Imperial College
London

Estimation of summary statistics

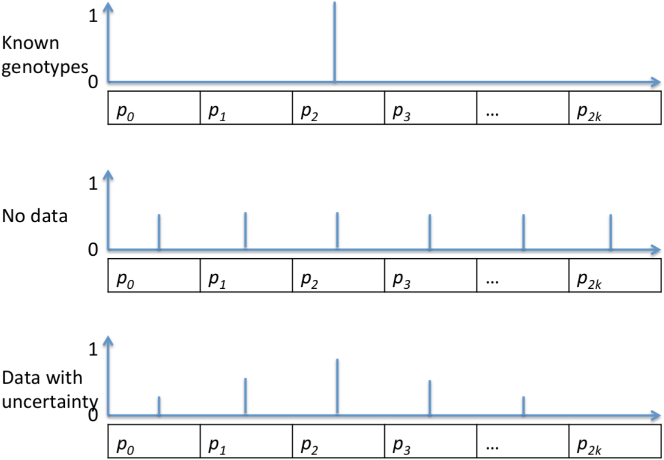
Matteo Fumagalli

Intended Learning Outcomes

By the end of this session you will be able to

- understand the theory underlying commonly used summary statistics
- appreciate how to extended such theory to low-coverage data
- acknowledge the process of inferring selection from sequencing data
- implement a pipeline in ANGSD to perform the aforementioned analyses

Sample allele frequency probabilities



Sample allele frequency posterior probabilities

$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$

- Estimating allele frequency

$$\hat{f} =$$

Sample allele frequency posterior probabilities

$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$

- Estimating allele frequency

$$\hat{f} = \sum_{i=0}^{2k} \left(\frac{i}{2k} \right) p(S = i)$$

Sample allele frequency posterior probabilities

With 6 chromosomes (3 diploids)

p_0=0.10	p_1=0.15	p_2=0.50	p_3=0.15	p_4=0.05	p_5=0.05	p_6=0.00

- SNP calling

$$p_{\text{var}} = \ ?$$

$$p_{\text{var}} > t$$

with *t* being 0.95, 0.99, 0.999 and so on.

Sample allele frequency posterior probabilities

p_0=0.10	p_1=0.15	p_2=0.50	p_3=0.15	p_4=0.05	p_5=0.05	p_6=0.00

- SNP calling

$$p_{\text{var}} = 1 - p(S=0) - p(S=2k) \quad = 0.90$$

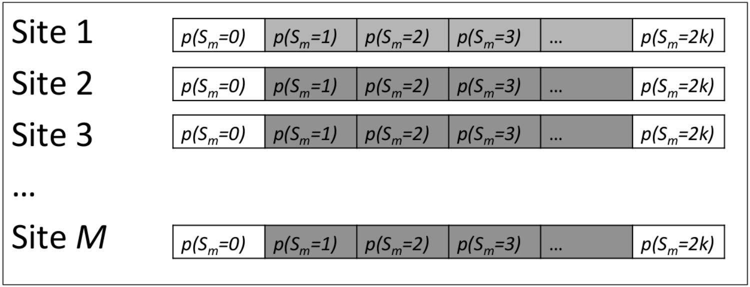
$$p_{\text{var}} > t$$

with *t* being 0.95, 0.99, 0.999 and so on.

Nr of segregating sites

Site 1	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 2	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 3	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
…						
Site M	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$

Nr of segregating sites



Site 1	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 2	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 3	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
…						
Site *M*	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$

Nr of segregating sites

Site 1	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 2	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 3	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
…						
Site M	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$

$$E[S]=\sum_{m=1}^{M} p_{\mathrm{var}}^{(m)}=\sum_{m=1}^{M}(1-p(S_m=0)-p(S_m=2k))$$

Nucleotide diversity

Site 1	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 2	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 3	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
…						
Site M	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$

$$D = 2f(1-f)$$

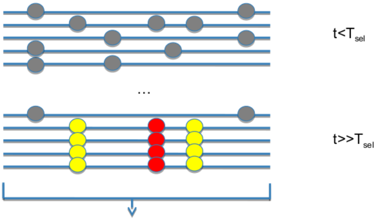
$$E[D] =$$

Nucleotide diversity

Site 1	*p(S* $p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 2	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 3	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
…						
Site *M*	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$

$$E[D]=\sum_{m=1}^{M}\sum_{j=0}^{2k}2\left(\frac{i}{2k}\right)\left(\frac{2k-i}{2k}\right)p(S_m=i)$$

Positive selection



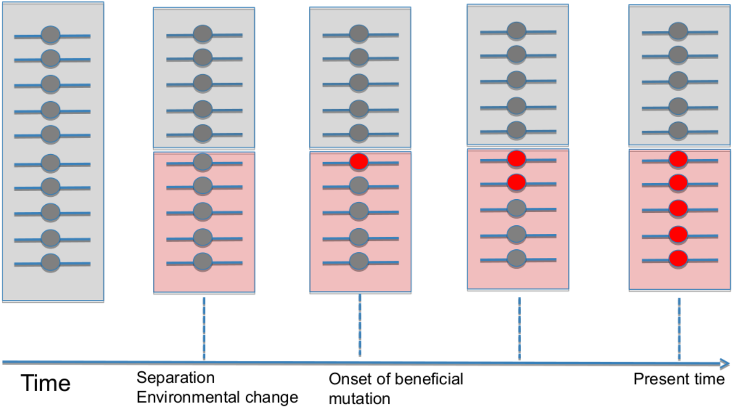
- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants (Pi)

Under neutrality, Theta and Pi are expected to be the same.
Tajima's D measures their difference.

$$D = \frac{\pi - \theta_W}{\sqrt{\hat{V}(\pi - \theta_W)}}$$

D<0 is suggestive of an excess of low-frequency variants

Allele frequency differentiation



Time

Separation
Environmental change

Onset of beneficial
mutation

Present time

F_{ST}

Common measure for quantifying population subdivision.

$$F_{ST} = H_B / (H_W + H_B)$$

H_B: between populations

H_W: average within populations

- if H_W<< H_B0 then F_{ST}~1
- if H_B=0 then F_{ST}=0

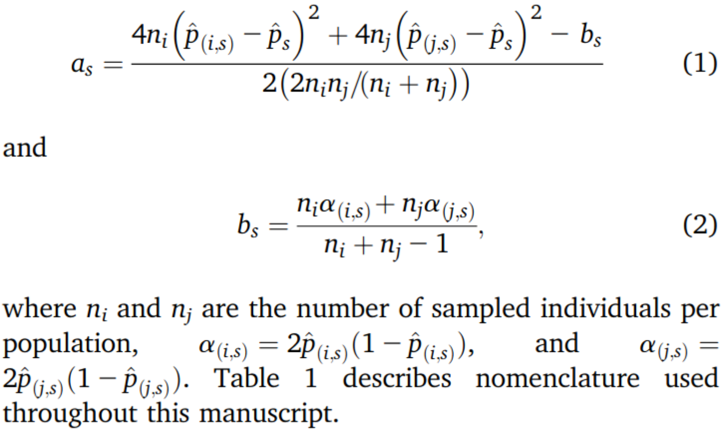
The estimate of F_{ST} for a single site is then

$$F_{\mathrm{ST}} = \frac{a_s}{a_s + b_s}$$

while for a *locus* of m sites it is

$$F_{\mathrm{ST}}^{(\mathrm{locus})} = \frac{\sum_{s=1}^{m} a_s}{\sum_{s=1}^{m} (a_s + b_s)}.$$

The genetic variance between and within populations at site s is, respectively,



$$a_s = \frac{4n_i\left(\hat{p}_{(i,s)} - \hat{p}_s\right)^2 + 4n_j\left(\hat{p}_{(j,s)} - \hat{p}_s\right)^2 - b_s}{2\left(2n_in_j/(n_i + n_j)\right)} \tag{1}$$

and

$$b_s = \frac{n_i\alpha_{(i,s)} + n_j\alpha_{(j,s)}}{n_i + n_j - 1}, \tag{2}$$

where n_i and n_j are the number of sampled individuals per population, $\alpha_{(i,s)} = 2\hat{p}_{(i,s)}(1 - \hat{p}_{(i,s)})$, and $\alpha_{(j,s)} = 2\hat{p}_{(j,s)}(1 - \hat{p}_{(j,s)})$. Table 1 describes nomenclature used throughout this manuscript.

Fumagalli et al. Genetics 2013

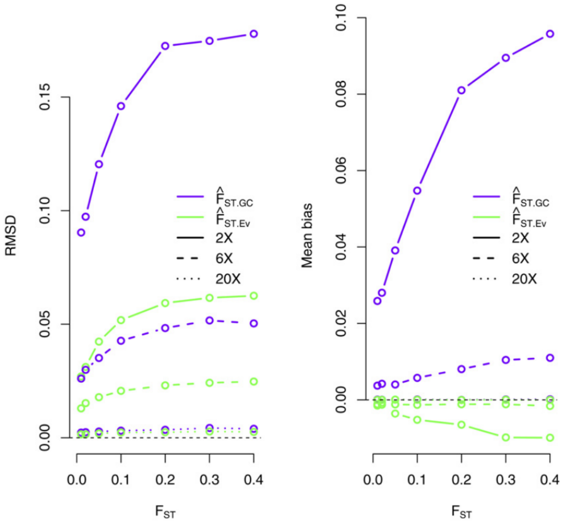
Method-of-moments estimation: Let $\pi_i^{(k)} = P(\widehat{p_i} = k/(2n_i)|Y_{(i,s)})$ be the posterior probability that a site in population i has derived sample allele frequency $\widehat{p_i} = k/(2n_i)$, in a sample of n_i diploid individuals, given the read data $Y_{(i,s)}$.

From these quantities, we compute the posterior expectation of the genetic variance between and within populations (see Equations 1 and 2) at site s as

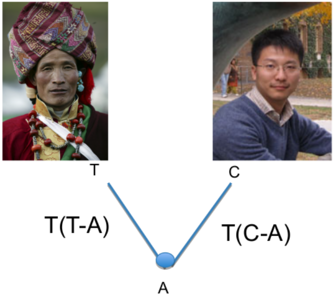
$$E[a_s|Y_s] = \sum_{k=0}^{2n_i} \sum_{z=0}^{2n_j} a_{(i,j)}^{(k,z)} \pi_{(i,j,s)}^{(k,z)} \tag{10}$$

and

$$E[b_s|Y_s] = \sum_{k=0}^{2n_i} \sum_{z=0}^{2n_j} b_{(i,j)}^{(k,z)} \pi_{(i,j,s)}^{(k,z)}, \tag{11}$$


RMSD
0.00
0.05
0.10
0.15
$\hat{F}_{ST.GC}$
$\hat{F}_{ST.Ev}$
2X
6X
20X
0.0
0.1
0.2
0.3
0.4
F_{ST}
Mean bias
0.00
0.02
0.04
0.06
0.08
0.10
$\hat{F}_{ST.GC}$
$\hat{F}_{ST.Ev}$
2X
6X
20X
0.0
0.1
0.2
0.3
0.4
F_{ST}

Population genetic differentiation



T

C

T(T-A)

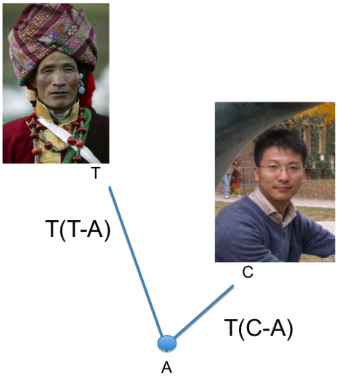
T(C-A)

A

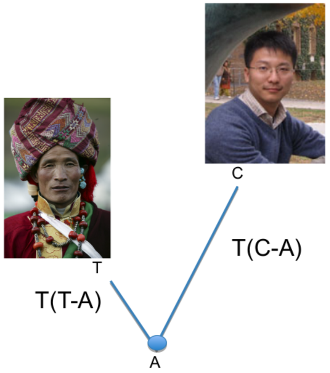
F_{ST}(T-C)~T(T-A-C)

Population genetic differentiation

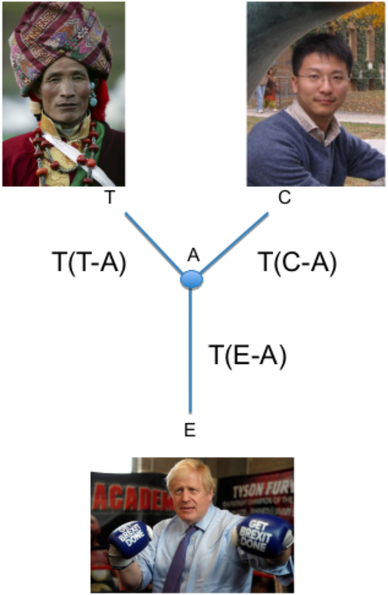
$F_{ST}(\text{T-C}) \sim \text{T(T-A-C)}$



?



Population branch statistic (PBS)



T

C

T(T-A)

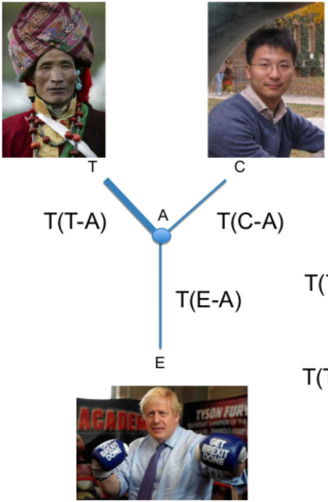
A

T(C-A)

T(E-A)

E

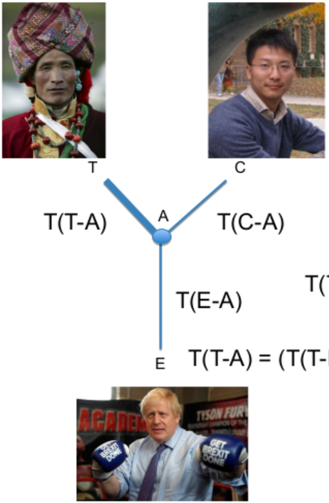
Population branch statistic (PBS)



T(T-A-C)=-log(1-F_{ST}(T-C))

T(T-A)?

Population branch statistic (PBS)



$$T(T\text{-}A\text{-}C)=-\log(1-F_{ST}(T\text{-}C))$$

$$T(T\text{-}A) = (T(T\text{-}E)+T(C\text{-}T)-T(C\text{-}E))/2$$

How can we estimate it from low-cov data?

Imperial College
London

Intended Learning Outcomes

At the end of this session you are now be able to

- understand the theory underlying commonly used summary statistics
- appreciate how to extended such theory to low-coverage data
- acknowledge the process of inferring selection from sequencing data
- implement a pipeline in ANGSD to perform the aforementioned analyses