« Environmental Genetics » doctoral course ABIES-GAIA

SELESTIM: Detecting and Measuring Selection from Gene Frequency Data

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

The data

Single Nucleotide Polymorphisms (SNPs) genotyped in a number of populations.

SNPs are bi-allelic, co-dominant markers.

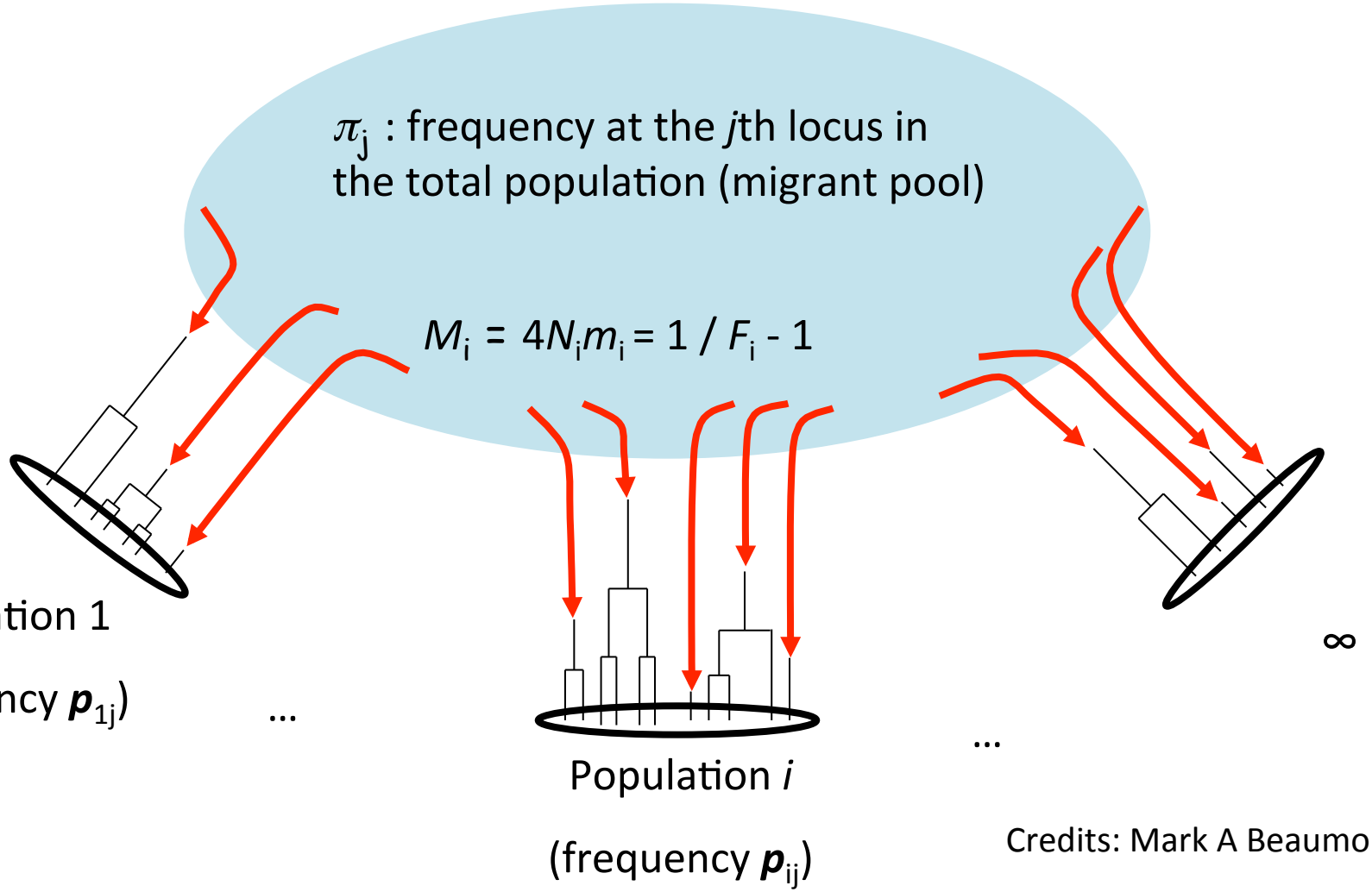
The data consist in allele counts $\boldsymbol{n_{ij}} = (x_{ij}\,;\, n_{ij} - x_{ij})$ at locus *j* in population *i*. The likelihood of a sample of genes reads:

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

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

Population model

We consider an infinite island model of population structure :



Credits: Mark A Beaumont

Selection model

We assume a simple model of selection

We assume that **all** marker loci are targeted by selection

In population *i*, at locus *j*, genotypes *AA*, *Aa* and *aa* have relative fitness:

AA	*Aa*	*aa*
$1 + s_{ij}$	$1 + s_{ij} / 2$	1

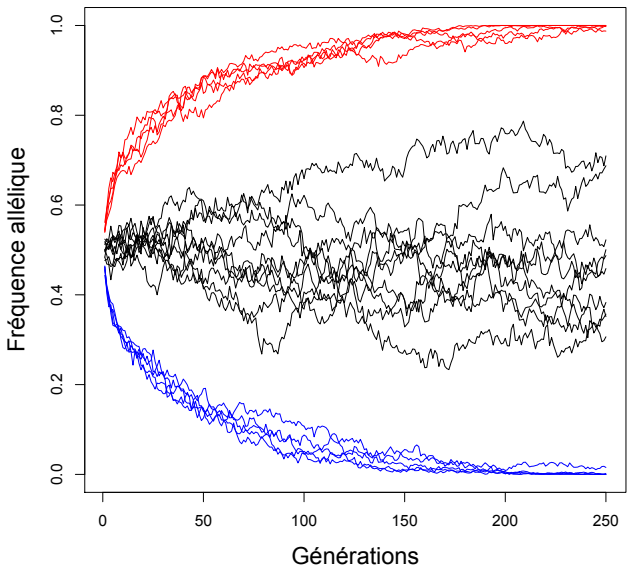
We define the scaled parameter $\sigma_{ij} = 2 N_i s_{ij}$

But which allele is *A*?

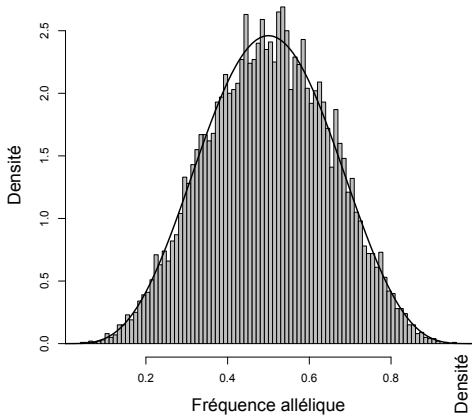
we introduce an indicator variable κ_{ij}: $\kappa_{ij} = 0$ if allele *A* is selected for, $\kappa_{ij} = 1$ if allele *a* is selected for…

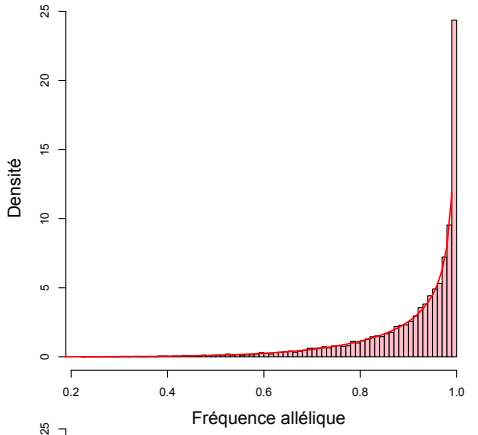
Diffusion approximation

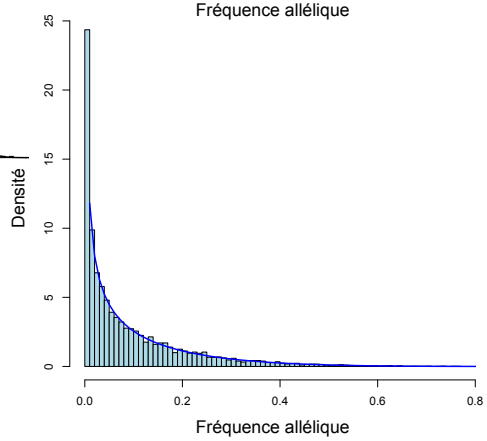
The diffusion approximation of the equilibrium density (plain lines) captures the joint effect of drift, migration and selection (histograms are from simulated data)



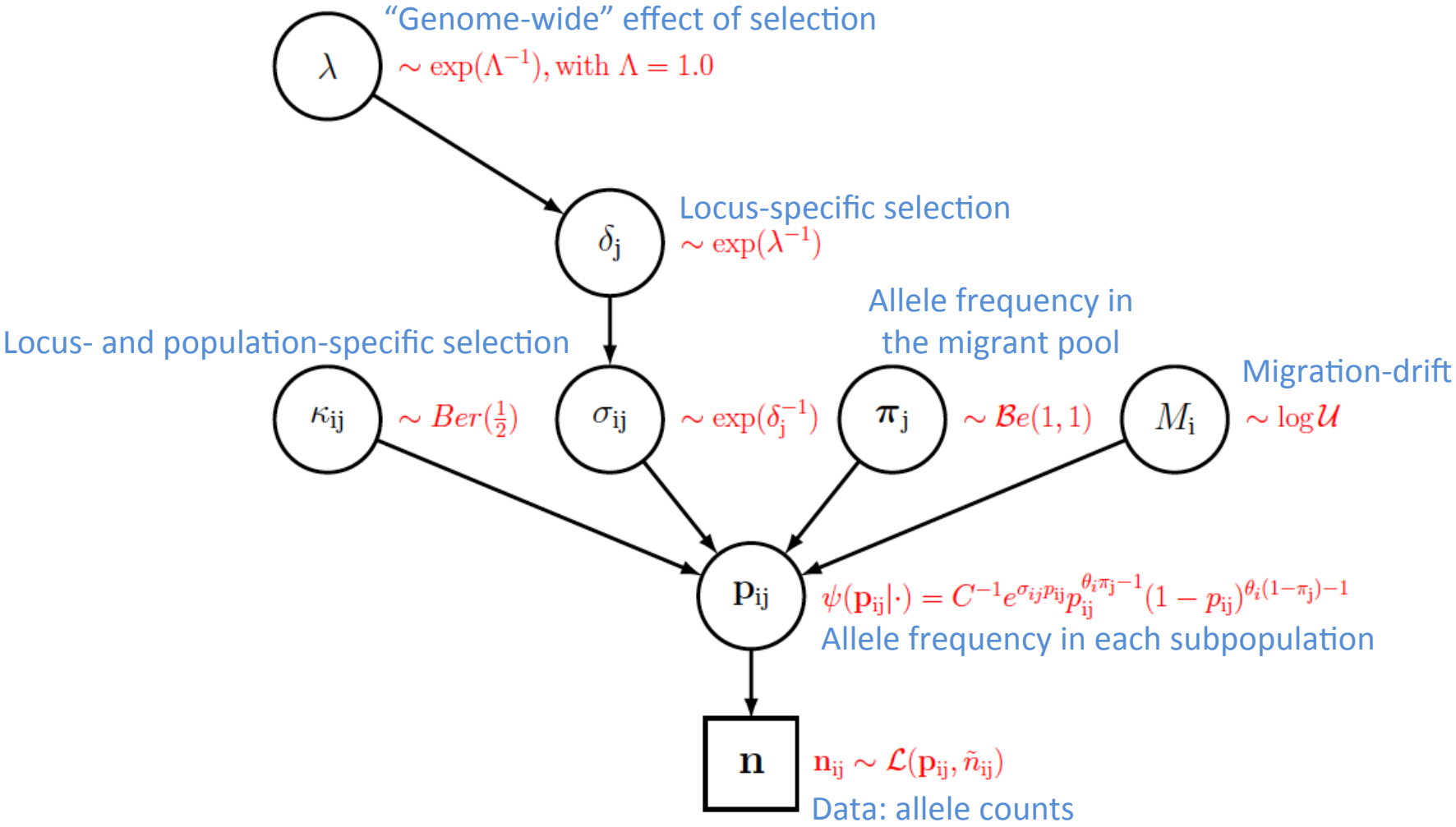
Simulated dynamics





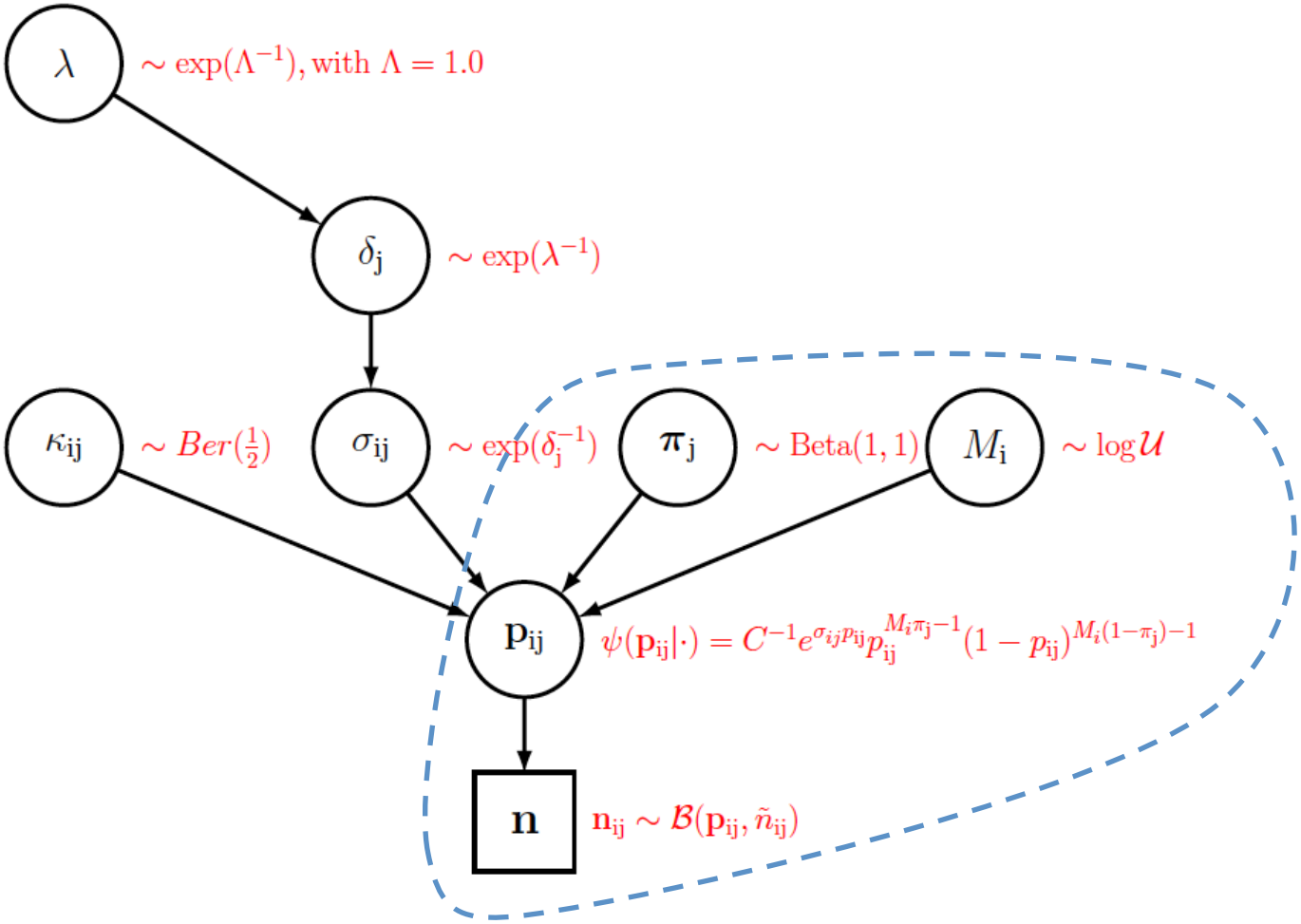


A hiearchical Bayesian model



Directed acyclic graph (DAG)

Relation to previous models

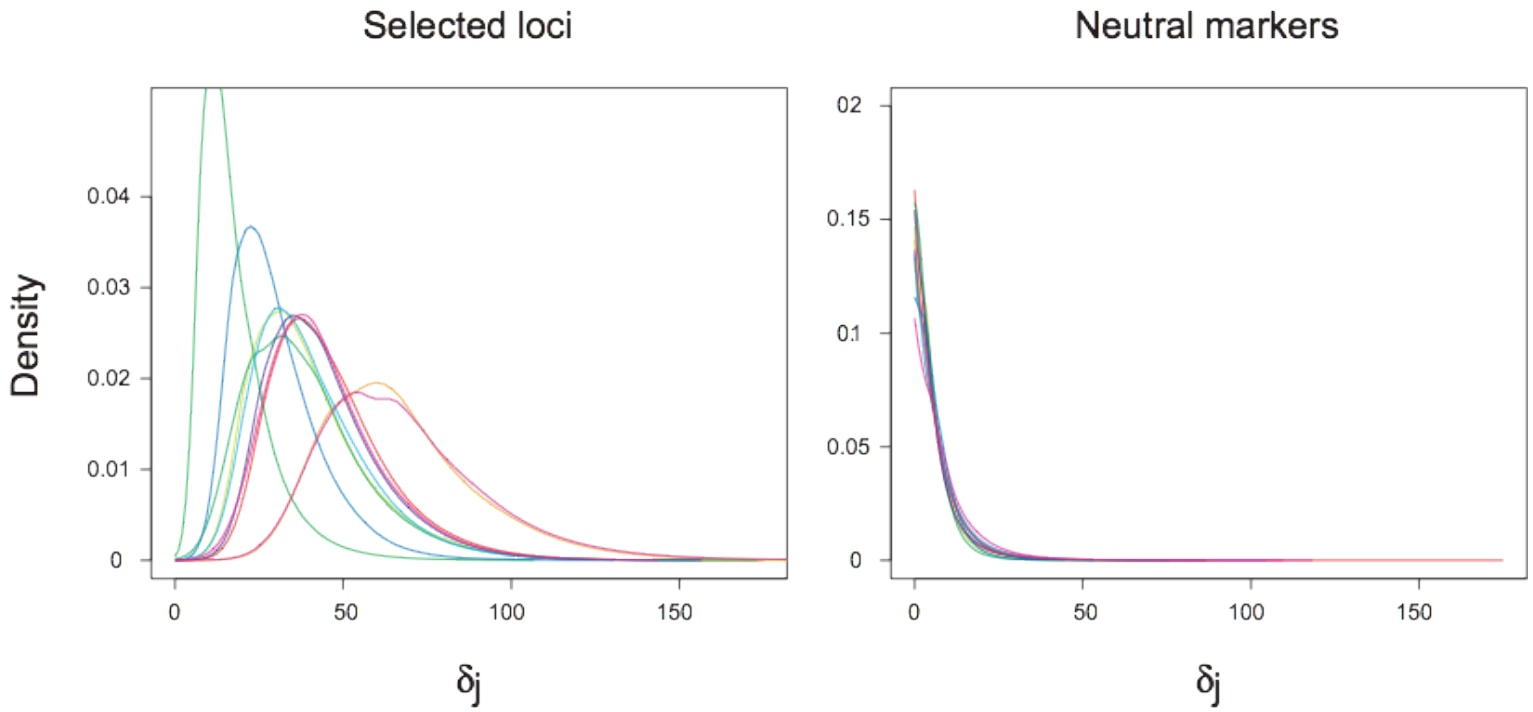


BayeScan (Foll and Gaggiotti 2008)

BayesF_{ST} (Beaumont and Balding 2004)

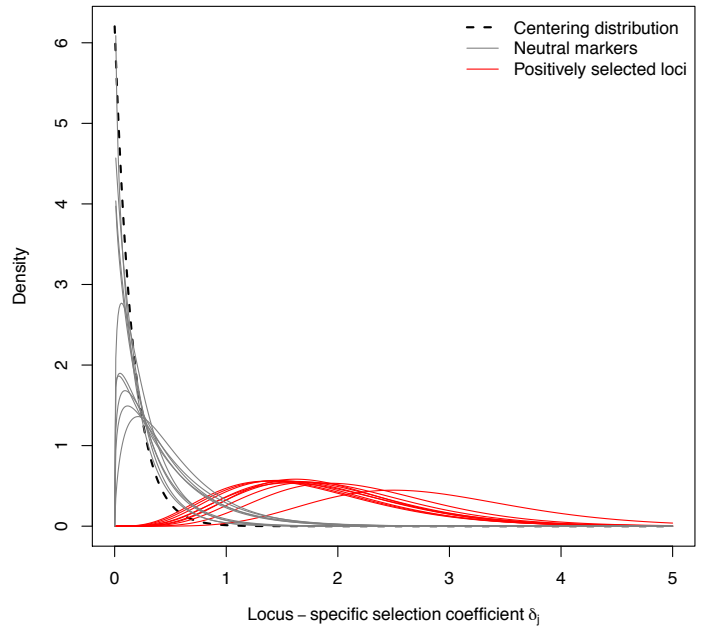
A criterion to detect selected markers?

Typically, we get the following posterior distributions for the locus-specific selection parameters:



How can we define a criterion to discriminate selected loci from presumably neutral markers?

Kullback-Leibler divergence (KLD)

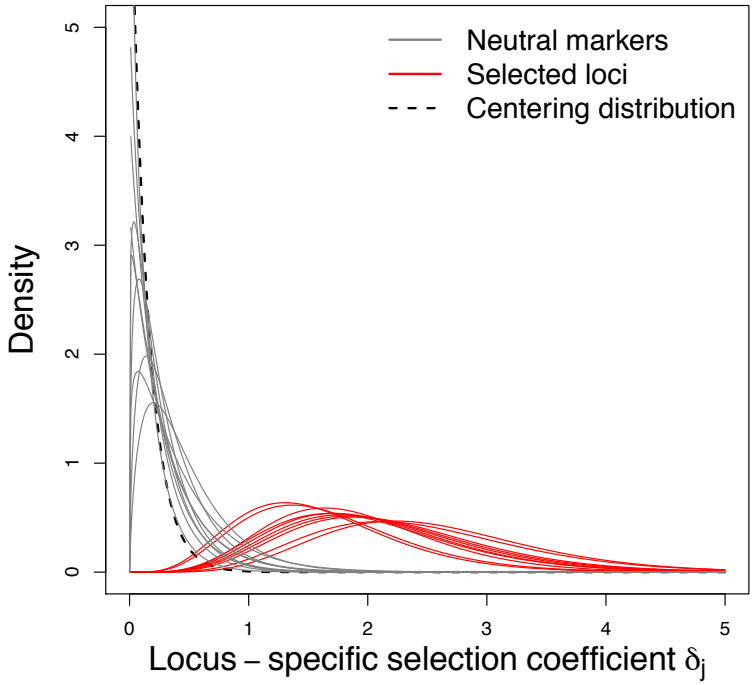


It is not sufficient to test whether the posterior distribution of δ_j at one locus departs from zero (this would neglect the genome-wide effects of selection or the departure from island model).

Since we assume that the locus-specific coefficients of selection are drawn from a common hyper-distribution with parameter λ, it is more appropriate to compare the posterior distribution of δ_j and its "centering distribution" (hyper-prior distribution of δ_j, with parameter λ equal to its posterior mean).

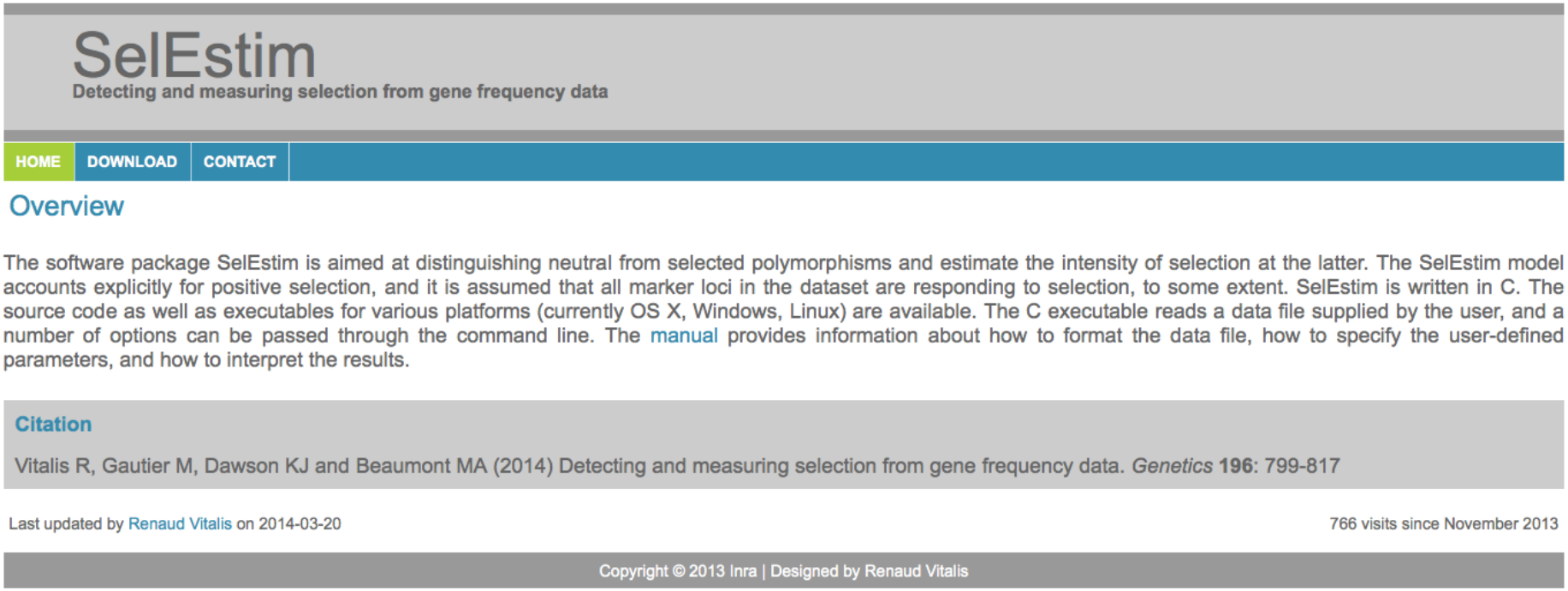
We use the KLD as a distance between these two distributions

Calibration of the KLD



We calibrate the KLD by generating pseudo observed data (pod), drawn from the posterior distribution of the model parameters. The pod is analysed, and the quantiles of the KLD distribution from the pod are then used as threshold values for the data of interest…

A software package



SelEstim

Detecting and measuring selection from gene frequency data

HOME | DOWNLOAD | CONTACT

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

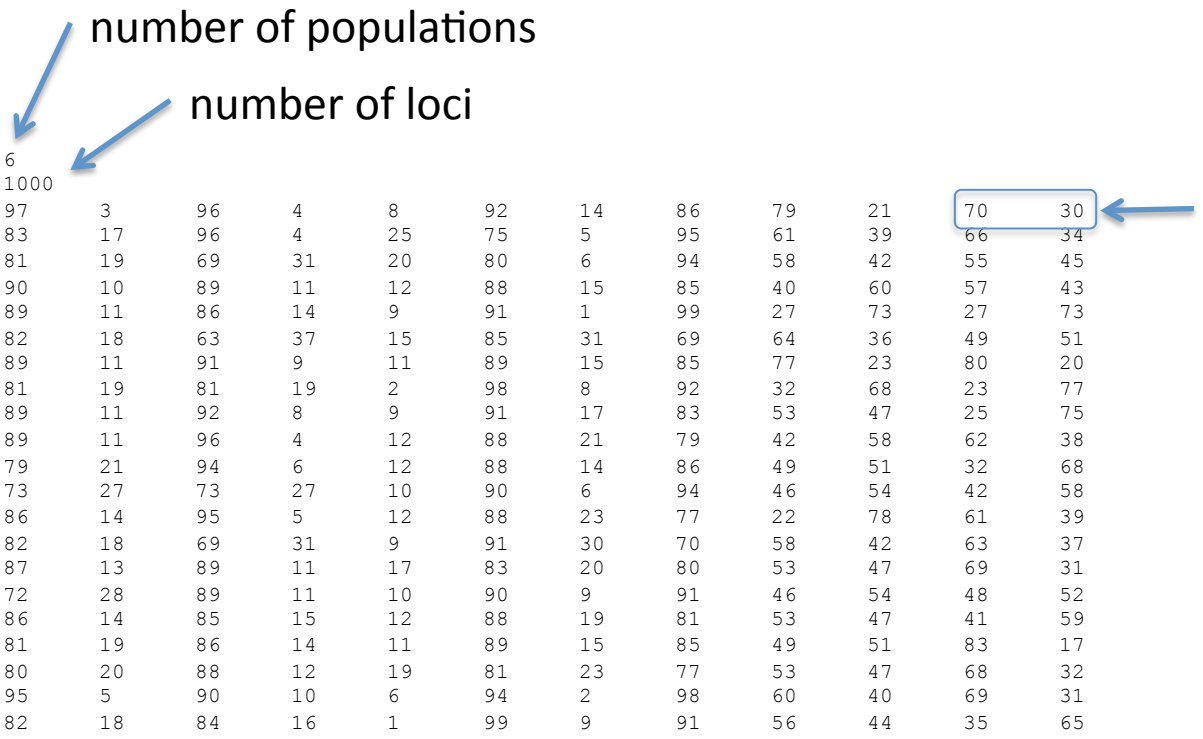
Last updated by Renaud Vitalis on 2014-03-20

766 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

Using SELESTIM: the data



Using SELESTIM

Using the command line:

```
./selestim -h
usage: ./selestim [ options ]
valid options are :
-help                print this message
-version             print version
-file                name of the input file (default: data.dat)
-outputs             directory where the outputs will be produced (default: current directory)
-seed                initial seed for the random number generator (default: computed from current time)
-threads             number of threads to be used (default: number of cpu available)
-length              run length of the Markov chain (default: 125000)
-thin                thinning interval size (default: 25)
-burnin              length of the burn-in period (default: 25000)
-npilot              number of pilot runs (default: 25)
-lpilot              length of each pilot run (default: 1000)
-pool                option to analyse data from pooled DNA samples (default: unset)
-fixed_beta          option to fix the shape parameters of the beta prior distribution of pi (default: unset)
-beta_a              shape parameter of the beta prior distribution of pi (default: 0.70)
-beta_b              shape parameter of the beta prior distribution of pi (default: 0.70)
-fixed_lambda        option to fix the value of lambda (default: unset)
-lambda_prior        prior distribution of lambda, which can only be inverse gamma ('invgam', by default) or an exponential ('exp')
-invgam_shape        shape parameter of the inverse gamma prior distribution of lambda (default: 3.00)
-invgam_rate         rate parameter of the inverse gamma prior distribution of lambda (default: 2.00)
-captl_lambda        rate parameter of the exponential prior distribution of lambda (default: 1.00)
-min_M               lower bound for the log-uniform prior on M (default: 0.001)
-max_M               upper bound for the log-uniform prior on M (default: 10000)
-max_sig             upper bound for the exponential prior on sigma (default: 700)
-dlt_cnt             half window width from which updates of allele counts are randomly drawn (default: 5)
-dlt_p               half window width from which updates of p are randomly drawn (default: 0.25)
-dlt_M               standard deviation of the lognormal distribution from which updates of M are drawn (default: 0.10)
-dlt_pi              half window width from which updates of pi are randomly drawn (default: 0.25)
-dlt_sig             standard deviation of the lognormal distribution from which updates of sigma are drawn (default: 2.50)
-dlt_del             standard deviation of the lognormal distribution from which updates of delta are drawn (default: 0.80)
-dlt_lam             standard deviation of the lognormal distribution from which updates of lambda are drawn (default: 0.05)
-dlt_beta_mu         half window width from which updates of the beta mu parameters are drawn (default: 0.03)
-dlt_beta_nu         standard deviation of the lognormal distribution from which updates of the beta nu parameters are drawn (default: 1.00)
-calibration         option to generate pseudo-observed data and calibrate the Kullback-Leibler divergence
-calibration_only    option to generate pseudo-observed data and calibrate the Kullback-Leibler divergence from previous analyses
-pod_nbr_loci                option to specify the number of loci to be simulated for calibration (if different from the dataset)
-verbose             option to print the traces of all parameters (generates big output files!)
```

Pilot runs are used to adjust the parameters of the proposal functions, in order to get acceptance rates between 0.25 and 0.40. The burnin corresponds to the preliminary part of the chain before it has reached stationarity.

```
./selestim -file data/1000_markers_sample_size_60.dat -outputs test-1/ -threads 8 -seed 12 -thin 20 -npilot 10 -lpilot 500 -burnin 1000 -length 20000
--------------------------------------------------------------------------------
Fri Sep  4 15:47:22 2015
--------------------------------------------------------------------------------

./selestim -file data/1000_markers_sample_size_60.dat -outputs test-1/ -threads 8 -seed 12 -thin 20 -npilot 10 -lpilot 500 -burnin 1000 -length 20000

This analysis was performed using selestim (version 1.1.4)

Checking file `data/1000_markers_sample_size_60.dat'... OK
The data consist in 1000 SNPs and 3 sampled populations

--------------------------------------------------------------------------------
Mean sample size (min, max) per sampled population:
--------------------------------------------------------------------------------
Population no.  1: 60.00 (60,60)
Population no.  2: 60.00 (60,60)
Population no.  3: 60.00 (60,60)
--------------------------------------------------------------------------------
Overall          : 60.00 (60,60)
--------------------------------------------------------------------------------

--------------------------------------------------------------------------------
Overall genetic differentiation (F_ST)                          = 0.1488
--------------------------------------------------------------------------------

--------------------------------------------------------------------------------
Prior distribution of lambda is inverse gamma (lambda_prior)
    with shape parameter (invgam_shape)                         = 3.000000
    and rate parameter (invgam_rate)                            = 2.000000

Number of threads used (threads)                                = 8
Random number generator's seed (seed)                           = 12

Length of the burn-in period (burnin)                           = 1000
Run length of the Markov chain (length)                         = 20000
Thinning interval (thin)                                        = 20
Number of MCMC samples (length / thin)                          = 1000
Number of pilot studies (npilot)                                = 10
Length of each pilot study (lpilot)                             = 500

Lower bound of the interval for M (min_M)                       = 0.001000
Upper bound of the interval for M (max_M)                       = 10000.00
Upper bound of the interval for sigma (max_sig)                 = 700.00
Initial half window width for updates of allele counts (dlt_cnt)      = 5
Initial half window width for updates of p (dlt_p)              = 0.250000
Initial SD of the lognormal for updates of M (dlt_M)            = 0.100000
Initial half window width for updates of pi (dlt_pi)            = 0.250000
Initial SD of the lognormal for updates of sigma (dlt_sig)      = 2.500000
Initial SD of the lognormal for updates of delta (dlt_del)      = 0.800000
Initial half window width for updates of mu (dlt_beta_mu)       = 0.025000
Initial SD of the lognormal for updates of nu (dlt_beta_nu)     = 1.000000
--------------------------------------------------------------------------------
```

```
-------------------------------------------------------------------------------
Pilot run # 1:
-------------------------------------------------------------------------------

        Allele frequencies p_ij's
                average value = 0.8316 [0.0008,1.0000]
                average updating parameter = 0.2361 [0.2000,0.3125]
                average acceptance rate = 0.2703 [0.0060,0.4620]
                1036 parameters have been scaled, out of 3000

        Population parameters M_i's
                average value = 11.2092 [6.0662,14.5225]
                average updating parameter = 0.1083 [0.1000,0.1250]
                average acceptance rate = 0.4013 [0.3940,0.4120]
                1 parameters have been scaled, out of 3

        Shape parameter (a) of the prior distribution of migrant allele frequencies pi_j's
                current value = 4.1726
                updating parameter = 0.0250
                average acceptance rate = 0.2880
                0 parameters have been scaled, out of 1

        Shape parameter (b) of the prior distribution of migrant allele frequencies pi_j's
                current value = 0.9407
                updating parameter = 0.8000
                average acceptance rate = 0.0600
                1 parameters have been scaled, out of 1

        Migrant allele frequencies pi_j's
                average value = 0.8345 [0.0610,0.9923]
                average updating parameter = 0.2749 [0.2000,0.3125]
                average acceptance rate = 0.3800 [0.1340,0.5260]
                505 parameters have been scaled, out of 1000

        Genome-wide coefficient of selection lambda
                current value = 2.5235

        Locus-specific selection coefficient delta_j's
                average value = 2.5577 [0.0004,19.2784]
                average updating parameter = 1.0000 [1.0000,1.0000]
                average acceptance rate = 0.5886 [0.4140,0.6800]
                1000 parameters have been scaled, out of 1000

        Locus- population-specific selection coefficient sigma_ij's
                average value = 2.6714 [0.0001,73.8689]
                average updating parameter = 3.1009 [2.0000,3.1250]
                average acceptance rate = 0.4542 [0.1800,0.5440]
                2897 parameters have been scaled, out of 3000

-------------------------------------------------------------------------------
Pilot run # 2:
-------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------

[…]
```

```
[…]
-------------------------------------------------------------------------------
Fri Sep  4 15:49:16 2015
------------------------
        Computing time elapsed since beginning = 14 secs.
        Estimated time until the MCMC stops    = 55 secs.
-------------------------------------------------------------------------------

-------------------------------------------------------------------------------
Running the MCMC
-------------------------------------------------------------------------------

 starting [.........]
 10% done [.........]
 20% done [.........]
 30% done [.........]
 40% done [.........]
 50% done [.........]
 60% done [.........]
 70% done [.........]
 80% done [.........]
 90% done [.........]
100% done !

-------------------------------------------------------------------------------
Computation of the effective sample size (ESS)

        log posterior density                          = 15.883337
        parameters M                                   = (24.352808,48.911723,16.465046)
        shape parameter (alpha) of the parameter pi    = 345.362524
        shape parameter (beta) of the parameter pi     = 212.063811
        (hyper-)parameter lambda                       = 14.920371

ESS is a measure of how well a Markov chain is mixing. ESS represents the number
of effectively independent draws from the posterior distribution that the Markov
chain is equivalent to [ESS must be compared to the chain length = 1000].

Warning! Low ESS (due to strong autocorrelation) indicates poor mixing of the
Markov chain. The ESS of the (hyper-)parameter lambda is typically lower than that
of the other parameters. You are strongly recommended to inspect the trace of the
lambda parameter in the 'trace_lambda.out' file. The trace shall show relatively
good mixing (low autocorrelation, AND no decreasing trend). Otherwise, you may want
to increase the length of the burn-in period and/or the total length of the Markov
chain.
-------------------------------------------------------------------------------

-------------------------------------------------------------------------------
Fri Sep  4 15:50:22 2015
------------------------
        Total computing time elapsed           = 1 min. 19 secs.
-------------------------------------------------------------------------------

The program has successfully terminated.
```

Example of outputs

summary_delta.out

locus	mean	std	KLD
1	5.624914	6.107495	0.010533
2	4.622427	4.766559	0.028151
3	4.965960	5.145243	0.015294
4	5.885860	5.532389	0.004609
5	4.716900	4.650066	0.022928
6	4.718736	4.909176	0.024757
7	4.672732	4.592066	0.024950
8	17.818480	8.843916	1.289237
9	5.179656	5.404096	0.010236
10	4.174291	4.553678	0.063355
11	4.812349	5.091915	0.023289
12	5.339638	5.501392	0.005762
13	4.954376	4.826432	0.014767
14	7.627897	6.562957	0.062287
15	5.381138	5.507947	0.004600
16	4.824512	5.326538	0.032416
17	4.852205	4.605712	0.020690
18	4.502991	4.941861	0.045023
19	4.717220	4.983770	0.026848
20	4.888352	4.885546	0.016129
21	5.570098	5.592022	0.001489
22	4.154663	4.208307	0.054263
23	5.093745	5.251101	0.011155
24	4.010893	4.266515	0.070106
25	5.370633	5.692714	0.008719

[…]

Using R scripts to analyse the outputs (e.g., the CODA package to test for convergence) and plot graphs (some ad-hoc functions in SelEstim.R)

KLD calibration

```
./selestim -file data/1000_markers_sample_size_60.dat -outputs test-1/ -threads 8 -seed 12 -thin 20 -npilot 10 -lpilot 500 -burnin 1000 -length 20000 -calibration_only
--------------------------------------------------------------------------------
Fri Sep  4 16:32:14 2015
--------------------------------------------------------------------------------

./selestim -file data/1000_markers_sample_size_60.dat -outputs test-1/ -threads 8 -seed 12 -thin 20 -npilot 10 -lpilot 500 -burnin 1000 -length 20000 -calibration_only

This analysis was performed using selestim (version 1.1.4)

--------------------------------------------------------------------------------
Calibration of the Kullback-Leibler divergence using pseudo-observed data
--------------------------------------------------------------------------------

Generating file `test-1/calibration/pod_1000_markers_sample_size_60.dat'...

 starting [..........]
 10% done [..........]
 20% done [..........]
 30% done [..........]
 40% done [..........]
 50% done [..........]
 60% done [..........]
 70% done [..........]
 80% done [..........]
 90% done [..........]
100% done !

[…]
```

KLD calibration

```
[…]

The pseudo-observed data consist in 1000 SNPs and 3 sampled populations

------------------------------------------------------------------------------------
Overall genetic differentiation (F_ST)                              = 0.1475
------------------------------------------------------------------------------------

------------------------------------------------------------------------------------
Prior distribution of lambda is inverse gamma (lambda_prior)
    with shape parameter (invgam_shape)                             = 3.000000
    and rate parameter (invgam_rate)                                = 2.000000

Number of threads used (threads)                                    = 8
Random number generator's seed (seed)                               = 12

Length of the burn-in period (burnin)                               = 1000
Run length of the Markov chain (length)                             = 20000
Thinning interval (thin)                                            = 20
Number of MCMC samples (length / thin)                              = 1000
Number of pilot studies (npilot)                                    = 10
Length of each pilot study (lpilot)                                 = 500

Lower bound of the interval for M (min_M)                           = 0.001000
Upper bound of the interval for M (max_M)                           = 10000.00
Upper bound of the interval for sigma (max_sig)                     = 700.00
Initial half window width for updates of allele counts (dlt_cnt)    = 5
Initial half window width for updates of p (dlt_p)                  = 0.250000
Initial SD of the lognormal for updates of M (dlt_M)                = 0.100000
Initial half window width for updates of pi (dlt_pi)                = 0.250000
Initial SD of the lognormal for updates of sigma (dlt_sig)          = 2.500000
Initial SD of the lognormal for updates of delta (dlt_del)          = 0.800000
Initial half window width for updates of mu (dlt_beta_mu)           = 0.025000
Initial SD of the lognormal for updates of nu (dlt_beta_nu)         = 1.000000

Calibration of the Kullback-Leibler divergence (calibration_only)
------------------------------------------------------------------------------------

[…]
```

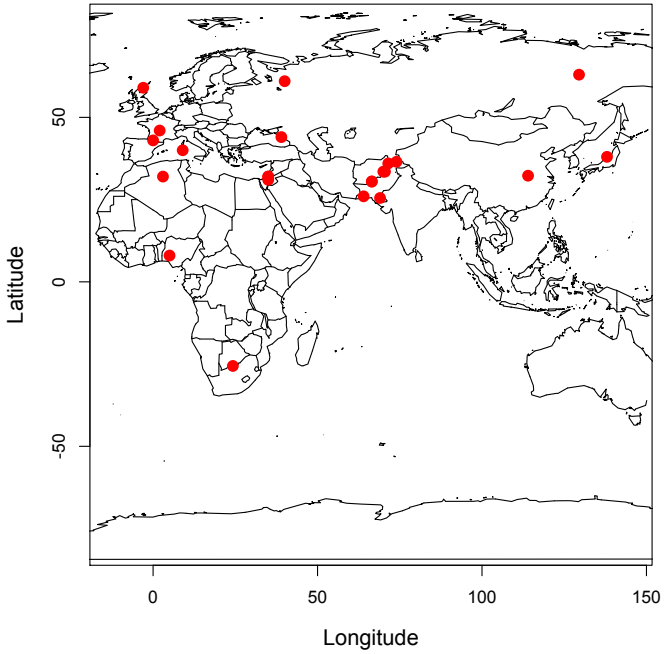
KLD calibration

calibration/KLD_quantiles.out

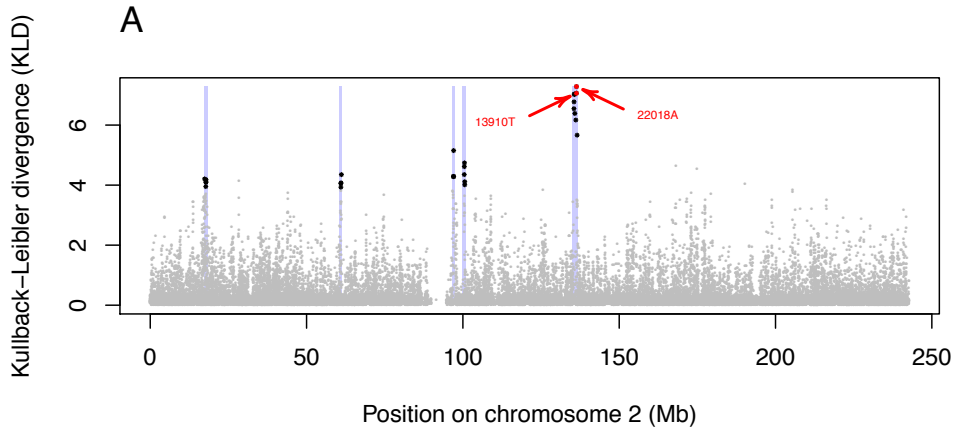
quantile	KLD
50.00%	0.025470
90.00%	0.169487
95.00%	0.530132
98.00%	1.041611
99.00%	1.345791
99.50%	1.602976
99.90%	1.885545
99.95%	1.985136
99.99%	2.064809

Application on human data (CEPH)

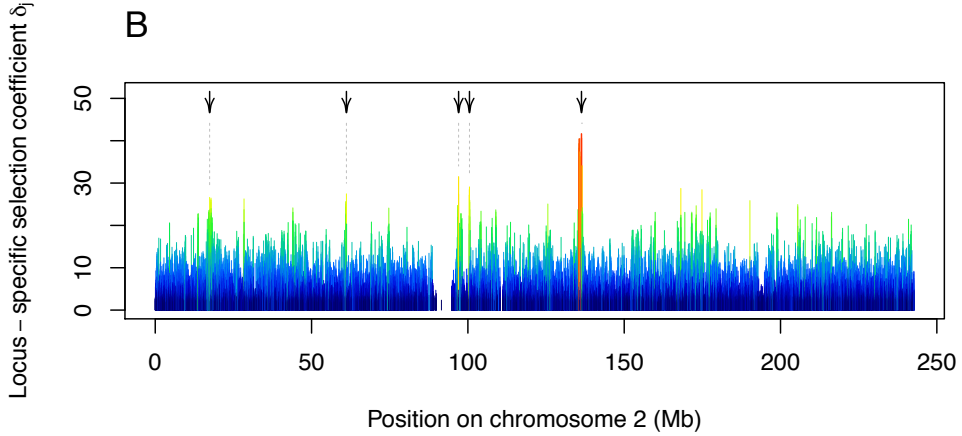
We have applied the method on a subset of the Stanford HGDP-CEPH SNP genotyping data from chromosome 2 (52,631 SNPs)



Application on human data (CEPH)

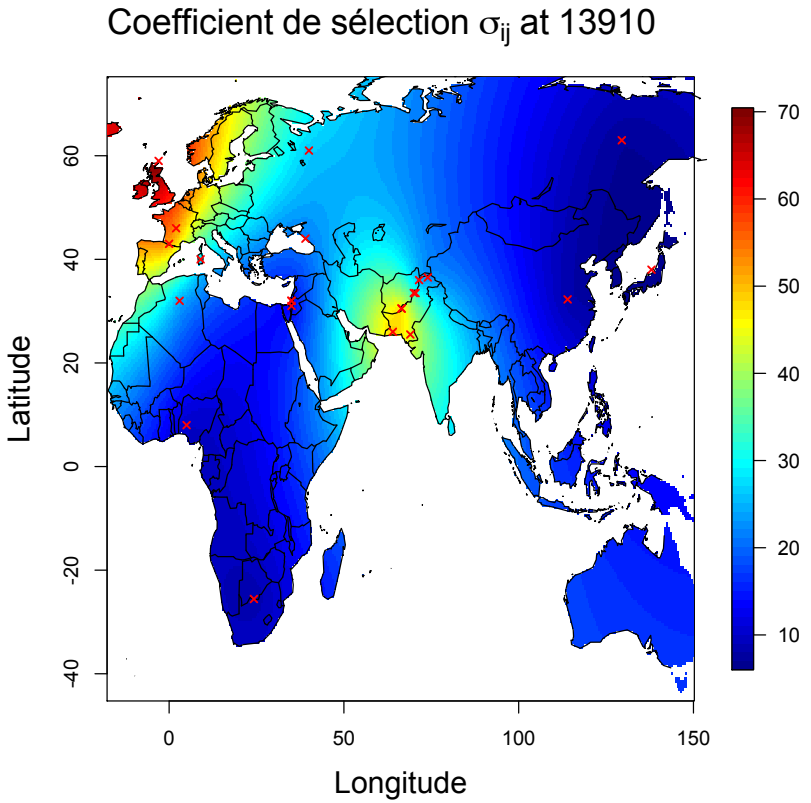


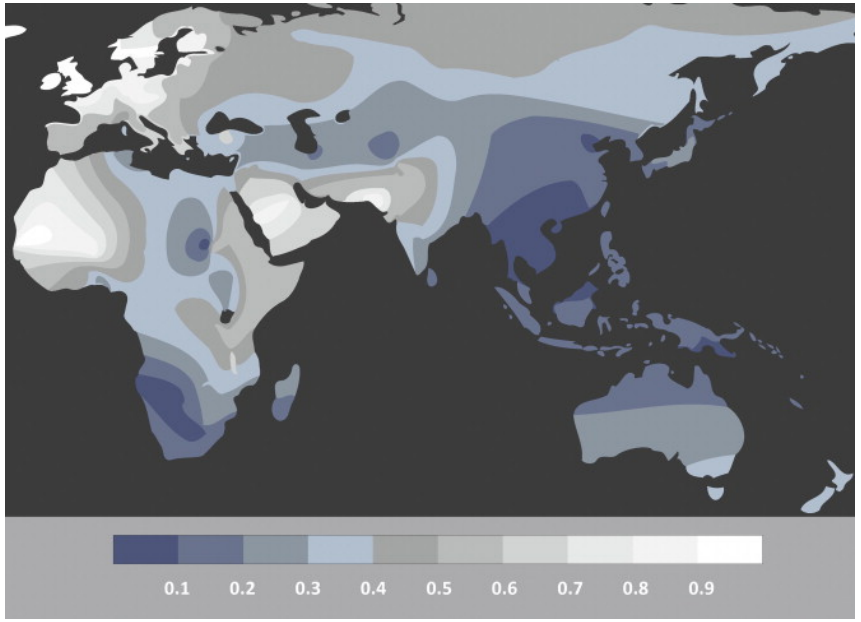
1Mb windows that contain at least 3 SNPs with KLD > 3.912



Strong signature of selection in the vicinity of the lactase gene *LCT*, in particular at 2 SNPs reported to be very tightly associated with lactase persistence (13919T and 22018A; see Bersaglieri *et al*. 2004).

Application on human data (CEPH)





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

The selection coefficient at 13910T (left) is stronger in milk-drinking populations. It correlates with lactase persistence in Europe and the Indus valley, not in Africa or the Near and Middle East: convergent evolution.

Take home messages

Bayesian methods: check for convergence and mixing properties! (see the R package CODA)

This family of approaches does not take LD into account yet

Be aware of the underlying population models and assumptions (equilibrium island model, etc.)