GENETICS

Graph-based pangenome analysis uncovers structural and functional impacts of allopolyploidization events

Victor Loegler[1], Anne Friedrich[2], Joseph Schacherer[1,2,3,4]*

Copyright © 2026 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution License 4.0 (CC BY).

The emergence of graph pangenomes has advanced the study of structural variants (SVs) across diverse species, yet the impact of complex ploidy on SV landscapes remains poorly explored. *Brettanomyces bruxellensis*, a yeast shaped by multiple allopolyploidization events, offers a compelling model due to its well-characterized genomic and phenotypic diversity. Here, we integrated subgenome-resolved assemblies with population-scale genomic and transcriptomic datasets to characterize SV distribution and functional impact at the species level. Leveraging a reference graph pangenome, we detected 5972 unique SVs across 1060 isolates. Allopolyploid genomes harbored significantly more SVs (320 on average) than diploids (116), reflecting hybridization's extensive impact on genome architecture. Functional association analysis revealed that 27.5% of SVs locally influence gene expression. Furthermore, a subset of SVs originating from ancestral divergence between subgenomes (hybridization-associated SVs) formed hotspots of structural diversity and disproportionately affect gene expression, with larger effect sizes than other SVs. These findings underscore allopolyploidization as a driver of structural and functional genomic diversification, highlighting the value of graph-based approaches in dissecting complex genome evolution.

INTRODUCTION

Polyploidization, defined as the duplication of whole genomes, is a major evolutionary force across eukaryotes, contributing to speciation, adaptation, and genomic innovation (*1*–*3*). One specific form, allopolyploidy, originates from the hybridization of related species, thereby introducing immediate genetic diversity by uniting divergent genomic copies inherited from distinct parental lineages within a single nucleus (*4*). In the resulting allopolyploid, the set of sequences that share a parental origin is called a subgenome. The novel allelic combinations and structural rearrangements introduced by allopolyploidization may drive phenotypic evolution and ecological adaptation (*5*–*7*).

The consequences of polyploidization have been widely explored through the analysis of small genomic variants or gene sequences (*8*, *9*). Its impact on larger structural variants (SVs; ≥50 bp), however, remains overlooked, partly due to the technical challenges associated with SV detection in polyploid genomes (*10*). Yet, the established functional impact of SVs in diploids indicates a potential for SVs to play a primary role in the diversification of polyploid species (*11*–*13*).

Recent advancements in long-read sequencing, haplotype-resolved assemblies, and graph-based pangenomes now offer powerful tools to identify SVs in polyploid genomes (*14*–*20*). These approaches facilitate the clear segregation of haplotypes (the individual genomic copies present within a nucleus) and subgenomes (sets of haplotypes sharing a common parental origin), which is essential to track the structural consequences of allopolyploidization.

The budding yeast *Brettanomyces bruxellensis* offers a valuable model to study allopolyploid genome evolution. Throughout its evolutionary trajectory, the species has undergone at least three independent allopolyploidization events, each originating from the hybridization of the primary lineage with a distinct and currently unidentified parental lineage (*21*, *22*). The resulting natural population comprises five genetic clusters (*23*). Two of these are diploid (D1 and D2) and carry only the primary subgenome, which is shared by all clusters. The remaining three are allotriploid (A1, A2, and A3): They retain two copies of the primary subgenome and harbor an additional copy of a cluster-specific acquired subgenome, each of which was introduced by an independent hybridization event. Although the species has been observed across a range of human-related environments including beer and kombucha brewing and wine spoilage (*24*–*28*), available collections are dominated by wine-associated isolates (*23*). Substantial resources are available for this species, including 1060 short-read sequenced isolates (*23*), 71 long-read sequenced isolates (*22*), and transcriptomic data for 87 isolates (*29*).

Here, we leverage these resources to investigate the structural and functional consequences of allopolyploidization. Using a newly constructed graph pangenome based on subgenome-resolved assemblies, in which each isolate is represented by one consensus sequence per subgenome, we conducted a comprehensive analysis of the SV landscape in this species. We identified structural differences between subgenomes, revealed their concentration in specific genomic hotspots, and demonstrated their substantial impact on gene expression.

RESULTS

A reference graph pangenome captures subgenome diversity

In a previous study, 71 long read genome assemblies with resolved subgenomes were generated, capturing the diversity of both diploid (D1 and D2) and allopolyploid (A1, A2, and A3) genetic clusters (*22*) (Fig. 1A and fig. S1). This collection of genomes, which includes 40 allopolyploid genomes, has been used to construct a reference pangenome graph. The Minigraph-Cactus algorithm (*15*) was used to build the graph, starting from the primary linear reference genome (*30*). The resulting graph exhibits a total length of 26,926,529 bp, distributed across 5,464,509 segments (nodes), each representing a DNA sequence shared by one or more genome assemblies. To assess the completeness of the graph, a rarefaction curve of the pangenome has

[1]CIC bioGUNE, Bizkaia Technological Park, Building 800, 48160 Derio, Spain. [2]Université de Strasbourg, CNRS, Inserm, IGBMC UMR 7104-UMR-S 1258, 67400 Illkirch, France. [3]Institut Universitaire de France (IUF), 75231 Paris, France. [4]Ikerbasque, Basque Foundation for Science, Bilbao, Spain.
*Corresponding author. Email: jschacherer@cicbiogune.es

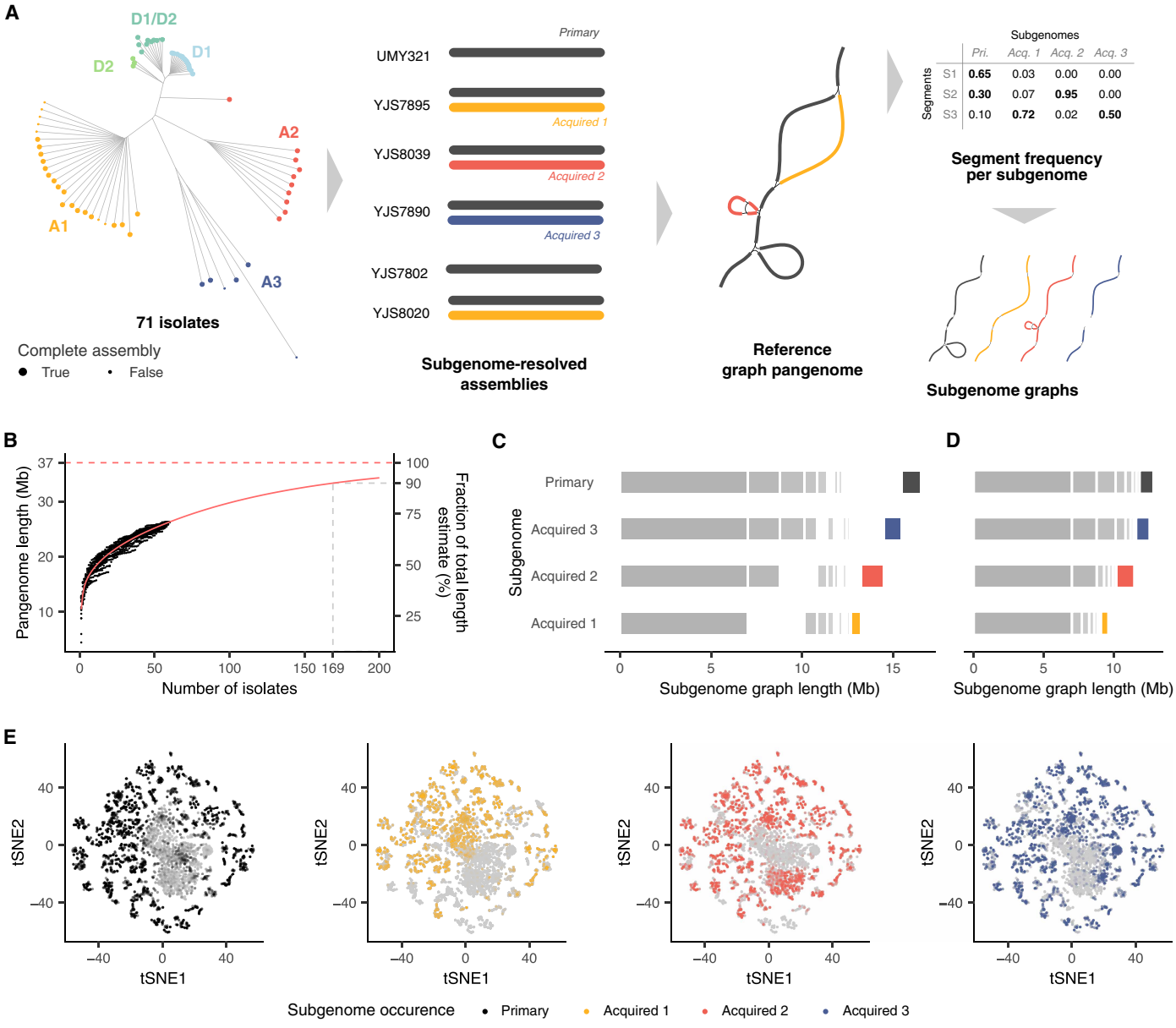


Fig. 1. Reference graph pangenome. (**A**) Neighbor-joining tree of the 71 isolates used for graph construction (*22*) and method for subgenome graph identification. The neighbor-joining tree was built from whole genome sequencing data. Colors indicate the genetic clusters assigned to each genome assembly in (*22*). (**B**) Rarefaction curve of the graph pangenome. The points represent observed rarefaction, the red curve correspond to the nonparametric rarefaction model and the red dashed line represents the estimated total pangenome length. (**C**) Shared length of subgenome graphs. (**D**) Total length of subgenome graphs. For (C) and (D), light gray rectangles correspond to sequences of the graph shared by multiple subgenomes. Colored and dark gray rectangles correspond to subgenome-specific sequences. (**E**) TSNE projection of segment occurrence across isolates. Segment occurrence within subgenome is color coded. Gray points correspond to segments absent from the subgenome of interest.

been modeled using nonparametric asymptotic estimators. The total estimated length was determined to be 37.0 Mb, of which 72.8% is captured by our graph (Fig. 1B). According to the rarefaction model, a total of 169 isolates would be required to reach 90% of the species pangenome length (table S1).

To identify subgenomic patterns within the graph, we computed the frequency of occurrence of each segment within the primary and all three acquired subgenomes (Fig. 1A). Four subgenome-specific graphs were extracted by retaining segments present in at least 20% of the assemblies assigned to the corresponding subgenome. The total length of subgenome-specific graphs ranges from 9.63, 11.49, and 12.58 Mb for acquired subgenomes 1, 2 and 3, respectively, up to 12.86 Mb for the primary subgenome (Fig. 1C). These results are consistent with the extent of acquired-to-primary subgenome conversion previously observed in each allopolyploid genetic cluster (*22*), a form of loss of heterozygosity in which acquired subgenome segments are replaced by primary subgenome homologous sequences, thereby reducing the effective length of acquired subgenome sequences (fig. S2). The A1 subgenome, which has undergone the most extensive conversion (*22*), is also the shortest and contributes the least nonreference

material. A total of 7.02 Mb is shared between all subgenome graphs, and 0.55 to 1.3 Mb is private to each graph (Fig. 1C, fig. S3, and table S2). A tSNE projection based on segment occurrence per isolate reveals clear segregation of subgenome-specific segments (Fig. 1D), revealing the complex genomic entanglement of the four subgenomes within the species.

Primary subgenome drives long-term pangenome expansion

We quantified the length of nonreference sequences brought by each genome and subgenome. On average, genomes contain 853 kb of nonreference sequences. This number greatly varies between strains, with diploid isolates bringing 314 kb nonreference sequences on average, while it goes up to an average of 1.3 Mb for allopolyploid isolates (Fig. 2A and table S3). The acquired subgenomes bring a larger amount of nonreference sequence to the graph compared to the primary subgenomes (785 versus 467 kb on average, two-sided rank-based Wilcoxon test, $P = 1.65 \times 10^{-4}$) (Fig. 2B and table S3), which is consistent with their higher divergence from the reference genome.

However, this early expansion rapidly reaches saturation in acquired subgenomes. Across all subgenomes in the population, 5, 7, and 2 isolates are required to reach 80% saturation of the A1, A2, and A3 pan-subgenomes, respectively, while 95 are required to reach this completion level for the primary pan-subgenome (Fig. 2C and table S1). This marked contrast can be attributed, in part, to the presence of the primary subgenome across all genetic clusters, while the acquired subgenomes are cluster specific. Nonetheless, even within allopolyploid clusters, more isolates are needed to reach saturation in the primary than in the acquired subgenome graphs. In the A1 cluster, 14 isolates allow to reach 80% of the primary genome estimated size, while this saturation is achieved with only 5 isolates for the acquired genome (fig. S4A). Similar results are observed in the A2 cluster, where 10 and 8 isolates are required, respectively (fig. S4B). These numbers reveal a lower diversity within the acquired genomes, hence a greater expansion of the primary genome as the number of isolates increases. This finding is coherent with previous observation of a lower level of diversity in the acquired subgenome in comparison to the primary one (*23*) and highlights the role of the primary subgenome in the long-term pangenome expansion.

Species-wide SV mapping through a graph pangenome

The construction of a graph pangenome for *B. bruxellensis* has enabled the first analysis of the SV landscape of the species. A total of 212,177 redundant SVs were detected across a collection of 1060 isolates (*23*) by mapping sequencing reads to the graph. This corresponds to 5972 unique SV events (≥50 bp), including 1385 insertions, 939 deletions, and 3648 complex SVs (Methods). The length of SVs extends up to 18 kb, with an average length of 345 bp (fig. S5). A substantial sequence redundancy is observed across SVs, with 655 insertions (47.3%) showing more than 90% sequence similarity with another (Fig. 3A). As such redundancy is often linked to the presence

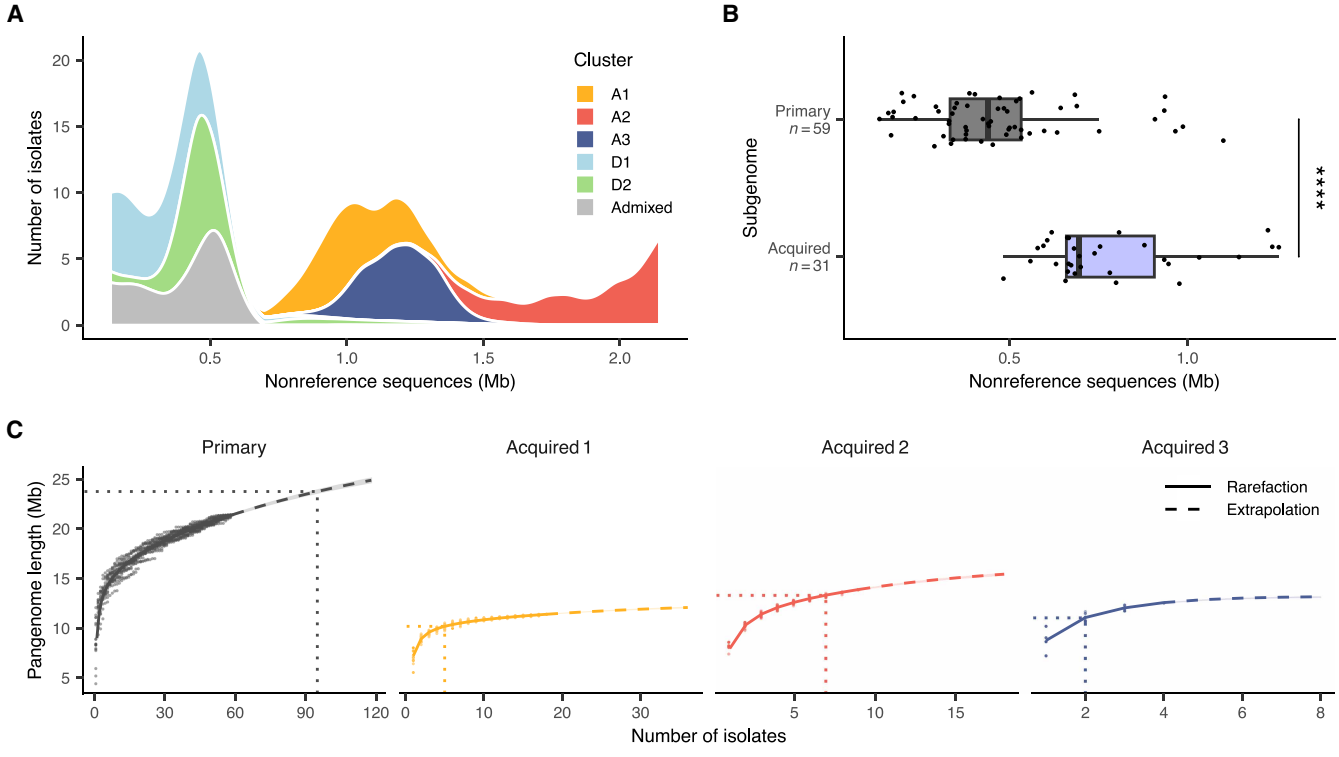


Fig. 2. Pangenome expansion. (**A**) Distribution of the amount of nonreference sequences found per genome colored by genetic cluster. (**B**) Amount of nonreference sequences found per subgenome. The middle bar of the box plots corresponds to the median; the upper and lower bounds correspond to the third and first quartiles, respectively. The whiskers correspond to the upper and lower bounds 1.5 times the interquartile range (IQR). The *P* value was computed using a two-sided rank-based Wilcoxon test. ****$P = 5.7 \times 10^{-9}$. (**C**) Rarefaction curve per subgenome. The points represent observed rarefaction, and the curves correspond to the nonparametric rarefaction model. The dotted lines indicate the pangenome length and number of isolates required to reach 80% of the pangenome saturation.

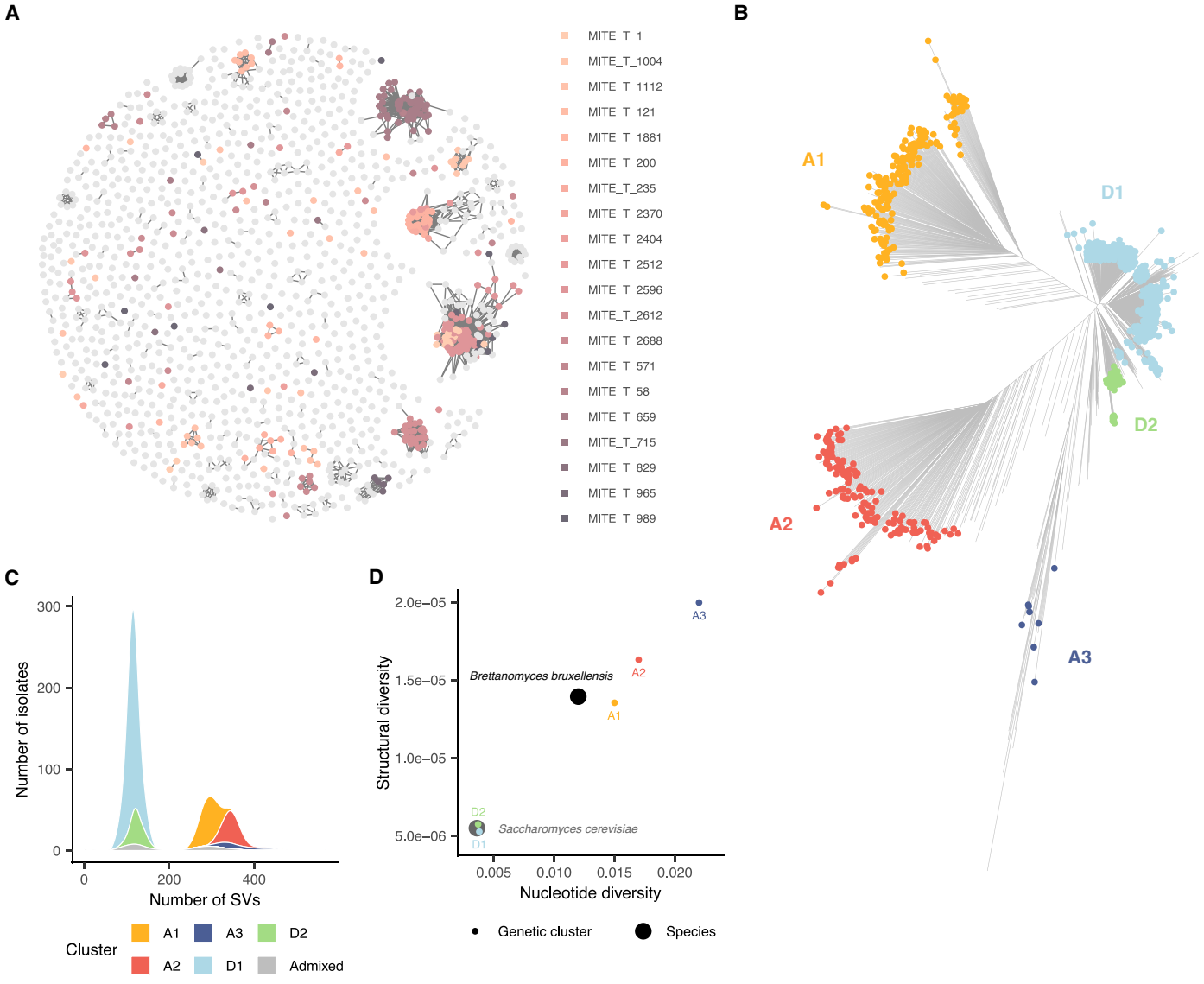


Fig. 3. Structural variant landscape. (**A**) Sequence similarity network of insertions. Each node represents one insertion, and edges indicate sequence identity ≥90% and ≥80% mutual coverage. MITE-related insertions are color coded. (**B**) Neighbor-joining tree based on 5972 nonredundant SVs. The SNP-based genetic cluster is color coded. (**C**) Distribution of the number of SVs per isolate, colored by genetic cluster. (**D**) Nucleotide and structural diversity of large collections of isolates for *Saccharomyces cerevisiae*, *Brettanomyces bruxellensis*, and genetic clusters of *B. bruxellensis*.

of transposable elements, we looked for the presence of miniature inverted-repeat transposable elements (MITEs) across all genomes. One MITE element has been previously reported in *B. bruxellensis* (*31*). Across the 71 genomes investigated, a total of 23 MITE families has been identified, with size ranging from 388 to 800 bp. Subsequent comparison with SV sequences enabled the characterization of 1092 MITE-related SVs, constituting 18.3% of the total number of SVs detected. The presence of MITE-related SVs explained nearly all the sequence redundancy observed within insertion sequences, suggesting that no other major transposable elements are related to SVs in the species (Fig. 3A). While many MITE families are shared across genetic clusters and subgenomes, we note a few exceptions: MITE families 4 and 5 are restricted to the primary subgenome in all clusters, and families 1, 9, 18, and 21 are shared between subgenomes but almost exclusive to a single genetic cluster (fig. S6).

Allopolyploidy enhances structural variation in the population

The distribution of SVs in a population generally mirrors the pattern observed for smaller genomic variants (*32*–*34*). To verify this statement in our population, an SV-based neighbor-joining tree was constructed and further compared with the genetic cluster previously established on an SNP-based analysis (*23*). The obtained phylogeny is in complete accordance with the genetic clusters (Fig. 3B). Furthermore, as with small variants, the number of SVs found in each isolate deeply varies between nonallopolyploid and allopolyploid states. While the former exhibits an average number of 116 SVs, the latter has an average number of 320 SVs (Fig. 3C) (two-sided ranked-based Wilcoxon test, $P < 2.2 \times 10^{-16}$). In addition, the average proportion of heterozygous SVs is 0.70 for allopolyploid isolates in comparison to 0.46 for nonallopolyploid isolates (fig. S7).

This represents a significant difference (two-sided ranked-based Wilcoxon test, $P < 2.2 \times 10^{-16}$), as expected by the inclusion of an additional distant haplotype in allopolyploid genomes.

To compare the level of structural diversity with a nonallopolyploid yeast species, we took advantage of a reference graph pangenome constructed using 500 haplotypes of *Saccharomyces cerevisiae* (*35*) and used the same genotyping strategy within a collection of 3034 natural isolates (*36*). The structural diversity, measured as the per-base average number of SVs between pairs of sequences (Methods), had a value of 1.39×10^{-5} in the *B. bruxellensis* population. The *S. cerevisiae* population, despite containing threefold more isolates, only exhibited a structural diversity of 5.49×10^{-6}, which is 0.4-fold that found in *B. bruxellensis*. This mirrors the observed difference in nucleotide diversity: 1.2×10^{-2} for *B. bruxellensis* (*23*) and 3.6×10^{-3} for *S. cerevisiae* (*36*) (Fig. 3D). The structural diversity was further computed within each genetic cluster. The A1, A2, A3, D1, and D2 clusters exhibit a structural diversity of 13.5, 16.2, 19.5, 5.0, and 5.6×10^{-6}, respectively (Fig. 3D). As for the nucleotide diversity (*23*), the structural diversity of allopolyploid cluster is higher than that of diploid clusters (two-sided ranked-based Wilcoxon test, $P < 2.2 \times 10^{-16}$). The level of structural diversity of nonallopolyploid clusters is of the same order of magnitude than the structural diversity of *S. cerevisiae* (Fig. 3D, fig. S8, and table S4), indicating allopolyploidization as a main driver of structural diversity.

Hotspots of SVs arise from ancestral subgenome divergence

By adding novel divergent haplotypes into a population, hybridization participates to an increase in structural diversity. SVs previously fixed between the two parental species are now polymorphic and heterozygous in the resulting population. We took advantage of the subgenome-aware graph pangenome to segregate SVs resulting from mutation and propagation in the population, and those gained through hybridization, which correspond to ancestral structural rearrangements between subgenomes. The latter can be identified by their high allele frequency within a specific acquired subgenome (allele frequency – AF > 0.75 among the assemblies of that acquired subgenome) and low frequency in the primary subgenome of the corresponding isolates (AF < 0.05). A subset of 31 allopolyploid isolates with qualitative primary and acquired genome assemblies was considered for this analysis. The distribution of allele frequency demonstrates a clear bias toward high frequencies for SVs located on the acquired genomes (Fig. 4A; two-sided rank-based Wilcoxon test, $P < 2.2 \times 10^{-16}$). Using the previously mentioned frequency thresholds, we identified 278 SVs (4.8%) corresponding to ancestral structural differences between subgenomes that were introduced into the population through hybridization, and which were fixed in the parental lineages before this event. For brevity, these SVs are hereafter referred to as hybridization-associated SVs. Within hybridization-associated SVs, 114, 135, and 35 were observed as ancestral differences between the primary subgenome and the A1, A2, and A3 subgenomes, respectively (i.e., present at high frequency in the corresponding acquired subgenome and nearly absent in the primary subgenome). Only five SVs met this criterion for more than one acquired subgenome, indicating that most hybridization-associated SVs are cluster specific. Despite the limited number of hybridization-associated SVs detected (278 among 5802 SVs), their large frequency within the population and their uneven distribution along the genome result in considerable hotspots of structural diversity, associated with ancestral structural differences between subgenomes (Fig. 4B). Among 21 hotspots identified, six are mainly related to hybridization-associated SVs (table S5). The distribution of these hybridization-associated hotspots differs between allopolyploid clusters, with some being cluster specific and other shared across clusters (fig. S9).

Hybridization-associated SVs disproportionately affect gene expression

Last, we sought to investigate the local effect of SVs on gene expression by leveraging a transcriptomic dataset generated and available for diverse 87 isolates (*29*). A total of 32,590 SV-gene pairs separated by less than 50 kb were considered for association testing, involving 4952 genes and 857 SVs. Using a linear mixed model to prevent the confounding effect of population structure, 436 SV-gene pairs showed

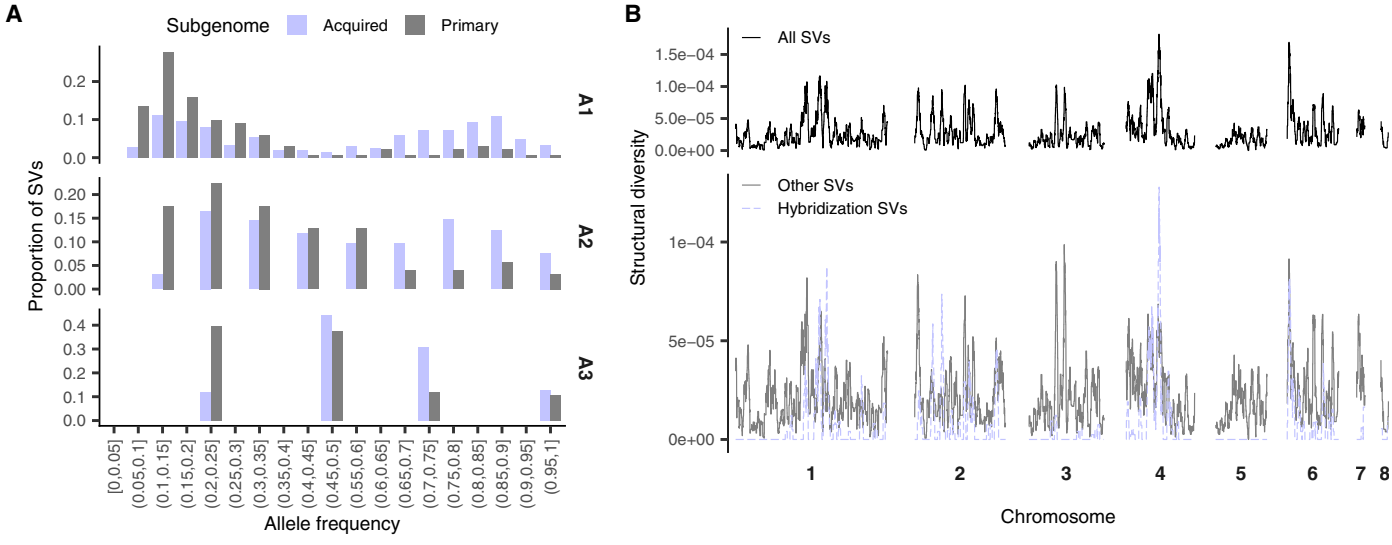


Fig. 4. Distribution of hybridization SVs. (**A**) Distribution of allele frequency per subgenome in the three allopolyploid clusters, using nonsingleton SVs. (**B**) Structural diversity along the genome using all SVs (top track, black line), hybridization-associated SVs (bottom track, purple line), and other SVs (bottom track, gray line).

significant association (Fig. 5A and table S6). Of the 857 SVs tested, 236 (27.5%) have a local effect on gene expression, affecting 298 genes (6.0% of all genes tested). Within associated SVs, there are 132 deletions (36.1% of all deletions tested), 53 insertions (26.5%), and 51 complex SVs (17.5%). Deletions are more frequently associated than insertions (two-sided Fisher's exact test, odds ratio = 1.56, $P = 0.024$) and complex SVs (two-sided Fisher's exact test, odds ratio = 2.65, $P = 1.2 \times 10^{-7}$). Despite a limited sample size ($n = 6$), we observe a tendency for SVs overlapping genes to have a higher effect size on gene expression than SVs located upstream and downstream the gene of interest (Fig. 5B) (two-sided ranked-based Wilcoxon test, $P = 0.021$ and 0.022, respectively).

To further explore the impact of allopolyploidization, hybridization-associated SVs were identified across all SVs tested for association using a competitive mapping strategy across subgenomes (Methods). A total of 40 were identified as hybridization-associated SVs, and 21 (51%) of these were associated with local differential gene expression. This represents a significant enrichment of hybridization-associated SVs (Fig. 5C; odds ratio = 2.92; two-sided Fisher's exact test, $P = 0.00105$). In addition, hybridization-associated SVs exhibit a higher effect size than other associated SVs (0.31 versus 0.24 on average; two-sided rank-based Wilcoxon test, $P = 0.0051$; Fig. 5D). Overall, this suggests that allopolyploidization has a substantial impact on gene expression diversification through the emergence of structural diversity.

DISCUSSION

Polyploidization events have profound consequences on species genomic diversity. In particular, polyploidy can increase structural diversity within populations in two ways. First, increasing the DNA content of cells amplifies the mutational input, as well as the masking of recessive deleterious variants. This is expected to result in a higher diversity in polyploids than in diploids. Consistently, recent studies demonstrated a higher SV load in tetraploid genomes than in diploid genomes in the plant genus *Cochlearia* and *Arabidopsis* (*37*, *38*). Second, in the context of allopolyploidization, the integration of a divergent genomic copy immediately increases structural variation. Structural differences fixed in the genomes of each parental species become suddenly part of

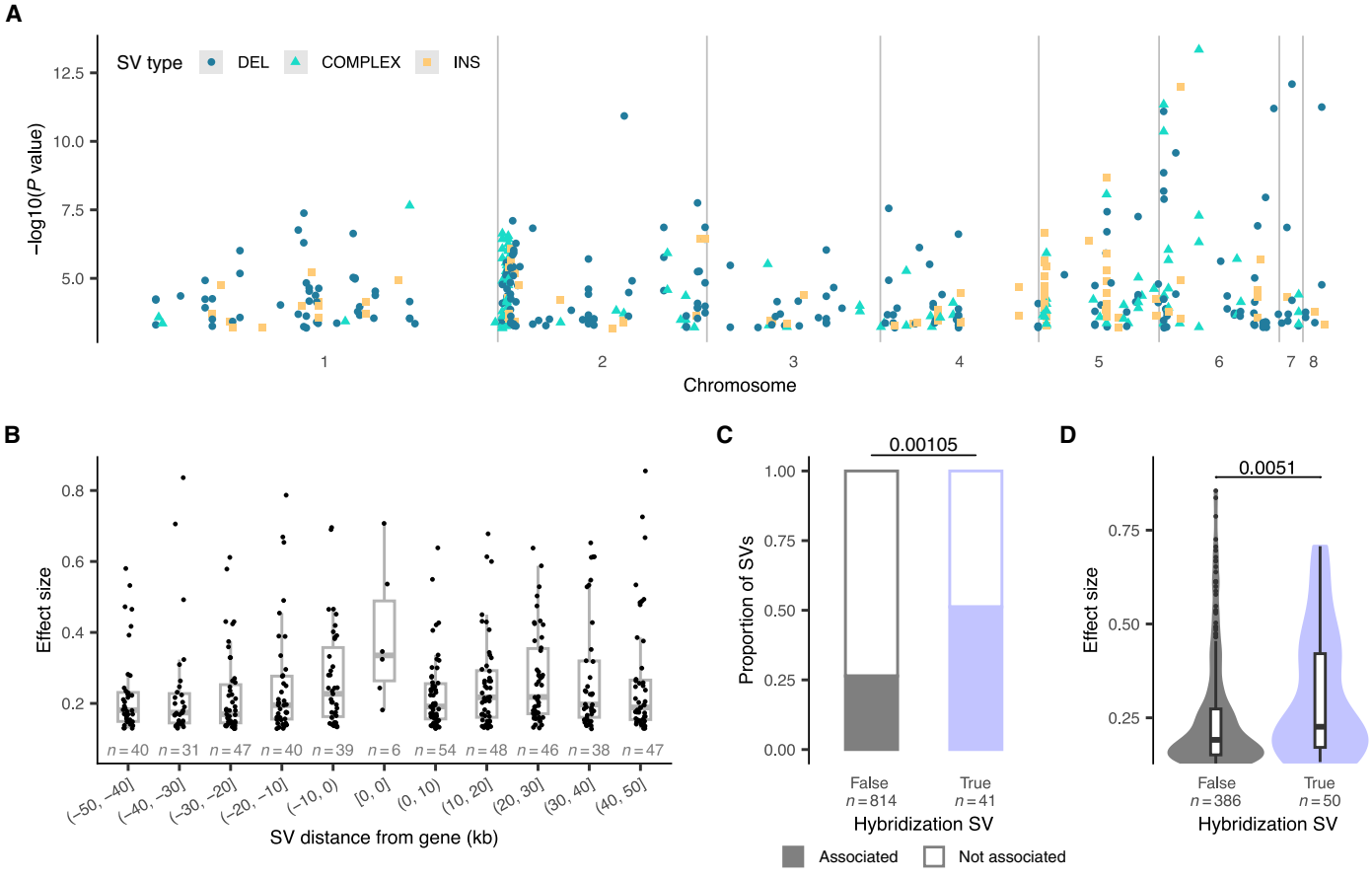


Fig. 5. SV associations with gene expression variation. (**A**) Associated SVs along the genome. The type of SV is color and shape coded. The *x* axis represents the −log10(*P* value) after FDR adjustment. (**B**) Effect size depending on the distance of the SV in relation to the gene whose expression is affected. The middle bar of the box plots corresponds to the median; the upper and lower bounds correspond to the third and first quartiles, respectively. The whiskers correspond to the upper and lower bounds 1.5 times the interquartile range (IQR). (**C**) Proportion of associated SVs among hybridization and other SVs tested for association. The *P* value was computed using a two-sided Fisher's exact test. (**D**) Effect size of hybridization SVs and other SVs. The middle bar of the box plots corresponds to the median; the upper and lower bounds correspond to the third and first quartiles, respectively. The whiskers correspond to the upper and lower bounds 1.5 times the interquartile range (IQR). The *P* value was computed using a two-sided rank-based Wilcoxon test.

the SV repertoire. Here, we demonstrated that these hybridization SVs strongly affect gene expression and may largely contribute to the phenotypic diversification of allopolyploid isolates (*23*).

Beyond the direct introduction of SVs, the three acquired subgenomes have been shaped differentially by acquired-to-primary genome conversion, resulting in cluster-specific differences in their total length and nonreference content. This variation is unlikely to reflect the age of each hybridization event, because the LOH pattern is highly uniform across isolates within a given allopolyploid cluster (*22*), and the primary and acquired subgenomes remain largely collinear and syntenic. These two observations contradict the hypothesis of a gradual, age-dependent erosion of the acquired subgenome. Instead, they suggest that conversion events occurred rapidly after allopolyploidization and were subsequently fixed within each cluster, most likely driven by adaptation.

Graph-based pangenomes played a pivotal role in enabling these findings. Although pangenomes are now routinely performed for diploid species, polyploid pangenomes remain scarce (*19*, *20*, *37*, *39*). Nevertheless, graph pangenomes built with appropriate phased genomes offer powerful tools to explore structural genome evolution in polyploids. For instance, the allopolyploid peanut pangenome revealed subgenome-biased structural diversity (*20*), and the potato pangenome revealed low haplotype diversity attributed to a recent population bottleneck (*19*), similar to that observed in *B. bruxellensis* acquired subgenome.

By generating a reference graph-based pangenome, this work also constitutes a valuable resource for further investigation of SVs in the species. Given the industrial significance of *B. bruxellensis* in beer and wine fermentation (*40*–*42*), our graph may support the characterization and selection of isolates for specific applications.

Despite these advances, major limitations should be acknowledged. First, the subgenome phasing method relies on sequence divergence (*22*), which is highly efficient to segregate parental subgenomes but may miss intersubgenome recombination events. More accurate phasing using long-range technologies, such as Hi-C, could resolve this issue and uncover additional structural complexity. Second, our current graph pangenome captures ~75% of the estimated species diversity and could be improved by the addition of unseen genomes. Third, the current population is biased toward wine-associated isolates (*23*), which precludes robust association analyses between structural variation and ecological niche. Broader sampling across environments will be required to address this question. Overall, this study highlights the structural and regulatory consequences of allopolyploidization, while demonstrating the power of graph pangenomes to elucidate genome evolution in complex systems.

METHODS

Graph pangenome construction

Long-read, subgenome-resolved assemblies were retrieved for 71 isolates from a previous study (*22*). To minimize the number of false-positive variants, the genomes were polished using paired-end Illumina sequencing data (*23*) with Hapo-G v1.3.2 (*43*). For each allopolyploid isolate, subgenome assemblies were concatenated into a single fasta file, polished, and contigs belonging to each subgenome were separated further. The graph pangenome reference was generated using Minigraph-Cactus v2.6.4 (*15*) with options *--reference UMY321 --vcf --gfa --gbz*, starting from the primary subgenome reference (*30*). Each was then added in a random order. For nonallopolyploid isolates, the primary genome was attributed to both haplotypes. For allopolyploid isolates, the acquired genome was attributed to the first haplotype, and the primary subgenome was attributed to the second.

Graph pangenome processing

The graph pangenome was converted into a presence-absence matrix with contigs as columns and segments as rows, using the GFA file from Minigraph-Cactus output. Briefly, a custom Python script was used to retrieve the presence of each segment in each contig from the walk lines in the GFA file. Because the subgenome and isolate corresponding to each contig can be traced back, this matrix allows one to check for the presence of segment per contig, subgenome, or isolate. Of the 71 isolates used for graph pangenome construction, a subset of 59 isolates with complete genome assemblies were retained (defined as having both an acquired and a primary subgenome longer than 8 Mb, which corresponds to the estimated size of the shortest acquired subgenome in the population). Subgenome-specific graphs were extracted by filtering out fragments that were present in less than 20% of the corresponding subgenome. This frequency-based threshold was chosen to make the four graphs comparable, despite differences in the number of assemblies available per subgenome. An absolute occurrence cutoff would cause subgenomes represented by more assemblies to accumulate more singleton segments, artificially increasing their graph size. For calculations requiring extensive computation time (e.g., tSNE and rarefaction modeling), a random subset of segments was used, and the results were scaled to match the actual pangenome length.

Rarefaction modeling

The R package iNext (v3.0.0) (*44*) with options $k = 400$ and $size = seq(1, 300)$ was used to model the rarefaction of the pangenome length with nonparametric asymptotic estimators. The modeling was performed using the presence-absence matrix of 10,000 nucleotides among isolates, and scaled back to the actual pangenome length. For subgenome rarefaction modeling, only contigs attributed to the corresponding subgenome were retained.

SV calling

The variant matrices were obtained in two ways. First, the Minigraph-Cactus output was used to obtain subgenome aware genotype calls for the 71 isolates considered in the graph pangenome construction. Second, a mapping strategy was used using the variation graph (vg) toolkit v1.54.0 (*45*) to generate genotypes in the population of 1060 isolates (*23*). The graph pangenome was first converted to GBZ format using the *vg autoindex* command with the option *--workflow giraffe*. Snarls in the graph were identified using the *vg snarls* command. Paired-end Illumina reads were further mapped to the graph using *vg giraffe* with options *--fragment-mean 350 --fragment-stdev 100 -p -b fast --rescue-algorithm gssw* to obtain a GAM file. Read support from the GAM file was obtained using the *vg pack* command, with option *-Q 5* to ignore mapping and base quality below 5. Last, a VCF file was generated for each sample using the *vg call* command with options *--genotype-snarls --all-snarls --snarls Graph.snarls.pb -k S.pack --sample S --ref-sample UMY321*, where *Graph.snarls.pb* is the snarls detected in the previous step, *S.pack* is the read support obtained previously and *S* is the sample name. The single-sample VCF files were merged into a single, multisample file using *bcftools merge* v1.18 (*46*) with option *--merge id*. Genotypes with DP < 2 were set as missing using *bcftools + setGT -- -t q -n. -e "FMT/DP > =2."*

The obtained multisample VCF files were processed to keep only SVs (≥50 bp) following these steps: First, loci without alleles larger than 50 bp were filtered out. Next, small variant alleles (<50 bp) were converted to the reference allele to avoid considering small variants that overlap SVs. Multiallelic loci were further split into biallelic records using *bcftools norm --multi-overlaps 0 -m -any*. To properly handle heterozygosity in the next step, diploid calls were split into two haploid calls. Similar SVs were merged using truvari v5.2.0 (*47*) with the command *truvari collapse --gt all*. Haploid calls were then merged back into diploid calls, and missing genotypes were set to the reference allele using the command *bcftools + setGT -- -t. -n 0* to ensure a conservative approach. Last, SVs were annotated as (i) insertion, when the reference allele length is equal to 1; (ii) deletion, when the alternative allele length is equal to 1; and (iii) complex, when the length of both alleles is >1. Any processing step not performed using bcftools was processed with custom Python scripts.

Subgenome occurrence of SVs

For the SV calls obtained from the Minigraph-Cactus output VCF file, the subgenome occurrence of SVs was determined from their presence among primary and subgenome assemblies. For the SV calls obtained from the graph mapping strategy, the subgenome occurrence of SVs was determined using a competitive mapping method. For each isolate, paired-end Illumina reads were mapped to the concatenation of the primary genome and corresponding acquired genome linear references, using bwa-mem2 v2.2.1 (*48*) with default parameters. Reads mapping to the primary and acquired subgenome references were separated using samtools v1.18 (*46*) with the command *samtools view --region-file contigs.txt -O BAM*, where *contigs.txt* corresponds to either the primary or the acquired subgenome contigs. *Samtools fastq -1 Reads_1.fastq.gz -2 Reads_2.fastq.gz -0 /dev/null -s /dev/null -n* was then used. The SV calling pipeline was then performed independently on the primary and the acquired subgenome reads, with the only exception being that the *--ploidy 1* option was used for the acquired subgenome with the *vg call* command.

Detection of acquired-to-primary genome conversion

For each allopolyploid isolate, regions of acquired-to-primary genome conversion were identified as a drop in nucleotide diversity, based on a short-read mapping strategy. Paired-end Illumina reads were mapped to the primary reference genome (*30*) using bwa-mem2 v2.2.1 (*48*). *Samtools sort* v1.18 (*46*) was used to sort the BAM files and GATK *AddOrReplaceReadGroups* command v4.5.0.0 (*49*) was used to add read groups. Variant calling was performed using Freebayes v1.3.2-dirty (*50*) with option *--ploidy 3*. SNPs were further extracted using bcftools v1.18 (*46*) and filtered for DP ≥ 3. Next, the number of SNPs within 50-kb windows sliding 1 kb was computed with VCFtools v0.1.16 (*51*) option *--SNPdensity 1000*. Peak and valley values were obtained from the kernel density of SNP across all windows. The discriminating threshold was chosen as the valley before the highest peak. A threshold of 19.16 SNP per kb was used to distinguish regions with the presence of both subgenomes from regions with acquired-to-primary genome conversion.

Detection of MITE families

Identification of MITE elements was performed on all genome assemblies using MITE-tracker (*52*) and vsearch v2.7.1 (*53*) with option *--id 0.5*. The representative sequence of each MITE family (file families_nr.fasta) was aligned on SV sequences (inserted sequence for insertions, deleted sequence for deletions, and both alleles for complex SVs) using blast v2.12.0 (*54*) with command *blastn -task blastn -outfmt "6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore qcovhsp qcovs" -query MITEsequences.fasta -subject SVsequences.fasta*. MITE-related SVs were defined by a percentage of identical position (*pident*) higher than 90% and a query coverage per subject (*qcovs*) higher than 80%.

Structural diversity calculation

We defined the structural diversity π_{SV} as the average number of structural differences per site between two DNA sequences, similarly to nucleotide diversity

$$\pi_{SV} = \frac{n}{n+1} \sum_{ij} x_i x_j \pi_{ij} = \frac{n}{n+1} \sum_{i=2}^{n} \sum_{j=1}^{i-1} 2 x_i x_j \pi_{ij}$$

where x_i and x_j are the frequencies of the sequences i and j, π_{ij} is the number of SVs between the two sequences, and n is the total number of sequences. The number of SVs was defined as the number of insertions, deletions, and complex SVs overlapping the sequence of interest. Structural diversity was computed over the whole genome, for the complete population and per genetic cluster, and over 50-kb windows sliding 1 kb with a custom Python script. Structural diversity hotspots were defined as 50-kb windows with a structural diversity higher than the 75th quantile of the genome-wide distribution plus 1.5 times the interquartile range.

Genotype-phenotype local associations

The TPM for each gene was obtained from the supplementary file 3 in (*29*) and was used as a molecular phenotype. TPMs were normalized per gene using a rank-based inverse normal transformation manually implemented in R. A linear mixed-model regression was performed for each SV-gene pair separated by less than 50 kb, using the *lmer* function of the lmerTest R package v3.1-3 (*55*) with genetic cluster assignment as random effect. The effect size was computed as η^2 using the *F_to_eta2* function of the effectsize R package v1.0.0 (*56*). *P* values were corrected using the FDR method implemented in the *adjust_pvalue* function of the rstatix R package v0.7.2 (*57*), and a threshold of 0.05 was used to detect significant associations.

Supplementary Materials

The PDF file includes:

Figs. S1 to S9

Legends for tables S1 to S6

Other Supplementary Material for this manuscript includes the following:

Tables S1 to S6

REFERENCES

1. S. P. Otto, J. Whitton, Polyploid incidence and evolution. *Annu. Rev. Genet.* **34**, 401–437 (2000).
2. D. E. Soltis, P. S. Soltis, Polyploidy: Recurrent formation and genome evolution. *Trends Ecol. Evol.* **14**, 348–352 (1999).
3. J. F. Wendel, D. Lisch, G. Hu, A. S. Mason, The long and short of doubling down: Polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Curr. Opin. Genet. Dev.* **49**, 1–7 (2018).
4. T. Qiu, Z. Liu, B. Liu, The effects of hybridization and genome doubling in plant evolution via allopolyploidy. *Mol. Biol. Rep.* **47**, 5549–5558 (2020).
5. K. Takahagi, K. Inoue, M. Shimizu, Y. Uehara-Yamaguchi, Y. Onda, K. Mochida, Homoeolog-specific activation of genes for heat acclimation in the allopolyploid grass *Brachypodium hybridum*. *Gigascience* **7**, giy020 (2018).

6. S. Volis, D. Ormanbekova, K. Yermekbayev, S. Abugalieva, Y. Turuspekov, I. Shulgina, Genetic architecture of adaptation to novel environmental conditions in a predominantly selfing allopolyploid plant. *Heredity* **116**, 485–490 (2016).
7. P. Wei, X.-L. Yu, Y.-J. Yang, Z.-Y. Chen, S.-Q. Zhao, X.-Z. Li, W.-C. Zhang, C.-L. Liu, X.-Y. Li, X. Liu, Biased gene expression reveals the contribution of subgenome to altitude adaptation in allopolyploid *Isoetes sinensis*. *Ecol. Evol.* **12**, e9677 (2022).
8. S. P. Otto, The evolutionary consequences of polyploidy. *Cell* **131**, 452–462 (2007).
9. Y. Van de Peer, E. Mizrachi, K. Marchal, The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).
10. Y. Fukasawa, Benchmarking long-read variant calling in diploid and polyploid genomes: insights from human and plants. *BMC Genomics* **27**, 46 (2026).
11. M. Alonge, X. Wang, M. Benoit, S. Soyk, L. Pereira, L. Zhang, H. Suresh, S. Ramakrishnan, F. Maumus, D. Ciren, Y. Levy, T. H. Harel, G. Shalev-Schlosser, Z. Amsellem, H. Razifard, A. L. Caicedo, D. M. Tieman, H. Klee, M. Kirsche, S. Aganezov, T. R. Ranallo-Benavidez, Z. H. Lemmon, J. Kim, G. Robitaille, M. Kramer, S. Goodwin, W. R. McCombie, S. Hutton, J. Van Eck, J. Gillis, Y. Eshed, F. J. Sedlazeck, E. van der Knaap, M. C. Schatz, Z. B. Lippman, Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145–161.e23 (2020).
12. C. Chiang, A. J. Scott, J. R. Davis, E. K. Tsang, X. Li, Y. Kim, T. Hadzic, F. N. Damani, L. Ganel, GTEx Consortium, S. B. Montgomery, A. Battle, D. F. Conrad, I. M. Hall, The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
13. S. O'Donnell, J.-X. Yue, O. A. Saada, N. Agier, C. Caradec, T. Cokelaer, M. De Chiara, S. Delmas, F. Dutreux, T. Fournier, A. Friedrich, E. Kornobis, J. Li, Z. Miao, L. Tattini, J. Schacherer, G. Liti, G. Fischer, Telomere-to-telomere assemblies of 142 strains characterize the genome structural landscape in *Saccharomyces cerevisiae*. *Nat. Genet.* **55**, 1390–1399 (2023).
14. E. Garrison, A. Guarracino, S. Heumos, F. Villani, Z. Bao, L. Tattini, J. Hagmann, S. Vorbrugg, S. Marco-Sola, C. Kubica, D. G. Ashbrook, K. Thorell, R. L. Rusholme-Pilcher, G. Liti, E. Rudbeck, A. A. Golicz, S. Nahnsen, Z. Yang, M. N. Mwaniki, F. L. Nobrega, Y. Wu, H. Chen, J. de Ligt, P. H. Sudmant, S. Huang, D. Weigel, N. Soranzo, V. Colonna, R. W. Williams, P. Prins, Building pangenome graphs. *Nat. Methods* **21**, 2008–2012 (2024).
15. G. Hickey, J. Monlong, J. Ebler, A. M. Novak, J. M. Eizenga, Y. Gao, Human Pangenome Reference Consortium, T. Marschall, H. Li, B. Paten, Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* **42**, 663–673 (2024).
16. W.-W. Liao, M. Asri, J. Ebler, D. Doerr, M. Haukness, G. Hickey, S. Lu, J. K. Lucas, J. Monlong, H. J. Abel, S. Buonaiuto, X. H. Chang, H. Cheng, J. Chu, V. Colonna, J. M. Eizenga, X. Feng, C. Fischer, R. S. Fulton, S. Garg, C. Groza, A. Guarracino, W. T. Harvey, S. Heumos, K. Howe, M. Jain, T.-Y. Lu, C. Markello, F. J. Martin, M. W. Mitchell, K. M. Munson, M. N. Mwaniki, A. M. Novak, H. E. Olsen, T. Pesout, D. Porubsky, P. Prins, J. A. Sibbesen, J. Sirén, C. Tomlinson, F. Villani, M. R. Vollger, L. L. Antonacci-Fulton, G. Baid, C. A. Baker, A. Belyaeva, K. Billis, A. Carroll, P.-C. Chang, S. Cody, D. E. Cook, R. M. Cook-Deegan, O. E. Cornejo, M. Diekhans, P. Ebert, S. Fairley, O. Fedrigo, A. L. Felsenfeld, G. Formenti, A. Frankish, Y. Gao, N. A. Garrison, C. G. Giron, R. E. Green, L. Haggerty, K. Hoekzema, T. Hourlier, H. P. Ji, E. E. Kenny, B. A. Koenig, A. Kolesnikov, J. O. Korbel, J. Kordosky, S. Koren, H. Lee, A. P. Lewis, H. Magalhães, S. Marco-Sola, P. Marijon, A. McCartney, J. McDaniel, J. Mountcastle, M. Nattestad, S. Nurk, N. D. Olson, A. B. Popejoy, D. Puiu, M. Rautiainen, A. A. Regier, A. Rhie, S. Sacco, A. D. Sanders, V. A. Schneider, B. I. Schultz, K. Shafin, M. W. Smith, H. J. Sofia, A. N. A. Tayoun, F. Thibaud-Nissen, F. F. Tricomi, J. Wagner, B. Walenz, J. M. D. Wood, A. V. Zimin, G. Bourque, M. J. P. Chaisson, P. Flicek, A. M. Phillippy, J. M. Zook, E. E. Eichler, D. Haussler, T. Wang, E. D. Jarvis, K. H. Miga, E. Garrison, T. Marschall, I. M. Hall, H. Li, B. Paten, A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
17. Q. Meng, P. Xie, Z. Xu, J. Tang, L. Hui, J. Gu, X. Gu, S. Jiang, Y. Rong, J. Zhang, J. A. Udall, C. E. Grover, K. Zheng, Q. Chen, J. Kong, M. Wang, X. Nie, Z. Lin, S. Jin, J. F. Wendel, X. Zhang, D. Yuan, Pangenome analysis reveals yield- and fiber-related diversity and interspecific gene flow in *Gossypium barbadense* L. *Nat. Commun.* **16**, 4995 (2025).
18. O. A. Saada, A. Friedrich, J. Schacherer, Towards accurate, contiguous and complete alignment-based polyploid phasing algorithms. *Genomics* **114**, 110369 (2022).
19. H. Sun, S. Tusso, C. I. Dent, M. Goel, R. Y. Wijfjes, L. C. Baus, X. Dong, J. A. Campoy, A. Kurdadze, B. Walkemeier, C. Sänger, B. Huettel, R. C. B. Hutten, H. J. van Eck, K. J. Dehmer, K. Schneeberger, The phased pan-genome of tetraploid European potato. *Nature* **642**, 389–397 (2025).
20. K. Zhao, H. Xue, G. Li, A. Chitikineni, Y. Fan, Z. Cao, X. Dong, H. Lu, K. Zhao, L. Zhang, D. Qiu, R. Ren, F. Gong, Z. Li, X. Ma, S. Wan, R. K. Varshney, C. Wei, D. Yin, Pangenome analysis reveals structural variation associated with seed size and weight traits in peanut. *Nat. Genet.* **57**, 1250–1261 (2025).
21. M. Avramova, A. Cibrario, E. Peltier, M. Coton, E. Coton, J. Schacherer, G. Spano, V. Capozzi, G. Blaiotta, F. Salin, M. Dols-Lafargue, P. Grbin, C. Curtin, W. Albertin, I. Masneuf-Pomarede, *Brettanomyces bruxellensis* population survey reveals a diploid-triploid complex structured according to substrate of isolation and geographical distribution. *Sci. Rep.* **8**, 4136 (2018).
22. C. Eberlein, O. Abou Saada, A. Friedrich, W. Albertin, J. Schacherer, Different trajectories of polyploidization shape the genomic landscape of the *Brettanomyces bruxellensis* yeast species. *Genome Res.* **31**, 2316–2326 (2021).
23. V. Loegler, J.-N. Runge, C. Eberlein, É. Peltier, J. Harrouard, C. Miot-Sertier, W. Albertin, A. Friedrich, J. Schacherer, Whole-genome sequencing of 1,060 *Brettanomyces bruxellensis* isolates reveals significant phenotypic impact of acquired subgenomes in allopolyploids. *Nat. Commun.* **16**, 5500 (2025).
24. J. De Roos, L. De Vuyst, Microbial acidification, alcoholization, and aroma production during spontaneous lambic beer production. *J. Sci. Food Agric.* **99**, 25–38 (2019).
25. K. M. Lynch, S. Wilkinson, L. Daenen, E. K. Arendt, An update on water kefir: Microbiology, composition and production. *Int. J. Food Microbiol.* **345**, 109128 (2021).
26. R. Suárez, J. A. Suárez-Lepe, A. Morata, F. Calderón, The production of ethylphenols in wine by yeasts of the genera *Brettanomyces* and *Dekkera*: A review. *Food Chem.* **102**, 10–21 (2007).
27. T. Tran, C. Grandvalet, F. Verdier, A. Martin, H. Alexandre, R. Tourdot-Maréchal, Microbial dynamics between yeasts and acetic acid bacteria in kombucha: Impacts on the chemical composition of the beverage. *Foods* **9**, 963 (2020).
28. D. Wedral, R. Shewfelt, J. Frank, The challenge of *Brettanomyces* in wine. *LWT - Food Sci. Technol.* **43**, 1474–1479 (2010).
29. A. Jallet, A. Friedrich, J. Schacherer, Impact of the acquired subgenome on the transcriptional landscape in *Brettanomyces bruxellensis* allopolyploids. *G3* **13**, jkad115 (2023).
30. T. Fournier, J.-S. Gounot, K. Freel, C. Cruaud, A. Lemainque, J.-M. Aury, P. Wincker, J. Schacherer, A. Friedrich, High-quality *de novo* genome assembly of the *Dekkera bruxellensis* yeast using nanopore MinION sequencing. *G3* **7**, 3243–3250 (2017).
31. O. P. Ishchuk, T. V. Zeljko, A. J. Schifferdecker, S. M. Wisén, Å. K. Hagström, E. Rozpędowska, M. R. Andersen, L. Hellborg, Z. Ling, A. A. Sibirny, J. Piškur, Novel centromeric loci of the wine and beer yeast *Dekkera bruxellensis* CEN1 and CEN2. *PLOS ONE* **11**, e0161741 (2016).
32. S. Gui, W. Wei, C. Jiang, J. Luo, L. Chen, S. Wu, W. Li, Y. Wang, S. Li, N. Yang, Q. Li, A. R. Fernie, J. Yan, A pan-*Zea* genome map for enhancing maize improvement. *Genome Biol.* **23**, 178 (2022).
33. Q. He, S. Tang, H. Zhi, J. Chen, J. Zhang, H. Liang, O. Alam, H. Li, H. Zhang, L. Xing, X. Li, W. Zhang, H. Wang, J. Shi, H. Du, H. Wu, L. Wang, P. Yang, L. Xing, H. Yan, Z. Song, J. Liu, H. Wang, X. Tian, Z. Qiao, G. Feng, R. Guo, W. Zhu, Y. Ren, H. Hao, M. Li, A. Zhang, E. Guo, F. Yan, Q. Li, Y. Liu, B. Tian, X. Zhao, R. Jia, B. Feng, J. Zhang, J. Wei, J. Lai, G. Jia, M. Purugganan, X. Diao, A graph-based genome and pan-genome variation of the model plant *Setaria*. *Nat. Genet.* **55**, 1232–1242 (2023).
34. S. Schloissnig, S. Pani, B. Rodriguez-Martin, J. Ebler, C. Hain, V. Tsapalou, A. Söylev, P. Hüther, H. Ashraf, T. Prodanov, M. Asparuhova, S. Hunt, T. Rausch, T. Marschall, J. O. Korbel, Long-read sequencing and structural variant characterization in 1,019 samples from the 1000 Genomes Project. bioRxiv 590093 [Preprint] (2024). https://doi.org/10.1101/2024.04.18.590093.
35. V. Loegler, P. Thiele, E. Teyssonnière, A. Tsouris, G. Brach, C. Cruaud, E. Payen, S. Engelen, M. J. Dunham, J. Hou, A. Friedrich, J. Schacherer, From genotype to phenotype with 1,086 near telomere-to-telomere yeast genomes. *Nature* **648**, 649–658 (2025).
36. V. Loegler, A. Friedrich, J. Schacherer, Overview of the *Saccharomyces cerevisiae* population structure through the lens of 3,034 genomes. *G3* **14**, jkae245 (2024).
37. T. Hämälä, C. Moore, L. Cowan, M. Carlile, D. Gopaulchan, M. K. Brandrud, S. Birkeland, M. Loose, F. Kolář, M. A. Koch, L. Yant, Impact of whole-genome duplications on structural variant evolution in Cochlearia. *Nat. Commun.* **15**, 5377 (2024).
38. J. Vlček, T. Hämälä, C. V. Cobo, E. Curran, G. Šrámková, T. Slotte, R. Schmickl, L. Yant, F. Kolář, Whole-genome duplication increases genetic diversity and load in outcrossing *Arabidopsis*. bioRxiv 632621 [Preprint] (2025). https://doi.org/10.1101/2025.01.12.632621.
39. G. Yildiz, S. F. Zanini, S. Weber, V. Kopalli, T. Kox, A. Abbadi, R. J. Snowdon, A. A. Golicz, Graphical pangenomics-enabled characterization of structural variant impact on gene expression in *Brassica napus*. *Theor. Appl. Genet.* **138**, 91 (2025).
40. M. Agnolucci, A. Tirelli, L. Cocolin, A. Toffanin, *Brettanomyces bruxellensis* yeasts: Impact on wine and winemaking. *World J. Microbiol. Biotechnol.* **33**, 180 (2017).
41. M. Avramova, A. Vallet-Courbin, J. Maupeu, I. Masneuf-Pomarède, W. Albertin, Molecular diagnosis of *Brettanomyces bruxellensis*' sulfur dioxide sensitivity through genotype specific method. *Front. Microbiol.* **9**, 1260 (2018).
42. J. Harrouard, C. Eberlein, P. Ballestra, M. Dols-Lafargue, I. Masneuf-Pomarede, C. Miot-Sertier, J. Schacherer, W. Albertin, *Brettanomyces bruxellensis*: Overview of the genetic and phenotypic diversity of an anthropized yeast. *Mol. Ecol.* **32**, 2374–2395 (2023).
43. J.-M. Aury, B. Istace, Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. *NAR Genom. Bioinform.* **3**, lqab034 (2021).
44. T. C. Hsieh, K. H. Ma, A. Chao, iNEXT: An R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods Ecol. Evol.* **7**, 1451–1456 (2016).
45. G. Hickey, D. Heller, J. Monlong, J. A. Sibbesen, J. Sirén, J. Eizenga, E. T. Dawson, E. Garrison, A. M. Novak, B. Paten, Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* **21**, 35 (2020).

46. P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, H. Li, Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
47. A. C. English, V. K. Menon, R. A. Gibbs, G. A. Metcalf, F. J. Sedlazeck, Truvari: Refined structural variant comparison preserves allelic diversity. *Genome Biol.* **23**, 271 (2022).
48. M. Vasimuddin, S. Misra, H. Li, S. Aluru, "Efficient architecture-aware acceleration of BWA-MEM for multicore systems," in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (IEEE, 2019), pp. 314–324.
49. G. A. Van der Auwera, B. D. O'Connor, *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra* (O'Reilly Media, 2020).
50. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907 [q-bio.GN] (2012).
51. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, 1000 Genomes Project Analysis Group, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
52. J. M. Crescente, D. Zavallo, M. Helguera, L. S. Vanzetti, MITE Tracker: An accurate approach to identify miniature inverted-repeat transposable elements in large genomes. *BMC Bioinformatics* **19**, 348 (2018).
53. T. Rognes, T. Flouri, B. Nichols, C. Quince, F. Mahé, VSEARCH: A versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
54. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
55. A. Kuznetsova, P. B. Brockhoff, R. H. B. Christensen, lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.* **82**, 1–26 (2017).
56. M. S. Ben-Shachar, D. Lüdecke, D. Makowski, effectsize: Estimation of effect size indices and standardized parameters. *J. Open Source Softw.* **5**, 2815 (2020).
57. A. Kassambara, *Rstatix: Pipe-Friendly Framework for Basic Statistical Tests* (2023; https://rpkgs.datanovia.com/rstatix/).
58. V. Loegler, A. Friedrich, J. Schacherer, Code and data for the construction and analysis of the *Brettanomyces bruxellensis* graph pangenome, Zenodo (2025); https://doi.org/10.5281/zenodo.15740803.

Acknowledgments: We thank all the members of the HaploTeam for valuable input and helpful discussions. **Funding:** This work was supported by a European Research Council Advanced grant (UnMiss) to J.S. In addition, this work was supported by a French National Research Agency grant (ANR-22-CE21-0011-03 UnderPhe and ANR-24-CE12-0998-01 PolyPhase). **Author contributions:** Conceptualization, design, and methodology: V.L. and J.S. Genome polishing, graph pangenome construction, SV detection, and association analyses: V.L. Data analyses: V.L., J.S., and A.F. Visualization: V.L. Supervision: J.S. Writing—original draft: V.L. and J.S. Writing—review and editing: V.L., J.S., and A.F. Funding acquisition: J.S. **Competing interests:** The authors declare that they have no competing interests. **Data, code, and materials availability:** All data and code needed to evaluate and reproduce the results in the paper are present in the paper and/or the Supplementary Materials. This study did not generate new materials. Previously published sequencing reads and genome assemblies used in this work are available from the original publications and public repositories cited in the manuscript. The polished genome assemblies and the reference graph pangenome, as well as the code used to construct and analyze the graph, have been deposited in a Zenodo repository (*58*) (https://doi.org/10.5281/zenodo.15740803).

Submitted 13 July 2025
Accepted 22 September 2026
Published 30 October 2026
10.1126/sciadv.aea5812