Summarizing Genetic Data

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

Synopsis

There are several ways you can summarize genetic data and here we will cover some simple approaches and introduce another class that aids in the analysis of population genetic data.

The Frequencies Class

The Frequencies class was designed to help out with allele frequency issues and provide a single interface from which you can extract frequency-related information. At its most basic level, a new Frequencies object is created from a list of Locus objects.

```
> require(gstudio)
> loc1 <- Locus( c(1,2) )
> loc2 <- Locus( c(2,2) )
> loc3 <- Locus( c(2,2) )
> freqs <- Frequencies( c( loc1, loc2, loc3) )
> freqs
```

```
Allele Frequencies:
  1 = 0.1666667
  2 = 0.8333333
```

Estimates of allele frequencies can be extracted from the Frequencies class using the get.frequencies method. This method needs to have the object and an optional list of alleles you are interested in getting frequencies for. If you do not pass the second parameter, it will give you the frequencies for all the alleles it currently has. If you do, it will give you the observed frequency of each (notice the value for the '42' allele)

```
> names(freqs)
```

```
[1] "1" "2"
```

```
> length(freqs)
```

```
[1] 2
```

```
> get.frequencies( freqs )
```

```
        1         2
0.1666667 0.8333333
```

```
> get.frequencies( freqs, c("1","42") )
```

```
        1        42
0.1666667 0.0000000
```

Heterozygosities

A fundamental component of many population genetic analysis is the estimation of heterozygosity. There are two basic types of heterozygosity, that which is expected under Hardy-Weinberg Equilibrium and that which was observed. For simplicity, these are denoted as H_e and H_o in many common texts.

Observed heterozygosity is probably the simplest of the two and it is simply the fraction of genotypes in the group you are looking at (could be a population or a region or a site) that are heterozygotes. In terms of the `Locus` class, the function `is.heterozygote` returns `TRUE` if the locus has at least two alleles (allowing for ploidy levels in excess of 2) and at least two different alleles are present. As part of the data accumulation process in the construction of an `AlleleFrequency` object, observed heterozygosity is recorded.

Expected heterozygosity requires an assumption of equilibrium (in the most simple case). For a diploid locus with alleles `A` & `B` and frequencies of each allele denoted as p_A & p_B, genotypes are expected to occur at a frequency of:

$$\begin{array}{rcl} AA & \rightarrow & p_A^2 \\ AB & \rightarrow & 2 * p_A * p_B \\ BB & \rightarrow & p_B^2 \end{array}$$

From the example set of loci we used above, the observed and expected frequencies are:

```
> ho( freqs )
```

```
       ho
0.3333333
```

```
> he( freqs )
```

```
       he
0.2777778
```

Allele Frequencies

The estimation of allele frequencies for a single site or population is probably one of the least informative summary approaches available. It is the differences among sites & populations and the various evolutionary and demographic processes that create these differences that are often of interest.

There are several helper functions and methods that can be used to examine allele frequencies across strata.

Getting Frequencies from Populations

The `Population` class has a method for returning an `AlleleFrequency` object for a particular locus. This is mostly a convenience method that goes through all the `Indiviudal` objects in the `Population` and creates a new `AlleleFrequency` object for you. As a single population you can grab it using the `allele.frequencies` routine.

```
> data(araptus_attenuatus)
> araptus.ltrs.freq <- allele.frequencies(araptus_attenuatus,"LTRS")
> araptus.ltrs.freq
```

```
$LTRS
Allele Frequencies:
  01 = 0.523416
  02 = 0.476584
```

If you do not pass `get.frequencies` the optional `loci` parameter, it will return a `list` of `Frequency` objects for all loci.

```
> all.freqs <- allele.frequencies(araptus_attenuatus)
> print(all.freqs[1:2])
```

```
$LTRS
Allele Frequencies:
  01 = 0.523416
  02 = 0.476584


$WNT
Allele Frequencies:
  01 = 0.3579545
  03 = 0.4303977
  04 = 0.02698864
  02 = 0.1818182
  05 = 0.002840909
```

With the `partition` method, you can take the entire data set and easily find allele frequencies for subsets of data.

```
> clades <- partition(araptus_attenuatus,"Species")
> names(clades)
```

```
[1] "CladeA" "CladeB" "CladeC"
```

```
> cladeC.freqs <- allele.frequencies(clades$CladeC)
> summary(cladeC.freqs)
```

```
     Length Class       Mode
LTRS  2     Frequencies S4
WNT   4     Frequencies S4
EN    5     Frequencies S4
EF    2     Frequencies S4
ZMP   2     Frequencies S4
AML  10     Frequencies S4
ATPS  6     Frequencies S4
MP20  8     Frequencies S4
```

```
> summary(cladeC.freqs$AML)
```

```
Class : Frequencies
N : 252
A : { 01, 02, 05, 06, 07, 08, 09, 10, 11, 13 }
ho : 0.4677419
he : 0.7284242
```

```
> get.frequencies(cladeC.freqs$AML, 11)
```

```
         11
0.002016129
```

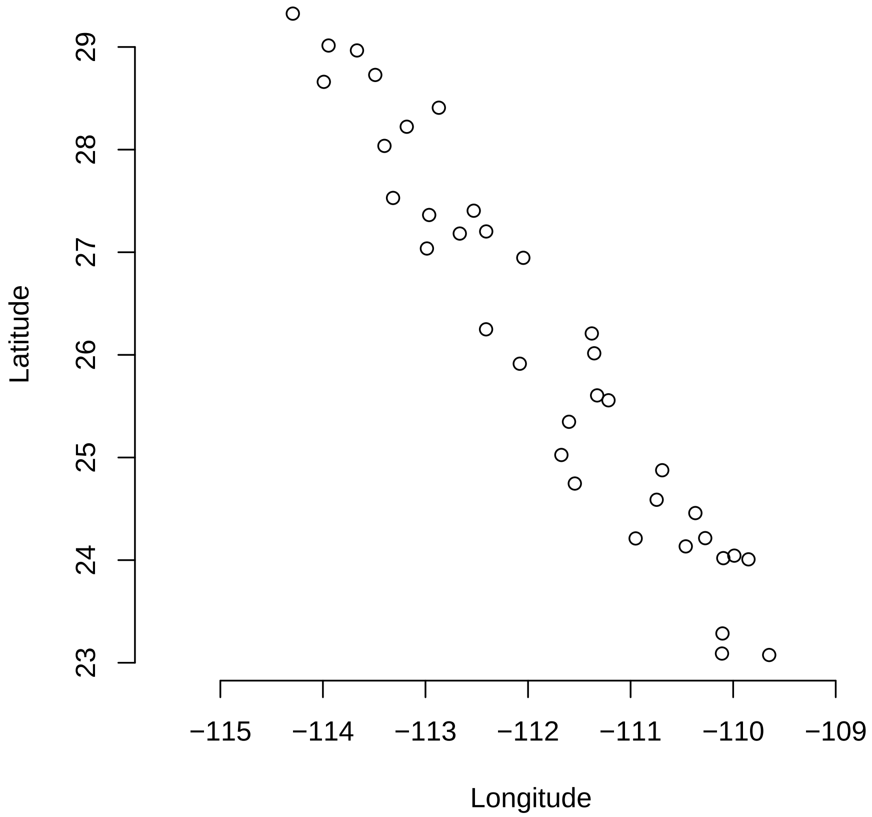
```
> allele.frequencies( araptus_attenuatus[ araptus_attenuatus$Lat > 26.3 ,], loci="AML" )
```

```
$AML
Allele Frequencies:
  08 = 0.308642
  09 = 0.2592593
  07 = 0.2407407
  10 = 0.02469136
  06 = 0.03703704
  11 = 0.08333333
  02 = 0.00308642
  13 = 0.00308642
  05 = 0.00308642
  01 = 0.00308642
  12 = 0.03395062
```

Plotting Frequencies

The combination of Population and Frequencies can easily be used to explore population structure. In the next snippet, we partition the dataset into populations along the Baja Peninsula and plot their locations (n.b., the bty option to plot removes the box around the image and the asp makes the axes equal).

```
> baja <- araptus_attenuatus[araptus_attenuatus$Species!="CladeB",]
> pop.coords <- unique( cbind( baja$Long, baja$Lat ) )
> plot(pop.coords, bty="n", xlab="Longitude", ylab="Latitude",asp=1)
```



Next, we can adjust the size of the symbol by diversity at any locus (below `LTRS` is used). Here the `lapply` function is used to apply a function to the elements of the `baja.pops` list. If you are not familiar with this function, you should look it up. The resulting heterozyosity estimates are scaled and used as symbol size (via `cex`; Figure 1).

```
> baja.pops <- partition( baja, "Pop" )
> pop.he <- lapply( baja.pops, function(x) he( Frequencies( x$LTRS ) ) )
> summary( unlist(pop.he) )
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.1800  0.2036  0.3457  0.4800
```

```
> plot(pop.coords, bty="n", xlab="Longitude", ylab="Latitude",asp=1,cex=2*unlist(pop.he)+1, main="Hetero
```

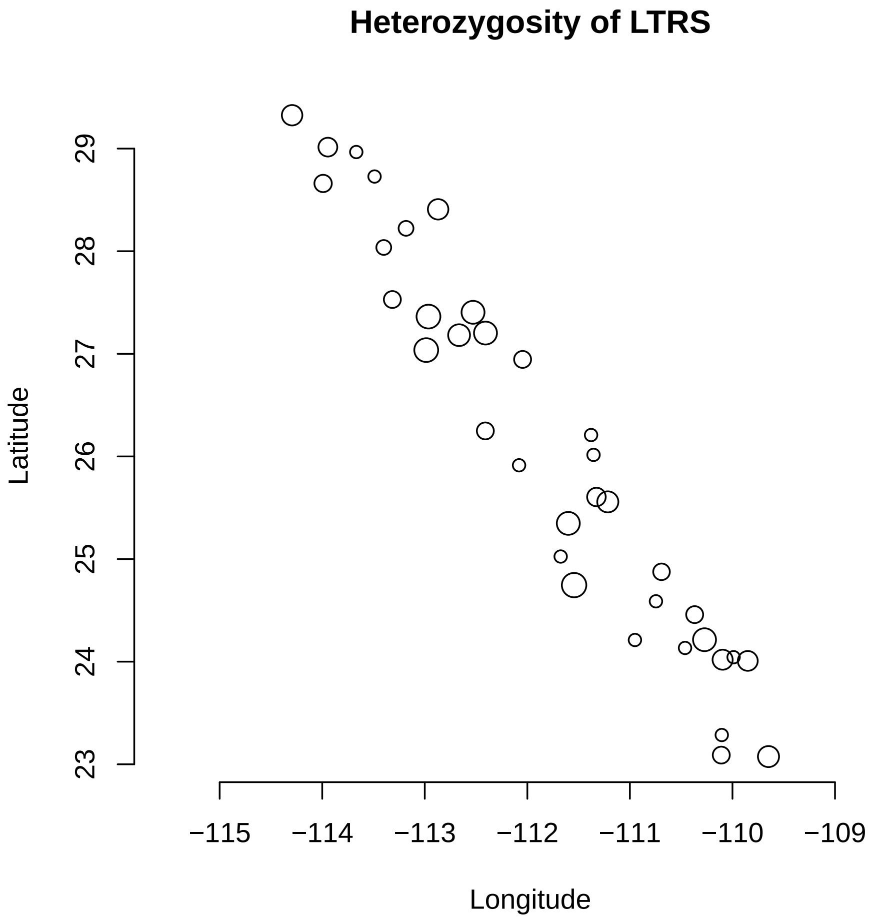


Figure 1: Heterozygosity of *Araptus attenuatus* populations (depicted by symbol size) on the peninsula of Baja California.