MMOD vignette

David Winter
david.winter@gmail.com

April 4, 2012

Contents

1 Why use mmod (or what's wrong with G_{ST}?)

Population geneticists, molecular ecologists and evolutionary biologists often want to be able to determine the degree to which populations are divided into smaller sub-populations. One very widely used approach to this question uses "F statistics" (measures based on Wrigtht's F_{ST}) to compare diversity within and between predefined sub-populations. Until recently, the most widely used of these statistics has been Nei's G_{ST}. Unfortunately, it has become increasingly clear that the value of G_{ST} is a at least partially dependent on the number of alleles at each locus and the number of populations sampled - making simple interpretation of this statistics difficult.

A number of "F_{ST} analogues" have been developed that compensate for these short comings, and give values that can be compared between studies. MMOD is a package that allows three of these statistics, G''_{ST}, D_{est} and φ'_{ST}, to be calculated from `genind` objects (the standard representation of genetic datasets in the `adegenet` library)

2 Which statistic should I use?

With the proliferation of F_{ST} analogues, it can be hard to decide on the most appropriate measure to use for your study. I encourage you to read Meirmans and Hedrick (2011 doi:10.1111/j.1755-0998.2010.02927.x), which includes a discussion on this topic. As you'll see in the demonstration below, the corrected statistics often tell a similar story. Interestingly, G''_{ST} can be directly related to the rate of migration between populations while D_{est} and φ'_{ST} are about partitioning distances or diversity between genes. You may consider which approach is most appropriate for the specific questions you wish to ask.

3 What statistics can't mmod calculate

There are at least two population genetic statistics related to the ones discussed above that MMOD can't calculate. R_{ST} was developed for micorsattelite data, and takes the relationship between alleles (and therefore the mutation rate) into account. It is not clear how the maximum potential value of R_{ST} for a given dataset can be calculated, so it is not possible to correct this statistic in a way similar to G''_{ST} and φ'_{ST}.

Similarly, the calculation of the maximum value of Weir and Cockerham's θ is complex (and not yet published). If you wish to calculate a corrected version of this statistic you can use RecodeData (http://www.bentleydrummer.nl/software/software/) to create a dataset in which all between-population differences are maximised.

You can then calculate θ for each dataset using `Fst` from the package `pegas`. The corrected statical is simply $\frac{\theta}{\theta_{\max}}$

4 An Example - differentiation in the nancycats data

With the description out of the way, let's see how these functions work in practice. As an example, we are going to examine the `nancycats` data that comes with `adegenet`. This dataset contains microsattelite genotypes taken from feral cats in Nancy, France. So let's start.

```
> library(mmod)
> data(nancycats)
> nancycats
```

```
   ####################
   ### Genind object ###
   ####################
- genotypes of individuals -

S4 class:  genind
@call: genind(tab = truenames(nancycats)$tab, pop = truenames(nancycats)$pop)

@tab:  237 x 108 matrix of genotypes

@ind.names: vector of  237 individual names
@loc.names: vector of  9 locus names
@loc.nall: number of alleles per locus
@loc.fac: locus factor for the  108 columns of @tab
@all.names: list of  9 components yielding allele names for each locus
@ploidy:  2
@type:  codom

Optionnal contents:
@pop:  factor giving the population of each individual
@pop.names:  factor giving the population of each individual

@other: a list containing: xy
```

The nancycats data comes in `adegenet`'s default class for genotypic data, the `genind` class. The functions in `mmod` work on genind objects, so you would usually start by reading in your data using `read.genepop`

Now that we have our data on hand, our goal is to see

- Whether this population is substantially differentiated into smaller sub-populations
- Whether such differentiation can be explained by the geographical distance between sub-populations.

We can look at several statistics to ask answer the first question by using the `diff_stats()` function:

```
> diff_stats(nancycats)
```

```
$per.locus
             Hs        Ht         Gst Gprime_st          D
fca8  0.7708277 0.8614311 0.10517782 0.4810570 0.42006021
fca23 0.7415102 0.7992621 0.07225650 0.2924881 0.23738411
fca43 0.7416796 0.7935120 0.06532017 0.2645865 0.21319208
fca45 0.7273320 0.7641204 0.04814486 0.1845960 0.14335289
fca77 0.7766369 0.8655618 0.10273670 0.4822798 0.42300076
fca78 0.6316202 0.6772045 0.06731245 0.1899390 0.13147655
fca90 0.7369587 0.8141591 0.09482221 0.3770880 0.31183460
fca96 0.6699736 0.7654561 0.12473941 0.3937947 0.30740024
fca37 0.5623259 0.6024354 0.06657894 0.1574662 0.09737005

$global
        Hs         Ht    Gst_est  Gprime_st      D_het     D_mean
0.70654052 0.77146027 0.08415178 0.29942062 0.23504860 0.20017978
```

OK, so what is that telling us? The first table has statistics calculated individually for each locus in the dataset. `Hs` and `Ht` are estimates of the heterozygosity expected for this population with and without the sub-populations defined in the `nancycats` data respectively. We need to use those to calculate the measures of population divergence so we might as well display them at the same time. `Gst` is the standard (Nei) G_{ST}, `Gprime_st` is Hedrick's G''_{ST} and `D` is Jost's D_{est}. Because all of these statistics are estimated from estimators of H_{S} and H_{T}, it's possible to get negative values for each of these differentiation measures. Populations can't be negatively differentiated, so you should think of these as estimates of a number close to zero (it's up to you and your reviewers to decide if you report the negative numbers of just zeros).

D_{est} is the easiest statistic to interpret, as you expect to find $D = 0$ for populations with no differentiation and $D = 1$ for completely differentiated populations. As you can see, different loci give quite different estimates of divergence but they range from $\sim$0.1–0.4.

MMOD can calculate another statistic of differentiation called φ'_{ST}. This statistic is based on the Analysis of Molecular Variance (AMOVA) method, which partitions the variance in genetic distances in a dataset to between-population and within-population components (it is possible to use this framework to partition variance using more than two levels of population structure, but that has not

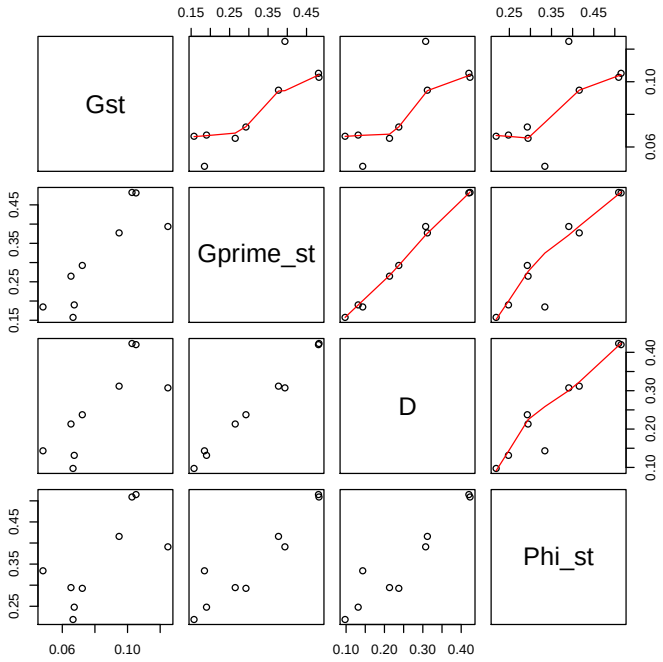

Figure 1: Comparison of differentiation measures

been implemented in MMOD yet). Because φ'_{ST} can take some time to calculate it's not included in `diff_stat` by default (but you can include it using `diff_stat(x, phi_st=TRUE)`).

You might want to see how all these different measures compare to each other across the loci we've looked at. You can see the corrected measures (all those) other than G_{ST} show a similar pattern, and G_{ST} is a bit strange (Figure˜1):

```
> nc.diff_stats <- diff_stats(nancycats, phi_st=TRUE)
> with(nc.diff_stats, pairs(per.locus[,3:6], upper.panel=panel.smooth))
```

The second part of the list returned by `diff_stat` contains global estimates of each of these statistics. For G_{ST} and G''_{ST} these are based on the average of `Hs` and `Ht` across loci. For D_{est} you get two, the harmonic mean of the D_{est} for each locus and, because that method won't work if you end up with negative estimates of D_{est}, one calculated as per G_{ST} and G''_{ST}.

You probably want to have some idea of how robust this result is. MMOD has a few functions for performing bootstrap samples of `genind` objects and calculating statistics from those samples. Because some of these functions can take a long time to run, we will create a very small (10 repetition) bootstrap

sample of the `nancycats` data, then calculate D_{est} from that sample:

```
> bs <- chao_bootstrap(nancycats, nreps=10)
> bs.D <- summarise_bootsrap(bs, D_Jost)
> bs.D

Estimates for each locus
Locus        Mean         95% CI
fca8        0.4094        (0.3413-0.4567)
fca23        0.2615        (0.2225-0.293)
fca43        0.2556        (0.1848-0.3054)
fca45        0.1867        (0.1407-0.253)
fca77        0.4466        (0.4077-0.48)
fca78        0.1525        (0.1176-0.1914)
fca90        0.3334        (0.2699-0.382)
fca96        0.3057        (0.2482-0.3437)
fca37        0.1092        (0.0751-0.1391)

Global Estimate based on average heterozygosity
0.2539        (0.2422-0.2666)

Global Estimate based on harmonic mean of statistic
0.2218        (0.204-0.2349)
```

As you can see, printing a summarised bootstrap sample gives us shows a basic overview of that data, but there is also quite a lot more there — use `str(bs.D)` to check it out. I don't think there is much point trying to interpret confidence intervals estimated from 10 samples, but the point estimates seem to show a population with some substantial differentiation.

Next, we want to know if geography can explain that differentiation. The nancycats data comes with coordinates for each populations. We can use these to get Euclidean distances:

```
> head(nancycats@other$xy, 4)

           x         y
P01 263.3498 171.10939
P02 183.5028 122.40790
P03 391.1050 254.70148
P04 458.6121  41.72336
```

```
> nc.pop_dists <- dist(nancycats@other$xy, method="euclidean")
```

`mmod` provides functions to calculate pairwise versions of each of the differentiation statistics. Because we want to perform a Mantel test, we'll use the "linearized" version of D_{est}, which is just $x/(1-x)$ (each of the pairwise stats has and argument to return this version).

```
> nc.pw_D <- pairwise_D(nancycats, linearized=TRUE)
```

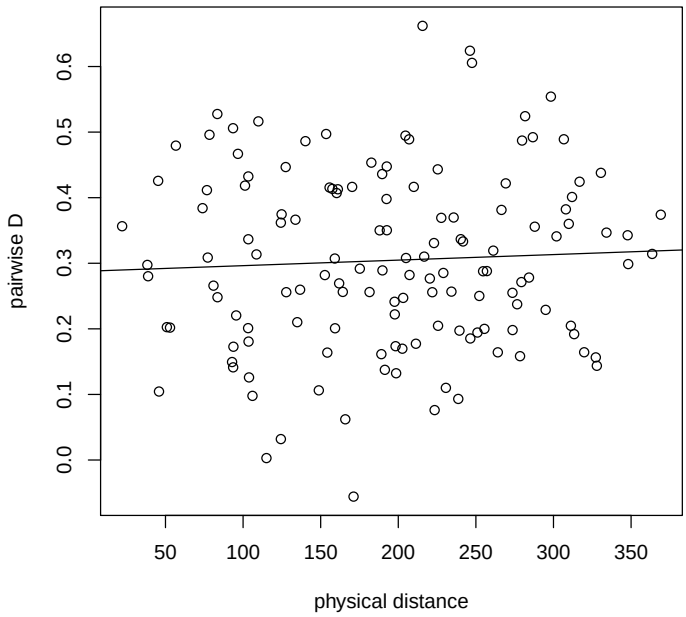

Figure 2: Geographic distance does not explain genetic differentiation

The library `ade4`, which is loaded with `mmod`, provides functions to perform Mantel tests on distance matrices.

```
> mantel.rtest(nc.pw_D, log(nc.pop_dists), 999)

Monte-Carlo test
Observation: 0.03194095
Call: mantelnoneuclid(m1 = m1, m2 = m2, nrepet = nrepet)
Based on 999 replicates
Simulated p-value: 0.376
```

So, the geographic distance between these populations can't explain the genetic divergences we see: the correlation is small and non-significant. If you like, we can also visualize this relationship (Figure˜2).

```
> fit <- lm(as.vector(nc.pw_D) ~ as.vector(nc.pop_dists))
> plot(as.vector(nc.pop_dists), as.vector(nc.pw_D),
+      ylab="pairwise D", xlab="physical distance")
> abline(fit)
```

There are a couple of other functions that are not used here, and a few of use the functions we have used have help messages that guide interpretation of their ressults - use `help(package="mmod")` to see the full documentation.