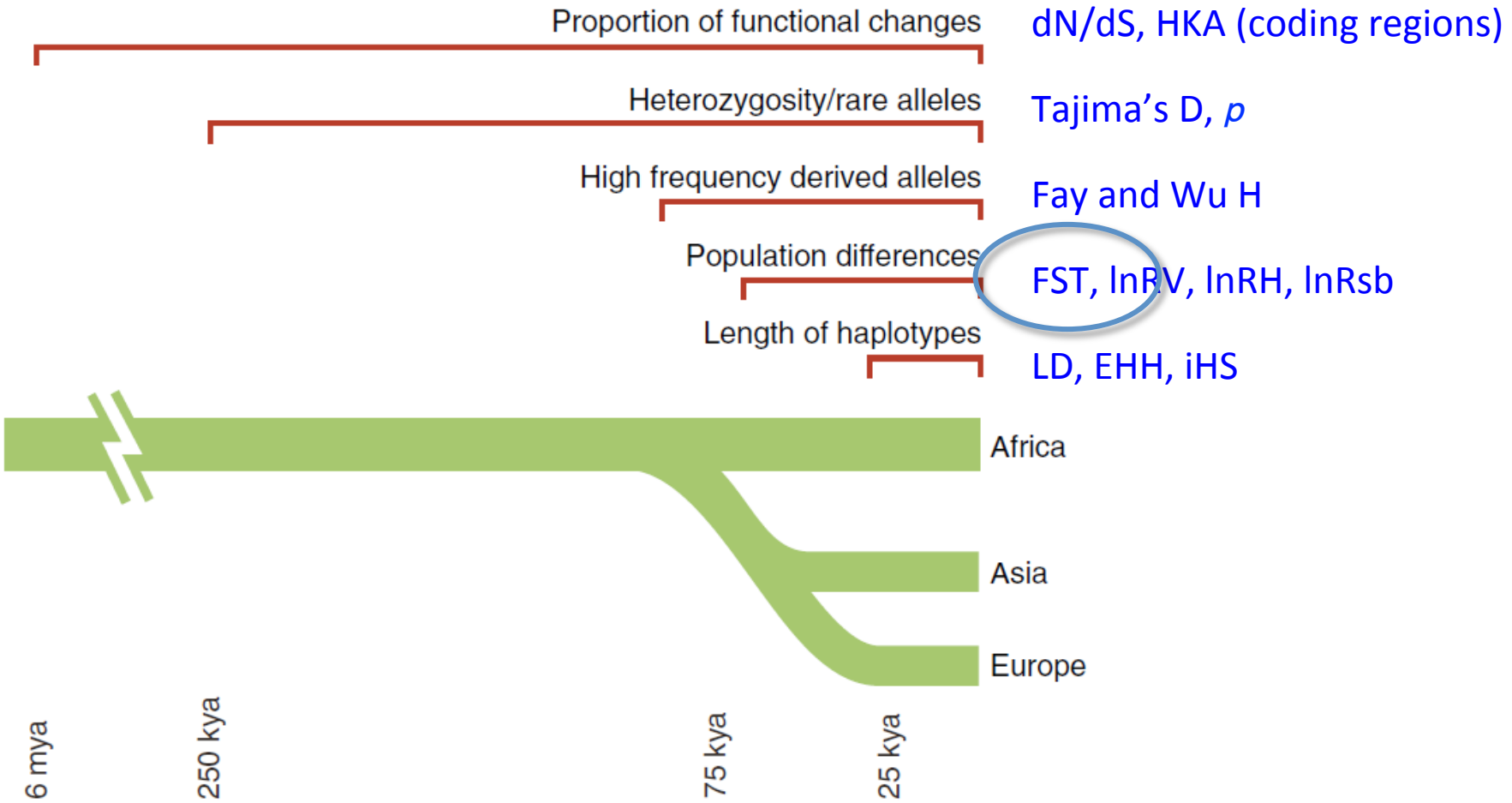
Detection of loci under selection from genome scans

Matthieu Foll

23.11.2011

Population Genomics course, Helsinki

Methods to detect loci under selection



Sabeti et al. 2006

Selection affects patterns of diversity between populations

Cavalli Sforza 1966 Population structure and human evolution. Proc Roy Soc B. 164: 362–379

Adaptation:

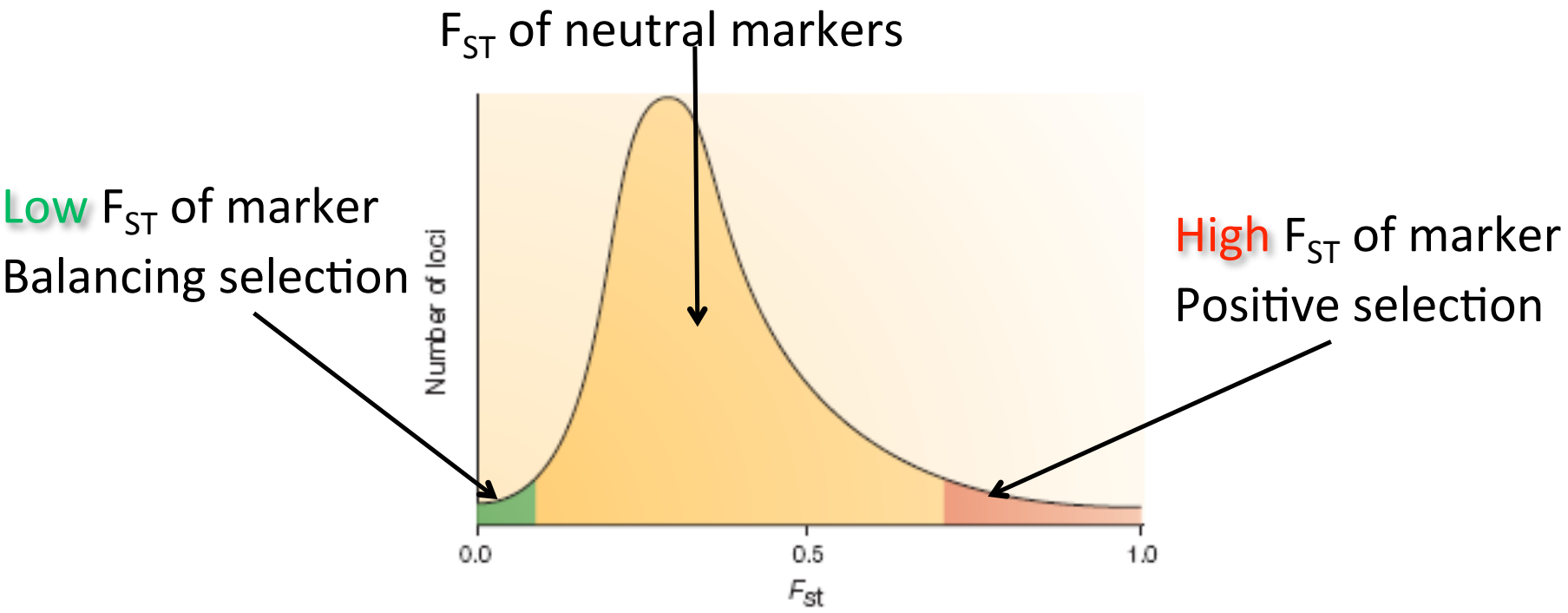
Recent selective sweep in a population or in a given region
will increase genetic differentiation (FST) between populations

Balancing selection:

Balancing selection may maintain alleles at low frequencies between populations (frequency dependent selection) or maintain particular alleles at identical frequencies in many populations (heterozygote advantage).

This will decrease genetic differentiation between populations
(low FST)

FST and selection



FST-based tests of selection

Lewontin and Krakauer (1973) have proposed to use the variance of F_{ST} across loci as a test of neutrality.

$$\sigma^2 = \frac{k\,F_{ST}^2}{n-1}$$

where $k \leq 2$, and n is the number of sampled populations.

This approach has been criticized:

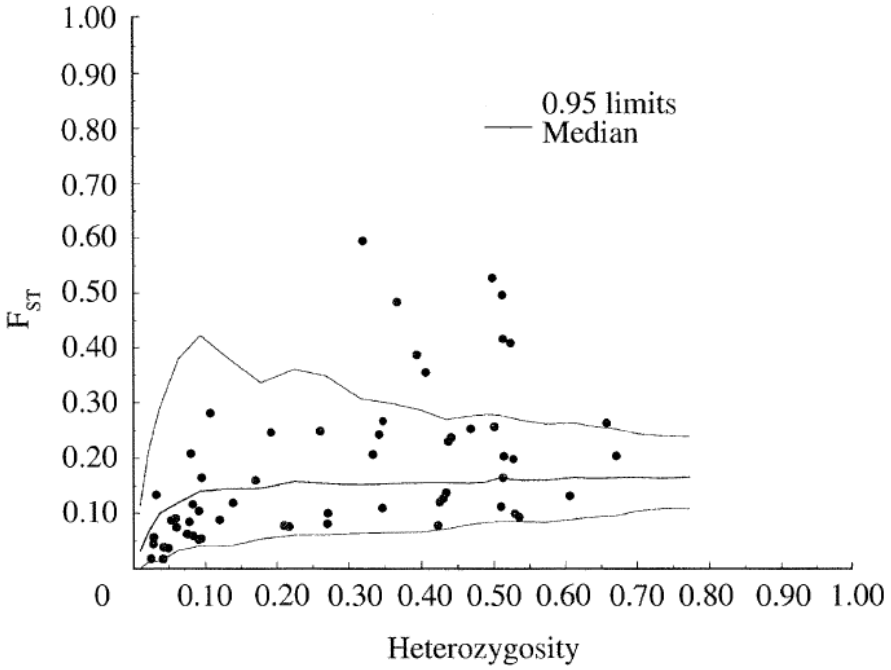
- This variance seemed often underestimated for some types of population structure
  - Isolation by distance (Nei and Maruyama, 1975)
  - Shared ancestry (Robertson 1975)
- The test has not been used very much, also due to a lack of appropriate data sets.

FST-based tests of selection

Beaumont and Nichols (1996) proposed to use the joint distribution of F_{ST} and heterozygosity between population, to detect outlier loci.

They used coalescent simulations under a finite island model to obtain the joint null distribution.

Method implemented into the **FDIST2** program



FDIST2 algorithm

- Calculate Fst from the observe data
- Converts Fst into migration rate using

$$F_{ST} = \frac{1}{1 + \frac{4Nmd}{(d-1)}}$$

- Simulate the joint null distribution of heterozygosity and Fst using coalescent simulations
- Calculate p-values for each locus based on the simulated distribution

FST-based tests of selection

The FDIST2 method was shown by Beaumont and Nichols (1996) to be relatively robust to alternative structures of population (colonization model, stepping-stone model).

The method of Beaumont and Nichols (1996) has been used extensively with the advent of the first genome scans (since 2002)

Problems with FST-based methods

A large number of outlier loci may be due to the same problem as incurred by the Lewontin and Krakauer test

This test was initially criticized by Robertson (1975) in that the variance of F_{ST} could be largely underestimated if the allele frequencies between sampled populations are correlated

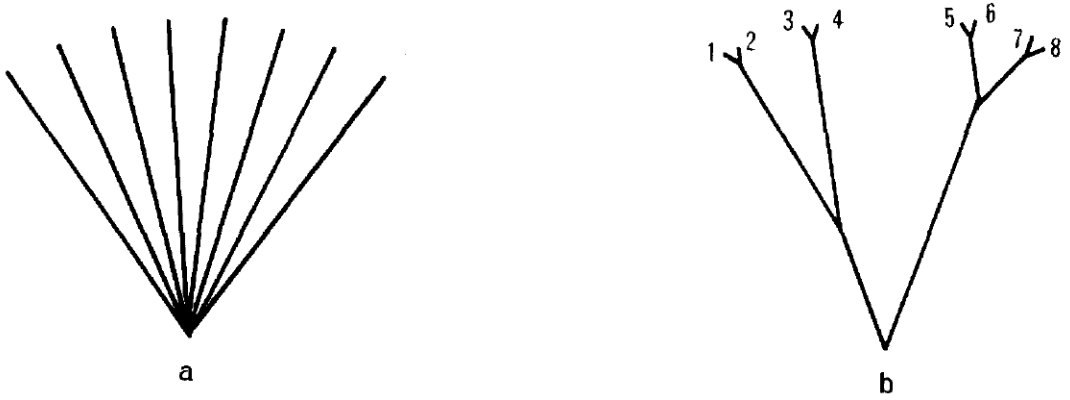


FIGURE 1.—Two hypothetical structures of relationship of populations within species.

But almost all current approaches assume that sampled populations are independent from each other

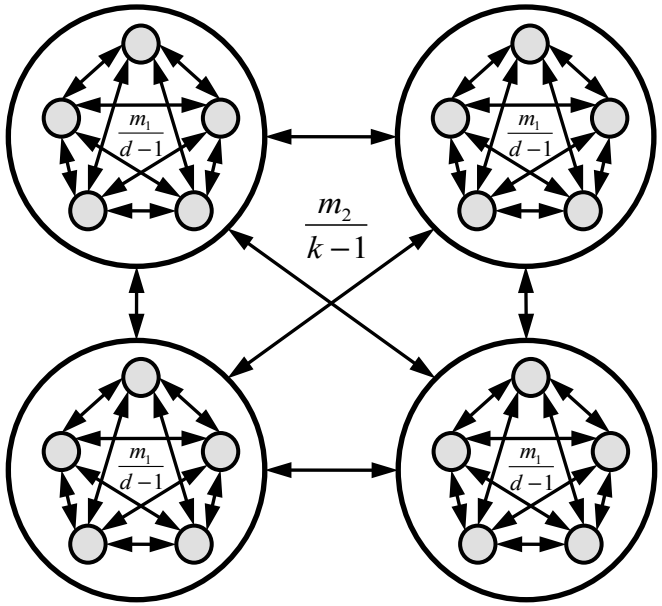
Extension of FST based-approaches: Hierarchical island model

Slatkin and Voelm (1991) have studied a hierarchical island model where they related hierarchical G-statistics to migration rates within and between groups of demes.

$$F_{SC} = \frac{1}{1+4Nm_1\dfrac{d}{d-1}}$$

$$F_{CT} = \frac{1}{1+4Nd\dfrac{k}{k-1}m_2+(d-1)\dfrac{k}{k-1}\dfrac{m_2}{m_1}}$$

$$F_{ST} \approx \frac{1}{1+4Nd\dfrac{k}{(k-1)}m_2}$$



This model was modified to get relationships between migration rates and F-statistics

These equations can be used to model a hierarchical island model leading to observed hierarchical F-statistics

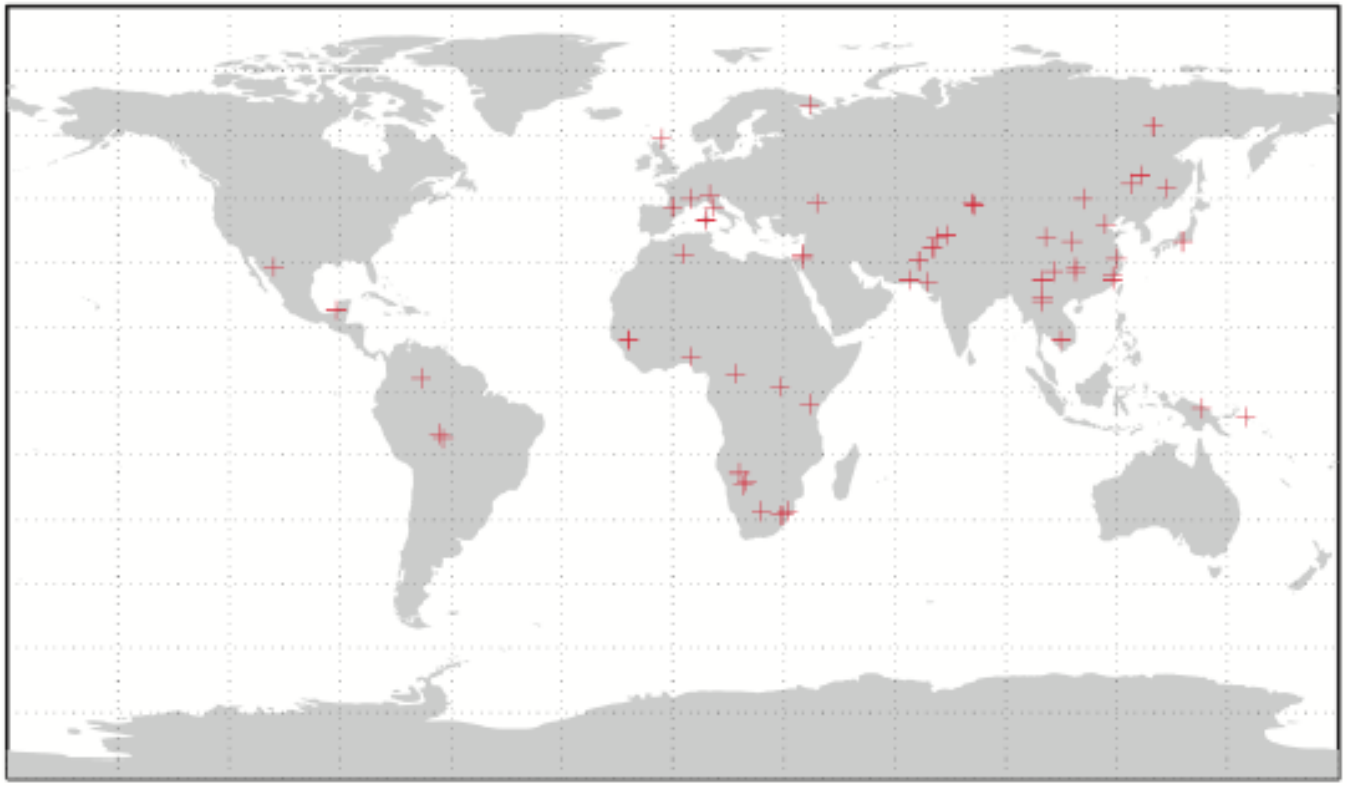
Excoffier *et al.* 2009

Detecting selection under a hierarchical island model

Procedure

1. Estimate F-statistics from observed data under a given hierarchical genetic structure (assumed known) of the populations
2. Convert F-statistics into migrations rates m_1 and m_2, assuming a given deme size N
3. Obtain the joint null distribution of heterozygosity and F_{ST} by simulating genetic diversity at neutral loci under this model
4. Compute the *p*-values of observed loci

HGDP data set

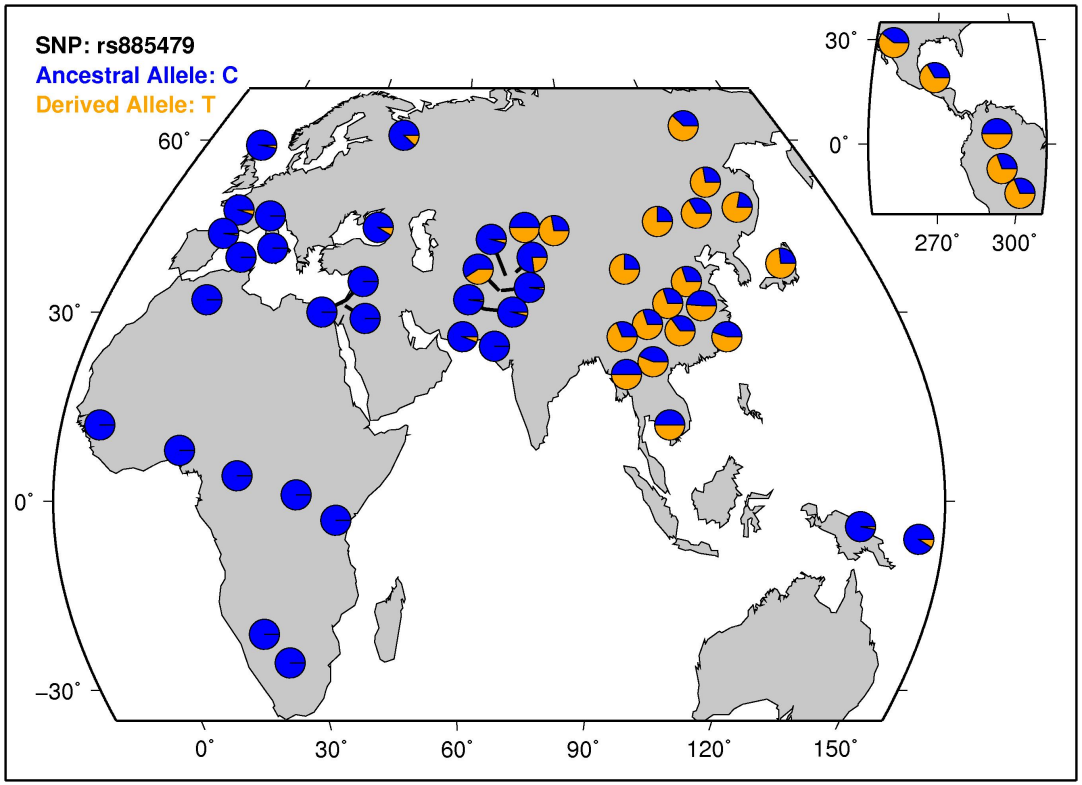


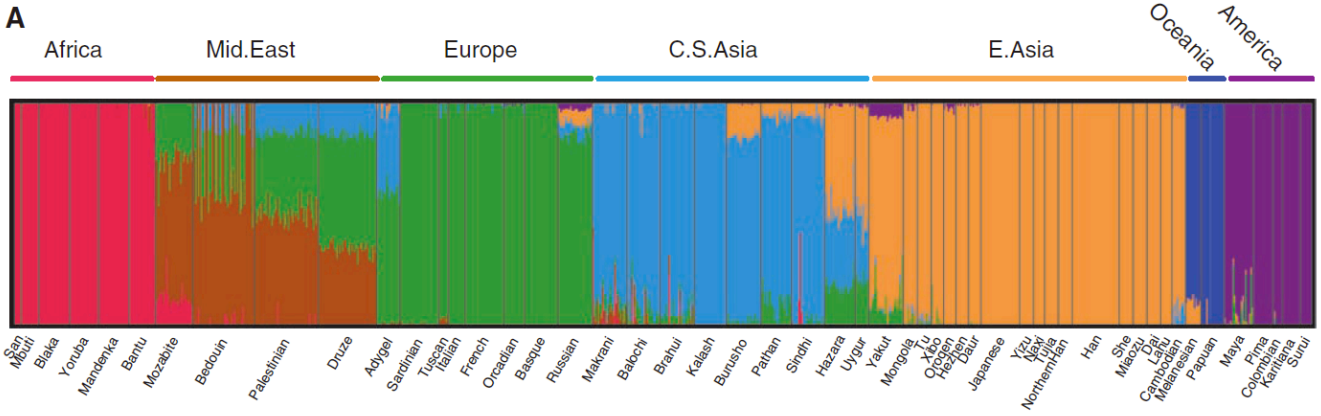
940 individuals in 53 populations

~600K SNPs in 22 chromosomes + X

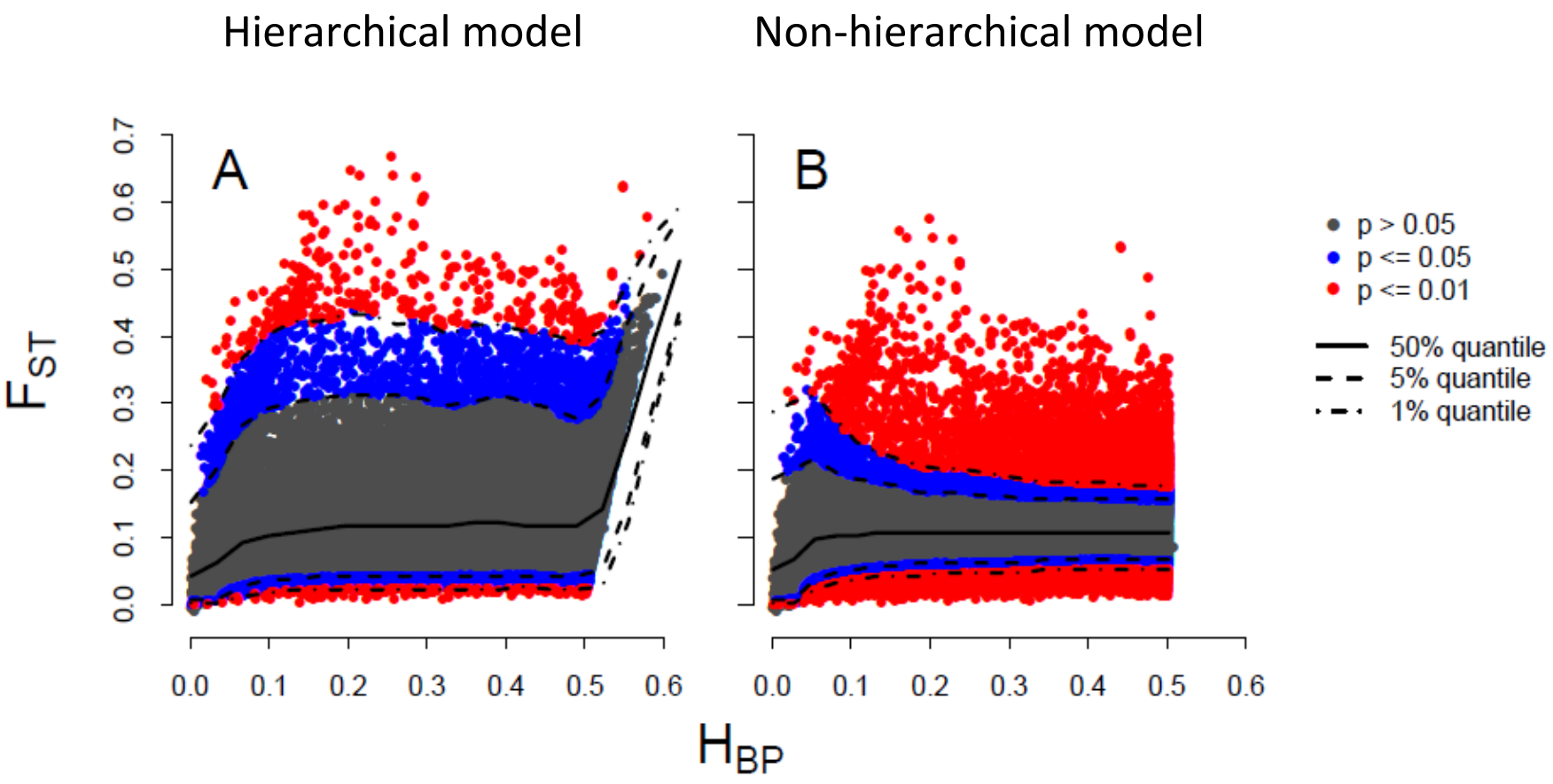
Li *et al.* 2008 (Illumina HumanHap 650K Beadchips)

HGDP data structure





Application to HGDP data set



Hofer et al, submitted

False positive rate

We studied the false positive rate when data generated under a hierarchical island model are analysed under a finite island model

STR data, 5 groups of 10 populations

F_{SC}=0.1; F_{CT}=0.05				
Expected false positive rate	*Balancing selection*		*Directional selection*	
	Finite island	Hierarchical island	Finite island	Hierarchical island
0.001	0.0155	0.0059	0.0083	0.0012
0.005	0.0301	0.0071	0.0233	0.0057
0.01	0.0412	0.0091	0.0350	0.0111
0.05	0.1023	0.0361	0.0875	0.0549

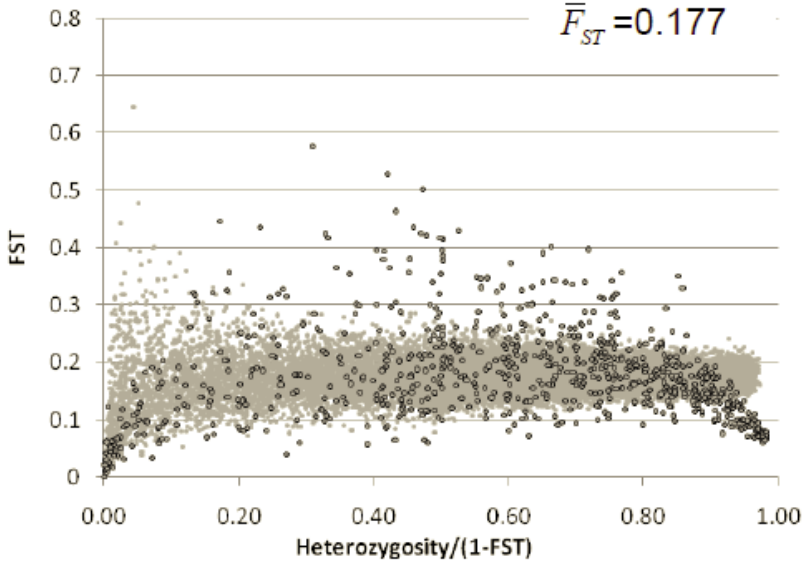
False positive rate

We studied the false positive rate when data generated under
a hierarchical island model are analysed under a finite island model
STR data, 5 groups of 10 populations

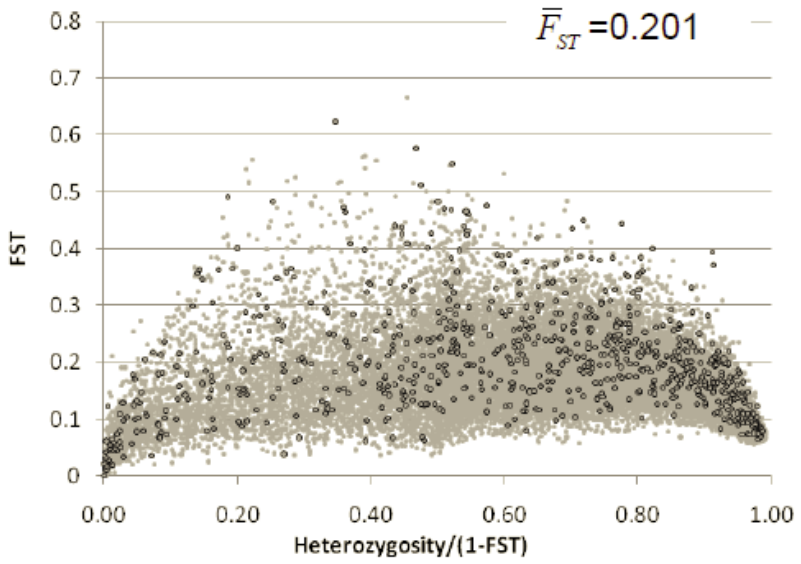
Expected false positive rate	F_{SC}=0.1; F_{CT}=0.05				F_{SC}=0.05; F_{CT}=0.2			
	Balancing selection		*Directional selection*		*Balancing selection*		*Directional selection*	
	Finite island	Hierarchical island	Finite island	Hierarchical island	Finite island	Hierarchical island	Finite island	Hierarchical island
0.001	0.0155	0.0059	0.0083	0.0012	0.1534	0.0065	0.0783	0.0012
0.005	0.0301	0.0071	0.0233	0.0057	0.2104	0.0089	0.1072	0.0046
0.01	0.0412	0.0091	0.0350	0.0111	0.2436	0.0130	0.1226	0.0091
0.05	0.1023	0.0361	0.0875	0.0549	0.3431	0.0495	0.1775	0.0455

False positive rate

We studied the false positive rate when data generated under
a hierarchical island model are analysed under a finite island model
STR data, 5 groups of 10 populations

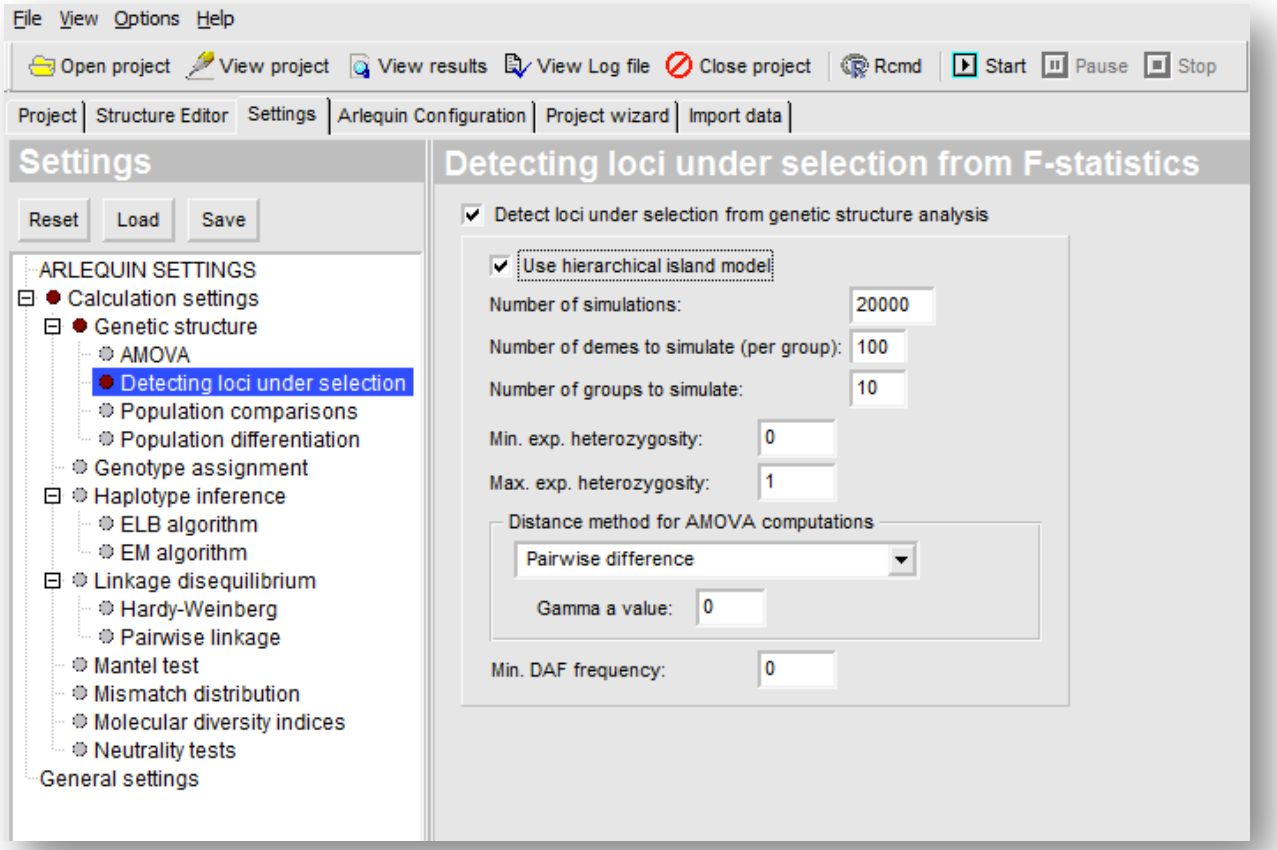


Finite island model

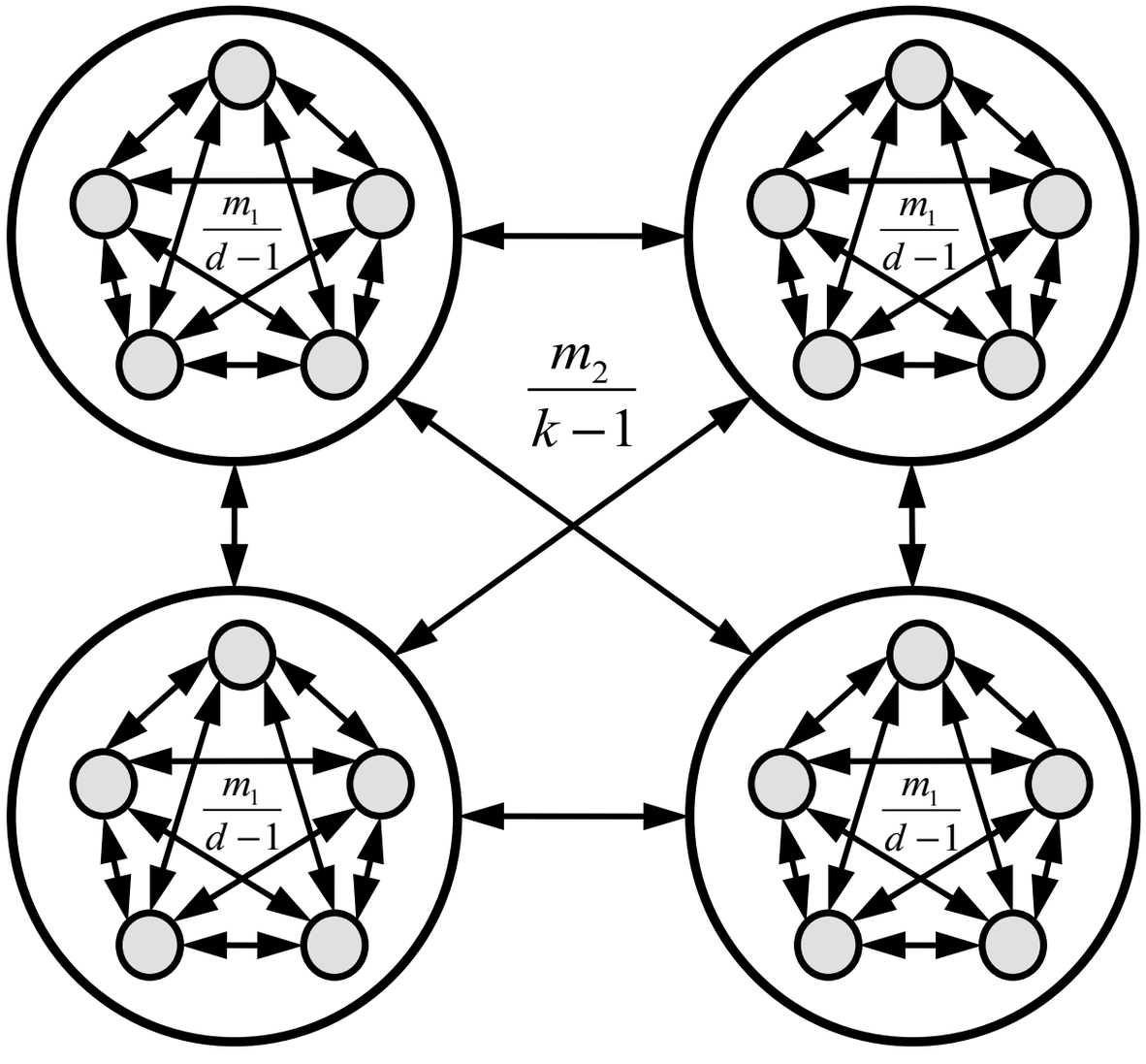


Hierarchical island model

Demo


File View Options Help
Open project
View project
View results
View Log file
Close project
Rcmd
Start
Pause
Stop
Project
Structure Editor
Settings
Arlequin Configuration
Project wizard
Import data
Settings
Reset
Load
Save
ARLEQUIN SETTINGS
Calculation settings
Genetic structure
AMOVA
Detecting loci under selection
Population comparisons
Population differentiation
Genotype assignment
Haplotype inference
ELB algorithm
EM algorithm
Linkage disequilibrium
Hardy-Weinberg
Pairwise linkage
Mantel test
Mismatch distribution
Molecular diversity indices
Neutrality tests
General settings
Detecting loci under selection from F-statistics
Detect loci under selection from genetic structure analysis
Use hierarchical island model
Number of simulations: 20000
Number of demes to simulate (per group): 100
Number of groups to simulate: 10
Min. exp. heterozygosity: 0
Max. exp. heterozygosity: 1
Distance method for AMOVA computations
Pairwise difference
Gamma a value: 0
Min. DAF frequency: 0

A realistic demographic model ?

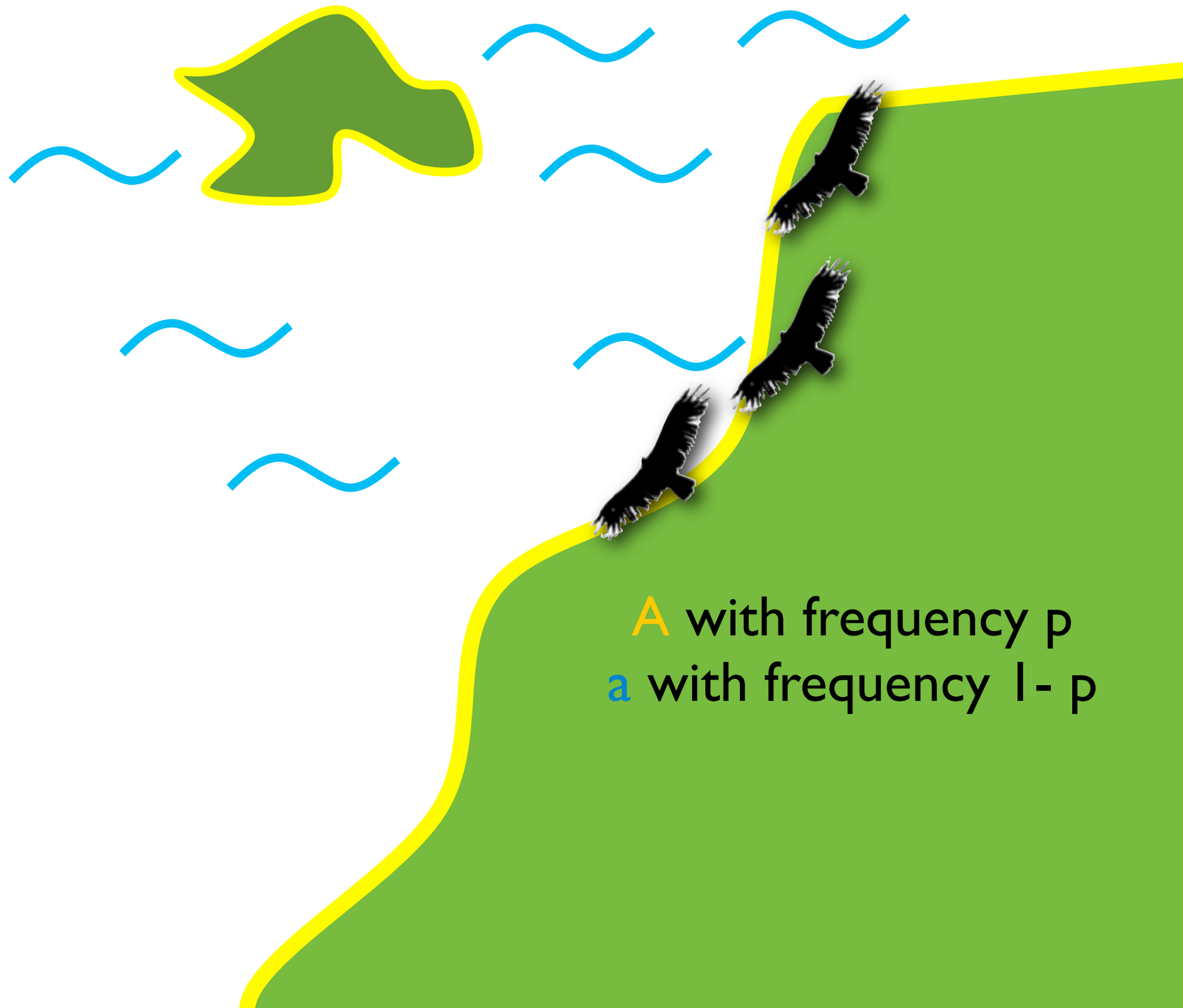


The F-model

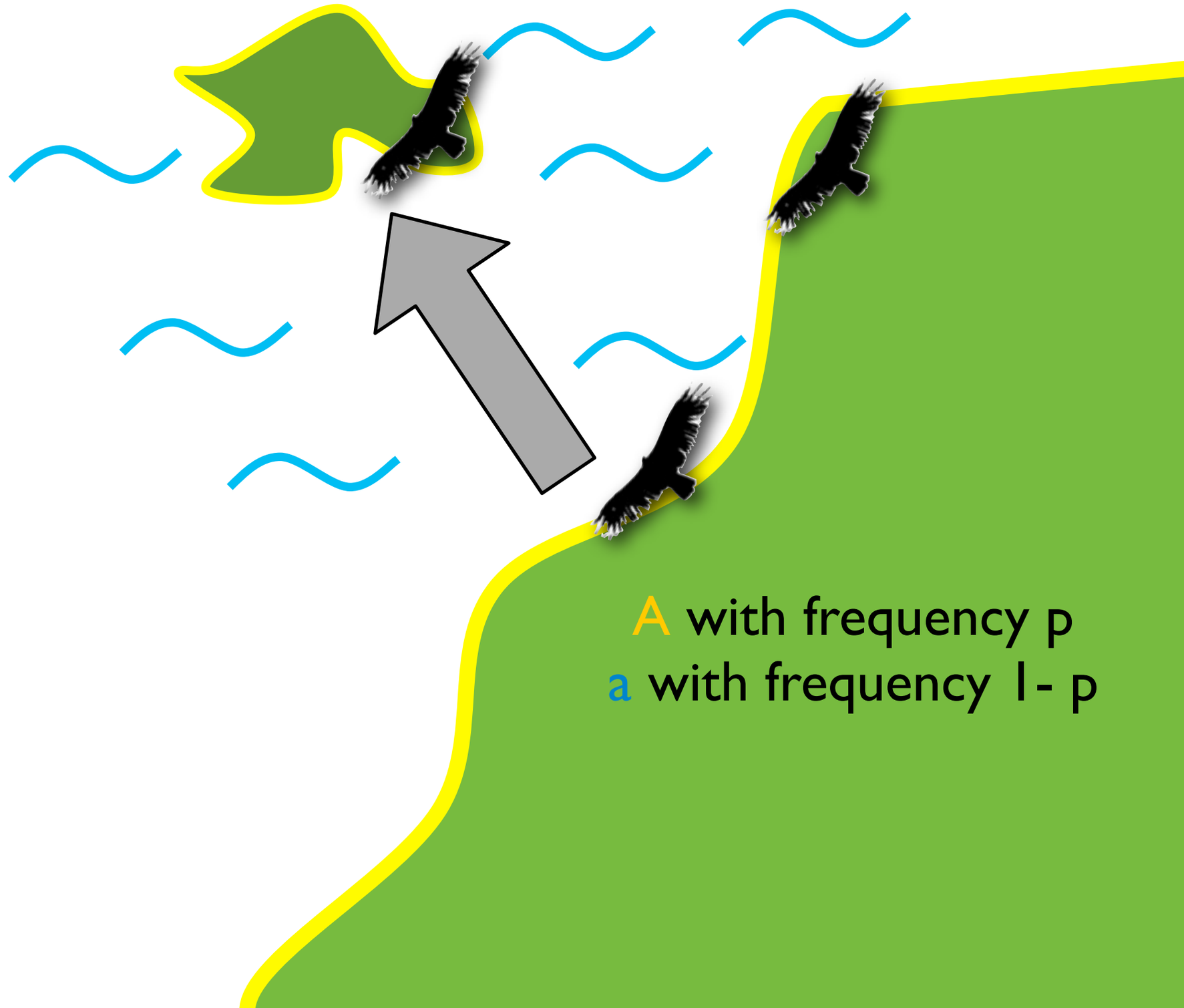
- Used in many software
  - STRUCTURE
  - Geneland
  - BAPS
  - Hickory
  - Bayenv
  - GESTE, BayeScan
  - ...
- Flexible Bayesian description of the genetic structure

Island - Mainland model

Island - Mainland model



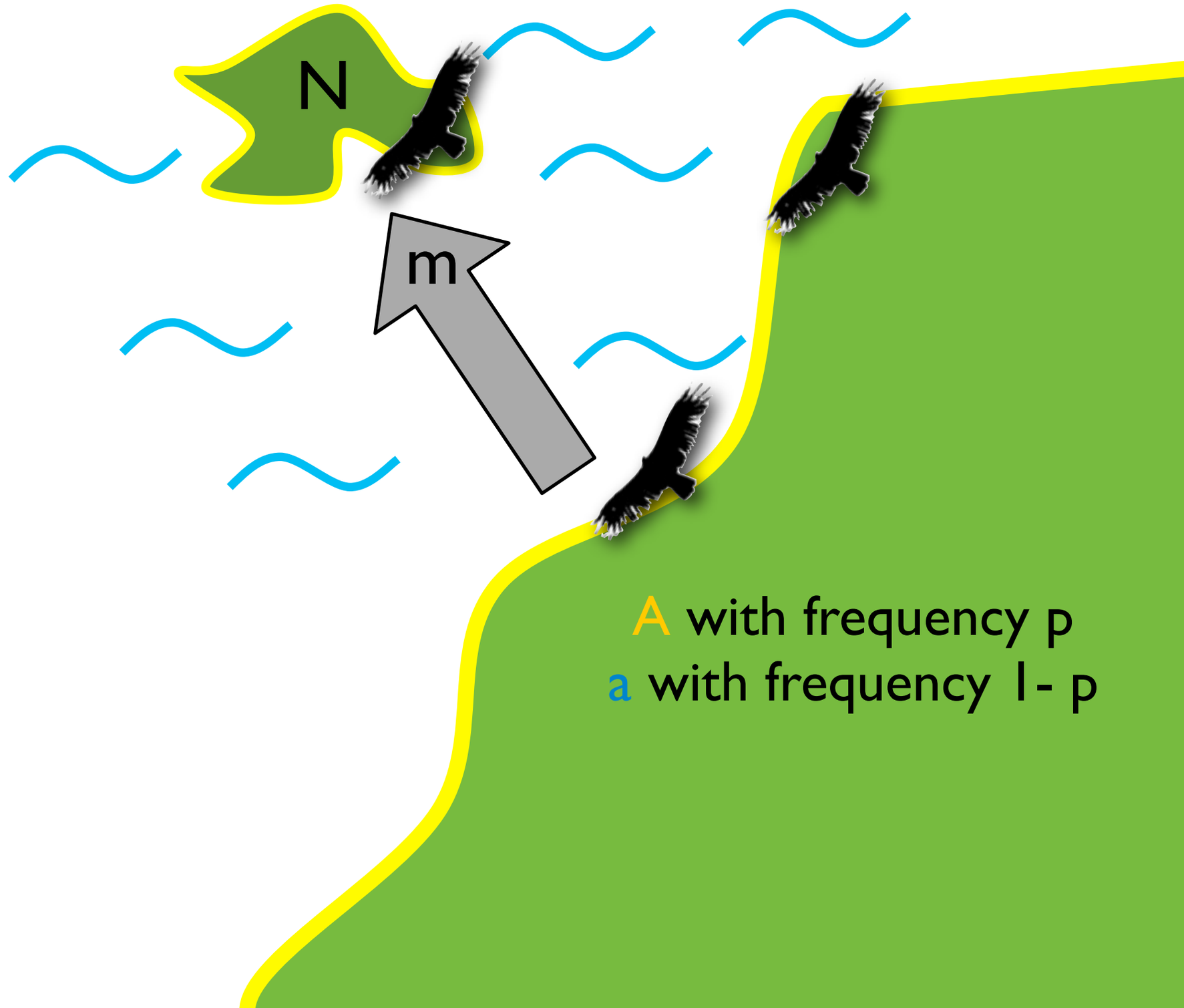
Island - Mainland model



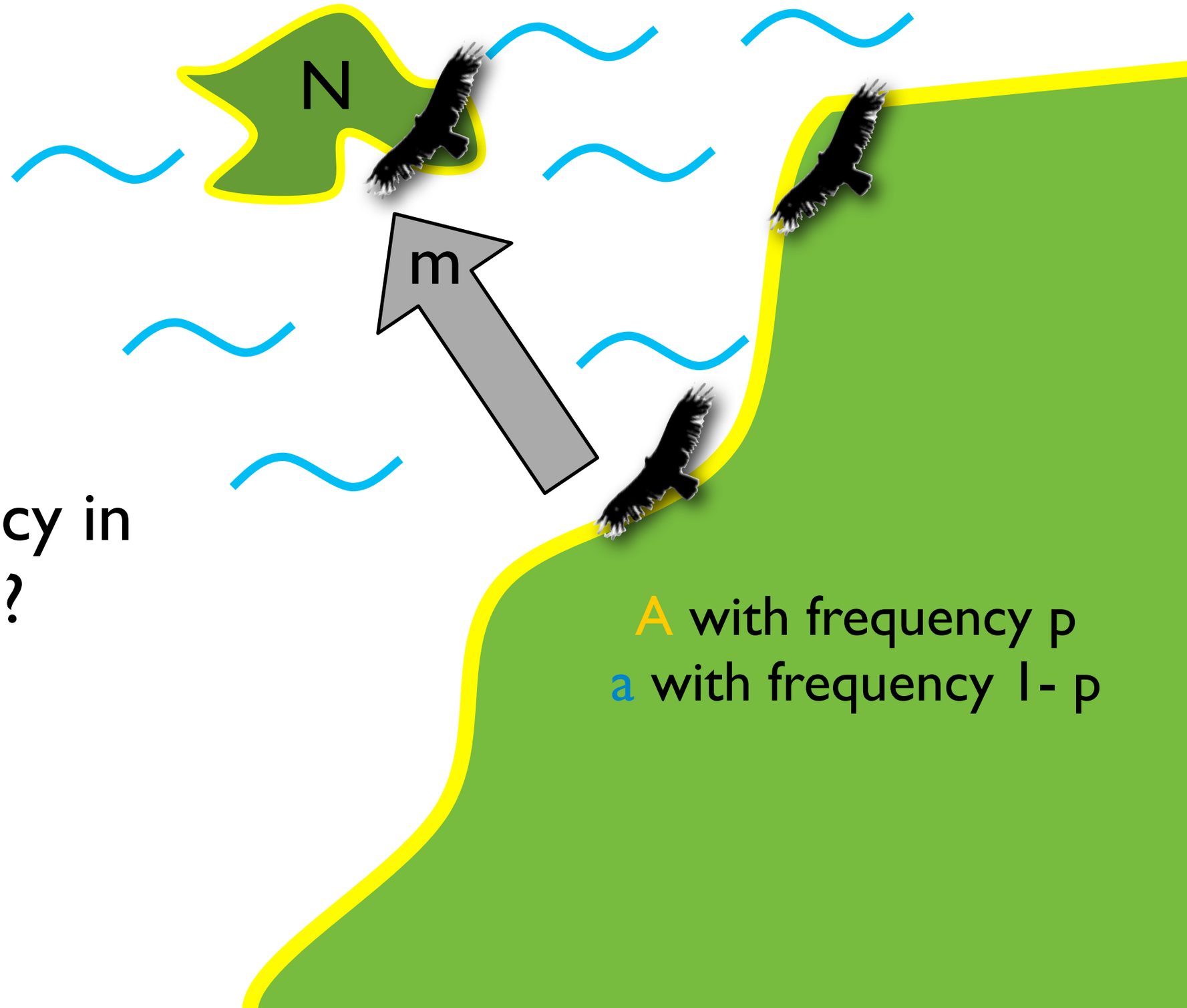
A with frequency p

a with frequency 1- p

Island - Mainland model



Island - Mainland model



Allele frequency in the island ?

Sewall Wright, 1931

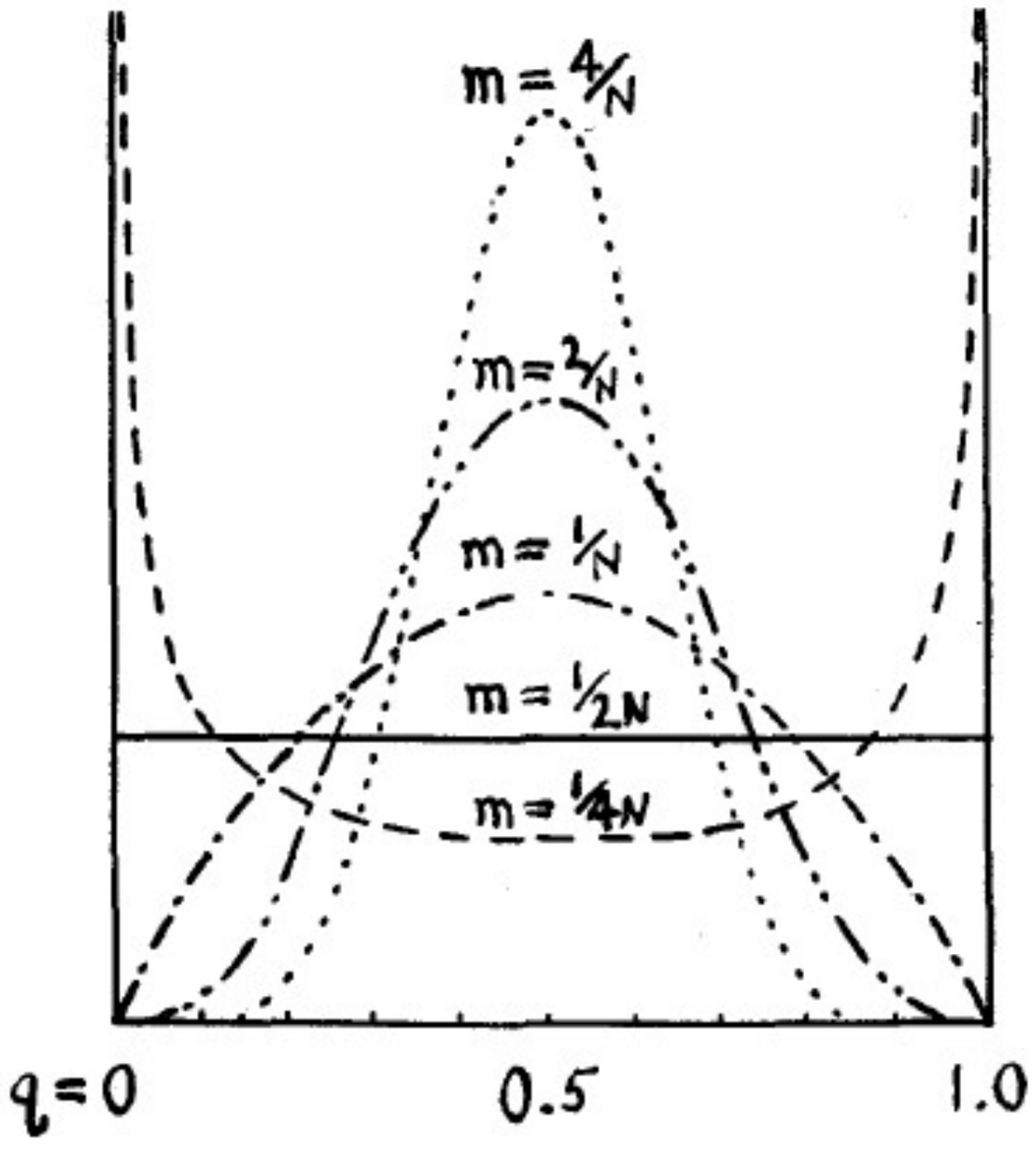


FIGURE 6.—Distribution of frequencies of a gene among subdivisions of a population in which $q_m = 1/2$ (or probability array of gene within a subdivision) under various amounts of intermigration. $y = Cq^{4Nmq_m - 1}(1-q)^{4Nm(1-q_m)-1}$.

Sewall Wright, 1931

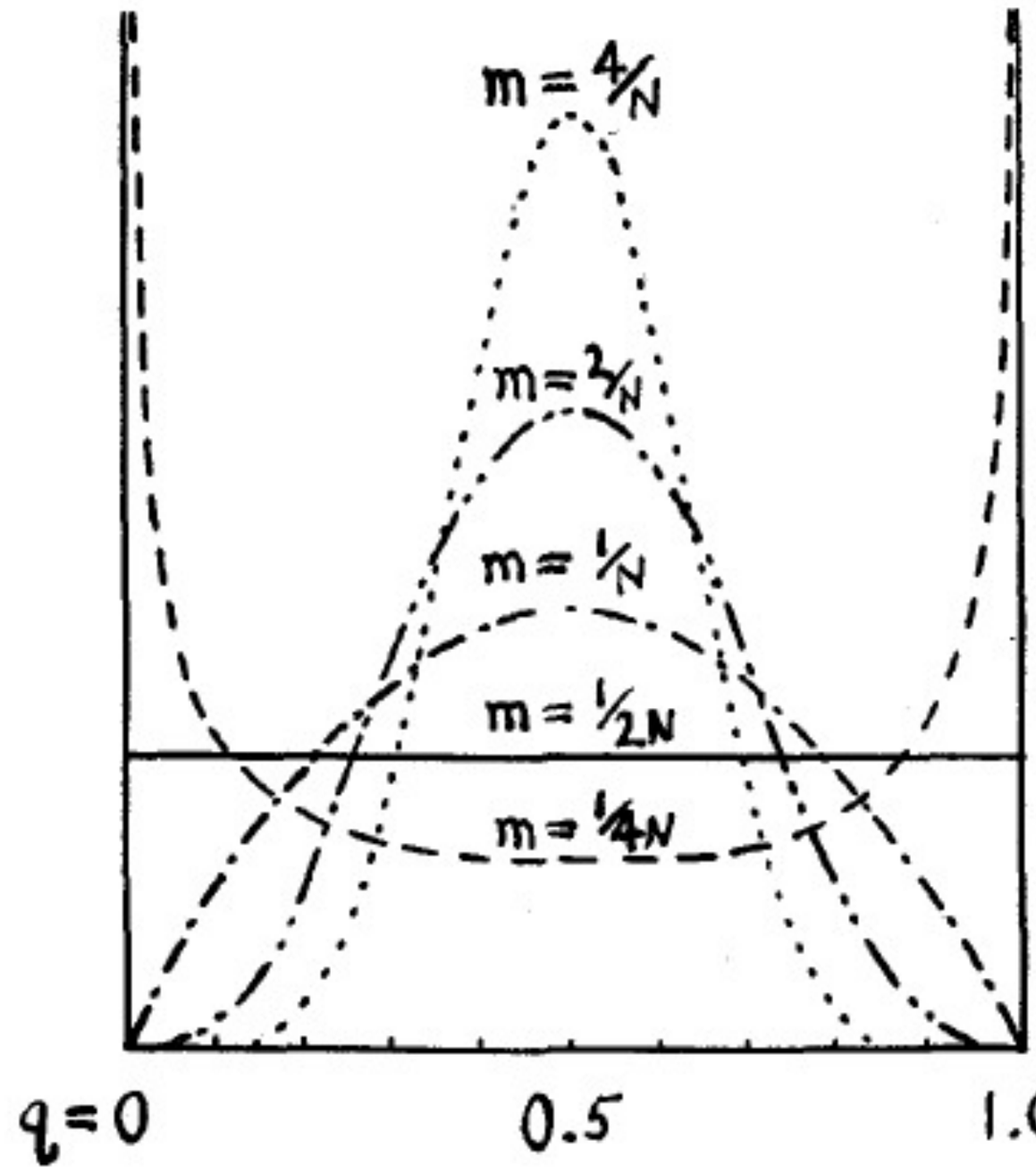


FIGURE 6.—Distribution of frequencies of a gene among subdivisions of a population in which $q_m = 1/2$ (or probability array of gene within a subdivision) under various amounts of intermigration. $y = Cq^{4Nmq_m - 1}(1-q)^{4Nm(1-q_m)-1}$.

$\widetilde{p} \sim \text{Beta}(\theta p, \theta(1-p))$

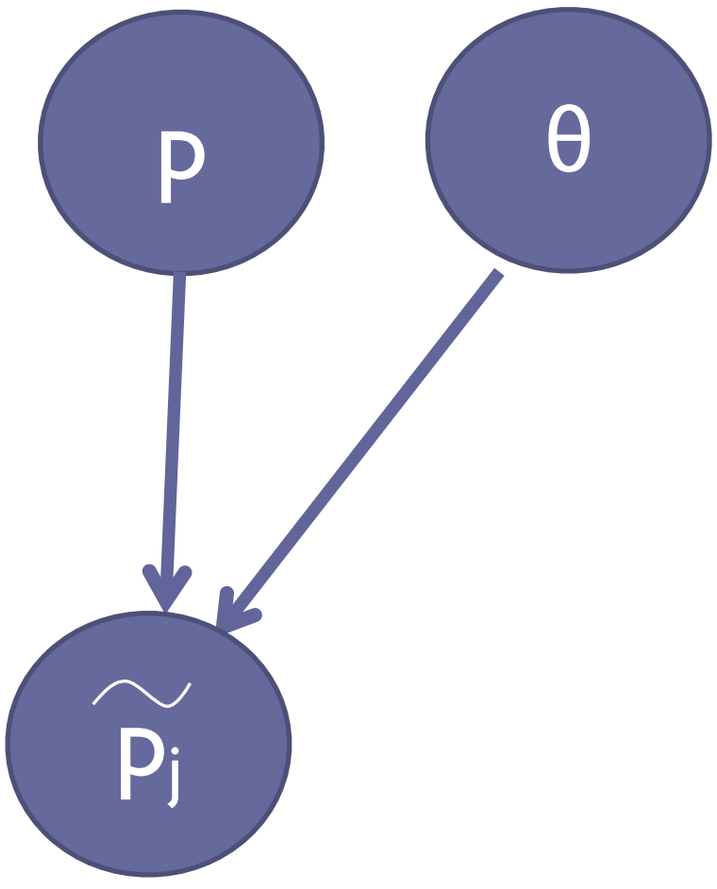
$\theta = 4Nm$

Sewall Wright, 1931

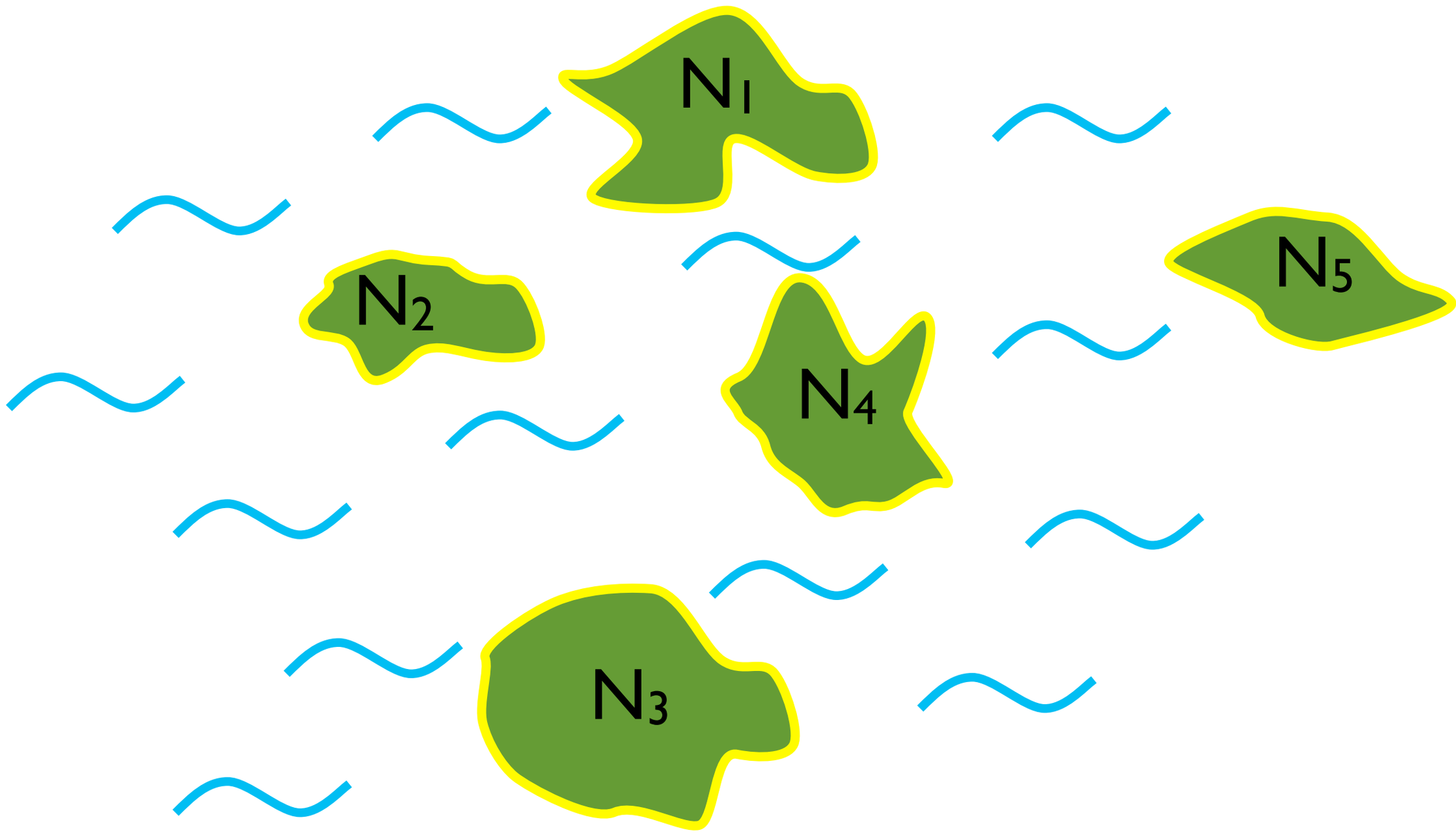
$\widetilde{p} \sim \text{Beta}(\theta p, \theta(1-p))$

$\theta = 4Nm$

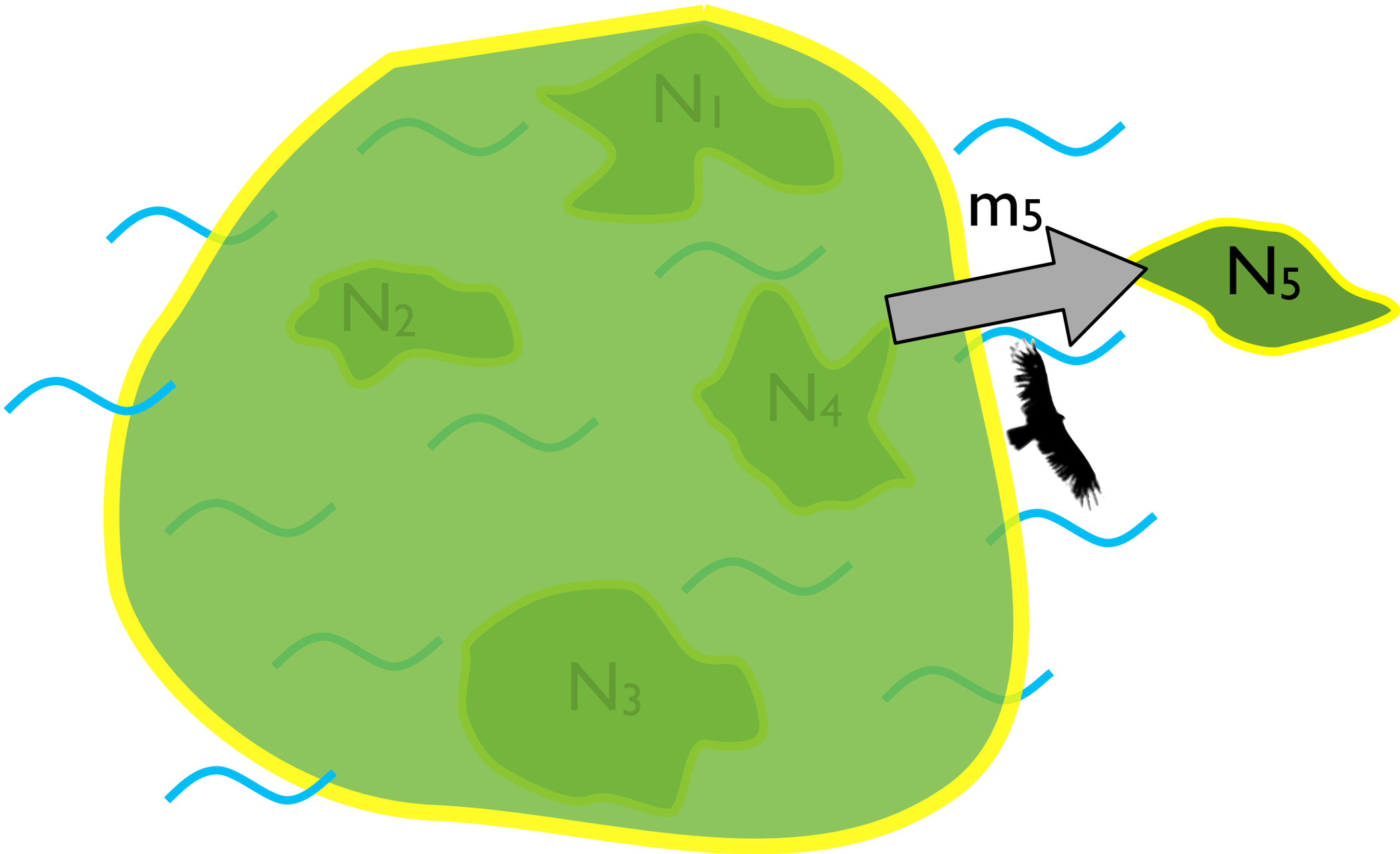
$F_{ST} = 1/(1+\theta)$



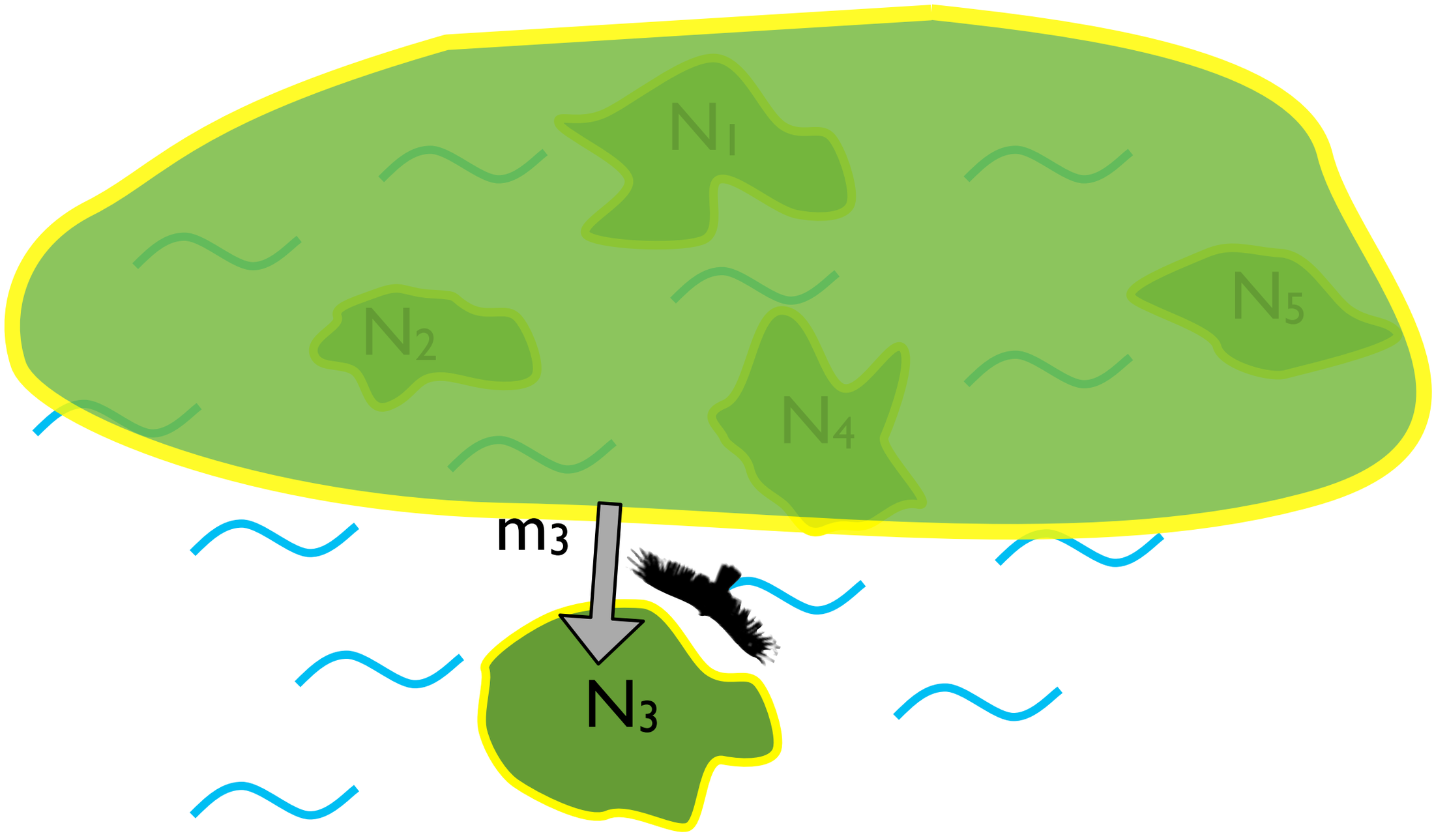
Infinite island approximation



Infinite island approximation

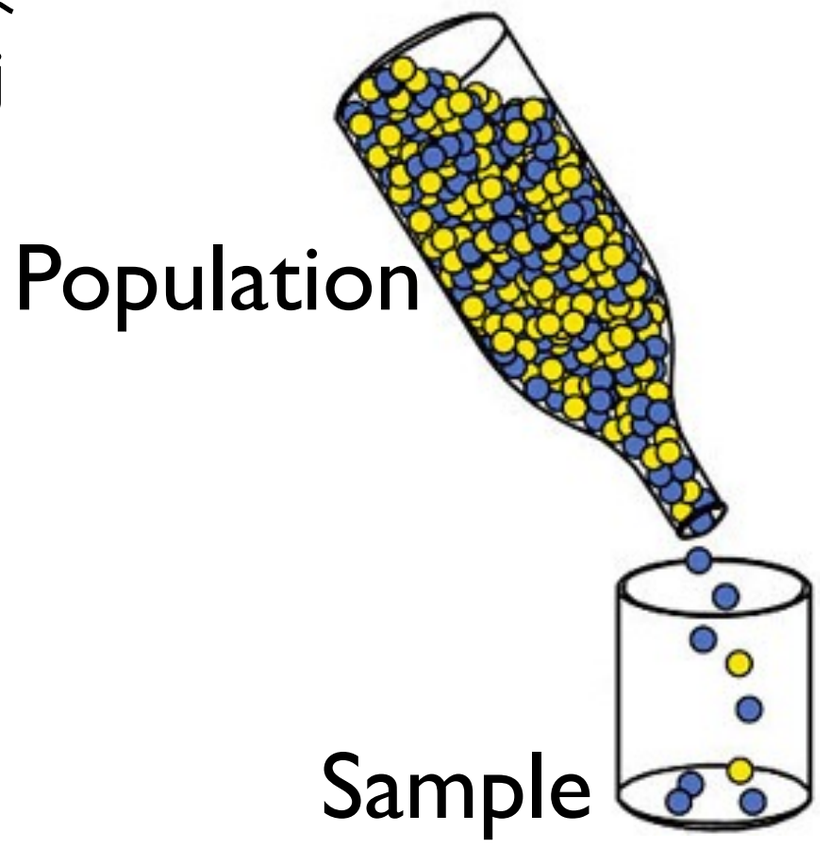


Infinite island approximation



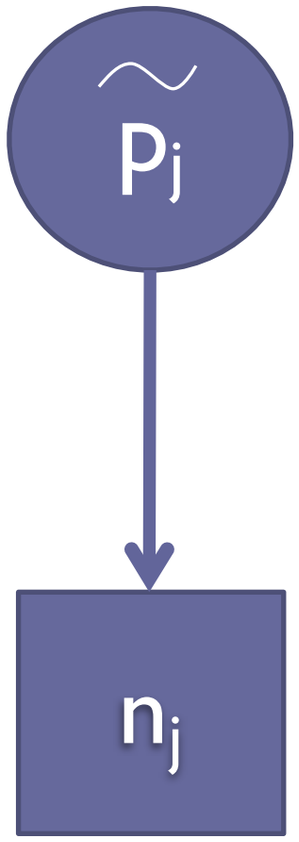
Observed data

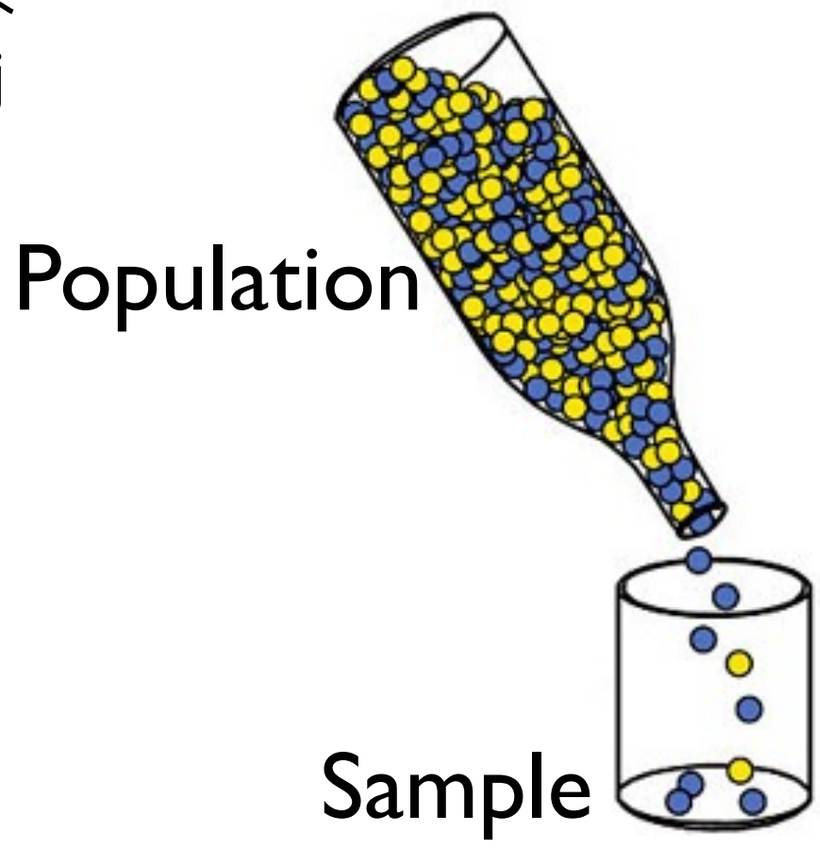
- In each population j:
  - Two alleles: A with frequency $\widetilde{p_j}$ and a 1 - $\widetilde{p_j}$
  - We sample n_j alleles randomly
  - Number of A is binomial with parameters n_j et $\widetilde{p_j}$

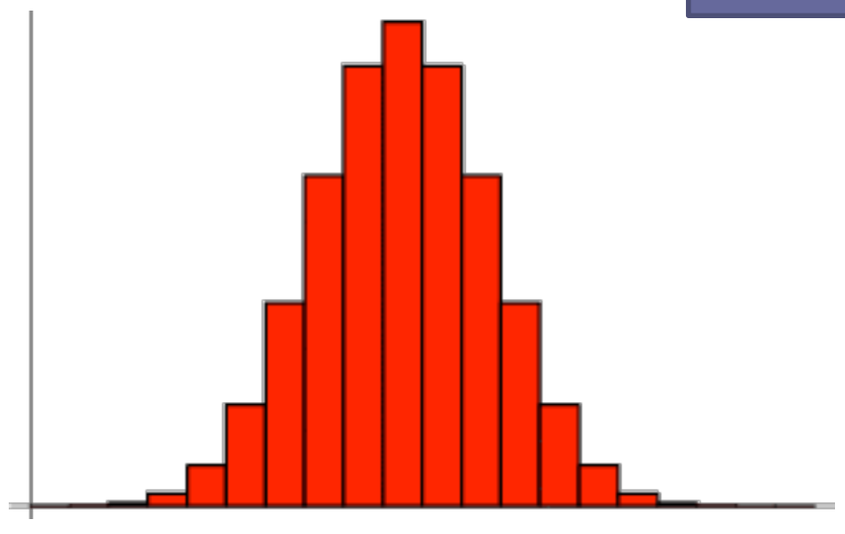


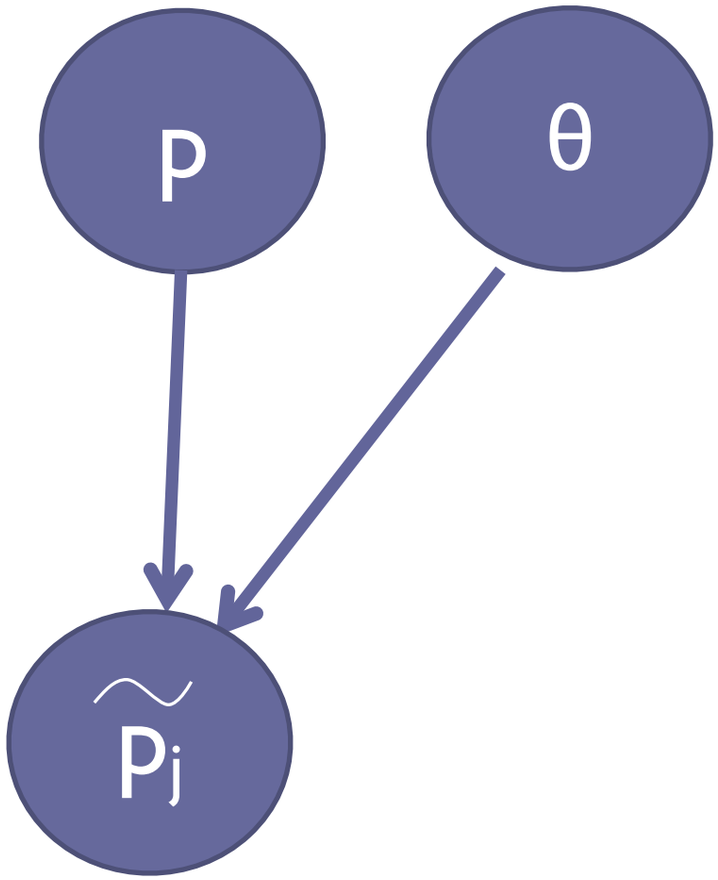
Observed data

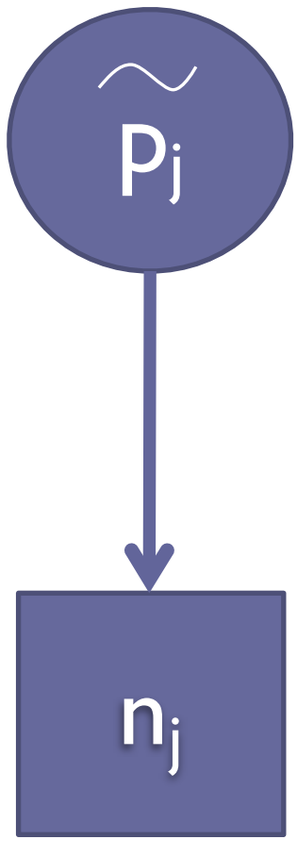
- In each population j:
  - Two alleles: A with frequency $\widetilde{p_j}$ and a 1 - $\widetilde{p_j}$
  - We sample n_j alleles randomly
  - Number of A is binomial with parameters n_j et $\widetilde{p_j}$

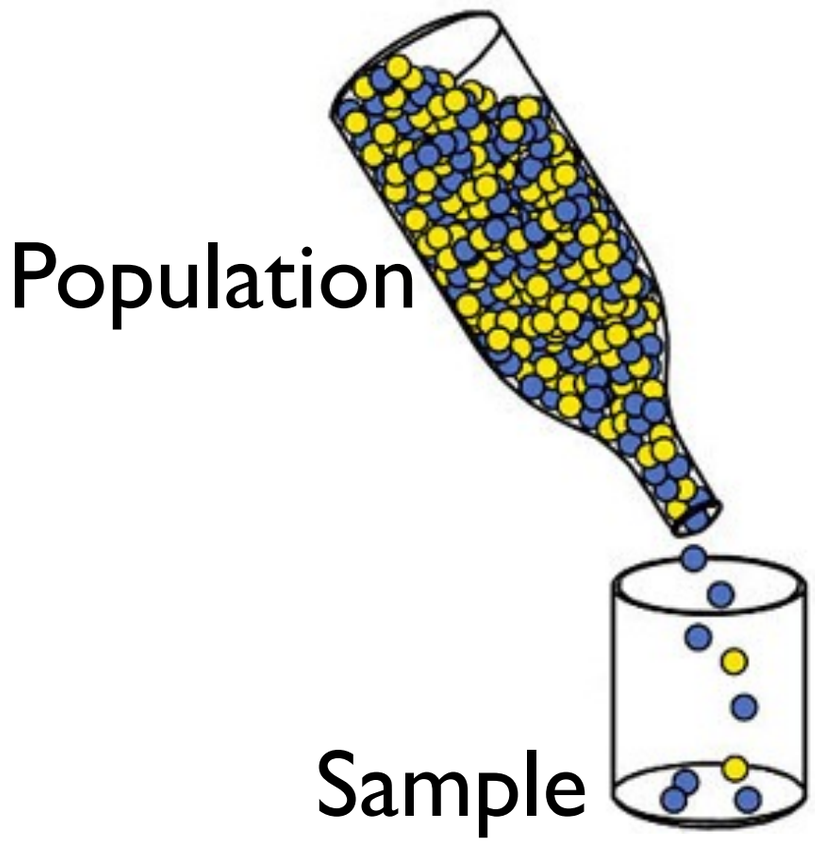


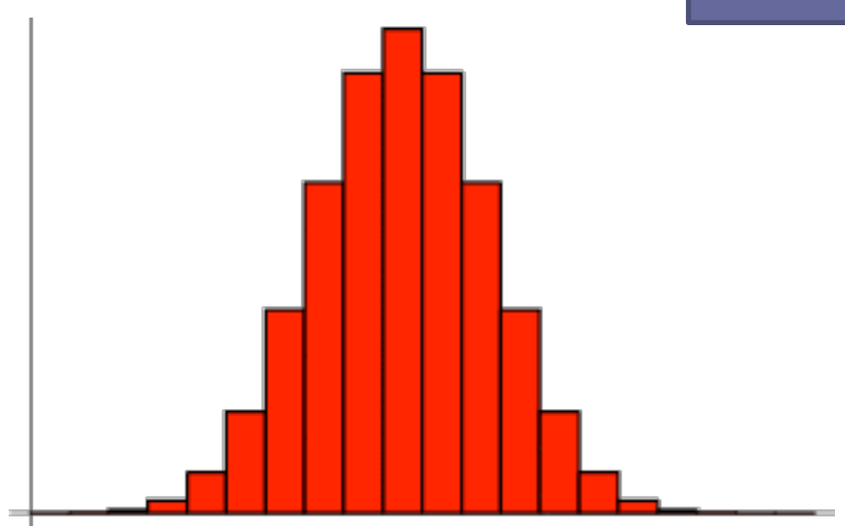




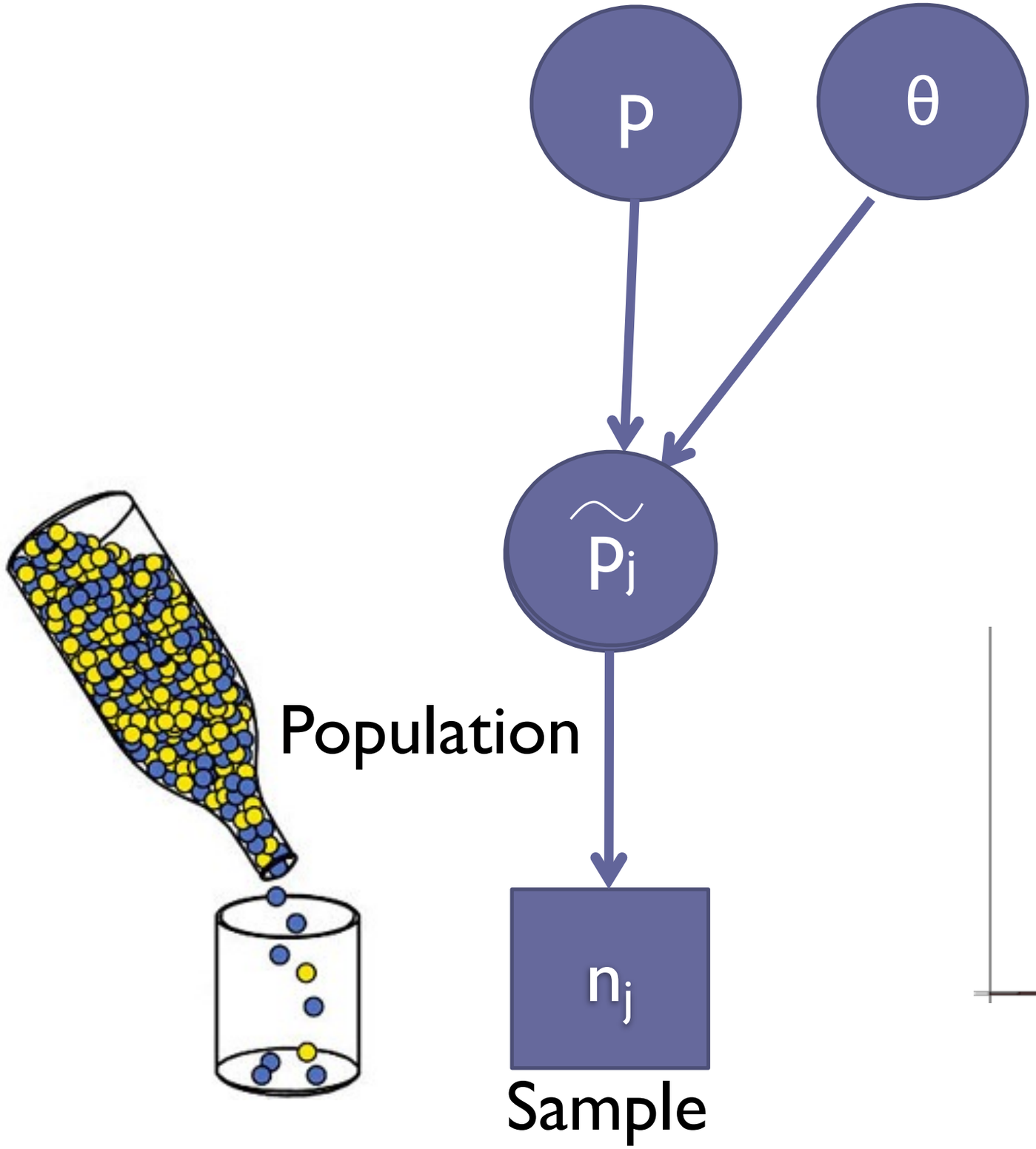

P
θ
$\tilde{P}_j$

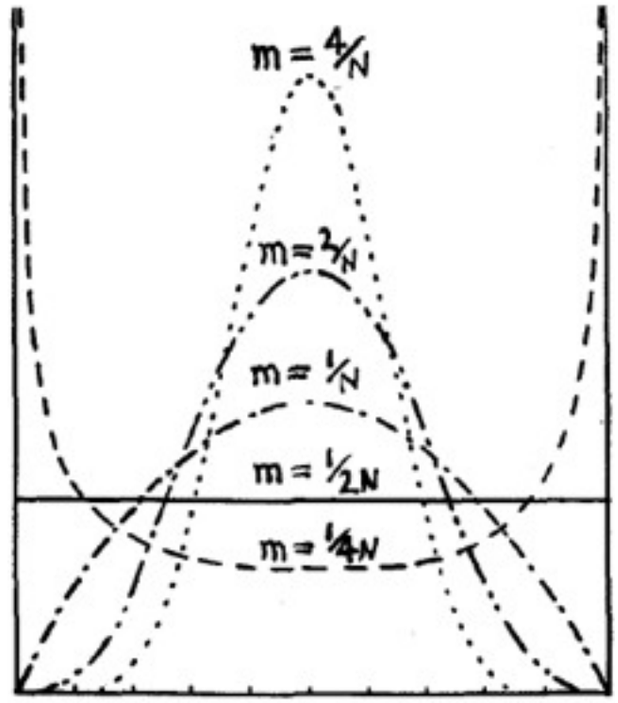

$\tilde{P}_j$
n_j

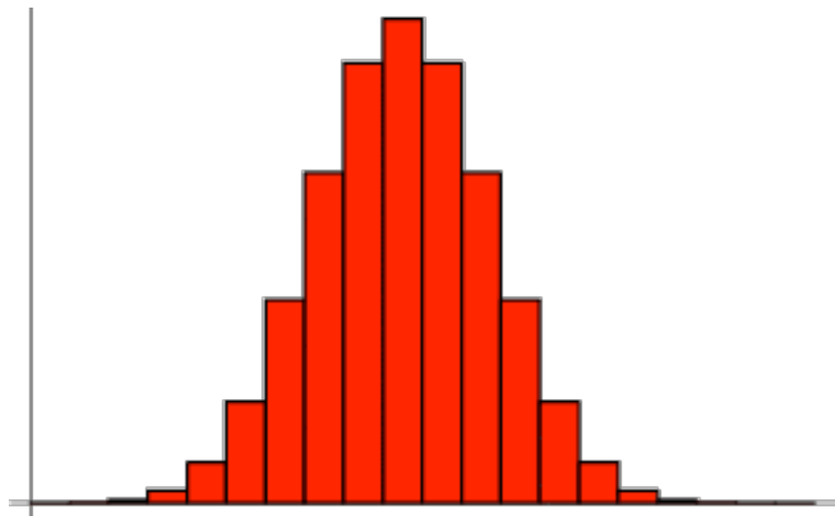

Population
Sample



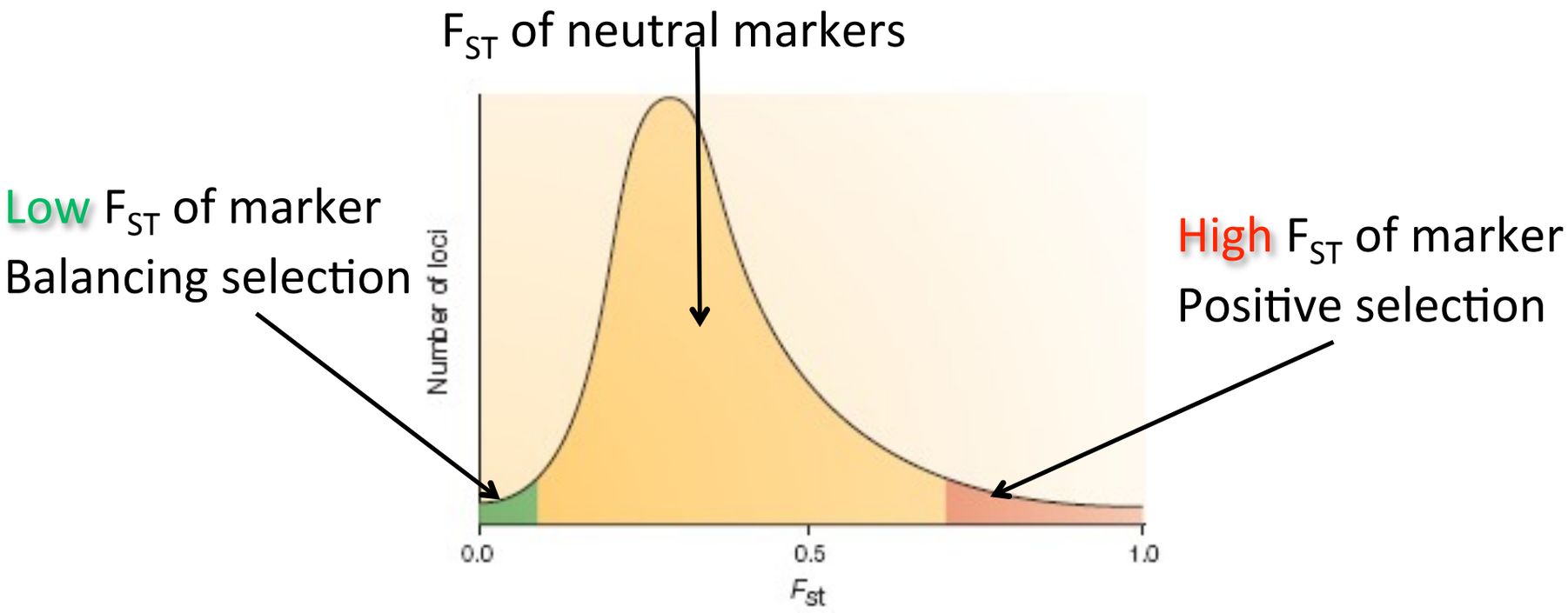
The F-model







Identifying selection



Identifying selection

- Decompose genome wide and locus specific effects
- For population j and locus i:

$$\log(\theta_{ij}) = \beta_j + \alpha_i$$

$$\theta = F_{ST}/(1-F_{ST})$$

Beaumont and Balding 2004

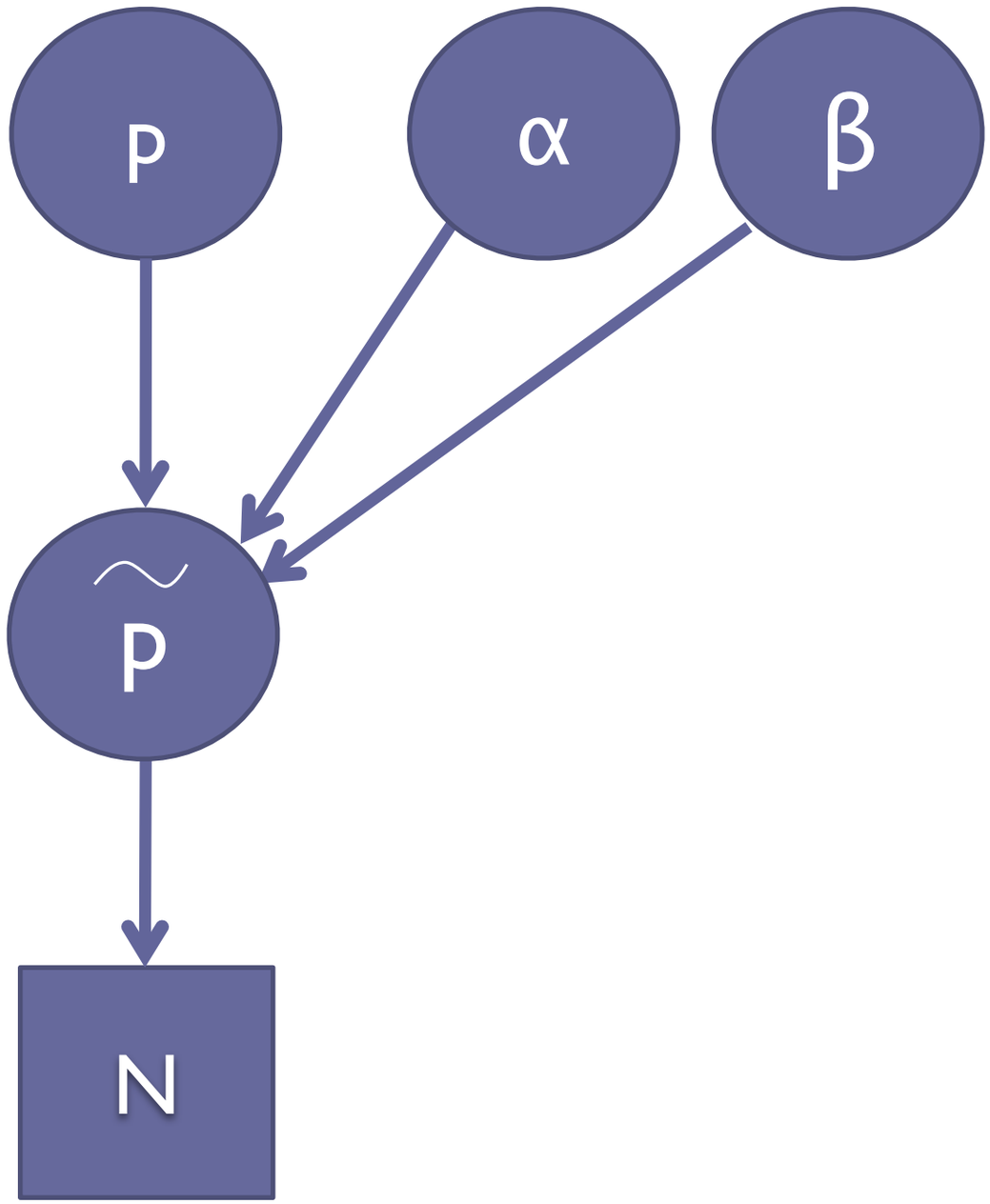
Identifying selection

- Decompose genome wide and locus specific effects

- For population j and locus i:

$$\log(\theta_{ij}) = \beta_j + \alpha_i$$

$$\theta = F_{ST}/(1-F_{ST})$$



Beaumont and Balding 2004

Extension

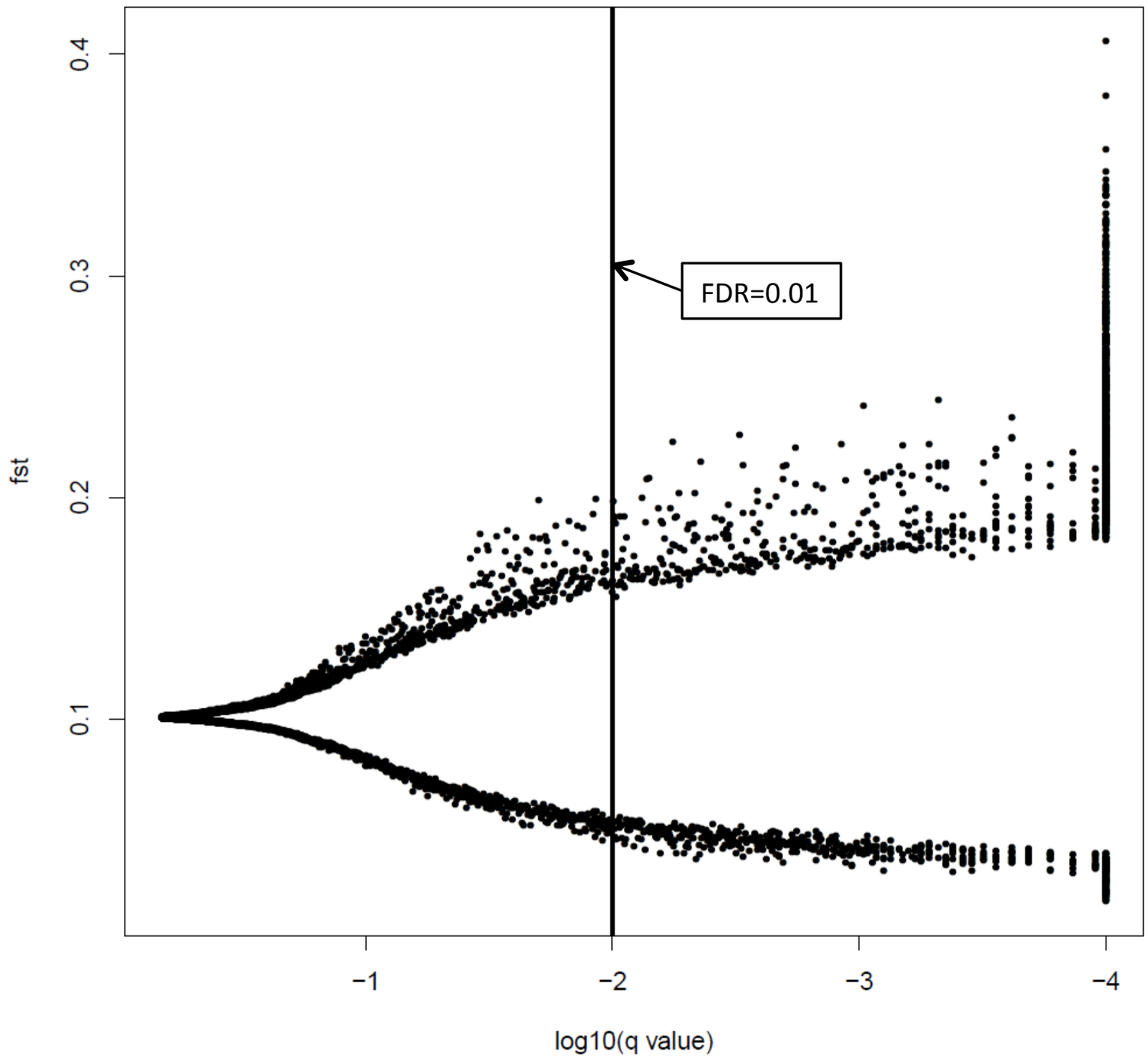
- Bayesian model choice, with two alternative models:
  - M_N: $\log(\theta_{ij})=\beta_j \quad (\alpha_i=0)$
  - M_S: $\log(\theta_{ij})=\beta_j+\alpha_i$
- Estimates probability of both models for each locus using RJ-MCMC

$$p = P(M_S|D) \text{ and } P(M_N|D)=1-p$$

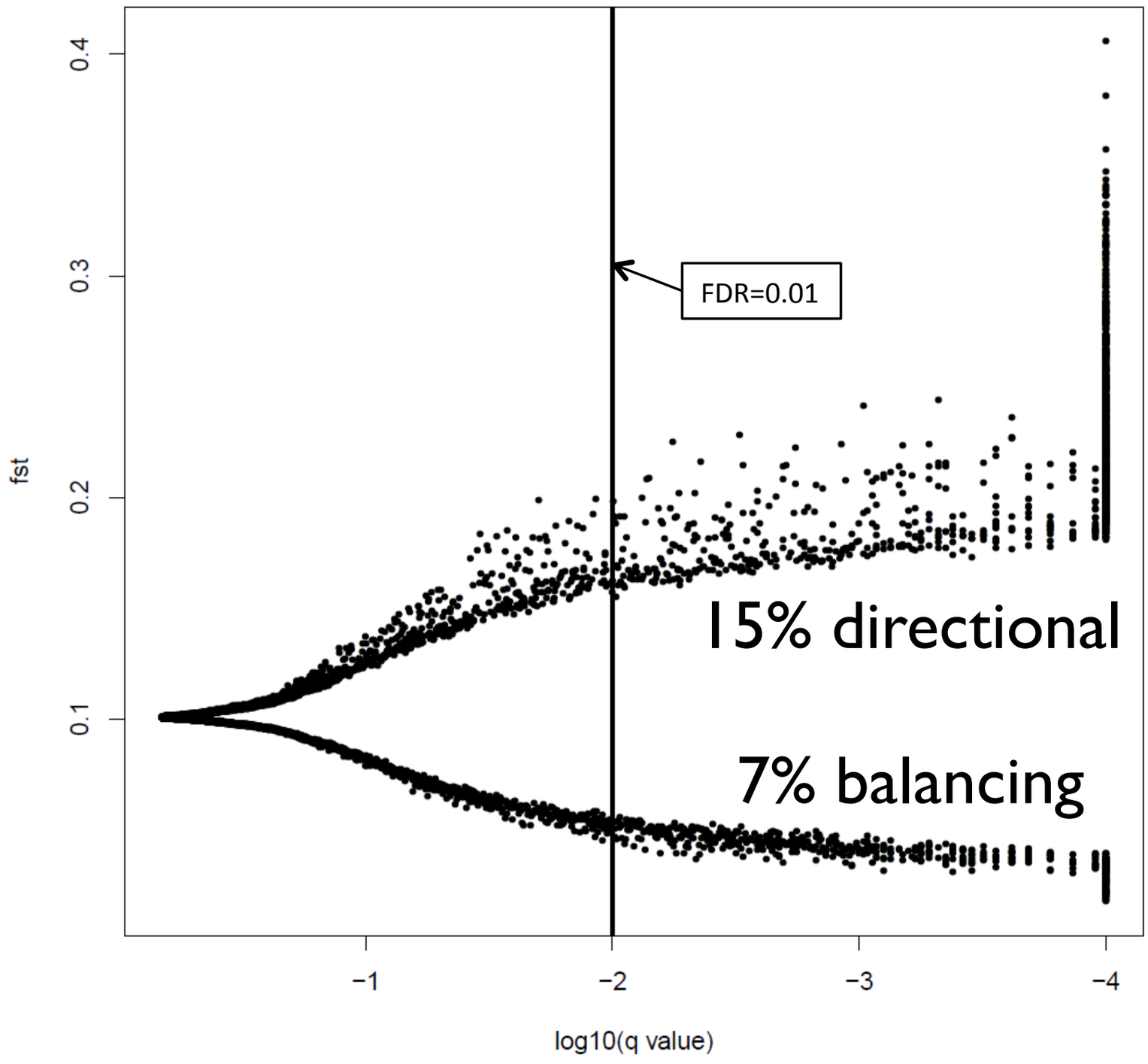
- Implemented in software BayeScan

Foll and Gaggiotti 2008

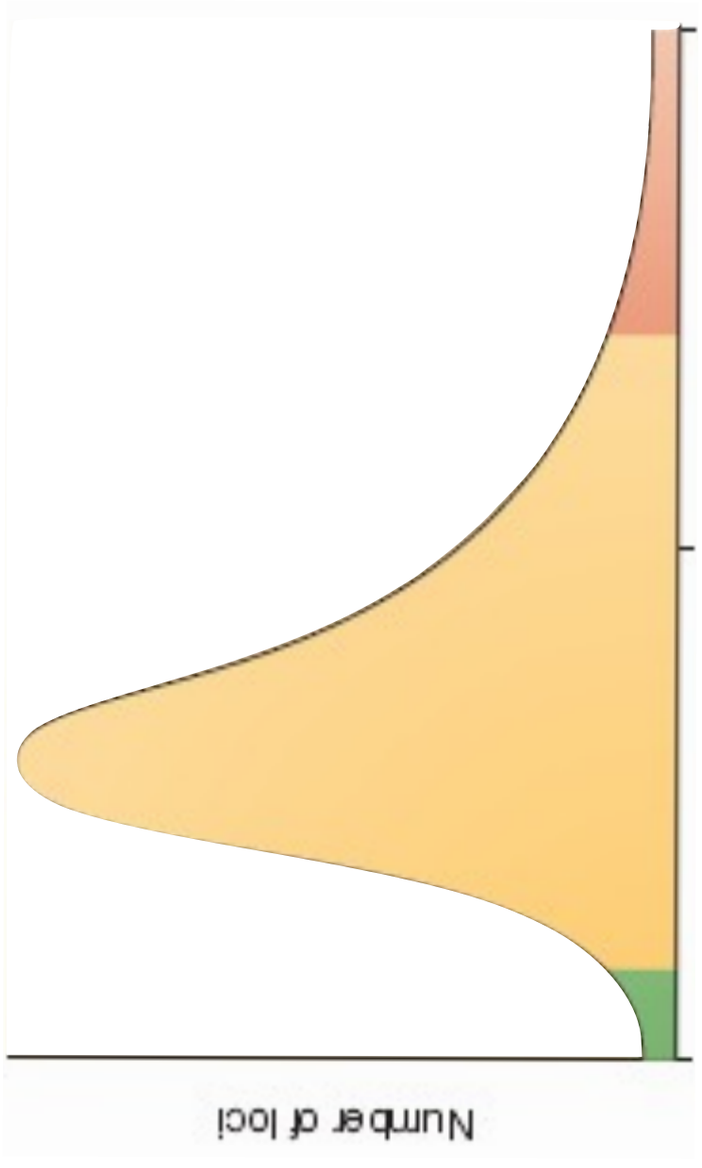
Results: chromosome 19

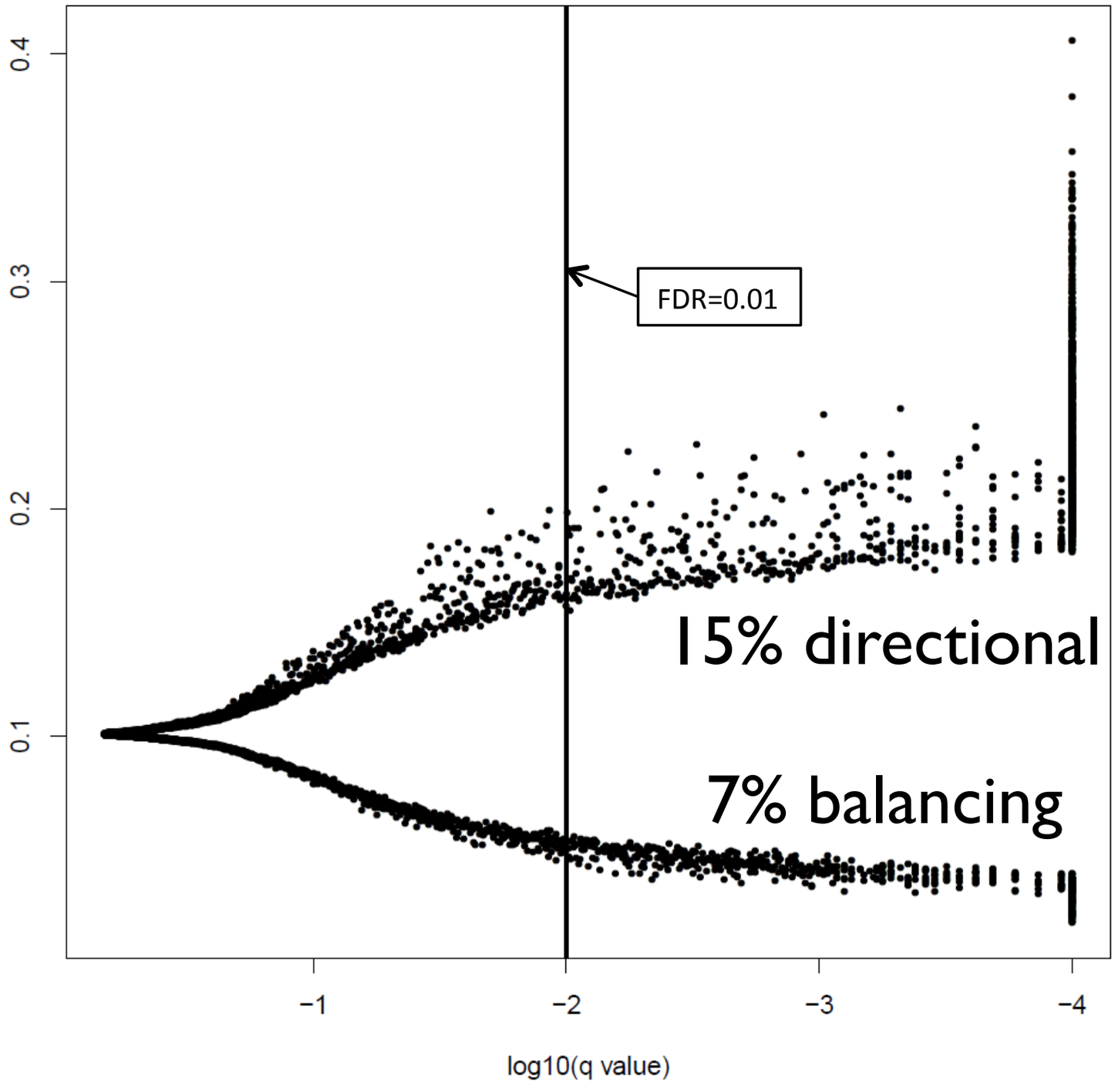


Results: chromosome 19

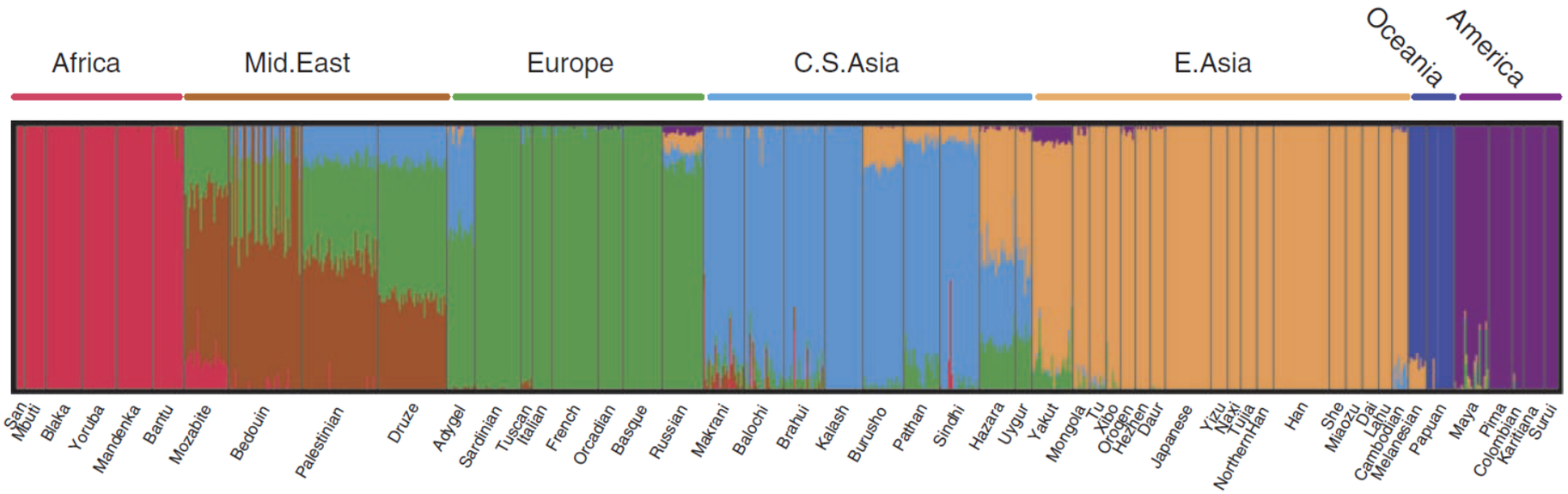


Results: chromosome 19



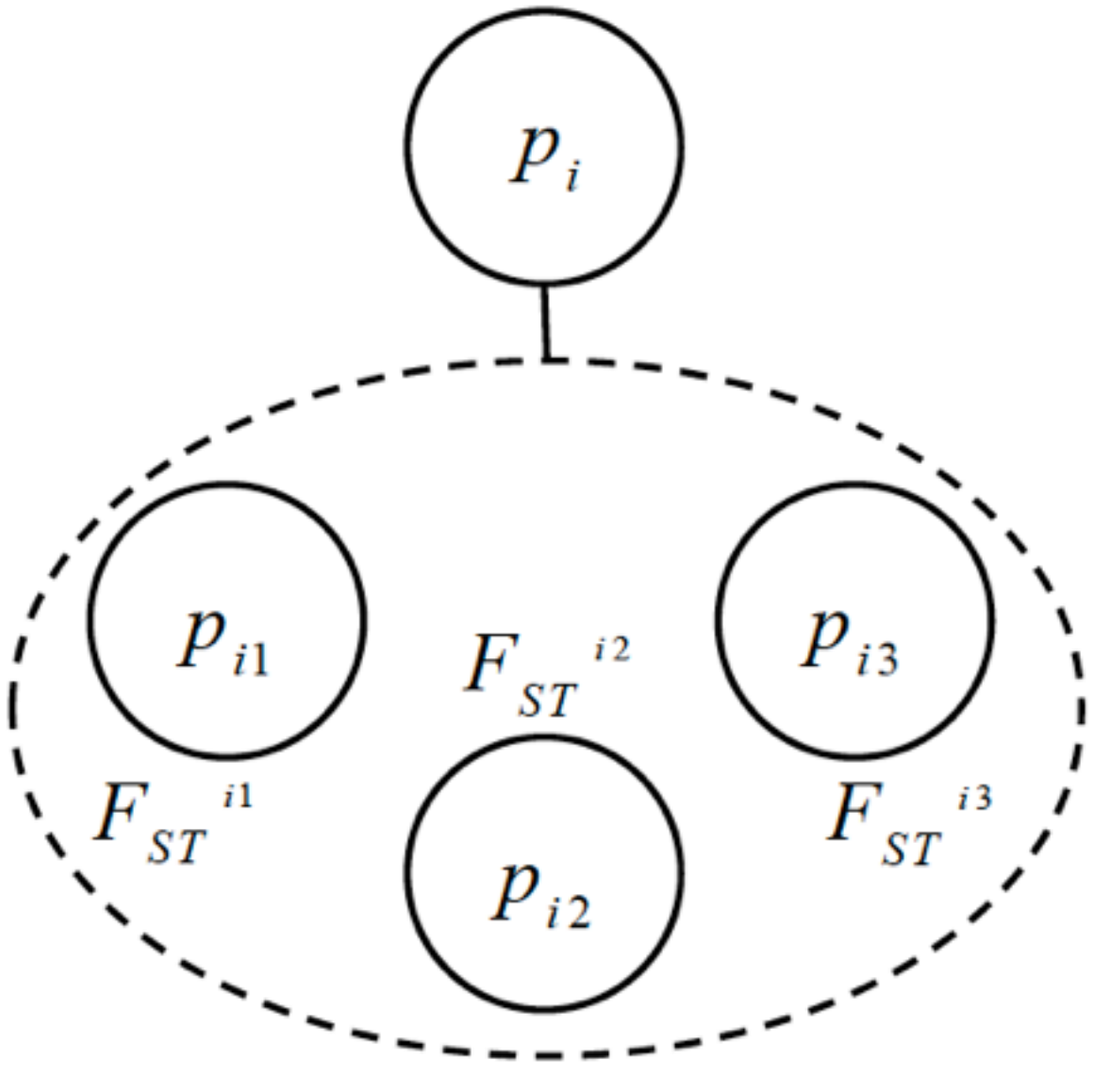


Hierachical population structure

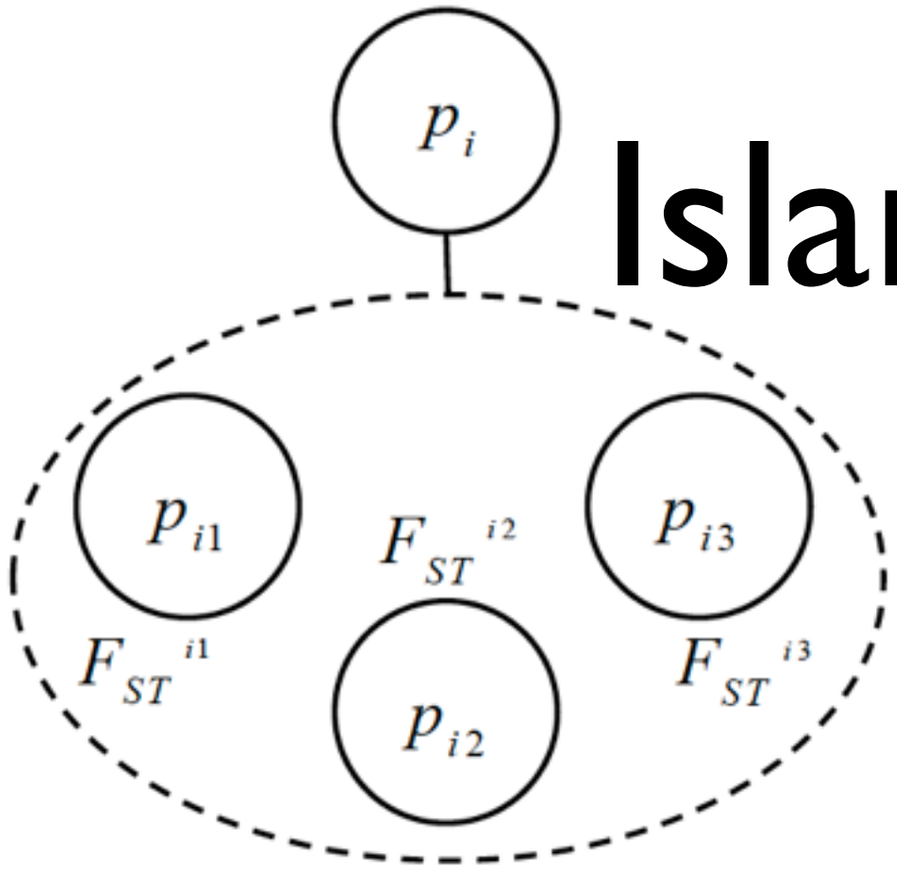


Li *et al.* 2008

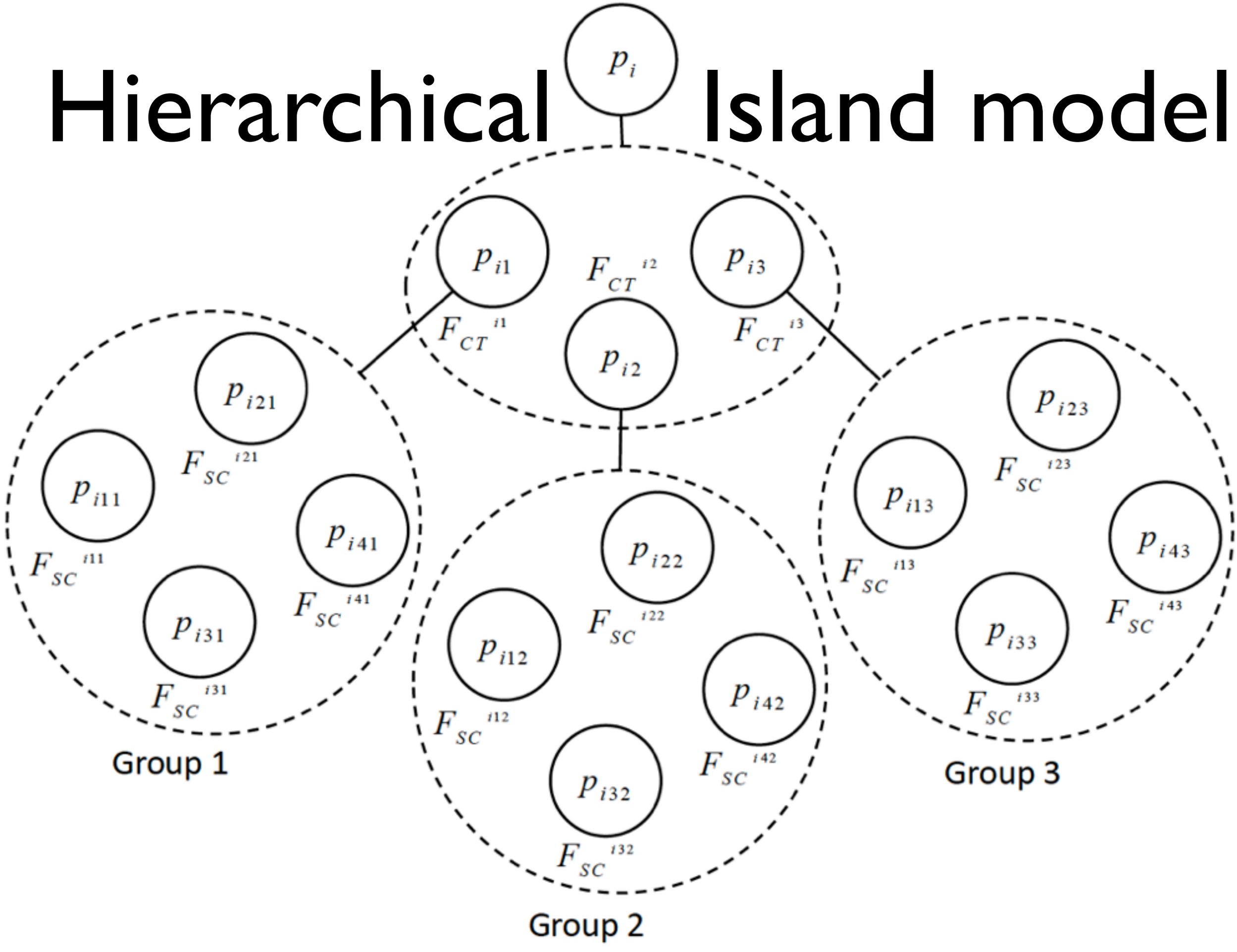
Island model



Island model

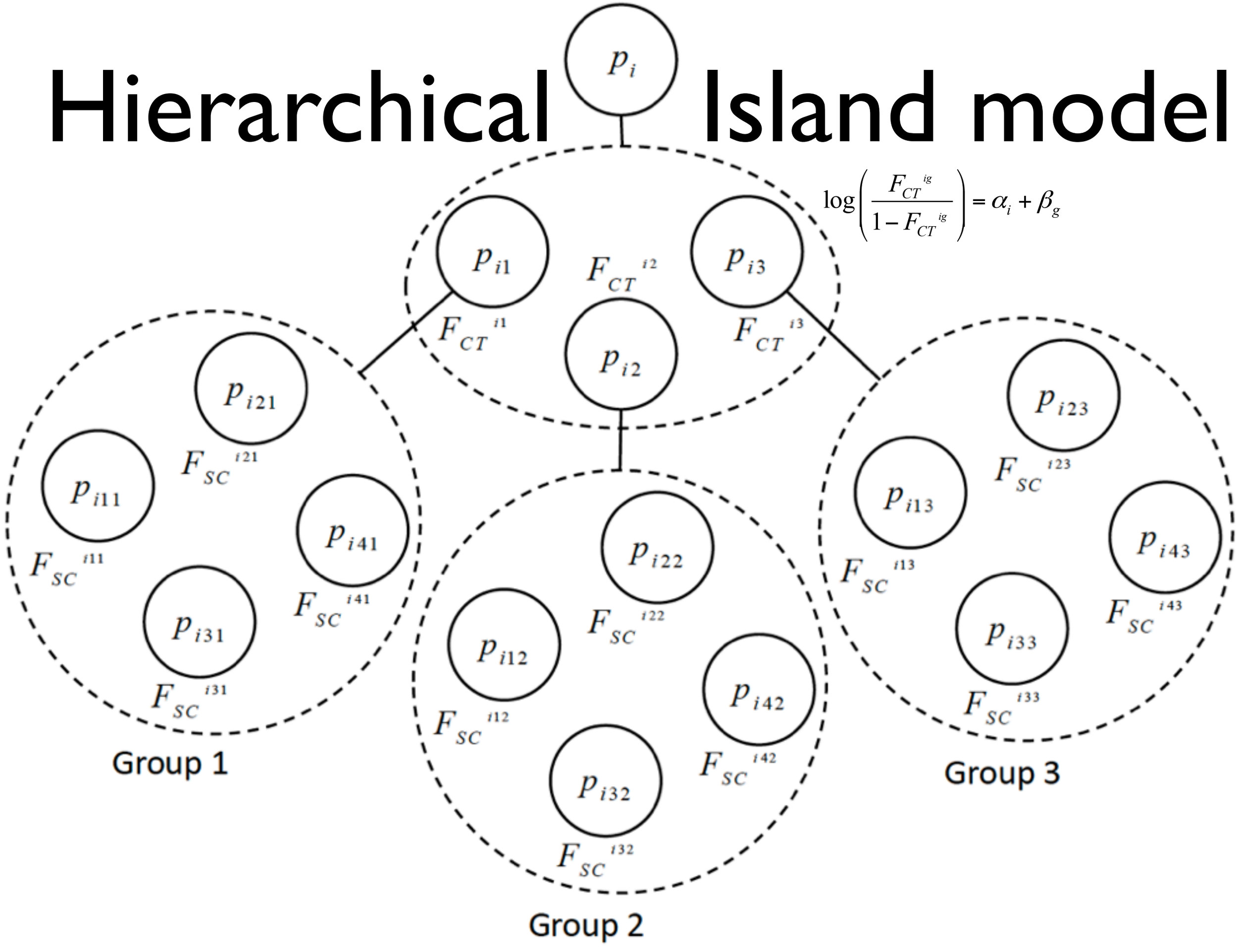


Hierarchical Island model

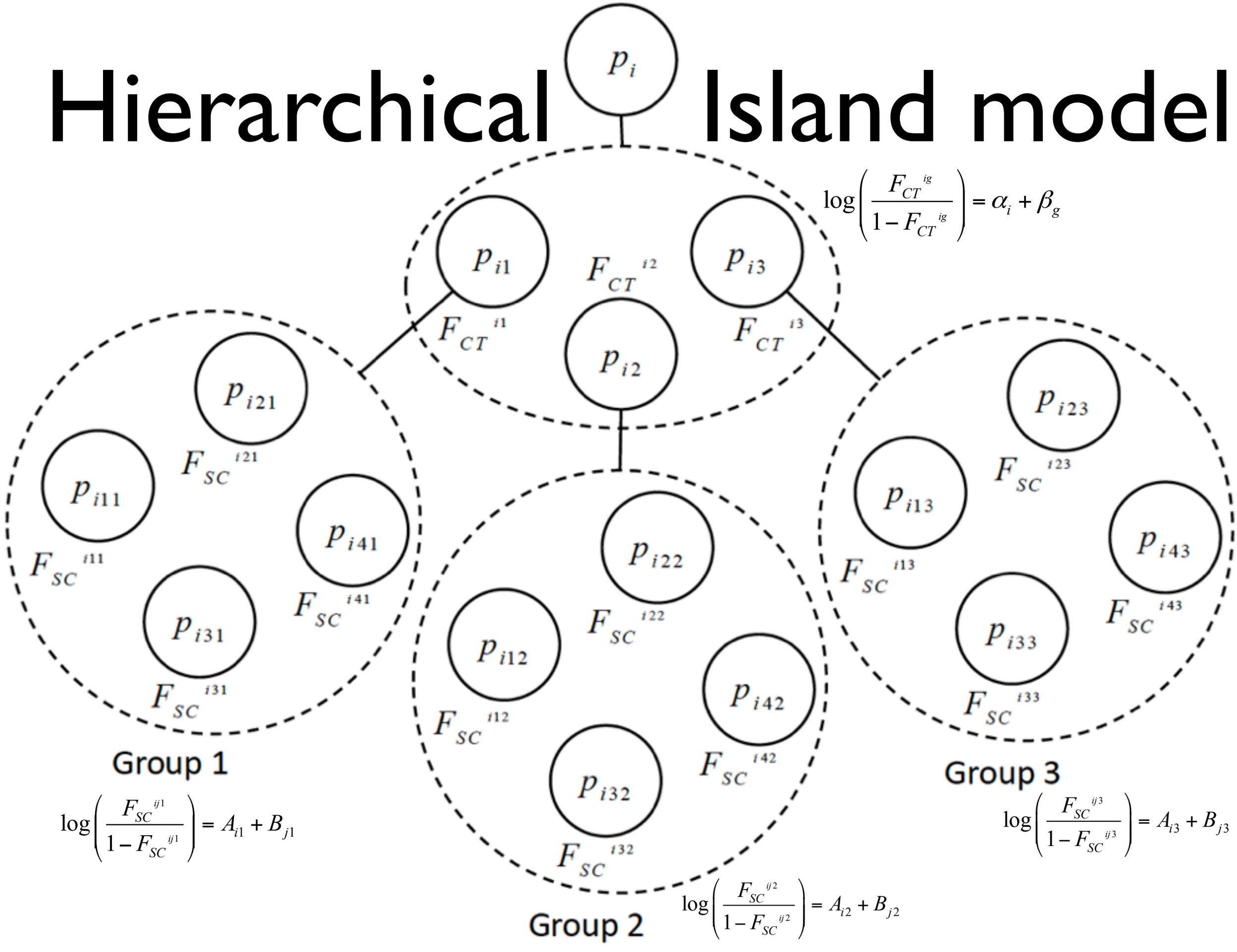


Hierarchical Island model

$$\log\left(\frac{F_{CT}^{ig}}{1-F_{CT}^{ig}}\right)=\alpha_i+\beta_g$$



Hierarchical Island model



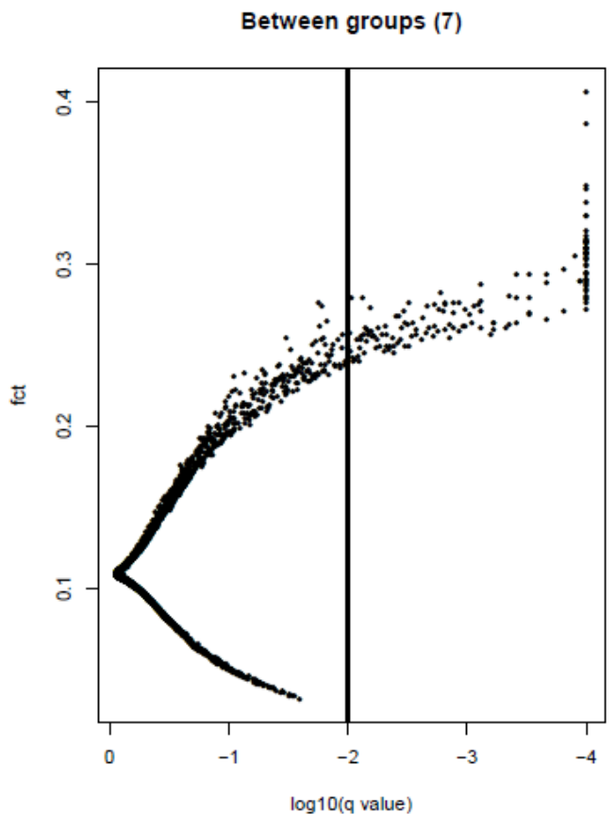
$$\log\left(\frac{F_{CT}^{ig}}{1-F_{CT}^{ig}}\right) = \alpha_i + \beta_g$$

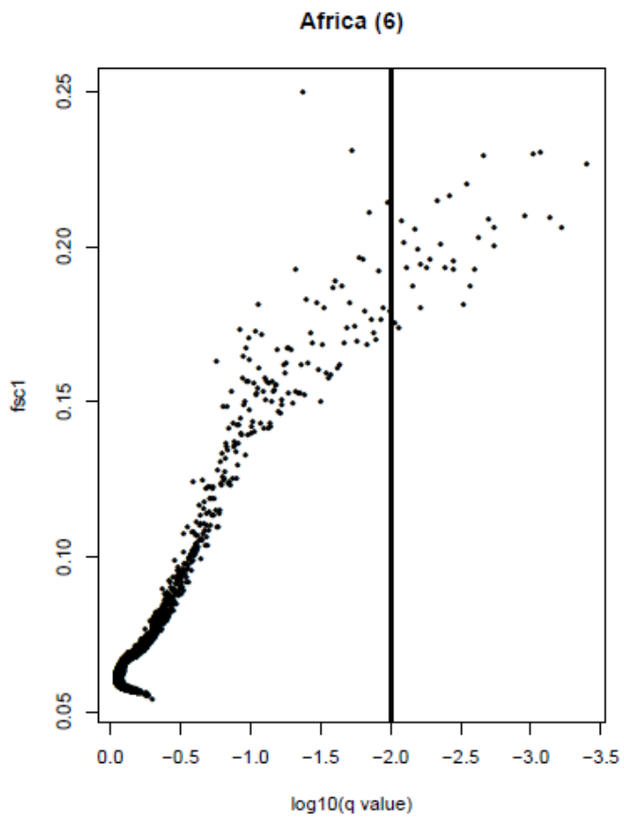
$$\log\left(\frac{F_{SC}^{ij1}}{1-F_{SC}^{ij1}}\right) = A_{i1} + B_{j1}$$

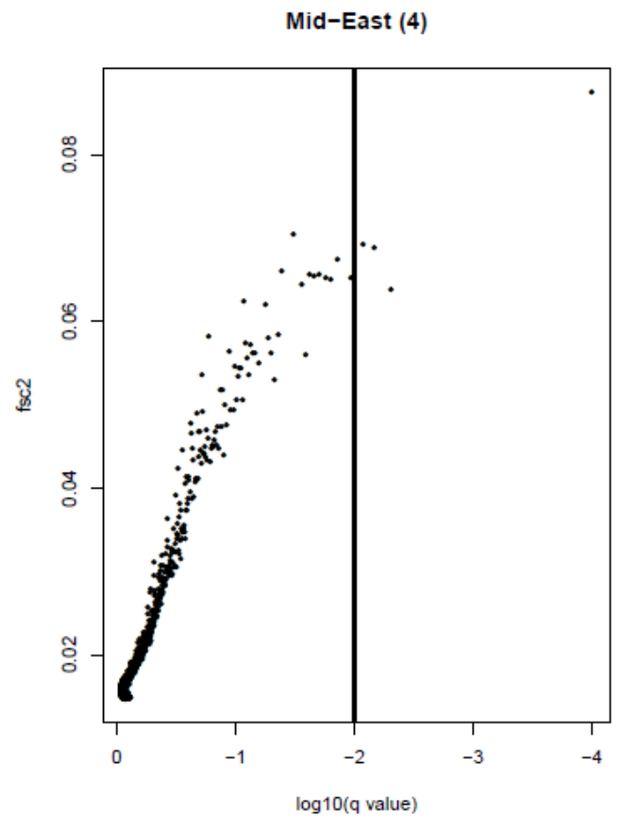
$$\log\left(\frac{F_{SC}^{ij2}}{1-F_{SC}^{ij2}}\right) = A_{i2} + B_{j2}$$

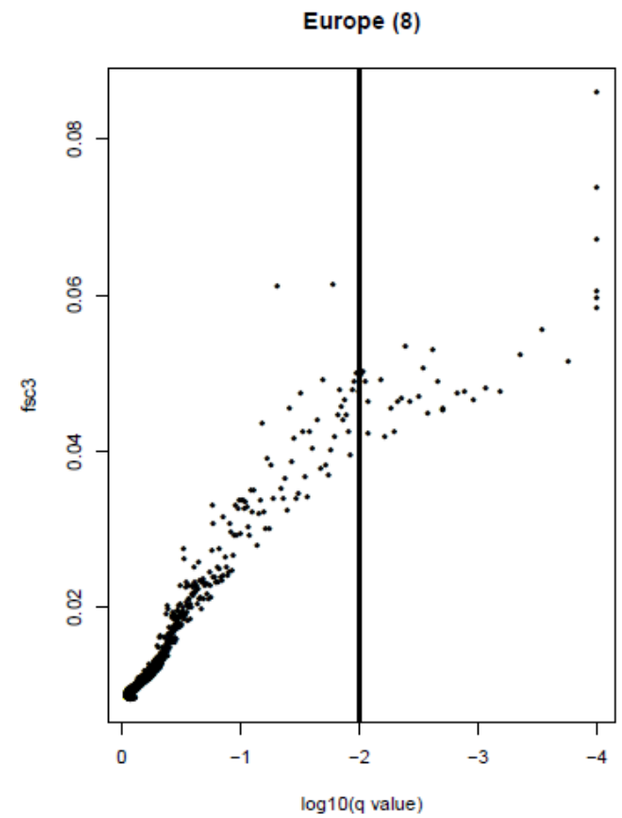
$$\log\left(\frac{F_{SC}^{ij3}}{1-F_{SC}^{ij3}}\right) = A_{i3} + B_{j3}$$

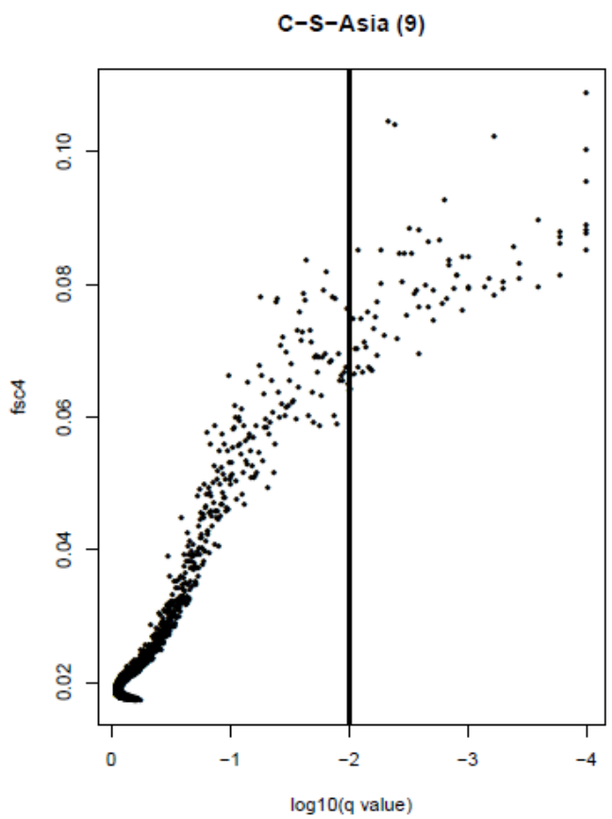
Results: chromosome 19

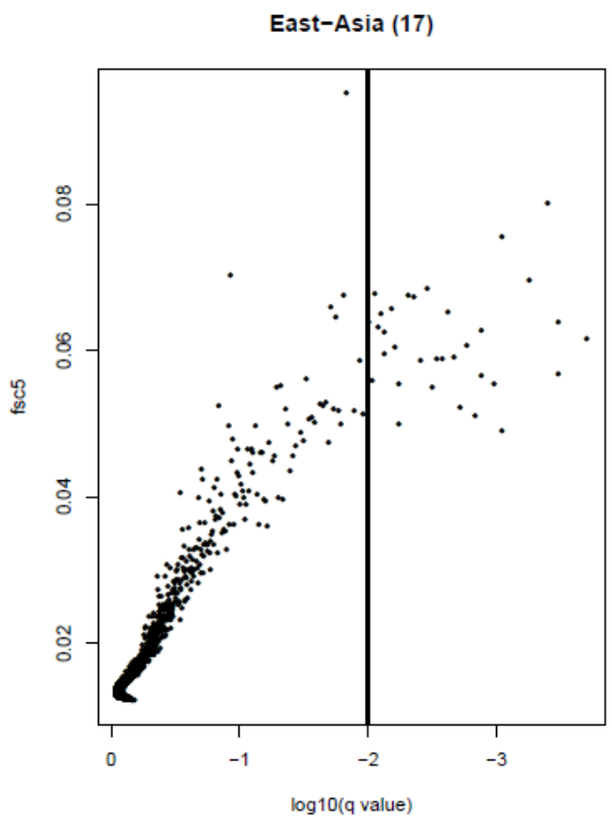


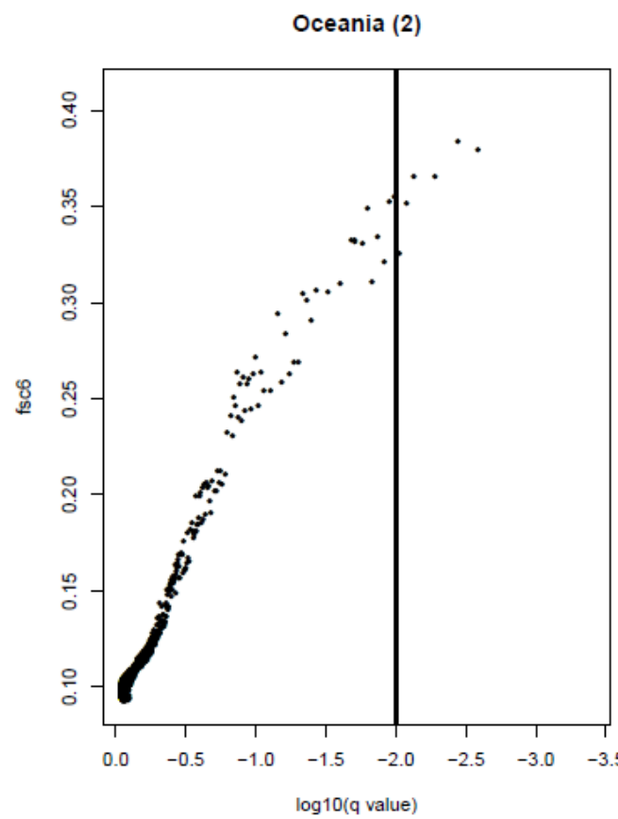


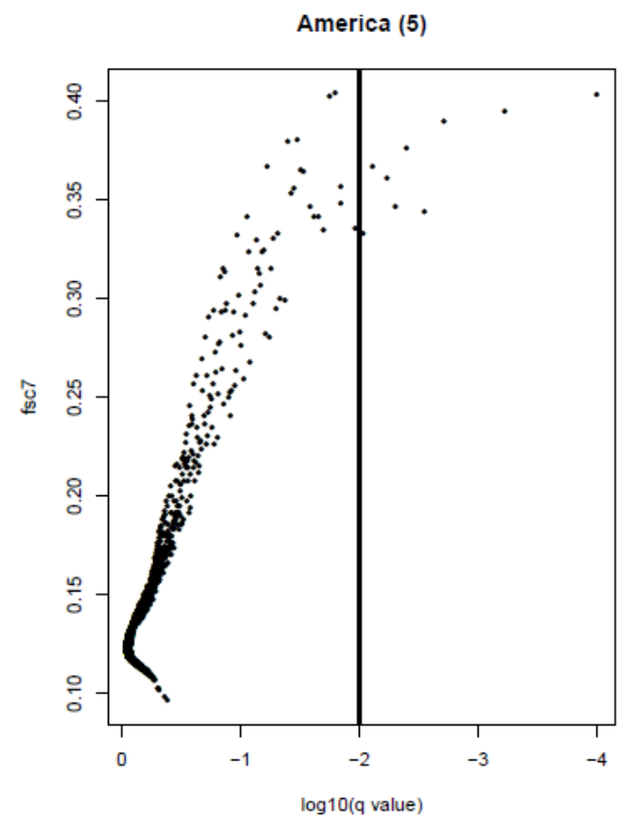












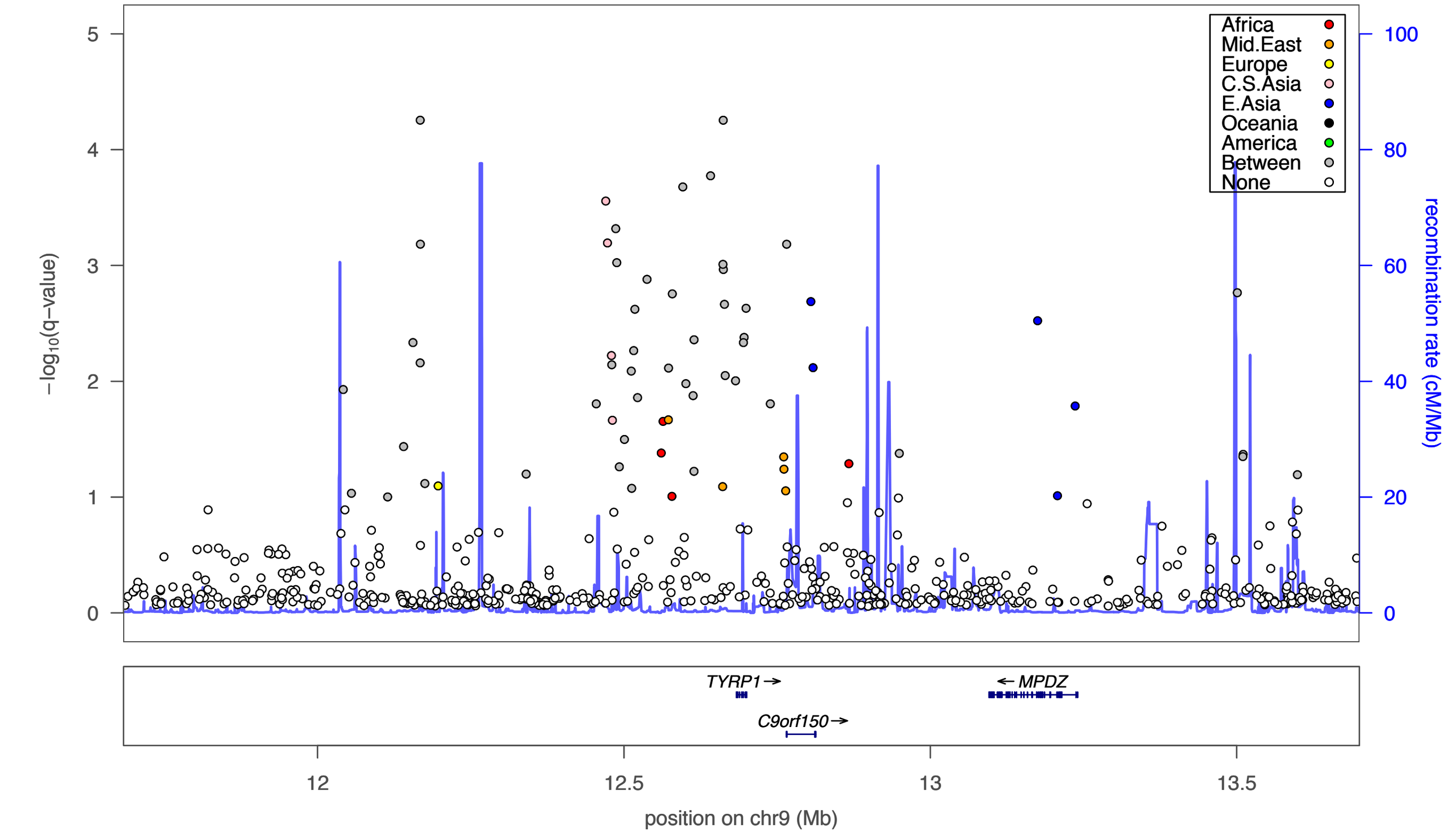
Results: chromosome 19

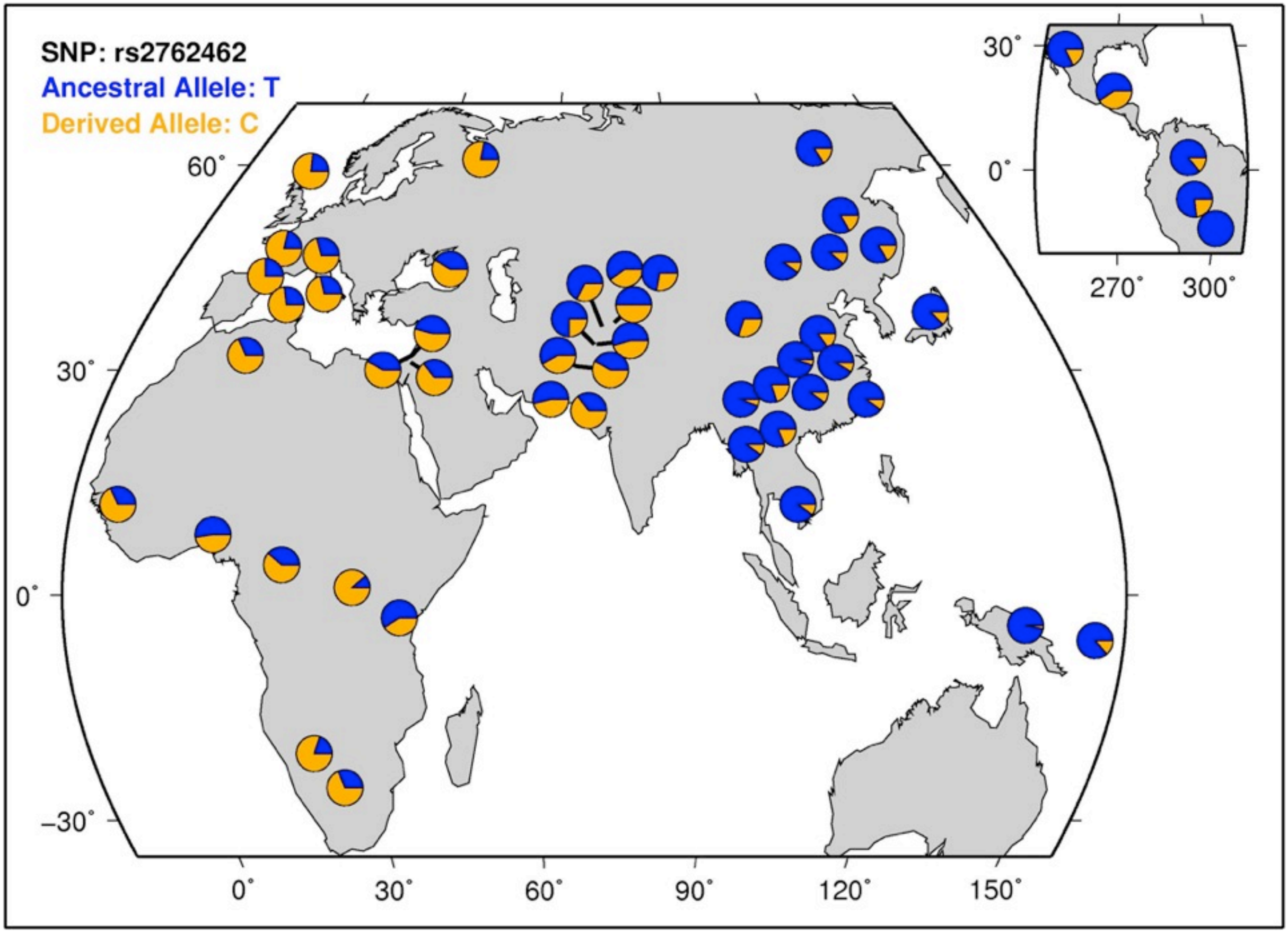
1.7% of SNPs under directional selection at FDR=0.01

no balancing selection

TYRP1 chr12 (pigmentation)

Plotted SNPs




SNP: rs2762462
Ancestral Allele: T
Derived Allele: C
60˚
30˚
0˚
−30˚
0˚
30˚
60˚
90˚
120˚
150˚
30˚
0˚
270˚
300˚

PIK3R3 chr1 (Pygmy size)

Plotted SNPs

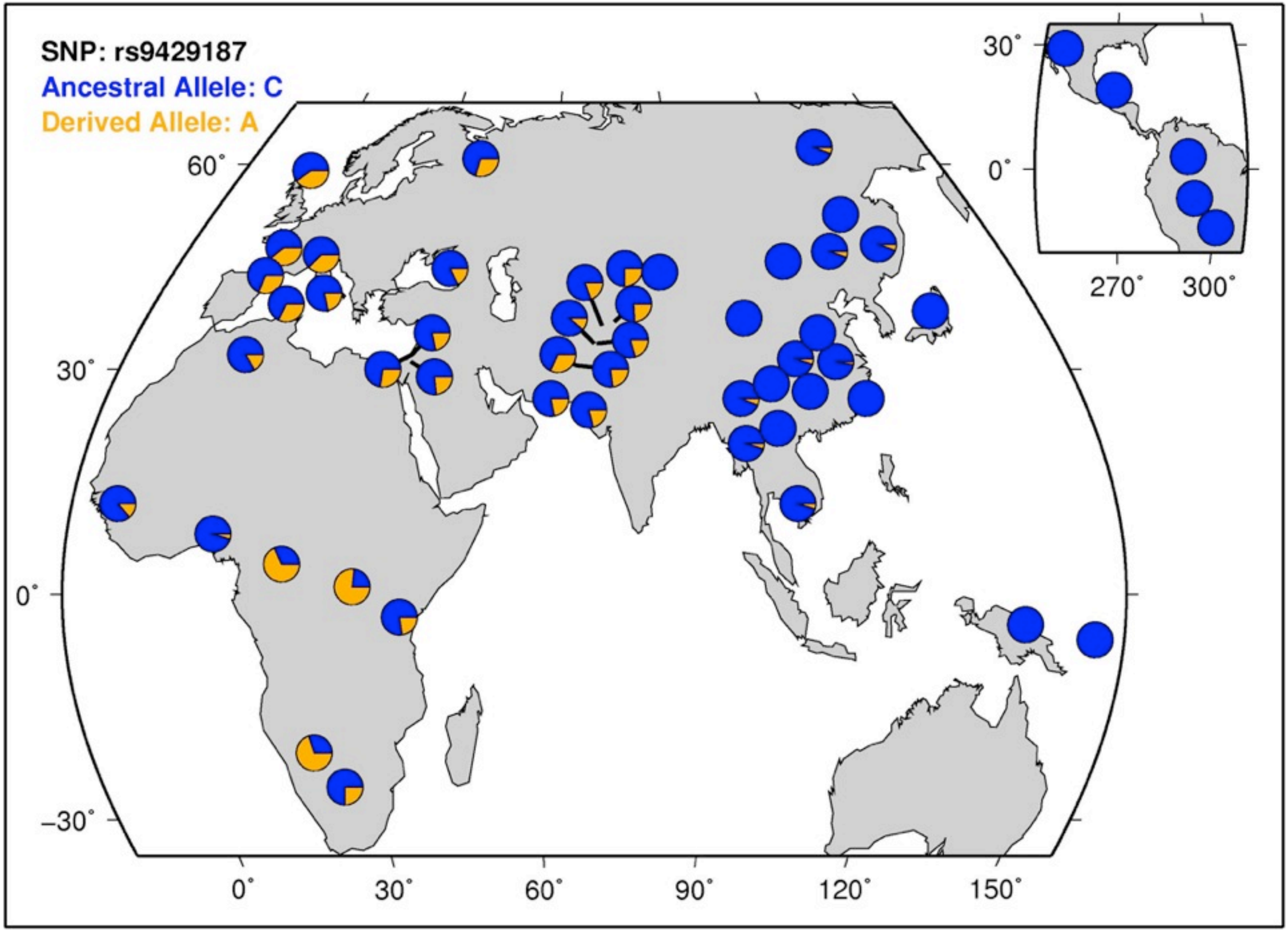
$-\log_{10}$(q−value)

recombination rate (cM/Mb)

Africa
Mid.East
Europe
C.S.Asia
E.Asia
Oceania
America
Between
None

ZSWIM5 HPDL MMACHC IPP MAST2 TSPAN1 NSUN4 KNCN CYP4B1 CYP4X1
LOC400752 PRDX1 TMEM69 PIK3R3 RAD54L DMBX1 KIAA0494 CYP4A11
MUTYH AKR1A1 POMGNT1 FAAH MKNK1 CYP4Z2P
TOE1 NASP C1orf190 MOBKL2C CYP4Z1
TESK2 CCDC17 LRRC41 ATPAF1
CCDC163P UQCRH C1orf223
GPBP1L1
RPS15AP10

position on chr1 (Mb)


SNP: rs9429187
Ancestral Allele: C
Derived Allele: A
60˚
30˚
0˚
−30˚
0˚
30˚
60˚
90˚
120˚
150˚
30˚
0˚
270˚
300˚

Conclusion

- Taking into accounts hierarchical population structure lowers number of false significants
- Useful when some populations are evolutionarily related or geographically close
- Hierarchical genetic structure is assumed known or needs to be estimated separately
    - Slightly wrong hierarchical genetic structure leads to less false positives that assuming populations are independent
- Hierarchical model is still an approximation. Reality may be more complex and false positives may emerge
    - Spatial expansions

Practicals

- We provide a small subset of 100 SNPs chosen at random (almost...) from chromosome 16 of HGDP data
- Use Arlequin to identify SNPs under selection
- Try both island and hierarchical-island models and see the effect on the number of outlier markers identified
- Find more details about significant SNPs at the 1% level under hierachical-island model:
  - http://www.ncbi.nlm.nih.gov/SNP/
  - http://hgdp.uchicago.edu