Genome-wide association and prediction at the population level using BayPass

Mathieu Gautier

UMR INRAE/CIRAD/IRD/SupAgro CBGP

9th September 2025

Introduction

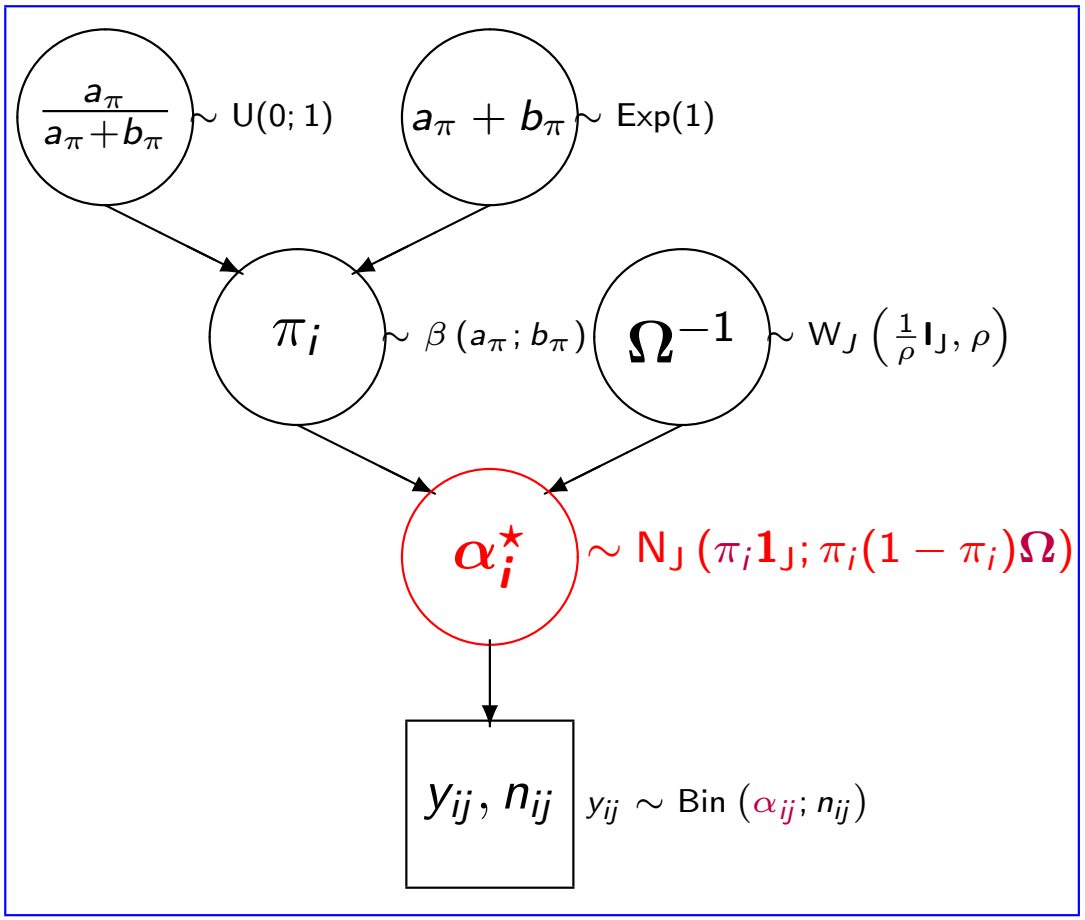
Genome-wide association with population-specific covariables

- Modelling the relationship between genetic diversity and population covariables of interest across several (differentiated) populations may allow
  - uncovering the nature of adaptive traits and their genetic architecture
  - Estimating population maladaptation
  - predicting covariate value from genomic information
- Different covariables of interest
  - Environmental (e.g., bioclimatic covariates, host plant, etc.) $\Rightarrow$ GEA
  - Phenotypic (e.g., mean height, mean weight, coat color) $\Rightarrow$ "pGWAS"

Demographic history : a critical confounding factor

- Shared population history $\Rightarrow$ covariance structure of allele freq.

The BayPass core model (Gautier, 2015)



- Multivariate Gaussian prior on the (unobserved) allele frequencies at I SNPs in J pops
- "instrumental" allele freq. $\alpha_{ij}^\star$ defined over the real line support :

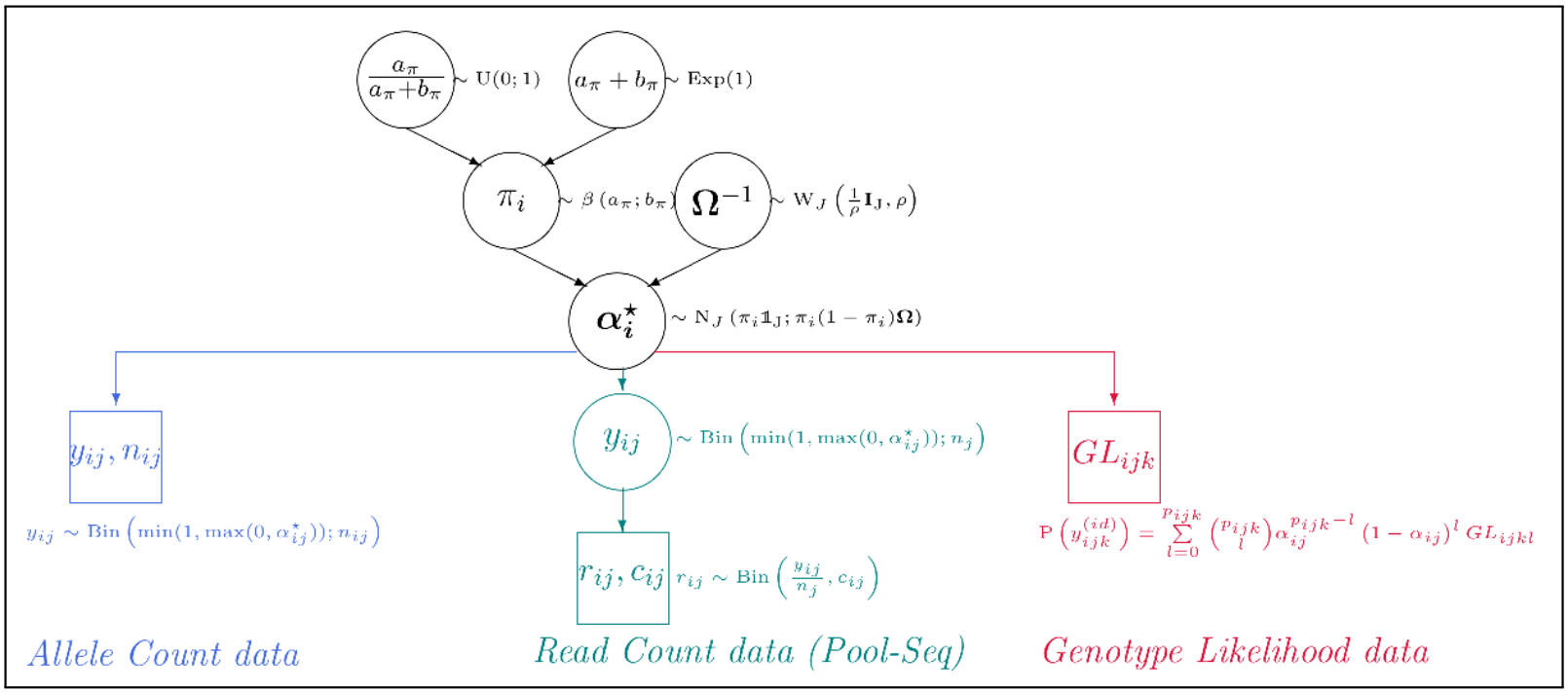
$$\alpha_{ij} = \begin{cases} \alpha_{ij}^\star, & \text{if } \alpha_{ij}^\star \in (0,1), \\ 0, & \text{if } \alpha_{ij}^\star < 0 \text{ (allele "lost")}, \\ 1, & \text{if } \alpha_{ij}^\star > 1 \text{ (allele "fixed")}. \end{cases}$$

- π_i may be interpreted as an "ancestral" ref. allele freq. of SNP i
- $\mathbf{\Omega}$ = J×J scaled cov. matrix of all. freq. $\Leftrightarrow$ "population relationship matrix" (captures the global effect of the demo.)

- BayPass (current=v3.1) : (adaptive) MCMC sampler coded in (modern) Fortran consisting of :
  - Pilot runs to adjust parameters of the proposals : *def : 10×100 iterations*
  - Burn-in period (to achieve stationary distributions) : *def : 1,000 iterations*
  - Parameters Sampling with thinning (to reduce auto-correlations) : *def : 20×250 iterations*

Some advantages of BayPass

- Handle various pop sample data (+hybrid data sets since ≥v3.0, Camus et al., 2024)
  - Pop. sample Allele Counts (e.g., $y_{ij} = 2n_{ij}^{hom_{ref}} + n_{ij}^{het}$)
  - Pop. sample Read Count (Pool–Seq) *(to integrating over unobs. allele count)*
  - Pop. sample Indiv. Genotype Likelihoods *(to deal with geno. uncertainty in low/medium Ind-Seq WGS)*



- Robust
  - Joint estimation of the (unobserved) π_i's and $\mathbf{\Omega}$ and other (hyper-)parameters
  - Unbalanced population origins and sample sizes ; missing data ; SNP ascertainement bias

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 (∼5 min on 1 CPU) :

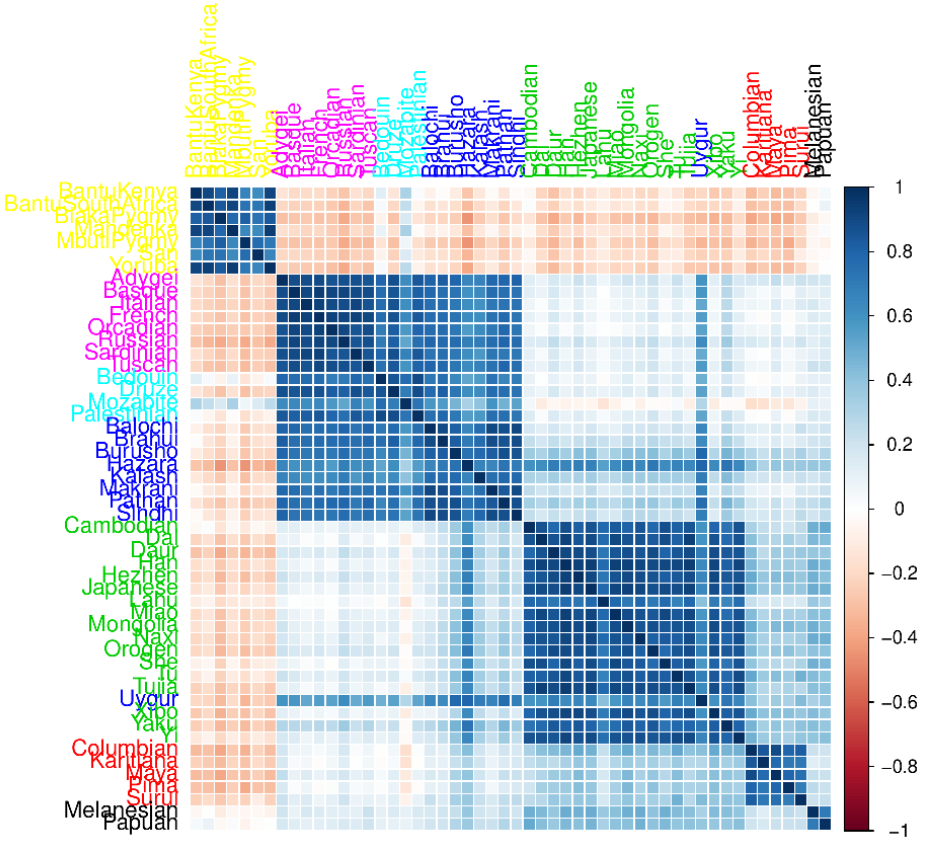
```
g_baypass -countdatafile hgdp.geno -outprefix corehgdp -nthreads 4
```

- Various options and flexible parametrization **(default should be appropriate for most, if not all, analyses)**
- If your are lost (see also the manual !) :

```
g_baypass -help
```

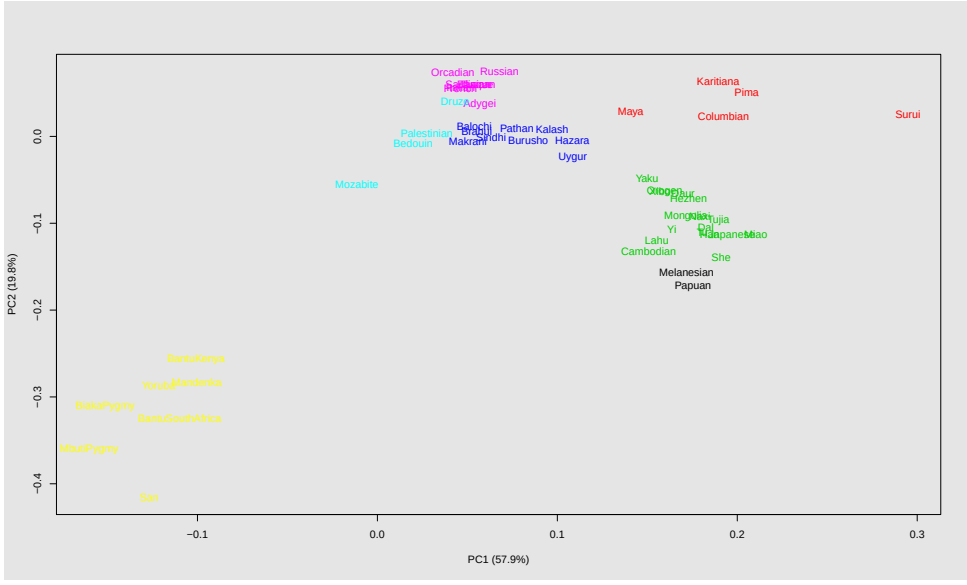
Example of visualization of Ω within R

A) Correlation Map




```
>require(corrplot)
>om=as.matrix(read.table("corehgdp_mat_omega.out"))
>corrplot(cov2cor(om))
```

B) Spectral Decomposition ($\Omega = U\Lambda U'$)




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

Correcting allele frequencies for demographic history

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

- $\mathbf{X_i} \simeq$ pop. allele freq. corrected for their joint demographic history
  - $\mathbf{X_i} = \{\widetilde{\alpha}_{ij}\}_{1\ldots J} = \mathbf{\Gamma^{-1}} \frac{\boldsymbol{\alpha_i^\star} - \pi_i}{\sqrt{\pi_i(1-\pi_i)}}$ (Guenther and Coop, 2013) where $\mathbf{\Omega} = \mathbf{\Gamma\Gamma'}$
  - if SNP i is "neutral", $\widetilde{\alpha}_{ij} \sim \mathrm{N}(0,1)$ for all populations j
- See *[outprefix_]summary_pij.out* BayPass output file

Related statistics (Guenther & Coop, 2013 ; Olazcuaga et al., 2020)

- $\mathrm{X^tX}$ and $\mathrm{X^tX^\star}$ for genome scan of adaptive differentiation
  $\rightarrow$ *[outprefix_]summary_pi_xtx.out*
- C_2 for association with binary covariable (`-contrastfile` option)
  $\rightarrow$ *[outprefix_]summary_contrast.out*

BayPass models for association studies (GEA/pGWAS)

- Equivalent to a Multivariate Linear Regression of the scaled allele frequencies $\widetilde{\alpha}_{ij}$ (SNP i ; pop. j) on pop. covariate vectors $\boldsymbol{Z}_{\boldsymbol{k}}^{(k)} = \{z_{jk}\}_{1..J}$ ($\Leftrightarrow$ "fixed" effect) :

$$\widetilde{\alpha}_{ij} = \sum_{k=1}^{K} \beta_{ik} z_{jk} + \epsilon_{ij} \text{ with } \epsilon_{ij} \sim \mathrm{N}(0,1)$$

- Accounts for the confounding ($\Leftrightarrow$ "random") effect of shared population history by the modeling of $\widetilde{\alpha}_{ij}$ (instead of α_{ij})
- If $\widehat{\beta}_{ik} \neq 0$, SNP i is deemed associated with the k$^{\text{th}}$ covariate
  - Formalize decision via Bayes Factor (comparing the two models $\widehat{\beta}_{ik} \neq 0$ vs. $\widehat{\beta}_{ik} = 0$)
  - e.g., Jeffreys's rule : BF>15 dB $\Rightarrow$ "decisive" evidence for association

Three procedures to estimate the β_i's and/or BF's

1. From $\widetilde{\alpha}_{ij}$'s sampled under the core model (`-efile` option) :

```
g_baypass -countdatafile hgdp.geno -efile hgdp.cov -outprefix core
```

Importance Sampling approximations (*core_summary_betai_reg.out*)

$\Rightarrow$ *"quick and dirty" & univariate regression on each covariable*

2. "covmcmc" mode (`-covmcmc -efile` & `-omegafile` i.e. needs a prior estimate of Ω) :

```
g_baypass -countdatafile hgdp.geno -efile hgdp.cov
          -omegafile core_mat_omega.out -covmcmc -outprefix covmc
```

MCMC sampling of the β_i's (*covmc_summary_betai.out*)

$\Rightarrow$ *accurate estimation (robust to cov. correlation) but decision made harder*

3. "auxmodel" mode (`-auxmodel -efile` & `-omegafile` i.e. needs a prior estimate of Ω) :

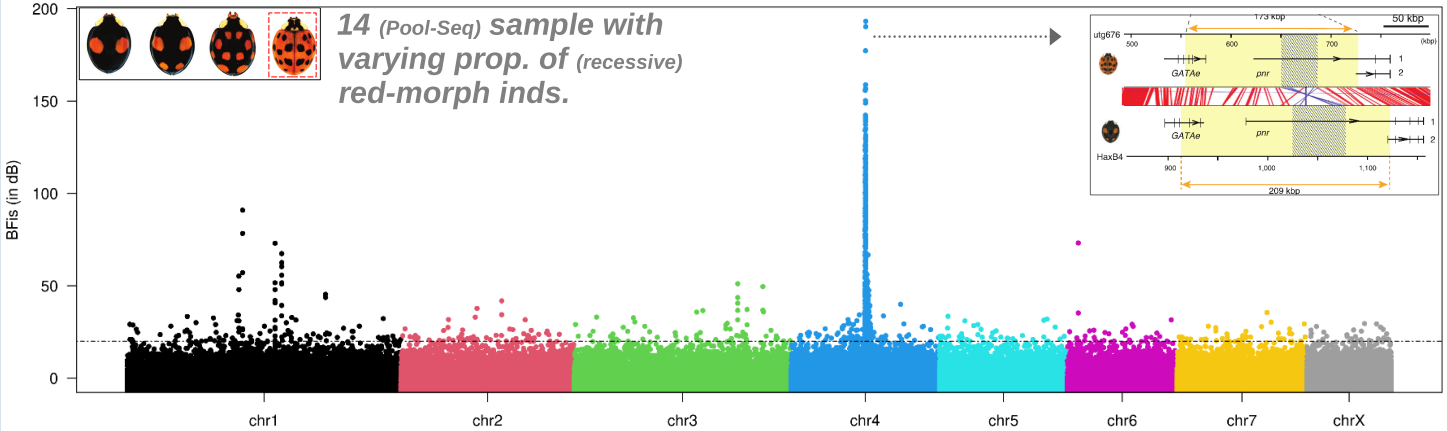
```
g_baypass -countdatafile hgdp.geno -efile hgdp.cov
          -omegafile core_mat_omega.out -auxmodel -outprefix aux
```

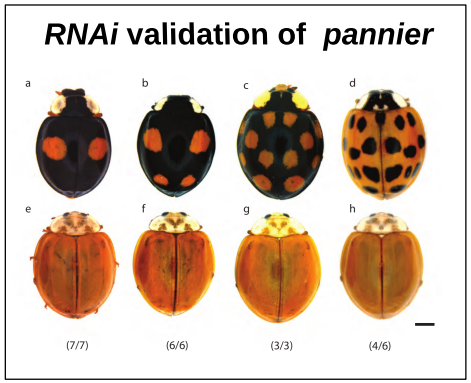
Penalized regression (*aux_summary_betai.out*)

$\Rightarrow$ *proper BF estimation but β_i's shrinked towards 0 (not robust to cov. correlation)*

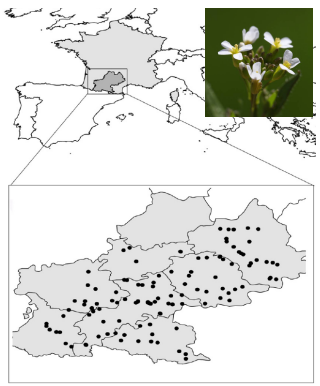
Example of applications

A) pGWAS and color morphs in the ladybird beetle *H. axyridis* (Gautier et al., 2018)



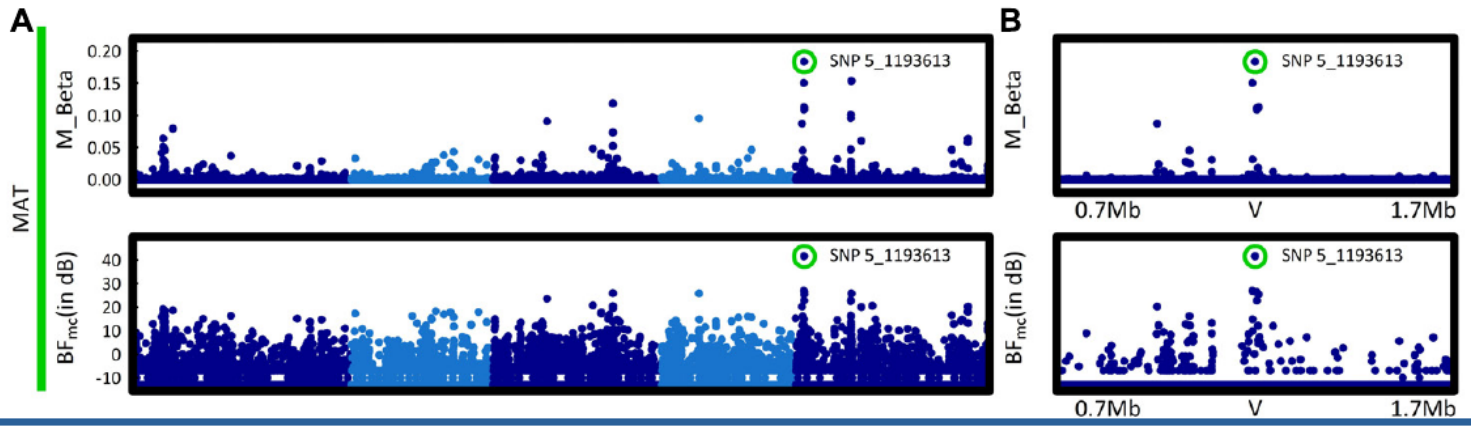


B) GEA and climate adaptation in *A. thaliana* (Frachon et al., 2018)



168 (Pool-Seq) samples (micro-geographic scale)

GEA with 6 non correlated env. Covariates *(e.g. Mean Annual Temperature)*



GEA models : beyond the hunt for genes...

Simple (but efficient) modeling of the relationship (across populations) between adaptive genomic composition and the environment

- In GEA linear models (e.g., BayPass) : the β's quantify the effect of (env.) covariates on the genetic diversity of adaptive variants

$$\widetilde{\alpha}_{ij} = \beta_i^{(1)} z_j^{(1)} + \ldots + \beta_i^{(K)} z_j^{(K)} + \epsilon_{ij}$$

- The ($n_{\text{snps}} \times n_{\text{cov}}$) matrix $\boldsymbol{B} = \{\beta_{ik}\}$ summarizes (linearly) the relationship between adaptive genetic diversity and environment (on a genome-wide basis)

Some assumptions to gain insights from $\boldsymbol{B}$ (Gain et al., 2023)

- Genotyped SNPs capture the whole-genome adaptive genetic diversity
- Sampled populations are representative of species diversity (for the geographical scale of interest) and locally adapted
- (some) covariables are (co)related to the (main) selective pressure
  $\boldsymbol{B}$ *may then give insights into those driving adaptation (e.g., via s.v.d.)*

Evaluating population maladaptation to a new environment

The (geometric) Genetic Offset (Gain et al., 2023)

- If $\boldsymbol{e}_o$ (resp. $\boldsymbol{e}^\star$) is the vector of the K covariable values (e.g., bioclim variables) for the original (resp. new) environment :

$$\text{GO} = \frac{1}{l} \left(\boldsymbol{e}_o - \boldsymbol{e}^\star\right)' \boldsymbol{B}'\boldsymbol{B} \left(\boldsymbol{e}_o - \boldsymbol{e}^\star\right) = \frac{1}{l} \sum_{i=1}^{l} \left(\tilde{e}_i - \tilde{e}_i^\star\right)^2$$

- GO $\Leftrightarrow$ (squared) euclidean environmental distance ("genetically") weighted by the env. effect on adaptive genetic diversity)
- Some restrictive assumptions but supported by simulated and empirical data (e.g., Laruson et al., 2022, Gain et al., 2023)

Computation with BayPass (details in manual pp29-30)

- Use β_i's estimated under the "covmcmc" mode → 2 BayPass analyses needed
  1. core model (default) to estimate $\boldsymbol{\Omega}$ (Bonus : provides BF_{IS} for GEA)

```
g_baypass -countdatafile ac.data -efile cov.data -outprefix anacore -nthreads 4
```

  2. "covmcmc" mode for accurate estimates of the β_i's

```
g_baypass -countdatafile ac.data -efile cov.data -covmcmc
          -omegafile anacore_mat_omega.out -outprefix anacvmc -nthreads 4
```

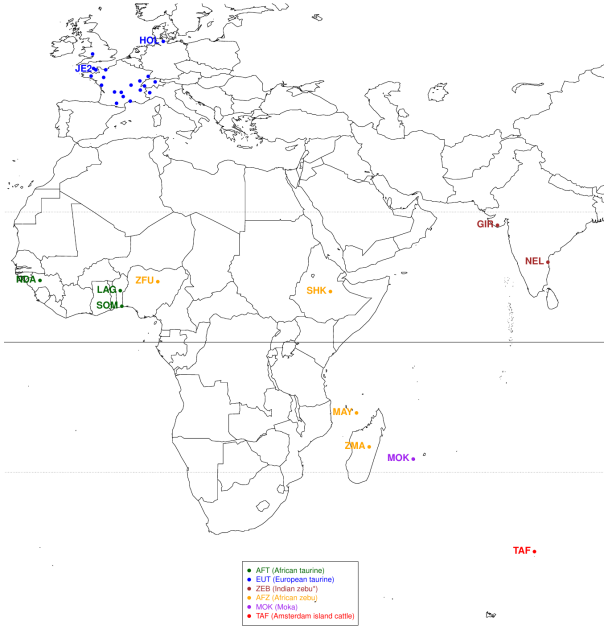
- In R with `compute_genetic_offset()` function (BayPass R utils)

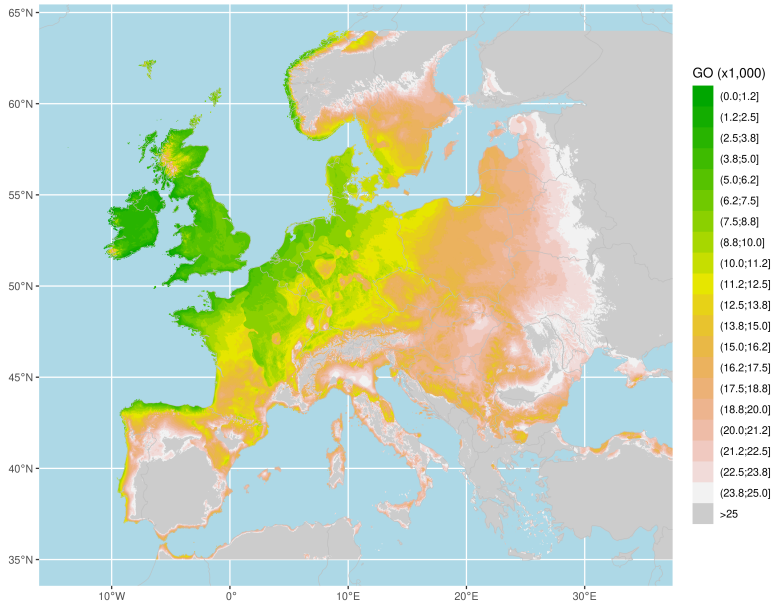
```
source("baypass/utils/baypass_utils.R")
res<-compute_genetic_offset(
   regfile="anacvmc_summary_betai.out", #BayPass output file
   covfile="cov.data", # orig. cov. file (for scaling and default ref. env): ncov x npop
   newenv=read.table("new.env"), #target env. (format=BayPass cov file): ncov x ntgt
   refenv=NULL) #use if you want to consider other ref. env other than original ones
```

  - `res$go` : here a npop × ntgt matrix of GO (for each orig. pop in each target env)
  - `res$covimp` : importance of each covariable ($\tau_k = \sum_{p=1}^{n_p} \lambda_p u_{kp}^2$; Gain et al., 2023)
  - If `compute.rona=T` ; `res$rona` gives RONA $= \frac{1}{n_{snp}} \sum_{s=1}^{n_{snp}} \mid \boldsymbol{b_s} \left(\boldsymbol{e} - \boldsymbol{e^\star} \right) \mid$ (Rellstab et al., 2016)

GO suggests some "pre-adaptation" of Amsterdam Island bovine ancestor (Gautier et al., MBE, 2024)

- 1871 : A farmer (from La Réunion) abandons 6 bovines on the isolated (58 km^2) subantarctic island of Amsterdam (very harsh environment)
- The cattle become feral (no active management) and grow up to 2,000 inds in 1953





- Estimating GO (∼"maldaptation") w.r.t. to Amsterdam Island (19 env. cov ; 31 cattle breeds)
- Reverse GO suggests breeds originating from North-Western Europe (main TAF ancestry) may have been somewhat pre-adapted

Conclusions and Future directions

Linear models : not as trendy as AI but still useful !

- Flexible, robust (to non-linearity)
- Competitive esp. with limited number of pop. samples (bias-variance trade-off)

Why bother with (old-school) Bayesian modeling as in BayPass ?

- Versatility makes it easy (but more computationally expensive) to account for
  - neutral structuring of genetic diversity (demographic history)
  - unbalanced designs, missing or noisy (e.g., Pool-Seq, pop. covariables) data
  - combined data sets *(Pool-Seq + Ind-Seq GL + count data since BayPass 3.0)*

Some perspectives

- Modeling uncertainty of the pop. covariates :
  full uncertainty ⇒ Genomic Prediction (*hidden BayPass option* ; Gautier, in prep.)
- Extend GEA/pGWAS/GP to categorical covariables (e.g., host plant)
- Accelerate Estimation (subsampling, HMC, VB)

BayPass in practice : decision table

Analysis	**Recommended models**	**Options**
Genome scan for adaptive differentiation	`core model` $\rightarrow$ XtX	*default*
Genome scan for association	`core model` $\rightarrow$ BF_{is} *(with binary cov.* $\rightarrow$ C_2 `aux. model` $\rightarrow$ BF_{mc}	*default* *-contrastfile)* `-auxmodel -omegafile`
Genetic Offset estimation (using baypass R utils)	`core model` $\rightarrow$ $\widehat{\Omega}$ + `cov. model` $\rightarrow$ $\widehat{\beta}$'s (accurate)	*default* `-covmcmc -omegafile`

Thank You For Your Attention

For more information, see the BayPass *GitLab public repo. :*

`https://forge.inrae.fr/mathieu.gautier/baypass_public`