## Probability of sampling two different alleles from sequence reads at a given individual and locus under Mendelian assumptions

Say we have a locus with ploidy *k* and number of alleles *j*. The allele frequencies are

where

The expected heterozygosity (i.e. the probability that two alleles randomly sampled from the population will be different) is

If we express the sum of the product of allele frequencies for all [*j* choose 2] pairs of non-identical alleles as

Then expected heterozygosity can also be defined as

For a given genotype, copy numbers of all alleles are defined as

where

and *k* is the ploidy.

Assuming polysomic inheritance and Hardy-Weinberg equilibrium, the frequency of any given genotype is

Assume we have sequenced a given individual at this locus, with infinite sequencing reads. The probability that if we sample two different reads with replacement, they will belong to different alleles is

Then across an infinite population of individuals, each with infinite reads, the average *Hind* is the sum of the product of *Hind* and *F* across all possible genotypes.

Once is expanded algebraically, it can always be factored to

Which can be simplified to

Thus, for any given ploidy and expected heterozygosity, we have an expected average within-individual read depth diversity.

For violations of Hardy-Weinberg equilibrium,

where *F* is the inbreeding coefficient.

Another way to approach this proof is that, if two locus copies are sampled with replacement from a genotype with *k* copies, the probability that the same copy is sampled twice is 1/*k*. Therefore (*k* – 1)/*k* is the probability that two sampled (with replacement) reads from one individual \* locus originate from different copies of the locus. This probability that two sampled reads come from different copies of the locus is multiplied by *HE*, the probability that two sampled copies of the locus correspond to different alleles, to get the probability that two sampled reads from one individual \* locus correspond to different alleles.

In practice the value of will be lower than this expectation due to non-infinite read depth and deviations from Hardy-Weinberg equilibrium. Values above the expectation can be used as an indication that a locus is actually an artefactual combination of two or more paralogous loci.

## Estimation of *Hind* / *HE* from empirical data

Say that we have sequence read depths across a set of alleles in an individual *m*.

Total read depth in one individual is

We can estimate read depth diversity within that individual using the Gini-Simpson index:

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

And expected heterozygosity is estimated as

For finite read depths, when two reads are sampled with replacement, the probability of sampling different reads is (*D* – 1)/*D*. Therefore, *HE* must be reduced by this factor to get a more accurate probability of sampling two different alleles on a per-individual basis.

The expectation from the proof in the previous section is then that

To avoid dividing by zero, individuals with fewer than two reads for a given locus are omitted from the calculation.

## Algorithm to optimize *Hind* / *HE*

We implemented an algorithm in Python 3 to adjust alignment locations of sequence tags in order to optimize *Hind* / *HE* for each locus, with the goal of getting the statistic at or below the expected value. Before running our Python pipeline, the GBSv2 pipeline from TASSEL5 (cite) 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. The depth matrix is exported to a text file using *GetTagTaxaDistFromDBPlugin*. Tag sequences are then aligned to the reference sequence using Bowtie2 (cite), with the *–k* parameter set to a value higher than the number of expected paralogous loci. For example, in the allohexaploid wheat reference (list particular reference and cite), 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 isoloci if they have the same 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 (default 100 individuals).

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 isolocus, with the tags assigned to that location being alleles. *Hind* / *HE* is then estimated for each putative isolocus.

If *Hind* / *HE* for any putative isolocus exceeds the expected value, tags are rearranged among isoloci based on negative associations among alleles, similarly to the method described by Clark and Schreier (2017; cite). 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 isoloci, 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:

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 *Hind* / *HE*, 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 *Hind* / *HE*. 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 isolocus. If any association group is split among the putative isoloci that were determined by alignment quality, the entire group is moved to the isolocus where the majority of tags had been previously assigned, or a random isolocus in the case of a tie.

The putative isoloci are again evaluated to determine if their *Hind* / *HE* exceeds the expected value. If any isolocus does exceed the expected value, a tabu search 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 isolocus to another. The neighbor solution where *Hind* / *HE* across all putative isoloci 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 *Hind* / *HE* 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 isolocus has a *Hind* / *HE* 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, isoloci are discarded from the output dataset if *Hind* / *HE* exceeds

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