





Measuring linkage disequilibrium and improvement of pruning and clumping in structured populations

Ulises Bercovich ^{1,†} Malthe Sebro Rasmussen ^{1,†} Zilong Li ² Carsten Wiuf ¹ Anders Albrechtsen^{2,*}

¹Department of Mathematical Sciences, University of Copenhagen, Copenhagen 2100, Denmark

²Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, Copenhagen 2200, Denmark

*Corresponding author: Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, Copenhagen 2200, Denmark. Email: aalbrechtsen@bio.ku.dk

[†]These authors contributed equally to this work.

Standard measures of linkage disequilibrium (LD) are affected by admixture and population structure, such that loci that are not in LD within each ancestral population appear linked when considered jointly across the populations. The influence of population structure on LD can cause problems for downstream analysis methods, in particular those that rely on LD pruning or clumping. To address this issue, we propose a measure of LD that accommodates population structure using the top inferred principal components. We estimate LD from the correlation of genotype residuals and prove that this LD measure remains unaffected by population structure when analyzing multiple populations jointly, even with admixed individuals. Based on this adjusted measure of LD, we can perform LD pruning to remove the correlation between markers for downstream analysis. Traditional LD pruning is more likely to remove markers with high differences in allele frequencies between populations, which biases measures for genetic differentiation and removes markers that are not in LD in the ancestral populations. Using data from moderately differentiated human populations and highly differentiated giraffe populations we show that traditional LD pruning biases F_{ST} and principal component analysis (PCA), which can be alleviated with the adjusted LD measure. In addition, we show that the adjusted LD leads to better PCA when pruning and that LD clumping retains more sites with the retained sites having stronger associations.

Keywords: linkage disequilibrium; heterogeneous populations; principal component analysis; Pearson's r^2 ; SNP markers

Introduction

Linkage disequilibrium (LD) is a measure of nonrandom association between alleles at different sites. In a homogeneous population, if the frequency of a haplotype carrying a particular pair of alleles is equal to the product of the frequencies of the two alleles, then the two alleles are independent of each other and are said to be in linkage equilibrium; otherwise, there is some degree of LD between the two alleles.

Drift and mutation will cause LD, whereas recombination tends to break it down. Therefore, alleles at sites located close to each other on the genome are likely to be in high LD since recombination events between close sites are rare. Conversely, alleles at sites located far apart or on different chromosomes generally have low levels of LD. However, various other biological processes create and maintain LD, including selection, inbreeding, and population structure (Slatkin 2008). In this study, we focus on the latter, where differences in allele frequencies among two or more homogeneous populations cause LD in heterogeneous populations obtained by mixing the homogeneous populations (Nei and Li 1973; Pfaff et al. 2001). LD created in this way may be observed at long genetic distances, including between chromosomes, and may change the baseline signal of no LD away from zero.

Accurate assessment of LD levels is important in many contexts. For example, the presence of LD is at the heart of genome-wide

association studies (GWAS), where the position of a nongenotyped causal SNP might be detected from LD patterns in a subset of genotyped SNPs, using a SNP-chip (Bush and Moore 2012). However, if multiple SNPs are in LD with a putative causal SNP, then it might be difficult to infer the position of the causal SNP with accuracy. Furthermore, the overall pattern of LD itself is of interest, since the way LD decays as a function of chromosomal distance is informative about the effective population size through time (Tenesa et al. 2007; Waples et al. 2016; Santiago et al. 2020) and the timing of admixture events (Moorjani et al. 2011; Loh et al. 2013). Therefore, it is tantamount to have a reliable assessment of LD levels from empirical samples.

Moreover, a wide variety of population genetic analyses assumes that sites are unlinked or are in low LD; see for example (Meisner and Albrechtsen 2022) in the case of admixture inference, and (Abdellaoui et al. 2013; Meisner and Albrechtsen 2022) in the context of detecting population structure using principal component analysis (PCA). Various methods exist to handle this issue; the most popular being LD pruning, where SNPs are removed such that all SNP pairs within a certain distance have estimated LD below a predefined threshold. However, if the estimated level of LD is wrong, one might remove or maintain the wrong SNPs too. In a sample consisting of individuals from multiple populations, SNPs with large differences in sub-population allele frequencies are more likely to be pruned away.

Received on 17 July 2024; accepted on 19 December 2024

© The Author(s) 2025. Published by Oxford University Press on behalf of The Genetics Society of America. All rights reserved. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

This is sometimes referred to as the two-locus Wahlund effect (Nei and Li 1973; Sinnock 1975; Waples and England 2011). Downstream analyses are also affected by this. For example, measures of population differentiation such as the fixation index (F_{ST}) (Li et al. 2019) become unreliable, and a range of other methods that quantify genetic difference turn untrustworthy as well, as we show in this study.

We address the above problems by introducing a true (theoretical) measure of population LD and a true measure of sample LD, based on an arbitrary evolutionary model. Secondly, we propose a way to estimate the sample LD and the population LD, and show how this leads to an analog of Pearson's correlation measure. In essence, the standard Pearson's correlation measure is the correlation coefficient between the residuals, calculated under the assumption of a homogeneous population. If this is not the case, we calculate the residuals under the proposed model, for example a model with population structure.

We apply the theory and method to a standard model of population structure (Pritchard et al. 2000; Alexander et al. 2009) and use PCA to estimate parameters and to predict individual genotypes. Then, we calculate the residual differences between observed and predicted genotypes to calculate the LD score. Using real data from populations with moderate differentiation such as humans, and high differentiation such as giraffes, we show that Pearson's standard LD measure is inflated by population structure, which causes biases in downstream analyses. We demonstrate that our proposed adjusted LD measure greatly reduces these biases.

Methods

Standard LD measures

For measuring LD between SNPs in a homogeneous population, standard statistics are based on haplotype frequencies (Hill and Robertson 1968). Given two diallelic SNPs at position s and t , the (theoretical or true) haplotype covariance is

$$(D_{std})_{st} = p_{st} - p_s p_t,$$

where p_{st} is the probability of having both reference alleles at positions s and t , and p_s and p_t are the probabilities of having the reference allele at s and t , respectively. This measure is then used to define the haplotype squared correlation as

$$(\rho_{std}^2)_{st} = \frac{(D_{std})_{st}^2}{(D_{std})_{ss}(D_{std})_{tt}}.$$

These two quantities might then be estimated using empirical haplotype frequencies (Weir 1997). However, both the theoretical quantities and the estimates fall short if the population is not homogeneous but has structure.

LD in admixed populations

For admixed populations, we are interested in a measure of LD that takes into account the genetic heterogeneity between sub-populations. The main difference from the homogeneous case is that the parameters describing an individual genetic composition are now in part private to the individual, and depend on the specific admixture proportions of the individual, that is, D_{std} and ρ_{std} are no longer meaningful quantities.

We propose an ancestry adjusted measure of LD between genomic sites. Since genomic data are generally unphased, the measure is defined from genotype data rather than haplotype data.

We first present a theoretical measure of true LD and afterwards provide means to estimate 'observed' genotypes.

Sample LD

Let n be the sample size and $G_{is}, G_{it} \in \{0, 1, 2\}$, $i = 1, \dots, n$, be the number of reference alleles in two sites, s and t , respectively. Furthermore, let $\mathbf{p}_{st}^i = (p_{st}^i, p_s^i, p_t^i)$ be the parameters of individual $i = 1, \dots, n$, where p_{st}^i is the probability that a haplotype carries both reference alleles (individual haplotype frequency), and similarly p_s^i, p_t^i are individual allele frequencies for the two sites. For convenience, we phrase the theory in terms of unrestricted parameters \mathbf{p}_{st}^i , that could all be different. However, in practise, the parameters depend on an underlying model; for example, if the model consists of a single homogeneous population, then the parameters are the same for each individual, and if the model consists of a mixture of homogeneous sub-populations, then the parameters for allele frequencies are the same within each sub-population, while the admixture proportions are private to the individual. As described in the section below, there are several ways to estimate these parameters but we suggest to use PCA. Whether one model or another is adopted, might subsequently influence how the parameters are estimated. It is however important to have a theoretical measure that does not rely on a particular model.

We assume the number of reference alleles in each site is a sum of independent parental gametic contributions $G_{is} = H_{is}^1 + H_{is}^2$ and $G_{it} = H_{it}^1 + H_{it}^2$. Further assuming that the two haplotypes are identically distributed, then the expression for the covariance between the genotype numbers is

$$\begin{aligned} \frac{1}{2} \text{Cov}(G_{is}, G_{it} | \mathbf{p}_{st}^i) &= \frac{1}{2} \text{Cov}(H_{is}^1, H_{it}^1 | \mathbf{p}_{st}^i) + \frac{1}{2} \text{Cov}(H_{is}^2, H_{it}^2 | \mathbf{p}_{st}^i) \\ &= p_{st}^i - p_s^i p_t^i. \end{aligned}$$

A measure of the (true) sample LD is thus the average covariance over all n individuals,

$$(D_{adj}^n)_{st} = \frac{1}{2n} \sum_{i=1}^n \text{Cov}(G_{is}, G_{it} | \mathbf{p}_{st}^i) = \frac{1}{n} \sum_{i=1}^n (p_{st}^i - p_s^i p_t^i), \quad (1)$$

which is adjusted for the heterogeneity in the sample. Clearly, $-1 \leq (D_{adj}^n)_{st} \leq 1$. If the parental haplotypes do not share the same distribution, then (1) becomes a sum over the $2n$ haplotypes. Moreover, if there are k separate sub-populations and n_ℓ individuals from the ℓ th sub-population, our measure of population LD agrees with that of (Nei and Li 1973),

$$(D_{adj}^n)_{st} = \frac{1}{n} \sum_{\ell=1}^k n_\ell (D_{std}^\ell)_{st}, \quad (2)$$

where D_{std}^ℓ is the standard measure of the ℓ th sub-population. Hence, the proposed measure of population LD also extends previous work on LD in sub-divided populations.

An adjusted squared correlation might be defined similarly to the standard haplotype squared correlation. If the parameter \mathbf{p}_{st}^i is the same for all the individuals, then $(D_{adj}^n)_{st} = (D_{std})_{st}$, which shows that the adjusted measure of sample LD and the standard measure of LD agrees in the case of a homogeneous population. In that case, there is no dependence on the sample size n .

Population LD

It is desirable to have a measure of LD that reflects the admixed population as such and is not attached to a specific sample of

individuals. One might imagine taking $n \rightarrow \infty$ in (1) assuming that the parameters \mathbf{p}_{st}^i for each individual follow a common distribution \mathbf{P}_{st} . It is thus natural to replace the average in (1) with an expectation. For this, let \mathbf{p}_{st} be a random draw from \mathbf{P}_{st} , and let $G_s, G_t \in \{0, 1, 2\}$ be the number of reference alleles in the two sites, drawn according to \mathbf{p}_{st} . Then, we define the (true) population LD as

$$(D_{adj})_{st} = \frac{1}{2} \mathbb{E}[\text{Cov}(G_s, G_t | \mathbf{p}_{st})],$$

where the expectation is with respect to \mathbf{P}_{st} . Also, the population LD might be written in terms of the haplotype covariance.

The measure for the sample LD and the population LD are linked through the law of large numbers that says $(D_{adj}^n)_{st} \rightarrow (D_{adj})_{st}$ as $n \rightarrow \infty$. Thus, population LD is reflected in sample LD, if the sample size is large enough. An adjusted squared correlation might be defined similarly to the standard haplotype squared correlation.

If the distribution \mathbf{P}_{st} is degenerate (always takes the same value), which is the case if the population is homogeneous, then $(D_{adj})_{st} = (D_{std})_{st}$. Thus, the proposed measure of population LD extends the standard measure of LD. In the case of k separate sub-populations with relative sub-population sizes w^ℓ , $\ell = 1, \dots, k$, then the population LD is a sum over the standard LDs of the sub-populations, similarly to (2) for the sample LD,

$$(D_{adj})_{st} = \sum_{\ell=1}^k w^\ell (D_{std}^\ell)_{st}, \quad (3)$$

where $(D_{std}^\ell)_{st}$ is the standard LD measure of the ℓ th sub-population. In fact, a stronger result holds.

Theorem 1. Let w_{st}^ℓ , $\ell = 1, \dots, k$, be the probability that the two alleles of a haplotype are both from sub-population ℓ . Then, $\sum_{\ell=1}^k w_{st}^\ell \leq 1$, and

$$(D_{adj})_{st} = \sum_{\ell=1}^k w_{st}^\ell (D_{std}^\ell)_{st}.$$

Furthermore, by Jensen's inequality, $(D_{adj})_{st}^2 \leq \sum_{\ell=1}^k w_{st}^\ell (D_{std}^\ell)_{st}^2$.

The proof of Theorem 1 and the proofs of the following statements can be found in the supplementary material.

Theorem 1 says that in an arbitrary admixed population, population LD might be seen as a weighted decomposition of LD within the sub-populations. If each individual has genetic material from only one sub-population, then $w_{st}^\ell = w^\ell$ is the relative population sizes, as in (3) (if your mother is from sub-population ℓ , then so is your father). If each individual chooses parents randomly from all sub-populations, then $w_{st}^\ell = (w^\ell)^2$ is the probability to choose both parents from the same sub-population (if your mother is from sub-population ℓ , then so is your father with probability w^ℓ).

In particular, $(D_{adj})_{st}$ always takes a value between the smallest and the largest value of $(D_{std}^\ell)_{st}$, $\ell = 1, \dots, k$. Similarly, $(D_{adj})_{st}^2$ is always smaller than the largest value of $(D_{std}^\ell)_{st}^2$, $\ell = 1, \dots, k$. In contrast, there is no lower bound but 0. For example, in the case where there is a balanced pooling of two sub-populations with $0 < (D_{std}^1)_{st} = -(D_{std}^2)_{st}$. Then,

$$(D_{adj})_{st} = \frac{1}{2} (D_{std}^1)_{st} + \frac{1}{2} (D_{std}^2)_{st} = 0,$$

and hence $(D_{adj})_{st}^2 = 0$, even though $(D_{std}^1)_{st}^2 = (D_{std}^2)_{st}^2 > 0$.

Estimation of LD in admixed populations

To estimate the sample LD, we first introduce a model and some notation. Let G be the observed genotype data matrix of n individuals across m SNPs, where each genotype consists of two alleles, hence each entry of G is 0, 1, or 2. Let $\Pi_{is} \in [0, 1]$ be the probability that individual i has the reference allele at position s , so both G and Π are matrices of dimension $n \times m$. Π is often called the individual allele frequency. We model the marginal distribution of G_{is} as a binomial distribution

$$G_{is} \sim \text{Bin}(2, \Pi_{is}),$$

and allow for dependence between G_{is} and G_{it} (within one individual), but we do not specify a model of this explicitly. In contrast, we assume genotypes from different individuals are independent, that is, G_{is} and G_{jt} are independent for $i \neq j$, given Π_{is} and Π_{jt} . In the context of the previous section, $\Pi_{is} = p_{is}^i$, and $\Pi_{it} = p_{it}^i$, whereas we leave unspecified the form of p_{st}^i .

Based on this general model, we introduce an estimate of the ancestry adjusted sample LD by first calculating the empirical covariance of the $n \times m$ residual matrix,

$$R = G - 2\hat{\Pi},$$

where $\hat{\Pi}$ is a matrix of predicted (estimated) values of Π . By subtracting the expected genotype for each individual ($2\hat{\Pi}$), correlation between pairs of SNPs is not due to population structure, but to true LD. In this way, we purge the component of covariance generated by population structure, while keeping the linkage between SNPs. If there is no population structure in the sample, then the individual allele frequency is the same for all individuals. In this case, the residual and the residual correlation between pairs of SNPs are the standard measures.

The empirical covariance between two SNPs (columns of R), calculated from the residual matrix, further using Bessel's correction, is an estimate of the ancestry adjusted sample LD,

$$(\hat{D}_{adj})_{st} = \frac{1}{2(n-k)} \sum_{i=1}^n (R_{is} - \bar{R}_{.s})(R_{it} - \bar{R}_{.t}), \quad (4)$$

where $\bar{R}_{.s} = \frac{1}{n} \sum_{i=1}^n R_{is}$, and k is the rank of Π . The empirical squared correlation, an estimate of the adjusted squared correlation, is given by Pearson's correlation calculated on the residuals, that is

$$(r_{adj}^2)_{st} = \frac{(\hat{D}_{adj})_{st}^2}{(\hat{D}_{adj})_{ss}(\hat{D}_{adj})_{tt}}. \quad (5)$$

The estimator in (4) is applicable whenever Π is estimable. We estimate Π assuming it has a specific structure, namely that the ancestry of each individual is composed of genetic material from k ancestral populations, where we take k to be a known parameter. Specifically, we assume that the matrix Π factorizes as $\Pi = QF$, where Q is an $n \times k$ matrix of rank $k \leq n$ consisting of (true) ancestral admixture proportions, such that the proportion of individual i 's genome from population ℓ is $Q_{i\ell}$ with $\sum_{\ell=1}^k Q_{i\ell} = 1$; and F is an $k \times m$ matrix of (true) ancestral SNPs frequencies, such that the frequency of the reference allele of SNP s in the ancestral population ℓ is $F_{\ell s}$. Hence, the probability that an individual i has the reference allele in site s is $\Pi_{is} = \sum_{\ell=1}^k Q_{i\ell} F_{\ell s}$. This is similar to

admixture models proposed in the literature (Pritchard et al. 2000; Alexander et al. 2009) and models based on PCA (Meisner and Albrechtsen 2019; Meisner et al. 2021). The rank condition on Q is for reasons of identifiability, to be able to disentangle the admixture proportions from all k ancestral populations for each individual.

To estimate $\Pi = QF$, different methods might be used. We distinguish between whether Q is known (e.g. when the population is homogeneous, $k = 1$ and Q is a vector of length n with only ones) or unknown. In the former case, we use linear regression to obtain an estimate of Π : $2\hat{\Pi} = PG$, where $P = Q(Q^T Q)^{-1} Q^T$ is an $n \times n$ matrix of rank k , which is the projection onto the column space of Q . Here, M^T denotes the transpose of a matrix M . If Q is unknown, one might use PCA to estimate Π by projecting G onto the first k principal components (Chen and Storey 2015; Conomos et al. 2016; Meisner and Albrechtsen 2019; Meisner et al. 2021). If so, then $2\hat{\Pi} = \hat{P}G$, where \hat{P} is an estimate of P (van Waaij et al. 2023). Alternatively, one might estimate Π by first estimating Q and F , as suggested in (Pritchard et al. 2000; Alexander et al. 2009). In that case, an estimate of Π can be obtained either as $\hat{\Pi} = \hat{Q}\hat{F}$ or as $2\hat{\Pi} = \hat{P}G$, where $\hat{P} = \hat{Q}(\hat{Q}^T \hat{Q})^{-1} \hat{Q}^T$ is an estimate of P .

Whenever an estimate \hat{P} of P is available, we have

$$\hat{D}_{\text{adj}} = \frac{1}{2(n-k)} G^T (I - \hat{P}) (I - J/n) (I - \hat{P}) G \quad (6)$$

(potentially with \hat{P} replaced by P , if Q is known), where J is an $n \times n$ matrix with all entries equal to one, see [Supplementary Lemma S1](#). In the case Q is known, (6) reduces to

$$\hat{D}_{\text{adj}} = \frac{1}{2(n-k)} G^T (I - P) G = \frac{1}{2(n-k)} (G^T G - (PG)^T PG), \quad (7)$$

using that $P^2 = P$ for a projection matrix. Since PG is an estimator of the expectation of G , then the second expression of \hat{D}_{adj} above resembles that of D_{adj}^n .

If the population is homogeneous ($k = 1$ and Q is a vector of ones), then \hat{D}_{adj}^n agrees with the unbiased estimator of D_{std} , based on unphased genotype data, suggested by (Ragsdale and Gravel 2019) (there is a factor 2 missing in their expression for $\hat{\Delta}$ on p931), and Burrow's estimator of D_{std} , also based on unphased genotype data (Cockerham and Weir 1977; Weir 1997). Hence, the estimator \hat{D}_{adj}^n is an extension of known estimators for LD. Similarly, in the case of a sub-divided population into k separate sub-populations with n_ℓ individuals from the ℓ th sub-population, with $\ell = 1, \dots, k$, and

$$Q = \begin{pmatrix} Q_1 & 0 & \dots & 0 \\ 0 & Q_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Q_k \end{pmatrix}, \quad (8)$$

where Q_ℓ is a column vector of length n_ℓ with all ones, and the 0s are null vectors of matching lengths, then $(\hat{D}_{\text{adj}}^n)_{\text{st}}$ agrees with the estimator suggested in Nei and Li (1973), when adapted to unphased genotype data using Burrow's estimator.

Properties of \hat{D}_{adj}

It remains to connect the estimated ancestry adjusted LD, \hat{D}_{adj} , to the true population LD, D_{adj} . Under mild conditions, we show that the empirical LD measure \hat{D}_{adj} stabilizes as the number of SNPs m become large (with fixed sample size n), and that for

unlinked SNP pairs, it converges on average to zero. If the admixture proportions are known (rather than estimated), then we show this average is plain zero, irrespective of the number of SNPs available; in accordance with our intuition that unlinked SNPs are not in LD (Theorem 2). In Theorem 3, we go further and show that if in addition, the number of individuals is large, then we recover the true population LD. In the following, we add superscripts n, m to emphasize the dependence on n, m , the size of the data matrix.

We explore the case where the sample size n and the matrix Q are fixed (none random), even though the admixture proportions might be known or unknown (and thus needs to be estimated). In that case, we simply write $\mathbb{E}[\hat{D}_{\text{adj}}^{n,m}]$ for the expectation of $\hat{D}_{\text{adj}}^{n,m}$. On the contrary, when n is large, that is, when $n \rightarrow \infty$, we consider the admixture proportions of each individual as a random draw from the distribution given by the population.

Theorem 2. Assume $\hat{P}^{n,m} \rightarrow P$ as $m \rightarrow \infty$ with n fixed. Then, it holds that

$$\begin{aligned} (\hat{D}_{\text{adj}}^{n,m})_{\text{st}} &\rightarrow \frac{1}{2(n-k)} G_{\text{s}}^T (I - P) G_{\text{t}} \quad \text{for } m \rightarrow \infty, \\ \mathbb{E}[(\hat{D}_{\text{adj}}^{n,m})_{\text{st}}] &\rightarrow C_{\text{st}} = \frac{1}{2(n-k)} \sum_{i=1}^n (I - P)_{\cdot i} H_{\text{st}} (I - P)_{\cdot i}^T, \end{aligned}$$

where H_{st} is the $n \times n$ diagonal matrix with diagonal elements $\text{Cov}(G_{1s}, G_{1t}), \dots, \text{Cov}(G_{ns}, G_{nt})$, the LD between the two SNPs for each individual.

In particular, if G_{s} and G_{t} are independent, then

$$\mathbb{E}[(\hat{D}_{\text{adj}}^{n,m})_{\text{st}}] \rightarrow 0 \quad \text{for } m \rightarrow \infty.$$

Moreover, if Q is known, then $\hat{P}^{n,m} = P$ and the above holds with \rightarrow replaced by $=$ without the need of $m \rightarrow \infty$.

The limiting expression of $(\hat{D}_{\text{adj}}^{n,m})_{\text{st}}$ in Theorem 2 is identical to that of $(D_{\text{adj}}^n)_{\text{st}}$ in (7) with the same interpretation. Moreover, in the case where we consider the pooling of k separate sub-populations, with n_ℓ individuals in sub-population ℓ as in (8), C_{st} is the pooled covariance of each sub-population with the Bessel's correction (see [Supplementary Lemma S2](#)), that is,

$$C_{\text{st}} = \frac{\sum_{\ell=1}^k (n_\ell - 1) (D_{\text{std}}^\ell)_{\text{st}}}{2(n-k)},$$

which we can compare with (3). In particular, C_{st} agrees with the sample LD D_{adj}^n when there is only one ancestral population.

Conditions for when $\hat{P}^{n,m}$ converges are given in van Waaij et al. (2023), as well as a comparison of the use of the different procedures to estimate \hat{P} from estimates $\hat{\Pi}$ or \hat{Q} of Π and Q , respectively. In particular, the PCA approach suggested by Chen and Storey (Chen and Storey 2015; Cabrer0s and Storey 2019) guarantees convergence. Under additional conditions, the same holds for standard PCA based on the mean normalized genotype data matrix (van Waaij et al. 2023). Empirically, convergence seems to hold irrespective the PCA approach used, or whether some other method, for example (Alexander et al. 2009), is applied to obtain $\hat{\Pi}$ or \hat{P} . This is important because in practice on large data sets, the PCA approach of Chen and Storey (2015); Cabrer0s and Storey (2019) has severe

computational limitations. In data analysis, we used mean normalized PCA, as this is conventionally used.

Theorem 3. *If Q is known, then for any pair of SNPs s and t ,*

$$(\widehat{D}_{\text{adj}}^{n,m})_{st} \rightarrow (D_{\text{adj}})_{st} \quad \text{for } n \rightarrow \infty,$$

the population LD.

When Q is known, as in Theorem 3, then $(\widehat{D}_{\text{adj}}^{n,m})_{st}$ does not depend on m . For the case of r_{adj}^2 , defined in (5), we can extend the theorem to show that, if $(D_{\text{adj}})_{ss}$ and $(D_{\text{adj}})_{tt}$ are both nonzero,

$$(r_{\text{adj}}^2)_{st} \rightarrow \frac{(D_{\text{adj}})_{st}^2}{(D_{\text{adj}})_{ss}(D_{\text{adj}})_{tt}} \quad \text{for } n \rightarrow \infty,$$

the square of the population correlation.

Sample size correction for mean r^2

When calculating the mean squared correlation coefficient the sample size becomes important because this measure is biased for finite sample sizes. There are several suggested methods for correcting this bias, but none of them perform perfectly (Ragsdale and Gravel 2019). We choose to use the method used in LD score regression (Bulik-Sullivan et al. 2015) due to its simplicity. The correction is given by

$$\tilde{r}^2 = r^2 - \frac{1 - r^2}{n - 2},$$

where r^2 is the calculated squared correlation coefficient and \tilde{r}^2 is the bias-corrected. However, it should be noted that it is not trivial to correct for this bias (Ragsdale and Gravel 2019) and that other methods also exist that perform similarly well (Waples 2006; Ragsdale and Gravel 2019) in mitigating the upward bias.

Results

In the following sections, we compare adjusted LD to standard LD on real data. We are interested in the measures themselves as well as their effects on downstream analyses when used for pruning and clumping.

Data

To illustrate the problems with standard LD in the presence of population structure, we use two datasets: one with moderately differentiated populations and one with a large amount of differentiation. In both cases, we have high-quality SNP and genotype calls from medium or high depth whole-genome sequencing data so that no prior SNP ascertainment was done other than quality control. An overview is shown in Fig. 1.

First, for the case of moderate population structure, we use the high-quality, human data from the 1000 Genomes Project (Byrska-Bishop et al. 2022). Specifically, we used 50 random, unrelated individuals from each of the CEU (Utah residents with Northern and Western European ancestry), YRI (Yoruba in Ibadan, Nigeria), and ASW (African ancestry in Southwest US) populations. The latter was chosen because the African Americans represent an admixed population with European and East African ancestry. We used PLINK (Purcell et al. 2007; Chang et al. 2015) to subset the data and remove sites with minor allele

frequency (MAF) less than 5%. The resulting dataset contains approximately 10 M common variants.

Secondly, to contrast the use of the adjusted measure in the first dataset, we chose nonhuman data with highly differentiated sub-populations to exemplify a case where there is a strong presence of population structure. This second dataset consists of whole-genome sequencing (20x coverage) of three populations of giraffes (Coimbra et al. 2021; Bertola et al. 2024), which we refer to as Masai ($n = 5$), Reticulated ($n = 12$), and Southern ($n = 12$). In comparison to the human dataset, the sample sizes are lower, and the population structure significantly greater—indeed, these groups might be considered different species (Bertola et al. 2024), though we use the term populations throughout.

Measures of LD

There are many ways to calculate LD (Ragsdale and Gravel 2019). We choose to focus on the squared correlation coefficient as this is an often used measure, and because it is used in LD pruning. We used the r^2 obtained directly from the covariance matrix and when adjusting the r^2 , we used PCA based on mean centered genotypes. However, other approaches for PCA and r^2 might be used, but in our analysis, they performed similarly on the above data sets (see Supplementary Figs. S1 and S2).

We begin by comparing r_{std}^2 and r_{adj}^2 LD between variants on different chromosomes. As described, standard measures of LD, including r_{std}^2 , are expected to find LD between sites on separate chromosomes in the presence of population structure, despite little or no LD being present in the ancestral populations prior to mixing. In contrast, we expect r_{adj}^2 to be generally close to zero in the cross-chromosome case, since it adjusts for this effect.

To investigate, for each dataset we thinned the data on chromosomes 1 and 2 to 10% and sampled 109,076 and 82,342 SNP pairs for human and giraffe, respectively. For each of these pairs, we calculated r_{std}^2 and r_{adj}^2 . The results (Fig. 1) confirm that the adjusted measure significantly reduces the amount of cross-chromosome LD measured. For example, on the giraffe dataset, the estimated mean r_{std}^2 is 0.19, whereas for the adjusted measure, the mean r_{adj}^2 is 0.018. As expected, the difference between adjusted and standard LD is greater on the giraffe dataset with greater population differentiation, but it is likewise apparent in the human data (Fig. 1).

We then turn to demonstrate that adjusted LD is meaningful within chromosomes, and not just a decreased measure of LD. When there is LD in the ancestral population, we expect both methods to capture this, but we expect standard LD to plateau to a higher level with increasing distance, since the relative importance of population structure on r_{std}^2 is larger at longer ranges. This can be seen in the sample size corrected LD decay curves shown in Fig. 1 for both standard and adjusted r^2 up to distances of 5 Mb. (The same data are shown without sample size correction in Supplementary Fig. S3. Standard LD curves for the separate populations are shown in Supplementary Figs. S4 and S5, with SNPs private to each population having been removed.) On the giraffe data, the standard LD curve is significantly shifted up relative to the adjusted measure. Again, the effect is less visually apparent on the human dataset, but we note that although smaller, the difference is due to sites that are particularly informative of population structure and hence are expected to exert an outside influence on certain analyses. We return to this point when looking at pruning below.

Additional studies to explore the behavior of the method under adversarial conditions are also presented. First, we analyzed the case where the convergence of $\widehat{D}_{\text{adj}}^{n,m}$ is challenged by lowering m ,

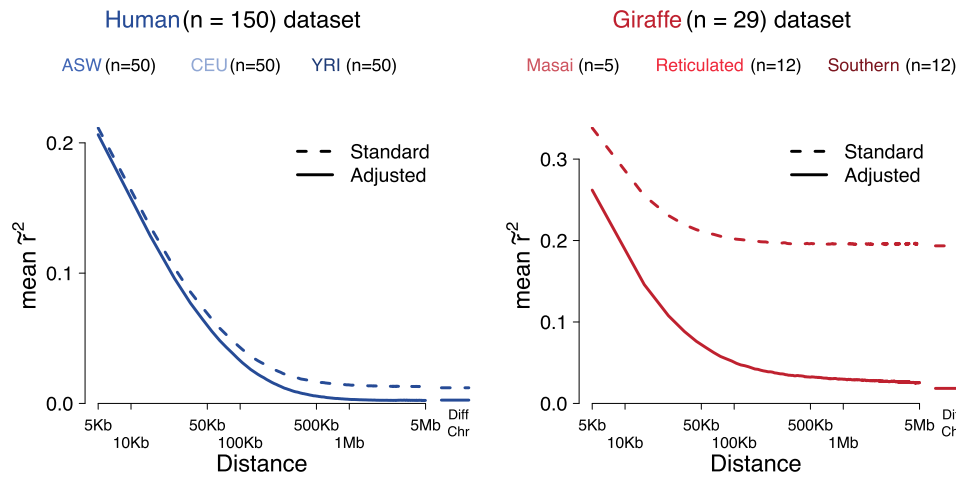


Fig. 1. Comparison of LD measures. An overview of the number of individuals in the two datasets and the three sub-populations included in each of them. LD decay curves of standard and adjusted \hat{r}^2 . Mean \hat{r}^2 shown in bins based on a sliding window over sites up to 5 Mb apart. Standard and adjusted \hat{r}^2 were also calculated for 109,076 (giraffe) and 82,342 (human) randomly sampled cross-chromosome pairs of sites.

the number of SNPs. Nevertheless, we see in [Supplementary Fig. S6](#) that the LD curves for the method stay unbiased for low m . We can also notice the loss of smoothness for the curves, something expected considering that the amount of pairs of SNPs for each distance decreases drastically when, for example, $m = 5000$. Secondly, we computed $\hat{D}_{adj}^{n,m}$ for different values of k , the parameter given by the number of ancestral populations. We noticed that underestimating k can be problematic as we still keep the influence of the population structure on the LD measure. On the other hand, overestimating k , and thus using too many principal components, has a minor effect compared to underestimating, as we can see in [Supplementary Fig. S7](#).

Effects of pruning

The LD measure has an effect on LD pruning and analyses based on pruned data. To investigate, we implemented a pruning algorithm like the one used in PLINK ([Purcell et al. 2007](#)). Briefly, for each SNP A, we consider all SNPs B in a window up to 100 kb ahead. For each SNP B in the window, starting with the closest, we calculate r^2 (either adjusted or standard) and remove the SNP with the lowest MAF if the r^2 value is above a certain threshold e.g. 0.5. The process is repeated until either the starting SNP A is removed in this way, or the end of the window is reached, and the window is then moved one SNP forward. Pruning occurs separately for each chromosome.

We pruned both datasets using either r_{std}^2 or r_{adj}^2 . To see the direct effects of pruning, we then calculated the standard LD decay curve from the jointly pruned data. In addition, we extracted genotypes from the jointly pruned data and calculated the standard LD curve for the common variants for each of the three populations. The resulting LD curves ([Fig. 2a](#), without sample size correction in [Supplementary Fig. S8](#)) show that using r_{adj}^2 over r_{std}^2 has a large effect on the joint LD curve (where there is population structure), but a comparatively smaller effect on the curves when calculating the remaining LD for each separate population. In other words, where there is population structure, pruning based on r_{std}^2 removes more LD (by standard measures) than r_{adj}^2 by removing sites in LD due to population differentiation while both methods remove a similar amount of within population LD.

As a consequence of removing sites in LD due to population structure, standard pruning also removes more sites than adjusted pruning ([Fig. 2b](#)). On the giraffe dataset, more than twice

as many sites are retained when pruning based on r_{adj}^2 compared to r_{std}^2 . Again, the differences on the human dataset look less stark. For example, only about 10% more sites are kept after pruning by r_{adj}^2 on the human dataset. However, those extra sites are likely to be exactly those that are most informative of population structure. As a result, our ability to infer population structure will be diminished by standard pruning. To illustrate, we used PLINK2 ([Chang et al. 2015](#)) to compute Hudson's estimator of F_{ST} ([Hudson et al. 1992](#); [Bhatia et al. 2013](#)) for all population pairs before and after pruning. The results, in [Fig. 2c](#), show significant differences in the estimates. For example, pruning based on standard and adjusted r^2 lead to F_{ST} estimates for CEU and YRI of 0.119 and 0.144, respectively, compared to a value of 0.156 using the unpruned data. On the giraffe dataset, the deviation is more extreme, with F_{ST} values after standard pruning approximately half those resulting from adjusted pruning.

As can be seen, F_{ST} remains lower after pruning even when using the adjusted measure. A possible explanation, which we ruled out, is that this is due to the direct effects of pruning on the 1D frequency spectrum, shown in [Supplementary Fig. S9](#) before and after pruning. To test this, we resampled sites after both standard and adjusted pruning in frequency bins of 0.001 to match the unpruned frequency spectrum ([Supplementary Fig. S10](#)). Recalculating F_{ST} on these resampled datasets results in broadly similar patterns ([Supplementary Fig. S11](#)). We note that LD pruning is not typically required for standard F_{ST} calculations, so these results mainly serve to illustrate the differences between the pruning methods in the context of inferring population structure. However, even for F_{ST} , the effects of LD pruning are sometimes important. For instance, if we wish to infer the F_{ST} of the ancestral components as inferred e.g. by ADMIXTURE ([Alexander et al. 2009](#)) software, pruning is assumed by the standard admixture model. To illustrate the effect of the LD adjustment in this context, we ran ADMIXTURE for 10 different seeds on each of the two pruned datasets and recorded the F_{ST} value for the run with the highest log-likelihood. The results are included in [Fig. 2c](#) and show a similar reduction of F_{ST} when using standard LD pruning.

Based on the differences in F_{ST} and the relationship between F_{ST} and PCA ([McVean 2009](#)), it is expected that a PCA to be likewise affected by the pruning method. Running PCA in PLINK confirms that this is the case on the giraffe data ([Fig. 2d](#)). As expected from the F_{ST} results, both methods of pruning affect the shape

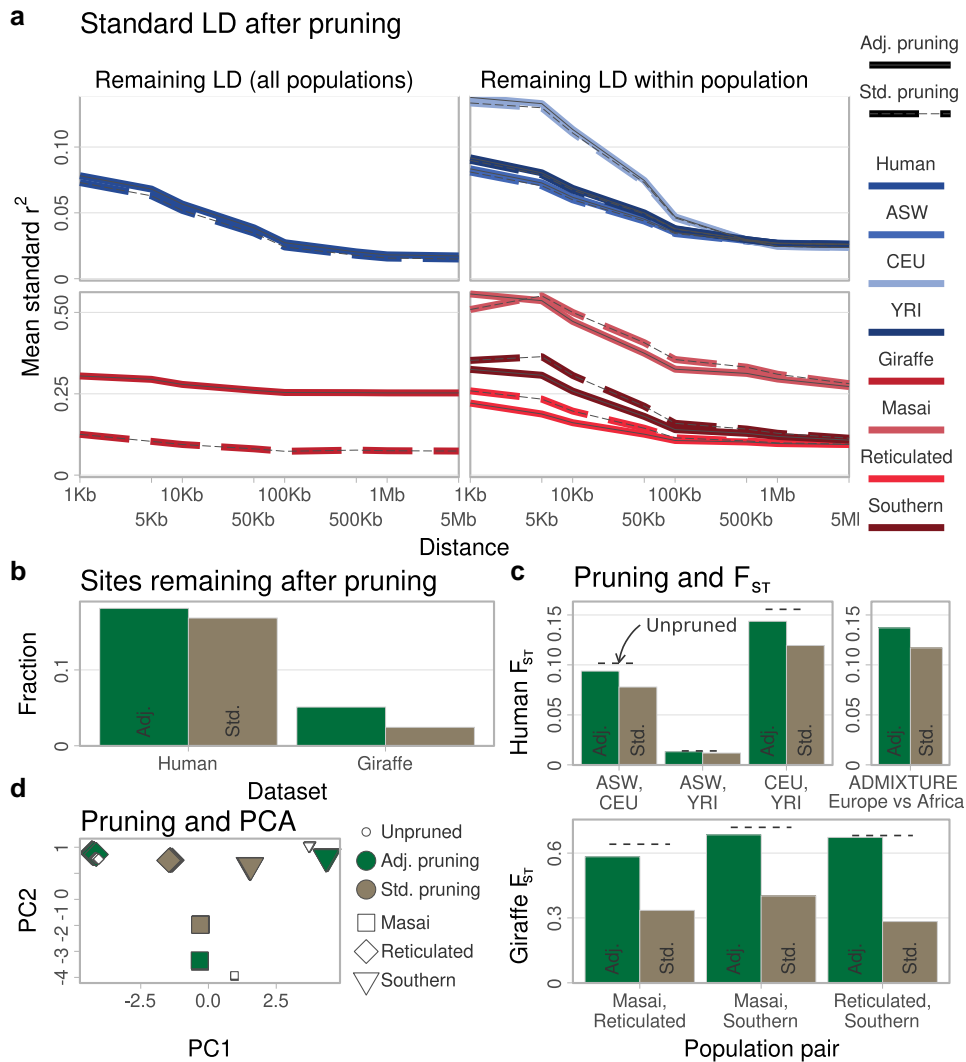


Fig. 2. Effects of pruning using different LD measures. a) Pruning is based on standard or adjusted r^2 , LD curve is standard r^2 after pruning. Standard LD decay curves after pruning based on either standard or adjusted r^2 . The joint datasets are shown, as well as LD decay curves for the constituent populations extracted from the pruned datasets. b) Fraction of sites remaining after pruning. c) F_{ST} after pruning for each pair of populations, with the unpruned F_{ST} shown for comparison. d) PCA after pruning on the giraffe dataset, with the unpruned data included for comparison. Eigenvectors scaled by corresponding eigenvalue shown.

of the principal components, though the adjusted LD pruning has a much smaller impact. The human data (Supplementary Fig. S12) show the same pattern along the first component but is harder to interpret along the second, since there is only one main axis of variation relating to population structure in the human dataset. In addition, the small size of both the giraffe and human datasets makes it hard to judge whether the eigenvectors themselves are impacted, or whether the pruning LD measure only influences the scaling.

Principal component analysis

To explore this question, we analyze the use of adjusted pruning for PCA on a larger and more complex human dataset. Specifically, based on the 1000 genomes dataset without close relatives (first and second degree) (Byrka-Bishop et al. 2022), we carried out both standard and adjusted pruning on the common variants ($MAF > 5\%$). We performed PCA on each of the resulting data sets, as well as on the unpruned data for comparison. For the adjusted LD, we chose the parameter of ancestral populations to be 9. This is based on the top principal components (PC) of the

unpruned data since these capture population structure (see Supplementary Fig. S13), where the PCs are the orthogonal directions that best seize the variation of the data.

The top four PCs are shown in Fig. 3a and b, comparing the two pruning methods to the unpruned PCs. The other top PCs are shown in Supplementary Figs. S13 and S14 which also shows that our standard pruning algorithm is comparable to the one in PLINK (Purcell et al. 2007; Chang et al. 2015). Looking at the raw eigenvectors unscaled by the eigenvalues, we do see subtle differences between the pruning methods. Moreover, as we move to the higher PCs, we begin to see that some of the PCs are shuffled, i.e. that the different axes of variation are captured in different order. The unpruned data also start to capture LD as indicated by locally high SNP loading. The question remains, however, whether these differences are meaningful and, if so, which method is preferable.

As a first answer to this question, Fig. 3c compares the cumulative variance explained for the first 10 PCs and clearly shows that PCA based on adjusted pruning explains more of the variance in the genotype data using fewer components. While informative

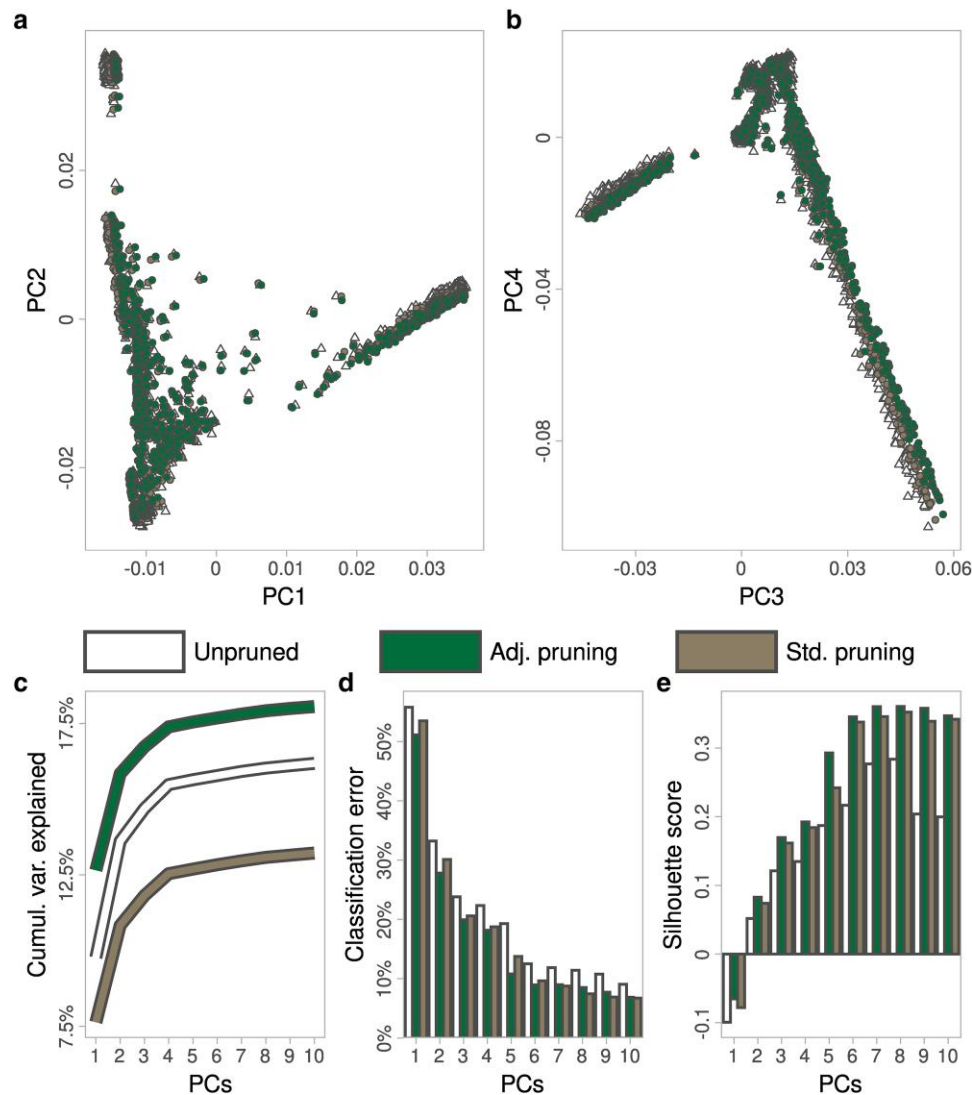


Fig. 3. PCA results on full 1000 G data set comparing no pruning, and pruning based on either standard or adjusted LD in a window of 1000 kb with a r^2 cutoff of 0.2. a, b) First four principal components, not scaled by eigenvalues. Population labels can be seen in [Supplementary Figs. S13](#) and [S14](#). c) The cumulative variance explained for the first 10 components. d) The classification error for clustering using `mclust` with the population labels. e) The mean silhouette score for clustering based on population labels.

of the structure of the decomposition overall, the variance explained measure ignores our knowledge of the population labels. In other words, it does not quantify how well each method captures population structure specifically. To investigate this, we examined specifically the separability and clusterability of the 26 populations in the PCAs, taking the given population labels from the 1000G project as ground truth. We took two distinct approaches to this.

First, we used the `mclust` R package ([Scrucca et al. 2023](#)) to cluster the populations using a variable number of the top PCs. For this, `mclust` does an automated model selection of different parametrizations of Gaussian mixtures to perform supervised clustering into the 23 population labels. Looking at the resulting classification errors ([Fig. 3d](#)), we see that adjusted pruning leads to lower classification errors for the first 5 PCs, after which the difference between adjusted and standard pruning tails off. While small, the benefit of adjusted pruning is consistently on the order of a couple of percentage points, which is a meaningful difference. The corresponding Brier scores are shown in [Supplementary Fig. S15](#) and show a similar pattern.

Second, we use `scikit-learn` ([Pedregosa et al. 2011](#)) to compute silhouette scores ([Rousseeuw 1987](#)) as a measure of clusterability by comparing distances between individuals from the same or different populations. This score is calculated from the silhouette value of each individual, which is based on the ratio between the mean distance of a fixed individual to the other individuals in the same cluster and the mean distance to the individuals from the second closest cluster for the fixed individual. Hence, a higher silhouette score means better clustering with clusters that are tighter and better separated among each other. As above, we varied the number of top PCs and averaged the silhouette score across all individuals, with results shown in [Fig. 3e](#) ([Supplementary Fig. S16](#) shows the data broken down by population). Again, the results indicate a consistent benefit of using adjusted LD pruning for PCA analysis compared to both no pruning and standard LD pruning.

Clumping

A final application for which the choice of LD measure matters is in the context of LD clumping. Here, each variant has an assigned value and the goal of LD clumping is to retain the largest subset set

of unlinked variant with the highest possible values. This is often used in association studies where clumping is performed to obtain association signals that are independent from each other. This is a standard practice when for example performing Mendelian randomization (Sanderson et al. 2022). While good methods exist for accounting for the effects of population structure while estimating the associations themselves (e.g. mixed models Zhou and Stephens 2014), these do not extend to the process of identifying unlinked loci afterwards. Briefly, clumping is one solution to this problem in which, starting from the most significant variants, all other variants within some distance are removed if they are in LD above some threshold, after which the procedure is iteratively applied to the next most significant variant that has not yet been removed. Clumping may be preferable to pruning for association studies, since clumping takes into account the inferred P -values to keep the most significant hit in each group of linked loci.

However, we anticipate that LD induced by population structure would serve to interfere with this process. Specifically, the idea is that clumping based on standard LD would discard SNPs that are not in LD when populations are considered separately and thus are associations that only appear correlated due to population structure. For most applications, the removal of such SNPs is undesired as the associated ones are needlessly removed and, as we show, because the retained SNPs have a weaker association to the traits.

To investigate, we applied clumping to summary statistics of a cross-population BMI GWAS study (Sakaue et al. 2021) (GWAS catalog (Sollis et al. 2022) accession: GCST90018947). To perform clumping, we calculated standard and adjusted LD based on the 1000 genomes data. We ran the clumping algorithm with various cutoffs on r^2 (0.0005, 0.001, 0.002, 0.005, 0.01, or 0.02) and various maximum distances within which SNPs can be removed (1 Mb, 5 Mb, or entire chromosomes). We refer to this procedure as adjusted or standard clumping, respectively, corresponding to the input LD measure. The r^2 thresholds chosen are stringent to illustrate the performance for clumping for Mendelian randomization. For reference, the default r^2 threshold used in the popular package `TwoSampleMR` (Hemani et al. 2018) is 0.001.

We find that for each combination of LD cutoff and clumping distance, adjusted clumping retains at least as many, and often more, association hits as standard LD (Fig. 4a). To quantify the strength of the association between these SNPs, Fig. 4b shows the sum of χ^2 scores for the kept SNPs, expressed as the difference between the two clumping methods. Across the range of LD cutoffs and clumping distances considered, this combined association strength of the kept SNPs is much higher using adjusted LD, suggesting that the quality of hits retained with adjusted clumping is preferable.

To visualize why adjusted LD performs better, Fig. 4c shows the clumped SNPs on chromosome 1 according to clumping methods at the 0.01 LD cutoff for the 1 Mb and entire chromosome cases (cf. Supplementary Fig. S17 for 5 Mb). Inspection of these figures supports the claim that where adjusted and standard clumping differ, adjusted clumping retains peaks with stronger associations in a particular group of linked signals. The case with no maximum distance clearly illustrates the explanation: during standard clumping, long range LD induced by population structure removes a large number of hits that are independent in each of the constituent populations. Of course, this happens to a lower degree with a short distance cap, but this is arguably an ad hoc solution to the problem that adjusted clumping addresses in a more principled way. Moreover, as the 1 and 5 Mb cases show, setting a cap only partially addresses the problem. Even though the choice of

distance seems arbitrary, it greatly influences the selection of which SNPs are kept with standard clumping; in comparison, adjusted clumping is fairly robust to the chosen cap (if any) within a reasonable band of LD cutoffs, as argued above.

Finally, we want to confirm that despite keeping more highly significant association hits, adjusted clumping removes at least as much LD as standard clumping in the constituent populations considered separately. To illustrate this, we extracted 390 individuals from Africa (ESN, GWD, MSL, YRI), Europe (CEU, GBR, IBS, TSI), and East Asia (CHB, CHS, JPT, KHV) and calculated standard r^2 among all intrachromosome pairs kept by each of the two clumping methods in the 1 Mb and $r^2 < 0.01$ scenario. The corresponding LD distributions are indeed very similar, as seen in Fig. 4d.

Discussion

In this study, we developed the mathematical foundation of a simple to use method that provides a measure of LD. This measure has some desirable properties when applied to datasets with individuals from multiple populations. Most importantly, this measure does not increase when individuals from different ancestries are analyzed jointly, unlike standard LD. We prove that for samples that come from a mixture of k ancestral populations, then the expected adjusted LD is zero ($D_{adj} = 0$) if the LD in each of the ancestral populations is also zero ($D_{std} = 0$ within ancestral populations). This is achieved by subtracting the predicted covariance given by the population structure from the standard covariance matrix of the genotypes. In practice, we estimate the predicted covariance given by the population structure inferred from the top principal components. We show that, even with a finite number of individuals, the measure is unbiased as the number of SNPs goes to infinity. The estimator is also consistent such that, if there is no LD in the ancestral populations, then each pairwise adjusted LD goes to zero as the number of individuals and SNPs goes to infinity. When there is LD in the ancestral populations, then the adjusted LD measure is also correlated if the genotype covariance is bigger than what is predicted from the population structure. In particular, for separate sub-populations, the expectation of the adjusted LD is the pooled covariance of the ancestral populations.

We evaluate the performance of the adjusted LD based on two data sets. A giraffe dataset consisting of a pooling of 3 populations with a large amount of differentiation and a dataset of 2 moderately differentiated human populations including individuals that are a mixture of the two. To evaluate the measures when there is little ancestral LD, we analyzed pairs of SNPs from different chromosomes. As expected the standard LD is high for many pairs of SNPs while it stays much closer to zero for the adjusted LD measure.

It is well known that analyzing individuals from multiple populations jointly increases standard LD. Hence, highly differentiated populations such as the giraffes, have a bigger increase in standard LD compared to more similar populations like humans. The difference is also observed on the LD decay curves, where the adjusted LD curves go towards zero as the distance between markers increases, while the standard LD curves levels at a much higher value. Because the mean estimated r^2 is biased upwards for finite sample sizes, we show results for \tilde{r}^2 , which is a simple correction based on the number of sampled individuals (Bulik-Sullivan et al. 2015). Other methods exist that in some instances perform better (Waples 2006; Ragsdale and Gravel 2019). However, we choose the simple correction due to its simplicity, but while taking into account that standard LD decay curves for the populations

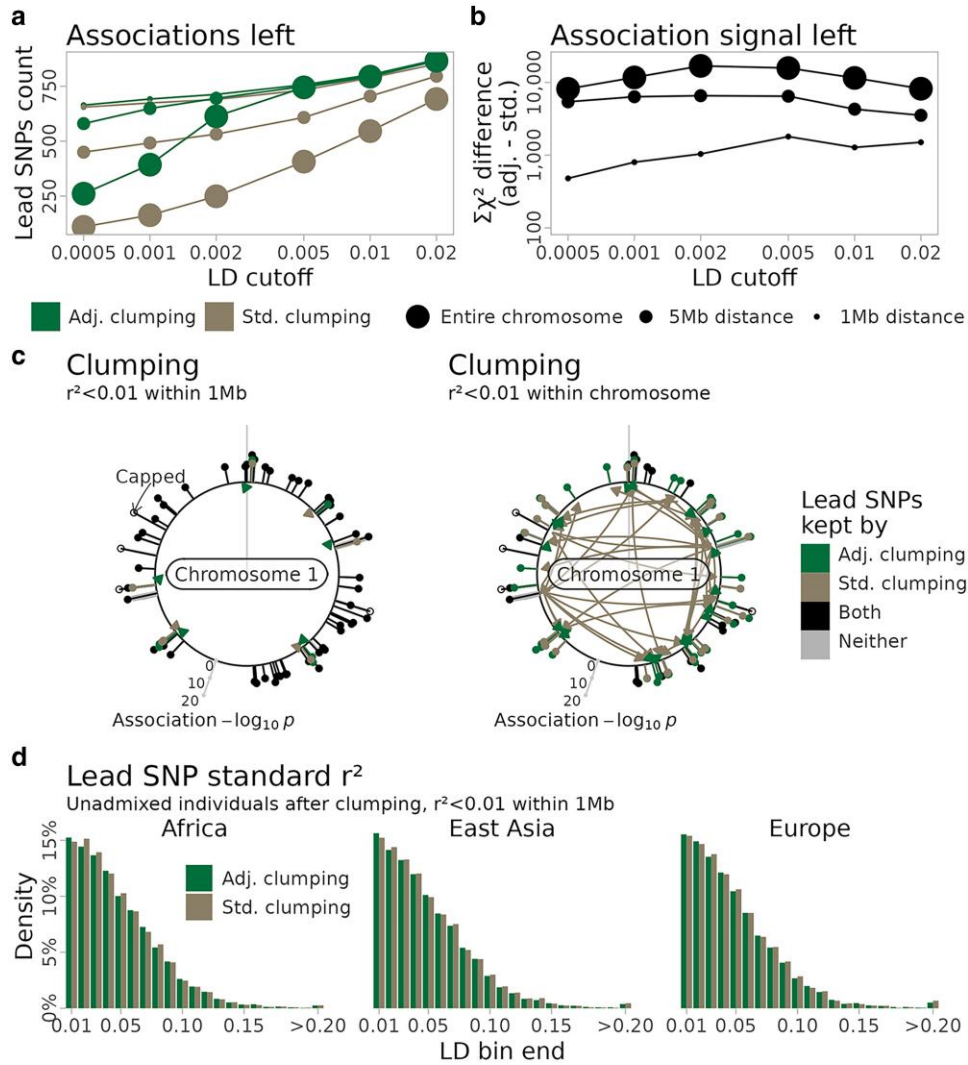


Fig. 4. Summary of stats for clumping from a multiethnic GWAS study comparing adjusted and standard LD estimated from the 1000G data. a) Inferred lead SNPs left after LD clumping using various r^2 thresholds and maximum pairwise distances. b) Association signal after clumping as quantified by the difference in the sum of the χ^2 statistics of the inferred lead SNPs. c) Chromosome 1 association signals colored by the LD measure that retains it when clumping. $-\log_{10} P$ -values are capped at 20 for open circles. Each SNP kept by exactly one method has an arrow showing from which SNP there was an association that caused the removal on the other method. For example, an adjusted clumping colored pin shows a SNP kept only with adjusted LD and has a standard clumping colored arrow from the SNP that removed it when using standard LD. d) Standard LD left after LD clumping among 390 individuals from each of the African, East Asian, and European populations.

considered separately do not always approach $1/(n-1)$ as predicted (Bulik-Sullivan et al. 2015).

Our method is not the first method that tries to overcome the issues of population structure when calculating LD. A previous study (Mangin et al. 2012) suggested first inferring the admixture proportions using STRUCTURE (Pritchard et al. 2000), and then using these as predictors in a linear model. From the linear regression, they obtain genotype residuals from which they calculate the correlation of residuals. This approach estimates the partial correlation (Lin et al. 2012) and assumes that the population structure is an observed variable. This makes the approach prone to noise in the estimation of the confounding variable. Also, the confounding variable must be linear. That is not so in our case, where we assume that population structure has been estimated.

Measuring LD is of interest in itself, but it is also often used to make datasets more appropriate for further analysis. Many methods assume that sites are independent such that there is no LD in the ancestral populations. This includes commonly used methods for inferring population structure such as ADMIXTURE (Alexander

et al. 2009), STRUCTURE (Pritchard et al. 2000), and PCA (Patterson et al. 2006). However, it also includes frequent measures in population genetics such as F_{ST} , D_{xy} (Nei 1973), heterozygosity, and kinship coefficients, which are often calculated under the assumption of no LD. Often it is not a big issue for the point estimates, since the correlation from LD mostly affects markers located close to each other. Thus, the estimators can still be consistent (Wiuf 2006). However, the uncertainty of any estimates increases and, therefore, it is often recommended to perform LD pruning. This is not an issue for most analyses that are performed within a single panmictic population. However, if the samples come from multiple ancestral populations then standard LD pruning can cause biases (Malomane et al. 2018; Li et al. 2019). This is because LD is created between alleles with different ancestral allele frequencies: the so-called two-locus Wahlund effect (Nei and Li 1973; Sinnock 1975; Waples and England 2011). Sites with a large difference in allele frequency are more likely to be pruned away because their standard LD increases more than sites with a small difference in allele frequency. Therefore, populations look genetically more similar after

standard LD pruning. We illustrate this issue by performing LD pruning on the human and giraffe populations using both the standard LD measure and the adjusted one. The standard LD pruning removed slightly more sites than the adjusted in the human populations, while in the more differentiated giraffe populations, the adjusted pruning retained more than twice the number of SNPs compared to the standard LD pruning. However, if we calculate standard LD in the ancestral populations on the remaining sites then we see that both methods have similar amounts of standard LD after pruning. Thus both methods are able to greatly reduce the amount of ancestral LD, but the adjusted LD pruning can do it while retaining more SNPs. When calculating F_{ST} from the pruned data, we see that standard LD pruning causes a huge bias for the giraffe with F_{ST} values being half of the value of the unpruned data. The effect is also apparent in the human where F_{ST} is reduced by about 20%. Using the adjusted LD for pruning alleviates most of this bias but there still appears to be some negative bias left with F_{ST} values being around 5% lower than with the unpruned data. This remaining bias could be due to allele frequency ascertainment bias which is known to bias F_{ST} (Albrechtsen et al. 2010) and PCA. However, even if we sub-sample the pruned sites to match the overall allele frequency distribution (Supplementary Fig. S9), the bias remained (Supplementary Fig. S11).

In addition to F_{ST} , we also explored the effect on PCA analysis, where standard pruned data showed fewer genetic differences between populations. This is not surprising since the top eigenvalues are proportional with F_{ST} (McVean 2009) when analyzing 3 populations. Nevertheless, the shape of the PCA was not affected for either the humans or the giraffes so in these cases the interpretation from the PCA would have remained the same.

Finally, we explored the performance of LD pruning and clumping on the diverse 1000 genomes project with individuals from 23 populations. We show that the PCA is better at reflecting population structure if the data are LD pruned, and that adjusted LD pruning performed better than standard LD pruning. The difference was very pronounced in the variance explained by each PC, but we also observed better performance in the clusterability of the populations. To illustrate the advantage of adjusted LD clumping, we applied it to GWAS summary statistics from an ethnically diverse study. Standard LD clumping on this data set removed independent association signals caused by the population structure. This can be seen when choosing large windows for clumping, where many of the strong association signals are removed due to long distance LD induced by the population structure. However, even when using a smaller maximum distance, the signals that remained after clumping were weaker compared to using standard LD clumping.

Data availability

We implemented the adjusted LD as well as pruning and clumping algorithms based on adjusted LD in the software PCAone (Li et al. 2023), which can be downloaded at <https://github.com/Zilong-Li/PCAone>.

Supplemental material available at GENETICS online.

Funding

UB and CW are supported by the Independent Research Fund Denmark (grant number: DFF-8021-00360B). MR and AA are supported by the Independent Research Fund Denmark (grant number: DFF-0135-00211B). ZL and AA are supported by the Novo Nordisk Foundation (grant number: NNF20OC0061343).

Conflicts of interest

The author declares no conflict of interest.

Literature cited

- Abdellaoui A, Hottenga JJ, de Knijff P, Nivard MG, Xiao X, Scheet P, Brooks A, Ehli EA, Hu Y, Davies GE, et al. 2013. Population structure, migration, and diversifying selection in the Netherlands. *Eur J Hum Genet.* 21(11):1277–1285. doi:10.1038/ejhg.2013.48.
- Albrechtsen A, Nielsen FC, Nielsen R. 2010. Ascertainment biases in SNP chips affect measures of population divergence. *Mol Biol Evol.* 27(11):2534–2547. doi:10.1093/molbev/msq148.
- Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19(9):1655–1664. doi:10.1101/gr.094052.109.
- Bertola LD, Quinn L, Hanghøj K, Garcia-Erill G, Rasmussen MS, Balboa RF, Meisner J, Bøggild T, Wang X, Lin L, et al. 2024. Giraffe lineages are shaped by major ancient admixture events. *Curr Biol.* 34(7):1576–1586. doi:10.1016/j.cub.2024.02.051.
- Bhatia G, Patterson N, Sankararaman S, Price AL. 2013. Estimating and interpreting F_{ST} : the impact of rare variants. *Genome Res.* 23(9):1514–1521. doi:10.1101/gr.154831.113.
- Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, Daly MJ, Price AL, Neale BM. 2015. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet.* 47(3):291–295. doi:10.1038/ng.3211.
- Bush WS, Moore JH. 2012. Chapter 11: Genome-wide association studies. *PLoS Comput Biol.* 8(12):e1002822. doi:10.1371/journal.pcbi.1002822.
- Byrskaa-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. 2022. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell.* 185(18):3426–3440.e19. doi:10.1016/j.cell.2022.08.004.
- Cabreros I, Storey J. 2019. A likelihood-free estimator of population structure bridging admixture models and principal components analysis. *Genetics.* 212(4):1009–1029. doi:10.1534/genetics.119.302159.
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 4(1):7. doi:10.1186/s13742-015-0047-8.
- Chen X, Storey J. 2015. Consistent estimation of low-dimensional latent structure in high-dimensional data. arXiv:1510.03497. doi:10.48550/arXiv.1510.03497.
- Cockerham C, Weir B. 1977. Digenic descent measures for finite populations. *Genet Res (Camb).* 30(2):121–147. doi:10.1017/S0016672300017547.
- Coimbra RTF, Winter S, Kumar V, Koepfli KP, Gooley RM, Dobrynin P, Fennessy J, Janke A. 2021. Whole-genome analysis of giraffe supports four distinct species. *Curr Biol.* 31(13):2929–2938. doi:10.1016/j.cub.2021.04.033.
- Conomos MP, Reiner AP, Weir BS, Thornton TA. 2016. Model-free estimation of recent genetic relatedness. *Am J Hum Genet.* 98(1):127–148. doi:10.1016/j.ajhg.2015.11.022.
- Hemani G, Zheng J, Elsworth B, Wade K, Baird D, Haberland V, Laurin C, Burgess S, Bowden J, Langdon R, et al. 2018. The MR-Base platform supports systematic causal inference across the human phenome. *Elife.* 7:e34408. doi:10.7554/eLife.34408.
- Hill WG, Robertson AP. 1968. Linkage disequilibrium in finite populations. *Theor Appl Genet.* 38(6):226–231. doi:10.1007/BF01245622.
- Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics.* 132(2):583–589. doi:10.1093/genetics/132.2.583.

- Li Z, Löytynoja A, Fraimout A, Merilä J. 2019. Effects of marker type and filtering criteria on q_{st} - f_{st} comparisons. *R Soc Open Sci*. 6(11):190666. doi:10.1098/rsos.190666.
- Li Z, Meisner J, Albrechtsen A. 2023. Fast and accurate out-of-core PCA framework for large-scale biobank data. *Genome Res*. 33(9):1599–1608. doi:10.1101/gr.277525.122.
- Lin CY, Xing G, Xing C. 2012. Measuring linkage disequilibrium by the partial correlation coefficient. *Heredity (Edinb)*. 109(6):401–402. doi:10.1038/hdy.2012.54.
- Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B. 2013. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*. 193(4):1233–1254. doi:10.1534/genetics.112.147330.
- Malomane DK, Reimer C, Weigend S, Weigend A, Sharifi AR, Simianer H. 2018. Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. *BMC Genomics*. 19(1):22. doi:10.1186/s12864-017-4416-9.
- Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C. 2012. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity (Edinb)*. 108(3):285–291. doi:10.1038/hdy.2011.73.
- McVean G. 2009. A genealogical interpretation of principal components analysis. *PLoS Genet*. 5(10):e1000686. doi:10.1371/journal.pgen.1000686.
- Meisner J, Albrechtsen A. 2019. Testing for Hardy–Weinberg equilibrium in structured populations using genotype or low-depth next generation sequencing data. *Mol Ecol Resour*. 19(5):1144–1152. doi:10.1111/men.v19.5.
- Meisner J, Albrechtsen A. 2022. Haplotype and population structure inference using neural networks in whole-genome sequencing data. *Genome Res*. 32(8):1542–1552. doi:10.1101/gr.276813.122.
- Meisner J, Liu S, Huang M, Albrechtsen A. 2021. Large-scale inference of population structure in presence of missingness using PCA. *Bioinformatics*. 37(13):1868–1875. doi:10.1093/bioinformatics/btab027.
- Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, Burns E, Ostrer H, Price AL, Reich D. 2011. The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genet*. 7:e1001373. doi:10.1371/journal.pgen.1001373.
- Nei M. 1973. Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci U S A*. 70:3321–3323. doi:10.1073/pnas.70.12.3321.
- Nei M, Li WH. 1973. Linkage disequilibrium in subdivided populations. *Genetics*. 75:213–219. doi:10.1093/genetics/75.1.213.
- Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet*. 2:e190. doi:10.1371/journal.pgen.0020190.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 12:2825–2830.
- Pfaff C, Parra E, Bonilla C, Hiester K, McKeigue P, Kamboh M, Hutchinson R, Ferrell R, Boerwinkle E, Shriver M. 2001. Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *Am J Hum Genet*. 68:198–207. doi:10.1086/316935.
- Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics*. 155:945–959. doi:10.1093/genetics/155.2.945.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 81(3):559–575. doi:10.1086/519795.
- Ragsdale AP, Gravel S. 2019. Unbiased estimation of linkage disequilibrium from unphased data. *Mol Biol Evol*. 37(3):923–932. doi:10.1093/molbev/msz265.
- Rousseeuw PJ. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 20:53–65. doi:10.1016/0377-0427(87)90125-7.
- Sakaue S, Kanai M, Tanigawa Y, Karjalainen J, Kurki M, Koshiba S, Narita A, Konuma T, Yamamoto K, Akiyama M, et al. 2021. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet*. 53:1415–1424. doi:10.1038/s41588-021-00931-x.
- Sanderson E, Glymour MM, Holmes MV, Kang H, Morrison J, Munafò MR, Palmer T, Schooling CM, Wallace C, Zhao Q, et al. 2022. Mendelian randomization. *Nat Rev Methods Primers*. 2:6. doi:10.1038/s43586-021-00092-5.
- Santiago E, Novo I, Pardiñas AF, Saura M, Wang J, Caballero A. 2020. Recent demographic history inferred by high-resolution analysis of linkage disequilibrium. *Mol Biol Evol*. 37(12):3642–3653. doi:10.1093/molbev/msaa169.
- Scrucca L, Fraley C, Murphy TB, Adrian ER. 2023. Model-based clustering, classification, and density estimation using mclust in R. Chapman and Hall/CRC.
- Sinnock P. 1975. The Wahlund effect for the two-locus model. *Am Nat*. 109(969):565–570. doi:10.1086/283027.
- Slatkin M. 2008. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat Rev Genet*. 9(6):477–485. doi:10.1038/nrg2361.
- Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, Groza T, Güneş O, Hall P, Hayhurst J, et al. 2022. The NHGRI-EBI GWAS catalog: knowledgebase and deposition resource. *Nucleic Acids Res*. 51(D1):D977–D985. doi:10.1093/nar/gkac1010.
- Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Res*. 17:520–526. doi:10.1101/gr.6023607.
- van Waaij J, Li S, Garcia-Erill G, Albrechtsen A, Wiuf C. 2023. Evaluation of population structure inferred by principal component analysis or the admixture model. *Genetics*. 225:1–15. doi:10.1093/genetics/iyad157.
- Waples RS. 2006. A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conserv Genet*. 7(2):167–184. doi:10.1007/s10592-005-9100-y.
- Waples RS, England PR. 2011. Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. *Genetics*. 189(2):633–644. doi:10.1534/genetics.111.132233.
- Waples RK, Larson WA, Waples RS. 2016. Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity (Edinb)*. 117:233–240. doi:10.1038/hdy.2016.60.
- Weir B. 1997. *Genetic Data Analysis II*. 2nd ed. Sunderland: Sinauer Associates.
- Wiuf C. 2006. Consistency of estimators of population scaled parameters using composite likelihood. *J Math Biol*. 53(5):821–841. doi:10.1007/s00285-006-0031-0.
- Zhou X, Stephens M. 2014. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods*. 11(4):407–409. doi:10.1038/nmeth.2848.