Genomic prediction

Andrea Rau

INRAE

Genomic prediction

Andrea Rau

Université Paris Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France

November 30, 2023

Outline

Installing packages

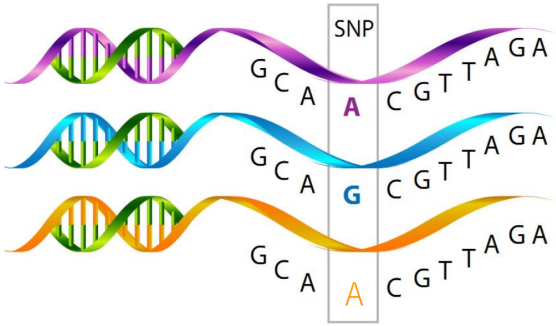
```
install.packages(c("glmnet", "BGLR", "tidyverse"))

library(glmnet)
library(BGLR)
library(tidyverse)
```

Aims for this class

- What is genomic prediction, and how is it used in agriculture and human health?
- What are some of the statistical challenges related to genomic prediction models?
- What models have been proposed to address these challenges, and what are their advantages/limitations?
- How are genomic prediction models evaluated?

Outline

Genomic information



http://neuroendoimmune.files.wordpress.com/2014/03/snp.png

- Mutation $<$ 1% $<$ Single nucleotide polymorphism (SNP)
- Construct genetic relationships, parentage determination, identification of quantiative trait loci (QTL), ...

Prediction of phenotypes from genotypes

Goal: given a training set of data (Y_i, X_i, Z_i) for $i = 1, \ldots, n$ individuals

- Y_i = phenotype
- X_i = vector of (usually genome-wide) genotypes
- Z_i = vector of covariates (age, location, sex, ...)

... predict the unobserved phenotype $Y_\star$ of a future individual with corresponding $X_\star$ and $Z_\star$

Prediction of phenotypes from genotypes

Goal: given a training set of data (Y_i, X_i, Z_i) for $i = 1, \ldots, n$ individuals

- Y_i = phenotype
- X_i = vector of (usually genome-wide) genotypes
- Z_i = vector of covariates (age, location, sex, ...)

... predict the unobserved phenotype $Y_\star$ of a future individual with corresponding $X_\star$ and $Z_\star$

Why?

- **Genomic selection** in plant/animal breeding: select individuals to mate or carry forward in breeding programs
- **Health care**: identify high-risk individuals for interventions/treatments/preventative care

Variable selection to prediction in genetics

For humans and plants/animals, shift in genetics studies from model selection (identifying associated genetic variants) to prediction (choosing optimal interventions and improving selection)

Genomic selection:

- Introduced by Meuwissen *et al.* (2001), successfully implemented in many plant/animal breeds for traits related to production, health, climate adaptation, ...
- Modest gains in predictions can have large economic impacts (reduced generation interval, reduced cost and labor for phenotyping)

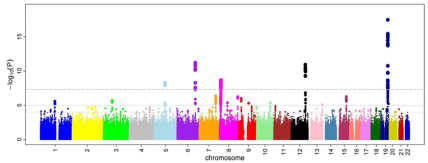
Human health:

- Less successful (need very high predictive accuracy to inform clinical decisions) but holds some promise for calculating risk scores

Variable selection versus prediction

Variable selection:

- Stringent multiple-testing corrections for genome-wide significance in GWAS



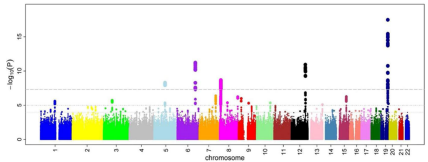
https://commons.wikimedia.org/wiki/File:Manhattan_Plot.png

Variable selection versus prediction

INRAE
science for people, life & earth

Variable selection:

- Stringent multiple-testing corrections for genome-wide significance in GWAS



https://commons.wikimedia.org/wiki/File:Manhattan_Plot.png

Prediction:

- Complex traits controlled by many genes with small effects + influenced by environments
- Little negative impact including (some...) uninformative variables
- Inference of average effects of allele substitution + variance components

Outline

Genotyping data

- Most abundant polymorphisms at DNA level are Single Nucleotide Polymorphisms (SNP)
  - About $\sim$ 3 billion nucleotides in the cattle genome, with over 30 million SNPs (introns, exons, promoters, enhancers, intergeneic regions, ...)
  - High-throughpout genotyping becoming cheaper (thousands of SNPS $\rightarrow$ 10k - 100k SNPs $\rightarrow$ whole genome sequencing
- Now possible to massively + accurately + economically read the same set of (biallelic) SNPs across several individuals $\rightarrow$ genotyping via SNP chips or whole genome sequencing
  - Possible alleles for SNP loci are all pairwise combinations among (A,C,G,T): A/C, A/G, A/T, C/G, C/T, G/T

Genotyping data

Image courtesy of Goto Morota (http://morotalab.org/guestlectures/2020/FREC5164-2020/FREC5164-2020.html)

Raw SNP genotyping file

```
[Header]
GSGT Version 1.9.4
Processing Date 3/16/2012 9:11 AM
Content OvineSNP50_B.bpm
Num SNPs 54241
Total SNPs 54241
Num Samples 36
Total Samples 36
[Data]
Sample ID Sample Name SNP Name Allele1 - Top Allele2 - Top GC Score
ES140000270478 PLACA_CIC_12_96 250506CS3900065000002_1238.1 G G 0.8932
ES140000270478 PLACA_CIC_12_96 250506CS3900140500001_312.1 A G 0.7341
ES140000270478 PLACA_CIC_12_96 250506CS3900176800001_906.1 A G 0.7532
ES140000270478 PLACA_CIC_12_96 250506CS3900211600001_1041.1 A A 0.9674
ES140000270478 PLACA_CIC_12_96 250506CS3900218700001_1294.1 G G 0.8178
ES140000270478 PLACA_CIC_12_96 250506CS3900283200001_442.1 C C 0.6684
ES140000270478 PLACA_CIC_12_96 250506CS3900371000001_1255.1 G G 0.4565
ES140000270478 PLACA_CIC_12_96 250506CS3900386000001_696.1 A A 0.4258
ES140000270478 PLACA_CIC_12_96 250506CS3900414400001_1178.1 G G 0.8690
ES140000270478 PLACA_CIC_12_96 250506CS3900435700001_1658.1 A A 0.5153
ES140000270478 PLACA_CIC_12_96 250506CS3900464100001_519.1 A G 0.8116
ES140000270478 PLACA_CIC_12_96 250506CS3900487100001_1521.1 A G 0.7448
ES140000270478 PLACA_CIC_12_96 250506CS3900539000001_471.1 G G 0.5248
```

SNP genotyping in condensed format

- map file: names of all markers and position (chromosome, bp)

```
1 F0100190 0 135098 2 1
1 TPM87 0 264710 2 1
1 TPM951 0 264740 1 2
1 F0100220 0 267940 1 2
1 RGX1000 0 349826 2 1
1 RGX2000 0 351236 2 1
```

- genotype file:

```
ES1400NAB40571 G G G G A A A C . . A G
ES1400NAB40573 G G G G G G A C G G A G
ES1400NAB40574 A G G G A G A C G G A A
ES1400NAB40159 G G G G A G A C G G A A
ES1400NAB40528 A G A G A G C C A G A A
ES1500VI492705 G G A G G G A C G G A G
ES1500SSA40533 A G G G A G C C G G A A
```

→ PLINK (https://www.cog-genomics.org/plink): .bed, .bim, .fam files

Considerations for genotyping data

- Typically recoded as number of copies of the minor allele (0, 1, or 2)
- Minor allele frequency (MAF) = frequency of the reference allele
- Call rate = number of observed genotypes (per individual, per marker)
- Linkage disequilibrium (LD): non-random association between alleles at different loci

$$\text{LD}_{k\ell} = \frac{Cov(\mathbf{x}_k, \mathbf{x}_\ell)^2}{Var(Cov(\mathbf{x}_k)Var(\mathbf{x}_\ell)} = \frac{(p_{ij} - p_i p_j)^2}{p_i(1-p_i)p_j(1-p_j)}$$

with p_{ij} the frequency of haplotype ij, p_i the frequency of allele i at locus k, and p_j the frequency of allele j at locus ℓ

- Imputation of missing genoypes (marginal allele distribution, full-sib family information)

$\rightarrow$ Typically filters applied on MAF, missing values, LD, ...

Illustration of LD

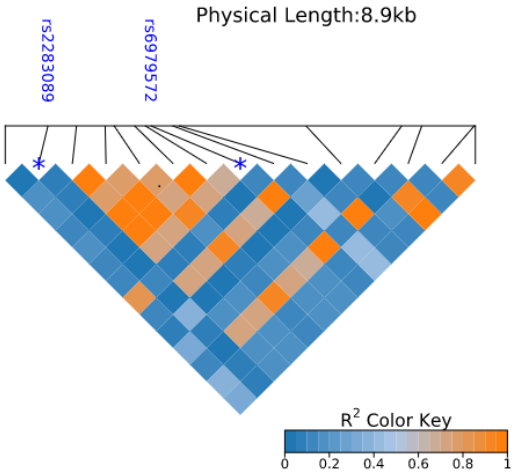


Image: https://www.bioinformatics.com.cn/plot_basic_LDheatmap_plot_094_en

CIMMYT Global Wheat Program

INRAE
science for people, life & earth

International Maize and Wheat Improvement Center (https://www.cimmyt.org): international organization for non-profit ag research + training

- Collection of n = 599 historical wheat lines from the CIMMYT Global Wheat Program from 4 main agroclimatic regions
  - Genotyping using 1447 Diversity Array Technology (DArT, https://www.diversityarrays.com)
  - Inbred lines $\Rightarrow$ DArT markers only take two values (presence/absence)
  - Pre-processing: filter MAF < 0.05, imputed missing genotypes
- Phenotype of interest = average grain yield

Outline

The linear model of genomic prediction

The workhorse of genomic prediction is the multiple linear regression model:

$$Y = \mathbf{Z}\theta + \mathbf{X}\beta + \varepsilon$$

- Y = n-vector of phenotypes
- $\mathbf{Z}$ = $n \times m$ matrix of covariates
- θ = m-vector of covariate effect parameters
- $\mathbf{X}$ = $n \times p$ matrix of (suitably coded) genotypes
- β = p-vector of genetic effect parameters
- ε = n-vector of errors representing noise, assumed iid and (usually) normally distributed

Modeling assumptions

- Most often only model additive and linear genetic effects and ignore dominance and epistasis
- Independence of ε assumes that kinship effects are accounted for through genetic markers

Covariates are often very important in prediction, but from now on we will ignore them to focus on prediction from genomic data alone...

Dimensionality of predictors

Many more variants p ($\sim$ 10k-1M) than individuals n ($\sim$ 1k) $\rightarrow p \gg n$!

- Including only significant GWAS hits usually leads to poor prediction: polygenic nature of complex traits, conservative testing thresholds, ...
- ... but including too many predictors in a model risks over-fitting and poor generalizability + non-existant ordinary least squares solution

Addressing dimensionality of predictors

Genomic best linear unbiased prediction (GBLUP):

$$Y = \mathbf{g} + \varepsilon, \quad \text{where} \quad \mathbf{g} \sim N(0, \mathbf{G}\sigma_g^2)$$

approximated by $Y = \mathbf{X}\beta + \varepsilon$

Variance-covariance matrix of Y is $\mathbf{V}_y = \mathbf{V}_g + \mathbf{V}_\varepsilon = \mathbf{X}\mathbf{X}'\sigma_a^2 + \mathbf{I}\sigma_\varepsilon^2$

- $\beta \sim N(0, \mathbf{I}\sigma_a^2)$, $\varepsilon \sim N(0, \mathbf{I}\sigma_\varepsilon^2)$
- Conditional mean of $\mathbf{g}$ given the data is extremely computationally efficient: $\text{BLUP}(\hat{\beta}) = \left(\mathbf{I} + (\mathbf{X}\mathbf{X}')^{-1}\frac{\sigma_\varepsilon^2}{\sigma_a^2}\right)^{-1} Y$

Outline

Penalization for curse of dimensionality

Many more variants p ($\sim$ 10k-1M) than individuals n ($\sim$ 1k) $\rightarrow p \gg n$!
$\Rightarrow$ Another solution is to use a penalized regression

- Penalty in the residual sum of squarers or log-likelihood "shrinks" parameter estimates towards 0
- Form of penalty function can be evaluated in terms of performance on test data (e.g., cross-validation predictive correlation or predictive log-likelihood)
- Bayesian framework for penalty to reflect known information about the distribution of variant effect sizes (prior distribution)

Penalized regression models

1. Ridge regression
2. Lasso regression
3. Elastic net regression
4. Partial least squares (PLS) regression
5. Bayesian methods

Ridge regression (RR)

- Maximum penalized likelihood approach with an independent mean-0 Gaussian prior on each genetic effect:

$$\hat{\beta}_{\text{ridge}} = \arg\min \beta \Big\{ \sum_{i=1}^{n} \varepsilon_i^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \Big\} \qquad \varepsilon_i = Y_i - \sum_{j=1}^{p} X_{ij}\beta_j$$

- Equivalent to a best linear unbiased predictor (BLUP) in a mixed model with allelic-correlation kinships computed from marker genotypes $\rightarrow$ in BLUP, λ is estimated from the data while in RR λ often treated as a tuning parameter

Lasso regression

- Similar to RR, but assumes a Laplace (double exponential) penalty on genetic effects, equivalent to a lienar term in the log-likelihoo (i.e. L_1 versus L_2 penalty):

$$\hat{\beta}_{\text{lasso}} = \arg\min \beta \Big\{ \sum_{i=1}^{n} \varepsilon_i^2 + \lambda \sum_{j=1}^{p} |\beta_j| \Big\}$$

- Due to sharp peak of Laplace distribution at 0, many genetic effects will be estimated at 0 $\Rightarrow$ model selection + prediction
- Note: number of non-zero effects constrained to be $\leq n$...
- In regions of high LD, typically only 1 SNP has a nonzero $\hat{\beta}_j$
- Many extensions: Bayesian lasso, HyperLasso, ...

Elastic net regression

- Combines RR and Lasso by weighting their penalities:

$$\hat{\beta}_{\text{enet}} = \arg\min \beta \Big\{ \sum_{i=1}^{n} \varepsilon_i^2 + \lambda \sum_{j=1}^{p} \alpha \beta_j^2 + (1-\alpha)|\beta_j| \Big\}$$

- Selects variables like Lasso + shrinks together coefficients of correlated predictors like RR
- Tuning (α, λ) can be time consuming when performed on a grid

PLS regression

- PLS identifies orthogonal linear combinations of genotypes $\mathbf{w}_1, \ldots, \mathbf{w}_k$ that maximize correlation with phenotype (rather than variance as in PCA) that are used as predictors:

$$\hat{\mathbf{b}}_{\text{pls}} = \arg\min \mathbf{b} \Big\{ \sum_{i=1}^{n} \left(Y_i - \mu - \sum_{j=1}^{k} w_{ij} b_j \right)^2 \Big\}$$

- Dimension reduction while including all indvidual SNPs as predictors (no need for a penalty, single parameter k to tune)
- ... but no estimates of individual genetic effects β

Outline

Bayesian methods

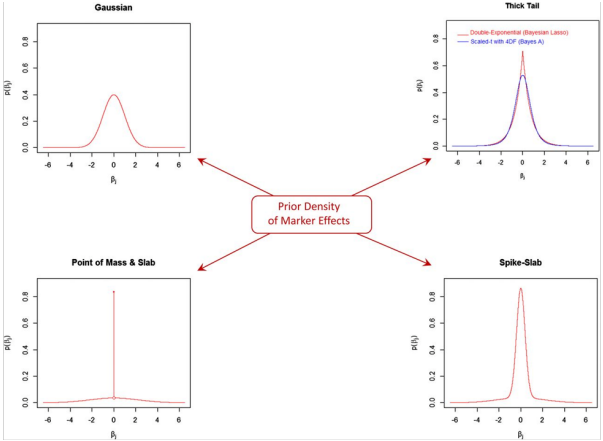
Bayesian models often have the form:

$$\prod_{i=1}^{n} N\left(Y_i \middle| \left(\mu + \sum_{j=1}^{p} X_{ij}\beta_j\right), \sigma^2\right) \quad \times \quad p(\sigma^2) \prod_{j=1}^{p} p(\beta_j|\Psi)$$

$$\text{likelihood} \quad \times \quad \text{prior}$$

- Ψ = vector of hyperparameters to specify the prior $\rightarrow$ can be fixed, integrated out with respect to a prior (fully Bayesian), or estimated from the data (empirical Bayes)
- σ^2 often assigned a $\chi^{-2}(\nu, S)$ prior distribution
- Gaussian prior for $\beta \Rightarrow$ posterior means are GBLUP estimates, Laplace prior for $\beta \Rightarrow$ Bayesian lasso

Prior distributions for Bayesian methods



Genetics, Volume 193, Issue 2, 1 February 2013, Pages 327–345, https://doi.org/10.1534/genetics.112.143313

Which prior to use?

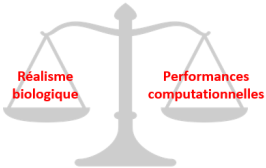


Image courtesy of Fanny Mollandin

Which prior to use?

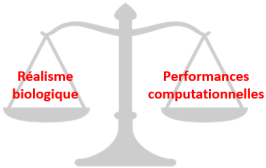


Image courtesy of Fanny Mollandin

- GBLUP: $\beta_i \sim N(0, \sigma^2_\beta)\ \forall i$

Which prior to use?

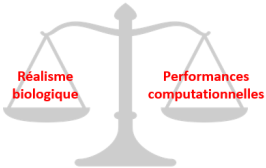


Image courtesy of Fanny Mollandin

- GBLUP: $\beta_i \sim N(0, \sigma^2_\beta)\ \forall i$
- BayesA: $\beta_i \sim N(0, \sigma^2_{\beta_i}), \sigma^2_{\beta_i} \sim \text{Inv } \chi^2(\nu, S^2)\ \forall i$

Which prior to use?

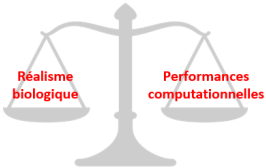


Image courtesy of Fanny Mollandin

- GBLUP: $\beta_i \sim N(0, \sigma^2_\beta)\ \forall i$
- BayesA: $\beta_i \sim N(0, \sigma^2_{\beta_i}), \sigma^2_{\beta_i} \sim \text{Inv } \chi^2(\nu, S^2)\ \forall i$
- BayesB: $\beta_i \sim N(0, \sigma^2_{\beta_i}), \sigma^2_{\beta_i} \sim \pi\delta(0) + (1-\pi)\text{Inv } \chi^2(\nu, S^2)\ \forall i$, π known

Which prior to use?

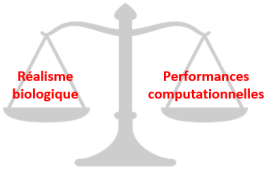


Image courtesy of Fanny Mollandin

- GBLUP: $\beta_i \sim N(0, \sigma^2_\beta)\ \forall i$
- BayesA: $\beta_i \sim N(0, \sigma^2_{\beta_i}), \sigma^2_{\beta_i} \sim \text{Inv } \chi^2(\nu, S^2)\ \forall i$
- BayesB: $\beta_i \sim N(0, \sigma^2_{\beta_i}), \sigma^2_{\beta_i} \sim \pi\delta(0) + (1-\pi)\text{Inv } \chi^2(\nu, S^2)\ \forall i$, π known
- BayesC: $\beta_i \sim \pi\delta(0) + (1-\pi)N(0, \sigma^2_\beta), \sigma^2_\beta \sim \text{Inv } \chi^2(\nu, S^2)\ \forall i$, π known

Which prior to use?

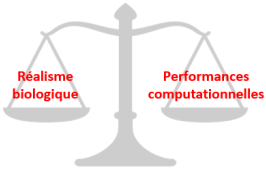


Image courtesy of Fanny Mollandin

- GBLUP: $\beta_i \sim N(0, \sigma^2_\beta)\ \forall i$
- BayesA: $\beta_i \sim N(0, \sigma^2_{\beta_i}), \sigma^2_{\beta_i} \sim \text{Inv}\ \chi^2(\nu, S^2)\ \forall i$
- BayesB: $\beta_i \sim N(0, \sigma^2_{\beta_i}), \sigma^2_{\beta_i} \sim \pi\delta(0) + (1-\pi)\text{Inv}\ \chi^2(\nu, S^2)\ \forall i$, π known
- BayesC: $\beta_i \sim \pi\delta(0) + (1-\pi)N(0, \sigma^2_\beta), \sigma^2_\beta \sim \text{Inv}\ \chi^2(\nu, S^2)\ \forall i$, π known
- BayesCπ: BayesC with $\pi \sim \text{Unif}(0, 1)$

Many other genomic prediction approaches...

- Random forest
- Neural networks
- Reproducing kernel Hilbert spaces
- Adaptive MultiBLUP (flexible shrinkage for promising genomic regions)
- Rank-based model averaging (minimize prediction errors made by a specific method, capture different kinds of genetic effects, ...)

Many other genomic prediction approaches...

- Random forest
- Neural networks
- Reproducing kernel Hilbert spaces
- Adaptive MultiBLUP (flexible shrinkage for promising genomic regions)
- Rank-based model averaging (minimize prediction errors made by a specific method, capture different kinds of genetic effects, ...)

In general, prediction accuracy depends on many factors:

- Size of training sample n
- Trait heritability
- Number of loci affecting the trait
- Genetic relatedness between training and test samples

How to measure success?

After fitting a prediction model on training data, we can measure success on independent test data with available phenotypes:

- Independent test dataset
- Withhold a random fraction of samples (say 10%) from training data $\rightarrow$ but test individuals are similar to training individuals, which may lead to inflated predictive accuracy with respect to future individuals
- Cross-validation to withhold multiple resampled fractions of samples
- Forward validation (train on year 1 data, test on year 2 data)

Predictive accuracy for continuous traits

INRAE
science for people, life & earth

Suppose in a test sample of size k, we have predictions $\hat{Y}_1, \ldots, \hat{Y}_k$ with observed values $Y_1, \ldots, Y_k$.

Goal: the closer $\hat{Y}_i$ to Y_i, the better!
The main goal of genomic selection is to select reproducing animals or new plant variets with better values of the trait of interest.

Predictive accuracy for continuous traits

Suppose in a test sample of size k, we have predictions $\hat{Y}_1, \ldots, \hat{Y}_k$ with observed values $Y_1, \ldots, Y_k$.

Goal: the closer $\hat{Y}_i$ to Y_i, the better!
The main goal of genomic selection is to select reproducing animals or new plant variets with better values of the trait of interest.

- Pearson correlation $cor(\hat{Y}, Y)$, or squared correlation, or Spearman correlation
- Mean absolute error or the (root) mean square error: $\frac{1}{k}\sum_{i=1}^{k}|\hat{Y}_i - Y_i|$ or $\frac{1}{k}\sum_{i=1}^{k}(\hat{Y}_i - Y_i)^2$

For binary traits: sensitivity, specificity, AUROC, positive predictive value, ...

Summary: Genomic prediction

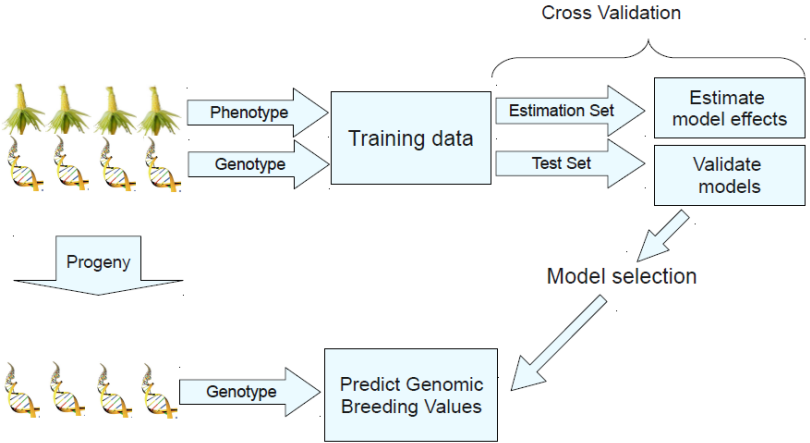


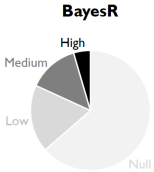
Image courtesy of Valentin Wimmer (Analysis pipeline for genomic prediction data using R and synbreed package)

Conclusions

- Genome-wide SNPs have opened the door to different ways to consider heritability and prediction
- Many genomic prediction models proposed in the literature (with different strengths + weakenesses) $\rightarrow$ different modesl may suit different trait architectures
  - Maximize over or integrate out genetic effets?
  - Prior/penalty for effect sizes?
  - Polygenic term (correlation structure or genome-wide distribution of effect sizes)

Future and ongoing work

- **(Overlapping) annotations in genomic prediction**



Genotype …0000010012010021002000101000010110010111l0…

…ACTCCGTAACTAGCCTACAAAGGCTAACTTACAAAAGATTTA…

Predict

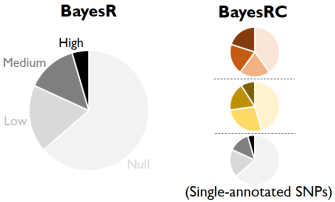
GBV

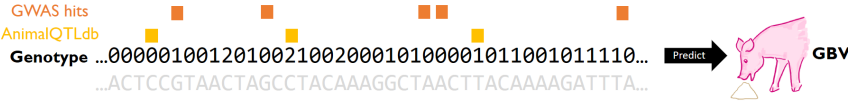
This project has received funding from the European Union's **Horizon 2020** **Research and Innovation** Programme under the grant agreement n° **817998**

GENE-SWitCH

Future and ongoing work

- (Overlapping) annotations in genomic prediction

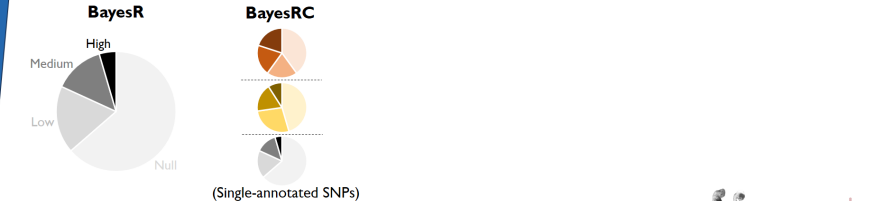




This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement n° 817998

Future and ongoing work

- **(Overlapping) annotations in genomic prediction**

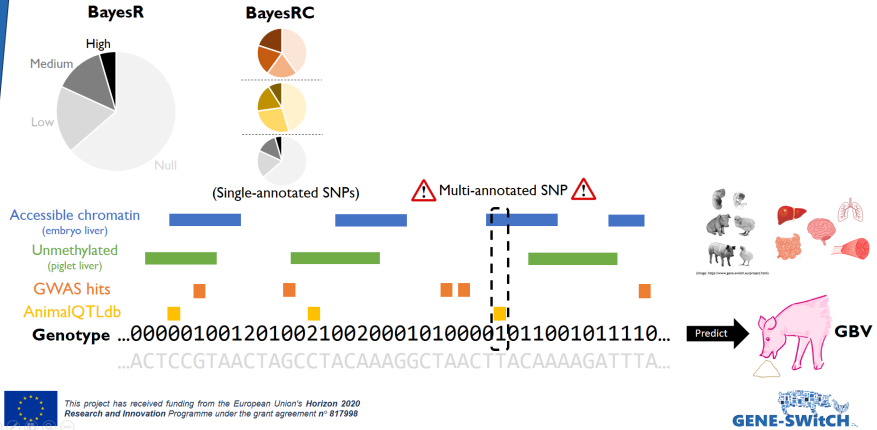


This project has received funding from the European Union's ***Horizon 2020 Research and Innovation*** *Programme under the grant agreement n° 817998*

GENE-SWitCH

Future and ongoing work

- **(Overlapping) annotations in genomic prediction**

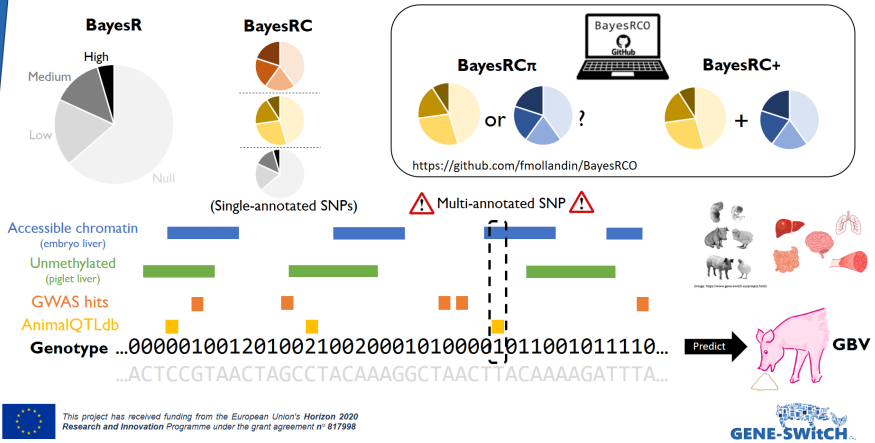


This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement n° 817998

GENE-SWitCH

Future and ongoing work

- **(Overlapping) annotations in genomic prediction**



This project has received funding from the European Union's ***Horizon 2020 Research and Innovation*** *Programme under the grant agreement n° 817998*

GENE-SWitCH

Acknowledgements & References

- Thanks to **Fanny Mollandin** (INRAE) and **Pascal Croiseau** (INRAE)

- Balding, *Introduction to Genomic Prediction* (Armidale Genetics Summer Course, 2016)
- Wray *et al.* (2013) Pitfalls of predicting complex traits from SNPs, *Nat Rev Genet* 14:507-515.
- Pérez and de los Campos (2013) Genome-wide regression and prediction with the BGLR statistical package, *Genetics* 198(2):483-495.
- Mollandin *et al.* (2022) Accounting for overlapping annotations in genomic prediction models of complex traits, *BMC Bioinformatics*, 23:65.