Supporting Information for: "polyRAD: Genotype calling with uncertainty from sequencing data in polyploids and diploids"

Methods

Algorithm initialization

Allele frequencies are used in Eqns. 1-2 for estimating genotype likelihoods, yet are not known if no genotype estimation has yet been performed. Therefore, the polyRAD algorithms are initialized using a rough estimation of allele frequencies based on read depth. For every individual *j* and allele *l*, a depth ratio *dr* is calculated using allelic read depths *a* and *b* as defined in the main manuscript. When a_{jl} and b_{jl} are both zero, dr_{jl} is treated as missing data.

Eqn 5: $dr_{jl} = \frac{a_{jl}}{a_{jl}+b_{jl}}$

For the mapping population and HWE pipelines, these depth ratios are then used for estimating allele frequencies, where n_{ind} is the number of individuals with reads:

Eqn. 6: $p_l = \frac{\sum_{j=1}^{n_{ind}} dr_{jl}}{n_{ind}}$

For the population structure pipeline, the matrix of depth ratios (individual*allele across all loci) is subjected to probabilistic principal components analysis using the R package pcaMethods (Stacklies *et al.* 2007). The first several principal components, up to an arbitrary threshold for the rate of change in their R^2 value, are retained. The depth ratios for each allele are then regressed on the retained principal components, and the fitted values are treated as local allele frequencies for the population of origin of each individual. These local allele frequencies are equivalent to the truncated singular value decomposition of the *dr* matrix. Mean local allele frequencies across all individuals are then estimated in order to be used in Eqn. 1.

Mapping populations

Any number of generations of backcrossing (gen_{bc}), intermating (gen_{int}), and self-fertilization (gen_{self}) can be specified. A donor and recurrent parent are specified, although these are interchangeable when $gen_{bc} = 0$. Where k_{maxD} is the maximum possible ploidy of the donor parent, k_{maxR} is the maximum possible ploidy of the recurrent parent, the expected allele frequencies in the population are:

Eqn. 7: $\left(\frac{m}{n}\right)_{m=0}^{n}$ where $n = (gen_{bc} + 1) * (k_{maxD} + k_{maxR})$

Allele frequencies estimated in Eqn. 6 are then rounded to the nearest expected allele frequency, and those frequencies are used for estimating genotype likelihoods in Eqns. 1-2.

For each possible ploidy, the genotypes with the highest likelihoods are identified for the parents. Where possible, parental genotypes are corrected if they do not match the corresponding progeny

allele frequency. For each possible ploidy combination, taking into account inheritance mode, gametes are simulated for each generation of backcrossing, intermating, and self-fertilization in order to estimate genotype prior probabilities ($P(G)$) for the progeny. Genotype posterior probabilities are then estimated using Eqn. 3.

Optionally, information from linked markers can be incorporated at this point in the pipeline. For a given allele, alleles from loci within a user-defined distance in basepairs are tested for linkage by estimating Pearson's correlation coefficient between weighted mean genotypes (Eqn. 4). If the correlation coefficient (r) is above a certain threshold (0.5 by default, and required to be positive), the markers are considered to be linked. If both alleles only have two possible genotypes segregating in the population, genotype priors for allele l based on genotype posterior probabilities for allele m are:

Eqn. 8: $$P\left(G_l \middle| P(G_m|a_m, b_m)\right) = r^2 * P(G_m|a_m, b_m) + (1 - r^2)/2$$

Given that r^2 represents the proportion of variance of allele l explained by allele m, it is used in Eqn. 8 as a mixing weight to determine how much influence the posterior genotype probabilities for allele m have on the prior genotype probabilities for allele l.

Otherwise, if either allele has more than two possible genotypes, it is unknown whether both alleles are linked in all parental haplotypes. In that case, the posterior probabilities for all genotypes for allele l are regressed on the posterior probabilities for all genotypes for allele m, and the fitted values are treated as priors for allele l. Although this method is somewhat ad hoc, it enables the use of linked alleles for predicting genotypes when linkage phase is unknown. We found that including this method reduced genotyping error by 4% in our *Miscanthus sinensis* F1 population, as compared to only utilizing linkage between alleles where only two genotypes were possible (data not shown).

Priors based on linkage across all alleles linked to a given allele are then obtained by multiplication, where M is the total number of linked alleles and i is a particular allele copy number (genotype):

Eqn. 9: $$P\left(G_{il} \middle| P(G|a, b)\right) = \frac{\prod_{m=1}^{M} P\left(G_{il} \middle| P(G_m|a_m, b_m)\right)}{\sum_{i=0}^{k} \prod_{m=1}^{M} P\left(G_{il} \middle| P(G_m|a_m, b_m)\right)}$$

Multiplication is used in Eqn. 9 because it causes alleles that are more tightly linked to a given allele, and/or have higher read depth, to have a larger influence on priors than alleles that are less tightly linked and/or have lower read depth. Additionally, if multiple linked alleles are in agreement about which allele copy number is most probable at a given allele, the prior probability of that allele copy number will be higher than if it were estimated from a single linked allele.

Genotype posterior probabilities are then re-estimated as:

Eqn. 10: $$P\left(G_{il} \middle| a_l, b_l, P(G|a, b)\right) = \frac{L(a_l, b_l|G_i) * P\left(G_{il} \middle| P(G|a, b)\right) * P(G_{il})}{\sum_{i=0}^{k} L(a_l, b_l|G_i) * P\left(G_{il} \middle| P(G|a, b)\right) * P(G_{il})}$$

Hardy-Weinberg equilibrium and inbreeding without population structure

For autopolyploids, genotype priors under HWE are:

Eqn. 11: $P(G_{il}) = \binom{k}{i} * p_l^i * (1 - p_l)^{k-i}$

If the self-fertilization rate, *s*, is above zero, priors are adjusted according to Equation 6 of de Silva et al. (2005):

Eqn. 12: $P\left(G_{l,self}\right) = (1 - s)(\boldsymbol{I} - s\boldsymbol{A})^{-1}P(G_l)$

where $\boldsymbol{I}$ is the identity matrix. $\boldsymbol{A}$ is a square matrix, with parental genotypes in columns and progeny genotypes in rows, indicating the frequency of progeny genotypes produced by the self-fertilization of each possible parental genotype [$\boldsymbol{A}^{\mathrm{T}}$ in de Silva et al. (2005)].

For allopolyploids, it is assumed that each allele only segregates in one subgenome. If *nsubgen* is the number of subgenomes, the allele frequency within that subgenome s is:

Eqn. 13: $p_{ls} = (p_l \bmod \frac{1}{n_{subgen}}) * n_{subgen}$

Genotype priors within subgenomes are then estimated as in Eqns. 11 and 12. The number of subgenomes that are fixed for allele *l* are estimated as

Eqn. 14: $p_l \operatorname{div} \frac{1}{n_{subgen}}$

in order to obtain overall priors for allele copy number. Genotype posterior probabilities are then estimated as in Eqn. 3. Posterior mean genotypes are estimated as in Eqn. 4. Allele frequencies are then re-estimated from posterior mean genotypes, where n_{ind} is the total number of individuals:

Eqn. 15: $p_l = \frac{\sum_{j=1}^{n_{ind}} pmg_{jl}}{n_{ind}}$

Re-estimation of genotype priors, genotype posterior probabilities, posterior mean genotypes, and allele frequencies then continues until allele frequencies converge.

Optionally, after the first round of posterior probability estimation, linkage between alleles at nearby loci can be estimated using Pearson's correlation coefficient between posterior mean genotypes as with mapping populations, with a default minimum correlation coefficient of 0.2. Using the same rationale as for Eqn. 8, priors based on genotype posterior probabilities at a linked allele are estimated as:

Eqn. 16: $P\left(G_{il}\middle|P(G_{im}|a_m, b_m)\right) = r^2 * P(G_{im}|a_m, b_m) + (1 - r^2)/(k + 1)$

Priors based on linkage across all alleles are estimated as in Eqn. 9, and posterior genotype probabilities are re-estimated as in Eqn. 10. Linkage is not re-estimated in subsequent iterations, in order to prevent overestimation, but genotype prior and posterior probabilties based on linkage are re-estimated in each iteration.

Population structure

The local allele frequencies estimated from PCA as described in "Algorithm initialization" are used for estimating local genotype frequencies under HWE or inbreeding as in Eqns. 11 and 12 in order to set genotype priors individually for each sample in the dataset. Genotype likelihoods, posterior probabilities, and posterior mean genotypes are then estimated according to Eqns. 2, 3, and 4, respectively. A new PCA is performed using posterior mean genotypes, and posterior mean genotypes are regressed on the PC axes in order to re-estimate local allele frequencies; the estimated local allele frequencies are equivalent to a truncated singular value decomposition of the *pmg* matrix, divided by the ploidy. Allele frequencies, genotype priors, genotype likelihoods, genotype posterior probabilities, posterior mean genotypes, and PCA are iteratively re-calculated until allele frequencies converge. Optionally, after the first round, linkage between alleles at nearby loci can be estimated. In order to estimate linkage disequilibrium that is not already explained by population structure, posterior mean genotypes are regressed on the PCA axes and the residuals are taken. Pearson's correlation coefficient is then estimated between these residuals and the weighted mean genotypes of nearby alleles. In all subsequent iterations, genotype priors based on linked alleles are estimated using Eqns. 16 and 9, then genotype posterior probabilities are estimated using Eqn. 10.

Multiple inheritance modes

When multiple possible inheritance modes are specified, genotype priors, likelihoods, and posterior probabilities are estimated for all inheritance modes. For each inheritance mode *h*, expected genotype frequencies across the whole dataset are taken from the genotype priors based on population parameters (i.e. not from the genotype priors based on linkage). Actual genotype counts are then estimated using genotype likelihoods:

Eqn. 17: $counts_{lh} = \sum_{j=1}^{n_{ind}} \frac{L(a_{jl}, b_{jl}|G_{ijlh})}{\sum_{i=0}^{k_h} L(a_{jl}, b_{jl}|G_{ijlh})}$

A pseudo-chi-squared statistic is then estimated for each allele and inheritance mode.

Eqn. 18: $\chi^2_{lh} = \frac{(counts_l - n_{taxa} * P(G_{il}))^2}{n_{taxa} * P(G_{il})}$

The pseudo-chi-squared statistic increases with the size of the deviation of the observed genotype frequencies from the expected genotype frequencies. Therefore, a larger value for χ^2_{lh} indicates a smaller likelihood that the inheritance mode is correct.

During pipeline iteration, and by default for the final output, posterior mean genotypes (Eqn. 3) are also weighted across the inverse of the pseudo-chi-squared values.

Eqn. 19: $pmg_l = \sum_h \left(pmg_{lh} * \frac{1/\chi^2_{lh}}{\sum_h 1/\chi^2_{lh}} \right)$

Results

Additional testing

To test polyRAD in a self-fertilizing diploid species with high marker density, SNP genotypes from 1179 *Glycine soja* (wild soybean) accessions were obtained from https://soybase.org/snps/ (Song *et al.* 2015). Chromosome 18, with 2957 SNPs, was selected for analysis and used for simulating RAD-seq data as had been done with *Miscanthus* and potato. For genotypes with > 0 reads, the lowest error was observed by using the GATK or diseq method of EBG, or by using any polyRAD method with an assumed selfing rate of 0.95 and linkage disequilibrium excluded from the model (Fig. S1A). When the assumed selfing rate was zero, accuracy of polyRAD was improved by incorporating population structure and linkage disequilibrium into the model, although polyRAD was less accurate than GATK or diseq due to miscalling homozygotes as heterozygotes (Fig. S1A). For genotypes with zero reads, the lowest error was observed using either rrBLUP or continuous genotypes output by polyRAD with population structure and linkage disequilibrium included in the model (Fig. S1B). Although assumed rate of self-fertilization had a large impact on polyRAD genotyping accuracy for genotypes with > 0 reads (Fig. S1A), it had little to no impact on polyRAD genotyping accuracy for genotypes with zero reads (Fig. S1B).

To simulate a self-fertilizing allohexaploid species similar to wheat, the same 2957 SNPs across 1179 *G. soja* accessions were used, and were treated as the genotypes within one subgenome, where the other two subgenomes were assumed to be fixed for the reference allele. RAD-seq read depth was simulated as before, but with scale = 15 rather than scale = 5 in order to get higher read depth. The EBG GATK method was tested, but not the EBG HWE or EBG diseq methods because they assume autopolyploidy. The EBG "alloSNP" model, which only allows two subgenomes, was run with subgenome 1 treated as tetraploid with all allele frequencies as zero, and subgenome 2 as diploid with unknown allele frequencies. polyRAD was run with an allohexaploid model (three diploid subgenomes) and an assumed self-fertilization rate of 0.95. Error (RMSE) was considerably lower across all read depths using polyRAD as compared to the GATK method and fitPoly, and somewhat lower as compared to EBG alloSNP and updog (Fig. S2A). On average, the polyRAD model with population structure, linkage disequilibrium, and continuous output gave a 70.3% (SE 0.3%) lower error rate than the GATK method, a 74.4% (SE 0.3%) lower error rate than fitPoly, a 62.3% (SE 0.3%) lower error rate than updog with continuous genotypes, and a 54.6% (SE 0.3%) lower error rate than the EBG alloSNP model on genotypes with more than zero reads. Accuracy of polyRAD was improved by assuming both population structure and linkage disequilibrium, both for genotypes with zero reads and > 0 reads (Fig. S2). Genotype imputation with rrBLUP using genotypes from the GATK method was similarly accurate to continuous genotype calls from polyRAD without assuming population structure or linkage disequilibrium, and considerably less accurate than polyRAD when population structure and linkage disequilibrium were assumed (Fig. S2B).

To test polyRAD in a domesticated but outcrossing species, we used data from 3650 SNPs across 96 diploid apple cultivars, available at https://www.rosaceae.org/ (Chagné *et al.* 2012), for simulation of RAD-seq data as was done in *Miscanthus*. polyRAD was similarly accurate to the

EBG HWE and diseq models and updog, and slightly more accurate than the EBG GATK model (Fig. S3A). Modeling population structure actually reduced accuracy of polyRAD in the apple dataset, perhaps due to the small number of accessions included in the analysis (Fig. S3). Linkage disequilibrium, however, improved genotyping accuracy (Fig. S3). For genotypes with zero reads, rrBLUP was the most accurate and LinkImpute the least accurate, with continuous output from polyRAD being similarly accurate to rrBLUP (Fig. S3B).

To test polyRAD in a domesticated autotetraploid species, we used data from 3762 SNPs across 221 potato cultivars, available at http://solcap.msu.edu/potato_infinium.shtml (Hamilton *et al.* 2011), for simulation of RAD-seq data as was done in *Miscanthus*. All polyRAD methods performed similarly to or slightly better than the EBG HWE and diseq methods, rrBLUP, and updog, and substantially better than the GATK or fitPoly method (Fig. S4). For genotypes with more than zero reads, using the polyRAD model with population structure, linkage disequilibrium, and continuous output, RMSE was reduced by 39.5% (SE 0.1%) with respect to the GATK method, 41.7% (SE 0.2%) with respect to fitPoly, 21.1% (SE 0.1%) with respect to the EBG "diseq" method, 18.3% (SE 0.2%) with respect to the EBG HWE method, and 9.7% (SE 0.4%) with respect to updog.

References

Chagné, D., R. N. Crowhurst, M. Troggio, M. W. Davey, B. Gilmore *et al.*, 2012 Genome-Wide SNP Detection, Validation, and Development of an 8K SNP Array for Apple (M. Bendahmane, Ed.). PLoS One 7: e31745.

Hamilton, J. P., C. N. Hansey, B. R. Whitty, K. Stoffel, A. N. Massa *et al.*, 2011 Single nucleotide polymorphism discovery in elite north american potato germplasm. BMC Genomics 12: 302.

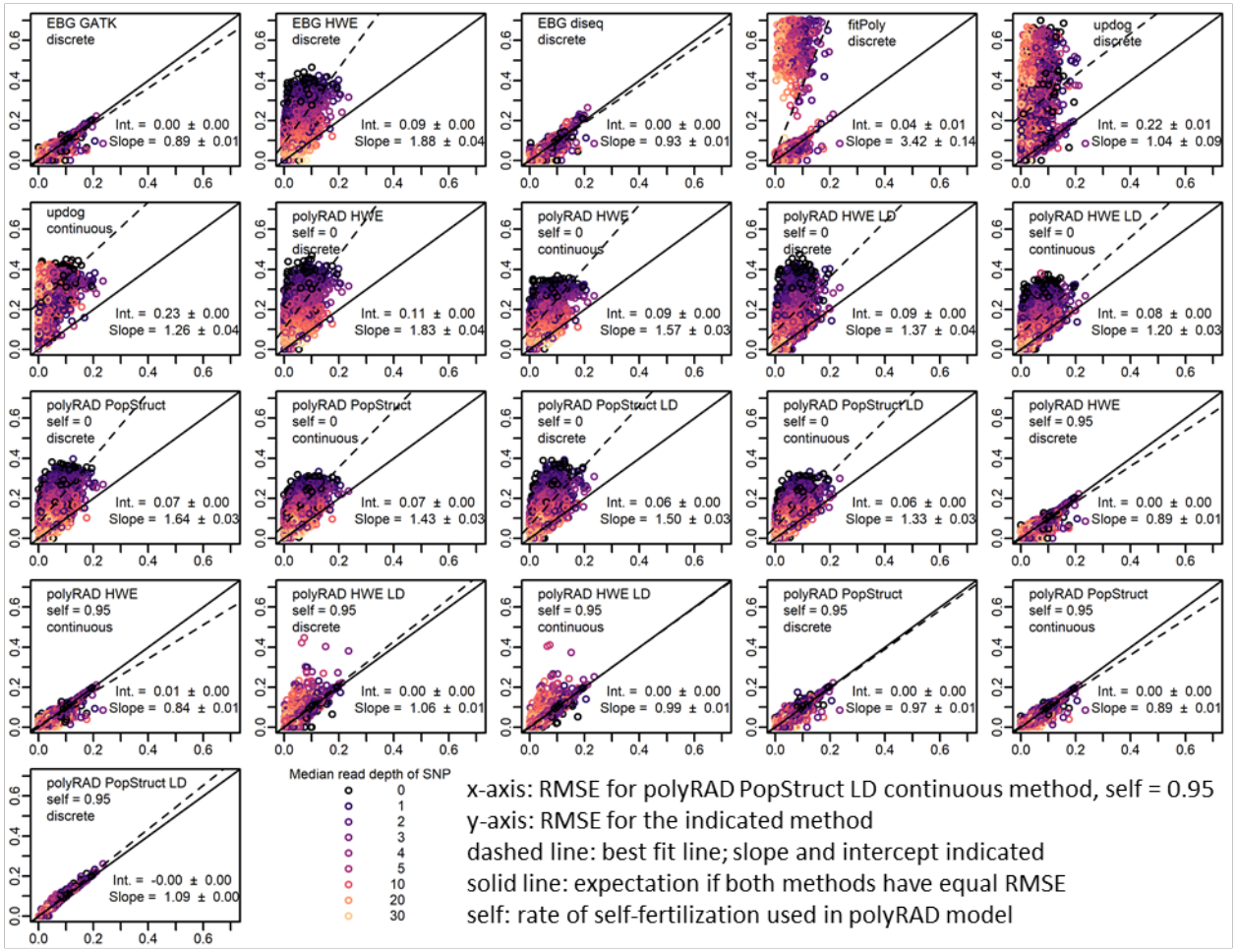
De Silva, H. N., A. J. Hall, E. Rikkerink, M. A. McNeilage, and L. G. Fraser, 2005 Estimation of allele frequencies in polyploids under certain patterns of inheritance. Heredity (Edinb). 95: 327–334.

Song, Q., D. L. Hyten, G. Jia, C. V. Quigley, E. W. Fickus *et al.*, 2015 Fingerprinting Soybean Germplasm and Its Utility in Genomic Research. G3 5: 1999–2006.

Stacklies, W., H. Redestig, M. Scholz, D. Walther, and J. Selbig, 2007 pcaMethods - A bioconductor package providing PCA methods for incomplete data. Bioinformatics 23: 1164–1167.

Figures

(A) Genotypes with read depth > 0



(B) Genotypes with read depth = 0

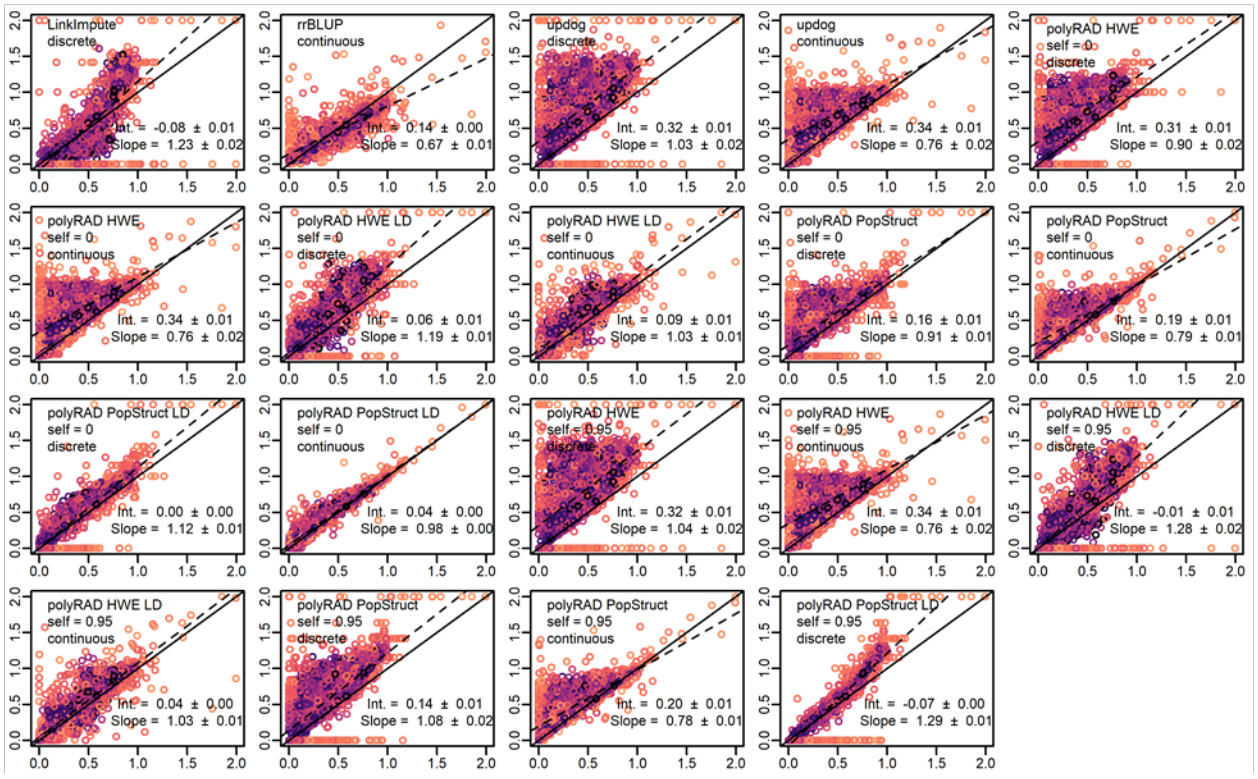


Fig. S1. Genotyping error of EBG, fitPoly, updog, polyRAD, LinkImpute, and rrBLUP in a diversity panel of 1179 diploid *Glycine soja* accessions. The benefits of incorporating population structure, linkage disequilibrium, and self-fertilization into the genotyping model and using continuous rather than discrete genotypes are illustrated. Genotypes were coded on a scale of 0 to 2. Root mean squared error (RMSE) was calculated between actual genotypes and genotypes ascertained from simulated RAD-seq reads at 2957 SNP markers on chromosome 18 (lower RMSE = higher accuracy). Each point represents one SNP. Median read depth is indicated by color, including genotypes with zero reads. The RMSE for continuous genotypes output by the polyRAD PopStruct LD method is shown on the x-axis, and the RMSE of other methods and types of genotypes (continuous or discrete) is shown on the y-axis. The dashed line indicates the ordinary least-squares regression with slope and intercept estimates, with standard errors. The "norm" model was used with updog. (A) RMSE calculated using only genotypes with more than zero reads. (B) RMSE calculated using only genotypes with zero reads, by genotyping or imputation method and genotype type.

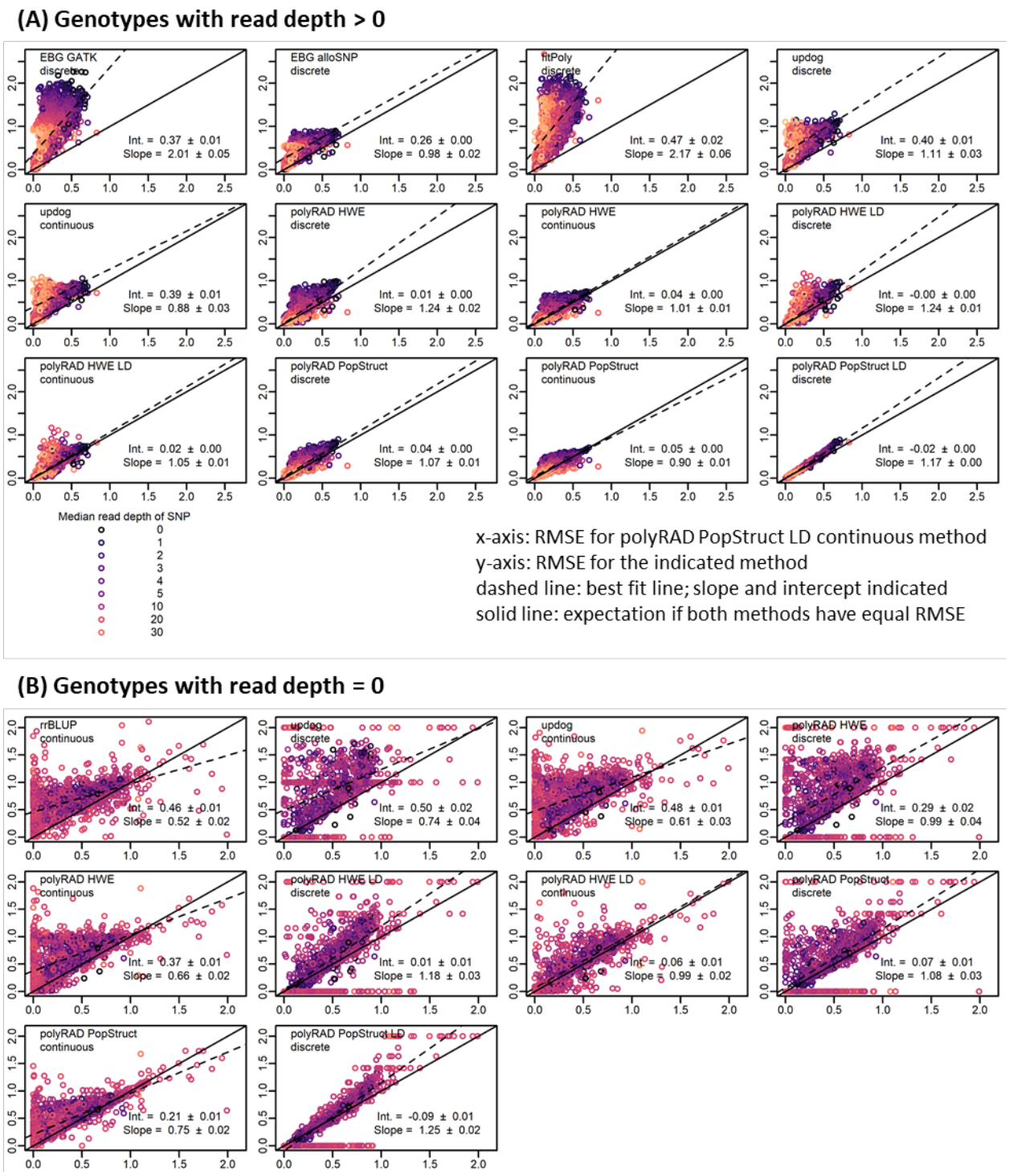


Fig. S2. Genotyping error of polyRAD, EBG, and rrBLUP on a diversity panel of a simulated allohexaploid species, generated from diploid SNP data from 1179 *Glycine soja* accessions using 2957 markers on chromosome 18. Two out of the three subgenomes were simulated as being fixed for the reference allele. The benefits of including inheritance mode, population structure, and linkage disequilibrium in the model are illustrated. polyRAD was run assuming a self-fertilization rate of 0.95. Root mean squared error (RMSE) was calculated between actual

genotypes (scored on a scale of 0 to 6) and genotypes ascertained from simulated RAD-seq reads. Each point represents one SNP. Median read depth is indicated by color, including genotypes with zero reads. The RMSE for continuous genotypes output by the polyRAD PopStruct LD method is shown on the x-axis, and the RMSE of other methods and types of genotypes (continuous or discrete) is shown on the y-axis. The dashed line indicates the ordinary least-squares regression with slope and intercept estimates, with standard errors. The "norm" model was used with updog. LinkImpute was not used because it was only designed for diploid genotypes. (A) RMSE calculated using only genotypes with more than zero reads. (B) RMSE calculated using only genotypes with zero reads, by genotyping or imputation method and genotype type.

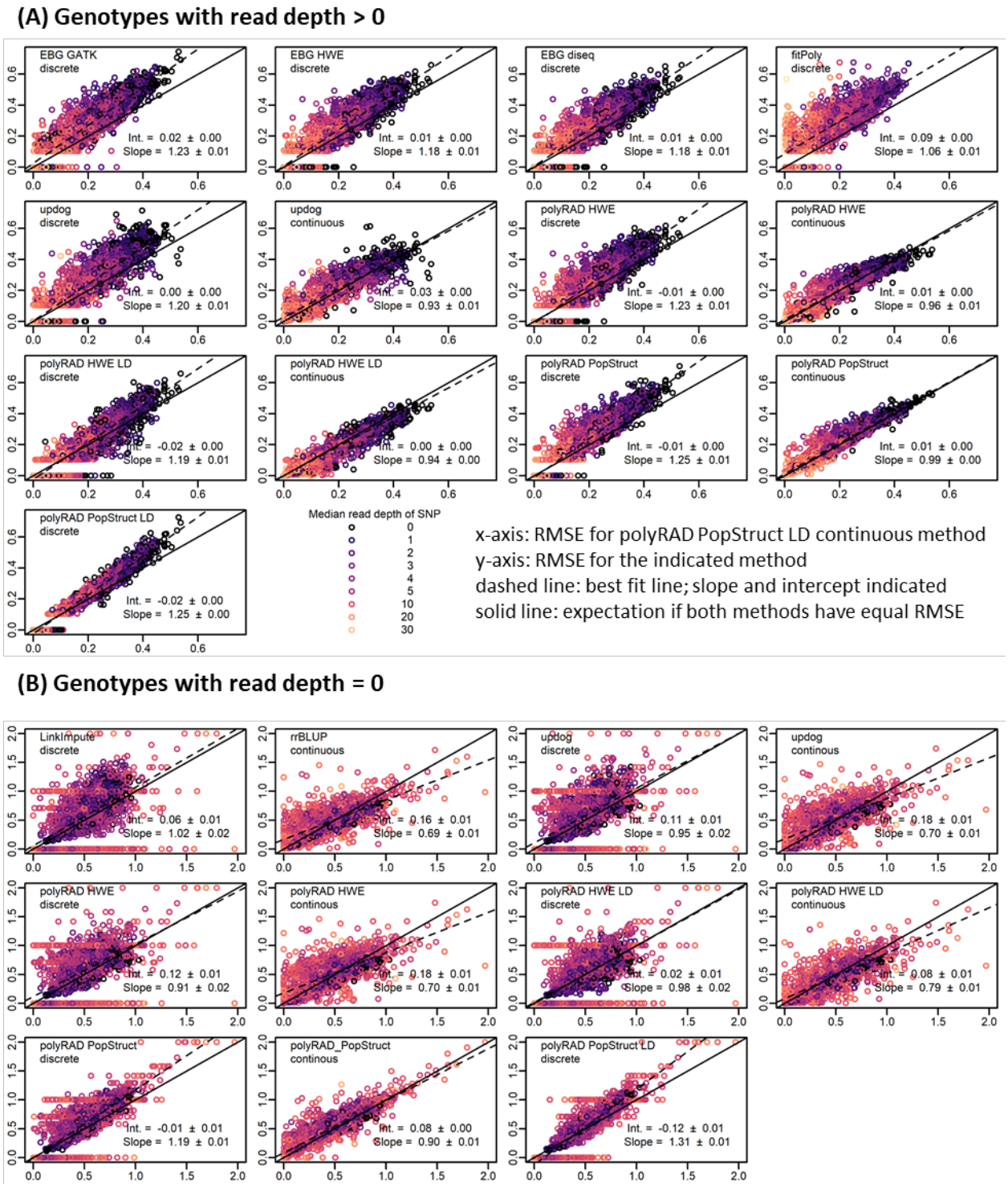


Fig. S3. Genotyping error of EBG, fitPoly, updog, polyRAD, LinkImpute, and rrBLUP in a diversity panel of 96 diploid apple cultivars. The benefits of incorporating linkage disequilibrium into the genotyping model and using continuous rather than discrete genotypes are illustrated. Genotypes were coded on a scale of 0 to 2. Root mean squared error (RMSE) was calculated between actual genotypes and genotypes ascertained from simulated RAD-seq

reads at 3650 SNP markers (lower RMSE = higher accuracy). Each point represents one SNP. Median read depth is indicated by color, including genotypes with zero reads. The RMSE for continuous genotypes output by the polyRAD PopStruct LD method is shown on the x-axis, and the RMSE of other methods and types of genotypes (continuous or discrete) is shown on the y-axis. The dashed line indicates the ordinary least-squares regression with slope and intercept estimates, with standard errors. The "norm" model was used with updog. (A) RMSE calculated using only genotypes with more than zero reads. (B) RMSE calculated using only genotypes with zero reads, by genotyping or imputation method and genotype type.

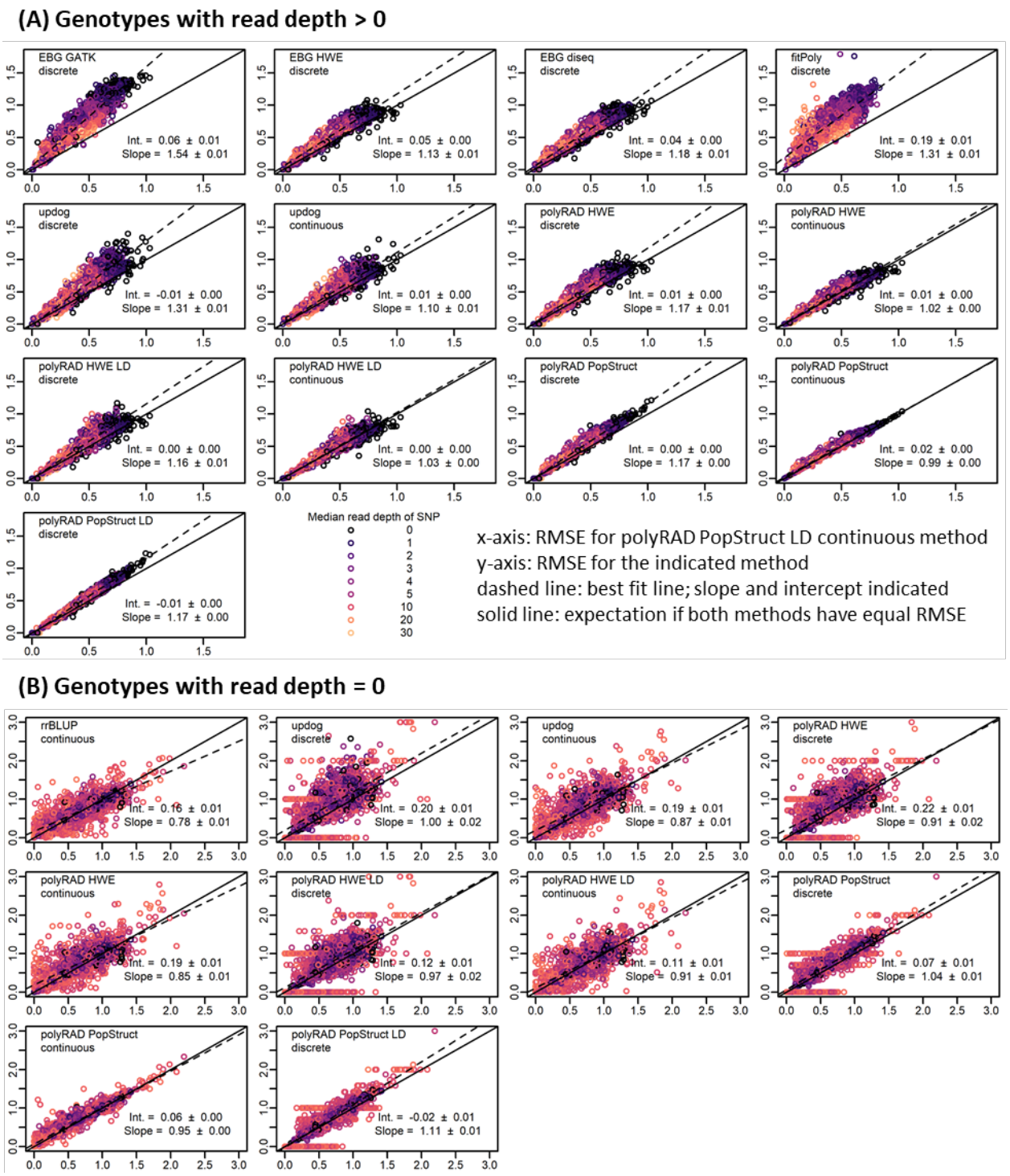


Fig. S4. Genotyping error of EBG, fitPoly, updog, polyRAD, and rrBLUP in a diversity panel of 221 tetraploid potato cultivars. The benefits of incorporating population structure into the genotyping model and using continuous rather than discrete genotypes are illustrated. Genotypes were coded on a scale of 0 to 4. Root mean squared error (RMSE) was calculated between actual genotypes and genotypes ascertained from simulated RAD-seq reads at 3762 SNP markers (lower RMSE = higher accuracy). Each point represents one SNP. Median read depth is

indicated by color, including genotypes with zero reads. The RMSE for continuous genotypes output by the polyRAD PopStruct LD method is shown on the x-axis, and the RMSE of other methods and types of genotypes (continuous or discrete) is shown on the y-axis. The dashed line indicates the ordinary least-squares regression with slope and intercept estimates, with standard errors. The "norm" model was used with updog. LinkImpute was not used because it was only designed for diploid genotypes. (A) RMSE calculated using only genotypes with more than zero reads. (B) RMSE calculated using only genotypes with zero reads, by genotyping or imputation method and genotype type.