



Incorporating biological information into genomic prediction models

Andrea Rau, Fanny Mollandin, Pascal Croiseau

► To cite this version:

Andrea Rau, Fanny Mollandin, Pascal Croiseau. Incorporating biological information into genomic prediction models. VistaMilk Artificial Intelligence in Agriculture Masterclass, Feb 2023, Online, Ireland. hal-04173250

HAL Id: hal-04173250

<https://hal.inrae.fr/hal-04173250>

Submitted on 28 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Incorporating biological information into genomic prediction models

Fanny Mollandin, Pascal Croiseau, Andrea Rau

VistaMilk

Artificial Intelligence in Agriculture Masterclass @ Zoom

February 8, 2023





INRAE Research Center @ Jouy en Josas:

- ✓ 1500+ staff
- ✓ Animal biology, microbiology, data science, systems biology

Animal Genetics & Integrative Biology (GABI) unit

- ✓ Understanding & exploiting animal genetic variability
- ✓ Construction of phenotypes and their interaction with microbial ecosystems and environments
- ✓ Agroecological transition

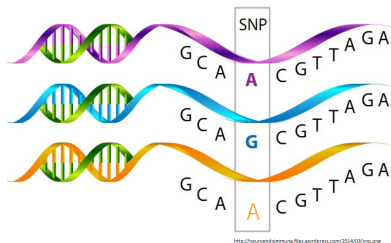


Genomic selection overview

Objective: select the best animals for reproduction to obtain **genetic improvement** of the population on **traits of interest**

Genomic selection overview

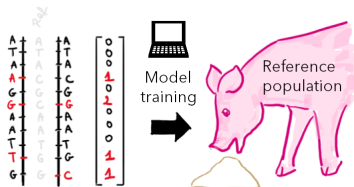
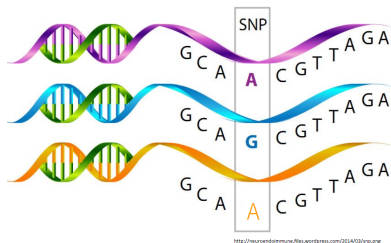
Objective: select the best animals for reproduction to obtain **genetic improvement** of the population on **traits of interest**



- Low- to high-density genotyping chips (10k-100k SNPs)
→ whole genome sequencing (10MM SNPs)

Genomic selection overview

Objective: select the best animals for reproduction to obtain **genetic improvement** of the population on **traits of interest**

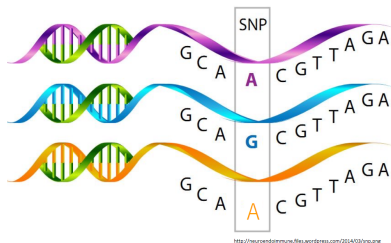


- Low- to high-density genotyping chips (10k-100k SNPs)
→ whole genome sequencing (10MM SNPs)

Image: F. Mollandin

Genomic selection overview

Objective: select the best animals for reproduction to obtain **genetic improvement** of the population on **traits of interest**



- Low- to high-density genotyping chips (10k-100k SNPs)
→ whole genome sequencing (10MM SNPs)

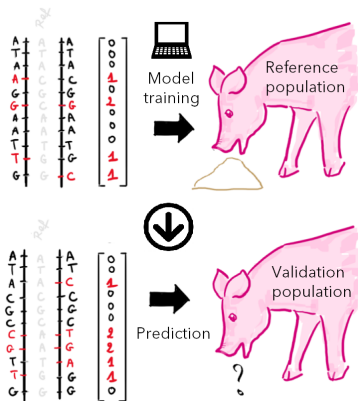


Image: F. Mollandin

Prediction models for genomic selection

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

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

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

Prediction models for genomic selection

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

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

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

- 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)

Challenges of genomic prediction models

- Non-random association between alleles at neighboring loci (aka LD)
- Polygenic nature of complex traits
- Many more SNPs (variables) than individuals (observations) \Rightarrow curse of dimensionality
 - Including too many predictors in a model risks **over-fitting**, **poor generalizability**, and **problems with model estimation**
 - ... but including only a small pre-identified subset of SNPs (e.g., significant GWAS hits) usually leads to **poor predictions**

→ Balance computational/statistical feasibility and biologically realistic assumptions

Challenges of genomic prediction models

- Non-random association between alleles at neighboring loci (aka LD)
- Polygenic nature of complex traits
- Many more SNPs (variables) than individuals (observations) \Rightarrow curse of dimensionality
 - Including too many predictors in a model risks **over-fitting**, **poor generalizability**, and **problems with model estimation**
 - ... but including only a small pre-identified subset of SNPs (e.g., significant GWAS hits) usually leads to **poor predictions**

\rightarrow Balance computational/statistical feasibility and biologically realistic assumptions

Can genomic prediction models be improved by better accounting for our **knowledge** about the **function** of certain regions of the genome?

Context: H2020 GENE-SWitCH project

The regulatory GENomE of Swine & Chicken: functional annotation during development

High-quality richly annotated maps of pig and chicken genomes:

- **Development:** early/late organogenesis, new born/hatched, adult
- **Sexes:** {M,F} \times 3 biological replicates
- **Tissues:** liver, skeletal muscle, small intestine, cerebellum, dorsal epidermis, lung, kidney
- **Assays:** RNA-seq, ATAC-seq, ChIP-seq, smRNA-seq, methylation, Hi-C



Integrate functional information with phenotypic + genotypic data in **genomic prediction models** for greater **power** and **interpretability**

Context: H2020 GENE-SWitCH project

The regulatory GENomE of Swine & Chicken: functional annotation during development

High-quality richly annotated maps of pig and chicken genomes:

- **Development:** early/late organogenesis, new born/hatched, adult
- **Sexes:** {M,F} × 3 biological replicates
- **Tissues:** liver, skeletal muscle, small intestine, cerebellum, dorsal epidermis, lung, kidney
- **Assays:** RNA-seq, ATAC-seq, ChIP-seq, smRNA-seq, methylation, Hi-C



Integrate functional information with phenotypic + genotypic data in **genomic prediction models** for greater **power** and **interpretability**

But how?

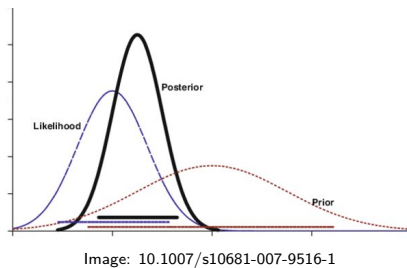
First, back to basics: the linear model

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

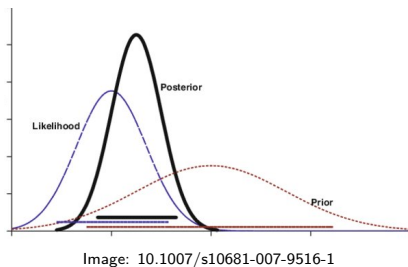
$$Y = \mathbf{Z}\theta + \mathbf{X}\beta + \varepsilon$$

- $Y = n$ -vector of traits
- $\mathbf{Z} = n \times m$ matrix of covariates
- $\theta = m$ -vector of covariate effect parameters
- $\mathbf{X} = n \times p$ matrix of (suitably coded) genotypes
- $\beta = p$ -vector of genetic effect parameters
- $\varepsilon = n$ -vector of errors representing noise, assumed to be iid and (usually) normally distributed

Bayesian methods for genomic prediction



Bayesian methods for genomic prediction



likelihood

×

prior

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

- σ^2 often assigned a χ^{-2} prior distribution
- Choice of prior for β_j should ideally reflect a trait's genetic architecture (and be computationally feasible...)

Which prior to use for β_j ?

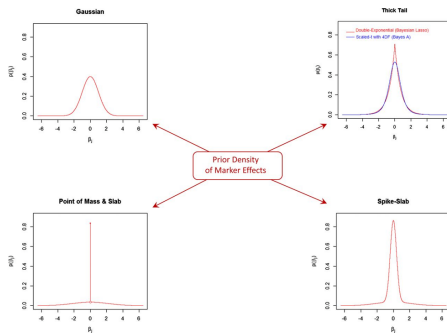


Image: 10.1543/genetics.112.143313

Which prior to use for β_j ?

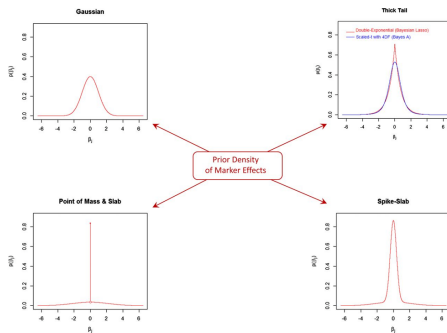


Image: 10.1543/genetics.112.143313

GBLUP: $\beta_i \sim N(0, \sigma_\beta^2)$

Which prior to use for β_j ?

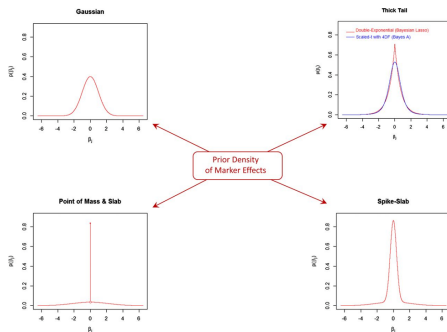


Image: 10.1543/genetics.112.143313

GBLUP: $\beta_i \sim N(0, \sigma_\beta^2)$

BayesA: $\beta_i \sim N(0, \sigma_{\beta_i}^2), \sigma_{\beta_i}^2 \sim \text{Inv } \chi^2(\nu, S^2)$

BayesB: $\beta_i \sim N(0, \sigma_{\beta_i}^2), \sigma_{\beta_i}^2 \sim \pi \delta(0) + (1 - \pi) \text{Inv } \chi^2(\nu, S^2), \pi \text{ fixed}$

Which prior to use for β_j ?

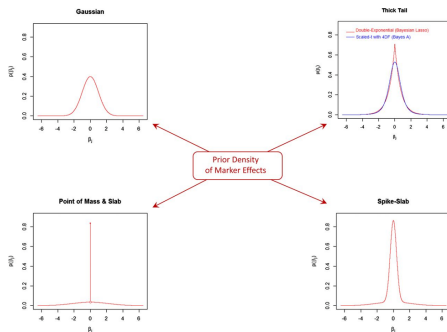


Image: 10.1543/genetics.112.143313

GBLUP: $\beta_i \sim N(0, \sigma_\beta^2)$

BayesA: $\beta_i \sim N(0, \sigma_{\beta_i}^2), \sigma_{\beta_i}^2 \sim \text{Inv } \chi^2(\nu, S^2)$

BayesB: $\beta_i \sim N(0, \sigma_{\beta_i}^2), \sigma_{\beta_i}^2 \sim \pi\delta(0) + (1 - \pi)\text{Inv } \chi^2(\nu, S^2), \pi$ fixed

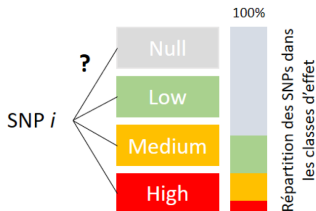
BayesC: $\beta_i \sim \pi\delta(0) + (1 - \pi)N(0, \sigma_\beta^2), \sigma_\beta^2 \sim \text{Inv } \chi^2(\nu, S^2), \pi$ fixed

BayesC π : BayesC with $\pi \sim \text{Unif}(0, 1)$

BayesR (Erbe *et al.*, 2012)

$$\beta_i \sim \underbrace{\pi_1 \delta(0)}_{\text{null}} + \underbrace{\pi_2 N(0, 0.0001\sigma_g^2)}_{\text{small}} + \underbrace{\pi_3 N(0, 0.001\sigma_g^2)}_{\text{medium}} + \underbrace{\pi_4 N(0, 0.01\sigma_g^2)}_{\text{large}}$$

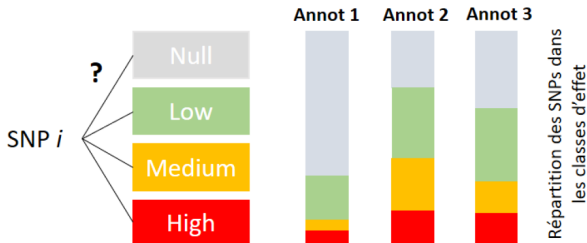
- $\pi \sim \text{Dirichlet}(\alpha)$, with $\alpha = (1, 1, 1, 1)$
- Gibbs sampler for estimation



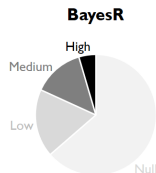
Back to annotations: BayesRC (MacLeod *et al.*, 2016)

$$f(\beta_i | C_i = c) = \sum_{k=1}^4 \pi_{c,k} f_k(\cdot | \theta_k)$$

- SNPs assigned to disjoint **“annotations”**, model is a factorized BayesR
- $\pi_c \sim \text{Dirichlet}(\alpha)$, with $\alpha = (1, 1, 1, 1)$
- Gibbs sampler for estimation



From BayesR to BayesRC ... and beyond



Genotype ...000001001201002100200010100001011001011110...
 ...ACTCCGTAAGTAGCCTACAAAGGCTAACTTACAAAAGATTTA...

Predict →



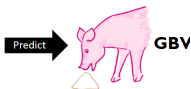
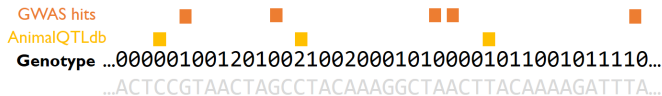
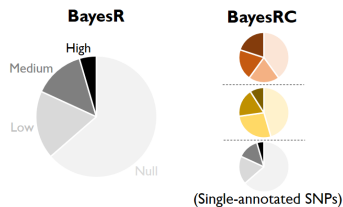
GBV



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



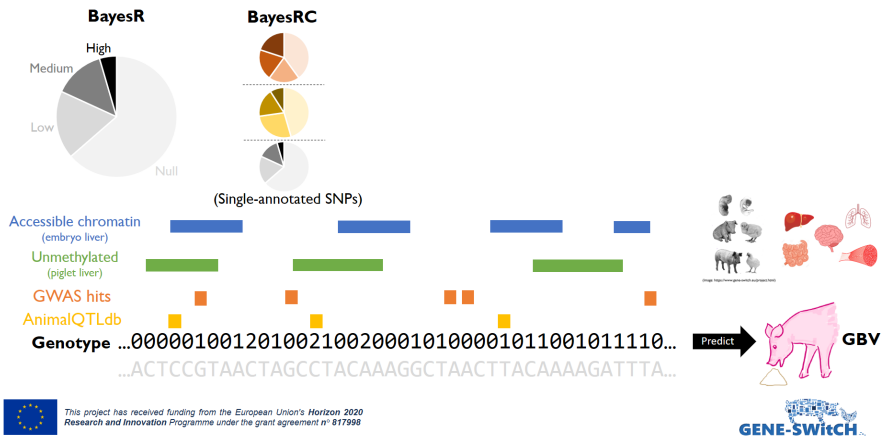
From BayesR to BayesRC ... and beyond



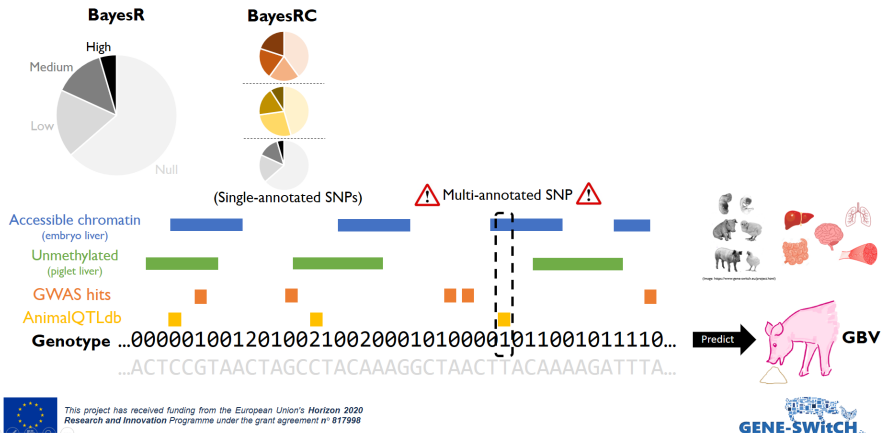
This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement n° 817998



From BayesR to BayesRC ... and beyond



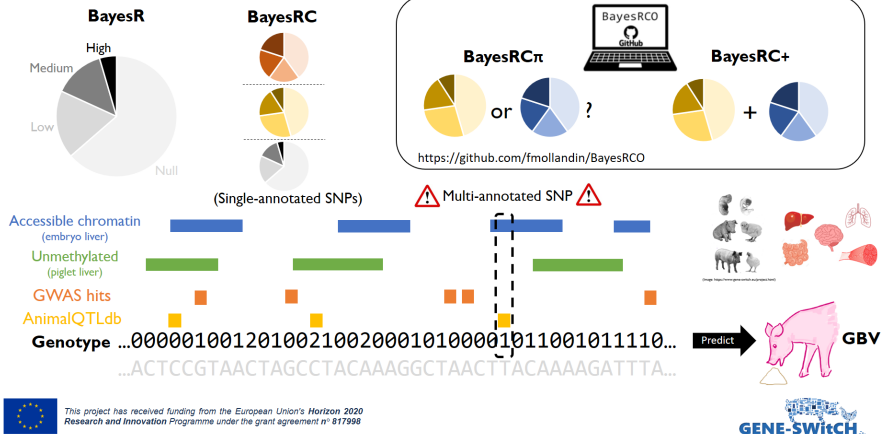
From BayesR to BayesRC ... and beyond



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



From BayesR to BayesRC ... and beyond



BayesRCO: BayesRC for Overlapping annotations

Two hypotheses = two models!

- 1 Multi-annotations represent **added confidence** → BayesRC+
- 2 Multi-annotations represent **uncertainty** → BayesRC π

$$A \in \{0,1\}^{S \times K} = \begin{pmatrix} & \text{Annot 1} & \text{Annot 2} & \dots & \text{Annot K} \\ \text{SNP}_1 & 0 & 1 & \dots & 0 \\ & 0 & 0 & \dots & 1 \\ & \cdot & \cdot & \cdot & \cdot \\ \text{SNP}_2 & 1 & 0 & \dots & 0 \\ & 1 & 0 & \dots & 0 \end{pmatrix} \quad \sum_{i=1}^K A(i,j) = 1$$

$$A \in \{0,1\}^{S \times K} = \begin{pmatrix} & \text{Annot 1} & \text{Annot 2} & \dots & \text{Annot K} \\ \text{SNP}_1 & 1 & 1 & \dots & 0 \\ & 0 & 0 & \dots & 1 \\ & \cdot & \cdot & \cdot & \cdot \\ \text{SNP}_2 & 1 & 0 & \dots & 1 \\ & 1 & 1 & \dots & 1 \end{pmatrix} \quad \sum_{i=1}^K A(i,j) \geq 1$$

	Method	SNP effect prior distribution	Annotations
	BayesR	$\beta_i \sim \sum_{K=1}^4 \pi_{K,a} \mathcal{N}(0, k\sigma_g^2)$	No
	BayesRC	$\beta_i a = A(i) \sim \sum_{K=1}^4 \pi_{K,a} \mathcal{N}(0, k\sigma_g^2)$	Yes, disjointed
Cumulative	BayesRC+	$\beta_i a \in A(i) \sim \sum_{a \in A(i)} \sum_{K=1}^4 \pi_{K,a} \mathcal{N}(0, k\sigma_g^2)$	Yes, overlapping
Preferential assignment	BayesRC π	$\beta_i a \in A(i) \sim \sum_{a \in A(i)} p_{i,a} \sum_{K=1}^4 \pi_{K,a} \mathcal{N}(0, k\sigma_g^2)$	Yes, overlapping

Simulation strategy

Phenotypes simulated from real cattle genotypes, 2500 animals:

- $h^2 = \{0.2, 0.5\}$
- 5 **large** QTLs representing $k = \{1\%, 2.5\%, 5\%\}$ of total additive variance σ_a^2
- 300 **medium** QTLs representing 0.1% of σ_a^2
- 4500 to **6500** **low** effect SNPs representing 0.01% of σ_a^2
- 50 datasets generated for each setting

Scenarios

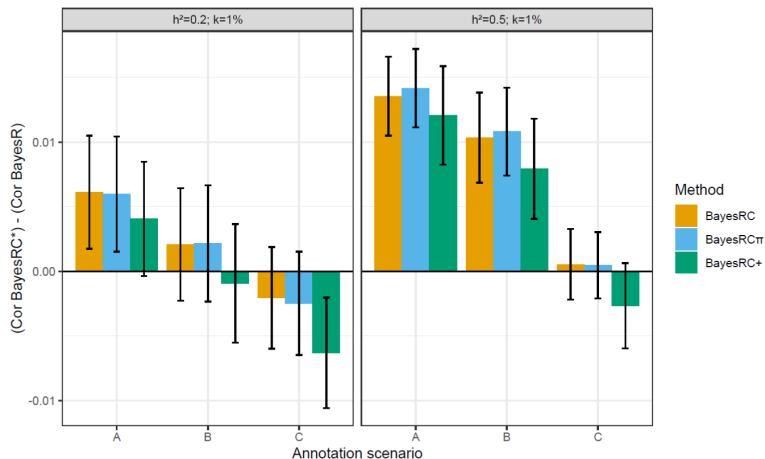
A B C



4 types of annotations possible:

- 1 strongly enriched:** 5 **large** QTLs + 300 **medium** QTLs + 150 **low** or null SNPs
- 2 moderately enriched:** 2 **large** QTLs + 100 **medium** QTLs + 300 **low** or null SNPs
- 3 weakly enriched:** 20 **medium** QTLs + 400 **low** or null SNPs
- 4 unenriched:** 450 **low** or null SNPs

Evaluating impact of using annotations on validation data



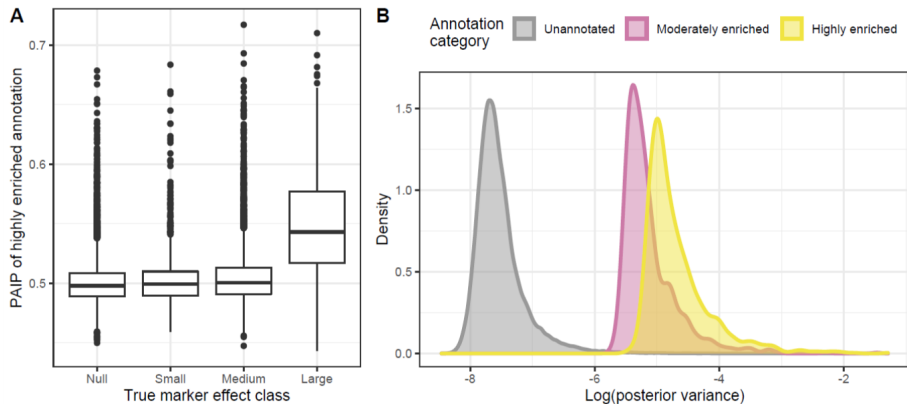
k= per-large QTL % of additive variance

A= 1 strongly enriched + 1 moderately enriched + unannotated;

B= 1 strongly enriched + 1 moderately enriched + 1 weakly enriched + 1 unenriched + unannotated

C= 2 strongly enriched + 2 moderately enriched + 3 weakly enriched + 2 unenriched + unannotated

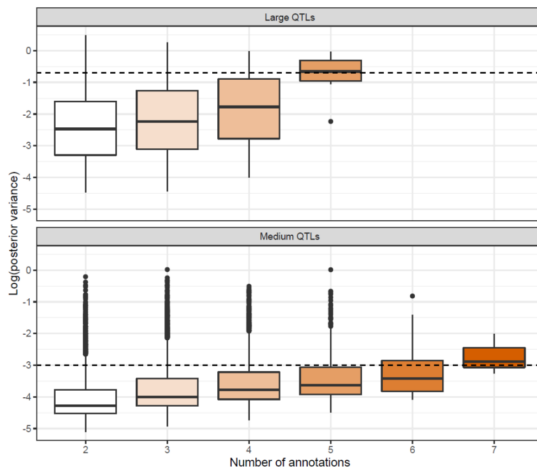
BayesRC π assigns informative annotations to QTLs



$h^2 = 0.5$, $k = 1\%$, scenario A

PAIP = posterior annotation inclusion probability (BayesRC π output)

BayesRC+ assigns more weight to multi-annotated variants



$h^2 = 0.5$, $k = 1\%$, scenario C

Application in backcross population of growing pigs

- $n = 1297$ backcross pigs (3/4 Large-White, 1/4 Creole), genetically related sows sired with 10 boars
 - Genotyped with Illumina Porcine 60k BeadChip array
 - Sibling-structured 10-fold cross validation procedure
- Traits pre-corrected for age, sex, farm
- Focus on average daily weight gain (**ADG**) and backfat thickness (**BFT**) at 23 weeks



Correlation of predicted traits in pig validation data

Annotations constructed using pigQTLdb for 11 trait sub-hierarchies

- Anatomy, behavioral, blood parameters, conformation, fatness, fatty acid content, feed conversion, growth, immune capacity, litter, reproductive organs
- Nearest up- and downstream neighboring markers also annotated

Correlation of predicted traits in pig validation data

Annotations constructed using pigQTLdb for 11 trait sub-hierarchies

- Anatomy, behavioral, blood parameters, conformation, fatness, fatty acid content, feed conversion, growth, immune capacity, litter, reproductive organs
- Nearest up- and downstream neighboring markers also annotated

	BayesR	BayesRC	BayesRC π	BayesRC+
ADG	0.21 (± 0.08)	+1.2 pts	+1.7 pts	+1.4 pts
BFT	0.26 (± 0.16)	-0.6 pts	-1 pts	+0.6 pts

Interpreting pigQTLdb annotations with BayesRC π



Conclusions: incorporating annotations with BayesRCO

BayesRCO:

- **BayesRC π** can assign informative annotations to multi-annotated SNPs to account for uncertainty in prior knowledge
- **BayesRC+** upweights multi-annotated SNPs and is robust to various annotation scenarios

- Fairly modest improvements in prediction (~ 1 -2 points) observed when incorporating biological annotations
 - Improved predictions and rankings of large QTLs in simulations, especially for highly informative annotations
 - Slight improvement in predictions for some traits in real data
 - Strategies for constructing annotation categories impact results

Take home messages

Can genomic prediction models be improved by better accounting for our **knowledge** about the **function** of certain regions of the genome?

Take home messages

Can genomic prediction models be improved by better accounting for our **knowledge** about the **function** of certain regions of the genome?

Yes, sometimes.

Take home messages

Can genomic prediction models be improved by better accounting for our **knowledge** about the **function** of certain regions of the genome?

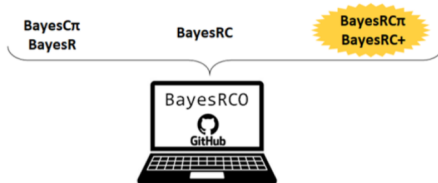
Yes, sometimes.

- **Models** → BayesRCO for overlapping annotation categories, extensions in progress to handle quantitative annotations
- **Genotyping data** → Capitalizing on annotation maps likely requires WGS resolution
- **Validation data** → Greater potential gains when prediction is performed on genetically distant populations
- **Traits** → Heritability, genetic architecture, link with annotations, ...
- **Annotations** → Which molecular assays, in which tissues?

Thank you!



Mollandin *et al.* (2022) Accounting for overlapping annotations in genomic prediction models of complex traits, *BMC Bioinformatics*, 23:65.



<https://github.com/FAANG/BayesRC0>