ScRAPdb: an integrated pan-omics database for the Saccharomyces cerevisiae reference assembly panel

Zepu Miao, Yifan Ren, Andrea Tarabini, Ludong Yang, Huihui Li, Chang Ye, Gianni Liti, Gilles Fischer, Jing Li, Jia-Xing Yue

Nucleic Acids Research, 2024, 1–12
https://doi.org/10.1093/nar/gkae955
Database issue

ScRAPdb: an integrated pan-omics database for the *Saccharomyces cerevisiae* reference assembly panel

Zepu Miao [1,†], Yifan Ren [1,†], Andrea Tarabini [2,†], Ludong Yang [1], Huihui Li [1], Chang Ye [3], Gianni Liti [4], Gilles Fischer [2], Jing Li [1,*] and Jia-Xing Yue [1,*]

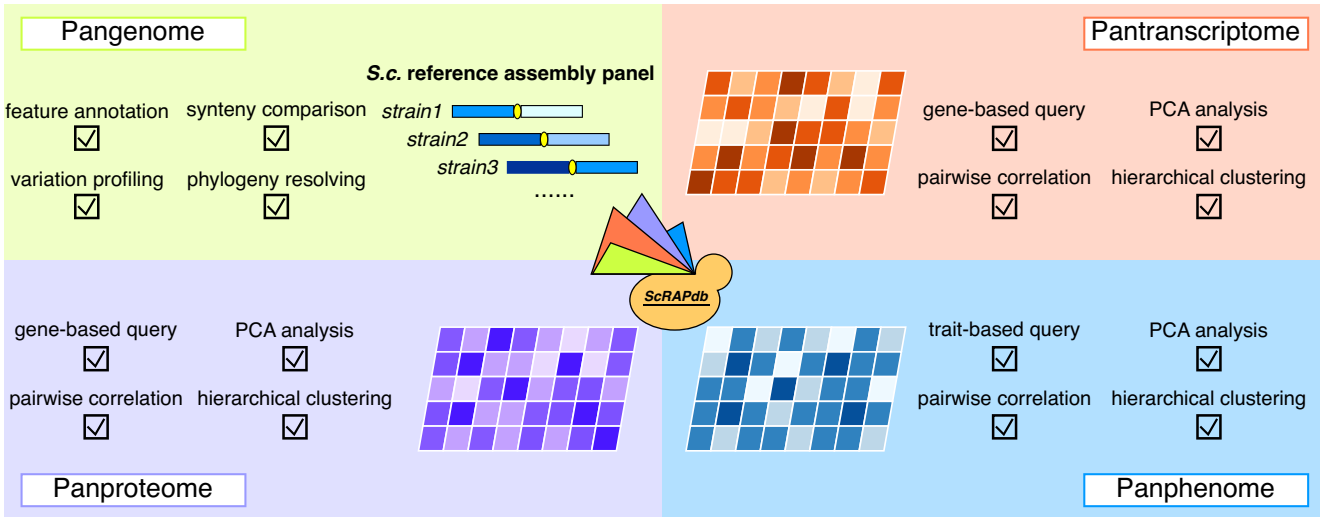
[1]State Key Laboratory of Oncology in South China, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, 651 Dongfeng East Road, Guangzhou 510060, China
[2]Sorbonne Université, CNRS, Institut de Biologie Paris-Seine, Laboratory of Computational and Quantitative Biology, 7-9 Quai Saint Bernard, Paris 75005, France
[3]Department of Chemistry, University of Chicago, 929 E 57th Street, Chicago, IL 60637, USA
[4]CNRS, INSERM, IRCAN, Université Côte d'Azur, 28 Avenue de Valombrose, Nice 06107, France

[*]To whom correspondence should be addressed. Tel: +86 20 87340150; Email: yuejiaxing@gmail.com
Correspondence may also be addressed to Jing Li. Tel: +86 20 87340260; Email: lijing3@sysucc.org.cn
[†]The first three authors should be regarded as Joint First Authors.

Abstract

As a unicellular eukaryote, the budding yeast *Saccharomyces cerevisiae* strikes a unique balance between biological complexity and experimental tractability, serving as a long-standing classic model for both basic and applied studies. Recently, *S. cerevisiae* further emerged as a leading system for studying natural diversity of genome evolution and its associated functional implication at population scales. Having high-quality comparative and functional genomics data are critical for such efforts. Here, we exhaustively expanded the telomere-to-telomere (T2T) *S. cerevisiae* reference assembly panel (ScRAP) that we previously constructed for 142 strains to cover high-quality genome assemblies and annotations of 264 *S. cerevisiae* strains from diverse geographical and ecological niches and also 33 outgroup strains from all the other *Saccharomyces* species complex. We created a dedicated online database, ScRAPdb (https://www.evomicslab.org/db/ScRAPdb/), to host this expanded pangenome collection. Furthermore, ScRAPdb also integrates an array of population-scale pan-omics atlases (pantranscriptome, panproteome and panphenome) and extensive data exploration toolkits for intuitive genomics analyses. All curated data and downstream analysis results can be easily downloaded from ScRAPdb. We expect ScRAPdb to become a highly valuable platform for the yeast community and beyond, leading to a pan-omics understanding of the global genetic and phenotypic diversity.

Graphical abstract



Received: August 14, 2024. **Revised:** October 5, 2024. **Editorial Decision:** October 8, 2024. **Accepted:** October 10, 2024

Introduction

Model organisms are invaluable for understanding life on earth, shedding lights on almost all aspects of modern biology. The budding yeast *Saccharomyces cerevisiae* has long been considered as a classic and powerful eukaryotic model organism for addressing a wide range of biological questions across diverse research fields, leading to many landmark discoveries with profound impacts. Moreover, *S. cerevisiae* has also been widely used as a proof-of-concept system for developing new experimental and computational methodologies, thanks to its unicellular nature, compact genome and relatively well-characterized regulatory machinery. Finally, *S. cerevisiae* is also an important workhorse in industrial settings like brewing, bakery, biofuel and pharmaceutics, highlighting its economic values for the human society. It is therefore of no coincidence that *S. cerevisiae* become the first eukaryotic organism to have its genome fully sequenced (1). Since then, the development of large-scale functional genomics panels such as yeast gene deletion (2), overexpression (3), tagging (4) and gene–gene interaction (5) further cast a comprehensive functional understanding of its genome biology. To accommodate and curate these reference-orientated resource, the *Saccharomyces* Genome Database (SGD, https://www.yeastgenome.org) (6) was developed >25 years ago and continued serving the yeast community (7), providing encyclopedic information regarding the genome, genes, proteins and other encoded features of *S. cerevisiae*.

While these landmark efforts were mostly achieved based on the reference strain S288C or closely related laboratory strains, large-scale population genomic and comparative studies were subsequently deployed for *S. cerevisiae* and its close relatives (e.g. *Saccharomyces paradoxus*) (8–16). For example, along with the advancing of sequencing technologies, three phases of the *Saccharomyces* Genome Resequencing Project (SGRP) were conducted with Sanger (8), short-read (9) and long-read sequencing (12), respectively, to obtain a more complete view of the genomic and phenotypic diversity of the *Saccharomyces* natural populations. In addition, the 1002 *S. cerevisiae* genome project (1002ScGP) represents another pivotal landmark by providing a massive panel of short-read-based genome assemblies for 1011 strains sampled globally together with a comprehensive pangenome open reading frames (ORFs) portrait for *S. cerevisiae* regarding its genome content evolution (13). Comparatively, long-read sequencing technologies such as PacBio and Oxford Nanopore substantially improves the quality and continuity of genome assembly, whose application on yeast genomes can often reach telomere-to-telomere (T2T) completeness for multiple chromosomes, without incorporating additional auxiliary technologies (e.g. Hi-C) (17). Recently, a long-read-sequenced panel of 142 geographically and ecologically representative strains was further assembled to delineate a high-resolution view on the structural genome evolution of *S. cerevisiae* (18). Long-read-based high-quality yeast genome assemblies for *S. cerevisiae* and its close relatives from the *Saccharomyces* species complex are still released on a regular basis (19–31). Given the strength of long-read sequencing in characterizing complex structural variants (SVs) and highly repetitive regions, it is expected that a curated compendium of these high-quality assemblies will provide a new foundation for unbiased population and comparative genomics studies toward a pangenome view of a species' global diversity.

Understanding how genomic variation translates into trait differences in the phenotypic space is a central aim in genetics. As a classic model organism with population-level scalability for multi-omics assays, *S. cerevisiae* is well suited for this undertake. The genome characterization of 1011 *S. cerevisiae* strains and their associated phenotyping data across multiple environments set the stage for an in-depth examination on how genetic diversity, population structure, domestication history, collectively shaped the phenotypic diversity at the species level (13,32). Recently, their associated high-throughput transcriptome (33) and proteome (34,35) data further rolled out, which bridges the gap between the genotypic and phenotypic space and opens up the opportunity for looking into how such genotype-to-phenotype translation is fulfilled at the mechanistic level.

Therefore, it is important to construct a centralized data hub enabling seamless cross-dataset exploration and knowledge integration. Toward this goal, here we introduce ScRAPdb, an integrated pan-omics database for *S. cerevisiae*, offering a multi-layered view on its pan-omics diversity at population scales. At the genomic level, we conducted an exhaustive search, curation, and annotation on all currently available long-read-based genome assemblies from the *Saccharomyces* species complex to form an expanded *S. cerevisiae* reference assembly panel (ScRAP) and performed comprehensive comparative genomics analyses. On top of this, we also gathered the *S. cerevisiae* pantranscriptome, panproteome and panphenome datasets and matched with our expanded ScRAP collection for multi-omics integration. Finally, rich visualization and analysis tools were developed and incorporated into ScRAPdb to facilitate interactive data exploration in real time. Taken together, we anticipate ScRAPdb to facilitate both basic and applied research on yeast and beyond, in terms of both biological discovery and technology development. The ScRAPdb is publicly available at https://www.evomicslab.org/db/ScRAPdb/.

Materials and methods

Exhaustive collection for long-read-based yeast genome assemblies

We collected all currently available (until 10 June 2024) long-read-based genome assemblies (for both nuclear and mitochondrial genomes if available) and their associated metadata for *S. cerevisiae* and its close relatives from the *Saccharomyces* species complex. These outgroup species include *S. paradoxus*, *Saccharomyces mikatae*, *Saccharomyces kudriavzevii*, *Saccharomyces arboricola*, *Saccharomyces eubayanus*, *Saccharomyces uvarum*, as well as the newly described *Saccharomyces jurei* (36) and *Saccharomyces chiloensis* (26). For each included *Saccharomyces* species, a combination of manual literature curation, GenBank query and automatic web crawling was used. For GenBank query, all chromosome-level *Saccharomyces* genome assemblies uploaded to GenBank (37) since 2015 were retrieved and those sequenced by long-read sequencing technologies were kept for downstream quality control. For automatic web crawling, we developed a python script that utilize the Biopython's Entrez module (https://biopython.org/docs/1.83/api/Bio.Entrez.html) (38) for information query and parsing. This in-house script allows to perform an extensive and in-depth search in literature for *S. cerevisiae* genome assemblies generated

by long-reads (39,40). Long-read-based yeast genome assemblies of meiotic spores from hybrid crosses were not considered (41).

Curation and annotation for the expanded ScRAP collection

All collected genome assemblies were processed with our previously developed LRSDAY pipeline (42) (v1.7.2) for assembly curation, chromosome-level scaffolding and genomic feature annotation (centromere, gene, transferRNA [tRNA], Ty transposable element, X element and Y' element). For nuclear assemblies from the outgroup species, two consecutive rounds of nuclear gene annotation were performed (the first round with default setting and the second round using the outputs from the first-round annotation as hints for potential coding regions). Assemblies that are too fragmented, noticeably incomplete, or carrying too many base-level errors (as suggested by excessive in-frame stop codons detected during our gene annotation) were eliminated after quality check. Manual assembly correction was occasionally applied for very clear mis-joining errors based on the dotplot generated by LRSDAY. For strains having multiple assemblies derived from the same read set (e.g. those from the original ScRAP project), the best assembly is kept based on the assembly quality. For assemblies derived from the same strain but independent reads sets (e.g. when two labs sequenced the same strain), we kept them altogether while using different assembly identifiers (e.g. asm01, asm02, etc.) to distinguish. The nuclear and mitochondrial reference assemblies of *S. cerevisiae* (denoted as 'SGDref') and *S. arboricola* (for which no long-read-assembly is currently available) were further added into our collection. The sequence start of all enrolled mitochondrial genome assemblies were uniformly reset to the start codon of the *ATP6* gene.

For each assembly hosted in ScRAPdb, its identifier consists of three parts: strain identifier (e.g. S288C), assembly identifier (e.g. asm01) and phasing status tag (e.g. HP0). The assembly identifiers are numbered sequentially as asm01, asm02, etc., which are used to differentiate genome assemblies generated from different reads sets (e.g. independently sequenced by different groups) for the same strain. The phasing status tag is defined in the same way as in the original ScRAP paper (18). For haploid and homozygous diploid/polyploid strains, a single haplotype can be recovered with no phasing process needed. We used 'HP0' as their phasing status tag, where 'HP' stands for haplotype. For strains that are heterozygous diploids, polyploids or with unknown zygosity and ploidy status, a single genome assembly is usually generated with different haplotypes being collapsed together. We used the 'collapsed' tag to denote their phasing status. For heterozygous diploid strains that were assembled in a haplotype-phased and separated manner, we used 'HP1' and 'HP2' to denote the two assemblies that correspond to the two separated haplotypes. For polyploid strains that were assembled in a haplotype-phased but unseparated manner, we used the 'HP' tag to denote the corresponding assemblies. All enrolled nuclear genome assemblies were evaluated by BUSCO (43) (v5.3.2) with BUSCO-associated lineage-specific database 'ascomycota_odb10' (creation date: 8 August 2024, number of genomes: 365, number of BUSCOs: 1706). BlobToolKit (44) (v4.3.11) was used for calculating assembly statistics such as N50, N90, GC%, etc.

Pairwise genome comparison, ANI calculation and full-spectral variant calling

For nuclear and mitochondrial genomes, pairwise genome comparison and average nucleotide identity (ANI) calculation were conducted by OrthoANI (45) (v0.50). Full-spectral genomic variants such as single-nucleotide variants (SNVs), insertions/deletions (INDEL) and SVs were detected using PAV (46) (v2.3.4) based on the *S. cerevisiae* reference genome (SGDref). The called variants were further processed with VEP (47) (v109.3) for variant effect prediction. The called SNVs, INDELs and SVs were further visualized via a built-in genome browser.

Phylogenetic reconstruction

For the 364 nuclear genome assemblies (after excluding the 13 phased but unseparated polyploidy assemblies from the original ScRAP) and 237 mitochondrial genome assemblies, we used Proteinortho (48) (v6.0.25; options: –check -selfblast -singles) to define 1-to-1 ortholog groups for four input subsets: (i) all 328 *S. cerevisiae* nuclear assemblies, (ii) all 36 outgroup nuclear assemblies + 11 representative *S. cerevisiae* nuclear assemblies, (iii) all 220 *S. cerevisiae* mitochondrial assemblies and (iv) all 17 outgroup mitochondrial assemblies together with 11 representative *S. cerevisiae* mitochondrial assemblies. For each 1-to-1 ortholog groups identified based on these subsets, the corresponding protein and CDS alignment were generated by MACSE (49) (v2.04; options: -prog alignSequences -gc_def 1 [for nuclear genomes] or 3 [for mitochondrial genomes] -seq $i.species_relabeled.fa -out_NT $i.macse_NT.aln.fa -out_AA $i.macse_AA.aln.fa and -prog exportAlignment -align $i.macse_NT.aln.fa -codonForFinalStop --- -codonForInternalStop NNN -codonForInternalFS NNN -codonForExternalFS --- -charForRemainingFS - -out_NT $i.macse_NT.aln.tidy.fa -out_AA $i.macse_AA.aln.tidy.fa). Afterwards, the concatenated supermatrix of the 1-to-1-ortholog-based CDS alignment was further generated. For mitochondrial assemblies, there are two identical copies of the *COX3* gene in strain UCD_61–190-6A (aka, CDN), for which only one copy was used for orthology identification. For each input subset, the CDS supermatrix and its associated partition definition (by first, second and third codon positions) were used for maximal likelihood tree building by IQ-TREE (50) (version: 2.3.4; options: -p $prefix.concatenated.cds.by_codon_position.partition.txt -s $prefix.concatenated.cds.tidy.fa -m MFP -bb 1000 -alrt 1000 –bnni -T $threads -pre $prefix.iqtree -safe). In total, 1000 rounds of ultrafast bootstrap (UB) and approximate likelihood-ratio test (aLRT) were used to assess the branch supports.

Pan-omics data curation and visualization

The previously collected pangenome ORF set (13) (1011 strains), pantranscriptome (33) (969 strains × 1 environment), panproteome (34,35) (942 strains × 1 environment and 796 strains × 1 environment, respectively) and panphenome data (13,32) (971 strain × 35 traits and 1011 strains × 99 traits, respectively) of the *S. cerevisiae* global populations were manually curated and matched with each other as well as our expanded ScRAPdb genome collection. For each of our previously characterized *S. cerevisiae* pangenome ORF set, its presence/absence status in the ScRAPdb genome collection as

well as in the 1002ScGP genome collection was evaluated by minimap2 (51) (v2.24). The pangenome ORFs with >80% alignment coverages and >80% sequence identities are considered as presence in the evaluated genome. For the pantranscriptome data, both the read count values calculated by featureCounts (52) and their corresponding transcript per million (TPM) values were retrieved from the original study, the latter of which can be used for direct comparison of gene expression levels across different strains and environments. For the panproteome data, the DIA-NN-inferred protein abundance values (53) obtained from SWATH mass spectrometry (54) were retrieved from the original studies. For the panphenome data, four classes of phenotypic traits data (i.e. generation time, yield, sporulation and chronological life span) were retrieved from the original studies. For each dataset, interactive plots such as heatmaps, violin plots and principal component analysis (PCA) plots were employed for data visualization.

Database implementation

The backend of ScRAPdb is written using the Django framework (https://www.djangoproject.com/; v3.2.4). On the frontend, Echarts (55) (v5.3.3), D3.js (https://d3js.org/; v6.7.0) and Plotly (https://plotly.com/; v2.33.0) were employed to for interactive graph plotting. Phylotree.js (56) (v1.0.16) was used for phylogenetic tree visualization. JBrowse (57) (v2.13.1) was used to power the genome browser. Bootstrap (https://getbootstrap.com/; v4.3.1) and jQuery (https://jquery.com/; v3.2.1) were further used for interactive query and rendering. The website was host via an Alibaba Simple Application Server equipped with 2 CPU, 4 Gb RAM and 280 Gb ESSD data storage, running with CentOS Linux (v8.2) as the operating system.

Results

Data collection and organization

Starting from the original ScRAP dataset (18), we performed an exhaustive search and curation on all currently available long-read-based genome assemblies for both *S. cerevisiae* and its eight outgroup species from the *Saccharomyces* species complex to form a significantly expanded ScRAP collection. This updated ScRAP collection consists of 341 *S. cerevisiae* and 36 outgroup nuclear assemblies as well as 220 *S. cerevisiae* and 17 outgroup mitochondrial assemblies. These assemblies are further derived from 264 *S. cerevisiae* strains and 33 outgroup strains (12 *S. paradoxus*, 2 *S. mikatae*, 2 *S. jurei*, 3 *S. kudriavzevii*, 1 *S. arboricola*, 2 *S. chiloensis sp. nov. AUS*, 4 *S. chiloensis sp. nov. SA-C*, 5 *S. eubayanus* and 2 *S. uvarum*) (Figure 1A). Following the same protocols (42), we performed systematic assembly curation, statistics evaluation and genomic feature annotation for all these genome assemblies, so that they can be directly compared with each other at both genome and gene levels. Detailed metadata for these strains and assemblies was also curated and compiled, such as their geographical origin, ecological niches, ploidy, zygosity, aneuploidy, mating type, marker gene genotypes and used sequencing technologies (Figure 1B). Comprehensive comparative genomics analyses such as gene orthology identification, genomic variant discovery, phylogeny reconstruction and synteny comparison were conducted accordingly, with their results rendered to the web-based front-end for interactive and intuitive exploration (Figure 1C). Finally, given the recent availability of pangenome ORF set (13), pantranscriptome (33), panproteome (34,35) and panphenome (13,32) data based on the 1002ScGP strain collection (13), we gathered and curated such multi-omics atlases and matched it to our expanded ScRAPdb genome collection, which enables multi-omics-based data exploration and comparison (Figure 1D–F). An overview of 1-to-1-orthologous-gene-based phylogeny of all ScRAPdb-enclosed strains together with their matched multi-omics atlases is further presented in Figure 2.

Web interface and usages

ScRAPdb comes with an intuitive web interface, which helps to navigate users to access different functional modules. A central horizontal navigation menu is provided in the front page with ten clickable tabs: 'Home', 'Strains', 'Phylogeny', 'Pangenome', 'Pantranscriptome', 'Panproteome', 'Panphenome', 'Tools', 'Download' and 'Help', each leading to either a dedicated page or a secondary-level menu (Figure 3). The 'Home' page gives a brief introduction to the ScRAPdb database with various summary plots for the strains and assemblies enrolled regarding their geographical distribution, ploidy and haplotype phasing, sequencing technologies, etc. The 'Strains' page provides a metadata table for all the strains enrolled, with clickable links leading to more strain-specific information. The 'Phylogeny' tab leads to phylogenetic trees based on the nuclear and mitochondrial genomes respectively. Users can easily interact with these trees with mouse clicks and drags. The 'Pangenome', 'Pantranscriptome', 'Panproteome' and 'Panphenome' pages present the pan-omics datasets corresponding to our ScRAPdb strains, with each dataset summarized in both graphical plots and informative tables. Gene-based query function (in both single-gene and batch-gene modes) is further built-in for fast search and analysis. The 'Tools' page leads to web-based tools for genome browsing, synteny comparison and homology search. The 'Download' page offers both batch and individual downloading links for all genome sequences and annotations as well as other precalculated results curated in ScRAPdb, including a compilation of full-spectral genomic variant call sets for SNVs, INDELs and SVs. Finally, the 'Help' page provides useful helping information regarding the data source, naming convention, contact information, etc.

Application demonstration 1: high-resolution discovery and characterization of structural variants

Long-read-based genome assemblies shine in their value to detect large and complex SVs, which can significantly impact functions and traits. The ScRAPdb compendium of such population-scale long-read-based *S. cerevisiae* genome assemblies offers a powerful platform to discover and characterize SVs accumulated in diverse evolutionary lineages of *S. cerevisiae*. With ScRAPdb, the existence of SVs can be captured and analyzed from three sources: (i) the primary-reference-based SV call set between each enrolled assembly and the SGD reference genome (*S. cerevisiae* strain: S288C); (ii) the interactive genome synteny comparison plots in multiple flavors (e.g. circular synteny plot, linear synteny plot and genome dotplot); (iii) the systematic characterization for the presence and

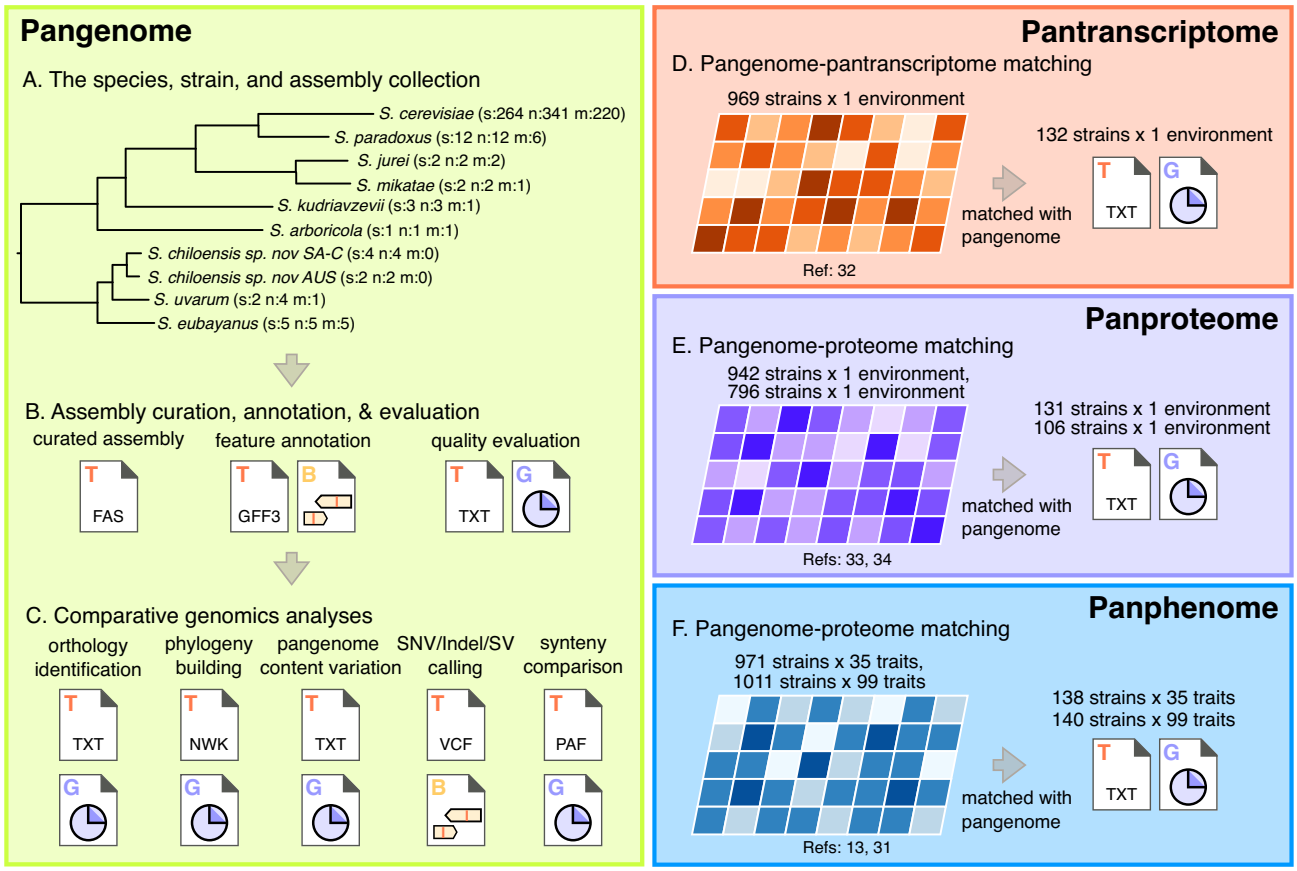


Figure 1. Overview of data collected and generated for ScRAPdb. (**A**) The species, strain and assembly collection. For each included species, the number of strains (s), nuclear assemblies (n) and mitochondrial assemblies (m) are indicated inside the corresponding parenthesis that appears after the species name. (**B**) Assembly curation, annotation and evaluation. (**C**) Comparative genomics analyses. (**D**–**F**) The pangenome-based matching with pantranscriptome, panproteome and panphenome data. For (B–F), major analyses and the corresponding result files generated were depicted. The specific file types ('T' for text files, 'G' for graphical files, 'B' for browser visualization) and formats (e.g. TXT, NWK, VCF and PAF) are further indicated. TXT: the plain text format; NWK: the Newick tree format; VCF: the variant call format; PAF: the pairwise mapping format.

absence variation (PAV) of the previously portraited *S. cerevisiae* pangenome ORF set across 1011 strains (13). Here, we used these three methods to highlight the polymorphic distribution of the killer toxin gene *KHR1* (SGD systematic name: YSC0002) among different *S. cerevisiae* strains (Figure 4). *KHR1* is absent from the SGD reference genome but exists in strains such as YJM789 (58). By searching *KHR1* on the genome content variation page under the 'Pangenome' menu of ScRAPdb, we can obtain a detailed description regarding its function and PAV status across different *S. cerevisiae* genome assemblies curated in ScRAPdb, which suggests the presence of *KHR1* in the strain SK1 (Figure 4A–C). By running BLAST search (natively supported by ScRAPdb) with the *KHR1* coding sequence (CDS; size = 891 bp) retrieved from SGD, we can easily identify the corresponding genomic location of *KHR1* in the SK1 genome assembly (chrIX:285388–286278) (Figure 4D). By zooming into this region in the interactive pairwise genome dotplot powered by ScRAPdb, we can not only verify this SV but also obtain its insertion location relative to the SGDref (chrIX:300544–300545) (Figure 4E and F). By examining this location along SGDref with ScRAPdb's built-in genome browser, we can further confirm a 1622-bp structural insertion containing the *KHR1* gene in SK1 as reflected by the pre-loaded SK1 genomic variant track (Figure 4G).

Application demonstration 2: understanding trait evolution across different omics layers

The domestication history of *S. cerevisiae* in food and beverage production dates back to thousands of years ago. Such long-term adaptation in specific anthropic environments has shaped their genome and trait evolution substantially. For example, sulfites are widely used in wine production as preservatives to maintain the flavor and freshness of wine. Accordingly, many *S. cerevisiae* strains from the Wine/European clade showed higher tolerance to SO_2, as a result from the use of sulfites during winemaking. The *SSU1* (*YPL092W*) gene encodes the sulfite efflux pump on the plasma membrane and is largely responsible for SO_2 tolerance in *S. cerevisiae* (59). With ScRAPdb, by comparing the gene expression level of *SSU1* across strains from different phylogenetic clades, we found four Wine/European strains with exceptionally high levels of *SSU1* expression, namely YJM981 (ADI), YJM978, C-6 (CRL) and CBS2807 (AIC) (Figure 5A and B). In parallel, the ScRAPdb genome synteny comparison between these strains and SGDref revealed clear chrVIII-chrXVI translocations in YJM981 (ADI), YJM978 and C-6 (CRL), which are also the top three strains with the highest levels of *SSU1* expression (Figure 5C). Previous studies have demonstrated that such chrVIII-chrXVI translocation can substantially elevate *SSU1* expression via a promoter hijack (60,61). Interestingly,

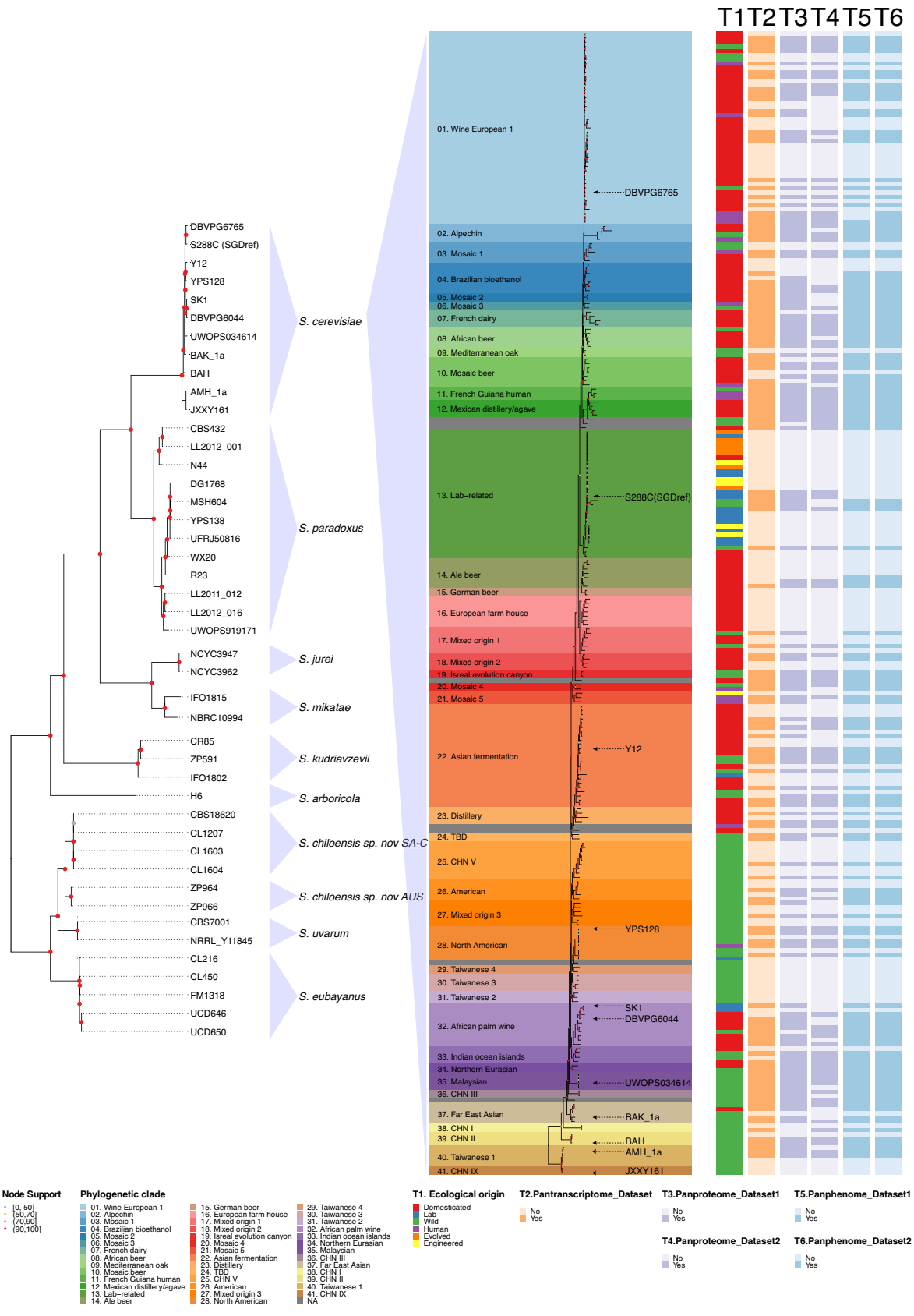


Figure 2. Phylogenetic distribution of yeast strains enclosed in ScRAPdb. The 1-to-1-orthologous-gene-based phylogenetic tree of yeast strains enclosed in ScRAPdb overlayed with their phylogenetic-clade assignments. The tracks T1-T6 represent the availability of the matched pantranscriptome, panproteome and panphenome datasets. The node supports (assessed by aLRT) for the phylogenetic tree are indicated by colored dots. Matched pan-omics datasets include Pantranscriptome Dataset (33), Panproteome_Dataset1 (34), Panproteome_Dataset2 (35), Panphenome_Dataset1 (13) and Panphenome_Dataset2 (32).

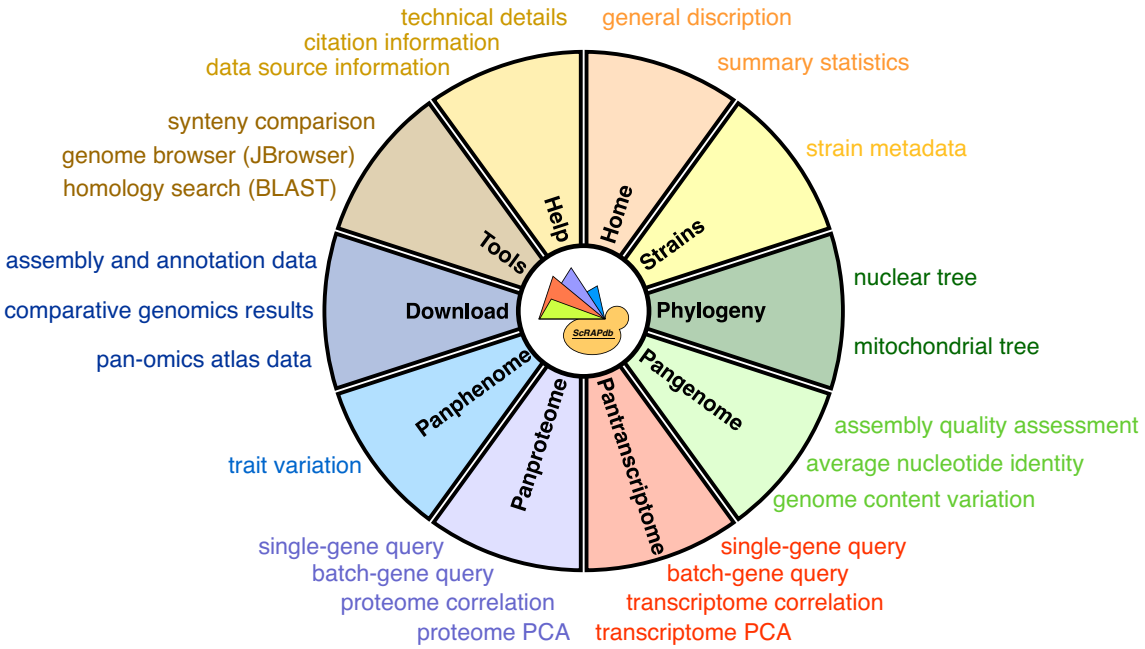


Figure 3. Menu organization for ScRAPdb. The top-level menus of ScRAPdb are shown in the inner cycle, while the corresponding features and functions are further called out in the outer cycle.

CBS2807 (AIC) is free from both this characteristic chrVIII-chrXVI translocation and two other known SVs related to *SSU1* (62,63), suggesting an alternative evolutionary strategy that is yet to be elucidated for *SSU1* up-regulation. In addition to winemaking, dairy fermentation is another typical domestication niche for *S. cerevisiae*. *GAL1* (*YBR020W*) and *GAL3* (*YDR009W*) are a pair of paralog genes in the *S. cerevisiae* genome that mediate galactose metabolic process. With ScRAPdb, we found these two genes show much higher expression levels in strains from the French dairy clade relative to all the other strains, which further agrees with the high growing yields of these strains in the galactose environment (Figure 5E). Therefore, higher expression levels, coupled with previously reported alternative *GAL* alleles (64,65), may synergistically contribute to enhanced galactose utilization. Taken together, these two examples showcased how ScRAPdb with its multi-omics integration can help researchers to better study trait evolution across different genomic and functional layers.

Discussion

While the budding yeast *S. cerevisiae* has been serving as an all-time classic model organisms in modern biology, the past 15 years truly witnessed its emergence as a leading system to study genomics and evolution at population scales. Till today, >3000 of strains that are geographically and ecologically distinct have been fully sequenced (66), exhibiting a surprisingly high level of genome diversity and plasticity. Considering their over 9000 years of association with the human society (67), it is expected that human domestication on *S. cerevisiae* in diverse industrial settings has played an important role along the way. Moreover, the recent release of large-scale transcriptomics, proteomics and phenomics data across hundreds of genetic backgrounds and environmental conditions further offered us invaluable resources and power to dissect complex genotype-to-phenotype interactions and to obtain a multi-layered view on how genomic variation translates into phenotypic differences. Therefore, it is critical to develop a centralized data hub to enable seamless integration, exploration and sharing of these multi-omics treasures. Such a unified platform is highly useful for not only the yeast enthusiasts but also a much broader research community for formulating and testing new hypotheses, models and tools in addressing general biological questions related to genotype-phenotype mapping.

In this study, we present ScRAPdb, an integrated pan-omics database for the ScRAP. Compared with the classic yeast genomics databases such as SGD (6) and YGOB (11), ScRAPdb shines with its unique emphasis on population-scaled pangenome characterization and multi-omics integration. In this sense, ScRAPdb filled the gap between the single-reference-centered SGD database and the inter-specific-comparison-motivated YGOB database by offering a novel platform for exploring intra-specific diversity of *S. cerevisiae* across multiple omics layers spanning between the genotype and phenotype space.

In the future, ScRAPdb will be updated on a regular basis to keep incorporating reference-quality genomes as well as large-scale functional assays for *S. cerevisiae* and its close relatives. Furthermore, towards the ultimate goal of elucidating the genetic basis of complex traits, we plan to add supports for user-uploaded phenotype data as well as built-in genome-wide association mapping tools to enable direct association tests across multi-omics layers. In addition, *S. cerevisiae* is well ahead of the time in genome editing, thanks to its highly effective homologous recombination system. As newer genome editing tools such as the clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) systems become increasingly accessible and effective, genome-wide editing and perturbation experiments are now possible for being applied to diverse natural strains of *S. cerevisiae* in parallel. ScRAPdb is expected to further incorporate such data when it becomes available, which will add another dimension for deciphering genotype-phenotype interaction and their

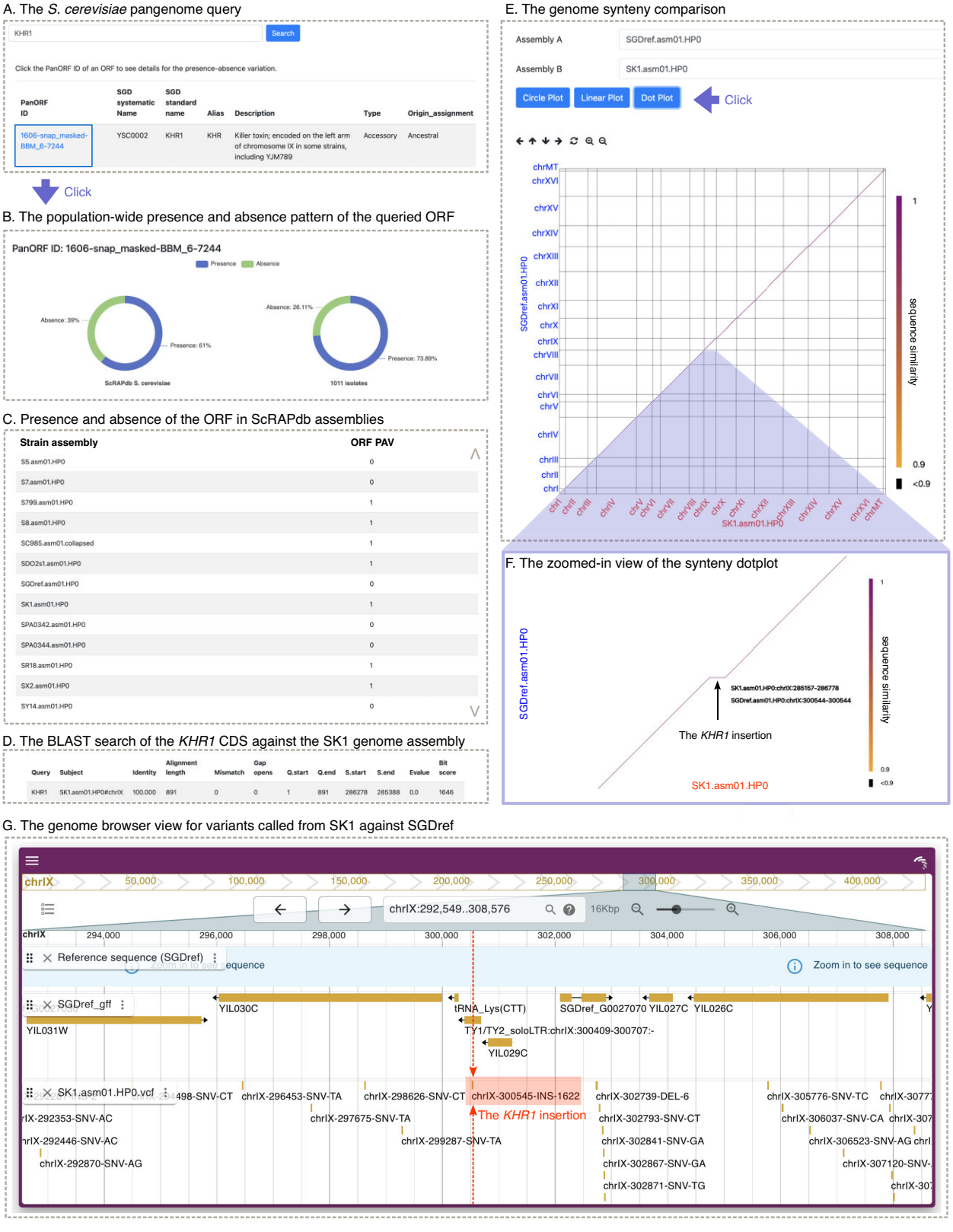


Figure 4. Structural variant characterization with ScRAPdb. (**A**) Pangenome ORF query of the *KHR1* gene. (**B**) The population-wide distribution of the query gene (*KHR1*) in the ScRAPdb and 1002ScGP strain panels, respectively. (**C**) The presence/absence status of the query gene (*KHR1*) in each ScRAPdb *S. cerevisiae* genome assembly. (**D**) The BLAST result of the query gene (*KHR1*) against the SK1 genome assembly in ScRAPdb. (**E**) The genome-wide dotplot for the pairwise comparison between SGDref and SK1 genome assemblies. (**F**) The zoom-in view of the SGDref versus SK1 genome dotplot. (**G**) The genome browser view of ScRAPdb for the *KHR1* insertion relative to SGDref and its flanking region at chrIX:292549–308576.

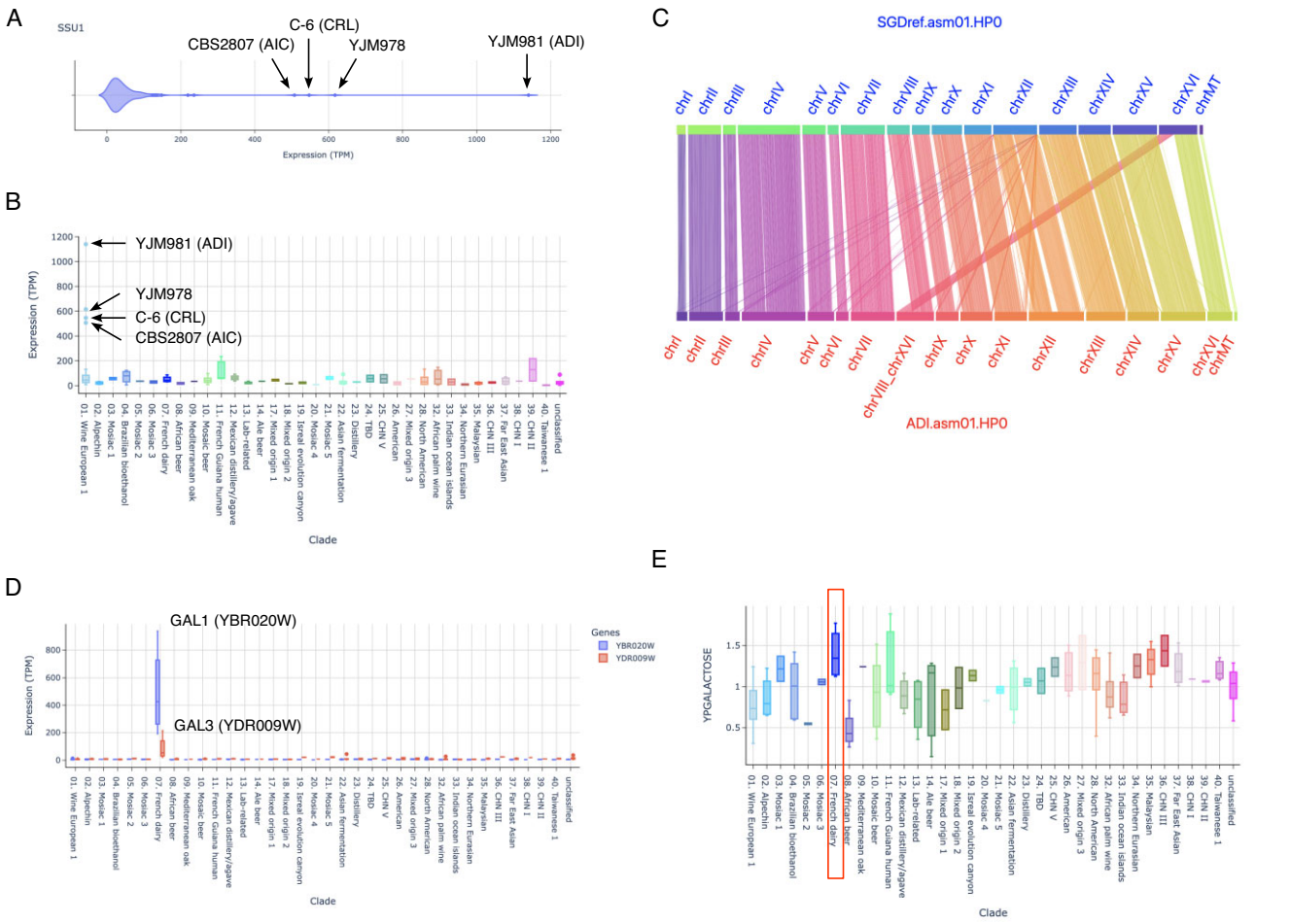


Figure 5. Multi-omics understanding of trait evolution with ScRAPdb. (**A**) The overall expression profile of the *SSU1* (*YPL092W*) gene in the 131 *S. cerevisiae* strains from the ScRAPdb strain panel with matched transcriptome data. (**B**) The expression levels of the *SSU1* (*YPL092W*) gene across different ScRAPdb phylogenetic clades. (**C**) The genome synteny comparison between SGDref and YJM981 (ADI). (**D**) The expression levels of the *GAL1* (*YBR020W*) and *GAL3* (*YDR009W*) genes across different phylogenetic clades defined in ScRAPdb. (**E**) The growth phenotype measured by yield of ScRAPdb across different ScRAPdb phylogenetic clades, with the French dairy clade highlighted.

regulatory machineries. In addition to actively seeking the opportunities to incorporate newly available population-level omics data on *S. cerevisiae*, we also welcome colleagues in the yeast community and beyond to contact us when such data are generated in their labs, and we are committed to host the data for public sharing.

Data availability

All data covered by ScRAPdb can be freely accessed and downloaded at https://www.evomicslab.org/db/ScRAPdb/. The genome sequences and annotation data for batch downloading are available in the Zenodo repository at https://doi.org/10.5281/zenodo.12580380.

Acknowledgements

We thank the valuable comments and suggestions from two anonymous reviewers, which helped to significantly improve the quality of this manuscript and the associated database. We appreciate the open-science culture of the yeast community for timely depositing many of these long-read yeast genome assemblies to the public domain. We thank the help of many researchers on communicating on and sharing the sequences and metadata of their long-read yeast genome assemblies to us. Special thanks to Dr Isheng Jason Tsai, Dr Kristoffer Krogerus, Dr Meru J. Sadhu, Dr Joshua Bloom, Dr Francisco A. Cubillos, Dr Kenneth H. Wolfe, Dr Anthony D. Long, Dr Stefano Campanaro, Dr Casey M. Bergman, Dr Douda Bensasson, Dr Lydia R. Heasley, Dr J.J. Emerson, Dr Xinwen Zhang, Dr Samina Naseeb, Dr Tomas Peña, Dr Chen-Guang Liu, and Dr Min-Seung Jeon (we apologize for any potential omission) for their expedite and informative response to our queries on data curation. We thank the valuable feedback from Dr Nicolò Tellini when trying out the developmental version of ScRAPdb. We thank the facility support from the Single-Molecule Sequencing Platform at Sun Yat-sen University Cancer Center.

Author contributions: Z.M.: Methodology, software, validation, formal analysis, investigation, data curation, visualization, writing—original draft, writing—review and editing. Y.R.: Methodology, validation, formal analysis, investigation, data curation, writing—original draft, writing—review and editing. A.T.: Methodology, software, validation, formal analysis, investigation, data curation, writing—original draft, writing—review and editing. L.Y.: Formal

analysis, investigation, visualization, data curation, writing—original draft, writing—review and editing. H.L.: Formal analysis, investigation, visualization, writing—review and editing. C.Y.: Methodology, software, visualization, writing—review and editing. G.L.: Data curation, writing—review and editing. G.F.: Data curation, writing—review and editing. J.L.: Validation, formal analysis, investigation, resources, data curation; writing—review and editing, funding acquisition. J.-X.Y.: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft, writing—review and editing, supervision, project administration, funding acquisition.

Funding

National Natural Science Foundation of China [32070592 to J.-X.Y., 32000395 to J.L.]; Guangdong Basic and Applied Basic Research Foundation [2022A1515010717 to J.-X.Y. and 2022A1515011873 to J.L.]; Guangdong Pearl River Talents Program [2019QN01Y183 to J.-X.Y., 2021QN02Y168 to J.L.]; Sun Yat-sen University Cancer Center [YTP-SYSUCC-0042 to J.-X.Y. and YTP-SYSUCC-0040 to J.L.]; Fundamental Research Funds for the Central Universities [24qnpy293 to J.-X.Y.]. Funding for open access charge: National Natural Science Foundation of China.

Conflict of interest statement

None declared.

References

1. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M., *et al.* (1996) Life with 6000 genes. *Science*, **274**, 546–567.
2. Giaever,G., Chu,A.M., Ni,L., Connelly,C., Riles,L., Véronneau,S., Dow,S., Lucau-Danila,A., Anderson,K., André,B., *et al.* (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, **418**, 387–391.
3. Sopko,R., Huang,D., Preston,N., Chua,G., Papp,B., Kafadar,K., Snyder,M., Oliver,S.G., Cyert,M., Hughes,T.R., *et al.* (2006) Mapping pathways and phenotypes by systematic gene overexpression. *Mol. Cell*, **21**, 319–330.
4. Huh,W.-K., Falvo,J.V., Gerke,L.C., Carroll,A.S., Howson,R.W., Weissman,J.S. and O'Shea,E.K. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.
5. Costanzo,M., VanderSluis,B., Koch,E.N., Baryshnikova,A., Pons,C., Tan,G., Wang,W., Usaj,M., Hanchard,J., Lee,S.D., *et al.* (2016) A global genetic interaction network maps a wiring diagram of cellular function. *Science*, **353**, aaf1420.
6. Cherry,J.M., Adler,C., Ball,C., Chervitz,S.A., Dwight,S.S., Hester,E.T., Jia,Y., Juvik,G., Roe,T., Schroeder,M., *et al.* (1998) SGD: saccharomyces genome database. *Nucleic Acids Res.*, **26**, 73–79.
7. Wong,E.D., Miyasato,S.R., Aleksander,S., Karra,K., Nash,R.S., Skrzypek,M.S., Weng,S., Engel,S.R. and Cherry,J.M. (2023) Saccharomyces genome database update: server architecture, pan-genome nomenclature, and external resources. *Genetics*, **224**, iyac191.
8. Liti,G., Carter,D.M., Moses,A.M., Warringer,J., Parts,L., James,S.A., Davey,R.P., Roberts,I.N., Burt,A., Koufopanou,V., *et al.* (2009) Population genomics of domestic and wild yeasts. *Nature*, **458**, 337–341.
9. Bergström,A., Simpson,J.T., Salinas,F., Barré,B., Parts,L., Zia,A., Nguyen Ba,A.N., Moses,A.M., Louis,E.J., Mustonen,V., *et al.* (2014) A high-definition view of functional genetic variation from natural yeast genomes. *Mol. Biol. Evol.*, **31**, 872–888.
10. Scannell,D.R., Zill,O.A., Rokas,A., Payen,C., Dunham,M.J., Eisen,M.B., Rine,J., Johnston,M. and Hittinger,C.T. (2011) The awesome power of yeast evolutionary genetics: new genome sequences and strain resources for the *Saccharomyces sensu stricto Genus*. *G3 (Bethesda)*, **1**, 11–25.
11. Byrne,K.P. and Wolfe,K.H. (2005) The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.*, **15**, 1456–1461.
12. Yue,J.-X., Li,J., Aigrain,L., Hallin,J., Persson,K., Oliver,K., Bergström,A., Coupland,P., Warringer,J., Lagomarsino,M.C., *et al.* (2017) Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat. Genet.*, **49**, 913–924.
13. Peter,J., De Chiara,M., Friedrich,A., Yue,J.-X., Pflieger,D., Bergström,A., Sigwalt,A., Barre,B., Freel,K., Llored,A., *et al.* (2018) Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature*, **556**, 339–344.
14. Duan,S.-F., Han,P.-J., Wang,Q.-M., Liu,W.-Q., Shi,J.-Y., Li,K., Zhang,X.-L. and Bai,F.-Y. (2018) The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. *Nat. Commun.*, **9**, 2690.
15. Wang,Q.-M., Liu,W.-Q., Liti,G., Wang,S.-A. and Bai,F.-Y. (2012) Surprisingly diverged populations of Saccharomyces cerevisiae in natural environments remote from human activity. *Mol. Ecol.*, **21**, 5404–5417.
16. Strope,P.K., Skelly,D.A., Kozmin,S.G., Mahadevan,G., Stone,E.A., Magwene,P.M., Dietrich,F.S. and McCusker,J.H. (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.
17. Burton,J.N., Adey,A., Patwardhan,R.P., Qiu,R., Kitzman,J.O. and Shendure,J. (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.*, **31**, 1119–1125.
18. O'Donnell,S., Yue,J.-X., Saada,O.A., Agier,N., Caradec,C., Cokelaer,T., De Chiara,M., Delmas,S., Dutreux,F., Fournier,T., *et al.* (2023) Telomere-to-telomere assemblies of 142 strains characterize the genome structural landscape in Saccharomyces cerevisiae. *Nat. Genet.*, **55**, 1390–1399.
19. Eberlein,C., Hénault,M., Fijarczyk,A., Charron,G., Bouvier,M., Kohn,L.M., Anderson,J.B. and Landry,C.R. (2019) Hybridization is a recurrent evolutionary stimulus in wild yeast speciation. *Nat. Commun.*, **10**, 923.
20. Lee,T.J., Liu,Y., Liu,W.-A., Lin,Y.-F., Lee,H.-H., Ke,H.-M., Huang,J.-P., Lu,M.-Y.J., Hsieh,C.-L., Chung,K.-F., *et al.* (2022) Extensive sampling of Saccharomyces cerevisiae in Taiwan reveals ecology and evolution of predomesticated lineages. *Genome Res.*, **32**,864–877.
21. Weller,C.A., Andreev,I., Chambers,M.J., Park,M., Program,N.C.S., Bloom,J.S., Sadhu,M.J., Barnabas,B.B., Black,S., Bouffard,G.G., *et al.* (2023) Highly complete long-read genomes reveal pangenomic variation underlying yeast phenotypic diversity. *Genome Res.*, **33**, 729–740.
22. Basile,A., De Pascale,F., Bianca,F., Rossi,A., Frizzarin,M., De Bernardini,N., Bosaro,M., Baldisseri,A., Antoniali,P., Lopreiato,R., *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.
23. Czaja,W., Bensasson,D., Ahn,H.W., Garfinkel,D.J. and Bergman,C.M. (2020) Evolution of Ty1 copy number control in yeast by horizontal transfer and recombination. *PLos Genet.*, **16**, e1008632.
24. Preiss,R., Fletcher,E., Garshol,L.M., Foster,B., Ozsahin,E., Lubberts,M., van der Merwe,G. and Krogerus,K. (2024) European farmhouse brewing yeasts form a distinct genetic group. *Appl. Microbiol. Biotechnol.*, **108**, 430.

25. Linder,R.A., Majumder,A., Chakraborty,M. and Long,A. (2020) Two synthetic 18-way outcrossed populations of diploid budding yeast with utility for complex trait dissection. *Genetics*, **215**, 323–342.
26. Peña,T.A., Villarreal,P., Agier,N., Chiara,M.D., Barría,T., Urbina,K., Villarroel,C.A., Santos,A.R.O., Rosa,C.A., Nespolo,R.F., *et al.* (2024) An integrative taxonomy approach reveals *Saccharomyces chiloensis* sp. nov. as a newly discovered species from Coastal Patagonia. *PLos Genet.*, **20**, e1011396.
27. Naseeb,S., Alsammar,H., Burgis,T., Donaldson,I., Knyazev,N., Knight,C. and Delneri,D. (2018) Whole genome sequencing, *de novo* assembly and phenotypic profiling for the new budding yeast species *Saccharomyces jurei*. *G3 (Bethesda)*, **8**, 2967–2977.
28. Mardones,W., Villarroel,C.A., Krogerus,K., Tapia,S.M., Urbina,K., Oporto,C.I., O'Donnell,S., Minebois,R., Nespolo,R., Fischer,G., *et al.* (2020) Molecular profiling of beer wort fermentation diversity across natural *Saccharomyces eubayanus* isolates. *Microb. Biotechnol.*, **13**, 1012–1025.
29. Bergin,S.A., Allen,S., Hession,C., Ó Cinnéide,E., Ryan,A., Byrne,K.P., Ó Cróinín,T., Wolfe,K.H. and Butler,G. (2022) Identification of European isolates of the lager yeast parent *Saccharomyces eubayanus*. *FEMS Yeast Res.*, **22**, foac053.
30. Chen,J., Garfinkel,D.J. and Bergman,C.M. (2023) Horizontal transfer and recombination fuel Ty4 retrotransposon evolution in Saccharomyces. bioRxiv doi: https://doi.org/10.1101/2023.12.20.572574, 20 December 2023, preprint: not peer reviewed.
31. Spealman,P., Avecilla,G., Matthews,J., Suresh,I. and Gresham,D. (2022) Complex genomic rearrangements following selection in a glutamine-limited medium over hundreds of generations. *Microbiol. Resour. Announc.*, **11**, e00729-22.
32. De Chiara,M., Barré,B.P., Persson,K., Irizar,A., Vischioni,C., Khaiwal,S., Stenberg,S., Amadi,O.C., Žun,G., Doberšek,K., *et al.* (2022) Domestication reprogrammed the budding yeast life cycle. *Nat. Ecol. Evol.*, **6**, 448–460.
33. Caudal,É., Loegler,V., Dutreux,F., Vakirlis,N., Teyssonnière,É., Caradec,C., Friedrich,A., Hou,J. and Schacherer,J. (2024) Pan-transcriptome reveals a large accessory genome contribution to gene expression variation in yeast. *Nat. Genet.*, **56**, 1278–1287.
34. Teyssonnière,E.M., Trébulle,P., Muenzner,J., Loegler,V., Ludwig,D., Amari,F., Mülleder,M., Friedrich,A., Hou,J., Ralser,M., *et al.* (2024) Species-wide quantitative transcriptomes and proteomes reveal distinct genetic control of gene expression variation in yeast. *Proc. Natl Acad. Sci. USA*, **121**, e2319211121.
35. Muenzner,J., Trébulle,P., Agostini,F., Zauber,H., Messner,C.B., Steger,M., Kilian,C., Lau,K., Barthel,N., Lehmann,A., *et al.* (2024) Natural proteome diversity links aneuploidy tolerance to protein turnover. *Nature*, **630**, 149–157.
36. Naseeb,S., James,S.A., Alsammar,H., Michaels,C.J., Gini,B., Nueno-Palop,C., Bond,C.J., McGhie,H., Roberts,I.N. and Delneri,D. (2017) *Saccharomyces jurei* sp. nov., isolation and genetic identification of a novel yeast species from Quercus robur. *Int. J. Syst. Evol. Microbiol.*, **67**, 2046–2052.
37. Sayers,E.W., Cavanaugh,M., Clark,K., Ostell,J., Pruitt,K.D. and Karsch-Mizrachi,I. (2020) GenBank. *Nucleic Acids Res.*, **48**, D84–D86.
38. Baxevanis,A.D. (2006) Searching the NCBI Databases using Entrez. *Curr. Protoc. Hum. Genet.*, **51**, 6.10.1–6.10.24.
39. Lin,J. (2009) Is searching full text more effective than searching abstracts?*BMC Bioinf.*, **10**, 46.
40. Kim,W., Yeganova,L., Comeau,D.C., Wilbur,W.J. and Lu,Z. (2022) Towards a unified search: improving PubMed retrieval with full text. *J. Biomed. Inform.*, **134**, 104211.
41. Li,J., Llorente,B., Liti,G. and Yue,J.-X. (2022) RecombineX: a generalized computational framework for automatic high-throughput gamete genotyping and tetrad-based recombination analysis. *PLos Genet.*, **18**, e1010047.
42. Yue,J.-X. and Liti,G. (2018) Long-read sequencing data analysis for yeasts. *Nat. Protoc.*, **13**, 1213–1231.
43. Manni,M., Berkeley,M.R., Seppey,M., Simão,F.A. and Zdobnov,E.M. (2021) BUSCO Update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.*, **38**, 4647–4654.
44. Challis,R., Richards,E., Rajan,J., Cochrane,G. and Blaxter,M. (2020) BlobToolKit – interactive quality assessment of genome assemblies. *G3 (Bethesda)*, **10**, 1361–1374.
45. Lee,I., Ouk Kim,Y., Park,S.-C. and Chun,J. (2016) OrthoANI: an improved algorithm and software for calculating average nucleotide identity. *Int. J. Syst. Evol. Microbiol.*, **66**, 1100–1103.
46. Ebert,P., Audano,P.A., Zhu,Q., Rodriguez-Martin,B., Porubsky,D., Bonder,M.J., Sulovari,A., Ebler,J., Zhou,W., Serra Mari,R., *et al.* (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, **372**, eabf7117.
47. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R.S., Thormann,A., Flicek,P. and Cunningham,F. (2016) The Ensembl Variant Effect Predictor. *Genome Biol.*, **17**, 122.
48. Klemm,P., Stadler,P.F. and Lechner,M. (2023) Proteinortho6: pseudo-reciprocal best alignment heuristic for graph-based detection of (co-)orthologs. *Front. Bioinform.*, **3**, 1322477.
49. Ranwez,V., Douzery,E.J.P., Cambon,C., Chantret,N. and Delsuc,F. (2018) MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol. Biol. Evol.*, **35**, 2582–2584.
50. Nguyen,L.-T., Schmidt,H.A., Haeseler,A. and Minh,B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.
51. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
52. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
53. Demichev,V., Messner,C.B., Vernardis,S.I., Lilley,K.S. and Ralser,M. (2020) DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods*, **17**, 41–44.
54. Messner,C.B., Demichev,V., Bloomfield,N., Yu,J.S.L., White,M., Kreidl,M., Egger,A.-S., Freiwald,A., Ivosev,G., Wasim,F., *et al.* (2021) Ultra-fast proteomics with Scanning SWATH. *Nat. Biotechnol.*, **39**, 846–854.
55. Li,D., Mei,H., Shen,Y., Su,S., Zhang,W., Wang,J., Zu,M. and Chen,W. (2018) ECharts: a declarative framework for rapid construction of web-based visualization. *Visual Inform.*, **2**, 136–146.
56. Shank,S.D., Weaver,S. and Kosakovsky Pond,S.L. (2018) phylotree.js - a JavaScript library for application development and interactive data visualization in phylogenetics. *BMC Bioinf.*, **19**, 276.
57. Diesh,C., Stevens,G.J., Xie,P., De,J., Martinez,T., Hershberg,E.A., Leung,A., Guo,E., Dider,S., Zhang,J., *et al.* (2023) JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biol.*, **24**, 74.
58. Wei,W., McCusker,J.H., Hyman,R.W., Jones,T., Ning,Y., Cao,Z., Gu,Z., Bruno,D., Miranda,M., Nguyen,M., *et al.* (2007) Genome sequencing and comparative analysis of Saccharomyces cerevisiae strain YJM789. *Proc. Natl Acad. Sci. USA*, **104**, 12825–12830.
59. Park,H. and Bakalinsky,A.T. (2000) SSU1 mediates sulphite efflux in *Saccharomyces cerevisiae*. *Yeast*, **16**, 881–888.
60. Goto-Yamamoto,N., Kitano,K., Shiki,K., Yoshida,Y., Suzuki,T., Iwata,T., Yamane,Y. and Hara,S. (1998) *SSU1-R*, a sulfite resistance gene of wine yeast, is an allele of *SSU1* with a different upstream sequence. *J. Ferment. Bioeng.*, **86**, 427–433.
61. Pérez-Ortín,J.E., Querol,A., Puig,S. and Barrio,E. (2002) Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. *Genome Res.*, **12**, 1533–1539.
62. Zimmer,A., Durand,C., Loira,N., Durrens,P., Sherman,D.J. and 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.
63. García-Ríos,E., Nuévalos,M., Barrio,E., Puig,S. and Guillamón,J.M. (2019) A new chromosomal rearrangement improves the adaptation of wine yeasts to sulfite. *Environ. Microbiol.*, **21**, 1771–1781.
64. Boocock,J., Sadhu,M.J., Durvasula,A., Bloom,J.S. and Kruglyak,L. (2021) Ancient balancing selection maintains incompatible versions of the galactose pathway in yeast. *Science*, **371**, 415–419.
65. Pontes,A., Paraíso,F., Liu,Y.-C., Limtong,S., Jindamorakot,S., Jespersen,L., Gonçalves,C., Rosa,C.A., Tsai,I.J., Rokas,A., *et al.* (2024) Tracking alternative versions of the galactose gene network in the genus *Saccharomyces* and their expansion after domestication. *iScience*, **27**, 108987.
66. Loegler,V., Friedrich,A. and Schacherer,J. (2024) Overview of the *Saccharomyces cerevisiae* population structure through the lens of 3,034 genomes. bioRxiv doi: https://doi.org/10.1101/2024.09.16.613241, 20 September 2024, preprint: not peer reviewed.
67. McGovern,P.E., Zhang,J., Tang,J., Zhang,Z., Hall,G.R., Moreau,R.A., Nuñez,A., Butrym,E.D., Richards,M.P., Wang,C., *et al.* (2004) Fermented beverages of pre- and proto-historic China. *Proc. Natl Acad. Sci. USA*, **101**, 17593–17598.

Received: August 14, 2024. **Revised:** October 5, 2024. **Editorial Decision:** October 8, 2024. **Accepted:** October 10, 2024