# Bayesian hierarchical models for GEA and prediction on a population scale

Mathieu Gautier

UMR INRA/CIRAD/IRD/SupAgro CBGP

$21^{st}$ September 2022

# Introduction

## General assumption

- Population allele freq. at loci underlying local adaptation are expected to co-vary with fitness-related traits or selective pressure intensity
- but see Lotterhos (2022) for a (simulation-based) critical evaluation

## Genome-wide association with population-specific covariates

- Modeling the relationship between genetic diversity (marker allele frequency variation) and covariates of interest across several (differentiated) populations :
  - insights into the genetic architecture of adaptive traits
  - predict covariate value from genomic information

## Different covariates of interest

- Environmental (e.g., climate, host plant, etc.) $\Rightarrow$ GEA
- Phenotypic (e.g., mean height, mean weight, coat color) $\Rightarrow$ 'pGWAS'

# Demographic history : an important confounding factor

## 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

- Demographic Factors (genetic drift, gene flow) expected to be common to all loci
  $\Longrightarrow$ Global Effect responsible for a correlation structure of pop. allele frequencies
- Selection (mutation and recombination) expected to vary across loci
  $\Longrightarrow$ Local Effect

Associating allele freq. differences with variation in environment or trait values among population requires accounting for possibly confounding demographic effects

# GEA/pGWAS model

## Historically

- covariate = environmental variables ⇒ proxies for ecological pressure

## Various approaches

- SAM (Joost et al., 2007) : univariate logistic regression of pop. all. freq. with the covariate ⇒ does not account for neutral all. freq. covariance
- Bayescenv (de Villemereuil et al., 2015) : association between the residuals of a logistic regression of marker and pop–specific $F_{ST}$ (with marker and population specific effects) and the covariate ⇒ 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 (Gautier, 2015) : robustly account for neutral all. freq. covariance and treat covariate as a 'fixed' effect.

See *de Villemereuil et al. (2014)* for a comparison under realistic simulation scenarios

! issues in Bayenv2 prog. penalized the Bayenv/BayPass model ; see Gautier (2015)

# The BayPass core model

- Central assumption : multivariate Gaussian distribution for population allele frequencies (bi-allelic) SNPs introduced by Coop et al. (2010) (extends the univariate 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 $\alpha_{ij}$ 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.)
  - $\pi_i$ : across pop. frequency (might be interpreted as the "ancestral" ref. allele frequency)
  - $\boldsymbol{\Omega}$ : scaled covariance ($J \times J$) matrix of pop. allele frequencies

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

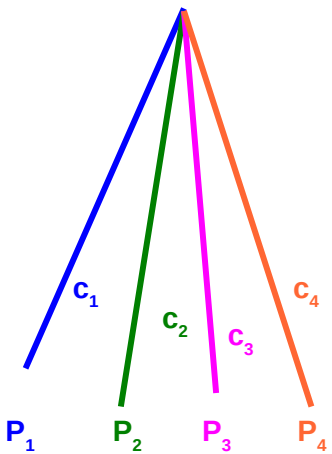
# Demographic interpretation of Ω

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

**Bifurcating tree**

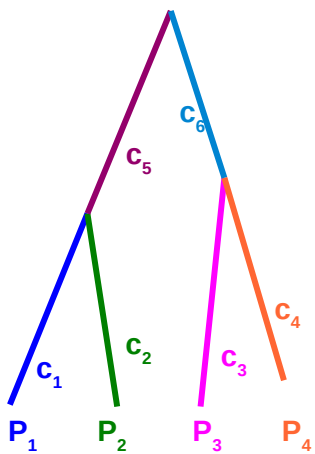
**Admixture graph**

*e.g., Patterson et al. (2012) (**ADMIXTOOLS**) ; Pickrell and Pritchard (2012) (**TreeMix**) ; Gautier et al. (2021) (**poolfstat**)*



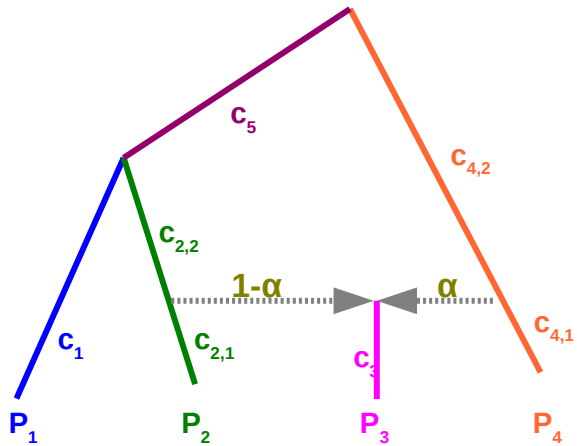

| | | | | |
|---|---|---|---|---|
| $P_1$ | $c_1$ | 0 | 0 | 0 |
| $P_2$ | 0 | $c_2$ | 0 | 0 |
| $P_3$ | 0 | 0 | $c_3$ | 0 |
| $P_4$ | 0 | 0 | 0 | $c_4$ |

**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**

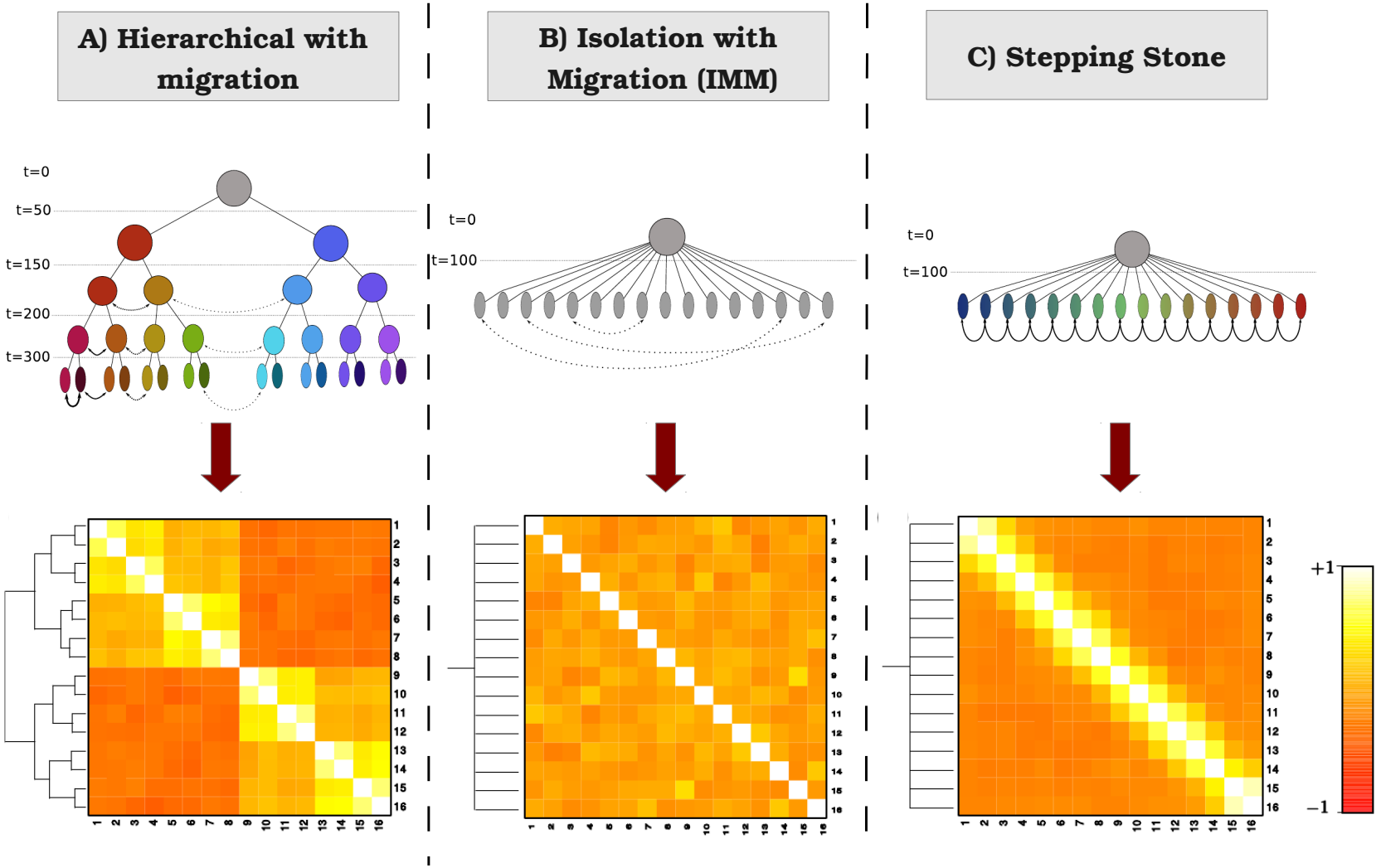



| | | | | |
|---|---|---|---|---|
| $P_1$ | $c_1+c_5$ | $c_5$ | $(1-\alpha)^2c_5$ | 0 |
| $P_2$ | $c_5$ | $c_{2,1}+c_{2,2}+c_5$ | $(1-\alpha)^2(c_{2,2}+c_5)$ | 0 |
| $P_3$ | $(1-\alpha)^2c_5$ | $(1-\alpha)^2(c_{2,2}+c_5)$ | $c_3+\alpha^2c_{4,2}+(1-\alpha)^2(c_{2,2}+c_5)$ | $\alpha^2c_{4,2}$ |
| $P_4$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

**Ω**

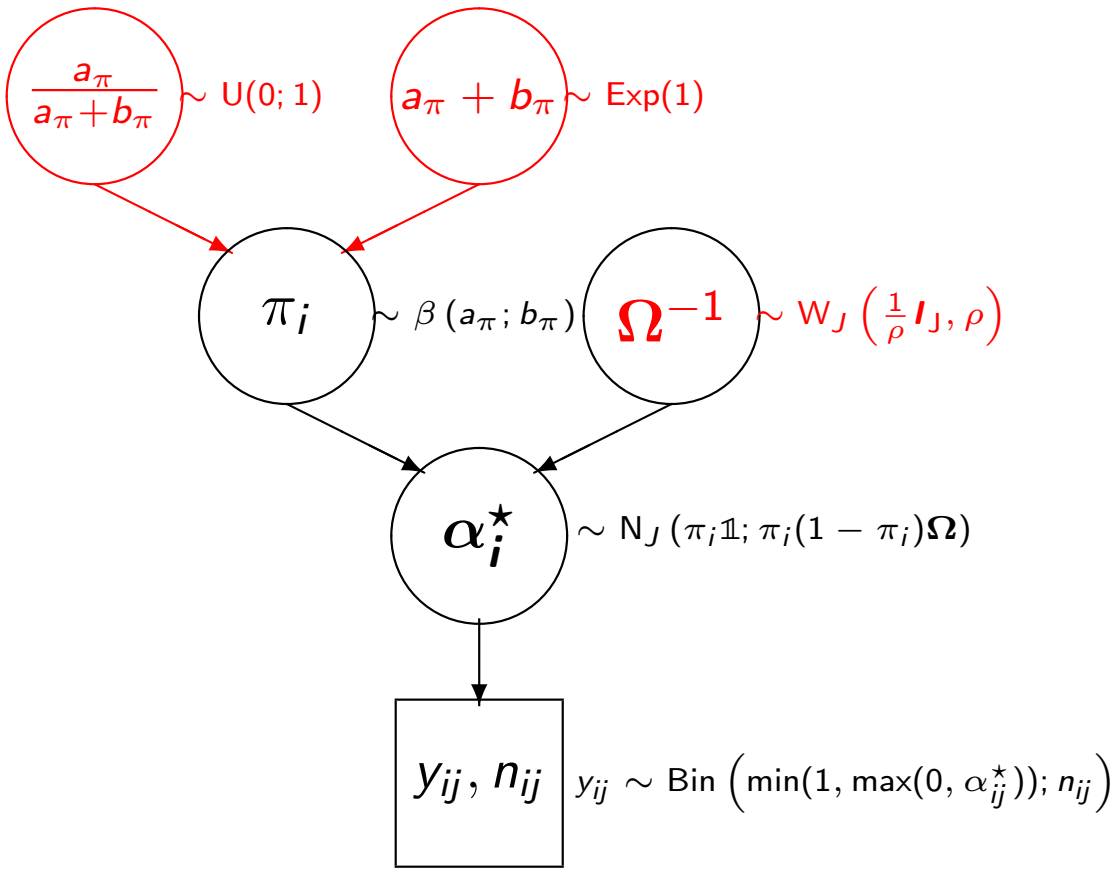
$$*c_j \simeq 1-\left(1-\frac{1}{2N_j}\right)^{t_j} \simeq \frac{t_j}{2N_j}$$

# 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 $\rho_{ij}=\omega_{ij}/(\omega_{ii}\omega_{jj})^{1/2}$

# Estimating of $\Omega$ under a Bayesian hierarchical model




- Joint estimation of the (unobserved) $\pi_i$'s and $\Omega$
- Robust to sampling bias *(e.g., unbalanced population origins, sample sizes, missing data, ascertainement bias)*
- Versatile : deal with read count data (Pool–Seq) *(by integrating over unobs. allele count)* or individual Genotype Likelihoods (from low/medium Ind-Seq WGS) or even combination of various data types (new to future version 3.0)

# Parameter estimation with a (MH 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 *(uniform proposals)*
  - the ($I$ SNPs $\times$ $J$ pops) $\alpha_{ij}^{\star}$'s *(Gaussian proposals)*
  - the matrix $\mathbf{\Omega}$ (actually $\mathbf{\Omega}^{-1}$) *(Gibbs update)*
  - the ($I$ SNPs) $\pi_i$'s *(uniform proposals)*
  - $a_\pi$ and $b_\pi$ (actually $\frac{a_\pi}{a_\pi+b_\pi}$ and $a_\pi + b_\pi$) *(uniform proposals)*
- A typical run consists of :
  - Several Pilot runs to adjust parameters of the proposals (e.g. targeted accept. rates between 0.25 and 0.4) : *e.g.* $20 \times 500$ *iterations*
  - A Burn-in period (to achieve stationary distributions) : *e.g. 5,000 iterations*
  - Parameters Sampling with thinning (to reduce auto-correlations) : *e.g.* $20 \times 1,000$ *iterations*
- The BayPass implementation
  - Coded in Fortran language with a flexible parametrization
  - Sampler extensively checked (simulated data + independant BUGS implementation (e.g., correct for implementation issues leading to inaccurate results with Bayenv2)
  - Reasonable computational times (+parallelization : far more efficient since v2.3) :
    $\sim 2.5h$ *(resp.* $< 1h$*) to analyze 18 pops* $\times$ *40,000 SNPs on 1 (resp. 4) CPU (*`-nthreads` *option)*

# Real Life Example : HSA allele count data

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

- $J = 52$ worldwide 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 -gfile hgdp.geno -outprefix corehgdp
```

- Changing some MCMC parameters : e.g. :

```
i_baypass -gfile hgdp.geno -pilotlength 1000 -burnin 1000
```

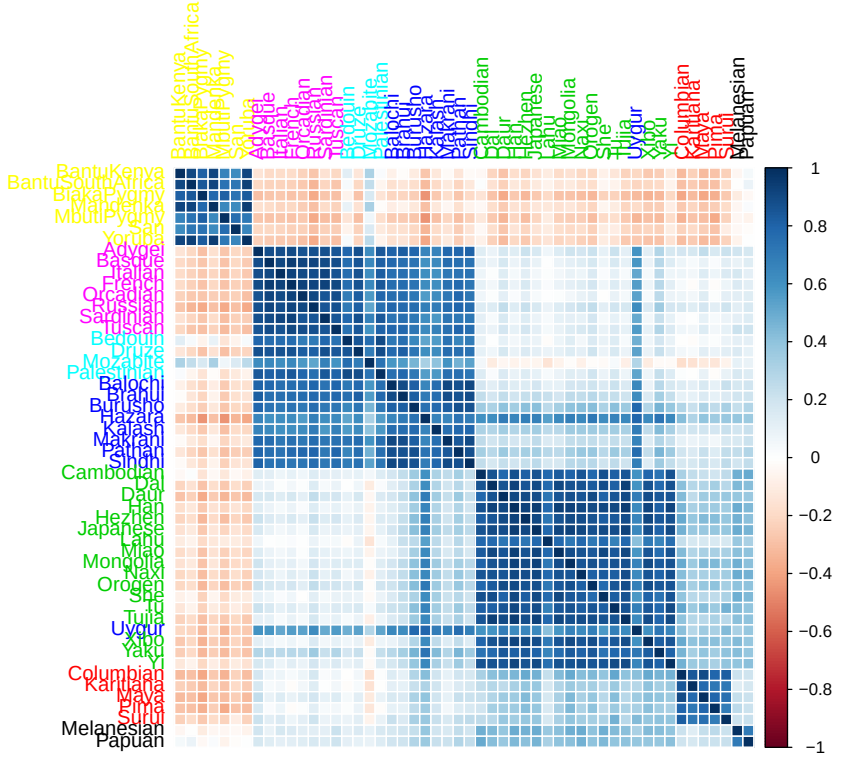
- Changing some modeling options (e.g. set Uniform informative prior on the distribution of $\pi_i$'s)) :

```
i_baypass -setpibetapar -betapiprior 1.0 1.0
```

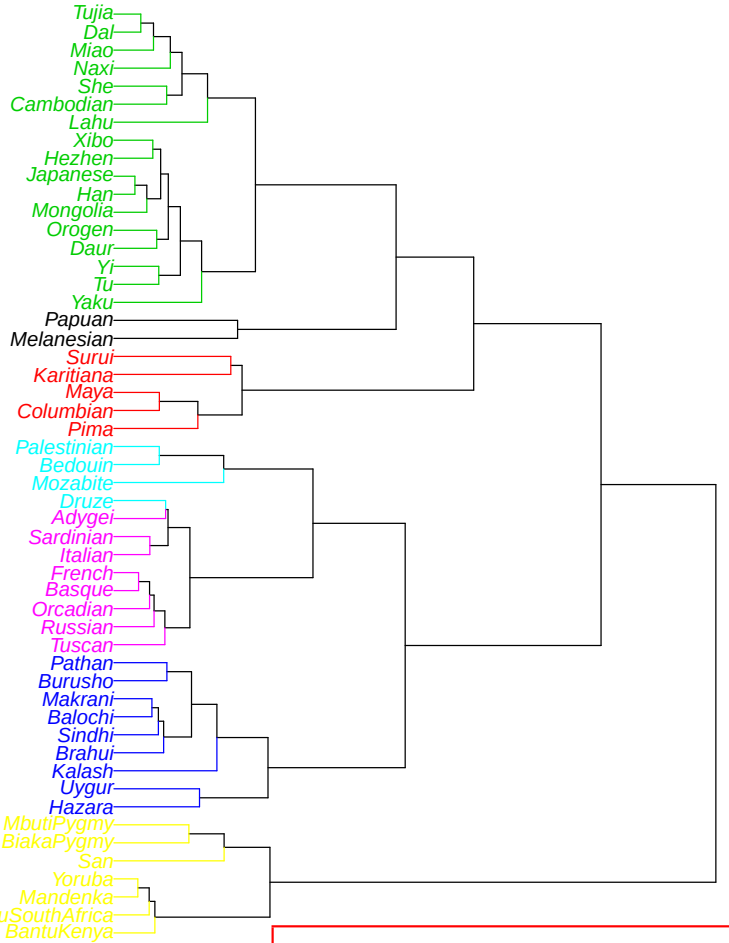
- If your are lost, use the option `-help` (and the manual)
- **N.B. : default MCMC options and model parameters should be appropriate for most (if not all) analyses**

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

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



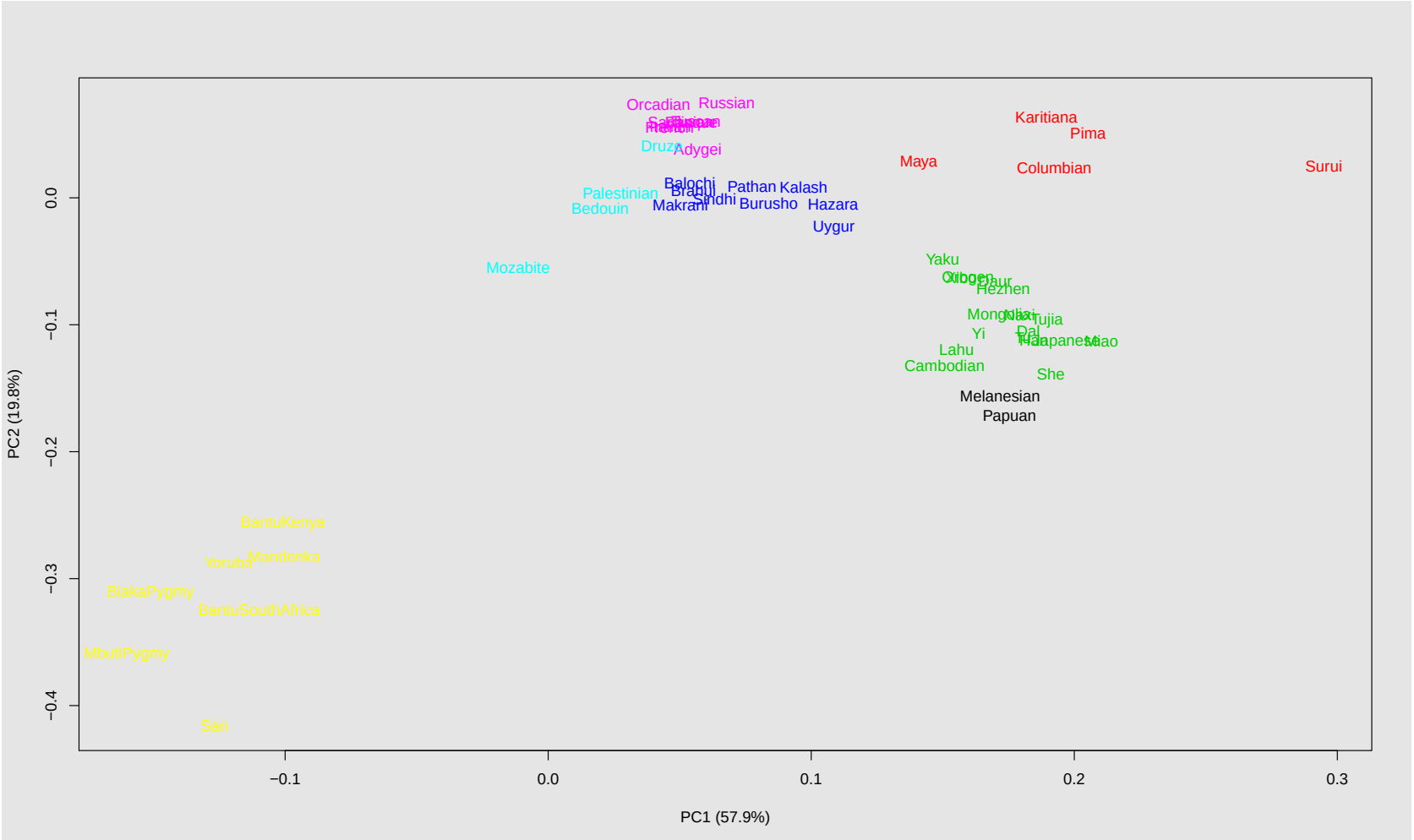

B) Hier. clust. tree based on $\hat{\Omega}^{bpas}_{HSA}$ ($d_{ij}$=1−$\rho_{ij}$)




```
>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
```

# PCA-like representation of $\boldsymbol{\Omega}$ ($= \boldsymbol{U\Lambda U'}$)




```
>source("baypass/utils/baypass_utils.R")
>omega=as.matrix(read.table("corehgdp_mat_omega.out"))
>plot.omega(omega,pop.names=hsa.pops,col=col.pops)
```

# Correcting allele frequencies for demographic history

## The vector $\boldsymbol{X}$ of scaled population allele frequencies

- Definition $\mathbf{X_i} = \{\widetilde{\alpha}_{ij}\}_{1\ldots J} = \mathbf{\Gamma^{-1}} \frac{\boldsymbol{\alpha_i^\star} - \pi_i}{\sqrt{\pi_i(1-\pi_i)}}$ (Guenther and Coop, 2013)
  - $\mathbf{\Omega} = \mathbf{\Gamma^{-1}\Gamma}$ (Cholesky decomposition)
  - e.g., if $\mathbf{\Omega}$ is diagonal (star-shaped pop. tree), $\mathbf{X_i} = \left\{ \frac{\alpha_{ij}^\star - \pi_i}{\sqrt{\omega_{ii}\pi_i(1-\pi_i)}} \right\}$
- $\mathbf{X_i} \simeq$ pop. allele freq. corrected for their joint demographic history
  - if SNP $i$ is "neutral", $\widetilde{\alpha}_{ij} \sim \mathrm{N}(0,1)$ for all populations $j$

## Computation in BayPass

- Post. mean (with associated variance) of the (post-burn-in and thinned) sampled values (column 'M_Pstd' of the *[outprefix_]summary_pij.out* output file)
- $\overline{\widehat{\widetilde{\alpha}_{ij}}} \simeq 0$ but $\mathrm{Var}\left(\widehat{\widetilde{\alpha}_{ij}}\right) < 1$ due to the hierarchical model structure
  $\Rightarrow$ shrinkage of the sampled values towards their prior mean
- Needs to be accounted for calibration of $\widetilde{\alpha}_{ij}$-derived statistics (Olazcuaga et al., 2020)

# The $X^tX$ statistics to identify "outlier" SNPs

## Definition (Guenther and Coop, 2013)

- $\mathbf{X^tX}_i = \mathbf{Var}(\mathbf{X_i}) = \sum_{j=1}^{J} \widetilde{\alpha}_{ij}^2 = \frac{(\boldsymbol{\alpha}_i^\star - \pi_i)\boldsymbol{\Omega}^{-1}(\boldsymbol{\alpha}_i^\star - \pi_i)}{\pi_i(1-\pi_i)}$
- $X^tX \simeq$ SNP–specific $F_{ST}$ (all. freq. variance) corrected for pop. history ($\boldsymbol{\Omega}$)
  $\approx$ immune to demographic factors confounding SNP differentiation
- High (resp. small) $X^tX \Rightarrow$ SNP affected by positive (resp. balancing) selection ?

## How extreme to be candidate ?

- $\widetilde{\alpha}_{ij} \sim N(0,1) \Rightarrow X^tX \sim \chi^2_J$ (i.e. $E(X^tX) = J$ and $Var(X^tX) = 2J$)
- While $\widehat{X^tX} = J$, but shrinkage of the $\widetilde{\alpha}_{ij} \Rightarrow Var\left(\widehat{X^tX}\right) \ll 2J$
- Two possible strategies for the calibration of the estimated $X^tX$
  - post. predictive check (Gautier, 2015) : obtain a distribution of $\widehat{X^tX}$ under $H_0$ from the analysis of PODs simulated under the fitted model ($\boldsymbol{\Omega}^{sim} = \widehat{\boldsymbol{\Omega}}$, $\pi^{sim} \sim B(\widehat{a_\pi}, \widehat{b_\pi})$)
  - $X^tX^\star$ (Olazcuaga et al., 2020) : $\widehat{X^tX}_i^\star = \sum_{j=1}^{J} \left(\widetilde{\alpha}_{ij}^{(u)}\right)^2 \simeq \chi^2_J$ under $H_0$

    $\rightarrow$ use empirically "unshrinked" $\widehat{\widetilde{\alpha}_{ij}^{(u)}} = \frac{\widehat{\widetilde{\alpha}_{ij}} - \hat{\mu}_\alpha}{\hat{\sigma}_\alpha}$ ($\hat{\mu}_\alpha$=mean and $\hat{\sigma}_\alpha$=s.d. over the I×J $\widehat{\widetilde{\alpha}_{ij}}$'s)

# Some limitations of $X^tX$

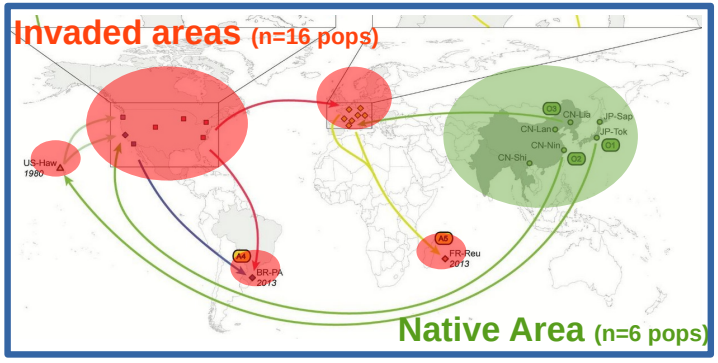
- No use of LD information
  - see window-based or (better) local-score (Fariello et al., 2017) analyses of $X^tX$
- An indirect genome-scan approach
  - no prior assumption on the driving selective pressure
  - makes biological interpretation of the results harder
- Possible strategies
  - using gene functions : cdts SNPs $\rightarrow$ genes (but LD) $\rightarrow$ biol. pathways
    - requires an annotated genome
    - highly prone to misleading story telling issues (e.g., Pavlidis et al., 2012)
  - GF or RDA on the outliers loci (whose variation is not explained by demography alone)
    - to identify candidate covariables driving selective pressure

# The $C_2$ contrast statistics : a first step toward GEA

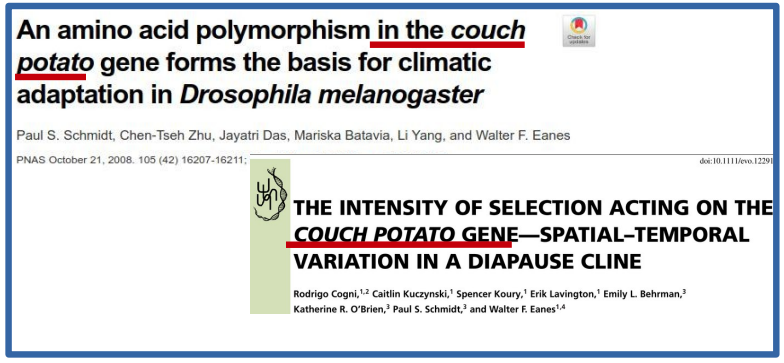
- Test association with a binary covariate (high/low altitude, salinity, etc.) that distinguishes two groups of populations $\mathcal{G}_1$ and $\mathcal{G}_2$ (Olazcuaga et al., 2020)
- $C_2$ = squared difference of (scaled) all. freq. between the two groups :
  - $\widetilde{\alpha}_{ij} \sim \mathrm{N}(0,1) \Rightarrow C = \frac{1}{\sqrt{\tilde{n}_p}} \left( \sum_{j \in \mathcal{G}_1} \widetilde{\alpha}_{ij} - \sum_{k \in \mathcal{G}_2} \widetilde{\alpha}_{ik} \right) \sim \mathrm{N}(0,1)$

    where $\tilde{n}_p = n(\mathcal{G}_1) + n(\mathcal{G}_2)$
  - Under $H_0$ (only neutral differentiation) : $C_2 = \frac{1}{\tilde{n}_p} \left( \sum_{j \in \mathcal{G}_1} \widetilde{\alpha}_{ij} - \sum_{k \in \mathcal{G}_2} \widetilde{\alpha}_{ik} \right)^2 \sim \chi^2_1$
- $\widehat{C_2}$ estimated from the (empirically) "unshrinked" $\widehat{\widetilde{\alpha}_{ij}}$'s (Olazcuaga et al., 2020)

# Association with the invasive status in *D. suzukii*




***Pool-Seq (>$10^6$ SNPs)***

An amino acid polymorphism in the *couch potato* gene forms the basis for climatic adaptation in *Drosophila melanogaster*

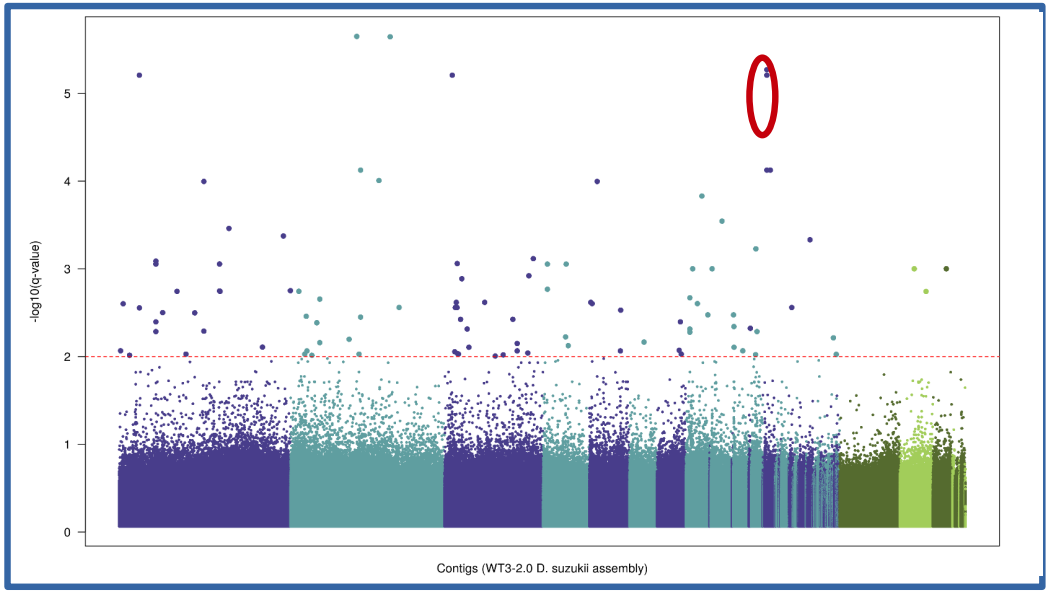
Paul S. Schmidt, Chen-Tseh Zhu, Jayatri Das, Mariska Batavia, Li Yang, and Walter F. Eanes

PNAS October 21, 2008. 105 (42) 16207-16211;

THE INTENSITY OF SELECTION ACTING ON THE *COUCH POTATO* GENE—SPATIAL–TEMPORAL VARIATION IN A DIAPAUSE CLINE

Rodrigo Cogni, Caitlin Kuczynski, Spencer Koury, Erik Lavington, Emily L. Behrman, Katherine R. O'Brien, Paul S. Schmidt, and Walter F. Eanes

***Genome-Scan***
*($C_2$ invasive vs. native)*

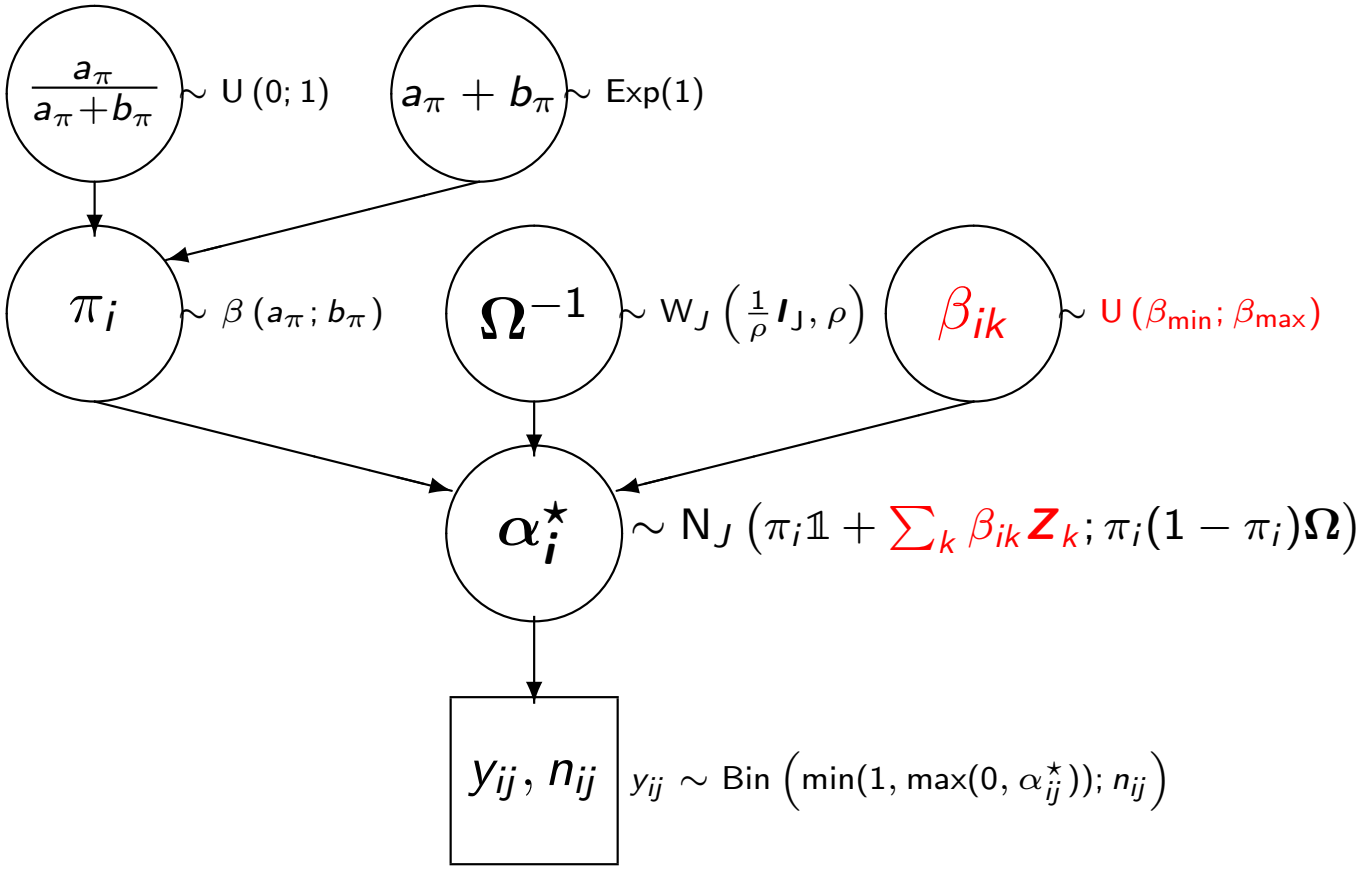



***Functional Annotation***

***[Olazcuaga et al., 2020]***

**See poster by Louise Camus** *(about $C_2$, GF and GO to characterize/predict biological invasion)*

# BayPass "standard" covariate model for GEA/pGWAS



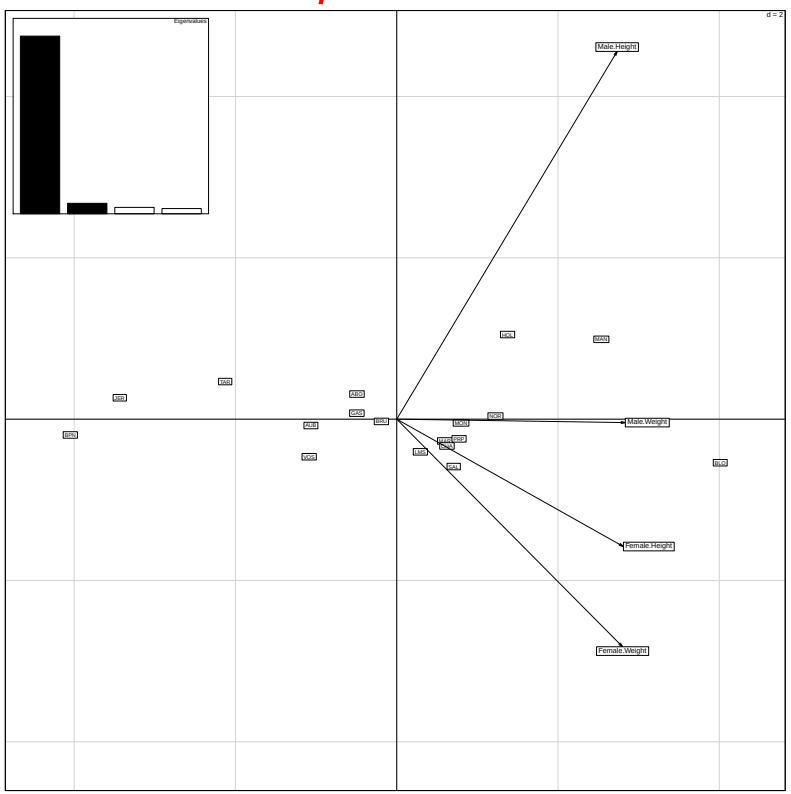

- $\Leftrightarrow$ (linear) regression of the (scaled) allele freq. on each covariate $k$ : $\boldsymbol{Z}_k = z_{jk}$
- Similar to BayEnv model (Coop et al., 2010) with additional extensions
  - Priors on $a_\pi$, $b_\pi$ and $\boldsymbol{\Omega}^{-1}$ (by default $\beta_{min} = -0.3$ instead of $-0.1$ and $\beta_{max} = 0.3$ instead of 0.1)
  - multivariate (pop. covariable assumed independent)

# Note : How does the covariable file look like ?

2 covariates in 18 cattle breeds (Gautier, 2015) : *Morpho. Score* and *Piebald color 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.]
```

*Ex. SMS Morpho score*



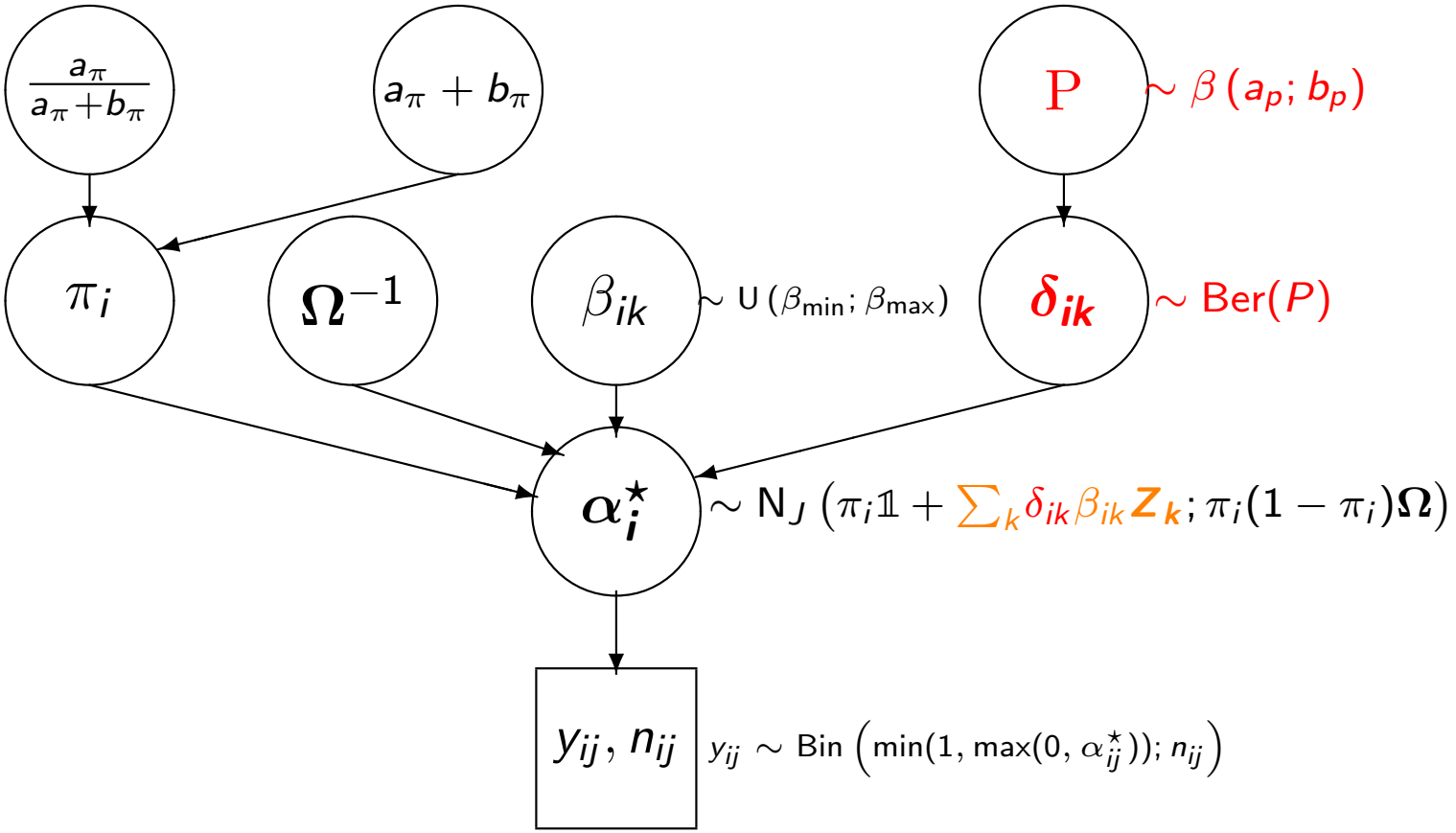

Best practices...

- Usually better to scale the covariables (*-scalecov* option)
- If several covariables, may use PCA to analyze only a few (uncorrelated) PCs (⇔ "synthetic" scores) →may hamper biological interpretation

# Estimating the $\beta_i$'s and assessing association significance

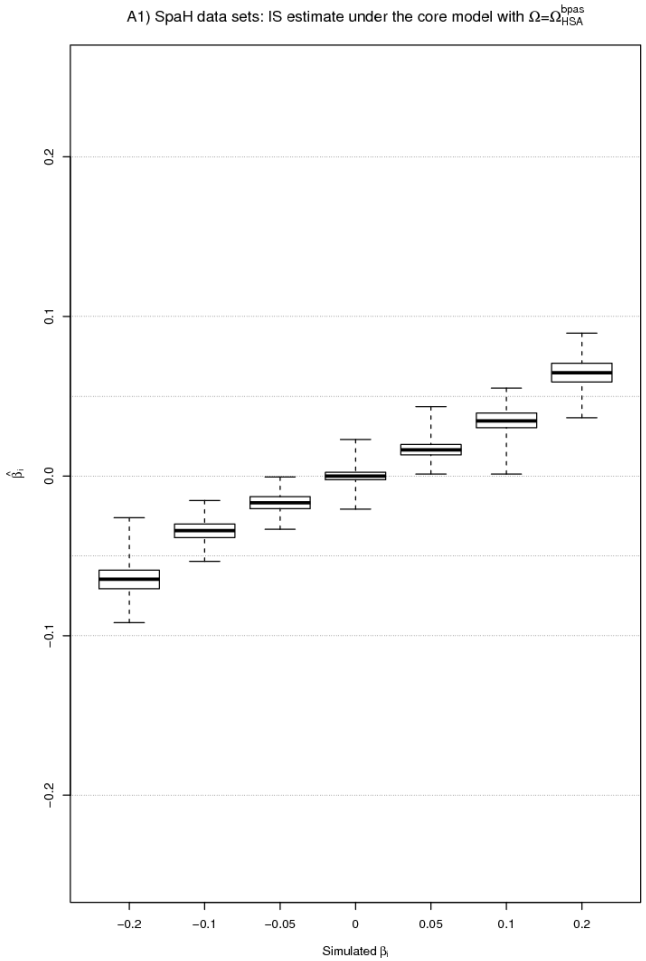
- A) Via Importance Sampling (only requires par. values sampled under the core model)
  - Gaussian approximation of the posterior distribution of the $\beta_i$'s
    - $\beta_i$ | data $\sim \mathrm{N}\left(\widehat{\mu(\beta_i)}, \widehat{\sigma(\beta_i)}\right)$
    - $Z_i = \frac{\widehat{\mu(\beta_i)}}{\widehat{\sigma(\beta_i)}}$ and $\mathrm{eBP}_{\mathrm{is}} = -\log_{10}\left(1 - 2\left|0.5 - \Phi\left(Z_i\right)\right|\right)$ (eBP>4 $\Leftrightarrow |Z_i| > 3.7$)
  - Direct approximation of the Bayes Factor ($\mathrm{BF}_{\mathrm{is}}$)
    - Two model comparison : with (i.e. $\beta_i \neq 0$) *vs.* without association (i.e. $\beta_i = 0$)
- B) Via MCMC (*-covmcmc* option)
  - Sampling from the posterior distribution of the $\beta_i$'s via MCMC
  - Posterior $\widehat{\mu\left(\beta_i\right)}$ and $\widehat{\sigma\left(\beta_i\right)} \Rightarrow \mathrm{eBP}_{\mathrm{mc}}$
- C) Using variable selection model (*-auxmodel* option)
  - To estimate BF via MCMC ($\mathrm{BF}_{\mathrm{mc}}$)

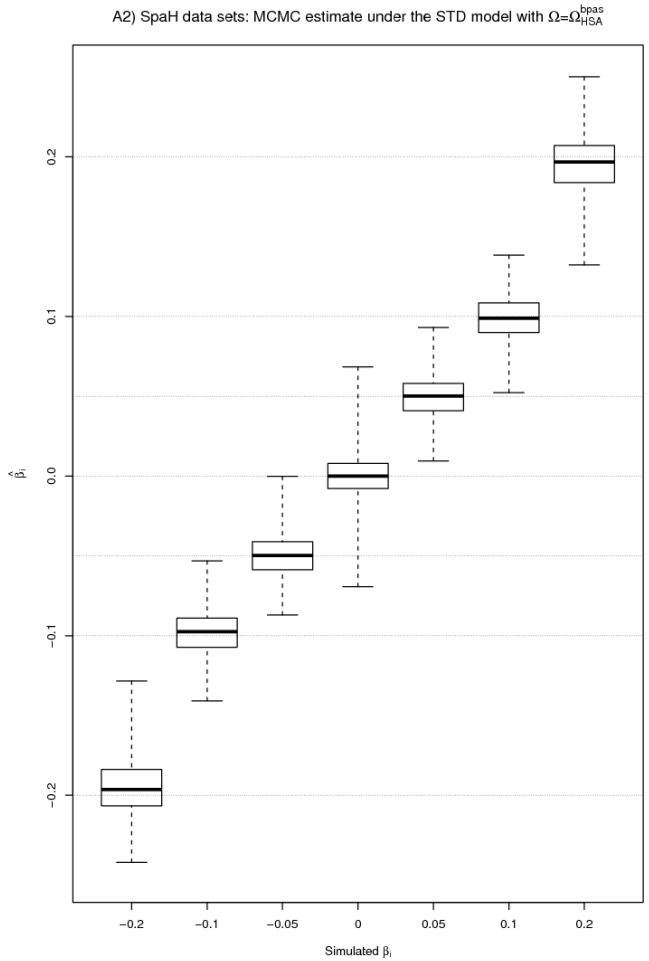
# The "AUX" covariate model (i.e., with 'auxiliary variable')

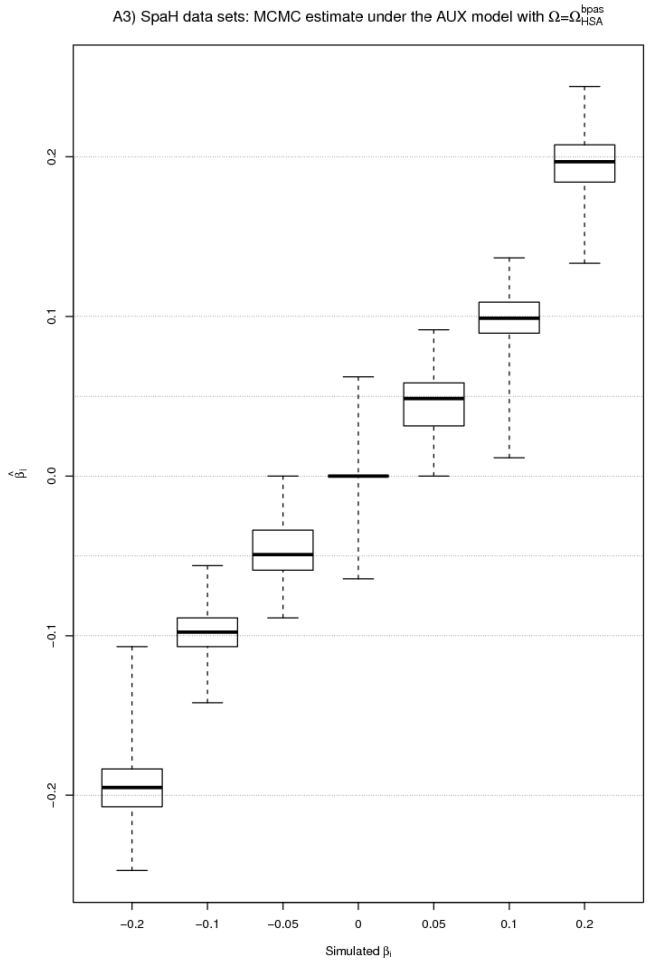



- The binary variable $\delta_i$ specifies whether the SNP is associated ($\delta_i = 1$) or not ($\delta_i = 0$)
- Integrating over $P$ (prop. of associated SNPs) allows dealing with multiple testing issues
- From $\mathrm{P}[\delta_i = 1|data]$ (a.k.a. PIP), $\mathbb{BF}_{\mathrm{mc}} = \frac{\text{Post. odds}}{\text{Prior odds}} = \frac{\text{PIP}}{1-\text{PIP}} \times \frac{1-\mathbb{E}[P]}{\mathbb{E}[P]}$ (with $\mathbb{E}[P] = \frac{a_P}{a_P+b_P}$)

# Accuracy of the different estimates of $\beta_i$'s

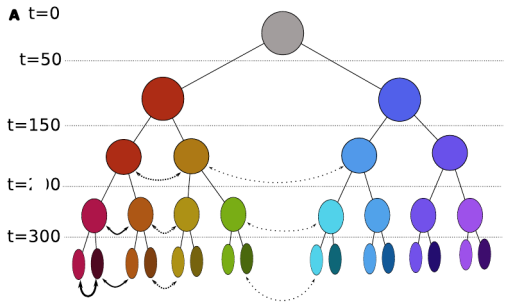


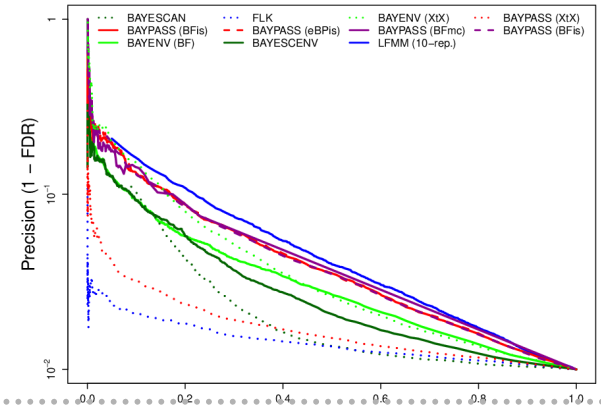



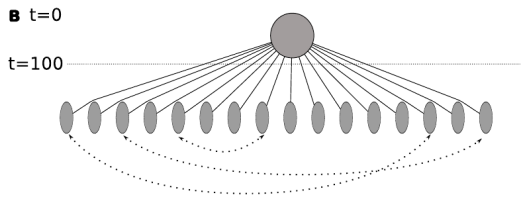
# Simulation study with polygenic selection (as in de Villemereuil et al., 2014)

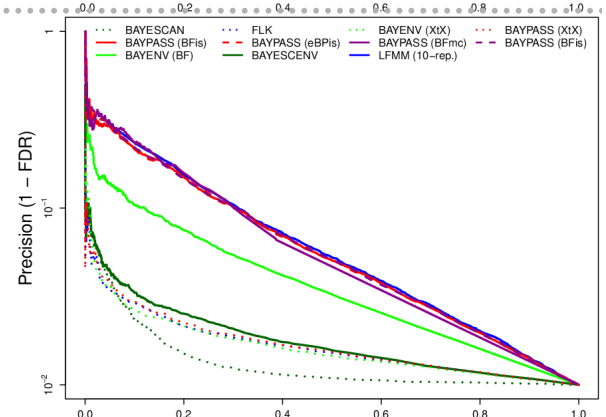
**A) Hierarchical with migration**



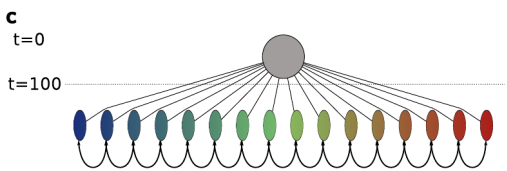



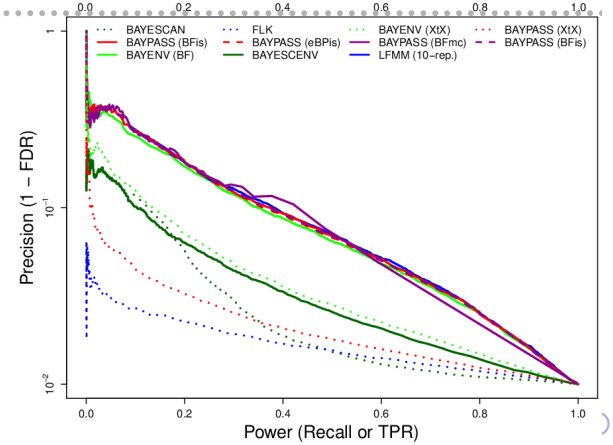

**B) Isolation with Migration (IMM)**




**C) Stepping Stone**



# General comments and limitations

## Decision rule

- eBP ($eBP_{is}$ or $eBP_{mc}$) not recommended
  - not well calibrated ($\neq$P-value) $\rightarrow$ may be removed in future BayPass release
- Use BF (explicit/rigorous model comparison) !
  - Jeffreys' rule : evidence "very strong" ("decisive") if $15 < 10 \log_{10}$ BF $< 20$ ($> 20$)
  - post. pred. checking ($\rightarrow$BF distrib. for "neutral" PODs SNPs) : usually consistent

## To sample the $\beta_i$'s or not ? IS *vs.* MCMC (`-covmcmc` and `-auxmodel`)

- Technical aspects
  - IS estimates ($BF_{is}$, etc.) are approximated : check consistency across (e.g., 3–5) independent chains (`-seed`) $\Rightarrow$ usually OK
  - If $> 1$ covariables : IS univariate vs. MCMC multivariate (for now, cov assumed independent)
  - "AUX" model ($BF_{mc}$) deals with multiple testing and MCMC more accurate
- In practice
  - MCMC require a prior estimates of $\widehat{\mathbf{\Omega}} \Rightarrow BF_{is}$ may always be available at $\sim$no extract cost
  - When *npop* is small (e.g., $\leq 8$) or highly differentiated, prefer IS (MCMC est. may be "unstable")
  - When data are not limiting, $BF_{mc}$ (MCMC) should be preferred for decision

# Genomic prediction of population covariate

## Rationale

- Relying on the relationship between genetic and covariate variation among populations to estimate the pop. covariate values
  $\Rightarrow$ pop-specific covariate treated as a random variable
- Interpretation : pop. mean phenotype or tolerance range (e.g., for env. covariable)

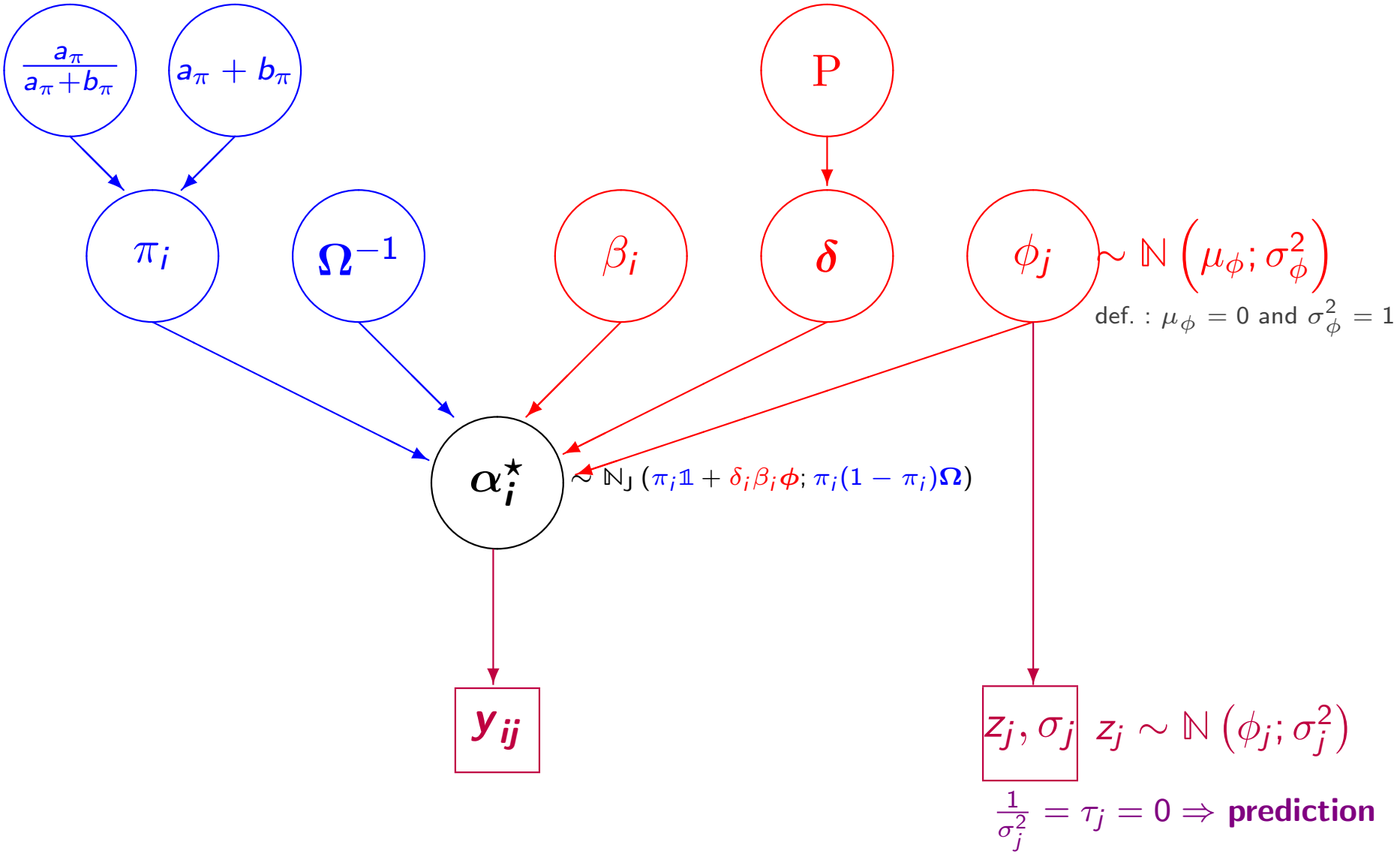
## Possible strategies

- First identify associated SNPs and estimates SNP effect, then predict
  - Analogy with Marker Assisted Selection in breeding programs
  - Strong limitations (but for simple architecture) : GEA/GWAS only identifies a fraction of QTNs and may overestimate their effect ($\Leftrightarrow$ "Beavis" effect)
- Joint modeling if all the genome and trait covariables variation
  - Analogy with Genomic Selection in breeding programs
  - (far) more efficient for complex traits (i.e., with a polygenic genetic architecture)

## Extending the BayPass model for genomic prediction

- Modeling uncertainty of the population covariate values
- full uncertainty $\Rightarrow$ prediction

# The 'AUX' genomic prediction model (univariate case)

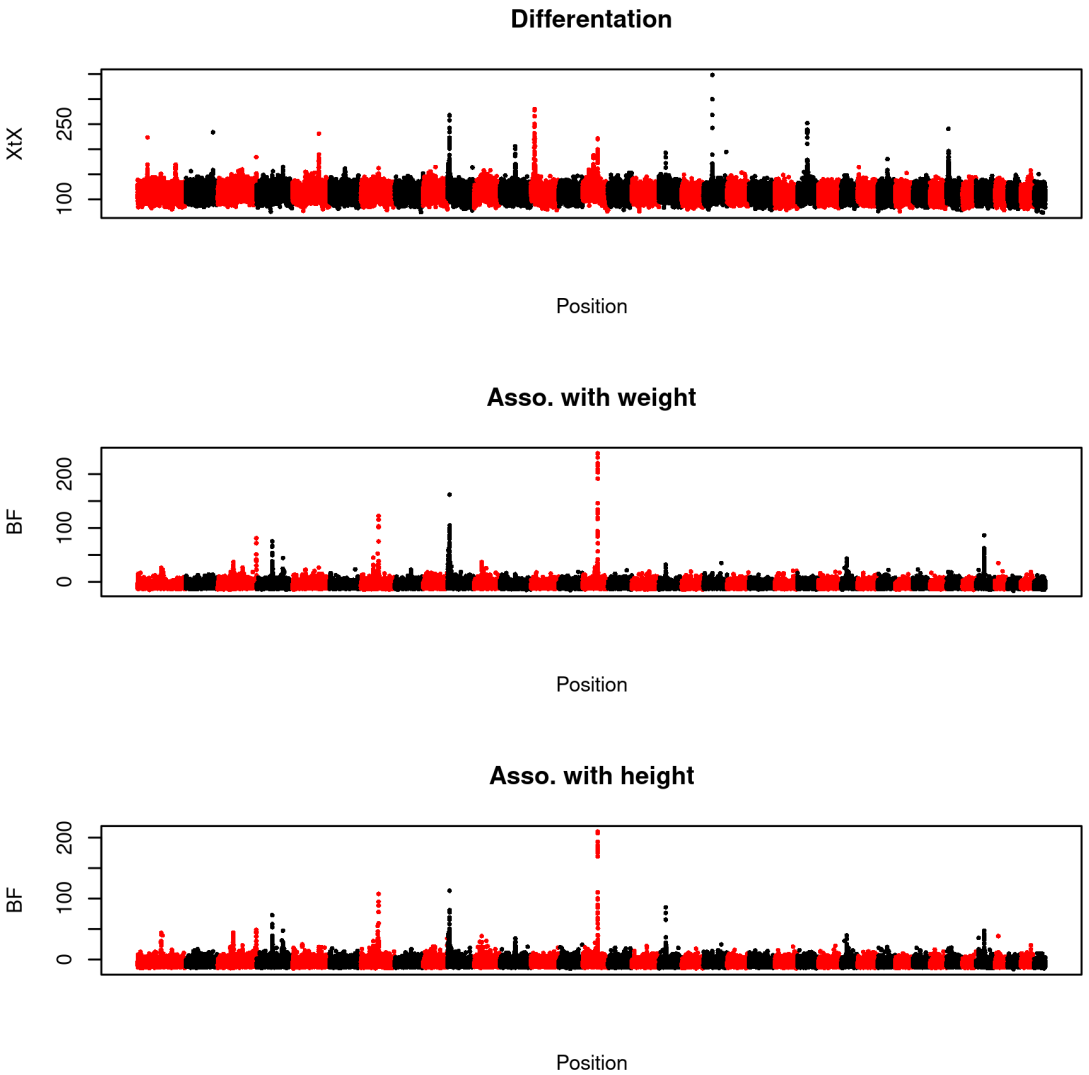


# Empirical evaluation : dog breeds morphology traits

## Data (Hayward et al., Nat. Comm., 2016)

- 111 breeds *(n=6–636 ; med=17)*
- 155,609 autosomal SNPs *(Illumina canineHD chip)*
  - At least 5 genotyped individuals per breed
  - Overall MAF$>0.01$
  - Moderate overall differentiation : $\widehat{F}_{ST} = 0.240$
- Two breed-average phenotypes
  - Male Height and Male Weight *($w^{0.38}$ transfo.)* scaled over all breeds
  - Ind. phenotypes obtained from the American Kennel Club ($\rightarrow \neq$ genotyped inds.)

# BayPass association results




ARTICLE
Received 13 Oct 2015 | Accepted 11 Dec 2015 | Published 22 Jan 2016
DOI: 10.1038/ncomms10460 OPEN

## Complex disease and phenotype mapping in the domestic dog

Jessica J. Hayward[1,*], Marta G. Castelhano[2,*], Kyle C. Oliveira[1], Elizabeth Corey[2], Cheryl Balkman[2], Tara L. Baxter[1], Margret L. Casal[3], Sharon A. Center[2], Meiying Fang[5], Susan J. Garrison[2], Sara E. Kalla[2], Pavel Korniliev[4], Michael I. Kotlikoff[1], N.S. Moise[2], Laura M. Shannon[1], Kenneth W. Simpson[2], Nathan B. Sutter[6], Rory J. Todhunter[2] & Adam R. Boyko[1]

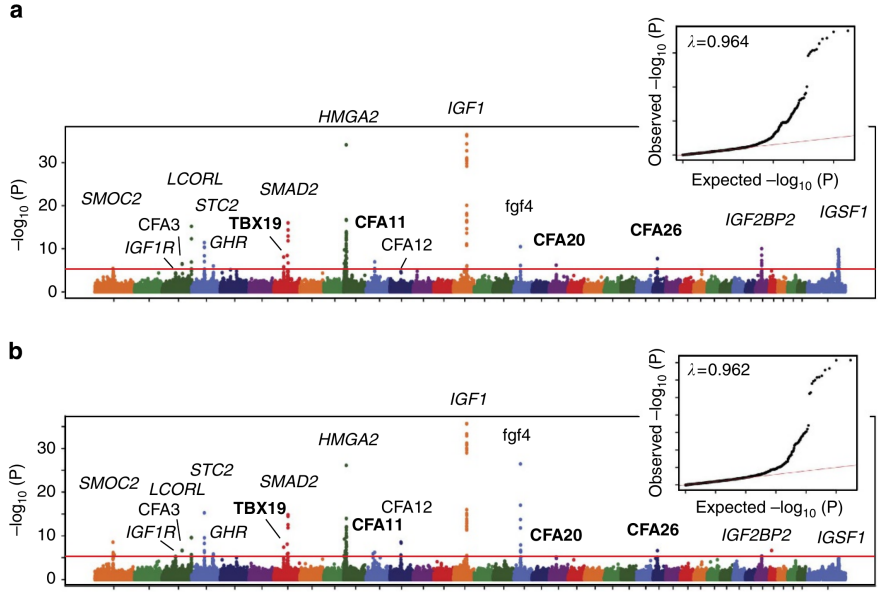
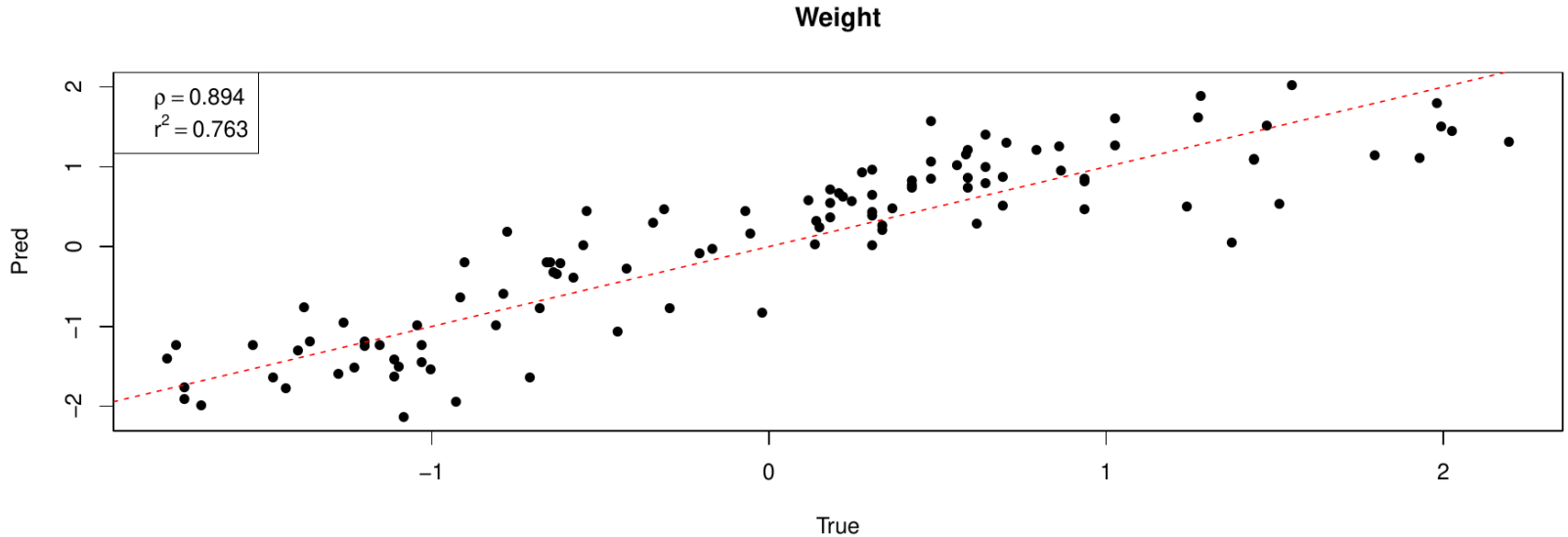



**Figure 3 | Body size association results.** Manhattan and quantile–quantile plots of (**a**) breed-average male weight$^{0.38}$ ($n=1,873$) and (**b**) breed-average male height ($n=1,873$), showing the 17 significant loci, four of which are novel (shown in bold). Red lines on the Manhattan plots are the significance thresholds, at $P=5\times10^{-6}$ (FDR of $<0.5\%$ and $<0.75\%$ for weight and height, respectively). Inflation factors ($\lambda$ values) are shown on the quantile–quantile plots. (**c**) Proportion of variance explained ($R^2$) by the 17 size loci in a linear model for weight (blue bars) and height (green bars), with sex and

# Evaluation of the AUX prediction model

## Leave-one out analysis

- 1 breed-phenotype assumed unknown to predict $(\frac{1}{\sigma_j} = 0)$
- The 110 others set to their actual values $\pm 0.01$ $(\sigma_j = 0.01)$
  $\Rightarrow$ 111 analyses per phenotype in total




# The AUX prediction model seems promising but...

- Still experimental (hidden in BayPass v2.3)
- Need to be tested far more extensively to evaluate the influence of
  - number of SNPs (w.r.t. across pop. LD) and pops
    $\Rightarrow$ *(very) poor performance on 18 cattle breeds with 40K SNPs (Gautier, 2015)*
  - genetic architecture of the trait/response
  - number of values to predict and uncertainty on the training covariables
- Extension needed to properly deal with categorical variable *(in progress)*
- Muli-trait (covariable) prediction (prior specification : e.g., )
  - for now : all covariable assumed Gaussian i.i.d.
    $\Rightarrow$ *OK in the dog example even if breed weight and height highly correlated*
  - Predict PC's and back-transform estimates to "natural" scale (but C.I. ?)
  - Try more informative prior specification
    *e.g., covariance* $\propto \mathbf{\Omega}$ *for phenotype trais ?*

# Conclusions (1)

## Why Bayesian modeling ? (philosophical considerations aside !)

- Versatility making it easier
  - to account for imperfection in the data (unbalanced designs, missing data, etc.)
  - to capitalize on prior knowledge
  - to model additional source of variation (e.g., Pool-Seq, Ind-Seq GL, pop. covariables)
  - to combine data sets *(Pool-Seq + Ind-Seq GL + count data in* BayPass *3.0)*
- Why BayPass ?
  - Accounts for the neutral structuring of genetic diversity (demographic history)
  - Robust approaches for genome scan ($X^tX$) and GEA/pGWAS ($C_2$, BF)
  - Provides other possibly useful estimates (e.g., scaled allele frequencies)

# Conclusions (2)

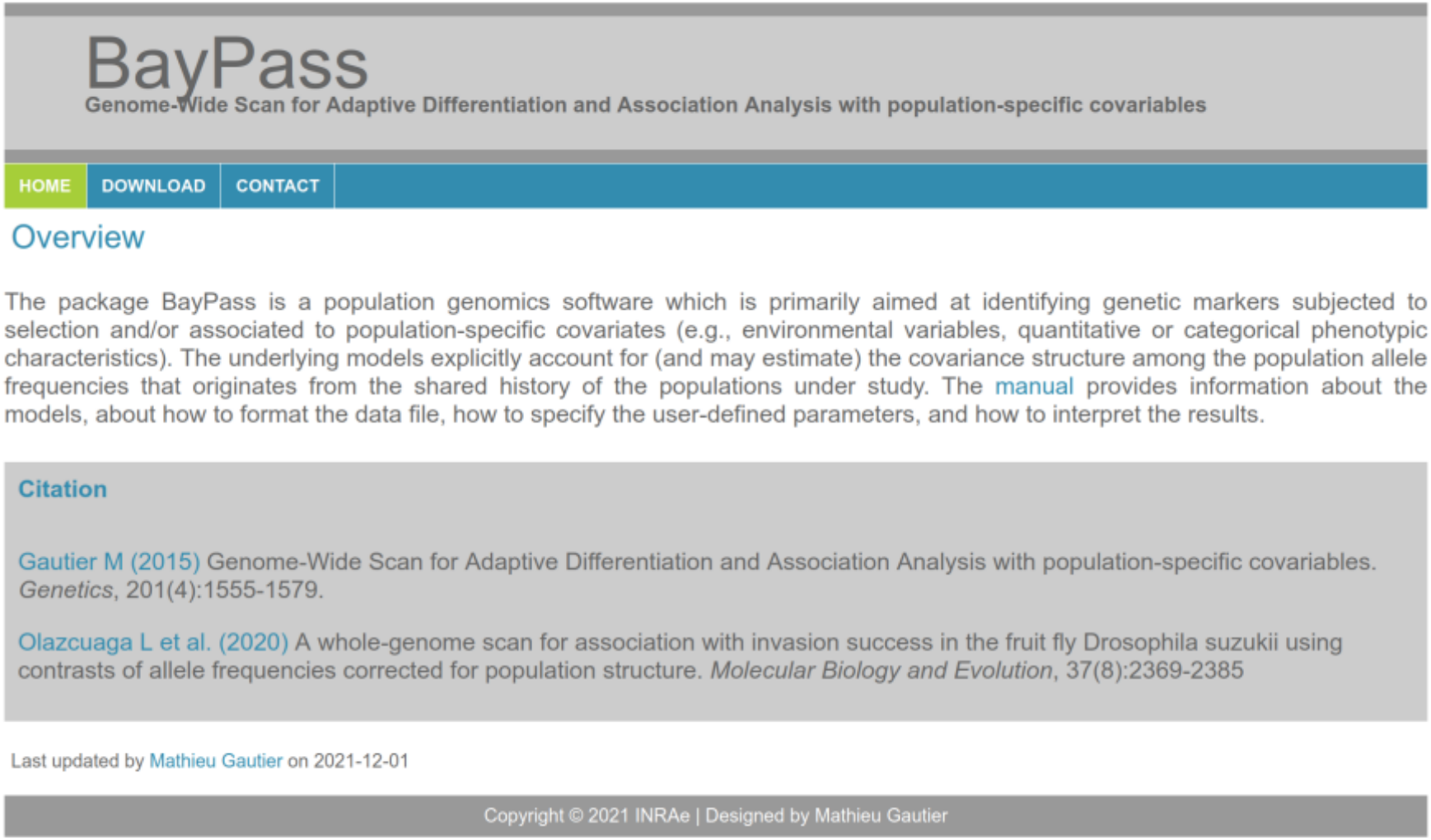
## Limitations of genome scans

- Covariate free (indirect) approaches ($X^tX$ for genome scan for adaptive differentiation)
  $\Rightarrow$ biological interpretation may be challenging
- GEA/pGWAS only provide access to a (small) fraction of genetic architecture of the associated trait or covariable *(see Lotterhos, 2022)*

## Some perspectives...

- identify ecological drivers of adaptation
  $\Rightarrow$ *GF or RDA ( ?) on extreme $X^tX$ loci, i.e., not fully explained by demography*
- correct for demography (scaled all. frequencies) in GF (or RDA)
  $\Rightarrow$ *benefits not clear yet (overcorrection ?)*
- predict population covariable
  $\Rightarrow$ *recall $\neq$ characterizing genetic architecture of the underlying traits*

# Thank You For Your Attention



BayPass

Genome-Wide Scan for Adaptive Differentiation and Association Analysis with population-specific covariables

HOME DOWNLOAD CONTACT

## Overview

The package BayPass is a population genomics software which is primarily aimed at identifying genetic markers subjected to selection and/or associated to population-specific covariates (e.g., environmental variables, quantitative or categorical phenotypic characteristics). The underlying models explicitly account for (and may estimate) the covariance structure among the population allele frequencies that originates from the shared history of the populations under study. The manual provides information about the models, about how to format the data file, how to specify the user-defined parameters, and how to interpret the results.

**Citation**

Gautier M (2015) Genome-Wide Scan for Adaptive Differentiation and Association Analysis with population-specific covariables. *Genetics*, 201(4):1555-1579.

Olazcuaga L et al. (2020) A whole-genome scan for association with invasion success in the fruit fly Drosophila suzukii using contrasts of allele frequencies corrected for population structure. *Molecular Biology and Evolution*, 37(8):2369-2385

Last updated by Mathieu Gautier on 2021-12-01

Copyright © 2021 INRAe | Designed by Mathieu Gautier

http://www1.montpellier.inra.fr/CBGP/software/baypass/