DNA replication errors are a major source of adaptive gene amplification

Julie N. Chuong[1], Nadav Ben Nun[2,3], Ina Suresh[1], Julia Matthews[1], Titir De[1], Grace Avecilla[4], Farah Abdul-Rahman[5], Nathan Brandt[6], Yoav Ram[2,3], David Gresham[1,7]

Affiliations:

[1]Department of Biology, Center for Genomics and Systems Biology, New York University
[2]School of Zoology, Faculty of Life Sciences, Tel Aviv University
[3]Edmond J. Safra Center for Bioinformatics, Tel Aviv University
[4]Biobus, Inc.
[5]Memorial Sloan Kettering Cancer Center
[6]Department of Biological Sciences, North Carolina State University
[7]Correspondence: dgresham@nyu.edu

Abstract

Detecting and understanding heritable changes in DNA that contribute to adaptive evolution is a primary goal of evolutionary biology. Copy number variants (CNVs)—gains and losses of genomic sequences—are an important source of genetic variation underlying rapid adaptation and genome evolution. However, despite their central role in evolution little is known about the factors that contribute to the structure, size, formation rate, and fitness effects of adaptive CNVs. Local genome elements are likely to be an important determinant of these properties. Whereas it is known that point mutation rates vary with genomic location and local sequence features, the role of genome architecture in the formation, selection, and the resulting evolutionary dynamics of CNVs is poorly understood. Previously, we have found that the *GAP1* gene in *Saccharomyces cerevisiae* undergoes frequent and repeated amplification and selection under long-term experimental evolution in glutamine-limiting conditions. The *GAP1* gene has a unique genomic architecture consisting of two flanking long terminal repeats (LTRs) and a proximate origin of DNA replication (autonomously replicating sequence, ARS), which are likely to promote rapid *GAP1* CNV formation. To test the role of these genomic elements on CNV-mediated adaptive evolution we performed experimental evolution in glutamine-limited chemostats using engineered strains lacking either the adjacent LTRs, ARS, or all elements. Using a CNV reporter system and neural network simulation-based inference (nnSBI) we quantified the formation rate and fitness effect of CNVs for each strain. We find that although *GAP1* CNVs repeatedly form and sweep to high frequency in all strains, removal of local DNA elements significantly impacts the rate and fitness effect of CNVs and the rate of adaptation. We performed genome sequence analysis to define the molecular mechanisms of CNV formation for 177 CNV lineages. We find that 49% of all *GAP1* CNVs are mediated by the DNA replication-based mechanism Origin Dependent Inverted Repeat Amplification (ODIRA) regardless of background strain. In the absence of the local ARS, a distal ARS can mediate ODIRA CNV formation. In the absence of local LTRs homologous recombination mechanisms still mediate gene amplification following *de novo* insertion of retrotransposon elements at the locus. Our study demonstrates the remarkable plasticity of the genome and reveals that DNA replication errors are a predominant source of adaptive CNVs.

Introduction

Defining the genetic basis and evolutionary dynamics of adaptation is a central goal in evolutionary biology. Mutations underlying adaptation or biological innovation can depend on multiple factors including genetic backgrounds, phenotypic states, and genome architecture (Blount et al., 2008, 2012). One important class of mutation mediating adaptive evolution are copy number variants (CNVs) which comprise duplications or deletions of genomic sequences that range in size from gene fragments to whole chromosomes. Quantifying the rates at which CNVs occur, the factors that influence their formation, and the fitness and functional effects of CNVs is essential for understanding their role in evolutionary processes.

CNVs play roles in rapid adaptation in multiple contexts and are an initiating event in biological innovation. In laboratory evolution experiments, a spontaneous tandem duplication captured a promoter for expression of a citrate transporter and resulted in *Escherichia coli* cells, typically unable to use citrate, to start metabolizing citrate as a carbon source (Blount et al., 2012). CNVs can be beneficial in cancer cells, promote tumorigenesis (Ben-David & Amon, 2020), enhance cancer cell adaptability (Rutledge et al., 2016), and accelerate resistance to anti-cancer therapies (Lukow et al., 2021). Over longer time scales, CNVs serve as substrate from which new genes evolve (Ohno, 1970; Taylor & Raes, 2004) as duplicated genes redundant in function can accumulate mutations and evolve to acquire new functions. For example, the globin gene family in mammals arose from rounds of gene duplication and subsequent diversification (Storz, 2016). CNVs also contribute to macro-evolutionary processes and thereby contribute to species differences, such as between humans and chimpanzees (Cheng et al., 2005) and speciation (Zuellig & Sweigart, 2018).

Mutations, including CNVs, occur in part because of errors made during DNA replication or DNA repair. Two general processes underlie CNV formation: (1) DNA recombination-based mechanisms and (2) DNA replication-based mechanisms (Harel et al., 2015; Hastings, Lupski, et al., 2009; Malhotra & Sebat, 2012; Pös et al., 2021; Zhang, Gu, et al., 2009). Recombination-mediated mechanisms of CNV formation include non-allelic homologous recombination (NAHR) and nonhomologous end joining. NAHR occurs via recombination between homologous sequences that are not allelic. As such, NAHR occurs more frequently with repetitive sequences due to improper alignment of DNA segments and can occur either between (interchromosomal) or within (intrachromosomal) a chromosome (Harel et al., 2015). One prevalent class of repetitive sequence are retrotransposons and both full length and partial sequences, such as long terminal repeats (LTR), are substrates for homologous recombination generating gene amplifications (Avecilla et al., 2023; Dunham et al., 2002; Gresham et al., 2008; Lauer et al., 2018; Spealman et al., 2022). DNA replication-based mechanisms include fork stalling template switching (FoSTes) and microhomology mediated break-induced repair (MMBIR) (Carvalho et al., 2013; Gu et al., 2008a; Hastings, Ira, et al., 2009; Lee et al., 2007). During FoSTes and MMBIR, the DNA replication fork stalls due to a single strand nick and a replication error occurs in which the lagging strand switches to an incorrect template strand mediated by microhomology. Reinitiation of DNA synthesis at the incorrect site can form CNVs. A particular type of DNA replication-based error is Origin-Dependent Inverted Repeat Amplification (ODIRA), in which short inverted repeats near an origin of DNA replication enable template switching of the leading strand to the lagging strand. Subsequent replication generates an

intermediate linear DNA molecule that can recombine into the original genome to form a triplication with an inverted middle copy (Brewer et al., 2015; Martin et al., 2024).

In microbes, CNVs can mediate rapid adaptation to selective conditions imposed through nutrient limitation in a chemostat. Selected CNVs often include genes encoding nutrient transporters that facilitate import of the limiting nutrient (Dunham et al., 2002; Gresham et al., 2008; Horiuchi et al., 1963; Payen et al., 2016; Sonti & Roth, 1989), likely as a result of improved nutrient transport capacity due to increased protein production. Previous studies have found amplification of the general amino acid permease gene, *GAP1,* when *Saccharomyces cerevisiae* populations are continuously cultured in glutamine-limited chemostats (Gresham et al., 2010; Lauer et al., 2018). Amplification of *GAP1* confers increased fitness in the selective environment (Avecilla et al., 2023). Sequence characterization of these CNVs revealed that a diversity of *de novo* CNV alleles arose and were selected for including tandem duplications of *GAP1,* complex large CNVs, aneuploidies, and translocations. However, little is known about the molecular basis of this genetic diversity.

Local genome elements are likely to be an important determinant of CNV formation rates and mechanisms. Genomic context can influence multiple properties including mutation rate, epigenetic regulation, chromatin state, transcription levels, DNA replication, and recombination rate (Arndt et al., 2005; Chuang & Li, 2004; Lang & Murray, 2011; Lercher & Hurst, 2002; Matassi et al., 1999; Nishant et al., 2009; Wolfe et al., 1989). Prior work has shown that CNVs occur more frequently in repetitive regions in the genome (Harel et al., 2015; Pentao et al., 1992; Stankiewicz et al., 2003; Turner et al., 2008). However, little is known about the role of local genomic architecture and organization on CNV formation rates, the types of CNVs that are generated, their associated fitness effects, and ultimately the paths taken during adaptive evolution.

Here, we aimed to directly investigate the effect of local genome architecture elements on *de novo GAP1* CNV formation and selection dynamics during adaptive evolution of *Saccharomyces cerevisiae*. We hypothesized that sequence elements proximate to *GAP1* potentiate CNV formation. The *GAP1* locus, which is located on the short arm of chromosome XI, consists of two flanking Ty1 long terminal repeats (LTRs) that share 82% sequence identity and an origin of DNA replication or autonomously replicating sequence (ARS) (**Figure 1A**). Both LTRs and ARS may facilitate *GAP1* CNV formation due to their proximity. First, the flanking LTRs can undergo inter-chromatid NAHR to form tandem duplications of *GAP1* on a linear chromosome (Lauer et al., 2018; Spealman et al., 2022). Second, intra-chromatid NAHR between the flanking LTRs can form an extrachromosomal circle containing *GAP1* and an ARS able to self-propagate and integrate into the genome (Gresham et al., 2010). Finally, *GAP1* triplications can form through ODIRA using short inverted repeats and the proximate ARS (Brewer et al., 2015; Lauer et al., 2018; Martin et al., 2024). These elements are thought to facilitate a high rate of *GAP1* amplification, estimated to be on the order of 10^{-4} per haploid genome per generation (Avecilla et al., 2022). To test our hypothesis we used a CNV reporter, wherein a constitutively expressed fluorescent GFP gene is inserted adjacent to *GAP1* (Lauer et al., 2018). We engineered strains that lacked either the ARS (ARS∆), both flanking LTRs (LTR∆), or all three elements (ALL∆) (**Figure 1A**). We performed experimental evolution using wildtype (WT) and genomic architecture mutant populations in glutamine-limited chemostats for 137 generations and quantified *GAP1* CNVs using flow cytometry (**Figure 1**). Surprisingly, we find that the proximate DNA elements are not required for *GAP1* CNV formation as *GAP1* CNVs were identified in all evolving

populations. We used neural network simulation-based inference (nnSBI) to infer the CNV formation rate and selection coefficient (Avecilla et al., 2022). We find that genomic architecture mutants have significantly reduced CNV formation rates relative to WT and significantly lower selection coefficients even though *GAP1* CNVs repeatedly form and sweep to high frequency in all strains. We performed genome sequence analysis to define the molecular mechanisms of CNV formation for 177 CNV lineages and found that 49% of *GAP1* CNVs are mediated by ODIRA regardless of background strain. In the absence of the local ARS, a distal ARS facilitates CNV formation through ODIRA. We also find that homologous recombination mechanisms still mediate gene amplification in the absence of LTRs in part initiated by *de novo* insertion of retrotransposon elements at the locus. Our study reveals the remarkable plasticity of the genome and that DNA replication errors are a predominant source of adaptive CNVs.

Results

Accurate estimation of CNV allele frequencies remains challenging using molecular methods such as DNA sequencing and qPCR. To address this challenge we developed a CNV reporter comprising a constitutively expressed fluorescent gene inserted upstream of *GAP1* and observed recurrent amplification and selection of *GAP1* in glutamine-limiting conditions (Lauer et al., 2018). Subsequently, we showed that a high rate of *GAP1* CNV formation and strong fitness effects explain the highly reproducible evolutionary dynamics (Avecilla et al., 2022). Noncoding sequence elements proximate to *GAP1*, including flanking LTRs in tandem orientation and an ARS, contribute to *GAP1* CNV formation (Gresham et al., 2010; Lauer et al., 2018). Many studies have shown that repetitive sequence regions and origins of replications are hotspots of CNVs (Arlt et al., 2012; Cardoso et al., 2016; Di Rienzi et al., 2009; Gresham et al., 2010; Lauer et al., 2018; Martin et al., 2024). Thus, we hypothesized that the local genomic architecture of *GAP1* facilitates its high rate of CNV formation.

To test the role of proximate genomic features we engineered strains deleted for each element and thus differing from our wildtype strain (WT) containing a *GAP1* CNV reporter by a single modification. Specifically, we constructed ARS∆, a strain lacking the single ARS, LTR∆, a strain lacking the flanking LTRs, and ALL∆, a strain lacking all three elements (**Figure 1A**). All strains contain the CNV reporter at the identical location as the WT strain. We confirmed scarless deletions of genetic elements using Sanger and whole-genome sequencing.

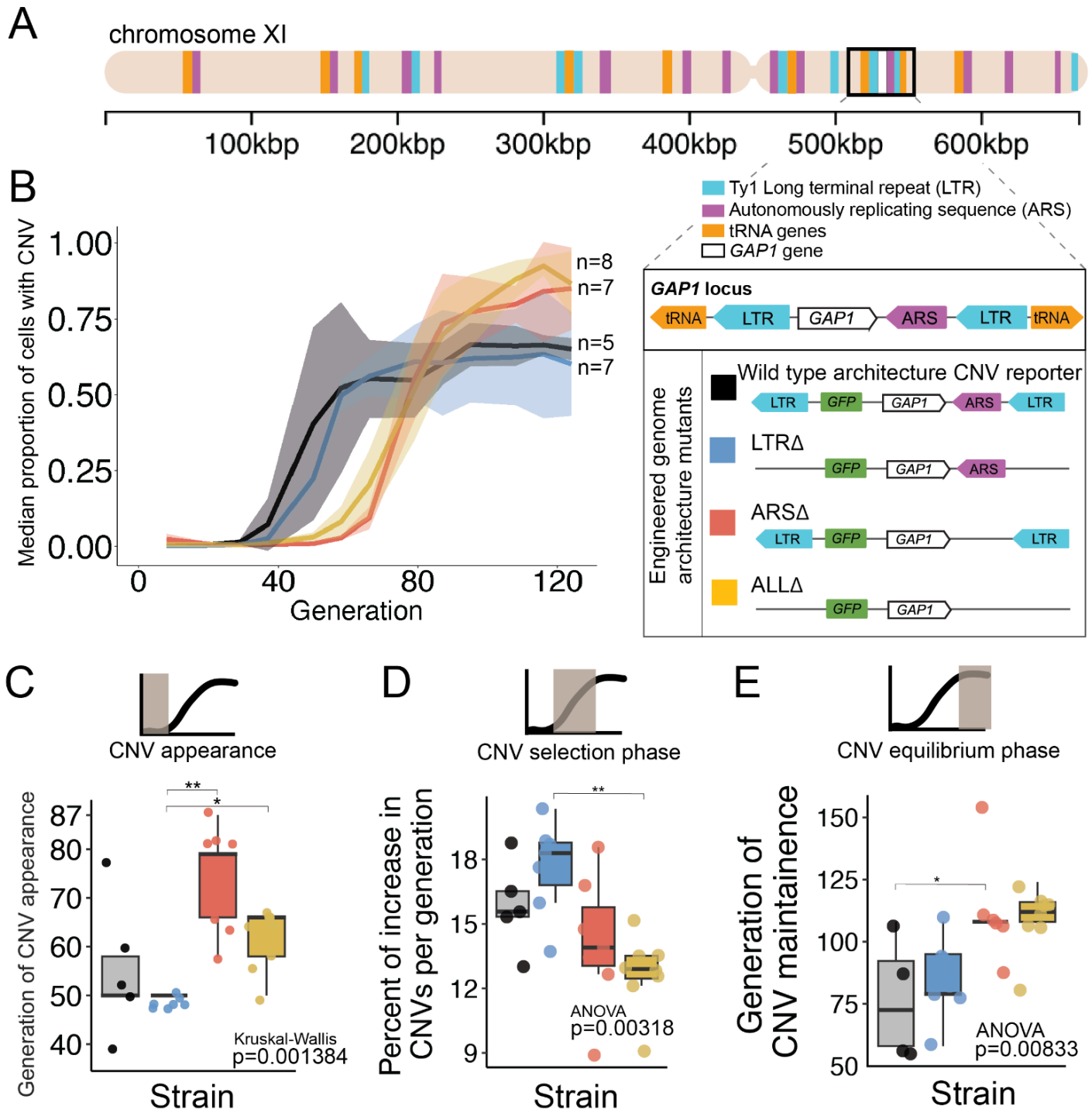


Figure 1. Local ARS contributes to CNV dynamics during adaptive evolution. (A) The *Saccharomyces cerevisiae GAP1* gene is located on the short arm of chromosome XI (beige rectangle). Light blue rectangle - Ty Long terminal repeats (LTR). Purple rectangle - Autonomously replicating sequences (ARS). Orange rectangles - tRNA genes. *GAP1* ORF - white rectangle. The *GAP1* gene (white rectangle) is flanked by Ty1 LTRs (YKRCδ11, YKRCδ12), which are remnants of retrotransposon events and is directly upstream of an autonomously replicating sequence (ARS1116). Variants of the *GAP1* locus were engineered to remove either both LTRs, the single ARS, or all three elements. All engineered genomes contain a CNV reporter. **(B)** We evolved the four different strains in 5-8 replicate populations, for a total of 27 populations, in glutamine-limited chemostats and monitored the formation and selection of *de novo GAP1*

CNVs for 137 generations using flow cytometry. Population samples were taken every 8-10 generations and 100,000 cells were assayed using a flow cytometer. Colored lines show the median proportion of cells in a population with *GAP1* amplifications across 5-8 replicate populations of the labeled strain. The shaded regions represent the median absolute deviation across the replicates. **(C)** We summarized CNV dynamics and found that strain has a significant effect on CNV appearance (Kruskal-Wallis, p =0.001384). There are significant differences in CNV appearance between LTR∆ (blue) and ARS∆ (red), and LTR∆ (blue) and ALL∆ (yellow) (pairwise wilcoxon test with Bonferroni correction, p=0.0059 and p=0.0124, respectively). **(D)** Strain has a significant effect on the per generation increase in proportion of cells with CNV (ANOVA, p = 0.00318). There is a significant difference between LTR∆ (blue) and ALL∆ (yellow) (pairwise t-test with Bonferroni correction, p = 0.0026). **(E)** Strain has a significant effect on time to CNV equilibrium phase (ANOVA, p = 0.00833). There is a significant difference in time to CNV equilibrium between WT and ARS∆ (pairwise t-tests with bonferroni correction, p =0.050).

Local genomic architecture contributes to *GAP1* CNV evolutionary dynamics

We founded independent populations with each of the three engineered strains lacking proximate genomic features and a WT strain. We studied *GAP1* CNV dynamics in populations maintained in glutamine-limited chemostats over 137 generations (**Figure 1**). For each of the four strains, we propagated 5-8 clonal replicate populations, each originating from the same inoculum (founder population) derived from a single colony. Approximately every 10 generations, we measured GFP fluorescence of sampled populations using a flow cytometer and quantified the proportion of cells containing *GAP1* CNVs (**Methods**). We observed similar CNV dynamics across independent populations within each strain (**Figure S3B**). Therefore, we summarized CNV dynamics for each strain using the median proportion of the population with a *GAP1* CNV. In every strain, *GAP1* CNVs are generated and selected resulting in qualitatively similar dynamics in WT and mutant strains (**Figure 1B**).

Deletion of the ARS, but not the flanking LTRs alters CNV dynamics

We quantified three phases of CNV dynamics 1) the time to CNV appearance, defined by the inflection point before rise in frequency (**Fig 1C**), 2) the selection phase, corresponding to the increase in proportion of CNVs per generation during the initial expansion of CNVs (**Figure 1D**), and 3) the equilibrium phase, corresponding to the plateau (**Figure 1E**). The time to CNV appearance (**Figure 1C**) and the CNV selection phase (**Figure 1B**) does not differ between WT and LTR∆ populations (pairwise wilcoxon test, adjusted p = 1, pairwise t-test p = 1, respectively). In the WT and LTR∆ populations, *GAP1* CNVs appear at generation 50 and increase in frequency at similar rates, ~15% per generation in WT and ~18% per generation in LTR∆. The two strains both reach their equilibrium phase at the same point around generation 75 (pairwise t-test, adjusted p =1). The absence of a significant difference in CNV adaptation dynamics between the two strains suggests that the LTRs are not a major determinant of *GAP1* CNV evolutionary dynamics.

By contrast, in ARS∆ and ALL∆ populations, we observe a delay in the time to CNV appearance. In both of these strains, CNVs are first detected at generations 65-80, whereas in WT and LTR∆ populations CNVs are first detected at generation 50 (ARS∆ vs. LTR∆, wilcoxon pairwise test, adjusted p = 0.0059, ALL∆ vs. LTR∆, wilcoxon pairwise test, adjusted, p = 0.0124) (**Figure 1**). Thus, the local ARS contributes to the initial *GAP1* CNV dynamics. Similarly, CNV selection is significantly different between the LTR∆ (18%) and ALL∆ (13%) (pairwise t-test, adjusted p-value =

0.0026). Finally, we also observe a significant delay (ANOVA, p = 0.00833) in the generation at which the CNV frequency reaches equilibrium in ARS∆ (~generation 112) compared to WT (pairwise t-test, adjusted p = 0.05) (**Figure 1E**). These observations suggest that absence of the ARS in the ARS∆ and ALL∆ strains delays the appearance of *GAP1* CNVs compared with the presence of the ARS in WT or LTR∆ strains.

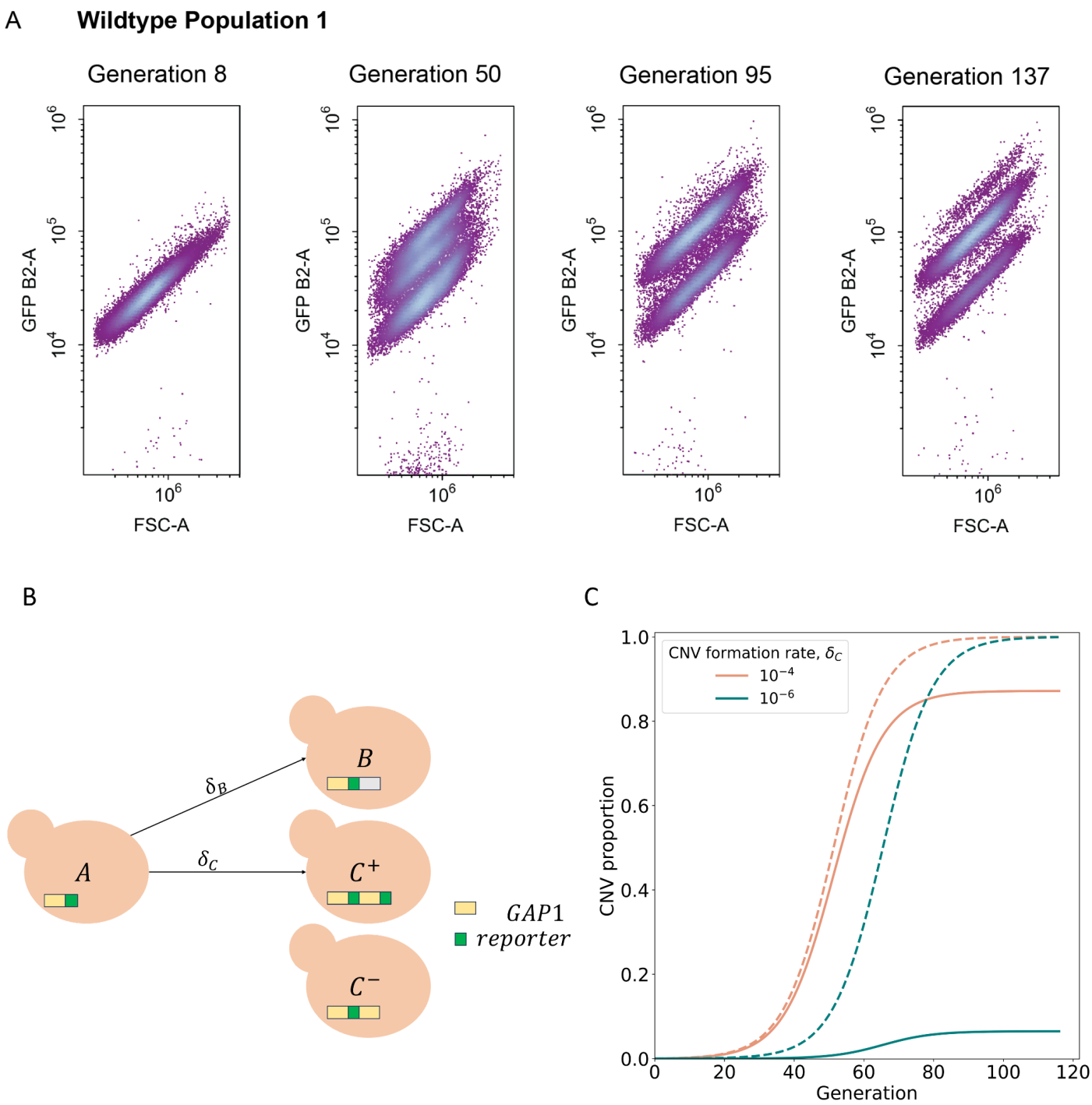


Figure 2. CNV reporter failure does not impact parameter inference. (**A**) Flow cytometry of a representative WT population with a persistent single-copy GFP subpopulation comprising ~25% of the population. (**B**) Model illustration. X_A is the frequency of ancestor cells in the chemostat; X_{C^+}, X_{C^-} are the frequencies of cells with *GAP1* duplications with two or one reporters, respectively, and a selection coefficient s_C; X_B is the frequency of cells with other beneficial mutations and a selection coefficient s_B. *GAP1* duplications form with a rate δ_C, other beneficial mutations occur with rate δ_B. At generation 0, only genotypes C^- and A are present, with frequencies of $X_{C^-} = \varphi$ and $X_A = 1 - \varphi$. (C) Examples of total CNV proportions (solid) and reported CNV proportions (dashed) for two parameter combinations, both with $s_C = 0.15$, $\varphi = 10^{-4}$.

GAP1 amplifications can occur without CNV reporter amplification

In both WT and LTRΔ populations we observed that *GAP1* CNV abundance stabilized around 75% during the equilibrium phase (**Figure 1B**). In these populations, inspection of raw flow cytometry data revealed a persistent single-copy GFP subpopulation (**Figure 2A**). These data could be explained by two possible scenarios: (1) the existence of a non-*GAP1* CNV subpopulation comprising beneficial variation at other loci with fitness effects equivalent to *GAP1* CNVs or (2) lineages with *GAP1* CNVs without co-amplification of the CNV reporter. To resolve these two possibilities we sequenced clones from the single-copy GFP subpopulation in each of the five WT populations (**Supplementary table 1)** and identified the presence of *GAP1* amplifications without co-amplification of the CNV reporter (**Supplementary figure S2A).** The presence of *GAP1* CNVs in four out of five populations lacking the CNV reporter amplification suggests it may have occurred in the founder population. We found eight distinct CNV breakpoint pairs (**Supplementary figure S2A**) indicating at least eight independent events occurred, either in the founder population or just after being propagated to independent populations. All clones from one of the five populations (i.e. population 3) contained one copy of GFP and one copy of *GAP1*, suggesting these clones have a beneficial mutation elsewhere in the genome allowing them to stably coexist with the *GAP1* CNV subpopulation.

Incorporating unreported pre-existing CNVs in an evolutionary model

To quantify the evolutionary parameters underlying empirically measured CNV dynamics (**Figure 1**) we built an evolutionary model with the goal of performing network simulation-based inference (nnSBI) (Avecilla et al., 2022; Cranmer et al., 2020; Gonçalves et al., 2020). Previously, our evolutionary model assumed the *GAP1* CNV reporter allowed us to detect all *GAP1* CNVs (Avecilla et al., 2022). However, our new experimental results indicate the existence of a small subpopulation of unreported *GAP1* CNVs present at the beginning of the experiments (**Figure 2A**). Therefore, we expanded the evolutionary model to include φ, the proportion of cells with *GAP1* CNVs without co-amplification of the reporter, at the commencement of the experiment (i.e. generation 0) (**Figure 2B**). The remaining model parameters are δ_C, the rate at which *GAP1* duplications form, and δ_B, the rate other beneficial mutations occur. We find that this updated evolutionary model can accurately describe the observed dynamics (**Supplementary Figure S3B**), which are clearly affected by the value of φ. When the total CNV proportion is very different from the reported proportion, e.g., when $\varphi \gg \delta_C > \delta_B$, a reduced CNV formation rate results in a greater discrepancy between reported and total CNV proportions (**Figure 2C**)

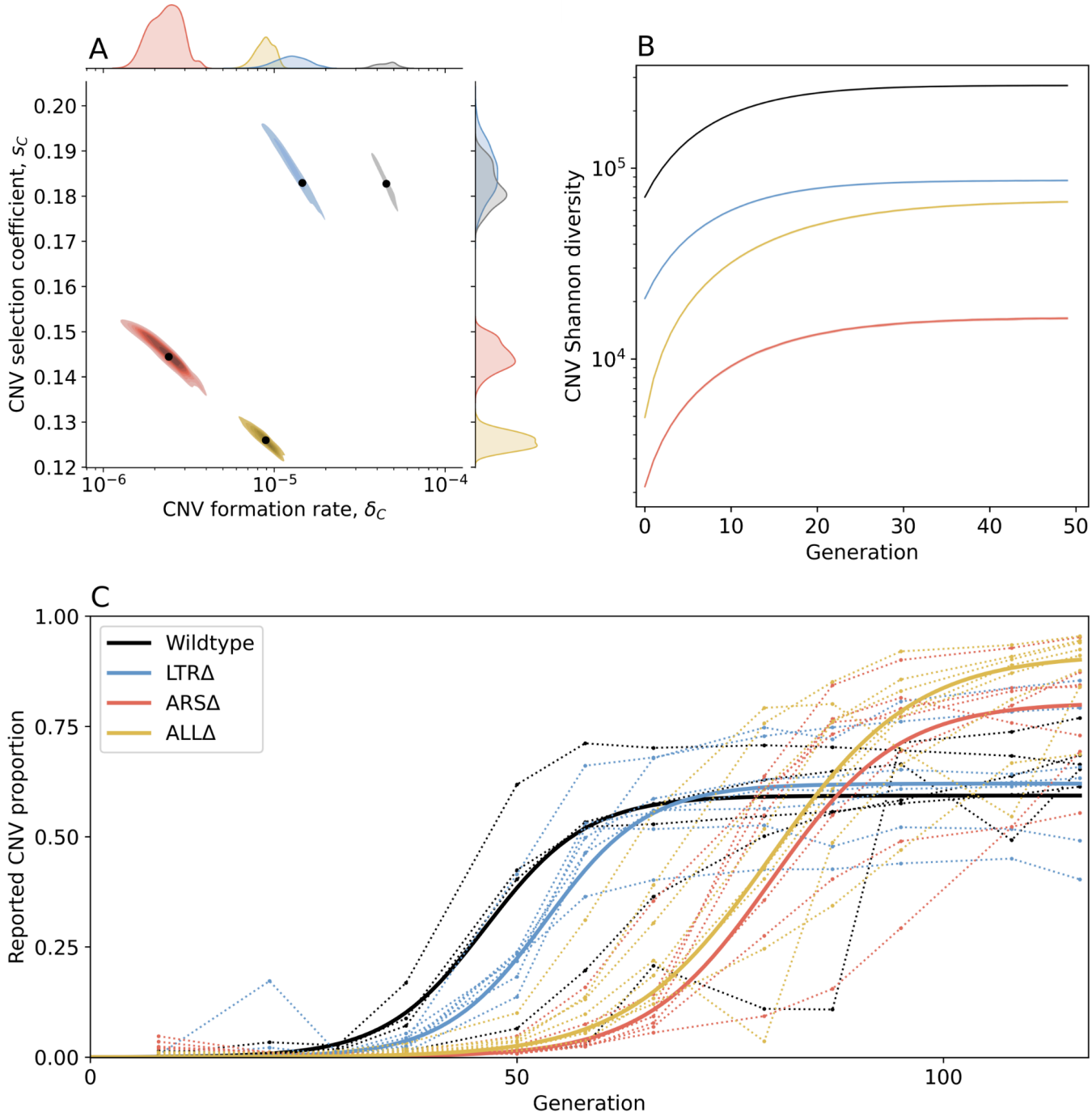


Figure 3. Inference of CNV formation rate and selection coefficient from experimental evolutionary data. (**A**) Collective MAP estimate (black markers) and 50% HDR (colored areas) of *GAP1* CNV formation rate, δ_C, and selection coefficient, s_C. Marginal posterior distributions are shown on the top and right axes. (**B**) Collective posterior prediction of Shannon diversity of CNV lineages ($e^{-\Sigma_i[p_i log(p_i)]}$, Jost, 2006). Line and shaded area show mean and 50% HDI. (**C**) CNV reported frequency (X_{C^+}) prediction using collective MAP (solid line) compared to empirical observations (dotted lines).

Decreased CNV formation rates in mutants suggest adjacent elements can drive *GAP1* CNV formation

We used nnSBI to estimate CNV formation rates and selection coefficients from evolutionary dynamics observed in glutamine-limited chemostats. We trained a neural density estimator using evolutionary simulations (**Methods**). This neural density estimator then allows us to infer posterior distributions over model parameters (formation rate, δ; selection coefficient, s) from the CNV dynamics in single replicates and a collective posterior distribution from a set of replicates of the same strain. δ_C is the *GAP1* CNV formation rate. s_C is the *GAP1* CNV selection coefficient, wherein the fitness effect is $1 + s_C$. We estimated the confidence of our inference approach on synthetic simulations by computing its coverage, i.e., the probability that the true parameter falls within the 95% highest density interval (HDI) of the posterior distribution. Our neural density estimator is slightly over-confident for φ with a coverage of 0.934, and under-confident for *GAP1* CNV selection coefficient and formation rate with a coverage of 0.992 for s_C and 0.995 for δ_C, respectively. (**Supplementary Table 2**). Despite this under-confidence, the posterior distributions are narrow in biological terms: the 95% HDI represents less than an order of magnitude for both s_C and δ_C. Thus, we did not apply post-training adjustments to the neural density estimator, such as calibration (Cook et al., 2006) or ensembles (Caspi et al., 2023; Hermans et al., 2022).

We find that the individual maximum a posteriori (MAP) estimates vary across strains and replicates (**Supplementary Figure S3A**). Overall, the CNV selection coefficient, s_C , ranges from 0.1 to 0.22 (with one exception of 0.3) whereas the CNV formation rate, δ_C , ranges from 10^{-6} to 10^{-4} (with one exception of 10^{-3} and two of 10^{-7}); and the proportion of pre-existing *GAP1* CNVs that do not amplify the reporter φ ranges from 10^{-6} to 10^{-2} (with two exceptions of 10^{-8}). We found that the MAP estimates for replicate populations within the same strain cluster together with some outliers (**Supplementary Figure S3A**). We performed posterior predictive checks, drawing parameter values from the posterior distributions and simulating the frequency dynamics (**Supplementary Figure S3H)**, which agree with the observed data (**Supplementary Figure S3B**). For each strain, we estimate the collective posterior distribution based on all individual posteriors, which generates a posterior distribution conditioned on all observations, $P(\theta|X_1, ..., X_n)$ (**Methods**). The collective posterior allows us to estimate whether there is a difference in CNV formation rate and fitness effects across the four strains.

Collective posterior HDIs are very narrow (**Figure 3A**) and fit empirical observations (**Figure 3C**). The collective MAP estimates of the CNV selection coefficient are similar for the WT and LTRΔ (0.182). For ARSΔ and ALLΔ, the selection coefficient is estimated to be lower, with values of 0.146 and 0.126, respectively. However, these selection coefficients are still large, consistent with these populations containing *GAP1* CNVs that are highly beneficial under glutamine-limitation. The collective MAP estimate for the CNV formation rate in WT is $4.5 \cdot 10^{-5}$. In contrast with selection

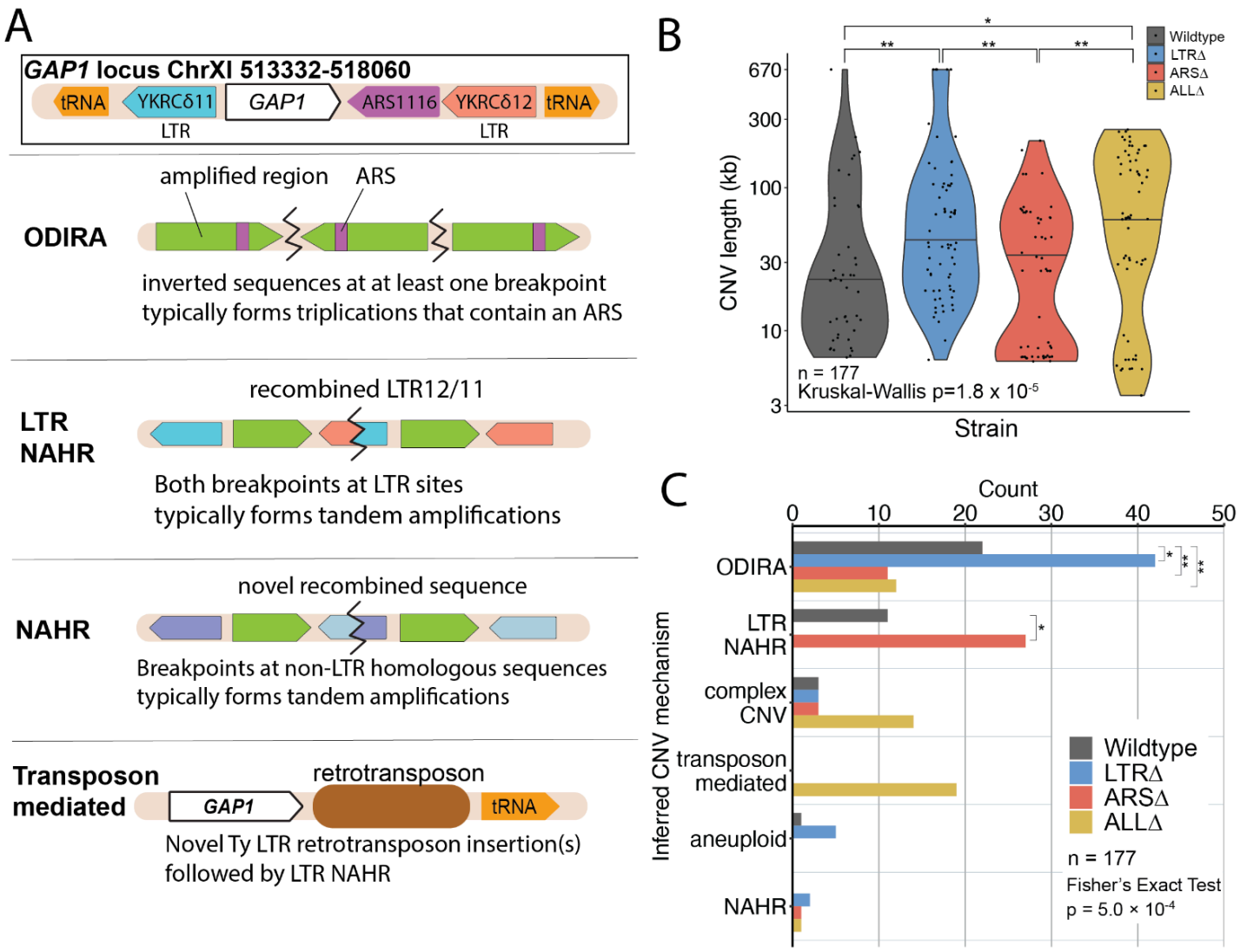
coefficients, the CNV formation rate is markedly lower in all strains ranging from $1{\cdot}10^{-5}$ for LTRΔ and ALLΔ to $2.4{\cdot}10^{-6}$ in ARSΔ. These results are consistent with our hypothesis that proximate sequence features facilitate *GAP1* CNV formation.

In all strains, collective parameter estimations are highly correlated, as expected for joint estimation of selection coefficients and beneficial mutation rates (Gitschlag et al., 2023). Thus, the collective MAP predictions can be interpreted as a range of parameter values for each strain. Indeed, other than the very final time point for the WT population, all collective MAP predictions lay within the interquartile ranges (**Supplementary Figure S3B**). The observed *GAP1* CNV frequency stabilizes at different levels in the different experiments (**Figure 1B**). This can be explained by pre-existing unreported CNVs with frequencies estimated to be between φ=$4{\cdot}10^{-6}$ to $1.6{\cdot}10^{-4}$ by the collective MAPs (**Supplementary Figure S3C**). Indeed, our model predicts that the total (reported and unreported) final CNV frequency is nearly one in all cases (**Supplementary Figure S3G**).

Next, we estimated the *de novo* CNV diversity. Previous work showed a diversity of CNV allele types formed under glutamine-limited selection including tandem duplications, segmental amplification, translocations, and whole chromosome amplification (Lauer et al., 2018), and that lineage richness decreases rapidly over the course of evolution due in part to competition and clonal interference (Lauer et al., 2018; Levy et al., 2015; Nguyen Ba et al., 2019). Our model does not include competition, clonal interference, or recurrent CNV formation. Therefore, diversity calculations are likely overestimations. Nonetheless, a comparison of diversity between strains is informative of whether proximate genome elements affect CNV allele diversity. Therefore, for each strain, we used its collective MAP to simulate a posterior prediction for the genotype frequencies (**Figure 3C**), which we then used to predict the posterior Shannon diversity (Jost, 2006). We predict the set of CNV alleles to be highly diverse: the final predicted Shannon diversity ranges from $1.6{\cdot}10^{4}$ in ARSΔ to $3.2{\cdot}10^{5}$ in WT (**Figure 3B**). Our model predicts that the diversity increases rapidly during the selection phase and stabilizes in the equilibrium phase. This is because CNV alleles that form towards the end of the experiment would have a low frequency with a minor effect on diversity. We observe the greatest diversity in WT populations with lower diversity in the three genomic architecture mutants. Moreover, diversity saturates faster in WT populations. This suggests that the WT strain is able to form more unique CNVs allele types earlier compared to the other three strains (**Figure 3B**). Shannon diversity is lower in LTRΔ and further lower in ALLΔ and ARSΔ (**Figure 3B**). This rank order is the same as the CNV formation rates (**Figure 3A**).

We used a modified version of the evolutionary model with the inferred parameters to simulate an evolutionary competition between WT and the three architecture mutant strains over 116 generations, a point in which CNVs have reached high frequency in the original experiment. To “win” these competitions, the competitor strains need to adapt to glutamine-limitation by producing CNVs (**Supplementary Figure S3J** and supplementary information). The results of the simulated evolutionary competitions predict that the WT would dominate over all three strains, as its predicted final proportion almost always exceeds its initial frequency of 0.5. The average predicted proportion of WT cells when competing with LTRΔ is 0.717. In contrast, ARSΔ and ALLΔ are predicted to be

almost eliminated by generation 116, as the average predicted WT proportion is 0.998 and 0.999, respectively. These simulated competitions further suggest that the ARS is a more important contributor to adaptive evolution mediated by *GAP1* CNVs.



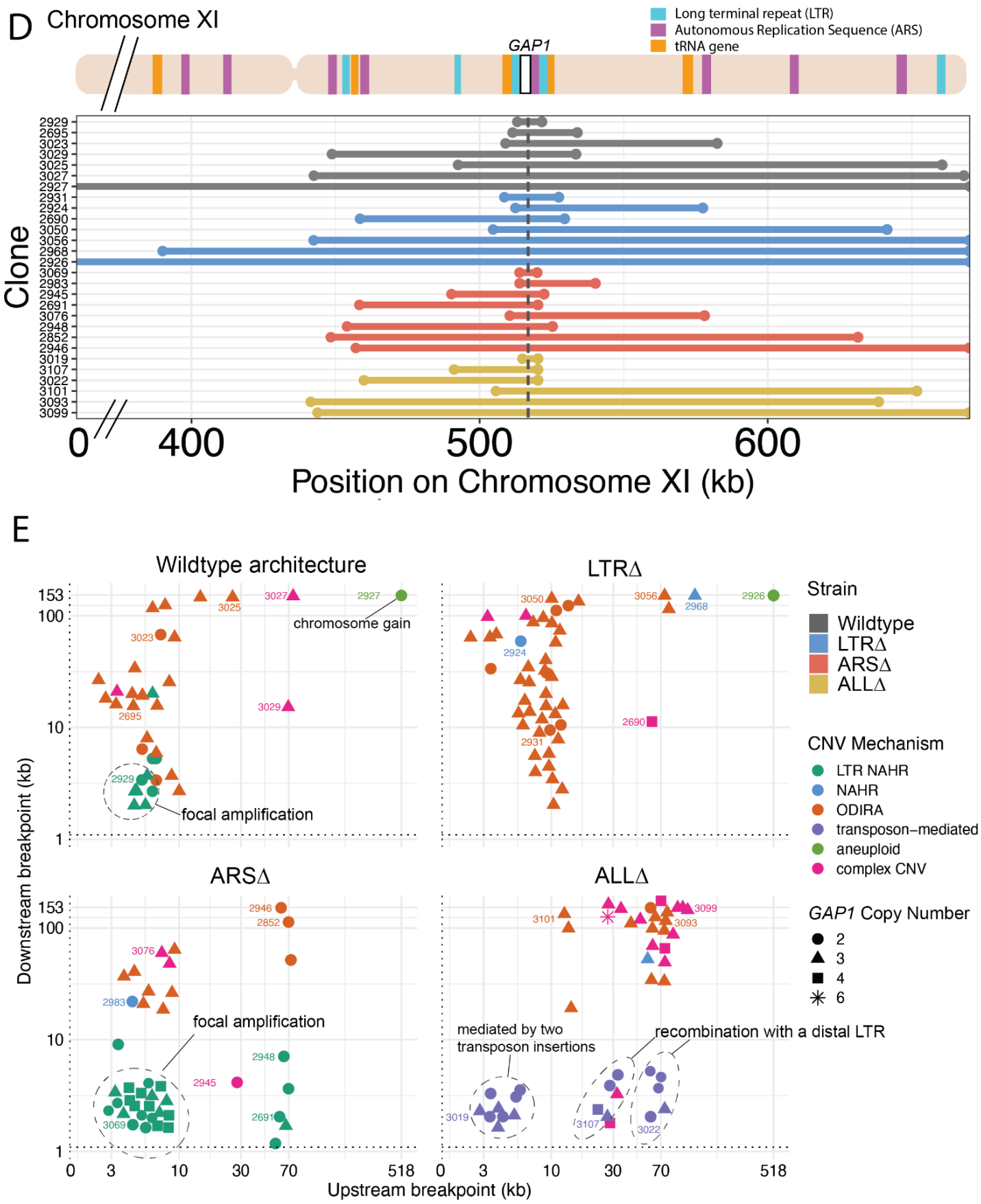

D
Chromosome XI
Long terminal repeat (LTR)
Autonomous Replication Sequence (ARS)
tRNA gene
GAP1
Clone
2929
2695
3023
3029
3025
3027
2927
2931
2924
2690
3050
3056
2968
2926
3069
2983
2945
2691
3076
2948
2852
2946
3019
3107
3022
3101
3093
3099
0
400
500
600
Position on Chromosome XI (kb)
E
Wildtype architecture
LTRΔ
ARSΔ
ALLΔ
chromosome gain
focal amplification
mediated by two transposon insertions
recombination with a distal LTR
Downstream breakpoint (kb)
Upstream breakpoint (kb)
153
100
10
1
0
3
10
30
70
518
Strain
Wildtype
LTRΔ
ARSΔ
ALLΔ
CNV Mechanism
LTR NAHR
NAHR
ODIRA
transposon–mediated
aneuploid
complex CNV
GAP1 Copy Number
2
3
4
6

Figure 4. Local and distal elements contribute to generation of *GAP1* CNV alleles. (A) Schematic of *Saccharomyces cerevisiae GAP1* locus on Chromosome XI: 513332-518060 with LTR, ARS elements and tRNA genes labeled. ODIRA is a DNA replication-error based CNV mechanism. Here, we classify a clone as ODIRA if it has an inverted sequence in at least one breakpoint. ODIRA typically forms tandem triplications with an inverted middle copy and contains an ARS. Long terminal repeat non-allelic homologous recombination (LTR NAHR) is a mechanism we define by having both CNV breakpoints at LTR sites. Sometimes we detect a hybrid sequence between two LTR sequences, a result of recombination between the two LTRs. Non-allelic homologous recombination (NAHR) is defined by having at least one CNV breakpoint not at LTR sites, ie. other homologous sequences in the genome. Sometimes we detect a hybrid sequence between the two homologous sequences. Transposon-mediated mechanisms observed involve at least one intermediate novel LTR retrotransposon insertion. The newly deposited LTR sequences recombines with other LTR sequences, either pre-existing or introduced by a second *de novo* retrotransposition, to form a resulting CNV. Complex CNV is defined by a clone having more than two breakpoints in chromosome XI, indicative of having more than one amplification event. (**B)** Violin plot of CNV length in each genome-sequenced clone, n = 177. Strain has a significant effect on CNV length, Kruskal-Wallis test, $p = 3.0 \times 10^{-4}$. (**C)** Barplot of inferred CNV mechanisms, described in A, for each CNV clone isolated from evolving populations. Inference came from a combination of read depth, split read, and discordant read analysis (see Methods). Strain is significantly associated with CNV Mechanism Fisher's Exact Test, $p = 5.0 \times 10^{-4}$. There is a significant increase in ODIRA prevalence between WT and LTR∆, chi-sq, p = 0.02469. There is a significant decrease in ODIRA prevalence from WT to ARS∆ and ALL∆, chi-sq, p = 0.002861 and 0.002196, respectively. There is a significant decrease of LTR NAHR from WT to LTR∆, chi-sq, p = 0.03083.
(D) **<u>Top</u>**: Schematic of *S. cerevisiae* chromosome XI, with LTR, ARS elements, tRNA genes annotated. LTR-blue, ARS-purple, tRNA-orange, *GAP1* ORF-white rectangle. Using a combination of read depth, split read, and discordant read analysis, we defined the extent of the amplified region, the precise CNV breakpoints, and *GAP1* copy number. *GAP1* copy numbers were estimated using read depth relative to the average read depth of chromosome XI. We define the upstream and downstream breakpoints as kilobases away from the start codon of the *GAP1* ORF (vertical dotted line). **<u>Bottom:</u>** Dumbbell plots represent the amplified region (>1 copy) relative to the WT architecture reference genome. The ends of the dumbbells mark the approximate CNV breakpoints. Select clones were chosen as representative of the observed diversity of amplifications.
(E) Scatterplots of CNV length for all genome-sequenced clones, n = 177. We defined the upstream and downstream breakpoints as kilobases away from the start codon of the *GAP1* ORF (vertical dotted line in dumbbell plot). CNV mechanisms are defined in **Figure 4A**.

Inference of CNV mechanisms in genome architecture mutants

Contrary to our expectations, removal of proximate genomic elements from the *GAP1* locus does not inhibit the formation of *GAP1* CNVs. We sought to determine the molecular basis by which *GAP1* CNVs form in the absence of these local elements. Therefore, we isolated ~40 *GAP1* CNV-containing clones from each population containing the four different strains at generations 79 and 125 and performed Illumina whole-genome sequencing. Using a combination of read depth, split read, and discordant read analysis, we defined the extent of the amplified region, the precise CNV breakpoints, and *GAP1* copy number. On the basis of these features, we inferred the CNV-forming mechanisms for each *GAP1* CNV (**Methods**). Over the 177 analyzed *GAP1* CNVs, we observed tandem amplifications, tandem triplications with an inverted middle copy, intra-chromosomal translocations, aneuploidy, and complex CNVs. *GAP1* copy numbers range from two to six in any given clone. Each of the four strains is able to produce a diversity of CNV alleles ranging from small (tens of kilobases) to large (~hundreds of kilobases) segmental amplifications (**Figure 4**). We identified six major CNV-forming mechanisms across the four strains: ODIRA, LTR NAHR, NAHR, transposon-mediated, complex CNVs, and whole chromosome duplication (aneuploidy) (**Figure 4A** and **Methods**).

ODIRA is a predominant mechanism of CNV formation

We inferred *GAP1* CNVs formed through ODIRA in all four genotypes at high frequencies: 22 out of 37 WT clones (59%), 42 out of 52 LTR∆ clones (81%), 11 out of 42 ARS∆ clones (26%), and 12 out of 46 ALL∆ clones (26%). Considering the set of all CNVs in all strains, ODIRA is the most common CNV mechanism comprising almost half of all CNVs (87/177, 49%). The second most common mechanism is NAHR between flanking LTRs (38/177, 21%), which generate tandem amplifications. In the WT background, ODIRA (22/37) and NAHR between LTRs (11/37) account for 89% of *GAP1* CNVs.

In LTR∆, *GAP1* CNVs form via ODIRA, chromosome missegregation, and NAHR using other sites. As expected, in LTR∆ clones we did not detect NAHR between LTRs in 52 clones and as a result focal amplifications were not detected (**Figure 4C).** In LTR∆**,** CNVs are formed predominantly by ODIRA (42/52, 81%) (**Supplementary Table 3**), a significant increase relative to WT clones (chi-sq, p = 0.02469) (**Figure 4C)**. By contrast, aneuploidy (5/52), complex CNV (3/52), and NAHR (2/52) account for less than 10% of *GAP1* CNVs in LTR∆. Consequently, we observe an increase in *GAP1* CNV size in LTR∆ relative to WT (**Figure 4B)** as there is an increased prevalence of segmental amplifications and aneuploidy (**Figure 4E**).

Aneuploidy was rarely observed. Whole amplification of chromosome XI was detected in six out of 177 clones (3.4%) (**Figure 4C**). It was detected only in 2 strains: WT and LTR∆ (**Figure 4D**).

ODIRA generates CNVs using distal ARS

Whereas removal of proximate LTRs abrogates the formation of CNVs through NAHR, removal of the local ARS does not prevent the formation of *GAP1* CNVs through ODIRA (**Figure 4C**). In the absence of the proximate ARS, distal ones are used to form ODIRA as all amplified regions of ODIRA clones contain a distal ARS (**Figure 4D),** with one exception (**see Methods, Supplementary Figure S4B**). We observe an increase of LTR NAHR in the ARS∆ clones (27/52, 52%) relative to WT clones (11/37, 39%) (**Figure 4C,** chi-sq, p = 0.03083). In ARS∆, we find two CNV size groups (**Figure 4B**). Small amplifications are formed via NAHR of the flanking LTRs (**Figure 4E**) and large segmental amplifications are formed via ODIRA and by NAHR between one local and one distal LTR (**Figure 4E**).

Novel retrotransposition events potentiate *GAP1* CNVs

CNVs in the ALL∆ clones form by two major mechanisms: 1) ODIRA using distal ARS sites to form large amplifications and 2) LTR NAHR following novel Ty LTR retrotransposon insertions to form focal amplifications (transposon-mediated, **Figure 4**). ALL∆ clones have larger amplifications formed by ODIRA than ODIRA-generated amplification in WT and LTR∆ (**Figure 4E**) because they encompass distal ARS and inverted repeats (**Figure 4D**). Surprisingly, we detected novel LTR retrotransposon events that generated new LTRs that subsequently formed *GAP1* CNVs through NAHR with a pre-existing LTR in the genome or an LTR from a second novel retrotransposition (**Figure 4E**). Regions upstream of tRNA genes are known to be hotspots for Ty retrotransposons (Ji et al., 1993; Mularoni et al., 2012). We find the novel retrotransposons insert near one or both of the previously deleted LTR sites (**Supplementary File 1**), which flank *GAP1* and are downstream of

tRNA genes (**Figure 1A**). We only detected novel retrotranspositions in ALL∆ populations. In total we detected 15 unique Ty retrotransposon insertion sites of which eight were upstream of the deleted LTR, YKRCδ11, and four were downstream of deleted LTR, YKRCδ12 (**Supplementary File 1**). The remaining two insertions were distal to the *GAP1* gene: one on the short arm and the second on the long arm of chromosome XI. Every insertion was upstream of an tRNA gene, consistent with the biased preference of Ty LTR insertions (Ji et al., 1993; Mularoni et al., 2012). Recombination between a new and preexisting LTR produces large amplifications whereas recombination between two newly inserted Ty1 flanking the *GAP1* gene forms focal amplifications of the *GAP1* gene (**Figure 4E**).

Discussion

In this study we sought to understand the molecular basis of repeated *de novo* amplifications and selection of the general amino acid permease gene, *GAP1*, in *S. cerevisiae* continuously cultured in glutamine-limited selection. We hypothesized that a high formation rate of *GAP1* CNVs is due to the unique genomic architecture at the locus, which comprises two flanking long terminal repeats and a DNA replication origin. We used genetic engineering, experimental evolution, and neural network simulation-based inference to quantify *de novo* CNV dynamics and estimate the CNV formation rate and selection coefficient in engineered mutants lacking the proximate genome elements. We find that removal of these elements has a significant impact on *de novo* CNV dynamics, CNV formation rate, and selection coefficients. However, CNVs are formed and selected in the absence of these elements highlighting the plasticity of the genome and diversity of mechanisms that generate CNVs during adaptive evolution.

Despite their proximity to *GAP1* we found that flanking LTRs are not an essential driver of CNV formation. The *de novo* CNV dynamics of WT and LTR∆ populations are similar and we find that although the CNV formation rate is reduced, the effect is small. In contrast, a significantly decreased CNV formation rate and delayed CNV appearance time was observed in the absence of the ARS in ARS∆ and ALL∆ populations, which suggests that the local ARS is a major determinant of *GAP1* CNV-mediated adaptive dynamics. Indeed, ODIRA was identified as the predominant CNV mechanism in sequence-characterized clones revealing that DNA replication errors are a major source of CNV formation during adaptive evolution.

The prevalence of ODIRA is a result of many replication origins and the pervasiveness of inverted repeat sequences throughout the chromosome (**Figure 4**). In particular, breakpoint analysis of LTR∆ CNV clones show that ODIRA produces a continuum of CNV sizes along the short arm of chromosome XI. Downstream breakpoints range from near the *GAP1* gene *(*~3 kilobases*)* to all the way to the right telomere of chromosome XI (153 kilobases) (**Figure 4E**). The *S. cerevisiae* genome contains a high number, 1-1000 per kilobase, of inverted repeats ranging from 3bp to 14bp (Martin et al., 2024). The longer repeats are more likely to be used in ODIRA (Martin et al., 2024). The ubiquity of inverted repeats is in stark contrast to the relative paucity of LTR sequences, which are dispersed

throughout the genome. Thus, ODIRA supplies a diverse and high number of gene amplifications for selection to act on, setting the stage for genome evolution and adaptation.

Consistent with previous reports of increased Ty insertions in *S. cerevisiae* under stress conditions (Morillon et al., 2000, 2002), we observed novel retrotransposon insertions in populations evolved in glutamine-limited chemostats. Transposon insertions could be harmful and lead to loss-of-function mutations but are also substrate for generating beneficial alleles including CNVs (Blanc & Adams, 2003; Dunham et al., 2002; Gresham et al., 2008; Wilke & Adams, 1992). We only detected novel Ty insertions in the ALL∆ strain. This is likely because regions upstream of tRNA genes unoccupied by LTRs are predisposed to transposition. Our detection of novel retrotransposon insertions is consistent with a previous experimental evolution study that suggested that Ty insertions were rare under constant nitrogen-limitation and substantially more common under fluctuating nitrogen limitation, in which cells experience total nitrogen starvation periodically (Hays et al., 2023). We observe a substantially lower frequency of novel Ty retrotransposition events, 15/177 clones, compared to 898/345 clones, with multiple insertions per genome, isolated from serial transfer ammonium-limited experimental evolution (Hays et al., 2023). Importantly, the role of Ty differs in the two studies, as in our case beneficial CNV formed after novel retrotransposition through recombination of newly introduced LTR sequences, whereas Hays et al. found Ty associated null alleles that are beneficial in nitrogen-limited conditions. Our results reveal the prevalence of Ty insertions even under constant nitrogen-limiting conditions and demonstrate how new retrotransposition events can contribute to CNV formation.

Aneuploidy was not a major source of adaptation in our experiments as it was infrequently detected (n = 6/177). This contrasts with studies suggesting aneuploidy is a rapid and transient route to adaptation over short evolutionary time scales (Chen, Bradford, et al., 2012; Chen et al., 2015; Chen, Rubinstein, et al., 2012; Pavelka et al., 2010; Selmecki et al., 2015; Yona et al., 2012). However, aneuploidy incurs a fitness cost (Robinson et al., 2023; Tsai et al., 2019; Yang et al., 2021) and therefore can be outcompeted by slow-forming but less costly beneficial mutations in large populations (Kohanovski et al., 2024). Our observations of higher frequencies of focal and segmental amplifications may be because they are less costly than whole-chromosome amplifications.

A variety of DNA replication errors generate CNVs. Replication slippage at palindromic DNA and DNA repeats can cause fork stalling and downstream CNV formation (Lee et al., 2007; Zhang, Khajavi, et al., 2009). DNA repeats can form secondary structures like R loops, cruciforms, non-B DNA structures, and hairpins which stimulate CNV formation (Gu et al., 2008b). Untimely replication, faulty fork progression, S-phase checkpoint dysfunction, defective nucleosome assembly, and DNA repeat sites including LTRs are sources of replication-associated genome instability (Aguilera & García-Muse, 2013). However, DNA replication alone might not be the only contributor. The *GAP1* gene is highly transcribed under glutamine-limitation (Airoldi et al., 2016). Transcription-replication collisions may fuel ODIRA CNV formation at this locus (Lauer & Gresham, 2019; Wilson et al., 2015). CNV formation can also be stimulated by transcription-associated replication stress and histone acetylation (Hull et al., 2017; Salim et al., 2021; Whale et al., 2022) and replication fork stalling at

tRNA genes (Osmundson et al., 2017; Yeung & Smith, 2020). Testing the role of transcription in promoting the formation of adaptive CNVs warrants further investigation.

Recent work has proposed that ODIRA CNVs are a major mechanism of CNVs in human genomes (Brewer et al., 2015; Martin et al., 2024). Studies of human and yeast genomes have typically considered homologous recombination as the predominant mechanism of CNV formation (Lupski & Stankiewicz, 2005). CNV hotspots identified in the human (Chance et al., 1994; Lupski, 1998; Lupski & Stankiewicz, 2005; Pentao et al., 1992) and yeast genomes are indeed mediated by NAHR of long repeat sequences (Green et al., 2010; Gresham et al., 2010). However, a focus on recombination-based mechanisms as a means of generating copy number variation may be the result of ascertainment bias or the comparative ease of studying the effect of long repeat sequences over short palindromic ones. Our study demonstrates that experimental evolution in yeast is a useful approach to elucidating the molecular mechanisms by which DNA replication errors generate CNVs.

Methods

Strains and Media

Each of the three architecture mutants were constructed independently starting with the *GAP1* CNV reporter strain (DGY1657). To construct each deletion strain, we performed two rounds of transformations both using PCR amplified donor templates designed for homology-directed repair. The first transformation used a repair template containing a nourseothricin resistance cassette to replace the pre-existing kanamycin resistance cassette and *GAP1* gene. The repair template was designed to also deleted the elements of interests (ie. ARS, both flanking LTRs, or both LTRs and ARS). The second transformation replaced the nourseothricin cassette with a kanamycin resistance cassette and *GAP1* gene thus yielding a genomic architecture Δ strain. We confirmed scarless deletions with sanger sequencing and whole-genome-sequencing. Final identifiers are DGY1657 for the WT architecture strain, DGY2076 for the LTRΔ strain, DGY2150 for the ARSΔ strain, and DGY2071 for the ALLΔ strain.

DGY1, DGY500, and DGY1315 are zero-, one-, two-, copy GFP controls, respectively, described in Lauer et al, 2018 and Spealman et al. 2023 (Lauer et al., 2018; Spealman et al., 2023). Briefly, GFP under the ACT1 promoter was inserted at neutral loci that do not undergo amplification in glutamine-limited continuous culture. 400μM glutamine-limited media is described in Lauer et al. 2018.

Long-Term Experimental Evolution

We performed experimental evolution of 30 *S. cerevisiae* populations in miniature chemostats (ministats) for ~137 generations under nitrogen limitation with 400μM glutamine as in Lauer et al. (2018). Of the 30 populations, there were three controls: one control population with no fluorescent reporter (DGY1), one with one GFP fluorescent reporter (DGY500), one with two GFP fluorescent

reporters (DGY1315). The remaining 27 populations have the *GAP1* CNV GFP reporter. Of these, five populations are WT (DGY1657), seven are LTR∆ (DGY2076), seven are ARS∆ (DGY2150), and eight are ALL∆ (DGY2071). We inoculated each ministat containing 20ml of glutamine-limited media with 0.5 ml culture from its corresponding genotype founder population. The founder population was founded by a single colony grown overnight in glutamine-limited media at 30°C. Replicate populations of the same strain were inoculated from the same founder population derived from a single colony. Strains were randomized among the 30-plex ministat setup to account for the possibility of systematic position effects. After inoculation, populations were incubated in a growth chamber at 30°C for 24 hours with the media inflow pump off. After 24 hours, the populations had reached early stationary phase and we turned on the media inflow pump and waited 4 hours for the populations to reach steady-state equilibrium, at which the population size was ~10^8 cells. This was generation zero. Ministats were incubated in a growth chamber at 30°C with a dilution rate of 0.12 culture volumes/hr. Since the ministats had a 20ml culture volume, the population doubling time was 5.8 hours. Approximately every 10 generations, we froze 2-ml samples of each population in 15% glycerol stored at -80°C. Approximately every 30 generations, we pelleted cells from 1-ml samples of each population and froze them at -80°C for genomic DNA extraction.

Flow Cytometry analysis to study *GAP1* CNV dynamics

To track *GAP1* CNV dynamics, we sampled 1-ml from each population approximately every 10 generations. We sonicated cell populations for 1 minute to remove any cell clumping and immediately analyzed samples on the Cytek Aurora flow cytometer. We sampled 100,000 cells per population and recorded forward scatter, side scatter, and GFP fluorescent signals for every cell. We performed hierarchical gating to define cells, single cells, unstained (zero-copy-GFP control) cells, cells with one copy of GFP (*GAP1*), and two or more copies of GFP (*GAP1*) (Spealman et al., 2023). First we gated for cells (filtered out any debris, bacteria) by graphing forward scatter area (FSC-A) against side scatter area (SSC-A). Second, we gated for single cells by graphing forward scatter area against forward scatter height and drawing along the resulting diagonal. Finally, we drew non-overlapping gates to define three subpopulations: zero copy, one copy, and two or more copies of GFP by graphing B2 channel area (B2-A), which detects GFP (excitation = 516 λ, emission = 529 λ), against forward scatter area (FSC-A). We note that the one copy and two copy events overlap some, which is a limitation in this experiment (Spealman et al., 2023).

We found that two architecture mutants, DGY2150 and DGY2071, had strain-specific GFP fluorescence even though they only harbored one copy of GFP. DGY2150 and DGY2071 had slightly higher fluorescence than the one copy GFP control strain, DGY500, but less than that of the two copy GFP control strain, DGY1315. The third architecture mutant, DGY2076, had the same GFP fluorescence as the one-copy GFP control strain (DGY500). We ruled out that they were spontaneous diploids by looking at forward scatter signals. The forward scatter signal was not different from that of the one copy control (a haploid) and was not as high as a diploid. Therefore due to strain-specific fluorescence, we decided to perform strain-based gating, ie. one set of gates for the WT strain, a second set of gates for the LTR∆ strain, and so on. Since the controls are also a strain of their own, they were not used to set universal gates for one-copy or two-copy. Thus, for each strain, we chose

the basis of our one-copy gate as the timepoint per strain in aggregate with the lowest median cell-sized normalized fluorescence. The two-or-more-copy (CNV) gate was drawn directly above and non-overlapping with the one-copy gate.

Quantification of dynamics

To obtain the proportion of CNVs for each population at each timepoint, we applied gates that correspond to zero-,one-, and two-or-more copy subpopulations. Using such proportion per population per timepoint, we summarize population CNV dynamics as follows (Lauer et al., 2018; Spealman et al., 2023). We calculate the generation of CNV appearance for each of the evolved populations. We defined CNV appearance as the generation where the proportion of CNV-containing cells first surpasses a threshold of 10% for three consecutive generations. Next, modified from Lang et al. (2011) (Lang et al., 2011) and Lauer et al. (2018) (Lauer et al., 2018), we calculate the percent increase in CNVs per generation for each evolved population. We compute the natural log of the proportion of the population with CNVs divided by the proportion of the population without CNVs for each timepoint. These proportions were obtained previously by gating. We plot these values across time and perform linear regression during the initial increase of CNVs. The slope of the linear regression is the percent increase in CNVs per generation. Finally, we calculate the time to CNV equilibrium, as defined by the generation at which a linear regression results in a slope < 0.005 after the selection phase.

Neural network simulation-based inference of evolutionary parameters

Evolutionary model. We developed a Wright-Fisher model that describes the evolutionary dynamics, similar to our previous study (Avecilla et al., 2022). In this study we have shown that a Wright-Fisher model is suitable for describing evolutionary dynamics in a chemostat. This is a discrete-time, non-overlapping generations model with a constant population size. Every generation has three stages: selection, in which the frequency of genotypes with beneficial alleles increases; mutation, in which genotypes can gain a single beneficial mutation or CNV; and drift, in which the population of the next generation is generated by sampling from a multinomial distribution. Our model follows the change in frequency of four genotypes (**Figure 2B**): A, the ancestor genotype; B, a cell with a non-CNV beneficial mutation; C^{+}, a genotype with two copies of *GAP1* and two copies of the CNV reporter; and C^{-}, a genotype with two copies of *GAP1* but only a single copy of the CNV reporter. CNV and non-CNV alleles are formed at a rate of δ_C and δ_B and have a selection coefficient of s_C and s_B, respectively. The frequency of genotype i is X_i. Unlike X_B and X_{C^+}, which may increase due to both mutation and selection, we assume that C^{-} is not generated after generation 0 (as experimental results suggest that the reporter is working properly). Hence, the C^{-} genotype only increases in frequency due to selection, with s_C as its selection coefficient. We assume C^{-} has an initial frequency φ. Model equations and further details are in the supplementary information.

Simulation-based inference. We use a neural network simulation-based inference method, Neural Posterior Estimation or NPE (Papamakarios, 2019) to estimate the joint posterior distribution of three model parameters, s_C, δ_C and φ, while the other parameters, s_B and δ_B are fixed to a specific value (**Table 1**). Inferring all five model parameters resulted in similar prediction accuracy and s_C and δ_C estimates.

Table 1. Model parameters and priors. Fixed parameters from (Avecilla et al., 2022; Hall et al., 2008; Joseph & Hall, 2004; Venkataram et al., 2016).

Parameter	Description	Prior / Fixed value
s_C	GAP1 CNV selection coefficient	$log_{10}(s_C) \sim U[-2, 0]$
δ_C	GAP1 CNV formation rate	$log_{10}(\delta_C) \sim U[-7, -0.3]$
φ	Proportion of pre-existing cells with GAP1 CNV	$log_{10}(\varphi) \sim U[-8, -2]$
s_B	Beneficial SNV selection coefficient	10^{-3}
δ_B	Beneficial SNV formation rate	10^{-5}

We applied NPE implemented in the Python package *sbi* (Tejero-Cantero et al., 2020) using a masked autoregressive flow (Papamakarios et al., 2018) as the neural density estimator: an artificial neural network that “learns” an amortized posterior of model parameters from a set of synthetic simulations. Posterior amortization allows us to estimate the posterior distribution P(θ|X) for a new observation X without the need to re-run the entire inference pipeline, i.e., generating new simulations and re-training the network (as is the case in sampling-based methods such as Markov chain Monte Carlo or MCMC).

We generated 100,000 synthetic observations simulated from our evolutionary model using parameters drawn from the prior distribution (Table 1). The neural density estimator was trained using early stopping with a convergence threshold of 100 epochs without decreases in minimal validation loss (the default in *sbi* is 20). Using 100 epochs as a threshold resulted in improved predictions. We validated that this improvement in prediction accuracy is not a result of over-fitting (**Supplementary Figure S3F**).

We validated the trained neural density estimator by measuring the coverage property: the probability of parameters to fall within the estimated posterior marginal 95% HDI. Then, we used the distribution of $(\frac{MAP}{True})$ (**Supplementary Figure S3D**) and posterior predictive checks (**Supplementary Figure S3B**) as quantitative and qualitative measures of prediction accuracy, respectively.

Collective posterior distribution. NPE estimates a posterior distribution per observation, i.e., $P(\theta|X)$. Given *n* observations $X_1, ...X_n$ generated from the same model distribution $P(\theta)$, where each observation is a time-series of GAP1 CNV frequency, NPE infers *n* individual posterior distributions, each conditioned on a single observation, $P\left(X_i\right)$. We estimate the *collective posterior distribution* based on n individual posteriors, that is, a posterior distribution conditioned on all observations,

$$P\left(X_1, ..., X_n\right)=\frac{P(\theta)^{1-n}\Pi_i\left[P\left(X_i\right)\right]}{\int P(\zeta)^{1-n}\Pi_i\left[P\left(X_i\right)\right]d\zeta} \text{ (eq. 1).}$$

This can be computed using the individual posteriors $P\left(X_i\right)$ and the prior $P(\theta)$ (see supplementary information for derivation).

However, as $P\left(X_i\right)$ could be infinitesimally small, a single observation could potentially reject a parameter value that is likely according to other observations. We want the collective posterior to be robust to such non-representative observations. Therefore, we define $P_\epsilon\left(X_i\right) = max(\epsilon, P\left(X_i\right))$ and use this quantity instead of $P\left(X_i\right)$ in eq. 1. For a correct choice of ϵ, the collective posterior mode should reflect a value with high posterior density for multiple observations, rather than a value that no individual posterior completely rejects. Thus, we implemented an amortized collective posterior: given a set of observations, $X_1, ...X_n$, an amortized posterior for individual observations, $P(\theta|X)$, a prior, $P(\theta)$, and ϵ, a collective posterior distribution $P\left(\theta|X_1, ..., X_n\right)$, can be computed. This distribution can then be used to evaluate the posterior distribution of specific parameters given the observations, generate random samples from the posterior distribution (e.g., using rejection sampling), or find the collective MAP.

We set $\epsilon = e^{-150}$ based on a visual grid-search. To find the normalizing factor (denominator in eq. 1), the integral is approximated by a dense Riemann sum (300^3 points). Maximizing the distribution, i.e., finding the collective MAP, is implemented using *scipy*'s *minimize* method with the Nelder-Mead algorithm.

Genetic diversity. Using our evolutionary model with the inferred parameters, we can estimate the diversity of CNV alleles in the experiments. For each strain, we used samples from its collective posterior to simulate a posterior prediction for the CNV allele frequencies (**Figure 3C**), which we then used to compute the posterior Shannon diversity (Jost, 2006), as detailed in the supplementary information.

Whole genome sequencing of isolated clones

Clones were isolated from archived populations and verified to harbor a *GAP1* CNV by measuring GFP fluorescence signal consistent with two or more copies. Populations of each strain from generation 79 were streaked out from the -80°C archive on YPD and incubated at 30°C for 2 days. Plates containing single colonies were viewed under a blue light to view GFP fluorescent colonies by

eye. Relative to the fluorescence of the 2 copy control strain, we picked single colonies that fluoresced as bright or brighter, reasoning that these colonies would likely contain *GAP1 CNVs*. Single colonies were used to inoculate cultures in glutamine-limited media and incubated at 30°C for 18 hours. The cultures were analyzed on the Cytek Aurora to verify they indeed harbored two or more copies of *GAP1* based on GFP fluorescence signal. For Illumina whole genome sequencing, genomic DNA was isolated using Hoffman-Winston method. Libraries were prepared using a Nextera kit and Illumina adapters. Libraries were sequenced on Illumina NextSeq 500 platform PE150 (2 x 150 300 Cycle v2.5) or Illumina NovaSeq 6000 SP PE150 (2x 150 300 Cycle v1.5). We also used custom Nextera Index Primers reported in S1 Table Baym et al. 2015 (Baym et al., 2015).

Breakpoint analysis and CNV mechanism inference in sequenced clones

Reference genomes. We created a custom reference genome for each of the genomic architecture mutants. The custom reference genome containing the *GAP1* CNV reporter in Lauer et al. 2018 (NCBI assembly R64) was modified to delete the flanking LTRs, single ARS, or all three elements.

Copy number estimation by read depth. The estimation of *GAP1* copy number from read depth used is described in Lauer et al. 2018, except we searched for ≥ 1000 base pairs of contiguous sequence. CNV boundaries were refined by visual inspection.

Structural variation calling and breakpoint analysis. Whole genome sequences of clones were run through CVish, a structural variant caller (Spealman, 2019). Structural variant calling was also done on each of the ancestor genomes: WT, ARS∆, LTR∆, and ALL∆. Output .bam files containing split reads and discordant reads of evolved clones and its corresponding ancestor were visualized on Jbrowse 2 to confirm locations of *de novo* CNV breakpoints and orientation of sequences at the novel breakpoint junctions. Novel contigs relative to the reference genome were outputted in addition to the supporting split reads that generated the contig. Blastn was used to verify orientation of contigs, namely inverted sequences used to define ODIRA (see Definitions of Inferred CNV Mechanisms).

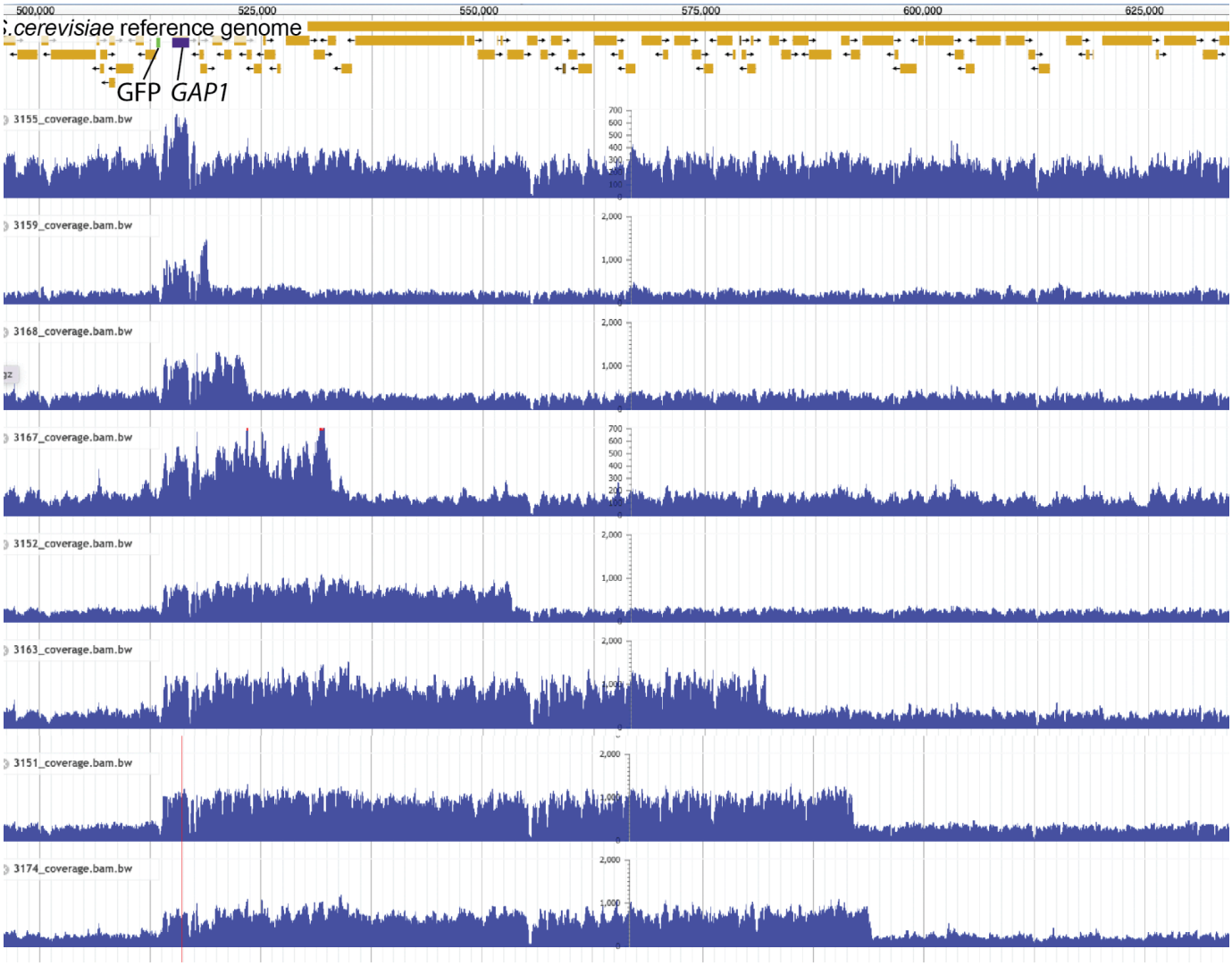
Definitions of Inferred CNV mechanisms. We used the following liberal classifications for each CNV category. We called a clone ODIRA if we found inverted sequences in at least one breakpoint (**Figure 4A**). We define LTR NAHR as having both breakpoints at LTR sites (**Figure 4A)**. This is evidence of recombination between the homologous LTR sequences. This mechanism typically forms tandem amplifications. In some cases, we find the hybrid sequence between two LTRs, but this is hard to retrieve in short-read-sequencing.
We define NAHR as having breakpoints at homologous sequences, with at least one breakpoint not at an LTR sequence (**Figure 4A**). We define transposon-mediated as a clone having a breakpoint at a novel LTR retrotransposon insertion and the other breakpoint at a different LTR site (**Figure 4A**). This supports that the newly deposited LTR sequence recombined with other LTR sequences (either pre-existing or introduced by a second *de novo* retrotransposition) to form CNVs. Sometimes we are able to recover the hybrid sequence between LTR sequences. We define complex CNV as having more than two breakpoints on chromosome XI and a read depth profile that looks like more than one

amplification event occurred (multi-step profile). For the complex CNV clones, we were not able to resolve the CNV mechanisms due to the limitations of short-read sequencing.

Supplement

Supplementary Table 1. Summary of genome sequence analysis of clones containing a single copy of the GAP1 CNV reporter. Estimated copy number of the GAP1 gene and inserted GFP gene of sequenced clones from five 1-copy-GFP subpopulations of the WT genome architecture strain. Copy number estimation is defined as the read depth of the target gene relative to the average read depth of the chromosome XI. Populations 1, 2, 4, 5 contain clones harboring GAP1 CNVs but only 1 copy of GFP. Population 3 and 5 contain clones containing 1 copy each of GAP1 and GFP suggesting these lineages have beneficial mutations elsewhere in the genome, allowing coexistence with the GAP1 CNV major subpopulation.

Sample	Generation	Population	Background Strain	*GAP1* copy number	GFP copy number
3150	182	1	WT	5	1
3171	182	1	WT	3	1
3172	182	1	WT	3	1
3173	182	1	WT	3	1
3174	182	1	WT	3	1
3151	153	2	WT	3	1
3152	153	2	WT	3	1
3153	153	2	WT	3	1
3154	153	2	WT	3	1
3155	153	2	WT	2	1
3156	182	3	WT	1	1
3157	182	3	WT	1	1
3158	182	3	WT	1	1
3175	182	3	WT	1	1
3176	182	3	WT	1	1
3177	182	3	WT	1	1
3178	182	3	WT	1	1
3179	182	3	WT	1	1
3180	182	3	WT	1	1
3181	182	3	WT	1	1
3182	182	3	WT	1	1
3159	166	4	WT	3	1
3160	166	4	WT	3	1
3162	166	4	WT	3	1
3163	166	4	WT	3	1
3164	182	5	WT	1	1
3165	182	5	WT	3	1
3166	182	5	WT	3	1
3167	182	5	WT	3	1
3168	182	5	WT	3	1



Supplementary figure S2A. **Independent *GAP1* amplifications lacking CNV reporter amplification**. Read depth plots of the *GAP1* CNV reporter locus of sequenced clones from five 1-copy-GFP subpopulations. Identification of eight distinct CNV breakpoint pairs, shown above, across the populations indicate the occurrence of at least eight independent amplifications of *GAP1* without GFP amplification. GFP reference gene - green rectangle, *GAP1* reference gene - purple rectangle.

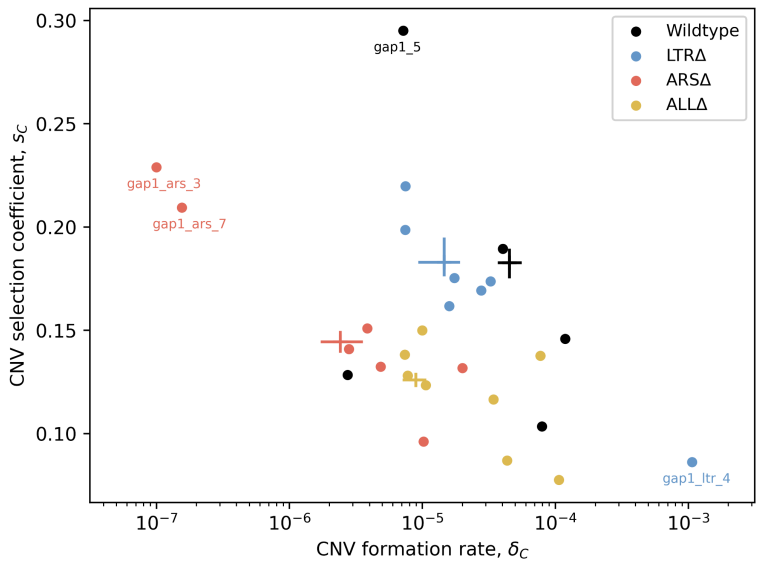


Figure S3A. MAP estimates of *GAP1* CNV formation rates (δ_C) and selection coefficients (s_C) for all replicate populations. Markers show MAP estimates from individual replicates, crosses show 50% HDI of collective posteriors. Extreme points are marked for comparison to data and posterior prediction, see Supplementary Figure S3B for posterior predictive checks.

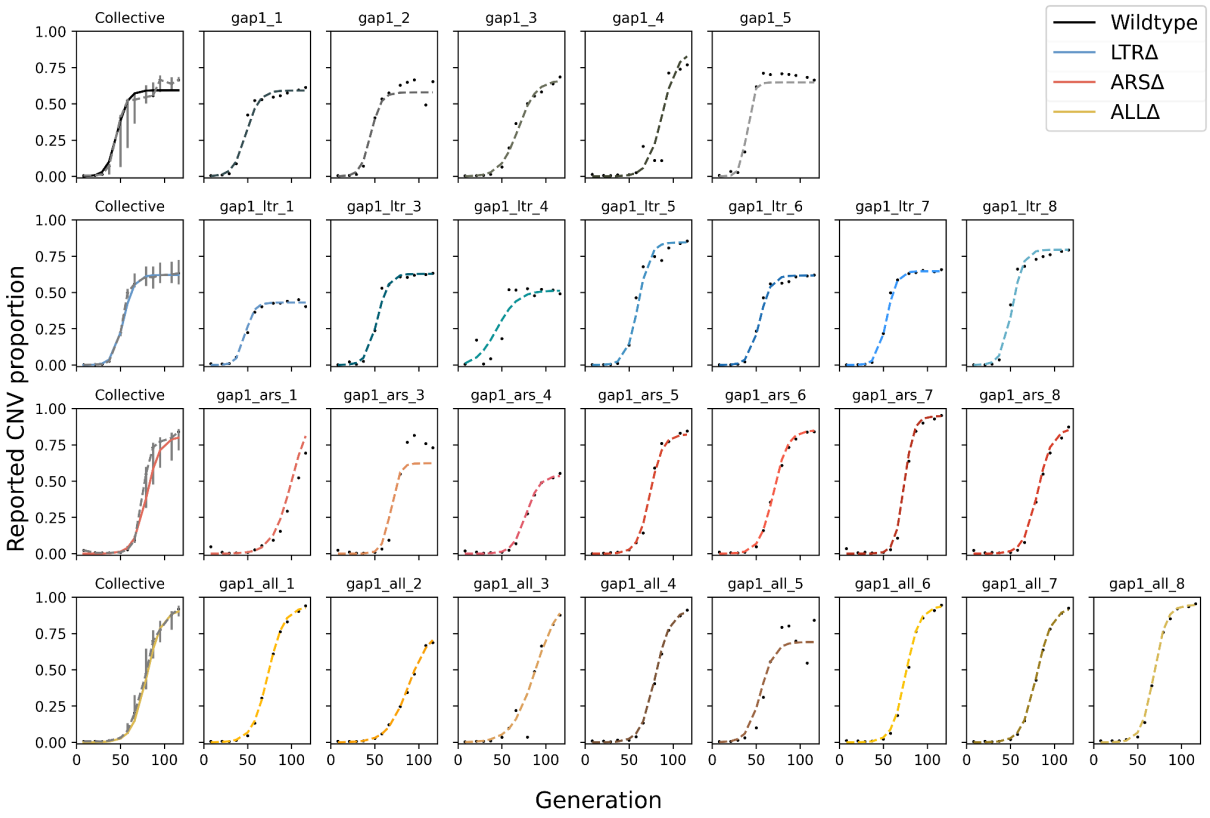


Figure S3B. Posterior predictive checks for all replicates. Black markers are the empirical observations, dashed line shows MAP prediction. The leftmost plot of each row shows the collective MAP prediction with empirical data's interquartile range (gray bars).

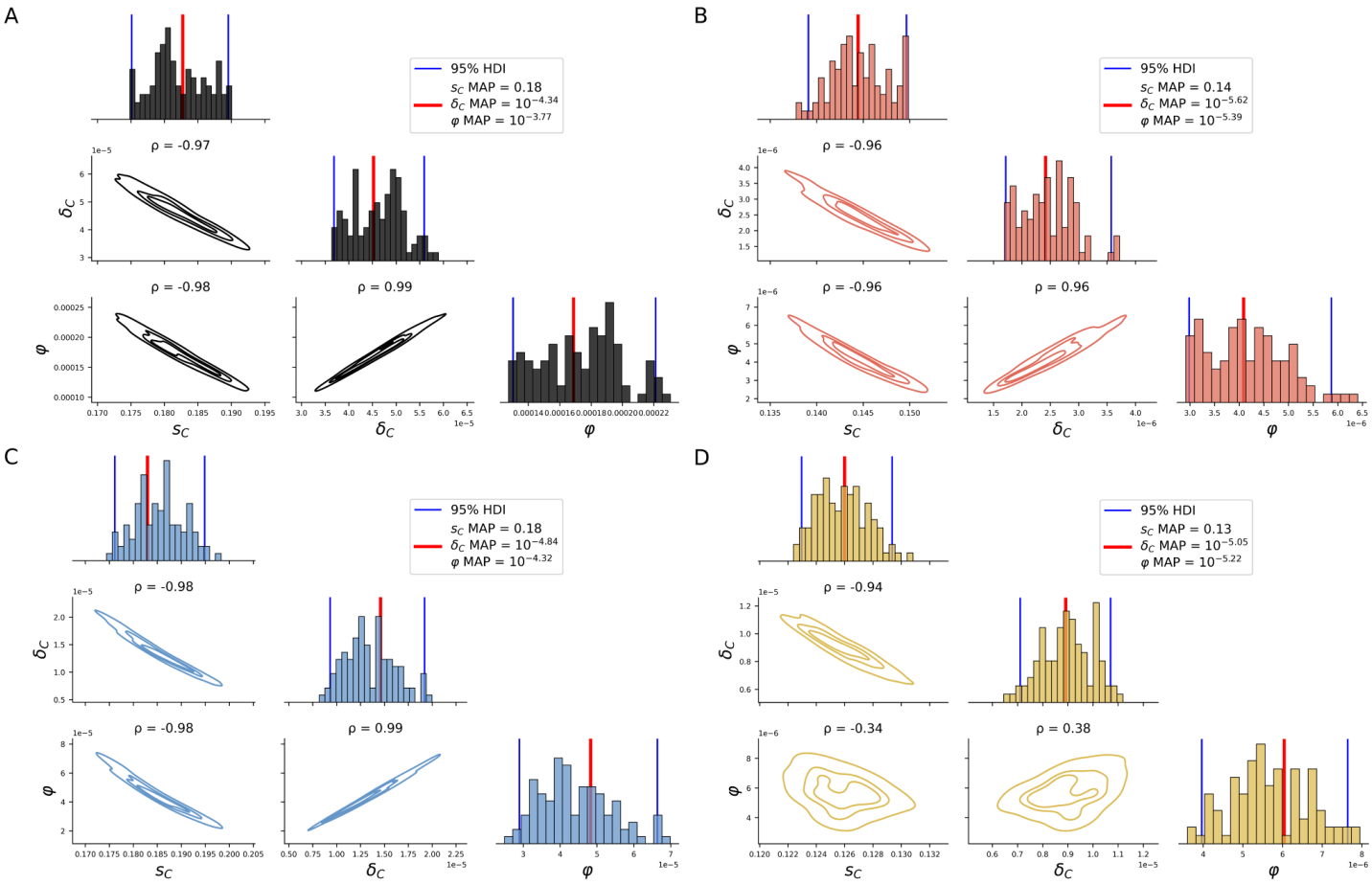


Figure S3C. Pairwise and marginal collective posteriors for all estimated model parameters. Diagonals show marginal collective posteriors per parameter per strain. Below-diagonal plots show pairwise KDEs for all pairs of model parameters. Collective joint MAPs (which may differ from collective marginal MAPs, as the marginal distribution integrates over all other parameters), are marked by a red vertical line. Panels are separated by strain: **(A)** WT, **(B)** ARSΔ, **(C)** LTRΔ, **(D)** ALLΔ.

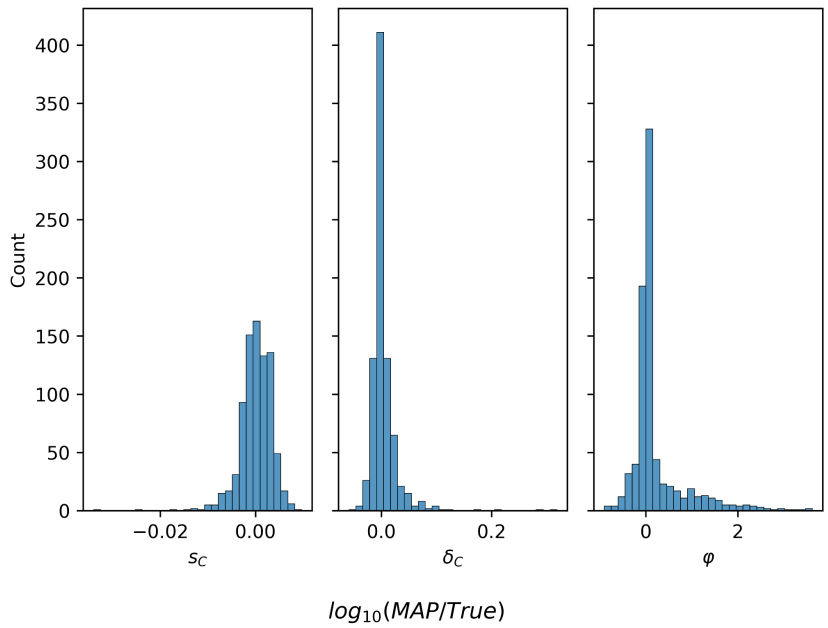


Figure S3D. Parameter estimation accuracy on synthetic data. Log-ratio of MAP estimate and true parameter value for 829 synthetic simulations in which the final reported *GAP1* CNV proportion is at least 0.3.

Parameter	Coverage
s_C	0.992
δ_C	0.995
φ	0.934

Supplementary Table 2. Estimation of network confidence. The coverage, defining the probability that the true parameter falls within the 95% highest density interval (HDI) of the posterior distribution, for 829 synthetic simulations in which the final reported *GAP1* CNV proportion is at least 0.3. 95% HDI was calculated for each simulation using 200 posterior samples. Our neural density estimator is slightly over-confident for φ (coverage of 0.934), and under-confident for *GAP1* CNV selection coefficient and formation rate (coverage of 0.992 for s_C and 0.995 for δ_C). Despite this under-confidence, the posterior distributions are narrow in biological terms: the 95% HDI represents less than an order of magnitude for both s_C and δ_C. Thus, we did not apply post-training adjustments to the neural density estimator, such as calibration (Cook et al., 2006) or ensembles (Caspi et al., 2023; Hermans et al., 2022).

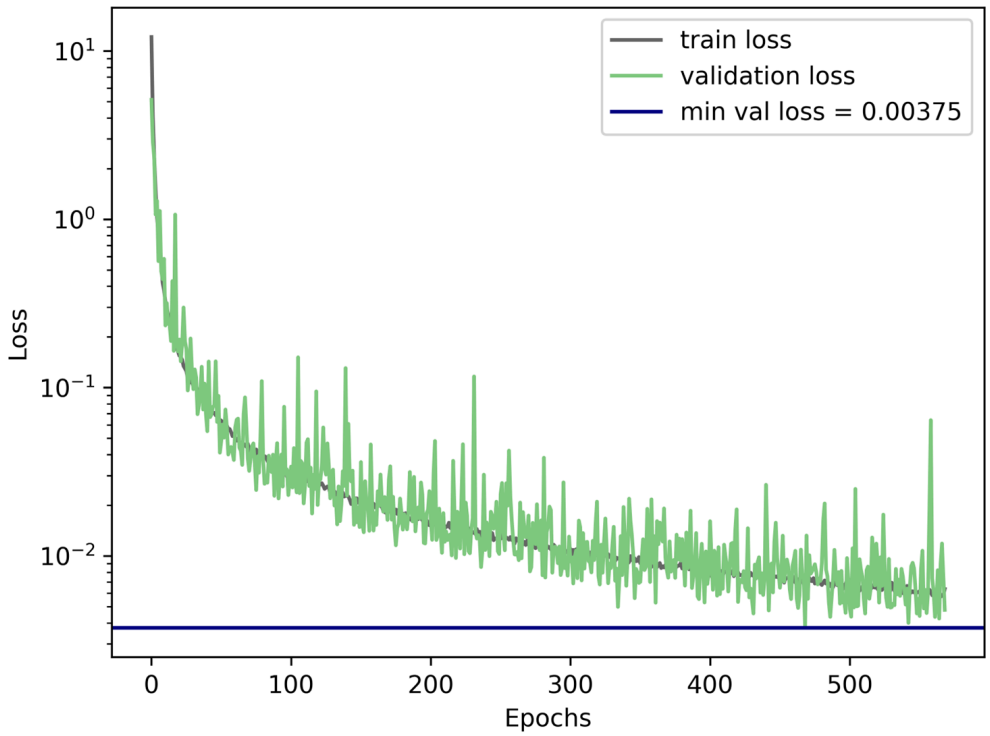


Figure S3F. Neural density estimator training and validation loss during training. Convergence threshold of 100 unimproved epochs (no decrease in minimal validation loss) was reached after 569 epochs.

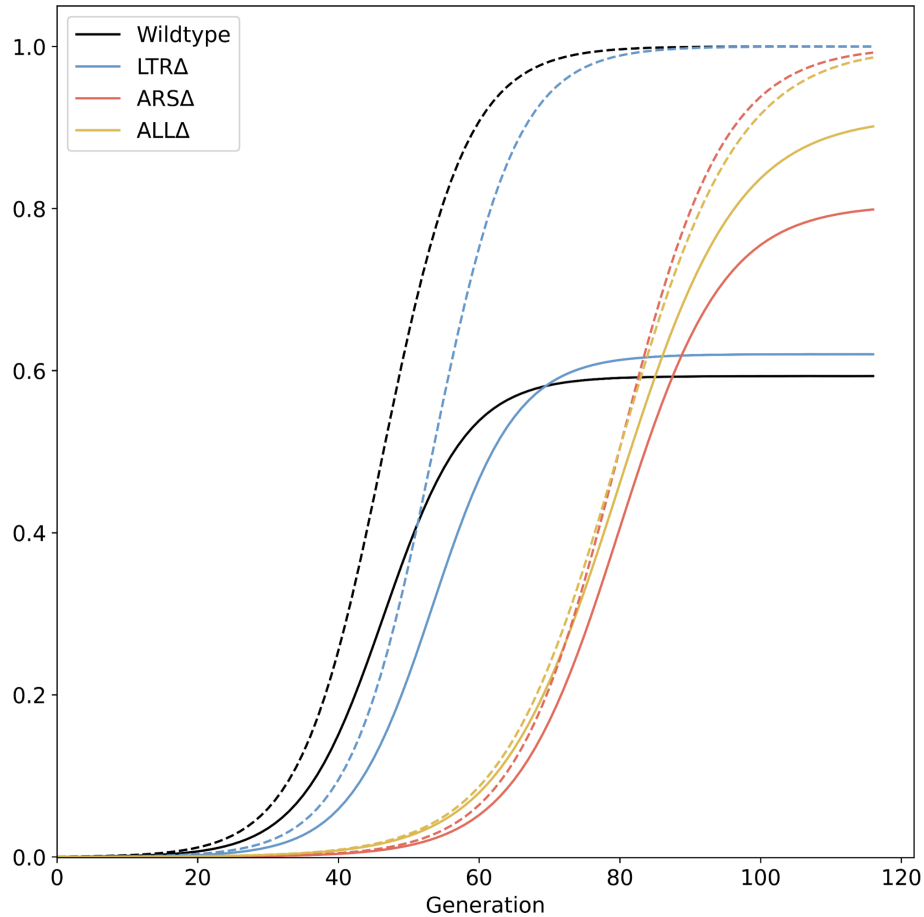


Figure S3G. Total *GAP1* CNV frequency. Solid lines show collective MAP predictions, dashed lines show the total proportion of *GAP1* CNVs, comprising unreported pre-existing CNVs and reported CNVs generated during the experiment, as predicted by the evolutionary model.

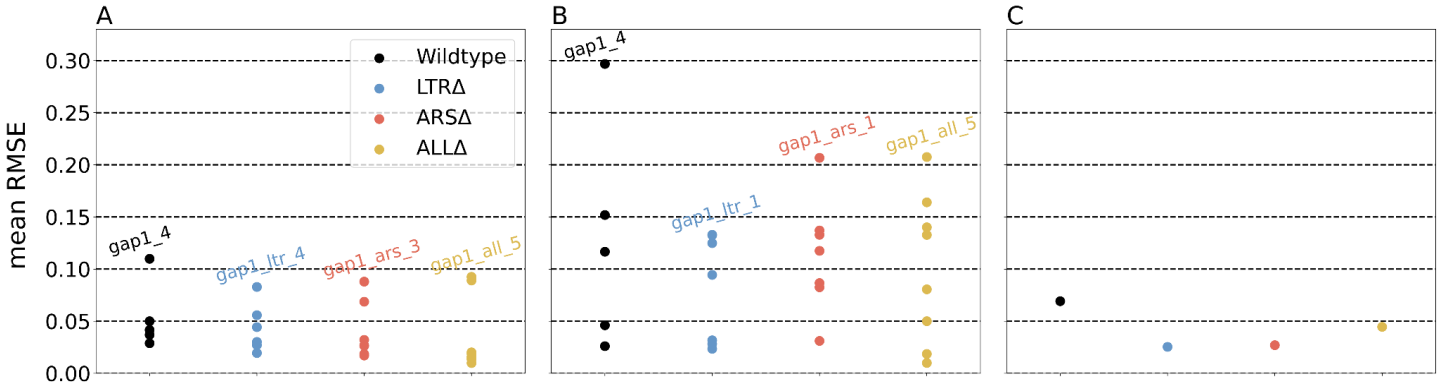


Fig S3H. Error estimation of parameter inference. Average root mean square errors (RMSE) of 50 posterior samples against the observed data. **(A)** Individual posteriors and individual replicates. **(B)** Collective posterior and individual replicates. **(C)** Collective posterior and empirical mean.

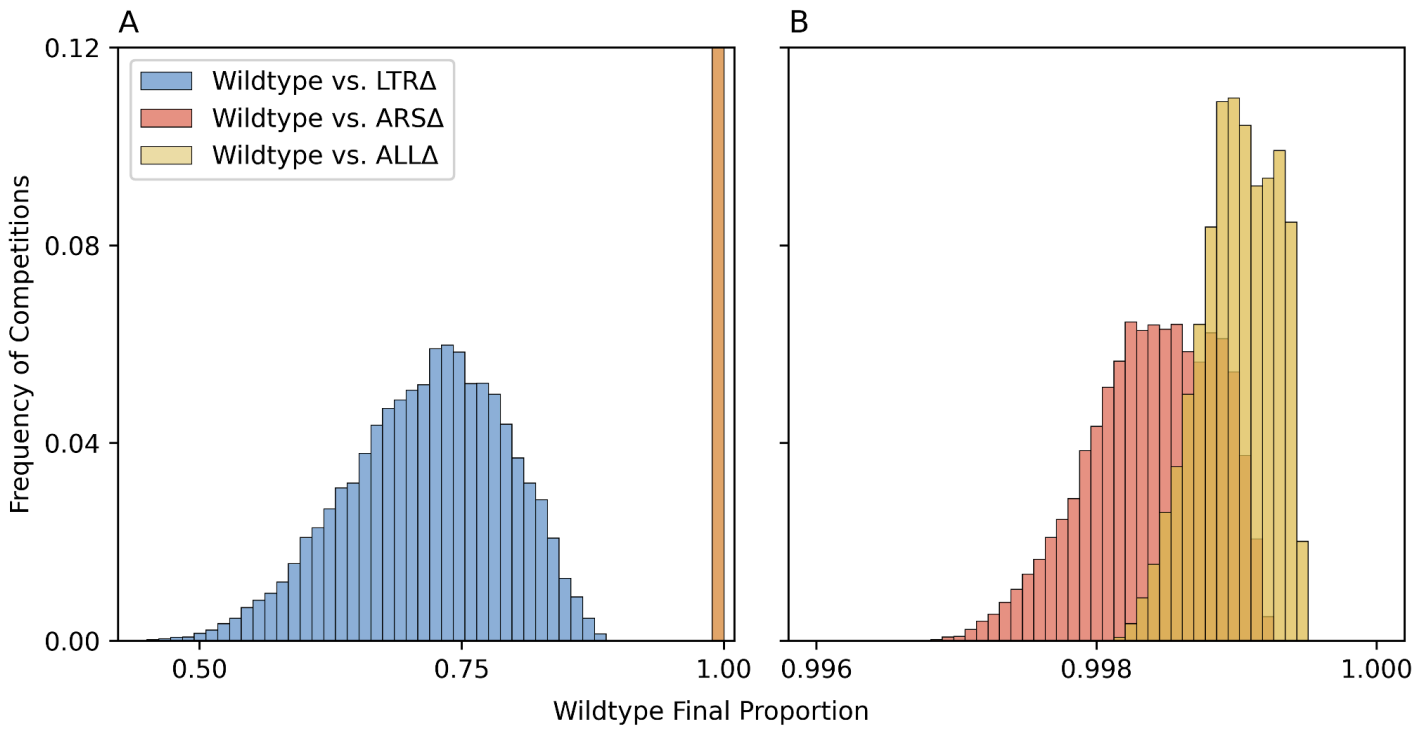
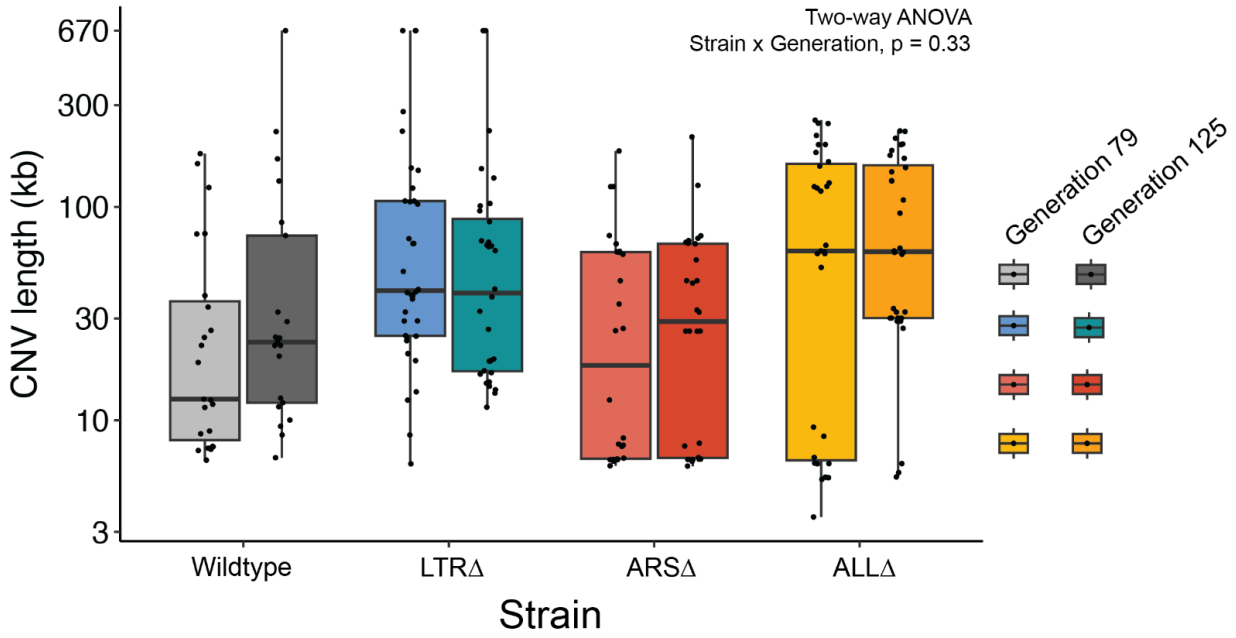
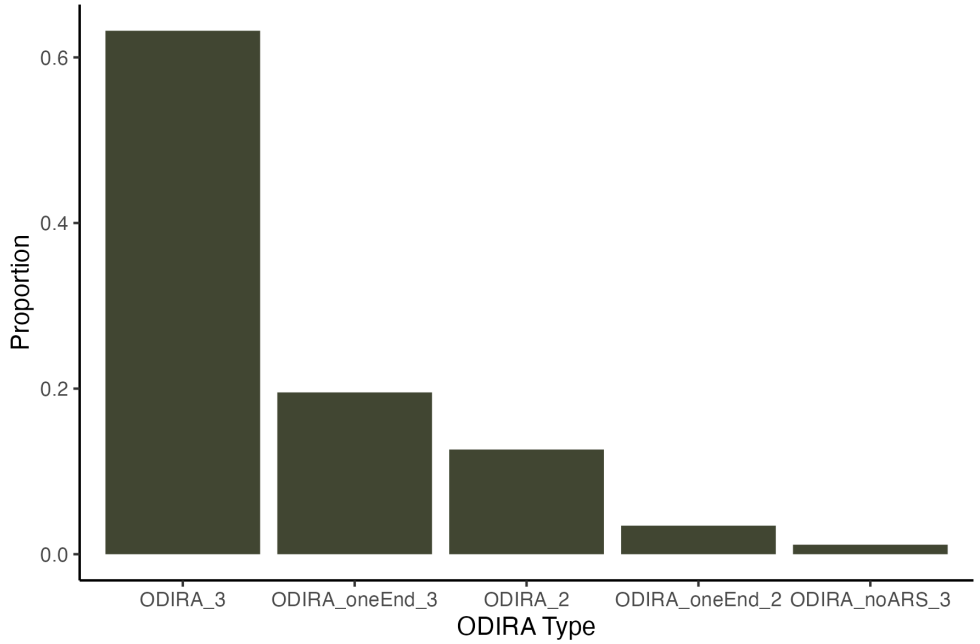


Figure S3J. Pairwise evolutionary competition predictions. We simulated evolutionary competitions in the experimental conditions of WT vs. genomic architecture mutants, starting from equal frequencies. The proportion at generation 116 of WT was predicted using 10,000 combinations of collective posterior samples for each pairwise competition. Overall, WT outcompetes all mutants because it adapts faster (due to faster CNV formation rate), but its advantage over ARSΔ and ALLΔ is much higher than its advantage over LTRΔ. (**A**) Histograms for three pairwise competitions. Note that ARSΔ and ALLΔ values overlap at this scale and are all in the rightmost bar. (**B)** High-resolution histograms for ARSΔ and ALLΔ.



Supplementary Figure S4A. No significant interaction between strain and generation on CNV length. Boxplot of CNV length of clones by strain and generation of isolation. There is no significant interaction between strain and generation of isolated clone, and no significant effect of generation on CNV length (Two-way ANOVA, Strain x Generation, p = 0.33)



Supplementary Figure S4B. **Types of ODIRA detected**. We found 87 ODIRA clones total regardless of strain. The majority of ODIRA clones fit the canonical definition of having two inverted junctions and 3 copies, 55/87 clones (63%) (ODIRA_3). We found four non-canonical types. We found 17 clones (20%) with only one inverted junction detected and 3 copies (ODIRA_oneEnd_3). We found 11 clones (13%) with two inverted junctions but only 2 copies (ODIRA_2). We found 3 clones (3.4%) with only one inverted junction detected and 2 copies (ODIRA_oneEnd_2). We found 1 clone (1.1%) with two inverted junctions but the amplified region did not contain an ARS.

Supplementary Table 3. Inferred CNV mechanisms by strain. Counts of inferred CNV mechanisms for each sequenced clone, n=177, separated by strain.

CNV Mechanism	WT	LTRΔ	ARSΔ	ALLΔ	Total
Aneuploid	1	5	0	0	6
Complex CNV	3	3	3	14	23
LTR NAHR	11	0	27	0	38
NAHR	0	2	1	1	4
ODIRA	22	42	11	12	87
Transposon-mediated	0	0	0	19	19
Total	37	52	42	46	177

Supplementary File 1. Ty-associated clones and locations of novel Ty insertions.

Supplementary File 2. CNV Clone Sequencing Analysis

Data availability

Sequencing data is available at SRA PRJNA1098800.
Other associated data are available here: https://osf.io/js7z8/.
Source code repository simulation-based inference: https://github.com/yoavram-lab/chuong_et_al.
Scripts for flow cytometry-based evolutionary dynamics and analysis of CNV clones: https://github.com/GreshamLab/local_arch_variants.

Acknowledgements

We thank all the members of the Gresham lab for helpful discussions, NYU Gencore for sequencing samples, and NYU High Performance Cluster for computing and storage. We thank Joshua Caleb Macdonald, Saharon Rosset, Uri Obolski, and Adi Stern for discussions and advice.
This work was supported by NSF GRFP DGE1839302 (JNC), NIGMS T32GM132037 (JNC), NIGMS R01GM134066 (DG), R01GM107466 (DG), and R35GM153419 (DG), NIAID R01AI140766 (DG), NSF 1818234 (DG), Israel Science Foundation (ISF, Y.R. 552/19), US–Israel Binational Science Foundation (Y.R. & D.G. 2021276), Minerva Center for Live Emulation of Evolution in the Lab (Y.R.) fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University (N.B.N), and fellowship from the AI and Data Science Center at Tel-Aviv University (N.B.N).

References

Aguilera, A., & García-Muse, T. (2013). Causes of genome instability. *Annual Review of Genetics*, *47*, 1–32. https://doi.org/10.1146/annurev-genet-111212-133232

Airoldi, E. M., Miller, D., Athanasiadou, R., Brandt, N., Abdul-Rahman, F., Neymotin, B., Hashimoto, T., Bahmani, T., & Gresham, D. (2016). Steady-state and dynamic gene expression programs in Saccharomyces cerevisiae in response to variation in environmental nitrogen. *Molecular Biology of the Cell*, *27*(8), 1383–1396. https://doi.org/10.1091/mbc.E14-05-1013

Arlt, M. F., Wilson, T. E., & Glover, T. W. (2012). Replication stress and mechanisms of CNV formation. *Current Opinion in Genetics & Development*, *22*(3), 204–210. https://doi.org/10.1016/j.gde.2012.01.009

Arndt, P. F., Hwa, T., & Petrov, D. A. (2005). Substantial Regional Variation in Substitution Rates in the Human Genome: Importance of GC Content, Gene Density, and Telomere-Specific Effects. *Journal of Molecular Evolution*, *60*(6), 748–763. https://doi.org/10.1007/s00239-004-0222-5

Avecilla, G., Chuong, J. N., Li, F., Sherlock, G., Gresham, D., & Ram, Y. (2022). Neural networks enable efficient and accurate simulation-based inference of evolutionary parameters from adaptation dynamics. *PLOS Biology*, *20*(5), e3001633. https://doi.org/10.1371/journal.pbio.3001633

Avecilla, G., Spealman, P., Matthews, J., Caudal, E., Schacherer, J., & Gresham, D. (2023). Copy number variation alters local and global mutational tolerance. *Genome Research*. https://doi.org/10.1101/gr.277625.122

Baym, M., Kryazhimskiy, S., Lieberman, T. D., Chung, H., Desai, M. M., & Kishony, R. (2015). Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes. *PLOS ONE*, *10*(5), e0128036. https://doi.org/10.1371/journal.pone.0128036

Ben-David, U., & Amon, A. (2020). Context is everything: Aneuploidy in cancer. *Nature Reviews. Genetics*, *21*(1), 44–62. https://doi.org/10.1038/s41576-019-0171-x

Blanc, V. M., & Adams, J. (2003). Evolution in Saccharomyces cerevisiae: Identification of Mutations Increasing Fitness in Laboratory Populations. *Genetics*, *165*(3), 975–983. https://doi.org/10.1093/genetics/165.3.975

Blount, Z. D., Barrick, J. E., Davidson, C. J., & Lenski, R. E. (2012). Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature*, *489*(7417), Article 7417. https://doi.org/10.1038/nature11514

Blount, Z. D., Borland, C. Z., & Lenski, R. E. (2008). Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. *Proceedings of the National Academy of Sciences*, *105*(23), 7899–7906. https://doi.org/10.1073/pnas.0803151105

Brewer, B. J., Payen, C., Rienzi, S. C. D., Higgins, M. M., Ong, G., Dunham, M. J., & Raghuraman, M. K. (2015). Origin-Dependent Inverted-Repeat Amplification: Tests of a Model for Inverted DNA Amplification. *PLOS Genetics*, *11*(12), e1005699. https://doi.org/10.1371/journal.pgen.1005699

Cardoso, A. R., Oliveira, M., Amorim, A., & Azevedo, L. (2016). Major influence of repetitive elements on disease-associated copy number variants (CNVs). *Human Genomics*, *10*, 30.

https://doi.org/10.1186/s40246-016-0088-9
Carvalho, C. M. B., Pehlivan, D., Ramocki, M. B., Fang, P., Alleva, B., Franco, L. M., Belmont, J. W., Hastings, P. J., & Lupski, J. R. (2013). Replicative mechanisms for CNV formation are error prone. *Nature Genetics*, *45*(11), 1319–1326. https://doi.org/10.1038/ng.2768
Caspi, I., Meir, M., Ben Nun, N., Abu Rass, R., Yakhini, U., Stern, A., & Ram, Y. (2023). Mutation rate, selection, and epistasis inferred from RNA virus haplotypes via neural posterior estimation. *Virus Evolution*, *9*(1), vead033. https://doi.org/10.1093/ve/vead033
Chance, P. F., Abbas, N., Lensch, M. W., Pentao, L., Roa, B. B., Patel, P. I., & Lupski, J. R. (1994). Two autosomal dominant neuropathies result from reciprocal DNA duplication/deletion of a region on chromosome 17. *Human Molecular Genetics*, *3*(2), 223–228. https://doi.org/10.1093/hmg/3.2.223
Chen, G., Bradford, W. D., Seidel, C. W., & Li, R. (2012). Hsp90 stress potentiates rapid cellular adaptation through induction of aneuploidy. *Nature*, *482*(7384), 246–250. https://doi.org/10.1038/nature10795
Chen, G., Mulla, W. A., Kucharavy, A., Tsai, H.-J., Rubinstein, B., Conkright, J., McCroskey, S., Bradford, W. D., Weems, L., Haug, J. S., Seidel, C. W., Berman, J., & Li, R. (2015). Targeting the Adaptability of Heterogeneous Aneuploids. *Cell*, *160*(4), 771–784. https://doi.org/10.1016/j.cell.2015.01.026
Chen, G., Rubinstein, B., & Li, R. (2012). Whole chromosome aneuploidy: Big mutations drive adaptation by phenotypic leap. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *34*(10), 893. https://doi.org/10.1002/bies.201200069
Cheng, Z., Ventura, M., She, X., Khaitovich, P., Graves, T., Osoegawa, K., Church, D., DeJong, P., Wilson, R. K., Pääbo, S., Rocchi, M., & Eichler, E. E. (2005). A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature*, *437*(7055), Article 7055. https://doi.org/10.1038/nature04000
Chuang, J. H., & Li, H. (2004). Functional Bias and Spatial Organization of Genes in Mutational Hot and Cold Regions in the Human Genome. *PLOS Biology*, *2*(2), e29. https://doi.org/10.1371/journal.pbio.0020029
Cook, S. R., Gelman, A., & Rubin, D. B. (2006). Validation of Software for Bayesian Models Using Posterior Quantiles. *Journal of Computational and Graphical Statistics*, *15*(3), 675–692. https://doi.org/10.1198/106186006X136976
Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, *117*(48), 30055–30062. https://doi.org/10.1073/pnas.1912789117
Di Rienzi, S. C., Collingwood, D., Raghuraman, M. K., & Brewer, B. J. (2009). Fragile Genomic Sites Are Associated with Origins of Replication. *Genome Biology and Evolution*, *1*, 350–363. https://doi.org/10.1093/gbe/evp034
Dunham, M. J., Badrane, H., Ferea, T., Adams, J., Brown, P. O., Rosenzweig, F., & Botstein, D. (2002). Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(25), 16144–16149. https://doi.org/10.1073/pnas.242624799
Gitschlag, B. L., Cano, A. V., Payne, J. L., McCandlish, D. M., & Stoltzfus, A. (2023). Mutation and

Selection Induce Correlations between Selection Coefficients and Mutation Rates. *The American Naturalist*, *202*(4), 534–557. https://doi.org/10.1086/726014

Gonçalves, P. J., Lueckmann, J.-M., Deistler, M., Nonnenmacher, M., Öcal, K., Bassetto, G., Chintaluri, C., Podlaski, W. F., Haddad, S. A., Vogels, T. P., Greenberg, D. S., & Macke, J. H. (2020). Training deep neural density estimators to identify mechanistic models of neural dynamics. *eLife*, *9*, e56261. https://doi.org/10.7554/eLife.56261

Green, B. M., Finn, K. J., & Li, J. J. (2010). Loss of DNA Replication Control Is a Potent Inducer of Gene Amplification. *Science*, *329*(5994), 943–946. https://doi.org/10.1126/science.1190966

Gresham, D., Desai, M. M., Tucker, C. M., Jenq, H. T., Pai, D. A., Ward, A., DeSevo, C. G., Botstein, D., & Dunham, M. J. (2008). The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genetics*, *4*(12), e1000303. https://doi.org/10.1371/journal.pgen.1000303

Gresham, D., Usaite, R., Germann, S. M., Lisby, M., Botstein, D., & Regenberg, B. (2010). Adaptation to diverse nitrogen-limited environments by deletion or extrachromosomal element formation of the GAP1 locus. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(43), 18551–18556. https://doi.org/10.1073/pnas.1014023107

Gu, W., Zhang, F., & Lupski, J. R. (2008a). Mechanisms for human genomic rearrangements. *PathoGenetics*, *1*(1), 4. https://doi.org/10.1186/1755-8417-1-4

Gu, W., Zhang, F., & Lupski, J. R. (2008b). Mechanisms for human genomic rearrangements. *PathoGenetics*, *1*(1), 4. https://doi.org/10.1186/1755-8417-1-4

Hall, D. W., Mahmoudizad, R., Hurd, A. W., & Joseph, S. B. (2008). Spontaneous mutations in diploid Saccharomyces cerevisiae: Another thousand cell generations. *Genetics Research*, *90*(3), 229–241. https://doi.org/10.1017/S0016672308009324

Harel, T., Pehlivan, D., Caskey, C. T., & Lupski, J. R. (2015). Chapter 1—Mendelian, Non-Mendelian, Multigenic Inheritance, and Epigenetics. In R. N. Rosenberg & J. M. Pascual (Eds.), *Rosenberg's Molecular and Genetic Basis of Neurological and Psychiatric Disease (Fifth Edition)* (pp. 3–27). Academic Press. https://doi.org/10.1016/B978-0-12-410529-4.00001-2

Hastings, P. J., Ira, G., & Lupski, J. R. (2009). A Microhomology-Mediated Break-Induced Replication Model for the Origin of Human Copy Number Variation. *PLOS Genetics*, *5*(1), e1000327. https://doi.org/10.1371/journal.pgen.1000327

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., & Ira, G. (2009). Mechanisms of change in gene copy number. *Nature Reviews Genetics*, *10*(8), 551–564. https://doi.org/10.1038/nrg2593

Hays, M., Schwartz, K., Schmidtke, D. T., Aggeli, D., & Sherlock, G. (2023). Paths to adaptation under fluctuating nitrogen starvation: The spectrum of adaptive mutations in Saccharomyces cerevisiae is shaped by retrotransposons and microhomology-mediated recombination. *PLOS Genetics*, *19*(5), e1010747. https://doi.org/10.1371/journal.pgen.1010747

Hermans, J., Delaunoy, A., Rozet, F., Wehenkel, A., Begy, V., & Louppe, G. (2022). *A Trust Crisis In Simulation-Based Inference? Your Posterior Approximations Can Be Unfaithful* (arXiv:2110.06581). arXiv. https://doi.org/10.48550/arXiv.2110.06581

Horiuchi, T., Horiuchi, S., & Novick, A. (1963). The genetic basis of hyper-synthesis of beta-galactosidase. *Genetics*, *48*, 157–169. https://doi.org/10.1093/genetics/48.2.157

Hull, R. M., Cruz, C., Jack, C. V., & Houseley, J. (2017). Environmental change drives accelerated

adaptation through stimulated copy number variation. *PLOS Biology*, *15*(6), e2001333. https://doi.org/10.1371/journal.pbio.2001333

Ji, H., Moore, D. P., Blomberg, M. A., Braiterman, L. T., Voytas, D. F., Natsoulis, G., & Boeke, J. D. (1993). Hotspots for unselected Ty1 transposition events on yeast chromosome III are near tRNA genes and LTR sequences. *Cell*, *73*(5), 1007–1018. https://doi.org/10.1016/0092-8674(93)90278-x

Joseph, S. B., & Hall, D. W. (2004). Spontaneous Mutations in Diploid Saccharomyces cerevisiae: More Beneficial Than Expected. *Genetics*, *168*(4), 1817–1825. https://doi.org/10.1534/genetics.104.033761

Jost, L. (2006). Entropy and diversity. *Oikos*, *113*(2), 363–375. https://doi.org/10.1111/j.2006.0030-1299.14714.x

Kohanovski, I., Pontz, M., Vande Zande, P., Selmecki, A., Dahan, O., Pilpel, Y., Yona, A. H., & Ram, Y. (2024). Aneuploidy can be an evolutionary diversion on the path to adaptation. *Molecular Biology and Evolution*, msae052. https://doi.org/10.1093/molbev/msae052

Lang, G. I., Botstein, D., & Desai, M. M. (2011). Genetic variation and the fate of beneficial mutations in asexual populations. *Genetics*, *188*(3), 647–661. https://doi.org/10.1534/genetics.111.128942

Lang, G. I., & Murray, A. W. (2011). Mutation Rates across Budding Yeast Chromosome VI Are Correlated with Replication Timing. *Genome Biology and Evolution*, *3*, 799–811. https://doi.org/10.1093/gbe/evr054

Lauer, S., Avecilla, G., Spealman, P., Sethia, G., Brandt, N., Levy, S. F., & Gresham, D. (2018). Single-cell copy number variant detection reveals the dynamics and diversity of adaptation. *PLOS Biology*, *16*(12), e3000069. https://doi.org/10.1371/journal.pbio.3000069

Lauer, S., & Gresham, D. (2019). An evolving view of copy number variants. *Current Genetics*, *65*(6). https://doi.org/10.1007/s00294-019-00980-0

Lee, J. A., Carvalho, C. M. B., & Lupski, J. R. (2007). A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell*, *131*(7), 1235–1247. https://doi.org/10.1016/j.cell.2007.11.037

Lercher, M. J., & Hurst, L. D. (2002). Human SNP variability and mutation rate are higher in regions of high recombination. *Trends in Genetics*, *18*(7), 337–340. https://doi.org/10.1016/S0168-9525(02)02669-0

Levy, S. F., Blundell, J. R., Venkataram, S., Petrov, D. A., Fisher, D. S., & Sherlock, G. (2015). Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature*, *519*(7542), 181–186. https://doi.org/10.1038/nature14279

Lukow, D. A., Sausville, E. L., Suri, P., Chunduri, N. K., Wieland, A., Leu, J., Smith, J. C., Girish, V., Kumar, A. A., Kendall, J., Wang, Z., Storchova, Z., & Sheltzer, J. M. (2021). Chromosomal instability accelerates the evolution of resistance to anti-cancer therapies. *Developmental Cell*, *56*(17), 2427-2439.e4. https://doi.org/10.1016/j.devcel.2021.07.009

Lupski, J. R. (1998). Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in Genetics*, *14*(10), 417–422. https://doi.org/10.1016/S0168-9525(98)01555-8

Lupski, J. R., & Stankiewicz, P. (2005). Genomic Disorders: Molecular Mechanisms for

Rearrangements and Conveyed Phenotypes. *PLoS Genetics*, *1*(6), e49. https://doi.org/10.1371/journal.pgen.0010049

Malhotra, D., & Sebat, J. (2012). CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics. *Cell*, *148*(6), 1223–1241. https://doi.org/10.1016/j.cell.2012.02.039

Martin, R., Espinoza, C. Y., Large, C. R. L., Rosswork, J., Bruinisse, C. V., Miller, A. W., Sanchez, J. C., Miller, M., Paskvan, S., Alvino, G. M., Dunham, M. J., Raghuraman, M. K., & Brewer, B. J. (2024). Template switching between the leading and lagging strands at replication forks generates inverted copy number variants through hairpin-capped extrachromosomal DNA. *PLOS Genetics*, *20*(1), e1010850. https://doi.org/10.1371/journal.pgen.1010850

Matassi, G., Sharp, P. M., & Gautier, C. (1999). Chromosomal location effects on gene sequence evolution in mammals. *Current Biology*, *9*(15), 786–791. https://doi.org/10.1016/S0960-9822(99)80361-3

Morillon, A., Bénard, L., Springer, M., & Lesage, P. (2002). Differential Effects of Chromatin and Gcn4 on the 50-Fold Range of Expression among Individual Yeast Ty1 Retrotransposons. *Molecular and Cellular Biology*, *22*(7), 2078–2088. https://doi.org/10.1128/MCB.22.7.2078-2088.2002

Morillon, A., Springer, M., & Lesage, P. (2000). Activation of the Kss1 Invasive-Filamentous Growth Pathway Induces Ty1 Transcription and Retrotransposition in Saccharomyces cerevisiae. *Molecular and Cellular Biology*, *20*(15), 5766–5776. https://doi.org/10.1128/MCB.20.15.5766-5776.2000

Mularoni, L., Zhou, Y., Bowen, T., Gangadharan, S., Wheelan, S. J., & Boeke, J. D. (2012). Retrotransposon Ty1 integration targets specifically positioned asymmetric nucleosomal DNA segments in tRNA hotspots. *Genome Research*, *22*(4), 693–703. https://doi.org/10.1101/gr.129460.111

Nguyen Ba, A. N., Cvijović, I., Rojas Echenique, J. I., Lawrence, K. R., Rego-Costa, A., Liu, X., Levy, S. F., & Desai, M. M. (2019). High-resolution lineage tracking reveals travelling wave of adaptation in laboratory yeast. *Nature*, *575*(7783), 494–499. https://doi.org/10.1038/s41586-019-1749-3

Nishant, K. T., Singh, N. D., & Alani, E. (2009). Genomic mutation rates: What high-throughput methods can tell us. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *31*(9), 912–920. https://doi.org/10.1002/bies.200900017

Ohno, S. (1970). *Evolution by Gene Duplication*. New York: Springer-Verlag.

Osmundson, J. S., Kumar, J., Yeung, R., & Smith, D. J. (2017). Pif1-family helicases cooperate to suppress widespread replication fork arrest at tRNA genes. *Nature Structural & Molecular Biology*, *24*(2), 162–170. https://doi.org/10.1038/nsmb.3342

Papamakarios, G. (2019). *Neural Density Estimation and Likelihood-free Inference* (arXiv:1910.13233). arXiv. https://doi.org/10.48550/arXiv.1910.13233

Papamakarios, G., Pavlakou, T., & Murray, I. (2018). *Masked Autoregressive Flow for Density Estimation* (arXiv:1705.07057). arXiv. https://doi.org/10.48550/arXiv.1705.07057

Pavelka, N., Rancati, G., Zhu, J., Bradford, W. D., Saraf, A., Florens, L., Sanderson, B. W., Hattem, G. L., & Li, R. (2010). Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. *Nature*, *468*(7321), 321–325. https://doi.org/10.1038/nature09529

Payen, C., Sunshine, A. B., Ong, G. T., Pogachar, J. L., Zhao, W., & Dunham, M. J. (2016).

High-Throughput Identification of Adaptive Mutations in Experimentally Evolved Yeast Populations. *PLoS Genetics*, *12*(10), e1006339. https://doi.org/10.1371/journal.pgen.1006339

Pentao, L., Wise, C. A., Chinault, A. C., Patel, P. I., & Lupski, J. R. (1992). Charcot-Marie-Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1.5 Mb monomer unit. *Nature Genetics*, *2*(4), 292–300. https://doi.org/10.1038/ng1292-292

Pös, O., Radvanszky, J., Buglyó, G., Pös, Z., Rusnakova, D., Nagy, B., & Szemes, T. (2021). DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. *Biomedical Journal*, *44*(5), 548–559. https://doi.org/10.1016/j.bj.2021.02.003

Robinson, D., Vanacloig-Pedros, E., Cai, R., Place, M., Hose, J., & Gasch, A. P. (2023). Gene-by-environment interactions influence the fitness cost of gene copy-number variation in yeast. *G3 Genes|Genomes|Genetics*, *13*(10), jkad159. https://doi.org/10.1093/g3journal/jkad159

Rutledge, S. D., Douglas, T. A., Nicholson, J. M., Vila-Casadesús, M., Kantzler, C. L., Wangsa, D., Barroso-Vilares, M., Kale, S. D., Logarinho, E., & Cimini, D. (2016). Selective advantage of trisomic human cells cultured in non-standard conditions. *Scientific Reports*, *6*, 22828. https://doi.org/10.1038/srep22828

Salim, D., Bradford, W. D., Rubinstein, B., & Gerton, J. L. (2021). DNA replication, transcription, and H3K56 acetylation regulate copy number and stability at tandem repeats. *G3 (Bethesda, Md.)*, *11*(6), jkab082. https://doi.org/10.1093/g3journal/jkab082

Selmecki, A. M., Maruvka, Y. E., Richmond, P. A., Guillet, M., Shoresh, N., Sorenson, A. L., De, S., Kishony, R., Michor, F., Dowell, R., & Pellman, D. (2015). Polyploidy can drive rapid adaptation in yeast. *Nature*, *519*(7543), 349–352. https://doi.org/10.1038/nature14187

Sonti, R. V., & Roth, J. R. (1989). Role of gene duplications in the adaptation of Salmonella typhimurium to growth on limiting carbon sources. *Genetics*, *123*(1), 19–28. https://doi.org/10.1093/genetics/123.1.19

Spealman, P. (2019). *CVish Structural Variant Breakpoint Identifier*. Github. https://github.com/pspealman/CVish

Spealman, P., Avecilla, G., Matthews, J., Suresh, I., & Gresham, D. (2022). Complex Genomic Rearrangements following Selection in a Glutamine-Limited Medium over Hundreds of Generations. *Microbiology Resource Announcements*, *11*(11), e00729-22. https://doi.org/10.1128/mra.00729-22

Spealman, P., De, T., Chuong, J. N., & Gresham, D. (2023). Best Practices in Microbial Experimental Evolution: Using Reporters and Long-Read Sequencing to Identify Copy Number Variation in Experimental Evolution. *Journal of Molecular Evolution*, *91*(3), 356–368. https://doi.org/10.1007/s00239-023-10102-7

Stankiewicz, P., Shaw, C. J., Dapper, J. D., Wakui, K., Shaffer, L. G., Withers, M., Elizondo, L., Park, S.-S., & Lupski, J. R. (2003). Genome architecture catalyzes nonrecurrent chromosomal rearrangements. *American Journal of Human Genetics*, *72*(5), 1101–1116. https://doi.org/10.1086/374385

Storz, J. F. (2016). Gene Duplication and Evolutionary Innovations in Hemoglobin-Oxygen Transport. *Physiology*, *31*(3), 223–232. https://doi.org/10.1152/physiol.00060.2015

Taylor, J. S., & Raes, J. (2004). Duplication and divergence: The evolution of new genes and old

ideas. *Annual Review of Genetics*, *38*, 615–643. https://doi.org/10.1146/annurev.genet.38.072902.092831

Tejero-Cantero, A., Boelts, J., Deistler, M., Lueckmann, J.-M., Durkan, C., Gonçalves, P. J., Greenberg, D. S., & Macke, J. H. (2020). *SBI -- A toolkit for simulation-based inference* (arXiv:2007.09114). arXiv. https://doi.org/10.48550/arXiv.2007.09114

Tsai, H.-J., Nelliat, A. R., Choudhury, M. I., Kucharavy, A., Bradford, W. D., Cook, M. E., Kim, J., Mair, D. B., Sun, S. X., Schatz, M. C., & Li, R. (2019). Hypo-osmotic-like stress underlies general cellular defects of aneuploidy. *Nature*, *570*(7759), 117–121. https://doi.org/10.1038/s41586-019-1187-2

Turner, D. J., Miretti, M., Rajan, D., Fiegler, H., Carter, N. P., Blayney, M. L., Beck, S., & Hurles, M. E. (2008). Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nature Genetics*, *40*(1), 90–95. https://doi.org/10.1038/ng.2007.40

Venkataram, S., Dunn, B., Li, Y., Agarwala, A., Chang, J., Ebel, E. R., Geiler-Samerotte, K., Hérissant, L., Blundell, J. R., Levy, S. F., Fisher, D. S., Sherlock, G., & Petrov, D. A. (2016). Development of a Comprehensive Genotype-to-Fitness Map of Adaptation-Driving Mutations in Yeast. *Cell*, *166*(6), 1585-1596.e22. https://doi.org/10.1016/j.cell.2016.08.002

Whale, A. J., King, M., Hull, R. M., Krueger, F., & Houseley, J. (2022). Stimulation of adaptive gene amplification by origin firing under replication fork constraint. *Nucleic Acids Research*, *50*(2), 915–936. https://doi.org/10.1093/nar/gkab1257

Wilke, C. M., & Adams, J. (1992). Fitness effects of Ty transposition in Saccharomyces cerevisiae. *Genetics*, *131*(1), 31–42. https://doi.org/10.1093/genetics/131.1.31

Wilson, T. E., Arlt, M. F., Park, S. H., Rajendran, S., Paulsen, M., Ljungman, M., & Glover, T. W. (2015). Large transcription units unify copy number variants and common fragile sites arising under replication stress. *Genome Research*, *25*(2), 189–200. https://doi.org/10.1101/gr.177121.114

Wolfe, K. H., Sharp, P. M., & Li, W.-H. (1989). Mutation rates differ among regions of the mammalian genome. *Nature*, *337*(6204), 283–285. https://doi.org/10.1038/337283a0

Yang, F., Todd, R. T., Selmecki, A., Jiang, Y., Cao, Y., & Berman, J. (2021). The fitness costs and benefits of trisomy of each Candida albicans chromosome. *Genetics*, *218*(2), iyab056. https://doi.org/10.1093/genetics/iyab056

Yeung, R., & Smith, D. J. (2020). Determinants of Replication-Fork Pausing at tRNA Genes in Saccharomyces cerevisiae. *Genetics*, *214*(4), 825–838. https://doi.org/10.1534/genetics.120.303092

Yona, A. H., Manor, Y. S., Herbst, R. H., Romano, G. H., Mitchell, A., Kupiec, M., Pilpel, Y., & Dahan, O. (2012). Chromosomal duplication is a transient evolutionary solution to stress. *Proceedings of the National Academy of Sciences*, *109*(51), 21010–21015. https://doi.org/10.1073/pnas.1211150109

Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy Number Variation in Human Health, Disease, and Evolution. *Annual Review of Genomics and Human Genetics*, *10*, 451–481. https://doi.org/10.1146/annurev.genom.9.081307.164217

Zhang, F., Khajavi, M., Connolly, A. M., Towne, C. F., Batish, S. D., & Lupski, J. R. (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex

rearrangements in humans. *Nature Genetics*, *41*(7), 849–853. https://doi.org/10.1038/ng.399
Zuellig, M. P., & Sweigart, A. L. (2018). Gene duplicates cause hybrid lethality between sympatric species of Mimulus. *PLoS Genetics*, *14*(4), e1007130. https://doi.org/10.1371/journal.pgen.1007130