# Population Genetics of Transposable Elements in 1,011 Strains of *S. cerevisiae*

Katherine Lawrence
(Dated: May 18, 2018)

The architectures of eukaryotic genomes are marked by the presence and signatures of transposable elements (TEs); far from passive "junk DNA", TEs are now known to exhibit complex evolutionary and ecological dynamics within their host genomes, with significant consequences for adaptation and survival of the host organisms. The budding yeast *Saccharomyces cerevisiae* contains a relatively simple landscape of TE families, and it represents one of the most extensively sequenced organisms, both in whole-genome coverage and diversity of strains. Here we analyze TE copy number distributions in a recent dataset of 1,011 fully sequenced isolates from dozens of wild and domesticated niches. We provide a brief review of analytical and computational models of TE dynamics, including mechanisms of transposition, excision, regulation, selection, and diversification. Opportunities for future data analysis to visualize TE dynamics and compare to such models are discussed.

## Introduction: Transposable Elements

Transposable elements (TEs), a type of mobile genetic element, have proved to be a ubiquitous feature of genomic architecture in every kingdom of life. Since their discovery by Barbara McClintock in the 1940s, TEs have been observed in dozens of organisms and have been shown to constitute significant or majority fractions of total DNA in many eukaryotes (15% in *Drosophila melanogaster*, 40% in *Homo sapiens*, and up to 80% in *Zea mays*) [1].

While TEs encompass a startling amount of functional and structural diversity, they are unified by their ability to move between different locations in the genome, either using enzymes that the TEs themselves encode or co-opting enzymes encoded in the host genome [1]. Often the transposition process involves replication of the TE sequence and insertion into a functional site such as a gene or regulatory region, which could confer a fitness cost to the organism; in this way, TEs can be thought of as a parasitic population of selfish genetic elements whose environment is their host organism's genome [2].

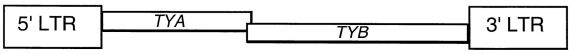
The maintenance and prevalence of TEs thus suggests a long history of co-adaptation over evolutionary timescales, with balancing forces for increase in TE copy number (replication rates, beneficial mutational effects or genomic rearrangements) and for decrease in TE copy number (deleterious mutational effects, excision rates, inactivating mutations, and host silencing mechanisms) [3]. TEs thus represent a promising field of study not only for understanding genetic architecture, genome flexibility and human disease [4], but also for modeling evolutionary dynamics and population genetics [5].

## TEs in *S. cerevisiae*

The genome of the model eukaryotic microbe *Saccharomyces cerevisiae* is quite compact, with TEs composing only 2-3% of the genome [6]. These TEs all belong to the same class, long terminal repeat (LTR) retrotransposons, and are divided into 5 families denoted Ty1-Ty5. Because of powerful yeast genetics tools developed over many decades and, more recently, large amounts of whole-genome sequencing data, yeast retrotransposons have been extensively characterized and have shown promise as a model system for studying TE dynamics [7]. While more complex organisms such as *Drosophila* are also promising systems due to their higher fraction of TE genomic content and wider diversity of TEs [8], the relative simplicity and experimental tractability of the yeast TEs make them an appealing context for studying TE dynamics.

### *Ty Element Structure*

All 5 TE families in yeast are LTR retrotransposons, and they share a similar structure and method of transposition. Similar to retroviruses, retrotransposons replicate by transcription into mRNA followed by reverse transcription back to cDNA, which then integrates into a new location in the genome. Structurally, a Ty element consists of a coding region, containing two open reading frames, bracketed on either side by an LTR region (see Fig. 1). The open reading frames *TYA* and *TYB* span 2.6-5.5 kb and are homologous to the *gag* and *pol* retrovirus loci, respectively; they code for enzymes including integrase, reverse transcriptase, and RNAse [7]. The LTR regions that bracket the coding region span 250-370 bp; they are identical to each other when the transposon is first inserted into the genome at a particular locus, though they can accumulate mutations as the TE remains at that locus. This allows for estimation of the age and transpositional activity of particular TE variants based on their accumulation of mutations in the LTR regions [9].




| | Size bp | | |
|---|---|---|---|
| Family | LTR | ORF | Group |
| Ty1 | 334 | 5,250 | copia |
| Ty2 | 332 | 5,300 | copia |
| Ty3 | 340 | 4,730 | gypsy |
| Ty4 | 371 | 5,480 | copia |
| Ty5 | 251 | 2,640 | copia |

FIG. 1: Structure of Ty retrotransposon families in *S. cerevisiae*. Reproduced from [7].

In addition to complete Ty elements, another common feature in the *S. cerevisiae* genome is "solo LTRs", where one LTR region appears on its own. These can arise from a process of excision by homologous recombination ("looping out"): a crossover occurs between the two LTR regions on either end of a Ty element, resulting in one LTR remaining at the genomic locus and a separate circular element containing the other LTR and the Ty coding region [10]. The solo LTR, having no remaining means of excision or transposition, can remain in the genome indefinitely; these traces of previous Ty insertions that have been lost provide a fossil record of Ty activity. In yeast, such solo LTRs far outnumber active, functional Ty elements [7].

### *The 1002 Yeast Genomes Project*

By 2013, from analysis of genomes of up to 41 different laboratory strains and natural isolates of *S. cerevisiae*, there was known to be a high degree of intraspecific diversity in Ty element copy number, family distribution, location, and activity [11, 12]. However, this represented only a small fraction of the known diversity of *S. cerevisiae* strains [13].

This year, a landmark dataset was published by the labs of Gianni Liti and Joseph Schacherer that contains complete genomes for 1,011 different *S. cerevisiae* isolates from a broad range of wild, domesticated/industrial, clinical, laboratory, and mosaic backgrounds, grouped into 30 clades [14]. In their paper, the authors analyze population structure and phylogeny, ploidy variation and aneuploidy, patterns in loss-of-heterozygosity, phenotypic diversity using GWAS, and the *S. cerevisiae* pangenome. The publicly available data (http://1002genomes.u-strasbg.fr/files/) includes raw Illumina reads, de novo assemblies, SNP matrices, and copy number counts for all pangenomic ORFs, including the 5 Ty families, for every strain. The authors briefly note differences in Ty copy number for certain clades, especially several domesticated clades, but without further analysis; in addition, there is no parsed data on copy numbers of solo LTRs, locations of active Ty elements and/or solo LTRs, or mutational relatedness/activity of Ty elements. Some of this data could potentially be extracted either from assemblies or raw Illumina reads, depending on how the copy number inference was done, but parsing such data on repetitive regions from short-read sequencing can be challenging. Due to time constraints this paper will only consider the copy number count data. As will become important below, the Ty copy numbers are defined per haploid genome (i.e. normalized by the number of whole-genome copies, which ranges from 1 to 5 in this dataset).

### Population Genetic Modeling of TEs

In the early 1980s, a series of papers by Charlesworth and others began to propose analytical population genetics models for the spread of transposable elements [15–18]. The simplest model describes the dynamics of the mean copy number per individual, $\bar{n}$, with only two terms, an effective replication rate $u$ and an effective excision rate $v$:

$$\Delta\bar{n} = u\bar{n} - v\bar{n} = (u - v)\bar{n}$$

Here, time is measured in discrete generations and the population size is infinite; the effective birth and death rates $u$ and $v$ can encompass multiple biological processes, but they are assumed to be independent of $\bar{n}$. We can see immediately that if $u < v$, the TE will go extinct, while if $u > v$, the TE copy number will grow exponentially (a biologically implausible scenario). Two mechanisms were proposed to limit the explosion of TEs and allow a steady-state nonzero mean copy number: (1) negative regulation of the birth rate such that $u(\bar{n})$ is a decreasing function of $\bar{n}$, and (2) natural selection on the deleterious effects of TEs, such that individuals with high $n$ are purged from the population.

The first case could be achieved by multiple biological factors, whether by TEs themselves preventig other insertions locally or globally, or via the host organism silencing or regulating TE activity once they are in sufficiently high copy number. Multiple different regulation mechanisms along these lines have indeed been observed [19]. We could model this as general negative autoregulation parametrized by

some $k$:

$$\Delta\bar{n} = (u(\bar{n}) - v)\bar{n} = \left(\frac{u}{1+k\bar{n}} - v\right)\bar{n}$$

This type of model, regardless of the choice of $u(\bar{n})$, predicts that at equilibrium, the steady-state copy number $\bar{n}^*$ will be such that $u(\bar{n}^*) = v$. However, this is not consistent with measurements of transposition and excision rates in multiple systems, where $v$ tends to be smaller than $u$ by at least an order of magnitude, so this is not likely to be the only mechanism [3].

For the second case, we must consider the fitness effects of TEs on the host organism, which can arise from several mechanisms. First, insertions can disrupt genes and regulatory regions. Second, TEs with repetitive elements (especially LTR retrotransposons) can mediate large-scale reorganization of the genome through ectopic recombination, where recombination between LTR regions from different TEs lead to duplications or deletions of large sections of chromosomes [10, 20]. The distribution of fitness effects of these types of mutations (insertions and gene copy number variation/aneuploidy) could vary in different organisms and in different environments (and indeed, selective effects ranging from lethality to beneficial adaptation have been observed [22]), though they are expected to be deleterious on average. We can write a general fitness function that depends on copy number, $w(n)$, and add a term to our model to denote the effect of natural selection:

$$\Delta\bar{n} = \left(u(\bar{n}) - v - \frac{\partial \ln w(\bar{n})}{\partial \bar{n}}\right)\bar{n}$$

One simple expression for the form of $w(n)$ is $w_n = 1 - sn^t$, where the parameter $s$ captures the average effect of each additional TE and $t$ represents the cumulative effect (additive for $t = 1$, multiplicative for $t > 1$, or sub-additive for $t < 1$). For insertion mutations of small deleterious effect, we might expect this function to be approximately additive, whereas for large-scale rearrangements, we may expect a multiplicative effect $\sim n^2$ since homologous regions on two different TEs must be colocalized for recombination to occur.

While simple models such as these were foundational in demonstrating that selfish genetic elements can, with the right parameters, invade genomes and be maintained in the population despite deleterious effects, there has not been a consensus on how to properly model regulation and/or selection in a way that agrees with complicated and often conflicting biological data [5]. More recently, computational/numerical models of stochastic dynamics have been developed for specific instances where the biological mechanisms and population genetic parameters are better characterized, most notably for *Drosophila* [5, 8, 21]. Additionally, some models began to consider multiple lineages of TEs as they accumulate mutations that reduce their transposition activity, motivated by the common observation that inactive TEs or their remnants are not only present in genomes but far outnumber active TEs [9, 22, 23]. The model in [22] also incorporates insertional fitness effects drawn from a distribution rather than given by a constant value. The authors of [23] construct phylogenies of their simulated TE populations and observe different structures under different parameter regimes: when selection against high TE copy numbers is strong, average copy numbers are low but a high proportion of them are active, and the phylogeny is strongly clustered; however, when selection against TEs is weak, copy numbers are high but many of them are inactive, and the phylogeny is star-shaped since most TEs are accumulating mutations without replicating.

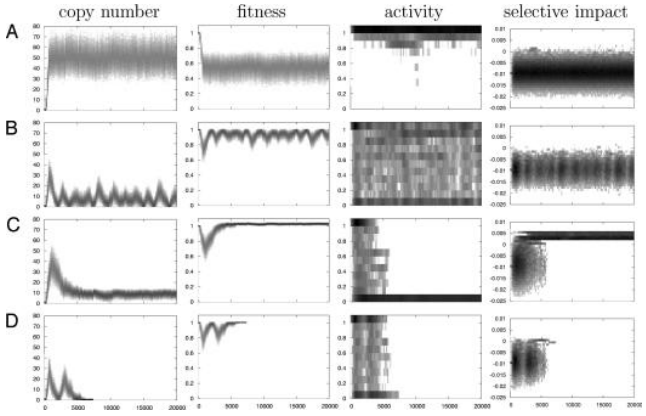



FIG. 2: Simulation results indicating potential outcomes of a TE invasion: TE copy number, population fitness, TE activity, and TE selective effect as a function of generations since invasion. Examples of outcomes under different values of mutation rate, activity, and selective effect: (A) equilibrium; (B) cycles; (C) domestication; (D) extinction. See [22] for model description and parameter values. Reproduced from [22].

A common feature of many of these more complex models is that steady states of copy number are rarely achieved, except for regimes of parameter values that seem inconsistent with experimental measurements [22–24]. Instead, common outcomes include large cycles or fluctuations in copy number over long times, "domestication" of inactive copies, or extinction of the TE after an initial invasion phase; we will not reproduce the details of different models and parameter regimes, but for a qualitative picture, examples of these outcomes are illustrated

in Fig. 2 (see [22] for details). Our experimental observations extend over very short timespans compared to the age of TEs, which for yeast have been estimated at millions of generations [9], and so it is quite conceivable that we are simply observing instantaneous copy numbers along an invasion-extinction trajectory that extends over evolutionary timescales. Phylogenetic models that infer past transposition frequency and activity may be more informative than copy-number-only models by constraining possible life histories of TEs, just as similar models are applied to infer past population structure of organisms [25].

### Data Analysis: Ty Copy Number Distributions

We can consider the 1,011 strain genomes from [14] as samples from the global population of *S. cerevisiae* (though not independent, due to their phylogenetic relatedness), or alternatively, we can consider each clade as its own subpopulation (with sizes ranging from 1 to 362 isolates per clade). There is also an intermediate level of organization: the clades can be categorized as wild (isolated from plants, fruits, insects, or soil samples) or domesticated (used for industrial production of wine, beer, sake, dairy products, bioethanol, and others), or not belonging clearly to either group (mosaic strains, clinical isolates, baking strains, and other strains of mixed origin). The domesticated strains and wild strains generally form two separate clusters in the phylogeny, with the mixed and mosaic clades in between (Fig. 3). Domesticated strains often show phenotypic and genotypic characteristics that reflect their distinct population size/structure and selective environment, making them interesting targets of study for evolutionary biologists as well as industry scientists [13, 26, 27]. In particular, domesticated yeasts such as wine and beer yeasts often show distinct karyotypes, including higher levels of ploidy (additional copies of the entire genome) as well as higher levels of aneuploidy (additional or missing chromosomes), and there is evidence in some cases that this is adaptive [28, 29].

We can first visualize the distribution of Ty elements in the wild, domesticated, and mixed populations separately. In Fig. 4 below, we show the distributions of the 5 Ty families in the three different categories (sample sizes $n = 60$ wild isolates, $n = 583$ domesticated isolates, and $n = 368$ mixed isolates). We see that for the most significant families, Ty1 and Ty2, the domesticated and mixed clades show both a higher mean copy number and higher variance in copy number than the wild populations. The simplest model of a birth-death process that has reached

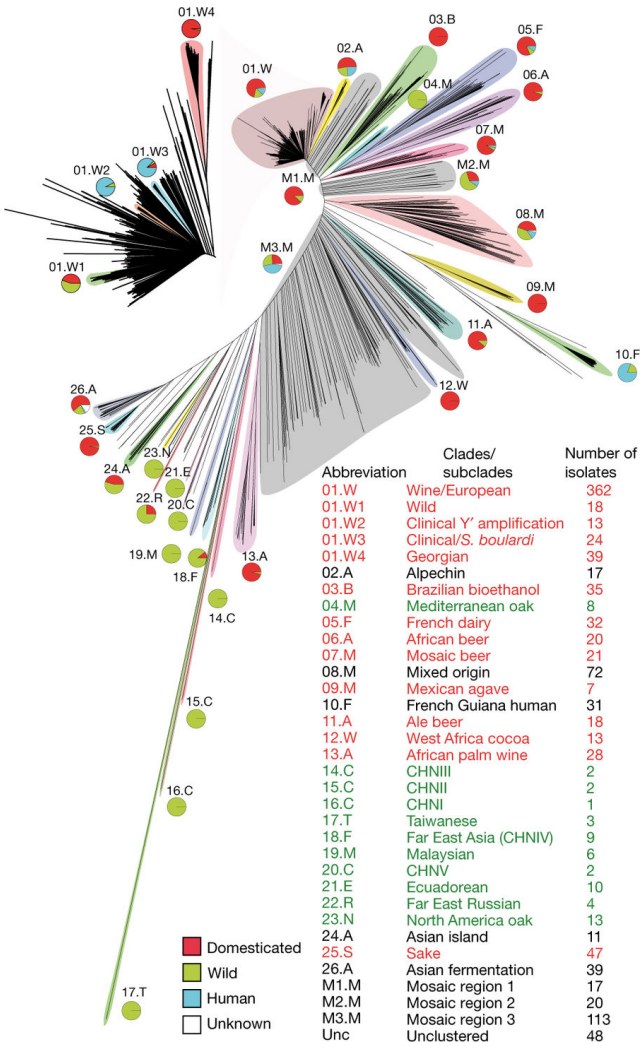



FIG. 3: Neighbor-joining tree constructed from SNPs showing the 1,011 strains grouped into 29 clades. Coloring of clade names indicates category assignment (red for domesticated, green for wild, black for mixed) and pie charts indicate origin of strains within each clade. Inset shows a magnified view of the Wine/European clade. Reproduced from [14].

equilibrium would predict a Poisson distribution of copy numbers within the population, with a variance in copy number equal to the mean. Here, we see much higher variance in almost all cases. However, we know the individual strains within each category were not independently sampled from one population, but rather belong to multiple distinct subpopulations as determined from their phylogenetic relatedness.

We can then examine the copy number distributions for each clade, under the assumption that individual isolates from a similar environmental niche and geographic region may perhaps be a more representative sample of a natural subpopulation. Most

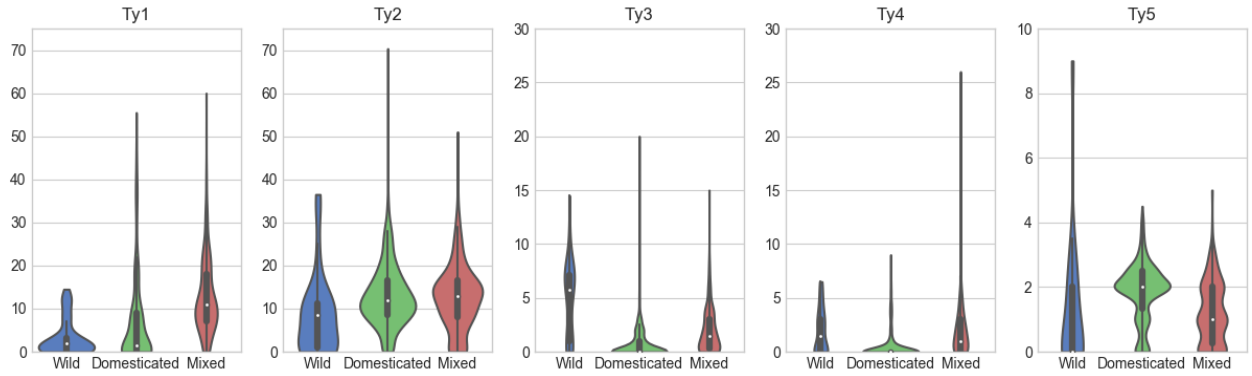



FIG. 4: Distributions of Ty element copy numbers in wild ($n = 60$), domesticated ($n = 583$), and mixed ($n = 368$) clades. Box plots indicate quartiles of the distribution. Note the different y-axis scalings for each family.

of the wild clades comprise 10 or fewer isolates, so it is difficult to assess the TE distribution; the mixed clades have many representative isolates, but because of their heterogeneous nature, it is difficult to interpret the distributions. The domesticated clades, however, comprise an average of 67 isolates per clade, and their population history and selective environment is better understood. We plot the Ty1 and Ty2 copy number distributions for the 10 domesticated clades in Fig. 5. We see that there is a large spread in the mean copy number across clades, but most clades show a relatively tight grouping around their mean. There are some exceptions that show very high variance, most notably French dairy for both Ty1 and Ty2 and the African beer and wine strains for Ty1. There is also no strong correlation between Ty1 copy number and Ty2 copy number, either positive or negative. Overall, the pattern is very heterogeneous from clade to clade even within the domesticated strains, perhaps reflecting either founder effects in the ancestral strains used during intial domestication or the widely different industries and environments in which these industrial yeasts have adapted.

We can also consider the prevalence of TEs in domesticated strains in relation to another pattern that differs systematically between domesticated and wild yeasts: ploidy levels and aneuploidies. Unlike wild strains, which in this dataset are only observed as diploids and (rarely) haploids, the domesticated and mixed isolates have ploidies ranging up to 5, and they also exhibit a significantly higher proportion of aneuploidies than wild strains. In Fig. 6, we show a density plot of strains by their total TE copy number and ploidy for wild, domesticated, and mixed strains. The TE copy number is plotted both as the per-haploid-genome normalized value (upper row) and as the true total copy number (normalized copy number times ploidy, bottom row). We observe that the domesticated strains with very high ploidy do show larger TE copy numbers in an absolute sense, but actually have quite low per-haploid-genome copy numbers. The mixed strains with high ploidy, by contrast, have fairly constant per-haploid-genome copy numbers even for the triploid and tetraploid strains, giving a roughly linear correlation in total TE copy number.

There are multiple ways in which ploidy and TE copy number can interact. As mentioned above, TEs are known to mediate large-scale genomic rearrangements, such as those that give rise to aneuploidies, through ectopic recombination; there are even observed examples of such Ty-mediated karyotype modifications in wine strains [28]. Alternatively, duplication (or deletion) of large portions of the genome will necessarily also duplicate (or delete) all TEs are contained therein. A whole-genome ploidy increase would not affect the per-haploid-genome normalized copy number; however, there could be effects that scale with absolute copy number rather than copy number per genome. For example, some absolute amount of transposition activity could be harmful to host organism fitness, such that TE copy numbers are reduced; or host organism silencing of TE activity could be upregulated once the TE copy number increases past a threshold, so that TE activity is reduced below what the copy number would suggest. Many of these effects could be acting simultaneously, and the TE copy number distribution is not the most sensitive variable to disentangle the various signs and magnitudes of the forces affecting TE dynamics, though it does elucidate the diversity of TE dynamics even within different isolates of one species.

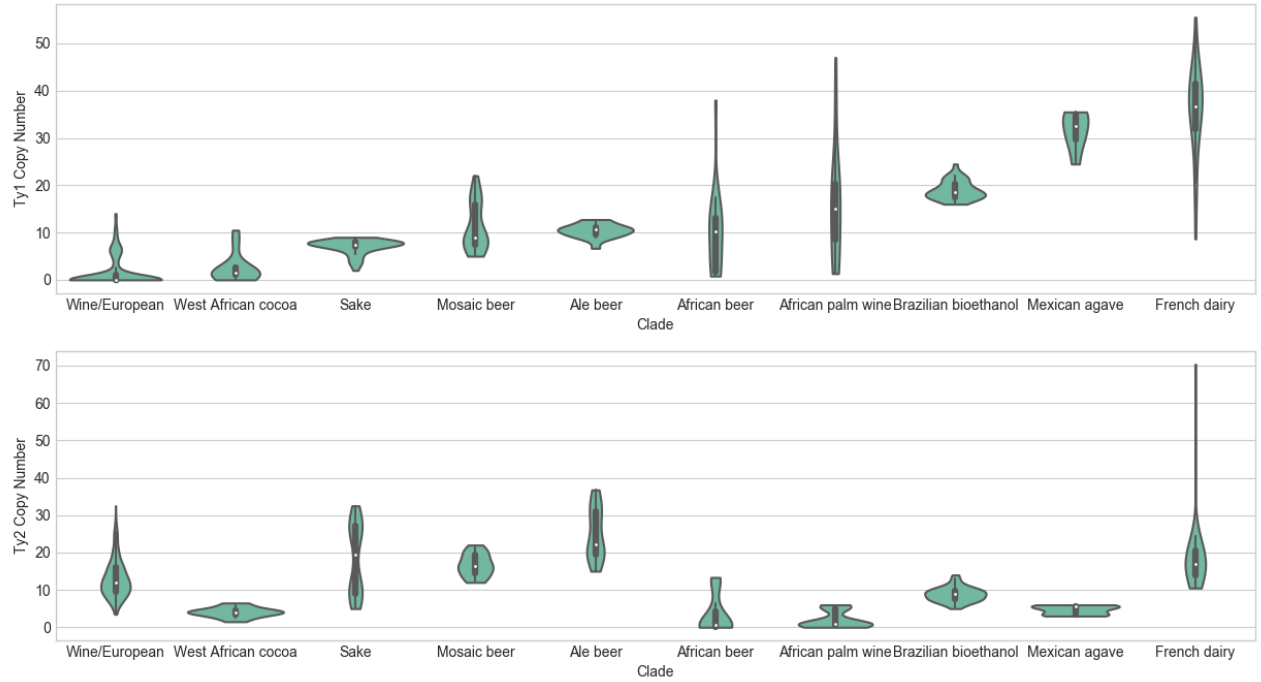



FIG. 5: Distributions of Ty1 and Ty2 copy numbers in domesticated clades: Wine/European ($n = 456$), West African cocoa ($n = 13$), Sake ($n = 47$), Mosaic beer ($n = 21$), Ale beer ($n = 18$), African beer ($n = 20$), African palm wine ($n = 28$), Brazilian bioethanol ($n = 35$), Mexican agave ($n = 7$), and French dairy ($n = 32$).

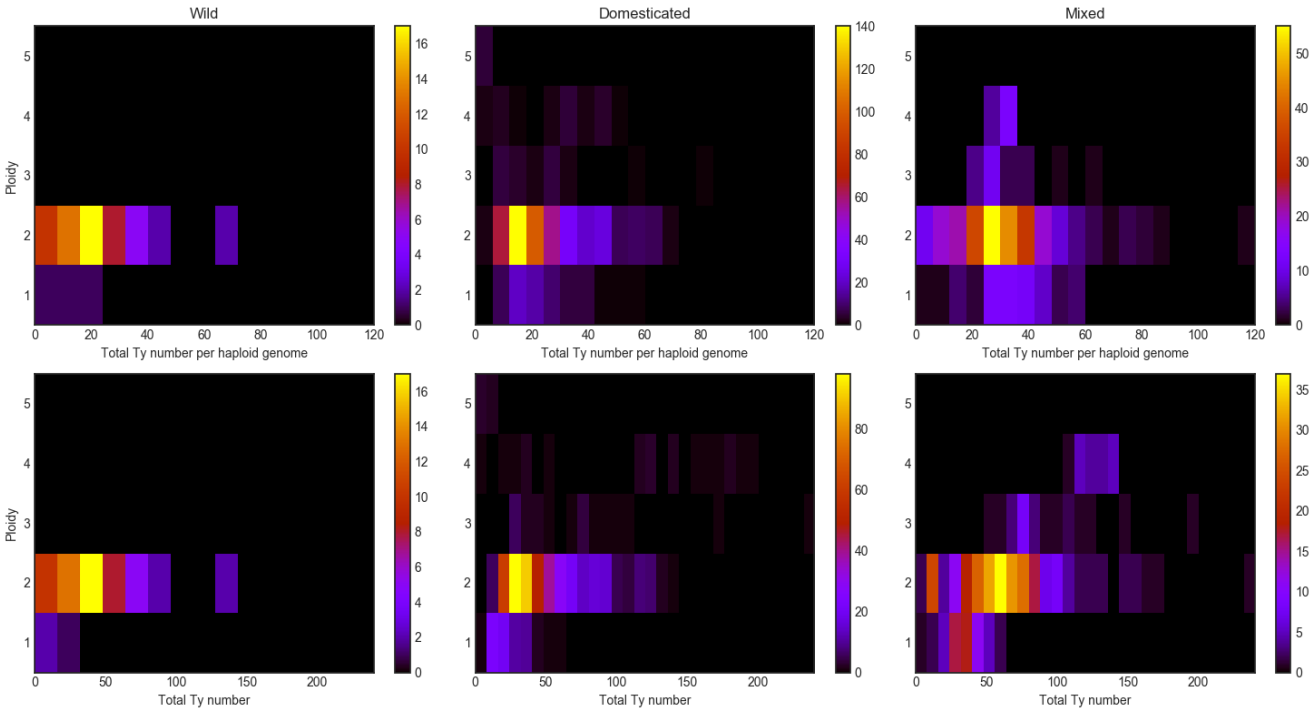



FIG. 6: Density plots of strains by ploidy and total TE copy number for wild (left), domesticated (center), and mixed (right) strains. Top row: TE copy number normalized by ploidy. Bottom row: total TE copy number.

### Future Work

Transposable elements exert influence on and are influenced by their host genomes via multiple mechanisms acting at multiple levels, ranging from DNA sequence to regulation/protein production to cell fitness to host population genetics. This complexity, in addition to the ubiquity of TEs and their signatures in eukaryotic genomes, has spurred decades of progress in biological experiments and population genetic modeling. Newly available whole genome sequencing datasets of organisms such as *S. cerevisiae*, with increasingly large sample sizes and increasingly high coverage, are enabling new explorations of TE biology. As a first glance, here we have presented copy number distribution data from 1,011 widely sampled strains of *S. cerevisiae* [14].

There are several types of additional data that could potentially be extracted from this dataset to address the questions raised above about TE dynamics in domesticated yeast strains. First, if one could compile data on TE locations throughout the genome in different strains, one could examine the aneuploidy locations to see if either TEs could have mediated the duplications or if TEs are preferentially present or absent in the aneuploid regions. Second, if the sequence identity of the TEs could be ascertained, then one could both measure the activity of TEs in different strains and the phylogenetic relationships between TEs in different strains. If TE activity was altered during the relatively recent period since domestication, that might provide evidence for roles that TEs could have played in the adaptation to industrial environments. Third, a survey of solo LTRs—the traces of former insertions—could provide a historical view of past activity in the genome, and perhaps shed light on whether particular Ty families are in a somewhat steady state or are progressing through an invasion-extinction cycle. Models such as those described in [22, 23] that explicitly account for inactivating mutations and distributions of selective effects could be useful in comparison with such data.


[1] N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, eds., *Mobile DNA II* (American Society for Microbiology Press, 2002).

[2] J. F. Y. Brookfield, Nature Reviews Genetics **6**, 128 EP (2005).

[3] B. Charlesworth, P. Sniegowski, and W. Stephan, Nature **371**, 215 EP (1994).

[4] O. Piskurek and D. J. Jackson, Genes **3**, 409 (2012).

[5] A. Le Rouzic and G. Deceliere, *Genetical research*, **85**, 171 (2005).

[6] J. M. Kim, S. Vanguri, J. D. Boeke, A. Gabriel, and D. F. Voytas, Genome Research **8**, 464 (1998).

[7] I. K. Jordan and J. F. McDonald, Genetics **151**, 1341 (1999), http://www.genetics.org/content/151/4/1341.full.pdf.

[8] T. J. McCullers and M. Steiniger, Mobile Genetic Elements **7**, 1 (2017).

[9] D. E. L. Promislow, I. K. Jordan, and J. E. McDonald, Proceedings of the Royal Society of London B: Biological Sciences **266**, 1555 (1999), http://rspb.royalsocietypublishing.org/content/266/1428/1555.full.pd

[10] P. A. Mieczkowski, F. J. Lemoine, and T. D. Petes, *Mechanisms of chromosomal translocation*, DNA Repair **5**, 1010 (2006).

[11] C. Bleykasten-Grosshans and C. Neuvéglise, *Ten years of genomic exploration in eukaryotes : strategy and progress of Genolevures*, Comptes Rendus Biologies **334**, 679 (2011).

[12] C. Bleykasten-Grosshans, A. Friedrich, and J. Schacherer, BMC Genomics **14**, 399 (2013).

[13] G. Liti, D. M. Carter, A. M. Moses, J. Warringer, L. Parts, S. A. James, R. P. Davey, I. N. Roberts, A. Burt, V. Koufopanou, I. J. Tsai, C. M. Bergman, D. Bensasson, M. J. T. O'Kelly, A. van Oudenaarden, D. B. H. Barton, E. Bailes, A. N. Nguyen, M. Jones, M. A. Quail, I. Goodhead, S. Sims, F. Smith, A. Blomberg, R. Durbin, and E. J. Louis, Nature **458**, 337 EP (2009).

[14] J. Peter, M. De Chiara, A. Friedrich, J.-X. Yue, D. Pflieger, A. Bergström, A. Sigwalt, B. Barre, K. Freel, A. Llored, C. Cruaud, K. Labadie, J.-M. Aury, B. Istace, K. Lebrigand, P. Barbry, S. Engelen, A. Lemainque, P. Wincker, G. Liti, and J. Schacherer, Nature **556**, 339 (2018).

[15] B. Charlesworth and D. Charlesworth, *Genetical Research*, **42**, 1 (1983).

[16] B. Charlesworth and C. H. Langley, Genetics **112**, 359 (1986).

[17] B. Charlesworth, *Genetical Research*, **57**, 127 (1991).

[18] J. F. Y. Brookfield and R. M. Badge, Genetica **100**, 281 (1997).

[19] M. Labrador and V. G. Corces, *Annual Review of Genetics*, Annual Review of Genetics **31**, 381 (1997).

[20] F. J. Lemoine, N. P. Degtyareva, K. Lobachev, and T. D. Petes, *Cell*, Cell **120**, 587 (2005).

[21] H. Quesneville and D. Anxolabéhère, Theoretical Population Biology **54**, 175 (1998).

[22] A. Le Rouzic, T. S. Boutin, and P. Capy, Proceedings of the National Academy of Sciences of the United States of America **104**, 19375 (2007).

[23] T. E. Kijima and H. Innan, Genetics **195**, 957 (2013).

[24] A. Le Rouzic and P. Capy, "Theoretical approaches to the dynamics of transposable elements in genomes, populations, and species," in *Transposons and the Dynamic Genome*, edited by D.-H. Lankenau and J.-N. Volff (Springer Berlin Heidelberg, Berlin, Heidelberg, 2009) pp. 1–19.

[25] A. Le Rouzic, T. Payen, and A. Hua-Van, Genome Biology and Evolution **5**, 77 (2013).

[26] M. Gonçalves, A. Pontes, P. Almeida, R. Barbosa, M. Serra, D. Libkind, M. Hutzler, P. Gonçalves, and J. Sampaio, *Current Biology*, Current Biology **26**, 2750 (2016).

[27] B. Gallone, J. Steensels, T. Prahl, L. Soriaga, V. Saels, B. Herrera-Malaver, A. Merlevede, M. Roncoroni, K. Voordeckers, L. Miraglia, C. Teiling, B. Steffy, M. Taylor, A. Schwartz, T. Richardson, C. White, G. Baele, S. Maere, and K. J. Verstrepen, *Cell*, Cell **166**, 1397 (2016).

[28] N. Rachidi, P. Barre, and B. Blondin, Molecular and General Genetics MGG **261**, 841 (1999).

[29] J. J. Infante, K. M. Dombek, L. Rebordinos, J. M. Cantoral, and E. T. Young, Genetics **165**, 1745 (2003).