eLife

Genetics and Genomics
Evolutionary Biology

The paradox of extremely fast evolution driven by genetic drift in multi-copy gene systems

Xiaopei Wang, Yongsen Ruan, Lingjie Zhang, Xiangnyu Chen, Zongkun Shi, Haiyu Wang, Bingjie Chen, Miles Tracy, Liying Huang, Chung-I Wu, Haijun Wen

State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China

Reviewed Preprint
v3 • March 6, 2025
Revised by authors

Reviewed Preprint
v2 • November 15, 2024

Reviewed Preprint
v1 • August 20, 2024

eLife Assessment

This study presents a **useful** theoretical model of molecular evolution of multi-copy gene systems by extending the classic Haldane model and applies the model to explain the surprisingly rapid evolution of rRNA genes. Although the conceptual model is intuitive and provides a new perspective for contextualizing this problem, the model presented does not adequately consider plausible biological constraints on the molecular and genetic processes. The lack of such constraints in the model, along with technical issues in the data analysis, provide **incomplete** support for the conclusion that the genetic variation patterns of rRNA genes in mouse is compatible with neutral evolution.

https://doi.org/10.7554/eLife.99992.3.sa3

Abstract

Multi-copy gene systems that evolve within, as well as between, individuals are common. They include viruses, mitochondrial DNAs, multi-gene families etc. The paradox is that neutral evolution in two stages should be far slower than single-copy systems but the opposite is often true, thus leading to the suggestion of natural selection. We now apply the new Generalized Haldane (GH) model to quantify genetic drift in the mammalian ribosomal RNA genes (or rDNAs). On average, rDNAs have $C \sim 150$ - 300 copies. A neutral mutation in rDNA should take $4NC^*$ generations to become fixed (N, the population size; C^*, the effective copy number). While $C > C^* >> 1$ is expected, the observed fixation time in mouse and human is $< 4N$, hence the paradox of $C^* < 1$. Genetic drift thus appears as much as 100 times stronger for rRNA genes as for single-copy genes. The large increases in genetic drift are driven by a host of molecular mechanisms such as gene conversion and unequal crossover. Although each mechanism of drift has been extremely difficult to quantify, the GH model permits the estimation of their total effects on genetic drift. In conclusion, the GH model can be generally applicable to multi-copy gene systems without being burdened by tracking the diverse molecular mechanisms individually.

Introduction

In this study, we focus on multi-copy gene systems, where the evolution takes place in two stages: both within (stage I) and between individuals (stage II). Multi-copy gene systems include viruses, transposons, mitochondria and multi-gene families (Alexandrov, et al. 2001; Szitenberg, et al. 2016; Xu, et al. 2019; Ruan, et al. 2021). Given the extra stage of within-host fixation, the neutral evolutionary rate of multi-copy systems should be much slower than in single-copy systems. However, the rapid evolution of multi-copy systems has been extensively documented (Charlesworth, et al. 1994; Eickbush and Eickbush 2007; Jurka, et al. 2007; Hou, et al. 2023) but the reason is unclear.

The speed of neutral evolution is the basis for determining how fast or slow all types of molecular evolution take place. Neutral evolution is driven by random transmission, gene conversion, stochastic replication etc., which collectively constitute genetic drift. Hence, genetic drift is the fundamental force of molecular evolution. All other evolutionary forces, such as selection, mutation and migration, may be of greater biological interest, but inferences are possible only when genetic drift is fully accounted for. In the companion study, we show that the standard Wright-Fisher (WF) model may often under-account genetic drift, thus leading to the over-estimation of selection.

The companion paper (Ruan, et al. 2024) re-introduces and generalizes the Haldane model to account for all random forces in molecular evolution. This generalized Haldane model is referred to as the GH model. The Haldane (as well as GH) model is based on the branching process. In haploids, each individual produces K progeny with the mean and variance of $E(K)$ and $V(K)$. Genetic drift is primarily $V(K)$ as there would be no drift if $V(K) = 0$. Gene frequency change in the population is then scaled by N (population size), expressed as $V(K)/N$(Chen, et al. 2017). In diploids, K would be the number of progenies to whom the gene copy of interest is transmitted. (The adjustments between haploidy and multi-ploidy are straightforward).

An important element missing in the WF models is the generation of N within the model. All WF models including subsequent modifications have to be supplied N's externally. The Haldane model is capable of generating $N' = N\,E(K)$ where N' is the population size of the next generation. However, the original Haldane (Haldane 1927; Haldane 1932) model reset $N' = N$ thus depriving the branching process an important advantage. The GH model permits $E(K)$ to depend on N and is thus capable of generating and regulating N as well as tracking gene frequencies (see (Ruan, et al. 2024)).

Furthermore, in the WF model, gene frequency is governed by $1/N$ (or $1/2N$ in diploids) because K would follow the Poisson distribution whereby $V(K) = E(K)$. As $E(K)$ is generally ~1 in evolutionary timescale, $V(K)$ would also be ~ 1. In this backdrop, many "modified WF" models have been developed (Der, et al. 2011), most of them permitting $V(K) \neq E(K)$ (Karlin and McGregor 1964; Chia and Watterson 1969; Cannings 1974). Nevertheless, paradoxes encountered by the standard WF model apply to these modified WF models as well, because all WF models share the key feature of being supplied N externally (see below and (Ruan, et al. 2024)). One of the paradoxes, first noted in (Chen, et al. 2017) is genetic drift during tumor growth whereby drift appears to become stronger as N increases (Wu, et al. 2016; Chen, Wu, et al. 2022; Zhai, et al. 2022). This trend is in stark opposite to the central tenet of the WF models.

A paradox requiring dedicated efforts to analyze concerns multi-copy gene systems, which are as diverse as viral epidemics (Huang, et al. 2021), transposons (Szitenberg, et al. 2016), mitochondrial DNAs (Xu, et al. 2019), satellite DNAs (Cabot, et al. 1993; Alexandrov, et al. 2001), and ribosomal RNA genes (van Sluis and McStay 2019; Hori, et al. 2021). In COVID-19,

the inability of the WF models to track both within- and between-host evolution simultaneously is a main reason for much confusion about the origin, spread and driving forces of SARS-CoV-2 (Ruan, et al. 2021; Deng, et al. 2022; Guan and Zhong 2022; Pan, Liu, et al. 2022; Ruan, Wen, et al. 2022; Zhou, et al. 2022; Hou, et al. 2023).

We now apply the GH model to multi-copy gene systems, using ribosomal RNA genes (rDNAs) as an example. In such systems, the molecular mechanisms driving neutral evolution within individuals are diverse. It has been difficult to model each mechanism individually, let alone in combination for rDNAs (Nagylaki 1983; Ohta and Dover 1983; Ohta and Dover 1984), viruses (Ruan, et al. 2021; Ruan, Hou, et al. 2022) and mitochondria DNA (Rand 2001; Xu, et al. 2019). Since our interest is to filter out the total effects of genetic drift in order to study the more interesting forces such as selection, mutation etc, the GH model is applied for this purpose. The details of each molecular mechanism will be missed but genetic drift, unlike selection, is really noises in biological evolution.

While the WF models have led to speculations of pervasive natural selection in multi-copy gene systems (Dover 1982; Arkhipova 2018; Chen, Yang, et al. 2022; Pan, Zhang, et al. 2022; Wang, et al. 2022), the GH model may reverse the view. For rDNAs, the GH model will show that neutrality is often a simpler explanation. If and when the model rejects the neutrality explanation, the rejection would also lead to the more convincing proof of selection,

Results

PART I presents a best-known multi-gene system of the rRNA genes. PART II (Theory) consolidates aspects of the Haldane model applied to polymorphism within species as well as divergence between species. In PART III (Data analyses), we apply the theory to rDNA evolution in mice and apes (human and chimpanzee).

PART I - The biology of rRNA gene clusters

The ribosomal RNA genes (rDNAs) are multi-copy gene clusters (Bowman, et al. 2020) that are arrayed as tandem repeats on multiple chromosomes as shown in **Fig. 1A** (Guillén, et al. 2004; Cazaux, et al. 2011). In humans, the copy number can vary from 60 to 1600 per individual (mean, 315; SD, 104; median, 301) (Parks, et al. 2018). For each haploid genome, $C \sim 150$ on average in humans and $C \sim 110$ in mice (Parks, et al. 2018). In humans, the five rRNA clusters are located on the short arm of the five acrocentric chromosomes (Smirnov, et al. 2021). Such an arrangement permits crossovers between chromosomes without perturbing the rest of the genomes. In *Mus*, the rDNAs are all located in the pericentromeric or sub-telomeric region, on the long arms of telocentric chromosomes (Cazaux, et al. 2011; Potapova and Gerton 2019). Thus, unequal crossovers between non-homologous chromosomes may involve centromeres while other genic regions are also minimally perturbed.

Each copy of rRNA gene has a functional and non-functional part as shown in **Fig. 1B**. The "functional" regions of rDNA, 18S, 5.8S, and 28S rDNA, are believed to be under strong negative selection, resulting in a slow evolution rate in animals (Salim and Gerton 2019). In contrast, the transcribed spacer (ETS and ITS) and the intergenic spacer (IGS) show significant sequence variation even among closely related species (Eickbush and Eickbush 2007). Clearly, these "non-functional" sequences are less constrained by negative selection. In this study of genetic drift, we focus on the non-functional parts. Data on the evolution of the functional parts will be provided only for comparisons.

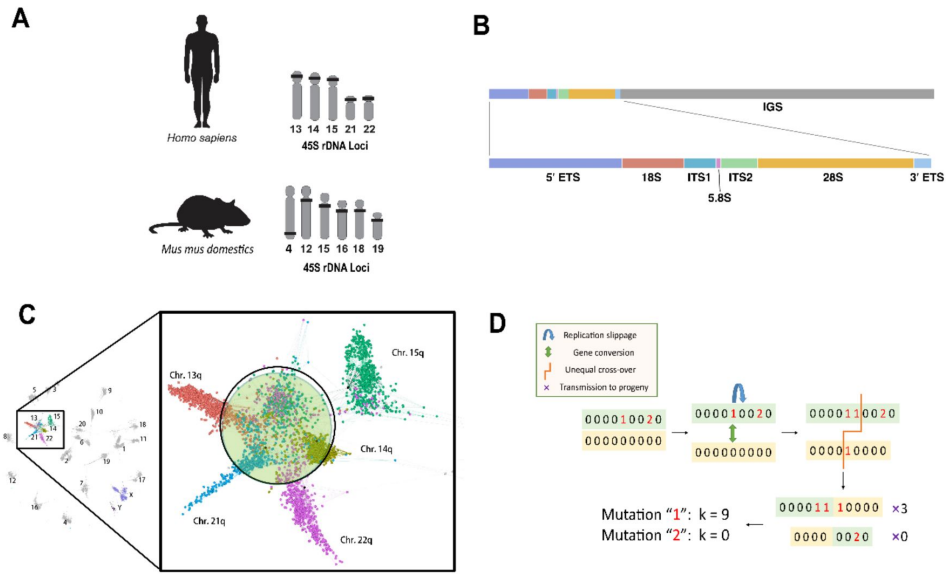


Figure 1.

The "chromosome community of rDNAs on five acrocentric chromosomes.

(A) The genomic locations of rDNA tandem repeats in human (Gibbons, et al. 2015) and mouse (Cazaux, et al. 2011). rDNAs are located on the short arms (human), or the proximal end of the long arms (mouse), of the chromosome. Either way, inter-chromosomal exchanges are permissible. (B) The organization of rDNA repeat unit. IGS (intergenic spacer) is not transcribed. Among the transcribed regions, 18S, 5.8S and 28S segments are in the mature rRNA while ETS (external transcribed spacer) and ITS (internal transcribed spacer) are excluded. (C) The pseudo-population of rRNA genes is shown by the "chromosomes community" map (Guarracino, et al. 2023), which indicates the divergence distance among chromosome segments. The large circle encompasses rDNAs from all 5 chromosomes. It shows the concerted evolution among rRNA genes from all chromosomes, which thus resemble members of a (pseudo-)population. The slightly smaller thin circle, from the analysis of this study, shows that the rDNA gene pool from each individual captures approximately 95% of the total diversity of human population. (D) A simple illustration that shows the transmissions of two new mutations (#1 and #2 in red letter). Mutation 1 experiences replication slippage, gene conversion and unequal crossover and grows to 9 copies (K = 9) after transmission. Mutation 2 emerges and disappears (K = 0). This shows how $V(K)$ may be augmented by the homogenization process.

The pseudo-population of ribosomal DNA copies within each individual

While a human haploid with 200 rRNA genes may appear to have 200 loci, the concept of "gene loci" cannot be applied to the rRNA gene clusters. This is because DNA sequences can spread from one copy to others on the same chromosome via replication slippage. They can also spread among copies on different chromosomes via gene conversion and unequal crossovers (Nagylaki 1983; Ohta and Dover 1983; Stults, et al. 2008; Smirnov, et al. 2021). Replication slippage and unequal crossovers would also alter the copy number of rRNA genes. These mechanisms will be referred to collectively as the homogenization process. Copies of the cluster on the same chromosome are known to be nearly identical in sequences (Hori, et al. 2021; Nurk, et al. 2022). Previous research has also provided extensive evidence for genetic exchanges between chromosomes (Krystal, et al. 1981; Arnheim, et al. 1982; van Sluis, et al. 2019).

In short, rRNA gene copies in an individual can be treated as a pseudo-population of gene copies. Such a pseudo-population is not Mendelian, but its genetic drift can be analyzed using the branching process (see below). The pseudo-population corresponds to the "chromosome community" proposed recently (Guarracino, et al. 2023). As seen in **Fig. 1C**, the five short arms harbor a shared pool of rRNA genes that can be exchanged among them. **Fig. 1D** presents the possible molecular mechanisms of genetic drift within individuals whereby mutations may spread, segregate or disappear among copies. Hence, rRNA gene diversity or polymorphism refers to the variation across all rRNA copies, as these genes exist as paralogs rather than orthologs. This diversity can be assessed at both individual and population levels according to the multi-copy nature of rRNA genes.

PART II - Theory

1. The Haldane model of genetics drift applied to multi-copy gene systems

The Haldane model of genetic drift based on the branching process is intuitively appealing. In the model, each copy of the gene leaves K copies in a time interval with the mean and variance of $E(K)$ and $V(K)$. If $V(K) = 0$, there is no gene frequency change and no genetic drift. In the standard WF model, $V(K) = E(K)$ whereas $V(K)$ is decoupled from $E(K)$ in the Haldane model; the latter thus being more flexible. [In the companion paper, we discuss the modified WF models with $V(K) \neq E(K)$ which, nevertheless, do not resolve the paradoxes generated from population sampling.]

Below, we compare the strength of genetic drift in rRNA genes vs. that of single-copy genes using the Haldane model (Ruan, et al. 2024). We shall use * to designate the equivalent symbols for rRNA genes; for example, $E(K)$ vs. $E^*(K)$. Both are set to 1, such that the total number of copies in the long run remains constant.

For simplicity, we let $V(K) = 1$ for single-copy genes. (If we permit $V(K) \neq 1$, the analyses will involve the ratio of $V^*(K)$ and $V(K)$ to reach the same conclusion but with unnecessary complexities.) For rRNA genes, $V^*(K) \geq 1$ may generally be true because K for rDNA mutations are affected by a host of homogenization factors including replication slippage, unequal cross-over, gene conversion and other related mechanisms not operating on single-copy genes. Hence,

$$N^* = \frac{NC}{V^*(K)} = N\left[\frac{C}{V^*(K)}\right] = NC^* \qquad Eq.(1)$$

where N^* and N are the effective population size of rDNAs and single-copy genes, C is the average number of rRNA genes in an individual, and $V^*(K)$ reflects the homogenization process on rRNA genes (**Fig. 1D**). Thus,

$$C^* = C/V^*(K)$$

represents the effective copy number of rRNA genes in the population, determining the level of genetic diversity relative to single-copy genes. Since C is in the hundreds and $V^*(K)$ is expected to be > 1, the relationship of $1 << C^* \leq C$ is hypothesized. **Fig. 1D** is a simple illustration that the homogenizing process may enhance $V^*(K)$ substantially over the WF model.

In short, genetic drift of rRNA genes would be equivalent to single-copy genes in a population of size NC^*(or N^*). Since $C^* >> 1$ is hypothesized, genetic drift for rRNA genes is expected to be slower than for single-copy genes.

2. rDNA polymorphism within species

A standard measure of genetic drift is the level of heterozygosity (H). At the mutation-selection equilibrium

$$H_{equi} = \frac{2N_e\mu}{2N_e\mu + 1}$$

where μ is the mutation rate of the entire gene and N_e is the effective population size. In this study, $N_e = N$ for single-copy gene and $N_e = C^*N$ for rRNA genes. The empirical measure of nucleotide diversity H is given by

$$H = \frac{\sum_{i=1}^{L} 2p_i(1-p_i)}{L} \qquad Eq.(2)$$

where L is the gene length (for each copy of rRNA gene, $L \sim$ 43kb) and p_i is the variant frequency at the i-th site.

We calculate H of rRNA genes at three levels – within-individual, within-species and then, within total samples (*Hi*, *Hs* and *Ht*, respectively). *Hs* and *Ht* are standard population genetic measures (Crow and Kimura 1970; Hartl, et al. 1997). In calculating *Hs*, all sequences in the species are used, regardless of the source individuals. A similar procedure is applied to *Ht*. The *Hi* statistic is adopted for multi-copy gene systems for measuring within-individual polymorphism. Note that copies within each individual are treated as a pseudo-population (see **Fig. 1** and text above). With multiple individuals, *Hi* is averaged over them.

Given the three levels of heterozygosity, there are two levels of differentiation. First, F_{IS} is the differentiation among individuals within the species, defined by

$$F_{IS} = [H_S - H_I]/H_S$$

F_{IS} is hence the proportion of genetic diversity in the species that is found only between individuals. We will later show $F_{IS} \sim 0.05$ in human rDNA (**Table 2**), meaning 95% of rDNA diversity is found within individuals.

Second, F_{ST} is the differentiation between species within the total species complex, defined as

$$F_{ST} = [H_T - H_S]/H_T$$

F_{ST} is the proportion of genetic diversity in the total data that is found only between species. Between mouse species, F_{ST} distribution is close to 1 (**Fig. 4C**), indicating a large genetic distance between species relative to within-species polymorphisms.

3. rDNA divergence between species

Whereas the level of genetic diversity is a function of the effective population size, the rate of divergence between species, in its basic form, is not. The rate of neutral molecular evolution (λ), although driven by mutation and genetic drift, is generally shown by **Eq. (3)** (Crow and Kimura 1970; Hartl, et al. 1997; Li 1997):

$$\lambda = N\mu\frac{1}{N} = \mu \qquad Eq.(3)$$

Note that the factor of $1/N$ in Eq. (3) indicates the fixation probability of a new mutation. For rDNA mutations, fixation must occur in two stages – fixation within individuals and then among individuals in the population. (We note again that new mutations can be fixed via homogenization in an individual, effectively forming a pseudo-population for rRNA genes.) Due to the cancelation of N in Eq. (3), the evolutionary rate of rRNA genes in the long run should be the same as single-copy genes.

Eq. (3) is valid in the long-term evolution. For shorter-term evolution, it is necessary to factor in the fixation time (**Fig. 2**), T_f, which is the time between the emergence of a mutation and its fixation. If we study two species with a divergent time (T_d) equal to or smaller than T_f, then few mutations of recent emergence would have been fixed as species diverge.

Note that T_d is about 6 million years (Myrs) between human and chimpanzee while T_f (as measured by coalescence) is 0.6 - 1 Myrs in humans. Mutations of single-copy genes would not get fixed during the more recent T_f, as indicated in **Figure 2**. Thus, the realized substitution rate may be 1/6 to 1/10 lower than the theoretical value. In comparison, T_f^* of rRNA mutations should be at least 4.8 - 8 Myrs based on the C^* estimate (see PATR III). Thus, the substitution rate could be at least 80% lower than calculated for single-copy genes.

Our own theoretical derivation has shown that the fixation time for rRNA genes is close to $4N^*$ as is the case for single-copy genes at $\sim 4N$. If $V^*(K)$ is sufficiently larger than $V(K)$, it is in fact possible for $N^* < N$ such that genetic drift is stronger for rRNA genes than for single-copy genes and $T_f^* < T_f$. Therefore, T_{f1} can represent the T_f^* of mutations in rRNA genes in **Fig. 2**. This is interesting because, if the homogenization is powerful enough, rRNA genes would have an effective copy number of $C^* < 1$. A short T_f^*, which can be obtained by large $V^*(K)$, would lead to a smaller T_f^*/T_d than T_f/T_d, and thus a higher substitution rate than single-copy genes, particularly in short-term evolution. However, even T_f^* approaches 0, the substitution rate can exceed that of single-copy genes but is still limited by fixation probability. As noted above, the rapidly fixed mutations may not be well represented in the polymorphic data but can accumulate in species divergence.

PART III - Data Analyses

Before presenting the main analyses, we provide some empirical observations on the rapid homogenization of rRNA genes within individuals (Stage I). These observations are needed for PART III that comprises i) the analyses of rDNA polymorphisms within species in mouse and human; and ii) the analysis of the divergence between species.

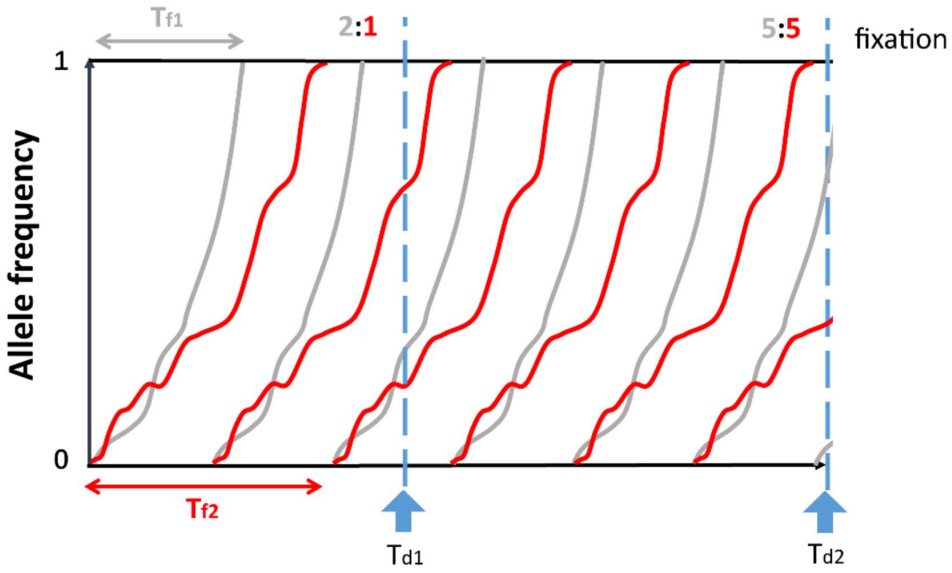


Figure 2.

Fixation of mutations at two levels of species divergence, (T_{d1}) and (T_{d2}).

(T_{f1}) and (T_{f2}) are mutations with a shorter and longer fixation time, respectively for single-copy and multi-copy genes. Note that mutations with a longer T_f would show a lower fixation rate in short-term evolution.

Empirical measurements of homogenization within cells

In an accompanying study (Wang, et al. unpublished data), the evolutionary rate of neutral rRNA variants within cells is measured. Here, genetic drift operates via the homogenizing mechanisms that include gene conversion, unequal crossover and replication slippage. In the literature, measurements of neutral rRNA evolution are usually based on comparisons among individuals. Therefore, the Mendelian mechanisms of chromosome segregation and assortment would also shuffle variants among individuals. Segregation and assortment would confound the measurements of homogenization effects within individuals.

In one experiment, the homogenization effects in rDNAs are measured in cultured cell lines over 6 months of evolution. For in vivo homogenization, we analyze the evolution of rRNA genes within solid tumors. We estimate the rate at which rRNA variants spread among copies within the same cells, which undergo an asexual process. The measurements suggest that, in the absence of recombination and chromosome assortment, the fixation time of new rRNA mutations within cells would take only 1 - 3 kyrs (thousand years). Since a new mutation in single-copy genes would take 300 - 600 kyrs to be fixed in human populations, the speed of genetic drift in Stage I evolution is orders faster than in Stage II. Therefore, despite having several hundred copies of rRNA genes per genome, the speed of genetic drift in rRNA genes may not be that much slower than in single-copy genes. This postulate will be tested below.

1. rDNA polymorphism within species

1) Polymorphism in mice

For rRNA genes, H_I of 10 individuals ranges from 0.0056 to 0.0067 while H_S is 0.0073 (**Table 1**). Thus, $F_{IS} = [H_S - H_I]/H_S$ for mice is 0.14, which means 86% of variation is within each individual. In other words, even one single randomly chosen individual would yield 85% of the diversity of the whole species. Hence, the estimated H_S should be robust as it is not affected much by the sampling.

H_S for *M. m. domesticus* single-copy genes is roughly 1.40 per kb genome-wide (Geraldes, et al. 2008) while H_S for rRNA genes is 7.25 per kb (**Table 1**), 5.2 times larger. In other words, $C^* = N^*/N \sim 5.2$. If we use the polymorphism data, it is as if rDNA array has a population size 5.2 times larger than single-copy genes. Although the actual copy number on each haploid is ~ 110, these copies do not segregate like single-copy genes and we should not expect N^* to be 100 times larger than N. The H_S results confirm the prediction that rRNA genes should be more polymorphic than single-copy genes.

Based on the polymorphism result, one might infer slower drift for rDNAs than for single-copy genes. However, the results from the divergence data in later sections will reveal the under-estimation of drift strength from polymorphism data. Such data would miss variants that have a fast drift process driven by, for example, gene conversion. Strength of genetic drift should therefore be measured by the long-term fixation rate.

2) Polymorphism in human

F_{IS} for rDNA among 8 human individuals is 0.059 (**Table 2**), much smaller than 0.142 in *M. m. domesticus* mice, indicating minimal genetic differences across human individuals and high level of genetic identity in rDNAs between homologous chromosomes among human population. Consisted with low F_{IS}, **Fig. 3** shows strong correlation of the polymorphic site frequency of rDNA transcribed region among each pair of individuals from three continents (2 Asians, 2 Europeans and 2 Africans). Correlation of polymorphic sites in IGS region is shown in Supplementary Fig. 1. The results suggest that the genetic variation due to the sampling of

Mouse strains	H_I (‰) - diversity within individuals	
	Functional parts (18S 5.8S and 28S)	**Non-functional parts (ETS, ITS and IGS)**
WSB/EiJ	0.36	6.45
ZALENDE/Ei	0.44	6.29
LEWES/EiJ	0.28	5.56
BALB/cJ	0.19	6.51
C57BL/6NJ	0.20	6.30
ST/bJ	0.24	6.73
NZW/LacJ	0.22	4.96
FVB/NJ	0.35	6.47
DBA/1J	0.22	6.44
C3H/HeJ	0.20	6.50
Averaged H_I diversity across individuals	0.27	**6.22**
H_S for rDNA in *M. m. domesticus*	0.34	7.25
H_S for single-copy genes in *M. m. domesticus*	---	1.40*

WSB/EiJ, ZALENDE/Ei and LEWES/EiJ are outbred strains, and the rest are inbred strains.
Sources: Wellcome Sanger Institute's Mouse Genome Project (MGP).
*Data from (Geraldes, et al. 2008).

Table 1.

rRNA gene nucleotide diversity in the 10 *M. m. domesticus* strains of a global collection.

chromosomes during sexual reproduction (e.g., segregation and assortment) is substantially attenuated by the effects of homogenization process within individual. Like those in mice, the pattern indicates that intra-species polymorphism is mainly preserved within individuals. The observed H_I of humans for rDNAs is 0.0064 to 0.0077 and the H_S is 0.0072 (**Table 2**). Research has shown that heterozygosity for the human genome is about 0.00088 (Zhao, et al. 2000), meaning the effective copy number of rDNAs is roughly, or $C^* \sim 8$. This reduction in effective copy number from 150 to 8 indicates strong genetic drift due to homogenization force.

2. rDNA divergence between species

We now consider the evolution of rRNA genes between species by analyzing the rate of fixation (or near fixation) of mutations. Polymorphic variants are filtered out in the calculation. Note that Eq. (3) shows that the mutation rate, μ, determines the long-term evolutionary rate, α. Since we will compare the α values between rDNA and single-copy genes, we first need to compare their mutation rates by analyzing their long-term divergence, which is hardly influenced by fixation time. As shown in Table S1, α falls in the range of 50-60 (differences per Kb) for single-copy genes and 40 – 70 for the non-functional parts of rRNA genes.

The data thus suggest that rRNA and single-copy genes are comparable in mutation rate. Differences between their α values will have to be explained by other means.

1) Between mouse species - Genetic drift as the sole driving force of the rapid divergence

We now use the F_{ST} statistic to delineate fixation and polymorphism. The polymorphism in *M. m. domesticus* is compared with two outgroup species, *M. spretus* and *M. m. castaneus*, respectively. There are hence two depths in the phylogeny with two T_d's, as shown in **Fig. 4A** (Rudra, et al. 2016; Kumar, et al. 2022). There is a fourth species, *M. m. musculus* (shown in grey in **Fig. 4A**), which yields very similar results as *M. m. domesticus* in these two comparisons. These additional analyses are shown in Supplement Table S2-S3.

The F_{IS} values of polymorphic sites in 3 outbred mice are primarily below 0.2 and rarely above 0.8 in **Fig. 4B**, indicating the low genetic differentiation in rDNAs within these 3 *M. m. domesticus*. While the F_{IS} distribution of 10 mice from **Table 1**, including 7 inbred and 3 outbred mouse strains, exhibits a noticeable right skewness, but does not exceed 0.8. This suggests that inbreeding to a certain extent limits the process of homogenization and enhances population differentiation. In comparison, the distribution of the F_{ST} of variant sites between *M. m. domesticus* and *Mus spretus* has a large peak near $F_{ST} = 1$. This peak in **Fig. 4C** represents species divergence not seen within populations (i.e., F_{IS}). We use $F_{ST} = 0.8$ as a cutoff for divergence sites between the two species. Roughly, when a mutant is > 0.95 in frequency in one species and < 0.05 in the other, F_{ST} would be > 0.80.

We first compare the divergence between *M. m. domesticus* and *M. m. castaneus* whereby T_d has been estimated to be less than 0.5 Myrs (Fujiwara, et al. 2022). In comparison, between *Mus m. domesticus* and *Mus spretus*, T_d is close to 3 Myrs (Rudra, et al. 2016). As noted above, the reduction in the divergence rate relative to that of Eq. (3) is proportional to T_f/T_d (for single copy genes) or T_f^*/T_d (for rRNA genes). As T_f and T_f^* are both from *M. m. domesticus* and T_d is 6 times larger in comparison with *M. spretus*, we expect the results to be quite different between the two sets of species comparisons.

Although T_f and T_f^* estimates are less reliable than T_d estimates, both comparisons use the same T_f and T_f^* from *M. m. domesticus*. Hence, the results should be qualitatively unbiased. For a demonstration, we shall use the estimates of T_f (i.e., the coalescence time) at 0.2 Myrs for single-copy genes by using an average nucleotide diversity of 0.0014 and the mutation rate of 5.7×10^{-9} per

Human Individuals	H_I (‰) - diversity within individuals	
	Functional parts (18S 5.8S and 28S)	Non-functional parts (ETS, ITS and IGS)
Asia CRC	0.53	7.69
Asia 1590	0.28	7.33
Europe 4929	0.42	6.53
Europe 1383	0.32	6.74
Africa 7733	0.36	6.65
Africa 9161	0.34	6.55
Asia F9551	0.87	6.37
Asia M9552	0.88	6.58
Averaged diversity H_I across individuals	0.50	6.81
H_S for rDNA in human population	0.68	7.24
H_S for single-copy genes in human population	---	0.88*

8 humans are from three continents (4 Asians, 2 Europeans and 2 Africans).
Source: National Center for Biotechnology Information (NCBI).
* Data from (Zhao, et al. 2000).

Table 2.

rRNA gene nucleotide diversity in the 8 humans of a global collection.

Figure 3.

Correlation of variant frequencies between human individuals.

The pairwise correlation of variant site frequency in the transcribed region of rDNAs among 6 individuals (2 Asians, 2 Europeans, and 2 Africans). The high correlations suggest that the diversity in each individual can well capture the population diversity. Each color represents a region of rDNA. The diagonal plots present the variant frequency distribution. The upper right section summarizes the Pearson correlation coefficients derived from the mirror symmetric plots in the bottom left. The analysis excluded the 18S, 3'ETS, and 5.8S regions due to the limited polymorphic sites. The result for IGS region is presented in Supplementary Figure S1.

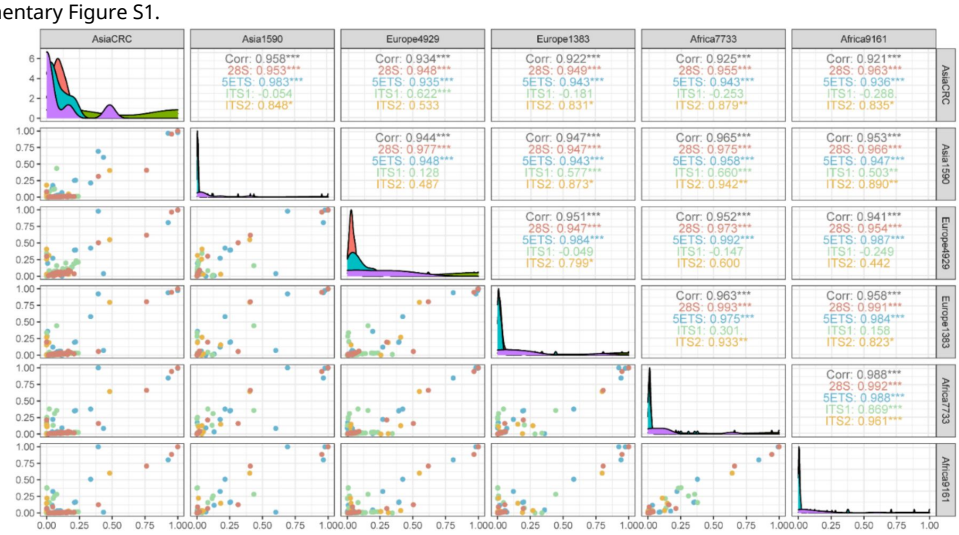
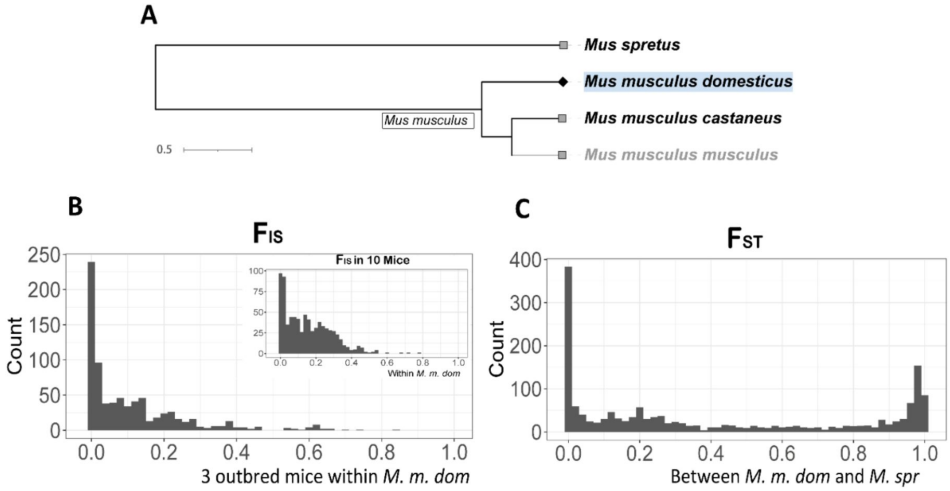


Figure 4.

Levels of polymorphism and divergence in mice.

(A) Phylogeny of *Mus musculus* and *Mus spretus* mice. The divergence times are obtained from *http://timetree.org/*. The line segment labeled 0.5 represents 0.5 Myrs. (B) F_{IS} distribution within *M. m. domesticus*. The distribution of F_{IS} for polymorphic sites in 3 outbred mouse strains or 10 mouse strains (including 7 inbred mice) in **Table 1** (Inset) is shown. (C) F_{ST} distribution between *M. m. domesticus* and *Mus spretus*. Note that the *F* values rise above 0.8 only in (C).



base pair per generation (Geraldes, et al. 2008; Phifer-Rixey, et al. 2020). Based on the estimated C^* above, we obtain T_f^* for rDNA mutations at 5×0.2 Myrs, or 1 Myrs. While some have estimated T_f to be close to 0.4 Myrs (Fujiwara, et al. 2022), we aim to show that the pattern of reduction in rRNA divergence is true even with the conservative estimates.

Between *M. m. domesticus* and *M. m. castaneus*, the reduction in substitution rate for single copy gene should be ~ 40% (T_f/T_d = 0.2/0.5), and the reduction for rRNA genes should be 100% (T_f^*/T_d = 1/0.5 > 1). **Table 3** on their DNA sequence divergence shows that rRNA genes are indeed far less divergent than single-copy genes. In fact, only a small fraction of rDNA mutations is expected to be fixed as T_f^* for rDNA at 1 Myrs is 2 times larger than the divergence time, T_d. We should note again that the non-negligible fixation of rRNA mutations suggests that C^* at 5 is perhaps an over-estimate.

Between *Mus m. domesticus* and *Mus spretus*, the reduction in actual substitution rate from theoretical limit for single-copy genes should be 6.7% (T_f/T_d = 0.2/3) and, for rRNA genes, should be 33% (T_f^*/T_d = 1/3). The evolutionary rate (i.e. the fixation rate) of IGS region is lower than single-copy genes, 0.01 in IGS and 0.021 in genome-wide (**Table 4**), as one would expect. However, ETS and ITS regions have evolved at a surprising rate that is 12% higher than single-copy genes. Note that the reduction in C^*, even to the lowest limit of C^* =1, would only elevate the rate of fixation in rRNA genes to a parity with single-copy genes. From **Eq. (1)**, the explanation would be that $V^*(K)$ has to be very large, such that C^* is < 1. With such rapid homogenization, the fixation time approaches 0 and the substitution rate in rRNA genes can indeed reach the theoretical limit of **Eq. (3)**. In such a scenario, the substitution rate in ETS and ITS, compared to single-copy genes in mice, may increase by 7%, $T_f/(T_d - T_f)$ = 0.2/(3-0.2). If we use T_f ~ 0.4 Myrs in an alternative estimation, the increase can be up to 15%.

In conclusion, the high rate of fixation in ETS and ITS may be due to very frequent gene conversions that reduce C^* to be less than 1. In contrast, IGS may have undergone fewer gene conversions and its long-term C^* is slightly larger than 1. Indeed, the heterozygosity in IGS region, at about 2-fold higher than that of ETS and ITS regions (8‰ for IGS, 5‰ for ETS and 3‰ for ITS), supports this interpretation.

2) Between Human and Chimpanzee - Positive selection in addition to rapid drift in rDNA divergence

Like the data of mouse studies, the polymorphism of rDNAs in humans would suggest a slower short-term evolution rate. The same caveat is that C^* estimated from the polymorphism data would have missed those rapidly fixed variants. Hence, the long-term C^* obtained from species divergence might be much smaller than 8.

Our results show that the evolutionary rate of rRNA genes between human and chimpanzee is substantially higher than that of other single-copy genes (**Table 5**). Especially, 5'ETS region shows a 100% rate acceleration, at 22.7‰ vs. 11‰ genome-wide. Even after removing CpG sites, their fixation rate still reaches 22.4‰. In this case, even if C^* <<1, the extremely rapid fixation will only increase the substitution rate by $T_f/(T_d - T_f)$ by 11%, compared to single-copy genes. Thus, the much accelerated evolution of rRNA genes between humans and chimpanzees cannot be entirely attributed to genetic drift. In the next and last section, we will test if selection is operating on rRNA genes by examining the pattern of gene conversion.

3) Positive selection for rRNA mutations in apes, but not in mice – Evidence from gene conversion patterns

For gene conversion, we examine the patterns of AT-to-GC vs. GC-to-AT changes. While it has been reported that gene conversion would favor AT-to-GC over GC-to-AT conversion (Jeffreys and Neumann 2002; Meunier and Duret 2004) at the site level, we are interested at the gene level

Table 3.

Divergence in rRNA genes between *M. m. domesticus* and *M. m. castaneus*.

	Length	mapping length (L) (w/o CpG sites)	Divergent site (D) (w/o CpG sites)	D/(L/1000) (w/o CpG sites)
Functional parts				
18S+5.8S + 28S	6757	6755 (5321)	4(4)	0.6(0.8)
Non-Functional parts				
5' ETS	4007	4006 (3165)	9(7)	2.2(2.2)
ITS1+ITS2	2088	2088 (1530)	11(7)	5.3(4.6)
3'ETS	551	539 (371)	1(1)	1.9(2.7)
IGS	31902	26101 (25461)	45(36)	1.7(1.4)
Genome-wide				
Single-copy genes	-	-	-	9.5

Note the larger divergence of single-copy genes (9.5) than for rRNA genes (1.7 - 5.3)

Table 4.

Divergence in rRNA genes between *M. m. domesticus* and *Mus spretus*.

	Length	mapping length L (w/o CpG sites)	Divergent site D (w/o CpG sites)	D/(L/1000) (w/o CpG sites)
Functional parts				
18S+5.8S + 28S	6757	6757(5321)	12(10)	1.8(1.9)
Non-Functional parts				
5' ETS	4007	4005(3165)	100(68)	25.0(21.5)
ITS1+ITS2	2088	2077(1530)	54(38)	26.0(24.8)
3'ETS	551	551(371)	17(13)	30.9(35.0)
IGS	31902	25516(24796)	270(241)	10.6(9.7)
Genome-wide				
Single-copy genes	-	-	-	21

Note the divergence of single-copy genes (21) is often smaller than rRNA genes (25.0 - 30.9)
The divergence between *M. m. musculus* and the two outgroups species of *M. m. castaneus* and *Mus spretus* are listed in Supplementary Table S2 – S3.

	Length	mapping length L* (w/o CpG sites)	Divergent site D† (w/o CpG sites)	D/(L/1000) ‡ (w/o CpG sites)
Functional parts				
18S+5.8S + 28S	7063	7036(5496)	13±0.53 (10±0.35)	1.85±0.08 (1.81±0.06)
Non-Functional parts				
5' ETS	3656	3608(2446)	82±1.07 (55±0.53)	22.73±0.30 (22.4±0.22)
ITS1+ITS2	2250	2234(1430)	24±3.34 (12+2.31)	10.74±1.49 (8.39±1.62)
3'ETS	345	345(215)	7±0.53 (7±0.53)	20.29±1.55 (32.56±2.47)
IGS	29685	29452(26650)	544±11.80 (410±8.01)	18.47±0.40 (15.38±0.30)
Genome-wide				
Single-copy genes	3.2billion	-	35million	11(Varki and Altheide 2005)

Note the divergence of single-copy genes (= 11) is usually smaller than rRNA genes.
Human data are from Table 2 and chimpanzee genome sequencing data are from (Tatsumoto, et al. 2017).
* The mapping length between two species; † Divergent sites ($F_{ST} > 0.8$) are shown as the mean ± SD; ‡ Divergent sites per kilobase are shown as the mean ± SD. SD, standard deviation.

Table 5.

Divergence in rRNA genes between Human and Chimpanzee.

by summing up all conversions across sites. We designate the proportion of AT-to-GC conversion as *f* and the reciprocal, GC-to-AT, as *g*. Both *f* and *g* represent the proportion of fixed mutations between species (see Methods). So defined, *f* and *g* are influenced by the molecular mechanisms as well as natural selection. The latter may favor a higher or lower GC ratio at the genic level between species. As the selective pressure is distributed over the length of the gene, each site may experience rather weak pressure.

Let *p* be the proportion of AT sites and *q* be the proportion of GC sites in the gene. The flux of AT-to-GC would be *pf* and the flux in reverse, GC-to-AT, would be *qg*. At equilibrium, *pf* = *qg*. Given *f* and *g*, the ratio of *p* and *q* would eventually reach $p/q = g/f$. We now determine if the fluxes are in equilibrium ($pf = qg$). If they are not, the genic GC ratio is likely under selection and is moving to a different equilibrium.

In these genic analyses, we first analyzed the human lineage (Brown and Jiricny 1989; Galtier and Duret 2007). Using chimpanzees and gorillas as the outgroups, we identified the derived variants that became nearly fixed in humans with frequency > 0.8 (**Table 6**). The chi-square test shows that the GC variants had a significantly higher fixation probability compared to AT. In addition, this pattern is also found in chimpanzees ($p < 0.001$). In *M. m. domesticus* (**Table 6**), the chi-square test reveals no difference in the fixation probability between GC and AT ($p = 0.957$). Further details can be found in Supplementary Figure 2. Overall, a higher fixation probability of the GC variants is found in human and chimpanzee, whereas this bias is not observed in mice.

Based on **Table 6**, we could calculate the value of *p*, *q*, *f* and *g* (see **Table 7**). Shown in the last row of **Table 7**, the $(pf)/(qg)$ ratio is much larger than 1 in both the human and chimpanzee lineages. Notably, the ratio in mouse is not significantly different from 1. Combining **Tables 4** and **7**, we conclude that the slight acceleration of fixation in mice can be accounted for by genetic drift, due to gene conversion among rRNA gene copies. In contrast, the different fluxes corroborate the interpretations of **Table 5** that selection is operating in both humans and chimpanzees.

Discussion

The Haldane model is an “individual-output” model of genetic drift (Chen, et al. 2017). Hence, it does not demand the population to follow the rules of WF populations. Indeed, it would be difficult to define WF populations in any multi-copy gene systems. The GH model is also sufficiently flexible for studying various stochastic forces other than the sampling errors that together drive genetic drift. In the companion study (Ruan, et al. 2024), we address the ecological forces of genetic drift and, in this study, we analyze the neutral evolution of rRNA genes. Both examples are amenable to the analysis by the Haldane model, but not by the WF model.

In multi-copy systems, there are several mechanisms of homogenization within individuals. For rRNA genes, whether on the same or different chromosomes (Gonzalez and Sylvester 2001; van Sluis and McStay 2019), the predominant mechanism of homogenization mechanism are gene conversion and unequal crossover. In the process of exchanging DNA sections in meiosis, gene conversions are an order of magnitude more common than crossover (Cole, et al. 2012; Williams, et al. 2015).

There have been many rigorous analyses that confront the homogenizing mechanisms directly. These studies (Smith 1974; Ohta 1976; Dover 1982; Nagylaki 1983; Ohta and Dover 1983) modeled gene conversion and unequal cross-over head on. Unfortunately, on top of the complexities of such models, the key parameter values are rarely obtainable. In the branching

Table 6.

The A/T to G/C and G/C to A/T changes in apes and mouse.

New mutants with a frequency >0.8 within an individual are considered as (nearly) fixed, except for humans, where the frequency was averaged over 8 individuals in the **Table 2**. The X-squared values for each species are as follows: 63.556 for human, 12.236 for chimpanzee, and 0.0029283 for *M. m. domesticus*.

	Direction of changes	Not fixed mutation counts	Fixed mutation counts	Chi-square test
Human	A/T to G/C	163	223	P = 1.559e-15
	G/C to A/T	159	48	
Chimpanzee	A/T to G/C	289	210	P = 0.000469
	G/C to A/T	201	83	
M. m. domesticus	A/T to G/C	99	10	P = 0.9568
	G/C to A/T	156	14	

Table 7.

The parameter values of *p*, *q*, *f* and *g* in the evolution between A/T and G/C.

The proportion of AT sites and GC sites are represented by *p* and *q* respectively. The proportion of AT-to-GC and GC-to-AT conversions (shown in **Table 6**) are represented by *f* and *g*. When (*pf*)/(*qg*) = 1, the evolution between A/T and G/C is in equilibrium.

Species	**Human**			**Chimpanzee**			***M. m. domesticus***		
Sequence segments	Non-functional	Total rDNA	Whole genome	Non-functional	Total rDNA	Whole genome	Non-functional	Total rDNA	Whole genome
p	0.43	0.42		0.43	0.42		0.54	0.51	
q	0.57	0.58	0.41	0.57	0.58	0.42	0.46	0.49	0.42
f	0.0142	0.0125		0.0136	0.012		0.0005	0.0004	
g	0.0023	0.0019		0.0041	0.0034		0.0007	0.0006	
(*pf*)/(*qg*)	4.66	4.76		2.5	2.56		0.84	0.69	

process, all these complexities are wrapped into $V^*(K)$ for formulating the evolutionary rate. In such a formulation, the *collective* strength of these various forces may indeed be measurable, as shown in this study.

By compressing all stochastic elements into $V(K)$, or $V^*(K)$, we will be able to filter out random factors and focus on forces of greater biological interest such as selection. Interestingly, while rDNA evolution in mice can be fully accounted for by genetic drift alone, its evolution in the great apes must have been driven by selection along with genetic drift. The contrast shows that the GH model only tries to delineate genetic drift more accurately than has been done by the WF model.

The branching process is a model for general processes. Hence, it can be used to track genetic drift in systems with two stages of evolution - within- and between-individuals, even though TEs, viruses or rRNA genes are very different biological entities. We use the rRNA genes to convey this point. Multi-copy genes, like rDNA, are under rapid genetic drift via the homogenization process. The drift is strong enough to reduce the copy number in the population from $\sim 150N$ to $< N$. A fraction of mutations in multi-copy genes may have been fixed by drift almost instantaneously in the evolutionary time scale. This acceleration is seen in mice but would have been interpreted to be due to positive selection by the convention. Interestingly, while positive selection may not be necessary to explain the mice data, it is indeed evident in human and chimpanzee, as the evolutionary rate of rRNA genes exceeds the limit of the strong drift.

In multi-copy systems, the copy number (designated C) is often in the hundreds or thousands per individual. Nevertheless, $C = 2$ as in all diploids is also a multi-copy gene system as the two copies may often evolve interactively via gene conversion or segregation distortion (Wu, et al. 1988; McDermott and Noor 2010). The WF models have been noted to be inadequate even in diploid systems (Charlesworth 2009; Chen, et al. 2017). After all, the WF model is essentially a haloid model with $2N$ copies in the population.

In conclusion, the GH model is far more general than the WF model as the two companion studies clearly demonstrate. Its $E(K)$ parameter would define N's while $V(K)$ would track random changes in gene frequency. It can quantify genetic drift in multi-gene systems including viruses such as SARS-CoV-2. The chaotic analyses of COVID-19 have shown why an accurate account of genetic drift is indispensable (Deng, et al. 2022; Pan, Liu, et al. 2022; Ruan, Wen, et al. 2022; Hou, et al. 2023).

Materials and Methods

Data Collection

We collected high-coverage whole-genome sequencing data for our study. The genome sequences of human (n = 8), chimpanzee (n = 1) and gorilla (n = 1) were sourced from National Center for Biotechnology Information (NCBI) (Supplementary Table 3). Human individuals were drawn from diverse geographical origins encompassing three continents (4 Asians, 2 Europeans and 2 Africans) and Asia CRC was the normal tissue of Case 1 patient from (Chen, Wu, et al. 2022). Genomic sequences of mice (n = 13) were sourced from the Wellcome Sanger Institute's Mouse Genome Project (MGP) (Keane, et al. 2011). Although some artificial selection has been performed on laboratory mouse strains (Yang, et al. 2011), the WSB/EiJ, ZALENDE/Ei and LEWES/EiJ strains were derived from wild populations. Incorporating these wild-derived laboratory strains, along with other inbred strains, a cohort of 10 mice was utilized to approximate the population representation of *M. m. domesticus*. Furthermore, the low F_{IS} of 0.14 for rDNA in *M. m. domesticus* found in this study suggests that each mouse covers 86% of the population's genetic diversity, thereby mitigating concerns about potential sampling biases. Accessions and the detailed information of samples used in this study are listed in the Supplementary Table 3 and **Table 4**.

Variant allele frequency

Following adapter trimming and the removal of low-quality sequences, these whole-genome sequencing data of apes and mice were mapped against respective reference sequences: the human rDNA reference sequence (Human ribosomal DNA complete repeating unit, GenBank: U13369.1) and the mouse rDNA reference sequence (*Mus musculus* ribosomal DNA, complete repeating unit, GenBank: BK000964.3). Alignment was performed using Burrows-Wheeler-Alignment Tool v0.7.17 (Li and Durbin 2009) with default parameters. All mapping and analysis are performed among individual copies of rRNA genes. The pipelines in variant calling are similar to the ones used in the literature (Ma, et al. 2022; Sun, et al. 2022). Each individual was considered as a psedo-population of rRNA genes and the diversity of rRNA genes was calculated using this psedo-population of rRNA genes. To determine variant frequency within individual, variants were called from multiple alignment files using bcftools (Danecek, et al. 2021) to ensure the inclusion of all polymorphic sites that appeared in at least one sample and to maintain consistent processing steps. Per-sample variant calling results were generated in a VCF file using the following settings: 'bcftools mpileup --redo-BAQ --max-depth 50000 --per-sample-mF --annotate' and 'bcftools call -mv'. Our analysis specifically focused on single nucleotide variants (SNVs) with only two alleles, then calculated the frequency of major allele, which is less susceptible to technical errors, to estimate the nucleotide diversity using Eq. (2), Variant information of interest per sample was extracted from the VCF files using the command 'bcftools view -V indels' and 'bcftools query –f '%CHROM\t%POS\t%REF\t%ALT\t[\t%GT]\t[\t%DP]\t[\t%AD]\n'\'. The AD (total high-quality bases of allelic depths) was used to obtain the number of reference-supporting reads (n.ref) and alternative-supporting reads (n.alt) in each sample. Sites with a depth < 10 in any sample were filtered out, resulting in an average of the minimum depths above 3000 for each remaining site across all samples. This filtering step enhances the robustness of variant frequency estimation, where the variant frequency was approximated by the ratio n.alt/(n.ref + n.alt). This process allowed us to identify all polymorphic sites present in the samples and the variant frequency within each individual. Then the population-level variant frequency was computed by averaging variant frequencies across all individuals.

Identification of Divergence Sites

Polymorphism represents the transient phase preceding divergence. With variant frequency within individual, within species, and between species, we obtained the F_{IS} and F_{ST} values for each site. Within the *M. m. domesticus* population, 779 polymorphic sites were identified. Among these, 744 sites (95.5%) exhibited an F_{IS} below 0.4 and rarely above 0.8 in **Fig. 4B**. In the comparison between *M. m. domesticus* and *Mus spretus*, 1579 variant sites were found, among which 453 sites displayed an F_{ST} above 0.8, indicating swift fixation of mutations during species divergence.

F_{ST} analysis between human and chimpanzee was conducted for each human individual, summarized in **Table 5**. We identified a range from 672 to 705 sites with F_{ST} values above 0.8 across individuals, depicting robust divergence sites. Considering the high mutation rate in CpG sites (Ehrlich and Wang 1981; Sved and Bird 1990) and predominantly GC content in rDNA (*e.g.* 58% GC in human), we further estimated the evolutionary rate at non-CpG sites during the interspecies divergence. To achieve this, mutations in CpG sites were manually removed by excluding all sites containing CpG in one species and TpG or CpA in the other; the reverse was similarly discarded. Additionally, the count of non-CpG sites within the mapping length, where site depth exceeded 10, was performed by Samtools (Danecek, et al. 2021) with the settings 'samtools mpileup -Q 15 -q 20'. As a result, the evolutionary rate of rDNA in non-CpG sites was ascertained.

For assessing diversity and divergence across gene segments, we used 'samtools faidx' to partition variants into a total of 8 regions within rRNA genes, including 5'ETS, ITS1, ITS2, 3'ETS, IGS, 18S, 5.8S, and 28S, aligning them with corresponding reference sequences for further analysis. The

functional parts (18S, 5.8S, and 28S) were subject to strong negative selection, exhibiting minimal substitutions during species divergence as expected. This observation is primarily used for comparison with the non-functional parts and reflects that the non-functional parts are less constrained by negative selection.

Genome-wide Divergence Estimation

To assess the genome-wide divergence between 4 mouse strain species, we downloaded their toplevel reference genomes from Ensembl genome browser (GenBank Assembly ID: GCA_001624865.1, GCA_001624775.1, GCA_001624445.1, and GCA_001624835.1). Then we used Mash tools (Ondov, et al. 2016) to estimate divergence across the entire genome (mainly single-copy genes) with 'mash sketch' and 'mash dist'. Additionally, the effective copy number of rRNA genes, denoted as C^*, can be estimated by calculating the ratio of population diversity observed in rDNA to that observed in single-copy genes.

Estimation of Site Conversion

To estimate site conversions, whether accidental or directional, it was essential to identify new mutant alleles in each lineage after divergence. New mutations were defined as derived alleles that differed from ancestral alleles shared by two outgroup species, where the ancestral state also exhibited high identity among copies. By focusing on new mutations with low initial frequency, we minimized the influence of their initial frequency on fixation probability and fixation time. Specifically, variants shared between chimpanzee and gorilla, humans and gorilla, *M. m. castaneus* and *M. spretus*, each with a frequency greater than 0.8, were considered as the ancestral for human, chimpanzee, and *M. m. domesticus*, respectively.

Derived variants were then categorized into two groups: (nearly) fixed (frequency > 0.8) and not fixed mutations in the lineages of humans, chimpanzees and *M. m. domesticus* (**Table 6**). The frequency threshold of >0.8 was chosen to balance the need for a sufficient number of sites to calculate of (f/g), and to ensure their reliability. We also applied a more stringent threshold of >0.9, which yielded similar results. In this study, six types of mutations were tabulated, representing ancestral-to-derived as depicted in Supplementary Fig. 2. For example, A-to-G represented the both A-to-G and T-to-C types of mutations. The C-to-G (or G-to-C) and A-to-T (or T-to-A) types of mutations were excluded in the subsequent analysis.

Specifically, f represents the proportion of fixed mutations where an A or T nucleotide has been converted to a G or C nucleotide. The numerator for f is the number of fixed mutations from A-to-G, T-to-C, T-to-G, or A-to-C. .Since the fixed sites accounted for less than 1% of the non-functional length of rDNA, the denominator is the total number of A or T sites in the rDNA sequence of the species lineage.

Similarly, g is defined as the proportion of fixed mutations where a G or C nucleotide has been converted to an A or T nucleotide. The numerator for g is the number of fixed mutations from G-to-A, C-to-T, C-to-A, or G-to-T. The denominator is the total number of G or C sites in the rDNA sequence of the species lineage. The consensus rDNA sequences for the species lineage were generated by Samtools consensus (Danecek, et al. 2021) from the bam file after alignment. The following command was used: 'samtools consensus -@ 20 -a -d 10 --show-ins no --show-del yes input_sorted.bam output.fa'.

The alternative hypotheses of GC-biased mutation process (Wolfe, et al. 1989; Francino and Ochman 1999) alone can be rejected in this study. According to the prediction of the mutational mechanism hypotheses, AT or GC variants should have equal fixation probabilities. We quantified the nearly fixed number of AT-to-GC and GC-to-AT types of mutations and conducted a chi-square test to assess their fixation probabilities.

Notably, we found the G (or C) variant had a significant higher fixation probability than A (or T) at site level in apes, but not in mice. In order to test whether there is an equilibrium state at the genic level in these three species, we computed the (*pf*)/(*qg*) ratio in **Table 7**, and a significant deviation of the ratio from 1 would imply biased genic conversion.

Acknowledgements

We are grateful for the helpful comments from many colleagues on the Chat Group “Cancer - The new evolving species”, in particular, Weiwei Zhai, Yong Zhang, GuoDong Wang of CAS and Jianrong Yang of SYSU. This work was supported by the National Natural Science Foundation of China (32150006, 32293193/32293190 to C.I.W., 32200493 to Y.R., and 82341092 to HJ Wen.), the National Key Research and Development Projects of the Ministry of Science and Technology of China (2021YFC2301300, 2021YFC0863400), and Guangdong Key Research and Development Program (No. 2022B1111030001).

Additional information

Data Availability

No new data were generated in this study. The genomic data used in this study are available from National Center for Biotechnology Information (NCBI) (*https://www.ncbi.nlm.nih.gov/*) and the Mouse Genomes Project (*https://www.sanger.ac.uk/data/mouse-genomes-project/*). The specific accession numbers recorded in Supplementary Table S4 and S5.

Additional files

supplementary information

References

Alexandrov I, Kazakov A, Tumeneva I, Shepelev V, Yurov Y 2001) **Alpha-satellite DNA of primates: old and new families** *Chromosoma* **110**:253–266

Arkhipova IR 2018) **Neutral Theory, Transposable Elements, and Eukaryotic Genome Evolution** *Molecular Biology and Evolution* **35**:1332–1337

Arnheim N, Treco D, Taylor B, Eicher EM 1982) **Distribution of ribosomal gene length variants among mouse chromosomes** *Proc Natl Acad Sci U S A* **79**:4677–4680

Bowman JC, Petrov AS, Frenkel-Pinter M, Penev PI, Williams LD 2020) **Root of the Tree: The Significance, Evolution, and Origins of the Ribosome** *Chemical Reviews* **120**:4848–4878

Brown T, Jiricny J 1989) **Repair of base-base mismatches in simian and human cells** *Genome / National Research Council Canada = Génome / Conseil national de recherches Canada* **31**:578–583

Cabot EL, Doshi P, Wu ML, Wu CI 1993) **Population genetics of tandem repeats in centromeric heterochromatin: unequal crossing over and chromosomal divergence at the Responder locus of Drosophila melanogaster** *Genetics* **135**:477–487

Cannings C 1974) **The latent roots of certain Markov chains arising in genetics: A new approach, I. Haploid models** *Advances in Applied Probability* **6**:260–290

Cazaux B, Catalan J, Veyrunes F, Douzery EJP, Britton-Davidian J 2011) **Are ribosomal DNA clusters rearrangement hotspots? A case study in the genus Mus (Rodentia, Muridae)** *BMC Evolutionary Biology* **11**:124

Charlesworth B 2009) **Effective population size and patterns of molecular evolution and variation** *Nature Reviews Genetics* **10**:195–205

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

Chen BJ, Wu XR, Ruan YS, Zhang YL, Cai QC, Zapata L, Wu CI, Lan P, Wen HJ 2022) **Very large hidden genetic diversity in one single tumor: evidence for tumors-in-tumor** *National Science Review* **9**:nwac250

Chen QP, Yang H, Feng X, Chen QJ, Shi SH, Wu CI, He ZW 2022) **Two decades of suspect evidence for adaptive molecular evolution-negative selection confounding positive-selection signals** *National Science Review* **9**:nwab217

Chen Y, Tong D, Wu CI 2017) **A New Formulation of Random Genetic Drift and Its Application to the Evolution of Cell Populations** *Mol Biol Evol* **34**:2057–2064

Chia AB, Watterson GA 1969) **Demographic effects on the rate of genetic evolution I. constant size populations with two genotypes** *Journal of Applied Probability* **6**:231–248

Cole F, Kauppi L, Lange J, Roig I, Wang R, Keeney S, Jasin M 2012) **Homeostatic control of recombination is implemented progressively in mouse meiosis** *Nature Cell Biology* **14**:424–

430

Crow JF, Kimura M 1970) **An Introduction to Population Genetics Theory** Blackburn Press

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021) **Twelve years of SAMtools and BCFtools** *Gigascience* **10**

Deng S, Xing K, He X 2022) **Mutation signatures inform the natural host of SARS-CoV-2** *National Science Review* **9**:nwab220

Der R, Epstein CL, Plotkin JB 2011) **Generalized population models and the nature of genetic drift** *Theoretical Population Biology* **80**:80–99

Dover G 1982) **Molecular drive: a cohesive mode of species evolution** *Nature* **299**:111–117

Ehrlich M, Wang RY. 1981) **5-Methylcytosine in eukaryotic DNA** *Science* **212**:1350–1357

Eickbush TH, Eickbush DG 2007) **Finely orchestrated movements: evolution of the ribosomal RNA genes** *Genetics* **175**:477–485

Francino MP, Ochman H 1999) **Isochores result from mutation not selection** *Nature* **400**:30–31

Fujiwara K, Kawai Y, Takada T, Shiroishi T, Saitou N, Suzuki H, Osada N 2022) **Insights into Mus musculus Population Structure across Eurasia Revealed by Whole-Genome Analysis** *Genome Biol Evol* **14**

Galtier N, Duret L 2007) **Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution** *Trends in Genetics* **23**:273–277

Geraldes A, Basset P, Gibson B, Smith KL, Harr B, Yu HT, Bulatova N, Ziv Y, Nachman MW 2008) **Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes** *Mol Ecol* **17**:5349–5363

Gibbons JG, Branco AT, Godinho SA, Yu S, Lemos B. 2015) **Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes** *PNAS* :2485

Gonzalez IL, Sylvester JE 2001) **Human rDNA: Evolutionary Patterns within the Genes and Tandem Arrays Derived from Multiple Chromosomes** *Genomics* **73**:255–263

Guan W-j, Zhong N-s 2022) **Strategies for reopening in the forthcoming COVID-19 era in China** *National Science Review* **9**:nwac054

Guarracino A, Buonaiuto S, de Lima LG, Potapova T, Rhie A, Koren S, Rubinstein B, Fischer C, Abel HJ, Antonacci-Fulton LL, et al. 2023) **Recombination between heterologous human acrocentric chromosomes** *Nature* **617**:335–343

Guillén AKZ, Hirai Y, Tanoue T, Hirai H 2004) **Transcriptional repression mechanisms of nucleolus organizer regions (NORs) in humans and chimpanzees** *Chromosome Research* **12**:225–237

Haldane JBS 1932) **The Causes of Evolution** London: Longmans, Green and Company

Haldane JBS 1927) **A mathematical theory of natural and artificial selection, part V: selection and mutation** Cambridge University Press

Hartl DL, Clark AG, Clark AG 1997) **Principles of population genetics** Sinauer Associates, Inc

Hori Y, Shimamoto A, Kobayashi T 2021) **The human ribosomal DNA array is composed of highly homogenized tandem clusters** *Genome Res* **31**:1971–1982

Hou M, Shi JR, Gong ZK, Wen HJ, Lan Y, Deng XZ, Fan QH, Li JJ, Jiang ML, Tang XP, et al. 2023) **Intra- vs. Interhost Evolution of SARS-CoV-2 Driven by Uncorrelated Selection-The Evolution Thwarted** *Molecular Biology and Evolution* **40**:msad204

Huang J, Liu X, Zhang L, Zhao Y, Wang D, Gao J, Lian X, Liu C 2021) **The oscillation-outbreaks characteristic of the COVID-19 pandemic** *National Science Review* **8**:nwab100

Jeffreys AJ, Neumann R 2002) **Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot** *Nat Genet* **31**:267–271

Jurka J, Kapitonov VV, Kohany O, Jurka MV 2007) **Repetitive sequences in complex genomes: structure and evolution** *Annu Rev Genomics Hum Genet* **8**:241–259

Karlin S, McGregor J 1964) **Direct Product Branching Processes and Related Markov Chains** *Proceedings of the National Academy of Sciences* **51**:598–602

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, et al. 2011) **Mouse genomic variation and its effect on phenotypes and gene regulation** *Nature* **477**:289–294

Krystal M, D'Eustachio P, Ruddle FH, Arnheim N 1981) **Human nucleolus organizers on nonhomologous chromosomes can share the same ribosomal gene variants** *Proceedings of the National Academy of Sciences of the United States of America* **78**:5744–5748

Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB 2022) **TimeTree 5: An Expanded Resource for Species Divergence Times** *Mol Biol Evol* :39

Li H, Durbin R 2009) **Fast and accurate short read alignment with Burrows-Wheeler transform** *Bioinformatics* **25**:1754–1760

Li W-H 1997) **Molecular evolution** Sunderland, Mass: Sinauer Associates

Ma Y, Mao X, Wang J, Zhang L, Jiang Y, Geng Y, Ma T, Cai L, Huang S, Hollingsworth P, et al. 2022) **Pervasive hybridization during evolutionary radiation of Rhododendron subgenus Hymenanthes in mountains of southwest China** *National Science Review* **9**:nwac276

McDermott SR, Noor MAF 2010) **The role of meiotic drive in hybrid male sterility** *Philosophical Transactions of the Royal Society B-Biological Sciences* **365**:1265–1272

Meunier J, Duret L 2004) **Recombination drives the evolution of GC-content in the human genome** *Molecular Biology and Evolution* **21**:984–990

Nagylaki T 1983) **Evolution of a large population under gene conversion** *Proc Natl Acad Sci U S A* **80**:5941–5945

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022) **The complete sequence of a human genome** *Science* **376**:44–53

Ohta T 1976) **Simple model for treating evolution of multigene families** *Nature* **263**:74–76

Ohta T, Dover GA 1984) **The Cohesive Population Genetics of Molecular Drive** *Genetics* **108**:501–521

Ohta T, Dover GA 1983) **Population genetics of multigene families that are dispersed into two or more chromosomes** *Proc Natl Acad Sci U S A* **80**:4079–4083

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM 2016) **Mash: fast genome and metagenome distance estimation using MinHash** *Genome Biology* **17**:132

Pan Y, Liu P, Wang F, Wu P, Cheng F, Jin X, Xu S 2022) **Lineage-specific positive selection on ACE2 contributes to the genetic susceptibility of COVID-19** *National Science Review* **9**:nwac118

Pan Y, Zhang C, Lu Y, Ning Z, Lu D, Gao Y, Zhao X, Yang Y, Guan Y, Mamatyusupu D, et al. 2022) **Genomic diversity and post-admixture adaptation in the Uyghurs** *National Science Review* **9**:nwab124

Parks MM, Kurylo CM, Dass RA, Bojmar L, Lyden D, Vincent CT, Blanchard SC 2018) **Variant ribosomal RNA alleles are conserved and exhibit tissue-specific expression** *Science Advances* **4**

Phifer-Rixey M, Harr B, Hey J 2020) **Further resolution of the house mouse (Mus musculus) phylogeny by integration over isolation-with-migration histories** *BMC Evol Biol* **20**:120

Potapova T, Gerton J 2019) **Ribosomal DNA and the nucleolus in the context of genome organization** *Chromosome Research* **27**

Rand DM 2001) **The units of selection on mitochondrial DNA** *Annual Review of Ecology and Systematics* **32**:415–448

Ruan Y, Hou M, Tang X, He X, Lu X, Lu J, Wu CI, Wen H 2022) **The Runaway Evolution of SARS-CoV-2 Leading to the Highly Evolved Delta Strain** *Mol Biol Evol* **39**

Ruan Y, Wang X, Hou M, Diao W, Tracy ME, Xu S, Liufu Z, Wen H, Wu C-I 2024) **Resolving Paradoxes in Molecular Evolution: The Integrated WF-Haldane (WFH) Model of Genetic Drift** *bioRxiv*

Ruan Y, Wen H, He X, Wu CI 2021) **A theoretical exploration of the origin and early evolution of a pandemic** *Sci Bull (Beijing* **66**:1022–1029

Ruan Y, Wen H, Hou M, He Z, Lu X, Xue Y, He X, Zhang YP, Wu CI 2022) **The twin-beginnings of COVID-19 in Asia and Europe-one prevails quickly** *Natl Sci Rev* **9**:nwab223

Rudra M, Chatterjee B, Bahadur M 2016) **Phylogenetic relationship and time of divergence of Mus terricolor with reference to other Mus species** *J Genet* **95**:399–409

Salim D, Gerton JL 2019) **Ribosomal DNA instability and genome adaptability** *Chromosome Research* **27**:73–87

Smirnov E, Chmúrčiaková N, Liška F, Bažantová P, Cmarko D 2021) **Variability of Human rDNA** *Cells* **10**

Smith GP 1974) **Unequal Crossover and the Evolution of Multigene Families** *Cold Spring Harb Symp Quant Biol* **38**:507–513

Stults DM, Killen MW, Pierce HH, Pierce AJ 2008) **Genomic architecture and inheritance of human ribosomal RNA gene clusters** *Genome Res* **18**:13–18

Sun N, Yang L, Tian F, Zeng H, He Z, Zhao K, Wang C, Meng M, Feng C, Fang C, et al. 2022) **Sympatric or micro-allopatric speciation in a glacial lake? Genomic islands support neither** *National Science Review* **9**:nwac291

Sved J, Bird AJPotNAoS 1990) **The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model** *PNAS* **87**:4692–4696

Szitenberg A, Cha S, Opperman CH, Bird DM, Blaxter ML, Lunt DH 2016) **Genetic Drift, Not Life History or RNAi, Determine Long-Term Evolution of Transposable Elements** *Genome Biology and Evolution* **8**:2964–2978

Tatsumoto S, Go Y, Fukuta K, Noguchi H, Hayakawa T, Tomonaga M, Hirai H, Matsuzawa T, Agata K, Fujiyama A 2017) **Direct estimation of de novo mutation rates in a chimpanzee parent-offspring trio by ultra-deep whole genome sequencing** *Sci Rep* **7**:13561

van Sluis M, Gailín M, McCarter JGW, Mangan H, Grob A, McStay B. 2019) **Human NORs, comprising rDNA arrays and functionally conserved distal elements, are located within dynamic chromosomal regions** *Genes Dev* **33**:1688–1701

van Sluis M, McStay B. 2019) **Nucleolar DNA Double-Strand Break Responses Underpinning rDNA Genomic Stability** *Trends in Genetics* **35**:743–753

Varki A, Altheide TK 2005) **Comparing the human and chimpanzee genomes: searching for needles in a haystack** *Genome Res* **15**:1746–1758

Wang X, He Z, Guo Z, Yang M, Xu S, Chen Q, Shao S, Li S, Zhong C, Duke NC, et al. 2022) **Extensive gene flow in secondary sympatry after allopatric speciation** *National Science Review* **9**:nwac280

Williams AL, Genovese G, Dyer T, Altemose N, Truax K, Jun G, Patterson N, Myers SR, Curran JE, Duggirala R, et al. 2015) **Non-crossover gene conversions show strong GC bias and unexpected clustering in humans** *eLife* **4**

Wolfe KH, Sharp PM, Li WH 1989) **Mutation rates differ among regions of the mammalian genome** *Nature* **337**:283–285

Wu C-I, Lyttle TW, Wu M-L, Lin G-F 1988) **Association between a satellite DNA sequence and the responder of segregation distorter in D. melanogaster** *Cell* **54**:179–189

Wu CI, Wang HY, Ling S, Lu X 2016) **The Ecology and Evolution of Cancer: The Ultra-Microevolutionary Process** *Annu Rev Genet* **50**:347–369

Xu J, Nuno K, Litzenburger UM, Qi YY, Corces MR, Majeti R, Chang HY 2019) **Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA** *eLife* **8**

Yang H, Wang JR, Didion JP, Buus RJ, Bell TA, Welsh CE, Bonhomme F, Yu AH-T, Nachman MW, Pialek J, et al. 2011) **Subspecific origin and haplotype diversity in the laboratory mouse** *Nat Genet* **43**:648–655

Zhai W, Lai H, Kaya NA, Chen J, Yang H, Lu B, Lim JQ, Ma S, Chew SC, Chua KP, et al. 2022) **Dynamic phenotypic heterogeneity and the evolution of multiple RNA subtypes in hepatocellular carcinoma: the PLANET study** *National Science Review* **9**:nwab192

Zhao Z, Jin L, Fu Y-X, Ramsay M, Jenkins T, Leskinen E, Pamilo P, Trexler M, Patthy L, Jorde LB, et al. 2000) **Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22** *Proceedings of the National Academy of Sciences* **97**:11354–11358

Zhou T, Shi T, Li A, Zhu L, Zhao X, Mao N, Qin W, Bi H, Yang M, Dai M, et al. 2022) **A third dose of inactivated SARS-CoV-2 vaccine induces robust antibody responses in people with inadequate response to two-dose vaccination** *National Science Review* **9**:nwac066

Author information

Xiaopei Wang
State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China
ORCID iD: 0009-0007-4587-4624

Yongsen Ruan
State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China
ORCID iD: 0000-0002-5573-4154

Lingjie Zhang
State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China

Xiangnyu Chen
State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China

Zongkun Shi
State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China

Haiyu Wang
State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China

Bingjie Chen
State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China

Miles Tracy
State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China

Liying Huang
State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China

Chung-I Wu
State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China

For correspondence: ciwu@uchicago.edu

Haijun Wen
State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China

For correspondence: wenhj5@mail.sysu.edu.cn

Editors

Reviewing Editor
Ziyue Gao
University of Pennsylvania, Philadelphia, United States of America

Senior Editor
George Perry
Pennsylvania State University, University Park, United States of America

Reviewer #1 (Public review):

The fundamental claim of the manuscript is that rRNA genes experience substitutions much too quickly, given that they are a multi-copy gene system. As clarified by the authors in their response, and as I think is relatively clear in the manuscript, they are collapsing all copies of the rRNA array down. They first quantify polymorphism (in this expanded definition, where polymorphism means variable at a given site across any copy). The authors find elevated levels of heterozygosity in rRNA genes compared to single copy genes, which isn't surprising, given that there is a substantially higher target size; that being said, the increase in polymorphism is smaller than the increase in target size. They then look at substitutions between mouse species and also between human and chimp, and argue that the substitution rate is too fast compared to single copy genes in many cases.

[Editors' note: we invite readers to consult the review in full from the previous version of the submission: https://doi.org/10.7554/eLife.99992.2.sa1]

https://doi.org/10.7554/eLife.99992.3.sa2

Reviewer #2 (Public review):

This revision has further improved the clarity of the paper, better articulating assumptions of the model and data analysis. I particularly appreciate the authors' thorough response to eLife assessment. However, the authors did not provide point-by-point response to the specific comments I had from last round of review and didn't revise the manuscript accordingly, so my major concerns remain.

At conceptual level, my biggest concern with the model is the lack of constraint on V*(K), which makes the null neutral model too "liberal". On the one hand, the number of descendants of each gene copy must be non-negative; on the other hand, even homogenizing process within an individual is extremely strong, it cannot "spread" gene copies across individuals, so the maximum number of descendants of one gene copy cannot exceed the number of offspring that individual has times C. For these reasons, I believe there must be a theoretical upper bound of the value of V*(K), and the actual V*(K) is likely much smaller under realistic strength of the homogenizing process. When I asked about modeling of the underlying homogenizing process, I did not mean the authors need to include specific molecular process in the model; instead, I am asking the authors to provide some realistic scenarios that can give rise to very large V*(K) values. As a result of the very "liberal" neutral model, although I do agree that rejection of null provides stronger evidence for selection in human, it is unclear whether there is no evidence of selection in mouse. Please see below for my specific comments regarding the definition and assumptions of V*(K) (copied from last review).

Regarding the data analysis, although I understand the authors' methodology and rationale behind, I am not convinced that high sequence similarity between rDNA copies guarantees no biases in alignment and variant calling. Furthermore, given divergence between species, I am particularly concerned about the practice of aligning reads of different species to human and mus musculus reference sequences. A separate issue is the calculation of divergence level. Instead of using Fst>0.8 as the criterion of calling fixed sites, the authors could calculate the pairwise average divergence between a random copy from one species and a random copy from another species. Mathematically, this could be calculated as p1(1-p2)+p2(1-p1). The observation that the estimated substitution rates for rDNA with and without CpG sites are so close seems to be an indication of technical error. Please also see below for my specific questions about data analysis (copied from last round of review).

Specific comments from last round of review:

Questions regarding V*(K)
(1) Another key parameter V*(K) was still not defined within the paper. In response 9, the authors explained that V*(K) refers to "the number of progeny to whom the gene copy of interest is transmitted (K) over a specific time interval". However, the meaning of "progeny" remains unclear. Are the authors referring to the descendent copies of a gene copy, or the offspring individuals (i.e., the living organisms)? For example, if a variant spreads horizontally through homogenizing processes and transmits vertically to multiple offspring individuals, the number of descent gene copies could differ substantially from the number of descendent individuals to whom a gene copy is transmitted to. This distinction needs to be clarified and clearly stated in the paper.

(2) The authors state that V*(K)>=1 for rDNA genes because of the homogenizing processes (lines 139-141) without providing justification. It is unclear, at least to me, whether homogenizing processes are expected increase or decrease the variance in "reproductive success" across gene copies. Moreover, the authors claim that V*(K) "can potentially reach values in the hundreds and may even exceed C, resulting in C*=C/V*(K)<1" (Response 7). This claim is unlikely to be true, as the minimum value of K is bounded by zero and E(K) is

assumed to be 1. Even in the extreme case that 1% gene copies leave large numbers of descends while the others leave none, V*(K) would still be less than 100. Such extreme case seems highly improbable, given realistic rates of the homogenizing processes.

(3) Regardless of how the authors define V*(K), it is not immediately clear why Equation 1 (N*=NC/V*(K)) holds. Both sides of the equation have their independent meanings, so the authors need to provide a step-by-step derivation demonstrating that they are equal. Only by doing this will the implicit underlying assumptions become clearer. I also strongly recommend that the authors conduct forward-in-time simulations with fixed N, C, V*(K) (however they define it) and μ to confirm that the right side of Equation 1 actually predicts the N* as calculated from the polymorphism level using the equation in line 165.

Questions about Ne* for multi-copy system

(1) While Ne is clearly defined in the standard single-copy gene model as the reciprocal of genetic drift (i.e., the decay in heterozygosity), its meaning for multiple-copy genes is unclear. Based on the context, it appears that the authors define Ne as the parameter that fits the population polymorphism level (Hs) using the equation in line 165. This definition is reasonable, but it should be explicitly clarified in the text."

(2) Without providing justification, the authors assumed that a certain number N* exists for rRNA such that it fits both the polymorphism level (line 156) in recent timescales and divergence level in longer timescales (i.e., in the comparison between Tf and Td). However, if N, C or any other relevant parameters have varied substantially throughout evolution, N* is expected to vary with time, and the same value may not fit both polymorphism and divergence data simultaneously.

Questions about data analysis

(1) A significant issue with aligning reads to a single reference genome is reference bias, referring to the phenomenon that reads carrying the reference alleles tend to align more easily than those with one or more non-reference alleles, thus creating a bias in genotype calling or variant allele frequency quantification. As a result, there may be an underrepresentation of non-reference alleles in called variants or an underestimate of non-reference allele frequency, particularly in regions with high genetic diversity. Simply focusing on bi-allelic SNVs is insufficient to minimize reference bias. Given the fourfold increase in diversity within rDNA, the authors must either provide evidence that reference bias is not a significant concern or adopt graph-based reference genomes or more sophisticated alignment algorithms to address this issue.

(2) The potential for reference bias also renders the analysis of divergence sites unreliable, as aligning reads from one species (e.g. chimpanzee) to the reference of another species (e.g., human) is likely to introduce biases in variant calling between the two. One commonly adopted approach to address this imbalance is to align reads from both species to a third reference genome that is expected to be equidistantly related to both.

(3) Although it is somewhat reassuring that the estimated divergence rate of rDNA between human and macaque is comparable to that of the rest of the genome, there still remains concern of a under-estimation of divergence in rDNA regions due to reference bias issue. Note that while the "third genome" approach reduces imbalance between two genomes in comparison, it may still under-estimate overall divergence level due to under-calling of non-reference variants.
(4) In response to my question about the similarity in rDNA substitution rates estimated with or without CpG sites, the authors suggest that this "may be due to strong homogenizing forces, which can rapidly fix or eliminate variants" (response17). However, this explanation is insufficient, because the observed substitution rate depends on the mutation rate multiplied by the fixation probability, and accelerated fixation or loss does not alter either. Unless the

authors can provide more convincing explanation, technical errors in calling of fixed sites still remain a concern.

Minor points:

Line 157: The statement "where μ is the mutation rate of the entire gene" must be wrong, as the heterozygosity calculated with such μ would correspond to the chance of seeing two different haplotypes at gene level, which is incompatible with the empirical calculation specified in Equation 2. Instead, μ must represent the mutation rate per site averaged over the entire gene.

In response 22, the authors explained that the allele frequency spectrum shown in Fig 3 is folded, because the ancestral allele was not determined. However, this is inconsistent with x-axis Fig 3 ranging between 0 and 1. I suspect the x-axis represents the frequency of the alternative (i.e., non-reference) allele. If so, the reported correlation is inflated, as the reference allele is somewhat random, and a variant at joint ALT allele frequencies of (0.9, 0.9) is no different from a variant at (0.1, 0.1). The proper way of calculate this correlation is to first determine the minor allele frequency across individuals and then calculate the correlation between minor allele frequencies.

Similarly, in response 14, it is unclear what the x-axis represents. Is it the ALT allele frequency or derived allele frequency? If the former, why are only variants with AF>0.8 defined as fixed variants, while those with AF<0.2 excluded? If it is the latter, please describe how ancestral state is determined.

https://doi.org/10.7554/eLife.99992.3.sa1

Author response:

The following is the authors' response to the previous reviews.

> ***eLife Assessment***
> *…. While intuitive, the model's underlying issue is grouping many factors under "variance in reproductive success" without explicitly modeling the molecular processes. This limitation, …, provides incomplete support for the authors' claim that the observed paradoxical patterns in rRNA genes can largely be explained by homogenizing processes, such as gene conversion, unequal crossover and replication slippage.*

This second paper addresses the genetic drift in multi-copy gene systems using rRNA genes as an example. Note that genetic drift happens in two stages here – within individuals and between individuals while the drift mechanisms are very different between the two stages. We now reply to the editors' decision that it would be more rigorous to model each molecular process, than to lump all stochastic forces into V(K). We respond to this criticism on three fronts.

First, for molecular evolutionists, there is NO NEED to model the detailed molecular processes. This is because we are only interested in knowing the totality of the stochastic variations. Interesting biological forces such as selection and meiotic drive are masked by such random forces. Our objective is precisely to lump all noises into a quantity that can be estimated.

Second, the homogenization process is the bulk, if not the totality of the within-individual random forces (i.e,, genetic drift). The criticism of incomplete support for drift as a sufficient account of the observations is curious because we did conclude that genetic drift is an insufficient explanation of the human data. Since drift only influences fixation time, which

can have a significant effect in short-term evolution (as shown in Fig 2), but it does not affect fixation rate itself. In contrast, selection influences the both. Thus, we can define the limitation of drift in evolutionary process. Even if the speed of drift-driven fixation is only a few generations, it is still too little for the human-chimpanzee divergence comparisons. In contrast, the speed of genetic drift in mice, as extrapolated from the polymorphism data, is sufficient to drive the divergence between *M. m. domesticus* and *Mus spretus*. The criticism appears to be that unbiased gene conversion, unequal crossover and replication slippage together may be insufficient to account for the observations. Since the contribution of each of these three forces is not central to our goal of filtering out the total contributions, we only conclude that the totality of within-individual drift in mice is sufficient to explain the data.

Third, even if we really want to dissect the molecular processes, previous attempts by prominent theorists like Tom Nagylaki and Tomoko Ohta could only model a small subset of such processes. In fact, Ohta often lumps a few of these forces into one process. More importantly, if we want to tackle other systems like viruses and mitochondria, we will have to develop a new set of theories for each molecular process. V(K) can take care of all such diverse systemes. In short, genetic drift is just noises and our goal is to quantify them in total across diverse systmes. By filtering out noises, we will be able to move on to something more important.

We now briefly comment on the WF models in relation to multi-gene systems. For example, in the case SARS-CoV-2, there are millions of virions in each patient among millions of patients. It is not possible to know what Ne acaully means in the WF modesl. Also, the rDNA population in each individual is not the sub-populations of the WF models. After all, the mechanisms of genetic drift within individuals by the homogenization processes are entirely different from the genetic drift between individuals. For a comparison, we published several papers (cited in #2) using the Haldane model to estimate the strength of genetic drift. It is also important to note that the parameters and assumptions of WF model cannot fully capture the evolutionary dynamics of the multi-copy genes.

> *... ., along with insufficient consideration of technical challenges in alignment and variants calling, provides incomplete support for the authors' claim ...*

Before delving into the technical details, we would like to summarize our defense. First, all rRNA gene copies belong in a pseudo-population, due to the homogenization process. The concept of specific locus with specific variants does not apply. Second, the levels of within-individual and within-species variation is so low that sequence alignment is not a problem at all. Third, thanks to the large number of sequence reads, occasional sequence errors (rarely encountered) should have minimal effects on the analyses. Now the technical details:

Regarding the concerns about the alignment and variant calling, we would like to clarify our methodology. While we acknowledge the technical challenges inherent in alignment and variant calling, particularly with respect to orthologous alignments to distinguish different copies, it is important to note that rDNA copies are subject to homogenization processes, meaning that there is no orthology among rDNA copies. Due to the high sequence similarity and frequent genetic exchange among rDNA units within species, we used the species-specific rDNA reference sequence for variant calling. We directly utilized the raw read depth from all rDNA copies within individuals to calculate the site frequency. For each site, we focused on the frequency of the major allele to calculate nucleotide diversity using the 2p(1-p), where p represents the frequency of the major allele. This approach helps capture genetic variation while minimizing the impact of alignment or variant calling errors, which primarily affect low-frequency variants (e.g., 0.800A, 0.199T, 0.001C, with A being the major allele). As for the divergence sites between species, we defined FST = 0.8 as a cutoff (roughly, when a mutant is > 0.95 in frequency in one species and < 0.05 in the other, FST would be > 0.80.), which is less likely to be influenced by low-frequency polymorphic sites within species.We believe this

method is more appropriate for estimating genetic diversity at rDNA than traditional variant calling pipelines designed to detect homozygotes and heterozygotes.

https://doi.org/10.7554/eLife.99992.3.sa0