Effects of Different Mutation Rates and Randomness on the Distribution of Linkage Disequilibrium

Rong Zhang

Yangtze University

ZhiQiang Du (✉ zhqdu@yangtzeu.edu.cn)

Yangtze University

Research

Keywords: Linkage disequilibrium, Mutation, Transition, Transversion, Randomness

Posted Date: November 15th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1066165/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

1 **Effects of Different Mutation Rates and Randomness on the**
2 **Distribution of Linkage Disequilibrium**

3

4 Rong Zhang^{1,2}, Cai-Xia Yang^{1,2}, Zhi-Qiang Du^{1,2*}

5

6 1. College of Animal Science, Yangtze University, Jingzhou 434025, Hubei, China

7 2. Center of Animal Breeding Technology Innovation, Hubei Province, China

8

9 * Correspondence:

10 Email: zhqdu@yangtzeu.edu.cn

Abstract

Background: Mutation has recently received much attention on its role in the evolution and genetics of complex trait. The linkage disequilibrium (LD) distribution can be affected by mutation as reported recently, in which the same mutation rates were adopted in the transition matrix. However, effects of different types, rates and randomness of mutation on LD distribution remain unexplored.

Results: Here, we considered in the transition matrix mutations at each locus to be of different types and rates (i.e. nucleotide transition or transversion treated differently), to examine how the LD distribution between two genetic loci was affected. After examining consecutively factors such as effective population size, recombination and selection, different mutation types and rates could further change the dynamics of LD distribution. However, at the current scale of mutation rate (weak at 10^{-9} - 10^{-8}), mutation seemed to play only a minor role, compared to recombination and selection. A simple model further showed that mutation randomness increased the ruggedness of LD curves, which fluctuated around the steady state.

Conclusions: Taken together, different mutation rates and randomness could further disturb the dynamics of LD distribution. Our findings can help better understand the role of mutation in molecular evolution and complex trait genetics.

Keywords: Linkage disequilibrium, Mutation, Transition, Transversion, Randomness

Background

Linkage disequilibrium (LD) reflects the status of non-random association between alleles of two loci in the genome [1-3]. LD information are widely employed to make important discoveries, such as inferring key parameters (e.g. effective population size, N_e) related to population structure and allelic evolutionary history in population genetics and genomics; detecting genetic markers linked with causal variants, and predicting genomic breeding values of elite breeders in quantitative genetics; and identifying key genes implied in human diseases [4].

Different terms (statistics) have been proposed to measure LD, for instance, allelic covariance (D) or squared correlation (r^2) [1,4]. Many factors contribute to affecting LD status and are under intensive examination, such as population size, recombination, selection, mutation, migration between populations, and the analytical analysis and unbiased estimation on unknown allelic haplotype states [4-8]. Recently, as an important element to LD distribution, mutation is discovered to be involved in the dynamic change of important phenotype [9-11], accumulate under direct selection in laboratory experimental evolution [12-14], and genetic improvement of agricultural traits [15-17].

To study the effects of population size, mutation, and selection on LD distribution, a transition matrix approach was proposed recently, in which the same mutation rates (10^{-9}) for all nucleotides were considered [6]. However, different mutation types (transition or transversion) and rates could affect LD distribution. In the field of computational molecular evolution, the transition matrix has also been used to calculate

the evolutionary distances between molecular sequences (nucleic acids and protein), and to build phylogenetic trees [18,19]. A variety of statistical models have been proposed (from JC69, K80 models, to the UNREST model), based on the assumption that different nucleotides between themselves have different exchanging rates in the transition matrix [18,19]. Moreover, rates, spectrum and evolution of mutation were discovered to be species-specific, distributed in a wide range (10^{-7} - 10^{-10}), and selectively maintained [20-25], which relied also on nucleotide types, and status of the studied organism (health, nutrition, environment, etc.) [26-31]. Novel insights on mutation and its role in population genetics [32,33] and disease incidence (especially driver mutations in cancer progression) were obtained, in large-scale cancer genomics and in-depth data mining studies [34-38].

In the present study, we examined the pattern of LD distribution assuming different rates for different types of nucleotide mutations (transition or transversion), adopting an approach similar to the K80 model. We found that under the assumed range of mutation rates (i.e. weak scale at 10^{-9} - 10^{-8}), patterns and dynamics of LD distribution were further affected, but in a minor scale, considering consecutively effective population size, recombination, and selection. Furthermore, when randomness was added into a simple model, LD distribution was discovered to fluctuate around the steady state.

Materials and methods

Term and definition

Factors such as genetic drift, recombination, selection and mutation affect the distribution of LD (r^2) [39]. Two statistics widely used to quantify the LD at two loci (diallelic system) are:

$$D = P_{A_1B_1} - P_{A_1} * P_{B_1} \quad (1)$$

$$r^2 = \frac{D^2}{P_{A_1}(1-P_{A_1})P_{B_1}(1-P_{B_1})} \quad (2)$$

where P_{A_i} and P_{B_j} represent the frequencies of the i th and j th alleles at genetic loci A and B, respectively; $P_{A_iB_j}$, the frequency of haplotype A_iB_j . Sometimes D may be negative, therefore, r^2 is increasingly used as a metric to quantify LD [4,39].

Genetic drift and recombination act on the expectation of r^2 , when at equilibrium in a finite population it is approximately:

$$r_E^2 = \frac{1}{1+4N_e c} \quad (3)$$

where c is the recombination rate; N_e is the effective population size (the number of individuals that effectively participates in producing the next generation) [2-5].

Mutation affects also the approximation for the equilibrium value $E(r^2)$, which can be described by a quantity (σ_d^2) called the standard linkage deviation (i.e. the ratio of the expectations of the numerator and the denominator of r^2) [5,24].

$$\sigma_d^2 = \frac{E(D^2)}{E(P_{A_1}(1-P_{A_1})P_{B_1}(1-P_{B_1}))} \quad (4)$$

$$\sigma_d^2 = \frac{10+\rho+4\theta}{22+13\rho+32\theta+\rho^2+6\rho\theta+8\theta^2} \quad (5)$$

where $\rho = 4N_e * c$, $\theta = 4N_e * u$, and u is the mutation rate per site.

Selection is important for the adaptation or genetic improvement of quantitative traits in the natural or artificial settings. As a major force, selection can either decrease or increase the LD given different scenarios [24,39].

Different type and rate of mutations

At generation t and $t + 1$, the relationship between probabilities of haplotype frequency counts is:

$$P_{t+1} = A * P_t \quad (6)$$

Where A is the $k \times k$ transition matrix, with mutation in the present study treated with different types and rates. By considering the same mutation rates for two loci [6], the same results on LD distribution were reproduced (Fig. S1). Generally, based on the transition matrix approach as recently reported [6], we introduced the idea of different types and rates of mutations into the model, and also considered the randomness of mutation in a simple model.

For two linked loci, A and B , it is assumed that the mutation rate from $A1$ to $A2$ is $u1$, and the rate of reverse mutation is $v1$. Likewise, the mutation rate from $B1$ to $B2$ is $u2$, and the rate of reverse mutation is $v2$. Four nucleotides (A, C, T, G) are divided into two types: purine (A, G) and pyrimidine (C, T). Mutation can happen between same types of nucleotides (transition: purine to purine, and pyrimidine to pyrimidine), or between different types of nucleotides (transversion: purine to pyrimidine, and pyrimidine to purine). Different mutation rates for transition and transversion mutations are observed, with transition mutations at higher frequency than transversion mutations due to different underlying molecular mechanisms [21-25].

Four possible situations regarding mutation rates were considered as follows:

1. $u1 = v1 = \alpha$, $u2 = v2 = \alpha$. Mutations between alleles are same types of bases (purines or pyrimidines) at both loci A and B .

2. $u1 = v1 = \alpha$, $u2 = v2 = \beta$. Mutations between alleles at locus A are same types of bases (purines or pyrimidines), but different at locus B .

3. $u1 = v1 = \beta$, $u2 = v2 = \alpha$. Mutations between alleles at locus B are same types of bases (purines or pyrimidines), but different at locus A .

4. $u1 = v1 = \beta$, $u2 = v2 = \beta$. Mutations between alleles are different types of bases (purines or pyrimidines) at both loci A and B .

Then, the four haplotype probabilities following mutation can be computed as:

$$\theta^{**} = \theta^* * T \quad (7)$$

Where T , the transition matrix, according to the aforementioned four possible situations, is built as follows:

$$T1 = \begin{bmatrix} (1-\alpha)^2 & (1-\alpha)\alpha & \alpha(1-\alpha) & \alpha^2 \\ (1-\alpha)\alpha & (1-\alpha)^2 & \alpha^2 & \alpha(1-\alpha) \\ \alpha(1-\alpha) & \alpha^2 & (1-\alpha)^2 & (1-\alpha)\alpha \\ \alpha^2 & \alpha(1-\alpha) & (1-\alpha)\alpha & (1-\alpha)^2 \end{bmatrix} \quad (8)$$

$$T2 = \begin{bmatrix} (1-\alpha)(1-\beta) & (1-\alpha)\beta & \alpha(1-\beta) & \alpha\beta \\ (1-\alpha)\beta & (1-\alpha)(1-\beta) & \alpha\beta & \alpha(1-\beta) \\ \alpha(1-\beta) & \alpha\beta & (1-\alpha)(1-\beta) & (1-\alpha)\beta \\ \alpha\beta & \alpha(1-\beta) & (1-\alpha)\beta & (1-\alpha)(1-\beta) \end{bmatrix} \quad (9)$$

$$T3 = \begin{bmatrix} (1-\beta)(1-\alpha) & (1-\beta)\alpha & \beta(1-\beta) & \beta\alpha \\ (1-\beta)\alpha & (1-\beta)(1-\alpha) & \beta\alpha & \beta(1-\alpha) \\ \beta(1-\alpha) & \beta\alpha & (1-\beta)(1-\alpha) & (1-\beta)\alpha \\ \beta\alpha & \beta(1-\alpha) & (1-\beta)\alpha & (1-\beta)(1-\alpha) \end{bmatrix} \quad (10)$$

$$T4 = \begin{bmatrix} (1-\beta)^2 & (1-\alpha)\alpha & \beta(1-\beta) & \beta^2 \\ (1-\beta)\beta & (1-\alpha)^2 & \beta^2 & \beta(1-\beta) \\ \beta(1-\beta) & \beta^2 & (1-\beta)^2 & (1-\beta)\beta \\ \beta^2 & \beta(1-\beta) & (1-\beta)\beta & (1-\beta)^2 \end{bmatrix} \quad (11)$$

Calculation of LD distribution

Based on the consideration of different types and rates of mutations, factors such as effective population size, recombination, and selection, were further consecutively incorporated in the following several different scenarios.

Scenario I & II: effective population size (N_e) + recombination (c)

Between locus A and locus B , haplotype frequency counts at i th row of X at generation t is $X_i^T = [f_{11}, f_{12}, f_{21}, f_{22}]$ (for $A1B1, A1B2, A2B1$ and $A2B2$, respectively). The recombination rate is designated as c .

At generation t , gametes with haplotype $A1B1$ can be produced in two ways, non-recombinant and recombinant. Non-recombinant is generated with the probability

$(1 - c) \times \frac{f_{11}}{2N_e}$, and recombinant with the following four ways.

1. Alleles $A1$ and $B1$, all from $A1B1$, with probability $c \times \frac{f_{11}}{2N_e} \times \frac{(f_{11}-1)}{2N_e-1}$;
2. $A1$ from $A1B1$, $B1$ from $A2B1$, with probability $c \times \frac{f_{11}}{2N_e} \times \frac{f_{21}}{2N_e-1}$;
3. $A1$ from $A1B2$, $B1$ from $A1B1$, with probability $c \times \frac{f_{12}}{2N_e} \times \frac{f_{11}}{2N_e-1}$;
4. $A1$ from $A1B2$, $B1$ from $A2B1$, with probability $c \times \frac{f_{12}}{2N_e} \times \frac{f_{21}}{2N_e-1}$.

Therefore, the overall probability for haplotype $A1B1$ is:

$$P(A1B1) = (1 - c) \frac{f_{11}}{2N_e} + c \frac{f_{11}}{2N_e} \left[\frac{(f_{11}-1)}{2N_e-1} + \frac{f_{21}}{2N_e-1} \right] + c \frac{f_{12}}{2N_e} \left[\frac{f_{11}}{2N_e-1} + \frac{f_{21}}{2N_e-1} \right] \quad (12)$$

For the other three haplotypes, their probabilities are:

$$P(A1B2) = (1 - c) \frac{f_{12}}{2N_e} + c \frac{f_{11}}{2N_e} \left[\frac{f_{12}}{2N_e-1} + \frac{f_{22}}{2N_e-1} \right] + c \frac{f_{12}}{2N_e} \left[\frac{(f_{12}-1)}{2N_e-1} + \frac{f_{22}}{2N_e-1} \right] \quad (13)$$

$$P(A2B1) = (1 - c) \frac{f_{21}}{2N_e} + c \frac{f_{21}}{2N_e} \left[\frac{f_{11}}{2N_e-1} + \frac{(f_{21}-1)}{2N_e-1} \right] + c \frac{f_{22}}{2N_e} \left[\frac{f_{11}}{2N_e-1} + \frac{f_{21}}{2N_e-1} \right] \quad (14)$$

$$P(A2B2) = (1 - c) \frac{f_{22}}{2N_e} + c \frac{f_{21}}{2N_e} \left[\frac{f_{12}}{2N_e-1} + \frac{f_{22}}{2N_e-1} \right] + c \frac{f_{22}}{2N_e} \left[\frac{f_{12}}{2N_e-1} + \frac{(f_{22}-1)}{2N_e-1} \right] \quad (15)$$

Based on these formulas, the effect of effective population size (Ne) on LD distribution is considered in Scenario I; and the effects of effective population size (Ne) + recombination (c) are examined in Scenario II.

Scenario III: effective population size (Ne) + recombination (c) + selection (s)

Furthermore, to incorporate selection in the model, a selection coefficient s that reduces the allele frequency of $A1$ at locus A , which is a causal variant, is considered.

Conditional on the haplotype probabilities in θ_i^{**} , i.e. P_{A1B1}^{**} , P_{A1B2}^{**} , P_{A2B1}^{**} , and P_{A2B2}^{**} .

the four haplotype probabilities following selection can be computed as:

$$\theta_i^{***} = \begin{bmatrix} \frac{(1-s)P_{A1B1}^{**}}{(1-s)(P_{A1B1}^{**}+P_{A1B2}^{**})+P_{A2B1}^{**}+P_{A2B2}^{**}} \\ \frac{(1-s)P_{A1B2}^{**}}{(1-s)(P_{A1B1}^{**}+P_{A1B2}^{**})+P_{A2B1}^{**}+P_{A2B2}^{**}} \\ \frac{(1-s)P_{A2B1}^{**}}{(1-s)(P_{A1B1}^{**}+P_{A1B2}^{**})+P_{A2B1}^{**}+P_{A2B2}^{**}} \\ \frac{(1-s)P_{A2B2}^{**}}{(1-s)(P_{A1B1}^{**}+P_{A1B2}^{**})+P_{A2B1}^{**}+P_{A2B2}^{**}} \end{bmatrix} \quad (16)$$

Finally, to compute the distribution of allele frequencies in generation $t + 1$ using Equation (6), where the forces of the recombination, mutation and selection are simultaneously considered, the matrix A is defined such that element j of column i contains the probability of getting X_j^T from a Multinomial ($2Ne$, θ_i^T) distribution, for $i, j = 1, \dots, k$.

Random mutation

To take into consideration the randomness of mutation into the framework, a simple model was constructed. At each generation, after completing the calculation of LD, mutation rates were randomly assigned, from the four possible situations as mentioned

previously in formulas 8-11. Then, the calculation of LD for the next generation was continued, and the whole process was repeated for all the generation.

Data Analysis

We followed a procedure similar to a recent report [6]. Briefly, populations with N_e ranging from 5 to 50 in intervals of 5 were compared under different recombination rates, mutation rates and selection coefficients for the causal variant ($s = 0.1$ or $s = 0.01$ for locus A). The recombination rates were in the ranges from 0.01 to 0.1 in intervals of 0.01, and from 0.1 to 0.5 in intervals of 0.1.

Furthermore, the exact distributions of r_E^2 derived under different scenarios as mentioned above were used to validate the approximate deterministic formulas [3,5].

$$r_E^2 = \frac{1}{a+b*N_e*c} \quad (17)$$

To generalize our results to larger effective population size, a non-linear regression model of the form following Sved's formula was considered. The parameters were estimated by non-linear least squares using the data generated from the transition-matrix approach as introduced above.

Results

Effects of different mutation rates on LD distribution

To explore how mutations (different types and rates) affect the dynamics of LD distribution, we compared four different types of mutation rate combinations for a pair of loci A and B : 1) all transitions ($\alpha = 10^{-8}$); 2) transition ($\alpha = 10^{-8}$) and transversion ($\beta = 10^{-9}$); 3) transition ($\alpha = 10^{-8}$) and transversion ($\beta = 10^{-9}$); and 4) all transversions ($\beta = 10^{-9}$) (note: 2) and 3) had the same results). Next, we examined

three different scenarios, with regard to effective population size (Ne), recombination (c), and selection (s), to see their effects on LD dynamics consecutively (details in Materials and methods).

Scenario I: effective population size (Ne)

We found that different mutation rates and their combinations could affect the distribution of LD. However, in the end, even the effective population sizes were different, the same final levels of LD were reached (Fig. 1).

With no regard on Ne , by looking at each panel in Fig. 1, we could easily see that mutation rates led to the different patterns of LD distribution. Compared to the previous study [6], first, we reproduced the effect of same mutation rates on LD distribution (violet lines: 10^{-9} , transversion mutations at two loci). In contrast, when mutation rates (for both transition and transversion) were all of the higher values (blue lines: 10^{-8}), LD started dropping at earlier generations, and reaching the stasis state faster. In addition, when one mutation rate was set at 10^{-8} , and the other at 10^{-9} , or vice versa, LD values distributed in between the above two curves (green and red lines overlapped).

Considering effective population sizes (Ne , from 15 to 45 in intervals of 10), we could see that the larger the Ne , the lower the LD can get. That is, the smaller population size can lead to bigger LD values (Fig. 1A). In addition, when Ne increases, the number of generations needed to reach the final steady state became bigger, from ~600 generations at $Ne=15$, ~1000 generations at $Ne=25$, ~1200 generations at $Ne=35$, to ~1500 generations at $Ne=45$ (Fig. 1A-D).

Scenario II: N_e + recombination (c)

Besides effective population size, we further added recombination rate into the framework, to study their combined effects on LD distribution. Recombination rate seemed to affect greatly the LD distribution, and also the final levels of LD when reaching the stasis state (Fig. 2). Generally, within each horizontal panel (Fig. 2A, B and C), we found the pattern that the larger the recombination rates, the smaller the LD values, and the narrower range LD (maximum and minimum) values could change in. Furthermore, three different relative ratios between transition and transversion mutation rates were also considered: transition mutation rate was fixed at 10^{-8} , and transversion mutation rates were set at 10^{-9} , 2×10^{-9} , and 5×10^{-9} , respectively. With the ratio of transition to transversion mutation rates decreasing ($\lambda=10, 5, 2$) (vertical panels, Fig. 2A, B and C), we could see that LD lines got closer, and changed only within a narrower range. Furthermore, as λ became smaller, LD values started to drop at earlier generation, and reached the stasis state faster.

Using the non-linear least squares method [6], to fit data generated from the transition-matrix approach considering the four different combinations of mutation types and rates, coefficients a and b were obtained for each 50 generations under different recombination rates (Tables S1-S3). When fitting the LD distribution according to N_e with a non-linear model, striking differences were found between the curves, for different mutation types and rates (Fig. 3A, B and C). Along with the increasing of population sizes, higher mutation rates (Ts and Tv all equal to 10^{-8} , blue line) could lead to the decrease of LD values (at generation 1500) (Fig. 3A and C), whereas

combinations of other mutation rates (green and red lines overlapped, violet line) exhibited the contrary trend. However, when using data at generation 3500 (i.e. reaching the final steady state of LD distribution), no difference could be seen (Fig. 3B and D).

In addition, 3-dimensional view on the effects of mutation and recombination on LD distribution was presented, to better understand its dynamic changes (Fig. 4). Similar to results in Fig. 2, λ values (10, 5, 2) did not affect the time when LD values reached to the stasis state. However, with the increasing of recombination rates, we could see that LD distribution reached the stasis state at an earlier generation, and finally at $c = 0.5$, LD values reached zero, i.e. the equilibrium state.

Scenario III: $Ne + c + \text{selection } (s)$

Selection plays a fundamental role in affecting the dynamics of allele frequency for causal genetic variant, and subsequently LD distribution. Thus, we analyzed the effect of selection, together with different ratios of mutation rates and recombination rates. The effective population size (Ne) examined here was set at 50.

Compared to previous results [6], we could clearly see different patterns of LD distribution. First, a higher relative ratio of mutation rates between the pair of genetic loci (λ (Ts/Tv) = 10) was considered. When selections on the mutation rates were absent or relatively weak (Fig. 5A and B), LD distribution showed curves similar to results previously presented (with only recombination and effective population size considered) (Fig. 2). However, even with a weak selection force on alleles at one of the genetic loci, LD distribution could clearly be seen affected, showing separated curves (green and red

lines), when under different combinations of transition and transversion mutation rates (Fig. 5B). Furthermore, when increasing the selection intensity to a relatively strong level ($s = 0.1$), abrupt changes of LD distribution appeared, with a totally different pattern (Fig. 5C). LD curves dropped precipitously at very earlier generations, and even down to a lower level of LD when reaching the steady state.

Next, we further computed the LD distribution, when lower relative ratios of mutation rates ($\lambda = 5$ and 2) were considered, and similar results were obtained (only were selection coefficients $s = 0.01$, $s = 0.1$ used) (Fig. S2 and S3). It suggested that compared to mutation, recombination and effective population size, selection on mutations could potentially play a major role in the determination of LD dynamics.

Furthermore, we filtered the minor allele frequencies at the two genetic loci, after which similar results were observed (Fig. S4, without selection; Fig. S5-S7, with selection, $\lambda = 10$, 5 and 2 , respectively).

Thus, mutation can further affect LD distribution when considering in concert the roles of selection, recombination and effective population size.

Random mutation on LD distribution

Mutation could arise in a random mode, varying with time, strength of selection, and tissue or cell-dependent [20-25]. To study the randomness of mutation and its effect on LD distribution, we modeled in a simpler manner the mode of mutation changes across generations (details in Materials and methods). We found that random mutation could only bring a small effect on LD distribution at the current (weak-mutation) scale, changing also in a random way (Fig. 6). Generally, LD distribution could still be seen

to have a similar pattern to previous results (Fig. 2 and 5), no matter how the randomness was brought into the modeling process of LD distribution. However, the LD curve was not smooth any more now, but showing increasing ruggedness. Especially when LD values reach the steady state at later generations, LD values went up and down along the equilibrium line, but only in a limited range. Thus, random mutation could only cause the fluctuation of LD values, and doesn't change the general pattern of LD distribution.

Discussion

With the fast development and wide adoption of genome sequencing technology in recent years, the importance of LD (an old concept in population genetics) has been fully recognized, in tracing population history and identifying genes for complex agricultural and human disease traits [4,6-8]. Since its invention, LD has been studied and dissected in great detail for its statistical and distributional properties, for factors like recombination, effective population size and selection [1-6,24,39]. Recently, in-depth analytical computation and mathematical unbiased estimation of LD on unphased genotype data were reported [7,8,40,41]. Due to its fundamental role in population and quantitative genetics, LD will stay as a fascinating research topic for quite some time. Mutation is considered as the only source to provide genetic variation, and has been extensively studied in fields like evolution, agriculture and human diseases, partially due to the relatively easy and large-scale identification and functional exploration by genome sequencing [15,34-38]. As we found here, different mutation types and rates, and their combination with factors like effective population size, recombination and

selection, do interfere with the dynamics of LD distribution. In addition, mutation at the weak scale in the current study can only affect LD distribution at a relatively minor level, compared to effective population size, recombination, and especially selection. However, recently, it was found that the strength of mutation (strong or weak) can affect the fitness landscape by breaking the limits of transition between saddle points [23,33]. Basically, transition between fitness lampposts comes down to the change of allele frequencies, and the break-down of the drift-barrier [9,23]. We suspect that the strength of mutation could affect LD distribution, but how and at what scale still needs further investigation.

Interaction between mutations (epistasis) was found to be important to the evolution of phenotype [42-45]. Alleles between mutations could be under positive or negative selection, and interact to promote the fitness of the organism or phenotype of interest. Several studies reported the negative epistasis between mutations, which can reduce mutational load, and amplify the fitness effects of deleterious mutations [43-45]. Epistatic variance can be converted into additive genetic variance, and utilized in the selection and improvement of important traits [46,47]. A recent simulation study also worked on the contribution of mutational variance to genetic variance and selection response [17]. However, the present study did not consider epistasis between mutations. It will be interesting to see how mutations interact to affect LD distribution, and also how phenotypic fitness can be affected by epistatic mutational variance.

Randomness of mutation was also shown to affect LD distribution, by adding noise and increasing the ruggedness of LD distribution. However, we only currently examined

the effect of randomness in a simple model. As discovered in the experimental research, the distribution of mutation rates and spectrum are quite different. At the species level, mutation rates distribute in a wide range, from 10^{-7} in vertebrates, to 100 or 1000 times lower in bacteria [22,23]. In experimental evolution, mutational spectrum is also discovered to be [9,43-45]. Furthermore, evolutionary forces play on proteins involved in replication fidelity, and affect the subsequent stochastic path taken by mutations [9]. Therefore, more complex or *de facto* design, considering the negative epistasis and stochasticity of mutations, might reveal more interesting results.

In the genetics standpoint, association analysis and genomic breeding value estimation at the genome level are all based on the LD status between genetic markers and causal variants [16]. Mutations can affect LD distribution, contribute to epistatic variance, or convert into genetic variance, which potentially affect the studies on the genetic basis, selection or evolution of complex traits. It is not uncommon that mutations are indicated to be involved in the complex trait of importance, e.g. agriculture or human disease studies [15,34-38]. Many questions remained to be answered. More specifically, how alleles mutate into different nucleotides, in which way the local LD block can be affected, and whether they can affect the mapping of genetic factors underlying complex traits, or estimation of breeding value.

Conclusions

Taken together, different mutation rates and randomness (at the current weak scale) could affect the dynamics of LD distribution, but only at a smaller scale. However, combined with other factors (e.g. selection), different mutation rates could

fundamentally disturb the LD distribution. Our findings can help better understanding the role of mutation in molecular evolution and complex trait genetics.

Abbreviations

LD: linkage disequilibrium; Ts: transitions; Tv: transversions; Ne: effective population size; C: recombination; S: selection

Supplementary Information

The online version contains supplementary material available at [.](#)

Additional file 1 Fig. S1. Reproduced comparisons of approximation to exact distribution of linkage disequilibrium (Sved's, Hill's and Calibrated) [6]. **(A)** Without mutation and selection; **(B)** with mutation, without selection. Recombination rates ($c = 0.1, 0.05, 0.01$); effective population size ($Ne = 5-50$).

Additional file 2 Fig. S2. Distribution of linkage disequilibrium when considering selection ($\lambda=5$). **(A)** Selection coefficient ($s = 0.01$); **(B)** $s = 0.1$. Recombination rates: $c = 0.1, 0.05, 0.01$; Effective population size ($Ne = 50$).

Additional file 3 Fig. S3. Distribution of linkage disequilibrium when considering selection ($\lambda=2$). **(A)** Selection coefficient ($s = 0.01$); **(B)** $s = 0.1$. Recombination rates: $c = 0.1, 0.05, 0.01$; Effective population size ($Ne = 50$).

Additional file 4 Fig. S4. Linkage disequilibrium dynamics after minor allele frequency filtering (without selection) ($\lambda (Ts/Tv) = 10$)

Additional file 5 Fig. S5. Linkage disequilibrium dynamics after minor allele frequency filtering ($\lambda = 10$). (A) $s = 0.01$; (B) $s = 0.1$

Additional file 6 Fig. S6. Linkage disequilibrium dynamics after minor allele frequency filtering ($\lambda = 5$). (A) $s = 0.01$; (B) $s = 0.1$

Additional file 7 Fig. S7. Linkage disequilibrium dynamics after minor allele frequency filtering ($\lambda = 2$). (A) $s = 0.01$; (B) $s = 0.1$

Additional file 8 Table S1.

Additional file 9 Table S2.

Additional file 10 Table S3.

Acknowledgements

We appreciate the help and support from other lab members in Drs. Du and Yang's groups.

Authors' Contributions

R. Zhang wrote the scripts, performed the data analysis, and drafted the manuscript. C.X. Yang helped with the discussion, and drafted the manuscript. Z.Q. Du conceived the idea, designed the work and wrote the manuscript. All authors have read and approved the final manuscript.

Funding

Funding supports were from the Start-up and the ESI Promotion grants provided by Yangtze University.

Availability of data and materials

The scripts used to generate and analyze the data can be found at <https://github.com/Rong-Zh/LD>.

Declarations

Ethics approval and consent to participate

No animal samples used in this study.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no conflict of interest.

References

1. Hill WG, Robertson A. Linkage disequilibrium in finite populations. Theor Appl Genet. 1968;38(6):226-31.
2. Sved JA. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor Popul Biol. 1971;2(2):125-41.

- 413 3. Sved JA, Feldman MW. Correlation and probability methods for one and two loci.
414 Theor Popul Biol. 1973;4(1):129-32.
- 415 4. Sved JA, Hill WG. One hundred years of linkage disequilibrium. Genetics.
416 2018;209(3):629-36.
- 417 5. Hill WG. Linkage disequilibrium among multiple neutral alleles produced by
418 mutation in finite population. Theor Popul Biol. 1975;8(2):117-26.
- 419 6. Qu J, Kachman SD, Garrick D, Fernando RL, Cheng H. Exact distribution of linkage
420 disequilibrium in the presence of mutation, selection, or minor allele frequency
421 filtering. Front Genet. 2020;21(11):362.
- 422 7. Song YS, Song JS. Analytic computation of the expectation of the linkage
423 disequilibrium coefficient r^2 . Theor Popul Biol. 2007;71(1):49-60.
- 424 8. Ragsdale AP, Gravel S. Unbiased estimation of linkage disequilibrium from
425 unphased data. Mol Biol Evol. 2020;37(3):923-32.
- 426 9. Liu H, Zhang J. The rate and molecular spectrum of mutation are selectively
427 maintained in yeast. Nat Commun. 2021;12(1):4044.
- 428 10. López-Cortegano E, Craig RJ, Chebib J, Samuels T, Morgan AD, Kraemer SA, et
429 al. De novo mutation rate variation and its determinants in *Chlamydomonas*. Mol
430 Biol Evol. 2021;38(9):3709–23.
- 431 11. Lu Z, Cui J, Wang L, Teng N, Zhang S, Lam HM, et al. Genome-wide DNA
432 mutations in *Arabidopsis* plants after multigenerational exposure to high
433 temperatures. Genome Biol. 2021;22(1):160.

- 434 12. Barrick JE, Lenski RE. Genome dynamics during experimental evolution. Nat Rev
435 Genet. 2013;14(12):827-39.
- 436 13. Tenaillon O, Barrick JE, Ribeck N, Deatherage DE, Blanchard JL, Dasgupta A, et
437 al. Tempo and mode of genome evolution in a 50,000-generation experiment. Nature.
438 2016;536(7615):165-70.
- 439 14. Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. The dynamics of
440 molecular evolution over 60,000 generations. Nature. 2017;551(7678):45-50.
- 441 15. Zhang H, Liang Q, Wang N, Wang Q, Leng L, Mao J, et al. Microevolutionary
442 dynamics of chicken genomes under divergent selection for adiposity. iScience.
443 2020;23(6):101193.
- 444 16. Hill WG. Is continued genetic improvement of livestock sustainable? Genetics.
445 2016;202(3):877-81.
- 446 17. Mulder HA, Lee SH, Clark S, Hayes BJ, van der Werf JHJ. The impact of genomic
447 and traditional selection on the contribution of mutational variance to long-term
448 selection response and genetic variance. Genetics. 2019;213(2):361-78.
- 449 18. Yang Z. Computational molecular evolution. Oxford University Press. 2006.
- 450 19. Yang Z. Molecular evolution: a statistical approach. Oxford University Press. 2014.
- 451 20. Denver DR, Morris K, Lynch M, Thomas WK. High mutation rate and
452 predominance of insertions in the *Caenorhabditis elegans* nuclear genome. Nature.
453 2004;430(7000):679-82.
- 454 21. Lynch M. Evolution of the mutation rate. Trends Genet. 2010;26(8):345-52.

- 455 22. Lynch M. Rate, molecular spectrum, and consequences of human mutation. *Proc*
456 *Natl Acad Sci U S A.* 2010;107(3):961-8.
- 457 23. Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, et al. Genetic
458 drift, selection and the evolution of the mutation rate. *Nat Rev Genet.*
459 2016;17(11):704-14.
- 460 24. Walsh B, Lynch M. Evolution and selection of quantitative traits. Oxford University
461 Press. 2018.
- 462 25. Chintalapati M, Moorjani P. Evolution of the mutation rate across primates. *Curr*
463 *Opin Genet Dev.* 2020;62:58-64.
- 464 26. Long H, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo W, et al.
465 Evolutionary determinants of genome-wide nucleotide composition. *Nat Ecol Evol.*
466 2018;2(2):237-40.
- 467 27. Argueso JL, Alani E. Hundreds of thousands of cell generations reveal a treasure
468 chest of genome alterations. *Proc Natl Acad Sci U S A.* 2020;117(50):31567-9.
- 469 28. Nguyen DT, Wu B, Long H, Zhang N, Patterson C, Simpson S, et al. Variable
470 Spontaneous Mutation and Loss of Heterozygosity among Heterozygous Genomes
471 in Yeast. *Mol Biol Evol.* 2020;37(11):3118-30.
- 472 29. Sui Y, Qi L, Wu JK, Wen XP, Tang XX, Ma ZJ, et al. Genome-wide mapping of
473 spontaneous genetic alterations in diploid yeast cells. *Proc Natl Acad Sci U S A.*
474 2020;117(45):28191-200.
- 475 30. Cherry JL. Extreme C-to-A Hypermethylation at a Site of Cytosine-N4 Methylation.
476 *mBio.* 2021;12(2):e00172-21.

- 477 31. DeWitt WS, Harris KD, Ragsdale AP, Harris K. Nonparametric coalescent
478 inference of mutation spectrum history and demography. *Proc Natl Acad Sci U S A*.
479 2021;118(21):e2013798118.
- 480 32. Loewe L, Hill WG. The population genetics of mutations: good, bad and indifferent.
481 *Philos Trans R Soc Lond B Biol Sci*. 2010;365(1544):1153-67.
- 482 33. Bakhtin Y, Katsnelson MI, Wolf YI, Koonin EV. Evolution in the weak-mutation
483 limit: Stasis periods punctuated by fast transitions between saddle points on the
484 fitness landscape. *Proc Natl Acad Sci U S A*. 2021;118(4):e2015665118.
- 485 34. Muiños F, Martínez-Jiménez F, Pich O, Gonzalez-Perez A, Lopez-Bigas N. In silico
486 saturation mutagenesis of cancer genes. *Nature*. 2021;596(7872):428-32.
- 487 35. Moore L, Cagan A, Coorens THH, Neville MDC, Sanghvi R, Sanders MA, et al.
488 The mutational landscape of human somatic and germline cells. *Nature*.
489 2021;597(7876):381-6.
- 490 36. Coorens THH, Moore L, Robinson PS, Sanghvi R, Christopher J, Hewinson J, et al.
491 Extensive phylogenies of human development inferred from somatic mutations.
492 *Nature*. 2021;597(7876):387-92.
- 493 37. Li R, Di L, Li J, Fan W, Liu Y, Guo W, et al. A body map of somatic mutagenesis
494 in morphologically normal human tissues. *Nature*. 2021;597(7876):398-403.
- 495 38. Park S, Mali NM, Kim R, Choi JW, Lee J, Lim J, et al. Clonal dynamics in early
496 human embryogenesis inferred from somatic mutation. *Nature*. 2021;597(7876):393-
497 7.

- 498 39. Falconer DS, Mackay TFC. Introduction to quantitative genetics (forth edition).
499 Pearson Education Ltd. 1996.
- 500 40. Gerard D. Scalable bias-corrected linkage disequilibrium estimation under
501 genotype uncertainty. *Heredity (Edinb)*. 2021;127(4):357-362.
- 502 41. Nelson D, Kelleher J, Ragsdale AP, Moreau C, McVean G, Gravel S. Accounting
503 for long-range correlations in genome-wide simulations of large cohorts. *PLoS Genet*.
504 2020;16(5):e1008619.
- 505 42. Kryazhimskiy S, Rice DP, Jerison ER, Desai MM. Microbial evolution. Global
506 epistasis makes adaptation predictable despite sequence-level stochasticity. *Science*.
507 2014;344(6191):1519-22.
- 508 43. Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF. Negative epistasis
509 between beneficial mutations in an evolving bacterial population. *Science*.
510 2011;332(6034):1193-6.
- 511 44. Zheng J, Guo N, Wagner A. Mistranslation Reduces Mutation Load in Evolving
512 Proteins through Negative Epistasis with DNA Mutations. *Mol Biol Evol*.
513 2021;38(11):4792-04.
- 514 45. Bendixsen DP, Østman B, Hayden EJ. Negative Epistasis in Experimental RNA
515 Fitness Landscapes. *J Mol Evol*. 2017;85(5-6):159-68.
- 516 46. Mäki-Tanila A, Hill WG. Influence of gene interaction on complex trait variation
517 with multilocus models. *Genetics*. 2014;198(1):355-67.

518 47. Hill WG. "Conversion" of epistatic into additive genetic variance in finite
519 populations and possible impact on long-term selection response. J Anim Breed
520 Genet. 2017;134(3):196-01.

Legends to Figures

Fig. 1. Effective population size (Ne) on linkage disequilibrium distribution when under different mutation schemes. (A-D), $Ne = 15, 25, 35, 45$. Blue lines, transition (Ts) ($\alpha = 10^{-8}$) mutations; green lines, transition ($\alpha = 10^{-8}$) and transversion (Tv) ($\beta = 10^{-9}$) mutations; red lines, transversion ($\beta = 10^{-9}$) and transition ($\alpha = 10^{-8}$) mutations; violet lines, transversion ($\beta = 10^{-9}$) mutations.

Fig. 2. Effects of relative ratios between transition (Ts) and transversion (Tv) mutation rates on linkage disequilibrium distribution. (A) λ (Ts/Tv) = 10; (B) $\lambda = 5$; (C) $\lambda = 2$. Effective population size (Ne) = 50, with recombination (c), but without selection. Recombination rates were set as $c = 0.1, 0.05, 0.01$.

Fig. 3. Non-linear relationship inferred. Between linkage disequilibrium and recombination at two representative generations ($Ne = 50$): (A) At generation 1500 (in decreasing); (B) At generation 3500 (equilibrium reached). Between effective population size (Ne) and linkage disequilibrium (r^2): (C) At generation 1500; (D) At generation 3500 (equilibrium reached).

Fig. 4. 3-Dimensional view on linkage disequilibrium dynamics under different scenarios (recombination, generation, mutation). (A) $Ts = 10^{-8}$, $Ts = 10^{-8}$; (B) $Ts = 10^{-8}$, $Tv = 10^{-9}$; (C) $Tv = 10^{-9}$, $Ts = 10^{-8}$; (D) $Tv = 10^{-9}$, $Tv = 10^{-9}$

Fig. 5. Distribution of linkage disequilibrium when considering selection. Three selection coefficients (s) were considered. (A) $s = 0$; (B) $s = 0.01$; (C) $s = 0.1$. Recombination rates: $c = 0.1, 0.05, 0.01$; Effective population size (Ne) = 50; Mutation rates λ (Ts/Tv) = 10.

543 **Fig. 6.** Random mutation on linkage disequilibrium without selection

544 **Fig. 7.** Random mutation on linkage disequilibrium with selection. **(A)** $\lambda = 10$; **(B)** $\lambda =$
545 5 ; **(C)** $\lambda = 2$. Effective population size (Ne) = 40; recombination rates (c) = 0.1, 0.05,
546 0.01; selection coefficients (s) = 0, 0.01, 0.1.

Figures

Figure 1

Effective population size (N_e) on linkage disequilibrium distribution when under different mutation schemes. (A-D), $N_e = 15, 25, 35, 45$. Blue lines, transition (Ts) ($\alpha = 10^{-8}$) mutations; green lines, transition

($\alpha=10^{-8}$) and transversion (Tv) ($\beta=10^{-9}$) mutations; red lines, transversion ($\beta=10^{-9}$) and transition ($\alpha=10^{-8}$) mutations; violet lines, transversion ($\beta=10^{-9}$) mutations.

Figure 2

Effects of relative ratios between transition (Ts) and transversion (Tv) mutation rates on linkage disequilibrium distribution. (A) λ (Ts/Tv) = 10; (B) λ = 5; (C) λ = 2. Effective population size (N_e) = 50, with recombination (c), but without selection. Recombination rates were set as c = 0.1, 0.05, 0.01.

Figure 3

Non-linear relationship inferred. Between linkage disequilibrium and recombination at two representative generations ($N_e = 50$): (A) At generation 1500 (in decreasing); (B) At generation 3500 (equilibrium reached). Between effective population size (N_e) and linkage disequilibrium (r^2): (C) At generation 1500; (D) At generation 3500 (equilibrium reached).

Figure 4

3-Dimensional view on linkage disequilibrium dynamics under different scenarios (recombination, generation, mutation). (A) $T_s = 10^{-8}$, $T_s = 10^{-8}$; (B) $T_s = 10^{-8}$, $T_v = 10^{-9}$; (C) $T_v = 10^{-9}$, $T_s = 10^{-8}$; (D) $T_v = 10^{-9}$, $T_v = 10^{-9}$

Figure 5

Distribution of linkage disequilibrium when considering selection. Three selection coefficients (s) were considered. (A) $s = 0$; (B) $s = 0.01$; (C) $s = 0.1$. Recombination rates: $c = 0.1, 0.05, 0.01$; Effective population size (N_e) = 50; Mutation rates λ (T_s/T_v) = 10.

Figure 6

Random mutation on linkage disequilibrium without selection

Figure 7

Random mutation on linkage disequilibrium with selection. (A) $\lambda = 10$; (B) $\lambda = 5$; (C) $\lambda = 2$. Effective population size (N_e) = 40; recombination rates (c) = 0.1, 0.05, 0.01; selection coefficients (s) = 0, 0.01, 0.1.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Fig.S1.jpg](#)
- [Fig.S2.jpg](#)
- [Fig.S3.jpg](#)
- [Fig.S4.jpg](#)
- [Fig.S5.jpg](#)
- [Fig.S6.jpg](#)
- [Fig.S7.jpg](#)
- [TableS1S3.docx](#)