Plan to meet weekly for 6-8 weeks (maybe more if popular demand)
Emphasis on my software but we may range further afield
Is this a good time to meet?
Plan to Cover

- F_{st}
- The Block Jackknife
- PCA, *smartpca*. Projection. Shrinkage
- f-statistics. ADMIXTOOLS. (at least 2 lectures).
- DATES (estimate admixture dates).

Emphasis on algorithms, program options and usage *not* applications.
Acknowledgements: Chris Chang
Arti Tandon
Sam Pollack
Niru Chennagiri
Mengyao Zhao
Shop Mallick
... and many others in the Reich Lab.

Resources:

Stephan Schiffels course (useful though more elementary than what I will cover)

`https://comppopgenworkshop2019.readthedocs.io/en/latest/contents/01_setting_up/setting_up.html#the-computational-infrastructure-for-this`

On O2:

`~np29/o2bin` `~np29/broaddatax/course20data` (Schiffels' data)

On Odyssey

source `~npatterson/setup`

`~npatterson/course20data`

If you have no access to either of these please see me. We can get things running on a Mac.

F_{st} what is it and how to estimate it

First defined by Sewall Wright and Gustav Malécot:
The correlation between random gametes, drawn from the same subpopulation, relative to the total
This is unfortunately not precise and this has caused much trouble.
Reference: Bhatia et al. (Genome Research (2013) Estimating F_{st} ...
But not everything I want to talk about is in there!

Definition
Consider a biallelic marker in two populations with allele frequencies p_1, p_2. $q_i = 1 - p_i$.
Define

$$\begin{aligned} N &= (p_1 - p_2)^2 \\ D &= p_1 q_2 + p_2 q_1 \\ F &= F_{st} = N/D \end{aligned}$$

We can also write D as

$$D = N + p_1(1 - p_1) + p_2(1 - p_2)$$

which makes F look like a ratio of variances and probably motivated Wright. This also shows that $0 \leq F \leq 1$.

For a set of markers $S_1, S_2, \ldots S_T$, define N_k, D_k for marker k as above and

$$F = F_{st} = \frac{\sum_i N_i}{\sum_i D_i}$$

Note that F_{st} is a model parameter *not* a statistic.

F_{st} and genetic drift

Standard timescale; Probability of coalescence of 2 samples by τ is $1 - e^{-\tau}$.

Theorem 1

$$X \xrightarrow{\tau} Y$$

and an allele has frequency x, $0 < x < 1$ in the population X and y in the population Y. Then:

$$E(y(1-y)|x) = x(1-x)e^{-\tau}$$

Proof [Myers]:
Consider 2 alleles chosen independently from Y. The probability that we have a heterozygote is $2E(y(1-y))$. On the other hand consider the MRCA. For a het, we cannot have coalescence more recently than the time of population X. Probability of two distinct ancestors at X is $e^{-\tau}$ and conditional on that, probability of a het is $2x(1-x)$. □

Exercise:

Show that

$$\frac{E(N|x)}{E(D|x)} = \frac{1 - e^{-\tau}}{2}$$

independent of the allele frequency x.

Estimating F_{st}

First consider a single SNP.

$$F = F_{st} = N/D$$

Probabilities p_1, p_2 for variant allele in populations P_1, P_2. $N = (p_1 - p_2)^2$ We observe allele counts a_1, a_2 for the variant
Conventionally, we will code the variant allele as 0, reference allele as 1. Counts b_1, b_2 for the reference allele. Take $n_i = a_i + b_i, \ i = 1, 2$. We of course find that

$$\hat{p}_i = a_i/n_i$$

is an unbiased estimate of p_i. We want an unbiased estimate of p_i^2. Set

$$h_i = p_i(1 - p_i)$$

the *heterozygosity*. We can write:

$$p_i^2 = p_i - h_i$$

Thus it is sufficient to find an unbiased estimate of h_i.

Case 1. No inbreeding

Pick 2 distinct alleles (u, v) independently and uniformly. Probability that $(u = 1, v = 0)$ is $p_1(1 - p_1) = h_1$. Thus setting $X = 1$ if $u = 1, v = 0$ else $X = 0$, X is an unbiased estimator of h_1. Now we average over all possible choices getting the estimator:

$$\hat{h}_1 = \frac{a_1 b_1}{n_1(n_1 - 1)}$$

[This is the 'Rao-Blackwell trick']. Here a_1, b_1 are sufficient statistics, and so $\hat{h}_1$ is the unique minimum variance unbiased estimator (MVUE). (Lehmann-Scheffé theorem. Bickel and Doksum Chapter 4).

Case 2. Inbreeding

Autosomes. x_0, x_1, x_2 are numbers of samples that have reference counts $0, 1, 2$, for population 1

$$\begin{aligned} s &= x_0 + x_1 + x_2 \\ \hat{p} &= \frac{x_1 + 2x_2}{2s} \end{aligned}$$

Pick 2 alleles u, v from *different* samples.

$$\hat{h} = P(u = 1, v = 0) = \frac{4x_0x_2 + 2(x_2 + x_0)x_1 + x_1(x_1 - 1)}{4X(X - 1)}$$

where $X = x_0 + x_1 + x_2$.

Case 3. Inbreeding and X-chromosome

Counts x_0, x_1, x_2 diploids (females)
Counts y_0, y_1 haploids (males)
$X = x_0 + x_1 + x_2$
$Y = y_0 + y_1$.

Exercise: Set

$$\begin{aligned} T &= 4x_0x_2 + 2(x_2 + x_0)x_1 + x_1(x_1 - 1) + (2x_0 + x_1)y_0 + (2x_2 + x_1)y_1 + y_0y_1 \\ B &= 4X(X - 1) + 4XY + Y(Y - 1) \end{aligned}$$

Show $\hat{h} = T/B$ is a MVUE estimator for heterozygosity h.

As before

$$\begin{aligned}
\hat{p}_1^2 &= \hat{p}_1 - \hat{h}_1 \\
\hat{p}_2^2 &= \hat{p}_2 - \hat{h}_2 \\
\hat{N} &= (\hat{p}_1 - \hat{p}_2)^2 + \widehat{p_1^2} - \hat{p}_1^2 + \widehat{p_2^2} - \hat{p}_2^2 \\
\hat{D} &= \hat{N} + \hat{h}_1 + \hat{h}_2
\end{aligned}$$

Note that $\hat{N}$ is unbiased and can therefore be negative.
Therefore so can F_{st}.

Highly negative F_{st} either indicates some artifact, or relatives in each of the two populations.

Meaning of F_{st}

What F_{st} can mean

Single SNP	N/D
Multiple SNPs	$\frac{\sum_i N_i}{\sum_i D_i}$
Ascertainment	$E(N)/E(D)$
Simple Demography.	$E(N)/E(D)$ (independent of ascertainment)

In no case is F_{st} a statistic, but a parameter that we can estimate.

Two options in *smartpca*

1. fstz: YES (lower triangle of F_{st} array are Z-scores assuming mean 0)
2. megaoutname: `<megafile>` (output suitable for *mega* graphics software)

Would like more software for plotting F_{st} based phylogenies.