Genome-wide association and prediction at the population level using Bayesian hierarchical models

Mathieu Gautier

UMR INRAE/CIRAD/IRD/SupAgro CBGP

23rd January 2025

Introduction

Forces driving the evolution of genetic diversity in populations

- Mutation : generates variability
- Drift : introduces stochasticity (Finite Population Size)
- Migration (gene flow)
- Selection

Different Influences of the evolutionary forces

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

Introduction

General assumption

- Diversity (pop. allele freq.) at loci underlying (genetic) adaptation of populations co-vary with fitness-related traits (but see Lotterhos, 2022)

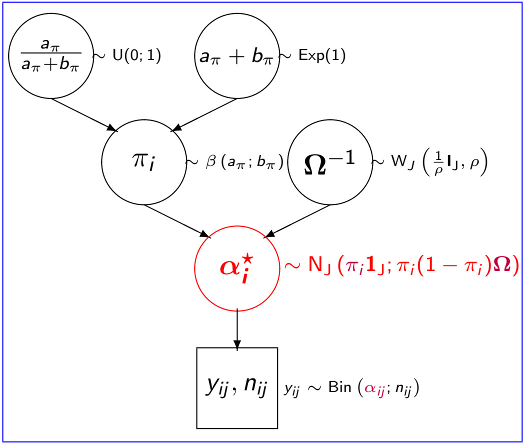
Genome-wide association with population-specific covariables

- 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
  - predicting covariate value from genomic information
- Different covariables of interest
  - Environmental (e.g., bioclimatic covariates, host plant, etc.) ⇒ GEA
  - Phenotypic (e.g., mean height, mean weight, coat color) ⇒ "pGWAS"

Demographic history : a critical confounding factor

- Shared population history ⇒ covariance structure of allele freq.

The BayPass core model (Gautier, 2015)


$\frac{a_\pi}{a_\pi+b_\pi} \sim \mathrm{U}(0;1)$
$a_\pi + b_\pi \sim \mathrm{Exp}(1)$
$\pi_i \sim \beta\left(a_\pi; b_\pi\right)$
$\boldsymbol{\Omega}^{-1} \sim \mathrm{W}_J\left(\frac{1}{\rho}\mathrm{I}_J, \rho\right)$
$\alpha_i^\star \sim \mathrm{N}_J\left(\pi_i \mathbf{1}_J; \pi_i(1-\pi_i)\boldsymbol{\Omega}\right)$
y_{ij}, n_{ij}
$y_{ij} \sim \mathrm{Bin}\left(\alpha_{ij}; n_{ij}\right)$

- Multivariate Gaussian prior on pop. (reference) allele frequencies (see Coop et al., 2010) of the I SNPs on 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 might be interpreted as the "ancestral" ref. allele freq. of SNP i
- $\boldsymbol{\Omega} = J \times J$ scaled covariance matrix of allele freq.

- $\boldsymbol{\Omega} \Leftrightarrow$ "population relationship matrix" (captures the global effect of the demography)
- Scaled allele frequencies (i.e., corrected for pop. demographic history) :

$$\mathbf{X}_i = \{\widetilde{\alpha}_{ij}\}_{1..J} = \boldsymbol{\Gamma}^{-1}\frac{\alpha_i^\star - \pi_i}{\sqrt{\pi_i(1-\pi_i)}}$$ with $\boldsymbol{\Omega} = \boldsymbol{\Gamma}'\boldsymbol{\Gamma}$ (Guenther & Coop, 2013 ; Olazcuaga et al., 2020)

BayPass models for association studies (GEA/pGWAS)

General Principles

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

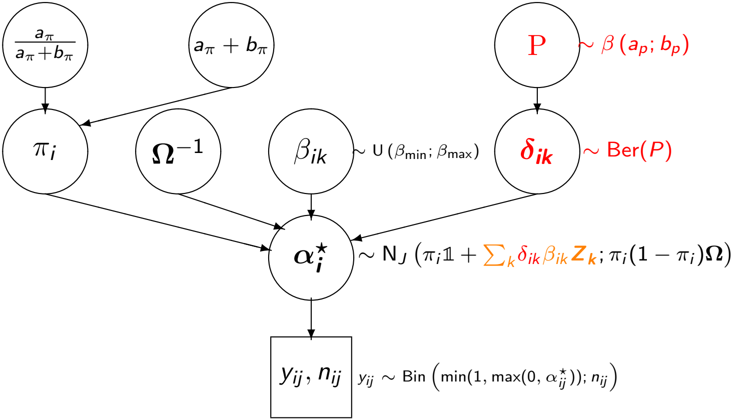
$$\widetilde{\alpha}_{ij} = \sum_{k=1}^{K} \beta_{ik} z_{jk} + \epsilon_{ij} \text{ with } \epsilon_{ij} \sim \mathrm{N}\left(0,1\right)$$

- 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

In BayPass : 3 procedures to estimate the β_i's and/or BF's

- From $\widetilde{\alpha}_{ij}$'s sampled under the core model with MCMC :
  - Importance Sampling approximation of the β_i's and BF
  - "quick and dirty" and $\Leftrightarrow$ univariate regression on each covariable in turn
- MCMC sampling of the β_i's $\Rightarrow$ accurately estimated but decision harder
- Penalized regression $\Rightarrow$ BF estimation (but some β_i's shrinked towards 0)

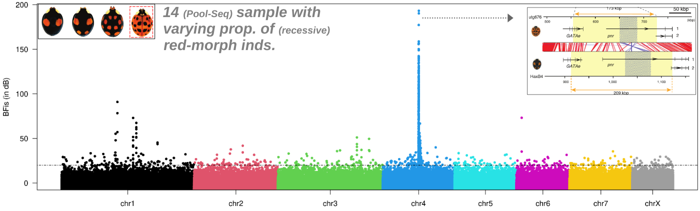
The “AUX” covariate model (i.e., with ‘auxiliary variable’)


$\frac{a_\pi}{a_\pi+b_\pi}$
$a_\pi + b_\pi$
P
$\sim \beta\left(a_P; b_P\right)$
π_i
Ω^{-1}
β_{ik}
$\sim \mathrm{U}\left(\beta_{\min}; \beta_{\max}\right)$
δ_{ik}
$\sim \mathrm{Ber}(P)$
$\alpha_i^\star$
$\sim \mathrm{N}_J\left(\pi_i \mathbb{1} + \sum_k \delta_{ik}\beta_{ik} \boldsymbol{Z}_k; \pi_i(1-\pi_i)\boldsymbol{\Omega}\right)$
y_{ij}, n_{ij}
$y_{ij} \sim \mathrm{Bin}\left(\min(1, \max(0, \alpha_{ij}^\star)); n_{ij}\right)$

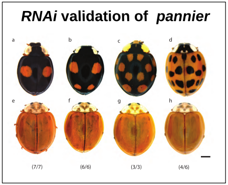
- The binary variable δ_i specifies whether the SNP is associated ($\delta_i = 1$) or not ($\delta_i = 0$)
- Integrating over P (prop. of associated SNPs) allows dealing with multiple testing issues
- From $\mathrm{P}[\delta_i = 1|data]$ (a.k.a. PIP), $\mathbb{BF}_{\mathrm{mc}} = \frac{\text{Post. odds}}{\text{Prior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

Example of applications

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

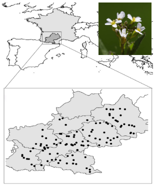


14 (Pool-Seq) sample with varying prop. of (recessive) red-morph inds.



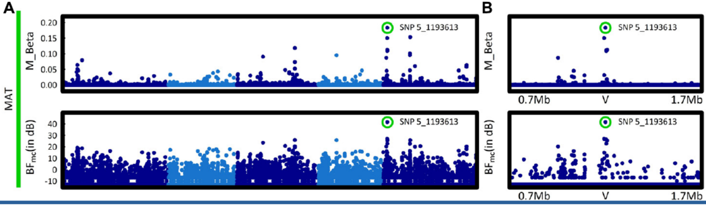
RNAi* validation of *pannier

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
  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 :

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

- $\tilde{\boldsymbol{e}} = \boldsymbol{B}\boldsymbol{e} = \left\{\sum_{k=1}^{K} \beta_{ik} e_k\right\}_i$ is the n_{snp}-length vector of global effect of environment on genetic div. at each SNP (NB : $\tilde{e}_i = 0$ if SNP i is "neutral")
- GO ⇔ (squared) euclidean environmental distance ("genetically") weighted by the env. effect on adaptive genetic diversity)

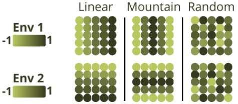
Properties of (geometric) GO

- $GO \propto -\log\left(w\left(x, x^\star\right)\right)$ where $w\left(x, x^\star\right) < 1$ is the relative fitness value of traits at equilibrium in e when placed in $e^\star$
- Supported by simulated and empirical data (e.g., Laruson et al., 2022, Gain et al., 2023)

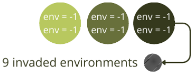
GO to predict population invasiveness (Camus et al., 2024)

Simulation Study

3 "native" environment grid (5 x 5 pop.) with **two environmental variables,** polygenic local adaptation during 3000 generations. → **10 individuals** from a source population are randomly chosen to **invade a new environment.** → Each invasion is replicated 50 times under a non-WF model

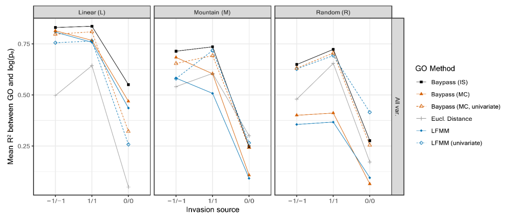


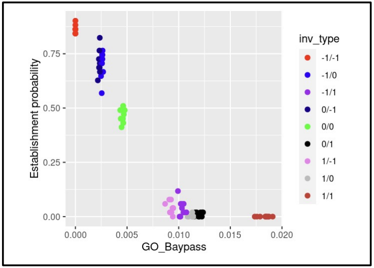
3 possible source populations :



9 invaded environments

$$PE = \frac{n_{establishment}}{n_{replicates}}$$





Predicting species invasiveness with genomic data: Is genomic offset related to establishment probability? Evolutionary Applications

Genomic prediction of population covariate

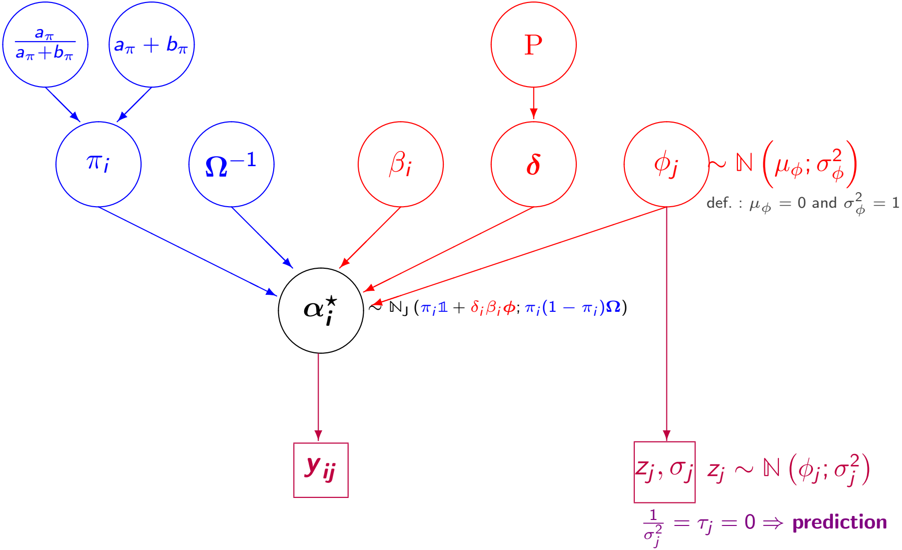
Rationale

- Rely on GEA modeling of the relationship between genetic and covariate variation across populations to estimate population covariate values
  ⇒ pop-specific covariate is treated as a random variable
- Interpretation : pop. mean phenotype or tolerance range (e.g., for env. covariable)

Extending the BayPass model for genomic prediction

- Modeling uncertainty of the population covariate values
- full uncertainty ⇒ prediction

The ‘AUX’ genomic prediction model (univariate case)



$\sim \mathbb{N}\left(\mu_\phi; \sigma_\phi^2\right)$

def. : $\mu_\phi = 0$ and $\sigma_\phi^2 = 1$

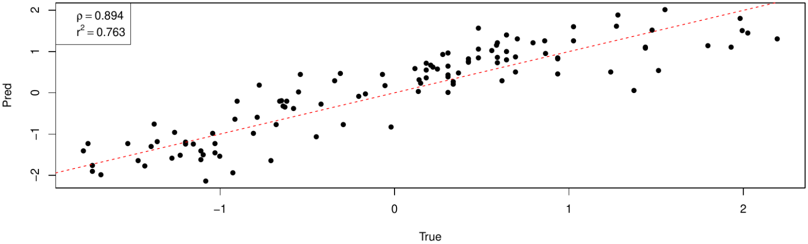
$\sim \mathbb{N}_J\left(\pi_i \mathbb{1} + \delta_i \beta_i \boldsymbol{\phi}; \pi_i(1-\pi_i)\boldsymbol{\Omega}\right)$

$z_j \sim \mathbb{N}\left(\phi_j; \sigma_j^2\right)$

$\frac{1}{\sigma_j^2} = \tau_j = 0 \Rightarrow$ **prediction**

Empirical evaluation : dog breeds weight (Gautier, in prep)

- Data (Hayward et al., 2016)
  - Genotypes : 155,609 SNPs genotyped on 111 dog breeds (n=6–636)
  - Phenotypes : mean male weight of each breed (American Kennel Club)
- 'Leave-one out' analysis (1 predicted pheno. vs. 110 known ±0.01)



Conclusions

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 data, additional source of variation (e.g., Pool-Seq, pop. covariables)
  - combined data sets *(Pool-Seq + Ind-Seq GL + count data in* BayPass *3.0)*
- Yet, urgent need to accelerate MCMC (subsampling, HMC)

Predictive approaches are promising but still need

- Further evaluation on real (e.g., *D. melanogaster*) and simulated data (SLiM)
  - GO : robustness to genetic architecture, demographic history (e.g., admixture), genetic load, etc.
  - Genomic Prediction : sensitivity to the nb. of SNPs (LD), genetic architecture, etc.
- New developments esp. for (pop-level) genomic prediction :
  - BayPass : extend to categorical variable (e.g., fruit) ; multivariate GP
  - Comp. with other (machine/deep learning) approaches (e.g., Random Forest or CNN)

Acknowledgements

CBGP

Louise Camus

Simon Boitard

Maria Bogaerts-Marquez

Laure Olazcuaga

Nicolas Rode

Arnaud Estoup

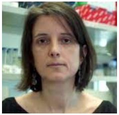
Anne Loiseau

IBDML *(CNRS, Marseille)*

Benjamin Prud'homme

Junichi Yamaguchi

IBE *(Barcelona)*; DrosEU



Josefa Gonzalez

LIPME *(INRAE, Toulouse)*

Fabrice Roux

Léa Frachon

… and you for your attention