Intégration d'annotations fonctionnelles dans des modèles de prédiction génomique bayésiens

Fanny Mollandin

université PARIS-SACLAY

Intégration d'annotations fonctionnelles dans des modèles de prédiction génomique bayésiens

Incorporation functional annotations into Bayesian genomic prediction models

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°581 : agriculture, alimentation, biologie, environnement, santé (ABIES)
Spécialité de doctorat : Mathématiques appliquées
Graduate School : Biosphera, Référent : AgroParisTech

Thèse préparée dans l'**UMR GABI** (Université Paris-Saclay, INRAE, AgroParisTech), sous la direction de **Andrea RAU**, Chargée de Recherche (HDR) et la co-direction de **Pascal CROISEAU**, Chargé de Recherche (HDR)

Thèse soutenue à Paris-Saclay, le 28 septembre 2022, par

Fanny Mollandin

Composition du jury

Xavier Rognon Professeur, AgroParisTech (Université Paris-Saclay)	Président
Emmanuelle GENIN Directrice de Recherche, INSERM (Université de Bretagne)	Rapporteur & Examinatrice
Andres LEGARRA Directeur de Recherche, INRAE (centre Occitanie-Toulouse)	Rapporteur & Examinateur
Sophie ALLAIS Maîtresse de Conférences, Institut Agro Rennes-Angers	Examinatrice
Etienne BIRMELE Professeur, Université de Strasbourg	Examinateur
Andrea RAU Chargée de Recherche (HDR), INRAE (Université Paris-Saclay)	Directrice de thèse
Pascal CROISEAU Chargé de Recherche (HDR), INRAE (Université Paris-Saclay)	Co-directeur de thèse

THESE DE DOCTORAT
NNT : 2022UPASB051

université PARIS-SACLAY

ÉCOLE DOCTORALE
Physique et ingénierie : Electrons, Photons, Sciences du vivant (EOBE)

Titre : Intégration d'annotations fonctionnelles dans des modèles de prédiction génomique bayésiens
Mots clés : prédiction génomique, modèles bayésiens, annotations fonctionnelles, multi-omique

Résumé : La disponibilité généralisée et la baisse des coûts des technologies de génotypage à haut débit et de séquençage génomique ont ouvert la voie à des méthodes d'évaluation génomique, qui ont accéléré la mise en œuvre de l'évaluation génomique dans l'élevage pour de nombreuses espèces. Les méthodes d'évaluation génomique partagent un objectif commun, à savoir estimer avec précision un phénotype ou une valeur d'élevage estimée à partir des effets d'un ensemble de polymorphismes nucléotidiques (single nucleotide polymorphisms ; SNP), c'est-à-dire de variations d'un nucléotide sur le génome. Les modèles de prédiction bayésiens ont rapidement été adoptés, capable d'évaluer simultanément les effets des SNPs, tout en étant flexibles. Ils ont aussi l'avantage de pouvoir incorporer des informations sur la distribution des SNPs par leur loi a priori. Une piste d'amélioration potentielle de ces modèles réside dans la hiérarchisation des SNPs potentiellement causaux. À cette fin, plusieurs actions et projets internationaux, dont le projet européen GENE-SWitCH, ont récemment commencé à concentrer des efforts importants pour mieux caractériser les processus fonctionnels intermédiaires reliant les génotypes aux phénotypes quantitatifs. En particulier, l'objectif est de compléter les données de génotypage par des données d'annotation fonctionnelle, telles que le niveau de méthylation ou l'accessibilité de la chromatine dans plusieurs tissus et à des stades de développement pertinents, afin de mieux identifier les SNP causaux. Un défi majeur dans l'exploitation de ces données fonctionnelles réside dans la gestion de leur hétérogénéité et de leur complexité. Dans ce projet de thèse, l'objectif est de développer et de valider des modèles bayésiens de prédiction génomique capables de pondérer les SNPs en fonction des informations extraites de ces annotations fonctionnelles. Nous visons à la fois une meilleure capacité prédictive et une meilleure interprétabilité des résultats. Dans ce but, nous avons étendu le modèle BayesRC, dans lesquelles les signaux des SNPs sont partitionnés en fonction d'une catégorisation disjointe, pour pouvoir utiliser des données d'annotations hétérogènes et chevauchantes. Nous proposons deux nouveaux modèles, BayesRC$_pi$ et BayesRC+, respectivement reposant sur une modélisation stochastique ou cumulative des annotations multiples, afin de prendre en considération les SNPs multi-annotés. Ces modèles ont été appliqués à des données simulées et réelles, et plusieurs façons de construire et d'interpréter les annotations ont été proposés.

Title : Incorporation functional annotations into Bayesian genomic prediction models
Keywords : genomic prediction, Bayesian models, functional annotations, multi-omics

Abstract : The widespread availability and decreasing costs of high-throughput genotyping and genomic sequencing technologies have paved the way for genomic evaluation methods, which have accelerated the implementation of genomic evaluation in breeding for many species. Genomic evaluation methods share a common goal of accurately estimating a phenotype or breeding value based on the effects of a set of single nucleotide polymorphisms (SNPs), i.e. variations of one nucleotide on the genome. Bayesian prediction models were quickly adopted, capable of simultaneously assessing the effects of SNPs, while being flexible. They also have the advantage of being able to incorporate information on the distribution of SNPs by their a prior distribution. A potential avenue for improvement of these models lies in the prioritisation of potentially causal SNPs. To this end, several international actions and projects, including the European GENE-SWitCH project, have recently begun to focus major efforts on better characterising the intermediate functional processes linking genotypes to quantitative phenotypes. In particular, the aim is to complement genotyping data with functional annotation data, such as methylation level or chromatin accessibility in several tissues and at relevant developmental stages, to better identify causal SNPs. A major challenge in exploiting these functional data is to manage their heterogeneity and complexity. In this thesis project, the objective is to develop and validate Bayesian genomic prediction models capable of weighting SNPs according to the information extracted from these functional annotations. We aim at both a better predictive capacity and a better interpretability of the results. To this end, we have extended the BayesRC model, in which SNP signals are partitioned according to a disjoint categorisation, to be able to use heterogeneous and overlapping annotation data. We propose two new models, BayesRCπ and BayesRC+, based on stochastic or cumulative modelling of multiple annotations, respectively, in order to consider multi-annotated SNPs. These models have been applied to simulated and real data, and several ways of constructing and interpreting annotations have been proposed.

Communications et papiers

Publications

Fanny Mollandin, Hélène Gilbert, Pascal Croiseau, Andrea Rau, 2022, *Accounting for overlapping annotations in genomic prediction models of complex traits*, https://doi.org/10.1186/s12859-022-04914-5, BMC Bioinformatics **(publié)**

Fanny Mollandin, Hélène Gilbert, Pascal Croiseau, Andrea Rau, 2022, *Capitalizing on complex annotations in Bayesian genomic prediction for a backcross population of growing pigs* **(publié, short paper)**

Fanny Mollandin, Andrea Rau, Pascal Croiseau, 2021, *An evaluation of the interpretability and predictive performance of the BayesR model for genomic prediction* https://doi.org/10.1101/2020.10.23.351700, G3 **(publié)**

Présentations poster

Fanny Mollandin, Hélène Gilbert, Pascal Croiseau, Andrea Rau, 2021, *Exploiting prior knowledge into genomic prediction models with BayesRCO*, European Mathematical Genetics Meeting, Cambridge (UK)

Fanny Mollandin, Hélène Gilbert, Pascal Croiseau, Andrea Rau, 2020, *Evaluating the predictive power and interpretability of the BayesR genomic prediction model*, Congrès Européen des Productions Animales (EAAP), Virtuel

Presentations orales

Fanny Mollandin, Hélène Gilbert, Pascal Croiseau, Andrea Rau, 2022, *Capitalizing on complex annotations in Bayesian genomic prediction for a backcross population of growing pigs*, WCGALP (Rotterdam, Netherlands)

Fanny Mollandin, Hélène Gilbert, Pascal Croiseau, Andrea Rau, 2021, *Extension of Bayesian genomic prediction models for the integration of functional annotations*, Congrès Européen des Productions Animales (EAAP) (Davos, Switzerland)

Fanny Mollandin, Hélène Gilbert, Pascal Croiseau, Andrea Rau, 2022, *Accounting for overlapping annotations as biological priors in genomic prediction models of complex traits*, Séminaire MIA Paris-Saclay (Virtuel)

Baber Ali, Pascal Croiseau, **Fanny Mollandin**, 2021, *Using a priori biological information to evaluate the ability of BayesRC model in genomic prediction*, Congrès Européen des Productions Animales (EAAP) (Davos, Switzerland)

Fanny Mollandin, Pascal Croiseau, Andrea Rau, 2021, *Evaluating the interpretability of SNP effect size classes in Bayesian genomic prediction models*, European Mathematical Genetics Meeting (Virtuel)

Contents

List of Figures

List of Tables

Abréviations

ADN = Acide désoxyribonucléique

ARN = Acide ribonucléique

GWAS = Genome wide association study

LD = Linkage disequilibrium

MAF = Minor allele frequency

MAP = Maximum a posteriori

PAIP = Posterior annotation inclusion probability

PIP = Posterior inclusion probability

QTL = Quantitative Trait loci

SNP = Single nucleotide polymorphism

WGS = Whole genome sequencing

Chapter 1

Introduction

1.1 Évaluation génomique

1.1.1 Histoire de la sélection génomique

La sélection des espèces animales repose sur un principe commun, l'amélioration des individus qui les constituent à travers les générations. Cette notion d'amélioration peut concerner des critères variés, changeant en fonction de l'espèce et de leur utilisation par l'homme. C'est ainsi qu'ont émergé de nouvelles races, répondant à différents besoins tels que l'alimentation (viandes, lait, œufs, etc), le travail (travail de la terre, chasse, etc) ou l'habillement (cuir, laine, fourrure), mais aussi le sport ou la compagnie. Les objectifs différents vont conduire à des races spécialisées, aujourd'hui pour la plupart définies par des standards de race et devant donc correspondre à des critères précis. Pour cela, nous allons chercher à conserver les individus correspondant au mieux aux objectifs liés à leur espèce ou race, dans l'espoir que leurs caractéristiques soient transmises à leur descendance. On comprend alors que le principe de sélection repose sur le choix des reproducteurs, en particulier via le reproducteur mâle. Parmi tous les animaux candidats à la reproduction, il devient nécessaire de les évaluer, afin de pouvoir les classer et sélectionner les meilleurs. Ce choix a pu reposer sur des critères arbitraires, notamment esthétiques, ou sur la base de performances visibles. Ainsi, certains éleveurs ont commencé à noter les performances des animaux afin d'assurer la conformité des individus reproducteurs à des standards de races, et conserver les qualités de ces races spécialisées. Parallèlement à cela, le début du XXème siècle voit l'émergence de la génétique, suite à la redécouverte des travaux de Mendel, résultant en l'introduction d'une science nouvelle, la génétique quantitative. Étymologiquement, on la définit comme la génétique des caractères dont l'observation passe par une mesure et, par extension, comme la génétique des caractères à déterminisme complexe (Sellier et al., 2019). La génétique quantitative a notamment bénéficié des travaux du statisticien et biologiste Ronald Fisher, qui a entre autres introduit l'idée que le phénotype (i.e. l'ensemble des caractères apparents d'un individu)

est affecté à la fois par son environnement et une valeur génétique qui lui est propre, obtenue comme la somme des petits effets d'une multitude de gènes (modèle polygénique) (Fisher, 1918). On voit alors que la génétique quantitative se doit de combiner statistique et génétique.

On définit l'héritabilité au sens large H^2 comme la part de variance phénotypique expliquée par la variance génétique, i.e. $H^2 = \frac{\sigma_g^2}{\sigma_p^2}$, avec $\sigma_p^2 = \sigma_g^2 + \sigma_e^2$, $\sigma_p^2, \sigma_g^2, \sigma_e^2$ étant respectivement la variance phénotypique, génétique et environnementale. Dans cette thèse, on réduira la variance génétique à la variance additive totale σ_a^2, soit $\sigma_g^2 = \sigma_a^2$, en excluant les effets de dominance ou d'interaction. On définit alors l'héritabilité au sens strict telle que $h^2 = \frac{\sigma_a^2}{\sigma_p^2}$. Par souci de simplification, le terme héritabilité sera utilisé pour dénommer l'héritabilité au sens strict. La qualité de la sélection, dépendant de la qualité de prédiction, est alors liée à l'héritabilité des caractères à prédire. Pour évaluer l'efficacité de la sélection, on utilise la formule du progrès génétique proposée par Dickerson and Hazel (1944). Soit i l'intensité de sélection, T l'intervalle de génération, σ_g l'écart type génétique et ρ la précision de la prédiction, on estime le progrès génétique ΔG tel que

$$\Delta G = \frac{i\sqrt{\rho}\sigma_g}{T}.$$

Pour améliorer le progrès génétique, il faut donc soit augmenter l'intensité de sélection, soit augmenter la précision génétique, soit diminuer l'intervalle de génération. Nous nous intéresserons ici au deuxième levier, l'amélioration de la précision génétique. Pour l'améliorer, des méthodologies statistiques de plus en plus complexes ont été développées afin d'attribuer aux animaux candidats à la reproduction des valeurs génétiques d'élevage (EBV, ou *estimated breeding value*). Pour réaliser ces évaluations génétiques, on s'appuie sur les informations des performances, mais aussi sur des généalogies et plus récemment sur des informations de génotypage. L'utilisation des génotypes, c'est-a-dire de l'information génétique portée par un individu, a révolutionné le domaine de l'évaluation génétique, et par extension de la sélection.

1.1.2 Utilisation de la génomique pour l'évaluation animale

1.1.2.1 Généralités sur la génomique

On définit la génomique comme la science qui étudie le génome, c'est-à-dire l'ensemble de l'information génétique d'un organisme contenu dans chacune de ses cellules sous la forme de chromosomes. Cette information génétique est portée par l'ADN (*acide désoxyribonucléique*), et est donc caractérisée par l'enchaînement de nucléotides qui le constitue. Des mutations survenant à des endroits du génome (ou *locus*), sous la forme de substitution, insertion ou délétion, peuvent être à l'origine de différentes versions du texte génétique d'un individu à l'autre. On appelle allèle chaque version de ce texte génétique. Au niveau d'un locus, il y a autant d'allèles que de variations présentes dans la population. On parle de polymorphisme quand plusieurs allèles sont présents pour un

même locus. Lorsqu'on observe la variation d'une seule paire de bases à une position spécifique du génome, provenant de la substitution d'un nucléotide par un autre, on parle de variant mononucléotidique (SNV, ou *single nucleotidique variant*), et de polymorphisme mononucléotidique (SNP, ou *single nucleotidique polymorphism*) si celui-ci est présent chez au moins 1% de la population. Le génotypage, permettant d'identifier les nucléotides variants d'un individu à l'autre, peut prendre différentes formes, comme le génotypage de microsatellites SSR (*simple sequence repeats*), le génotypage ISBP (*insertion site based polymorphism*) ou le génotypage SNP. Les SNPs étant les variations génétiques les plus fréquentes (Kruglyak and Nickerson, 2001), ce dernier type de génotypage s'est rapidement imposé, et connaît un essor dans les années 1990, grâce aux progrès de la biologie moléculaire, des biotechnologies et de l'informatique. Par simplification, nous réduirons le terme génotypage SNP à génotypage dans cette thèse. Enfin, on différencie le génotypage, qui identifie les variations du génome pour des marqueurs biologiques fixés (par exemple des SNPs), du séquençage qui permet de connaître l'agencement des nucléotides, sans connaissance en amont d'information génétique.

Le coût d'un génotypage dépend du nombre de SNPs génotypés (ou densité de génotypage). Au début élevés, et limitant donc le nombre de marqueurs génotypés, ces coûts ont drastiquement baissé ces dernières années, permettant alors de génotyper la totalité du génome par séquençage haut-débit (WGS, *whole genome sequencing*), s'assurant ainsi d'avoir un accès direct aux mutations causales. Cela permet aussi le séquençage de plus d'individus, ainsi que l'étude des espèces ou races de petites effectifs qui étaient jusqu'à là mises à l'écart des évaluations génomiques. L'accès au génome entier est aussi facilité par le développement des techniques d'imputation, qui infère les SNPs non observés, à l'aide d'autres génomes entièrement séquençés, et permet donc d'augmenter la densité du génome. L'accumulation de ces données de très grande dimension a bénéficié d'une capacité de stockage et calcul accrue, résultant d'une amélioration des techniques et de l'infrastructure informatique (Kahn, 2011), sans lesquelles il n'aurait pas été possible de les traiter.

1.1.2.2 Utilité des données de génotypage

Pour un phénotype ayant une héritabilité suffisante, l'analyse de la variabilité génétique peut alors permettre d'expliquer ou prédire ce phénotype. Ces deux approches, quoique pouvant être effectuée de façon simultanée, sont en général utilisées dans des cadres différents avec des méthodologies différentes. La prédiction génomique sert principalement dans le milieu agronome, en permettant la sélection plus précises des animaux et des plantes. Parallèlement, le monde médical exploite les données de génotypage pour mieux comprendre des maladies, en procédant à des GWAS (*genome wide association studies*) le plus souvent. Ces deux approches reposent néanmoins sur l'hypothèse commune que certaines régions du génome ont un effet sur le phénotype. On définit alors un QTL (*quantitative trait loci*) comme une région du génome (de taille variable) qui explique une grande partie de la variabilité d'un caractère quantitatif.

Les données de génotypage sont aussi utilisées dans d'autres types d'analyses, notamment en génétique des populations, qui étudie l'évolution et les divergences entre les différents groupes d'une espèce, ou encore pour l'étude de la résistance de certains microorganismes, comme les bactéries.

Définitions utiles

Acide désoxyribonucléique (ADN): Macromolécule biologique porteuse de l'information génétique.

Génome: Ensemble du matériel génétique d'un organisme.

Nucléotide: Molécule biologique composant les acides nucléiques (ADN ou ARN), représenté par les lettres A, C, T, G et U.

Mutation: Modification du matériel biologique dans le génome, par substitution, insertion ou délétion. Peut aussi être appelé "variant".

Allèle: Version donnée d'un texte génétique;

Locus: Position du génome.

Polymorphisme: Présence en un *locus* de plusieurs allèles.

Single nucleotide variant (SNV): Variation d'un seul nucléotide entre les génomes d'individus d'une même population.

Single nucleotide polymorphism (SNP): SNV devant apparaître chez au moins 1% de la population.

Quantitative trait loci: Région du génome expliquant une partie de la variabilité d'un caractère quantitatif.

1.1.2.3 Les génotypes sous forme de données statistiques

Un allèle sera considéré comme référent pour une population, tandis que le ou les autre(s) seront considérés comme alternatif(s). Pour la majorité des SNPs, seulement deux variants sont observés, nous simplifierons donc l'encodage des génotypes en ne considérant que deux allèles (bi-allélisme). Chez un individu diploïde, il est alors possible d'observer 0, 1 ou 2 copie(s) d'allèles alternatifs. On représente donc les données de génotypage sous forme de matrice $\boldsymbol{X} \in \{0, 1, 2\}^{n \times p}$, avec n le nombre d'individus, p le nombre de SNPs, et chaque élément de matrice indiquant le nombre de copie de l'allèle alternatif (Figure 1.1).

Les données de génotypage comportent quelques spécificités, leur induisant une modélisation statistique adaptée. Elles se retrouvent sous la forme de matrices de très grande dimension, avec $n \ll p$, assez creuses (i.e comportant beaucoup de zéros). On retrouve aussi généralement une structuration importante de données, que ce soit entre individus ou entre les SNPs. Les premiers sont corrélés en raison de leur proximité génétique, en particulier chez les animaux d'élevage qui vont souvent avoir des liens de parentés favorisés par la sélection antérieure.

Les seconds sont corrélés les uns aux autres, notamment en fonction de leur proximité dans le génome. Lors de la méiose, un phénomène de recombinaison nommé *crossing over* modifie la suite de nucléotides des chromosomes

concernés, par l'échange de portion de nucléotides entre deux chromosomes homologues. Lors de ces recombinaisons, deux marqueurs consécutifs ont moins de probabilité d'être séparés que d'autres plus éloignés, il y aura donc une probabilité plus élevée qu'ils soient co-transmis à leur descendance. Deux marqueurs sont en déséquilibre de liaison (LD, *linkage disequilibrium*) quand la fréquence d'association de leurs allèles ne correspond pas à la fréquence d'association liée à une association aléatoire des nucléotides. Le LD peut être estimé de multiples façons, une des plus courantes étant le r^2 (Pritchard and Przeworski, 2001). Soit $\{f_A, f_a\}$, $\{f_B, f_b\}$ les fréquences des allèles pour deux marqueurs et f_{AB}, f_{aB}, f_{Ab} et f_{ab} les fréquences d'associations de chaque paire d'allèles (i.e fréquence des haplotypes), alors on calcule le r^2 tel que:

$$r^2 = \frac{(f_{AB} - f_A f_B)^2}{f_A f_a f_B f_b}.$$

Plus r^2 est proche de 1, plus le LD est fort. Le LD est aussi lié à d'autres forces évolutives, telles que la dérive génétique, la sélection historique et la mutation, et peut être un avantage comme un inconvénient pour l'analyse des données génomiques (Qanbari, 2020). Ainsi, le LD est nécessaire pour pouvoir prédire des caractères sans avoir directement accès à la mutation causale (ce qui est fréquent en densité de marqueurs restreinte), en exploitant un ou des SNPs génotypés en LD avec la mutation causale. C'est notamment sur ce principe que sont bâties les puces à ADN, en utilisant des marqueurs réguliers tout au long de l'ADN pour tenter d'exploiter au mieux les mutations causales via leur déséquilibre de liaison, tout en permettant de réduire les coûts et la complexité d'analyses à l'aide d'une densité plus faible. A l'inverse, le LD induit des blocs de corrélation forte entre les variables, ce qui peut se révéler problématique dans l'utilisation des modèles statistiques, à cause d'une forte multicolinéarité. En pratique, cela peut nécessiter de filtrer en amont les SNPs pour n'en garder qu'un sous-ensemble représentatif au sein d'une région en fort LD.

1.1.3 Prédiction des phénotypes à partir de génotypes

L'utilisation des données de génotypage pour la prédiction de caractères génétiques divers (Figure 1.1) nécessite la construction de modèles de prédiction robustes, adaptés aux spécificités des données de génotypage. On définit un modèle de prédiction comme une description mathématique d'un processus, liant un élément déterministe du processus (variable de réponse, ici un phénotype) à des prédicteurs (ou variables explicatives, ici les SNPs) par une équation. Il doit aussi comporter, comme tout modèle statistique, une partie aléatoire, avec une distribution de probabilité associée. On peut l'écrire comme:

$$y_i = f(x_i) + e_i, \forall i \in 1, 2, ..., n \tag{1.1}$$

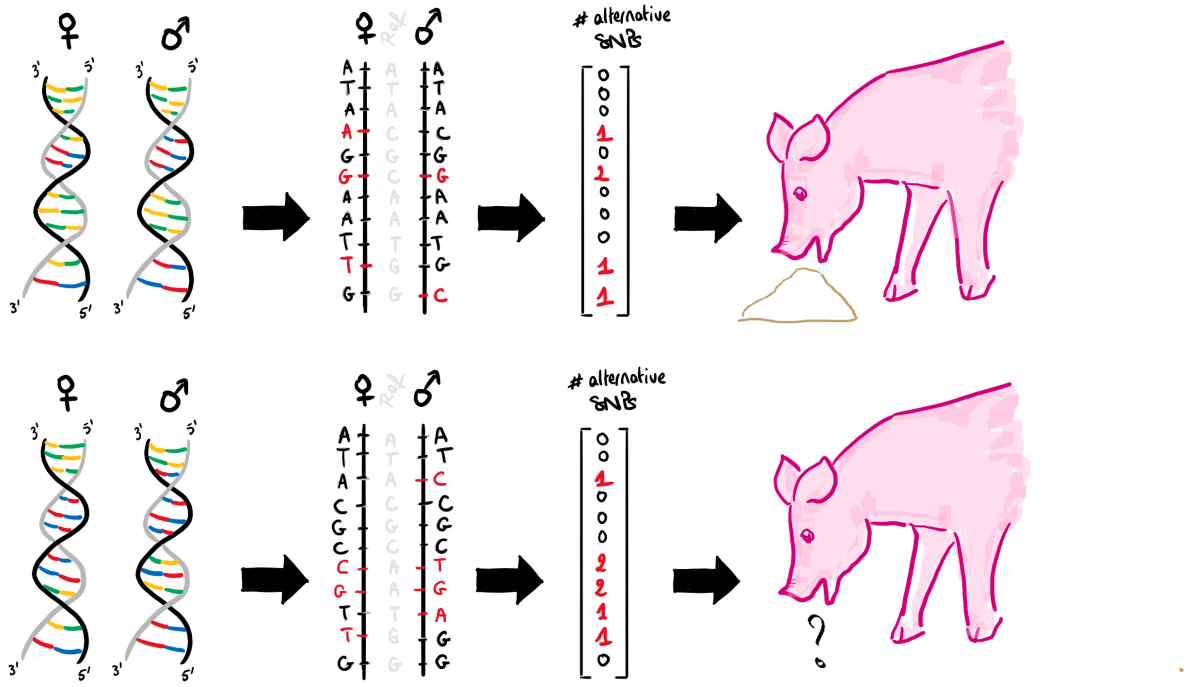


Figure 1.1: **Utilisation de l'ADN pour la prédiction de caractère de production**
L'ADN est séquencé de tel façon à identifier les nucléotides pour chaque individu. Ces nucléotides sont représentées sous la forme d'une matrice $\boldsymbol{X} \in \{0,1,2\}^{n \times p}$, avec n le nombre d'animaux et p le nombre de SNPs, représentant le nombre d'allèles alternatifs (rouge) possédés par l'individu, par rapport à un allèle considérée comme référent (noir). La prédiction s'opère en 2 étapes, (1) l'apprentissage sur un premier groupe d'animaux, de phénotype connu (haut), (2) la prédiction sur un second groupe dont le phénotype est inconnu.

y_i représentant la variable de réponse, x_i les prédicteurs et e_i le terme d'erreur aléatoire. Ce dernier peut être perçu comme l'unicité de chaque individu. La fonction f peut prendre plusieurs formes, et se doit d'être choisie de manière à être cohérente aux données observées. Il n'existe pas de fonction parfaite pour tous les processus, il est donc pertinent d'évaluer et comparer les modèles correspondants à différents choix de f pour chaque jeu de données.

1.1.3.1 Modèle linéaire général

Pour la prédiction génomique, sous l'hypothèse de l'additivité des SNPs, on utilise classiquement l'équation linéaire de prédiction suivante:

$$\begin{aligned} \boldsymbol{y} &= \mu \mathbb{1}_n + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \\ \boldsymbol{e} &\sim N(0, \boldsymbol{I}_n \sigma_e^2) \end{aligned} \tag{1.2}$$

avec $\boldsymbol{y}_{n\times 1}$ le vecteur des n phénotypes observés (i.e la variable de réponse), $\boldsymbol{\mu}_{1\times 1}$ l'intercept, $\boldsymbol{X}_{n\times p}$ la matrice de génotypage (i.e les variables prédictives), $\boldsymbol{\beta}_{p\times 1}$ les effets des p marqueurs (variable aléatoire) et $\boldsymbol{e}_{n\times 1}$ le vecteur des résidus (terme d'erreur aléatoire). Soit $\hat{\boldsymbol{\beta}}$ et $\hat{\mu}$ respectivement les effets et la moyenne estimés par le modèle, on estime le caractère $\boldsymbol{y}$ par $\hat{\boldsymbol{y}} = \hat{\boldsymbol{\mu}} + \boldsymbol{X}\hat{\boldsymbol{\beta}}$. On évaluera alors la qualité de prédiction en utilisant la corrélation de Pearson:

$$\rho(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\text{Cov}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\sigma_y \sigma_{\hat{y}}},$$

avec σ_y et $\sigma_{\hat{y}}$ les écart-types respectifs de $\boldsymbol{y}$ et $\hat{\boldsymbol{y}}$. S'il est aussi courant d'utiliser d'autres critères, comme le MSE (*mean squared error*) dans d'autres problématiques de prédiction, la corrélation de Pearson est dans notre cas adaptée car nous nous intéressons principalement au classement des animaux.

Le modèle général ainsi défini, on différencie les différents modèles abordés dans cette thèse par la distribution des effets de marqueurs β_j pour SNP j, et notamment de leur variance associée $\sigma^2_{\beta_j}$. On cherchera alors à estimer au mieux le vecteur $\boldsymbol{y}$, en estimant au mieux le vecteur $\boldsymbol{\beta}$. Pour la suite, on note σ^2_g la variance additive totale des marqueurs, qui sous nos hypothèses est défini comme $\sigma^2_g = \sum_{j=1}^{p} 2f_j(1-f_j)\sigma^2_{\beta_j}$ avec f_j la fréquence allélique du marqueur j.

1.1.3.2 Un premier modèle de prédiction génomique: GBLUP

Le modèle le plus largement utilisé dans le cadre de la sélection génomique est le *genomic best linear unbiased predictor*, ou GBLUP (Habier et al., 2007). Il est défini comme un modèle linéaire classique, reposant sur l'hypothèse forte que les marqueurs ont *a priori* tous la même variance, tels que $\sigma^2_{\beta_j} = \sigma^2_\beta$, pour tout j et $\beta_j \overset{i.i.d}{\sim} N(0, \sigma^2_\beta)$. Cela signifie *a priori* que tous les SNPs ont un effet non-nul, peu importe la densité de génotypage, et que ces effets sont petits.

Bien que reposant sur une hypothèse forte, peu plausible en réalité, le GBLUP reste le modèle de référence en prédiction génomique. Il est relativement peu coûteux en temps de calcul, facile à mettre en place, et présente des performances de prédiction souvent similaires à des modèles plus sophistiqués. C'est un modèle ainsi souvent utilisé dans les évaluations génétiques de routine. On peut cependant émettre des réserves sur sa modélisation, qui pousse l'ensemble des variants à avoir des effets non-nuls, et donc à répartir la variance additive totale entre tous les variants. Cela a pour conséquence d'une part de surestimer des variants nuls, majoritaires dans les données, et d'autre part de sous-estimer des QTLs, qui vont se voir attribuer une distribution à très petite variance. Néanmoins, cette sous-estimation des effets des QTLs est compensée par l'exploitation des marqueurs voisins en fort LD, ce qui peut expliquer les bonnes capacités de prédiction pour le GBLUP malgré une mauvaise représentation de l'architecture génétique du caractère étudié.

1.1.3.3 Diversité dans les approches de prédiction génomique

De multiples modèles de prédiction ont été proposés ces vingt dernières années. Les modèles se sont complexifiés, cherchant à résoudre les différents problèmes rencontrés en prédiction génomique. Dans l'ensemble, on peut les diviser en trois grandes familles :

- les modèles fréquentistes
- les modèles bayésiens
- les modèles types machine learning

Les modèles bayésiens étant au coeur de cette thèse, ils seront développés plus en détail dans la suite. Nous présenterons ici les autres approches existantes, de façon non exhaustive.

Dans le cadre fréquentiste, les effets des marqueurs sont considérés comme des effets fixes. En raison du grand nombre de marqueurs, induisant des problèmes d'estimation liées à la grande dimension, les méthodes de régressions pénalisées ont été privilégiées. On retrouve ainsi l'utilisation de modèles pénalisés classiques tels que le lasso (Tibshirani, 1996), le ridge (Hoerl and Kennard, 1970), le lasso adaptatif (Zou, 2006) ou l'elastic-net (Zou and Hastie, 2005). Bien que montrant de bons résultats de prédiction (Ogutu et al., 2012; Usai et al., 2009), ils ont leurs propres limites. Tout d'abord, ils nécessitent de fixer le ou les paramètres de pénalisation, ce qui peut se révéler difficile à faire en pratique. Si la régression lasso est adaptée à la grande dimension, elle peut rencontrer des difficultés quand les variables prédictives sont très inter-corrélées par blocs car certaines variables peuvent être privilégiées au détriment d'autres. De plus, le nombre de variables sélectionnées par le lasso est limité au nombre d'individus au maximum. Cela peut poser un problème pour les données génomiques, où les SNPs sont corrélés à cause du LD, et où $n \ll p$, en particulier pour les données animales (Waldmann et al., 2013). La pénalisation ridge permet de gérer plus aisément les variables inter-corrélées, mais n'est pas adaptée si une partie de ces variables n'ont pas d'effet. Elle aura aussi tendance à sous-évaluer les SNPs à effet fort. L'utilisation d'une régression elastic-net permet de combiner à la fois les avantages des pénalisations lasso et ridge, donc à faire de la sélection de variable tout en prenant en compte leurs corrélations. Les performances de la pénalisation elastic-net dépendent encore ici des paramètres de pénalisation choisis, qu'il faut identifier pour les parties ridge et lasso.

Plus récemment, on observe de plus en plus d'applications de méthodes d'apprentissage automatique (machine learning). Parmis les plus utilisées, on retrouve le gradient boosting machine (GBM), les forêts aléatoires, ou encore le support vector machine (SVM). Bien que leur utilisation n'aient pour l'instant pas montré d'augmentation notable de la qualité de prédiction (Bellot et al., 2018), elles semblent néanmoins être de bonnes pistes pour des données à structure hétérogènes et non linéaires (Nayeri et al., 2019; Montesinos-López et al., 2021; van Dijk et al., 2021). En particulier, elles semblent plus adaptées à des architectures génétiques complexes, non additives, avec des effets de dominance ou d'épistasie (Zingaretti et al., 2020). Une réserve à l'égard de ces modèles, non

spécifique à la prédiction génomique, est l'aspect "boîte noire" qui les caractérise, et qui rend difficile l'interprétation des résultats hors prédiction.

1.1.3.4 Stratégies de validation de modèle

Pour évaluer la qualité du modèle de prédiction, les individus doivent être divisés entre une population d'apprentissage, sur laquelle le modèle prédictif va être ajusté, et une population de validation, sur laquelle la qualité de prédiction va être évaluée. Deux grands types de validation existent:

- **Validation non-croisée** (*holdout method*). Cela consiste en la division des n individus en une proportion q pour l'apprentissage et $1-q$ pour la validation, généralement avec q aux alentours de 70-80%.
- **Validation croisée à k blocs** (k*-fold cross validation*). Les n individus sont divisés en k blocs, afin d'utiliser les $k-1$ blocs pour l'apprentissage, et le dernier pour la validation. On répète l'opération k fois, de tel sorte que chaque bloc est utilisé en validation une fois. Un cas particulier de *cross-validation* est $k=n$, où chaque individu est prédit individuellement à l'aide des $n-1$ en apprentissage (validation *leave-one-out*).

Si ces partitionnements peuvent être fait de façon aléatoire, il est courant dans un cadre de sélection de choisir les individus les plus jeunes pour la validation. D'autres approches peuvent être envisagée selon les particularités des données, il est par exemple possible d'utiliser les individus d'une race (ou famille génétique) en apprentissage et utiliser les individus d'une autre race (ou famille génétique) en validation.
Dans le cas de validation croisée, les résultats de chaque *fold* sont synthétisés pour obtenir une performance de validation, souvent à l'aide de moyenne et d'écart-type.

1.2 Utilisation d'une approche bayésienne pour la prédiction génomique

L'approche bayésienne s'est rapidement imposée pour la prédiction génomique. Nous rappelons en premier lieu dans ce chapitre des généralités sur les statistiques bayésiennes, afin de mieux comprendre les modèles bayésiens développés pour la prédiction génomique.

1.2.1 Généralités sur les statistiques bayésiennes

L'approche bayésienne, à l'instar de l'approche fréquentiste, suppose l'utilisation de données échantillonnées à partir d'une loi de paramètres inconnus. Cependant, alors que ces paramètres inconnus sont considérés fixes par les statisticiens fréquentistes, ils sont considérés comme variables par les statisticiens bayésiens, nécessitant alors une distribution *a priori* pour chacun de ces paramètres inconnus $\boldsymbol{\theta}$. Le modèle d'échantillonnage est donné par la probabilité de distribution $f(y|\boldsymbol{\theta})$, de vraisemblance associée $L(\boldsymbol{\theta}|y)$. On écrit $\pi(\boldsymbol{\theta})$ distribution *a priori* de $\boldsymbol{\theta}$ qui

sera parfois nommée seulement *prior* par la suite. On obtient alors sa distribution *a posteriori*, ou *posterior*, par l'utilisation du théorème de Bayes (Bayes, 1763):

$$p(\boldsymbol{\theta}|y) = \frac{p(y,\boldsymbol{\theta})}{p(y)} = \frac{p(y,\boldsymbol{\theta})}{\int p(y,\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{f(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{1.3}$$

On note que le *posterior* correspond au produit de la vraisemblance et du prior, renormalisé par une constante de façon à que son intégrale soit égale à 1. On peut alors simplifier l'équation 1.3 par

$$p(\boldsymbol{\theta}|y) \propto L(\boldsymbol{\theta}|y)\pi(\boldsymbol{\theta}) \tag{1.4}$$

signifiant que le *posterior* est proportionnel à la vraisemblance par le *prior*.

Le choix du *prior* est donc déterminant à l'estimation des paramètres. Il peut être spécifié en fonction de connaissances antérieures ou d'hypothèses faites par des spécialistes, mais est souvent déterminé de telle façon que la distribution *a posteriori* soit facilement identifiable. On appelle ainsi un *prior* conjugué à une vraisemblance $L(\boldsymbol{\theta}|y)$ un *prior* appartenant à une famille de distribution à laquelle son *posterior* associé appartient aussi. Il n'est cependant pas rare de n'avoir aucune information en amont sur la forme que devrait prendre le *prior*, ou bien de vouloir faire de l'inférence à partir des données uniquement. On considère un *prior* comme non informatif quand il ne favorise pas de valeur de $\boldsymbol{\theta}$. Dans la réalité, il est difficile de considérer qu'il existe des *priors* ne représentant réellement aucune information, on cherche donc une forme minimisant l'information qu'elle apporte, et donc laisse aux données observées plus d'importance relative. Plusieurs façons de construire des *priors* non informatifs ont été proposées, on peut par exemple citer le *prior* de Jeffreys ou le *prior* de référence (Bernardo, 1979).

Les *priors* peuvent eux-mêmes être paramétrés par ce que l'on appelle des hyperparamètres, qui peuvent être considérés comme fixes ou variables. Il est ainsi possible de rajouter des couches de paramétrage successives, nous obtenons alors un modèle bayésien à plusieurs niveaux, ou dit de structure hiérarchique. Formellement, un modèle est hiérarchique à ℓ niveaux quand la loi jointe des données observées et des ℓ paramètres est telle que

$$p(y,\theta_1,\theta_2,...,\theta_\ell) = f(y|\theta_1)\pi_1(\theta_1|\theta_2)\pi_2(\theta_2|\theta_3)...\pi_\ell(\theta_\ell) \tag{1.5}$$

pour le vecteur $\boldsymbol{\theta} = (\theta_1,\theta_2,...,\theta_\ell)$ regroupant les paramètres et hyperparamètres du modèle.
Soit $\tilde{\theta}$ un estimateur de θ et $K(\theta,\tilde{\theta})$ une fonction de coût. On appelle l'estimateur bayésien l'estimateur $\hat{\theta}$ qui respecte la règle de Bayes, c'est à dire qui minimise le risque *a posteriori* $E_\pi[K(\theta,\hat{\theta})]$

$$\hat{\theta} = \min_{\tilde{\theta}\in\theta} \int K(\theta,\tilde{\theta})\pi(\theta|x)d\theta$$

Les fonctions de coûts les plus populaires sont :

- la fonction de coût quadratique, qui a pour estimateur la moyenne
- la fonction de coût absolue, qui a pour estimateur la médiane
- la fonction de coût 0-1, qui a pour estimateur le maximum à posteriori

Selon la fonction de coût, un estimateur bayésien n'est pas toujours directement disponible, en particulier quand θ est multidimensionnel, ce qu'on retrouve souvent dans les modèles bayésiens hiérarchiques complexes. Il faut donc mettre en place des stratégies indirectes pour l'approcher.

1.2.2 Algorithme MCMC et implémentation

Le calcul de l'estimateur bayésien, particulièrement pour un $\boldsymbol{\theta}$ multidimensionnel, est parfois impossible à calculer, nécessitant son approximation. Pour résoudre ce problème d'inférence bayésienne, il est courant d'utiliser des algorithmes de Monte Carlo par chaînes de Markov (MCMC, Markov chain Monte Carlo).

1.2.2.1 Chaînes de Markov

On définit une chaîne de Markov $(Z_n)_{n\geq 1}$ à valeurs dans un ensemble E comme un processus aléatoire dont les transitions entre les éléments sont définis par une matrice stochastique $\mathbb{P}(Z_n, Z_{n+1})$, et dont les processus vérifient la propriété de Markov, pour tout $(z_0, z_1, ..., z_{k+1})$ dans E:

$$\mathbb{P}(Z_{k+1} = z_{k+1}|Z_k = z_k, ..., Z_1 = z_1) = \mathbb{P}(Z_{k+1} = z_{k+1}|Z_k = z_k), \forall k \in \mathbb{N}$$

Une chaîne de Markov est donc un processus stochastique dont la probabilité d'un élément ne dépend que de l'état de l'élément précédent, ou un processus sans mémoire où le futur dépend uniquement du présent et non du passé. C'est donc une forme simple de suites de variables aléatoires, qui permettent de construire des algorithmes itératifs.

1.2.2.2 Méthodes de Monte Carlo

Théorème (Loi forte des grands nombres): Soit $(U_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes, identiquement distribuées, telles que $\mathbb{E}[|U_1|] < \infty$. Soit $\mathbb{E}[|U_1|] = m$, alors

$$\frac{1}{n}\sum_{k=1}^{n} U_k \xrightarrow{p.s.} m$$

Les méthodes de Monte Carlo sont des applications directes de ce théorème, permettant l'approximation de quantités, notamment d'intégrales, à partir de simulations de variables aléatoires. Soit $U = h(V)$, avec h fonction borélienne et $V \sim f$, l'estimateur de la quantité $\delta = \mathbb{E}[h(V)] = \int h(v) f_V(v) dv$ par la loi forte des grandes nombres

est

$$\hat{h}_N = \frac{1}{N}\sum_{k=1}^{N} h(V_k),$$

pour $(V_1, V_2, ..., V_N)$ générés sous la loi de densité f. Si h est intégrable par rapport à la mesure f alors

$$\hat{h}_N \xrightarrow{p.s.} \delta.$$

Plus le paramètre N est grand, plus l'erreur d'approximation est réduite. Bien que cette procédure soit relativement simple, elle requiert des échantillons indépendants et identiquement distribués (*iid*). S'il existe des méthodes pour s'assurer d'avoir un échantillon respectant cette hypothèse, telles que la méthode par inversion ou la méthode de rejet, elles s'avèrent insuffisantes dans de nombreux cas. La méthode par inversion nécessite par exemple la connaissance de l'inverse de la fonction de répartition, ce qui n'est pas toujours accessible, particulièrement dans des modèles bayésiens hiérarchiques.

1.2.2.3 Méthodes de Monte Carlo par chaînes de Markov

Par définition, les méthodes de MCMC sont la combinaison des deux principes précédents. Elles comprennent une classe d'algorithmes pour l'échantillonnage à partir d'une distribution de probabilité, via la construction d'une chaîne de Markov, et en générant des échantillons aléatoires. A la différence des méthodes de Monte Carlo classique, les échantillons X_i n'ont plus besoin d'être générés indépendamment, et peuvent-être corrélés. Parmi les méthodes MCMC les plus connues et diffusées, on peut citer l'algorithme de Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) et l'échantillonneur de Gibbs (Geman and Geman, 1984).

Dans ce travail de thèse, seul ce deuxième a été utilisé dans l'implémentation des modèles, et sera donc développé dans cette introduction. On peut néanmoins noter que l'échantillonneur de Gibbs est un cas particulier de l'algorithme de Metropolis-Hastings.

Le principe de l'échantillonneur de Gibbs est de décomposer un problème général, multivarié, en une série de problèmes plus simples (typiquement tous univariés). Dans le cadre bayésien, cela revient à remplacer la modélisation via une loi jointe par une série de lois conditionnelles.

Soit $\theta = (\theta_1, \theta_2, ..., \theta_\ell)$ le vecteur des paramètres à échantillonner suivant la loi *a posteriori* $\pi(\theta|y)$. A partir d'une

initialisation du vecteur $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_\ell^{(0)})$, générer pour tout $t \geq 0$:

$$\begin{aligned}
&\theta_1^{(t+1)} \sim \pi_1(\theta_1|\theta_2^{(t)}, \theta_3^{(t)}, ..., \theta_\ell^{(t)}; y) \\
&\theta_2^{(t+1)} \sim \pi_2(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, ..., \theta_\ell^{(t)}; y) \\
&... \\
&\theta_\ell^{(t+1)} \sim \pi_\ell(\theta_\ell|\theta_1^{(t+1)}, \theta_2^{(t+1)}, ..., \theta_{\ell-1}^{(t+1)}; y)
\end{aligned}$$

Il est commun d'utiliser un échantillonneur de Gibbs quand la distribution jointe est intractable, soit car elle n'est pas connue, soit parce qu'il est difficile d'échantillonner à partir de cette loi. On le retrouve donc dans des problèmes en grande dimension, et les problèmes où les distributions conditionnelles sont suffisamment faciles à identifier. Comme pour les méthodes MCMC en général, il est courant de procéder à l'estimation des paramètres en utilisant seulement les échantillonnages à partir d'un temps n_T donné, que l'on appelle étape de *burn-in* (Gilks et al., 1995). Cela permet de ne pas prendre en compte une initialisation loin de la vraie valeur du paramètre, et ainsi laisser le temps à l'algorithme d'atteindre sa distribution d'équilibre. Pour des soucis computationnels et de stockage, il est aussi répandu d'utiliser un *thinning rate*, qui consiste à conserver uniquement les échantillonnages toutes les q itérations. Ces deux pratiques sont néanmoins sujettes à débat (Link and Eaton, 2012), et d'autres alternatives leurs ont été proposées.

1.2.3 État de l'art des modèles de prédiction génomique bayésiens

1.2.3.1 Introduction de l'"alphabet bayésien"

De multiples modélisations bayésiennes pour la prédiction génomique ont été proposées ces deux dernières décennies, tentant de représenter au mieux la réalité, et ainsi potentiellement améliorer la qualité de prédiction. La plupart de ces méthodes se distinguent par la distribution a priori proposée pour les effets de SNPs, nous rentrons alors dans un cadre bayésien qui nous permet d'introduire nos hypothèses directement dans les modèles via les *priors*. Les deux premiers modèles proposés sont BayesA et BayesB, dans un même papier par Meuwissen et al. (2001). Le premier opte pour une hypothèse proche de la réalité, où chaque effet de marqueur possède une variance qui lui est propre dans sa distribution *a priori*. Cependant, cette méthode est très coûteuse en temps de calcul, particulièrement dans des contextes de densification des génotypes et d'augmentation du nombre d'animaux génotypés. De plus, il est raisonnable de considérer qu'une grande majorité des marqueurs présents dans les données n'ont aucun effet sur le caractère étudié, ou ne sont pas en LD avec un marqueur qui a un effet non négligeable, et ainsi ne contribue pas à la variance génétique. Il est donc intéressant de considérer plutôt des *priors zero-inflated*, i.e. dont la distribution permet de nombreuses observations nulles. Cela correspond à un modèle de mélange comportant un terme de masse en 0. C'est dans cette optique qu'a d'abord été développé

BayesB. Un paramètre π est introduit pour représenter la proportion de SNPs avec un effet nul des données. La proportion $(1 - \pi)$ des SNPs inclus dans le modèles se partagent la variance additive, π a donc un impact sur le rétrécissement (*shrinkage*) des effets des SNPs inclus. Dans BayesB, ce paramètre est connu et fixé à l'avance, ce qui a été critiqué par Habier et al. (2011) comme un "manque d'aprentissage bayésien", et potentiellement responsable de la sous-estimation des SNPs à effet fort. Pour répondre à ce problème, Habier et al. (2011) ont proposé BayesCπ et BayesDπ, en modélisant le paramètre π comme un effet variable, suivant *a priori* une loi Beta. Ces méthodes *zero-inflated* ont permis une amélioration notable du temps de calcul.

Les différents modèles introduits dans cette thèse, non exhaustifs, sont regroupés sous l'appellation d' "alphabet bayésien" (Gianola et al., 2009), qui continue encore aujourd'hui à être étendu (Table 1.2). Il convient de souligner que ces modèles restent peu utilisés dans les évaluations génétiques de routine utilisées pour la sélection, mais ont un rôle essentiel dans la recherche d'une modélisation retranscrivant au mieux la biologie sous-jacente des caractères étudiés.

Model	Hierarchical Prior Density			Prior Density conditional on the parameters $p(\mathrm{u_i}\|\omega,\cdots)$	Terms & description
	Prior of $\mathbf{u_i}$ conditional on the variance $\sigma^2_{u_i}$ $(N(\mathrm{u_i}\|0,\sigma^2_{u_i}))$	Prior of $\sigma^2_{u_i}$ conditional on hyper-parameters ω $(p(\sigma^2_{u_i}\|\omega))$	Hyper-parameters $(p(\omega))$		
Gaussian	$u_i \sim N(0,\sigma^2_u)$	σ^2_u was fixed	-	$u_i \sim N(0,\sigma^2_u)$	Following uniform normal distribution e.g. SNP-BLUP (Meuwissen *et al.* 2001), GBLUP (VanRaden 2008)
Thick tail	$u_i \sim N(0,\sigma^2_{u_i})$	$\sigma^2_{u_i} \sim \chi^{-2}(v,S)$	v,S were fixed	$u_i \sim t(0,v,S)$	Following student distribution e.g. BayesA (Meuwissen *et al.* 2001)
		$\sigma^2_{u_i} \sim DE(0,\lambda)$	λ was fixed	$u_i \sim DE(0,\lambda)$	Following Double exponential distribution (DE) e.g. BayesLasso (Tibshirani 1996; Park & Casella 2008)
Spike-around -zero & Slab	$u_i \sim \pi N(0,\sigma^2_u+\sigma^2_b) + (1-\pi)N(0,\sigma^2_b)$	σ^2_u, σ^2_b were fixed;	π: uniform prior	$u_i \sim \pi N(0,\sigma^2_u+\sigma^2_b) + (1-\pi)N(0,\sigma^2_b)$	The mixture of two normal distributions e.g. BSLMM (Zhou *et al.* 2013b)
	$u_i \sim \pi N(0,\sigma^2_{u_i})+(1-\pi)N(0,0.01\sigma^2_{u_i})$	$\sigma^2_{u_i} \sim \chi^{-2}(v,S)$	$\pi \sim uniform(0,1)$ v,S were fixed	$u_i \sim \pi t(0,v,S)+(1-\pi)t(0,v,0.01S)$	The mixture of t distributions e.g. BayesSSVS (Verbyla *et al.* 2009; Verbyla *et al.* 2010)
Spike-at-zero & Slabs	$u_i \sim N(0,\sigma^2_{u_i})$	$\sigma^2_{u_i} \sim \pi(\sigma^2_{u_i}=0) + (1-\pi)\chi^{-2}(v,S)$	$\pi \sim uniform(0,1)$ v,S were fixed	$u_i \sim \pi(u_i=0)$ $+(1-\pi)\mathrm{t}(0,v,S)$	The mixture of point mass at zero and t distribution e.g. BayesB (Meuwissen *et al.* 2001)
	$u_i \sim \pi(u_i=0) + (1-\pi)N(0,\sigma^2_{u_i})$	$\sigma^2_{u_i} \sim \chi^{-2}(v,S)$	$\pi \sim uniform(0,1)$ v,S were fixed		The mixture of point mass at zero and t distribution but with the same variance e.g. BayesC(π) (Habier *et al.* 2011)
		$\sigma^2_{u_i} \sim \chi^{-2}(v,S)$	$S \sim gamma(1,1)$ $\pi \sim uniform(0,1)$		The mixture of point mass at zero and t distribution but with the same variance e.g. BayesD (Habier *et al.* 2011), BayesDπ(Habier *et al.* 2011)
	$u_i \sim \pi_1 N(0,0.0001\sigma^2_u) + \pi_2 N(0,0.001\sigma^2_u) + \pi_3 N(0,0.01\sigma^2_u) + \pi_4(u_i=0)$	σ^2_u was fixed	$\sum_{i=1}^{4}\pi_i = 1$ $\pi_i \sim Dirichlet(\alpha)$	$u_i \sim \pi_1 N(0,0.0001\sigma^2_u) + \pi_2 N(0,0.001\sigma^2_u) + \pi_3 N(0,0.01\sigma^2_u) + \pi_4(u_i=0)$	The mixture of point mass at zero and three normal distributions e.g. BayesR (Erbe *et al.* 2012; Moser *et al.* 2015), BayesRC (MacLeod *et al.* 2016)

Figure 1.2: **Alphabet bayésien**
Présentation d'une partie des modèles (liste non-exhaustive) constituant l'"alphabet bayésien", extrait de la thèse de Wang (2016). Ces modèles se différencient sur la distribution *a priori* attribuée aux effets de SNPs, de leur variance, et éventuellement de leurs *hyperpriors*.

1.2.3.2 Focus sur le modèle BayesR

Modélisation de BayesR Malgré l'introduction d'un paramètre de mélange pour modéliser des SNPs à effet nul et non-nul, les effets des SNPs restent fréquemment sous-estimés, notamment avec l'augmentation de la densité des génotypes. L'hypothèse que tous les effets des marqueurs inclus dans le modèle suivent *a priori* la même distribution semble par ailleurs peu plausible. Il a aussi été montré que l'utilisation d'un mélange gaussien permet d'estimer des nombreuses distributions via son paramètre de mélange (McLachlan and Basford, 1988). Dans cette optique, Erbe et al. (2012) ont proposé BayesR, qui cherche à estimer au mieux la distribution des effets des marqueurs en les répartissant en différentes classes d'effets. Un mélange de quatre distributions est utilisé, une première représentant la classe des marqueurs nuls, modélisée par un Dirac en 0, et les trois autres par des lois gaussiennes centrées, de variance représentant un pourcentage fixé de la variance additive totale σ_g^2 (respectivement de 0.01%, 0.1%, 1%):

$$f(\beta_j) = \sum_{k=1}^{4} \pi_k f_k(\cdot|\theta_k)$$

$$\text{telle que } f_k = \begin{cases} \delta(0), \text{ si } k = 1 \\ \phi(\cdot|0, \theta_k) \text{ sinon} \end{cases}$$

où $\boldsymbol{\theta} = (\theta_2, \theta_3, \theta_4) = (0.0001\sigma_g^2, 0.001\sigma_g^2, 0.01\sigma_g^2)$, $\sum_{k=1}^{4} \pi_k = 1$, $\delta(0)$ représente un Dirac en 0, et ϕ est la densité de probabilité gaussienne centrée. Par la nature multimodale de la distribution des effets des SNPs, BayesR permet donc à la fois de prendre en compte la sparcité des effets des marqueurs, mais aussi permet aux marqueurs à effets non-nuls de se distinguer des autres. Le modèle est alors suffisamment flexible pour couvrir diverses distributions pour divers caractères. Moser et al. (2015) ont montré le potentiel de BayesR pour à la fois mieux comprendre les caractères génétiques prédits, en plus d'une bonne qualité de prédiction. Cela suggère un potentiel d'intreprétabilité accru par rapport aux méthodes précédentes, et de mise en valeur de l'information biologique sous-jacente du caractères étudié.

BayesR est un modèle hiérarchique bayésien complexe comportant plusieurs niveaux de *priors* (Figure 1.3). Toutes les inconnues nécessitent une distribution *a priori*:

- La moyenne de la population π suit une loi uniforme non informative, $\pi(\mu) \propto 1$
- Les effets des SNPs β_i suivent un *a priori* un mélange gaussien *zero-inflated* défini précédemment (Equation 1.6) pour tout i
- Le paramètre de mélange $\boldsymbol{\pi} \sim$ Dirichlet(α), avec $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ défini comme un paramètre connu tel que $\alpha = (1, 1, 1, 1)$

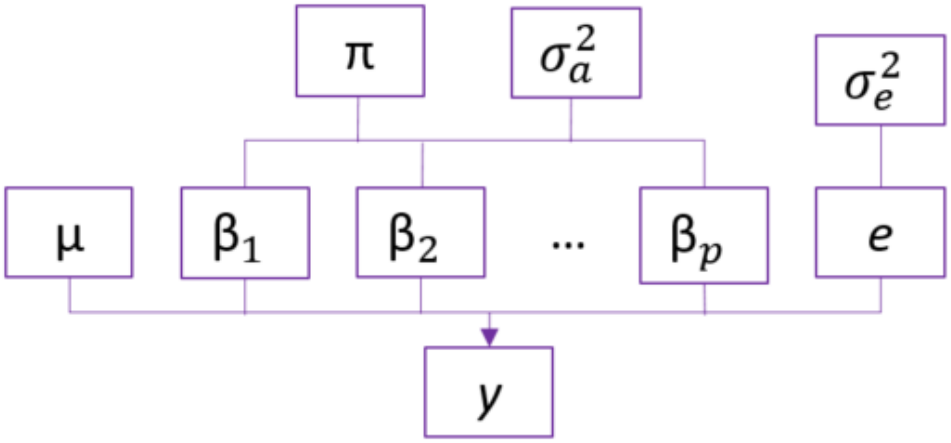


Figure 1.3: **Représentation du modèle hiérarchique bayésien de BayesR**

- La variance additive totale $\sigma_g^2 \sim$ inv $-\chi^2(v_0, S_0^2)$, avec v_0 et S_0^2 paramètres connus tels que $v_0 = -2$ et $S_0^2 = 0$
- Les résidus $e_i \sim N(0, \sigma_e^2)$ pour tout i
- La variance résiduelles $\sigma_e^2 \sim$ inv $-\chi^2(v_0, S_0^2)$, avec v_0 et S_0^2 paramètres connus tels que $v_0 = -2$ et $S_0^2 = 0$

Les hyperparamètres des lois $\boldsymbol{\pi}$, σ_g^2 et σ_e^2 sont choisis de telle sorte que les distributions des *priors* soient considérés comme "flat", c'est-à-dire comme étant non informatifs.

Algorithme et implémentation de BayesR La loi jointe des paramètres de BayesR est intractable, et nécessite donc l'utilisation d'un échantillonneur de Gibbs pour inférencer les paramètres. Les distributions conditionnelles *a posteriori* sont plus accessibles, grâce à l'utilisation de *priors* conjugués. Des variables latentes ont été utilisées:

1. $\boldsymbol{b} = (b_1, b_2, b_3, b_4)| \cdot \sim$ Multinomiale$(p, 4, \pi)$, qui quantifie le nombre de marqueurs appartenant à chaque classe de taille d'effet
2. $\boldsymbol{k}_j| \cdot \sim$ Categorielle$(4, r^{(j)}) \, \forall j \in 1, 2, ...p$, avec $r^{(j)} = (r_1^{(j)}, r_2^{(j)}, r_3^{(j)}, r_4^{(j)})$ la probabilité que le SNP j appartienne à chacune des quatre classes de taille d'effet

On définit les probabilités $\boldsymbol{r}^{(j)}$ par part de vraisemblance conditionnel de marqueur j d'être dans la classe k:

$$r_k^{(j)} = \frac{L(j|k)}{\sum_{\ell=1}^{4} L(j|\ell)} \tag{1.6}$$

Par simplification, on utilisation la notation $X|\cdot$ pour exprimer la conditionnalité d'une variable (ici X) par rapport à toutes les autres. On estime donc les paramètres en suivant les lois conditionnelles suivantes :

- $\mu|\cdot \sim N(\frac{\sum_{i=1}^{n}(y_i - \boldsymbol{X}_{i.}\boldsymbol{\beta})}{n}, \frac{\sigma_e^2}{n})$, avec la notation $\boldsymbol{X}_{i.}$ indiquant le vecteur ligne i de la matrice $\boldsymbol{X}$
- $\forall j,\ \ \beta_j|\cdot \sim N(\frac{\sum_{i=1}^{n} X_{ij}\tilde{y_i}}{\sum_{i=1}^{n} X_{ij}^2 + \sigma_e^2/(k_j\sigma_g^2)}, \frac{\sigma_e^2}{\sum_{i=1}^{n} X_{ij}^2 + \sigma_e^2/(k_j\sigma_g^2)})$, où $\tilde{y_i} = y_i - \mu - \boldsymbol{X}_{i.}\boldsymbol{\beta} + X_{ij}\beta_j$ le phénotype corrigé pour la moyenne et tous les marqueurs exceptés j.
- $\boldsymbol{\pi}|\cdot \sim$ Dirichlet$(1+b_1, 1+b_2, 1+b_3, 1+b_4)$
- $\sigma_g^2|\cdot \sim \text{inv}-\chi^2((p-b_1)-2, \frac{\sum_{j=1}^{p}\beta_j^2}{(p-b_1)-2})$
- $\sigma_e^2|\cdot \sim \text{inv}-\chi^2(n-2, \frac{\sum_{i=1}^{n}(y_i - \mu - \boldsymbol{X}_{i.}\boldsymbol{\beta})^2}{n-2})$

Concrètement, pour chaque itération, chaque marqueur est dans un premier temps assigné à une des quatre classes d'effet, avec une probabilité proportionnelle à la vraisemblance conditionnelle respective, avant d'avoir un effet estimé en fonction de la variance associée à cette classe d'effet. Les grandes étapes de l'algorithme utilisé pour l'échantillonneur Gibbs sont explicitées sous forme du pseudo-code 1.

Algorithm 1 BayesR pseudo-code

```
Initialization
for each iteration do
    update μ|.
    for i in 1:p do
        ỹ = ỹ + x_{.,i}β_i
        for k in 1:4 do
            compute LogL(i, k|π)
        end for
        assign SNP i to k̂ ∈ {1,2,3,4} |LogL(i, .|π)
        update β_i|k̂
        ỹ = ỹ − x_{.,i}β_i
    end for
    update σ_g^2, σ_e^2, π
end for
```

L'estimateur bayésien des différents paramètres, sous la fonction de coût quadratique, consiste en l'espérance de la distribution *posteriori* des paramètres. On l'approche ici comme la moyenne des échantillonages obtenus au cours des itérations, en retirant les 20000 premières itérations comme étape de *burn-in*, et en prenant en compte les valeurs toutes les 10 itérations (*thinning rate*).

1.3 Exploitations des connaissances biologiques acquises

Les modèles présentés jusqu'ici supposent que les variants ont une chance équivalente d'avoir un effet sur le caractère. Cependant, des connaissances accumulées au fil des études semblent suggérer que certaines régions du génome ont des fonctions et des rôles différents. Il serait donc pertinent de relâcher cette hypothèse d'équivalence, et que les chances pour les SNPs d'être inclus dans le modèle varient en fonction des régions du

génome, en introduisant une pondération. En particulier, cela pourrait être intéressant dans des cas de données en séquençage complet - WGS - qui ont montré peu d'amélioration de la prédiction en raison des difficultés à prioriser les mutations causales parmi les autres SNPs. De plus, cela pourrait améliorer la prédiction dans des situations de prédiction plus complexes, telles que la prédiction inter-races, ou de petit effectif, ou encore pour de faible héritabilité. L'utilisation d'informations biologiques connues est alors une piste pour améliorer la qualité de la prédiction, et mieux comprendre l'architecture génétique des caractères.

1.3.1 Une grande hétérogénéité d'informations biologiques disponibles

La notion d'information biologique est définie ici de façon assez large. A partir du moment où une région est caractérisée fonctionnellement à partir d'une étude, c'est une information biologique que nous souhaiterions potentiellement exploiter dans les modèles de prédiction génomique. Nous utiliserons principalement le terme générique d'"annotation", qui représente une classification binaire de l'appartenance ou non des SNPs à une région caractérisée par une fonctionnalité spécifique. On peut donc avoir accès à de multiples annotations, hétérogènes, et provenant d'études variées. Pour clarifier nos propos, nous présentons par la suite le type d'annotations exploitables, de façon non-exhaustive.

1.3.1.1 Description des données type -omiques

On appelle "omiques" (*omics* en anglais) le regroupement des disciplines biologiques associées aux différents mécanismes moléculaires s'opérant du génome au phénotype. Elles ont ainsi pour but d'identifier, caractériser, et quantifier toutes les molécules biologiques impliquées dans la structure, la fonction et les dynamismes d'une cellule, d'un tissu ou d'un organisme (Vailati-Riboni et al., 2017). En règle générale, on qualifie ces données de cellule simple (*single-cell*) si elles ont été étudiées au niveau de la cellule, et *bulk* si elles ont été étudiées au niveau de cellules regroupées. Cela comprend:

- la génomique, que l'on a déjà définie en 1.1.2;
- la transcriptomique, qui est l'étude de l'expression de gènes, via la quantification de l'ARN messager;
- la protéomique, qui quantifie les protéines présentes, et étudie leur propriétés biochimiques et fonctionnelles;
- la métabolomique, qui étudie les processus chimique impliquant des métabolites, donc les traces chimiques des mécanismes cellulaires;
- l'épigénomique, qui étudie les variations de l'ADN ou de l'ARN qui interfèrent avec l'expression des gènes, recouvrant tous les phénomènes de régulation génomique;
- la microbiomique, qui caractérise la composition bactériale et virale d'un échantillon.

	Assay	Platform	Main advantages and disadvantages	Standard bioinformatics pipelines
Genomics	Identify nucleotide variants (SNPs) in the whole genome associated with clinical traits (GWAS)	Genotyping arrays, whole-exome sequencing	SNP variability is stable during life; provides limited information in complex diseases due to several loci implicated	GWAS protocol review [10]
Transcriptomics	Quantify expression levels of cellular transcripts (*e.g.* mRNA)	Expression arrays, RNA sequencing	Widely used due to its high information content on cell status; differences in mRNA expression do not imply differences in proteins; does not take into account post-transcriptional modifications	RNA sequencing pipelines review [11]
Proteomics	Characterise protein expression levels of cells/samples	MS-based approaches	Expected to be closer to the phenotype; not widely used, expensive and more cumbersome analysis	Next-generation proteomics review [12]
Metabolomics	Characterise abundance profile of metabolites and their relative ratios	MS-based approaches	Representative of the cellular status; applicable to many biological fluids (*i.e.* breath, blood, urine, *etc.*); not widely used	Review of analytical methods for metabolomics [13]
Epigenomics	Determine modifications in DNA and small RNA that interfere with gene expression	DNA methylation analysis with arrays (Infinium MethylationEPIC 850K; Illumina, San Diego, CA, USA), next-generation sequencing, small RNA sequencing, arrays, *etc.*	Provides additional information to transcriptomics; related to exposures; more expensive than transcriptomics; sequencing-based approaches have computational tools in active development	Bioinformatics aspect of DNA methylation studies [14]
Microbiomics	Characterise bacterial (and viral) composition of a sample	Targeted sequencing of 16S rRNA gene, shotgun metagenomics sequencing	Provides information of external factors likely to be associated with disease; 16S sequencing does not differentiate between the presence of live/dead bacteria	Bioinformatics analysis for the characterisation of the human microbiome [15]

SNP: single nucleotide polymorphism; GWAS: genome-wide association study; MS: mass spectrometry.

Figure 1.4: **Récapitulatif des types de donnés omiques**
Extrait de Noell et al. (2018). Ce tableau regroupe l'utilité, les technologies, les avantages et désavantages et les techniques bioinformatiques pour les 6 types de données omiques présentées (génomique, transcriptomique, protéomique, métabolomique, épigénomique et microbiomique).

Ces différentes types d'approches font appel à des technologies qui sont détaillées dans le tableau 1.4 (Noell et al., 2018). Il est donc possible de construire des annotations du génome à partir de ces données, qui captent des informations complémentaires aux données de génotypage. On peut par exemple considérer un SNP contenu dans ou à proximité d'un gène différentiellement exprimé pour un phénotype ayant un lien avec le caractère de production à prédire comme étant dans l'annotation, ou annoté. En effet, on peut faire l'hypothèse qu'un variant ayant un lien avec un gène exprimé a plus de chance d'avoir un effet sur le caractère à prédire qu'un proche d'un gène qui ne s'exprime pas. Il est possible d'utiliser d'autres critères que la proximité, comme les eQTLs, les régions *upstream*, ou autres informations qui lieraient certains SNPs à ces gènes différentiellement exprimés. Nous pouvons procéder de façon analogue pour les autres types de données omiques, en cherchant à associer des SNPs aux résultats de ces études, notamment en priorisant les variants non méthylés, ou les zones de chromatine accessible. Une grande variété existe donc pour caractériser le génome, conduisant à une hétérogénéité des données, potentiellement multi-omiques, multi-tissus, multi-temporelles, etc.

1.3.1.2 Caractériser le génome à l'aide d'annotations structurelles

En fonction de leur positionnement par rapport à des gènes, ou dans un codon, les variants génétiques peuvent avoir une fonction différente dans un processus biologique cellulaire. Ces dernières années, un effort a été fait par des consortiums internationaux (Functional Annotation of Animal Genomes, FAANG, https://www.faang.org/) pour annoter les génomes de façon structurelle et fonctionnelle. Par souci d'uniformité lexicale, la Sequence Ontology (SO) a été développée, avec des descriptions détaillées des attributs de chaque catégorie de variants. Certaines catégories ont un effet plus fort sur le processus de transcription (Figure 1.5). Par exemple, SO décrit un "stop_lost" comme un variant séquentiel dont au moins une des bases du codon stop est modifiée, ce qui résulte en un transcript allongé, et ayant un effet fort sur le mécanisme de transcription. A l'inverse un "synonymous_variant", défini comme un variant de séquence dont la modification n'a aucun impact sur l'acide aminé produit, est considéré comme ayant un impact faible sur le processus de transcription. On peut aisément faire l'hypothèse que les SNPs annotés dans des catégories structurelles à effet modéré ou fort sont plus susceptibles d'avoir un impact sur le phénotype.

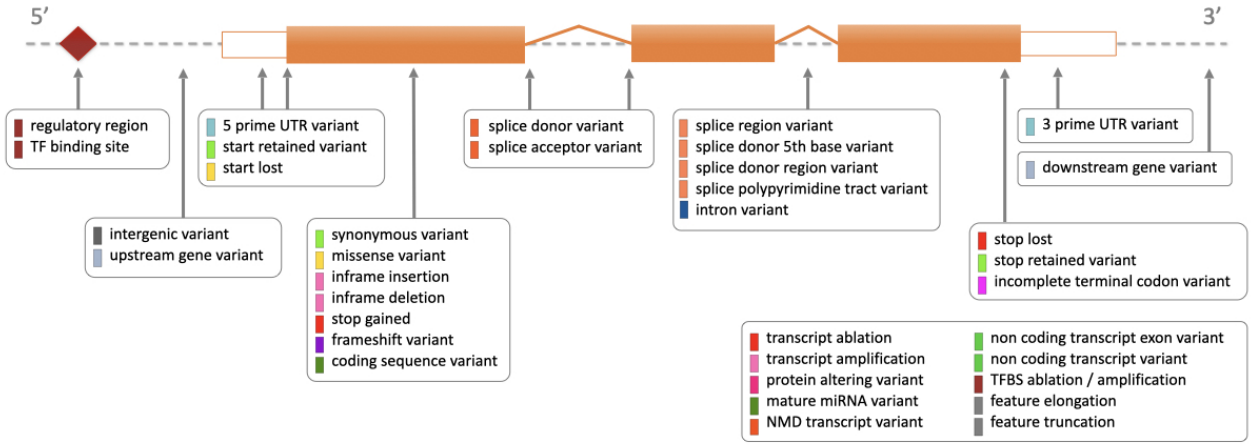


Figure 1.5: **Régions structurelles de transcription**
Description de la caractérisation des régions autour d'un gène, en fonction de leur positionnement, et qui peuvent avoir un impact sur le mécanisme de transcription dudit gène. Source: https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html.

1.3.1.3 Origine et construction des annotations fonctionnelles

Il est possible de construire des annotations soit à partir de son propre jeu de données à prédire (même génotypes), soit à partir d'un jeu de données provenant des mêmes individus (toutes informations issues d'une même étude), ou encore à partir de bases de données externes. Dans le premier cas, de premières analyses du génotype peuvent être effectuées, pour pouvoir les exploiter ensuite dans des modèles de prédiction. Il est par

exemple possible d'effectuer une ANOVA (ANalysis Of VAriance) univariée entre le phénotype et chaque SNP et de récupérer les p-valeurs (Gao et al., 2015), ou encore d'exploiter les déséquilibres de liaisons (Ramstein et al., 2016) ou la MAF (Speed et al., 2017). Il y a donc une multiple exploitation sur les mêmes données de génotypage, ce qui peut mener à un surapprentissage du modèle. Le deuxième cas correspond à l'accessibilité à d'autres données disponibles en plus des données de génotypages, collectées sur les mêmes animaux, pour une même étude, et pouvant donc être utilisées en complément direct pour la prédiction. Cependant, ces données appariées de phénotypes, génotypes, et autres omiques sont rarement de taille suffisamment importante pour permettre des modèles intégratifs de prédiction. Le troisième cas correspond aux multiples bases de données qui se développent rapidement, en exploitant l'accumulation de connaissances de ces dernières années. Celles-ci sont souvent publiques, et ont pour but de donner des outils pour mieux comprendre le génome. Certaines de ces bases de données sont issus d'efforts internationaux pour produire des annotations du génome, et ce pour différentes espèces. On peut citer le projet ENCODE (Harrow et al., 2012) chez l'humain, ou le consortium FAANG (Giuffra et al., 2019) chez les animaux. Des bases de données comme QTLdb (Hu et al., 2022) sont potentiellement de bonne sources d'annotations, mais regroupent différents résultats d'analyses de données, par conséquent très hétérogènes.

1.3.1.4 Formalisation binaire des annotations fonctionnelles

Dans cette thèse, nous nous intéressons majoritairement à ce que l'on nomme annotations binaires, où les SNPs peuvent être attribués ou non à une ou plusieurs catégories d'annotation. Si des SNPs ne sont présents dans aucune annotation, une dernière "annotation" les regroupant doit être créée afin de les prendre en compte dans le modèle de prédiction. Nous pouvons représenter cette information sous la forme d'une matrice d'annotations $A \in \{0,1\}^{\{p \times m\}}$ pour m annotations utilisées, sous condition que $\sum_{c=1}^{m} A_{j,c} \geq 1 \; \forall j \in \{1,2,..,p\}$ (Figure 1.6). On appellera A matrice d'annotations non chevauchantes si $\sum_{i=1}^{m} A_{j,c} = 1 \; , \forall j \in \{1,2,..,p\}$, et A matrice d'annotations chevauchantes si $\exists j \in \{1,2,..,p\}$ tel que $\sum_{i=1}^{m} A_{j,c} > 1$.

1.3.2 Comment intégrer ces informations ?

Dans un premier temps, ces informations étaient principalement utilisées en validation des résultats de prédiction. Plus récemment, de nouvelles approches ont été proposées afin de les prendre en compte directement dans les modèles de prédiction, dans l'espoir de mieux retranscrire les processus biologiques impliqués entre le génotype et le phénotype.

Figure 1.6: **Matrice d'annotations**
Mise en correspondance d'annotations fonctionnelles hétérogènes (multi-tissus, multi-omiques, multi-stades de développement) et des SNPs présents dans le génome. On peut formaliser cela sous la forme d'une matrice d'annotations binaires, avec 1 si le SNP appartient à une zone annotée, 0 sinon. Les SNPs non annotés sont regroupés en une dernière liste (sur la figure "unannotated").

1.3.2.1 État de l'art des méthodes de prédiction génomique intégrant des annotations fonctionnelles

Les approches utilisées pour l'intégration d'informations biologiques peuvent se diviser en deux types. On peut tout d'abord regrouper les méthodes exploitant de concert des données génomiques et des données omiques d'un autre type (RNA-seq, ChIPseq, ATACseq, etc.), que nous qualifierons de méthodes de prédiction multiomiques. Une façon répandue de procéder est d'ajouter des composantes omiques (souvent transcriptomiques et/ou métabolomiques) à l'équation de prédiction définie en 1.2, de façon additive (Xu et al., 2017; Hu et al., 2021; Wang et al., 2019) ou en prenant en compte les intéractions. Des méthodes de prédiction décrites dans la section 1.2.3 sont alors appliquées sur cette nouvelle équation, par exemple la régression lasso ou encore des méthodes de *machine learning*, comme le *support vector machine*. Ces méthodes nécessitent d'avoir les données multi-omiques pour tous les individus à prédire, ce qui en pratique reste une situation peu courante.

D'autres méthodes ont utilisé un partitionnement du génome, qui consiste à découper le génome en parties non vides, deux à deux disjointes, et à utiliser cette information dans les modèles de prédiction. Parmis ces méthodes de partitionnement, certaines séparent les annotations pour les utiliser une à une dans des modèles de prédiction classique, opérant ainsi un filtrage et pouvant les comparer, ou sélectionner les meilleurs (Do et al., 2015; Morota

et al., 2014; Abdollahi-Arpanahi et al., 2016). Une autre option est d'intégrer simultanément ces annotations dans un même modèle, en allouant des distributions différentes aux effets des SNPs appartenant à ces différentes catégories (Speed et al., 2017; Zhu and Stephens, 2018). En particulier, MacLeod et al. (2016) a adapté BayesR pour la prise en compte d'annotations, et a proposé un nouveau modèle, BayesRC, dans la continuité de l'alphabet bayésien. Flexible et performant, ce sera le point de départ des modèles développés dans cette thèse.

1.3.2.2 Focus sur le modèle BayesRC

BayesRC est un modèle de référence pour l'intégration d'informations hétérogènes, et présente deux avantages non négligeables. D'une part, les données introduites en plus du séquençage des SNPs proviennent d'autres analyses, recueillies sur d'autres individus. Cela est intéressant car en pratique il est rare d'avoir accès au séquençage génomique et à d'autres types d'informations cellulaires lors des évaluations qui mènent à la sélection. Deuxièmement, les informations sont utilisées sous forme de catégorisation binaire, cela offre la possibilité d'introduire de très diverses sortes d'informations, à partir du moment où il est possible d'obtenir une classification binaire pour les SNPs.

Modélisation de BayesRC BayesRC exploite des informations fonctionnelles en séparant les marqueurs en catégories disjointes, à partir du modèle existant BayesR. Le paramètre de mélange π de ce dernier lui permettait d'être flexible aux différents caractères complexes étudiés, en pouvant s'ajuster à l'enrichissement des SNPs à effet faible, moyen et fort présents dans le génotypage. Il est cependant raisonnable de faire l'hypothèse que cet enrichissement peut varier selon les régions du génome. BayesRC introduit alors les paramètres $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, ..., \boldsymbol{\pi}_m$ pour m annotations utilisées, tel que $\boldsymbol{\pi}_\ell \sim$ Dirichlet$(1,1,1,1),$ pour tout $\ell \in 1, 2, ..., m$. Ces paramètres de mélange, similairement à BayesR, s'ajustent à l'enrichissement des effets de SNPs dans chacune des annotations, permettant de les différencier, et de les pondérer. Formellement :

$$f(\beta_j | \boldsymbol{C}_j = c) = \sum_{k=1}^{4} \pi_{k,c} f_k(\cdot|\theta_k),$$

$$\text{telle que } f_k = \begin{cases} \delta(0), \text{ si } k = 1 \\ \phi(\cdot|0, \theta_k) \text{ sinon} \end{cases}$$

avec $C_j \in \{1, 2, ..., m\},\ \forall j$ l'annotation du SNP j, et les autres paramètres définis identiquement à (Equation 1.6). BayesR est un cas particulier de BayesRC, en considérant la présence d'une seule catégorie de SNPs, soit $m = 1$. La Figure 1.7 représente la structure hiérarchique de BayesRC (à deux annotations par souci de simplification).

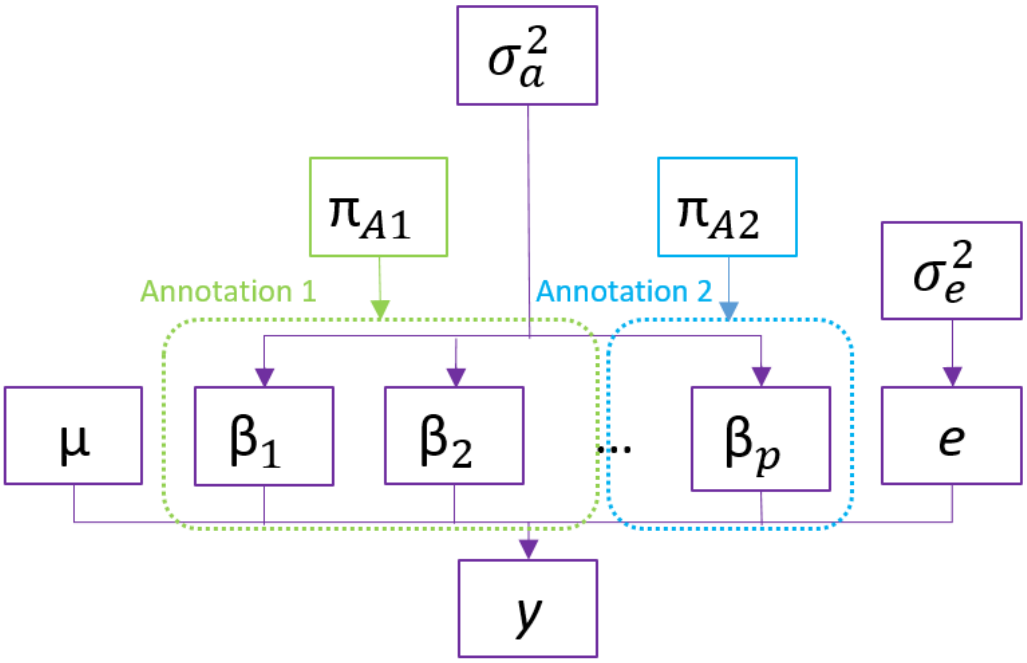


Figure 1.7: **Représentation du modèle bayésien hiérarchique à 2 annotations de BayesRC**

Algorithme et implémentation de BayesRC Comme pour BayesR, la loi jointe *posterior* est intractable, nécessitant un échantillonneur de Gibbs. Les étapes sont similaires à celle de BayesR, la principale différence se jouant sur le calcul des vraisemblances des SNPs, qui dépend du paramètre de mélange π_c.

Algorithm 2 BayesRC pseudo-code

Initialization
for each iteration **do**
  update $\mu|$.
  for i in 1:p **do**
    $\tilde{y} = \tilde{y} + x_{.,i}\beta_i$
    $c = C_i$
    for k in 1:4 **do**
      compute $LogL(i, k|c, \pi_c)$
    end for
    assign SNP i to $\hat{k} \in \{1, 2, 3, 4\}$ $|LogL(i, .|c, \pi_c)$
    update $\beta_i|\hat{k}$
    $\tilde{y} = \tilde{y} - x_{.,i}\beta_i$
  end for
  update $\sigma_g^2, \sigma_e^2, \pi_{c_1}, \pi_{c_2}, ...\pi_{c_m}$
end for

1.3.2.3 Extensions et limites d'utilisation de BayesRC

BayesRC est un modèle flexible qui peut être utilisé dans des contextes variés avec différents types d'annotations. Il a par ailleurs été adapté notamment pour l'étude des traits complexes humains, via l'exploitation de statistiques

résumés (*summary statistics*) issues de GWAS pour mieux prioriser les SNPs à potentiel effet fort (Zeng et al., 2021), mais aussi à l'analyse de survie, en combinant BayesRC et fonction de survie (Ojavee et al., 2022).
Pourtant, un inconvénient non négligeable de BayesRC n'a pas encore été abordé, qui est l'aspect non-chevauchant du partitionnement des SNPs en catégories d'annotation. Avec l'accumulation des annotations que l'on voudrait intégrer dans le modèle, nous augmentons la probabilité d'avoir des marqueurs appartenant à plusieurs annotations, et donc d'avoir des annotations chevauchantes. Plusieurs stratégies peuvent alors être mises en places pour utiliser BayesRC dans ces conditions comme :

1. l'attribution des SNPs multi-annotés dans une seule annotation, et ce de façon aléatoire ou basée sur d'autres critères

2. la création de nouvelles annotations, construites comme la jointure des annotations chevauchantes

La première stratégie implique de faire un choix, naïf ou non, qui a pour conséquence la perte d'une partie de l'information. La seconde stratégie génère des annotations potentiellement de très petit effectif, voir constitué d'un singleton. Cette situation pose problème pour l'inférence du paramètre de mélange par annotation, qui sera alors plus influencé par son hyperparamètre. De plus, il semble pertinent de se poser la question de la raison derrière l'existence d'un SNP multi-annoté, et de ne pas rejeter cette information.

Chapter 2

An evaluation of the predictive performance and mapping power of the BayesR model for genomic prediction

2.1 Résumé

Avant de chercher à intégrer des annotations fonctionnelles dans des modèles de prédiction génomique nous avons cherché à évaluer de quelle façon l'architecture génétique de caractères complexes est retranscrite avec BayesR. Trois principales raisons ont motivé ce choix de modèle:

1. BayesR est un des derniers modèles de l'alphabet bayésien largement diffusé, pourtant peu de publications sont disponibles sur ses capacités dans des configurations variées.

2. La distribution des effets des SNPs proposées par BayesR est plus fine que celles avancées par d'autres modèles, et pourrait donc mieux représenter l'architecture génétiques des caractères complexes. Un modèle permettant de bien retransmettre les mécanismes biologiques sous-jacents semble être un candidat pour exploiter des annotations fonctionnelles.

3. Bien comprendre et manipuler BayesR, ses avantages et inconvénients, est nécessaire à une bonne utilisation de BayesRC

Pour répondre à notre question, nous avons simulé un grand nombre de données simulées à partir de données de génotypages réelles, en prenant en compte un vaste panel d'architecture génétiques et d'héritabilité. Nous avons de plus évalué l'impact de l'exclusion ou de l'inclusion de variants causaux parmi les génotypes, en exploitant les

jeux de données simulés avec ou sans les QTLs les plus forts. Nous définissons enfin plusieurs critères statistiques, au niveau du SNP individuel et par fenêtres coulissantes, et comparons leur capacité à hiérarchiser précisément les QTLs simulés.

Nous avons trouvé que BayesR permettait de bonnes qualités de prédiction et cartographie de régions causales, ainsi qu'une bonne flexibilité face aux différentes architectures génétiques simulées. L'introduction de ses quatre classes d'effet (nul, faible, moyen et fort) offre de nouvelles façons de comprendre et caractériser l'architecture génomique d'un caractère complexe. Ainsi, on peut chercher à caractériser les SNPs en fonction de la variance *a posteriori* mais aussi à partir de leur fréquence d'inclusion à une de ces quatre classes. Nous avons aussi proposé un nouveau critère statistique efficace, le CIP (*weighted cumulative inclusion probability*), permettant à la fois de prendre en compte le LD via l'utilisation de fenêtres coulissantes, et d'exploiter les fréquences d'inclusion des SNPs dans chaque classes d'effet.

Ces résultats, présentés dans la suite de ce chapitre, ont été publiés dans la revue *G3: Genes, Genomes, Genetics* en 2021 (https://doi.org/10.1093/g3journal/jkab225).

G3, 2021, **11(11)**, jkab225
DOI: 10.1093/g3journal/jkab225
Advance Access Publication Date: 7 July 2021
Investigation

An evaluation of the predictive performance and mapping power of the BayesR model for genomic prediction

Fanny Mollandin,[1,*,†] Andrea Rau ,[1,2] and Pascal Croiseau [1]

[1]INRAE, AgroParisTech, GABI, Université Paris-Saclay, Jouy-en-Josas 78350, France
[2]BioEcoAgro Joint Research Unit, INRAE, Université de Liège, Université de Lille, Université de Picardie Jules Verne, Peronne 80203, France

*Corresponding author: INRAE, AgroParisTech, GABI, Université Paris-Saclay, Allée de Vilvert, Jouy-en-Josas 78350, France. Email: fanny.mollandin@inrae.fr
[†]Author contributed equally to this work.

Abstract

Technological advances and decreasing costs have led to the rise of increasingly dense genotyping data, making feasible the identification of potential causal markers. Custom genotyping chips, which combine medium-density genotypes with a custom genotype panel, can capitalize on these candidates to potentially yield improved accuracy and interpretability in genomic prediction. A particularly promising model to this end is BayesR, which divides markers into four effect size classes. BayesR has been shown to yield accurate predictions and promise for quantitative trait loci (QTL) mapping in real data applications, but an extensive benchmarking in simulated data is currently lacking. Based on a set of real genotypes, we generated simulated data under a variety of genetic architectures and phenotype heritabilities, and we evaluated the impact of excluding or including causal markers among the genotypes. We define several statistical criteria for QTL mapping, including several based on sliding windows to account for linkage disequilibrium (LD). We compare and contrast these statistics and their ability to accurately prioritize known causal markers. Overall, we confirm the strong predictive performance for BayesR in moderately to highly heritable traits, particularly for 50k custom data. In cases of low heritability or weak LD with the causal marker in 50k genotypes, QTL mapping is a challenge, regardless of the criterion used. BayesR is a promising approach to simultaneously obtain accurate predictions and interpretable classifications of SNPs into effect size classes. We illustrated the performance of BayesR in a variety of simulation scenarios, and compared the advantages and limitations of each.

Keywords: genomic prediction; QTL mapping; Bayesian model

Introduction

The primary objective of genomic prediction is to use genomic variation, usually single nucleotide polymorphisms (SNPs), to predict phenotypes, *i.e.*, an observable trait of an individual. In particular, genomic prediction models are widely used as an evaluation tool for genomic selection in plant (Heslot *et al.* 2015; Voss-Fels *et al.* 2019) and animal breeding (Meuwissen *et al.* 2001), and for the calculation of polygenic risk scores for human diseases (Wray *et al.* 2019). As genotyping costs have declined (Mardis 2017), there has been a corresponding increase in the amount of genotyping data available for analysis. In addition, lower costs and better data storage capacity have allowed for increasingly dense genotypes, up to and including whole-genome sequences (WGSs), which in turn have enabled sequence-level genotypes to be imputed for individuals genotyped using lower density chips (Marchini *et al.* 2007). However, analyzing these increasingly large genotype data can come at a high computational cost and requires suitable statistical methods. Although the use of higher density genotypes was initially thought to hold promise for improved prediction accuracy, their performance was not found to improve that of high-density chips in real data, due to the inclusion of a large number of noncausative SNPs (Pérez-Enciso *et al.* 2015). While the exhaustive use of WGS variants has not led to meaningful improvements in prediction, they do allow for the direct inclusion of candidate, or even causal, mutations (Liu *et al.* 2020). For simplicity, we refer to such mutations as quantitative trait loci (QTL) throughout. If such QTLs are known a priori or can be directly identified through variable selection in the model itself, this could potentially lead to the double advantage of improving both the accuracy and interpretability of genomic prediction models (Brøndum *et al.* 2015; Van den Berg *et al.* 2016). With this in mind, custom chips, which include SNPs from a medium-density chip (intended to cover the genome) as well as candidates or causal mutations for a set of traits, have been developed, offering the cost and computational advantages of a reasonably sized chip with the increased predictive ability and interpretability provided by the inclusion of potential causal mutations.

Most models used in routine genomic selection are based on linear models, notably best linear unbiased prediction (BLUP) and genomic BLUP (GBLUP). These models assume that all SNPs contribute equally to the genomic variance, with each SNP effect following a normal distribution with common variance. Although the assumption about common SNP effects allows for great computational efficiency, it is quite strong and can limit the biological

Received: March 02, 2021. **Accepted:** June 27, 2021

interpretability of results. To address this limitation, although deep learning models have recently started to appear (Bellot *et al.* 2018; Abdollahi-Arpanahi *et al.* 2020), a more frequent alternative is the set of Bayesian models comprising the so-called Bayesian alphabet. These include, among others, BayesA (Meuwissen *et al.* 2001), BayesB (Meuwissen *et al.* 2001), BayesCπ (Habier *et al.* 2011), BayesR (Erbe *et al.* 2012), and BayesRC (MacLeod *et al.* 2016). The aim of all of these models is to improve predictive accuracy by better estimating SNP effects through more flexible prior specifications. For instance, in the earliest model introduced, BayesA, all markers are assumed to be drawn from a normal distribution whose variance follows an Inv $-\chi^2$ distribution. Although the assumptions of BayesA are arguably closer to reality than BLUP or GBLUP, it is computationally expensive to estimate variances for every SNP in dense genotyping data. Instead, a useful alternative is to assume that a (potentially large) portion of markers contributes no genetic variance. This is the strategy employed by both BayesB and BayesC, which model marker effect variances as a zero-inflated distribution by assigning null effects with a fixed probability, and assuming the variance of nonnull SNPs respectively follow a per-SNP or common Inv $-\chi^2$ distribution. BayesCπ further assumes that the proportion of null SNP effects is itself a random variable, and otherwise uses a common prior distribution for nonnull SNP effects. BayesR provides additional flexibility by defining four classes of SNP effect size (null, small, medium, and large), where SNP effects are modeled using a four-component normal mixture model. The related BayesRC model further allows for SNPs to be grouped into disjoint categories (*e.g.*, according to prior biological information), for which the BayesR model is subsequently fit independently.

Although these Bayesian genomic prediction models are mainly used for phenotype prediction, they also provide valuable per-SNP information, including posterior estimates of effect size and variance, which could be used for QTL mapping. In contrast to genome-wide association study (GWAS) methods, SNP effects are estimated simultaneously and make use of variable selection within the model itself, rather than relying on univariate hypothesis tests and corrections for multiple testing. As the quantity and quality of prior biological knowledge continues to improve and the identification of causal mutations from WGS data (Sanchez *et al.* 2016) becomes increasingly feasible, the flexible model definition of BayesR and BayesRC thus make them interesting candidates for simultaneously providing good predictability and biologically interpretable QTL mapping results. In this spirit, Moser *et al.* (2015) showed encouraging results for the use of BayesR in complex traits for prediction and QTL mapping in real data. However, a comprehensive simulation study investigating the interpretability and performance of BayesR in a wide variety of settings is currently lacking in the literature. In addition, to date there has been little discussion of the various criteria that can potentially be used to map QTLs using the BayesR model output.

To address this gap, our goal in this study is to identify the coherence between the BayesR model specification and known QTL effects in simulated data under a variety of conditions. The BayesR approach is of particular interest here, as it has been shown in the literature to improve prediction accuracy (Zhu *et al.* 2019), but its ability to correctly assign QTLs to the appropriate effect size categories has not yet been extensively evaluated in simulations. We focus on the case where a prior categorization of markers (*i.e.*, the BayesRC approach) is not available. Using simulated data, we evaluate the robustness of BayesR under a wide variety of genetic architectures, phenotype heritabilities, and polygenic variances, and we illustrate the conditions under which BayesR successfully identifies known QTLs while maintaining high accuracy for phenotypic prediction. Finally, we describe and compare several statistical criteria that can be used to perform QTL mapping using BayesR output. Based on the results of our simulation study, we discuss the optimal framework for an accurate and interpretable analysis using BayesR, as well as its limitations.

Materials and methods

Data simulation based on real genotypes

To maintain a realistic linkage disequilibrium (LD) structure among SNPs, we generated simulated data based on a set of genotypes assayed using Illumina Bovine SNP50 BeadChip arrays from $n = 2605$ Montbéliarde bulls. We divided individuals into learning and validation sets (*i.e.*, the "holdout method"), with the 80% oldest bulls ($n_{\text{learning}} = 2083$) in the former and the 20% youngest ($n_{\text{validation}} = 522$) in the latter to reflect the strategy typically used in routine genomic selection. We excluded SNPs with a minor allele frequency (MAF) less than 0.01, leaving a total of $P = 46{,}178$ SNPs.

To simulate phenotypes $\mathbf{y}$ for the $n = 2605$ bulls, we made use of a standard linear model:

$$\mathbf{y} = \mu 1_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \mathbf{e} \sim \mathrm{N}(0, \mathbf{I}_n \sigma_e^2) \quad (1)$$

where μ denotes the intercept, $\boldsymbol{\beta}$ the vector of effects for the p SNPs, and $\mathbf{e}$ the vector of residuals which is assumed to follow a multivariate normal distribution with mean 0 and variance covariance matrix $\sigma_e^2\mathbf{I}$. $\mathbf{X}$ is the marker matrix, centered and scaled as: $x_{ij} = (w_{ij} - 2f_j)/\sqrt{2f_j(1-f_j)}$, with w_{ij} the number of copies of the reference allele {0,1,2} and f_j the frequency of the reference allele. Parameters for this linear model were set as follows. For each simulated dataset we sampled from the available SNPs a set of n_{QTL} QTLs and a set of n_{poly} polygenic SNPs, as well as their corresponding genetic variances for each selected marker. To reduce the impact of extreme MAFs on genomic prediction (Uemoto *et al.* 2015) and QTL detection, we focused on frequent QTLs by drawing the n_{QTL} and n_{poly} SNPs from those with a MAF ≥ 0.15. In all simulations, we selected a total of $n_{\text{QTL}} = 5$ large QTLs, varying the corresponding proportion k of total genetic additive variance σ_g^2 as described below. The phenotypic variance and mean were respectively set to $\sigma_y^2 = 100$ and $\mu = 0$, and SNP heritability $h^2 = \frac{\sigma_g^2}{\sigma_y^2}$ was varied across simulation settings.

We constructed 13 scenarios with different proportions k of genetic variance attributed to the QTLs, with 10 independent datasets created for each (Table 1). For the SNPs randomly selected as QTLs and polygenics SNPs, the corresponding effect β_i for selected SNP i was set as follows:

$$\beta_i = \begin{cases} \frac{1}{2} u_i \sqrt{\frac{10^{-4}\sigma_g^2}{2MAF_i(1-MAF_i)}} & \text{if SNP}_i \text{ is polygenic} \\ \frac{1}{2} u_i \sqrt{\frac{k\sigma_g^2}{2MAF_i(1-MAF_i)}} & \text{if SNP}_i \text{ is a QTL} \end{cases},$$

where u_i was drawn from a discrete Uniform$\{-1, 1\}$ distribution to allow nonnull effects to take on positive or negative values. For unselected SNPs (*i.e.*, null SNPs), β_i was set to 0. We varied the proportion of genetic variance attributed to each QTL between $k = 0.725$ and 5%, with a greater density of values evaluated between 0.725 and 2%; we focused in particular on this range as it

Table 1 Simulation settings for each of the 13 QTL effect-size scenarios considered for each given level of heritability, $h^2 = \{0.1, 0.3, 0.5, 0.8\}$

Number of QTLs	5	5	5	5	5	5	5	5	5	5	5	5	5
Number of polygenic SNPs	9637	9550	9500	9450	9350	9250	9100	9000	8750	8500	8250	8000	7500
Per-QTL % of σ^2_g	0.725	0.9	0.10	0.11	0.13	0.15	0.18	2	2.5	3	3.5	4	5
Per-polygenic SNP % of σ^2_g	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

The number of simulated QTLs, number of polygenic SNPs, percentage of genetic variance attributed to each QTL, and percentage of genetic variance attributed to each polygenic SNP are provided. Summing the percentage of genetic variance explained by the total number of QTLs and polygenic SNPs yields 100%.

corresponds to more plausible QTL sizes and facilitated a study of the sensitivity of BayesR to small changes. For each value of k, the same $n_{QTL} = 5$ QTLs were used across scenarios, but the number (and thus the subset) of polygenic SNPs used varied (see Table 1). As the same 5 QTLs were simulated across scenarios for each of the 10 independent datasets, a total of 50 QTLs was considered. Finally, each scenario was run for four different levels of heritability $h^2 = \{0.1, 0.3, 0.5, 0.8\}$, and we evaluated the performance of BayesR for two alternatives: (1) using genotype data that excludes the 5 known QTLs, resembling a classic 50k genotyping array ("50k data"); and (2) using genotype data that includes the 5 known QTLs, which mimics a custom 50k genotyping array ("50k custom data"). In total, this corresponds to $13 \times 10 \times 4 \times 2 = 1040$ simulated datasets.

To investigate the sensitivity of BayesR to a different underlying genetic architecture, we also simulated a secondary set of simulations in an analogous manner with 5 large QTLs as well as 5 additional intermediate QTLs, whose percentage of genetic variance was set to 10% of that of the large QTLs. The settings for these additional simulations are described in Supplementary Table S1.

Statistical analysis

BayesR genomic prediction model: The models of the Bayesian alphabet are all based on the linear model in Equation (1). BayesR assumes that SNP effects β_i follow a four-component normal mixture, making it well-aligned to our simulations (for which SNPs fall into null, weak, and strong classes). The effect of SNP i is assumed to be distributed as

$$\beta_i \sim \pi_1(\beta_i = 0) + \pi_2 N(0, 0.0001\sigma^2_g) + \pi_3 N(0, 0.001\sigma^2_g) + \pi_4 N(0, 0.01\sigma^2_g), \quad (2)$$

where as before, σ^2_g represents the total additive genetic variance (*i.e.*, the cumulative variance of all SNP effects) and $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$ the mixing proportions such that $\sum_{i=1}^{4} \pi_i = 1$. The mixing proportions π are assumed to follow a Dirichlet prior, $\pi \sim \text{Dirichlet}(\alpha + \gamma)$, with α representing a vector of pseudocounts and γ the cardinality of each component. In this study, we used a flat Dirichlet distribution, with $\alpha = (1, 1, 1, 1)$, for the prior. As suggested by Moser *et al.* (2015), σ^2_g is assumed to be a random variable following an $\text{Inv} - \chi^2(\upsilon_0, S^2_0)$ distribution, with hyperparameters $\upsilon_0 = -2$ and $S^2_0 = 0$, which leads to an improper flat prior distribution.

As exact computation of the posterior distribution is intractable for this model, Bayesian inference is performed by obtaining draws of the posterior using a Gibbs sampler; full details of the algorithm can be found in Moser *et al.* (2015) and Kemper *et al.* (2015). In practice, at each iteration of the algorithm, SNPs are assigned to one of the four categories, and their effect is subsequently sampled from the full conditional posterior distribution for the corresponding mixture component. Model parameters are then estimated using the posterior mean across iterations, after excluding the burn-in phase and thinning draws. Here, the Gibbs sampler was run for a total of 50,000 iterations, including 20,000 as a burn-in and a thinning rate of 10.

In this study, we used the open source Fortran 90 code described in Moser *et al.* (2015) and available at https://github.com/syntheke/bayesR. We made a few modifications to this code, notably adding the posterior variance of estimated SNP effects at each iteration to the output; our modified BayesR code may be found at https://github.com/fmollandin/BayesR_Simulations.

Prediction accuracy for BayesR was quantified using the Pearson correlation between the true phenotypic values ($\mathbf{y}$) and those estimated using BayesR ($\hat{\mathbf{y}}$) in the validation set.

Statistical criteria for QTL mapping: In this section, we present several potential criteria based on BayesR output that can be used for the purpose of QTL mapping. We have sub-divided these criteria into those defined for (1) each SNP individually; (2) neighborhoods, or sliding windows, around each marker; and (3) those used for ranking potential QTLs.

Mapping criteria for individual SNPs: BayesR is unique in the Bayesian alphabet, in that it assigns SNPs to one of four effect size classes at each iteration by weighting according to their likelihood of belonging to each. We thus have access to the posterior frequency with which SNPs were assigned to each class, which can be interpreted as an inclusion probability. We denote the posterior inclusion probability (PIP) of SNP i belonging to class j as $\text{PIP}_i^{(j)}$, such that $\sum_{j=1}^{4} \text{PIP}_i^{(j)} = 1 \; \forall i \in \{1, \ldots, p\}$. In the following, we interchangeably refer to the null, small, medium, and large classes as $j = 1$, 2, 3, and 4, respectively. The PIP provides a straightforward method for classifying SNPs as having a null, small, medium, or large effect. We define the maximum a posteriori (MAP) rule for SNP i as

$$\text{MAP}_i = \underset{j}{\text{argmax}} \; \text{PIP}_i^{(j)}, \quad (3)$$

implying that SNPs are assigned to their most frequently assigned class. Since SNPs may move frequently from one class to another, the MAP in Equation (3) may not detect SNPs that are predominantly included in the model but move between the three nonnull classes. Merging the nonnull classes addresses this problem, and leads to a less stringent criterion, the nonnull MAP:

$$\text{MAP}_i^{\text{non-null}} = \begin{cases} 1 & \text{if } \text{PIP}_i^{(1)} < \sum_{j \in \{2,3,4\}} \text{PIP}_i^{(j)} \\ 0 & \text{else} \end{cases} \quad (4)$$

Based on this criterion, SNP i is thus included in the model if $1 - \text{PIP}_i^{(1)} > 0.5$. In this way, all SNPs preferentially assigned to the null class take on a value of $\text{MAP}_i^{\text{non-null}} = 0$, while those assigned to any nonnull class (small, medium, or large) take on a value of $\text{MAP}_i^{\text{non-null}} = 1$.

The BayesR model definition explicitly allows for some SNPs to have larger estimated variances than methods such as GBLUP, which tends to shrink the variance of causal markers due to the assumption of a common variance (Kemper *et al.* 2015). As such, BayesR has the potential for more closely approximating the true variance of QTLs. The posterior variance of SNP i corresponds to

$$V_i = \beta_i^2 \mathbf{X_i^T X_i}, \tag{5}$$

where $\mathbf{X_i}$ represents the i^{th} column of the centered and scaled genotype design matrix. As the SNP effects are computed on the scaled and centered genotype design matrix $\mathbf{X}$, the per-SNP posterior variance can be estimated using

$$\hat{V}_i = \hat{\beta}_i^2 \mathbf{X_i^T X_i} = \hat{\beta}_i^2,$$

where $\hat{\beta}_i^2$ corresponds to the posterior mean of β_i^2, $\hat{\beta}_i^2 = \frac{1}{n}\sum_{\ell=1}^{n} \beta_i^{(\ell)2}$, where n is the number of iterations and $\beta_i^{(\ell)2}$ the value of β_i^2 at iteration ℓ. We indirectly estimated this parameter as the sum of the posterior variance and squared posterior mean of each per-SNP effect. We can then estimate a posteriori the proportion of genetic variance of a SNP i as $\hat{V}_i / \sum_{j=1}^{p} \hat{V}_j$.

Neighborhood-based mapping criteria: LD represents a preferential association between two alleles and can have a large impact on how estimated variances are distributed among SNPs in an LD block. This in turn affects the evaluation of the variance in the neighborhood of a causal mutation, as well as the ability to perform QTL mapping using the aforementioned criteria, for several reasons. First, SNPs in close proximity to a QTL are likely to be in high LD with it, and thus may erroneously have their own effects overestimated to the detriment of the QTLs. The per-SNP criteria defined above risk incorrectly identifying a QTL as null in such cases. An alternative approach is to define a neighborhood-based criteria around each marker (Fernando *et al.* 2017), thus mapping QTLs when one or more or its close neighbors is detected. Here, we define each neighborhood as a sliding window of 15 SNPs (covering approximately 1 Mb) centered around each marker.

Using these neighborhoods, we define the vector of PIPs for a neighborhood centered on SNP i as follows:

$$\begin{aligned} \mathrm{PIP}_i &= (\mathrm{PIP}_i^{(1)}, \ldots, \mathrm{PIP}_i^{(4)}) = \mathrm{PIP}_{i'}, \text{with} \\ i' &= \underset{\ell \in \{i-7,\ldots,i,\ldots,i+7\}}{\operatorname{argmax}} \; 1 - \mathrm{PIP}_\ell^{(1)}, \end{aligned} \tag{6}$$

with the corresponding neighborhood inclusion probability equal to

$$\mathrm{IP}_i = (1 - \mathrm{PIP}_i^{(1)}). \tag{7}$$

The criteria proposed in Equations (3)–(5) can thus be adapted to accommodate neighborhoods as follows:

$$\begin{aligned} \mathrm{MAP}_i = \mathrm{MAP}_{i'} \text{and} \mathrm{MAP}_i^{\text{non-null}} &= \mathrm{MAP}_{i''}^{\text{non-null}}, \text{with} \\ i' &= \underset{\ell \in \{i-7,\ldots,i,\ldots,i+7\}}{\operatorname{argmax}} \; \mathrm{IP}_\ell, \end{aligned} \tag{8}$$

where SNP indices are assumed to be ordered according to their physical location. Similarly, the estimated variance of a neighborhood is fixed to the maximal value of its individual markers:

$$V_i = \max_{\ell \in \{i-7,\ldots,i,\ldots,i+7\}} V_\ell. \tag{9}$$

LD structure raises an additional related problem—in some cases, the BayesR algorithm may alternate assigning different SNPs in an LD block to the large effect class, which has the consequence of diluting variance over a region rather than for a single marker. The window-based criteria in Equations (8) and (9) successfully flag regions where a single SNP sufficiently stands out, but not necessarily those including several diluted effects. In addition, it can be difficult to accurately assess the variance over a region, due to the covariance among SNPs. To provide a neighborhood-level summary of SNP assignments to the four effect classes, we propose the following sliding-window statistic for SNP i, that we will call Weighted Cumulative Inclusion Probability (CIP_i):

$$\mathrm{CIP}_i = \sum_{\ell=i-7}^{i+7} (0 \times \mathrm{PIP}_\ell^{(1)} + 10^{-4}\mathrm{PIP}_\ell^{(2)} + 10^{-3}\mathrm{PIP}_\ell^{(3)} + 10^{-2}\mathrm{PIP}_\ell^{(4)}). \tag{10}$$

Finally, we used the Lewontin D' statistic (Lewontin 1964) to quantify the LD between SNPs. Briefly, the LD coefficient D_{AB} between SNPs A and B is defined as $D_{AB} = p_{AB} - p_A p_B$, where p_A, p_B, and p_{AB}, respectively denote the frequency of allele A in the first locus, allele B in the second, and the frequency of simultaneously having both. D′ normalizes D so that $D' = \frac{D}{D_{\max}}$, with

$$D_{\max} = \begin{cases} \max\{-p_A p_B, -(1-p_A)(1-p_B)\}, & \text{if } D < 0 \\ \min\{p_A(1-p_B), (1-p_A)p_B\}, & \text{otherwise.} \end{cases}$$

We will use the maximum value of the LD of a QTL with its neighboring SNPs as a reference for the LD in the region.

Criteria ranking for QTL mapping: For the quantitative criteria V_i and CIP_i defined in Equations (9), (5), and (10), we propose the use of rankings for SNP prioritization rather than fixing value thresholds. For QTL mapping based the estimated posterior variance V_i, we focus on the ten SNPs with the highest V_i. As CIP_i represents a sum over 15 SNPs in the neighborhood of SNP i, SNPs adjacent to those that are frequently categorized as nonnull are likely to share large values for this criterion. As such, to address this redundancy, we focus on the 150 SNPs with the highest CIP_i value.

Results and discussion

Results

In the following, we first investigate the sensitivity of BayesR to parameter specification. We next evaluate the model's performance for phenotype prediction and QTL mapping, based on the statistical criteria defined in the previous section, using simulated data that include a set of $n_{\mathrm{QTL}} = 5$ QTLs, as well as polygenic SNPs and null SNPs with no effect on the phenotype.

Sensitivity of BayesR parameter specification. Although the proportion of additive genetic variance assigned to the small, medium, and large effect classes is typically set to 0.01, 0.1, and 1%, respectively [see Equation (2) and Erbe *et al.* (2012)], these prior parameters can be varied by the user. To evaluate the impact on downstream results, we varied the latter between 0.5, 1, and 2% for all scenarios with $h^2 = 0.5$, leaving those of the small and medium effect classes at their default values. Modifying the proportion of genetic variance of the large effect class did not appear to

have a strong impact on the validation correlation; nevertheless, we observed differences in correlation among the three prior values that reached up to 2.6 and 1% for the 50k and 50k custom data respectively. However, the posterior mean of the number of SNPs assigned in each class and its associated posterior estimated variance do appear to be somewhat affected by this parameterization (Table 2). To assess the impact of the prior specification on per-SNP effect estimates, we calculated the Pearson correlation between the estimated posterior means $\hat{\beta}_i$ across SNPs, simulated scenarios and datasets. Among the three prior specifications, the correlation of estimated SNP effects (between pairs of prior parameter settings for a given proportion of additive genetic variance) was between 97.4 and 98.6% for all SNPs. We further evaluated the sensitivity to prior specification on our secondary simulations including both large and intermediate QTLs (Supplementary Table S1), and similar results across settings were observed for both the validation correlations and concordance of SNP effect estimates (results not shown).

Based on these results, we consider that the prior specification appears to have little practical impact on the performance of BayesR, whether for its predictive performance or for per-SNP effect estimates. For the remainder, we therefore use the default prior specification for proportion of genetic variance in each effect class. Note that the choice of the variance used for each component of the BayesR prior mixture distribution is primarily intended to improve mixing of the Markov chain, and no theoretical justification is provided by Erbe *et al.* (2012).

Predictive power of BayesR in varied simulation settings. We next sought to investigate the predictive power of BayesR across simulation scenarios, varying the contribution of QTLs to the additive genetic variance (which we refer to as scenarios below), heritability, and use of 50k or 50k custom genotype data.

The mean validation correlation (over the ten independent datasets simulated for each) for each simulation scenario illustrates the expected drop in prediction quality for decreasing heritabilities, whether 50k or 50k custom data are used (Figure 1). For the former, the mean (±sd) validation correlation across scenarios is 0.125 (±0.048), 0.301 (±0.057), 0.447 (±0.058), and 0.650 (±0.049) for $h^2 = \{0.1, 0.3, 0.5, 0.8\}$. For the latter, the inclusion of the true QTLs among the genotypes unsurprisingly leads to higher validation correlations, with mean (±sd) values across scenarios equal to 0.128 (±0.049), 0.312 (±0.058), 0.466 (±0.059), and 0.680 (±0.046) for $h^2 = \{0.1, 0.3, 0.5, 0.8\}$.

Although the trends of the mean validation correlation are nonlinear as the QTL effects take on an increasing percentage of genetic variance for both types of data, we do remark an increasing disparity in performance between the 50k and 50k custom data, particularly as the heritability itself increases (in Supplementary Figure S1). In particular, as expected the potential gain in including the true causal mutations among genotypes (as is the case of the 50k custom data) appears to be particularly strong for moderate to large heritabilities and QTL effects. For h^2=0.01, the average difference in validation correlation was 0.003 (±0.009), and in some cases, the use of the 50k custom data actually corresponded to a slightly worse prediction. Similar results are observed at this level of heritability regardless of the simulated effect size of the QTLs. However, for $h^2 = \{0.3, 0.5, 0.8\}$, 50k custom data led to a nearly systematic gain in performance: the average increase in validation correlation was 0.011 (±0.014), 0.019 (±0.020), and 0.031 (±0.030) across QTL effect size scenarios, and attained maximum values of 0.076, 0.112, and 0.160, respectively. For a given heritability, Supplementary Figure S1 also shows marked improvements in prediction when including QTLs simulated with large shares of additive genetic variance.

QTL mapping using BayesR. A natural first tool to investigate for QTL mapping is the neighborhood PIP defined in Equation (6). We focus on the behavior of the neighborhood PIPs for the true QTLs across scenarios (Figure 2), averaging over the 50 QTLs available for each (5 QTLs × 10 independent datasets); note that as this is a window-based measure, this measure can be computed for the true QTLs whether the 50k or 50k custom data are used. As shown in Figure 2, the allotment of true QTL neighborhoods to effect classes varies widely across heritabilities, proportion of genetic variance for each QTL, and type of data used. Globally, assigning QTL neighborhoods to nonnull effect classes, particularly the large effect class, is more frequent for larger heritabilities and simulated QTL effect sizes, as well as for 50k custom compared to 50k data. However, this difference disappears for small heritabilities; when $h^2 = 0.1$, the average (±sd) neighborhood PIP for the null class across scenarios is 0.91 (±0.009) and 0.90 (±0.013) for the 50k and 50k custom data, respectively. Across scenarios, we observe a similar usage of the small effect class, with an average corresponding neighborhood PIP of 0.08 (±0.007) regardless of the genotyping data used. When $h^2 = \{0.3, 0.5, 0.8\}$, as the simulated share of genetic variances for QTLs increases for both the 50k and 50k custom data, the null neighborhood PIP decreases and the large-effect neighborhood PIP increases. Across all simulated datasets and scenarios, the average (±sd) small- and medium-effect neighborhood PIPs are 0.117 (±0.053) and 0.058 (±0.040), respectively, illustrating that these two classes appear to be less often filled compared to the null and large classes (although all four classes do appear to be used outside of the lowest heritability setting).

The neighborhood PIP results provide a preview of how QTLs are grouped into nonnull effect classes according to the neighborhood MAP rule [Equation (8); Figure 3]. In all simulation settings, no QTL neighborhoods were assigned to the small effect class using this criterion. When $h^2 = 0.1$, without surprise, all QTLs were classified as null. For $h^2 = 0.5$, a very small number of QTL neighborhoods were assigned to the medium effect class for the 50k data; increasing to $h^2 = 0.8$ led to a larger number moving to this

Table 2 Average (across all simulation scenarios and independent datasets) of the posterior mean cardinality of each BayesR SNP effect class (null, small, medium, large) for three parameterizations of the prior large effect class variance

Prior large class variance (%)	Null	Small	Medium	Large	V_{null}	V_{small}	V_{medium}	V_{large}
0.5	40,783.25	5,054.51	300.44	39.80	0	25.55	14.89	9.60
1 (default)	40,568.94	5,256.72	336.12	16.23	0	26.65	6.91	10.15
2	40,501.21	5,307.33	361.08	8.38	0	26.81	17.97	5.45

For a given dataset, each class size (#) is computed as the posterior mean of the number of SNPs assigned to each class across iterations, and V_j is the posterior estimated cumulative variance of each class j.

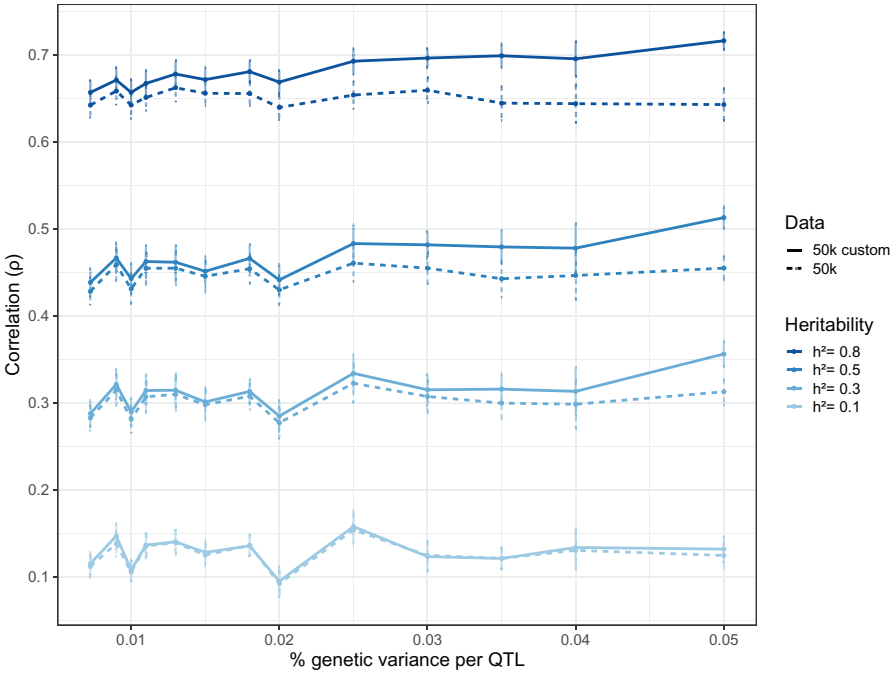


Figure 1 BayesR predictive performance across simulation settings. For each setting (h^2 and percentage of genetic variance assigned to each QTL), points represent mean validation correlations across 10 independent datasets. Heritability values are represented by dark to light blue ($h^2 = 0.8$ to 0.1), and solid and dotted lines represent results for the 50k and 50k custom datasets, respectively. The error bars correspond to the Monte Carlo standard errors computed across the 10 datasets for each setting.

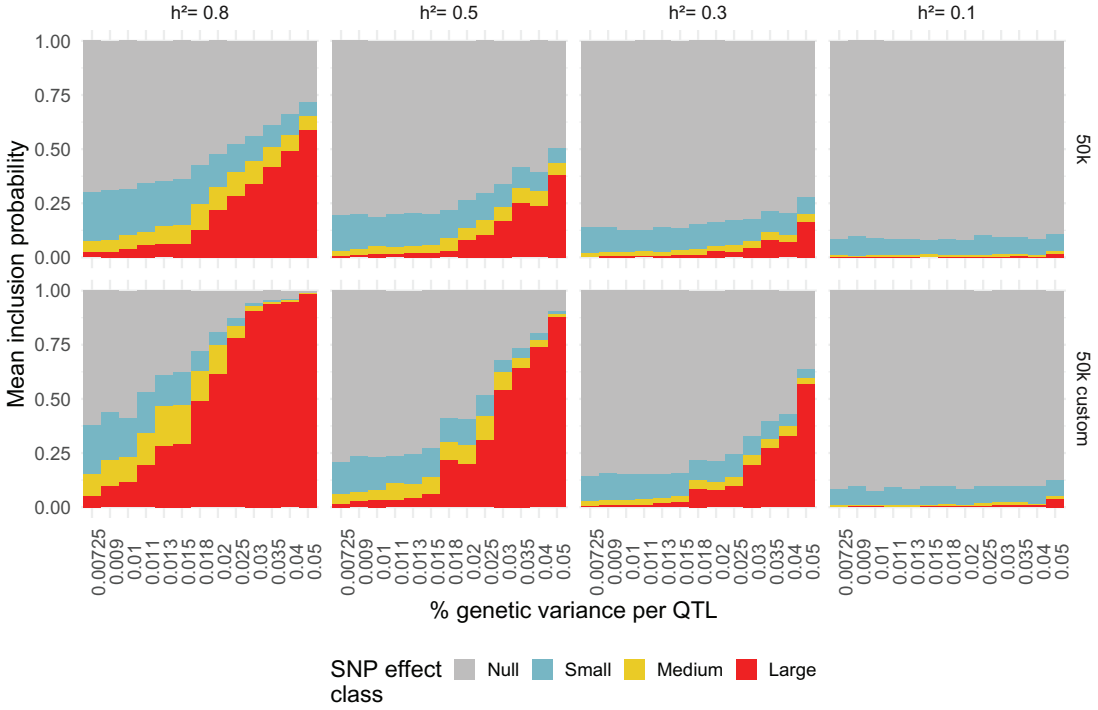


Figure 2 Neighborhood posterior inclusion probabilities across simulation settings. Panels represent combinations of heritability (columns; $h^2 = 0.8$ to 0.1) and type of data used (rows; 50k or 50k custom). Bars represent average (across 5 QTLs × 10 independent datasets) neighborhood PIP values for the four BayesR effect size classes: null (gray), small (blue), medium (yellow), and large (red).

class for both the 50k and 50k custom data. When not assigned to the null class, it was much more common to attribute QTL neighborhoods to the large effect class; the number of correctly identified QTL neighborhoods increased with the simulated effect size and/or heritability, as well as when the causal markers were included among the genotypes; what's more, these gains tend to accumulate when taken together. Correctly detecting at least one QTL window with the MAP rule required the proportion of genetic variance simulated for each QTL be $k \geq 3\%$ for $h^2 = 0.3$ using the 50k data, increasing to up to 6 QTL windows for larger simulated effects. A larger heritability of $h^2 = 0.5$ for the same data required only $k \geq 0.9\%$ to correctly identify at least one QTL window,

which increases to 22 for $k = 5\%$. However, including the causal markers in the genotype data enabled detection of QTL windows at $k \geq 1.3\%$ for $h^2 = 0.3$, with up to 30 correctly detected at $k = 5\%$. In the most favorable scenario, with $h^2 = 0.8$ and 50k custom data, QTL windows are detected for all values k, and they are exhaustively assigned to the large effect class for $k = 5\%$.

Given these results, it is not surprising that the neighborhood $MAP^{non-null}$ in Equation (8) will tend to detect more QTL windows as being nonnull. However, it is also useful to consider the behavior of this criterion while considering the LD blocks specific to each simulated QTL. In Figure 4, we visualize the neighborhood inclusion probability IP_i [defined in Equation (7)] for each of the 50 simulated QTL windows across scenarios for $h^2 = 0.5$, illustrating the proportion that are correctly included as nonnull in the model (*i.e.*, when the neighborhood inclusion probability >0.5). The $MAP^{non-null}$ appears to require a minimum LD of 55% to correctly recover QTL windows using the 50k data. Below this threshold, a large portion of QTL windows are not detected. Above this threshold, QTL window detection appears to become feasible once the simulated per-QTL percentage of genetic variance attains about $k = 2\%$. In the 50k custom data, QTL window detection does not however depend on the amount of LD, although we do note lower inclusion probabilities for QTLs in very high LD with their neighbors as compared to the 50k data. Similar to the 50k data, there is an effect size threshold at about $k = 1.8\%$ at which QTL windows are more frequently detected.

Because the same five QTLs are simulated in each independent dataset across effect size scenarios, Figure 5 also allows for their specific detection to be followed across configurations. Thus, it can be seen that some QTLs windows are not detected in any of the scenarios, while others are more easily detected, even for lower shares of the genetic variance. That said, there are occasionally discontinuities in detection observed for increasing shares of the variance (*i.e.*, a QTL window correctly identified for $k = 0.02$ but not 0.025). With the exception of $h^2 = 0.1$, which had very weak detection in all scenarios and datasets, we found similar conclusions for $h^2 = 0.3$ and 0.8, with respectively slightly smaller and larger overall inclusion probabilities than those shown in Figure 5.

Beyond the assignment of SNPs to effect classes using the neighborhood PIPs (and corresponding MAP rules), BayesR also provides posterior estimates of variability at several levels, including the additive genetic variance $\hat{\sigma}_g$, the cumulative variance for each of the three nonnull effect classes, and the variance of each SNP. Before discussing the latter (arguably the most pertinent for QTL mapping), we verify the estimation quality of the additive genetic variance. In the 50k genotype data, on average (±sd) across scenarios, $\hat{\sigma}_g$ was 9.06 (±3.32), 30.85 (±3.93), 50.12 (±4.30), and 77.36 (±4.61) for $h^2 = \{0.1, 0.3, 0.5, 0.8\}$ respectively; the corresponding true value of σ_g for each were 10, 30, 50 and 80. In the case of the 50k custom data, this same parameter was estimated to be 9.11 (±3.27), 31.01 (±3.97), 50.27 (±4.32), and 77.54 (±4.49), respectively.

Given that the total additive genetic variance appears to be well-estimated for both types of genotype type, we turn our attention to the posterior variance $\hat{V}_i / \sum_j \hat{V}_j$ of each neighborhood as defined in Equation (9). We focus in particular on the case where $h^2 = 0.5$ and proportions of genetic variance per QTL equal to $k = \{1\%, 2.5\%, 5\%\}$ (Figure 5); similar trends were observed for $h^2 = \{0.3, 0.8\}$. We note that the estimated proportion of genetic variance per SNP window are largely shrunk toward zero, clearly distinguishing those included in the model. In the 50k data, certain true QTL windows are clearly prioritized and easily identifiable. Of the 5 simulated QTLs, we observe one that can be visually identified for $k = 1\%$, and three for $k = \{2.5\%, 5\%\}$; more moderated peaks are observed for the remaining QTLs. In addition, the estimated posterior SNP window variance is about 3%, regardless of the share of variance for the simulated QTLs. When

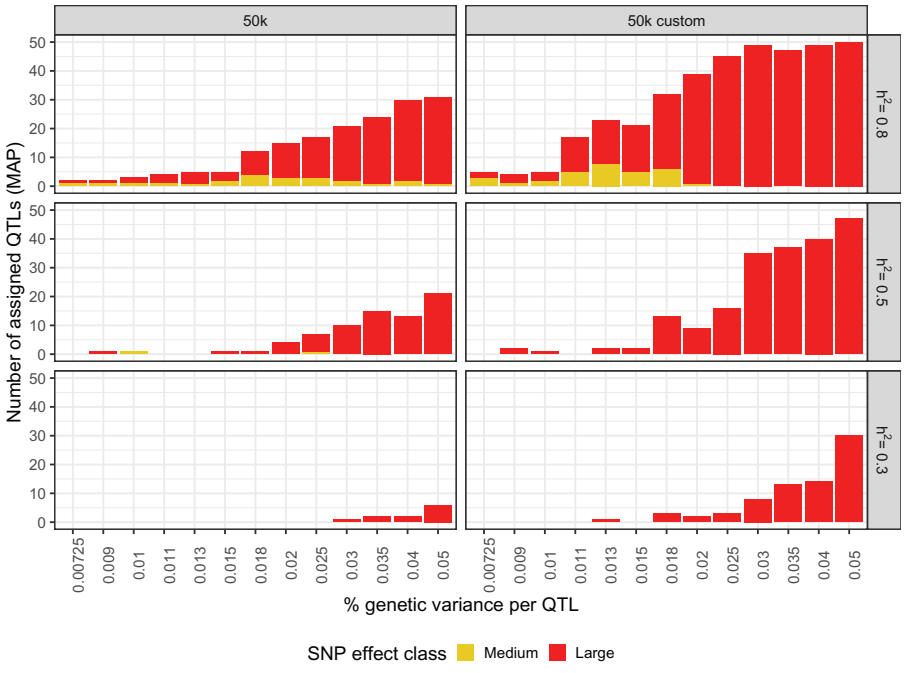


Figure 3 Neighborhood MAP rule for QTL mapping across simulation settings. Number of true QTL windows (out of 5 QTLs × 10 independent datasets simulated for each scenario, corresponding to a total of 50) correctly assigned to the medium (yellow) and large (red) effect size class using the neighborhood MAP rule. Panels represent data type (columns; 50k and 50k custom) and heritability (rows; $h^2 = 0.8$ to 0.1). The small effect class is not represented because it was empty across all simulation configurations.

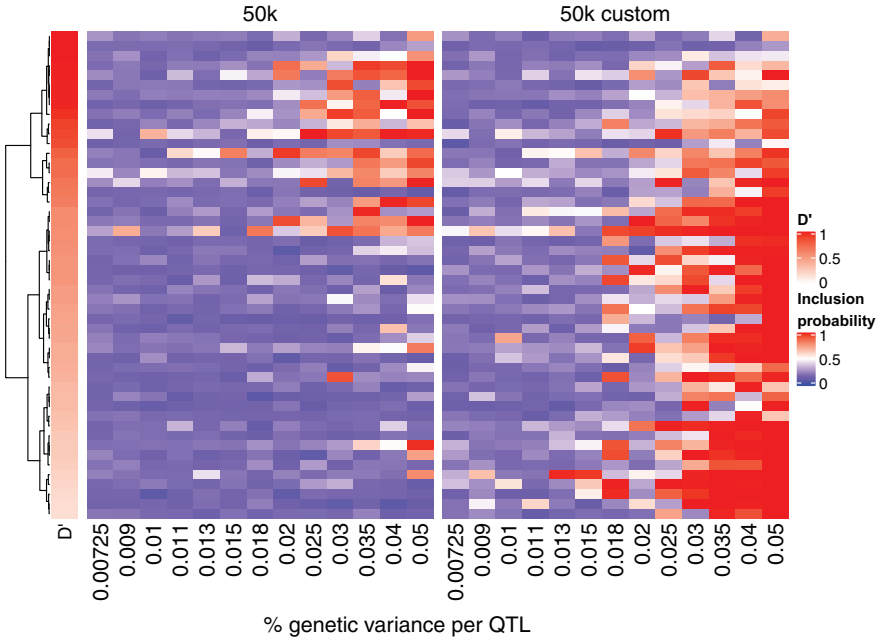


Figure 4 QTL window mapping using the neighborhood inclusion probability across different effect sizes and LD strengths for $h^2 = 0.5$. Neighborhood inclusion probabilities 1-PIP$_i^{(1)}$ for each of the 50 simulated QTLs (heatmap rows) for the 50k (left) and 50k custom (right) data across scenarios (heatmap columns). QTLs are sorted in descending order according to their LD, as measured by D' (left annotation, with deeper reds representing larger values). QTL windows that are represented by white to red cells are correctly detected using the neighborhood nonnull MAP.

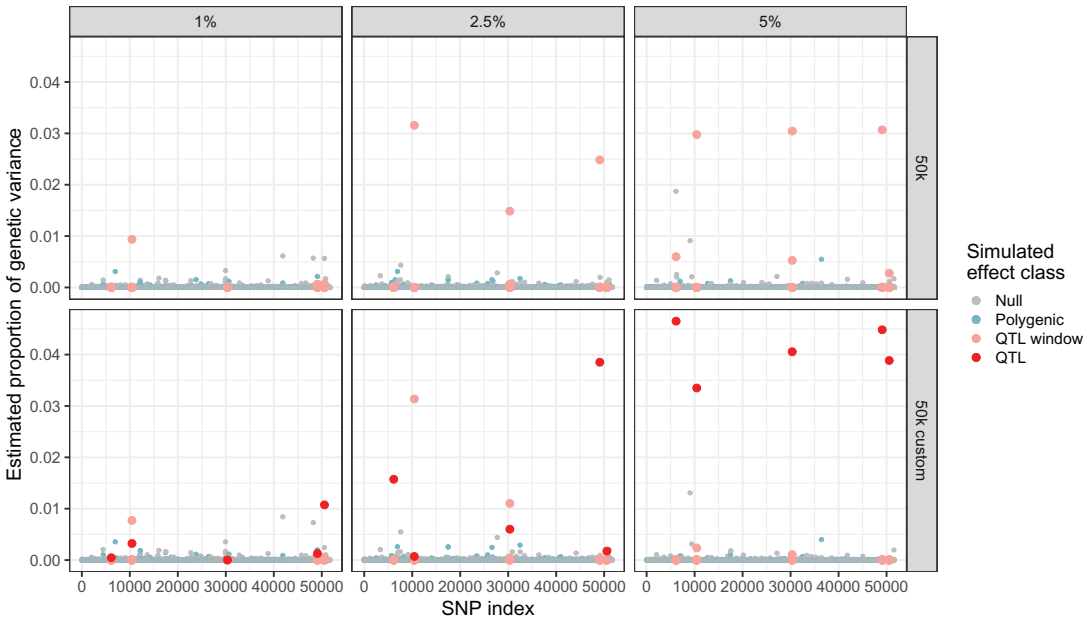


Figure 5 Genome-wide posterior estimate of the proportion of genetic variance per SNP for a single dataset with $h^2 = 0.5$. Posterior estimates of the per-SNP proportion of genetic variance across all $P = 46,178$ SNPs for one of the simulated independent datasets. Panels represent a given simulation setting for percentage of genetic variance per QTL (columns; $k = \{1\%, 2.5\%, 5\%\}$) and data type (rows; 50k *vs* 50k custom). Points represent individual SNPs, and are colored according to their true effect class (null, polygenic, in the neighborhood of a true QTL, and true QTL). The same five QTLs appear in each panel; true QTLs are only present in the 50k custom data.

$k = \{1\%, 2.5\%\}$, the prioritized QTL windows appear to have estimated variances close to the true simulated values. These estimates further improve when the 50k custom data are used, and a larger number of QTLs are clearly prioritized: we note that 2, 4, and 5 QTLs have visibly distinct peaks for $k = \{1\%, 2.5\%, 5\%\}$, respectively.

As a final criterion, we investigate the weighted cumulative inclusion probability statistic CIP$_i$ defined in Equation (10) as a way to prioritize neighborhoods where the assignment of SNPs to nonnull classes is somewhat diluted. This statistic tends to upweight regions as SNPs in the neighborhood are assigned to nonnull classes (potentially in the place of the primary QTL, which may be in tight LD with its neighbors). We expect QTL windows already detected by the neighborhood MAP to similarly have large CIP$_i$ values; however, it may facilitate the detection of those for

which a cumulative integration of nonnull SNPs across the window provides additional information.

To evaluate this point, we compared the QTL mapping performance of BayesR using the following three criteria: the neighborhood $MAP_i^{non-null}$, and the rankings of the neighborhood V_i (top 10) and neighborhood CIP_i (top 150). We chose to use $MAP_i^{non-null}$ here rather than MAP_i as it is less stringent. Across simulation scenarios and heritabilities, all QTL windows correctly detected by the nonnull neighborhood MAP were also identified by the other two criteria (Figure 6). Similarly, all QTL windows correctly detected by the posterior neighborhood variance V_i ranking were all also flagged by the CIP_i ranking. The sliding window statistic thus appears to provide the greatest detection sensitivity, while the MAP criterion is the most conservative.

For all three criteria, the number of detected QTLs increases with the simulated effect size and heritability, as well as with their inclusion among the genotypes (50 k custom data), with the exception of the lowest considered heritability, $h^2 = 0.1$. In this case, no QTL windows are detected with the $MAP^{non-null}$, and the number of QTLs identified does not greatly increase for larger QTL effect sizes. Using the CIP_i rankings, about half of the true QTL windows can be recovered using the 50k data when $h^2 = 0.8$ in the 50k chip, and similar results are possible with the 50k custom data for $h^2 = 0.5$. When the true QTLs are excluded from the genotypes, at most 46 of the 50 true QTL windows can be identified with CIP_i, even in ideal circumstances ($h^2 = 0.8$ and $k = 4\%$). However, using the 50k custom data that include these QTLs allows for universal detection when $h^2 = 0.5$ and $k = \{3\%, 4\%, 5\%\}$, or $h^2 = 0.8$ for $k \geq 2.5\%$.

An additional point of interest is to investigate the extent to which more intermediate QTL effects can be identified using these QTL mapping criteria. Using our secondary set of simulations (Supplementary Table S1), in which a set of five large and five intermediate QTLs were included, we calculated the estimated proportion of genetic variance for each QTL (Supplementary Figure S5). Unsurprisingly, the posterior variances for intermediate QTLs are smaller than those for large QTLs; however the discrepancy is considerably larger than the differential that was in fact simulated (*i.e.*, intermediate QTLs with 10% of the proportion of genetic variance assigned to large QTLs), particularly for moderate to large heritabilities. Although the estimated posterior variances of intermediate QTLs do tend to increase with larger simulated effect sizes and heritabilities, the pattern is muted compared to that of large QTLs. In addition, with the exception of the most favorable scenario ($h^2 = 0.8$, $k = 5\%$), intermediate QTLs were not assigned to the small or medium BayesR class. This suggests that the mapping criteria discussed above would be unlikely to prioritize such intermediate QTLs in all but the most highly favorable scenarios.

Evaluation of QTL mapping power vs error rate. In the previous section, our primary interest was in the detection power (*i.e.*, true positives) of BayesR for identifying QTLs. A critical related issue to contextualize these results is the corresponding error rate. We will focus on an evaluation of the neighborhood-based criteria, which generally led to more detections across scenarios compared to the $MAP^{non-null}$ (see Figure 6). Quantifying true positive discoveries is fairly straightforward here (*i.e.*, QTLs and their immediate neighborhoods are known); however, the quantification of negatives and false positives for the neighborhood-based criteria can lead to some ambiguity. In particular, the cumulative nature of the CIP, which sums weighted inclusion probabilities across a window, leads to highly dependent values for contiguous SNPs whose neighborhoods overlap a large-effect SNP. In cases where the signal of a QTL is carried by a neighboring SNP due to LD structure, any neighborhood overlapping the latter (even those not containing the true QTL) will thus tend to have inflated

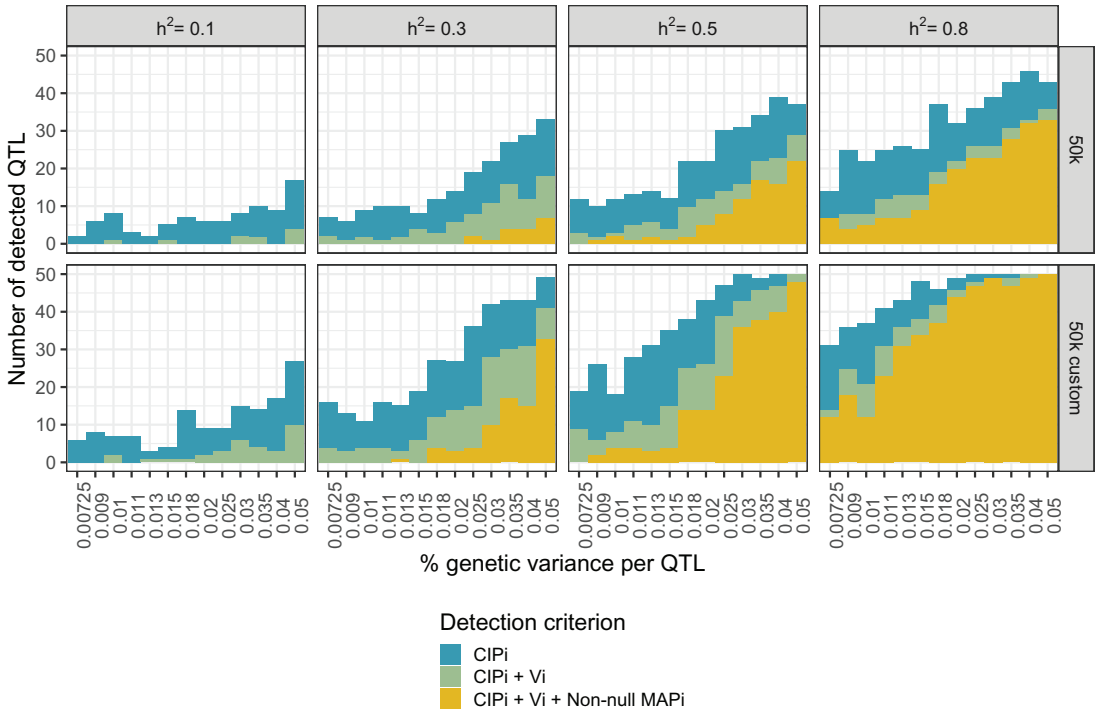


Figure 6 QTL window mapping using three different criteria across simulation settings. Number of true QTL windows (out of 5 QTLs × 10 independent datasets simulated for each scenario, corresponding to a total of 50) corrected identified using the CIP_i ranking (top 150), V_i (top 10), and $MAP_i^{non-null}$ neighborhood criteria. Panels represent data type (rows; 50k and 50k custom) and heritability (columns; $h^2 = 0.1$ to 0.8).

CIP values. Similarly, for the posterior variance V_i, LD structure can also lead to cases where neighboring SNPs have large estimated values; determining whether and how these should each be individually counted or aggregated is not clear-cut. For simplicity here, each marker located in the 15-SNP window centered on one of the 5 QTLs was individually considered to represent a positive, while all others were considered to represent negatives.

Because the CIP and estimated variance are both quantitative criteria, we sought to identify whether QTLs or their immediate neighborhoods tend to be more highly ranked than other SNPs using an Area Under the Receiver Operating Characteristic curve (AUROC). When considering genome-wide results, there is a considerable imbalance between positive and negative cases (*e.g.*, 5×15 *vs* ~40 k). As our focus was on the ranking of the top SNPs, we instead calculated AUROC values based on the 10 and 150 most highly ranked SNPs using V_i and the CIP, respectively. AUROC values were averaged across the 10 datasets for each simulation setting, and undefined values (*i.e.*, cases where no positives were included among the top SNPs) were set to be zero. As expected, AUROC values were very small in cases of low heritability or small QTL effect sizes (Figure 7). However, as heritability ($h^2 > 0.3$) and QTL effect sizes increase ($k > 2\%$), a marked increase in AUROC can be observed. For example, for sufficiently large values of heritability and QTL effect sizes, ($h^2 \geq 0.5$, $k \geq 2\%$) AUROC values attained nearly 0.80 for the 50 k custom data using either the CIP or the posterior variance. Finally, we note a slight advantage to the CIP criterion compared to the posterior variances, both for 50k and 50k custom datasets. Taken together with the previous results shown in Figure 5, these results suggest that the top rankings of SNPs provided by the CIP and posterior variance indeed tend to prioritize true positives, particularly in intermediate to favorable scenarios.

Comparison of BayesR with BayesCπ. Although BayesR has been our primary focus in this study, it is also of interest to compare its performance to that of a related widely used method, BayesCπ (Habier *et al.* 2011). BayesCπ is also based on Equation (1), but unlike BayesR, BayesCπ assumes that SNP effects β_i follow a two-component normal mixture including null and nonnull effects:

$$\beta_i \sim (1-\pi)(\beta_i = 0) + \pi N(0, \sigma^2_\beta),$$

where σ^2_β corresponds to the total genetic variance σ^2_g divided by the number of SNPs attributed to the nonnull class. In addition to the use of two rather than four effect classes, a major difference between BayesCπ and BayesR is thus that the variance of the nonnull category varies as a function of the number of SNPs included in the model.

With respect to the predictive performance of BayesCπ *vs* BayesR, as in previous studies (Zhu *et al.* 2019) we observed similar validation correlations across simulation settings, with a slight advantage for BayesR for increasing heritability or percentage of genetic variance per QTL (Supplementary Figure S2). The QTL mapping criteria previously defined for BayesR can be readily adapted to the case of BayesCπ, although considerable differences in their behavior can be observed. With the exception of cases of very large heritability with large simulated QTL effects, the per-SNP inclusion probabilities of BayesCπ tend to be much larger genome-wide than those observed for BayesR (Supplementary Figure S3). This is due to the fact that BayesCπ has a single nonnull class that tends to include a larger number of SNPs, each assigned a small proportion of the genetic variance.

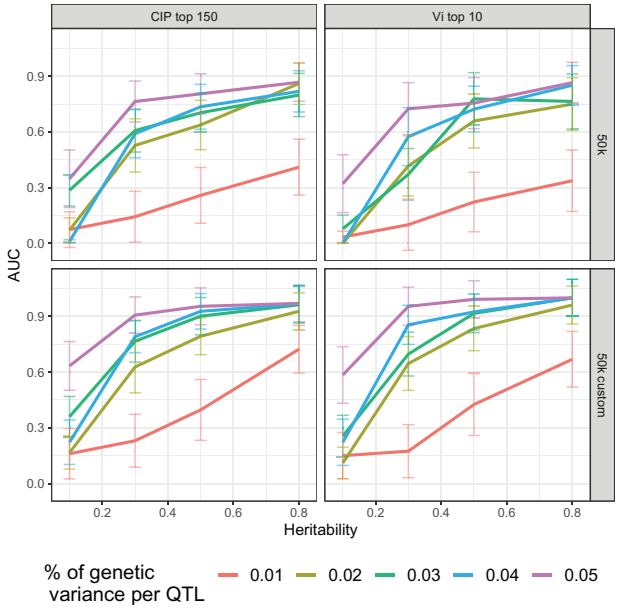


Figure 7 AUROC for the CIP and posterior variance across simulation settings. AUROC values *vs* heritability for the CIP (left column; based on top 150 values) and posterior variance (right column; based on top 10 values) by data type (rows; 50k *vs* 50k custom) for BayesR. Percentage of genetic variance per QTL is represented by different colored lines, and individual points represent averages across the 10 simulated datasets for each setting. The error bars correspond to the Monte Carlo standard errors computed across the 10 datasets for each setting.

This implies that the use of the nonnull MAP criterion generally lacks interpretability in the case of BayesCπ. In the case of the posterior variance V_i (Supplementary Figure S4), we remark that although the ranking of SNPs by BayesCπ generally mirrors that of BayesR, the estimated proportion of genetic variance is largely under-estimated. This phenomenon is another reflection of the consequences of a single, nonnull class made up of many markers with small variances. Finally, we note that the CIP criterion previously defined in Equation (10) is not immediately applicable to BayesCπ as the variance of the nonnull class is itself a random variable and not fixed as in BayesR; a similar statistic could be defined for BayesCπ, but such an adaptation is out of the scope of this study.

Discussion

In this study, we evaluated the performance of the BayesR Bayesian genomic prediction model for prediction quality and QTL mapping performance on simulated data under a variety of scenarios, including varying QTL effect sizes, heritabilities, and the use of 50k *vs* 50k custom genotype data. Simulated phenotypes were generated using SNPs from a real set of genotype data in cattle that were divided into three categories (null, polygenic SNPs, and QTLs), with variable corresponding shares of the additive genetic variance. In our study, polygenic SNPs were simulated to have the same share of genetic additive variance as the default BayesR small effect class, *i.e.*, $10^{-4} \times \sigma^2_g$. QTLs were assigned variances ranging from $7.25.10^{-3} \times \sigma^2_g$ to $5.10^{-2} \times \sigma^2_g$, constituting an interval that includes the default prior variance of the BayesR large effect class, *i.e.*, $10^{-2} \times \sigma^2_g$. These scenarios were simulated at different levels of heritability $h^2 = \{0.1, 0.3, 0.5, 0.8\}$, and we considered both genotype data that excluded (50k data) or included (50k custom data) the true simulated QTLs. As the BayesR model definition includes four

different effect size classes (null, small, medium, and large), it is of particular interest to evaluate how well the model itself adapts to the underlying genomic architecture of the data. Finally, we note that within each of the ten simulated repetitions in our study, the same set of five QTLS was selected across scenarios (heritability, QTL effect size), corresponding to a total of 50 across repetitions; this allowed for a consistent set of QTLs across scenarios for a given repetition, thus facilitating matched comparisons across settings. This is an important point, as it enabled a control of the variability due to QTL minor allele frequencies and LD patterns across settings.

The specific parameterization of BayesR (*e.g.*, number and magnitude of nonnull effect classes) can be adapted for different applications. In this study, we investigated the sensitivity of BayesR results based on the magnitude of the large effect class, and we found that the performance of BayesR (predictive power, estimations of per-SNP effects) was relatively robust. This suggests a limited benefit to modifying the priors based on prior biological knowledge. A more promising approach to integrate such prior knowledge is the related BayesRC model (MacLeod *et al.* 2016). In the BayesRC approach, SNPs are divided by the user into two or more nonoverlapping subsets, each of which represents a biologically relevant grouping with a potentially different proportion of QTLs. For each subset, the four BayesR SNP effect classes are used, with proportions modeled using an independent Dirichlet prior (*i.e.*, varying among subsets). As this flexibility can help prioritize informative SNP subsets that contain a larger proportion of QTLs, it would be of great interest to evaluate the impact of the choice of SNP subsets on QTL mapping with BayesRC, using the criteria we investigated here.

With the exception of very low heritability ($h^2 = 0.1$), validation correlation unsurprisingly increases when QTLs are included among the genotypes (*i.e.*, the 50 k custom data); this increase is particularly marked for highly heritable phenotypes as well as for QTLs with large effects. We note that the predictive power of the BayesR model varied both across simulated scenarios, as well as within a given scenario, suggesting that the specific position of simulated QTLs and polygenic SNPs appears to have an influence on the behavior of BayesR.

We presented several statistics for QTL mapping and interpretation using BayesR results, but we note that accurately assessing and quantifying the importance of a particular genomic region remains a challenge. One major obstacle is the presence of LD between SNPs. On one extreme, low LD among neighboring SNPs can impede the detection of regions if causal mutations are not directly included among genotypes, while on the other, strong LD blocks can dilute the signal among adjacent SNPs, leading to alternating assignments to nonzero effect classes (and subsequently lower estimated PIPs and variances). While the $MAP_i^{non-null}$ appears to be overly conservative for the detection of QTL neighborhoods, the V_i has the advantage of facilitating an estimation of the proportion of variability corresponding to each QTL neighborhood, given the overall estimated genetic additive variability. On the other hand, the CIP_i statistic better takes LD into account by incorporating the cumulative importance of an entire region, perhaps explaining why it can better identify QTL neighborhoods than the other criteria considered, even under nonoptimal conditions (*e.g.*, $h^2 = 0.1$).

There are several limits to our current study that should be taken into consideration. First, we note that some of our simulation scenarios could be considered to represent optimal conditions (*e.g.*, large heritabilities and QTL effect sizes) that would be rare in real applications. However, studying these extreme scenarios enables the behavior of the BayesR model to be established in ideal cases. All of our simulations made use of a constant number of individuals in both the training and validation sets, but a future study evaluating the impact of the training population sample size on QTL mapping ability, particularly for cases with low heritability (*e.g.*, $h^2 = 0.1$), could provide insight on this point. Finally, when sampling SNPs to represent QTLs in our simulations, we chose to limit the choice to those with a MAF > 0.15, thus excluding those with rare alleles. Although this allowed us to avoid edge cases that would arise with very low MAFs, making it easier to homogenize simulated datasets across different selections of QTLs, this however is an important consideration in QTL mapping.

Conclusions

BayesR is a powerful tool for simultaneously providing accurate phenotypic predictions and mapping causal regions. Our simulation results illustrate the flexibility of BayesR for different genomic architectures for all but very low heritabilities ($h^2 = 0.1$) or small QTL effects (<2% share of the additive genomic variance). Although the four effect size classes (null, small, medium, large) defined in BayesR do not themselves always reflect the true categorization of SNPs, they do offer a new approach to understanding and characterizing the genomic architecture underlying a phenotype. To this end, we presented a variety of statistical criteria that can be used to perform QTL mapping using the output of the BayesR model, including neighborhood-based nonnull maximum a posteriori rules, posterior estimated variances, and cumulative inclusion probabilities. We showed that some of the challenges in QTL mapping posed by strong LD blocks could be overcome using the latter criterion, which focuses on the assignment to nonnull effect classes of SNPs in an entire neighborhood. By ranking SNPs using this criterion, we demonstrated that QTL windows could more easily be detected, even in simulation scenarios with more challenging conditions.

Acknowledgments

The authors thank Mario Calus, Marco Bink, and Bruno Perez for helpful discussions, and Didier Boichard for providing the simulation software used to generate simulated phenotypes from a set of genotype data.

Funding

This study is part of the GENE-SWitCH project that has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement no 817998. This study also benefited from the clustering activities organized with the BovReg project, part of the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement no 815668.

Data availability

The Montbéliarde genotyping data on which simulations are based originate from a private French genomic selection program

and were funded by the users (breeding companies and breeders). They are thus proprietary data that cannot be publicly disseminated to the scientific community.

All code used to simulate and analyze the data, as well as the scripts to implement BayesR are available on GitHub (https://github.com/fmollandin/BayesR_Simulations). The repository is divided into three parts:

- *Simulations:* Fortran source code of the software used to simulate data based on real genotypes. An example of parameters and description of their use are also provided.
- *bayesR:* The modified version of BayesR (available at https://github.com/syntheke/bayesR), including recovery of the estimated per-SNP effects for each iteration, which in turn facilitates the estimation of per-SNP posterior variances. An example of the use of this software is also provided.
- *codes_R:* Partial R scripts used to analyze the BayesR model output and visualize the corresponding results. Scripts to reproduce all figures presented in the article are also included.

Conflicts of interest

None declared.

Literature cited

Abdollahi-Arpanahi R, Gianola D, Peñagaricano F. 2020. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genet Sel Evol. 52:15.

Bellot P, de los Campos G, Pérez-Enciso M. 2018. Can deep learning improve genomic prediction of complex human traits? Genetics. 210:809–819.

Brøndum RF, Su G, Janss L, Sahana G, Guldbrandtsen B, *et al.* 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. J Dairy Sci. 98:4107–4116.

Erbe M, Hayes B, Matukumalli L, Goswami S, Bowman P, *et al.* 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 95:4114–4129.

Fernando R, Toosi A, Wolc A, Garrick D, Dekkers J. 2017. Application of whole-genome prediction methods for genome-wide association studies: a bayesian approach. JABES. 22:172–193.

Habier D, Fernando RL, Kizilkaya K, Garrick DJ. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics. 12:186.

Heslot N, Jannink J-L, Sorrells ME. 2015. Perspectives for genomic selection applications and research in plants. Crop Sci. 55:1–12.

Kemper KE, Reich CM, Bowman PJ, Vander Jagt CJ, Chamberlain AJ, *et al.* 2015. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. Genet Sel Evol. 47:29.

Lewontin R. 1964. The interaction of selection and linkage. I. general considerations; heterotic models. Genetics. 49:49–67.

Liu A, Lund MS, Boichard D, Karaman E, Fritz S, *et al.* 2020. Improvement of genomic prediction by integrating additional single nucleotide polymorphisms selected from imputed whole genome sequencing data. Heredity (Edinb). 124:37–49.

MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, *et al.* 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics. 17:144.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 39:906–913.

Mardis ER. 2017. DNA sequencing technologies: 2006–2016. Nat Protoc. 12:213–218.

Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157:1819–1829.

Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, *et al.* 2015. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS Genet. 11: e1004969.

Pérez-Enciso M, Rincón JC, Legarra A. 2015. Sequence-vs. chip-assisted genomic selection: accurate biological information is advised. Genet Sel Evol. 47:14.

Sanchez MP, Govignon-Gion A, Ferrand M, Gelé M, Pourchet D, *et al.* 2016. Whole-genome scan to detect quantitative trait loci associated with milk protein composition in 3 french dairy cattle breeds. J Dairy Sci. 99:8203–8215.

Uemoto Y, Sasaki S, Kojima T, Sugimoto Y, Watanabe T. 2015. Impact of QTL minor allele frequency on genomic evaluation using real genotype data and simulated phenotypes in Japanese black cattle. BMC Genet. 16:134.

Van den Berg I, Boichard D, Lund MS. 2016. Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. Genet Sel Evol. 48:83.

Voss-Fels KP, Cooper M, Hayes BJ. 2019. Accelerating crop genetic gains with genomic selection. Theor Appl Genet. 132:669–686.

Wray NR, Kemper KE, Hayes BJ, Goddard ME, Visscher PM. 2019. Complex trait prediction from genome data: contrasting EBV in livestock to PRS in humans: genomic prediction. Genetics. 211: 1131–1141.

Zhu B, Guo P, Wang Z, Zhang W, Chen Y, *et al.* 2019. Accuracies of genomic prediction for twenty economically important traits in Chinese simmental beef cattle. Anim Genet. 50:634–643.

Communicating editor: G. de los Campos

2.7 Matériel supplémentaire

Supplementary Materials: An evaluation of the interpretability and predictive performance of the BayesR model for genomic prediction

Fanny Mollandin*, Andrea Rau*,†, Pascal Croiseau*

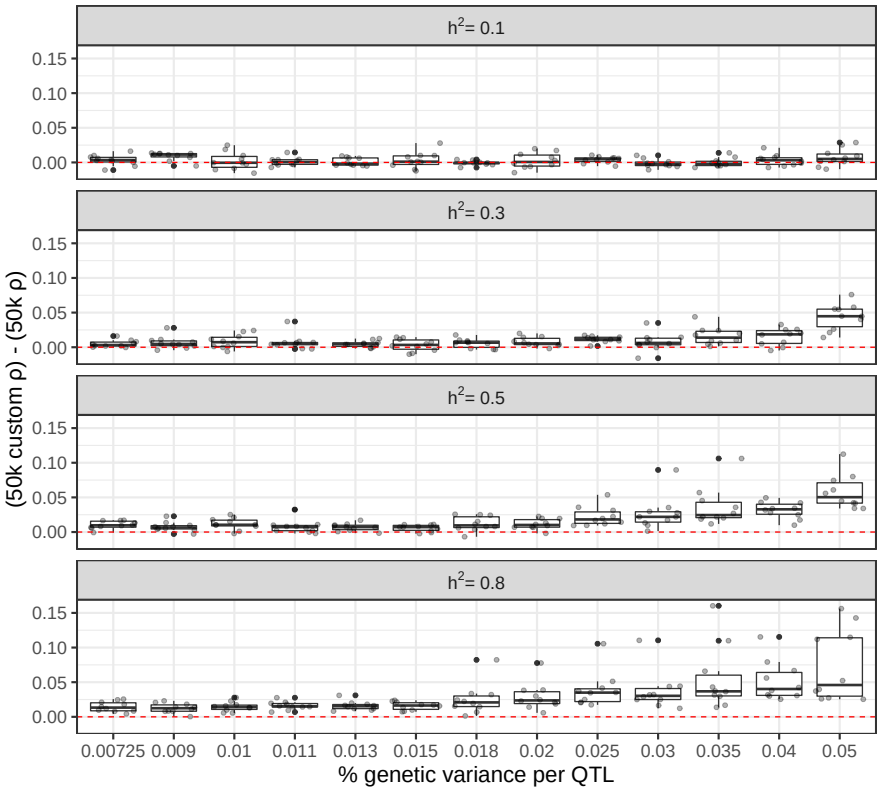
* Université Paris-Saclay, INRAE, AgroParisTech, GABI
† BioEcoAgro Joint Research Unit, INRAE, Université de Liège, Université de Lille, Université de Picardie Jules Verne

1 Supplementary Tables

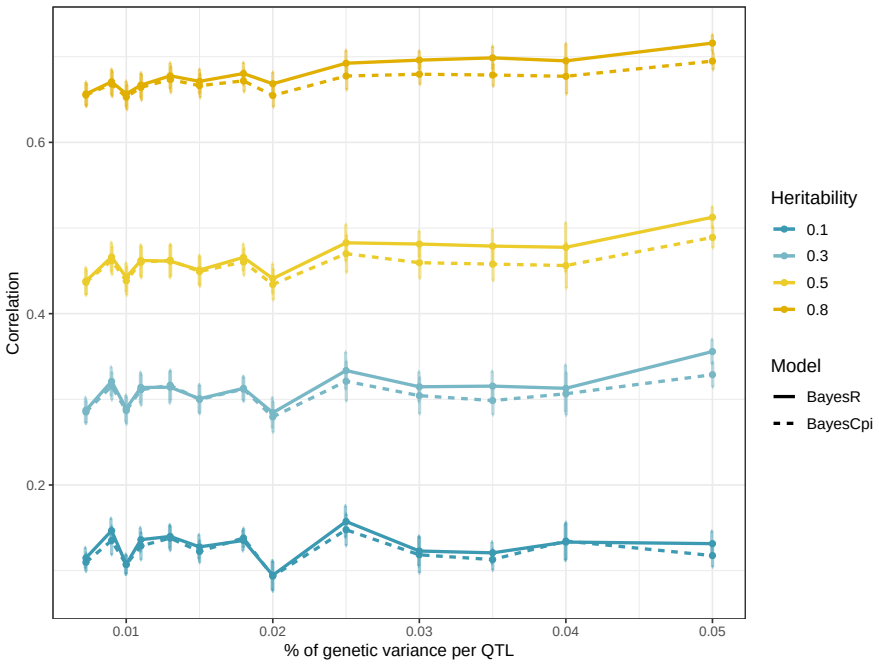
Number of large QTLs	5	5	5
Number of medium QTLs	5	5	5
Number of polygenic SNPs	9450	8625	7250
Per-large QTL % of σ_g^2	1	2.5	5
Per-medium QTL % of σ_g^2	0.1	0.25	0.5
Per-polygenic SNP % of σ_g^2	0.01	0.01	0.01

Table 1: Simulation settings for each of the 3 QTL effect-size alternative scenarios, including medium QTLs, considered for each given level of heritability, $h^2 = \{0.1, 0.3, 0.5, 0.8\}$. The number of simulated QTLs, number of polygenic SNPs, percentage of genetic variance attributed to each QTL, and percentage of genetic variance attributed to each polygenic SNP are provided. Summing the percentage of genetic variance explained by the total number of large and medium QTLs and polygenic SNPs yields 100%.

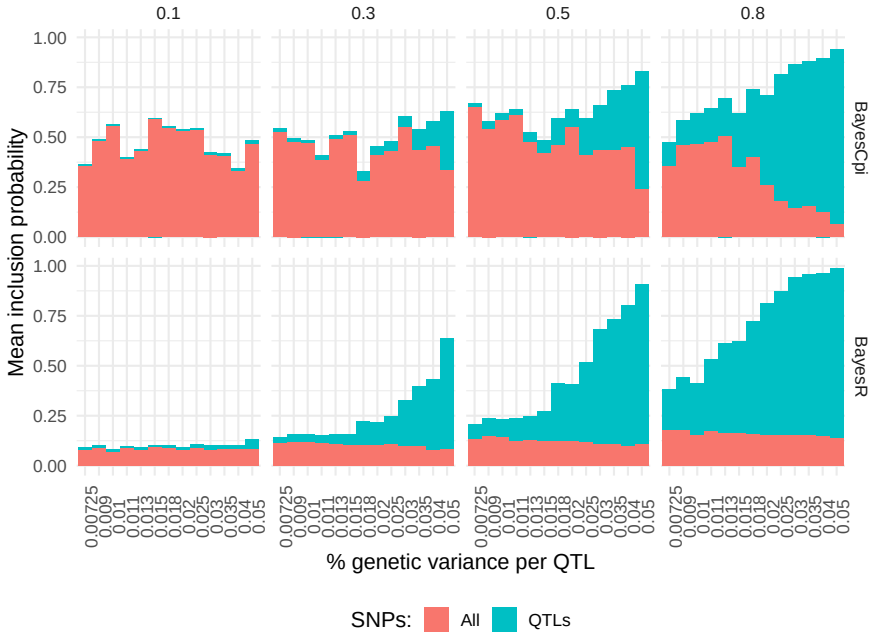
2 Supplementary Figures



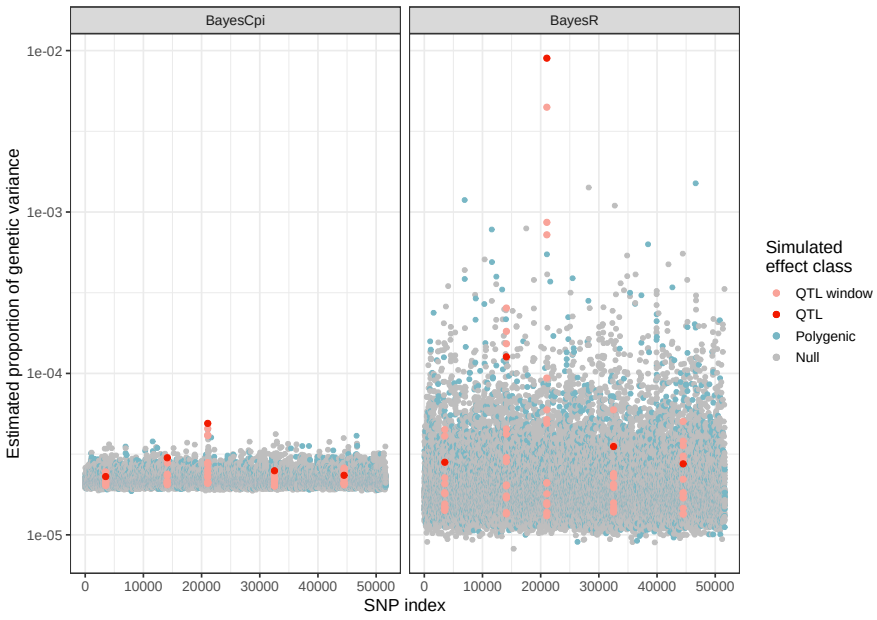
Supplementary Figure 1: **Difference in BayesR predictive performance for the 50k versus 50k custom genotypes across simulation settings.** Each panel from top to bottom represents a given heritability ($h^2 = 0.1$ to 0.8), and boxplots represent the distribution of differences in validation correlation between the 50k and 50k custom datasets for each independent dataset (i.e., for which the same 5 QTLs are simulated). The red dotted line indicates a baseline of 0.



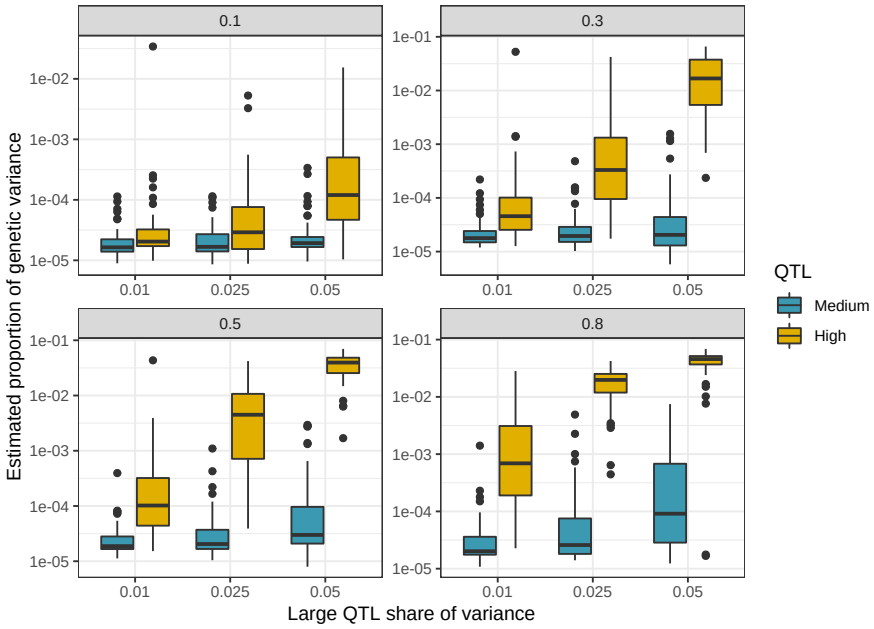
Supplementary Figure 2: **BayesR and BayesCπ predictive performance across simulation settings for the 50k custom datasets.** For each setting (h^2 and percentage of genetic variance assigned to each QTL), points represent mean validation correlations across 10 independent datasets. Heritability values are represented by dark gold to dark blue (h^2 = 0.8 to 0.1), and solid and dotted lines represent results for the BayesR and BayesCπ custom datasets, respectively. The error bars correspond to the Monte Carlo standard errors computed across the 10 datasets for each setting.



Supplementary Figure 3: **BayesCπ and BayesR model inclusion probabilities across simulation settings for the 50k custom datasets.** For each setting (h^2 and percentage of genetic variance assigned to each QTL), bars represent the average posterior inclusion probability across all SNPs (red) or QTLs (blue) for the BayesCπ (top row) or BayesR (bottom row) models.



Supplementary Figure 4: **BayesCπ and BayesR model estimated proportion of variance for a simulated dataset with 50k custom genotypes.** Results for BayesCπ (left) and BayesR (right) are shown for one simulated 50k custom dataset with $h^2 = 0.5$ and percentage of genetic variance per QTL = 1%. Each point represents a SNP, with colors indicating true nulls (grey), polygenic effects (light blue), or SNPs simulated to be true QTLs (red) or within a 15-SNP window around true QTLs (pink).



Supplementary Figure 5: **Proportion of genetic variance estimated by BayesR for scenarios including intermediate QTL with 50k custom genotypes.** Each panel represents a given heritability ($h^2 = 0.1$ to 0.8), and boxplots represent the estimated posterior variance for intermediate (blue) and large (orange) QTLs.

Chapter 3

Accounting for overlapping annotations in genomic prediction models of complex traits

3.1 Résumé

L'étude de simulations sur BayesR nous a montré sa flexibilité et ses capacités de prédiction face à diverses architectures génétiques. Dans la suite de cette thèse, nous continuerons de l'utiliser en tant que modèle référent, qui n'exploite pas d'annotations fonctionnelles. BayesRC se pose en suite logique pour une première intégration d'annotations fonctionnelles. Si son approche par partitionnement des SNPs est une modélisation innovante et prometteuse de l'utilisation d'annotations fonctionnelles, nous nous sommes rapidement confrontée à la limite d'une seule annotation par SNP. Avec une augmentation du nombre d'informations fonctionnelles, la probabilité d'avoir des SNPs multi-annotés augmente, ce qui implique dans le cas de BayesRC de devoir faire un choix, perdant une partie de l'information exploitable.

Nous avons identifié deux hypothèses sur l'interprétation à avoir d'un SNP multi-annoté:

1. Un SNP multi-annoté peut circuler entre ses différentes annotations en fonction de sa proximité avec leur distribution
2. Un SNP multi-annoté devrait être priorisé, et avoir plus de chances d'être inclus dans le modèle

Pour répondre à la première hypothèse, nous avons proposé le modèle BayesRCπ, qui définit la distribution *a priori* des effets des SNPs comme une mélange de mélanges gaussiens. Pour la seconde, nous avons proposé le modèle BayesRC+, qui définit la distribution *a priori* des effets des SNPs comme un cumul de mélanges

gaussiens. Ces deux modèles ont été évalués sur des données simulées, ainsi que des données réelles porcines. Similairement au chapitre précédent, nous avons simulé des phénotypes à partir de génotypes existants, en considérant plusieurs architectures génétiques et héritabilités. Nous avons de plus simulés différents scénarios d'annotations, en combinant des listes d'annotations d'enrichissements variés, pour évaluer la flexibilité des modèles face à des annotations peu informatives, et avec une multiplication des chevauchements. L'amélioration de la qualité de prédiction dépend de la qualité des annotations intégrées, et reste modeste (au maximum +2 points de corrélations en moyenne). En revanche, les modèles intégrant les annotations, et en particulier BayesRC+, permettent de prioriser les marqueurs annotés, et ainsi estimer plus précisément l'effet des QTLs simulés. BayesRCπ offre pour sa part des possibilités d'interprétation des annotations et de leur utilisation dans le modèle.

Pour les données porcines issues du projet PigHeat (Gourdine et al., 2019), constituées de génotypes et phénotypes (2 caractères de production, BFT *backfat thickness* et ADG *average daily gain*), nous avons constitués des listes d'annotations à partir de la base de données publique pigQTLdb, regroupant des listes de QTLs liés à des caractères de production. Plusieurs fenêtres d'annotations ont de plus été définies (*regular*, *fuzzy* et *hard*). Nous avons fait le choix de mettre toutes les annotations (11 au total) dans les modèles, indifféremment de leur cohérence au caractère de production prédit. Si ADG a pu voir sa qualité de prédiction augmenter en utilisant les annotations pigQTLdb, les résultats de prédiction ont été plus mitigés pour BFT. Nous avons aussi pu observer des différences dans la qualité de prédiction entre les différentes fenêtres utilisées pour construire les annotations.

Dans l'ensemble, bien que modestes, BayesRCπ et BayesRC+ ont réalisé des gains de prédiction, que ce soit pour le cadre de simulation ou réel sous couvert d'annotations adaptées. Ils permettent aussi la priorisation des marqueurs multi-annotés, qui peuvent apporter de nouvelles perspectives pour la compréhension de caractères complexes. Que ce soit pour la prédiction, ou la représentation de l'architecture génétique des caractères complexes, la construction des annotations semblent avoir un impact important.

Ces résultats, ont été publiés par la revue BMC Bioinformatics en 2022, sous le DOI https://doi.org/10.1186/s12859-022-04914-5.

Mollandin *et al. BMC Bioinformatics* (2022) 23:365
https://doi.org/10.1186/s12859-022-04914-5

BMC Bioinformatics

RESEARCH Open Access

Accounting for overlapping annotations in genomic prediction models of complex traits

Fanny Mollandin[1*], Hélène Gilbert[2], Pascal Croiseau[1] and Andrea Rau[1,3]

*Correspondence:
fanny.mollandin@inrae.fr

[1] INRAE, AgroParisTech, GABI, Université Paris-Saclay, Allée de Vilvert, 78350 Jouy-en-Josas, France
[2] GenPhySE, INRAE, ENVT, Université de Toulouse, 31320 Castanet Tolosan, France
[3] BioEcoAgro Joint Research Unit, INRAE, Université de Liège, Université de Lille, Université de Picardie Jules Verne, 50136 Estrée-Mons, France

Abstract

Background: It is now widespread in livestock and plant breeding to use genotyping data to predict phenotypes with genomic prediction models. In parallel, genomic annotations related to a variety of traits are increasing in number and granularity, providing valuable insight into potentially important positions in the genome. The BayesRC model integrates this prior biological information by factorizing the genome according to disjoint annotation categories, in some cases enabling improved prediction of heritable traits. However, BayesRC is not adapted to cases where markers may have multiple annotations.

Results: We propose two novel Bayesian approaches to account for multi-annotated markers through a cumulative (BayesRC+) or preferential (BayesRCπ) model of the contribution of multiple annotation categories. We illustrate their performance on simulated data with various genetic architectures and types of annotations. We also explore their use on data from a backcross population of growing pigs in conjunction with annotations constructed using the PigQTLdb. In both simulated and real data, we observed a modest improvement in prediction quality with our models when used with informative annotations. In addition, our results show that BayesRC+ successfully prioritizes multi-annotated markers according to their posterior variance, while BayesRCπ provides a useful interpretation of informative annotations for multi-annotated markers. Finally, we explore several strategies for constructing annotations from a public database, highlighting the importance of careful consideration of this step.

Conclusion: When used with annotations that are relevant to the trait under study, BayesRCπ and BayesRC+ allow for improved prediction and prioritization of multi-annotated markers, and can provide useful biological insight into the genetic architecture of traits.

Keywords: Genomic prediction, Functional annotation, Bayesian models

Background

It is now widespread in plant and animal breeding [1] and agriculture [2, 3] to predict phenotypes, i.e., observable traits in an individual, from genotypes. In recent years, improvements in sequencing technologies and their decreasing cost [4], combined with increased computational and storage capacities, have further accelerated the use of genomic prediction. Since the early 2000's, a variety of genomic prediction models

have been proposed, including the genomic best linear unbiased predictor (GBLUP; [5]) and the family of methods constituting the "Bayesian alphabet" [6], such as BayesA [1], BayesB [1], BayesCπ [7] and BayesR [8]. These models rely on different assumptions on the distribution of single nucleotide polymorphism (SNP) effects, striking a balance between flexibility and computational efficiency. BayesR in particular has been shown to be a powerful and flexible model, generally yielding high quality predictions while simultaneously facilitating quantitative trait loci (QTL) mapping, though some marker effects remain underestimated, especially for traits with low heritabilities [9, 10].

One interesting strategy for improvement to guide genomic prediction models is the use of prior biological information [11]. An increasing amount of such prior information is available, ranging from publicly available trait mapping information to functional or structural annotations of the genome [12]. By appropriately including these heterogeneous and complex data, it is hoped that causal mutations could be more readily identified and prioritized, thus potentially improving model predictions and interpretability. Such prior biological information can be collected on the same individuals used for genomic prediction, or on an independent population (e.g., from publicly available databases, such as FAANG "Functional Annotation of Animal Genome" [12]). Several methods have been introduced for the former case, such as GTBLUP [13] and GTiBLUP [14], although such fully coupled datasets remain rare. However, prior biological information from independent sources are much more widely accessible, and such information can be used to assign variants to annotation categories.

To make use of such information, the BayesRC model [15] factorizes the genome according to a prior categorization of markers. Each annotation category is independently modeled according to the mixture prior defined by BayesR, where SNPs may have a null, small, medium or large effect. In practice, BayesRC is generally used for a small number of disjoint annotation categories, where each marker is assigned to a single category. Considering a greater number of potentially overlapping annotations would likely lead to the presence of multi-annotated markers. To use BayesRC in this case would necessitate the choice of a single annotation for each multi-annotated marker, which may lead to an undesired loss of information. There thus remains a need to define robust models that can handle multi-annotated SNPs.

Depending on the context, choice of annotations, and desired interpretation of multi-annotated SNPs, in this work we propose two different models. First, if multiple annotation categories are thought to represent ambiguity in the appropriate annotation assignment, we directly model the probability of category assignment for multi-annotated SNPs. For this, we propose the BayesRCπ model with a mixture of mixtures prior distribution on SNP effects, thus allowing multi-annotated SNPs to be assigned a posteriori to the most informative annotation. Alternatively, if the number of annotations for a given marker is assumed to be informative (e.g., a larger number of annotations implies a stronger belief that a marker may be causal), we may wish to systematically assign greater weight to multi-annotated markers in the model. To this end, we propose the BayesRC+ model, with a cumulative mixture prior distribution on SNP effects.

As the proposed BayesRCπ and BayesRC+ models incorporate biological information in different ways, their performance is likely to be highly dependent on the underlying genetic architecture of the studied traits, the construction of annotation categories, and

the biological relevance of the prior information. However, given the potential difficulty in defining annotation categories in practice, both models must be robust to the inclusion of noisy or irrelevant annotations. To evaluate our models, we simulated data with various genetic architectures and annotation configurations. In the same perspective, we applied both methods to real pig data in conjunction with annotations constructed from a public database, again varying the way annotation categories were constructed. This study thus highlights the interest of BayesRCπ and BayesRC+ in different scenarios, while proposing a preliminary, non-exhaustive framework for their use in practice.

Methods

Bayesian genomic prediction without annotations

The general statistical model for genomic prediction can be defined as

$$\begin{aligned} &\mathbf{y} = \mu \mathbb{1}_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \\ &\mathbf{e} \sim N(0, \mathbf{I}_n \sigma_e^2) \end{aligned} \tag{1}$$

where $\mathbf{y}$ is the vector of corrected phenotypes, μ the intercept, $\boldsymbol{\beta}$ the vector of the SNP effects and $\mathbf{e}$ the vector of residuals. We assume that $\mathbf{e}$ follows a multivariate normal distribution with mean 0 and variance covariance matrix $\mathbf{I}_n\sigma_e^2$. $\mathbf{X}$ is the marker matrix, centered and scaled such that $X_{ij} = (w_{ij} - 2f_j)/\sqrt{2f_j(1-f_j)}$, with $w_{ij} \in \{0, 1, 2\}$ the number of copies of the alternative allele of marker j in individual i and f_j the frequency of the alternative allele of marker j in the full population. We note σ_g^2 the total additive variance, i.e., the cumulative variance of all SNP effects.

By defining different prior distributions on the SNP effect vector $\boldsymbol{\beta}$, models in the Bayesian alphabet seek to overcome model overparametrization in various ways. BayesR [8] assumes that SNP effects β_i follow a four-component normal mixture:

$$f(\beta_i) = \sum_{k=1}^{4} \pi_k f_k(\cdot|\theta_k) \tag{2}$$

$$\text{such that} f_k = \begin{cases} \delta(0), & \text{if } k = 1 \\ \phi(\cdot|0, \theta_k) & \text{otherwise} \end{cases} \tag{3}$$

where $\boldsymbol{\theta} = (\theta_2, \theta_3, \theta_4) = (0.0001\sigma_g^2, 0.001\sigma_g^2, 0.01\sigma_g^2)$, $\sum_{k=1}^{4} \pi_k = 1$, $\delta(0)$ represents a point mass at 0, and ϕ is the centered Gaussian probability density function.

Practically, the BayesR model implies that markers are assigned to one of four different effect size classes: null, small, medium or large, corresponding respectively to 0%, 0.01%, 0.1% and 1% of the total additive genetic variance σ_g^2. The mixing proportions $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)$ are assumed to follow a Dirichlet prior, corresponding to the posterior $f(\boldsymbol{\pi}|.) \sim \text{Dirichlet}(\boldsymbol{\alpha} + \boldsymbol{\gamma})$, with $\boldsymbol{\alpha}$ representing a vector of pseudocounts and $\boldsymbol{\gamma}$ the cardinality of each component. In this work, we used a flat Dirichlet pior distribution, with $\boldsymbol{\alpha} = (1, 1, 1, 1)$, and σ_g^2 is assumed to be a random variable following an Inv$-\chi^2$ distribution.

Formalizing annotation categories

There are several potential ways that biological annotations could be formalized for use as prior information. Here, we assume that markers are categorized in a binary fashion for each category (annotated or not), where SNPs with no known annotation are aggregated together under an "other" category. We note $C_i \subseteq \{c_1, c_2, ..., c_m\}$ the set of annotations corresponding to SNP i. Depending on the case, marker i can have a single annotation (i.e., $|C_i| = 1$) or be multi-annotated (i.e. $|C_i| \geq 1$).

Bayesian genomic prediction with disjoint annotations

BayesRC [15] extends the BayesR model prior in Eq. (3) by dividing the genome into disjoint annotations such that $|C_i| = 1$, each with a potentially different proportion of small, medium, and large QTLs. BayesRC thus exploits the same four SNP effect size classes as BayesR, but the mixing proportions $\boldsymbol{\pi}_c$ for each annotation c are estimated separately:

$$f(\beta_i|\mathbf{C}_i = c) = \sum_{k=1}^{4} \pi_{k,c} f_k(\cdot|\theta_k),$$
$$\text{such that} f_k = \begin{cases} \delta(0), & \text{if } k = 1 \\ \phi(\cdot|0, \theta_k) & \text{otherwise} \end{cases}$$

where $\boldsymbol{\theta}$, $\delta(0)$, and ϕ are defined as before, and $\sum_{k=1}^{4} \pi_{k,c} = 1$ for all $c \in \{c_1, c_2, ..., c_m\}$. The mixing proportions $\boldsymbol{\pi}_c$ are assumed to follow a Dirichlet prior, yielding the posterior $f(\boldsymbol{\pi}_c|.) \sim \text{Dirichlet}(\boldsymbol{\alpha} + \boldsymbol{\gamma}_c)$, with $\boldsymbol{\alpha}$ representing a vector of pseudocounts and $\boldsymbol{\gamma}_c$ the cardinality of each component in annotation c. As for BayesR, we used a flat Dirichlet distribution, with $\boldsymbol{\alpha} = (1, 1, 1, 1)$, for the mixing proportion priors. In order to limit the impact this prior can have on the posterior, MacLeod et al. [15] recommend using relatively common annotations (i.e., including more than about 1000 markers).

Bayesian genomic prediction with overlapping annotations

By increasing the number and variety of (potentially redundant) annotations, it becomes increasingly likely to have multi-annotated markers where $|C_i| \geq 1$ for some i. By definition, BayesRC cannot directly account for such overlapping annotations, limiting the full use of available information. To address this, we propose two novel methods for exploiting overlapping annotations in different contexts: BayesRCπ and BayesRC+.

BayesRCπ

In cases where multiple annotation categories include potential ambiguity for multi-annotated SNPs, it may be of interest to model the probability of category assignment. To this end, we propose the BayesRCπ model to allow multi-annotated markers to preferentially associate with annotations, according to their coherence with the respective SNP effect distributions. Specifically, we define a mixture of mixtures prior distribution for SNP effects:

$$f(\beta_i|\mathbf{C}_i) = \sum_{c \in \mathbf{C}_i} p_{i,c} \sum_{k=1}^{4} \pi_{k,c} f_k(\cdot|\theta_k)$$
$$\text{such that} f_k = \begin{cases} \delta(0), & \text{if } k = 1 \\ \phi(\cdot|0, \theta_k) & \text{otherwise} \end{cases}$$

where $\boldsymbol{\theta}$, $\delta(0)$, and ϕ are defined as before and $\sum_{k=1}^{4} \pi_{k,c} = 1$ for all $c \in \{c_1, c_2, ..., c_m\}$. We have thus introduced the mixing parameter $\mathbf{p}_i \in]0,1]^{|C_i|}$ for SNP i in its set of annotations $\mathbf{C}_i$, such that $\sum_{c \in \mathbf{C}_i} p_{i,c} = 1$ for all i. Once again, the mixing proportions $\boldsymbol{\pi}_c$ are assumed to follow a Dirichlet prior, giving the posterior $f(\boldsymbol{\pi}_c|.) \sim \text{Dirichlet}(\boldsymbol{\alpha} + \boldsymbol{\gamma}_c)$, with $\boldsymbol{\alpha} = (1,1,1,1)$. The mixing proportions $\mathbf{p}_i$ are assumed to follow a Dirichlet prior, with size depending on the cardinality of the annotation set of each SNP i.

BayesRC+

An alternative way of interpreting a multi-annotated marker is to assume that a greater number annotations implies that more weight should be attributed to the marker in the model. In this spirit, we propose the BayesRC+ model to assign an additive impact of multiple annotation categories on estimated SNP effects. Multi-annotated variants will thus tend to have a greater chance to be included as non-null in the model, and as such a larger estimated effect. Specifically, we define a cumulative mixture prior distribution for the effect of SNP i:

$$f(\beta_i|\mathbf{C}_i) = \sum_{c \in \mathbf{C}_i} \sum_{k=1}^{4} \pi_{k,c} f_k(\cdot|\theta_k)$$
$$\text{such that} f_k = \begin{cases} \delta(0), & \text{if } k = 1 \\ \phi(\cdot|0, \theta_k) & \text{otherwise} \end{cases}$$

where $\boldsymbol{\theta}$, $\delta(0)$ and ϕ are defined as before and $\sum_{k=1}^{4} \pi_{k,c} = 1$ for all $c \in \{c_1, c_2, ..., c_m\}$. Prior and posterior distributions for the mixing proportions $\boldsymbol{\pi}_c$ are as described above in the BayesRC and BayesRCπ models.

Gibbs sampling

As the full posterior distributions for all models described above are intractable, model parameters are estimated with a Gibbs sampler, using the posterior mean of their estimations. Algorithm implementation details for BayesR were previously described by Kemper et al. [16] and Moser et al. [9] and were used as a base to implement BayesRC, BayesRCπ and BayesRC+. Broad steps of the algorithms for each are shown in pseudocode Additional file 1: Algorithms 1–4.

Concretely for BayesRCπ, within a given iteration of the Gibbs sampler SNPs are assigned to the annotation category with probability proportional to its conditional likelihood given the current estimates of other model parameters. Note that this step is analogous to that in the standard BayesR algorithm of assigning SNPs to one of the four effect classes, based on a conditional likelihood calculation given the current estimates of model parameters. For BayesRC+, at each iteration of the Gibbs sampler the conditional effect of a given SNP is estimated for each of its associated annotation categories in turn, and its total effect is subsequently calculated as the sum over all of its per-annotation effects.

For all models, we ran the Gibbs sampler algorithm for 50,000 iterations, discarding 20,000 for burning, and using a thinning rate of 10.

BayesRCO package on Github

We propose the *BayesRCO* (BayesRC for Overlapping annotations) software, which implements five different Bayesian genomic prediction models, including three state-of-the art approaches (BayesCπ [7], BayesR, and BayesRC) and our two novel algorithms, BayesRCπ and BayesRC+. The implementation of our two new models builds on that of the *bayesR* software found at https://github.com/syntheke/bayesR [9]. Since BayesR can be seen as a special case of BayesRC with a single annotation category, and BayesRC as a special case of BayesRCπ and BayesRC+ using non-overlapping annotations, the *BayesRCO* algorithm is divided into three independent modules: BayesCπ, BayesRCπ and BayesRC+.

Metrics for evaluation

All the metrics defined here are used to evaluate the models on the simulated and real data defined in the next section.

Prediction accuracy

Prediction accuracy for all models was quantified using the Pearson correlation between the true ($\mathbf{y}$) and estimated ($\hat{\mathbf{y}}$) phenotypic values in the validation set.

Posterior variance

The posterior variance of SNP i can be estimated as:

$$\widehat{V}_i = \widehat{\beta_i^2} Var(X_{.i}),$$

where $X_{.i}$ represents the ith column of the centered and scaled genotype design matrix and $\widehat{\beta_i^2}$ corresponds to the posterior mean of β_i^2, estimated by $\widehat{\beta_i^2} = \frac{1}{N}\sum_{\ell=1}^{N} \beta_i^{(\ell)2}$, where N is the number of iterations and $\beta_i^{(\ell)2}$ the value of β_i^2 at iteration ℓ. As the SNP effects are computed on the scaled and centered genotype design matrix X, the per-SNP posterior variance can be estimated using $\widehat{V}_i = \widehat{\beta_i^2}$.

Assignment of annotation categories using BayesRCπ

The specificity of BayesRCπ is that it models the assignment of multi-annotated markers to different annotation categories. To quantify these assignments, we introduce the posterior annotation inclusion probability (PAIP), representing the frequency (across Gibbs sampler iterations) of assignment for each multi-annotated marker to each of its annotations. We note $\text{PAIP}_i = \{\text{PAIP}_{i,c_1}, \ldots, \text{PAIP}_{i,c_m}\}$ the PAIP of marker i such that, for all i, $\text{PAIP}_{i,c} \in]0, 1]$ for all $c \in \mathbf{C}_i$, and $\sum_{c\in\mathbf{C}_i} \text{PAIP}_{i,c} = 1$ for all i.

Simulation framework

We next sought to simulate phenotypes associated with genomic data and associated annotations (as described below) to evaluate our models. For this purpose, we used real Illumina Bovine SNP50 BeadChip genotyping data from $n = 2605$ Montbéliarde bulls. Using these data as a base for our simulations has the advantage of including realistic population and linkage disequilibrium (LD) structures in our simulations.

We excluded SNPs with a minor allele frequency (MAF) less than 0.01, leaving a total of $p = 46{,}178$ SNPs. We divided individuals into learning and validation sets, respectively consisting of 80% of the oldest (2083 bulls) and 20% of the youngest bulls (522 bulls).

Phenotype simulation

To simulate phenotypes $\mathbf{y}$ for the $n = 2605$ bulls, we used the linear model in Equation (1), with parameters set as follows. For each simulated dataset, we randomly sampled a set of SNPs among those with a MAF ≥ 0.15; by focusing on frequent variants, we sought to reduce the impact of extreme MAFs on genomic prediction [17]. For selected SNPs, the corresponding effect β_i for selected SNP i was set as follows:

$$\beta_i = \frac{1}{2} u_i \sqrt{\frac{k\sigma_g^2}{2\text{MAF}_i(1-\text{MAF}_i)}},$$

where $k \in \{k_{\text{small}}, k_{\text{medium}}, k_{\text{large}}\}$ corresponds to the proportion of the total additive variance σ_g^2 for a given effect size class (described below), MAF represents the frequency of the alternative allele in the population, and u_i is drawn from a discrete Uniform$\{-1, 1\}$ distribution to allow non-null effects to take on positive or negative values. For remaining (unselected) SNPs, β_i was set to 0. Note that this ensures that the explained variance is the same for each simulated QTL regardless of its frequency. In addition, this guarantees that the sum of all explained variances per SNP is equal to the fixed total additive variance.

In all simulations, we selected a total of $n_{\text{large}} = 5$ large QTLs, varying the corresponding proportion of the total genetic additive variance σ_g^2 such that $k_{\text{large}} \in \{1\%, 2.5\%, 5\%\}$. We also selected $n_{\text{medium}} = 300$ medium QTLs, each representing $k_{\text{medium}} = 0.1\%$ of σ_g^2. We filled in the remaining genetic additive variance with small effect SNPs representing k_{small} =0.01% of σ_g^2. The number of these small effect SNPs varied according to the chosen value of k_{large}, respectively corresponding to $n_{\text{small}} = \{6500, 5750, 4500\}$. Finally, the phenotypic variance and mean were respectively set to $\sigma_y^2 = 100$ and $\mu = 0$, and SNP heritability $h^2 = \frac{\sigma_g^2}{\sigma_y^2}$ was set to one of two levels: $h^2 = \{0.2, 0.5\}$. For each simulation setting, 50 independent datasets were simulated.

Simulation of annotations

Annotations are defined here as informative when they are enriched in (i.e., contain a large proportion of) non-null markers, thus explaining a non-negligible portion of the total variance. To evaluate the impact of different annotation configurations on our models, we introduce four types of annotations: unenriched (i.e., uninformative), weakly enriched, moderately enriched and strongly enriched. Each annotation is constructed as shown in the upper half of Table 1 by randomly assigning different effect size SNPs (as well as their immediate neighbors). We note that each annotation constructed in this fashion contains around 1200–1300 markers.

These individual annotations can then be mixed and matched to form (partially) overlapping annotation sets. To simulate scenarios with different combinations of annotations, these individual annotations were then mixed and matched to form (partially)

Table 1 Simulation settings for the annotation scenarios

	Annotation enrichment			
	Strongly	**Moderately**	**Weakly**	**Unenriched**
SNP effect class				
Large	5	2	–	–
Medium	300	100	20	–
Low/null	150	300	400	450
Scenario				
A	1	1	–	–
B	1	1	1	1
C	–	2	1	1
D	2	2	3	2

We defined 4 levels of non-null SNP enrichment for simulated annotations (top part of the table) that can be mixed and matched to construct various scenarios (bottom part of the table). The top part shows the number of SNPs of each size effect class (rows) used to construct each level of annotation enrichment (columns). The bottom part indicates the number of each type of annotation enrichment (columns) used to construct each scenario (rows)

overlapping annotation sets. We focus on 4 annotation scenarios defined in the lower half of Table 1. Scenario A consists of one strongly and one moderately enriched annotations. Scenario B builds on scenario A by adding noise via two less enriched annotations. Scenario C represents a potentially less advantageous case, with no strongly enriched annotation. Finally, scenario D combines nine annotations with varying levels of enrichments, thus creating more overlaps and greater ambiguity. Given the number of simulated total large, medium and low QTL effects (5, 300 and more than 4500, respectively, see previous section), many large and medium-effect QTLs are then multi-annotated in all the scenarios.

Recall that overlapping annotations cannot be exploited for BayesRC. To include it in our comparisons, we used a naive work-around for this issue to randomly select a single annotation for each multi-annotated marker for BayesRC. As such, annotations used for BayesRC in the following results are not quite the same as those used for BayesRCπ and BayesRC+.

Production traits genomic prediction using QTL public database for growing pigs

Data description and pre-processing

A set of $n = 634$ and $n = 664$ animals (from 60 and 70 Large-White sows, respectively) from a population of 75% Large-White × 25% Creole crossbred pigs were raised in a temperate or tropical environment [18]. These offspring were descendants of a common batch of 10 boars that were themselves crossbred 50% Large-White × 50% Creole. A variety of traits were measured in this experiment using a common recording protocol in the two environments; trait measurements were pre-corrected for environment, age and sex effects. In this paper, we focus in particular on back fat thickness (BFT) and average daily weight gain (ADG), both measured at 23 weeks. For these traits, a total of $n = 1147$ and $n = 1146$ animals were respectively phenotyped. Animals were genotyped with the Illumina Porcine 60k BeadChip array.

To establish the potential impact of our models on prediction accuracy, we used a sibling-structured 10-fold cross validation procedure. For the descendants from each sire

in turn, we calculated the correlation between their observed corrected phenotypes and those predicted from models constructed on the descendants of the remaining 9 sires; validation correlations were averaged across the ten folds. As the number of offspring per boar was relatively homogeneous, we thus obtained an approximate split of 90–10% between training and validation sets. Using PLINK [19], we filtered out genotypes with a MAF < 0.01 for each training set independently, and retained only markers across all ten training sets ($p = 46,908$ and 4, 6881 SNPs for BFT and ADG, respectively).

Strategies for constructing annotations from pigQTLdb

Animal QTLdb (https://www.animalgenome.org/QTLdb) groups together curated results from genotype-phenotype association studies in several livestock species [20]. Cross-experiment QTL data from PigQTLdb (Release 45; SS11.1) for traits relevant to pig production were downloaded for eleven trait sub-hierarchy categories (anatomy, behavioral, blood parameters, conformation, fatness, fatty acid content, feed conversion, fowth, immune capacity, litter traits, reproductive organs). An additional "other" category was created for markers not included in PigQTLdb.

The potential utility of annotations depends on several factors, including their quality, relevance to the trait considered, LD around annotated mutations, and concordance in genotyping and annotation density (e.g., low density genotypes versus sequence-level information). An interesting strategy to consider is the use of expanded annotated windows around markers of interest, as well as the appropriate size of such a window; adding too many neighboring markers to annotations runs the risk of diluting the information they contribute. In this study, we explored three strategies for constructing annotations for genotyped markers: (1) using the exact position of known PigQTLdb markers ("regular"); (2) using the position of known PigQTLdb markers extended by a hard window, i.e., including the nearest up- and downstream neighbors ("hard"); and (3) using the position of known PigQTLdb markers as before but instead extended by a fuzzy window, where neighboring markers were allowed ambiguous assignment to both trait-specific and "other" categories ("fuzzy"). This latter strategy is particularly suited to BayesRCπ, as it allows markers in the neighborhood of annotated SNPs the possibility or not of inclusion with the respective annotation. "Regular" and "hard" annotations were used with BayesRC (with downsampling as before to avoid multiple annotations), BayesRCπ and BayesRC+, while "fuzzy" annotations were used with BayesRCπ alone. In the three strategies, 1.3%, 4.9% and 17.7% of markers were respectively assigned to two or more categories (Additional file 1: Fig. S4). The same three sets of annotations were used for both the BFT and ADG traits.

Results

Simulation results

We evaluated our proposed BayesRCπ and BayesRC+ models, compared to BayesR (without annotations) and BayesRC (with downsampled annotations to remove overlaps). We focused on their predictive power, as well as their ability to prioritize true QTLs, multi-annotated markers, and informative annotation categories.

Impact of annotation scenarios on prediction accuracy

We calculated the Pearson correlation between simulated and estimated phenotypes in the validation data for each model in each simulation scenario (annotation configuration, heritability, large QTL effect size). Results were averaged across the 50 simulated datasets for each setting. As a baseline, we consider the results for BayesR, which ignores annotation information. Hence, we tested the difference of correlation for each of the following pairs of models: {BayesRC vs BayesR}, {BayesRCπ vs BayesR} and {BayesRC+ vs BayesR}. In the case of $h^2 = 0.2$, average ($\pm$ sd) BayesR prediction accuracy was 0.211 ($\pm$ 0.050), 0.224 ($\pm$ 0.053) and 0.234 ($\pm$ 0.054) for k_{large} of 1%, 2.5% and 5%. For $h^2 = 0.5$, these values were respectively 0.447 ($\pm$ 0.048), 0.464 ($\pm$ 0.051) and 0.497 ($\pm$ 0.045). As expected, we observed improved predictions with higher heritability and stronger QTL effects. However, as suggested by the 95% level confidence intervals shown in Fig. 1 and Additional file 1: Fig. S1, not all of these differences are significant (paired t-test at 95% level). In particular, if for scenarios A and B in most settings (except for h^2 and small k), one can conclude that the methods incorporating functional annotations significantly improve the correlation, this is not always the case for scenarios C and D. In general, the tests were significant in cases where graphically the confidence interval did not reach the x axis.

We next turn to the impact observed for models incorporating annotations. The average differences in correlation with respect to BayesR are shown for BayesRC, BayesRCπ and BayesRC+ for all settings with $k_{\text{large}} = 1\%$ in Fig. 1. Results for $k_{\text{large}} = 2.5\%$ and 5% are shown in the Additional file 1. First, we observe that incorporating annotations in BayesRC does not lead to a universal gain in prediction accuracy compared to BayesR across scenarios. For $h^2 = 0.2$, BayesRC gains on average 0.6 ($\pm$1.6) and 0.2 ($\pm$1.6) points for scenarios A and B, but loses 0.2 ($\pm$1.4) and 0.3 ($\pm$1.5) correlation points for

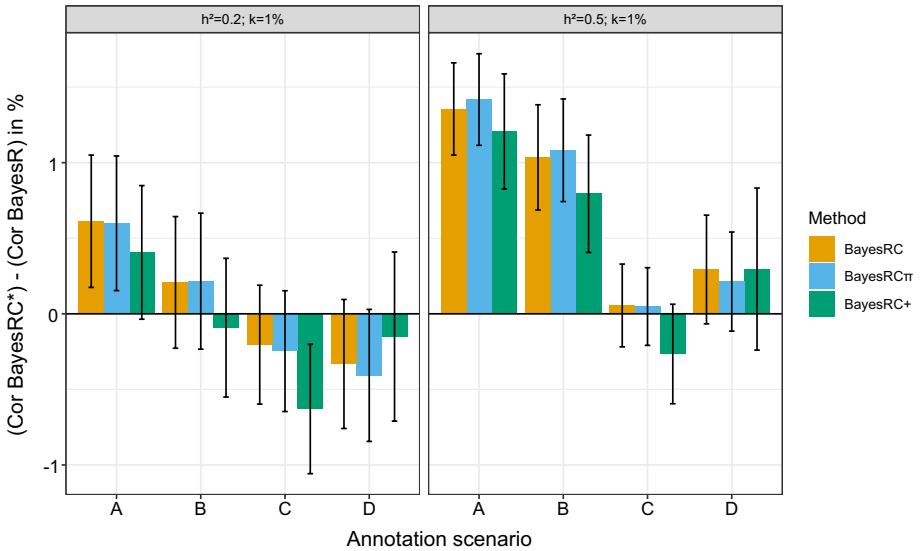


Fig. 1 Differences in validation correlation with respect to BayesR for four annotation scenarios. For $h^2 = 0.2$ and $h^2 = 0.5$ and $k_{\text{large}} = 1\%$, the difference in validation correlation between the three models including annotations (BayesRC, BayesRCπ and BayesRC+, gathered under the BayesRC* label) and BayesR, which does not include annotations. Colored bars and error bars represent averages and 95 % confidence interval across 50 simulated datasets

scenarios C and D. With higher heritability ($h^2 = 0.5$), BayesRC leads to average gains of 1.6 ($\pm$1.1), 1.0 ($\pm$1.2), 0.1 ($\pm$1.0), and 0.3 ($\pm$1.3) for scenarios A, B, C, and D. BayesRC thus seems to perform best when the provided annotations are informative and contain little noise (scenario A). For larger QTL sizes (Additional file 1: Fig. S1), similar trends are observed, with potentially higher prediction gains (up to 2 correlation points for $h^2 = 0.2$, $k_{\text{large}} = 5\%$ in scenario A). Moreover, a positive average gain is observed in all scenarios with BayesRC for sufficiently large QTLs and/or heritability.

We next looked at whether a better use of overlapping annotations could improve prediction accuracy. Overall, the differences in prediction between BayesRC and BayesRCπ are slight, corresponding to an average gain or loss of about 0.1 correlation points depending on the scenario and setting. At most, BayesRCπ led to a 1-point gain in correlation, for one dataset with $h^2 = 0.2$, $k_{\text{large}} = 5\%$ and scenario A. The small differences here can likely be explained by the construction of annotation sets, where the random downsampling of annotations for BayesRC still tends to categorize multi-annotated markers in an enriched annotation, a favorable situation for BayesRC; as such markers are already well-ranked by BayesRC, the impact of BayesRCπ will be limited. On the other hand, the underlying additive hypothesis of BayesRC+ distinguishes it more from BayesRC, so we can expect to see larger differences. Once again, predictions are better with more informative annotation scenarios (A and B) than with the noisier annotation sets in scenarios C and D. BayesRC+ underperforms BayesRC for scenarios A, B and C, but shows better results in scenario D, reaching an average gain of 0.5 points for $h^2 = 0.2$ and $k_{\text{large}} = 5\%$. BayesRC+ thus seems to be more robust to the addition of noise in annotations, given that there are some that are sufficiently informative. In the contrary case (scenario C), BayesRC+ risks too strongly prioritizing unimportant markers, thus deteriorating the prediction accuracy.

Model behavior for multi-annotated markers

BayesRC+ and BayesRCπ are designed to handle multi-annotated SNPs differently. With BayesRCπ, we aim to reclassify multi-annotated markers to the annotation whose enrichment which best matches their estimated effect. This has the added advantage of providing useful information about the probability of assignment for each annotation across iterations of the Gibbs sampling algorithm (via the PAIP statistic). On the other hand, BayesRC+ assumes that multi-annotated markers should be more likely to have a non-null effect (and thus, potentially a higher variance) in the model, counterbalancing an underestimation of QTL effects.

We focus on Scenario A here, as it provides the simplest illustration of model behavior for multi-annotated markers. In particular, it is constructed using only two annotations, one highly and one moderately enriched; as such, multi-annotated markers are necessarily included in both. In Fig. 2A, we represent the PAIP of the highly enriched annotation as a function of simulated marker size category ($h^2 = 0.5$, $k_{\text{large}} = 1\%$). We clearly distinguish the large effect QTLs, which have an average highly enriched PAIP of 0.556. In fact, 89.4% of these large QTLs were predominantly assigned (PAIP > 0.5) to the highly enriched rather than the moderately enriched annotation. In comparison, we observe an average highly enriched PAIP of {0.505,0.502,0.500} for the medium, small and null marker effect sizes respectively, and a corresponding proportion of preferential

assignment to the highly enriched annotation of 52.9%, 49.3% and 45.0%. This suggests that for QTLs with sufficiently large effects, using a maximum a posteriori (MAP) classification rule could provide useful insight into annotation enrichment. In Fig. 2B we show the densities of log posterior variance by assigned annotation category (via a MAP rule for the PAIP) for one representative dataset (scenario A, $h^2 = 0.5$, $k_{\text{large}} = 1\%$). The distributions of log $\widehat{V}_i$ are distinct for each category, though that of unannotated SNPs is clearly separated from those of the moderately and strongly enriched annotation categories. This seems to be consistent with the simulated distributions of annotations in scenario A, for which strong and medium QTLs were found in both annotations; larger differences in annotation enrichments may lead to a greater effect on PAIP values. For example, moving from $k_{\text{large}} = 1\%$ to $k_{\text{large}} = 5\%$, the share of σ_g^2 contributed by large QTLs in the strongly and moderately enriched annotations increases respectively from 5% to 25%, and from 2% to 10%. In this case, the strongly enriched PAIP for large QTLs increases drastically (to an average of 0.770) compared to the $k_{\text{large}} = 1\%$ case (Additional file 1: Fig. S2). Larger QTL effects thus lead to higher values for the strongly enriched PAIP, and a systematic assignment (100%) of large QTLs to the strongly enriched annotation using a MAP rule for PAIPs. At the same time, this proportion decreases to 49.4%, 43.1% and 40.6% respectively for the medium, low and null categories of simulated markers.

With BayesRC+, we instead seek to explicitly prioritize multi-annotated markers; this prioritization depends both on the number and quality of annotations for each marker. We now turn our attention to Scenario D, which is composed of a set of 9 annotations. In Fig. 3, we represent the log $\widehat{V}_i$ of large and medium QTLs as a function of the number of associated annotations ($h^2 = 0.5$, $k_{\text{large}} = 1\%$; results for $h^2 = 0.5$, $k_{\text{large}} = 2.5\%$ and 5% are shown in Additional file 1: Fig. S3). We first remark that the posterior variances of markers with smaller numbers of annotations tend to be underestimated. However, once a sufficient number of annotations are available (about 4 for large QTLs, and 7 for medium QTLs), estimated posterior variances approach the true simulated values. In

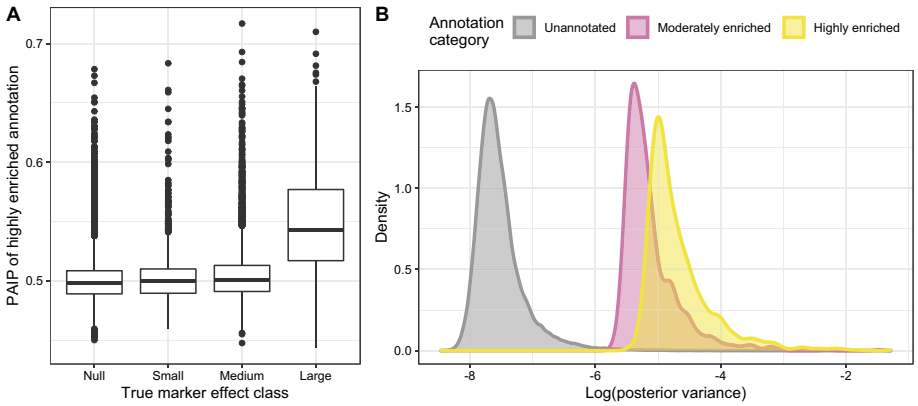


Fig. 2 Using the PAIP to interpret annotation importance for BayesRCπ. The PAIP, "Posterior Annotation Inclusion Probability", is defined as the frequency of marker assignment in each annotation across iterations. Results are shown for $h^2 = 0.5$, $k_{\text{large}} = 1\%$, and scenario A. **A** Posterior mean frequency of marker assignment to the strongly enriched annotation (i.e., strongly enriched PAIP) by simulated effect size category (null, small, medium, high). Results are averaged across 50 independent datasets. **B** Distribution of the log posterior variance of markers by PAIP-assigned annotation for one illustrative dataset

this scenario, no QTLs overlap all 9 annotations, and several configurations of multi-annotations are possible, each with a different potential downstream impact. Thus, a marker included in two weakly enriched annotations is less likely to be assigned a strong or medium effect in the model, compared to one included in two highly enriched annotations.

Impact of directly modeling multi-annotated markers versus down-sampling annotations

BayesRCπ shares greater similarity to BayesRC than BayesRC+, so its impact on prediction accuracy and marker variance estimation may potentially be more limited. However, assigning multi-annotated markers to a single annotation during data pre-processing, as we have done for BayesRC in this work, can have a negative effect on marker variance estimation. For example, inadvertently assigning a large-effect QTL to an uninformative annotation may lead to an underestimation of its effect. In Fig. 4 ($h^2 = 0.5$, $k_{\text{large}} = 1\%$, scenario D), we show the difference in estimated posterior variance for BayesRCπ and BayesRC+ on the full set of large QTLs (250 across simulated datasets) with respect to BayesRC. We recall that in this scenario, all large QTLs were multi-annotated (systematically for the two strongly enriched annotations, often for moderately enriched annotations, and rarely for weak or unenriched annotations). Each large QTL was randomly

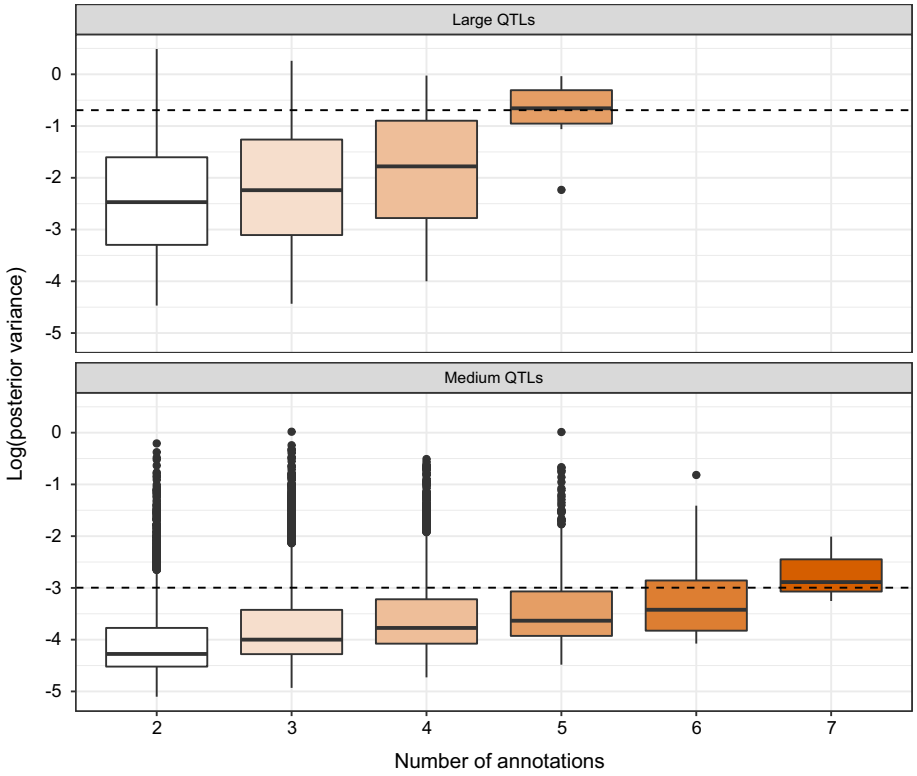


Fig. 3 Impact of number of annotations on markers for BayesRC+ model. Log posterior variance of large (top panel) and medium (bottom panel) effect QTLs by the number of associated annotations. All QTLs across the 50 independent datasets are represented. Results are shown for $h^2 = 0.5$, $k_{\text{large}} = 1\%$ and scenario D (including 9 annotations). The black dotted lines represent the true simulated value of log V_i for large and medium QTLs

assigned to a single annotation (unenriched, weakly, moderately, or strongly enriched) prior to fitting BayesRC, allowing an evaluation of the impact of these random assignments. As expected, estimated posterior variances are similar between BayesRC and BayesRCπ for large QTLs that were correctly randomly assigned to a strongly enriched annotation. However, those randomly assigned to a moderately enriched annotation saw an average gain of 0.016 ($\pm$ 0.032), and a gain of 0.058 ($\pm$ 0.058) for those erroneously assigned to a weakly enriched or unenriched annotation. By allowing multi-annotated markers to navigate among annotations, BayesRCπ thus avoids an underestimation of their effect related to an incorrect upstream assignment. A similar but stronger trend is observed for BayesRC+, though improved variance estimates are observed even for "correctly" assigned large QTLs. Average gains for BayesRC+ are 0.097 ($\pm$ 0.13), 0.120 ($\pm$ 0.14) and 0.171 ($\pm$ 0.12) for large QTLs randomly assigned to highly, moderately, and weakly/unenriched annotations, respectively. Exploiting multiple annotations in an additive manner thus has a strong effect on variance estimation, in addition to compensating for potential misassignment of important QTLs to a less informative annotation.

Improved rankings of large-effect QTLs by posterior variances when incorporating annotations

One way to prioritize markers is to focus on those with the largest estimated posterior variances. Figure 4 suggests that BayesRCπ and BayesRC+ both yield larger estimated

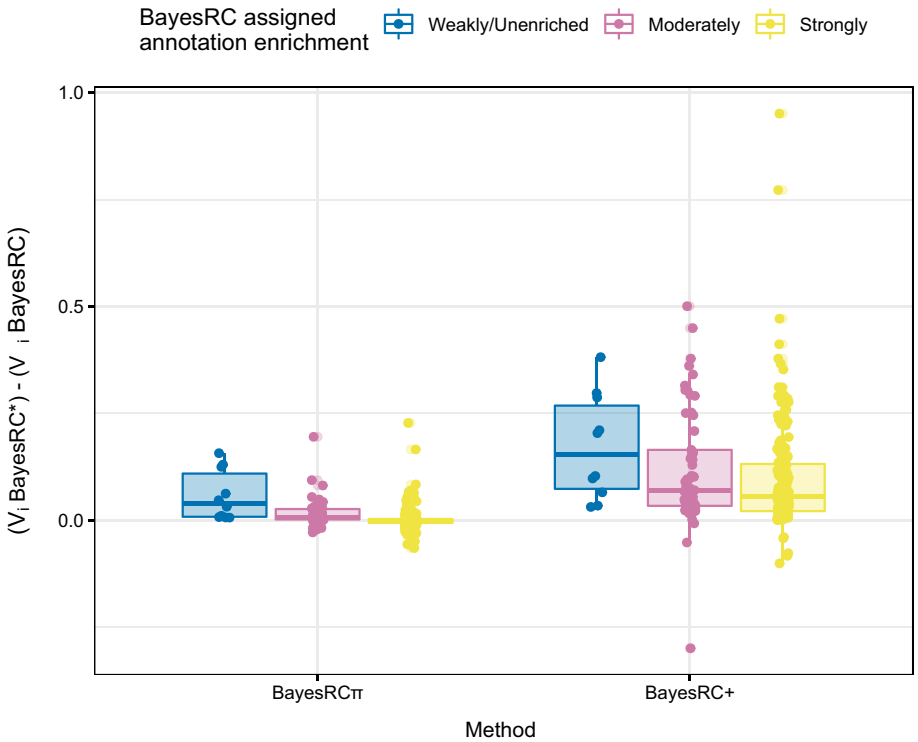


Fig. 4 Impact of BayesRC random annotation assignment on large-effect QTL variance estimation. Boxplots represent the difference with BayesRC in estimated posterior variances for large-effect QTLs for BayesRCπ and BayesRC+. Results are shown according to the randomly assigned annotation categories (strongly, moderately, weakly/unenriched) used for BayesRC. All large QTLs across the 50 independent datasets are represented ($h^2 = 0.5$, $k_{large} = 1\%$, scenario D with 9 annotations)

posterior variances to multi-annotated large QTLs than BayesRC; in turn, this tends to lead to a better average ranking for large QTLs (Table 2) for most simulation settings, especially for BayesRC+. One exception is the setting with $h^2 = 0.5$ and $k_{\text{large}} = 5\%$, a favorable situation where BayesR readily prioritizes the simulated large QTLs without use of annotations. Otherwise in scenarios A, B and D, large QTL rankings are systematically improved in BayesRCπ compared to BayesRC, and in BayesRC+ compared to BayesRCπ. Scenario C behaves somewhat differently, where large QTL rankings are generally worse and the best performing method depends on the settings. This is perhaps unsurprising, as scenario C is composed of the least informative annotation set. Overall, large QTL rankings for BayesRC and BayesRCπ are best in scenario A, followed by scenarios B and D; rankings for BayesRC+ appear to be more stable across scenarios for a given simulation setting. This robustness can likely be explained by the fact that BayesRC+ takes full advantage of the strongly enriched annotations present in scenarios A, B and D, while by design BayesRCπ may allow QTLs to be assigned to less enriched annotations in at least some iterations of the algorithm. This suggests that for the purposes of prioritizing QTLs, BayesRCπ may be more sensitive to the addition of noise in annotations.

Genomic prediction results for pigs data

In the following, we illustrate the performance of our models on real data from a population of growing pigs, in conjunction with annotations related to multiple production traits from a public database.

Impact of pigQTLdb annotation strategies on prediction accuracy

We first sought to determine whether the pigQTLdb annotations appear to contribute useful information for predicting BFT and ADG, and if so, whether the "hard" or "fuzzy" window-based annotations were beneficial (Fig. 5). As a baseline without annotations, BayesR yielded an average correlation of 0.21 ($\pm$ 0.08) and 0.26 ($\pm$ 0.16) for ADG and BFT, respectively. Two different behaviors are observed for model performance in the two traits when incorporating the pigQTLdb annotations. For ADG, prediction accuracy is deteriorated by the "regular" annotation ($-$ 1.3, $-$ 1.2, and $-$ 0.6 correlation points

Table 2 Average rankings of large QTLs by estimated posterior variance

Annotation scenario		None	A (2 annot.)			B (4 annot.)			C (4 annot.)			D (9 annot.)		
h^2	k_{large} (%)	R	RC	RCπ	RC+	RC	RCπ	RC+	RC	RCπ	RC+	RC	RCπ	RC+
0.2	1	10286	501	479	**342**	652	617	**433**	4930	4446	**4212**	1268	1112	**322**
	2.5	2991	140	120	**91**	188	167	**110**	**1577**	1933	1755	340	303	**93**
	5	597	21	19	**14**	29	25	**18**	392	**361**	425	44	41	**16**
0.5	1	2711	162	140	**88**	194	152	**102**	1482	**1303**	1436	365	261	**102**
	2.5	127	12	11	**8**	21	12	**10**	**73**	74	104	23	20	**10**
	5	4	**3**	**3**	**3**	**3**	**3**	**3**	**4**	26	55	**3**	**3**	**3**

Mean rank (by decreasing estimated posterior variance) of large QTLs, averaged across 50 independent datasets for each setting (heritability, k_{large}) and each method (R = BayesR, RC = BayesRC, RCπ = BayesRCπ and RC+ = BayesRC+). With the exception of BayesR, which does not use annotations, results are presented by annotation scenario (A, B, C and D). Boldface is used to indicate the best ranking obtained for each setting. As each dataset contains 5 large QTLs, the highest average ranking that can be obtained is equal to 3 (average of 1 to 5)

for BayesRC, BayesRCπ and BayesRC+). However, by extending annotations to include the nearest neighboring markers ("hard") led to respective gains of + 1.2, + 1.7, and + 1.4 points compared to BayesR. A similar gain in correlation (+ 1.4 points) was also achieved with BayesRCπ and "fuzzy" annotations allowing for an ambiguous neighborhood extension. For BayesRCπ, we thus observe a difference of 2.9 correlation points between the "regular" and "hard" annotation strategies, highlighting the potential impact this upstream step plays. In the case of ADG, it is thus possible to identify a useful strategy for constructing and including annotation sets from pigQTLdb to improve trait prediction. On the contrary, for BFT these same pigQTLdb annotations appear to be less relevant for the task of prediction. For all annotation strategies, a largely equivalent or deteriorated prediction performance compared to BayesR is observed. Thus, for BayesRC, we go respectively from a zero average gain using "regular" annotations to a loss of − 0.6 points for "hard". For BayesRCπ, we go from a gain of + 0.1 points to a loss of − 1.0 points and − 0.2 points respectively with the "regular", "hard" and "fuzzy" annotations. BayesRC+ presents the best results for this trait, with a gain of + 0.9 points and + 0.6 points for the "regular" and "hard" annotations; however, the results are highly variable, and appear to be insignificant for this trait.

PigQTLdb annotation category interpretation using BayesRCπ

In this study, we used sets of relatively common annotations from pigQTLdb without any upstream relevance selection; it may thus be of interest to evaluate the contribution of each to prediction using outputs from BayesRCπ (Fig. 6, "fuzzy" annotations). The average proportion of medium- or large-effect SNPs assigned to each annotation using the PAIP highlights a difference in estimated enrichment across annotation categories

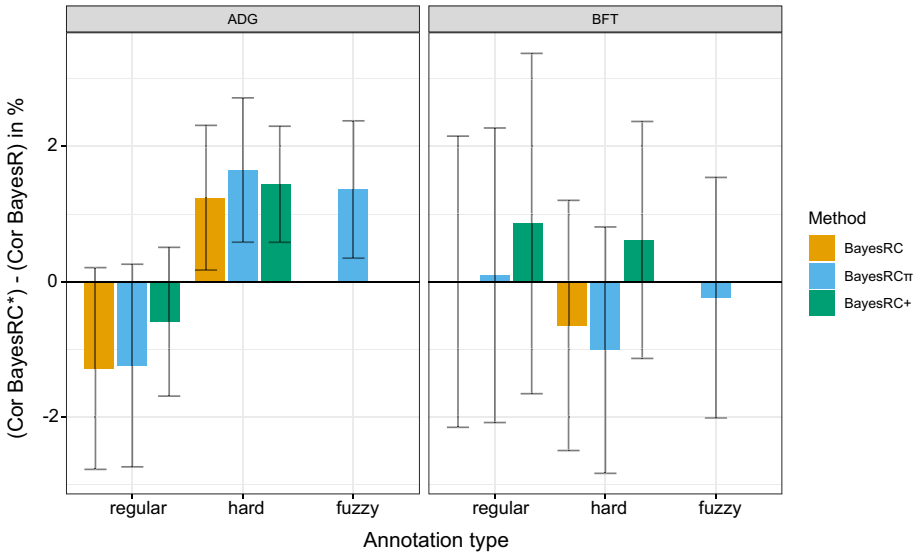


Fig. 5 Differences in validation correlation with respect to BayesR for backcross pig data and various pigQTLdb annotations. Difference of correlation between the 3 methods including functional annotation (BayesRC, BayesRCπ and BayesRC+, gathered under the BayesRC* label) and BayesR, for average daily gain (ADG) and and back fat thickness (BFT) for three strategies of annotation construction: regular, hard, and fuzzy. Colored bars and error bars represent averages and standard deviations across 10-fold validation datasets

and between traits. For ADG, 7.0% of markers are classified as having medium or large effects in the Behavorial annotation compared to 1.8% for Immune capacity, i.e. almost 4 times less. The Unannotated category is made up of a large number of markers estimated to have null or small effects, and thus features few medium- to large-effect SNPs. Annotation category ranks were found to be identical when ordered by the median variance of assigned SNPs. Moreover, in the top 1% of markers (i.e., 469 SNPs, ranked by $\widehat{V}_i$), more than half were assigned to the Anatomy (21%), Behavorial (17%) and Conformation (15%) annotations. Conversely, Immune capacity represents only 0.4% of these top SNPs, less than the Unannotated category (1.4%). Taken together, we can for example question the relevance of the Immune capacity annotation for predicting this trait.

With respect to BFT, given the poor prediction quality observed for all annotation strategies (Fig. 5), we must interpret the relevance of annotations with caution. The most enriched annotation for BFT is Feed conversion, while Behavorial drops from the first position in ADG to the 7th here. Immune capacity again shows very low enrichment (1.4%). In general, the annotations appear to be more weakly enriched in BFT (3.5% of markers in medium to high classes) than in ADG (4.5%). On the contrary, the enrichment of the Unannotated category increases slightly in BFT (0.3%) compared to ADG (0.2%), suggesting the possibility of important markers for this trait being unannotated in PigQTLdb.

Fuzzy and hard expanded windows for annotation construction

The "fuzzy" strategy we have explored here is essentially best adapted to BayesRCπ, as it allows for ambiguity in annotation assignment by design. In the case of ADG, this annotation strategy provided good predictions. To give additional insight into the behavior of BayesRCπ for "fuzzy" annotations, we provide an illustration for ADG in Additional file 1: Fig. S5 of the posterior variance of markers, averaged over the ten cross-validation datasets, according to their relative position (zoomed in between 8000 and 12000 for

Fig. 6 Interpretation of pigQTLdb annotations using BayesRCπ according to the trait for backcross pig data. For each annotation, we represent the proportion of markers assigned to the medium and large SNP effect size class, average on the 10-fold datasets. The two panels shows the results for the two studied trait, ADG and BFT. The order of annotation on y-axis is based on the decreasing value of the sum of large and medium proportion

clarity). We note that a large discontinuous block of markers were directly annotated within pigQTLdb. We focus our attention on markers with ambiguous annotations (i.e., those neighboring the pigQTLdb markers, and thus included with annotations in the "hard" and "fuzzy" strategies). These "ambiguously annotated" SNPs systematically have larger estimated values for $\widehat{V}_i$ in both the "fuzzy" and "hard" strategies. For example, the marker at relative position 11554 (which is a direct neighbor of a pigQTLdb-annotated marker) is estimated to have a posterior variance of 0.85 with the "regular" annotation, 16.81 with "fuzzy" and 24.78 with "hard". To avoid overestimating the effect of unimportant neighboring markers, the "fuzzy" strategy allows for their potential assignment to the Unannotated category, thus giving them less chance to be estimated as non-null in the model. This dampening effect can be seen in Additional file 1: Fig. S5, where some of the ambiguously annotated markers have estimated posterior variances intermediate to those the unannotated and annotated markers.

Comparison of top ranked SNPs by estimated posterior variance

Incorporating biological information in genomic prediction models has the potential to improve their interpretability by better prioritizing informative markers. In a similar spirit to genome-wide association studies (GWAS), where markers are prioritized by *p*-value from a univariate test of association, we can use the estimated posterior variances from the Bayesian models to rank markers. To evaluate the similarities and differences in top-ranked SNPs for each method, we selected for each the top 100 markers according to the average (across 10 cross-validation datasets) estimated posterior variances. Results are shown in Fig. 7 for ADG and "hard" pigQTLdb annotations (a setting with good prediction results) for BayesRC, BayesRCπ and BayesRC+; BayesR, without annotations, is also included for reference. 41 markers were ranked in the top 100 by all methods, with most (40 markers) included in at least one annotation. More than half (53 %) of the markers prioritized by BayesR were not highly ranked by the other methods, including 31 unannotated markers. On the other hand, unsurprisingly all 30 markers prioritized by BayesRC, BayesRCπ and BayesRC+ were annotated. Finally, our newly proposed models BayesRCπ and BayesRC+ highly ranked 32 markers not highly ranked by the others. Of these, all of which were annotated, 18 featured multiple annotations: 8 markers with 2 annotations, 3 with 3 annotations, 2 with 5 annotations and 1 with 7 annotations.

Discussion

In this work, we presented two new genomic prediction methods (BayesRCπ and BayesRC+) that allow for the use of multiple, overlapping annotations, which until now has been a limiting factor in the BayesRC reference method. This led us to compare the three methods for integrating annotations in simulated and real data with different genomic architectures and sets of annotations. To evaluate the interest of adding annotation information to genomic prediction models, we used BayesR as a reference. We constructed a variety of simulations to study the impact of favorable and unfavorable annotation sets and identify appropriate use cases for each of the proposed models. Since BayesRCπ operates by stochastically classifying multi-annotated markers to a single annotation, we often observed little difference with BayesRC. We

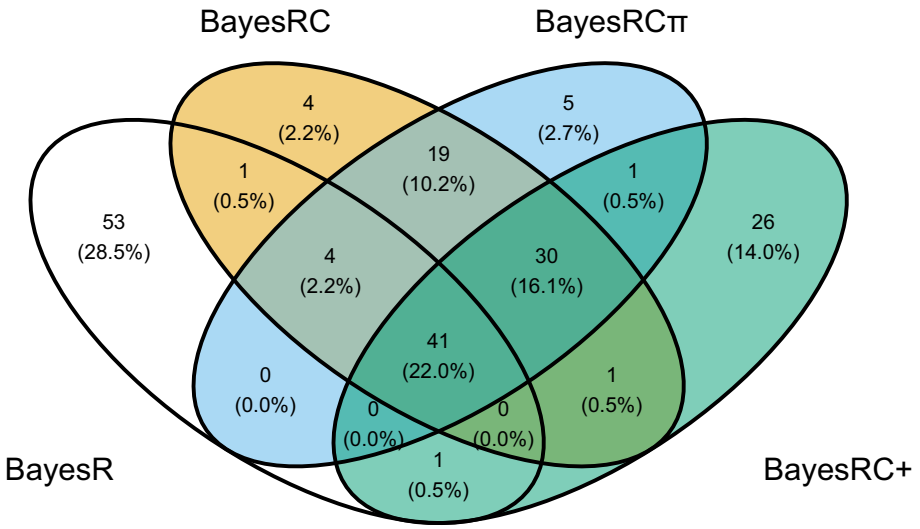


Fig. 7 Top-ranked SNPs by estimated posterior variance in backcross pig data. Venn diagram of the lists of 100 most highly ranked SNPs by estimated posterior variance for each Bayesian method, averaged across ten cross-validation datasets

remark that building relevant and irrelevant annotations in our simulations was not a straightforward task, as the incorporation of a single large QTL (or a single marker in LD with a large QTL) de facto changes the enrichment of a given annotation. Thus, large QTLs were mostly multi-annotated for highly or moderately enriched annotations, a favorable situation for BayesRC despite the random assignment we used. It is also important to note that if annotations contain few overlaps, differences with BayesRC would be further reduced for both BayesRCπ and BayesRC+. However, the annotation scenarios we considered in our simulation study nevertheless allowed us to highlight the behavior of BayesRCπ and BayesRC+ in a variety of situations.

In addition to the simulated data, we illustrated our approaches on data from a backcross population of growing pigs in conjunction with annotations from pigQLTdb. All available annotations were used for two traits (ADG and BFT) to study differences in model behavior. As for the simulations, we observed situations where the use of annotations was not advantageous, as was the case for BFT. Another choice that was decisive in the prediction accuracy of the models was the sibling-structured cross-validation procedure we used. Animals were grouped into ten sets by their fathers, meaning that animals within the validation set tended to be genetically similar to one another, and potentially distant from the training data. This necessarily complicates the prediction task compared to a fully randomized cross-validation procedure. It has been previously suggested that the use of annotations in situations where the validation population is genetically distant from the learning population could lead to improvement; this holds on average for the ADG trait when appropriate annotations are used.

The optimal use of our proposed models appears to depend on several factors, including the genomic architecture, the relevance and construction strategy of annotations, the number and interpretation of overlapping annotations, and the desired

goal. There were in fact cases where BayesR (without annotations) yielded better results than models including annotations. These likely correspond to situations where annotations are not relevant to the trait of interest. In such cases, BayesRCπ provides tools to interpret the relevance of annotations a posteriori using the PAIP statistic. On the other hand, although its assumption of additivity may be overly strong in some cases, BayesRC+ tends to compensate for the underestimation of QTL posterior variances and encourages a prioritization of multi-annotated markers. As such, for BayesRC+ it is important that overlapping annotations represent not uncertainty but rather complementary information leading to greater confidence in the impact of SNPs on the phenotype.

Via the "regular", "hard" and "fuzzy" window annotation strategies, we saw that there are cases where it may be interesting to expand annotations to include neighboring markers. This is likely linked to uncertainty of the precise location of causal mutations and as well as LD structures with nearby markers, especially when SNP density is low to medium (under 100K) and/or LD extent is large (familial structures). The fuzzy annotation strategy represents a potentially interesting approach to capitalize on the modeling specificity of BayesRCπ, as it provides indicators of the relevance of multiple annotations. It could also be potentially interesting to use BayesRCπ outputs to refine and adapt the annotation set, for example by merging, splitting or deleting some annotations. A more exhaustive exploration of different strategies for constructing annotations would be a useful avenue for future research.

The methods we have proposed here take into account annotations coded as binary classifications, and thus do not reflect the potentially continuous nature of underlying annotations (e.g., GWAS *p*-values). It would therefore be interesting to extend these approaches to handle continuous annotations, in particular to allow the flexibility to give greater weight to certain markers in annotations, or to certain annotations for a given marker. This future development has the potentially to more fully exploit the heterogeneous and complex functional information that is increasingly available.

Conclusions

The full use of complex and potentially overlapping annotations can improve genomic prediction and the estimation of posterior variance for markers. These annotations impact the results of the different prediction methods used, and their use must make sense with respect to the studied trait. We proposed two new methods based on different assumptions on the interpretation of multi-annotated markers: allowing such SNPs to be assigned to the most representative annotation (BayesRCπ) or cumulatively assigning a greater weight for such SNPs (BayesRC+). These models each perform well in different settings, and lead to different downstream analyses. We observed average gains of up to 2 correlation points compared to a model ignoring annotations in simulated data, and up to 1.7 points on real data. Models integrating annotations, and in particular BayesRCπ and BayesRC+, are particularly good at prioritizing medium and large QTLs according to their estimated posterior variances. BayesRCπ and BayesRC+, in addition to BayesCπ, BayesR and BayesRC, have been implemented in an open-source software package called *BayesRCO*. Many strategies for constructing annotations are possible, and we compared several approaches based on extended hard or fuzzy windows around

known pigQTLdb hits. In future work, a promising approach would be to further extend our models to fully account for continuous, rather than categorical, annotations.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-04914-5.

Additional file 1: Figures S1–S5 and pseudocode for the BayesR, BayesRC, BayesRCπ, and BayesRC+ algorithms.

Acknowledgements
Not applicable.

Author contributions
PC and AR conceived and designed the study. FM developed the proposed models, implemented them in the BayesRCO software, and performed all analyses. HG provided the real data and contributed to the interpretation of results. FM took the lead in writing the manuscript. All authors discussed results, contributed to the final manuscript, and approved the submitted version. All authors read and approved the final manuscript.

Funding
This work is part of the GENE-SWitCH project that has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement n^{o}817998. This work also benefited from the clustering activities organized with the BovReg project, part of the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement n^{o}815668. The financial support of the French National Agency of Research (ANR PigHeaT, ANR-12-ADAP-0015) is also gratefully acknowledged.

Availability of data and materials
We implemented the two novel methods BayesRCπ and BayesRC+ in a Fortran software package named ***BayesRCO*** (BayesRC for Overlapping annotations), available at https://github.com/fmollandin/BayesRCO, https://doi.org/10.5281/zenodo.6809653. In addition, the package also implements BayesCπ, BayesR, and BayesRC. The associated GitHub repository includes a full user's guide, detailing the different types of parameterization possible. Code used to simulate phenotypes based on genotype data is available in the "Simulation" folder of the following Github repository: https://github.com/fmollandin/BayesR_Simulations. An R script to extract and format pigQTLdb annotations is available in the following repository: https://github.com/andreamrau/tidyqtldb.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 16 February 2022 Accepted: 25 August 2022
Published online: 06 September 2022

References

1. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
2. Heslot N, Jannink J-L, Sorrells ME. Perspectives for genomic selection applications and research in plants. Crop Sci. 2015;55(1):1–12.
3. Voss-Fels KP, Cooper M, Hayes BJ. Accelerating crop genetic gains with genomic selection. Theor Appl Genet. 2019;132(3):669–86.
4. Mardis ER. DNA sequencing technologies: 2006–2016. Nat Protoc. 2017;12(2):213.
5. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91(11):4414–23.
6. Gianola D, de Los Campos G, Hill WG, Manfredi E, Fernando R. Additive genetic variability and the Bayesian alphabet. Genetics. 2009;183(1):347–63.
7. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics. 2011;12(1):1–12.
8. Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, Mason BA, Goddard ME. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012;95(7):4114–29. https://doi.org/10.3168/jds.2011-5019.
9. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLOS Genet. 2015;11(4):1004969. https://doi.org/10.1371/journal.pgen.1004969.

10. Mollandin F, Rau A, Croiseau P. An evaluation of the predictive performance and mapping power of the BayesR model for genomic prediction. G3. 2021;11(11):225.
11. Edwards SM, Sørensen IF, Sarup P, Mackay TF, Sørensen P. Genomic prediction for quantitative traits is improved by mapping variants to gene ontology categories in drosophila melanogaster. Genetics. 2016;203(4):1871–83.
12. Giuffra E, Tuggle CK, Consortium F. Functional annotation of animal genomes (FAANG): current achievements and roadmap. Annu Rev Anim Biosci. 2019;7:65–88.
13. Li Z, Gao N, Martini JWR, Simianer H. Integrating gene expression data into genomic prediction. Front Genet. 2019;10:126. https://doi.org/10.3389/fgene.2019.00126.
14. Morgante F, Huang W, Sørensen P, Maltecca C, Mackay TF. Leveraging multiple layers of data to predict drosophila complex traits. G3. 2020;10(12):4599–613.
15. MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ, Schrooten C, Hayes BJ, Goddard ME. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics. 2016;17(1):144. https://doi.org/10.1186/s12864-016-2443-6.
16. Kemper KE, Reich CM, Bowman PJ, van der Jagt CJ, Chamberlain AJ, Mason BA, Hayes BJ, Goddard ME. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. Genet Sel Evol. 2015;47(1):29. https://doi.org/10.1186/s12711-014-0074-4.
17. Uemoto Y, Sasaki S, Kojima T, Sugimoto Y, Watanabe T. Impact of QTL minor allele frequency on genomic evaluation using real genotype data and simulated phenotypes in Japanese Black Cattle. BMC Genet. 2015;16(1):134.
18. Gourdine J-L, Riquet J, Rosé R, Poullet N, Giorgi M, Billon Y, Renaudeau D, Gilbert H. Genotype by environment interactions for performance and thermoregulation responses in growing pigs. J Anim Sci. 2019;97(9):3699–713.
19. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
20. Hu Z-L, Park CA, Reecy JM. Bringing the animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. Nucleic Acids Res. 2022;50(D1):956–61.

Publisher's Note

3.10 Matériel supplémentaire

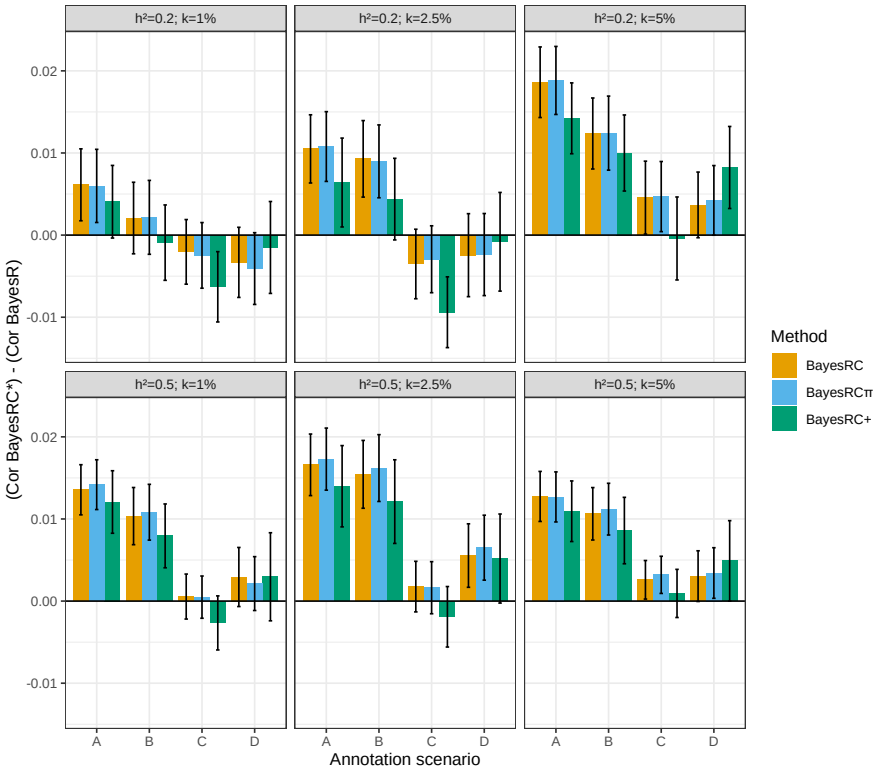
3.10.1 Materiel supplémentaire article

Supplementary Materials: Accounting for overlapping annotations in genomic prediction models of complex traits

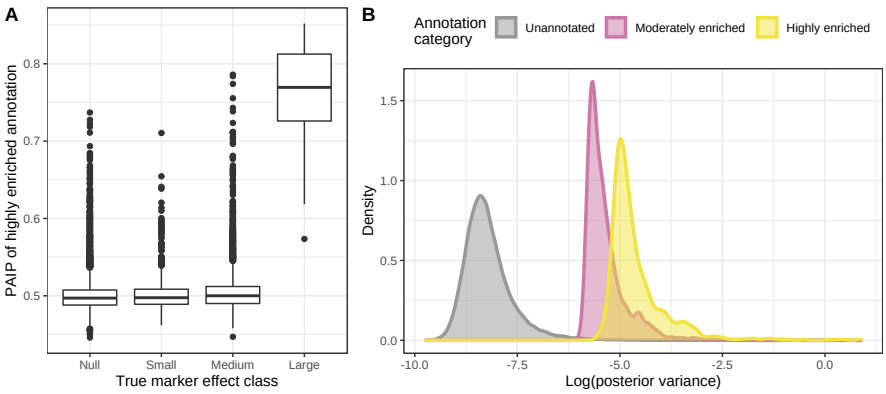
Fanny Mollandin[1], Hélène Gilbert[2], Pascal Croiseau[1], and Andrea Rau[1,3]

[1] Université Paris-Saclay, INRAE, AgroParisTech, GABI
[2] GenPhySE, Université de Toulouse, INRAE, ENVT
[3] BioEcoAgro Joint Research Unit, INRAE, Université de Liège, Université de Lille, Université de Picardie Jules Verne

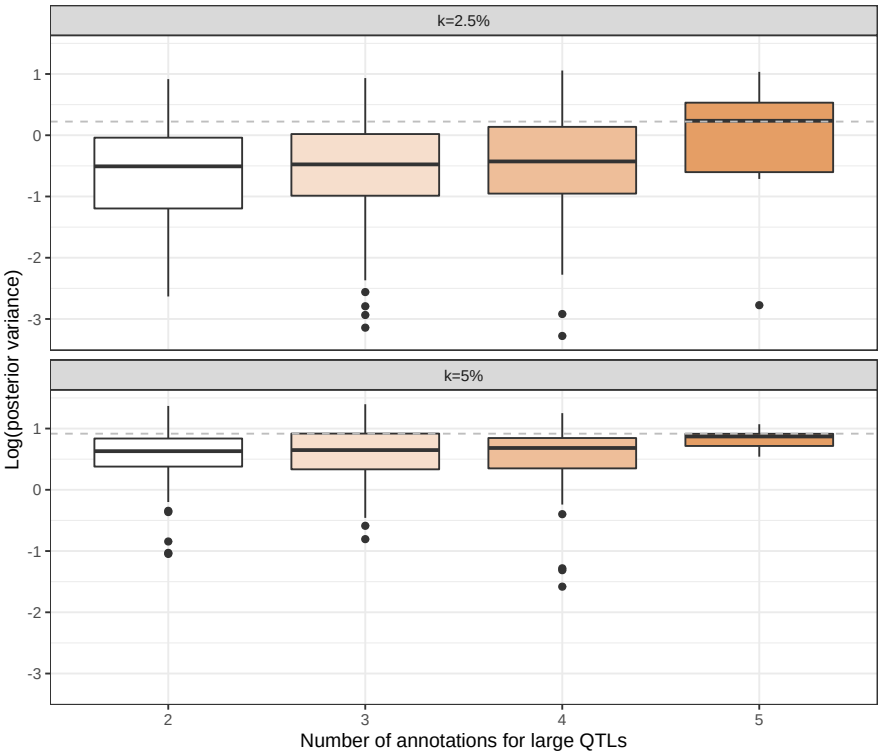
S1 Supplementary Figures



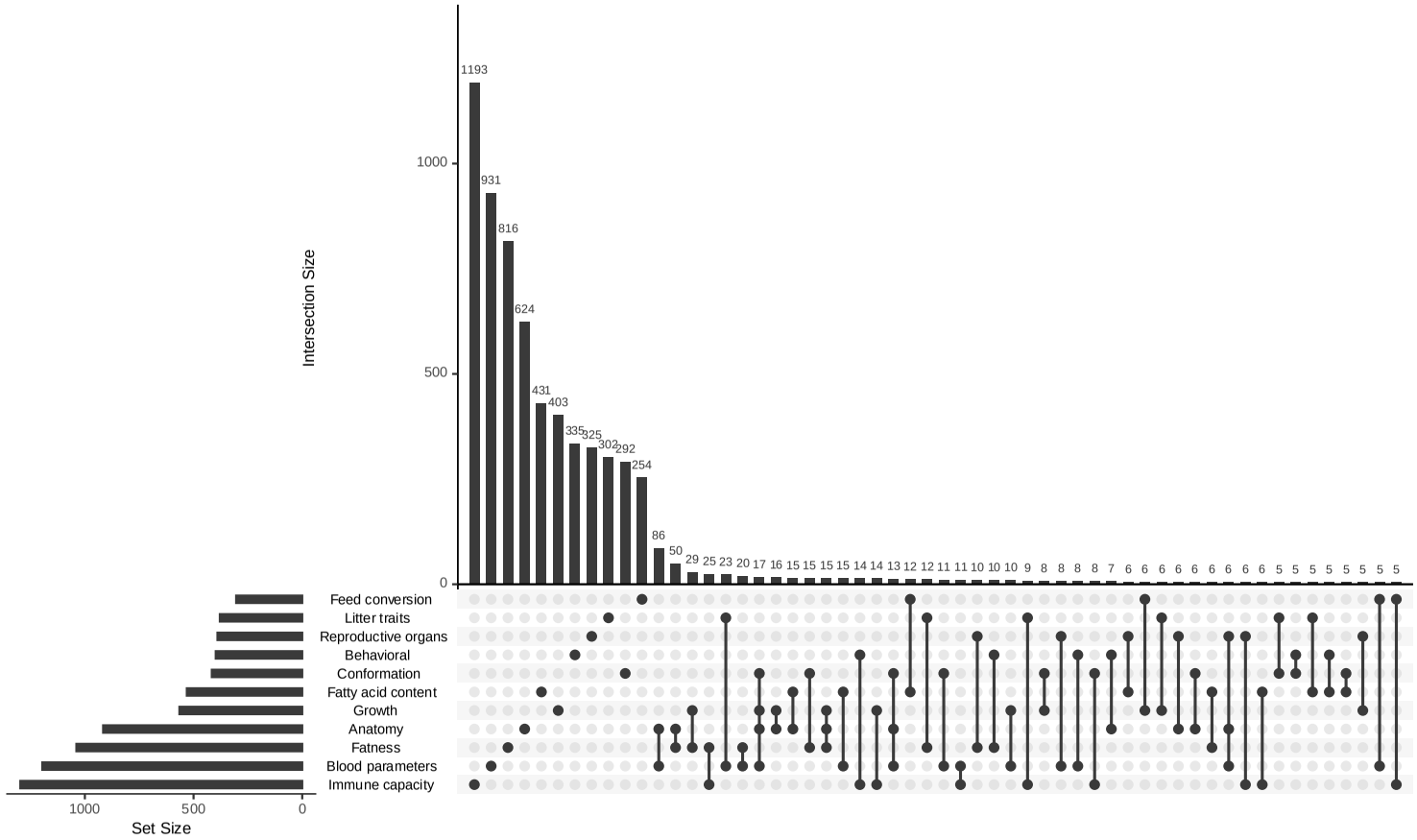
Supplementary Figure S1: Differences in validation correlation with respect to BayesR for four annotation scenarios. For $h^2 = 0.2$ and $h^2 = 0.5$ and $k_{\text{large}} = 1\%$, $k_{\text{large}} = 2.5\%$ and $k_{\text{large}} = 5\%$, the difference in validation correlation between the three models including annotations (BayesRC, BayesRCπ and BayesRC+) and BayesR, which does not include annotations. Colored bars and error bars represent averages and the 95% confidence interval across 50 simulated datasets.



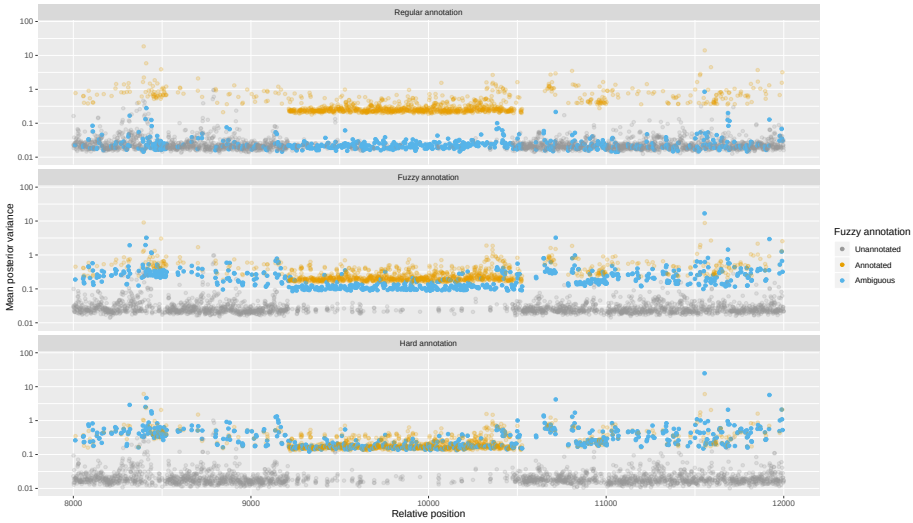
Supplementary Figure S2: Using the PAIP to interpret annotation importance for BayesRCπ. Results are shown for $h^2 = 0.5$, $k_{\text{large}} = 5\%$, and scenario A. (A) Posterior mean frequency of marker assignment to the strongly enriched annotation (i.e., strongly enriched PAIP) by simulated effect size category (null, small, medium, high). Results are averaged across 50 independent datasets. (B) Distribution of the log posterior variance of markers by PAIP-assigned annotation for one illustrative dataset.



Supplementary Figure S3: Impact of number of annotations on markers for BayesRC+ model. Log posterior variance of large effect QTLs by the number of associated annotations. All QTLs across the 50 independent datasets are represented. Results are shown for $h^2 = 0.5$ and scenario D (including 9 annotations), and for $k_{\text{large}} = 2.5\%$ (top panel) and $k_{\text{large}} = 5\%$ (bottom panel). The black dotted lines represent the true simulated value of $\log V_i$ for each k_{large}.



Supplementary Figure S4: pigQTLdb annotation structure. Upset plot of the 11 pigQTLdb annotations, representing their respective numbers and the overlaps between annotations. For the purpose of visibility, only overlaps between annotations concerning more than 5 markers have been represented.



Supplementary Figure S5: Impact of annotation construction on variance estimation in BayesRCπ for ADG trait. The three panels correspond to "regular" (top), "fuzzy" (middle), and "hard" (bottom) annotations for BayesRCπ. For each, the posterior variance of markers, averaged on the 10-fold datasets, is represented and colored according to its fuzzy window annotation: unannotated, annotated or ambiguously annotated (i.e both annotated and unannotated).

S2 Algorithm pseudocode and details

Algorithm 1 BayesR pseudo-code

Initialization
for each iteration **do**
  update $\mu|$.
  for i in 1:p **do**
    $\tilde{y} = \tilde{y} + x_{.,i}\beta_i$
    for k in 1:4 **do**
      compute $LogL(i, k|\pi)$
    end for
    assign SNP i to $\hat{k} \in \{1, 2, 3, 4\}$ $|LogL(i, .|\pi)$
    update $\beta_i|\hat{k}$
    $\tilde{y} = \tilde{y} - x_{.,i}\beta_i$
  end for
  update $\sigma_g^2, \sigma_e^2, \pi$
end for

Algorithm 2 BayesRC pseudo-code; $\#C_i = 1\ \forall i$

Initialization
for each iteration **do**
  update $\mu|$.
  for i in 1:p **do**
    $\tilde{y} = \tilde{y} + x_{.,i}\beta_i$ $c = C_i$
    for k in 1:4 **do**
      compute $LogL(i, k|c, \pi_c)$
    end for
    assign SNP i to $\hat{k} \in \{1, 2, 3, 4\}$ $|LogL(i, .|c, \pi_c)$
    update $\beta_i|\hat{k}$
    $\tilde{y} = \tilde{y} - x_{.,i}\beta_i$
  end for
  update $\sigma_g^2, \sigma_e^2, \pi_{c_1}, \pi_{c_2}, ...\pi_{c_m}$
end for

Algorithm 3 BayesRCπ pseudo-code; $\#C_i \geq 1\ \forall i$

Initialization
for each iteration **do**
  update $\mu|$.
  for i in 1:p **do**
    for $c \in C_i$ **do**
      compute $LogL(i, c|p_i)$
    end for
    assign SNP i to $\hat{c} \in C_i\ |LogL(i, .|p_i)$
    $\tilde{y} = \tilde{y} + x_{.,i}\beta_i$
    for k in 1:4 **do**
      compute $LogL(i, k|\hat{c}, \pi_{\hat{c}})$
    end for
    assign SNP i to $\hat{k} \in \{1, 2, 3, 4\}\ |LogL(i, .|\hat{c}, \pi_{\hat{c}})$
    update $\beta_i|\hat{k}$
    $\tilde{y} = \tilde{y} - x_{.,i}\beta_i$
  end for
  update $\sigma_g^2, \sigma_e^2, \pi_{c_1}, \pi_{c_2}, ...\pi_{c_m}, p_1, p_2, ..., p_p$
end for

Algorithm 4 BayesRC+ pseudo-code; $\#C_i \geq 1\ \forall i$

Initialization
for each iteration **do**
  update $\mu|$.
  for i in 1:p **do**
    $\tilde{y} = \tilde{y} + x_{.,i}\beta_i$
    for $c \in c_i$ **do**
      for k in 1:4 **do**
        compute $LogL(i, k|c, \pi_c)$
      end for
      assign SNP i to $\hat{k} \in \{1, 2, 3, 4\}\ |LogL(i, .|c, \pi_c)$
      update $\beta_{i,c}|\hat{k}$
      $\tilde{y} = \tilde{y} - x_{.,i}\beta_{i,c}$
    end for
    $\beta_i = \sum_{c \in c_i} \beta_{i,c}$
  end for
  update $\sigma_g^2, \sigma_e^2, \pi_{c_1}, \pi_{c_2}, ...\pi_{c_m}$
end for

3.10.2 Schéma résumé des modèles utilisés

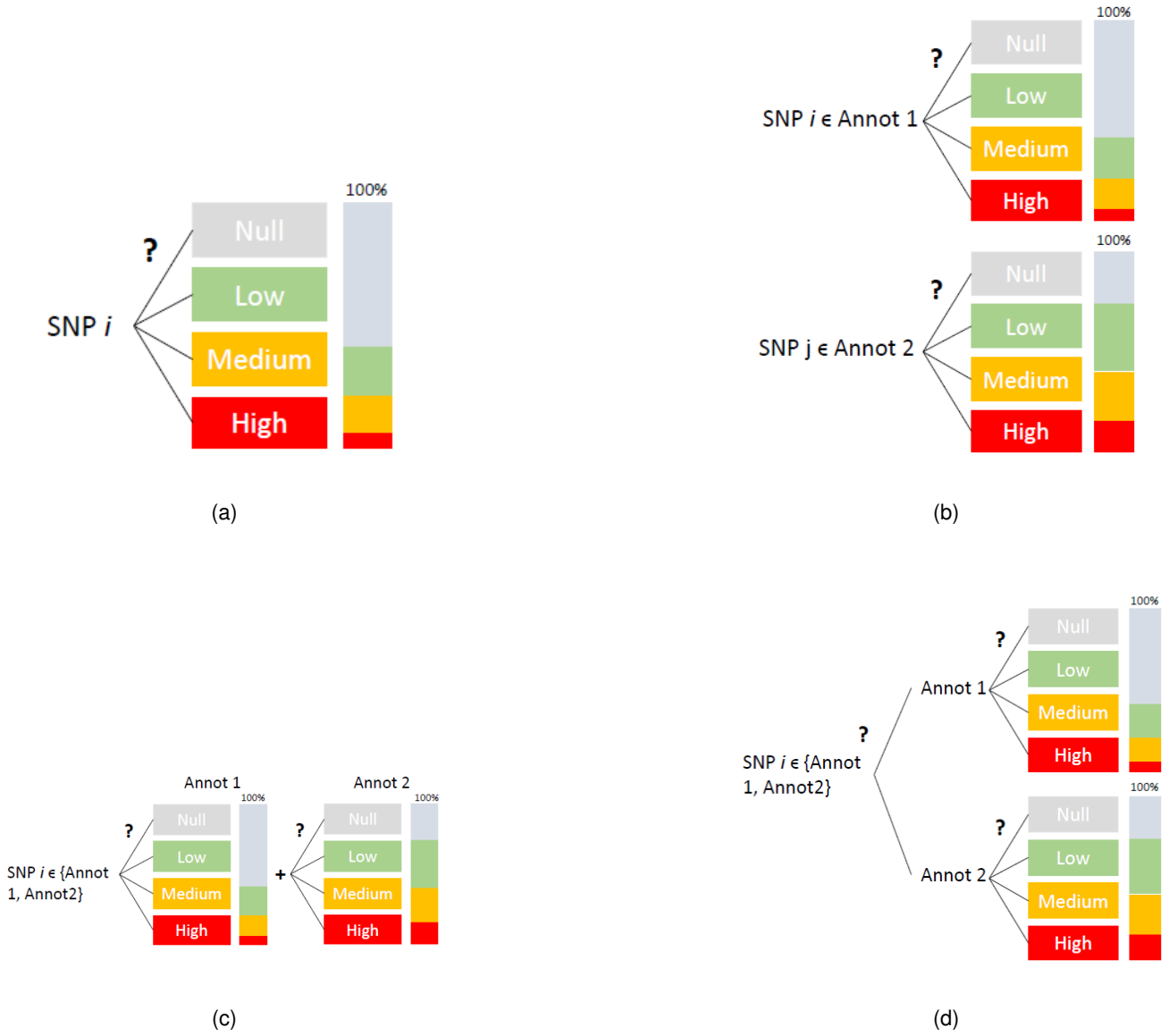


Figure 3.1: **Représentation de la distribution des effets des SNPs en fonction des modèles.** Schéma simplifié de la distribution des effets de SNPs entre les quatre classes d'effet pour les modèles (a) BayesR, (b) BayesRC, (c) BayesRC+ et (d) BayesRCπ. La jauge à doite de chaque répartition correspond à la proportion de SNPs dans chacune des catégories ajustées par le modèle. Pour les méthodes d'intégration d'annotations fonctionnelles, seules deux annotations ont été représentées par simplification.

Chapter 4

Prédiction de l'expression génique spécifique à un tissu à l'aide des SNPs d'un unique chromosome et d'annotations fonctionnelles

4.1 Introduction

Projet GENE-SWitCH Le projet EU-H2020 GENE-SWitCH (European Union's Horizon 2020 Research and Innovation Programme, accord de subvention n° 817998) cherche à générer de nouvelles annotations fonctionnelles chez le cochon et le poulet, dans le but de pouvoir les exploiter dans le secteur agricole. Le projet est actif dans le consortium *Functional Annotation of Animal Genomes* (FAANG), et est composé de trois grands piliers : l'identification et la caractérisation d'éléments génomiques fonctionnels, l'implémentation d'innovations

FAANG pour la production animale et la standardisation des données, des processus et leur diffusion auprès des acteurs agraires, politiques et sociaux. C'est dans ce cadre de coopération européenne que s'inscrit cette thèse, et une partie de son financement. Dans ce projet, de nombreuses données omiques fonctionnelles ont été générées (Figure 4.1), y compris des données ATAC-seq, ChIP-seq, RNA-seq, de méthylation et Hi-C. Pour ces analyses, jusqu'à sept tissus ont été exploités, et ce à plusieurs stades de développement des porcelets et des poulets.

Assays		**Annotations**	**Number of assays**	**Number of samples**				
				Early organogenesis 7 tissues * 2 species * 4 replicates (WP1)	**Late organogenesis** 7 tissues * 2 species * 4 replicates (WP1) 2 tissues * 72 replicates (WP5)	**Newborn/ hatched** 7 tissues * 2 species * 4 replicates (WP1)	**Weaned piglets** 2 tissues * 72 replicates (WP5)	**Fattening pigs** 3 tissues * 300 replicates (WP4)
				Pillar 1	**Pillar 1&2**	**Pillar 1**	**Pillar 1&2**	**Pillar2**
ATAC-seq	ATAC-seq	Accessible chromatin regions.	168 + 288	56	56 + 144	56	144	
ChIP-seq	H3K4me3	Promoters, enhancers, repressors.	168	56	56	56		
	H3K4me1		168	56	56	56		
	H3K27me3		168	56	56	56		
	H3K27Ac	Chromatin domain boundaries.	168	56	56	56		
	CTCF		168	56	56	56		
RNA-seq	mRNA-seq	Characterisation of transcripts structure (start and stop sites) and categories.	168	56	56	56		
	small RNA-seq		168	56	56	56		
	lrRNA-seq		42	14 (1 sample)	14 (1 sample)	14 (1 sample)		
	3' Quantseq	Transcripts expression levels.	900 + 288		144		144	900
DNA methylation	RRBS	Active and inactive promoters	126	42 (3 replicates)	42 (3 replicates)	42 (3 replicates)		
	WGBS	Active and inactive chromatin	42	14 (1 sample)	14 (1 sample)	14 (1 sample)		
Hi-C	Capture Hi-C	Structural domain annotations Enhancer/promoter interactions	24	8 (2 tissues, 2 species, 2 replicates)	8 (2 tissues, 2 species, 2 replicates)	8 (2 tissues, 2 species, 2 replicates)		
WGS	Whole Genome Sequencing	Genetic variants	300					300
TOTAL	**14 different assays**		**3354 molecular assays**	**Pillar 1 : Deep molecular characterisation of 168 tissues Pillar 2 : Specific molecular characterisation of 288 and 900 samples representing respectively 2 tissues (skeletal muscle and liver) and 3 tissues (skeletal muscle, ileum, liver)**				

Core FAANG assays

Figure 4.1: **Données générées dans le cadre du projet GENE-SWitCH**

Contexte de l'étude Dans la filière porcine, les caractères liés à la santé animale sont de plus en plus importants à prendre en compte dans les objectifs de sélection, afin de répondre aux problématiques de résistance aux antibiotiques ainsi qu'aux attentes de la société sur le bien-être animal (Merks et al., 2012). Dans ce contexte, certains gènes semblent particulièrement pertinents à utiliser comme phénotype intermédiaire, i.e un caractère

quantitatif fiable et suffisamment héritable positionné entre les variants génétiques et une maladie (Flint et al., 2014; Preston and Weinberger, 2022).
L'expression de gène représente un phénotype intermédiaire et héritable, entre le génome et un caractère observé. Elle est impactée par des mécanismes de régulation, qui ont des propriétés d'inhibition, de modulation, ou de promotion; on peut par exemple citer les promoteurs, les amplificateurs ou les inactivateurs, situés majoritairement en régions non codantes, et dans de plus rares cas dans des régions codantes. Certaines séquences de l'ADN possèdent ces caractéristiques de régulation, soit en modulant directement l'expression du gène (facteur *cis*, généralement à proximité du gène) ou indirectement en agissant sur un facteur *cis* (facteur *trans*, plus éloignées du gène et potentiellement sur un autre chromosome) (Wittkopp et al., 2004). Ces séquences régulatrices peuvent être impactée par des mutations, impactant par conséquent l'expression génique.

Identifier ces séquences régulatrices et les variants à fort effet parmi elles peut être effectué à l'aide de méthodes d'association (*genome wide association studies*; GWAS), mais certaines restent difficile à détecter, en raison d'une puissance insuffisante ou de la rareté de ces variants. L'utilisation de données WGS offre la possibilité de la découverte de ces variants rares (Cirulli and Goldstein, 2010). Une meilleure identification des variants rares via l'exploitation de données WGS pourrait induire une amélioration de la qualité de prédiction. Cependant, les gains constatés jusqu'à présent restent modestes. L'exploitation d'annotations fonctionnelles, en priorisant certaines régions du génome, pourrait permettre d'améliorer la précision des modèles de prédiction. La prédiction de l'expression du gène à partir de données WGS et l'intégration des informations sur les mécanismes régulateurs pourrait donc révéler des informations sur les différents processus de régulation (Avsec et al., 2021), et ainsi aider à mieux comprendre les processus biologiques derrière des caractères complexes.

Peu de modèles de prédiction génomique sont aujourd'hui capables d'utiliser des annotations fonctionnelles externes, qui ne nécessitent pas d'avoir été prélevées sur les mêmes individus. En partitionnant les SNPs en fonction des annotations, BayesRC (MacLeod et al., 2016) permet une intégration flexible et directe d'informations régulatrices pour la prédiction de phénotypes. Cependant, en contraignant les SNPs à n'appartenir qu'à une classe d'annotation fonctionnelle, elle se heurte à la complexité des annotations dans de vastes études telles que GENE-SWitCH, où plusieurs tissus, temporalités et technologies sont exploités. Les méthodes BayesRCO (Mollandin et al., 2022) ont été proposées pour dépasser cette limitation d'unicité d'annotation par SNP, en introduisant deux modèles bayésiens, BayesRCπ et BayesRC+, un modèle de mélange de mélanges et un modèle cumulatif de mélanges. Ces modèles permettent à la fois de prédire des caractères complexes, d'identifier des mécanismes génétiques, et d'apporter de nouvelles informations sur le lien entre les annotations et le phénotype.

L'intégration d'annotations fonctionnelles dans les modèles de prédiction génomique n'a cependant pas montré

une amélioration constante de la qualité et précision de prédiction. Le cadre d'une bonne utilisation de ces informations biologiques est pour l'instant vague, et il est nécessaire de parvenir à identifier quels jeux d'annotations sont adaptés pour chaque trait. Plusieurs scénarios sont alors envisageables pour évaluer le potentiel de BayesRCπ pour la prédiction d'expression de gènes. L'expression des gènes revêt un caractère temporel, et l'utilisation d'informations sur la régulation à un temps proche pourrait être adaptée, ainsi que des informations sur le dynamisme de ces mécanismes de régulation (*switches*). Ici, nous nous focalisons sur des annotations issues de technologie d'accessibilité à la chromatine et de quantification de la méthylation, toutes deux ayant un impact direct sur la possibilité de transcription d'un gène et représentant ainsi de bons candidats pour la construction des annotations. Le choix des animaux utilisés pour l'apprentissage des modèles, en fonction de ceux dont on veut prédire le phénotype, a aussi un effet non négligeable sur la qualité des résultats de prédiction. Cependant, il n'est pas toujours possible d'avoir l'effectif et la structure familiale optimale, particulièrement pour les races de petit effectif. Des stratégies de validation entre races, ou en les mélangeant, peuvent être envisagées. Dans cette optique, l'utilisation d'annotations fonctionnelles pourrait être une plus-value pour la prédiction génomique, en permettant de prioriser des mécanismes cellulaires communs entre les races.

Dans cette étude, nous appliquons BayesRCπ pour la prédiction de l'expression d'un sous-ensemble de gènes ciblés, à partir de données de génotypage (WGS) par chromosome et de régulation (méthylation d'ADN et ATAC-seq) générées dans le cadre du projet GENE-SWitCH. Nous utilisons des annotations prélevées à trois différents stades de développement, et dans des tissus identiques à ceux utilisés pour mesurer l'expression des gènes. Deux stratégies de validation sont utilisées, la validation inter-race et la validation avec toutes les races regroupées. Nous explorons alors la qualité de prédiction, la priorisation des SNPs et l'utilisation des annotations par le modèle.

4.2 Matériels et méthodes

4.2.1 Plan d'étude

Des échantillons provenant de trois tissus différents (duodénum, foie et muscle) ont été prélevés à l'abattage chez n=300 porcs de 7 mois ($\sim$ 208.5 jours) de trois races différentes (Duroc, DU ; Landrace, LD ; Large White, LW; (Crespo-Piazuelo, 2022)), avec n=100 animaux par race. Dans cette étude, nous nous concentrons en particulier sur les données provenant du foie et du muscle. L'ADN génomique a été extrait du sang (DU, LD) ou du foie (LW), et l'ARN a été extrait du duodénum, du foie et du muscle. L'ADN et l'ARN ont ensuite été séquencés en paires (2 x 150 pb) à l'aide de la plateforme Illumina NovaSeq6000. Dans cette étude, seuls les échantillons issus du muscle et du foie ont été utilisés.

4.2.2 Données WGS

Les séquences d'ADN du génome entier ont été cartographiés sur l'assemblage du génome de référence *Sscrofa11.1* avec BWA-MEM (Li, 2012). Après avoir filtré les variants en retirant ceux dont la proportion de valeurs manquantes (*missing call rate*) était inférieure à 10% et la fréquence des allèles mineurs (MAF, *minor allele frequency*) inférieure à 5% sur l'ensemble des n=300 animaux, un total de 25,315,878 variants génétiques a été obtenu à l'aide de GATK (McKenna et al., 2010). Dans un deuxième filtrage, d'autres étapes de prétraitement des données ont été réalisées à l'aide de PLINK (Chang et al., 2015) afin d'éliminer les variants qui comportaient des valeurs manquantes et d'effectuer un filtrage sur le LD (taille de la fenêtre = 100 kb, taille du pas = 50, r^2 = 99%) pour éliminer les variants redondants. Enfin, les variants ont été séparés par chromosome.

4.2.3 Données transcriptomiques

Les séquences ARN ont été cartographiées de manière similaire à l'assemblage du génome de référence *Sscrofa11.1* à l'aide de STAR (Dobin et al., 2013). L'expression des gènes a été quantifiée et normalisée en comptant les lectures (*reads*) alignées sur les gènes à l'aide de RSEM (Li and Dewey, 2011), ce qui a permis d'obtenir des valeurs logarithmiques de comptage par million (CPM) pour 15,710 et 13,887 gènes dans le foie et le muscle, respectivement, après filtrage des gènes à faible expression.

4.2.4 Choix des gènes

Dans cette étude, l'objectif était de se concentrer sur les prédictions génomiques de l'expression d'un sous-ensemble de gènes d'intérêt ; nous avons cherché en particulier à identifier les gènes qui étaient de bons candidats à être régulés par des variants génétiques ou épigénétiques. Une étude antérieure d'association pangénomique avec l'expression génique (eGWAS ; Crespo-Piazuelo (2022)) a identifié des polymorphismes significativement associés (p-valeurs corrigées selon la procédure de Bonferroni, seuil de significativité à 5%) à l'expression des gènes, qui ont ensuite été regroupés en régions eQTL (*expression quantitative trait loci*). Les régions *cis*-eQTL ont été définies comme étant celles trouvées dans une fenêtre de $\pm$1Mb autour de leur gène respectif, et les autres régions eQTL ont été désignées comme étant des *trans*-eQTL, ces derniers pouvant appartenir ou non au même chromosome que le gène. Sur la base de ces analyses, nous avons identifié un ensemble de 8 gènes exprimés pour lesquels les meilleurs polymorphismes dans leurs régions cis-eQTL respectives étaient partagés dans les trois tissus (duodénum, foie, muscle) : CELF2, DET1, HUS1, L3HYPDH, NUDT22, R3HCC1, SLA-7 et SUPT3H. Nous avons également exploités 3 gènes supplémentaires, qui ont été identifiés comme ayant des signaux *cis* dans le eGWAS et dont de précédentes études les avaient identifiés comme étant méthylés : IGF2, LEPR et PRKAG1. Cet ensemble de 11 gènes a donc été utilisé dans nos analyses ultérieures ; une description et un résumé des résultats de l'eGWAS pour ces gènes sont présentés dans le tableau 4.1.

Ensembl ID	Gene	Chr	Pos	*Cis* (liver)	*Trans* (liver)	h^2 (liver)	*Cis* (muscle)	*Trans* (muscle)	h^2 (muscle)
ENSSSCG00000005103	DET1	1	191,263,447-191,284,968	3	124+1	0.450834	3	177+4	0,575891
ENSSSCG00000037854	L3HYPDH	1	188,584,377-188,632,585	811	1798+20	0.835491	1274	4404+224	0.996073
ENSSSCG00000013039	NUDT22	2	7,880,258-7,883,945	217	0+6	0.871961	295	27+3	0.542974
ENSSSCG00000035293	IGF2	2	1,469,132-1,496,442	18	0+1	0.257796	354	17+79	0.978421
ENSSSCG00000000185	PRKAG1	5	15,033,053-15,049,660	488	0+5	0.348881	0	2+2	0.555389
ENSSSCG00000025188	LEPR	6	146,798,979-146,896,108	342	2+16	0.619324	—	—	—
ENSSSCG00000001398	SLA-7	7	23,634,639-23,649,314	9	53+1	0.362549	13	44+1	0.49011
ENSSSCG00000001709	SUPT3H	7	39,751,927-40,161,114	25	0+0	0.579295	247	9+10	0.702241
ENSSSCG00000011121	CELF2	10	60,506,482-61,084,861	4	0+0	0.198011	5	12+6	0.311422
ENSSSCG00000039915	R3HCC1	14	7,405,792-7,434,935	1348	54+2	0.426799	2947	42+7	0.498386
ENSSSCG00000028523	HUS1	18	48,522,997-48,538,788	0	0+0	0.17501	7443	70+1	0.550025

Table 4.1: **Description des gènes utilisés dans l'étude.**
Le tableau comprend l'identifiant Ensembl, le nom du gène, le chromosome et la position, le nombre d'eQTLs *cis* et *trans* dans le foie et le muscle, et l'héritabilité estimée dans le foie dans le muscle. Les eQTLs cis sont définis comme des variants situés à $\pm$1 Mb du gène. Les eQTL trans sont définis comme la somme des variants situés à plus de $\pm$1 Mb sur le même chromosome que le gène et ceux situés sur d'autres chromosomes. Comme LEPR a une expression spécifique au foie, il n'a aucune valeur estimée dans le muscle.

4.2.5 Données d'annotations fonctionnelles

Le projet GENE-SWitCH (Acloque, 2022) a produit de nombreuses informations génomiques fonctionnelles, dans une variété de tissus différents, et au cours des premiers stades de développement du porc et du poulet (Figure 4.1). En particulier, des échantillons post-mortem ont été prélevés sur des embryons de porc au début de l'organogenèse (30 jours après la fécondation ; dpf) et à la fin de l'organogenèse (70 dpf) ainsi que sur des porcelets nouveau-nés (NB) LW, avec 2 mâles et 2 femelles à chaque stade. Notez que ces animaux sont complètement indépendants de ceux utilisés pour le séquençage du génome entier et les séquençage transcriptomique décrits ci-dessus. Dans ce travail, nous nous concentrons sur les annotations générées à chacun des trois stades de développement dans nos deux tissus d'intérêt (foie et muscle) pour deux types de données omiques fonctionnels différents :

- Profilage de la méthylation par séquençage bisulfite pangénomique (WGBS), où les régions génomiques sont classées comme étant non méthylées (UMR ; correspondant approximativement aux promoteurs), faiblement méthylées (LMR ; correspondant aux amplificateurs putatifs), totalement méthylées (FMR), ou non dans un tissu et un stade de développement donnés.

- Profilage de l'accessibilité de la chromatine par analyse de la chromatine accessible par transposase avec séquençage à haut débit (ATAC-seq), où les régions génomiques sont classées comme étant accessibles (représentant potentiellement des exhausteurs, des promoteurs, des isolateurs ou des sites de début de transcription) ou non dans un tissu et un stade de développement donnés.

Trois stratégies différentes ont été envisagées pour catégoriser les variants situés dans une ou plusieurs régions annotées dans un tissu donné : (1) annotations ATAC-seq seules (3 catégories, représentant les trois stades de

développement) ; (2) annotations de méthylation (9 catégories, représentant les régions UMR, LMR ou FMR x 3 stades de développement) ; (3) annotations ATAC-seq et de méthylation ensemble (12 catégories, représentant accessible, UMR, LMR, FMR x 3 stades de développement). Tous les variants non annotés ont été affectés à une catégorie finale “autre".

4.2.6 Modèles

Un modèle linéaire général pour la prédiction génomique peut être défini tel que

$$\begin{aligned} \mathbf{y} &= \mu \mathbb{1}_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \\ \mathbf{e} &\sim N(0, \mathbf{I}_n \sigma_e^2), \end{aligned} \tag{4.1}$$

où $\mathbf{y}$ est le vecteur des valeurs logarithmiques du CPM pour un gène d'intérêt donné, μ l'intercept, $\mathbf{X}$ la matrice des marqueurs centrée réduite, $\boldsymbol{\beta}$ le vecteur des effets des SNPs, $\mathbf{e}$ le vecteur des résidus, et $\mathbf{e}$ suit une distribution normale multivariée centrée et de matrice de variance-covariance $\mathbf{I}_n \sigma_e^2$. Dans ce travail, nous cherchons à comparer (1) les prédictions génomiques de l'expression des gènes obtenues en utilisant uniquement les variants génétiques spécifiques des chromosomes à (2) celles obtenues en incluant en plus des catégorisations préalables des variants obtenues à partir de cartes d'annotation fonctionnelle. À cette fin, nous faisons appel à deux modèles apparentés : BayesR (pour le premier point) et BayesRCπ (pour le second point). Tous deux utilisent *a priori* des distributions de mélange à quatre composantes pour $\boldsymbol{\beta}$ afin de modéliser les variants ayant des effets nuls, faibles, moyens ou importants ; BayesRCπ introduit en outre un *prior* de mélange de mélanges afin de discriminer les annotations des variants ayant plusieurs annotations (comme c'est le cas dans les catégorisations d'annotation fonctionnelle décrites dans la section précédente). Des détails supplémentaires sur les deux modèles peuvent être trouvés dans Mollandin et al. (2022); les deux modèles sont mis en œuvre dans le logiciel BayesRCO (https://github.com/fmollandin/BayesRCO).

4.2.7 Stratégie de validation

L'ensemble entier des données des n=300 porcs a été divisé en sous-ensembles d'apprentissage et de validation pour estimer les paramètres du modèle et évaluer la qualité de la prédiction, respectivement. Étant donné la structure des échantillons de porcs en races distinctes, nous avons envisagé deux stratégies différentes pour définir les sous-ensembles d'apprentissage et de validation : (1) prédictions *toutes-races*, où une approche de validation croisée à partir de 10 échantillons (*10-fold cross-validation*) a été utilisée pour créer les sous-ensembles d'apprentissage et de validation de n=270 et n=30 animaux, répartis de manière équilibrée entre les trois races ; et (2) prédictions *inter-races*, où chaque paire de races (DU+LD ; DU+LW ; LD+LW) a été utilisée

comme un ensemble d'entraînement de n=200 animaux, et la race restante comme un ensemble de validation (LW ; LD ; DU) de n=100. La précision de la prédiction pour les modèles utilisant différentes stratégies d'annotation fonctionnelle (aucune, atacseq, méthylation, atacseq+méthylation) a été évaluée en utilisant la corrélation de Pearson entre les valeurs d'expression génique réelles (y) et estimées () dans l'ensemble de validation.

4.3 Résultats

4.3.1 Distribution des annotations fonctionnelles sur le génome.

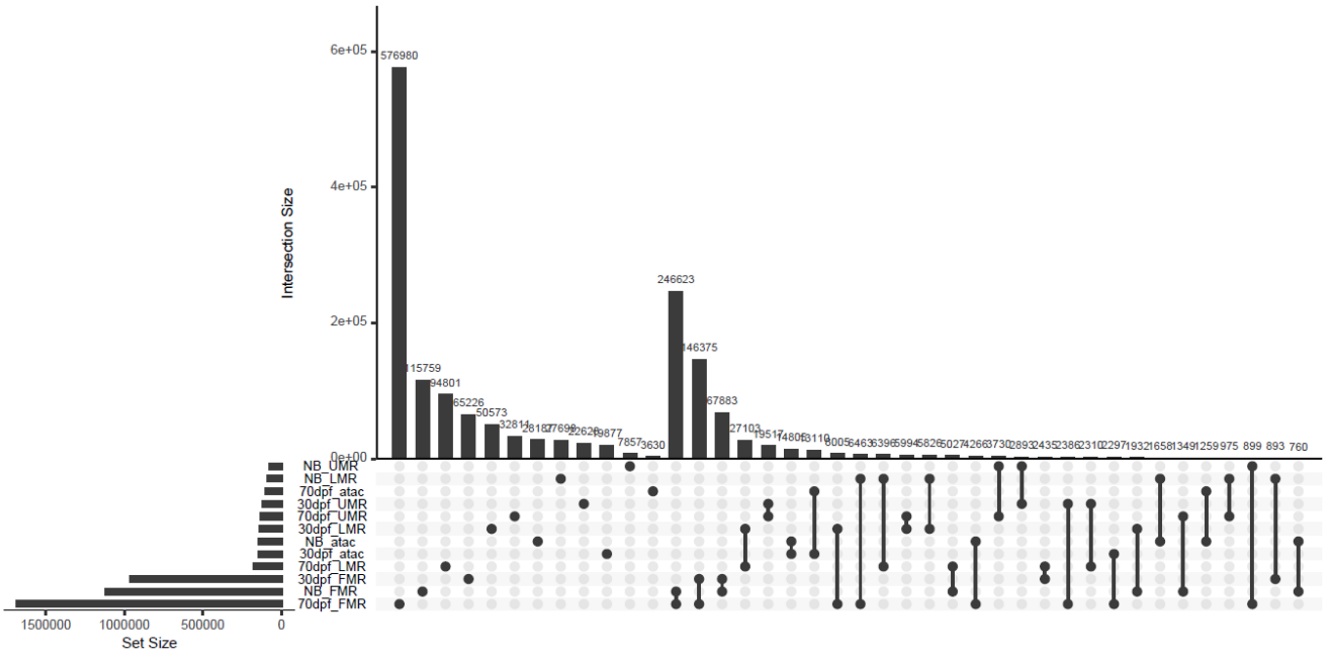


Figure 4.2: **Représentation graphique UpsetR des intersections de catégories d'annotations pour les chromosomes 1, 2, 5, 6, 7, 10, 14 et 18 (concaténés) du foie.**
Les catégories d'annotation correspondent aux variants se trouvant dans les régions de chromatine accessible (atac), non méthylées (UMR), faiblement méthylées (LMR) ou totalement méthylées (FMR) à trois stades de développement : 70 jours après la fécondation (dpf), 30 dpf, ou chez les porcelets nouveau-nés (NB). Seuls les 40 ensembles et intersections les plus grands sont représentés.

Au maximum, 12 annotations ont été utilisées pour la prédiction de l'expression de gènes avec BayesRCπ. Dans ce cas, on s'attend à une augmentation du nombre de chevauchements entre les annotations, qui justifient l'utilisation de BayesRCπ. Les effectifs du nombre de SNPs annotés pour chaque combinaison d'annotations dans le foie sont représentés Figure 4.2, pour les 8 chromosomes utilisés pour la prédiction (note : le graphique représente ici les 40 combinaisons majoritaires). Les annotations FMR sont largement les plus représentées avec 1,680,095, 1,119,483, 959,105 pour les stades de développement 70dpf, NB et 30dpf respectivement. Les annotations restantes ont des effectifs de taille plus modeste, de 74525 à 173414 SNPs. Sans surprise en raison de leur grande taille, les combinaisons d'annotations ayant le plus fort effectif sont là encore pour les annotations

FMR, avec les combinaisons 70dpf + NB, 70dpr + 30dpf et NB + 30dpf. Ensuite viennent d'autres combinaisons de 2 annotations, aucune combinaison plus complexe n'est assez suffisamment fréquente pour avoir été présentée sur ce graphique. Parmi les plus fréquentes, nous observons un appariement des annotations liées à la méthylation (UMR, LMR ou FMR), et un appariement de celles liées aux données ATACseq. Les chevauchements entre méthylation et ATACseq sont plus rares. Peu de SNPs sont annotés identiquement aux 3 stades de développement, suggérant des évolutions de la méthylation ou de l'accessibilité à la chromatine en fonction du temps. Cette configuration d'annotations est relativement homogène d'un chromosome à l'autre. Entre 34,1% et 46,5% des SNPs sont annotés dans le foie pour chaque chromosome exploité, on retrouve des résultats proches dans le muscle (Figure 4.3). Cette proportion d'annotations relativement élevée est en partie due aux annotations FMR.

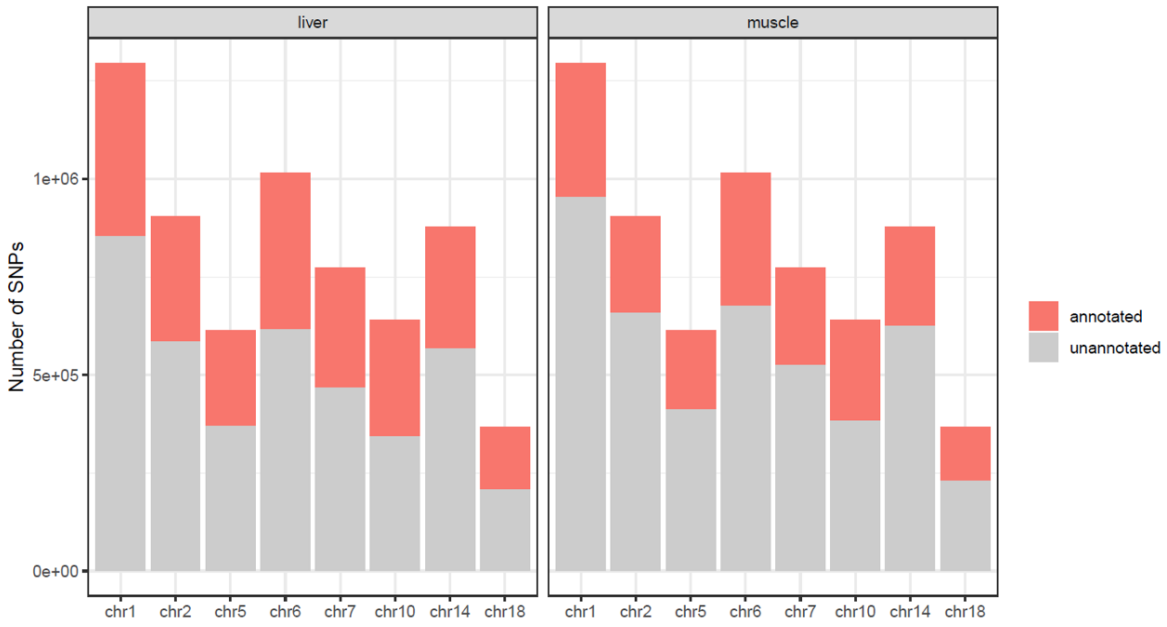


Figure 4.3: **Proportion de SNPs annotés par chromosome et par tissu.**
Nombre total de variants génétiques annotés (en rouge ; régions d'accessibilité à la chromatine chromatine ou de niveaux de méthylation spécifiques dans les trois stades de développement) et non annotés (en gris) dans chaque chromosome par tissu.

4.3.2 Impact des annotations fonctionnelles pour la prédiction pour toutes-races

La qualité de prédiction de l'expression de gènes à partir du séquençage du chromosome associé à chaque gène, sans annotations fonctionnelles, et toute-races, est très variable d'une configuration à l'autre. On observe Figure 4.4, qu'en moyenne (sur 10-folds), avec BayesR, certaines corrélations sont quasiment nulles, notamment PRKAG1 et DET1 dans le foie, avec une corrélation moyenne respective de 0.01 et -0.08. HUS1 a les meilleurs résultats de corrélation dans le foie et dans le muscle, respectivement de 0.59 et 0.58. Certains gènes ont des résultats bien plus différents entre les différents tissus, PRKAG1 dont la prédiction est quasi-nulle dans le foie a une prédiction bien plus honorable dans le muscle, avec une moyenne de 0.46. A l'inverse, la prédiction de

l'expression du gène NUDT22 est meilleure dans le foie que dans le muscle, avec 0.5 contre 0.1.

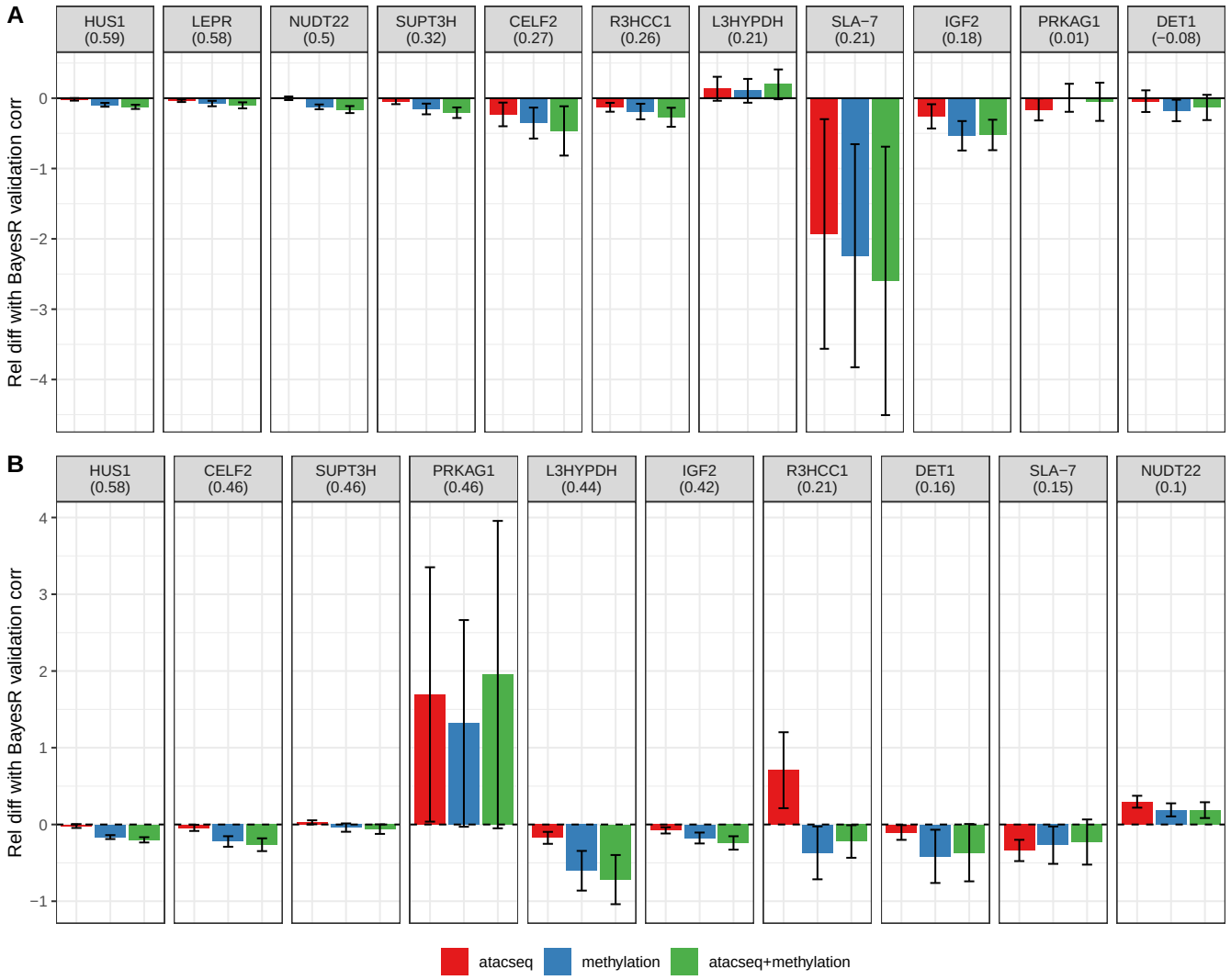


Figure 4.4: **Qualité de prédiction de la prédiction toute-races intégrant des annotations fonctionnelles.** Différence relative de corrélation de validation par rapport à BayesR (moyenne sur 10 *folds*, avec barres d'erreur standard) pour les prédictions toutes races confondues utilisant BayesRCπ et les annotations ATAC-seq, méthylation, ou ATAC-seq + méthylation. Les résultats sont présentés pour chacun des 11 gènes considérés (colonnes) dans deux tissus (foie et muscle dans les panneaux A et B, respectivement). Les gènes sont classés de gauche à droite en fonction des valeurs de corrélation de validation moyennes décroissantes pour BayesR (indiquées entre parenthèses dans les titres des facettes).

Dans l'ensemble, l'utilisation d'annotations fonctionnelles ne semble pas aider à la prédiction de l'expression des gènes, la majorité des situations gènes/tissu/annotations présente des résultats inférieurs ou sensiblement identiques à ceux issus de BayesR. Dans les cas les plus extrêmes, on observe une différence relative de -2,6 pour le gène SLA-7 dans le foie avec les données de *atac+méthylation*, les autres annotations étant de même défavorable à la prédiction, avec -1,9 et -2,2 pour les annotations *atac* et *méthylation*. Une situation montre cependant une amélioration notable de la qualité de prédiction à l'aide d'annotations fonctionnelles, PRKAG dans le muscle. La différence relative obtenue avec les données de méthylation et d'accessibilité à la chromatine atteint près de 2 en

moyenne. Là encore, on observe une tendance similaire avec les deux autres types d'annotations. En comparant les résultats entre les annotations pour toutes les configurations, on observe la plupart du temps les mêmes tendances (gain ou perte), mis à part pour le gène R3HCC1 dans le muscle, pour qui seules les annotations ATAC-seq semblent lui conférer un gain de prédiction, quand les deux autres annotations apportent une perte de prédiction.

4.3.3 Impact des annotations fonctionnes pour la prédiction inter-races

Le type de validation utilisé impacte substantiellement la qualité de prédiction. Avec BayesR, on observe déjà des différences de corrélations moyennes notables entre les valeurs présentées Figure 4.4 et Figure 4.5. Sur 11 gènes exprimés dans le foie, 8 sont moins bien prédits dans la stratégie de validation inter-races que toutes-races. Similairement, sur les 10 gènes exprimés dans le muscle, 7 sont moins bien prédits dans la stratégie de validation inter-races. Certaines différences sont non négligeables. LEPR par exemple, bien prédit dans le cadre toutes-races (0.58 en moyenne), chute à 0 dans le cadre inter-races. La prédiction moyenne (sur les 3 races de validation) est pour certains gènes et tissus quasiment nulle, ou faible. 3 gènes ont les meilleurs résultats de prédiction à la fois pour le foie et le muscle, HUS1, NUDT22 et R3HCC1, avec un avantage dans le cadre du muscle.

L'apport des annotations fonctionnelles dépend beaucoup du set de validation. La prédiction de l'expression de SUPT3H semble fortement influencée par l'introduction d'annotations fonctionnelles, en moyenne positivement dans le foie pour les 3 types d'annotations, et positivement dans le muscle pour les annotations *atac* uniquement. Néanmoins, les résultats sont très variables pour ce gène, on observe une différence relative jusqu'à 11 pour la prédiction de la race DU dans le foie avec les annotations *méthylation*. Le gain pour les autres races reste modéré, mais positif. Pour ce même gène dans le muscle, la validation sur chacune des races apportent une évolution de la prédiction différente: la validation sur les porcs DU voit une différence relative de 48 avec les données ATACseq, mais descend de 13 pour les mêmes annotations mais sur une validation sur les porcs LW.

Dans l'ensemble, l'expression des gènes ayant été bien prédits par BayesR sont peu impactés par l'intégration d'annotations fonctionnelles. Les différences de résultats entre annotations semblent plus notables que dans la prédiction toutes-races, avec la présence de cas où l'ajout d'un type d'annotations ou l'autre a des effets contraires sur la qualité de prédiction. Par exemple dans le foie, les gènes SLA-7 et LEPR voient leur prédiction s'améliorer pour les porcs Duroc avec les annotations *méthylation* et *atac+méthylation*, mais aussi se détériorer pour les annotations *atac*. Toujours dans le foie et en validation Duroc, on observe tout à fait l'inverse pour le gène PRKAG. Par ailleurs, le set de validation DU semble avoir des résultats de prédiction bien plus impactés que les autres races face à l'ajout d'annotations fonctionnelles.

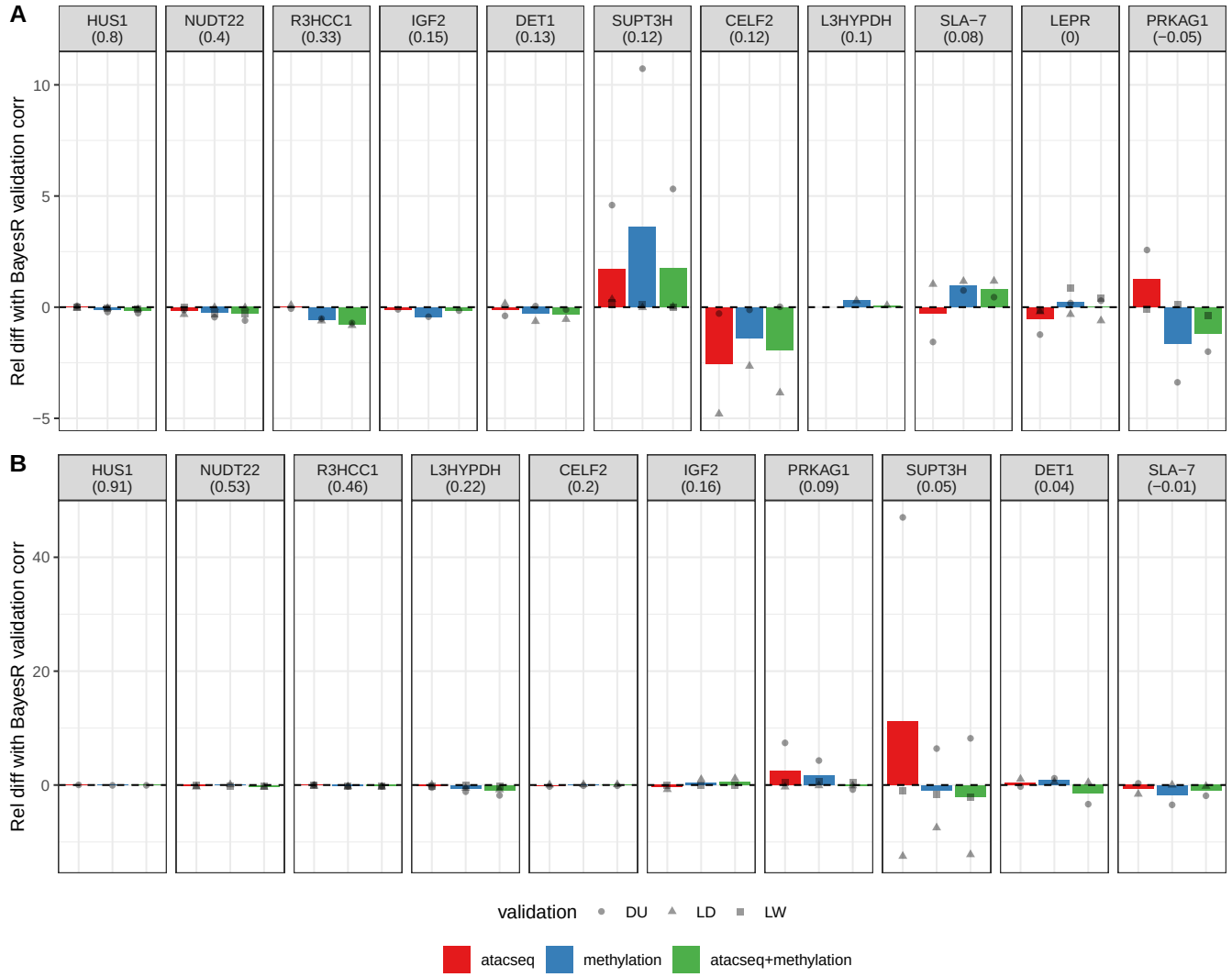


Figure 4.5: **Qualité de prédiction de la prédiction inter-races intégrant des annotations fonctionnelles.** Différence relative de corrélation de validation par rapport à BayesR (moyenne sur les trois races) pour les prédictions inter-races utilisant BayesRCπ et les annotations ATAC-seq, méthylation, ou ATAC-seq + méthylation. Les résultats sont présentés pour chacun des 11 gènes considérés (colonnes) dans deux tissus (foie et muscle dans les panneaux A et B, respectivement). Les gènes sont classés de gauche à droite en fonction des valeurs de corrélation de validation moyennes décroissantes pour BayesR (indiquées entre parenthèses dans les titres des facettes).

4.3.4 Estimation de l'effet des SNPs pour la prédiction de l'expression de SUPT3H dans le foie à l'aide d'annotations fonctionnelles complexes

SUPT3H est un gène identifié comme étant régulé par des processus de méthylation. S'il présente des résultats de prédiction corrects pour le cas de la validation toutes-races dans le muscle et dans le foie, ce n'est pas le cas de la validation inter-races. En revanche, il semblerait que l'ajout d'annotations fonctionnelles pour prédire l'expression de SUPT3H, dans la configuration inter-races et foie, augmente la qualité de prédiction, en particulier avec les annotations *méthylation*. Dans le foie, 25 *cis*-eQTLs ont été identifiés, et aucun *trans*-eQTLs (Table 4.1), que ce soit

sur le même chromosome ou un autre. Ces 25 *cis*-eQTLs sont représentés en rouge Figure 4.6, qui représente la variance *a posteriori* des SNPs du chromosome 7 estimées dans les 3 découpages inter-races.

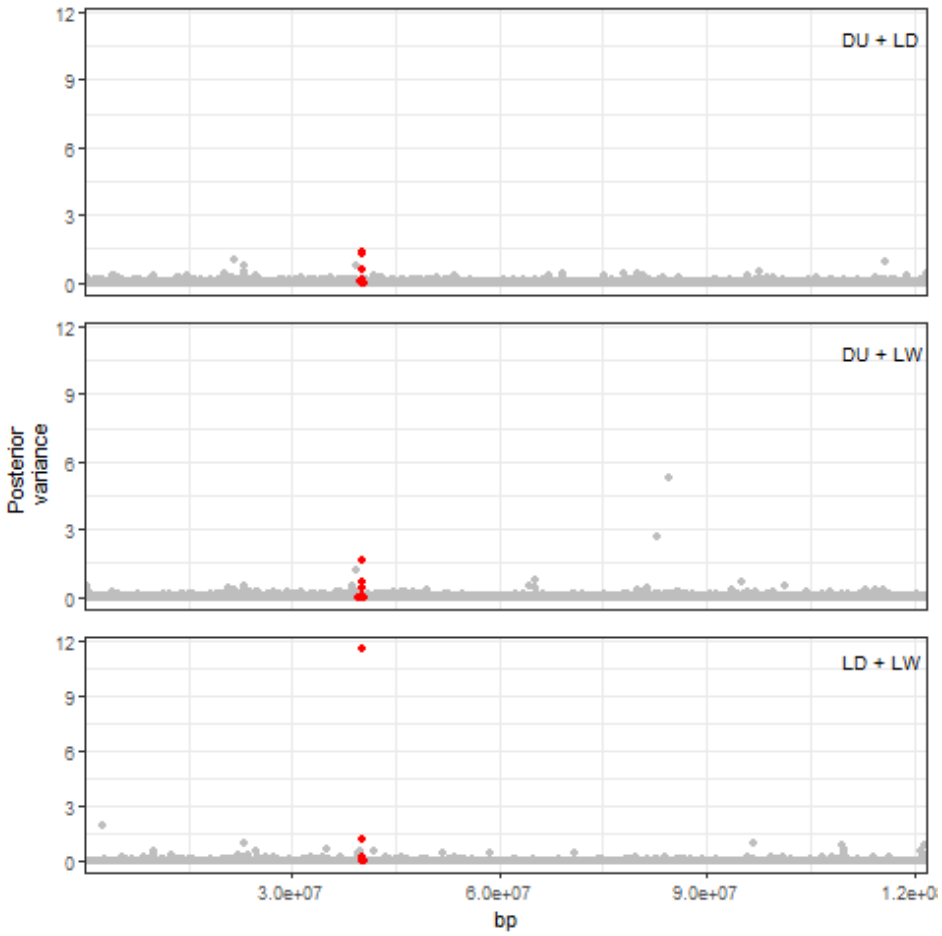


Figure 4.6: **Variance *a posteriori* des SNP sur le chromosome 7 pour la prédiction de l'expression de SUPT3H dans le foie pour trois populations d'apprentissage.**
DU=Duroc, LD= Landrace, LW=Large white En haut : DU + LD ; au milieu : DU + LW, en bas : LD + LW. Les variantes identifiées comme des eQTLs significatifs pour SUPT3H sont surlignées en rouge. Résultats obtenus à partir des données de méthylation.

Dans les 3 prédictions par races, une partie des eQTLs ont été priorisés par BayesRCπ. Dans le cas du set d'apprentissage DU+LD, 3 eQTLs précédemment identifiés (bp=40,153,808; bp=40,206,820; bp=40,161,091) sont placés respectivement en 1ère, 2ème et 7ème position (variance *a posteriori* décroissante). Dans le set DU+LW, on retrouve ces 3 eQTLs encore mis en avant, cette fois-ci en 7ème, 3ème et 14ème position respectivement. Enfin, dans le dernier set (LD+LW), le eQTL à la position 40,161,091 est estimé particulièrement haut, avec une variance *a posteriori* estimée à 11.5, contre 0.6 et 0.4 pour les deux autres jeux de données. D'autres SNPs en dehors de la région proche du gène sont aussi mis en avant, notamment les SNPs aux positions 82,886,488 et 84,685,744 du set DU+LW, aux variances *a posteriori* respectives de 5.3 et 2.7. Le SNP bp=82,886,488 est annoté dans les catégories *70dpf_FMR* et *NB_LMR* (PAIP 44.8% vs 55.2%), et le SNP bp=84,685,744 est annoté uniquement pour *NB_LMR*. Ces deux variants n'ont pour autant pas été priorisés par BayesRCπ dans les deux

autres jeux d'apprentissage. De même, le SNP bp=2,508,205, non annoté, est priorisé uniquement pour le jeu de données LD+LW. Certaines régions, non identifiées en tant que eQTLs, sont néanmoins estimées comme semblant avoir un effet sur l'expression du gène SUPT3H, notamment aux alentours des positions 21,000,000 bp - 25,000,000 bp.

Le gène SUPT3H s'étend de la position 39,751,927 bp à 40,161,114 bp (ENSEMBL, 2022). Parmi les 25 eQTLs détectés, 7 font partie du gène, les autres se trouvant à proximité. La Figure 4.7 propose une focalisation autour de ce gène, entre 39,500,000 bp et 40,300,000 bp. Une majorité des variants présents dans cette région sont FMR, tous stades confondus. Les SNPs annotés dans les régions UMR ou LMR sont en petit nombre, mais ils semblent concorder en partie avec les SNPs mis en avant par BayesRCπ. Toute une zone de 13 variants consécutifs entre 39,674,532 bp et 39,675,278 bp, annotés *70dpf_UMR* et *NB_LMR* constitue un petit pic pour les trois jeux de données. Enfin, parmi les 3 eQTLs discutés précédemment (bp=40,153,808; bp=40,161,091; bp=40,206,820), tous sont multi-annotés. Le eQTL bp=40,153,808, présent dans le gène, est annoté pour *30dpf_LMR* et *70dpf_LMR*, et les deux autres (bp=40,161,091 et 40,206,820) sont annotés pour *30dpf_UMR*, *70dpf_UMR* et *NB_UMR*, c'est-à-dire comme étant non méthylés aux trois stades de développement.

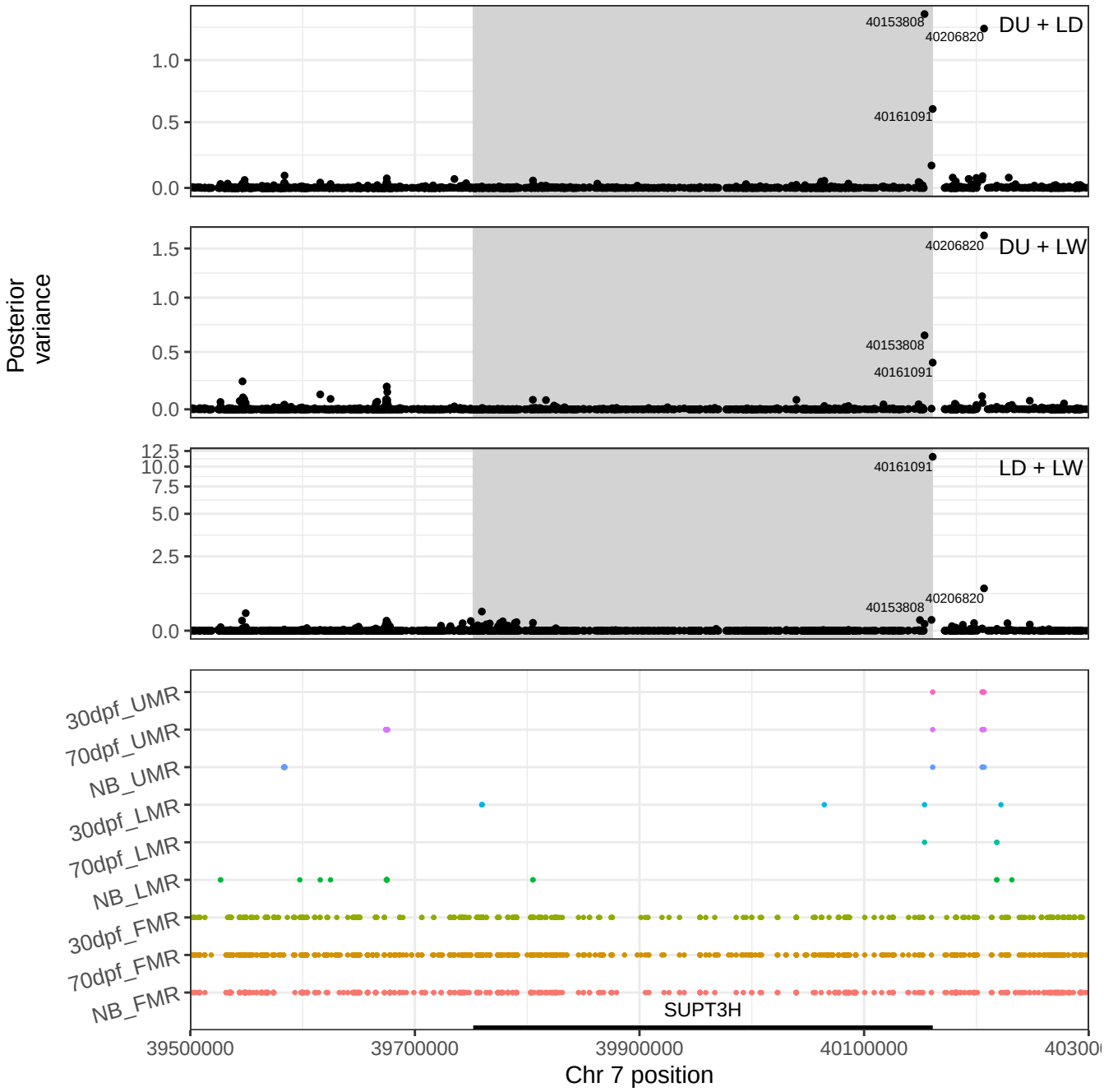


Figure 4.7: **Représentation des variances *a posteriori* pour chaque population d'apprentissage et annotations correspondantes dans le foie dans le voisinage (39,500,000 - 40,300,000) du gène SUPT3H sur le chromosome 7.**

Les catégories d'annotations correspondent aux variants tombant dans les régions non méthylées (UMR), faiblement méthylées (LMR) ou totalement méthylées (FMR) à trois stades de développement : 30 jours après la fécondation (dpf, *days post-fertilization*), 70 dpf, ou chez les porcelets nouveau-nés (NB). L'échelle des ordonnées à été mise à l'échelle *pseudo-log*.

4.3.5 Les régions génomiques non méthylées des porcelets nouveau-nés jouent le rôle le plus important dans la prédiction de l'expression de SUPT3H dans le foie

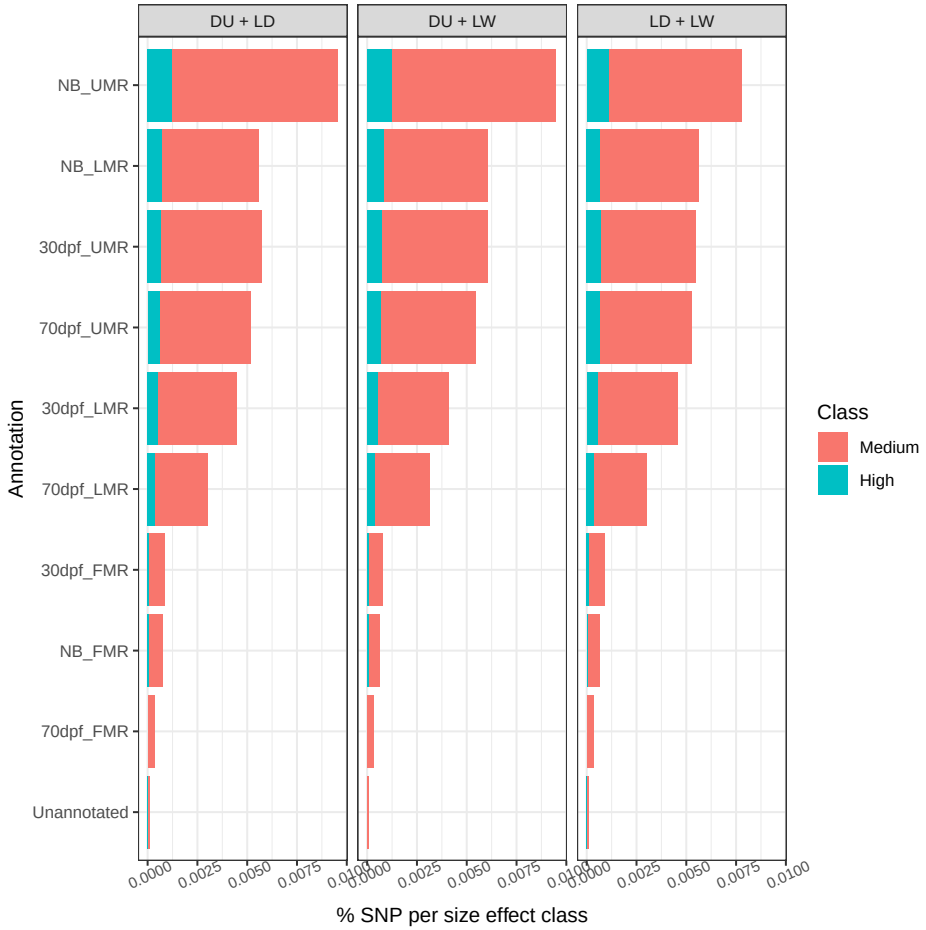


Figure 4.8: **Interprétation des annotations pour la prédiction du gène SUPT3H.**
Proportion de SNPs à effet fort (bleu) et moyen (rouge) du chromosome 7 assignés à chacune des catégories d'annotation par BayesRCπ pour l'expression de SUPT3H, dans le cadre de la prédiction entre races.

L'utilisation d'annotations fonctionnelles avec BayesRCπ n'a un impact que si les annotations ont un enrichissement de SNPs d'effet nul, faible, moyen et fort différent de celle des données entières. Si cela peut être difficile à évaluer précisément en amont, les *outputs* de BayesRCπ proposent des statistiques utiles pour explorer l'utilisation des annotations par le modèle. Dans le cas du gène SUPT3H dans le foie, l'amélioration de la qualité de prédiction avec les annotations fonctionnelles *méthylation* suggère une bonne exploitation de ces annotations par le modèle. On observe dans la Figure 4.8, une différence effective d'enrichissement entre les annotations et la catégorie des SNPs non annotés. Pour les 3 jeux de données, le classement des annotations par proportion d'enrichissement est similaire. L'annotation *NB_UMR* est à chaque fois la plus enrichie, entre 0.8% et 1% de SNPs à effet moyen ou fort. Après, se trouvent de façon assez équivalente *NB_LMR*, *30dpf_UMR* et *70dpf_UMR*, puis *30dpf_LMR* et *70dpf_LMR*. Les 3 annotations de méthylation totale *30dpf_FMR*, *NB_FMR*, *70dpf_FMR*, ont un

enrichissement moins important que les autres annotations de méthylation, mais celui-ci reste supérieur à la catégorie regroupant les SNPs non annotés. Il peut néanmoins avoir un effet de taille sur cette différence d'enrichissement, les annotations de méthylation étant bien plus grandes que les autres. Ce classement d'enrichissement suggère des tendances UMR > LMR > FMR, et dans une moindre mesure NB > 30dpf > 70dpf. La temporalité du prélèvement des données et la stratification du niveau de méthylation semble donc apporter des informations différentes au modèle.

4.4 Discussion

Nous avons utilisé des annotations fonctionnelles en plus des données de séquences ADN pour prédire un phénotype intermédiaire; l'expression de gène. Simultanément, nous avons cherché à identifier les variants impliqués dans la variation de l'expression génique. Nous avons pour cela exploité des données génomiques porcines, en considérant uniquement le chromosome sur lequel est positionné le gène dont nous souhaitons prédire l'expression. Le choix d'utiliser ce chromosome uniquement repose sur deux arguments (1) une majeure partie des autres chromosomes devrait avoir un impact très limité sur l'expression du gène, (2) des limites computationnelles peuvent résulter de l'utilisation du génome entier. Cependant, les caractéristiques des gènes présentés dans la Table 4.1 ont été calculées sur le génome entier et pour tous les animaux, l'héritabilité peut donc être bien inférieure en considérant le chromosome support du gène uniquement. En particulier, certains gènes ont un nombre non négligeable de eQTLs identifiés comme *trans* et appartenant à un autre chromosome, on en compte par exemple 79 pour IGF2 et 224 pour L3HYPDH dans le muscle. Pour ces deux gènes et dans le tissu musculaire, la qualité de prédiction avec BayesR reste relativement bonne, peut-être compensée par une héritabilité forte et une grande présence de eQTL *cis* et *trans* positionnés sur le chromosome support du gène. Il pourrait donc être intéressant pour certains gènes, connus pour avoir beaucoup de eQTLs en *trans* sur d'autres chromosomes, d'utiliser le génome entier afin d'exploiter ces régions d'intérêts.

L'expression de certains gènes, pourtant possédant une héritabilité suffisante et peu d'eQTLs identifiés sur d'autres chromosomes, a été mal prédite même dans le cas de la validation toutes-races (par exemple DET1 et PRKAG1 dans le foie). Ces mauvais résultats peuvent résulter du petit nombre d'animaux utilisés pour l'apprentissage et la prédiction, altérant la puissance des modèles de prédiction. Le découpage apprentissage/validation peut impacter fortement la qualité de prédiction par ailleurs, notamment dans le cas de profils d'expression très différents entre ces deux sous-jeux de données, ou des fréquences alléliques des variants majeurs dans les populations utilisées. La prédiction inter-races est une piste intéressante pour la recherche génomique chez les petites races, mais montrent encore une qualité des prédiction inférieure à la stratégie de validation toutes-races. Néanmoins, l'expression de certains gènes à été efficacement prédite, notamment pour le

gène HUS1. Pour d'autres gènes, il serait pertinent d'envisager une autre stratégie pour prédire des petites races, par exemple en ajoutant quelques individus de cette race dans le set d'apprentissage.

L'apport des annotations fonctionnelles dans la prédiction a très rarement abouti à une amélioration de la qualité de la prédiction. Leur construction est une étape cruciale, et il serait possible d'affiner nos listes d'annotations dans un prochain temps. Tout d'abord, ces annotations sont issues uniquement de porcelets LW, et peuvent donc dévoiler des mécanismes de régulation différents que les races DU et LD. Ensuite, si tous les stades d'évolution ont été utilisés ici, il pourrait être raisonnable de faire l'hypothèse que seule l'annotation (antérieure) la plus proche dans le temps de l'expression pourrait être intégrée (ici le stade NB). Une autre hypothèse plausible, est de considérer que les SNPs passant par des changements de méthylation, ou d'accessibilité à la chromatine, que nous qualifions de *switches*, peuvent constituer une annotation à part entière car représentant des dynamismes de régulation potentiellement impliqués dans l'expression du gène. Dans le projet GENE-SWitCH, de nouvelles annotations, issues de technologie ChIP-seq, vont bientôt être disponibles. Si il est possible de les intégrer directement dans la prédiction, nous voyons qu'en pratique ce n'est pas toujours pertinent d'utiliser le plus d'annotations possibles sans discrimination au préalable, et qu'il serait pertinent de mettre en place des stratégies pour évaluer quelles sont les meilleurs combinaisons d'annotations.

Avant le prétraitement des annotations fonctionnelles, un bon prétraitement des données génomiques est important pour la prédiction génomique. Nous avons fait le choix de réduire le nombre de SNPs issus des données WGS en filtrant sur la fréquence allélique, sur le taux de valeurs manquantes et le LD entre les SNPs appartenant à une fenêtre coulissante définie. Si nous avons choisi des paramètres standards pour le seuil de fréquence allélique et le LD, il pourrait être affiné en fonction des données. L'intérêt d'utiliser des données WGS est entre autres d'avoir un accès direct aux mutations causales, il serait donc regrettable de les éliminer dans ces étapes de filtrage. De même, une autre option que de retirer tout SNP ayant au moins une valeur manquant est d'imputer ces valeurs et ainsi d'éviter de perdre de l'information.

Un autre approche permettant d'utiliser des annotations fonctionnelles complexes (BayesRC+, Mollandin et al. (2022)) permet de mettre en avant les SNPs multi-annotés. Ce modèle repose sur l'hypothèse d'annotations cumulatives, ce qui pourrait être raisonnable pour les annotations ATACseq et de méthylation. Ainsi ces deux types d'annotations semblent apporter des informations complémentaires, en étant peu chevauchantes. Cette hypothèse d'additivité des annotations semble en revanche moins adaptée aux différentes temporalités pour un type de données omiques (ATACseq ou méthylation), tout du moins pour l'expression du gène qui s'inscrit lui-même dans une temporalité.

4.5 Conclusion

Si les modèles de prédiction génomiques sont largement utilisés pour la prédiction de caractères de production ou de risque de maladie, ils peuvent être intéressants à exploiter pour la prédiction de l'expression de gènes d'intérêts. L'expression de 11 gènes d'intérêt chez le cochon a été prédite à partir de données WGS et de données transcriptomiques pour trois races porcines (Duroc, Landrace et Large White), ainsi que d'annotations fonctionnelles représentant l'accessibilité à la chromatine et le niveau de méthylation, dans deux tissus (muscle et foie), et pour trois niveaux de développement (30 jours après fécondation, 70 jours après fécondation, nouveaux-nés). Afin de prendre en compte ces données d'annotations fonctionnelles complexes, nous avons utilisé le modèle de prédiction BayesRCπ, un modèle de mélange de mélange gaussien bayésien. Deux stratégies de validation, toutes-races ou inter-races ont été appliquées pour jauger des qualités de prédiction et d'apprentissage. Pour une grande partie des gènes, la qualité de prédiction n'a pas été améliorée, voire détériorée, avec l'utilisation des annotations fonctionnelles. En revanche, des différences relatives notables ont pu être observées dans certaines configurations, par exemple pour le gène SUPT3H dans le foie, pour la prédiction inter-races et avec des annotations de méthylation. L'apprentissage de ce gène avec les annotations fonctionnelles montre l'implication de SNPs déjà identifiés comme *cis*-eQTL dans une autre étude, mais aussi de SNPs plus éloignés du gène qui se sont vu priorisés par les annotations de méthylation. Parmi les annotations de méthylation, certaines semblent être plus enrichies en variants important pour l'expression de SUPT3H, notamment les annotations d'absence de méthylation totale, et prélevée à la naissance des porcelets. Ces résultats ouvrent la voie à un affinement de l'utilisation d'annotations fonctionnelles pour la prédiction de l'expression de gènes, afin de mieux comprendre les processus de régulation mis en place au niveau cellulaire.

4.6 Remerciements

Ce projet fait partie de GENE-SWitCH (https://www.gene-switch.eu) et a reçu un financement du programme de recherche et d'innovation Horizon 2020 de l'Union Européenne dans le cadre de la convention de subvention n° 817998. Il fait également partie d'EuroFAANG (https://eurofaang.eu), une synergie de cinq projets Horizon 2020 qui partagent l'objectif commun de découvrir des liens entre le génotype et le phénotype chez les animaux d'élevage et d'atteindre les objectifs mondiaux d'annotation fonctionnelle des génomes d'animaux (FAANG). En particulier, nous remercions les collègues suivants (classés par ordre alphabétique) pour leurs importantes contributions à ce travail : Maria Ballester, Marco Bink, Mario Calus, Pascal Croiseau, Daniel Crespo, Sylvain Foissac, Hélène Gilbert, Cervin Guyomar, Ole Madsen, Juan Pablo Sanchez, Bruno Perez, Andrea Rau, and Jani de Vos.

Chapter 5

Utilisation d'annotations quantitatives dans le modèle BayesRCπ

Jusqu'à présent, seule des informations binaires ont été exploitées dans les modèles de prédiction. Cependant, une partie des informations à exploiter proviennent de données continues, qui perdent alors en précision lors de l'étape de classification binaire. En particulier, l'information représentant la significativité d'association entre un SNP et une annotation est perdue. Cette information semble être particulièrement importante dans un modèle tel que BayesRCπ, où les marqueurs multi-annotés peuvent potentiellement être assignés préférentiellement à une annotation. Jusqu'à présent, les *hyperpriors* du modèles ont été peu abordés. Ils présentent cependant l'opportunité de réintégrer ce type d'informations. Dans ce chapitre, nous allons nous focaliser sur le paramètrage du mélange d'annotations de BayesRCπ, pour pondérer l'attribution préférentielle des SNPs multi-annotés dans une annotation via l'utilisation des informations quatitatives.

5.1 Matériels et méthodes

5.1.1 Distribution de Dirichlet

La distribution de Dirichlet est une distribution de probabilité continue et multinomiale. La loi de Dirichlet d'ordre $K \geq 2$ de paramètre $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_K) > 0$ a pour support $x_1, x_2, ..., x_K \in [0, 1]$, $\sum_{i=1}^{K} x_i = 1$ et pour densité de

probabilité

$$f(x_1, x_2, ..., x_K; \alpha_1, \alpha_2, ..., \alpha_K) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} \alpha x_k^{\alpha_i - 1}$$
$$\text{avec } B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}$$

B et Γ respectivement la loi Beta multivariée et la loi Gamma.

La loi Beta est un cas particulier de la loi de Dirichlet à $K = 2$ dimensions. La distribution de Dirichlet est fréquemment utilisée en inférence bayésienne, car elle est conjuguée aux lois Multinomiale et Catégorielle :

- Si $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_K) \sim$ Dirichlet(α) et $X = (X_1, X_2, ..., X_K)|\pi \sim$ Multinomiale(n, π) alors $\pi|X \sim$ Dirichlet$(\alpha + X)$. Le *posterior* de π est paramétré à partir de α et du nombre d'éléments dans chacune des classes de X.

- Si $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_K) \sim$ Dirichlet$(\boldsymbol{\alpha})$ et $\boldsymbol{X} = (X_1, X_2, ..., X_K)|\pi \sim$ Catégorielle(π) alors $\pi|X \sim$ Dirichlet$(\boldsymbol{\alpha} + \boldsymbol{X})$. La distribution *a posteriori* est donc la même que pour la loi multinomiale, ce qui est cohérent car on peut voir la loi catégorielle comme un cas particulier de la loi multinomiale avec $n = 1$. L'interprétation du *posterior* de $\boldsymbol{\pi}$ est un peu différente, étant paramétrée à partir de $\boldsymbol{\alpha}$, auquel on ajoute 1 uniquement dans la classe de X.

Le *posterior* de π est donc dépendant du paramètre α, qu'on peut exploiter pour ajouter de l'information dans le modèle.

Soit $X \sim$ Dirichlet(α). Le paramètre α contrôle la concentration ainsi que la pondération des différentes composantes de X. Plus la valeur de α_i est élevée, plus cela donne du poids à X_i. En particulier, si $\alpha_1 = \alpha_2 = ... = \alpha_K$ alors la distribution est symétrique. Si $\alpha_i < 1$, alors cela peut être vu comme un contrepoids, qui pousse X_i vers des valeurs extrêmes. On montre Figure 5.1 l'impact de cette paramétrisation, dans le cas d'une distribution de Dirichlet à trois composantes.

Dans les modèles BayesR (1.6), BayesRC (4.1), BayesRCπ et BayesRC+, les paramètres de mélange suivant une distribution de Dirichlet sont paramétrés de manière à être non-informatif, en utilisant la loi uniforme, c'est-à-dire $\boldsymbol{\alpha} = (1, 1, ..., 1)$. C'est une modélisation courante de ce *prior*, qui permet de ne favoriser *a priori* aucune de composantes. Il est aussi parfois proposé d'utiliser un aute *flat prior*, le prior de Jeffreys, qui suppose $\boldsymbol{\alpha} = (\frac{1}{2}, \frac{1}{2}, ..., \frac{1}{2})$. Il peut sembler raisonnable d'utiliser un *flat prior* pour le paramètre π, introduit par BayesR pour le mélange entre les quatres classes d'effets de SNPs, afin que le *posterior* soit déterminé par les données. En effet, utiliser $\alpha_k = 1$ pour tout $k \in \{1, 2, 3, 4\}$ a peu d'impact sur le *posterior* quand n le nombre de SNPs dans le modèle est suffisamment grand. Cependant, ajuster les priors des paramètres de mélange p_i entre annotations introduits par BayesRCπ pourrait être une piste d'amélioration du modèle, en pouvant intégrer des annotations fonctionnelles continues.

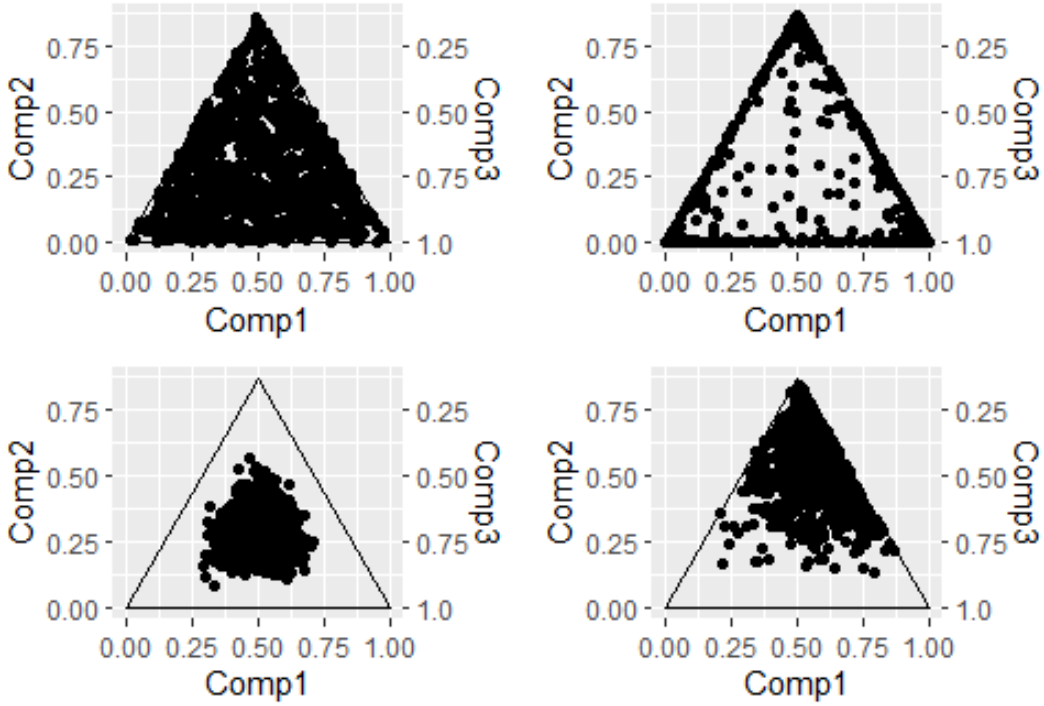


Figure 5.1: **Impact du paramètre α sur la distribution de Dirichlet.**
Pour $n = 1000$ échantillons, de gauche à droite, et de haut en bas respectivement: $\alpha_1 = \alpha_2 = \alpha_3 = 1$; $\alpha_1 = \alpha_2 = \alpha_3 = 0.1$; $\alpha_1 = \alpha_2 = \alpha_3 = 10$; $\alpha_1 = 1, \alpha_2 = 2$ et $\alpha_3 = 5$

5.1.2 Intégration d'annotations quantitatives

Jusqu'à présent, BayesRCπ exploite une information binaire pour chaque SNP suscebtible d'appartenir ou non à chaque annotation. Si cela semble direct à partir de certaines informations, telles que les annotations structurelles, cela demande l'utilisation d'un seuil à fixer pour classifier les SNPs chez d'autres annotations, par exemple issues de données OMICS. Le niveau de confiance qu'un SNP appartienne à une annotation pourrait être une piste pour la pondéreration des annotations dans le modèle. Dans la suite nous discuterons de l'utilisation de p-valeurs pour quantifier ce niveau de confiance, mais d'autres annotations quantitatives sont également possibles (e.g., quantification du niveau de méthylation ou d'accessibilité de la chromatine).

Sous l'hypothèse qu'un marqueur multi-annoté est mieux représenté par l'annotation où il est le plus fortement associé, nous proposons l'utilisation de $\alpha_j^{(c)} = -\log_{10}(\rho_j^{(c)})$, avec $\alpha_j^{(c)}$ la composante c de l'hyperprior α_j du paramètre de mélange p_j du SNP j, et $\rho_j^{(c)}$ la p-valeur du SNP j dans l'annotation c. Par souci computationnel, les p-valeurs supérieures à un seuil de significativité (par exemple, 5%) sont fixées à 1. Cette stratégie correspond donc à un affinement de la classification binaire utilisée précédemment.

5.1.3 Simulations

Nous reprenons les simulations utilisées chapitre 3, pour les niveaux d'héritabilités $h^2 = \{0.2, 0.5\}$. Pour rappel, 5 QTLs forts (1% de σ_g^2), 300 moyens (0.1% de σ_g^2) et 6500 SNPs à effet faible (0.01% de σ_g^2) indépendants ont été simulés, et ce pour 50 datasets indépendants. On échantillonne quatre annotations suivant différents niveaux d'enrichissement:

- Une annotations très enrichie, contenant les 5 QTLs forts, les 300 moyens et 150 SNPs à faible ou nul effet
- Deux annotations moyennement enrichies, contenant 2 QTLs forts, 20 moyens, et 400 SNPs à faible ou nul effet
- Une annotation faiblement enrichie, contenant un QTL fort et 450 SNPs à faible ou nul effet

Les annotations sont étendues aux SNPs directement voisins (± 1). De plus, les annotations sont construites de façon à ce que chaque QTL fort soit chevauchant dans l'annotation très enrichie ainsi qu'une des 3 autres.
Afin de voir l'impact de la modification de l'*hyperprior* sur la priorisation des QTLs, les QTLs forts et leurs voisins sont supposés plus significatifs dans l'annotation très enrichie que dans les autres, avec respectivement des p-valeurs à 10^{-10}, 10^{-7} et 10^{-6} dans les annotations très enrichie, moyennement enrichie ou faiblement enrichie. En utilisant la transformation $-\log_{10}(\rho_j^{(c)})$, on a utilise donc des valeurs d'*hyperpriors* pour les QTLs forts chevauchant à 10, 7 ou 6 respectivement. Ces paramétrages, désignés comme "priors pondérés" sont comparés aux priors uniformes.
Nous remarquons qu'un des 50 jeux de données simulées a mené à un problème de calcul qui n'a pas encore pu être résolu. Nous explorons toujours les conditions qui ont provoqué cette erreur, et dans la suite nous présentons uniquement des résultats basés sur les 49 jeux de données restants.

5.2 Résultats

5.2.1 Qualité de prédiction

Les moyennes ($\pm$ écart-types) des corrélations de Pearson sur set de validation des 49 datasets sont respectivement de 0.215 ($\pm$ 0.050) et 0.216 ($\pm$ 0.049) pour les modèles avec priors uniformes et pondérés respectivement, pour $h^2 = 0.2$. Au niveau $h^2 = 0.5$, ces moyennes atteignent respectivement 0.458 (± 0.047) et 0.458 (± 0.047). A l'aide d'un test de Student apparié, on n'observe aucune différence significative entre les corrélations sur set de validation en utilisant des priors pondérés sur les QTLs forts plutôt que des priors uniformes, avec des p-valeurs de 0.71 et 0.121 aux niveaux d'héritabilité h^2 respectifs de 0.2 et 0.5. Cependant, on observe une variation plus importante entre la qualité de prédiction obtenue avec priors pondérés et priors

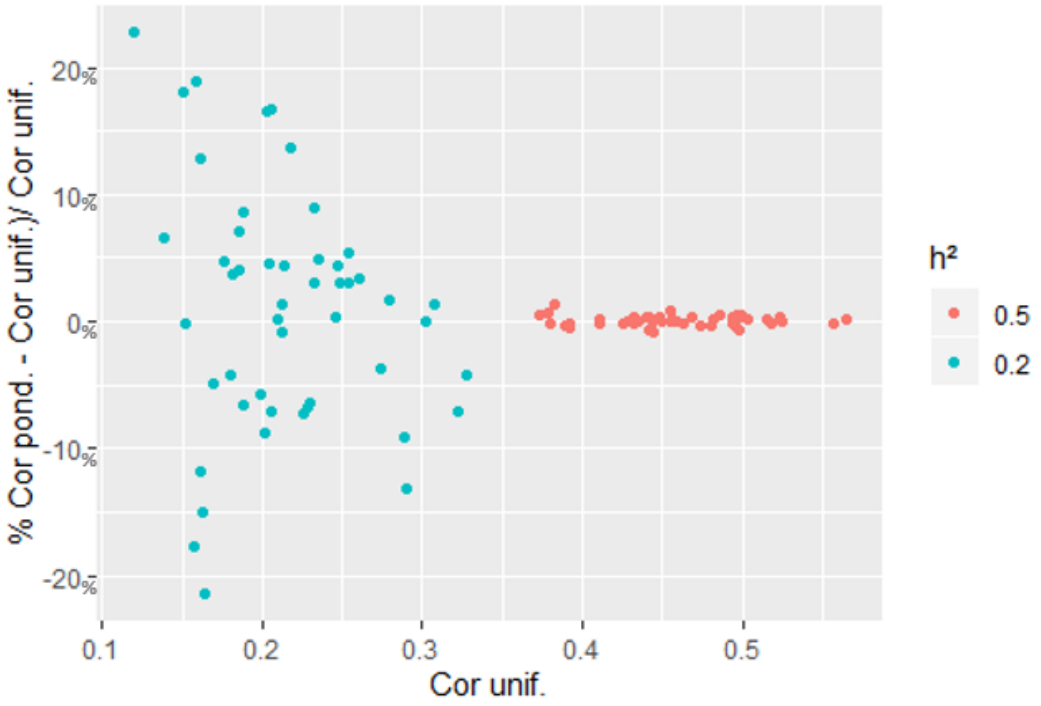


Figure 5.2: **Variation de la différence de la qualité de prédiction**
La différence relative de la corrélation de validation obtenue pour BayesRCπ avec des *priors* pondérés et des *priors* uniformes est représenté en fonction de la corrélation référente de BayesRCπ avec les *priors* uniformes. Ces résultats sont présentés pour les deux niveaux d'héritabilité $h^2 = 0.2$ (en bleu), et $h^2 = 0.5$ (en rouge).

uniformes dans les cas plus difficiles (i.e., où l'héritabilité est plus faible). On montre Figure 5.2 le taux de variation entre la corrélation du modèle avec priors pondérés et du modèle avec priors uniformes, en fonction des corrélations avec priors uniformes, utilisés comme référence. Au niveau $h^2 = 0.2$, et pour des petites valeurs de la corrélation de référence, on peut observer jusqu'à $\pm$22% environ. Cette variation tend à diminuer avec la qualité de corrélation, jusqu'au niveau $h^2 = 0.5$ où elle atteint au plus 1.30%.

5.2.2 Assignation à une annotation

Tous les QTLs forts sont chevauchants, et ont donc la possibilité d'être assignés à différentes annotations. On peut classifier chaque SNPs chevauchant dans une annotation dont sa probabilité d'appartenance *a posteriori* est la plus élevée, de façon analogue à la règle du MAP présenté chapitre 2. Dans la Table 5.1, on montre le nombre de QTLs forts assignés majoritairement dans chacune des quatre annotations introduites dans le modèle, pour les deux niveaux d'héritabilité, et les deux paramétrisations des priors. Avec des priors uniformes, les QTLs forts sont majoritairement assignés à l'annotation très enrichie, représentant respectivement 65.3% et 94.3% de l'effectif total pour $h^2 = 0.2$ et $h^2 = 0.5$. Avec des annotations pondérés, cela passe à 100% pour les deux héritabilités, suggérant un effet du prior pour donner du poids à l'annotation très enrichie comparativement aux autres, et ainsi

mieux classer les QTLs forts.

		Classification d'annotations (maximum a posteriori)			
h^2	Priors	Très enrichie	Moyennement enrichie (1)	Moyennement enrichie (2)	Faiblement enrichie
0.2	Uniformes	160	32	32	21
	Pondérés	245	-	-	-
0.5	Uniformes	231	5	6	3
	Pondérés	245	-	-	-

Table 5.1: **Effectifs des 245 QTLs forts assignés majoritairement à chaque annotation**
Nombre de QTLs forts assignés majoritairement à chacune des quatre annotations (1 très enrichie, 2 moyennement enrichie, 1 faiblement enrichie), pour les deux niveaux d'héritabilités $h^2 = \{O.2, 0.5\}$ et les priors uniformes soit pondérés.

5.2.3 Assignation aux classes d'effets de SNP

L'assignation d'un marqueur à une annotation ou une autre a potentiellement un effet sur son inclusion dans le modèle, à partir du moment où les annotations ont des enrichissements différents. On montre Table 5.2 l'assignation des QTLs forts à chacune des quatre catégorie d'effet de SNPs, déterminées à partir de la règle du MAP. A l'héritabilité $h^2 = 0.2$, 22 QTLs forts sont assignés majoritairement dans une classe non nulle (dont 3 dans la classe forte) avec des priors uniformes, ce nombre augmente à 33 (dont 4 dans la classe forte) avec des priors pondérés. On retrouve la même tendance avec $h^2 = 0.5$. L'assignation des QTLs dans l'annotation très fortement enrichie semble avoir favoriser leur assignation dans des classes d'effets non-nuls. En moyenne, les QTLs forts ont une inclusion dans le modèle 5% plus élevé avec des priors pondérés qu'uniformes, aux deux niveaux d'héritabilités (test de Student, p-valeur respectivement à 6.10^{-10} et $< 2.210^{-16}$).

		Classification par effet de SNPs (maximum a posteriori)			
h^2	Priors	Fort	Moyen	Faible	Nul
0.2	Uniformes	3	-	19	223
	Pondérés	4	1	28	212
0.5	Uniformes	13	54	40	138
	Pondérés	14	63	65	103

Table 5.2: **Effectifs des 245 QTLs forts assignés majoritairement dans chaque classe d'effet de SNPs**
Nombre de QTLs forts assignés majoritairement à chacune des quatre classe d'effet de SNPs (nul, faible, moyen ou fort), pour les deux niveaux d'héritabilités $h^2 = \{O.2, 0.5\}$ et les priors uniformes soit pondérés.

5.2.4 Priorisation des QTLs forts

Les QTLs forts, plus fréquemment inclus dans le modèle, ont une variance *a posteriori* plus élevée, augmentant en moyenne de 5.3% et 5.2% pour $h^2 = 0.2$ et $h^2 = 0.5$ respectivement (test de Student, p-valeur respectivement à 0.0035 et 2.16310^{-14}). Cette augmentation de la variance *a posteriori* induit un meilleur classement des QTLs forts, en considérant leur rang moyen comme leur rang médian (Tableau 5.3. En moyenne, l'utilisation de priors pondérés ont fait gagner 83 et 27 rangs pour les héritabilités respectives de 0.2 et 0.5. En médiane, la progression est de 64 et 5 rangs respectivement.

		Classements des QTLs forts	
h²	Priors	Moyenne	Médiane
0.2	Uniformes	848	418
	Pondérés	757	371
0.5	Uniformes	242	51
	Pondérés	213	46

Table 5.3: **Classement des 245 QTLs forts**
Classement (par variance décroissante) moyen et médian des 245 QTLs forts simulés, pour les deux niveaux d'héritabilités $h^2 = \{O.2, 0.5\}$ et les priors définis comme uniformes ou pondérés.

En utilisant le top 10 des meilleurs SNPs pour chaque jeu de données (par variance décroissante) comme dans le chapitre 2, on identifie 6 nouveaux QTLs comme forts au niveau $h^2 = 0.5$ avec les priors pondérés (69 QTLs contre 63). Au niveau $h^2 = 0.2$ aucun nouveau QTL n'est détecté à ce seuil (22 QTLs pour les deux paramétrisations).

5.3 Discussion

Les résultats suggèrent que l'utilisation des annotations de façon continue, en les injectant dans les paramètres de mélange *a priori* de BayesRCπ, ont un effet sur l'assignation des QTLs dans les différentes annotations, ce qui entraîne une modification de l'estimation de la variance *a posteriori* et donc de leur identification en tant que QTLs. Les SNPs voisins ($\pm$ 1 autour des QTLs forts), paramétrisés de la même façon par cohérence avec la construction des annotations par fenêtre, ont aussi vu leur inclusion et leur variance augmenter.

Si la variance *a posteriori* des QTLs forts est mieux estimée avec des *priors* pondérés, la qualité de prédiction n'est pas améliorée pour autant. De plus, on observe aucun impact de la pondération des *priors* sur la corrélation de validation quand la qualité de prédiction est déjà bonne avec des *priors* quantitatifs. Une explication possible est que dans des conditions favorables, tel qu'avec une bonne héritabilité, les QTLs sont facilement identifiables, et leur mise en avant par de nouvelles pondérations n'est pas nécessaire. Néanmoins, on observe une différence allant jusqu'à ± 3.8 points de corrélation pour $h^2 = 0.2$, ce qui suggère un impact de la paramétrisation sur la qualité de prédiction dans un contexte moins favorable. Seuls les QTLs forts et leur voisins ont été pondérés, représentant 15 variants sur 46178 et environ 5% de la variance additive totale, ce qui limite aussi l'impact de leur estimation sur la qualité de prédiction. Des valeurs plus extrêmes des *hyperpriors* aurait aussi pu avoir un effet plus fort sur les résultats de prédiction.

Il est ici fait l'hypothèse qu'un variant impliqué dans la variabilité d'un caractère donné ait une significativité plus élevée dans une annotation importante pour ce même caractère, et plus enrichie, qu'une autre. En pratique, cela pourrait être discutable, en particulier en cas d'annotations de provenances hétérogènes, où la significativité dépend de la puissance de l'étude qui la calcule. Il est aussi possible dans ce cas d'utiliser ces *hyperpriors* pour pondérer les annotations en fonction de leur puissance, ou dans la confiance qu'on leur porte, et ainsi en prioriser

certaines. Ainsi, si cette étude permet l'intégration de données continues dans les modèles de prédiction génomique, elle peut aussi être exploitée pour affiner l'assignation des SNPs dans ces différentes annotations, sur divers critères.

Dernièrement, il est possible d'utiliser d'autres valeurs que $-\log_{10}(\rho_j^{(c)})$ pour le SNP j dans l'annotation c, dans la mesure où ces valeurs entre les différentes annotations restent comparables. Par exemple, utiliser une mesure de comptage transcriptomique d'une part et une p-valeur d'une autre créerait un déséquilibre évident pour le premier dans une grande majorité des cas. Il est alors nécessaire de projeter les valeurs continues dans un espace comparable entre annotations.

5.4 Conclusion

L'exploitation des *hyperpriors* de BayesRCπ pour influencer l'attribution des SNPs multi-annotés à une seule annotation est une première façon d'intégrer des annotations continues. En donnant plus de chances aux QTLs forts simulés d'être assignés à une annotation fortement enrichie, ceux-ci sont plus fréquemment inclus dans le modèle et priorisés par rapport aux autres SNPs. Cela a pourtant peu de conséquence sur la qualité de prédiction, sans hausse significative de la corrélation de validation, en particulier quand les QTLs sont relativement bien identifiés avec des *hyperpriors* continus. Si cette approche de pondération fonctionne "mécaniquement", il reste indispensable de l'adapter en fonction du type d'annotations traitées.

Chapter 6

Conclusion et perspectives

Deux modèles bayésiens de référence, BayesR et BayesRC, ont été utilisés comme point de départ de ce travail. Dans une première étude, BayesR, un modèle génomique bayésien exploitant quatre classes d'effet de SNPs (nul, faible, moyen et fort) à l'aide d'une distribution *a priori* suivant un mélange gaussien, a montré ses qualités pour à la fois prédire des phénotypes et détecter des QTLs dans une large variété de simulations, avec des héritabilités et des tailles et effectifs de QTL variés, reproduisant ainsi une multitude d'architectures génomiques. BayesRC est une extension directe de BayesR permettant de partitionner les SNPs entre différentes annotations, en supposant qu'elles soient pourvues d'un enrichissement de SNPs à effet nul, faible, moyen ou fort qui leur est propre. Étudier BayesR et BayesRC en premier lieu a donc permis de déterminer le potentiel et les limites de ces modèles, tout en les gardant en référence par rapport aux nouveaux modèles à développer, BayesR comme modèle comparatif sans annotation fonctionnelle, et BayesRC comme modèle comparatif avec annotations fonctionnelles mais simplifiées pour éviter les chevauchements, impliquant une perte d'information. BayesRCπ et BayesRC+ ont alors été développés pour surmonter la principale limitation de BayesRC, la limite d'une annotation par SNP, excluant l'exploitation directe des annotations chevauchantes. Ces deux modèles reposent sur deux hypothèses différentes sur la façon de comprendre un SNP multi-annotés, (1) un variant multi-annoté devrait pouvoir évoluer entre ses différentes annotations, (2) un variant multi-annoté devrait avoir plus de chances d'être intégré dans le modèle, et ainsi être priorisé. Pour répondre à cette première hypothèse, BayesRCπ propose une distribution des effets de SNPs comme un mélange de mélange gaussien. Pour la deuxième hypothèse, BayesRC+ mise sur une distribution cumulative de mélange.

Ces deux modèles ont été implémentés en Fortran 90 dans un logiciel disponible sur GitHub, BayesRCO (pour "BayesRC Overlap"), afin de pouvoir les évaluer, à la fois sur données simulées et données réelles. Sur données simulées, ils ont montré des améliorations de la qualité de prédiction modestes, jusqu'à 2 points en moyenne pour certaines configuration favorables, mais aussi une priorisation des QTLs simulés et placés dans une ou plusieurs annotations. Sur données réelles, les résultats sont bien plus variables. Deux applications sur deux jeux de

données porcins distincts ont été effectuées, dont une dans le cadre du projet co-finançant ce travail de thèse, GENE-SWitCH. Ces deux applications ont eu des buts et des plans d'études très différents. Le premier jeu de donnée, issu du projet PigHeat, avait pour objectif de mieux prédire des caractères de production, à partir d'annotations provenant de pigQTLdb, une base d'annotations publique regroupant les QTLs détectés dans une variété d'étude et catégorisée en fonction de leur lien avec des traits de production. Le deuxième jeu de données, issu des données GENE-SWitCH, cherche à prédire l'expression d'un panel de gène d'intérêts, en exploitant des annotations, plus complexes que dans la première application, liées à des mécanismes de régulation (accessibilité à la chromatine et méthylation) à plusieurs stades et dans plusieurs tissus. Dans ces deux applications en données réelles, les résultats se sont montrés beaucoup plus variables, avec des prédictions parfois améliorées, mais qui semble rester dans des cas spécifiques, avec des annotations adaptées. Pour d'autres phénotypes, l'apport des annotations a détérioré la qualité de prédiction.

Perspectives de modélisation L'utilisation de modèles bayésiens permet l'intégration d'annotations fonctionnelles pour la prédiction génomique à travers l'utilisation de ses *priors*. C'est une approche relativement naturelle et intuitive, qui permet une certaine flexibilité face aux différents jeux de données. Les modèles BayesRCπ et BayesRC+ autorisent l'utilisation de multiples annotations fonctionnelles en prenant en compte le caractère multi-annoté de certains SNPs. S'ils répondent à des hypothèses différentes, on pourrait aussi utiliser leurs qualités respectives en un nouveau modèle hybride, ou certaines annotations seraient additives, tandis que d'autres nécessitent un modèle de mélange. Cela pourrait répondre à des données d'annotations plus hétérogènes que celles détaillées dans cette thèse, et pouvant mêler multi-omiques, multi-tissus et différentes temporalités. Afin de tirer le meilleur parti de BayesRCπ et BayesRC+, on pourrait modéliser les effets du SNP i, ayant un set d'annotations $\mathbf{C}_i$ partitionné en G sous-ensembles d'annotations additifs $\mathbf{C}_i = \{C_i^{(1)}, C_i^{(2)}, ..., C_i^{(G)}\}$, eux-mêmes constitués d'annotations non additives, par la distribution

$$f(\beta_i|\mathbf{C}_i) = \sum_{g=1}^{G} \sum_{c \in \mathbf{C}_i} 1_{c \in C_i^{(g)}} p_{i,c} \sum_{k=1}^{4} \pi_{k,c} f_k(\cdot|\theta_k)$$

$$\text{tel que } f_k = \begin{cases} \delta(0), \text{ si } k = 1 \\ \phi(\cdot|0, \theta_k) \text{ sinon} \end{cases},$$

avec δ, ϕ, θ_k, $\pi_{k,c}$, définit comme précédemment, $\sum_{g=1}^{G} \sum_{c \in \mathbf{C}_i} 1_{c \in C_i^{(g)}} p_{i,c} = 1$ pour tout SNP i et partitionnement à G sous-ensembles. Dans l'application aux données GENE-SWitCH (chapitre 4), il aurait par exemple pu être envisageable de considérer les annotations de méthylation et d'accessibilité à la chromatine comme additives et au sein de celles-ci procéder à un mélange des trois temporalités.

D'autres adaptations des modèles peuvent aussi être envisagés pour mieux répondre aux différents cas de figure

de la prédiction génomique. Notamment, il peut être souhaitable de faire de la prédiction *multi-trait*, qui peuvent aider à la prédiction de caractères corrélés les uns avec les autres (Guo et al., 2014). En ce sens, il existe une extension *multi-trait* de BayesR, BayesMV (Kemper et al., 2018). Les méthodes BayesRCO étant basées sur BayesR, l'extension de BayesMV à l'intégration d'annotations fonctionnelles pourrait être relativement facile à modéliser et implémenter.

De plus, d'autres travaux de modélisation, notamment sur les distributions *a priori* des phénotypes, peuvent être envisagées. Si les méthodes BayesRCO sont adaptées à une variété de phénotypes, quantitatifs et suivant *a priori* un loi gaussienne, certains phénotypes ont une distribution toute autre. Par exemple, les caractères dichotomiques, tel que la présence ou non d'une maladie, nécessite en général une distribution logistique, et certains modèles bayésiens ont pu être développés pour ce type de phénotype (Technow and Melchinger, 2013). On peut aussi avoir à faire à des prédictions de caractères temporels (*age-at-onset*), où des modèles propres à l'analyse de survie sont employés (Ojavee et al., 2021).

Si l'utilisation des *hyperpriors* de BayesRCπ ont montré une première voie (chapitre 5) pour exploiter l'aspect quantitatif de certaines annotations, et ainsi mieux les exploiter, cela reste limité aux cas des SNPs multi-annotés, en pondérant l'une par rapport à l'autre. BayesRC+ lui n'exploite en aucun cas cette information. Une prochaine étape dans le travail de modélisation de modèles génomiques prédictifs intégrant des annotations fonctionnelles serait donc la capacité d'exploiter pleinement le côté quantitatif de toutes les données annexes que l'on souhaite injecter dans les modèles.

Optimisation du logiciel D'autre part, un travail d'implémentation serait nécessaire pour optimiser l'algorithme, et ainsi mieux exploiter les modèles. Une optimisation de l'algorithme utilisé pour BayesRCπ et BayesRC+ pourrait notamment permettre d'utiliser une plus grande densité de génotypage dans un temps de calcul raisonnable. En effet, si l'utilisation d'algorithme MCMC permet l'implémentation directe des modèles, elle reste coûteuse computationnellement pour des modèles bayésiens complexes comme ceux développés ici, cela est donc limitant à l'utilisation d'une grande densité de génotypage, voir d'analyses sur séquence complète (*whole genome sequencing*), où BayesRCπ et BayesRC+ sont pourtant adaptés pour permettre d'avoir à la fois les mutations causales, et de prioriser certaines régions du génome.

Une option répandue pour diminuer le temps de calcul réside dans la parallélisation de l'algorithme, qui consiste en la division des tâches afin qu'elles soient traitées en simultanées, et non en séquentiel, et ce sur plusieurs noeuds (Zhao et al., 2020; Calus et al., 2016). Les méthodes MCMC, bien qu'adaptées pour l'échantillonnage des distributions *posteriors*, peuvent se révéler computationnellement lourdes. L'utilisation d'un algorithme EM (*expectation-maximization algorithm*), comme emBayesB (Shepherd et al., 2010) ou emBayesR (Wang et al., 2015), semblent offrir des qualités de prédiction proches de leurs alternatives MCMC, mais en améliorant le temps de calcul dans de grands jeux de données. Des méthodes hybrides MCMC et EM ont aussi été proposées, avec

l'objectif d'une part d'exploiter le gain computationnel des méthodes EM, avec la précision des algorithmes MCMC. On peut citer Hyb_BR (Wang et al., 2016), dont l'hybridation EM-MCMC promet une diminution drastique du temps de calcul. Si l'optimisation computationnelle du *software* BayesRCO n'est pas au coeur de ce travail de modélisation, on voit que son adaptation à des données de très grandes dimensions pourrait lui permettre d'exploiter au mieux son potentiel.

De la bonne utilisation des annotations Un point clef de l'amélioration de la qualité de prédiction, indépendamment de la modélisation des méthodes de prédiction et de leur implémentation, réside dans la façon de constituer les annotations. Tout d'abord, la génération de listes d'annotations à partir de données hétérogènes, issues de méthodologie de puissance variable, n'est pas si directe. Si certaines listes semblent plus naturelles à constituer, par exemple à partir d'annotations structurelles (e.g. le SNP est dans un intron ou non), d'autres demandent une étape de classification en amont. Il faut alors faire des choix sur ce que l'on considère comme informatif pour le caractère, ce qui peut être subjectif, particulièrement pour des caractères dont le mécanisme est peu connu. Pour des données de méthylation, quelle est la meilleure classification, méthylé vs non-méthylé, ou une combinaison de trois listes binaires de non-méthylé, faiblement méthylé, fortement méthylé ? Pour des données transcriptomiques où nous pouvons nous intéresser aux gènes différentiellement exprimés, faut-il annoter les SNPs contenus dans ces gènes, ou considérer une fenêtre autour du gène pour prendre en compte les zones régulatrices à proximité ? Quel tissu ou temporalité choisir ? Est-il informatif d'utiliser des annotations générées sur d'autres races, ou tout du moins sur des populations éloignées génétiquement ? Pour toutes ces questions et bien d'autres, il est nécessaire de faire des choix qui peuvent impacter la qualité de prédiction, ainsi que la potentielle détection d'une région causale. Il n'est pas rare d'être confronté.e.s à plus d'une de ces interrogations en voulant intégrer le plus d'informations dans le modèle, ce qui multiplie les possibilités de scénarios d'annotations.

Par ailleurs, même en faisant l'hypothèse qu'une annotation est bien "construite", cela ne la rend pas pour autant pertinente à utiliser dans toute situation. En cas d'utilisation d'annotations non informatives, nous avons pu constater des détériorations de la qualité de prédiction, en priorisant des régions qui n'auraient pas dû l'être. Dans des recherches futures, il serait donc nécessaire de mettre en place des stratégies en amont pour évaluer l'intérêt des annotations pour la prédiction d'un caractère complexe donné. Une partie de la décision d'utilisation d'une annotation peut découler des connaissances générales sur le caractère, en essayant de cibler des annotations ayant un lien connu avec un caractère de production, ou dans un tissu associé. Des méthodes statistiques peuvent offrir aussi des informations sur l'information portée par les annotations. Par exemple, la distribution des p-valeurs des SNPs contenus dans une annotation, issues d'un GWAS sur les données (si la puissance est suffisante) peut donner une première indication sur l'association des SNPs des annotations et le caractère prédit. Une autre stratégie pourrait être d'appliquer une méthode de prédiction n'intégrant pas d'annotations fonctionnelles, tel qu'un

GBLUP ou BayesR, et de comparer la qualité de prédiction en utilisant le génotype entier, le génotype filtré des SNPs issus des annotations, ou au contraire en utilisant uniquement les SNPs des annotations.
Les modèles BayesRCO semblent capables de différencier les annotations utilisées en leur attribuant des enrichissements différents, mais cela ne suffit pas à annuler l'impact négatif provenant du bruit apporté par des annotations non informatives. Ils possèdent néanmoins des *outputs* intéressants dans l'étude de l'information portée par les informations, en particulier BayesRCπ, tel que la proportion de SNPs dans chacune des quatre classes d'effets, ou la densité des variances *a posteriori* dans chaque annotation. Ces *outputs* peuvent alors être utilisés pour filtrer ou fusionner des annotations dans un second temps.

BayesRCπ ou BayesRC+ ? Les modélisations de BayesRCπ et BayesRC+ reposent sur des hypothèses différentes, par conséquent il est là encore nécessaire de déterminer leur utilisation selon le type de données et annotations exploitées. BayesRC+, en particulier, repose sur une hypothèse forte, qui peut influer fortement sur les résultats de prédiction, et l'hypothèse d'additivité des annotations n'est pas cohérente biologiquement dans tous les cas. Ainsi, cette hypothèse peut paraître raisonnable dans l'application sur données réelles (chapitre 3), avec des annotations regroupant les QTLs par association avec des caractères de production, et où les SNPs multi-annotés sont ainsi identifiés comme importants dans différentes études. En revanche, cette hypothèse est moins plausible pour les données utilisées dans le chapitre 4, qui cherche à prédire l'expression d'un gène à un moment *t*, en fonction des données de méthylation et d'accessibilité à la chromatine à trois stades de développement antérieurs, où il est moins justifiable biologiquement de prioriser un SNP méthylé pour les trois temporalité de méthylation par exemple. Pour les deux méthodes, il serait utile de déterminer un cadre pour leur bonne utilisation.
Si l'utilisation de BayesRCπ et BayesRC+ ne montre pas toujours une amélioration notable de la qualité de prédiction, ils semblent avoir le potentiel pour identifier et estimer des QTLs, comme nous l'avons constaté lors de l'étude de simulation du chapitre 3. Cela a une conséquence sur la qualité de prédiction par la suite pour deux raisons. Tout d'abord, il a été montré que l'inclusion des mutations causales dans des puces SNPs augmentait la précision de prédiction dans la population de validation (Meuwissen and Goddard, 2010). Deuxièmement, l'exploitation de ces SNPs causaux peut améliorer la persistance de prédiction (MacLeod et al., 2014). Ainsi, les recombinaisons alléliques ont un impact sur le LD, et de génération en génération on peut observer un LD en baisse entre un SNP causal et d'autres voisins. L'utilisation d'un SNP en LD complet avec un SNP causal au lieu de celui-ci, n'aura pas d'effet négatif sur la prédiction. Plusieurs générations et recombinaisons plus tard, l'utilisation de ce SNP proxy n'est plus équivalente à l'utilisation du SNP causal, ce qui altère la prédiction. Une meilleure identification des SNPs causaux par BayesRCπ et BayesRC+ peut donc avoir un effet positif sur la qualité de prédiction dans le temps.

Utilisation de BayesRCO dans l'évaluation génomique Aujourd'hui, des méthodes *single-step*, et en particulier le ssGBLUP (*single-step genetic best linear unbiaised prediction* (Misztal et al., 2009; Aguilar et al., 2010), sont principalement utilisées dans les évaluations génomiques de routine, permettant de combiner des scores génomiques et sur pedigree des candidats à la sélection. Il a été montré que certains modèles *single-step* bayésiens ont obtenu de meilleurs résultats que le ssGBLUP, notamment sur des caractères déterminés par une petite quantité de QTLs à fort effet (Zhou et al., 2018). Cependant, le gain de précision des méthodes bayésiennes reste pour l'instant trop modeste pour qu'il soit intéressant d'utiliser ces méthodes en routine, surtout pour des temps de calcul plus élevés qu'un ssGBLUP classique. De même pour les modèles BayesRCO, il faudrait montrer un gain de prédiction suffisant pour envisager leur utilisation par les entreprises de sélection. Ce gain de prédiction peut être constaté dans certains cadres où les annotations apportent une réelle information pour un caractère observé, mais que nous avons encore du mal à définir. Il n'est donc pas intéressant de développer une méthodologie *single-step* BayesRCO tant qu'un cadre général des cas où l'on observe une meilleure prédiction n'est pas défini. En revanche, ces modèles offrent un potentiel d'utilisation dans un cadre plus fondamental, et notamment pour la meilleure compréhension des mécanismes biologiques sous-jacents à un caractère complexe.

Par leur flexibilité à différentes données de génotypages, de types de phénotypes, et de construction des annotation, BayesRCπ et BayesRC ouvrent la voie à de nouvelles possibilités d'application et de compréhension des mécanismes cellulaire liées aux phénotypes complexes, dans le cadre de l'élevage, mais aussi dans d'autres domaines d'étude, tel que la médecine. Ils ont aussi permis un gain de prédiction dans certaines configurations, modérés en simulation mais pouvant être important en données réelles, mais ponctuel. Cela promet un espoir d'amélioration de la qualité de prédiction dans le futur, sous couvert d'une bonne identification de leur cadre d'utilisation optimal.

Appendix A

BayesRCO: notice d'utilisation

BayesRCO User Guide

Version 0.0.1

Fanny Mollandin[1], Pascal Croiseau[1], Andrea Rau[1,2]

[1] Université Paris-Saclay, INRAE, AgroParisTech, GABI
[2] BioEcoAgro Joint Research Unit, INRAE, Université de Liège, Université de Lille, Université de Picardie Jules Verne

January 4, 2022

Contents

1 Overview

The BayesRCO software includes five different Bayesian genomic prediction models, including three state-of-the art approaches and two novel algorithms:

- BayesCπ (Habier et al., 2011)
- BayesR (Erbe et al., 2012)
- BayesRC (MacLeod et al., 2016)
- BayesRCπ
- BayesRC+

All five models are Bayesian Gaussian mixture models for the genomic prediction of complex traits using genetic variation such as single nucleotide polymorphisms (SNPs), with parameters estimated using a Markov Chain Monte Carlo (MCMC) algorithm. These prediction methods also facilitate a study of the underlying genomic architecture of these traits, in particular by enabling QTL mapping. The two new methods implemented in BayesRCO, BayesRC+ and Bayesπ, both aim to integrate prior categorizations of SNPs arising from multiple, potentially overlapping annotations.

This document is intended to describe the underlying models of BayesRCO, provide help for download and compilation of the software, and describe the various inputs, outputs, and options provided by the software.

> *Note*: The core of the BayesRCO software is based on version 0.75 of the BayesR software described and implemented by Moser et al. (2015), although further functionalities and outputs have been added (including options for the BayesRCπ and BayesRC+ algorithms). As many of the input arguments in BayesRCO are the same as those of BayesR, there are many similarities between this document and the BayesR User Manual.

2 Bayesian genomic prediction models

2.1 SNP effect prior distributions

All five Bayesian genomic prediction models included in BayesRCO exploit the same underlying linear model, which aims to obtain an accurate prediction of a vector of phenotypes **y** by best estimating a vector of SNP effects β:

$$\mathbf{y} = \mu \mathbb{1}_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$
$$\mathbf{e} \sim N(0, \mathbf{I}_n \sigma_e^2)$$

The five Bayesian models included in BayesRCO can be differentiated by the prior distribution attributed to $\boldsymbol{\beta}$, as indicated in the table below. In each model, SNP effects are assumed to follow a Gaussian mixture distribution with varying numbers of components: 2 (null and non-null) for BayesCπ, or 4 (null, low, medium and high) for all of the other methods. In addition, three of the models (BayesRC, BayesRCπ and BayesRC+) additionally incorporate a prior known categorization of SNPs (e.g., according to functional information, or lists of candidate or causal mutations).

Method	SNP effect prior distribution	# effect classes	Prior Annotations	|A(i)|
BayesCπ	$\beta_i \sim \pi\mathcal{N}(0,0) + (1-\pi)\mathcal{N}(0,\sigma_\beta^2)$	2	No	-
BayesR	$\beta_i \sim \sum_{\ell=1}^{4} \pi_\ell \mathcal{N}(0, k\sigma_g^2)$	4	No	-
BayesRC	$\beta_i \vert A(i) \sim \sum_{\ell=1}^{4} \pi_{\ell,a} \mathcal{N}(0, k\sigma_g^2)$	4	Yes, disjointed	=1
BayesRC+	$\beta_i \vert A(i) \sim \sum_{a\in A(i)} \sum_{\ell=1}^{4} \pi_{\ell,a} N(0, k\sigma_g^2)$	4	Yes, overlapping	$\geq$1
BayesRCπ	$\beta_i \vert A(i) \sim \sum_{a\in A(i)} p_{i,a} \sum_{\ell=1}^{4} \pi_{\ell,a} N(0, k\sigma_g^2)$	4	Yes, overlapping	$\geq$1

where σ_g^2 is the total additive genetic variance, $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$ the mixing proportions such that $\sum_{\ell=1}^{4} \pi_\ell = 1$, $p_{i,a}$ the mixing proportions of SNP i in its set of annotations $A(i)$ such that $\sum_{a\in A(i)} p_{i,a} = 1$, and $k = \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$.

2.2 Gibbs sampler algorithm

As an exact computation of the posterior distribution is intractable for this set of models, Bayesian inference is performed in all cases by obtaining draws from the posterior distribution using a Gibbs sampler. Model parameters are subsequently estimated using the posterior mean across iterations, after excluding a burn-in phase and thinning draws. By default, the Gibbs sampler runs for a total of 50,000 iterations, including 20,000 as a burn-in and a thinning rate of 10.

2.3 Novelty of BayesRCπ and BayesRC+

We developed BayesRCπ and BayesRC+ as an extension of BayesRC to handle cases where prior categorizations of SNPs are overlapping rather than disjointed (i.e., where SNPs can potentially be assigned to multiple categories).

In the case of BayesRCπ, SNP effects are assumed to follow a mixture of mixtures distribution; that we assume that SNPs follow a mixture distribution over their corresponding annotation categories, and within a given annotation in turn, SNP effects are modeled with a 4-component Gaussian mixture distribution as in the BayesR model. Concretely, within a given iteration of the Gibbs sampler used for estimation, SNPs are assigned to the annotation category which maximizes its likelihood given the current estimates of the other model parameters. Note that this step is analogous to that in the standard BayesR algorithm of assigning SNPs to one of the four SNP effect classes based on a likelihood calculation and the current estimates of model parameters.

In the case of BayesRC+, we assume that multiple annotation categories cumulatively impact the estimate of SNP effects; that is, we assume that multiple annotation categories have an additive impact on estimated SNP effects. At each iteration of the Gibbs sampler, the conditional effect of a given SNP is estimated for each of its associated annotation categories in turn, and its total effect is subsequently calculated as the sum over all of its per-annotation effects. Although this assumption of additivity may be strong, it may be useful for avoiding the underestimation of SNP effects in cases where multi-annotated SNPs can be expected to have larger effects than those with a single (or no) annotation category.

3 Download & Compilation

The core of the BayesRCO software is based on version 0.75 of the BayesR software by Moser et al. (2015). As such, a very similar file structure is used:

- *RandomDistributions.f90*: auxiliary file containing various random generator
- *baymodsRCO.f90*: support module for BayesRCO containing common variables and routines (note: unchanged from version 0.75 of BayesR)
- *bayesRCO.f90*: main program

BayesRCO can be compiled with a FORTRAN95 compiler on a Unix operating system using the following command:

```
gfortran RandomDistributions.f90 baymodsRCO.f90 bayesRCO.f90 –o bayesRCO
```

4 Inputs

4.1 Data

BayesRCO requires PLINK binary ped file format. It requires *.bim and *.fam files to determine the number of SNPs and the number of individuals, and a *.bed file for the genotype information.

Genotype data: BayesRCO requires genotypes in PLINK binary format in default-SNP major mode. Since BayesRCO includes all genotypes in the model, samples missing a genotype call cannot simply be omitted. Missing genotypes are replaced by the mean genotype value of a given marker.

Phenotype data: The program reads column 6 as the phenotype column from a PLINK *.fam file. A different phenotype column can be specified by using the `–n [num]` option, where `–n 1` uses the original 6th column (default), `–n 2` uses column 7 and so forth. Missing phenotypes (or phenotypes to be predicted) must be coded as `NA`.

4.2 Prior annotation categories for SNPs

We can represent SNP annotation categories as a binary design matrix, with SNPs in rows and annotation categories in columns. We differentiate two types of annotation matrix, non-overlapping (for BayesRC) or potentially overlapping (for BayesRC+ or BayesRCπ). An example of a **non-overlapping annotation** matrix, such that all SNPs are assigned to a single annotation, is as follows:

$$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ . & . & . \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

An example of an **overlapping** annotation matrix, such that all SNPs are assigned to *at least* one annotation, is as follows:

$$\begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 0 \\ . & . & . \\ 1 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

In the latter example, the first SNP has been categorized as belonging to annotations 2 and 3, while the second SNP has been categorized as belonging only to annotation 2.

As recommended by MacLeod et al. (2016), it is important to have sufficiently large annotation categories ($\geq$ 1000 SNPs) to avoid difficulties to estimate the π_a parameters.

4.3 General Inputs

Input	Description	Default
-bfile	prefix PLINK binary files	None
-out	prefix for output	None
-n	phenotype column	1
-vara	SNP variance prior	0.01
-vare	error variance prior	0.01
-dfvara	degrees of freedom Va	-2.0
-dfvare	degrees of freedom Ve	-2.0
-delta	prior for Dirichlet	1.0
-msize	number of SNPs in reduced update	0
-mrep	number of full cycles in reduced update	5000
-numit	length of MCMC chain	50000
-burnin	burnin steps	20000
-thin	thinning rate	10
-ndist	number of mixture distributions	4
-gpin	effect sizes of mixtures (% x Va)	0.0,0.0001,0.001,0.01
-seed	initial value for random number	0
-predict	perform prediction	f
-snpout	output detailed SNP info	f
-cat	output SNP categories per iteration	None
-beta	output SNP effect per iteration	None
-permute	permute order of SNP	f
-model	model summary file (for prediction)	None
-freq	SNP frequency file (for prediction)	None
-param	SNP effect file (for prediction)	None
-ncat	number of SNP annotations	1
-catfile	SNP annotation matrix file	None
-additive	run BayesRC+	f
-bayesCpi	run BayesCπ	f

4.4 Running BayesRCO

BayesRCO is run in two steps: a first for the training data (and thus the estimation of model parameters) and a second for prediction; we thus use two separate datasets, one including phenotype values, and one without (be careful, the SNPs must match between the two datasets!). We illustrate here how to launch these two features. By default, the software runs a BayesRCπ model.

4.4.1 BayesRCπ

```
path/bayesRCO -bfile [prefix_learning] -out
[prefix_learning] -ncat [number of annotations]
-catfile [annotation_matrix]

path/bayesRCO -bfile [prefix_validation]
-predict -out [prefix_validation] -model
[prefix_learning].model -freq [prefix_learning].frq
-param [prefix_learning].param -ncat [number of
annotations] -catfile [annotation_matrix]
```

4.4.2 BayesRC+

To run BayesRC+, use the flag `-additive` in the training step:

```
path/bayesRCO -bfile [prefix_learning] -out
[prefix_learning] -ncat [number of annotations]
-catfile [annotation_matrix] -additive

path/bayesRCO -bfile [prefix_validation]
-predict -out [prefix_validation] -model
[prefix_learning].model -freq [prefix_learning].frq
-param [prefix_learning].param -ncat [number of
annotations] -catfile [annotation_matrix]
```

4.4.3 BayesRC

As BayesRC is a special case of BayesRCπ or BayesRC+ where no SNPs are assigned to more than one prior annotation category, you can simply run

either of the two previous methods with the appropriate disjoint annotation matrix.

4.4.4 BayesR

As BayesR is a special case of BayesRC with a single prior annotation category to which all SNPs are assigned, it can be run in the same way as for BayesRC using an annotation matrix corresponding to a vector (the same length as the number of SNPs) of 1's. In this case, as the default number of `ncat` is 1, there is no need to specify this option.

4.4.5 BayesCπ

Finally, BayesCπ can be run in a similar manner as for BayesR with the additional flag `-bayesCpi`.

4.4.6 Options

Prior distributions for variance components: Prior inverted-chi squared distribution can be specified for both additive and residual variances (σ^2_g and ^2_e). Scale and degrees of freedom (df) for the variance components are required. Flat (improper) distributions can be specified by setting `df` to -2. It is also possible to specify the heritability of the trait by setting `dfvara` to -3.0 (i.e. `-dfvara -3.0`). In this case the scale parameter is treated as the heritability and the SNP-based variance is set (fixed) to σ^2_g=heritability$\times\sigma^2_y$ (σ^2_y being the phenotypic variance).

Effect size Dirichlet prior (all): The default is to use a uniform and almost uninformative prior for the mixture distribution with a pseudo-observation of 1 (SNP) for each class. Different priors can be specified using the `delta [num]` option. For example, `-delta 3,2,1` specifies a prior with 3, 2 and 1 pseudo-observations for classes 1 to 3 of a 3-component mixture model, `-delta 2` sets the prior to 2 for all mixture components.

Annotation Dirichlet prior (BayesRCπ) For the moment there is no parameter to change the value of the annotation assignment prior. Such an option may be added in future versions.

Mixture model: The BayesR, BayesRC, BayesRC+ and BayesRCπ models assume that the true SNP effect is derived from a series of normal distributions. The default models uses 4 mixture distributions with SNP variances of 0, 0.0001, 0.001 and 0.01, so that the variance (S) of

the j[th] SNP has 4 possible values: S1=0, S2=0.0001$\times\sigma_g^2$, S3=0.001$\times\sigma_g^2$, S4=0.01$\times\sigma_g^2$. Different mixture models can be specified using the –ndist [num] and –gpin [num] options. For example, –ndist 3 –gpin 0.0,0.001,0.05 fits a 3 component mixture with SNP variances S1=0, S2=0.001$\times\sigma_g^2$, S3=0.05$\times\sigma_g^2$.

MCMC sampling: The default is to use a chain length of 50,000 samples (–numit) with the first 20,000 samples (–burnin) being discarded, and using every 10[th] sample (–thin) for posterior inference. To improve mixing, one can use the option –permute to update SNP effects in random order.

5 Outputs

The outputs all have the same name as specified when launching the software, followed by a suffix corresponding to their type, as follows:

5.1 Frequency file

name_output.type: One column containing the SNP allele 2 frequency.

5.2 Log File

The file name prefix is as specified by –out [prefix_training]. The suffix '.log' is appended to give the file name. This is a descriptive file and provides a summary of the run parameters used and the number of records processed.

5.3 Frequency File

Contains allele frequency of the '2' allele. The suffix '.frq' is appended to the prefix. This file is required for scaling and centering genotypes for prediction analysis. The SNP order has to be the same as the genotype input file.

5.4 Model File

The suffix 'model' is appended to the output prefix. This file contains means of the posterior samples of model parameters:

Mean: intercept

Nsnp: number of SNPs in model

Va: genetic variance explained by SNPs (σ_g^2)

Nk1_1,...,Nkk_j: residual variance (σ_e^2)

Pk1_1,...,Pkk_j: proportion of SNPs in mixture component 1 to *k* and annotation 1 to *j*

Vk1_1,...,Vkk_j: sum of squares of SNP effects in mixture component 1 to *k* and annotation 1 to *j*

5.5 Hyperparameter file

The file *prefix.hyp* gives posterior parameter estimates for each MCMC sample:

Replicate: iteration number

Nsnp: number of SNPs in model

Va: genetic variance explained by SNPs

Ve: residual variance

Nk1_1,...,Nkk_j: number of SNPs in mixture components 1 to *k* and annotation 1 to *j*

Pk1_1,...,Pkk_j: proportion of SNPs in mixture component 1 to *k* and annotation 1 to *j*

Vk1_1,...,Vkk_j: sum of squares of SNP effects in mixture component 1 to *k* and annotation 1 to *j*

5.6 Parameter file

The suffix 'param' is appended to the output prefix. The SNP order is the same as the genotype input file. This file contains mean posterior estimates for each individual SNP:

PIP_1, ..., PIP_k: Posterior inclusion probabilities of the SNP in mixture classes 1 to *k*

beta: posterior SNP effect

PAIP_1, ..., PAIP_j: Posterior annotation inclusion probabilities of the SNP in annotation 1 to *j* (useful for BayesRCπ, otherwise gives the annotations each SNP belong to)

Vbeta: variance of the posterior SNP effects across iterations

Vi: posterior variance of the SNP effects

5.7 Genetic value file

This file outputs the predicted genomic values (GVs). The output prefix is used to give the file name *prefix.gv*.

5.8 Optional files

`-snpout`: provide output in sparse format `mixture class:SNP:effect size`. The SNP number (SNP #) corresponds to the row number of the SNP in the PLINK *.bim file.

`-cat`: output SNP categories per iteration

`-beta`: output SNP effect per iteration

References

M. Erbe, B. Hayes, L. Matukumalli, S. Goswami, P. Bowman, C. Reich, B. Mason, and M. Goddard. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science*, 95(7):4114–4129, July 2012. ISSN 00220302. doi: 10.3168/jds.2011-5019. URL https://linkinghub.elsevier.com/retrieve/pii/S0022030212003918.

D. Habier, R. L. Fernando, K. Kizilkaya, and D. J. Garrick. Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, 12 (1):186, Dec. 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-186. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-186.

I. M. MacLeod, P. J. Bowman, C. J. Vander Jagt, M. Haile-Mariam, K. E. Kemper, A. J. Chamberlain, C. Schrooten, B. J. Hayes, and M. E. Goddard. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics*, 17

(1):144, Dec. 2016. ISSN 1471-2164. doi: 10.1186/s12864-016-2443-6. URL http://www.biomedcentral.com/1471-2164/17/144.

G. Moser, S. H. Lee, B. J. Hayes, M. E. Goddard, N. R. Wray, and P. M. Visscher. Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. *PLOS Genetics*, 11(4):e1004969, Apr. 2015. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004969. URL https://dx.plos.org/10.1371/journal.pgen.1004969.

Funding

This work is part of the GENE-SWitCH project that has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement number 817998.

Appendix B

WCGALP short paper

Capitalizing on complex annotations in Bayesian genomic prediction for a backcross population of growing pigs

F. Mollandin[1*], H. Gilbert[2], P. Croiseau[1] and A. Rau[1,3]

[1] INRAE, AgroParisTech, GABI, Université Paris-Saclay, Jouy-en-Josas 78350, France;
[2] GenPhySE, Université de Toulouse, INRAE, ENVT, Castanet Tolosan 31320, France;
[3] BioEcoAgro Joint Research Unit, INRAE, Université de Liège, Université de Lille, Université de Picardie Jules Verne, Estrées-Mons 80203, France ; [*]fanny.mollandin@inrae.fr

Abstract

Prior biological information has the potential to guide and inform genomic prediction models, but the BayesRC approach is currently limited to the use of disjoint categorizations of genetic markers. We propose two novel Bayesian approaches to model cumulative (BayesRC+) or preferential (BayesRCπ) contributions of multiple biological categories for multi-annotated SNPs. We illustrate the performance of these approaches on data from a backcross population of growing pigs in conjunction with several different sets of annotations related to multiple production traits constructed using the PigQTLdb. On the two traits predicted, ADG and BFT, we observed improved prediction quality on ADG (up to 1.7-gain point) with both BayesRCpi and BayesRC+, and suitable annotation set.

Introduction

In plant and animal breeding, genomic prediction models have been widely developed and deployed in recent years to predict polygenic traits using genetic variants, typically single nucleotide polymorphisms (SNP). An interesting and potentially useful approach to improve upon existing genomic prediction models is to combine the use of genotype and phenotype data with prior biological information to better guide models. Most routinely used genomic prediction models are based on linear models, including notably genomic best linear unbiased prediction. Another family of models, known as the Bayesian alphabet (Habier *et al.*, 2011), uses a flexible set of assumptions about how individual SNPs contribute to the overall genomic variance. Among these, BayesR (Erbe *et al.*, 2012) assumes SNP effects arise from one of four groups (null, small, medium, or large variance) and has been shown to perform well for both prediction and quantitative trait loci (QTL) mapping (Moser *et al.*, 2015; Mollandin *et al.*, 2021). BayesRC extends BayesR to further incorporate prior biological information in the form of disjoint annotation categories (MacLeod *et al.*, 2016), but SNPs can only be assigned to a single annotation category. There thus remains a need for genomic prediction models able to capitalize on annotations of greater complexity, in particular those for which SNPs may potentially be assigned to multiple categories.

In this work, we present two novel extensions to BayesRC to deal with such complex, overlapping annotations, and we illustrate their utility on data from an experimental backcross population in growing pigs. This project is part of EuroFAANG (https://eurofaang.eu), a synergy of five Horizon 2020 projects that share the common goal to discover links between genotype to phenotype in farmed animals and meet global Functional Annotation of ANimal Genomes (FAANG) objectives.

Materials & Methods

Bayesian genomic prediction with complex, overlapping annotations. The general statistical model for genomic prediction can be defined as

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{X\beta} + \boldsymbol{e}, \quad e_i \sim N(0, \sigma_e^2) \tag{1}$$

where $\boldsymbol{y}$ is a vector of phenotypes, $\boldsymbol{\mu}$ an intercept, $\boldsymbol{\beta}$ the vector of SNP effects, $\boldsymbol{X}$ the centered and scaled marker matrix, and σ_e^2 the variance of the residuals $\boldsymbol{e}$. We further assume that $\boldsymbol{C} = (C_{i,j})$ denotes annotation categories, such that $C_{i,j} = 1$ if SNP *i* is included in category *j* and 0 otherwise.

Using the four-component mixture of BayesR as a base, we propose two alternative models to account for overlapping annotations (where $\sum_j C_{i,j} > 1$ for some *i*). The first, BayesRCπ, defines a mixture-of-mixtures prior for SNP effects to assign multi-annotated markers to the single annotation category that maximizes its conditional likelihood:

$$\beta_i \sim \sum_{j \in C_{i,j}=1} p_{i,j}\left(\pi_{1,j}\delta(0) + \pi_{2,j}N(0, 10^{-4}\sigma_g^2) + \pi_{3,j}N(0, 10^{-3}\sigma_g^2) + \pi_{4,j}N(0, 10^{-2}\sigma_g^2)\right), \tag{2}$$

such that $\delta(0)$ represents the dirac function at 0, $\sum_k \pi_{k,j} = 1$ for all annotations *j*, σ_g^2 the total additive genetic variance, and $p_{i,j}$ the annotation mixing parameter with $\sum_j p_{i,j} = 1$ for all *i*. The second, BayesRC+, instead defines a cumulative mixture prior across categories for SNP effects:

$$\beta_i \sim \sum_{j \in C_{i,j}=1}\left(\pi_{1,j}\delta(0) + \pi_{2,j}N(0, 10^{-4}\sigma_g^2) + \pi_{3,j}N(0, 10^{-3}\sigma_g^2) + \pi_{4,j}N(0, 10^{-2}\sigma_g^2)\right). \tag{3}$$

In both models, all mixing proportions are assumed to follow flat Dirichlet priors and σ_g^2 an inverse χ^2 prior. A Gibbs sampler is used for inference as posterior distributions are not tractable. Both BayesRCπ and BayesRC+ have been implemented in the BayesRCO software in Fortran; additional details can be found in the User's Guide (https://github.com/fmollandin/BayesRCO).

Genotype and phenotype data from a backcross pig population. A backcross (BC) population between Large White (LW; 3/4) and Creole (CR;1/4) pigs was established as previously described (Gourdine *et al.*, 2019). BC (*n* = 1,297 from 130 LW sows) growing pigs raised in two environments (tropical and temperate) were related via genetically related sows sired with the same 10 F1 × CR LW boars. A common trait recording protocol was used in the two environments for phenotypic data. Phenotypes were pre-corrected for age, sex, and farm; we focus here on measures at 23 weeks for backfat thickness (BFT) and average daily weight gain (ADG). Animals were genotyped using the Illumina Porcine 60k BeadChip array; markers with minor allele frequencies greater than 0.01 were retained for the analysis (corresponding to 46,908 and 46,881 markers for ADG and BFT, respectively). To establish the potential impact of our models on prediction accuracy, we used a sibling-structured 10-fold cross validation procedure. For the descendants

from each sire in turn, we calculated the correlation between their observed corrected phenotypes and those predicted from models constructed on the descendants of the remaining 9 sires; validation correlations were averaged across the ten folds.

PigQTLdb annotations. Animal QTLdb (https://www.animalgenome.org/QTLdb) groups together curated results from genotype-phenotype association studies in several livestock species (Hu *et al.*, 2021). Cross-experiment QTL data from PigQTLdb (Release 45; SS11.1) for traits relevant to pig production were downloaded for eleven trait sub-hierarchy categories (anatomy, behavioral, blood parameters, conformation, fatness, fatty acid content, feed conversion, fowth, immune capacity, litter traits, reproductive organs). An additional "other" category was created for markers not included in PigQTLdb. Genotyped markers in our data were subsequently assigned to one or more annotation categories using three different strategies: (1) using the position of known PigQTLdb markers; (2) using the extended position of known PigQTLdb markers, including the nearest up- and downstream neighbors ("extended PigQTLdb"); and (3) using the extended position of known PigQTLdb markers as before, where neighboring markers were allowed ambiguous assignment to both trait-specific and "other" categories ("fuzzy extended PigQTLdb"). In the three annotation construction strategies, 1.3%, 4.9% and 17.7% of markers were respectively assigned to two or more categories.

Results

We compared the prediction accuracy of BayesRCπ and BayesRC+ to that of BayesR (ignoring annotation categories) and BayesRC (where a single category is allowed per marker). For the latter, multi-annotated SNPs were randomly assigned to a single category. We notably observed different trends for the two traits (Table 1). For ADG, we remark a loss in prediction accuracy compared to BayesR for all annotation-based models with straightforward pigQTLdb annotations; however, extending these annotations to include neighboring markers (extended and fuzzy extended pigQTLdb) led to improvements in prediction quality, with a 1.7-point gain in correlation for BayesRCπ with extended annotations. On the other hand, for BFT the use of annotations, regardless of how they are constructed, did not appear to lead to a marked improvement in prediction. This suggests that categorizations constructed from PigQTLdb contribute little pertinent information for the genomic prediction of BFT in our data, or that environment-dependent categories should be added to the model to account for the significant GxE affecting this trait (Gourdine *et al.*, 2019).

Discussion

In this work we have proposed two new approaches, BayesRCπ and BayesRC+, to fully capitalize on complex, overlapping annotations in genomic prediction. Both methods showed promise for incorporating partially overlapping categories from pigQTLdb in genomic prediction for a growing pig population, although a gain in predictive accuracy was observed for only one (ADG) of the two traits considered here. We also compared three strategies for constructing prior biological

categories by extending pigQTLdb annotations in various ways to include neighboring markers, which has the potential to better exploit linkage disequilibrium around relevant markers. Taken together, these results suggest that the incorporation of complex annotations can lead to modest gains in prediction performance in some cases, even for moderate marker density SNP chips, but such gains depend strongly on the choice and construction of annotations and are unlikely to be universal across traits.

Table 1. Validation correlation for two traits in pig data for BayesRCπ and BayesRC+ with different annotation strategies, as compared to BayesR and BayesRC.

Method	Annotations	ADG Mean (SD)	BFT Mean (SD)
BayesR	—	0.213 (± 0.081)	0.265 (± 0.161)
BayesRC	PigQTLdb (random)	0.200 (± 0.105)	0.265 (± 0.159)
	Extended pigQTLdb (random)	0.225 (± 0.098)	0.258 (± 0.157)
BayesRCπ	PigQTLdb	0.200 (± 0.100)	0.266 (± 0.157)
	Extended pigQTLdb	0.229 (± 0.095)	0.254 (± 0.162)
	Fuzzy extended pigQTLdb	0.226 (± 0.096)	0.262 (± 0.159)
BayesRC+	PigQTLdb	0.207 (± 0.097)	0.273 (± 0.163)
	Extended pigQTLdb	0.227 (± 0.095)	0.271 (± 0.158)

This work is part of the GENE-SWitCH project that has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement n° 817998. The financial support of the French National Agency of Research (ANR PigHeaT, ANR-12-ADAP-0015) is also gratefully acknowledged.

References

Erbe, M., Hayes, B.J., Matukumalli, L.K., Goswami, S., Bowman, P.J., *et al*. (2012) Journal of Dairy Science 95(7): 4114–29. https://doi.org/10.3168/jds.2011-5019

Gourdine, J.-L., Riquet, J., Rosé, R. Poullet, N., Giorgi, M. *et al.* (2019) Journal of Animal Science 97(9): 3699–3713. https://doi.org/10.1093/jas/skz245

Habier, D., Fernando, R.L., Kizilkaya, K., and Garrick, D.J. (2011). BMC Bioinformatics 12(1): 186. https://doi.org/10.1186/1471-2105-12-186

Hu, Z.-L., Park, C.A., and Reecy, J.M. (2021) Nucleic Acids Research, gkab1116. https://doi.org/10.1093/nar/gkab1116

MacLeod, I.M., Bowman, P.J., Vander Jagt, C.J., Haile-Mariam, M., Kemper, K.E., *et al.* (2016) BMC Genomics. 17(1): 144. https://doi.org/10.1186/s12864-016-2443-6

Mollandin, F, Rau, A., and Croiseau, P. (2021) G3 Genes|Genomes|Genetics 11(11): jkab225. https://doi.org/10.1093/g3journal/jkab225

Moser, G., Lee, S.H., Hayes, B.J., Goddard, M.E., Wray, N.R. *et al*. (2015) PLOS Genetics 11(4): e1004969. https://doi.org/10.1371/journal.pgen.1004969

Bibliography

Abdollahi-Arpanahi, R., Morota, G., Valente, B. D., Kranis, A., Rosa, G. J., and Gianola, D. (2016). Differential contribution of genomic regions to marked genetic variation and prediction of quantitative traits in broiler chickens. *Genetics Selection Evolution*, 48(1):1–13.

Acloque, H. e. a. (July 2022). Extensive functional genomics information from early developmental time points for pig and chicken. *12th World Congress on Genetics Applied to Livestock Production*.

Aguilar, I., Misztal, I., Johnson, D., Legarra, A., Tsuruta, S., and Lawlor, T. (2010). Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of holstein final score. *Journal of dairy science*, 93(2):743–752.

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203.

Bayes, T. (1763). Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s. *Philosophical transactions of the Royal Society of London*, (53):370–418.

Bellot, P., de los Campos, G., and Pérez-Enciso, M. (2018). Can Deep Learning Improve Genomic Prediction of Complex Human Traits? *Genetics*, 210(3):809–819.

Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):113–128.

Calus, M. P., Bouwman, A. C., Schrooten, C., and Veerkamp, R. F. (2016). Efficient genomic prediction based on whole-genome sequence data using split-and-merge bayesian variable selection. *Genetics Selection Evolution*, 48(1):1–19.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):s13742–015.

Cirulli, E. T. and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, 11(6):415–425.

Crespo-Piazuelo, D., e. a. (July 2022). Deciphering genetic variants from whole genome affecting duodenum, liver and muscle transcriptomes in pigs. *12th World Congress on Genetics Applied to Livestock Production*.

Dickerson, G. and Hazel, L. (1944). Effectiveness of selection on progeny performance as a supplement to earlier culling. *J. agric. Res.*, 69:459.

Do, D. N., Janss, L. L., Jensen, J., and Kadarmideen, H. N. (2015). Snp annotation-based whole genomic prediction and selection: an application to feed efficiency and its component traits in pigs. *Journal of Animal Science*, 93(5):2056–2063.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21.

ENSEMBL (2022). Gene: Supt3h, ensssscg00000001709.

Erbe, M., Hayes, B., Matukumalli, L., Goswami, S., Bowman, P., Reich, C., Mason, B., and Goddard, M. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science*, 95(7).

Fisher, R. (1918). The correlation between relatives on the supposition of mendelian inheritance. trans. roy. soc.

Flint, J., Timpson, N., and Munafò, M. (2014). Assessing the utility of intermediate phenotypes for genetic mapping of psychiatric disease. *Trends in neurosciences*, 37(12):733–741.

Gao, N., Li, J., He, J., Xiao, G., Luo, Y., Zhang, H., Chen, Z., and Zhang, Z. (2015). Improving accuracy of genomic prediction by genetic architecture based priors in a bayesian model. *BMC genetics*, 16(1):1–11.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741.

Gianola, D., de Los Campos, G., Hill, W. G., Manfredi, E., and Fernando, R. (2009). Additive genetic variability and the bayesian alphabet. *Genetics*, 183(1):347–363.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC press.

Giuffra, E., Tuggle, C. K., and Consortium, F. (2019). Functional annotation of animal genomes (faang): current achievements and roadmap. *Annual review of animal biosciences*, 7:65–88.

Gourdine, J.-L., Riquet, J., Rosé, R., Poullet, N., Giorgi, M., Billon, Y., Renaudeau, D., and Gilbert, H. (2019). Genotype by environment interactions for performance and thermoregulation responses in growing pigs. *Journal of animal science*, 97(9):3699–3713.

Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., and Su, G. (2014). Comparison of single-trait and multiple-trait genomic prediction models. *BMC genetics*, 15(1):1–7.

Habier, D., Fernando, R. L., and Dekkers, J. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389–2397.

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC bioinformatics*, 12(1):1–12.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1).

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hu, H., Campbell, M. T., Yeats, T. H., Zheng, X., Runcie, D. E., Covarrubias-Pazaran, G., Broeckling, C., Yao, L., Caffe-Treml, M., Gutiérrez, L., et al. (2021). Multi-omics prediction of oat agronomic and seed nutritional traits across environments and in distantly related populations. *Theoretical and Applied Genetics*, 134(12):4043–4054.

Hu, Z.-L., Park, C. A., and Reecy, J. M. (2022). Bringing the animal qtldb and corrdb into the future: meeting new challenges and providing updated services. *Nucleic acids research*, 50(D1):D956–D961.

Kahn, S. D. (2011). On the future of genomic data. *science*, 331(6018):728–729.

Kemper, K. E., Bowman, P. J., Hayes, B. J., Visscher, P. M., and Goddard, M. E. (2018). A multi-trait bayesian method for mapping qtl and genomic prediction. *Genetics Selection Evolution*, 50(1):1–13.

Kruglyak, L. and Nickerson, D. A. (2001). Variation is the spice of life. *Nature genetics*, 27(3):234–236.

Li, B. and Dewey, C. N. (2011). Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):1–16.

Li, H. (2012). Exploring single-sample snp and indel calling with whole-genome de novo assembly. *Bioinformatics*, 28(14):1838–1844.

Link, W. A. and Eaton, M. J. (2012). On thinning of chains in mcmc. *Methods in ecology and evolution*, 3(1):112–115.

MacLeod, I. M., Bowman, P. J., Vander Jagt, C. J., Haile-Mariam, M., Kemper, K. E., Chamberlain, A. J., Schrooten, C., Hayes, B. J., and Goddard, M. E. (2016). Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics*, 17(1):144.

MacLeod, I. M., Hayes, B. J., and Goddard, M. E. (2014). The effects of demography and long-term selection on the accuracy of genomic prediction with sequence data. *Genetics*, 198(4):1671–1684.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*, volume 38. M. Dekker New York.

Merks, J., Mathur, P., and Knol, E. (2012). New phenotypes for new breeding goals in pigs. *Animal*, 6(4):535–543.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Meuwissen, T. and Goddard, M. (2010). Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics*, 185(2):623–631.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, page 11.

Misztal, I., Legarra, A., and Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *Journal of dairy science*, 92(9):4648–4655.

Mollandin, F., Gilbert, H., Croiseau, P., and Rau, A. (2022). Accounting for overlapping annotations in genomic prediction models of complex traits. *BMC Bioinformatics*, 23(1):365.

Montesinos-López, O. A., Montesinos-López, A., Pérez-Rodríguez, P., Barrón-López, J. A., Martini, J. W., Fajardo-Flores, S. B., Gaytan-Lugo, L. S., Santana-Mancilla, P. C., and Crossa, J. (2021). A review of deep learning applications for genomic selection. *BMC genomics*, 22(1):1–23.

Morota, G., Abdollahi-Arpanahi, R., Kranis, A., and Gianola, D. (2014). Genome-enabled prediction of quantitative traits in chickens using genomic annotation. *BMC genomics*, 15(1):1–10.

Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2015). Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. *PLOS Genetics*, 11(4):e1004969.

Nayeri, S., Sargolzaei, M., and Tulpan, D. (2019). A review of traditional and machine learning methods applied to animal breeding. *Animal Health Research Reviews*, 20(1):31–46.

Noell, G., Faner, R., and Agustí, A. (2018). From systems biology to p4 medicine: applications in respiratory medicine. *European Respiratory Review*, 27(147).

Ogutu, J. O., Schulz-Streeck, T., and Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. In *BMC proceedings*, volume 6, pages 1–6. Springer.

Ojavee, S. E., Kousathanas, A., Trejo Banos, D., Orliac, E. J., Patxot, M., Läll, K., Mägi, R., Fischer, K., Kutalik, Z., and Robinson, M. R. (2021). Genomic architecture and prediction of censored time-to-event phenotypes with a bayesian genome-wide analysis. *Nature communications*, 12(1):1–17.

Ojavee, S. E., Maksimova, E. S., Läll, K., Sadler, M. C., Mägi, R., Kutalik, Z., and Robinson, M. R. (2022). Novel discoveries and enhanced genomic prediction from modelling genetic risk of cancer age-at-onset. *medRxiv*.

Preston, G. A. and Weinberger, D. R. (2022). Intermediate phenotypes in schizophrenia: a selective review. *Dialogues in clinical neuroscience*, 7(2):165–179.

Pritchard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics*, 69(1):1–14.

Qanbari, S. (2020). On the extent of linkage disequilibrium in the genome of farm animals. *Frontiers in Genetics*, 10:1304.

Ramstein, G. P., Evans, J., Kaeppler, S. M., Mitchell, R. B., Vogel, K. P., Buell, C. R., and Casler, M. D. (2016). Accuracy of genomic prediction in switchgrass (panicum virgatum l.) improved by accounting for linkage disequilibrium. *G3: Genes, Genomes, Genetics*, 6(4):1049–1062.

Sellier, P., Boichard, D., and Verrier, E. (2019). La génétique animale à l 'inra. soixante ans d'une histoire scientifique en prise avec le monde de la sélection et riche en rebondissements technologiques. *Histoire de la recherche contemporaine. La revue du Comité pour l'histoire du CNRS*, 8(1):86–97.

Shepherd, R. K., Meuwissen, T. H., and Woolliams, J. A. (2010). Genomic selection and complex trait prediction using a fast em algorithm applied to genome-wide markers. *Bmc Bioinformatics*, 11(1):1–12.

Speed, D., Cai, N., Johnson, M. R., Nejentsev, S., and Balding, D. J. (2017). Reevaluation of snp heritability in complex human traits. *Nature genetics*, 49(7):986–992.

Technow, F. and Melchinger, A. E. (2013). Genomic prediction of dichotomous traits with bayesian logistic models. *Theoretical and applied genetics*, 126(4):1133–1143.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Usai, M. G., Goddard, M. E., and Hayes, B. J. (2009). Lasso with cross-validation for genomic selection. *Genetics research*, 91(6):427–436.

Vailati-Riboni, M., Palombo, V., and Loor, J. J. (2017). What are omics sciences? In *Periparturient diseases of dairy cows*, pages 1–7. Springer.

van Dijk, A. D. J., Kootstra, G., Kruijer, W., and de Ridder, D. (2021). Machine learning in plant science and plant breeding. *Iscience*, 24(1):101890.

Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C., and Sölkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in genetics*, 4:270.

Wang, S., Wei, J., Li, R., Qu, H., Chater, J. M., Ma, R., Li, Y., Xie, W., and Jia, Z. (2019). Identification of optimal prediction models using multi-omic data for selecting hybrid rice. *Heredity*, 123(3):395–406.

Wang, T. (2016). *Computationally Efficient Genomic Prediction From Whole Genome Sequence Data In Dairy Cattle*. PhD thesis, La Trobe University.

Wang, T., Chen, Y.-P. P., Bowman, P. J., Goddard, M. E., and Hayes, B. J. (2016). A hybrid expectation maximisation and MCMC sampling algorithm to implement Bayesian mixture model based genomic prediction and QTL mapping. *BMC Genomics*, 17(1):744.

Wang, T., Chen, Y.-P. P., Goddard, M. E., Meuwissen, T. H., Kemper, K. E., and Hayes, B. J. (2015). A computationally efficient algorithm for genomic prediction using a Bayesian model. *Genetics Selection Evolution*, 47(1):34.

Wittkopp, P. J., Haerum, B. K., and Clark, A. G. (2004). Evolutionary changes in cis and trans gene regulation. *Nature*, 430(6995):85–88.

Xu, Y., Xu, C., and Xu, S. (2017). Prediction and association mapping of agronomic traits in maize using multiple omic data. *Heredity*, 119(3):174–184.

Zeng, J., Xue, A., Jiang, L., Lloyd-Jones, L. R., Wu, Y., Wang, H., Zheng, Z., Yengo, L., Kemper, K. E., Goddard, M. E., et al. (2021). Widespread signatures of natural selection across human complex traits and functional genomic categories. *Nature communications*, 12(1):1–12.

Zhao, T., Fernando, R., Garrick, D., and Cheng, H. (2020). Fast parallelized sampling of bayesian regression models for whole-genome prediction. *Genetics Selection Evolution*, 52(1):1–11.

Zhou, L., Mrode, R., Zhang, S., Zhang, Q., Li, B., and Liu, J.-F. (2018). Factors affecting gebv accuracy with single-step bayesian models. *Heredity*, 120(2):100–109.

Zhu, X. and Stephens, M. (2018). Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nature communications*, 9(1):1–14.

Zingaretti, L. M., Gezan, S. A., Ferrão, L. F. V., Osorio, L. F., Monfort, A., Muñoz, P. R., Whitaker, V. M., and Pérez-Enciso, M. (2020). Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. *Frontiers in plant science*, 11:25.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.