Inferring the evolutionary history of populations

Renaud Vitalis

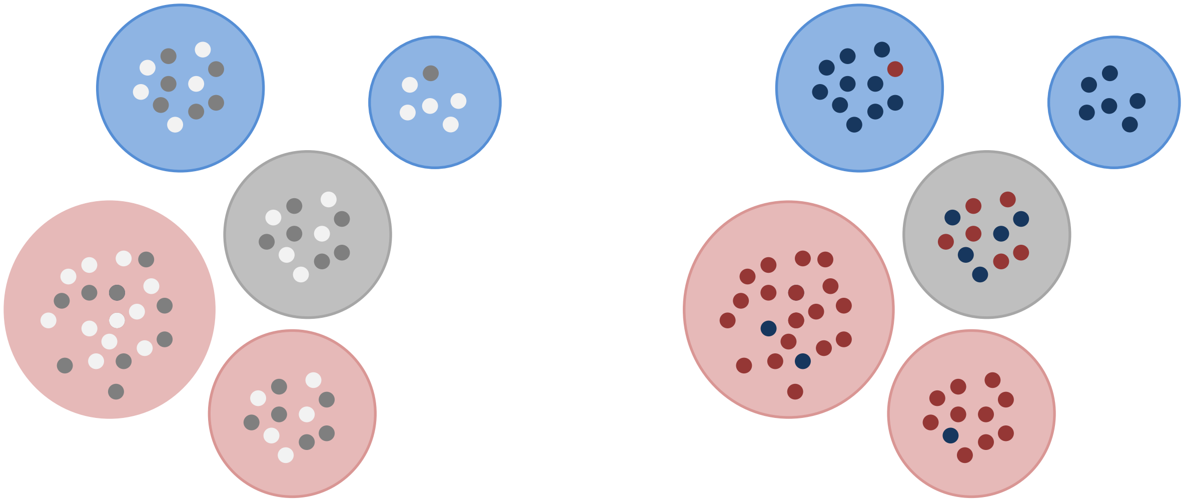
Centre de Biologie pour la Gestion des Populations
INRA, Montpellier

A long standing question

“One fairly obvious attack [*to this problem*] is to investigate […] the expected consequences of drift by examining the variation of gene frequencies in time, or space. […] The likelihood that [*selection*] simulate exactly [*the amount of variation due to drift*] will become smaller the more independent gene systems we examine, as the expectation of drift, unlike selective variation, will be the same for all genes”

(Cavalli-Sforza 1966 *Proc. Roy. Soc. Lond. B Biol. Sci.*)

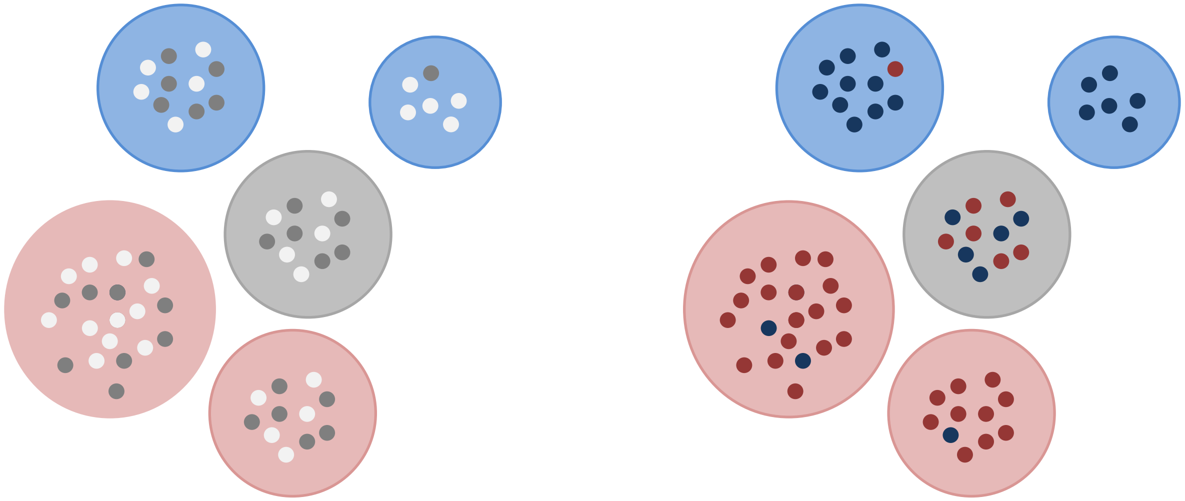
Neutral *vs*. locally adapted genes



Characterizing the expected variation due to drift:

- $T_{\mathrm{LK}} = (n-1)\ F_{\mathrm{ST}}/\bar{F}_{\mathrm{ST}}$ (Lewontin and Krakauer 1973; Bonhomme *et al*. 2010)

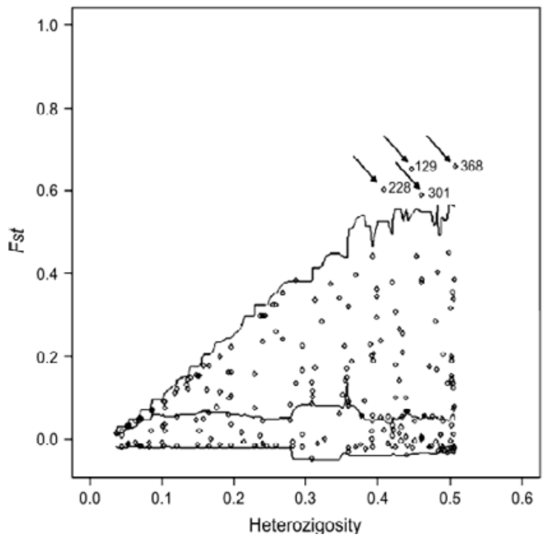
Neutral *vs*. locally adapted genes



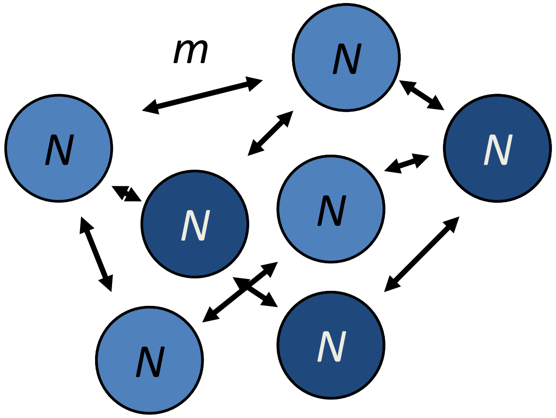
Characterizing the expected variation due to drift:

- $T_{\mathrm{LK}} = (n-1)\,F_{\mathrm{ST}}/\bar{F}_{\mathrm{ST}}$ (Lewontin and Krakauer 1973; Bonhomme *et al*. 2010)
- Coalescent simulations (Beaumont and Nichols 1996; Vitalis et al. 2001; Excoffier *et al*. 2009)

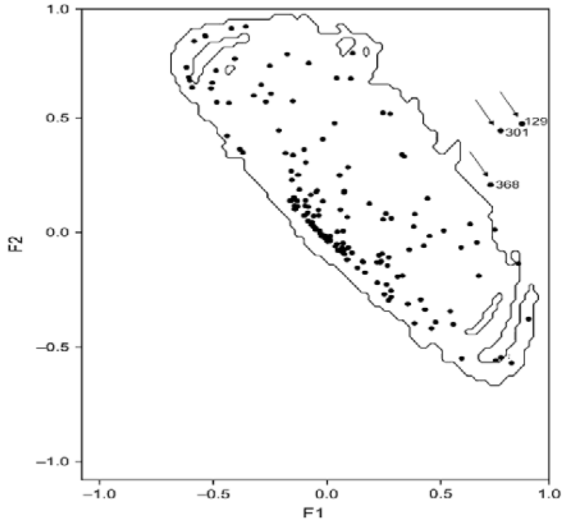
Coalescent-based simulations



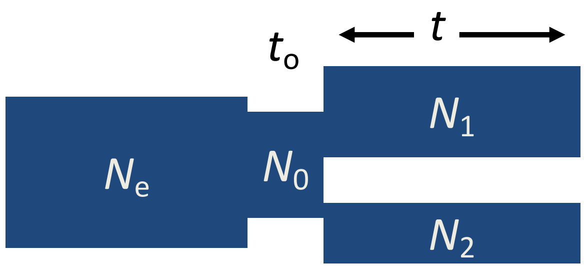
FDIST – Beaumont & Nichols 1996



symmetrical population differentiation (F_{ST}), as a function of heterozygosity



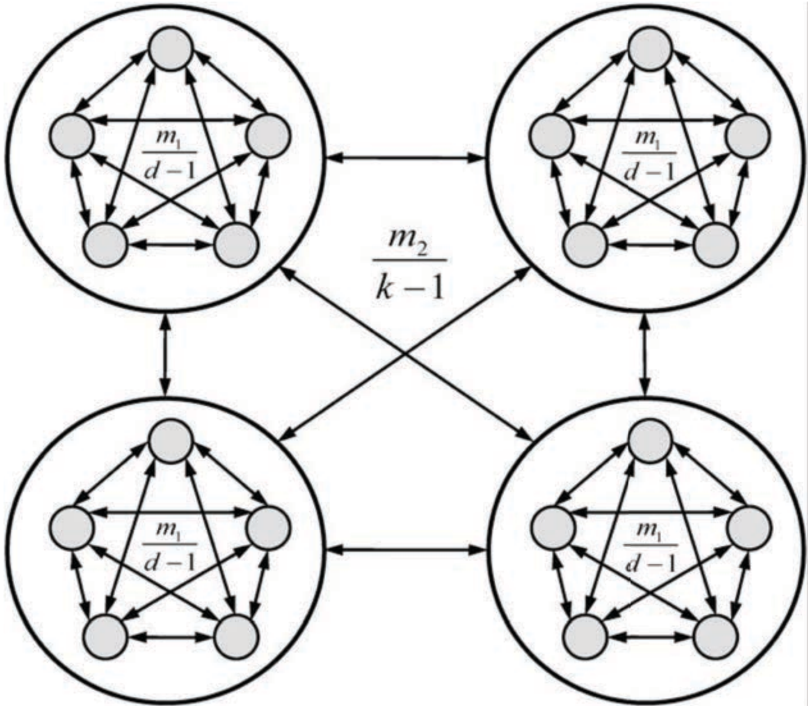
DETSEL – Vitalis *et al*. 2001

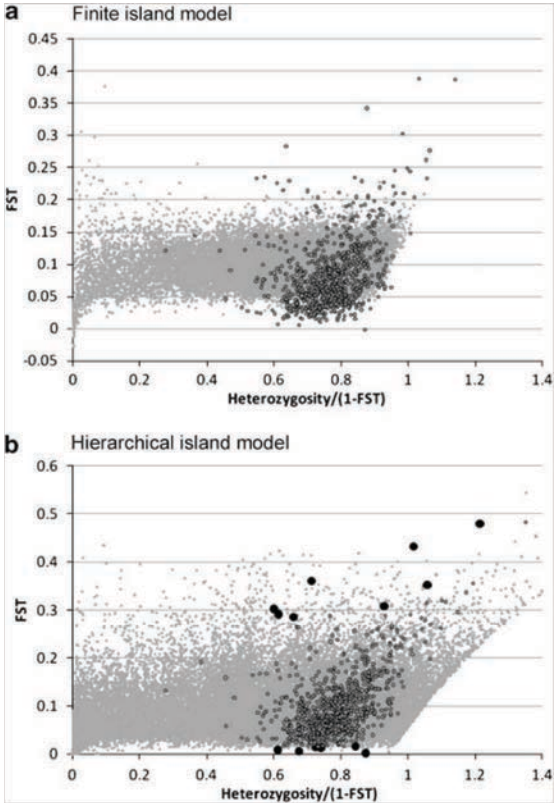


Joint distribution of F_1 and F_2 (which measure the divergence of populations 1 and 2 from their ancestor)

Credits : Bonin *et al.* (2006) *Mol Biol Evol* **23**: 773-783 (and R. Butlin)

Ignoring hierarchical structure

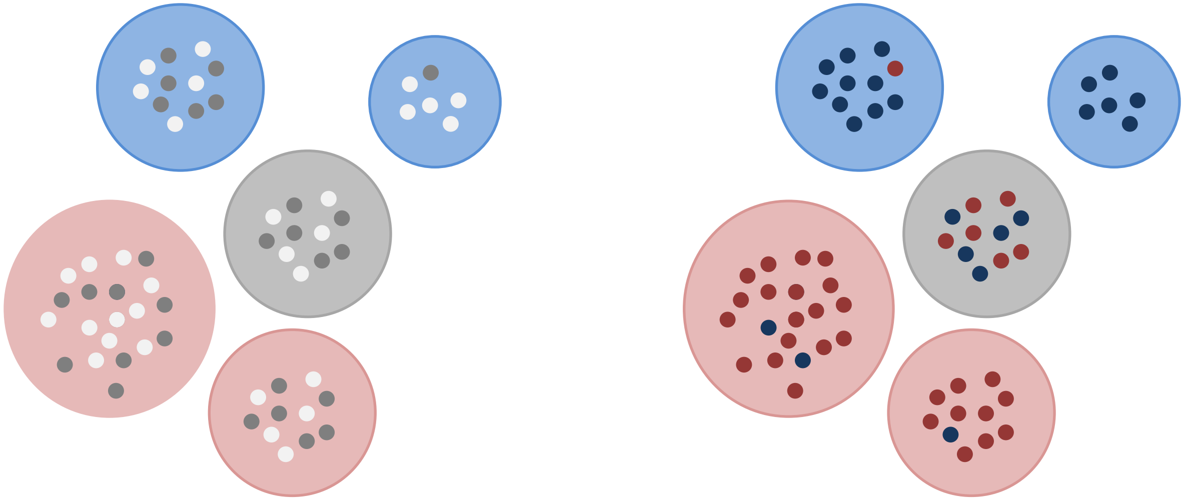




Ignoring higher levels of structure increases the rate of false-positives…

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

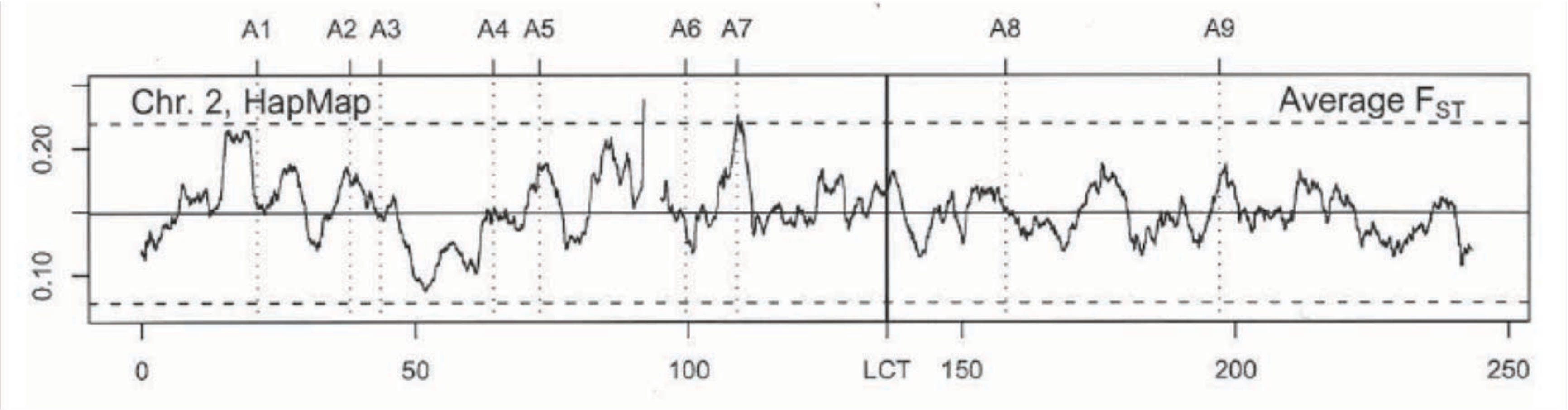
Neutral *vs*. locally adapted genes



Characterizing the expected variation due to drift:

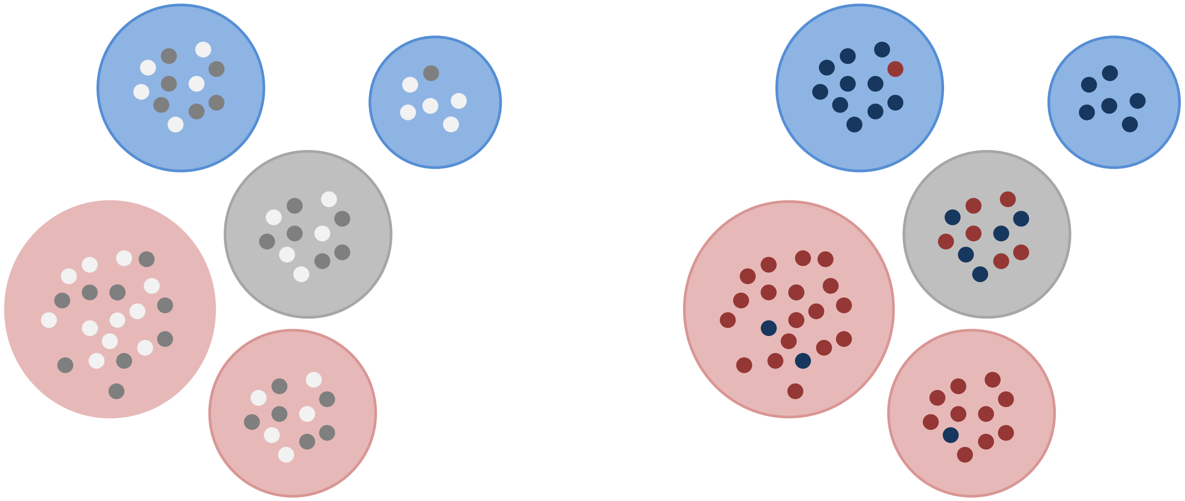
- $T_{\mathrm{LK}} = (n-1)\, F_{\mathrm{ST}}/\bar{F}_{\mathrm{ST}}$ (Lewontin and Krakauer 1973; Bonhomme *et al*. 2010)
- Coalescent simulations (Beaumont and Nichols 1996; Vitalis et al. 2001; Excoffier *et al*. 2009)
- Using empirical distributions (Akey *et al*. 2002; Weir *et al*. 2005)

Empirical distributions



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

Model-based approaches



Characterizing the distribution of allele frequencies, conditionally on some model and parameters

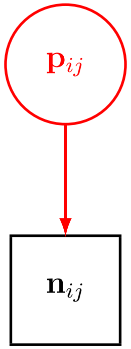
- Island model: Beaumont and Balding (2004); Riebler *et al*. (2008); Foll and Gaggiotti (2008)
- Hierarchical island model: Gompert and Buerkle (2011); Foll *et al*. (2014)
- **Explicit modelling of selection**: SelEstim (Vitalis *et al*. 2014)

The data

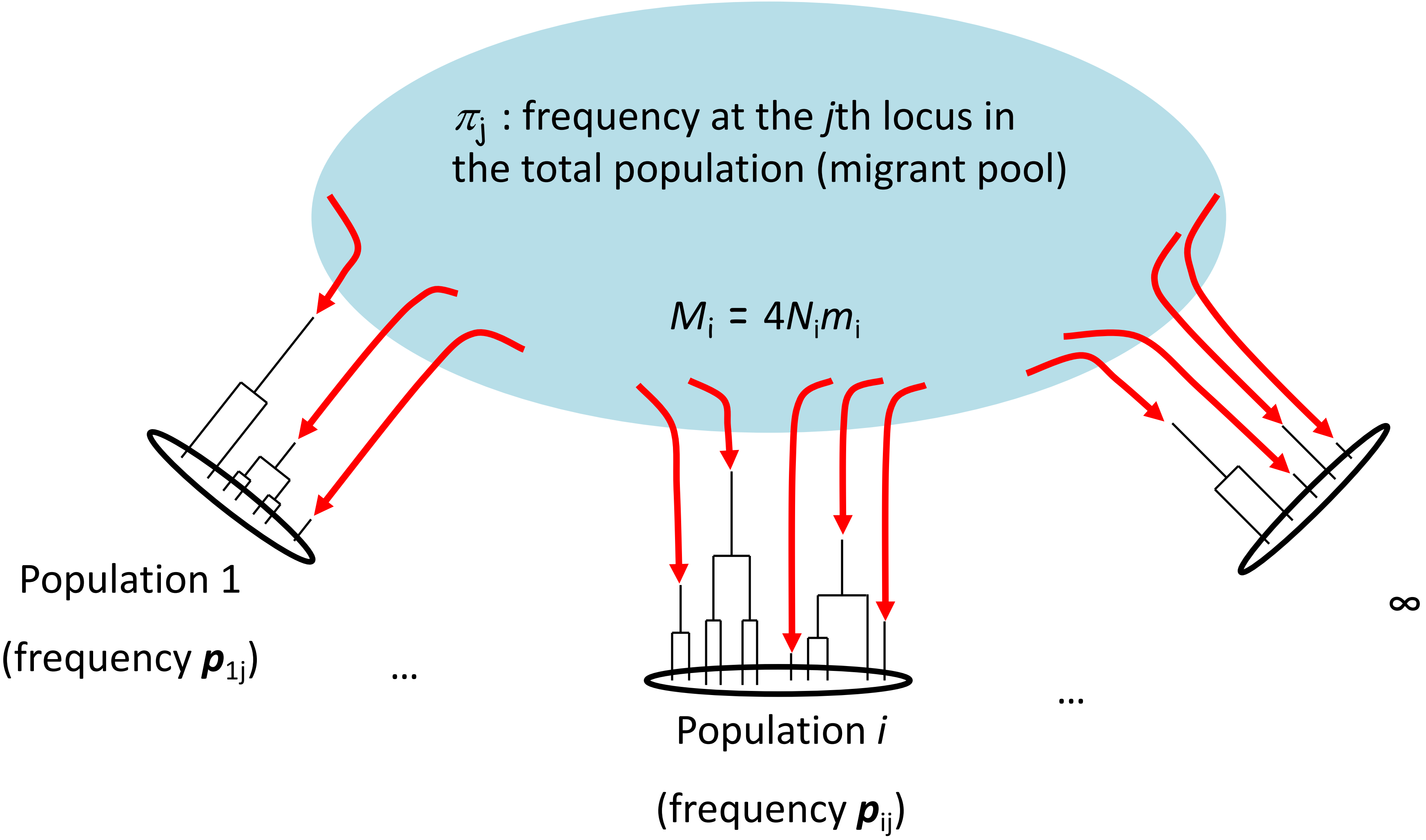
$$\boxed{\mathbf{n}_{ij}}$$

SNPs at many loci, in several populations (allele counts)

Population allele frequencies



Binomial likelihood that depends upon (unknown) population frequencies



π_j : frequency at the *j*th locus in the total population (migrant pool)

$M_i = 4N_i m_i$

Population 1

(frequency $\boldsymbol{p}_{1j}$)

…

Population *i*

(frequency $\boldsymbol{p}_{ij}$)

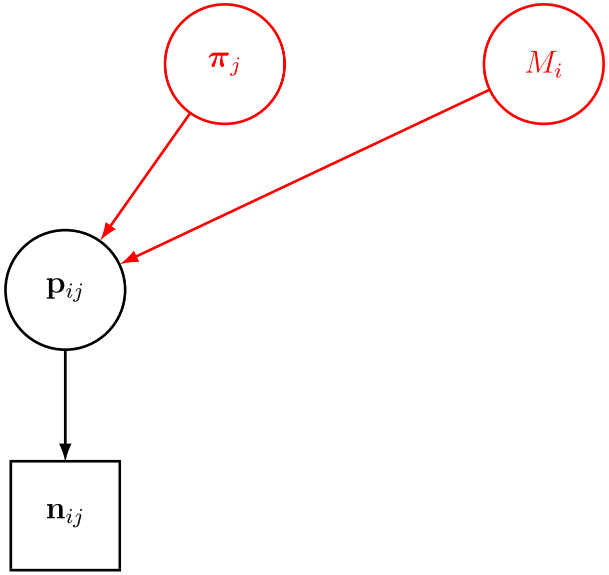
…

∞

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

The population model

Infinite island model: the population frequencies depend on $M_i = 4N_i m_i$ and the frequencies in the migrant pool

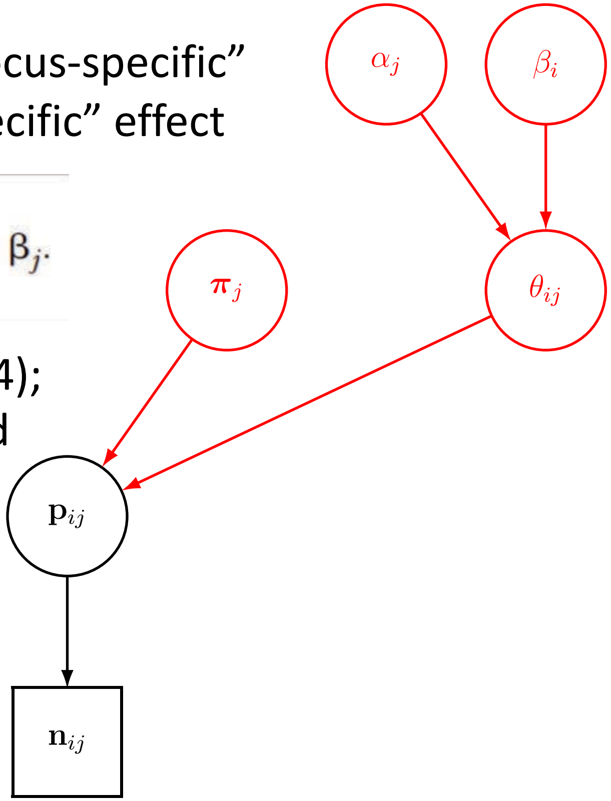


Relation to previous models

- Logistic regression model
- F_{ST} is decomposed into a “locus-specific” effect and a “population-specific” effect

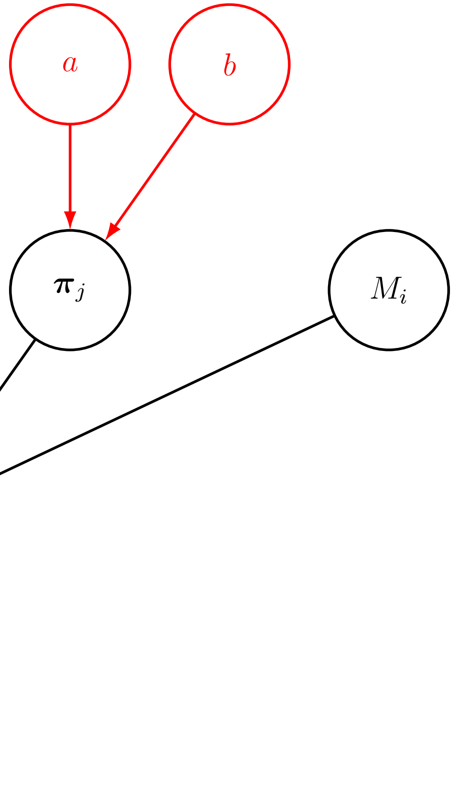
$$\log\left(\frac{F_{\mathrm{ST}}^{ij}}{1-F_{\mathrm{ST}}^{ij}}\right)=\log\left(\frac{1}{\theta_{ij}}\right)=\alpha_i+\beta_j.$$

- Beaumont and Balding (2004); Riebler *et al*. (2008); Foll and Gaggiotti (2008)

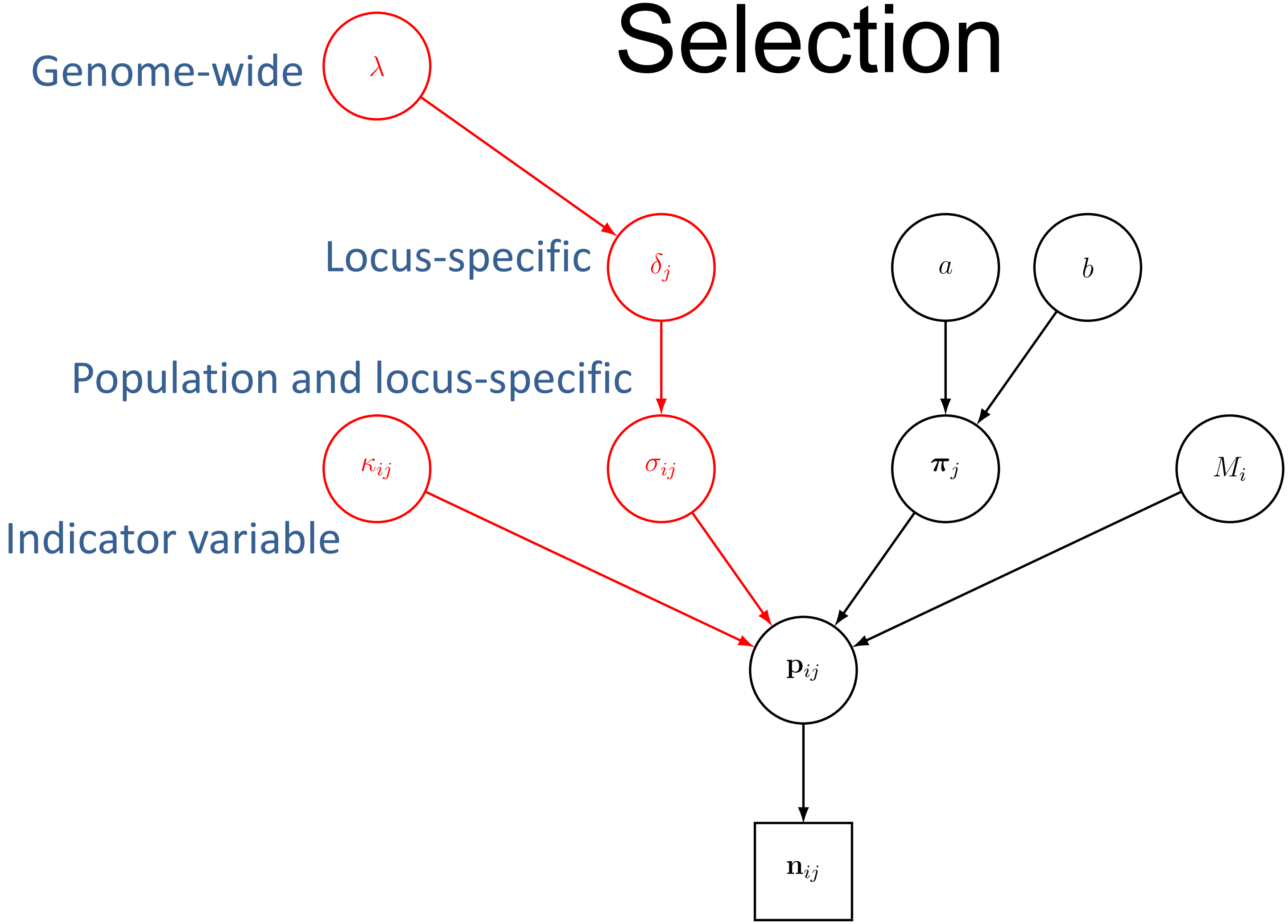


Allele frequencies in the migrant pool

Shape parameters of the beta distribution of (migrant) allele frequencies



Selection



Genome-wide

Locus-specific

Population and locus-specific

Indicator variable

- Allele frequencies = stationary density of the diffusion process (Wright 1949)
- All marker loci are targeted by selection, to some extent
- Sampling from the joint posterior distribution of the parameters using MCMC

Markov chain Monte Carlo (MCMC)

We use the Metropolis – Hastings algorithm to sample from the joint posterior distribution of the model parameters:

Likelihood Prior distributions

$$f(\mathrm{M},\boldsymbol{\pi},\boldsymbol{\kappa},\boldsymbol{\sigma},\boldsymbol{\delta},\lambda|\mathrm{n}) \propto \prod_{i=1}^{n_\mathrm{d}} \prod_{j=1}^{L} \mathcal{L}(p_{ij};\mathrm{n}_{ij})\psi(p_{ij};M_i,\boldsymbol{\pi}_j,\kappa_{ij},\sigma_{ij})\times$$

$$f(\mathrm{M})f(\boldsymbol{\pi})f(\boldsymbol{\kappa})f(\boldsymbol{\sigma}|\boldsymbol{\delta})f(\boldsymbol{\delta}|\lambda)f(\lambda)$$

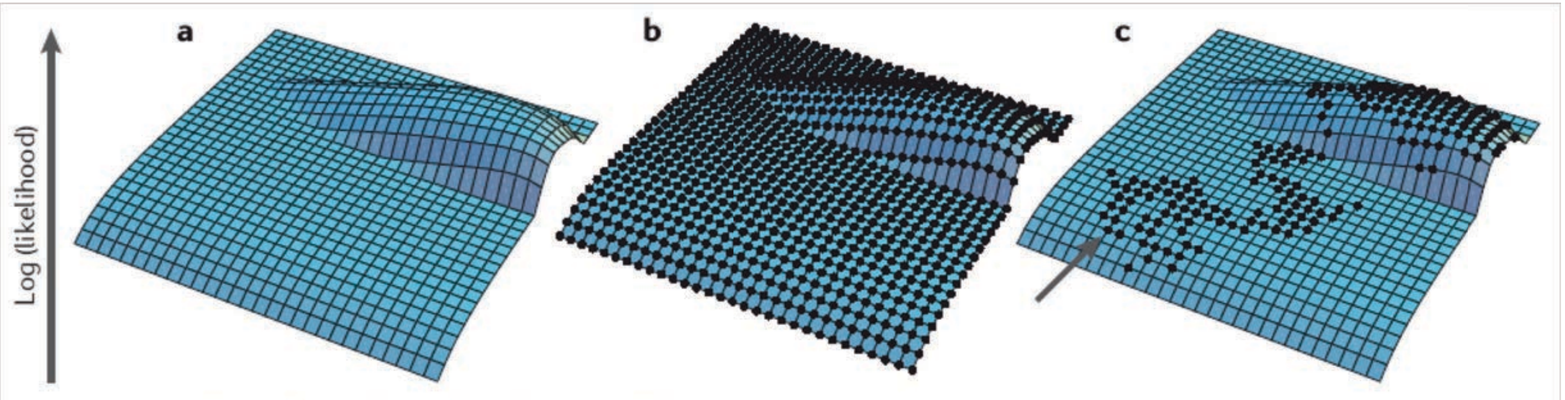
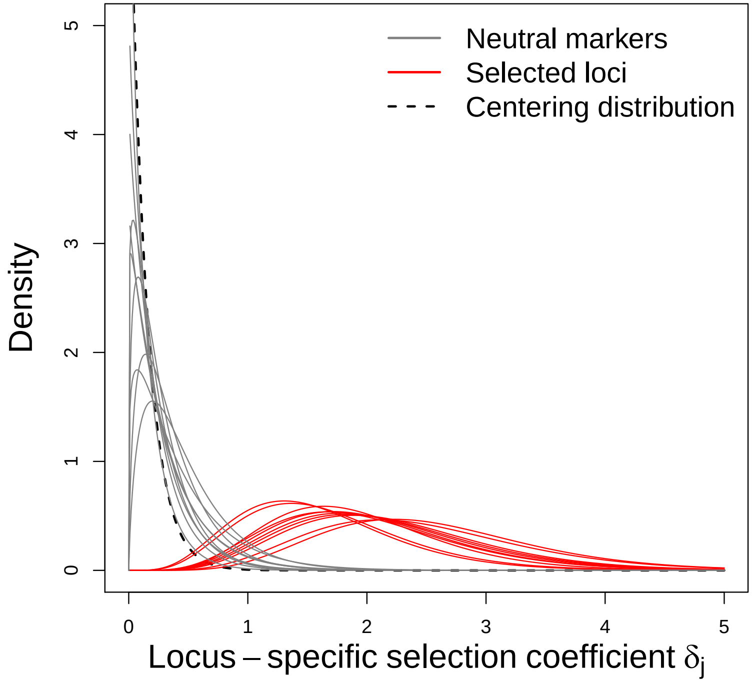


Image courtesy of Peter Beerli, Florida State University, USA.

Credits : Excoffier et Heckel (2006) *Nature Reviews Genetics* **7** : 745-758

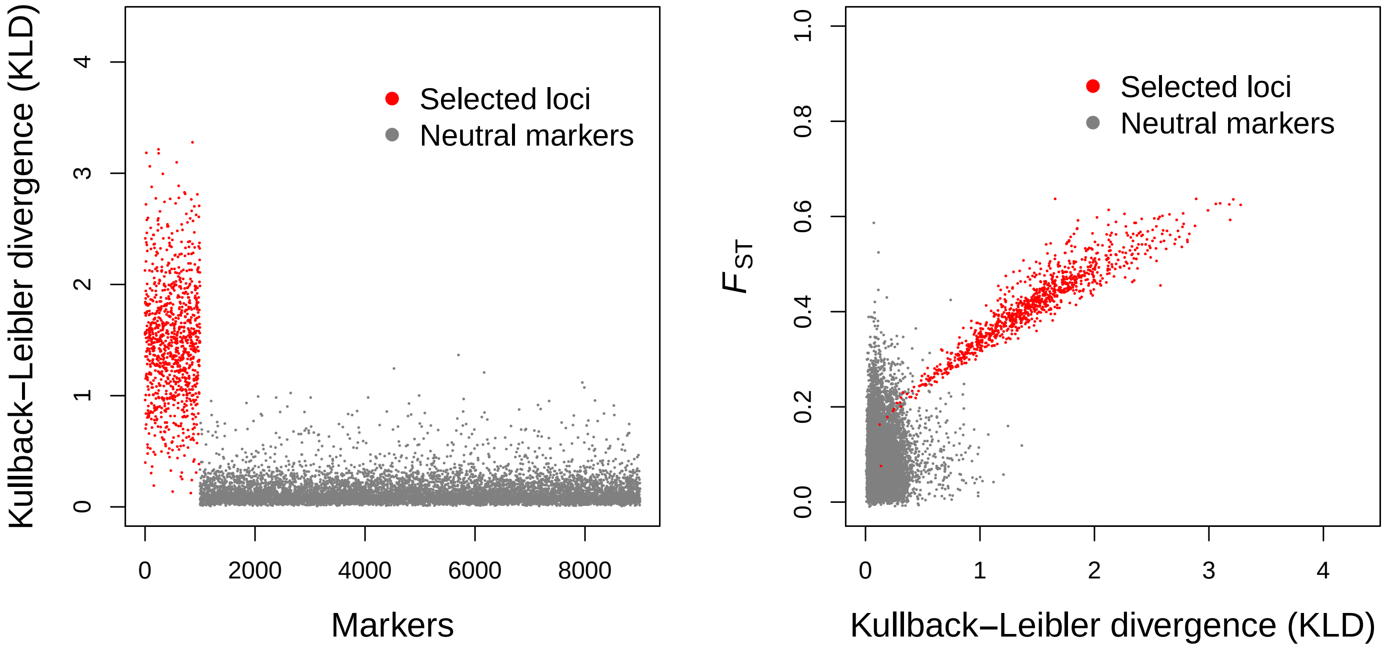
Decision criterion



- We compare the posterior distribution of δ_{ij} to a “centering distribution” that integrates over the overall departure from neutrality

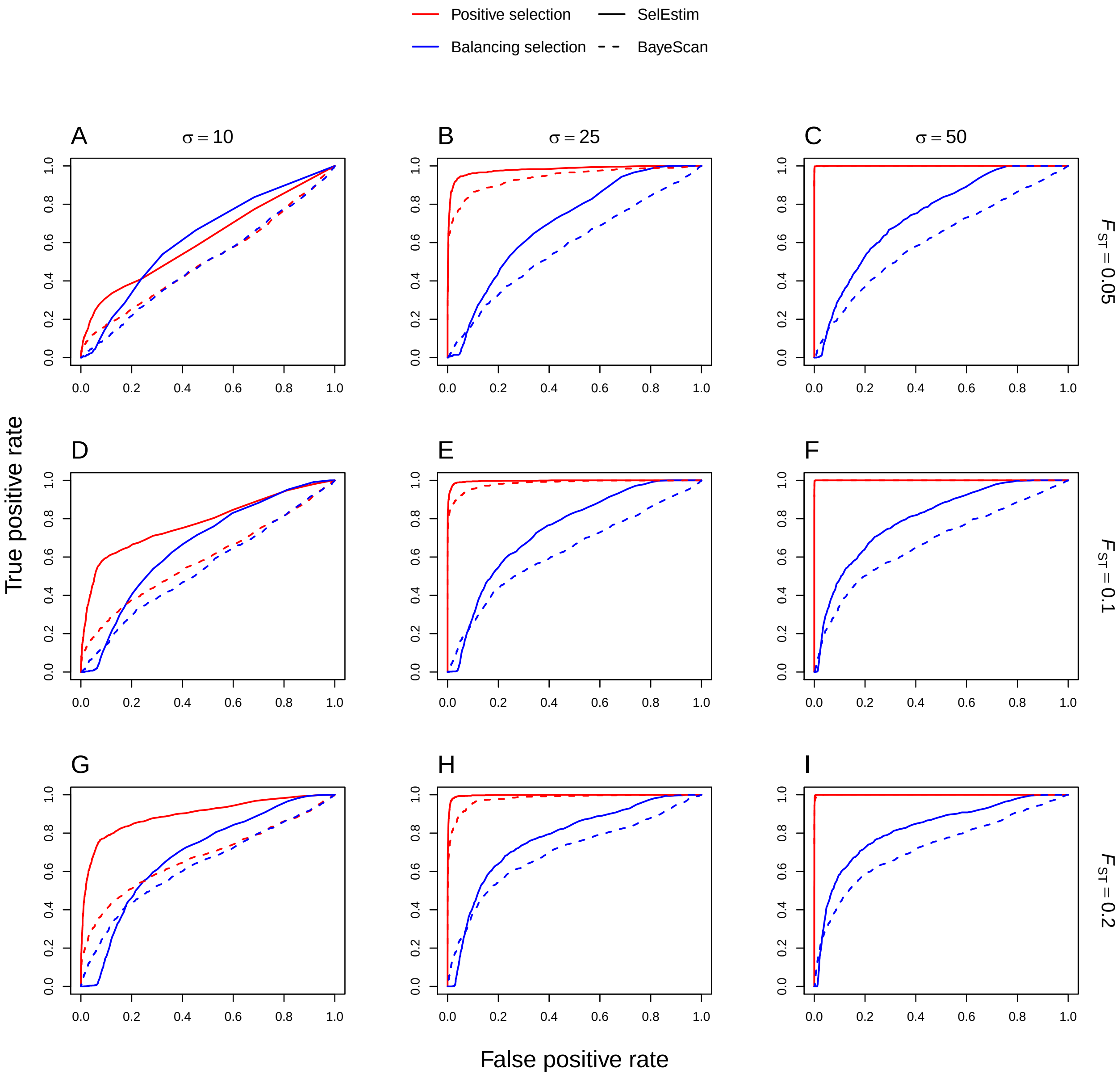
- We use the Kullback-Leibler divergence (KLD) as a distance between these distributions, calibrated using pseudo-observed datasets

Simulation-based tests

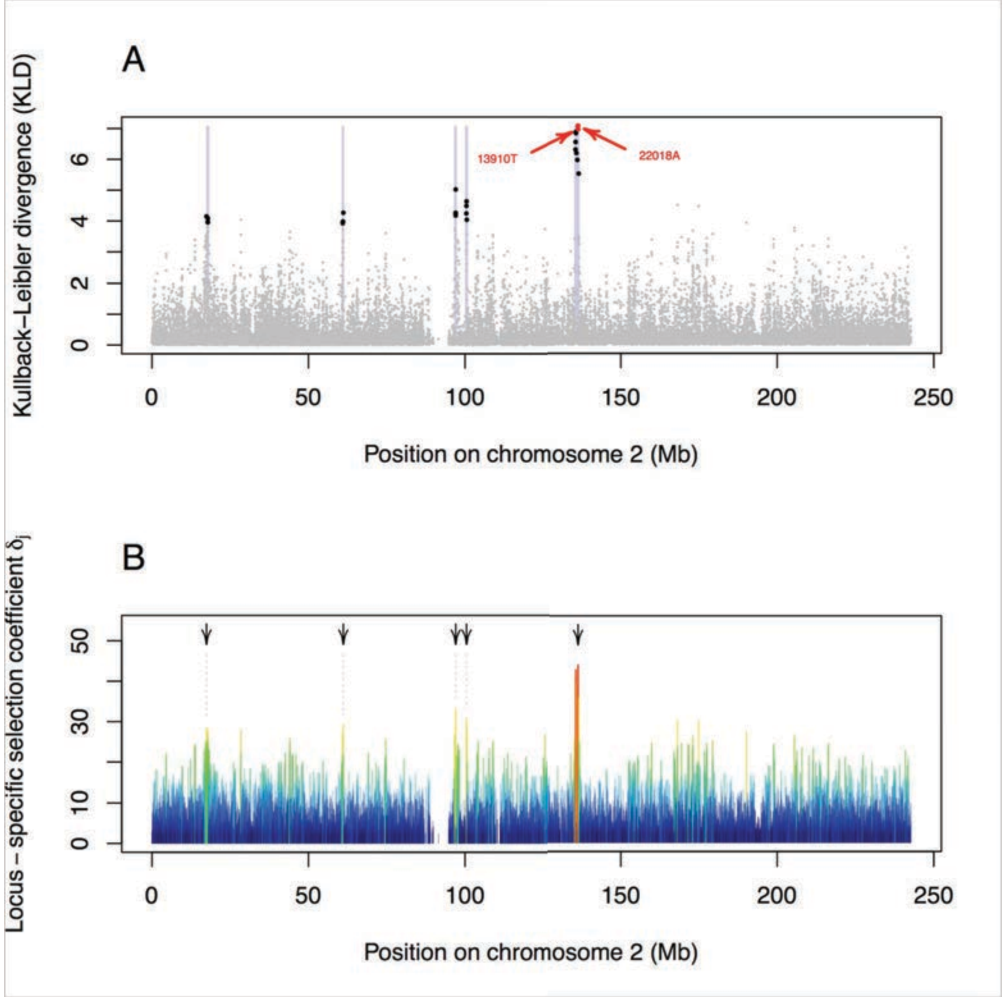


An example of application on simulated data (F_{ST} = 0.10): the distribution of the KLD for positively selected markers departs from that of neutral markers (and correlates with F_{ST}).

Comparison with BayeScan

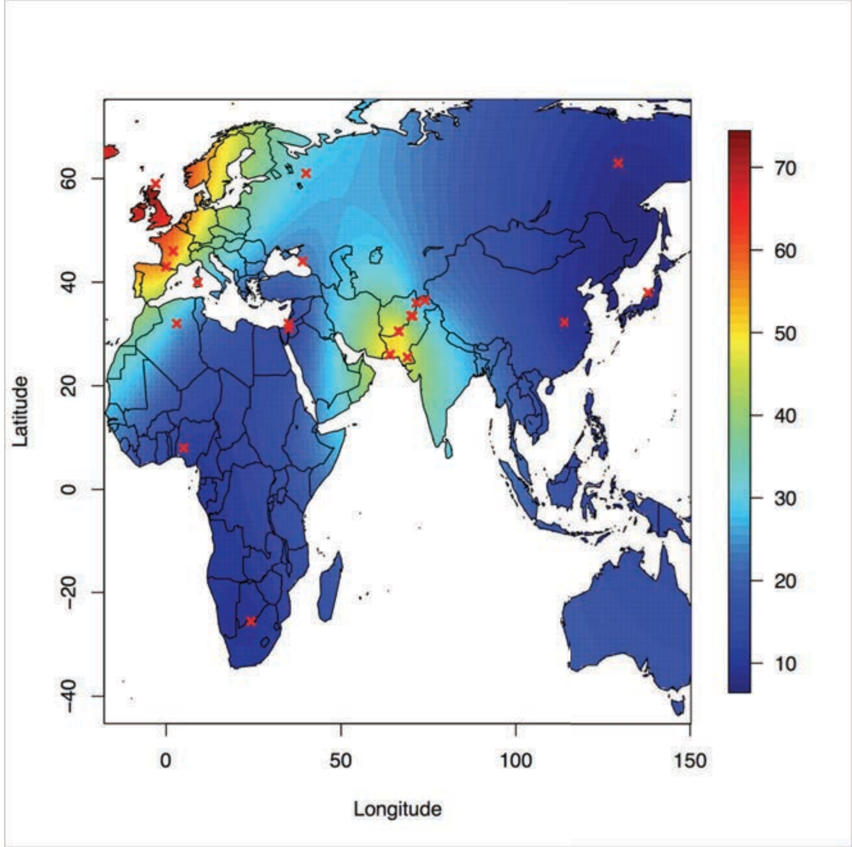


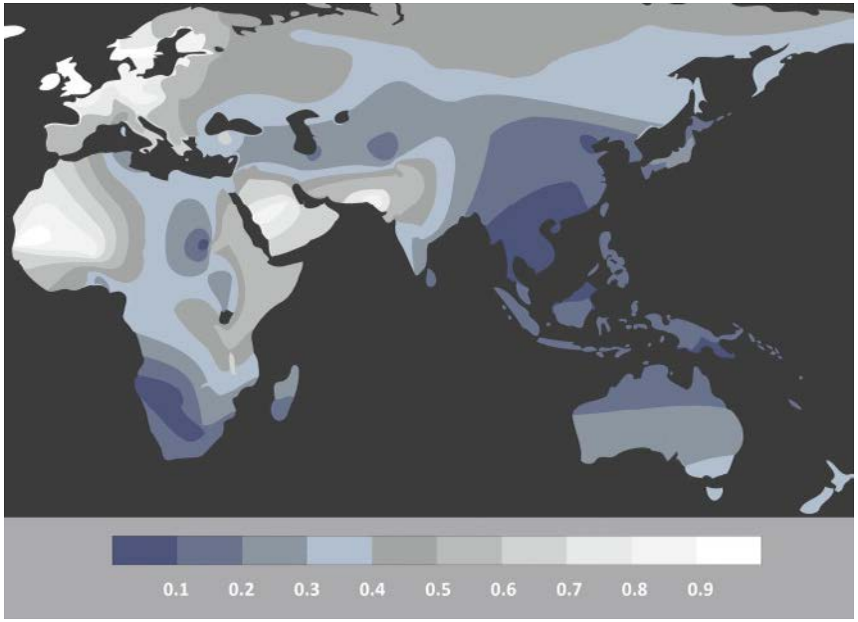
Application on human data (CEPH)



- Strong signature of selection in the vicinity of the lactase gene *LCT*
- Strongest KLD at at 2 SNPs reported to be tightly associated with lactase persistence (13910T and 22018A; see Bersaglieri *et al*. 2004)

Application on human data (CEPH)

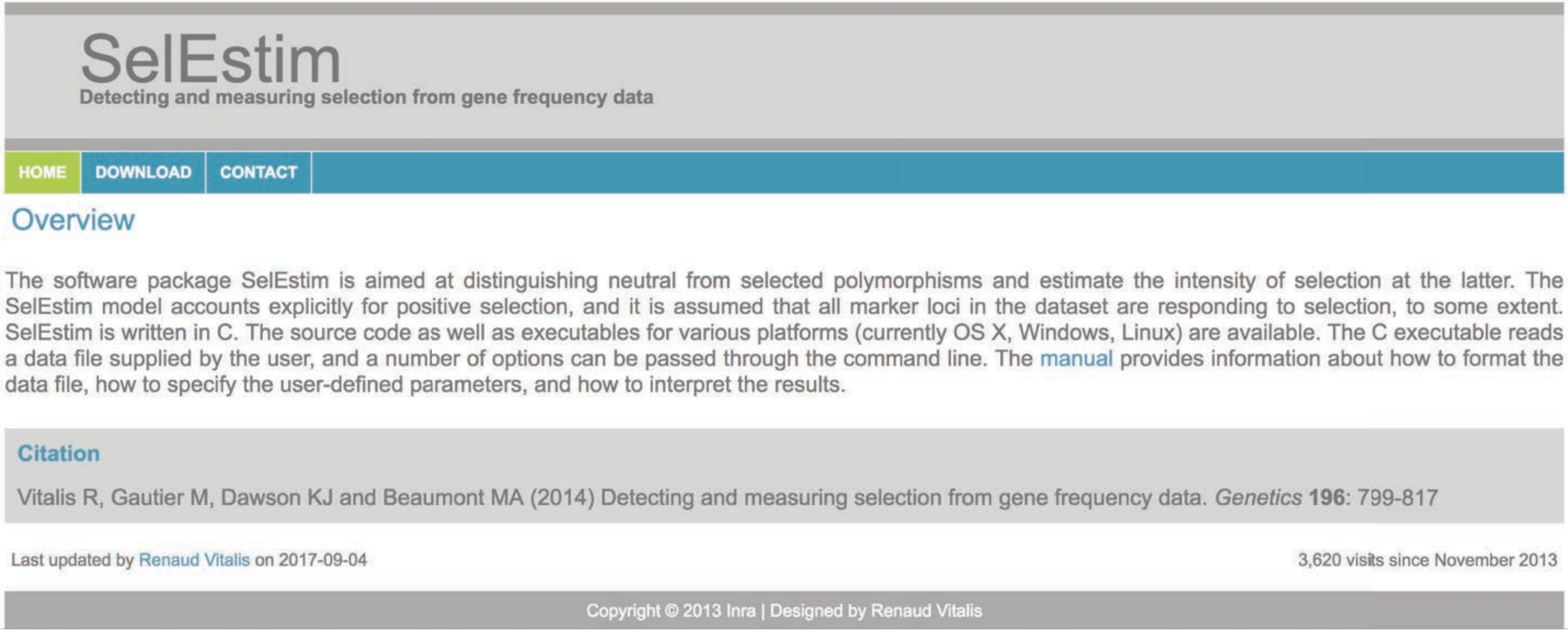




Distribution of lactase persistence phenotype (Itan *et al*. 2010)

- Population-specific selection coefficient at 13910T (left) correlates with lactase persistence phenotype, particularly in Europe and the Indus valley

A software package



Overview

The software package SelEstim is aimed at distinguishing neutral from selected polymorphisms and estimate the intensity of selection at the latter. The SelEstim model accounts explicitly for positive selection, and it is assumed that all marker loci in the dataset are responding to selection, to some extent. SelEstim is written in C. The source code as well as executables for various platforms (currently OS X, Windows, Linux) are available. The C executable reads a data file supplied by the user, and a number of options can be passed through the command line. The manual provides information about how to format the data file, how to specify the user-defined parameters, and how to interpret the results.

Citation

Vitalis R, Gautier M, Dawson KJ and Beaumont MA (2014) Detecting and measuring selection from gene frequency data. *Genetics* **196**: 799-817

Last updated by Renaud Vitalis on 2017-09-04

3,620 visits since November 2013

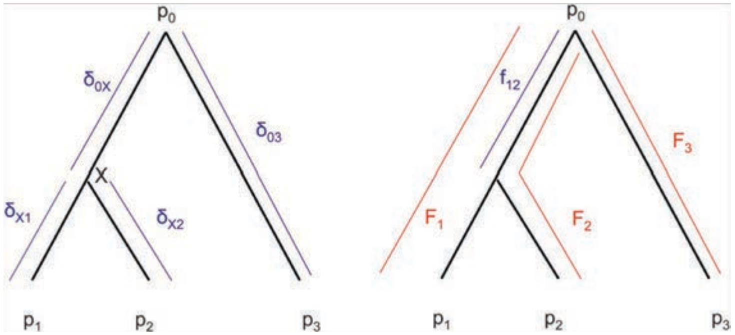
Copyright © 2013 Inra | Designed by Renaud Vitalis

A command-line, parallelized (OpenMP), interface:
http://www1.montpellier.inra.fr/CBGP/software/selestim/index.html

How to use the information brought by haplotype structure?

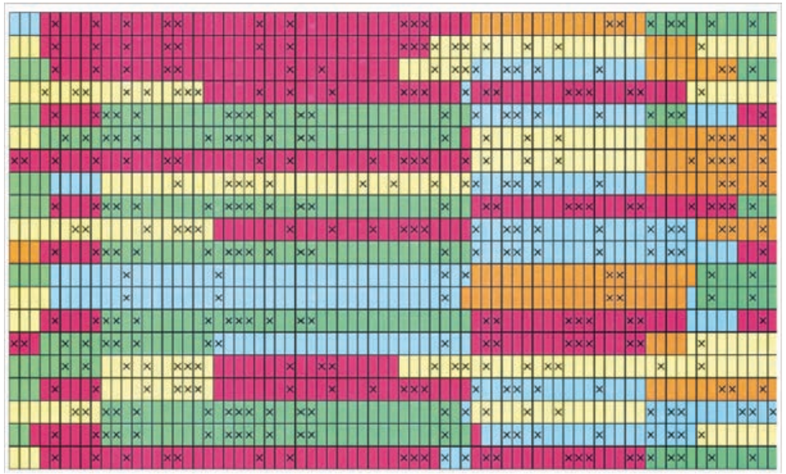
Valentin Hivert’s PhD (2015 – 2018)

F_{ST}-based tests using haplotype data



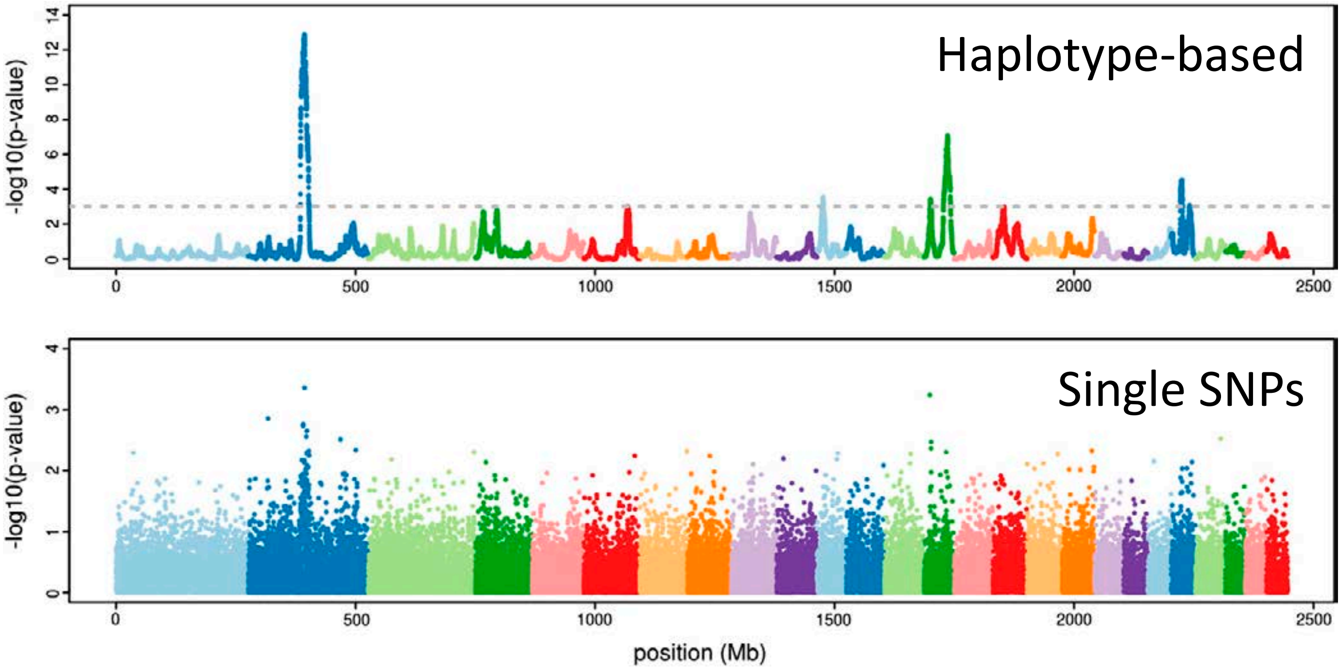
Bonhomme *et al*. (2010): a generalization of the Lewontin-Krakauer test that accounts for a tree-like history

Fariello *et al*. (2013): application on haplotypes obtained by local clustering (fastPHASE)



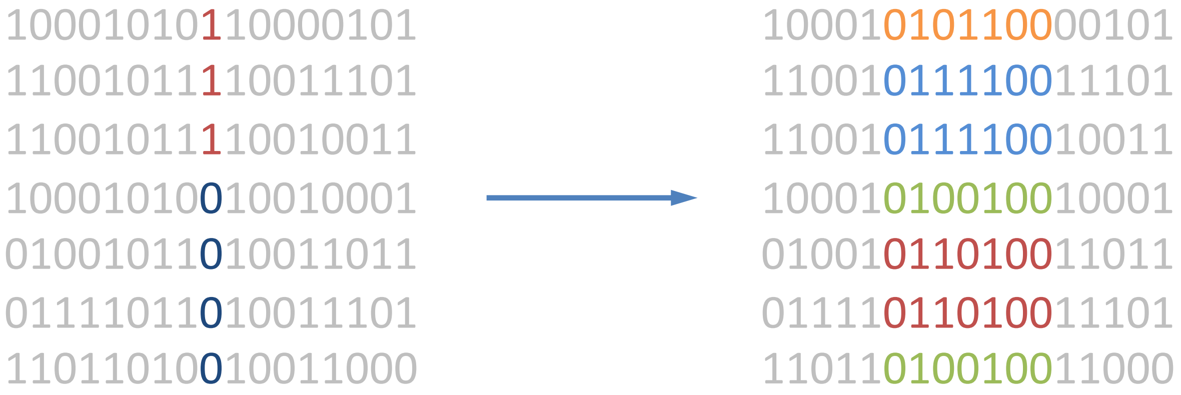
Credits: Fariello *et al*. (2013) *Genetics* **193**:929-941

F_{ST}-based tests using haplotype data



Credits: Fariello *et al*. (2013) *Genetics* **193**:929-941

SELESTIM with haplotypes



10001010110000101
11001011110011101
11001011110010011
10001010010010001
01001011010011011
01111011010011101
11011010010011000

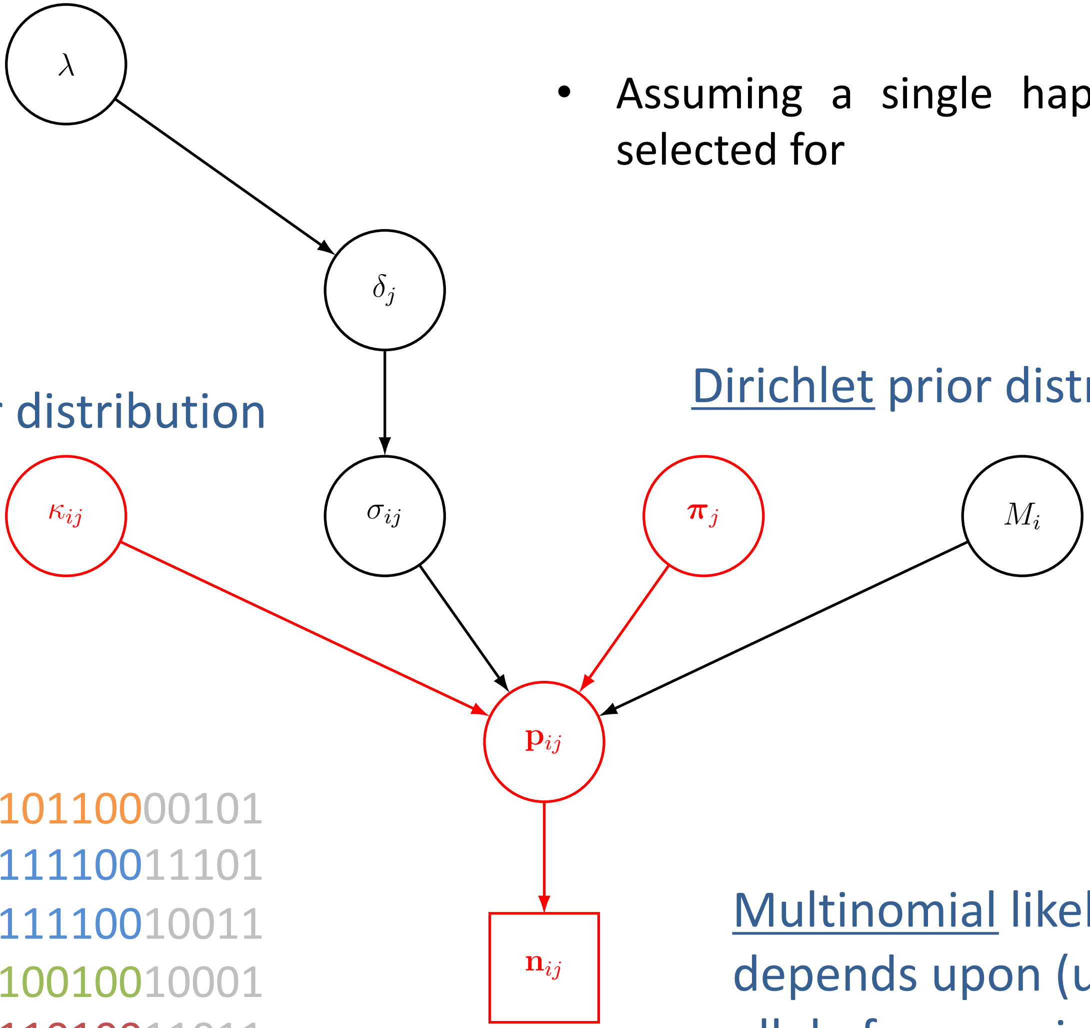
10001010110000101
11001011110011101
11001011110010011
10001010010010001
01001011010011011
01111011010011101
11011010010011000

SNPs

haplotypes
(multiallelic markers)

SelEstim with haplotypes

- Assuming a single haplotype is selected for

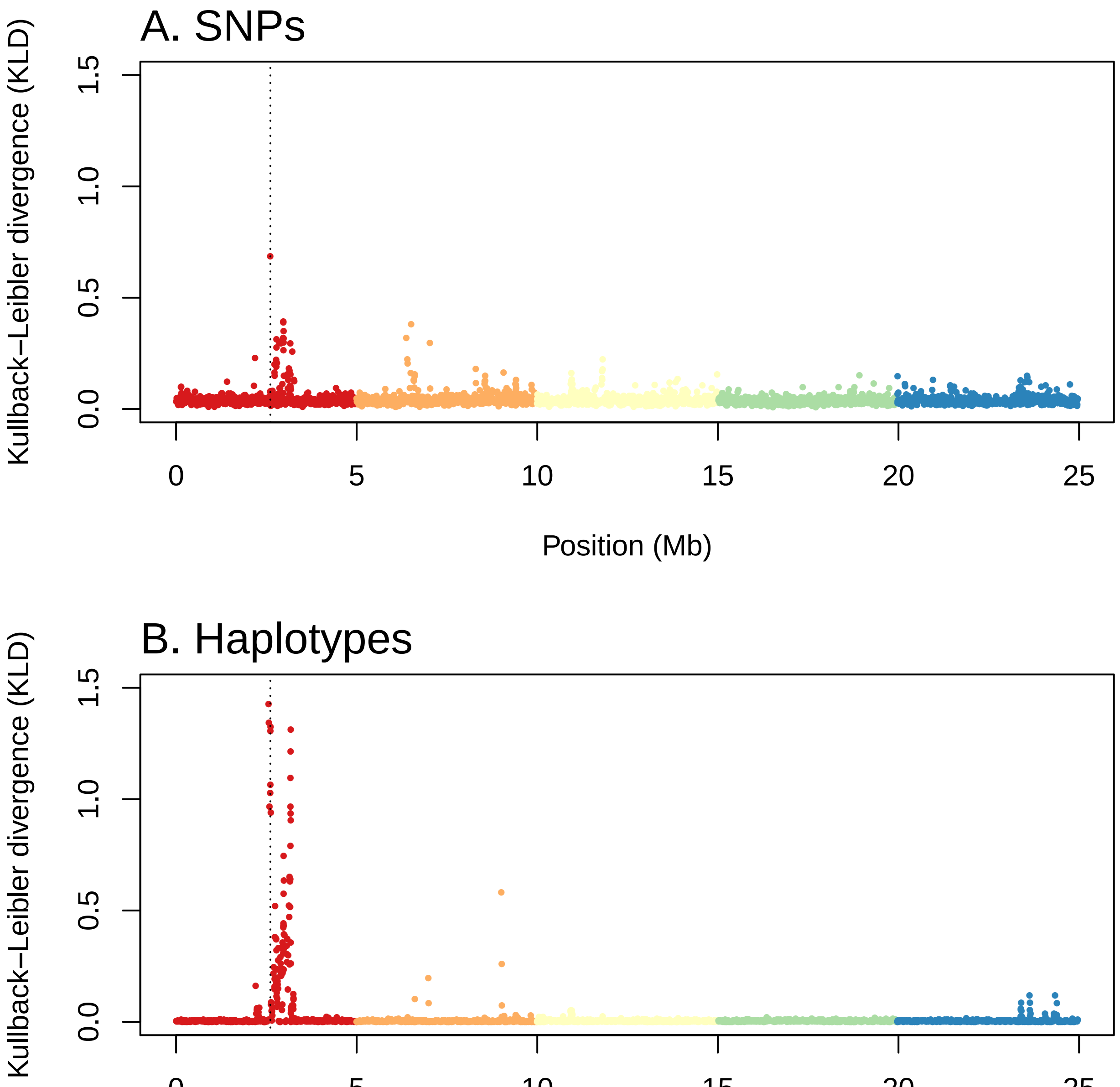


Categorical prior distribution

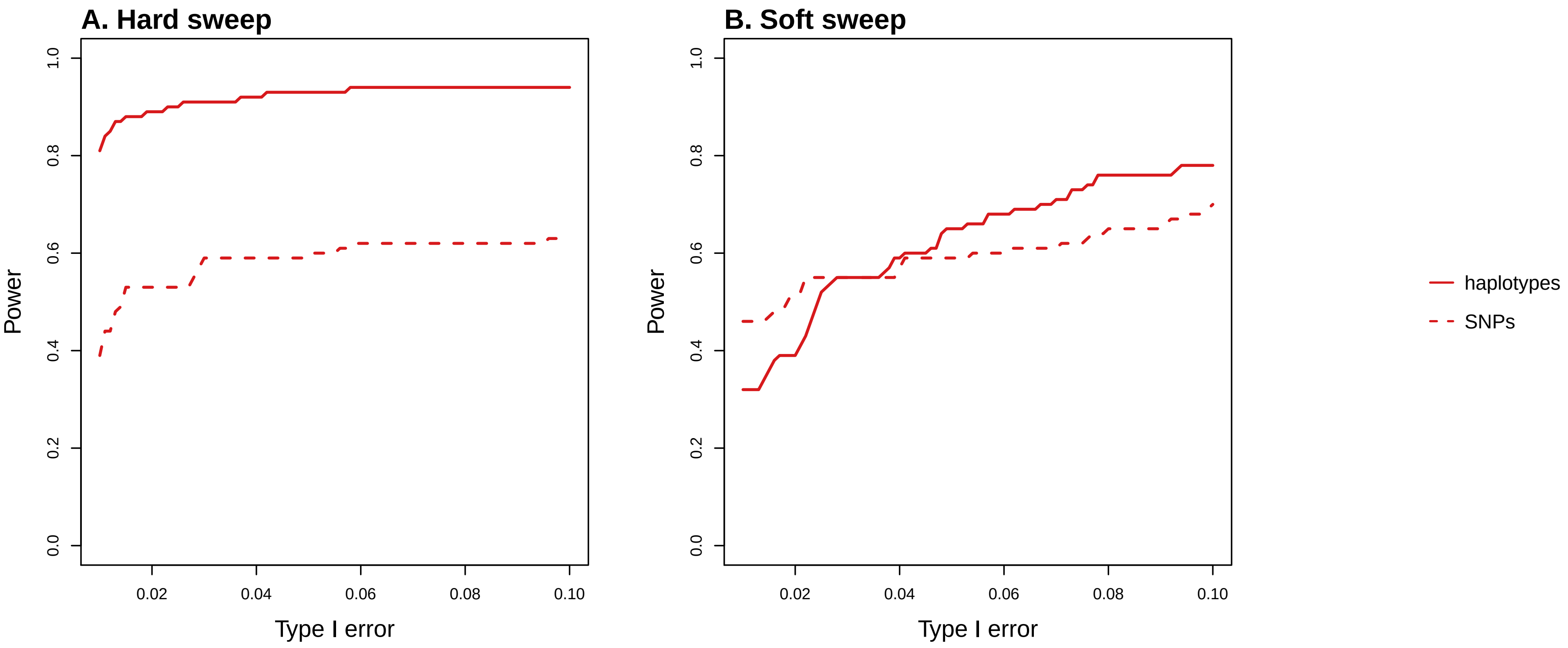
Dirichlet prior distribution

10001010110000101
11001011110011101
11001011110010011
10001010010010001
01001011010011011
01111011010011101
11011010010011000

Multinomial likelihood that depends upon (unknown) allele frequencies

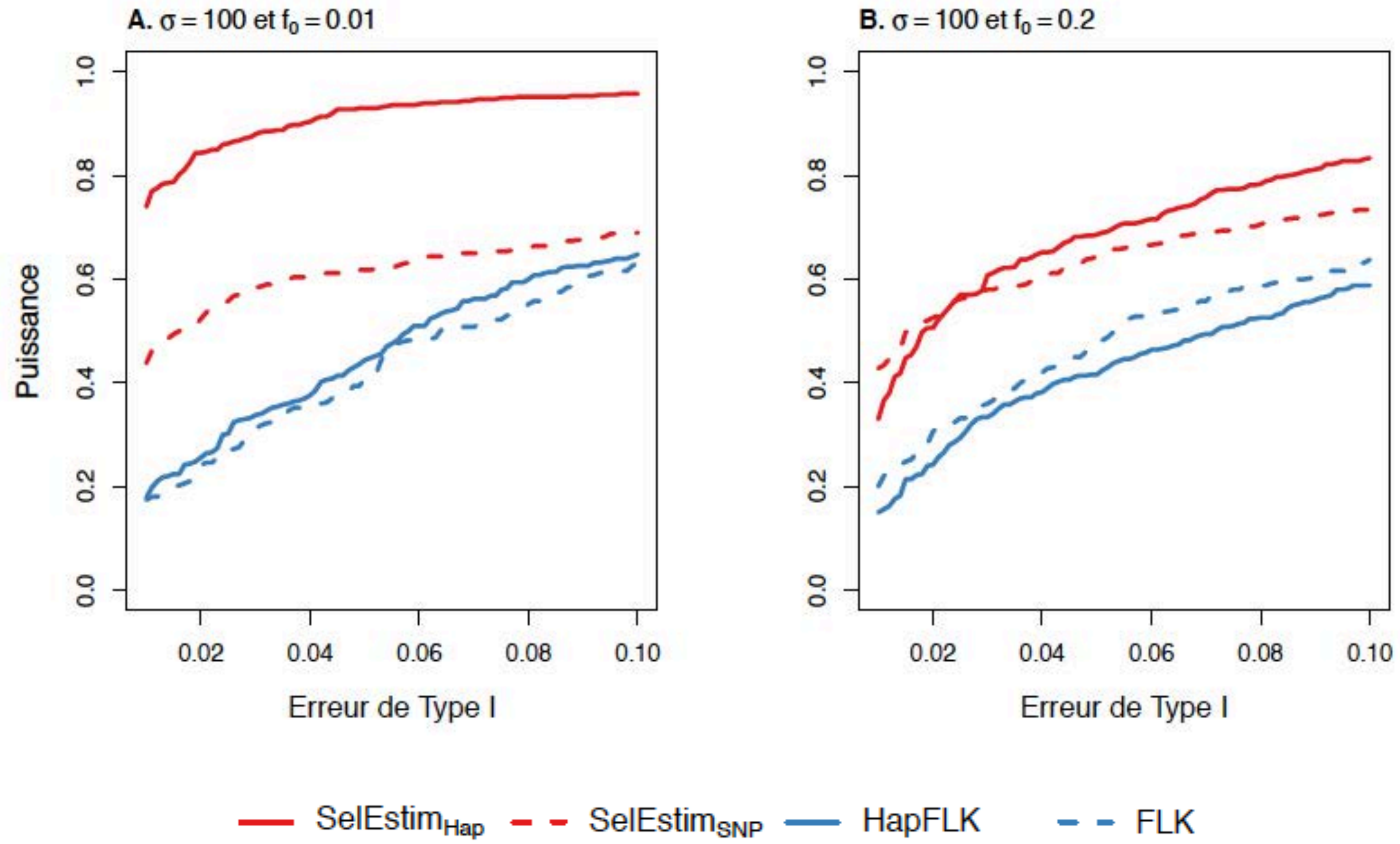

A. SNPs
Kullback–Leibler divergence (KLD)
0.0
0.5
1.0
1.5
0
5
10
15
20
25
Position (Mb)
B. Haplotypes
Kullback–Leibler divergence (KLD)
0.0
0.5
1.0
1.5
0
5
10
15
20
25
Position (Mb)

Performance in the island model



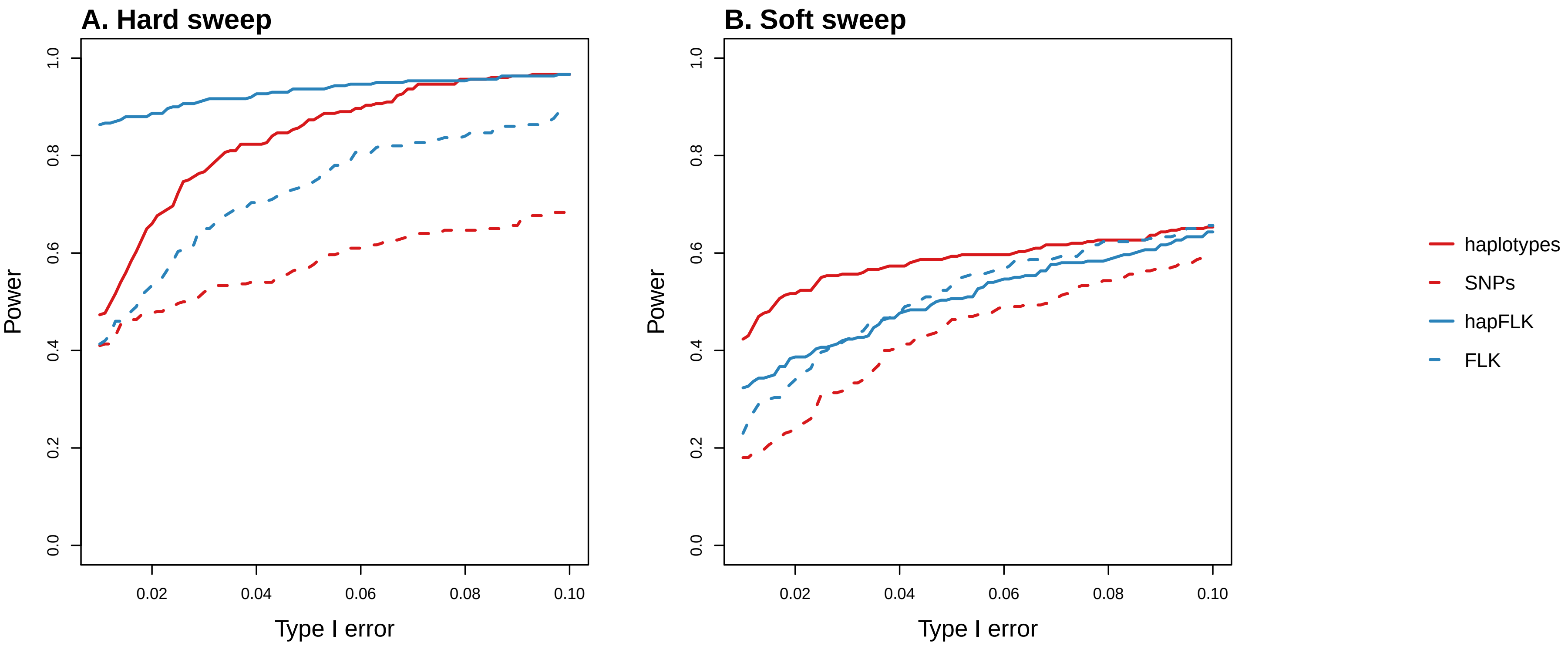
- Improved statistical power with haplotype-based analyses (*vs*. SNPs)

Performance in the island model



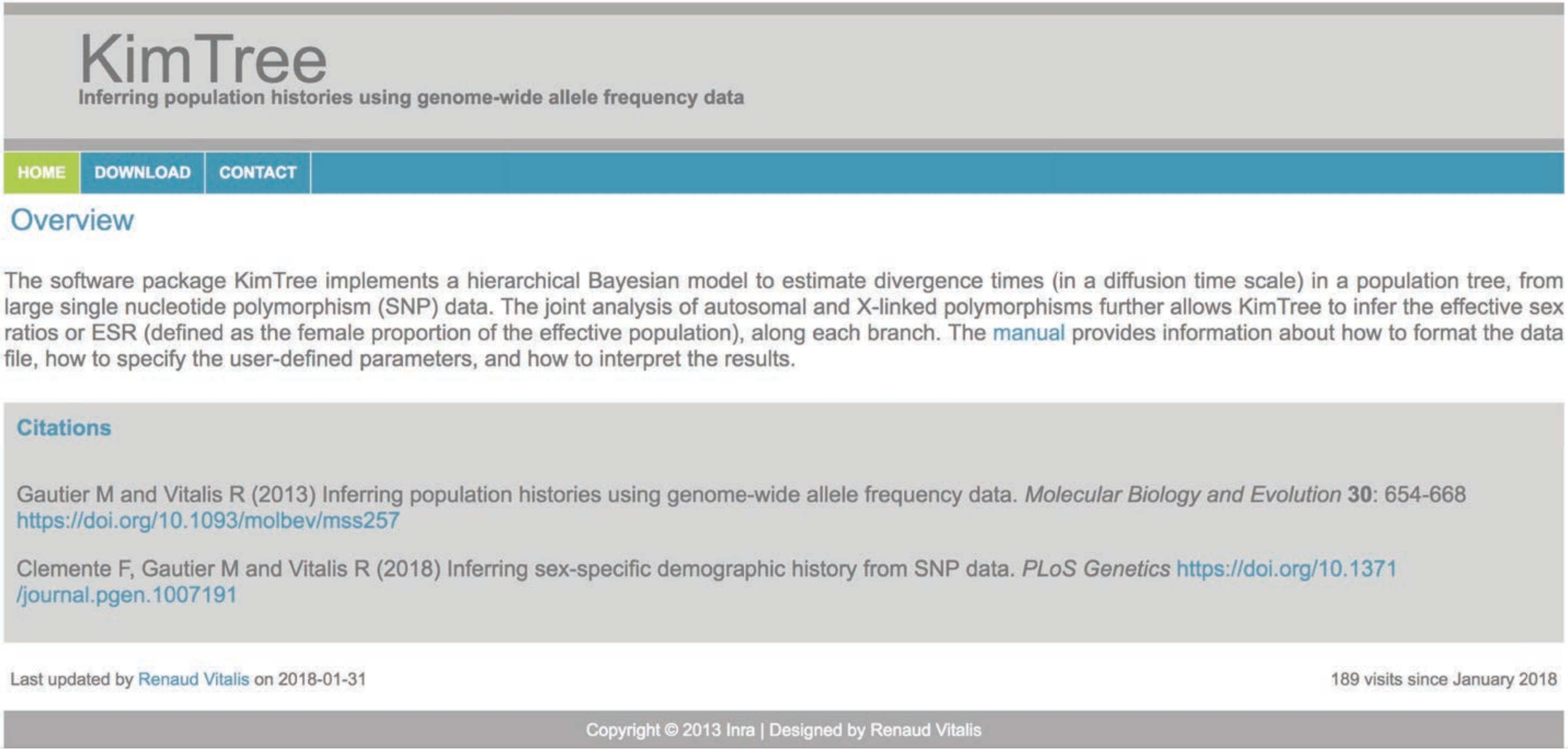
- Improved statistical power with haplotype-based analyses (*vs*. SNPs)
- Better performance than FLK (Bonhomme *et al*. 2010) and hapFLK (Fariello *et al*. 2013)

Performance in divergence models



- Improved statistical power with haplotype-based analyses (*vs*. SNPs)
- Poorer performance than FLK (Bonhomme *et al*. 2010) and hapFLK (Fariello *et al*. 2013)

A software package



KimTree

Inferring population histories using genome-wide allele frequency data

Overview

The software package KimTree implements a hierarchical Bayesian model to estimate divergence times (in a diffusion time scale) in a population tree, from large single nucleotide polymorphism (SNP) data. The joint analysis of autosomal and X-linked polymorphisms further allows KimTree to infer the effective sex ratios or ESR (defined as the female proportion of the effective population), along each branch. The manual provides information about how to format the data file, how to specify the user-defined parameters, and how to interpret the results.

Citations

Gautier M and Vitalis R (2013) Inferring population histories using genome-wide allele frequency data. *Molecular Biology and Evolution* **30**: 654-668 https://doi.org/10.1093/molbev/mss257

Clemente F, Gautier M and Vitalis R (2018) Inferring sex-specific demographic history from SNP data. *PLoS Genetics* https://doi.org/10.1371/journal.pgen.1007191

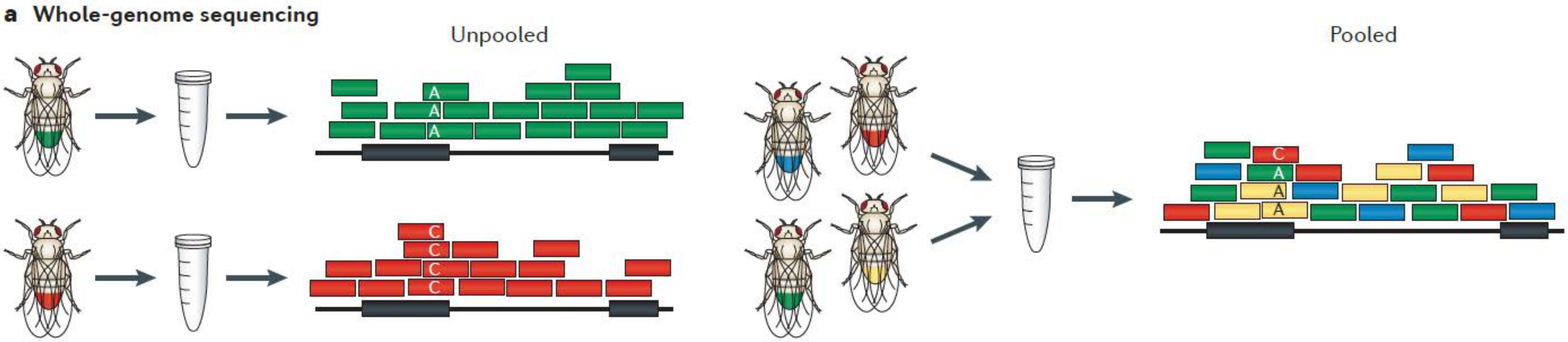
Last updated by Renaud Vitalis on 2018-01-31

189 visits since January 2018

Copyright © 2013 Inra | Designed by Renaud Vitalis

A command-line, parallelized (OpenMP), interface:
http://www1.montpellier.inra.fr/CBGP/software/kimtree/index.html

A digression on Pool-seq…

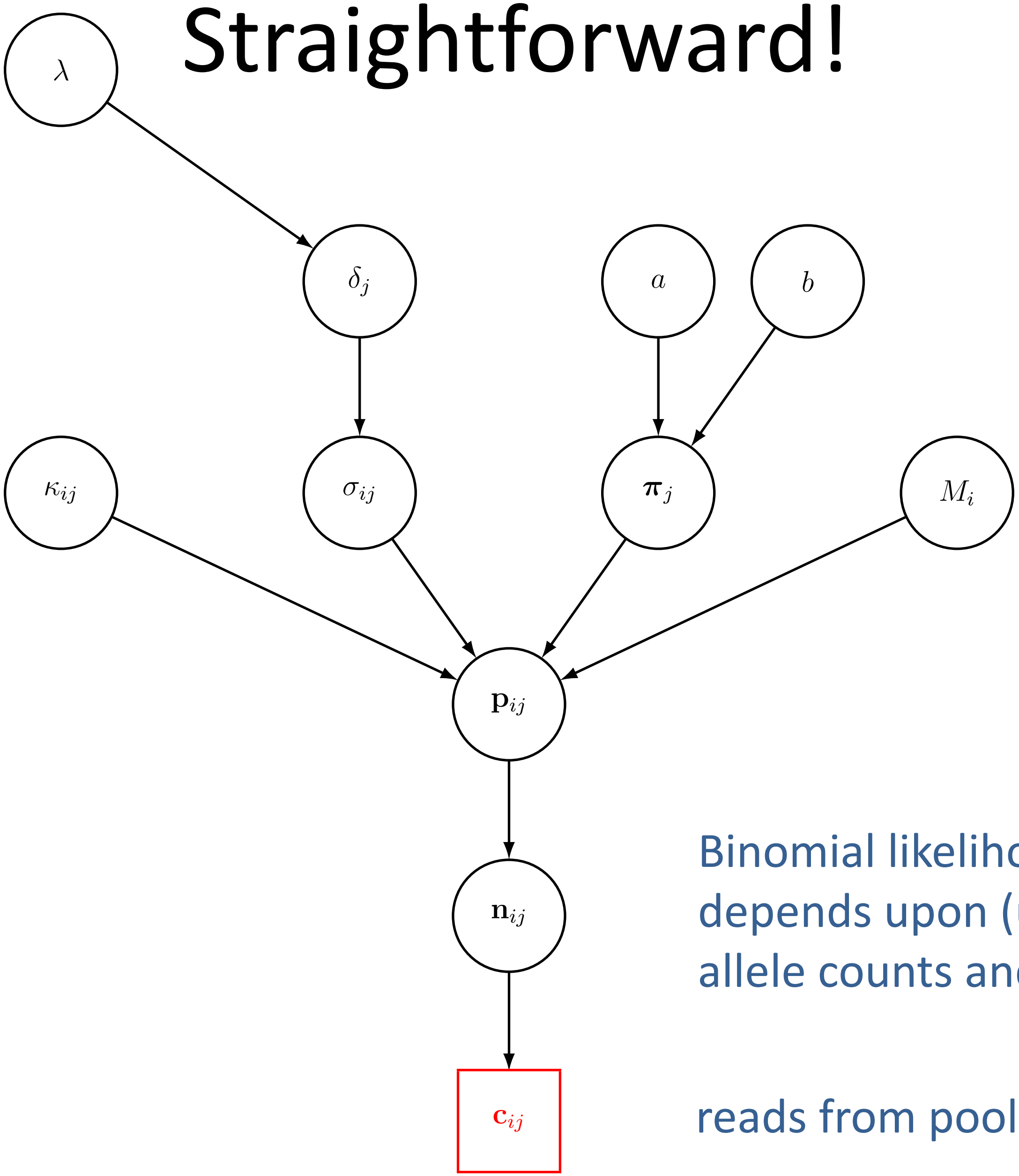


Individual genotypes

Read numbers for the entire pool

Credits: Schlötterer *et al*. (2014) *Nature Reviews Genetics* **15**: 749-763

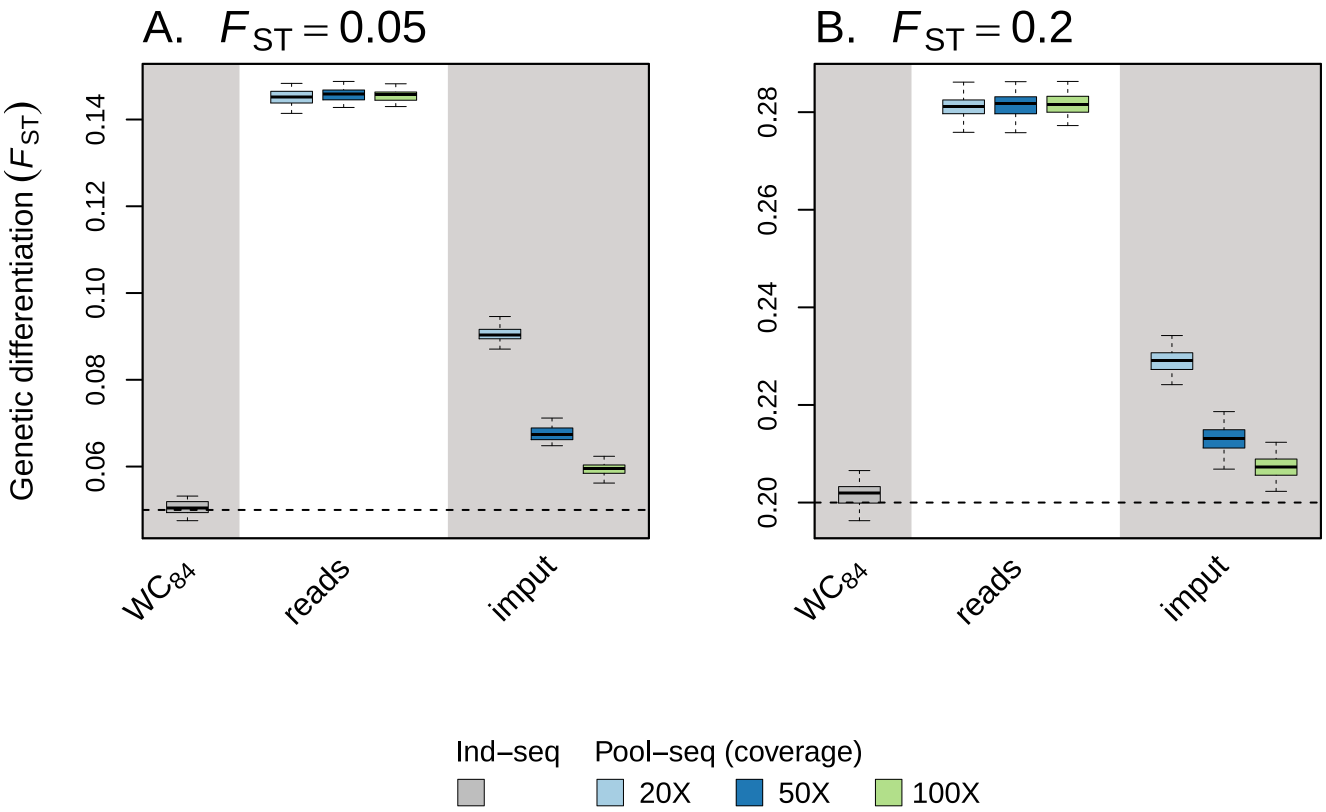
Straightforward!



Binomial likelihood that depends upon (unknown) allele counts and coverage

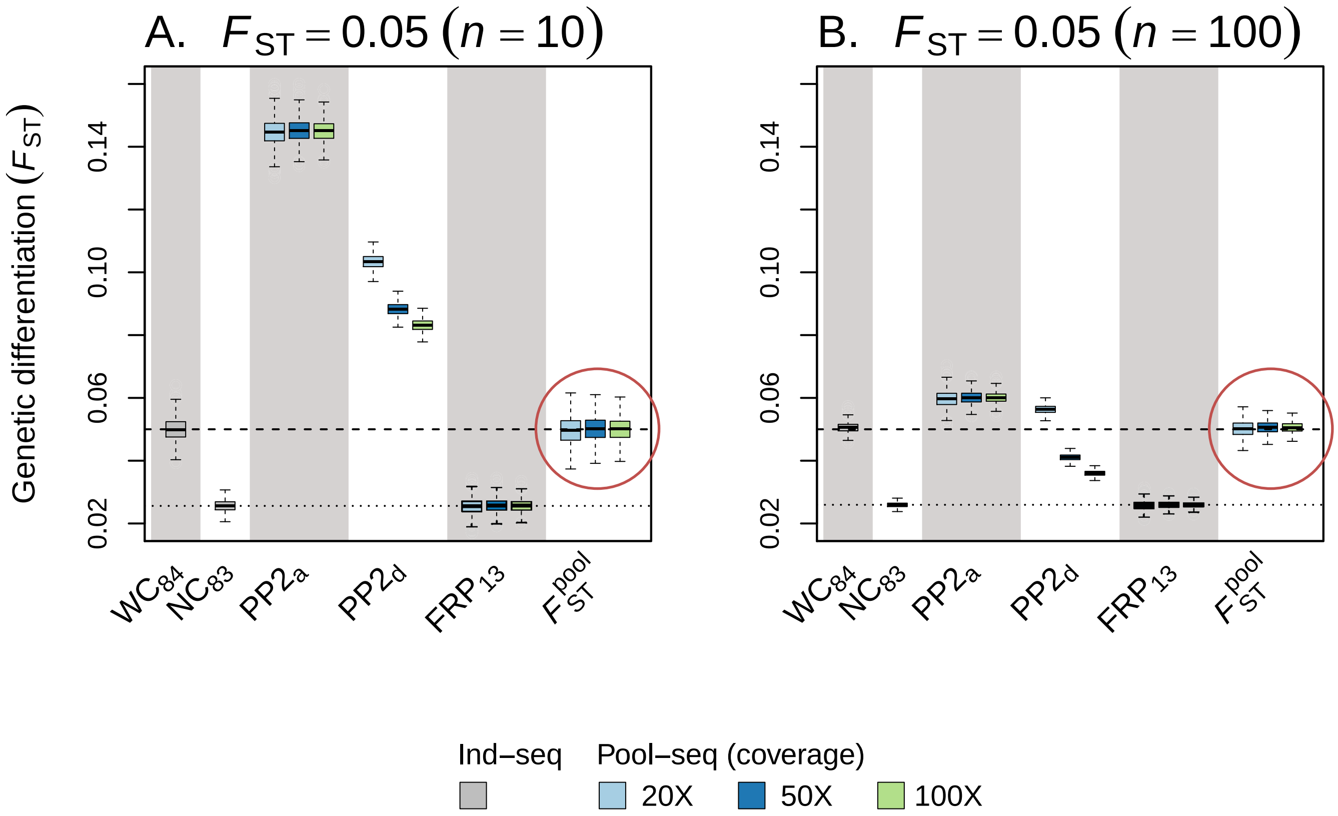
reads from pooled samples

More tricky: F_{ST} from pooled data…



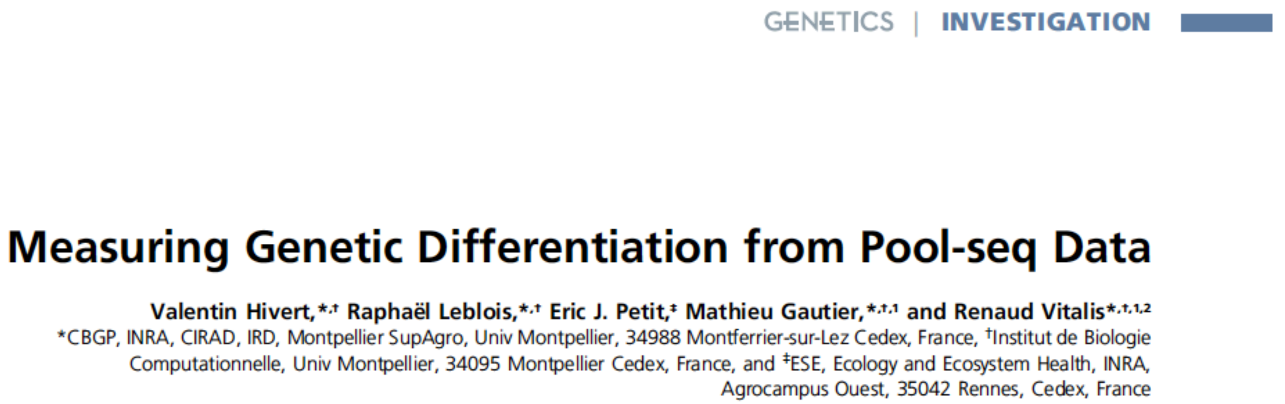
- Naive approaches may fail…
  - ✓ Considering reads as allele counts
  - ✓ Imputing allele counts using a maximum likelihood argument

A new estimator of F_{ST} for pooled data



- Method-of-moments estimator, based on an analysis-of-variance framework
  - ✓ No bias
  - ✓ Performs better than any other estimator available (PoPoolation2, etc.)

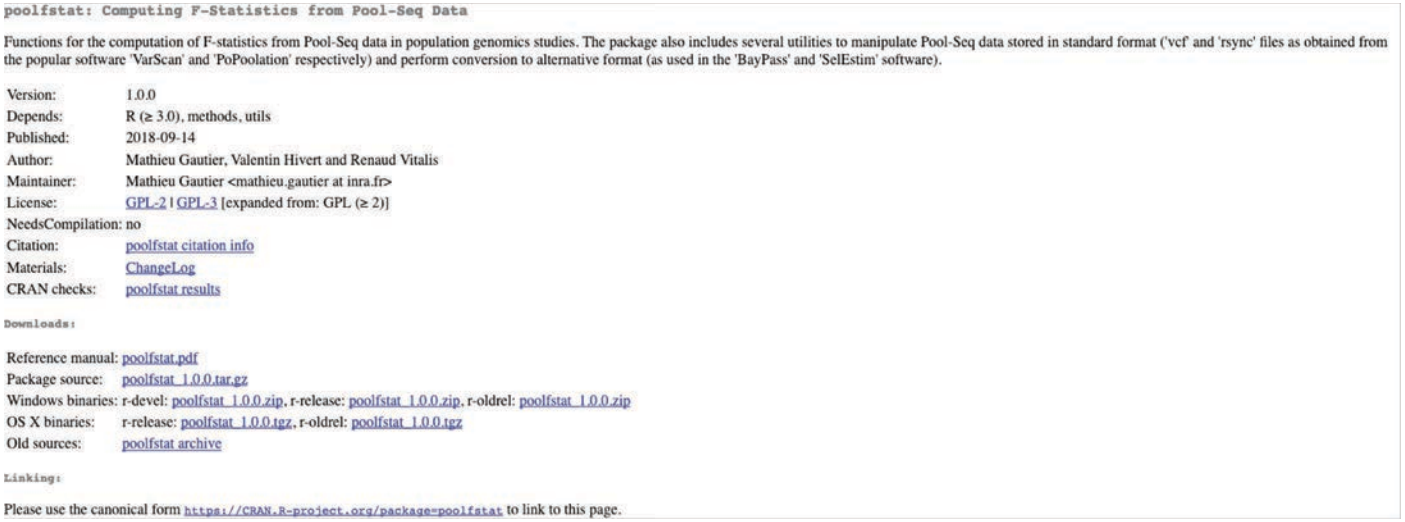
A new estimator of F_{ST} for pooled data



GENETICS | INVESTIGATION

Measuring Genetic Differentiation from Pool-seq Data

Valentin Hivert,*,† Raphaël Leblois,*,† Eric J. Petit,‡ Mathieu Gautier,*,†,1 and Renaud Vitalis*,†,1,2
*CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ Montpellier, 34988 Montferrier-sur-Lez Cedex, France, †Institut de Biologie Computationnelle, Univ Montpellier, 34095 Montpellier Cedex, France, and ‡ESE, Ecology and Ecosystem Health, INRA, Agrocampus Ouest, 35042 Rennes, Cedex, France



poolfstat: Computing F-Statistics from Pool-Seq Data

Functions for the computation of F-statistics from Pool-Seq data in population genomics studies. The package also includes several utilities to manipulate Pool-Seq data stored in standard format ('vcf' and 'rsync' files as obtained from the popular software 'VarScan' and 'PoPoolation' respectively) and perform conversion to alternative format (as used in the 'BayPass' and 'SelEstim' software).

Version:	1.0.0
Depends:	R (≥ 3.0), methods, utils
Published:	2018-09-14
Author:	Mathieu Gautier, Valentin Hivert and Renaud Vitalis
Maintainer:	Mathieu Gautier <mathieu.gautier at inra.fr>
License:	GPL-2 | GPL-3 [expanded from: GPL (≥ 2)]
NeedsCompilation:	no
Citation:	poolfstat citation info
Materials:	ChangeLog
CRAN checks:	poolfstat results

Downloads:

Reference manual:	poolfstat.pdf
Package source:	poolfstat_1.0.0.tar.gz
Windows binaries:	r-devel: poolfstat_1.0.0.zip, r-release: poolfstat_1.0.0.zip, r-oldrel: poolfstat_1.0.0.zip
OS X binaries:	r-release: poolfstat_1.0.0.tgz, r-oldrel: poolfstat_1.0.0.tgz
Old sources:	poolfstat archive

Linking:

Please use the canonical form `https://CRAN.R-project.org/package=poolfstat` to link to this page.

Hivert *et al*. (2018) *Genetics* **210**: 315-330

Take home messages

- All these methods are designed to identify overly differentiated marker loci: local adaptation or intrinsic genetic incompatibilities?

- Be aware of the underlying population models and assumptions (e.g., island model *vs*. divergence models) and the robustness of the methods to model misspecifications

- Pool-seq experiments: random sampling of reads from allele counts must be (properly) accounted for!

Acknowledgements and credits

- Mark A. Beaumont, Florian Clemente (former postdoc), Kevin J. Dawson, Mathieu Gautier, Valentin Hivert (former PhD student)