frontiers in Genetics

METHODS
published: 21 April 2020
doi: 10.3389/fgene.2020.00362

Exact Distribution of Linkage Disequilibrium in the Presence of Mutation, Selection, or Minor Allele Frequency Filtering

Jiayi Qu[1], *Stephen D. Kachman*[2], *Dorian Garrick*[3], *Rohan L. Fernando*[4] *and Hao Cheng*[1]*

[1] Department of Animal Science, University of California, Davis, Davis, CA, United States, [2] Department of Statistics, University of Nebraska Lincoln, Lincoln, NE, United States, [3] School of Agriculture, Massey University, Wellington, New Zealand, [4] Department of Animal Science, Iowa State University, Ames, IA, United States

Linkage disequilibrium (LD), often expressed in terms of the squared correlation (r^2) between allelic values at two loci, is an important concept in many branches of genetics and genomics. Genetic drift and recombination have opposite effects on LD, and thus r^2 will keep changing until the effects of these two forces are counterbalanced. Several approximations have been used to determine the expected value of r^2 at equilibrium in the presence or absence of mutation. In this paper, we propose a probability-based approach to compute the exact distribution of allele frequencies at two loci in a finite population at any generation t conditional on the distribution at generation $t-1$. As r^2 is a function of this distribution of allele frequencies, this approach can be used to examine the distribution of r^2 over generations as it approaches equilibrium. The exact distribution of LD from our method is used to describe, quantify, and compare LD at different equilibria, including equilibrium in the absence or presence of mutation, selection, and filtering by minor allele frequency. We also propose a deterministic formula for expected LD in the presence of mutation at equilibrium based on the exact distribution of LD.

Keywords: linkage disequilibrium, effective population size, mutation rate, selection, minor allele frequency filtering

OPEN ACCESS

Edited by:
Jacob A. Tennessen, Harvard University, United States

Reviewed by:
Guanglin He, Sichuan University, China
Dan Skelly, Jackson Laboratory, United States

***Correspondence:**
Hao Cheng
qtlcheng@ucdavis.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *08 November 2019*
Accepted: *25 March 2020*
Published: *21 April 2020*

Citation:
Qu J, Kachman SD, Garrick D, Fernando RL and Cheng H (2020) Exact Distribution of Linkage Disequilibrium in the Presence of Mutation, Selection, or Minor Allele Frequency Filtering. Front. Genet. 11:362. doi: 10.3389/fgene.2020.00362

1. INTRODUCTION

Linkage disequilibrium (LD), the non-random association of alleles at two or more loci, is an important concept in various areas of genetics, including evolutionary, quantitative, and statistical genetics. In evolutionary genetics, LD is used to detect the genomic locations of historical selection and to estimate the time of divergence and geographic subdivision between populations (Sved and Hill, 2018). For example, LD was used to date the divergence of the European population from the African population using the HapMap data (Sved et al., 2008). In the field of quantitative and statistical genetics, methods for genomic prediction and genome-wide association studies (GWAS) rely on the existence of LD between the molecular markers and the unobserved quantitative trait loci (QTL).

Two statistics that have been widely used to quantify the LD between allelic values at two loci are: their covariance (D); or their squared correlation (r^2) (Hill and Robertson, 1968). For a diallelic system, they can be expressed as

$$D = p_{A_1B_1} - p_{A_1}p_{B_1} \tag{1}$$

and

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

respectively, where p_{A_i} represents the frequency of the *ith* allele at locus A, p_{B_j} represents the frequency of the *jth* allele at locus B, and $p_{A_iB_j}$ represents the frequency of haplotype A_iB_j in the population. Although the covariance of different pairs of alleles depends on their corresponding haplotype frequencies and the respective allele frequencies, a single value of D is sufficient to characterize the disequilibrium between two loci in a diallelic system, i.e., $D = D_{A_1B_1} = -D_{A_1B_2} = -D_{A_2B_1} = D_{A_2B_2}$. The value of D depends on how the alleles are coded (i.e., alleles A_1 and A_2 could be coded as 0 and 1 or as 1 and 0) while the value of r^2 is invariant to this choice. Therefore, r^2 is increasingly used as a metric to quantify LD in the literature.

In a finite population, allele frequencies at two loci are subject to random fluctuations due to the stochastic sampling of a finite number of gametes across generations. This random change of the gene frequency between one generation and the next is genetic drift. Due to the randomness of the genetic drift, one initial random-mating population can evolve into one of a finite collection of sub-populations or lines with different allele frequencies at the two loci, and thus with different LD. Usually, this finite collection of possible lines is jointly considered to understand the distribution of the allele frequencies and LD of two linked loci over generations. This is equivalent to considering a finite collection of pairs of loci in one line. The consequences of the genetic drift, mutation, and selection for the collection of lines of two loci equally apply to a collection of pairs of loci in one line. The former conceptualization is used in this paper to understand the effects of genetic drift, mutation, selection, and minor allele frequency (MAF) cutoff on the distribution of LD.

Generally, in a finite population, the opposite effects on LD of genetic drift and recombination lead to an equilibrium value for the expectation of r^2 when these two forces are counterbalanced. It has been shown that for large values of N_ec, the equilibrium value for the expectation of r^2 (r_E^2) is approximately:

$$r_E^2 = \frac{1}{1+4N_ec} \tag{3}$$

where c is the recombination rate and N_e is the effective population size (Sved, 1971; Sved and Feldman, 1973; Hill, 1975).

The above formula, however, does not consider the variation introduced by mutation in the population. When mutation is considered, a balance is expected between the loss of variation by the fixation of one or other allele, and the replenishment of an extinct allele by mutation. An approximation for the equilibrium value of r^2 in the presence of mutation in a finite population has been developed in Ohta and Kimura (1971) and Hill (1975). Instead of using $E(r^2)$, they used a quantity (σ_d^2) named the standard linkage deviation (Ohta and Kimura, 1969) to describe the status of this equilibrium, where σ_d^2 is an approximation of the expected value of r^2 expressed as the ratio of the expectations of the numerator and the denominator of r^2 (Ohta and Kimura, 1971):

$$\sigma_d^2 = \frac{E(D^2)}{E(p_{A_1}(1-p_{A_1})p_{B_1}(1-p_{B_1}))}. \tag{4}$$

The equilibrium value of σ_d^2 is approximated as:

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

where $\rho = 4N_ec$, $\theta = 4N_eu$ and u is the mutation rate per site (Ohta and Kimura, 1971). The equilibrium value of σ_d^2 as an approximation of r_E^2 has been described in Walsh and Lynch (2018). Similarly, this value is considered in this paper as an approximation of r_E^2 in the presence of mutation.

Selection is another important force affecting LD that was not considered in either Sved's or Hill's formulas. Selection occurs at causal variants or QTL during evolution (natural selection) or selective breeding (artificial selection). The impact of selection can either decrease or increase the LD in a population given different scenarios (Mitchell-Olds et al., 2007). The equilibrium among mutation and selection is of special interest in terms of the distribution of allele frequencies and of r^2.

In contrast to the approximate methods mentioned above, in this paper we will derive the exact distribution of allele frequencies and LD of two linked loci at equilibrium in the absence or presence of mutation, selection, and MAF cutoff. The distribution of LD in the presence of filtering by MAF is considered in our study due to the prevalent practice of filtering out marker loci with low MAF in genomic analyses used for genetic evaluation or QTL discovery (GWAS).

Our exact distributions are first used to validate the approximate deterministic formulas for r_E^2, such as Sved's and Hill's approximations. Next we describe, quantify and compare r_E^2 at different equilibria: including equilibrium in the presence of mutation; equilibrium in the presence of mutation and selection; with or without filtering by MAF. Finally, we calibrate Sved's deterministic formula for r_E^2 in the presence of mutation based on the exact distribution of LD. The objective of this paper is to present a computational approach to derive the exact distribution of LD over generations and use it to study the distribution of LD whether or not there is mutation, selection, or filtering by MAF.

2. MATERIALS AND METHODS

In this section, we will show how to compute the exact distribution of allele frequencies at two linked loci in generation t, given their distribution in generation $t-1$, as a function of the effective population size, the recombination rate between the loci, the mutation rate, the selection coefficient, and the MAF threshold. As LD is a function of the allele frequencies at the two

loci, its distribution over generations can be calculated based on the distribution of the allele frequencies.

2.1. Transition-Matrix Approach in the Presence of Mutation, Selection, or Filtering by Minor Allele Frequency

Two diallelic loci A and B, with alleles A_1 and A_2 at the A locus and alleles B_1 and B_2 at the B locus, are considered. To incorporate selection and MAF threshold, we consider the A locus as the causal variant under selection and B locus as the molecular marker under the MAF filtering. Four possible haplotypes (i.e., A_1B_1, A_1B_2, A_2B_1, and A_2B_2) are possible at these two loci. In a population of size N_e, the frequency counts of these four haplotypes can take on $k = \frac{(2N_e+3)!}{3!(2N_e)!}$ possible values, where for each of these possibilities, the sum of the four counts is $2N_e$. For example, when N_e is equal to two, all possible combinations of haplotype frequency counts are given in **Figure S1**. In general, let $\mathbf{X}$ be a $k \times 4$ matrix with each row representing a possible combination of haplotype frequency counts for some value of N_e. Thus, the rows of $\mathbf{X}$ represent the collection of k lines with different allele frequencies at two loci. Let $\mathbf{P}_t$ denote a $k \times 1$ vector with element i indicating the probability of the frequency counts in row i of $\mathbf{X}$ at generation t. Thus, $\mathbf{P}_t$ gives the distribution of allele frequencies at generation t. We show below that the distribution in generation $t+1$ can be written in general as

$$\mathbf{P}_{t+1} = \mathbf{A}\mathbf{P}_t, \tag{6}$$

where $\mathbf{A}$ is a $k \times k$ transition matrix that is derived below in various circumstances of recombination, mutation, and selection.

First, consider a line with haplotype frequency counts $\mathbf{x}_i^T$ from ith row of $\mathbf{X}$ at generation t, e.g., $\mathbf{x}_i^T = [f_{11}, f_{12}, f_{21}, f_{22}]$, where the frequency counts for the four haplotypes, A_1B_1, A_1B_2, A_2B_1, and A_2B_2, are denoted by f_{11}, f_{12}, f_{21} and f_{22}. Ignoring the effects of recombination, mutation, or selection, sampling of $2N_e$ gametes from this line can be modeled by a multinomial process with sample size $n = 2N_e$ and probability vector $\boldsymbol{\theta}_i = \frac{\mathbf{x}_i^T}{2N_e}$ for the four haplotype probabilities. Thus, the distribution of the frequency count in the next generation is given by the $k \times 1$ vector $\mathbf{m}_i$, where the element j of $\mathbf{m}_i$ is the probability of getting $\mathbf{x}_j^T$ from the Multinomial(n, $\boldsymbol{\theta}_i$) distribution. Given that element i of $\mathbf{P}_t$ is the probability of $\mathbf{x}_i^T$ in generation t, the distribution of allele frequencies in the next generation is given by:

$$\mathbf{P}_{t+1} = \mathbf{M}\mathbf{P}_t, \tag{7}$$

where $\mathbf{M}$ is the $k \times k$ matrix with columns $\mathbf{m}_i$, as described above, for $i = 1, \ldots, k$. The above formula shows how the distribution of allele frequencies change due to genetic drift, ignoring recombination, mutation, and selection. Now we accommodate recombination in computing $\mathbf{P}_{t+1}$. In gametes produced from generation t, the probability of a non-recombinant A_1B_1 is $(1-c) \times \frac{f_{11}}{2N_e}$, where c is the recombination rate between locus A and locus B. On the other hand, a recombinant A_1B_1 can be produced in one of four ways. They and their associated probabilities are:

1. Alleles A_1 and B_1 originate from two different A_1B_1 haplotypes with probability $c \times \frac{f_{11}}{2N_e} \times \frac{f_{11}-1}{2N_e-1}$.
2. Allele A_1 originates from an A_1B_1 haplotype and B_1 originates from an A_2B_1 haplotype with probability $c \times \frac{f_{11}}{2N_e} \times \frac{f_{21}}{2N_e-1}$.
3. Allele A_1 originates from an A_1B_2 haplotype and B_1 originates from an A_1B_1 haplotype with probability $c \times \frac{f_{12}}{2N_e} \times \frac{f_{11}}{2N_e-1}$.
4. Allele A_1 originates from an A_1B_2 haplotype and B_1 originates from an A_2B_1 haplotype with probability $c \times \frac{f_{12}}{2N_e} \times \frac{f_{21}}{2N_e-1}$.

Thus, accounting for recombination, the probability of a A_1B_1 haplotype changes from $\frac{f_{11}}{2N_e}$ to:

$$\begin{aligned}\Pr(A_1B_1) = (1-c)\frac{f_{11}}{2N_e} + \\ c\frac{f_{11}}{2N_e}\left[\frac{(f_{11}-1)}{2N_e-1} + \frac{f_{21}}{2N_e-1}\right] + \\ c\frac{f_{12}}{2N_e}\left[\frac{f_{11}}{2N_e-1} + \frac{f_{21}}{2N_e-1}\right].\end{aligned} \tag{8}$$

Similarly, due to recombination, the probabilities of the other three haplotypes become:

$$\begin{aligned}\Pr(A_1B_2) = (1-c)\frac{f_{12}}{2N_e} + \\ c\frac{f_{11}}{2N_e}\left[\frac{f_{12}}{2N_e-1} + \frac{f_{22}}{2N_e-1}\right] + \\ c\frac{f_{12}}{2N_e}\left[\frac{(f_{12}-1)}{2N_e-1} + \frac{f_{22}}{2N_e-1}\right],\end{aligned} \tag{9}$$

$$\begin{aligned}\Pr(A_2B_1) = (1-c)\frac{f_{21}}{2N_e} + \\ c\frac{f_{21}}{2N_e}\left[\frac{f_{11}}{2N_e-1} + \frac{(f_{21}-1)}{2N_e-1}\right] + \\ c\frac{f_{22}}{2N_e}\left[\frac{f_{11}}{2N_e-1} + \frac{f_{21}}{2N_e-1}\right],\end{aligned} \tag{10}$$

and

$$\begin{aligned}\Pr(A_2B_2) = (1-c)\frac{f_{22}}{2N_e} + \\ c\frac{f_{21}}{2N_e}\left[\frac{f_{12}}{2N_e-1} + \frac{f_{22}}{2N_e-1}\right] + \\ c\frac{f_{22}}{2N_e}\left[\frac{f_{12}}{2N_e-1} + \frac{(f_{22}-1)}{2N_e-1}\right].\end{aligned} \tag{11}$$

Now, to see how mutation alters these probabilities, we let $\boldsymbol{\theta}_i^*$ be a vector of the four probabilities from Equations (8) through (11) computed using the four haplotype frequencies in $\mathbf{x}_i^T$. In modeling mutation, we assume that an A_1 or B_1 allele mutates to an A_2 or B_2 allele with probability u and an A_2 or B_2 allele mutates to an A_1 or B_1 allele with probability v. Then, haplotype probabilities following mutation can be computed as:

$$\boldsymbol{\theta}_i^{**} = \mathbf{T}\boldsymbol{\theta}_i^*, \tag{12}$$

where

$$\mathbf{T} = \begin{bmatrix} (1-u)^2 & (1-u)v & v(1-u) & v^2 \\ (1-u)u & (1-u)(1-v) & vu & v(1-v) \\ u(1-u) & uv & (1-v)(1-u) & (1-v)v \\ u^2 & u(1-v) & (1-v)u & (1-v)^2 \end{bmatrix}.$$

Furthermore, to incorporate selection in the model, a selection coefficient s that reduces the allele frequency of A_1 at locus A, which is a causal variant, is considered. Conditional on the haplotype probabilities in $\boldsymbol{\theta}_i^{**}$, i.e., $p_{A_1B_1}^{**}$, $p_{A_1B_2}^{**}$, $p_{A_2B_1}^{**}$, and $p_{A_2B_2}^{**}$, the haplotype probabilities following selection can be computed as

$$\boldsymbol{\theta}_i^{***} = \begin{bmatrix} \frac{(1-s)p_{A_1B_1}^{**}}{(1-s)(p_{A_1B_1}^{**}+p_{A_1B_2}^{**})+p_{A_2B_1}^{**}+p_{A_2B_2}^{**}} \\ \frac{(1-s)p_{A_1B_2}^{**}}{(1-s)(p_{A_1B_1}^{**}+p_{A_1B_2}^{**})+p_{A_2B_1}^{**}+p_{A_2B_2}^{**}} \\ \frac{p_{A_2B_1}^{**}}{(1-s)(p_{A_1B_1}^{**}+p_{A_1B_2}^{**})+p_{A_2B_1}^{**}+p_{A_2B_2}^{*}} \\ \frac{p_{A_2B_2}^{**}}{(1-s)(p_{A_1B_1}^{**}+p_{A_1B_2}^{**})+p_{A_2B_1}^{**}+p_{A_2B_2}^{**}} \end{bmatrix}$$

Finally, to compute the distribution of allele frequencies in generation $t+1$ using Equation (6), where the forces of the genetic drift, recombination, mutation and selection are simultaneously considered, the matrix $\mathbf{A}$ is defined such that element j of column i contains the probability of getting $\mathbf{x}_j^T$ from a Multinomial $(2N_e, \boldsymbol{\theta}_i^{***T})$ distribution, for $i, j = 1, \ldots, k$.

The value of r^2 can be computed for a sub-population or line with haplotype frequency counts in any row of $\mathbf{X}$. For example, consider the frequency counts in row 11 of the $\mathbf{X}$ matrix given in **Figure S1**, where $f_{11} = 0$, $f_{12} = 2$, $f_{21} = 1$ and $f_{22} = 1$. From these frequency counts, $Pr(A_1) = 1/2$, $Pr(B_1) = 1/4$, and $Pr(A_1B_1) = 0$ are obtained, and r^2 calculated using Equations (1) and (2) is 1/3 for a line with haplotype frequency counts in $\mathbf{x}_{11}^T$ from **Figure S1**. The distribution of r^2 in generation t is given by values of r^2 corresponding to the frequency counts in each row of $\mathbf{X}$ together with the probabilities for haplotype frequency counts given in $\mathbf{P}_t$. Note that haplotype frequency counts in rows of $\mathbf{X}$ that have indeterminate values of r^2 (i.e., when one or other allele is extinct) are not used to compute the distribution of r^2. Similarly, the distribution of r^2 with MAF cutoff is given by considering only the values of r^2 and their probabilities corresponding to the rows of $\mathbf{X}$ with MAF at the B locus $\geq$ 5%. MAF of 0.05 is used as a threshold in the present study.

Starting with an allele frequency of 0.5 at each locus and linkage equilibrium between the two loci, the expected value of r^2 was computed over generations given some values of the mutation rate, recombination rate, selection coefficient and effective population size. Mutation rate of $u = v = 0$ is used to represent the absence of mutation, while $u = v = 1 \times 10^{-9}$ is used to represent the existence of mutation. Similarly, a selection coefficient of $s = 0$ is used to represent the absence of selection, while $s = 0.1$ or $s = 0.01$ is used to represent the existence of selection.

2.2. Data Analysis

In our analysis, different population parameters are considered. Populations with N_e ranging from 5 to 50 in intervals of 5 are compared under different recombination rates, mutation rates (i.e., $u = 0$ or $u = 1e^{-9}$) and selection coefficients for the causal variant (i.e., $s = 0.1$ or $s = 0.01$ for locus A), in the absence or presence of filtering by MAF (MAF$\geq$ 0.05 for locus B). The recombination rates ranged from 0.01 to 0.1 in intervals of 0.01 and from 0.1 to 0.5 in intervals of 0.1.

The recombination rate for adjacent markers in a bovine 777 k SNP panel was considered to mimic a realistic scenario. The recombination rate for adjacent markers was estimated to be 6.25×10^{-5} using the Kosambi map function given that the mean distance between adjacent markers is around 5 kb (Espigolan et al., 2013), and the average genetic distance per unit of physical distance in bovine genome is 1.25 cM/Mb (Arias et al., 2009). A period of more than 3,000 generations was used to ensure that the haplotype frequencies had reached their equilibrium values. At equilibrium, the distribution of LD stays constant over generations (i.e., $P_{t+1} = P_t$). That distribution was used to describe, quantify, and compare r_E^2 at different equilibria in the absence or presence of mutation, selection or MAF filtering. Results from a population similar to the international black and white Holstein dairy cattle are presented in these cases, for which the effective population size is estimated to be about 50 (Kim and Kirkpatrick, 2009; Wray et al., 2019).

Furthermore, the exact distributions of r_E^2 derived under different scenarios were used to validate the approximate deterministic formulas from Sved and Feldman (1973) and Hill (1975). To generalize our results to larger effective population size, a non-linear regression model of the form $\frac{1}{a+b\times Ne\times c}$, following Sved's formula, was considered. The parameters in that model were estimated by non-linear least squares using the data generated from our transition-matrix approach.

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article.

3. RESULTS

3.1. Comparison of Sved's and Hill's Approximations to the Exact Distributions of r_E^2 Derived From the Transition-Matrix Approach

In this section, Sved's and Hill's approximations were compared to the exact distribution of r_E^2 derived from our transition-matrix approach. The relationship between equilibrium value of r^2 (r_E^2), recombination rate and mutation rate is shown in **Figure 1** for these three approaches. In the absence of mutation and selection, Sved's formula showed consistency with the exact values from our transition-matrix approach (**Figure 1A**). On the other hand, Hill's formula had a better fit to the exact values of r_E^2 (**Figure 1B**) in the presence of mutation. However, neither Sved's nor Hill's

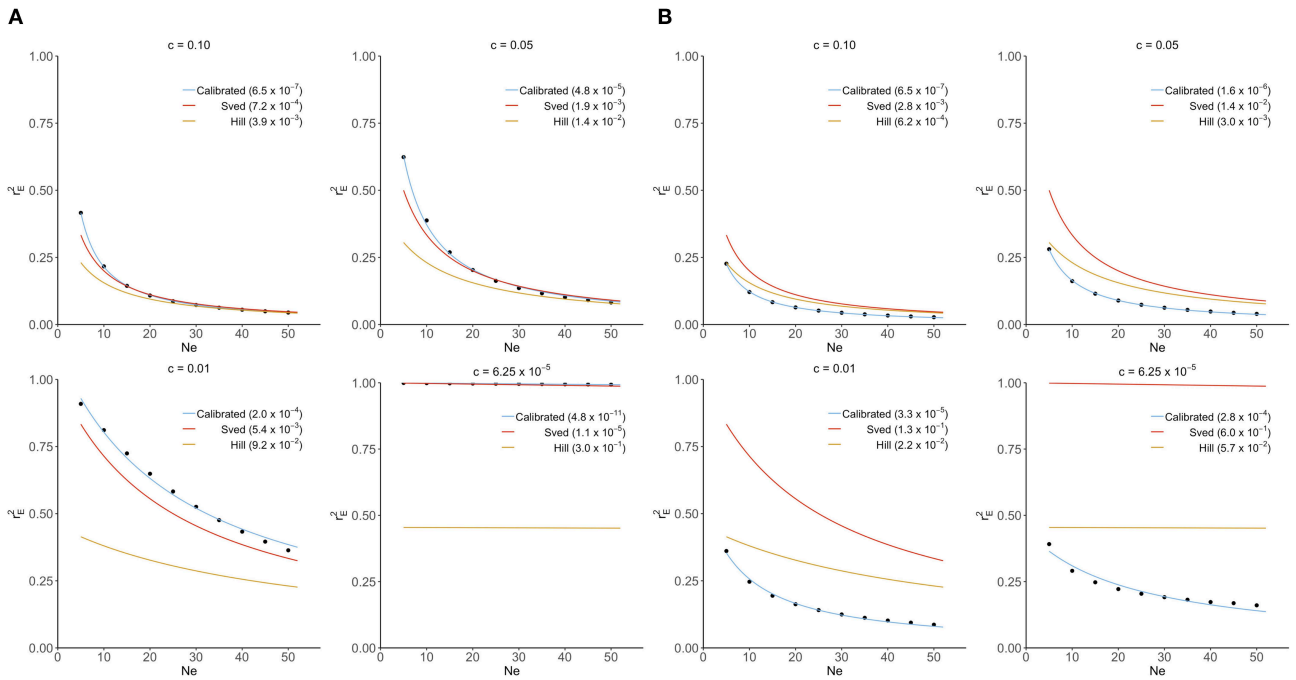

FIGURE 1 | Comparison of Sved's and Hill's approximation to exact distribution of r_E^2 (scatter points) derived from transition-matrix approach. Mean square errors are shown in the parentheses. "Calibrated" denotes the non-linear regression formula derived from the transition-matrix approach. **(A)** In the absence of mutation and selection. **(B)** In the presence of mutation but no selection.

deterministic formulae are accurate enough to describe r_E^2 in the presence of mutation. Therefore, we proposed an adjusted non-linear regression model to correct Sved's approximation to predict r_E^2 for larger effective population sizes in the presence of mutation. The mean square of errors is used to evaluate the fit of these three models and shows that our calibrated non-linear regression model was significantly better than Sved's or Hill's approximations.

3.2. Trajectory of LD Over the Generations Until Equilibrium Is Reached

The trajectory of evolving LD over generations computed from our transition-matrix approach is shown in this section, in the absence or presence of mutation and selection, where selection was applied with either a low selection coefficient of 0.01 or a high selection coefficient of 0.1. Starting with an allele frequency of 0.5 at each locus and linkage equilibrium between the two loci, expected values of LD [i.e., $E(r^2)$] over generations are displayed in **Figure 2**. Note that the trajectory may be quite different, depending on the initial conditions used.

In the absence of mutation and selection, the expected value of r^2 over generations increases to its maximum and reaches its "equilibrium" status. Value of r_E^2 increases as c decreases due to the decreasing breakdown of LD with lower recombination rates. Typically, when the recombination rate is relatively small (e.g., $c = 6.25 \times 10^{-5}$), the expected value of LD at "equilibrium" is almost equal to 1. However, at this "equilibrium" stage, frequencies of lines ($\mathbf{P}_t$) keep changing though the distribution of allele frequencies remain unchanged. This is because forces of recombination and drift are balanced in lines with determinate r^2, and frequencies of lines with determinate r^2 almost proportionally decrease toward 0.

In the absence of selection but with mutation, the expected value of r^2 increases and stays at an apparent equilibrium for several generations and then decreases to its true equilibrium value, where the mutation-drift equilibrium is reached. During the apparent equilibrium that is initially observed, the expected value of r^2 is identical to the equilibrium value in the absence of mutation and selection, where the forces of drift and recombination are balanced. When this stage is first reached, the effect of mutation is negligible because the mutation rate is low relative to the frequency of lines with segregating loci. The apparent equilibrium ends when the frequency is high for lines where one or both loci are fixed (r^2 is indeterminate) and the frequency of lines with segregating loci becomes close to the rate of mutation. Then, alleles introduced due to mutation into these lines have a noticeable effect on the distribution of allele frequencies and expected LD changes. At the end of the apparent equilibrium, the most frequent lines have only one haplotype. In the lines with two haplotypes, which have low frequencies, r^2 is either indeterminate or has value 1.0. When mutation introduces a third haplotype in a line where r^2 is 1.0, it drops in value. Further, when mutation introduces a third haplotype in a line where r^2 is indeterminate, the value of r^2 will be close to zero. Thus, as mutation becomes noticeable, the expected value of LD decreases. When the true equilibrium

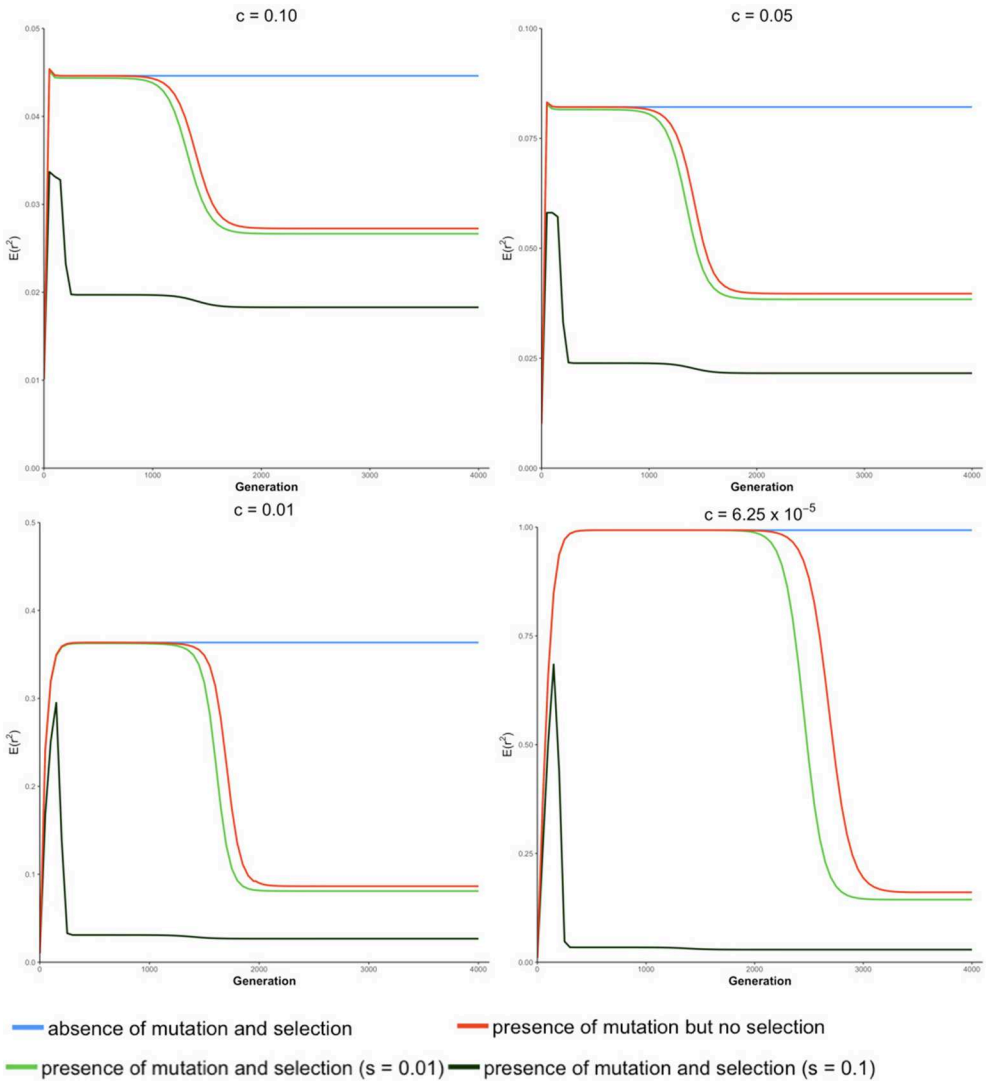

FIGURE 2 | Expected value of LD ($E(r^2)$) over generations under four different conditions at different recombination rates (c) for a population of effective size $Ne = 50$.

is reached, the loss of variation by fixation is balanced by its replenishment by mutation. In other words, the frequency of lines where loci are segregating will remain non-zero. Therefore, when mutation is present, the probabilities of allele frequencies (P_t) stay constant over generations when the true equilibrium is reached.

In the presence of both mutation and selection, lines with segregating loci become low in frequency at an earlier stage due to selection, particularly when the selection coefficient is large (e.g., s = 0.1), and thus, mutation starts to have a noticeable effect on allele frequency at an earlier stage in the presence of selection relative to when selection is absent. When selection coefficient is relatively large (e.g., s = 0.1), the expected value of r^2 reaches its maximum value early and then decreases sharply to two plateaus. When the recombination rate is low, e.g., $c \leq 0.01$, the difference between these two plateaus is small.

In the absence of selection, allele frequencies at two loci tend to be similar. However, when locus A is under selection, allele frequencies at these two loci tend to be different such that LD is lower than when selection is absent. This effect, however, is negligible when the selection coefficient is small (e.g., s = 0.01). Thus, a selection coefficient of 0.1 is used to present results when mutation and selection are present in the next section.

The expected values of r^2 over generations in the presence of filtering by MAF (i.e., MAF $\geq$ 0.05 for locus B) are shown in **Figure S2**, where a similar pattern as for $E(r^2)$ over generations are observed.

3.3. Distributions of LD at Equilibrium

In a population where mutation is present and $N_e = 50$, the equilibrium value of r^2 (r^2_E) is shown in **Figure 3**, when selection is absent or present, for different recombination rates. As explained in last section, the equilibrium values of

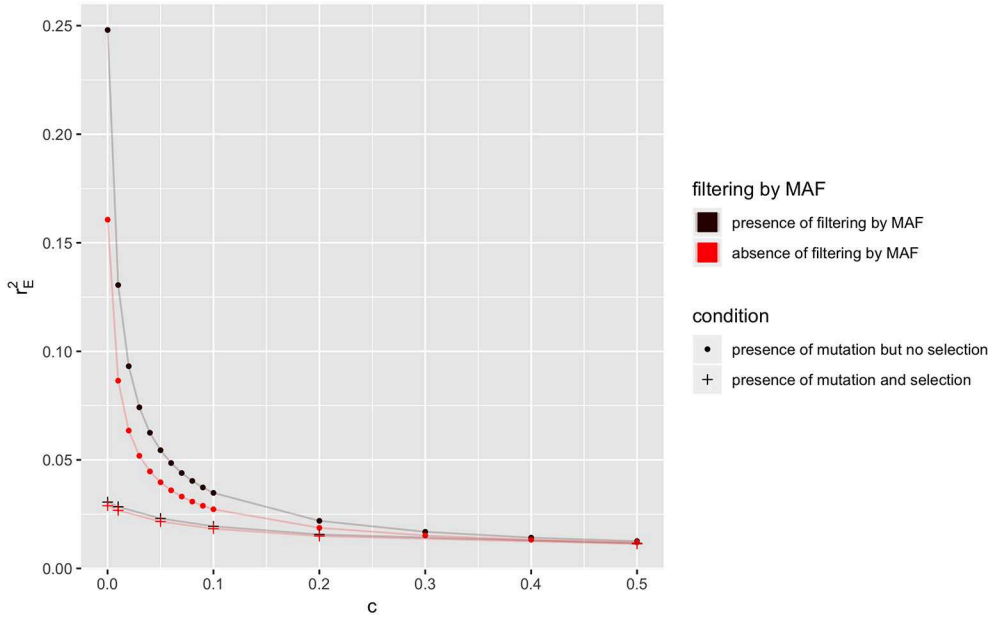

FIGURE 3 | Relationship between the expectation of LD at equilibrium (r^2_E) and recombination rate (c) in the absence or presence of selection or filtering by MAF at locus B for a population of effective size $N_e = 50$ with mutation rate $u = 1.0 \times 10^{-9}$. Only locus A is under selection with selection coefficient $s = 0.1$.

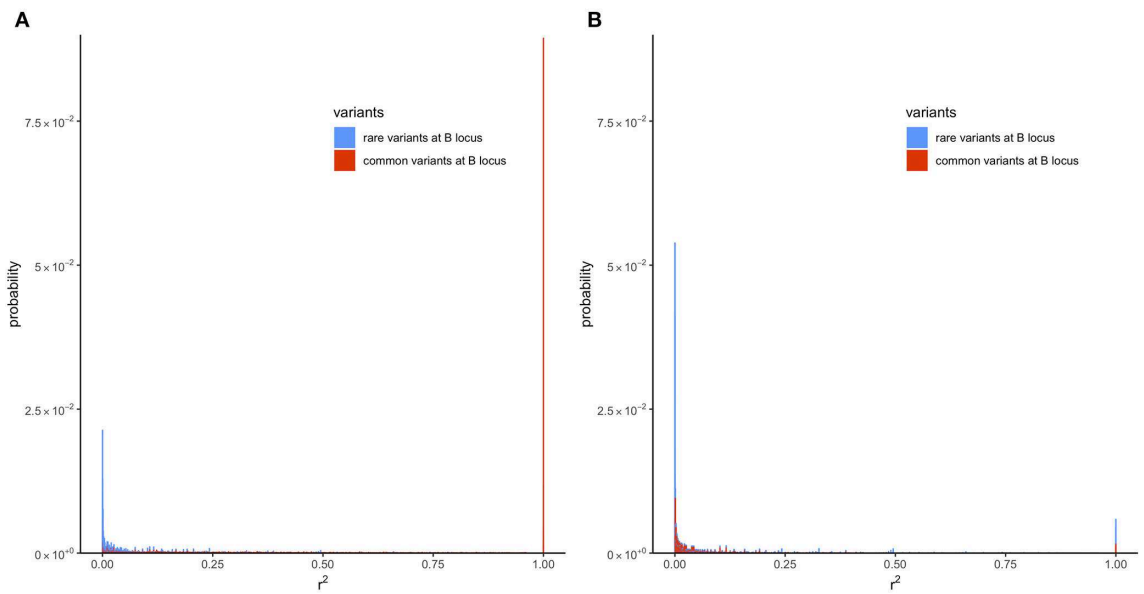

FIGURE 4 | Distribution of LD (r^2) with recombination rate (c) of 6.25×10^{-5} at equilibrium in the presence of mutation but no selection or in the presence of mutation and selection in a population of effective size $Ne = 50$. Only locus A is under selection. **(A)** Presence of mutation but no selection (s = 0). **(B)** Presence of mutation and selection (s = 0.1).

LD in the absence of selection are always higher than those in the presence of selection. The equilibrium value of LD (i.e., r^2_E) with filtering by MAF (i.e., MAF $\geq$ 5% at locus B) is higher than its corresponding value in the absence of filtering.

To understand why filtering by MAF increases r^2_E, the k lines with different haplotype frequencies were divided into two groups: lines where MAF at the locus B is $<$ 5% and lines where MAF $\geq$ 5%. Next, the transition-matrix approach was used to compute the frequency of each of these lines at equilibrium. In

lines where r^2 is defined, the equilibrium frequency of each line is plotted against its r^2 value in **Figure 4**; the color blue is used for the first group of lines where MAF at the locus B is $< 5\%$ and red is used for the second group of lines where MAF is $\geq 5\%$; frequencies in the absence of selection are given in **Figure 4A** and those in the presence of selection are given in **Figure 4B**.

In the absence of selection (**Figure 4A**), a proportion of about 0.41 of the lines were in the first group and a proportion of about 0.59 were in the second group. It can be seen from this figure that most of the r^2 values in the first group were small, where around 45% of the lines had an r^2 value <0.001. In contrast, the second group had many lines with large r^2 values; around 15% of the lines had $r^2 = 1$. This difference in the distribution of r^2 in the two groups shows why filtering by MAF at the B locus (removing lines from the first group, which had an abundance of low r^2 values) would increase the expected value of r^2.

The reason for the lower value of r^2 in the first group is that, due to filtering by MAF at the B locus, this group has a large proportion of lines with recent mutations at the B locus. Consider a line where alleles are segregating at the A locus but fixed at the B locus. Such a line would belong to neither of the groups because r^2 is not defined in a line where one locus is fixed. However, the introduction of a new allele at the B locus due to mutation will result in r^2 becoming defined, but, typically, at a very low level because it results from the association in a single haplotype. The MAF in this line with the new mutation will be $\frac{1}{2N_e}$, and it will belong to the first group provided that $\frac{1}{2N_e} < 0.05$.

Figure 4B gives the distribution of r^2 for the two groups in the presence of selection, and it can be seen that still there is a greater abundance of low r^2 values in the first group. The difference between the two groups, however, is smaller than in the absence of selection, and this explains the smaller effect of filtering on equilibrium value of r^2 when selection is present (**Figure 3**).

3.4. Extrapolation of Exact r_E^2 to Larger Population Size by Non-linear Modeling

The two existing deterministic formulas (i.e., Sved's and Hill's formula) for r_E^2 are not accurate as shown in **Figure 1**. Thus, a non-linear regression formula by recombination rate was calibrated using the exact values of r_E^2 calculated from our transition-matrix approach. r_E^2 is computed as the expectation of r^2 at equilibrium (i.e., mean r^2 weighted by the frequency of each of the k possible lines at equilibrium). In the following, this formula is referred to as the calibrated non-linear regression formula. The form of the non-linear regression formula follows that of Sved's formula as presented below:

$$r_E^2 = 1/\left(\beta_0 + \beta_1 N_e c\right). \tag{13}$$

To study the extrapolation of r_E^2 from small population sizes (i.e., $N_e \leq 50$) to predict those for a larger population size, we split the values of r_E^2 calculated from our transition-matrix approach into training and validation sets. Exact values of r_E^2 with $N_e \leq 40$ are included in the training set, and the remaining with $N_e > 40$ are used for validation. The prediction accuracy is assessed using mean square error (MSE). As shown in **Table S1**, prediction accuracy from the calibrated non-linear regression formula is substantially higher than those from Sved's or Hill's formulas. Finally, parameters in Equation (13) are estimated using all available values of r_E^2, and parameter estimates are shown in **Table 1**. Typically, the estimates of β_0 and β_1 are inversely related to the values of recombination rates.

TABLE 1 | Calibration (estimation of β_is) of the non-linear regression model under different recombination rates (c).

c	β_0	β_1
6.25×10^{-5}	2.26	1557.11
0.01	1.75	21.41
0.02	1.49	15.13
0.03	1.31	12.58
0.04	1.17	11.10
0.05	1.05	10.09
0.06	0.96	9.33
0.07	0.87	8.74
0.08	0.80	8.25
0.09	0.73	7.84
0.1	0.67	7.49
0.2	0.28	5.48
0.3	0.02	4.49
0.4	−0.19	3.85
0.5	−0.38	3.38

To further validate the accuracy of the extrapolation approach, the expected value of LD at equilibrium between adjacent markers in a 777k SNPs chip (i.e.,$c = 6.25 \times 10^{-5}$) was predicted for a Nellore cattle population, for which the effective population size is about 100 (Brito et al., 2013), using Equation (13) with $\beta_0 = 2.26$ and $\beta_1 = 1557.11$. In a real genotyped Nellore cattle population (Espigolan et al., 2013), the estimated mean and standard deviation for r_E^2 was 0.17 and 0.20, respectively. The predicted r_E^2 is 0.08, 0.449, and 0.976 using the calibrated non-linear regression formula, Hill's formula, and Sved's formula, respectively. Only the predicted r_E^2 from the calibrated non-linear regression formula falls in the range 0.17 ± 0.2, while the other two are out of the measured range from this real population.

We further studied the accuracy of the extrapolation approach for very large effective population size, e.g., $Ne = 10,000$ in human. However, r_E^2 did not align with the biological expectation, which indicates that our non-linear formulas for extrapolation of r_E^2 to extremely large population size does not work because the non-linear regression formulas are derived from data with $N_e \leq 50$.

4. DISCUSSION

A contribution of this article is to propose a computational transition-matrix approach to deriving the distribution of LD between two loci over generations in the presence of multiple genetic forces including drift, mutation, and selection. These distributions of LD are also studied in the presence of filtering by MAF (i.e., MAF $\geq$ 0.05). In addition to deriving exact

distribution of LD, several critical results emerge from the proposed approach.

1. **The expected value of LD at equilibrium decreases in the presence of selection.** Decrease of LD caused by mutation is magnified in the presence of selection due to a higher fixation rate of the favorable allele. In the presence of mutation but no selection (or in the presence of mutation and weak selection), allele frequencies at both loci are similar due to genetic drift, and the LD between them tends to remain high. Conversely, in the presence of mutation and strong selection, this phenomenon is disrupted by the selection process resulting in diverse gene frequencies at the two loci,and low LD between them is observed.
2. **Caution is needed when LD between a causal variant and a marker is inferred after filtering out marker loci with low MAF.** LD between a causal variant and a marker variant with high MAF (i.e., MAF $\geq$ 5%) is higher than that between a causal variant and all marker variants, especially in the presence of mutation but no selection. That is, r_E^2 attributed to LD between a causal variant and marker variants with high MAF is relatively higher than that between a causal variant and marker variant with low MAF which leads to the reduction of the overall expected LD. In practice, LD is sometimes inferred from molecular marker panel with MAF cutoff applied. The inferred value is usually used to estimate important population parameters (e.g., effective population size). Here we have demonstrated the potential increase of LD brought by MAF cutoff and caution is needed when inferring LD in the presence of filtering by MAF.
3. **"Fake" equilibrium may appear in the presence of mutation.** Two equilibrium stages are observed in the presence of mutation. The first "equilibrium" indicates the balance between fixation increasing LD and recombination decreasing LD in lines with determinant r^2 given the effect of mutation is negligible. When most loci become fixed at the end of the first "equilibrium" stage, change of allele frequency caused by mutation is of importance and reduces the expected LD. At the second equilibrium stage, the balance between mutation decreasing LD and fixation increasing LD is reached. Note that linkage equilibrium between two loci of allele frequency 0.5 is assumed in the initial population in **Figure 2**, and distribution of allele frequencies in the initial population affects the existence of the "fake" equilibrium.

4.1. Exact r_E^2 for Large Effective Population Sizes

The maximum effective population size (N_e) presented in this paper is 50 (Kim and Kirkpatrick, 2009; Wray et al., 2019), which is the estimated N_e for the international black and white Holstein dairy cattle population. Larger N_e was not studied in this paper due to computational limitations. When $N_e = 50$, the transition matrix $\mathbf{A}$ is square of order $k = 4N_e^2 + 8N_e = 176,851$, which requires more than 250 gigabytes memory to store. In addition, it takes around 5 h to complete the computational process for the analysis of 3,500 generations. Note that the memory complexity to store $\mathbf{A}$ is $O(N_e^6)$. When $N_e = 100$, the estimated processing memory to store the transition matrix $\mathbf{A}$ is more than 14 terabytes. This computational problem may be addressed by parallel computing. This idea, however, needs further investigation and is out of the scope of this paper. Thus, to generalize the transition-matrix approach a non-linear regression model was calibrated using exact values of r_E^2. The proposed calibrations may provide a better description of the relationship between LD, effective population size, recombination rate, and mutation.

DATA AVAILABILITY STATEMENT

The scripts used to generate the data can be found at https://github.com/Jiayi-Qu/Distribution_of_LD.

AUTHOR CONTRIBUTIONS

HC and RF conceived the study. JQ, HC, and RF undertook the analysis and wrote the draft. DG and SK contributed to the analysis. All authors contributed to the final version of manuscript, read, and approved the final manuscript.

FUNDING

Financial support was provided by the United States Department of Agriculture, Agriculture and Food Research Initiative National Institute of Food and Agriculture Competitive grant no. 2018-67015-27957.

ACKNOWLEDGMENTS

RF is grateful to Prof. William G. Hill for extensive discussions of early results from the approach presented here. This manuscript has been released as a Pre-Print at https://www.biorxiv.org/content/10.1101/794347v1.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00362/full#supplementary-material

REFERENCES

Arias, J. A., Keehan, M., Fisher, P., Coppieters, W., and Spelman, R. (2009). A high density linkage map of the bovine genome. *BMC Genet.* 10:18. doi: 10.1186/1471-2156-10-18

Brito, F., Sargolzaei, M., Braccini Neto, J., Cobuci, J., Pimentel, C., Barcellos, J., et al. (2013). In-depth pedigree analysis in a large Brazilian nellore herd. *Genet. Mol. Res.* 12, 5758–5765. doi: 10.4238/2013.November.22.2

Espigolan, R., Baldi, F., Boligon, A. A., Souza, F. R., Gordo, D. G., Tonussi, R. L., et al. (2013). Study of whole genome linkage disequilibrium

in nellore cattle. *BMC Genomics* 14:305. doi: 10.1186/1471-2164-14-305

Hill, W., and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38, 226–231. doi: 10.1007/BF01245622

Hill, W. G. (1975). Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theor. Popul. Biol.* 8, 117–126. doi: 10.1016/0040-5809(75)90028-3

Kim, E. S., and Kirkpatrick, B. W. (2009). Linkage disequilibrium in the North American Holstein population. *Anim. Genet.* 40, 279–288. doi: 10.1111/j.1365-2052.2008.01831.x

Mitchell-Olds, T., Willis, J. H., and Goldstein, D. B. (2007). Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nat. Rev. Genet.* 8, 845–856. doi: 10.1038/nrg2207

Ohta, T., and Kimura, M. (1969). Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics* 63, 229–238.

Ohta, T., and Kimura, M. (1971). Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics* 68, 571–580.

Sved, J. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol.* 2, 125–141. doi: 10.1016/0040-5809(71)90011-6

Sved, J., and Feldman, M. (1973). Correlation and probability methods for one and two loci. *Theor. Popul. Biol.* 4, 129–132. doi: 10.1016/0040-5809(73)90008-7

Sved, J. A., and Hill, W. G. (2018). One hundred years of linkage disequilibrium. *Genetics* 209, 629–636.

Sved, J. A., McRae, A. F., and Visscher, P. M. (2008). Divergence between human populations estimated from linkage disequilibrium. *Am. J. Hum. Genet.* 83, 737–743. doi: 10.1016/j.ajhg.2008.10.019

Walsh, B., and Lynch, M. (2018). *Evolution and Selection of Quantitative Traits*. Oxford:Oxford University Press.

Wray, N. R., Kemper, K. E., Hayes, B. J., Goddard, M. E., and Visscher, P. M. (2019). Complex trait prediction from genome data: contrasting EBV in livestock to PRS in humans. *Genetics* 211, 1131–1141. doi: 10.1534/genetics.119.301859

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.