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

Mathieu Gautier

UMR INRA/CIRAD/IRD/SupAgro CBGP

6 septembre 2017

BayPass Overview

Comparisons of SNP allele frequencies across populations

- GSD : Genome-Scan for extremely Differentiated SNPs ("outliers")
- pGWAS : Genome-Wide Association analyses with population-specific covariates (e.g., Ecological Association)

Data

- For GSD/pGWAS : Population Allele Count (Read Count in the Pool–Seq mode)
- For pGWAS : Population Covariables (e.g., environmental variables, quantitative or categorical phenotypic characteristics)
- For pGWAS : Optional : : Map order to (roughly) account for LD via an Ising model in the pGWAS

Bayesian Hierarchical Model

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 : $\boldsymbol{\alpha_i^{\star}} = \left\{\alpha_{ij}^{\star}\right\}_{(1..J)}$

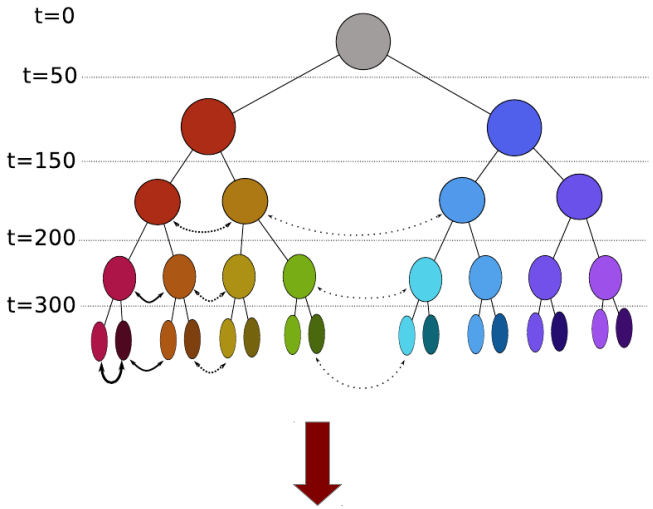
$$\boldsymbol{\alpha_i^{\star}} \sim \mathrm{N}_J\left(\pi_i \mathbb{1}; \pi_i(1-\pi_i)\boldsymbol{\Omega}\right)$$

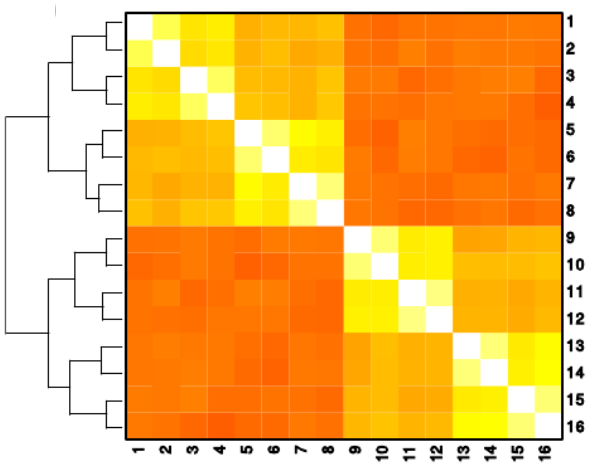
  - $\mathbb{1}$: identity vector of length J (number of pops.)
  - π_i : across pop. frequency (might be interpreted as the "ancestral" ref. allele frequency)
  - $\boldsymbol{\Omega}$: scaled covariance ($J \times J$) matrix of pop. allele frequency

$\boldsymbol{\Omega}$ captures the covariance structure of allele frequencies that originates from the population shared history (global effect of the demography)

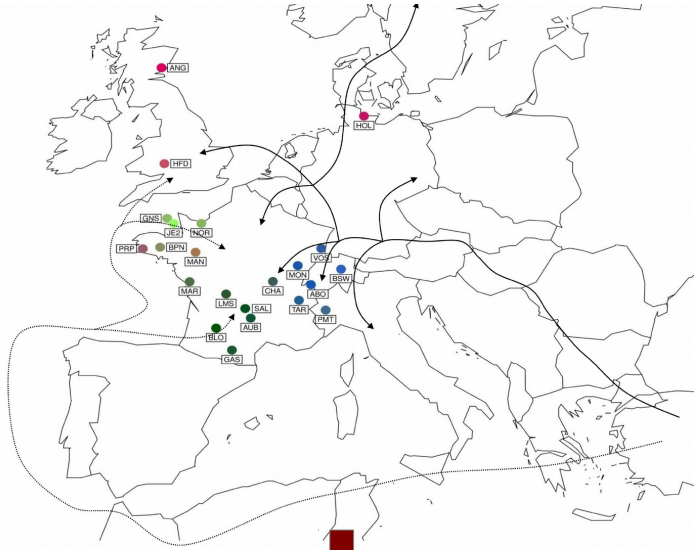
Example of realized Ω

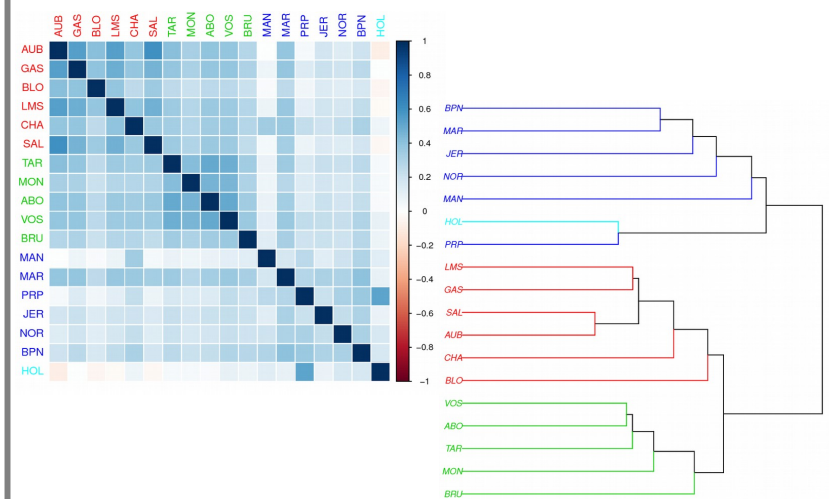
Simulated Data: Hierarchical model with migration
(from de Villemereuil et al., 2014)



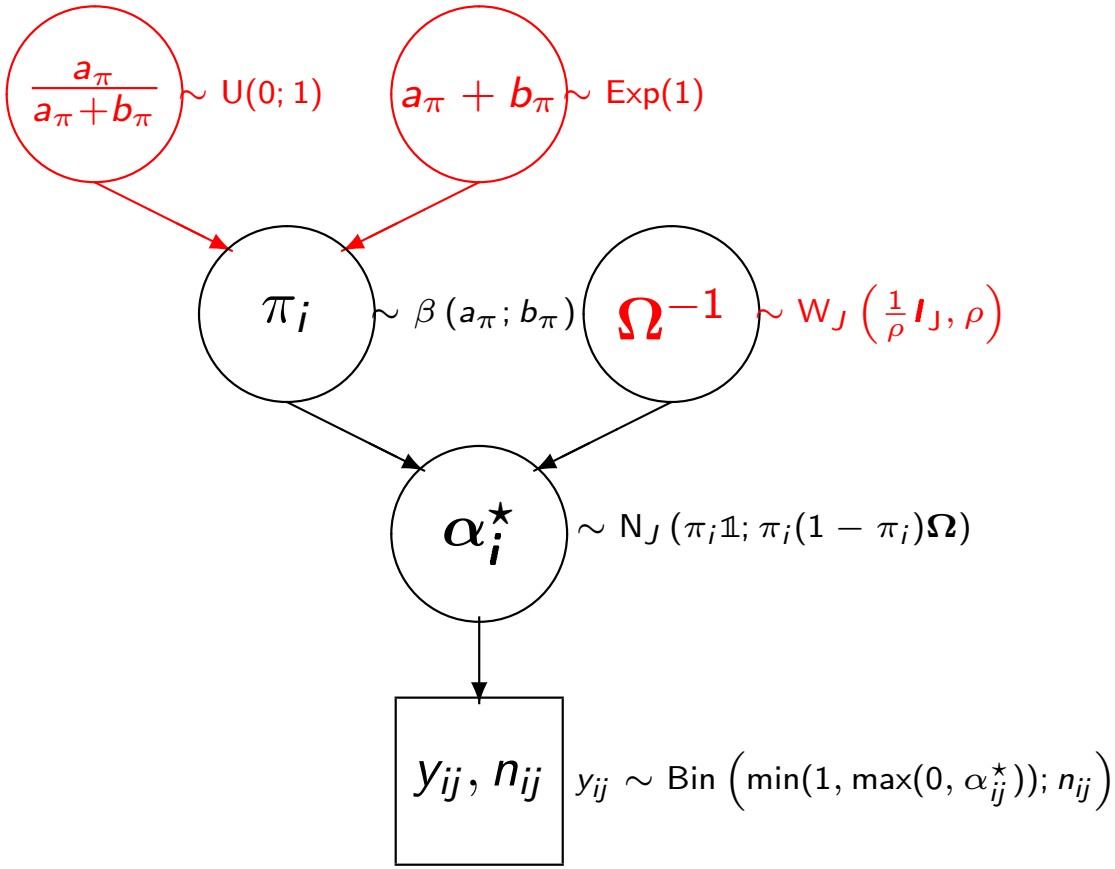


Real Data: 18 cattle breeds (42,046 SNPs)
(from Gautier et al., 2010)





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 $\rho = J$)

Identifying outliers with the XtX statistic

Definition (Guenther and Coop, 2013) and computation

- Let $\mathbf{X_i} \simeq$ vector of scaled pop. allele freq. ($\mathbf{X}_i = \mathbf{\Gamma}^{-1} \frac{\boldsymbol{\alpha}_i^\star - \pi_i}{\sqrt{\pi_i(1-\pi_i)}}$ with $\mathbf{\Omega} = \mathbf{\Gamma}^{-1}\mathbf{\Gamma}$)
- $\mathrm{X^tX}_i = \mathrm{Var}\left(\mathbf{X_i}\right) = \frac{(\boldsymbol{\alpha}_i^\star - \pi_i)\mathbf{\Omega}^{-1}(\boldsymbol{\alpha}_i^\star - \pi_i)}{\pi_i(1-\pi_i)}$ ($\simeq$ *FLK* by Bonhomme et al. (2010))

Calibration (How extreme to be outlier ?)

- From the model : $\mathrm{X^tX} \sim \chi^2\left(J\right)$ *(i.e.* $\mathrm{E}\left(\mathrm{X}^t\mathrm{X}\right) = \frac{1}{2}\mathrm{Var}\left(\mathrm{X}^t\mathrm{X}\right) = J$*)*
- From the MCMC samples : $\mathrm{E}\left(\widehat{\mathrm{X^tX}}\right) = J$ but $\mathrm{Var}\left(\widehat{\mathrm{X^tX}}\right) \ll 2J$
- Calibration by analysing PODs generated under the inference model
  - $\mathbf{\Omega}^{\mathrm{sim}} = \widehat{\mathbf{\Omega}}$, $a_\pi^{\mathrm{sim}} = \widehat{a_\pi}$ and $b_\pi^{\mathrm{sim}} = \widehat{b_\pi}$ (see *simulate_baypass()* R function)
- Normalizing transformation of the $\widehat{\mathrm{X^tX}}$ (NEW and not extensively validated)
  - Based on Wilson–Hilferty transform of rescaled $\widehat{\mathrm{X^tX}}$ (see *standardize.xtx()* R function)

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)

GWAS with population–specific covariates

(Very) brief overview of approaches

- 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

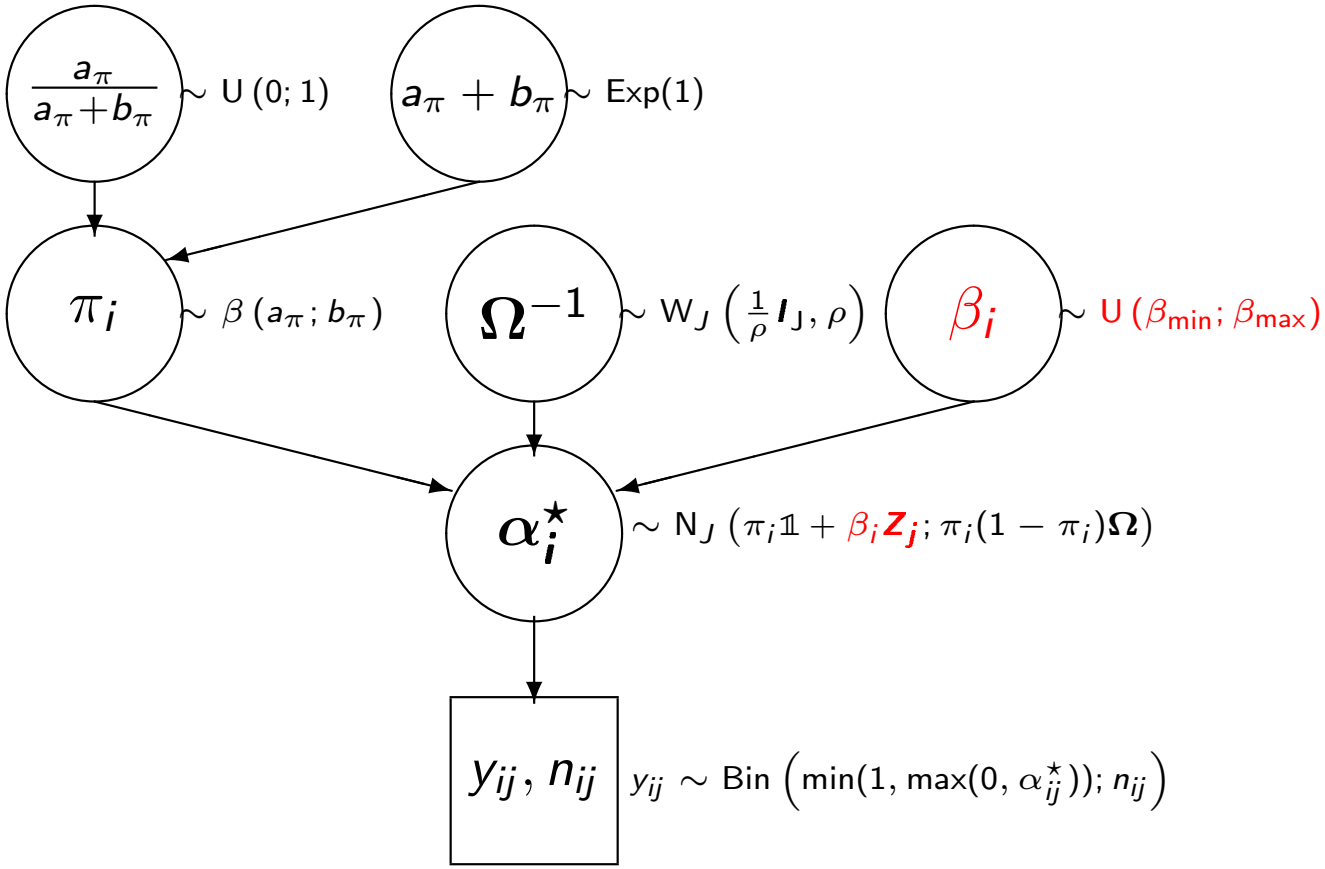
- Bayescenv (de Villemereuil et al., 2015) : association between residuals of a logistic regression of marker and pop specific F_{ST} (with marker and population specific effects) and the environmental variable

  $\Rightarrow$ 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.

The BayPass "standard" covariate model



- Similar to BayEnv model (Coop et al., 2010) with additional extensions
  - Priors on a_π, b_π and Ω^{-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)

Estimating the β_i's and assessing association significance

- A) Via Importance Sampling (only requires samples drawn under the core model)
  - Direct 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. $\beta_i = 0$)
  - $\widehat{\mu\left(\beta_i\right)}$ and $\widehat{\sigma\left(\beta_i\right)} \Rightarrow \text{eBP}_{\text{is}} = -\log_{10}\left(1 - 2\left|0.5 - \Phi\left(\frac{\widehat{\mu(\beta_i)}}{\widehat{\sigma(\beta_i)}}\right)\right|\right)$
- B) Via MCMC (`covmcmc` option)
  - Sampling from the posterior distribution of the β_i's via MCMC
  - Posterior $\widehat{\mu\left(\beta_i\right)}$ and $\widehat{\sigma\left(\beta_i\right)} \Rightarrow \text{eBP}_{\text{mc}}$
- C) Via MCMC with the aux. variable model (`auxmodel` option)
  - $\beta_i = \delta_i\beta_i^\star$:$\delta_i = 1$ ($\delta_i = 0$) if the SNP is (not) associated ($\delta_i \sim \text{Ber}(P)$ allowing to integrate over the unknown prop. P of associated SNPs to deal with multiple testing issues).
  - $\text{BF}_{mc} = \frac{\text{Post. odds}}{\text{Prior odds}} = \frac{\text{P}[\delta_i=1|data]}{[1-\text{P}(\delta_i=1|data)]} \times \frac{1-\mathbb{E}[P]}{\mathbb{E}[P]}$

In practice...

To sample or not to sample the β_i's ? (i.e., IS or MCMC ?)

- When *npop* is small (e.g., $\leq$ 8 and/or pops are highly differentiated), AUX and STD models may be "unstable" (seemingly due to identifiability problems)
  $\Rightarrow$ BF_{is} (or eBP_{is}) should then be preferred.
- Specific recommendation regarding BF_{is} (and eBP_{is})
  - Estimates rely on (Importance Sampling) approximations
  - Check consistency across several (e.g., 3–5) independent runs (`-seed`)
- When data are not limiting, sampling the β_i's should be preferred
  $\Rightarrow$ $\underline{BF_{mc}}$ (`-auxmodel`) or eBP_{mc} (`-covmcmc`)

Decision Rule

- Jeffreys' rule : $15 < BF < 20 \Rightarrow$ *"very strong evidence"* ; $BF > 20 \Rightarrow$ *"decisive evidence"*
- Calibration with PODs (e.g., eBP_{is}, eBP_{mc} or BF_{is})
- $\underline{BF_{mc}}$ accounts for multiple testing issues
  AUX model $\Rightarrow$ Model Averaging : $P[\delta_i = 1|data]$ =Posterior Inclusion Probability (PIP)

French cattle breeds example

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

- $J = 18$ (mostly) French cattle breeds $I = 42,046$ SNPs
- (partial) view of the allele count file : "bta.geno"

```
8 36 11 25 24 36 15 29 12 46 45 47 10 28 26 62 18 26 18 22 16 ....[2x18=36 col.]
13 31 16 20 30 30 19 25 44 14 26 66 15 23 28 60 13 31 8 32 11 ....[2x18=36 col.]
6 38 0 36 1 59 3 41 15 43 9 83 8 30 10 78 2 42 3 37 6 34 4 40 ....[2x18=36 col.]
19 25 25 11 35 25 20 24 13 45 68 24 22 16 47 41 29 15 26 14 22....[2x18=36 col.]
..............................
[42046 rows in total]
```

Covariate file

- Ex. 18 cattle breeds and 2 covariates : Morpho. Score and Piebald 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]
```

- Best Practices
  - Scale the covariables (done by the *-scalecov* option)
  - Use PCA to decorrelate variables (analyze PC's="synthetic" scores)

Estimating Ω and XtX (+ BFis/eBPis if covariate file)

Command Lines (Both lead to the same estimates of XtX and Ω)

- Running with default parameters (XtX only) :

```
i_baypass -npop 18 -gfile bta.geno -outprefix ana_core \
          -nthreads 4 -pilotlength 500 -burnin 2500 > ana_core.log
```

- Running with default parameters (XtX + IS estimates of BF and eBP) :

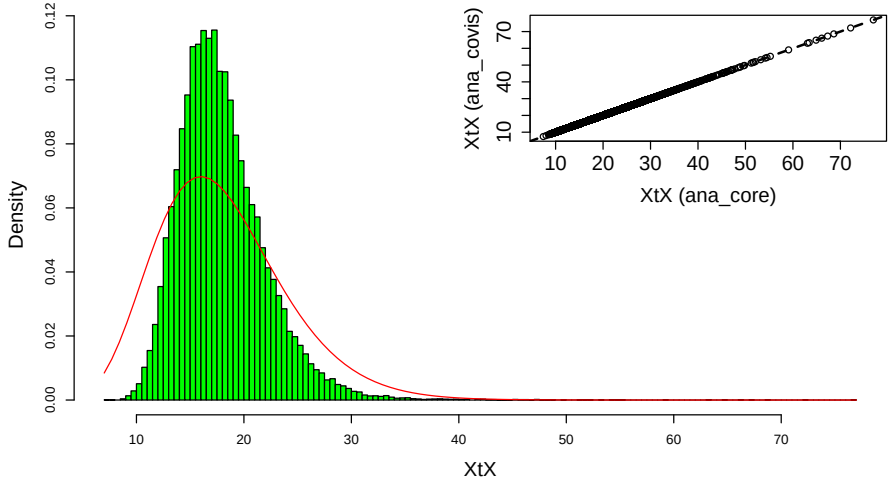
```
i_baypass -npop 18 -gfile bta.geno -efile trait.dat -outprefix ana_covis \
          -nthreads 4 -pilotlength 500 -burnin 2500 > ana_covis.log
```

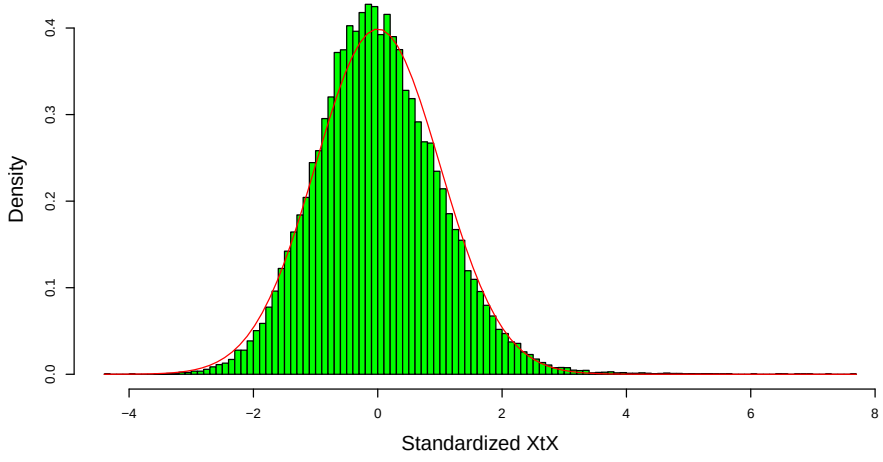
Some output files (with *=ana_core or *=ana_covis)

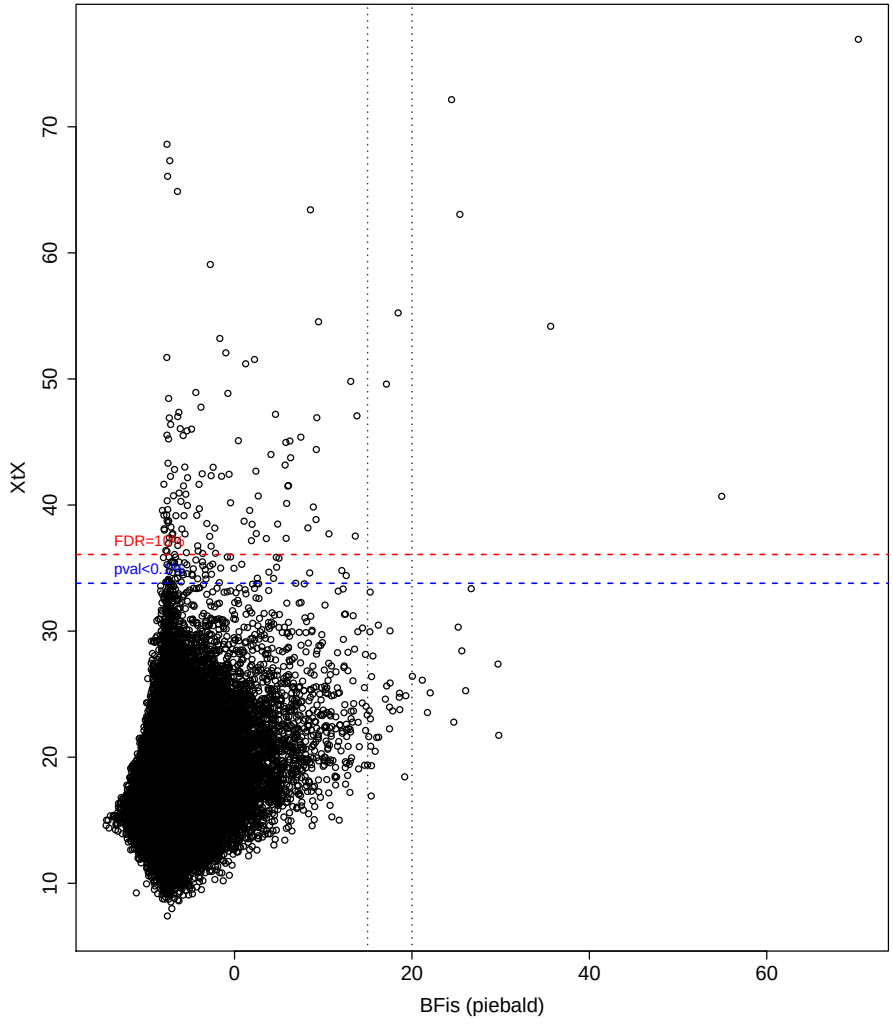
- $\widehat{\Omega}$: `*_mat_omega.out` (and `*_summary_lda_omega.out` for more info.)
- $\widehat{X^tX}$ (and $\widehat{BF_{is}}$) : `*_summary_pi_xtx.out` (`ana_covis_summary_betai_reg.out`)

Calibrate the XtX after normalizing transform (in R)

```
xtx=read.table("ana_covis_summary_pi_xtx.out",h=T)$M_XtX
xtx.std=standardize.xtx(xtx,npop=18)  ; thr.pval1ppm=min(xtx[xtx.std$xtx.pval.pos<0.001])
library(qvalue)
xtx.qval=qvalue(xtx.std$xtx.pval.pos) ; thr.FDR10pcent=min(xtx[xtx.qval$qvalue<0.1])
```


Density
XtX
XtX (ana_covis)
XtX (ana_core)


Density
Standardized XtX


XtX
BFis (piebald)
FDR=10%
pval<0.1%

Calibration of the XtX (and BFis/eBPis if covariate file) with PODs

Generate a POD with R (e.g., 100,000 SNPs)

```
source("YOUR_PATH/baypass/utils/baypass_utils.R")
om.bta=as.matrix(read.table("ana_core_mat_omega.out"))
pi.beta.coef=read.table("ana_core_summary_beta_params.out",h=T)$Mean
geno.data<-geno2YN("bta.geno")
simu.bta<-simulate.baypass(omega.mat=om.bta,nsnp=100000,sample.size=geno.data$NN,
                           beta.pi=pi.beta.coef,pi.maf=0,suffix="bta.pods")
```

Analyze the POD (e.g., with covariate file)

```
i_baypass -npop 18 -gfile G.bta.pods -efile trait.dat -outprefix ana_pod \
-nthreads 4 -pilotlength 500 -burnin 2500 > ana_pod.log
```

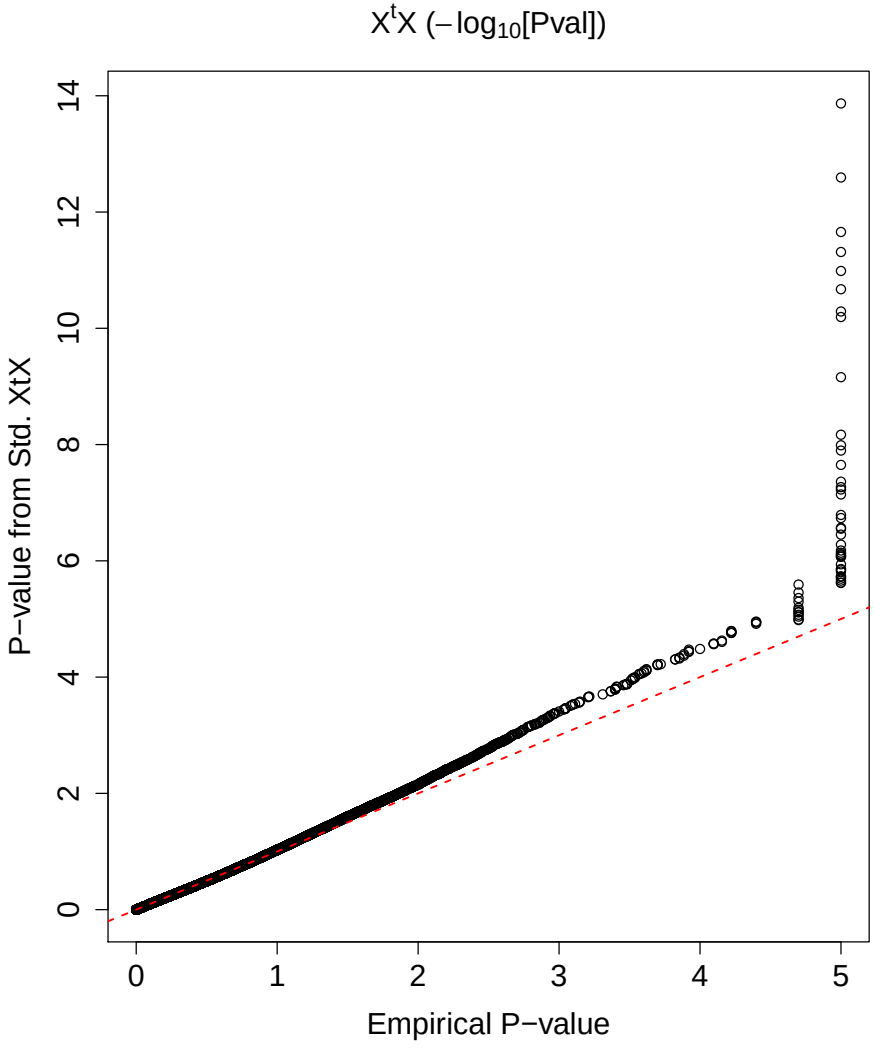
Calibrate the statistics (e.g., with R)

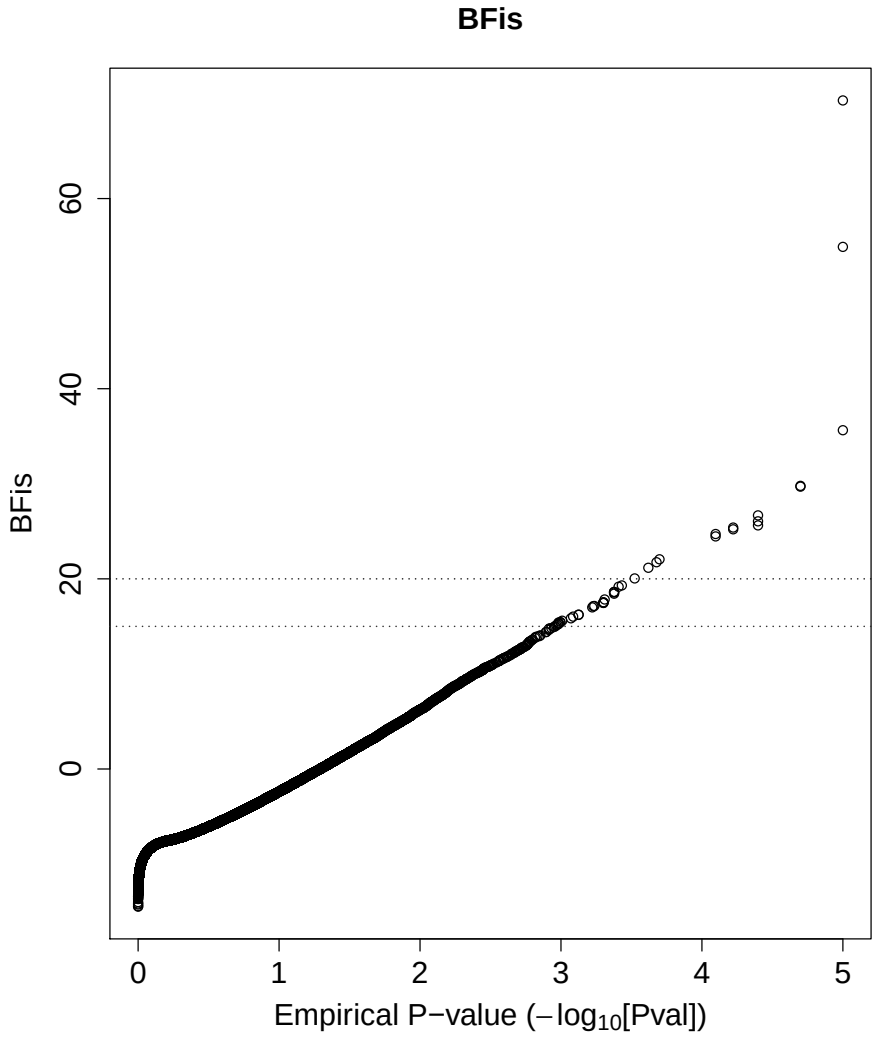
- For the XtX

```
xtx.pods=read.table("ana_pod_summary_pi_xtx.out",h=T)$M_XtX
xtx.threshold=quantile(xtx.pods,probs=c(0.0001,0.001,0.5,0.999,0.9999))
xtx.empval=empPvals(stat = xtx,stat0=xtx.pods) #With qvalue package
```

- For the BFis

```
res.pods=read.table("ana_pod_summary_betai_reg.out",h=T)
bfis.pods=res.pods$BF.dB.[res.pods$COVARIABLE==1] #for the second covariable
bfis.threshold=quantile(bfis.pods,probs=c(0.999,0.9999))
bfis.empval=empPvals(stat = bfis,stat0=bfis.pods) #With qvalue package
```


X^tX ($-\log_{10}$[Pval])
P−value from Std. XtX
Empirical P−value
0
2
4
6
8
10
12
14
0
1
2
3
4
5


BFis
BFis
Empirical P−value ($-\log_{10}$[Pval])
0
20
40
60
0
1
2
3
4
5

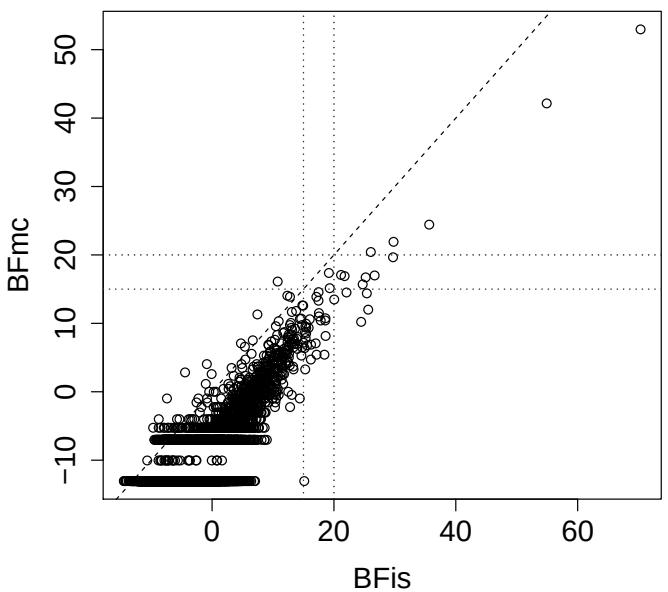
Estimating BFmc

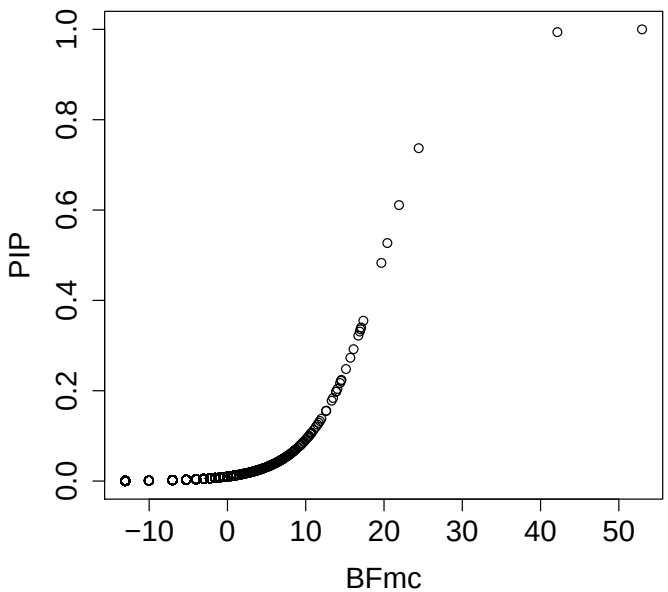
- Running the AUX model :

```
i_baypass -npop 18 -gfile bta.geno -efile trait.dat \
      -omegafile ana_covis_mat_omega.out -auxmodel  \
      -nthreads 4 -pilotlength 500 -burnin 2500 -outprefix ana_covaux > ana_covaux.log
```

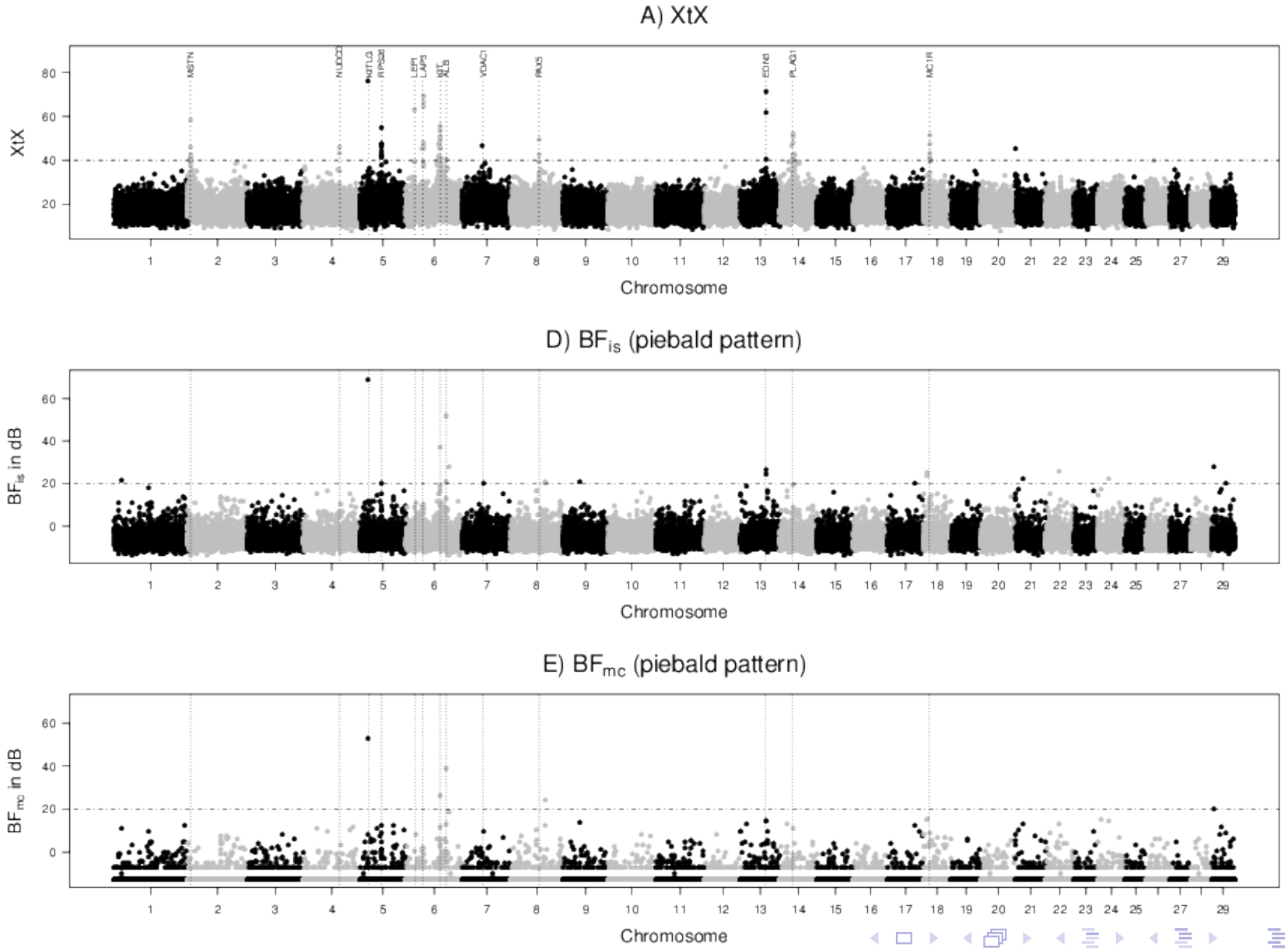
- Plotting results

```
res.aux=read.table("ana_covaux_summary_betai.out",h=T)
bfmc=res.aux$BF.dB.[res.aux$COVARIABLE==1] #for the second covariable
pip=res.aux$M_Delta[res.aux$COVARIABLE==2] #for the second covariable
```





Manhattan plots (piebald pattern)



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 $\sim$ allows 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
- For more details :
  - The most important option : *-help*
  - The manual : `http://www1.montpellier.inra.fr/CBGP/software/baypass/`