Inferring the Selective History of CNVs Using a Maximum Likelihood Model

Seyed Amir Malekpour [1,*], Ata Kalirad [2], Sina Majidian [3,4,*]

[1]School of Biological Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran 19395-5746, Iran
[2]Department for Integrative Evolutionary Biology, Max Planck Institute for Biology Tübingen, Tübingen 72076, Germany
[3]SIB Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland
[4]Department of Computational Biology, University of Lausanne, Lausanne 1015, Switzerland

*Corresponding authors: E-mails: a.malekpour@ipm.ir; sina.majidian@unil.ch.

Accepted: March 13, 2025

Abstract

Copy number variations (CNVs)—structural variations generated by deletion and/or duplication that result in a change in DNA dosage—are prevalent in nature. CNVs can drastically affect the phenotype of an organism and have been shown to be both involved in genetic disorders and be used as raw material in adaptive evolution. Unlike single-nucleotide variations, the often large and varied effects of CNVs on phenotype hinders our ability to infer their selective advantage based on the population genetics data. Here, we present a likelihood-based approach, dubbed PoMoCNV (POlymorphism-aware phylogenetic MOdel for CNVs), that estimates the evolutionary parameters such as mutation rates among different copy numbers and relative fitness loss per copy deletion at a genomic locus based on population genetics data. As a case study, we analyze the genomics data of 40 strains of *Caenorhabditis elegans*, representing four different populations. We take advantage of the data on chromatin accessibility to interpret the mutation rate and fitness of copy numbers, as inferred by PoMoCNV, specifically in open or closed chromatin loci. We further test the reliability of PoMoCNV by estimating the evolutionary parameters of CNVs for mutation-accumulation experiments in *C. elegans* with varying levels of genetic drift.

Key words: copy number variations, likelihood-based inference, *C. elegans*, polymorphism-aware phylogenetic model.

Significance

Inferring the mutation rate and fitness of copy numbers based on population genetics data is crucial to understand their role in evolution. However, given the diversity in the size and effects of CNVs, such inference poses a challenge. We developed a likelihood-based approach called PoMoCNV to address this issue.

Introduction

Structural variations, i.e. genomic changes spanning more than 1 kilo bases (kb) have been widely recognized as a major source of genetic variability within and between species (Freeman et al. 2006; Conrad and Hurles 2007). Copy number variations (CNVs) are a subset of structural variations that are specifically generated by duplication and deletion (Malekpour et al. 2018). Given the considerable length of CNVs, their effects on the phenotype, compared to single-nucleotide variations (SNVs), could be substantial. For example, recent studies have shown that deletions are more deleterious than single nucleotide variants (Hämälä et al. 2021; Aqil et al. 2023). In fact, the ever-growing body of literature on CNVs in humans has delivered a hefty catalog of

such variations in the human genome (Feuk et al. 2006; Zarrei et al. 2015) and the myriad of diseases linked to this group of structural changes (Dumas and Sikela 2009; Zhang et al. 2009; Conrad et al. 2010; Tang and Amon 2013). Aside from such detrimental effects, the co-option of gene duplicates has long been regarded as a short path to the emergence of novel functions (Ohno 1970; Conant and Wolfe 2008; Ponting 2008). In that respect, CNVs have been shown to provide just such raw material for adaptive evolution in a diverse array of organisms. For example, the duplication of hexose transporter in *Saccharomyces cerevisiae* confers adaptive benefit (i.e. higher fitness) in stressful conditions (reviewed in Kondrashov 2012), and in humans, copy number of the salivary amylase gene seems to have increased in population accustomed to high-starch diets (Perry et al. 2007; Yılmaz et al. 2023; Bolognini et al. 2024). Even in plants, CNVs appear to provide crucial genetic raw material for selection during domestication (Lye and Purugganan 2019). Furthermore, cognitive advancements that emerged during the transition between the genus *Australopithecus* to *Homo*—usually attributed to the expansion of neocortex—can be traced back to the repeated duplication of *SRGAP2* during the hominin evolution (Charrier et al. 2012; Dennis et al. 2012; Gómez-Robles et al. 2024).

In spite of their prevalence in nature, understanding the evolutionary dynamics of copy numbers (e.g. rate of changes and relative fitness loss) has proved a Herculean task. Whereas an SNV can be linked to a given biological function, a CNV, given its length, would more likely have detrimental functional impact. Deletion or duplication of a long genomic region will decrease or amplify the number of copies of genes and/or regulatory elements located on the affected region. Since lengthy regions cover more genes, such alterations could involve genes participating in various molecular processes and diseases (Auwerx et al. 2024a, 2024b; Fang and Edwards 2024). For instance, 16p11.2 BP4-5 CNVs in humans are associated with changes in more than forty traits, including anemia, sleep apnea, schizophrenia, and puberty timing (Auwerx et al. 2024a, 2024b). Consequently, the available data on SNVs can be used to draw evolutionary conclusions. For example, it is relatively simple to link SNVs to changes in protein structure across phyla to understand compensatory evolution (Kondrashov et al. 2002; Ivankov et al. 2014), whereas making similar evolutionary inferences with respect to CNVs is not straightforward. However, cases of CNVs involved in compensatory evolution have been reported. For instance, Szamecz et al. (2014) postulated that a deletion of *RPL6B*, which produces a component of 60S subunit, during an experiential evolution of *Saccharomyces cerevisiae* was compensated by an increase in the copy number of *RPLA6A*.

One widely used approach to understanding the fitness effects of SNVs has been to use their current frequencies in a population or populations, combined with *a priori* information on their demographic histories, i.e. population growth or recent bottlenecks, to infer the patterns of selection in the past. This approach has been applied to the polymorphism data in humans to infer the evolutionary consequences of amino acid substitutions (Williamson et al. 2005; Boyko et al. 2008). De Maio et al. (2013) combined the available genetic polymorphism data on humans and their great ape relatives, chimpanzees, and orangutans to infer their evolutionary history, i.e. fixation rate and mutation rate of a single base. Here, we extended their analysis of single base mutation to copy number mutation. The latter refers to the changes in the number of copies of a genomics locus. We examine this by calculating the rate of changes as a parameter called copy number mutation rate. We also considered a fitness loss ($\mathbb{S}$) parameter that quantifies the relative disadvantage caused by each copy deletion at a genomic locus, in comparison to a locus with two copies having a fitness of $S_2 = 1$, see Table 1 and Equation (4). In this context, the fitness of genomic loci with homozygous and heterozygous deletions is considered as $S_0 = 1 - 2\,\mathbb{S}$ and $S_1 = 1 - \mathbb{S}$, respectively. Additionally, the fitness of genomic loci with duplicated copies (>2 copies for diploid organisms) is given by $S_3 = 1 - 0.5\,\mathbb{S}$. This formulation of fitness values is inspired by a recent study indicating that deletions appear to be approximately four times as deleterious as duplications (Hujoel et al. 2022). In the above equations, the fitness of a locus with *i* copies in the genome, denoted as S_i, is a function of the parameter $\mathbb{S}$. This simplification reduces the number of parameters in the model, leading to more reliable estimations from the available data. In this contribution, we propose a novel POlymorphism-aware phylogenetic MOdel (PoMo) for CNV datasets, dubbed PoMoCNV, which infers mutation rates of copy numbers and fitness associated with each copy number based on genomic data (see Materials and Methods for details).

As a case study, we applied PoMoCNV to understand the evolution of CNVs in *Caenorhabditis elegans*, as a diploid organism. Nevertheless, it is worth noting that the PoMoCNV framework is applicable to other organisms, whether they are haploid or polyploid. The sheer number of duplicated genes in *C. elegans* (∼ 8,971) dwarfs those of *Saccharomyces cerevisiae* (∼ 1,858) or *Drosophila melanogaster* (∼ 5,536) (Rubin et al. 2000), with an estimated rate of duplication of ∼ 0.02 per gene per million years, compared to ∼ 0.002 in *D. melanogaster* (Lynch and Conery 2000). Given such a high rate of gene duplication, *C. elegans* has been used to study the evolutionary dynamics of CNVs during laboratory experimental evolution. For example, (Konrad et al. 2018) took advantage of the short generation time of *C. elegans*—roughly 3 days (Wood 1988)—and followed the emergence and loss of duplicates in a mutation accumulation (MA) experiment

Table 1. Description of parameters and symbols.

Parameter	Description
Copy numbers	There are four possible copy numbers (0, 1, 2, and 3) for a genomic locus. 0: for homozygous deletions, 1: for heterozygous deletions, 2: for diploid or normal copies, 3: for duplicated locus with more than 2 copies.
N	Number of individuals per population
State	For each genomic locus, a state is uniquely defined by the types of copy numbers and their frequencies within a population of $N = 10$ individuals [a].
$\begin{pmatrix} i & I \\ j & J \end{pmatrix}$	For each genomic locus, this represents a state within a population of N individuals, where I and J are alleles with respective frequencies i and j. For integer numbers $i > 0$ and $j > 0$, satisfying the condition $i + j = N$, this configuration indicates a bi-allelic state.
α_{ij}	Mutation (change) rate in the copy number of a genomic region from i copies to j copies within an individual per generation.
$M^{N-}_{I,J}$	The transition rate (probability) of moving from a mono-allelic state with N individuals, all having a copy number I in a genomic locus, to a bi-allelic state where one individual with copy number I is replaced by a new individual with copy number J [b].
$M^{i-}_{I,J}$	The transition rate (probability) of moving from a bi-allelic state with i and $j = N - i$ individuals having copy numbers I and J, respectively, in a genomic locus, to another state where one individual with copy number I is replaced by a new individual with copy number J, in one generation [c].
$\mathbb{S}$	This represents the relative fitness loss for each copy deletion of a locus. Given a fitness value of $S_2 = 1$ for a locus with two copies, the fitness values for loci with heterozygous or homozygous deletions are $S_1 = 1 - \mathbb{S}$ and $S_0 = 1 - 2\mathbb{S}$, respectively.
S_i	This represents the fitness of a locus with i copies in the genome, for $i = 0, 1, 2, 3$.

Note: [a] For example, it can be a mono-allelic state where all $N = 10$ individuals have copy number 2 or a bi-allelic state where 8 individuals have copy number 2 and the others have 3 copies in the locus, such bi-allelic state is indicated by $\begin{pmatrix} 8 & I=2 \\ 2 & J=3 \end{pmatrix}$. [b] This is a transition in the PoMoCNV's state governed by the mutation rate. [c] This is a transition in the PoMoCNV's state governed by copy number frequencies and their fitness values.

for ∼ 400 generations. To vary the relative strength of selection to genetic drift, Konrad *et al.* used three different bottleneck sizes, $B = 1$, $B = 10$, and $B = 100$, to propagate different lines in the MA experiment. The bottleneck size can influence the efficacy of selection by affecting the genetic diversity and the ability to remove deleterious mutations from the population. Thus, changes in bottleneck size can impact the strength of selection observed in these MA experiments. The dataset generated on the dynamics of the CNVs during this MA experiment, provides a perfect testing bed for *PoMoCNV*, since no speculation concerning the demographic history is required given the known the timing and the intensity of the bottlenecks, which consequently determines the strength of selection in each lineage. In addition, the available in-depth knowledge concerning the genome regulation in *C. elegans*, specifically chromatin accessibility (Valouev et al. 2008; Gerstein et al. 2010; Evans et al. 2016; Daugherty et al. 2017), can be used to provide a more accurate inference of the evolutionary dynamics of CNVs, in open or closed chromatin loci.

Results

PoMoCNV is employed to estimate the evolutionary parameters for three different real *C. elegans* datasets and one synthetic dataset mimicking bottlenecking processes in the population. The evolutionary parameters include rate of changes in (mutation of) copy numbers, and relative fitness loss ($\mathbb{S}$) of every copy deletion at a genomic locus.

i. The first dataset includes CNVs of four *C. elegans* populations (Africa, Australia, France, and Hawaii), each consisting of 10 strains. Each strain represents a specific *C. elegans* individual with unique genetic characteristics (see Lee et al. 2021 for further details).
ii. We consider CNVs of *C. elegans* populations and utilized ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) data (Jänes et al. 2018) to categorize chromatin into open or closed segments. Evolutionary rates are then estimated separately for the open and closed chromatin segments.
iii. We also benefited from three MA experiments with bottleneck sizes $B = 1$, $B = 10$, or $B = 100$, in *C. elegans* populations (Konrad et al. 2018). These MA experiments allow for copy gain and loss events under varying selection intensities in *C. elegans*.
iv. To further evaluate the reliability of parameter estimation, including mutation rates and relative fitness loss ($\mathbb{S}$) per copy deletion, we generated a synthetic dataset involving populations undergoing bottleneck processes in each generation. Compared to the previous MA experiment in *C. elegans*, this simulated dataset contains more CNV calls, specifically enhancing the reliability of mutation rate estimations. See the Results section for further details.

CNV Evolution in *C. elegans*

The evolution of copy numbers is studied using DNA sequencing data. We benefited from publicly available short read DNA sequencing data of *C. elegans* (Lee et al. 2021). The first step is to align reads to the reference genome (upper panel of Fig. 1). We aligned the short read DNA sequences from four *C. elegans* populations (Africa, Australia, France, and Hawaii), each consisting of 10 strains, to the N2 reference genome (WS270). CNVs are then inferred with CNVnator (Abyzov et al. 2011). After

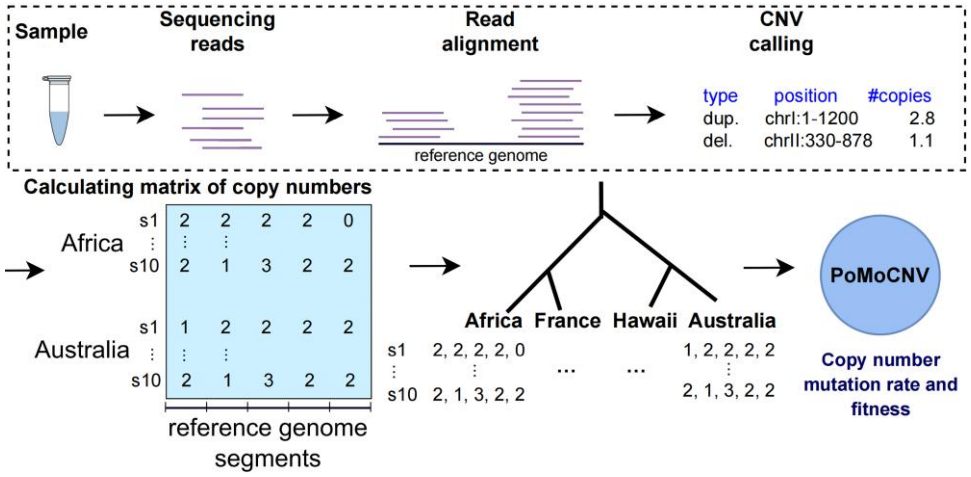

Fig. 1. PoMoCNV is employed to estimate fitness parameters and mutation rates between different copy numbers in *C. elegans*. Short reads from 10 strains of *C. elegans* for each population (Africa, France, Hawaii, and Australia) are aligned to the N2 reference genome, and used to infer copy numbers for 1kb genomic segments. Utilizing the phylogenetic tree of populations and estimated copy numbers, PoMoCNV infers the parameters governing CNV evolution along branches.

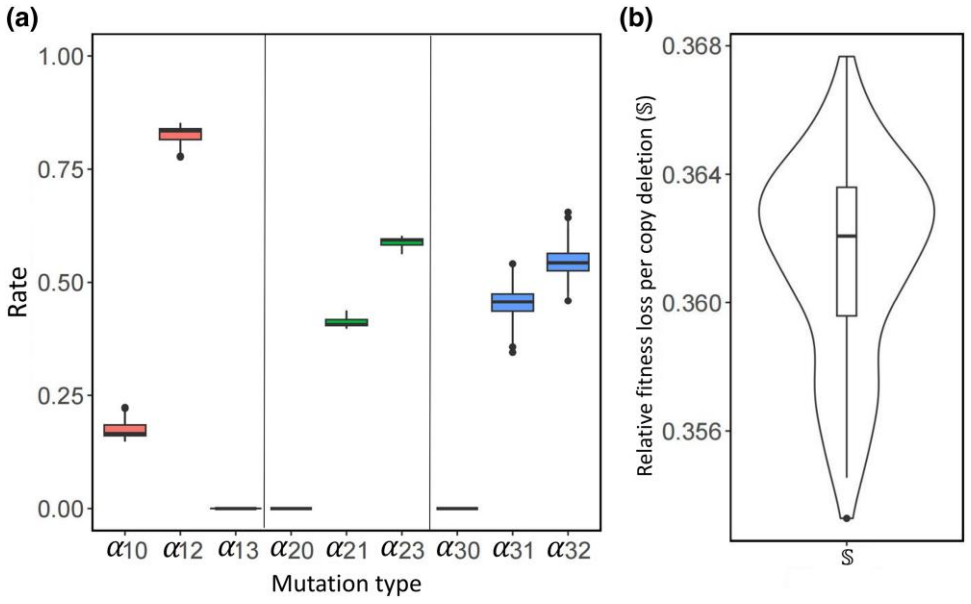

Fig. 2. The relative mutation rates and relative fitness loss per copy deletion (𝕊) at a genomic locus, in *C. elegans*. a) PoMoCNV calculates relative mutation rates, which represent the frequency of copy number changes within a genomic segment per generation. b) PoMoCNV also estimates the relative fitness loss that occurs with each copy number deletion in a genomic segment.

dividing the genome into nonoverlapping segments with a length of 1kb, copy numbers, including homozygous and heterozygous deletions (zero and one copy), normal (2 copies), and duplications (more than 2 copies), are assigned to each segment using CNV calls from CNVnator. Supplementary figure S1, Supplementary Material online displays copy number distribution for one representative strain from each population. While an increase in copy numbers of a gene could provide fitness benefits and facilitate adaptation (e.g. increase in the number of cellulase gene in a nematode Han et al. 2022), the prevalence of genes with 2 copies in *C. elegans* suggests that, at least on average, copy number variations are under stabilizing selection (Lenormand et al. 2020).

Our proposed method, PoMoCNV, infers the mutation rates of copy numbers and relative fitness loss (𝕊) per

copy deletion at a genomic locus along branches in the phylogenetic tree. The relative fitness loss ($\mathbb{S}$) quantifies the relative disadvantage caused by each copy deletion at a genomic locus, in comparison to a locus with two copies having a fitness of $S_2 = 1$, see Table 1 and Equation (4). With 4 populations of 10 *C. elegans* strains at the leaves of the tree, the copy number evolution of genomic segments is modeled forward in time, starting from the ancestral population (root node) and proceeding to the present-day populations with a Moran birth–death model (Moran 1962; Nowak et al. 2004). In each generation forward in time, one strain is selected to reproduce and transmit its allele (copy number) to an offspring, while another strain is selected, and its allele is removed from the population. In PoMoCNV, the likelihood of this birth–death process is modeled per genomic segment, taking into account the allele fitness and frequencies. For more computational details see Materials and Methods section.

We assessed the robustness of the estimated parameters by utilizing 100 bootstrap samples from genomic segments. Figure 2a presents boxplots depicting the estimated mutation rates between copy numbers across 100 bootstrap samples. As Fig. 2a illustrates, mutations from each copy number to other copies—immediately higher or lower than the original copy—exhibit the highest rates. For instance, mutations from copy number 1 predominantly occur to copy number 0 or 2, from copy number 2 to 1 or 3, and from copy number 3 to 2 or 1. Furthermore, mutations leading to higher copy numbers exhibit higher rates, such as $\alpha_{12} > \alpha_{10}$, $\alpha_{23} > \alpha_{21}$, and $\alpha_{32} > \alpha_{31}$. This preference for copy gains is likely driven by the potential benefits associated with increased gene dosage, while copy losses are generally more detrimental to the organism (Sung et al. 2016; Chunduri et al. 2022; Gonzalez et al. 2019). This observation is in line with the estimates of gene duplication rate across the Tree of Life based on comparative analyses (Katju and Bergthorsson 2013). Additionally, direct measurements in the MA experiment in *C. elegans* under varying selective regimes indicate a gene duplication rate of 2.9×10^{-5} per gene per generation, and a deletion rate of 5×10^{-6} per gene per generation (Konrad et al. 2018). The frequency of duplication is expected to be higher than deletion since the latter is assumed to be under stronger purifying selection in exonic and intronic regions, probably since a deletion can disrupt the gene itself or its splicing, resulting in loss of function (Conrad et al. 2006; Redon et al. 2006). Figure 2b illustrates the distribution of the relative fitness loss ($\mathbb{S}$) per copy deletion (calculated using 100 bootstrap samples of genomic segments). The plot reveals that the first quartile (Q1) and third quartile (Q3) of the relative fitness loss fall within the interval of (0.36, 0.364). The tight clustering of the data points within this interval indicates a low variability in the relative fitness loss, reinforcing the robustness of the observed trend.

CNV Evolution in Open and Closed Chromatin Segments

Understanding the CNV evolution in open (accessible) and closed (less accessible) chromatin segments is important in genomics research. The chromatin state of a gene regulatory segment, whether it is open or closed, plays a crucial role in regulating gene expression (Miyamoto et al. 2018; Yoshida et al. 2019; Wong et al. 2023) and determining the functional consequences of CNVs. Then, investigating the CNV evolution specifically in open and closed chromatin segments, (i) furthers our insights into the dynamics of genomic changes, which involves factors such as mutation between different copy numbers and relative fitness loss for each copy deletion at a genomic locus, and their impact on gene regulation, (ii) sheds light on the interplay between genetic and epigenetic factors (Shi et al. 2020), (iii) paves the way for elucidating the underlying mechanisms driving CNV formation and selection, providing valuable information for understanding the genetic basis of diseases and the evolutionary processes shaping genomic diversity.

For this purpose, we utilized ATAC-seq data to identify the open and closed chromatin regions (Daugherty et al. 2017; Jänes et al. 2018; Thibodeau et al. 2021). The ATAC-seq peak data from *C. elegans* (Jänes et al. 2018), includes 42,245 regions of high chromatin accessibility. Collectively, these peaks cover approximately 6.4% of the *C. elegans* (N2) genome. In order to investigate the evolution of copy numbers in open and closed regions, the N2 reference genome, comprising approximately 100.3 million bps, is partitioned into genomic loci with a length of 50 bps. In African strains, the average copy numbers for open and closed chromatin loci were 1.99 and 1.93, respectively, and the difference between them is statistically significant (Wilcoxon rank-sum test, *P*-value <2.2e – 16). Similarly, the copy number in the other three populations of France, Australia, and Hawaii was significantly higher in the open chromatin loci than in the closed loci. Figure 3 illustrates how copy numbers are distributed among chromatin loci in African strains. This indicates that open chromatin loci harbor a lower proportion of deletions (both homozygous and heterozygous) but a higher proportion of duplications compared to closed chromatin loci. Moreover, multiallelic variants are more likely to be present in closed chromatin loci, whereas mono-allelic variants are more prevalent in open chromatin loci. In open chromatin loci, the percentages of mono-, bi-, and three-allelic variants are 42.4%, 52.4%, and 5.1%, respectively, while in closed chromatin loci, the percentages are 39.9%, 52.3%, and 7.8%. As expected, the majority of mono-allelic variants in open or closed loci have a copy number of 2. The percentages of the three-allelic variants in open and closed chromatin loci (5.1%, 7.8%) were significantly different (Two-sample proportions test, *P*-value <2.2e – 16). These disparities in average copy numbers or percentages of

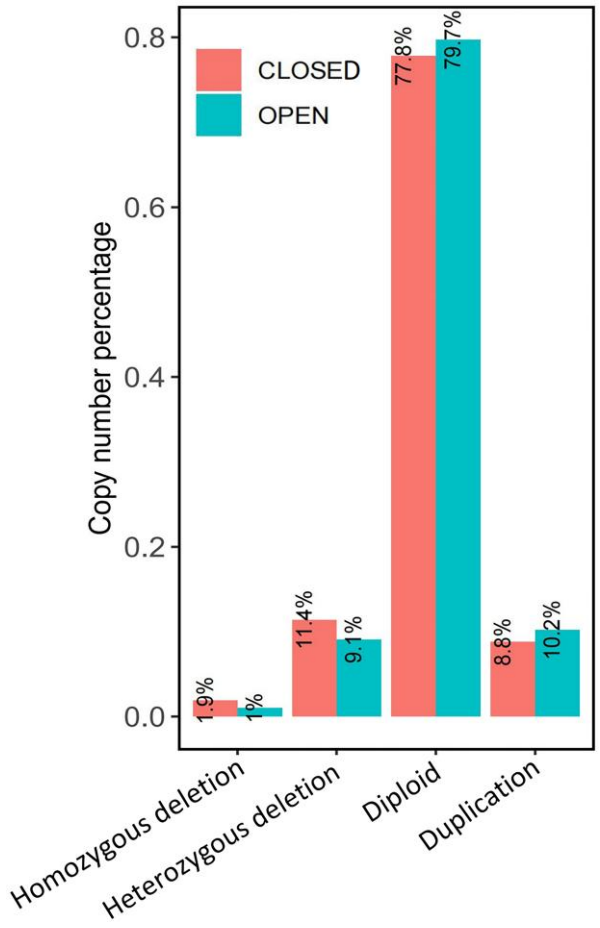

Fig. 3. Distribution of copy number percentage for African strains stratified by type (Homozygous/heterozygous deletion, diploid and duplication). In open loci, duplication (copy gains) are more frequent, while deletion (copy losses) are less frequent, compared to closed chromatin loci in *C. elegans*.

multi-allelic variants reflect the differential selection pressures acting on open and closed chromatin loci. Copy losses in open chromatin loci can have significant impacts on genome functionalities, intensifying selection pressure to eliminate such variants.

Among all loci, around 7% exhibit multi-allelic characteristics, with more than two possible alleles (copy numbers) observed among N=10 individuals in at least one population of *C. elegans* strains from Africa, Australia, France, and Hawaii. Subsequent analysis focused on the remaining 1,902,566 segments, which were found to be either mono- or bi-allelic across all populations. These loci were further classified into two groups based on their overlap with 42,245 regions of high chromatin accessibility. Specifically, 120,562 loci were located in regions of high chromatin accessibility (open), while the remaining 1,782,004 loci were situated in regions of low chromatin accessibility (closed). Next, we have taken 100 bootstrap samples from the open and closed chromatin segments to examine the evolution of CNVs in each segment. In each bootstrap repeat, we sampled a total of 1,902,566 loci. Subsequently, these loci were divided into open and closed classes, which were used separately for estimating mutation rates and assessing the fitness of copy numbers within each class. Figure 4 shows the boxplots for the mutation rates and fitness parameters estimated using PoMoCNV, specifically for the open and closed chromatin segments. The observations presented in Fig. 4a indicate that open chromatin segments (depicted in blue) exhibit higher mutation rates with increased copy numbers compared to closed chromatin segments (depicted in red). Specifically, the values of α_{12}, α_{23}, and α_{32} are higher for open segments in comparison to the closed segments. Conversely, open chromatin segments (blue) demonstrate lower mutation rates with decreased copy numbers compared to closed chromatin segments (red). In this case, the values of α_{10}, α_{21}, and α_{31} are lower for open segments relative to the closed segments. Additionally, as depicted in Fig. 4b, the relative fitness loss ($\mathbb{S}$) per copy deletion is notably higher for the open chromatin segments compared to the closed segments. According to the Wilcoxon rank sum test, the difference in the relative fitness loss per copy deletion between open and closed segments is statistically significant at the *alpha* = 0.05 level.

It is important to note that the majority of functional segments, such as enhancers and promoters, primarily reside within the open chromatin segments (Ernst and Kellis 2010, 2012; Thurman et al. 2012; Denny et al. 2016; Cusanovich et al. 2018; Klemm et al. 2019). These functional segments play a crucial role in transcriptional regulation by acting as binding sites for Transcription Factors (TFs). Specifically, TFs facilitate the spatial proximity between enhancers and promoters, enabling the precise regulation of downstream genes in the 3D genomic landscape (Pliner et al. 2018; Luo et al. 2022). Consequently, any loss of copy numbers within these open chromatin segments can have a more profound impact on the genome. These copy losses in functional segments disrupt the intricate regulatory interactions and can lead to significant impairments in gene expression, potentially resulting in more severe consequences for genomic stability and overall functionality.

CNV Evolution in the *C. elegans* Data with Distinct Bottleneck Sizes

The performance of PoMoCNV is further assessed in CNV datasets from a recent study (Konrad et al. 2018), in which a mutation accumulation (MA) framework was employed to investigate the emergence rate of gene CNVs under varying selection intensities in *C. elegans*. These MA lines were derived from a single hermaphrodite ancestor (N2). In each

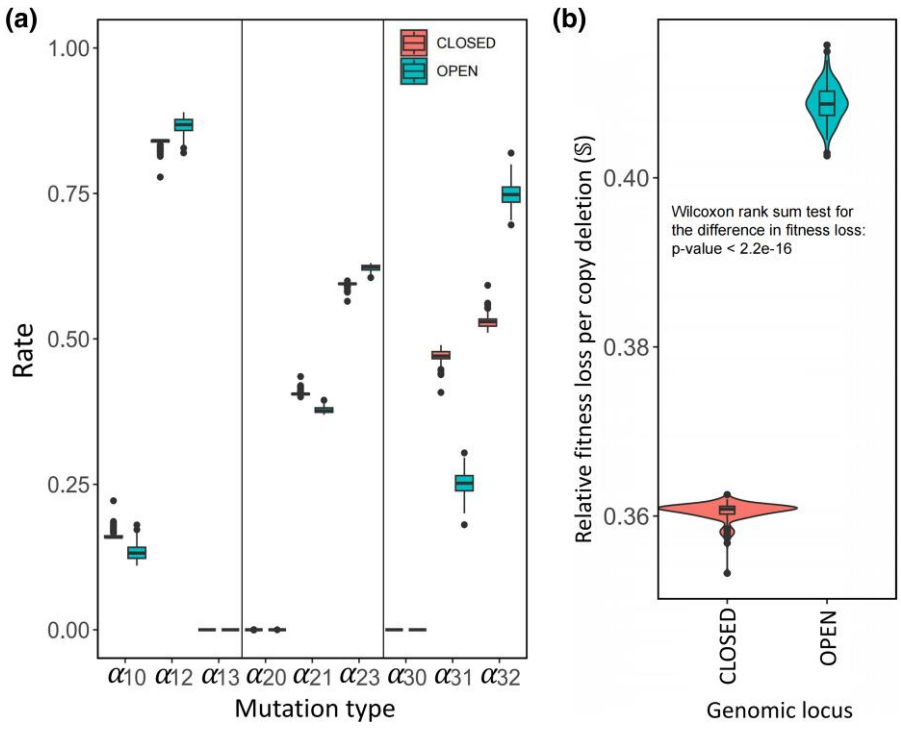

Fig. 4. The relative mutation rates and relative fitness loss (𝕊) per copy deletion in open and closed genomic segments in *C. elegans*. PoMoCNV calculates the relative mutation rates a) and estimates the relative fitness loss resulting from each copy deletion in both open and closed segments b).

generation, a bottlenecking process was implemented, resulting in a reduced population size of either $B = 1$, $B = 10$, or $B = 100$ hermaphrodites. Specifically, only one, ten, or one hundred hermaphrodites were selected to reproduce and pass on their genetic information to the next generation, creating the MA lines $B = 1$, $B = 10$, or $B = 100$. After repeating this bottlenecking process for over 400 generations, 18, 40, and 25 descendant individuals from the final generations of each MA line were selected for performing oaCGH (oligonucleotide array comparative genome hybridization) experiments. The aim of these experiments was to assess the presence of copy gains and losses in genomic segments under neutrality ($B = 1$) and with increasing selection intensity ($B = 10$, $B = 100$). These experiments provide a controlled setting where the selection strengths are already known, offering a benchmark against which we can evaluate the performance of our method in inferring selection strengths. *C. elegans* individuals from each bottleneck size were clustered based on similarities in their copy number patterns using Ward's method and the heatmap.2 package in R. Supplementary figure S2, Supplementary Material online displays the heatmaps showing copy numbers in genomic segments with CNVs and the dendrograms representing similarities among *C. elegans* individuals from each bottleneck size. In supplementary figure S2, Supplementary Material online, individuals with similar copy gain or copy loss patterns are closer to each other in the dendrogram.

The *C. elegans* individuals from each bottleneck size were further divided into four populations, each consisting of five individuals with similar copy gain or loss patterns based on Ward's clustering results. Out of the 25 available individuals for the $B = 100$ MA line, we utilized a subset of 20 individuals in each bootstrap repeat for our PoMoCNV analysis. This approach allows us to assess how PoMoCNV handles the inclusion or exclusion of a few individuals within each bootstrap repeat, thus evaluating its robustness. In bottleneck sizes $B = 10$ and $B = 100$, there are no common individuals among the four populations. However, in $B = 1$ with 18 individuals, two individuals are shared between two populations to create four populations of size five. The grouping of *C. elegans* individuals into four populations for each B serves the purpose of facilitating the estimation of evolutionary parameters, such as mutation rates, within PoMoCNV. PoMoCNV relies on the populations to trace their evolutionary relationships back to their ancestors, allowing for the inference of how copy numbers have changed over time. Additionally, individuals within each population are expected to exhibit greater similarity to one another compared to individuals from different populations, aiding in the analysis of evolutionary dynamics within and between populations.

To assess the sensitivity of the results to the *C. elegans* individuals included in the constructed populations, we performed 100 bootstrap repeats to sample similar individuals using Ward's score and assigned them to the same

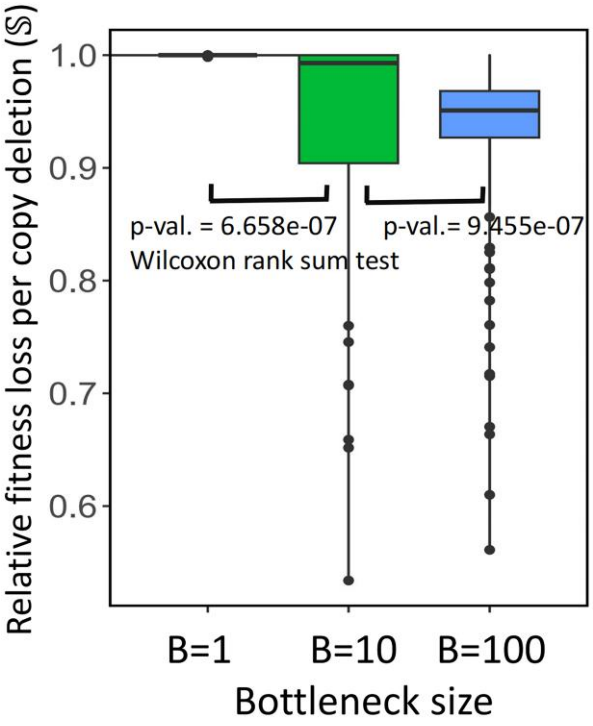

Fig. 5. The relative fitness loss per copy deletion (𝕊) estimated with PoMoCNV for three *C. elegans* datasets with bottleneck sizes $B = 1$, $B = 10$, and $B = 100$. For the $B = 1$ line, in the absence of selection, more CNVs are observed in the population evolved over 400 generations, which could negatively impact its fitness. PoMoCNV correctly inferred $\mathbb{S} = 1$, indicating a higher fitness loss for the $B = 1$ line, as expected, when compared to the $B = 10$ and $B = 100$ lines. Increasing B corresponds to stronger purifying selection, resulting in fewer CNVs in the population evolved under these regimes and, consequently, lower inferred values of $\mathbb{S}$. Estimations were performed using 100 bootstrap repeats, dividing *C. elegans* individuals into four populations of size five for each bottleneck size. Individuals within each population share similarities in copy gains and losses compared to the N2 reference genome.

population. For each bottleneck size ($B = 1$, $B = 10$, $B = 100$), we used PoMoCNV to estimate the relative fitness loss (𝕊) per copy deletion from each bootstrap repeat with four populations of five *C. elegans* individuals. Due to the limited number of CNVs observed in the MA experiments with bottleneck sizes $B = 1$, $B = 10$, and $B = 100$, specifically with 41, 28, and 14 duplication calls, and 39, 16, and 13 deletion calls shared among the sampled *C. elegans* individuals, the estimation of mutation rates between different copy numbers is not reported for this dataset. Figure 5 presents boxplots illustrating the relative fitness loss per copy deletion for $B = 1$, $B = 10$, and $B = 100$, showing that these estimates statistically differ among bottleneck sizes based on the Wilcoxon rank sum test. Here, $\mathbb{S}$ shows how much a deletion, given the CNVs in a population, would negatively affect the fitness relative to a diploid genotype (two copies). Also, a genotype with two deletions has $2\mathbb{S}$ lower fitness relative to a diploid genotype. Note that these genotypes have nonzero absolute fitness. For $B = 1$, in the absence of purifying selection, more CNVs are observed in the final population, which, on average, have a more negative effect on fitness. In comparison, under the B=100 regime, fewer copy gains and losses are observed, indicating strong selection. Consequently, a lower $\mathbb{S}$ is inferred for B=100.

CNV Evolution in the Synthetic Data

In PoMoCNV, the transition probabilities from mono-allelic or bi-allelic genomic loci are influenced by distinct mechanisms. At a mono-allelic genomic locus where all individuals carry allele (copy number) I, a transition to a bi-allelic locus with alleles I and J can occur within a single generation through a mutation (change) in the copy number of the locus. Conversely, transitions from a bi-allelic state with alleles I and J depend on the allele fitness and their frequencies. See full details in the Materials and Methods section. Then, through temporal disentanglement of the effects of copy number mutations and selection events, the PoMoCNV pipeline is capable of estimating pertinent parameters without confounding them. To further evaluate the effectiveness of this strategy in estimating copy-number mutation rates and the impact on relative fitness loss (𝕊) from each copy deletion, we simulated a dataset of populations that evolved under different bottleneck sizes.

The procedure for generating the synthetic data is described in full detail in the Materials and Methods section. We have simulated datasets considering two bottleneck sizes, $B = 1{,}000$ and $B = 3{,}000$, for a total of $L = 10{,}000$ genes. In the simulation, we considered $p_{dup} = 0.01$ and $p_{del} = 0.005$ as the probabilities of copy duplication and deletion for each locus, in each generation, see Materials and Methods for details. We utilized PoMoCNV to estimate mutation rates and relative fitness loss (𝕊) per copy deletion for these two bottleneck sizes, with 100 bootstrap repeats. Figure 6a shows the estimated mutation rates and relative fitness loss (𝕊) for each copy deletion for $B = 1{,}000$. Figure 6b illustrates the relative fitness loss (𝕊) in a population characterized by a larger bottleneck size of $B = 3{,}000$. Overall, Fig. 6 demonstrates that the estimated mutation rates remain consistent with the findings from previous sections. Furthermore, when the population undergoes less stringent bottleneck sizes with $B = 3{,}000$ compared to $B = 1{,}000$, the relative fitness loss per copy deletion is skewed towards lower values. Indeed, a less stringent bottleneck size results in higher fitness and selection rates for copy number gains and losses in genomic segments.

Discussion

It is generally assumed that the Dynamics of CNVs in populations, mirroring that of SNVs, is shaped by genetic drift and demographic history, as well as selection (Coop et al.

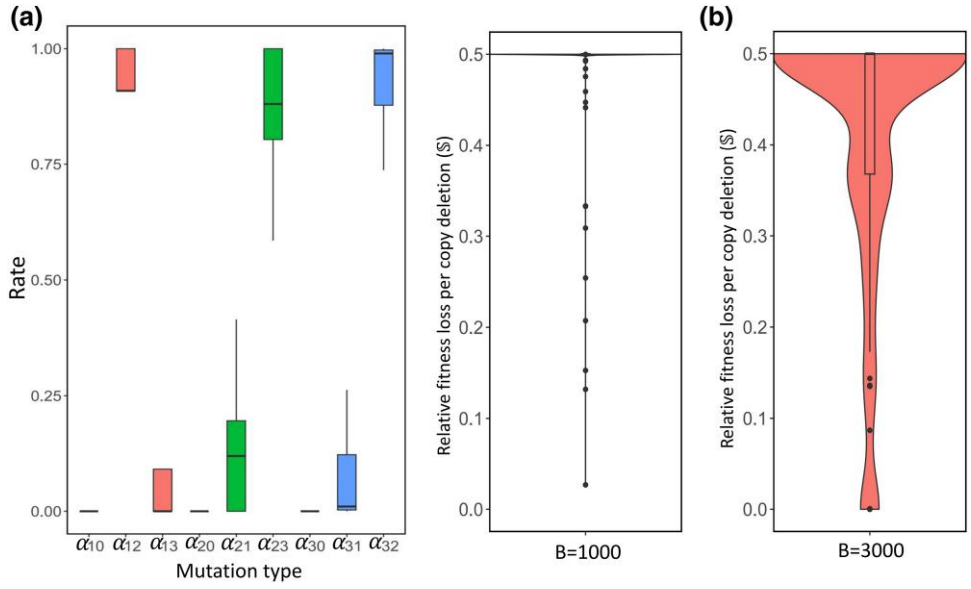

Fig. 6. Parameter estimation for two bottleneck sizes, $B = 1{,}000$ and $B = 3{,}000$, on simulated data. (a) Estimated mutation rates and relative fitness loss ($\mathbb{S}$) per copy deletion for $B = 1{,}000$. (b) Estimated relative fitness loss ($\mathbb{S}$) for $B = 3{,}000$. Mutation rate and relative fitness loss estimates under varying bottleneck sizes align with previous findings, affirming PoMoCNV's reliability in estimating both parameters without confounding them.

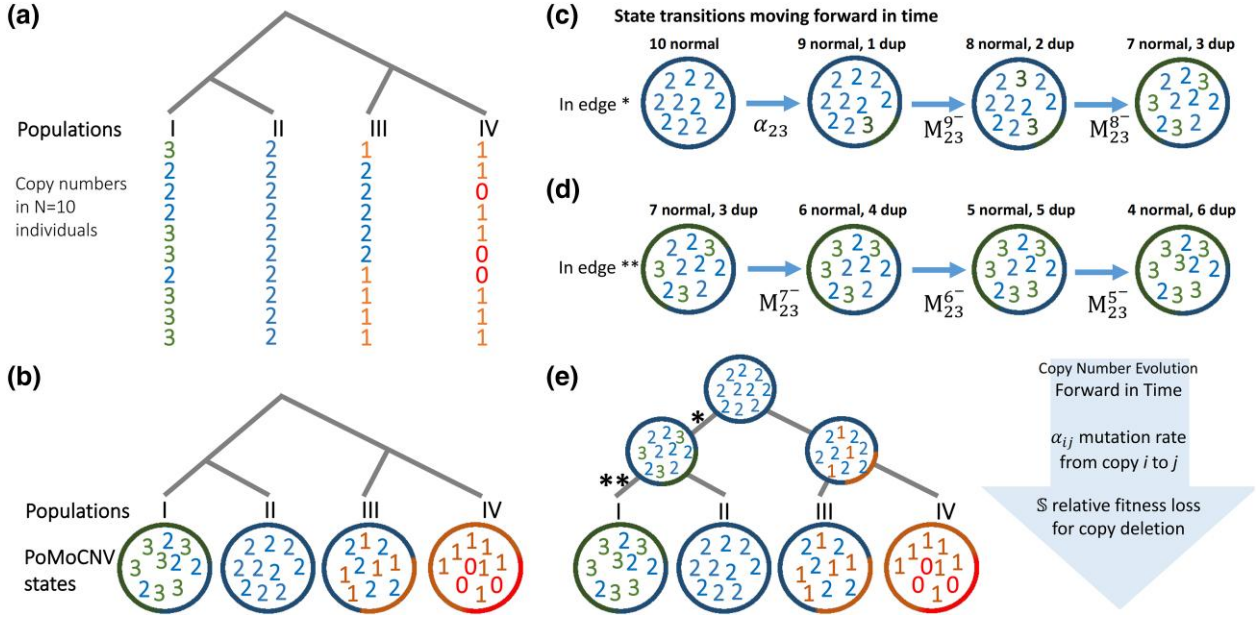

Fig. 7. Pipeline for PoMoCNV (POlymorphism-aware phylogenetic MOdel (PoMo) for CNV). a) The copy numbers (alleles) of $N = 10$ individuals from four *C. elegans* populations are displayed along the phylogenetic tree for a specific genomic segment. b) Each genomic segment within each *C. elegans* population is associated with a PoMoCNV state based on the copy number distribution. c) In mono-allelic states, such as at the root node of edge ⋆, where only one copy number exists in a population, PoMoCNV utilizes a mutation rate (α) to introduce a new allele (copy number) resulting in a potential bi-allelic population at the inner nodes of the phylogenetic tree. d) In bi-allelic states, such as at the root node of edge ⋆⋆, which include distinct copy numbers within a population, PoMoCNV utilizes a Moran model with a state transition rate matrix (M_N) to simulate forward movement between different states over time. e) The Felsenstein pruning algorithm is utilized by PoMoCNV to aggregate the likelihoods across all potential states at the inner nodes. This allows for the estimation of the most probable fitness and mutation rate associated with copy numbers.

2009; Hollox et al. 2022). In addition, given the large disruptive effects of structural variations on phenotype, CNVs are expected to be nonuniformly distributed across the genome and lay away from functional segments (Conrad and Hurles 2007). The distribution of CNVs in the human genome seemingly abides by these expectations (Sudmant et al. 2015). However, the pattern of CNV distribution does vary from species to species. For example, an analysis of CNVs in two dairy cattle breeds showed a higher overlap between segments of CNVs and genes than expected (Lee et al. 2020).

The fundamental obstacle to a thorough and satisfying evolutionary account of CNVs in a given species is the plethora of mechanisms—i.e. retrotransposons, nonallelic homologous recombination, and nonhomologous end joining—that can result in CNVs of different sizes and functions. Our knowledge of demographic history of a lineage, which inevitably leaves its fingerprints on the frequency of CNVs, is generally based on a speculated ancestral state that is the result of comparison with a related lineage (Conrad and Hurles 2007). Odd copy numbers typically arise from events like unequal crossovers, deletions, and duplications, which are influenced by the genetic architecture and reproductive strategies of the organism. In selfing organisms like *C. elegans*, where individuals can reproduce asexually, reduced genetic recombination and other inherent mechanisms of self-fertilization can lead to the loss of genetic diversity (Barrett et al. 2014; Greer et al. 2022) and increased homozygosity. Furthermore, the genetic bottleneck associated with selfing can lead to the fixation of certain alleles. Then, selfing may reduce the probability of observing odd copy numbers, impacting the genomic landscape and evolutionary trajectories within such organisms.

PoMoCNV is an attempt to infer aspects of the evolutionary dynamics of CNVs from population genomic data. In the case study of the MA experiment on *C. elegans* (Konrad et al. 2018), given the known ancestral state and varying levels of genetic drift, imposed via varying the bottleneck size in the experiment, enabled us to comprehensively examine the ability of *PoMoCNV* to infer the selective values of CNVs in contrast to our expectation. Based on the available data on chromatin accessibility in *C. elegans*, *PoMoCNV* indicates that CNVs tend to emerge, propagate, and contract more easily in closed chromatin regions. Moreover, the change in copy number in these regions, which are less rich in functional elements, appears to have a comparatively less detrimental effect than CNVs occurring in open chromatin regions. In addition, the fitness effects of CNVs, as inferred by *PoMoCNV*, reflect the interplay between selection and drift, that is, the inferred negative effects of CNVs on fitness decreases in concert with the increase in population size, and consequently the increase in the efficacy of natural selection.

Materials and Methods

In this study, we model the evolution of CNVs along a phylogenetic tree relating populations of individuals. For each individual, the model considers four possible copy numbers for a genomic locus including homozygous deletion, heterozygous deletion, normal (two copies for diploid organisms), and duplication. We also assume that there exist a maximum of two distinct alleles (representing copy numbers) for each locus in a population of N individuals. Here, we trace the evolution of copy numbers forward in time.

We consider two parameters to investigate the evolutionary dynamics of copy numbers, (i) mutation rates between alleles (copy numbers), and (ii) fitness coefficient associated with each allele, indicating the relative advantage or disadvantage of a specific copy number at a genomic locus for the organism, see Equation (4) and Table 1. The mutation rate refers to the probability of a change in the copy number of a genomic locus of an individual in a single generation. For instance, the mutation rate can specify the chance that a normal locus (with two copies) duplicates, resulting in three copies. Such a mutation can serve as a driving force to introduce a new allele in a mono-allelic population, leading to a transition to a bi-allelic population, at this locus (see Equation (1)). However, changes in the frequency of alleles in a bi-allelic population are influenced by the frequencies of each allele and their respective fitness values following the Moran framework (see Equations (2) and (3)). In the Moran framework (Moran 1962; Nowak et al. 2004), the population evolves via a stochastic birth–death process for each genomic locus forward in time. In such a process, at each time step, one individual is selected to reproduce and transmit all its copies (allele) to an offspring. Concurrently, another individual is randomly chosen to die and be removed from the population. For each genomic locus, the reproducing and dying individuals each have their defined copy numbers. This cycle of reproduction and death, in combination with fitness values, allows the copy numbers to evolve over generations, as individuals with different copy number profiles are chosen to pass on their genetic material or die out in each time step.

In summary, in our model, the transition from a mono-allelic to a bi-allelic state results solely from mutation, without involvement of selection. Conversely, alterations in allele frequencies within a bi-allelic state are solely influenced by selection. Therefore, we consider a temporal separation between mutation and selection events that act on the mono-allelic and bi-allelic loci within a population, enabling a more reliable estimation of mutation rate and copy number fitness without confounding them. This integrative approach advances existing literature which has largely focused on sequence features like SNVs and their frequencies in a population in isolation. It will help address long-standing questions regarding CNV evolutionary roles,

including whether they serve as reservoirs of functional variation or transient neutral changes. Overall, this research exemplifies how modeling the interplay between selection and copy number changes can further our understanding of genome evolution. By leveraging information on population-level allele (copy number) frequencies in different present-day populations/species, our model aims to disentangle the relative contributions of selection and mutation in driving observed CNV patterns.

States for Copy Number Frequencies

We assume that the phylogenetic tree of the populations or species is known and we use the Moran model (Nowak 2006; De Maio et al. 2013) to estimate the evolution of alleles (copy numbers) along branches in a phylogenetic tree. The modeling process of PoMoCNV is depicted in Fig. 7 for a single genomic locus (e.g. gene), with $p = 4$ populations represented as four leaves of the tree. Each population includes $N = 10$ diploid individuals (*C. elegans* strains), Fig. 7a. For each individual, the model considers four possible alleles (copy numbers) that a gene or genomic segment can adopt—homozygous and heterozygous deletions, normal, and duplication—which are denoting copy numbers 0, 1, 2 and 3, respectively. See Table 1 for definitions.

Figure 7b demonstrates how PoMoCNV defines a state. It considers the allele types and their frequencies within each population for a single genomic segment. For instance, in population II, we have a mono-allelic state in which all 10 individuals have allele 2 (two copies). However, in population I, 4 individuals carry allele 2 and 6 individuals carry allele 3, resulting in a bi-allelic state. Our modeling enforces that at each genomic segment, up to two alleles are allowed in a population. The possible alleles are denoted with I and J, representing the four copy numbers 0, 1, 2 and 3. In Table 1, a state is uniquely defined by the allele frequency distribution denoted by $\begin{pmatrix} i & I \\ N-i & J \end{pmatrix}$, where i and $N - i$ refer to the counts of alleles I and J, respectively, in the population of N individuals. This allows representation of all possible allele configurations ranging from mono-allelic states with all I as $\begin{pmatrix} N & I \\ 0 & J \end{pmatrix}$, or with all J as $\begin{pmatrix} 0 & I \\ N & J \end{pmatrix}$. The intermediate bi-allelic states can be as $\begin{pmatrix} 1 & I \\ N-1 & J \end{pmatrix}$, $\begin{pmatrix} 2 & I \\ N-2 & J \end{pmatrix}$, …, $\begin{pmatrix} N-1 & I \\ 1 & J \end{pmatrix}$. For a population with $N = 10$ diploid individuals and four possible allele types, there are a total of $4 + 6 \times (N - 1) = 58$ distinct allele frequency states, encompassing the spectrum from mono- to bi-allelic loci. More precisely, there are four possible mono-allelic states where all $N = 10$ individuals can have 0, 1, 2, or 3 copies in their locus. In bi-allelic states, there can be six allele (copy number) combinations: {0, 1}, {0, 2}, {0, 3}, {1, 2}, {1, 3}, and {2, 3}, at a locus. For each combination, such as {0, 1}, among $N = 10$ individuals, the possible bi-allelic states are $\begin{pmatrix} 1 & I=0 \\ 9 & J=1 \end{pmatrix}$, $\begin{pmatrix} 2 & I=0 \\ 8 & J=1 \end{pmatrix}$,…, or $\begin{pmatrix} 9 & I=0 \\ 1 & J=1 \end{pmatrix}$, in a bi-allelic locus.

Transition Probabilities

By tracking transitions between frequency states over generations, we can estimate key parameters such as mutation rates of copy numbers and relative fitness loss per copy deletion at a genomic locus that govern the transition and selection processes affecting copy number variation. This analysis specifically focuses on copy number variation at the genomic segment within the branches of the phylogenetic tree that connects different populations. The transition from a mono-allelic state to a bi-allelic state can occur through the introduction of a new copy number in the population, which arises from a mutation event, specifically, a change in the copy number of a locus. Specifically, Equation (1) shows $M_{I,J}^{N^-}$, which is the probability of transitioning from a mono-allelic state $\begin{pmatrix} N & I \\ 0 & J \end{pmatrix}$ (where all N individuals have a copy number of I), to a bi-allelic state $\begin{pmatrix} N-1 & I \\ 1 & J \end{pmatrix}$ (characterized by copy numbers I and J, where the frequencies are $\{N - 1, 1\}$).

$$M_{I,J}^{N^-} = N \times \alpha_{IJ} \tag{1}$$

where α_{IJ} represents the mutation rate of an individual with copy number I to copy number J within a single generation (see Table 1).

Consider a bi-allelic genomic segment with copy numbers I and J which are present in the population, the Moran model allows calculation of transition probabilities between allele frequency states. Specifically, the probability of transitioning from state $\begin{pmatrix} i & I \\ N-i & J \end{pmatrix}$ to $\begin{pmatrix} i+1 & I \\ N-i-1 & J \end{pmatrix}$ in one generation is denoted as $M_{I,J}^{i^+}$ and is given by the following equation, for details see (De Maio et al. 2013; Alexandre et al. 2025).

$$M_{I,J}^{i^+} = \frac{N-i}{N} \times \frac{i(1+S_I-S_J)}{i(1+S_I-S_J)+N-i} \tag{2}$$

In this formulation, $\frac{N-i}{N}$ represents the probability of randomly selecting an individual with allele J to die, while the term $\frac{i(1+S_I-S_J)}{i(1+S_I-S_J)+N-i}$ gives the probability of choosing an individual with allele I to reproduce. The formulation of reproduction probability in Equation (2) considers the frequencies i and $N - i$, along with the fitness values S_I and S_J, of alleles I and J at a bi-allelic locus in a population of N individuals. Also, S_I is the relative advantage or disadvantage of copy

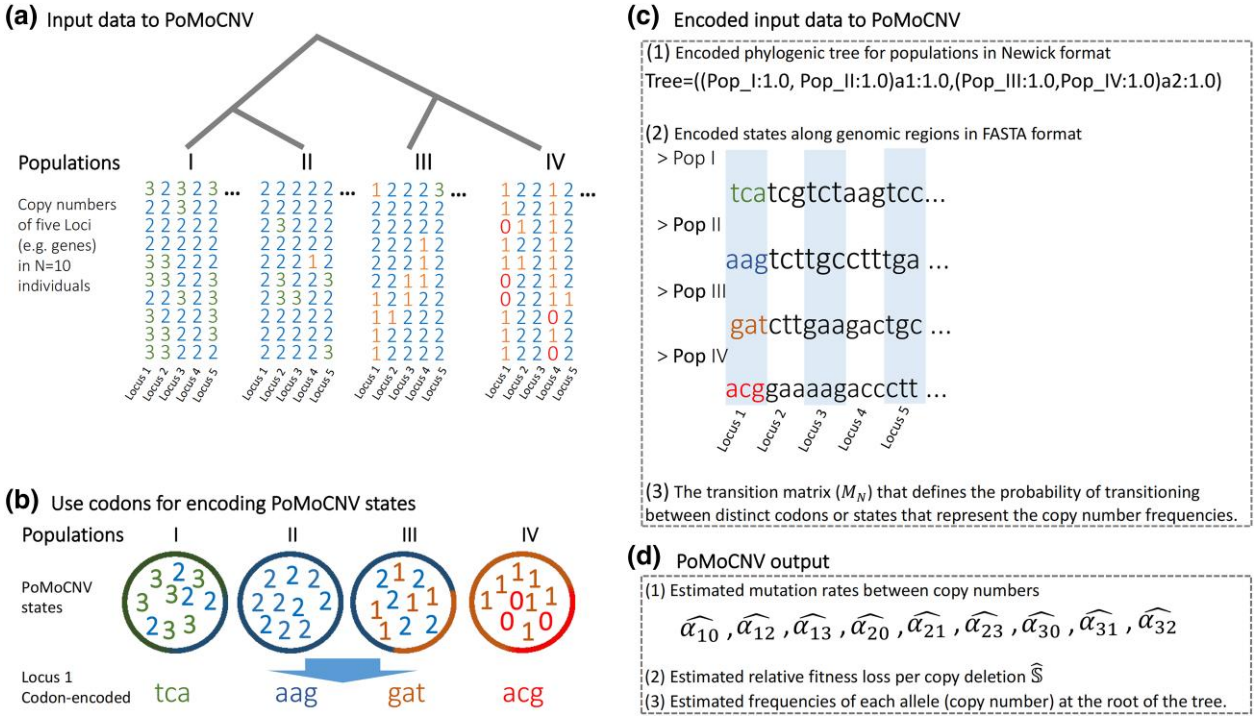

Fig. 8. Triplets (codons) for encoding PoMoCNV states. a) Copy numbers for five genomic loci are depicted across four populations, with $N = 10$ individuals per population. It is assumed that the phylogenetic tree representing the evolutionary relationships among populations is known. b) Triplets (codons) are used to encode the PoMoCNV states. For instance, in locus 1, the frequencies of copy numbers among 10 individuals in each population uniquely define a state within PoMoCNV, linked to a corresponding triplet. c) The phylogenetic tree of the four populations, along with the PoMoCNV states, is encoded in Newick and triplets-encoded FASTA files. These files are input into PoMoCNV. d) PoMoCNV utilizes the HyPhy package to fit a maximum likelihood evolutionary model to the triplet-encoded data. This process estimates parameters of mutation rate and relative fitness loss per copy deletion within the transition matrix.

number I at a genomic locus for the organism, see Equation (4) and Table 1. The intuition behind the above formulation is that when the fitness values are equal ($S_I = S_J$), it simplifies to i/N, showing that selection of individuals with allele I for reproduction is solely based on its frequency i among N individuals. However, if allele I has a fitness advantage over allele J ($S_I > S_J$), its frequency i is adjusted by a factor of $(1 + S_I - S_J)$ to enhance its selection for reproduction, and conversely for $S_I < S_J$. Similarly, the reverse transition probability $M^{i-}_{I,J}$ from state $\begin{pmatrix} i & I \\ N-i & J \end{pmatrix}$ to $\begin{pmatrix} i-1 & I \\ N-i+1 & J \end{pmatrix}$ is:

$$M^{i-}_{I,J} = \frac{i}{N} \times \frac{N-i}{i(1+S_I-S_J)+N-i} \tag{3}$$

These transition probabilities depend on both the current allele frequencies i and j, as well as the relative fitness values S_I and S_J associated with each allele. The fitness of copy numbers are considered as:

$$S_0 = 1 - 2\mathbb{S}, \quad S_1 = 1 - \mathbb{S}, \quad S_2 = 1, \quad S_3 = 1 - 0.5\mathbb{S} \tag{4}$$

In this formulation, we designate genomic segments with normal copies (copy number=2 for diploid organism) as a reference, assigning them a fitness value of $S_2 = 1$ (See Table 1). The relative fitness loss ($\mathbb{S}$) parameter quantifies the relative disadvantage caused by each copy deletion at a genomic locus, in comparison to a locus with two copies having a fitness of $S_2 = 1$. It should be noted that we assume genotypes to have nonzero absolute fitness since changes in copy numbers affect the fitness of a given genotype proportion to $\mathbb{S}$, relative to the diploid genotype (with two copies). In Equation (4), the relative fitness loss for homozygous and heterozygous deletions (corresponding to 2-copy and 1-copy deletions, respectively) is set at $2\mathbb{S}$ and $\mathbb{S}$. Additionally, each copy gain incurs a relative fitness loss of $0.5\mathbb{S}$, making a homozygous deletion four times more deleterious than a copy gain. We then applied a gradual increase in relative fitness loss per copy deletion to reflect the varying impact of different copy number alterations on fitness based on the literature (Schrider and Hahn 2010; Sung et al. 2016; Gonzalez et al. 2019; Chunduri et al. 2022). This modeling is also inspired by a recent study indicating that deletions appear to be approximately four times as deleterious as duplications (Hujoel et al. 2022). For example, spinal muscular atrophy (SMA) is caused by homozygous deletion of SMN1 (Lefebvre et al. 1995), and individuals with the mildest form of SMA often possess three or more copies of SMN2

(Feldkötter et al. 2002; Mailman et al. 2002), enabling their survival into adulthood. Moreover, homozygous deletions exhibit a higher rate of occurrence within genomic regions characterized by a lower gene density, compared to other types of genetic variations (Girirajan et al. 2011). In other complex diseases such as autism, schizophrenia, bipolar disorder, amyotrophic lateral sclerosis, attention deficit hyperactivity disorder (ADHD), and Tourette syndrome, large deletions play a crucial role (Girirajan et al. 2011). In individuals experiencing developmental delay, the occurrence of deletions is more prevalent compared to reciprocal duplications (Girirajan et al. 2011). However, copy gains still impose a burden on the genome by increasing its size. Consequently, genomic segments with copy gains will exhibit lower fitness compared to segments with normal copies (Simon-Loriere and Holmes 2013).

Figure 7c and d demonstrates how transitions between distinct states can occur forward in time across generations. At the root node of edge ⋆, as depicted in Fig. 7e, the transition from a mono-allelic locus to a bi-allelic locus within a single generation necessitates a mutation in the copy number of an individual carrying allele 2. This mutation introduces a new allele type (such as copy number 3 in this case) to the population. The rate of this mutation is denoted as α_{23}. In a bi-allelic state like that at the root node of edge ⋆⋆, the transition to a state with a change in allele 2 frequency is determined by the fitness parameters S_2 and S_3 of alleles 2 and 3, along with their frequencies. Moving forward in time, the transitions from bi-allelic states in other generations are influenced by the selection process (see Fig. 7c and d). The parameters (mutation rates and relative fitness loss ($\mathbb{S}$) per copy deletion) that impact the transitions between different states are illustrated in Fig. 7e.

Equation (5) shows a subset of the transition matrix (M_N) for different mono-allelic and bi-allelic states, consisting of alleles I and J, e.g. $I = 0$ and $J = 1$. In this matrix, the rows represent the original states, and the columns represent the possible states reached in a single generation. Here, the last row illustrates the transition from a mono-allelic state with allele I to a state where one locus is mutated to allele J in the next generation, governed by the mutation rate α_{IJ} (Equation (1)). However, other rows illustrate the transitions originating from a bi-allelic locus and are governed by the allele frequencies and their fitness values, as shown in Equations (2), and (3).

Then, in our framework, mutation solely operates on the mono-allelic loci, and mutation events lead to the transition from mono-allelic to bi-allelic loci in one generation. Upon the emergence of bi-allelic locus, selection acts to either retain or eliminate the mutated locus from the population. In other words, transitions originating from bi-allelic loci are shaped by selection events that factor in allele frequencies and fitness values. By introducing this temporal separation, we aim to disentangle the effects of mutation and selection, allowing us to effectively estimate both mutation rate and fitness parameters.

$$
M_N \quad
\begin{matrix}
 & \binom{0}{N}\ \binom{I}{J} & \binom{1}{N-1}\ \binom{I}{J} & \binom{2}{N-2}\ \binom{I}{J} & \binom{3}{N-3}\ \binom{I}{J} & \cdots & \binom{N-2}{2}\ \binom{I}{J} & \binom{N-1}{1}\ \binom{I}{J} & \binom{N}{0}\ \binom{I}{J} \\
\binom{1}{N-1}\ \binom{I}{J} & & & & & & & & \\
\binom{2}{N-2}\ \binom{I}{J} & & & & & & & & \\
\binom{3}{N-3}\ \binom{I}{J} & & & & & & & & \\
\cdots & & & & & & & & \\
\binom{N-1}{1}\ \binom{I}{J} & & & & & & & & \\
\binom{N}{0}\ \binom{I}{J} & & & & & & & &
\end{matrix}
=
\begin{pmatrix}
M_{I,J}^{1-} & \star & M_{I,J}^{1+} & & & & & & \\
 & M_{I,J}^{2-} & \star & M_{I,J}^{2+} & & & & & \\
 & & M_{I,J}^{3-} & \star & M_{I,J}^{3+} & & & & \\
 & & & \cdots & & & & & \\
 & & & & & & M_{I,J}^{N-1-} & \star & M_{I,J}^{N-1+} \\
 & & & & & & & M_{I,J}^{N-} & \star
\end{pmatrix}
\tag{5}
$$

The final stage in formulating our model involves the conversion of the discrete-time Markov chain (measured in generations) into a continuous-time framework. A continuous-time Markov chain is characterized by its instantaneous rate matrix Q_N. For our continuous-time process, we define the instantaneous rate matrix as $Q_N := N(M_N - I)$, where I represents the identity matrix (for details see De Maio et al. 2013). Consequently, the transition probabilities at coalescent time t/N are expressed as $P(t/N) = e^{Q_N(t/N)}$, where t denotes the virtual generation count, now permitting noninteger values.

Encoding States and Phylogenetic Tree

The PoMoCNV states are represented using three letters (triplet, which is called a codon in the core software). For $N = 10$, we utilized 58 triplets to encode states corresponding to distinct allele frequency patterns per genomic locus for a population. In Fig. 8, the encoding process is detailed. Figure 8a illustrates the copy numbers of $N = 10$ individuals from each population across five genomic loci and four populations, assuming a known phylogenetic tree representing their evolutionary relationships. In each population, copy number frequencies uniquely define a state in PoMoCNV, which is encoded with the corresponding triplet (Fig. 8b). PoMoCNV states across all genomic loci are encoded with a sequence of triplets for each population (Fig. 8c). Inputs into PoMoCNV are (1) the phylogenetic tree of populations defined in Newick format, (2) the triplet sequences encoding states, in FASTA format, and (3) the transition matrix (M_N), its mutation rate and fitness parameters (Fig. 8c).

PoMoCNV utilizes the HyPhy package (Pond et al. 2004), for conducting maximum likelihood analysis of evolutionary models within a continuous-time Markov framework, applied to the evolving triplets (Fig. 8d). These triplets represent the pattern of copy number frequency (state) within a population for genomic loci. The transition matrix, including all transition probabilities, is obtained using Equations (1)–(3). These probabilities depend on mutation rates and the fitness of each copy number S_I for $I = 0, 1, 2, 3$. We benefited from the assumption that S_I values are linear functions of $\mathbb{S}$ (detailed in Materials and Methods) to make the search space limited. The goal is to maximize the likelihood function to determine the optimal values of $\mathbb{S}$ and mutation rates that best fit the observed data (phylogenetic tree and CNV states of genomic regions). This optimization employs Bayesian approximation algorithms. For more details on implementation, see the HyPhy tutorials (Pond et al. 2004).

Simulation of the Evolution of CNVs

To test the suitability of the PoMoCNV framework when evolutionary parameters are known *a priori*, we constructed a basic model to simulate the evolution of CNVs in finite populations. In this model, the genotype of an individual is represented by **G**, as a vector of length $2L$, assuming L genes each with 2 copies. The fitness of an individual is defined as a function of the number of copies of L genes:

$$w = \prod_{i}^{L} 1 - e^{-\lambda\left[\ln(c_i) - \ln(c_{opt})\right]^2}, \quad (6)$$

where c_i is number of copies for gene i and c_{opt} is the optimal number of copies, which we assume to be 2, and λ determines the intensity of the stabilizing selection on the copy number of genes (unless otherwise specified, $\lambda = 0.1$). This model was inspired by a recent model of chromosome degradation (Lenormand et al. 2020). At the start, there are 2 copies for each of the L genes. At each generation, each locus can duplicate with probability p_{dup} or lose a copy with probability p_{del}. We assume that the probability of duplication and deletion is uniform across the genome. We assume no recombination, i.e. the offspring is an exact copy of the parent. The offspring undergoes duplication and deletion events. Afterwards, next generation is created by a Wright–Fisher process, i.e. sampling B offspring with replacement proportional to their fitness.

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Funding

This work did not receive any specific funding.

Data Availability

The HyPhy codes and data analyzed in the article are available on the GitHub page of the project "PoMoCNV" at the following URL: https://github.com/CompBioIPM/PoMoCNV.

Literature cited

Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011:21(6):974–984. https://doi.org/10.1101/gr.114876.110.

Alexandre A, Abbara A, Fruet C, Loverdo C, Bitbol A-F. Bridging Wright-Fisher and Moran models. J Theor Biol. 2025:599:112030. https://doi.org/10.1016/j.jtbi.2024.112030.

Aqil A, Speidel L, Pavlidis P, Gokcumen O. Balancing selection on genomic deletion polymorphisms in humans. Elife. 2023:12:e79111. https://doi.org/10.7554/eLife.79111.

Auwerx C, Kutalik Z, Reymond A. The pleiotropic spectrum of proximal 16p11.2 CNVs. Am J Hum Genet. 2024a:111(11):2309–2346. https://doi.org/10.1016/j.ajhg.2024.08.015.

Auwerx C, Moix S, Kutalik Z, Reymond A. Disentangling mechanisms behind the pleiotropic effects of proximal 16p11.2 bp4-5 CNVs. Am J Hum Genet. 2024b:111(11):2347–2361. https://doi.org/10.1016/j.ajhg.2024.08.014.

Barrett SCH, Arunkumar R, Wright SI. The demography and population genomics of evolutionary transitions to self-fertilization in plants. Philos Trans R Soc Lond B Biol Sci. 2014:369(1648):20130344. https://doi.org/10.1098/rstb.2013.0344.

Bolognini D, Halgren A, Lou RN, Raveane A, Rocha JL, Guarracino A, Soranzo N, Chin J, Garrison E, Sudmant PH. Global diversity, recurrent evolution, and recent selection on amylase structural haplotypes in humans. bioRxiv. https://doi.org/10.1101/2024.02.07.579378, 2024, preprint: not peer reviewed.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR,

et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 2008:4(5):1–13. https://doi.org/10.1371/journal.pgen.1000083.
Charrier C, Joshi K, Coutinho-Budd J, Kim J-E, Lambert N, de Marchena J, Jin W-L, Vanderhaeghen P, Ghosh A, Sassa T, et al. Inhibition of srgap2 function by its human-specific paralogs induces neoteny during spine maturation. Cell. 2012:149(4):923–935. https://doi.org/10.1016/j.cell.2012.03.034.
Chunduri NK, Barthel K, Storchova Z. Consequences of chromosome loss: why do cells need each chromosome twice? Cells. 2022:11(9):1530. https://doi.org/10.3390/cells11091530.
Conant GC, Wolfe KH. Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet. 2008:9(12):938–950. https://doi.org/10.1038/nrg2482.
Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. Nat Genet. 2006:38(1):75–81. https://doi.org/10.1038/ng1697.
Conrad DF, Hurles ME. The population genetics of structural variation. Nat Genet. 2007:39(S7):S30–S36. https://doi.org/10.1038/ng2042.
Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010:464(7289):704–712. https://doi.org/10.1038/nature08516.
Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, Myers RM, Cavalli-Sforza LL, Feldman MW, Pritchard JK. The role of geography in human adaptation. PLoS Genet. 2009:5(6):e1000500. https://doi.org/10.1371/journal.pgen.1000500.
Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova GN, Huang X, Christiansen L, DeWitt WS, et al. A single-cell atlas of in vivo mammalian chromatin accessibility. Cell. 2018:174(5):1309–1324.e18. https://doi.org/10.1016/j.cell.2018.06.052.
Daugherty AC, Yeo RW, Buenrostro JD, Greenleaf WJ, Kundaje A, Brunet A. Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans. Genome Res. 2017:27(12):2096–2107. https://doi.org/10.1101/gr.226233.117.
De Maio N, Schlötterer C, Kosiol C. Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. Mol Biol Evol. 2013:30(10):2249–2262. https://doi.org/10.1093/molbev/mst131.
Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H, et al. Evolution of human-specific neural srgap2 genes by incomplete segmental duplication. Cell. 2012:149(4):912–922. https://doi.org/10.1016/j.cell.2012.03.033.
Denny SK, Yang D, Chuang C-H, Brady JJ, Lim JS, Grüner BM, Chiou S-H, Schep AN, Baral J, Hamard P-J, et al. Nfib promotes metastasis through a widespread increase in chromatin accessibility. Cell. 2016:166(2):328–342. https://doi.org/10.1016/j.cell.2016.05.052.
Dumas L, Sikela JM. Duf1220 domains, cognitive disease, and human brain evolution. Cold Spring Harb Symp Quant Biol. 2009:74(0):375–382. https://doi.org/10.1101/sqb.2009.74.025.
Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010:28(8):817–825. https://doi.org/10.1038/nbt.1662.
Ernst J, Kellis M. Chromhmm: automating chromatin-state discovery and characterization. Nat Methods. 2012:9(3):215–216. https://doi.org/10.1038/nmeth.1906.
Evans KJ, Huang N, Stempor P, Chesney MA, Down TA, Ahringer J. Stable *Caenorhabditis elegans* chromatin domains separate broadly expressed and developmentally regulated genes. Proc Natl Acad Sci U S A. 2016:113(45):E7020–E7029. https://doi.org/10.1073/pnas.1608162113.
Fang B, Edwards SV. Fitness consequences of structural variation inferred from a house finch pangenome. Proc Natl Acad Sci U S A. 2024:121(47):e2409943121. https://doi.org/10.1073/pnas.2409943121.
Feldkötter M, Schwarzer V, Wirth R, Wienker TF, Wirth B. Quantitative analyses of SMN1 and SMN2 based on real-time lightCycler PCR: fast and highly reliable carrier testing and prediction of severity of spinal muscular atrophy. Am J Hum Genet. 2002:70(2):358–368. https://doi.org/10.1086/338627.
Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006:7(2):85–97. https://doi.org/10.1038/nrg1767.
Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, et al. Copy number variation: new insights in genome diversity. Genome Res. 2006:16(8):949–961. https://doi.org/10.1101/gr.3677206.
Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. Integrative analysis of the *Caenorhabditis elegans* genome by the modencode project. Science. 2010:330(6012):1775–1787. https://doi.org/10.1126/science.1196914.
Girirajan S, Campbell CD, Eichler EE. Human copy number variation and complex genetic disease. Annu Rev Genet. 2011:45(1):203–226. https://doi.org/10.1146/genet.2011.45.issue-1.
Gómez-Robles A, Nicolaou C, Smaers JB, Sherwood CC. The evolution of human altriciality and brain development in comparative context. Nat Ecol Evol. 2024:8(1):133–146. https://doi.org/10.1038/s41559-023-02253-z.
Gonzalez CE, Roberts P, Ostermeier M. Fitness effects of single amino acid insertions and deletions in TEM-1 beta-lactamase. J Mol Biol. 2019:431(12):2320–2330. https://doi.org/10.1016/j.jmb.2019.04.030.
Greer SU, Wright SI, Eckert CG. Population bottleneck associated with but likely preceded the recent evolution of self-fertilization in a coastal dune plant. Evolution. 2022:77(2):454–466. https://doi.org/10.1093/evolut/qpac047.
Hämälä T, Wafula EK, Guiltinan MJ, Ralph PE, de Pamphilis CW, Tiffin P. Genomic structural variants constrain and facilitate adaptation in natural populations of theobroma cacao, the chocolate tree. Proc Natl Acad Sci U S A. 2021:118(35):e2102914118. https://doi.org/10.1073/pnas.2102914118.
Han Z, Sieriebriennikov B, Susoy V, Lo W-S, Igreja C, Dong C, Berasategui A, Witte H, Sommer RJ. Horizontally acquired cellulases assist the expansion of dietary range in pristionchus nematodes. Mol Biol Evol. 2022:39(2):msab370. https://doi.org/10.1093/molbev/msab370.
Hollox EJ, Zuccherato LW, Tucci S. Genome structural variation in human evolution. Trends Genet. 2022:38(1):45–58. https://doi.org/10.1016/j.tig.2021.06.015.
Hujoel MLA, Sherman MA, Barton AR, Mukamel RE, Sankaran VG, Terao C, Loh PR. Influences of rare copy-number variation on human complex traits. Cell. 2022:185(22):4233–4248.e27. https://doi.org/10.1016/j.cell.2022.09.028.
Ivankov DN, Finkelstein AV, Kondrashov FA. A structural perspective of compensatory evolution. Curr Opin Struct Biol. 2014:26:104–112. New constructs and expression of proteins / Sequences and topology. https://doi.org/10.1016/j.sbi.2014.05.004.
Jänes J, Dong Y, Schoof M, Serizay J, Appert A, Cerrato C, Woodbury C, Chen R, Gemma C, Huang N, et al. Chromatin accessibility dynamics across *C. elegans* development and ageing. Elife. 2018:7:e37344. https://doi.org/10.7554/eLife.37344.
Katju V, Bergthorsson U. Copy-number changes in evolution: rates, fitness effects and adaptive significance. Front Genet. 2013:4:273. https://doi.org/10.3389/fgene.2013.00273.

Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet. 2019:20(4):207–220. https://doi.org/10.1038/s41576-018-0089-8.

Kondrashov AS, Sunyaev S, Kondrashov FA. Dobzhansky–Muller incompatibilities in protein evolution. Proc Natl Acad Sci U S A. 2002:99(23):14878–14883. https://doi.org/10.1073/pnas.232565499.

Kondrashov FA. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc Biol Sci. 2012:279(1749):5048–5057. https://doi.org/10.1098/rspb.2012.1108.

Konrad A, Flibotte S, Taylor J, Waterston RH, Moerman DG, Bergthorsson U, Katju V. Mutational and transcriptional landscape of spontaneous gene duplications and deletions in *Caenorhabditis elegans*. Proc Natl Acad Sci U S A. 2018:115(28):7386–7391. https://doi.org/10.1073/pnas.1801930115.

Lee D, Zdraljevic S, Stevens L, Wang Y, Tanny RE, Crombie TA, Cook DE, Webster AK, Chirakar R, Baugh LR, et al. Balancing selection maintains hyper-divergent haplotypes in *Caenorhabditis elegans*. Nat Ecol Evol. 2021:5(6):794–807. https://doi.org/10.1038/s41559-021-01435-x.

Lee Y-L, Bosse M, Mullaart E, Groenen MAM, Veerkamp RF, Bouwman AC. Functional and population genetic features of copy number variations in two dairy cattle populations. BMC Genomics. 2020:21:89.

Lefebvre S, Bürglen L, Reboullet S, Clermont O, Burlet P, Viollet L, Benichou B, Cruaud C, Millasseau P, Zeviani M, et al. Identification and characterization of a spinal muscular atrophy-determining gene. Cell. 1995:80(1):155–165. https://doi.org/10.1016/0092-8674(95)90460-3.

Lenormand T, Fyon F, Sun E, Roze D. Sex chromosome degeneration by regulatory evolution. Curr Biol. 2020:30(15):3001–3006.e5. https://doi.org/10.1016/j.cub.2020.05.052.

Luo Z, Zhang R, Hu T, Zhu Y, Wu Y, Li W, Zhang Z, Yao X, Liang H, Song X. NicE-C efficiently reveals open chromatin-associated chromosome interactions at high resolution. Genome Res. 2022:32(3):534–544. https://doi.org/10.1101/gr.275986.121.

Lye ZN, Purugganan MD. Copy number variation in domestication. Trends Plant Sci. 2019:24(4):352–365. https://doi.org/10.1016/j.tplants.2019.01.003.

Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000:290(5494):1151–1155. https://doi.org/10.1126/science.290.5494.1151.

Mailman MD, Heinz JW, Papp AC, Snyder PJ, Sedra MS, Wirth B, Burghes AH, Prior TW. Molecular analysis of spinal muscular atrophy and modification of the phenotype by SMN2. Genet Med. 2002:4(1):20–26. https://doi.org/10.1097/00125817-200201000-00004.

Malekpour SA, Pezeshk H, Sadeghi M. MSeq-CNV: accurate detection of copy number variation from sequencing of multiple samples. Sci Rep. 2018:8(1):4009. https://doi.org/10.1038/s41598-018-22323-8.

Miyamoto K, Nguyen KT, Allen GE, Jullien J, Kumar D, Otani T, Bradshaw CR, Livesey FJ, Kellis M, Gurdon JB. Chromatin accessibility impacts transcriptional reprogramming in oocytes. Cell Rep. 2018:24(2):304–311. https://doi.org/10.1016/j.celrep.2018.06.030.

Moran P. The statistical processes of evolutionary theory. Clarendon Press; 1962.

Nowak MA. Evolutionary dynamics: exploring the equations of life. Belknap Press; 2006.

Nowak MA, Sasaki A, Taylor C, Fudenberg D. Emergence of cooperation and evolutionary stability in finite populations. Nature. 2004:428(6983):646–650. https://doi.org/10.1038/nature02414.

Ohno S. Evolution by gene duplication. Springer New York; 1970.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. Diet and the evolution of human amylase gene copy number variation. Nat Genet. 2007:39(10):1256–1260. https://doi.org/10.1038/ng2123.

Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, Srivatsan S, Qiu X, Jackson D, Minkina A, et al. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. Mol Cell. 2018:71(5):858–871.e8. https://doi.org/10.1016/j.molcel.2018.06.044.

Pond SLK, Frost SDW, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2004:21(5):676–679. https://doi.org/10.1093/bioinformatics/bti079.

Ponting CP. The functional repertoires of metazoan genomes. Nat Rev Genet. 2008:9(9):689–698. https://doi.org/10.1038/nrg2413.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. Global variation in copy number in the human genome. Nature. 2006:444(7118):444–454. https://doi.org/10.1038/nature05329.

Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, et al. Comparative genomics of the eukaryotes. Science. 2000:287(5461):2204–2215. https://doi.org/10.1126/science.287.5461.2204.

Schrider DR, Hahn MW. Gene copy-number polymorphism in nature. Proc R Soc Lond B Biol Sci. 2010:277(1698):3213–3221. https://doi.org/10.1098/rspb.2010.1180.

Shi X, Radhakrishnan S, Wen J, Chen JY, Chen J, Lam BA, Mills RE, Stranger BE, Lee C, Setlur SR. Association of CNVs with methylation variation. NPJ Genom Med. 2020:5(1):41. https://doi.org/10.1038/s41525-020-00145-w.

Simon-Loriere E, Holmes EC. Gene duplication is infrequent in the recent evolutionary history of RNA viruses. Mol Biol Evol. 2013:30(6):1263–1269. https://doi.org/10.1093/molbev/mst044.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015:526(7571):75–81. https://doi.org/10.1038/nature15394.

Sung W, Ackerman MS, Dillon MM, Platt TG, Fuqua C, Cooper VS, Lynch M. Evolution of the insertion-deletion mutation rate across the tree of life. G3 (Bethesda). 2016:6(8):2583–2591. https://doi.org/10.1534/g3.116.030890.

Szamecz B, Boross G, Kalapis D, Kovács K, Fekete G, Farkas Z, Lázár V, Hrtyan M, Kemmeren P, Groot Koerkamp MJA, et al. The genomic landscape of compensatory evolution. PLoS Biol. 2014:12(8):1–16. https://doi.org/10.1371/journal.pbio.1001935.

Tang Y-C, Amon A. Gene copy-number alterations: a cost-benefit analysis. Cell. 2013:152(3):394–405. https://doi.org/10.1016/j.cell.2012.11.043.

Thibodeau A, Khetan S, Eroglu A, Tewhey R, Stitzel ML, Ucar D. CoRE-ATAC: a deep learning model for the functional classification of regulatory elements from single cell and bulk ATAC-seq data. PLoS Comput Biol. 2021:17(12):1–32. https://doi.org/10.1371/journal.pcbi.1009670.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. The accessible chromatin landscape of the human genome. Nature. 2012:489(7414):75–82. https://doi.org/10.1038/nature11232.

Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, et al. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res. 2008:18(7):1051–1063. https://doi.org/10.1101/gr.076463.108.

Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci U S A. 2005:102(22):7882–7887. https://doi.org/10.1073/pnas.0502300102.

Wong YY, Harbison JE, Hope CM, Gundsambuu B, Brown KA, Wong SW, Brown CY, Couper JJ, Breen J, Liu N, et al. Parallel recovery of chromatin accessibility and gene expression dynamics from frozen human regulatory T cells. Sci Rep. 2023:13(1):5506. https://doi.org/10.1038/s41598-023-32256-6.

Wood WB. The nematode Caenorhabditis elegans. Cold Spring Harbor Laboratory Press; 1988.

Yılmaz F, Karageorgiou C, Kim K, Pajic P, Scheer K, Consortium HGSV, Beck CR, Torregrossa A-M, Lee C, Gokcumen O. Paleolithic gene duplications primed adaptive evolution of human amylase locus upon agriculture. bioRxiv 2023–11. https://doi.org/10.1101/2023.11.27.568916, 2023, preprint: not peer reviewed.

Yoshida H, Lareau CA, Ramirez RN, Rose SA, Maier B, Wroblewska A, Desland F, Chudnovskiy A, Mortha A, Dominguez C, et al. The cis-regulatory atlas of the mouse immune system. Cell. 2019:176(4):897–912.e20. https://doi.org/10.1016/j.cell.2018.12.036.

Zarrei M, Macdonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015:16(3):172–183. https://doi.org/10.1038/nrg3871.

Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 2009:10(1):451–481. PMID: 19715442. https://doi.org/10.1146/genom.2009.10.issue-1.

Associate editor: David Enard