The paradox of extremely fast evolution driven in multi-copy gene systems - A resolution

Xiaopei Wang[1], Yongsen Ruan[1], Lingjie Zhang[1], Xiangnyu Chen[1], Zongkun Shi[1], Haiyu Wang[1], Bingjie Chen[1], Miles Tracy[1], Haijun Wen[1,*], Chung-I Wu[1,*]

[1]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China

[*] Corresponding authors: wenhj5@mail.sysu.edu.cn (H. Wen); ciwu@uchicago.edu (C.I. Wu)

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 evolution in two stages should be far slower than single-copy systems but the opposite is often true. The paradox has been unresolvable because the standard Wright-Fisher (WF) model for molecular evolution cannot track multi-copy genes. We now apply the newly expanded WF-Haldane (WFH) model to such systems, first on ribosomal rRNA genes. On average, rRNAs have $C \sim 150$ - 300 copies per haploid in mammals. While a neutral mutation would take $4N$ (N being the population size) generations to become fixed, the time should be $4NC$ generations for rRNAs. However, the observed fixation time in mouse and human is $< 4N$; thus the paradox means, effectively, $C < 1$. Genetic drift (i.e., all random neutral forces driving molecular evolution by the WFH model) of rRNA genes appears 200-300 times higher than single-copy genes, thus reducing C to < 1. The large increases in genetic drift are driven by the homogenizing forces of unbiased gene conversion, unequal crossover and replication slippage within individuals. This study is one of the first applications of the WFH model to track random neutral forces of evolution. Such random forces, outside of the standard WF model, are often mis-interpreted as the working of natural selection.

Introduction

In this study, we focus on multi-copy gene systems, where the evolution takes place both within and between individuals in two stages. Multi-copy gene systems include viruses, transposons, mitochondria and multi-gene family (Alexandrov, et al. 2001; Szitenberg, et al. 2016; Xu, et al. 2019; Ruan, Wen, et al. 2021). Given the large number of copies in the entire population and the two-stage evolution, the neutral evolutionary rate should be much slower than in the standard single-copy systems. However, the rapid evolution of multi-copy gene systems has been extensively documented (Charlesworth, et al. 1994; Eickbush and Eickbush 2007; Jurka, et al. 2007; Hou, et al. 2023). The main reason for this paradox is that the speed of neutral evolution of multi-gene systems is beyond the power of the standard WF (Wright-Fisher) model of molecular evolution.

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, unbiased 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 their inferences are possible only when genetic drift is fully accounted for. Under-estimation of genetic drift has been the major cause of the over-estimation of other forces. In the companion study(Ruan, et al. 2024), we show that the conventional Wright-Fisher (WF) model may often under-account genetic drift.

Ruan, et al. (Ruan, et al. 2024) propose the integration of the WF model with the Haldane model of genetic drift, referred to as the WFH model. 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 thus 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 [see (Ruan, et al. 2024)])

In the WF model, gene frequency is governed by $1/N$ (or $1/2N$ in diploids) because K would follow the Poisson distribution whereby $V(K) = E(K)$. Since $E(K)$ is generally ~1 in the long run, $V(K)$ would be ~ 1 as well. (We note that $N_e = N/V(K)$ can be a solution for the WF model, but such a solution would be biologically nonsensical albeit mathematically feasible.) Instead, the WFH model is biologically based on the Haldane model, but with mathematical solutions approximated by the WF model when feasible.

The WF model has led to a number of paradoxes as discussed in (Ruan, et al. 2024). They include i) the paradox of changing N's whereby the strength of drift may increase as N increases; ii) the paradox of sex chromosomes whereby genetic drift depends on the sex, even when N is fully accounted for; iii) the paradox of drift under selection whereby genetic drift depends on $V(K)$ but not on N. These paradoxes are resolved by the integrated WFH model.

In addition to the many paradoxes, there exist a number of genetic systems that are

incompatible with the WF model. The influence of stochastic forces are hence rarely defined in such systems. These may include somatic cell evolution (Ling, et al. 2015; Wu, et al. 2016; Chen, Wu, et al. 2022), domestication (Wang, et al. 2014; Ostrander, et al. 2019) and even interspecific competitions, which do have stochastic elements as well.

A common multi-copy system is viral evolution, evident in the recent pandemic of COVID-19 (Ruan, Luo, et al. 2021; Ruan, Wen, et al. 2021; Ruan, et al. 2022; Hou, et al. 2023). Another example is systems of large number of nearly-identical genes within the same genome such as transposons (Szitenberg, et al. 2016), mitochondrial DNAs (Xu, et al. 2019), satellite DNAs (Wu, et al. 1989; Cabot, et al. 1993; Alexandrov, et al. 2001), and ribosomal RNA genes (van Sluis and McStay 2019; Hori, et al. 2021). In these examples, 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; Tao, et al. 2001; Wyckoff, et al. 2002; McDermott and Noor 2010). The WF model has been noted to be inadequate even in such diploid systems (Charlesworth 2009; Charlesworth and Charlesworth 2010; Chen, et al. 2017). After all, the WF model is an essentially a haloid model with $2N$ copies in the population.

We now apply the Haldane model to track the genetic drift in multi-copy gene systems, using ribosomal RNA (rRNA) genes as an example. The absence of an adequate model has led to speculations of adaptation by natural selection (Dover 1982; Lu and Wu 2005; Nair, et al. 2008; Arkhipova 2018; Chen, Yang, et al. 2022). We will show that neutral stochasticity would be a simpler explanation if the more powerful Haldane model is adopted. The analysis also reveals a new paradox of genetic drift; namely, the "effective population size" within individuals can in fact be smaller than 1.

Results

PART I presents a best-known multi-gene system of the rRNA genes. PART II (Theory) consolidates aspects of the Haldane model with respect 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, $C \sim 150$ 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.

Fig. 1A-D here

Each copy of rRNA gene has functional and non-functional parts 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 since the divergence of animals and plants (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). This suggests that 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 "chromosome community" of ribosomal RNA genes

While a human individual with 300 rRNA genes may appear to have 300 loci, the concept of distinct "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 all other copies on the same or 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, thus accounting for the copy number variation among individuals. These mechanisms will be referred to collectively as the homogenization process, resulting in a non-Mendelian pattern of inheritance.

Copies of the cluster on the same chromosome are widely known to be nearly identical in sequences (Hori, et al. 2021; Nurk, et al. 2022). Previous research has also provided extensive evidence that rDNA clusters among nonhomologous chromosomes share the same variants, thus confirming genetic exchange between chromosomes (Krystal, et al. 1981; Arnheim, et al. 1982; van Sluis, et al. 2019). For example, Nurk, et al. (Nurk, et al. 2022) report a 5-kb segment that covers mostly non-functional sites showing a 98.7% identity among the short arms of the five acrocentric chromosomes. Additionally, variants are often found in all human rDNAs but not in other great apes, suggesting rapid fixation (Arnheim, et al. 1980). The rapid fixation is unexpected because, for a new rDNA mutation to be fixed, it has to be fixed among all ~ 300 copies in all individuals in the population. In this process, both rapid drift and positive selection can contribute to these fixations. However, the main difference between them is that genetic drift only shorten the fixation time of mutations without changing their fixation probability, whereas selection can alter both. As a result, rapid drift can only accelerate the evolution rate up to neutral molecular evolutionary rate (Eq. (3) below), while positive selection is not constrained by this limit.

With the extensive exchanges via the homogenization process, copies of rRNA genes in an individual can be treated as a sub-population of gene copies. Such sub-populations do not

have to be in the sense of a Mendelian population when using the branching process to track genetic drift (see Discussion and Chen et al. 2017). (Guarracino, et al. 2023) recently proposed the view of "chromosome community" for the short arms of the 5 acrocentric chromosomes, where rRNA genes are located. As seen in Fig. 1C, these five short arms harbor a pool of rRNA genes that do not have strong associations with genomic locations. Fig. 1C further illustrates that each human individual's rRNA gene pool can capture ~ 95% of the total diversity of the human populations (see PART III). Fig. 1D presents the molecular mechanisms of genetic drift within individuals (i.e., Stage I evolution) in relation to the multi-copy gene systems.

PART II - Theory

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

The Haldane model of genetic drift based on the branching process is intuitively more 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 WF model, $V(K) = E(K)$ whereas $V(K)$ is decoupled from $E(K)$ in the Haldane model; the latter thus being more flexible.

Below, we compare the strength of genetic drift in rRNA genes vs. that of single-copy genes using the branching process. $E(K)$ is set to 1 so that we can focus on $V(K)$. For single-copy genes, the effective population size that determines genetic drift would be $N_e = N/V(K)$ (Kimura and Crow 1963). Details are presented in the companion paper (Ruan, et al. 2024). We shall use * to designate the equivalent symbols for rRNA genes. Hence,

$$N_e^* = N^*/V^*(K) = C^*N / V^*(K) \qquad \text{Eq. (1)}$$

where N^* is the number of rRNA genes in the entire population and, therefore, $N^* = C^*N$. For rRNA genes, $C^* >> 1$ but < 150, C^* being the effective copy number of rRNA genes. (C =1 for single copy genes.) The value of C^* is uncertain because each haploid in humans, on average, has ~ 150 copies that do not segregate but assort among chromosomes (Parks, et al. 2018). $V^*(K)$ for rRNA genes is likely to be much larger than $V(K)$ because K for rRNA mutations may be affected by a host of factors including replication slippage, unequal cross-over, gene conversion and other related mechanisms not operating on single copy genes. We thus designate $Vh = V^*(K)/V(K)$, as the homogenizing factor of rRNA genes.

Fig. 1D is a simple illustration that the homogenizing process may enhance $V(K)$ substantially over the conventional WF model as each mechanism amplifies the variance in K. Importantly, the homogenization process is considered stochastic rather than deterministic because it affects both the mutants and wild type equally, as expected of genetic drift.

We then designate $R^* = (1/N_e^*)/(1/N_e)$ as the strength of genetic drift of rRNA genes, relative to that of single copy genes. R^* is measurable, as will be done below.

$$R^* = (1/N_e^*)/(1/N_e) = [V^*(K)/ V(K)]/C^* = Vh/C^* \qquad \text{Eq. (2)}$$

In short, R^* is the ratio of two factors. A larger Vh [$=V^*(K)/V(K)$] speeds up the drift strength of rRNA genes whereas a larger C^* slows down the process of fixation by drift. It would be most interesting if $R^* > 1$ as it would mean that the effective population size of rRNA genes is smaller than single-copy genes. In this case, the homogenization force should be strong enough to overcome the large copy number of rRNA genes (~ 150) during the fixation of new mutations.

2. rDNA polymorphism within species

For rDNAs, there are 3 levels of diversity: H_I, H_S and H_T. They represent, respectively, the heterozygosity within individual (I), within species (S) and in the total data set (T, which often comprises two species or more). Briefly, $H = 2p(1-p)$ if there are two variants with the frequency of p and $1-p$. In an equilibrium haploid population,

$$H_S = 2N_e u/[2N_e u + 1]$$

Given u, H_S is a measure of genetic drift within a population. For rRNA genes, H_S is expected to be larger than that single-copy genes as $N^* > N$. With the three levels of heterozygosity for multi-copy genes, there are two levels of differentiation. First, F_{IS} is the differentiation among individuals within the species, defined by

$$F_{IS} = [H_S - H_I]/H_S \qquad (1)$$

Second, F_{ST} is the differentiation between populations (or species) in the context of total diversity, defined as

$$F_{ST} = [H_T - H_S]/H_T \qquad (2)$$

The two F statistics will inform about the distribution of nucleotide diversity among individuals, or between species. H_T is the expected heterozygosity when there is no subdivision between species and H_S is the observed heterozygosity. Therefore, F_{ST} will be equal to 0 when there is no subdivision and will be positive as long as there is variation in allele frequency across subdivisions.

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 is (λ), although driven by mutation and genetic drift, is generally shown by Eq. (3) below (Crow and Kimura 1970; Hartl, et al. 1997; Li 1997):

$$\lambda = Nu[1/N] = u \qquad \text{Eq. (3).}$$

Hence, we might view the evolution of rRNA genes in the long run to be the same as single-copy genes. However, Eq. (3) is valid only 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, λ 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. Thus, λ may be at least 80% lower than calculated for single-copy genes.

Fig. 2 here

Our own theoretical derivation has shown that the fixation time for rRNA genes is close to $4N_e^*$ as is the case for single-copy genes at ~ $4N_e$. If $V^*(K)$ is sufficiently larger than $V(K)$, it is in fact possible for $N_e^* < N_e$ such that genetic drift is stronger for rRNA genes than for single-copy genes and $T_f^* < T_f$. Therefore, T_{f1} can represent the T_f^* of mutations in rRNA genes in Fig. 2. This is interesting because, if the homogenization is powerful enough, rRNA genes would have an effective copy number of $C^* < 1$. A short T_f^*, which can be obtained by large $V^*(K)$, would lead to a small T_f^*/T_d and thus a higher λ, particularly in short-term evolution. However, even T_f^* approaches 0, λ 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 orders faster than in Stage II. Therefore, despite having several hundred copies of rRNA genes per genome, the speed of genetic drift in rRNA genes may not be that much slower than in single-copy genes. This postulate will be tested below.

1. rDNA polymorphism within species

1) Polymorphism in mice

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

H_S for *M. m. domesticus* single-copy genes is roughly 1.40 per kb genome-wide (Geraldes, et al. 2008) while H_S for rRNA genes is 7.25 per kb, 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.

Tables 1- 2 and Fig. 3 here

2) Polymorphism in human

F_{IS} for rDNA among 8 human individuals is 0.059, much smaller than 0.142 in *M. m. domesticus* mice, suggesting a major role of stage I evolution in humans. Consisted with low F_{IS}, Fig. 3 shows strong correlation of the polymorphic site frequency of rDNA transcribed region among each pair of individuals from three continents (2 Asians, 2 Europeans and 2 Africans). Correlation of polymorphic sites in IGS region is shown in Supplementary Fig. 1. Like those in mice, the pattern indicates that intra-species polymorphism is mainly preserved within individuals. In addition, the 6 panels on the diagonal line given in Fig. 3 demonstrated that the allele frequency spectrum is highly concentrated at very low frequency within individuals, indicating rapid homogenization of rDNA mutations among different copies even different chromosomes. This homogenization can lead to the elimination of the genetic variation of this region. 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.

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 S1- S2.

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

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

Although T_f and T_f^* estimates are less reliable than T_d estimates, both comparisons use the same T_f and T_f^* from *M. m. domesticus*. Hence, the results should be qualitatively unbiased. For a demonstration, we shall use the estimates of T_f (i.e., the coalescence time) at 0.2 Myrs for single-copy genes by using an average nucleotide diversity of 0.0014 and the mutation rate of 5.7×10^{-9} per base pair per generation (Geraldes, et al. 2008; Phifer-Rixey, et al. 2020). Based on the estimated C^* above, we obtain T_f^* for rDNA mutations at 5×0.2 Myrs, or 1 Myrs. While some have estimated T_f to be close to 0.4 Myrs (Fujiwara, et al. 2022), we aim to show that the pattern of reduction in rRNA divergence is true even with the conservative estimates.

Between *M. m. domesticus* and *M. m. castaneus*, the reduction in substitution rate for single copy gene should be ~ 40% (T_f/T_d = 0.2/0.5), and the reduction for rRNA genes should be

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

Tables 3, 4, 5 and Fig. 4 here

Between *Mus m. domesticus* and *Mus spretus*, the reduction in actual evolution rate from theoretical limit for single-copy genes should be 6.7% (T_f/T_d = 0.2/3) and, for rRNA genes, should be 33% (T_f^*/T_d = 1/3). The evolutionary rate (i.e. the fixation rate) of IGS region is lower than single-copy genes, 0.01 in IGS and 0.021 in genome-wide (Table 4), as one would expect. However, ETS and ITS regions have evolved at a surprising rate that is 12% higher than single-copy genes. Note that the reduction in C^*, even to the lowest limit of C^* =1, would only elevate the rate of fixation in rRNA genes to a parity with single-copy genes. From Eq. 2, the explanation would be that the homogenization factor $V^*(K)/V(K)$ has to be very large, such that C^* is < 1. With such rapid homogenization, the fixation time approaches 0 and the evolutionary rate in rRNA genes can indeed reach the theoretical limit of Eq. (3). In such a scenario, the substitution rate in ETS and ITS, compared to single-copy genes in mice, may increase by 7%, $T_f/(T_d - T_f)$ = 0.2/(3-0.2). If we use $T_f \sim$ 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 long-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 that are considered to have a ten-fold higher mutation rate, their fixation rate still reaches 22.4‰. In this case, even if C^* <<1, the extremely rapid fixation will only increase the substitution rate by $T_f/(T_d - T_f)$ by 11%, compared to single-copy genes. Thus, the much accelerated evolution of rRNA genes between humans and chimpanzees cannot be entirely attributed to genetic drift. Other forces, such as biased gene conversion, must have contributed to this acceleration as well.

3) Biased gene conversion - Positive selection for rRNA mutations in apes, but not in mice

As stated, gene conversion should increase $V(K)$ and speed up genetic drift. Conversion is commonly observed in highly recombinogenic regions, especially in repetitive sequences. If gene conversion is unbiased, the effect on the rate of divergence between species would be quite modest unless $T_d \leq T_f$. This is because unbiased conversion would only reduce the fixation time, T_f, as demonstrated in Fig. 2.

In this context, it is necessary to define unbiased gene conversion as it has different meanings at the site level vs. at the gene level. At the site level, when an AT site is paired with a GC site, it has been commonly reported that gene conversion would favor AT-to-GC over GC-to-AT conversion (Jeffreys and Neumann 2002; Meunier and Duret 2004). We first designate the proportion of AT-to-GC conversion as f and the reciprocal, GC-to-AT, as g. Given $f \neq g$, this bias is true at the site level. However, at the genic level, the conversion may in fact be unbiased for the following reason. 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 $p/q = g/f$. Thus, gene conversion at the genic level would be unbiased when the AT/GC ratio is near the equilibrium. In other words, there is no selection driving either AT-to-GC or GC-to-AT changes.

We now show that *genic* conversion is indeed unbiased (or nearly unbiased) between mouse species. In contrast, between human and chimpanzee, the fluxes are unequal. It appears that selection has favored the AT-to-GC changes in the last 6 million years that separate the two species.

In these genic analyses, we first determine whether there is biased conversion in the human lineage (Brown and Jiricny 1989; Galtier and Duret 2007). Using the same variant found in chimpanzees and gorillas as the ancestral state, 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 ($p < 0.001$). Between *M. m. domesticus* and *M. m. castaneus* (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 as shown in Table 7. In particular, the $(pf)/(qg)$ ratio is much larger than 1 in both the human and chimpanzee lineages. In contrast, the ratio in mouse is not significantly different from 1. The ratio for total rDNAs in mouse is lower than 1, but that is likely due to the influence of selective constraints on the functional parts of these genes.

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 is recombination that may lead to gene conversion or 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 copes 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 ~ $150N$ 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.

Materials and Methods

Data Collection.

We collected high-coverage whole-genome sequencing data for our study. The genome sequences of human, chimpanzee and gorilla were sourced from National Center for Biotechnology Information (NCBI). Human individuals were drawn from diverse geographical origins encompassing three continents (4 Asians, 2 Europeans and 2 Africans) and Asia CRC was the normal tissue of Case 1 patient from (Chen, Wu, et al. 2022). Genomic sequences of mice were sourced from the Wellcome Sanger Institute's Mouse Genome Project (MGP). Although some artificial selection have been performed on laboratory mouse strains (Yang, et al. 2011), the WSB/EiJ, ZALENDE/Ei and LEWES/EiJ strains were derived from wild populations. Incorporating these wild-derived laboratory strains, along with other inbred strains, a cohort of 10 mice was utilized to approximate the population representation of *M. m. domesticus*. Furthermore, the low F_{IS} of 0.14 for rDNA in *M. m. domesticus* found in this study suggests that each mouse covers 86% of population's genetic diversity, thereby mitigating concerns about potential sampling biases. Accessions and the detailed information of samples used in this study are list in Supplementary Table 4 and Table 5.

Data Processing.

Following adapter trimming and the removal of low-quality sequences, these whole-genome sequencing data of apes and mice were mapped against respective reference sequences: the human rDNA reference sequence (Human ribosomal DNA complete repeating unit, GenBank: U13369.1) and the mouse rDNA reference sequence (*Mus musculus* ribosomal DNA, complete repeating unit, GenBank: BK000964.3). Alignment was performed using Burrows-Wheeler-Alignment Tool v0.7.17 (Li and Durbin 2009) with default parameters. Each individual was considered as a pool of rRNA genes, thus there were three levels of diversity.

To determine the variant frequency within individual, bcftools (Danecek, et al. 2021) was employed to call variants from multiple alignment files. Per-sample allele depths were obtained using following settings 'bcftools mpileup --redo-BAQ --max-depth 50000 --per-sample-mF --annotate FORMAT/AD,FORMAT/ADF,FORMAT/ADR,FORMAT/DP,INFO/AD,INFO/ADF,INFO/ADR' and 'bcftools call -mv'. The sites with depth <10 were filtered out in subsequent analysis, and the average of the minimum depths for each remaining site across all samples is above 3000, thereby enhancing the robustness of variant frequency estimation. In this process, we can identity all the polymorphic sites present in these samples and the variant frequency within each individual. The population-level variant frequency was computed by averaging variant frequencies across all individuals. Our analysis specifically focused on single nucleotide variants (SNVs), while other mutation types were discarded for consistency.

Identification of Divergence Sites.

Polymorphism represents the transient phase preceding divergence. With variant frequency within individual, within species, and between species, we obtained the F_{IS} and F_{ST} values for

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

F_{ST} analysis between human and chimpanzee was conducted for each human individual, summarized in Table 5. We identified a range from 672 to 705 sites with F_{ST} values above 0.8 across individuals, depicting robust divergence sites. Considering the high mutation rate in CpG sites (Ehrlich and Wang 1981; Sved and Bird 1990) and predominantly GC content in rDNA (*e.g.* 58% GC in human), we further estimated the evolutionary rate at non-CpG sites during the interspecies divergence. To achieve this, mutations in CpG sites were manually removed by excluding all sites contained 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, variants shared between chimpanzees and gorillas, humans and gorillas, *M. m. castaneus* and *M. spretus* were considered as the ancestral state for human, chimpanzee, and *M. m. domesticus*, respectively. We identified the derived variants that became nearly fixed (frequency > 0.8) in the lineages of humans, chimpanzees and *M. m. domesticus*. 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.

The alternative hypotheses of biased mutation pattern (Wolfe, et al. 1989; Francino and Ochman 1999) alone can be rejected in this study. According to the prediction of the mutational mechanism hypotheses, AT or GC variants should have equal fixation probabilities. We quantified the nearly fixed number of AT-to-GC and GC-to-AT types of mutations and conducted a chi-square test to assess their fixation probabilities. Notably, we found the GC variant had significant higher fixation probability than AT at site level in apes, but not in mice. Since the fixed sites accounted for less than 1% of non-functional length of rDNA, we calculated the AT-to-GC conversion rate, f, by dividing the number of fixed G or C sites by the number of AT base pairs, and the GC-to-AT conversion rate, g, was computed using a comparable approach. In order to test whether there is an equilibrium state at the genic level in these three species, we computed the $(pf)/(qg)$ ratio in Table 7, and a significant deviation of the ratio from 1 would imply biased genic conversion.

Data Availability

No new data were generated in this study. The genomic data used in this study are available from National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/) and the Mouse Genomes Project (https://www.sanger.ac.uk/data/mouse-genomes-project/). The specific accession numbers recorded in Supplementary Table S3 and S4.

Acknowledgments

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.

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, Krystal M, Schmickel R, Wilson G, Ryder O, Zimmer E. 1980. Molecular evidence for genetic exchanges among ribosomal genes on nonhomologous chromosomes in man and apes. Proc Natl Acad Sci U S A 77:7323-7327.

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.
Cazaux B, Catalan J, Veyrunes F, Douzery EJP, Britton-Davidian J. 2011. Are ribosomal DNA clusters rearrangement hotspots? A case study in the genus Mus (Rodentia, Muridae). BMC Evolutionary Biology 11:124.
Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics 10:195-205.
Charlesworth B, Charlesworth D. 2010. Elements of evolutionary genetics: Springer.
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:nwac250.
Chen Q, Yang H, Feng X, Chen Q, Shi S, Wu CI, He Z. 2022. Two decades of suspect evidence for adaptive molecular evolution-negative selection confounding positive-selection signals. Natl Sci Rev 9:nwab217.
Chen Y, Tong D, Wu CI. 2017. A New Formulation of Random Genetic Drift and Its Application to the Evolution of Cell Populations. Mol Biol Evol 34:2057-2064.
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, US A. 1970. An Introduction to Population Genetics Theory Harper & Row.
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.
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.201416878.
Gonzalez IL, Sylvester JE. 2001. Human rDNA: Evolutionary Patterns within the Genes and Tandem Arrays Derived from Multiple Chromosomes. Genomics 73:255-263.
Guarracino A, Buonaiuto S, de Lima LG, Potapova T, Rhie A, Koren S, Rubinstein B, Fischer C, Abel HJ, Antonacci-Fulton LL, et al. 2023. Recombination between heterologous human acrocentric chromosomes. Nature 617:335-343.
Guillén AKZ, Hirai Y, Tanoue T, Hirai H. 2004. Transcriptional repression mechanisms of nucleolus organizer regions (NORs) in humans and chimpanzees. Chromosome Research 12:225-237.
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, Shi JR, Gong ZK, Wen HJ, Lan Y, Deng XZ, Fan QH, Li JJ, Jiang ML, Tang XP, et al. 2023. Intra- vs. Interhost Evolution of SARS-CoV-2 Driven by Uncorrelated Selection-The Evolution Thwarted. Molecular Biology and Evolution 40:msad204.

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.

Kimura M, Crow JF. 1963. THE MEASUREMENT OF EFFECTIVE POPULATION NUMBER. Evolution 17:279-288.

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, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. 2022. TimeTree 5: An Expanded Resource for Species Divergence Times. Mol Biol Evol 39.

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

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

Ling S, Hu Z, Yang Z, Yang F, Li Y, Lin P, Chen K, Dong L, Cao L, Tao Y, et al. 2015. Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. Proc Natl Acad Sci U S A 112:E6496-6505.

Lu J, Wu C-I. 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. Proceedings of the National Academy of Sciences 102:4063-4067.

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.

Nair S, Miller B, Barends M, Jaidee A, Patel J, Mayxay M, Newton P, Nosten F, Ferdig MT, Anderson TJC. 2008. Adaptive Copy Number Evolution in Malaria Parasites. Plos Genetics 4:e1000243.

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

Ohta T. 1976. Simple model for treating evolution of multigene families. Nature 263:74-76.

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

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

Ostrander EA, Wang G-D, Larson G, vonHoldt BM, Davis BW, Jagannathan V, Hitte C, Wayne RK, Zhang Y-P, Dog KC. 2019. Dog10K: an international sequencing effort to advance studies of canine domestication, phenotypes and health. National Science Review 6:810-824.

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

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

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

Ruan Y, Hou M, Tang X, He X, Lu X, Lu J, Wu CI, Wen H. 2022. The runaway evolution of SARS-CoV-2 leading to the highly evolved Delta strain. Mol Biol Evol.

Ruan Y, Luo Z, Tang X, Li G, Wen H, He X, Lu X, Lu J, Wu CI. 2021. On the founder effect in COVID-19 outbreaks: how many infected travelers may have started them all? Natl Sci Rev 8:nwaa246.

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.

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.

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. 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.

Sved J, Bird AJPotNAoS. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. 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.

Tao Y, Hartl DL, Laurie CC. 2001. Sex-ratio segregation distortion associated with reproductive isolation in Drosophila. Proceedings of the National Academy of Sciences 98:13183-13188.

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

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

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

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

Wang G-D, Xie H-B, Peng M-S, Irwin D, Zhang Y-P. 2014. Domestication Genomics: Evidence from Animals. Annual Review of Animal Biosciences 2:65-84.

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

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

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

Wu C-I, True JR, Johnson N. 1989. Fitness reduction associated with the deletion of a satellite DNA array.

Nature 341:248-251.
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.
Wyckoff GJ, Li J, Wu C-I. 2002. Molecular Evolution of Functional Genes on the Mammalian Y Chromosome. Molecular Biology and Evolution 19:1633-1636.
Xu J, Nuno K, Litzenburger UM, Qi Y, Corces MR, Majeti R, Chang HY. 2019. Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. Elife 8.
Yang H, Wang JR, Didion JP, Buus RJ, Bell TA, Welsh CE, Bonhomme F, Yu AH-T, Nachman MW, Pialek J, et al. 2011. Subspecific origin and haplotype diversity in the laboratory mouse. Nat Genet 43:648-655.
Zhao Z, Jin L, Fu Y-X, Ramsay M, Jenkins T, Leskinen E, Pamilo P, Trexler M, Patthy L, Jorde LB, et al. 2000. Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proceedings of the National Academy of Sciences 97:11354-11358.

Figures and Tables

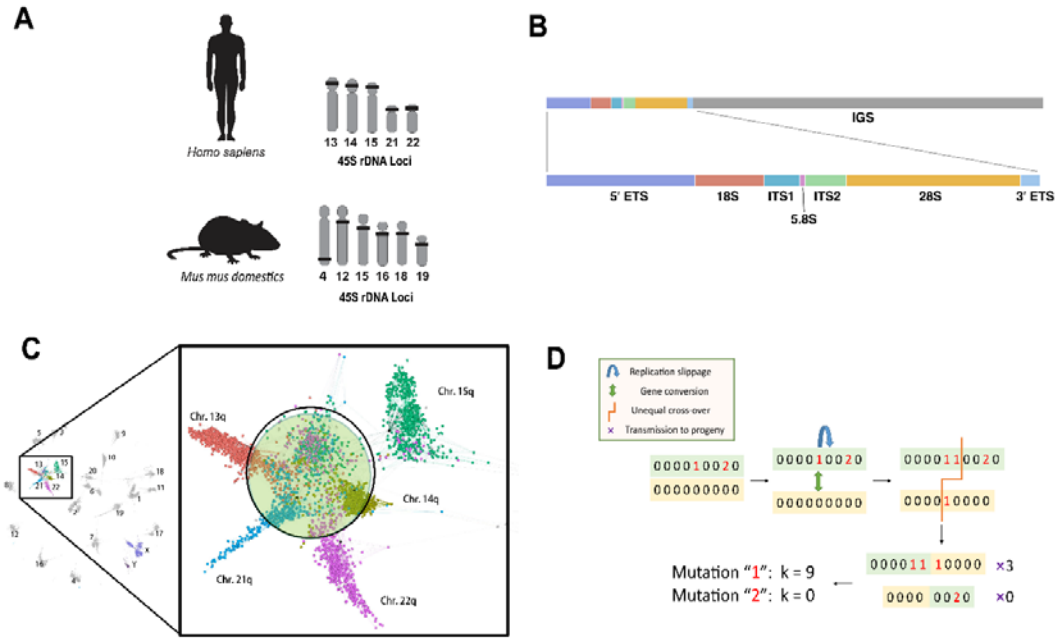

Figure 1. The "chromosome community of rDNAs on five acrocentric chromosomes. (A) The genomic locations of rDNA tandem repeats in human (Gibbons, et al. 2015) and mouse (Cazaux, et al. 2011). rDNAs are located on the short arms (human), or the proximal end of the long arms (mouse), of the acrocentric 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 "chromosomes community" map modified from Fig. 1 of (Guarracino, et al. 2023). The map shows the distance of divergence among chromosome segments. The large circle encompasses rDNAs from all 5 chromosomes with color dots indicating the chromosome origin. It is clear that rDNAs experience concerted evolution regardless of their genomic locations. 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.

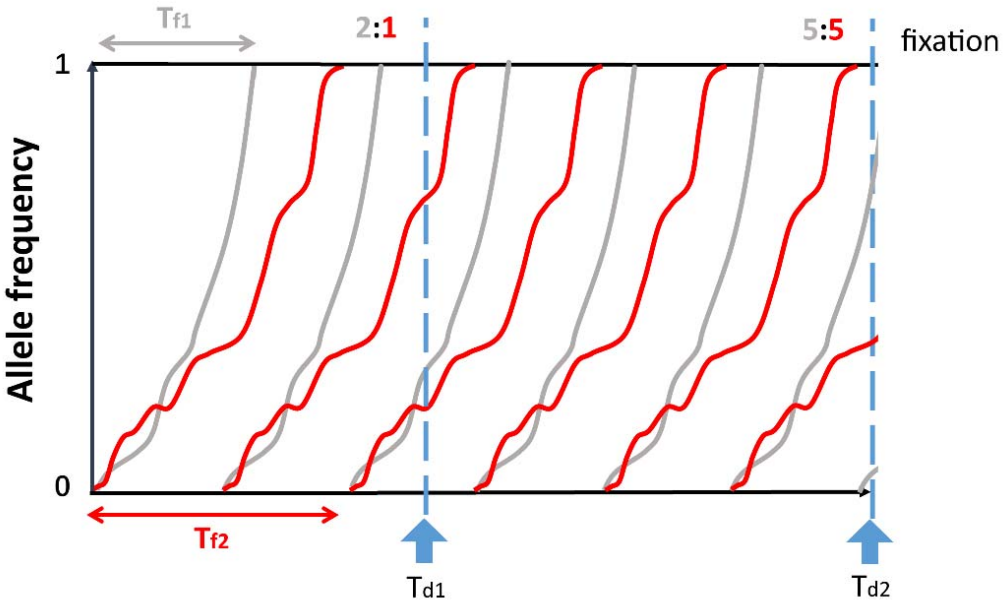

Figure 2. Fixation of mutations at two levels of species divergence, (T_{d1}) and (T_{d2}). (T_{f1}) and (T_{f2}) are mutations with a shorter and longer fixation time, respectively for single-copy and multi-copy genes. Note that mutations with a longer T_f would show a lower fixation rate in short-term evolution.

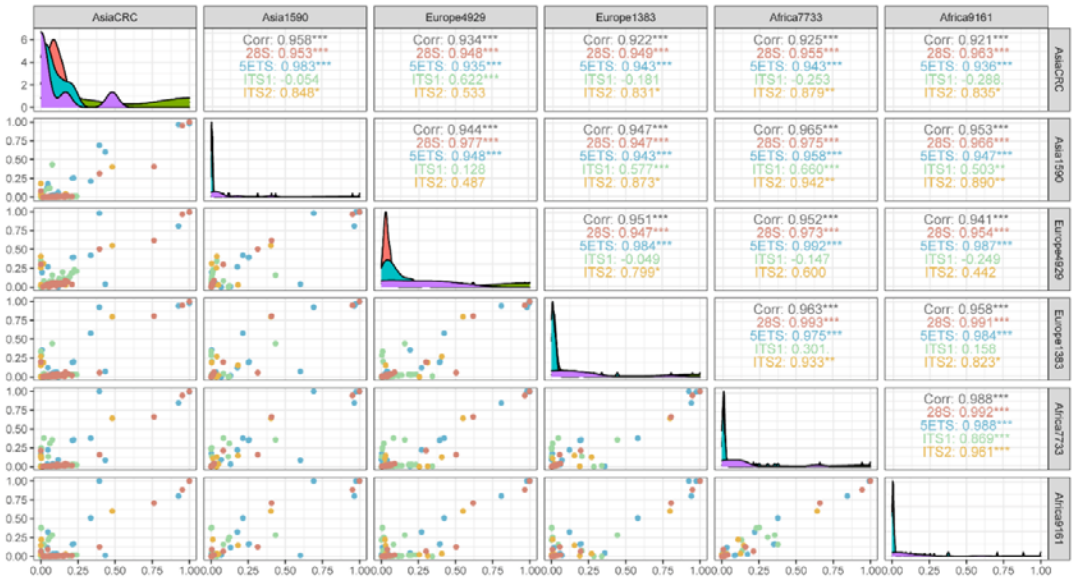

Figure 3. Correlation of variant frequencies between human individuals. The pairwise correlation of variant site frequency in the transcribed region of rDNAs among 6 individuals (2 Asians, 2 Europeans, and 2 Africans). The high correlations suggest that the diversity in each individual can well capture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

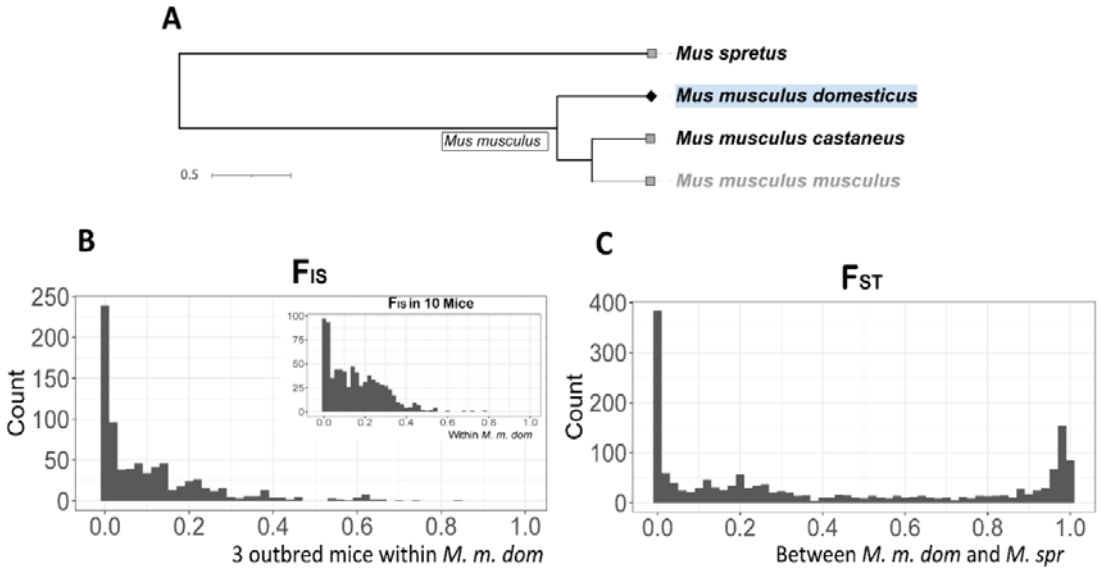

Figure 4. Levels of polymorphism and divergence in mice. (A) Phylogeny of *Mus musculus* and *Mus spretus* mice. The divergence times are obtained from http://timetree.org/. The line segment labelled 0.5 represents 0.5 Myrs. (B) F_{IS} distribution within *M. m. domesticus*. The distribution of F_{IS} for polymorphic sites in 3 outbred mouse strains or 10 mouse strains (including 7 inbred mice) in Table 1 (Inset) as shown. (C) F_{ST} distribution between *M. m. domesticus* and *Mus spretus*. Note that the *F* values rise above 0.8 only in (C).

Table 1. rRNA gene diversity in the 10 *M. m. domesticus* strains of a global collection.

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

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

Table 2. rRNA gene diversity in the 8 humans of a global collection.

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

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

Table 3. Divergence in rRNA genes between *M. m. domesticus* and *M. m. castaneus*.

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

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

Table 4. Divergence in rRNA genes between *M. m. domesticus* and *Mus spretus*.

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

Note the divergence of single-copy genes (21) is often smaller than rRNA genes (25.0 - 30.9) The divergence between *M. m. musculus* and the two outgroups species of *M. m. castaneus* and *Mus spretus* are listed in Supplementary Table S1 - S2.

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.

Table 6. 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	163	223	P = 1.559e-15
	G/C to A/T	159	48	
Chimpanzee	A/T to G/C	289	210	P = 0.000469
	G/C to A/T	201	83	
M. m. domesticus	A/T to G/C	99	10	P = 0.9568
	G/C to A/T	156	14	

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

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

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.