MMOD vignette

David Winter
david.winter@gmail.com

November 4, 2024

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 analogues" (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, 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. This makes simple interpretations of G_{ST} difficult, and comparisons between studies (or even between loci in the same population) potentially misleading.

A number of new 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 Which statistics can MMOD not 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 when measuring between-allele distances. 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`. If the statistic calculated form the recoded data is θ_{max} then the corrected statistic is simply $\frac{\theta}{\theta_{\mathrm{max}}}$.

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 //////////

 // 237 individuals; 9 loci; 108 alleles; size: 150.5 Kb

 // Basic content
   @tab:  237 x 108 matrix of allele counts
   @loc.n.all: number of alleles per locus (range: 8-18)
   @loc.fac: locus factor for the 108 columns of @tab
   @all.names: list of allele names for each locus
   @ploidy: ploidy of each individual  (range: 2-2)
   @type:  codom
   @call: genind(tab = truenames(nancycats)$tab, pop = truenames(nancycats)$pop)

 // Optional content
   @pop: population of each individual (group size range: 9-23)
   @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.genpop or read.fstat depending on the format it's in.

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  0.7740044 0.8616180 0.10168493 0.4750445 0.41190817
```

```
fca23 0.7415102 0.7992621 0.07225650 0.2956688 0.23738411
fca43 0.7416796 0.7935120 0.06532017 0.2675766 0.21319208
fca45 0.7085554 0.7642248 0.07284422 0.2653163 0.20374594
fca77 0.7766369 0.8655618 0.10273670 0.4855829 0.42300076
fca78 0.6316202 0.6772045 0.06731245 0.1933327 0.13147655
fca90 0.7369587 0.8141591 0.09482221 0.3807578 0.31183460
fca96 0.6725600 0.7656083 0.12153507 0.3913924 0.30192942
fca37 0.5623259 0.6024354 0.06657894 0.1609576 0.09737005


$global
         Hs         Ht    Gst_est  Gprime_st      D_het     D_mean
 0.70509459 0.77150953 0.08608441 0.30848948 0.23928310 0.20931242
```

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 ∼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 among-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 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
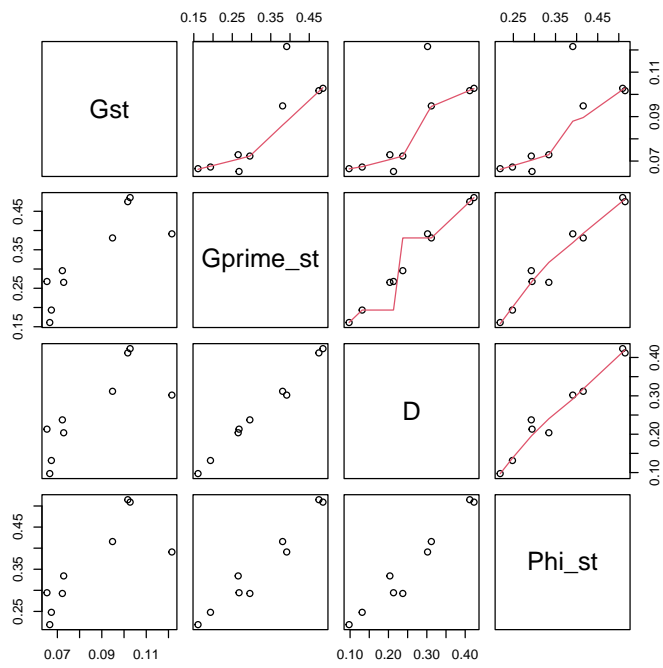
Figure 1: Comparison of differentiation measures

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}. The global estimate of φ'_{ST} is calculated from the average distance among individuals across all loci.

Now that we have a point-estimate for how differentiated these populations are we will 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_bootstrap(bs, D_Jost)
> bs.D

Estimates for each locus
Locus         Mean          95% CI
fca8          0.4119        (0.327-0.497)
fca23         0.2374        (0.147-0.328)
fca43         0.2132        (0.173-0.253)
fca45         0.2037        (0.137-0.271)
fca77         0.4230        (0.356-0.490)
fca78         0.1315        (0.069-0.194)
fca90         0.3118        (0.251-0.372)
fca96         0.3019        (0.247-0.357)
fca37         0.0974        (0.063-0.131)


Global Estimate based on average heterozygosity
0.2393        (0.217-0.262)


Global Estimate based on harmonic mean of statistic
0.2093        (0.183-0.236)
```

As you can see, printing a summarised bootstrap sample gives us shows a basic overview of that data. In this case the confidence intervals are calculated using the "normal method", which it to say the the intervals are the observed value statistic +/- 1.96 x the standard error of the boostrap sample. There is more to these objects than gets printed — 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 population. 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 differen-
tiation 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)
```

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
Call: mantelnoneuclid(m1 = m1, m2 = m2, nrepet = nrepet)


Observation: -0.02584796


Based on 999 replicates
Simulated p-value: 0.633
Alternative hypothesis: greater

     Std.Obs  Expectation     Variance
-0.332466390  0.004175431  0.008155001
```

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
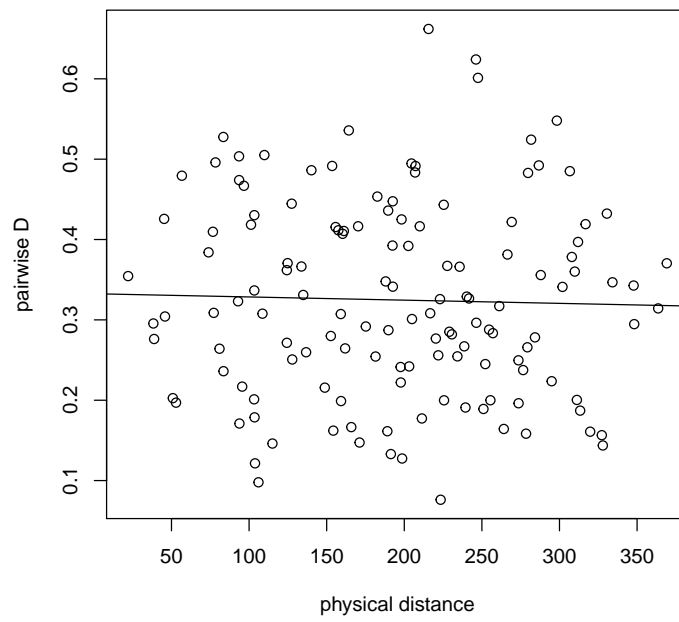ressults - use help(package="mmod") to see the full documentation.

Figure 2: Geographic distance does not explain genetic differentiation