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

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, 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.

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 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 *HE* and *Hind***

Say that we have sequence read depths across a set of alleles.

Total read depth is

We can estimate read depth diversity using the Simpson index:

*HE* is estimated using this method, with allelic read depths summed across all individuals in the population. *Hind* is estimated using within-individual depths, then averaged across individuals with non-zero depth in order to estimate .

I am also exploring estimating *Hind*/*HE* on a per-individual basis before averaging across individuals. To correct for low read depth in some individuals, it is estimated as: