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***

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.