the geneticssociety

www.nature.com/hdy

ARTICLE

Scalable bias-corrected linkage disequilibrium estimation under genotype uncertainty

David Gerard[1]

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, but are almost as fast as naive approaches based only on posterior mean genotypes. This opens up bias-corrected LD estimation to genome-wide applications. In addition, we provide standard errors for these moment-based estimators. All methods discussed in this manuscript are implemented in the `ldsep` package, available on the Comprehensive R Archive Network (https://cran.r-project.org/package=ldsep).

Heredity (2021) 127:357–362; https://doi.org/10.1038/s41437-021-00462-5

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 that loci are less dependent than in reality (Gerard 2021).

This bias is particularly strong in polyploids, organisms with more than two complete sets of chromosomes. Unlike diploids, polyploids exhibit multiple levels of heterozygosity. For example, at a biallelic locus with alleles A and a, a heterozygous diploid would have genotype Aa, whereas a heterozygous tetraploid might have genotypes Aaaa, AAaa, or AAAa. These multiple levels of heterozygosity make polyploid dosage more difficult to estimate, and exacerbate the impact on estimation of data-specific quirks, such as allelic bias and overdispersion (Gerard et al. 2018). This all increases genotype uncertainty in polyploid organisms, increasing the effect of LD attenuation. Therefore, in Gerard (2021) we derived maximum likelihood estimates (MLEs) that have lower bias and are consistent estimates of LD. This approach was particularly helpful for polyploids.

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.

Here, we derive scalable approaches to estimate LD that account for genotype uncertainty ("Materials and methods"). 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 ("Simulations") and real data ("LD estimates for *Solanum tuberosum*"), 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.

MATERIALS AND 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). There are two types of LD measures considered in the literature, "haplotypic" (called "gametic" in the diploid literature) and "composite." Haplotypic LD measures are more familiar, representing the association between loci that reside on the same haplotype (Hedrick et al. 1978), whereas composite LD measures aggregate the associations between alleles on all haplotypes between two loci (Cockerham and Weir 1977; Weir 1979). As obtaining estimates of haplotypic LD from unphased genotypes typically requires additional assumptions (such as Hardy–Weinberg

[1]Department of Mathematics and Statistics, American University, Washington, DC, USA. email: dgerard@american.edu

Received: 20 April 2021 Revised: 19 July 2021 Accepted: 19 July 2021
Published online: 9 August 2021

equilibrium), we will only consider estimating composite measures of LD. Advantageously, these composite measures are appropriate LD measures for generic autopolyploid, allopolyploid, and segmental allopolyploid populations, even in the absence of Hardy–Weinberg equilibrium (Gerard 2021). 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.

We will now review these composite measures of LD at biallelic loci. Let $\boldsymbol{G} = (G_A, G_B)$ be the random variable of genotypes of a K-ploid individual at loci A and B, where each G_j is the dosage (from 0 to K) of an allele at locus j. A sample of individuals, $\boldsymbol{G}_1, \boldsymbol{G}_2, \ldots, \boldsymbol{G}_n$ is assumed to be independent and identically distributed to $\boldsymbol{G}$. The composite measure of correlation between loci A and B is just the Pearson correlation,

$$\rho = \mathrm{cor}(G_A, G_B). \tag{1}$$

The composite LD coefficient is the covariance divided by the ploidy K,

$$\Delta = \frac{1}{K}\mathrm{cov}(G_A, G_B). \tag{2}$$

We divide by the ploidy in Eq. (2) so that, for a population in Hardy–Weinberg equilibrium, the composite LD coefficient equals the well-known haplotypic LD coefficient. The possible values of Δ are bounded, with the size of this bound depending on the allele frequencies at each locus, making it difficult to compare LD across loci. To create a measure of LD that is less dependent on allele frequencies, we have the composite standardized LD coefficient,

$$\Delta' = \Delta/\Delta_m, \text{ where} \tag{3}$$

$$\Delta_m = \begin{cases} \min\{E[G]_A E[G]_B, (K - E[G]_A)(K - E[G]_B)\}/K^2 & \text{if } \Delta < 0, \text{ and} \\ \min\{E[G]_A(K - E[G]_B), (K - E[G]_A)E[G]_B\}/K^2 & \text{if } \Delta > 0. \end{cases} \tag{4}$$

One can show that Δ' is free to vary between $-K$ and K, but is constrained between -1 and 1 for populations in Hardy–Weinberg equilibrium. For further details of these measures see Gerard (2021).

We wanted to create LD estimators of Eqs. (1)–(3) that account for genotype uncertainty while also being agnostic to the genotyping technology, e.g., microarrays (Fan et al. 2003), next-generation sequencing (Baird et al. 2008; Elshire et al. 2011), or mass spectrometry (Oeth et al. 2009). One way to do this is to use only the genotype posterior distributions for each individual, which are often provided by different genotyping software that analyze data from different genotyping technologies (e.g. Clark et al. 2019; Gerard and Ferrão 2019; Gerard et al. 2018; Serang et al. 2012; Voorrips et al. 2011; Zych et al. 2019). We will thus assume that the user provides the posterior means and variances for the genotypes for each individual at two loci, which can be easily obtained from the full posterior distributions for each individual. An advantage of this approach is its modularity. That is, as genotyping platforms improve and become better calibrated, the approach below will still be usable without having to create a tailor-made method to estimate LD directly from these new genotyping platforms.

To define our estimators of LD, let X_{iA} and X_{iB} be the posterior mean genotypes at loci A and B for individual $i \in \{1, \ldots, n\}$. Let Y_{iA} and Y_{iB} be the posterior variances of genotypes 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.

$$u_{xA} := \frac{1}{n}\sum_{i=1}^{n} X_{iA}, \;\; u_{xB} := \frac{1}{n}\sum_{i=1}^{n} X_{iB}, \tag{5}$$

$$v_{xA} := \frac{1}{n-1}\sum_{i=1}^{n} (X_{iA} - u_{xA})^2, \;\; v_{xB} := \frac{1}{n-1}\sum_{i=1}^{n} (X_{iA} - u_{xB})^2, \tag{6}$$

$$c_x := \frac{1}{n-1}\sum_{i=1}^{n} (X_{iA} - u_{xA})(X_{iB} - u_{xB}), \tag{7}$$

$$u_{yA} := \frac{1}{n}\sum_{i=1}^{n} Y_{iA}, \text{ and } u_{yB} := \frac{1}{n}\sum_{i=1}^{n} Y_{iB}. \tag{8}$$

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

$$\hat{\Delta} := \left(\frac{u_{yA} + v_{xA}}{v_{xA}}\right)\left(\frac{u_{yB} + v_{xB}}{v_{xB}}\right)\left(\frac{c_x}{K}\right). \tag{9}$$

The estimated Pearson correlation is as follows:

$$\hat{\rho} := \sqrt{\frac{u_{yA} + v_{xA}}{v_{xA}}}\sqrt{\frac{u_{yB} + v_{xB}}{v_{xB}}}\frac{c_x}{\sqrt{v_{xA} v_{xB}}}. \tag{10}$$

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

$$\hat{\Delta}' := \hat{\Delta}/\hat{\Delta}_m, \text{ where} \tag{11}$$

$$\hat{\Delta}_m := \begin{cases} \min\left\{u_{xA} u_{xB}, (K - u_{xA})(K - u_{xB})\right\}/K^2 & \text{if } c_x < 0, \text{ and} \\ \min\left\{u_{xA}(K - u_{xB}), (K - u_{xA})u_{xB}\right\}/K^2 & \text{if } c_x > 0. \end{cases} \tag{12}$$

We can compare our new estimators to those researchers typically use in practice. Since the population LD parameters (1)–(3) are population moments of the individual genotypes, researchers typically set X_{iA} and X_{iB} as estimates of G_{iA} and G_{iB}, and then use the sample moments of the X_{iA}'s and X_{iB}'s to estimate these population moments. That is

$$\hat{\rho}^{(naive)} := \frac{c_x}{\sqrt{v_{xA} v_{xB}}}, \tag{13}$$

$$\hat{\Delta}^{(naive)} := \frac{1}{K} c_x, \text{ and} \tag{14}$$

$$\hat{\Delta}'^{(naive)} := \frac{\hat{\Delta}^{(naive)}}{\hat{\Delta}_m}. \tag{15}$$

Comparing Eqs. (13)–(15) to Eqs. (9)–(11), we see that our new estimators take the naive estimators most researchers use in practice and inflate these by a multiplicative effect. Such multiplicative effects are sometimes called "reliability ratios" in the measurement error models literature (Fuller 2009).

Standard errors are important for hypothesis testing (Brown 1975), read-depth suggestions (Maruki and Lynch 2014), and shrinkage (Dey and Stephens 2018). Because estimators (9)–(11) 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 of the Supplementary Material).

Section S3 of the Supplementary Material contains practical considerations for improving our estimates of LD. We apply hierarchical shrinkage (Stephens 2016) on the log of the reliability ratios to improve estimation performance (Section S3.1). As we have observed unstable behavior when SNPs are mostly monoallelic, we apply a thresholding strategy to mitigate the effects of unusually large reliability ratios (Section S3.2). We also truncate LD estimates when sampling variability causes estimates (9)–(11) to lie outside their theoretical boundaries (Section S3.3). Section S4 of the Supplementary Material contains some theoretical discussions on why our methods perform as well as the MLE in the simulations of Section 3.1.

All methods are implemented in the `ldsep` package on the Comprehensive R Archive Network https://cran.r-project.org/package=ldsep.

RESULTS

Simulations

Comparison to the MLE and the standard approach. We compared our moment-based estimators (9)–(11) to those of the MLE of Gerard (2021) as well as the naive estimators that calculate the sample covariance and sample correlation between posterior mean genotypes at two loci (13)–(15). 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 and Ferrão 2019; Gerard et al. 2018) to generate read-counts at read-depths of either 10 or 100, a sequencing error rate of 0.01, an overdispersion value of 0.01, and no allele bias. `Updog` was then used to generate genotype likelihoods and genotype posterior distributions for

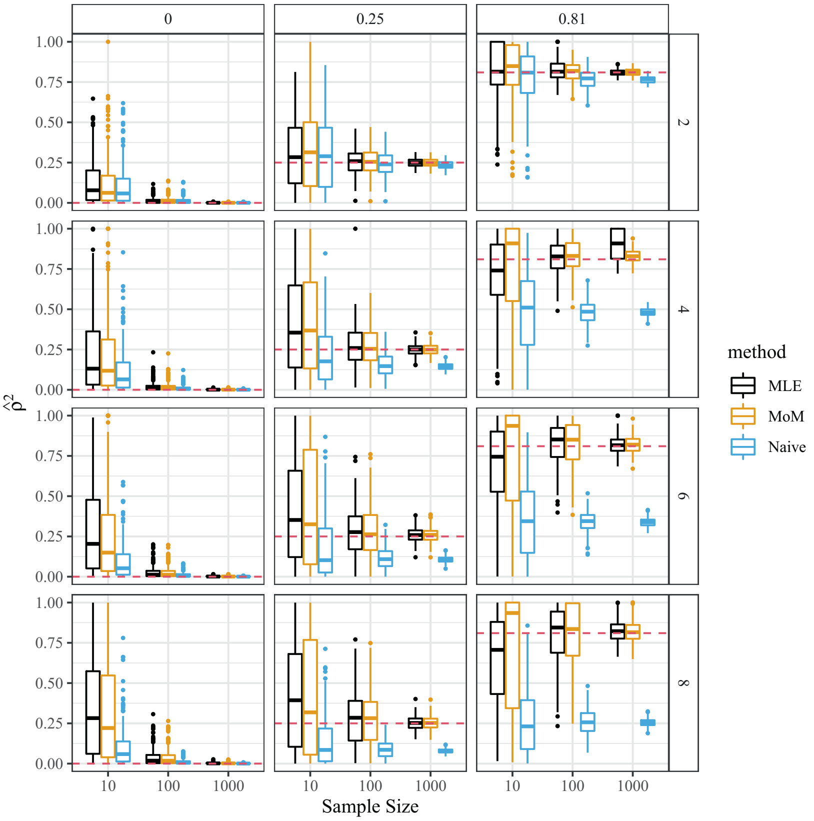

Fig. 1 Estimate of ρ^2 (y-axis) for the maximum likelihood estimator (Gerard 2021) (MLE), our new moment-based estimator (Eq. (10)) (MoM), and the naive squared sample correlation coefficient between posterior mean genotypes (Eq. (13)) (Naive). The x-axis indexes the sample size, the row-facets index 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.

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 Fig. 1. The results for other scenarios are similar and may be found in Figs. S5–S21 of the Supplementary Material. 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 when the sample size is moderate to large ($n = 100$ or 1000) but can be unstable for very small sample sizes ($n = 10$) (Figs. S1 and S2 of the Supplementary Material). This is not unexpected as the standard errors rely on asymptotic approximations (Section S2).

Additional simulation results, exploring our estimators when applied to rare variants, are presented in Section S5 of the Supplementary Material. The conclusions of that section are the same as here: the naive approach performs better at complete linkage equilibrium due to its attenuation bias, but performs worse at larger ploidies and larger levels of LD. However, we note that LD between rare variants is, in general, difficult to estimate.

The effect of using different genotyping strategies. Our new methods rely on accurate genotyping priors, which can be obtained adaptively using empirical Bayes approaches using sufficiently many samples. We therefore wished to study the effects of using either a fixed prior or a different genotyping platform. To do this, we generated posterior genotype probabilities under four scenarios: (i) the empirical Bayes approach of estimating the prior implemented by `updog` (Gerard and Ferrão 2019; Gerard et al. 2018), (ii) the empirical Bayes approach of estimating the prior implemented by `polyRAD` (Clark et al. 2019), (iii) a Bayesian approach assuming an unrealistic uniform prior on the genotypes, as implemented by `updog`, and (iv) a Bayesian approach assuming an unrealistic "horseshoe-like" prior on the genotypes that puts most mass on genotypes 0 and K, as implemented by `updog`. Specifically, for the "horseshoe-like" prior, the prior probability of a dosage of 0 or K was set to 0.45 each and the prior probability of dosages 1, …, $(K - 1)$ was set to $0.1/(K - 1)$ each.

We ran simulations under the same parameter settings of "Comparison to the MLE and the standard approach", where genotyping uncertainty had the greatest effect on LD estimation: higher ploidy species ($K = 8$) with $p_A = p_B = 0.5$ and a Pearson correlation $\rho = 0.9$. We simulated $n \in \{10, 100, 1000, 10000\}$ individuals, with a sequencing depth of 5, 10, or 100. As in "Comparison to the MLE and the standard approach", we generated genotypes and read-counts using the `updog` software at a sequencing error rate of 0.01, an overdispersion parameter of 0.01, and no allele bias. We then used the above four procedures to generate genotype posterior probabilities. These were fed into `ldsep` to obtain estimates of ρ. We replicated each simulation setting 200 times.

The results are presented in Fig. 2. There, we find that for larger sequencing depths (e.g., ≈100×), one can essentially use a uniform prior and normalize the genotype likelihoods to be posterior probabilities. The genotype posteriors using this simple approach

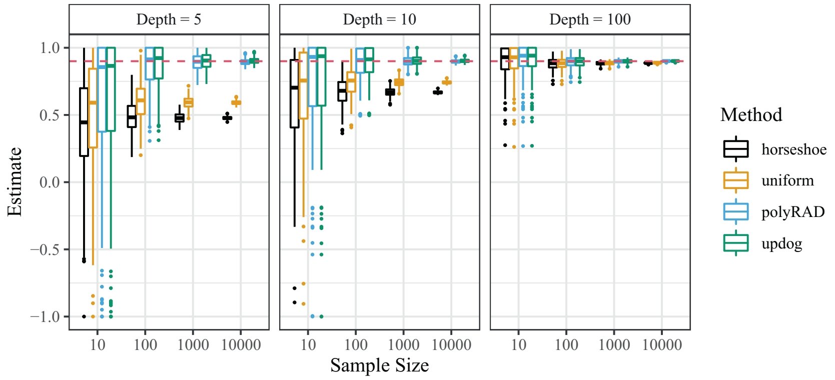

Fig. 2 Estimates of ρ using Eq. (10) (y-axis) when the true ρ is 0.9 (red dashed line) for different sample sizes (x-axis), different read-depths (facets) and different methods for obtaining the genotype posterior probabilities. The `updog` software (Gerard and Ferrão 2019; Gerard et al. 2018) was used either with an empirical Bayes approach to estimate the prior ("updog"), a fixed uniform prior ("uniform") or a fixed unrealistic "horseshoe-like" prior ("horseshoe"). The `polyRAD` software (Clark et al. 2019) was also used to obtain posterior genotype probabilities ("polyRAD").

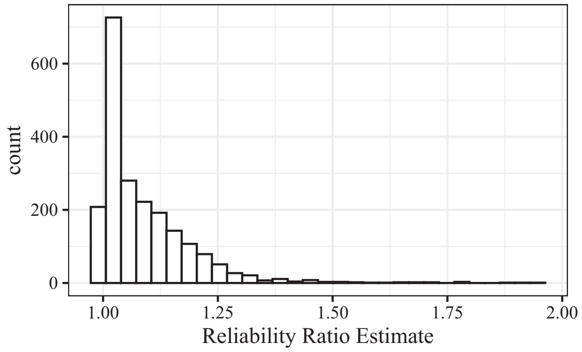

Fig. 3 Reliability ratio estimates. Histogram of estimated reliability ratios (S69) using the data from Uitdewilligen et al. (2013).

are close enough to those using adaptive approaches to provide decent LD estimates. However, for smaller read-depths, using a fixed prior has a deleterious effect. In such cases, one should use an adaptive genotyping approach that can consistently estimate the prior for larger sample sizes, even at lower read-depths. Many approaches that accomplish this exist, but for our analyses we found that two designed specifically for sequencing data work well in practice: `updog` (Gerard and Ferrão 2019; Gerard et al. 2018) and `polyRAD` (Clark et al. 2019). For non-sequencing data, there exist adaptive methods as well (Serang et al. 2012; Voorrips et al. 2011; Zych et al. 2019).

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 and Ferrão 2019; Gerard et al. 2018) 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 >0.95 or <0.05, and filtered out SNPs with a variance of posterior means <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 (Eq. (10)).

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 s it took to estimate all pairwise correlations using the naive approach. In Gerard (2021), we found that the MLE approach took about 0.1 s 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.

The histogram of estimated reliability ratios is presented in Fig. 3. We see there that the reliability ratios of most SNPs only increase their correlation estimates by <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 Fig. 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.

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 from, for example, repeated measurements on the same individuals) (Degracie and Fuller 1972; Koopmans 1937), using instrumental variables (Carter and Fuller 1980), and using distributional assumptions (Pal 1980). See Fuller (2009) for a detailed introduction to this vast field. In order to accommodate different data types (Baird et al. 2008; Elshire et al. 2011; Fan et al. 2003; Oeth et al. 2009) and different genotyping programs (Clark et al. 2019; Gerard and Ferrão 2019; Gerard et al. 2018; Serang et al. 2012; Voorrips et al. 2011; Zych et al. 2019), and therefore increase the generality of our methods, we limited ourselves to using just posterior genotype probabilities to calculate LD. This excluded using these previous approaches. Our solution, then, 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 S6 of the Supplementary Material). This supports using the posterior mean as a covariate in simple linear

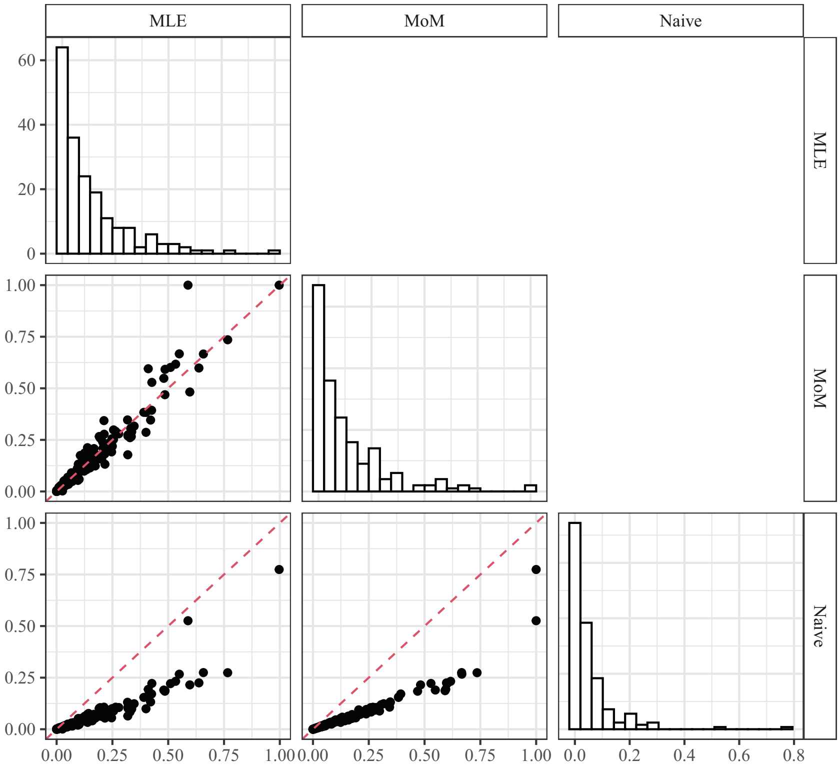

Fig. 4 Pairs plot for ρ^2 estimates between the twenty SNPs from Uitdewilligen et al. (2013) with the largest estimated reliability ratios when using either maximum likelihood estimation (MLE) (Gerard 2021), our new moment-based approach (Eq. (10)) (MoM), or the naive approach using just posterior means (Naive). The dashed line is the $y = x$ line. The MLE and the moment-based approach result in much more similar LD estimates.

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 nonlinear models (Carroll et al. 2006). Developing methods to account for genotype uncertainty in these more complicated settings is a research interest of the author, and a topic for future work.

We would not recommend using our methods to analyze diploid genomes. As seen in the simulations of "Comparison to the MLE and the standard approach," diploid approaches that do not account for genotype uncertainty perform fine, even at low depths, because genotype uncertainty is much less of an issue for diploids. Furthermore, phasing approaches are well-established and highly effective in the diploid literature (Browning and Browning 2007; Li et al. 2010; Scheet and Stephens 2006; Swarts et al. 2014), and our approach would likely not perform comparatively well against haplotype-aware LD estimation methods that use such phased information. However, in polyploids, haplotype estimation is much harder to achieve (Cheng et al. 2021; Mollinari and Garcia 2019; Shen et al. 2016; Zheng et al. 2016), and so accurate approaches that leverage only read-based information between two SNPs are important.

In this article, we demonstrated that naive LD estimates are typically attenuated toward zero in higher ploidy organisms due to the effects of genotype uncertainty. To correct for this bias, we presented moment-based approaches that perform as well as principled likelihood-based approaches, but only take a fraction of the computation time. Possible future directions include (i) extending our methods to multiallelic loci and (ii) evaluating the downstream consequences of using our improved LD estimates, such as for effective population size estimation (Ragsdale and Gravel 2019; Waples 2006) or admixture estimation (Loh et al. 2013). Our moment-based estimators will allow researchers to use de-biased LD estimators for such tasks at scale.

DATA AVAILABILITY

All methods discussed in this manuscript are implemented in the `ldsep` package, available on the Comprehensive R Archive Network (https://cran.r-project.org/package=ldsep) under a GPL-3 license. Scripts to reproduce the results of this research are available on Zenodo (https://doi.org/10.5281/zenodo.4543473). 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, and https://doi.org/10.1371/journal.pone.0062355.s010.

REFERENCES

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA et al. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3(10):1–7

Brown A (1975) Sample sizes required to detect linkage disequilibrium between two or three loci. Theor Popul Biol 8(2):184–201

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81(5):1084–1097

Carroll R, Ruppert D, Stefanski L, Crainiceanu C (2006) Measurement error in nonlinear models: a modern perspective, second edition. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, Boca Raton, FL

Carter RL, Fuller WA (1980) Instrumental variable estimation of the simple errors-in-variables model. J Am Stat Assoc 75(371):687–692

Cheng H, Concepcion GT, Feng X, Zhang H, Li H (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods 18(2):170–175

Clark LV, Lipka AE, Sacks EJ (2019) polyRAD: genotype calling with uncertainty from sequencing data in polyploids and diploids. G3: Genes, Genomes, Genet 9(3):663–673

Cockerham CC, Weir BS (1977) Digenic descent measures for finite populations. Genet Res 30(2):121–147

Degracie JS, Fuller WA (1972) Estimation of the slope and analysis of covariance when the concomitant variable is measured with error. J Am Stat Assoc 67(340):930–937

Dey KK, Stephens M (2018) CorShrink: empirical Bayes shrinkage estimation of correlations, with applications. bioRxiv

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES et al. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6(5):1–10

Fan J, Oliphant A, Shen R, Kermani BG, García F, Gunderson KL et al. (2003) Highly parallel SNP genotyping. Cold Spring Harb Symp Quant Biol 68:69–78

Fox EA, Wright AE, Fumagalli M, Vieira FG (2019) ngsLD: evaluating linkage disequilibrium using genotype likelihoods. Bioinformatics 35(19):3855–3856

Fuller WA (2009) Measurement error models. John Wiley & Sons, New York, NY

Gerard D (2021) Pairwise linkage disequilibrium estimation for polyploids. Mol Ecol Resour 21(4):1230–1242

Gerard D, Ferrão LFV (2019) Priors for genotyping polyploids. Bioinformatics 36 (6):1795–1800

Gerard D, Ferrão LFV, Garcia AAF, Stephens M (2018) Genotyping polyploids from messy sequencing data. Genetics 210(3):789–807

Hedrick P, Jain S, Holden L (1978) Multilocus systems in evolution. In: Hecht MK, Steere WC, Wallace B (eds), Evolutionary biology, vol 11. Springer, New York, NY, p 101–184

Hill W, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38(6):226–231

Koopmans TC (1937) Linear regression analysis of economic time series, vol 20. De erven F. Bohn nv, Haarlem, Netherlands

Lewontin R (1964) The interaction of selection and linkage. I. general considerations; heterotic models. Genetics 49(1):49

Lewontin RC, Kojima K (1960) The evolutionary dynamics of complex polymorphisms. Evolution 14(4):458–472

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34 (8):816–834

Loh P, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D et al. (2013) Inferring admixture histories of human populations using linkage disequilibrium. Genetics 193(4):1233–1254

Maruki T, Lynch M (2014) Genome-wide estimation of linkage disequilibrium from population-level high-throughput sequencing data. Genetics 197(4):1303–1313

Mollinari M, Garcia AAF (2019) Linkage analysis and haplotype phasing in experimental autopolyploid populations with high ploidy level using hidden markov models. G3: Genes, Genomes, Genet 9(10):3297–3314

Oeth P, del Mistro G, Marnellos G, Shi T, van den Boom D (2009) Qualitative and quantitative genotyping using single base primer extension coupled with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MassARRAY®). In: Komar A (ed) Single nucleotide polymorphisms. Humana Press, Totowa, NJ, p 307–343

Pal M (1980) Consistent moment estimators of regression coefficients in the presence of errors in variables. J Econom 14(3):349–364

R Core Team (2021). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Ragsdale AP, Gravel S (2019) Unbiased estimation of linkage disequilibrium from unphased data. Mol Biol Evol 37(3):923–932

Rosyara UR, De Jong WS, Douches DS, Endelman JB (2016) Software for genome-wide association studies in autopolyploids and its application to potato. Plant Genome 9 (2):1–10

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78(4):629–644

Serang O, Mollinari M, Garcia AAF (2012) Efficient exact maximum *a posteriori* computation for Bayesian SNP genotyping in polyploids. PLoS ONE 7(2):1–13

Shen J, Li Z, Chen J, Song Z, Zhou Z, Shi Y (2016) SHEsisPlus, a toolset for genetic studies on polyploid species. Sci Rep 6:24095

Slatkin M (2008) Linkage disequilibrium-understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9(6):477

Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15(1):72–101

Stephens M (2016) False discovery rates: a new deal. Biostatistics 18(2):275–294

Sved JA, Hill WG (2018) One hundred years of linkage disequilibrium. Genetics 209 (3):629–636

Swarts K, Li H, Navarro JAR, An D, Romay MC, Hearne S et al. (2014) Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Genome 7(3):1–12

Uitdewilligen JGAML, Wolters AA, D'hoop BB, Borm TJA, Visser RGF, van Eck HJ (2013) A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLoS ONE 8(5):1–14

Voorrips RE, Gort G, Vosman B (2011) Genotype calling in tetraploid species from bi-allelic marker data using mixture models. BMC Bioinform 12(1):172

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

Weir BS (1979) Inferences about linkage disequilibrium. Biometrics 35(1):235–254

Wen X, Stephens M (2010) Using linear predictors to impute allele frequencies from summary or pooled genotype data. Ann Appl Stat 4(3):1158–1182

Wientjes YCJ, Veerkamp RF, Calus MPL (2013) The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193 (2):621–631

Zheng C, Voorrips RE, Jansen J, Hackett CA, Ho J, Bink MC (2016) Probabilistic multilocus haplotype reconstruction in outcrossing tetraploids. Genetics 203 (1):119–131

Zhu X, Stephens M (2018) Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. Nat Commun 9(1):1–14

Zych K, Gort G, Maliepaard CA, Jansen RC, Voorrips RE (2019) FitTetra 2.0—improved genotype calling for tetraploids with multiple population and parental data support. BMC Bioinform 20(1):148

ACKNOWLEDGEMENTS

Most analyses were performed using the R statistical language (R Core Team 2021).

AUTHOR CONTRIBUTIONS

David Gerard developed the methodology, wrote the software, implemented the study, and wrote the manuscript.

COMPETING INTERESTS

The author declares no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41437-021-00462-5.

Correspondence and requests for materials should be addressed to D.G.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.