« Environmental Genetics » doctoral course ABIES-GAIA

Footprints of Selection

Renaud Vitalis
Centre de Biologie pour la Gestion des Populations
INRA ; Montpellier
E-mail : vitalis@supagro.inra.fr

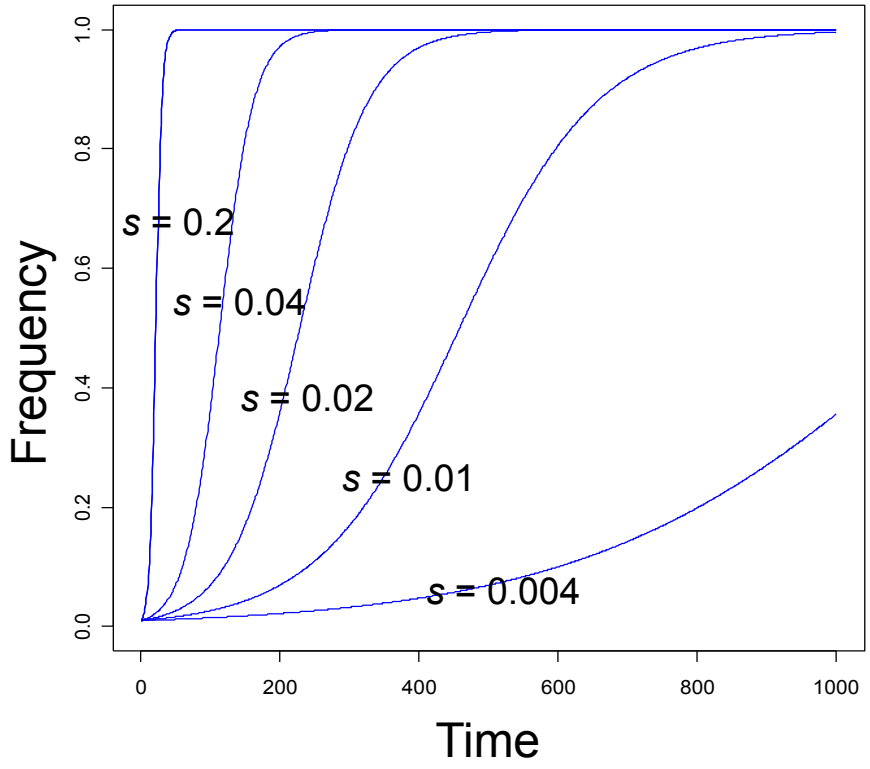
Next-generation genotyping

- Huge amounts of data in model and non-model species…
- These data contains lots of information on the evolutionary history of species…
- **One big question then is: can we distinguish demography from selection (or: what are the targets of natural, or artificial, selection?)**

How does selection act?

Selection at the molecular level

- Allele A is selected for in an infinite population (with relative fitness 1 + *s*)



- Allele frequency change as a function of time:

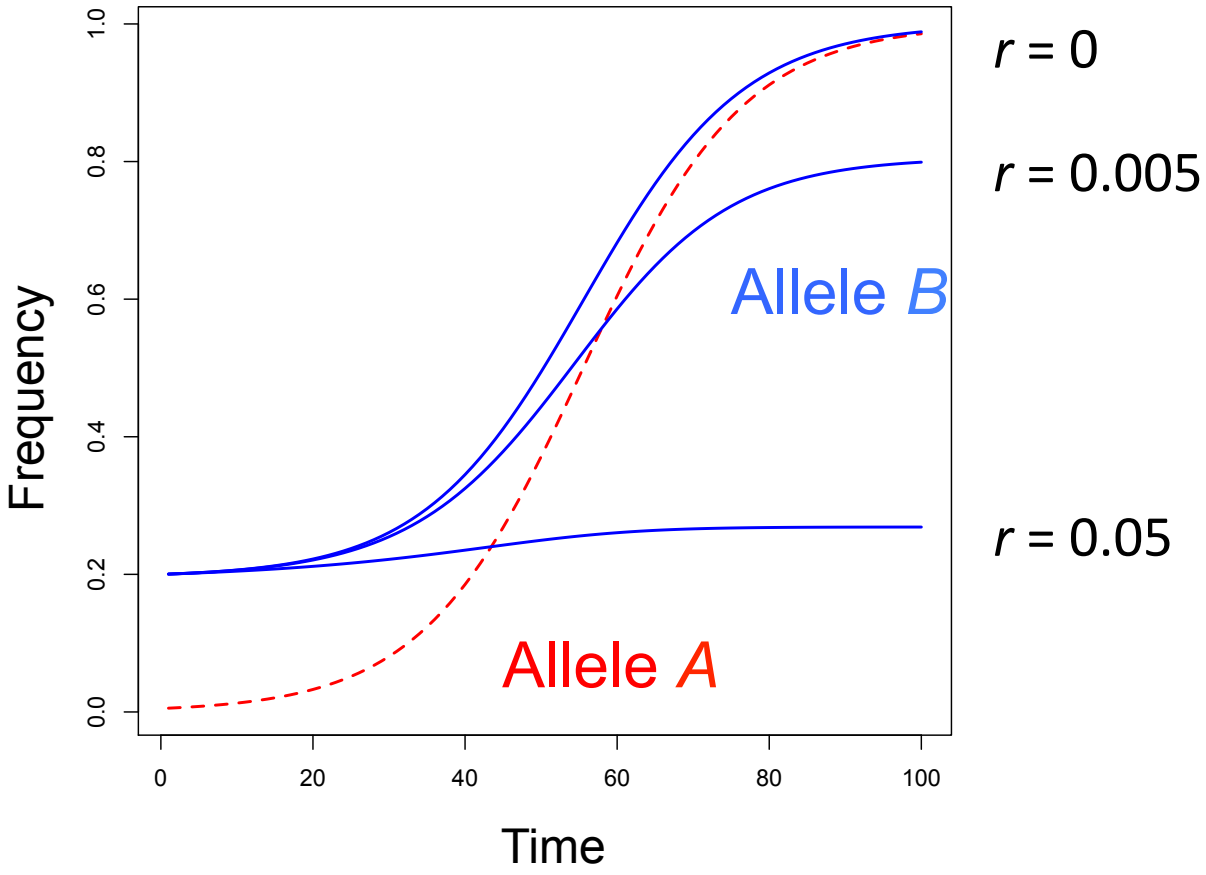
$$\Delta p = \frac{sp[t]q[t]}{1+sp[t]}$$

with $W_A = (1+s)$ and $W_a = 1$

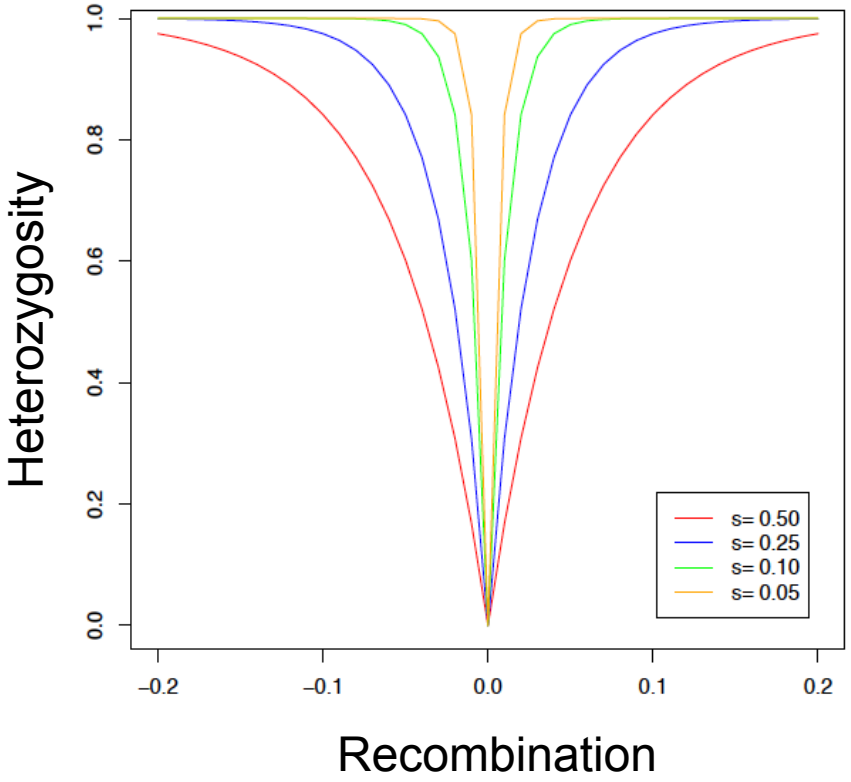
$$\frac{p[t]}{q[t]} = \frac{p[0]}{q[0]}\left(\frac{W_A}{W_a}\right)^t = \frac{p[0]}{q[0]}(1+s)^t = \frac{p[0]}{q[0]}e^{st}$$

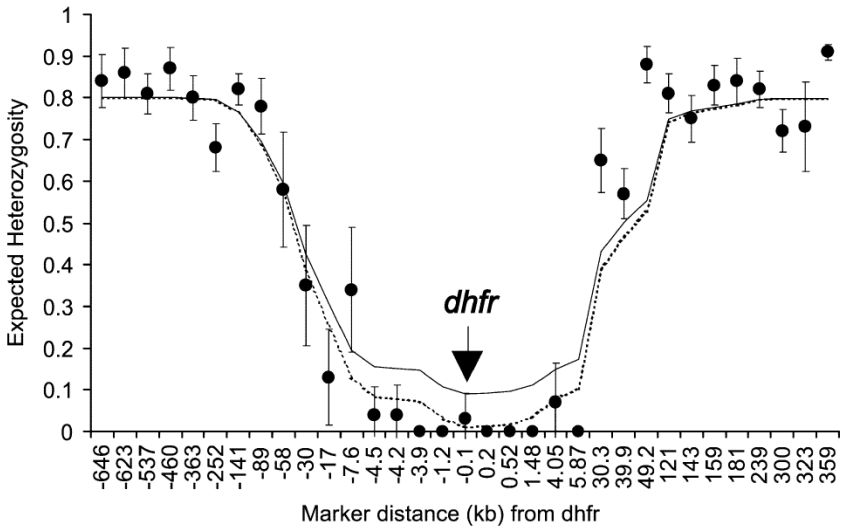
Selective sweeps

- Effect of selection at linked polymorphisms



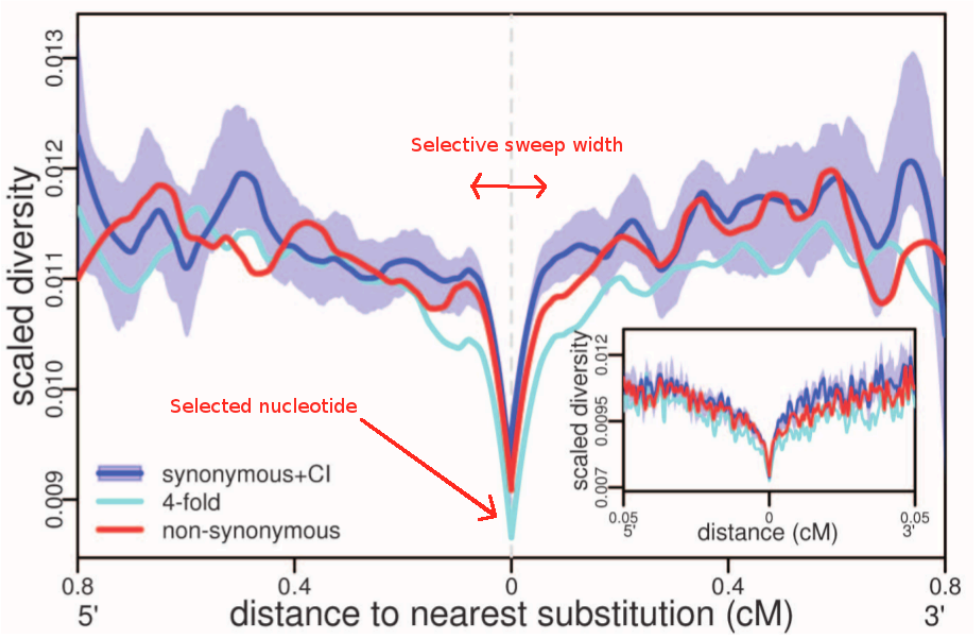
Selective sweeps





- An advantageous mutation in the *dhfr* gene in *Plasmodium falciparum* (vector for malaria), involved in the resistance to anti-parasite treatments

Selective sweeps are not so frequent?

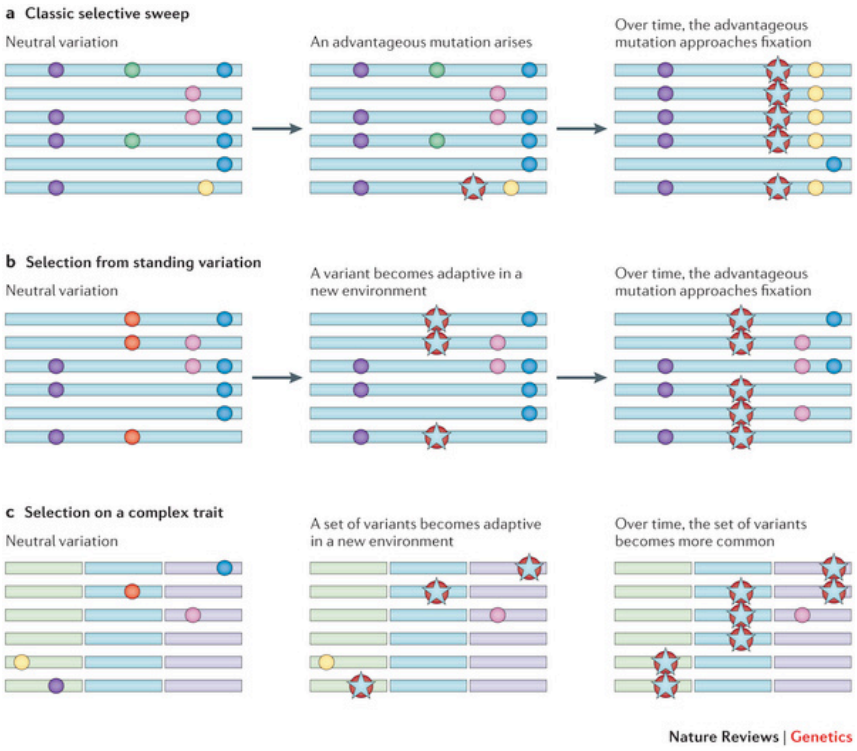


- Only few “classical” sweeps detected in Humans.
- Hernandez *et al*. (2011) *Science* **331**: 920-924

Alternative models

- "soft sweeps" (Hermisson and Pennings 2005), where advantageous variants segregate in the population before they respond to selection

- Polygenic adaptation (Chevin and Hospital 2008)



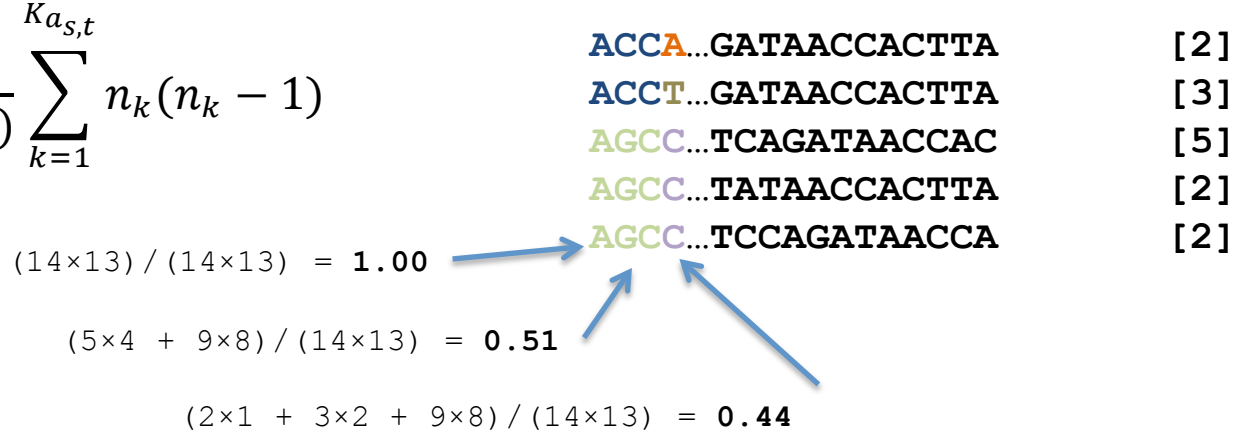
- **More tricky to detect!**

- Scheinfeldt et Tishkoff (2013) *Nature Reviews Genetics* **14**: 692-702

Extended haplotype homozygosity (*EHH*)

- *EHH* is the probability that two randomly chosen chromosomes carrying the core haplotype of interest (*s*) are identical by descent (as assayed by homozygosity at all SNPs) for the entire interval from the core region to the point *t*. *EHH* thus detects the transmission of an extended haplotype without recombination.

$$EHH_{s,t} = \frac{1}{n_{a_s}(n_{a_s}-1)} \sum_{k=1}^{K a_{s,t}} n_k(n_k - 1)$$

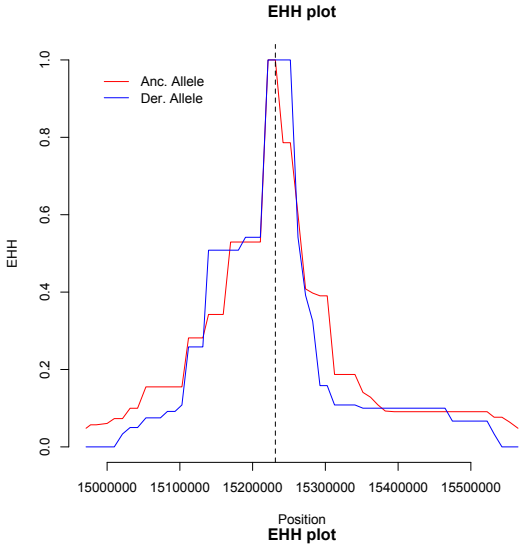


- Where n_{as} gives the number of haplotypes with allele a_s ; $K_{as,t}$ gives the number of unique extended haplotypes carrying allele a_s within the interval from SNP *s* to SNP *t* ; n_k gives the number of copies of a given haplotype *k*

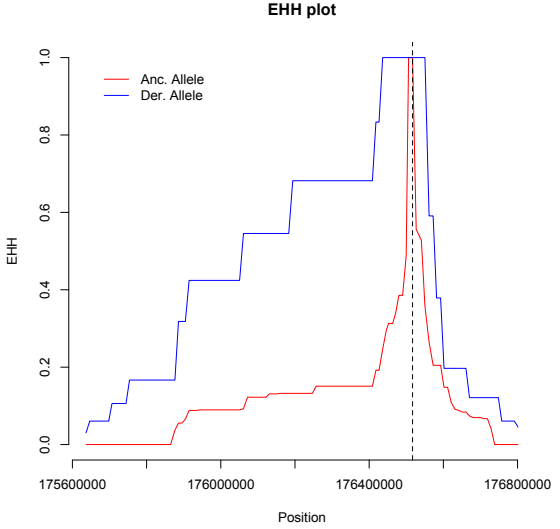
Credits: Sabeti *et al*. (2002) *Nature* **419**: 832-837

Extended haplotype homozygosity (*EHH*)

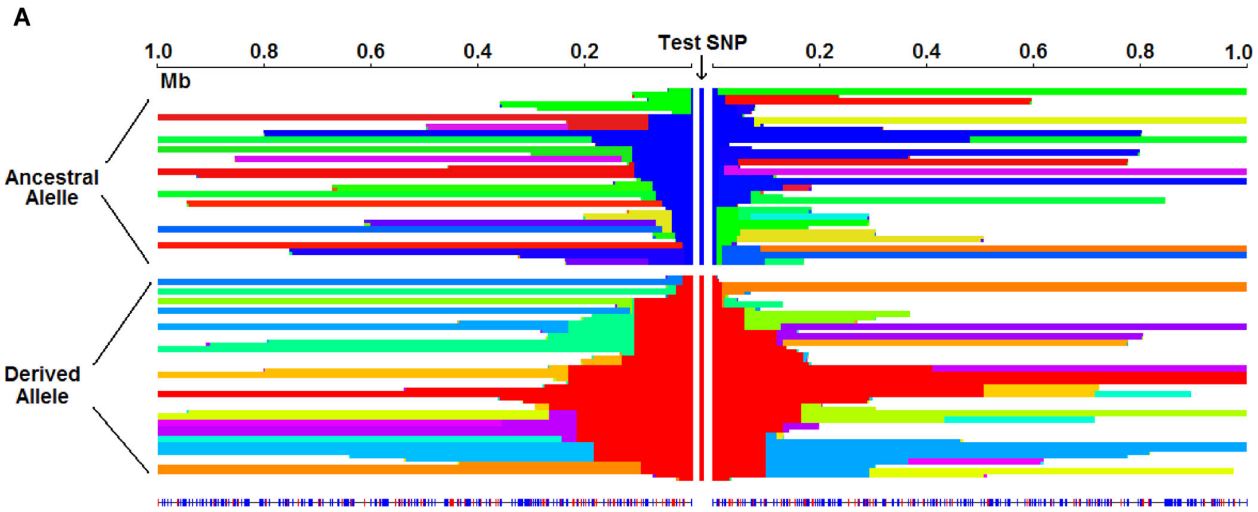
- **Neutrality**: haplotypes associated with ancestral and derived alleles have balanced frequencies : the *EHH* decays at the same rate

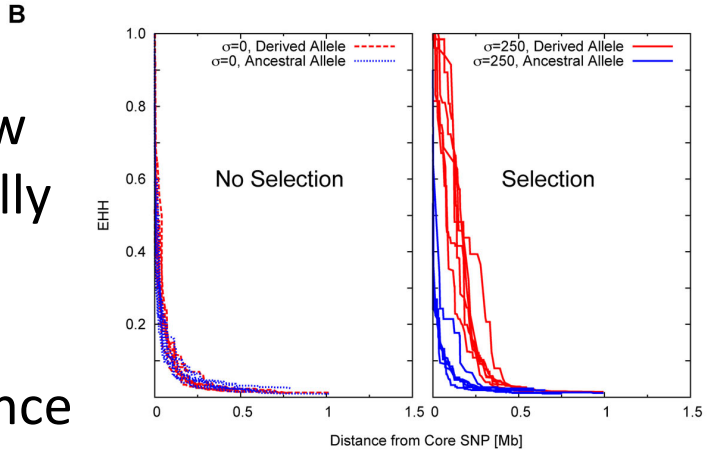


- **Positive selection**: few (extended) haplotypes associated with the derived allele have unusually high frequencies : the *EHH* for the derived allele at the focal SNP decays much more slowly than that of the ancestral allele…



Extended haplotype homozygosity (*EHH*)





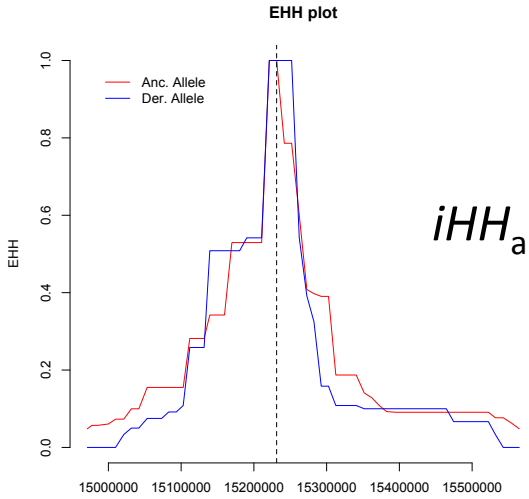
- Limits: in neutral models, low frequency alleles are generally younger and are associated with longer haplotypes than higher frequency alleles. Hence it might be difficult to compare the *EHH* at different positions…

Credits: Voight *et al*. (2006) *PLoS Biol* **4(3)**: e72

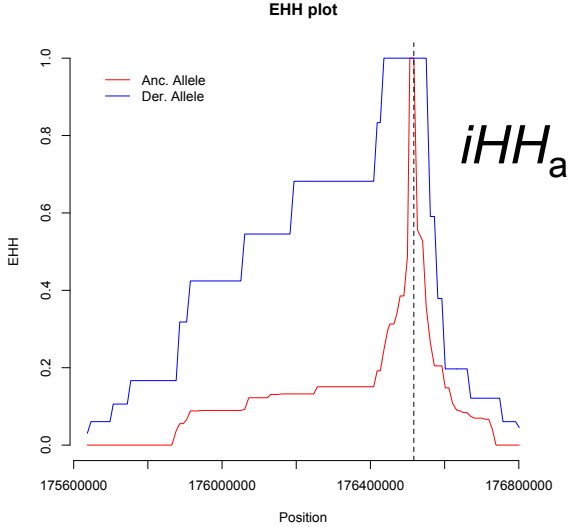
Standardized measure: *iHS*

- To get a standardized measure of extended haplotype homozygosity, Voight *et al*. (2006) defined:
- iHH_a and iHH_d, as the areas under the *EHH* curves ("*integrated EHH*")
- *unstandardized iHS* = $\log(iHH_a / iHH_d)$
- The *iHS* is standardized using the empirical distribution of $\log(iHH_a / iHH_d)$ at SNPs whose derived allele frequency *p* matches the frequency at the core SNP:

$$iHS^{(s)} = \frac{\log\left(\frac{iHH_a^{(s)}}{iHH_d^{(s)}}\right) - E_{p_s}\left[\log\left(\frac{iHH_a}{iHH_d}\right)\right]}{SD_{p_s}\left[\log\left(\frac{iHH_a}{iHH_d}\right)\right]}$$

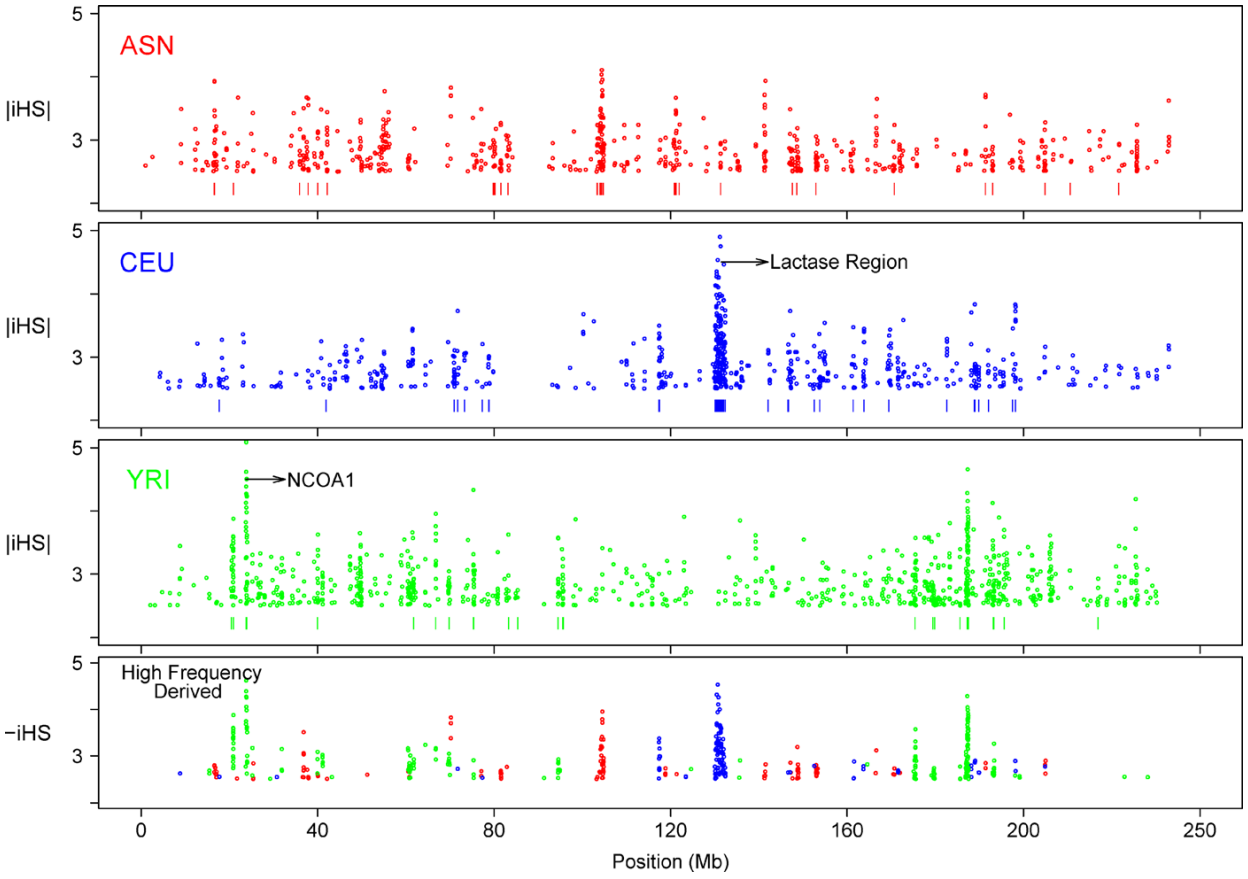


$iHH_a \approx iHH_d$



$iHH_a < iHH_d$

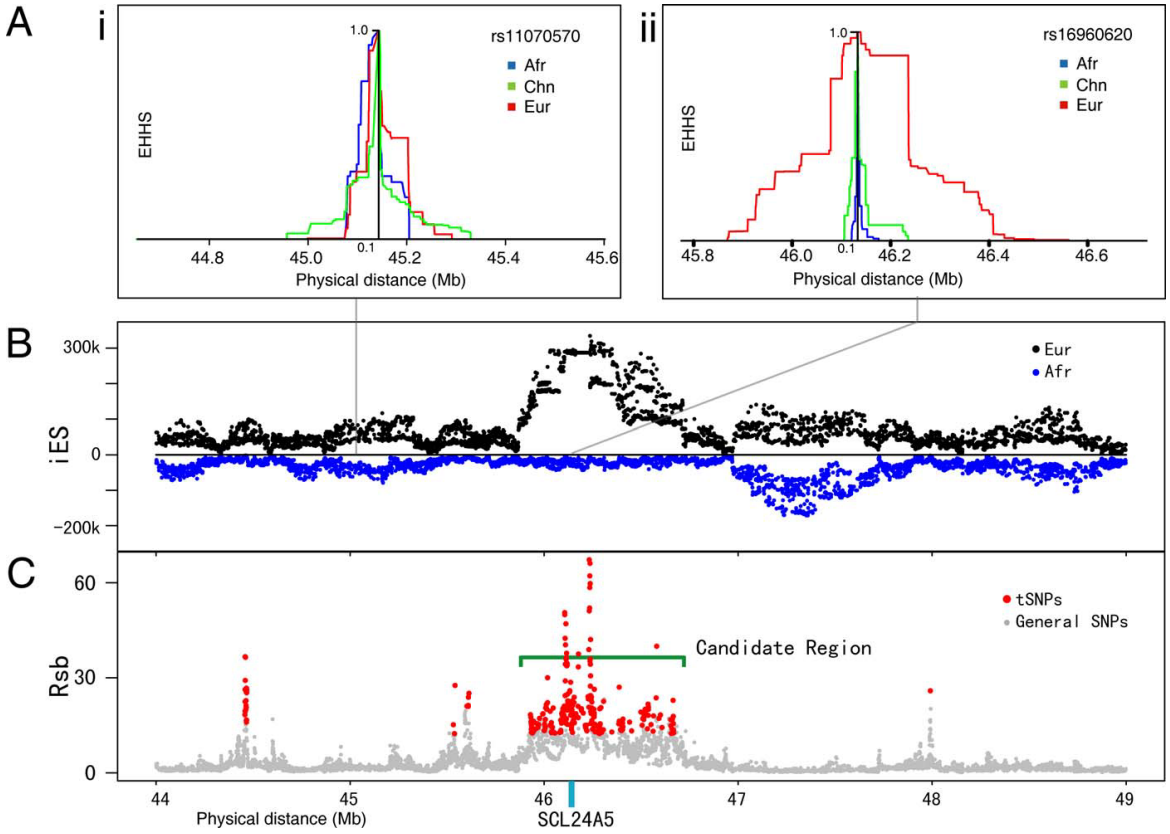
Standardized measure: *iHS*



- Plots of SNPs on chromosome 2 with extreme *iHS* values indicate discrete clusters of signals

Between-population comparisons

- *EHH*-based approaches are for single populations, yet statistics have been derived for between-populations comparisons:



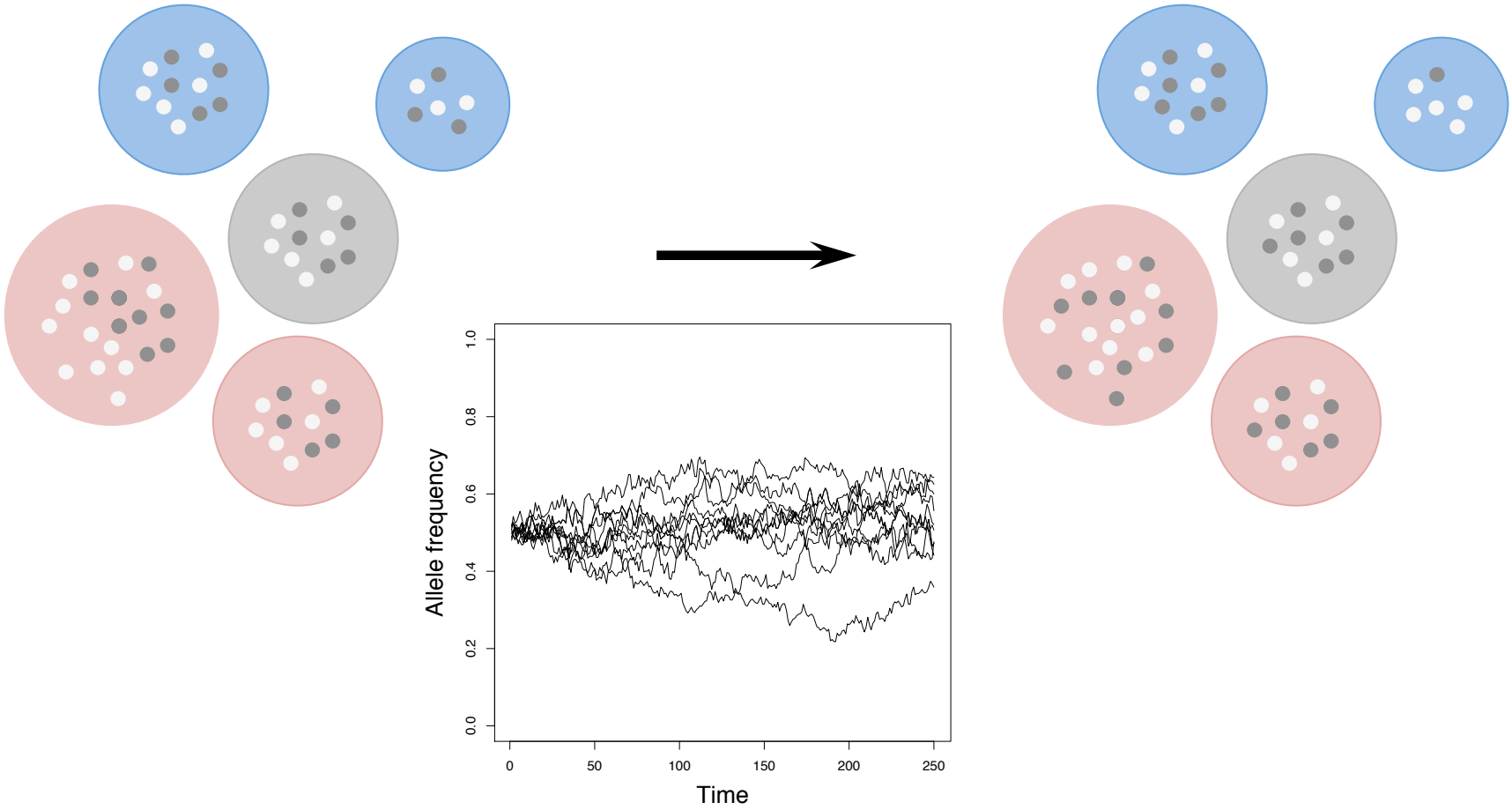
- Use REHH! (https://cran.r-project.org/web/packages/rehh/index.html)

Credits: Tang *et al*. (2007) *PLoS Biol* **5(7)**: e171

How can we detect local adaptation in subdivided populations?

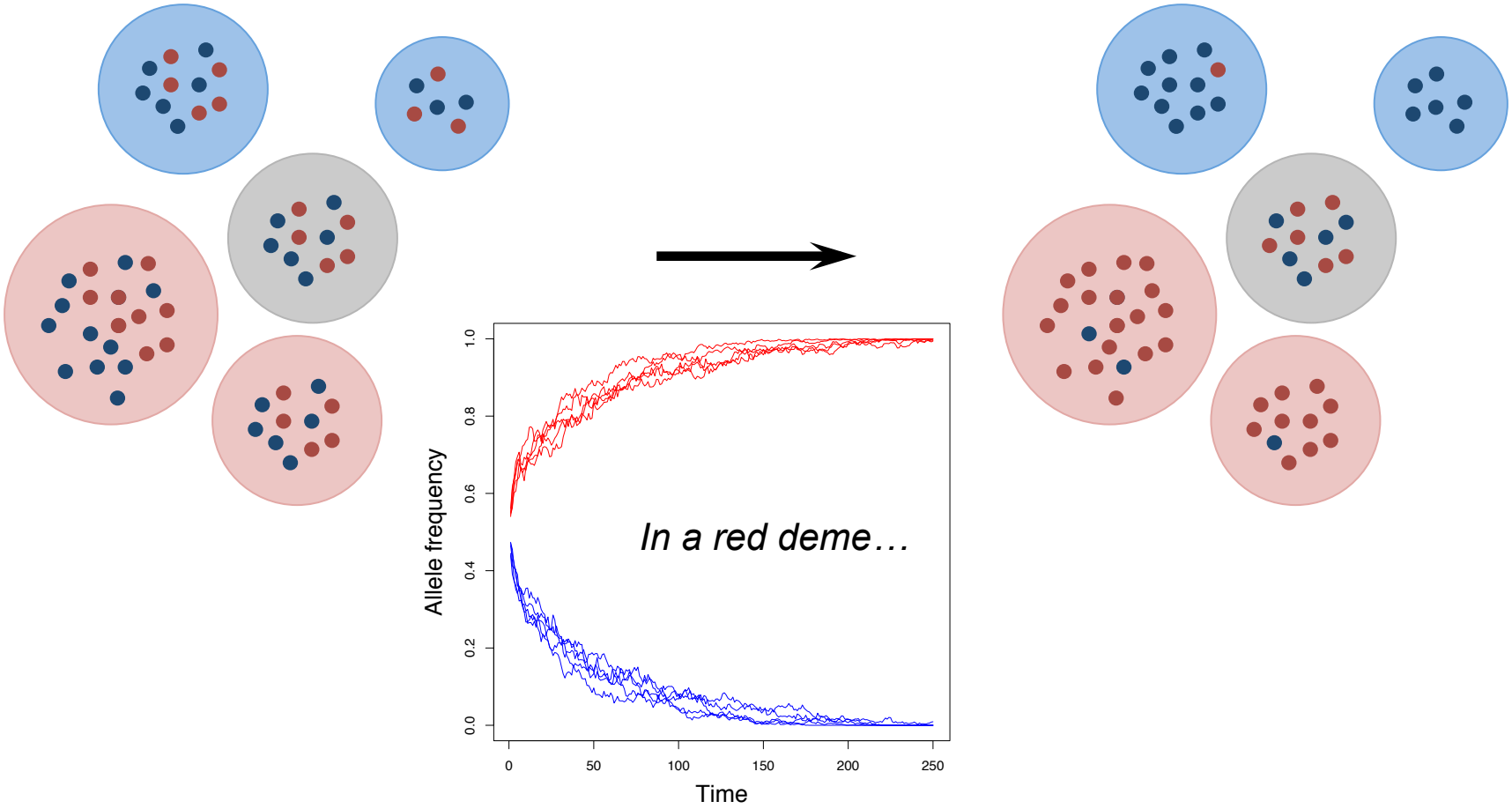
Signatures of local adaptation?

Neutral polymorphisms



Signatures of local adaptation?

Locally adapted genes

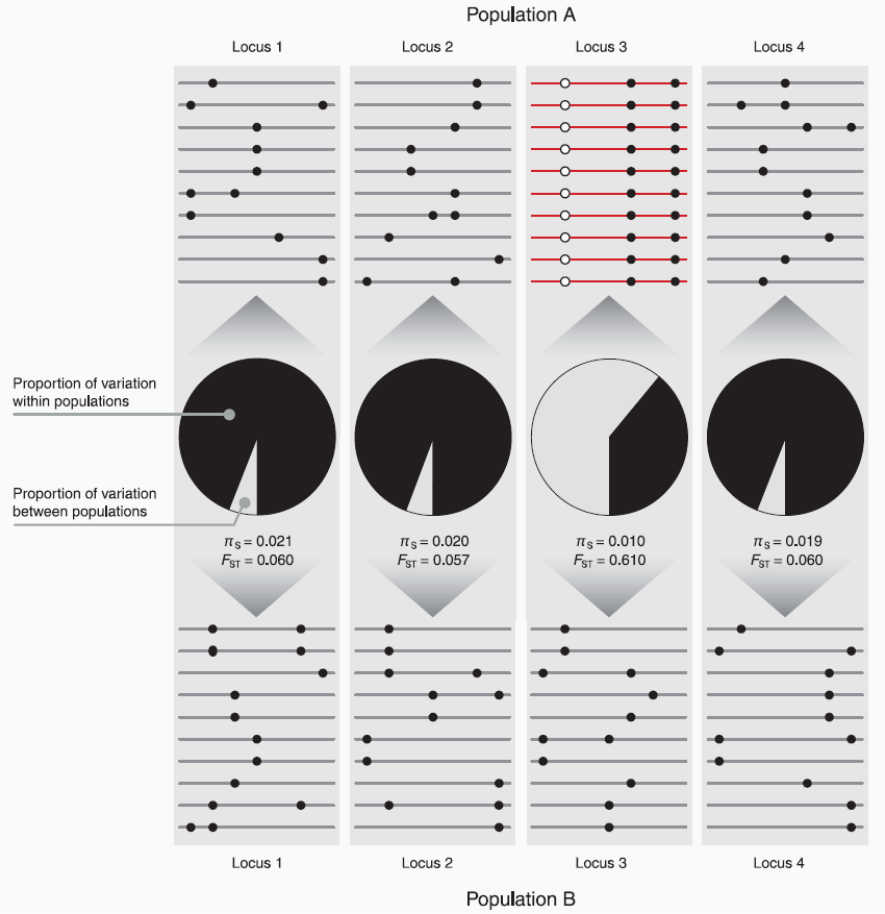


Genome scans: a little bit of history…

- Demography (drift, gene flow, etc.) influences genetic polymorphism at all loci in the same way (on average); not selection, which acts in a locus-specific way (Cavalli-Sforza 1966)

- **The question then is: how to distinguish genome-wide effects from locus-specific effects?**

Signatures of local adaptation?



Within-variation decreased

Between-variation increased

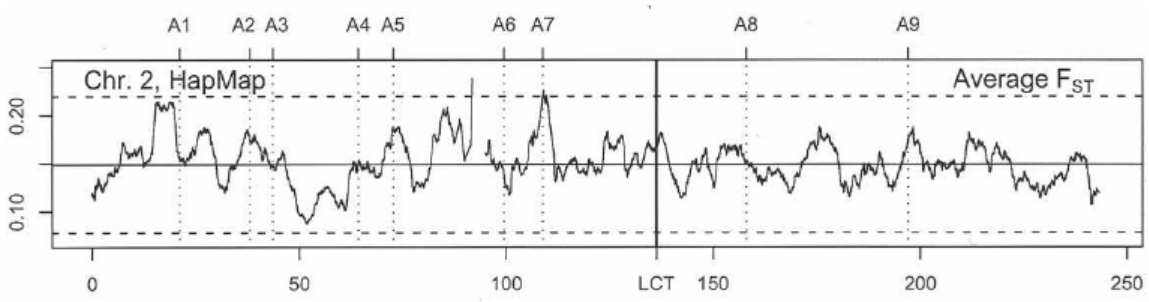
F_{ST} (differentiation) increased

Credits: Storz (2005) *Mol Ecol* **14**: 671-688

Detecting outlier loci in subdivided populations

- General idea: detect outlier loci that depart from the expected distribution of metrics such as F_{ST}. These outlier loci are supposed to be the target of selection.

- "expected distribution": from the simulation of a simple (or not so simple) population model, from the empirical distribution using a large number of marker loci, etc.



Credits: Weir *et al*. (2005) *Genome Research* **15**: 1468-1476

Lewontin and Krakauer's test (1973)

DISTRIBUTION OF GENE FREQUENCY AS A TEST OF THE THEORY OF THE SELECTIVE NEUTRALITY OF POLYMORPHISMS[1,2]

R. C. LEWONTIN AND JESSE KRAKAUER

Department of Theoretical Biology and Department of Biology, University of Chicago, Chicago, Illinois 60637

Manuscript received February 14, 1972
Revised copy received January 16, 1973
Transmitted by T. PROUT

F_{ST} may be defined as: $$F_{ST} = \frac{s_p^2}{\bar{p}(1-\bar{p})} = \frac{\left(1/\left(n-1\right)\right)\sum_{i=1}^{n}(p_i-\bar{p})^2}{\bar{p}(1-\bar{p})}$$

where $\bar{p}$ and s_p^2 are the sampling estimates of the mean and variance of the vector *p* of allele frequencies. Lewontin and Krakauer's test statistic is:

$$T_{LK} = \frac{n-1}{\overline{F}_{ST}} F_{ST}$$

which was shown to be distributed as a X^2 with (*n* - 1) d.f.

Credits: Lewontin et Krakauer (1973) *Genetics* **74**: 175-195

Lewontin and Krakauer's test (1973)

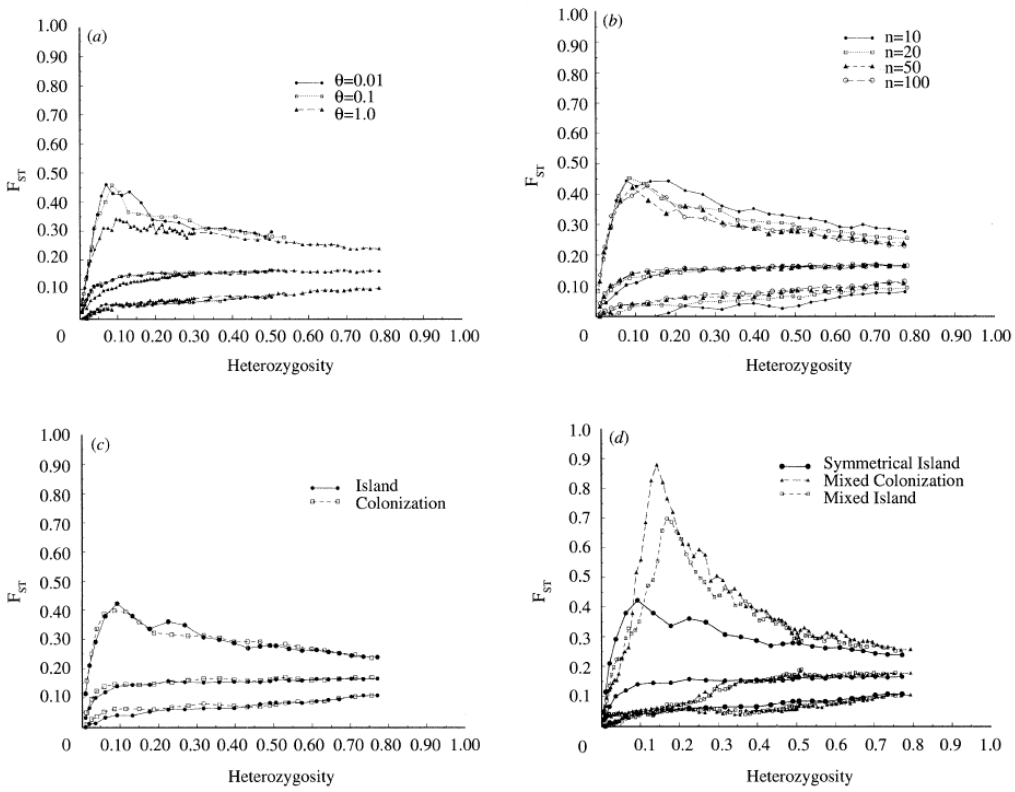
Severe criticisms by Robertson (1975), and Nei and Maruyama (1975)

1. Only a (small set of) ad-hoc distributions of *p were considered*

2. The approach does not account for (realistic) demographic history *(which may result in correlated gene frequencies across demes)*

3. The approach does not allow to identify which locus is targetted by selection

Beaumont and Nichols (1996)



- The joint distribution of F_{ST} and heterozygosity is generally robust, in particular for $H_e > 0.2$… but:
- *This assumes that mutation rates are small as compared to migration rates*
- *The distribution may be altered when the demography departs from a symmetrical island model (and is not accounted for)*

Credits: Beaumont et Nichols (1996) *Proc Roy Soc Lond B* **263**: 1619-1626

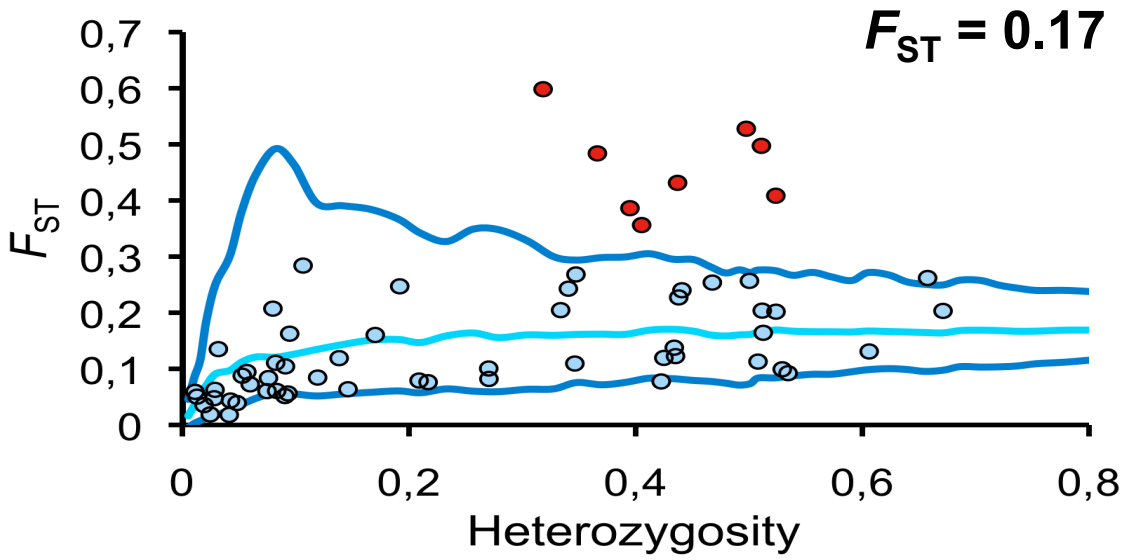
Beaumont and Nichols (1996)

Since the joint distribution of F_{ST} and H_e does not depend much on the nuisance parameters (and on the details of the true history), Beaumont and Nichols have suggested to:

- Measure F_{ST} from the full dataset (multi-locus estimate)
- Simulate artificial data in the island model with $4Nm = 1 / F_{ST} - 1$ (coalescent-based)
- Compute the joint distribution of F_{ST} and H_e
- Identify those loci that depart from this neutral distribution (outliers)

Beaumont and Nichols (1996)

- *Drosophila melanogaster* (15 populations, 61 enzymatic loci)



- The joint distribution F_{ST} and H_e (median and 95% confidence limits), conditional to the observed multilocus estimate of F_{ST}, obtained by means of coalescent simulation of an island model…

Beware hierarchical structure!

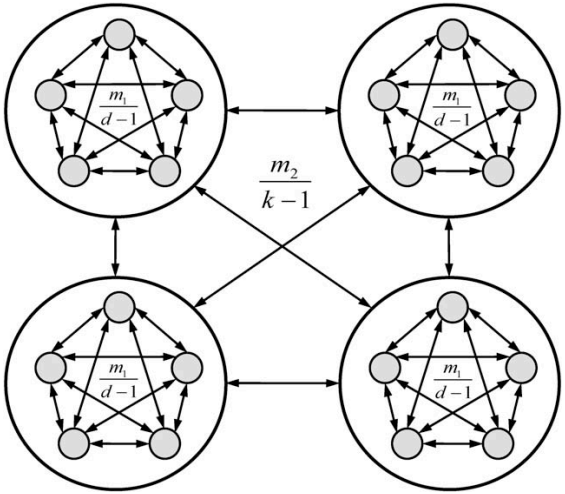
NEWS AND COMMENTARY

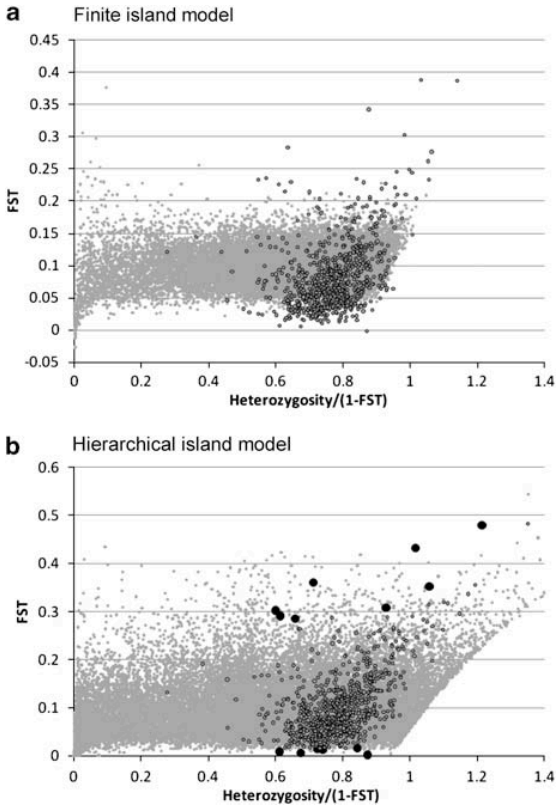
Searching for signatures of selection

Who believes in whole-genome scans for selection?

J Hermisson

Heredity (2009) **103,** 283–284; doi:10.1038/hdy.2009.101; published online 5 August 2009





- Ignoring higher levels of structure tightens the distribution: this increases the rate of false-positives…

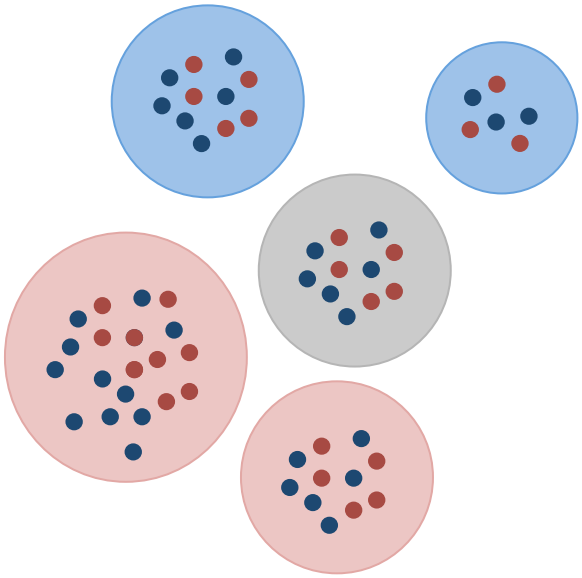
Credits : Excoffier *et al*. (2009) *Heredity* **103**: 285-298

In practice…

- There is a number of software packages:

Fdist2	Beaumont and Nichols (1996)	http://www.maths.bris.ac.uk/~mamab/software/
Dfdist	Beaumont and Nichols (1996)	http://www.maths.bris.ac.uk/~mamab/stuff/
DetSel	Vitalis *et al*. (2001)	http://cran.r-project.org/web/packages/DetSel/index.html
Lositan	Antao *et al*. (2008)	http://popgen.net/soft/lositan/
Mcheza	Antao and Beaumont (2011)	http://popgen.net/soft/mcheza/
Arlequin	Excoffier and Lischer (2010)	http://cmpg.unibe.ch/software/arlequin35/

Alternative, model-based approaches



- The idea is to characterize the distribution of allele frequencies in a model (*e.g.*, the island model) and estimate its parameters from observed data (allele counts).

- The model is parameterized so that the genetic differentiation (F_{ST}) is decomposed into population-specific and locus-specific effects: see, e.g., Beaumont and Balding (2004), Foll and Gaggiotti (2008), Riebler *et al*. (2008), Gompert and Buerkle (2011), etc.

The data

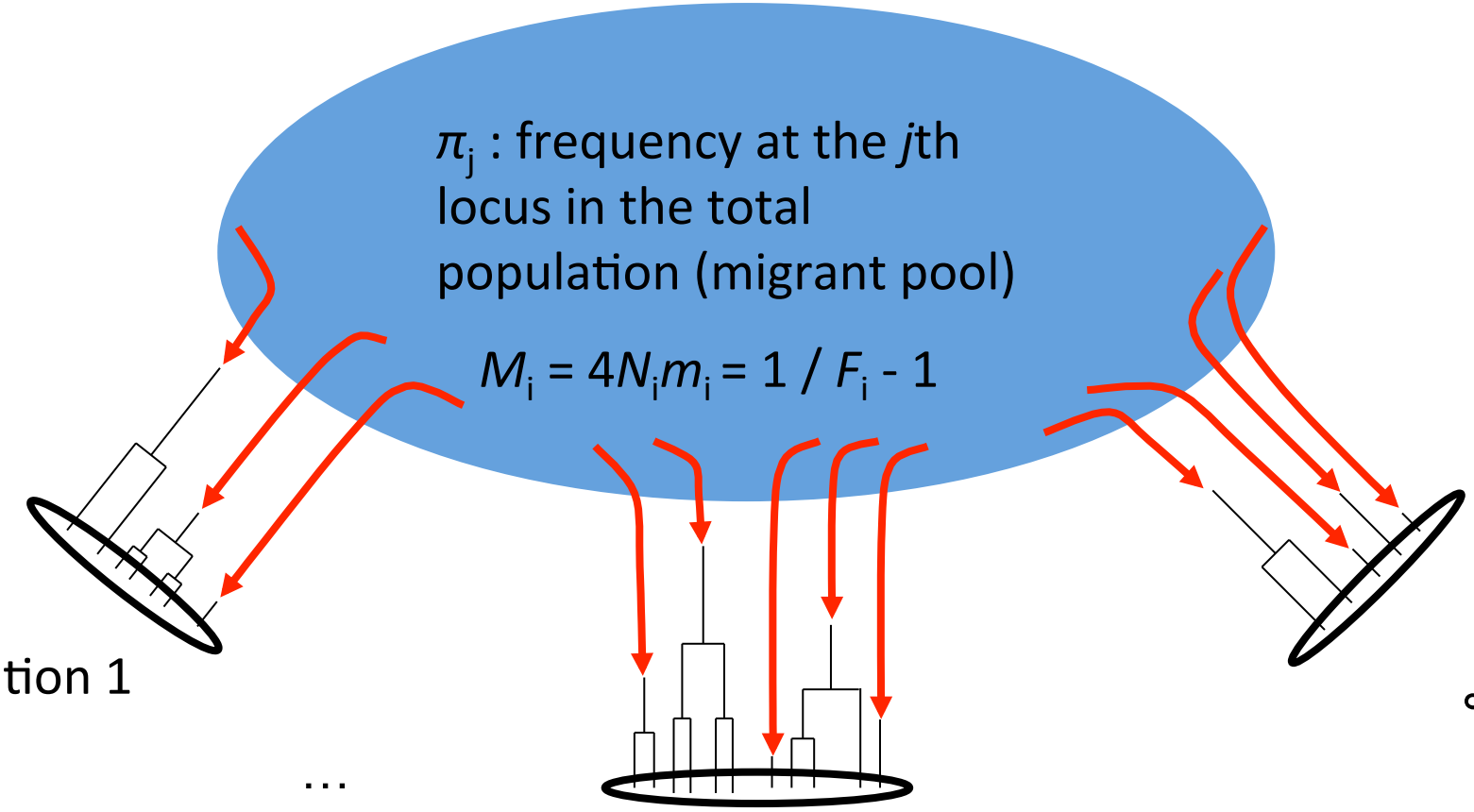
- Single Nucleotide Polymorphisms (SNPs) genotyped in a number of populations.
- SNPs are bi-allelic, co-dominant markers.
- The data consist in allele counts $\boldsymbol{n}_{ij}$ = (x_{ij} ; n_{ij} - x_{ij}) at locus *j* in population *i*. The likelihood of a sample of genes reads:

$$\mathcal{L}(p_{ij}; \mathbf{n}_{ij}) = \binom{n_{ij}}{x_{ij}} p_{ij}^{x_{ij}} (1 - p_{ij})^{(n_{ij} - x_{ij})}$$

- Where p_{ij} is the (unknown) allele frequency at the *j*th locus in the *i*th population

Bayesian model

Island model:



Credits: Beaumont (2005) *Trends Ecol Evol* **20**: 435-440

Bayesian logistic regression: BAYESFST

Note that $M = 4Nm = 1 / F_{ST} - 1$ and assume:

$$\log\left(\frac{F_{ST}}{1 - F_{ST}}\right) = \alpha_i + \beta_j + \gamma_{ij},$$

where:

α_i is a locus effect

β_j is a population effect

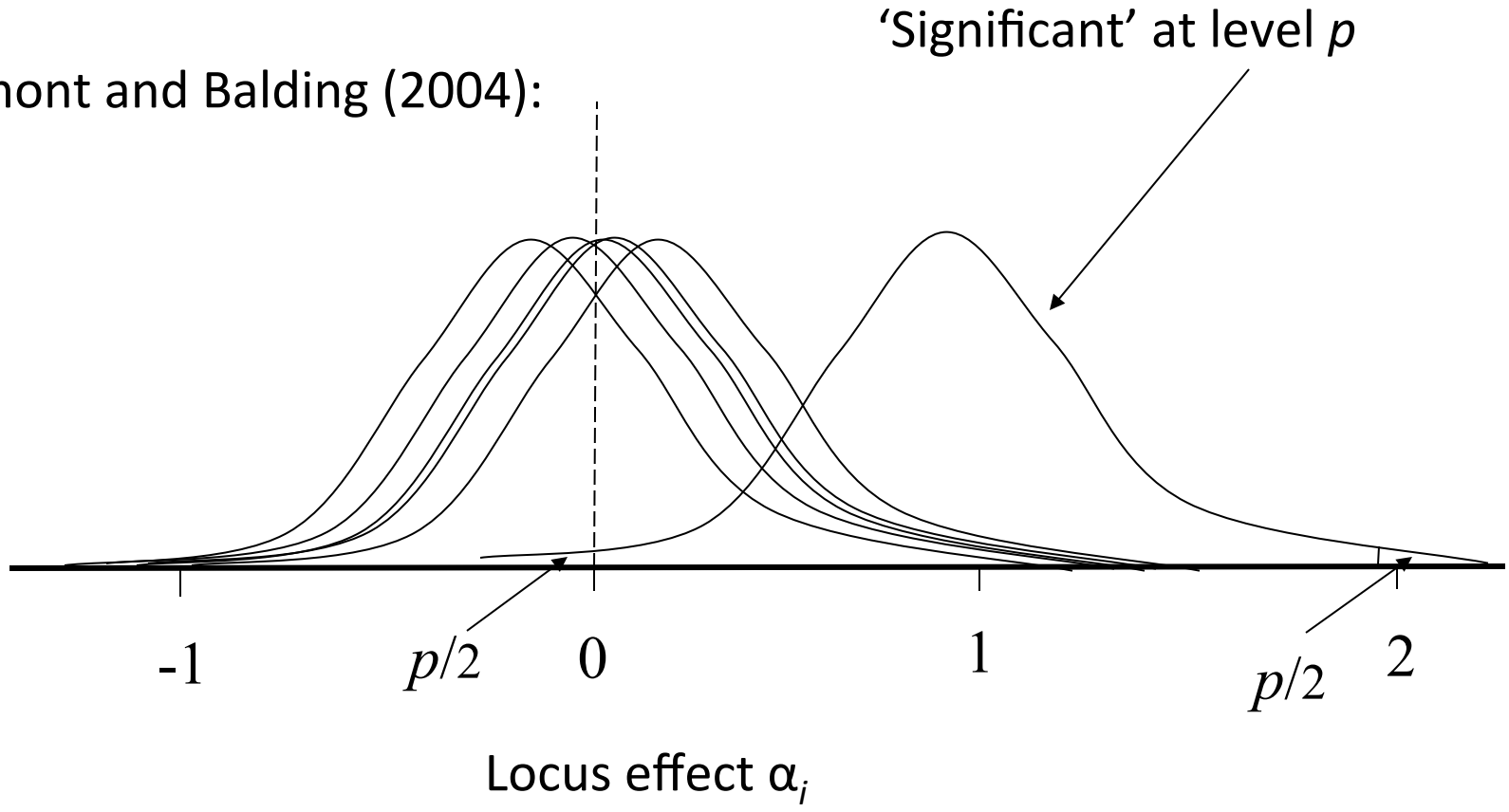
γ_{ij} is a specific locus-by-population effect

Sampling from the posterior distribution (MCMC) assuming normal prior distributions for α_i , β_j and γ_{ij}

Credits: Beaumont et Balding (2004) *Mol Ecol* **13**: 969-980

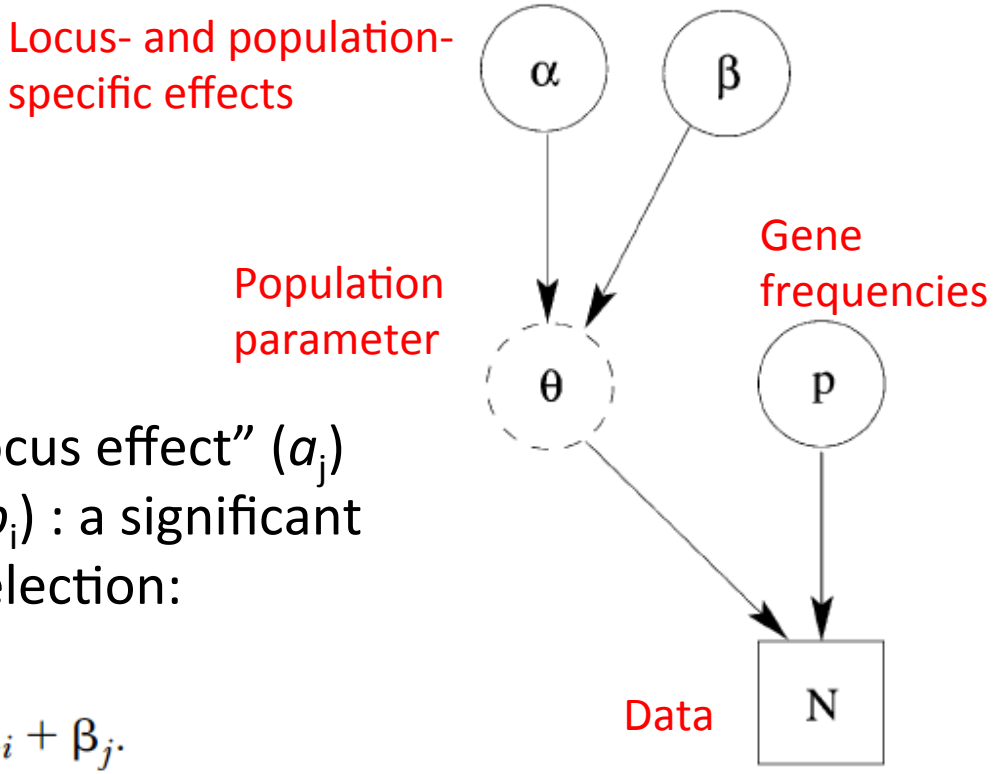
Hypothesis testing

Beaumont and Balding (2004):



« we define α_i to be "significant at level *P*" if its equal-tailed 100(1 - *P*)% posterior interval excludes zero »

Alternative model: BAYESSCAN



- Consider a model with a "locus effect" (a_j) and a "population effect" (b_i) : a significant locus effect is a proxy for selection:

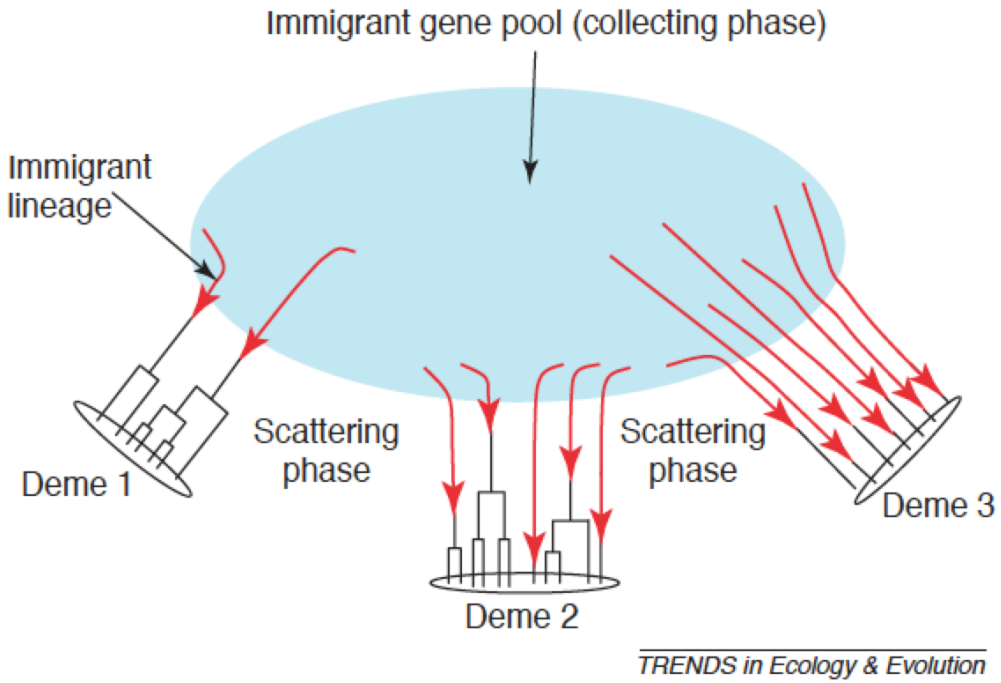
$$\log\left(\frac{F_{\mathrm{ST}}^{ij}}{1-F_{\mathrm{ST}}^{ij}}\right) = \log\left(\frac{1}{\theta_{ij}}\right) = \alpha_i + \beta_j.$$

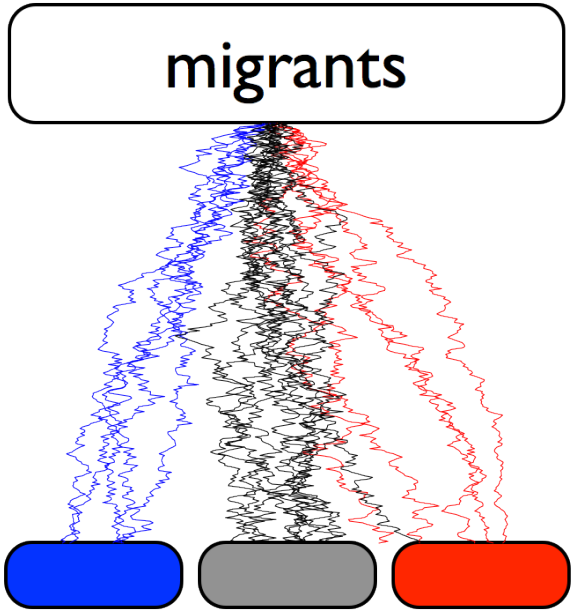
- Foll and Gaggiotti (2008) consider a RJ-MCMC algorithm to decide whether a locus is targeted by selection (or not)

Credits: Foll et Gaggiotti (2008) *Genetics* **180**: 977-993

A change of perspective: SELESTIM

- From neutrality tests… to the inference of selection strength…

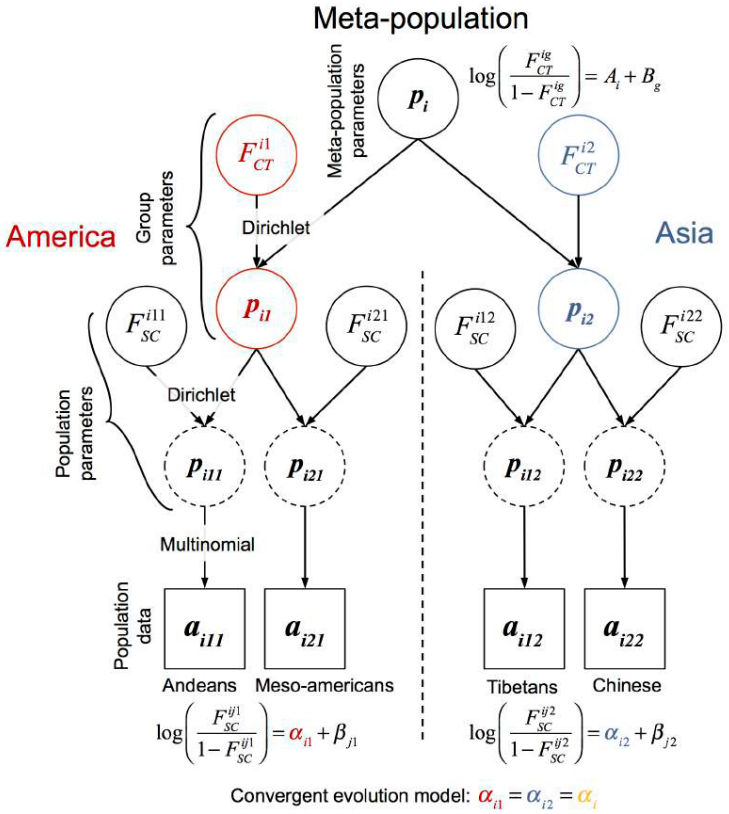


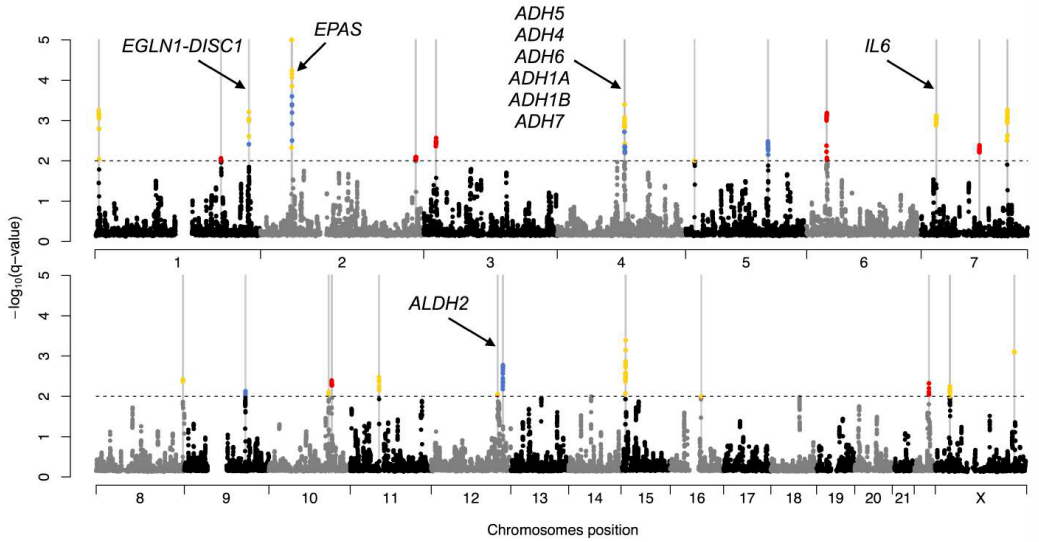


Credits: Vitalis *et al*. (2014) *Genetics.* 196: 799-817

Accounting for hierarchical population structure

- Gompert and Buerkle (2011), Foll *et al*. (2014): accounting for a hierarchical population model (assumed to be known)

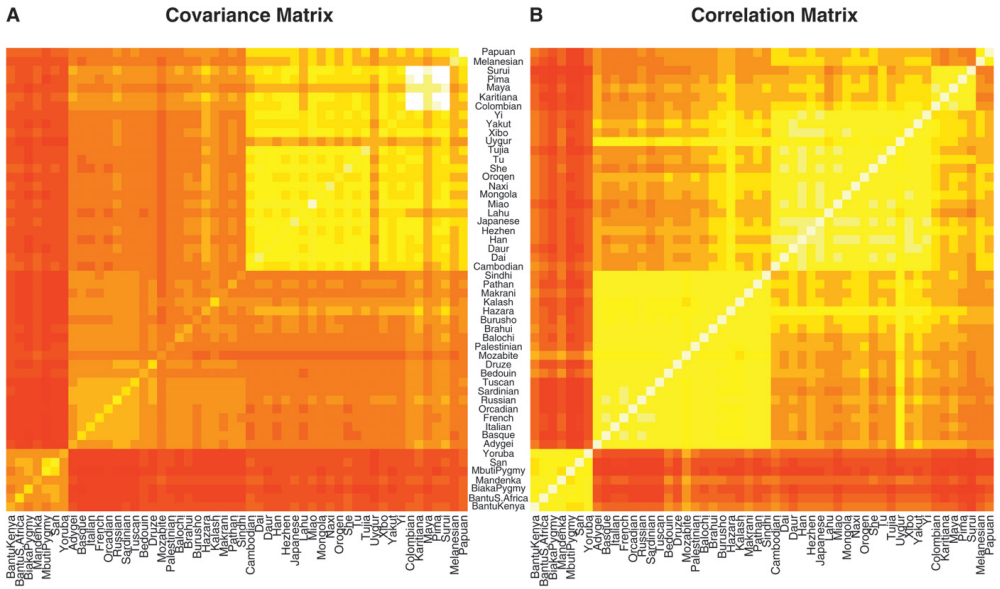




Credits : Foll *et al*. (2014) *AJHG* **95**: 394-407

Accounting for any population structure

- BAYENV (Coop *et al*. (2010) and BAYPASS (Gautier 2015): a Bayesian method that estimates the pattern of covariance in allele frequencies between populations from a set of markers, and then uses this as a null model for a test at individual SNPs

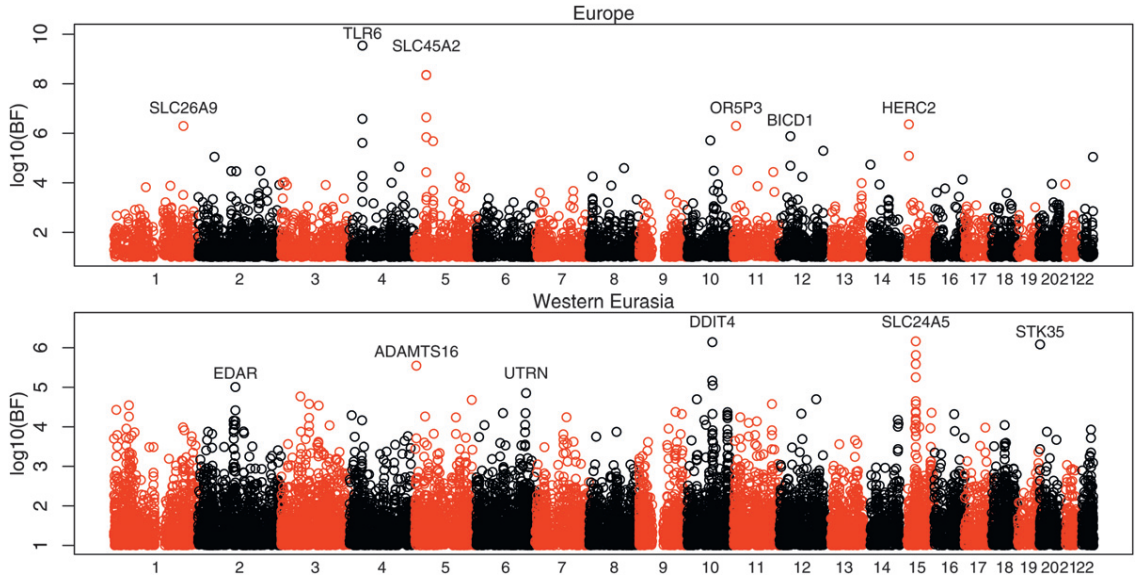


Credits: Coop *et al*. (2010) *Genetics* **185**: 1411-1423

Correlations with environmental variables

- Conditional on SNP frequency variation across populations, the model is used to investigate whether allele frequencies at a SNP of interest are significantly correlated with an environmental variable *Y*

- Support of the model with an environmental variable, compared with the null model for each SNP along the human genome, for (A) a « European » effect and (B) a « western » Eurasian effect:



In practice…

- There is a number of software packages:

BayesFST	Beaumont and Balding (2004)	http://www.reading.ac.uk/Statistics/genetics/software.html
BayeScan	Foll and Gaggiotti (2008)	http://cmpg.unibe.ch/software/BayeScan/
SelEstim	Vitalis *et al*. (2014)	http://www1.montpellier.inra.fr/CBGP/software/selestim/index.html
Bamova	Gompert and Buerkle (2011)	http://www.uwyo.edu/buerkle/software/bamova/
Bayenv	Coop *et al*. (2010)	http://www.eve.ucdavis.edu/gmcoop/Software/Bayenv/Bayenv.html
LFMM	Frichot *et al*. (2013)	http://membres-timc.imag.fr/Eric.Frichot/lfmm/index.htm
BayPass	Gautier (2015)	http://www1.montpellier.inra.fr/CBGP/software/baypass/

- All these methods are Bayesian: you must check for convergence and mixing properties. They are based on assumptions for the population model.

Take home messages

- Genome scans are very popular

- Yet outlier loci may be due to endogenous genetic barriers rather than to local adaptation (Bierne *et al*. 2011 *Mol. Ecol*. **20**, 2044-2072)

- All models are wrong… Be aware of their limits, their robustness to violations of the model assumptions, etc.

- Different methods use different aspects of the data… (allele frequencies, haplotype information, etc.): different time scales?

- Poor agreement among studies (on the same data!): only biological information will ultimately permit to distinguish between false positives and true signals