Whole genome scan for adaptive divergence and association analysis with population specific covariates using BayPass

Mathieu Gautier

UMR INRA/CIRAD/IRD/SupAgro CBGP

26^{th} May 2016

Genome Scan for Adaptive Differentiation

Forces driving Allele frequencies evolution

- Mutation (and recombination when considering haplotypes) : generate variability
- Drift : introduces stochasticity (Finite Population Size)
- Migration (in terms of gene flow)
- Selection

Different Influences of the evolutionary forces (*Cavalli-Sforza, 1966*)

- Demographic Factors (genetic drift, gene flow) expected to be common to all loci
  $\Longrightarrow$ Global Effect
- Selection (mutation and recombination) expected to vary across loci
  $\Longrightarrow$ Local Effect

Identifying Footprints of Selection $\leftrightarrow$ disentangle locus-specific from demographic effects on allele frequency differences

Various Approaches in the population genomics era (1)

Likelihood–free approaches

- Evaluate differentiation for each marker

  (i.e. compute an F_{ST}–like summary statistic)

- Identify overly (or poorly) differentiated loci (outliers) relative to a distribution expected under neutrality
- Defining the neutral distribution (of the differentiation summary stats)
  - Theoretically (Lewontin et Krakauer, 1973 ; Bonhomme et al., 2010 ; Fariello et al., 2013)

    $\Longrightarrow$ model assumptions
  - Via Simulation (Bowcock et al., 1991)

    $\Longrightarrow$ demographic assumptions
  - Empirically (Akey et al., 2002 ; Flori et al., 2009)

    $\Longrightarrow$ control of FDR/FNR

Various Approaches in the population genomics era (2)

(Bayesian) model–based approaches

- Develop a (statistical) model to disentangle locus-specific from demographic effect on the variance of population allele frequencies

  $\Longrightarrow$ Approaches generally conceived as test of departure from a neutral model

- Prior assumption on the population allele frequency distribution
  - Beta distribution $\Longleftarrow$ $\mathcal{F}$–model (migration drift equilibrium)

    Beaumont & Balding (2004) ; Riebler et al. (2008) ; Foll & Gaggiotti (2008)...
  - Beta-Hypergeometric Distribution $\Longleftarrow$ $\mathcal{F}$–model with selection (migration–drift–selection equilibrium)

    Vitalis et al., 2014
  - Gaussian distribution $\Longleftarrow$ $\sim$ pure drift model (cf. Nicholson et al., 2002)

    Coop et al., 2010 ; Gautier et al., 2010 ; Gautier 2015

Multivariate Gaussian distribution assumption for population allele frequencies

- Introduced by Coop et al. (2010) as a generalization of the univariate Gaussian model by Nicholson et al. (2002)
- Let $\alpha_{ij}^{\star}$ the (unobserved) "instrumental" freq. of the ref. allele at SNP i in pop j defined over the real line support and related to α_{ij} by :
  - $\alpha_{ij} = \alpha_{ij}^{\star}$ if $\alpha_{ij}^{\star} \in (0, 1)$
  - $\alpha_{ij} = 0$ if $\alpha_{ij}^{\star} < 0$ (allele absent or "lost")
  - $\alpha_{ij} = 1$ if $\alpha_{ij}^{\star} > 1$ (allele "fixed")
- Prior distribution for pop allele freq. vectors : $\boldsymbol{\alpha_i^{\star}} = \left\{\alpha_{ij}^{\star}\right\}_{(1..J)}$

$$\boldsymbol{\alpha_i^{\star}} \sim \mathrm{N}_J\left(\pi_i \mathbb{1}; \pi_i(1-\pi_i)\boldsymbol{\Omega}\right)$$

  - $\mathbb{1}$: identity vector of length J (number of pops.)
  - π_i : across pop. frequency (might be interpreted as the "ancestral" ref. allele frequency)
  - $\boldsymbol{\Omega}$: scaled covariance ($J \times J$) matrix of pop. allele frequency

Ω captures the covariance structure of allele frequencies that originates from the population shared history (global effect of the demography)

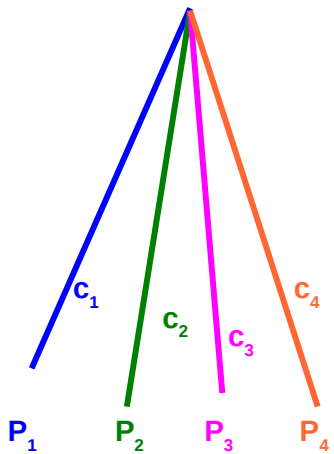
Demographic interpretation of Ω

Star-shaped
(e.g., Nicholson et al., 2002)

Bifurcating tree

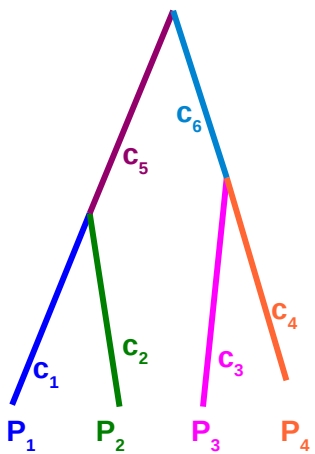
Admixture graph

*e.g., Patterson et al. (2012) (**ADMIXTOOLS**) ; Pickrell and Pritchard (2012) (**TreeMix**)*



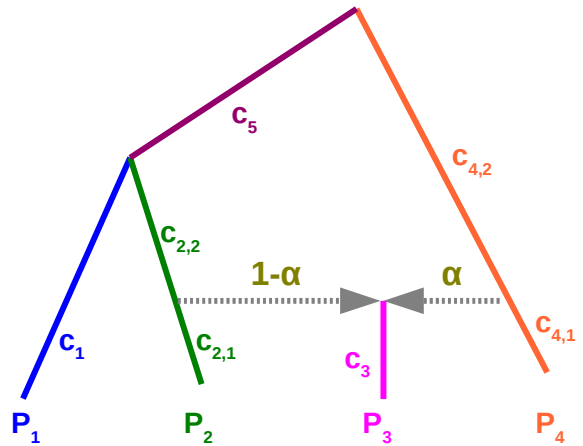
P_1	c_1	0	0	0
P_2	0	c_2	0	0
P_3	0	0	c_3	0
P_4	0	0	0	c_4

Ω



P_1	c_1+c_5	c_5	0	0
P_2	c_5	c_2+c_5	0	0
P_3	0	0	c_3+c_6	c_6
P_4	0	0	c_6	c_4+c_6

Ω

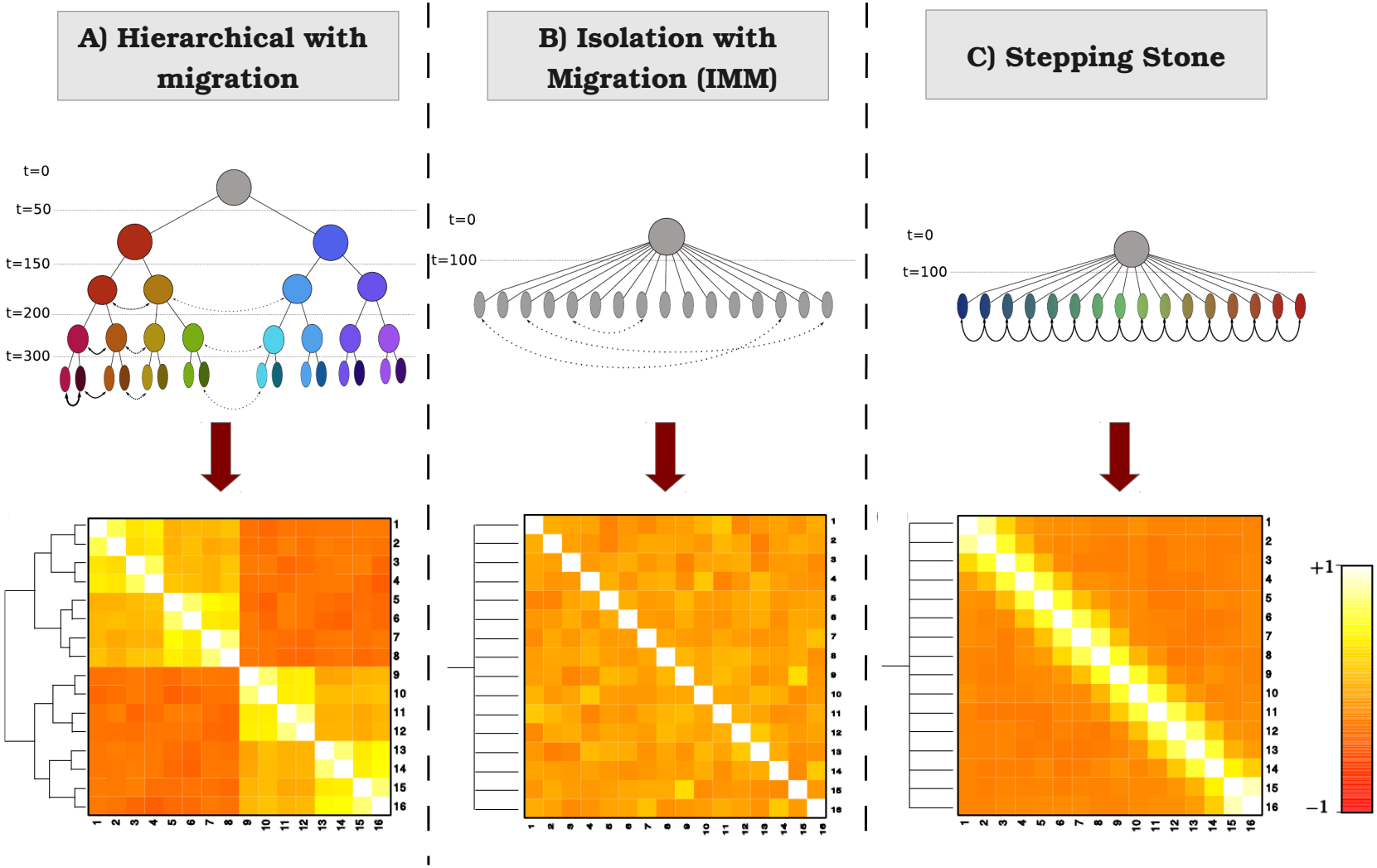


P_1	c_1+c_5	c_5	$(1-\alpha)^2c_5$	0
P_2	c_5	$c_{2,1}+c_{2,2}+c_5$	$(1-\alpha)^2(c_{2,2}+c_5)$	0
P_3	$(1-\alpha)^2c_5$	$(1-\alpha)^2(c_{2,2}+c_5)$	$c_3+\alpha^2c_{4,2}+$ $(1-\alpha)^2(c_{2,2}+c_5)$	$\alpha^2c_{4,2}$
P_4	0	0	$\alpha^2c_{4,2}$	$c_{4,1}+c_{4,2}$

Ω

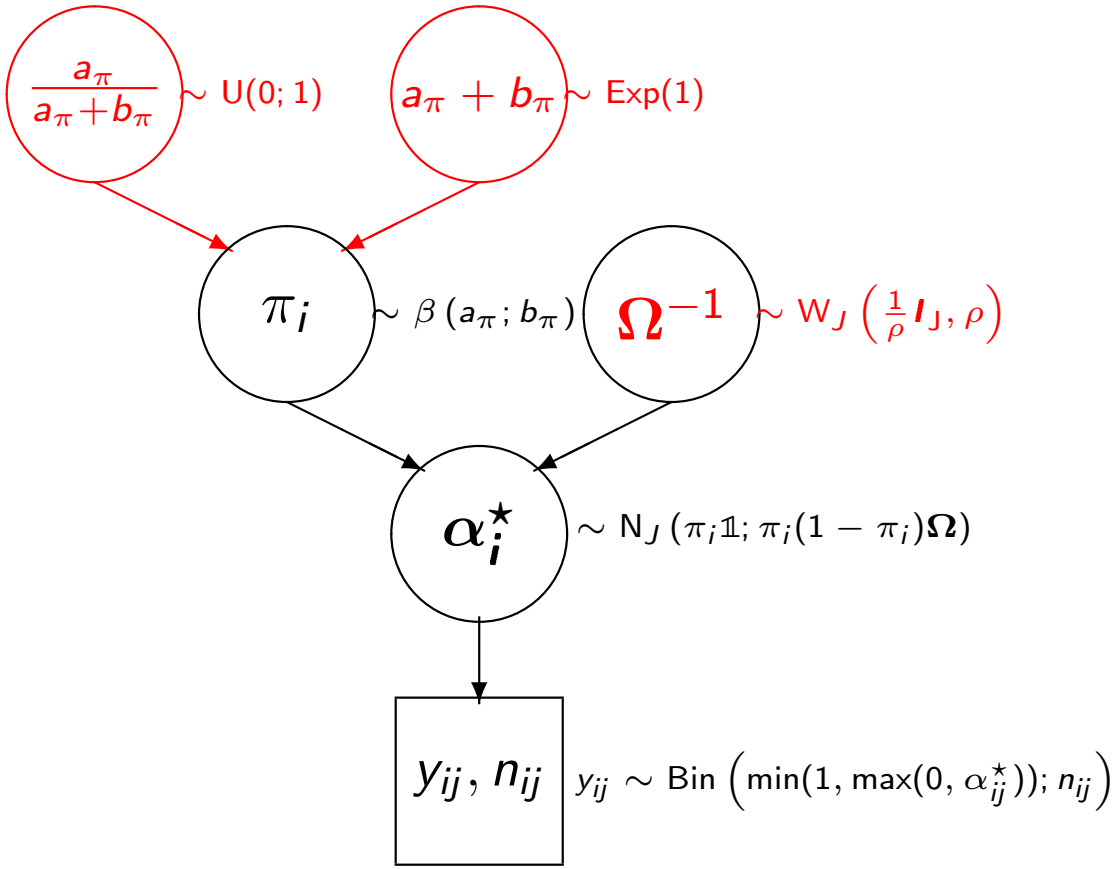
$$*c_j \simeq 1-\left(1-\frac{1}{2N_j}\right)^{t_j} \simeq \frac{t_j}{2N_j}$$

Realized Ω in more complex models De Villemereuil et al., 2014



NB: Heatmap of the correlation matrix $\mathbf{\Gamma}=\{\rho_{ij}\}$ related to $\mathbf{\Omega}=\{\omega_{ij}\}$ by $\rho_{ij}=\omega_{ij}/(\omega_{ii}\omega_{jj})^{1/2}$

The core BayPass Bayesian hierarchical model



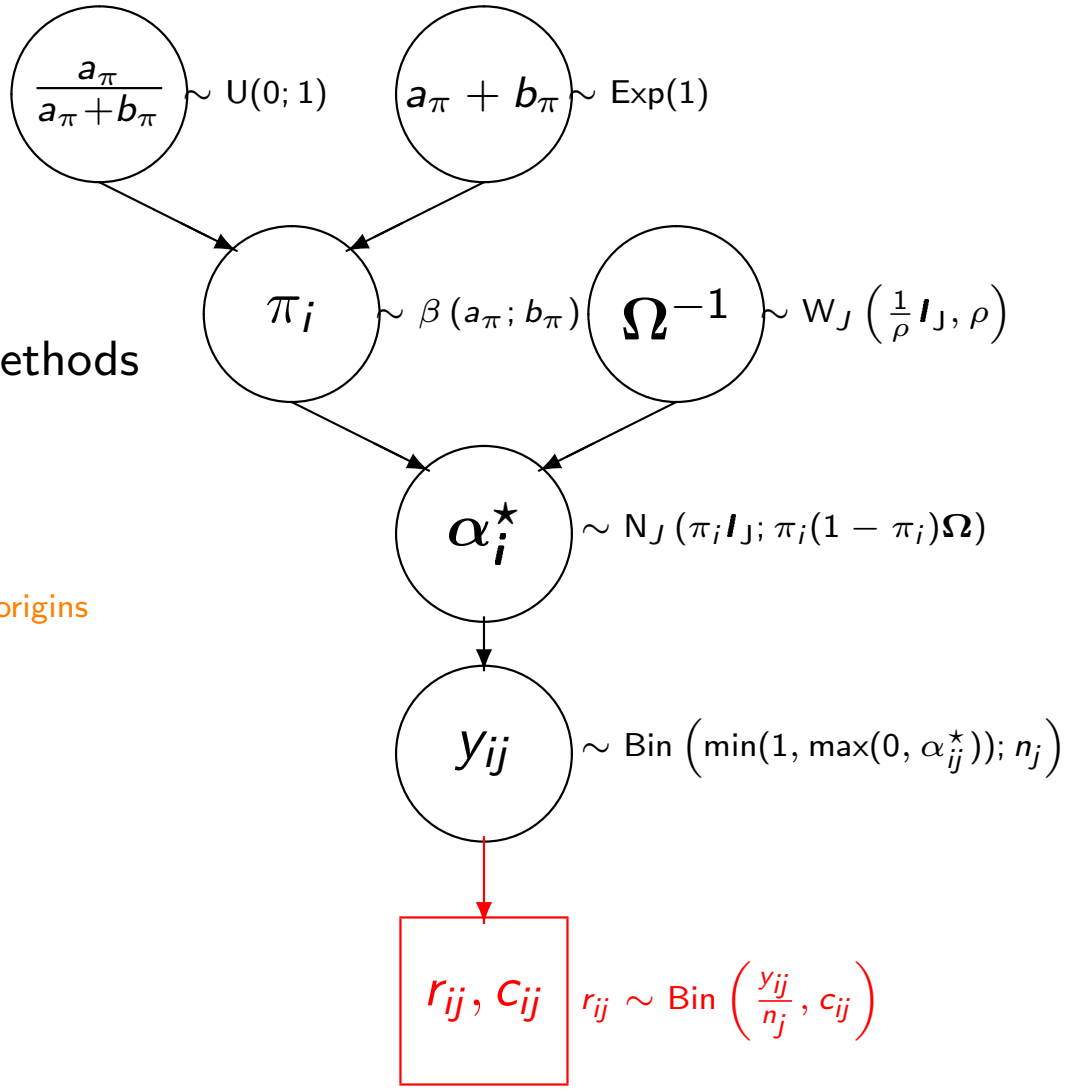
- Similar to the core Bayenv model (Coop et al., 2010) with additional extensions
  - Priors on a_π and b_π (instead of setting $a_\pi = b_\pi = 1$)
  - Less informative (e.g., singular) Wishart prior on Ω^{-1} (e.g., setting $\rho = 1$ instead of $\rho = J$)

Estimating of Ω under a Bayesian hierarchical model

Overcome (potentially serious) limitations of methods based the observed covariance matrix

(e.g. TreeMix, PCA–related approaches...) :

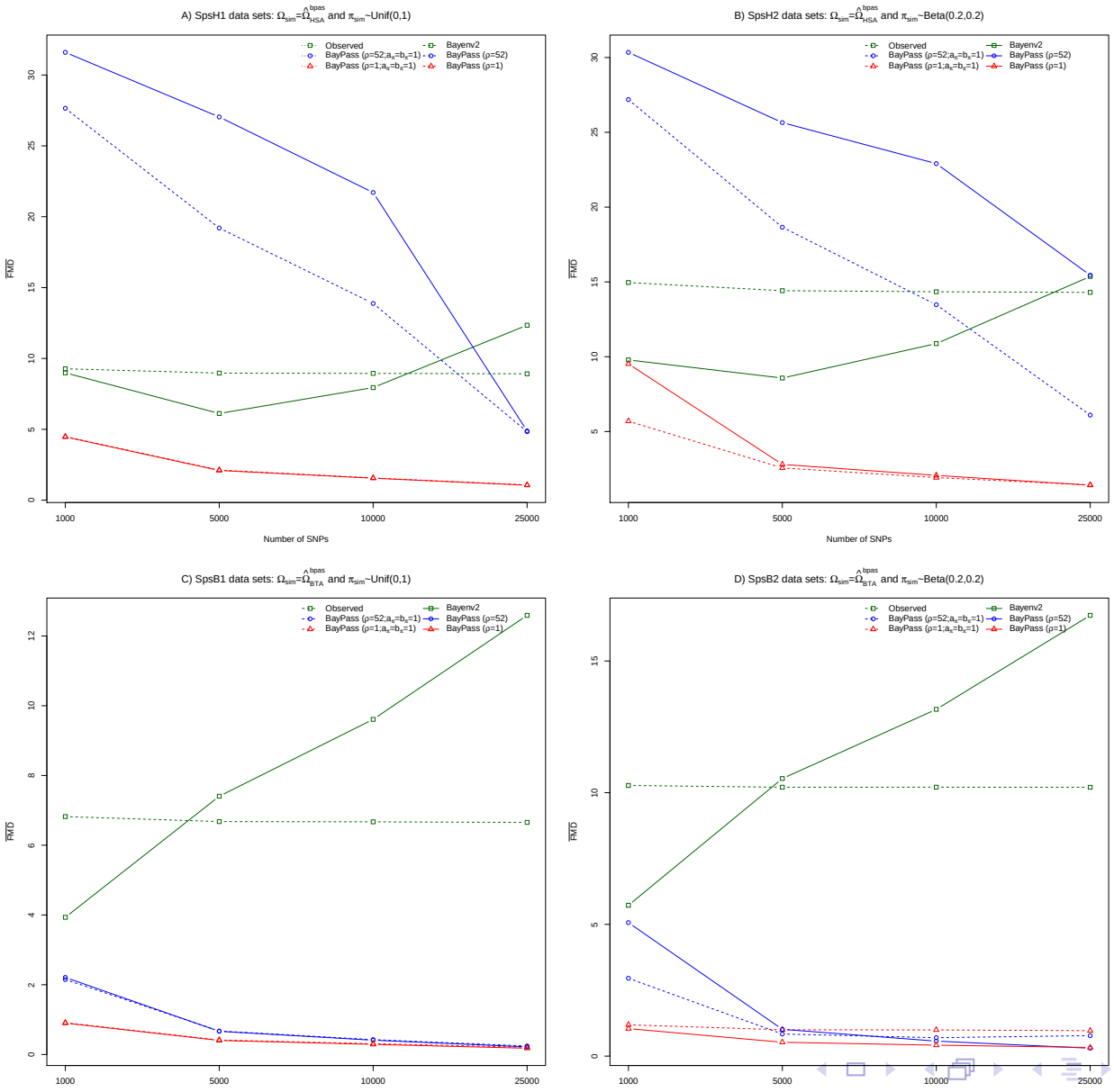
- Joint estimation of the (unobserved) π_i's and $\mathbf{\Omega}$
- Robust to sample representativeness of population origins *(e.g., unbalanced trees)*
- Robust to variation in population sample sizes *(incl. missing data)*
- Simple to deal with read count data (Pool–Seq) *(by integrating over uncertainty in allele count)*



Parameter estimation : a standard M–H within Gibbs MCMC algorithm

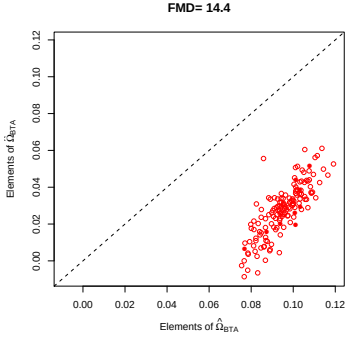
- Initialize all the parameter values (e.g., methods of moments estimators)
- Sample one parameter at a time (from their full conditional distribution) :
  - *If read count data* : the (I SNPs $\times$ J pops) y_{ij}'s *(M–H updates with uniform proposals)*
  - the (I SNPs $\times$ J pops) $\alpha_{ij}^{\star}$'s *(M–H updates with Gaussian proposals)*
  - the matrix $\mathbf{\Omega}$ (actually $\mathbf{\Omega}^{-1}$) *(Gibbs update)*
  - the (I SNPs) π_i's *(M–H updates with uniform proposals)*
  - a_π and b_π (actually $\frac{a_\pi}{a_\pi+b_\pi}$ and $a_\pi + b_\pi$) *(M–H updates with uniform proposals)*
- A typical run consists in :
  - Several Pilot runs to adjust parameters of the proposals (e.g. targeted accept. rates between 0.25 and 0.4) : *e.g.* $20 \times 1,000$ *iterations*
  - A Burn-in period (to achieve stationary distributions) : *e.g. 5,000 iterations*
  - Parameters Sampling with thinning (to reduce auto-correlations) : *e.g.* $50 \times 1,000$ *iterations*
- The BayPass implementation
  - Coded in Fortran90 language with a flexible parametrization : *(sampler checked via analyses of simulated data under the inference models + comparisons with an independent BUGS implementation)*
  - Reasonable computational times : $2.5h$ *to analyze a data sets of* $J = 18$ *and* $I = 40,000$ *SNPs on a single processor (ifort compiled binary)* ($1h05$ *with 4CPUs : option* `-nthreads 4`)
  - Correct for some implementation issues that lead to inaccurate results in the Bayenv2 code

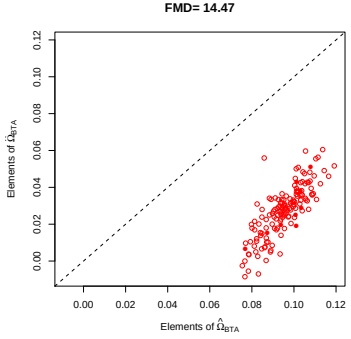
Precision in the estimation of Ω : model/implementation comparisons

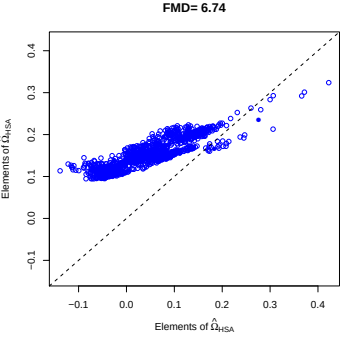


Note : FMD distance between covariance matrices

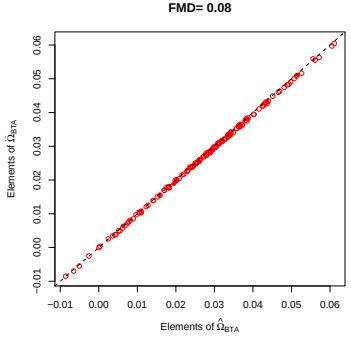
Bayenv2

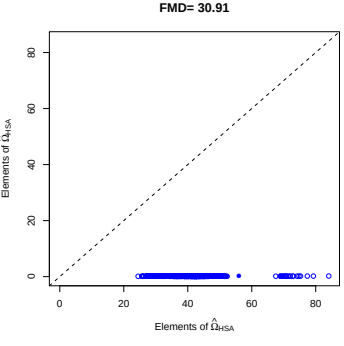


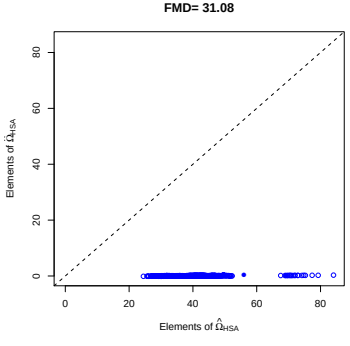




BayPass (ρ=1)







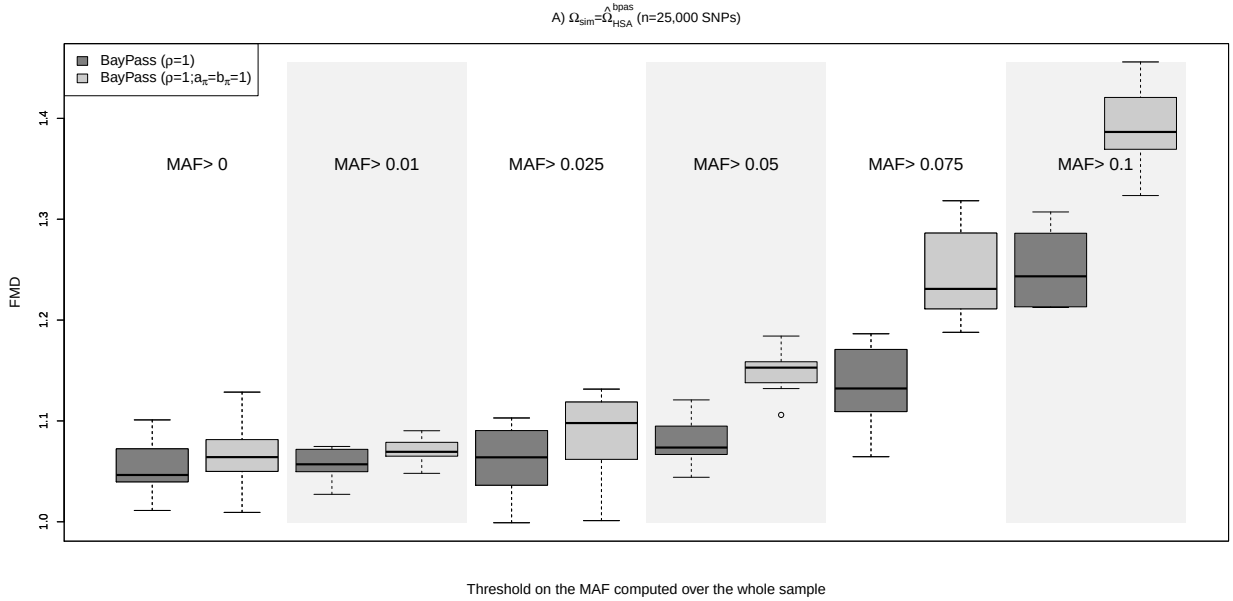
BayPass (ρ=J)

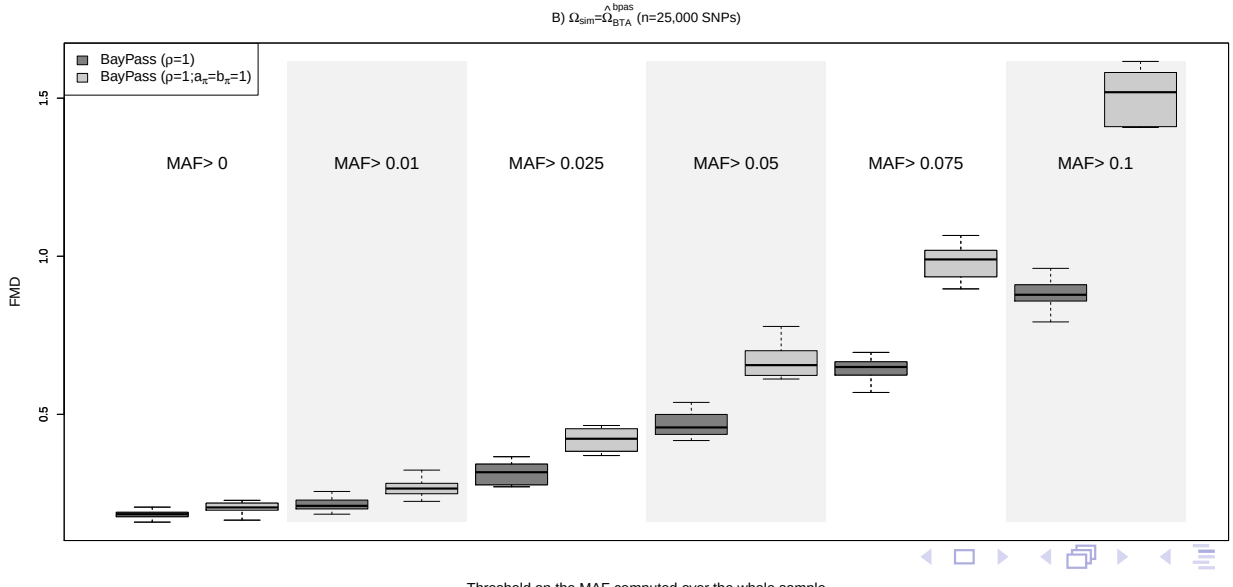
Definition (Forstner and Moonen, 2003)

$$\mathrm{FMD}\left(\boldsymbol{\Omega_1}, \boldsymbol{\Omega_2}\right) = \sqrt{\sum_{j=0}^{J} \ln^2 \lambda_j \left(\boldsymbol{\Omega_1}, \boldsymbol{\Omega_2}\right)}$$

- λ_j =j$^{\text{th}}$ generalized eigenvalues ($\det(\boldsymbol{\Omega_1} - \lambda \boldsymbol{\Omega_2}) = 0$)
- Symmetric
- In practive, FMD $< 1 \Rightarrow$ elements of the two matrices are similar
- See the *fmd.dist()* R function in the BayPass package (*baypass_utils.R*)

Estimation of a_π and b_π improves robustness to SNP ascertainment bias





Real Life Example : HSA allele count data

The allele count data file (from Coop et al., 2010)

- $J = 52$ worldwide human populations from the Human Genome Diversity Panel genotyped at $I = 2,333$ SNPs
- (partial) view of the allele count file : "hgdp.geno"

```
0 22 0 16 0 44 0 42 0 24 0 12 0 44 1 29 1 47 0 24 4 52 0 26 1 47....[2x52=104 col.]
0 22 0 16 0 46 0 42 0 24 0 12 0 44 3 27 12 36 3 21 9 47 2 24 12.. ..[2x52=104 col.]
14 8 12 4 38 8 35 7 21 3 11 1 33 11 5 25 16 32 8 16 18 38 8 18 5....[2x52=104 col.]
..............................
[2333 rows in total]
```

Examples of command lines

- Running with default parameters :

```
i_baypass -npop 52 -gfile hgdp.geno -outprefix corehgdp
```

- Changing some MCMC parameters : e.g. :

```
i_baypass -npop 52 -gfile hgdp.geno -pilotlength 500 -burnin 10000
```

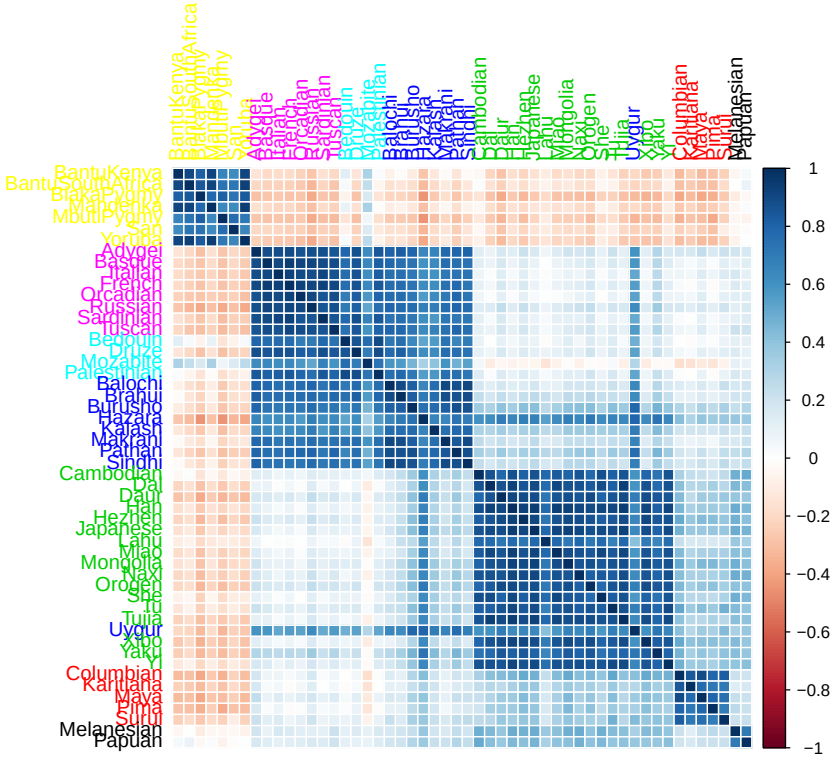
- Changing some modeling options : e.g. :

```
i_baypass -npop 52 -rho 52
```

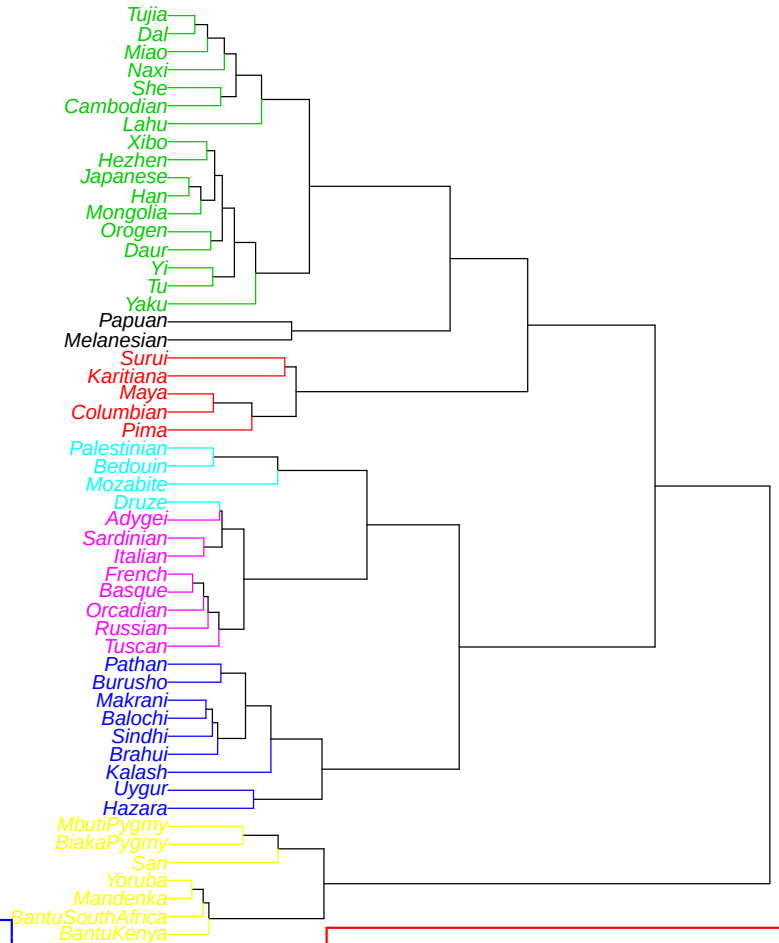
- If your are lost, use the option : *-help*

Results (default options : 35min on 1CPU) visualized within R

A) Correlation map based on $\hat{\Omega}^{bpas}_{HSA}$ (with ρ =1)



B) Hier. clust. tree based on $\hat{\Omega}^{bpas}_{HSA}$ ($d_{ij}=1-\rho_{ij}$)




```
>require(corrplot) ; require(ape)
>omega=as.matrix(read.table("corehgdp_mat_omega.out"))
>cor.mat=cov2cor(omega)
>corrplot(cor.mat)
>plot(as.phylo(hclust(as.dist(1-cor.mat))))
```

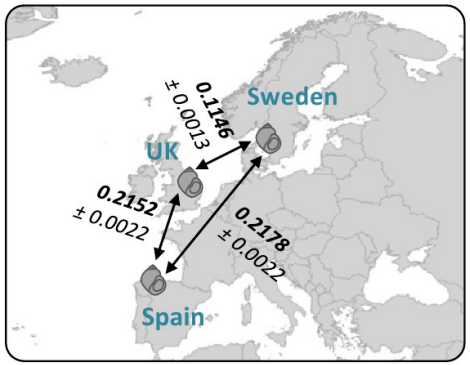
```
more corehgdp_summary_beta_params.out")
PARAM Mean SD
a_beta_pi       1.054237        0.040531
b_beta_pi       1.783356        0.067600
```

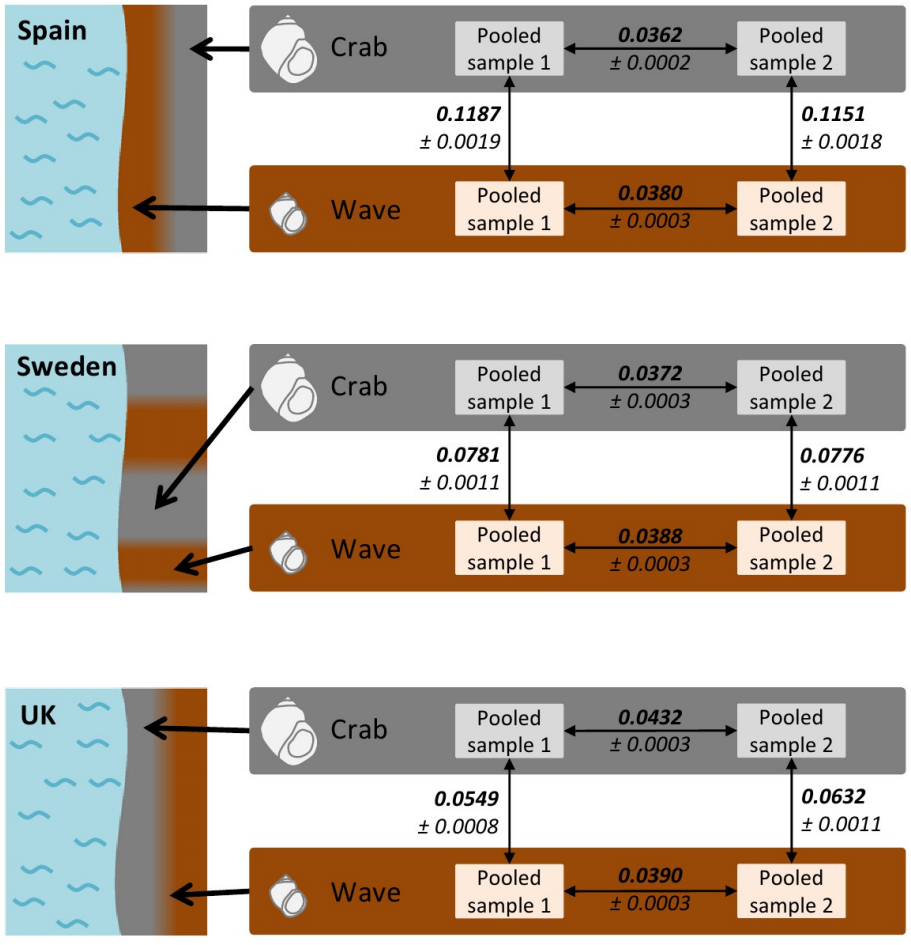
Real Life Example : Littorina Pool–Seq read count data

Do the same genes underlie parallel phenotypic divergence in different *Littorina saxatilis* populations?

MOLECULAR ECOLOGY
Molecular Ecology (2014) **23**, 4603–4616

A. M.WESTRAM,* J. GALINDO,† M. ALM ROSENBLAD,‡ J. W. GRAHAME,§ M. PANOVA¶ and R. K. BUTLIN*¶





‘Crab’ habitat

‘Wave’ habitat

Real Life Example (2) : Littorina Pool–Seq read count data

The read count data file (adapted from Westram et al., 2010)

- $J = 12$ LSA pops (Pool Size $\simeq 100$) ; $I = 53,387$ ascertained SNPs
- (partial) view of the allele count file : "lsa.geno"

```
0 64 0 49 0 39 0 81 0 57 0 57 0 34 0 45 1 23 4 34 5 98 0 40        [2x12=24 col.]
101 0 77 0 96 2 138 0 65 3 98 7 70 2 83 4 53 0 73 2 168 2 54 0     [2x12=24 col.]
59 3 27 0 49 3 49 2 59 0 74 2 48 1 52 0 27 1 47 1 105 10 42 2     [2x12=24 col.]
..............................
[53,387 rows in total]
```

- Haploid pool size file : "lsa.poolsize"

```
100 100 100 100 100 100 100 100 100 100 100 100                   [12 col.]
```

Examples of command lines

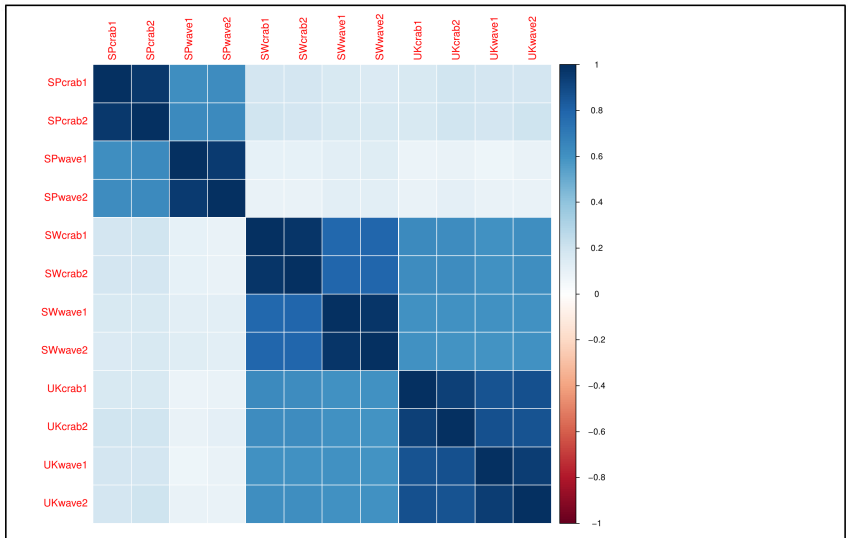
- Running with default parameters :

```
i_baypass -npop 12 -gfile lsa.poolsize -poolsizefile lsa.poolsize -outprefix corelsa
```

- Changing some MCMC parameters (e.g., $\delta_{y,0}$) :

```
i_baypass -npop 12 -gfile lsa.poolsize -poolsizefile lsa.poolsize -d0yij 10
```

Results (default options : 2h30 on 1CPU) visualized within R




```
more corelsa_summary_beta_params.out")
 PARAM Mean SD
a_beta_pi       0.208126      0.005269
b_beta_pi       0.213517      0.005343
```

Identifying overly differentiated SNPs : the XtX statistic

Definition (Guenther and Coop, 2013)

- Let the vector $\mathbf{X_i} = \mathbf{\Gamma^{-1}} \frac{\boldsymbol{\alpha}_i^\star - \pi_i}{\sqrt{\pi_i(1-\pi_i)}}$

  where $\mathbf{\Gamma^{-1}}$ *results from the Cholesky decomposition of* $\mathbf{\Omega}$ *(i.e.,* $\mathbf{\Omega} = \mathbf{\Gamma^{-1}\Gamma}$*)*

- $\mathbf{X_i} \simeq$ vector of scaled pop. allele frequencies

  e.g., if $\mathbf{\Omega}$ *diagonal (i.e.,* $\omega_{i \neq j} = 0$*),* $\mathbf{X_i} = \left\{ \frac{\alpha_{ij}^\star - \pi_i}{\sqrt{\omega_{ii}\pi_i(1-\pi_i)}} \right\}$

- $\mathrm{X^tX}_i = \mathrm{Var}\left(\mathbf{X_i}\right) = \frac{(\boldsymbol{\alpha}_i^\star - \pi_i)\mathbf{\Omega^{-1}}(\boldsymbol{\alpha}_i^\star - \pi_i)}{\pi_i(1-\pi_i)}$

- This statistic is thus homogeneous to a SNP–specific FST (variance) while explicitly correcting for $\mathbf{\Omega}$

  share some similarities with the FLK (Bonhomme et al., 2001) except $\mathbf{\Omega}$ *is not computed in the same way*.

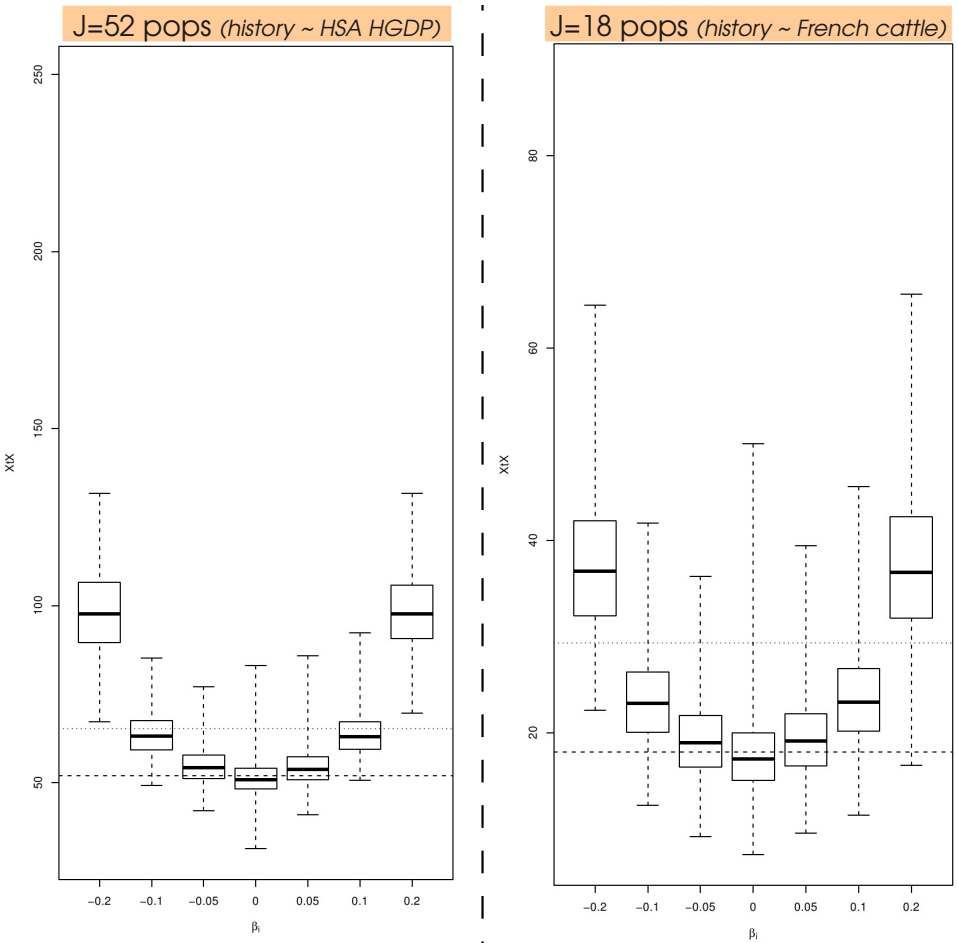
Computation in BayPass

- $\widehat{\mathrm{X^tX}_i}$: post. mean over the post burn–in samples (in the core model integrate over Ω samples $\neq$ BayEnv2)
- Available in the *[outprefix_]summary_pi_xtx.out* output file.

Example of *[outprefix_]summary_pi_xtx.out* file

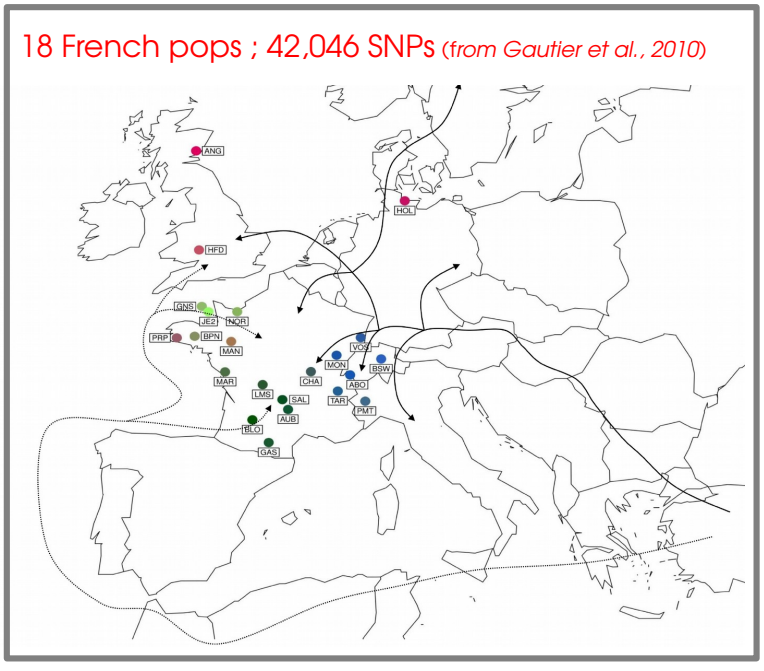
MRK	M_P	SD_P	DELTA_P	ACC_P	M_XtX	SD_XtX
1	0.51470535	0.07950854	0.40000000	0.55100000	12.40573883	3.16409933
2	0.29002817	0.07174040	0.40000000	0.48740000	12.88555815	3.93091494
3	0.40434515	0.07927629	0.40000000	0.53428000	12.97591190	3.77959282
4	0.03868250	0.02223511	0.10485760	0.39708000	16.71037046	5.22830252
5	0.18624566	0.05480118	0.32000000	0.46132000	15.19206580	4.50959616
6	0.14491668	0.05090899	0.25600000	0.47300000	17.38096802	4.80911100
7	0.66119177	0.07470559	0.40000000	0.52116000	13.45442225	3.22967630
8	0.24679971	0.05521047	0.32000000	0.47768000	27.75047460	4.94576745
9	0.31537781	0.06458381	0.40000000	0.44564000	29.63308572	4.67469577
10	0.07504550	0.03442540	0.16384000	0.43752000	16.88825304	5.06646171
11	0.30210380	0.07034558	0.40000000	0.48780000	14.70875284	3.74375902
12	0.63045381	0.07588395	0.40000000	0.53700000	13.11244089	3.69764101
13	0.26493616	0.06936325	0.40000000	0.45708000	15.36246327	4.26431306
14	0.13312173	0.05001387	0.25600000	0.44716000	14.19822577	4.64740382
15	0.37049327	0.06443027	0.40000000	0.46468000	29.45952453	4.71207933
16	0.56428117	0.06831302	0.40000000	0.48980000	27.71959099	4.43895572
17	0.18101669	0.05679940	0.32000000	0.45896000	14.15783797	3.94264955
18	0.13141500	0.04532795	0.25600000	0.44468000	16.01465502	4.44044949
19	0.30642070	0.06822290	0.40000000	0.47304000	18.72378974	4.23774213
20	0.37771103	0.07482055	0.40000000	0.50848000	18.10229741	3.96054024
21	0.55726408	0.08274862	0.40000000	0.54704000	13.86471704	3.51338586
22	0.54216634	0.08066086	0.40000000	0.54860000	13.11173563	3.14209717
23	0.29936304	0.06576716	0.40000000	0.46512000	19.95493751	4.58003052
24	0.03898583	0.02551602	0.13107200	0.34188000	15.39678543	5.12916276
25	0.31676502	0.07429964	0.40000000	0.50088000	13.03886301	3.88243549
26	0.24400538	0.06706918	0.40000000	0.45816000	11.25712465	3.77212442
27	0.30220266	0.06742214	0.32000000	0.54956000	17.78647873	3.75171990
28	0.33676440	0.07152014	0.40000000	0.47908000	19.05758989	4.04334918
29	0.45159756	0.07549984	0.50000000	0.44748000	18.45704617	3.92666774
30	0.76027493	0.06514416	0.32000000	0.52580000	15.65486452	3.74914608
31	0.17869438	0.05368651	0.25600000	0.51152000	19.70089054	4.84022180

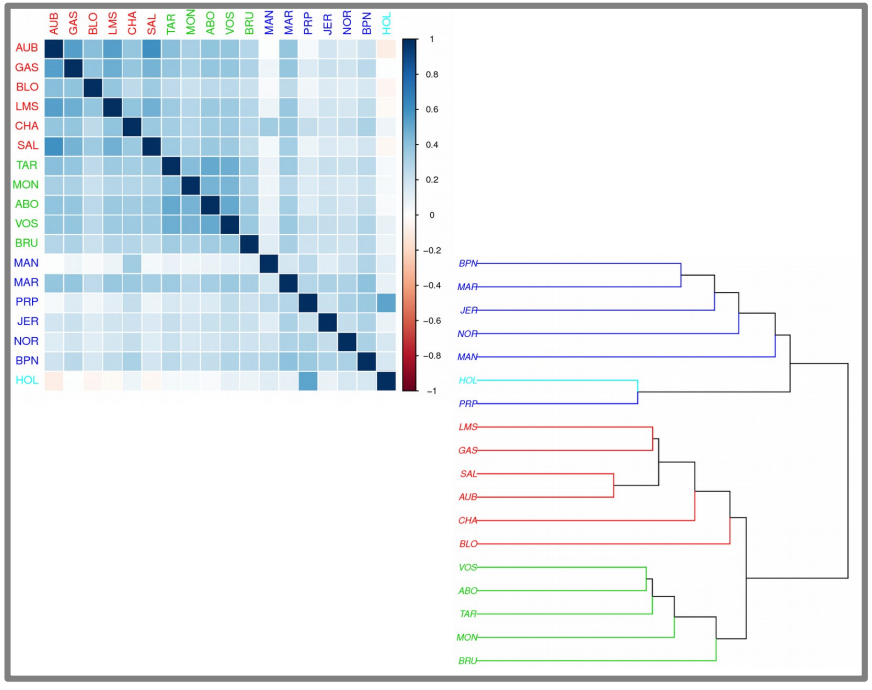
Using X^tX to identify SNPs under selection

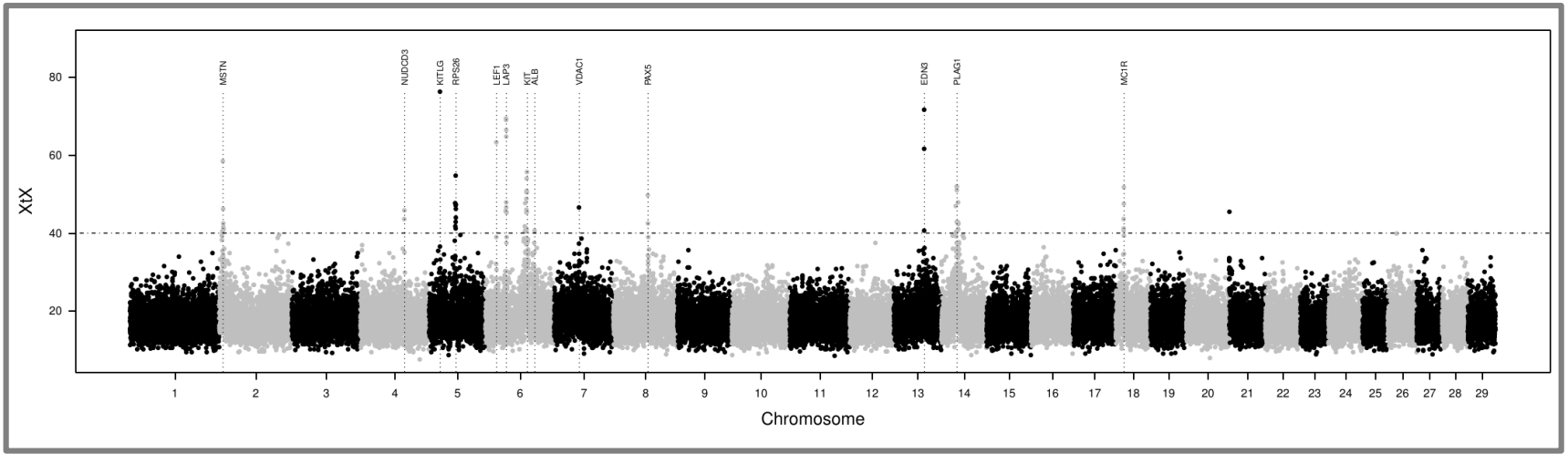


- Selected SNPs = Extreme $\widehat{X^tX_i}$
  (overly differentiated SNPs)
- Correct for population history
  ($\approx$ immune to demographic confounding factors)
- How extreme to be candidate ?
  - In theory, $X^tX \sim \chi^2(J)$
    (i.e. $\mathrm{E}(X^tX) = \mathrm{Var}(X^tX) = J$*)*
  - Here, $\mathrm{E}\left(\widehat{X^tX}\right) = J$ but $\mathrm{Var}\left(\widehat{X^tX}\right) \ll J$
    (shrinkage towards to the mean)
- Calibration by analysing PODs (e.g. 100,000 SNPs) generated under the inference model
  - $\boldsymbol{\Omega}^{\mathrm{sim}} = \widehat{\boldsymbol{\Omega}}$, $a_\pi^{\mathrm{sim}} = \widehat{a_\pi}$ and $b_\pi^{\mathrm{sim}} = \widehat{b_\pi}$
  - See *simulate_baypass()* R function
  - Comparing $\widehat{\boldsymbol{\Omega}^{\mathrm{sim}}}$, $\widehat{a_\pi^{\mathrm{sim}}}$ and $\widehat{b_\pi^{\mathrm{sim}}}$ (PODs) with $\widehat{\boldsymbol{\Omega}}$, $\widehat{a_\pi}$ and $\widehat{b_\pi}$ ensure that the PODs faithfully mimic the real data set.

Cattle Example







Using X^tX to identify SNPs under selection

Key characteristics

- Robust to demographic history (via $\mathbf{\Omega}$)
- No prior information about population history needed ($\neq$ Hierarchical island model)
- But...do not account for haplotype information (see hapFLK)

Limitations...common to all indirect genome scan approaches

- Biological interpretations (underlying selective pressure ?) require an annotated genome for the species of interest (or a closely related one)
- Highly prone to misleading story telling issues (e.g., Pavlidis et al., 2012).
- Experimental validation (if possible) $\Rightarrow$ reverse ecology (e.g., Li et al., 2008)

Genome wide association with population–specific covariates

- Historically presented with environmental variables

  $\Rightarrow$ proxies for ecological pressure

- SAM (Joost et al., 2007) : univariate logistic regression of pop. all. freq. with env. variable

  $\Rightarrow$ does not account for neutral all. freq. covariance

- BayEscenv (de Villemereuil et al., 2015) : association between residuals of a logistic regression of marker and pop specific F_{ST} (with marker and population specific effects) and the environmental variable

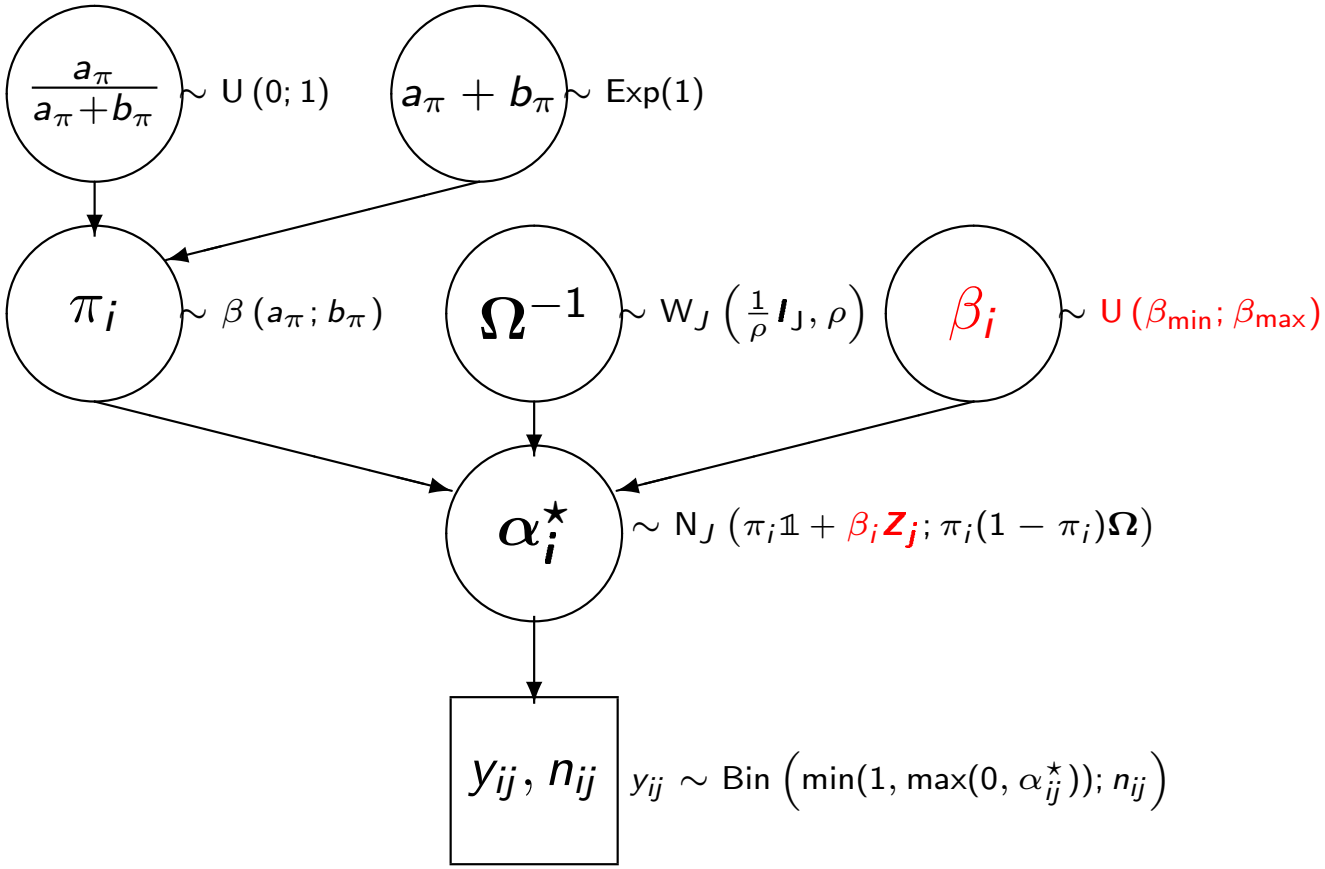
  $\Rightarrow$ basic modeling of the pop. structure (F–model)

- LFMM (Frichot et al.,2013) : assess association via a mixed model with latent factors to account for population structure

- Bayenv (Coop et al.,2010) and BayPass : extent the previous model to include a "fixed" environmental effect.

see de Villemereuil et al. (2014) for a comparison of the approaches via simulation under realistic scenarios (! problems in Bayenv might have penalized the Bayenv model)

The BayPass "standard" covariate model



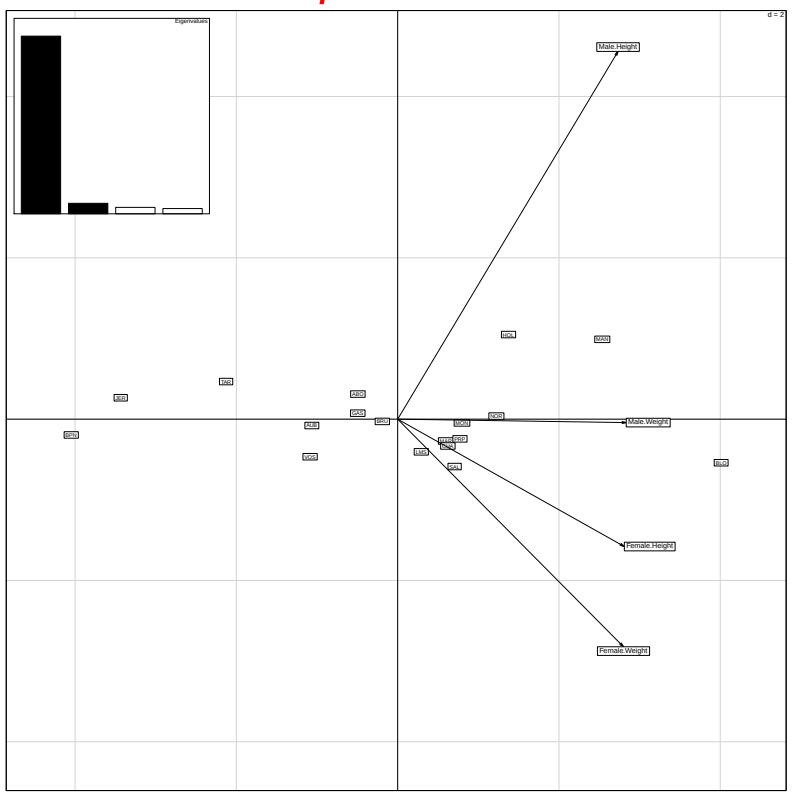
- Similar to BayEnv model (Coop et al., 2010) with additional extensions
  - Priors on a_π, b_π and $\boldsymbol{\Omega}^{-1}$ (see above)
  - β_{min} and β_{max} can be set by the user
    (by default $\beta_{min} = -0.3$ instead of -0.1 and $\beta_{max} = 0.3$ instead of 0.1)
  - Rk. : Gaussian priors on the β_i's were found to perform poorly

Note : How does the covariable file look like ?

Ex. : *18 French cattle breeds and two covariates : Morpho. Score and Pieblad pattern*

```
-0.5484 -1.0961 0.411 -0.2549 2.0671 1.3074 0.3085 0.1509 -0.2542....[18 col.]
-1. -1. 1. -1. -1. 1. -1. -1. 1. 1. -1. 1. 1. -1. 1. -1. -1. 1.  ....[18 col.]
[2 rows in total]
```

Ex. SMS Morpho score



Best practices...

- Generally, better to scale the covariables (automatically done by the *-scalecov* option)
- If several covariables are considered, use PCA to analyze only a few (uncorrelated) PC ("synthetic" scores to analyze)

Estimating the β_i's and assessing association significance : A) Via Importance Sampling

- Very elegant (and computationally efficient) algorithm proposed by Coop et al. (2010) to estimate the Bayes Factors
- Only requires samples drawn under the core model
- In BayPass an IS algorithm is used by default with pop. covariables (*-efile* option) to estimate for each SNP (*[outprefix_]summary_betai_reg.out* file) :
  - $\widehat{\mu(\beta_i)}$ and $\widehat{\sigma(\beta_i)}$ allowing to derive a Z–score $Z_i = \frac{\widehat{\mu(\beta_i)}}{\widehat{\sigma(\beta_i)}}$
  - "empirical Bayesian P–value" : $\text{eBP}_{\text{is}} = -\log_{10}\left(1 - 2\left|0.5 - \Phi\left(Z_i\right)\right|\right)$
    to assess association significance : eBP>4 $\Leftrightarrow$ $|Z_i| > 3.7$.
  - An estimate of the Bayes Factor $\text{BF}_{\text{is}} = 10\log_{10}\left(\widehat{\text{BF}}\right)$ (in deciban units) comparing models with *vs.* without association (i.e. $\beta_i = 0$)
    Jeffreys' rule : $15 < BF < 20 \Rightarrow$ "very strong evidence" ; BF> $20 \Rightarrow$ "decisive evidence"
  - All these values are also computed for read count data (Pool–Seq mode) and (an estimate of) $\boldsymbol{\Omega}$ is not required ($\neq$ Bayenv2)
    integrate over the posterior of $\boldsymbol{\Omega}$ (unless it is given : -omega_file option)

Estimating the β_i's and assessing association significance :

A) Via Importance Sampling (example of *[outprefix_]summary_betai_reg.out* file)

COVARIABLE	MRK	M_Pearson	SD_Pearson	BF(dB)	Beta_is	SD_Beta_is	eBPis
1	1	0.17418897	0.14673005	-6.61213338	0.02117456	0.01800173	0.62070462
1	2	-0.40375551	0.12943690	-1.79324212	-0.04506824	0.01660997	2.17644913
1	3	-0.33337571	0.13540831	-3.36446055	-0.04081512	0.01834861	1.58302963
1	4	0.05681107	0.17623422	-11.03967159	0.00342988	0.00898525	0.15325083
1	5	0.33921744	0.12334830	-3.80163833	0.03455217	0.01379059	1.91263533
1	6	0.19531076	0.12352730	-7.12411706	0.01959901	0.01310393	0.87049561
1	7	-0.26656129	0.13578744	-5.40375137	-0.03142045	0.01683963	1.20718524
1	8	0.04049091	0.09606448	-8.19767856	0.00764043	0.01493099	0.21548991
1	9	0.00706596	0.08722841	-8.10817601	0.00248200	0.01512992	0.06063299
1	10	0.03066293	0.13386872	-10.38498439	0.00235689	0.00960991	0.09352600
1	11	-0.10185271	0.14885197	-7.51011413	-0.01258991	0.01753351	0.32538926
1	12	-0.11326002	0.15547431	-7.27371838	-0.01368219	0.01807884	0.34759355
1	13	0.12822636	0.13420140	-7.57908519	0.01519291	0.01591720	0.46873542
1	14	0.16353677	0.14531857	-8.25471224	0.01403572	0.01207165	0.61092056
1	15	0.36084318	0.08994236	2.53452965	0.06462972	0.01728093	3.73504008
1	16	-0.10281822	0.10291052	-6.84646320	-0.01806784	0.01857580	0.48053302
1	17	0.24449648	0.12257616	-6.78250180	0.02379334	0.01213777	1.30134313
1	18	-0.18873606	0.14149988	-7.39670603	-0.01786422	0.01348850	0.73196281
1	19	-0.22342308	0.11604000	-5.49662447	-0.03030307	0.01684859	1.14212810
1	20	0.03284065	0.12396529	-7.76765270	0.00396459	0.01756071	0.08545367
1	21	0.20732700	0.15049855	-6.04338770	0.02463278	0.01914579	0.70281459
1	22	0.24403160	0.14665890	-5.67688519	0.02847890	0.01838142	0.91613110
1	23	-0.20895894	0.13867526	-5.25078649	-0.02904647	0.01785843	0.98360791
1	24	-0.44572874	0.13271740	-1.92752090	-0.02234165	0.00981050	1.64269059
1	25	0.22684063	0.14512108	-6.00823399	0.02603745	0.01760658	0.85642004
1	26	-0.20981631	0.16270380	-6.97050032	-0.01977467	0.01595457	0.66719006
1	27	0.44667791	0.10449866	2.45356609	0.05986790	0.01630654	3.61758512
1	28	0.44434051	0.12019628	1.85460901	0.06053785	0.01830191	3.02662319

Estimating the β_i's and assessing association significance : B) Via MCMC

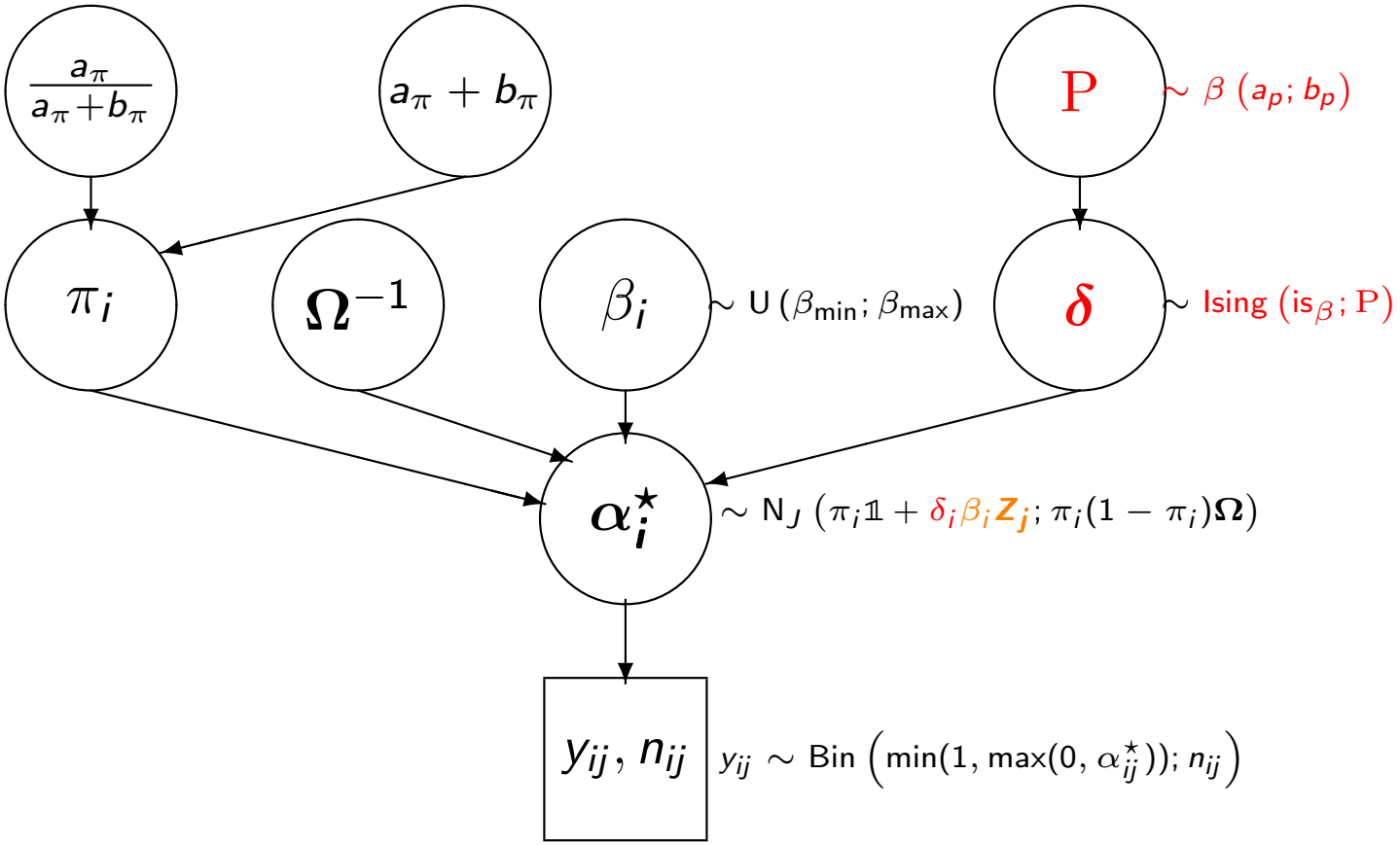
- Sampling of the β_i's added to the MCMC sampler (M–H update) : *-covmcmc* option
- (At the moment) require an estimate of $\mathbf{\Omega}$ given using the *-omega_file* option
- $\widehat{\mu(\beta_i)}$ and $\widehat{\sigma(\beta_i)}$ are estimated using the samples drawn from the corresponding posterior distribution $\Rightarrow$ eBP_{mc}
- (*[outprefix_]summary_betai.out* file)

Estimating the β_i's and assessing association significance :

B) Via MCMC (example of *[outprefix_]summary_betai.out* file)

COVARIABLE	MRK	M_Beta	SD_Beta	AccRateB	DeltaB	eBPmc
1	1	0.00792721	0.46144000	0.15258789	0.04131320	0.07168833
1	2	-0.04053921	0.42800000	0.15258789	0.03961729	0.51402324
1	3	-0.02826815	0.36964000	0.19073486	0.04154143	0.30434326
1	4	0.00290478	0.34920000	0.07812500	0.01941971	0.05497636
1	5	0.06980853	0.44400000	0.12207031	0.03262616	1.48968098
1	6	0.03823162	0.41276000	0.12207031	0.02984712	0.69848490
1	7	-0.03731710	0.44520000	0.15258789	0.03759355	0.49365427
1	8	0.01855285	0.39824000	0.15258789	0.03458441	0.22793749
1	9	0.00362525	0.42952000	0.15258789	0.03644473	0.03585206
1	10	0.00497661	0.41176000	0.09765625	0.02304751	0.08142262
1	11	-0.01163830	0.42588000	0.15258789	0.03918325	0.11551651
1	12	-0.02872653	0.45736000	0.15258789	0.04223415	0.30417259
1	13	0.02460031	0.41328000	0.15258789	0.03639899	0.30178196
1	14	0.02076340	0.38260000	0.12207031	0.02674876	0.35891418
1	15	0.08634889	0.44400000	0.15258789	0.03829315	1.61731923
1	16	-0.03533303	0.37676000	0.19073486	0.04101240	0.41010466
1	17	0.03391202	0.42744000	0.12207031	0.02968959	0.59625706
1	18	-0.03090144	0.40728000	0.12207031	0.02996948	0.51928146
1	19	-0.03241210	0.44172000	0.15258789	0.03969611	0.38277846
1	20	0.02259973	0.45732000	0.15258789	0.04276294	0.22390909
1	21	0.03462215	0.37604000	0.19073486	0.04215259	0.38568787
1	22	0.04424623	0.37080000	0.19073486	0.04004690	0.56989238
1	23	-0.02800792	0.43476000	0.15258789	0.03934238	0.32191389
1	24	-0.03300417	0.37620000	0.07812500	0.02162415	0.89638800
1	25	0.03683909	0.42832000	0.15258789	0.03830703	0.47338862
1	26	-0.03894193	0.40048000	0.15258789	0.03629202	0.54780891
1	27	0.07834605	0.42896000	0.15258789	0.03752925	1.43375166
1	28	0.07813112	0.43344000	0.15258789	0.03713981	1.45094315
1	29	-0.01757049	0.36784000	0.19073486	0.04137818	0.17320950

The BayPass "AUX" ("auxiliary variable") covariate model



- The binary variable δ_i indicates whether the SNP is associated ($\delta_i = 1$) or not ($\delta_i = 0$)
  $\mathrm{P}\,[\delta_i = 1 \mid \textit{data}]$: *posterior prob. the locus is associated to the covariable* $\boldsymbol{Z}$

- P = Prior proportion of associated SNPs
  - Beta prior distribution (def. $a_P = 0.02$ and $b_P = 1.98 \Rightarrow \mathbb{E}\,(\mathrm{P})$ =1% but $\mathrm{P}\,[P > 10\%] = 2.8\%$).
  - Integrating over the uncertainty on P allows to deal with multiple testing issues.

The Ising prior on the vector of δ_i's (AUX model)

Definition (see Duforet-Frebourg et al., 2014)

$$\pi\left(\boldsymbol{\delta} \mid P, \text{is}_\beta\right) \propto P^{s_1}(1-P)^{s_0} e^{\eta \text{is}_\beta}$$

- $s_1 = \sum_{i=1}^{I} \mathbb{I}_{\delta_i=1}$ (respectively $s_0 = I - s_1$) are the number of SNPs associated (respectively not associated) with the covariable
- $\eta = \sum_{i \sim j} \mathbb{I}_{\delta_i=\delta_j}$ is the number of pairs of consecutive markers (neighbors) that are in the same state at the auxiliary variable (i.e., $\delta_i = \delta_{i+1}$).
- is_β determines the level of spatial homogeneity of the auxiliary variables between neighbors.

Special cases

- default : $\text{is}_\beta = 0$ (no spatial dependency : e.g., no mapping information available) $\Leftrightarrow \delta_i \sim \text{Ber}(P)$ (see Riebler et al., 2008)
- $\text{is}_\beta > 0 \Leftrightarrow$ similar δ_i's are assumed to cluster in the genome (the higher the is_β, the higher the level of spatial homogeneity).

Estimating the β_i's, the δ_i's and assessing association significance under the AUX model

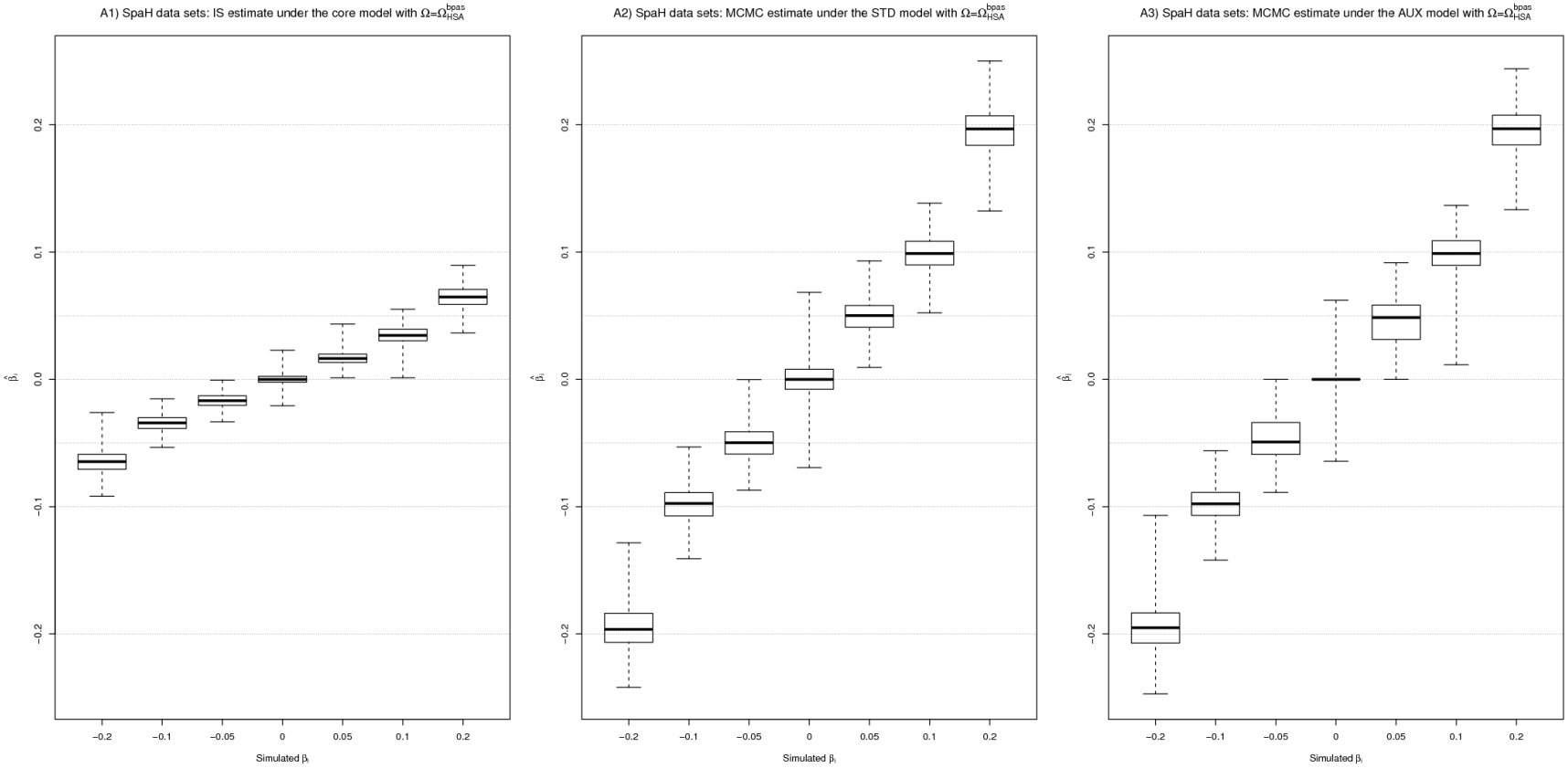
- δ_i's and P sampled using MCMC (Gibbs updates) : *-auxmodel* option
- (At the moment) require an estimate of $\mathbf{\Omega}$ given using the *-omega_file* option
- $\widehat{\mu(\beta_i)}$ are estimated using the samples drawn from the corresponding posterior distribution
- Straightforward to derive a Bayes Factor ($\mathrm{BF}_{\mathrm{mc}}$) from $\mathrm{P}[\delta_i = 1 \mid \textit{data}]$ (e.g., Gautier et al., 2009)
  - Prior odds$= \mathrm{P}[\delta_i = 1]/[1 - \mathrm{P}(\delta_i = 1)] = \frac{\mathbb{E}[P]}{1-\mathbb{E}[P]}$
  - Posterior odds$= \mathrm{P}[\delta_i = 1|\textit{data}]/[1 - \mathrm{P}(\delta_i = 1|\textit{data})]$
  - $\mathrm{BF}_{dB} = 10\log_{10}\left(\frac{\text{Post. odds}}{\text{Prior odds}}\right)$
- (*[outprefix_]summary_betai.out* file)

The AUX model : example of *[outprefix_]summary_betai.out* file

COVARIABLE	MRK	M_Beta	SD_Beta	M_Delta	BF
1	1	0.00000000	0.00000000	0.00000000	-1.30517760
1	2	0.00000000	0.00000000	0.00000000	-1.30517760
1	3	0.00000000	0.00000000	0.00000000	-1.30517760
1	4	0.00000000	0.00000000	0.00000000	-1.30517760
1	5	0.00000000	0.00000000	0.00000000	-1.30517760
1	6	0.00000000	0.00000000	0.00000000	-1.30517760
1	7	0.00000000	0.00000000	0.00000000	-1.30517760
1	8	0.00000000	0.00000000	0.00000000	-1.30517760
1	9	0.00000000	0.00000000	0.00000000	-1.30517760
1	10	0.00000000	0.00000000	0.00000000	-1.30517760
1	11	0.00000000	0.00000000	0.00000000	-1.30517760
1	12	0.00000000	0.00000000	0.00000000	-1.30517760
1	13	0.00000000	0.00000000	0.00000000	-1.30517760
1	14	0.00000000	0.00000000	0.00000000	-1.30517760
1	15	0.00000000	0.00000000	0.00000000	-1.30517760
1	16	0.00000000	0.00000000	0.00000000	-1.30517760
1	17	0.00000000	0.00000000	0.00000000	-1.30517760
1	18	0.00000000	0.00000000	0.00000000	-1.30517760

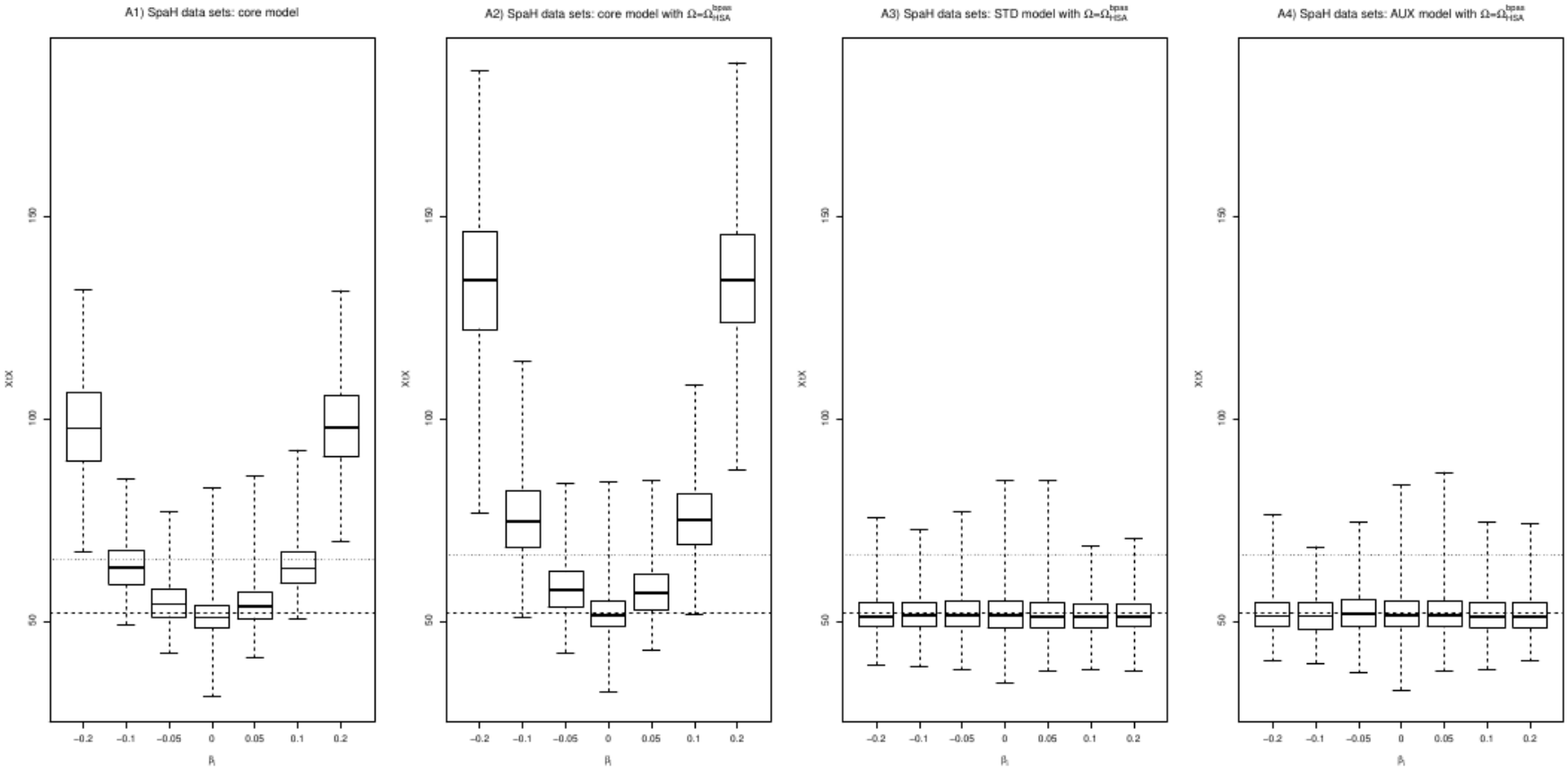
Comparison of models/implementations on simulated data :

A) Estimation of the β_i's



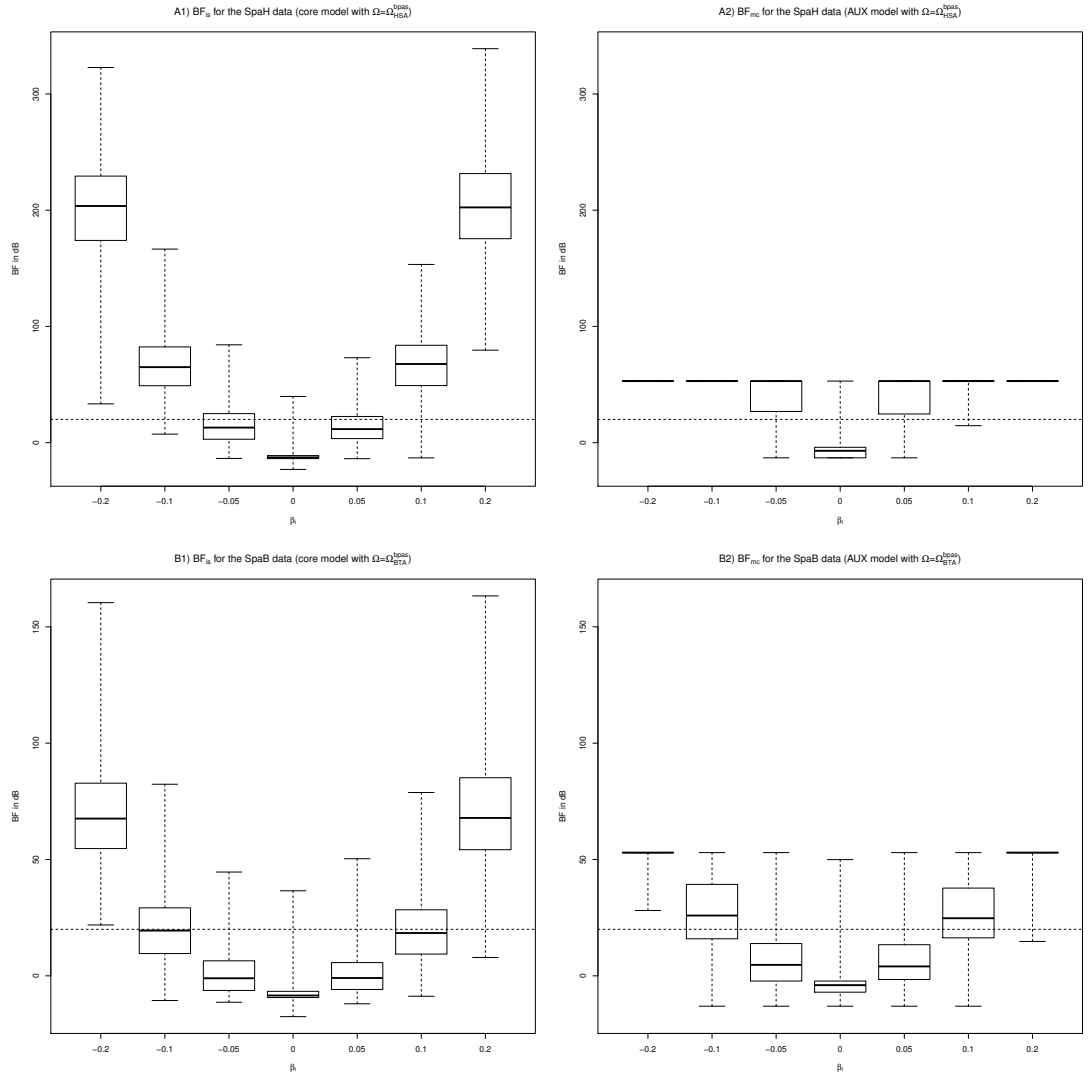
Comparison of models/implementations on simulated data :

B) covariate models correct X^tX for fixed covariable effect



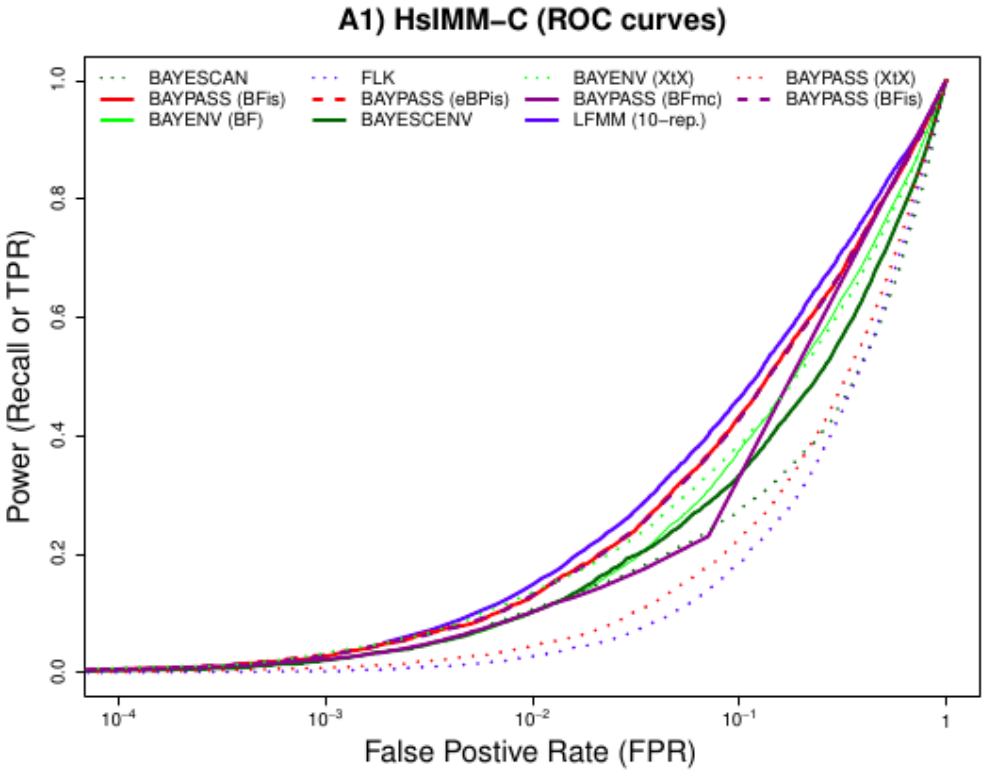
Comparison of models/implementations on simulated data :

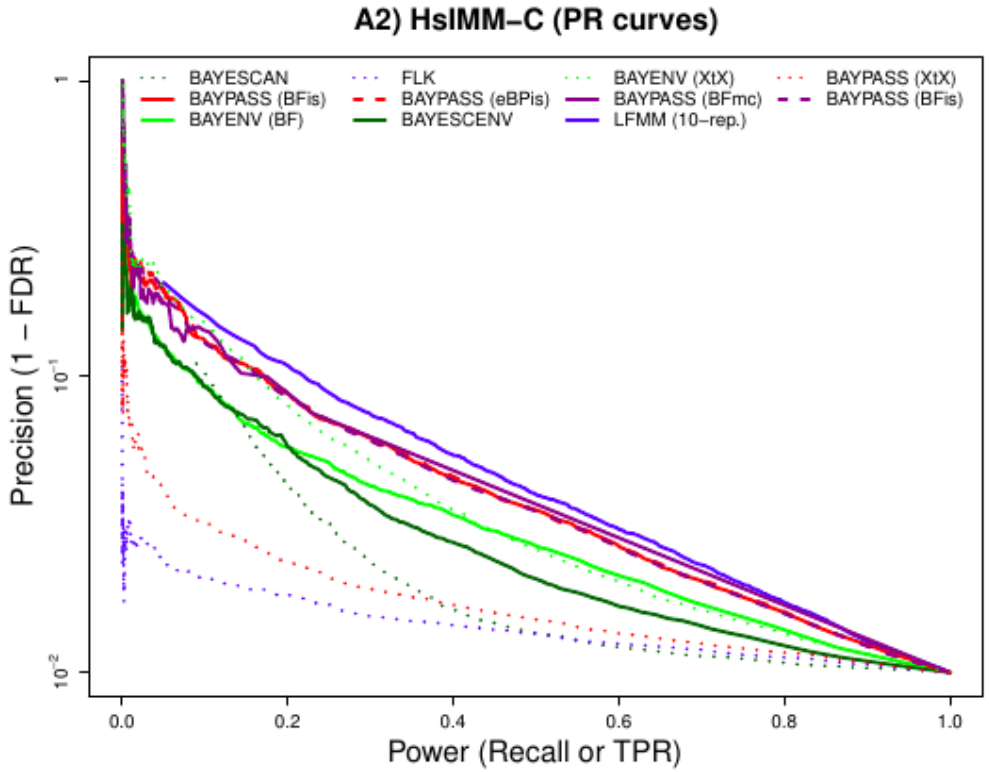
C) Computation of Bayes Factors



Performance on (realistic) simulated data (as in De Villemereuil et al., 2014)

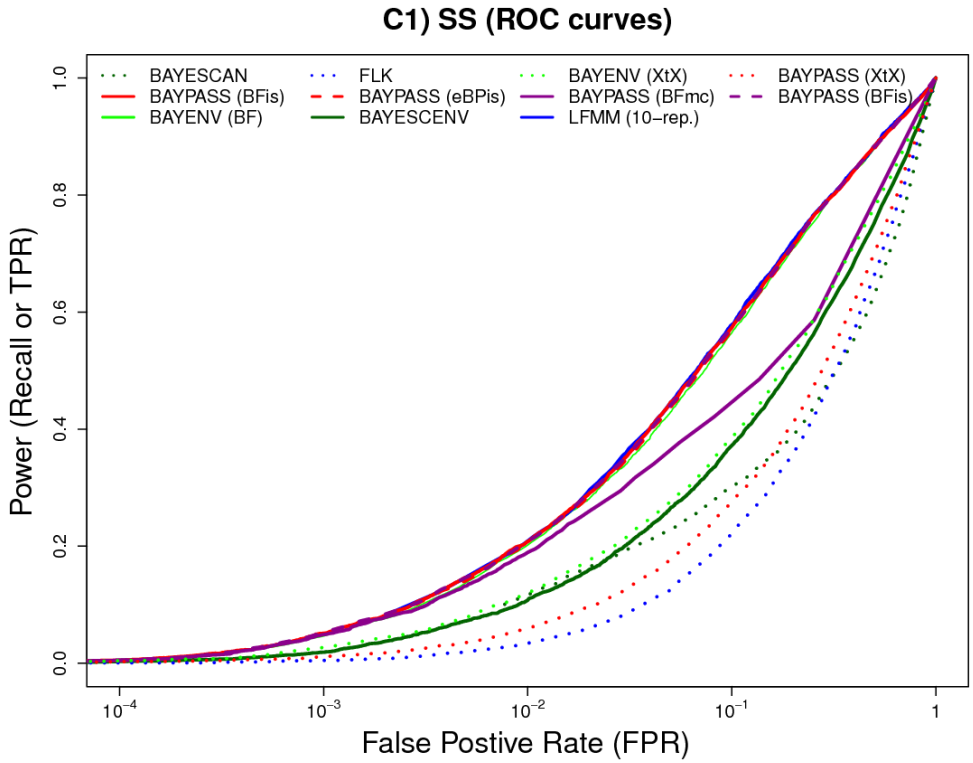
A) HsIMM ; Polygenic Selection (according to an environmental gradient)

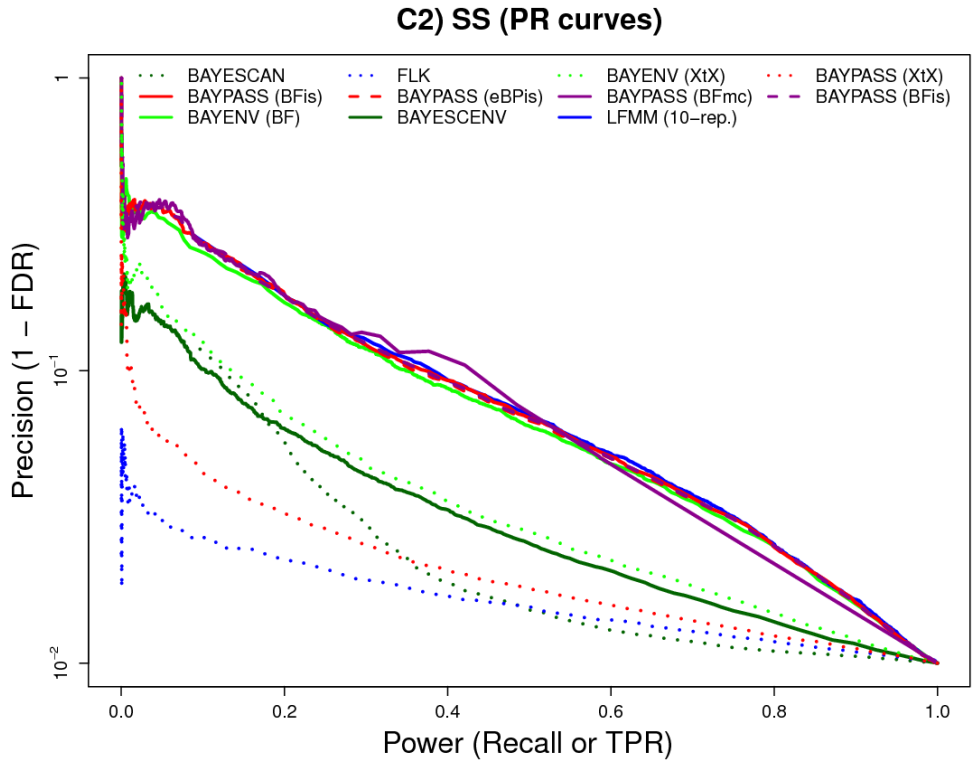




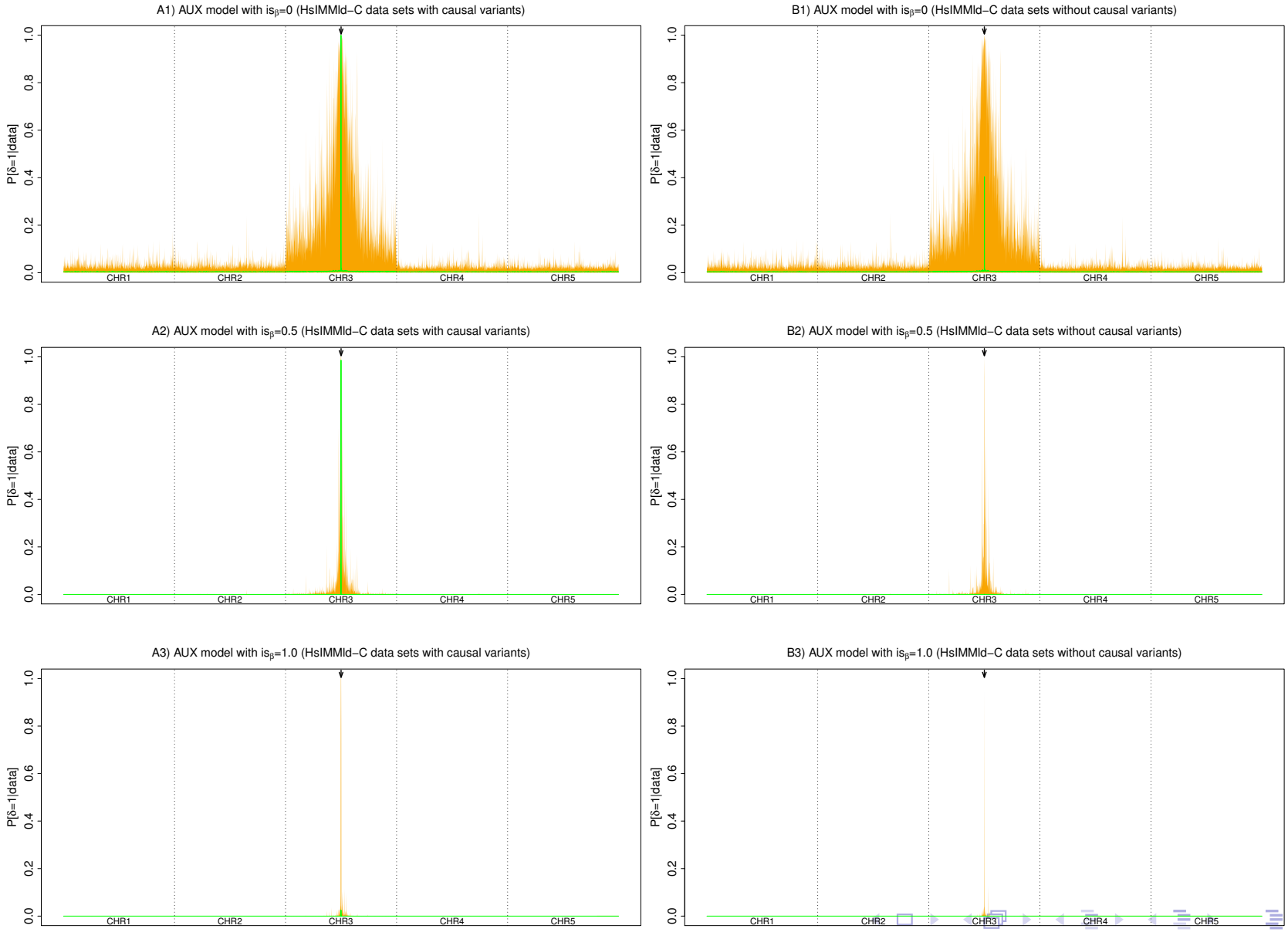
Performance on (realistic) simulated data (as in De Villemereuil et al., 2014)

B) Stepping Stone ; Polygenic Selection (according to an environmental gradient)





AUX model : Ising prior



French cattle example : command lines

- Running with default parameters (Imp. Sampling estimates) :

```
i_baypass -npop 18 -gfile bta.geno -efile trait.dat -outprefix covis
```

- Running the covariate model (MCMC mode)

```
i_baypass -npop 18 -gfile bta.geno -efile trait.dat -omegafile covis_mat_omega.out
          -covmcmc -outprefix covmcmc
```

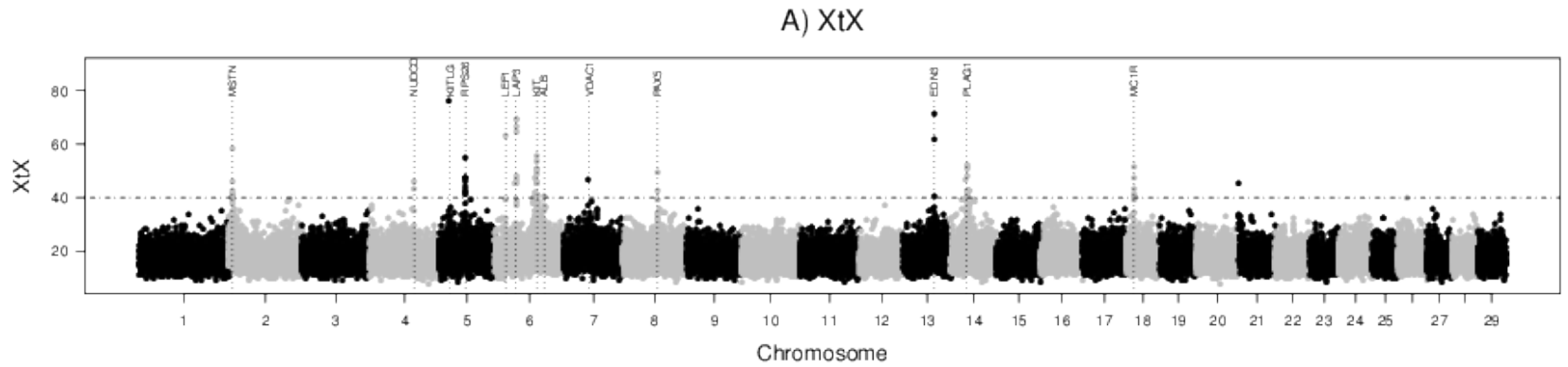
- Running the AUX covariate model (no SNP spatial dependency)

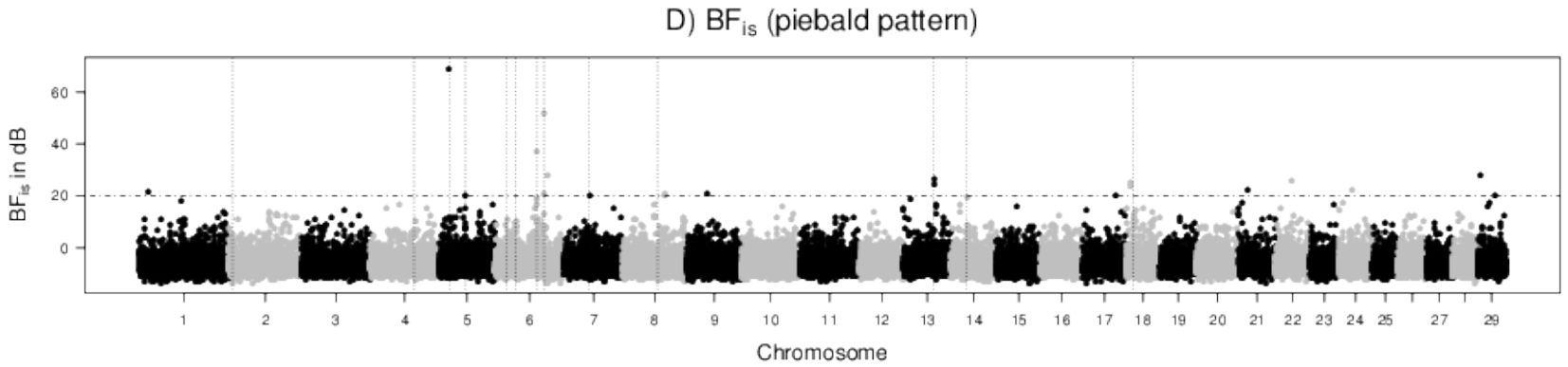
```
i_baypass -npop 18 -gfile bta.geno -efile trait.dat -omegafile covis_mat_omega.out
          -auxmodel -outprefix covaux
```

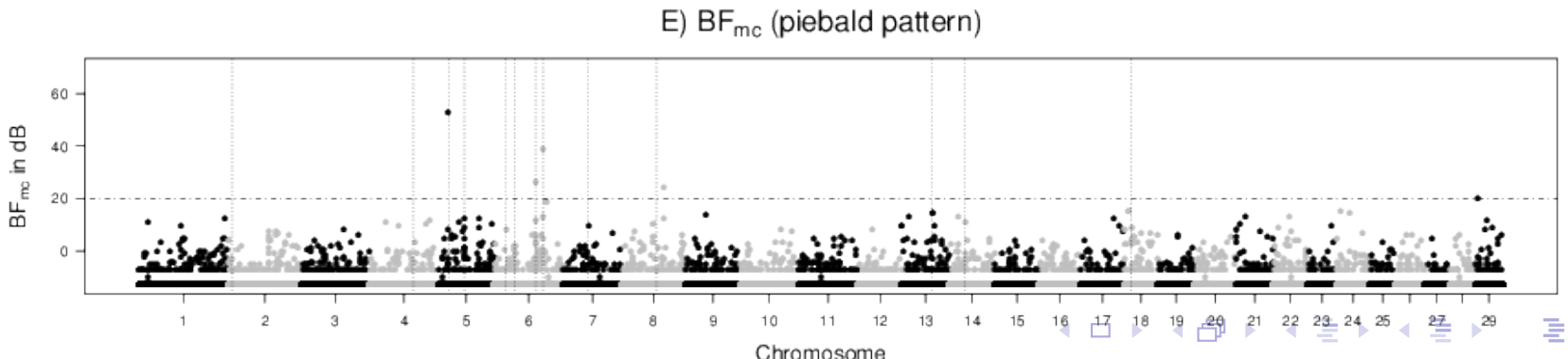
- Running the AUX covariate model with spatial SNP dependency *(SNPs are assumed to be ordered in the file bta.geno)*

```
i_baypass -npop 18 -gfile bta.geno -efile trait.dat -omegafile covis_mat_omega.out
          -auxmodel -isingbeta 1.0 -outprefix covauxisb1
```

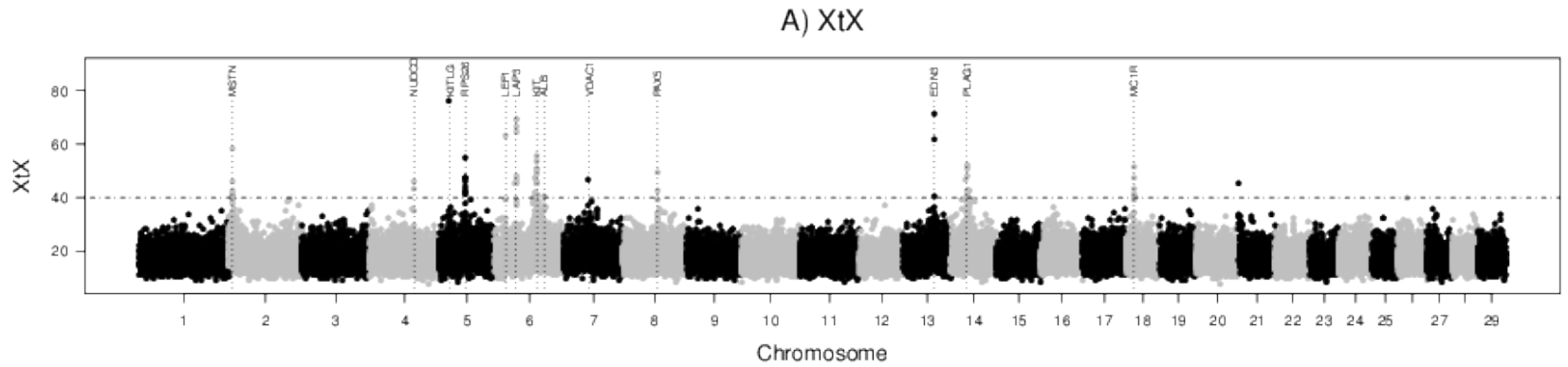
Results (piebald pattern)

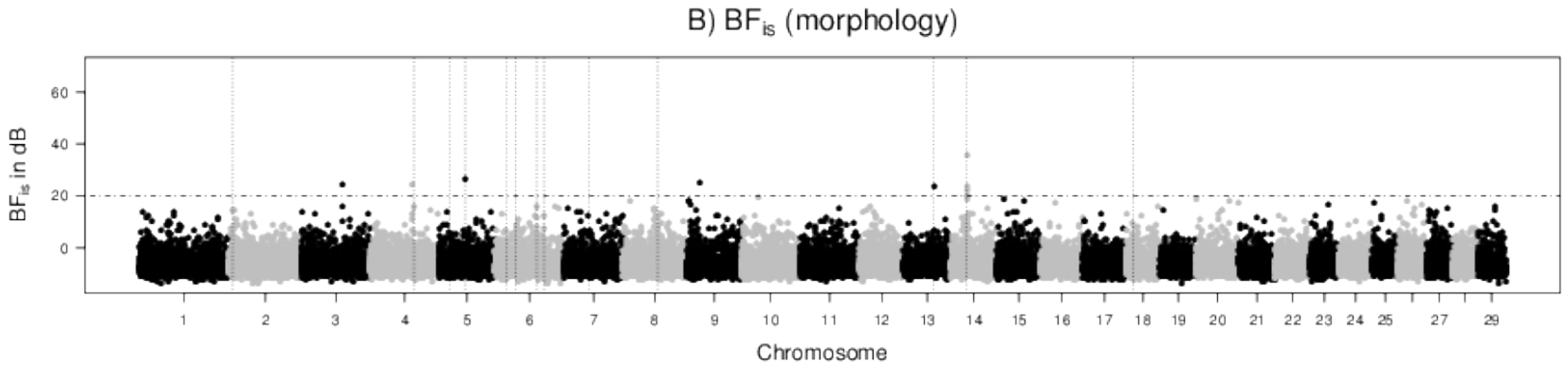


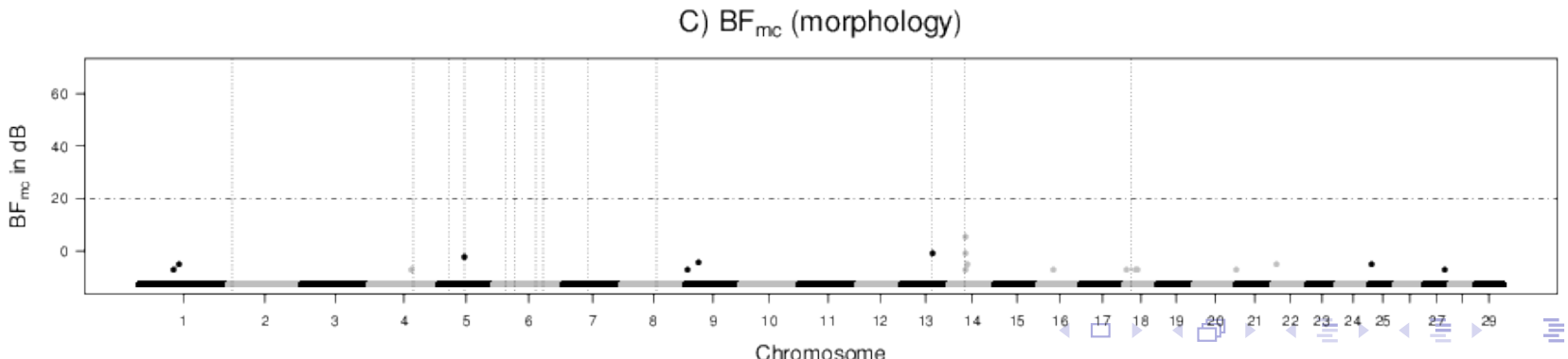




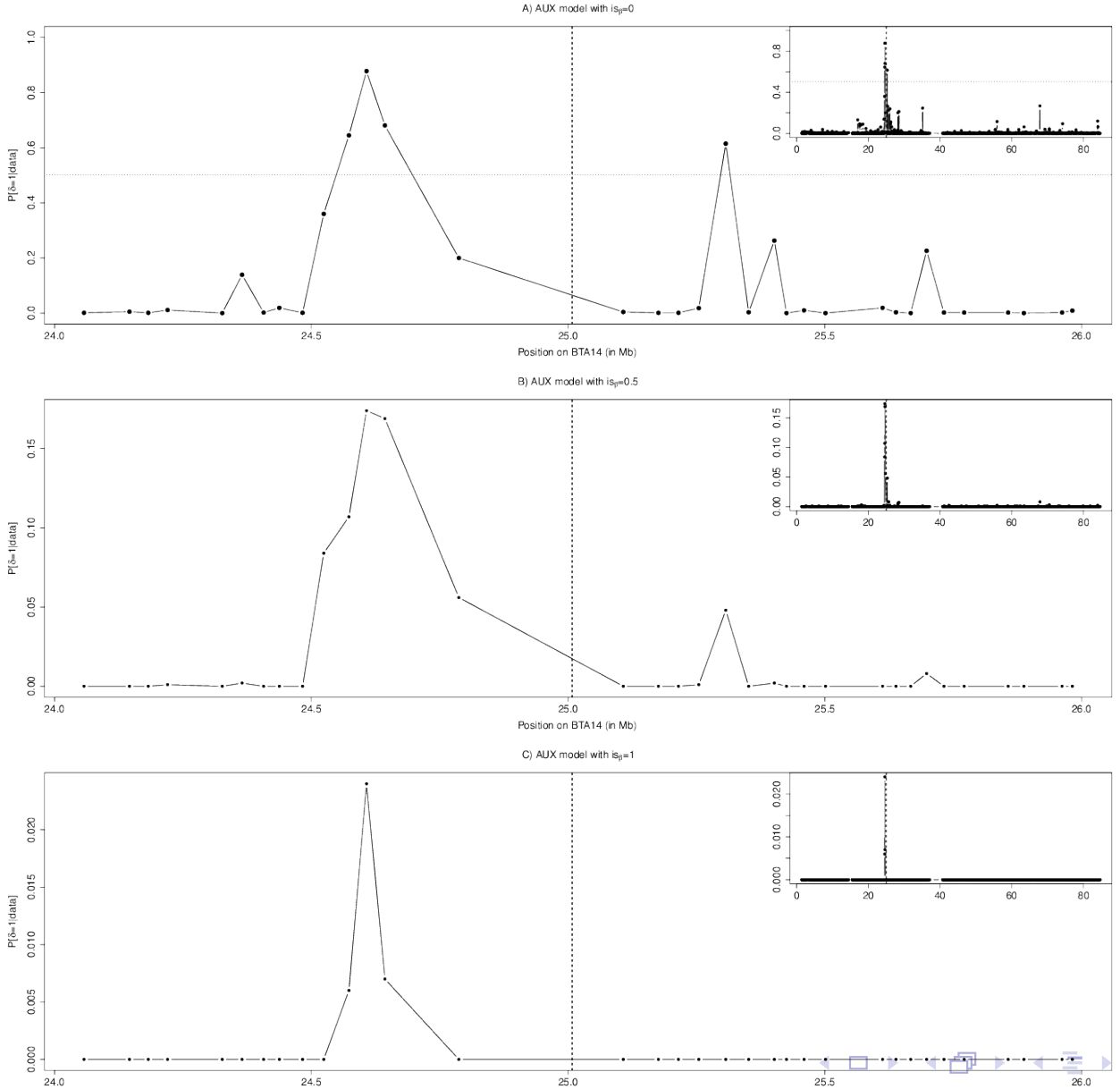
Results (morphology)







BTA14 (PLAG1 region)



Littorina example : command lines

- Running with default parameters (Imp. Sampling estimates) :

```
i_baypass -npop 12 -gfile lsa.geno -efile lsa.ecotype
          -poolsizefile lsa.poolsize -outprefix lsacovis
```

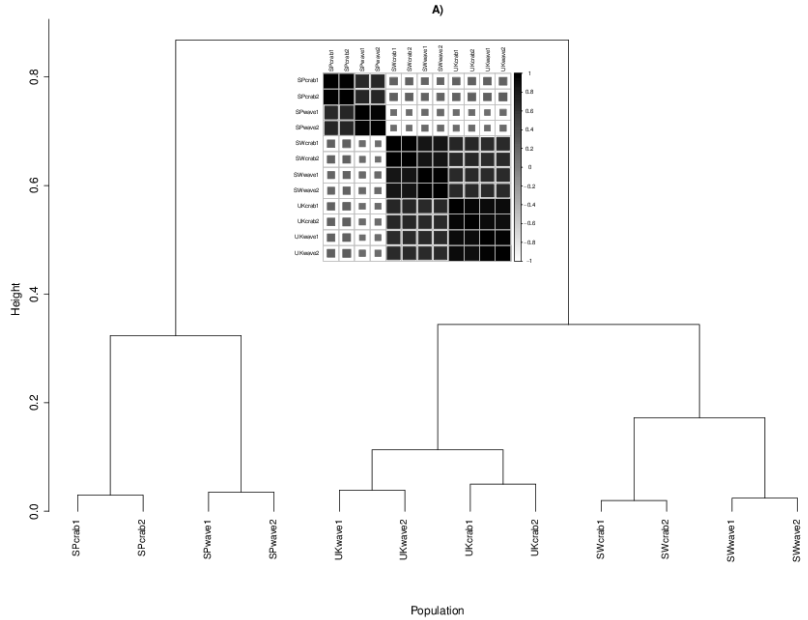
- Running the covariate model (MCMC mode)

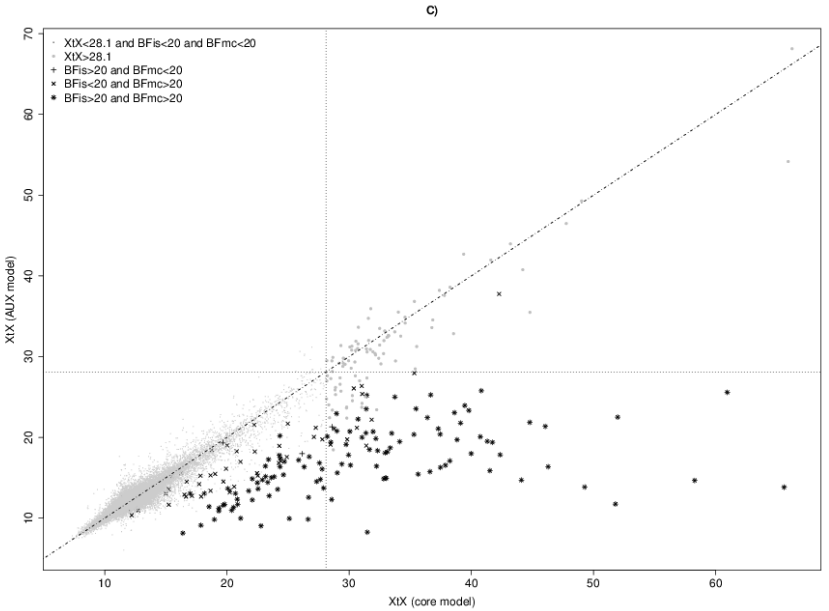
```
i_baypass -npop 12 -gfile lsa.geno -efile lsa.ecotype -poolsizefile lsa.poolsize
          -omegafile covis_mat_omega.out -covmcmc -outprefix lsacovmc
```

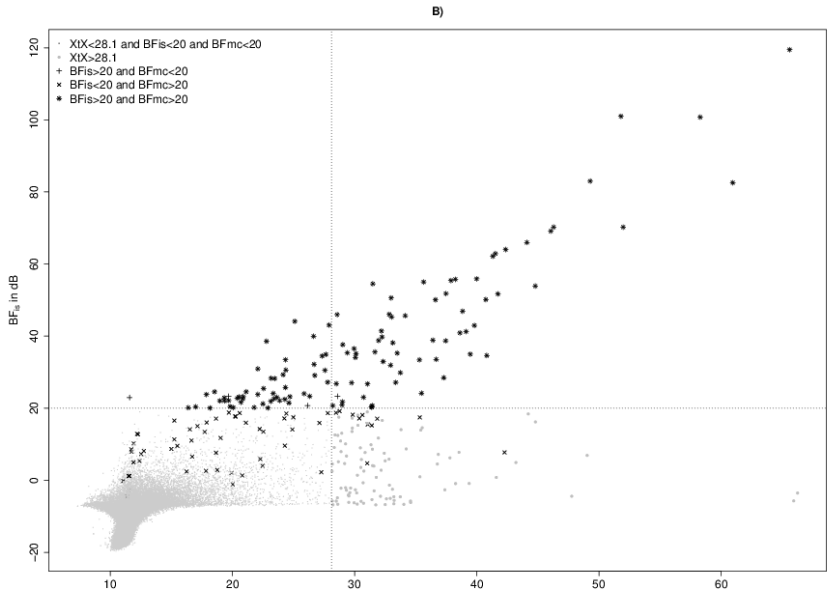
- Running the AUX covariate model (no SNP spatial dependency)

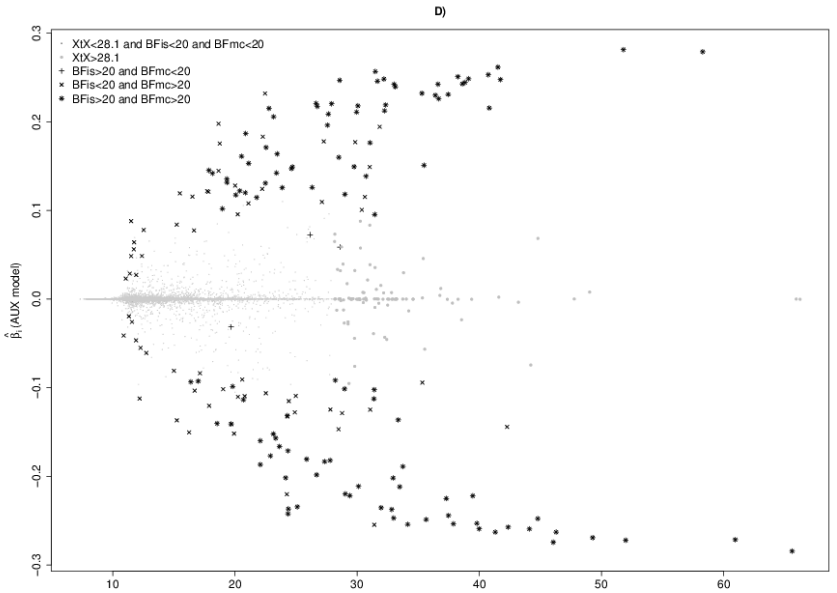
```
i_baypass -npop 12 -gfile lsa.geno -efile lsa.ecotype -poolsizefile lsa.poolsize
          -omegafile covis_mat_omega.out -auxmodel -outprefix lsacovaux
```

Results


A)
Height
Population
SPcrab1
SPcrab2
SPwave1
SPwave2
UKwave1
UKwave2
UKcrab1
UKcrab2
SWcrab1
SWcrab2
SWwave1
SWwave2


C)
XtX<28.1 and BFis<20 and BFmc<20
XtX>28.1
BFis>20 and BFmc<20
BFis<20 and BFmc>20
BFis>20 and BFmc>20
XtX (AUX model)
XtX (core model)


B)
XtX<28.1 and BFis<20 and BFmc<20
XtX>28.1
BFis>20 and BFmc<20
BFis<20 and BFmc>20
BFis>20 and BFmc>20
BFis in dB


D)
XtX<28.1 and BFis<20 and BFmc<20
XtX>28.1
BFis>20 and BFmc<20
BFis<20 and BFmc>20
BFis>20 and BFmc>20
β̂i (AUX model)

Key features of BayPass

- Accurate estimation of $\mathbf{\Omega}$ ($\Leftrightarrow$ account for pop. demographic history) :
  - without any prior information
  - $\Leftrightarrow$ improved estimation of the related statistics and decision criteria
- Implementation of complementary approaches :
  - covariate free (indirect) approaches ($\mathrm{X^tX}$ for genome scan for adaptive differentiation) with calibration procedure
  - association with pop. specific covariates
    - Different decision criteria (eBPis, BFis, eBPmc and BFmc)
    - the AUX model deals with multiple testing issues and allows (to some extent) to account for spatial dependency of markers (but $\sim$ smoothing approach)
- Flexible :
  - Computationnally efficient (e.g., parallel computing)
  - Accomodate PoolSeq data in a rigorous way

Current developments and Future directions

- Alternative priors on $\boldsymbol{\Omega}$
  - Nicholson prior (diagonal) : relevant for E&R experiments (incl. information on ancestral allele frequency i.e. in the base population) : *done but not yet released*
  - Other prior spatially or phylogenetically explicit
- Accomadating multi-allelic markers (e.g. haplotypes or sequences)
- Integrating over the uncertainties in population-specific covariates (Gautier, in prep.)

Downloading (and citing) the program