bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Scalable Bias-corrected Linkage Disequilibrium Estimation Under Genotype Uncertainty

David Gerard

Department of Mathematics and Statistics, American University, Washington, DC, 20016, USA

Abstract

Linkage disequilibrium (LD) estimates are often calculated genome-wide for use in many tasks, such as SNP pruning and LD decay estimation. However, in the presence of genotype uncertainty, naive approaches to calculating LD have extreme attenuation biases, incorrectly suggesting that SNPs are less dependent than in reality. These biases are particularly strong in polyploid organisms, which often exhibit greater levels of genotype uncertainty than diploids. A principled approach using maximum likelihood estimation with genotype likelihoods can reduce this bias, but is prohibitively slow for genome-wide applications. Here, we present scalable moment-based adjustments to LD estimates based on the marginal posterior distributions of the genotypes. We demonstrate, on both simulated and real data, that these moment-based estimators are as accurate as maximum likelihood estimators, and are almost as fast as naive approaches based only on posterior mean genotypes. This opens up bias-corrected LD estimation to genome-wide applications. Additionally, we provide standard errors for these moment-based estimators. All methods are implemented in the `ldsep` R package on GitHub https://github.com/dcgerard/ldsep.

1 Introduction

Pairwise linkage disequilibrium (LD), the statistical association between alleles at two different loci, has applications in genotype imputation [Wen and Stephens, 2010], genome-wide association studies [Zhu and Stephens, 2018], genomic prediction [Wientjes et al., 2013], population genetics [Slatkin, 2008], and many other tasks [Sved and Hill, 2018]. LD is often estimated from next-generation sequencing technologies, where the genotypes and haplotypes are not known with certainty [Gerard et al., 2018]. Thus, researchers typically use estimated genotypes, such as posterior mean genotypes [Fox et al., 2019], to estimate LD. However, this can cause biased LD estimates, attenuated toward zero, implying loci are less dependent than in reality. This bias is particularly strong in polyploids, and so in Gerard [2020] we derived maximum likelihood estimates (MLEs) that have lower bias and are consistent estimates of LD.

Unfortunately, the MLE approach is prohibitively slow. Researchers typically calculate pairwise LD at genome-wide scales, and the MLE approach takes on the order of a tenth of a second. Thus, for many genome-wide applications, containing millions of SNPs, LD estimation using the MLE approach would take years of computation time. This is not conducive to large-scale applications.

Keywords and phrases: attenuation bias, genotype likelihood, linkage disequilibrium, polyploidy, reliability ratio.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Here, we derive scalable approaches to estimate LD that account for genotype uncertainty (Section 2). Our methods use only the first two moments of the marginal posterior genotype distribution for each individual at each locus, which are often provided or easily obtainable from many genotyping programs. We calculate sample moments from these posterior moments, and use these to multiplicatively inflate naive LD estimates. We show, through simulations (Section 3.1) and real data (Section 3.2), that our estimates can reduce attenuation bias and improve LD estimates when genotypes are uncertain. All calculations have computational complexities that are linear in the sample size, and so these estimates are scalable to genome-wide applications.

2 Methods

In this section, we will define moment-based estimators of the LD coefficient Δ [Lewontin and Kojima, 1960], the standardized LD coefficient Δ' [Lewontin, 1964], and the Pearson correlation ρ [Hill and Robertson, 1968]. We will only consider estimating the "composite" versions of these LD measures which, advantageously, are appropriate LD measures for generic autopolyploid, allopolyploid, and segmental allopolyploid populations, even in the absence of Hardy-Weinberg equilibrium [Gerard, 2020]. We will also only consider biallelic loci, where the genotype for each individual is the dosage (from 0 to the ploidy) of one of the two alleles.

To define our estimators of LD, we assume the user provides the posterior means and variances for the genotypes for each individual at two loci. The full posterior genotype distribution for each individual is often provided by genotyping software [Gerard et al., 2018, Gerard and Ferrão, 2019, e.g.], from which these posterior moments can be obtained. If genotype posteriors are not provided, genotype likelihoods may be normalized to posterior probabilities (assuming a uniform prior) and used in what follows. Let X_{iA} and X_{iB} be the posterior means at loci A and B for individual $i \in \{1, \ldots, n\}$. Let Y_{iA} and Y_{iB} be the posterior variances at loci A and B for individual i. Our estimators are based entirely on the following sample moments of these posterior moments, which may be calculated in linear time in the sample size, n.

$$u_{xA} := \frac{1}{n}\sum_{i=1}^{n} X_{iA}, \quad u_{xB} := \frac{1}{n}\sum_{i=1}^{n} X_{iB}, \tag{1}$$

$$v_{xA} := \frac{1}{n-1}\sum_{i=1}^{n}(X_{iA} - u_{xA})^2, \quad v_{xB} := \frac{1}{n-1}\sum_{i=1}^{n}(X_{iA} - u_{xB})^2, \tag{2}$$

$$c_x := \frac{1}{n-1}\sum_{i=1}^{n}(X_{iA} - u_{xA})(X_{iB} - u_{xB}), \tag{3}$$

$$u_{yA} := \frac{1}{n}\sum_{i=1}^{n} Y_{iA}, \text{ and } u_{yB} := \frac{1}{n}\sum_{i=1}^{n} Y_{iB}. \tag{4}$$

For a K-ploid species, our LD estimators, which we derive in Section S1, are as follows. The estimated LD coefficient is

$$\hat{\Delta} := \left(\frac{u_{yA} + v_{xA}}{v_{xA}}\right)\left(\frac{u_{yB} + v_{xB}}{v_{xB}}\right)\left(\frac{c_x}{K}\right). \tag{5}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The estimated Pearson correlation is

$$\hat{\rho} := \sqrt{\frac{u_{yA} + v_{xA}}{v_{xA}}}\sqrt{\frac{u_{yB} + v_{xB}}{v_{xB}}}\frac{c_x}{\sqrt{v_{xA}v_{xB}}}. \tag{6}$$

Note that $c_x/\sqrt{v_{xA}v_{xB}}$ is the sample Pearson correlation between posterior mean genotypes. The estimated standardized LD coefficient is

$$\hat{\Delta}' := \hat{\Delta}/\hat{\Delta}_m, \text{ where} \tag{7}$$

$$\hat{\Delta}_m := \begin{cases} \min\{u_{xA}u_{xB}, (K - u_{xA})(K - u_{xB})\}/K^2 & \text{if } c_x < 0, \text{ and} \\ \min\{u_{xA}(K - u_{xB}), (K - u_{xA})u_{xB}\}/K^2 & \text{if } c_x > 0. \end{cases} \tag{8}$$

Equations (5)–(7) take the naive estimators most researchers use in practice (the sample covariance/correlation of posterior means) and inflate these by a multiplicative effect. Such multiplicative effects are sometimes called "reliability ratios" in the measurement error models literature [Fuller, 2009]. Due to sampling variability, this inflation could result in estimates that lie beyond the theoretical bounds of the parameters being estimated. In such cases, we apply the following truncations.

$$\tilde{\rho} := \begin{cases} \max\{\hat{\rho}, -1\} & \text{if } \hat{\rho} < 0 \\ \min\{\hat{\rho}, 1\} & \text{if } \hat{\rho} > 0 \end{cases} \tag{9}$$

$$\tilde{\Delta} := \begin{cases} \max\{\hat{\Delta}, -\sqrt{(v_{xA} + u_{yA})(v_{xB} + u_{yB})}/K\} & \text{if } \hat{\Delta} < 0 \\ \min\{\hat{\Delta}, \sqrt{(v_{xA} + u_{yA})(v_{xB} + u_{yB})}/K\} & \text{if } \hat{\Delta} > 0 \end{cases} \tag{10}$$

$$\tilde{\Delta}' := \begin{cases} \max\{\hat{\Delta}', -K\} & \text{if } \hat{\Delta}' < 0 \\ \min\{\hat{\Delta}', K\} & \text{if } \hat{\Delta}' > 0 \end{cases} \tag{11}$$

Standard errors are important for hypothesis testing [Brown, 1975], read-depth suggestions [Maruki and Lynch, 2014], and shrinkage [Dey and Stephens, 2018]. Because estimators (5)–(7) are functions of sample moments, deriving their standard errors can be accomplished by appealing to the central limit theorem, followed by an application of the delta method (Section S2).

Additional considerations for improving our estimates of the reliability ratios, such as using hierarchical shrinkage [Stephens, 2016], are considered in Section S3.

All methods are implemented in the `ldsep` R package on GitHub `https://github.com/dcgerard/ldsep`.

3 Results

3.1 Simulations

We compared our moment-based estimators (5)–(7) to those of the MLE of Gerard [2020] as well as the naive estimator that calculates the sample covariance and sample correlation between posterior mean genotypes at two loci. Each replication, we generated genotypes for $n \in \{10, 100, 1000\}$ individuals with ploidy $K \in \{2, 4, 6, 8\}$ under Hardy-Weinberg equilibrium at two loci with major allele frequencies $(p_A, p_B) \in \{(0.5, 0.5), (0.5, 0.75), (0.9, 0.9)\}$ and Pearson correlation $\rho \in \{0, 0.5, 0.9\}$. We then used `updog`'s `rflexdog()` function [Gerard et al., 2018, Gerard and Ferrão, 2019] to generate read-counts at read-depths of either 10 or 100, a sequencing error rate of 0.01, an overdispersion

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

value of 0.01, and no allele bias. `Updog` was then used to generate genotype likelihoods and genotype posterior distributions for each individual at each SNP. These were then fed into `ldsep` to obtain the MLE, our new moment-based estimator, and the naive estimator. Simulations were replicated 200 times for each unique combination of simulation parameters.

The accuracy of estimating ρ^2 when $p_A = p_B = 0.5$ at a read-depth of 10 is presented in Figure 1. The results for other scenarios are similar and may be found on GitHub (`https://github.com/dcgerard/ldfast_sims`). We see that the moment-based estimator and the MLE perform comparably, even for small read-depth and sample size. The naive estimator has a strong attenuation bias toward zero. This bias is particularly prominent for higher ploidy levels. For example, for an octoploid species where the true ρ^2 is 0.81, the naive estimator appears to converge to a ρ^2 estimate of around 0.25. This bias does not disappear with increasing sample size. Estimated standard errors are reasonably well-behaved, except for $\hat{\rho}$ and $\hat{\rho}^2$ when the sample size is small and the LD is large (Figure 2).

3.2 LD estimates for *Solanum tuberosum*

We evaluated our methods on the autotetraploid potato (*Solanum tuberosum*, $2n = 4x = 48$) genotyping-by-sequencing data from Uitdewilligen et al. [2013]. We used `updog` [Gerard et al., 2018, Gerard and Ferrão, 2019] to obtain the posterior moments for each individual's genotype at each SNP on a single super scaffold (PGSC0003DMB000000192). To remove monoallelic SNPs, we filtered out SNPs with allele frequencies either greater than 0.95 or less than 0.05, and filtered out SNPs with a variance of posterior means less than 0.05. This resulted in 2108 SNPs. We then estimated the squared correlation between each SNP using either the naive approach of calculating the sample Pearson correlation between posterior means, or using our new moment-based approach (6).

Our estimators are scalable. On a 1.9 GHz quad-core PC running Linux with 32 GB of memory, it took a total of 1.9 seconds to estimate *all* pairwise correlations using our new moment-based approach, which is a small increase over the 0.7 seconds it took to estimate all pairwise correlations using the naive approach. In Gerard [2020], we found that the MLE approach took about 0.1 seconds for *each* pair of SNPs for a tetraploid individual. Extrapolating this to 2108 SNPs would indicate that the MLE approach would take about 2.5 days of computation time to calculate all pairwise LD estimates on this dataset ($\binom{2108}{2} \times 0.1\text{sec} \times 1\text{min}/60\text{sec} \times 1\text{hr}/60\text{min} \times 1\text{d}/24\text{hr} = 2.57\text{d}$).

The histogram of estimated reliability ratios are presented in Figure 3. We see there that the reliability ratios of most SNPs only increase their correlation estimates by less than 10%. But a not insignificant portion have reliability ratios that increase the correlation estimates by more than 10%. To evaluate the LD estimates of high reliability ratio SNPs, we calculated the MLEs for ρ^2 between the twenty SNPs with the largest reliability ratios. A pairs plot for ρ^2 estimates between the three approaches is presented in Figure 4. We see there that the MLE and new moment-based approach result in very similar ρ^2 estimates, while the naive approach using posterior means results in much smaller ρ^2 estimates.

4 Discussion

It has been known since at least the time of Spearman that the sample correlation coefficient (or, similarly, the ordinary least squares estimator in simple linear regression) is attenuated in the presence of uncertain variables [Spearman, 1904]. Methods to adjust for this bias include assuming prior knowledge on the measurement variances or the ratio of measurement variances (resulting

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from, for example, repeated measurements on the same individuals) [Koopmans, 1937, Degracie and Fuller, 1972], using instrumental variables [Carter and Fuller, 1980], and using distributional assumptions [Pal, 1980]. See Fuller [2009] for a detailed introduction to this vast field. Our solution was to use sample moments of marginal posterior moments which, to our knowledge, has never been proposed before.

It is natural to ask if our methods could be used to account for uncertain genotypes in genome-wide association studies. However, the moment-based techniques we used in this manuscript, when applied to simple linear regression with an additive effects model (where the SNP effect is proportional to the dosage), result in the standard ordinary least squares estimates when using the posterior mean as a covariate (Section S4). This supports using the posterior mean as a covariate in simple linear regression with an additive effects model. This is not to say, however, that using the posterior mean is also appropriate for more complicated models of gene action [Rosyara et al., 2016], or for non-linear models.

Acknowledgments

Most analyses were performed using the R statistical language [R Core Team, 2020].

Data availability

All methods discussed in this manuscript are implemented in the `ldsep` R package, available on GitHub (`https://github.com/dcgerard/ldsep`) under a GPL-3 license. Scripts to reproduce the results of this research are available on GitHub (`https://github.com/dcgerard/ldfast_sims`). All datasets used in this manuscript are publicly available [Uitdewilligen et al., 2013] and may be downloaded from:

- `https://doi.org/10.1371/journal.pone.0062355.s004`
- `https://doi.org/10.1371/journal.pone.0062355.s007`
- `https://doi.org/10.1371/journal.pone.0062355.s009`
- `https://doi.org/10.1371/journal.pone.0062355.s010`

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5 Figures

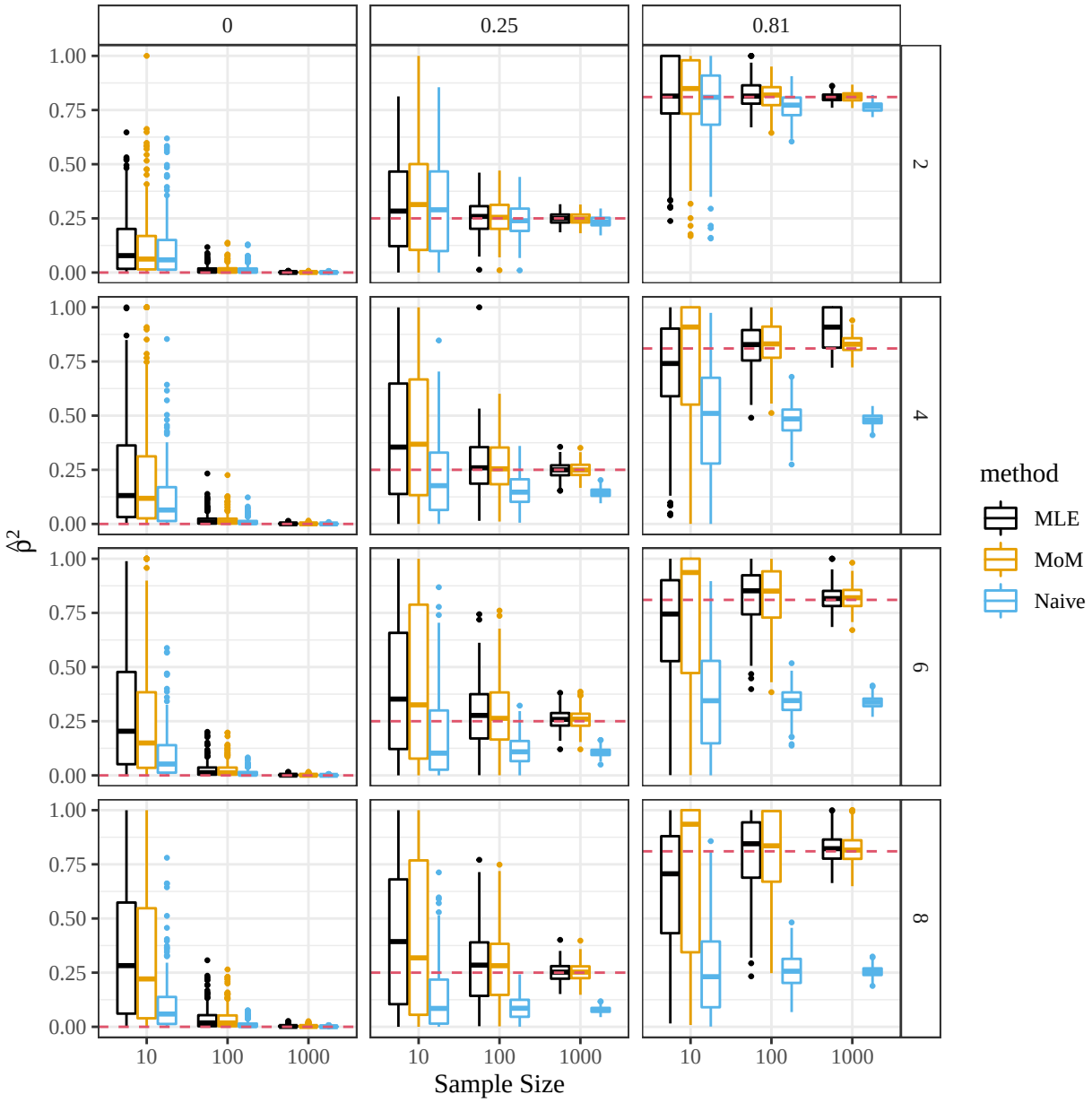

Figure 1: Estimate of ρ^2 (y-axis) for the maximum likelihood estimator [Gerard, 2020] (MLE), our new moment-based estimator (6) (MoM), and the naive squared sample correlation coefficient between posterior mean genotypes (Naive). The x-axis indexes the sample size, the row-facets index the ploidy, and the column-facets index the true ρ^2, which is also presented by the horizontal dashed red line. These simulations were performed using a read-depth of 10, and major allele frequencies of 0.5 at each locus. The naive estimator presents a strong attenuation bias toward 0, particularly for higher ploidy regimes.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

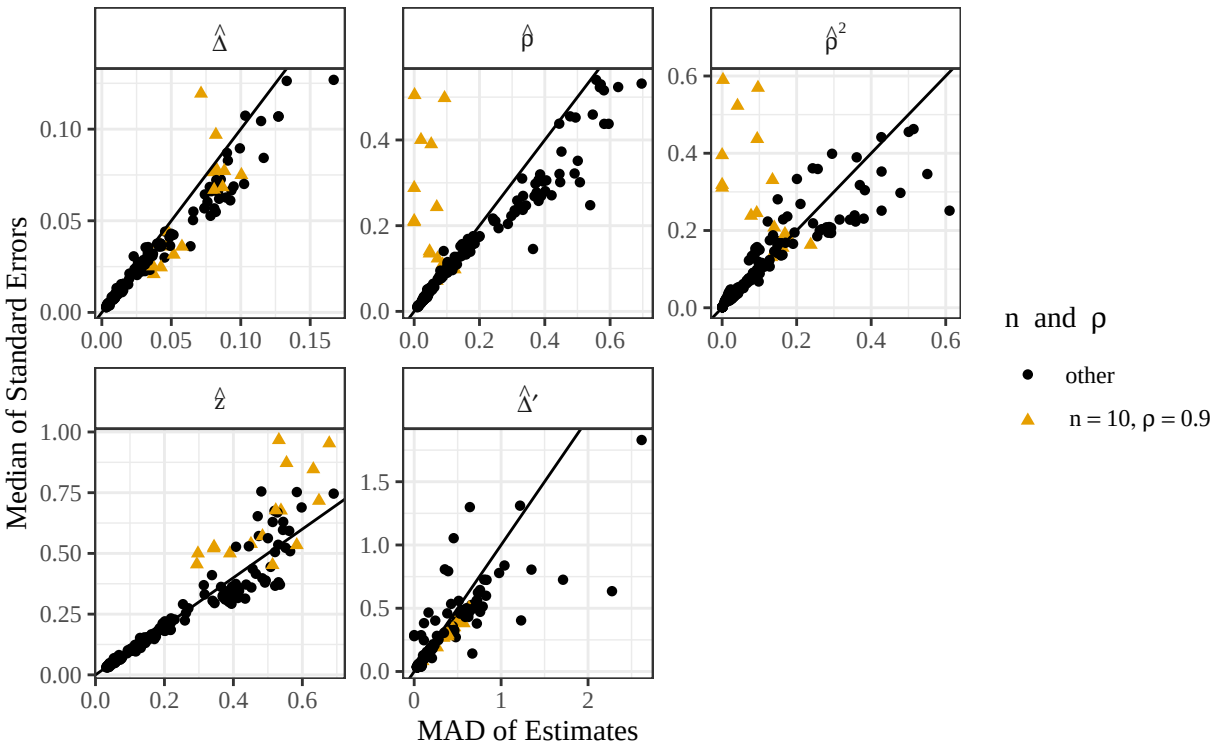

Figure 2: Median of estimated standard errors (y-axis) versus median absolute deviations (x-axis) of each of the moment-based LD estimators (facets). The line is the $y = x$ line, and points above this line indicate that the estimated standard errors are typically larger than the true standard errors. Estimated standard error are reasonably unbiased except for $\hat{\rho}$ and $\hat{\rho}^2$ in scenarios with small sample sizes ($n = 10$) and a large levels of LD ($\rho = 0.9$) (color and shape).

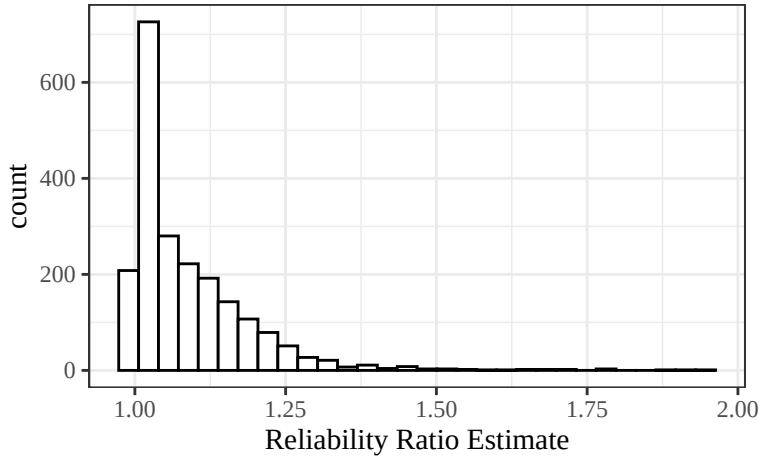

Figure 3: Histogram of estimated reliability ratios (S69) using the data from Uitdewilligen et al. [2013].

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

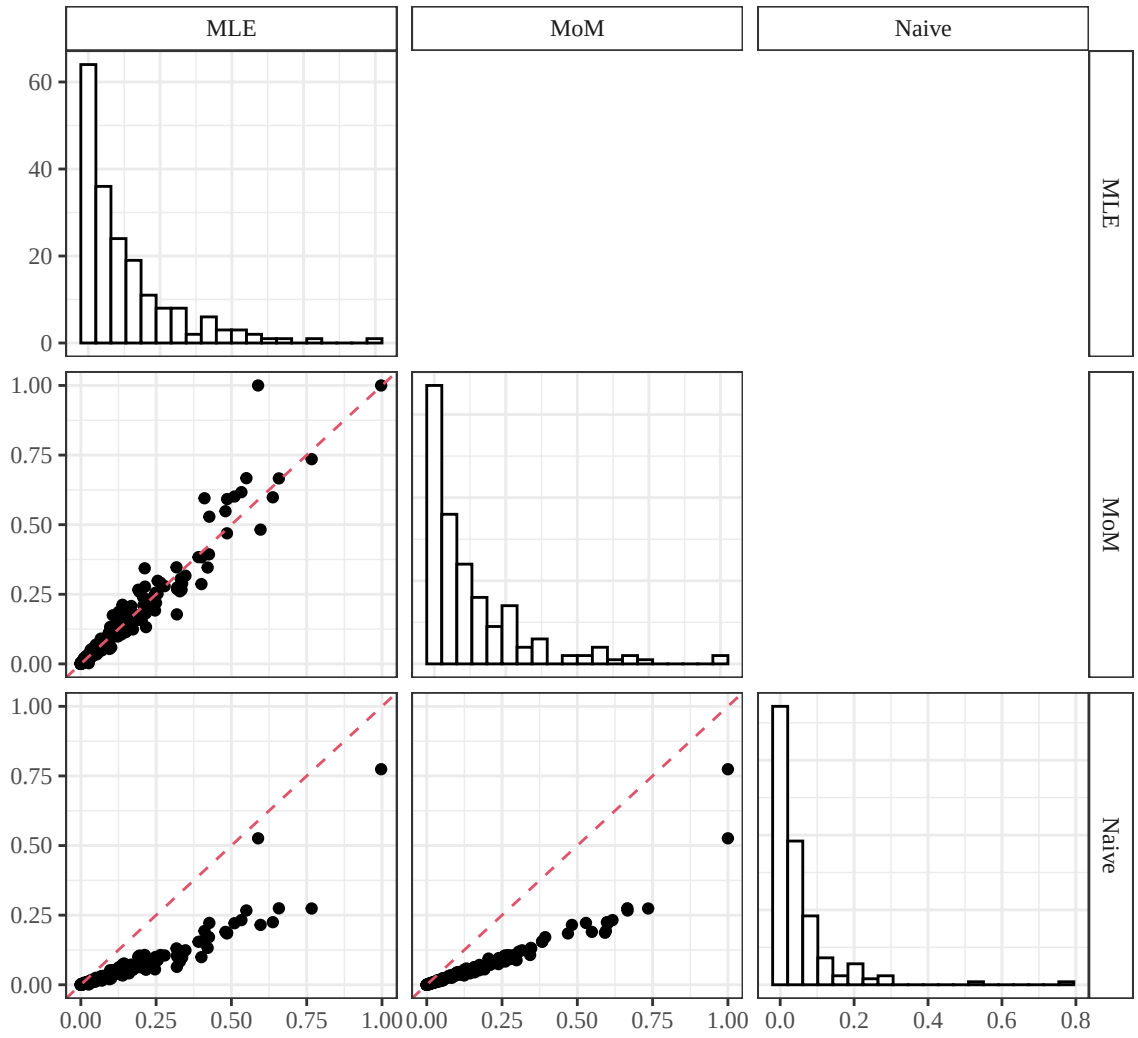

Figure 4: Pairs plot for ρ^2 estimates between the twenty SNPs from Uitdewilligen et al. [2013] with the largest estimated reliability ratios when using either maximum likelihood estimation (MLE) [Gerard, 2020], our new moment-based approach (6) (MoM), or the naive approach using just posterior means (Naive). The dashed line is the $y = x$ line. The MLE and the moment-based approach result in much more similar LD estimates.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Material

S1 Derivation of LD estimators

In this section, we derive estimators (5)–(7). We do this by assuming a normal model on the data and the genotypes. This is obviously not appropriate when using genotypes and sequencing data, but our simulations in Section 3.1 were also accomplished using sequencing data and resulted in very good performance.

Let $\boldsymbol{G}_i = (G_{iA}, G_{iB})^\intercal$ be the genotype for individual i at loci A and B. Let $\boldsymbol{Z}_i = (Z_{iA}, Z_{iB})^\intercal$ be the data for individual i at loci A and B. Then we let

$$\boldsymbol{G}_i \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ and} \tag{S1}$$

$$\boldsymbol{Z}_i | \boldsymbol{G}_i \sim N_2(\boldsymbol{G}_i, \boldsymbol{S}), \text{ where} \tag{S2}$$

$$\boldsymbol{\mu} = (\mu_1, \mu_2)^\intercal, \tag{S3}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}, \text{ and} \tag{S4}$$

$$\boldsymbol{S} = \begin{pmatrix} s_{11} & 0 \\ 0 & s_{22} \end{pmatrix}. \tag{S5}$$

To interpret these terms, μ_1/K and μ_2/K are the allele frequencies at each locus, σ_{11} and σ_{22} are the variances of the genotypes at each locus, s_{11} and s_{22} are the variances of the genotyping errors at each locus, and σ_{12} is covariance between genotypes. By elementary methods, we have the well-known result that, marginally,

$$\boldsymbol{Z}_i \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{S}). \tag{S6}$$

We assume the user has provided posterior moments on the genotypes

$$X_{iA} = E[G_{iA}|Z_{iA}], X_{iB} = E[G_{iB}|Z_{iB}], Y_{iA} = \text{var}(G_{iA}|Z_{iA}), \text{ and } Y_{iB} = \text{var}(G_{iB}|Z_{iB}). \tag{S7}$$

These posterior moments are marginal in that they only condition on either Z_{iA} or Z_{iB}, but not both. Thus, we assume they are well-approximated by the model

$$G_{iA} \sim N(\mu_1, \sigma_{11}) \tag{S8}$$

$$Z_{iA}|G_{iA} \sim N(G_{iA}, s_{11}) \tag{S9}$$

$$G_{iB} \sim N(\mu_2, \sigma_{22}) \tag{S10}$$

$$Z_{iB}|G_{iB} \sim N(G_{iB}, s_{22}). \tag{S11}$$

By standard methods, this results in

$$G_{iA}|Z_{iA} \sim N\left[\left(\frac{1}{\sigma_{11}} + \frac{1}{s_{11}}\right)^{-1}\left(\frac{1}{\sigma_{11}}\mu_1 + \frac{1}{s_{11}}Z_{iA}\right), \left(\frac{1}{\sigma_{11}} + \frac{1}{s_{11}}\right)^{-1}\right], \text{ and} \tag{S12}$$

$$G_{iB}|Z_{iB} \sim N\left[\left(\frac{1}{\sigma_{22}} + \frac{1}{s_{22}}\right)^{-1}\left(\frac{1}{\sigma_{22}}\mu_2 + \frac{1}{s_{22}}Z_{iB}\right), \left(\frac{1}{\sigma_{22}} + \frac{1}{s_{22}}\right)^{-1}\right]. \tag{S13}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Treating only $\boldsymbol{Z}_i$ as random from distribution (S6), we have

$$u_{xA} \approx E\left[\left(\frac{1}{\sigma_{11}}+\frac{1}{s_{11}}\right)^{-1}\left(\frac{1}{\sigma_{11}}\mu_1+\frac{1}{s_{11}}Z_{iA}\right)\right] \tag{S14}$$

$$= \left(\frac{1}{\sigma_{11}}+\frac{1}{s_{11}}\right)^{-1}\left(\frac{1}{\sigma_{11}}\mu_1+\frac{1}{s_{11}}E[Z_{iA}]\right) \tag{S15}$$

$$= \left(\frac{1}{\sigma_{11}}+\frac{1}{s_{11}}\right)^{-1}\left(\frac{1}{\sigma_{11}}\mu_1+\frac{1}{s_{11}}\mu_1\right) \tag{S16}$$

$$= \mu_1. \tag{S17}$$

Similarly,

$$u_{xB} \approx \mu_2. \tag{S18}$$

Furthermore,

$$v_{xA} \approx \mathrm{var}\left[\left(\frac{1}{\sigma_{11}}+\frac{1}{s_{11}}\right)^{-1}\left(\frac{1}{\sigma_{11}}\mu_1+\frac{1}{s_{11}}Z_{iA}\right)\right] \tag{S19}$$

$$= \left(\frac{1}{\sigma_{11}}+\frac{1}{s_{11}}\right)^{-2}\frac{1}{s_{11}^2}\,\mathrm{var}(Z_{iA}) \tag{S20}$$

$$= \left(\frac{1}{\sigma_{11}}+\frac{1}{s_{11}}\right)^{-2}\frac{\sigma_{11}+s_{11}}{s_{11}^2} \tag{S21}$$

$$= \left(\frac{1}{\sigma_{11}}+\frac{1}{s_{11}}\right)^{-1}\frac{\sigma_{11}}{s_{11}}. \tag{S22}$$

Similarly,

$$v_{xB} \approx \left(\frac{1}{\sigma_{22}}+\frac{1}{s_{22}}\right)^{-1}\frac{\sigma_{22}}{s_{22}}. \tag{S23}$$

Now, using the posterior variances, we have

$$u_{yA} \approx \left(\frac{1}{\sigma_{11}}+\frac{1}{s_{11}}\right)^{-1}, \text{ and} \tag{S24}$$

$$u_{yB} \approx \left(\frac{1}{\sigma_{22}}+\frac{1}{s_{22}}\right)^{-1}. \tag{S25}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, the expectation of the sample covariance of posterior means is

$$c_x \approx \operatorname{cov}\left[\left(\frac{1}{\sigma_{11}}+\frac{1}{s_{11}}\right)^{-1}\left(\frac{1}{\sigma_{11}}\mu_1+\frac{1}{s_{11}}Z_{iA}\right), \left(\frac{1}{\sigma_{22}}+\frac{1}{s_{22}}\right)^{-1}\left(\frac{1}{\sigma_{22}}\mu_2+\frac{1}{s_{22}}Z_{iB}\right)\right] \tag{S26}$$

$$= \left(\frac{1}{\sigma_{11}}+\frac{1}{s_{11}}\right)^{-1}\left(\frac{1}{\sigma_{22}}+\frac{1}{s_{22}}\right)^{-1}\frac{1}{s_{11}}\frac{1}{s_{22}}\operatorname{cov}(Z_{iA}, Z_{iB}) \tag{S27}$$

$$= \left(\frac{1}{\sigma_{11}}+\frac{1}{s_{11}}\right)^{-1}\left(\frac{1}{\sigma_{22}}+\frac{1}{s_{22}}\right)^{-1}\frac{1}{s_{11}}\frac{1}{s_{22}}\sigma_{12}. \tag{S28}$$

Using a method-of-moments approach, we now have a system of five equations and five unknowns:

$$v_{xA} = \left(\frac{1}{\sigma_{11}}+\frac{1}{s_{11}}\right)^{-1}\frac{\sigma_{11}}{s_{11}}, \tag{S29}$$

$$v_{xB} = \left(\frac{1}{\sigma_{22}}+\frac{1}{s_{22}}\right)^{-1}\frac{\sigma_{22}}{s_{22}}, \tag{S30}$$

$$u_{yA} = \left(\frac{1}{\sigma_{11}}+\frac{1}{s_{11}}\right)^{-1}, \tag{S31}$$

$$u_{yB} = \left(\frac{1}{\sigma_{22}}+\frac{1}{s_{22}}\right)^{-1}, \text{ and} \tag{S32}$$

$$c_x = \left(\frac{1}{\sigma_{11}}+\frac{1}{s_{11}}\right)^{-1}\left(\frac{1}{\sigma_{22}}+\frac{1}{s_{22}}\right)^{-1}\frac{1}{s_{11}}\frac{1}{s_{22}}\sigma_{12}. \tag{S33}$$

Solving for s_{11}, s_{22}, σ_{11}, σ_{22}, and σ_{12}, we obtain:

$$\hat{s}_{11} = \frac{u_{yA}(u_{yA}+v_{xA})}{v_{xA}} \tag{S34}$$

$$\hat{s}_{22} = \frac{u_{yB}(u_{yB}+v_{xB})}{v_{xB}} \tag{S35}$$

$$\hat{\sigma}_{11} = u_{yA}+v_{xA} \tag{S36}$$

$$\hat{\sigma}_{22} = u_{yB}+v_{xB} \tag{S37}$$

$$\hat{\sigma}_{12} = \frac{u_{yA}+v_{xA}}{v_{xA}}\frac{u_{yB}+v_{xB}}{v_{xB}}c_x. \tag{S38}$$

Using (S14)–(S18), we also have

$$\hat{\mu}_1 = u_{xA}, \text{ and} \tag{S39}$$

$$\hat{\mu}_2 = u_{xB}. \tag{S40}$$

The LD coefficient estimates (5)–(7) can be obtained by substituting in parameter estimates in

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the following equations [Gerard, 2020]

$$\Delta = \sigma_{12}/K, \tag{S41}$$

$$\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}, \text{ and} \tag{S42}$$

$$\Delta' = \Delta/\Delta_m, \text{ where} \tag{S43}$$

$$\Delta_m = \begin{cases} \min\{\mu_1\mu_2, (K-\mu_1)(K-\mu_2)\}/K^2 & \text{if } \Delta < 0, \text{ and} \\ \min\{\mu_1(K-\mu_2), (K-\mu_1)\mu_2\}/K^2 & \text{if } \Delta > 0. \end{cases} \tag{S44}$$

S2 Derivation of standard errors

Let

$$\boldsymbol{M}_i := (X_{iA}, X_{iA}^2, X_{iB}, X_{iB}^2, X_{iA}X_{iB}, Y_{iA}, Y_{iB})^\intercal. \tag{S45}$$

Then, by the central limit theorem, we have for

$$\bar{\boldsymbol{M}} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{M}_i, \tag{S46}$$

that $\sqrt{n}\bar{\boldsymbol{M}}$ is asymptotically multivariate normal with some limiting covariance, say, $\boldsymbol{\Omega}$. Finite variances are guaranteed by the finite support of the genotypes. We can estimate $\boldsymbol{\Omega}$ with the sample covariance matrix

$$\hat{\Omega} := \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{M}_i - \bar{\boldsymbol{M}})(\boldsymbol{M}_i - \bar{\boldsymbol{M}})^\intercal. \tag{S47}$$

Estimators (5)–(7) are approximately functions of $\bar{\boldsymbol{M}}$. Namely

$$\hat{\Delta} \approx \left(\frac{\bar{M}_6 + \bar{M}_2 - \bar{M}_1^2}{\bar{M}_2 - \bar{M}_1^2}\right)\left(\frac{\bar{M}_7 + \bar{M}_4 - \bar{M}_3^2}{\bar{M}_4 - \bar{M}_3^2}\right)\left(\frac{\bar{M}_5 - \bar{M}_1\bar{M}_3}{K}\right) \tag{S48}$$

$$\hat{\rho} \approx \left(\frac{\sqrt{\bar{M}_6 + \bar{M}_2 - \bar{M}_1^2}}{\bar{M}_2 - \bar{M}_1^2}\right)\left(\frac{\sqrt{\bar{M}_7 + \bar{M}_4 - \bar{M}_3^2}}{\bar{M}_4 - \bar{M}_3^2}\right)\left(\bar{M}_5 - \bar{M}_1\bar{M}_3\right) \tag{S49}$$

$$\hat{\Delta}' \approx \left(\frac{\bar{M}_6 + \bar{M}_2 - \bar{M}_1^2}{\bar{M}_2 - \bar{M}_1^2}\right)\left(\frac{\bar{M}_7 + \bar{M}_4 - \bar{M}_3^2}{\bar{M}_4 - \bar{M}_3^2}\right)\left(\frac{\bar{M}_5 - \bar{M}_1\bar{M}_3}{K}\right)/\hat{\Delta}_m, \text{ where} \tag{S50}$$

$$\hat{\Delta}_m = \begin{cases} \min\{\bar{M}_1\bar{M}_3, (K-\bar{M}_1)(K-\bar{M}_3)\}/K^2 & \text{if } \bar{M}_5 - \bar{M}_1\bar{M}_3 < 0, \text{ and} \\ \min\{\bar{M}_1(K-\bar{M}_3), (K-\bar{M}_1)\bar{M}_3\}/K^2 & \text{if } \bar{M}_5 - \bar{M}_1\bar{M}_3 > 0. \end{cases} \tag{S51}$$

These are smooth functions of $\bar{M}$ (except on a space of Lebesgue measure zero), and so admit the

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

following gradients, calculated in Mathematica [Wolfram Research, Inc., 2020]:

$$\boldsymbol{g}_\Delta := \frac{d\hat{\Delta}}{d\bar{\boldsymbol{M}}} = \begin{pmatrix} -\frac{\left(\bar{M}_1^4\bar{M}_3-2\bar{M}_1\bar{M}_5\bar{M}_6+\bar{M}_1^2\bar{M}_3(-2\bar{M}_2+\bar{M}_6)+\bar{M}_2\bar{M}_3(\bar{M}_2+\bar{M}_6)\right)\left(\bar{M}_3^2-\bar{M}_4-\bar{M}_7\right)}{K\left(\bar{M}_1^2-\bar{M}_2\right)^2\left(\bar{M}_3^2-\bar{M}_4\right)} \\ \frac{(\bar{M}_1\bar{M}_3-\bar{M}_5)\bar{M}_6\left(\bar{M}_3^2-\bar{M}_4-\bar{M}_7\right)}{K\left(\bar{M}_1^2-\bar{M}_2\right)^2\left(\bar{M}_3^2-\bar{M}_4\right)} \\ -\frac{\left(\bar{M}_1^2-\bar{M}_2-\bar{M}_6\right)\left(-2\bar{M}_3\bar{M}_5\bar{M}_7+\bar{M}_1\left(\bar{M}_3^4+\bar{M}_3^2(-2\bar{M}_4+\bar{M}_7)+\bar{M}_4(\bar{M}_4+\bar{M}_7)\right)\right)}{K\left(\bar{M}_1^2-\bar{M}_2\right)\left(\bar{M}_3^2-\bar{M}_4\right)^2} \\ \frac{(\bar{M}_1\bar{M}_3-\bar{M}_5)\left(\bar{M}_1^2-\bar{M}_2-\bar{M}_6\right)\bar{M}_7}{K\left(\bar{M}_1^2-\bar{M}_2\right)\left(\bar{M}_3^2-\bar{M}_4\right)^2} \\ \frac{\left(-\bar{M}_1^2+\bar{M}_2+\bar{M}_6\right)\left(-\bar{M}_3^2+\bar{M}_4+\bar{M}_7\right)}{K\left(\bar{M}_1^2-\bar{M}_2\right)\left(\bar{M}_3^2-\bar{M}_4\right)} \\ \frac{(-\bar{M}_1\bar{M}_3+\bar{M}_5)\left(-\bar{M}_3^2+\bar{M}_4+\bar{M}_7\right)}{K\left(\bar{M}_1^2-\bar{M}_2\right)\left(\bar{M}_3^2-\bar{M}_4\right)} \\ \frac{(-\bar{M}_1\bar{M}_3+\bar{M}_5)\left(-\bar{M}_1^2+\bar{M}_2+\bar{M}_6\right)}{K\left(\bar{M}_1^2-\bar{M}_2\right)\left(\bar{M}_3^2-\bar{M}_4\right)} \end{pmatrix}, \tag{S52}$$

$$\boldsymbol{g}_\rho := \frac{d\hat{\rho}}{d\bar{\boldsymbol{M}}} = \begin{pmatrix} \frac{\left(\bar{M}_1^3\bar{M}_5+\bar{M}_1^2\bar{M}_3(-\bar{M}_2+\bar{M}_6)+\bar{M}_2\bar{M}_3(\bar{M}_2+\bar{M}_6)-\bar{M}_1\bar{M}_5(\bar{M}_2+2\bar{M}_6)\right)\sqrt{-\bar{M}_3^2+\bar{M}_4+\bar{M}_7}}{\left(\bar{M}_1^2-\bar{M}_2\right)^2\left(\bar{M}_3^2-\bar{M}_4\right)\sqrt{-\bar{M}_1^2+\bar{M}_2+\bar{M}_6}} \\ \frac{(\bar{M}_1\bar{M}_3-\bar{M}_5)\left(\bar{M}_1^2-\bar{M}_2-2\bar{M}_6\right)\sqrt{-\bar{M}_3^2+\bar{M}_4+\bar{M}_7}}{2\left(\bar{M}_1^2-\bar{M}_2\right)^2\left(\bar{M}_3^2-\bar{M}_4\right)\sqrt{-\bar{M}_1^2+\bar{M}_2+\bar{M}_6}} \\ -\frac{\sqrt{-\bar{M}_1^2+\bar{M}_2+\bar{M}_6}\left(\bar{M}_1\bar{M}_3^2(\bar{M}_4-\bar{M}_7)-\bar{M}_1\bar{M}_4(\bar{M}_4+\bar{M}_7)+\bar{M}_3\bar{M}_5\left(-\bar{M}_3^2+\bar{M}_4+2\bar{M}_7\right)\right)}{\left(\bar{M}_1^2-\bar{M}_2\right)\left(\bar{M}_3^2-\bar{M}_4\right)^2\sqrt{-\bar{M}_3^2+\bar{M}_4+\bar{M}_7}} \\ \frac{(\bar{M}_1\bar{M}_3-\bar{M}_5)\sqrt{-\bar{M}_1^2+\bar{M}_2+\bar{M}_6}\left(\bar{M}_3^2-\bar{M}_4-2\bar{M}_7\right)}{2\left(\bar{M}_1^2-\bar{M}_2\right)\left(\bar{M}_3^2-\bar{M}_4\right)^2\sqrt{-\bar{M}_3^2+\bar{M}_4+\bar{M}_7}} \\ \frac{\sqrt{-\bar{M}_1^2+\bar{M}_2+\bar{M}_6}\sqrt{-\bar{M}_3^2+\bar{M}_4+\bar{M}_7}}{\left(\bar{M}_1^2-\bar{M}_2\right)\left(\bar{M}_3^2-\bar{M}_4\right)} \\ \frac{(-\bar{M}_1\bar{M}_3+\bar{M}_5)\sqrt{-\bar{M}_3^2+\bar{M}_4+\bar{M}_7}}{2\left(\bar{M}_1^2-\bar{M}_2\right)\left(\bar{M}_3^2-\bar{M}_4\right)\sqrt{-\bar{M}_1^2+\bar{M}_2+\bar{M}_6}} \\ \frac{(-\bar{M}_1\bar{M}_3+\bar{M}_5)\sqrt{-\bar{M}_1^2+\bar{M}_2+\bar{M}_6}}{2\left(\bar{M}_1^2-\bar{M}_2\right)\left(\bar{M}_3^2-\bar{M}_4\right)\sqrt{-\bar{M}_3^2+\bar{M}_4+\bar{M}_7}} \end{pmatrix}, \tag{S53}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and

$$\boldsymbol{g}_{\Delta'} := \frac{d\hat{\Delta}}{d\bar{\boldsymbol{M}}} = \boldsymbol{g}_{\Delta}/\hat{\Delta}_m - \boldsymbol{A}, \text{ where} \tag{S54}$$

$$\boldsymbol{A} = \begin{pmatrix} \hat{\Delta} C_1(\bar{M}_1, \bar{M}_3, \bar{M}_5)/\hat{\Delta}_m^2 \\ 0 \\ \hat{\Delta} C_3(\bar{M}_1, \bar{M}_3, \bar{M}_5)/\hat{\Delta}_m^2 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{S55}$$

$$C_1(\bar{M}_1, \bar{M}_3, \bar{M}_5) = \begin{cases} \bar{M}_3/K^2 & \text{if } \bar{M}_5 < \bar{M}_1\bar{M}_3 \text{ and } \bar{M}_1\bar{M}_3 < (K-\bar{M}_1)(K-\bar{M}_3) \\ -(K-\bar{M}_3)/K^2 & \text{if } \bar{M}_5 < \bar{M}_1\bar{M}_3 \text{ and } \bar{M}_1\bar{M}_3 > (K-\bar{M}_1)(K-\bar{M}_3) \\ -\bar{M}_3/K^2 & \text{if } \bar{M}_5 > \bar{M}_1\bar{M}_3 \text{ and } \bar{M}_1(K-\bar{M}_3) > (K-\bar{M}_1)\bar{M}_3 \\ (K-\bar{M}_3)/K^2 & \text{if } \bar{M}_5 > \bar{M}_1\bar{M}_3 \text{ and } \bar{M}_1(K-\bar{M}_3) < (K-\bar{M}_1)\bar{M}_3 \end{cases} \tag{S56}$$

$$C_3(\bar{M}_1, \bar{M}_3, \bar{M}_5) = \begin{cases} \bar{M}_1/K^2 & \text{if } \bar{M}_5 < \bar{M}_1\bar{M}_3 \text{ and } \bar{M}_1\bar{M}_3 < (K-\bar{M}_1)(K-\bar{M}_3) \\ -(K-\bar{M}_1)/K^2 & \text{if } \bar{M}_5 < \bar{M}_1\bar{M}_3 \text{ and } \bar{M}_1\bar{M}_3 > (K-\bar{M}_1)(K-\bar{M}_3) \\ (K-\bar{M}_1)/K^2 & \text{if } \bar{M}_5 > \bar{M}_1\bar{M}_3 \text{ and } \bar{M}_1(K-\bar{M}_3) > (K-\bar{M}_1)\bar{M}_3 \\ -\bar{M}_1/K^2 & \text{if } \bar{M}_5 > \bar{M}_1\bar{M}_3 \text{ and } \bar{M}_1(K-\bar{M}_3) < (K-\bar{M}_1)\bar{M}_3 \end{cases} \tag{S57}$$

Though these gradients are rather complicated, they are not computationally intensive and may be calculated in constant time in the sample size.

The asymptotic variances of $\hat{\Delta}$, $\hat{\rho}$, and $\hat{\Delta}'$ are

$$\frac{1}{n}\boldsymbol{g}_{\Delta}^{\intercal}\hat{\boldsymbol{\Omega}}\boldsymbol{g}_{\Delta}, \ \frac{1}{n}\boldsymbol{g}_{\rho}^{\intercal}\hat{\boldsymbol{\Omega}}\boldsymbol{g}_{\rho}, \text{ and } \frac{1}{n}\boldsymbol{g}_{\Delta'}^{\intercal}\hat{\boldsymbol{\Omega}}\boldsymbol{g}_{\Delta'}, \tag{S58}$$

respectively.

To accommodate missing data, we use only pairwise complete observations for the sample covariance matrix (S47). This ensures that $\hat{\boldsymbol{\Omega}}$ is positive definite and, thus, the resulting standard errors are non-negative. However, we use all non-missing observations for $\bar{\boldsymbol{M}}$. That is, let

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$\Theta_A, \Theta_B \subseteq \{1, 2, \ldots, n\}$ be the index sets of non-missing values at loci A and B, respectively. Then

$$\bar{M}_1 = \frac{1}{|\Theta_A|} \sum_{i \in \Theta_A} X_{iA} \tag{S59}$$

$$\bar{M}_2 = \frac{1}{|\Theta_A|} \sum_{i \in \Theta_A} X_{iA}^2 \tag{S60}$$

$$\bar{M}_3 = \frac{1}{|\Theta_B|} \sum_{i \in \Theta_B} X_{iB} \tag{S61}$$

$$\bar{M}_4 = \frac{1}{|\Theta_B|} \sum_{i \in \Theta_B} X_{iB}^2 \tag{S62}$$

$$\bar{M}_5 = \frac{1}{|\Theta_A \cap \Theta_B|} \sum_{i \in \Theta_A \cap \Theta_B} X_{iA} X_{iB} \tag{S63}$$

$$\bar{M}_6 = \frac{1}{|\Theta_A|} \sum_{i \in \Theta_A} Y_{iA} \tag{S64}$$

$$\bar{M}_7 = \frac{1}{|\Theta_B|} \sum_{i \in \Theta_B} Y_{iB} \tag{S65}$$

$$\bar{\boldsymbol{M}}^* = \frac{1}{|\Theta_A \cap \Theta_B|} \sum_{i \in \Theta_A \cap \Theta_B} \boldsymbol{M}_i \tag{S66}$$

$$\hat{\boldsymbol{\Omega}} = \frac{1}{|\Theta_A \cap \Theta_B| - 1} \sum_{i \in \Theta_A \cap \Theta_B} (\boldsymbol{M}_i - \bar{\boldsymbol{M}}^*)(\boldsymbol{M}_i - \bar{\boldsymbol{M}}^*)^\intercal \tag{S67}$$

The asymptotic variances of $\hat{\Delta}$, $\hat{\rho}$, and $\hat{\Delta}'$ are then

$$\frac{1}{|\Theta_A \cap \Theta_B|} \boldsymbol{g}_\Delta^\intercal \hat{\boldsymbol{\Omega}} \boldsymbol{g}_\Delta, \ \frac{1}{|\Theta_A \cap \Theta_B|} \boldsymbol{g}_\rho^\intercal \hat{\boldsymbol{\Omega}} \boldsymbol{g}_\rho, \text{ and } \frac{1}{|\Theta_A \cap \Theta_B|} \boldsymbol{g}_{\Delta'}^\intercal \hat{\boldsymbol{\Omega}} \boldsymbol{g}_{\Delta'}, \tag{S68}$$

respectively.

S3 Adjusting the reliability ratios

S3.1 Adaptive shrinkage on the reliability ratios

Each SNP has an estimated reliability ratio,

$$b_j := \frac{u_{yj} + v_{xj}}{v_{xj}}, \tag{S69}$$

which corresponds to the multiplicative adjustment to all LD estimates that include that SNP (see (5)). These reliability ratios might have high variance due to (i) lower sequencing depth or (ii) containing fewer individuals with non-missing data. Thus, some reliability ratios may be noisy. Hierarchical shrinkage is a statistical technique that allows high-variance observations to borrow strength from low-variance observations and thus improve estimation performance. Adaptive shrinkage (*ash*) [Stephens, 2016] is a recently proposed general-purpose hierarchical shrinkage

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

technique that we can use to model the distribution of reliability ratios flexibly, only constraining them to be unimodal. In this section, we will use *ash* to improve our reliability ratio estimates.

We will now describe the procedure for applying *ash* to shrink the reliability ratios. Our strategy will be to derive the standard errors for the log of the reliability ratios (S69) and apply *ash* on the log-scale using these standard errors. To begin, let X_{ij} be the posterior mean for individual i at SNP j. Let Y_{ij} be the posterior variance for individual i at SNP j. Finally, let

$$\boldsymbol{M}_{ij} = (X_{ij}, X_{ij}^2, Y_{ij}), \tag{S70}$$

$$\bar{\boldsymbol{M}}_j = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{M}_{ij}, \text{ so} \tag{S71}$$

$$\bar{M}_{j1} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, \tag{S72}$$

$$\bar{M}_{j2} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}^2, \text{ and} \tag{S73}$$

$$\bar{M}_{j3} = \frac{1}{n}\sum_{i=1}^{n} Y_{ij}. \tag{S74}$$

Then the log of the reliability ratio for SNP j is

$$L_j := \log\left(\frac{\bar{M}_{j3} + \bar{M}_{j2} - \bar{M}_{j1}^2}{\bar{M}_{j2} - \bar{M}_{j1}^2}\right) \tag{S75}$$

$$= \log(\bar{M}_{j3} + \bar{M}_{j2} - \bar{M}_{j1}^2) - \log(\bar{M}_{j2} - \bar{M}_{j1}^2). \tag{S76}$$

Let the sample covariance be

$$\hat{\boldsymbol{\Omega}}_j := \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{M}_{ij} - \bar{\boldsymbol{M}}_j)(\boldsymbol{M}_{ij} - \bar{\boldsymbol{M}}_j)^\intercal. \tag{S77}$$

Then we have by the central limit theorem that $\sqrt{n}\bar{\boldsymbol{M}}_j$ is asymptotically multivariate normal, and we can use $\hat{\boldsymbol{\Omega}}_j$ as the estimate of the covariance matrix. The gradients for (S75) are

$$g_{j1} := \frac{dL_j}{d\bar{M}_{j1}} = \frac{-2\bar{M}_{j1}}{\bar{M}_{j3} + \bar{M}_{j2} - \bar{M}_{j1}^2} + \frac{2\bar{M}_{j1}}{\bar{M}_{j2} - \bar{M}_{j1}^2} \tag{S78}$$

$$g_{j2} := \frac{dL_j}{d\bar{M}_{j2}} = \frac{1}{\bar{M}_{j3} + \bar{M}_{j2} - \bar{M}_{j1}^2} - \frac{1}{\bar{M}_{j2} - \bar{M}_{j1}^2} \tag{S79}$$

$$g_{j3} := \frac{dL_j}{d\bar{M}_{j3}} = \frac{1}{\bar{M}_{j3} + \bar{M}_{j2} - \bar{M}_{j1}^2} \tag{S80}$$

Then, with $\boldsymbol{g}_j := (g_{j1}, g_{j2}, g_{j3})^\intercal$, the variance for L_j is

$$\hat{s}_j^2 := \frac{1}{n}\boldsymbol{g}_j^\intercal \hat{\boldsymbol{\Omega}}_j \boldsymbol{g}_j. \tag{S81}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We apply *ash* to $(L_1, \hat{s}_1), \ldots, (L_m, \hat{s}_m)$ to obtain shrunken log reliability ratios $\hat{L}_1, \ldots, \hat{L}_m$. Because *ash*'s grid-based scheme for estimating the mode is not the most computationally efficient, we used the half-sample mode estimator of Robertson and Cryer [1974] prior to running *ash*.

This procedure seems to result in improved performance for SNPs with unusually variable reliability ratios (Figure S1).

S3.2 Thresholding the reliability ratios

If a researcher accidentally provides a monoallelic SNP, its reliability ratio could explode due to having a denominator close to zero in (S69). For example, the right panel of Figure S2 contains a monoallelic SNP (PotVar0080327) whose reliability ratio estimate (S69) is 100.92. This can provide unstable estimates of LD as some SNPs will, due to sampling variability, have correlations with these monoallelic SNPs on the order of 0.01. For example, the sample correlation between posterior means of PotVar0080327 and PotVar0078678 (left facet of Figure S2) -0.0098. But due to the extreme reliability ratio of PotVar0080327, the genotype-error adjusted correlation estimate is -1. This is, of course, unsettling. So by default, our software will take all reliability ratio estimates (S69) above a user-provided value (default of 10) and assign these to have reliability ratios of the median reliability ratio in the dataset.

S4 Genome-wide Association Studies

In this section, we demonstrate that the techniques used in Section S1, when applied to simple linear regression with an additive effects model [Rosyara et al., 2016], result in the standard ordinary least squares estimate when using the posterior mean as a covariate. This indicates that for genome-wide association studies, using the posterior mean is appropriate in a linear regression context when using an additive model for gene action.

Let G_i be the genotype for individual i at a locus. Let Z_i be the data that lead to the genotyping for individual i at the same locus. Let W_i be some quantitative trait of interest for individual i. Then we let

$$W_i|G_i \sim N(\beta_0 + \beta_1 G_i, \sigma^2) \tag{S82}$$

$$Z_i|G_i \sim N(G_i, s^2) \tag{S83}$$

$$G_i \sim N(\mu, \tau^2). \tag{S84}$$

We suppose the user is only provided the posterior means and variances of each $G_i|Z_i$. Let $X_i = E[G_i|Z_i]$ and $Y_i = \text{var}(G_i|Z_i)$. From elementary methods, we have

$$Z_i \sim N(\mu, s^2 + \tau^2) \tag{S85}$$

$$G_i|Z_i \sim N\left[\left(\frac{1}{\tau^2} + \frac{1}{s^2}\right)^{-1}\left(\frac{1}{\tau^2}\mu + \frac{1}{s^2}Z_i\right), \left(\frac{1}{\tau^2} + \frac{1}{s^2}\right)^{-1}\right]. \tag{S86}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Let

$$u_w = \frac{1}{n}\sum_{i=1}^{n} W_i \tag{S87}$$

$$u_x = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{S88}$$

$$c_{xw} = \frac{1}{n-1}\sum_{i=1}^{n}(W_i - u_w)(X_i - u_x) \tag{S89}$$

$$v_x = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - u_x)^2 \tag{S90}$$

$$v_w = \frac{1}{n-1}\sum_{i=1}^{n}(W_i - u_w)^2. \tag{S91}$$

We have that

$$c_{wx} \approx \operatorname{cov}(W_i, X_i) \tag{S92}$$

$$\approx \operatorname{cov}\left(W_i, \left(\frac{1}{\tau^2} + \frac{1}{s^2}\right)^{-1}\left(\frac{1}{\tau^2}\mu + \frac{1}{s^2}Z_i\right)\right) \tag{S93}$$

$$= \left(\frac{1}{\tau^2} + \frac{1}{s^2}\right)^{-1}\frac{1}{s^2}\operatorname{cov}(W_i, Z_i) \tag{S94}$$

$$= \left(\frac{1}{\tau^2} + \frac{1}{s^2}\right)^{-1}\frac{1}{s^2}\beta_1 \operatorname{var}(G_i) \tag{S95}$$

$$= \left(\frac{1}{\tau^2} + \frac{1}{s^2}\right)^{-1}\frac{\tau^2}{s^2}\beta_1. \tag{S96}$$

We also have from (S19)–(S22) that

$$v_x \approx \left(\frac{1}{\tau^2} + \frac{1}{s^2}\right)^{-1}\frac{\tau^2}{s^2}. \tag{S97}$$

Using method of moments with equations (S96) and (S97), we have the following estimator for β_1

$$\hat{\beta}_1 = c_{wx}/v_x \tag{S98}$$

$$= \frac{c_{wx}}{\sqrt{v_x v_w}}\frac{\sqrt{v_w}}{\sqrt{v_x}}. \tag{S99}$$

Equation (S99) is the sample correlation between the W_i's and the X_i's ($c_{wx}/\sqrt{v_x v_w}$) multiplied by the ratio of the sample standard deviations of the W_i's and the X_i's ($\sqrt{v_w}/\sqrt{v_x}$). This is the well-known formula for the ordinary least squares estimate of β_1 from a regression of W_i on X_i.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S5 Supplementary figures

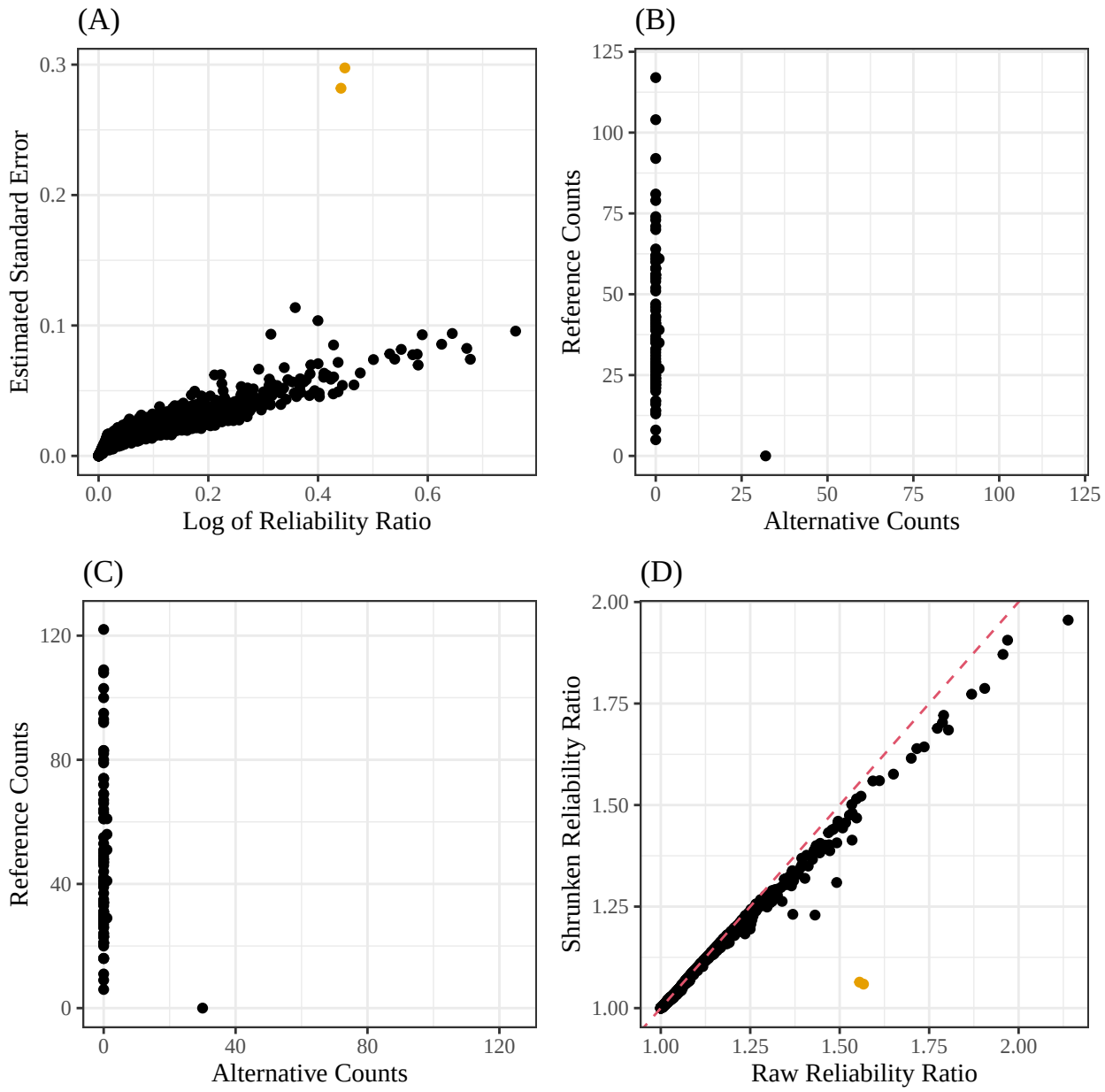

Figure S1: **(A)** The log of the reliability ratios (x-axis) versus their estimated standard errors (y-axis). The two highlighted points do not seem to fit the trend. When we plot the read-counts for these highlighted points (**(B)** and **(C)**), we notice that these two SNPs are almost monoallelic, providing doubts on their unusually large reliability ratios. We plot the shrunken reliability ratios (y-axis) against their original values (x-axis) in **(D)**, noting that the problem SNPs (color) have their reliability ratios highly adjusted.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

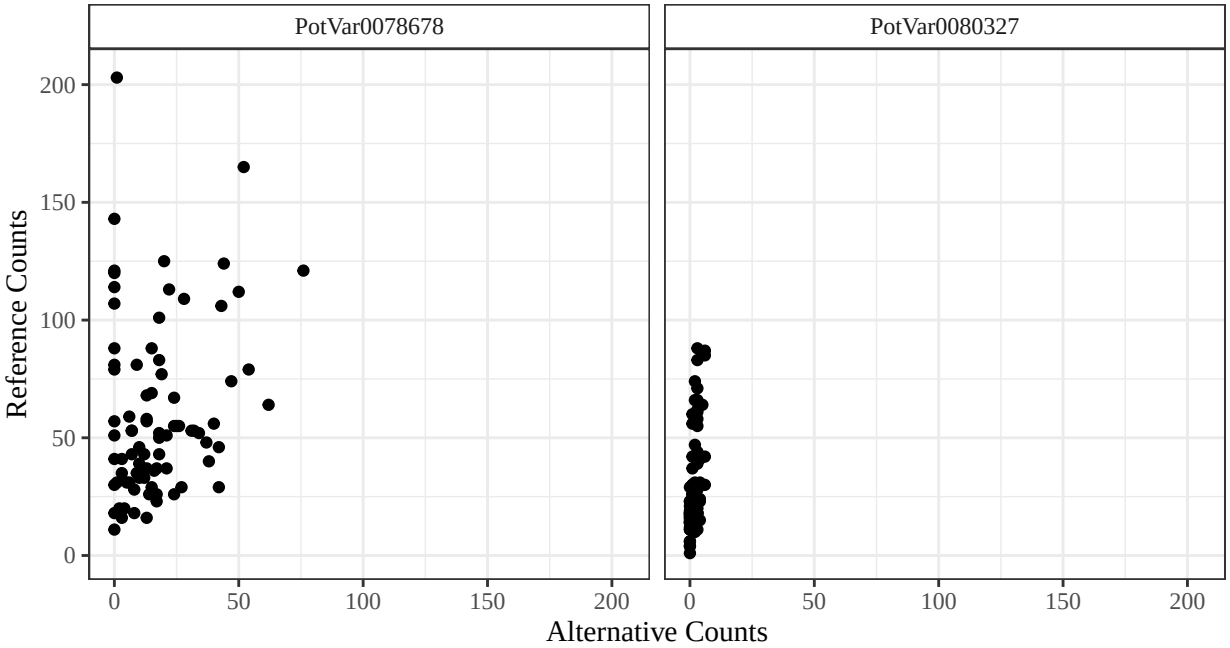

Figure S2: Plots of read-counts of two SNPs (facets) from Uitdewilligen et al. [2013]. Alternative counts lie on the x-axis and reference counts lie on the y-axis. The right SNP is monoallelic and because of this the estimated correlation between the two SNPs using raw reliability ratios is -1, even though the sample correlation between posterior means is only -0.0098.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

A. Brown. Sample sizes required to detect linkage disequilibrium between two or three loci. *Theoretical Population Biology*, 8(2):184 – 201, 1975. ISSN 0040-5809. doi: 10.1016/0040-5809(75)90031-3.

R. L. Carter and W. A. Fuller. Instrumental variable estimation of the simple errors-in-variables model. *Journal of the American Statistical Association*, 75(371):687–692, 1980. doi: 10.1080/01621459.1980.10477534.

J. S. Degracie and W. A. Fuller. Estimation of the slope and analysis of covariance when the concomitant variable is measured with error. *Journal of the American Statistical Association*, 67 (340):930–937, 1972. doi: 10.1080/01621459.1972.10481321.

K. K. Dey and M. Stephens. CorShrink: Empirical Bayes shrinkage estimation of correlations, with applications. *bioRxiv*, 2018. doi: 10.1101/368316.

E. A. Fox, A. E. Wright, M. Fumagalli, and F. G. Vieira. ngsLD: evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics*, 35(19):3855–3856, 03 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz200.

W. A. Fuller. *Measurement error models*. John Wiley & Sons, 2009.

D. Gerard. Pairwise linkage disequilibrium estimation for polyploids. *bioRxiv*, 2020. doi: 10.1101/2020.08.03.234476.

D. Gerard and L. F. V. Ferrão. Priors for genotyping polyploids. *Bioinformatics*, 36(6):1795–1800, 11 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz852. bioRxiv: 751784.

D. Gerard, L. F. V. Ferrão, A. A. F. Garcia, and M. Stephens. Genotyping polyploids from messy sequencing data. *Genetics*, 210(3):789–807, 2018. ISSN 0016-6731. doi: 10.1534/genetics.118.301468.

W. Hill and A. Robertson. Linkage disequilibrium in finite populations. *Theoretical and applied genetics*, 38(6):226–231, 1968. doi: 10.1007/BF01245622.

T. C. Koopmans. *Linear regression analysis of economic time series*, volume 20. De erven F. Bohn nv, 1937.

R. Lewontin. The interaction of selection and linkage. i. general considerations; heterotic models. *Genetics*, 49(1):49, 1964. URL `https://www.genetics.org/content/49/1/49`.

R. C. Lewontin and K.-i. Kojima. The evolutionary dynamics of complex polymorphisms. *Evolution*, 14(4):458–472, 1960. doi: 10.1111/j.1558-5646.1960.tb03113.x.

T. Maruki and M. Lynch. Genome-wide estimation of linkage disequilibrium from population-level high-throughput sequencing data. *Genetics*, 197(4):1303–1313, 2014. ISSN 0016-6731. doi: 10.1534/genetics.114.165514.

M. Pal. Consistent moment estimators of regression coefficients in the presence of errors in variables. *Journal of Econometrics*, 14(3):349 – 364, 1980. ISSN 0304-4076. doi: 10.1016/0304-4076(80)90032-9.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

T. Robertson and J. D. Cryer. An iterative procedure for estimating the mode. *Journal of the American Statistical Association*, 69(348):1012–1016, 1974. doi: 10.1080/01621459.1974.10480246.

U. R. Rosyara, W. S. De Jong, D. S. Douches, and J. B. Endelman. Software for genome-wide association studies in autopolyploids and its application to potato. *The Plant Genome*, 9(2), 2016. doi: 10.3835/plantgenome2015.08.0037.

M. Slatkin. Linkage disequilibrium-understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477, 2008. doi: 10.1038/nrg2361.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.08.430270; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

C. Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904. doi: 10.2307/1422689.

M. Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 10 2016. ISSN 1465-4644. doi: 10.1093/biostatistics/kxw041.

J. A. Sved and W. G. Hill. One hundred years of linkage disequilibrium. *Genetics*, 209(3):629–636, 2018. ISSN 0016-6731. doi: 10.1534/genetics.118.300642.

J. G. A. M. L. Uitdewilligen, A.-M. A. Wolters, B. B. D'hoop, T. J. A. Borm, R. G. F. Visser, and H. J. van Eck. A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLOS ONE*, 8(5):1–14, 05 2013. doi: 10.1371/journal.pone.0062355.

X. Wen and M. Stephens. Using linear predictors to impute allele frequencies from summary or pooled genotype data. *The annals of applied statistics*, 4(3):1158–1182, 2010. ISSN 1932-6157. doi: 10.1214/10-aoas338.

Y. C. J. Wientjes, R. F. Veerkamp, and M. P. L. Calus. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics*, 193(2):621–631, 2013. ISSN 0016-6731. doi: 10.1534/genetics.112.146290.

Wolfram Research, Inc. Mathematica, Version 12.2, 2020. URL `https://www.wolfram.com/mathematica`. Champaign, IL.

X. Zhu and M. Stephens. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nature communications*, 9(1):1–14, 2018. doi: 10.1038/s41467-018-06805-x.