Genetic Structure

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

Synopsis

Estimation of genetic structure is a fundamental process in population genetic analyses. Broadly defined, structure can be defined as the non-random association of genotypes and alleles in populations due to evolutionary processes such as gene flow, drift, selection, and inbreeding. For this, the *Araptus attenuatus* data set and will be used again.

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

Genotype Frequencies

The manner by which alleles are arranged into genotypes tells us a lot about the history of a species. The structure statistic that are presented below all rely upon estimation of genotype frequencies so a brief digression to talk about genotype frequencies is in order.

Under a model of random mating, a locus with ℓ alleles whose frequencies are denoted by $p_1, p_2, \ldots, p_\ell$, homozygotes for the i^{th} allele are expected to occur at a frequency of p_i^2 and ij-heterozygotes are expected at $2p_ip_j$.

The expected frequencies are estimated from the allele frequencies assuming Hardy-Weinberg Equilibrium. If you were only interested in the proportion of heterozygotes, you can use the `ho` and `he` functions.

```
> freq.ltrs <- allele.frequencies(baja, "LTRS")
> he(freq.ltrs$LTRS)*length(baja$LTRS)
```

```
      he
161.7324
```

```
> ho(freq.ltrs$LTRS)*length(baja$LTRS)
```

```
ho
69
```

However, at times, it is of interest to look at all genotypes. If you use the `as.character` method for `Locus` objects, you can easily tabulate the counts of each genotypic state[1].

```
> obs <- genotype.counts( araptus_attenuatus, "LTRS")
> obs
```

[1] This function does take into consideration the non-sorting nature of the `Locus` object so that a `3:4` locus and a `4:3` locus will be counted as the same heterozygote.

```
01:01 01:02 02:02
  147    86   130
```

```
> obs/sum(obs)
```

```
    01:01     01:02     02:02
0.4049587 0.2369146 0.3581267
```

Below they are denoted as a matrix, the values on the diagonal of `exp` are the expected number of homozygotes and off-diagonal estimates are the expected frequency of heterozygotes.

```
> p <- get.frequencies( freq.ltrs$LTRS )
> p
```

```
       01        02
0.5519878 0.4480122
```

```
> exp.freq <- p %*% t(p)
> row.names(exp.freq) <- colnames(exp.freq)
> exp <- exp.freq * length(baja$LTRS)
> exp
```

```
         01       02
01 99.63379 80.86621
02 80.86621 65.63379
```

As you can see there are fewer heterozygotes than expected ($N_{hets;\,exp} = 162$, $N_{hets;\,obs} = 86$).

Hardy-Weinberg Equilibrium

While the `gstudio` package provides the basic units for population genetic analyses, there are already some very good packages that conduct analyses like testing for Hardy-Weinberg Equilbrium[2].

```
> require(HardyWeinberg)
> ltrs.genoytpes <- genotype.counts( araptus_attenuatus, "LTRS")
> HWChisq(ltrs.genoytpes,verbose=T)
```

```
Chi-square test with continuity correction for Hardy-Weinberg equilibrium
Chi2 =  98.52808 p-value =  0 D =  -47.55096
$chisq
[1] 98.52808

$pval
[1] 0

$D
    01:02
-47.55096

$p
   01:01
0.523416
```

[2]There are many other functional packages on cran.r-project.org and you should always make sure someone hasn't already solved a problem for you before you try to code up a solution.

Structure Parameters

Population structure parameters are fundamental tools for population genetics and have been perhaps, the most poorly understood and misused as well. At the end of this section, some examples of the differences between the parameters is given.

These structure parameters are estimated using the function `genetic.structure` and requires a `Population` object, a `stratum`, the `loci` you want to estimate parameters from, and a `mode` (the parameter you want). If you leave off the `loci` parameter, all loci will be used. There is also an optional parameter, `num.perm` that is used to test significance.

Finally, of note here is that all these parameters use a sample-size corrected estimates of heterozygosity.

$$\hat{H}_S = \frac{2\mu}{2\mu - 1} H_S$$
$$\hat{H}_T = H_T + \frac{\hat{H}_S}{2k\mu}$$

Where μ is the harmonic mean strata size and k is the number of stratum. As you can see as μ gets larger $\hat{H}_S \rightarrow H_S$, which translates to "if you have more samples, you can get a better estimate of the average heterozygosity" and as k get larger, $\hat{H}_T \rightarrow H_T$ which says the same thing about the number of populations. The take-home here is that you need many samples from many places.

The G_{ST} Parameter

The parameter G_{ST} is an estimate of the reduction in heterozygosity due to individuals being in different populations. It is functionally equivalent to F_{ST} from Wright and as he points out, it is not a measure of differentiation in the way that we think of differentiation. Rather it is a measure of the extent to which populations have gone to fixation. It is estimated as:

$$G_{ST} = 1 - \frac{\hat{H}_S}{\hat{H}_T}$$

where H_S is the average expected heterozygosity at each stratum $[1 - \sum_{i=1}^{\ell} p_i^2]/K$ and H_T is the expected heterozygosity across the entire dataset.

For the *EN* locus in the Baja California dataset, G_{ST} is estimated by:

```
> gst.baja <- genetic.structure(baja,stratum="Pop",loci="EN",mode="Gst",num.perm=999)
> print(gst.baja)

Geneic Structure Analysis:
  Estimator: Gst
  Stratum: Pop
  Loci: { EN }
   - EN ;  Gst = 0.345786963051191 ; P = 0.001
```

The G'_{ST} Parameter

The parameter G'_{ST} was introduced by Hedrick (20XX) in response to the observation that the parameter G_{ST} is not insensitive to the number of alleles at a locus. Fixing this is done by standardizing the estimate

of G_{ST} by the maximal is can be given the number of alleles present, essential a restandardization to the [0,1] range. This is done by:

$$G'_{ST} = \frac{G_{ST}(k - 1 + \mu)}{(k-1)(1 - \hat{H}_S)}$$

For the same locus, we get a larger

```
> gst.prime.baja <- genetic.structure(baja,stratum="Pop","EN",mode="Gst.prime",num.perm=999)
> print(gst.prime.baja)

Geneic Structure Analysis:
  Estimator: Gst.prime
  Stratum: Pop
  Loci: { EN }
   - EN ;  Gst.prime = 0.459618108931204 ; P = 0.001
```

The D_{EST} Parameter

It has been pointed out that even with the corrections for large numbers of alleles, G'_{ST} may not be acting like a statistic of "differentiation" in the way that we think of differentiation. For example consider the following code where I make three populations, the first one fixed for the "1" allele and the next fixed for the "2" (sure this is an extreme point, but Wright originally made it and it should be repeated).

```
> locus1 <- list()
> for(i in 1:50)
+         locus1[i] <- Locus( c(1,1) )
> for(i in 51:150)
+         locus1[i] <- Locus( c(2,2) )
> strata <- c(rep("Pop-A",50), rep("Pop-B",50), rep("Pop-C",50) )
> pop <- Population(strata=strata, loci=locus1)
> summary(pop)

    strata           loci
 Length:150         1:1: 50
 Class :character   2:2:100
 Mode  :character
```

When we estimate either G_{ST} or G_{ST} on these data we get:

```
> genetic.structure(pop,"strata","loci",mode="Gst")

Geneic Structure Analysis:
  Estimator: Gst
  Stratum: strata
  Loci: { loci }
   - loci ;  Gst = 1

> genetic.structure(pop,"strata","loci",mode="Gst.prime")

Geneic Structure Analysis:
  Estimator: Gst.prime
  Stratum: strata
  Loci: { loci }
   - loci ;  Gst.prime = 1
```

Now, intuitively, if it were just "Pop-A" and "Pop-B" then this would make sense but look at the differences between "Pop-B" and "Pop-C", this should be $G_{ST} = G'_{ST} = 0$! In fact, if you had only one population fixed

for the "1" allele and a thousand populations fixed for the other, these parameters would still equal unity. This is because, as Wright originally pointed out, these population parameters are not meant to measure differentiation but fixation. The parameter D_{est} was introduced by Joost (20XX) to address this issue (n.b., Gregorious proposed this back in the 80's but was not taken serious about it then, perhaps Joost can have better luck).

The parameter is defined as:

$$D_{est} = \frac{k-1}{k} \frac{\hat{H}_T - \hat{H}_S}{1 - \hat{H}_S}$$

For the contrived data set, it gives:

```
> genetic.structure(pop,"strata","loci","Dest")

Geneic Structure Analysis:
  Estimator: Dest
  Stratum: strata
  Loci: { loci }
   - loci ;  Dest = 0.296296296296296
```

Which is what would be expected, roughly a third, if we have two populations that are identical and one that is differentiated from the rest. I recommend looking at the several papers that go over these issues for more clarity.

For completeness, the results of the Baja California data set, under D_{est} are:

```
> Dest.baja <- genetic.structure(baja,stratum="Pop","EN",mode="Dest",num.perm=999)
> print(Dest.baja)

Geneic Structure Analysis:
  Estimator: Dest
  Stratum: Pop
  Loci: { EN }
   - EN ;  Dest = 0.164464814867075 ; P = 0.001
```

Pairwise Structure

The `genetic.structure` function can also be used to estimate pairwise estimates of each parameter using the optional `pairwise` flag.

```
> sonora <- araptus_attenuatus[araptus_attenuatus$Species=="CladeB",]
> genetic.structure(sonora,"Pop",loci="EN",mode="Gst.prime", pairwise=TRUE)

          101       32       102
101 0.0000000 0.4136727 0.3661894
32  0.4136727 0.0000000 0.1245850
102 0.3661894 0.1245850 0.0000000
```