# Genetic Diversity

Rodney J. Dyer
Department of Biology
Virginia Commonwealth University
http://dyerlab.bio.vcu.edu

## Synopsis

Genetic diversity is measure of within stratum variance and there are several methods available for the estimation of diversity. In a general sense, we will be using measures of allelic richness from the Baja California data set, which can easily be found my examining the `Frequencies` of the loci.

```
> require(gstudio)
> data(araptus_attenuatus)
> baja <- araptus_attenuatus[araptus_attenuatus$Species != "CladeB",]
> freqs <- allele.frequencies(baja)
> freqs$LTRS
```

```
Allele Frequencies:
  01 = 0.5519878
  02 = 0.4480122
```

```
> freqs$MP20
```

```
Allele Frequencies:
  07 = 0.2892308
  05 = 0.2969231
  15 = 0.001538462
  08 = 0.02769231
  06 = 0.08923077
  04 = 0.009230769
  18 = 0.1784615
  19 = 0.009230769
  17 = 0.04153846
  10 = 0.01384615
  11 = 0.04153846
  16 = 0.001538462
```

In this data set, the raw allelic diversity across all the samples range from 2 - 12 alleles. However, using a base approach such as this falls short for several reasons:

1. We are only looking at the number of alleles across the entire data set and there are many cases where it may be of interest to look at allelic diversity within substrata. It is possible to use the `partition` function along with `allele.frequencies` to get to the number of alleles at partitions but the problem with that is:
2. The raw number of alleles depends upon the number of individuals sampled. It is not statistically sound to compare raw diversity of stratum with different numbers of individuals. This is where *rarefaction* comes in.

3. The sole number of alleles present may not be as important as other measures of genetic diversity such as the diversity of non-rare alleles, or the average 'effective' number of alleles.

To overcome both of these issues, the `genetic.diversity` function is used.

### Rarefaction

Before we get into the nitty-gritty, the basic concept of rarefaction should be examined. Rarefaction is a permutation technique that can be used to standardize samples based upon sample allocation and is an old friend to ecologists.

For our purposes, we will consider rarefaction as a subsampling of alleles in strata standardized by the size of the smallest stratum. So if we have one population with 10 individuals (20 alleles if the locus is diploid) and the rest of the populations have 50 individuals (100 alleles), a rarefied comparison of diversity should be based upon sampling of 20 alleles.

The function `genetic.diversity` takes random samples of the alleles within each population and recomputes the requested allelic diversity statistic. While in many ecological studies, rarefaction is depicted as an accumulation curve (they are generally interested in sampling intensity), `genetic.diversity` only reports the distribution at the largest size where all strata are equal (e.g., the number of alleles present in the smallest population).

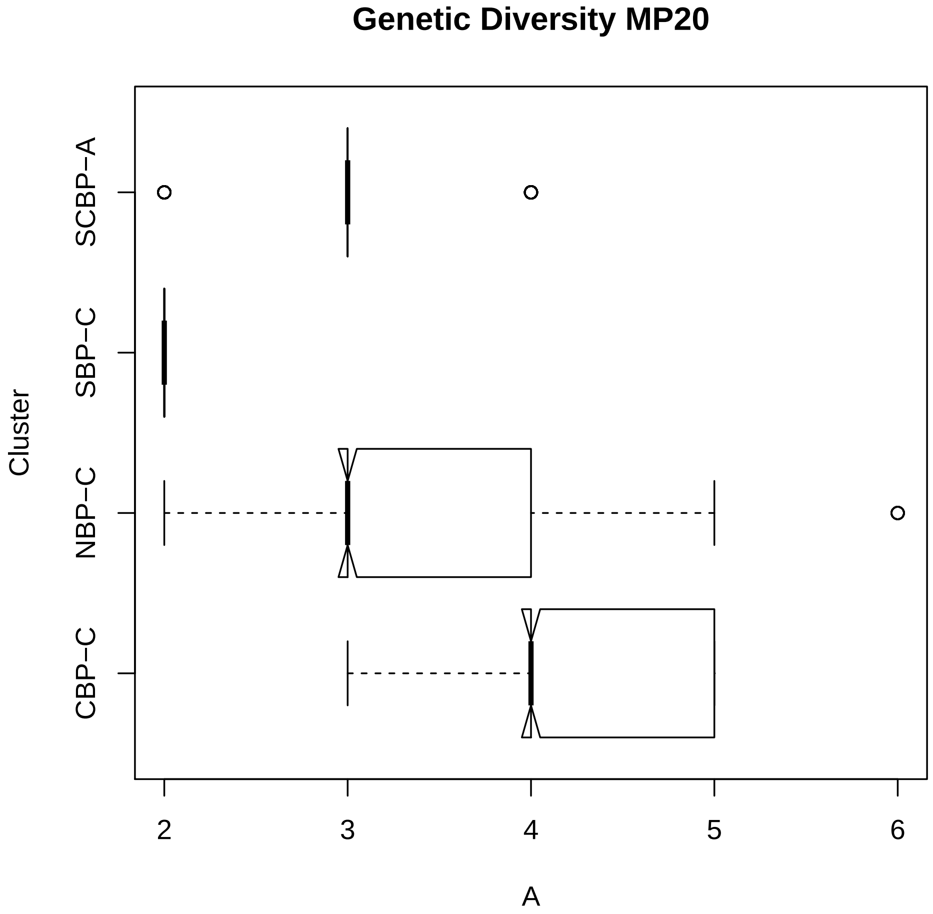
## Allelic Diversity: $A$

The parameter $A$ is solely a measure of the number of alleles at a locus. If a population has a single individual with a single copy of allele $A$ and everyone else has allele $C$, $A = 2$, which is the same case as if half the population was homozygous for $A$ and the remaining individuals were homozygous for $C$. The function `genetic.diversity` returns an object that can be both printed and examined in plot fashion (by default it is a boxplot)

```
> A <- genetic.diversity(baja,stratum="Cluster",loci="MP20",mode="A")
> A
```

```
Geneic Diversity:
  Estimator: A
  Stratum: Cluster
  Loci: { MP20 }
  Locus = MP20
    CBP-C A = 5 ; Rarefaction A = 4.1961961961962
    NBP-C A = 6 ; Rarefaction A = 3.49249249249249
    SBP-C A = 2 ; Rarefaction A = 2
    SCBP-A A = 4 ; Rarefaction A = 2.99299299299299
```

```
> plot(A)
```



The plot itself is a horizontal `boxplot`. If you conduct the analysis with either the `loci` missing or as a list of loci, the results from each locus will be displayed in the terminal and the plotting will cycle through each locus requiring some input from the keyboard. It is also possible to plot just a single locus by passing the locus name as a second parameter to the `plot` command.

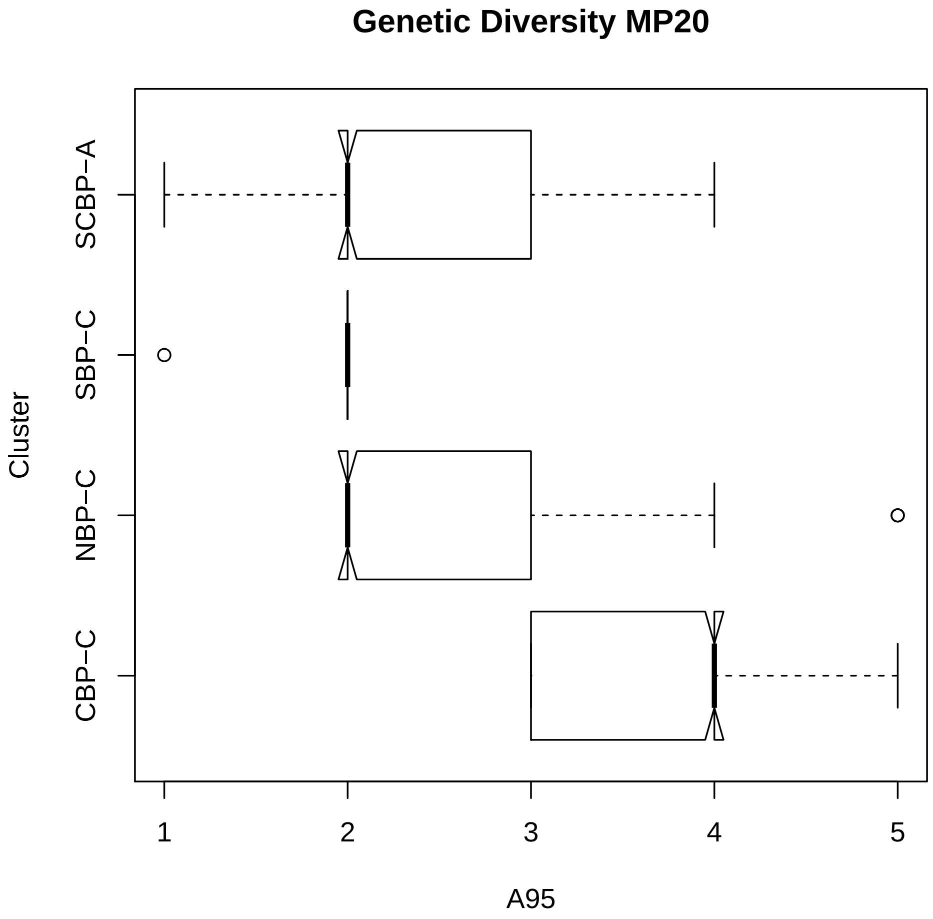
## Allelic Diversity of Non-Rare Alleles: $A_{95}$

The parameter $A_{95}$ ignores rare alleles by not counting those whose frequencies are below 95% *within* the stratum. So alleles locally rare will not be counted and in general $A >= A_{95}$.

```
> A95 <- genetic.diversity(baja,stratum="Cluster",loci="MP20",mode="A95")
> A95
```

```
Geneic Diversity:
  Estimator: A95
  Stratum: Cluster
  Loci: { MP20 }
  Locus = MP20
    CBP-C A95 = 4 ; Rarefaction A95 = 3.67167167167167
    NBP-C A95 = 2 ; Rarefaction A95 = 2.33633633633634
    SBP-C A95 = 2 ; Rarefaction A95 = 1.998998998999
    SCBP-A A95 = 2 ; Rarefaction A95 = 2.4034034034034
```

```
> plot(A95)
```



## Effective Allelic Diversity: $A_e$

The last diversity statistic is $A_e$, which is another frequency corrected allelic diversity statistic. For a locus with $\ell$ alleles, each of which occurs at a frequency of $p_i$, the effective number of alleles is:

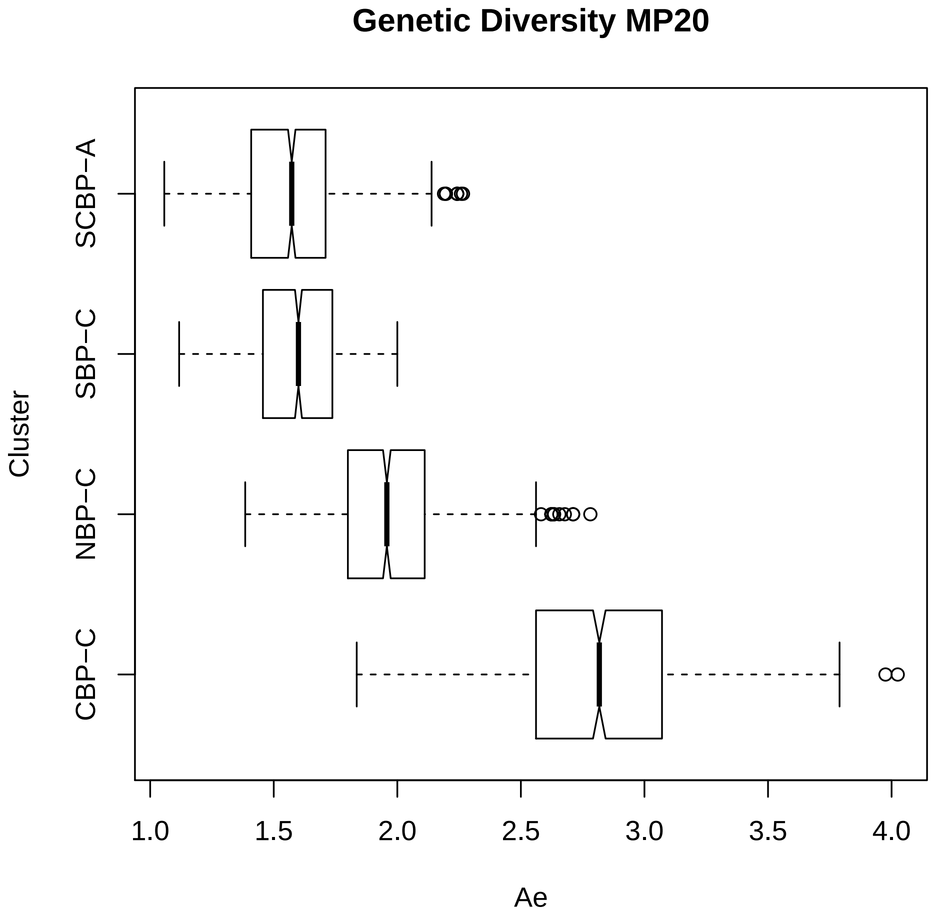
$$A_e = \frac{1}{\sum_{i=1}^{\ell} p_i^2} \tag{1}$$

And for the example data:

```
> Ae <- genetic.diversity(baja,stratum="Cluster",loci="MP20",mode="Ae")
> Ae
```

```
Geneic Diversity:
  Estimator: Ae
  Stratum: Cluster
  Loci: { MP20 }
  Locus = MP20
```

```
    CBP-C Ae = 2.93481610504455 ; Rarefaction Ae = 2.81591355900569
    NBP-C Ae = 1.97536394176932 ; Rarefaction Ae = 1.95956819490084
    SBP-C Ae = 1.6 ; Rarefaction Ae = 1.59153994280315
    SCBP-A Ae = 1.58205596962453 ; Rarefaction Ae = 1.57547483909162
> plot(Ae)
```



One obvious difference in $A_e$ from the others is that it is not an integer value (both $A$ and $A95$ are integers) and as such can show a bit more granularity.