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

Lindsay V. Clark[1*], Wittney Mays[2], 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

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

Abstract

Given the economic and environmental importance of allopolyploids and other species with highly duplicated genomes, there is a need for accurate genotyping methodology that distinguishes paralogs in order to yield Mendelian markers. Methods such as comparing observed and expected heterozygosity are frequently used for identifying 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. 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. We also introduce an algorithm that can choose among multiple alignment locations for a given sequence tag in order to optimize the value of H_{ind}/H_E for each locus, correcting alignment errors that frequently occur in highly duplicated genomes. Our methodology is implemented in polyRAD v1.2, available at https://github.com/lvclark/polyRAD.

Introduction

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. 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. 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 (Dufresne, Stift, Vergilino, & Mable, 2014). This separation of paralogs is especially important in variant calling, because SNPs and indels will not behave in a Mendelian fashion if the reads from which they originate were not all aligned to a single Mendelian locus

(Kaur, Francki, & Forster, 2012). 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 (Dufresne et al., 2014; Meirmans & Van Tienderen, 2013; Nadukkalam Ravindran, Bentzen, Bradbury, & Beiko, 2018; Verdu et al., 2016).

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 (Arruda et al., 2016; Hohenlohe, Amish, Catchen, Allendorf, & Luikart, 2011; Tinker, Bekele, & Hattori, 2016). 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 (Catchen, Amores, Hohenlohe, Cresko, & Postlethwait, 2011), 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. (2012) developed a method, extended by Willis et al. (2017), 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 (Gayral et al., 2013) as well as read depth ratios in combination with observed heterozygosity (McKinney, Waples, Seeb, & Seeb, 2017). Lastly, multiple methods identify putative paralogs based on networks of similarity among sequence tags (Lu et al., 2013; Nadukkalam Ravindran et al., 2018).

Although the methods listed above aim to remove paralogous loci from genetic datasets, it has been observed that stringent filtering of paralogs can result in large amounts of potentially useful data being discarded (Limborg, Seeb, & Seeb, 2016; McKinney, Waples, Pascal, Seeb, & Seeb, 2018; Tinker et al., 2016). To recover paralogous markers without the need for a reference genome, the software Haplotag uses observed heterozygosity to sort tags into their respective paralogous loci, but is only designed to work on highly inbred species (Tinker et al., 2016). In species for which a reference genome has been assembled that includes all homeologs or other duplicated regions, one might perform conventional sequence alignment and variant calling, with the assumption that the alignment software can distinguish paralogs. However, such pipelines only utilize one alignment per sequence tag, which could be incorrect given that alignments are typically selected at random if a tag aligns equally well to multiple locations. One solution is to only retain sequence tags that align uniquely (J. Clevenger, Chavarro, Pearl, Ozias-Akins, &

Jackson, 2015), although this can result in a substantial loss of data, particularly in gene-rich regions that are conserved between homeologs (Kaur et al., 2012). SWEEP overcomes this issue by using a sliding window approach with a reference genome to distinguish true SNPs from differences between paralogs, although its dependency on all paralogs being sequenced in all individuals precludes its use at low read depth (J. P. Clevenger & Ozias-Akins, 2015). PolyCat and PolyDog align sequence tags separately to each subgenome and assume the best alignment to be the correct one (Page & Udall, 2015), but do not examine heterozygosity or other parameters to confirm that loci are behaving in a Mendelian fashion. Alignment of sequences to reference genomes of diploid progenitor species is also possible, but only in recent allopolyploids for which the progenitor species are extant (Kaur et al., 2012)

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 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 or correcting loci before any genotype calling is performed, saving computation time. For species with an available reference genome with separated paralogs, we also present an algorithm that utilizes our statistic to sort sequence tags into their correct alignment locations. We extend our Bayesian genotype calling software polyRAD (Clark, Lipka, & Sacks, 2019) to implement the novel statistic and algorithm.

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 mean 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. The reasoning behind this is H_E that 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 (Simpson, 1949):

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.

Algorithm to correct tag alignment locations via optimization of H_{ind}/H_E

We implemented an algorithm in Python 3 to adjust alignment locations of sequence tags in order to optimize H_{ind}/H_E for each locus in a natural population, with the goal of getting the statistic at or below the expected value. The algorithm assumes that paralogous loci are represented separately in the reference genome, and represent "diverged duplicates" according to the terminology of Limborg et al. (2016), meaning that they do not have any alleles in common. (When paralogous loci do have alleles in common, the algorithm is generally expected to fail to optimize H_{ind}/H_E, and as a result will filter the loci from the dataset.) Where there may be multiple "best" alignments of a tag to a reference genome, or the best alignment may be incorrect, the algorithm uses H_{ind}/H_E to choose the correct alignment location for each tag.

Before running our Python pipeline, the GBSv2 pipeline from TASSEL5 (Bradbury et al., 2007) is used to identify all unique sequence tags in a reduced representation sequencing dataset, as well as their depth in each individual in the dataset, using the *GBSSeqToTagDBPlugin*. The depth matrix is exported to a text file using *GetTagTaxaDistFromDBPlugin*, and unique tag sequences are exported to FASTQ using *TagExportToFastqPlugin*. Tag sequences are then aligned to the reference sequence using Bowtie2 (Langmead & Salzberg, 2012), with the *–k* parameter set to a value higher than the number of expected paralogous loci. For example, in the allohexaploid wheat reference (Clavijo et al., 2017), one would want *–k 4* or higher in order to allow the aligner to identify one alignment location for each of three putative paralogs. If the aligner found four (or more) alignments for a given tag, the tag would be assumed to be from a region that is more highly duplicated than expected, and would be discarded by our algorithm. Otherwise, our algorithm groups tags into putative sets of paralogous loci if they have the same, or overlapping, unordered set of alignment locations. Groups of tags are then filtered out of the analysis if they don't meet a minimum threshold for number of individuals with more than zero reads (100 individuals by default).

Within each group of tags identified in the above step, our algorithm then attempts to assign tags to alignment locations. Tags are initially assigned to the location where they best aligned, as measured by the number of mutations (*NM* values from Bowtie2 output), or a random location if they aligned equally well to more than one. Each location is then treated as a putative Mendelian locus, with the tags assigned to that location being alleles. H_{ind}/H_E is then estimated for each putative Mendelian locus.

If H_{ind}/H_E for any putative locus exceeds the expected value, tags are rearranged among paralogous loci based on negative associations among alleles, similarly to the method described by Clark and Schreier (2017). This step is based on the principle that if two alleles belong to different loci, their copy numbers in the organism will be independent, whereas if two alleles belong to the same locus, more copies of one allele will be associated with fewer copies of the other and vice versa. For every pair of tags *g* and *h* within a group of putative loci, in every

individual a depth ratio r is calculated for both tags, indicating what proportion of reads belonged to that tag, excluding reads from the other tag:

Eqn. 9: $r_{ghm} = \frac{d_{gm}}{\left(\sum_{i=1}^{j} d_{im}\right) - d_{hm}}$

A one-tailed Kendall's Tau test is then used to determine if these read depth ratios are negatively correlated between tags g and h. Tags are initially grouped if the p-value for their association is 0.1 or less, but if any of these groups exceed the expected value for H_{ind}/H_E, the p-value threshold is reduced by a factor of 10 until no groups can be made or no groups exceed the expected value for H_{ind}/H_E. These association groups are then treated as one unit throughout any remaining tag swapping steps, *i.e.* negatively associated tags will always be assigned to the same locus. If any association group is split among the putative loci that were determined by alignment quality, the entire group is moved to the locus where the majority of tags had been previously assigned, or a random locus in the case of a tie.

The putative loci are again evaluated to determine if their H_{ind}/H_E exceeds the expected value. If any locus does exceed the expected value, a tabu search (Glover, 1989) is performed to attempt to find a better arrangement of tags among isoloci. For each iteration of the tabu search, all "neighbors" to the current solution are examined, where a neighbor solution involves one tag (or group of negatively associated tags) being relocated from one locus to another. The neighbor solution at which H_{ind}/H_E across all putative loci exceeds the expected value by the minimum amount, or, in the case of a tie, where the number of mutations between tags and the reference sequence is lowest, is considered the best solution. Twenty-five iterations of the tabu search are performed, with the five most recent best solutions being considered "tabu" to encourage exploration of the solution space. The best solution, again first in terms of H_{ind}/H_E and then in terms of number of mutations from the reference, found across all 25 iterations is output as the optimal solution.

Ideally, after this algorithm has run, all tags for a group of alignment locations have been arranged such that no locus has a H_{ind}/H_E greater than the expected value of $(1 - F)(k - 1)/k$. However, to filter out non-Mendelian loci, allowing for some Mendelian loci to exceed the expected value due to sampling error, loci are discarded from the output dataset if H_{ind}/H_E exceeds

Eqn. 10: $\exp\left(\frac{\log\left(\frac{(1-F)(k-1)}{k}\right)+\log\left(\frac{(1-F)(2k-1)}{2k}\right)}{2}\right)$

i.e. if the putative locus more closely resembles a Mendelian locus from twice the expected ploidy than one from the expected ploidy.

Implementation in polyRAD

Functions for estimating H_{ind}/H_E and $H_{ind}/H_{E.map}$ are available in polyRAD v1.2 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 paralogous loci.

Python scripts for correction of tag alignment are also distributed with polyRAD v1.2, along with a vignette describing their use. The *process_sam_multi.py* script takes a SAM file and *TagTaxaDist* file and outputs sorted and filtered sets of tags and their depths, grouped by sets of alignment locations. The *readProcessSamMulti* function in polyRAD can import data from these files to a *RADdata* object, allowing the user to get a preliminary look at the distribution of H_{ind}/H_E for the sake of filtering samples and estimating inbreeding. The *process_isoloci.py* script can then be used to determine the correct alignment location for each tag, and the output of the script can be imported to polyRAD using the *readProcessIsoloci* function, after which genotype calling can be performed using one of several existing polyRAD functions.

PolyRAD v1.2 is currently available on GitHub, and can be installed using *devtools::install_github("lvclark/polyRAD")*.

Datasets for testing

Previously published RAD-seq data for *Miscanthus sacchariflorus* (Clark, Jin, et al., 2019) were used for testing the H_{ind}/H_E statistic and alignment correction algorithm. Raw sequence reads are available on the NCBI Sequence Read Archive under accessions SRP026347, SRP048207, SRP063572, and SRP087645. 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 (Kim et al., 2012; Ma et al., 2012; Swaminathan et al., 2012). Moreover, some populations of *M. sacchariflorus* display autotetraploidy in addition to this genome duplication (4x = 76) (Clark, Jin, et al., 2019; Clark et al., 2015), 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-pollenated and self-incompatible (Hirayoshi, Nishikawa, & Kato, 1955), meaning that low levels of heterozygosity cannot be used to identify paralogs as they 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, where *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 (Bradbury et al., 2007) to identify unique tag sequences and their depths in all individuals. Tag sequences were then aligned to the *Miscanthus sinensis* v7.1 reference genome (DOE-JGI, http://phytozome.jgi.doe.gov/) and the *Sorghum bicolor* v3.1.1 reference genome (McCormick et al., 2018) using Bowtie 2 (Langmead & Salzberg, 2012). The tag manager feature of TagDigger (Clark & Sacks, 2016) 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.

To test the pipeline for correcting alignment locations, *process_sam_multi.py* was run on the same *M. sacchariflorus* dataset, using a SAM file output by Bowtie 2 allowing multiple alignments to the *M. sinensis* reference genome. Inbreeding (*F*) was estimated at 0.15 using polyRAD. H_{ind}/H_E was estimated on a per-individual basis using the first 10,000 loci. The sets of 356 diploid and 268 tetraploid individuals were then run separately with *process_isoloci.py* to assign tags to alignment locations. A log file was generated to record H_{ind}/H_E before and after sorting, as well as the number of tabu reps needed to find the optimal solution.

Results

Utility of H_{ind}/H_E for detecting collapsed paralogs

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 variance of H_{ind}/H_E increased considerably as read depth decreased, suggesting a mean read depth of five to ten as a reasonable cutoff for making use of the statistic for filtering loci (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 0.5 for diploids and 0.75 for tetraploids (Fig. 2), indicating some inbreeding, likely due to population structure (Clark, Jin, et al., 2019). 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 (Figs. 1 and 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 (Fig. 3). 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.

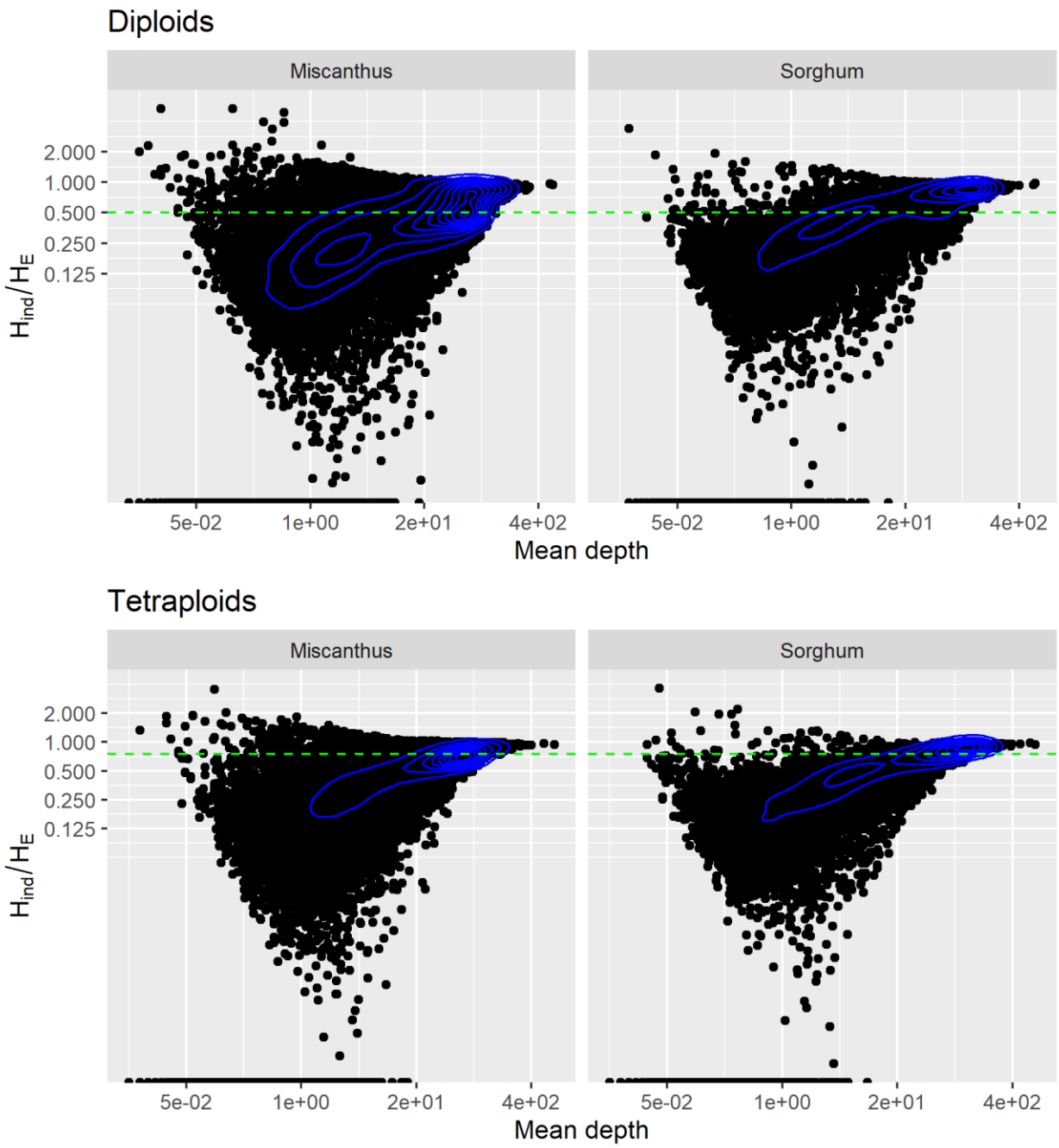

Figure 1. Relationship between H_{ind}/H_E statistic and mean sequence read depth per locus across 356 diploid and 268 tetraploid *Miscanthus sacchariflorus*, with loci called based on alignments to the *M. sinensis* (36,804 loci) or *Sorghum bicolor* (18,402 loci) reference genome. The expected value for a Mendelian locus in Hardy-Weinberg equilibrium is shown with a green dashed line. Blue lines show 2D density of the scatter plot.

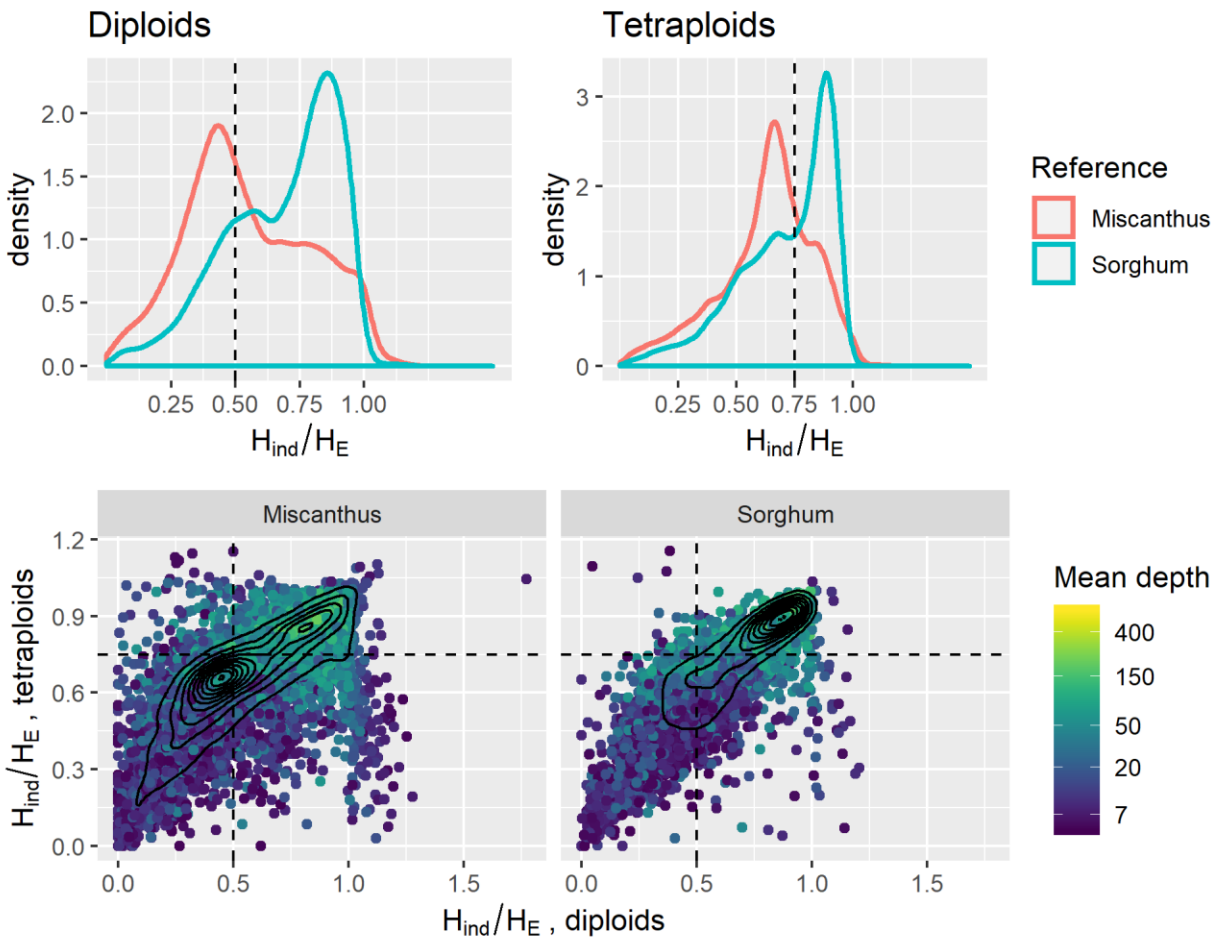

Figure 2. Distribution of H_{ind}/H_E statistic among loci in diploid and tetraploid *Miscanthus sacchariflorus*, using either *M. sinensis* or *S. bicolor* as the reference genome. Loci with a mean read depth below five were omitted, leaving 11,516 loci aligned to the *M. sinensis* reference and 8,820 loci aligned to the *S. 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.

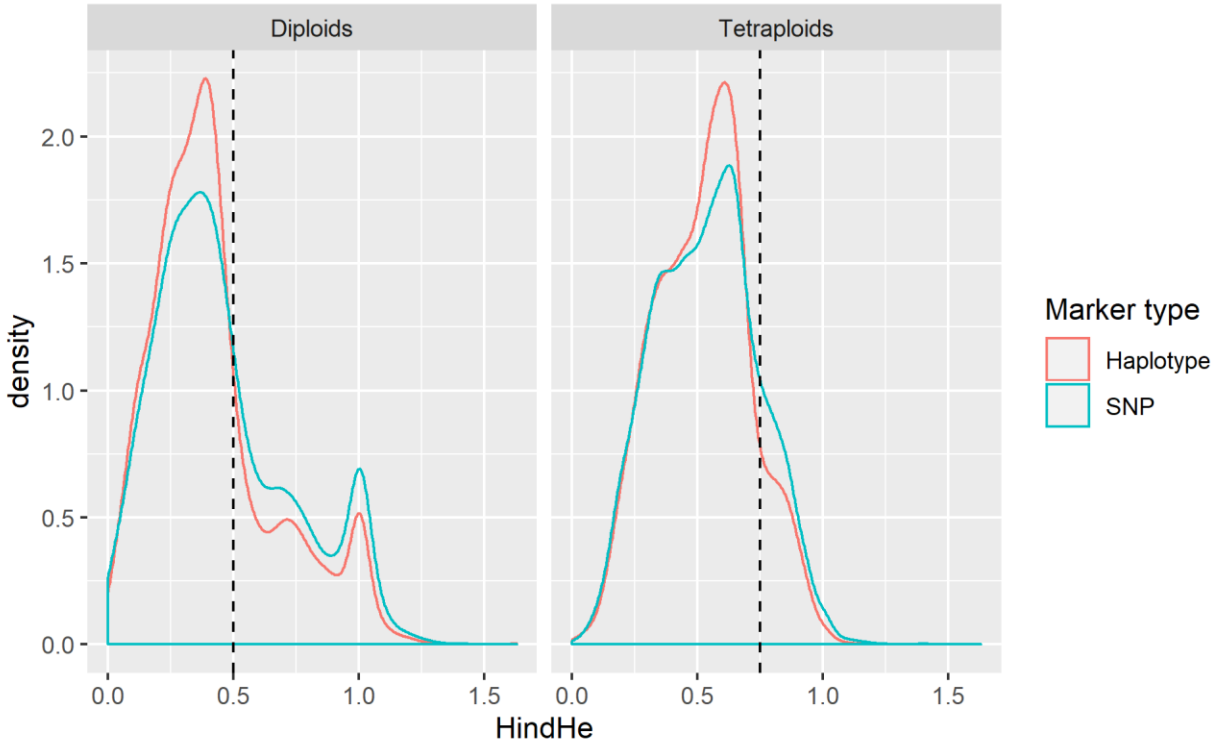

Figure 3. Distribution of H_{ind}/H_E among SNPs and haplotype-based loci in diploid and tetraploid *M. sacchariflorus*. The expected value for Mendelian loci under Hardy-Weinberg equilibrium is shown with the dashed line. The statistic was estimated for 10,458 SNPs, as well as those same SNPs phased into 3710 haplotype-based, multiallelic loci.

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. 4). 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 dataset included many hybrid 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. 4).

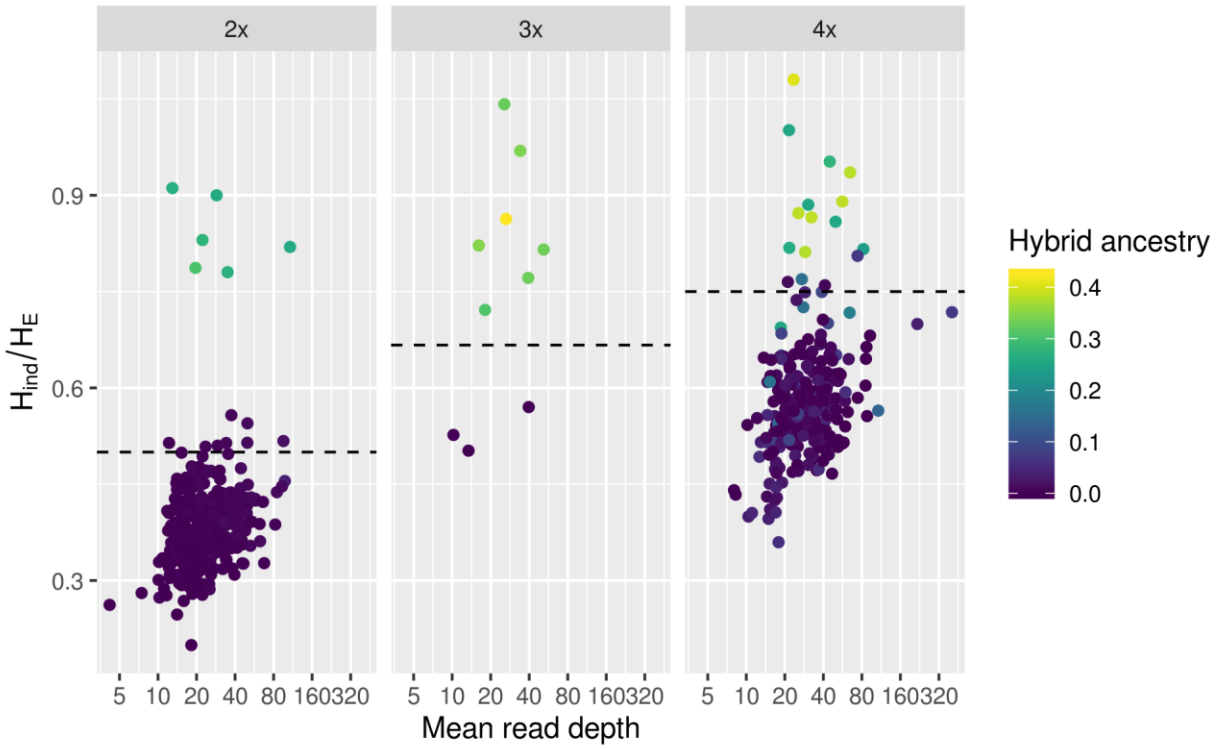

Figure 4. Relationship between ploidy, sequence read depth, hybrid ancestry, and H_{ind}/H_E among 620 *M. sacchariflorus* individuals. Ploidy and proportion *M. sinensis* ancestry (hybrid ancestry) were determined previously (Clark, Jin, et al., 2019). Read depth and H_{ind}/H_E were averaged across 10,000 loci. The expected value for H_{ind}/H_E under Hardy-Weinberg equilibrium is shown with the dashed line.

Ability of the sorting pipeline to optimize H_{ind}/H_E

Using correlations between alleles followed by a tabu search, our algorithm was able to optimize H_{ind}/H_E at or below the expected value for both loci in 51% of all pairs of loci (Fig. 5). In 63% of these cases, rearranging by allele correlations alone was sufficient to find the optimum solution, while for the remainder very few tabu reps were typically necessary (Fig. 5). When the algorithm was not able to lower H_{ind}/H_E to the expected value, more tabu reps tended to be necessary to arrange sequence tags into putative loci with the minimum H_{ind}/H_E (Fig. 5).

Future versions of this preprint will evaluate the accuracy of the sorting pipeline on simulated data, demonstrate improved power for genome-wide association when the pipeline is used to recover loci that would otherwise be discarded, and explore the properties of $H_{ind}/H_{E.map}$ in mapping populations.

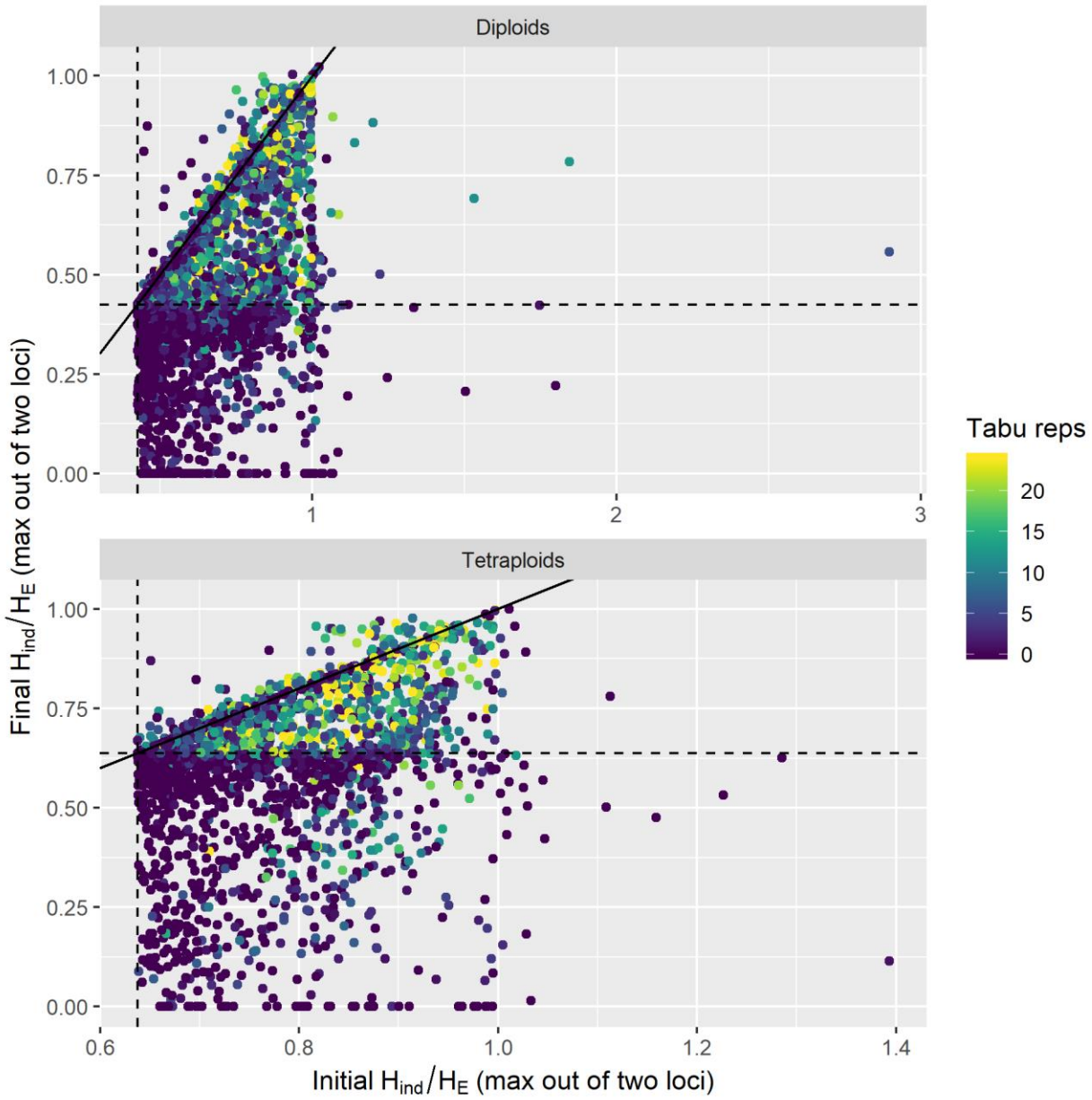

Figure 5. Optimization of H_{ind}/H_E by tabu search. Groups of sequence tags from *M.* sacchariflorus aligning to both chromosomes 1 and 2 in *M. sinensis* were sorted into alignment locations in an attempt to generate loci with H_{ind}/H_E at or below the expected value (dashed lines) given ploidy and inbreeding. Only pairs of loci in which one or more loci exceeded the expected value were processed, including 2870 pairs in the diploid dataset and 2522 pairs in the tetraploid dataset. The x-axis and y-axis show the maximum H_{ind}/H_E for each pair before and after, respectively, any swapping of tags by allele correlation groups and/or tabu search. Color indicates the tabu rep in which the best solution was found. The solid line has an intercept of zero and slope of one; points below the solid line indicate pairs of loci in which the maximum H_{ind}/H_E was reduced by the algorithm. Note that the algorithm seeks to minimize the sum of the amounts by which each of the two loci exceed the expected value.

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. 2, 3, 4). A benefit of this, however, is that as long as ploidy is known, H_{ind}/H_E can be used to estimate inbreeding, either at the population or individual level, directly from sequence read depth. Conversely, 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. 4). 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.

The expected value of H_{ind}/H_E is independent of read depth, number of individuals sampled, and the allele frequency. However, we expect all of these factors to influence the variance of the estimate. 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. In future versions of polyRAD we plan to provide such a simulation function, and in future versions of this preprint we will quantify variance due the above factors. We expect that variance will increase as read depth, sample size, and minor allele frequency decrease. For this reason, the *process_sam_multi.py* script included with polyRAD by default ignores sets of haplotypes for which fewer than 100 samples have reads, and the *readProcessIsoloci* function by default does not import loci with a median read depth below ten; alleles may be assigned inaccurately to such loci due to the high variance in H_{ind}/H_E estimation. Depending on the downstream application, we recommend weighing the number of markers needed versus the importance of marker quality

when determining a median read depth threshold for filtering markers generated by *process_sam_multi.py*.

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 (Fig. 3). 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 (J. P. Clevenger & Ozias-Akins, 2015; Tinker et al., 2016).

Uses of the H_{ind}/H_E statistic

We anticipate that locus filtering to be the most common application of the H_{ind}/H_E statistic, with major advantages being that it is not confounded 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 more 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. The use of observed heterozygosity, read depth ratios within genotypes, and number of haplotypes per individual, however, are redundant with H_{ind}/H_E and unnecessary if it has already been used for filtering. In addition to its use 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.

Because H_{ind}/H_E is an indicator of whether a group of alleles belong to the same Mendelian locus, we developed a variant calling pipeline that chooses among multiple alignment locations for sequence tags in an effort to optimize the value of H_{ind}/H_E. Using this pipeline, loci can be recovered that would otherwise be discarded due to paralogy. Output of our variant calling pipeline can be passed directly to polyRAD for genotype calling, or, because the output contains alignment location and read depth in user-friendly CSV format, to a custom genotype calling pipeline.

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 (Kyriakidou, Tai, Anglin, Ellis, & Strömvik, 2018). 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.

Acknowledgements

We thank Jiale He for testing various methods to detect paralogous loci in *Miscanthus*. This material is based upon work supported by the National Science Foundation under Grant No. 1661490.

References

Arruda, M. P., Brown, P., Brown-Guedira, G., Krill, A. M., Thurber, C., Merrill, K. R., … Kolb, F. L. (2016). Genome-wide association mapping of *Fusarium* head blight resistance in wheat using genotyping-by-sequencing. *Plant Genome*, *9*(1). https://doi.org/10.3835/plantgenome2015.04.0028

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: building and genotyping loci de novo from short-read sequences. *G3*, *1*(3), 171–182. https://doi.org/10.1534/g3.111.000240

Clark, L. V., Lipka, A. E., & Sacks, E. J. (2019). polyRAD: Genotype Calling with Uncertainty from Sequencing Data in Polyploids and Diploids. *G3*, *9*(3), 663–673. https://doi.org/10.1534/g3.118.200913

Clark, L. V., & Schreier, A. D. (2017). Resolving microsatellite genotype ambiguity in populations of allopolyploid and diploidized autopolyploid organisms using negative correlations between allelic variables. *Molecular Ecology Resources*, *17*(5), 1090–1103. https://doi.org/10.1111/1755-0998.12639

Clark, L. V, Jin, X., Petersen, K. K., Anzoua, K. G., Bagmet, L., Chebukin, P., … Sacks, E. J. (2019). 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. *Annals of Botany*, *124*(4), 731–748. https://doi.org/10.1093/aob/mcy161

Clark, L. V, & Sacks, E. J. (2016). TagDigger: user-friendly extraction of read counts from GBS and RAD-seq data. *Source Code for Biology and Medicine*, *11*, 11. https://doi.org/10.1186/s13029-016-0057-7

Clark, L. V, Stewart, J. R., Nishiwaki, A., Toma, Y., Kjeldsen, J. B., Jørgensen, U., … Sacks, E. J. (2015). Genetic structure of *Miscanthus sinensis* and *Miscanthus sacchariflorus* in Japan indicates a gradient of bidirectional but asymmetric introgression. *Journal of Experimental Botany*, *66*(14), 4213–4225. https://doi.org/10.1093/jxb/eru511

Clavijo, B. J., Venturini, L., Schudoma, C., Accinelli, G. G., Kaithakottil, G., Wright, J., … Clark, M. D. (2017). An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for

chromosomal translocations. *Genome Research*, *27*(5), 885–896. https://doi.org/10.1101/gr.217117.116

Clevenger, J., Chavarro, C., Pearl, S. A., Ozias-Akins, P., & Jackson, S. A. (2015). Single nucleotide polymorphism identification in polyploids: A review, example, and recommendations. *Molecular Plant*, *8*(6), 831–846. https://doi.org/10.1016/j.molp.2015.02.002

Clevenger, J. P., & Ozias-Akins, P. (2015). SWEEP: A Tool for Filtering High-Quality SNPs in Polyploid Crops. *G3*, *5*(9), 1797–1803. https://doi.org/10.1534/g3.115.019703

Dufresne, F., Stift, M., Vergilino, R., & Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology*, *23*(1), 40–69. https://doi.org/10.1111/mec.12581

Gayral, P., Melo-Ferreira, J., Glémin, S., Bierne, N., Carneiro, M., Nabholz, B., … Galtier, N. (2013). Reference-Free Population Genomics from Next-Generation Transcriptome Data and the Vertebrate-Invertebrate Gap. *PLoS Genetics*, *9*(4). https://doi.org/10.1371/journal.pgen.1003457

Glover, F. (1989). Tabu Search—Part I. *ORSA Journal on Computing*, *1*(3), 190–206. https://doi.org/10.1287/ijoc.1.3.190

Hirayoshi, I., Nishikawa, K., & Kato, R. (1955). Cytogenetical Studies on forage plants. (IV) Self-incompatibility in *Miscanthus*. *Japanese Journal of Plant Breeding*, *5*(3), 167–170. https://doi.org/10.1270/jsbbs1951.5.167

Hohenlohe, P. a, Amish, S. J., Catchen, J. M., Allendorf, F. W., & Luikart, G. (2011). Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, *11 Suppl 1*, 117–122. https://doi.org/10.1111/j.1755-0998.2010.02967.x

Kaur, S., Francki, M. G., & Forster, J. W. (2012). Identification, characterization and interpretation of single-nucleotide sequence variation in allopolyploid crop species. *Plant Biotechnology Journal*, *10*(2), 125–138. https://doi.org/10.1111/j.1467-7652.2011.00644.x

Kim, C., Zhang, D., Auckland, S. a, Rainville, L. K., Jakob, K., Kronmiller, B., … Paterson, A. H. (2012). SSR-based genetic maps of *Miscanthus sinensis* and *M. sacchariflorus*, and their comparison to sorghum. *Theoretical and Applied Genetics*, *124*(7), 1325–1338. https://doi.org/10.1007/s00122-012-1790-1

Kyriakidou, M., Tai, H. H., Anglin, N. L., Ellis, D., & Strömvik, M. V. (2018). Current strategies of polyploid plant genome sequence assembly. *Frontiers in Plant Science*, *9*, 1660. https://doi.org/10.3389/fpls.2018.01660

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Limborg, M. T., Seeb, L. W., & Seeb, J. E. (2016). Sorting duplicated loci disentangles complexities of polyploid genomes masked by genotyping by sequencing. *Molecular Ecology*, *25*(10), 2117–2129. https://doi.org/10.1111/mec.13601

Lu, F., Lipka, A. E., Glaubitz, J., Elshire, R., Cherney, J. H., Casler, M. D., … Costich, D. E. (2013). Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genetics*, *9*(1), e1003215. https://doi.org/10.1371/journal.pgen.1003215

Ma, X.-F., Jensen, E., Alexandrov, N., Troukhan, M., Zhang, L., Thomas-Jones, S., … Flavell, R. (2012). High resolution genetic mapping by genome sequencing reveals genome duplication and tetraploid genetic structure of the diploid *Miscanthus sinensis*. *PLoS ONE*, *7*(3), e33821. https://doi.org/10.1371/journal.pone.0033821

McCormick, R. F., Truong, S. K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., … Mullet, J. E. (2018). The *Sorghum bicolor* reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant Journal*, *93*(2), 338–354. https://doi.org/10.1111/tpj.13781

McKinney, G. J., Waples, R. K., Pascal, C. E., Seeb, L. W., & Seeb, J. E. (2018). Resolving allele dosage in duplicated loci using genotyping-by-sequencing data: A path forward for population genetic analysis. *Molecular Ecology Resources*, *18*(3), 570–579. https://doi.org/10.1111/1755-0998.12763

McKinney, G. J., Waples, R. K., Seeb, L. W., & Seeb, J. E. (2017). Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources*, *17*(4), 656–669. https://doi.org/10.1111/1755-0998.12613

Meirmans, P. G., & Van Tienderen, P. H. (2013). The effects of inheritance in tetraploids on genetic diversity and population divergence. *Heredity*, *110*(2), 131–137. https://doi.org/10.1038/hdy.2012.80

Nadukkalam Ravindran, P., Bentzen, P., Bradbury, I. R., & Beiko, R. G. (2018). PMERGE: Computational filtering of paralogous sequences from RAD-seq data. *Ecology and Evolution*, *8*(14), 7002–7013. https://doi.org/10.1002/ece3.4219

Page, J. T., & Udall, J. A. (2015). Methods for mapping and categorization of DNA sequence reads from allopolyploid organisms. *BMC Genetics*, *16*(Suppl 2), S4. https://doi.org/10.1186/1471-2156-16-S2-S4

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PloS One*, *7*(5), e37135. https://doi.org/10.1371/journal.pone.0037135

Simpson, E. H. (1949). Measurement of diversity. *Nature*, *163*(1943), 688.

Swaminathan, K., Chae, W. B., Mitros, T., Varala, K., Xie, L., Barling, A., … Moose, S. P. (2012). A framework genetic map for *Miscanthus sinensis* from RNAseq-based markers shows recent tetraploidy. *BMC Genomics*, *13*(1), 142. https://doi.org/10.1186/1471-2164-13-142

Tinker, N. A., Bekele, W. A., & Hattori, J. (2016). Haplotag: Software for Haplotype-Based Genotyping-by-Sequencing Analysis. *G3*, *6*(4), 857–863.

https://doi.org/10.1534/g3.115.024596

Verdu, C. F., Guichoux, E., Quevauvillers, S., De Thier, O., Laizet, Y., Delcamp, A., … Mariette, S. (2016). Dealing with paralogy in RADseq data: in silico detection and single nucleotide polymorphism validation in *Robinia pseudoacacia* L. *Ecology and Evolution*, *6*(20), 7323–7333. https://doi.org/10.1002/ece3.2466

Willis, S. C., Hollenbeck, C. M., Puritz, J. B., Gold, J. R., & Portnoy, D. S. (2017). Haplotyping RAD loci: an efficient method to filter paralogs and account for physical linkage. *Molecular Ecology Resources*, *17*(5), 955–965. https://doi.org/10.1111/1755-0998.12647