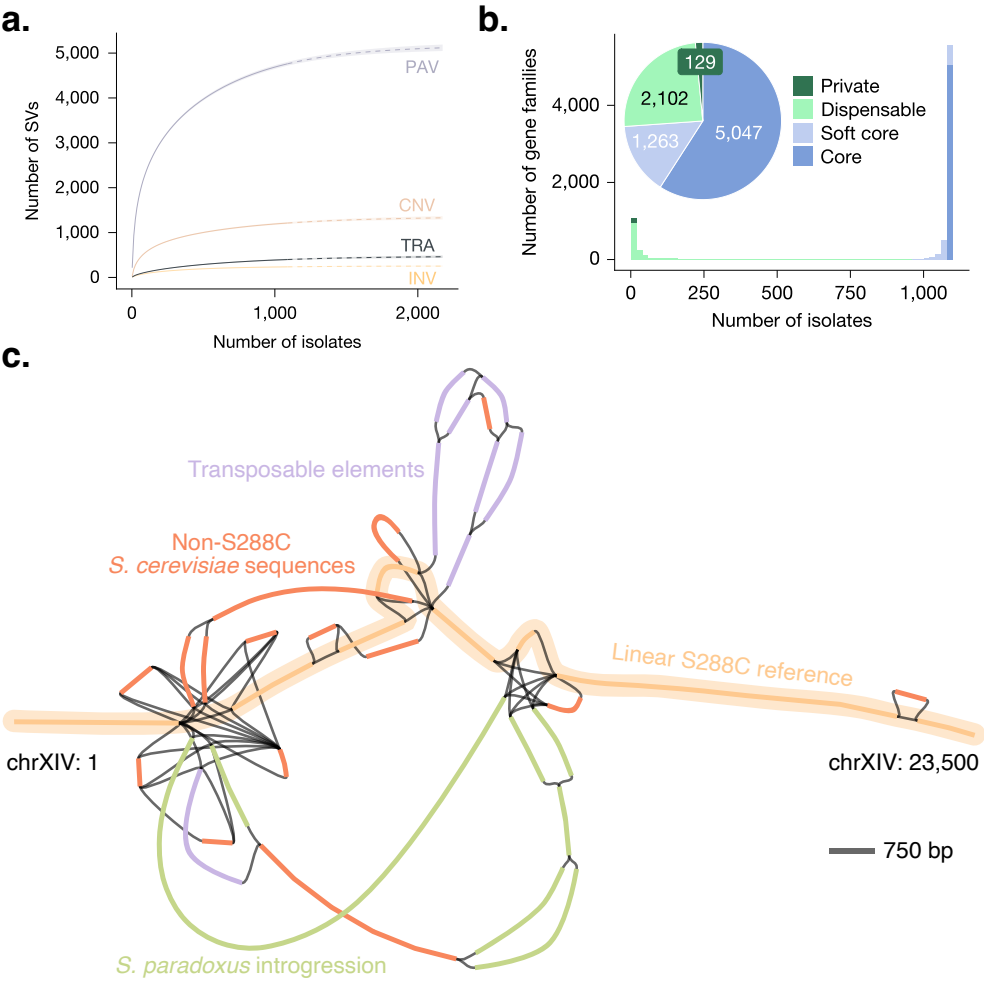


Figure 9 | A graph pangenome resolves variation invisible to a single linear reference.

a. Accumulation curves for four SV classes as the number of sampled isolates increases: presence-absence variants (PAV), copy-number variants (CNV), translocations (TRA) and inversions (INV); adapted from (Loegler et al., 2025).

b. Gene-based pangenome. The species contains 5,047 core and 3,494 accessory genes; adapted from (Loegler et al., 2025).

c. Graph representation of chromosome XIV positions 1 to 23,500 built from T2T *Saccharomyces cerevisiae* assemblies (Loegler et al., 2025); alternative nodes capture sequence divergent from or absent from S288C: non-S288C *S. cerevisiae* sequences (orange), transposable elements (purple), and a *S. paradoxus* introgression (green). Each isolate follows a distinct path through the structure.

the SNP-based tree, further indicating that human cultural practices shape the species (Ruffieux et al., 2026). The copy number of the *MAL33*-like YX00354 gene (Loegler et al., 2025) correlated with maltose growth rate, showing a concrete link between accessory gene variation and fermentation phenotype (Ruffieux et al., 2026). This highlights the necessity of species-wide catalogues.

Variation is not just nucleotide variation. Genome structure itself (including inversions, translocations, copy-number variation, accessory gene content, and telomeric caps formed by horizontally transferred cassettes) varies across natural isolates and is biologically important (O'Donnell et al., 2023; Loegler et al., 2025). These structural features account for a meaningful fraction of phenotypic variance that is invisible to a SNP-only view. A graph pangenome now provides a less reference-biased representation of much of this diversity (Eizenga et al., 2020; Loegler et al., 2025). Thus, three decades after the S288C reference, the field is slowly transitioning from a single genome to a structured pangenome of thousands of isolates, resolved at both the nucleotide and structural levels. The technological trajectory from Sanger sequencing to microarrays, short reads, and finally long reads has paced each successive wave of insights, with each wave exposing a class of variation invisible to its predecessors. The next step is to turn this catalogue of diversity into a systematic predictive understanding of how genotype determines phenotype.

PART IV. From variation to mapping and function

Parts I to III followed the assembly of a species-wide catalogue of variation, from the single S288C reference to a structured pangenome of thousands of isolates capturing variation at both the nucleotide and structural levels (Loegler et al., 2024, 2025). Such a catalogue, however, says very little on its own about function (Nowak, 1995). The natural next step is to map this variation directly to quantitative phenotypes: which variants shift a trait? If they do, by how much?

Part IV first defines the genotype–phenotype relationship and the architecture of quantitative traits. It then turns to the surveys of the molecular layers that sit between genotype and phenotype and the assays used to measure them. Next, it reviews the mapping strategies that can connect variants to traits (from single crosses to genome-wide association and multi-parent designs), and finally turns to integrating these layers and building a species-wide functional toolkit.

Conceptual foundations of genotype–phenotype mapping

The relationship between genotype and phenotype is a central question in genetics, and might be one of the most complex in biology. From Gregor Mendel's pea plants to the modern attempts at building a genotype-to-phenotype map, the discipline has been shaped by the effort to link observable traits (phenotypes) to heritable factors (genotypes) (Mendel, 1865; Morgan, 1911; van Dijk et al., 2022; Stenseth et al., 2022). The specific vocabulary emerged at the turn of the twentieth century. After Mendel's work had been largely forgotten for 35 years, his segregation laws were independently rediscovered by the botanists Carl Correns, Erich von Tschermak, and Hugo de Vries (Rheinberger, 2008). Soon after, Bateson coined the term genetics, and Johannsen introduced the terms gene, genotype, and phenotype. The chromosomal theory of inheritance followed. Thanks to cytological work in sea urchins and grasshoppers, it was established that chromosomes segregate according to Mendel's laws and must therefore carry heritable factors (Hegreness & Meselson, 2007; Laubichler & Davidson, 2008).

For much of the twentieth century, genes underlying simple traits could be mapped through pedigrees and crosses. But the genetic basis of quantitative variation in complex traits (which encompass most traits) remained inaccessible. The conceptual framework had nevertheless been established early by Fisher's infinitesimal model. In this model, continuous traits arise from the cumulative effects of many independently segregating loci (Fisher, 1918, 1919; Visscher & Bruce Walsh, 2019). Work in *D. melanogaster* then connected Mendelian inheritance to chromosomal linkage (Morgan et al., 1915), and a decade later, Muller demonstrated that X-ray exposure could induce mutations (Muller, 1927). Genes were later linked to biochemical activity through the one-gene-one-enzyme experiments in *Neurospora* (Beadle & Tatum, 1941). DNA was subsequently established as the hereditary material in bacteria and bacteriophages (Avery et al., 1944; Hershey & Chase, 1952). X-ray diffraction data from Franklin and Gosling, together with related work by Wilkins, Stokes, and Wilson, then contributed

to the elucidation of DNA's structure, presented by Watson and Crick as a double helix (Franklin & Gosling, 1953; Watson & Crick, 1953; Wilkins et al., 1953).

From that point onward, the genotype–phenotype relationship could be framed as a chain of measurable biochemical events rather than an abstract mapping. DNA is transcribed into mRNA, which is translated by ribosomes into protein. Crick formalized these transfers of sequence information (and the prohibition of transfer from protein back to nucleic acid or protein) as the central dogma (Crick, 1958, 1970). The familiar DNA-to-RNA-to-protein dogma was later expanded by reverse transcription, while recent reports of protein-templated DNA synthesis may represent a genuine challenge to the dogma (Mrnjavac & Martin, 2026; Song et al., 2026). Even so, it remains an organizing principle of molecular biology. An individual could now be described through its genome, whose nucleotide sequence shaped, layer after layer, its RNA, protein, and metabolite pools, ultimately contributing to cellular and organismal phenotypes (Buccitelli & Selbach, 2020).

Genotype-to-phenotype studies most often measure organismal-level traits, including morphology, physiology, behavior, and disease susceptibility. Yet, variation at this level often stems from differences at the underlying intermediary molecular layers. In turn, understanding the genotype-to-phenotype relationship requires insight into these regulatory layers. The abundance of each transcript or protein, for example, can therefore be considered a phenotype in its own right (Brem & Kruglyak, 2005). This layered view also turns the genotype-to-phenotype mapping question into a problem of measurement. Each layer requires its own assays, and the development of those assays would pace the field over the next half-century.

What are traits, really?

Observable traits can be placed on a continuum between two limit cases. On one side sit Mendelian traits (Fig. 10a), where a single locus partitions the population into discrete classes (Mendel, 1865). At the other end of the spectrum lies the limit of infinite polygenicity (Fig. 10b), in which the cumulative effects of thousands of small-effect variants produce an approximately normal phenotypic distribution (Fisher, 1918, 1919; Visscher & Bruce Walsh, 2019). As of August 7, 2026, the Online Mendelian Inheritance in Man (OMIM) catalogue (Amberger et al., 2019) listed 7,157¹ phenotype descriptions with a known molecular basis, with cystic fibrosis and Duchenne muscular dystrophy as familiar examples. Among nonpathological traits, the ABO blood-group system, classically described by three main alleles at a single locus on chromosome 9, is a familiar Mendelian example (Yamamoto et al., 1990). Unsurprisingly, things are not always that simple. Most quantifiable traits lie toward the polygenic end of the spectrum. Autism spectrum disorder and human height, for example, are each shaped by hundreds to thousands of small-effect variants, alongside a smaller set of rare, high-impact alleles (Marouli et al., 2017; Yang et al., 2026b). The environment also has a major influence. For instance, a temperature-sensitive tyrosinase allele in Himalayan rabbits permits pigment production only in cooler tissues, so the same allele yields a pale body and dark

¹OMIM Entry Statistics, accessed August 11, 2026.

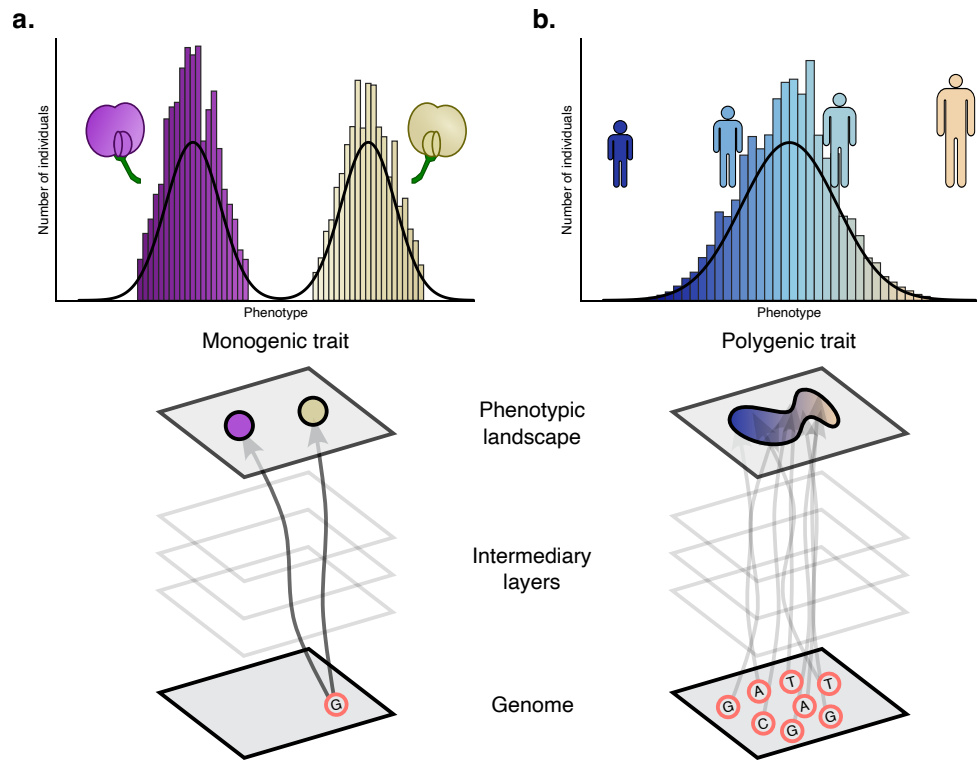


Figure 10 | The two limit cases of trait architecture.

- a.** A monogenic trait controlled by a single locus partitions a population into discrete classes, illustrated here by Mendel's purple and white-flowered peas. The phenotype distribution is bimodal, each phenotypic class forming its own peak.
- b.** A polygenic trait combines many small-effect loci, producing a continuous, approximately normal distribution, illustrated here by human height.

$$\sigma^2_{\text{Phenotypic}} = \sigma^2_{\text{Genetic}} + \sigma^2_{\text{Environmental}} + \sigma^2_{\text{GxE}}$$
$$\sigma^2_{\text{Additive}} + \sigma^2_{\text{Dominance}} + \sigma^2_{\text{Epistasis}}$$

Figure 11 | Partitioning phenotypic variance.

Phenotypic variance is decomposed into genetic, environmental, and genotype-by-environment terms. Genetic variance is further partitioned into additive, dominance, and epistatic contributions: additive variance is the sum of independent per-locus effects, dominance variance captures interactions between alleles within a locus, and epistatic variance captures interactions between loci.



extremities depending on local skin temperature (Kwon et al., 1989; Aigner et al., 2000). However, this simple-to-complex scale appears to be rather fluid. For instance, alleles that segregate cleanly in one yeast cross can produce continuous distributions in another because the penetrance of the same variant changes across genetic backgrounds (Hou et al., 2016). Across natural backgrounds in yeast, gene essentiality itself is similarly context dependent (Bosch-Guiteras & van Leeuwen, 2022), with 39 genes that are essential in the laboratory reference proving dispensable in at least one of 18 diverse natural strains, typically through single modifier loci that indirectly compensate for loss of the essential function (Batté et al., 2026).

Quantifying the relative contributions of these components requires partitioning phenotypic variance. Fisher's infinitesimal model provided the early quantitative-genetic framework for this decomposition (Fisher, 1918, 1919). Phenotypic variance ($\sigma_{\text{Phenotypic}}^2$) can then be partitioned into genetic ($\sigma_{\text{Genetic}}^2$), environmental ($\sigma_{\text{Environmental}}^2$), and genotype-by-environment ($\sigma_{\text{G} \times \text{E}}^2$) terms (Fig. 11). The genetic component itself can be partitioned further into additive ($\sigma_{\text{Additive}}^2$), dominance ($\sigma_{\text{Dominance}}^2$), and epistatic variance components ($\sigma_{\text{Epistasis}}^2$) (Visscher et al., 2008).

"Dominance" and "epistasis" describe interactions within and between loci, respectively. The G×E term captures cases where the same allele acts differently across environments. For instance, biallelic pathogenic variants in the *PAH* gene can cause phenylketonuria, but dietary phenylalanine restriction can largely prevent its severe neurological manifestations (Scriver, 2007).

From this partitioning, we can derive two measures: broad-sense and narrow-sense heritability.

Broad-sense heritability, $H^2 = \sigma_G^2 / \sigma_P^2$, is the proportion of phenotypic variance attributable to all genetic effects, including additive, dominance, and epistatic components (Visscher et al., 2008).

Narrow-sense heritability, $h^2 = \sigma_A^2 / \sigma_P^2$, is the proportion of phenotypic variance attributable to additive genetic effects, the component targeted by most standard association models. In outbred species, these measures have traditionally been estimated indirectly through twin or family designs. Twin studies, for example, place the heritability of adult human height near $h^2 = 0.8$ (Silventoinen et al., 2003).

Yeast is an ideal model because environmental conditions can be tightly controlled. Replicate measurements of genetically diverse strains within controlled conditions allow environmental variance to be estimated and the genetic components to be partitioned more directly. Across yeast panels, the additive component consistently dominates, though its share varies with the design: from 43% in haploid segregants, where dominance cannot arise and pairwise epistasis adds a further 9% (Bloom et al., 2013, 2015), to 73% in phased diploid lines, alongside 10% dominance, 7% second-order epistasis, and 1.7% third-order epistasis (Hallin et al., 2016). Diploid panels additionally capture dominance. In a diallel of hybrids, approximately one-third of phenotypic variance was attributed to non-additive effects, with complete dominance playing a key role (Fournier et al., 2019).

Explaining this variance and mapping it back to specific loci, however, requires a comprehensive catalogue of variants. Genetic variants span a wide range of sizes (Fig. 12). At the smallest scale, SNPs and InDels are the most abundant variant classes and are typically detected using short-read whole-genome sequencing. The 1,011-genome project in yeast catalogued 1,625,809 SNPs and 125,701 InDels (Peter et al., 2018). The 1000 Genomes Project detected 84.7 million SNPs and 3.6 million InDels across 2,504 humans, with each genome differing from the reference at 4.1 to 5 million sites (Auton et al., 2015).

Variation at larger scales includes structural variants, such as large deletions, insertions, duplications, inversions, and reciprocal or non-reciprocal translocations, as well as karyotypic changes such as aneuploidies and whole-genome duplications (Fig. 12). These variants arise through errors in DNA replication or repair, recombination between repetitive sequences, and chromosome missegregation, and are a recurring force in adaptive evolution (Collins & Talkowski, 2025). In *S. cerevisiae*, the *SSU1-R* allele that confers sulfite tolerance to wine strains is the product of a reciprocal translocation between chromosomes VIII and XVI, whose selection was probably favored by the long-standing use of sulfite as a wine preservative (Pérez-Ortín et al., 2002). Population-scale surveys historically focused on short variants because short reads provided limited access to large or complex structural variants. Long-read sequencing has now enabled population-wide catalogues spanning both short and structural variants. Beyond yeast, a long-read survey of 3,622 Icelanders detected a median of 22,636 SVs per individual (13,353 insertions and 9,474 deletions), revealing variation that previous sequencing technologies had missed (Beyter et al., 2021).

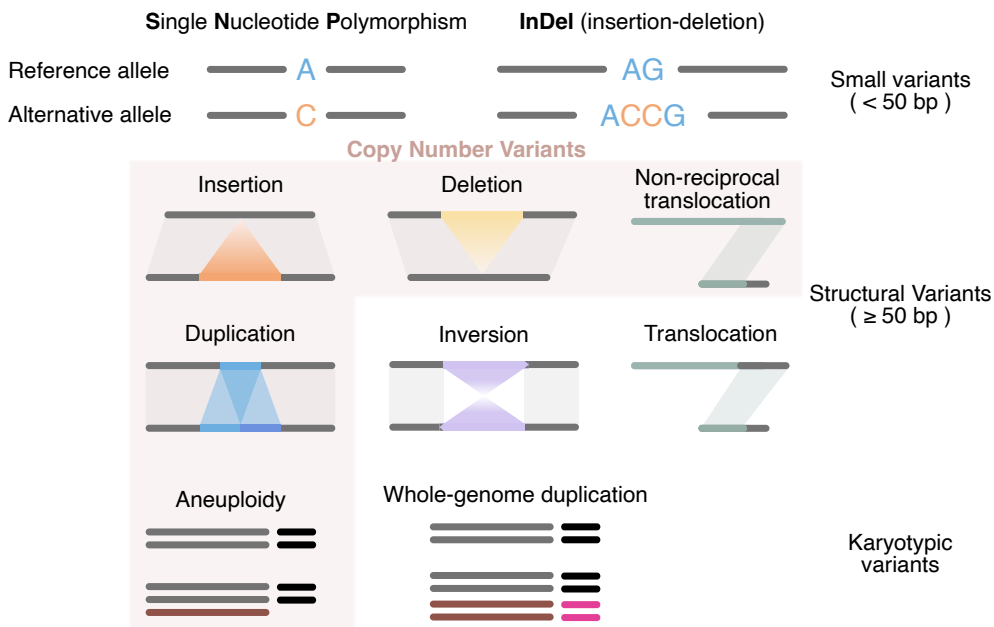


Figure 12 | The size spectrum of genetic variation. Variants span scales from single bases to whole chromosome sets. Small variants (< 50 bp) are single-nucleotide polymorphisms (SNPs) and short insertions–deletions (InDels). Structural variants (≥ 50 bp) include insertions, deletions, duplications, inversions, and reciprocal and non-reciprocal translocations. Karyotypic variants alter whole chromosomes (aneuploidy) or the entire complement (whole-genome duplication). Pink shading marks copy-number variants that can alter dosage. Gray bars denote reference sequence; colored segments mark the variant.

Measuring biological layers

The study of each successive biological layer required its own instruments, and the history of functional genomics is largely one of increasing scale: from measuring one transcript or protein at a time to surveying whole transcriptomes and proteomes, and from profiling one strain at a time to analyzing species-wide panels.

(i) *Detection of transcripts*

Transcripts were usually quantified one at a time using Northern blotting, ribonuclease protection assays, and, eventually, reverse-transcription quantitative PCR (RT-qPCR) (VanGuilder et al., 2008).

Unfortunately, none of these methods scaled past a handful of targets per assay, making multiplexing nearly impossible. DNA microarrays would solve that issue by immobilizing complementary DNA (cDNA) probes on glass slides and detecting hybridization through two-color fluorescence, a principle demonstrated with a 45-gene *Arabidopsis* array (Schena et al., 1995) and readily scaled to thousands of probes. As discussed in Part I, the approach was soon pushed to genome scale in *S. cerevisiae* (DeRisi et al., 1997), querying the entire reference transcriptome and enabling the characterization of the diauxic shift from glucose to ethanol.

Microarrays would be replaced as quickly as they had displaced Northern blots by RNA sequencing (RNA-seq) (Wang et al., 2009). Direct counting of short reads removed the need for predesigned probes. This made it possible to capture novel and unannotated transcripts and to extend the dynamic range by orders of magnitude (Stark et al., 2019).

RNA-seq has been a key technology in modern biology, supporting population-scale transcriptome catalogues across human tissues (GTEx Consortium, 2020), species-scale yeast resources spanning hundreds of natural isolates (Caudal et al., 2024), and structured panels of inter-strain hybrids (Tsouris et al., 2024a, b). Yeast has also been a testing ground for transcriptome-resolution upgrades that conventional short-read RNA-seq cannot deliver. For example, transcript isoform sequencing (TIF-seq) linked the 5' and 3' ends of individual transcript molecules and revealed extensive variation in transcription start and termination sites in yeast (Pelechano et al., 2013).

Short reads, however, often cannot unambiguously resolve full isoform structure. Indeed, from a single genetic locus, distinct transcripts can be produced. They can vary through alternative splicing, alternative transcription start or termination sites, or alternative polyadenylation, for instance. It is known that more than half of annotated human transcripts exceed 2,500 bp (Frankish et al., 2019), and ambiguity at exon junctions often complicates isoform-level quantification (Djebali et al., 2012; Monzó et al., 2025). Long-read RNA sequencing addresses many of these limitations because its reads can span entire transcripts, without the need for fragmentation. As such, PacBio Iso-Seq and ONT cDNA platforms can recover full-length cDNAs and are now widely used for *de novo* transcriptome assembly and isoform annotation (Sharon et al., 2013; Amarasinghe et al., 2020; Monzó et al., 2025).

In addition, ONT is able to sequence native RNA molecules directly, through direct RNA sequencing (dRNA-seq). Because dRNA-seq analyzes native RNA rather than a cDNA copy, modification-dependent signals are retained and can be interrogated for modifications such as m⁶A, pseudouridine, and m⁵C. In turn, this extends transcriptome profiling to the epitranscriptome (Garalde et al., 2018; Workman et al., 2019; Begik et al., 2021; White & Hesselberth, 2022; Kim et al., 2025; Luo et al., 2026). Because the poly(A) tail enters the pore first, direct RNA sequencing can estimate tail lengths for individual molecules and resolve transcript-specific distributions, allowing their dynamic regulation under stress to be tracked (Tudek et al., 2021).

At the same time, technologies are moving toward quantification in individual cells. With single-cell RNA-seq, it is now possible to profile transcriptomes from tens of thousands of individual cells per experiment, enabling transcriptional noise to be assessed. With droplet-based approaches, it is (relatively) straightforward to scale to tens of thousands of cells per run using 3'-end counting with unique molecular identifiers (UMIs) (Macosko et al., 2015; Zheng et al., 2017). Adapting these approaches to yeast, however, requires the cell wall to be weakened or removed by spheroplasting, either before encapsulation or within droplets (Jariani et al., 2020). The technique has since been used in yeast crosses, revealing genetic effects on expression noise that bulk sequencing averages out (Boocock et al., 2025). Single-cell profiling was subsequently extended to the deletion collection using an RNA-barcoded redesign to profile the transcriptomes of more than 3,500 mutants (75% of the nonessential genes) under control and osmotic-stress conditions (Nadal-Ribelles et al., 2025).

(ii) Protein detection

Protein quantification scaled later than transcript quantification. Before modern mass spectrometry (MS)-based proteomics, protein abundance was measured using Western blotting, colorimetric assays, and two-dimensional gel electrophoresis. These methods remained common through the 1990s but were limited in throughput, dynamic range, or reproducibility (Lowry et al., 1951; O'Farrell, 1975; Bradford, 1976; Towbin et al., 1979; Gygi et al., 1999). As previously discussed, proteome-wide abundance and localization catalogues were soon assembled in yeast using TAP-tagged (Ghaemmaghami et al., 2003) and GFP-tagged (Huh et al., 2003) collections.

MS-based proteomics removed the need for strain engineering. Liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) (Aebersold & Mann, 2003) and the move to data-independent acquisition (DIA) (Gillet et al., 2012), followed by deep-learning-based analysis with DIA-NN (Demichev et al., 2020), helped bring proteome coverage closer to transcriptome-scale coverage. It is now possible to perform species-scale proteomics across hundreds of yeast isolates (Messner et al., 2023; Muenzner et al., 2024; Teyssonnière et al., 2024b). These first surveys show that natural aneuploidy is broadly buffered at the protein level, with 70.5% of proteins encoded on aneuploid chromosomes showing dosage compensation and enhanced protein turnover contributing to this buffering (Muenzner et al., 2024).

Beyond yeast, large-scale proteome maps have been assembled for *E. coli* (Schmidt et al., 2016), *A. thaliana* (Mergner et al., 2020), and humans (Wilhelm et al., 2014; Uhlén et al., 2015), with affinity-based blood plasma proteomics now profiling cohorts of tens of thousands of individuals (Sun et al., 2023). Efforts have also begun to map the human “dark proteome.” Across 95,520 proteomics experiments, approximately one-quarter of 7,264 noncanonical ORFs yielded detectable peptides. The same study formalized the new “peptidein” category for confirmed protein molecules of indeterminate functional status (Deutsch et al., 2026). However comprehensive these efforts become, other layers still need to be explored to obtain a fuller picture.

(iii) Other layers can be explored too

Between transcript and protein sits translation. In 2009, ribosome profiling was developed in yeast. Cycloheximide-arrested polysomes are digested with RNase I and the ~30-nt mRNA fragments protected by elongating ribosomes are isolated and sequenced (Ingolia et al., 2009). The positions of translating ribosomes can thus be mapped at single-codon resolution due to the characteristic 3-nt periodicity. In turn, translation efficiency (footprints per transcript) spans nearly 100-fold across genes; in consequence, footprint density predicts protein abundance substantially better than mRNA abundance does (Ingolia et al., 2009). The same approach has been extended to the yeast meiotic program (Brar et al., 2012), to translation-level mapping in a BY × RM cross (Albert et al., 2014), and across eight natural yeast isolates (Teyssonnière et al., 2024a), but no species-scale ribosome-profiling panel has been assembled yet. Metabolomics can also be explored. Once again from a BY × RM cross, 114 segregants were profiled by targeted LC-MS/MS, with 74 compounds retained for linkage analysis (Breunig et al., 2014). Studies keep pushing to finer resolution (Mülleder et al., 2016), with a recent study quantifying 19 amino acids in 27 populations across nine budding-yeast species, most of them *S. cerevisiae* (Tengölics et al., 2024). Other layers, such as chromatin accessibility or post-translational modifications, remain less explored.

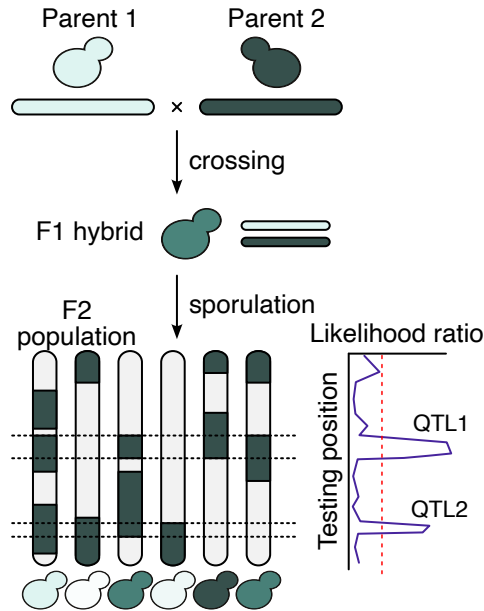


Figure 13 | Linkage mapping in a biparental cross.

Two genetically distinct parents differing in a trait are crossed to produce an F1 hybrid. Sporulation produces an F2 population in which meiotic recombination creates unique mosaics of the parental chromosomes. Each segregant is genotyped at markers spanning the genome and scored for the trait. At each genomic position, the association between local genotype and phenotype is evaluated as a likelihood ratio; values above the significance threshold (red dashed line) mark quantitative trait loci. Peak height reflects the strength of the genotype–phenotype association.

Genetic mapping strategies

With both a variant catalogue and a record of phenotypic variation, the remaining goal is to connect the two: to map specific variants to the phenotypic variation they help produce, where the trait can be organismal or, as with the transcript and protein levels just surveyed, molecular. Complementary strategies have shaped yeast quantitative genetics: linkage mapping in single crosses, bulk-segregant approaches, genome-wide association studies, and multi-parent designs.

(i) Linkage mapping in single crosses

Quantitative trait loci (QTL) are genomic loci that affect a trait and therefore co-segregate with variation in it. QTL mapping extends classical linkage analysis from discrete Mendelian traits to complex ones. It is based on the co-inheritance logic that placed single genes on chromosomes through tetrad dissection. Briefly, two genetically distinct haploid parental strains are mated to form a diploid hybrid (Fig. 13). Afterwards, the sporulation of the hybrid is induced, and meiotic recombination shuffles the two parental genomes into the haploid spores (segregants). Each segregant carries a unique mosaic of its parents. A phenotype of interest is then measured across a few hundred to a few thousand segregants, and the segregants are sequenced. Chromosomal segments

whose parental origin co-segregates with trait value more often than chance across the panel mark candidate QTL.

This technique was successfully employed before whole-genome sequencing, notably using restriction fragment length polymorphisms (Botstein et al., 1980; Lander & Botstein, 1989), and has since been applied across model organisms such as *A. thaliana* and *Caenorhabditis elegans* (Brachi et al., 2010; Rockman et al., 2010). *S. cerevisiae*, having a short sexual generation time, efficient sporulation, and a compact genome, is an ideal candidate for this technique (Lynch, 2006; Liti & Louis, 2012). Indeed, three key factors for QTL mapping are the density of markers along the genome, the size of the offspring population, and the recombination rate. A higher recombination rate implies more crossovers and, in turn, shorter genomic segments, leading to more precise localization of causal variants.

In 2002, QTL mapping was used to dissect the genetic architecture of a complex trait in yeast. The genetic basis of the high-temperature-growth phenotype was dissected using a BY × YJM145 cross (Steinmetz et al., 2002). Two intervals spanning 51.6 kb on chromosome XIV and 8.1 kb on XVI were identified. These intervals were further refined by reciprocal-hemizygosity analysis: in each of two otherwise identical hybrids, one parental copy of the candidate gene is deleted, leaving the other expressed alone. In consequence, any phenotypic difference between the two diploids tested can be attributed to the allelic difference at the candidate gene. Three causal genes were identified, notably *MKT1*, with a 40-fold growth difference between MKT1-YJM145/Δ and MKT1-BY/Δ (Steinmetz et al., 2002).

A parallel study used DNA hybridization microarrays to quantify the expression of thousands of transcripts. This technique enabled high-throughput surveys of transcription regulation, with the abundance of each transcript treated as its own complex trait. From a single BY4716 × RM11-1a cross, 40 haploid segregants were used to map the genetic basis of expression variation. Of 6,215 genes profiled, more than 1,500 were differentially expressed between the parents. Variation in transcript abundance for 570 genes was linked to loci referred to as expression quantitative trait loci (eQTL) (Brem et al., 2002). A block of 28 genes, half involved in heme, fatty acid, or sterol metabolism, implicated the *HAP1* gene as their regulator. BY carries a Ty1 insertion at that locus. Segregants inheriting the BY allele therefore shared a specific transcriptional signature (Brem et al., 2002).

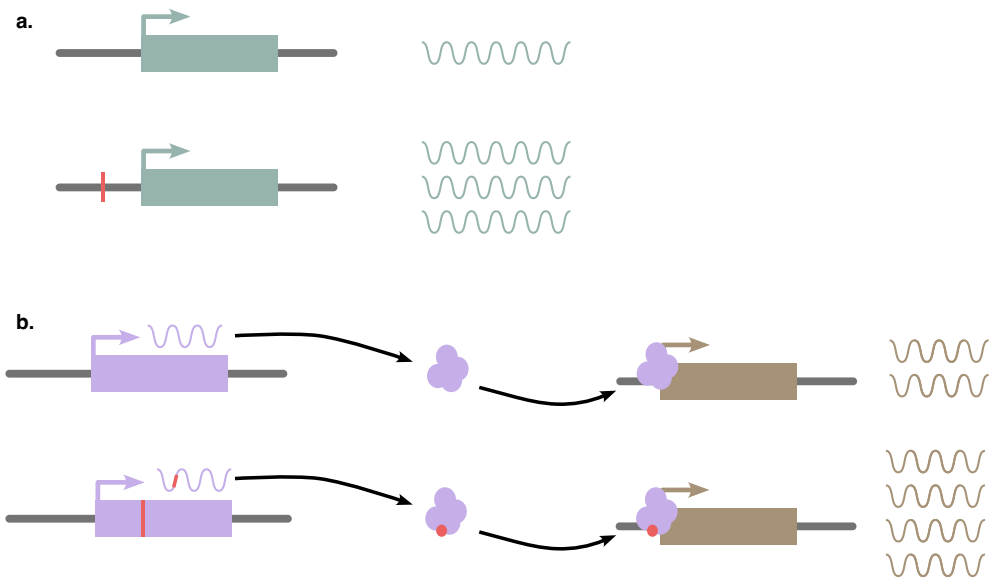


Figure 14 | Cis- and trans-acting regulatory variation.

Schematic of the two mechanisms by which a sequence variant (red) alters gene expression.

a. Cis-acting: a variant affecting a gene locally and in an allele-specific manner increases transcription of that gene (teal), raising transcript output (here from one to three copies). The variant is drawn upstream but can lie anywhere that affects only the linked allele, including within the gene body; cis-acting variants map to the gene they regulate and underlie cis-eQTL.

b. Trans-acting: a variant in a gene encoding a diffusible regulator (purple), such as a transcription factor, alters the regulator and, in turn, the transcription of a distant target gene (brown; here from two to four copies). The variant is drawn within the coding sequence but may instead alter the regulator's abundance or activity through its own regulatory elements. Because the regulator is diffusible, trans-acting variants can influence genes anywhere in the genome and underlie trans-eQTL.

These eQTL can be split into two categories based on their positional proximity to their target genes, with distance thresholds varying across organisms. Local eQTL are in close proximity to the gene they affect. By contrast, distant eQTL, as their name suggests, are located further away and often act through diffusible intermediates such as transcription factors or chromatin modifiers. In diploids and higher-ploidy organisms, a further distinction can be made. Regulation occurs at both local and distant levels, but it can also be separated into *cis* and *trans* components. Although *cis* and *trans* are often used interchangeably with local and distant, respectively, the terms carry different mechanistic meanings (Fig. 14). A *cis*-acting variant affects expression from the allele on the same DNA molecule, potentially leading to allelic expression imbalance. These imbalances have, for example, been linked to disease risk at both the single-locus and genome-wide scales. As an example, at the 1p13 human locus, a noncoding variant creates a CCAAT/enhancer-binding protein (C/EBP) binding site that drives allele-specific hepatic *SORT1* expression, having a large effect on plasma low-density lipoprotein (LDL) cholesterol and myocardial infarction risk (Musunuru et al., 2010). More broadly, cis-eQTL colocalize with a substantial share of hits across hundreds of

complex traits (GTEx Consortium, 2020). A *trans*-acting variant acts through a diffusible regulator and therefore affects both homologous alleles of its target; it does not by itself create allelic imbalance. More often than not, *cis*-eQTL have larger effect sizes than *trans*-eQTL (Võsa et al., 2021).

Expanding the same BY $\times$ RM panel to 112 segregants revealed that most heritable expression traits are highly polygenic. The strongest QTL explained a median of only 27% of trait variance (Brem & Kruglyak, 2005). A few years later, gene expression was profiled in 109 BY $\times$ RM segregants grown in glucose- or ethanol-containing media. Three phenotypes were analyzed for each transcript: (i) expression in glucose, (ii) expression in ethanol, and (iii) the ethanol-minus-glucose difference per segregant as a derived trait. This allowed G $\times$ E eQTL (genotype-by-environment eQTL) to be identified, as loci where the effect of genotype on expression depends on the environment. More than 1,500 G $\times$ E eQTL were identified; 963 had effects in both conditions, and 72% of these had opposite directions of effect between glucose and ethanol. *IRA2*, a Ras GTPase-activating protein upstream of protein kinase A (PKA), was identified as a carbon-source hotspot, and swapping the *IRA2* coding region between BY and RM backgrounds recapitulated the locus effects on expression across both media (Smith & Kruglyak, 2008).

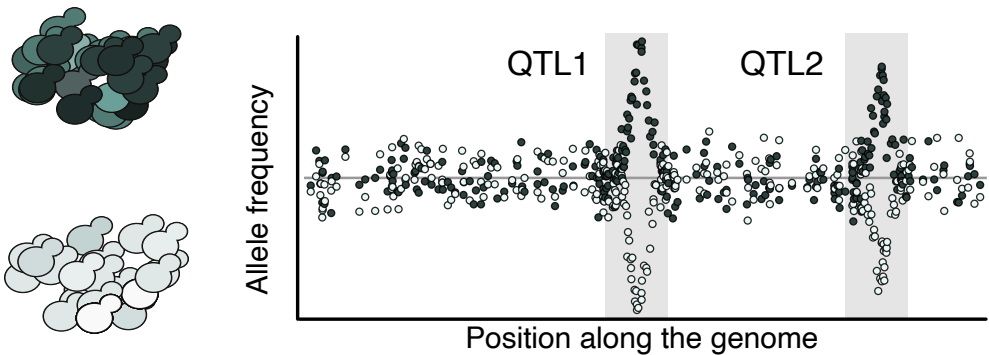


Figure 15 | Bulk segregant analysis (BSA).

Rather than genotyping segregants individually, a large segregant population is split into two pools at the phenotypic extremes. Each pool is sequenced, and the frequency of each parental allele is measured along the genome. Across most of the genome the two bulks share similar allele frequencies, but at loci controlling the trait the alleles diverge sharply, producing opposing frequency peaks that mark quantitative trait loci.

(ii) Bulk segregant approaches

Classical QTL mapping required every segregant to be both individually phenotyped and individually genotyped. The genotyping cost, scaling linearly with the panel size, was therefore a bottleneck that capped how many recombinants could be screened. An alternative approach, called bulk segregant analysis (BSA), was developed. Briefly, segregants from a cross are individually phenotyped, and the extreme individuals at the tails of the distribution are pooled together before being genotyped in bulk (Fig. 15). Alleles contributing to the phenotype become enriched in one bulk relative to the other, so strong differences in parental-allele frequency mark candidate QTL. In

yeast, BSA was notably applied using oligonucleotide microarrays to map QTL for a range of phenotypes (Brauer et al., 2006; Segrè et al., 2006), and a formal statistical framework for short-read sequencing readouts followed (Magwene et al., 2011). To push the resolution further, a new flavor of BSA, called extreme QTL mapping (X-QTL), was developed (Ehrenreich et al., 2010). It scaled BSA from hundreds of phenotyped segregants to pools of more than one million, enabling gene-level resolution.

While QTL mapping is a powerful approach and has even been extended to SVs (Weller et al., 2023), it can be limited by its relatively low resolution. A large number of QTL have been mapped, but very few have been resolved even to the level of a causal gene. Furthermore, a common criticism of classical biparental linkage studies is that each cross explores variation between only two parental genomes, which represent only a small fraction of the natural diversity within a species.

(iii) Genome-Wide Association Studies (GWAS)

By 2006, dense-SNP arrays (and, within a few more years, short-read WGS) had reduced genotyping costs enough to profile the allelic composition of hundreds to thousands of individuals across their entire genome. While QTL mapping relies on the generation of a large offspring population, starting from phenotypically distinct parental strains, GWAS leverage the ancestral recombination accumulated over many generations in a natural population to ask which variants are associated with phenotypic variation across the species without designing a cross (Hirschhorn & Daly, 2005; Visscher et al., 2012, 2017).

The first GWAS was performed in 2005 in humans, dissecting the genetic basis of age-related macular degeneration (AMD) in 150 elders, in which a common intronic variant of the complement factor H (CFH) gene was found to be associated with the disease; follow-up resequencing of CFH then flagged the Tyr402His allele (Klein et al., 2005). The next year, another AMD-associated variant was identified in a promoter using a Chinese cohort (DeWan et al., 2006). These studies were made possible by the International HapMap Project that had spent a few years cataloguing common variation across human populations. Around 1 million SNPs (later extended to 3 million) were genotyped in more than 260 individuals from four populations (The International HapMap Consortium, 2003, 2005; Frazer et al., 2007). The project also characterized patterns of linkage disequilibrium (LD). LD is the nonrandom association of alleles at nearby loci. Because recombination rarely breaks up short chromosomal intervals, SNPs within the same LD block tend to be correlated and inherited together; genotyping a representative SNP can therefore provide information about its neighbors. A tag-SNP array exploits this correlation by genotyping a set of representative SNPs, capturing much of the common variation with far fewer markers than the total number of known SNPs. With that catalogue in hand, the field truly moved away from candidate-gene studies, criticized for their reproducibility issues (Hirschhorn et al., 2002), toward hypothesis-free scanning of the entire genome (at least for common variants) at once through GWAS.

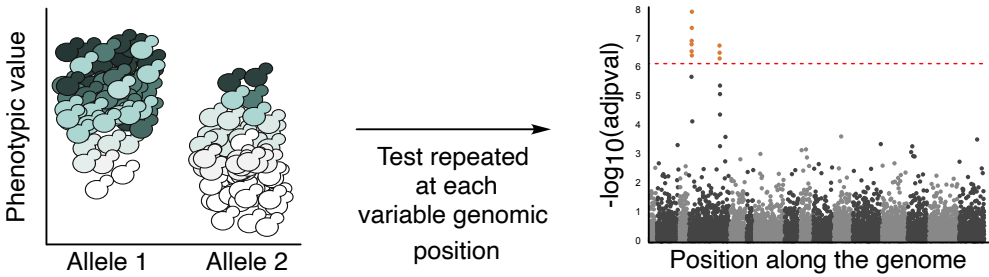


Figure 16 | Genome-wide association study (GWAS).

In a panel of ideally unrelated individuals, samples are grouped at each variable genomic position by allele, and their phenotypic values are compared. The test is repeated genome-wide, and association strengths are plotted as $-\log_{10}(\text{p-value})$ against genomic position (Manhattan plot). Variants whose association exceeds the significance threshold (red dashed line; orange points) mark loci contributing to the trait.

The principle is simple (Fig. 16). At each biallelic position in the genome, individuals are grouped by genotype (or, in haploids, by allele). A test then determines whether phenotypic values differ significantly among the groups, thereby testing whether the variant is associated with phenotypic variation. The process is repeated independently at each polymorphic site in the genome. Therefore, in theory, any common biallelic marker with a measurable effect on the trait can be detected, without requiring a cross. Because the test is repeated across potentially hundreds of thousands of markers, significance is held to a stringent genome-wide threshold (classically $p < 5 \times 10^{-8}$ in humans, or a permutation-based cutoff in smaller panels such as yeast) rather than the nominal 0.05. The tested variants were almost exclusively SNPs in the first generation of scans; CNVs were included in genome-wide association scans a few years later (Wellcome Trust Case Control Consortium et al., 2010), while most other forms of structural variation remained largely invisible until recently (Loegler et al., 2025). As of August 2, 2026, the GWAS Catalog listed 7,784 publications, 1,188,619 top associations, and 193,739 full summary statistics², but GWAS have also been performed in classic model organisms such as *A. thaliana*, *Mus musculus*, *C. elegans*, or *D. melanogaster* (Atwell et al., 2010; Harbison et al., 2013; Alonso-Blanco et al., 2016; Gonzales et al., 2018). With sequencing costs dropping, GWAS have also notably been used on a wide range of agriculturally important (and delicious) species: rice, tomato, corn, strawberries, peaches, and pineapples (Cao et al., 2016; Kim et al., 2021; Muñoz et al., 2024; Nashima et al., 2024; Tang et al., 2024; Kordi et al., 2025). Interestingly enough, any phenotype with variation can theoretically be studied, from autism spectrum disorders (Grove et al., 2019) to type 2 diabetes (Xue et al., 2018) to more unconventional ones such as educational attainment, musical beat synchronization, proneness to anger, or the Big Five personality traits (Rietveld et al., 2013; Mick et al., 2014; Niarchou et al., 2022; Schwaba et al., 2026). The first GWAS of 2005-2006 relied on simple case-control tests in deliberately homogeneous cohorts, with genomic control used to absorb residual confounding (Devlin & Roeder, 1999), but this simple design did not statistically survive larger and more diverse cohorts.

²GWAS Catalog, accessed August 11, 2026.

Two sources of confounding need to be defined here. (i) Population structure consists of systematic allele-frequency differences between subgroups (different ancestries or clades); consequently, any SNP that differs between subgroups can appear associated with any phenotype that also differs between them, whether causal or not. (ii) Cryptic relatedness is the presence of unrecognized close relatives in the sample, whose genotypes are more similar than expected and whose phenotypes may therefore also be correlated.

Both come down to the same thing: isolates that are not statistically independent. They are usually treated separately in human cohorts, but in a predominantly clonal species that outcrosses only rarely, such as *S. cerevisiae*, the distinction becomes blurred: near-clonal lineages and diverged clades reflect different scales of one and the same genealogy, so methods that model relatedness often absorb both (Connolly & Akey, 2012; Peter et al., 2018). Structure correction is then an umbrella term for the statistical methods that absorb this relatedness, so that association reflects causal signal rather than shared ancestry.

Scaling to larger and more admixed panels required more advanced statistical models. Early approaches modeled structure explicitly, either by estimating individuals' membership in inferred ancestral subpopulations (Pritchard et al., 2000) or by adding the leading principal components of the genotype matrix as covariates (Price et al., 2006).

The simplest association test fits, marker by marker, an ordinary linear model defined as:

$$y = X\beta + s\gamma + \varepsilon$$

where y is the phenotype measured across individuals, X is the fixed-effect design matrix (intercept and any covariates) and β contains the corresponding coefficients, s is the allele dosage at the marker (0, 1, or 2 copies of the minor allele; 0 or 1 in haploids), γ is the per-allele effect tested against the null $\gamma = 0$, and ε is the residual.

This model assumes that residuals are independent across isolates, which is often not the case in genetic data, especially in *S. cerevisiae*, a species with strong population structure.

Linear mixed models (LMMs) provide a way to absorb both structure and finer-scale relatedness in a single fit, by using a random effect whose covariance is set by a genome-wide kinship matrix (Yu et al., 2006):

$$y = X\beta + s\gamma + g + \varepsilon, \quad g \sim \mathcal{N}(0, \sigma_g^2 K), \quad \varepsilon \sim \mathcal{N}(0, \sigma_e^2 I)$$

where K is the kinship matrix estimated from genome-wide markers and σ_g^2 , σ_e^2 are the genetic and residual variance components, respectively.

However, a naive LMM fit requires matrix operations that scale with the cube of the number of individuals, which becomes prohibitive when repeated at every variant. Algorithmic refinements made such calculations doable. Efficient Mixed-Model Association (EMMA) reduced the per-variant variance-component fit to an efficient one-dimensional optimization, using a spectral decomposition of the kinship matrix to make each likelihood evaluation relatively inexpensive computationally. But it still re-estimated the variance components at every marker, leaving it impractical for cohorts of thousands (Kang et al., 2008). EMMA eXpedited (EMMAX) dealt with this issue by estimating the variance components once under the null model and reusing them genome-wide, cutting computation times (Kang et al., 2010). Factored Spectrally Transformed Linear Mixed Models (FaST-LMM) went further, using factored spectral transformations and a low-rank representation of the kinship matrix (Lippert et al., 2011). This allowed runtime to scale linearly rather than cubically with cohort size when the kinship matrix was low rank, meaning that relatedness could be represented using fewer independent genetic dimensions than there were individuals. Unlike EMMAX's null-model approximation, the test remained exact. LMMs and their fast implementations have since become a central part of the GWAS toolkit. For more information on GWAS and a new graphics processing unit (GPU) implementation of FaST-LMM, see Part I of the appendices, p. 250.

The first population-based GWAS in *S. cerevisiae* was performed in 2011: 88 clinical and environmental isolates were genotyped on oligonucleotide tiling arrays and scanned for associations with clinical origin (Muller et al., 2011), identifying several candidate loci despite the small panel. Simulations on the 36- and 63-strain panels available at the time helped quantify the issue. Even with mixed-model corrections, genome-wide false-positive rates were substantially inflated because the complex and heterogeneous population structure of *S. cerevisiae* was not fully captured by global corrections (Connelly & Akey, 2012). Through the mid-2010s, yeast GWAS remained underpowered at the sample sizes then available. Indeed, power was limited by the panel size and by the absence of a curated species-scale natural-isolate collection (Peter & Schacherer, 2016). The 1,011-isolate resource provided both genotypes and phenotypes at species scale (Peter et al., 2018). GWAS across 971 of those isolates identified 35 genetic variants associated with 14 growth conditions. Copy number variation of *CUP1* (a copper-binding metallothionein) explained 44.5% of variation in copper sulfate resistance. Across associated traits, CNVs explained a median of 36.8% of phenotypic variance, compared with 4.49% for SNPs.

Genome-wide association studies have become a standard tool to dissect the genotype–phenotype relationship (Uffelmann et al., 2021) and are now applied across a wide range of organisms. They nevertheless carry notable weaknesses that limit their detection power (Tam et al., 2019; Ciochetti et al., 2023). When the genetic variance captured by genome-wide-significant hits is measured against the narrow-sense heritability (h^2) estimated from twin and family studies, a persistent gap appears (Wainschein et al., 2022, 2026; Palmer et al., 2023). Human height is a classic example, with variation in height estimated to be ~80% heritable (Silventoinen et al., 2003). A 2018 meta-analysis of 700,000 individuals recovered ~25% of the phenotypic

variance in height through genome-wide-significant SNPs (Yengo et al., 2018). A 2022 follow-up, scaled to 5.4 million individuals of diverse ancestries, identified 12,111 independent SNPs across more than 7,000 genomic segments; common SNPs within these loci accounted for an astonishing $\sim 100\%$ of SNP-based heritability (h_{SNP}^2) in European-ancestry populations (Yengo et al., 2022). Notwithstanding this result, SNP-based heritability is not the same as the full narrow-sense heritability, and GWAS still do not recover the entire genetic architecture of complex traits. The remaining discrepancy between heritability estimates and the variance explained by variants identified through GWAS has been named “missing heritability” (Maher, 2008; Manolio et al., 2009).

Many explanations for this discrepancy have been put forward, both technical and biological (Génin, 2020; Matthews & Turkheimer, 2022). Frequently cited sources include (i) rare genetic variants, (ii) SVs, (iii) non-additive effects, and (iv) pangenomic gene-content variation. In that sense, some of the missing heritability may simply be hiding outside the variants that GWAS can have access to.

(i) Population studies have shown that the vast majority of variants are present at low frequency. Across surveyed species, there is a clear bias toward lower allele frequencies, with around 90% of the variants having a minor allele frequency (MAF) of less than 5% (Cao et al., 2011; Gibson, 2012; Auton et al., 2015; Peter et al., 2018). GWAS are well powered for common alleles but lose power with diminishing allele frequency. This leaves much low-frequency variation outside the detectable range of GWAS unless sample sizes are very large or effect sizes are strong.

(ii) Structural variants are another source of the missing heritability. Indeed, they have historically been underrepresented in GWAS because they are difficult to identify and genotype consistently at scale. Short-read sequencing can detect many SVs but often misses variants in repetitive or complex regions, while SNP arrays represent most SVs only indirectly through LD with assayed SNPs (Wellcome Trust Case Control Consortium et al., 2010). As a result, single-marker GWAS can miss the SV-borne heritability, especially when the underlying signal is not tagged by nearby SNPs. In yeast, however, this gap is beginning to close: the 1,086 near-T2T genomes recently produced from long-read sequencing made structural variants detectable across the species. GWAS performed using these SV calls showed that SVs have a major phenotypic impact: they were more frequently associated with traits and more pleiotropic than SNPs or InDels. Including SVs and small InDels alongside SNPs raised heritability estimates by $\sim 14\%$ on average, recovering signal that short-read panels had been missing (Loegler et al., 2025). In humans, recent developments have enabled such studies, revealing that SVs contributed to at least 4.7% of the common genetic variation for complex traits (Bai et al., 2026).

(iii) Standard GWAS, including most LMM-based implementations, generally assume that allelic effects on a phenotype combine additively. Non-additive effects come in two forms. Dominance is the interaction between alleles at the same locus in a diploid context. In these cases, one allele can mask another. The other is epistasis, which corresponds to the interaction between alleles at different loci, where the combined

effect of two or more variants does not equal the sum of their individual effects, producing an effect not predicted by those individual effects. Most standard scans do not explicitly model epistasis, because they test one variant at a time. Even when taking pairwise interactions into account, unfortunately, the number of interacting loci isn't limited to two, and higher-order interactions naturally become rapidly harder to detect as the number grows. Structured crosses in yeast put a substantial share of genetic variance beyond additivity. In diploid panels, where dominance is expressed, one study found that non-additive effects accounted for up to a third of the variance, primarily through dominance (Fournier et al., 2019), while another estimated ~19%, partitioned into ~10% dominance, ~7% pairwise epistasis, and ~2% third-order interactions (Hallin et al., 2016). Haploid segregants, blind to dominance, place pairwise epistasis lower, around 9% (Bloom et al., 2013, 2015). Either way, this non-additive variance is largely inaccessible to standard single-variant GWAS. Furthermore, the yeast deletion collection allows a complementary, gene-by-gene view. The pairwise screen detected close to one million genetic interactions (Costanzo et al., 2016). A subsequent triple-mutant scan suggests that the global trigenic interaction network may be about 100-fold larger than the digenic network (Kuzmin et al., 2018, 2020). Higher-order epistasis between three, four, or more loci has also been characterized (Taylor & Ehrenreich, 2014, 2015). However, the number of possible combinations grows combinatorially with the number of loci, placing most higher-order interactions beyond the reach of variant-by-variant testing. This makes non-additive effects another potential source of missing heritability.

(iv) The pangenome causes a less obvious problem. A scan that tests only reference-aligned positions cannot see sequence the reference lacks, including variation inside accessory genes. Their presence or absence can still be encoded as a marker and tested directly when an accessory gene segregates at sufficient frequency, as done across the 1,011-isolate panel; but rare and private accessory genes, and variation internal to those genes, escape any reference-anchored test (Caudal et al., 2024; Loegler et al., 2025). Recent developments in pangenome-graph GWAS, using variation graphs that integrate multiple reference genomes into a single mapping structure, are a promising route to the accessory and structural variation that reference-anchored GWAS cannot reach (Eizenga et al., 2020; Li et al., 2023; Wang et al., 2024, 2026; Yang et al., 2025; Naji et al., 2026).

Moreover, several other contributors remain difficult to capture.

(1) We can cite the influence of the genetic background as one. The genetic background a variant finds itself in can shape inheritance patterns. Indeed, traits that segregate as Mendelian in one individual can lose that simple pattern in another, as the same focal allele shows different expressivity across genetic backgrounds (Hou et al., 2016, 2018).

(2) Mitochondrial genomes are another source of missing heritability (Hudson et al., 2014). They carry variants that standard GWAS often omit even though mitochondria are not only powerhouses of the cell but central metabolic hubs that substantially contribute to multiple cellular processes.

(3) Regulatory variants in non-coding regions are functionally harder to interpret than coding variants and resist fine-mapping to single causal nucleotides, even when detected with adequate statistical support (Visscher et al., 2017).

As discussed, one limitation of GWAS is inherited from whatever population they're used on. Natural populations carry population structure, systematic allele-frequency differences between clades that make any variant separating two clades appear associated with any phenotype that also separates them. *S. cerevisiae* is a very structured species, with a well-defined stratification, meaning that correcting for it limits the power of GWAS. Rather than correct for it statistically, multi-parent crosses have been designed in order to break clade structure while preserving allelic breadth. Furthermore, most variants in the population are singletons, private to each individual, making them impossible to study through classic GWAS. Therefore, other methods have to be explored in order to obtain deeper insights into the genotype–phenotype relationship.

(iv) Multi-parent designs

We can also find a third class, somewhere between the two-parent cross and the full species-wide associations, designed to recover what GWAS misses, namely rare variants and masked allelic effects, while breaking the population structure that confounds natural panels.

A limited panel of founders is chosen to span the species' allelic diversity rather than to represent a single phenotypic contrast. These founders are then crossed under different mating schemes in order to shuffle the haplotypes. Across generations, thousands of segregants are produced, with chromosomal blocks that can be traced to the parental founders. This represents a combination of the allelic breadth of a natural population with the controlled kinship structure of a cross, with each segregant having a known pedigree. In *S. cerevisiae*, three mating schemes have mainly been used. A round-robin pairs each founder with its two neighbors once. A diallel crosses every pair of founders. Finally, funnels are designed to halve the number of separate pools at each successive mating step.

An early study crossed four divergent founders from the *Saccharomyces* Genome Resequencing Project in all pairwise combinations and mapped QTLs across 23 growth conditions in 576 recombinant strains. Most QTLs were specific to a single cross combination: A × B identified one set of significant loci, A × C a different set, and so on across the six pairwise crosses. No single cross managed to capture the full picture, and crosses between phenotypically similar founders produced more offspring variation than expected (Cubillos et al., 2011).

Two 2019 studies pushed this class of designs in complementary directions. A round-robin scheme (Fig. 17a) with 16 genetically and ecologically diverse parents yielded 13,950 individually genotyped haploid segregants and mapped 4,552 QTLs across 38 growth conditions (Bloom et al., 2019). It tackled the question of the impact of rare variants across the population. These variants are largely inaccessible through GWAS because their low minor-allele frequency leaves almost no statistical signal in a natural panel. A round-robin design breaks that limitation: a variant carried by one of

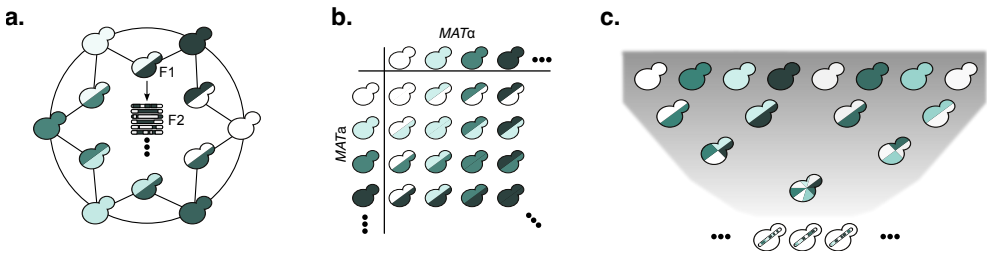


Figure 17 | Multi-parent population designs.

Extending mapping beyond a single biparental cross, several founders are combined so that more alleles segregate in one population.

- a.** Round-robin (cyclic) design: founders arranged in a circle are crossed to their neighbors, and each F1 hybrid is sporulated to produce an F2 population.
- b.** Diallel design: a panel of MATa strains is crossed to a panel of MATα strains in all pairwise combinations, generating the full matrix of hybrids.
- c.** Funnel design: multiple founders are intercrossed over successive generations, progressively blending their genomes into highly recombinant mosaics that sharpen mapping resolution.

the 16 founders will segregate at 50% frequency among the haploid spores of the two crosses where that founder participates, no matter how rare it is in the wild. Therefore, a rare allele in the natural population becomes common inside the cross. Across all traits, rare variants (MAF < 1% in yeast) constituted only 28% of all variants but explained a median of 52% of the additive genetic variance per trait. This shows that low-frequency alleles tend to carry larger per-variant effects (Bloom et al., 2019).

In parallel, a complete diallel panel (Fig. 17b) of 55 strains was used to generate 3,025 pairwise hybrids that were phenotyped on 49 growth conditions (Fournier et al., 2019). Since every hybrid is a heterozygous diploid at every site where the parents differ, the design allowed phenotypic variance to be partitioned into 55% additive and 29% non-additive (with dominance as the main non-additive mode). A GWAS using this artificial hybrid population was performed, mapping 281 low-frequency variants whose per-variant contribution to heritability was statistically indistinguishable from common variants (Fournier et al., 2019). Thanks to the two designs, it was made clear that the variants that natural-population GWAS cannot access don't necessarily have smaller effects but remain undetected because statistical power is limited at low allele frequencies.

Later, a 26-parent half-diallel carried the same logic. The rationale was twofold. First, extend the rare-variant heritability question from growth traits to gene expression, and second, bring SVs in as candidate eQTL drivers (Tsouris et al., 2024a, b). The parents were sequenced using ONT long reads, and 323 high-quality transcriptomes were generated. Both SVs and low-frequency SNPs were found to shape gene expression at the population scale (Tsouris et al., 2024b), with SNPs having larger per-variant effects than SVs but low-frequency alleles being overrepresented among lead eQTLs. This complemented the rare-variant heritability picture previously reported for growth traits (Bloom et al., 2019; Fournier et al., 2019).

Nevertheless, round-robin and diallel panels still inherit some clade structure from their natural-isolate founders (Cromie et al., 2025b). A different mating scheme is therefore needed to further reduce it. Indeed, round-robin designs leave each founder-specific rare variant confined to the two crosses involving its carrying founder (Bloom et al., 2019), and diallel hybrids are unrecombined (Fournier et al., 2019).

In order to improve genotype–phenotype mapping, further mixing is required. CYClones (Collaborative Yeast Cross clones), a library generated through a funnel-cross design, was constructed from 8 genetically diverse founder strains (Fig. 17c): four Chinese isolates and one individual each from the Philippines, the United States of America, Nigeria, and South Africa, with no European representation (Cromie et al., 2025a, b). Together, they captured a sizable fraction (56% of biallelic single-nucleotide variants (SNVs) at global MAF > 0.05, and 32% at MAF > 0.005) of the currently known genetic diversity of the species. Three successive rounds of pooled mating and sporulation were used in a funnel-like manner, without inbreeding steps. The resulting library of 11,392 recombinant mosaics is able to deliver more than 95% mapping power at heritabilities as low as 0.36%, with potential sub-gene resolution once a variant accounts for 9% or more of the phenotypic variance ($h^2 \geq 0.09$). A first application phenotyped 9,344 euploid segregants across ten growth conditions and recovered 349 QTL, of which 37 resolved to a single causal gene (Cromie et al., 2025a). For example, *PCA1*, a cadmium efflux pump carried as a non-*cerevisiae* introgression in one founder, mapped to a 2.4 kb interval explaining 72% of the variance in growth on cadmium. At *GAL3*, two distinct missense alleles produced the same loss-of-function phenotype. This allelic heterogeneity was invisible to a per-SNP GWAS, which would test each variant separately rather than aggregate them at the gene level. From these and similar cases, the authors coined the term “allele-dependent stratification”: allelic variation at a few major loci splits the panel into regulatory or physiological subgroups, each with a distinct genetic architecture. A model that fits the majority therefore is unable to describe the minorities.

Allele-dependent stratification suggests that mapping alone, even at the resolution provided by CYClones, cannot fully describe the genetic architecture of complex traits: variants also act through the molecular layers that sit between genotype and final phenotype (Fig. 18). When subgroups of strains have distinct architectures, orthogonal molecular data can help characterize the regulatory or physiological states that distinguish them. Measurements that report on cellular state rather than on the genotype itself (such as gene expression, protein abundance, chromatin accessibility, or any molecular phenotype downstream of the stratifying alleles) then provide information that genotype-based mapping alone does not directly capture (Cromie et al., 2025a).

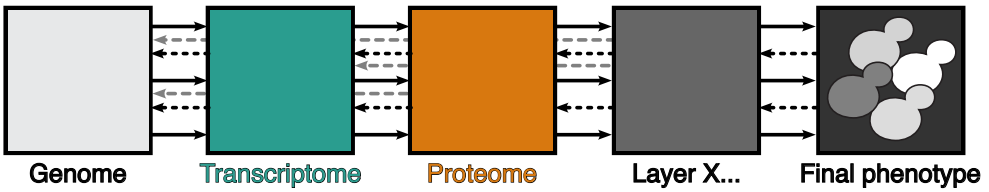


Figure 18 | The molecular layers between genotype and phenotype.

Genetic variation is relayed to the final phenotype through a cascade of measurable molecular layers. Studying them is necessary to understand the genotype–phenotype relationship.

Gene expression and protein abundance are themselves quantitative traits, and their eQTLs and protein quantitative trait loci (pQTLs) are now mapped by GWAS across organisms, from human cohorts to other model systems (Gonzales et al., 2018; GTEx Consortium, 2020; Lan et al., 2021; Gudjonsson et al., 2022; Wongsurawat et al., 2022; Zhang et al., 2022; Suhre et al., 2025; Wingo et al., 2025). The yeast community has begun assembling these layers at species scale, taking advantage of the powerful model system. At the species level, studies of transcriptomes profiled during exponential growth on glucose (Caudal et al., 2024) and proteomes from the 1,011-isolate panel (Muenzner et al., 2024; Teyssonnière et al., 2024b) mapped thousands of eQTLs and pQTLs using GWAS, but found an overlap of only around 4%, suggesting that the genetic architectures governing transcript and protein abundance are largely distinct (Teyssonnière et al., 2024b). Surprisingly, a recent reanalysis suggests the opposite. Much of the apparent gap may reflect statistical power: protein abundances are quantified less precisely than transcripts, so a locus acting on both is often detected on only one side and scored as layer-specific. When the two layers are profiled in the same cross and matched for power, as in the BY $\times$ RM cross, they appear largely concordant. Across thirteen such datasets, genetic effects on mRNA and protein mostly agree, with no hotspot reproducibly specific to either layer; the rare layer-specific effects, globally an order of magnitude less frequent, concentrate at a few recurrent loci such as *MKT1*, *BUL2* and *IRA2*, where protein-specific signals seem to reflect turnover, modification or complex stoichiometry (Van Dyke et al., 2026).

Two gaps still remain. First, as extensively discussed, a single genetic background isn't enough to understand the genotype–phenotype relationship. Studying multiple strains therefore is a natural and necessary step toward more precise genetic insights. Second, the molecular depth available across such species-wide panels remains limited. The genotype–phenotype relationship is tightly linked to the molecular layers, as they sit between DNA and the final phenotype. Transcript and protein abundances have been explored at species scale, but other layers, such as translation, metabolite abundance, chromatin accessibility, transcript isoforms, or post-translational modifications, have each been explored only in one or a few backgrounds, with no species-scale survey to date (Ingolia et al., 2009; Pelechano et al., 2013; Tudek et al., 2021; Teyssonnière et al., 2024a).

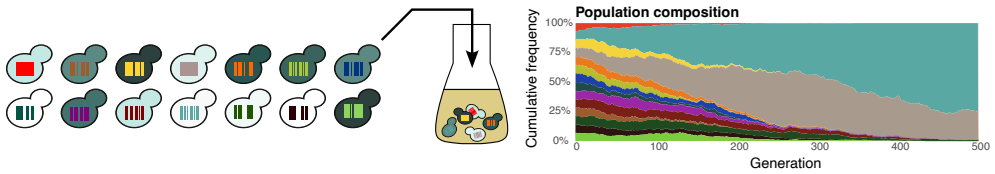


Figure 19 | Barcoded lineage tracking during experimental evolution.

A library of otherwise isogenic strains, each tagged with a unique chromosomal DNA barcode (colored bars), is pooled and propagated for generations. Barcode sequencing at successive time points tracks each lineage over 400 simulated generations, shown as cumulative population composition. Most lineages decline as a few adaptive clones expand and compete (clonal interference), with one lineage (teal) ultimately sweeping toward fixation.

Toward a species-wide functional toolkit

Every mapping framework above was applied to a species in which the tools (knockouts, tagged proteins, double-mutant arrays) were single-strain projections of a multi-strain reality. In order to get deeper insights into the genotype–phenotype map, using the full diversity of *S. cerevisiae* might prove to be a necessity.

Barcoded lineage tracking offers another route into this diversity (Fig. 19). A first species-wide chemical-genomic screen provided a proof of concept. A barcoded collection of 520 *S. cerevisiae* natural isolates was grown competitively in liquid medium in the presence of more than 600 bioactive and natural compounds. Relative fitnesses were inferred from changes in barcode frequencies across generations measured by pooled barcode sequencing. Six major fitness-based compound clusters, whose response patterns tracked population structure, were identified. This is a new view into the species-wide variability that loss-of-function screens cannot access (Dutta et al., 2026). GWAS applied to those compound-response profiles implicated variants in 46 genes linked to genome maintenance, ribosome biogenesis, vesicular trafficking and stress tolerance. The resulting chemical-response space proved to be broader than the deletion-collection equivalent built almost two decades earlier in the S288C background (Hillenmeyer et al., 2008). Only a species-wide panel can expose this range of responses, hence the necessity of species-wide explorations.

Recently, a wild-focused panel of 33 Taiwanese haploids spanning pre-domesticated lineages was established by targeted *HO* disruption (Yang et al., 2026a). At the species scale, a collection of 364 barcode-tagged stable haploid strains, derived from genetically diverse natural diploids and available in MATa and MATα flavors as both wildtype and *ura3*-deleted versions, was released (Becerra-Rodríguez et al., 2026). This laboratory-friendly resource derived from natural isolates is a concrete step toward species-wide perturbation screens and finer variant mapping. On this scale, for the first time, lab-friendly strains are available for the community to use, recapitulating the diversity of the species itself, rather than a single genetic background.

Outlook

Now, thirty years after André Goffeau and his consortium published the first *Saccharomyces cerevisiae* genome, the field is slowly moving from that single reference to a near-complete catalogue of variants across more than a thousand isolates. From Sanger sequencing to microarrays, short-read population surveys, and long-read T2T assemblies, each new technology helped the community to gather genotypes and phenotypes.

Connecting the two is the real problem, and all this data serves two complementary ends. On one side, mapping locates causal variants as discussed earlier; on the other side, prediction asks whether a strain's phenotype can be estimated from its genotype alone. A perfect prediction model can be opaque to causality while a short list of validated variants will under-predict because it captures only a fraction of heritability. What models now have is training material that reaches well beyond the variant catalogue with transcriptomes (Caudal et al., 2024) and proteomes (Teyssonnière et al., 2024b) across the same isolates, providing measurements of the molecular layers that sit between genotype and phenotype, as well as multi-parent panels of tens of thousands of recombinants (Nguyen Ba et al., 2022; Cromie et al., 2025b). Together, these resources can be interpreted against the thousand-plus T2T genomes that fix a near-complete variant set (Loegler et al., 2025).

These supply training material for models that bypass linear additivity altogether. A study in this spirit surveyed 223 life-history and phenotypic traits in the 1,011 isolates and identified gene presence/absence and gene-disruption scores as the strongest predictors, once again implicating the accessory genome as one of the major contributors to the genotype–phenotype map (Khaiwal et al., 2025). Through deep neural networks trained on massive assays in yeast, we can already predict promoter activity directly from sequence, with accuracies substantially exceeding those of linear models, capturing the combinatorial syntax of transcription-factor binding sites (de Boer et al., 2020; Vaishnav et al., 2022). The current state of the art extends this sequence-to-expression lineage to exogenous DNA expressed in yeast cells for synthetic-biology design (Chao et al., 2025; Mandl & Orenstein, 2026). Toward organism-level prediction, combining yeast-specific predictors with general-purpose models such as AlphaFold for protein structure (Jumper et al., 2021; Abramson et al., 2024) and, potentially, AlphaGenome-like architectures for non-coding sequences (Avsec et al., 2026) is a plausible route toward bridging the genotype–phenotype gap.

Nonetheless, predictions are still far from matching real biology and need to be pushed further to bring new knowledge to the community. But doing so might prove to be a real challenge. Indeed, heritability is dynamic, epistasis is pervasive, trait architectures shift across environments, and the panels built to map variants have limited detection power. More fundamentally, at the species level, the catalogue itself is not yet complete either. As under-sampled habitats are surveyed, divergent populations are likely to keep surfacing. Nevertheless, some of the components needed to close the genotype–phenotype gap are now in place: pangenome graphs, species-wide transcriptomes and proteomes, multi-parent and barcoded recombinant

panels, designed-variant screens, and sequence-to-function models. Few (if any) other eukaryotes bring this combination together. With the compact genome that made *S. cerevisiae* the first eukaryote ever sequenced, the species is one of the strongest candidates to become the first with a near-complete genotype-to-phenotype map.

However that promise comes with a large caveat. The catalogue of variants appears nearly complete, yet the variants whose effects we can detect still explain far less than the heritability measured. As the preceding parts set out, a large part of that gap comes from sources that association mapping has the least power to detect: low-frequency and structural variants, together with non-additive, background- and environment-dependent effects. On top of this, the effects of variants can depend on complex interactions across loci and are further relayed through the molecular layers between genotype and phenotype, blurring the picture further. This is where my thesis project sits: recovering a small part of that missing heritability.

References

- Aa E, Townsend JP, Adams RI, Nielsen KM, Taylor JW (2006) Population structure and gene evolution in *Saccharomyces cerevisiae*. *FEMS Yeast Res* **6**: 702–715.
- Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**: 493–500.
- Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. *Nature* **422**: 198–207.
- Aigner B, Besenfelder U, Müller M, Brem G (2000) Tyrosinase gene variants in different rabbit strains. *Mamm Genome* **11**: 700–702.
- Akao T, Yashiro I, Hosoyama A, Kitagaki H, Horikawa H, Watanabe D, et al. (2011) Whole-Genome Sequencing of Sake Yeast *Saccharomyces cerevisiae* Kyokai no. 7. *DNA Res* **18**: 423–434.
- Albert FW, Muzzey D, Weissman JS, Kruglyak L (2014) Genetic Influences on Translation in Yeast. *PLoS Genet* **10**: e1004692.
- Almeida P, Barbosa R, Zalar P, Imanishi Y, Et. A (2015) A population genomics insight into the Mediterranean origins of wine yeast domestication. *Mol Ecol*.
- Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, et al. (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. *Cell* **166**: 481–491.
- Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21**: 30.
- Amberger JS, Bocchini CA, Scott AF, Hamosh A (2019) OMIM.org: Leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Res* **47**: D1038–D1043.
- Annaluru N, Muller H, Mitchell LA, Ramalingam S, Stracquadanio G, Richardson SM, et al. (2014) Total Synthesis of a Functional Designer Eukaryotic Chromosome. *Science* **344**: 55–58.
- Argueso JL, Carazzolle MF, Mieczkowski PA, Duarte FM, Netto OVC, Missawa SK, et al. (2009) Genome structure of a *Saccharomyces cerevisiae* strain widely used in bioethanol production. *Genome Res* **19**: 2258–2270.
- Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, et al. (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**: 627–631.
- Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. (2015) A global reference for human genetic variation. *Nature* **526**: 68–74.
- Avelar-Rivas JA, Sedeño I, García-Ortega LF, Urban Aragon JA, López-Gallegos C, Aguirre-Dugua X, et al. (2026) Recurrent introgression and geographical stratification shape *Saccharomyces cerevisiae* in the Neotropics. *Nat Commun* **17**: 3024.
- Avery OT, MacLeod CM, McCarty M (1944) STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES: INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *J Exp Med* **79**: 137–158.
- Avsec Ž, Latysheva N, Cheng J, Novati G, Taylor KR, Ward T, et al. (2026) Advancing regulatory variant effect prediction with AlphaGenome. *Nature* **649**: 1206–1218.

- Bai W-Y, Liu S, Duan Z, Yang J-J, Chen J, Hou J, et al. (2026) Genome-wide associations of structural variants with human traits through imputation from long-read assemblies. *Nat Genet* **58**: 1258–1267.
- Batté A, Bosch-Guiteras N, Pons C, Ota M, Lopes M, Sharma S, et al. (2026) The modifiers that cause changes in gene essentiality. *Cell Syst* **17**.
- Beadle GW, Tatum EL (1941) Genetic Control of Biochemical Reactions in *Neurospora**. *Proc Natl Acad Sci* **27**: 499–506.
- Becerra-Rodríguez C, Marsit S, Galeote V (2020) Diversity of Oligopeptide Transport in Yeast and Its Impact on Adaptation to Winemaking Conditions. *Front Genet* **11**.
- Becerra-Rodríguez C, Taghouti G, Portier P, Dequin S, Casal M, Paiva S, et al. (2021) Yeast Plasma Membrane Fungal Oligopeptide Transporters Display Distinct Substrate Preferences despite Their High Sequence Identity. *J Fungi* **7**: 963.
- Becerra-Rodríguez C, Thiele P, Brach G, Dutta A, Garin M, Tan G, et al. (2026) Domestication drives repeated evolution of sexual–asexual life cycle trade-offs in yeast. *Proc Natl Acad Sci* **123**: e2526682123.
- Begik O, Lucas MC, Pryszcz LP, Ramirez JM, Medina R, Milenkovic I, et al. (2021) Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat Biotechnol* **39**: 1278–1291.
- Ben-Aroya S, Coombes C, Kwok T, O'Donnell KA, Boeke JD, Hieter P (2008) Toward a Comprehensive Temperature-Sensitive Mutant Repository of the Essential Genes of *Saccharomyces cerevisiae*. *Mol Cell* **30**: 248–258.
- Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59.
- Bergström A, Simpson JT, Salinas F, Barré B, Parts L, Zia A, et al. (2014) A High-Definition View of Functional Genetic Variation from Natural Yeast Genomes. *Mol Biol Evol* **31**: 872–888.
- Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, et al. (2021) Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat Genet* **53**: 779–786.
- Bigey F, Segond D, Friedrich A, Guezenec S, Bourgaïs A, Huyghe L, et al. (2021) Evidence for Two Main Domestication Trajectories in *Saccharomyces cerevisiae* Linked to Distinct Bread-Making Processes. *Curr Biol* **31**: 722–732.e5.
- Bleykasten-Grosshans C, Fabrizio R, Friedrich A, Schacherer J (2021) Species-Wide Transposable Element Repertoires Retrace the Evolutionary History of the *Saccharomyces cerevisiae* Host. *Mol Biol Evol* **38**: 4334–4345.
- Bloom JS, Boocock J, Treusch S, Sadhu MJ, Day L, Oates-Barker H, et al. (2019) Rare variants contribute disproportionately to quantitative trait variation in yeast. *eLife* **8**: e49212.
- Bloom JS, Ehrenreich IM, Loo WT, Lite T-LV, Et. A (2013) Finding the sources of missing heritability in a yeast cross. *Nature*.
- Bloom JS, Kottenko I, Sadhu MJ, Treusch S, Albert FW, Kruglyak L (2015) Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. *Nat Commun* **6**: 8712.
- Boer CG de, Vaishnav ED, Sadeh R, Abeyta EL, Friedman N, Regev A (2020) Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat Biotechnol* **38**: 56–65.

- Boocock J, Alexander N, Alamo Tapia L, Walter-McNeill L, Patel SP, Munugala C, et al. (2025) Single-cell eQTL mapping in yeast reveals a tradeoff between growth and reproduction. *eLife* **13**: RP95566.
- 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**: 415–419.
- Borneman AR, Desany BA, Riches D, Affourtit JP, Forgan AH, Pretorius IS, et al. (2011) Whole-Genome Comparison Reveals Novel Genetic Elements That Characterize the Genome of Industrial Strains of *Saccharomyces cerevisiae*. *PLOS Genet* **7**: e1001287.
- Borneman AR, Forgan AH, Kolouchova R, Fraser JA, Schmidt SA (2016) Whole Genome Comparison Reveals High Levels of Inbreeding and Strain Redundancy Across the Spectrum of Commercial Wine Strains of *Saccharomyces cerevisiae*. *G3 GenesGenomesGenetics* **6**: 957–971.
- Borneman AR, Forgan AH, Pretorius IS, Chambers PJ (2008) Comparative genome analysis of a *Saccharomyces cerevisiae* wine strain. *FEMS Yeast Res* **8**: 1185–1195.
- Bosch-Guiteras N, Leeuwen J van (2022) Exploring conditional gene essentiality through systems genetics approaches in yeast. *Curr Opin Genet Dev* **76**: 101963.
- Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32**: 314–331.
- Brachi B, Faure N, Horton M, Flahauw E, Vazquez A, Nordborg M, et al. (2010) Linkage and Association Mapping of *Arabidopsis thaliana* Flowering Time in Nature. *PLOS Genet* **6**: e1000940.
- Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* **72**: 248–254.
- Brar GA, Yassour M, Friedman N, Regev A, Et. A (2012) High-Resolution View of the Yeast Meiotic Program Revealed by Ribosome Profiling. *Science*.
- Brauer MJ, Christianson CM, Pai DA, Dunham MJ (2006) Mapping Novel Traits by Array-Assisted Bulk Segregant Analysis in *Saccharomyces cerevisiae*. *Genetics* **173**: 1813–1816.
- Brem RB, Kruglyak L (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci* **102**: 1572–1577.
- Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* **296**: 752–755.
- Breslow DK, Cameron DM, Collins SR, Schuldiner M, Stewart-Ornstein J, Newman HW, et al. (2008) A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. *Nat Methods* **5**: 711–718.
- Breunig JS, Hackett SR, Rabinowitz JD, Kruglyak L (2014) Genetic Basis of Metabolome Variation in Yeast. *PLOS Genet* **10**: e1004142.
- Buccitelli C, Selbach M (2020) mRNAs, proteins and the emerging principles of gene expression control. *Nat Rev Genet* **21**: 630–644.
- Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet* **43**: 956–963.
- Cao K, Zhou Z, Wang Q, Guo J, Zhao P, Zhu G, et al. (2016) Genome-wide association study of 12 agronomic traits in peach. *Nat Commun* **7**: 13246.
- Caudal É, Loegler V, Dutreux F, Vakirlis N, Teyssonnière É, Caradec C, et al. (2024) Pan-transcriptome reveals a large accessory genome contribution to gene expression variation in yeast. *Nat Genet* **56**: 1278–1287.

- Chao K-H, Magzoub MM, Stoops E, Hackett S, Linder J, Kelley DR (2025) Predicting dynamic expression patterns in budding yeast with a fungal DNA language model. *bioRxiv*: 2025.09.19.677475.
- Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, et al. (1998) SGD: Saccharomyces Genome Database. *Nucleic Acids Res* **26**: 73–79.
- Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al. (2012) Saccharomyces Genome Database: The genomics resource of budding yeast. *Nucleic Acids Res* **40**: D700–D705.
- Chervitz SA, Hester ET, Ball CA, Dolinski K, Dwight SS, Harris MA, et al. (1999) Using the Saccharomyces Genome Database (SGD) for analysis of protein similarities and structure. *Nucleic Acids Res* **27**: 74–78.
- Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, et al. (1998) The Transcriptional Program of Sporulation in Budding Yeast. *Science* **282**: 699–705.
- Ciochetti NP, Lugli-Moraes B, Silva BS da, Rovaris DL (2023) Genome-wide association studies: Utility and limitations for research in physiology. *J Physiol* **601**: 2771–2799.
- Collins RL, Talkowski ME (2025) Diversity and consequences of structural variation in the human genome. *Nat Rev Genet*: 1–20.
- Colón-González M, Aguirre-Dugua X, Guerrero-Osornio MG, Avelar-Rivas JA, DeLuna A, Mancera E, et al. (2025) Thriving in Adversity: Yeasts in the Agave Fermentation Environment. *Yeast* **42**: 16–30.
- Conant GC, Wolfe KH (2007) Increased Glycolytic Flux as an Outcome of Whole-genome Duplication in Yeast. *Mol Syst Biol* **3**: MSB4100170.
- Conant GC, Wolfe KH (2008) Turning a hobby into a job: How duplicated genes find new functions. *Nat Rev Genet* **9**: 938–950.
- Connelly CF, Akey JM (2012) On the Prospects of Whole-Genome Association Mapping in *Saccharomyces cerevisiae*. *Genetics* **191**: 1345–1353.
- Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, et al. (2010) The Genetic Landscape of a Cell. *Science* **327**: 425–431.
- Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, et al. (2016) A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**: aaf1420.
- Crick F (1970) Central Dogma of Molecular Biology. *Nature* **227**: 561–563.
- Crick FH (1958) On protein synthesis. *Symp Soc Exp Biol* **12**: 138–163.
- Cromie GA, Hyma KE, Ludlow CL, Garmendia-Torres C, Gilbert TL, May P, et al. (2013) Genomic Sequence Diversity and Population Structure of *Saccharomyces Cerevisiae* Assessed by RAD-seq. *G3 GenesGenomesGenetics* **3**: 2163–2171.
- Cromie GA, Lo RS, Ames L, Morgan TS, Owens K, Clark AE, et al. (2025a) Pervasive context-dependent effects in the genetic architecture of complex and quantitative traits revealed by a powerful multiparent mapping population in yeast. : 2025.10.27.683165.
- Cromie GA, Lo RS, Morgan TS, Clark AE, Ashmead J, Timour MS, et al. (2025b) CYClones: A highly powered, fully genotyped, 8-parent yeast mapping population. : 2025.10.15.682626.
- Cubillos FA, Billi E, Zörgö E, Parts L, Fargier P, Omholt S, et al. (2011) Assessing the complex architecture of polygenic traits in diverged yeast populations. *Mol Ecol* **20**: 1401–1413.
- D’Angiolo M, De Chiara M, Yue J-X, Irizar A, Stenberg S, Persson K, et al. (2020) A yeast living ancestor reveals the origin of genomic introgressions. *Nature* **587**: 420–425.
- De Chiara M, Friedrich A, Barré B, Breitenbach M, Schacherer J, Liti G (2020) Discordant evolution of mitochondrial and nuclear yeast genomes at population level. *BMC Biol* **18**: 49.

- De Deken RH (1966) The Crabtree Effect: A Regulatory System in Yeast. *Microbiology* **44**: 149–156.
- Dean EJ, Davis JC, Davis RW, Petrov DA (2008) Pervasive and Persistent Redundancy among Duplicated Genes in Yeast. *PLOS Genet* **4**: e1000113.
- Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing Chromosome Conformation. *Science* **295**: 1306–1311.
- DeLuna A, Vetsigian K, Shoresh N, Hegreness M, Colón-González M, Chao S, et al. (2008) Exposing the fitness contribution of duplicated genes. *Nat Genet* **40**: 676–681.
- Demichev V, Messner CB, Vernardis SI, Lilley KS, Ralser M (2020) DIA-NN: Neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* **17**: 41–44.
- DeRisi JL, Iyer VR, Brown PO (1997) Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science* **278**: 680–686.
- Deutsch EW, Kok LW, Mudge JM, Valls CF, Jungreis I, Ruiz-Orera J, et al. (2026) Expanding the human proteome with microproteins and peptideins. *Nature*: 1–13.
- Devlin B, Roeder K (1999) Genomic Control for Association Studies. *Biometrics* **55**: 997–1004.
- DeWan A, Liu M, Hartman S, Zhang SS-M, Liu DTL, Zhao C, et al. (2006) HTRA1 Promoter Polymorphism in Wet Age-Related Macular Degeneration. *Science* **314**: 989–992.
- Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, et al. (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304**: 304–307.
- Dijk PJ van, Jessop AP, Ellis THN (2022) How did Mendel arrive at his discoveries? *Nat Genet* **54**: 926–933.
- Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. (2012) Landscape of transcription in human cells. *Nature* **489**: 101–108.
- Doniger SW, Kim HS, Swain D, Corcuera D, Williams M, Yang S-P, et al. (2008) A Catalog of Neutral and Deleterious Polymorphism in Yeast. *PLOS Genet* **4**: e1000183.
- Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, et al. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**: 78–81.
- Duan S-F, Han P-J, Wang Q-M, Liu W-Q, Shi J-Y, Li K, et al. (2018) The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. *Nat Commun* **9**: 2690.
- Dujon B (1992) Altogether now – sequencing the yeast genome. *Curr Biol* **2**: 279–281.
- Dujon B (1996) The yeast genome project: What did we learn? *Trends Genet* **12**: 263–270.
- Dujon B (2010) Yeast evolutionary genomics. *Nat Rev Genet* **11**: 512–524.
- Dujon B (2015) Basic principles of yeast genomics, a personal recollection. *FEMS Yeast Res* **15**: fov047.
- Dujon B, Cooper TG (2026) Yeasty memories, yeast genomes and beyond—from controversy to collaboration. *FEMS Yeast Res* **26**: foaf073.
- Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, et al. (2004) Genome evolution in yeasts. *Nature* **430**: 35–44.
- Dunn B, Levine RP, Sherlock G (2005) Microarray karyotyping of commercial wine yeast strains reveals shared, as well as unique, genomic signatures. *BMC Genomics* **6**: 53.
- Dunn B, Richter C, Kvitek DJ, Pugh T, Sherlock G (2012) Analysis of the *Saccharomyces cerevisiae* pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. *Genome Res* **22**: 908–924.

- Dutta A, Garin M, Loegler V, Brach G, Friedrich A, Yoshimura M, et al. (2026) Population-scale chemical response revealed by a barcoded yeast collection. *Nat Commun*: 2025.12.03.691428.
- Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, et al. (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res* **30**: 69–72.
- Dymond JS, Richardson SM, Coombes CE, Babatz T, Muller H, Annaluru N, et al. (2011) Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. *Nature* **477**: 471–476.
- Ehrenreich IM, Torabi N, Jia Y, Kent J, Martis S, Shapiro JA, et al. (2010) Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature* **464**: 1039–1042.
- Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133–138.
- Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, et al. (2020) Pangenome Graphs. *Annu Rev Genomics Hum Genet* **21**: 139–162.
- Engel SR, Aleksander S, Nash RS, Wong ED, Weng S, Miyasato SR, et al. (2025) Saccharomyces Genome Database: Advances in genome annotation, expanded biochemical pathways, and other key enhancements. *Genetics* **229**: iyae185.
- Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, et al. (2013) The Reference Genome Sequence of *Saccharomyces cerevisiae*: Then and Now. *G3 GenesGenomesGenetics* **4**: 389–398.
- Engel SR, Wong ED, Nash RS, Aleksander S, Alexander M, Douglass E, et al. (2022) New data and collaborations at the Saccharomyces Genome Database: Updated reference genome, alleles, and the Alliance of Genome Resources. *Genetics* **220**: iyab224.
- Fay JC, Benavides JA (2005) Evidence for Domesticated and Wild Populations of *Saccharomyces cerevisiae*. *PLOS Genet* **1**: e5.
- Fay JC, Liu P, Ong GT, Dunham MJ, Cromie GA, Jeffery EW, et al. (2019) A polyploid admixed origin of beer yeasts derived from European and Asian wine populations. *PLOS Biol* **17**: e3000147.
- Fisher R (1918) The correlation between relatives on the supposition of mendelian inheritance. *Trans. Roy. Soc.*
- Fisher RA (1919) The causes of human variability. *Eugen Rev* **10**: 213–220.
- Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. (1995) Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd. *Science* **269**: 496–512.
- Fogel S, Welch JW, Cathala G, Karin M (1983) Gene amplification in yeast: CUP1 copy number regulates copper resistance. *Curr Genet* **7**: 347–355.
- Fournier T, Abou Saada O, Hou J, Peter J, Caudal E, Schacherer J (2019) Extensive impact of low-frequency variants on the phenotypic landscape at population-scale. *eLife* **8**.
- Foury F, Roganti T, Lecrenier N, Purnelle B (1998) The complete sequence of the mitochondrial genome of *Saccharomyces cerevisiae*. *FEBS Lett*.
- Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**: D766–D773.
- Franklin RE, Gosling RG (1953) Molecular configuration in sodium thymonucleate. *Nature* **171**: 740–741.
- Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851–861.

- Gaisne M, Bécam A-M, Verdière J, Herbert CJ (1999) A 'natural' mutation in *Saccharomyces cerevisiae* strains derived from S288c affects the complex regulatory gene HAP1 (CYP1): *Curr Genet* **36**: 195–200.
- Gallone B, Steensels J, Prah T, Soriaga L, Saels V, Herrera-Malaver B, et al. (2016) Domestication and Divergence of *Saccharomyces cerevisiae* Beer Yeasts. *Cell* **166**: 1397–1410.e16.
- Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, et al. (2018) Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* **15**: 201–206.
- Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, et al. (2000) Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Mol Biol Cell* **11**: 4241–4257.
- Gavin A-C, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**: 631–636.
- Génin E (2020) Missing heritability of complex diseases: Case solved? *Hum Genet* **139**: 103–113.
- Ghaemmaghami S, Huh W-K, Bower K, Howson RW, Belle A, Dephoure N, et al. (2003) Global analysis of protein expression in yeast. *Nature* **425**: 737–741.
- Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, et al. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**: 387–391.
- Giaever G, Flaherty P, Kumm J, Proctor M, Nislow C, Jaramillo DF, et al. (2004) Chemogenomic profiling: Identifying the functional interactions of small molecules in yeast. *Proc Natl Acad Sci* **101**: 793–798.
- Gibson G (2012) Rare and common variants: Twenty arguments. *Nat Rev Genet* **13**: 135–145.
- Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, et al. (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics MCP* **11**: O111.016717.
- Giordano F, Aigrain L, Quail MA, Coupland P, Bonfield JK, Davies RM, et al. (2017) De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Sci Rep* **7**: 3935.
- Goddard MR, Anfang N, Tang R, Gardner RC, Jun C (2010) A distinct population of *Saccharomyces cerevisiae* in New Zealand: Evidence for local dispersal by insects and human-aided global dispersal in oak barrels. *Environ Microbiol* **12**: 63–73.
- Goffeau A (2000) Four years of post-genomic life with 6000 yeast genes. *FEBS Lett* **480**: 37–41.
- Goffeau A, Barrell B. G., Bussey H., Davis R. W., Dujon B., Feldmann H., et al. (1996) Life with 6000 Genes. *Science* **274**: 546–567.
- Goffeau A, Vassarotti A (1991) The European project for sequencing the yeast genome. *Res Microbiol* **142**: 901–903.
- Gojković Z, Knecht W, Zameitat E, Warneboldt J, Coutelis J-B, Pynyaha Y, et al. (2004) Horizontal gene transfer promoted evolution of the ability to propagate under anaerobic conditions in yeasts. *Mol Genet Genomics* **271**: 387–393.
- Gonçalves M, Pontes A, Almeida P, Barbosa R, Serra M, Libkind D, et al. (2016) Distinct Domestication Trajectories in Top-Fermenting Beer Yeasts and Wine Yeasts. *Curr Biol* **26**: 2750–2761.
- Gonzales NM, Seo J, Hernandez Cordero AI, St. Pierre CL, Gregory JS, Distler MG, et al. (2018) Genome wide association analysis in a mouse advanced intercross line. *Nat Commun* **9**: 5162.
- González SS, Barrio E, Querol A (2008) Molecular Characterization of New Natural Hybrids of *Saccharomyces cerevisiae* and *S. kudriavzevii* in Brewing. *Appl Environ Microbiol* **74**: 2314–2320.

- Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR (2015) Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res* **25**: 1750–1756.
- Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: Ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**: 333–351.
- Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. (2019) Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet* **51**: 431–444.
- GTEx Consortium (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**: 1318–1330.
- Guan Y, Dunham MJ, Troyanskaya OG (2007) Functional Analysis of Gene Duplications in *Saccharomyces cerevisiae*. *Genetics* **175**: 933–943.
- Gudjonsson A, Gudmundsdottir V, Axelsson GT, Gudmundsson EF, Jonsson BG, Launer LJ, et al. (2022) A genome-wide association study of serum proteins reveals shared loci with common diseases. *Nat Commun* **13**: 480.
- Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between Protein and mRNA Abundance in Yeast. *Mol Cell Biol* **19**: 1720–1730.
- Hall C, Brachet S, Dietrich FS (2005) Contribution of Horizontal Gene Transfer to the Evolution of *Saccharomyces cerevisiae*. *Eukaryot Cell* **4**: 1102–1115.
- Hallin J, Märtens K, Young AI, Zackrisson M, Salinas F, Parts L, et al. (2016) Powerful decomposition of complex traits in a diploid model. *Nat Commun* **7**: 13311.
- Han D-Y, Han P-J, Rumbold K, Koricha AD, Duan S-F, Song L, et al. (2021) Adaptive Gene Content and Allele Distribution Variations in the Wild and Domesticated Populations of *Saccharomyces cerevisiae*. *Front Microbiol* **12**: 631250.
- Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**: 99–104.
- Harbison ST, McCoy LJ, Mackay TF (2013) Genome-wide association study of sleep in *Drosophila melanogaster*. *BMC Genomics* **14**: 281.
- Hartwell LH, Culotti J, Pringle JR, Reid BJ (1974) Genetic Control of the Cell Division Cycle in Yeast. *Science* **183**: 46–51.
- Hegreness M, Meselson M (2007) What Did Sutton See?: Thirty Years of Confusion Over the Chromosomal Basis of Mendelism. *Genetics* **176**: 1939–1944.
- Hénault M, Fogg V, Heasley LR (2026) From junk to deleterious: Natural subtelomeric repeat amplifications impact fitness and cellular phenotypes in yeast. : 2026.06.29.735415.
- Hershey AD, Chase M (1952) INDEPENDENT FUNCTIONS OF VIRAL PROTEIN AND NUCLEIC ACID IN GROWTH OF BACTERIOPHAGE. *J Gen Physiol* **36**: 39–56.
- Higgins P, Grace CA, Lee SA, Goddard MR (2021) Whole-genome sequencing from the New Zealand *Saccharomyces cerevisiae* population reveals the genomic impacts of novel microbial range expansion. *G3 GenesGenomesGenetics* **11**: jkaa027.
- Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, et al. (2008) The Chemical Genomic Portrait of Yeast: Uncovering a Phenotype for All Genes. *Science* **320**: 362–365.
- Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**: 95–108.
- Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. *Genet Med* **4**: 45–61.
- Hou J, Leeuwen J van, Andrews BJ, Boone C (2018) Genetic Network Complexity Shapes Background-Dependent Phenotypic Expression. *Trends Genet TIG* **34**: 578–586.

- Hou J, Sigwalt A, Fournier T, Pflieger D, Peter J, Montigny J de, et al. (2016) The Hidden Complexity of Mendelian Traits across Natural Yeast Populations. *Cell Rep* **16**: 1106–1114.
- Hudson G, Gomez-Duran A, Wilson IJ, Chinnery PF (2014) Recent Mitochondrial DNA Mutations Increase the Risk of Developing Common Late-Onset Human Diseases. *PLOS Genet* **10**: e1004369.
- Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, et al. (2000) Functional Discovery via a Compendium of Expression Profiles. *Cell* **102**: 109–126.
- Huh W-K, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, et al. (2003) Global analysis of protein localization in budding yeast. *Nature* **425**: 686–691.
- Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS (2009) Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **324**: 218–223.
- Istace B, Friedrich A, Agata L d', Faye S, Payen E, Beluche O, et al. (2017) De novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *GigaScience* **6**: giw018.
- Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci* **98**: 4569–4574.
- Jacobus AP, Stephens TG, Youssef P, González-Pech R, Ciccotosto-Camp MM, Dougan KE, et al. (2021) Comparative Genomics Supports That Brazilian Bioethanol *Saccharomyces cerevisiae* Comprise a Unified Group of Domesticated Strains Related to Cachaça Spirit Yeasts. *Front Microbiol* **12**.
- Jacquier A, Legrain P, Dujon B (1992) Sequence of a 10.7 kb segment of yeast chromosome XI identifies the APN1 and the BAF1 loci and reveals one tRNA gene and several new open reading frames including homologs to RAD2 and kinases. *Yeast* **8**: 121–132.
- Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**: 338–345.
- Jain M, Olsen HE, Paten B, Akeson M (2016) The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community. *Genome Biol* **17**: 239.
- Jariani A, Vermeersch L, Cerulus B, Perez-Samper G, Voordeckers K, Van Brussel T, et al. (2020) A New Protocol for Single-Cell RNA-seq Reveals Stochastic Gene Expression during Lag Phase in Budding Yeast. *eLife* **9**: e55320.
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**: 583–589.
- Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, et al. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**: 348–354.
- Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. (2008) Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics* **178**: 1709–1723.
- Kasianowicz JJ, Brandin E, Branton D, Deamer DW (1996) Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci U S A* **93**: 13770–13773.
- Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**: 617–624.
- Khaiwal S, De Chiara M, Barré BP, Barrio-Hernandez I, Stenberg S, Beltrao P, et al. (2025) Predicting natural variation in the yeast phenotypic landscape with machine learning. *Mol Syst Biol* **21**: 1466–1489.

- Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF (1998) Transposable elements and genome organization: A comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res* **8**: 464–478.
- Kim KE, Peluso P, Babayan P, Yeadon PJ, Et. A (2014) Long-read, whole-genome shotgun sequence data for five model organisms. *Sci Data*.
- Kim M, Nguyen TTP, Ahn J-H, Kim G-J, Sim S-C (2021) Genome-wide association study identifies QTL for eight fruit traits in cultivated tomato (*Solanum lycopersicum* L.). *Hortic Res* **8**: 203.
- Kim Y, Saville L, O'Neill K, Garant J-M, Liu Y, Haile-Merhu S, et al. (2025) Nanopore direct RNA sequencing of human transcriptomes reveals the complexity of mRNA modifications and crosstalk between regulatory features. *Cell Genomics* **5**: 100872.
- Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, et al. (2005) Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science* **308**: 385–389.
- Kordi M, Farrokhi N, Ahmadikhah A, Ingvarsson PK, Saidi A, Jahanfar M (2025) Genome-wide association study of rice (*Oryza sativa* L.) inflorescence architecture. *Plant Sci* **352**: 112382.
- Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, et al. (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**: 637–643.
- Kuzmin E, VanderSluis B, Nguyen Ba AN, Wang W, Koch EN, Usaj M, et al. (2020) Exploring whole-genome duplicate gene retention with complex genetic interaction analysis. *Science* **368**: eaaz5667.
- Kuzmin E, VanderSluis B, Wang W, Tan G, Deshpande R, Chen Y, et al. (2018) Systematic analysis of complex genetic interactions. *Science* **360**: eaao1729.
- Kwon BS, Halaban R, Chintamaneni C (1989) Molecular basis of mouse Himalayan mutation. *Biochem Biophys Res Commun* **161**: 252–260.
- Lan Y, Sun R, Ouyang J, Ding W, Kim M-J, Wu J, et al. (2021) AtMAD: Arabidopsis thaliana multi-omics association database. *Nucleic Acids Res* **49**: D1445–D1451.
- Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**–85–199.
- Laubichler MD, Davidson EH (2008) Boveri's long experiment: Sea urchin merogones and the establishment of the role of nuclear chromosomes in development. *Dev Biol* **314**: 1–11.
- Le Jeune C, Lollier M, Demuyter C, Erny C, Legras J-L, Aigle M, et al. (2007) Characterization of natural hybrids of *Saccharomyces cerevisiae* and *Saccharomyces bayanus* var. *uvarum*. *FEMS Yeast Res* **7**: 540–549.
- Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, et al. (2002) Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*. *Science* **298**: 799–804.
- Lee TJ, Liu Y-C, Liu W-A, Lin Y-F, Lee H-H, Ke H-M, et al. (2022) Extensive sampling of *Saccharomyces cerevisiae* in Taiwan reveals ecology and evolution of predomesticated lineages. *Genome Res* **32**: 864–877.
- Legras J-L, Galeote V, Bigey F, Camarasa C, Marsit S, Nidelet T, et al. (2018) Adaptation of *S. cerevisiae* to Fermented Food Environments Reveals Remarkable Genome Plasticity and the Footprints of Domestication. *Mol Biol Evol* **35**: 1712–1727.
- Legras J-L, Merdinoglu D, Cornuet J-M, Karst F (2007) Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. *Mol Ecol*.
- Li N, He Q, Wang J, Wang B, Zhao J, Huang S, et al. (2023) Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nat Genet* **55**: 852–860.

- Li Z, Vizeacoumar FJ, Bahr S, Li J, Warringer J, Vizeacoumar FS, et al. (2011) Systematic exploration of essential yeast gene function with temperature-sensitive mutants. *Nat Biotechnol* **29**: 361–367.
- Libkind D, Hittinger CT, Valério E, Gonçalves C, Dover J, Johnston M, et al. (2011) Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. *Proc Natl Acad Sci* **108**: 14539–14544.
- Lindgren CC (1949) *The Yeast Cell: Its Genetics and Cytology*. Education Publishers Inc.
- Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011) FaST linear mixed models for genome-wide association studies. *Nat Methods* **8**: 833–835.
- Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, et al. (2009) Population genomics of domestic and wild yeasts. *Nature* **458**: 337–341.
- Liti G, Louis EJ (2012) Advances in Quantitative Trait Analysis in Yeast. *PLOS Genet* **8**: e1002912.
- Liu H, Styles CA, Fink GR (1996) *Saccharomyces Cerevisiae* S288C Has a Mutation in *FL08*, a Gene Required for Filamentous Growth. *Genetics* **144**: 967–978.
- Liu L, Wang J, Levin MJ, Sinnott-Armstrong N, Zhao H, Zhao Y, et al. (2019) The origins of specialized pottery and diverse alcohol fermentation techniques in Early Neolithic China. *Proc Natl Acad Sci* **116**: 12767–12774.
- Llorente B, Durrens P, Malpertuy A, Aigle M, Artiguenave F, Blandin G, et al. (2000) Genomic Exploration of the Hemiascomycetous Yeasts: 20. Evolution of gene redundancy compared to *Saccharomyces cerevisiae*. *FEBS Lett* **487**: 122–133.
- Loegler V, Friedrich A, Schacherer J (2024) Overview of the *Saccharomyces cerevisiae* population structure through the lens of 3,034 genomes. *G3 GenesGenomesGenetics* **14**: jkae245.
- Loegler V, Thiele P, Teyssonnière E, Tsouris A, Brach G, Cruaud C, et al. (2025) From genotype to phenotype with 1,086 near telomere-to-telomere yeast genomes. *Nature*: 1–10.
- Louis EJ (2016) Historical Evolution of Laboratory Strains of *Saccharomyces cerevisiae*. *Cold Spring Harb Protoc* **2016**: pdb.top077750.
- Lowe TM, Eddy SR (1997) tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res* **25**: 955–964.
- Lowry OH, Rosebrough NJ, Farr AL, Randall RJ (1951) Protein measurement with the Folin phenol reagent. *J Biol Chem* **193**: 265–275.
- Luckey JA, Drossman H, Kostichka AJ, Mead DA, D’Cunha J, Norris TB, et al. (1990) High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Res* **18**: 4417–4421.
- Ludlow CL, Cromie GA, Garmendia-Torres C, Sirr A, Hays M, Field C, et al. (2016) Independent Origins of Yeast Associated with Coffee and Cacao Fermentation. *Curr Biol* **26**: 965–971.
- Luo T, Xu M, Wang M, Chen F, Shi J (2026) Systematic evaluation of computational tools for multitype RNA modification detection using nanopore direct RNA sequencing. *Nat Methods* **23**: 438–450.
- Lynch M (2006) The Origins of Eukaryotic Gene Structure. *Mol Biol Evol* **23**: 450–468.
- Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, et al. (2012) The *Drosophila melanogaster* Genetic Reference Panel. *Nature* **482**: 173–178.
- Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**: 1202–1214.
- Magwene PM, Willis JH, Kelly JK (2011) The Statistics of Bulk Segregant Analysis Using Next Generation Sequencing. *PLOS Comput Biol* **7**: e1002255.
- Maher B (2008) Personal genomes: The case of the missing heritability. *Nature* **456**: 18–21.

- Mandl J, Orenstein Y (2026) ExoShorkie: Predicting RNA-seq Coverage of Exogenous Genomes in Yeast by Transfer Learning. : 2026.01.25.701486.
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. (2009) Finding the missing heritability of complex diseases. *Nature* **461**: 747–753.
- Marcet-Houben M, Gabaldón T (2015) Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. *PLOS Biol* **13**: e1002220.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–380.
- Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, Kjaer TR, et al. (2017) Rare and low-frequency coding variants alter human adult height. *Nature* **542**: 186–190.
- Marr RA, Moore J, Formby S, Martiniuk JT, Hamilton J, Ralli S, et al. (2023) Whole genome sequencing of Canadian *Saccharomyces cerevisiae* strains isolated from spontaneous wine fermentations reveals a new Pacific West Coast Wine clade. *G3 GenesGenomesGenetics* **13**: jkad130.
- Marsit S, Dequin S (2015) Diversity and adaptive evolution of *Saccharomyces* wine yeast: A review. *FEMS Yeast Res* **15**: fov067.
- Masneuf I, Hansen J, Groth C, Piskur J, Dubourdieu D (1998) New Hybrids between *Saccharomyces Sensu Stricto* Yeast Species Found among Wine and Cider Production Strains. *Appl Environ Microbiol* **64**: 3887–3892.
- Matthews LJ, Turkheimer E (2022) Three Legs of the Missing Heritability Problem. *Stud Hist Philos Sci* **93**: 183–191.
- Maxam AM, Gilbert W (1977) A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74**: 560–564.
- Maxam AM, Gilbert W (1980) Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol* **65**: 499–560.
- Mendel G (1865) Versuche über Pflanzen-Hybriden. *Verhandlungen Naturforschenden Vereines Brunn*: 41.
- Mergner J, Frejno M, List M, Papacek M, Chen X, Chaudhary A, et al. (2020) Mass-spectrometry-based draft of the *Arabidopsis* proteome. *Nature* **579**: 409–414.
- Messner CB, Demichev V, Muenzner J, Aulakh SK, Barthel N, Röhl A, et al. (2023) The proteomic landscape of genome-wide genetic perturbations. *Cell* **186**: 2018-2034.e21.
- Metzker ML (2010) Sequencing technologies – the next generation. *Nat Rev Genet* **11**: 31–46.
- Mewes HW, Albermann K, Bähr M, Frishman D, Gleissner A, Hani J, et al. (1997) Overview of the yeast genome. *Nature* **387**: 7–8.
- Mick E, McGough J, Deutsch CK, Frazier JA, Kennedy D, Goldberg RJ (2014) Genome-Wide Association Study of Proneness to Anger. *PLoS One* **9**: e87257.
- Mnaimneh S, Davierwala AP, Haynes J, Moffat J, Peng W-T, Zhang W, et al. (2004) Exploration of Essential Gene Functions via Titratable Promoter Alleles. *Cell* **118**: 31–44.
- Monzó C, Liu T, Conesa A (2025) Transcriptomics in the era of long-read sequencing. *Nat Rev Genet* **26**: 681–701.
- Morgan TH (1911) Random Segregation Versus Coupling in Mendelian Inheritance. *Science* **34**: 384–384.
- Morgan TH, Sturtevant AH, Muller J, Bridges CB (1915) The Mechanism of Mendelian Heredity. *Science* **44**: 536–543.

- Mortimer RK, Johnston JR (1986) GENEALOGY OF PRINCIPAL STRAINS OF THE YEAST GENETIC STOCK CENTER. *Genetics* **113**: 35–43.
- Mrnjavac N, Martin WF (2026) The central dogma in reverse. *Proc Natl Acad Sci* **123**: e2604888123.
- Muenzner J, Trébulle P, Agostini F, Zauber H, Messner CB, Steger M, et al. (2024) Natural proteome diversity links aneuploidy tolerance to protein turnover. *Nature* **630**: 149–157.
- Müllder M, Calvani E, Alam M, Wang R, Eckerstorfer F, Zelezniak A, et al. (2016) Functional Metabolomics Describes the Yeast Biosynthetic Regulome. *Cell* **167**: 553–565.e12.
- Muller HJ (1927) Artificial Transmutation of the Gene. *Science* **66**: 84–87.
- Muller L a. H, Lucas JE, Georgianna DR, McCUSKER JH (2011) Genome-wide association analysis of clinical vs. nonclinical origin provides insights into *Saccharomyces cerevisiae* pathogenesis. *Mol Ecol*.
- Muñoz P, Roldán-Guerra FJ, Verma S, Ruiz-Velázquez M, Torreblanca R, Oiza N, et al. (2024) Genome-wide association studies in a diverse strawberry collection unveil loci controlling agronomic and fruit quality traits. *Plant Genome* **17**: e20509.
- Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, et al. (2010) From Noncoding Variant to Phenotype via SORT1 at the 1p13 Cholesterol Locus. *Nature* **466**: 714–719.
- Nadal-Ribelles M, Solé C, Díez-Villanueva A, Stephan-Otto Attolini C, Matas Y, Steinmetz L, et al. (2025) A single-cell resolved genotype-phenotype map using genome-wide genetic and environmental perturbations. *Nat Commun* **16**: 2645.
- Naji MM, Sorin V, Grohs C, Fritz S, Klopp C, Faraut T, et al. (2026) Integrating Structural Variants into Sequence-Based GWAS Using a Pangenome and Imputation Framework in French Dairy Cattle. : 2026.01.18.700144.
- Nashima K, Omine Y, Shirasawa K, Sato T, Yamada M, Shoda M, et al. (2024) Genome-wide association study of pineapple breeding population. *Sci Hortic* **338**: 113757.
- Nature Methods (2023) Method of the Year 2022: Long-read sequencing. *Nat Methods* **20**: 1–1.
- Naumov G I., Masneuf I, S. Naumova E, Aigle M, Dubourdiou D (2000) Association of *Saccharomyces bayanus* var. *uvarum* with some French wines: Genetic analysis of yeast populations. *Res Microbiol* **151**: 683–691.
- Nguyen Ba AN, Lawrence KR, Rego-Costa A, Gopalakrishnan S, Temko D, Michor F, et al. (2022) Barcoded bulk QTL mapping reveals highly polygenic and epistatic architecture of complex traits in yeast. *eLife* **11**: e73983.
- Niarchou M, Gustavson DE, Sathirapongsasuti JF, Anglada-Tort M, Eising E, Bell E, et al. (2022) Genome-wide association study of musical beat synchronization demonstrates high polygenicity. *Nat Hum Behav* **6**: 1292–1309.
- Novick P, Schekman R (1979) Secretion and cell-surface growth are blocked in a temperature-sensitive mutant of *Saccharomyces cerevisiae*. *Proc Natl Acad Sci* **76**: 1858–1862.
- Novo M, Bigey F, Beyne E, Galeote V, Gavory F, Mallet S, et al. (2009) Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* EC1118. *Proc Natl Acad Sci* **106**: 16333–16338.
- Nowak R (1995) Entering the Postgenome Era. *Science* **270**: 368–369.
- O'Donnell S, Yue J-X, Saada OA, Agier N, Caradec C, Cokelaer T, et al. (2023) Telomere-to-telomere assemblies of 142 strains characterize the genome structural landscape in *Saccharomyces cerevisiae*. *Nat Genet* **55**: 1390–1399.
- O'Farrell PH (1975) High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* **250**: 4007–4021.

- Oliver SG, Aart QJM van der, Agostoni-Carbone ML, Aigle M, Alberghina L, Alexandraki D, et al. (1992) The complete DNA sequence of yeast chromosome III. *Nature* **357**: 38–46.
- Olson MV, Dutchik JE, Graham MY, Brodeur GM, Helms C, Frank M, et al. (1986) Random-clone strategy for genomic restriction mapping in yeast. *Proc Natl Acad Sci* **83**: 7826–7830.
- Palmer DS, Zhou W, Abbott L, Wigdor EM, Baya N, Churchhouse C, et al. (2023) Analysis of genetic dominance in the UK Biobank. *Science* **379**: 1341–1348.
- Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, Loose M (2021) Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat Biotechnol* **39**: 442–450.
- Pelechano V, Wei W, Steinmetz LM (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**: 127–131.
- Percudani R, Pavesi A, Ottonello S (1997) Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. *J Mol Biol* **268**: 322–330.
- Pérez-Ortín JE, Querol A, Puig S, Barrio E (2002) Molecular Characterization of a Chromosomal Rearrangement Involved in the Adaptive Evolution of Yeast Strains. *Genome Res* **12**: 1533–1539.
- Peris D, Lopes CA, Arias A, Barrio E (2012) Reconstruction of the Evolutionary History of *Saccharomyces Cerevisiae* x *S. kudriavzevii* Hybrids Based on Multilocus Sequence Analysis. *PLoS One* **7**: e45527.
- Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, et al. (2018) Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **556**: 339–344.
- Peter J, Schacherer J (2016) Population genomics of yeasts: Towards a comprehensive view across a broad evolutionary scale. *Yeast*.
- Pfeiffer T, Morley A (2014) An evolutionary perspective on the Crabtree effect. *Front Mol Biosci* **1**.
- Piotrowski JS, Li SC, Deshpande R, Simpkins SW, Nelson J, Yashiroda Y, et al. (2017) Functional Annotation of Chemical Libraries across Diverse Biological Processes. *Nat Chem Biol* **13**: 982–993.
- Piškur J (2001) Origin of the duplicated regions in the yeast genomes. *Trends Genet* **17**: 302–303.
- Planta RJ, Mager WH (1998) The list of cytoplasmic ribosomal proteins of *Saccharomyces cerevisiae*. *Yeast* **14**: 471–477.
- Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, et al. (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122**: 517–527.
- Preiss R, Tyrawa C, Krogerus K, Garshol LM, Merwe G van der (2018) Traditional Norwegian Kveik Are a Genetically Distinct Group of Domesticated *Saccharomyces cerevisiae* Brewing Yeasts. *Front Microbiol* **9**.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909.
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.
- Pryde FE, Gorham HC, Louis EJ (1997) Chromosome ends: All the same under their caps. *Curr Opin Genet Dev* **7**: 822–828.
- Rheinberger H-J (2008) Heredity and its entities around 1900. *Stud Hist Philos Sci Part A* **39**: 370–374.
- Richardson SM, Mitchell LA, Stracquandano G, Yang K, Dymond JS, DiCarlo JE, et al. (2017) Design of a synthetic yeast genome. *Science* **355**: 1040–1044.

- Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, Martin NW, et al. (2013) GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment. *Science* **340**: 1467–1471.
- Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Séraphin B (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol* **17**: 1030–1032.
- Rockman MV, Skrovanek SS, Kruglyak L (2010) Selection at Linked Sites Shapes Heritable Phenotypic Variation in *C. elegans*. *Science* **330**: 372–376.
- Ronaghi M, Uhlén M, Nyrén P (1998) A sequencing method based on real-time pyrophosphate. *Science* **281**: 363, 365.
- Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**: 348–352.
- Ruffieux M, Brandt N, Gutierrez A, Wolfe BE, Dunn RR, Heil CS (2026) Insights from sourdough redefine the domestication landscape of baker's yeast. : 2026.04.29.721752.
- Salazar AN, Vries AR, Gorter de, Broek M van den, Wijsman M, Torre Cortés P de la, Brickwedde A, et al. (2017) Nanopore sequencing enables near-complete de novo assembly of *Saccharomyces cerevisiae* reference strain CEN.PK113-7D. *FEMS Yeast Res* **17**: fox074.
- Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, et al. (1977a) Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**: 687–695.
- Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**: 441–448.
- Sanger F, Coulson AR, Friedmann T, Air GM, Barrell BG, Brown NL, et al. (1978) The nucleotide sequence of bacteriophage ϕ X174. *J Mol Biol* **125**: 225–246.
- Sanger F, Nicklen S, Coulson AR (1977b) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**: 5463–5467.
- Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L (2009) Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* **458**: 342–345.
- Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* **270**: 467–470.
- Schindler D, Walker RSK, Jiang S, Brooks AN, Wang Y, Müller CA, et al. (2023) Design, construction, and functional characterization of a tRNA neochromosome in yeast. *Cell* **186**: 5237–5253.e22.
- Schmidt A, Kochanowski K, Vedelaar S, Ahrné E, Volkmer B, Callipo L, et al. (2016) The quantitative and condition-dependent *Escherichia coli* proteome. *Nat Biotechnol* **34**: 104–110.
- Schwaba T, Clapp Sullivan ML, Akingbuwa WA, Ilves K, Tanksley PT, Williams CM, et al. (2026) Robust inference and correlates from genetic associations with personality. *Nature*.
- Scriber CR (2007) The PAH gene, phenylketonuria, and a paradigm shift. *Hum Mutat* **28**: 831–845.
- Segrè AV, Murray AW, Leu J-Y (2006) High-Resolution Mutation Mapping Reveals Parallel Experimental Evolution in Yeast. *PLOS Biol* **4**: e256.
- Seoighe C, Wolfe KH (1999) Updated map of duplicated regions in the yeast genome. *Gene* **238**: 253–261.
- Sharon D, Tilgner H, Grubert F, Snyder M (2013) A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31**: 1009–1014.
- Shen X-X, Opulente DA, Kominek J, Zhou X, Steenwyk JL, Buh KV, et al. (2018) Tempo and mode of genome evolution in the budding yeast subphylum. *Cell* **175**: 1533–1545.e20.

- Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. (2017) DNA sequencing at 40: Past, present and future. *Nature* **550**: 345–353.
- Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotechnol* **26**: 1135–1145.
- Sherman D, Durrens P, Beyne E, Nikolski M, Souciet J-L (2004) Génolevures: Comparative genomics and molecular evolution of hemiascomycetous yeasts. *Nucleic Acids Res* **32**: D315–D318.
- Shevchenko A, Jensen ON, Podtelejnikov AV, Sagliocco F, Wilm M, Vorm O, et al. (1996) Linking Genome and Proteome by Mass Spectrometry: Large-scale Identification of Yeast Proteins from Two Dimensional Gels. *Proc Natl Acad Sci* **93**: 14440–14445.
- Silventoinen K, Sammalisto S, Perola M, Boomsma DI, Cornes BK, Davis C, et al. (2003) Heritability of adult body height: A comparative study of twin cohorts in eight countries. *Twin Res Off J Int Soc Twin Stud* **6**: 399–408.
- Smith EN, Kruglyak L (2008) Gene–Environment Interaction in Yeast Gene Expression. *PLOS Biol* **6**: e83.
- Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, et al. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* **321**: 674–679.
- Sniegowski PD, Dombrowski PG, Fingerman E (2002) *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. *FEMS Yeast Res* **1**: 299–306.
- Song X-Y, Xia Y, Zhang J-T, Wei X-Y, Qi H, Li L, et al. (2026) Antiviral reverse transcriptase–primase synthesizes protein-templated DNA. : 2026.04.18.719102.
- Souciet J-L, Aigle M, Artiguenave F, Blandin G, Bolotin-Fukuhara M, Bon E, et al. (2000) Genomic Exploration of the Hemiascomycetous Yeasts: 1. A set of yeast species for molecular evolution studies 1. *FEBS Lett* **487**: 3–12.
- Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. (1998) Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Mol Biol Cell* **9**: 3273–3297.
- Stanojević D, Lin D, Nurk S, Sessions PF de, Šikić M (2026) Telomere-to-Telomere Assembly Using HERRO-Corrected Simplex Nanopore Reads. *Nature*: 1–3.
- Stark R, Grzelak M, Hadfield J (2019) RNA sequencing: The teenage years. *Nat Rev Genet* **20**: 631–656.
- Steenwyk J, Rokas A (2017) Extensive Copy Number Variation in Fermentation-Related Genes Among *Saccharomyces cerevisiae* Wine Strains. *G3 GenesGenomesGenetics* **7**: 1475–1485.
- Steinmetz LM, Sinha H, Richards DR, Spiegelman JJ, Oefner PJ, McCusker JH, et al. (2002) Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416**: 326–330.
- Stenseth NChr, Andersson L, Hoekstra HE (2022) Gregor Johann Mendel and the development of modern evolutionary biology. *Proc Natl Acad Sci* **119**: e2201327119.
- Strope PK, Skelly DA, Kozmin SG, Mahadevan G, Stone EA, Magwene PM, et al. (2015) 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.
- Suhre K, Chen Q, Halama A, Mendez K, Dahlin A, Stephan N, et al. (2025) A genome-wide association study of mass spectrometry proteomics using a nanoparticle enrichment platform. *Nat Genet* **57**: 2987–2996.
- Sun BB, Chiou J, Traylor M, Benner C, Hsu Y-H, Richardson TG, et al. (2023) Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**: 329–338.

- Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D (2019) Benefits and limitations of genome-wide association studies. *Nat Rev Genet* **20**: 467–484.
- Tang R, Zhuang Z, Bian J, Ren Z, Ta W, Peng Y (2024) GWAS and Meta-QTL Analysis of Kernel Quality-Related Traits in Maize. *Plants* **13**: 2730.
- Taylor MB, Ehrenreich IM (2014) Genetic Interactions Involving Five or More Genes Contribute to a Complex Trait in Yeast. *PLOS Genet* **10**: e1004324.
- Taylor MB, Ehrenreich IM (2015) Transcriptional Derepression Uncovers Cryptic Higher-Order Genetic Interactions. *PLOS Genet* **11**: e1005606.
- Tengölics R, Szappanos B, Mülleder M, Kalapis D, Grézal G, Sajben C, et al. (2024) The metabolic domestication syndrome of budding yeast. *Proc Natl Acad Sci* **121**: e2313354121.
- Teyssonnière EM, Shichino Y, Mito M, Friedrich A, Iwasaki S, Schacherer J (2024a) Translation variation across genetic backgrounds reveals a post-transcriptional buffering signature in yeast. *Nucleic Acids Res* **52**: 2434–2445.
- Teyssonnière EM, Trébulle P, Muenzner J, Loegler V, Ludwig D, Amari F, et al. (2024b) Species-wide quantitative transcriptomes and proteomes reveal distinct genetic control of gene expression variation in yeast. *Proc Natl Acad Sci* **121**: e2319211121.
- The International HapMap Consortium (2003) The International HapMap Project. *Nature* **426**: 789–796.
- The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* **437**: 1299–1320.
- Thierry A, Fairhead C, Dujon B (1990) The complete sequence of the 8.2 kb segment left of MAT on chromosome III reveals five ORFs, including a gene for a yeast ribokinase. *Yeast* **6**: 521–534.
- Tong AHY, Evangelista M, Parsons AB, Xu H, Bader GD, Pagé N, et al. (2001) Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants. *Science* **294**: 2364–2368.
- Towbin H, Staehelin T, Gordon J (1979) Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: Procedure and some applications. *Proc Natl Acad Sci U S A* **76**: 4350–4354.
- Townsend JP, Cavalieri D, Hartl DL (2003) Population Genetic Variation in Genome-Wide Gene Expression. *Mol Biol Evol* **20**: 955–963.
- Tsouris A, Brach G, Friedrich A, Hou J, Schacherer J (2024a) Diallel panel reveals a significant impact of low-frequency genetic variants on gene expression variation in yeast. *Mol Syst Biol* **20**: 362–373.
- Tsouris A, Brach G, Schacherer J, Hou J (2024b) Non-additive genetic components contribute significantly to population-wide gene expression variation. *Cell Genomics* **4**.
- Tsukada M, Ohsumi Y (1993) Isolation and characterization of autophagy-defective mutants of *Saccharomyces cerevisiae*. *FEBS Lett* **333**: 169–174.
- Tudek A, Krawczyk PS, Mroczek S, Tomecki R, Turtola M, Matylla-Kulińska K, et al. (2021) Global view on the metabolism of RNA poly(A) tails in yeast *Saccharomyces cerevisiae*. *Nat Commun* **12**: 4951.
- Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, et al. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**: 623–627.
- Uffelmann E, Huang QQ, Munung NS, Vries J de, Okada Y, Martin AR, et al. (2021) Genome-wide association studies. *Nat Rev Methods Primer* **1**: 1–21.
- Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. (2015) Tissue-based map of the human proteome. *Science* **347**: 1260419.
- Vaishnav ED, Boer CG de, Molinet J, Yassour M, Fan L, Adiconis X, et al. (2022) The evolution, evolvability and engineering of gene regulatory DNA. *Nature*: 1–9.

- Van Dyke K, Feraru M, Albert FW (2026) High concordance between genetic effects on mRNA and protein abundance. : 2026.05.26.727960.
- VanGuilder HD, Vrana KE, Freeman WM (2008) Twenty-Five Years of Quantitative PCR for Gene Expression Analysis. *BioTechniques* **44**: 619–626.
- Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five Years of GWAS Discovery. *Am J Hum Genet* **90**: 7–24.
- Visscher PM, Bruce Walsh J (2019) Commentary: Fisher 1918: The foundation of the genetics and analysis of complex traits. *Int J Epidemiol* **48**: 10–12.
- Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era – concepts and misconceptions. *Nat Rev Genet* **9**: 255–266.
- Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**: 5–22.
- Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. (2021) Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* **53**: 1300–1310.
- Wainschtein P, Jain D, Zheng Z, Cupples LA, Shadyab AH, McKnight B, et al. (2022) Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat Genet* **54**: 263–273.
- Wainschtein P, Zhang Y, Schwartzentruber J, Kassam I, Sidorenko J, Fiziev PP, et al. (2026) Estimation and mapping of the missing heritability of human phenotypes. *Nature* **649**: 1219–1227.
- Wang J, Hu H, Jiang X, Zhang S, Yang W, Dong J, et al. (2024) Pangenome-Wide Association Study and Transcriptome Analysis Reveal a Novel QTL and Candidate Genes Controlling both Panicle and Leaf Blast Resistance in Rice. *Rice* **17**: 27.
- Wang Q-M, Liu W-Q, Liti G, Wang S-A, Bai F-Y (2012) Surprisingly diverged populations of *Saccharomyces cerevisiae* in natural environments remote from human activity. *Mol Ecol* **21**: 5404–5417.
- Wang W, Wu T, Fan G, Zhang S, Liu S, Jiang S, et al. (2026) Integrating pan-genome analysis, GWAS, and interpretable machine learning to prioritize trait-associated structural variations in *Setaria italica*. *Plant Commun* **7**: 101626.
- Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet* **10**: 57–63.
- Ward CM, Onetto CA, Heuvel SVD, Cuijvers KM, Hale LJ, Borneman AR (2024) Recombination, admixture and genome instability shape the genomic landscape of *Saccharomyces cerevisiae* derived from spontaneous grape ferments. *PLOS Genet* **20**: e1011223.
- Warringer J, Forsmark A, Durbin R, Omholt SW, Louis EJ, Liti G, et al. (2011) Trait Variation in Yeast Is Defined by Population History. *PLoS Genet* **7**: 15.
- Watson JD, Crick FHC (1953) Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**: 737–738.
- Wei W, McCusker JH, Hyman RW, Jones T, Ning Y, Cao Z, et al. (2007) Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. *Proc Natl Acad Sci* **104**: 12825–12830.
- Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, et al. (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**: 713–720.

- Weller CA, Andreev I, Chambers MJ, Park M, Program NCS, Bloom JS, et al. (2023) Highly complete long-read genomes reveal pangenomic variation underlying yeast phenotypic diversity. *Genome Res* **33**: 729–740.
- Weng S, Dong Q, Balakrishnan R, Christie K, Costanzo M, Dolinski K, et al. (2003) Saccharomyces Genome Database (SGD) provides biochemical and structural information for budding yeast proteins. *Nucleic Acids Res* **31**: 216–218.
- Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37**: 1155–1162.
- Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**: 872–876.
- White LK, Hesselberth JR (2022) Modification mapping by nanopore sequencing. *Front Genet* **13**.
- Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, et al. (2014) Mass-spectrometry-based draft of the human proteome. *Nature* **509**: 582–587.
- Wilkins MHF, Stokes AR, Wilson HR (1953) Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids. *Nature* **171**: 738–740.
- Wingo AP, Liu Y, Vattathil SM, Gerasimov ES, Mei Z, Ravindran SP, et al. (2025) Multiancestry brain pQTL fine-mapping and integration with genome-wide association studies of 21 neurologic and psychiatric conditions. *Nat Genet* **57**: 2156–2165.
- Winston F, Dollard C, Ricupero-Hovasse SL (1995) Construction of a set of convenient *Saccharomyces cerevisiae* strains that are isogenic to S288C. *Yeast* **11**: 53–55.
- Winzeler EA, Castillo-Davis CI, Oshiro G, Liang D, Richards DR, Zhou Y, et al. (2003) Genetic Diversity in Yeast Assessed With Whole-Genome Oligonucleotide Arrays. *Genetics* **163**: 79–89.
- Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, et al. (1999) Functional Characterization of the *S. cerevisiae* Genome by Gene Deletion and Parallel Analysis. *Science* **285**: 901–906.
- Wolfe KH (2015) Origin of the Yeast Whole-Genome Duplication. *PLOS Biol* **13**: e1002221.
- Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**: 708–713.
- Wongsurawat T, Jenjaroenpun P, Wanchai V, Nookaew I (2022) Native RNA or cDNA Sequencing for Transcriptomic Analysis: A Case Study on *Saccharomyces cerevisiae*. *Front Bioeng Biotechnol* **10**: 842299.
- Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, et al. (2019) Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods* **16**: 1297–1305.
- Xue A, Wu Y, Zhu Z, Zhang F, Kemper KE, Zheng Z, et al. (2018) Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat Commun* **9**: 2941.
- Yamamoto F, Clausen H, White T, Marken J, Hakomori S (1990) Molecular genetic basis of the histo-blood group ABO system. *Nature* **345**: 229–233.
- Yang C-J, Yeh Y-C, Hsiao C, Liu Y-H, Lu MR, Liu Y-C, et al. (2026a) A haploid wild yeast resource for exploring the natural ecology of *Saccharomyces cerevisiae*. : 2026.02.13.705822.
- Yang K, Geng Y, Zhou W, Zhang L, Xu Z, Wang J, et al. (2026b) Closing the Gap in Autism Genetics: Population-Specific Variants and the Imperative for Global Inclusion. *Biol Psychiatry*.

- Yang L, He W, Zhu Y, Lv Y, Li Y, Zhang Q, et al. (2025) GWAS meta-analysis using a graph-based pan-genome enhanced gene mining efficiency for agronomic traits in rice. *Nat Commun* **16**: 3171.
- Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. (2018) Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum Mol Genet* **27**: 3641–3649.
- Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, et al. (2022) A saturated map of common genetic variants associated with human height. *Nature* **610**: 704–712.
- Yofe I, Weill U, Meurer M, Chuartzman S, Zalckvar E, Goldman O, et al. (2016) One library to make them all: Streamlining the creation of yeast libraries via a SWAp-Tag strategy. *Nat Methods* **13**: 371–378.
- Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**: 203–208.
- Yuan G-C, Liu Y-J, Dion MF, Slack MD, Wu LF, Altschuler SJ, et al. (2005) Genome-Scale Identification of Nucleosome Positions in *S. cerevisiae*. *Science* **309**: 626–630.
- Yue J-X, Li J, Aigrain L, Hallin J, Persson K, Oliver K, et al. (2017) Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat Genet* **49**: 913–924.
- Zhang G, Roberto NM, Lee D, Hahnel SR, Andersen EC (2022) The impact of species-wide gene expression variation on *Caenorhabditis elegans* complex traits. *Nat Commun* **13**: 3462.
- Zhang T, Li H, Jiang M, Hou H, Gao Y, Li Y, et al. (2024) Nanopore sequencing: Flourishing in its teenage years. *J Genet Genomics* **51**: 1361–1374.
- Zhao Y, Coelho C, Hughes AL, Lazar-Stefanita L, Yang S, Brooks AN, et al. (2023) Debugging and consolidating multiple synthetic chromosomes reveals combinatorial genetic interactions. *Cell* **186**: 5220-5236.e16.
- Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. (2017) Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**: 14049.
- Zhu YO, Sherlock G, Petrov DA (2016) Whole Genome Analysis of 132 Clinical *Saccharomyces cerevisiae* Strains Reveals Extensive Ploidy Variation. *G3 GenesGenomesGenetics* **6**: 2421–2434.

Overview of the project

Understanding how genetic variation gives rise to the phenotypic diversity observed within a species is a central goal of modern biology. Genotype and phenotype are connected through successive molecular layers: genetic variation can affect the abundance and structure of transcripts, the proteins they encode, and ultimately cellular and organismal traits, while regulation at each layer reshapes the variation passed downstream. Transcript abundance is both a quantitative molecular phenotype and one of the earliest molecular layers that can be measured systematically across a population, making it a natural entry point for dissecting the genetic basis of complex traits. This relationship between genotype and phenotype is rarely simple: trait architectures span a continuum from single Mendelian determinants to highly polygenic networks, and the same allele can behave differently depending on the genetic background in which it sits. In humans, two decades of GWAS have mapped thousands of variants contributing to complex traits such as stature or disease susceptibility, yet the variants recovered still account for only a fraction of the heritability estimated from families and twins.

This shortfall, known as missing heritability (Manolio et al., 2009), is one motivation for my thesis work, which asks what molecular variation is missed when gene expression is studied using a restricted set of variants, in a single condition, or at a single molecular layer. As detailed in the previous sections, possible sources include limited power to detect small-effect variants, the difficulty of estimating non-additive effects, and the poor representation of rare, low-frequency and structural variants in association studies.

In this context, my thesis examines sources of molecular variation that standard approaches can miss, taking advantage of the powerful yeast model system and a collection of more than a thousand fully sequenced *S. cerevisiae* natural isolates spanning the ecological and genetic diversity of the species (Peter et al., 2018; Loegler et al., 2025).

My work takes the transcriptome as its central axis: it examines the effects of genetic variation on transcript abundance, variation in transcript structure, and the relationship between transcript and protein abundance. The first two chapters focus on the transcriptome, and the third couples it to the proteome (Fig. 1).

In the first chapter, we asked how much of the variation in gene expression between strains is influenced by structural variants, a class often excluded from association studies. Building on a multi-parent design, a half-diallel panel of 26 isolates was built and 323 hybrid transcriptomes covering close to 6,000 genes generated. Short-read genotypes and long-read assemblies of the parental strains provided catalogues of SNPs and SVs that were then tested as candidate eQTLs. Low-frequency SNPs were enriched among eQTLs despite having smaller effects than common SNPs, and 20 SV-eQTLs were detected. The limited number of SVs that could be tested, however, prevented a comprehensive estimate of their contribution. This chapter extends to transcript abundance the low-frequency-variant enrichment previously observed for growth (Fournier et al., 2019), and provides evidence that structural variation contributes to variation in gene expression. This collaborative work, with Andreas Tsouris and me as co-first authors, resulted in a publication (Tsouris et al., 2024); my own contribution covered the genome sequencing, genome assembly and SV detection that underlie

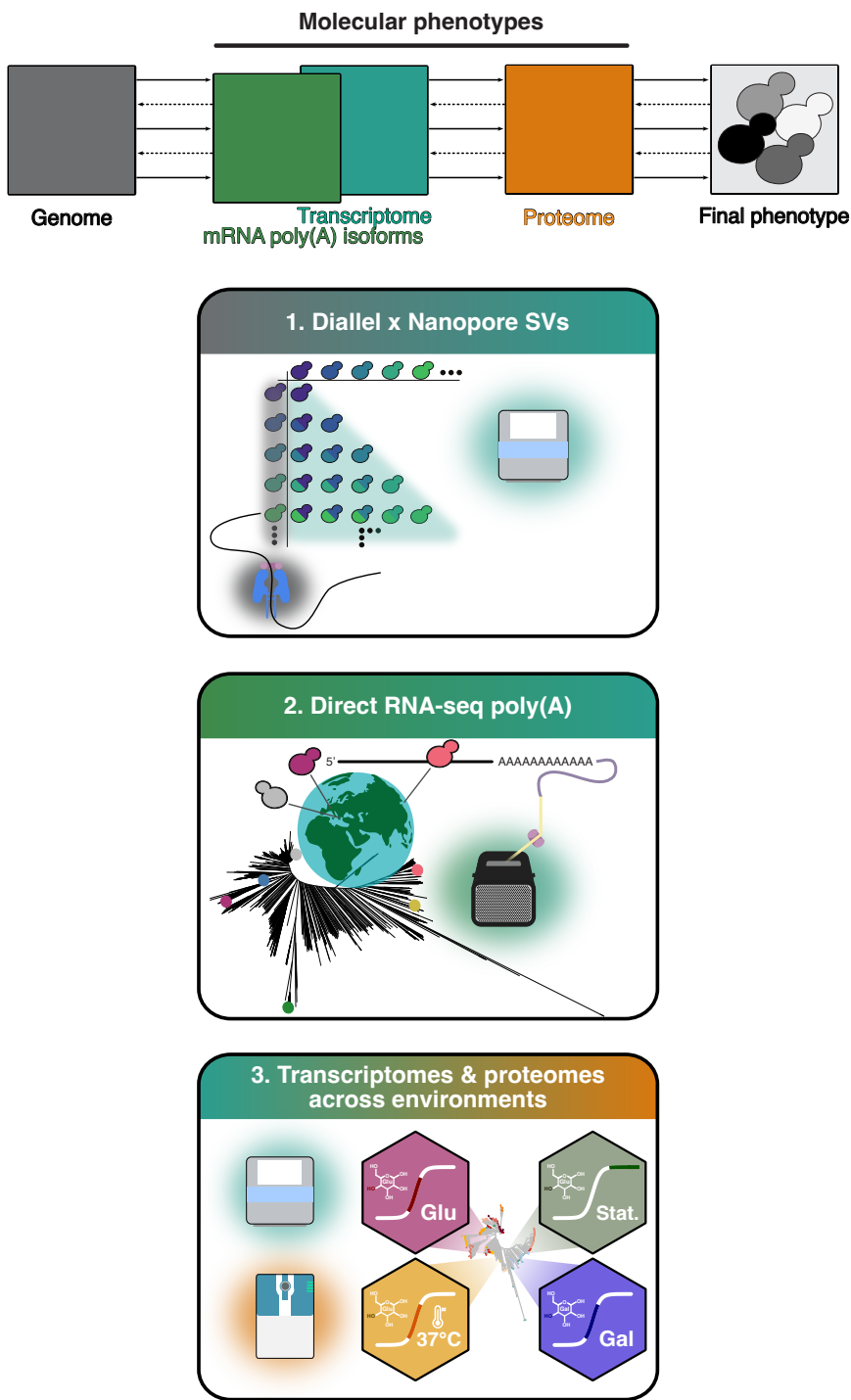


Figure 1 | Overview of the project.

Three axes were explored: genetic variants, mRNA isoforms, and transcript–protein coupling across environments.

the variant catalogue, together with the bench-side generation of part of the hybrid transcriptomes.

In the second chapter, I examine variation within the transcriptome that gene-level abundance does not capture. Beyond its abundance, an mRNA varies in its structure, including its splicing and the position and polyadenylation of its 3' end. Yet while isoform diversity and poly(A) tails have been described in reference strains, these features remain poorly characterized across natural isolates (Pelechano et al., 2013; Tudek et al., 2021). I therefore generated transcriptomes for six natural isolates using Oxford Nanopore direct RNA sequencing, an approach that reads native RNA molecules directly and enables molecule-level measurement of 3' end positions and poly(A)-tail lengths. This chapter compares 3' end regulation across the six isolates and asks which features are conserved and which show strain-associated differences, without establishing the genetic causes of that variation.

In the third chapter, I extend this transcript-centered view to the proteome, measuring both layers from the same cultures. I also add a dimension that previous species-wide catalogues lacked: the environment. At the scale of the species, transcript and protein abundance have each been mapped, but the overlap between their genetic determinants is strikingly low, on the order of a few percent (Caudal et al., 2024; Teyssonnière et al., 2024). These previous resources were generated independently in a single glucose condition. To address these limitations, I developed a protocol to obtain matched transcriptomes and proteomes across four environmental states spanning carbon source, temperature and growth phase: exponential growth in glucose at 30°C as the reference, exponential growth in galactose, exponential growth in glucose at 37°C, and early stationary phase in glucose. The resulting resource comprises 2,721 transcriptomes and 2,668 proteomes, with same-culture measurements for 555 isolates. This allowed transcript and protein variation to be compared across molecular layers and environments. The aim is to determine how the genetic control of each layer changes with the environment, to test the concordance of transcript and protein effects in matched data at the species scale, and to identify genetic effects that single-condition catalogues leave unresolved.

Taken together, my PhD work examines three sources of molecular variation that standard designs can miss: low-frequency and structural variants at the genome-to-transcript step; transcript structure and 3' end usage beyond gene-level abundance; and the species-wide coupling of transcriptome and proteome across environments. Together, these chapters show how expanding the variants, transcript features, environments and molecular layers under study reveals variation that narrower designs leave unresolved.

References

- Caudal É, Loegler V, Dutreux F, Vakirlis N, Teyssonnière É, Caradec C, et al. (2024) Pan-transcriptome reveals a large accessory genome contribution to gene expression variation in yeast. *Nat Genet* **56**: 1278–1287.
- Fournier T, Abou Saada O, Hou J, Peter J, Caudal E, Schacherer J (2019) Extensive impact of low-frequency variants on the phenotypic landscape at population-scale. *eLife* **8**.
- Loegler V, Thiele P, Teyssonnière E, Tsouris A, Brach G, Cruaud C, et al. (2025) From genotype to phenotype with 1,086 near telomere-to-telomere yeast genomes. *Nature*: 1–10.
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. (2009) Finding the missing heritability of complex diseases. *Nature* **461**: 747–753.
- Pelechano V, Wei W, Steinmetz LM (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**: 127–131.
- Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, et al. (2018) Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **556**: 339–344.
- Teyssonnière EM, Trébulle P, Muenzner J, Loegler V, Ludwig D, Amari F, et al. (2024) Species-wide quantitative transcriptomes and proteomes reveal distinct genetic control of gene expression variation in yeast. *Proc Natl Acad Sci* **121**: e2319211121.
- Tsouris A, Brach G, Friedrich A, Hou J, Schacherer J (2024) Diallel panel reveals a significant impact of low-frequency genetic variants on gene expression variation in yeast. *Mol Syst Biol* **20**: 362–373.
- Tudek A, Krawczyk PS, Mroczek S, Tomecki R, Turtola M, Matylla-Kulińska K, et al. (2021) Global view on the metabolism of RNA poly(A) tails in yeast *Saccharomyces cerevisiae*. *Nat Commun* **12**: 4951.

Diallel panel reveals a significant impact of low-frequency genetic variants on gene expression variation in yeast

Andreas Tsouris^{1,*}, Gauthier Brach^{1,*}, Anne Friedrich¹, Jing Hou¹, Joseph Schacherer^{1,2}

¹ Université de Strasbourg, CNRS, GMGM UMR 7156, Strasbourg, France

² Institut Universitaire de France (IUF), Paris, France

* These authors contributed equally.

Published in *Molecular Systems Biology* (2024),

doi:10.1038/s44320-024-00021-0

| Abstract

Unraveling the genetic sources of gene expression variation is essential to better understand the origins of phenotypic diversity in natural populations. Genome-wide association studies identified thousands of variants involved in gene expression variation, however, variants detected only explain part of the heritability. In fact, variants such as low-frequency and structural variants (SVs) are poorly captured in association studies. To assess the impact of these variants on gene expression variation, we explored a half-diallel panel composed of 323 hybrids originated from pairwise crosses of 26 natural *Saccharomyces cerevisiae* isolates. Using short- and long-read sequencing strategies, we established an exhaustive catalog of single nucleotide polymorphisms (SNPs) and SVs for this panel. Combining this dataset with the transcriptomes of all hybrids, we comprehensively mapped SNPs and SVs associated with gene expression variation. While SVs impact gene expression variation, SNPs exhibit a higher effect size with an overrepresentation of low-frequency variants compared to common ones. These results reinforce the importance of dissecting the heritability of complex traits with a comprehensive catalog of genetic variants at the population level.

Introduction

Gene expression variation among individuals corresponds to an essential step linking genetic variation and phenotypic diversity observed in natural populations (Rockman & Kruglyak, 2006; Albert & Kruglyak, 2015; Hill et al., 2021). Dissecting the genetic basis of gene expression variation at the population level is therefore crucial to better understand the genotype–phenotype relationship. Genetic variants or loci associated with gene expression variation (i.e., expression Quantitative Trait Loci, eQTL) have been detected using different mapping strategies, such as linkage and genome-wide association studies (GWAS) (Mackay et al., 2009). Large-scale transcriptomic surveys in model and non-model systems have highlighted the preponderance of genetic variants impacting gene expression variation between individuals (Schadt et al., 2003; West et al., 2007; Rockman et al., 2010; Zhang et al., 2011, 2022; Battle et al., 2014; Vu et al., 2015; Kawakatsu et al., 2016; Aguet et al., 2017; Kita et al., 2017; Albert et al., 2018; Ferraro et al., 2020; Caudal et al., 2024). In general, the transcript level of every gene appears to be influenced by one or more eQTL.

In humans, the genetic basis of gene expression variation was dissected via GWAS on a large dataset of 49 tissues obtained for up to 838 individuals (Ferraro et al., 2020). This study clearly highlighted tissue-specific gene expression patterns and detected a large number of eQTL across tissues. The detected variants are mostly local eQTL (i.e., located close to the genes they influence), while distant eQTL (i.e., located far from the genes they influence) remains difficult to identify in such context due to the large genomes, limited sample size and low statistical power. As most of gene expression variation does not arise from local eQTL, part of the variance is therefore still unexplored. More recently, a population-scale transcriptomic analysis was performed on more than 1000 *Saccharomyces cerevisiae* yeast isolates leading to a deeper view of the genetic control of gene expression (Caudal et al., 2024). Overall, local eQTL were less frequent, representing 26% of the total set of eQTL detected, which is consistent with previous observation in a yeast cross (Albert et al., 2018). Nevertheless, the detected local and distant eQTL together only explained a small fraction of the gene expression heritability (Caudal et al., 2024), which is usually the case for all complex traits (Hindorf et al., 2009; Manolio et al., 2009). Several sources can potentially be considered as the origin of this missing heritability, i.e., the part of the phenotypic variance not explained by associated causal loci. These sources include the low power to detect small effect variants, to estimate non-additive effects (Cordell, 2009; Zuk et al., 2012; Mackay, 2014) but also the fact that rare and low-frequency variants (Pritchard, 2001; Hindorf et al., 2009; Manolio et al., 2009; Gibson, 2012; Walter et al., 2015) as well as structural variants (Peter et al., 2018) are not systematically taken into account in GWAS.

Unlike common variants, rare and low-frequency variants are not adequately captured or tested in standard GWAS analyses, as they are only present in less than 1% or 5% of the population, respectively. In humans, these variants have been shown to play a role and alter adult height for example (Marouli et al., 2017; Akiyama et al., 2019). In addition, recent estimates of the impact of rare variants on the heritability for two human traits (height and body mass index) from whole-genome sequence data on

25,465 unrelated individuals of European ancestry confirmed that these variants are probably a major source of the missing heritability (Wainschtein et al., 2022). In the *S. cerevisiae* yeast model, the effect of rare and low-frequency variants on phenotypic variance has also been tested, since a bias towards this type of variants is observed, with more than 90% of the SNPs having a minor allele frequency (MAF) lower than 0.05 in a dataset of 1011 yeast genomes (Peter et al., 2018). Based on this resource, two independent surveys have shown that these variants contribute disproportionately to growth variation in a large number of conditions observed across yeast natural isolates (Bloom et al., 2019; Fournier et al., 2019). Although there is now strong evidence for the impact of such variants on organismal traits, their effect on molecular traits, such as gene expression variation, has been poorly studied at the population level.

Regarding the structural variants (SVs), they have long been recognized as an important source of genetic diversity and phenotypic variation in all organisms (Weischenfeldt et al., 2013). A few studies have provided catalogs of SVs in large populations of various organisms, including human, tomato, and yeast (Alonge et al., 2020; Li et al., 2023; Liao et al., 2023; O'Donnell et al., 2023), however, these catalogs are still non-exhaustive. Indeed, SVs are difficult to systematically and exhaustively detect for large natural populations and thus they are often excluded from complex trait association studies. In humans, certain SVs have been identified as involved in various diseases and are generally presumed to act through their effects on gene expression (Weischenfeldt et al., 2013). The impact of SVs on gene expression variation was estimated by exploring the effect of 61,668 SVs in 613 individuals on gene expression (GTEx Consortium, 2017; Scott et al., 2021). It was found that common SVs are causal for 2.66% of the eQTL and that SVs often affect multiple nearby genes (Scott et al., 2021). The impact of SVs on neighboring genes has also been recently studied in other organisms, such as tomato and yeast, and the results have shown variable effects on gene expression depending on the type of SVs (Alonge et al., 2020; O'Donnell et al., 2023). Although these studies represent the most comprehensive analysis of the impact of SVs on gene expression to date, they are still limited because either only part of the SVs are detected and not all SV-gene expression associations can be tested. Consequently, these studies do not provide a comprehensive view of the effect of SVs on the transcriptional landscape at the population level.

Understanding the effect of low-frequency variants as well as SVs on gene expression variation remains a challenge but it should provide deeper insights into the molecular basis of phenotypic diversity. Here, we took advantage of a diallel hybrid design in the *Saccharomyces cerevisiae* yeast model, which allows for exhaustive characterization of these two types of variants and their associations with gene expression variation in the population. Our findings show the need for a comprehensive catalog of genetic variants to identify the genetic basis of trait variation.

Results

Diallel design and transcript abundance variation

To better understand the sources of missing heritability of complex traits and more specifically of molecular traits such as transcript abundance, we sought to investigate the underlying genetic architecture of gene expression variation through the use of a diallel panel. In this context, a diallel cross panel was constructed by crossing a set of 26 genetically diverse *S. cerevisiae* natural isolates (Fig. 1, Dataset EV1). To capture a wide range of genetic diversity, different isolates from various ecological (e.g., beer brewery, fruit, clinical samples), and different geographical locations (e.g., Ireland, China, USA) were selected (Fig. S1A-B). The pairwise nucleotide divergence between parental isolates ranges from 0.03 to 1.1%, with an average of 0.59% (Dataset EV2). Parental isolates were crossed in pairs in all possible non-reciprocal combinations, a configuration often referred to as a non-reciprocal diallel cross. This diallel panel resulted in a total of 351 hybrids with 325 heterozygous hybrids coming from the cross of two different parents, and 26 homozygous hybrids resulting from the cross of genetically identical parental isolates.

To quantify gene expression variation in this population, RNA sequencing was performed on all the 351 hybrids (Tsouris et al., 2024). High-quality transcriptomes were obtained for 323 unique hybrids with over one million reads mapped to the reference genome. We obtained the expression levels (as transcript per million or tpm) for 6186 genes that are expressed in at least half of the samples (tpm > 0), including 5740 core ORFs as well as 422 accessory ORFs that are variably present in the generated hybrids (Dataset EV3). The expression abundance (as the mean of $\log_2(\text{tpm}+1)$) and dispersion (as the mean absolute deviation of $\log_2(\text{tpm}+1)$) of each gene in the diallel population correlates with those determined in a large population of 969 *S. cerevisiae* natural isolates ($R=0.79$ and $R=0.72$, respectively) (Fig. S1C-D) (Caudal et al., 2024). In addition, gene expression levels are also correlated between the homozygous hybrids generated in the current dataset and the original parental isolates that were characterized in the previously generated population-level dataset (Caudal et al., 2024) ($R=0.73$) (Fig. S1E).

Detection of the SNP-eQTL using the diallel panel

To determine the genotypes of all the hybrids, we retrieved the biallelic genetic variants of the 26 parental strains from the genomic sequences of the 1011 yeast genomes project (Peter et al., 2018). As a diallel panel leads to a highly structured population, it may introduce some biases when performing genome-wide associations. Singletons, i.e. genetic variants that are only present in a parental genetic background, show a strong linkage throughout the genome. We therefore removed these variants from the genotype matrix, resulting in a SNP matrix of 31,909 variants with a minor allele frequency (MAF) higher than 5%. Using this genotype matrix, we quantified the genome-wide heritability (h^2_g) across the 6186 gene expression traits (Fig. S2A, Dataset EV4). The median h^2_g is 0.28, which is similar to heritability of gene expression estimated in linkage mapping

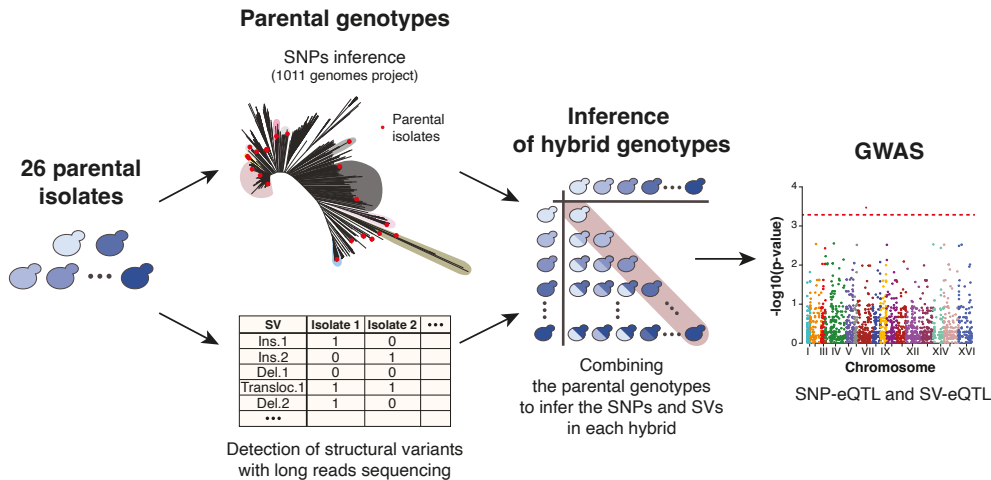


Figure 1 | Workflow for genome-wide association using a diallel panel.

A set of 26 natural isolates were selected as parents for a diallel panel. The SNPs in each isolate were inferred from the data of the 1011 yeast genomes project (Peter et al., 2018) while their SVs were detected using long-reads Oxford Nanopore sequencing. The parental genotypes were combined pairwise to infer the SNPs and SVs in each of the hybrids produced by the diallel panel. Finally, the SNP and SV genotypes of the hybrids were used to perform GWAS on the transcript abundance of 6186 genes.

and GWAS in yeast (median = 0.26) as well as in other organisms such as in humans (mean = 0.03–0.26) (Huan et al., 2015; Albert et al., 2018; Ouwens et al., 2020; Zhang et al., 2022).

To identify SNPs that influence gene expression variation (SNP-eQTL), we performed GWAS with the transcript abundance traits of 6,186 genes and using the SNP matrix as the genotype (Methods). We detected a total of 3,039 SNP-eQTL that influence the level of transcript abundance of 1714 genes (Fig. 2A, Dataset EV5). Most SNP-eQTL impact only one or a few traits while few SNP-eQTL influence more than 20 phenotypes (Fig. S2B). In fact, we detected a total of 4 eQTL hotspots located in the *BCY1*, *GAL2* and *MLP1* genes, while another one is located in the promoter of *GSH2* and are associated to the expression of 83, 65, 62, 57 genes, respectively (Fig. 2B). The number of hotspots detected here is intermediate between the number identified by linkage mapping using single cross segregants and GWAS across a large set of natural isolates (Albert et al., 2018; Caudal et al., 2024). In addition, most traits are influenced by a single SNP-eQTL while very few are controlled by more than 5 (Fig. S2C). Overall, a trait was associated to 1.77 eQTL while each eQTL influenced 2.94 different phenotypes on average. A total of 1032 unique SNPs were associated with the 3039 eQTL (Dataset EV5).

We then distinguished the SNP-eQTL according to their relative position with respect to the gene they impact, with local eQTL being close to the affected gene and distant eQTL located further away or on different chromosomes (see Methods). We found 2952 distant and 36 local eQTL, while 51 eQTL involved HGT genes for which the precise locations of the genes were undetermined (Dataset EV5). Consistent with previous observations,

local SNP-eQTL display larger effect sizes compared to the distant ones (Wilcoxon test, $p\text{-value} = 1.5e^{-4}$) (Fig. 2C) Albert et al., 2018; Caudal et al., 2024).

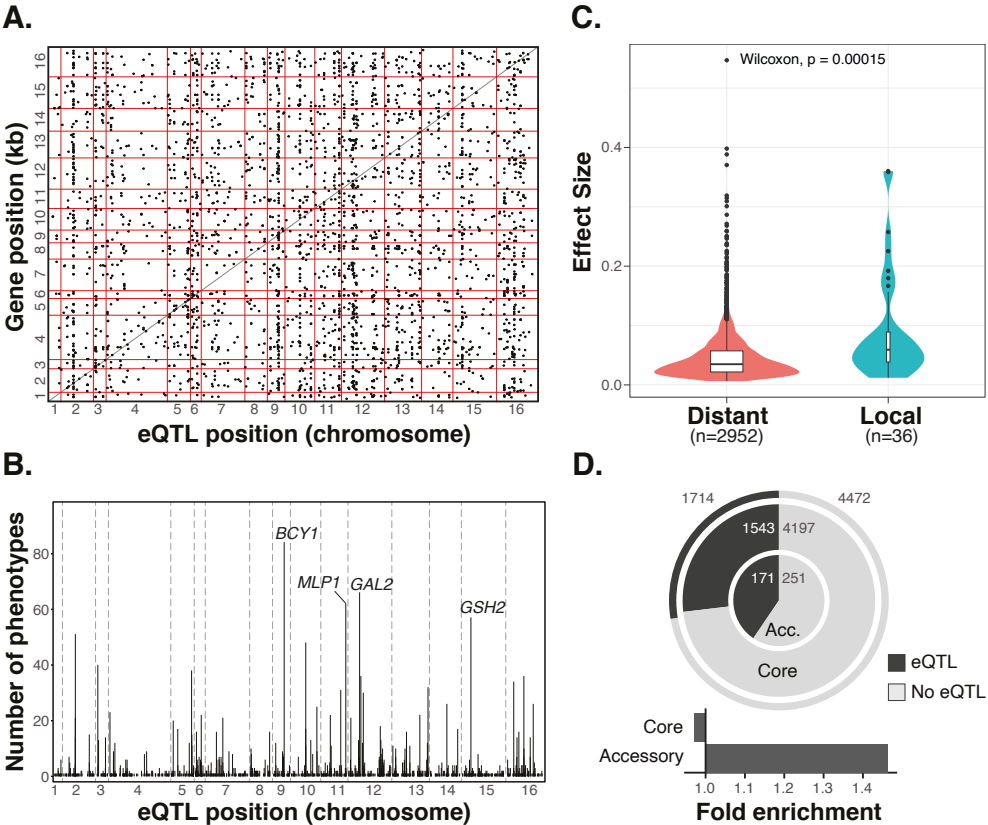


Figure 2 | Genome-wide association results using the SNP matrix.

(A.) Positions of the SNP-eQTL and their associated genes along the genome. Each point represents an eQTL and the horizontal and vertical red lines mark the margins of the chromosomes. The diagonal black line indicates the positions where the eQTL position and gene position are the identical, therefore eQTLs on that line are very close to the gene they influence. (B.) Positions of SNP-eQTL and the number of genes they are associated with. (C.) Comparison of the effect sizes of distant and local-eQTL inferred from GWAS (two-sided Mann-Whitney-Wilcoxon test, $p\text{-value} = 1.5e^{-4}$). For distant-eQTL effect sizes, min = 0.006, 1st quartile = 0.022, median = 0.035, 3rd quartile = 0.057, max = 0.546. For local-eQTL effect sizes, min = 0.012, 1st quartile = 0.038, median = 0.059, 3rd quartile = 0.088, max = 0.36. (D.) The number and proportions of core (outer circle) and accessory genes (inner circle) associated with eQTLs. Fold enrichment of core and accessory genes associated to eQTLs. Significance was calculated using a two-sided Fisher test.

Gene expression regulation of accessory genes has been shown to differ from that of the core genome (Caudal et al., 2024). We therefore compared the likelihood of accessory and core genes to be associated to SNP-eQTL (Fig. 2D). Accessory genes are more likely to be associated with SNP-eQTL (Fisher test, $p\text{-value} = 3.7e^{-3}$). In total, 40.5% of the accessory genes (171 out of 422 accessory genes) considered in GWAS are associated with an eQTL, compared to a 26.7% of all genes (1543 out of 5740 genes), with a 1.52-fold enrichment. In addition, SNP-eQTL associated with

accessory genes have a larger effect size on average than those associated with core genes (Wilcoxon test, $p\text{-value} = 3.7\text{e}^{-6}$). Specifically, the effect size of distant SNP-eQTL controlling accessory genes is larger than those controlling core genes (Wilcoxon test, $p\text{-value} = 3\text{e}^{-10}$). However, no significant difference was observed between the local SNP-eQTL associated with the expression of accessory and core genes (Wilcoxon test, $p\text{-values} = 0.48$) (Fig. S2D).

Impact of the low-frequency SNPs on gene expression variation

Next, we sought to explore the contribution of rare and low-frequency genetic variants (i.e., with a $\text{MAF} < 0.01$ and < 0.05 , respectively) to the observed gene expression variation. In fact, genetic variants considered by GWAS must have a relatively high frequency in the population to be detectable, usually over 0.05 for relatively small datasets (Visscher et al., 2017). Rare and low-frequency variants are therefore excluded from association studies. However, our diallel panel provides a powerful strategy to assess the impact of low-frequency variants that are initially present in the population, i.e., in the 1011 *S. cerevisiae* collection here (Peter et al., 2018). In fact, each parental genotype is present several times, increasing the frequency of these variants in this new population and allowing their detection in GWAS. While the parental genomes carry 1228 low-frequency variants (48 of which are rare variants) as defined in the population of 1011 isolates, none of the genetic variants has a MAF lower than 0.058 in our diallel panel (Datafile 1).

Interestingly, the MAF of the genetic variants identified as SNP-eQTL is much lower than that of the entire set of SNPs considered for GWAS, suggesting that low-frequency variants have a large impact on transcript abundance variation (Fig. 3A). In fact, 8.4% (104 out of 1228) of the significant SNP-eQTL have a MAF below 0.05, whereas only 3.0% (928 out of 30,681) of SNP matrix used for GWAS have a MAF below 0.05 in the natural population (Fig. 3B). This nearly threefold enrichment in low-frequency SNPs in the set of associated variants is consistent with previous studies focusing on the influence of low-frequency variants on yeast growth traits (Bloom et al., 2019; Fournier et al., 2019). Using a diallel panel, we previously have found that 16.8% of growth SNP-QTL were low-frequency variants in the collection of 1011 *S. cerevisiae* isolates (Fournier et al., 2019). Finally, we measured the respective effect size of each SNP-eQTL and we found that even if enrichment is observed, low-frequency variants have a lower effect size compared to the common variants (Fig. 3C).

To further characterize the functional effects of the associated variants, we annotated all variants in the SNP matrix using SnpEff (Cingolani et al., 2012). The majority of variants falls into three annotation groups, namely intergenic (10,098/31,909), synonymous_variant (15,893/31,909) and missense_variant (5501/31,909) (Dataset EV6). The remaining 417 variants mainly belonged to intron related variants and some rare cases of loss of start codon (start_loss, 13 variants) and gain of premature stop codon (stop_gained, 37 variants) (Dataset EV6). Using these annotations, we analyzed the distribution of different variant groups across common vs. low frequency variants (Dataset EV6). We observed no significant differences between

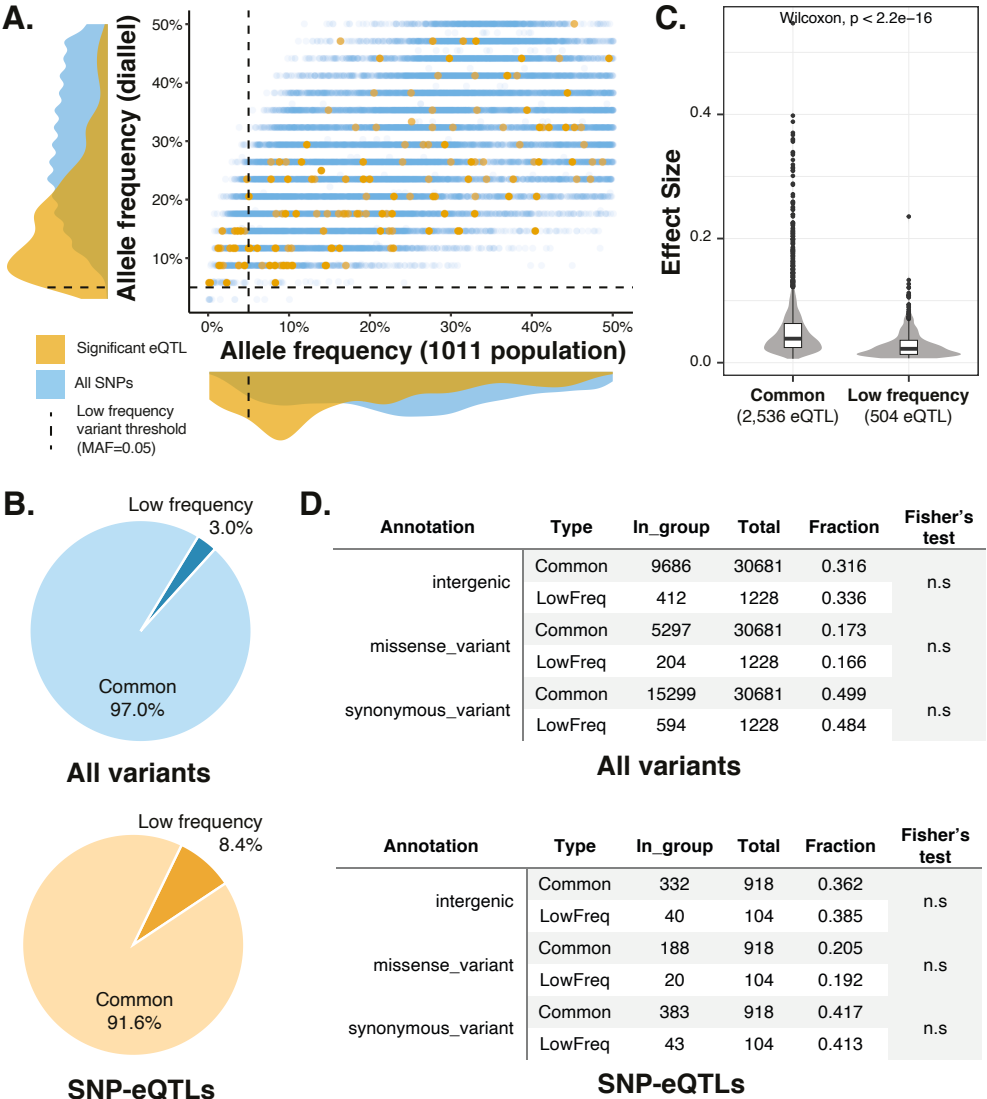


Figure 3 | Low frequency variants and gene expression variation.

(A.) Comparison of the MAF of the SNPs used in GWAS in the diallel population and the natural population (1011 isolates). Orange points represent significant eQTL SNPs whereas blue points represent the remaining SNPs. In the distributions along the axes, the orange distribution covers the significant SNPs, and the blue distribution covers all the SNPs used in GWAS. Vertical and horizontal dashed lines show the threshold of MAF (5%) used to define low-frequency variants. (B.) Comparison of the fraction of low frequency variants in all the SNPs used in GWAS (blue) and the significant eQTLs (orange). (C.) Comparison of the effect size of common and low-frequency SNP-eQTL. The number of associated variants (n) in each category are indicated. For common variants associated eQTL, the effect sizes show min = 0.006, 1st quartile = 0.024, median = 0.039, 3rd quartile = 0.063, max = 0.547. For low frequency variants associated eQTL, the effect sizes show min = 0.007, 1st quartile = 0.013, median = 0.022, 3rd quartile = 0.036, max = 0.235. (D.) Comparison of functional annotations between low frequency and common variants across the full SNP matrix (upper panel) and across the associated SNPs (lower panel). Fisher's exact tests were performed for all common vs. low frequency for each annotation group.

common and low frequency variants across all annotation groups (Dataset EV6, Fig. 3D), indicating that low frequency variants are a random sampling of all SNPs. We performed the same comparison for all eQTL associated SNPs and observed the same trend (Dataset EV7, Fig. 3D). While there is no difference in terms of functional annotation between common and low frequency variants, we do observe that missense variants and intergenic variants are more likely to be associated with an eQTL than synonymous variants. Specifically, missense variants are 1.18-fold overrepresented in eQTL associated SNPs (Fisher's test P-value=0.03), and intergenic variants are 1.15-fold overrepresented in eQTL associated SNPs (Fisher's test P-value=0.02) compare to the full variant matrix.

Characterization of the structural variants in the diallel panel

We then focused on the impact of SVs on gene expression variation using our diallel panel. Accurate and exhaustive identification of SVs in a large population can be very laborious and, in some cases, practically impossible. These variants are therefore only very rarely taken into account in association studies. However, the structure of a yeast diallel panel makes this task much easier because the SVs present in the hybrids correspond to the combination of the SVs present in their parents. Long-read genome sequencing and SV detection thus provide an exhaustive list of the SVs present in each of the 323 hybrids.

We therefore sequenced the genomes of 24 out of 26 parental isolates using a long-read Oxford Nanopore sequencing strategy, as two of the parental strains, S288C and Σ 1278b, were already completely sequenced and assembled (Goffeau et al., 1996; Dowell et al., 2010). On average, we obtained 108,351 reads with a length of 12.8 kb for each parental genome (Dataset EV8). The mean coverage was 66X on average, ranging from 19.7X to 150X for the ABA and ACS isolates, respectively. We then used these long reads to generate de novo assemblies for the entire set of strains (see Methods). Overall, we obtained contiguous assemblies with a N50 of 835 kb and an average genome size of 12.2 Mb (Dataset EV9). In most cases, a single contig covered more than 90% of a chromosome, while in other cases, only a few contigs were needed. (Fig. S3A). The assembly of each parent was then compared to the genome of the reference strain (S288C) using MUM&co to detect all the SVs present in the parental genomes (O'Donnell & Fischer, 2020). We define SVs as variants that are at least 50 bp in size. As the SVs are detected independently in each parent, we also had to identify the presence of the same event in the several genetic backgrounds in order to merge them into the same occurrence. For this purpose, we used Jasmine (Kirsche et al., 2021, 2023), a tool that compares the genomic position as well as the sequences of SVs, and we found a total of 1953 SVs in our set of parental genomes (Dataset EV10).

The SVs were classified into 6 types, namely insertions ($n=1032$ variants), deletions ($n=543$), duplications ($n=250$), contractions ($n=65$), translocation ($n=33$) and inversions ($n=30$) (Fig. 4A). The number of SVs in each parent correlates with the nucleotide divergence between that parent and the reference genome ($R=0.71$, $p\text{-value}=7.30e^{-05}$) (Fig. S3B). While Σ 1278b has the lowest nucleotide diversity

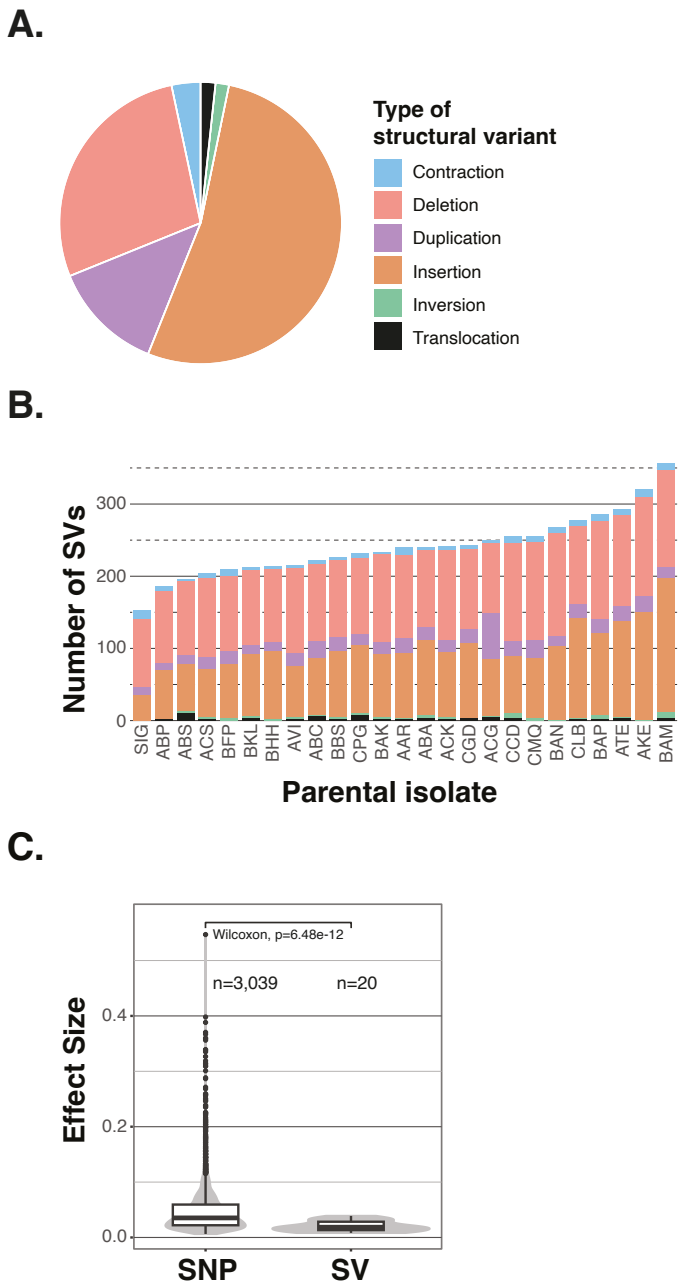
compared to S288C and the lowest number of SVs ($n=153$), the BAM isolate has the highest number of SVs ($n=357$) and corresponds to the most divergent strain with respect to S288C (Fig. 4B). On average, we detected 241 SVs per genome. The distribution of SV types in each parent is similar to their overall abundance. Most SVs are insertions and deletions, while the remaining SVs types only represent a small fraction. As with SNPs, SVs exhibit a bias towards low-frequency alleles (Appendix Fig. S3C). A total of 60.1% of SVs ($n=1190$) are low-frequency variants, with only 763 SVs have a MAF greater than 0.05 (Fig. S3C).

Finally, we also identified SVs that are related to transposable elements, a major source of structural variation (see Methods). Overall, 869 SVs are Ty-related (44.5%), while 1084 (55.5%) are not (Fig. S3D). Ty-related SVs can be separated into two groups according to their length. The first group of SVs are about 340 bp in length and contains the Ty Long Terminal Repeats (LTR), while the second group of SVs are about 6 kb in length and contains the entire or partial Ty elements (Fig. S3D). A total of 64.2% ($n=558$) of Ty related SVs only contain LTR sequences, while the remaining 35.8% ($n=311$) contain partial or complete Ty elements.

Impact of SVs on gene expression phenotypes

Before performing genome-wide associations, we inferred the genotypes of the 323 hybrids by combining the SVs of their parents as it was done previously for the SNPs. We excluded the singletons, i.e., all variants present in only one of the 26 parents, because they would cause many false positive associations due to severe linkage throughout the genome. As a result, we obtained a genotype matrix of 763 SVs across the 323 hybrids. The h^2_g of transcript abundance traits using the SV genotype matrix was 0.021, an order of magnitude lower than the h^2_g obtained with the SNP matrix ($h^2_g=0.28$) (Fig. S4A).

To identify SVs that influence gene expression variation (SV-eQTL), we performed GWAS with the transcript abundance traits of 6186 genes and using the SV matrix as the genotype (see Methods). We only detected 20 significant SV-eQTL that influence the expression of 13 genes (Dataset EV11). A total of 19 unique SVs were associated. The transcript abundance of 10 genes was influenced by one SV-eQTL, while the expression of two genes was influenced by multiple SV-eQTL (Fig. S4B). All SV-eQTL are associated to a single gene except for one that was associated to the transcript abundance of two genes (Fig. S4C). Most SV-eQTL are insertions and deletions ($n=13$ and $n=4$, respectively), while one SV-eQTL is a duplication located on chromosome 13 and another one a contraction on chromosome 15. SV-eQTL involving deletions show higher effect sizes than insertions, contraction and duplications, (Fig. S4D). Half of the SV-eQTL are Ty-related (10 out of the 20 SV-eQTL) but this is not significant compared to the overall abundance of Ty-related SVs (fisher test, $p\text{-value}=1$). Overall, Ty-related SV-eQTL tend to show lower effect sizes compared to non-Ty SV-eQTL for the same variant type (Fig. S4D). Finally, since we assessed the relationship of SNPs and SVs to the same phenotypes across the same population, it was possible to compare the effect of SNPs and SVs on gene expression. In fact, we found that the effect sizes of SV-eQTL are overall smaller than those of SNP-eQTL (Fig. 4C).



1

Figure 4 | Structural variants detected in the parental genomes and effect size comparison. (A.) Fractions of the different types of structural variants identified in the parental strains. (B.) Number of the SVs detected in each parental strain. Each parent is pictured in each stacked bar and colors of the bars represent the different types of SV as above. (C.) Comparison of the effect sizes of SNP-eQTL and SV-eQTL identified through GWAS (two-sided Mann-Whitney-Wilcoxon test, p -value = $6.48e^{-12}$). The sample size n corresponds to the number of SNP- and SV-eQTL and are indicated on the plot.

Discussion

Understanding the genetic basis of gene expression variation is essential to have a better insight into the genotype–phenotype relationship. While large-scale studies detected thousands of genetic variants involved in gene expression variation via genome-wide associations, these variants explain only a small fraction of gene expression variation (GTEx Consortium, 2017; Ferraro et al., 2020). In this context, we sought to investigate the impact of potential sources of this so-called missing heritability. Using a diallel design, we were able to assess the influence of low-frequency as well as structural variants on the variation in transcript abundance. On the one hand, this strategy makes it possible to artificially increase the frequency of genetic variants in a diallel panel and thus to take low-frequency variants into account in association studies. On the other hand, it is possible to easily establish the exhaustive list of SVs present in the hybrids from the parental genotypes. Based on a generated yeast diallel panel composed of 323 hybrids and associated with the corresponding transcriptomes, we carried out a GWAS for each gene expression trait and identified 3,039 SNP-eQTL and 20 SV-eQTL in total.

Regarding the low-frequency and rare variants, we found a strong enrichment of these variants in the set of significantly associated SNPs, i.e., SNP-eQTL. While only 3.0% of the SNPs considered in the association tests have a low frequency in the natural yeast population, we found that 8.4% of all SNPs-eQTL are low-frequency variants, showing a ~3-fold enrichment. These results clearly show the same trend and enrichment already observed for yeast growth phenotypes (Bloom et al., 2019; Fournier et al., 2019). This observation highlights that low-frequency variants strongly contribute to gene expression variation in a natural population. This is very important as much of the detected genetic polymorphisms in a population, such as the 1011 yeast genomes dataset, are low-frequency variants with almost 92.7% of the polymorphic sites associated with a MAF lower than 0.05 (Peter et al., 2018). Moreover, low frequency variants do not differ from common variants in terms of functional annotations, indicating that they are a random sample of all variants. Interestingly, low frequency variants are consistently enriched for both QTL related to growth phenotypes and gene expression traits. Alongside this enrichment, we also found that overall, the effect sizes of these variants are smaller than those of the common variants for gene expression traits, contrasting to the growth phenotypes (Bloom et al., 2019; Fournier et al., 2019). This difference could possibly be due to the rich and lenient growth condition used for the gene expression measurements, contrasting to various stress conditions used for growth phenotypes. Further transcriptomic analyses across different culture condition are still needed to explore more deeply the effects of low-frequency variants in a population.

Our diallel panel, however, showed some limitations in exploring the impact of SVs on gene expression variation. In fact, we could only detect a total of 19 SV-eQTL probably because very few SVs were taken into account compared to SNPs in the genotype matrices (763 SVs versus 31,909 SNPs). By taking into account the size of the genotypic matrices, we can observe that the same fraction of SNP-eQTL and SV-QTL was globally detected, corresponding respectively to 3.2% and 2.5% of associated genetic variants, respectively. It is clear that SVs are widely present in natural populations, as in *S.*

cerevisiae for example. A recent study of 142 telomere-to-telomere genomes from natural isolates characterized approximately 4800 unique SVs and identified 97 SVs that influence the expression of neighboring genes (O'Donnell et al., 2023). The impact of SVs on gene expression has also been shown in humans where SVs are responsible for 2.66% of eQTL and they often affect multiple nearby genes (Scott et al., 2021). The impact of SVs is not limited to gene expression traits, a survey of over 100 tomato genomes has indeed identified hundreds of SV-QTLs impacting several volatile flavor and fruit metabolites (Alonge et al., 2020; Li et al., 2023).

Our diallel design also comes with other limitations. First, the diallel hybrid population is very structured, and in order to avoid biases we had to eliminate many variants that were in linkage, therefore limiting the number of variants used in GWAS. This limit could possibly be circumvented by significantly increasing the size of the panel. Second, the hybrids were not sequenced directly but their genotypes were inferred from their parents. Therefore, any potential genomic modifications that have arisen since the formation of the hybrids were generated (such as any mutation, loss of heterozygosity or other structural events) are not considered in our design. Moreover, as the mitochondrial genomes in the hybrid cannot be easily inferred as the nuclear genome due to potential recombination between the parental types, the contribution of the mitochondrial genome on gene expression is not considered here. Finally, as already mentioned, the number of SVs used for the association studies was probably insufficient to have a global view of the impact of SVs on gene expression.

A population-scale comprehensive survey would now be essential to have an accurate and exhaustive view of the impact of SVs on gene expression variation. The collection of 1011 natural yeast isolates could be an excellent resource for such exploration. Long-read sequencing of this large collection would lead to a species-wide view of the structural variants as well as a comprehensive pangenome graph (Liao et al., 2023; Hickey et al., 2024). The exhaustive catalog of SVs generated could then be used to perform GWAS on a large number of traits. Many traits have already been measured for this large collection of *S. cerevisiae* isolates, including growth phenotypes, as well as molecular traits like transcript and protein abundance (Peter et al., 2018; Caudal et al., 2024; Muenzner et al., 2024; Teyssonnière et al., 2024). A better view of the impact of SVs on traits would definitely improve our understanding of the genotype–phenotype relationship by revealing their real role in missing heritability.

Methods

Diallel panel generation and cell harvesting

The hybrid diallel panel was produced from a series of pairwise crosses between 26 haploid parental strains. Two stable haploid lines were generated for each parental strain, the *MATa* lines carry a KanMX cassette and the *MAT α* lines carry a NatMX cassette, each cassette replacing the HO locus (Fournier et al., 2019). For each cross cells of opposite mating type were patched together on solid YPD media (1% yeast extract, 2% peptone, and 2% glucose) and incubating them at 30 °C overnight. To

select for hybrid cells by selecting the combination of both cassettes the cell patches were then transferred to YPD media containing G418 (200 mg/mL) and nourseothricin (100 mg/mL) and incubated at 30 °C overnight. All procedures were done using the replicating robot ROTOR (Singer Instruments). In total, we obtained 351 genetically unique hybrids.

Hybrids were then incubated in liquid synthetic complete (SC) media with 2% glucose in 96-deep-well plates for cell harvesting. The optical density of each hybrid's culture was measured systematically using a 96-well microplate reader (Tecan Infinite F200 Pro). During log-phase growth ($OD_{600nm} \sim 0.3$), cells were harvested by filtration in 96-well filter plate (Norgen, #40008) where the media was eliminated by applying vacuum (VWR, #16003-836). The filter plates were immediately flash frozen in liquid nitrogen and stored at -80 °C.

cDNA library preparation

We used the Dynabeads® mRNA Direct Kit (ThermoFisher #61012) to extract the mRNA of each hybrid. The cells were lysed using glass beads and were then incubated at 65 °C for 2 min. mRNA was selected with two rounds of hybridization of their polyA tails to magnetic beads coupled to oligo(dT) residues.

cDNA sequencing libraries were prepared with the NEBNext® Ultra™ II Directional RNA Library Prep Kit (NEB, #E7765L) and using the manufacturer's protocol. The concentration of cDNA in each library was quantified using the Qubit™ dsDNA HS Assay Kit (Invitrogen™) in a 96-well plate using a microplate reader (Tecan Infinite F200 Pro) with an excitation frequency of 485 nm and emission of 528 nm. Fragment size was assessed with Bioanalyzer 2100 (Agilent™) using the High sensitivity DNA kit (#5067-4626). We generated sequencing pools containing equimolar fragments from each sample. Lastly, the pools were sequenced for 75 bp single-end with Nextseq 550 (Illumina™) sequencer at the EMBL Genomics Core Facility.

mRNA abundance quantification

For each sample the raw sequencing reads were mapped to a custom reference genome using STAR (Dobin et al., 2013) with the following parameters:

```
--outSAMtype BAM SortedByCoordinate \
--outFilterType BySJout \
--outFilterMultimapNmax 20 \
--outFilterMismatchNmax 4 \
--alignIntronMin 20 \
--alignIntronMax 2000 \
--alignSJoverhangMin 8 \
--alignSJDBoverhangMin 1
```

The custom reference genome combines the 16 chromosomes of the R64_nucl reference genome with the accessory ORFs present in the parental strains ($n=665$) according to the data from the 1011 yeast genomes project (Peter et al., 2018), each accessory ORF is considered as a separate contig. In total, 323 hybrids had enough reads and were used in the subsequent analyses. The reads aligning to each gene of the reference ($n=6285$) and accessory ($n=665$) genomes were counted using the featureCounts function of the R package subread (Liao et al., 2014) with the parameter countMultiMappingReads=F in order to not take multi-mapped reads into account. For a given hybrid, if accessory genes that have orthologs in the reference genome were annotated as absent, we merged their reads counts to those of their reference genome counterparts. mRNA abundance was then normalized by calculating the transcripts per million value (tpm) of each gene. This gave us a list of tpm values for 6917 genes. We filtered out genes that have a zero tpm value in more than half of all samples, leading to a final dataset of 6186 genes (Tsouris et al., 2024).

Functional annotation of the SNP variants

SnEff (v5.1) (Cingolani et al., 2012) was used to annotate variants located in coding regions in the vcf file (reference genome R64-1-1.105). The annotation file in .vcf format can be found in Datafile 2.

Long-reads sequencing of the parental genomes

To carry out long-reads sequencing of the parental isolates with Oxford Nanopore Technologies (ONT) sequencing, we first incubated each parental isolate in 40 mL of YPD (1% yeast extract, 2% peptone and 2% glucose) at 30 °C for 48 h. The cells were then harvested by centrifugation at $7000 \times g$ for 3 min and then resuspended in water before a second round of centrifugation. The cell pellet was resuspended in 4 mL of sorbitol 1 M containing 250 μ L of zymolyase (1000 U/mL) and incubated at 30 °C for 2 h with agitation. Protoplasts were recovered by centrifuging for 5 min at $2000 \times g$ at 4 °C and discarding the supernatant. They were then resuspended in 4 mL of lysis buffer, a solution of Tris-HCl 125 mM, EDTA 62.5 mM, NaCl 0.625 M, 0.05 g PVP40, 1.25% SDS and 10 mg/mL RNaseA. The cell lysis suspension was incubated

at 50 °C for 3 h and was then chilled on ice for 2 min. We then added 5 mL TE and 3 mL ACK 5 M to the suspension and centrifuged twice at 2000 × g for 15 min at 4 °C. After each centrifugation the supernatant was transferred to a new Falcon tube where we then added 12 mL of isopropanol to precipitate the genomic DNA. After a 5-min centrifugation at 500 × g the DNA pellet was washed with ice cold 70% ethanol and incubated for 5 min on ice. To precipitate the DNA pellet, we carried out a centrifugation of 5 min at 500 × g at 4 °C and then removed all ethanol from the tube. Finally, we added 250 µL of TE without disturbing the pellet and incubated overnight at room temperature before recovering the resuspended DNA without recovering any of the remaining pellet.

Some parental isolates were sequenced in-house using MinION R9.4.1 flowcells (10 isolates) whereas 14 isolates were sequenced by the Genoscope using PromethION R9.4.1 flowcells. All libraries were prepared with the Ligation Sequencing Kit SQK-LSK109. In order to sequence multiple isolates in the same run, the isolates sequenced with the MinION and PromethION were barcoded with the EXP-NBD104 and NBD114.96 Native Barcoding Kits respectively. Basecalling was carried out with guppy 5.0.16 (MinION data) and 5.0.11 (PromethION data).

Genome assembly and structural variant detection

Raw fastq files were processed using Porechop (v.0.2.4; [Prrwick/Porechop](#)) in order to remove adapters and barcodes. We further treated these reads using Filtlong (v.0.2.1; [Prrwick/Filtlong](#)) to remove reads shorter than 1000 nucleotides. For each strain, reads were then assembled using both Canu (v. 2.2) ([Koren et al., 2017](#)) and SMARTdenovo ([Liu et al., 2021](#)). Canu assemblies were generated using the options `-nanopore-raw` and `genomeSize=12m` while we used the default options for SMARTdenovo. We used the same fastq reads to polish both assemblies per strain using Medaka (v. 1.7.2; [nanoporetech/medaka](#)).

We then set out to detect SVs using MUM&Co (v 3.8) ([O'Donnell & Fischer, 2020](#)) separately on each genome assembly, using the option `-g 12000000`. Unique identifiers were also added to each SV using a combination of bcftools ([Danecek et al., 2021](#)) and the R package vcfr (v1.12) ([Knaus & Grünwald, 2017](#)). We then used Jasmine ([Alonge et al., 2020](#); [Kirsche et al., 2023](#)) to collapse SV calls during two rounds. First, for each strain, SV calls from both Canu and SMARTdenovo assemblies were collapsed into 24 strains specific VCF files. During the second round of merging, the 24 VCF files and a 25th containing direct SV calls from the publicly available Σ 1278b genome ([Dowell et al., 2010](#)) were collapsed into one final VCF file for downstream analyses. Then, since the 26th parent is a haploid version of the reference strain we assigned all the SVs as absent in that parent. Finally, SVs linked to transposable elements were detected using BLAST ([Johnson et al., 2008](#)) and a database containing LTR and Ty sequences ([Bleykasten-Grosshans et al., 2021](#)). Any SV covered by Ty-related elements on more than 50% of its length was flagged as Ty-related.

Generation of the genotype matrix

We recovered the genotypes for all the hybrids in our diallel panel from Fournier et al., 2019 where the parental genotype had been combined to infer the genotypes of the hybrids (Fournier et al., 2019). We retained biallelic variants using `vcftools` with the option `--min-alleles 2` and excluded singletons with the `vcftools --singleton` and `--exclude` commands, resulting in a final matrix of 31,818 SNPs. The genotype matrix was then recoded to a bed file with the `recode12` function of PLINK (Chang et al., 2015). The SVs were detected using MUM&co were subjected to the same filtering and recoding process, in order to be used for GWA analyses. The gene expression phenotypes, measured as transcripts per million, of 6186 genes were recovered from (Tsouris et al., 2024). The z-scores of each phenotype were then calculated and later used in GWAS.

Genome-wide heritability estimation

The genotype matrices used to estimate the genome-wide heritability were generated from the genotype data of the 1011 yeast genomes project (for the SNP matrix) and from the SVs identified using long-reads sequencing (for the SV matrix). To decrease linkage between the individuals we removed the variants present in only one parent. Furthermore, the variants with strong linkage disequilibrium ($r^2 > 0.8$) were removed for the calculation of the genome-wide heritability estimation. We calculated the weights of each variant using `ldak` with the `--cut-weights` and `--calc-weights-all` arguments and the default parameters (Zhang et al., 2021). All variants with non-zero weights were used to generate a filtered `vcf` matrix of 5493 SNPs that was then recoded with the `plink --make-bed` command. The filtered and recoded matrix was used to calculate the kinship between the individuals using the `popkin` function of the R package `popkin` with the default parameters. To estimate genome-wide heritability (h^2_g), from the kinship matrix mentioned above we used the `hglm` R package (Rönnegård et al., 2010) using the default parameters (Tsouris et al., 2024).

Genome-wide association

Genome-wide association analyses on transcript abundance z-scores were performed with the `single_snp` function of the `FaST-LMM` python package (Lippert et al., 2011; Widmer et al., 2014). We calculated condition-specific p-values by permuting the phenotypic values of the individuals 100 times and setting average 5% quantile (5% lowest p-values) as the threshold. This method doesn't provide us with a false discovery rate (FDR). Independent GWA analyses were carried out for SNPs and SVs but the p-value thresholds for the GWA on SNPs and on SVs were very different due to the important difference between the number of SNP and SV used. To correct for this discrepancy, we adjusted the condition-specific p-values of each GWA by normalizing to the total number of variants (SNPs and SVs).

The GWA eQTL displayed some genetic linkage associated patterns/signatures driving groups of linked variants to pass the significance threshold. We calculated genetic

linkage between variants using the plink `--r` command (Chang et al., 2015) and for each linkage group with an R value over 0.8 we only retained the variant with the lowest p-value. This variant filtering steps were independently run for the GWASs using the SNP and SV.

To distinguish between local and distant eQTLs we used a threshold of 25 kb, where the variants withing 25 kb of the start site of a gene were considered as local whereas those further away were considered as distant. We tested the differences in effect sizes or absolute variant weight between groups of eQTLs with Wilcoxon-Mann-Whitney tests using `wilcox.test` R function with the default parameters.

Analysis of low frequency variants

The MAF of the SNPs in the hybrids and the natural population of 1011 isolates was calculated with the `vcfR` using the `maf` command (Knaus & Grünwald, 2017) using the genotype matrix that was used for GWAS and the genotype matrix of the 1011 yeast genomes project respectively. The fold enrichment values of eQTLs from low frequency and common variants were carried out manually in R and tested using the `fisher.test` command.

Supplementary figures

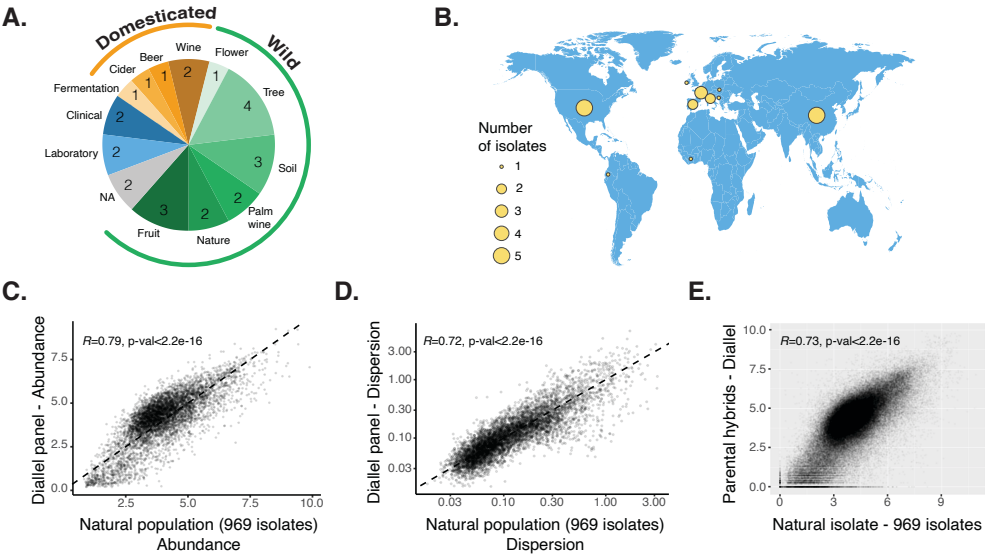


Figure S1 | Selection of the parental isolates and correlation between the transcript abundance in the diallel and the natural population.

(A., B.) Ecological and geographical origin of the parental isolates that were used to generate the diallel crossing panel. Pearson's correlation of the average transcript abundance (C.) and dispersion (D.) for each gene between the diallel population and the population of 969 natural isolates. (E.) Pearson's correlation of the transcript abundance of each homozygous hybrid and the respective homozygous diploid natural isolate.

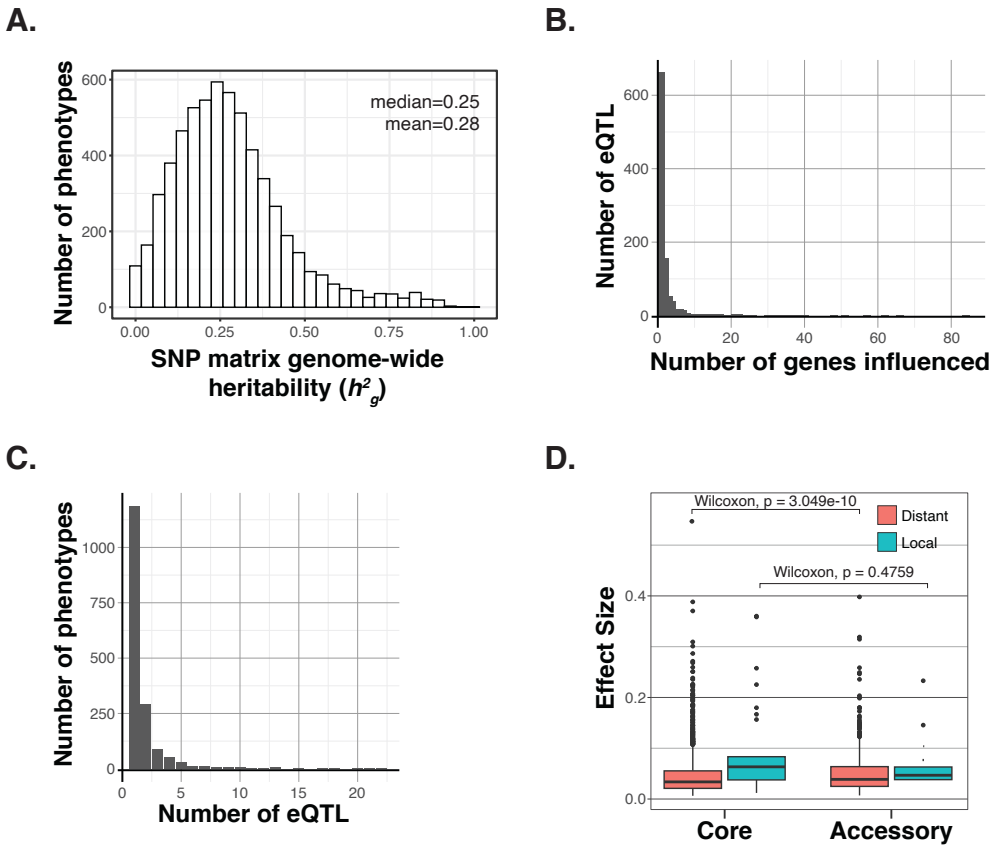


Figure S2 | Genome-wide heritability and GWAS results using the SNP matrix.

(A.) Distribution of genome-wide heritability (h^2_g) of all transcript abundance phenotypes calculated based on the kinship matrix of the hybrids. (B.) Distribution of the number of genes influenced by each SNP-eQTL. (C.) Distribution of the number of eQTLs associated to each gene. (D.) Effect size of eQTLs based on their position, distant or local in red and cyan, and the type of gene that is influenced (core or accessory gene). Significant difference is observed only between the distant eQTLs impacting core and accessory genes (two-sided Mann-Whitney-Wilcoxon test, p -value = 3.049×10^{-10}). Quantiles for local eQTL effect sizes for core genes ($n = 31$): min = 0.012, 1st quartile = 0.040, median = 0.056, 3rd quartile = 0.081, max = 0.360. Quantiles for local eQTL effect sizes for accessory genes ($n = 5$): min = 0.034, 1st quartile = 0.035, median = 0.105, 3rd quartile = 0.192, max = 0.257. Quantiles for distant eQTL effect sizes for core genes ($n = 2,555$): min = 0.006, 1st quartile = 0.021, median = 0.034, 3rd quartile = 0.056, max = 0.547. Quantiles for distant eQTL effect sizes for accessory genes ($n = 397$): min = 0.009, 1st quartile = 0.028, median = 0.044, 3rd quartile = 0.068, max = 0.319.

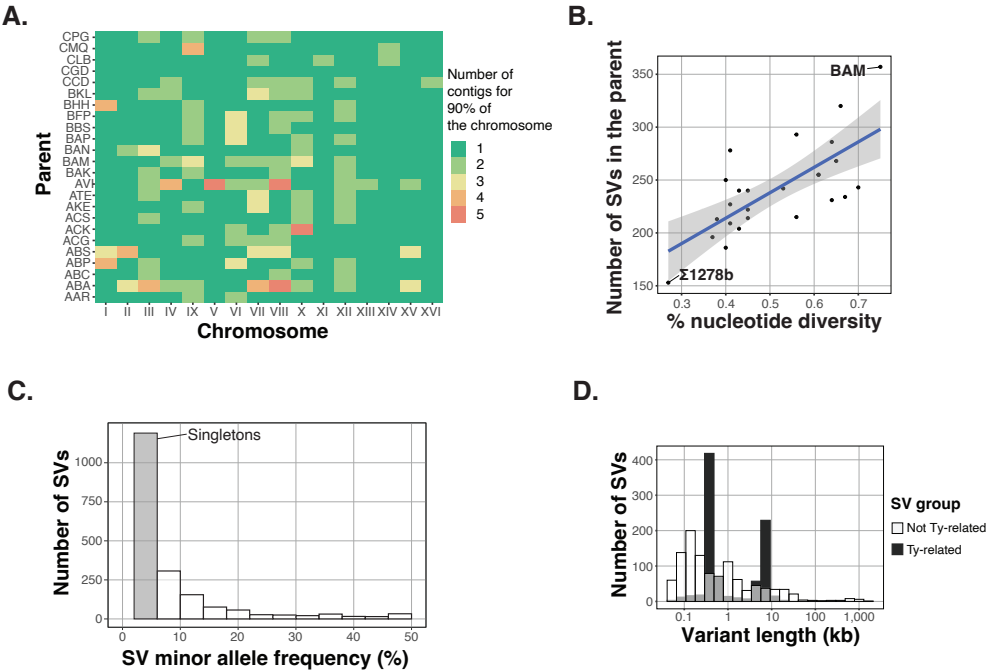
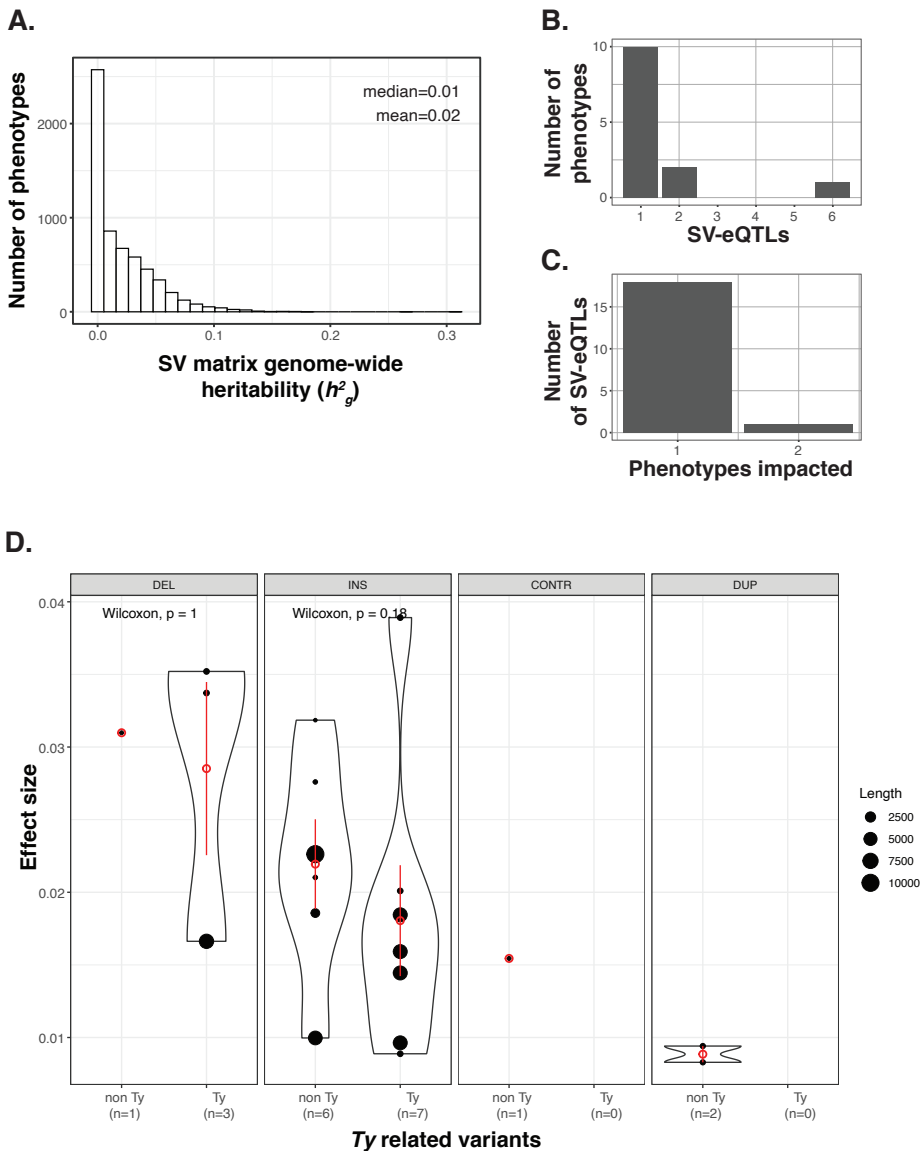


Figure S3 | Genome assemblies and structural variant detection statistics.

(A.) Number of contigs, obtained by the genome assemblies of the parents, needed to cover more than 90% of each chromosome's length. (B.) Pearson's correlation between the number of SVs detected in each parent and the parents nucleotide diversity from the reference strain (S288c) (Pearson's correlation $R = 0.7$, $p\text{-value} = 7.3e^{-5}$). The blue line represents the linear regression line between the two variables. (C.) Distribution of the SVs' MAF across the 26 parental isolates. The grey bar highlights the SVs present in only one individuals (singletons). (D.) Comparison of the variant length distributions of the Ty-related SVs (black bars) and not Ty-related SVs (white bars).



1

Figure S4 | SV genome-wide heritability and GWAS results.

(A.) Distribution of genome-wide heritability (h_g^2) based on the SV genotype matrix, excluding the SVs present in only one of the parents. (B.) Distribution of the number of eQTLs associated to each phenotype. (C.) Distribution of the number of phenotypes associated to each eQTL. (D.) Effect sizes for Ty and non-Ty related SVs across variant types. Error bars correspond to mean standard error from bootstrap.

Acknowledgements

This work was supported by a National Institutes of Health (NIH) grant R01 (GM147040-01), a European Research Council (ERC) Consolidator grant (772505) to JS and a French National Research Agency (ANR) young investigator grant (ANR-22-CE12-0023-01) to JH. It is also part of Interdisciplinary Thematic Institutes (ITI) Integrative Molecular and Cellular Biology (IMCBio), as part of the ITI 2021-to-2028 program of the University of Strasbourg, CNRS, and Inserm, supported by IdEx Unistra (ANR-10-IDEX-0002). JS is a Fellow of the University of Strasbourg Institute for Advanced Study (USIAS) and a member of the Institut Universitaire de France.

Author contributions

Andreas Tsouris: Conceptualization; Formal analysis; Investigation; Methodology; Writing – original draft. Gauthier Brach: Formal analysis; Investigation; Methodology. Anne Friedrich: Formal analysis; Methodology. Jing Hou: Conceptualization; Formal analysis; Funding acquisition; Investigation; Methodology; Writing – review and editing. Joseph Schacherer: Conceptualization; Supervision; Funding acquisition; Investigation; Methodology; Writing – original draft; Writing – review and editing.

Data availability

All Oxford Nanopore sequencing reads are available in the European Nucleotide Archive (ENA) under the accession number [PRJEB64478](#).

The 1002 Yeast Genome website http://1002genomes.u-strasbg.fr/files/Diallel_RNAseq provides access to:

- [Datafile 1](#): MAF of the SNPs across the diallel population and the population of 969 natural isolates.
- [Datafile 2](#): Functional annotation of all SNP variants.

References

- Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. (2017) Genetic effects on gene expression across human tissues. *Nature* **550**: 204–213.
- Akiyama M, Ishigaki K, Sakaue S, Momozawa Y, Horikoshi M, Hirata M, et al. (2019) Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nat Commun* **10**.
- Albert FW, Bloom JS, Siegel J, Day L, Kruglyak L (2018) Genetics of trans-regulatory variation in gene expression. *eLife* **7**: e35471.
- Albert FW, Kruglyak L (2015) The role of regulatory variation in complex traits and disease. *Nat Rev Genet* **16**: 197–212.
- Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. (2020) Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**: 145–161.e23.
- Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. (2014) Characterizing the Genetic Basis of Transcriptome Diversity through RNA-sequencing of 922 Individuals. *Genome Res*.
- Bleykasten-Grosshans C, Fabrizio R, Friedrich A, Schacherer J (2021) Species-Wide Transposable Element Repertoires Retrace the Evolutionary History of the *Saccharomyces cerevisiae* Host. *Mol Biol Evol* **38**: 4334–4345.
- Bloom JS, Boockvar J, Treusch S, Sadhu MJ, Day L, Oates-Barker H, et al. (2019) Rare variants contribute disproportionately to quantitative trait variation in yeast. *eLife* **8**: e49212.
- Caudal É, Loegler V, Dutreux F, Vakirlis N, Teyssonnière É, Caradec C, et al. (2024) Pan-transcriptome reveals a large accessory genome contribution to gene expression variation in yeast. *Nat Genet* **56**: 1278–1287.
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**.
- Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**: 80–92.
- Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* **10**: 392–404.
- Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. (2021) Twelve years of SAMtools and BCFtools. *GigaScience* **10**: giab008.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. (2013) STAR: Ultrafast Universal RNA-seq Aligner. *Bioinformatics* **29**: 15–21.
- Dowell RD, Ryan O, Jansen A, Cheung D, Agarwala S, Danford T, et al. (2010) Genotype to Phenotype: A Complex Problem. *Science* **328**: 469–469.
- Ferraro NM, Strober BJ, Einson J, Abell NS, Aguet F, Barbeira AN, et al. (2020) Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science*.
- Fournier T, Abou Saada O, Hou J, Peter J, Caudal E, Schacherer J (2019) Extensive impact of low-frequency variants on the phenotypic landscape at population-scale. *eLife* **8**.
- Gibson G (2012) Rare and common variants: Twenty arguments. *Nat Rev Genet* **13**: 135–145.

- Goffeau A, Barrell B. G., Bussey H., Davis R. W., Dujon B., Feldmann H., et al. (1996) Life with 6000 Genes. *Science* **274**: 546–567.
- GTEX Consortium (2017) The impact of structural variation on human gene expression. *Nat Genet* **49**: 692–699.
- Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, et al. (2024) Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat Biotechnol* **42**: 663–673.
- Hill MS, Vande Zande P, Wittkopp PJ (2021) Molecular and evolutionary processes generating variation in gene expression. *Nat Rev Genet* **22**: 203–215.
- Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* **106**: 9362–9367.
- Huan T, Liu C, Joehanes R, Zhang X, Chen BH, Johnson AD, et al. (2015) A systematic heritability analysis of the human whole blood transcriptome. *Hum Genet* **134**: 343–358.
- Johnson M, Zaretskaya I, Raytselis Y, Merezuk Y, McGinnis S, Madden TL (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res* **36**:W5–W9.
- Kawakatsu T, Huang S-SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, et al. (2016) Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. *Cell* **166**: 492–505.
- Kirsche M, Prabhu G, Sherman R, Ni B, Aganezov S, Schatz MC (2021) Jasmine: Population-scale Structural Variant Comparison and Analysis. *bioRxiv*: 2021.05.27.445886.
- Kirsche M, Prabhu G, Sherman R, Ni B, Battle A, Aganezov S, et al. (2023) Jasmine and Iris: Population-scale structural variant comparison and analysis. *Nat Methods* **20**: 408–417.
- Kita R, Venkataram S, Zhou Y, Fraser HB (2017) High-resolution mapping of cis-regulatory variation in budding yeast. *Proc Natl Acad Sci* **114**: E10736–E10744.
- Knaus BJ, Grünwald NJ (2017) Vcfr: A package to manipulate and visualize variant call format data in R. *Mol Ecol Resour* **17**: 44–53.
- Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM (2017) Canu: Scalable and Accurate Long-Read Assembly via Adaptive k-Mer Weighting and Repeat Separation. *Genome Res*: gr.215087.116.
- Li N, He Q, Wang J, Wang B, Zhao J, Huang S, et al. (2023) Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nat Genet* **55**: 852–860.
- Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. (2023) A draft human pangenome reference. *Nature* **617**: 312–324.
- Liao Y, Smyth GK, Shi W (2014) featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinforma Oxf Engl*.
- Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011) FaST linear mixed models for genome-wide association studies. *Nat Methods* **8**: 833–835.
- Liu H, Wu S, Li A, Ruan J, Wu S, Li A, et al. (2021) SMARTdenovo: A de novo assembler using long noisy reads. *Gigabyte* **2021**: 1–9.
- Mackay TFC (2014) Epistasis and quantitative traits: Using model organisms to study gene-gene interactions. *Nat Rev Genet* **15**: 22–33.
- Mackay TFC, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: Challenges and prospects. *Nat Rev Genet* **10**: 565–577.
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. (2009) Finding the missing heritability of complex diseases. *Nature* **461**: 747–753.

- Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, Kjaer TR, et al. (2017) Rare and low-frequency coding variants alter human adult height. *Nature* **542**: 186–190.
- Muenzner J, Trébulle P, Agostini F, Zaubner H, Messner CB, Steger M, et al. (2024) Natural proteome diversity links aneuploidy tolerance to protein turnover. *Nature* **630**: 149–157.
- O'Donnell S, Fischer G (2020) MUM&Co: Accurate detection of all SV types through whole-genome alignment. *Bioinformatics* **36**: 3242–3243.
- O'Donnell S, Yue J-X, Saada OA, Agier N, Caradec C, Cokelaer T, et al. (2023) Telomere-to-telomere assemblies of 142 strains characterize the genome structural landscape in *Saccharomyces cerevisiae*. *Nat Genet* **55**: 1390–1399.
- Ouwens KG, Jansen R, Nivard MG, Dongen J, Frieser MJ, Hottenga J-J, et al. (2020) A Characterization of Cis- and Trans-Heritability of RNA-Seq-based Gene Expression. *Eur J Hum Genet EJHG* **28**: 253–263.
- Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, et al. (2018) Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **556**: 339–344.
- Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* **69**: 124–137.
- Rockman MV, Kruglyak L (2006) Genetics of global gene expression. *Nat Rev Genet* **7**: 862–872.
- Rockman MV, Skrovanek SS, Kruglyak L (2010) Selection at Linked Sites Shapes Heritable Phenotypic Variation in *C. elegans*. *Science* **330**: 372–376.
- Rönnegård L, Shen X, Alam M (2010) Hglm: A package for fitting hierarchical generalized linear models. *R J* **2**: 20–28.
- Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, et al. (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**: 297–302.
- Scott AJ, Chiang C, Hall IM (2021) Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Res* **31**: 2249–2257.
- Teyssonnière EM, Trébulle P, Muenzner J, Loegler V, Ludwig D, Amari F, et al. (2024) Species-wide quantitative transcriptomes and proteomes reveal distinct genetic control of gene expression variation in yeast. *Proc Natl Acad Sci* **121**: e2319211121.
- Tsouris A, Brach G, Schacherer J, Hou J (2024) Non-additive genetic components contribute significantly to population-wide gene expression variation. *Cell Genomics* **4**.
- Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**: 5–22.
- Vu V, Verster AJ, Schertzberg M, Chuluunbaatar T, Spensley M, Pajkic D, et al. (2015) Natural variation in gene expression modulates the severity of mutant phenotypes. *Cell* **162**: 391–402.
- Wainshtein P, Jain D, Zheng Z, Cupples LA, Shadyab AH, McKnight B, et al. (2022) Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat Genet* **54**: 263–273.
- Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, et al. (2015) The UK10K project identifies rare variants in health and disease. *Nature*.
- Weischenfeldt J, Symmons O, Spitz F, Korb J (2013) Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nat Rev Genet* **14**: 125–138.
- West MAL, Kim K, Kliebenstein DJ, Leeuwen H, Michelmore RW, Doerge RW, et al. (2007) Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. *Genetics*.
- Widmer C, Lippert C, Weissbrod O, Fusi N, Kadie C, Davidson R, et al. (2014) Further improvements to linear mixed models for genome-wide association studies. *Sci Rep* **4**.

- Zhang G, Roberto NM, Lee D, Hahnel SR, Andersen EC (2022) The impact of species-wide gene expression variation on *Caenorhabditis elegans* complex traits. *Nat Commun* **13**: 3462.
- Zhang Q, Privé F, Vilhjálmsson B, Speed D (2021) Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat Commun* **12**.
- Zhang X, Cal AJ, Borevitz JO (2011) Genetic architecture of regulatory variation in *Arabidopsis thaliana*. *Genome Res* **21**: 725–733.
- Zuk O, Hechter E, Sunyaev L SR, E.S. (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci USA* **109**: 1193–1198.

Natural variation in alternative polyadenylation across *Saccharomyces cerevisiae* isolates

Gauthier Brach¹, Christos Papadopoulos², José Carlos Montañés², M. Mar Albà^{2,3}, Joseph Schacherer^{1,4,5,6}

¹ Université de Strasbourg, CNRS, Inserm, IGBMC UMR 7104 – UMR-S 1258, 67400 Illkirch, France

² Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Research Institute, Barcelona, Spain

³ Catalan Institute for Research and Advanced Studies (ICREA), Barcelona, Spain

⁴ Institut Universitaire de France (IUF), 75231 Paris, France

⁵ CIC bioGUNE, Bizkaia Technological Park, Building 800, 48160 Derio, Spain

⁶ Ikerbasque, Basque Foundation for Science, Bilbao, Spain

| Abstract

Understanding how genetic variation shapes post-transcriptional regulation is essential for linking genotype to phenotype. Here, we used Oxford Nanopore direct RNA sequencing and strain-specific genome assemblies to characterize mRNA 3' end regulation across six genetically diverse natural isolates of *Saccharomyces cerevisiae*. Poly(A) tail length was strongly gene-specific and largely conserved between isolates, and downstream-region sites proximal to the annotated stop codon and distal sites farther away in transcript orientation showed no common directional difference in tail length. The approximate position of transcript-end regions was also generally conserved, while matched end-region usage showed quantitative, strain-associated differences at 921 genes under predefined statistical and effect-size criteria. Site detection and apparent strain specificity remained sensitive to read depth and matching resolution. Direct RNA sequencing also identified a limited set of unannotated transcript models, including 191 high-quality calls across 75 canonical host genes. Together, these data show broad conservation of gene-level poly(A)-tail properties alongside quantitative variation in transcript-end usage, without establishing the causal genetic basis of that variation.

Introduction

Although genome-wide genetic studies have identified thousands of loci associated with molecular and organismal traits, the mechanisms linking DNA variation to phenotype remain incompletely understood. A major reason is that transcript abundance provides only a partial description of gene regulation, as numerous post-transcriptional processes shape mRNA fate before translation. Understanding natural variation in these processes is therefore essential for explaining how genetic diversity is translated into phenotypic diversity.

Among post-transcriptional processes, mRNA 3' end formation occupies a central position. Cleavage and polyadenylation determine both the position of the transcript end and the length of the poly(A) tail, two features that influence mRNA stability, translation, nuclear export and subcellular localisation (Proudfoot, 2011; Mayr, 2016; Tian & Manley, 2017). Poly(A) tail shortening by the Ccr4-Not and Pan2-Pan3 deadenylase complexes is the first and generally rate-limiting step of the major mRNA decay pathway (Decker & Parker, 1993; Tucker et al., 2001; Wahle & Winkler, 2013; Webster et al., 2018). In parallel, alternative polyadenylation (APA) generates transcript isoforms with distinct 3' untranslated regions through the selection of different cleavage sites (Di Giammartino et al., 2011; Elkon et al., 2013). APA is pervasive across eukaryotes, dynamically regulated during development and stress responses, and its dysregulation has been implicated in cancer, neurological disorders and immune diseases (Mayr & Bartel, 2009; Gruber & Zavolan, 2019; Mitschka & Mayr, 2022). In human populations, genetic variants affecting polyadenylation site usage explain a substantial fraction of complex-trait associations independently of eQTL, highlighting APA as an important regulatory layer (Li et al., 2021).

The yeast *Saccharomyces cerevisiae* provides an ideal model to investigate natural variation in 3' end regulation. Population genomic studies have characterized the extensive genetic diversity of this species (Peter et al., 2018; Loegler et al., 2024), while large-scale transcriptomic and proteomic analyses have revealed widespread regulatory variation across natural isolates (Albert et al., 2018; Caudal et al., 2024; Muenzner et al., 2024; Tsouris et al., 2024a, b). However, comparisons between transcript and protein QTL reveal remarkably limited overlap, and ribosome profiling has shown that nearly half of transcriptional differences are buffered before translation (McManus et al., 2014; Blevins et al., 2019; Teyssonnière et al., 2024a, b). These observations suggest that post-transcriptional regulation represents a major but still poorly characterized component of natural molecular variation.

Genome-wide analyses in the laboratory strain S288C have shown that yeast transcripts exhibit gene-specific poly(A) tail lengths and widespread alternative polyadenylation, with many genes producing multiple transcript isoforms differing in their 3' ends (Ozsolak et al., 2010; Moqtaderi et al., 2013; Pelechano et al., 2013; Geisberg et al., 2014, 2023; Tudek et al., 2021). The choice of polyadenylation site influences transcript stability and RNA-binding protein interactions, demonstrating important functional consequences of 3' end selection. However, all current knowledge derives almost exclusively from the S288C laboratory background. Whether poly(A) tail length,

transcript end usage and APA are conserved across natural populations remains essentially unknown.

Recent advances now make this question accessible. Near telomere-to-telomere genome assemblies for more than one thousand *S. cerevisiae* isolates enable accurate strain-specific transcript mapping without reference bias (O'Donnell et al., 2023; Loegler et al., 2025). Combined with Oxford Nanopore direct RNA sequencing (dRNA-seq), which measures native RNA molecules and provides molecule-resolved 3' end positions and poly(A)-tail estimates (Garalde et al., 2018; Wang et al., 2018; Workman et al., 2019; Tudek et al., 2021; Esfahani et al., 2025; Kim et al., 2025), these resources provide an opportunity to characterize post-transcriptional regulation across natural isolates.

Here, we use Oxford Nanopore direct RNA sequencing to compare mRNA 3' end regulation across six *S. cerevisiae* natural isolates using strain-specific genome assemblies. We quantify gene expression, transcript-end usage, poly(A) tail length and transcript isoform diversity at single-molecule resolution. This design tests which properties are conserved across isolates and which show quantitative, strain-associated differences, while explicitly accounting for the strong effects of sequencing depth and operational site definitions.

Results

Direct RNA sequencing generates a high-quality transcriptomic atlas of six natural isolates

We performed Oxford Nanopore direct RNA sequencing (dRNAseq) on six *S. cerevisiae* natural isolates selected from distinct clades: BBQ (Mediterranean oak), BED (French Guiana human), BGQ (French dairy), BLR (Wine/European), CCE (Ecuadorean) and CLI (Sake) (Fig. 1A; Supplementary Table S1) (Peter et al., 2018; Loegler et al., 2024). Because dRNAseq measures native RNA molecules directly, the approach avoids PCR amplification and cDNA-synthesis artefacts, preserves strand information and retains poly(A) tail information. Reads were first error-corrected with FMLRC2 (Wang et al., 2018; Mak et al., 2023) using strain-matched short-read data (Caudal et al., 2024), further refined with TranscriptClean (Wyman & Mortazavi, 2019), and aligned with minimap2 in splice-aware mode. For expression quantification and raw-versus-corrected comparisons, reads were aligned to the common S288C R64-3-1 reference; 3'-end analyses used the corresponding strain-specific assembly. After filtering reads outside the 75-16,000 nt range, we retained an average of 14 million corrected reads per strain (Fig. S1A; Supplementary Data 1). Read-length distributions were highly similar among isolates, with most reads falling between 500 and 3,000 nt (Fig. S1B).

Most reads contained a detectable poly(A) tail, and the depth-weighted pooled tail-length distributions had medians of 27–33 nt (Fig. 1B). BBQ had the lowest pooled median (27 nt); its original and replicate libraries had medians of 31 and 27 nt, respectively, and equal weighting of the libraries gave 28 nt (Supplementary Table S3). Library imbalance therefore contributes to, but does not fully explain, the shorter pooled BBQ distribution. These estimates are consistent with the 30–50 nt steady-state tails previously reported in yeast (Subtelny et al., 2014; Tudek et al., 2021) and with the approximately 35 nt median obtained by dRNAseq in the S288C background (Workman et al., 2019). Biological replicates were also highly concordant at the gene-count level, with Pearson correlations ranging from 0.928 for BBQ to 0.980 for CCE (Fig. S1C), supporting the reproducibility of the sequencing and processing procedures.

We next evaluated whether read correction altered transcript quantification. Mapping rates increased from 94.5-97.6% for raw reads to 99.6-99.9% after FMLRC2 and TranscriptClean correction (Fig. S1D). Despite this gain in alignment, gene-level counts were almost unchanged: raw and corrected datasets showed Pearson correlations above 0.98 in all strains (Fig. S1E). The magnitude of correction was also evenly distributed across expression levels, with no evidence that highly or weakly expressed genes were preferentially affected. Thus, the correction pipeline improved read mapping without a detectable systematic shift in gene-level expression estimates.

Gene expression is broadly conserved but retains strain-specific signatures

To establish the transcriptional background against which 3' end variation could be interpreted, we quantified gene expression from corrected, merged reads aligned to the common S288C R64-3-1 reference. After counts-per-million (CPM) normalization, pairwise correlations of log₂-transformed expression values were uniformly high, ranging from 0.908 for BBQ-BED to 0.962 for BLR-CCE (Fig. S1F). The number of genes supported by at least five reads was also comparable among isolates (Fig. S1G), and UpSet analysis showed that the overwhelming majority of expressed genes were shared by all six strains, with only a limited set detected in a strain-specific manner (Fig. 1C).

We therefore asked which genes differed most in expression across isolates. Ranking the 5,051 genes above 10 CPM by their coefficient of variation, the most variable decile (506 genes; Supplementary Table S2) was enriched for carbon and energy metabolism and for vitamin biosynthesis, including carbohydrate metabolism (2.9-fold, $\text{padj} = 8.4 \times 10^{-11}$), primary alcohol metabolism (7.3-fold, 9.2×10^{-12}), thiamine biosynthesis (9.1-fold, 4.2×10^{-9}) and generation of precursor metabolites and energy (2.7-fold, 1.5×10^{-8} ; Fig. 1D). Together, these results indicate that natural isolates share a common expression landscape while maintaining modest and reproducible strain-specific transcriptional signatures (Caudal et al., 2024).

Poly(A) tail length is gene-specific and largely conserved across strains

We next examined the poly(A) tail landscape independently of transcript-end choice. Although global tail-length distributions were similar among strains (Fig. 1B), median tail length varied substantially among genes. A marginal two-factor sums-of-squares summary assigned 7.1% of the total poly(A)-tail sum of squares to gene identity and 1.0% to strain. The remaining 91.8% is a residual category containing within-gene and within-strain molecule-to-molecule variation, gene-by-strain structure, measurement variation and other unmodelled effects; it is not an independently attributable third biological factor (Fig. 2A).

Across all six strains, gene expression was inversely related to median poly(A) tail length: highly expressed genes tended to carry shorter tails (Fig. 2B). Gene-level median tail lengths were also correlated across all strain pairs (Spearman $\rho = 0.66$ – 0.82 ; all $P < 2.2 \times 10^{-16}$; Fig. 2C, Fig. S2A), while the fraction of tail-length variation explained by strain was low for most individual genes (Fig. S2B). Fig. S2A displays unweighted gene-level medians, centred at 33 nt in BBQ and 39–43 nt in the other strains, whereas Fig. 1B pools and therefore weights individual reads; the inverse expression–tail relationship makes the pooled distribution more strongly influenced by abundant, shorter-tailed genes. Together, these results indicate that gene identity is a major determinant of steady-state poly(A) length and that the expression–tail relationship is a general feature of the transcriptome rather than a strain-specific effect. This pattern is consistent with the poly(A) paradox described in other systems, in which stable and efficiently translated mRNAs undergo rapid co-translational deadenylation and reach a shorter steady-state

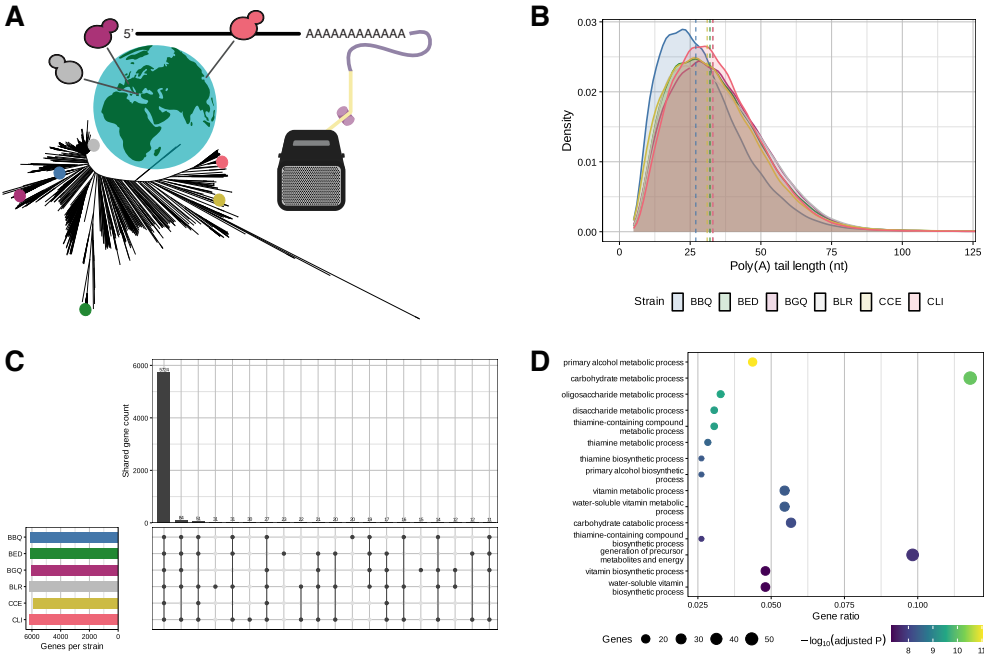


Figure 1 | Direct RNA-sequencing overview and the shared expression landscape.

(A.) Six *Saccharomyces cerevisiae* natural isolates analyzed were chosen to span the diversity of the species. (B.) Global distributions of per-read poly(A)-tail length, restricted to 0–120 nt. (C.) UpSet plot of expressed genes, defined as genes supported by at least five reads, showing per-strain totals and the 20 largest intersections. (D.) Gene Ontology enrichment among the 506 genes in the most variable expression decile of the 5,051 genes above 10 CPM.

tail length (Subtelny et al., 2014; Passmore & Collier, 2022). A comparable inverse relationship has also been reported from fission-yeast dRNAseq data (Montañés et al., 2022).

Tail-length heterogeneity within genes was extensive. The coefficient of variation of read-level tail lengths differed markedly among genes (Fig. S2C), implying that some transcripts maintain a narrow tail-length distribution whereas others occupy a much broader range. To distinguish unusually broad or narrow distributions from the expected mean-variance relationship, we fitted a loess regression and analysed the residuals, which showed no remaining dependence on expression level (Fig. S2D). Genes with large positive residuals had broader distributions than their expression-matched peers whereas genes with strongly negative residuals were unusually sharp (Fig. S2E). Neither the top 10% of genes with the longest tails nor genes with the broadest distributions showed significant functional enrichment (Fig. S2F).

We next asked whether the broadest distributions were bimodal. Hartigan’s dip test applied to all available reads called 4,392 of 7,880 genes (55.7%) bimodal in at least three strains, but this reflected sequencing depth rather than tail-length distribution: above approximately 2,000 reads per gene per strain every gene-strain pair was called bimodal (Fig. S2G), while the dip statistic itself decreased with depth (Fig. S2H).

Subsampling every pair to a common depth of 100 reads left no gene bimodal in three or more strains, and on a fixed set of pairs the call rate rose from 0% to 99.6% as the imposed depth increased from 100 to 1,600 reads. We therefore treat tail-distribution shape as a graded rather than a binary property. At matched depth, the per-gene dip statistic remained reproducible between isolates (Spearman $\rho = 0.27$ - 0.46 across strain pairs; Fig. S2I), indicating that the breadth of a gene's tail distribution is gene-intrinsic and conserved even where no discrete two-state structure can be resolved.

By contrast, cross-strain differences in a gene's overall tail length were generally small. Although read-level Kruskal-Wallis tests identified significant strain effects for 5,405 of 5,505 genes ($\text{FDR} < 0.05$; Fig. 2D), the effects were modest: the median gene differed by 1.25 fold ($\log_2$ fold change 0.32, interquartile range 0.26 to 0.39) between its most extreme strains, or 8.6 nt, against median within-gene minimum and maximum strain means of 35.3 and 44.3 nt, and only 55 genes exceeded a $\log_2$ fold change of 0.75. As with the dip test, the near-universal significance reflects the very high number of reads per gene rather than large biological effects. Overall, poly(A) tail length is highly heterogeneous within genes but is largely conserved for a given gene across natural isolates.

A robust map of alternative polyadenylation across natural isolates

We then mapped transcript end sites (TESs) to define the alternative polyadenylation landscape. Our custom pipeline extracts strand-specific 3'-end positions from aligned reads, estimates local TES density by Gaussian kernel density estimation, and separates peaks by level-set island carving. Nearby cleavage events are thereby grouped into contiguous polyadenylation islands that represent functional transcript-end regions rather than individual nucleotide positions (Fig. 3A). Candidate islands were retained according to density relative to the within-gene maximum, valley separation, minimum peak height and density, and the fraction of the gene's reads they contained; the complete thresholds are given in Methods. All 3'-end and per-gene poly(A) analyses were run on read sets subsampled to a common depth of 6,840,171 primary alignments ($\text{MAPQ} \geq 20$) per strain, so that island counts and site usage are not confounded by differences in sequencing effort.

At equalized sequencing depth, we detected 8,067-8,852 polyadenylation islands per strain. Most expressed genes had one dominant island, whereas 23.7-26.7% had at least two supported islands and were therefore classified as APA (Alternative PolyAdenylation) genes (Fig. 3B). These values are lower than the rates reported by the two earlier genome-wide surveys of laboratory yeast, and the difference is largely methodological. Both used Helicos direct RNA sequencing, which records every individual cleavage position without amplification. Ozsolak et al. (2010) reported that 72.1% of yeast genes carried polyadenylation positions separated by at least 50 nt, and Moqtaderi et al. (2013), working in a BY4741 derivative, found a median of 62 distinct polyadenylated termini per protein-coding gene distributed over an 'end zone' of median 269 nt, within which the most-used site accounted for only about 20% of

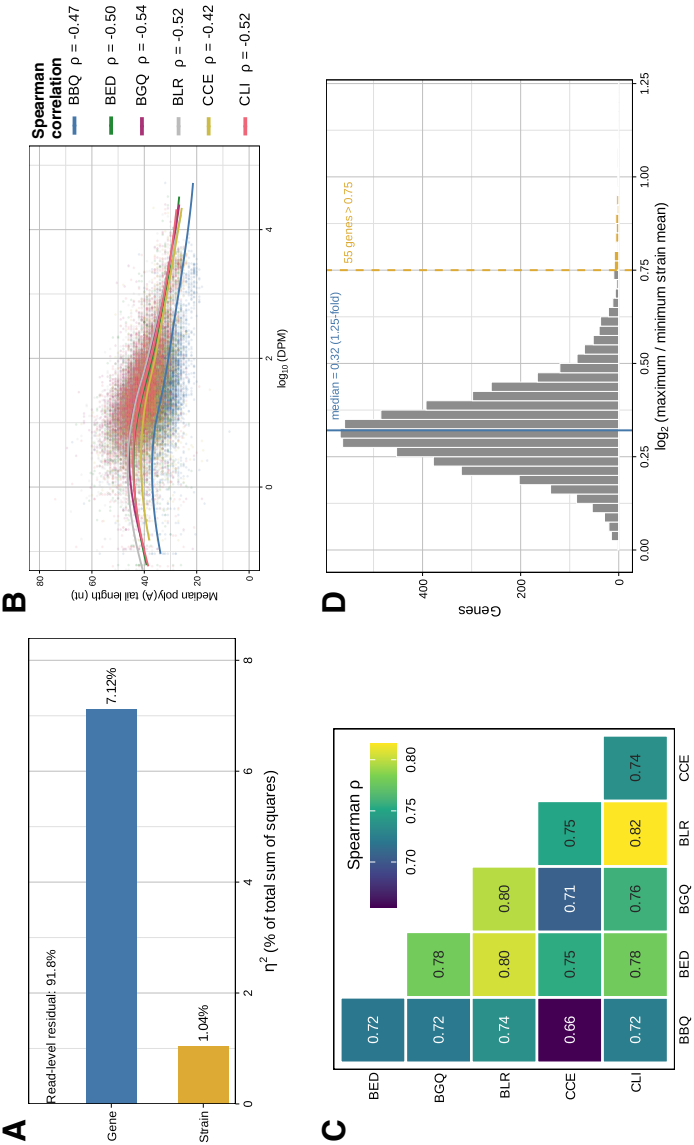


Figure 2 | Gene-level poly(A)-tail length is largely conserved across strains.

(A.) Marginal gene and strain components of the total read-level poly(A)-tail sum of squares. The residual includes gene-by-strain structure and read-level variation. (B.) Gene-level median poly(A)-tail length against expression for each strain, with loess fits and Spearman correlations. (C.) Pairwise Spearman correlations of gene-level median poly(A)-tail length across strains. (D.) Distribution of the $\log_2$ ratio between the largest and smallest strain mean for each gene supported by at least 20 reads in at least three strains, with 0.75 as the effect-size reference.

a gene's molecules. Our pipeline instead requires an island to be prominent and well supported before it is counted, so the minor positions in an end zone fall below our thresholds and never become islands. Island widths were narrow (medians of 19-22 bp; Fig. 3C) and the median island carried 55-60% of its gene's 3' ends, meaning we quantify the dominant core of a cleavage region rather than every position within it. Sequencing depth contributes as well, since APA rates rose to 23.9-38.9% when the full libraries were used.

Genes with two or more islands carry a small number of closely spaced transcript-end regions rather than a broad cleavage zone. For downstream-region comparisons in genes with exactly two islands, proximal denotes the site closer to the annotated stop codon and distal the site farther away in transcript orientation. Usage ratios varied widely between genes, with 55–59% of genes favouring one site by more than two-fold, but all six strains showed a modest population-level preference for the distal site (median log₂ proximal/distal ratio -0.11 to -0.35, Wilcoxon $P < 0.05$ in every strain; Fig. S3A). Consecutive islands were separated by a median of 27–31 bp (interquartile range 17–47 bp; Fig. S3B), close to the 20-bp merge and sub-peak separation thresholds used during island calling, so this spacing reflects the resolution limit of our approach as much as any biological constraint. The caller's window around the coverage-extended annotated end also captured 666 of 39,508 accepted islands (1.69%) with a summit upstream of the stop codon. However, it was not designed to survey 5' UTRs or the first 10% of coding regions systematically; upstream and early-CDS termination described by Moqtaderi et al. (2013) and Montañés et al. (2026) therefore remains outside the main scope of this analysis.

We validated the detected islands using three complementary approaches. First, 90.7–92.2% of deduplicated TES peaks contained at least one of 11 known yeast polyadenylation-related sequence patterns in the -90- to +40-nt window around the peak, compared with 60.9–61.8% of matched random genomic positions (Fisher's exact test, $P < 2.2 \times 10^{-16}$; approximately 30-percentage-point excess; Fig. 3D). The set comprised TA-rich efficiency-like patterns (TATATA, TATAAA and ATATAT), U/T-rich patterns (TTTTTT, TTTTAT, TTTTAA, TTCTTT and TCTTTC) and A-rich positioning-like patterns (AATAAA, AAAAAA and AAATAA). Here, deduplication removes identical genomic peak coordinates generated through overlapping gene or transcript models, preferentially retaining systematic ORF assignments. De novo 6-mer analysis independently identified TATATA, ATATAT and TTTTTT as the most enriched motifs (Z-scores > 120), while TATAAA and AATAAA were also significantly enriched. Motif enrichment was highly conserved across strains, with coefficients of variation below 0.08 for the most variable motifs (Fig. S3C). These observations are consistent with the degenerate and distributed nature of the yeast polyadenylation signal (Guo & Sherman, 1995; Graber et al., 1999; Liu et al., 2017; Stroup & Ji, 2024).

Second, we compared TES calls in the S288C reference background with the 91,891 sites mapped by Helicos direct RNA sequencing (Ozsolak et al., 2010). Of our 10,652 S288C TES islands, 80.0% were located within 25 bp of an Ozsolak site, 87.3% within 50 bp and 91.4% within 100 bp (Fig. S3D). In the reverse comparison, 35.4% of Ozsolak sites were within 25 bp and 48.9% within 50 bp of one of our islands,

as expected from the much greater density of nucleotide-level sites in the Oszolak catalogue. At the gene level, 109 genes were classified as APA in both datasets, whereas 1,873 were APA-specific to our data and 4,184 to the Oszolak dataset. This asymmetry reflects the different objectives of the two approaches: the Oszolak dataset captures numerous low-frequency cleavage positions, whereas our pipeline emphasises prominent, well-supported islands.

Third, we quantified the effect of sequencing depth. In the full-depth datasets APA rates rose from 23.7–26.7% to 23.9–38.9%, with the largest gains in the deepest libraries (BLR, 23.7% to 37.5%; BED, 24.6% to 38.0%) and none in the shallowest, CCE, whose full library sets the equalization target and is therefore the same read set in both analyses (23.9% in each; Fig. S3E). Across the other five strains, genes gaining an island outnumbered those losing one by 2.9- to 5.3-fold, and 66–82% of genes kept the same island count. The equalized-depth analysis therefore provides a conservative estimate of APA prevalence, while deeper sequencing primarily reveals additional minor sites without changing the core landscape.

APA architecture is only partially conserved across natural isolates

We next asked how strongly the APA landscape is conserved. The total number of APA genes was similar among strains (1,472–1,727), but the gene sets overlapped only partially: of 3,214 genes with APA in at least one strain, 546 were shared by all six and 1,049 were detected in a single strain (Fig. 4A; Supplementary Data 2). Detection breadth was associated with read support: the median of each gene's per-strain read counts rose stepwise from 141 reads for APA detected in one strain to 525 reads for APA detected in all six (Supplementary Table S3). The data therefore support a conserved core of alternatively polyadenylated genes together with a coverage-sensitive set of strain-associated calls; presence in one strain alone is not sufficient to establish a strain-restricted biological event.

The distinction between site presence and quantitative site usage was particularly important. We compared per-gene APA profiles across strains by calculating cosine similarity between 5-bp-binned individual 3'-end histograms indexed from each gene-strain's strand-specific window start; that window was anchored on the coverage-extended annotated transcript end. This fine-scale endpoint analysis retains the complete positional distribution without imposing discrete island boundaries, whereas islands provide robust units for site counting, isoform assignment and differential-usage testing. Among 5,601 genes represented in at least four strains with at least 10 reads, the median pairwise cosine similarity was 0.524 (Fig. 4B; per-strain-pair values and profile variability in Fig. S4A). Only 314 genes (5.6%) had highly conserved profiles with cosine similarity above 0.95, whereas 2,593 genes (46.3%) had low-similarity profiles below 0.5 (Fig. 4C). As a typical example, for the *YDR064W* gene (Fig. S4B), reads are divided between sites 73 and 102 bp past the stop codon in the same ratio in every strain, with 81 to 87% at the proximal site. At the *YGR267C* locus (Fig. S4C), covered by 620 to 956 reads in all six strains, all reads were assigned to the proximal matched class in BBQ and BLR and to the distal class

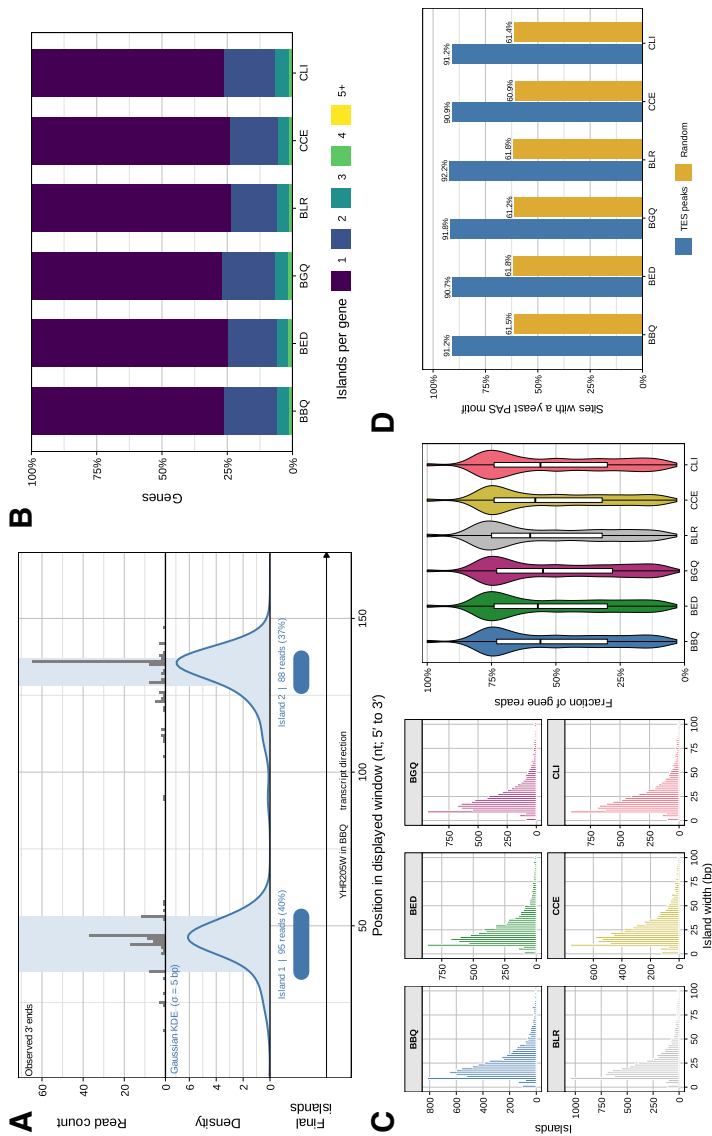


Figure 3 | Detection and validation of polyadenylation islands.

(A.) Island calling illustrated for *YHR205W* in BBQ. Observed 3'-end counts are smoothed by Gaussian kernel density estimation ($\sigma = 5$ bp), and retained islands are shown with their read support in transcript orientation. (B.) Fraction of expressed genes with one to five or more retained islands in each strain. Genes with at least two islands were classified as alternative-polyadenylation genes. (C.) Island-width distributions (left) and the fraction of a gene's reads assigned to each island (right). (D.) Fraction of deduplicated TES peaks and matched random positions with at least one known yeast polyadenylation-signal motif in the 100-nt upstream window.

in CCE, while CLI divided 28% versus 72% between them (Fig. 4D). The raw BBQ profile contained nearby ends at 70 and 91.5 bp, which were grouped into the same matched class. Most transcript-end regions are therefore shared between strains but used in different proportions.

We also assessed conservation of the overall 3' end position by calculating, for each gene, the median TES relative to the stop codon in each strain. Among 4,819 genes with sufficient coverage in at least four strains, the median cross-strain range was only 11.9 nt, although the most divergent genes shifted by as much as 319 nt (Fig. 4E). Applying predefined significance and effect-size criteria classified 727 genes as conserved ($q \geq 0.05$ and median range < 25 nt), 2,308 as intermediate and 1,784 as strain-divergent ($q < 0.05$ and median range ≥ 25 nt; Fig. S4D). The two classes are visibly distinct at individual loci: in conserved genes the six per-strain distributions overlap almost completely, while in divergent genes the median position differs by up to several hundred nucleotides between backgrounds (Fig. 4F).

3' UTR SNP load did not distinguish the classes, averaging 1.34 SNPs per 3' UTR in divergent genes and 1.28 in conserved genes (1,784 and 727 genes; Wilcoxon $P = 0.102$; Fig. S4E). Local SNP burden alone therefore did not explain TES position shifts, although this count-based comparison does not test the effects of particular variants or distinguish cis from trans contributions.

Together, these analyses suggest a hierarchy of conservation. The approximate position of a gene's polyadenylation region is usually stable across isolates, the presence of individual APA sites is only partially conserved, and the quantitative allocation of transcripts among those sites is more variable still.

APA generates substantial isoform diversity, including a limited set of novel transcripts

We next quantified the transcript isoforms generated by the detected TES islands. The island-derived transcriptome contained one or two transcript models for 85-87% of expressed genes, split between 32-35% with a single model and 50-54% with two (Fig. 5A). Among models receiving at least one assigned read in Salmon quantification (Patro et al., 2017), expressed transcript counts exceeded expressed gene counts by 78.9–83.1% (Fig. S5A). TIF-seq reported a median of 26 major transcript isoforms per protein-coding gene (Pelechano et al., 2013); our island-based approach yields fewer because it groups nearby 3' ends and does not resolve independent 5' end variation. Among multi-isoform genes, the dominant isoform accounted for 60–78% of gene expression across the interquartile range (Fig. S5B), although a subset of genes showed more balanced usage. Isoform diversity, measured as Shannon entropy, correlated modestly with expression level (Spearman ρ 0.27-0.34; Fig. S5C), which may reflect both biological complexity and greater power to detect low-abundance isoforms in deeply sequenced genes.

Independently of the island-derived 3'-end models, we reconstructed transcript structures with IsoQuant and classified them with SQANTI3 (Tardaguila et al.,

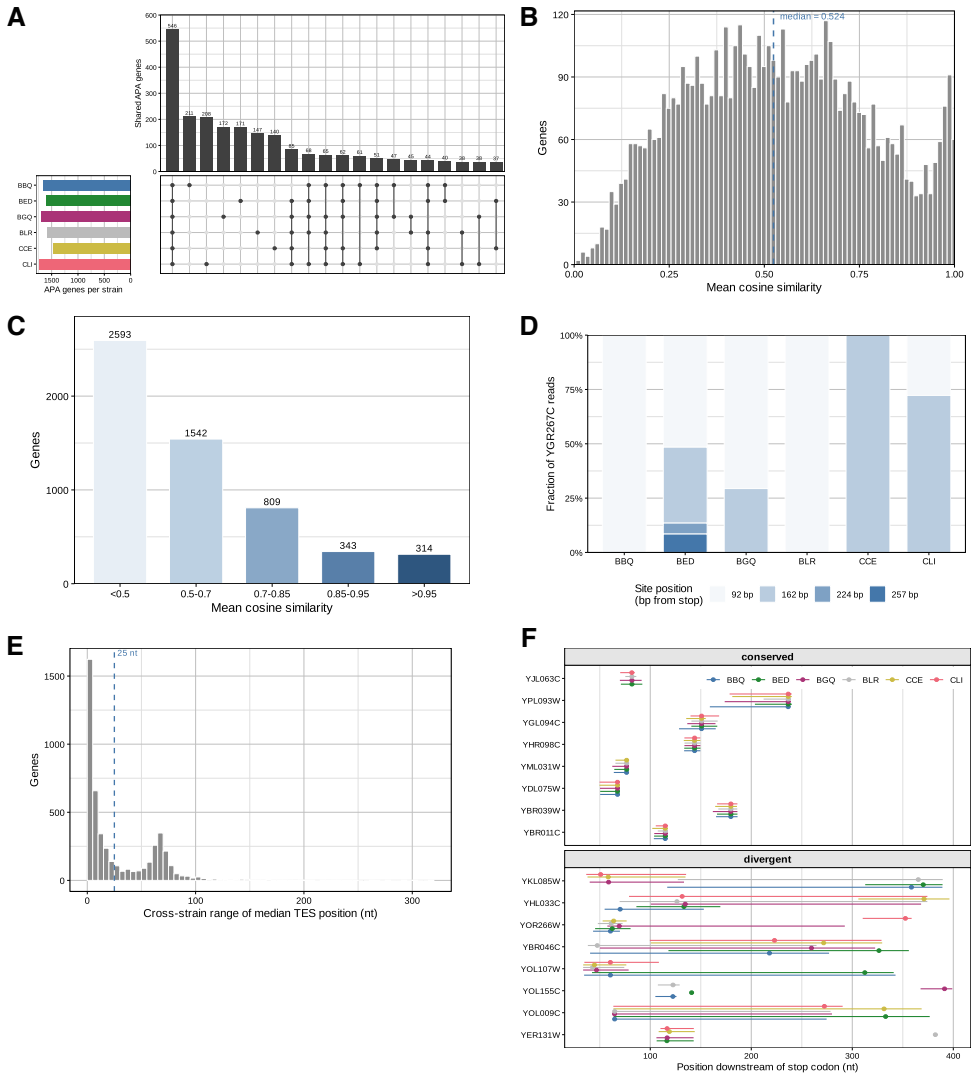


Figure 4 | Conservation and quantitative variation of transcript-end profiles.

(A.) UpSet summary of genes classified as alternatively polyadenylated in each strain, showing per-strain totals and the 20 largest intersections. (B.) Distribution of per-gene mean pairwise cosine similarity between 5-bp-binned transcript-end profiles for genes with at least 10 reads in at least four strains. (C.) Numbers of genes in five mean-cosine-similarity intervals. (D.) Fraction of *YGR267C* reads assigned to matched site classes across strains; positions are measured downstream of the stop codon. (E.) Cross-strain range of the median TES position among genes supported by at least 30 reads in at least four strains, with a 25-nt effect-size threshold. (F.) Selected conserved and divergent loci, summarized by strain median and 10th–90th percentiles of read-end positions relative to the stop codon.

2018; Pardo-Palacios et al., 2024) using the strain-specific near-telomere-to-telomere assemblies and annotations from the 1,086-genome yeast project (Loegler et al., 2025). SQANTI3 categories describe splice-junction and exon–intron structure, while transcript boundaries were evaluated separately. Across strains, 83–91% of structural models matched a reference annotation, whereas 6–13% mapped to genes absent from the S288C reference and therefore represented strain-specific accessory-genome loci (Fig. S5D). Only a small fraction constituted candidate new structural isoforms of annotated genes: 22–40 high-quality (HQ) novel isoforms and 130–202 lower-quality candidates were detected per strain (Fig. 5B). In total, 1,240 novel-isoform calls were assigned to 952 strain-specific annotation identifiers. The 191 HQ calls involved 178 such identifiers corresponding to 75 canonical host genes (Supplementary Data 3).

HQ novel isoforms were not restricted to negligible expression levels. Across 147 matched pairs, their abundance was comparable to that of the corresponding full-splice-match transcript pool, the per-strain median ranging from 0.53 to 4.15 times the reference median (Fig. 5C), with no consistent overall difference (paired Wilcoxon $P = 0.07$). When expressed as a fraction of host-gene reads, the median contribution of an HQ novel isoform ranged from 20% in BED to 56% in CCE, and 39–55% of HQ novel isoforms accounted for more than half of their gene’s expression (Fig. 5D). Many of these transcripts therefore represent major rather than incidental isoforms.

Structurally, the HQ novel isoforms fell into two groups (Fig. 5E). Half (95 of 191; 12–20 per strain) were classified as genic, meaning they overlap both exons and introns of an annotated gene without reproducing its structure. The remainder were dominated by intron retention: all 73 novel-in-catalogue models (6–18 per strain) retained an annotated intron, 61 of them as fully unspliced single-exon transcripts, as did the four incomplete splice matches (Fig. S5E). Only 19 models (2–5 per strain) used a splice site absent from the annotation and were therefore classified as novel-not-in-catalogue. Thus 40% of HQ novel isoforms are unspliced or partly unspliced forms of annotated genes rather than new gene structures. Host genes with HQ retained-intron models were not enriched for ribosomal-protein genes relative to the eligible background of expressed intron-containing genes (10 of 33 versus 84 of 244; odds ratio 0.81, 95% confidence interval 0.32–1.87; Fisher’s exact $P = 0.695$; Supplementary Table S3). Their termination sites were displaced from annotated gene boundaries by a median of 126–145 nt (Fig. S5F). Because the strain annotations contain gene, transcript, exon and CDS records but no explicit UTR features, these are offsets from ORF/CDS-like boundaries rather than empirically annotated 3′-UTR endpoints. Direct RNA sequencing also truncates 5′ ends. The offsets could therefore reflect alternative transcript boundaries, strain-specific annotation or model differences and do not alone establish new APA events. Novel genes, by contrast, ranged from 328 in CCE to 875 in CLI and scaled with library depth (Spearman $\rho = 0.94$, $P = 0.017$), reinforcing the need to interpret discovery counts in light of sequencing coverage.

Genes carrying a new isoform had shorter median poly(A) tails than reference-only genes in every strain (Wilcoxon $P < 7.8 \times 10^{-5}$), a difference of 6–8 nt (Fig. S5G). A linear model with strain as a covariate estimated a -5.9-nt difference ($P = 1.3 \times 10^{-37}$). The association persisted after additionally modelling expression, both linearly (-3.08

nt, $P = 2.3 \times 10^{-15}$) and with a three-degree-of-freedom spline (-2.14 nt, $P = 8.3 \times 10^{-9}$; Supplementary Table S3). Higher expression therefore contributes to, but does not fully explain, the shorter tails of novel-isoform host genes.

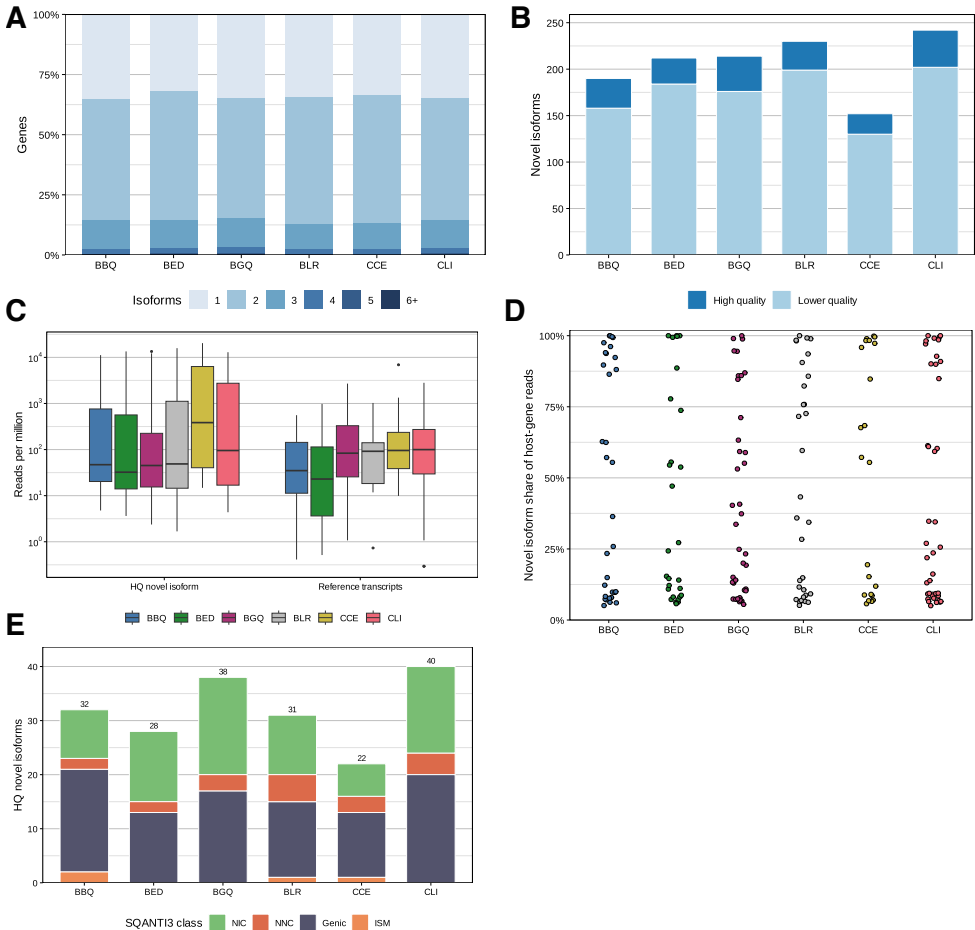


Figure 5 | Isoform diversity and candidate novel transcripts.

(A.) Fraction of expressed genes with one to six or more quantified island-derived transcript models. (B.) Numbers of high-quality and lower-quality novel isoforms identified by IsoQuant and SQANTI3 in each strain. (C.) Read abundance of high-quality novel isoforms and the summed full-splice-match reference transcripts of their host genes. (D.) High-quality novel-isoform abundance as a fraction of all reads assigned to the host gene. (E.) Structural classes of high-quality novel isoforms. NIC, novel in catalogue; NNC, novel not in catalogue; ISM, incomplete splice match.

Poly(A) tail length and APA are associated in gene-specific and strain-associated patterns

We then investigated whether poly(A) tail length depends on the polyadenylation site used by a transcript. Each read was assigned to a TES island by its 3' end, and tail-length distributions were compared between proximal and distal islands in genes with exactly

two islands and at least five reads at each. Median tail lengths at the two sites were within 0-1 nt of each other in every strain (Fig. S6A). The per-gene difference, defined as distal minus proximal median tail length, had a median of 0 nt in BBQ, BLR and CLI, +0.5 nt in BGQ and CCE, and +1 nt in BED. The fraction of genes with longer distal tails ranged from 45.9% in CLI to 54.2% in BED, and per-gene medians clustered tightly around the diagonal (Fig. 6A). Paired signed-rank tests reached significance in four of six strains (BED, BGQ, BLR and CCE; $\text{padj} \leq 2.0 \times 10^{-4}$), but no median effect exceeded 1 nt. In human cells, distal APA isoforms carry longer poly(A) tails than proximal isoforms of the same gene (Legnini et al., 2019); we detect no equivalent genome-wide directional relationship in budding yeast.

Individual genes departed from this flat average in both directions, with per-gene differences spanning 35 nt shorter to 26 nt longer at the distal site (Fig. 6B). Most genes used both sites: 94.3% assigned 10–90% of reads to the proximal island, whereas only 1.8% used either site almost exclusively ($\leq 5\%$ or $\geq 95\%$ proximal; Fig. 6C). APA in budding yeast is therefore quantitative, varying continuously in isoform proportions and in the tail lengths associated with each site.

To test whether poly(A)-tail length differed among TES islands within individual genes, we compared read-level tail lengths by Kruskal-Wallis tests with Benjamini-Hochberg correction within each strain. Between 27.7% of multi-island genes in CCE and 54.5% in CLI showed a significant association between TES identity and tail length ($\text{FDR} < 0.05$; Fig. S6B). The direction of the effect varied from gene to gene, with the longer tail at the distal site in some genes and at the proximal site in others, matching the near-symmetric spread of per-gene differences (Fig. 6B). TES identity is therefore frequently associated with a distinct tail-length distribution even though the relationship has no common genome-wide direction.

We further defined large site-associated tail-length differences as an absolute proximal–distal median difference of at least 10 nt. Between 40 and 97 genes per strain and direction met this threshold. The two directions were balanced in BBQ, BGQ, CCE and CLI, whereas BED and BLR carried an excess of genes with longer distal tails (64.6% and 62.0% of their large-effect genes; binomial $\text{padj} = 0.015$ and 0.016 ; Fig. 6D). Large-effect gene sets were almost entirely strain-associated: pairwise Jaccard indices ranged from 0 to 0.033, and no strain pair overlapped more than expected by chance (1,000 permutations, minimum $\text{padj} = 0.14$; Fig. S6C). The 883 large-effect gene–strain combinations resolved into 818 gene–direction pairs across 791 genes, of which 60 recurred in at least two strains and five in three (*YER169W*, *YHR166C*, *YLL008W*, *YDR142C* and *YMR059W*); none recurred in four or more (Fig. S6D). The lack of recurrence argues against a conserved gene-intrinsic pattern, but these data cannot identify the cis, trans or technical sources of the strain-associated differences.

Matched APA usage differs quantitatively among isolates

Finally, we quantified population-level differences in isoform usage. Each deduplicated island was represented by the highest raw-count local peak within it. Within each gene, sorted summit positions relative to the stop codon were greedily chained when every

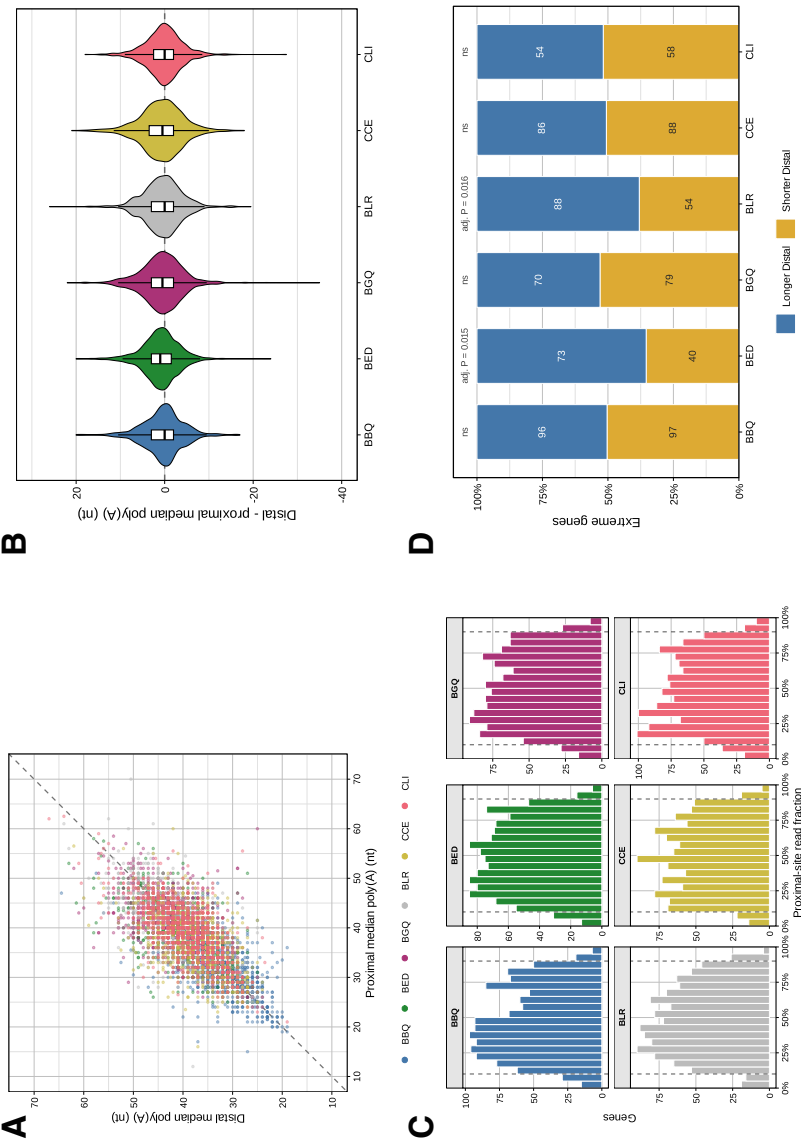


Figure 6 | Association between polyadenylation-site choice and tail length.

(A.) Per-gene median poly(A)-tail length at proximal and distal islands in genes with exactly two islands supported by at least five reads each. (B.) Distribution of the per-gene difference between distal- and proximal-island median tail lengths. (C.) Distribution of the fraction of reads assigned to the proximal island, with 10% and 90% usage thresholds. (D.) Direction of extreme coupling, defined as an absolute proximal–distal median difference of at least 10 nt. Benjamini–Hochberg-adjusted binomial tests assess directional imbalance.

consecutive gap was at most 50 bp; consequently, a chained class can span more than 50 bp. Of 1,028 genes carrying at least two matched island classes in at least three strains, 1,018 (99.0%) showed significant heterogeneity in island proportions (chi-squared test, Monte Carlo P values for sparse tables, BH-adjusted $P < 0.05$; Fig. S7A). Requiring in addition a Cramér's V of at least 0.15 and a maximum usage difference of at least 20 percentage points left 921 genes (89.6%) with differential usage of 3'-end isoforms (DIU; Fig. 7A); the effect-size criterion was the limiting one, since every gene that passed it was also significant. In 673 of these, the dominant island class differed between strains, with 226–374 such genes per strain pair (Fig. S7B).

A further 1,371 islands collapsed to 1,324 positional classes with no counterpart in another strain under the 50-bp rule. Of these classes, 1,025 (77.4%) were on genes detected in more than one strain and 299 (22.6%) were on genes called in a single strain (Fig. S7C). The first group is consistent with a strain-associated difference in end position, whereas the second cannot be separated from a gene-detection difference. These counts describe unmatched operational classes rather than 1,324 independently validated strain-specific biological events.

To describe whether the molecular layers arranged the six isolates similarly, we built and clustered separate strain-distance matrices from gene expression, poly(A) tail length, novel-isoform sharing and the number of TES islands per gene (Fig. 7B). Pairwise cophenetic correlations between the four dendrograms ranged from -0.18 to 0.88 and were highest between the poly(A)-tail and APA island-count trees (Fig. S7D). We then compared APA distances with an independent genomic-distance matrix constructed from 119,062 biallelic SNPs. No positive concordance was detected, either with all islands or after requiring at least 10, 20 or 50 supporting reads per island (Spearman $r = -0.21$ to -0.39 ; exact permutation $P = 0.32$ – 0.53 ; Supplementary Table S3). Focusing on the strongest APA calls therefore did not recover a phylogenetic signal in this six-isolate panel.

Complete switches accounted for 432 of the 863 DIU genes whose stop-codon anchors agreed within 200 bp, where one matched class carried all of a gene's assigned reads in at least one strain and none in another. In the more stringent well-covered subset, 227 anchor-consistent DIU genes had at least 200 assigned reads in every strain and 66 included a complete switch. The twenty largest well-covered changes therefore ranged from complete to partial reallocations of matched-site usage and were supported by at least 456 reads per strain (Fig. 7C). Overall, 921 genes met the predefined statistical and effect-size criteria for differential 3'-end usage, and 1,025 unmatched positional classes occurred on genes detected in multiple strains. Both quantities depend on matching resolution: at a 20-bp tolerance, 2,175 genes were testable, including 1,167 that did not meet the default two-class testing requirement at 50 bp. These results show strain-associated quantitative variation in transcript-end choice, without establishing its heritability or causal genetic basis.

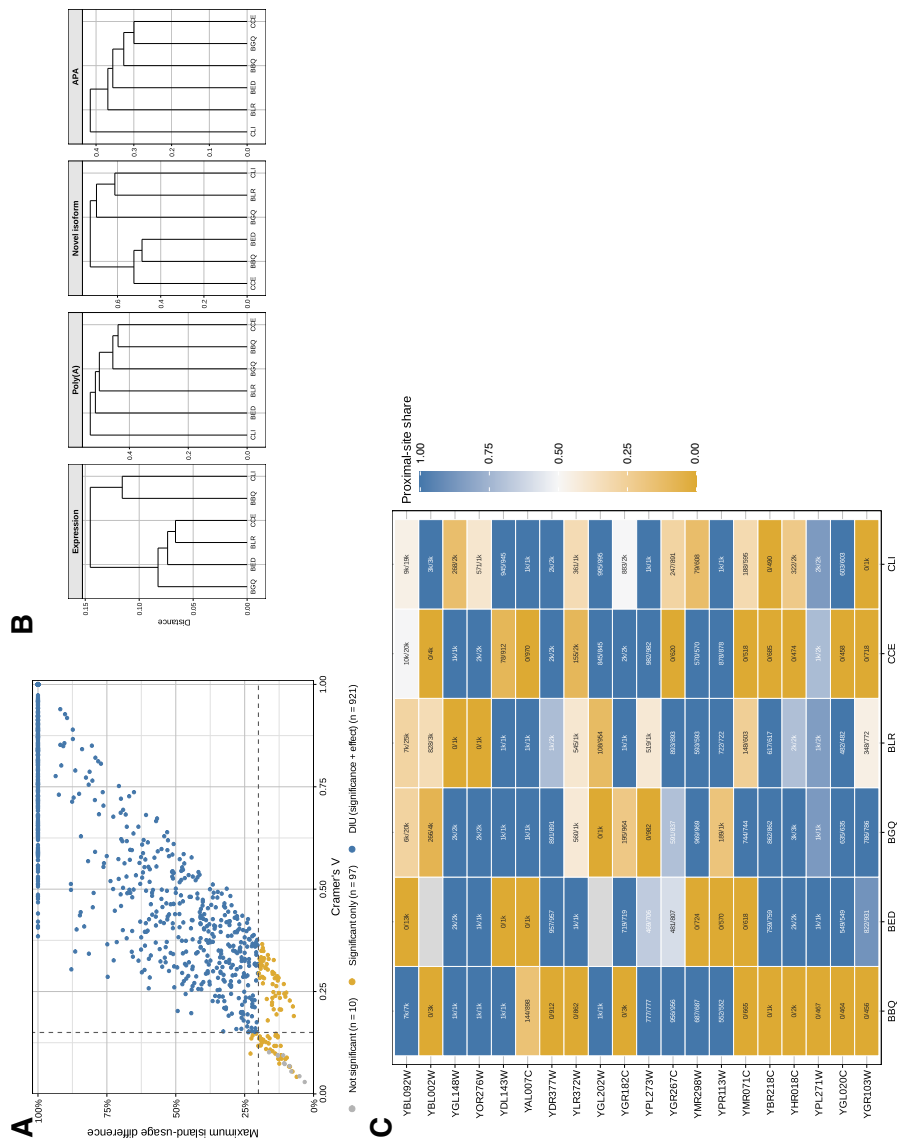


Figure 7 | Quantitative variation in matched transcript-end usage.

(A.) Effect-size landscape for the 1,028 genes with at least two matched island classes in at least three strains. Differential isoform usage (DIU) required an adjusted chi-squared-test $P < 0.05$, Cramér's $V \geq 0.15$ and a maximum class-usage difference ≥ 20 percentage points. (B.) Ward-clustered strain dendrograms derived from gene expression, gene-level poly(A)-tail length, novel-isoform sharing and APA island counts. (C.) Proximal-site usage among the twenty largest well-covered, anchor-consistent DIU changes, with assigned read counts.

Discussion

Most genome-wide maps of poly(A) tail length and transcript-end usage in *S. cerevisiae* derive from S288C or related laboratory strains, leaving their generality across natural isolates unclear. Using direct RNA sequencing and strain-specific near-telomere-to-telomere assemblies (Loegler et al., 2025), we compared six natural isolates. Gene expression and overall gene-level poly(A)-tail patterns were broadly conserved, whereas 3' end profiles and relative end-region usage differed at hundreds of genes.

At the gene level, poly(A)-tail central tendency was more strongly associated with gene identity than strain (7.1% versus 1.0% of read-level sum of squares), although most variation occurred among reads from the same gene and isolate. Pooled read-level cross-isolate tests were significant for many genes, but typical shifts were modest (median 8.6 nt), so significance here has to be read alongside effect size. Binary dip-test rejection was strongly depth-dependent and should not be treated as evidence for bimodal populations. At matched depth, positive correlations of per-gene dip statistics between isolates suggest a partly gene-associated component to departures from unimodality. The inverse expression–tail association was shared across isolates and is consistent with, but does not demonstrate, co-translational deadenylation (Subtelny et al., 2014; Montañés et al., 2022; Passmore & Collier, 2022).

At equal depth, we called 8,067–8,852 operational polyadenylation islands per strain, with 23.7–26.7% of expressed genes carrying at least two. These rates are lower than earlier coordinate-level Helicos estimates (Ozsolak et al., 2010; Moqtaderi et al., 2013), but the quantities are not directly comparable: our caller groups recurrent nearby ends into prominent, supported islands and may not represent minor positions as separate sites. This operational definition provides a consistent unit for cross-isolate comparison. Similar enrichment of known yeast polyadenylation-related motifs across strains supports the biological plausibility of the called regions (Guo & Sherman, 1995; Graber et al., 1999; Liu et al., 2017; Stroup & Ji, 2024). A small fraction of accepted islands lay upstream of the annotated stop codon, but our transcript-end windows were not designed to interrogate 5'-UTR or early-CDS alternative termination systematically. Such sites, including those associated with translated upstream ORFs (Moqtaderi et al., 2013; Montañés et al., 2026), remain an important extension for a dedicated pipeline.

Across isolates, median transcript-end position was often similar (median cross-strain range 11.9 nt), whereas island detection remained coverage-sensitive and matched end-region usage frequently differed. Among 1,028 genes with matched island classes, 921 met the predefined criteria for differential island usage; 432 complete switches occurred among the 863 DIU genes with stop-codon anchors agreeing within 200 bp. These values are operational rather than absolute: the consecutive-gap chaining rule can join classes spanning more than 50 bp, while a 20-bp tolerance makes an additional 1,167 genes testable.

In pooled read analyses, TES-island identity was associated with tail-length distributions in 27.7–54.5% of tested multi-island genes, but aggregate proximal–distal differences

did not exceed 1 nt, unlike the directional relationship reported in human cells (Legnini et al., 2019). We also identified 191 high-quality candidate novel-model calls relative to strain-matched annotations across 75 canonical host genes. Many had appreciable read support, intron retention was common, and their host genes had shorter tails than reference-only genes. The association persisted after flexible adjustment for expression, and retained-intron host genes were not enriched for ribosomal-protein genes, but this between-gene comparison still does not identify a mechanism. In *Schizosaccharomyces pombe*, retained-intron isoforms instead had longer tails than reference isoforms of the same gene (Montañés et al., 2022); that within-gene comparison and our between-gene comparison address different contrasts and are not directly equivalent.

The four molecular-layer dendrograms did not recover a common isolate arrangement, and an independent SNP analysis showed no positive relationship between genomic and APA distances, even for islands supported by at least 50 reads. This absence of a clear phylogenetic signal could reflect the small number of isolates and pairwise distances, the one-isolate-per-clade design, environmental or trans-regulatory effects, depth-sensitive detection of minor sites, or simply that quantitative APA regulation does not track neutral genomic divergence over these lineages. With six isolates and no hybrids, we cannot map causal variants or separate cis from trans effects. Island calls merge nearby cleavage positions and remain depth-dependent, so the reported island counts are conservative relative to the fine-scale cleavage repertoire. Direct RNA sequencing also limits inference at transcript 5' ends, and measured 3' offsets of candidate models from ORF/CDS-like annotation boundaries do not alone distinguish APA from annotation or model differences (Soneson et al., 2019; Jain et al., 2022; Prawer et al., 2023).

Hybrid allele-specific direct RNA sequencing and larger, deeply sequenced panels would distinguish cis and trans contributions and map variants associated with 3' end use (Kita et al., 2017; Albert et al., 2018; Mittleman et al., 2020; Li et al., 2021; Tsouris et al., 2024a, b), and measurements of RNA stability, translation and protein abundance would then connect the APA differences observed here to function. Together, these data show that 3' end isoform usage varies among natural yeast isolates in ways that gene-level abundance alone cannot capture.

Methods

Yeast strains, growth and RNA extraction

Six *Saccharomyces cerevisiae* natural isolates were selected from the 1,011 genomes collection (Peter et al., 2018) to represent distinct ecological and genetic backgrounds: BBQ (Mediterranean oak), BED (French Guiana human), BGQ (French dairy), BLR (Wine/European), CCE (Ecuadorean) and CLI (Sake) (Supplementary Table S1). Two independent biological replicates were prepared per strain. Cultures (50 mL) were grown in 250 mL flasks containing synthetic complete medium supplemented with glucose at 30°C and 120 rpm to mid-log phase (OD600 approximately 0.7). Cells were collected by vacuum filtration through nitrocellulose filters, flash-frozen in liquid nitrogen and stored at -80°C.

Total RNA was extracted with the RNeasy Mini Kit (Qiagen #74104), including on-column DNase I treatment (Qiagen #79254), and polyadenylated RNA was enriched with the Dynabeads mRNA Direct Kit (ThermoFisher #61012). RNA concentration was measured with the Qubit RNA High Sensitivity Assay (ThermoFisher #32852).

Direct RNA sequencing and basecalling

Libraries were prepared from 750 ng polyadenylated RNA with the Oxford Nanopore SQK-RNA002 protocol, including reverse-transcription-adapter ligation, SuperScript III reverse transcription and RMX sequencing-adapter ligation. Libraries were sequenced on MinION FLO-MIN106D or PromethION FLO-PRO002 flow cells. Dorado v0.5.3 was used for basecalling with the `rna002_70bps_hac@v3` model, a minimum Q score of 7 and `--estimate-poly-a` to obtain per-read poly(A)-tail estimates from the ionic-current signal.

Read correction, alignment and gene expression

Dorado BAM files were converted to FASTQ with `dorado demux --emit-fastq`, and reads outside 75-16,000 nt were discarded. Reads were corrected with FMLRC2 v0.1.7 (Wang et al., 2018; Mak et al., 2023) using strain-matched Illumina data (Caudal et al., 2024), then refined with TranscriptClean v2.1 (Wyman & Mortazavi, 2019). Corrected reads were aligned with minimap2 v2.30 in splice-aware mode (`-ax splice -uf -k14 -G1500`), and alignments were sorted and indexed with samtools v1.20. Expression quantification and raw-versus-corrected comparisons used alignments to the common S288C R64-3-1 reference. Transcript-end, APA and poly(A) analyses used separate alignments to the corresponding near-telomere-to-telomere strain assembly (Loegler et al., 2025). Unless stated otherwise, downstream analyses used corrected reads pooled across the two replicates of each strain.

Gene-level counts were obtained with featureCounts (Subread v2.1.1) (Liao et al., 2014) in long-read and stranded mode (`-L -s 1 -O -M --minOverlap 10`) against the R64-3-1 GTF reduced to `gene_id` attributes. Replicate concordance was calculated by

Pearson correlation of $\log_2(\text{count} + 1)$. Cross-strain comparisons used nuclear ORF counts normalised to counts per million (CPM) and $\log_2(\text{CPM} + 1)$ -transformed. A gene was considered expressed when supported by at least five reads. For analyses based on Salmon isoform estimates, gene abundance was expressed as direct reads per million (DPM): summed assigned reads per gene divided by the total assigned reads and multiplied by 10^6 .

Genes with mean CPM at least 10 were ranked by their coefficient of variation across strains. The most variable decile was tested for Gene Ontology enrichment with `clusterProfiler::enrichGO` using `org.Sc.sgd.db`, all genes passing the CPM threshold as background, `pvalueCutoff = 0.05`, `qvalueCutoff = 0.2` and Benjamini-Hochberg correction. Of 506 target and 5,051 background ORFs, 458 and 4,874 respectively mapped to the annotation database. Equivalent tests were performed for genes with the longest tails, broad tail distributions, divergent transcript-end positions and extreme APA-tail coupling, using all genes tested in the corresponding analysis as background.

Depth equalization

Corrected pooled BAMs were restricted to primary alignments with MAPQ at least 20 (`samtools view -F 0x904`) and subsampled to 6,840,171 reads, the number present in the shallowest library (CCE). All main transcript-end, APA and per-gene poly(A) analyses used these equalized BAMs. The same transcript-end pipeline was also run on the full-depth BAMs to assess coverage sensitivity. For the global poly(A)-tail distribution, the main estimate pooled the two libraries within each strain and was therefore depth-weighted. As a replicate-balance sensitivity analysis, up to 500,000 tail-bearing reads were sampled independently from each library before the two samples were combined with equal weight.

Transcript-end detection and polyadenylation islands

Per-base strand-specific coverage was generated with `bedtools v2.31.1` (Quinlan & Hall, 2010). Annotated transcript ends were extended to the last position at which coverage exceeded the larger of five reads or 10% of coverage at the annotated end. Read 3' ends were intersected with windows extending 500 bp upstream and 150 bp downstream of the coverage-extended annotated end, and polyadenylation islands were called per gene with `call_TES.py`. The strain annotations contain gene, mRNA, exon and CDS records but no explicit UTR features; their transcript boundaries are therefore ORF/CDS-like rather than empirically annotated UTR endpoints. The windows can capture limited sites upstream of the stop codon but were not designed for a systematic 5'-UTR or early-CDS termination analysis.

The caller first required at least five 3'-end reads and mean coverage of at least 0.1 reads per bp. Read-end positions were smoothed by Gaussian kernel density estimation (sigma 5 bp), and contiguous regions above 30% of the within-gene maximum were carved into candidate islands, with a maximum width of 300 bp. Sub-peaks were

separated when the intervening valley was no more than 30% of the lower peak, their summits were at least 20 bp apart and the smaller sub-peak represented at least 3% of the gene's reads. Non-primary islands required a density of at least 0.75 reads per bp and a peak of at least eight reads. Adjacent islands were merged across gaps of at most 20 bp subject to the valley criterion, or when their summits were within 25 bp and the shorter summit was at least half the taller. Retained islands were trimmed to the shortest span containing 90% of their read-end mass. The final filters required each island to represent at least 3% of gene reads and allowed at most five islands per gene. Parameters were selected from a 72-combination main sweep, followed by four lower-bandwidth combinations, across the six isolates using positional concordance with Ozsolak et al. (2010) and consistency of APA rates. A gene with at least two retained islands was classified as an APA gene.

Transcript-end validation

Calls in the S288C reference background were compared with the Helicos poly(A)-site catalogue of Ozsolak et al. (2010). Ozsolak sites were assigned to the nearest same-strand annotated gene end within 500 bp, and nearest-site distances were computed in both directions. For gene-level comparison, Ozsolak positions separated by no more than 25 bp were grouped into one site; genes with at least two Ozsolak sites or at least two islands in our data were classified as APA in the corresponding dataset.

For sequence validation, exact duplicate peak coordinates produced by overlapping gene or transcript models were counted once, preferring systematic ORF assignments and then the model whose CDS anchor was nearest. Sequence composition was analysed strand-aware from 90 nt upstream to 40 nt downstream of each retained peak using 11 literature-derived patterns: TA-rich efficiency-like elements (TATATA, TATAAA and ATATAT), U/T-rich elements (TTTTTT, TTTTAT, TTTTAA, TTCTTT and TCTTTC), and A-rich positioning-like elements (AATAAA, AAAAAA and AAATAA) (Guo & Sherman, 1995; Graber et al., 1999; Liu et al., 2017; Stroup & Ji, 2024). Presence of at least one pattern was compared with equally sized sets of matched random genomic positions by Fisher's exact test. De novo 6-mer enrichment in the same window was calculated as a Z score against the random positions.

Isoform quantification and classification

One transcript model per retained island was written to GFF and converted to FASTA with gffread v0.12.7. Corrected reads were aligned to this transcriptome with minimap2 (`-ax map-ont`), retaining primary alignments, and quantified with Salmon v1.10.3 in alignment-based mode. Because models assigned to one gene generally differ only at the 3' end, Salmon estimates were interpreted as 3'-end-choice usage rather than independent whole-transcript abundance. Genes with at least one assigned read were retained for gene-level model counts, largest-isoform read fractions and Shannon entropy; Fig. S5A separately counted transcript models with at least one assigned read.

Transcript structures were independently reconstructed with IsoQuant v3.4.1 (`--data_type nanopore --matching_strategy precise --splice_correction_strategy default_ont`) using strain-matched short-read STAR alignments and the corresponding strain annotation, then classified with SQANTI3 v5.2.1 (Tardaguila et al., 2018; Pardo-Palacios et al., 2024). This structural analysis was distinct from the island-derived 3'-end models described above. Models were curated into reference isoforms, high-quality or lower-quality novel structural isoforms of annotated genes, and models at loci absent from the annotation. SQANTI3 structural categories and subcategories were retained to distinguish full- and incomplete-splice matches, novel combinations of annotated splice sites, unannotated splice sites, intron-retention models and other genic models; transcript boundaries were evaluated separately. Expression of each high-quality novel isoform was compared with the summed full-splice-match expression of its host gene by paired Wilcoxon test. Its host-gene read fraction and its offsets from annotated 5' and 3' boundaries were obtained from SQANTI3. Genes carrying a high-quality novel isoform were compared with reference-only genes for median poly(A)-tail length by per-strain Wilcoxon tests and linear models containing strain and novel-isoform status, with sensitivity models additionally including either linear $\log_2(\text{DPM} + 0.01)$ or a three-degree-of-freedom natural spline of $\log_2(\text{DPM} + 0.01)$. Enrichment of ribosomal-protein genes among high-quality intron-retention host genes was tested by two-sided Fisher's exact test using IsoQuant-expressed, R64-annotated intron-containing genes as the background.

Poly(A)-tail analyses

Per-read tail estimates were taken from the Dorado `pt` tag, retaining positive values. A 300-nt upper bound was applied to the global distribution and variance-decomposition analyses; values above this bound represented at most 0.025% of reads in a strain. Reads were assigned to genes and islands by their 3' alignment coordinate. Per gene and island, we calculated median, mean, standard deviation and coefficient of variation of tail length. Cross-strain reproducibility was measured by Spearman correlation of gene medians, and expression dependence by Spearman correlation between $\log_{10}$ DPM and median tail length. To identify unusually broad or sharp distributions independently of the mean-variance relationship, per-gene tail-length coefficient of variation was regressed on mean tail length by loess (span 0.5), and the upper and lower 2% of residuals were retained as examples.

Read-level tail lengths from genes supported by at least 10 reads in at least three strains were summarised by marginal gene and strain sums of squares calculated from group means. The residual was the total sum of squares minus these two components and therefore includes gene-by-strain structure and read-level variation. Reads were stratified by gene and capped at 500,000 per strain. Hartigan's dip test (Hartigan & Hartigan, 1985) was applied per gene and strain when at least 100 reads were available, with Benjamini-Hochberg correction within strain. Depth dependence was evaluated by repeating the test after subsampling each eligible gene-strain pair to 100, 200, 400, 800 and 1,600 reads, with ten resamples per depth. For genes with at least 20 reads in at least three strains, read-level tails were compared across strains by Kruskal-Wallis test; effect

size was the log₂ ratio between the largest and smallest strain mean. These pooled-read tests describe molecule-level differences among the sampled cultures; effect sizes and depth analyses were used to avoid interpreting read-count-driven significance alone.

Cross-strain transcript-end comparisons

For downstream-region comparisons in genes with exactly two islands, sites were designated proximal or distal according to whether they were closer to or farther from the annotated stop codon in transcript orientation. Usage was summarised as log₂(proximal/distal reads) and as the proximal read fraction. Cross-strain profile similarity was calculated from 5-bp-binned individual 3'-end histograms indexed from each gene-strain's strand-specific window start. The window was anchored on the coverage-extended annotated transcript end and generally spanned 500 bp upstream to 150 bp downstream. This endpoint analysis retained fine-scale distributions independently of island boundaries. Genes required at least 10 reads in at least four strains, and similarity was measured by the mean pairwise cosine similarity.

Conservation of overall transcript-end position was assessed for nuclear ORFs using read ends 1-400 nt downstream of the stop codon. Genes required at least 30 reads in at least four strains. For each gene and strain we calculated weighted position percentiles, and tested read-level positions across strains by Kruskal-Wallis test after capping each strain at 3,000 reads. Genes were classified as divergent when $q < 0.05$ and the range of strain medians was at least 25 nt, conserved when q was at least 0.05 and the range was below 25 nt, and intermediate otherwise.

Variants relative to R64 were extracted from strain BCF files with bcftools v1.20. Canonical ORFs were parsed from the Name attribute of R64 CDS records, and putative 3' UTRs were defined as the 500 bp downstream of the final CDS coordinate. Non-reference SNPs were counted per gene and strain. SNP load was compared between genes with strain-specific and shared APA and between divergent and conserved transcript-end classes by Wilcoxon or Kruskal-Wallis tests as specified.

Differential isoform usage

To compare islands across strain-specific assemblies, each deduplicated BED island was counted once using its island support and was represented by the highest raw-count local peak within it. The summit's distance from the gene's stop codon was used as the shared coordinate. Within a gene, sorted summit positions were greedily chained into one matched class when every consecutive gap was at most 50 bp; this rule allows the total span of a chained class to exceed 50 bp. Genes with strain medians differing by more than 200 bp were flagged as anchor-inconsistent. Genes with at least two classes in at least three strains were tested for heterogeneity of island counts by chi-squared test, using 2,000 Monte Carlo replicates when any expected count was below five. Differential isoform usage required Benjamini-Hochberg-adjusted $P < 0.05$, Cramér's V at least 0.15 and a maximum class-usage difference of at least 20 percentage points. Dominant-class and complete switches were recorded, and matching sensitivity

was assessed by repeating the analysis with a 20-bp tolerance. Unmatched events were counted once per gene, strain and matched positional class rather than once per underlying local peak.

Coupling between polyadenylation site and tail length

For genes with exactly two islands supported by at least five reads each, proximal and distal median tail lengths were compared within strain by paired Wilcoxon signed-rank test. Differences in poly(A)-tail length among islands within individual genes were tested for genes with at least two islands supported by at least 10 reads each using a per-gene Kruskal-Wallis test on read-level tails, with Benjamini-Hochberg correction within strain. Large site-associated tail-length differences were defined as an absolute proximal–distal median difference of at least 10 nt. Directional balance was tested by binomial test, and overlap of large-effect gene sets was compared with 1,000 permutations that retained the number of genes tested per strain.

Strain relationships and statistical analysis

Strain distance matrices were calculated for gene expression and poly(A)-tail length as one minus Spearman correlation, for novel-isoform sharing as one minus Jaccard similarity, and for APA as one minus the Spearman correlation of island counts per gene. Matrices were clustered by Ward's method. Similarity between molecular-layer dendrograms was described by pairwise correlations of their cophenetic distances. Independently, genetic distances were calculated from 119,062 selected biallelic SNPs as the mean pairwise allelic-dosage difference over callable sites. Concordance between the 15 upper-triangle genetic and APA distances was measured by Spearman correlation, with an exact two-sided P value from all 720 permutations of strain labels. APA distance was recalculated after requiring each island to have at least 1, 10, 20 or 50 supporting reads. These analyses were treated as descriptive because only six isolates were available.

Analyses were performed in R v4.4.1 unless otherwise stated. Reported adjusted P values use the Benjamini-Hochberg method. Figures were generated with ggplot2 and associated R packages.

Data availability

The supplementary tables and data supporting this chapter are deposited on [Zenodo](#).

Supplementary figures

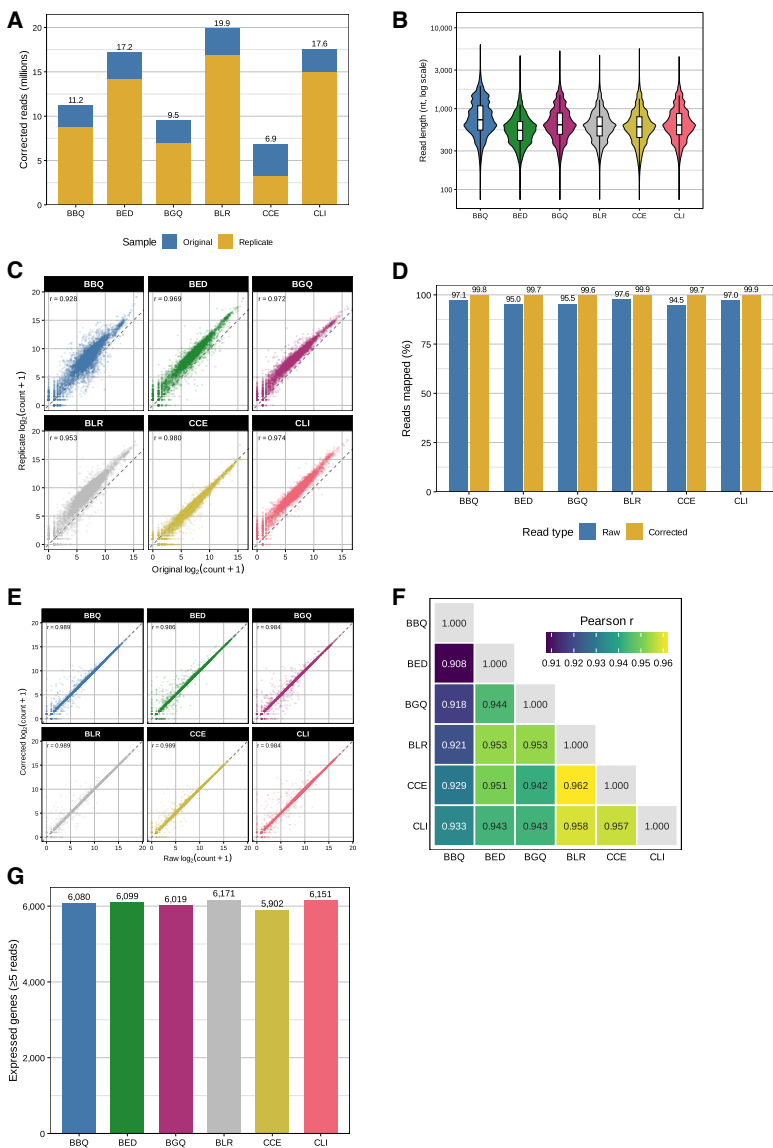


Figure S1 | Sequencing, correction and gene-expression quality controls.

(A.) Numbers of corrected reads from the original and replicate libraries, with per-strain totals. (B.) Read-length distributions on a log₁₀ scale. (C.) Original-versus-replicate gene counts after log₂ transformation, with Pearson correlations. (D.) Mapping rates of raw and corrected reads. (E.) Raw-versus-corrected gene counts after log₂ transformation, with Pearson correlations. (F.) Pairwise Pearson correlations of log₂ CPM expression values. (G.) Numbers of genes supported by at least five reads.

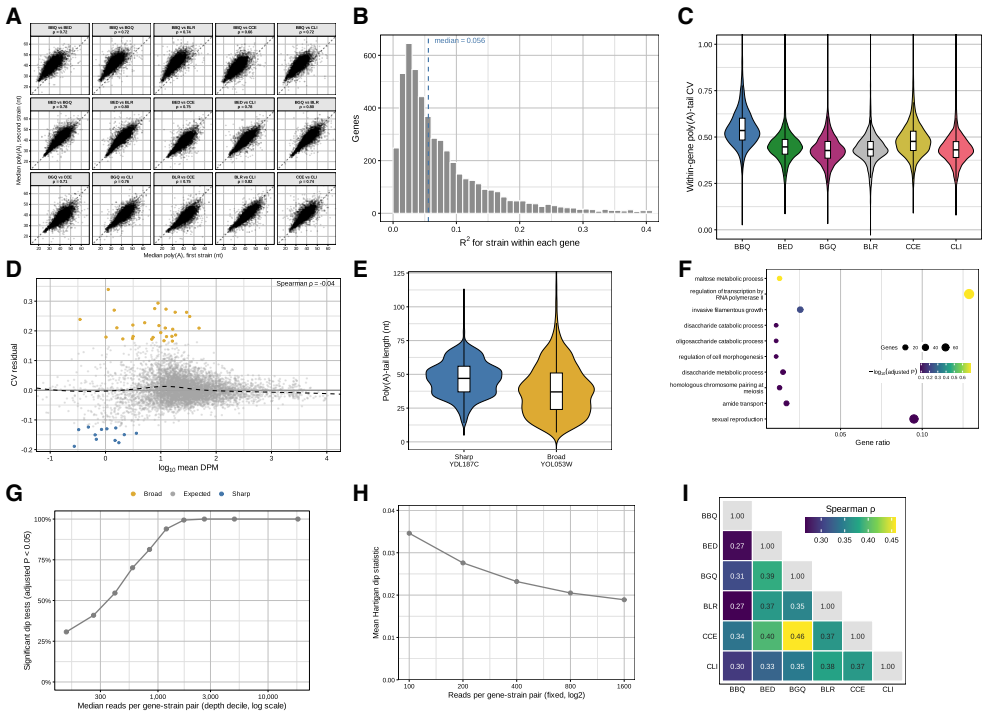


Figure S2 | Extended analyses of gene-level poly(A)-tail properties.

(A.) Pairwise comparisons of gene-level median poly(A)-tail length across all strain pairs, with Spearman correlations. (B.) Distribution of the fraction of within-gene tail-length variation explained by strain. (C.) Within-gene coefficient of variation of poly(A)-tail length by strain. (D.) Tail-length CV residual after removing its loess relationship with mean tail length, plotted against expression. The upper and lower 2% of residuals define broad and sharp distributions, respectively; the Spearman correlation is reported. (E.) Representative sharp (*YDL187C*) and broad (*YOL053W*) per-read tail-length distributions. (F.) Gene Ontology terms for the 10% of genes with the longest median tails; no displayed term reached $FDR < 0.05$. (G.) Fraction of significant Hartigan dip tests across read-depth deciles. (H.) Mean dip statistic after subsampling eligible gene-strain pairs to fixed depths. (I.) Pairwise Spearman correlations of per-gene dip statistics after subsampling each pair to 100 reads.

2

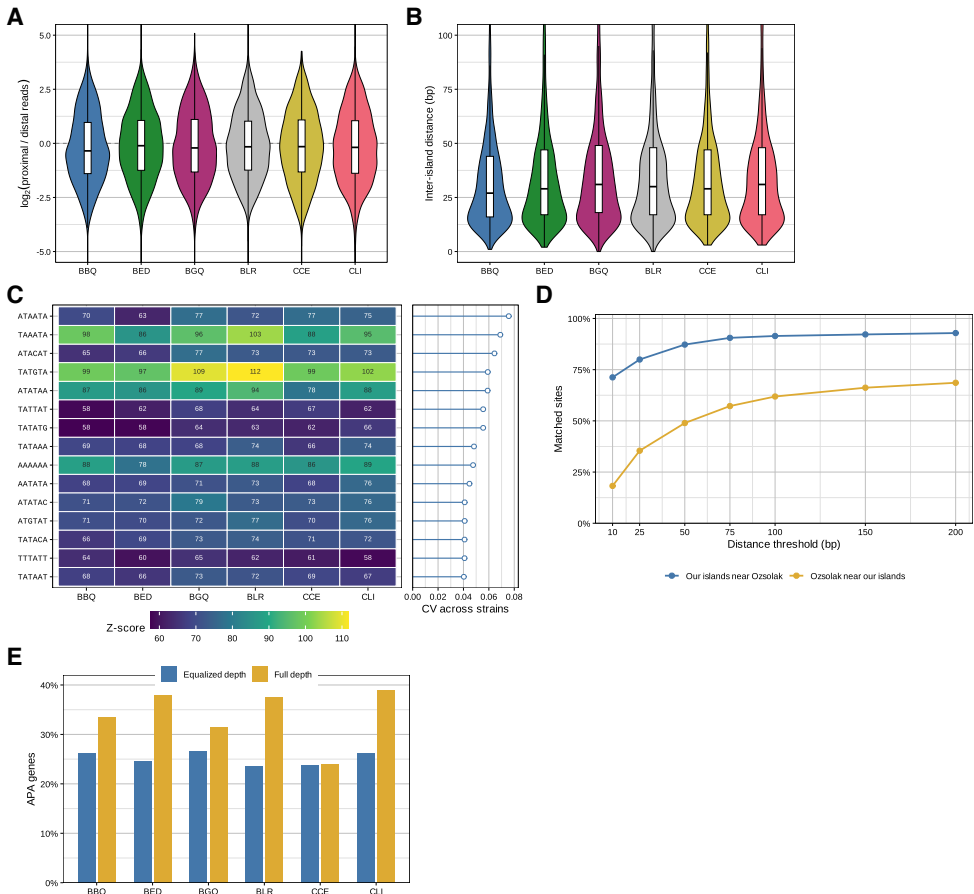


Figure S3 | Extended validation and properties of polyadenylation islands.

(A.) Log2 ratio of proximal- to distal-island reads for genes with exactly two islands. (B.) Distances between consecutive islands within genes. (C.) Cross-strain enrichment Z-scores and coefficients of variation for 15 upstream six-mers selected for high between-strain variability. (D.) Bidirectional positional concordance between TES islands detected in the S288C background and the Ozsolak et al. (2010) catalogue across increasing distance thresholds. (E.) Fraction of APA genes at the equalized depth of 6,840,171 primary alignments and at full sequencing depth.

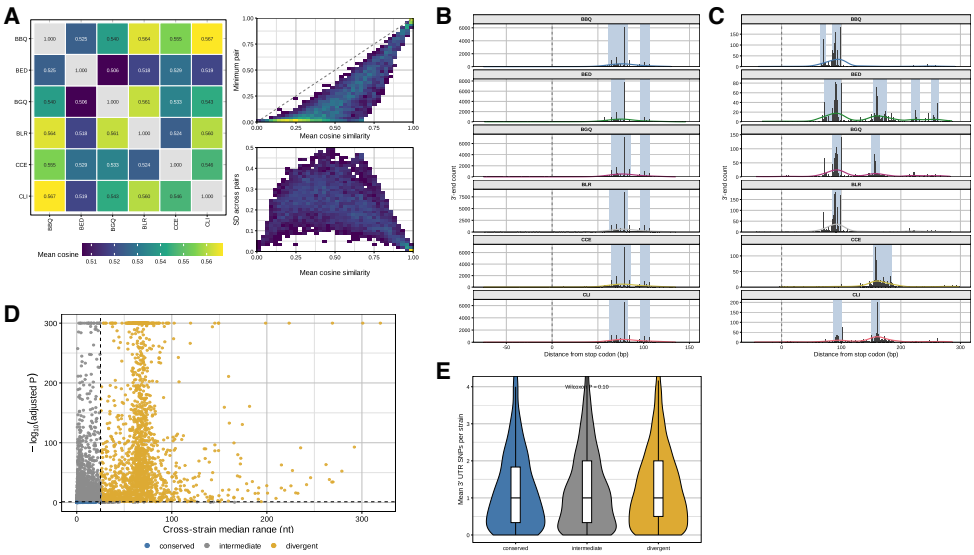


Figure S4 | Extended analyses of transcript-end conservation.

(A.) Pairwise and per-gene variation in transcript-end profiles, summarized by mean cosine similarity, minimum pairwise similarity and the standard deviation across pairs. (B.) *YDR064W*, a conserved-usage example, and (C.) *YGR267C*, a divergent-usage example, showing observed and kernel-smoothed 3'-end profiles, retained islands and the stop codon. (D.) Classification of TES-position conservation from the cross-strain median range and Kruskal–Wallis adjusted P value. Divergent genes have adjusted $P < 0.05$ and a range ≥ 25 nt, conserved genes have adjusted $P \geq 0.05$ and a range < 25 nt, and all others are intermediate. (E.) Mean number of non-reference SNPs per strain in the 500-bp region downstream of the final CDS coordinate for conserved and divergent genes, with a two-sided Wilcoxon rank-sum test.

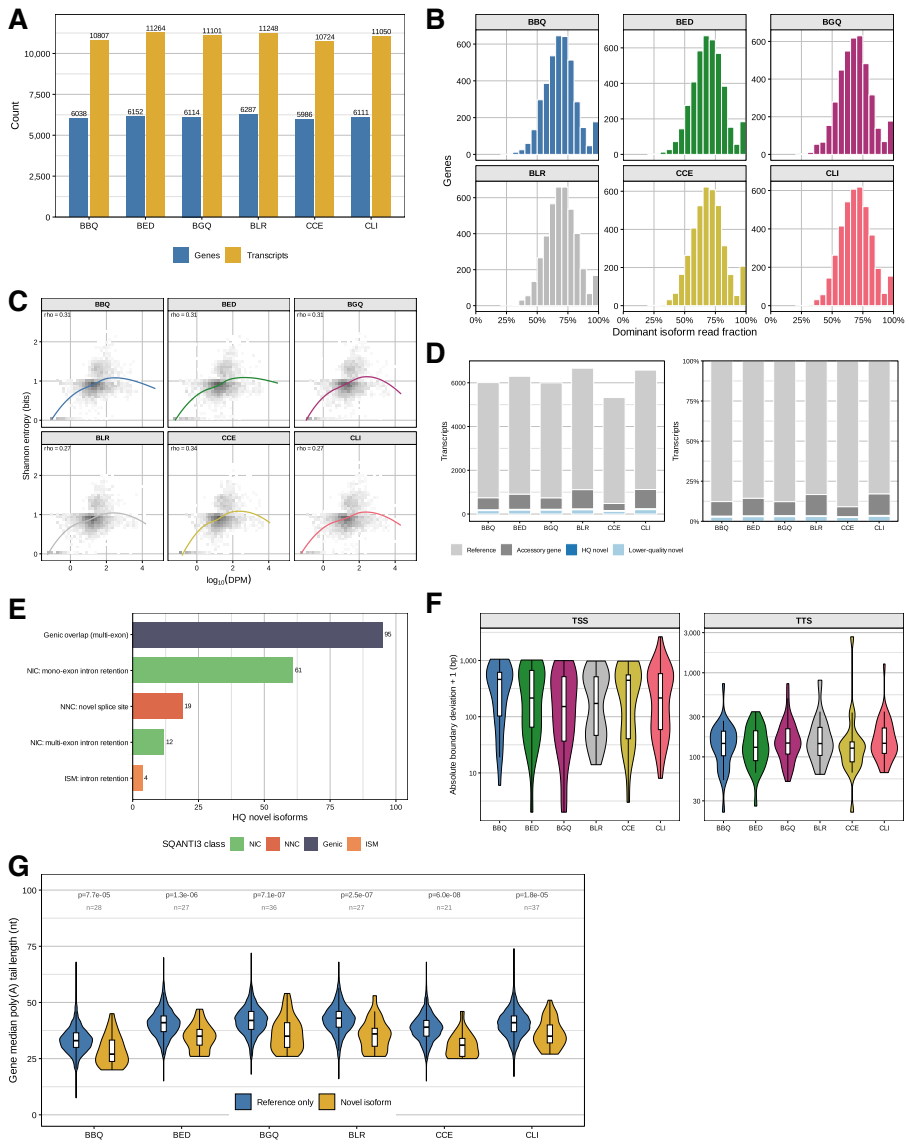


Figure S5 | Extended isoform and novel-transcript analyses.

(A.) Numbers of genes and transcript models receiving at least one assigned read. (B.) Distribution of the dominant-isoform read fraction among multi-isoform genes. (C.) Shannon entropy of isoform usage against expression, with per-strain loess fits and Spearman correlations. (D.) Counts and proportions of reference transcripts, accessory-gene transcripts, high-quality novel isoforms and lower-quality novel isoforms. (E.) Intron-retention and splice-structure subtypes among high-quality novel isoforms, grouped by SQANTI3 structural class. (F.) Absolute deviations of high-quality novel-isoform transcript start sites (TSSs) and transcript termination sites (TTSs) from annotated gene boundaries, $\log_{10}$ -transformed after adding 1 bp. (G.) Gene-level median poly(A)-tail length in genes with and without a high-quality novel isoform, with per-strain two-sided Wilcoxon rank-sum tests.

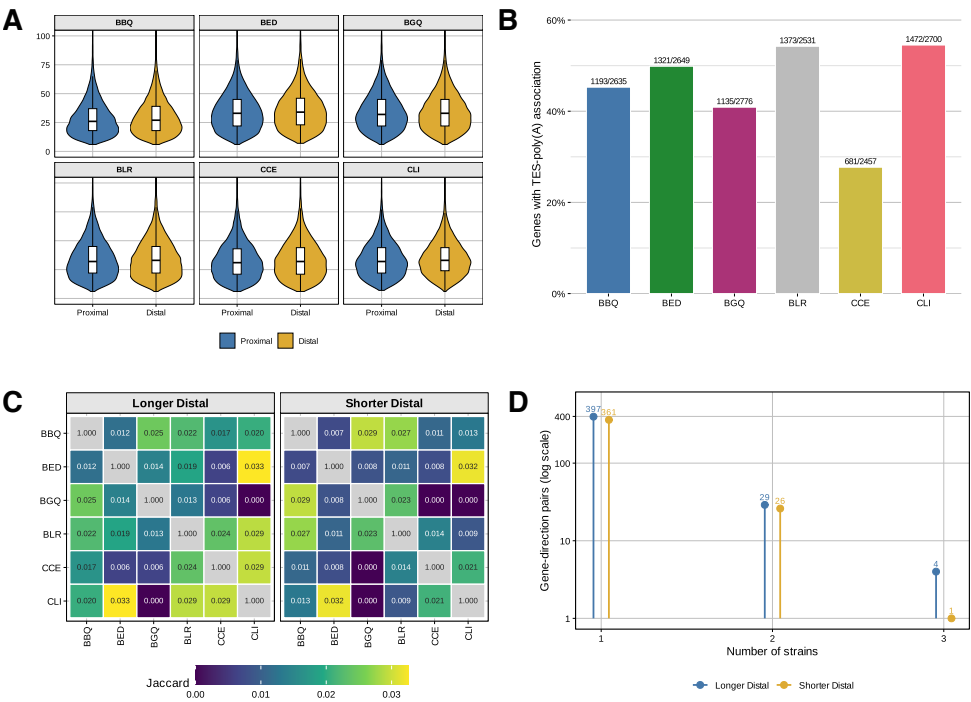


Figure S6 | Extended analyses of polyadenylation-site and tail-length coupling.

(A.) Per-read poly(A)-tail distributions at proximal and distal islands by strain. (B.) Fraction of multi-island genes in which read-level tail length differed among islands by Kruskal–Wallis test at Benjamini–Hochberg FDR < 0.05. Each island required at least 10 reads; significant and tested gene counts are reported. (C.) Pairwise Jaccard similarity of extreme-gene sets, separated into longer-distal and shorter-distal effects. (D.) Recurrence of extreme gene-direction pairs across one, two or three strains.

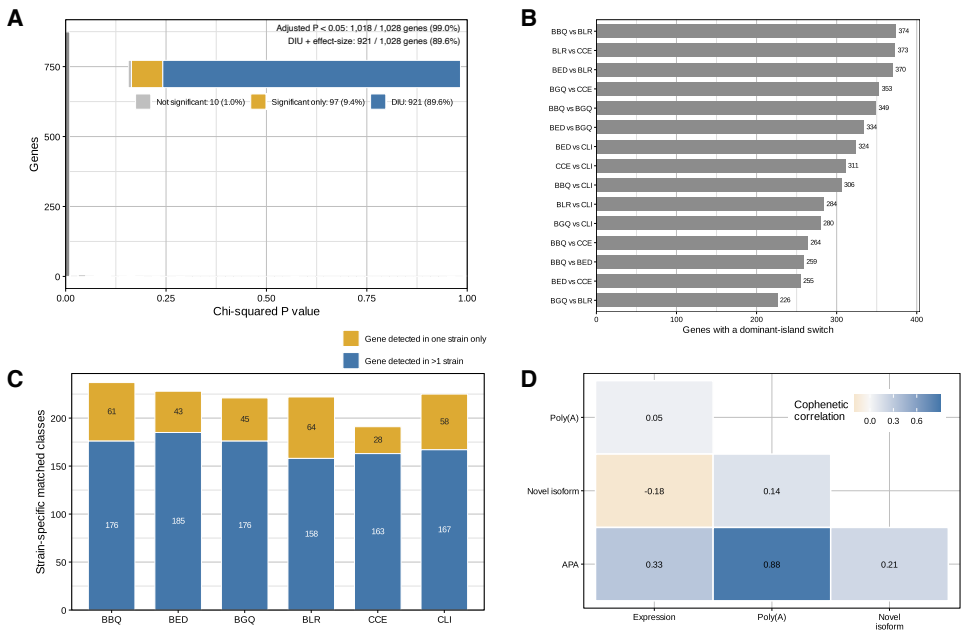


Figure S7 | Extended analyses of strain-associated transcript-end usage.

(A.) Partition of the 1,028 tested genes into not significant, significant without the required effect sizes, and DIU, with counts and percentages. (B.) Number of genes whose dominant matched island class switched for each strain pair. (C.) Numbers of positional classes unmatched within 50 bp in every other strain, separated according to whether the host gene was detected in one or multiple strains. (D.) Pairwise correlations between the cophenetic-distance vectors of the four molecular-layer dendrograms shown in Fig. 7B.

References

- Albert FW, Bloom JS, Siegel J, Day L, Kruglyak L (2018) Genetics of trans-regulatory variation in gene expression. *eLife* **7**: e35471.
- Blevins WR, Tavella T, Moro SG, Blasco-Moreno B, Closa-Mosquera A, Díez J, et al. (2019) Extensive post-transcriptional buffering of gene expression in the response to severe oxidative stress in baker's yeast. *Sci Rep* **9**: 11005.
- Caudal É, Loegler V, Dutreux F, Vakirlis N, Teyssonnière É, Caradec C, et al. (2024) Pan-transcriptome reveals a large accessory genome contribution to gene expression variation in yeast. *Nat Genet* **56**: 1278–1287.
- Decker CJ, Parker R (1993) A turnover pathway for both stable and unstable mRNAs in yeast: Evidence for a requirement for deadenylation. *Genes Dev* **7**: 1632–1643.
- Di Giammartino DC, Nishida K, Manley JL (2011) Mechanisms and consequences of alternative polyadenylation. *Mol Cell* **43**: 853–866.
- Elkon R, Ugalde AP, Agami R (2013) Alternative cleavage and polyadenylation: Extent, regulation and function. *Nat Rev Genet* **14**: 496–506.
- Esfahani NG, Stein AJ, Akeson S, Tzadikario T, Jain M (2025) Evaluation of Nanopore direct RNA sequencing updates for modification detection. *bioRxiv*: 2025.05.01.651717.
- Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, et al. (2018) Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* **15**: 201–206.
- Geisberg JV, Moqtaderi Z, Fan X, Oszolak F, Struhl K (2014) Global Analysis of mRNA Isoform Half-Lives Reveals Stabilizing and Destabilizing Elements in Yeast. *Cell* **156**: 812–824.
- Geisberg JV, Moqtaderi Z, Struhl K (2023) Condition-specific 3' mRNA isoform half-lives and stability elements in yeast. *Proc Natl Acad Sci* **120**: e2301117120.
- Graber JH, Cantor CR, Mohr SC, Smith TF (1999) Genomic detection of new yeast pre-mRNA 3'-end-processing signals. *Nucleic Acids Res* **27**: 888–894.
- Gruber AJ, Zavolan M (2019) Alternative cleavage and polyadenylation in health and disease. *Nat Rev Genet* **20**: 599–614.
- Guo Z, Sherman F (1995) 3'-End-forming signals of yeast mRNA. *Mol Cell Biol* **15**: 5983.
- Hartigan JA, Hartigan PM (1985) The Dip Test of Unimodality. *Ann Stat* **13**: 70–84.
- Jain M, Abu-Shumays R, Olsen HE, Akeson M (2022) Advances in nanopore direct RNA sequencing. *Nat Methods* **19**: 1160–1164.
- Kim Y, Saville L, O'Neill K, Garant J-M, Liu Y, Haile-Merhu S, et al. (2025) Nanopore direct RNA sequencing of human transcriptomes reveals the complexity of mRNA modifications and crosstalk between regulatory features. *Cell Genomics* **5**: 100872.
- Kita R, Venkataram S, Zhou Y, Fraser HB (2017) High-resolution mapping of cis-regulatory variation in budding yeast. *Proc Natl Acad Sci* **114**: E10736–E10744.
- Legnini I, Alles J, Karaikos N, Ayoub S, Rajewsky N (2019) FLAM-seq: Full-Length mRNA Sequencing Reveals Principles of Poly(A) Tail Length Control. *Nat Methods* **16**: 879–886.
- Li L, Huang K-L, Gao Y, Cui Y, Wang G, Elrod ND, et al. (2021) An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nat Genet* **53**: 994–1005.

- Liao Y, Smyth GK, Shi W (2014) featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinforma Oxf Engl*.
- Liu X, Hoque M, Larochelle M, Lemay J-F, Yurko N, Manley JL, et al. (2017) Comparative analysis of alternative polyadenylation in *S. cerevisiae* and *S. pombe*. *Genome Res* **27**: 1685–1695.
- Loegler V, Friedrich A, Schacherer J (2024) Overview of the *Saccharomyces cerevisiae* population structure through the lens of 3,034 genomes. *G3 GenesGenomesGenetics* **14**: jkae245.
- Loegler V, Thiele P, Teyssonnière E, Tsouris A, Brach G, Cruaud C, et al. (2025) From genotype to phenotype with 1,086 near telomere-to-telomere yeast genomes. *Nature*: 1–10.
- Mak QXC, Wick RR, Holt JM, Wang JR (2023) Polishing De Novo Nanopore Assemblies of Bacteria and Eukaryotes With FMLRC2. *Mol Biol Evol* **40**: msad048.
- Mayr C (2016) Evolution and Biological Roles of Alternative 3'UTRs. *Trends Cell Biol* **26**: 227–237.
- Mayr C, Bartel DP (2009) Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell* **138**: 673–684.
- McManus CJ, May GE, Spealman P, Shteyman A (2014) Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res* **24**: 422–430.
- Mitschka S, Mayr C (2022) Context-specific regulation and function of mRNA alternative polyadenylation. *Nat Rev Mol Cell Biol*.
- Mittleman BE, Pott S, Warland S, Zeng T, Mu Z, Kaur M, et al. (2020) Alternative polyadenylation mediates genetic regulation of gene expression. *eLife* **9**: e57492.
- Montañés JC, Huertas M, Moro SG, Blevins WR, Carmona M, Ayté J, et al. (2022) Native RNA sequencing in fission yeast reveals frequent alternative splicing isoforms. *Genome Res* **32**: 1215–1227.
- Montañés JC, Papadopoulos C, Al-Obaidi S, Szegedi A, Blevins WR, Talló-Parra M, et al. (2026) Evolutionary emergence and preservation of microproteins encoded by upstream ORFs.
- Moqtaderi Z, Geisberg JV, Jin Y, Fan X, Struhl K (2013) Species-specific factors mediate extensive heterogeneity of mRNA 3' ends in yeasts. *Proc Natl Acad Sci* **110**: 11073–11078.
- Muenzner J, Trébulle P, Agostini F, Zauber H, Messner CB, Steger M, et al. (2024) Natural proteome diversity links aneuploidy tolerance to protein turnover. *Nature* **630**: 149–157.
- O'Donnell S, Yue J-X, Saada OA, Agier N, Caradec C, Cokelaer T, et al. (2023) Telomere-to-telomere assemblies of 142 strains characterize the genome structural landscape in *Saccharomyces cerevisiae*. *Nat Genet* **55**: 1390–1399.
- Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, et al. (2010) Comprehensive Polyadenylation Site Maps in Yeast and Human Reveal Pervasive Alternative Polyadenylation. *Cell* **143**: 1018–1029.
- Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomás J, Amorín R, et al. (2024) SQANTI3: Curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nat Methods* **21**: 793–797.
- Passmore LA, Collier J (2022) Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression. *Nat Rev Mol Cell Biol* **23**: 93–106.
- Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**: 417–419.
- Pelechano V, Wei W, Steinmetz LM (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**: 127–131.
- Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, et al. (2018) Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **556**: 339–344.

- Prawer YDJ, Gleeson J, De Paoli-Iseppi R, Clark MB (2023) Pervasive effects of RNA degradation on Nanopore direct RNA sequencing. *NAR Genomics Bioinforma* **5**: lqad060.
- Proudfoot NJ (2011) Ending the message: Poly(A) signals then and now. *Genes Dev* **25**: 1770–1782.
- Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.
- Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD, Hussain S (2019) A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat Commun* **10**: 3359.
- Stroup EK, Ji Z (2024) Delineating yeast cleavage and polyadenylation signals using deep learning. *Genome Res* **34**: 1066–1080.
- Subtelny AO, Eichhorn SW, Chen GR, Sive H, Bartel DP (2014) Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508**: 66–71.
- Tardaguila M, Fuente L de la, Marti C, Pereira C, Pardo-Palacios FJ, Risco H del, et al. (2018) SQANTI: Extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res* **28**: 396–411.
- Teyssonnière EM, Shichino Y, Mito M, Friedrich A, Iwasaki S, Schacherer J (2024a) Translation variation across genetic backgrounds reveals a post-transcriptional buffering signature in yeast. *Nucleic Acids Res* **52**: 2434–2445.
- Teyssonnière EM, Trébulle P, Muenzner J, Loegler V, Ludwig D, Amari F, et al. (2024b) Species-wide quantitative transcriptomes and proteomes reveal distinct genetic control of gene expression variation in yeast. *Proc Natl Acad Sci* **121**: e2319211121.
- Tian B, Manley JL (2017) Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* **18**: 18–30.
- Tsouris A, Brach G, Friedrich A, Hou J, Schacherer J (2024a) Diallel panel reveals a significant impact of low-frequency genetic variants on gene expression variation in yeast. *Mol Syst Biol* **20**: 362–373.
- Tsouris A, Brach G, Schacherer J, Hou J (2024b) Non-additive genetic components contribute significantly to population-wide gene expression variation. *Cell Genomics* **4**.
- Tucker M, Valencia-Sanchez MA, Staples RR, Chen J, Denis CL, Parker R (2001) The Transcription Factor Associated Ccr4 and Caf1 Proteins Are Components of the Major Cytoplasmic mRNA Deadenylase in *Saccharomyces cerevisiae*. *Cell* **104**: 377–386.
- Tudek A, Krawczyk PS, Mroczek S, Tomecki R, Turtola M, Matylla-Kulińska K, et al. (2021) Global view on the metabolism of RNA poly(A) tails in yeast *Saccharomyces cerevisiae*. *Nat Commun* **12**: 4951.
- Wahle E, Winkler GS (2013) RNA decay machines: Deadenylation by the Ccr4–Not and Pan2–Pan3 complexes. *Biochim Biophys Acta BBA - Gene Regul Mech* **1829**: 561–570.
- Wang JR, Holt J, McMillan L, Jones CD (2018) FMLRC: Hybrid Long Read Error Correction Using an FM-index. *BMC Bioinformatics* **19**: 50.
- Webster MW, Chen Y-H, Stowell JAW, Alhusaini N, Sweet T, Graveley BR, et al. (2018) mRNA Deadenylation Is Coupled to Translation Rates by the Differential Activities of Ccr4–Not Nucleases. *Mol Cell* **70**: 1089–1100.e8.
- Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, et al. (2019) Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods* **16**: 1297–1305.
- Wyman D, Mortazavi A (2019) TranscriptClean: Variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. *Bioinformatics* **35**: 340–342.

Environmental context shapes the cross-omic consequences of natural genetic variation in yeast

| Abstract

The genotype–phenotype relationship is generally studied at the level of organismal traits, and these traits emerge from variation in intermediary molecular layers such as transcriptome and proteome. Understanding how genetic effects act across these layers and environments is therefore essential to explain phenotypic variation. Here, we generated matched quantitative transcriptomes and proteomes across glucose, galactose, elevated temperature and stationary phase in a species-wide panel of *Saccharomyces cerevisiae* isolates, alongside growth phenotypes. The resource comprises 2,721 transcriptomes and 2,668 proteomes, with same-culture measurements for 555 isolates. The four environments produced distinct transcriptomic and proteomic profiles, with additional variation associated with population structure. Genotype-by-environment interactions were widespread, and most molecular quantitative trait loci (QTL) were private to each environment. The genetic bases of transcript and protein abundance overlapped only modestly, yet the corresponding protein-response effects usually had the same direction as the transcript effects, although they were smaller. Their transmission varied across environments and was weakest in stationary phase. At a galactose-specific *GAL2* locus, 23 of 26 measured protein responses followed the direction of a 44-gene transcript response signature enriched for vacuolar acidification and the tricarboxylic acid cycle. Together, our results establish a species-wide, cross-environment multi-omic resource and show that environment determines both the regulatory targets of genetic variants and the extent to which transcript effects persist at protein level.

Introduction

Genetic variation contributes to phenotypic diversity among individuals, from molecular and cellular traits to complex organismal phenotypes such as growth, metabolism, drug response and disease susceptibility. At the population level, phenotypic variance can be partitioned into genetic, environmental and genotype-by-environment components, together with residual variation (Visscher et al., 2008). Understanding how these components act through molecular layers to produce phenotypes is a central question in biology (Brem et al., 2002; Albert & Kruglyak, 2015). Yeast mapping studies showed that gene-product abundance is shaped by local variants and distant regulatory hotspots at both transcript and protein levels (Yvert et al., 2003; Albert et al., 2014, 2018). Organismal traits emerge through molecular changes within cells, and transcript and protein abundance provide complementary views of how genetic variation shapes cellular state (Parts et al., 2014; Battle et al., 2015). These layers are closely connected but not equivalent because protein abundance additionally integrates translation and degradation (Vogel & Marcotte, 2012; Liu et al., 2016; Buccitelli & Selbach, 2020). Foundational yeast studies showed that mRNA abundance does not fully predict protein abundance and that integrating both layers exposes regulation missed by either alone (Gygi et al., 1999; Ideker et al., 2001). Genetic effects can be preserved, attenuated or reshaped between transcript and protein (Foss et al., 2007; Wu et al., 2013; Parts et al., 2014; Battle et al., 2015; Chick et al., 2016). Jointly measuring both layers across genetically diverse populations is therefore essential for connecting genetic variation to phenotypic diversity.

Environmental conditions add a further dimension to genetic regulation. Cells extensively remodel gene expression in response to their surroundings, changing the molecular context in which genetic effects are expressed (Gasch et al., 2000; Causton et al., 2001). The effects of individual variants can consequently differ among environments, producing genotype-by-environment (GxE) interactions (Smith & Kruglyak, 2008; Hodgins-Davis et al., 2012; Gagneur et al., 2013). Regulatory associations vary across human tissues and after immune stimulation, demonstrating the importance of biological context at population scale (Fairfax et al., 2014; Lee et al., 2014; GTEx Consortium, 2020). However, how species-wide molecular GxE is transmitted from the transcriptome to the proteome across environments remains unresolved (Smith & Kruglyak, 2008; Teyssonnière et al., 2024).

Previous population-scale studies of genetic effects on transcript and protein abundance profiled each genotype in one reference condition (Foss et al., 2007; Wu et al., 2013; Parts et al., 2014; Battle et al., 2015; Chick et al., 2016; Teyssonnière et al., 2024). Such designs cannot reveal whether the targets of regulatory loci change across environments or whether environment-dependent transcript effects remain detectable in the proteome. Answering these questions requires measuring the same genotypes across environments and following their effects from transcript to protein (Smith & Kruglyak, 2008; Gagneur et al., 2013).

Foundational population-genomic surveys resolved extensive structure among wild and domesticated *Saccharomyces cerevisiae* lineages (Liti et al., 2009; Schacherer et al.,

2009). Yeast genomics has since expanded from a single reference strain to thousands of isolates, revealing extensive genetic, ecological and geographical diversity across the species (Brach & Schacherer, 2026). Species-wide genomic surveys of 1,011 and 3,034 isolates further resolved lineages shaped by geography, ecology and human activity (Peter et al., 2018; Loegler et al., 2024). Near telomere-to-telomere assemblies of 1,086 isolates resolved structural and pangenome variation across the species (Loegler et al., 2025). Within this population framework, a pan-transcriptomic analysis revealed a disproportionate contribution of accessory genes to expression variation and identified local and distant expression quantitative trait loci (eQTL) affecting both core and accessory genes (Caudal et al., 2024). A complementary proteomic analysis showed that transcript-protein correlations across isolates were modest and that only a small fraction of eQTL and protein quantitative trait loci (pQTL) overlapped (Teyssonnière et al., 2024). Together, these studies established a species-wide genomic framework and molecular resources in glucose. However, the transcriptomic and proteomic resources were generated independently and did not provide matched measurements from the same cultures or across environmental states (Caudal et al., 2024; Teyssonnière et al., 2024).

Here, we generated a species-scale dataset of matched transcriptomes and proteomes from the same cultures across glucose, galactose, elevated temperature (37°C) and early stationary phase, alongside growth phenotypes. These environments span well-characterized transitions in carbon utilization, thermal stress and growth state, each accompanied by extensive remodelling of gene expression and protein abundance (Gasch et al., 2000; Ideker et al., 2001; Marguerat et al., 2012). This design enabled direct comparisons across isolates, environments and molecular layers, providing a more complete view of the genotype–phenotype relationship. We resolved lineage-specific environmental responses, mapped eQTL and pQTL within each environment, identified transcript-response QTL (gxe(e)QTL) across environmental transitions and tested their corresponding protein-response effects (gxe(p)QTL). By moving population-scale molecular genetics beyond a single reference condition, this resource establishes a framework for understanding how genetic diversity is converted into environment-specific molecular phenotypes.

Results

A population-scale, cross-environment, matched multi-omic resource

To build a population-scale, cross-environment resource, we profiled 587 distinct *S. cerevisiae* isolates across overlapping transcriptomic and proteomic panels. Cultures were split at harvest so that paired transcript and protein measurements came from the same culture and physiological state (Fig. 1a). The collection captured the ecological and geographical diversity of the species, spanning 70 countries and including wild and domesticated subpopulations distributed across 34 clades and five superclades (Fig. 1b-c). We quantified transcript and protein abundance under glucose, galactose, temperature (37°C) and stationary-phase conditions, together with growth phenotypes in the first three conditions. The transcriptomic dataset contained 2,721 samples from 572 isolates, including 659 in glucose, 767 in galactose, 660 in temperature and 635 in stationary phase. The proteomic dataset contained 2,668 samples, including 682 in glucose, 691 in galactose, 649 in temperature and 646 in stationary phase, from 576 isolates, including 15 not present in the transcriptomic panel. In total, 555 isolates had matched same-culture transcriptomic and proteomic data retained for cross-layer analyses (Fig. 1d; Supplementary Table 1).

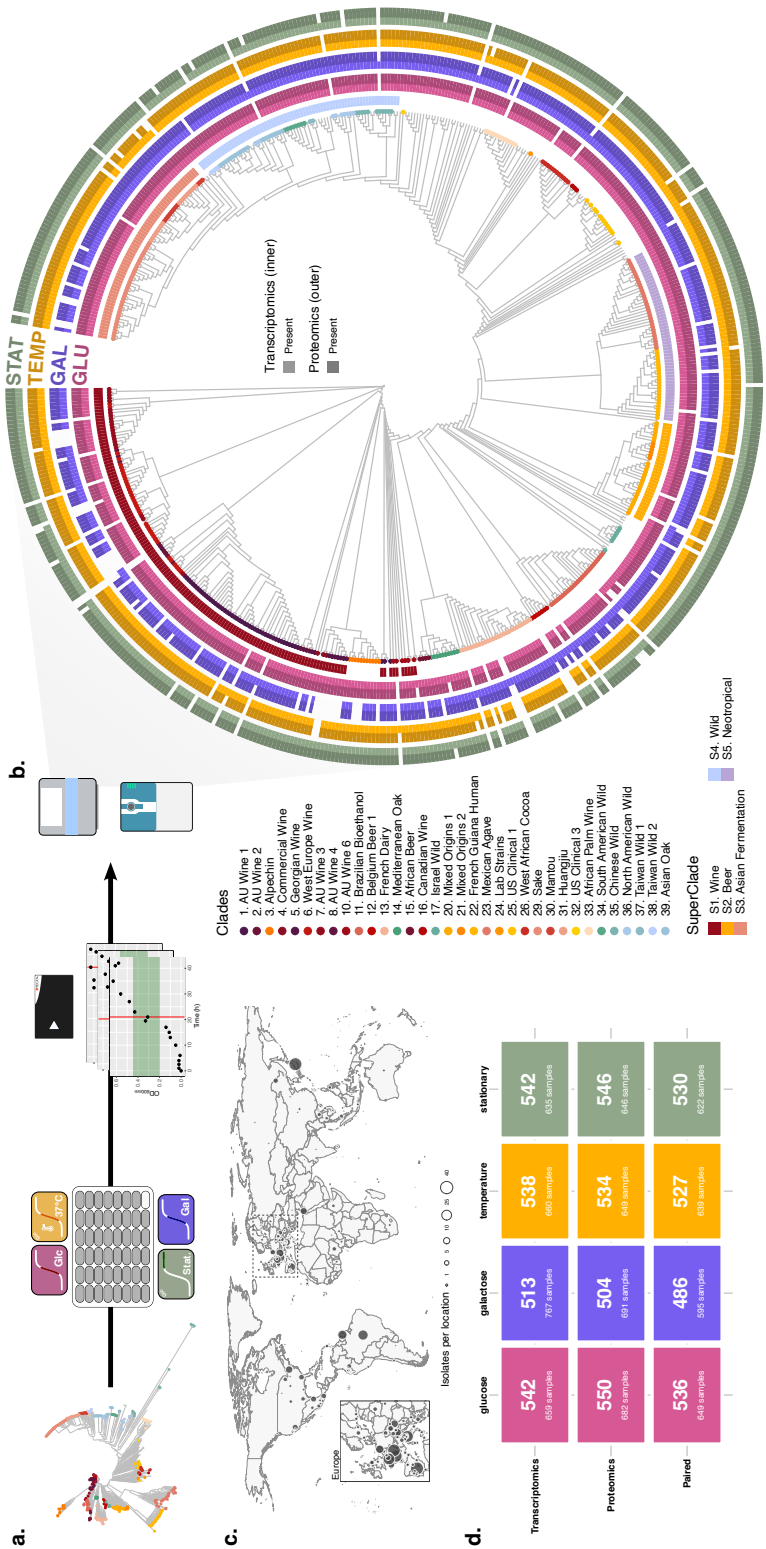
The resulting molecular profiles showed strong agreement across replicates, sequencing batches and independent datasets. RNA libraries contained a median of 5.1 million mapped reads per sample and yielded 13.2 billion uniquely mapped reads in total (Fig. S1a). Sample identities were verified against reference genotypes, and expression was quantified with a pangenome-augmented annotation spanning rare accessory to core genes (Loegler et al., 2025, Fig. S1b). Replicate coverage was extensive, with 381 strain-environment groups sampled more than once: 76 in glucose, 183 in galactose, 67 at 37°C and 55 in stationary phase (Fig. S1c). Together, these groups produced 2,479 RNA replicate pairs with a mean correlation of 0.96, showing the robustness of our dataset (Fig. S1d). Across 500 isolates also profiled by Caudal et al. (2024), glucose expression remained highly concordant despite a slightly modified experimental protocol to allow for proteomics measurements (median Pearson $r = 0.946$; Fig. S1e), supporting reproducibility across studies. An independent validation set and repeated sequencing of the control strain (XTRA_FGV, Mexican Agave clade) showed similar agreement (Fig. S1f). Same-strain correlations across culture plates remained high, with condition medians of 0.958 to 0.962, although within-plate pairs were modestly more concordant, with medians of 0.978 to 0.983 (Fig. S1g).

At the protein level, median coefficients of variation were 6.0% for technical controls and 9.6% for within-strain replicates, compared with 25.6% across isolates (Fig. S2a). Within conditions, same-strain replicate proteomes had a median Pearson correlation of 0.980, compared with 0.961 for an equal-sized sample of randomly paired isolates (Fig. S2b). Across conditions, 4,308 to 4,399 proteins were detected, of which 2,294 to 2,659 met the 75% sample-completeness threshold that we set (Fig. S2c). Transcript-protein agreement was high across genes within isolates but substantially weaker for the same gene across isolates. Across genes within an isolate, median

correlations ranged from 0.49 in stationary phase to 0.68 in galactose. Within individual genes across isolates, the corresponding medians ranged from 0.08 in stationary phase to 0.22 at 37°C (Fig. S2d). The glucose median of 0.20 was close to the 0.165 reported in the larger glucose-grown population (Teyssonnière et al., 2024). Thus, differences among genes were preserved between the transcriptome and proteome much more strongly than population variation within the same gene. Together, the extensive replication, cross-study concordance and scale-dependent transcript-protein agreement establish the robustness of the matched dataset across environments.

Figure 1 | Matched transcriptomes and proteomes across yeast diversity and environments.

(a.) Experimental workflow showing selection of the natural-isolate panel, four environmental states, matched transcript and protein collection and growth phenotypes where available. (b.) Population-tree placement of profiled isolates, with clade, superclade and condition-specific data availability. (c.) Geographic origin of the profiled isolates, including a Europe inset. Isolates without usable coordinates are shown at 0 degrees longitude and 0 degrees latitude. (d.) Number of isolates and samples represented in the transcriptome, proteome and paired same-culture dataset in each environment. Detailed transcriptomic and proteomic quality controls are shown in Supplementary Figures 1 and 2.



Environmental and population structure shape the multi-omic landscape

Environment produced the dominant structure in both transcriptome and proteome t-SNE maps, although the two molecular layers formed distinct clustering patterns (Fig. 2a). Samples grown in glucose at 30°C and 37°C occupied overlapping regions of each t-SNE map, consistent with a comparatively mild global response to heat stress. Galactose and stationary-phase samples instead formed distinct clusters. Clade membership produced local structure within each state but did not replace environment as the dominant axis of variation. An environmental stress response (ESR) score, calculated as the mean standardized abundance of genes in the induced Gasch ESR programme minus that of genes in its repressed PAC and ribosomal-protein arms, increased across the temperature and stationary-phase regions (Gasch et al., 2000). A GAL programme score, calculated from seven core galactose-pathway genes, specifically marked the galactose state (Fig. S3a, ESR and GAL maps). Thus, sample separation reflected coordinated biological responses to the growth conditions, including ESR activation and GAL induction.

To integrate the same samples and features across molecular layers, we fitted MOFA to 2,496 matched isolate-condition samples from 554 strains, using transcript and protein measurements for 2,664 genes and growth traits where available (see Methods). The model resolved 11 factors that explained 77.0% of mRNA variance, 40.6% of protein variance and 9.1% of growth variance (Argelaguet et al., 2018, 2020, Fig. 2b). Projection of the first five factor scores onto the common transcriptome t-SNE space related these axes to the environmental states (Fig. S3b). Correlations with broad biological gene-set scores connected the factors to growth, ribosome biogenesis, carbon metabolism, respiration and stress responses in both matched transcripts and proteins (Fig. S4a). Gene-set enrichment along the negative and positive factor-loading poles, evaluated in matched transcripts, matched proteins and the full transcriptome, further established the biological identity of each factor (Fig. S4b). Factor 1 accounted for 45.4% of mRNA and 10.5% of protein variance, whereas Factor 2 accounted for 5.5% and 14.8%, respectively, showing that the main axes contributed differently to the two molecular layers. Environment explained 84% and 85% of variation in Factors 1 and 2, whereas most subsequent factors varied predominantly within environments. Factor 3 provided the clearest layer-specific example, explaining 17.0% of mRNA variance but only 0.15% of protein and 0.01% of growth variance. Factors 1 and 2 provided the clearest separation among environments, globally recapitulating ESR activation and GAL induction. Factors 1-8 were each recovered in at least 15 of 16 retrainings on random 85% isolate subsets, whereas Factors 9-11 were recovered in 14, 13 and 6 retrainings, respectively (Fig. S4c). The first three factors closely matched axes from a separate sensitivity model fitted to all retained transcripts rather than only the 2,664 genes shared with proteomics, with absolute correlations of 0.96 to 0.99 (Fig. S4d). Together, these analyses distinguished stable environmental axes from molecular signatures that were specific to one regulatory layer or varied among strains within the same environment.

Population structure shaped expression variation within each environment. Median factor scores for each clade and environment showed how individual clades differed from the overall population response (Fig. 2c). Although environment explained 84.6% of Factor 2 variation across the full dataset (Fig. 2b), clade membership explained 53.4% of variation along this galactose-associated axis within the galactose condition (Fig. 2d). This effect was driven largely by French Dairy isolates: removing this clade reduced the adjusted R^2 from 53.4% to 28.5%. French Dairy isolates form a closely related group recovered from French cheeses and carry a dairy-adapted GAL system that is constitutively expressed in glucose (Legras et al., 2018; Peter et al., 2018; Caudal et al., 2024). In galactose, these isolates differed by -2.74 standard deviations from the remaining population on Factor 2. Factor 5 showed a broader lineage-associated architecture. Overall, 63.2% of its variation occurred within rather than between environments, and clade membership explained 18.0% to 45.4% of its within-environment variation, with the strongest contribution in stationary phase.

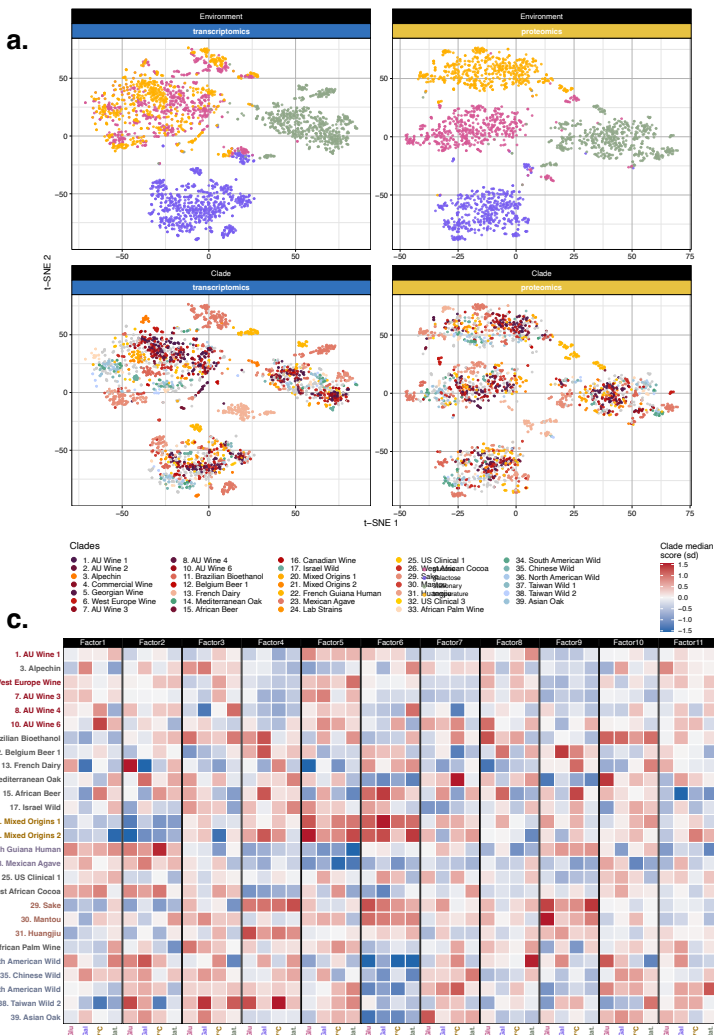
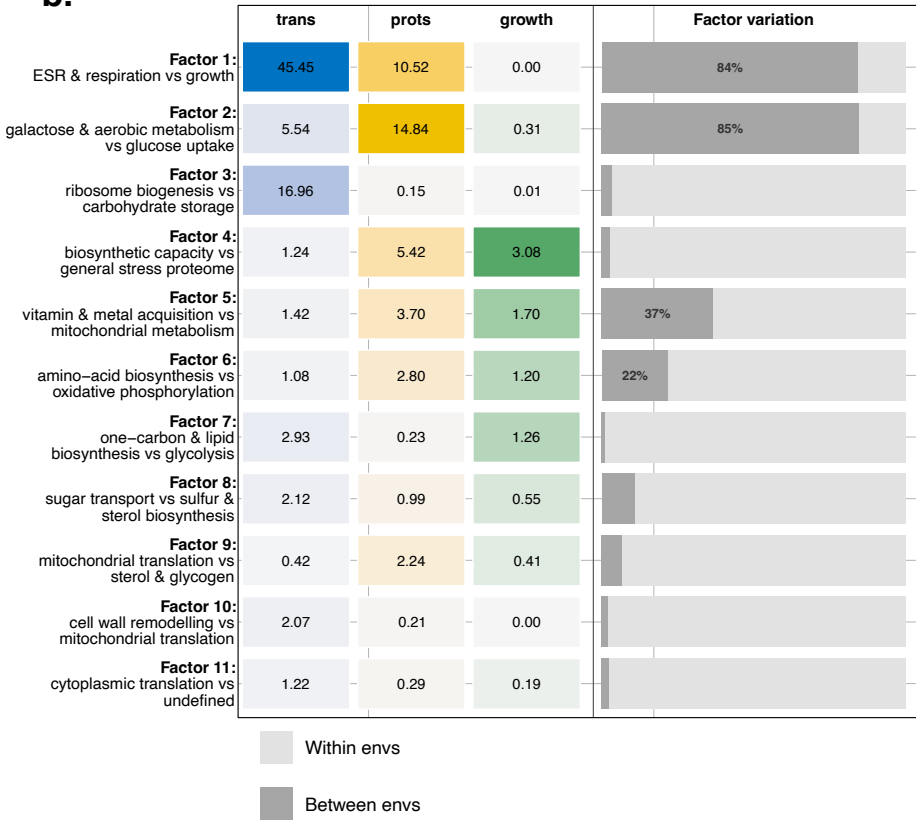


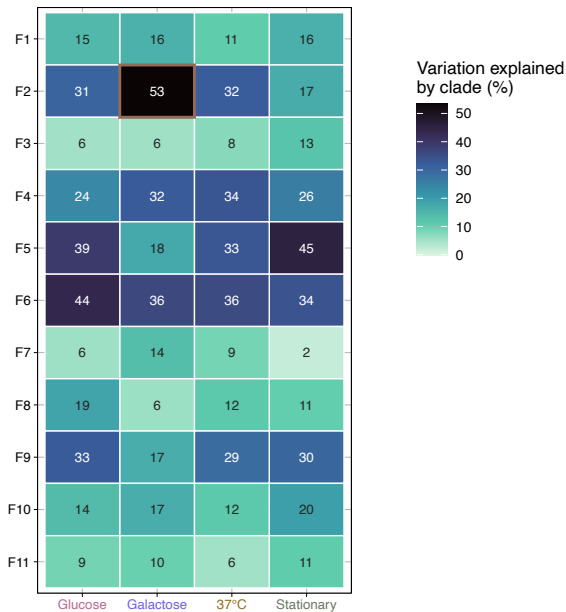
Figure 2 | Environment dominates multi-omic structure while population history shapes variation within states.

(a.) Independent transcriptome and proteome t-SNE maps by environment and population clade. Coordinates were calculated independently for the two molecular layers; the protein vertical axis was reflected for visual alignment only. (b.) MOFA factor atlas showing variance explained in transcript, protein and growth data, and the partition of factor scores between and within environments. Factor labels summarize the gene-set correlations and factor-pole enrichments shown in Fig. S4a-b. (c.) Median standardized MOFA scores by clade and environment. (d.) Adjusted R^2 for clade membership within each environment. Clade membership explained 53.4% of Factor 2 variation in galactose, decreasing to 28.5% after French Dairy was excluded. All associations passed BH FDR < 0.05 except Factor 7 in stationary phase.

b.



d.



Environmental responses are coordinated but lineage dependent

To determine how environmental responses at transcript level were reflected at protein level, we calculated within-isolate changes in abundance for galactose, 37°C and stationary phase relative to glucose, then averaged each gene's response across isolates with matched measurements. We classified genes as RNA responsive when their mean transcript change was at least 0.1 log₂ units in absolute magnitude and passed a BH-adjusted $P < 0.05$. Among these genes, transcript-protein response correlations were 0.634 in galactose, 0.225 at 37°C and 0.474 in stationary phase, indicating weakest agreement in the relative ordering of gene responses at 37°C. The slope of protein on transcript response, fitted through zero, captured response magnitude instead: a one-unit population-average transcript change corresponded to 0.561 protein units in galactose, 0.169 at 37°C and 0.171 in stationary phase (Fig. 3a). Although the 37°C and stationary-phase slopes were nearly identical, they described response distributions with different dynamic ranges. Median absolute transcript and protein changes were 0.246 and 0.104 at 37°C, compared with 0.865 and 0.213 in stationary phase. Consistent with this difference, 68.5% of RNA-responsive genes were buffered or attenuated in stationary phase, compared with 53.5% at 37°C. Stationary phase therefore showed a clear quantitative decoupling between the transcriptome and proteome: extensive transcriptome rewiring was compressed into a substantially smaller proteome response. By contrast, the 37°C response showed weaker gene-wise concordance but a less pronounced difference in response magnitude.

Relative to glucose, transcriptome rewiring involved 2,521 differentially expressed genes in stationary phase, compared with 472 in galactose and 230 at 37°C, whereas proteome rewiring involved similar numbers of differentially abundant proteins in galactose and stationary phase, 325 and 312, respectively. At 37°C, 90 transcripts decreased and 140 increased, while 54 proteins decreased and 30 increased (Fig. S5a). In galactose, 144 transcripts decreased and 328 increased, while 152 proteins decreased and 173 increased. In stationary phase, 958 transcripts decreased and 1,563 increased, while 206 proteins decreased and 106 increased. Stationary phase therefore caused extensive transcriptome rewiring and broad proteome remodelling, although the number of affected proteins was similar to galactose. The corresponding transcript enrichments distinguished carbon metabolism, respiration, translation and membrane-associated processes among environments (Fig. S5b). Population structure also influenced expression within each environment (Fig. S5c-d). Lineage-associated effects were detected in both molecular layers but were more extensive in the transcriptome, and the extent, direction and gene composition of these signatures changed with environment. Sake isolates showed large transcript signatures in every environment, ranging from 360 to 464 genes. Alpechin yeasts showed a qualitatively different temperature state, with 193 genes expressed at higher levels and 64 at lower levels than in the remaining population, whereas their signatures in the other three environments were dominated by lower expression. French Guiana Human isolates showed the largest stationary-phase profile, involving 541 genes.

French Dairy contributed 509 differentially expressed genes in galactose, among the largest environment-specific profiles, and showed elevated core GAL programme activity outside galactose but markedly reduced induction from glucose to galactose. At the four environmental endpoints, its core GAL programme differed from that of other isolates in every condition (two-sided Wilcoxon rank-sum tests, Bonferroni-adjusted $P \leq 1.76 \times 10^{-12}$; Fig. 3b). The programme was markedly elevated in glucose, at 37°C and in stationary phase, but slightly lower in galactose. Endpoint expression, however, does not measure induction from the glucose state. We therefore calculated within-strain condition-minus-glucose responses. Considering the three responses jointly in the 390 isolates measured in all four environments, a MANOVA provided strong evidence of lineage-level GxE (18 French Dairy and 372 other isolates; Pillai trace = 0.871, $P = 1.9 \times 10^{-171}$). We then tested each response separately using two-sided Wilcoxon rank-sum tests, with Bonferroni correction across the three contrasts. Mean galactose responses were 0.38 in French Dairy and 2.31 in other lineages, a difference of -1.93 standardized programme units (adjusted $P = 1.29 \times 10^{-13}$). Responses at 37°C and in stationary phase were modestly higher in French Dairy (adjusted $P = 0.00688$ and 1.04×10^{-4} , respectively). Together, these analyses recover the known constitutive GAL state and show that French Dairy divergence encompasses both its glucose baseline and its response to other environments.

A stringent clade-resolved analysis identified 2,221 lineage-level GxE events across 891 genes, with events in 22 of the 24 tested clades (Fig. 3c; Supplementary Data 1). We first screened genes with a clade-by-environment interaction model and then resolved the responsible clade and environmental contrast in the 370 isolates measured in all four environments. For each clade represented by at least five complete isolates, we compared its paired condition-minus-glucose response on the DESeq2 variance-stabilized expression scale (VST) with the response of all other clades. We retained clade-gene-environment events with clade-resolved FDR < 0.05 and an absolute response difference greater than one VST unit, provided that the gene also passed the initial interaction screen. The VST criterion was used as an effect-size threshold on the transformed expression scale rather than as a literal fold-change cutoff. These comprised 720 events involving 407 genes in galactose, 384 events involving 286 genes at 37°C and 1,117 events involving 698 genes in stationary phase. Across the 891 genes, 525 were detected in one response contrast, 232 in two and 134 in all three. Representative *HXK2* and *THI5* reaction norms showed Mexican Agave-specific stationary-phase and Sake-specific temperature responses, respectively (Fig. S6a-b). The events were unevenly distributed among populations and across the genome (Fig. 3d; Fig. S6c). South American Wild isolates accounted for 196 of the 384 temperature events, of which 189 showed a more negative response than other clades: 140 genes were repressed more strongly, and 49 switched from activation in other clades to repression in South American Wild (Fig. S6d). Thus, analysing expression at 37°C alone concealed a much broader divergence in how South American Wild isolates responded relative to glucose. Lineage-dependent environmental responses therefore cannot be inferred from single-condition differential expression, highlighting the need to characterize molecular variation across multiple environments.

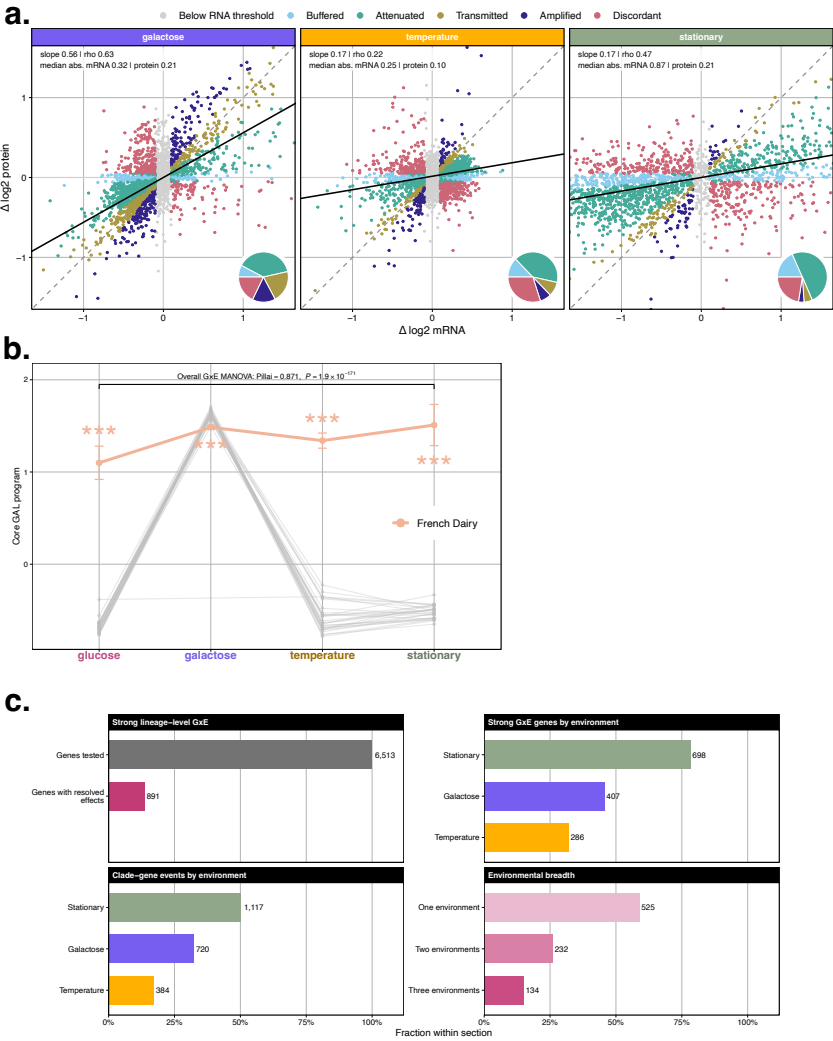
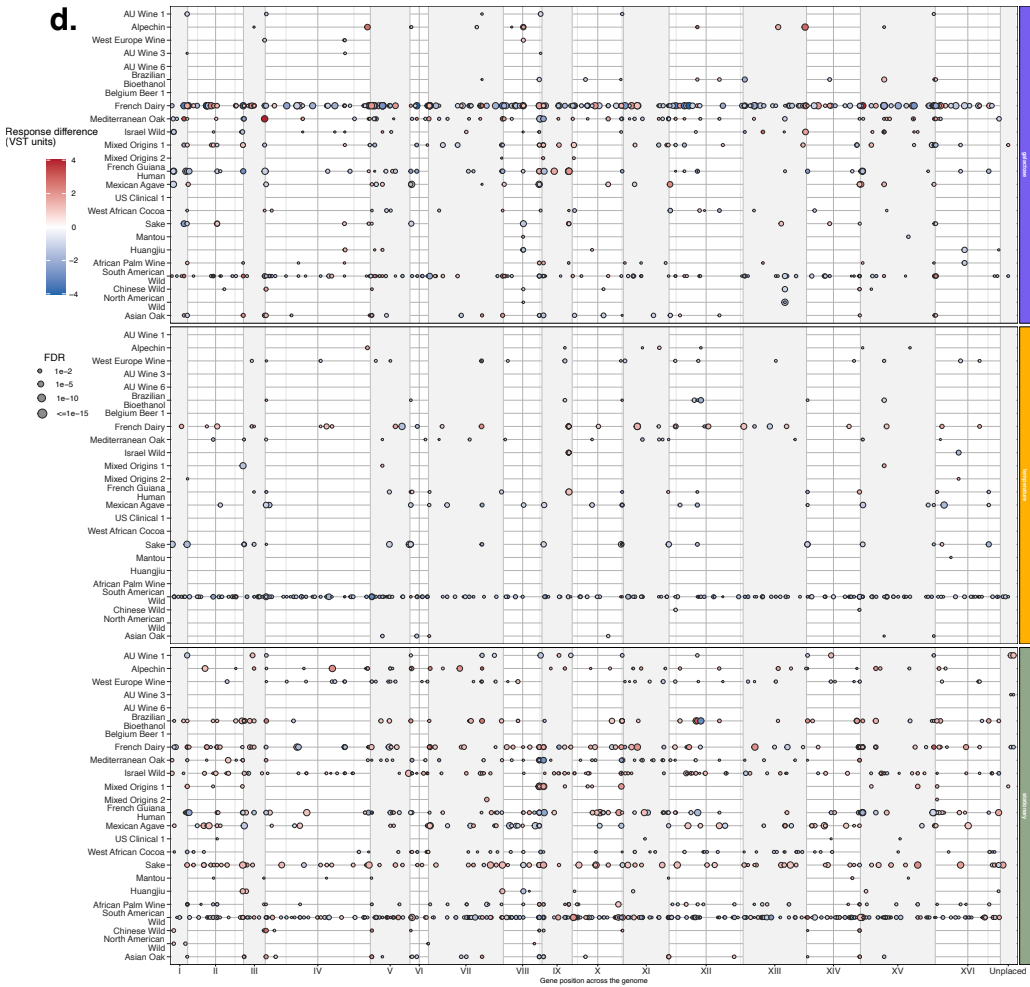


Figure 3 | Transcript and protein responses differ across environments and lineages.

(a.) Population-average transcript and protein changes for galactose, temperature and stationary phase relative to glucose. (b.) Mean core GAL programme across four environments for clades represented by at least three isolates. French Dairy was compared with the other isolates using Bonferroni-adjusted two-sided Wilcoxon tests; a joint MANOVA tested the three condition-minus-glucose responses (18 French Dairy and 372 other isolates; Pillai trace = 0.871, $P = 1.9 \times 10^{-171}$). (c.) Hierarchy from 6,513 tested genes to globally significant GxE genes and 891 genes with strong clade-resolved effects, including their distribution across environments. (d.) Genome-wide positions of strong clade-gene-environment events.



Regulatory architecture is restricted by environment and molecular layer

Genome-wide association studies across the four environments detected 25,179 transcript and 7,487 protein locus-target pairs after linkage-disequilibrium grouping and exclusion of variants within 20 kb of telomeres (Fig. 4a; Supplementary Data 1). The transcriptomics GWAS dataset included 542, 513, 538 and 540 isolates in glucose, galactose, elevated temperature and stationary phase, respectively; the corresponding proteomics dataset included 513, 485, 524 and 511 isolates. The transcriptome GWAS used 95,804 genetic markers, comprising 88,971 SNPs, 1,468 InDels, 767 structural variants (SVs) and 4,598 CNV dosage markers. The proteome GWAS used 95,829 markers, comprising 88,971 SNPs, 1,499 InDels, 769 SVs and 4,590 CNV dosage markers. The 3.36-fold difference in raw pair counts partly reflected the broader transcriptomic phenotype set: across the union of traits tested in the four environments, the scans included 6,816 transcripts and 2,714 proteins. This corresponded to 3.69 distinct locus-target pairs per assayed transcript and 2.76 per assayed protein, a 1.34-fold difference that should also be interpreted in light of lower proteomic coverage and power. Diagonal concentrations in the genome maps represented local associations, whereas vertical concentrations marked regulatory hotspots connecting individual loci to multiple distant targets. The maps suggested that genetic associations were more concentrated at high-reach transcript hotspots, although the raw cross-layer comparison was affected by differences in phenotype coverage. Among positioned pairs, 3,337 RNA associations were local and 21,775 were distant, whereas 382 protein associations were local and 7,091 were distant. Using a 20-kb gene window, 13.3% of RNA pairs and 5.1% of protein pairs were local. This proportion varied by variant class: 41.2% of RNA CNV pairs and 11.9% of RNA SNP pairs were local, whereas more than 93% of protein pairs were distant for every variant class (Fig. S7a). Local associations had larger standardized effects than distant associations in every environment in both molecular layers (Bonferroni-adjusted Wilcoxon $P \leq 8.0 \times 10^{-4}$; Fig. 4b). Variant class was also associated with effect magnitude in every environment and molecular layer (Bonferroni-adjusted Kruskal-Wallis $P \leq 1.16 \times 10^{-7}$). SNPs or InDels had the largest median absolute effect in all eight layer-environment comparisons, whereas SVs or CNVs had the smallest (Fig. S7b).

The genetic bases of expression were largely environment specific, as most genetic associations were detected in only one environment. Empirical random-allocation thresholds defined a regulatory hotspot as an LD-defined locus associated with at least five transcript targets or four protein targets in one environment (see Methods). We refer to each threshold-passing locus-environment combination as a hotspot instance and to its unique genomic locus as a hotspot position. Hotspot reach varied sharply among environments and was generally broader for transcript than protein abundance (Fig. S8a). Across the four environments, 855 unique transcript hotspot positions produced 1,367 condition-specific hotspot instances, whereas 225 protein hotspot positions produced 270 instances. At 37°C, the two largest transcript hotspot instances reached 81 and 60 targets and were led by variants 149 bp apart that both fell within *ECM9*; 25 targets were shared. To determine whether the broader cross-layer difference was explained by the larger transcript phenotype set, we restricted each

environment to genes tested in both molecular layers and repeatedly matched the number of significant locus-target associations. Under a common five-target hotspot cutoff, the median reach of the ten largest transcript loci exceeded that of protein loci in all four environments. Transcript hotspot density was also greater in galactose, at 37°C and in stationary phase, whereas protein hotspot density was greater in glucose (Fig. S8d). Target-gene coverage and discovery count therefore did not fully explain the concentration of transcript associations at regulatory hotspots, although differences in sample composition and measurement precision remain potential contributors. To test positional sharing between molecular layers, we paired transcript and protein hotspot instances from the same environment when their lead positions were no more than 5 kb apart, irrespective of target identity. Depending on the environment, 6.3% to 12.7% of transcript hotspot instances and 26.7% to 47.4% of protein hotspot instances had a positional counterpart in the other layer (Fig. S8b). These loci connected to distinct functional target sets in each environment. These included carbohydrate and energy metabolism in glucose (74 targets at *PRP16*), mitochondrial translation in galactose and at elevated temperature (71 targets near *YIR017W-A* and 66 at *ECM9*, respectively), and cytoplasmic translation in stationary phase (41 targets at *PFK26*; Fig. S8c). Across LD-defined regulatory loci paired with the same target gene, 21,729 of 25,179 RNA pairs (86.3%) and 7,172 of 7,487 protein pairs (95.8%) occurred in one environment, whereas 985 RNA pairs (3.9%) and 50 protein pairs (0.7%) occurred in all four (Fig. 4c). Hotspot positions recurred across at least two environments more often than the same locus-target association: 33.0% compared with 13.7% for transcript abundance and 16.9% compared with 4.2% for protein abundance (Fig. 4c; Fig. S9a-c). Thus, regulatory positions could recur across environments while changing their target genes.

Effect magnitude increased with the environmental breadth of the same locus-target association (Fig. S9d). Between one and four environments, median RNA effects increased from 0.315 to 0.413 for local associations and from 0.307 to 0.416 for distant associations. Protein effects increased from 0.318 to 0.434 locally and from 0.300 to 0.399 distantly. This trend is partly expected because associations detected in several environments must repeatedly pass the discovery threshold. By contrast, hotspots affecting more genes did not have larger effects. Hotspot target number was weakly negatively correlated with median effect in RNA ($\rho = -0.092$, $P = 6.3 \times 10^{-4}$) and was unrelated in protein ($\rho = 0.073$, $P = 0.23$; Fig. S9e).

Exact SNP-target associations rarely overlapped between the transcriptome and proteome, but shared associations had concordant effects. Using exact SNP-target matches, 2.1% to 5.2% of pQTL matched an eQTL in the same environment; the same definition yielded 2.8% (15 of 528) in glucose in the earlier species-wide analysis (Fig. 4d; Teyssonnière et al., 2024). Most significant associations therefore differed between the two layers. Searching across all four transcriptomic environments modestly increased the fraction of pQTL with a matching eQTL to 5.1% in glucose (119 of 2,326), 6.7% in galactose (102 of 1,528), 4.1% at 37°C (86 of 2,117) and 5.0% in stationary phase (73 of 1,461). The largest relative increase occurred in stationary phase, from 2.1% to 5.0%, but no condition exceeded 6.7%. Environmental turnover of eQTL therefore explained only part of the limited overlap between molecular layers. Sharing

was concentrated among local associations: 17.7% to 26.8% of local pQTL, but only 0.6% to 3.5% of distant pQTL, matched an eQTL in the same environment (Fig. S10a). Relaxing exact identity to LD $r^2 \geq 0.2$ increased the pooled shared-locus fraction only from 11.0% to 12.2% among gene-environment combinations with QTL in both layers (Fig. S10b). When local and distant exact shared associations were combined within each environment, signed allele-aligned Spearman correlations ranged from 0.758 to 0.964, and 97.9% to 100% had the same direction. Local and distant signed effects were positively correlated in every condition with sufficient shared associations for estimation (Fig. S10c). Thus, the significant associations detected for transcript and protein abundance were largely distinct, whereas associations detected in both layers almost uniformly acted in the same direction.

Among these cross-layer associations, the SNP rs_16_375375_T_A at the chromosome-XVI *SSU1-GLR1* interval was associated with *GLR1* transcript and protein abundance in every condition (Fig. S11a). This interval contains a breakpoint of the adaptive VIII-XVI *SSU1-R* translocation (Pérez-Ortín et al., 2002). The variant lies 206 bp downstream of *SSU1* and in the proximal *GLR1* promoter, 127 bp upstream of the start codon and 93 or 102 bp upstream of the two transcript starts annotated from YPD and galactose data, respectively. Reference-sequence inspection placed it between a CRE/Yap response-element pair and the TATA-like sequence, rather than directly within these motifs. Zoomed association scans around the interval identified the same peak, led by rs_16_375375_T_A, in transcript and protein data from all four environments (Fig. S11b). Transcript effects ranged from 0.699 to 0.787 standardized phenotype units and protein effects from 0.602 to 0.742; all eight associations exceeded their trait-specific permutation thresholds. The focal allele and the catalogued VIII-XVI *SSU1* translocation occupied partly overlapping but distinct subsets of isolates across the transcriptome and proteome expression maps (Fig. S11c). Only 24 of 217 focal-allele carriers also carried the translocation, while 10 of the 34 translocation carriers lacked the focal allele, resulting in weak marker-dosage correlation ($r^2 = 0.029$). Across all eight layer-environment combinations, the SNP effect remained positive and similar after adjustment for the translocation, restriction to translocation non-carriers, adjustment for the original technical covariates, exclusion of chromosome-XVI aneuploids and estimation within clades. Leaving out each major clade in turn yielded effects of 0.649 to 0.922 standardized units across the eight contexts. Together, these analyses distinguish the persistent *GLR1* association from the nearby *SSU1* rearrangement and illustrate how a local variant can remain detectable from transcript to protein across cellular states.

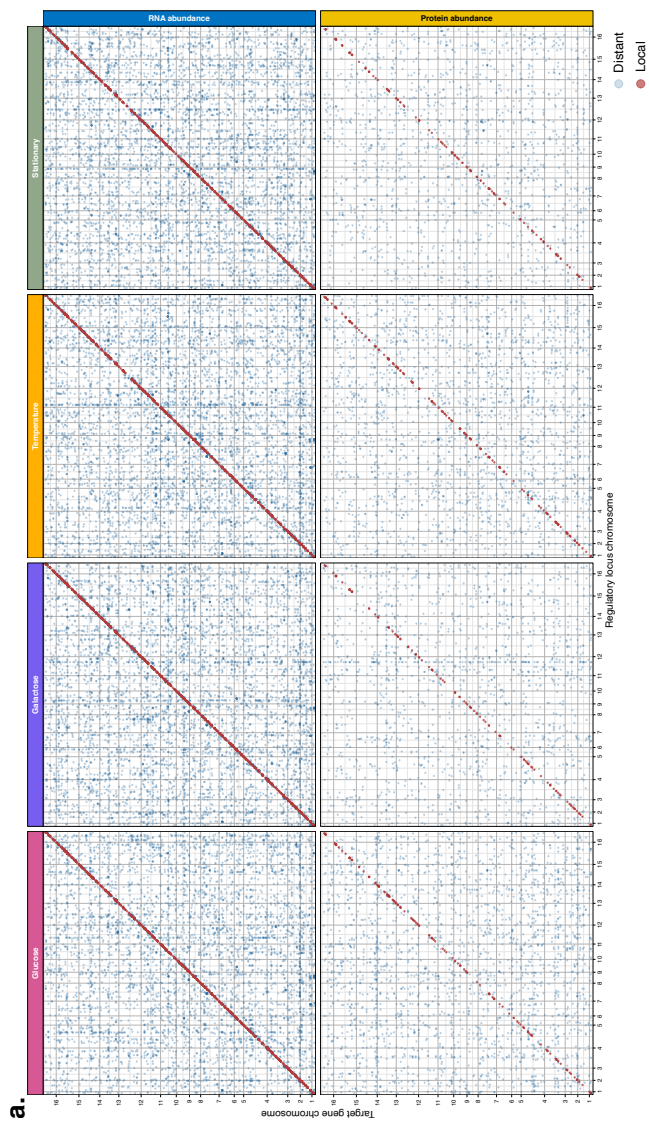
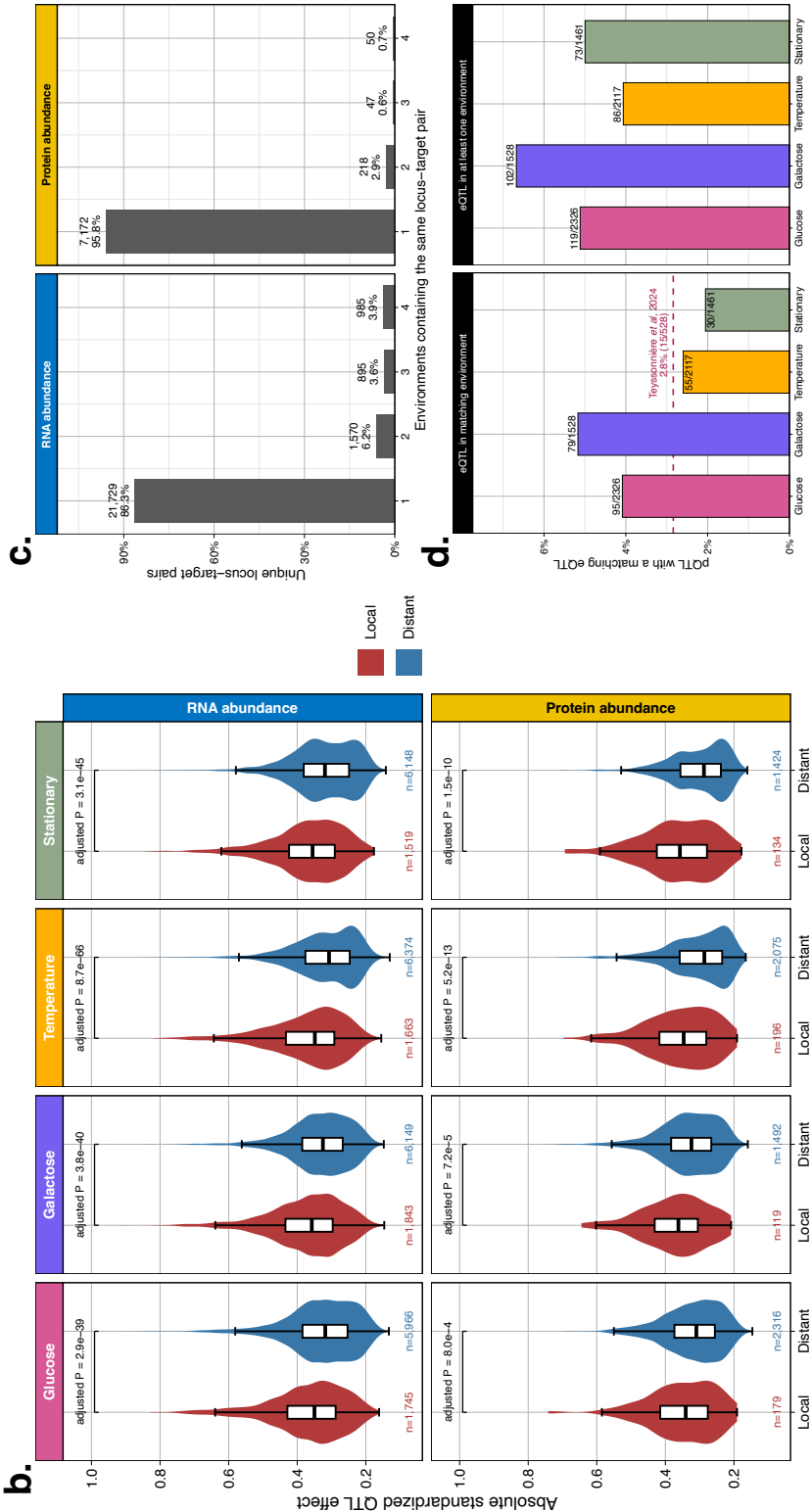


Figure 4 | Expression QTL are predominantly environment specific.

(a.) Genome-wide positions of regulatory loci and target genes by molecular layer and environment. The input matrices contained 95,804 transcriptome and 95,829 proteome markers. (b.) Condition-stratified standardized-effect distributions for local and distant associations in the transcriptome and proteome. P values are from two-sided Wilcoxon rank-sum tests with Bonferroni correction across the eight comparisons. (c.) Number and fraction of unique LD-defined locus-target pairs discovered in one, two, three or four environments. Each pair links one regulatory locus to one target gene and is distinct from a hotspot position. (d.) Fraction of pQTL sharing an exact SNP-target eQTL in the matching environment or in at least one transcriptomic environment, using pQTL as the denominator. The glucose estimate from Teyssonniere et al. is included for comparison.



Cross-omic transmission of genetic effects is incomplete and context dependent

Strict eQTL-pQTL overlap captures only associations that independently reach significance in both molecular layers. To quantify cross-layer transmission more broadly, we estimated the effect of every discovered eQTL on the abundance of its corresponding protein, retaining all 7,507 matched variant-gene estimates regardless of pQTL significance (Fig. 5a). For the same variant-gene pairs, standardized allelic effects were smaller at protein than transcript level, with zero-intercept protein-to-transcript slopes of 0.354 in glucose, 0.365 in galactose, 0.344 at elevated temperature and 0.219 in stationary phase. Signed Spearman correlations were 0.706, 0.724, 0.733 and 0.553, and direction agreement remained high at 84.4%, 85.5%, 86.6% and 72.5%, respectively. Within the RNA-selected lookup set, 612, 540, 526 and 140 protein effects passed BH FDR < 0.05 in the four environments. Direct effect lookup therefore recovered widespread cross-layer concordance that was not apparent from overlap between significant eQTL and pQTL alone, while showing substantial attenuation of effect magnitude, particularly in stationary phase.

GxE GWAS of within-strain transcript responses identified 2,674 galactose, 3,917 temperature and 3,561 stationary-phase LD-defined locus-target pairs. Response phenotypes were defined as galactose-minus-glucose, temperature-minus-glucose or stationary-minus-glucose differences. A gxe(e)QTL therefore marks a variant whose alleles are associated with different magnitudes or directions of transcript response between environments, rather than simply different abundance in one condition. The three scans included 487, 530 and 529 isolates, respectively, and involved 1,664, 2,401 and 2,354 target genes. Across the three contrasts, 9,907 of 10,029 distinct locus-target pairs (98.8%) were detected in one response contrast, 121 in two and one in all three (Fig. S12a). As a technical validation, we reconstructed each RNA interaction effect by subtracting the glucose eQTL effect from the corresponding condition-specific eQTL effect. These reconstructed effects closely agreed with those obtained by directly mapping within-strain environmental responses (Spearman ρ = 0.83-0.90; Fig. S12b).

For 3,264 transcript response QTLs with a measured protein target, we estimated the corresponding protein-response effect without requiring protein significance: 634 in galactose, 1,562 at 37°C and 1,068 in stationary phase (Fig. 5b). Protein effects frequently retained the transcript-predicted direction but only part of its magnitude (Fig. 5c). In galactose, 75.6% of locus-target pairs agreed in direction and the protein-to-transcript slope was 0.449. The corresponding values were 71.8% and 0.222 at 37°C, and 63.0% and 0.138 in stationary phase. Effect correlations showed the same descriptive ordering across the distinct discovery sets (0.610, 0.429 and 0.325, respectively). All nine condition-wise measures exceeded gene-block randomization nulls (Holm-adjusted P < 0.001), and alternative model implementations gave concordant protein estimates (Fig. S13a-c). Transcript response QTLs therefore frequently remained visible at protein level, but their transmission was incomplete and varied among environmental contrasts. This genome-wide pattern provided a baseline for identifying response hotspots with unusually strong cross-layer retention.

gxe(e)QTL hotspots differed in environmental specificity and protein retention (Fig. S13d). A locus near *MNN1* was restricted to the 37°C condition, whereas loci near *KGD1* occurred in galactose and stationary phase. By contrast, *rs_12_612729_C_T*, located within *THI7*, was detected in all three response contrasts, with 12, 17 and 19 transcript targets in galactose, temperature and stationary phase, respectively. Recurrence of the locus did not imply recurrence of its targets: no target was shared between galactose and temperature, only two between galactose and stationary phase and one between temperature and stationary phase. Protein effects retained the *gxe(e)QTL*-predicted direction for 8 of 10, 4 of 7 and 8 of 9 measured targets, respectively, with weaker retention at elevated temperature. The widest galactose *gxe(e)QTL* hotspot was *rs_12_291599_C_G* within *GAL2*, with 44 transcript targets. Of the 26 targets measured at protein level, 23 retained the transcript-predicted direction. This combination of broad transcript reach and strong protein retention motivated detailed analysis of the locus.

A *GAL2*-linked allele marks a cross-omic vacuolar response

The chromosome-XII signal mapped within *GAL2* (Fig. 6a). Its lead SNP, *rs_12_291599_C_G*, is a missense substitution encoding P463R. We used the RNA-defined target directions to construct aggregate transcript and protein response scores and compare 369 variants across *GAL2* and 25 kb of flanking sequence on either side. P463R was the only variant exceeding the local Bonferroni threshold in both score scans, with additive effects of 1.15 standardized score units on the transcript score ($P = 9.05 \times 10^{-11}$) and 1.04 on the transcript-defined protein score ($P = 7.30 \times 10^{-10}$). No other local variant had $r^2 \geq 0.8$ with P463R, as also indicated by the LD colour scale in Fig. 6a, and no secondary variant passed the local Bonferroni threshold after including P463R genotype as a covariate in the protein-response scan (Fig. S14a). Carriers were concentrated in Mixed Origins 1 and Mexican Agave, with smaller groups in three additional clades and three unassigned isolates (Fig. S14b). These two principal carrier clades belong to the Beer and Neotropical superclades, respectively. The focal effect remained positive after excluding either superclade (Fig. S14c). The association was therefore not restricted to one major carrier ancestry.

The *GAL2*-linked response extended across the genome to 44 coordinate-resolved transcript targets whose galactose-minus-glucose response differed by genotype (Fig. 6b). Protein-response effects were estimated for 26 transcript-selected targets, of which 23 followed the transcript-predicted direction and 20 passed protein FDR correction as *gxe(p)QTL*. Separate endpoint estimates showed that the absolute effect was larger in galactose than glucose for all 26 transcript effects and 20 of the 26 protein effects (Fig. S15a). The targets formed a functionally coherent set spanning vacuolar acidification, carbon-state metabolism and transmembrane transport (Fig. 6c). Their protein-to-transcript response-effect slope was 1.68 and remained between 1.60 and 1.78 after removing each target in turn. Coordinated negative responses included *VMA4*, *VMA7*, *VMA8* and *VMA10*, whereas *ACO1* had the strongest positive protein response.

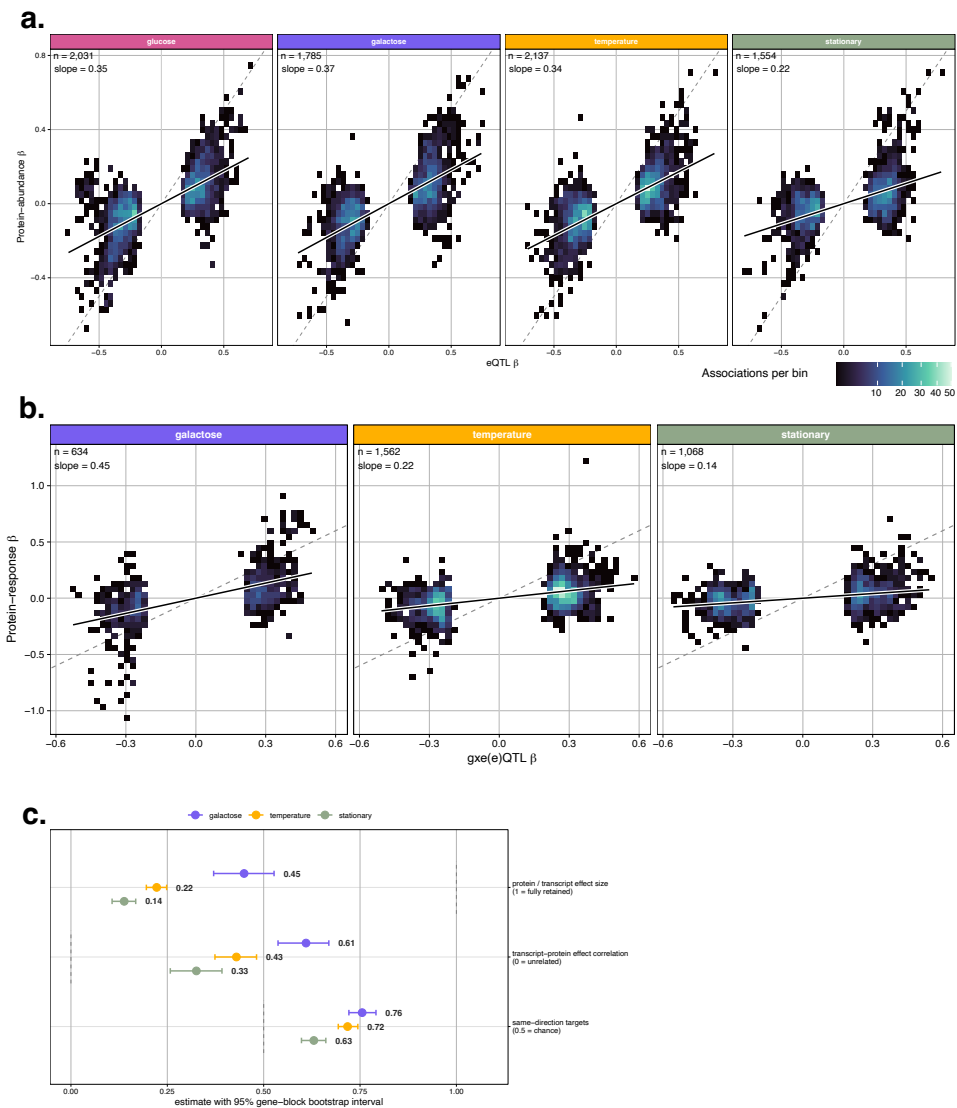


Figure 5 | RNA genetic effects are incompletely retained at protein level.

(a.) Standardized protein effects corresponding to eQTL discoveries, plotted against transcript effects in each environment. All protein estimates were retained, with a zero-intercept slope fitted in each environment. (b.) The same comparison for transcript-response QTL, or $gxe(e)QTL$, with each response measured relative to glucose. All corresponding protein-response estimates are retained; only associations passing protein FDR are termed $gxe(p)QTL$. (c.) Transmission of $gxe(e)QTL$ effects to protein, summarized by retained effect magnitude, transcript-protein effect correlation and same-direction targets, with 95% target-gene-block bootstrap intervals and gene-block randomization nulls. All nine observed estimates exceeded their matched nulls (Holm-adjusted $P < 0.001$).

Vacuolar acidification was enriched among both transcript and protein response targets, whereas the tricarboxylic acid cycle was enriched at transcript level (Fig. 6c; Fig. S15b). Among 42 GO-annotated *gxe(e)*QTL targets, vacuolar acidification and the tricarboxylic acid cycle each had $FDR = 0.0047$. Among the 20 *gxe(p)*QTL targets, four belonged to vacuolar acidification ($FDR = 1.77 \times 10^{-4}$), and the set was also enriched for proton transmembrane transport ($FDR = 0.0081$). The *GAL2*-linked response therefore connected carbon-state variation with vacuolar pH regulation at the transcriptome and proteome levels.

To distinguish environmental response from abundance within galactose, we separately mapped transcript and protein levels measured in galactose. The scan identified 88 eQTL targets and 80 pQTL targets; 86 and 79, respectively, were coordinate resolved in the genome map (Fig. S15c). Twenty-seven genes were detected in both layers, with concordant paired effects (Fig. S15d). Across these shared targets, aggregate local evidence in both molecular layers peaked at P463R (Fig. S15e). The observed overlap exceeded the hypergeometric expectation of 1.94 genes ($P = 1.97 \times 10^{-26}$; Fig. S15f). Among 26 GO-annotated shared targets, seven belonged to aerobic respiration ($FDR = 1.90 \times 10^{-7}$), and their placement across respiratory complexes and linked metabolism is shown in Fig. S15g. Only *ACO1* belonged to both this 27-gene abundance signature and the 26 protein-measured response targets. Thus, the same linked allele marked two largely distinct molecular consequences: a vacuolar environmental response and a respiratory abundance state in galactose.

The direction-oriented protein-response score showed a recessive-like association with P463R genotype across population backgrounds (Fig. 6d). Among 455 isolates, comprising 426 C/C, 17 C/G and 12 G/G isolates, the overall Kruskal-Wallis test gave $P = 3.56 \times 10^{-8}$. G/G isolates had higher scores than C/C isolates (Hodges-Lehmann shift = 1.397; Bonferroni-adjusted Wilcoxon $P = 2.34 \times 10^{-8}$) and C/G isolates (shift = 1.325; adjusted $P = 3.87 \times 10^{-5}$), whereas C/G and C/C isolates did not differ (adjusted $P = 0.961$). The genotype association was also detected within Mixed Origins 1 and Mexican Agave (Bonferroni-adjusted omnibus $P = 0.00504$ and 0.00456 , respectively). No association with the six available growth traits in glucose, galactose or elevated temperature survived multiple-testing correction. UniProt topology placed residue 463 in the cytoplasmic loop between transmembrane helices 10 and 11. Across five AlphaFold 3 predictions per allele, the global Gal2 fold was conserved, whereas P463R models placed residue 463 in close contact ($<5 \text{ \AA}$) with T325 and E535; neither contact was present in the wild-type predictions (Fig. S16a-b). The P463R haplotype nevertheless marked the widest galactose-response hotspot, with coordinated effects across dozens of transcript and protein targets. This broad cross-omic association nominates P463R for direct tests of Gal2 transport and signalling. Together with the genome-wide *gxe(e)*QTL landscape, the *GAL2*-linked response shows that genetic variants can reshape distributed transcript and protein responses in an environment-specific manner, including molecular effects not captured by the available growth traits. Resolving intermediate molecular phenotypes across environments is therefore necessary to understand genotype–phenotype relationships at population scale.

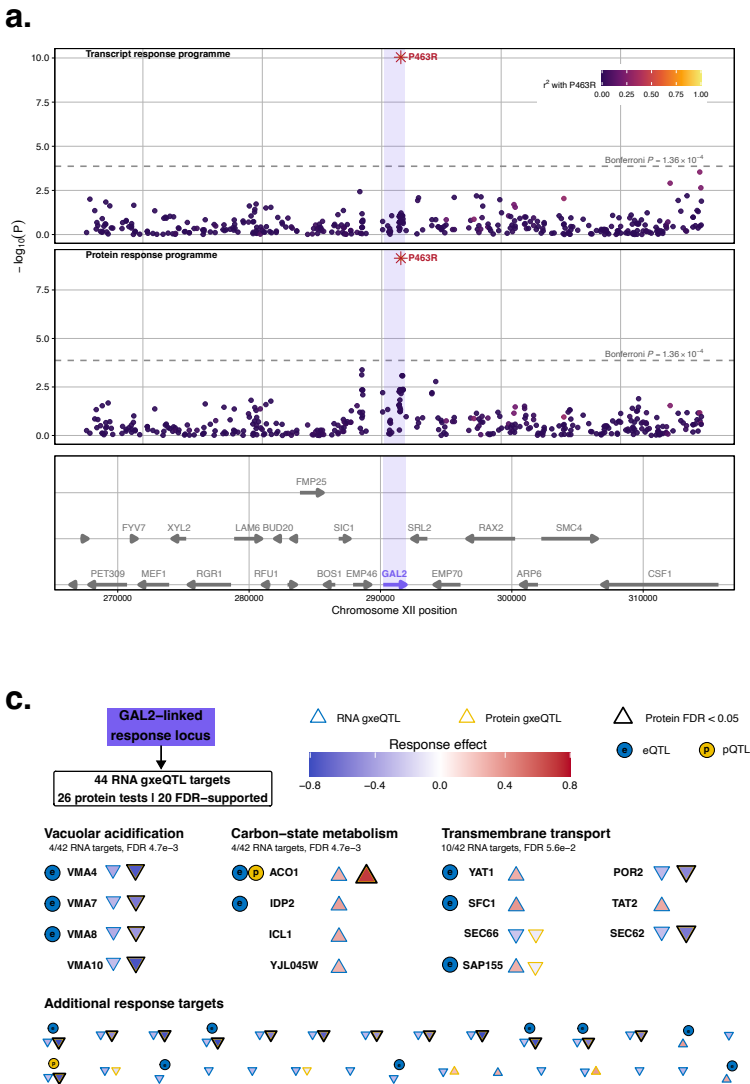
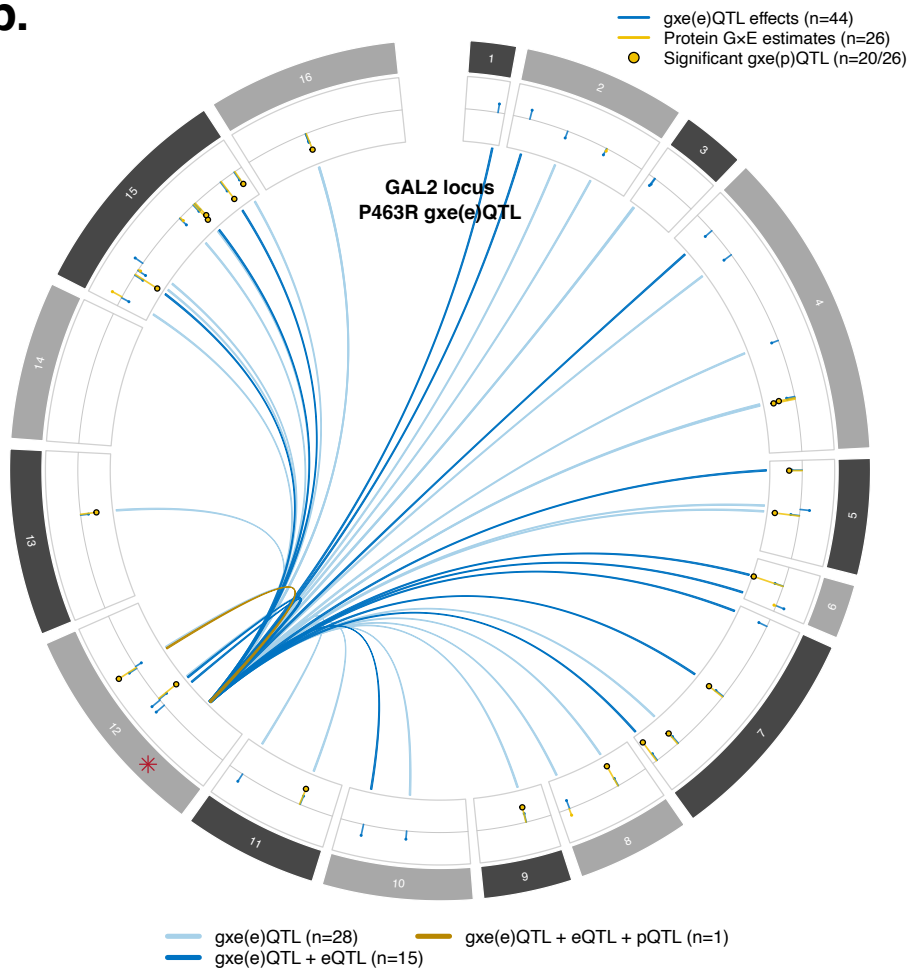


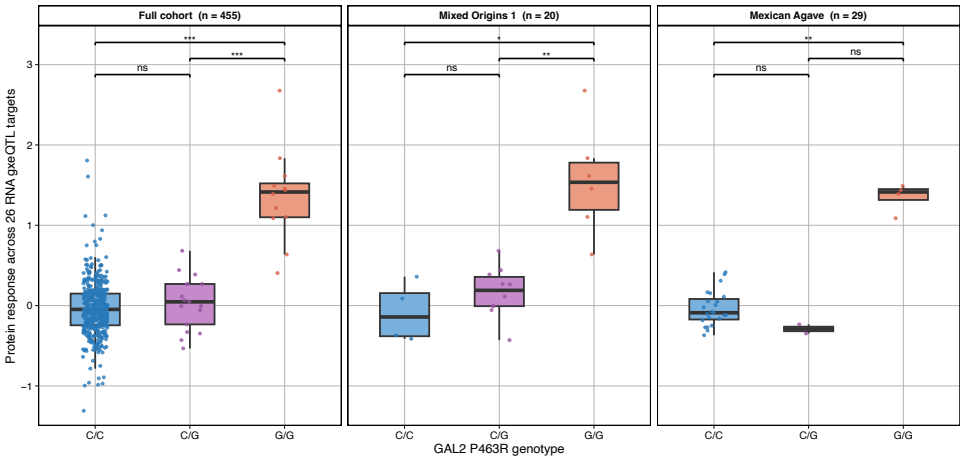
Figure 6 | *GAL2* variation links galactose response to cross-omic vacuolar regulation.

(a.) Local chromosome-XII scans of transcript and protein response scores constructed from the RNA-defined *GAL2* target set, with neighbouring genes, linkage disequilibrium with P463R and local Bonferroni thresholds. (b.) Circular map linking *GAL2* to 44 coordinate-resolved transcript-response targets: 27 response-only targets, 15 also detected as galactose eQTL, *VPS74*, also detected as a pQTL target, and *ACO1*, detected as both an eQTL and pQTL target. Twenty of the 26 protein-tested targets were gxe(p)QTL. (c.) Functional summary of the 44 response targets, grouped into vacuolar acidification, carbon-state metabolism and transmembrane transport, with transcript and protein response effects and galactose eQTL or pQTL status. (d.) Protein response scores across P463R genotypes in the full cohort, Mixed Origins 1 and Mexican Agave. The score summarizes galactose-minus-glucose changes across 26 proteins, oriented to the P463R-associated transcript effect. Omnibus Kruskal-Wallis tests were followed by Bonferroni-adjusted two-sided Wilcoxon rank-sum tests.

b.



d.



Discussion

To our knowledge, this study provides the first species-scale matched transcriptomic and proteomic analysis across multiple growth environments. It comprises 2,721 transcriptomes and 2,668 proteome samples across four environmental states, with a same-culture matched subset of 555 isolates. Measuring both molecular layers from the same cultures and repeatedly profiling the same genotypes across environments allowed direct comparison of two dimensions of regulatory variation: changes in genetic effects between environments and their retention from transcript to protein. Earlier cross-layer studies in yeast crosses, human cells and mouse populations established that genetic variants can act differently on transcript and protein abundance (Foss et al., 2007; Parts et al., 2014; Battle et al., 2015; Chick et al., 2016). Our design extends previous species-wide transcriptomic and proteomic resources generated in glucose (Caudal et al., 2024; Muenzner et al., 2024; Teyssonnière et al., 2024) by measuring both molecular layers from the same cultures across four environments. Across these environments, 86.3% of transcript-level and 95.8% of protein-level locus-target pairs were detected in one environment, whereas 3.9% and 0.7%, respectively, were detected in all four. The genetic bases of transcript and protein abundance were therefore largely private to individual environments. A small regulatory core nevertheless persisted across cellular states, including local associations detected in both molecular layers across all four environments. Around this core, most hotspot positions were restricted to one environment, and recurrent positions varied in their target reach across environments. The genetic architecture therefore combined persistent effects with a much larger set of condition-specific locus-target relationships. This environmental restriction extends earlier evidence for context-dependent regulatory effects in yeast crosses, human tissues and perturbed human cell populations (Smith & Kruglyak, 2008; Gagneur et al., 2013; Fairfax et al., 2014; GTEx Consortium, 2020; Lea et al., 2022).

Population history also structured environmental responses. Natural trait variation in yeast has long been known to follow population history, and the laboratory reference background is not representative of the phenotypic diversity of the species (Warringer et al., 2011). More recent species-wide transcriptomic and proteomic surveys similarly identified lineage-associated molecular profiles (Caudal et al., 2024; Teyssonnière et al., 2024). Here, MOFA separated dominant environmental factors from lineage-associated expression patterns, and a stringent clade-resolved interaction analysis assigned 891 genes to 2,221 large clade-gene-environment effects. French Dairy isolates recovered the constitutive GAL state described previously, with increased expression in glucose and reduced induction in galactose (Caudal et al., 2024). This molecular state complements the dairy-associated adaptation of galactose growth reported for cheese strains (Legras et al., 2018). South American Wild isolates showed a broadly more negative transcriptional response at 37°C, dominated by stronger repression and activation-to-repression reversals that were not apparent when the 37°C samples were analysed alone. A single environmental measurement therefore does not capture population variation in expression responses, consistent with molecular GxE

observed in yeast and other population panels (Smith & Kruglyak, 2008; Lea et al., 2022).

Transcript and protein abundance had largely distinct genetic bases, but low overlap between significant eQTL and pQTL did not imply an absence of transmission between layers. Exact SNP-target overlap ranged from 2.1% to 5.2% within the same environment, reproducing the approximately 3% reported in glucose (Teyssonnière et al., 2024). A shared QTL call requires the same association to exceed independent significance thresholds in both datasets. This statistic therefore combines weak protein effects that remain below pQTL significance with pQTL that have no detectable transcript association, while differences in assay coverage and precision further influence discovery power. Direct lookup addressed the first component without requiring protein significance: 72.5% to 86.6% of transcript-selected abundance effects and 63.0% to 75.6% of transcript-response effects retained the same direction at protein level. High-powered yeast linkage analyses similarly showed that overlap between significant QTL underestimated effect-level sharing (Albert et al., 2018). Searching across transcriptomic environments and relaxing exact marker identity only modestly increased sharing in our data, however, indicating that discovery thresholds and allelic resolution did not explain the entire separation between layers.

The separation also had biological structure. Sharing was concentrated among local associations: 17.7% to 26.8% of local pQTL matched an eQTL in the same environment, compared with only 0.6% to 3.5% of distant pQTL. Direct local effects may therefore propagate more reliably than distant effects mediated through regulatory networks. In the previous species-wide glucose dataset, proteins encoded by genes with shared eQTL and pQTL had lower turnover and longer half-lives, linking cross-layer concordance to protein stability (Teyssonnière et al., 2024). Other matched studies showed that transcript effects can be attenuated downstream of ribosome occupancy and that protein-specific QTL can arise without corresponding transcript effects (Battle et al., 2015; Chick et al., 2016; Brion et al., 2020). The extent of buffering or enhancement can also depend on the functional target programme (Grossbach et al., 2022). Protein turnover changes across carbon sources and contributes to dosage compensation of genomic copy-number changes, illustrating how proteostasis can reshape an upstream molecular perturbation (Martin-Perez & Villén, 2017; Muenzner et al., 2024). In our data, protein-to-transcript slopes ranged from 0.219 to 0.365 for abundance QTL and from 0.138 to 0.449 for response QTL. Their descriptive ordering was consistent across retention measures, with the highest estimates in galactose and the lowest in stationary phase. Environment-dependent translation, degradation, complex assembly and growth dilution provide plausible mechanisms for this variation; global turnover measurements in yeast show that apparent protein lifetimes are strongly shaped by dilution through cell division (Christiano et al., 2014). Turnover may influence steady-state sharing and response kinetics differently: the association with stable proteins does not imply that a long-lived protein pool will adjust rapidly after an environmental transition. Persistent proteins and reduced growth dilution could instead delay protein adjustment in stationary phase. Because synthesis, degradation and response kinetics were not measured directly, their relative contributions remain to be resolved.

The *GAL2*-linked haplotype was associated with a galactose-specific transcript response that was strongly retained at protein level but not with the measured growth traits. Natural variation in the GAL network can alter adaptation to galactose and its growth consequences through alleles of the galactose sensor *GAL3* and coadapted alleles of *GAL2*, *GAL1/GAL10/GAL7* and *PGM1* (Lee et al., 2017; Boocock et al., 2021). Our association of the *GAL2* haplotype with coordinated vacuolar and mitochondrial responses at both transcriptome and proteome levels suggests consequences beyond canonical galactose induction. Gal2 itself undergoes glucose-triggered catabolite inactivation through ubiquitination, endocytosis and vacuolar degradation, providing a precedent for a connection between transporter state and vacuolar regulation (Horak & Wolf, 1997). The absence of a detectable growth association despite coherent molecular effects suggests that this variation may be buffered at the level of bulk growth or expressed through physiological traits not captured by the assay. Indeed, the relationship between gene expression and fitness can itself change across carbon environments (Siddiq et al., 2024). This separation between molecular and organismal phenotypes illustrates why environmental responses and intermediate molecular layers are needed to trace the path from genotype to phenotype.

Environmental context determined which genetic associations were detected, how many targets regulatory loci reached and how strongly transcript effects were retained at protein level. Across four selected environments, a small core of persistent cross-condition and cross-layer effects was accompanied by extensive condition-specific and layer-specific genetic control. This combination reveals a highly complex regulatory architecture between genetic variation and the establishment of organismal phenotypes. Consequently, analyses performed in a single condition, at one molecular layer or only at the level of a terminal phenotype recover only a tiny subset of the genetic effects acting across environmental states. Context-dependent eQTL and heritability have likewise been observed across cellular exposures and complex traits (Dahl et al., 2020; Lea et al., 2022). Such conditional effects provide one route by which genetic variation can escape marginal association tests and may contribute to the broader problem of missing heritability (Manolio et al., 2009; Dahl et al., 2020). Species-wide matched multi-omic profiling across environmental states is therefore necessary to connect genetic diversity to the intermediate molecular changes from which genotype–phenotype relationships emerge.

Methods

Isolate panel and environmental conditions

The transcriptomic panel contained 572 predominantly diploid *S. cerevisiae* isolates, comprising 547 isolates from the collection profiled by Caudal et al. (2024) and 25 additional Mexican Agave isolates from the population described by Avelar-Rivas et al. (2026). The proteomic panel contained 576 isolates, including 15 not present in the transcriptomic panel, giving 587 distinct isolates across either molecular layer. Strain metadata, clade assignments, ecological origins, ploidy, heterozygosity and aneuploidy states are provided in Supplementary Table 1. Clade assignments and colours followed the published population classification (Loegler et al., 2024).

Strains were cultured in synthetic complete medium under four conditions: exponential growth with 2% glucose at 30°C, exponential growth with 2% galactose at 30°C, exponential growth with 2% glucose at 37°C, and early stationary phase in 2% glucose at 30°C.

Preculture and growth monitoring

Precultures were established in 96-well plates containing 200 µL of the appropriate medium supplemented with ampicillin and incubated at 30°C for approximately 24 h to reach saturation. Using an Eppendorf epMotion 96XL liquid-handling system, 100 µL of each preculture was transferred to a flat-bottom 96-well plate (Sarstedt, 83.3924) and optical density (OD) was measured with a Tecan Infinite M Nano+ plate reader. Based on these measurements, strains were inoculated into 6-mL 48-well polypropylene plates (Thermo Scientific Abgene, AB0988) containing 6 mL of culture medium per well at a target starting OD of 0.05. Each culture was established across three parallel plates used for OD monitoring, backup and sample harvesting. Plates were sealed with breathable membranes (Greiner 676051) and incubated at 30°C or 37°C with constant agitation at 180 rpm.

In addition, every well on all culture plates was mixed by repeated pipette aspiration and dispensing at 2-h intervals throughout the 36-h monitoring period. For growth monitoring, 100 µL was transferred from each culture in the OD-monitoring plate to a flat-bottom 96-well plate of the same type and measured with the Tecan plate reader. Each aliquot was then returned to its original well to maintain constant culture volumes. Exponential cultures were harvested individually when the monitoring culture reached an OD of approximately 0.3, denoting the exponential growth-phase. Stationary-phase cultures were harvested approximately 8 h after the transition to stationary phase, as determined from the OD trajectory.

Sample harvesting

At harvest, 4 mL of each culture was transferred to a new 48-well plate of the same type for proteomics and centrifuged at $2,000 \times g$ for 20 s. After removal of the supernatant,

the cell pellet was resuspended and transferred to a 2-mL 96-well polypropylene storage plate with square V-bottom wells (PlateOne, STARLAB, S1896-2000), centrifuged again at $2,000 \times g$ for 20 s, and the supernatant discarded. For RNA sequencing, a total of 1,600 μL of the same culture was divided between two storage plates of the same type, centrifuged at $2,000 \times g$ for 20 s, and the supernatant removed. Storage plates were sealed with pierceable aluminium adhesive foil (4titude, 4Ti-0550), and cell pellets for both molecular layers were flash-frozen in liquid nitrogen and stored at -80°C until processing.

RNA extraction, library preparation and sequencing

Polyadenylated RNA was isolated with the Dynabeads mRNA DIRECT Purification Kit (Invitrogen, 61012) after disruption with acid-washed 425-600 μm glass beads (Sigma-Aldrich, G8772). Strand-specific libraries were prepared in 96-well format with the NEBNext Ultra II Directional RNA Library Prep Kit (New England Biolabs, E7765L) and unique dual indexes. Libraries were quantified, pooled equimolarly, size selected and sequenced on Illumina NextSeq 2000 P2/P3 flow cells, with one run generated on an Illumina NovaSeq S2 flow cell. The dataset contained both 120-nt single-end and approximately 60-nt paired-end libraries.

Adapters and low-quality sequence were removed with fastp v1.0.1 (Chen et al., 2018). Reads shorter than 40 bp after trimming or with mean Phred quality below 25 were discarded. FastQC v0.12.1 and MultiQC v1.31 were used for quality reporting (Ewels et al., 2016). Read processing, identity verification and quantification were orchestrated with Snakemake v9.12.0 (Mölder et al., 2021). Libraries with fewer than one million mapped reads were excluded from downstream analyses.

Sample identity verification and contamination screening

RNA-seq reads were aligned to the R64-3-1 reference genome with STAR v2.7.11b for genotype verification (Dobin et al., 2013). Alignments and variant records were processed with SAMtools/BCftools v1.21 (Danecek et al., 2021) and GATK v4.6.1.0 (McKenna et al., 2010), and sample genotypes were compared with a reference panel of 3,039 sequenced isolates. Candidate identities were ranked using all informative alternate sites and, independently, low-frequency variants. Samples were validated when the expected strain ranked first in both comparisons; concordant alternative matches were reassigned and ambiguous samples were inspected manually or discarded. Unmapped reads were screened for contamination with Kraken2 v2.1.3 and Kaiju v1.10.1 (Menzel et al., 2016; Wood et al., 2019).

Pangenome-aware RNA quantification

Expression was quantified against an augmented reference containing the R64-3-1 genome and 2,199 supplementary pangenome ORFs, each represented as a separate contig with 100 bp of flanking sequence. Reads were aligned with STAR and counted with featureCounts v2.0.6 in reverse-stranded mode, allowing fractional assignment

of multimapping reads (Dobin et al., 2013; Liao et al., 2014). Supplementary ORFs were aligned to reference coding sequences with BLAST+ blastn v2.16.0 (Camacho et al., 2009). Redundant supplementary ORFs were collapsed onto reference genes when nucleotide alignments covered at least 95% of the reference coding sequence, differed in length by no more than 20% and had $E < 10^{-100}$. This yielded a nonredundant annotation of 7,909 features. Technical resequencing replicates were summed at count level. TPM values were calculated from merged counts for descriptive abundance analyses where stated. Differential-expression significance and log₂ fold changes were estimated from raw counts with DESeq2. For global expression maps, MOFA, programme scores, reaction norms and clade-resolved GxE effect sizes, counts were also transformed using the DESeq2 variance-stabilizing transformation (VST), which reduces the dependence of variance on mean abundance and approaches a log₂ scale for sufficiently expressed genes (Love et al., 2014). Within each condition, genes were retained when TPM was greater than zero in at least 50% of samples unless otherwise stated.

Proteomics acquisition and processing

Protein pellets from the matched cultures were lysed, reduced, alkylated and digested with trypsin in a semiautomated plate-based workflow. Peptides were measured by analytical-flow liquid chromatography and Scanning-SWATH mass spectrometry (Messner et al., 2021).

DIA-NN v2.2.0 precursor reports were filtered at 1% FDR across precursor and protein-group confidence metrics (Demichev et al., 2020). Precursors with MS2 quantity below one-third of the global median were removed, together with run-level signal or identification outliers. Intensities were aggregated to gene level using the maximum MaxLFQ value per gene and log₂ transformed (Cox et al., 2014). Two processed protein matrices were retained. For differential-abundance and environmental-response analyses, samples with more than 50% missing values were discarded, proteins observed in at least 80% of retained samples were kept, samples were median centred and missing values were not imputed. For pQTL mapping, protein t-SNE and MOFA, processing was performed separately within each environment: proteins observed in at least 75% of runs were retained, samples with more than 50% missingness were removed, missing log₂ values were imputed by k-nearest neighbours ($k = 10$; Troyanskaya et al., 2001) using `impute::impute.knn`, and the resulting matrices were quantile normalized (Bolstad et al., 2003) using `preprocessCore::normalize.quantiles`. Direct cross-omic response tests and the *GAL2* response scores used the nonimputed measurements.

Growth phenotypes

Growth curves were measured in glucose, galactose and elevated-temperature conditions in flat-bottom 96-well plates (Sarstedt, 83.3924) using the Tecan Infinite M Nano+ plate reader, with OD600 recorded every 10 min for approximately 48 h. After joining successive acquisition segments and baseline correcting each well, six

traits were derived in R using *gcplyr* (Blazanin, 2024): maximum per-capita growth rate, carrying-capacity amplitude, lag duration, area under the curve, time to 90% of carrying capacity and time to mid-log phase. Artefactual readings and derivative edge effects were excluded from the relevant estimates, and replicate estimates were summarized by their median for each strain and condition. Cultures that did not reach the minimum carrying-capacity amplitude and area-under-the-curve thresholds were classified as non-growing. Quantitative traits were rank-based inverse-normal transformed within condition for integrative analyses.

Global expression maps and programme scores

RNA and protein t-SNE maps were generated independently using t-SNE (van der Maaten & Hinton, 2008) with the Barnes-Hut approximation (van der Maaten, 2014), as implemented in *Rtsne*. Both used perplexity 30, theta 0.25, 2,500 iterations; the transcript map used no PCA initialization, whereas the protein map used PCA initialization. For programme scores, each gene was standardized across samples before averaging. The ESR gene sets were taken from the 2017-updated annotation of the programme defined by Gasch et al. (2000), comprising 283 induced and 596 repressed genes (Gasch, 2026). The repressed set contains PAC/ribosome-biogenesis and ribosomal-protein arms. The ESR score was the mean induced-ESR z score minus the mean repressed-arm z score, so larger values indicate stronger stress-response activation. The GAL score was the mean z score of *GAL1*, *GAL2*, *GAL3*, *GAL7*, *GAL10*, *GAL80* and *PGM2*. Protein ESR scores were calculated identically using members measured in all four environments. The protein GAL map was not used because the defining proteins were not measured across all four environments without imputation.

Multi-omics factor analysis

MOFA+ was fitted with MOFA2 v1.20.2 and its *mofapy2* v0.7.4 backend to integrate 2,496 matched isolate-condition samples from 554 isolates using VST-transformed transcript abundance, condition-specific imputed and quantile-normalized protein abundance, and inverse-normal-transformed growth traits (Argelaguet et al., 2018, 2020). The transcript and protein data layers comprised the same 2,664 genes, selected because they were quantified in the proteomics resource; the growth layer comprised six traits where available. All three layers were scaled to unit variance and modelled as continuous Gaussian data. Automatic relevance determination was used to shrink factors towards zero in data layers where they explained little variation. We initialized 15 factors and retained 11 that explained at least 1% of the variance in one or more layers. To assess stability, we refitted the model 16 times after randomly retaining 85% of isolates. A factor was considered recovered when its scores had an absolute correlation of at least 0.8 with the best-matching factor from the complete model across shared samples. We also fitted a sensitivity model in which the transcript layer was expanded to all filtered genes while the protein and growth layers remained unchanged. GO Biological Process enrichment was evaluated separately for positive and negative

factor-loading poles with `MOFA2::run_enrichment`, using its parametric test and BH adjustment.

Functional enrichment

GO Biological Process memberships were obtained through the Bioconductor annotation framework (Huber et al., 2015), using AnnotationDbi with the SGD-derived `org.Sc.sgd.db` resource and `GO.db`, and augmented with curated annotations for supplementary ORFs (The Gene Ontology Consortium et al., 2021; Engel et al., 2025). For condition-level enrichment, transcripts and proteins were ranked by the DESeq2 Wald statistic and limma moderated t statistic, respectively. Gene-set enrichment analysis was performed with clusterProfiler v4.18 using its fgsea backend, gene-set sizes of 10 to 500 and Benjamini-Hochberg adjustment (Subramanian et al., 2005; Xu et al., 2024).

Environmental responses and lineage differential expression

Environmental responses were defined within isolate as abundance in each alternative condition minus abundance in glucose. These comparisons used $\log_2(\text{TPM} + 1)$ transcript values and nonimputed, median-centred $\log_2$ protein values. Genes required paired measurements from at least 40 isolates. A transcript response was retained when its mean differed from zero in a two-sided one-sample t test, with BH-adjusted $P < 0.05$ within the contrast, and its absolute mean change was at least 0.1. Cross-layer plots retained the remaining genes as background. For each gene, we separated the mean response across isolates from each isolate's deviation from that mean. Confidence intervals were estimated by resampling whole genes, keeping all observations from the same gene together.

3

Global environmental differential expression was tested relative to glucose. Transcript contrasts were fitted to raw counts with DESeq2 using `glmGamPoi` dispersion estimation and sequencing run and condition as predictors, and $\log_2$ fold changes were shrunk with `apegglm` (Love et al., 2014; Zhu et al., 2019; Ahlmann-Eltze & Huber, 2021). Differentially expressed transcripts required an absolute $\log_2$ fold change greater than 1 and BH-adjusted $P < 1 \times 10^{-5}$. Protein contrasts were fitted to the nonimputed, median-centred $\log_2$ matrix with limma, including MS batch and using robust empirical-Bayes moderation (Ritchie et al., 2015; Phipson et al., 2016). Differentially abundant proteins required an absolute $\log_2$ fold change greater than 0.5 and BH-adjusted $P < 1 \times 10^{-5}$.

Within each environment, lineage-associated RNA expression was tested with DESeq2 negative-binomial Wald models fitted to raw counts (Love et al., 2014). VST values were used to visualize expression patterns but not to estimate differential-expression significance or $\log_2$ fold changes. Genes were retained when they had a nonzero count in at least 50% of samples in that environment, and clades were tested when represented by at least three profiled samples. For each clade-environment combination, the focal clade was contrasted with all other samples using the design `~ library +`

clade_vs_all. Size factors were inherited from the baseline RNA model, and dispersions estimated once per environment with glmGamPoi were reused across clade contrasts (Ahlmann-Eltze & Huber, 2021). Positive log₂ fold changes indicate higher expression in the focal clade. The transcript catalogue required an absolute log₂ fold change greater than 1 and a Benjamini-Hochberg adjusted P value below 1×10^{-5} , with adjustment performed across genes within each clade-environment contrast. Protein models used limma with MS batch as a covariate where estimable (Ritchie et al., 2015). The protein catalogue required an absolute log₂ fold change greater than 0.5 and a Benjamini-Hochberg adjusted P value below 0.05 within each clade-environment contrast.

Global and clade-resolved GxE

RNA GxE was first screened for 6,513 genes with a DESeq2 count model containing clade, environment and their interaction, using glmGamPoi dispersion estimation (Love et al., 2014; Ahlmann-Eltze & Huber, 2021). Eligible clades contained at least four samples in each environment. A likelihood-ratio test compared the full model with one lacking the interaction, and genes with BH FDR below 10^{-5} passed this initial screen. Specific lineage responses were then resolved among the 370 isolates measured in all four environments. For each alternative condition, limma compared the mean within-isolate VST response of each clade represented by at least five isolates with the unweighted mean response of all other eligible clades, followed by robust empirical-Bayes moderation. P values were BH adjusted across genes within each clade-condition comparison. A clade-gene-environment event was retained when the gene passed the initial interaction screen, its clade-resolved FDR was below 0.05 and the absolute response difference exceeded one VST unit. The initial interaction screen used raw counts, whereas the clade-resolved analysis used VST values to quantify response magnitude.

Static eQTL and pQTL mapping

BCFtools v1.21 was used to subset and filter biallelic SNPs, InDels and structural variants, and PLINK v1.90b6.21 was used to convert and merge these markers with gene-CNV encodings into binary genotype matrices (Chang et al., 2015; Danecek et al., 2021). Genome-wide association was performed independently for each environment and molecular layer with the FaST-LMM linear mixed model (Lippert et al., 2011). Phenotypes were transformed to a rank-based inverse-normal distribution, following previous yeast GWAS (Caudal et al., 2024; Tsouris et al., 2024; Loegler et al., 2025). Genotypes had a minor allele frequency of at least 5% in the relevant population. The transcriptome genotype matrix contained 95,804 marker encodings: 88,971 SNPs, 1,468 InDels, 767 SVs and 4,598 CNV dosage thresholds representing 1,740 genes. The proteome matrix contained 95,829 marker encodings: 88,971 SNPs, 1,499 InDels, 769 SVs and 4,590 CNV dosage thresholds representing 1,731 genes. Population structure was controlled using a genetic relatedness matrix calculated without the chromosome containing the tested variant, an approach known as leave-one-chromosome-out.

Aneuploidy states and technical batch indicators were included as covariates. Following the same empirical strategy, each phenotype was permuted 100 times and the minimum P value was retained from each genome-wide scan; the fifth-lowest of these 100 minima defined the trait-specific 5% threshold.

For regulatory-architecture analyses, significant variants within 20 kb of chromosome ends were excluded. Pairwise LD was calculated with PLINK v2.0.0-a.6.9LM (Chang et al., 2015). Starting from the most significant unassigned variant, markers within 50 kb and with $r^2 > 0.5$ to that lead were grouped into the same locus; this lead-centred procedure was repeated until every significant variant was assigned. Associations were classified as local when the lead marker lay within 20 kb of the annotated target-gene interval and distant otherwise. We retained the most significant association for each LD-defined locus-target pair, together with its environment and molecular layer. Local and distant effect distributions were compared within each of the eight layer-environment combinations using two-sided Wilcoxon rank-sum tests with Bonferroni correction. Effect distributions among variant classes were compared using Kruskal-Wallis tests with Bonferroni correction across the same eight combinations.

Regulatory-hotspot thresholds were calibrated separately for each molecular layer and environment. We randomly reassigned the observed locus-target associations among eligible LD-defined loci 2,000 times while keeping the locus set and total association count fixed. The threshold was the smallest target count reached by no more than 1% of randomized loci. This gave five targets in every transcript environment and three or four targets in the protein environments. We used the more conservative common protein threshold of four targets in all environments.

A hotspot instance was an LD-defined locus that passed the relevant target-count threshold in one environment. A hotspot position was the corresponding unique locus across environments, and a recurrent position passed the threshold in at least two environments. Environmental breadth was the number of environments in which a hotspot position passed the threshold. These quantities are distinct from a locus-target association, which links one regulatory locus to one target gene. An exact-variant hotspot analysis using the same thresholds was performed as a sensitivity analysis. Positional sharing between layers required transcript and protein hotspot instances in the same environment with lead positions no more than 5 kb apart, irrespective of target identity. The stricter same-target definition additionally required at least one common target gene.

To compare hotspot architecture at matched molecular coverage, we restricted each environment to genes tested in both layers and randomly subsampled significant locus-target associations to the smaller layer-specific count. This was repeated 1,000 times using a common threshold of five targets. We compared the number of hotspots per 1,000 tested loci, maximum target reach, median reach of the ten largest loci and the fraction of associations assigned to hotspots. This analysis matched target-gene coverage and discovery counts but not sample composition or measurement precision.

Cross-environment and cross-layer QTL comparisons

Environmental breadth was defined as the number of conditions in which an LD-defined locus-target pair was significant. Its association with log median absolute effect was tested separately by molecular layer and locality using lme4, with environment count as a fixed effect and regulatory locus as a random intercept (Bates et al., 2015). To match the published single-environment analysis, SNP-QTLs were pruned at $r^2 > 0.6$ and sharing required the same SNP position and target gene (Teyssonnière et al., 2024). Same-environment overlap required the SNP-target association to be significant at both transcript and protein levels in the same condition. Cross-environment lookup allowed the matching eQTL to occur in any transcriptomic environment. These percentages used pQTL as the denominator and therefore differ from gene-level overlap statistics. A sensitivity analysis relaxed exact marker identity by allowing RNA and protein variants in LD at $r^2 \geq 0.2$. Effect-concordance analyses aligned effect alleles and retained one shared SNP for each gene, environment and local/distant class: the SNP with the largest combined RNA and protein log10 significance margin above the respective permutation thresholds.

For the *SSU1-GLR1* interval, association statistics were examined from 365 to 385 kb on chromosome XVI, and LD was calculated as the squared correlation between each marker dosage and rs_16_375375_T_A. Focal-SNP and VIII-XVI *SSU1* rearrangement genotypes were projected onto the independently generated transcriptome and proteome t-SNE maps. Rearrangement carriers were identified using the structural-variant marker with a chromosome-XVI breakpoint at approximately 373.7 kb. Robustness analyses used the same rank-based inverse-normal-transformed *GLR1* phenotypes and focal-SNP dosage. Separate analyses adjusted for five genotype PCs, the original GWAS technical covariates or the rearrangement marker; excluded rearrangement carriers or chromosome-XVI aneuploids; or estimated effects within clades. Within-clade estimates centred phenotype and dosage within clades containing at least five isolates and more than one SNP genotype. Leave-one-clade analyses sequentially excluded clades represented by at least ten isolates.

Direct cross-omic transmission analysis

For abundance QTL, we took every RNA locus-target discovery retained after subtelomeric filtering and estimated the effect of the same allele on the corresponding protein in the same environment. Transcript and protein effects were aligned to the same allele and expressed in rank-based inverse-normalized phenotype units. All finite protein estimates were retained irrespective of protein significance, yielding 7,507 tests: 2,031 in glucose, 1,785 in galactose, 2,137 at 37°C and 1,554 in stationary phase. Protein P values were adjusted by the BH method within each environment.

RNA response phenotypes were within-strain log2(TPM + 1) differences between galactose, 37°C or stationary phase and glucose. Traits with essentially no variation were removed when their median absolute deviation was zero, at least half of responses

were exactly zero, or at least 80% had an absolute response no greater than 0.05 and an interquartile range no greater than 0.1.

For each significant RNA locus-target response association whose protein was measured, the protein response was the corresponding nonimputed within-strain abundance difference. We included only isolates with protein measurements for that gene in both glucose and the contrasted environment, with a minimum of 300 isolates. The isolate set could therefore differ among genes and contrasts. Protein responses were transformed by the Blom rank-based inverse-normal transformation and tested with a mixed model using a relatedness matrix that excluded the tested chromosome and the same aneuploidy covariates as the RNA response scan. Protein P values were BH adjusted within each response contrast. Within this targeted lookup, associations passing $FDR < 0.05$ were termed $gxe(p)QTL$; this label does not denote an independent genome-wide protein-response scan.

Cross-layer transmission was summarized by the allele-aligned Spearman correlation, the fraction of effects with the same sign and the protein-to-transcript slope from a regression constrained through zero. This slope measures the fraction of the standardized transcript effect retained at protein level; it is not a biochemical translation coefficient. Uncertainty was estimated from 1,000 bootstrap samples of target genes, with all associations involving the same gene resampled together. Null distributions came from 10,000 randomizations in which each target gene was assigned a random sign that was applied to all its associations. One-sided empirical P values were Holm corrected across the nine condition-by-metric tests in Fig. 5c. A separate wide-matrix analysis restricted to fully observed proteins served as an implementation sensitivity check.

***GAL2* locus and response analyses**

The focal *rs_12_291599_C_G* association, encoding Gal2 P463R, was selected as the galactose RNA-response hotspot with the largest number of targets before protein effects were examined. Cross-omic response scores were constructed from the same 26 RNA-defined targets that were measured at protein level. For each gene, its galactose-minus-glucose response was standardized across isolates and multiplied by the sign of its RNA response-QTL effect. The score was the mean of these oriented values within an isolate and required at least 20 measured transcript targets or 15 measured protein targets. Protein responses were not imputed.

The local scan included variants with a minor allele frequency of at least 2% across *GAL2* and 25 kb on either side, giving 369 variants. Transcript and protein scores were Blom inverse-normalized and tested with chromosome-XII leave-one-chromosome-out mixed models including the RNA-response aneuploidy covariates. LD with P463R was the squared correlation between genotype dosages, and the local threshold was Bonferroni $0.05/369$. To search for a secondary signal, the P463R G-allele dosage was added as a covariate and every other local variant was retested. Robustness analyses repeated the P463R test after excluding the Beer or Neotropical superclade and within Mixed Origins 1 and Mexican Agave.

C/C, C/G and G/G response-score distributions were compared with a Kruskal-Wallis test followed by pairwise two-sided Wilcoxon rank-sum tests. The three pairwise comparisons were Bonferroni corrected within each population, and the two carrier-clade omnibus tests were corrected together.

Biological-process enrichment was tested with clusterProfiler separately for the 44 transcript-response targets and the 20 protein-response QTL against the 5,804 GO-annotated genes retained in the galactose-response scan (The Gene Ontology Consortium et al., 2021; Xu et al., 2024). One-sided hypergeometric P values were BH corrected within each target set. For the separate galactose-abundance analysis, transcript and protein target overlap was tested among the 2,476 genes measured in both layers; this common set contained 60 transcript targets, 80 protein targets and 27 shared targets. Biological-process enrichment of the 26 GO-annotated shared targets used the 2,353 GO-annotated genes in this common set, retained terms containing 5 to 500 genes and applied one-sided hypergeometric tests followed by BH correction.

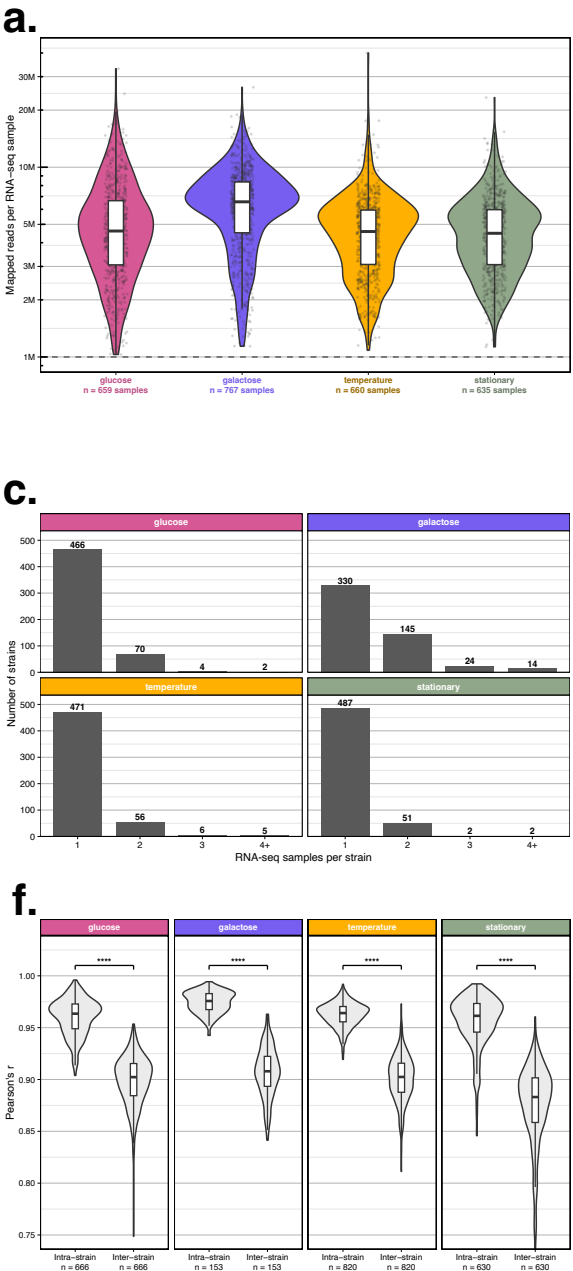
Growth associations used six rank-based inverse-normalized traits measured in glucose, galactose and at 37°C, together with the six paired galactose-minus-glucose response traits. Kinship-adjusted models used a SNP genomic relationship matrix calculated with LDAK v6.1 using the LD-adjusted kinship approach of Speed et al. (2012), included separate C/G and G/G genotype terms relative to C/C, and applied Bonferroni correction across the resulting 48 coefficients.

Gal2 membrane topology was obtained from the UniProt annotation for P13181 (The UniProt Consortium et al., 2025). The full-length reference sequence and a sequence differing only by P463R were modelled separately with AlphaFold 3, retaining five predictions per allele (Abramson et al., 2024). Models were aligned using shared C α atoms with pLDDT $\geq$ 70 in both structures, excluding residues 457 to 472 around the focal loop. Core and focal-loop root mean square deviation, residue-wise C α displacement, solvent accessibility and minimum heavy-atom distances between residue 463 and non-neighbouring residues were calculated with the Bio.PDB module of Biopython (Cock et al., 2009); distances below 5 Å were classified as contacts. The highest-ranked prediction for each allele was rendered in PyMOL. The five predictions per allele were used to assess whether structural features recurred across model outputs and were not treated as biological replicates.

Statistical analysis

All tests were two-sided unless specified. Small group comparisons used Bonferroni correction, whereas genome-wide analyses used Benjamini-Hochberg FDR. Exact tests, model covariates, sample sizes and effect scales are reported with each analysis and in source tables. No significance threshold was applied to protein estimates in RNA-selected cross-omic lookup analyses.

Supplementary figures



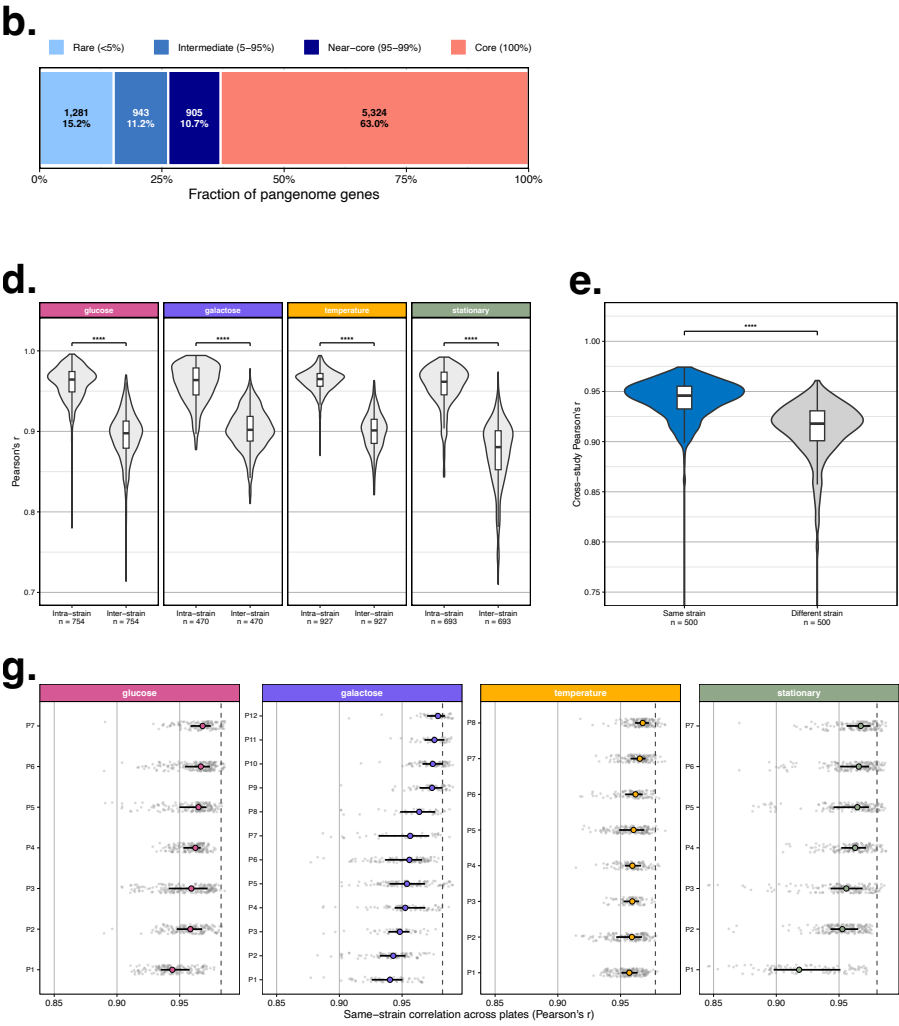


Figure S1 | Transcriptomic resource quality.

(a.) Distribution of mapped RNA-sequencing reads after filtering. (b.) Pangenome gene-frequency spectrum across the 536 isolates represented in the copy-number matrix. Core genes were present in all isolates; accessory genes were absent from at least one. (c.) Number of RNA samples available per strain. (d.) Within-strain RNA replicate correlations across environments. (e.) Cross-study transcript-abundance correlations across 500 matched glucose-grown isolates from the Caudal et al. dataset. (f.) RNA correlations in the independent XTRA_FGV validation set, including repeated sequencing of this Mexican Agave control strain. (g.) Plate-resolved same-strain transcriptome correlations across culture plates.

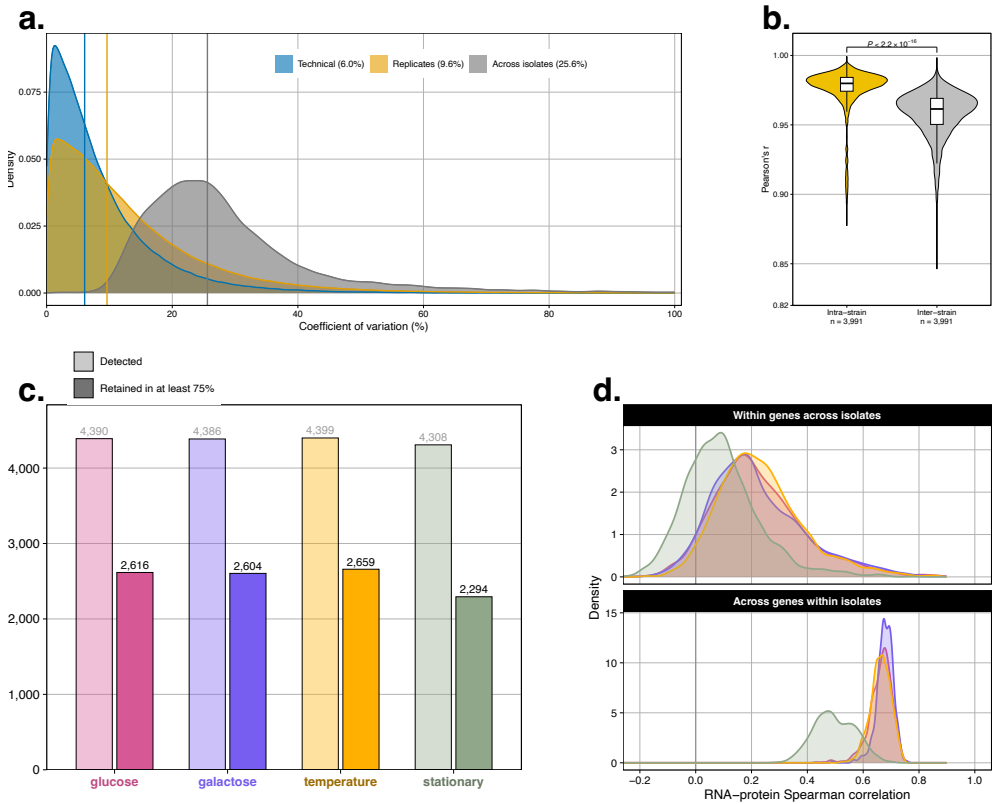


Figure S2 | Proteomic and cross-layer resource quality.

(a.) Coefficient-of-variation distributions across technical controls, within-strain biological replicates and the isolate population. (b.) Within-condition protein-abundance correlations for same-strain replicate pairs and equal-sized samples of randomly paired isolates. (c.) Numbers of proteins detected in each condition and retained at the 75% sample-completeness threshold used for condition-specific analyses. (d.) Transcript-protein correlations within genes across isolates and across genes within isolates, shown for each environment.

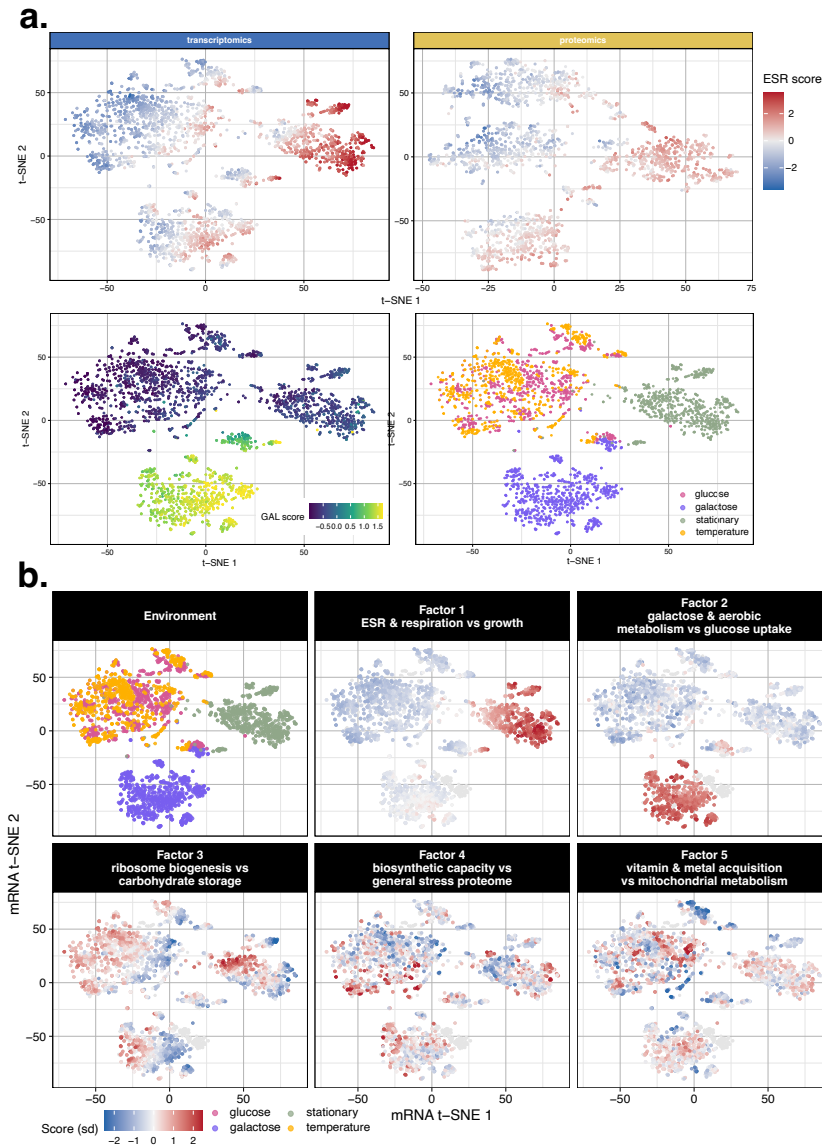


Figure S3 | Expression programmes in low-dimensional maps.

(a.) Transcriptome and proteome t-SNE maps coloured by environmental stress response activity, together with the transcriptome map coloured by GAL programme activation. Environmental stress response activity is the mean z score of the induced Gasch ESR genes minus the mean of the repressed PAC and ribosomal-protein genes; GAL activity is the mean z score of seven core galactose-pathway genes. (b.) The transcriptome t-SNE map used in Fig. 2, coloured by environment and the standardized scores of MOFA Factors 1–5. Factor colours are capped at the 99th percentile of the absolute standardized score.

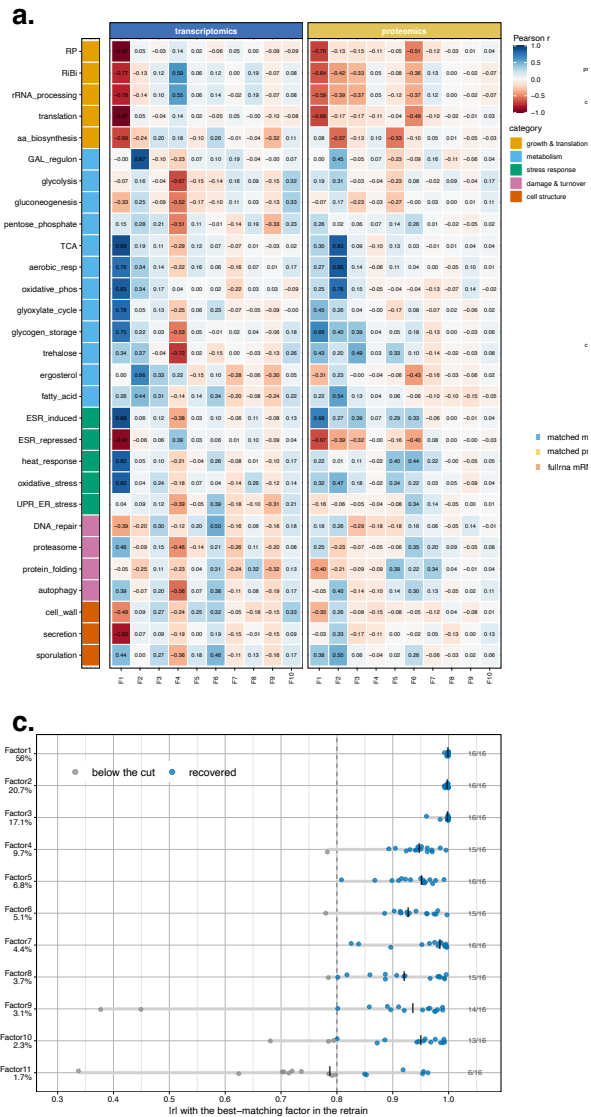
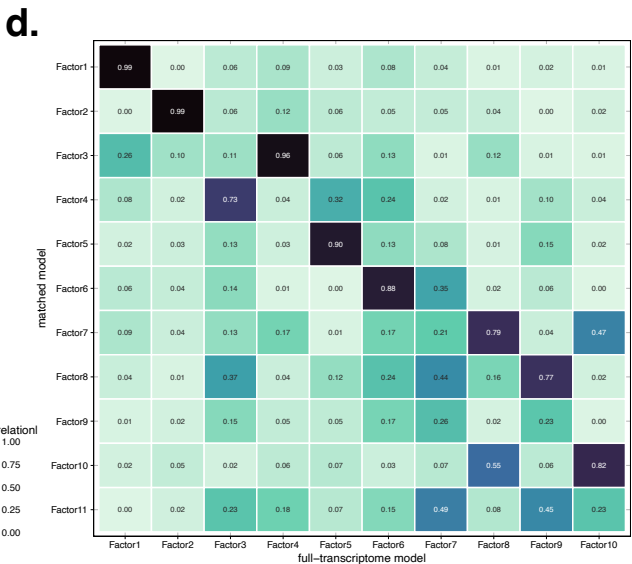
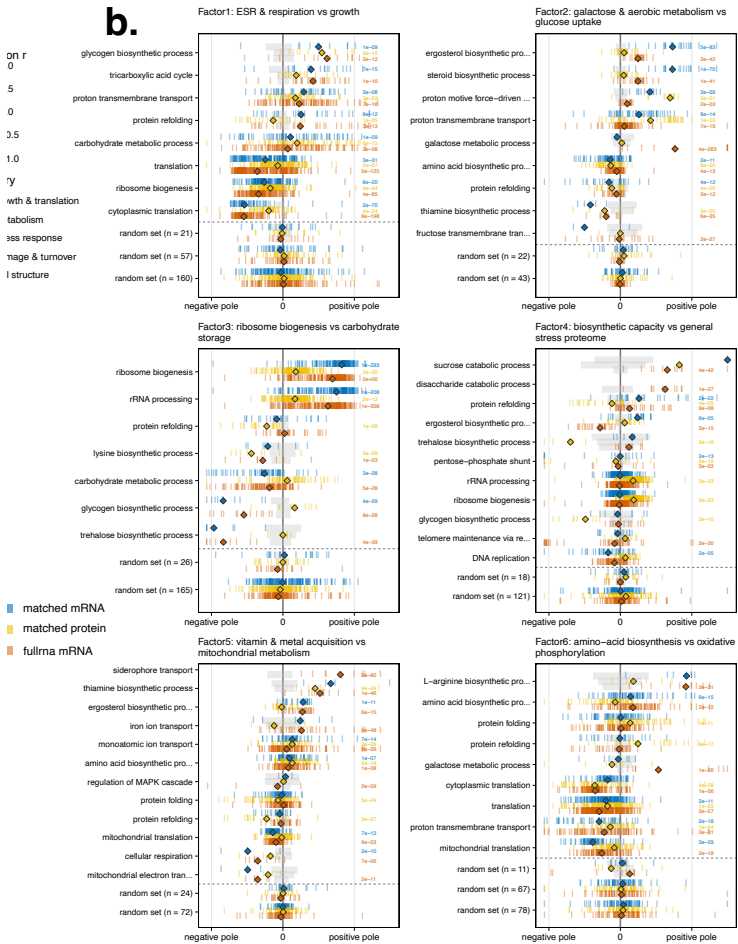


Figure S4 | MOFA factor interpretation and robustness.

(a.) Pearson correlations between MOFA factor scores and broad biological programme scores in matched transcript and protein data. (b.) Gene-set enrichment along the negative and positive loading poles of Factors 1 to 6 in matched transcripts, matched proteins and the full-transcriptome model, assessed against size-matched random gene sets. (c.) Stability of the 11 MOFA factors across 16 retrainings on random 85% isolate subsets. (d.) Absolute factor-score correlations between the matched multi-omic model and the full-transcriptome model across shared samples.



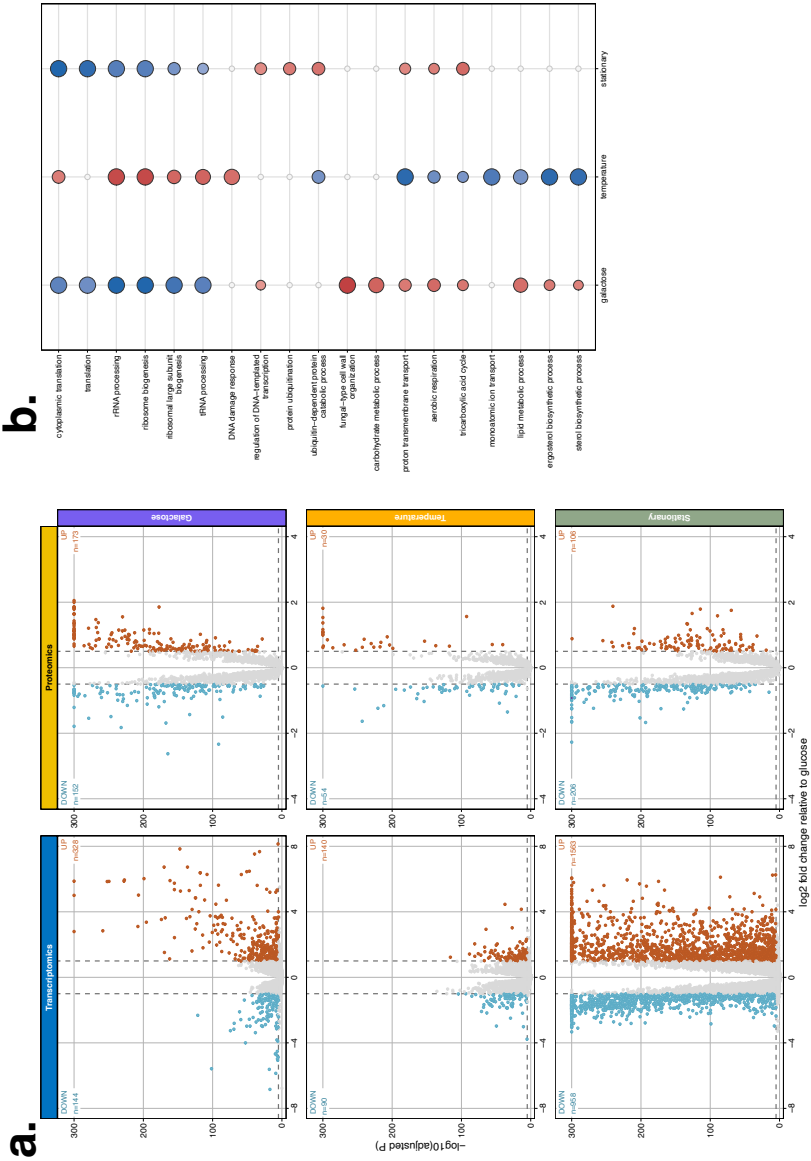
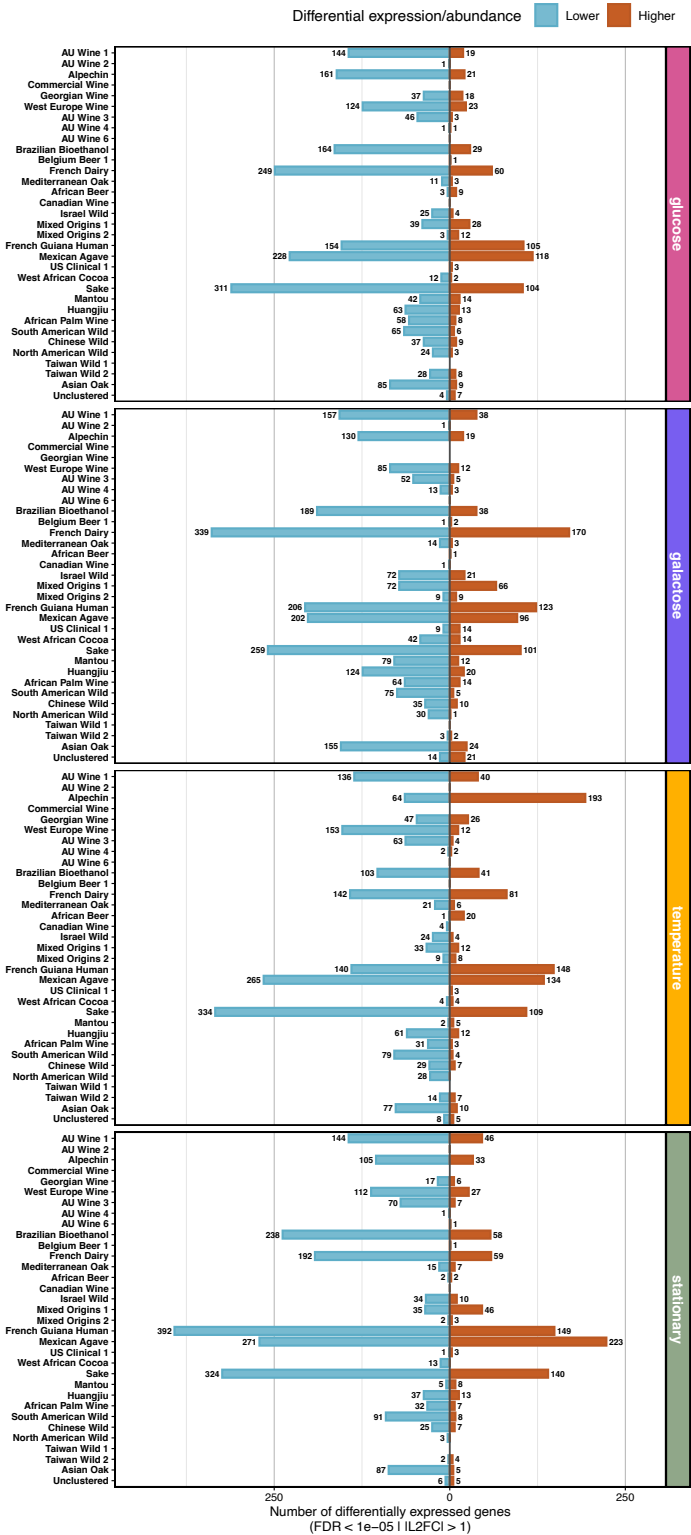


Figure S5 | Environmental and lineage differential expression.

(a.) Transcript differential-expression and protein differential-abundance volcano plots for galactose, elevated temperature and stationary phase relative to glucose; counts include all tested genes. (b.) Gene-set enrichment of transcript responses across the three environmental contrasts. (c.) Transcript differential-expression catalogue across clades and environments, using BH FDR $< 1 \times 10^{-5}$ and $|\log_2 \text{ fold change}| > 1$. (d.) Corresponding protein differential-abundance catalogue, using BH FDR < 0.05 and $|\log_2 \text{ fold change}| > 0.5$.

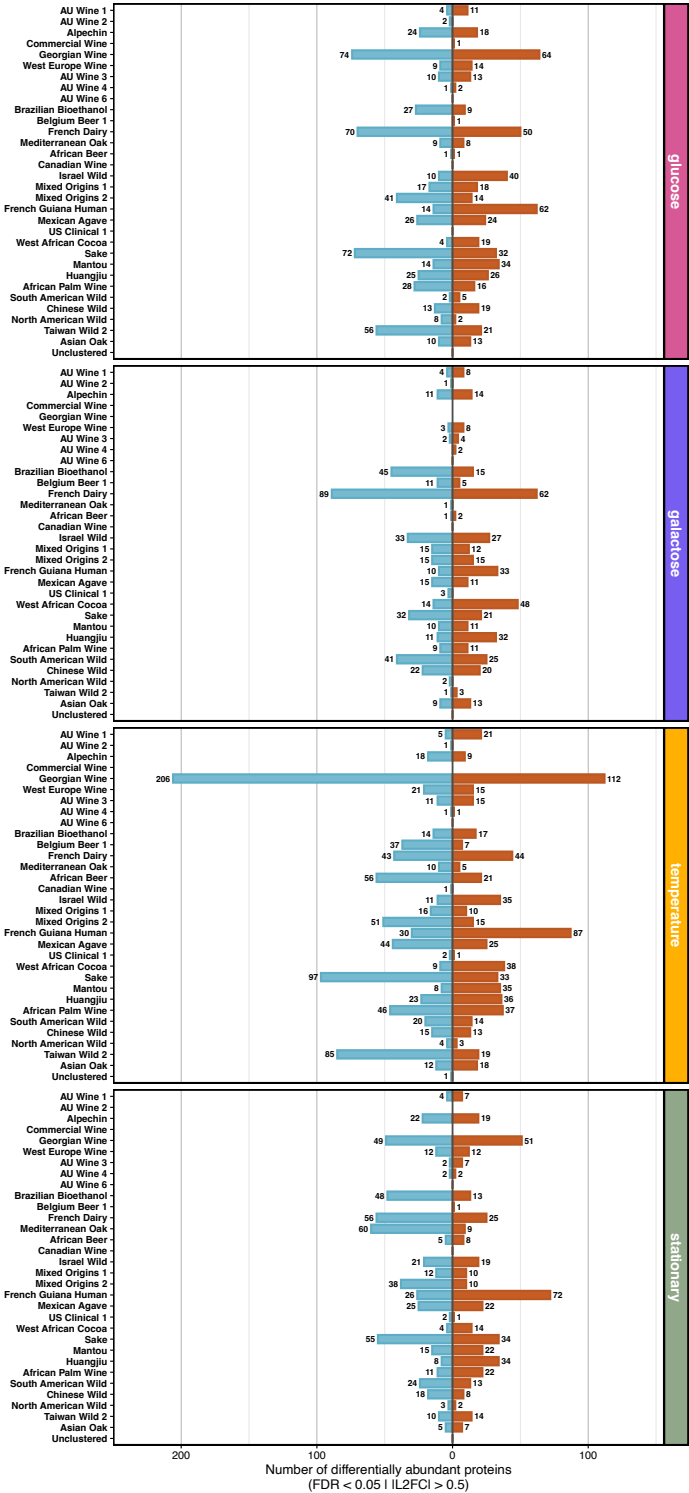
C.



d.

Differential expression/abundance

Lower Higher



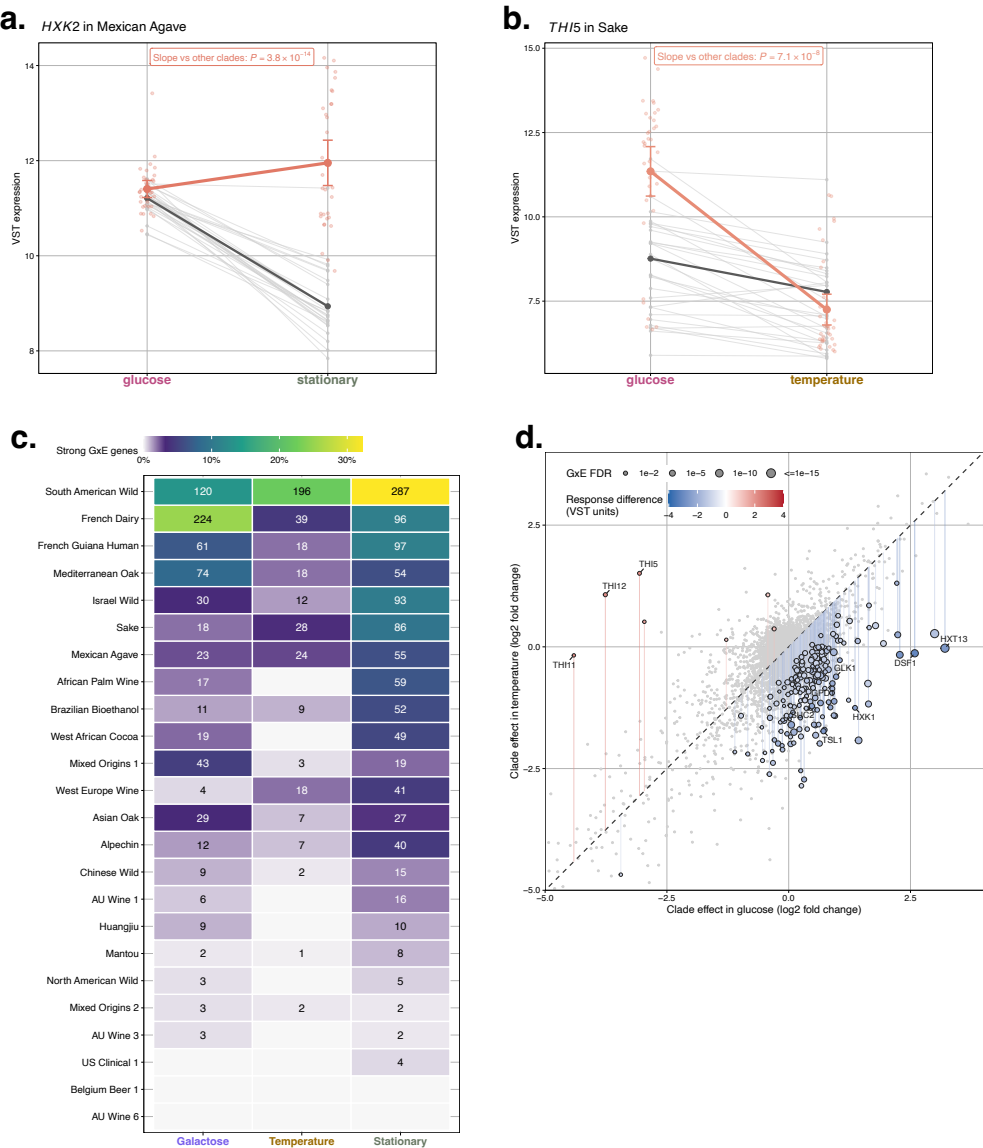


Figure S6 | Lineage GxE examples and extended atlas.

(a.) *HXK2* reaction norm in Mexican Agave. (b.) *THI5* reaction norm in Sake. (c.) Numbers of strong transcript GxE genes by clade and response contrast. (d.) Change in the South American Wild expression profile between glucose and 37°C, highlighting the 196 strong temperature GxE events.

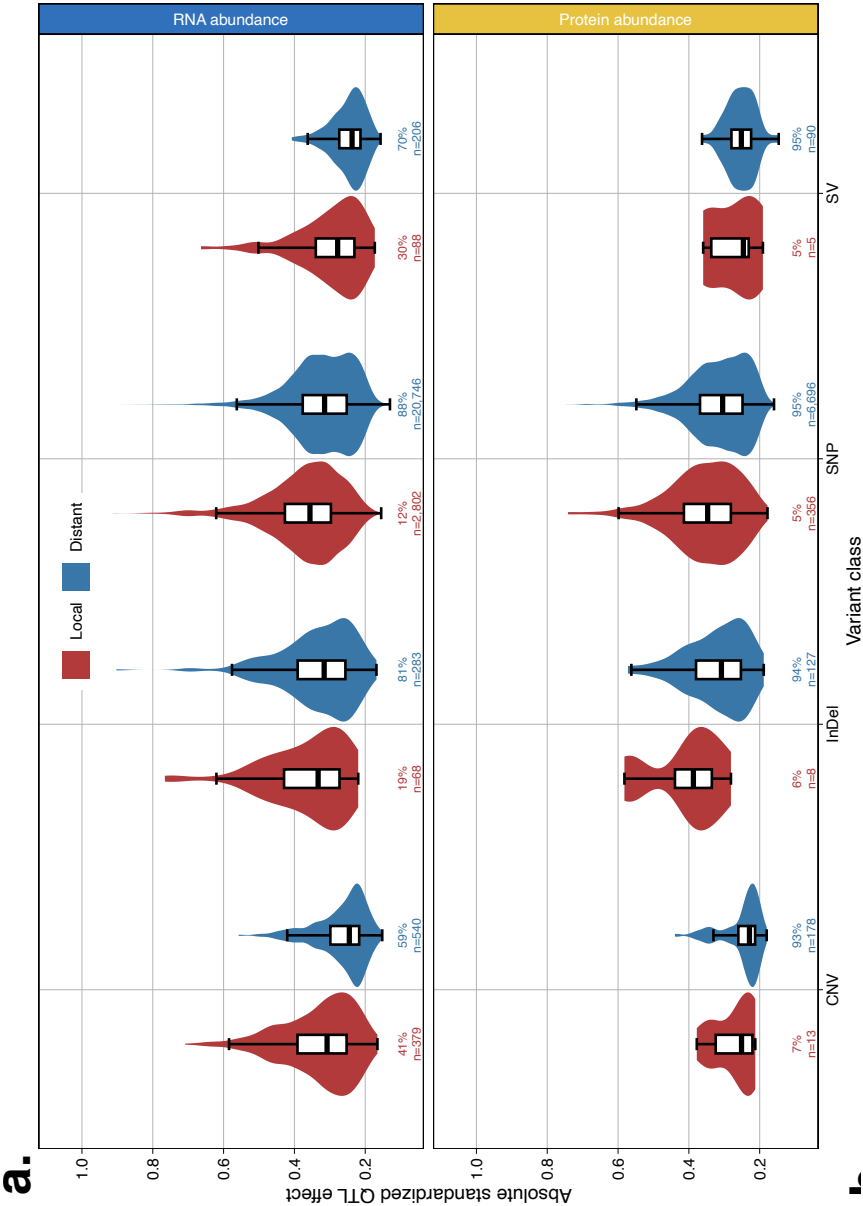
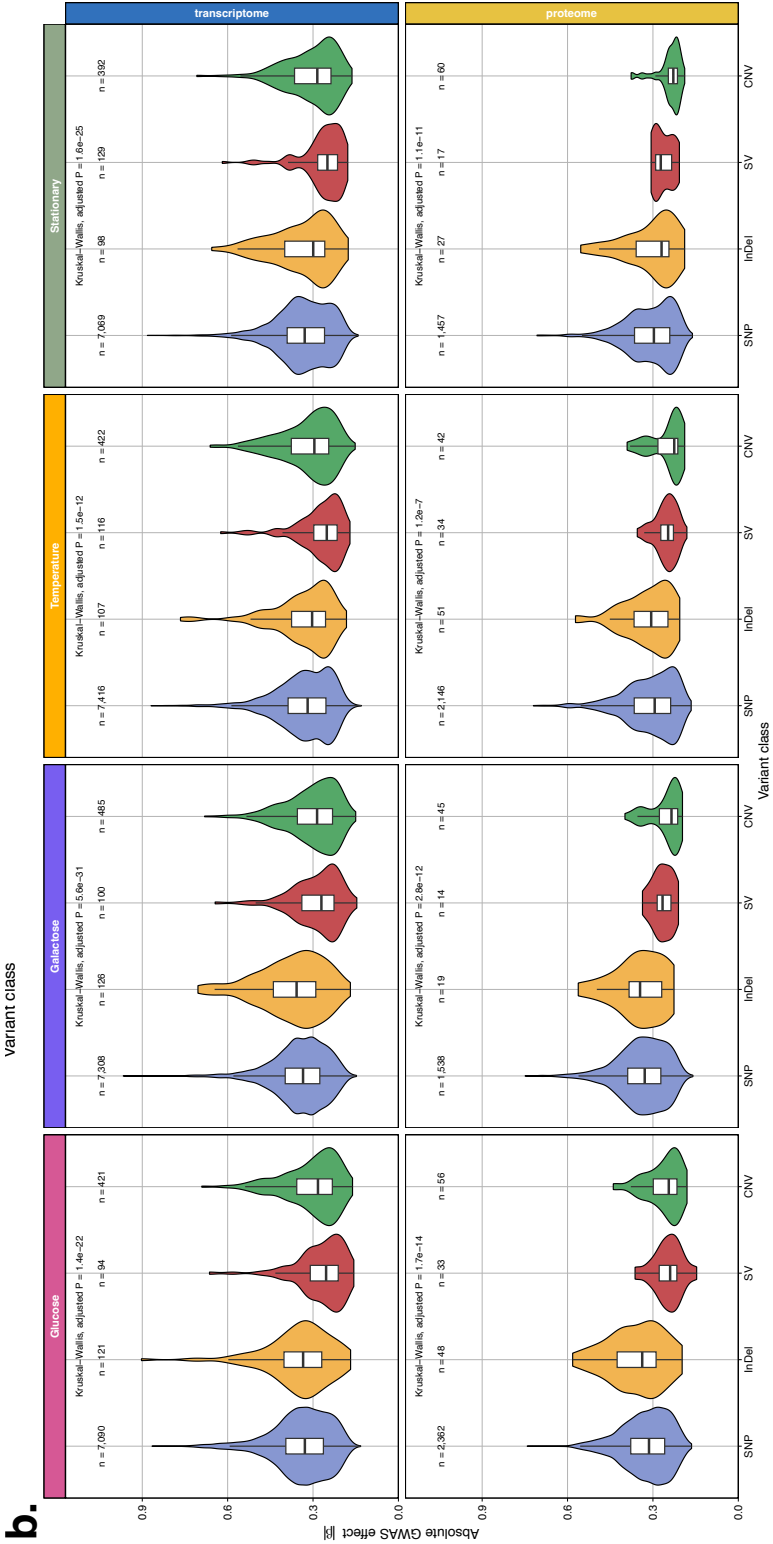


Figure S7 | Genome-wide QTL architecture.

(a.) Local and distant locus-target composition and standardized-effect distributions across variant classes and molecular layers. (b.) Standardized-effect distributions by variant class, molecular layer and environment. Facets report Bonferroni-adjusted Kruskal-Wallis tests across the four variant classes.



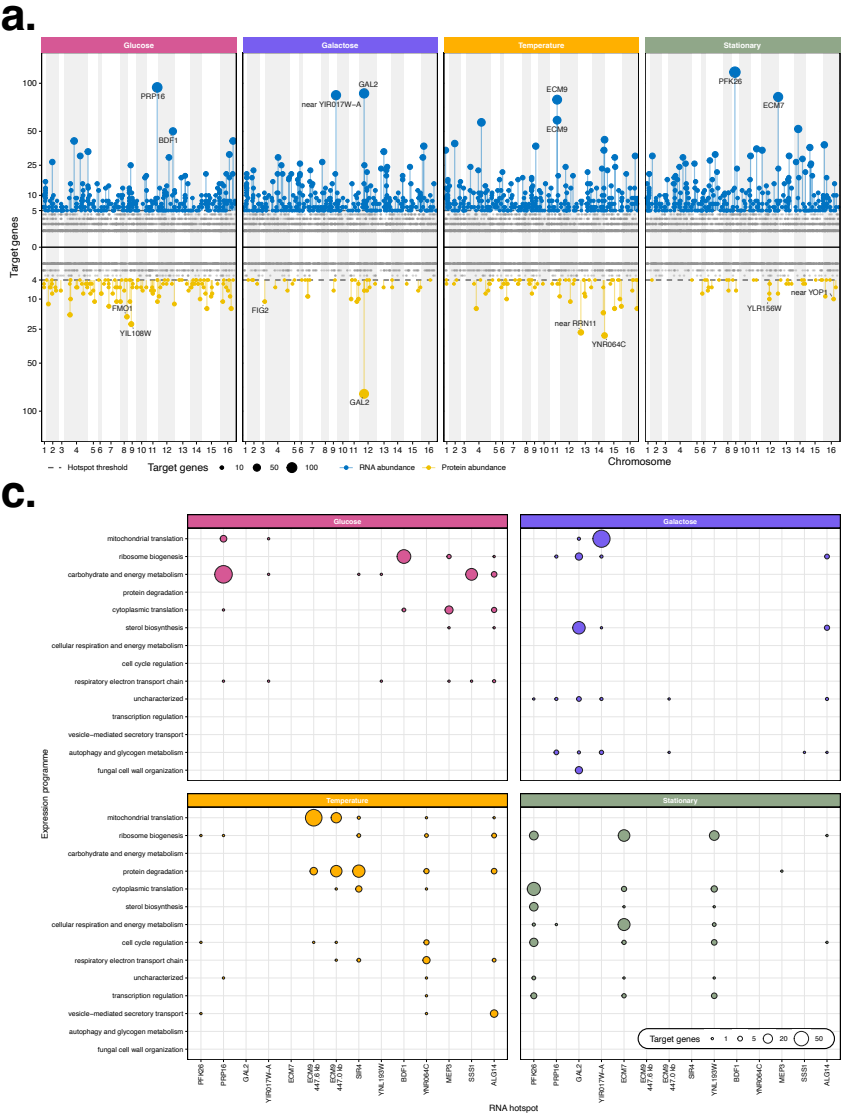
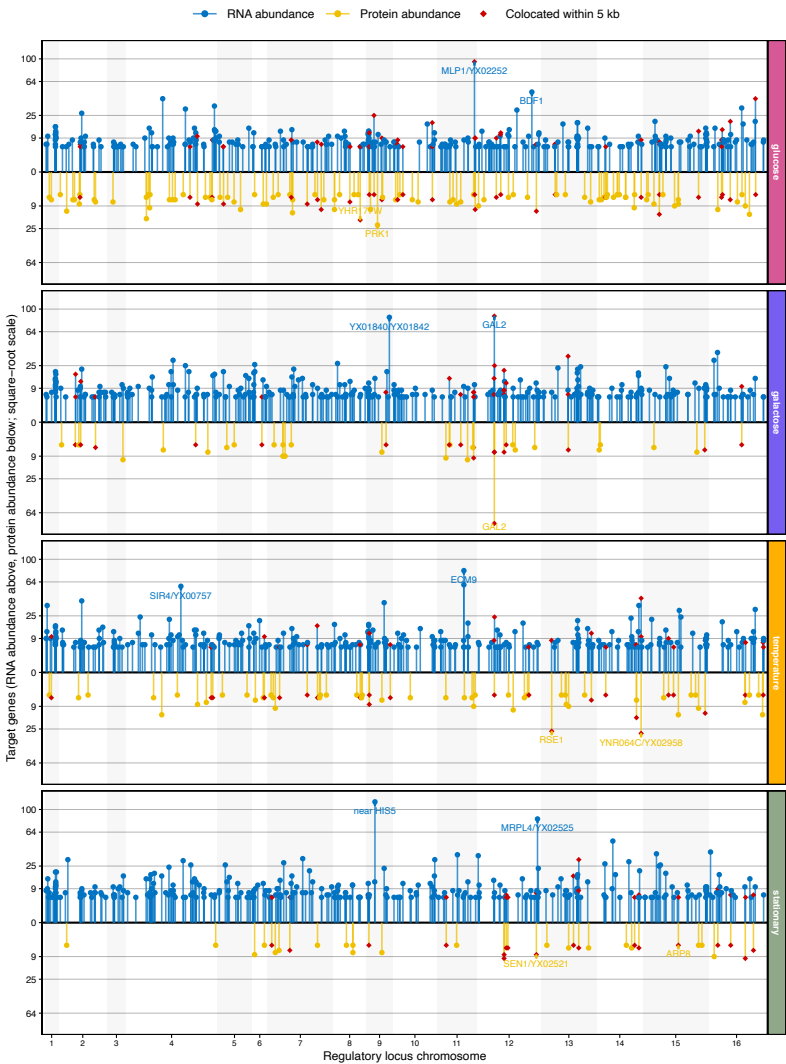
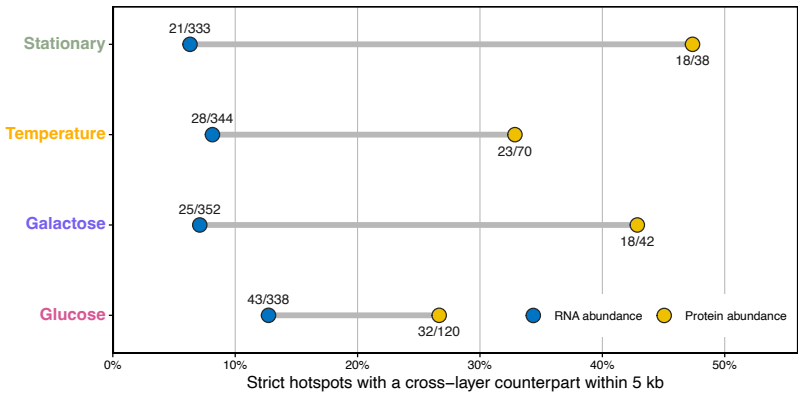


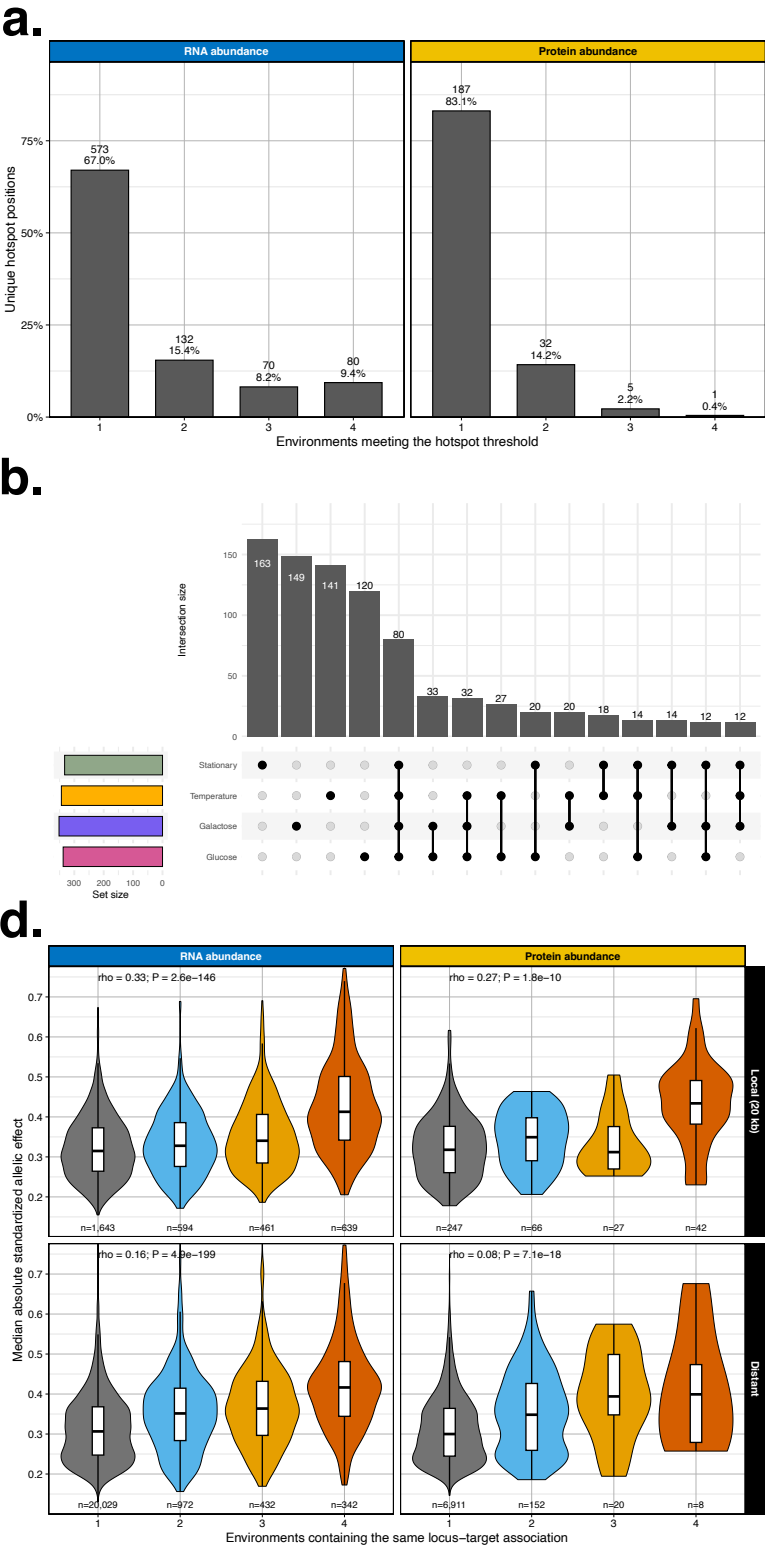
Figure S8 | Hotspot reach and regulatory programmes.

(a.) Genome-wide target reach of transcript and protein regulatory loci in each environment, with hotspots defined by at least five transcript targets or four protein targets. The two widest transcript hotspots at 37°C are led by variants 149 bp apart within *ECM9*. (b.) Cross-layer positional colocalization of hotspot instances whose lead positions lie within 5 kb in the same environment, without requiring the same lead variant or shared targets. (c.) Network linking regulatory loci, target programmes and environments. (d.) Hotspot sensitivity after restricting each environment to genes tested in both molecular layers and matching association counts over 1,000 resamplings under a common five-target cutoff.

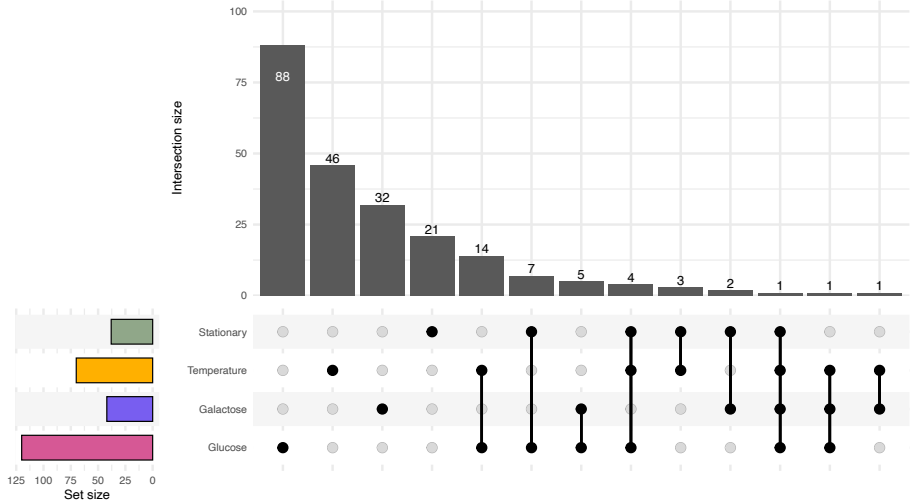
b.



3



C.



e.

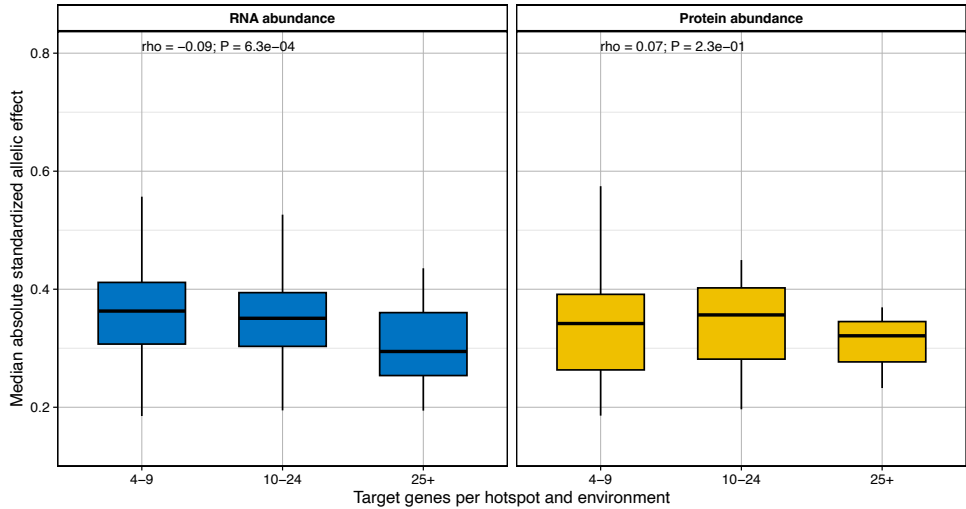


Figure S9 | Hotspot recurrence and effect magnitude.

(a.) Number of environments in which each unique transcript or protein hotspot position passed its layer-specific target-count threshold. (b.) Intersections of transcript hotspot positions across environments. (c.) Corresponding intersections for protein hotspot positions. (d.) Absolute locus-target effect size as a function of the number of environments in which the association was detected. (e.) Relationship between the number of targets per hotspot and median absolute effect.



210

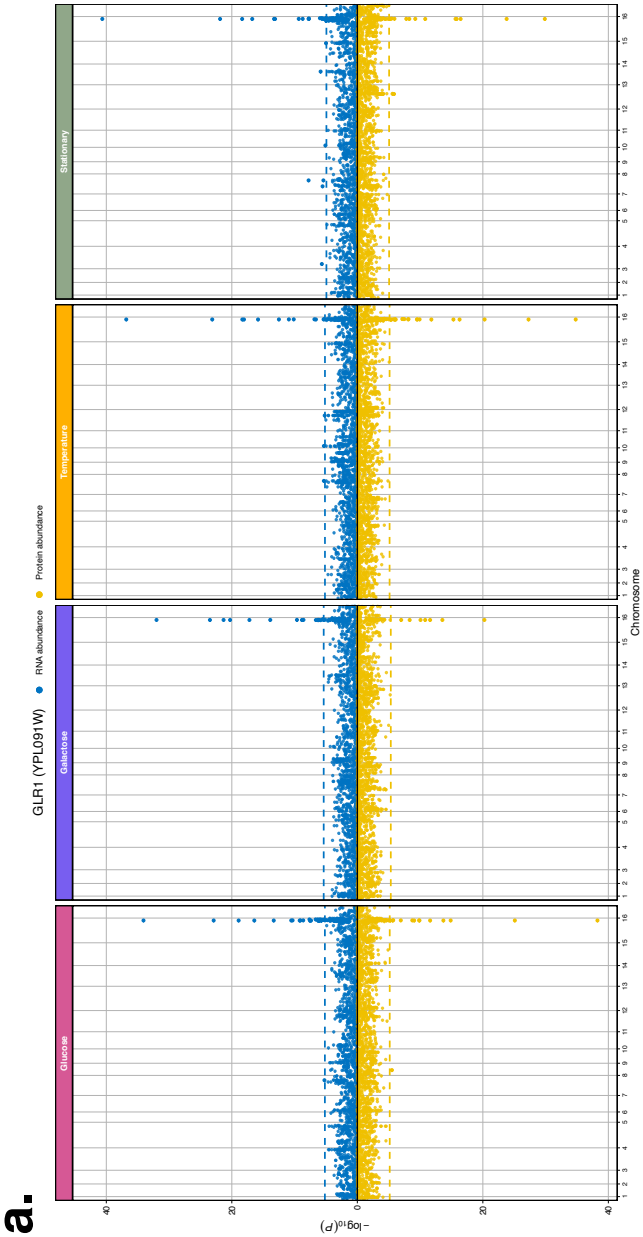
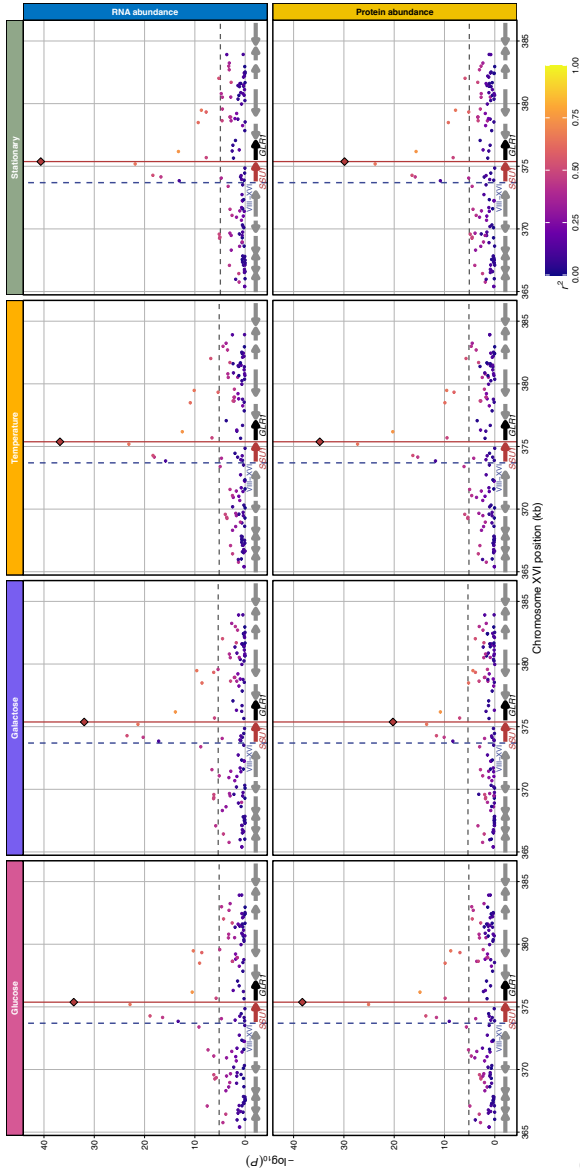


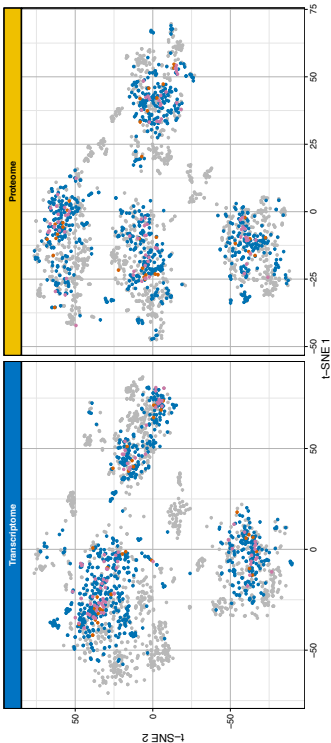
Figure S11 | Persistent cross-layer association at the *SSU1-GLR1* interval.

(a.) Miami plots of *GLR1* transcript and protein associations across four environments. Associations below $-\log_{10}(P) = 3$ were progressively subsampled. (b.) Regional *GLR1* association scans across four environments and two molecular layers, with linkage disequilibrium to rs_16_375375_T_A and the local gene map. (c.) Joint distribution of the rs_16_375375_T_A allele and the catalogued VIII-XVI *SSU1* translocation across transcriptome and proteome t-SNE maps.

b.



c.



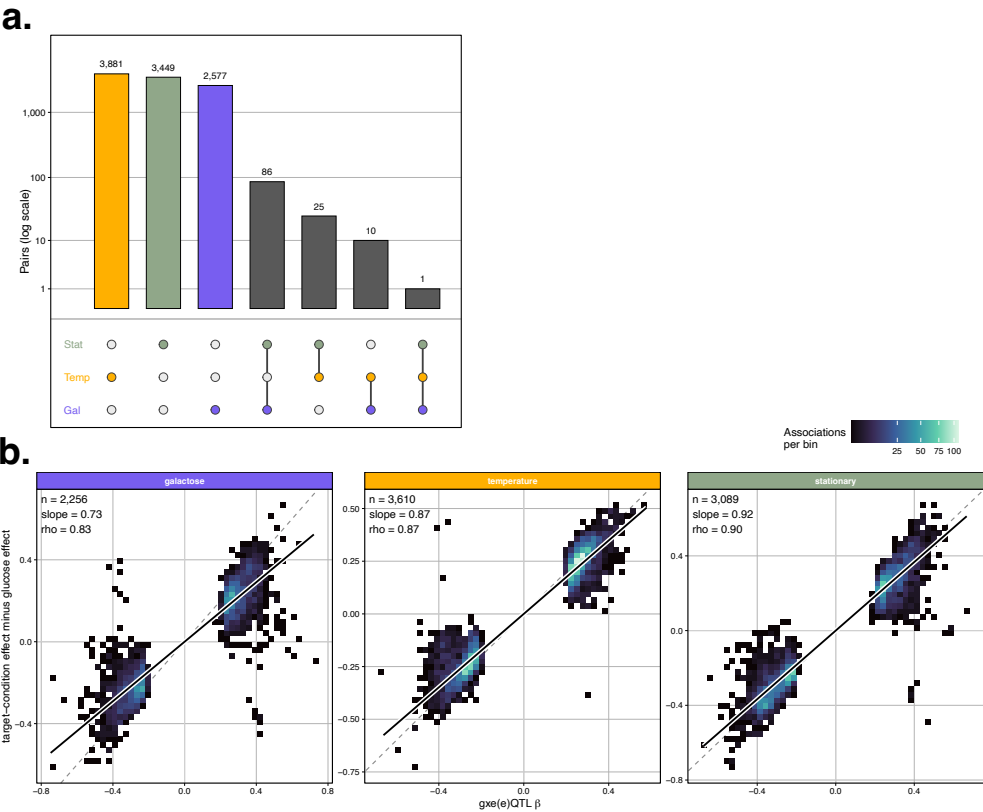


Figure S12 | gxe(e)QTL discovery and calibration.

(a.) Overlap of gxe(e)QTL locus-target pairs among the galactose-minus-glucose, elevated-temperature-minus-glucose and stationary-phase-minus-glucose responses. (b.) Calibration of the transcript-response gxe(e)QTL effects used in the cross-layer lookup, with the number of tests, origin-constrained slope and Spearman correlation.

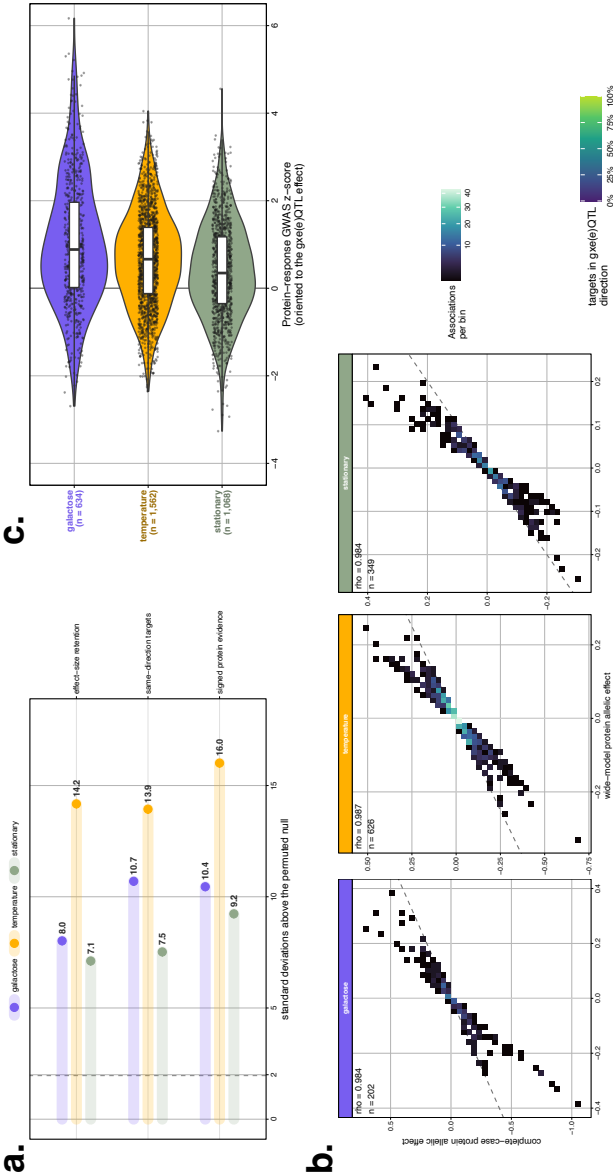
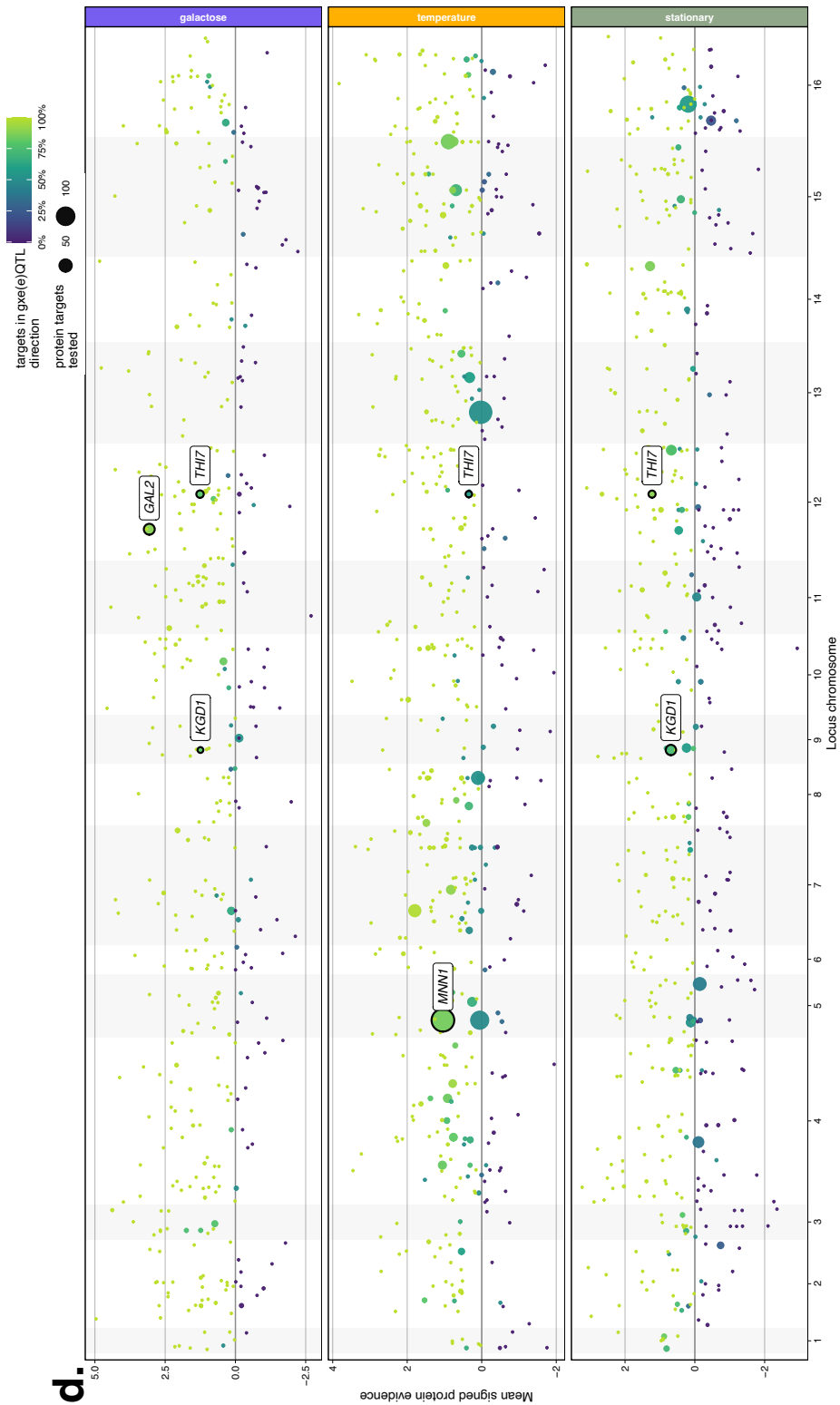


Figure S13 | Validation and genomic distribution of cross-layer response effects.

(a.) Cross-layer transmission statistics relative to gene-block sign-randomization nulls, expressed as standard deviations above the null. (b.) Agreement between pair-specific protein-response estimates and an alternative wide-matrix implementation. (c.) Protein-response evidence oriented to the corresponding transcript response QTL. (d.) Genome-wide locations of transcript-response QTL hotspots and their direction-oriented protein support.



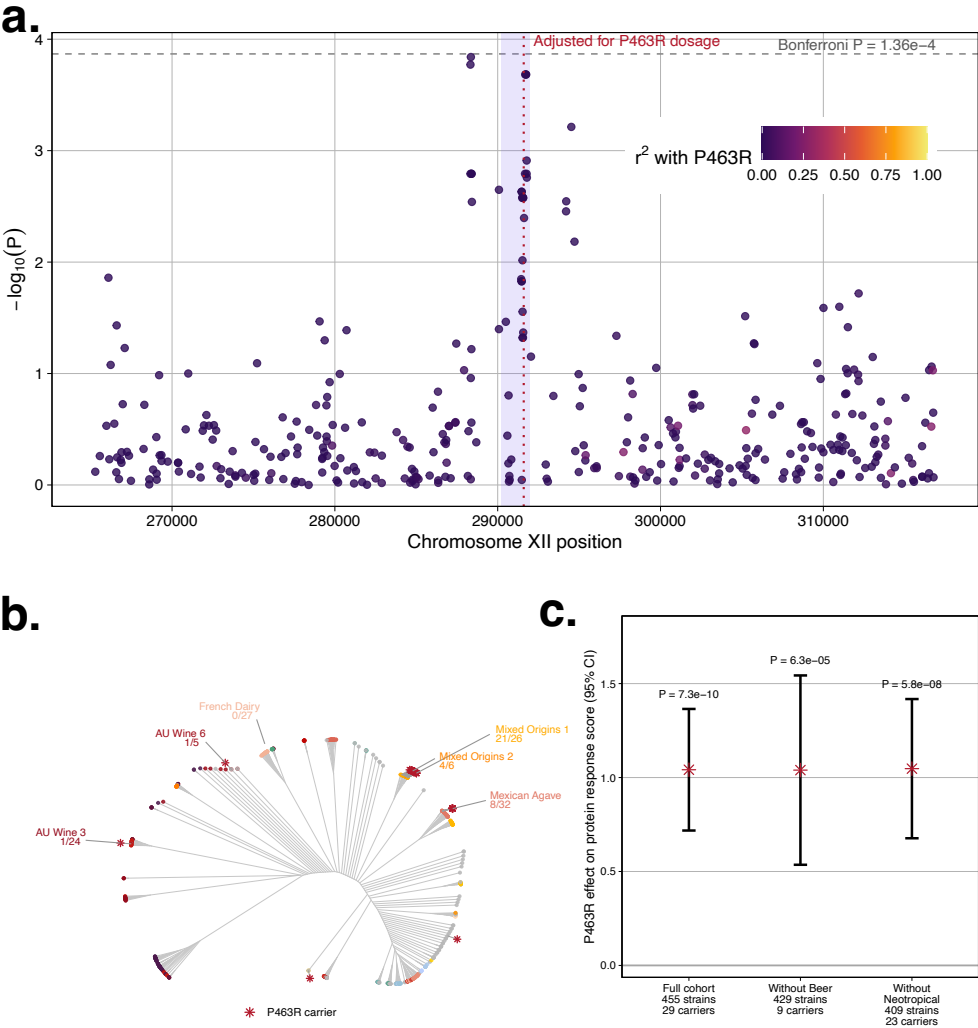
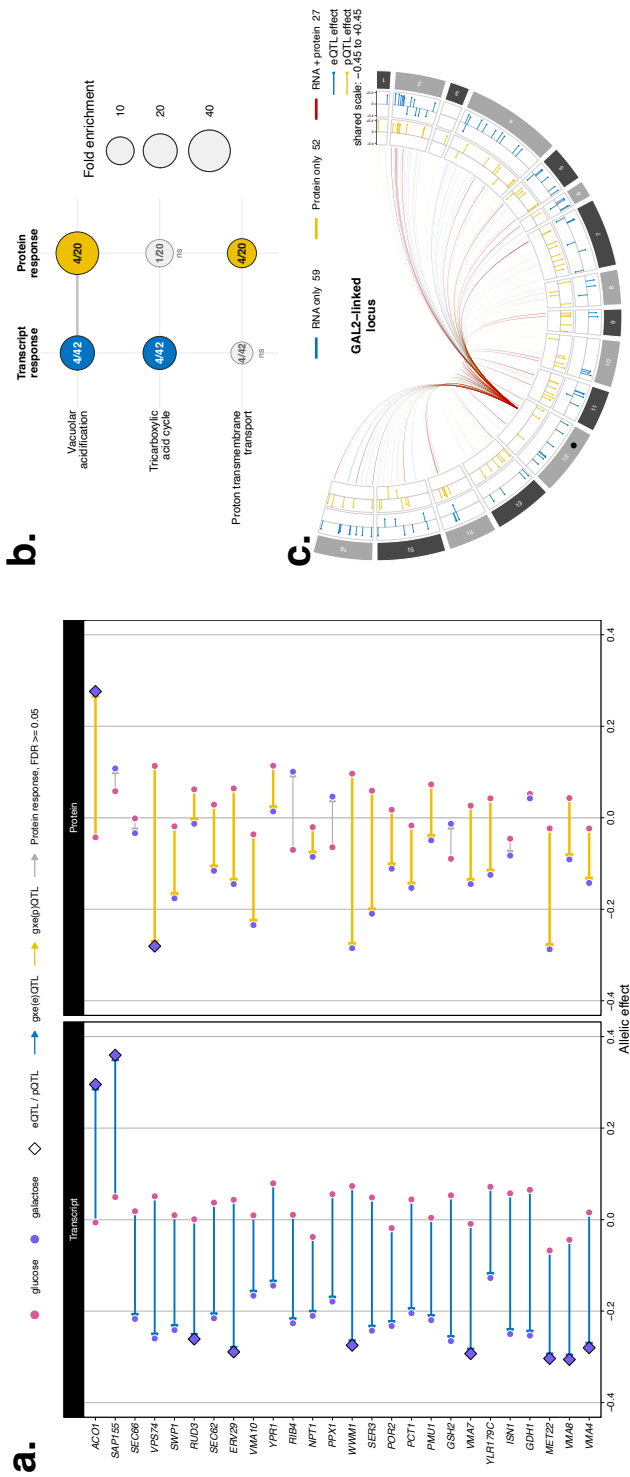


Figure S14 | *GAL2* locus and population context.

(a.) Association of local variants with the transcript-defined protein-response score after adjustment for P463R dosage, together with their linkage disequilibrium with P463R. (b.) Position of P463R carriers among the 571 isolates in the population tree. (c.) Additive P463R effects in the full cohort and after excluding either the Beer or Neotropical superclade, with 95% confidence intervals.



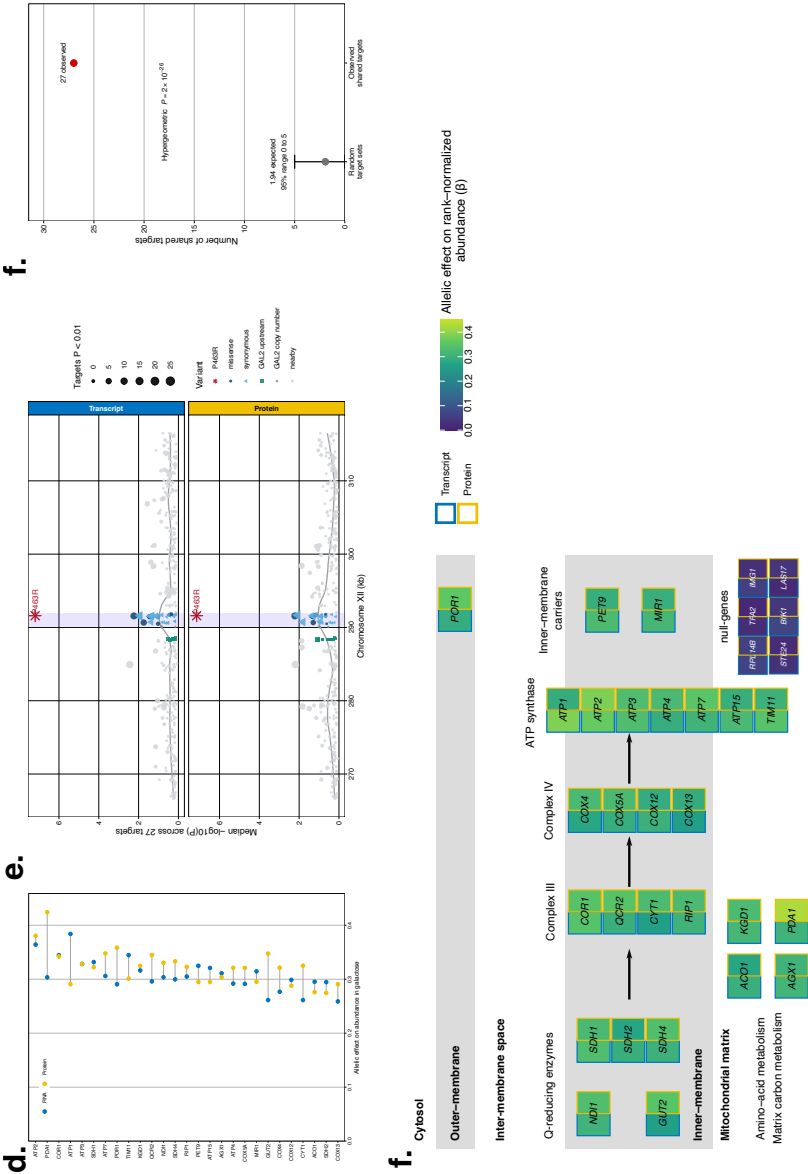


Figure S15 | *GAL2* response and abundance programmes.

(a.) Glucose and galactose allelic effects and condition-specific eQTL or pQTL status for the 26 transcript-response targets measured at protein level. (b.) Representative GO enrichments among *gxe*(e)QTL targets and the *gxe*(p)QTL subset. (c.) Genome-wide map of galactose-abundance targets linked to *GAL2*: 59 transcript-only, 52 protein-only and 27 shared coordinate-resolved targets. (d.) Allele-aligned transcript and protein effects for the 27 shared targets. (e.) Aggregate local association evidence across these 27 targets. (f.) Hypergeometric expectation and central 95% range for the eQTL-pQTL target overlap, compared with the 27 genes observed in both layers. (g.) Schematic of the 27 shared targets across mitochondrial respiration, matrix carbon metabolism and linked amino-acid metabolism. Six non-target genes with $P > 0.5$ in both layers provide near-zero references.

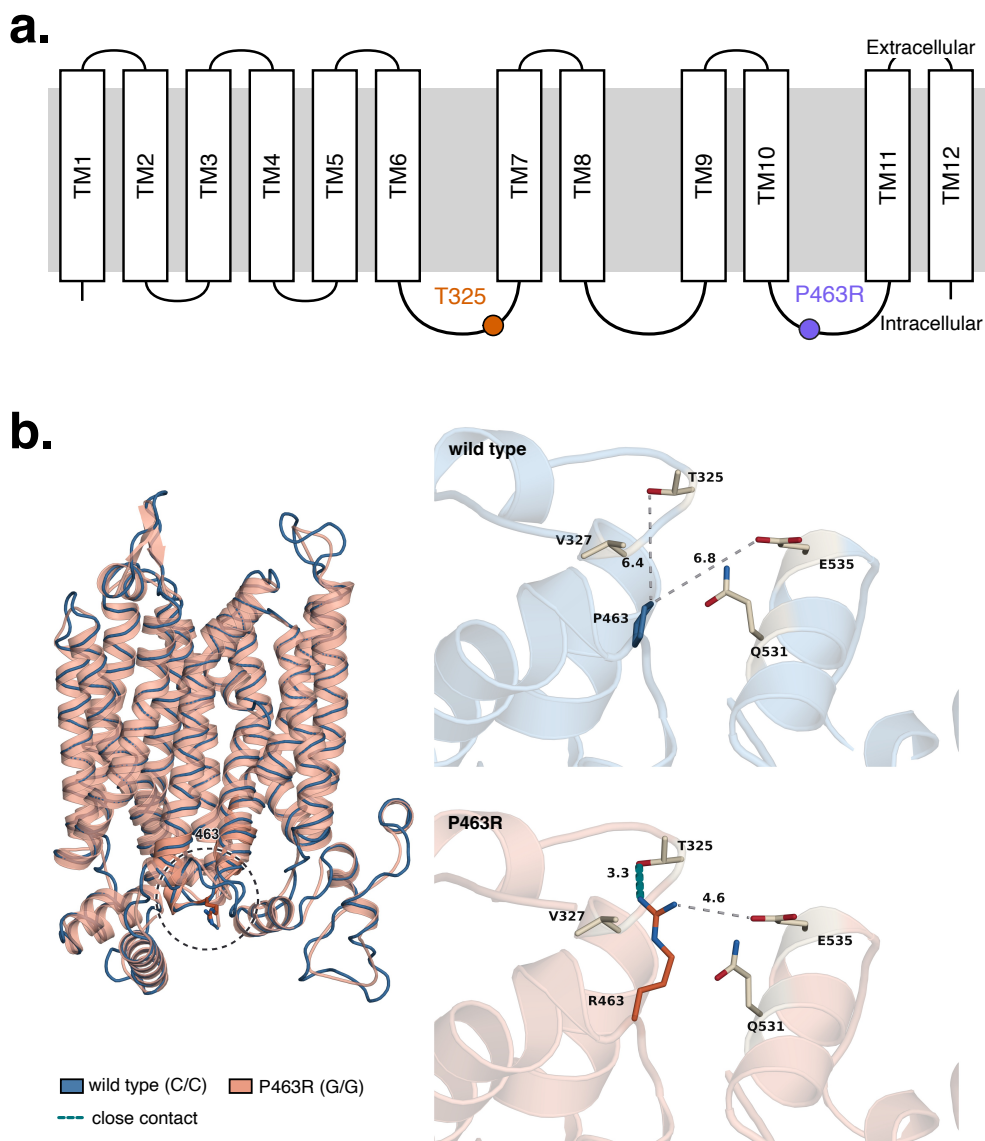


Figure S16 | Predicted structural context of Gal2 P463R.

(a.) UniProt topology annotation for Gal2 (P13181). T325 lies between TM6 and TM7, and residue 463 between TM10 and TM11. (b.) Superposition of the top-ranked AlphaFold 3 models of wild-type Gal2 P463 (C/C, blue) and Gal2 P463R (G/G, salmon). Selected residues highlight the predicted neighbourhood of position 463, including the 3.3-Å close contact between R463 and T325.

Data and code availability

The supplementary tables and data supporting this chapter are deposited on [Zenodo](#).

Author contributions

J.S. conceived and supervised the project. G.B. and M.G. performed the experiments. L.R.S., A.L. and D.L. generated the proteomic data. M.R. and J.S. contributed resources and analytical tools. G.B. analysed the data, with supervision from J.S. G.B. and J.S. wrote the manuscript.

Competing interests

The authors declare no competing interests.

References

- Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**: 493–500.
- Ahlmann-Eltze C, Huber W (2021) glmGamPoi: Fitting Gamma-Poisson generalized linear models on single cell count data. *Bioinformatics* **36**: 5701–5702.
- Albert FW, Bloom JS, Siegel J, Day L, Kruglyak L (2018) Genetics of trans-regulatory variation in gene expression. *eLife* **7**: e35471.
- Albert FW, Kruglyak L (2015) The role of regulatory variation in complex traits and disease. *Nat Rev Genet* **16**: 197–212.
- Albert FW, Treusch S, Shockley AH, Bloom JS, Kruglyak L (2014) Genetics of single-cell protein abundance variation in large yeast populations. *Nature* **506**: 494–497.
- Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. (2020) MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* **21**: 111.
- Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. (2018) Multi-Omics Factor Analysis—a Framework for Unsupervised Integration of Multi-omics Data Sets. *Mol Syst Biol* **14**: e8124.
- Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using **lme4**. *J Stat Softw* **67**.
- Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, et al. (2015) Impact of regulatory variation from RNA to protein. *Science* **347**: 664–667.
- Blazanin M (2024) Gcplyr: An R package for microbial growth curve data analysis. *BMC Bioinformatics* **25**: 232.
- Bolstad BM, Irizarry RA, Åstrand M, Speed TP (2003) A comparison of normalization methods for high densityoligonucleotide array data based on variance and bias. *Bioinformatics* **19**: 185–193.
- Boock J, Sadhu MJ, Durvasula A, Bloom JS, Kruglyak L (2021) Ancient balancing selection maintains incompatible versions of the galactose pathway in yeast. *Science* **371**: 415–419.
- Brach G, Schacherer J (2026) From one genome to thousands, and beyond. *FEMS Yeast Res* **26**: foag038.
- Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* **296**: 752–755.
- Brion C, Lutz SM, Albert FW (2020) Simultaneous quantification of mRNA and protein in single cells reveals post-transcriptional effects of genetic variation. *eLife* **9**: e60645.
- Buccitelli C, Selbach M (2020) mRNAs, proteins and the emerging principles of gene expression control. *Nat Rev Genet* **21**: 630–644.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. (2009) BLAST+: Architecture and applications. *BMC Bioinformatics* **10**: 421.
- Caudal É, Loegler V, Dutreux F, Vakirlis N, Teyssonnière É, Caradec C, et al. (2024) Pan-transcriptome reveals a large accessory genome contribution to gene expression variation in yeast. *Nat Genet* **56**: 1278–1287.

- Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, et al. (2001) Remodeling of Yeast Genome Expression in Response to Environmental Changes. *Mol Biol Cell* **12**: 323–337.
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**.
- Chen S, Zhou Y, Chen Y, Gu J (2018) Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**: i884–i890.
- Chick JM, Munger SC, Simecek P, Huttlin EL, Choi K, Gatti DM, et al. (2016) Defining the consequences of genetic variation on a proteome-wide scale. *Nature* **534**: 500–505.
- Christiano R, Nagaraj N, Fröhlich F, Walther Tobias C (2014) Global Proteome Turnover Analyses of the Yeasts *S. cerevisiae* and *S. pombe*. *Cell Rep* **9**: 1959–1965.
- Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. (2009) Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**: 1422–1423.
- Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M (2014) Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Mol Cell Proteomics* **13**: 2513–2526.
- Dahl A, Nguyen K, Cai N, Gandal MJ, Flint J, Zaitlen N (2020) A Robust Method Uncovers Significant Context-Specific Heritability in Diverse Complex Traits. *Am J Hum Genet* **106**: 71–91.
- Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. (2021) Twelve years of samtools and bcftools. *Gigascience* **10**:giab008.
- Demichev V, Messner CB, Vernardis SI, Lilley KS, Ralser M (2020) DIA-NN: Neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* **17**: 41–44.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. (2013) STAR: Ultrafast Universal RNA-seq Aligner. *Bioinformatics* **29**: 15–21.
- Engel SR, Aleksander S, Nash RS, Wong ED, Weng S, Miyasato SR, et al. (2025) *Saccharomyces* Genome Database: Advances in genome annotation, expanded biochemical pathways, and other key enhancements. *Genetics* **229**: iyae185.
- EWELS P, Magnusson M, Lundin S, KÄLLER M (2016) MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**: 3047–3048.
- Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, et al. (2014) Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. *Science* **343**: 1246949.
- Foss EJ, Radulovic D, Shaffer SA, Ruderfer DM, Et. A (2007) Genetic basis of proteome variation in yeast. *Nat Genet*.
- Gagneur J, Stegle O, Zhu C, Jakob P, Tekkedil MM, Aiyar RS, et al. (2013) Genotype-Environment Interactions Reveal Causal Pathways That Mediate Genetic Effects on Phenotype. *PLoS Genet* **9**: e1003803.
- Gasch AP (2026) Twenty-Five Years of the Environmental Stress Response and the Enduring Power of Yeast in Stress Biology. *Yeast*: yea.70028.
- Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, et al. (2000) Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Mol Biol Cell* **11**: 4241–4257.
- Grossbach J, Gillet L, Clément-Ziza M, Schmalohr CL, Schubert OT, Schütter M, et al. (2022) The impact of genomic variation on protein phosphorylation states and regulatory networks. *Mol Syst Biol* **18**: MSB202110712.

- GTEx Consortium (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**: 1318–1330.
- Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between Protein and mRNA Abundance in Yeast. *Mol Cell Biol* **19**: 1720–1730.
- Hodgins-Davis A, Adomas AB, Warringer J, Townsend JP (2012) Abundant Gene-by-Environment Interactions in Gene Expression Reaction Norms to Copper within *Saccharomyces cerevisiae*. *Genome Biol Evol* **4**: 1061–1079.
- Horak J, Wolf DH (1997) Catabolite inactivation of the galactose transporter in the yeast *Saccharomyces cerevisiae*: Ubiquitination, endocytosis, and degradation in the vacuole. *J Bacteriol* **179**: 1541–1549.
- Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12**: 115–121.
- Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, et al. (2001) Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. *Science* **292**: 929–934.
- Lea AJ, Peng J, Ayroles JF (2022) Diverse environmental perturbations reveal the evolution and context-dependency of genetic effects on gene expression levels. *Genome Res*: genome;gr.276430.121v1.
- Lee KB, Wang J, Palme J, Escalante-Chong R, Hua B, Springer M (2017) Polymorphisms in the yeast galactose sensor underlie a natural continuum of nutrient-decision phenotypes. *PLOS Genet* **13**: e1006766.
- Lee MN, Ye C, Villani A-C, Raj T, Li W, Eisenhaure TM, et al. (2014) Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells. *Science* **343**: 1246980.
- Legras J-L, Galeote V, Bigey F, Camarasa C, Marsit S, Nidelet T, et al. (2018) Adaptation of *S. cerevisiae* to Fermented Food Environments Reveals Remarkable Genome Plasticity and the Footprints of Domestication. *Mol Biol Evol* **35**: 1712–1727.
- Liao Y, Smyth GK, Shi W (2014) featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinforma Oxf Engl*.
- Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011) FaST linear mixed models for genome-wide association studies. *Nat Methods* **8**: 833–835.
- Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, et al. (2009) Population genomics of domestic and wild yeasts. *Nature* **458**: 337–341.
- Liu Y, Beyer A, Aebersold R (2016) On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**: 535–550.
- Loegler V, Friedrich A, Schacherer J (2024) Overview of the *Saccharomyces cerevisiae* population structure through the lens of 3,034 genomes. *G3 GenesGenomesGenetics* **14**: jkae245.
- Loegler V, Thiele P, Teyssonnière E, Tsouris A, Brach G, Cruaud C, et al. (2025) From genotype to phenotype with 1,086 near telomere-to-telomere yeast genomes. *Nature*: 1–10.
- Love MI, Huber W, Anders S (2014) Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2. *Genome Biol* **15**: 550.
- Maaten L van der (2014) Accelerating t-SNE using Tree-Based Algorithms. *J Mach Learn Res* **15**: 3221–3245.
- Maaten L van der, Hinton G (2008) Visualizing Data using t-SNE. *J Mach Learn Res* **9**: 2579–2605.
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. (2009) Finding the missing heritability of complex diseases. *Nature* **461**: 747–753.
- Marguerat S, Schmidt A, Codlin S, Chen W, Aebersold R, Bähler J (2012) Quantitative Analysis of Fission Yeast Transcriptomes and Proteomes in Proliferating and Quiescent Cells. *Cell* **151**: 671–683.

- Martin-Perez M, Villén J (2017) Determinants and Regulation of Protein Turnover in Yeast. *Cell Syst* **5**: 283–294.e5.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–1303.
- Menzel P, Ng KL, Krogh A (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun* **7**: 11257.
- Messner CB, Demichev V, Bloomfield N, Yu JSL, White M, Kreidl M, et al. (2021) Ultra-fast proteomics with Scanning SWATH. *Nat Biotechnol* **39**: 846–854.
- Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. (2021) Sustainable data analysis with Snakemake. *F1000Research* **10**: 33.
- Muenzner J, Trébulle P, Agostini F, Zauber H, Messner CB, Steger M, et al. (2024) Natural proteome diversity links aneuploidy tolerance to protein turnover. *Nature* **630**: 149–157.
- Parts L, Liu Y-C, Tekkedil MM, Steinmetz LM, Caudy AA, Fraser AG, et al. (2014) Heritability and genetic basis of protein level variation in an outbred population. *Genome Res* **24**: 1363–1370.
- Pérez-Ortín JE, Querol A, Puig S, Barrio E (2002) Molecular Characterization of a Chromosomal Rearrangement Involved in the Adaptive Evolution of Yeast Strains. *Genome Res* **12**: 1533–1539.
- Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, et al. (2018) Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **556**: 339–344.
- Phipson B, Lee S, Majewski IJ, Alexander WS, Smyth GK (2016) Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Ann Appl Stat* **10**.
- Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. (2015) Limma Powers Differential Expression Analyses for RNA-sequencing and Microarray Studies. *Nucleic Acids Res* **43**: e47–e47.
- Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L (2009) Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* **458**: 342–345.
- Siddiq MA, Duveau F, Wittkopp PJ (2024) Plasticity and environment-specific relationships between gene expression and fitness in *Saccharomyces cerevisiae*. *Nat Ecol Evol* **8**: 2184–2194.
- Smith EN, Kruglyak L (2008) Gene–Environment Interaction in Yeast Gene Expression. *PLOS Biol* **6**: e83.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* **102**: 15545–15550.
- Teyssonnière EM, Trébulle P, Muenzner J, Loegler V, Ludwig D, Amari F, et al. (2024) Species-wide quantitative transcriptomes and proteomes reveal distinct genetic control of gene expression variation in yeast. *Proc Natl Acad Sci* **121**: e2319211121.
- The Gene Ontology Consortium, Carbon S, Douglass E, Good BM, Unni DR, Harris NL, et al. (2021) The Gene Ontology resource: Enriching a GOLD mine. *Nucleic Acids Res* **49**: D325–D334.
- The UniProt Consortium, Bateman A, Martin M-J, Orchard S, Magrane M, Adesina A, et al. (2025) UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* **53**: D609–D617.
- Tsouris A, Brach G, Friedrich A, Hou J, Schacherer J (2024) Diallel panel reveals a significant impact of low-frequency genetic variants on gene expression variation in yeast. *Mol Syst Biol* **20**: 362–373.

- Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era – concepts and misconceptions. *Nat Rev Genet* **9**: 255–266.
- Vogel C, Marcotte EM (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **13**: 227–232.
- Warringer J, Forsmark A, Durbin R, Omholt SW, Louis EJ, Liti G, et al. (2011) Trait Variation in Yeast Is Defined by Population History. *PLoS Genet* **7**: 15.
- Wood DE, Lu J, Langmead B (2019) Improved metagenomic analysis with Kraken 2. *Genome Biol* **20**: 257.
- Wu L, Candille SI, Choi Y, Xie D, Jiang L, Li-Pook-Than J, et al. (2013) Variation and genetic control of protein abundance in humans. *Nature* **499**: 79–82.
- Xu S, Hu E, Cai Y, Xie Z, Luo X, Zhan L, et al. (2024) Using clusterProfiler to characterize multiomics data. *Nat Protoc* **19**: 3292–3320.
- Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, et al. (2003) Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat Genet* **35**: 57–64.
- Zhu A, Ibrahim JG, Love MI (2019) Heavy-tailed prior distributions for sequence count data: Removing the noise and preserving large differences. *Bioinformatics* **35**: 2084–2092.

Discussion and perspectives

Using yeast as a powerful model system to study the genotype–phenotype relationship.

From genotype to phenotype through molecular layers

The *S. cerevisiae* reference genome turned thirty this year. The 1996 sequence of S288C revealed what the species was made of: about six thousand genes, most of them without a known function at the time (Goffeau et al., 1996). Functional genomics closed much of that gap, gene by gene, while advances in sequencing replaced the single reference with a catalogue of thousands of natural isolates (Peter et al., 2018; Loegler et al., 2025). The question is no longer what the species is made of, but how its diversity arises and how two individuals with similar genotypes develop different phenotypes.

Genotype and phenotype are connected through successive molecular layers, and transcription is among the earliest that can be measured systematically. It is one of the first measurable manifestations of genetic variation and a phenotype in its own right. Transcript abundance can be measured for every gene in every individual and captures a substantial component of the heritable variation between individuals. If some of the missing heritability of complex traits has molecular sources, these may be detectable at this layer before they become apparent further downstream. Yeast is a powerful model system for studying these genetic effects: more than 1,000 natural isolates have been sequenced with both Illumina short reads and ONT long reads, providing a species-wide collection with known genotypes and phylogeny, as well as sufficient phenotypic diversity for mapping.

My doctoral work approached the genotype–phenotype relationship from three angles, with the transcriptome as the common axis. First, we described how an entire class of genetic variation, structural variants (SVs), contributes to variation in gene expression. Second, we examined the polyadenylation landscape across six genetic backgrounds using nanopore direct RNA sequencing, with a particular focus on poly(A) tails. Finally, we investigated environmental effects on transcriptomes and proteomes at the species level.

Looking back at the three angles

1. Effects of structural and low-frequency variants on the transcriptome

Structural and low-frequency variants are difficult to associate with phenotypes for different reasons. First, association studies have limited power for variants that occur at low frequencies in the population. In the 1,011 genomes collection, 92.7% of polymorphic sites have a MAF below 5%, so most of the genetic variation in the species is poorly captured by conventional GWAS (Peter et al., 2018). A diallel panel partly overcomes this limitation by design. Here, we used a half-diallel of 26 parental isolates and obtained high-quality transcriptomes for 323 hybrids. A variant carried by a single parent is present in every hybrid involving that parent (Fournier et al., 2019). Second, because SVs are difficult to identify across large populations, they are often excluded from association studies. We therefore performed long-read sequencing and genome assembly of the 26 parental genomes to obtain a catalogue that included SVs alongside SNPs and InDels. We then used these variants in association studies, with the expression levels of close to 6,000 genes as molecular phenotypes (Tsouris et al., 2024a).

Low-frequency variants in the natural population made up 3.0% of the SNPs tested but accounted for 8.4% of the SNP-eQTL, a nearly threefold enrichment. A similar enrichment had already been established for growth (Bloom et al., 2019; Fournier et al., 2019). Their effect sizes, however, appeared smaller than those of common variants, in contrast to the growth data. One possible explanation is the culture medium: the transcriptomes were collected under a single, permissive, nutrient-rich condition, so a variant that matters in another environment may have no detectable effect. This hypothesis was one motivation for conducting the third study across several environments.

The long-read assemblies revealed 1,953 SVs in the parental genomes, but 60% were singletons carried by a single parent. These had to be removed because variants private to the same parent have identical segregation patterns across the diallel and cannot be distinguished statistically. After filtering, 763 SVs remained alongside 31,909 SNPs. We detected 20 SV-eQTL (19 unique SVs), compared with 3,039 SNP-eQTL (1,032 unique SNPs). Normalised by the number of variants tested, the two classes were associated at almost the same rate: 2.5% of SVs and 3.2% of SNPs were eQTL for at least one gene. However, the two classes were not equally represented. SVs made up 2% of the genotype matrix, compared with 98% for SNPs, and the significance threshold was set by the total number of variants tested. The SNPs therefore raised the bar that an SV had to clear to be detected as significantly associated. Furthermore, hybrid genotypes were inferred from the parents rather than sequenced directly, so aneuploidies, LOH events or rearrangements arising after the cross were absent from the matrix even if their effects were reflected in the transcriptome. Consistent with this possibility, 117 of the original 351 hybrids showed expression bias across entire chromosomes or large chromosomal segments. The diallel design nevertheless recovered signals from low-frequency variants and provided evidence that SVs are associated with variation in gene expression.

2. The transcript layer at full length

Gene-level RNA abundance does not capture variation in transcript structure. Cleavage determines transcript ends and initiates polyadenylation. Deadenylation then remodels poly(A)-tail length. Both features can influence RNA stability, translation, export and localisation (Mayr, 2016; Tian & Manley, 2017; Passmore & Collier, 2022). Because most genome-wide maps of yeast 3' ends derive from S288C or related laboratory strains, whether these ends are conserved across natural populations remained unclear. We therefore applied Oxford Nanopore direct RNA sequencing to six genetically diverse isolates and aligned the reads to strain-specific, near-telomere-to-telomere assemblies (O'Donnell et al., 2023; Loegler et al., 2025). This enabled joint, single-molecule quantification of gene expression, transcript-end use, poly(A)-tail length and transcript-model diversity while reducing reference-mapping bias.

Gene expression was concordant among isolates (Pearson correlations, 0.908–0.962), and global median tail lengths ranged from 27 to 33 nt. Variance decomposition attributed 7.1% of read-level tail-length variation to gene identity and 1.0% to strain; the remaining 91.8% included gene-by-strain structure and read-level variation. Genes therefore had characteristic tail lengths despite extensive within-gene heterogeneity. Although pooled tests detected cross-isolate differences for 5,405 of 5,505 genes, the median difference between the most divergent strain means was only 8.6 nt, so these statistical differences generally had small effect sizes. The inverse association between expression and tail length was reproduced across all isolates and is consistent with, but does not demonstrate, co-translational deadenylation (Subtelny et al., 2014; Passmore & Collier, 2022).

Transcript-end usage showed greater isolate specificity. At equalised depth, 23.7–26.7% of expressed genes carried at least two operational polyadenylation islands. Although median end position was often conserved (median cross-strain range, 11.9 nt), 644 of 905 genes with matched island classes met the predefined criteria for differential usage, including 104 complete matched-site switches. TES-island identity was associated with tail-length distributions in 27.7–54.5% of tested multi-island genes, but aggregate proximal–distal differences did not exceed 1 nt, arguing against a common directional coupling between cleavage position and tail length. We also identified 191 high-quality candidate novel-model calls across 75 canonical host genes, frequently involving intron retention.

Transcript composition therefore varies among isolates beyond differences in gene-level abundance, but its genetic causes and functional consequences remain unknown. Six isolates without hybrids cannot support variant mapping or distinguish *cis*- from *trans*-acting effects; island calls remained depth-dependent, and direct RNA sequencing resolves 5' ends incompletely. Larger panels and allele-specific measurements in hybrids are needed to map the determinants of 3'-end use, followed by measurements of RNA stability, translation and protein abundance to establish their consequences (Mittleman et al., 2020; Li et al., 2021).

3. Matched transcriptomes and proteomes across four environments

The first two studies expanded the usual scope of association studies by considering structural variants and transcript structure. The third examined what is missed when molecular phenotypes are measured in a single environment and at a single layer. We generated 2,721 transcriptomes and 2,668 proteome samples across glucose, galactose, elevated temperature and early stationary phase, with matched measurements from the same cultures for 555 isolates. These matched samples allowed us to compare environmental responses across the diversity of the species and genetic effects at the RNA and protein levels.

Environment was not a minor perturbation of a stable regulatory architecture. The four conditions produced distinct molecular states, and population history further structured how isolates responded to them. In particular, French Dairy isolates showed a constitutive galactose-expression state, whereas South American Wild isolates showed a stronger transcriptional response to elevated temperature. The associations were also environment-dependent: 86.3% of transcript-level and 95.8% of protein-level locus–target pairs were detected in only one environment, and very few persisted across all four. A genetic effect observed in glucose should therefore not be treated as a fixed property of an allele. The environment affects whether that effect is detected and which molecular targets are associated with it.

The comparison between layers led to a related conclusion. Exact overlap between significant eQTL and pQTL was low, ranging from 2.1% to 5.2% within an environment, but this did not mean that transcript effects vanished at protein level. When effects selected at the transcript level were examined in the proteome, 72.5% to 86.6% of abundance effects and 63.0% to 75.6% of environmental response effects retained the same direction. Local associations were shared more often than distant ones, suggesting that direct effects propagate between layers more reliably than effects mediated through regulatory networks. The degree of retention also varied among environments and was weakest in stationary phase. Translation, protein degradation, complex assembly and growth dilution could all contribute to this variation, but the present steady-state measurements cannot separate their respective roles.

The galactose-specific *GAL2* locus illustrates both the value and the limits of resolving intermediate phenotypes. Its P463R-linked haplotype was associated with a broad transcript response involving vacuolar acidification and carbon metabolism, and 23 of the 26 measured protein responses followed the same direction. Yet the locus was not associated with the six available growth traits. This coherent molecular response may be buffered before affecting growth, or it may affect physiological traits that were not measured. More generally, a single condition, molecular layer or final phenotype captures only a fraction of the paths connecting genotype to phenotype. Additional time points would be needed to distinguish delayed transmission from genuine buffering and to understand how environment-dependent effects unfold.

Going further

A larger diallel: sequencing hybrids rather than parents

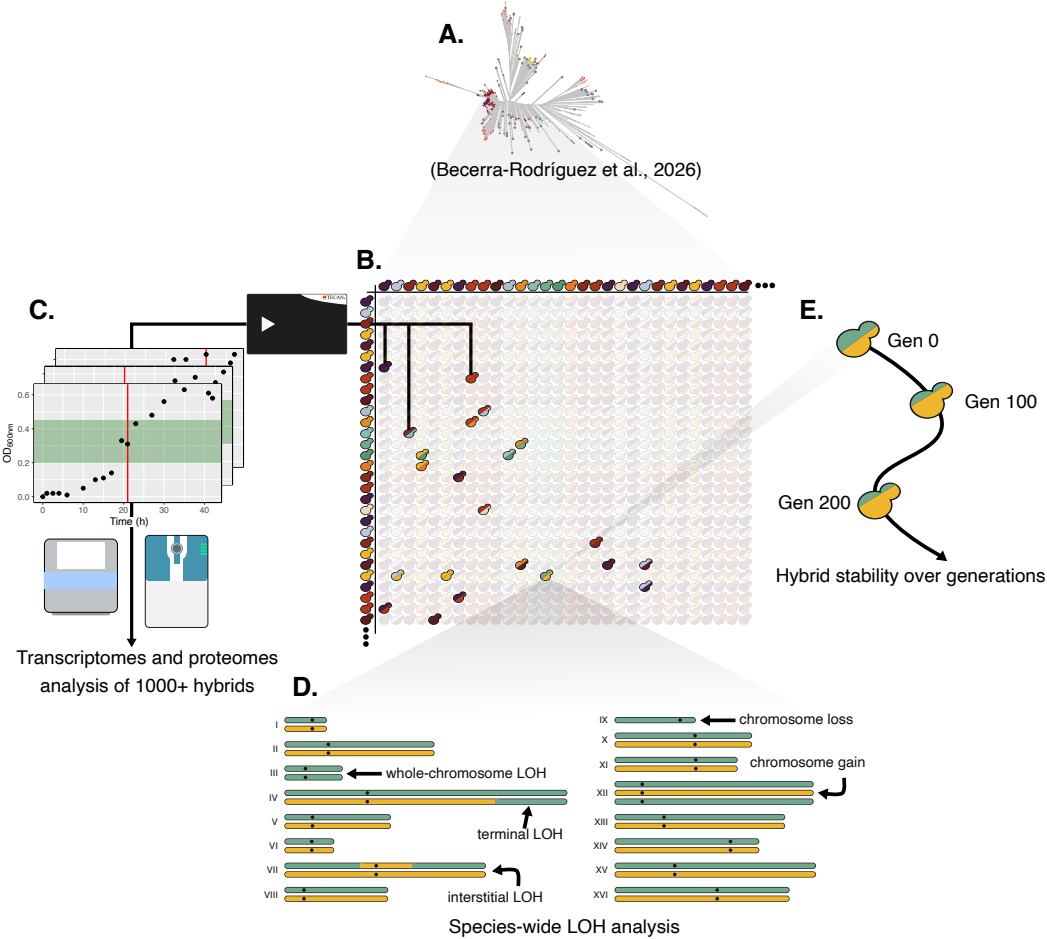
One improvement to our diallel scheme would be to increase the number of parents and sequence the hybrid genomes directly rather than infer them from the parents. The laboratory now has a collection of 364 stable haploid strains derived from natural isolates, available in both mating types and in prototrophic and *ura3* flavours (Becerra-Rodríguez et al., 2026). This collection covers most of the diversity of the species, with only a few clades unrepresented. Each background also carries a 20-bp barcode in place of its *HO* locus.

Crossing so many parents and harvesting every hybrid is out of reach, as it would represent more than 66,000 individuals. Crossing half of the parents in a sparse pairwise design, after removing many of the nearly clonal wine isolates, could produce around a thousand hybrids while covering far more genetic diversity than the 26 parents we used (Fig. 1.) In a small, complete diallel, each parent is represented in many hybrids, so its variants remain linked and LD stays high. Distributing the same number of hybrids across many more parents could reduce this structure, while a rare variant would still occur in every hybrid involving its parent of origin.

Isolating a single hybrid from each cross would make every strain the descendant of a unique hybridisation event, allowing differences from its two parental genomes to be assigned to that lineage. Sequencing the hybrids could reveal these differences directly, including aneuploidies, ploidy changes, terminal or interstitial LOH and rearrangements. Even when starting from a single colony, the cross or subsequent propagation could trigger major genomic changes. In our panel, 117 of the original 351 hybrids showed expression bias across entire chromosomes or large chromosomal segments, suggesting that such events may be frequent, although direct genome sequencing is required for confirmation. A thousand independent crosses could reveal which events arise after mating, at what rate, and whether particular regions are recurrent hotspots.

Whether any new karyotypes remain stable over time is a separate question. Aneuploid chromosomes are costly, and yeast may lose them or compensate for their effects (Gilchrist & Stelkens, 2019; Muenzner et al., 2024). Propagating a subset for a few hundred generations while tracking their karyotypes could distinguish transient states from stable new lineages.

The paired transcriptomic and proteomic protocol developed in Chapter 3 could link these genomic changes to molecular phenotypes measured in the same cultures, alongside colony growth across conditions. These traits could be associated with a genotype matrix containing SNPs, InDels, SVs, aneuploidies, ploidy changes and LOH. Because the parents have already been phenotyped, parent–hybrid comparisons could also quantify non-additive trait components (Tsouris et al., 2024b).



D

Figure 1 | A species-wide diallel to study molecular phenotypes and genome stability. **A.** A collection of 364 stable, barcoded haploid strains spans most of the phylogenetic diversity of *S. cerevisiae* (Becerra-Rodríguez et al. 2026). **B.** A subset of genetically diverse parents could be crossed in a sparse pairwise design. **C.** More than 1,000 hybrids could be profiled through transcriptomics and proteomics. **D.** Sequencing the hybrids directly could reveal whole-chromosome, terminal and interstitial LOH, as well as chromosome losses and gains, allowing these events to be analysed across the species. **E.** A subset of hybrids could be propagated over hundreds of generations to determine whether newly formed karyotypes are transient or stable.

Looking at the transcriptome at single-cell resolution

One major limitation of the work presented here is its reliance on bulk measurements. At harvest, each transcriptomic sample contained about 8×10^6 cells and each proteomic sample about 4×10^7 cells. Every measurement was performed on the entire population, so both the transcriptomes and the proteomes I analysed were population averages.

A culture of genetically identical cells is far from uniform. Two major sources of cell-to-cell variation are relevant here. The first is the cell cycle. Cells in a culture are not synchronised and, at any given time, occupy different stages of the cycle. This is particularly important for the transcriptome because many genes are expressed in a cell-cycle-dependent manner (Spellman et al., 1998). The second source is the stochastic nature of gene expression. Even in a constant environment, cells of the same clone differ substantially in their regulatory state (Gasch et al., 2017). For example, individual cells respond to a change in carbon source at different times (Jariani et al., 2020), and the expression of metabolic genes varies strongly from one cell to the next (Nadal-Ribelles et al., 2019).

Bulk RNA sequencing cannot resolve such differences, regardless of their magnitude. Harvesting cultures at matched optical densities and growth phases standardises sampling as far as possible, but cannot eliminate cell-cycle confounding. The approach implicitly assumes that the distribution of cells across cell-cycle stages is similar across genetic backgrounds, which may not be the case. A more slowly growing strain, for example, might be harvested with a different distribution of cell-cycle stages even at a matched optical density. Part of what appears to be an expression difference between two isolates might therefore reflect population composition as well as gene regulation. Growth rates vary considerably across a collection spanning the diversity of the species, so this confounding could be systematic.

The second source of variability is harder to control because it is intrinsic to the process itself. Transcription often occurs in bursts separated by intervals of random length. Regulation sets the probability of these events, so a gene can be tightly controlled and still yield different numbers of transcripts in two otherwise identical cells (Raser & O'Shea, 2005). This variability can be advantageous. For example, a population that already contains a fraction of cells primed for a future challenge may better withstand an environmental shock (Acar et al., 2008; Levy et al., 2012). The heterogeneity that confounds a comparison between two strains could, in this light, itself be a target of selection. Expression noise also appears to be under genetic control. Both *cis*- and *trans*-acting mutations can change the noise of a gene independently of its mean expression (Raser & O'Shea, 2004), and genome-wide patterns of noise reflect protein function: environmental-response proteins are noisier than the more tightly regulated components of the translation machinery (Newman et al., 2006). Noise could therefore be a phenotype in its own right, with a genetic basis that could be mapped just like transcript or protein abundance.

Two strains with the same mean expression can have very different distributions across their cells. Standard bulk RNA sequencing cannot distinguish whether a change arises

from the fraction of cells expressing a gene or from the expression level within those cells, and it can miss differences that preserve the population mean. The problem is not specific to expression levels. Bulk isoform proportions have the same ambiguity: a gene that appears to produce an even mixture of two isoforms could reflect either every cell using both or two subpopulations each committed to one. This raises a central question about the genotype–phenotype relationship: is some heritable variation reflected in the distribution of phenotypes rather than in their mean? If so, mean-based bulk measurements may miss it. Addressing this possibility would require resolving phenotypes at single-cell resolution and measuring the distribution of expression across cells rather than only its average. Single-cell RNA sequencing (scRNA-seq) now makes this possible and has already been applied in several yeast studies (Gasch et al., 2017; Boocock et al., 2024; Nadal-Ribelles et al., 2025).

A species-wide study of these traits may now be feasible. GWAS would require a large natural population representing as much of the species' genetic diversity as possible (Fig. 2A). Strains would be cultured at high throughput in 96-well plates, with OD measured at regular intervals in a microplate reader. In principle, cells from different strains could be pooled in a single sequencing run and subsequently assigned to their strain of origin using the genetic variants carried by each isolate (Kang et al., 2018). Because every isolate in the 1,011-genome collection has been sequenced, this assignment could be performed across the species, provided that every pair within a pool can be distinguished by enough expressed variants (Fig. 2B). Pools containing equal numbers of cells from each strain would then enter a standard scRNA-seq pipeline, with individual cells encapsulated alongside oligo-coated beads in a water-in-oil emulsion (Fig. 2B). Following enzymatic digestion of the cell wall and membrane permeabilisation, mRNAs would be reverse-transcribed and the resulting cDNAs amplified before Illumina sequencing (Fig. 2C). Cells would then be assigned to strains before analysis.

Resolving expression cell by cell would also allow us to account for the confounders discussed above. Because many genes show periodic expression, each cell could be assigned to a stage of the cycle based on its expression of these genes (Spellman et al., 1998), as shown in (Fig. 2C). Strains could then be compared within a stage, or cell-cycle effects could be regressed out, instead of assuming that every culture contains the same distribution of cells across stages, as the bulk approach I employed implicitly does. That distribution could itself become measurable and could be treated as a potentially mappable phenotype. Each gene could then be represented by several phenotypes per strain rather than one: the mean, which is what we already use; the variance among cells, as a proxy for the noise discussed above; the fraction of cells expressing the gene; and its covariance with other genes in the same cells, among other traits. Each could be used in GWAS and might illuminate the genotype–phenotype relationship beyond the population mean. Single-cell eQTL mapping has already been carried out using segregants from a cross between two backgrounds (Boocock et al., 2024; Alexander et al., 2026); extending it to the scale of the species would be the next step.

Such an approach poses challenges. Although *S. cerevisiae* has often been at the forefront of technological development partly because it is easy to use, single-cell

RNA sequencing is an exception. A yeast cell contains an order of magnitude fewer mRNA molecules than a mammalian cell, and its cell wall must be digested before lysis. Consequently, few transcripts are captured per cell and many genes are detected in only a fraction of cells. Protocols for yeast were designed around this problem (Nadal-Ribelles et al., 2019; Jariani et al., 2020), and analyses must account for these limitations. Standard droplet protocols also generate short 3' tags with limited depth per cell, so the fine-scale isoform proportions examined in the second chapter would be difficult to recover in this way. The barcoded cDNA can, however, be sequenced on a nanopore rather than on an Illumina flow cell, recovering full-length molecules without losing their cell of origin (Wang et al., 2021; Shiao et al., 2023). Single-cell protein abundance is a more distant prospect, but may soon become accessible. Nanopores can now distinguish the twenty amino acids, along with some of their modifications (Restrepo-Pérez et al., 2019; Wang et al., 2024), and peptides can be profiled in parallel to identify their protein of origin (Wang et al., 2026). The objective is to “sequence” whole proteins one molecule at a time (Lu et al., 2025; Ritmejeris et al., 2025), but the throughput required for single-cell proteomics is not yet available (Motone & Nivala, 2023). Nonetheless, a species-wide single-cell transcriptomic survey may now be within reach and could extend genotype–phenotype mapping to cell-to-cell variation.

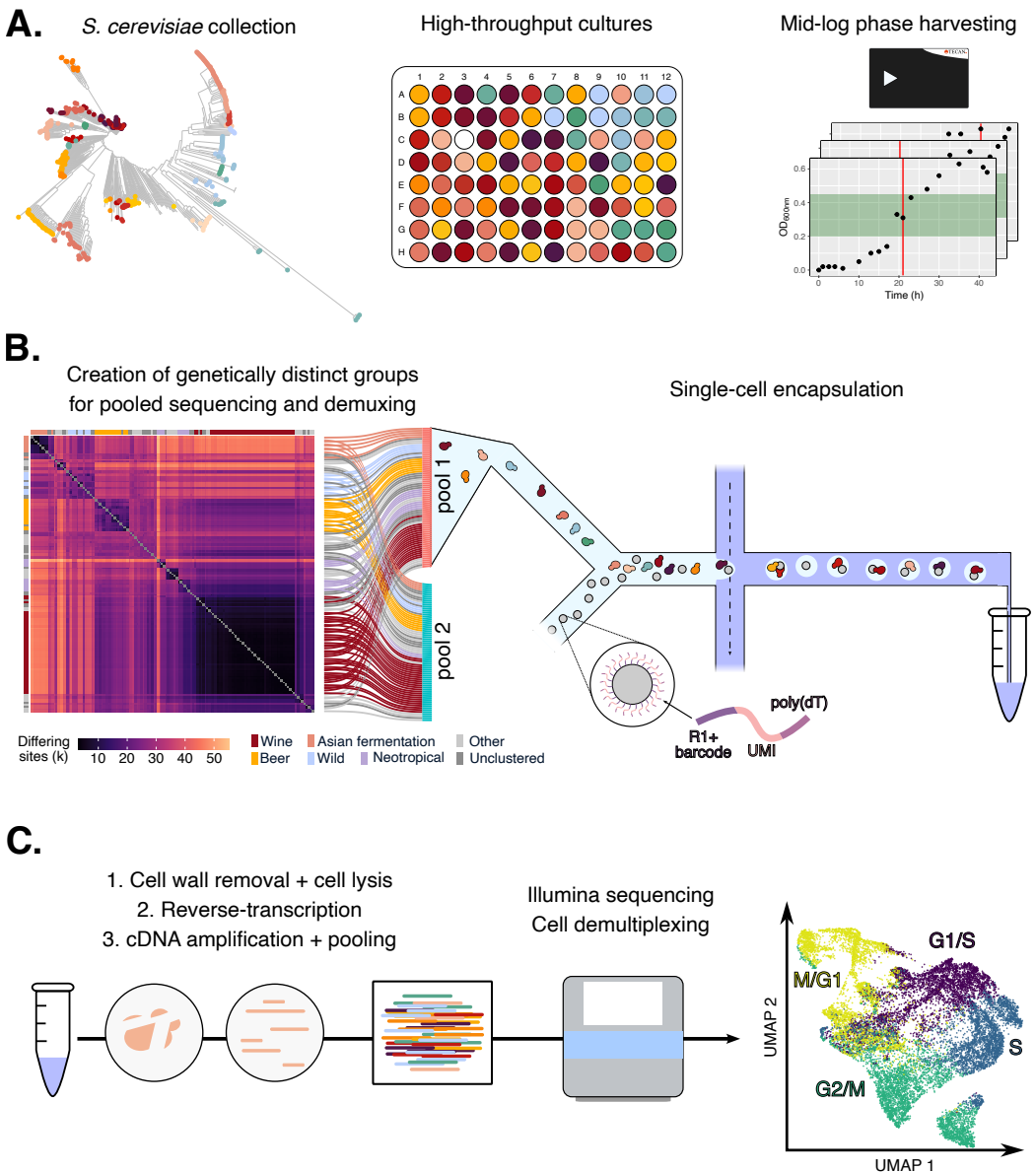


Figure 2 | Resolving the transcriptome of the species one cell at a time.

A. Isolates could be selected across the phylogeny of the collection and grown in microplates, with each strain harvested in mid-log phase. **B.** Strains could be divided into pools designed so that every pair would be distinguished by enough expressed variants after sequencing. The matrix represents the number of sites that differ between each pair in the panel, with the clade of origin shown along the margin. Individual cells would be encapsulated in droplets together with a bead carrying a cell barcode, a unique molecular identifier (UMI) and a poly(dT) capture sequence. **C.** Inside each droplet, the cell wall would be digested and the cell lysed; mRNAs would be reverse-transcribed and cDNAs amplified. The emulsion would then be broken, and the cDNA pooled for Illumina sequencing. Each cell could be assigned to its strain of origin using SNPs detected in the transcriptome (Kang et al. 2018). After strain assignment, the UMAP plot shows 15,314 cells coloured by cell-cycle phase, inferred using the genes described in Spellman et al. 1998.

Limitations of linear models for deciphering the genotype–phenotype relationship

Within each association test, the effect of a variant is modelled as a fixed increment, independent of the rest of the genome. This is a simplifying assumption. Background dependence is pervasive in yeast, with the same allele having different effects in different genetic contexts (Hou et al., 2018). Genetic interactions form a dense network that has been mapped systematically at the genome scale (Costanzo et al., 2010); non-additive components account for a substantial share of population-wide expression variation (Tsouris et al., 2024b); and dominance effects have also been detected at individual loci for human traits (Guindo-Martínez et al., 2021; Palmer et al., 2023). Nevertheless, this assumption is a necessary approximation. Testing groups of k variants among m segregating sites means running

$$\binom{m}{k} = \frac{m!}{k!(m-k)!}$$

tests, a number that grows combinatorially as soon as k exceeds 1. With 100,000 segregating sites, close to what we would obtain from a yeast panel filtered at a MAF of 5%, this amounts to 10^5 single-variant tests, 5×10^9 pairwise tests, 1.67×10^{14} triplet tests, and so on. The significance threshold becomes correspondingly more stringent: a Bonferroni-adjusted threshold falls from 5×10^{-7} for single variants to 10^{-11} for pairs, 3×10^{-16} for triplets and 1.2×10^{-20} for quadruplets, so statistical power decreases rapidly. This is only the statistical side of the problem; computational demand grows accordingly, even with mathematical shortcuts and the GPU implementation of the scan described in Part I of the appendices, p. 250. Moreover, interactions need not stop at two or three variants, so k can increase far beyond what can be enumerated. Yet no amount of computing power would rescue the design because the fundamental limit is not the speed of the tests but the size of the panel. Each individual contributes a single measurement per trait, so a thousand strains provide a thousand observations, and a thousand observations cannot support billions of interaction terms, regardless of the procedure used to fit them.

It is impossible to test all possible interactions by brute force, so a model capable of representing some of them could complement linear analyses. Instead of testing each combination, we could fit a model flexible enough to represent non-linear functions of the genotype and determine which combinations contribute to prediction. The aim would be to assess how well an individual's phenotype can be inferred from its genome sequence within a defined environment.

Sequence-based prediction has already succeeded in other contexts. Neural networks stack layers of artificial neurons, with each layer feeding the next (Rosenblatt, 1958; Rumelhart et al., 1986). This architecture allows them to represent interactions among their inputs rather than only linear combinations. Neural networks became widely used in the early 2010s, once data availability and computing power caught up (Krizhevsky et al., 2017). Protein structure is the best-known biological example: sequence was

known to carry the information required for folding (Anfinsen, 1973), and AlphaFold 2 achieved near-experimental accuracy for many targets in the blind CASP14 assessment (Moult et al., 1995; Jumper et al., 2021; Kryzhafovykh et al., 2021).

The same shift is now occurring at the genome level. Early networks addressed narrow questions over a few hundred bases, such as whether a given protein binds a stretch of DNA or whether a region of chromatin is open (Alipanahi et al., 2015). Their successors expanded the sequence window first to one hundred kilobases (Kelley et al., 2018) and then to two hundred (Avsec et al., 2021); AlphaGenome now processes one million bases at once to predict expression, splicing and chromatin accessibility from sequence alone (Avsec et al., 2026). This window is designed for mammalian enhancers, which can act on their target genes over megabase distances (Schoenfelder & Fraser, 2019). In yeast, it would span most of a chromosome: regulatory sequences are compact and lie mostly within a few hundred bases upstream of the coding sequence, and long-range enhancer regulation is much rarer (Hahn & Young, 2011). Models of this kind have nonetheless begun to be trained on fungal genomes, and one predicts expression in budding yeast from sequence (Chao et al., 2025).

The next question concerns what distinguishes two individuals of the same species, whose genomes are almost identical. The input would be the list of variants carried by each individual; the output, a whole transcriptome with one value per gene. Framed in this way, a genome-wide model would have an advantage over GWAS. Association tests consider one variant at a time, and testing pairwise interactions quickly creates power problems, whereas a network receiving the entire genotype at once could learn which combinations matter. Predicting every gene jointly would also allow covarying genes to share an internal representation instead of being fitted independently. There is, however, an important caveat. A model with more parameters than observations is at high risk of overfitting the training data and failing on unseen samples. Deep networks usually contain millions of parameters, and even after restricting the input to the 100,000 sites in a yeast panel that pass a MAF filter of 5%, a thousand strains would probably be insufficient without strong constraints. Yeast genomics produces large datasets by the standards of the field, with thousands of isolates sequenced and phenotyped, but a thousand observations per gene is not the scale for which these models were developed. It would therefore make sense to start with simpler models. Penalised regressions containing a limited number of interaction terms, or small networks with a few thousand parameters, could provide initial evidence of non-additive signal.

What could such a model look like? For diploid, biallelic sites, each variant could be encoded in three columns, one for each possible genotype, with only one active for a given strain (Fig. 3A). This would triple the width of an input that already spans hundreds of thousands of positions. A one-dimensional convolution could reduce its dimensionality by sliding a window a few dozen variants wide along the variant vector, applying the same weights at every position and pooling along the way (Fig. 3B). Successive convolution and pooling layers would encode local combinations of variants and reduce the sequence to a shorter set of features. An attention layer applied after this reduction could then combine features from distant genomic regions, allowing long-range interactions without a fully connected layer over every input variant (Fig. 3C).

The resulting latent space of a few hundred neurons would form a bottleneck intended to favour features shared across strains over a separate representation memorised for each strain (Fig. 3C). A linear regression could first be fitted and its coefficients frozen; the network could then learn a residual correction to the expression prediction. Its hidden layers would be randomly initialised, while only the final projection of the residual branch would be initialised to zero. The initial correction would therefore be zero, so training would begin exactly at the linear fit (Fig. 3D). Finally, two output heads would use the latent representation: one to predict how strongly each gene is expressed and another to predict whether it is expressed at all (Fig. 3E), with every gene predicted simultaneously (Fig. 3F).

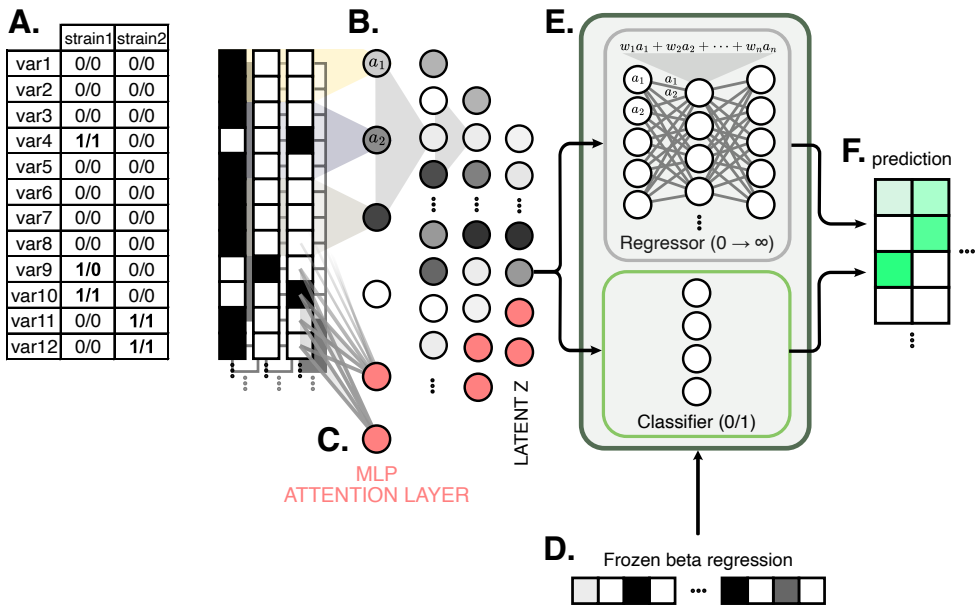


Figure 3 | A possible non-linear model for predicting a full transcriptome from genotype and an additive baseline.

A. For diploid, biallelic sites, each strain would be represented by its genotype (SNPs, InDels and SVs together), one-hot encoded in three columns per variant. A reference homozygote 0/0 would fill the first column, a heterozygote 0/1 the second and an alternative homozygote 1/1 the third, with the other two columns set to zero. **B.** A one-dimensional convolution could slide a window along this vector, allowing neighbouring variants to be processed together, while successive convolution and pooling layers would reduce the sequence to a shorter set of local features. **C.** After downsampling, an attention layer could weight these features across the genome, allowing distant loci to contribute jointly without directly connecting every input variant to every hidden unit. The result would be a compact latent space (Z) intended to favour features shared across strains. **D.** An additive model would be fitted in advance and frozen, providing a baseline expression prediction to which the network could add a learned residual correction. The hidden layers would be randomly initialised, while only the final projection of the residual branch would be initialised to zero. The combined model would therefore start exactly at the additive fit while remaining trainable; performance on held-out strains would determine whether the non-linear correction genuinely helps. **E.** Two heads would use the latent representation. A regressor would predict the expression level of each gene when expressed, and a classifier would predict whether the gene is expressed at all. **F.** The outputs would be an expression probability and a predicted level for every gene in each strain.

Concluding remarks

The questions asked in my thesis are not specific to yeast. How are phenotypic layers connected? Could exploring more layers improve our understanding of the genotype–phenotype relationship? How does a variant act across a population? Corresponding questions are now being asked in human genetics at a scale larger than is possible in yeast.

The UK Biobank now holds whole genomes for nearly 500,000 participants (Carss et al., 2025), and the All of Us program also includes genomic data from 500,000 participants drawn from a more diverse population (Bick et al., 2024; Sharma et al., 2025). Even at this scale, molecular phenotyping lags far behind genomic profiling: plasma proteomes, for example, have been measured for “only” 50,000 individuals (Sun et al., 2023), an order of magnitude fewer than the genomes. Project JAGUAR (jaguar.liigh.unam.mx/), for instance, is assembling a deep atlas of immune-cell composition, transcriptomes, protein expression and matching genotypes. Its 1,000 donors are being recruited across seven Latin American countries, and the cohort represents a mixture of Indigenous American, European and African ancestries that remains poorly represented in catalogues built from European-ancestry cohorts.

The species went from a single reference genome in 1996 to a handful of sequenced isolates, then to hundreds, and finally to the thousand-genome collections used throughout this thesis (Goffeau et al., 1996; Peter et al., 2018; Caudal et al., 2024; Loegler et al., 2024, 2025; Muenzner et al., 2024; Teyssonnière et al., 2024; Brach & Schacherer, 2026). That expansion was initially enabled by the small genome of *S. cerevisiae*, which for two decades made population-scale sequencing far cheaper than in almost any other organism. Costs have since fallen, narrowing the practical gap between yeast and human genomes. Cheap sequencing is no longer the decisive advantage for yeast that it once was. Even the largest yeast multi-omic cohorts remain modest by current human standards. Indeed, matched genomic, transcriptomic and proteomic data are available for only hundreds of isolates, compared with panels in the thousands or tens of thousands in human studies.

The advantage of yeast now lies in the range of experiments that few other systems permit at a comparable scale. Crosses can be designed and thousands of segregants recovered; environments can be controlled; loci can be edited; data can be generated quickly; measurements can be repeated on the same individuals; and experimental evolution can be conducted at scale. Such controlled designs are unavailable to human cohorts, regardless of their size, and this is likely where future advances in experimental design will emerge. The community that took this species from one genome to thousands is best placed to realize this potential and, as at every stage of that expansion, no single strain will be enough. By isolating, sequencing and phenotyping ever more strains while exploiting the species’ unique advantages, we can continue to push the limits of what yeast can teach us.

References

- Acar M, Mettetal JT, Oudenaarden A van (2008) Stochastic switching as a survival strategy in fluctuating environments. *Nat Genet* **40**: 471–475.
- Alexander N, Boocock J, Marquez H, Bloom JS, Kruglyak L (2026) The genetics of transcriptional responses to stress in yeast. : 2026.05.29.728873.
- Alipanahi B, DeLong A, Weirauch MT, Frey BJ (2015) Predicting the Sequence Specificities of DNA- and RNA-binding Proteins by Deep Learning. *Nat Biotechnol* **33**: 831–838.
- Anfinsen CB (1973) Principles that Govern the Folding of Protein Chains. *Science* **181**: 223–230.
- Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. (2021) Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**: 1196–1203.
- Avsec Ž, Latysheva N, Cheng J, Novati G, Taylor KR, Ward T, et al. (2026) Advancing regulatory variant effect prediction with AlphaGenome. *Nature* **649**: 1206–1218.
- Becerra-Rodríguez C, Thiele P, Brach G, Dutta A, Garin M, Tan G, et al. (2026) Domestication drives repeated evolution of sexual–asexual life cycle trade-offs in yeast. *Proc Natl Acad Sci* **123**: e2526682123.
- Bick AG, Metcalf GA, Mayo KR, Lichtenstein L, Rura S, Carroll RJ, et al. (2024) Genomic data in the All of Us Research Program. *Nature* **627**: 340–346.
- Bloom JS, Boocock J, Treusch S, Sadhu MJ, Day L, Oates-Barker H, et al. (2019) Rare variants contribute disproportionately to quantitative trait variation in yeast. *eLife* **8**: e49212.
- Boocock J, Alexander N, Tapia LA, Walter-McNeill L, Munugala C, Bloom JS, et al. (2024) Single-cell eQTL mapping in yeast reveals a tradeoff between growth and reproduction. *eLife* **13**.
- Brach G, Schacherer J (2026) From one genome to thousands, and beyond. *FEMS Yeast Res* **26**: foag038.
- Carss K, Halldorsson BV, Hou L, Liu J, Wheeler E, Lo Y, et al. (2025) Whole-genome sequencing of 490,640 UK Biobank participants. *Nature* **645**: 692–701.
- Caudal É, Loegler V, Dutreux F, Vakirlis N, Teyssonnière É, Caradec C, et al. (2024) Pan-transcriptome reveals a large accessory genome contribution to gene expression variation in yeast. *Nat Genet* **56**: 1278–1287.
- Chao K-H, Magzoub MM, Stoops E, Hackett S, Linder J, Kelley DR (2025) Predicting dynamic expression patterns in budding yeast with a fungal DNA language model. *bioRxiv*: 2025.09.19.677475.
- Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, et al. (2010) The Genetic Landscape of a Cell. *Science* **327**: 425–431.
- Fournier T, Abou Saada O, Hou J, Peter J, Caudal E, Schacherer J (2019) Extensive impact of low-frequency variants on the phenotypic landscape at population-scale. *eLife* **8**.
- Gasch AP, Yu FB, Hose J, Escalante LE, Place M, Bacher R, et al. (2017) Single-cell RNA sequencing reveals intrinsic and extrinsic regulatory heterogeneity in yeast responding to stress. *PLOS Biol* **15**: e2004050.
- Gilchrist C, Stelkens R (2019) Aneuploidy in yeast: Segregation error or adaptation mechanism? *Yeast Chichester Engl* **36**: 525–539.

- Goffeau A, Barrell B. G., Bussey H., Davis R. W., Dujon B., Feldmann H., et al. (1996) Life with 6000 Genes. *Science* **274**: 546–567.
- Guindo-Martínez M, Amela R, Bonàs-Guarch S, Puiggròs M, Salvoro C, Miguel-Escalada I, et al. (2021) The impact of non-additive genetic associations on age-related complex diseases. *Nat Commun* **12**: 2436.
- Hahn S, Young ET (2011) Transcriptional Regulation in *Saccharomyces cerevisiae*: Transcription Factor Regulation and Function, Mechanisms of Initiation, and Roles of Activators and Coactivators. *Genetics* **189**: 705–736.
- Hou J, Leeuwen J van, Andrews BJ, Boone C (2018) Genetic Network Complexity Shapes Background-Dependent Phenotypic Expression. *Trends Genet TIG* **34**: 578–586.
- Jariani A, Vermeersch L, Cerulus B, Perez-Samper G, Voordeckers K, Van Brussel T, et al. (2020) A New Protocol for Single-Cell RNA-seq Reveals Stochastic Gene Expression during Lag Phase in Budding Yeast. *eLife* **9**: e55320.
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**: 583–589.
- Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, et al. (2018) Multiplexed Droplet Single-Cell RNA-sequencing Using Natural Genetic Variation. *Nat Biotechnol* **36**: 89–94.
- Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J (2018) Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res* **28**: 739–750.
- Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. *Commun ACM* **60**: 84–90.
- Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J (2021) Critical assessment of methods of protein structure prediction (CASP)–Round XIV. *Proteins Struct Funct Bioinforma* **89**: 1607–1617.
- Levy SF, Ziv N, Siegal ML (2012) Bet Hedging in Yeast by Heterogeneous, Age-Correlated Expression of a Stress Protectant. *PLOS Biol* **10**: e1001325.
- Li L, Huang K-L, Gao Y, Cui Y, Wang G, Elrod ND, et al. (2021) An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nat Genet* **53**: 994–1005.
- Loegler V, Friedrich A, Schacherer J (2024) Overview of the *Saccharomyces cerevisiae* population structure through the lens of 3,034 genomes. *G3 GenesGenomesGenetics* **14**: jkae245.
- Loegler V, Thiele P, Teyssonnière E, Tsouris A, Brach G, Cruaud C, et al. (2025) From genotype to phenotype with 1,086 near telomere-to-telomere yeast genomes. *Nature*: 1–10.
- Lu C, Bonini A, Viel JH, Maglia G (2025) Toward single-molecule protein sequencing using nanopores. *Nat Biotechnol* **43**: 312–322.
- Mayr C (2016) Evolution and Biological Roles of Alternative 3'UTRs. *Trends Cell Biol* **26**: 227–237.
- Mittleman BE, Pott S, Warland S, Zeng T, Mu Z, Kaur M, et al. (2020) Alternative polyadenylation mediates genetic regulation of gene expression. *eLife* **9**: e57492.
- Motone K, Nivala J (2023) Not if but when nanopore protein sequencing meets single-cell proteomics. *Nat Methods* **20**: 336–338.
- Moult J, Pedersen JT, Judson R, Fidelis K (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins Struct Funct Bioinforma* **23**: ii–iv.
- Muenzner J, Trébulle P, Agostini F, Zauber H, Messner CB, Steger M, et al. (2024) Natural proteome diversity links aneuploidy tolerance to protein turnover. *Nature* **630**: 149–157.

- Nadal-Ribelles M, Islam S, Wei W, Latorre P, Nguyen M, Nadal E de, et al. (2019) Sensitive High-Throughput Single-Cell RNA-seq Reveals within-Clonal Transcript Correlations in Yeast Populations. *Nat Microbiol* **4**: 683–692.
- Nadal-Ribelles M, Solé C, Díez-Villanueva A, Stephan-Otto Attolini C, Matas Y, Steinmetz L, et al. (2025) A single-cell resolved genotype-phenotype map using genome-wide genetic and environmental perturbations. *Nat Commun* **16**: 2645.
- Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, et al. (2006) Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441**: 840–846.
- O'Donnell S, Yue J-X, Saada OA, Agier N, Caradec C, Cokelaer T, et al. (2023) Telomere-to-telomere assemblies of 142 strains characterize the genome structural landscape in *Saccharomyces cerevisiae*. *Nat Genet* **55**: 1390–1399.
- Palmer DS, Zhou W, Abbott L, Wigdor EM, Baya N, Churchhouse C, et al. (2023) Analysis of genetic dominance in the UK Biobank. *Science* **379**: 1341–1348.
- Passmore LA, Collier J (2022) Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression. *Nat Rev Mol Cell Biol* **23**: 93–106.
- Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, et al. (2018) Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **556**: 339–344.
- Raser JM, O'Shea EK (2004) Control of Stochasticity in Eukaryotic Gene Expression. *Science* **304**: 1811–1814.
- Raser JM, O'Shea EK (2005) Noise in Gene Expression: Origins, Consequences, and Control. *Science* **309**: 2010–2013.
- Restrepo-Pérez L, Wong CH, Maglia G, Dekker C, Joo C (2019) Label-Free Detection of Post-translational Modifications with a Nanopore. *Nano Lett* **19**: 7957–7964.
- Ritmeijeris J, Chen X, Dekker C (2025) Single-molecule protein sequencing with nanopores. *Nat Rev Bioeng* **3**: 303–316.
- Rosenblatt F (1958) The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol Rev* **65**: 386–408.
- Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature* **323**: 533–536.
- Schoenfelder S, Fraser P (2019) Long-range enhancer–promoter contacts in gene expression control. *Nat Rev Genet* **20**: 437–455.
- Sharma S, Nagar SD, Pemu P, Zuchner S, Mariño-Ramírez L, Meller R, et al. (2025) Genetic ancestry and population structure in the All of Us Research Program cohort. *Nat Commun* **16**: 4123.
- Shiau C-K, Lu L, Kieser R, Fukumura K, Pan T, Lin H-Y, et al. (2023) High throughput single cell long-read sequencing analyses of same-cell genotypes and phenotypes in human tumors. *Nat Commun* **14**: 4124.
- Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. (1998) Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Mol Biol Cell* **9**: 3273–3297.
- Subtelny AO, Eichhorn SW, Chen GR, Sive H, Bartel DP (2014) Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508**: 66–71.
- Sun BB, Chiou J, Traylor M, Benner C, Hsu Y-H, Richardson TG, et al. (2023) Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**: 329–338.

- Teyssonnière EM, Trébulle P, Muenzner J, Loegler V, Ludwig D, Amari F, et al. (2024) Species-wide quantitative transcriptomes and proteomes reveal distinct genetic control of gene expression variation in yeast. *Proc Natl Acad Sci* **121**: e2319211121.
- Tian B, Manley JL (2017) Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* **18**: 18–30.
- Tsouris A, Brach G, Friedrich A, Hou J, Schacherer J (2024a) Diallel panel reveals a significant impact of low-frequency genetic variants on gene expression variation in yeast. *Mol Syst Biol* **20**: 362–373.
- Tsouris A, Brach G, Schacherer J, Hou J (2024b) Non-additive genetic components contribute significantly to population-wide gene expression variation. *Cell Genomics* **4**.
- Wang J, Chen J, Pan H, Luo F, Qin W, Zeng H, et al. (2026) Nanopore-based massively parallel sensing for peptide profiling and protein identification. *Nat Commun* **17**: 3058.
- Wang K, Zhang S, Zhou X, Yang X, Li X, Wang Y, et al. (2024) Unambiguous discrimination of all 20 proteinogenic amino acids and their modifications by nanopore. *Nat Methods* **21**: 92–101.
- Wang Q, Bönigk S, Böhm V, Gehring N, Altmüller J, Dieterich C (2021) Single-cell transcriptome sequencing on the Nanopore platform with ScNapBar. *RNA* **27**: 763–770.

Appendices

F

FaST-ER-LMM: a GPU implementation of FaST-LMM for our lab's GWAS

| Abstract

A genome-wide association scan is, at its core, the same regression repeated across genetic markers. In our workflow, the entire scan is then repeated for thousands of phenotypes and their permutations. At our lab's (still modest) scale, a full transcriptome-wide mixed-model scan takes days on a cluster, so even a small change to the analysis requires another long run. FaST-ER-LMM keeps the FaST-LMM statistical model unchanged but reorganizes its calculations into large batches suited to graphics processing units (GPUs). The same scan runs in about ten minutes on a single V100S GPU, a speedup of close to $800\times$, and closely matches the FaST-LMM reference output ($r = 0.999998$ for $-\log_{10}(p)$). This runtime makes gene-by-environment and epistasis scans practical at the scale used in the lab.

Introduction

A genome-wide association study (GWAS) repeats the same basic question across a large number of markers: does the genotype at this position explain variation in the phenotype, after accounting for the covariates? For a single phenotype, each marker is tested using the allele dosage carried by each individual, and the resulting statistic is compared with a genome-wide significance threshold. That threshold can be analytical, such as a Bonferroni correction, or empirical, as in the phenotype permutations used here. The individual test is cheap; each test uses data from N individuals, and the cost comes from applying it across M markers and P phenotypes, then repeating much of this work for phenotype permutations and leave-one-chromosome-out (LOCO) analysis. This becomes expensive quickly when thousands of transcript or protein abundances are analyzed.

In its simplest form, the per-variant test is an ordinary linear regression relating the phenotype to allele dosage while accounting for covariates:

$$y = X\beta + s\gamma + e,$$

where

- y : phenotype measured across the N individuals ($N \times 1$)
- X : covariate matrix, with effects β ($N \times C$)
- s : allele dosage at the tested variant ($N \times 1$)
- γ : effect of the tested variant
- e : residual noise, assumed independent across individuals ($N \times 1$)

The independence assumption is a poor approximation in a structured genetic panel. This is particularly relevant for *S. cerevisiae*, whose natural isolates form well-defined populations and can also share substantial recent ancestry (Connelly & Akey, 2012; Peter et al., 2018; Loegler et al., 2024). If relatedness is ignored, an allele that follows ancestry can appear associated with any phenotype that follows the same ancestry, even when the allele itself is not causal.

Genomic control addresses this confounding by rescaling association statistics with an inflation factor λ , usually estimated from the ratio of the observed median test statistic to its null expectation (Devlin & Roeder, 1999). Principal-component methods such as EIGENSTRAT instead include the leading axes of genotype variation as fixed-effect covariates (Price et al., 2006). Both account for broad population structure, but a small set of principal components may miss finer family structure or cryptic relatedness (Price et al., 2010). Here, genome-wide relatedness is modeled directly with a linear mixed model.

For one phenotype y measured across N individuals, the model is (Yu et al., 2006; Kang et al., 2008)

$$y = X\beta + s\gamma + g + e, \quad g \sim \mathcal{N}(0, \sigma_g^2 K), \quad e \sim \mathcal{N}(0, \sigma_e^2 I),$$

where

- X : covariates such as sequencing batch or aneuploidy, with effects β ($N \times C$)
- s : allele dosage at the tested variant ($N \times 1$)
- γ : fixed effect of the tested variant, tested against $H_0 : \gamma = 0$
- g : polygenic random effect capturing phenotypic similarity that follows genome-wide relatedness
- e : residual noise, assumed independent across individuals ($N \times 1$)
- K : kinship or genomic relationship matrix ($N \times N$)
- I : identity matrix ($N \times N$), indicating that residual variance is independent across individuals
- σ_g^2 and σ_e^2 : polygenic and residual variance components

The Gaussian (normal) distributions assigned to g and e are modeling assumptions, not claims that every underlying biological effect is normally distributed. For g , this approximation is related to Fisher's infinitesimal model, in which many small genetic effects combine into an approximately Gaussian polygenic contribution (Fisher, 1918, 1919; Visscher & Bruce Walsh, 2019). The random effect g captures covariance that follows genome-wide relatedness, while e collects variation not explained by the fixed effects or the kinship term. Under this model,

$$y \sim \mathcal{N}(X\beta + s\gamma, \sigma_g^2 K + \sigma_e^2 I).$$

The likelihood can therefore be evaluated for any given pair of variance components. The model itself is compact; the computational problem is evaluating it repeatedly across many markers, phenotypes, LOCO iterations and permutations.

The kinship matrix K is built from the genome-wide genotype matrix Z ($N \times M$). In the scaling used here, each marker column is centered by its sample mean and divided by its sample standard deviation, after which

$$K = \frac{ZZ^\top}{M}.$$

Genetically similar strains therefore have larger off-diagonal entries in K . In turn, the model can account for phenotypic similarity that follows genome-wide relatedness while estimating the effect of the tested marker. This is a standardized-marker form of a genomic relationship matrix (VanRaden, 2008; Yang et al., 2010). Under

Hardy–Weinberg equilibrium, scaling by the observed sample standard deviation is closely related to scaling by $\sqrt{2p_j(1-p_j)}$, where p_j is the allele frequency at marker j , but the two are not identical in finite or structured populations. The standardization places markers on a common variance scale; allele frequency and linkage disequilibrium (LD) still affect the realized structure of K .

Provided that K is scaled appropriately, the ratio

$$\frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

is the proportion of phenotypic variance attributed to the fitted random effect. In association-mapping papers this quantity is often called *pseudoheritability*; “SNP-based heritability” is appropriate when the relationship matrix is specifically constructed from SNPs under the corresponding scaling (Kang et al., 2010). Work aggregating the effects of many common SNPs also showed that substantial genetic variance can be spread across variants whose individual effects are too small to reach genome-wide significance (Yang et al., 2010). The variance components can be estimated by maximum likelihood or restricted maximum likelihood (REML). REML accounts for the degrees of freedom used to estimate the fixed effects and therefore usually reduces small-sample bias in variance-component estimates.

If K is built from all genome-wide markers, it also contains the tested marker and nearby markers in LD with it. The random effect can then absorb part of a true local signal before γ is estimated, reducing power. This is proximal contamination (Listgarten et al., 2012). A common way to reduce it is leave-one-chromosome-out (LOCO): when a marker on one chromosome is tested, K is built from the remaining chromosomes. This excludes the tested marker and its same-chromosome LD neighborhood from the kinship calculation, greatly reducing contamination from physically linked variants.

Early mixed-model implementations were computationally expensive. The unified mixed model required repeated variance-component fitting, with matrix operations whose cost grows with the cube of the number of individuals: doubling the panel size makes these operations approximately eight times more expensive (Yu et al., 2006). EMMA made exact fitting more efficient with spectral methods, but marker-by-marker refitting remained costly (Kang et al., 2008). EMMAX and the closely related P3D approach (“population parameters previously determined”) instead estimate variance components under the null model and reuse them across markers (Kang et al., 2010; Zhang et al., 2010). FaST-LMM extends this strategy by reusing the expensive transformation of a kinship matrix throughout a scan (Lippert et al., 2011).

FaST-LMM

The computational bottleneck lies in fitting the mixed model. Each marker test compares the model including the marker effect γ with the null model $H_0 : \gamma = 0$. In FaST-LMM, association significance is assessed with a likelihood-ratio test, which measures how much the model fit improves when γ is allowed to differ from zero.

The expensive part of the calculation is accounting for the covariance between related strains. Two near-clonal isolates carrying the same allele should not contribute as if they were independent observations. The covariance among the phenotype measurements of all strains is described by

$$V = \sigma_g^2 K + \sigma_e^2 I,$$

and the corresponding generalized least-squares calculation uses V^{-1} so that correlated observations are weighted appropriately.

The structure of V is known once K has been built, but the variance components are not. In a naive implementation, likelihood optimization evaluates many candidate values of (σ_g^2, σ_e^2) , and each evaluation requires solving linear systems involving V . Direct factorization of an $N \times N$ covariance matrix costs $O(N^3)$, which becomes prohibitive if it is repeated for every marker and phenotype.

FaST-LMM avoids repeating that cubic operation inside every likelihood evaluation (Lippert et al., 2011). For a given kinship matrix it computes the eigendecomposition

$$K = U \Lambda U^\top,$$

where the columns of U are the eigenvectors of K and Λ is diagonal, with eigenvalues λ_i . Rotating y , X and the tested genotype by $U^\top$ re-expresses them along the orthogonal axes of genetic relatedness defined by K . In these coordinates,

$$U^\top V U = \sigma_g^2 \Lambda + \sigma_e^2 I,$$

which is diagonal. The required linear solves now use element-wise divisions, and the log-determinant becomes a sum over the diagonal entries. Once the eigendecomposition is available, evaluating the profile likelihood for a fixed number of covariates is therefore linear rather than cubic in N .

The decomposition is performed once for each kinship matrix, which in the workflow used here means once per LOCO iteration. The rotated genotype data can then be reused across likelihood evaluations and marker tests. Rotating a full genotype matrix is still substantial work when both N and M are large, but the rotation is performed once per kinship matrix rather than repeated during every likelihood evaluation.

When the kinship is built from $k \ll N$ markers, its relatedness structure can be represented in k dimensions rather than N . FaST-LMM can then work directly with the smaller $N \times k$ genotype matrix as a low-rank factor instead of explicitly forming and decomposing a full $N \times N$ kernel. This matters for the large- N benchmarks below, where the kinship is pruned to $k = 5,000$ markers and N reaches 100,000.

A second simplification reduces the variance-component search from two parameters to one. At a fixed ratio of residual to polygenic variance

$$\delta = \frac{\sigma_e^2}{\sigma_g^2},$$

the best-fitting fixed effects and overall variance scale can be calculated directly, without a numerical search. The remaining numerical optimization is one-dimensional in δ , rather than a two-dimensional search over (σ_g^2, σ_e^2) . Once δ is fixed, the two variance components follow from the fitted scale.

FaST-LMM stands for *Factored Spectrally Transformed Linear Mixed Models*. Separating the spectral decomposition from the likelihood calculations allows the same eigendecomposition to be reused even when δ is re-estimated for different marker tests, depending on the association mode used (Lippert et al., 2011).

FaST-ER-LMM: making it faster on GPUs

Why the reference implementation is too slow for our needs

Even with these optimizations, the production workflow is expensive because the scan is repeated across phenotypes, permutations and LOCO iterations. At the scale used in our team (approximately 500 strains, 100,000 variants, 6,500 transcript phenotypes, 100 permutations and 16 LOCO iterations), one phenotype takes about 720 seconds on 10 CPU cores. This is admittedly a worst-case benchmark, as it was run on our aging (but fantastic) computing cluster. With ten such jobs running concurrently, a transcriptome-wide scan takes about 130 hours of wall-clock time and consumes approximately 13,000 CPU-hours.

The difference between wall-clock time and total CPU use matters in practice. A 130-hour run ties up 100 CPU cores for more than five days, and the same analysis across four environments approaches 52,000 CPU-hours before the proteomic phenotypes are added. It also discourages iteration: changing a covariate, transformation or filtering rule means waiting days for another complete scan. FaST-ER-LMM grew out of that bottleneck.

Porting the association calculations to GPUs

FaST-ER-LMM stands for *Factored Spectrally Transformed Embarrassingly Reducible Linear Mixed Model*. The added *ER* is both a description and a joke: it spells out the project's entire ambition, which was simply to go fast*ER*. The term itself borrows from *embarrassingly parallel*, which describes a workload that splits into independent tasks with no communication between them. Once the phenotypes and genotypes have been rotated into the spectral basis, no test depends on the result of any other, whether across markers, phenotypes, permutations or LOCO iterations. The result is a very large number of small, independent reductions.

I ported the core association calculations to graphics processing units (GPUs) with PyTorch (Paszke et al., 2019). The expensive parts of the scan are dominated by dense linear algebra and large batches of repeated calculations: eigendecompositions, matrix rotations and per-marker tests across many phenotypes and permutations. These operations map naturally onto GPU hardware.

I chose PyTorch as the backend because it provides CUDA tensors and GPU implementations of the linear-algebra operations needed here. Matrix multiplication is handled by NVIDIA's CUDA libraries, while routines such as `torch.linalg.eigh` use specialized CUDA backends, including cuSOLVER. Phenotypes, individuals, variants and permutations can also be represented directly as tensor dimensions, allowing thousands of tests to be evaluated in batches instead of Python loops.

I deliberately kept the statistical behavior of the port close to that of the reference implementation. It reimplements the association calculations used in the reference Python FaST-LMM workflow while keeping the same LOCO scheme, and adds the preprocessing and output conventions used routinely in our team: permutation-based thresholds, rank-based inverse normal transformation (RINT) and covariate handling (Peter et al., 2018; Fournier et al., 2019; Caudal et al., 2024; Teyssonnière et al., 2024; Tsouris et al., 2024; Loegler et al., 2025a, b; Becerra-Rodríguez et al., 2026; Dutta et al., 2026). In the production mode benchmarked below, δ is fitted once per phenotype and LOCO iteration under the null model and then held fixed across marker tests.

Transcript-abundance phenotypes are often strongly skewed and contain outliers. In a panel of only a few hundred strains, a single extreme value can noticeably affect an association statistic, so our team routinely applies RINT. Each phenotype value is replaced by the normal score corresponding to its rank: a strain with tie-averaged rank r_i among n measured strains receives

$$\Phi^{-1}\left(\frac{r_i - 3/8}{n + 1/4}\right),$$

where Φ^{-1} is the standard normal quantile function and $3/8$ is Blom's offset (Beasley et al., 2009). RINT is a preprocessing choice, not a guarantee of better calibration or power in every setting, so its behavior still needs to be checked for the analysis at hand.

Running a scan

FaST-ER-LMM runs the full GWAS from one command instead of splitting phenotypes across manually orchestrated cluster jobs. The `gwas` subcommand takes a PLINK genotype dataset, a wide phenotype table with one column per phenotype, an optional PLINK-style covariate file and an output directory:

```
fasterlmm gwas --geno matrix.plink --pheno glucose_wide.tsv.gz \
              --covar covariates_glucose.tab --outdir results/glucose/ \
              --n-perm 100 --rint
```

The benchmarking results presented below were generated with this configuration. The default analysis uses LOCO, RINT, 100 permutations, an empirical significance threshold given by the 5% quantile of the minimum p -value across permutations, and a fixed random seed (09091993). The fixed seed makes the permutation assignment reproducible for a given input; small floating-point differences can still arise across hardware or numerical backends.

The permutations provide an empirical genome-wide significance threshold. For each permuted phenotype, the smallest p -value across all tested variants is retained, generating the null distribution of the most significant association expected by chance. Taking the 5% quantile of these minima therefore gives a threshold corresponding to an approximate 5% family-wise error rate across the variants tested for that phenotype while accounting for the number of markers tested and their LD structure. Our team routinely uses 100 permutations as a compromise between computational cost and threshold estimation.

By default, the scan writes one directory per phenotype containing the association table, permutation minima and empirical threshold, matching the layout already used by our team. It can also write the associations as a single Parquet dataset, which avoids creating thousands of small files and makes downstream aggregation much simpler.

The principal runtime options are:

- `--device`: selects the computational backend. On a multi-GPU node, `cuda` dispatches phenotype blocks across all visible GPUs. A specific device can be selected with, for example, `cuda:0`, while `cpu` runs the same implementation without GPU acceleration.
- `--phenos-per-job`: sets the number of phenotypes that share one eigendecomposition. Larger blocks reuse the genotype-dependent calculations across more phenotypes, but require more GPU memory for the rotated phenotype matrix.
- `--shard i/n`: partitions the phenotype list across independent jobs, for example for orchestration with Snakemake or Nextflow.
- `--dry-run`: reports the planned work, panel dimensions, output layout and resolved device without starting the scan.

Sharing one eigendecomposition across phenotypes

Moving the arithmetic to a GPU is only part of the speedup. The expensive genotype-dependent work, especially construction and factorization of the kinship matrix, does not depend on which phenotype is being tested. Within a LOCO iteration,

the same decomposition can therefore be shared by many phenotypes and all of their permutations.

FaST-ER-LMM groups B phenotypes and their $B \times n_{\text{perm}}$ permutation columns (where n_{perm} is the number of permutations) into a matrix of dimension $N \times B(1 + n_{\text{perm}})$. That block is rotated once per chromosome, after which the phenotypes are tested against variants with batched matrix operations. With $B = 64$ and 100 permutations, one block contains 6,464 columns. A study with 6,500 phenotypes contains 656,500 observed and permuted phenotype columns in total, so the full matrix is processed in blocks when it does not fit in GPU memory. This sharing across phenotypes is the *Embarrassingly Reducible* part of the name. Unless stated otherwise, the benchmarks below use $B = 256$.

The resulting per-phenotype timing has the form

$$t(B) = \frac{t_{\text{block}}}{B} + t_{\text{pheno}},$$

where t_{block} is the time spent once per block (reading and decoding genotypes, building and factorizing K , and rotating the genotype matrix by U^T), while t_{pheno} is the time spent on work repeated for each phenotype (rotating the observed and permuted phenotype columns, fitting δ under the null model, and testing the markers). As B increases, the block-level cost is shared across more phenotypes and contributes less to each phenotype's runtime. Once that contribution becomes small, available GPU memory rather than arithmetic efficiency sets the practical limit.

The benchmarks were run on an NVIDIA Tesla V100S PCIe GPU with 5,120 CUDA cores, 32 GB of HBM2 memory, 1,134 GB/s of memory bandwidth and 16.4 TFLOPS of peak single-precision throughput. The memory bandwidth matters because large genotype and phenotype blocks are moved repeatedly, while the 32 GB capacity limits how many phenotype columns can be kept on the device at once. The V100S is now several generations behind current data-center GPUs, so these measurements are a conservative hardware baseline rather than a best-case runtime.

Using more than one GPU

Multi-GPU execution partitions the phenotype list across devices, with one process per GPU. The devices work independently, but each process still loads and decodes the full set of M variants. A second GPU therefore helps only once the additional computation is large enough to outweigh that duplicated input work.

At $B = 64$, the median measured two-GPU speedup across the tested values of N is 0.97 (Fig. 1), so the second GPU provides no overall benefit at this block size. Reading and decoding the PLINK data dominates wall time: two processes duplicate that work while each receives only half of the phenotypes. At $N = 5,000$, the speedup rises with B , reaching 1.21 at $B = 256$, 1.72 at $B = 1,024$ and 1.93 when all 6,500 phenotypes are processed together. For the full transcriptome block, the measured two-GPU speedup ranges from 1.78 to 1.93 across the tested values of N .

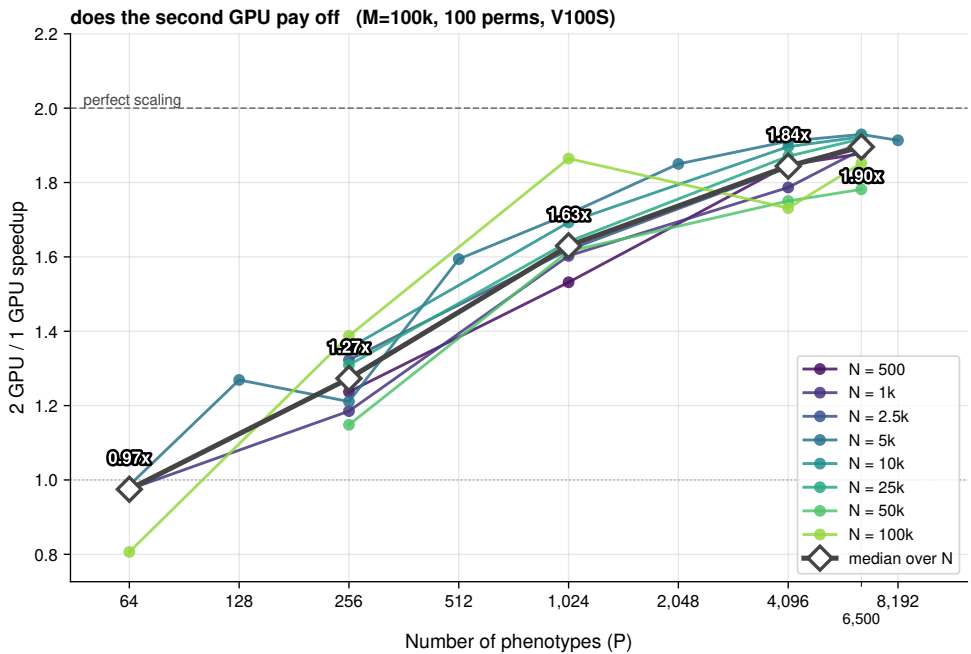


Figure 1 | Speedup using two GPUs
Two-GPU over one-GPU speedup as a function of phenotype block size B , measured on V100S GPUs at $M = 100,000$ variants, $k = 5,000$ and 100 permutations.

Results

For 542 strains, 95,804 variants filtered at a minor allele frequency of 5%, 6,484 phenotypes, 100 permutations and 16 LOCO iterations, FaST-ER-LMM completes the scan in about 10 minutes on one V100S GPU. The CPU baseline takes approximately 130 hours of wall-clock time, giving a $780\times$ speedup. That baseline consumes approximately 13,000 CPU-hours; the GPU run requires only modest CPU support and no large multi-core allocation.

I also benchmarked the scan on simulated data with $M = 100,000$ variants, N ranging from 500 to 100,000, 100 permutations, 16 LOCO iterations and a kinship matrix pruned to $k = 5,000$ markers in both implementations. Table 1 reports the per-phenotype wall time, while (Fig. 2) shows throughput and total wall time across panel sizes N and phenotype block sizes B for one and two GPUs.

At the upper end of the benchmark, with $N = 100,000$ individuals and $M = 100,000$ variants, the GPU implementation takes only a few seconds per phenotype even with 100 permutations. Multiplying the measured per-phenotype times by 6,500 traits gives about 4.9 hours on one GPU and 2.7 hours on two.

(Fig. 3) shows a non-monotonic speedup curve, with a sharp drop around 5,000 strains. This comes from the rank-dependent branch in the reference FaST-LMM implementation. When the number of markers used to build the kinship matrix is smaller than the number

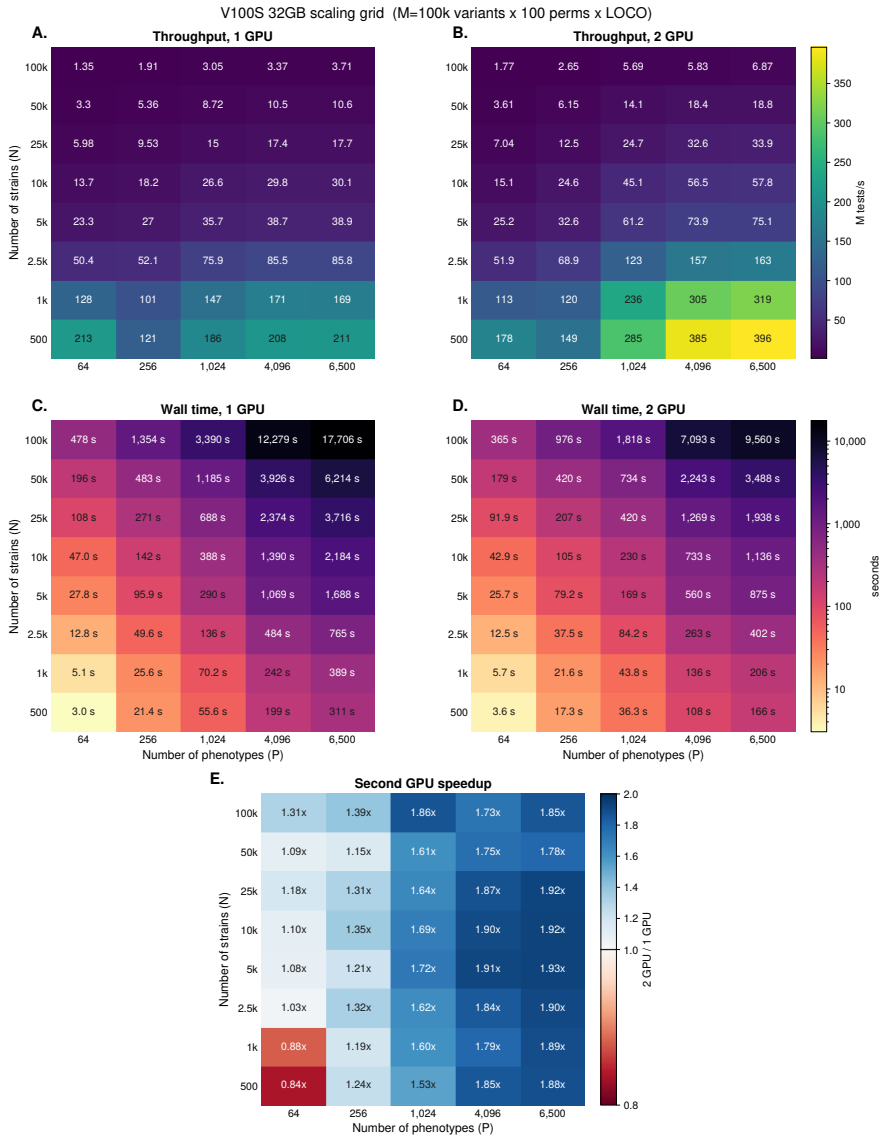
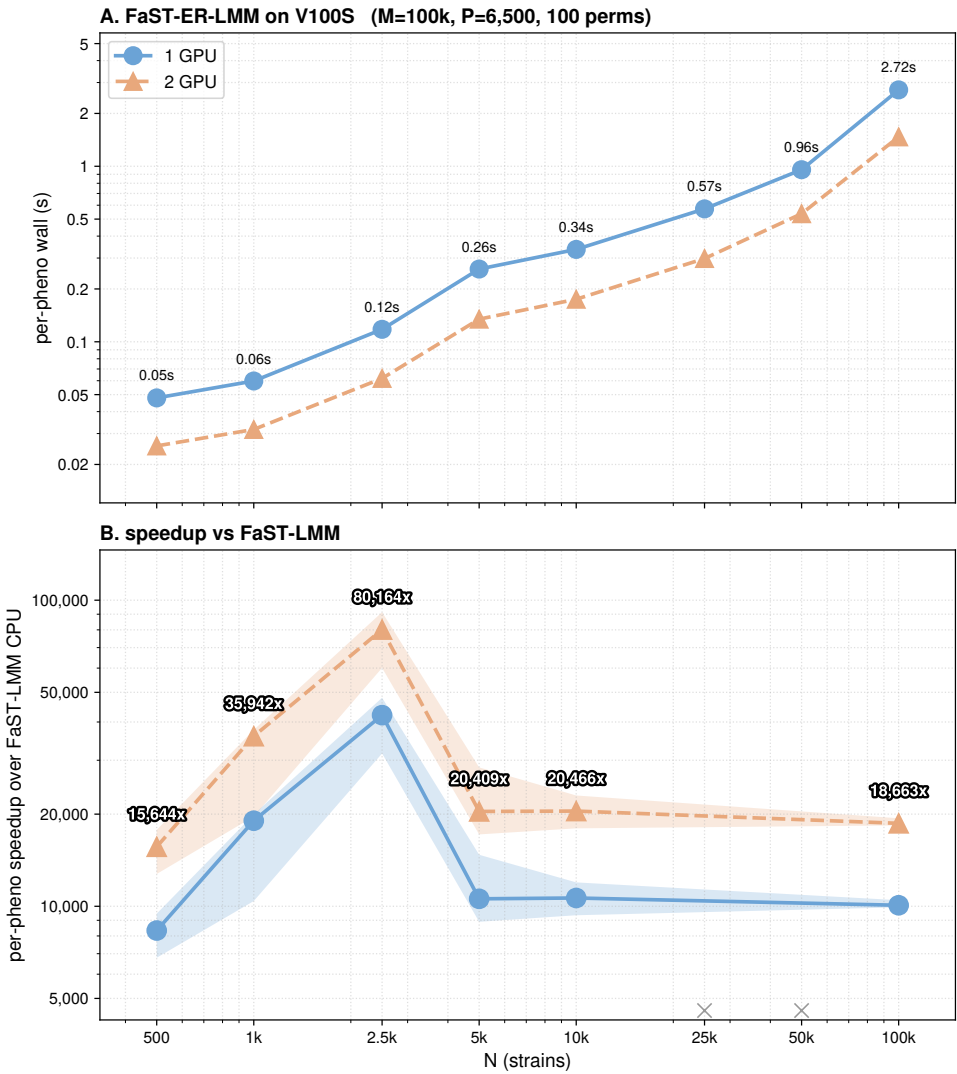


Figure 2 | Throughput and wall time across strain and phenotype block sizes. Measurements on V100S 32 GB GPUs with simulated data at $M = 100,000$ variants, $k = 5,000$, 100 permutations and LOCO, using one GPU (A, C) or two (B, D). One test is one variant evaluated against one phenotype column, including permutation columns. Panel E shows the two-GPU speedup.



A

Figure 3 | FaST-ER-LMM runtime and speedup on simulated data. Per-phenotype wall time (A) and speedup over the FaST-LMM CPU baseline (B) at $M = 100,000$ variants, $k = 5,000$, $P_{\text{total}} = 6,500$ phenotypes and 100 permutations, with N ranging from 500 to 100,000. Every point in panel A is measured. Bands in panel B show the range over five baseline replicates. Crosses on the axis mark the two panel sizes omitted from panel B; the curve therefore connects $N = 10,000$ directly to $N = 100,000$.

N	FaST-LMM	FaST-ER-LMM ($P_{\text{total}} = 6,500$)	
	1 process	1 GPU	2 GPUs
500	247	0.048	0.026
1,000	974	0.060	0.032
2,500	4,735	0.118	0.062
5,000	2,558	0.260	0.135
10,000	3,375	0.336	0.175
25,000	7,491	0.572	0.298
50,000	13,946	0.956	0.537
100,000	26,869	2.724	1.471

Table 1 | Seconds per phenotype

Each per-phenotype time includes the observed phenotype and its 100 permutations. $M = 100,000$, $k = 5,000$ and LOCO are used throughout. Values are medians of five replicates. The decrease in FaST-LMM runtime between $N = 2,500$ and $N = 5,000$ reflects a branch in the reference implementation, discussed below.

of individuals, FaST-LMM can use a low-rank representation; otherwise it falls back to a full-rank kernel. With $k = 5,000$ before LOCO filtering, that switch occurs between $N = 2,500$ and $N = 5,000$, exactly where the reference runtime decreases. The GPU implementation does not slow down at this point; the measured speedup drops because the CPU baseline itself becomes faster.

Timing the two approaches directly supports this explanation. At $N = 5,000$, the low-rank approach took 400.8 seconds, compared with 8,585.0 seconds for the forced full-rank approach, a 21-fold difference.

Parity with the reference implementation

A faster implementation is useful only if it reproduces the reference results! I therefore compared both implementations across all 26,414 transcript phenotypes from the four environments analyzed in the third chapter. Both implementations tested the same 95,804 variants, for 2,530,566,856 variant–phenotype pairs. Across all pairs, the Pearson correlation is 0.999998 for $-\log_{10}(p)$ and 0.999999 for the signed marker effect (Fig. 4a-b). Because a global correlation over billions of mostly null tests can hide phenotype-specific discrepancies, I also examined agreement phenotype by phenotype.

The Bonferroni-significant sets are identical. At a threshold of $0.05/M$, both implementations identify the same 64,748 significant variant–phenotype associations, with no pair crossing the threshold for only one of the two implementations. I use Bonferroni correction for this parity check rather than the empirical permutation threshold because the two comparison runs were not forced to use identical permutation draws.

A

For most phenotypes, the differences are effectively at the level of numerical precision. For the median phenotype, the correlation with the reference rounds to 1.0 and the largest difference across variants is 3×10^{-7} in $-\log_{10}(p)$.

A tiny tail of phenotypes shows larger discrepancies (Fig. 4c). Of the 26,414 phenotypes, 165 have a correlation more than 10^{-6} below 1, six are more than 10^{-4} below 1, and the lowest correlation is 0.9857. These were low-variance phenotypes, but I found no other pattern. The two implementations are therefore interchangeable for this analysis.

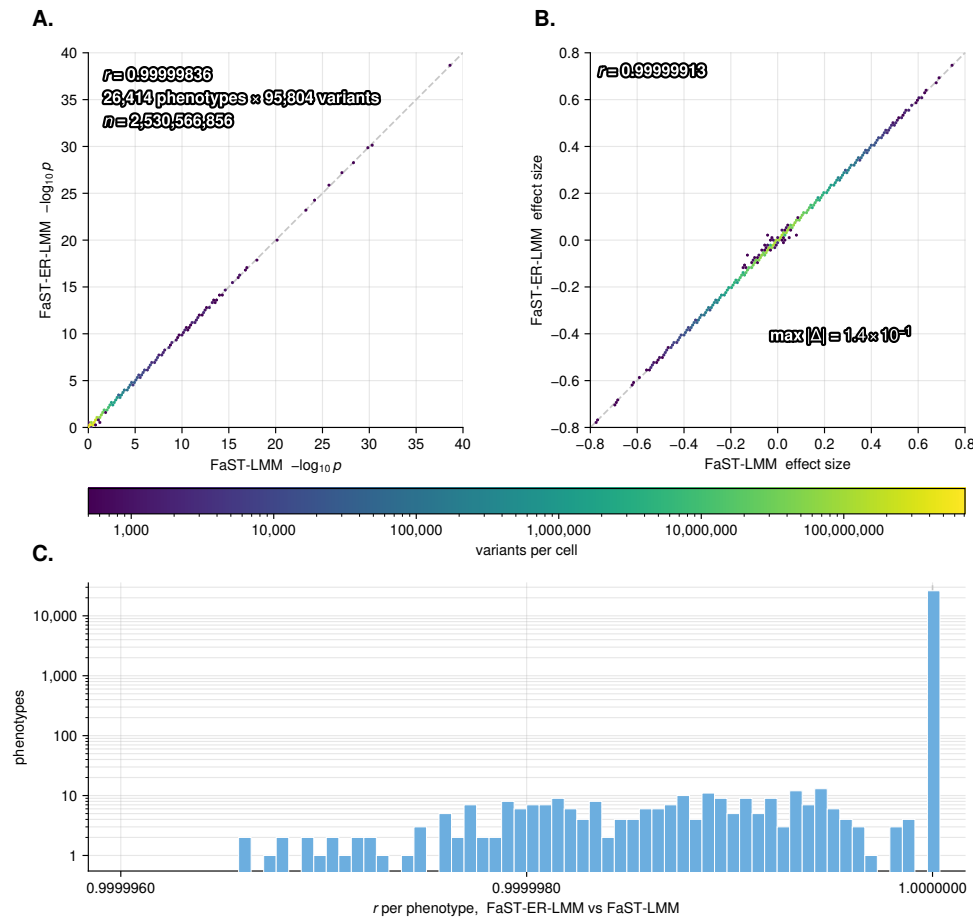


Figure 4 | FaST-ER-LMM compared with FaST-LMM across all STARLIGHT phenotypes. Four conditions, 26,414 phenotypes and 95,804 variants. Panels **A** and **B** show per-variant $-\log_{10}(p)$ and signed per-variant effects, respectively. Both are displayed as hexagonal bins with cell counts on a shared logarithmic scale; the dashed line corresponds to $y = x$. Correlations are calculated over all 2.53 billion pairs. Displayed cells are based on a uniform one-in-500 sample, with counts rescaled to represent the full variant density. Panel **C** shows the distribution of per-phenotype correlations for $-\log_{10}(p)$, zoomed to the narrow range near 1; the seven phenotypes below 0.999995 are excluded from the displayed range.

Conclusion

FaST-ER-LMM, in a nutshell, is faster. It does not introduce a new association model; it reorganizes the FaST-LMM computation for GPU hardware and shares genotype-dependent work across phenotypes. In our usual workflow, a transcriptome-wide scan falls from about 130 hours on CPUs to ten minutes on one V100S GPU, a wall-clock speedup of approximately 780× that avoids a 13,000-CPU-hour cluster allocation.

The complete scan fits on a single GPU workstation instead of requiring a large cluster allocation. The phenotype block size B can be adjusted to the available GPU memory: smaller GPUs process smaller blocks and share the fixed calculations across fewer phenotypes, but the statistical analysis itself is unchanged.

The accelerated implementation remains close to the reference, with an overall correlation above 0.999998 for association statistics and identical Bonferroni-significant variant–phenotype pairs in the full-scale comparison. Reducing the runtime from days to minutes allows alternative covariates, quality-control thresholds, phenotype transformations and environmental contrasts to be compared instead of frozen because each new scan is too expensive. It also makes larger interaction analyses, including genotype-by-environment and epistasis models with more strains, computationally feasible, although those models require statistical machinery beyond the single-marker scan implemented here.

The code is available at [gbrach/FaST-ER-LMM](https://github.com/gbrach/FaST-ER-LMM).

The underlying LMM is from Lippert et al. (2011), available at [fastlmm/FaST-LMM](https://github.com/fastlmm/FaST-LMM).

References

- Beasley TM, Erickson S, Allison DB (2009) Rank-Based Inverse Normal Transformations are Increasingly Used, But are They Merited? *Behav Genet* **39**: 580–595.
- Becerra-Rodríguez C, Thiele P, Brach G, Dutta A, Garin M, Tan G, et al. (2026) Domestication drives repeated evolution of sexual–asexual life cycle trade-offs in yeast. *Proc Natl Acad Sci* **123**: e2526682123.
- Caudal É, Loegler V, Dutreux F, Vakirlis N, Teyssonnière É, Caradec C, et al. (2024) Pan-transcriptome reveals a large accessory genome contribution to gene expression variation in yeast. *Nat Genet* **56**: 1278–1287.
- Connelly CF, Akey JM (2012) On the Prospects of Whole-Genome Association Mapping in *Saccharomyces cerevisiae*. *Genetics* **191**: 1345–1353.
- Devlin B, Roeder K (1999) Genomic Control for Association Studies. *Biometrics* **55**: 997–1004.
- Dutta A, Garin M, Loegler V, Brach G, Friedrich A, Yoshimura M, et al. (2026) Population-scale chemical response revealed by a barcoded yeast collection. *Nat Commun*: 2025.12.03.691428.
- Fisher R (1918) The correlation between relatives on the supposition of mendelian inheritance. *Trans. Roy. Soc.*
- Fisher RA (1919) The causes of human variability. *Eugen Rev* **10**: 213–220.
- Fournier T, Abou Saada O, Hou J, Peter J, Caudal E, Schacherer J (2019) Extensive impact of low-frequency variants on the phenotypic landscape at population-scale. *eLife* **8**.
- Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, et al. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**: 348–354.
- Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. (2008) Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics* **178**: 1709–1723.
- Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011) FaST linear mixed models for genome-wide association studies. *Nat Methods* **8**: 833–835.
- Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D (2012) Improved linear mixed models for genome-wide association studies. *Nat Methods* **9**: 525–526.
- Loegler V, Friedrich A, Schacherer J (2024) Overview of the *Saccharomyces cerevisiae* population structure through the lens of 3,034 genomes. *G3 GenesGenomesGenetics* **14**: jkae245.
- Loegler V, Runge J-N, Eberlein C, Peltier É, Harrouard J, Miot-Sertier C, et al. (2025a) Whole-genome sequencing of 1,060 *Brettanomyces bruxellensis* isolates reveals significant phenotypic impact of acquired subgenomes in allopolyploids. *Nat Commun* **16**: 5500.
- Loegler V, Thiele P, Teyssonnière E, Tsouris A, Brach G, Cruaud C, et al. (2025b) From genotype to phenotype with 1,086 near telomere-to-telomere yeast genomes. *Nature*: 1–10.
- Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. (2019) PyTorch: An imperative style, high-performance deep learning library. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 8026–8037.
- Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, et al. (2018) Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **556**: 339–344.

- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909.
- Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11**: 459–463.
- Teyssonnière EM, Trébulle P, Muenzner J, Loegler V, Ludwig D, Amari F, et al. (2024) Species-wide quantitative transcriptomes and proteomes reveal distinct genetic control of gene expression variation in yeast. *Proc Natl Acad Sci* **121**: e2319211121.
- Tsouris A, Brach G, Friedrich A, Hou J, Schacherer J (2024) Diallel panel reveals a significant impact of low-frequency genetic variants on gene expression variation in yeast. *Mol Syst Biol* **20**: 362–373.
- VanRaden PM (2008) Efficient Methods to Compute Genomic Predictions. *J Dairy Sci* **91**: 4414–4423.
- Visscher PM, Bruce Walsh J (2019) Commentary: Fisher 1918: The foundation of the genetics and analysis of complex traits. *Int J Epidemiol* **48**: 10–12.
- Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**: 565–569.
- Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**: 203–208.
- Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, et al. (2010) Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* **42**: 355–360.

R

D'un seul génome à des milliers, et au-delà...

Gauthier Brach¹ et Joseph Schacherer^{1,2,3,4}

¹ Université de Strasbourg, CNRS, Inserm, IGBMC
UMR 7104 – UMR-S 1258, 67400 Illkirch, France

² Institut Universitaire de France (IUF), 75231 Paris, France

³ CIC bioGUNE, Bizkaia Technological Park, Building 800, 48160 Derio, Espagne

⁴ Ikerbasque, Basque Foundation for Science, Bilbao, Espagne

| Résumé

En trois décennies, la levure *Saccharomyces cerevisiae* est passée du statut de premier eucaryote dont le génome a été entièrement séquencé à celui de l'un des systèmes biologiques les plus étudiés. Les progrès du séquençage, de la génomique fonctionnelle et de la génomique des populations ont étendu l'étude d'un seul génome de référence à des milliers d'isolats naturels et au pangénome de l'espèce. Ils ont révélé une diversité génétique et structurale considérable, façonnée par l'évolution, l'écologie et la domestication, tandis que des ressources expérimentales à grande échelle faisaient de la levure un organisme central en biologie des systèmes. Nous retraçons ici ce passage du génome de référence à la diversité des populations, puis examinons comment les assemblages télomère-à-télomère, les graphes de pangénome et les approches multi-omiques permettent de relier la variation génomique au phénotype. Grâce à ces ressources, *S. cerevisiae* constitue aujourd'hui l'un des principaux modèles eucaryotes pour l'étude et la prédiction des relations génotype-phénotype.

La version anglaise de cette revue a été publiée dans *FEMS Yeast Research*, 2026
(doi 10.1093/femsyr/foag038).

Pour satisfaire aux exigences d'un quota de 10% de texte en français, la traduction de cette revue a été réalisée à l'aide du modèle Opus 5 d'Anthropic. Le texte a ensuite fait l'objet d'une révision complète pour en assurer l'exactitude.

L'intégralité du reste de ce manuscrit a été rédigée sans l'assistance d'outils de génération de texte.

La publication du génome de *Saccharomyces cerevisiae* en 1996 a ouvert une nouvelle période de la biologie eucaryote (Goffeau et al., 1996). Ce premier génome eucaryote séquencé a donné à la communauté une référence autour de laquelle se sont organisées des ressources génomiques, génétiques et fonctionnelles. En trois décennies, le champ d'étude s'est étendu d'une souche de laboratoire à des milliers d'isolats naturels, puis aux pangénomes de l'espèce et aux représentations en graphes (Fig. 1) (Peter et al., 2018; Loegler et al., 2025). Les outils expérimentaux développés en parallèle ont fait de la levure une plateforme unique pour relier la variation génétique aux phénotypes moléculaires et cellulaires. Cette revue suit ces évolutions, du génome de référence à la génomique des populations et à la pangénomique, pour montrer comment *S. cerevisiae* est devenue l'une des espèces les mieux caractérisées.

Construire le cadre génomique de référence de la levure

Le génome compact de *S. cerevisiae* et les outils construits autour de sa séquence en ont fait l'un des systèmes biologiques les mieux caractérisés. Au cours des trois décennies suivant la publication de la référence, la levure a servi de terrain d'essai à presque toutes les avancées majeures de la génomique fonctionnelle, de la biologie des systèmes et de l'ingénierie des génomes (Botstein & Fink, 2011). Le séquençage de S288C n'était que le point de départ : son influence tient surtout aux ressources et aux outils expérimentaux qui ont suivi. Le cadre ainsi établi a ensuite façonné les approches appliquées à l'ensemble des eucaryotes.

Le premier génome de référence eucaryote

Le séquençage du génome de *S. cerevisiae* fut une étape majeure de la génomique. Lancé en 1989 sous l'impulsion d'André Goffeau, le projet a adopté une organisation alors inédite en biologie. Au lieu de concentrer le travail dans quelques plateformes dédiées, le consortium a réparti des segments de chromosomes entre des dizaines de laboratoires en Europe, en Amérique du Nord et au Japon. Cette organisation décentralisée reflétait à la fois les limites technologiques de l'époque et la tradition de collaboration de la communauté des levuristes (Goffeau & Vassarotti, 1991; Dujon, 1992; Dujon & Cooper, 2026).

Le séquençage pilote du chromosome III a démontré à la fois la faisabilité et la valeur scientifique d'une séquence génomique complète. Alors que *S. cerevisiae* figurait déjà parmi les organismes eucaryotes les mieux étudiés, l'analyse du chromosome III a révélé que près de 80 % des cadres ouverts de lecture (ORF, *open reading frames*) prédits ne correspondaient pas à des gènes déjà connus (Oliver et al., 1992). Le résultat était frappant : de nombreux gènes restaient entièrement non caractérisés même chez l'eucaryote le plus intensivement étudié, ce qui suggérait que le séquençage pouvait révéler de vastes pans de biologie inexplorée.

Quatre ans plus tard, le consortium publiait la séquence complète de 12,07 Mb, couvrant 16 chromosomes et 5 885 gènes prédits comme codant des protéines, dans une publication simplement intitulée « Life with 6000 genes » (Goffeau et al., 1996;

Mewes et al., 1997). Il s'agissait du premier génome eucaryote complet assemblé et, à l'époque, du plus grand génome séquencé. La référence révélait une architecture compacte, caractérisée par une forte densité génique, très peu d'introns et de courtes régions intergéniques. Environ 40 % des gènes pouvaient se voir attribuer une fonction sur la base des données existantes, tandis que des milliers restaient entièrement non caractérisés (Dujon, 1996; Mewes et al., 1997). Le génome de référence est aussi devenu le système de coordonnées utilisé pour organiser les données suivantes. Une nomenclature standardisée a été élaborée pour identifier de manière univoque chaque gène, chaque position chromosomique et chaque élément génomique (Thierry et al., 1990; Jacquier et al., 1992). Parallèlement, la base de données *Saccharomyces* Genome Database (SGD, <https://www.yeastgenome.org/>) est devenue le dépôt central des séquences, des annotations, des interactions génétiques et des données fonctionnelles (Cherry et al., 1998, 2012). Elle est aujourd'hui l'une des bases de données biologiques les plus complètes et rassemble des informations génomiques, transcriptomiques, protéomiques et phénotypiques (Engel et al., 2022).

L'influence de cette référence s'est étendue bien au-delà de la biologie de la levure. La séquence de S288C a démontré que les approches à l'échelle du génome pouvaient être appliquées systématiquement aux eucaryotes et a fourni un modèle pour le séquençage ultérieur des plantes, des animaux et de l'humain.

Établir un cadre évolutif

Une séquence génomique complète a également permis d'étudier l'évolution des génomes. Des études comparatives ont rapidement montré que *S. cerevisiae* portait la signature d'une ancienne duplication complète du génome (WGD, *whole-genome duplication*), une découverte qui a profondément modifié les conceptions de l'évolution des génomes eucaryotes (Wolfe & Shields, 1997; Seoighe & Wolfe, 1999).

Les blocs chromosomiques dupliqués dispersés dans tout le génome indiquaient que l'ancêtre de la levure bourgeonnante *S. cerevisiae* avait subi un doublement complet de son matériel génétique il y a environ cent millions d'années. Les comparaisons avec des espèces apparentées n'ayant pas subi la duplication, telles que *Ashbya gossypii* et *Kluyveromyces waltii* ont confirmé ce scénario. Elles ont permis de reconstruire les événements évolutifs qui ont suivi la duplication (Dietrich et al., 2004; Kellis et al., 2004).

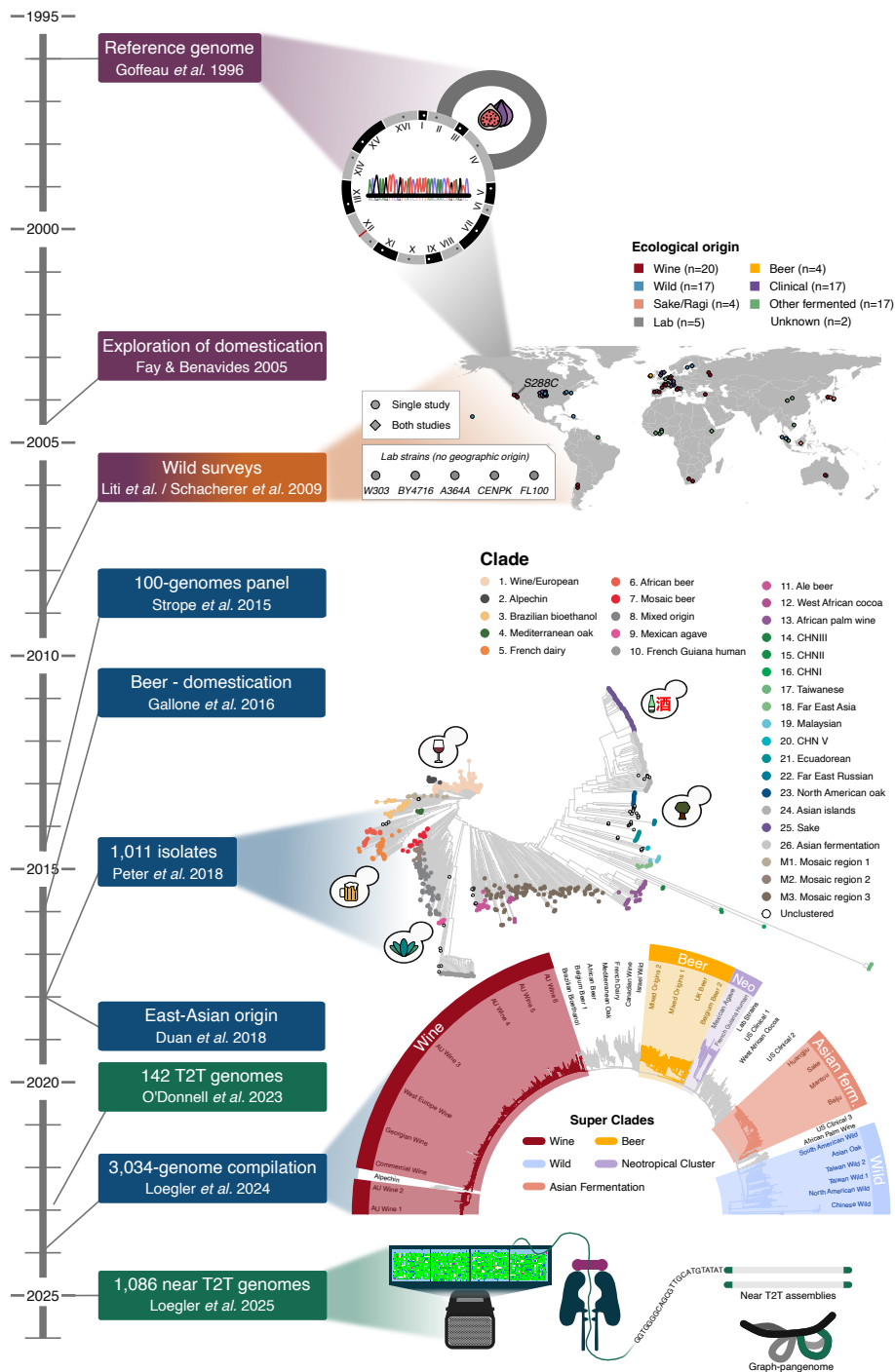


Figure 1 | Chronologie des principales ressources ayant élargi la génomique de la levure.

Les études marquantes sont disposées le long d'un axe temporel vertical (1995–2026). Les étiquettes indiquent la technologie dominante : Sanger, puces, lectures courtes ou lectures longues. Le génome de référence de S288C a fourni le système de coordonnées fondateur du domaine (Goffeau et al. 1996; représentation circulaire des 16 chromosomes avec un tracé Sanger représentatif). Le séquençage multilocus d'isolats écologiquement divers a d'abord distingué les populations domestiquées des populations sauvages (Fay & Benavides 2005). Les premières prospections de populations (carte du monde; Liti et al. 2009 et Schacherer et al. 2009) ont révélé des populations sauvages et domestiquées structurées. Le réséquençage en lectures courtes a élargi à la fois l'échantillonnage et la portée phénotypique de l'espèce, à travers un panel de 100 génomes cliniques et environnementaux (Strope et al. 2015) et l'histoire de domestication des levures de bière industrielles (Gallone et al. 2016). Le projet 1002 Yeast Genomes (1 011 isolats; Peter et al. 2018) a résolu la structure à l'échelle de l'espèce en 26 clades et trois groupes mosaïques (arbre par neighbor-joining). Un échantillonnage dédié a en outre identifié l'Extrême-Orient asiatique comme centre d'origine des lignées domestiquées (Duan et al. 2018). Une compilation de 3 034 génomes (Loegler et al. 2024) a ensuite consolidé les collections en lectures courtes accumulées, dont l'organisation en super-clades est montrée dans le dendrogramme semi-circulaire (Wine, Beer, Wild, Asian fermentation, Neotropical; le groupe Neotropical est décrit plus finement par Avelar-Rivas et al. 2026), avant que le séquençage en lectures longues ne produise des panels de référence complets (142 assemblages télomère-à-télomère; O'Donnell et al. 2023) et des assemblages quasi télomère-à-télomère pour 1 086 isolats ainsi qu'un graphe de pangénome (Loegler et al. 2025).

Bien que la plupart des gènes dupliqués aient ensuite été perdus, plus de cinq cents paires d'ohnologues, des duplicats issus de la WGD, subsistent dans le génome de *S. cerevisiae*. Leur étude a permis de distinguer plusieurs devenir évolutifs. Certains ont réparti les fonctions ancestrales entre les copies, un processus appelé sous-fonctionnalisation, tandis que d'autres ont acquis de nouvelles fonctions par néofonctionnalisation (Dujon et al., 2004; Conant & Wolfe, 2008). Il a été proposé que, dans certains cas, la rétention reflète une sélection en faveur d'une augmentation du dosage génique plutôt qu'une sous- ou néofonctionnalisation. Par exemple, les copies conservées de gènes liés à la glycolyse augmentent le flux glycolytique, un changement associé à l'évolution de l'effet Crabtree et au mode de vie fermentaire de *S. cerevisiae* (Piškur, 2001; Conant & Wolfe, 2007).

La WGD a également fourni une explication évolutive à de nombreuses observations qui émergeraient plus tard de la génomique fonctionnelle. Les gènes conservant des paralogues présentaient souvent une forte redondance génétique, et de nombreuses interactions de létalité synthétique impliquaient des paires de gènes dupliqués qui continuaient à se compenser mutuellement plus de cent millions d'années après l'événement de duplication (Guan et al., 2007; Dean et al., 2008; DeLuna et al., 2008).

Au-delà de la duplication, des projets de génomique comparative tels que Génolevures ont étendu l'analyse à l'ensemble du sous-embranchement des Saccharomycotina (Souciet et al., 2000; Dujon et al., 2004; The Génolevures Consortium et al., 2009; Shen et al., 2018b; Opulente et al., 2024). La comparaison de plusieurs espèces a identifié d'importantes pertes de gènes, des réarrangements chromosomiques et des acquisitions horizontales. Elle a montré que mutations ponctuelles et changements

structuraux de grande échelle modifient conjointement le contenu génique et l'architecture des chromosomes. Le génome de la levure apparaissait ainsi non comme une référence statique, mais comme un système évolutif dynamique.

De la séquence à la fonction

Une fois le génome de référence disponible, l'attention s'est déplacée de la découverte des gènes vers leur caractérisation fonctionnelle, portée par des innovations technologiques qui ont lancé l'ère de la génomique fonctionnelle. L'une des premières avancées a été le développement du profilage transcriptionnel à l'échelle du génome. Dans les deux ans qui ont suivi la publication du génome, des puces à ADN (*microarrays*) étaient déjà disponibles, permettant de suivre simultanément les variations d'expression de presque tous les gènes de levure (DeRisi et al., 1997). Ces expériences ont révélé les programmes transcriptionnels sous-jacents aux grandes fonctions biologiques, dont la transition diauxique, la progression du cycle cellulaire, la sporulation et les réponses au stress environnemental (Chu et al., 1998; Spellman et al., 1998; Gasch et al., 2000).

La capacité à mesurer des milliers de gènes simultanément a changé la façon dont les biologistes concevaient la fonction. Au lieu d'étudier les gènes individuellement, il devenait possible d'identifier des groupes de gènes co-régulés et d'inférer des fonctions biologiques à partir de profils d'expression partagés (Hughes et al., 2000). Peu après, l'immunoprécipitation de la chromatine couplée aux puces à ADN a permis de produire les premières cartes à l'échelle du génome des sites de fixation des facteurs de transcription et une vision systémique de la régulation génique (Lee et al., 2002; Harbison et al., 2004).

L'abondance des transcrits ne décrivait toutefois qu'un niveau de l'organisation cellulaire. La protéomique a rapidement apporté une mesure complémentaire (Gygi et al., 1999). L'identification des protéines par spectrométrie de masse à l'échelle du génome a montré que l'abondance des ARNm prédit souvent mal celle des protéines (Foss et al., 2007; Teyssonnière et al., 2024). Résoudre cette discordance demeure un obstacle majeur pour la biologie prédictive et montre que les mécanismes post-transcriptionnels peuvent façonner les phénotypes cellulaires.

D'autres travaux ont porté sur le génome mitochondrial (Foury et al., 1998), l'architecture tridimensionnelle du génome et la chromatine. La capture de conformation chromosomique (3C), développée chez la levure, a permis d'étudier directement l'architecture tridimensionnelle du génome et a ouvert la voie aux études modernes des interactions de la chromatine (Dekker et al., 2002). Les premières cartes du positionnement des nucléosomes et des modifications d'histones à l'échelle du génome ont ensuite montré que même un génome compact possède une organisation de la chromatine fortement structurée (Pokholok et al., 2005; Yuan et al., 2005).

Au milieu des années 2000, il était ainsi possible d'analyser simultanément chez la levure la séquence génomique, l'abondance des transcrits et des protéines, l'état de la chromatine et l'architecture nucléaire. Le génome n'était plus étudié comme une

séquence linéaire isolée, mais comme l'un des niveaux d'un système de régulation cellulaire. Peu d'autres systèmes offraient alors une vision intégrée comparable de la fonction cellulaire eucaryote.

Construire une boîte à outils expérimentale à l'échelle du génome

Si le génome de référence représentait le plan, la génomique fonctionnelle exigeait la capacité d'en manipuler systématiquement chaque composant. L'efficacité de la recombinaison homologue chez la levure a rendu cela possible et a conduit à la création d'une collection de ressources à l'échelle du génome sans équivalent chez d'autres organismes.

L'une des ressources les plus influentes fut la collection de mutants de délétion de la levure (*yeast knockout collection*) (Winzeler et al., 1999 ; Giaever et al., 2002). Presque tous les gènes non essentiels du génome ont été remplacés par une cassette de délétion portant des séquences codes-barres uniques. La collection de délétions pouvait donc être regroupée et analysée simultanément, ce qui permettait de mesurer les effets sur la fitness à l'échelle du génome. Comme environ un cinquième des gènes de levure sont essentiels, des ressources complémentaires ont été développées pour étudier les gènes qui ne pouvaient pas être délétés. Des allèles répressibles par la tétracycline ont permis la déplétion conditionnelle de protéines essentielles (Mnaimneh et al., 2004), tandis que les souches DAmP réduisaient l'expression en déstabilisant les transcrits (Breslow et al., 2008). Des collections de mutants thermosensibles ont ensuite étendu la couverture en permettant l'inactivation conditionnelle de centaines de gènes essentiels (Ben-Aroya et al., 2008 ; Li et al., 2011). Ces ressources rendaient ainsi presque tous les gènes accessibles à une perturbation expérimentale.

En parallèle, de vastes projets de marquage GFP ont produit le premier atlas de localisation à l'échelle du protéome, et ont cartographié simultanément la distribution subcellulaire de milliers de protéines (Huh et al., 2003). Les collections TAP-tag ont étendu l'approche à la quantification et à la purification des protéines et ont rendu possible la reconstruction des complexes protéiques à grande échelle (Ghaemmaghami et al., 2003 ; Gavin et al., 2006 ; Krogan et al., 2006).

Ces ressources ont permis de reconstruire des réseaux d'interactions moléculaires plutôt que d'analyser les protéines une à une. Des cribles double-hybride complémentaires ont cartographié systématiquement les interactions binaires protéine-protéine (Uetz et al., 2000 ; Ito et al., 2001).

La réussite la plus ambitieuse de cette période fut sans doute la technologie Synthetic Genetic Array (SGA) (Tong et al., 2001). En automatisant la construction de doubles mutants, la SGA a rendu possible l'exploration systématique des interactions génétiques. Les générations successives de cartes d'interaction sont passées de centaines à des millions de paires de gènes testées, et ont produit le plus grand réseau d'interactions génétiques disponible à ce jour (Costanzo et al., 2010, 2016). Ces études ont révélé que les processus cellulaires sont organisés en modules fonctionnels étroitement

interconnectés et ont démontré que les profils d'interaction génétique prédisent souvent la fonction mieux que les seules interactions physiques.

Ces mêmes collections ont aussi permis l'émergence de la chimiogénomique. Le profilage de fitness à l'échelle du génome en présence de composés bioactifs a permis d'identifier de manière systématique des cibles de composés bioactifs et des voies de résistance (Giaever et al., 2004; Hillenmeyer et al., 2008; Piotrowski et al., 2017). Cela a fait de la levure une plateforme à la fois pour la biologie fondamentale et pour la découverte pharmaceutique.

L'aboutissement de cette tradition d'ingénierie est incarné par le projet Sc2.0 : depuis 2011, il vise à reconcevoir et reconstruire de manière synthétique l'intégralité du génome de la levure (Dymond et al., 2011; Annaluru et al., 2014; Richardson et al., 2017; Zhao et al., 2023). En démontrant progressivement la faisabilité de la construction d'un génome eucaryote synthétique, Sc2.0 fournit une plateforme pour l'évolution expérimentale et l'ingénierie des génomes à grande échelle, et prolonge dans une ère nouvelle la vision originelle de la biologie à l'échelle du génome (Shen et al., 2018a; Jin et al., 2020).

Un cadre pour la génomique moderne

Au début des années 2010, aucun autre eucaryote ne réunissait un ensemble comparable de ressources : une séquence génomique complète, une base de données richement annotée, des atlas d'expression et de protéome, des collections de mutants presque exhaustives, de vastes réseaux d'interactions et des mesures à l'échelle de l'espèce de plus en plus élaborées. Toutes étaient organisées autour du génome de référence S288C et formaient un cadre intégré pour la génomique moderne.

Découvrir que la souche S288C n'est pas représentative de l'espèce

La première décennie qui a suivi la publication du génome de référence de *S. cerevisiae* a par conséquent été dominée par le développement de ressources de génomique fonctionnelle bâties presque exclusivement dans le fond génétique S288C (Baker Brachmann et al., 1998; Schacherer et al., 2007). La stratégie a bien fonctionné, mais elle a aussi renforcé l'hypothèse implicite selon laquelle la souche de référence représentait correctement la biologie de l'espèce dans son ensemble. La multiplication des études à l'échelle des populations au début des années 2000 a progressivement infirmé cette hypothèse. Ce qui ressemblait à une espèce assez homogène s'est révélé être une population fortement structurée, façonnée par l'écologie, la domestication, l'hybridation et l'isolement géographique. Le domaine s'est donc déplacé de la compréhension d'un génome de référence à la compréhension d'une espèce.

Bâtir tout un domaine de recherche sur un seul génome entraîne trois limites distinctes. Premièrement, la souche de référence est elle-même atypique. Deuxièmement, la cartographie fondée sur une référence ne détecte que ce que la référence contient déjà. Enfin, un système de coordonnées linéaire ne peut représenter des séquences que certains isolats portent et que d'autres ne possèdent pas. La première limite est devenue manifeste dès que S288C a été comparée à des isolats naturels.

S288C, une représentante peu fidèle de l'espèce

Les premiers signes que S288C pourrait ne pas être représentative sont venus de la caractérisation détaillée de la souche elle-même. Bien qu'elle dérive en grande partie de l'isolat sauvage EM93 isolé d'une figue en décomposition, des décennies de propagation en laboratoire et de croisements sélectifs avaient entraîné l'accumulation de mutations qui la distinguaient de nombreux isolats naturels (Mortimer & Johnston, 1986; Engel et al., 2013). Plusieurs de ces mutations affectaient des caractères écologiquement importants dans une population sauvage. La souche de référence porte un allèle *HAP1* défectueux résultant de l'insertion d'un élément transposable Ty1; *HAP1* code pour un activateur transcriptionnel dépendant de l'hème qui couple l'expression des gènes respiratoires à la disponibilité en oxygène (Gaisne et al., 1999). Elle contient également des mutations dans *FLO1* et *FLO8* qui abolissent la floculation et la croissance filamenteuse, deux caractères qui contribuent à la formation de biofilms, à la colonisation des surfaces et à la persistance environnementale (Liu et al., 1996; Louis, 2016). D'autres allèles de perte de fonction affectent l'utilisation du galactose, le métabolisme du mélibiose et la biosynthèse de la biotine, des caractères fréquemment intacts chez les isolats naturels (Louis, 2016). Certaines de ces différences sont précisément ce qui rend S288C commode en laboratoire. Ainsi, la perte de la floculation et de la croissance filamenteuse maintient les cellules uniformément dispersées en culture liquide. Cette dispersion permet des mesures de densité optique stables et précises, et donc des expériences plus reproductibles. De même, les auxotrophies fournissent des marqueurs de sélection essentiels, nécessaires

à la manipulation génétique. Les souches de laboratoire standard, comme les souches BY dérivées de S288C, ont été construites avec des déficiences spécifiques dans des gènes de biosynthèse tels que *URA3*, *LEU2*, *HIS3* et *MET15*, permettant de sélectionner les transformants plasmidiques et les remplacements de gènes sur milieux synthétiques carencés (*drop-out*) (Baker Brachmann et al., 1998). D'autres différences, comme l'allèle *HAP1* défectueux, ne constituent pas des avantages de laboratoire mais sont simplement tolérées sur milieu riche. Cette commodité s'accompagnait d'une réserve : la souche de référence avait été façonnée par son histoire en laboratoire et non choisie pour représenter la diversité de l'espèce. La question devenait alors : quelle part de la biologie observée chez S288C reflète des propriétés communes à *S. cerevisiae*, et quelle part tient aux particularités d'une lignée de laboratoire ?

Les premiers aperçus de la diversité des populations

Les premières tentatives de réponse à cette question reposaient sur des approches antérieures au séquençage de génomes complets. Des analyses sur puces à ADN ont détecté des polymorphismes de sonde unique (SFP, *single-feature polymorphisms*) au sein de petites collections de souches et ont montré qu'une variation génétique substantielle existait même entre des isolats jusqu'alors réputés étroitement apparentés (Winzeler et al., 2003 ; Schacherer et al., 2007). En 2005, l'hybridation génomique comparative sur puces (aCGH) a étendu la comparaison des polymorphismes de séquence au nombre de copies de gènes. Quatre souches de vin commerciales ont été profilées par rapport à S288C, et chacune portait des amplifications ou des délétions de gènes isolés, dont beaucoup codaient des transporteurs tels que les transporteurs d'hexoses *HXT* ou les pompes à sodium *ENA*. Les différences étaient modestes en nombre, mais suffisantes pour identifier chaque souche de manière unique (Dunn et al., 2005).

À la même période, des prospections écologiques ont isolé *S. cerevisiae* dans des environnements naturels tels que les habitats associés aux chênes, ce qui a remis en question l'hypothèse répandue selon laquelle l'espèce était principalement domestiquée et associée à l'humain (Sniegowski et al., 2002).

Une image plus systématique a émergé des études multilocus. En examinant une poignée de loci non liés dans des souches issues de vignobles, de fermentations de saké, de boissons africaines traditionnelles et de contextes cliniques, ces études ont démontré que *S. cerevisiae* est composée de populations génétiquement distinctes présentant une structure nette (Fay & Benavides, 2005). Elles ont identifié des lignées domestiquées et sauvages, montrant que les populations associées à l'humain ne constituaient qu'une partie de la diversité évolutive de l'espèce.

Les études écologiques et de génomique des populations ultérieures ont renforcé cette conclusion. Les isolats associés aux chênes des forêts nord-américaines formaient des populations génétiquement cohérentes, clairement différenciées des souches de vignoble (Aa et al., 2006). Des analyses de microsatellites portant sur des centaines d'isolats ont révélé des regroupements distincts associés aux fermentations de vin,

de bière, de pain et de saké, preuve que les pratiques culturelles humaines avaient profondément façonné l'évolution de la levure (Legras et al., 2007).

Deux résultats ont structuré la suite des travaux. La niche écologique peut organiser les populations aussi fortement que la géographie (Peña et al., 2025), et la domestication laisse des signatures génomiques détectables pendant des siècles ou des millénaires.

Le séquençage de génomes complets révèle l'étendue de la diversité génétique

Les études multilocus révélaient une structure de population nette, mais ne décrivaient qu'une petite partie de la diversité génétique. Le séquençage de génomes complets a élargi cette vue.

La première séquence génomique complète d'une souche autre que la référence, YJM789, a démontré l'ampleur de la variation pouvant exister au sein de l'espèce. La comparaison avec S288C a identifié environ 60 000 polymorphismes mononucléotidiques (SNP, *single nucleotide polymorphisms*) et des milliers de polymorphismes d'insertion-délétion, correspondant à environ 0,6 % de divergence nucléotidique (Wei et al., 2007). Elle contenait également des traces d'introgression en provenance de l'espèce sœur *Saccharomyces paradoxus*, suggérant que le flux de gènes interspécifique contribue à la diversité intraspécifique. Des comparaisons supplémentaires portant sur des isolats de vignoble et de chêne ont mis au jour des dizaines de milliers de polymorphismes affectant à la fois des régions codantes et régulatrices (Doniger et al., 2008). L'année suivante, la souche de vin AWRI1631 a été séquencée et comparée à la fois à S288C et à YJM789. Les SNP distinguaient les trois souches à une densité moyenne d'un SNP toutes les 150 pb. Le nombre et l'agencement des éléments Ty différaient également entre elles. De plus, plusieurs régions d'ADN spécifiques à AWRI1631 étaient prédites comme codant des protéines propres à cette souche industrielle (Borneman et al., 2008). Il était désormais clair que les isolats naturels diffèrent sur l'ensemble du génome plutôt qu'en quelques loci.

En 2009, deux études ont donné la première description de la diversité de *S. cerevisiae* à l'échelle du génome. À l'aide de puces à ADN de type tiling (*tiling arrays*), une prospection de 63 isolats écologiquement et géographiquement divers a identifié plus de 100 000 sites ségrégeants et près de 4 000 grands événements de délétion, révélant une vaste variation nucléotidique et structurale (Schacherer et al., 2009). En parallèle, le *Saccharomyces* Genome Resequencing Project (SGRP) a séquencé plus de 70 isolats de *S. cerevisiae* et *S. paradoxus*. Il a distingué cinq lignées majeures de *S. cerevisiae* (Wine/European, Sake, West African, Malaysian et North American) et montré que de nombreuses souches possèdent des génomes mosaïques issus de l'admixture entre des populations auparavant isolées (Liti et al., 2009). Ces résultats ont établi le rôle de la domestication et de la recombinaison dans la diversité de l'espèce, tout en montrant que S288C n'en capture qu'une fraction.

D'un génome de référence à une perspective d'espèce

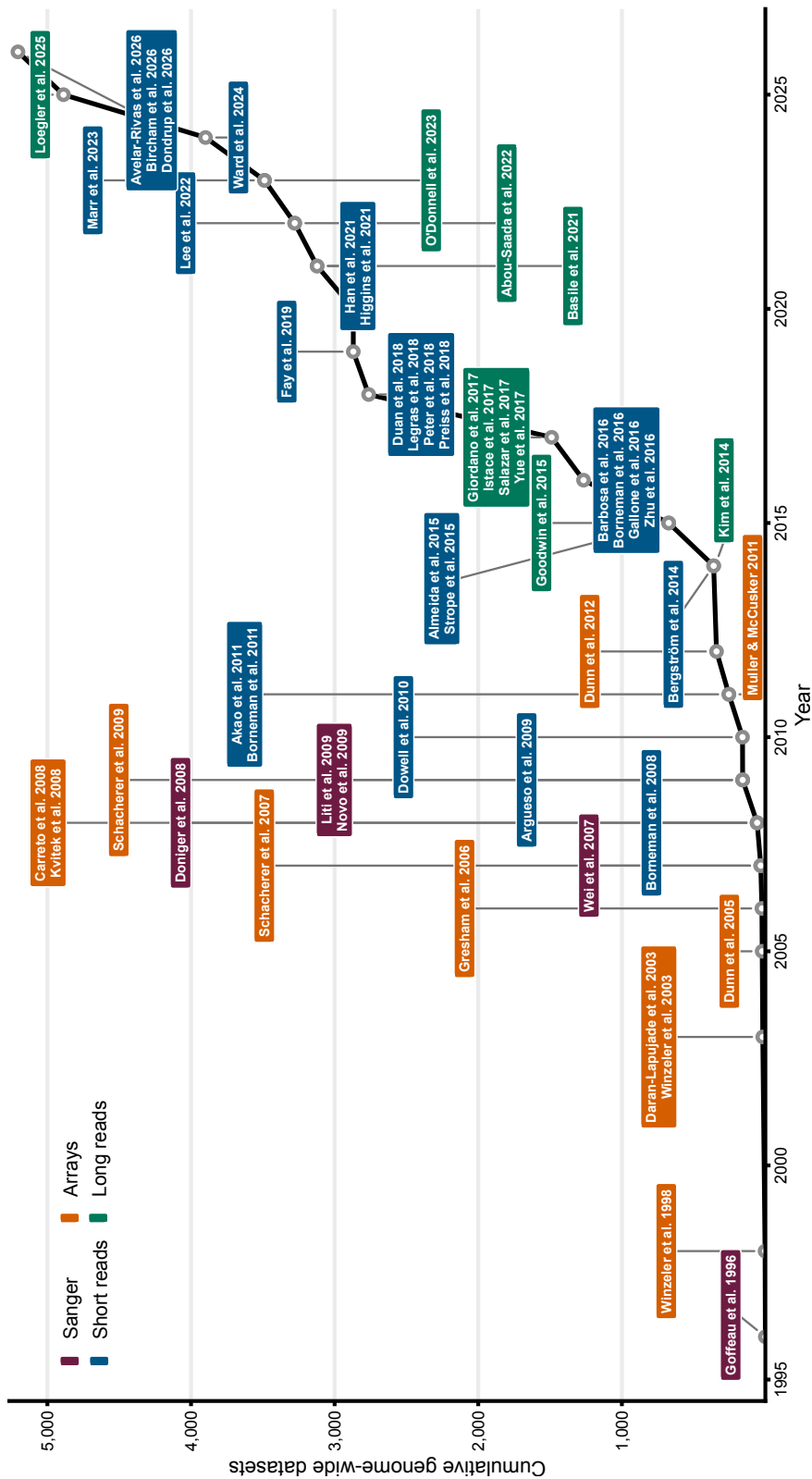
Le génome de référence demeurait essentiel comme cadre expérimental, mais ne pouvait plus tenir lieu d'espèce à lui seul. Il fallait désormais échantillonner la diversité des populations à une échelle suffisante pour en étudier les conséquences fonctionnelles et évolutives, ce qui imposait de passer de dizaines de génomes à des centaines, puis à des milliers.

Passer de centaines à des milliers de génomes

La plupart des études reposaient encore sur des dizaines, ou au mieux quelques centaines, d'isolats. Cet effectif suffisait à identifier les grands groupes de populations, mais pas à résoudre l'histoire évolutive de l'espèce ni à fournir la puissance statistique requise par les études génotype-phénotype (Liti et al., 2009; Schacherer et al., 2009; Connelly & Akey, 2012). Le passage à des milliers de génomes a précisé la structure des populations et les histoires de domestication, tout en fournissant les données nécessaires aux études d'association et à la pangénomique. Outre les études discutées dans cette revue, de nombreuses autres prospections à l'échelle du génome ont contribué à cette trajectoire, des premiers génotypes sur puces aux efforts de séquençage les plus récents (Fig. 2) (Winzeler et al., 1998; Daran-Lapujade et al., 2003; Gresham et al., 2006; Carreto et al., 2008; Kvitek et al., 2008; Dowell et al., 2010; Muller et al., 2011; Barbosa et al., 2016; Borneman et al., 2016; Fay et al., 2019; Basile et al., 2021; Han et al., 2021; Higgins et al., 2021; Lee et al., 2022; Marr et al., 2023; Ward et al., 2024).

Figure 2 | Augmentation cumulative du nombre de souches de *Saccharomyces cerevisiae* caractérisées à l'échelle du génome.

Chaque point correspond à une étude; les étiquettes sont colorées selon la technologie dominante : Sanger, puces, lectures courtes ou lectures longues.



Élargir le recensement génomique de l'espèce

La génomique des populations s'est développée relativement tard chez la levure par rapport à d'autres organismes modèles. Au début des années 2010, des consortiums dédiés avaient déjà assemblé de vastes collections représentatives de la diversité dans d'autres systèmes modèles : 168 (puis 205) lignées consanguines pour *Drosophila melanogaster* (Mackay et al., 2012; Huang et al., 2014), 2 504 individus pour l'humain (Auton et al., 2015) et 1 135 génomes pour *Arabidopsis thaliana* (Alonso-Blanco et al., 2016), tandis que les études de population chez *S. cerevisiae* restaient encore limitées à quelques centaines de génomes, dont beaucoup dérivés d'isolats de laboratoire haploïdes ou de dérivés monospores (Liti et al., 2009; Schacherer et al., 2009; Connelly & Akey, 2012). Ces collections ne capturaient qu'une fraction de la diversité naturelle présente au sein de l'espèce et masquaient souvent des caractéristiques importantes telles que l'hétérozygotie, la polyploidie et la variation structurale. La relation génotype-phénotype chez la levure était principalement disséquée par des croisements expérimentaux et la génétique inverse, des voies qui ne requéraient que deux parents et permettaient de mesurer directement l'effet des variants (Liti & Louis, 2012); de vastes panels d'isolats naturels n'étaient donc pas l'outil évident. Son écologie à l'état sauvage, en outre, n'a été reconnue que tardivement : longtemps considérée comme un organisme « fabriqué par l'homme », lié à la fermentation et dépourvu de niche écologique claire, *S. cerevisiae* n'a commencé à être échantillonnée systématiquement dans la nature que dans les années 2000 (Goddard & Greig, 2015; Liti, 2015).

Les premiers efforts pour accroître l'échantillonnage se sont concentrés sur des niches écologiques et des environnements industriels spécifiques. Des prospections par séquençage de souches de bière, de vin, de bioéthanol, de boulangerie et cliniques ont révélé que la variation du contenu génique, les changements de nombre de copies et les introgressions étaient bien plus répandus qu'on ne l'avait cru jusque-là (Borneman et al., 2011; Dunn et al., 2012; Bergström et al., 2014; Strope et al., 2015). Ces études ont aussi introduit le concept de pangénome de levure fondé sur les gènes, et reconnu qu'aucun génome seul ne pouvait contenir le répertoire complet des gènes présents au sein de l'espèce (Loegler et al., 2026).

Ces collections ont aussi montré que la diversité des populations ne se limite pas aux variants nucléotidiques. Les régions subtélomériques sont enrichies en variants du nombre de copies, en gains et pertes de gènes et en séquences acquises horizontalement. Beaucoup de ces gènes participent à l'adaptation environnementale, notamment au métabolisme des sucres, à la résistance au stress, à la floculation et à l'acquisition des nutriments. L'espèce comprend ainsi un noyau génomique très conservé et un génome accessoire qui évolue rapidement (Brown et al., 2010; Dunn et al., 2012; Bergström et al., 2014; Peter et al., 2018).

Le projet 1002 Yeast Genomes

Le projet 1002 Yeast Genomes a porté la génomique des populations à l'échelle de l'espèce et fourni le premier jeu de données de cette ampleur pour *S. cerevisiae* (Peter et al., 2018). Il a séquençé plus de mille isolats issus de nombreuses niches écologiques et régions géographiques. Plus de 1,6 million de SNP ont été identifiés, et la structure des populations a été résolue en 26 clades et trois grands groupes mosaïques. Les variations de ploïdie, l'aneuploïdie, l'introggression et la variation du contenu génique se sont révélées fréquentes plutôt qu'exceptionnelles.

La caractérisation du génome accessoire a révélé qu'un noyau d'environ 5 000 gènes est partagé par tous les isolats, tandis que près de 3 000 gènes supplémentaires varient en fréquence entre les souches. Ces loci accessoires sont enrichis dans les régions subtelomériques et associés à l'adaptation environnementale, aux réponses au stress et aux interactions intercellulaires. La variation du nombre de copies et du contenu génique constitue donc un moteur de la diversité phénotypique des populations.

L'introggression s'est également révélée être une caractéristique déterminante de l'espèce (Steensels et al., 2021). Chaque isolat portait des traces de flux de gènes en provenance d'espèces apparentées de *Saccharomyces*, en particulier *S. paradoxus* (Peter et al., 2018). Ces régions introgressées contenaient fréquemment des gènes à valeur adaptative potentielle et illustraient l'importance de l'hybridation comme source d'innovation évolutive.

La variation de la ploïdie était tout aussi fréquente. Les génomes polyploïdes étaient communs dans les populations domestiquées, en particulier dans les lignées de brasserie, tandis que l'aneuploïdie touchait près d'un cinquième de tous les isolats. Les souches cliniques présentaient des taux particulièrement élevés d'instabilité chromosomique, ce qui suggère que la plasticité du génome pourrait faciliter l'adaptation aux environnements associés à l'humain (Zhu et al., 2016; Peter et al., 2018).

Le jeu de données 1002 Genomes est devenu la référence pour la génomique des populations de levure et a fourni la puissance statistique nécessaire pour étudier l'histoire des populations, l'adaptation et la variation des caractères quantitatifs à une échelle jusque-là inaccessible.

Reconstruire les histoires de domestication

L'augmentation de l'échantillonnage a permis de reconstruire l'histoire de la domestication de la levure. De nombreuses populations de *S. cerevisiae* sont étroitement associées à la production alimentaire humaine, et leurs génomes conservent des signatures de ces pratiques (Marsit & Dequin, 2015; Legras et al., 2018).

Les souches de vin en sont l'un des exemples les plus clairs. L'intégration d'isolats de chènes méditerranéens nouvellement séquençés a démontré que ces populations représentent les plus proches parents sauvages connus de la lignée Wine/European

(Almeida et al., 2015), ce qui est cohérent avec les pratiques de vinification autour du bassin méditerranéen. Des analyses démographiques suggéraient que la séparation entre populations sauvages et populations vinicoles domestiquées s'est produite il y a environ 1300 à 10000 ans, ce qui concorde globalement avec le développement de la viticulture (Pretorius, 2000; Almeida et al., 2015). Les souches de vin portent des marques de domestication sous forme de transferts horizontaux de gènes (HGT), de variants du nombre de copies (CNV) et de translocations. Trois segments chromosomiques totalisant 120 kb et 39 gènes prédits ont été acquis auprès d'autres espèces de levures, parmi lesquels une paire de gènes en tandem codant des transporteurs d'oligopeptides *FOT* provenant de *Torulaspora microellipsoides*; leur délétion réduit la biomasse de 12 % et laisse du fructose résiduel en fin de fermentation (Novo et al., 2009; Marsit & Dequin, 2015; Becerra-Rodríguez et al., 2020). Un autre exemple concerne une translocation réciproque entre les chromosomes VIII et XVI, qui place *SSU1* sous le promoteur d'*ECM34* et confère une résistance aux sulfites, un caractère potentiellement sélectionné par l'usage du sulfite comme conservateur du vin, commun à la plupart des souches de vin (Pérez-Ortín et al., 2002; Marsit & Dequin, 2015). Enfin, la tolérance au cuivre varie directement avec le nombre de copies de *CUP1*, qui va de zéro à quatorze parmi 132 souches de vin, dont 68 % portent plus d'une copie; cette distribution est attribuée à l'usage du sulfate de cuivre comme fongicide de vignoble (Fogel et al., 1983; Steenwyk & Rokas, 2017) (Fig. 3a).

Les souches de bioéthanol ont accumulé des amplifications de gènes impliqués dans la réponse au stress et de vastes réarrangements chromosomiques associés à la fermentation industrielle. Chez JAY270, le dérivé diploïde de la souche brésilienne PE-2, au moins six chromosomes diffèrent en longueur entre homologues, et les points de cassure tombent dans des répétitions subtélomériques. Les gènes de réponse au stress *SNO* et *SNZ* sont présents en neuf copies contre quatre chez S288C, avec une expression multipliée par quatre (Argueso et al., 2009) (Fig. 3b).

Les souches associées aux produits laitiers portent une version divergente de la voie du galactose. Le cluster natif *GAL7-GAL10-GAL1* est remplacé par des orthologues plus proches de ceux de *S. uvarum* et *S. eubayanus* que de n'importe quel allèle de *S. cerevisiae*, ce qui a été interprété comme une introgression en provenance d'une lignée divergente du genre *Saccharomyces* (Legras et al., 2018). Une modélisation démographique de ces mêmes allèles a par la suite rejeté une introgression récente et daté leur divergence des allèles de référence à environ 3 milliards de générations, antérieure au dernier ancêtre commun du genre. Elle a aussi attribué leur persistance à une sélection équilibrante multilocus, agissant peut-être sur un compromis entre une croissance plus rapide sur galactose et une croissance plus lente sur glucose (Boocock et al., 2021). L'origine de la voie alternative demeure donc contestée, ce qui n'est pas le cas de son association aux environnements laitiers (Fig. 3c).

Les souches de bière se répartissent en deux lignées qui rassemblent ensemble plus de 85 % des isolats de brasserie et semblent issues de deux événements de domestication indépendants (Gallone et al., 2016; Gonçalves et al., 2016; Abou Saada et al., 2022). Leurs génomes présentent une forte incidence de polyploïdie et d'aneuploïdie, et portent des duplications des gènes d'utilisation du maltose (Fig. 3d). Des mutations

de perte de fonction dans *PAD1* et *FDC1*, deux gènes adjacents situés dans une région subtélomérique du chromosome IV et contrôlant ensemble la production de 4-vinylguaïacol, se retrouvent de manière récurrente parmi les souches de bière. Elles abolissent l'arôme phénolique de clou de girofle, indésirable dans la plupart des styles de bière, mais pas dans tous. Toutes les levures de bière ne sont pas industrielles. Alors que le brassage moderne repose désormais souvent sur quelques souches de départ pures, les brasseurs fermiers traditionnels d'Europe du Nord et de l'Est entretiennent des cultures vivantes mixtes qui préservent une diversité de *S. cerevisiae* largement disparue des stocks industriels, avec des signatures de domestication propres, dont un cycle sexué conservé et une plus grande tolérance au stress (Bircham et al., 2026 ; Schacherer, 2026). Les kveik norvégiennes en sont un exemple : ensemencées dans le moût à 28–40 °C, 19 des 25 isolats croissent à 40 °C, une température tolérée par une seule des quatre souches témoins (Preiss et al., 2018), et les prospections génomiques les placent parmi les lignées domestiquées ayant divergé le plus anciennement et peut-être les plus anciennes encore en usage continu (Dondrup et al., 2026).

Les souches de brassage du saké présentent une vaste perte d'hétérozygotie (LOH) sur l'ensemble de leurs génomes (Fig. 3e), une mosaïque de segments homozygotes déjà détectée en 2011 chez la souche de saké Kyokai n° 7, couvrant plus de 80 % du génome (Akao et al., 2011 ; Peter et al., 2018).

Ces prospections de populations ont redéfini la domestication comme un ensemble d'événements évolutifs distincts et parallèles plutôt que comme un événement unique et uniforme. En effet, les environnements associés à l'humain ont imposé de façon répétée de fortes pressions de sélection, ce qui a conduit à des lignées spécialisées adaptées à des niches de fermentation distinctes (Marsit et al., 2017 ; Sampaio & Pontes, 2025). La structure des populations chez *S. cerevisiae* ne peut donc s'expliquer par la seule géographie. Elle reflète les effets combinés de la spécialisation écologique, de la domestication, de la migration et de l'histoire culturelle humaine.

L'Asie de l'Est et la recherche de l'origine de l'espèce

L'échantillonnage à grande échelle a montré que les populations caractérisées jusqu'alors ne représentaient qu'une petite fraction de la diversité mondiale. Les prospections en Chine ont élargi l'étendue évolutive connue de l'espèce et remis en question les hypothèses sur son origine géographique. Les premières études indiquaient déjà que l'Asie de l'Est abritait des populations exceptionnellement diverses (Wang et al., 2012). Des prospections plus vastes dans les forêts et les environnements de fermentation chinois ont ensuite identifié de nombreuses lignées profondément divergentes. Plusieurs sont séparées des populations connues par des distances génétiques supérieures à celles qui distinguent les grands clades domestiqués ailleurs dans le monde (Duan et al., 2018).

Les populations sauvages et domestiquées différaient nettement. Les souches sauvages d'Asie de l'Est présentaient une forte diversité nucléotidique, une faible hétérozygotie et une reproduction sexuée efficace. À l'inverse, les souches domestiquées présentaient une hétérozygotie élevée, une efficacité de sporulation réduite, une forte variation

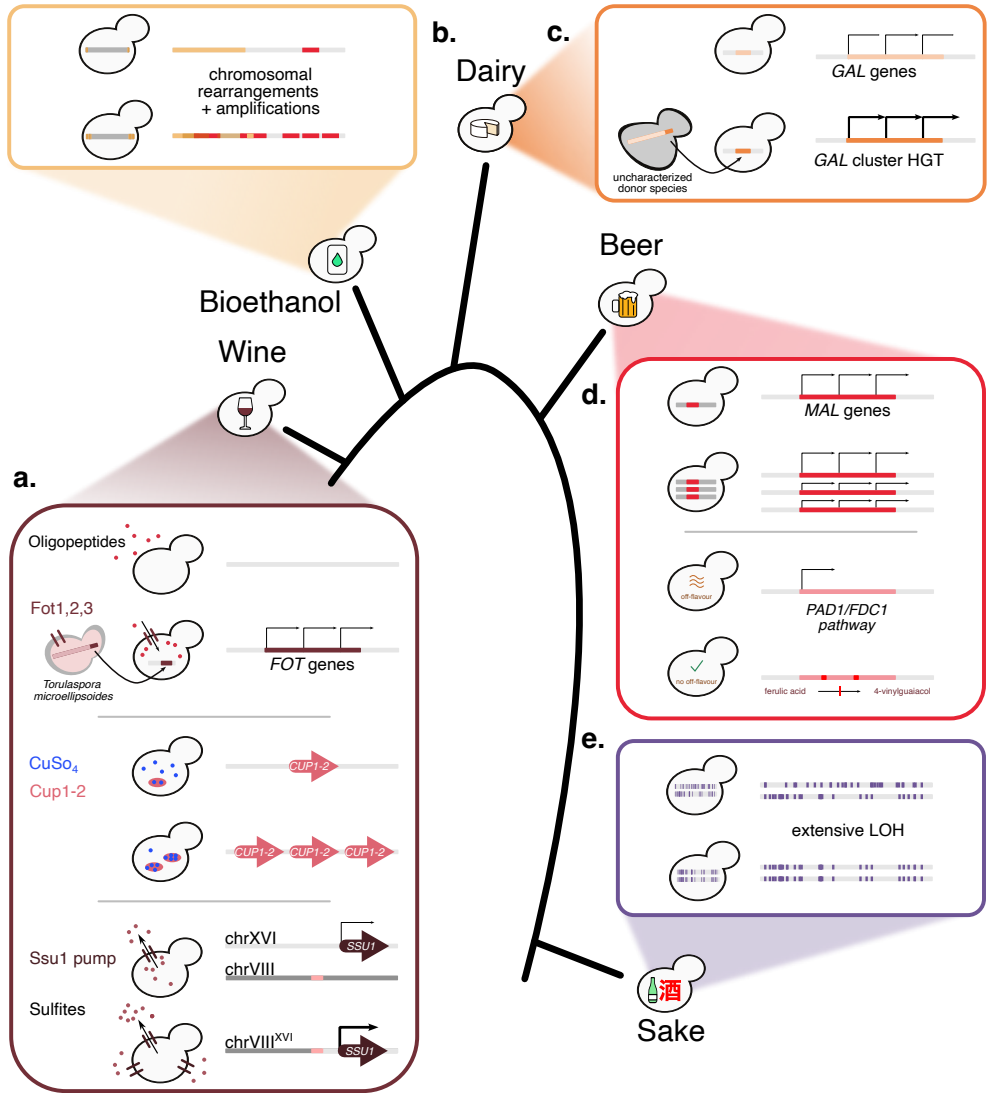


Figure 3 | Signatures génomiques de la domestication à travers les lignées de *Saccharomyces cerevisiae*.

Phylogénie schématique des grands clades domestiqués, chaque branche étant étiquetée par environnement (Wine, Bioethanol, Dairy, Beer, Sake) avec les variations adaptatives caractéristiques indiquées dans les encadrés. **a.** Vin : transporteurs d'oligopeptides *FOT1,2,3* acquis horizontalement depuis *Torulaspora microellipsoides*, amplification de *CUP1-2* (tolérance au cuivre) et translocation *chrVIII*–*chrXVI* entraînant la surexpression de *SSU1* (efflux de sulfites). **b.** Bioéthanol : réarrangements chromosomiques et amplifications subtélomériques. **c.** Produits laitiers : haplotype divergent du cluster de gènes *GAL*, dont l'origine évolutive demeure débattue. **d.** Bière : amplification de gènes *MAL*, polyploïdie et perte de fonction de *PAD1/FDC1* abolissant l'arôme phénolique indésirable. **e.** Saké : vaste perte d'hétérozygotie (LOH).

du nombre de copies et des signatures de sélection artificielle (Duan et al., 2018; Becerra-Rodríguez et al., 2026). La domestication a donc modifié à la fois l'architecture du génome et la biologie de la reproduction.

Les analyses démographiques placent les populations d'Asie de l'Est près de la racine de la phylogénie de l'espèce (Peter et al., 2018). Les lignées chinoises contiennent ensemble plus de diversité que le reste de l'espèce connue, ce qui soutient l'hypothèse d'un réservoir ancestral est-asiatique à l'origine des populations modernes de *S. cerevisiae*. L'origine géographique précise demeure cependant débattue, mais les prospections génomiques à grande échelle ont déplacé le centre de l'histoire évolutive de la levure vers l'Asie.

Vers un cadre à l'échelle de l'espèce

À la fin des années 2010, des milliers d'isolats répartis dans le monde avaient remplacé la poignée de génomes initiale. La structure des populations était mieux résolue, plusieurs histoires de domestication avaient été reconstruites, et l'Asie de l'Est apparaissait comme un réservoir probable de diversité ancestrale. Ces travaux montraient aussi qu'un seul génome ne peut représenter la diversité d'une espèce. Les SNP, les changements de nombre de copies et les introgressions avaient été catalogués, mais une partie de la variation restait hors de portée. Les grandes collections de reséquençage, dont le projet 1002 Yeast Genomes et le premier pangénome fondé sur les gènes, reposaient principalement sur des lectures courtes. Celles-ci détectent précisément les SNP et les petites insertions-délétions, mais résolvent mal les régions répétées ou structurellement complexes (Goodwin et al., 2016).

Les lectures longues ont permis de traverser ces régions et de reconstruire des chromosomes entiers, ouvrant la voie aux assemblages télomère-à-télomère (T2T) et aux pangénomes.

L'ère pangénomique : assemblages télomère-à-télomère et variation structurale

Le séquençage en lectures longues a donné accès aux régions répétées du génome et révélé une forte complexité structurale chez *S. cerevisiae*, notamment des réarrangements chromosomiques et des expansions subtélomériques fréquents. Une part substantielle de cette variation ne pouvait pas être représentée dans une référence linéaire unique. À mesure que les assemblages complets s'accumulaient, la comparaison à S288C a donc été complétée par des représentations à l'échelle de l'espèce, fondées sur plusieurs assemblages ou sur des graphes et couvrant l'ensemble des classes de variation génomique.

Les limites d'un génome de référence unique

Le succès du génome de référence S288C a engendré un paradoxe inévitable. La même référence qui a porté deux décennies de découvertes a aussi, en retour, contraint

l'interprétation de la diversité génomique. Presque toutes les études de génomique des populations reposaient sur l'alignement des lectures sur l'unique référence linéaire (Eizenga et al., 2020; Martiniano et al., 2020), de sorte que la détection était fortement favorisée pour les séquences déjà présentes dans la référence. Cette stratégie a introduit un biais de référence intrinsèque. Les variants nucléotidiques uniques étaient relativement faciles à détecter, mais les grandes insertions, les régions répétées, les translocations ainsi que les réarrangements subtelomériques étaient fréquemment manqués, et les séquences absentes de S288C ne pouvaient pas être représentées directement dans son système de coordonnées. La conséquence en était visible depuis une décennie dans les génomes de souches d'intérêt industriel. Par exemple, la souche de vin EC1118 porte des dizaines de gènes absents de S288C, dont plusieurs contribuent à la performance de fermentation et à l'adaptation à l'environnement du vin (Novo et al., 2009; Marsit & Dequin, 2015). La variation structurale, la variation du contenu génique ainsi que les séquences acquises horizontalement sont précisément les classes de variants les plus fortement associées à l'adaptation, et donc les moins accessibles au travers d'une seule référence linéaire. Pour comprendre pleinement la diversité de l'espèce, des assemblages génomiques complets devaient remplacer les collections de variants cartographiés sur une référence.

Le séquençage en lectures longues change l'échelle de l'assemblage des génomes

Les stratégies de séquençage en lectures longues ont été la solution à ce problème. Contrairement au séquençage en lectures courtes, qui génère des fragments de seulement quelques centaines de paires de bases, les technologies Pacific Biosciences (PacBio) et Oxford Nanopore Technologies (ONT) produisent des lectures pouvant couvrir des milliers, voire des centaines de milliers de bases (Eid et al., 2009; Jain et al., 2016). Ces lectures longues peuvent franchir les régions répétées et permettre la reconstruction directe de séquences à l'échelle du chromosome.

Les premiers jeux de données PacBio ont produit des assemblages beaucoup plus continus que les approches en lectures courtes (Kim et al., 2014). Peu après, ONT a permis le premier assemblage *de novo* quasi complet d'un génome eucaryote à partir de seules lectures nanopore (Goodwin et al., 2015).

Ces premiers assemblages ont révélé des caractéristiques génomiques restées cachées pendant des décennies. Les répétitions en tandem de l'ADN ribosomique, les éléments transposables, les répétitions subtelomériques et les télomères pouvaient désormais être reconstruits directement plutôt qu'inférés indirectement. Des analyses comparatives ont montré que les assemblages en lectures longues restituaient des chromosomes entiers avec une complétude inatteignable par les technologies de lectures courtes (Giordano et al., 2017).

Même les souches classiques de laboratoire continuaient de révéler une complexité inattendue. L'assemblage en lectures longues de la souche très utilisée CEN.PK113-7D a récupéré environ 800 kb de séquence auparavant non résolue et a identifié une

translocation à l'échelle du chromosome qui avait échappé à la détection malgré des années d'étude intensive (Salazar et al., 2017). Il s'est avéré que l'assemblage incomplet n'était pas seulement un problème pour les isolats sauvages : il touchait aussi les souches de laboratoire les mieux caractérisées.

La variation structurale s'impose comme une source majeure de diversité

Appliqué aux populations naturelles, le séquençage en lectures longues a révélé que la variation structurale constitue une composante centrale de la diversité de la levure. Aux SNP qui dominaient les analyses précédentes s'ajoutaient des inversions, des translocations, des duplications et des variants de présence-absence (PAV).

Les premières études à l'échelle des populations en lectures longues ont fourni une démonstration de cette complexité. Les assemblages nanopore de 21 isolats divers ont identifié des dizaines de réarrangements chromosomiques auparavant inconnus et ont généré les premières cartes exhaustives de la diversité des éléments transposables dans les populations naturelles (Istace et al., 2017). Dans le même temps, les assemblages PacBio de *S. cerevisiae* et *S. paradoxus* ont établi que les régions subtélomériques évoluent considérablement plus vite que les noyaux chromosomiques et servent de réservoirs majeurs de variation adaptative (Yue et al., 2017).

La variation structurale n'était plus une rare exception mais bien une composante omniprésente de l'évolution des génomes. De nombreux phénotypes adaptatifs auparavant attribués aux seuls SNP sont désormais compris comme des conséquences de gains, de pertes ou de duplications de gènes, ainsi que de réarrangements chromosomiques.

Des assemblages télomère-à-télomère vers une vision pangénomique

Les assemblages quasi télomère-à-télomère (T2T) à l'échelle des populations ont fourni la première vue d'ensemble de cette diversité dans toute l'espèce. Le *Saccharomyces cerevisiae* Reference Assembly Panel (ScRAP) a produit des assemblages de haute qualité pour des souches représentatives des grandes lignées. Il a montré que l'architecture du génome varie fortement entre isolats, en particulier dans les régions subtélomériques (O'Donnell et al., 2023).

Des assemblages quasi T2T ont ensuite été générés pour 1 086 isolats naturels échantillonnés dans le monde (Loegler et al., 2025). Cette ressource comprend 1 482 assemblages haplotypiques de haute qualité, dont des haplotypes phasés pour des centaines de souches hétérozygotes, et résout plus de 97 % des chromosomes en contigs uniques. Elle a permis d'identifier 6 587 variants structuraux (SV) uniques couvrant plus de 27 Mb de séquence et a révélé de fortes différences d'organisation du génome entre populations. Ces assemblages décrivent la diversité de l'espèce sans la réduire aux écarts par rapport à S288C. Le pangénome rassemble le répertoire complet de l'information génétique observée chez *S. cerevisiae*, avec des éléments conservés, partagés par tous les individus, et des éléments variables, restreints à certaines souches.

Graphes de pangénome contre référence unique

Si les pangénomes fondés sur les gènes englobent le répertoire codant total d'une espèce, ils restent contraints par leur dépendance à de simples listes de gènes. À l'inverse, les graphes de pangénome vont plus loin en combinant de multiples assemblages génomiques en une seule structure. Au lieu de décrire la diversité comme des écarts par rapport à une séquence de référence, les approches fondées sur les graphes représentent simultanément les allèles alternatifs, les insertions, les délétions, les réarrangements et les séquences propres à chaque lignée. Les génomes individuels sont donc représentés comme des chemins distincts à travers le graphe plutôt que comme une liste de différences par rapport à S288C (Eizenga et al., 2020; Loegler et al., 2026).

Le graphe construit à partir de l'ensemble mondial d'assemblages quasi T2T capture presque toute la variation génomique commune de l'espèce et incorpore environ 2,5 Mb de séquence absente de S288C (Loegler et al., 2025) (Fig. 4). Il réduit le biais de référence et fournit une représentation commune des SNP, des variants structuraux et des séquences propres à certaines lignées pour les analyses de génomique des populations et de génotype-phénotype. Après près de trois décennies organisées autour de S288C, la diversité de l'espèce peut ainsi servir elle-même de référence (The Computational Pan-Genomics Consortium, 2018; Eizenga et al., 2020).

Vers une représentation complète de la diversité de l'espèce

Le passage d'un génome de référence unique à un catalogue presque complet de la variation de l'espèce ne résout pas le principal problème : il faut maintenant interpréter cette diversité. Quels variants influencent le phénotype? Quels sont les effets des changements structuraux? Dans quelle mesure le génome permet-il de prédire le comportement cellulaire? Ces questions déplacent l'analyse du recensement de la variation vers ses conséquences fonctionnelles et vers la relation génotype-phénotype.

De la diversité génétique aux relations génotype-phénotype

Les ressources génomiques à l'échelle de l'espèce ont fait de *S. cerevisiae* l'un des eucaryotes les mieux caractérisés et révélé une variation génétique qui ne peut plus être résumée par S288C. Il faut maintenant déterminer comment elle façonne les phénotypes moléculaires, cellulaires et à l'échelle de l'organisme. Peu d'organismes réunissent, comme la levure, une grande diversité naturelle, une génétique expérimentale efficace, des ressources de génomique fonctionnelle exhaustives et de vastes jeux de données de population (Botstein & Fink, 2011). Au cours des deux dernières décennies, la cartographie de liaison, la cartographie de loci de caractères quantitatifs (QTL), les études d'association pangénomiques et le profilage multi-omique ont relié la variation génétique aux phénotypes avec une résolution croissante (Peter & Schacherer, 2016). Le domaine est ainsi passé du catalogue de la variation génomique à l'étude de ses conséquences fonctionnelles.

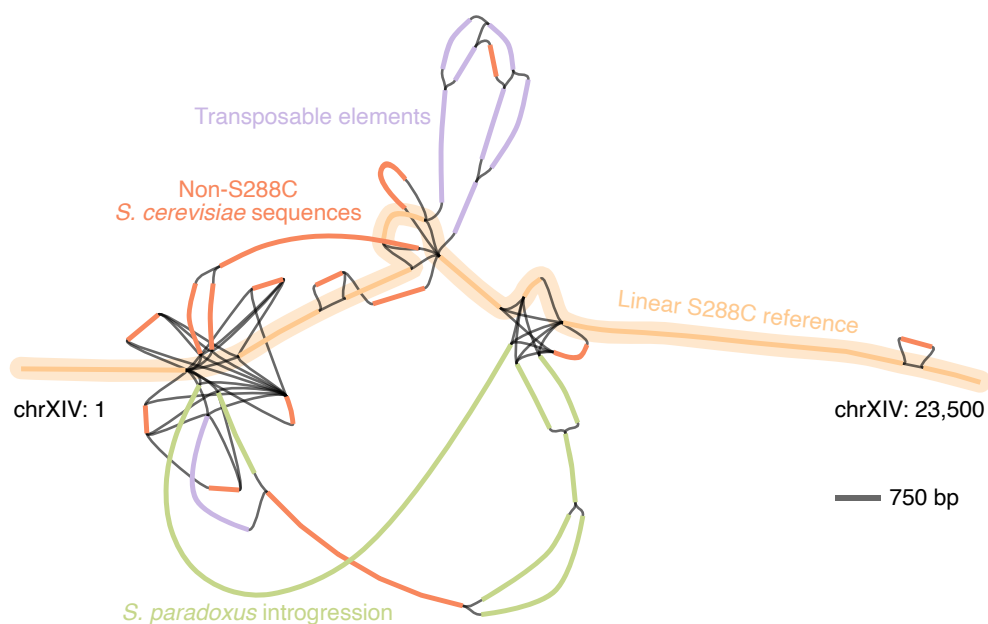


Figure 4 | Un graphe de pangénome résout une variation invisible à une référence linéaire unique.

Représentation d'une région subtélomérique du chromosome XIV (positions 1 à 23 500) construite à partir d'assemblages T2T de *S. cerevisiae* (Loegler et al. 2025); les nœuds alternatifs capturent une séquence divergente ou absente de S288C : séquences de *S. cerevisiae* non-S288C, éléments transposables et une introgression de *S. paradoxus*. Chaque isolat peut être résolu comme un chemin distinct à travers la structure.

Les fondements conceptuels de la génétique quantitative

Relier le génotype au phénotype est un défi classique. La génétique mendélienne classique a expliqué avec succès les caractères contrôlés par un seul gène, mais la plupart des caractéristiques biologiques ne suivent pas de schémas d'hérédité simples. Le taux de croissance, la tolérance au stress, la performance de fermentation, la morphologie ou la longévité varient tous de manière continue et sont influencés simultanément par de nombreux facteurs génétiques et environnementaux. Le cadre conceptuel permettant de comprendre de tels caractères a été établi par le modèle infinitésimal de Fisher, qui proposait que les phénotypes quantitatifs résultent de la contribution combinée de nombreux loci, chacun exerçant un faible effet (Fisher, 1919). Cette intuition a réconcilié l'hérédité mendélienne avec la variation continue et a posé les fondations de la génétique quantitative moderne.

Dans ce cadre, la variation phénotypique peut être décomposée en trois grandes composantes : les effets génétiques, les effets environnementaux et les interactions génotype × environnement (Mackay et al., 2009). En pratique, cela signifie qu'un même génotype peut exprimer des phénotypes différents selon les environnements, tandis que

des fonds génétiques distincts peuvent converger vers un même phénotype dans des conditions environnementales uniformes.

La levure offre de grands avantages expérimentaux pour étudier ces phénomènes. Les individus peuvent être propagés de manière clonale, les conditions environnementales peuvent être précisément contrôlées, et un seul croisement peut produire des millions de descendants (Ehrenreich et al., 2010). Par conséquent, bon nombre des avancées conceptuelles et méthodologiques de la génétique quantitative ont été développées ou affinées chez *S. cerevisiae* (Liti & Louis, 2012 ; Swinnen et al., 2012).

La transition vers les études d'association pangénomiques

Bien que la cartographie de liaison demeure un outil puissant (voir (Liti & Louis, 2012 ; Swinnen et al., 2012) pour revue), sa portée est intrinsèquement limitée par la diversité génétique présente dans un croisement donné. À mesure que les ressources de génomique des populations s'étendaient, une stratégie différente est devenue de plus en plus attrayante : les études d'association pangénomiques (GWAS, *genome-wide association studies*) (Visscher et al., 2017 ; Uffelmann et al., 2021).

Plutôt que de générer une descendance expérimentale, les GWAS exploitent la recombinaison naturelle accumulée à l'échelle des temps évolutifs. L'association entre variants génétiques et variation phénotypique est testée dans de larges collections d'individus, en tenant compte de la structure de population et de l'apparentement. Comme la recombinaison historique a fragmenté les haplotypes ancestraux en segments plus petits, les GWAS peuvent souvent atteindre une résolution de cartographie nettement supérieure à celle des analyses de liaison traditionnelles (Hirschhorn & Daly, 2005).

La faisabilité des GWAS dépend étroitement de la taille de l'échantillon et de la couverture génomique (Connelly & Akey, 2012 ; Peter & Schacherer, 2016). Le passage de centaines à des milliers de génomes, notamment avec le projet 1002 Yeast Genomes et les jeux de données ultérieurs, a fourni la puissance statistique nécessaire aux études d'association à grande échelle (Peter et al., 2018).

La levure occupe une position inhabituelle parmi les organismes modèles : la structure des populations est forte, les histoires de recombinaison sont complexes, et la domestication a créé un fort déséquilibre de liaison au sein de certaines lignées. Ces caractéristiques compliquent les analyses d'association mais permettent aussi d'étudier l'adaptation dans les populations naturelles. Les GWAS ont identifié des loci associés à la croissance dans différents environnements, à la production de métabolites, à la tolérance au stress, aux caractères de fermentation et à la morphologie cellulaire (Peter et al., 2018 ; Sardi et al., 2018 ; De Chiara et al., 2022 ; Kessi-Pérez et al., 2023). Leurs résultats indiquent que la plupart de ces caractères sont hautement polygéniques.

Plutôt que d'être contrôlés par quelques gènes majeurs, les phénotypes quantitatifs émergent généralement des effets cumulés de nombreux variants répartis dans le génome. Cette observation reflète les résultats issus de la génétique humaine et suggère

que la polygénicité pourrait être une caractéristique générale des systèmes biologiques complexes (Boyle et al., 2017 ; Visscher et al., 2017).

Au-delà des SNP : variation structurale et génome accessoire

L'une des leçons les plus importantes de l'ère pangénomique est que les SNP ne capturent qu'une partie de l'architecture génétique sous-tendant la variation phénotypique. Les SV exercent souvent des effets nettement supérieurs à ceux des substitutions nucléotidiques individuelles (Pérez-Ortín et al., 2002 ; Zimmer et al., 2014). Les duplications de gènes peuvent augmenter les niveaux d'expression, les délétions peuvent éliminer des voies entières, et les réarrangements chromosomiques peuvent altérer les interactions régulatrices (Moore et al., 2026). De même, le génome accessoire introduit des gènes entièrement absents chez certains individus et qui, par conséquent, ne peuvent pas être représentés dans les analyses conventionnelles fondées sur les SNP. Les exemples abondent dans les populations de levure. L'amplification du locus *CUP1* modifie la tolérance au cuivre (Fogel et al., 1983). La variation des clusters *ARR* influence la résistance à l'arsenic (Bobrowicz et al., 1997 ; Maciaszczyk et al., 2004). Les gènes d'utilisation du maltose contribuent directement à la performance de brasserie (Gallone et al., 2016 ; Gonçalves et al., 2016). Des loci transférés horizontalement acquis par les souches de vin améliorent l'acquisition des oligopeptides pendant la fermentation (Marsit & Dequin, 2015 ; Becerra-Rodríguez et al., 2020). Les régions introgressées héritées d'espèces apparentées de *Saccharomyces* codent fréquemment des fonctions associées à l'adaptation environnementale (D'Angiolo et al., 2020 ; Avelar-Rivas et al., 2026).

À mesure que les assemblages télomère-à-télomère se multiplient, ils élargissent le champ des études génotype-phénotype. Les SV, la variation du contenu génique et l'architecture des chromosomes peuvent désormais être analysés avec les SNP conventionnels. Cette perspective sera probablement essentielle pour comprendre les caractères complexes : de nombreux phénotypes adaptatifs pourraient dépendre davantage du gain, de la perte ou du nombre de copies de gènes que des seuls SNP.

Multi-omique et phénotypes intermédiaires

L'intégration d'autres couches moléculaires complète les données génomiques. La séquence d'ADN n'agit pas directement sur le phénotype : ses effets passent par de nombreux processus intermédiaires qu'il faut mesurer pour reconstruire la relation génotype-phénotype (Civelek & Lusi, 2014).

La variation génétique agit d'abord sur la transcription, qui influe ensuite sur l'abondance des protéines, l'activité métabolique, l'organisation de la chromatine et la physiologie cellulaire avant de façonner les phénotypes de l'organisme. Chaque niveau ajoute de la complexité, mais permet aussi une compréhension mécanistique. La transcriptomique et la protéomique à l'échelle des populations mesurent plusieurs de ces intermédiaires. Elles ont permis d'identifier des QTL d'expression (eQTL) et des

QTL protéiques (pQTL), qui relie la variation génétique à des caractères moléculaires (Caudal et al., 2024 ; Muenzner et al., 2024 ; Teyssonnière et al., 2024).

Les phénotypes intermédiaires sont souvent plus directement liés à la variation génétique que les caractères à l'échelle de l'organisme. L'abondance des transcrits peut être plus facile à prédire à partir du génotype que le taux de croissance ou l'efficacité de fermentation, qui intègrent les effets de milliers de processus moléculaires (Albert & Kruglyak, 2015). La combinaison de plusieurs couches omiques permet de retracer de mieux en mieux les voies causales qui relient la séquence d'ADN au phénotype. Cette approche systémique constitue l'une des voies les plus prometteuses vers une biologie prédictive.

Conclusion et perspectives

L'objectif ultime de la génétique quantitative est la prédiction : étant donné une séquence génomique, peut-on inférer le phénotype d'un organisme? Pour la plupart des eucaryotes, ce but reste lointain, car les génomes sont grands, les effets environnementaux difficiles à contrôler et la validation expérimentale souvent limitée. La levure occupe cependant une position unique parmi les organismes modèles.

L'espèce possède l'un des catalogues de variation génétique les plus complets disponibles chez un eucaryote. Des milliers de génomes ont été séquencés. Les assemblages télomère-à-télomère ont catalogué les variants structuraux et les gènes accessoires, les annotations fonctionnelles couvrent la grande majorité des gènes, et de vastes réseaux d'interactions génétiques relient les voies et les processus cellulaires. Les phénotypes moléculaires à l'échelle des populations sont aussi de plus en plus nombreux. Ces ressources permettent de dépasser l'identification d'associations pour construire des modèles prédictifs (Khailwal et al., 2025). Des méthodes d'apprentissage automatique intègrent déjà les données génomiques, transcriptomiques, protéomiques et phénotypiques. La prédiction reste imparfaite, mais des données plus complètes et de meilleurs modèles réduisent progressivement l'écart entre génotype et phénotype.

Le problème génotype-phénotype constitue ainsi l'aboutissement des avancées décrites dans cette revue. Trente ans après la publication du premier génome de *Saccharomyces cerevisiae* par André Goffeau et son consortium, la référence unique a été complétée par un catalogue presque exhaustif de variants couvrant plus d'un millier d'isolats phénotypés dans des dizaines d'environnements. Ce jalon historique, les ressources de génomique fonctionnelle, la découverte de la diversité des populations et les pangénomes convergent vers une même question : comment la variation génomique engendre-t-elle la diversité biologique? La faisabilité de modèles génotype-phénotype prédictifs complets demeure incertaine. Néanmoins, aucun autre système eucaryote n'est actuellement mieux placé que *S. cerevisiae* pour aborder cette question.

References

- Aa E, Townsend JP, Adams RI, Nielsen KM, Taylor JW (2006) Population structure and gene evolution in *Saccharomyces cerevisiae*. *FEMS Yeast Res* **6**: 702–715.
- Abou Saada OA, Tsouris A, Large C, Friedrich A, Dunham MJ, Schacherer J (2022) Phased polyploid genomes provide deeper insight into the multiple origins of domesticated *Saccharomyces cerevisiae* beer yeasts. *Curr Biol* **32**: 1350–1361.e3.
- Akao T, Yashiro I, Hosoyama A, Kitagaki H, Horikawa H, Watanabe D, et al. (2011) Whole-Genome Sequencing of Sake Yeast *Saccharomyces cerevisiae* Kyokai no. 7. *DNA Res* **18**: 423–434.
- Albert FW, Kruglyak L (2015) The role of regulatory variation in complex traits and disease. *Nat Rev Genet* **16**: 197–212.
- Almeida P, Barbosa R, Zalar P, Imanishi Y, Et. A (2015) A population genomics insight into the Mediterranean origins of wine yeast domestication. *Mol Ecol*.
- Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, et al. (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. *Cell* **166**: 481–491.
- Annaluru N, Muller H, Mitchell LA, Ramalingam S, Stracquadanio G, Richardson SM, et al. (2014) Total Synthesis of a Functional Designer Eukaryotic Chromosome. *Science* **344**: 55–58.
- Argueso JL, Carazzolle MF, Mieczkowski PA, Duarte FM, Netto OVC, Missawa SK, et al. (2009) Genome structure of a *Saccharomyces cerevisiae* strain widely used in bioethanol production. *Genome Res* **19**: 2258–2270.
- Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. (2015) A global reference for human genetic variation. *Nature* **526**: 68–74.
- Avelar-Rivas JA, Sedeño I, García-Ortega LF, Urban Aragon JA, López-Gallegos C, Aguirre-Dugua X, et al. (2026) Recurrent introgression and geographical stratification shape *Saccharomyces cerevisiae* in the Neotropics. *Nat Commun* **17**: 3024.
- Baker Brachmann C, Davies A, Cost GJ, Caputo E, Li J, Hieter P, et al. (1998) Designer Deletion Strains Derived from *Saccharomyces Cerevisiae* S288C: A Useful Set of Strains and Plasmids for PCR-mediated Gene Disruption and Other Applications. *Yeast* **14**: 115–132.
- Barbosa R, Almeida P, Safar SVB, Santos RO, Morais PB, Nielly-Thibault L, et al. (2016) Evidence of Natural Hybridization in Brazilian Wild Lineages of *Saccharomyces cerevisiae*. *Genome Biol Evol* **8**: 317–329.
- Basile A, De Pascale F, Bianca F, Rossi A, Frizzarin M, De Bernardini N, et al. (2021) Large-scale sequencing and comparative analysis of oenological *Saccharomyces cerevisiae* strains supported by nanopore refinement of key genomes. *Food Microbiol* **97**: 103753.
- Becerra-Rodríguez C, Marsit S, Galeote V (2020) Diversity of Oligopeptide Transport in Yeast and Its Impact on Adaptation to Winemaking Conditions. *Front Genet* **11**.
- Becerra-Rodríguez C, Thiele P, Brach G, Dutta A, Garin M, Tan G, et al. (2026) Domestication drives repeated evolution of sexual–asexual life cycle trade-offs in yeast. *Proc Natl Acad Sci* **123**: e2526682123.
- Ben-Aroya S, Coombes C, Kwok T, O'Donnell KA, Boeke JD, Hieter P (2008) Toward a Comprehensive Temperature-Sensitive Mutant Repository of the Essential Genes of

Saccharomyces cerevisiae. *Mol Cell* **30**: 248–258.

- Bergström A, Simpson JT, Salinas F, Barré B, Parts L, Zia A, et al. (2014) A High-Definition View of Functional Genetic Variation from Natural Yeast Genomes. *Mol Biol Evol* **31**: 872–888.
- Bircham PW, Cortona AD, Garshol LM, Tawfeeq MT, Herrera-Malaver B, Mannaerts S, et al. (2026) Distinctive domestication of farmhouse beer yeasts preserved pre-industrial genetic and phenotypic diversity. *Curr Biol* **36**: 2028–2042.e5.
- Bobrowicz P, Wysocki R, Owsianik G, Goffeau A, Ułaszewski S (1997) Isolation of three contiguous genes, ACR1, ACR2 and ACR3, involved in resistance to arsenic compounds in the yeast *Saccharomyces cerevisiae*. *Yeast Chichester Engl* **13**: 819–828.
- 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**: 415–419.
- Borneman AR, Desany BA, Riches D, Affourtit JP, Forgan AH, Pretorius IS, et al. (2011) Whole-Genome Comparison Reveals Novel Genetic Elements That Characterize the Genome of Industrial Strains of *Saccharomyces cerevisiae*. *PLOS Genet* **7**: e1001287.
- Borneman AR, Forgan AH, Kolouchova R, Fraser JA, Schmidt SA (2016) Whole Genome Comparison Reveals High Levels of Inbreeding and Strain Redundancy Across the Spectrum of Commercial Wine Strains of *Saccharomyces cerevisiae*. *G3 GenesGenomesGenetics* **6**: 957–971.
- Borneman AR, Forgan AH, Pretorius IS, Chambers PJ (2008) Comparative genome analysis of a *Saccharomyces cerevisiae* wine strain. *FEMS Yeast Res* **8**: 1185–1195.
- Botstein D, Fink GR (2011) Yeast: An Experimental Organism for 21st Century Biology. *Genetics* **189**: 695–704.
- Boyle EA, Li YI, Pritchard JK (2017) An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**: 1177–1186.
- Breslow DK, Cameron DM, Collins SR, Schuldiner M, Stewart-Ornstein J, Newman HW, et al. (2008) A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. *Nat Methods* **5**: 711–718.
- Brown CA, Murray AW, Verstrepen KJ (2010) Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Curr Biol CB* **20**: 895–903.
- Carreto L, Eiriz MF, Gomes AC, Pereira PM, Schuller D, Santos MA (2008) Comparative genomics of wild type yeast strains unveils important genome diversity. *BMC Genomics* **9**: 524.
- Caudal É, Loegler V, Dutreux F, Vakirlis N, Teyssonnière É, Caradec C, et al. (2024) Pan-transcriptome reveals a large accessory genome contribution to gene expression variation in yeast. *Nat Genet* **56**: 1278–1287.
- Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, et al. (1998) SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res* **26**: 73–79.
- Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al. (2012) *Saccharomyces* Genome Database: The genomics resource of budding yeast. *Nucleic Acids Res* **40**: D700–D705.
- Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, et al. (1998) The Transcriptional Program of Sporulation in Budding Yeast. *Science* **282**: 699–705.
- Civelek M, Lusi AJ (2014) Systems genetics approaches to understand complex traits. *Nat Rev Genet* **15**: 34–48.
- Conant GC, Wolfe KH (2007) Increased Glycolytic Flux as an Outcome of Whole-genome Duplication in Yeast. *Mol Syst Biol* **3**: MSB4100170.
- Conant GC, Wolfe KH (2008) Turning a hobby into a job: How duplicated genes find new functions. *Nat Rev Genet* **9**: 938–950.

- Connelly CF, Akey JM (2012) On the Prospects of Whole-Genome Association Mapping in *Saccharomyces cerevisiae*. *Genetics* **191**: 1345–1353.
- Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, et al. (2010) The Genetic Landscape of a Cell. *Science* **327**: 425–431.
- Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, et al. (2016) A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**: aaf1420.
- D’Angiolo M, De Chiara M, Yue J-X, Irizar A, Stenberg S, Persson K, et al. (2020) A yeast living ancestor reveals the origin of genomic introgressions. *Nature* **587**: 420–425.
- Daran-Lapujade P, Daran J-M, Kötter P, Petit T, Piper MDW, Pronk JT (2003) Comparative genotyping of the *Saccharomyces cerevisiae* laboratory strains S288C and CEN.PK113-7D using oligonucleotide microarrays. *FEMS Yeast Res* **4**: 259–269.
- De Chiara M, Barré BP, Persson K, Irizar A, Vischioni C, Khaiwal S, et al. (2022) Domestication reprogrammed the budding yeast life cycle. *Nat Ecol Evol* **6**: 448–460.
- Dean EJ, Davis JC, Davis RW, Petrov DA (2008) Pervasive and Persistent Redundancy among Duplicated Genes in Yeast. *PLOS Genet* **4**: e1000113.
- Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing Chromosome Conformation. *Science* **295**: 1306–1311.
- DeLuna A, Vetsigian K, Shores N, Hegreness M, Colón-González M, Chao S, et al. (2008) Exposing the fitness contribution of duplicated genes. *Nat Genet* **40**: 676–681.
- DeRisi JL, Iyer VR, Brown PO (1997) Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science* **278**: 680–686.
- Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, et al. (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304**: 304–307.
- Dondrup M, Martinussen AO, Haugland LK, Brandenburg J, Inanli O, Elameen A, et al. (2026) Genomic evidence of early-diverging domesticated lineages in Norwegian farmhouse yeast. : 2026.03.16.711853.
- Doniger SW, Kim HS, Swain D, Corcuera D, Williams M, Yang S-P, et al. (2008) A Catalog of Neutral and Deleterious Polymorphism in Yeast. *PLOS Genet* **4**: e1000183.
- Dowell RD, Ryan O, Jansen A, Cheung D, Agarwala S, Danford T, et al. (2010) Genotype to Phenotype: A Complex Problem. *Science* **328**: 469–469.
- Duan S-F, Han P-J, Wang Q-M, Liu W-Q, Shi J-Y, Li K, et al. (2018) The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. *Nat Commun* **9**: 2690.
- Dujon B (1992) Altogether now – sequencing the yeast genome. *Curr Biol* **2**: 279–281.
- Dujon B (1996) The yeast genome project: What did we learn? *Trends Genet* **12**: 263–270.
- Dujon B, Cooper TG (2026) Yeasty memories, yeast genomes and beyond—from controversy to collaboration. *FEMS Yeast Res* **26**: foaf073.
- Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, et al. (2004) Genome evolution in yeasts. *Nature* **430**: 35–44.
- Dunn B, Levine RP, Sherlock G (2005) Microarray karyotyping of commercial wine yeast strains reveals shared, as well as unique, genomic signatures. *BMC Genomics* **6**: 53.
- Dunn B, Richter C, Kvitek DJ, Pugh T, Sherlock G (2012) Analysis of the *Saccharomyces cerevisiae* pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. *Genome Res* **22**: 908–924.

- Dymond JS, Richardson SM, Coombes CE, Babatz T, Muller H, Annaluru N, et al. (2011) Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. *Nature* **477**: 471–476.
- Ehrenreich IM, Torabi N, Jia Y, Kent J, Martis S, Shapiro JA, et al. (2010) Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature* **464**: 1039–1042.
- Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133–138.
- Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, et al. (2020) Pangenome Graphs. *Annu Rev Genomics Hum Genet* **21**: 139–162.
- Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, et al. (2013) The Reference Genome Sequence of *Saccharomyces cerevisiae*: Then and Now. *G3 GenesGenomesGenetics* **4**: 389–398.
- Engel SR, Wong ED, Nash RS, Aleksander S, Alexander M, Douglass E, et al. (2022) New data and collaborations at the *Saccharomyces* Genome Database: Updated reference genome, alleles, and the Alliance of Genome Resources. *Genetics* **220**: iyab224.
- Fay JC, Benavides JA (2005) Evidence for Domesticated and Wild Populations of *Saccharomyces cerevisiae*. *PLOS Genet* **1**: e5.
- Fay JC, Liu P, Ong GT, Dunham MJ, Cromie GA, Jeffery EW, et al. (2019) A polyploid admixed origin of beer yeasts derived from European and Asian wine populations. *PLOS Biol* **17**: e3000147.
- Fisher RA (1919) The causes of human variability. *Eugen Rev* **10**: 213–220.
- Fogel S, Welch JW, Cathala G, Karin M (1983) Gene amplification in yeast: CUP1 copy number regulates copper resistance. *Curr Genet* **7**: 347–355.
- Foss EJ, Radulovic D, Shaffer SA, Ruderfer DM, Et. A (2007) Genetic basis of proteome variation in yeast. *Nat Genet*.
- Foury F, Roganti T, Lecrenier N, Purnelle B (1998) The complete sequence of the mitochondrial genome of *Saccharomyces cerevisiae*. *FEBS Lett*.
- Gaisne M, Bécam A-M, Verdière J, Herbert CJ (1999) A ‘natural’ mutation in *Saccharomyces cerevisiae* strains derived from S288c affects the complex regulatory gene HAP1 (CYP1): *Curr Genet* **36**: 195–200.
- Gallone B, Steensels J, Prah T, Soriaga L, Saels V, Herrera-Malaver B, et al. (2016) Domestication and Divergence of *Saccharomyces cerevisiae* Beer Yeasts. *Cell* **166**: 1397–1410.e16.
- Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, et al. (2000) Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Mol Biol Cell* **11**: 4241–4257.
- Gavin A-C, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**: 631–636.
- Ghaemmaghami S, Huh W-K, Bower K, Howson RW, Belle A, Dephoure N, et al. (2003) Global analysis of protein expression in yeast. *Nature* **425**: 737–741.
- Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, et al. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**: 387–391.
- Giaever G, Flaherty P, Kumm J, Proctor M, Nislow C, Jaramillo DF, et al. (2004) Chemogenomic profiling: Identifying the functional interactions of small molecules in yeast. *Proc Natl Acad Sci* **101**: 793–798.
- Giordano F, Aigrain L, Quail MA, Coupland P, Bonfield JK, Davies RM, et al. (2017) De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Sci Rep* **7**: 3935.

- Goddard MR, Greig D (2015) *Saccharomyces cerevisiae*: A nomadic yeast with no niche? *FEMS Yeast Res* **15**: fov009.
- Goffeau A, Barrell B. G., Bussey H., Davis R. W., Dujon B., Feldmann H., et al. (1996) Life with 6000 Genes. *Science* **274**: 546–567.
- Goffeau A, Vassarotti A (1991) The European project for sequencing the yeast genome. *Res Microbiol* **142**: 901–903.
- Gonçalves M, Pontes A, Almeida P, Barbosa R, Serra M, Libkind D, et al. (2016) Distinct Domestication Trajectories in Top-Fermenting Beer Yeasts and Wine Yeasts. *Curr Biol* **26**: 2750–2761.
- Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR (2015) Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res* **25**: 1750–1756.
- Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: Ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**: 333–351.
- Gresham D, Ruderfer DM, Pratt SC, Schacherer J, Dunham MJ, Botstein D, et al. (2006) Genome-Wide Detection of Polymorphisms at Nucleotide Resolution with a Single DNA Microarray. *Science* **311**: 1932–1936.
- Guan Y, Dunham MJ, Troyanskaya OG (2007) Functional Analysis of Gene Duplications in *Saccharomyces cerevisiae*. *Genetics* **175**: 933–943.
- Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between Protein and mRNA Abundance in Yeast. *Mol Cell Biol* **19**: 1720–1730.
- Han D-Y, Han P-J, Rumbold K, Koricha AD, Duan S-F, Song L, et al. (2021) Adaptive Gene Content and Allele Distribution Variations in the Wild and Domesticated Populations of *Saccharomyces cerevisiae*. *Front Microbiol* **12**: 631250.
- Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**: 99–104.
- Higgins P, Grace CA, Lee SA, Goddard MR (2021) Whole-genome sequencing from the New Zealand *Saccharomyces cerevisiae* population reveals the genomic impacts of novel microbial range expansion. *G3 GenesGenomesGenetics* **11**: jkaa027.
- Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, et al. (2008) The Chemical Genomic Portrait of Yeast: Uncovering a Phenotype for All Genes. *Science* **320**: 362–365.
- Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**: 95–108.
- Huang W, Massouras A, Inoue Y, Peiffer J, Ràmia M, Tarone AM, et al. (2014) Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome Res* **24**: 1193–1208.
- Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, et al. (2000) Functional Discovery via a Compendium of Expression Profiles. *Cell* **102**: 109–126.
- Huh W-K, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, et al. (2003) Global analysis of protein localization in budding yeast. *Nature* **425**: 686–691.
- Istace B, Friedrich A, Agata L d', Faye S, Payen E, Beluche O, et al. (2017) De novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *GigaScience* **6**: giw018.
- Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci* **98**: 4569–4574.

- Jacquier A, Legrain P, Dujon B (1992) Sequence of a 10.7 kb segment of yeast chromosome XI identifies the APN1 and the BAF1 loci and reveals one tRNA gene and several new open reading frames including homologs to RAD2 and kinases. *Yeast* **8**: 121–132.
- Jain M, Olsen HE, Paten B, Akeson M (2016) The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community. *Genome Biol* **17**: 239.
- Jin J, Jia B, Yuan Y-J (2020) Yeast chromosomal engineering to improve industrially-relevant phenotypes. *Curr Opin Biotechnol* **66**: 165–170.
- Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**: 617–624.
- Kessi-Pérez EI, Acuña E, Bastías C, Fundora L, Villalobos-Cid M, Romero A, et al. (2023) Single nucleotide polymorphisms associated with wine fermentation and adaptation to nitrogen limitation in wild and domesticated yeast strains. *Biol Res* **56**: 43.
- Khaiwal S, De Chiara M, Barré BP, Barrio-Hernandez I, Stenberg S, Beltrao P, et al. (2025) Predicting natural variation in the yeast phenotypic landscape with machine learning. *Mol Syst Biol* **21**: 1466–1489.
- Kim KE, Peluso P, Babayan P, Yeaton PJ, Et. A (2014) Long-read, whole-genome shotgun sequence data for five model organisms. *Sci Data*.
- Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, et al. (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**: 637–643.
- Kvitek DJ, Will JL, Gasch AP (2008) Variations in Stress Sensitivity and Genomic Expression in Diverse *S. cerevisiae* Isolates. *PLOS Genet* **4**: e1000223.
- Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, et al. (2002) Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*. *Science* **298**: 799–804.
- Lee TJ, Liu Y-C, Liu W-A, Lin Y-F, Lee H-H, Ke H-M, et al. (2022) Extensive sampling of *Saccharomyces cerevisiae* in Taiwan reveals ecology and evolution of predomesticated lineages. *Genome Res* **32**: 864–877.
- Légras J-L, Galeote V, Bigey F, Camarasa C, Marsit S, Nidelet T, et al. (2018) Adaptation of *S. cerevisiae* to Fermented Food Environments Reveals Remarkable Genome Plasticity and the Footprints of Domestication. *Mol Biol Evol* **35**: 1712–1727.
- Légras J-L, Merdinoglu D, Cornuet J-M, Karst F (2007) Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. *Mol Ecol*.
- Li Z, Vizeacoumar FJ, Bahr S, Li J, Warringer J, Vizeacoumar FS, et al. (2011) Systematic exploration of essential yeast gene function with temperature-sensitive mutants. *Nat Biotechnol* **29**: 361–367.
- Liti G (2015) The fascinating and secret wild life of the budding yeast *S. cerevisiae*. *eLife* **4**: e05835.
- Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, et al. (2009) Population genomics of domestic and wild yeasts. *Nature* **458**: 337–341.
- Liti G, Louis EJ (2012) Advances in Quantitative Trait Analysis in Yeast. *PLOS Genet* **8**: e1002912.
- Liu H, Styles CA, Fink GR (1996) *Saccharomyces Cerevisiae* S288C Has a Mutation in *FL08* , a Gene Required for Filamentous Growth. *Genetics* **144**: 967–978.
- Loegler V, Friedrich A, Schacherer J (2024) Overview of the *Saccharomyces cerevisiae* population structure through the lens of 3,034 genomes. *G3 GenesGenomesGenetics* **14**: jkae245.
- Loegler V, Friedrich A, Schacherer J (2026) Dynamics of genome evolution in the era of pangenome analysis. *Cell Genomics* **6**.
- Loegler V, Thiele P, Teyssonnière E, Tsouris A, Brach G, Cruaud C, et al. (2025) From genotype to phenotype with 1,086 near telomere-to-telomere yeast genomes. *Nature*: 1–10.

- Louis EJ (2016) Historical Evolution of Laboratory Strains of *Saccharomyces cerevisiae*. *Cold Spring Harb Protoc* **2016**: pdb.top077750.
- Maciaszczyk E, Wysocki R, Golik P, Lazowska J, Ulaszewski S (2004) Arsenical resistance genes in *Saccharomyces douglasii* and other yeast species undergo rapid evolution involving genomic rearrangements and duplications. *FEMS Yeast Res* **4**: 821–832.
- Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, et al. (2012) The *Drosophila melanogaster* Genetic Reference Panel. *Nature* **482**: 173–178.
- Mackay TFC, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: Challenges and prospects. *Nat Rev Genet* **10**: 565–577.
- Marr RA, Moore J, Formby S, Martiniuk JT, Hamilton J, Ralli S, et al. (2023) Whole genome sequencing of Canadian *Saccharomyces cerevisiae* strains isolated from spontaneous wine fermentations reveals a new Pacific West Coast Wine clade. *G3 GenesGenomesGenetics* **13**: jkad130.
- Marsit S, Dequin S (2015) Diversity and adaptive evolution of *Saccharomyces* wine yeast: A review. *FEMS Yeast Res* **15**: fov067.
- Marsit S, Leducq J-B, Durand É, Marchant A, Filteau M, Landry CR (2017) Evolutionary biology through the lens of budding yeast comparative genomics. *Nat Rev Genet* **18**: 581–598.
- Martiniano R, Garrison E, Jones ER, Manica A, Durbin R (2020) Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome Biol* **21**: 250.
- Mewes HW, Albermann K, Bähr M, Frishman D, Gleissner A, Hani J, et al. (1997) Overview of the yeast genome. *Nature* **387**: 7–8.
- Mnaimneh S, Davierwala AP, Haynes J, Moffat J, Peng W-T, Zhang W, et al. (2004) Exploration of Essential Gene Functions via Titratable Promoter Alleles. *Cell* **118**: 31–44.
- Moore J, Marr RA, Montpetit R, Montpetit B, Measday V (2026) From Oak to Wine: Evolution, Admixture, and Adaptation of California and British Columbia *Saccharomyces Cerevisiae* Wine Strains. *Genome Biol Evol* **18**: evag149.
- Mortimer RK, Johnston JR (1986) GENEALOGY OF PRINCIPAL STRAINS OF THE YEAST GENETIC STOCK CENTER. *Genetics* **113**: 35–43.
- Muenzner J, Trébulle P, Agostini F, Zauber H, Messner CB, Steger M, et al. (2024) Natural proteome diversity links aneuploidy tolerance to protein turnover. *Nature* **630**: 149–157.
- Muller L a. H, Lucas JE, Georgianna DR, McCUSKER JH (2011) Genome-wide association analysis of clinical vs. nonclinical origin provides insights into *Saccharomyces cerevisiae* pathogenesis. *Mol Ecol*.
- Novo M, Bigey F, Beyne E, Galeote V, Gavory F, Mallet S, et al. (2009) Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* EC1118. *Proc Natl Acad Sci* **106**: 16333–16338.
- O'Donnell S, Yue J-X, Saada OA, Agier N, Caradec C, Cokelaer T, et al. (2023) Telomere-to-telomere assemblies of 142 strains characterize the genome structural landscape in *Saccharomyces cerevisiae*. *Nat Genet* **55**: 1390–1399.
- Oliver SG, Aart QJM van der, Agostoni-Carbone ML, Aigle M, Alberghina L, Alexandraki D, et al. (1992) The complete DNA sequence of yeast chromosome III. *Nature* **357**: 38–46.
- Opulente DA, LaBella AL, Harrison M-C, Wolters JF, Liu C, Li Y, et al. (2024) Genomic factors shape carbon and nitrogen metabolic niche breadth across *Saccharomycotina* yeasts. *Science* **384**: eadj4503.
- Peña JJ, Scopel EFC, Ward AK, Bensasson D (2025) Footprints of Human Migration in the Population Structure of Wild Baker's Yeast. *Mol Ecol* **34**: e17669.

- Pérez-Ortín JE, Querol A, Puig S, Barrio E (2002) Molecular Characterization of a Chromosomal Rearrangement Involved in the Adaptive Evolution of Yeast Strains. *Genome Res* **12**: 1533–1539.
- Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, et al. (2018) Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **556**: 339–344.
- Peter J, Schacherer J (2016) Population genomics of yeasts: Towards a comprehensive view across a broad evolutionary scale. *Yeast*.
- Piotrowski JS, Li SC, Deshpande R, Simpkins SW, Nelson J, Yashiroda Y, et al. (2017) Functional Annotation of Chemical Libraries across Diverse Biological Processes. *Nat Chem Biol* **13**: 982–993.
- Piškur J (2001) Origin of the duplicated regions in the yeast genomes. *Trends Genet* **17**: 302–303.
- Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, et al. (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122**: 517–527.
- Preiss R, Tyrawa C, Krogerus K, Garshol LM, Merwe G van der (2018) Traditional Norwegian Kveik Are a Genetically Distinct Group of Domesticated *Saccharomyces cerevisiae* Brewing Yeasts. *Front Microbiol* **9**.
- Pretorius IS (2000) Tailoring wine yeast for the new millennium: Novel approaches to the ancient art of winemaking. *Yeast* **16**: 675–729.
- Richardson SM, Mitchell LA, Stracquadanio G, Yang K, Dymond JS, DiCarlo JE, et al. (2017) Design of a synthetic yeast genome. *Science* **355**: 1040–1044.
- Salazar AN, Vries AR Gorter de, Broek M van den, Wijsman M, Torre Cortés P de la, Brickwedde A, et al. (2017) Nanopore sequencing enables near-complete de novo assembly of *Saccharomyces cerevisiae* reference strain CEN.PK113-7D. *FEMS Yeast Res* **17**: fox074.
- Sampaio JP, Pontes A (2025) Yeast domestication. *Curr Biol* **35**: R575–R586.
- Sardi M, Paithane V, Place M, Robinson DE, Hose J, Wohlbach DJ, et al. (2018) Genome-wide association across *Saccharomyces cerevisiae* strains reveals substantial variation in underlying gene requirements for toxin tolerance. *PLOS Genet* **14**: e1007217.
- Schacherer J (2026) Microbial domestication: Farmhouse brewing preserves a hidden reservoir of yeast diversity. *Curr Biol* **36**: R478–R480.
- Schacherer J, Ruderfer DM, Gresham D, Dolinski K, Botstein D, Kruglyak L (2007) Genome-Wide Analysis of Nucleotide-Level Variation in Commonly Used *Saccharomyces cerevisiae* Strains. *PLoS One* **2**: e322.
- Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L (2009) Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* **458**: 342–345.
- Seoighe C, Wolfe KH (1999) Updated map of duplicated regions in the yeast genome. *Gene* **238**: 253–261.
- Shen MJ, Wu Y, Yang K, Li Y, Xu H, Zhang H, et al. (2018a) Heterozygous diploid and interspecies SCRaMbLEing. *Nat Commun* **9**: 1934.
- Shen X-X, Oplente DA, Kominek J, Zhou X, Steenwyk JL, Buh KV, et al. (2018b) Tempo and mode of genome evolution in the budding yeast subphylum. *Cell* **175**: 1533–1545.e20.
- Sniegowski PD, Dombrowski PG, Fingerman E (2002) *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. *FEMS Yeast Res* **1**: 299–306.
- Souciet J-L, Aigle M, Artiguenave F, Blandin G, Bolotin-Fukuhara M, Bon E, et al. (2000) Genomic Exploration of the Hemiascomycetous Yeasts: 1. A set of yeast species for molecular evolution studies 1. *FEBS Lett* **487**: 3–12.

- Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. (1998) Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Mol Biol Cell* **9**: 3273–3297.
- Steensels J, Gallone B, Verstrepen KJ (2021) Interspecific hybridization as a driver of fungal evolution and adaptation. *Nat Rev Microbiol* **19**: 485–500.
- Steenwyk J, Rokas A (2017) Extensive Copy Number Variation in Fermentation-Related Genes Among *Saccharomyces cerevisiae* Wine Strains. *G3 GenesGenomesGenetics* **7**: 1475–1485.
- Strope PK, Skelly DA, Kozmin SG, Mahadevan G, Stone EA, Magwene PM, et al. (2015) 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.
- Swinnen S, Thevelein JM, Nevoigt E (2012) Genetic mapping of quantitative phenotypic traits in *Saccharomyces cerevisiae*. *FEMS Yeast Res* **12**: 215–227.
- Teyssonnière EM, Trébulle P, Muenzner J, Loegler V, Ludwig D, Amari F, et al. (2024) Species-wide quantitative transcriptomes and proteomes reveal distinct genetic control of gene expression variation in yeast. *Proc Natl Acad Sci* **121**: e2319211121.
- The Computational Pan-Genomics Consortium (2018) Computational pan-genomics: Status, promises and challenges. *Brief Bioinform* **19**: 118–135.
- The Génolevures Consortium, Souciet J-L, Dujon B, Gaillardin C, Johnston M, Baret PV, et al. (2009) Comparative genomics of protoploid *Saccharomycetaceae*. *Genome Res*.
- Thierry A, Fairhead C, Dujon B (1990) The complete sequence of the 8.2 kb segment left of MAT on chromosome III reveals five ORFs, including a gene for a yeast ribokinase. *Yeast* **6**: 521–534.
- Tong AHY, Evangelista M, Parsons AB, Xu H, Bader GD, Pagé N, et al. (2001) Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants. *Science* **294**: 2364–2368.
- Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, et al. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**: 623–627.
- Uffelmann E, Huang QQ, Munung NS, Vries J de, Okada Y, Martin AR, et al. (2021) Genome-wide association studies. *Nat Rev Methods Primer* **1**: 1–21.
- Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**: 5–22.
- Wang Q-M, Liu W-Q, Liti G, Wang S-A, Bai F-Y (2012) Surprisingly diverged populations of *Saccharomyces cerevisiae* in natural environments remote from human activity. *Mol Ecol* **21**: 5404–5417.
- Ward CM, Onetto CA, Heuvel SVD, Cuijvers KM, Hale LJ, Borneman AR (2024) Recombination, admixture and genome instability shape the genomic landscape of *Saccharomyces cerevisiae* derived from spontaneous grape ferments. *PLOS Genet* **20**: e1011223.
- Wei W, McCusker JH, Hyman RW, Jones T, Ning Y, Cao Z, et al. (2007) Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. *Proc Natl Acad Sci* **104**: 12825–12830.
- Winzeler EA, Castillo-Davis CI, Oshiro G, Liang D, Richards DR, Zhou Y, et al. (2003) Genetic Diversity in Yeast Assessed With Whole-Genome Oligonucleotide Arrays. *Genetics* **163**: 79–89.
- Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, McCullough MJ, et al. (1998) Direct Allelic Variation Scanning of the Yeast Genome. *Science* **281**: 1194–1197.
- Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, et al. (1999) Functional Characterization of the *S. cerevisiae* Genome by Gene Deletion and Parallel Analysis. *Science* **285**: 901–906.

- Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**: 708–713.
- Yuan G-C, Liu Y-J, Dion MF, Slack MD, Wu LF, Altschuler SJ, et al. (2005) Genome-Scale Identification of Nucleosome Positions in *S. cerevisiae*. *Science* **309**: 626–630.
- Yue J-X, Li J, Aigrain L, Hallin J, Persson K, Oliver K, et al. (2017) Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat Genet* **49**: 913–924.
- Zhao Y, Coelho C, Hughes AL, Lazar-Stefanita L, Yang S, Brooks AN, et al. (2023) Debugging and consolidating multiple synthetic chromosomes reveals combinatorial genetic interactions. *Cell* **186**: 5220–5236.e16.
- Zhu YO, Sherlock G, Petrov DA (2016) Whole Genome Analysis of 132 Clinical *Saccharomyces cerevisiae* Strains Reveals Extensive Ploidy Variation. *G3 GenesGenomesGenetics* **6**: 2421–2434.
- Zimmer A, Durand C, Loira N, Durrens P, Sherman DJ, Marullo P (2014) QTL Dissection of Lag Phase in Wine Fermentation Reveals a New Translocation Responsible for *Saccharomyces cerevisiae* Adaptation to Sulfite. *PLoS One* **9**: e86298.

P

Publications, communications and teaching



Happy 30th birthday, S288C. Made from pigmented yeast strains spotted onto agar by the SINGER PIXL robot, colour corrected in Siril³.

³Richard C, Hourdin V, Melis C, Knagg-Baugh A (2024) Siril: An Advanced Tool for Astronomical Image Processing. J Open Source Softw 9: 7242.

List of publications

- 202x** **Environmental context shapes the cross-omic consequences of natural genetic variation in yeast**
In preparation.
- 202x** **Natural variation in alternative polyadenylation across *Saccharomyces cerevisiae* isolates**
Brach G., Papadopoulos C., Montañés J.C., Albà M.M., Schacherer J.
In preparation.
- 2026** **From one genome to thousands, and beyond**
Brach G., Schacherer J.
FEMS Yeast Research. [doi 10.1093/femsyr/foag038](https://doi.org/10.1093/femsyr/foag038)
- 2026** **Population-scale chemical response revealed by a barcoded yeast collection**
Dutta A., Garin M., Loegler V., **Brach G.**, Friedrich A., Yoshimura M., Hirano H., Osada H., Boone C., Yashiroda Y., Hou J., Schacherer J.
Nature Communications. [doi 10.1038/s41467-026-73532-z](https://doi.org/10.1038/s41467-026-73532-z)
- 2026** **Domestication drives repeated evolution of sexual–asexual life cycle trade-offs in yeast**
Becerra-Rodríguez C., Thiele P., **Brach G.**, Dutta A., Garin M., Tan G., Loegler V., Friedrich A., Andrews B., Boone C., Schacherer J., Hou J.
PNAS. [doi 10.1073/pnas.2526682123](https://doi.org/10.1073/pnas.2526682123)
- 2025** **From genotype to phenotype with 1,086 near telomere-to-telomere yeast genomes**
Loegler V., Thiele P., Teyssonnière E., Tsouris A., **Brach G.**, Cruaud C., Payen E., Engelen S., Dunham M.J., Hou J., Friedrich A., Schacherer J.
Nature. [doi 10.1038/s41586-025-09637-0](https://doi.org/10.1038/s41586-025-09637-0)
- 2024** **Multiple independent losses of crossover interference during yeast evolutionary history**
Dutta A., Dutreux F., Garin M., Caradec C., Friedrich A., **Brach G.**, Thiele P., Gaudin M., Llorente B., Schacherer J.
PLOS Genetics. [doi 10.1371/journal.pgen.1011426](https://doi.org/10.1371/journal.pgen.1011426)
- 2024** **Diallel panel reveals a significant impact of low-frequency genetic variants on gene expression variation in yeast**
Tsouris A.[†], **Brach G.**[†], Friedrich A., Hou J., Schacherer J.
Molecular Systems Biology. [doi 10.1038/s44320-024-00021-0](https://doi.org/10.1038/s44320-024-00021-0)
- 2024** **Non-additive genetic components contribute significantly to population-wide gene expression variation**
Tsouris A., **Brach G.**, Schacherer J., Hou J.
Cell Genomics. [doi 10.1016/j.xgen.2023.100459](https://doi.org/10.1016/j.xgen.2023.100459)

[†] These authors contributed equally.

Oral communications

- 2026** **Population-scale quantitative transcriptomes and proteomes across environments in *Saccharomyces cerevisiae***
PYFF9, Conference on Physiology of Yeasts and Filamentous Fungi. Talk.
València, Spain, 8–11 September 2026
- 2025** **Population-scale quantitative transcriptomes and proteomes across environments in yeast**
Yeast 2025 (ICYGMB32). Main-stage talk.
Paris, France, 21–24 July 2025
- 2024** **Exploration of transcript diversity in a yeast natural population using direct RNA sequencing**
EMBO Evolutionary and Comparative Genomics. Poster. Best Poster Award.
Nafplion, Greece, 3–6 November 2024

Teaching

- 2023–2024** **Génétique fondamentale et appliquée**
Tutorials for second-year bachelor's students: phenotypic segregation in crosses, genetic linkage and auxotrophy, taught through worked examples.
Licence Sciences de la vie, parcours Biologie moléculaire et cellulaire, Université de Strasbourg
- 2022–2023** **Propriétés et mécanismes des enzymes**
Tutorials for third-year bachelor's students: enzyme properties and catalytic mechanisms, taught through worked examples.
Licence Sciences de la vie, parcours Biologie moléculaire et cellulaire, Université de Strasbourg

Abstract

Genetic variation shapes phenotypes through molecular layers, yet association studies often focus on common variants, gene-level RNA abundance and one environment. Using the natural diversity of the baker's yeast *Saccharomyces cerevisiae*, this thesis examines how structural variants (SVs), transcript structure and environmental context shape molecular phenotypes. A 26-parent half-diallel looked at the influence of low-frequency variants and SVs on gene expression variation. Oxford Nanopore direct RNA sequencing of six isolates revealed variation in transcript-end use and poly(A)-tail length beyond gene-level abundance. Matched transcriptomes and proteomes from more than 550 isolates across four environments showed genotype-by-environment interactions and largely condition-specific QTL. Transcript effects retained at protein level were concordant, but their transmission varied among environments. Together, this work tackles hidden dimensions of molecular variation and shows that understanding the genotype–phenotype relationship requires integrating variant classes, transcript structures, molecular layers and environmental context.

Gauthier Brach

Team intraspecific variation and genome evolution
Institute of Genetics and Molecular and Cellular Biology
UMR 7104 – UMR-S 1258 – CNRS, Inserm
University of Strasbourg
Ph.D. advised by Prof. Joseph Schacherer

