

Genetic Distance

The math free version

Distance, a familiar word

- Geographic Distance- Essentially a way of describing how different two physical locations are.
- There are many different ways to measure distance
 - As the crow flies
 - Mileage in a car
 - Driving time

Genetic Distance

- Same idea, except now we're talking about how different some nucleic acids are...
- Could be individuals, populations, species
- Similarly, there are tons of ways to express these differences, each with its own merits

Dissimilarity is the key

- “Different” is a binary statistic, either 1 or 0, true or false, yes or no
- Either or statistics are not very informative
- Dissimilarity (or Similarity) is vastly more informative, comes in shades of grey!
 - Close to zero, not so dissimilar
 - Close to one, not much in common

Inputs

- Any genetic data will do
 - Allele frequencies, Genotype frequencies, Allele sizes
 - Allozymes, SNPs, RAPDs, AFLPs, mtDNA
 - And everyones favorite- microsatellites
- The more loci the merrier
- The more alleles per locus the merrier
- The more individuals the merrier

User beware!

- It is possible to have too much power
- Statistically significant does not equal BIOLOGICALLY significant!!!
- Especially with m-sats
 - Loci galore
 - Alleles galore
 - Cheap enough for individuals galore (assuming you don't have to develop your own of course...)

Assumptions

- Markers must be Neutral
 - Selection can lead to over (directional selection) or underestimation (stabilizing selection)
 - Test for this looking for deviations from HWE
- Markers must be Unlinked
- Avoid Null alleles
 - Leads to overestimation of homozygotes

Models of Evolution

Each with its own distances, of course

- IAM- Infinite alleles model
- Any allele can mutate to any other allele
- Doesn't allow homoplasy, leads to overestimation
- SMM/TPM – Stepwise mutation model/ Two phase model
- SMM orders alleles, and mutations can only proceed one step forward or back
- Can correct for homoplasy
- TPM is just SMM with rare multi-step events

Measures of Distance

- D_{SA} : proportion shared alleles
- D_S : Nei's D , allele frequencies
- D_{CH} : Chord (linear) distance
- Θ : Angular Distance



IAM

- S_B : sum of squares of difference in allele sizes
- $(\delta\mu)^2$: delta mu squared



SMM/TPM

Some random comments gleaned from putting this together...

- D_{SA} : Good for assigning individuals to populations
- D_S : Assumes instantaneous fragmentation
- D_{CH} : Good for dendrograms, but you need lotsa loci (>30) and distantly related populations
- Θ : Highly sensitive to population size
- S_B : SMM distances are best for distantly related populations and higher mutation rates
- $(\delta\mu)^2$: In principle, independent of pop. size

The bottom line

- You want linearity, and low variance
- Choice of a distance depends on the model of evolution
 - IAM tends to work for less diverged pops, or lower mutation rates
 - SMM is often best for m-sats, which have high mutation rates
- All of these distance metrics are sensitive to demography
- But the good news- all of them are improved with more loci, more alleles, and more individuals (so at least there's that)

So what do we do with these again?

- Assigning individuals to populations
- Estimating divergence
- Estimating gene flow
- Reconstructing phylogenies
- Or Mantel's Tests...

Mantel's Test

Guess who came up with this one
(In 1967)

So what is it?

- Basically, it's a correlation between matrices consisting of some measurements, estimations, or predictions of dissimilarity (or similarity of course)
- Parametric analyses (the conventional kinds of statistics) are confounded by auto-correlation among variables, and this is the work around

Matrices can come from anywhere

- Experimental estimates of distance
- Predicted distances generated from a theoretical model
- Geographic variables
- Environmental predictor variables
- You get the picture- the key is that it can be adapted for variables of different logical types (eg categorical, rank, interval-scale data...)

What does it mean?

- The operative question- Do samples that are similar for metric 1 tend to be similar for metric 2?
- Don't worry, I'll *try* to make this more clear in a minute
- Significance must be assessed using permutation
- Importantly, all non-linear relationships are lost in these tests

Mantel's Test on Geographic Distances

- Species similarity is the dependent distance matrix (for example)
- Geographic distance (spatial dissimilarity) is the predictor matrix
- The question- Are samples that are found close together similar in their species composition?

More simple Mantel's

- Species similarity and rain fall- Does similar precipitation levels lead to similar species composition?
- Observed and predicted species similarity- Is the predicted composition similar to the observed composition?
- And so on...

Partial Mantel's Test

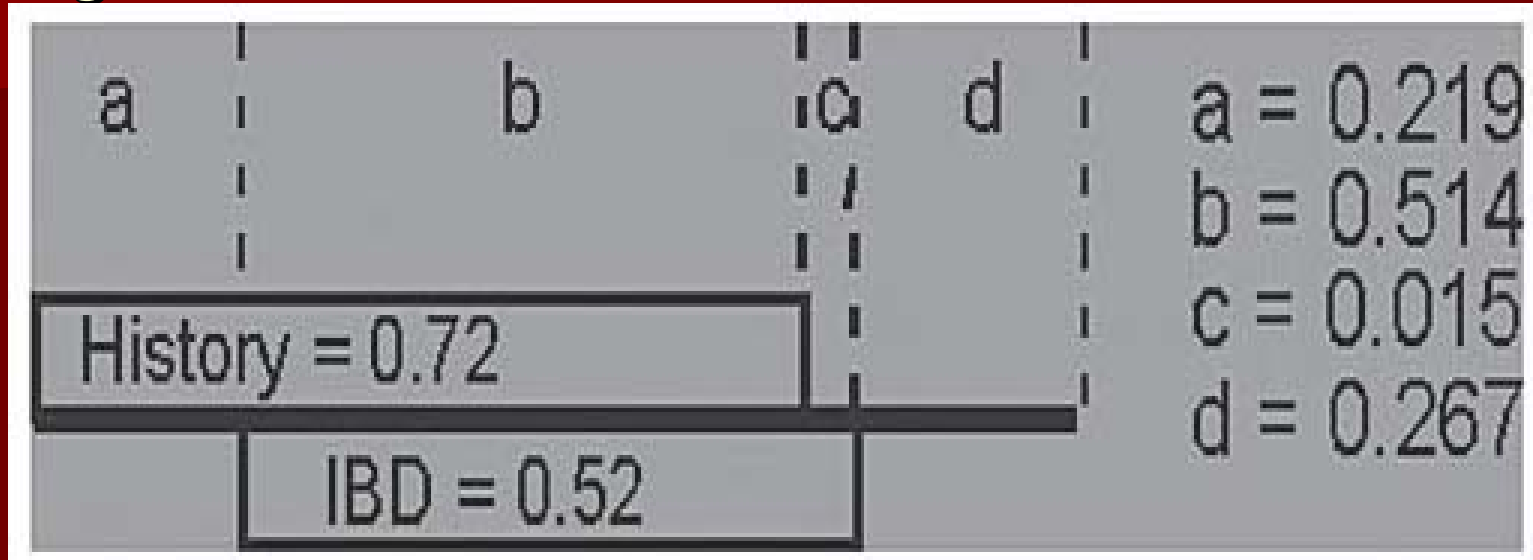
- Things like geographic location and climatic variables are obviously not unrelated, which will confound analysis
- Ideally we'd like to know how much is explained by the predictors, and whether the residuals themselves are similar to another matrix...

Genetic distance, isolation-by-distance, and historical divergence

- Telles and Diniz-Filho 2005
- Isolation-by-distance explained 52% of the variation in genetic distance
- Historical divergence (of *Eugenia dysenterica* in Brazil) using a binary matrix of eastern and western groups explains 72% of the variation
- To resolve these contributions, they used a partial Mantel's test

Interpreting the results

Figure 2 from Telles and Diniz-Filho 2005



- Long term divergence matrix is itself correlated with geography, therefore large overlap between the two is expected
- a is the variation explained by historical divergence alone, and c is from IBD alone, and they clearly show that a simple Mantel's test of IBD alone, though plenty significant, does not tell the whole story!

Biological intuition required!

- The beauty of these tests lies in their versatility
- Correlation does not necessarily mean causation, but path analysis can help (see Leduc *et al.* 1992)
- Thanks Bev, for helping me get this together
- And thank gawd for google!