

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



Reviewed Preprint

v2 • November 15, 2024

Revised by authors

Reviewed Preprint

v1 • August 20, 2024

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

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

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

eLife Assessment

This study presents a **useful** theoretical model of molecular evolution and attempts to use it to resolve the paradox of rapid evolution of ribosomal RNA genes. While intuitive, the model's underlying issue is grouping many factors under "variance in reproductive success" without explicitly modeling the molecular processes. This limitation, along with insufficient consideration of technical challenges in alignment and variants calling, 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.

<https://doi.org/10.7554/eLife.99992.2.sa3>

Abstract

Multi-copy gene systems that evolve within, as well as between, individuals are common. They include viruses, mitochondrial DNAs, transposons and multi-gene families. The paradox is that their (neutral) evolution in two stages should be far slower than single-copy systems but the opposite is often true. As the paradox cannot be resolved by the standard Wright-Fisher (WF) model, we now apply the newly expanded WF-Haldane (WFH; (Ruan, et al. 2024)) model to mammalian ribosomal RNA (rRNA) genes. On average, rDNAs have $C \sim 150 - 300$ copies per haploid in humans. While a neutral mutation of a single-copy gene would take $4N$ generations (N being the population size of an ideal population) to become fixed, the time should be $4NC^*$ generations for rRNA genes (C^* being the effective copy number). Note that $C^* \gg 1$, but $C^* < (or >) C$ would depend on the drift strength. Surprisingly, the observed fixation time in mouse and human is $< 4N$, implying the paradox of $C^* < 1$. Genetic drift that encompasses all random neutral evolutionary forces appears as much as 100 times stronger for rRNA genes as for single-copy genes, thus reducing C^* to < 1 . The large increases in genetic drift are driven by the homogenizing forces of gene conversion, unequal crossover and replication slippage within individuals. This study is one of the first applications of the WFH model to track random genetic drift in multi-copy gene systems. Many random forces, often

stronger than the WF model prediction, could be mis-interpreted as the working of natural selection.

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 [\[1\]](#); Szitenberg, et al. 2016 [\[2\]](#); Xu, et al. 2019 [\[3\]](#); Ruan, et al. 2021 [\[4\]](#)). 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 [\[5\]](#); Eickbush and Eickbush 2007 [\[6\]](#); Jurka, et al. 2007 [\[7\]](#); Hou, et al. 2023 [\[8\]](#)). A reason for this paradox, as well as many others (Ruan, et al. 2024 [\[9\]](#)), is that the speed of neutral evolution of multi-gene systems is not known.

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 (Ruan, et al. 2024 [\[9\]](#)), we show that the standard Wright-Fisher (WF) model may often under-account genetic drift, thus leading to the over-estimation of selection.

We propose the integration of the WF model with the Haldane model, referred to as the WFH model of genetic drift (Ruan, et al. 2024 [\[9\]](#)). The Haldane 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 $V(K)/N$. In diploids, K would be the number of progeny to whom the gene copy of interest is transmitted. (The adjustments between haploidy and multi-ploidy are straightforward).

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 , $V(K)$ would also be ~ 1 . In this backdrop, many “modified WF” models have been developed (Der, et al. 2011 [\[10\]](#)), most of them permitting $V(K) \neq E(K)$ (Karlin and McGregor 1964 [\[11\]](#); Chia and Watterson 1969 [\[12\]](#); Cannings 1974 [\[13\]](#)). 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 gene sampling (see below and (Ruan, et al. 2024 [\[9\]](#))). One of the paradoxes, first noted in (Chen, et al. 2017 [\[14\]](#)) is genetic drift during tumor growth whereby drift appears to become stronger as N increases (Wu, et al. 2016 [\[15\]](#); Chen, Wu, et al. 2022 [\[16\]](#); Zhai, et al. 2022 [\[17\]](#)). 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 [\[18\]](#)), transposons (Szitenberg, et al. 2016 [\[2\]](#)), mitochondrial DNAs (Xu, et al. 2019 [\[3\]](#)), satellite DNAs (Cabot, et al. 1993 [\[19\]](#); Alexandrov, et al. 2001 [\[1\]](#)), and ribosomal RNA genes (van Sluis and McStay 2019 [\[20\]](#); Hori, et al. 2021 [\[21\]](#)). 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 [\[4\]](#); Deng, et al. 2022 [\[22\]](#); Guan and Zhong 2022 [\[23\]](#); Pan, Liu, et al. 2022 [\[24\]](#); Ruan, et al. 2022 [\[25\]](#); Zhou, et al. 2022 [\[26\]](#); Hou, et al. 2023 [\[8\]](#)).

In multi-copy systems, the copy number (designated C) is 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.

We now apply the Haldane model to multi-copy gene systems, using ribosomal RNA (rRNA) genes as an example. While the WF models have led to speculations of pervasive natural selection (Dover 1982; Arkhipova 2018; Chen, Yang, et al. 2022; Pan, Zhang, et al. 2022; Wang, et al. 2022), neutral stochasticity would be a simpler explanation if the more powerful Haldane model is adopted. Since the WF model can yield results that are often good approximations for the Haldane model, the integration is referred to as the WFH model.

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.

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

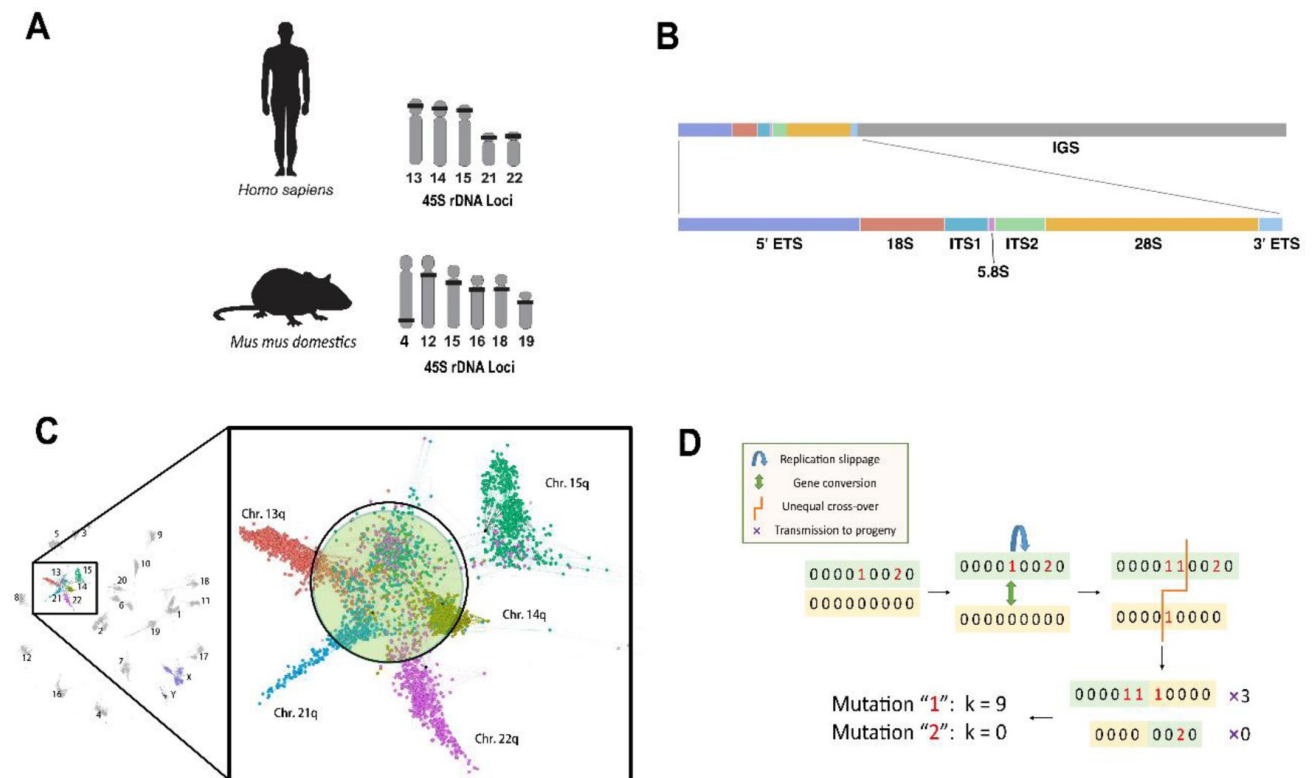


Figure 1.

The “chromosome community of rDNAs on five acrocentric chromosomes.

(A) The genomic locations of rDNA tandem repeats in human (Gibbons, et al. 2015 [\[1\]](#)) and mouse (Cazaux, et al. 2011 [\[2\]](#)). 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 [\[3\]](#)), 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.

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.]

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,

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

where 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 $C^* \gg 1$ 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 $N_e = C \cdot N$ for rRNA genes. The empirical measure of nucleotide diversity H is given by

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

where L is the gene length (for each copy of rRNA gene, $L \sim 43\text{kb}$) 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 (H_I , H_S and H_T , respectively). H_S and H_T are standard population genetic measures (Hartl, et al. 1997 [\[1\]](#); Crow and Kimura 2009 [\[2\]](#)). In calculating H_S , all sequences in the species are used, regardless of the source individuals. A similar procedure is applied to H_T . The H_I 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** [\[3\]](#) and text above). With multiple individuals, H_I 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** [\[4\]](#)), 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** [\[5\]](#)), 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)** [\[6\]](#) (Crow and Kimura 1970 [\[7\]](#); Hartl, et al. 1997 [\[8\]](#); Li 1997 [\[9\]](#)):

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

Mouse strains	<i>H_I</i> (‰) - diversity within individuals	
	Functional parts	Non-functional parts
	(18S 5.8S and 28S)	(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 <i>H_I</i> diversity across individuals	0.27	6.22
<i>H_S</i> for rDNA in <i>M. m.</i> <i>domesticus</i>	0.34	7.25
<i>H_S</i> for single-copy genes in <i>M. m.</i> <i>domesticus</i>	---	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.

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 $T^* < T$. Therefore, T 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 small T_f^*/T_d and thus a higher substitution rate, 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.

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

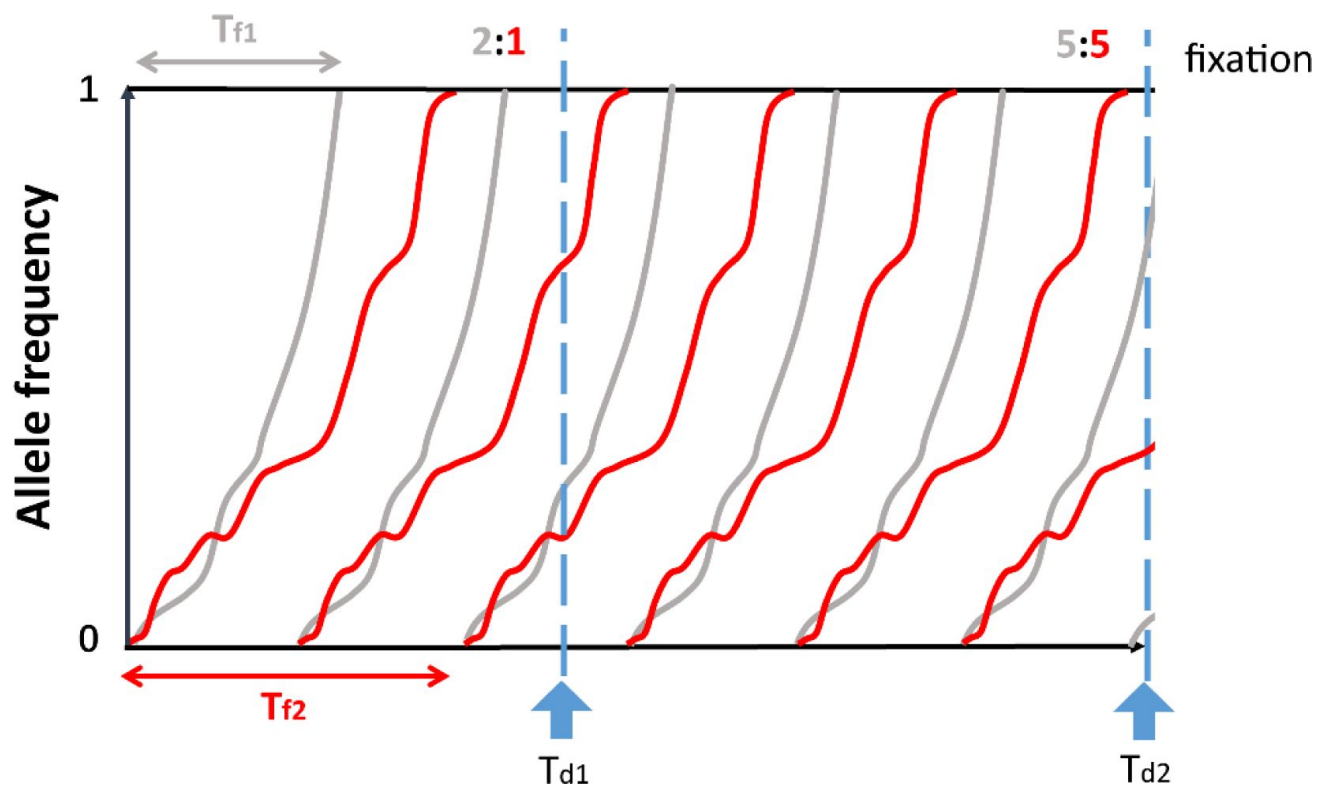


Figure 2.

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.

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 (Gerald, 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. Consistent 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 drift due to the sampling of chromosomes during sexual reproduction (e.g., segregation and assortment) is augmented substantially 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, λ . Since we will compare the λ values between rRNA and single-copy genes, we have to compare their mutation rates first by analyzing their long-term evolution. As shown in Table S1, λ 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 λ values will have to be explained by other means.

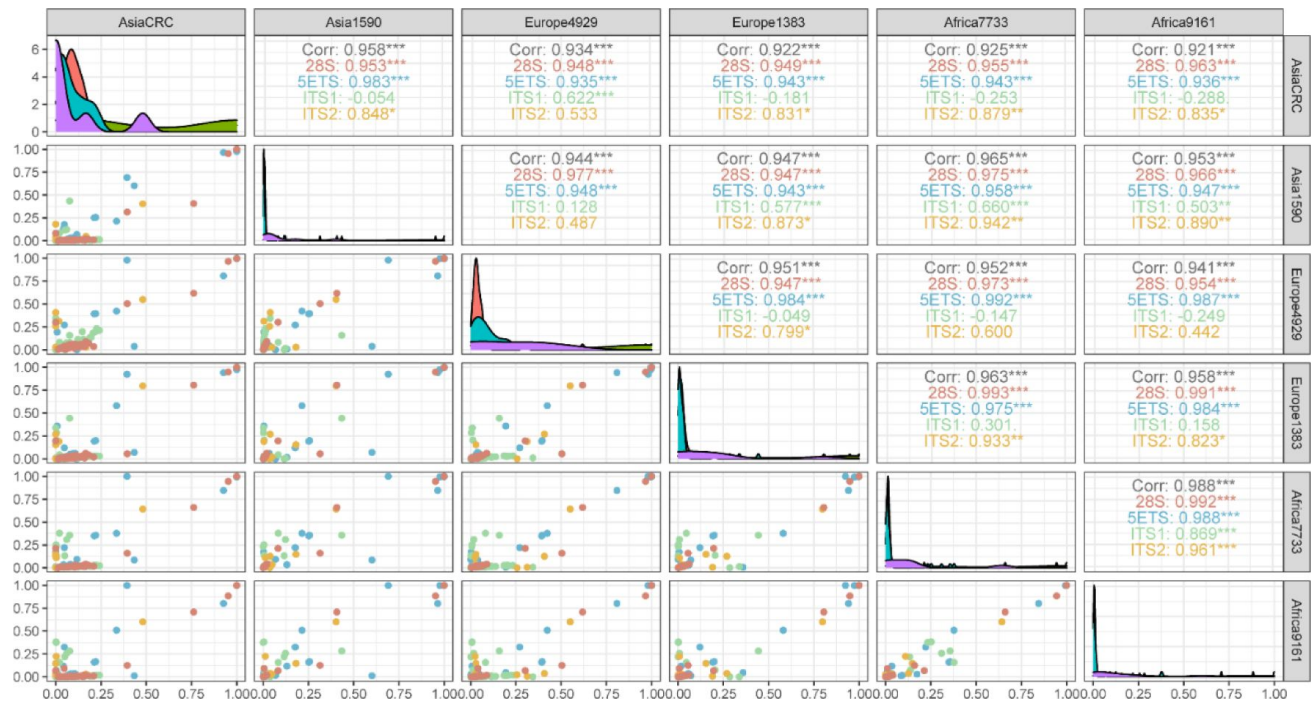


Figure 3.

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.

Human Individuals	<i>H_I</i> (‰) - diversity within individuals	
	Functional parts	Non-functional parts
	(18S 5.8S and 28S)	(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 <i>H_I</i> across individuals	0.50	6.81
<i>H_S</i> for rDNA in human population	0.68	7.24
<i>H_S</i> 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.

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^*/T_d (for rRNA genes). As T_f and T^* are both from *M. m. domesticus* and T 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^* 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 base pair per generation ([Gerald, 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 $\sim 40\%$ ($T_f/T_d = 0.2/0.5$), 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^* 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% ($T^*/T_d = 1/3$). 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 $C^* < 1$. With such rapid homogenization, the fixation time approaches 0 and the substitution rate in rRNA genes can indeed reach the theoretical limit of [Eq.](#)

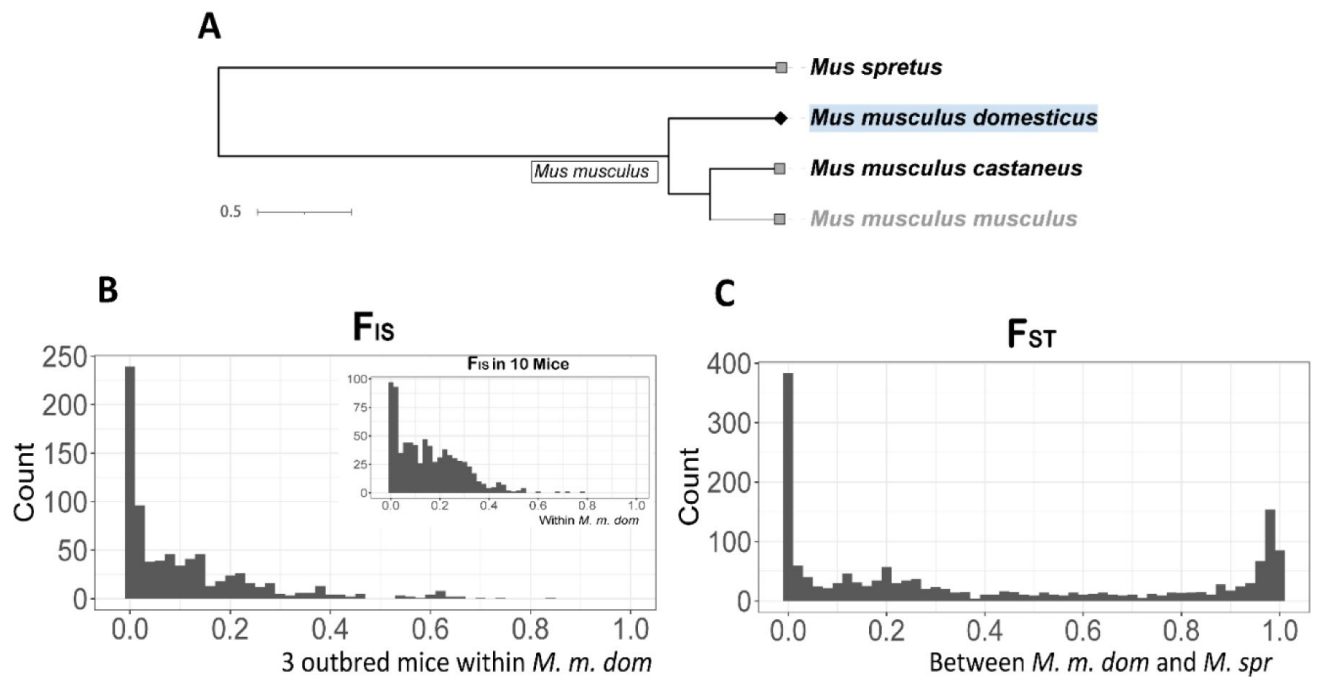


Figure 4.

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).

	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 3.

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

(3) [↗](#). In such a scenario, the substitution rate in ETS and ITS, compared to single-copy genes in mice, may increase by 7%, $T_f/(T_d - T_p) = 0.2/(3 - 0.2)$. If we use $T_f \sim 0.4$ 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 $T_f/(T_d - T_p)$ 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 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 analyze 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.

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.

Table 5.

Divergence in rRNA genes between Human and Chimpanzee.

	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.

	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	
<i>M. m. domesticus</i>	A/T to G/C	99	10	P = 0.9568
	G/C to A/T	156	14	

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*.

Table 6.

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

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 Mendelian populations. It 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](#)). It is not clear how large a role is played by replication slippage which affects copies of the same cluster.

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 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.

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 conclusion, the Haldane model is far more general than the WF model as this and the companion study clearly demonstrate. Its $E(K)$ parameter (which is usually set to 1) should be equivalent to the single parameter of the WF model (i.e., N). The other parameter of the Haldane model, i.e., $V(K)$, is then free to track genetic drift, whereas the WF model, setting $V(K) = E(K)$, is highly constrained. Nevertheless, the vast literature using the WF model has led to substantial understandings of the neutral process via the diffusion process ([Crow and Kimura 1970](#)) or coalescence ([Kingman 1982](#); [Fu 2006](#)). In this sense, the Haldane model should be built on the WF model by introducing a second parameter and permit the analyses of a broader range of stochastic ecological and evolutionary forces.

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

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.

Table 7.

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

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 [DOI](#)) 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 [DOI](#); Francino and Ochman 1999 [DOI](#)) 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** [DOI](#), and a significant deviation of the ratio from 1 would imply biased genic conversion.

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/> [DOI](#)) and the Mouse Genomes Project (<https://www.sanger.ac.uk/data/mouse-genomes-project/> [DOI](#)). The specific accession numbers recorded in Supplementary Table S4 and S5.

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).

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**
- 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 B, Wu X, Ruan Y, Zhang Y, Cai Q, Zapata L, Wu CI, Lan P, Wen H (2022) **Very large hidden genetic diversity in one single tumor: evidence for tumors-in-tumor** *Natl Sci Rev* **9**
- 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**
- 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 J, Kimura MJNY (1970) **An Introduction to Population Genetics Theory** Harper & Row
- Crow JF, Kimura M (2009) **An Introduction to Population Genetics Theory** Blackburn Press
- Danecek P *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**
- 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
- Fu Y-X (2006) **Exact coalescent for the Wright–Fisher model** *Theoretical Population Biology* **69**:385–394
- 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** *Proc Natl Acad Sci U S A* **112**:2485–2490
- 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**
- Guarracino A *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

- Hartl DL, Clark AG, Clark AG (1997) **Principles of population genetics** Sinauer associates Sunderland
- 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 *et al.* (2023) **Intra- vs. Interhost Evolution of SARS-CoV-2 Driven by Uncorrelated Selection-The Evolution Thwarted** *Molecular Biology and Evolution* **40**
- 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**
- 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 *et al.* (2011) **Mouse genomic variation and its effect on phenotypes and gene regulation** *Nature* **477**:289–294
- Kingman JFC (1982) **On the Genealogy of Large Populations** *Journal of Applied Probability* **19**:27–43
- 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, Kaspruwicz 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 *et al.* (2022) **Pervasive hybridization during evolutionary radiation of Rhododendron subgenus Hymenantes in mountains of southwest China** *National Science Review* **9**
- 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 *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 (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**

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**

Pan Y *et al.* (2022) **Genomic diversity and post-admixture adaptation in the Uyghurs** *National Science Review* **9**

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**

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

Ruan Y, Wang X, Hou M, Diao W, Xu S, 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**

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 *et al.* (2022) **Sympatric or micro-allopatric speciation in a glacial lake? Genomic islands support neither** *National Science Review* **9**

Sved J, Bird A (1990) **The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model** *Proc Natl Acad Sci U S A* **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**
- 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 *et al.* (2022) **Extensive gene flow in secondary sympatry after allopatric speciation** *National Science Review* **9**
- Williams AL *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 *et al.* (2011) **Subspecific origin and haplotype diversity in the laboratory mouse** *Nat Genet* **43**:648–655
- Zhai W *et al.* (2022) **Dynamic phenotypic heterogeneity and the evolution of multiple RNA subtypes in hepatocellular carcinoma: the PLANET study** *National Science Review* **9**
- Zhao Z *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 *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**

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](https://orcid.org/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](https://orcid.org/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 E Tracy

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 revision by Wang et al is a much more clear and readable manuscript than the original version, which I think was a bit too terse and hard to parse. In this version, I think I basically understand all the analyses that the authors undertake and how they argue that those analyses support their conclusions.

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.

I think that this is an interesting problem and one that obviously occupies some space in the literature. As the authors point out, one possibility for explaining the elevated fixation rate is that there is some kind of positive selection in these putatively non-functional regions. The authors, instead, argue that the elevated rate of evolution is due to neutral homogenizing processes. I'm sympathetic to this argument, I'm a neutralist myself :)

That being said, I find the whole analysis and the connection with the WFH model very strange. As I stated in my previous review, it feels very odd to chalk everything up to variance in reproductive success, rather than explicitly modeling the molecular processes that may lead to the homogenization. For example, the authors bring up gene conversion, and even do a small test of gene conversion. But a force like biased gene conversion is perhaps better modeled as a deterministic force, rather than a stochastic force. Indeed, I think that explicit modeling of mutation dynamics has been very helpful in understanding the role of replicative vs damage-related mutation in humans, as seen in Gao et al (2016) and Spisak et al (2024). I realize, as the authors say in their cover letter, that this is hard! But a major concern with this manuscript is that it's about whether drift can plausibly explain the pattern, but then it's basically impossible to know if it really can, because we have no way to compare the estimated parameters with biophysical or biochemical measurements of the rates of homogenizing forces, because the homogenizing forces are just wrapped up under "variance in reproductive success". I think a much more interesting manuscript would have a more explicit model of homogenizing forces.

I also have some concerns about the data analysis, echoing some concerns of the other reviewer. The biggest issue is that traditional read mapping and SNP calling pipelines for highly duplicated loci don't really make sense. I don't fully understand the variant calling pipeline. The authors state that "All mapping and analysis are performed among individual

copies of rRNA genes." which makes it sound like the reads mapping to different copies were somehow deconvolved, which is what you'd need to do to use "normal" variant calling approaches that call look for homozygotes and heterozygotes. But I don't know enough about this literature to understand how they did that and if it makes any sense. If, instead, they called variants against collapsed rRNA copies, then using a standard variant calling approach does not make sense. If you have a variant in 2 out of 100 copies, a standard variant calling algorithm would very likely call that a homozygous ancestral site. Conditional on the variant calls being reasonable, however, I'm basically okay with their use of read counts to estimate "allele frequencies" within individuals.

I have some more minor comments:

(1) In the paragraph starting line 61, the authors say that WF models are unable to handle things like viral epidemics and transposons. I don't think that's really fair: the issue here isn't WF dynamics or not, it's that there is fundamentally evolution on two levels (which is also the case in the rRNA case considered in this manuscript). I certainly agree with the authors that you can't just naively apply standard pop gen theory in these systems, but I think the arrow at the WF model is misaimed, as the real issue is drift and selection on multiple levels.

(2) Line 268-269: The authors argue that the long term rate of evolution in rRNA genes is roughly similar to single copy genes, suggesting not a big influence of increased mutation rate. I'm not sure I understand where this number comes from, as opposed to the divergence numbers they look at in Table 3. These seem to be two different conclusions from roughly the same measurement? Surely I am misunderstanding something.

References:

Gao, Z., Wyman, M. J., Sella, G., & Przeworski, M. (2016). Interpreting the dependence of mutation rates on age and time. *PLoS biology*, 14(1), e1002355.

Spisak, N., de Manuel, M., Milligan, W., Sella, G., & Przeworski, M. (2024). The clock-like accumulation of germline and somatic mutations can arise from the interplay of DNA damage and repair. *PLoS biology*, 22(6), e3002678.

<https://doi.org/10.7554/eLife.99992.2.sa2>

Reviewer #2 (Public review):

I appreciate the authors' efforts in addressing previous feedback by correcting typos, clarifying terms, and expanding the methodological descriptions. The revisions have notably improved the manuscript's clarity and readability. However, despite these positive changes, I still have several significant concerns, both conceptual and technical, that need to be addressed to strengthen the conclusions of the paper.

The key idea of this paper is the treatment of rDNA copies in an individual as a pseudo-population and model their sequence evolution with the WFH framework by introducing the parameter $V^*(K)$. With this modeling framework, the authors claim that the molecular evolution rate of rDNA relative to that of single-copy genes can be expressed as a simple function $V^*(K)$ and C (the copy number per individual). Moreover, when $V^*(K)$ is sufficiently large, the neutral molecular evolution of rDNA can be faster than expected under a naïve model without considering horizontal, homogenizing processes and thus be potentially compatible with empirical data. However, several issues persist in the definition, assumptions, and derivation of the model:

(1) Several terms in the model remain undefined. While N_e 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 N_e as the parameter that fits the population polymorphism level (H_s) using the equation in line 165. This definition is reasonable, but it should be explicitly clarified in the text."

(2) 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.

(3) The authors state that $V^*(K) \geq 1$ 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.

(4) 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.

(5) 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 T_f and T_d). 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.

The authors also provided more detailed description of their data analysis methods, but some of my major concerns remain:

(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.99992.2.sa1>

Author response:

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

eLife Assessment

This study attempts to resolve an apparent paradox of rapid evolutionary rates of multi-copy gene systems by using a theoretical model that integrates two classic population models. While the conceptual framework is intuitive and thus useful, the specific model is perplexing and difficult to penetrate for non-specialists. The data analysis of rRNA genes provides inadequate support for the conclusions due to a lack of consideration of technical challenges, mutation rate variation, and the relationship between molecular processes and model parameters.

Overall Responses:

Since the eLife assessment succinctly captures the key points of the reviews, the reply here can be seen as the overall responses to the summed criticisms. We believe that the overview should be sufficient to address the main concerns, but further details can be found in the

point-by-point responses below. The overview covers the same grounds as the provisional responses (see the end of this rebuttal) but is organized more systematically in response to the reviews. The criticisms together fall into four broad areas.

First, the lack of engagement with the literature, particularly concerning Cannings models and non-diffusive limits. This is the main rebuttal of the companion paper (eLife-RP-RA-2024-99990). The literature in question is all in the WF framework and with modifications, in particular, with the introduction of $V(K)$. Nevertheless, all WF models are based on population sampling. The Haldane model is an entirely different model of genetic drift, based on gene transmission. Most importantly, the WF models and the Haldane model differ in the ability to handle the four paradoxes presented in the two papers. These paradoxes are all incompatible with the WF models.

Second, the poor presentation of the model that makes the analyses and results difficult to interpret. In retrospect, we fully agree and thank all the reviewers for pointing them out. Indeed, we have unnecessarily complicated the model. Even the key concept that defines the paradox, which is the effective copy number of rRNA genes, is difficult to comprehend. We have streamlined the presentation now. Briefly, the complexity arose from the general formulation permitting $V(K) \neq E(K)$ even for single copy genes. (It would serve the same purpose if we simply let $V(K) = E(K)$ for single copy genes.) The sentences below, copied from the new abstract, should clarify the issue. The full text in the Results section has all the details.

“On average, rDNAs have $C \sim 150 - 300$ copies per haploid in humans. While a neutral mutation of a single-copy gene would take $4N$ generations (N being the population size of an ideal population) to become fixed, the time should be $4NC^*$ generations for rRNA genes (C^* being the effective copy number). Note that $C^* \gg 1$, but $C^* < (or >) C$ would depend on the drift strength. Surprisingly, the observed fixation time in mouse and human is $< 4N$, implying the paradox of $C^* < 1$.”

Third, the confusion about which rRNA gene is being compared with which homology, as there are hundreds of them. We should note that the effective copy number C^* indicates that the rRNA gene arrays do not correspond with the “gene locus” concept. This is at the heart of the confusion we failed to remove clearly. We now use the term “pseudo-population” to clarify the nature of rDNA variation and evolution. The relevant passage is reproduced from the main text shown below.

“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.”

Fourth, the lack of consideration of many technical challenges. We have responded to the criticisms point-by-point below. One of the main criticisms is about mutation rate differences between single-copy and rRNA genes. We did in fact allude to the parity in mutation rate between them in the original text but should have presented this property more prominently as is done now. Below is copied from the revised text:

“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, m , determines the long-term evolutionary rate, l . Since we will compare the l values between rRNA and single-copy genes, we have to compare their mutation rates first by analyzing their long-term evolution. As shown in Table S1, l 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 l values will have to be explained by other means.”

While the overview should address the key issues, we now present the point-by-point response below.

Public Reviews:

Reviewer #1 (Public Review):

The manuscript by Wang et al is, like its companion paper, very unusual in the opinion of this reviewer. It builds off of the companion theory paper's exploration of the "Wright-Fisher Haldane" model but applies it to the specific problem of diversity in ribosomal RNA arrays.

The authors argue that polymorphism and divergence among rRNA arrays are inconsistent with neutral evolution, primarily stating that the amount of polymorphism suggests a high effective size and thus a slow fixation rate, while we, in fact, observe relatively fast fixation between species, even in putatively non-functional regions.

They frame this as a paradox in need of solving, and invoke the WFH model.

The same critiques apply to this paper as to the presentation of the WFH model and the lack of engagement with the literature, particularly concerning Cannings models and non-diffusive limits. However, I have additional concerns about this manuscript, which I found particularly difficult to follow.

Response 1: We would like to emphasize that, despite the many modified WF models, there has not been a model for quantifying genetic drift in multi-copy gene systems, due to the complexity of two levels of genetic drift – within individuals as well as between individuals of the population. We will address this question in the revised manuscript (Ruan, et al. 2024) and have included a mention of it in the text as follows:

“In the WF model, gene frequency is governed by $1/N$ (or $1/2_N$ in diploids) because K would follow the Poisson distribution whereby $V(K) = E(K)$. As $E(K)$ is generally ~ 1 , $V(K)$ would also be ~ 1 . In this backdrop, many "modified WF" models have been developed (Der, et al. 2011), most of them permitting $V(K) \neq E(K)$ (Karin 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 gene sampling (see below and (Ruan, et al. 2024)).”

My first, and most major, concern is that I can never tell when the authors are referring to diversity in a single copy of an rRNA gene compared to when they are discussing diversity across the entire array of rRNA genes. I admit that I am not at all an expert in studies of rRNA diversity, so perhaps this is a standard understanding in the field, but in order for this manuscript to be read and understood by a larger number of people, these issues must be clarified.

Response 2: We appreciate the reviewer’s feedback and acknowledge that the distinction between the diversity of individual rRNA gene copies and the diversity across the entire array of rRNA genes may not have been clearly defined in the original manuscript. The diversity in our manuscript is referring to the genetic diversity of the population of rRNA genes in the cell. To address this concern, we have revised the relevant paragraph in the text:

“Hence, rRNA gene diversity or polymorphism refer 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.”

Additionally, we have updated the Methods section to include a detailed description of how diversity is measured as follows:

“All mapping and analysis are performed among individual copies of rRNA genes.

Each individual was considered as a pseudo-population of rRNA genes and the diversity of rRNA genes was calculated using this pseudo-population of rRNA genes.”

The authors frame the number of rRNA genes as roughly equivalent to expanding the population size, but this seems to be wrong: the way that a mutation can spread among rRNA gene copies is fundamentally different than how mutations spread within a single copy gene. In particular, a mutation in a single copy gene can spread through vertical transmission, but a mutation spreading from one copy to another is fundamentally horizontal: it has to occur because some molecular mechanism, such as slippage, gene conversion, or recombination resulted in its spread to another copy. Moreover, by collapsing diversity across genes in an rRNA array, the authors are massively increasing the mutational target size.

For example, it's difficult for me to tell if the discussion of heterozygosity at rRNA genes in mice starting on line 277 is collapsed or not. The authors point out that Hs per kb is ~5x larger in rRNA than the rest of the genome, but I can't tell based on the authors' description if this is diversity per single copy locus or after collapsing loci together. If it's the first one, I have concerns about diversity estimation in highly repetitive regions that would need to be addressed, and if it's the second one, an elevated rate of polymorphism is not surprising, because the mutational target size is in fact significantly larger.

Response 3: As addressed in previous Response2, the measurement of diversity or heterozygosity of rRNA genes is consistently done by combining copies, as there is no concept of single gene locus for rDNAs. We agree that by combining the diversity across multiple rRNA gene copies into one measurement, the mutational target size is effectively increased, leading to higher observed levels of diversity than one gene. This is in line with our text:

“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 HS results confirm the prediction that rRNA genes should be more polymorphic than single-copy genes.”

Under this consensus, the reviewer points out that the having a large number of rRNA genes is not equivalent to having a larger population size, because the spreading of mutations among rDNA copies within a species involves two stages: within individual (*horizontal transmission*) and between individuals (*vertical transmission*). Let’s examine how the mutation spreading mechanisms influence the population size of rRNA genes.

First, an increase in the copy number of rRNA genes dose increase the actual population size (CN) of rRNA genes. If reviewer is referring to the effective population size of rRNA genes in the context of diversity ($N^* = CN/V^*(K)$), then an increase in C would also increase N^* . In addition, the linkage among copies would reduce the drift effect, leading to increase diversity. Conversely, homogenization mechanism, like gene conversion and unequal crossing-over would reduce genetic variations between copies and increase $V^*(K)$, leading to lower diversity. Therefore, the $C^* = C/V^*(K)$ in mice is about 5 times larger for rRNA genes than the rest of the genome (which mainly single-copy genes), even though the actual copy number is about 110, indicating a high homogenization rate.

Even if these issues were sorted out, I'm not sure that the authors framing, in terms of variance in reproductive success is a useful way to understand what is going on in rRNA arrays. The authors explicitly highlight homogenizing forces such as gene conversion and replication slippage but then seem to just want to incorporate those as accounting for variance in reproductive success. However, don't we usually want to dissect these things in terms of their underlying mechanism? Why build a model based on variance in reproductive success when you could instead explicitly model these homogenizing processes? That seems more informative about the mechanism, and it would also serve significantly better as a null model, since the parameters would be able to be related to in vitro or in vivo measurements of the rates of slippage, gene conversion, etc.

In the end, I find the paper in its current state somewhat difficult to review in more detail, because I have a hard time understanding some of the more technical aspects of the manuscript while so confused about high-level features of the manuscript. I think that a revision would need to be substantially clarified in the ways I highlighted above.

Response 4: We appreciate your perspective on modeling the homogenizing processes of rRNA gene arrays.

We employ the WFH model to track the drift effect of the multi-copy gene system. In the context of the Haldane model, the term K is often referred to as reproductive success, but it might be more accurate to interpret it as “transmission rate” in this study. As stated in the caption of Figure 1D, two new mutations can have very large differences in individual output (K) when transmitted to the next generation through homogenization process.

Regarding why we did not explicitly model different mechanisms of homogenization, previous elegant models of multigene families have involved mechanisms like unequal crossing over (Smith 1974a; Ohta 1976; Smith 1976) or gene conversion (Nagylaki 1983; Ohta 1985) for concerted evolution, or using conversion to approximate the joint effect of conversion and crossing over (Ohta and Dover 1984). However, even when simplifying the gene conversion mechanism, modeling remains challenging due to controversial assumptions, such as uniform homogenization rate across all gene members (Dover 1982; Ohta and Dover 1984). No models can fully capture the extreme complexity of factors, while these unbiased mechanisms are all genetic drift forces that contribute to changes in mutant transmission. Therefore, we opted for a more simplified and collective approach using $V^*(K)$ to see the overall strength of genetic drift.

We have discussed the reason for using $V^*(K)$ to collectively represent the homogenization effect in Discussion. As stated in our manuscript:

“There have been many rigorous analyses that confront the homogenizing mechanisms directly. These studies (Smith 1974b; 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 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.”

Reviewer #2 (Public Review):

Summary:

Multi-copy gene systems are expected to evolve slower than single-copy gene systems because it takes longer for genetic variants to fix in the large number of gene copies in the entire population. Paradoxically, their evolution is often observed to be surprisingly fast. To explain this paradox, the authors hypothesize that the rapid evolution of multi-copy gene systems arises from stronger genetic drift driven by homogenizing forces within individuals, such as gene conversion, unequal crossover, and replication slippage. They formulate this idea by combining the advantages of two classic population genetic models -- adding the $V(k)$ term (which is the variance in reproductive success) in the Haldane model to the Wright-Fisher model. Using this model, the authors derived the strength of genetic drift (i.e., reciprocal of the effective population size, N_e) for the multi-copy gene system and compared it to that of the single-copy system. The theory was then applied to empirical genetic polymorphism and divergence data in rodents and great apes, relying on comparison between rRNA genes and genome-wide patterns (which mostly are single-copy genes). Based on this analysis, the authors concluded that neutral genetic drift could explain the rRNA diversity and evolution patterns in mice but not in humans and chimpanzees, pointing to a positive selection of rRNA variants in great apes.

Strengths:

Overall, the new WFH model is an interesting idea. It is intuitive, efficient, and versatile in various scenarios, including the multi-copy gene system and other cases discussed in the companion paper by Ruan et al.

Weaknesses:

Despite being intuitive at a high level, the model is a little unclear, as several terms in the main text were not clearly defined and connections between model parameters and biological mechanisms are missing. Most importantly, the data analysis of rRNA genes is extremely over-simplified and does not adequately consider biological and technical factors that are not discussed in the model. Even if these factors are ignored, the authors' interpretation of several observations is unconvincing, as alternative scenarios can lead to similar patterns. Consequently, the conclusions regarding rRNA genes are poorly supported. Overall, I think this paper shines more in the model than the data analysis, and the modeling part would be better presented as a section of the companion theory paper rather than a stand-alone paper. My specific concerns are outlined below.

Response 5: We appreciate the reviewer's feedback and recognize the need for clearer definitions of key terms. We have made revisions to ensure that each term is properly defined upon its first use.

Regarding the model's simplicity, as in the Response4, our intention was to create a framework that captures the essence of how mutant copies spread by chance within a population, relying on the variance in transmission rates for each copy ($V(K)$). By doing so, we aimed to incorporate the various homogenization mechanisms that do not affect single-copy genes, highlighting the substantially stronger genetic drift observed in multi-copy systems compared to single-copy genes. We believe that simplifying the model was necessary to make it more accessible and practical for real-world data analysis and provides a useful approximation that can be applied broadly. It is clearly an underestimate the actual rate as some forces with canceling effects might not have been accounted for.

(1) Unclear definition of terms

Many of the terms in the model or the main text were not clearly defined the first time they occurred, which hindered understanding of the model and observations reported. To name a few:

(i) In Eq(1), although C^ is defined as the "effective copy number", it is unclear what it means in an empirical sense. For example, N_e could be interpreted as "an ideal WF population with this size would have the same level of genetic diversity as the population of interest" or "the reciprocal of strength of allele frequency change in a unit of time". A few factors were provided that could affect C^* , but specifically, how do these factors impact C^* ? For example, does increased replication slippage increase or decrease C^* ? How about gene conversion or unequal cross-over? If we don't even have a qualitative understanding of how these processes influence C^* , it is very hard to make interpretations based on inferred C^* . How to interpret the claim on lines 240-241 (If the homogenization is powerful enough, rRNA genes would have $C^* < 1$)? Please also clarify what C^* would be, in a single-copy gene system in diploid species.*

Response 6: We apology for the confusion caused by the lack of clear definitions in the initial manuscript. We recognize that this has led to misunderstandings regarding the concept we presented. Our aim was to demonstrate the concerted evolution in multi-copy gene systems, involving two levels of "effective copy number" relative to single-copy genes: first, homogenization within populations then divergence between species. We used C^* and N_e^* to try to designated the two levels driven by the same homogenization force, which complicated the evolutionary pattern.

To address these issues, we have simplified the model and revised the abstract to prevent any misunderstandings:

"On average, rDNAs have $C \sim 150 - 300$ copies per haploid in humans. While a neutral mutation of a single-copy gene would take $4N$ (N being the population size) generations to become fixed, the time should be $4N_{C^*}$ generations for rRNA genes where $1 < C^*$ (C^* being the effective copy number; $C^* < C$ or $C^* > C$ would depend on the drift strength). However, the observed fixation time in mouse and human is $< 4N$, implying the paradox of $C^* < 1$. Genetic drift that encompasses all random neutral evolutionary forces appears as much as 100 times stronger for rRNA genes as for single-copy genes, thus reducing C^* to < 1 ."

Thus, it should be clear that the fixation time as well as the level of polymorphism represent the empirical measures of C^* . We have also revised the relevant paragraph in the text to define C^* and $V^*(K)$ and removed Eq. 2 for clarity:

"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,

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

where 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 $1 \ll C^* \leq C$ 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 $C^* \gg 1$ is hypothesized, genetic drift for rRNA genes is expected to be slower than for single copy genes.”

(ii) In Eq(1), what exactly is $V^(K)$? Variance in reproductive success across all gene copies in the population? What factors affect $V^*(K)$? For the same population, what is the possible range of $V^*(K)/V(K)$? Is it somewhat bounded because of biological constraints? Are $V^*(K)$ and $C^*(K)$ independent parameters, or does one affect the other, or are both affected by an overlapping set of factors?*

Response 7: - In Eq(1), what exactly is $V^*(K)$? In Eq(1), $V^*(K)$ refers to the variance in the number of progeny to whom the gene copy of interest is transmitted (K) over a specific time interval. When considering evolutionary divergence between species, $V^*(K)$ may correspond to the divergence time.

- What factors affect $V^*(K)$? For the same population, what is the possible range of $V^*(K)/V(K)$? Is it somewhat bounded because of biological constraints? “ $V^*(K)$ for rRNA genes is likely to be much larger than $V(K)$ for single-copy genes, because K for rRNA mutations may be 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. For simplicity, we let $V(K) = 1$ (as in a WF population) and $V^*(K) \geq 1$.” Thus, the $V^*(K)/V(K) = V^*(K)$ can potentially reach values in the hundreds, and may even exceed C , resulting in $C^*(= C/V^*(K))$ values less than 1. Biological constraints that could limit this variance include the minimum copy number within individuals, sequence constraints in functional regions, and the susceptibility of chromosomes with large arrays to intrachromosomal crossover (which may lead to a reduction in copy number)(Eickbush and Eickbush 2007), potentially reducing the variability of K .

- Are $V^*(K)$ and $C^*(K)$ independent parameters, or does one affect the other, or are both affected by an overlapping set of factors? There is no $C^*(K)$, the C^* is defined as follows in the text:

“ $C^* = C/V^*(K)$ represents the effective copy number of rRNA genes, reflecting 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 $1 \ll C^* \leq C$ is hypothesized.” The factors influencing V^* (K) directly affect C^* due to this relationship.

(iii) In the multi-copy gene system, how is fixation defined? A variant found at the same position in all copies of the rRNA genes in the entire population?

Response 8: We appreciate the reviewer's suggestion and have now provided a clear definition of fixation in the context of multi-copy genes within the manuscript.

“For rDNA mutations, fixation must occur in two stages – fixation within individuals and among individuals in the population. (Note that a new mutation can be fixed via homogenization, thus making rRNA gene copies in an individual a pseudo-population.)”

The evolutionary dynamics of multi-copy genes differ from those of single-copy (Mendelian) genes, which mutate, segregate and evolve independently in the population. Fixation in multi-copy genes, such as rRNA genes, is influenced by their ability to transfer genetic information among their copies through nonreciprocal exchange mechanisms, like gene conversion and unequal crossover (Ohta and Dover 1984). These processes can cause fluctuations in the number of mutant copies within an individual's lifetime and facilitate the spread of a mutant allele across all copies even in non-homologous chromosomes. Over time, this can result in the mutant allele replacing all preexisting alleles throughout the population, leading to fixation (Ohta 1976) meaning that the same variant will eventually be present at the corresponding position in all copies of the rRNA genes across the entire population. Without such homogenization processes, fixation would be unlikely to be obtained in multi-copy genes.

(iv) Lines 199-201, H_I , H_S , and H_T are not defined in the context of a multi-copy gene system. What are the empirical estimators?

Response 9: We appreciate the reviewer's comment and would like to clarify the definitions and empirical estimators for within the context of a multi-copy gene system in the text:

“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

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

where L is the gene length (for each copy of rRNA gene, $L \sim 43\text{kb}$) 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 (H_I , H_S and H_T , respectively). H_S and H_T are standard population genetic measures (Hartl, et al. 1997; Crow and Kimura 2009). In calculating H_S , all sequences in the species are used, regardless of the source individuals. A similar procedure is applied to H_T . The H_I 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.”

(v) Line 392-393, *f* and *g* are not clearly defined. What does “the proportion of AT-to-GC conversion” mean? What are the numerator and denominator of the fraction, respectively?

Response 10: We appreciate the reviewer's comment and have revised the relevant text for clarity as well as improved the specific calculation methods for *f* and *g* in the Methods section.

“We first designate the proportion of AT-to-GC conversion as *f* and the reciprocal, GC-to-AT, as *g*. Specifically, *f* represents the proportion of fixed mutations where an A or T nucleotide has been converted to a G or C nucleotide (see Methods). Given $f \neq g$, this bias is true at the site level.”

Methods:

“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. The denominator is the total number of A or T sites in the rDNA sequence of the specie 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 specie 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’.`

(2) *Technical concerns with rRNA gene data quality*

Given the highly repetitive nature and rapid evolution of rRNA genes, myriads of things could go wrong with read alignment and variant calling, raising great concerns regarding the data quality. The data source and methods used for calling variants were insufficiently described at places, further exacerbating the concern.

(i) What are the accession numbers or sample IDs of the high-coverage WGS data of humans, chimpanzees, and gorillas from NCBI? How many individuals are in each species? These details are necessary to ensure reproducibility and correct interpretation of the results.

Response 11: We apologize for not including the specific details of the sample information in the main text. All accession numbers and sample IDs for the WGS data used in this study, including mice, humans, chimpanzee, and gorilla, are already listed in Supplementary Tables S4-S5. We have revised the table captions and referenced them at the appropriate points in the Methods to ensure clarity.

“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 4). ... Genomic sequences of mice ($n = 13$) were sourced from the Wellcome Sanger Institute’s Mouse Genome Project (MGP) (Keane, et al. 2011).

The concern regarding the number of individuals needed to support the results will be addressed in Response 13.

(ii) Sequencing reads from great apes and mice were mapped against the human and mouse rDNA reference sequences, respectively (lines 485-486). Given the rapid evolution of rRNA genes, even individuals within the same species differ in copy number and sequences of these genes. Alignment to a single reference genome would likely lead to incorrect and even failed alignment for some reads, resulting in genotyping errors. Differences in rDNA sequence, copy number, and structure are even greater between species, potentially leading to higher error rates in the called variants. Yet the authors provided no justification for the practice of aligning reads from multiple species to a single reference genome nor evidence that misalignment and incorrect variant calling are not major concerns for the downstream analysis.

Response 12: While the copy number of rDNA varies in each individuals, the sequence identity among copies is typically very high (median identity of 98.7% (Nurk, et al. 2022)). Therefore, all rRNA genes were aligned against to the species-specific reference sequences, where the consensus nucleotide nearly accounts for >90% of the gene copies in the population. In minimize genotyping errors, our analysis focused exclusively on single nucleotide variants (SNVs) with only two alleles, discarding other mutation types.

Regarding sequence divergence between species, which may have greater sequence variations, we excluded unmapped regions with high-quality reads coverage below 10. In calculation of substitution rate, we accounted for the mapping length (L), as shown in the column 3 in Table 3-5.

We appreciate the reviewer's comments and have provide details in the Methods.

(vi) It is unclear how variant frequency within an individual was defined conceptually or computed from data (lines 499-501). The population-level variant frequency was calculated by averaging across individuals, but why was the averaging not weighted by the copy number of rRNA genes each individual carries? How many individuals are sampled for each species? Are the sample sizes sufficient to provide an accurate estimate of population frequencies?

Response 13: Each individual was considered as a psedo-population of rRNA genes, varaint frequency within an individual was the proportions of mutant allele in this psedo-population. The calculation of varaint frequency is based on the number of supported reads of each individual.

The reason for calculating population-level variant frequency by averaging across individuals is relevant in the calculation of FIS and FST. In calculating FST, the standard practice is to weigh each population equally. So, when we show FST in humans, we do not consider whether there are more Africans, Caucasians or Asians. There is a reason for not weighing them even though the population sizes could be orders of magnitude different, say, in the comparison between an ethnic minority and the main population. In the case of FIS, the issue is moot. Although copy number may range from 150 to 400 per haploid, most people have 300 – 500 copies with two haploids.

As for the concern regarding the number the individuals needed to support of the results:

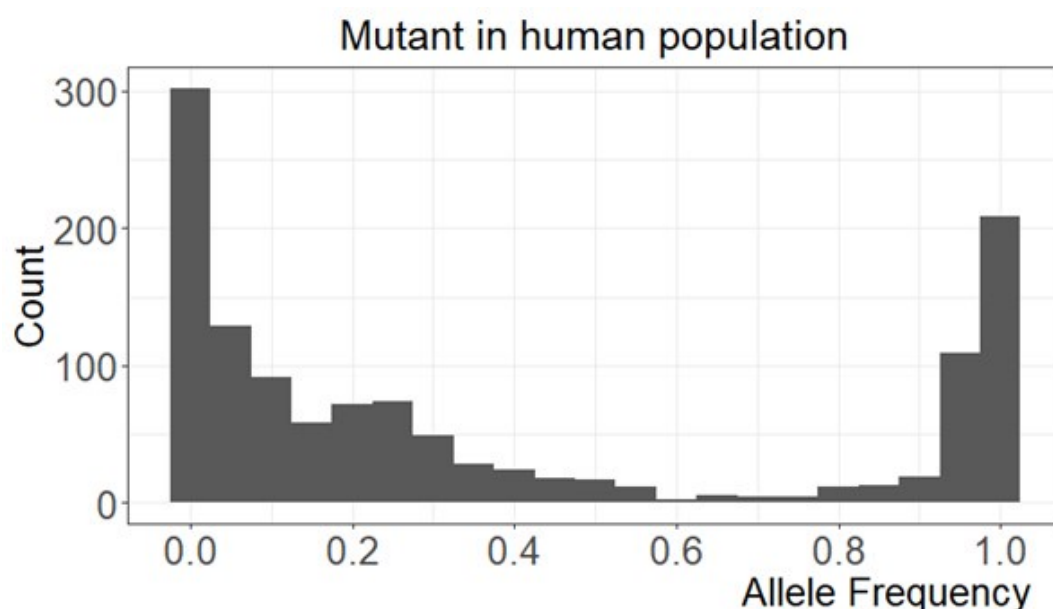
Considering the nature of multi-copy genes, where gene members undergo continuous exchanges at a much slower rate compared to the rapid rate of random distribution of chromosomes at each generation of sexual reproduction, even a few variant copies that arise during an individual's lifetime would disperse into the gene pool in the next generation (Ohta

and Dover 1984). Thus, there is minimal difference between individuals. Our analysis is also aligns with this theory, particularly in human population ($FIS = 0.059$), where each individual carries the majority of the population's genetic diversity. Therefore, even a single chimpanzee or gorilla individual carries sufficient diversity with its hundreds of gene copies to calculate divergence with humans.

(vii) Fixed variants are operationally defined as those with a frequency >0.8 in one species. What is the justification for this choice of threshold? Without knowing the exact sample size of the various species, it's difficult to assess whether this threshold is appropriate.

Response 14: First, the mutation frequency distribution is strongly bimodal (see Figure below) with a peak at zero and the other at 1. This high frequency peak starts to rise slowly at 0.8, similar to FST distribution in Figure 4C. That is why we use it as the cutoff although we would get similar results at the cutoff of 0.90 (see Table below). Second, the sample size for the calculation of mutant frequency is based on the number of reads which is usually in the tens of thousands. Third, it does not matter if the mutation frequency calculation is based on one individuals or multiple individuals because 95% of the genetic diversity of the population is captured by the gene pool within each individual.

Author response image 1.



Author response table 1.

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

		Direction of changes	Not fixed mutation counts	Fixed mutation counts	Chi-square test
Human		A/T to G/C	179	207	P = 2.247e-14
		G/C to A/T	164	43	
Chimpanzee		A/T to G/C	313	186	P = 0.004864
		G/C to A/T	207	77	
M. domesticus	m.	A/T to G/C	105	4	P = 0.3555
		G/C to A/T	158	12	

New mutants with a frequency >0.9 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: 58.303 for human, 7.9292 for chimpanzee, and 0.85385 for M. m. domesticus.

(viii) It is not explained exactly how FIS, FST, and divergence levels of rRNA genes were calculated from variant frequency at individual and species levels. Formulae need to be provided to explain the computation.

Response 15: After we clearly defined the HI, HS, and HT in Response9, understanding FIS and F_ST_ becomes straightforward.

“Given the three levels of heterozygosity, there are two levels of differentiation. First, FIS is the differentiation among individuals within the species, defined by

$$FIS = [HS - HI]/HS$$

FIS is hence the proportion of genetic diversity in the species that is found only between individuals. We will later show FIS ~ 0.05 in human rDNA (Table 2), meaning 95% of rDNA diversity is found within individuals.

Second, FST is the differentiation between species within the total species complex, defined as

$$FST = [HT - HS]/HT$$

FST is the proportion of genetic diversity in the total data that is found only between species.”

(3) Complete ignorance of the difference in mutation rate difference between rRNA genes and genome-wide average

Nearly all data analysis in this paper relied on comparison between rRNA genes with the rest (presumably single-copy part) of the genome. However, mutation rate, a key parameter determining the diversity and divergence levels, was completely ignored in the comparison. It is well known that mutation rate differs tremendously along the genome, with both fine and large-scale variation. If the mutation rate of rRNA genes differs substantially from the genome average, it would invalidate almost all of the analysis results. Yet no discussion or justification was provided.

Response 16: We appreciate the reviewer's observation regarding the potential impact of varying mutation rates across the genome. To address this concern, we compared the long-

term substitution rates on rDNA and single-copy genes between human and rhesus macaque, which diverged approximately 25 million years ago. Our analysis (see Table S1 below) indicates that the substitution rate in rDNA is actually slower than the genome-wide average. This finding suggests that rRNA genes do not experience a higher mutation rate compared to single-copy genes, as stated in the text:

“Note that Eq. (3) shows that the mutation rate, m , determines the long-term evolutionary rate, l . Since we will compare the l values between rRNA and single-copy genes, we have to compare their mutation rates first by analyzing their long-term evolution. As shown in Table S1, l 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 l values will have to be explained by other means.”

However, given the divergence time (T_d) being equal to or smaller than T_f , even if the mutation rate per nucleotide is substantially higher in rRNA genes, these variants would not become fixed after the divergence of humans and chimpanzees without the help of strong homogenization forces. Thus, the presence of divergence sites (Table 5) still supports the conclusion that rRNA genes undergo much stronger genetic drift compared to single-copy genes.

Related to mutation rate: given the hypermutability of CpG sites, it is surprising that the evolution/fixation rate of rRNA estimated with or without CpG sites is so close (2.24% vs 2.27%). Given the 10 - 20-fold higher mutation rate at CpG sites in the human genome, and 2% CpG density (which is probably an under-estimate for rDNA), we expect the former to be at least 20% higher than the latter.

Response 17: While it is true that CpG sites exhibit a 10-20-fold higher mutation rate, the close evolution/fixation rates of rDNA with and without CpG sites (2.24% vs 2.27%) may be attributed to the fact that fixation rates during short-term evolutionary processes are less influenced by mutation rates alone. As observed in the Human-Macaque comparison in the table above, the substitution rate of rDNA in non-functional regions with CpG sites is 4.18%, while it is 3.35% without CpG sites, aligning with your expectation of 25% higher rates where CpG sites are involved.

This discrepancy between the expected and observed fixation rates may be due to strong homogenization forces, which can rapidly fix or eliminate variants, thereby reducing the overall impact of higher mutation rates at CpG sites on the observed fixation rate. This suggests that the homogenization mechanisms play a more dominant role in the fixation process over short evolutionary timescales, mitigating the expected increase in fixation rates due to CpG hypermutability.

Among the weaknesses above, concern (1) can be addressed with clarification, but concerns (2) and (3) invalidate almost all findings from the data analysis and cannot be easily alleviated with a complete revamp work.

Recommendations for the authors:

Reviewing Editor Comments:

Both reviewers found the manuscript confusing and raised serious concerns. They pointed out a lack of engagement with previous literature on modeling and the presence of ill-defined terms within the model, which obscure understanding. They also noted a significant disconnection between the modeling approach and the biological processes involved. Additionally, the data analysis was deemed problematic due to the failure to consider essential biological and technical factors. One reviewer suggested that the

modeling component would be more suitable as a section of the companion theory paper rather than a standalone paper. Please see their individual reviews for their overall assessment.

Reviewer #2 (Recommendations For The Authors):

Beyond my major concerns, I have numerous questions about the interpretation of various findings:

Lines 62-63: Please explain under what circumstance $N_e = N/V(K)$ is biologically nonsensical and why.

Response 18: “Biologically non-sensical” is the term used in (Chen, et al. 2017). We now used the term “biologically untenable” but the message is the same. How does one get $V(K) \neq E(K)$ in the WF sampling? It is untenable under the WF structure. Kimura may be the first one to introduce $V(K) \neq E(K)$ into the WF model and subsequent papers use the same sort of modifications that are mathematically valid but biologically dubious. As explained extensively in the companion paper, the modifications add complexities but do not give the WF models powers to explain the paradoxes.

Lines 231-234: The claim about a lower molecular evolution rate (λ) is inaccurate - under neutrality, the molecular evolution rate is always the same as the mutation rate. It is true that when the species divergence T_d is not much greater than fixation time T_f , the observed number of fixed differences would be substantially smaller than $2\mu T_d$, but the lower divergence level does not mean that the molecular evolution is slower. In other words, in calculating the divergence level, it is the time term that needs to be adjusted rather than the molecular evolution rate.

Response 19: Thanks, we agree that the original wording was not accurate. It is indeed the substitution rate rather than the molecular evolution rate that is affected when species divergence time T_d is not much greater than the fixation time T_f . We have revised the relevant text in the manuscript to correct this and ensure clarity.

Lines 277-279: Hs for rRNA is 5.2x fold than the genome average. This could be roughly translated as $N_e^/N_e = 5.2$. According to Eq 2: $(1/N_e^*)/(1/N_e) = V_h/C^*$, it can be derived that mean $N_e^*/N_e = C^*/V_h$. Then why do the authors conclude " $C^* = N^*/N \sim 5.2$ " in line 278? Wouldn't it mean that C^*/V_h is roughly 5.2?*

Response 20: We apologize for the confusion. To prevent misunderstandings, we have revised Equation 1 and deleted Equation 2 from the manuscript. Please refer to the Response6 for further details.

Lines 291-292: What does "a major role of stage I evolution" mean? How does it lead to lower FIS?

Response 21: We apologize for the lack of clarity in our original description, and we have revised the relevant content to make them more directly.

“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).”

“FIS 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. ... Correlation of polymorphic sites in IGS region is shown in Supplementary Fig. 1. The results suggest that the genetic drift due to the sampling of chromosomes during

sexual reproduction (e.g., segregation and assortment) is augmented substantially by the effects of homogenization process within individual. Like those in mice, the pattern indicates that intra-species polymorphism is mainly preserved within individuals.”

Line 297-300: why does the concentration at very allele frequency indicate rapid homogenization across copies? Suppose there is no inter-copy homogenization, and each copy evolves independently, wouldn't we still expect the SFS to be strongly skewed towards rare variants? It is completely unclear how homogenization processes are expected to affect the SFS.

Response 22: We appreciate the reviewer’s insightful comments and apologize for any confusion in our original explanation. To clarify:

If there is no inter-copy homogenization and each copy evolves independently, it would effectively result in an equivalent population size that is C times larger than that of single-copy genes. However, given the copies are distributed on five chromosomes, if the copies within a chromosome were fully linked, there would be no fixation at any sites. Considering the data presented in Table 4, where the substitution rate in rDNA is higher than in single-copy genes, this suggests that additional forces must be acting to homogenize the copies, even across non-homologous chromosomes.

Regarding the specific data presented in the Figure 3, the allele frequency spectrum is based on human polymorphism sites and is a folded spectrum, as the ancestral state of the alleles was not determined. High levels of homogenization would typically push variant mutations toward the extremes of the SFS, leading to fewer intermediate-frequency alleles and reduced heterozygosity. The statement that “allele frequency spectrum is highly concentrated at very low frequency within individuals” was intended to emphasize the localized distribution of variants and the high identity at each site. However, we recognize that it does not accurately reflect the role of homogenization and this conclusion cannot be directly inferred from the figure as presented. Therefore, we have removed the sentence in the text.

The evidence of gBGC in rRNA genes in great apes does not help explain the observed accelerated evolution of rDNA relative to the rest of the genome. Evidence of gBGC has been clearly demonstrated in a variety of species, including mice. It affects not only rRNA genes but also most parts of the genome, particularly regions with high recombination rates. In addition, gBGC increases the fixation probability of $W>S$ mutations but suppresses the fixation of $S>W$ mutations, so it is not obvious how gBGC will increase or decrease the molecular evolution rate overall.

Response 23: We have thoroughly rewritten the last section of Results. The earlier writing has misplaced the emphasis, raising many questions (as stated above). To answer them, we would have to present a new set of equations thus adding unnecessary complexities to the paper. Here is the streamlined and more logical flow of the new section.

First, Tables 4 and 5 have shown the accelerated evolution of the rRNA genes. We have now shown that rRNA genes do not have higher mutation rates. Below is copied from the revised text:

“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, m , determines the long-term evolutionary rate, l . Since we will compare the l values between rRNA and single-copy genes, we have to compare their mutation rates first by analyzing their long-term evolution. As shown in Table S1 l 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 l values will have to be explained by other means.”

Second, we have shown that the accelerated evolution in mice is likely due to genetic drift, resulting in faster fixation of neutral variants. We also show that this is unlikely to be true in humans and chimpanzees; hence selection is the only possible explanation. The section below is copied from the revised text. It shows the different patterns of gene conversions between mice and apes, in agreement with the results of Tables 4 and 5. In essence, it shows that the GC ratio in apes is shifting to a new equilibrium, which is equivalent to a new adaptive peak. Selection is driving the rDNA genes to move to the new adaptive peak.

Revision - “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.

1. 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 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 analyze 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.

Tables 6-7 here

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.”

References

- 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.
- 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.
- 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.
- 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.
- Crow JF, Kimura M. 2009. *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.
- Datson NA, Morsink MC, Atanasova S, Armstrong VW, Zischler H, Schlumbohm C, Dutilh BE, Huynen MA, Waagele B, Ruepp A, et al. 2007. Development of the first marmoset-specific DNA microarray (EUMAMA): a new genetic tool for large-scale expression profiling in a non-human primate. *Bmc Genomics* 8:190.
- 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.
- Eickbush TH, Eickbush DG. 2007. Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* 175:477-485.
- Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in Genetics* 23:273-277.
- Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, et al. 2007. Evolutionary and Biomedical Insights from the Rhesus Macaque Genome. *Science* 316:222-234.
- 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.
- Hartl DL, Clark AG, Clark AG. 1997. *Principles of population genetics*: Sinauer associates Sunderland.
- Hori Y, Shimamoto A, Kobayashi T. 2021. The human ribosomal DNA array is composed of highly homogenized tandem clusters. *Genome Res* 31:1971-1982.
- Jeffreys AJ, Neumann R. 2002. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet* 31:267-271.
- 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.

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, Bizikadze 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. 1985. A model of duplicative transposition and gene conversion for repetitive DNA families. *Genetics* 110:513-524.

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.

Ruan Y, Wang X, Hou M, Diao W, Xu S, Wen H, Wu C-I. 2024. Resolving Paradoxes in Molecular Evolution: The Integrated WF-Haldane (WFH) Model of Genetic Drift. *bioRxiv:2024.2002.2019.581083*.

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

Smith GP. 1976. Evolution of Repeated DNA Sequences by Unequal Crossover. *Science* 191:528-535.

Smith GP. 1974a. Unequal crossover and the evolution of multigene families. *Cold Spring Harbor symposia on quantitative biology* 38:507-513.

Smith GP. 1974b. Unequal Crossover and the Evolution of Multigene Families. 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.

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.

Wall JD, Frisse LA, Hudson RR, Di Rienzo A. 2003. Comparative linkage-disequilibrium analysis of the beta-globin hotspot in primates. *Am J Hum Genet* 73:1330-1340.

<https://doi.org/10.7554/eLife.99992.2.sa0>