A population-level statistic for assessing Mendelian behavior of genotyping-by-sequencing data from highly duplicated genomes

Lindsay V. Clark[1*], Wittney Mays[23], Alexander E. Lipka[2], and Erik J. Sacks[2]

[1]Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[2]Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[3]Sandia National Laboratories, Livermore, CA 94551, USA

[*]To whom correspondence should be addressed. Email: lvclark@illinois.edu

Abstract

Background

Given the economic and environmental importance of allopolyploids and other species with highly duplicated genomes, there is a need for methods to distinguish paralogs, i.e. duplicate sequences within a genome, from Mendelian loci, i.e. single copy sequences that pair at meiosis. Methods such as comparing observed and expected heterozygosity are frequently used to identify collapsed paralogs, but have limitations in genotyping-by-sequencing datasets, in which observed heterozygosity is difficult to estimate due to undersampling of alleles. These limitations are especially pronounced when the species is highly heterozygous or the expected inheritance is polysomic.

Results

We introduce a novel statistic, H_{ind}/H_E, that uses the probability of sampling reads of two different alleles at a sample*locus, instead of observed heterozygosity. The expected value of H_{ind}/H_E is the same across all loci in a dataset, regardless of read depth or allele frequency. In contrast to methods based on observed heterozygosity, it can be estimated and used for filtering loci prior to genotype calling. In addition to filtering paralogs, it can be used to filter loci with null alleles or high overdispersion, and identify individuals with unexpected ploidy and hybrid status. We demonstrate that the statistic is useful at read depths as low as five to 10, well below the depth needed for accurate genotype calling in polyploid and outcrossing species.

Conclusions

Our methodology for estimating H_{ind}/H_E across loci and individuals, as well as determining reasonable thresholds for filtering loci, is implemented in polyRAD v1.5, available at https://github.com/lvclark/polyRAD. In large sequencing datasets, we anticipate that the ability to filter markers and identify problematic individuals prior to genotype calling will save researchers considerable computational time.

Keywords

Polyploidy, single nucleotide polymorphism (SNP), heterozygosity, next generation DNA-sequencing (NGS), genome duplication

Background

Highly duplicated genome sequences are common throughout the plant kingdom. These include recent allopolyploids such as wheat, cotton, canola, strawberry, and coffee, as well as species with evidence of ancient whole genome duplication such as maize and legumes [1]. This

phenomenon is also present in the animal kingdom, for example allopolyploidy in the model frog *Xenopus*, as well as an ancient tetraploidization event followed by diploidization in salmonid fishes [2, 3]. For species in which paralogous sequences no longer pair at meiosis, accurate separation of paralogs in DNA and RNA sequence analysis, including reference genome assembly, remains challenging [4]. This separation of paralogs is especially important in variant calling, because SNPs and indels will not behave in a Mendelian fashion if the reads originate from more than one locus yet are erroneously attributed to a single locus [5]. Accurate variant calling therefore impacts all downstream analysis that assumes Mendelian inheritance, including linkage and QTL mapping, association studies, genomic selection, population genetics, and parentage analysis. For example, failure to remove paralogs from downstream analysis has been demonstrated to bias estimates of allele frequency and inbreeding as well as population structure [4, 6–8].

Due in part to the difficulty of assembling highly duplicated reference genomes, several methods have been published for filtering collapsed paralogous loci from genotyping-by-sequencing (GBS, including restriction-site associated DNA sequencing (RAD) approaches) datasets without the need for a reference genome. The most straightforward approach is to call genotypes and then determine if observed heterozygosity exceeds expected heterozygosity [9–11]. However, sampling error at low read depth can confound this filtering step by causing heterozygotes to be miscalled as homozygotes, lowering the observed heterozygosity. Moreover, estimating observed heterozygosity becomes complicated when polysomic inheritance is expected, due to the challenge of estimating allele copy number, as well as the question of how to quantify heterozygosity when multiple types of heterozygote are possible. Another approach is to filter loci that have read depth above an arbitrary threshold [12], although due to differences in

amplification efficiency based on fragment size and GC content, this method could fail to filter some paralogs while filtering other non-paralogous loci. Peterson et al. [13] developed a method, extended by Willis et al. [14], that involved counting the number of unique haplotypes per individual for a putative locus, with the idea that in a collapsed paralog, the number of haplotypes would exceed the ploidy. However, this method can be confounded by sequencing error and cross-contamination among samples, and its sensitivity depends on allele frequencies and ploidy. Other approaches have examined read depth ratios within individual genotypes [15] as well as read depth ratios in combination with observed heterozygosity [16]. Lastly, multiple methods identify putative paralogs based on networks of similarity among sequence tags [6, 17].

We present a novel statistic, H_{ind}/H_E, for evaluating marker quality, in particular for assessing whether a marker represents one Mendelian locus or multiple collapsed paralogous loci, based upon read depth distribution in a population. For a Mendelian locus, the statistic has the same expected value regardless of number of alleles, allele frequency, and total read depth. As a result, the distribution of the statistic can be visualized across loci in order to identify threshold values for filtering. The expected value can be calculated from ploidy (assuming disomic or polysomic inheritance) and the inbreeding coefficient, or the mode value of the statistic in a population can be used to estimate ploidy or inbreeding. Notably, because genotype calls are not needed in order to estimate this statistic, it can be used for filtering loci before any genotype calling is performed, saving computation time. We extend our Bayesian genotype calling software polyRAD [18] to implement the novel statistic and determine appropriate cutoffs.

Results

Utility of H_{ind}/H_E for detecting collapsed paralogs in a diversity panel

To compare the distribution of H_{ind}/H_E values for Mendelian loci versus collapsed paralogs, we aligned *M. sacchariflorus* tag sequences to the *M. sinensis* reference genome, in which they should align to the correct paralog most of the time, and to the *S. bicolor* reference genome, in which two paralogs from *Miscanthus* correspond to one alignment location. We found that loci with a mean read depth less than five had very low estimates of H_{ind}/H_E, likely due to restriction site polymorphisms or other technical issues (Fig. 1). As mean read depth increased above 100 in our dataset, however, loci tended to have H_{ind}/H_E values above the expectation for a Mendelian locus, suggesting that most loci at this depth and higher were in fact collapsed paralogs (Figs. 1 and 2).

When a mean depth of five was used as a cutoff and the *M. sinensis* genome was used as a reference, the peak H_{ind}/H_E value was slightly below the expected values of 0.5 for diploids and 0.75 for tetraploids (Fig. 2), indicating some inbreeding, likely due to population structure [19]. A second peak was observed at a higher value of H_{ind}/H_E (Fig. 2), likely representing sets of tags that belonged to different Mendelian loci despite aligning to the same location (i.e. misalignments). When *S. bicolor* was used as the reference genome, the opposite trend was observed, where most loci had a H_{ind}/H_E above the expected value, indicating collapsed paralogs, but a second peak was observed closer to the expected value, indicating regions in the *S. bicolor* genome that may only have synteny with one region of the *M. sinensis* genome (Fig. 2). Although the peaks overlapped somewhat, they were distinct enough that a reasonable threshold for identifying putative collapsed paralogs could be visually determined (Fig. 2, top panel). Moreover, although the diploid and tetraploid datasets were processed separately, they were

largely in agreement about which loci were Mendelian and which were collapsed paralogs (Fig. 2, bottom panel), suggesting that the filtering performed in one population can be applied to another population, which could be especially useful for populations that are too small for accurate estimation of H_{ind}/H_E.

In both the diploid and tetraploid datasets, the distribution and peak values of H_{ind}/H_E were similar regardless of whether biallelic SNPs or multiallelic, haplotype-based markers were used (Additional File 1: Fig. S1). However, the variance of H_{ind}/H_E was approximately 20% higher when SNPs were used, suggesting that the higher information content of multiallelic markers improves the precision of H_{ind}/H_E estimates.

By individual, H_{ind}/H_E reflects ploidy and hybrid status

In addition to evaluating the mean H_{ind}/H_E within loci, we also obtained the mean statistic within individuals in order to assess the utility of the statistic for determining ploidy. We found that H_{ind}/H_E increased with ploidy, largely independent of read depth (Fig. 3). Although the distributions overlapped too much for H_{ind}/H_E to be a conclusive indicator of ploidy, it could still potentially be used to identify outlier individuals whose ploidy should be confirmed by other means (e.g. flow cytometry). Additionally, because our empirical dataset included many natural interspecific (*M. sacchariflorus* × *M. sinensis*) F1 hybrid and backcross individuals, we were also able to observe that H_{ind}/H_E values were considerably higher in hybrids than in non-hybrids, reflecting higher heterozygosity (Fig. 3).

Variance and bias in the H_{ind}/H_E statistic using simulated data

Using simulated data resembling a diversity panel or natural population, variance and bias in the statistic were minimized if there were at least 500 samples, minor allele frequency was 0.05 or

higher, and read depth was at least 5 (Fig. 4). Ploidy had negligible impact on variance and bias (Fig. 4). As expected, inbreeding and overdispersion reduced the mean H_{ind}/H_E estimate, with the effect of overdispersion being greater at higher minor allele frequencies (Additional File 1: Fig. S2). The H_{ind}/H_E estimate also decreased linearly as null allele frequency increased (Additional File 1: Fig. S3).

In simulated F1 mapping populations, the standard deviation of the H_{ind}/H_E ranged from 0.01 to 0.07 depending on the marker type (Fig. 5). In tetraploids, marker types with high expected heterozygosity in the progeny, such as triplex x nulliplex and triplex x simplex, had lower variance in the estimate than marker types with lower expected heterozygosity in the progeny, such as simplex x nulliplex and simplex x simplex (Fig. 5). A few rare markers had H_{ind}/H_E estimates that deviated very far from the expected value, indicating that the parents were incorrectly genotyped (Fig. 5).

Discussion

Properties of the H_{ind}/H_E statistic

While H_{ind}/H_E can be used, in combination with other metrics, to assess locus quality, this should be performed with an understanding of what biological and technical phenomena can cause it to deviate from the expected value. Inbreeding from any source will lower the expected value below $(k - 1)/k$; this includes not only self-fertilization and preferential mating with relatives, but also population structure, which is why we observed values below $(k - 1)/k$ even in self-incompatible, wind-pollinated *M. sacchariflorus* (Figs. 1-3). A benefit of this, however, is that as long as ploidy is known and overdispersion can be reasonably estimated (e.g. with the *TestOverdispersion* function in polyRAD), H_{ind}/H_E can be used to estimate inbreeding, either at

the population or individual level, directly from sequence read depth. Additionally, individuals that are hybrids between species or between highly diverged populations, as well as DNA samples that are an accidental mix of two or more individuals, may have H_{ind}/H_E above the expected value (Fig. 3). Strong selection for homozygotes or heterozygotes at particular loci would be expected to lower and raise H_{ind}/H_E, respectively.

At the locus level, a H_{ind}/H_E that exceeds the expected value can be an indication that alleles are derived from paralogous loci rather than a true Mendelian locus. More broadly, if all alleles truly belong to a single locus, then the expected value is $(1 - F)(k - 1)/k$. However, if a set of random, independent alleles were assigned to one putative locus, the expected value of H_{ind}/H_E would be one, because the probability of sampling reads from two different alleles within one individual would be the same as the probability of sampling reads from two different alleles in the general population. A H_{ind}/H_E of zero could indicate a cytoplasmic marker, because while there may be variation in the population, each individual would only be expected to possess reads from one allele. Loci with highly overdispersed read depth distributions due to technical issues such as differential fragment size or variation in library preparation would also be expected to have H_{ind}/H_E below expectations; it may be advantageous to filter these from the dataset as they will tend to yield poor-quality genotype calls. Lastly, loci with common null alleles have lower than expected H_{ind}/H_E (Additional File 1: Fig. S3). Null alleles can be the result of restriction cut site polymorphism in RAD-based techniques, primer binding site mutations in amplicon sequencing, or deletion mutations using any genotyping method. Because they are a common problem, H_{ind}/H_E can be used to identify and filter loci with null alleles.

The expected value of H_{ind}/H_E is independent of read depth, number of individuals sampled, and the allele frequency. However, all of these factors influence the variance of the estimate (Fig. 4).

As there is no generalized formula to estimate the variance of a ratio, the variance of H_{ind}/H_E cannot be estimated mathematically. We recommend instead simulating data for Mendelian loci given the ploidy, inbreeding, sample size, and distribution of read depth and allele frequency observed in the dataset of interest. The distribution of H_{ind}/H_E across simulated loci then can be used to determine cutoff values for filtering loci in the empirical dataset. The *ExpectedHindHe* and *ExpecteHindHeMapping* functions are available in polyRAD for this purpose, and suggest cutoffs for filtering loci in order to retain 95% of Mendelian loci. Depending on the downstream application, we recommend considering the number of markers needed versus the importance of marker quality when determining thresholds for read depth, allele frequency, and H_{ind}/H_E.

H_{ind}/H_E is more useful for detecting paralogs when haplotypes are treated as alleles (i.e. loci can be multiallelic), as opposed to when all loci are treated as biallelic SNPs, simply due to the fact that multiallelic markers are more information-rich than biallelic markers for the same distribution of minor allele frequencies. We observed that, for the same set of SNPs in *M. sacchariflorus*, the median value of H_{ind}/H_E per locus was very similar regardless of whether they were phased into haplotypes within the span of a single RAD tag, but the variance in H_{ind}/H_E was about 20% higher for SNPs vs. haplotypes (Additional File 1: Fig. S1). This improved power and information content is why polyRAD generally imports multiallelic, haplotype-based genotypes rather than SNPs as the default. Other methods for marker calling in highly duplicated genomes have also benefitted from the use of haplotype information [11, 20], and multiallelic markers have been found to be advantageous over biallelic SNPs for linkage mapping in polyploids [21].

Uses of the H_{ind}/H_E statistic

We anticipate locus-filtering to be the most common application of the H_{ind}/H_E statistic, with major advantages being that it is not biased by read depth or allele frequency and can be estimated prior to genotype calling. Stringency of filtering should depend on the genotype quality needed for downstream analysis; for example, parentage analysis and QTL mapping are sensitive to genotyping errors, whereas genome-wide association studies and estimations of population structure from principal components analysis are less sensitive. Missing data rate, median read depth, and minor allele frequency are common criteria that should be used in combination with H_{ind}/H_E to determine which loci to retain for downstream analysis. In our empirical dataset, we found the loci ranging in depth from five to 100 had the best distribution of H_{ind}/H_E (Fig. 1), but a higher minimum depth may be required for applications that require accurate genotype calling, and the optimal maximum depth used in filtering depends on the overall depth of the dataset. The use of observed heterozygosity, read depth ratios within genotypes, and number of haplotypes per individual are redundant with H_{ind}/H_E and unnecessary if it has already been used for filtering. In addition to its use for detecting paralogs in highly duplicated genomes, H_{ind}/H_E can be used for marker filtering in less duplicated genomes where occasional paralogs are still an issue. Additionally, in any species, markers with low values of H_{ind}/H_E (e.g. below the 95% confidence interval generated by simulated data) are likely to have null alleles, high overdispersion, or other technical issues and should generally be removed from the dataset.

As an example use case with our *M. sacchariflorus* empirical datasets, the mode values H_{ind}/H_E of 0.4 and 0.65 for diploids and tetraploids can be used to estimate inbreeding to be 0.15 and 0.05, respectively, assuming an overdispersion parameter of 9 (Fig. 2 and Additional File 1: Fig.

S2). Based on these parameters as well as the sample size, read depth distribution, and allele frequency distribution of the datasets after removing markers with a mean read depth below five, the *ExpectedHindHe* function of polyRAD suggests retaining markers with H_{ind}/H_E ranging from 0.26 to 0.57 for diploids and 0.54 to 0.78 for tetraploids. These cutoffs capture the main peak of the distribution well, while omitting most of the peak that represents collapsed paralogs (Fig. 2). Using these cutoffs, 47% and 50% of markers would be retained for diploids and tetraploids, respectively.

Although less accurate for determining ploidy than techniques such as flow cytometry, when averaged within individuals, H_{ind}/H_E can be used to identify individuals whose ploidy might deviate from expectations and should be confirmed. If flow cytometry is not an option, several other tools exist for the estimation of ploidy directly from next-generation sequencing data [22]. Lastly, H_{ind}/H_E could be potentially useful for improving reference genome assemblies, increasing the value of complementing a de novo assembly with a resequencing or genotyping-by-sequencing effort in a large population or diversity panel. Regions of the reference genome that contain collapsed paralogs are expected to have inflated H_{ind}/H_E values, which could be visualized in a smoothed plot of H_{ind}/H_E vs. alignment position.

At a minor allele frequency of 0.05, a read depth of five or higher is sufficient to estimate H_{ind}/H_E with minimal variance (Fig. 4). It is notable that a read depth of five is too low to call genotypes with confidence, to some extent in diploids but especially in polyploids. However, using the H_{ind}/H_E statistic, such low depth data are useful for a variety of applications such as identification of outlier individuals in terms of ploidy and hybridity, estimation of inbreeding, identification of loci with technical issues, and assessment of reference genome quality. This in

turn can enable researchers to reduce sequencing costs by generating preliminary, low-depth datasets to evaluate these issues before (or instead of) sequencing more deeply.

Conclusions

Here we introduce the H_{ind}/H_E statistic, which can be used evaluating marker and sample quality in genotyping-by-sequencing datasets for a variety of downstream applications. We demonstrate that reads from paralogous loci cause the statistic to be above the expected value, whereas technical issues such as overdispersion and null alleles cause the statistic to be below the expected value. In typical datasets (hundreds of individuals, read depth above five) the statistic has sufficiently low variance to be useful for filtering loci. The polyRAD R package can estimate H_{ind}/H_E, suggest filtering cutoffs based on simulated data, and perform genotyping after filtering.

Materials and Methods

The H_{ind} statistic

Here we describe a novel statistic, H_{ind}, that is based on sequence read depth across all alleles at a given locus and sample, and is agnostic of genotype calls, inheritance mode, and ploidy. It is related to observed heterozygosity, H_O, which in a diploid can be thought of as a matrix of ones and zeros indicating whether the genotype at each sample*locus is heterozygous. H_{ind} is instead a number ranging from zero to one, indicating the probability that if two sequencing reads were sampled without replacement at that sample*locus, they would represent different alleles. The abbreviation “ind” stands for “individual”, as it is calculated for each individual before averaging across a population. It can be calculated for SNP loci or for multiallelic haplotype-based loci, as long as allelic read depth is available.

The expected value for H_{ind} in a natural population of diploids or polysomic polyploids is:

Eqn. 1: $\overline{H_{ind}} = \frac{k-1}{k} H_E \ (1 - F)$

where k is the ploidy, H_E is the expected heterozygosity at the same locus, and F is the inbreeding coefficient. H_E is the probability that two alleles drawn at random from the population will be different, $(1 - F)$ is the probability that two alleles randomly drawn from an individual will not be identical by descent, and $(k - 1)/k$ is the probability that two sequencing reads originated from different chromosome copies. Multiplied together, these three terms yield the probability that two sequence reads from one sample at one locus will be different from each other.

If we divide H_{ind} by H_E:

Eqn. 2: $\overline{H_{ind}}/H_E = \frac{k-1}{k} \ (1 - F)$

we now have a statistic that is only dependent on ploidy and inbreeding, two parameters that we will assume to be consistent across samples and loci.

In a mapping population, the term $H_E * (1 - F)$ must be replaced by the probability, for a given locus, that two locus copies in a progeny will be different alleles. This requires knowledge of the ploidy, parental genotypes, and population design including number of generations of backcrossing and self-fertilization. This probability, which we will call $H_{E.map}$, can be estimated by simulation of the cross. The expectation is then:

Eqn. 3: $\overline{H_{ind}}/H_{E.map} = \frac{k-1}{k}$

Empirical estimation of H_{ind}/H_E

Say that we have sequence read depths, $\{d_{1m} \dots d_{jm}\}$, across a set of *j* alleles at a single locus in an individual *m*. Total read depth in one individual is

Eqn. 4: $D_m = \sum_{i=1}^{j} d_{im}$

As long as there are two or more reads, we can estimate H_{ind} within that individual using the Gini-Simpson index [23]:

Eqn. 5: $\widehat{H}_{ind,m} = \left(1 - \sum_{i=1}^{j} \left(\frac{d_{im}}{D_m}\right)^2\right) \frac{D_m}{D_m - 1}$

For a population of *n* individuals with sequencing reads, allele frequencies are estimated from average within-individual read depth ratios:

Eqn. 6: $\hat{p}_i = \frac{\sum_{m=1}^{n} \frac{d_{im}}{D_m}}{n}$

And expected heterozygosity is estimated as

Eqn. 7: $\widehat{H}_E = 1 - \sum_{i=1}^{j} \hat{p}_i^2$

Averaged across *n* individuals with two or more reads at a given locus in a natural population, the expectation is then:

Eqn. 8: $\widehat{H}_{ind} / \widehat{H}_E = \frac{\sum_{m=1}^{n} \widehat{H}_{ind,m} / \widehat{H}_E}{n} \cong \frac{k-1}{k}(1 - F)$

In a mapping population, $\widehat{H}_{ind,m}$ is estimated in the same way. $H_{E.map}$ is estimated from parental genotypes and population design, and the expected average ratio within a locus is given in Eqn. 3.

Implementation in polyRAD

Functions for estimating H_{ind}/H_E and $H_{ind}/H_{E.map}$ are available in polyRAD v1.2 and later, and are named *HindHe* and *HindHeMapping*, respectively. Both utilize an internal Rcpp function for fast calculation, take a *RADdata* object as input, and return a matrix of values, with samples in rows and loci in columns. The mean value across rows can then be used to get a per-sample estimate, for identifying individuals that are interspecies hybrids or unexpected ploidies. The mean value across columns can be used to get a per-locus estimate for filtering loci. Additionally, polyRAD v1.5 and later includes the *ExpectedHindHe* and *ExpectedHindHeMapping* functions, which simulate data to emulate the sample size, allele frequency distribution or parental genotypes, and read depth distribution of an empirical dataset, and return the distribution of H_{ind}/H_E as if all loci were Mendelian, giving the user reasonable thresholds to use for filtering loci.

PolyRAD v1.5 is currently available on CRAN, and can be installed using *install.packages("polyRAD").*

Datasets for testing

Two types of datasets were used to test H_{ind}/H_E: (1) empirical data from a diversity panel of *Miscanthus sacchariflorus*, and (2) simulated datasets of diversity panels and of biparental F1 mapping populations. Previously published RAD-seq data for an *M. sacchariflorus* diversity panel [19] were used for the empirical tests. All species in the *Miscanthus* genus share an

ancient genome duplication, increasing the chromosome number to 19 from the base of 10 in the Andropogoneae tribe [24–26]. Moreover, some populations of *M. sacchariflorus* display autotetraploidy in addition to this genome duplication (4x = 76) [19, 27], allowing us to test our algorithm in situations where tetrasomic inheritance is expected, in addition to the more typical disomic inheritance. *Miscanthus* is also highly heterozygous due to being wind-pollinated and self-incompatible [28], thus heterozygosity cannot be used to identify paralogs as it could in an inbred crop species. Together, these factors make *M. sacchariflorus* an ideal test case.

To compare values of H_{ind}/H_E in putatively Mendelian markers versus collapsed paralogs, markers were called from the same dataset using either *Miscanthus sinensis* or *Sorghum bicolor* as a reference because *M. sinensis* has a whole genome duplication with respect to *S. bicolor*. Raw sequence reads from *M. sacchariflorus* were processed by the TASSEL-GBSv2 pipeline [29] to identify unique tag sequences and their depths in all individuals. Tag sequences were then aligned to the *Miscanthus sinensis* v7.1 reference genome [30] and the *Sorghum bicolor* v3.1.1 reference genome [31] using Bowtie 2 [32]. The tag manager feature of TagDigger [33] was used to process the SAM files, recording the alignment location of each tag in both reference genomes. Tag alignment locations within the *S. bicolor* reference were retained for further analysis if they corresponded to two alignment locations in the *M. sinensis* reference matching the known synteny between chromosomes. Under this filtering, 239,501 tags were retained at 18,402 *S. bicolor* alignment locations corresponding to 36,804 *M. sinensis* alignment locations, in a set of 356 diploid and 268 tetraploid individuals. H_{ind}/H_E was then estimated per-locus in polyRAD for both the *M. sinensis* and *S. bicolor* alignments.

To compare the variance of H_{ind}/H_E when biallelic SNPs were used versus multiallelic, haplotype-based markers, the TASSEL-GBSv2 pipeline was used to call SNP variants from *M.*

sacchariflorus and export them to VCF. Markers from chromosome 1 were imported to polyRAD using *VCF2RADdata*, with and without the option to phase SNPs into haplotypes, yielding 3710 and 10,458 loci, respectively. H_{ind}/H_E was then estimated by locus in polyRAD separately for diploids and tetraploids.

Simulated diversity panel datasets were generated in order to assess the effect of minor allele frequency, sample size, read depth, overdispersion, inbreeding, ploidy, and null alleles on variance and bias of the H_{ind}/H_E statistic, using the *SimGenotypes* and *SimAlleleDepth* functions in polyRAD v1.5. See Clark et al. [18] (Eqn. 2) for a definition of the overdispersion parameter; lower values result in allelic read depths that deviate further from the ratios expected based on allelic dosage. Three sets of data were simulated. (1) Minor allele frequencies of 0.01, 0.05, 0.1, and 0.2; sample sizes of 100, 500, and 1000; and sample*locus read depths of 2, 5, 10, 20, 50, 100, and 200 were simulated in all combinations under diploidy and tetraploidy, with no inbreeding and an overdispersion parameter of 20. For each combination, 5000 biallelic loci were simulated. (2) Minor allele frequencies of 0.01 and 0.05, overdispersion spanning all integers from 5 to 20, and inbreeding (F; the probability that two locus copies in an individual are identical by descent) spanning all intervals of 0.1 from 0 to 1 were simulated in all combinations under diploidy and tetraploidy, with a sample size of 500 and a read depth of 20. For each combination, 20,000 biallelic loci were simulated. (3) Minor non-null allele frequencies of 0.01 and 0.05 and null allele frequencies of 0.01, 0.05, 0.1, and 0.2 were simulated in all combinations under diploidy and tetraploidy, with a sample size of 500, a read depth of 20, overdispersion of 20, and no inbreeding. For each combination, 5000 triallelic (with one allele being null, i.e. having all of its reads discarded) loci were simulated.

Simulated F1 mapping population datasets were generated in order to assess the effect of ploidy and marker type on variance of the H_{ind}/H_E statistic. For diploids, testcross (homozygote x heterozygote) and F2 (heterozygote x heterozygote) markers were evaluated. For tetraploids, simplex x nulliplex (AAAB x AAAA), duplex x nulliplex (AABB x AAAA), triplex x nulliplex (ABBB x AAAA), simplex x simplex (AAAB x AAAB), simplex x duplex (AAAB x AABB), simplex x triplex (AAAB x ABBB), and duplex x duplex (AABB x AABB) markers were evaluated. For each marker type, 5000 biallelic markers were simulated in a population with 500 offspring, with a read depth of 20 and overdispersion parameter of 20.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Raw sequence reads for *M. sacchariflorus* are available on the NCBI Sequence Read Archive under accessions SRP026347, SRP048207, SRP063572, and SRP087645. Genotype calls and read depths for *M. sacchariflorus* are available on the Illinois Data Bank at https://doi.org/10.13012/B2IDB-8170405_V1. Tag alignments and counts for *M. sacchariflorus* using the *M. sinensis* and *S. bicolor* reference genomes are available on the Illinois Data Bank at https://doi.org/10.13012/B2IDB-4814898_V1. All scripts for testing the H_{ind}/H_E statistic are available on GitHub at https://github.com/lvclark/paralog_id, archived on Zenodo at https://doi.org/10.5281/zenodo.5425343.

Project name: polyRAD

Project home page: https://github.com/lvclark/polyRAD

Archived version: https://doi.org/10.5281/zenodo.1143744

Operating system: Platform independent

Programming language: R, C++ via Rcpp

Other requirements: R ≥ 3.5.0; CRAN packages fastmatch, Rccp, and stringi; Bioconductor package pcaMethods

License: GNU GPL (≥2)

Any restrictions to use by non-academics: None

Competing interests

The authors declare they have no competing interests.

Funding

This material is based upon work supported by the National Science Foundation under Grant No. 1661490. The funding body was not involved in the design, analysis, or interpretation of the study.

Authors' contributions

LVC designed the H_{ind}/H_E statistic, wrote the polyRAD software, performed the analysis, and wrote the manuscript. WM performed literature review and tested the software and statistic. AEL gave statistical advice. EJS provided the *M. sacchariflorus* datasets. All authors read and approved the final manuscript.

Acknowledgements

We thank Jiale He for testing various methods to detect paralogous loci in *Miscanthus*.

References

1. Renny-Byfield S, Wendel JF. Doubling down on genomes: Polyploidy and crop plants. Am J Bot. 2014;101:1711–25. doi:10.3732/ajb.1400119.

2. Gregory TR, Mable BK. Polyploidy in Animals. In: Gregory TR, editor. The Evolution of the Genome. San Diego: Elsevier; 2005. p. 427–517.

3. Session AM, Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S, et al. Genome evolution in the allotetraploid frog *Xenopus laevis*. Nature. 2016;538:336–43. doi:10.1038/nature19840.

4. Dufresne F, Stift M, Vergilino R, Mable BK. Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. Mol Ecol. 2014;23:40–69. doi:10.1111/mec.12581.

5. Kaur S, Francki MG, Forster JW. Identification, characterization and interpretation of single-nucleotide sequence variation in allopolyploid crop species. Plant Biotechnol J. 2012;10:125–38. doi:10.1111/j.1467-7652.2011.00644.x.

6. Nadukkalam Ravindran P, Bentzen P, Bradbury IR, Beiko RG. PMERGE: Computational filtering of paralogous sequences from RAD-seq data. Ecol Evol. 2018;8:7002–13. doi:10.1002/ece3.4219.

7. Verdu CF, Guichoux E, Quevauvillers S, De Thier O, Laizet Y, Delcamp A, et al. Dealing with paralogy in RADseq data: in silico detection and single nucleotide polymorphism validation in *Robinia pseudoacacia* L. Ecol Evol. 2016;6:7323–33. doi:10.1002/ece3.2466.

8. Meirmans PG, Van Tienderen PH. The effects of inheritance in tetraploids on genetic diversity and population divergence. Heredity (Edinb). 2013;110:131–7. doi:10.1038/hdy.2012.80.

9. Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. Mol Ecol Resour. 2011;11 Suppl 1:117–22. doi:10.1111/j.1755-0998.2010.02967.x.

10. Arruda MP, Brown P, Brown-Guedira G, Krill AM, Thurber C, Merrill KR, et al. Genome-wide association mapping of Fusarium head blight resistance in wheat using genotyping-by-sequencing. Plant Genome. 2016;9. doi:10.3835/plantgenome2015.04.0028.

11. Tinker NA, Bekele WA, Hattori J. Haplotag: Software for Haplotype-Based Genotyping-by-Sequencing Analysis. G3. 2016;6:857–63. doi:10.1534/g3.115.024596.

12. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. Stacks: building and genotyping loci de novo from short-read sequences. G3. 2011;1:171–82. doi:10.1534/g3.111.000240.

13. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLoS One. 2012;7:e37135. doi:10.1371/journal.pone.0037135.

14. Willis SC, Hollenbeck CM, Puritz JB, Gold JR, Portnoy DS. Haplotyping RAD loci: an efficient method to filter paralogs and account for physical linkage. Mol Ecol Resour. 2017;17:955–65. doi:10.1111/1755-0998.12647.

15. Gayral P, Melo-Ferreira J, Glémin S, Bierne N, Carneiro M, Nabholz B, et al. Reference-Free Population Genomics from Next-Generation Transcriptome Data and the Vertebrate-Invertebrate Gap. PLoS Genet. 2013;9. doi:10.1371/journal.pgen.1003457.

16. McKinney GJ, Waples RK, Seeb LW, Seeb JE. Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. Mol Ecol Resour. 2017;17:656–69. doi:10.1111/1755-0998.12613.

17. Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD, et al. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genet. 2013;9:e1003215. doi:10.1371/journal.pgen.1003215.

18. Clark L V., Lipka AE, Sacks EJ. polyRAD: Genotype Calling with Uncertainty from Sequencing Data in Polyploids and Diploids. G3. 2019;9:663–73. doi:10.1534/g3.118.200913.

19. Clark L V, Jin X, Petersen KK, Anzoua KG, Bagmet L, Chebukin P, et al. Population structure of *Miscanthus sacchariflorus* reveals two major polyploidization events, tetraploid-mediated unidirectional introgression from diploid *M. sinensis*, and diversity centred around the Yellow Sea. Ann Bot. 2019;124:731–48. doi:10.1093/aob/mcy161.

20. Clevenger JP, Ozias-Akins P. SWEEP: A Tool for Filtering High-Quality SNPs in Polyploid Crops. G3. 2015;5:1797–803. doi:10.1534/g3.115.019703.

21. Mollinari M, Olukolu BA, Da Pereira GS, Khan A, Gemenet D, Craig Yencho G, et al. Unraveling the hexaploid sweetpotato inheritance using ultra-dense multilocus mapping. G3 Genes, Genomes, Genet. 2020;10:281–92. doi:10.1534/g3.119.400620.

22. Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömvik M V. Current strategies of polyploid plant genome sequence assembly. Front Plant Sci. 2018;9:1660. doi:10.3389/fpls.2018.01660.

23. Simpson EH. Measurement of diversity. Nature. 1949;163:688. doi:10.1038/163688a0.

24. Ma X-F, Jensen E, Alexandrov N, Troukhan M, Zhang L, Thomas-Jones S, et al. High resolution genetic mapping by genome sequencing reveals genome duplication and tetraploid genetic structure of the diploid *Miscanthus sinensis*. PLoS One. 2012;7:e33821. doi:10.1371/journal.pone.0033821.

25. Swaminathan K, Chae WB, Mitros T, Varala K, Xie L, Barling A, et al. A framework genetic map for *Miscanthus sinensis* from RNAseq-based markers shows recent tetraploidy. BMC Genomics. 2012;13:142. doi:10.1186/1471-2164-13-142.

26. Kim C, Zhang D, Auckland SA, Rainville LK, Jakob K, Kronmiller B, et al. SSR-based genetic maps of *Miscanthus sinensis* and *M. sacchariflorus*, and their comparison to sorghum. Theor Appl Genet. 2012;124:1325–38. doi:10.1007/s00122-012-1790-1.

27. Clark L V, Stewart JR, Nishiwaki A, Toma Y, Kjeldsen JB, Jørgensen U, et al. Genetic structure of *Miscanthus sinensis* and *Miscanthus sacchariflorus* in Japan indicates a gradient of bidirectional but asymmetric introgression. J Exp Bot. 2015;66:4213–25. doi:10.1093/jxb/eru511.

28. Hirayoshi I, Nishikawa K, Kato R. Cytogenetical Studies on forage plants. (IV) Self-incompatibility in *Miscanthus.* Japanese J Plant Breed. 1955;5:167–70. doi:10.1270/jsbbs1951.5.167.

29. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633–5. doi:10.1093/bioinformatics/btm308.

30. Mitros T, Session AM, James BT, Wu GA, Belaffif MB, Clark L V., et al. Genome biology of the paleotetraploid perennial biomass crop *Miscanthus*. Nat Commun. 2020;11:5442. doi:10.1038/s41467-020-18923-6.

31. McCormick RF, Truong SK, Sreedasyam A, Jenkins J, Shu S, Sims D, et al. The *Sorghum bicolor* reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. Plant J. 2018;93:338–54. doi:10.1111/tpj.13781.

32. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. doi:10.1038/nmeth.1923.

33. Clark L V, Sacks EJ. TagDigger: user-friendly extraction of read counts from GBS and RAD-seq data. Source Code Biol Med. 2016;11:11. doi:10.1186/s13029-016-0057-7.

Figure captions

Figure 1. Relationship between H_{ind}/H_E statistic and mean sequence read depth per locus. Loci were called across 356 diploid and 268 tetraploid *Miscanthus sacchariflorus* based on alignments to the *M. sinensis* reference genome. The number of loci in each depth category is indicated. The expected value for a Mendelian locus in Hardy-Weinberg equilibrium is shown with a dashed line.

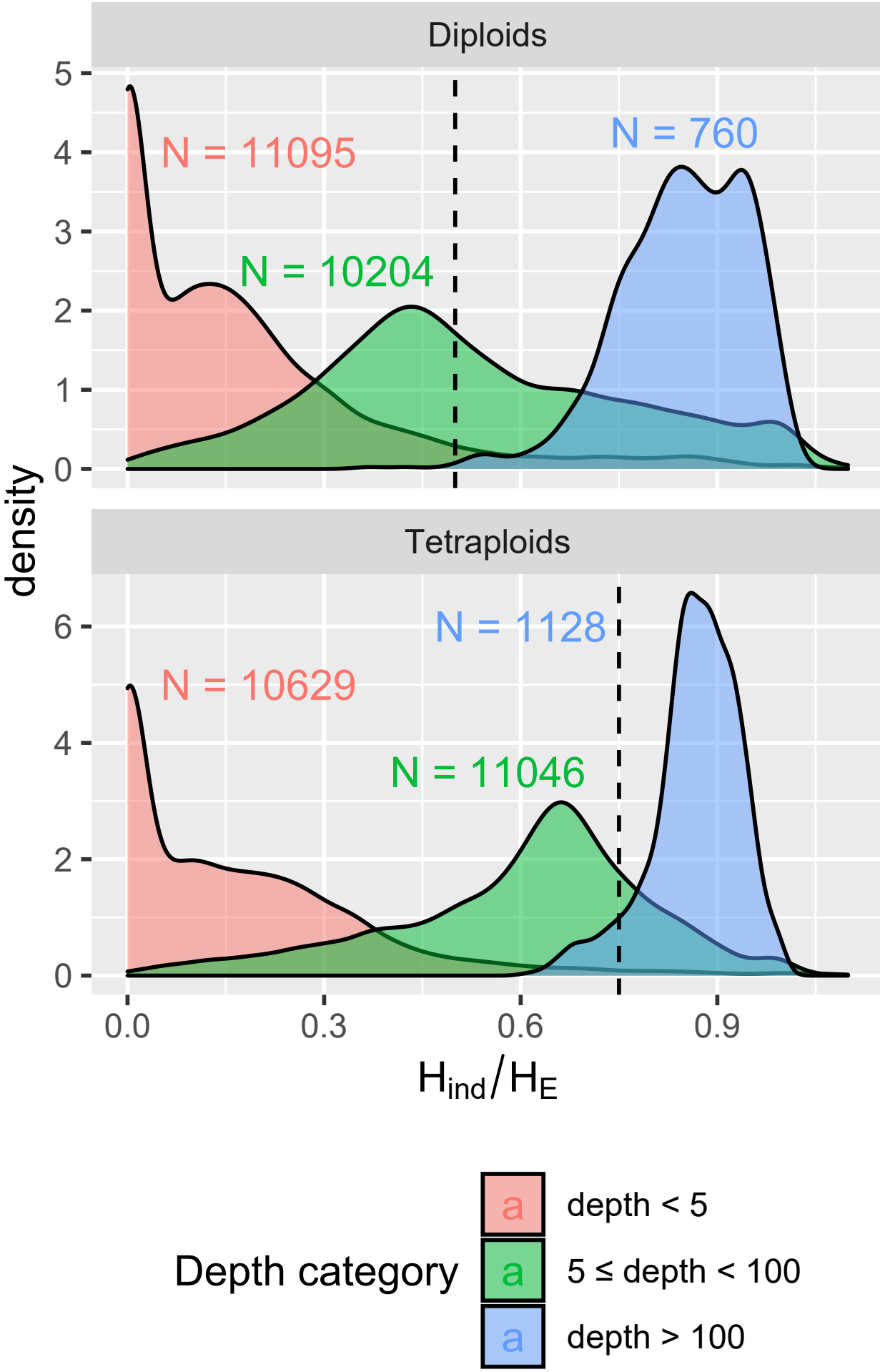
Figure 2. Effect of reference genome and ploidy on H_{ind}/H_E per locus in *Miscanthus sacchariflorus*. Loci with a mean read depth below five were omitted, leaving 11,516 loci aligned to the *M. sinensis* reference and 8,820 loci aligned to the *Sorghum bicolor* reference. Expected values for Mendelian loci under Hardy-Weinberg equilibrium are shown with dashed

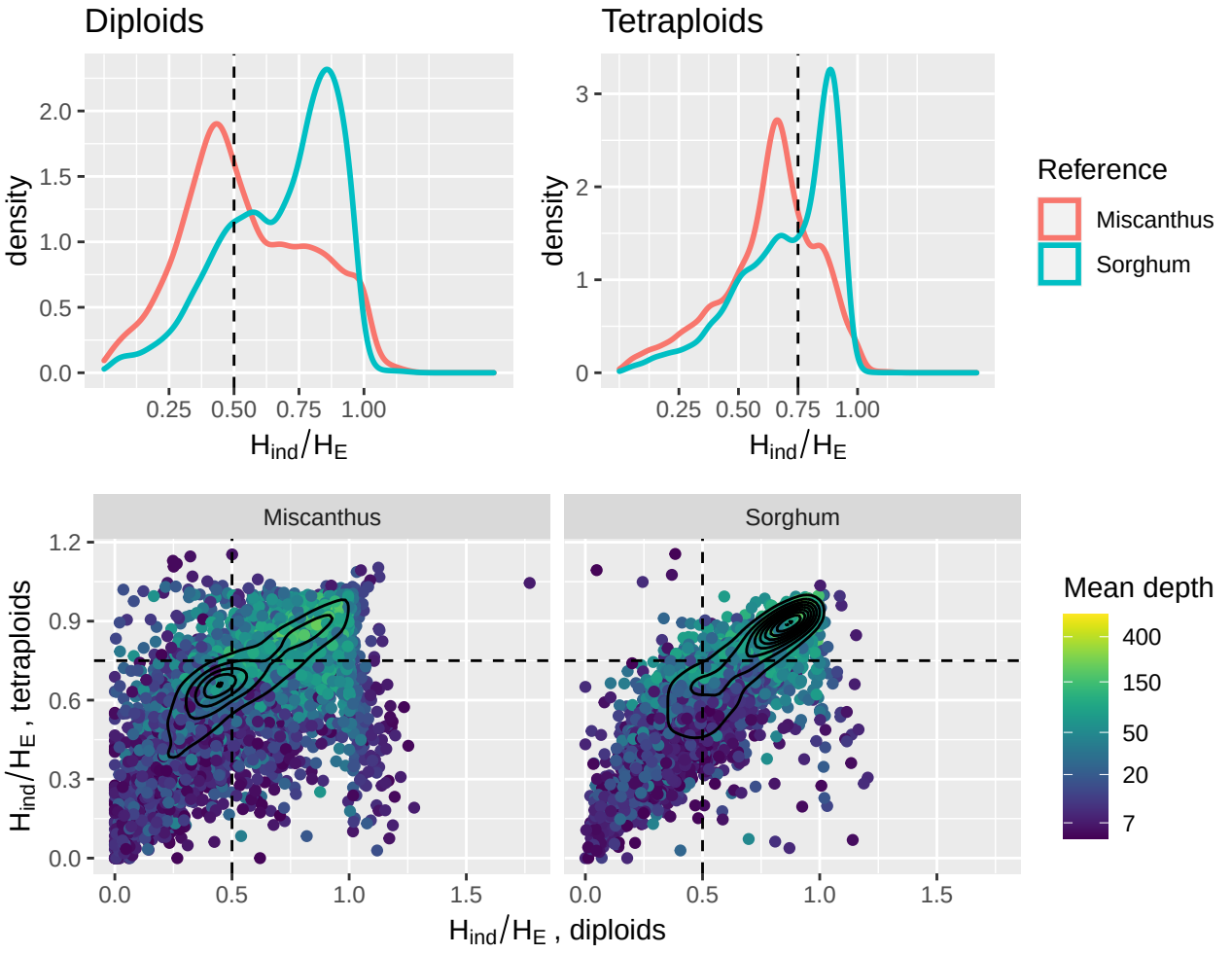
lines. Density plots are shown in the top panel, and scatter plots comparing results in diploids and tetraploids, overlaid with 2D density plots, are shown in the bottom panel.

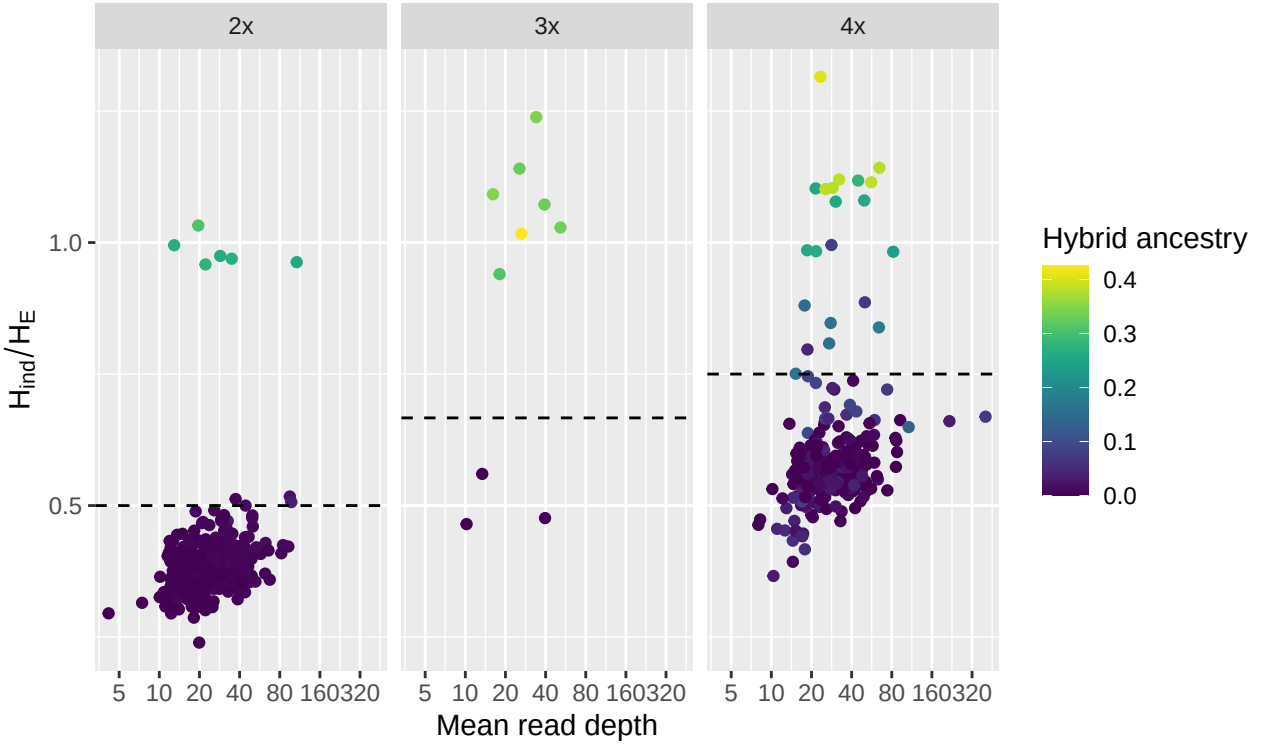
Figure 3. Relationship between ploidy, sequence read depth, hybrid ancestry, and H_{ind}/H_E among 620 *M. sacchariflorus* individuals. Ploidy and proportion *M. sinensis* ancestry (hybrid ancestry) were determined previously [19]. Read depth and H_{ind}/H_E were averaged across 10,000 loci. The expected value for H_{ind}/H_E under Hardy-Weinberg equilibrium is shown with the dashed line.

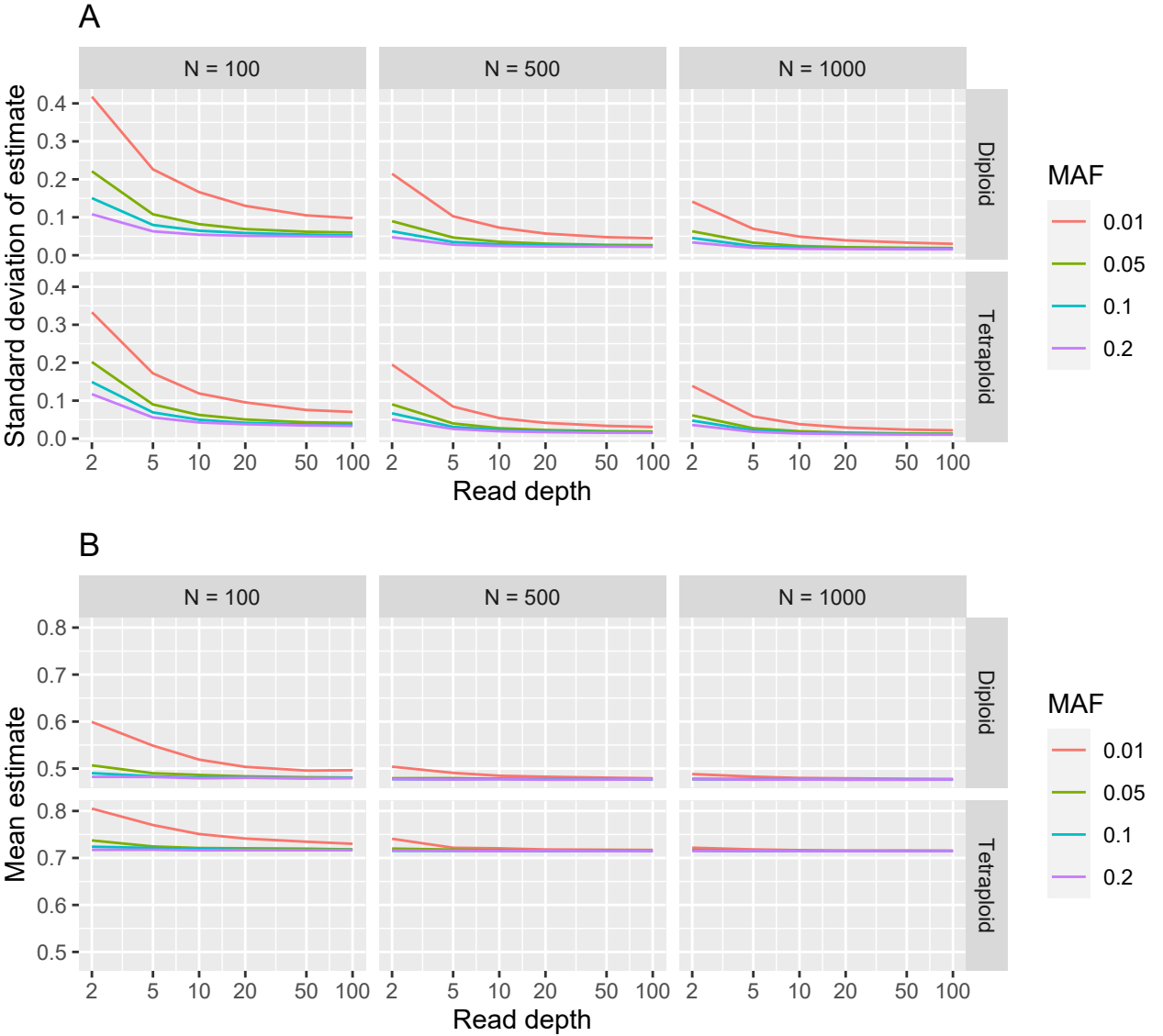
Figure 4. Effect of sample size, read depth, and minor allele frequency on variance and bias of estimates of H_{ind}/H_E. For each combination of ploidy, sample size, read depth, and minor allele frequency (MAF), 5000 biallelic loci were simulated under Hardy-Weinberg Equilibrium with an overdispersion parameter of 20. (A) Standard deviation of H_{ind}/H_E estimates. (B) Mean H_{ind}/H_E estimates. Expected values are 0.5 for diploids and 0.75 for tetraploids; deviations from these values indicate bias in estimation.

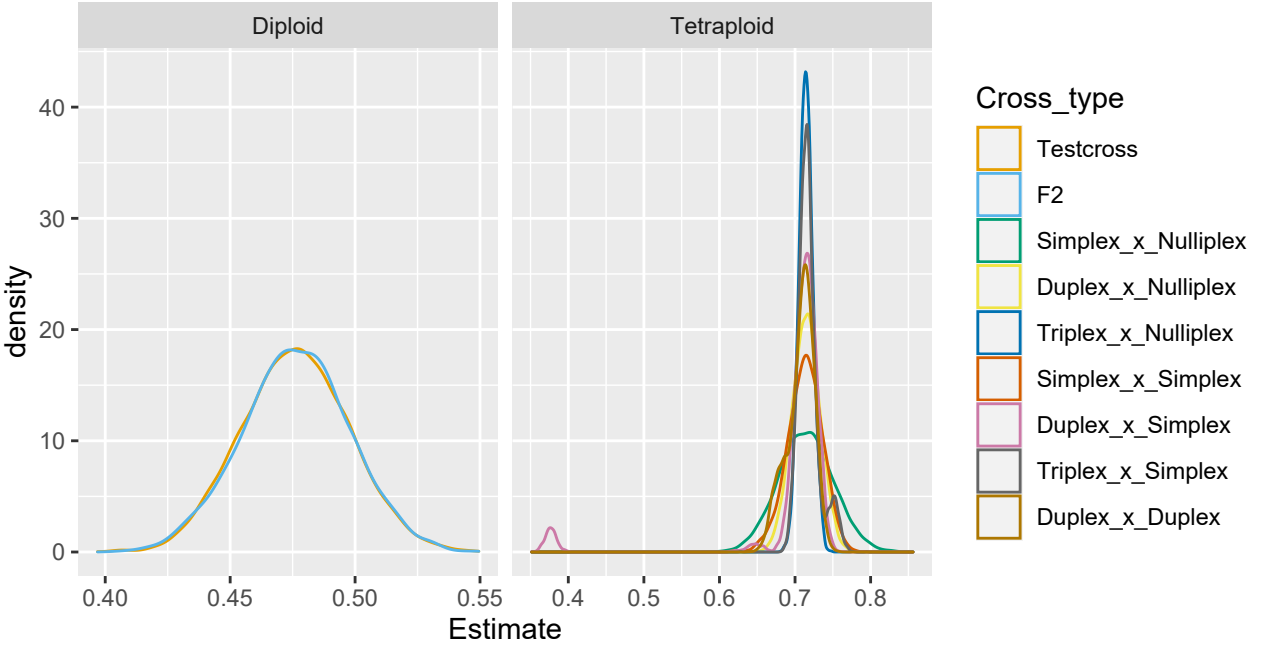
Figure 5. Distribution of H_{ind}/H_E estimates in simulated F1 mapping populations. For each cross type, 5000 biallelic loci with a read depth of 20 and overdispersion parameter of 20 were simulated across 500 individuals.


Diploids
N = 11095
N = 10204
N = 760
Tetraploids
N = 10629
N = 11046
N = 1128
density
0
1
2
3
4
5
0
2
4
6
0.0
0.3
0.6
0.9
H_{ind}/H_E
Depth category
a
depth < 5
a
5 ≤ depth < 100
a
depth > 100


Diploids
Tetraploids
density
2.0
1.5
1.0
0.5
0.0
3
2
1
0
0.25
0.50
0.75
1.00
H_{ind}/H_E
Reference
Miscanthus
Sorghum
Miscanthus
Sorghum
1.2
0.9
0.6
0.3
0.0
H_{ind}/H_E , tetraploids
0.0
0.5
1.0
1.5
H_{ind}/H_E , diploids
Mean depth
400
150
50
20
7


2x
3x
4x
1.0
0.5
H_{ind}/H_E
5 10 20 40 80 160320
Mean read depth
Hybrid ancestry
0.4
0.3
0.2
0.1
0.0


A
N = 100
N = 500
N = 1000
Diploid
Tetraploid
Standard deviation of estimate
0.0
0.1
0.2
0.3
0.4
2
5
10
20
50
100
Read depth
MAF
0.01
0.05
0.1
0.2
B
N = 100
N = 500
N = 1000
Diploid
Tetraploid
Mean estimate
0.5
0.6
0.7
0.8
2
5
10
20
50
100
Read depth
MAF
0.01
0.05
0.1
0.2


Diploid
Tetraploid
density
Estimate
0
10
20
30
40
0.40
0.45
0.50
0.55
0.4
0.5
0.6
0.7
0.8
Cross_type
Testcross
F2
Simplex_x_Nulliplex
Duplex_x_Nulliplex
Triplex_x_Nulliplex
Simplex_x_Simplex
Duplex_x_Simplex
Triplex_x_Simplex
Duplex_x_Duplex