The effect of hybridization on transposable element accumulation in an undomesticated fungal species

Mathieu Hénault[1245*], Souhir Marsit[12345], Guillaume Charron[1345#a] and Christian R. Landry[12345]

[1]Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec, QC, Canada

[2]Département de biochimie, microbiologie et bioinformatique, Université Laval, Québec, QC, Canada

[3]Département de biologie, Université Laval, Québec, QC, Canada

[4]Quebec Network for Research on Protein Function, Engineering, and Applications (PROTEO), Université Laval, Québec, QC, Canada

[5]Université Laval Big Data Research Center (BDRC_UL), Québec, QC, Canada

[#a]Current address: Centre de foresterie des Laurentides, Ressources naturelles Canada, Québec, QC, Canada

[*]Corresponding author

E-mail: mathieu.henault.1@ulaval.ca

Abstract

Transposable elements (TEs) are mobile genetic elements that can profoundly impact the evolution of genomes and species. A long-standing hypothesis suggests that the merging of diverged genomes within hybrids could alter the regulation of TEs and increase transposition. Higher transposition rates could potentially fuel hybrid evolution with rare adaptive TE insertions, but also cause postzygotic reproductive isolation if maladaptive insertion loads render hybrids sterile or inviable. Mixed evidence for higher TE activity in hybrids was reported in many animal and plant species. Here, we tested for increased TE accumulation in hybrids between incipient species of the undomesticated yeast *Saccharomyces paradoxus*. Population genomic data revealed no increase in TE content in the natural hybrid lineages. As we could not rule out the elimination of past transposition increase signatures by natural selection, we performed a laboratory evolution experiment on a panel of artificial hybrids to measure TE accumulation in controlled conditions and in the near absence of selection. Changes in TE copy numbers were highly dependent on the individual hybrid genotypes and were not predicted by the evolutionary divergence between the parents of a hybrid genotype. Rather, our data suggested that initial TE copy numbers in hybrids negatively impacted transposition rate, suggesting that TE self-regulation could play a predominant role on TE accumulation in yeast hybrids.

Introduction

Hybridization is increasingly recognized as an important component of species evolution [1–3]. Abundant empirical evidence shows that hybridization frequently leads to transgressive phenotypes in plants, animals and fungi [4,5]. Mutation rates are traits for which transgressive segregation can have a major impact on the evolutionary fate of hybrids, with potential implications for the emergence of reproductive barriers and speciation. Insertions caused by transposable elements (TEs) are a class of mutations that could be particularly responsive to hybridization. TEs are diverse and ubiquitous DNA sequences that can duplicate and spread within genomes, forming families of dispersed repeated sequences [6]. They can either replicate through “copy-and-paste” (retrotransposons) or “cut-and-paste” (DNA transposons) mechanisms. TEs were first identified in cytological studies of chromosome instability in *Zea mays* [7]. In her seminal 1984 perspective, McClintock summarized evidence suggesting that TEs respond to various types of “genome shocks”, one of which she suspected was hybridization [8]. Since McClintock’s early insights, many studies investigated the association between TEs and hybridization across taxonomic groups and documented increased TE activity in hybrids. We refer to this increase as TE reactivation, but our use of this term implies no mechanistic basis and other terms like derepression are used in the literature. Reactivation was evidenced by increased TE abundance, novel TE insertions, TE epigenetic changes or higher TE transcript levels [9–15]. However, hybrid reactivation is not universal, as many studies across a similar taxonomic range found no evidence for it [16–19]. Thus, TE reactivation exhibits no clear taxonomic distribution, and the conditions favorable for its manifestation are still unclear.

Studies of hybrids in controlled crosses revealed genetic bases to TE reactivation, but less focus was put on cases of natural hybrids (but see [18,20]). In addition to the genetic structure

of populations, environmental pressures and demographic factors can influence the action of evolutionary forces on TEs, thus impacting the genomic landscapes of TEs in natural populations, including hybrids. Population genetic models have investigated the action of the major evolutionary forces on mean TE copy number (CN) per individual [21,22]. Three such forces are deterministic [23]. First, TE CNs increase at a certain transposition rate, which incorporates the various forms of transposition regulation (either by the host or by TEs themselves). Second, TEs are removed from genomes at a certain excision rate. Third, most TE insertions are deleterious, such that individual fitness is assumed to be a decreasing function of CN and natural selection acts to purge TE insertions from populations. The fourth important force is genetic drift, a stochastic force that impacts the efficiency of natural selection by increasing the probability of fixation of deleterious mutations [24]. The intensity of drift is modulated by the effective population size N_e. In populations with small N_e, increased drift renders natural selection inefficient, such that deleterious mutations are more likely to fix by chance. In principle, any of those evolutionary forces could affect hybrids in a way that differs from their parents, leading to hybrid-specific dynamics in the accumulation of TEs.

Population data in *Drosophila* indicate that deterministic forces have a predominant influence on the allele frequency spectrum of TE insertions [22], consistent with a minimal contribution from drift. Among the deterministic forces, transposition rate is arguably the most likely to be impacted by hybridization, for instance if genetic incompatibilities impair transposition regulation mechanisms. A classic illustration of this is hybrid dysgenesis in *Drosophila*, a syndrome caused by mismatches between paternal TEs and egg-loaded pools of Piwi-interacting RNAs (piRNAs) that lead to inefficient TE repression [25–28]. Similarly, hybrids between *Arabidopsis thaliana* and *A. arenosa* show overexpression of a TE family from the *A. arenosa* subgenome that is dependent on the ploidy level of the parental accessions involved, suggesting that dosage

imbalance of epigenetic regulators causes reactivation [29]. It can also be expected that the number of genetic incompatibilities increases with genetic divergence, yielding a positive relationship between evolutionary divergence and TE deregulation [30,31]. Additionally, evidence from *Drosophila* suggests that hybridization may impact excision rates for specific TE families [16]. Along with these deterministic factors, reductions in natural selection efficiency should also be considered. Evidence from *Arabidopsis* [32,33], *Caenorhabditis* [34] and *Daphnia pulex* [35] suggests that long-term differences in N_e caused by demography or mating systems are associated with variation in TE content [36]. A reduction in N_e can be expected when a novel hybrid lineage is formed. Indeed, it has been hypothesized that population bottlenecks inherent with hybrid speciation in *Helianthus* might have played a role in the accumulation of TEs through a decreased efficiency of purifying selection [20].

An emerging eukaryotic model system in ecological and evolutionary genomics is the yeast *Saccharomyces paradoxus*, the undomesticated sister species of the budding yeast *Saccharomyces cerevisiae*. *S. paradoxus* has a complex admixture history in North America, making it an ideal model system to test hybridization-related evolutionary hypotheses in natural populations and in the laboratory (Fig 1) [37]. North american *S. paradoxus* lineages comprise the Eurasian lineage *SpA*, the endemic incipient species *SpB* and *SpC* (hereafter termed pure lineages), and the incipient hybrid species *SpC**, SpD_1 and SpD_2 (hereafter termed hybrid lineages; Fig 2a). *SpC** is an introgressed lineage that arose from a post-glacial secondary contact between the endemic lineages *SpB* and *SpC* and consist of mostly *SpC*-like genomes with ~5% introgressions from *SpB* [38], while SpD_1 and SpD_2 arose more recently from hybridization between *SpC** and *SpB* and have roughly equal genomic contributions from both parents [39] (Fig 2b).

Yeasts of the *Saccharomyces* genus are valuable model species to investigate the interplay between hybridization and transposition. *Saccharomyces* genomes contain many long terminal repeats (LTR) retrotransposons of the Ty1/copia (Ty1, Ty2, Ty4 and Ty5) and Ty3/gypsy (Ty3) families. Ty1/copia and Ty3/gypsy retrotransposon superfamilies [40] respectively correspond to the *Pseudoviridae* and *Metaviridae* virus families [41] and have a broad taxonomic distribution. Ty elements comprise two overlapping open reading frames (ORFs) within a ~5 kb internal sequence, which is flanked by two ~300 bp LTRs in direct orientation. The two LTRs of full-length elements are prone to intra-element recombination, leaving behind solo LTRs in the genome [42]. Ty1 in the budding yeast *S. cerevisiae* is among the best characterized eukaryotic TEs, benefiting from decades of research on its molecular and cellular biology and, more recently, on its population genomics [43,44]. Although *Saccharomyces* genomes harbor a relatively modest fraction of Ty elements (~1-3% in size), their streamlined genomes (~12 Mbp) and short generation time (~2h) provide great power to perform TE comparative genomics and laboratory evolutionary experiments.

Here, we use population genomic data and laboratory evolution experiments on *S. paradoxus* to investigate the hybrid reactivation hypothesis and, more generally, the factors governing TE accumulation in natural and experimental lineages. Our results show that natural and experimental hybrids exhibit no reactivation of TEs. Rather, we show that deterministic factors like population structure and the properties of individual hybrid genotypes are major determinants of TE content evolution in hybrid genomes.

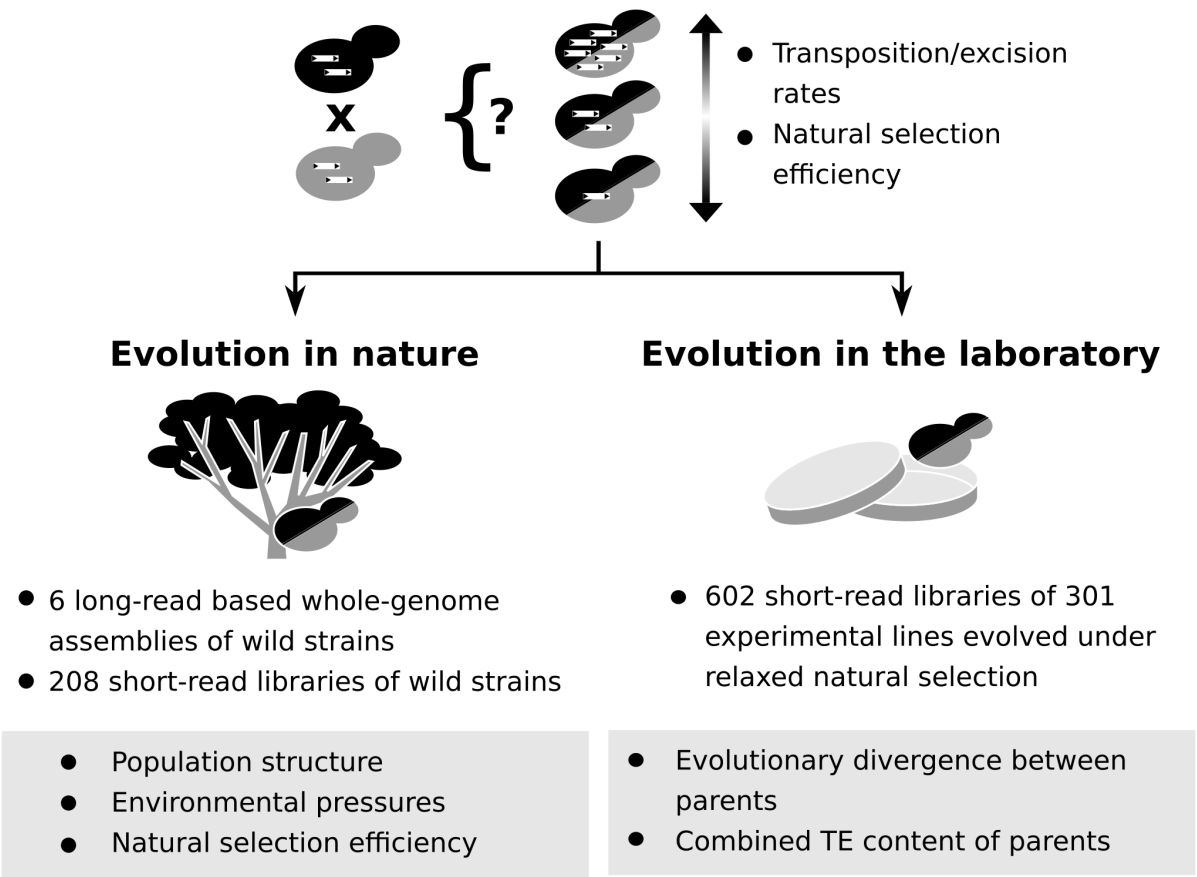

Fig 1. Population genomic data and laboratory evolution experiments allow to uncover the effects of various factors on TE accumulation in hybrids. Upon hybridization, the TE content of a hybrid is determined by equal contributions from the TE contents of its parents. Subsequent evolution can either leave the hybrid TE content unchanged, or drive it to increase or decrease. These changes can be driven by biases in transposition or excision rates or by variation in natural selection efficiency. Two complementary approaches can be used to investigate which factors drive TE dynamics in hybrids. First, genomic data from natural populations can be harnessed to understand how population structure, environmental pressures and natural selection efficiency shaped natural variation in TE content. Second, evolution experiments in laboratory controlled conditions (constant environment and relaxed natural

selection) allow to test the effect of properties of artificial hybrid genotypes on TE accumulation, namely evolutionary divergence between parents and initial TE content.

Results

S. paradoxus lineages exhibit contrasted TE evolutionary dynamics but no reactivation in hybrids

To investigate the TE content and dynamics in North American *S. paradoxus* pure and hybrid lineages, we annotated six highly contiguous whole-genome assemblies previously generated from Oxford Nanopore long reads (PRJNA514804) [39]. We used a database comprising reference sequences for *S. cerevisiae* Ty1, Ty2, Ty3, Ty4 and Ty5 families, along with *S. paradoxus* Ty3-related Ty3p [45] and Ty4-related Tsu4, the latter originating from a horizontal transfer from the *S. eubayanus*/*S. uvarum* clade to *S. paradoxus* [46]. These annotations revealed counts of full-length elements an order of magnitude lower than solo LTRs (Fig 2c, Fig S1), consistent with previously reported data on *S. cerevisiae* and *S. paradoxus* [45,47–49]. To overcome the difficulty of comparing multiple families of short and often degenerated LTR sequences within a phylogenetic framework, we built intra-genome LTR sequence similarity networks. Sequences within the Ty1-Ty2 and Ty3p-Ty3 LTR family pairs were indistinguishable (Fig S2), in agreement with Ty2 and Ty3 being specific to *S. cerevisiae* [45,50]. Consequently, Ty2 and Ty3 LTR annotations were merged with Ty1 and Ty3p, respectively. Although they were clearly distinct, Ty4 and Tsu4 LTRs showed a close relatedness, which justified their combination into the Ty4/Tsu4 family. No Tsu4 LTR was annotated in the *SpA* genome, consistent with this family being specific to North American *S. paradoxus* and absent from Eurasian *S. paradoxus* [46].

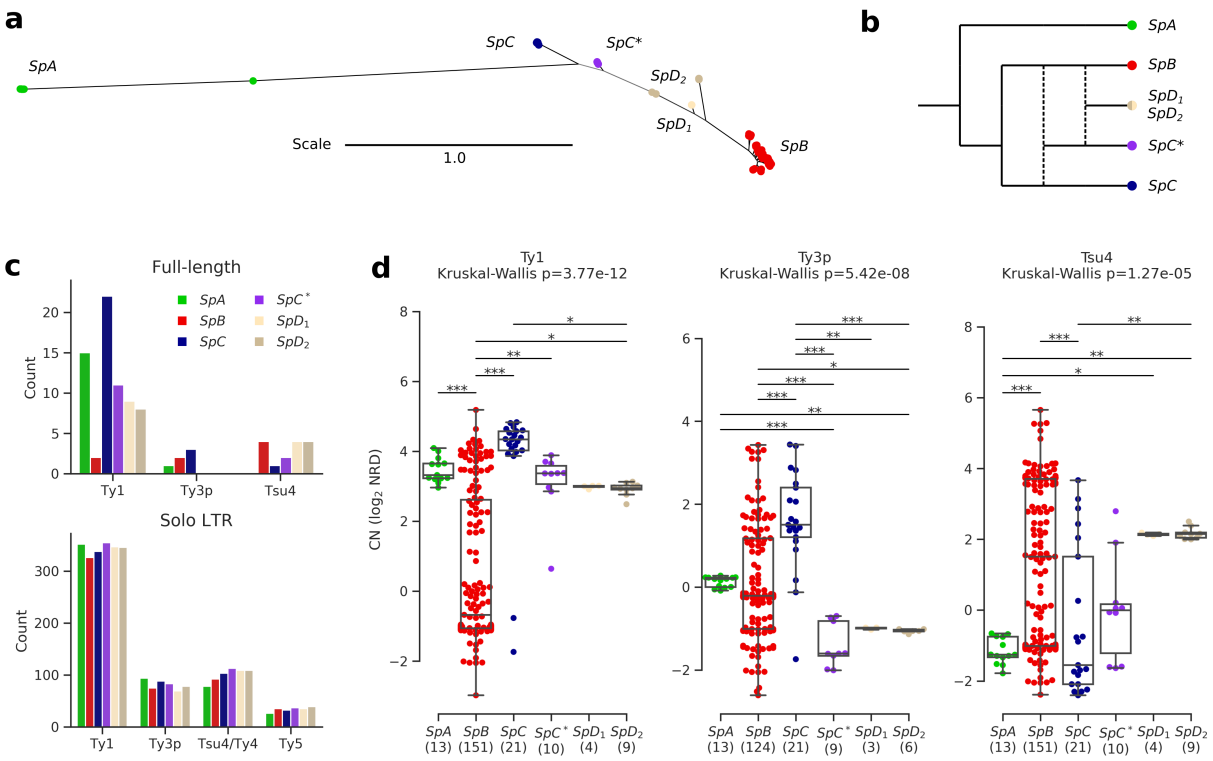

Fig 2. Ty elements showed no reactivation in hybrid lineages and contrasted abundances among pure lineages. a. Bayesian phylogenetic tree of the North American *S. paradoxus* lineages based on genome-wide SNPs. **b.** Schematic admixture graph showing the evolutionary relationships between pure and hybrid lineages as inferred by [38,39]. Dotted lines represent admixture events. **c.** Counts of full-length elements (top) and solo LTRs (bottom) in six whole-genome assemblies based on long reads. No full-length Ty4 or Ty5 were found in any assembly, thus these families were considered extinct. **d.** CN variation in active Ty families (i.e. families which have full-length elements) measured as $\log_2$ normalized read depth (NRD) over Ty reference internal sequences for 208 wild strains. Sample size for each lineage (excluding strains for which there was no coverage) is indicated between parentheses. Whiskers span 1.5 times the interquartile range. P-values of Kruskal-Wallis tests are shown. Stars on bars show the results of Conover post-hoc tests. *:p<0.05, **:p<0.01, ***:p<0.001.

Based on the assembly annotations, Ty1 was the most abundant family and exhibited the largest CN variation among lineages, with a striking difference between *SpB* and *SpC* (Fig 2c). The generality of CN patterns revealed by genome assemblies was confirmed by normalized short-read depth (NRD) over Ty reference sequences for 208 natural isolates (Fig 2d, Fig S3, Table S1; PRJNA277692, PRJNA324830, PRJNA479851). The hybrid lineages *SpC**, SpD_1 and SpD_2 had intermediate full-length Ty1 CNs compared to their parental lineages, indicating no apparent reactivation in these lineages. Notably, Ty3p appeared extinct in all hybrid lineages, with no signal for internal sequences. The most outstanding profile was that of the pure lineage *SpC*, which had significantly higher full-length Ty1 and Ty3p CNs compared to the other lineages. For Tsu4, *SpB* was the lineage with the highest CN. Thus, lineage-wide CNs were consistent with an absence of reactivation in hybrids, the most contrasted profiles being found among pure lineages.

CNs alone may not reflect the actual dynamics of a TE family within a genome. For instance, a recently expanding family may not have had the time to reach the CN of another family (or the same family in a different lineage) that would be more abundant but at equilibrium. We thus explored the evolutionary dynamics of Ty families within each genome. If high CNs were caused by recent family expansions, they should be associated with a signature of low divergence [51]. We used the sequence divergence to the most closely related LTR for each LTR annotation to approximate evolutionary divergence since transposition. The importance of low-divergence peaks (≤1%) for Ty1 was significantly higher in *SpC* and *SpA* compared to *SpB* (Fisher's exact test, FDR corrected, Fig 3a; see Fig S4 for other divergence thresholds). *SpA* also had a significant underrepresentation of low-divergence LTRs for the composite Ty4/Tsu4 family due to the absence of Tsu4. The height of low-divergence peaks was consistent with CNs of full-length elements (Fig 2d). Hybrid lineages *SpC**, SpD_1 and SpD_2 had Ty1 peaks similar to *SpC*,

although less important. They also had no peak for Ty3p, consistent with the absence of full-length elements. Overall, this result indicated that CN variation reflected variation in family dynamics and confirmed that Ty elements exhibited no reactivation in hybrid lineages.

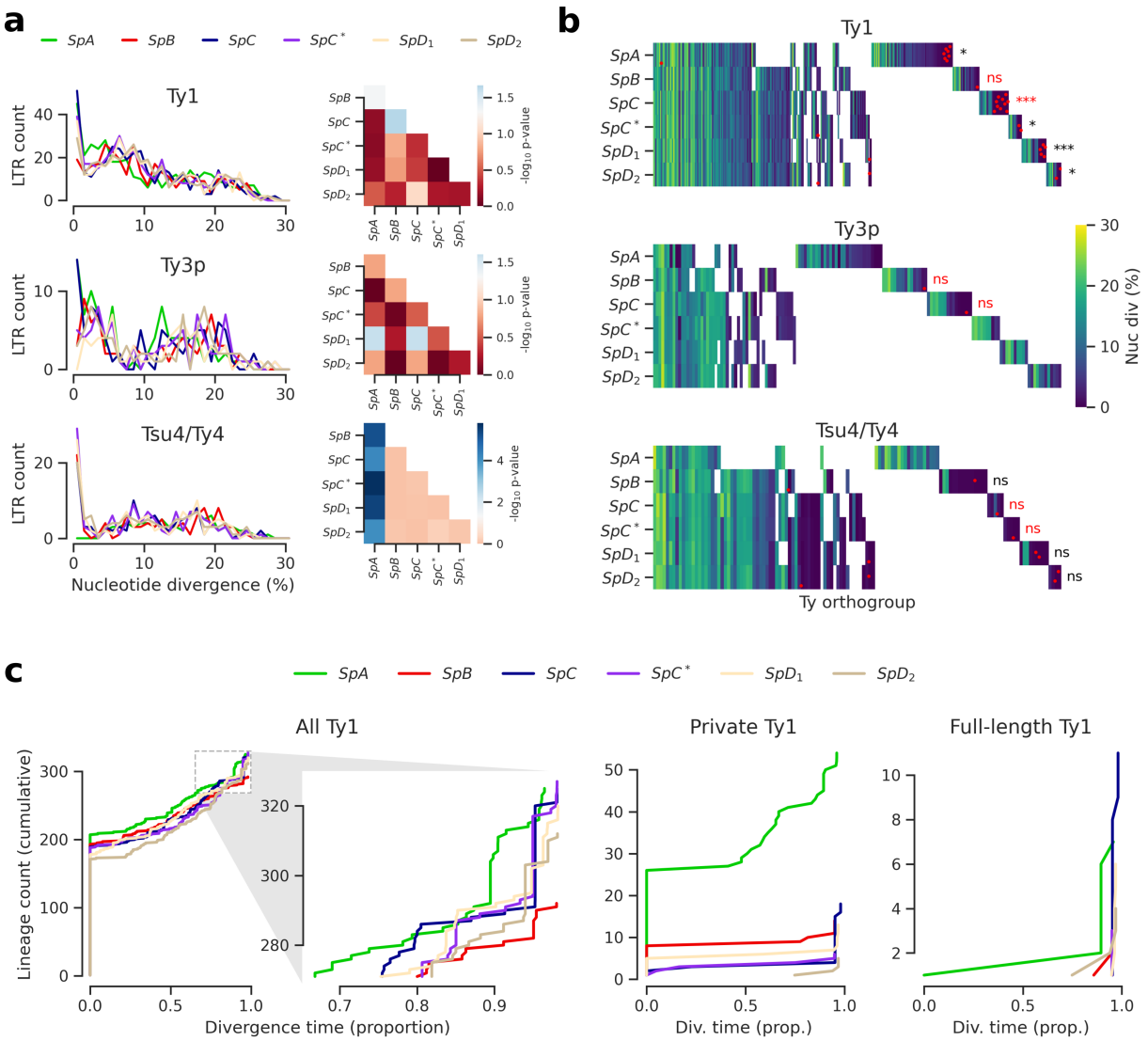

Fig 3. Evolutionary dynamics of Ty families in *S. paradoxus*. a. Distributions of minimum nucleotide divergence between LTR sequences. Heatmaps show FDR-corrected p-values for pairwise Fisher's exact tests between ratios of elements ≤1% and >1% nucleotide divergence. Color maps are centered at the significance threshold of 0.05. See Fig S4 for other divergence

thresholds. **b.** LTR divergence and conservation levels. Horizontal positions depict Ty orthogroups clustered from conserved (left) to private (right). Color map shows minimum nucleotide divergence. Scattered red dots show LTRs that belong to full-length elements. FDR-corrected p-values of Fisher's exact tests for the ratio of full-length elements in private orthogroups are shown (red marks denote cases where counts of full-length elements in non-private orthogroups were artificially inflated from 0 to 1). ns: $p \geq 0.05$, *:$p<0.05$, **:$p<0.01$, ***:$p<0.001$. **c.** Lineages through time curves for Ty1 ultrametric bayesian phylogenetic trees, using a conserved Ty4 LTR orthogroup as an outgroup. From left to right, distributions show all LTRs (with a close-up on recent dynamics), LTRs from private elements and LTRs from full-length elements.

Next, we tested whether orthology between genome-wide Ty insertion loci was consistent with transposition activity in all lineages. We defined Ty orthogroups from a whole-genome multiple alignment of the six assemblies (Fig 3b). Most orthogroups were either private to one genome or conserved across the six (Fig S5a), suggesting that a major fraction of Ty insertion polymorphisms segregated at low frequency in North American lineages. This was consistent with studies in *S. cerevisiae* showing that insertions are mostly fixed or at low frequency [44,45]. Private orthogroups comprised LTRs that had low divergence, consistent with recent transposition activity (Fig S5b). Private orthogroups were significantly enriched in full-length Ty1 elements (Fig 3b), including hybrid lineages. This indicated that all lineages had active transposition for at least some Ty families. It also suggested that an absence of transposition in hybrid lineages could not explain the absence of reactivation. We called lineage-wide insertion polymorphisms from discordant mappings of the 208 short read libraries to estimate the distribution of insertion allele frequencies. This calling method is noisy and only allows to call non-reference alleles, but should nevertheless provide rough estimates of the insertion allele

frequency spectra. Principal component analysis (PCA) on non-reference insertions yielded the expected grouping of strains into well-defined lineages (Fig S6a). The allele frequency distributions showed a similar excess of very low frequency alleles for the six lineages (Fig S6b), suggesting that Ty elements were subject to largely uniform evolutionary forces across lineages.

We focused on the most abundant Ty1 to investigate the transposition dynamics in a phylogenetic framework. Bayesian phylogenetic trees of Ty1 LTRs were converted into ultrametric phylogenies to approximate the timing of branching events using a conserved Ty4 LTR orthogroup as an outgroup (Fig S7). Branching times revealed that lineages had largely similar Ty1 dynamics (Fig 3c). Due to the shortness of LTR sequences, most branches collapsed with the outgroup, indicating an expected saturation of the molecular substitution signal. A focus on recent dynamics revealed a clear expansion signature in *SpC* that was shared by *SpC** (Fig 3c), which was reflected by the peaks of recent private and full-length elements in *SpC*. The rapid expansion in *SpC* and *SpC** can be contrasted with the more gradual expansion in *SpA*, which harbored the second highest Ty1 CN. This contrasted with the assumption that TE abundances can be interpreted as equilibrium copy numbers under the transposition-selection balance hypothesis [21] and rather suggested a recent transposition rate increase in *SpC*-related lineages.

Population structure best explains TE content variation, with minor contributions from climatic variation

Our results showed that hybrid lineages exhibited no outstanding Ty CN profiles or sequence dynamics. However, potential hybrid-specific differences in transposition rate might be masked by other factors that can be expected to impact Ty dynamics. To exclude this possibility, we

investigated which factors best explain the variation in Ty CN in natural lineages. First, population structure incorporates both host-encoded [43] and Ty-encoded [52,53] determinants of Ty transposition rates. Second, environmental factors can influence Ty transposition rate, as demonstrated by temperature [54] and nitrogen starvation [55] for Ty1 in *S. cerevisiae*. If hybrid lineages experience distinct environmental conditions in nature, this may impact their transposition rates. Finally, the efficiency at which natural selection purges deleterious Ty insertions (which depends on N_e) could vary between lineages, ultimately affecting the transposition-selection balance. To help discriminate global from lineage-specific patterns arising from these factors, analyses with all lineages aggregated were complemented by lineage-specific analyses. However, hybrid lineages were represented by relatively few strains (≤10 in each case), which made it hard to reach meaningful conclusions for some analyses. Thus, some of the lineage-specific analyses were focused on the more abundant *SpB* and *SpC* pure lineages.

We first investigated the relative contribution of population structure and environmental factors. We combined genome-wide single nucleotide variants from our population genomic data with climatic (temperature, precipitation, solar radiation, vapor pressure, wind) and water balance data [56] and reduced the dimensionality of both datasets by PCA, using coordinates on the major principal components (PCs) to describe population structure (Fig S8) and climatic variation (Fig S9, Fig S10). We then fitted linear models to compare how genotypes-based PCs (gPCs) and environment-based PCs (ePCs) can explain CN variation in Ty1, Ty3p and Tsu4 (Fig 4a). Globally, gPCs reached higher levels of statistical significance as linear predictors of CN compared to ePCs. gPC1 both explained a high proportion of variance (44%, 9% and 32% for all lineages, *SpB* and *SpC*, respectively; Fig S8) and was a highly significant linear predictor of Ty1 CN. Linear models on all lineages that explained Ty1 and Tsu4 CN with only gPCs were

more parsimonious than models that included both gPCs and ePCs (as indicated by lower Akaike information criterion [AIC] values; Table S2), suggesting that population structure was sufficient to explain CN variation at a global scale for these families. The model incorporating both gPCs and ePCs for global Ty3p CN as well as lineage-specific models for all families had lower AIC values, suggesting that climatic variation may impact CNs in a family- or lineage-specific manner (Table S2). Unsupervised clustering of strains along gPCs yielded more consistent clusters than ePCs, with lower within-cluster variance in Ty CN (Fig 4b, Fig S11). Overall, these results show that the genetic structure of North American *S. paradoxus* is generally a much better predictor of Ty CNs across lineages than climatic variables. Thus, exposure to specific environmental conditions appears unlikely to explain the absence of observable Ty reactivation in hybrid lineages.

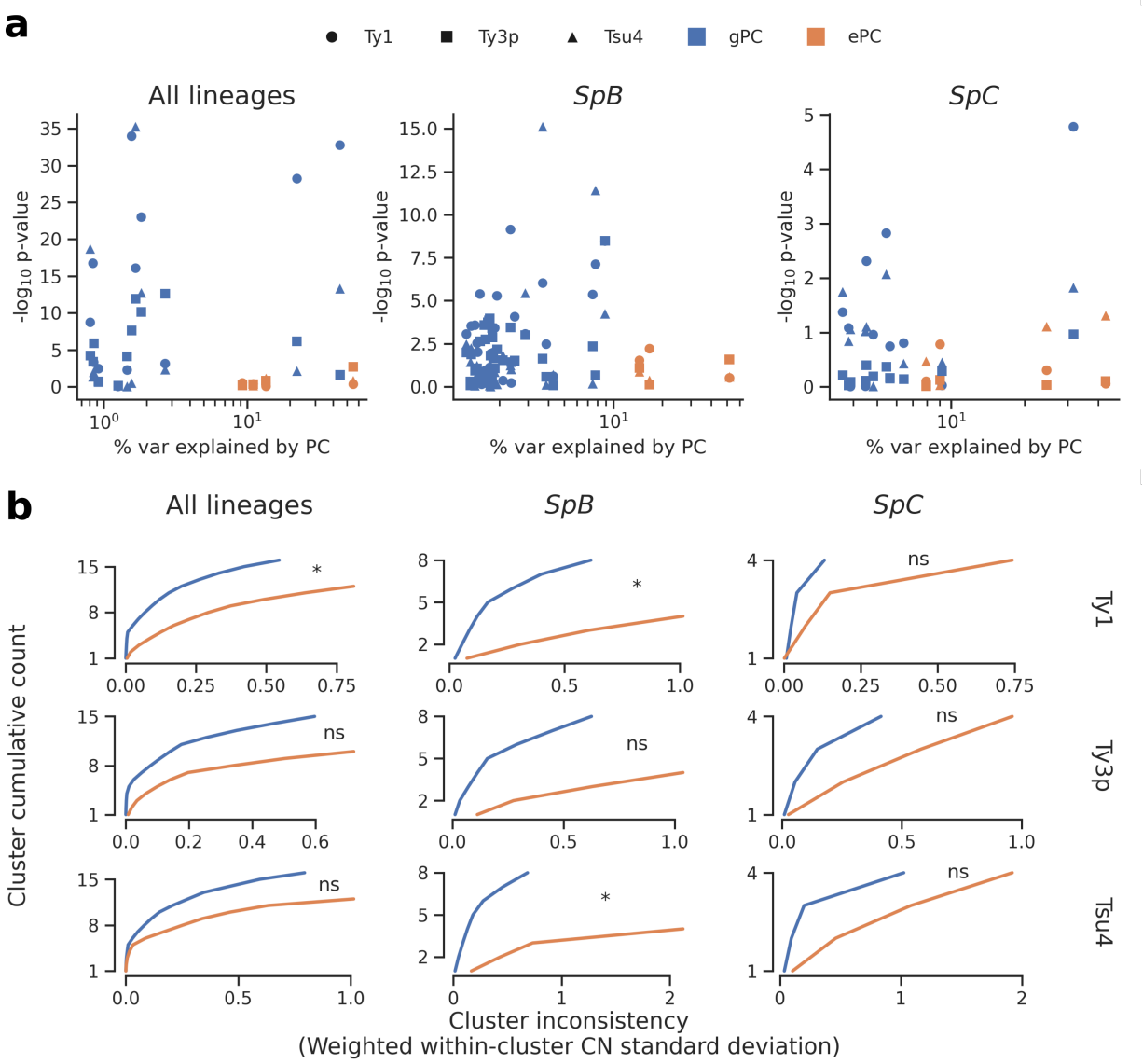

Fig 4. Associations between population structure, climatic variables and Ty CNs for all lineages (left), *SpB* only (middle) or *SpC* only (right). a. Linear models fitted on Ty CNs with gPCs and ePCs as independent variables. For each PC, the p-value in the linear model is plotted against the proportion of variance explained in the respective PCA. **b.** Unsupervised clustering performed on either gPC or ePC coordinates. Cumulative distributions show the standard deviation in Ty CN within each gPC-cluster or ePC-cluster (weighted by the

corresponding cluster size). FDR-corrected p-values of two-tailed Mann-Whitney U tests between gPC-cluster and ePC-cluster distributions are shown. ns: p≥0.05, *:p<0.05.

Variation in natural selection efficiency cannot explain TE abundance variation in wild lineages

Most TE insertions are expected to have deleterious effects on the individual in which they occur [57], although some can be advantageous [58–61]. Natural selection acts to purge deleterious insertions, and its efficiency depends on the strength of genetic drift opposing it (which is inversely proportional to N_e). Ty reactivation in hybrid lineages might be counterbalanced by higher efficiency of natural selection in these lineages, although there is no *a priori* reason to expect higher N_e in hybrids. We thus sought to investigate if CN variation across lineages can be explained by variation in N_e. Pure lineages exhibit variation in genome-wide nucleotide diversity at four-fold degenerate codon positions (π_S; Table S3), which is a rough estimation for long-term N_e [36]. However, there was no obvious negative relationship between Ty CNs and π_S (Fig S12), as would be expected under the assumption of transposition-selection equilibrium. For instance, nucleotide diversity was higher in *SpC* than in *SpA*, while *SpC* exhibited the highest Ty1 and Ty3p CN. This suggested that natural selection efficiency is likely not the main determinant in CN variation. However, our data indicated that some Ty families experienced recent expansions and thus may not be at equilibrium. Recent hybrid lineages also cannot be meaningfully characterized by long-term N_e, further limiting the usefulness of π_S as a measure of the efficiency of natural selection.

We can expect that most of the fitness effects associated with TE insertions are caused by alterations of host functions encoded in *cis*, for instance through altered transcription rates following insertion in untranslated or regulatory regions. Indeed, Ty elements contain many

transcription factor binding sites and can affect the expression levels of adjacent genes [55,62,63]. Thus, functional properties of the genes linked with TE insertions should reflect the efficiency of natural selection acting against the latter. Perhaps the TE insertions with the largest functional alteration potential are those which interrupt ORFs of host genes. Across the six genome assemblies, no Ty insertion was found to interrupt an ORF. Lineage-wide Ty insertion calls revealed four Ty insertions predicted to be within host ORFs, three of which are found in *SpC* strains and one in *SpB*. All the ORF-disrupting insertions were private alleles, and in all cases the strain harboring them had a CN higher than the median of its lineage (Fig S13). Examination of depth of coverage at these four loci revealed narrow peaks in the corresponding strains that were consistent with target site duplications (TSDs) arising from Ty integration (Fig S14). The absence of ORF-disrupting Ty insertions in hybrid lineages and their scarcity in pure lineages suggests that natural selection is globally very efficient at purging these mutations.

The overrepresentation of ORF-disrupting Ty insertion alleles in *SpC* was suggestive of a slight bias towards more deleterious insertion alleles in this lineage. There were striking patterns of functional relatedness among genes harboring a ORF-disrupting insertion. Two *SpC* strains had independent Ty1 insertions at different loci within YDL024C (*DIA3*, which encodes a protein of unknown function [64]), as evidenced by the corresponding TSDs (Fig S14). The two other genes were involved in resistance to arsenic compounds: LL2011_004 (*SpC*) had a Ty1 insertion disrupting YPR201W (*ARR3*, which encodes an arsenite/H^+ antiporter [65]), while 16_037 (*SpB*) had a Tsu4 insertion in YPR200C (*ARR2*, which encodes an arsenate reductase [66]). This functional relatedness could be explained by positive selection; however, this scenario is incompatible with the extremely low frequency at which these insertions segregate. Thus, the most parsimonious explanations for these patterns appears to be either a mutational bias favoring Ty insertions at these loci or simply a lack of fitness effect.

We investigated the fitness effect of the ORF-disrupting Ty1 insertion in *ARR3* by measuring the growth of multiple *S. paradoxus* strains in culture medium supplemented with arsenite. *SpC* and *SpC** lineages exhibited poor growth in 0.8 mM arsenite compared to *SpB* and *SpA* (Fig 5a). This phenotype was generalized to the whole lineages, and strain LL2011_004 was not outstanding in the *SpC* distribution (Fig 5b). This indicated that the disruption of a major gene involved in arsenite resistance had no further deleterious effect on the poor growth ability of *SpC*-related strains in presence of arsenite. The absence of growth defect in arsenite for LL2011_004, combined with the extremely low frequencies of the other ORF-disrupting alleles, suggests that these insertions are effectively neutral and are not observed because of an important reduction in purifying selection efficiency.

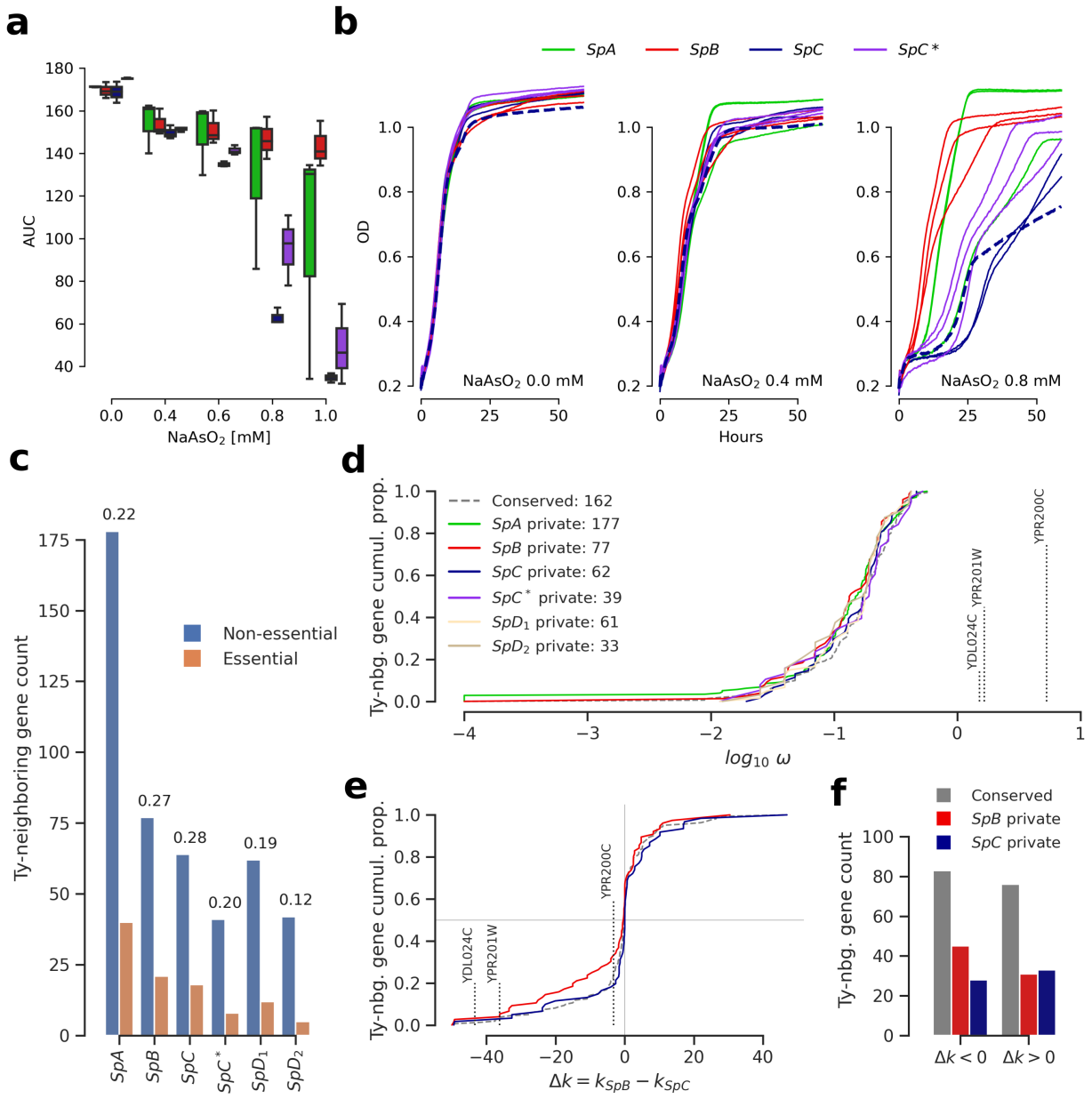

Fig 5. Genomic patterns of Ty insertions indicate that natural selection efficiency cannot explain CN variation in natural lineages. a. Growth curve area under the curve (AUC) at various concentrations of arsenite. Three representative strains per lineage were assayed. **b.** Growth curves showing optical density (OD) of cultures through time at 0, 0.4 and 0.8 mM $NaAsO_2$. The *SpC* strain LL2011_004 (which harbors a Ty1 insertion in *ARR3*) is highlighted

with dotted curves. **c.** Counts of essential and non-essential genes in the immediate vicinity of Ty insertions across the six whole-genome assemblies. Essential to non-essential ratios are shown on top of the bars. **d.** Cumulative distribution of ω for genes neighboring conserved (dotted line) or private (solid lines) Ty insertions. Genes containing ORF-disrupting insertions are labeled. **e.** Cumulative distributions of Δk between *SpB* and *SpC.* **f.** Counts of neighboring genes with positive or negative Δk values.

To gain a more comprehensive view of the strength of selection acting on Ty insertions, we compared the properties of genes in their adjacency. If relaxed purifying selection is responsible for Ty CN variation, we reason that high CNs in specific lineages should be associated with neighboring genes that otherwise experience stronger purifying selection overall in *S. paradoxus*. Such genes can be assumed more functionally critical for the cell, and harmful mutations in their vicinity should accumulate in populations in which purifying selection is weaker. Focusing on the Ty insertions private to each lineage, we found no significant difference in the proportion of neighboring genes annotated as essential in *S. cerevisiae* ($p>0.05$, pairwise Fisher's exact tests, FDR correction), although *SpC* did exhibit the highest essential to non-essential ratio (Fig 5c). We fitted the ratio of non-synonymous to synonymous substitutions (ω) to each neighboring gene to estimate the strength of purifying selection on protein-coding sequences. The ω distributions for genes neighboring private Ty insertions exhibited no significant bias ($p>0.05$, pairwise two-tailed Mann-Whitney U tests, FDR correction). In particular, the distribution in *SpC* exhibited no skew towards lower values, suggesting no association with genes under strong purifying selection (Fig 5d). Notably, genes harboring ORF-disrupting insertions show no signature of purifying selection (Fig 5d).

We next investigated whether lineage-specific selection signatures on genes are associated with Ty insertions. We used the RELAX branch-site codon model [67] to derive estimates of

selection intensity (k) along individual gene trees. Within the RELAX framework, k<1 indicates relaxed selection and k>1 indicates strengthened (positive or purifying) selection. We define $\Delta k = k_{SpB}-k_{SpC}$, corresponding to the difference between the average k values for clades comprising *SpB* and *SpC* strains, respectively. Overall, the distribution of Δk values for genes neighboring private *SpC* insertions was very similar to that for conserved insertions (Fig 5e), while the distribution for private *SpB* insertions was slightly biased towards genes relaxed in *SpB.* For both *SpB* and *SpC*, there were more private insertions adjacent to genes experiencing relaxed selection in the corresponding lineage, although comparisons with conserved insertions were not statistically significant (p>0.05, Fisher's exact tests, Fig 5f). This is consistent with efficient purifying selection acting against insertions near genes in both lineages. While genes harboring ORF-disrupting insertions have signatures of stronger selection in *SpC* (Fig 5e, Fig S15), this observation is mitigated by their far-right position in the ω distribution (Fig 5d). Overall, these results show that variation in natural selection intensity is unlikely to explain a large part of the variation in Ty CNs in natural lineages.

TE content variation in experimental evolution under relaxed selection reveal no systematic effect of hybridization but major effects of individual genotypes

Thus far, our results showed that no reactivation is apparent in natural hybrid lineages, that population structure over environmental variables best explains variation in Ty CNs and that natural selection efficiency is unlikely to explain much variation in Ty genomic landscapes. However, natural selection seems equally efficient at purging deleterious Ty insertions in all lineages, leaving the possibility that strong purifying selection could mask any reactivation in hybrids. Moreover, natural hybrids are somewhat limited in their ability to uncover the interplay between hybridization and transposition, as they are relatively recent introgressed lineages that resulted from admixture between closely related lineages. To systematically explore the effect of

hybridization on Ty transposition rates, we performed a large-scale mutation accumulation (MA) experiment on 9 artificial hybrid backgrounds, six of which were previously published [68]. A MA experiment with yeast consists of subjecting a population to periodic single-cell bottlenecks (typically during hundreds to thousands of mitotic generations) by streaking a randomly chosen single colony, thus amplifying genetic drift and making natural selection inefficient [69] (Fig 6a). *S. paradoxus* was inferred to reproduce mitotically 99.9% of the time in the wild [70], making MA experiments a reasonably realistic approximation of its natural life cycle. Pairs of haploid derivatives of natural *S. paradoxus* strains were chosen to span a range of intraspecific evolutionary divergence (Fig 6b), yielding four categories of crosses (from the least to the most divergent): *SpC*×*SpC* (CC), *SpB*×*SpB* (BB), *SpB*×*SpC* (BC) and *SpB*×*SpA* (BA; Fig 6c, Tables S4-6). The crosses used also spanned a range of variation in parental Ty CNs (Fig 6d). For each cross, 48 to 96 MA lines derived from independent matings were propagated for ~770 mitotic divisions.

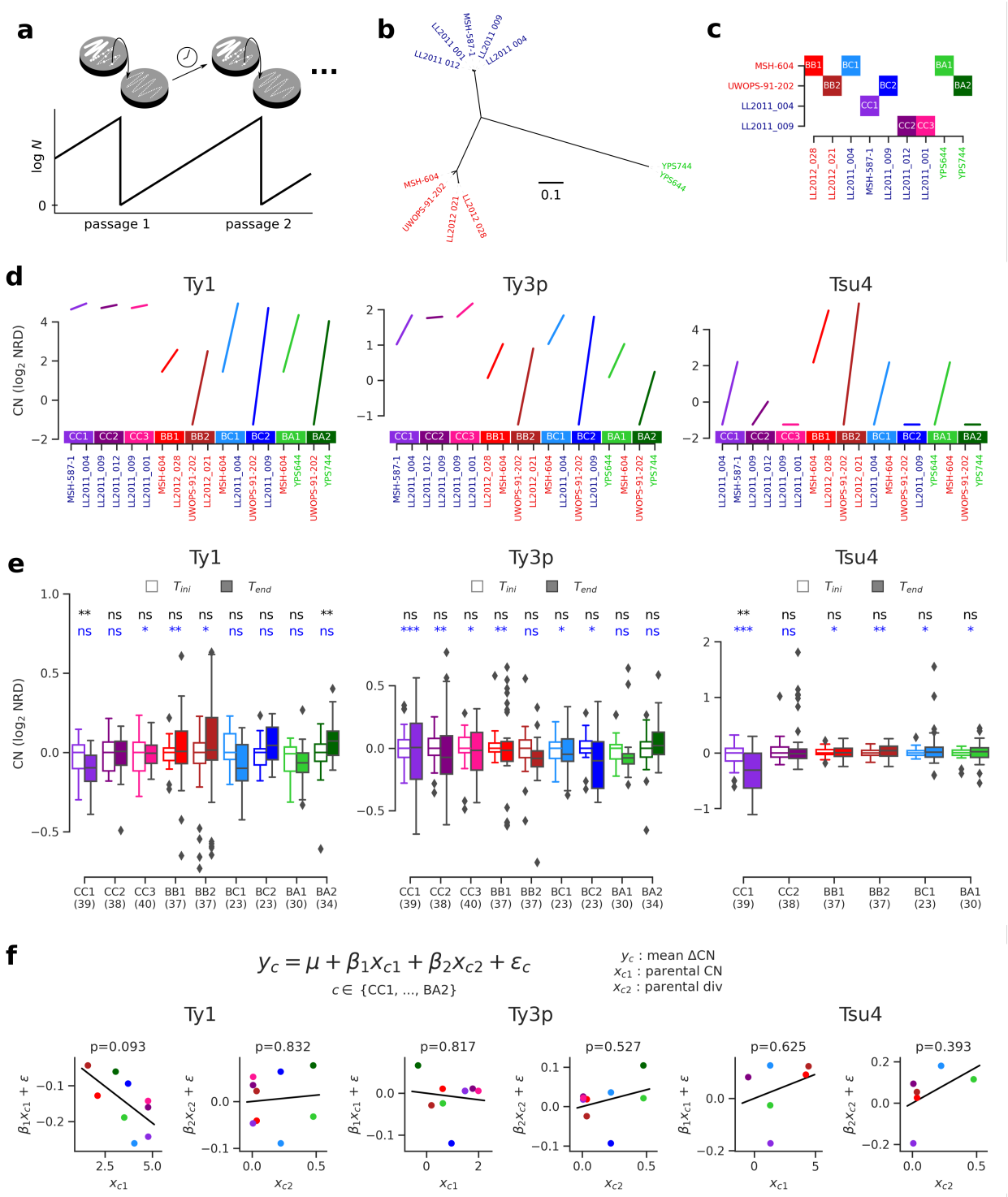

Fig 6. Ty CN variation in MA lines reveal no effect of parental divergence and major genotype-specific effects. a. Principle of a MA experiment on a yeast population. *N*

corresponds to population size. **b.** Bayesian phylogenetic tree showing nucleotide divergence between the parental haploid strains used to generate artificial hybrids. **c.** Design of the crosses for the MA experiment. **d.** Ty CNs in the haploid parental strains of each cross in c. **e.** CN at the onset (T_{ini}, empty boxes) and at the end (T_{end}, full boxes) of the MA experiment. Whiskers span 1.5 times the interquartile range. FDR-corrected p-values of Wilcoxon signed-rank tests and Brown–Forsythe tests for the equality of variances are shown in black and blue, respectively. ns: $p \geq 0.05$, *:$p<0.05$, **:$p<0.01$, ***:$p<0.001$. **f.** Linear models explaining the mean variation in CN ($\Delta CN = CN_{Tend}-CN_{Tini}$) for a given cross as a function of the joint CN and genetic divergence of its parental strains. The p-value for each term in the models is shown. Scatter plots (CCPR plots) show the partial residuals for each independent variable, along with the fitted regression.

We selected a random subset of 23 to 40 MA lines per cross and performed whole-genome resequencing on stocks frozen at T_{ini} and T_{end}, totaling 301 strains sequenced at each timepoint (PRJNA515073). All these strains were confirmed to be diploid at T_{ini} and T_{end} by quantifying total DNA content using flow cytometry ([68], Marsit *et al.* in prep). Focusing on full-length Ty1, Ty3p and Tsu4, no global trend emerged from the variation of NRD before and after MA (Fig 6e, Fig S16). Mixed effects linear models fitting cross-specific intercepts and slopes showed no statistically significant support for a global effect of time on CNs. Nevertheless, specific combinations of crosses and Ty families showed significant differences in either central tendency or variance. Often, crosses within a category of parental strain divergence showed contrasting patterns. For example, although CC1, CC2 and CC3 are genetically very similar *SpC*×*SpC* crosses, only CC1 exhibits a significant decrease in Ty1 and Tsu4 CNs. There were also cases of opposite trends within the same category of cross, for instance Ty1 CN in BC and BA crosses. This showed that individual parental genotypes, even if they are evolutionarily closely related, can have very different dynamics, both in sign and magnitude.

We next sought to understand how these dynamics are determined by properties of the different hybrid genotypes tested. We fitted linear models to the variation in CN between T_{ini} and T_{end} (ΔCN), using the sum of CNs of both parental strains of a cross and the evolutionary divergence between them as independent variables. Although these models explain little global variance (R^2=0.41, 0.15 and 0.27 for Ty1, Ty3p and Tsu4 respectively), we found that neither independent variable was a significant predictor of ΔCN (Fig 6f). Although all regression slopes were slightly positive with respect to levels of evolutionary divergence, none of them was significant, indicating that divergence between parents of a hybrid has no large influence on its CN dynamics. On the other hand, Ty1 ΔCN had a negative relationship with parental Ty1 CN and was marginally significant (p=0.093), suggesting that the initial CN of a hybrid genotype can dictate the sign and magnitude of its CN change. However, this result should be interpreted with care as some of the crosses shared common parental strains and were thus not entirely independent. This relatedness may in fact explain part of the pattern seen for Ty1. Crosses that shared the *SpB* parental strain UWOPS-91-202 (BB2, BC2, BA2) seemed to have systematically higher Ty1 ΔCN values than the corresponding crosses that shared the *SpB* parental strain MSH-604 (BB1, BC1, BA1). UWOPS-91-202 appeared to be a naturally Ty-less strain, as no full-length element was found for any family in its genome (Fig 6d). Previous reports of such Ty-less strains were shown to be permissive to the accumulation of Ty elements when assayed in the laboratory using a reporter assay [53,71]. This observation also supports the hypothesis that genotype-specific characteristics can be major determinants of variation in Ty CN.

Discussion

The literature on TE reactivation in hybrids highlights its relevance to the field of speciation [72]. Reactivation can be a source of postzygotic reproductive isolation by compromising genome stability [73], which translates into a spectrum of deleterious hybrid phenotypes spanning sterility, developmental abnormalities and inviability. Many studies investigated the TE reactivation hypothesis at the interspecific level and yielded equivocal evidence. For instance, mixed evidence for TE reactivation was found within the *Drosophila* genus [11,16,17,74], where reactivation seems to depend on the species pair involved. On the other hand, three well-established hybrid species that emerged from parallel hybridization between the same species pair in the genus *Helianthus* all exhibit important reactivation of LTR retrotransposons, both at the genomic [15,20] and transcriptomic [75] levels. However, a study of hybrid zones between the two parental species revealed that no reactivation is apparent in recent hybrid populations, casting doubt on the idea that hybridization alone triggered reactivation and suggesting the involvement of other factors like abiotic stress [18]. Overall, this data nuances the hypothesis that hybridization could act as a systematic trigger of TE proliferation and reproductive isolation.

In this study, we used population genomic data on North American lineages of the yeast *S. paradoxus* to test whether TEs were reactivated in natural hybrids. It was previously shown that these lineages experienced two hybridization events that resulted in the establishment of admixed lineages that have partial postzygotic reproductive isolation with their respective parental lineages [38,39]. We found no evidence of TE reactivation in any of these natural hybrid lineages. Our data showed that while some hybrid lineages may have relatively high CNs and exhibit evolutionary dynamics consistent with recent expansions, these signatures were never more pronounced than that of their parental lineages (*SpC* in the case of *SpC** and *SpC** in the case of SpD_1 and SpD_2). Moreover, our analyses suggest that environmental and

demographic factors are unlikely to be major factors affecting TE CNs in natural lineages, which therefore likely reflect the underlying transposition rates. Our observations would seem consistent with a role of TE reactivation in speciation, as natural hybrids with high levels of reactivation would likely be unobserved. However, they contrast with long-established hybrid species in *Helianthus*, showing that reactivation and long-term maintenance can be compatible.

Many factors are susceptible to determine whether a hybrid genotype exhibits reactivation or not. According to the Bateson-Dobzhansky-Muller (BDM) model, genetic incompatibilities should accumulate as populations diverge [30]. Additionally, the evolutionary conflict between host genomes and TEs results in an arms race that can accelerate the evolution of the involved genes [76,77]. Thus, it can be expected that incompatibilities affecting transposition rates are more likely to be observed in hybrids between highly diverged genotypes compared to closely related ones. Here, we report data from mutation accumulation (MA) experiments on artificial hybrids that probe the relationship between evolutionary divergence and transposition rates. We show that contrary to the expectation from the BDM model, hybrids between more divergent genotypes exhibit no deregulation of TEs. Instead of reactivation, many hybrid genotypes exhibited marked decreases in TE CNs after evolution. This trend is especially important for Ty3p, as seven out of nine genotypes exhibit a negative change in median CN. Ty3p is also the least abundant Ty family in natural lineages, as it is found in rarely more than ~5 copies and extinct in all hybrid strains sampled. Such low abundance families are likely more prone to extinction, as stochastic fluctuation in CN would more frequently lead to the loss of all copies. Stochastic loss of Ty copies may come from intra-element LTR recombination or loss of heterozygosity at a locus where a Ty element is present in one parent only, although the mechanistic causes of CN decrease remain to be investigated. This result also seems consistent with the reported Ty CN decrease in mitotically propagated experimental populations

of *S. cerevisiae* [78], suggesting a potential intrinsic bias towards TE loss. However, it should be noted that the latter study used haploid populations in a regime with efficient natural selection [79], thus mitigating comparisons with our results.

Transposition is a complex cellular process, with multiple steps regulated by both TEs and host genomes. Genetic screens in *S. cerevisiae* revealed that genes involved in multiple cellular functions (chromatin organization, transcription, translation, RNA and protein metabolism) affect Ty1 transposition rate, either positively or negatively (reviewed in [43]). Thus, it is possible that divergent hybrids suffer from misregulation at different steps of the transposition process that have antagonistic effects, such that they cancel out at the level of transposition rates. It is also possible that such antagonistic interactions exist between transposition and excision rates, yielding no net change in CN. Answering these questions will require a thorough investigation of novel Ty insertion loci in the MA lines, which is beyond the scope of this study. Transposition regulation being a polygenic trait with a large mutational target is also consistent with the strong genotype specificity we observed in our artificial hybrids. Some of the contrasted dynamics between crosses that share high genetic similarity might originate from cryptic genetic variation in genes involved in Ty regulation, whose effect on transposition rate would be uncovered in hybrids.

TEs are structurally very diverse and populate the genomes of a vast taxonomic range, and the control mechanisms in response to their mobilization may exhibit a correspondingly large diversity. In many eukaryotes, TEs are repressed epigenetically through mechanisms that include DNA methylation, chromatin modification and mRNA degradation, processes in which small RNAs play central roles [80,81]. In many fungi, TE repression acts via multiple mechanisms involving RNA interference (RNAi) or cytosine methylation, including repeat-induced point mutation (RIP) [82]. Notably, the genomes of introgressed species of the fungal

genus *Coccidioides* harbor signatures of efficient RIP [83], suggesting that this mechanism may remain active in the face of hybridization. Unlike most eukaryotes, *S. cerevisiae* and its closely related species lost the genes of the RNAi pathway [84] and have no cytosine methylation [85]. Instead, *S. cerevisiae* Ty1 has a potent self-repression mechanism, which strength depends on the actual Ty1 CN in the host genome (thus termed CN control, CNC) [71]. CNC is a post-transcriptional mechanism that relies on p22, an N-truncated form of the Gag (p49) Ty1-encoded capsid protein which is translated from an alternative transcript of Ty1 [52]. p22 inhibits transposition by interfering with the assembly of the cytoplasmic virus-like particles that are essential to the replication cycle of Ty1 [52]. This unique repression mechanism may explain some of the patterns we observed in the artificial hybrids. In particular, our data suggests a negative relationship between the initial Ty1 CN in hybrid genotypes and the change in Ty1 CN between the onset and the end of the MA experiment. This pattern seems consistent with an efficient action of CNC, although our results would then suggest that CNC efficiency is largely independent of the degree of evolutionary divergence between the subgenomes of a hybrid. This is surprising given the recent finding that sequence divergence in p22 has dramatic impacts on the action of CNC [53]. Thus, further experimental work is needed to disentangle the mechanistic contributions of Ty1 CN and sequence divergence in the regulation of transposition rates in *Saccharomyces paradoxus* hybrids.

In conclusion, our work illustrates how the use of alternative model systems can bring novel insights on long-standing evolutionary hypotheses. Our results show that natural hybrid lineages of the wild yeast *S. paradoxus* exhibit no TE reactivation and suggest that dynamics of TE content evolution in hybrids depend on deterministic factors that evolutionary divergence between parents can hardly predict.

Materials and Methods

General tools

GNU Parallel version 20170922 [86] and SnakeMake version 5.5.4 [87] were used extensively in the following analyses. Unless otherwise stated, Python version 3.7.6 [88] was used for the custom scripts used in the following analyses. These scripts were run on JupyterLab version 2.0.1 [89] and made extensive use of the following packages: BioPython version 1.76 [90], DendroPy version 4.4.0 [91], Matplotlib version 3.2.0 [92], NetworkX version 2.4 [93], NumPy version 1.18.1 [94], pandas version 1.0.2 [95], SciPy version 1.4.1 [96], seaborn version 0.10.0 [97].

Protein-coding genes annotation and orthogroup definition

Whole-genome *de novo* assemblies were previously generated for six North American *S. paradoxus* strains using a combination of long (Oxford Nanopore) and short (Illumina) reads (PRJNA514804). Protein-coding genes were predicted *ab initio* using AUGUSTUS version 3.3.1 with parameters species=saccharomyces genemodel=partial singlestrand=false. Predicted ORF sequences for the six assemblies were combined with ORF annotated from the *S. cerevisiae* strains S288c and YPS128 from [48] as references. Protein-coding genes orthogroups were inferred from sequence similarity and synteny using the program SynChro version Jan2015 [98] with parameters a=0 DeltaRBH=1. Orthogroups were defined as genes with one-to-one relationships across all pairwise comparisons between the eight genomes. Protein-coding genes systematic names were propagated from S288c to the corresponding orthogroups.

Annotation of Ty LTR retrotransposons

Ty LTR retrotransposons were annotated from the six whole-genome assemblies using RepeatMasker version open-4.0.7 [99] with parameters -s -low -gccalc and the library of

reference LTR and internal sequences assembled by [45] (containing Ty1, Ty2, Ty3, Ty4 and Ty5 from *S. cerevisiae* and Ty3p from *S. paradoxus*), additionned with Tsu4 from *S. uvarum* [46]. RepeatMasker output was defragmented using REannotate version 26.11.2007 [100] with parameters -s 15000 -d 10000, yielding defragmented elements classified as full-length, solo LTR, truncated or orphan internal sequences. The -f parameter was used to allow defragmentation of Ty1/Ty2 hybrid elements following [45]. The defragmented annotations were manually curated to split defragmented elements spanning entire protein-coding genes or forming concatemers.

Network-based Ty LTR sequence similarity analysis

We analyzed the sequence diversity of LTR annotations within each of our six *S. paradoxus* assemblies using a network approach [101]. The sequences of all LTR annotations within a genome were extracted, All-against-all search was performed between LTR sequences in a genome using BLASTN version 2.7.0 [102] with parameters word_size 7 dust no. Networks were built using the Python package Networkx version 2.4 [93] by keeping hits with a bit score higher or equal to the 10th percentile of all bit scores for a given LTR. Visualization of the resulting networks was done using Cytoscape version 3.7.1 [103] using the Prefuse Force Directed layout algorithm. Inspection of these networks justified the merging of the following LTR families: Ty1-Ty2, Ty3p-Ty3 and Tsu4-Ty4.

Estimation of evolutionary divergence since transposition

Evolutionary divergence since transposition was estimated from inter-element LTR sequence identity, assuming that the LTRs of a newly inserted element are identical to those of the source element. From the sequence similarity graphs, the closest relative of each element was retrieved by keeping the hit with the highest bit score. Hits between the two LTRs of full-length

elements were excluded, and only one LTR per full-length element was considered. For each element, the evolutionary divergence since transposition was extracted as the percent sequence identity to the closest relative.

Definition of Ty orthogroups from genome assemblies

Orthology of LTR retrotransposons in the six whole-genome assemblies was assessed as follows. Whole-genome multiple sequence alignment (MSA) with rearrangements was performed using Mugsy version 1r2.3 [104] with parameters -d 20000 -nucmeropts "-b 1000 -l 15" --fullsearch --refine mugsy. MSA blocks for which at least one genome had no corresponding sequence were trashed. The remaining alignment blocks were scanned in 10kb sliding, non-overlapping windows in both directions and windows were trimmed until at least 80% of sites were informative in a MSA window. Ty annotation coordinates were transformed into an MSA coordinates system and clustered using the DBSCAN algorithm implemented in the Python package scikit-learn version 0.22.2 with parameters min_samples=1 eps=500. This one-dimensional clustering approach allowed to define orthologous clusters containing at least one Ty annotation in at least one genome.

We defined a set of confident orthogroups to derive an empirical distribution of physical distance between start and end coordinates of orthologs. Clusters were selected if they contained at most one annotation per genome and if all annotations were on the same strand and belonged to the same family. Pairwise BLASTN searches within each cluster were used to keep elements that produced at least one hit with ≥80% nucleotide identity and an alignment length between 90% and 110% of both query and subject lengths. From the 36 resulting confident orthogroups, we extracted the distribution of differences between annotation start and end coordinates and used the 95th percentile of that distribution (17 bp) as threshold for the following orthogroup definition procedure. Within each cluster, pairs of annotations meeting the following two criteria

were considered orthologous: the pair had to include at least one best BLASTN hit, and either start or end coordinates (or both) had to be within the defined threshold of 17 nt. This method allowed to define 720 Ty orthogroups, 20 of which (3%) had >1 element for at least one genome and were discarded.

Phylogenetic analysis of Ty1 evolution

For each of our six genome assemblies, we extracted the LTR sequence of all Ty1 annotations, keeping only one LTR at random for full-length elements. As an outgroup, a conserved Ty4 LTR orthogroup was added on the basis of minimal variance in start and end coordinates across genomes and minimal difference between the length of the reference Ty4 sequence and the average length of the annotations. MSAs were computed using MUSCLE version 3.8.31 [105] and filtered to maximize the proportion of informative sites while keeping the outgroup, first discarding sequences with >25% gaps and then removing sites with ≥25% gaps. Bayesian phylogenetic trees were computed using MrBayes version 3.2.5 [106] with model parameters Nst=6 Rates=invgamma. Markov chain Monte Carlo parameters were set to 10 million generations for two runs of four chains each with 25% burn-in fraction and temperature parameter of 0.03 to improve the efficiency of Metropolis coupling. Bayesian phylogenetic trees were converted into ultrametric phylogenetic trees using PATHd8 [107] and rooted at the Ty4 LTR outgroup. The branching time of each internal node in the ultrametric trees was computed as the distance from the root. Branching events were counted as the number of children nodes for each internal node, as the bayesian phylogenetic reconstructions yielded many polytomies at the base and tips of the trees.

Construction of a library of reference Ty sequences for S. paradoxus

We combined the Ty LTR sequences of all families and full-length sequences of Ty1, Ty3p and

Tsu4 from the defragmented annotations of our six *S. paradoxus* assemblies and aligned the sequences of each family and type using MUSCLE. The resulting MSAs were filtered to remove sequences and sites that had ≥70% gaps. Consensus sequences from those MSAs were computed using EMBOSS cons version 6.6 [108] and the annotation sequence with the lowest BLASTN E value against the resulting consensus was included in the reference library. LTRs of the full-length reference elements were manually trimmed using MEGA version 7.0.26 [109]. The *S. paradoxus* sequences were complemented with *S. cerevisiae* internal sequences for Ty2, Ty3, Ty4 and Ty5 [45].

Ty insertion allele calling and population structure analysis

Illumina PE 150 libraries for 208 wild *S. paradoxus* strains were downloaded from NCBI (PRJNA277692, PRJNA324830, PRJNA479851) and trimmed using Trimmomatic version 0.36 [110] with parameters ILLUMINACLIP:combined_adapters.fa:2:30:10 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 and a custom library of Illumina adapter sequences. Trimmed reads were mapped against the masked version of our *SpB* genome assembly (strain MSH-604, from the RepeatMasker output) using bwa mem version 0.7.16a-r1181 [111] and sorted using samtools sort version 1.9 [112]. Duplicates were removed using Picard RemoveDuplicates version 2.18.5-SNAPSHOT [113] with parameter REMOVE_DUPLICATES=true. RetroSeq version 1.5 [114] was used to call Ty insertion alleles from discordant pairs of mapped reads. The -discover command was used with parameter -eref and our custom reference Ty sequences library. The -call command was run with parameter -soft. Visual inspection of the resulting calls revealed several instances of multiple, tightly clustered calls of the same family, likely reflecting a single true positive call. To minimize false positives, the calls were transformed into MSA coordinates and clustered using DBSCAN with parameters min_samples=1 eps=500. Resulting clusters were considered true positives if they

contained at least one original call of high-quality (flag 8). This strategy provided the best balance between false positive rates and false negative rates in an analysis made with various simulated short-read datasets (data not shown). Insertion allele frequency spectra were computed within each lineage. Principal component analysis (PCA) was performed on Ty alleles using the Python package scikit-learn version 0.22.2.post1.

Population structure based on genome-wide SNPs

Trimmed reads from the 208 wild strains were mapped against the unmasked version of our *SpB* genome assembly (strain MSH-604) using bwa mem version 0.7.16a-r1181 and sorted using samtools sort. Duplicates were removed using Picard RemoveDuplicates with parameter REMOVE_DUPLICATES=true. Variants were called using freebayes version 1.3.1-17-gaa2ace8 [115]. Variants were filtered using bcftools view version 1.9 [112] with parameter -e 'QUAL<=20', along with the tools vcfallelicprimitives, vcfbreakmulti and vcffilter with parameter -f 'TYPE = snp' from vcflib version 1.0.0_rc2 [116]. PCA on genome-wide SNPs was performed using the Python package scikit-allel version 1.2.1 [117], keeping only biallelic SNPs and excluding singletons. From the filtered SNPs of the 208 wild strains, 10000 SNPs were randomly selected using downsamplevcf version cf73d1a8 from the JVarkit library [118] and converted to NEXUS format using vcf2phylip version 2.0 [119]. A bayesian phylogenetic tree was computed using MrBayes with model parameters Nst=6 Rates=invgamma. Markov chain Monte Carlo parameters were set to 1 million generations for two runs of four chains each with 25% burn-in fraction and temperature parameter of 0.1.

Depth of coverage of wild strains on Ty reference sequences

Trimmed reads from the 208 wild strains were mapped against our custom reference Ty sequences library using bwa mem version 0.7.16a-r1181 and sorted using samtools sort version

1.9. Only internal sequences extracted from full-length reference sequences were used. Depth of coverage was extracted using samtools depth. Median depth values of Ty elements were transformed in $\log_2$ and normalized by the $\log_2$ genome-wide depth computed from the mapped bases on the genome-wide mappings previously described. Kruskal-Wallis tests were performed on normalized $\log_2$ depth values for families with full-length elements in natural lineages (Ty1, Ty3p, Tsu4). Conover post-hoc tests were performed to test for pairwise differences between lineages using the Python package scipy-posthocs version 0.6.3 [120].

Climatic data

We used the TerraClimate climatic dataset [56] (accessed in October 2019), which incorporates high spatial and temporal resolution climatic interpolation data from primary and derived climatic variables. For the sampling locations of our 208 wild strains, we aggregated the 12 climatic variables over the time period from 1958 to 2015 and computed the mean, standard deviation, minimum and maximum values over time for a given variable variable at a given location. The resulting dataset was normalized into z-scores and summarized by PCA using scikit-learn.

Association analysis between Ty CNs, population structure and climatic variables

SNP-based population structure PCAs and climatic variable PCAs were combined in linear models as predictors of CNs for the three active families Ty1, Ty3p and Tsu4. In each case, PCs explaining a cumulative sum of 80% of the total variance in the dataset were kept and named gPCs (from PCAs on SNP data) or ePCs (from PCAs on climatic variables). Ordinary least squares models were fitted using the Python package statsmodels version 0.10.1 [121]. Two types of models were fitted: OLS using both gPCs and ePCs as predictors, and OLS using only gPCs as predictors. The quality of the two model types was compared using their AIC values. Strains were clustered according to their coordinates on gPCs and ePCs using the

AffinityPropagation algorithm from scikit-learn with parameters max_iter=10000 damping=0.8.

Nucleotide diversity at synonymous sites

Long-term effective population size for pure lineages was estimated by computing nucleotide diversity at synonymous sites in protein-coding genes. Only SNPs at four-fold degenerate sites in protein-coding genes (third positions of codons CTN, GTN, TCN, CCN, ACN, GCN, CGN and GGN) were considered. Nucleotide diversity was computed using the mean_pairwise_difference function from scikit-allel.

Growth assay in $NaAsO_2$

Three strains were chosen at random in each of the lineages *SpA*, *SpB* and *SpC** and two *SpC* strains were chosen at random, totalling 12 strains with the *SpC* strain LL2011_004. Precultures were grown in synthetic complete medium (SC; 2% glucose, 0.174% yeast nitrogen base [without amino acids without ammonium sulfate], 0.5% ammonium sulfate, 0.134% complete drop-out) until they reached a 595 nm optical density (OD595) of ~0.5-1.0. Precultures were diluted at an OD_{595} of 0.1 ml^{-1} in SC medium supplemented with 0, 0.4, 0.6, 0.8 and 1 mM $NaAsO_2$ in a single 96-well microplate, with border columns and rows filled with sterile SC medium. OD_{595} was monitored during 60 hours at 15 minutes intervals at 25°C in a Tecan Spark microplate reader. OD_{595} values were normalized by the mean of the sterile wells at each timepoint.

Properties of genes neighboring Ty insertions

The analysis of properties of neighboring genes considered Ty-centric immediate adjacency relationships with protein-coding genes, meaning that a single gene could be represented more than once if many Ty annotations are immediately adjacent to it. From the whole-genome MSAs, protein-coding genes adjacent to Ty annotations were retrieved on the basis of synteny

with the *S. cerevisiae* S288c genome. Genes annotated as essential in *S. cerevisiae* were extracted from YeastMine (accessed in August 2019).

The strength of purifying selection acting on individual genes was assessed at two levels: at the species-wide level (i.e. including all North-American *S. paradoxus* lineages) by fitting a single ω ratio (non-synonymous to synonymous substitution rates) per gene; and at the lineage level, by evaluating the relative strength of natural selection in the main pure lineages *SpB* and *SpC*. A randomly selected subset of 34 strains from the 208 wild strains were chosen. The VCF matrix from the SNP calling step above was used to derive the DNA sequences of protein-coding genes in these 34 strains. Neighbor-joining gene trees were produced using the nj function from the R package ape version 5.0 [122] and R version 3.4.3 [123]. ω ratios were fitted using codeml from the PAML suite version 4.8 [124] with parameters seqtype=1 model=0. Selection strength at the lineage level was evaluated using the RELAX framework from the HyPhy software suite version develop-0d598ad [67]. The RELAX general descriptive model was fitted using the RELAX-scan script (https://github.com/veg/hyphy-analyses/tree/master/RELAX-scan). From the model output, the fitted value of the *k* parameter (describing the strength of selection) was extracted for each branch of the gene trees. For *SpB* and *SpC*, an average *k* value was computed by selecting the clade containing all the strains of the corresponding lineage (from the random subset) and weighting *k* values by the length of the corresponding branch. The difference in selection intensity between *SpB* and *SpC* was defined as $\Delta k = k_{SpB}-k_{SpC}$. For the three genes with ORF-disrupting Ty insertions, the RELAX framework was fitted on gene trees comprising all 208 wild strains.

Mutation accumulation

Crosses BB, BC and BA were previously published [68]. CC crosses were performed following the same methodology. The *ADE2* gene was deleted in haploid *MATa hoΔ::kanMX4* or *MATa*

hoΔ::natMX4 derivatives of the *SpC* strains LL2011_001, LL2011_012 and MSH-587-1 using a *hphNT1* cassette that was amplified from plasmid pFA6a-hphNT1 [125] using primers CLOP97-F5 and CLOP97-F6 (Tables S3 and S4). The three strains were paired with the previously published *SpC* strains LL2011_004 *MATα hoΔ::kanMX4* and LL2011_009 *MATα hoΔ::natMX4* to generate three CC crosses by selection on medium containing 100 µg ml^{-1} geneticin and 10 µg ml^{-1} nourseothricin. For each cross, 64 MA lines were initiated by selecting single colonies from independent mating cultures that were spotted on selective medium. Diploid hybrids were propagated mitotically on Petri dishes containing solid YPD medium (1% yeast extract, 2% tryptone, 2% glucose, 2% agar) by streaking for single colonies. Every three days, a single colony was randomly chosen to inoculate a new plate by picking the colony that grew the closest to a predefined mark on the Petri plate. Colonies that had an obvious difference in red pigmentation were excluded to avoid the propagation of mitochondrial petite mutants. MA lines were propagated for 35 passages (approx. 770 mitotic generations), keeping glycerol stocks of all the lines every three passages.

Illumina library preparation for MA lines

For each cross, 23 to 40 MA lines were randomly chosen among the subset for which diploidy was confirmed at T_{ini} and T_{end} [68] (Marsit *et al.* in prep.). The glycerol stocks from passage 1 (T_{ini}) and passage 35 (T_{end}) were streaked on solid YPD medium. Single colonies were used to inoculate liquid YPD cultures. DNA was extracted from saturated cultures using QIAGEN Blood and Tissue kit with modifications for yeast cell wall digestion. DNA was treated with RNase A and purified on Axygen Axygen AxyPrep Mag PCR Clean-up SPRI beads. Genomic DNA libraries were prepared using the Illumina Nextera Tagmentase and Nextera index kits following a modified version of the protocol described in [126]. The quality of randomly selected libraries was controlled using an Agilent BioAnalyzer 2100 electrophoresis system. Sequencing was

performed at the CHU Sainte-Justine and Génome Québec Integrated Centre for Pediatric Clinical Genomics (Montreal, Canada) on an Illumina NovaSeq 6000 sequencer with SP flow cells.

Depth of coverage of MA lines on Ty reference sequences

Illumina PE 150 libraries for 301 MA lines and their 11 parental haploid strains (deposited at NCBI Sequence Read Archive under PRJNA515073) were trimmed using Trimmomatic version 0.33 with parameters ILLUMINACLIP:nextera.fa:6:20:10 MINLEN:40 and a library of Illumina Nextera adapter sequences. Trimmed reads were mapped against our *SpB* genome assembly (strain MSH-604) using bwa mem version 0.7.17 and sorted using samtools sort version 1.8. Duplicates were removed using Picard RemoveDuplicates version 2.18.29-SNAPSHOT with parameter REMOVE_DUPLICATES=true. We extracted the list of *S. cerevisiae* genes with paralogs from YeastMine (accessed in December 2018) [127] and randomly selected 50 genes that didn't figure in this list. The protein sequences of these genes were used as TBLASTN queries against our *SpB* genome assembly and the best hit loci were extracted. Depth of coverage over these loci was extracted using samtools depth version 1.9 and depth values were transformed in $\log_2$. Depth values were transformed in $\log_2$ and normalized by the $\log_2$ genome-wide depth computed from the mapped bases on the genome-wide mappings.

The trimmed reads of the 301 MA lines and their 11 parental haploid strains were mapped against our custom reference Ty sequences library using bwa mem version 0.7.16a-r1181 and sorted using samtools sort version 1.9. Only internal sequences extracted from full-length reference sequences were used. Median depth values of Ty elements were transformed in $\log_2$ and normalized by the $\log_2$ genome-wide computed from the mapped bases on the genome-wide mappings. Normalized $\log_2$ depth values for a given cross and Ty family were discarded if the median was lower than the 10th percentile of the normalized $\log_2$ depth values for non-

paralogous genes in the corresponding cross, thus considering the family absent from the initial MA genotypes. Similarly, the 10th percentile of normalized $\log_2$ depth values for non-paralogous genes in all crosses aggregated (-1.25) was used as a lower threshold for depth values of the parental haploid strains, replacing lower values by -1.25.

Phylogenetic analysis of the MA lines haploid parents

From the mappings of the MA lines parental haploid strains, variants were called using freebayes. Variants were filtered using bcftools view with parameter -e 'QUAL<=20 -g ^het', along with the tools vcfallelicprimitives, vcfbreakmulti and vcffilter with parameter -f 'TYPE = snp' from vcflib. From the filtered SNPs, 10000 SNPs were randomly selected using downsamplevcf from the JVarkit library and converted to NEXUS format using vcf2phylip. A bayesian phylogenetic tree was computed using MrBayes with model parameters Nst=6 Rates=invgamma. Markov chain Monte Carlo parameters were set to 1 million generations for two runs of four chains each with 25% burn-in fraction and temperature parameter of 0.1. From the resulting consensus tree, the evolutionary divergence between parental strains was measured as the sum of branch lengths connecting each pair.

Association analysis between Ty ΔCN, evolutionary divergence between parents and initial Ty CN

For each cross, the average ΔCN was analysed in a linear model incorporating parental evolutionary divergence and initial Ty CN of the hybrid genotypes as predictors. For each cross, the Ty CNs of parental haploid strains were summed and used as a measure of the initial Ty CN in all MA lines of the cross. An OLS model was fitted using statsmodels.

Acknowledgements

We thank Hélène Martin for contributing to sequencing data analyses and Philippe Després, Anne-Marie Dion-Côté, Marika Drouin, Alexandre Dubé, Clara Dubernard and Anna Fijarczik for reading and providing useful comments on draft versions of this manuscript.

References

1. Mallet J. Hybridization as an invasion of the genome. Trends in Ecology and Evolution. Elsevier Ltd; 2005. pp. 229–237. doi:10.1016/j.tree.2005.02.010

2. Mallet J, Besansky N, Hahn MW. How reticulated are species? Bioessays. 2016;38: 140–149.

3. Stukenbrock EH. The role of hybridization in the evolution and emergence of new fungal plant pathogens. Phytopathology. 2016;106: 104–112.

4. Rieseberg LH, Archer MA, Wayne RK. Transgressive segregation, adaptation and speciation. Heredity . 1999;83 (Pt 4): 363–372.

5. Langdon QK, Peris D, Baker EP, Opulente DA, Nguyen H-V, Bond U, et al. Fermentation innovation through complex hybridization of wild and domesticated yeasts. Nat Ecol Evol. 2019;3: 1576–1586.

6. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biol. 2018;19: 199.

7. McClintock B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A. 1950;36: 344–355.

8. McClintock B. The significance of responses of the genome to challenge. Science. 1984;226: 792–801.

9. Kidwell MG, Kidwell JF, Sved JA. Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: A Syndrome of Aberrant Traits Including Mutation, Sterility and Male Recombination. Genetics. 1977;86: 813–833.

10. Picard G. Non-mendelian female sterility in Drosophila melanogaster: hereditary transmission of I factor. Genetics. 1976;83: 107–123.

11. Labrador M, Farré M, Utzet F, Fontdevila A. Interspecific hybridization increases transposition rates of Osvaldo. Mol Biol Evol. 1999;16: 931–937.

12. O'Neill RJ, O'Neill MJ, Graves JA. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. Nature.

1998;393: 68–72.

13. Dion-Côté AM, Renaut S, Normandeau E, Bernatchez L. RNA-seq reveals transcriptomic shock involving transposable elements reactivation in hybrids of young lake whitefish species. Mol Biol Evol. 2014;31: 1188–1199.

14. Liu B, Wendel JF. Retrotransposon activation followed by rapid repression in introgressed rice plants. Genome. 2000;43: 874–880.

15. Ungerer MC, Strakosh SC, Zhen Y. Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. Curr Biol. 2006;16: R872–3.

16. Coyne JA. Mutation rates in hybrids between sibling species of Drosophila. Heredity . 1989;63 (Pt 2): 155–162.

17. Hey J. Speciation via hybrid dysgenesis: negative evidence from the Drosophila affinis subgroup. Genetica. 1988;78: 97–103.

18. Kawakami T, Dhakal P, Katterhenry AN, Heatherington CA, Ungerer MC. Transposable element proliferation and genome expansion are rare in contemporary sunflower hybrid populations despite widespread transcriptional activity of LTR retrotransposons. Genome Biol Evol. 2011;3: 156–167.

19. Göbel U, Arce AL, He F, Rico A, Schmitz G, De Meaux J. Robustness of transposable element regulation but no genomic shock observed in interspecific Arabidopsis hybrids. Genome Biol Evol. 2018;10: 1403–1415.

20. Ungerer MC, Strakosh SC, Stimpson KM. Proliferation of Ty3/gypsy-like retrotransposons in hybrid sunflower taxa inferred from phylogenetic data. BMC Biol. 2009;7: 1–13.

21. Charlesworth B, Charlesworth D. The population dynamics of transposable elements. Genet Res. 1983;42: 1–27.

22. Charlesworth B, Langley CH. The population genetics of Drosophila transposable elements. Annu Rev Genet. 1989;23: 251–287.

23. Charlesworth B, Sniegowski P, Stephan W. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature. 1994;371: 215–220.

24. Lynch M. The Origins of Genome Architecture. Sinauer Associates; 2007.

25. Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, Hannon GJ. An epigenetic role for maternally inherited piRNAs in transposon silencing. Science. 2008;322: 1387–1392.

26. Erwin AA, Galdos MA, Wickersheim ML, Harrison CC, Marr KD, Colicchio JM, et al. piRNAs Are Associated with Diverse Transgenerational Effects on Gene and Transposon Expression in a Hybrid Dysgenic Syndrome of D. virilis. PLoS Genet. 2015;11: e1005332.

27. Rubin GM, Kidwell MG, Bingham PM. The molecular basis of P-M hybrid dysgenesis: the

nature of induced mutations. Cell. 1982;29: 987–994.

28. Bucheton A, Paro R, Sang HM, Pelisson A, Finnegan DJ. The molecular basis of I-R hybrid dysgenesis in Drosophila melanogaster: identification, cloning, and properties of the I factor. Cell. 1984;38: 153–163.

29. Josefsson C, Dilkes B, Comai L. Parent-dependent loss of gene silencing during interspecies hybridization. Curr Biol. 2006;16: 1322–1328.

30. Orr HA. The population genetics of speciation: The evolution of hybrid incompatibilities. Genetics. 1995;139: 1805–1813.

31. Parhad SS, Theurkauf WE. Rapid evolution and conserved function of the piRNA pathway. Open Biol. 2019;9: 180181.

32. Lockton S, Gaut BS. The evolution of transposable elements in natural populations of self-fertilizing Arabidopsis thaliana and its outcrossing relative Arabidopsis lyrata. BMC Evol Biol. 2010;10: 10.

33. Lockton S, Ross-Ibarra J, Gaut BS. Demography and weak selection drive patterns of transposable element diversity in natural populations of Arabidopsis lyrata. Proc Natl Acad Sci U S A. 2008;105: 13965–13970.

34. Dolgin ES, Charlesworth B, Cutter AD. Population frequencies of transposable elements in selfing and outcrossing Caenorhabditis nematodes. Genet Res . 2008;90: 317–329.

35. Ye Z, Xu S, Spitze K, Asselman J, Jiang X, Ackerman MS, et al. A New Reference Genome Assembly for the Microcrustacean Daphnia pulex. G3 . 2017;7: 1405–1416.

36. Lynch M, Conery JS. The Origins of Genome Complexity. Science. 2003;302: 1401–1404.

37. Hénault M, Eberlein C, Charron G, Durand É, Nielly-Thibault L, Martin H, et al. Yeast Population Genomics Goes Wild: The Case of Saccharomyces paradoxus. Springer, Cham; 2017. pp. 1–24.

38. Leducq JB, Nielly-Thibault L, Charron G, Eberlein C, Verta JP, Samani P, et al. Speciation driven by hybridization and chromosomal plasticity in a wild yeast. Nature Microbiology. 2016;1: 1–10.

39. Eberlein C, Hénault M, Fijarczyk A, Charron G, Bouvier M, Kohn LM, et al. Hybridization is a recurrent evolutionary stimulus in wild yeast speciation. Nat Commun. 2019;10: 923.

40. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8: 973–982.

41. Krupovic M, Blomberg J, Coffin JM, Dasgupta I, Fan H, Geering AD, et al. Ortervirales: New Virus Order Unifying Five Families of Reverse-Transcribing Viruses. J Virol. 2018;92. doi:10.1128/JVI.00515-18

42. Parket A, Inbar O, Kupiec M. Recombination of Ty elements in yeast can be induced by a double-strand break. Genetics. 1995;140: 67–77.

43. Curcio MJ, Lutz S, Lesage P. The Ty1 LTR-Retrotransposon of Budding Yeast, Saccharomyces cerevisiae. Microbiology Spectrum. 2015;3: MDNA3–0053–2014.

44. Bleykasten-Grosshans C, Friedrich A, Schacherer J. Genome-wide analysis of intraspecific transposon diversity in yeast. BMC Genomics. 2013;14: 399.

45. Carr M, Bensasson D, Bergman CM. Evolutionary Genomics of Transposable Elements in Saccharomyces cerevisiae. Stajich JE, editor. PLoS One. 2012;7: e50978.

46. Bergman CM. Horizontal transfer and proliferation of Tsu4 in Saccharomyces paradoxus. Mob DNA. 2018;9: 18.

47. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 Genes Published by : American Association for the Advancement of Science Linked references are available on JSTOR for this article : Life with 6000 Genes. Science. 1996;274: 563–567.

48. Yue JX, Li J, Aigrain L, Hallin J, Persson K, Oliver K, et al. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. Nat Genet. 2017;49: 913–924.

49. Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. Genome Res. 1998;8: 464–478.

50. Naumov GI, Naumova ES, Lantto RA, Louis EJ, Korhola M. Genetic homology betweenSaccharomyces cerevisiae and its sibling speciesS. paradoxus andS. bayanus: Electrophoretic karyotypes. Yeast. 1992;8: 599–612.

51. Bergman CM, Bensasson D. Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in Drosophila melanogaster. Proc Natl Acad Sci U S A. 2007;104: 11340–11345.

52. Saha A, Mitchell JA, Nishida Y, Hildreth JE, Ariberre JA, Gilbert WV, et al. A trans-dominant form of Gag restricts Ty1 retrotransposition and mediates copy number control. J Virol. 2015;89: 3922–3938.

53. Czaja W, Bensasson D, Ahn HW, Garfinkel DJ, Bergman CM. Evolution of Ty1 copy number control in yeast by horizontal transfer and recombination. PLoS Genet. 2020;16: e1008632.

54. Paquin CE, Williamson VM. Temperature effects on the rate of Ty transposition. Science. 1984;226: 53–55.

55. Morillon A, Springer M, Lesage P. Activation of the Kss1 invasive-filamentous growth pathway induces Ty1 transcription and retrotransposition in Saccharomyces cerevisiae. Mol Cell Biol. 2000;20: 5766–5776.

56. Abatzoglou JT, Dobrowski SZ, Parks SA, Hegewisch KC. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958-2015. Sci Data. 2018;5: 170191.

57. Wilke CM, Adams J. Fitness effects of Ty transposition in Saccharomyces cerevisiae. Genetics. 1992;131: 31–42.

58. Gresham D, Desai MM, Tucker CM, Jenq HT, Pai DA, Ward A, et al. The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. PLoS Genet. 2008;4. doi:10.1371/journal.pgen.1000303

59. Filteau M, Hamel V, Pouliot MC, Gagnon-Arsenault I, Dube AK, Landry CR. Evolutionary rescue by compensatory mutations is constrained by genomic and environmental backgrounds. Mol Syst Biol. 2015;11: 832.

60. Hof AEV t., Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, et al. The industrial melanism mutation in British peppered moths is a transposable element. Nature. 2016;534: 102–105.

61. Aminetzach YT, Macpherson JM, Petrov DA. Pesticide resistance via transposition-mediated adaptive gene truncation in Drosophila. Science. 2005;309: 764–767.

62. Coney LR, Roeder GS. Control of yeast gene expression by transposable elements: maximum expression requires a functional Ty activator sequence and a defective Ty promoter. Mol Cell Biol. 1988;8: 4009–4017.

63. Todeschini A-L, Morillon A, Springer M, Lesage P. Severe Adenine Starvation Activates Ty1 Transcription and Retrotransposition in Saccharomyces cerevisiae. Mol Cell Biol. 2005;25: 7459–7472.

64. Palecek SP, Parikh AS, Kron SJ. Genetic analysis reveals that FLO11 upregulation and cell polarization independently regulate invasive growth in Saccharomyces cerevisiae. Genetics. 2000;156: 1005–1023.

65. Wysocki R, Bobrowicz P, Ułaszewski S. The Saccharomyces cerevisiae ACR3 gene encodes a putative membrane protein involved in arsenite transport. J Biol Chem. 1997;272: 30061–30066.

66. Mukhopadhyay R, Rosen BP. Saccharomyces cerevisiae ACR2 gene encodes an arsenate reductase. FEMS Microbiol Lett. 1998;168: 127–136.

67. Wertheim JO, Murrell B, Smith MD, Pond SLK, Scheffler K. RELAX: Detecting relaxed selection in a phylogenetic framework. Mol Biol Evol. 2015;32: 820–832.

68. Charron G, Marsit S, Hénault M, Martin H, Landry CR. Spontaneous whole-genome duplication restores fertility in interspecific hybrids. Nat Commun. 2019;10. doi:10.1038/s41467-019-12041-8

69. Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, et al. A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci U S A. 2008;105:

9272–9277.

70. Tsai IJ, Bensasson D, Burt A, Koufopanou V. Population genomics of the wild yeast Saccharomyces paradoxus: Quantifying the life cycle. Proceedings of the National Academy of Sciences. 2008;105: 4957–4962.

71. Garfinkel DJ, Nyswaner K, Wang J, Cho JY. Post-transcriptional cosuppression of Ty1 retrotransposition. Genetics. 2003;165: 83–99.

72. Serrato-Capuchina A, Matute DR. The role of transposable elements in speciation. Genes . 2018;9. doi:10.3390/genes9050254

73. Dion-Côté A-M, Barbash DA. Beyond speciation genes: an overview of genome stability in evolution and speciation. Curr Opin Genet Dev. 2017;47: 17–23.

74. Vela D, Fontdevila A, Vieira C, Pilar García Guerreiro M. A genome-wide survey of genetic instability by transposition in Drosophila hybrids. PLoS One. 2014;9. doi:10.1371/journal.pone.0088992

75. Renaut S, Rowe HC, Ungerer MC, Rieseberg LH. Genomics of homoploid hybrid speciation: diversity and transcriptional activity of long terminal repeat retrotransposons in hybrid sunflowers. Philos Trans R Soc Lond B Biol Sci. 2014;369: 20130345–20130345.

76. Van Valen L. A new evolutionary law. Evolutionary Theory. 1973;1: 1–30.

77. Blumenstiel JP, Erwin AA, Hemmer LW. What Drives Positive Selection in the Drosophila piRNA Machinery? The Genomic Autoimmunity Hypothesis. Yale J Biol Med. 2016;89: 499–512.

78. Bast J, Jaron KS, Schuseil D, Roze D, Schwander T. Asexual reproduction reduces transposable element load in experimental yeast populations. Elife. 2019;8. doi:10.7554/eLife.48548

79. McDonald MJ, Rice DP, Desai MM. Sex speeds adaptation by altering the dynamics of molecular evolution. Nature. 2016;531: 233–236.

80. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 2007;8: 272–285.

81. Lisch D. Epigenetic regulation of transposable elements in plants. Annu Rev Plant Biol. 2009;60: 43–66.

82. Gladyshev E. Repeat-Induced Point Mutation and Other Genome Defense Mechanisms in Fungi. Microbiol Spectr. 2017;5. doi:10.1128/microbiolspec.FUNK-0042-2017

83. Neafsey DE, Barker BM, Sharpton TJ, Stajich JE, Park DJ, Whiston E, et al. Population genomic sequencing of Coccidioides fungi reveals recent hybridization and transposon control. Genome Res. 2010;20: 938–946.

84. Drinnenberg IA, Weinberg DE, Xie KT, Mower JP, Wolfe KH, Fink GR, et al. RNAi in

budding yeast. Science. 2009;326: 544–550.

85. Proffitt JH, Davie JR, Swinton D, Hattman S. 5-Methylcytosine is not detectable in Saccharomyces cerevisiae DNA. Mol Cell Biol. 1984;4: 985–988.

86. Tange O. GNU Parallel: the command-line power tool. ;login: The USENIX Magazine. 2011;36: 42–47.

87. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012;28: 2520–2522.

88. Van Rossum G, Drake FL. Python 3 Reference Manual: (Python Documentation Manual Part 2). CreateSpace Independent Publishing Platform; 2009.

89. Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, Carol Willing, Jupyter Development Team. Jupyter Notebooks – a publishing format for reproducible computational workflows. In: Fernando Loizides BS, editor. Proceedings of the 20th International Conference on Electronic Publishing. IOS Press; 2016. pp. 87–90.

90. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25: 1422–1423.

91. Sukumaran J, Holder MT. DendroPy: a Python library for phylogenetic computing. Bioinformatics. 2010;26: 1569–1571.

92. Hunter JD. Matplotlib: A 2D Graphics Environment. Computing in Science Engineering. 2007;9: 90–95.

93. Hagberg A, Swart P, S Chult D. Exploring network structure, dynamics, and function using NetworkX. In: Gäel Varoquaux, Travis Vaught, and Jarrod Millman, editor. Proceedings of the 7th Python in Science Conference (SciPy2008). 2008. pp. 11–15.

94. van der Walt S, Colbert SC, Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science Engineering. 2011;13: 22–30.

95. McKinney W. Data Structures for Statistical Computing in Python. In: van der Walt S, Millman J, editors. Proceedings of the 9th Python in Science Conference. 2010. pp. 56–61.

96. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17: 261–272.

97. Waskom M, Botvinnik O, Ostblom J, Lukauskas S, Hobson P, MaozGelbart, et al. seaborn. 2020. doi:10.5281/zenodo.3629446

98. Drillon G, Carbone A, Fischer G. SynChro: A Fast and Easy Tool to Reconstruct and Visualize Synteny Blocks along Eukaryotic Chromosomes. Fairhead C, editor. PLoS One.

2014;9: e92621.

99. Smit AFA, Hubley R, Green P. RepeatMasker. 2015.

100. Pereira V. Automated paleontology of repetitive DNA with REANNOTATE. BMC Genomics. 2008;9: 1–21.

101. Levy O, Knisbacher BA, Levanon EY, Havlin S. Integrating networks and comparative genomics reveals retroelement proliferation dynamics in hominid genomes. Science Advances. 2017;3: 1–13.

102. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421.

103. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13: 2498–2504.

104. Angiuoli SV, Salzberg SL. Mugsy: Fast multiple alignment of closely related whole genomes. Bioinformatics. 2011;27: 334–342.

105. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32: 1792–1797.

106. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012;61: 539–542.

107. Britton T, Anderson CL, Jacquet D, Lundqvist S, Bremer K. Estimating divergence times in large phylogenetic trees. Syst Biol. 2007;56: 741–752.

108. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16: 276–277.

109. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016;33: 1870–1874.

110. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120.

111. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013;00: 1–3.

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

113. Broad Institute. Picard. Available: https://broadinstitute.github.io/picard/

114. Keane TM, Wong K, Adams DJ. RetroSeq: Transposable element discovery from next-generation sequencing data. Bioinformatics. 2013;29: 389–390.

115. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012; 1–9.

116. Garrison E. vcflib. 2018. Available: https://github.com/vcflib/vcflib

117. Miles A, Ralph P, Rae S, Pisupati R. scikit-allel. 2019. Available: https://github.com/cggh/scikit-allel

118. Lindenbaum P. JVarkit: java-based utilities for Bioinformatics. FigShare. 2015. Available: https://github.com/lindenb/jvarkit

119. Ortiz EM. vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis. 2019. Available: https://github.com/edgardomortiz/vcf2phylip

120. Terpilowski M. scikit-posthocs: Pairwise multiple comparison tests in Python. JOSS. 2019;4: 1169.

121. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference. Scipy; 2010. p. 61.

122. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35: 526–528.

123. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available: https://www.R-project.org/

124. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24: 1586–1591.

125. Janke C, Magiera MM, Rathfelder N, Taxis C, Reber S, Maekawa H, et al. A versatile toolbox for PCR-based tagging of yeast genes: new fluorescent proteins, more markers and promoter substitution cassettes. Yeast. 2004;21: 947–962.

126. Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony RK. Inexpensive multiplexed library preparation for megabase-sized genomes. PLoS One. 2015;10: 1–15.

127. Balakrishnan R, Park J, Karra K, Hitz BC, Binkley G, Hong EL, et al. YeastMine--an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. Database . 2012;2012: bar062.

Supporting information

Table S1. Description of the wild strains used in this study. CNs of full-length Ty1, Ty3p and Tsu4 (expressed as $\log_2$ NRD) are indicated for individual strains.

Table S2. AIC values for linear models testing the association between population structure and environmental variation.

Table S3. Nucleotide diversity at four-fold degenerate sites in protein-coding genes (π_S).

Table S4. Primers used for ADE2 deletion in the parental *SpC* strains generated in this study.

Table S5. Parental strains used in the MA experiment made in this study.

Table S6. Description of the MA lines used in this study. CNs of full-length Ty1, Ty3p and Tsu4 (expressed as $\log_2$ NRD) are indicated for individual strains.

Figure S1. Counts of LTR retrotransposon annotations in the six whole-genome assemblies.

Figure S2. LTR sequence similarity networks. Networks were built using blastn bit scores as a similarity metric, keeping the top 10% interactions for each node. Networks were plotted using Cytoscape with the Prefuse Force Directed layout algorithm.

Figure S3. Depth of coverage on reference internal sequences for Ty1, Ty3p and Tsu4. Row z-scores are shown for 200 bp-wide non-overlapping windows. Phylogenetic lineage assignment and Ty CN ($\log_2$ normalized read depth) are shown in the two first columns. White bars correspond to CN=0 (i.e. one copy).

Figure S4. Enrichment of low-divergence peaks in distributions of minimum nucleotide divergence between LTR sequences. Heatmaps show FDR-corrected p-values for pairwise Fisher's exact tests between ratios of elements below and strictly above thresholds of nucleotide divergence from 0 to 10%. Color maps are centered at the significance threshold of 0.05.

Figure S5. Ty orthogroups defined from the six genome assemblies. a. Distributions of Ty orthogroup sizes, from private (1) to conserved (6). **b.** Distributions of nucleotide divergence to the most closely related LTR sequence for each LTR within an orthogroup size category. Whiskers span 1.5 times the interquartile range.

Figure S6. Population structure based on Ty insertions called from discordant short read mappings. a. PCA on Ty insertion calls showing the first four PCs and the distribution of explained variance. **b.** Cumulative distirbutions of Ty allele frequencies within lineages. Triangle marks on the top indicate allele frequencies corresponding to one strain. Numbers in parentheses correspond to counts of alleles.

Figure S7. Bayesian phylogenetic trees of Ty1 LTR sequences. Bayesian phylogenetic trees are shown on the left and derived ultrametric phylogenetic trees are shown on the right.

Figure S8. Population structure based on genome-wide SNPs. PCA on SNP calls showing the first four PCs and the distribution of explained variance. Dot face colors correspond to the clusters used in Figure S10 and Figure 4b and dot edge colors correspond to the phylogenetic lineage assignment.

Figure S9. Climatic variation among natural lineages. PCA was performed on high-resolution climatic data (Abatzoglou et al. Sci Data 2018) at the sampling site of each strain.

The first four or three PCs are plotted with the distribution of explained variance. Dot face colors correspond to the clusters used in Figure S10 and Figure 4b and dot edge colors correspond to the phylogenetic lineage assignment.

Figure S10. Loadings of the PCAs on climatic variation data. Eigenvectors associated with the two first PCs of each dataset are plotted.

Figure S11. Strains clustering performed on either gPC or ePC coordinates. Columns within a heatmap correspond to individual strains. Ty CNs are represented by top rows, using the continuous color maps defined below each case. “Lineage” rows correspond to the established phylogenetic lineage assignment using the color code used throughout this paper. “gPCs” and “ePCs” correspond to clusters on gPCs and ePCs respectively, with discrete colors indicating the membership to different clusters. Cluster color mappings are consistent within datasets (All lineages, *SpB* and *SpC*).

Figure S12. Synonymous diversity at four-fold degenerate codon positions (π_S). a. Genome-wide π_S. **b.** Cumulative distributions of π_S. **c.** Median Ty CN per lineage as a function of π_S.

Figure S13. Ty CNs of strains harboring ORF-disrupting insertions in relation with lineage-wide distributions of CN. Strains with ORF-disrupting insertions are labelled with the corresponding gene and Ty family. Median Ty1 CN for SpC and median Tsu4 CN for SpB are shown as dotted lines.

Figure S14. Normalized depth of coverage at loci predicted to harbor Ty insertions. Rows correspond to the 208 wild strains. Strains with predicted insertions are labelled on the y axes. Positions on the x axes correspond to 400 bp windows centered at the predicted loci.

Figure S15. Neighbor joining phylogenetic trees of genes with ORF-disrupting Ty insertions. Branch color corresponds to the k parameter fitted by RELAX.

Figure S16. Ty CNs in MA lines with predicted values for the combination of parental CNs. CN values are the same as shown in Figure 6e, except that they are not normalized for median T_{ini} CN. Whiskers span 1.5 times the interquartile range. Colored lines denote individual strains.

Author contributions

Conceptualization: Christian R. Landry, Mathieu Hénault

Data curation: Mathieu Hénault

Formal analysis: Mathieu Hénault

Funding acquisition: Christian R. Landry, Mathieu Hénault, Souhir Marsit, Guillaume Charron

Investigation: Mathieu Hénault, Souhir Marsit, Guillaume Charron

Methodology: Mathieu Hénault, Guillaume Charron, Souhir Marsit

Visualization: Mathieu Hénault

Writing - Original draft preparation: Mathieu Hénault

Writing - Review and editing: Mathieu Hénault, Christian R. Landry, Souhir Marsit

Funding Disclosure

This work was supported by NSERC Discovery Grant to CRL, FRQNT and NSERC Alexander Graham-Bell scholarships to MH and GC and FRQS post-doctoral fellowship to SM. CRL holds the Canada Research Chair in Evolutionary Cell and Systems Biology.

Data availability

Illumina short read sequencing data of the MA lines is available at NCBI under accession PRJNA515073.