MMOD vignette

David Winter
david.winter@gmail.com

January 27, 2012

This is a short demo of `mmod` library for differentiation statistics in R. 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 explain 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's all that then? The first table has statistics calculated individually for each locus in the dataset. H_S and H_T are estimates of the heterozgosity 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 decided to 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.

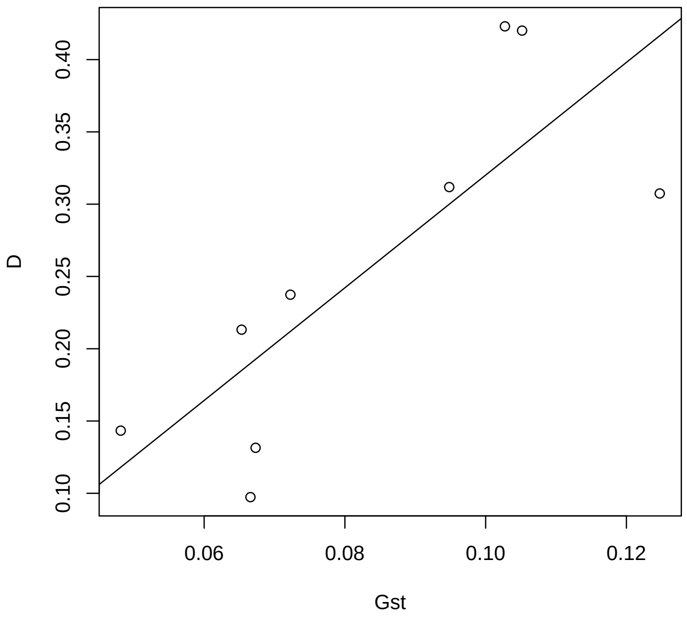

Figure 1: Nei's Gst against Jost's D

You might also want to see how regular old G_{ST} compares with with D (Figure 1):

```
> nc.diff_stats <- diff_stats(nancycats)
> with(nc.diff_stats, plot(per.locus[, "Gst"], per.locus[, "D"],
+     xlab = "Gst", ylab = "D"))
> fit <- with(nc.diff_stats, (lm(per.locus[, "D"] ~ per.locus[,
+     "Gst"])))
> abline(fit)
```

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 $\mathtt{H}_S$ and $\mathtt{H}_T$ 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, let's create a whopping great 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        0.2892        (0.1042-0.4769)
fca23        0.2701        (0.141-0.4196)
fca43        0.2825        (0.131-0.4802)
fca45        0.2567        (0.1099-0.4374)
fca77        0.288        (0.1121-0.4586)
fca78        0.266        (0.1442-0.4545)
fca90        0.2863        (0.1315-0.4501)
fca96        0.2778        (0.1104-0.4736)
fca37        0.2711        (0.1284-0.4901)

Global Estimate based on average heterozygosity
0.2581        (0.247-0.2677)

Global Estimate based on harmonic mean of statistic
0.2257        (0.1987-0.241)
```

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))` or the classic bs.D$ [tab tab] 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, 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
Observation: 0.03194095
Call: mantelnoneuclid(m1 = m1, m2 = m2, nrepet = nrepet)
Based on 999 replicates
Simulated p-value: 0.38
```

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 for each statistic (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, use help(package="mmod") to read about them.

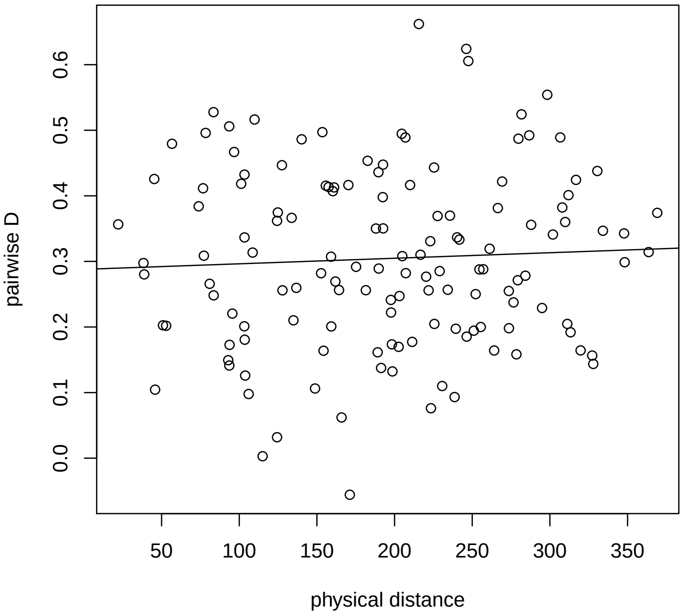

Figure 2: Geographic distance does not explain genetic differentiation