



UNIL | Université de Lausanne

Unicentre

CH-1015 Lausanne

<http://serval.unil.ch>

Year : 2022

Genetic basis of common complex traits

Patxot Bertran Marion

Patxot Bertran Marion, 2022, Genetic basis of common complex traits

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive <http://serval.unil.ch>

Document URN : urn:nbn:ch:serval-BIB_4C2394C34F245

Droits d'auteur

L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

Copyright

The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.



UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Département de Biologie Computationnelle

Genetic basis of common complex traits

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Marion Patxot Bertran

Maîtrise universitaire en sciences moléculaires du vivant, Université de Lausanne

Jury

Prof. Curdin Conrad, Président
Prof. Zoltán Kutalik, Directeur de thèse
Prof. Matthew Robinson, Co-directeur de thèse
Prof. Rachel Freathy, Experte
Prof. Sven Bergmann, Expert

Lausanne
(2022)



UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Département de Biologie Computationnelle

Genetic basis of common complex traits

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Marion Patxot Bertran

Maîtrise universitaire en sciences moléculaires du vivant, Université de Lausanne

Jury

Prof. Curdin Conrad, Président
Prof. Zoltán Kutalik, Directeur de thèse
Prof. Matthew Robinson, Co-directeur de thèse
Prof. Rachel Freathy, Experte
Prof. Sven Bergmann, Expert

Lausanne
(2022)



UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Ecole Doctorale

Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président·e	Monsieur	Prof.	Curdin	Conrad
Directeur·trice de thèse	Monsieur	Prof.	Zoltán	Kutalik
Co-directeur·trice	Monsieur	Prof.	Matthew	Robinson
Expert·e·s	Madame	Prof.	Rachel	Freathy
	Monsieur	Prof.	Sven	Bergmann

le Conseil de Faculté autorise l'impression de la thèse de

Marion Patxot Bertran

Maîtrise universitaire ès Sciences en sciences moléculaires du vivant, Université de Lausanne

intitulée

Genetic basis of common complex traits

Lausanne, le 13 juillet 2022

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Curdin Conrad

*Pour ce qui est de l'avenir,
il ne s'agit pas de le prévoir mais de le rendre possible.*

— Antoine de Saint-Exupéry

Acknowledgments

Le doctorat est une intrigue où on fait les cent pas, où on cherche l'essentiel, où on rit et on pleure aussi. Après quatre ans, je suis fière d'avoir vécu cette histoire. Ici, je tiens à remercier les personnes avec qui je l'ai partagé et qui m'ont aidé tout au long de l'aventure.

First, I am grateful to my supervisor **Matthew** for giving me the opportunity to embark on this journey. Thank you for being my mentor and for your guidance. I have learned a lot working with you. I also thank past and present members of the group. **Thanasis** and **Daniel** for their advice. **Alex** for his good humour and Australian accent. And especially **Sven**. You are a passionate, kind and brilliant person. Tu t'es adapté à la vie lausannoise comme un poisson dans l'eau et je suis heureuse d'avoir partagé ce chemin avec toi. Thank you for the countless explanations, the beers by the lake and for being a good friend. I am also very grateful to **Zoltan** for his support over the past two years. Thank you to the members of his group. It was a real pleasure to work with all of you. I would like to specifically thank **Ninon** and **Liza** for their warm welcome, kindness and the Schrödinger's altar, a real highlight. **Chiara** et **Marie**, pour leur entrain et leur compétitivité aux jeux de cartes. And, **Eleonora** for her advice and for comforting me at a difficult time in my PhD. Last but not least, I am also thankful to the **UK Biobank participants** as well as **David**, **Milos** and especially **Rosanna** without whom I would not have been able to work on genetics in maternal health.

Ensuite, je tiens à remercier **mes amis** pour tous leurs encouragements sans jamais vraiment comprendre mon sujet. L'équipe de Touch Rugby des **Lakers**, playing with you always gives me lots of joy and energy to keep going in life. **Flavia**, avec qui j'ai commencé cette aventure en 1ère année de Bachelor à l'UNIL. You are the most positive person I know, finding joy in the smallest things. Un vero raggio di sole. Merci pour ton écoute, nos CP et MB du mercredi, le top du top. J'attends avec impatience nos prochaines aventures post-PhD. Je tiens également à remercier **Nicolas** pour sa bonne humeur et ses blagues très astucieuses du mercredi. Ma famille, **mes parents et mes soeurs Anna, Juliette et Ines**, pour leur soutien inconditionnel peu importe la situation. Pour tous les appels et les vacances au soleil. I uppala, aquest doctorat s'està acabant i no puc esperar per celebrar-ho amb vosaltres. **Yann**, no habría terminado esta aventura sin ti. Tu en as vu de toutes les couleurs pendant ce doctorat et je te remercie de m'avoir soutenu et challengé dans toutes mes décisions. De m'avoir rappelé que pour aller de l'avant, il faut aussi prendre le temps de s'amuser. Merci pour tous les rires, les high-fives motivants et pour ta joie de vivre. Merci d'être toi et de faire les cent pas avec moi.

Abstract

Common complex traits are driven by multiple genetic and environmental factors. The genetic basis of such traits can be studied through their genetic architecture, which aims to describe the underlying mechanism involved in the creation of phenotypic variation within the population. Despite efforts to estimate the contribution of genomic regions to complex trait variation, the distribution of effect sizes across functional annotations remains unknown. Most studies estimate enrichment between annotations in downstream analysis following genome-wide association study rather than using functional information to assess enrichment conditional on the rest of the genome. In this thesis, I present BayesRR-RC, a scalable Bayesian model that utilizes genomic annotations and individual-level data to jointly estimate marker effects while accounting for correlation among genetic markers. I then introduce the CHUV Maternity Cohort, a unique cohort with haematological longitudinal measures to study maternal health and maternal-fetal outcomes at delivery. The BayesRR-RC model is firstly applied to explore the genetic architecture of height, body mass index, type-2-diabetes and coronary artery disease in the UK Biobank and secondly, to predict four major pregnancy-related complications in the CHUV Maternity Cohort. This work provides LD-unbiased estimates of annotation enrichment, determines which genomic regions are influential and improves disease risk prediction. It also provide a comprehensive description of the haematological changes that occur in pregnancies from the CHUV Maternity Cohort to improve differentiation between normal physiological changes and disease pathology in pregnancy. Together, these studies contribute to the improvement of quantitative genetic methods and their application in maternal health, which may lead to promising advancements in personalised medicine.

Résumé

Les traits complexes communs sont déterminés par de multiples facteurs génétiques et environnementaux. La base génétique de ces traits peut être étudiée à travers leur architecture génétique, qui vise à décrire le mécanisme impliqué dans la création de la variation phénotypique au sein de la population. Malgré les efforts déployés pour estimer la contribution des régions génomiques à la variation des traits complexes, la distribution des tailles d'effets entre les annotations fonctionnelles reste inconnue. La plupart des études estiment l'enrichissement entre les annotations avec des analyses à la suite d'une étude d'association pangénomique plutôt que d'utiliser des informations fonctionnelles pour évaluer l'enrichissement conditionnel au reste du génome. Dans cette thèse, je présente BayesRR-RC, un modèle bayésien qui utilise les annotations génomiques et les données au niveau individuel pour estimer conjointement les effets des marqueurs, tout en tenant compte de la corrélation entre les marqueurs génétiques. Je présente ensuite la Cohorte de la Maternité du CHUV, une cohorte unique avec des mesures longitudinales hématologiques pour étudier la santé maternelle pendant la grossesse et à l'accouchement. Le modèle BayesRR-RC est premièrement appliqué pour explorer l'architecture génétique de la taille, de l'indice de masse corporelle, du diabète de type 2 et de la maladie coronarienne dans la UK Biobank. Il est ensuite appliqué pour prédire quatre complications majeures liées à la grossesse dans la Cohorte de la Maternité du CHUV. Cette recherche apporte des estimations non biaisées de l'enrichissement des annotations, détermine quelles régions génomiques contribuent et améliore la prédiction du risque de maladie. Elle apporte également une description complète des changements hématologiques dans la cohorte de maternité du CHUV afin d'améliorer la différenciation entre les changements physiologiques normaux et pathologiques pendant la grossesse. Ensemble, ces études contribuent à l'amélioration des méthodes de génétique quantitative et à leur application à la santé maternelle, ce qui pourrait conduire à des avancées prometteuses en médecine personnalisée.

List of Abbreviations

General abbreviations

1000G : 1000 Genomes Project

CBC : Complete blood count

CER-VD : Commission cantonale d'éthique de la recherche sur l'être humain

CHUV : Centre Hospitalier Universitaire Vaudois

GMRM : Bayesian grouped mixture of regressions model

GRM : Genomic relationship matrix

GWAS : Genome-wide association studies

HapMap : Haplotype Mapping Project

HRC : Haplotype Reference Consortium

LD : Linkage disequilibrium

LDSC : LD score regression

MAF : Minor allele frequency

MCMC : Markov chain Monte Carlo

MR : Mendelian randomisation

QTL : Quantitative trait loci

REML : Restricted maximum-likelihood

s-LDSC : stratified LD score regression

SNP : Single nucleotide polymorphism

TOPMed : Trans-Omics for Precision Medicine

TWAS : Transcriptome-wide association study

PCA : Principal component analysis

PDR : Pleiotropic decomposition regression

PPWV : Posterior probability of window variance

PRS : Polygenic risk score

Complex trait abbreviations

BASO : Basophil count

BMD : Heel bone mineral density T-score

BMI : Body mass index

BP : Blood pressure

CAD : Cardiovascular disease outcomes

CHOL : Cholesterol

CRET : Creatinine
DBP : Diastolic blood pressure
EOSI : Eosinophil count
FVC : Forced vital capacity
GDM : Gestational diabetes mellitus
GLU : Glucose
HAC : Haematocrit count
HbA1c : Glycated haemoglobin
HDL : High-density lipoprotein cholesterol
HDP : Hypertensive disorders of pregnancy
HMC : Haemoglobin count
HT : Height
LDL : Direct low-density lipoprotein cholesterol
LYMPH : Lymphocyte count
MCH : Mean corpuscular haemoglobin
MCHC : Mean corpuscular haemoglobin concentration
MCV : Mean corpuscular volume
MDD : Major depression disorder
MONO : Monocyte count
MPV : Mean platelet volume
NEUT : Neutrophil count
PA : placental abruption
PIH : Pregnancy-induced hypertension
PLATE : Platelet count
PPH : Post-partum hemorrhage
PTB : Preterm birth
RBC : Red blood cell count
RCDW : Red cell distribution width
SBP : Systolic blood pressure
T2D : Type-2-diabetes
WBC : White blood cell count

Contents

Introduction	2
Genetics of complex traits	2
Human genome sequencing	3
GWAS and the genetic architecture of traits	5
Explicitly modelling LD	8
Fitting SNPs jointly	11
From association studies to health care	13
Pregnancy and maternal health	14
Chapter 1	17
Bayesian penalized regression for complex trait analysis	17
BayesRR-RC model validation	18
Application to complex traits in the UK Biobank	19
Chapter 2	21
Maternal cohort study	21
Haematological changes in low and high risk pregnancies	21
Genetic risk prediction of maternal complex diseases	22
Methods	22
DNA extraction and SNP genotyping	22
Quality control of the genotype data	22
Genotype imputation	23
Prediction into the Maternity CHUV cohort	24
Results	28
Prediction of GDM, HDP, PTB and PPH	28
Zooming in on GDM	29
Limitations of the prediction analysis	30
Discussion	32
BayesRR-RC: extensions and future directions	32
Maternal complex traits	34
Conclusion	37
References	38
Appendix A - Probabilistic inference of the genetic architecture underlying functional enrichment of complex traits	
Appendix B - Haematological changes from conception to childbirth: an indicator of major pregnancy complications	

List of Tables

1	List of GWAS in maternal health.	15
2	Steps to format genotype data	23
3	GMRM predictors from the UK Biobank.	26

List of Figures

1	Genome-wide association studies.	5
2	From genetic data to the architecture of common complex traits.	7
3	GWAS identified SNPs in a high-LD architecture	9
4	Difference in s-LDSC and SumHer functional enrichment estimates	10
5	BayesRR-RC model design	17
6	Initial simulation study	18
7	The Maternity CHUV cohort	25
8	Out-of-sample prediction of major maternal complications.	27
9	Maximising the prediction of GDM in European maternal samples.	29

Introduction

SNP-trait association studies have greatly improved our understanding of the genetic architecture of many common complex traits in the human population. Among the most studied traits are human height (HT), body mass index (BMI), type-2-diabetes (T2D) and coronary artery disease (CAD). However, most methods struggle to fully account for correlations in genetic data and explicitly model linkage disequilibrium (LD) when estimating contributions from genetic markers and functional annotations. This introduction provides current knowledge in complex trait genetics, an overview of the different methods implemented since the introduction of genome-wide association studies (GWAS), and shows the need for better models to obtain unbiased and reliable estimates. I specifically focus on methods developed to unravel the bridge between identified genomic signals and the underlying biological pathways of disease.

Genetics of complex traits

The genotype is the set of genetic information of an individual that determines part of who we are biologically and is encoded in DNA. DNA is a macro-molecule and consists of four very specific bases called adenine (A), thymine (T), guanine (G) and cytosine (C) that work in base pairs, A-T and G-C, to form a double-stranded structure. DNA is found in the nucleus of eukaryotic cells where it is tightly packed into chromosomes. As diploid organisms, humans have two copies of each chromosome, a maternal and a paternal one, and at each genetic position, we find alleles on each chromosome copy. Chromosomes consist of approximately 3 billion base pairs, of which only a small part is readable. These parts are genes, DNA sequences scattered throughout the genome and encoding for specific molecules and proteins, which in turn perform one or more functions in the body and can influence the phenotype. A phenotype refers to any trait that can be measured or observed in an individual. In human quantitative genetics, we aim to study and explain the phenotypic variance of complex traits among individuals and more precisely, we aim to better understand the genetic basis of human diseases and continuous phenotypes [J Rowe and Tenesa, 2012]. A complex trait is determined by a mixture of multiple genetic and environmental factors. Height is a good example, it is well known that children's height resembles that of their parents, even though it varies from one individual to another along a spectrum [Visscher, 2008]. This phenotypic variation can be explained by environmental factors such as nutrition, but more widely by genetics, through genes and variations found in the human genome. In contrast to complex traits, Mendelian traits are determined by single genetic changes that result in large phenotypic differences [Zwick et al., 2000]. Known examples of such single-gene diseases in the human population are sickle cell anemia or Huntington's disease.

Genetic variance can be further partitioned in three components [Falconer and Mackay, 1983]. The first is the additive genetic variance, in which we assume that genes contribute to the phenotype in an additive fashion. Additive genes provide a "*what you see is what you get*" approach as the effect of individual alleles on the phenotype is continuous and becomes measurable. The second and third are the dominance and the epistatic genetic variance. Both describe non-additive effects. Dominance is driven by interactions within a single locus (genetic position). For example, let's consider the shape of a pea, as Gregor Mendel studied in the 19th century. The round and wrinkled shapes are associated with the A and a alleles respectively. We observe that peas with the AA and Aa genotypes are round while peas with the aa genotype are wrinkled. The presence of the A allele cancels out the other and we can conclude that there is a complete dominance of the A allele. This dominance effect may vary dependent on the phenotype studied. Epistasis is of greater complexity and results from interactions within and between multiple loci. The proportion of phenotypic variance explained by genetics is called the heritability. It is also a measure of resemblance among individuals and there are two perspectives to it [Falconer and Mackay, 1983]. The *broad* sense heritability, which includes additive, dominant and epistatic effects and the *narrow* sense heritability taking only additive effects into account. Traditionally, the broad sense heritability has been measured in family or twin studies. However, non-additive effects are highly discussed in the field because they are difficult to estimate and their contribution to the total genetic variance is expected to be very small [Hivert et al., 2021]. Genetic quantitative studies mainly assume additive effects only and estimate the heritability in the narrow sense to describe the genetic variance of complex traits.

Human genome sequencing

Genome sequencing is the reading of base pairs and a great step forward in the study of genetics. The first DNA sequencing technique was invented by Frederick Sanger in 1977, and since then, it continues to progress with advances in genomics, information technology, computer science and biotechnology. In the 90's, the Human Genome Project, an international scientific research project aiming to sequence the entire human genome, began. The first draft is published in 2001 before completion of the project in 2003, suggesting that the human genome includes 30,000 to 40,000 protein-coding genes [Lander et al., 2001]. The sequencing of the first genome sparked others large-scale projects to reference and study human genome variation such as the Haplotype Map-

ping Project (HapMap) in 2003, whose phases I [Altshuler et al., 2005], II [Consortium et al., 2007] and III [GENOMICS, 2010] were published successively, and the 1000 Genomes Project (1000G) [Durbin and Altshuler, 2010, Consortium et al., 2015] initiated in 2008. Combined they now include up to 3,901 individuals from 28 global populations with different ancestries. Following the completion of 1000G in 2015, 88 million variant sites were identified, of which 84.7 million are single-nucleotide variants (SNPs), 3.6 million short indels (insertions or deletions) and 60,000 structural variants such as large deletions or copy number variants [Consortium et al., 2015]. Genetic variants are DNA changes as a result of mutations and are often identified using a reference genome. Because genetic variants can be more or less common in the population, we use minor allele frequency (MAF) to measure their frequency. Results from 1000G identified approximately 64 million autosomal rare variants with $MAF < 0.5\%$ and 8 million common ones with $MAF > 5\%$. [Consortium et al., 2015] report that we find variation in about 4 to 5 million sites with a majority of common SNPs in a single human genome and that this number varies across different population ancestries.

With a rapidly growing catalog of SNPs, researchers have naturally turned toward the possibility of using LD to study complex traits and human diseases. Using HapMap, recombination sites were shown to be associated with boundaries of LD regions emphasizing that the human genome is structured in blocks of varying size [Gabriel et al., 2002]. LD reflects non-random loci associations and measures the correlation between neighbouring alleles. LD levels are affected by recombination but also natural selection, genetic drift and mutations, which is why local and genome-wide LD patterns are indicative of past events in the genetic and evolutionary history of populations [Ardlie et al., 2002, Sawyer et al., 2005]. Since the recombination frequency is lower when genetic distance is reduced, two neighbouring loci tend to be inherited together and to have higher LD. This is extremely useful as it implies that we can sequence a subset of SNPs that may inform us about other linked SNPs [Slatkin, 2008]. This approach is called SNP genotyping. Its goal is to determine the genotype at specific positions, which allows to optimize the cost of generating human genetic data. Genotyping micro-arrays target common SNPs as well as specific sets of clinically-relevant SNP with different MAF. These micro-arrays are updated as genetic discoveries are made.

Today, several initiatives and consortia are focused on generating and aggregating data for genomic research. Among these is the UK Biobank, which has been collecting a massive amount of environmental, lifestyle and genetic data from 500,000 participants since 2006 [Sudlow et al., 2015, Bycroft et al., 2018]. There is also the Haplotype Reference Consortium (HRC), which describes over 60,000 human haplotypes (group of alleles on the same chromosome that are trans-

mitted together) [McCarthy et al., 2016]. Haplotypes are a valuable resource because they allow imputation of genotypic data, which consists in completing the data with a number of SNPs inherited together thereby increasing the number of variants studied. Finally, a third and recent initiative to mention is the Trans-Omics for Precision Medicine (TOPMed) program, a valuable resource for personalized medicine research with approximately 180,000 participants [Taliun et al., 2021]. It aggregates sequencing data, omics data such as metabolic profiles or protein expression patterns, and other types of environmental and clinical data with the goal of better understanding heart, lung, blood, and sleep disorders. Taken together, these efforts have led to a better understanding of the patterns of genetic variation in humans and have provided quality data for estimating their contribution to phenotypic variation in the human population.

GWAS and the genetic architecture of traits

SNP-trait association studies play a powerful role in understanding the genetic basis influencing an individual's phenotype, disease risk and response to the environment. The objective of GWAS is to identify genetic variants associated with a trait or disease [Uffelmann et al., 2021] (Figure 1). We want to test the association between an allele and a given phenotype, i.e. whether a particular

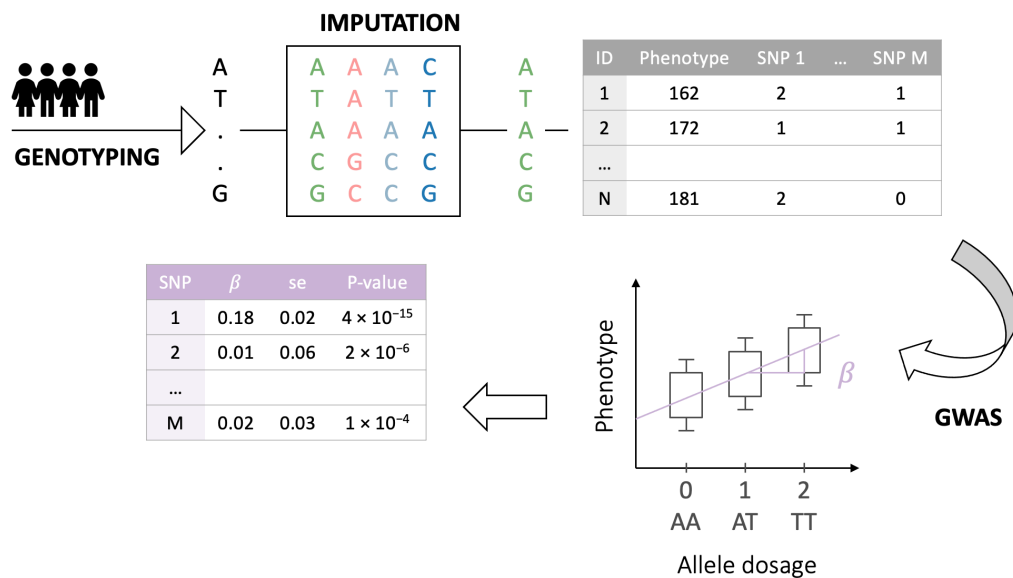


Figure 1. Genome-wide association studies. GWAS consists in identifying SNPs associated with a phenotype in a population. First, a number of individuals are genotyped. Genotyping micro-arrays mainly target common SNPs. Second, genotype data is usually imputed. Imputation uses haplotypes to complete the data with a number of SNPs that are transmitted together. These haplotypes are used as references and come from initiatives such as 1000G [Durbin and Altshuler, 2010, Consortium et al., 2015] or HRC [McCarthy et al., 2008]. Then, each SNP is tested individually for association as described in the main text. Finally, the results are combined in a summary statistics that is generally made publicly available.

allele is found more often than expected in individuals with T2D. The simplest and widely used GWAS approach is to apply a linear regression model per SNP such as for each individual i :

$$Y = Z\alpha + X\beta + \varepsilon \tag{1}$$

where Y is a vector with the individual's phenotypes, Z is a matrix of covariates for each individual and α is a vector with the corresponding covariate effect sizes. GWAS typically adjust for sex, age, genotyping batch, if any, and ancestry to account for stratification in genetic data but these can vary according to the studied phenotype. X is a matrix with the genotype value coded as the dosage effect allele 0, 1 or 2 at a single SNP for each individual. For imputed SNPs dosage is between 0 and 2. β is a vector with the corresponding SNP effect size. $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ is a vector of residual errors. GWAS linear models are applied to continuous complex traits, i.e. height. In case of binary traits, i.e. presence or absence of T2D, we can (i) apply logistic regression instead or (ii) assume that the trait is continuous with an underlying continuous liability. Heritability is then estimated linearly on the observed scale and we can subsequently transform it to the liability scale, taking into account the disease prevalence as cases could be over-represented relative to the population prevalence [Falconer, 1965, Lee et al., 2011]. This second approach allows a better comparison of the genetic basis between complex traits [Ojavee et al., 2022].

In the past 15 years, single-SNP GWAS has led to a substantial number of discoveries and progress in complex-trait genetics and translational medicine [Visscher et al., 2017]. To date, over 4,300 GWAS papers have identified more than 55,000 SNPs associated with approximately 5,000 complex traits and diseases [Loos, 2020], raising several questions about their interpretability: what can we say from a genome-wide significant SNP? How much do "top hits" actually contribute to i.e. T2D or schizophrenia? Through which biological pathways? Numerous studies have focused on answering these questions to better understand the genetic architecture of complex traits, by taking a closer look at GWAS results. The genetic architecture refers to a more detailed understanding of the genetic contributions to a given trait through different characteristics such as: the number of variants, their frequency, if there is a genomic function involved or interactions with other genetic components [Timpson et al., 2018]. In the literature, genetic architecture is often described as monogenic, oligogenic, or polygenic depending on the number of genetic contributions [Timpson et al., 2018]. Monogenic traits are characterized by single, rare genetic changes with high penetrance, whereas complex traits are more polygenic involving multiple variants, and GWAS typically identify common SNPs with low to moderate effect sizes [McCarthy et al., 2008]. Other proposed architectures include the omnigenic model, which suggests that a modest number of *core* genes are directly involved in disease etiology and that an infinite number of small *periph-*

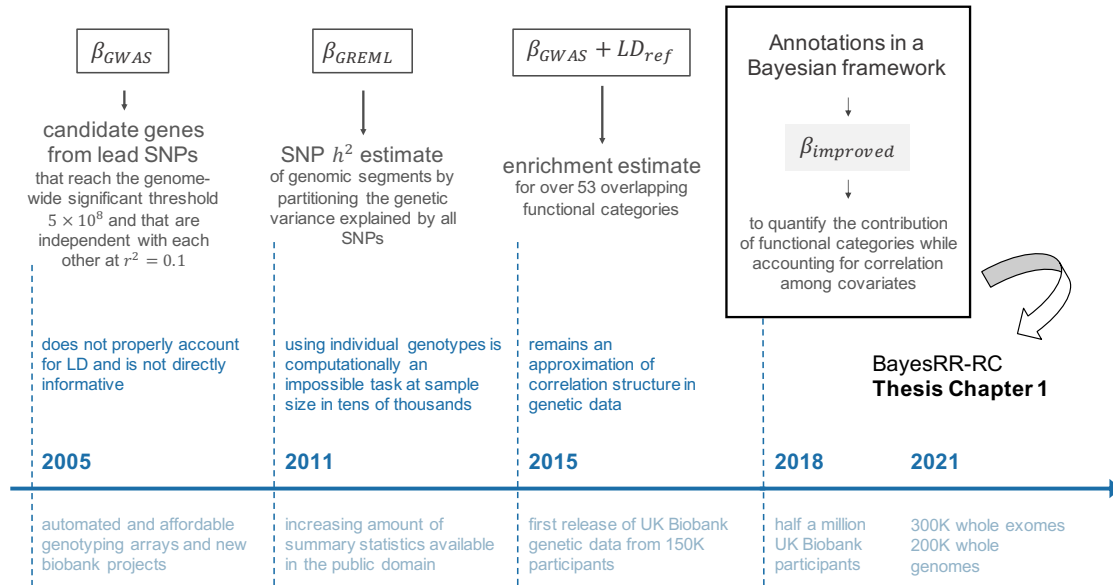


Figure 2. From genetic data to the architecture of common complex traits. A chronological overview of what has been done to better understand the underlying biology of common complex traits. The timeline highlights key progress in method development with the different challenges encountered since the experimental design of GWAS. It also describes the rapid increase of data experienced in the last 15 years.

eral contributions also influence disease through the core genes [Boyle et al., 2017]. This model is much discussed [Wray et al., 2018] and differs from the infinitesimal model, where each loci influences and is useful to describe the polygenic architecture of complex traits. SNP estimates obtained from GWAS hold great potential to better understand the underlying genetic architecture of complex traits which remains one of the biggest challenge in the field. Figure 2 summarizes the main models and challenges discussed in this introduction, for exploring genetic annotations or functions involved in complex diseases.

So, as GWAS became popular to identify susceptible markers contributing to a phenotype, GWAS-hits were further investigated using pathways analysis to discover potential causal genes involved [Lango Allen and *et al.*, 2010, Ripke and Neale, 2014, Locke and *et al.*, 2015, Hao et al., 2018, Gong et al., 2018, Vösa et al., 2018]. For example, [Locke and *et al.*, 2015] published a GWAS meta-analysis of BMI, commonly used to define obesity, in 339,224 individuals. They identified 97 BMI-associated SNPs and investigated the genetic architecture of BMI by (i) manually examining SNPs in high LD (r -squared > 0.7) and all genes within ± 500 bp from all BMI-associated-SNPs; (ii) applying DEPICT [Pers et al., 2015] and MAGENTA [Segrè et al., 2010], two widely used integrative tools for pathway analyses, to identify specific pathways and potential causal gene sets. DEPICT was also used to identify specific tissues and cell-types based on gene expression near the 97 BMI-associated SNPs and tested for significant enrichment. Using this complementary

approach [Locke and *et al.*, 2015] showed strong enrichment in the central nervous system and isolated the hypothalamus, pituitary gland, hippocampus and limbic system known to be linked with appetite regulation, cognitive functions and emotions. Over the years, such complementary approaches highlighted potential causal genes and provided strong genetic evidence for particular biological pathways [Ripke and Neale, 2014, Locke and *et al.*, 2015].

There are a variety of gene prioritization methods similar to DEPICT [Pers et al., 2015] and MAGENTA [Segrè et al., 2010] that also use different omics data. One of them is to perform a transcriptomic association study (TWAS) [Gusev et al., 2016, Xu et al., 2017] to discover possible functional annotations from the GWAS statistics. The idea behind this approach is to integrate gene expression data mostly from GTEx [GTEx Consortium, 2015] with GWAS summary statistics and an LD reference panel from 1000G to find any evidence of expression-trait associations. An example of TWAS application is a recent study on birth weight to identify any potential eQTLs (quantitative trait loci affecting gene expression) underlying associations of birth weight with maternal and fetal effects [Warrington and Beaumont, 2019]. However, TWAS mostly applies to expression data only, although it can be extended to incorporate other plausible regulatory variants such as splicing or histone marks [Gusev et al., 2016]. The results are dependent upon the quality of the gene expression data and TWAS does not exclude the possibility that identified eQTLs might be caused by the phenotype instead of the SNP.

Explicitly modelling LD

There are a number of issues in the experimental design of GWAS that prohibit a full characterization of genetic effects and how they might influence disease. Among these issues is the fact that millions of markers are tested one at a time and so, the significance threshold needs to be adjusted to account for multiple testing [Tam et al., 2019]. Typically, a threshold of $P - value \leq 5 \cdot 10e - 8$ is used assuming 1 million independent tests and using Bonferroni correction (false discovery rate at 5%). Due to the stringent threshold, smaller effects may be difficult to identify and larger sample sizes are required. Moreover, a number of factors need to be considered when running a GWAS such as population structure or assortative mating, which can affect the distribution of genetic variance and bias the results if ignored. Another major concern is controlling for LD, which utterly fails to comply with the independence assumptions made in GWAS, resulting in an overestimation of SNP estimates [Vilhjálmsón and *et al.*, 2015, Maier et al., 2018]. Because of LD, GWAS-lead SNPs found to be associated with a trait are not directly informative with respect to the target gene or mechanism driving the studied phenotype [Visscher et al., 2017]. Any functional

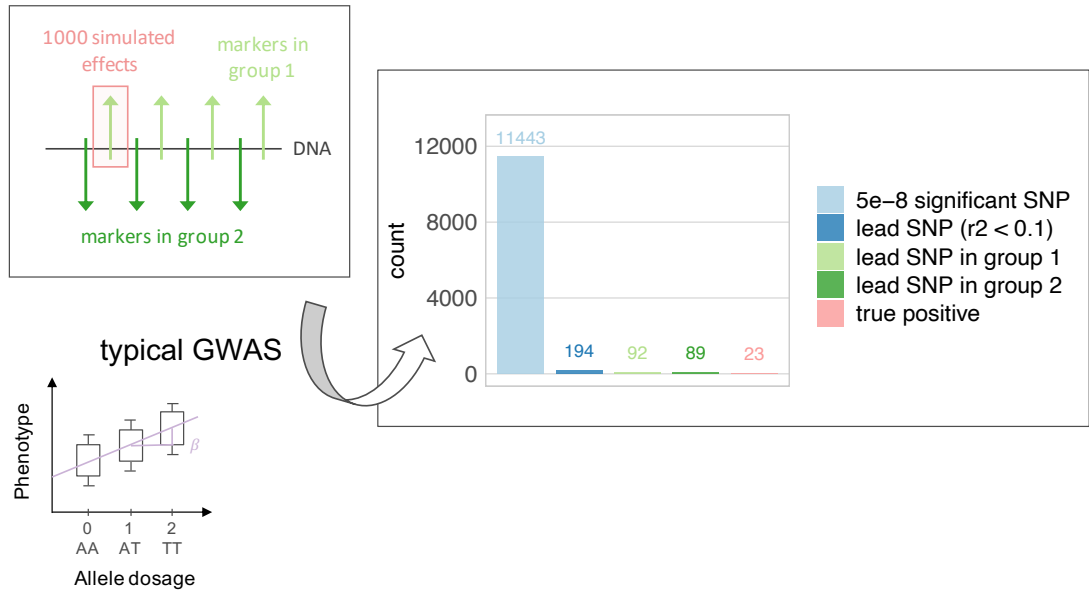


Figure 3. GWAS identified SNPs in a high-LD architecture. This example aims to illustrate that lead SNPs identified from a typical GWAS are not directly informative with respect to the tagging genomic region. Using 87,000 markers and 20,000 individuals from the UK biobank, we assigned every second marker to group 1 while the others were assigned to group 2. Groups are thereby in high LD because for every marker in group 1, adjacent markers are in group 2. We simulated a genetic variance of 0.6 explained by 1000 randomly assigned SNPs in group 1 only. Among the 194 SNPs to be considered independent (r^2 measure of LD between SNPs < 0.1) half of them belong to group 2, which does not contribute to the genetic variance simulated. Furthermore, we only identify 23 out of the 1000 true causal SNPs.

annotations in the human genome might be correlated and complementary analyses cannot fully account for this correlation structure, further resulting in biased enrichment estimates, i.e. [Locke and *et al.*, 2015] and [Ongen *et al.*, 2017] highlight the central nervous system as a significantly enriched tissue but how can we be certain that we are not just tagging regions in high LD? As an example, I ran a single-SNP GWAS simulation of 20,000 individuals and 87,000 markers from the UK Biobank (Figure 3), where 1000 SNP effects are assigned to group 1 that is in high LD with markers in group 2. Groups can be perceived as two functional annotations with all the genetic variance explained by annotation 1. We identified 194 independent lead SNPs, of which 89 in the zero-variance component group 2. This provides evidence that GWAS ignores LD and further pathway analysis from these 194 trait-associated SNPs could show biased enrichment estimates.

With GWAS being so successful, there has been an overwhelming and continuous increase of GWAS summary statistics available in the public domain [Visscher *et al.*, 2017]. Consequently, new methods such as LD score regression (LDSC) were developed to: (i) properly and explicitly model LD and (ii) estimate the enrichment of specific genomic regions from summary statistics. LDSC was first introduced by [Bulik-Sullivan *et al.*, 2015] in 2015 to quantify SNP heritability from summary statistics and then extended to stratified LDSC (s-LDSC) by [Finucane *et al.*, 2015]

to partition the SNP heritability and estimate enrichment of heritability in functional annotations. By combining LD reference scores and prior-biological information, s-LDSC has facilitated the discovery of functional elements in the genome and provided us with key information about the underlying genetic architecture of many complex traits. In 2018, a recent study on major depressive disorder (MDD) [Wray et al., 2018] identified 44 SNPs associated with MDD clinical features. They also provided potential clues to common biological mechanisms that may influence other common psychiatric diseases using s-LDSC. The study combined GWAS summary statistics from seven different cohorts and investigated the contribution of several functional annotations constructed from two public projects on functional elements in the human genome, ENCODE [Consortium, 2012] and Roadmap [Kundaje and Meuleman, 2015].

While s-LDSC is able to detect enrichment, it remains an approximation and shows three major limitations. First, it requires a well referenced LD panel that matches the population studied; i.e. if you have a European ancestry population, you need an adequate European reference panel to correct for LD in your data [Finucane et al., 2015, Bulik-Sullivan et al., 2015]. Second, it assumes that a rigorous quality control of the genetic data was applied and that GWAS summary statistics are properly combined. Any presence of heterogeneity in the data can easily affect SNP heritability and enrichment estimates [Marees et al., 2018]. Third, recent work has shown striking differences between LDSC and SumHer [Speed and Balding, 2019], implemented in the LDAK software to estimate (i) SNP heritability, (ii) enrichment of heritability, (iii) confounding bias and (iv) genetic

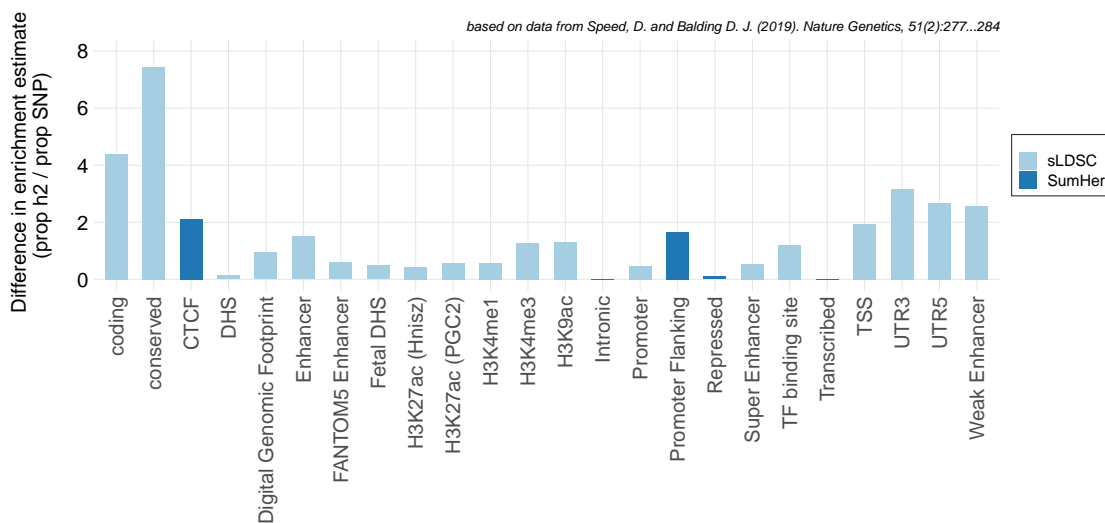


Figure 4. Difference in s-LDSC and SumHer functional enrichment estimates. This figure shows the difference in the estimates of functional enrichment from sLDSC (53-part model) and SumHer-GC (25-part model). In Speed et al. (2019), average enrichment estimates were calculated for the 24 functional categories from Finucane et al. (2015) and showed striking differences.

correlations applied to GWAS summary statistics. When considering the 24 functional categories described in [Finucane et al., 2015], s-LDSC reports an average 13-fold enrichment of heritability in conserved regions across 17 traits whereas SumHer shows no significant enrichment above two-fold for any functional category (Figure 4). However, if we look at their modelling assumptions, s-LDSC assumes that each SNP contributes equally across the genome, whereas in SumHer, SNPs in high LD regions are expected to contribute less than SNPs in low LD. These discrepancies indicate that enrichment estimates are sensitive as to how we model LD [Speed and Balding, 2019].

Fitting SNPs jointly

Methods based on GWAS summary statistics as LDSC approximate a mixed effect model and should perform similarly to a restricted maximum-likelihood (REML) model, which uses individual level data and fits SNPs jointly. So, instead of considering a single SNP at a time as in equation 1, each SNP is estimated conditionally on the joint effects of all the other SNPs in the model, to accommodate for linkage between them such as:

$$Y = Z\alpha + X\beta + \mathbf{g} + \varepsilon \quad (2)$$

where $\mathbf{g} \sim N(0, \sigma_G^2 W)$ is a random effect to account for other SNPs. σ_G^2 measures the additive genetic variance and W is the genomic relationship matrix (GRM), describing the genetic relationship between individuals from SNP data. The software GCTA is commonly used and applies a GREML (REML fitting a GRM) model [Yang et al., 2011a]. Initially developed to estimate SNP-heritability, the software was subsequently extended to partition the genetic variance component by chromosomes and genomic segments [Yang et al., 2011b, Gusev and *et al.*, 2014]. However, GREML has its own limitations starting with the model specification. Like s-LDSC, it assumes that we are in linkage equilibrium, which poorly reflects reality and might influence SNP estimates particularly in high LD regions [Speed and Balding, 2019]. Second, because the number of variants is much larger than the number of individuals included, running GREML becomes a computational nightmare specially when analyzing multiple functional categories with tens of thousands of individuals [Finucane et al., 2015]. To overcome these issues, a paper introduced RHE-reg-mc a fast and scalable method-of-moments to jointly estimate multiple variance components [Pazokitoroudi et al., 2020]. Simulation work showed a 400 fold reduction in computation time compared to REML-based methods. SNP heritability was computed in only 40 min for 300,000 individuals and 500,000 SNPs partitioned in 250 components. As a trade-off, RHE-reg-mc unfortunately shows larger standard errors and is statistically less efficient.

On top of the computational challenge, the GREML model assumes an infinitesimal genetic architecture where the genetic variance is explained by an infinite number of small effects from all markers. Given the number of markers, several studies have proposed using prior distributions to account for markers that may have no effect [Meuwissen et al., 2001, Erbe et al., 2012]. BayesR, a Bayesian mixture model for genomic data, was first developed to improve genomic phenotype prediction in dairy cattle breeds [Erbe et al., 2012]. In 2015, [Moser et al., 2015a] further introduced BayesR to estimate the number of trait-associated SNPs and their heritability from human individual level data. BayesR is coupled with a Markov chain Monte Carlo (MCMC) scheme and has shown robust results for inferring the genetic variation of traits [Erbe et al., 2012, Moser et al., 2015a]. In this framework, SNPs are jointly estimated and treated as random variables. SNP effects (β) are then drawn from four mixture distributions: a spike at zero and three zero-mean distributions, each with a fixed variance. This last feature allows the user to distinguish SNPs according to their effect size. Some will not enter the model, some will have small effects, some will have moderate effects and some will have larger effects. The advantage of using a Bayesian approach is that it is flexible and can be applied to sequencing data while including prior-information on the distribution of effects. Modelling a spike at zero allows for selection and only non-zero SNP effects can be used if one wishes to predict the phenotype. BayesR simultaneously informs about us on the number of variants involved and their contribution to the genetic variance at different effect magnitudes.

BayesR was subsequently extended to BayesRC to estimate the variance explained of multiple mixtures of SNP categories [MacLeod et al., 2016]. The model uses prior biological information to divide SNPs into classes that are likely to be relevant and show enrichment for a trait. BayesRC has mostly been applied to investigate complex traits in plant and animal breeding and an important caveat of this method is handling bigger sample sizes in human genetics. In 2020, [Banos et al., 2020] introduced BayesRR to estimate associations between measured epigenetic marks and disease risk in humans. In contrast to BayesRC coded in Fortran, BayesRR was implemented in C++-11.0 and using the matrix algebra library Eigen [Guennebaud et al., 2015]. The implementation handled over 5,000 individuals and is computationally extremely promising for multiple human genetic applications [Banos et al., 2020]. A combination of the two Bayesian methods would therefore provide a more complete characterization of the underlying biological pathways while reducing computational cost. Moreover, if we integrate functional annotations into this Bayesian framework, we could simultaneously estimate the SNP heritability, the genetic contribution of these annotations and each SNP effect size while taking into account the correlation among all parameters. With this strategy and as annotations accumulate, SNP and enrichment estimates

would improve and the genetic architecture of common complex diseases could be fully explored in an unbiased and objective manner [Moser et al., 2015a, MacLeod et al., 2016].

The first chapter of my thesis focuses on the development of a novel Bayesian model which integrates functional annotations to better understanding the genetic architecture of complex traits. My interest then turn towards the application of association studies and more specifically to applications in maternal health.

From association studies to health care

As discussed, numerous pipeline investigate the genetic architecture of complex diseases from genetic loci identified in GWAS. These pipelines aim to improve our understanding of disease etiology and determine potential drug targets, that would then be validated in molecular experiments and clinical trials [Loos, 2020]. For example, in [Dwivedi et al., 2019], they demonstrate through a genotype study and metabolic profiling *in vivo* that a rare allele in the *SLC30A8* gene has a protective effect against T2D. They described that the allele is enriched in Western Finland and that this protection comes from a better response to glucose leading to an insulin secretion that is more efficient. Their findings on *SLC30A8* provide candidate drug targets for maintaining insulin secretion in patients with T2D. This is a follow-up study to a previous paper published in 2014 where [Flannick et al., 2014] identified 12 protein-truncating variants in *SLC30A8*. The gene also has a common variant that is associated with T2D, glucose and proinsulin levels. These studies are motivated by initial GWAS results on T2D.

Results from SNP-trait association studies have also been widely used in epidemiology. A popular application is to explore causality between an exposure, i.e. systolic blood pressure (SBP) and an outcome, i.e. pregnancy-induced hypertension (PIH) through Mendelian randomisation (MR) analysis where genetic variants robustly associated with the exposure, are used as a proxy to assess causality, i.e. the causal effect of SBP on PIH. MR is an excellent alternative to randomized controlled trials, where one would need to expose individuals to different exposures such as smoking that are often neither practical nor ethical [Lawlor et al., 2008]. MR draws its benefits from the principle that alleles are transmitted randomly from parents to offspring, creating a framework in which genetic instruments would be independent of any confounding factors. Furthermore, because genotypes are assumed to be fixed from birth, associations with an outcome would not be expected to be due to reverse causation [Lawlor et al., 2008]. The example mentioned above has been studied by [Ardissino et al., 2022] who performed MR to explore the causal effects of

maternal cardiovascular risk factors on preeclampsia and on eclampsia, two hypertension diseases in pregnancy affecting both maternal and fetal health. They find evidence that SBP as well as BMI and T2D contribute to cause both complications during pregnancy.

Finally, SNP estimates from association studies can be used to predict an individual's disease risk, a key feature in personalised medicine. Typically, disease risk prediction is quantified using a polygenic risk score (PRS) where we aggregate the effect of genetic variants estimated using SNP-trait associations [Loos, 2020]. Such scores have shown to be predictive for various complex diseases. In a clinical setting, they are combined with other non-genetic risk factors like age, cell blood count, disease family history or smoking status, to help physicians in their decisions [Lambert et al., 2019]. Predictive analysis can also shed light on the underlying and overlapping genetic architecture of complex traits. For example, [Steinhorsdottir et al., 2020] published a large meta-analysis on preeclampsia which included a PRS analysis. They predicted preeclampsia and gestational hypertension in Icelandic individuals from the deCODE cohort [Gudbjartsson et al., 2015] using a PRS constructed from GWAS results on hypertension in the UK Biobank data. Their results indicate a correlation between hypertension and preeclampsia suggesting a common genetic background. However, the effect of the hypertension PRS on gestational hypertension was almost twice as large, which probably implies the existence of other risk factors specific to preeclampsia.

Pregnancy and maternal health

Among the most studied pregnancy complications in genetics are preeclampsia, gestational diabetes mellitus (GDM), preterm birth (PTB), hyperemesis gravidarum (HG), placental abruption (PA) and miscarriage. An overview of standard GWAS conducted in maternal health is given in Table 1. Current results show that having one complication most likely increases the risk of another and the risk of adverse outcomes at delivery. They also suggest that complex maternal traits have a genetic basis, which appears to be highly polygenic, and for this reason larger sample sizes are needed to explore the role of genetics in maternal health [Barbitoff et al., 2020]. In a study just published in March 2022, the Genetics of Diabetes in Pregnancy Consortium aggregated several studies on GDM and performed the largest and most diverse GWAS meta-analysis available [Pervjakova et al., 2022]. The study includes 5,458 cases and 347,856 controls. They identified 5 genes associated with GDM, 4 of which are also associated with other traits. For instance, the gene *MTNR1B* has already been associated with T2D and fasting blood glucose characteristics in non-diabetic individuals. Maternal SNPs at the *MTNR1B* gene were also associated with fasting blood glucose in pregnant women and birth weight of the offspring. Following these results,

they performed (1) an enrichment analysis in which they identified significant enrichment mapping to protein-coding exons, chromatin immuno-precipitation sequence (ChIP-seq) binding sites for 3 transcription factors, and chromatin states marking enhancers and transcribed regions in adipose tissue and skeletal muscle; and (2) an MR analysis where they identified a causal link between BMI and GDM risk.

Table 1. List of GWAS in maternal health.

Trait	Study	Cases	Controls	Sample	Associations
Preeclampsia	[Johnson et al., 2012]	538	540	maternal	<i>INHBB</i>
	[McGinnis et al., 2017]	4,380	310,238	fetal	<i>FLT1</i>
	[Steinthorsdottir et al., 2020]	6,775	375,372	fetal	<i>FLT1</i>
		9,515	157,719	maternal	<i>ZNF831</i> <i>FTO</i> <i>MECOM</i> <i>FGF5</i> <i>SH2B3</i>
GDM	[Kwak et al., 2012]	931	783	maternal	<i>CDKAL1</i> <i>MTNR1B</i>
	[Wu et al., 2021]	103	115	maternal	<i>SYNPR</i> <i>CDH18</i> <i>CTIF</i> <i>PTGIS</i>
	[Pervjakova et al., 2022]	5,458	347,856	maternal	<i>MTNR1B</i> <i>TCF7L2</i> <i>CDKAL1</i> <i>CDKN2A-CDKN2B</i> <i>HKDC1</i>
PTB	[Zhang et al., 2017]	3,331	39'237	maternal	<i>EBF1</i> <i>EEFSEC</i> <i>AGTR2</i>
	[Rappoport et al., 2018]	1,349	12,595	fetal	2 SNPs
	[Liu et al., 2019]	4,775	79'914	fetal	2 SNPs
	[Tiensuu et al., 2019]	247	419	fetal	<i>SLIT2</i>
	[Taliun et al., 2021]	18'797	260'246	maternal	6 SNPs
HG	[Fejzo et al., 2018]	1'306	15'756	maternal	<i>GDF15</i> <i>IGFBP7</i>
PA	[Workalemahu et al., 2018]	959	1,553	maternal	12 candidate genes
Miscarriage	[Laisk et al., 2020]	69,054	359,469	maternal	5 SNPs

Our knowledge on the etiology of obstetric complications is currently limited, and there are few ways to effectively prevent them [Barbitoff et al., 2020]. SNP-trait association studies can help to better understand the mechanisms underlying these complications and identify promising therapeutic targets. Such studies can also help predict complications even before pregnancy occurs, using a PRS constructed from the mother's genetics. Genetic scores could then be combined with other risk factors to improve prediction, i.e. to predict preeclampsia one could combine PRS

with maternal clinical factors such as BMI, parity, or history of preeclampsia, as well as biomarkers [Antwi et al., 2020]. PRSs would allow a better stratification of women in low and high-risk pregnancies [Ma and Zhou, 2021], and would assist clinicians in improving maternal care to minimise adverse outcomes, i.e. your pregnancy risk status can help decide where to be monitored and where to give birth. With the availability of genotyping, quantitative genetic research on maternal health has increased in recent years. This topic is extremely important but also complex because of ethical considerations related to maternal and fetal data, the context of the pregnancy itself, and the modest sample sizes available [Barbitoff et al., 2020]. In Table 1, some GWAS use a looser p-value threshold of $p < 10^{-5}$ instead of $p < 10^{-8}$ to further explore associations. As we move forward, larger studies will be needed to identify new genetic associations and take advantage GWAS in obstetric medicine.

To date, there is very little emphasis on what makes a healthy pregnancy from the mother's perspective, the physiological changes she undergoes, and the genetic basis of maternal health. For example, are physiological measurements in early pregnancy indicative of measurements later on? Are women more likely to experience complications if measurements fluctuate? Can risk groups be predicted early in pregnancy? In the second chapter of my thesis, I explore data from the Lausanne maternity hospital and apply different statistical models with the aim of finding physiological and genetic risk factors, to predict major pregnancy-related complications and support clinicians in maternal health care.

Chapter 1

Bayesian penalized regression for complex trait analysis

This first chapter outlines my contribution and involvement to the development of BayesRR-RC, which builds on previous work from [Banos et al., 2020], and is presented in the manuscript: **Probabilistic inference of the genetic architecture underlying functional enrichment of complex traits** (Patxot *et al.* (2021) - see Appendix A). The project is motivated by the desire to better understand the genetic architecture of traits, to identify genomic regions, functions or genes that contribute most to a trait whilst accounting for all structure in the data. To achieve that, we integrated functional genome annotation information into the BayesR framework, which uses Bayesian statistics coupled with an MCMC scheme [Moser et al., 2015b], to jointly estimate the contribution of SNP-marker groups to complex traits as described in Figure 5. As a starting point, I adapted the BayesR algorithm of [Banos et al., 2020] to infer the variance explained of multiple genetic components and implemented these changes in C++11.0. The software was developed using hybrid-parallel algorithms by Etienne Orliac, an expert in HPC applications and Daniel Trejos Banos, a postdoc in the group. I then contributed to testing the influence of increasing parallelism in our algorithm, wrote a wiki description and implemented a user-friendly example on GitHub.

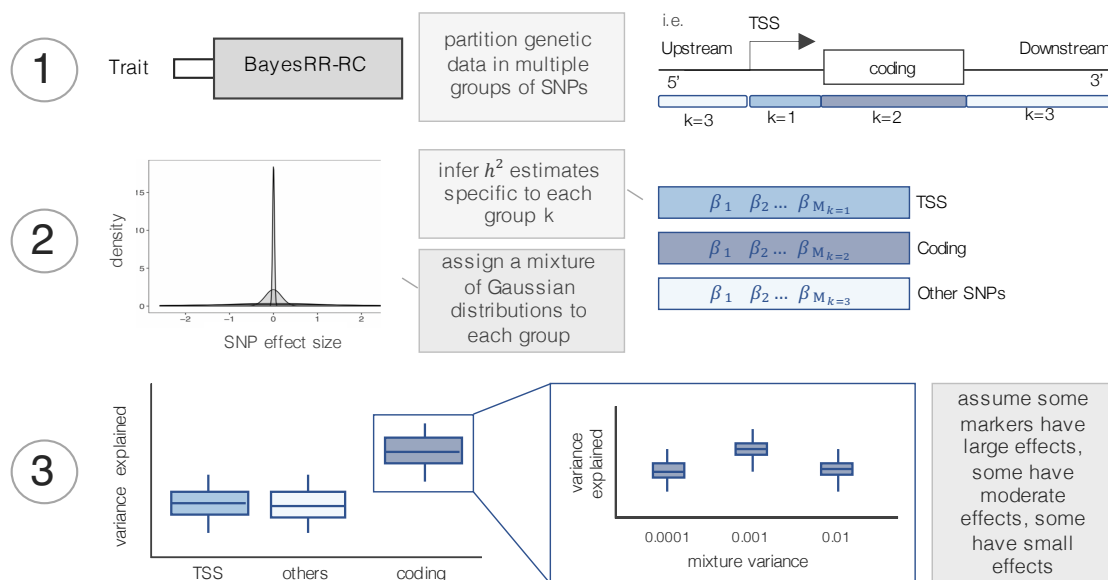


Figure 5. BayesRR-RC model design. Our extended BayesR method uses prior biological knowledge to (1) partition SNPs into multiple annotations, i.e. coding, TSS and non-coding regions and (2) quantify their unique contribution enabling us to improve our understanding of common complex diseases. We further assign prior mixture distributions to each SNP-marker group including a discrete spike at zero to highlight the underlying genetic architecture within each group in (3). In the manuscript, we demonstrate through large-scale simulation that adding LD bins to each annotation group takes into account all levels of LD in the data, which further improves the performance of the model.

BayesRR-RC model validation

To evaluate the performance of the model, I initially completed a set of simulations as described in Figure 6. With Prof. Matthew Robinson, we further validated the model including MAF and LD sub-groups, in multiple simulation settings as described in the manuscript. I specifically contributed to validating the inference of simulated genetic architectures using empirical annotations, exploring effects of relatedness on the estimates obtained from the model, examining the ability of BayesRR-RC to recover effect sizes in specific groups compared to BayesR, and validating the use of the posterior probability window variance (PPWV) approach for downstream analysis. In our analysis, the latter provides the posterior inclusion probability of a genomic region to contribute at least 0.001% to the SNP heritability.

In summary, BayesRR-RC allows us to accurately compare and contrast the inferred underlying genetic distribution of complex phenotypes. When enrichment is specified using prior knowledge, the genetic architecture is accurately inferred, except in a very low polygenic setting where the

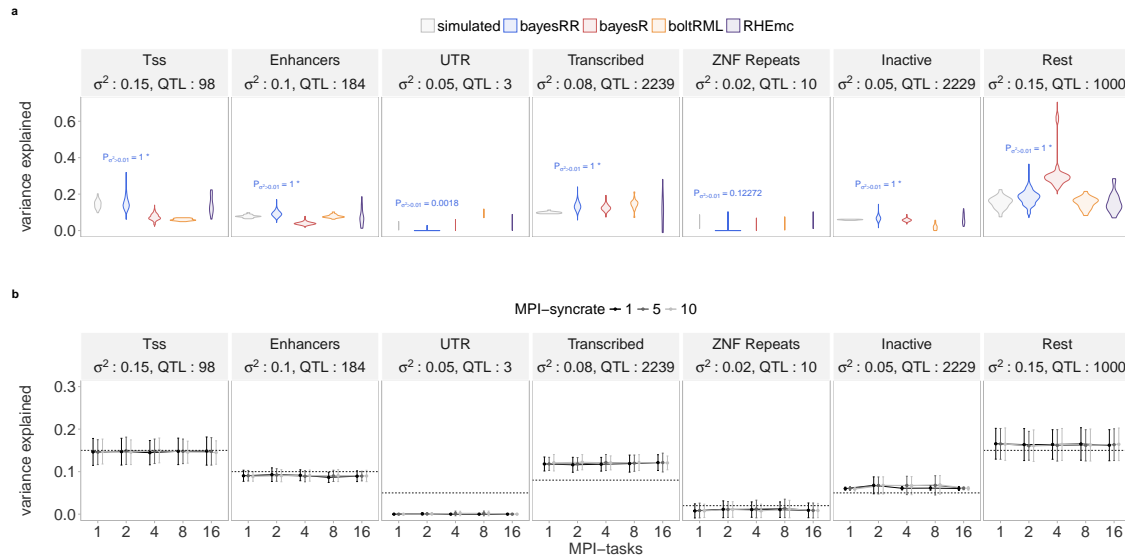


Figure 6. Initial simulation study. Simulation with $N = 20,000$ UK Biobank unrelated European individuals and $M = 328,385$ markers from chromosome 2. We used epigenome E062, primary mononuclear cells from peripheral blood from the Epigenome Roadmap Project, to split SNPs into 7 groups based on 15 chromatin states: (1) transcription start site (TSS), (2) active enhancers, (3) untranslated transcribed regions (UTR), (4) actively transcribed regions, (5) actively transcribed zinc finger genes (ZNF Repeats), (6) inactive states and (7) others SNPs (Rest). 5% of the SNPs in each group are QTLs, except for the unmapped SNPs where we randomly sampled 1000 QTLs, allowing different polygenicity level in each group. **(a)** Violin plot of genetic variance estimates from 20 simulation replicates of BayesRR (chain length: 10500, burn-in: 500, thin: 4). We compare our software to (1) multi-variance component methods, Bolt-REML [Loh et al., 2015] and RHE-mc [Pazokitoroudi et al., 2019], that use individual-level data but provide single heritability estimates per genetic component, and (2) the total genetic variance estimated with BayesR and calculated post-analysis for each group. $P_{\sigma^2 > 0.01}^2 > 0.01$ consists in quantifying the proportion of variance estimates above a threshold of 0.01; if less than 95% of estimates are > 0.01 , we will consider this group to have a null contribution at current sample size. **(b)** Mean and standard deviation of genetic variance estimates from 10 simulation replicates of BayesRR (chain length: 2000, burn-in: 1000, thin: 1), specifying the same annotations and using 1, 2, 4, 8 and 16 tasks at 1, 5 and 10 message passing rate for our sampling scheme.

few contributing SNPs might randomly not be picked up by the MCMC chain, resulting in a loss of power (Figure 6). Inference is improved over other approaches with partitioning of SNPs by genomic regions and especially by LD-informed bins, in a BayesR framework with a Dirac spike and slab prior set for each group. Validating the PPWV approach offers an alternative to standard GWAS for locating genetic associations at the regional level. We explicitly demonstrate that summary statistics approaches are less efficient for variance component estimation and for SNP-marker group enrichment, than individual-level methods. And finally, we reduce the computational complexity of applying a Bayesian model to large-scale genomic data such as the UK Biobank.

Application to complex traits in the UK Biobank

After performing the simulation work, we investigated the genetic architecture of CAD, T2D, BMI, and HT measured in 382,466 unrelated European individuals from the UK Biobank. We used 8,433,421 imputed SNP markers with $MAF > 0.0002$. Here, I applied BayesRR-RC to the four traits along with other methods commonly used in quantitative genetics. Among them, the Bolt-REML software [Loh et al., 2015] did not reach the end of the estimation after 7 days, illustrating the importance of scalable solutions, especially with the forthcoming sequencing data. In the manuscript, we report results obtained with RHE-mc [Pazokitoroudi et al., 2020] and two widely used methods based on GWAS summary statistics, s-LDSC [Finucane et al., 2015] and SumHer [Speed and Balding, 2019]. In addition, I performed all downstream analyses to identify significant associations and compare the number of results between methods. Finally, we used our estimates in the UK Biobank to predict the same four complex traits in the Estonian Genome Centre data. I specifically contributed to the comparison with the MegaPRS software [Zhang et al., 2021] where I used a boltLMM prediction. Following this analysis, Prof. Matthew Robinson conducted a follow-up study to further improve genomic prediction accuracy and put into perspective the application of BayesRR-RC implemented in the GMRM software [Orliac et al., 2021].

The results showed a similar distribution of variance among the groups between the four traits. Due to the Bayesian nature of BayesRR-RC, it is not obvious how our results can be directly compared to frequentist methods. If we consider the associations based on a $PIP > 95\%$, across the four traits, we identified 391 SNP associations, of which 53% have already been listed by the fastGWAS UK Biobank summary statistic data with a $p\text{-value} < 5 \times 10^{-8}$ [Jiang et al., 2019]. When comparing the estimated SNP heritability across the four traits using the PPWV approach, we identified that 32 to 44% of the genetic variance is attributed to intronic regions and 12 to 25% to coding regions with over 3,100 independent regions having $\geq 95\%$ probability of contributing

$\geq 0.001\%$ to the genetic variance of these four traits. The distal 10-500kb regions appear to be more polygenic with SNP heritability estimates ranging from 22 to 28% and more than 5,400 independent regulatory regions identified. Surprisingly, less than 10% of the genetic variance is captured by proximal regions in the 10kb upstream of the genes. Finally, prediction results showed that we obtained higher prediction accuracy compared to MegaPRS [Zhang et al., 2021].

Chapter 2

Maternal cohort study

We obtained local cantonal ethics approval from the CER-VD in May 2019 for the reuse of data collected at the CHUV Maternity Hospital between 2009 and 2014 to study maternal and fetal toxoplasmosis infections. The CHUV data extraction unit retrieved maternal medical records in a de-identified format for 4,347 pregnancies and made available the corresponding maternal and umbilical cord blood samples collected at delivery. Of the latter, 1,628 samples were processed for SNP genotyping with the help of our laboratory technician, Rosanna Pescini Gobert, Microsynth AG and iGE3 Genomics Platform. I was actively involved in setting up the maternity cohort to better utilize existing data, I contributed to the application at the CER-VD, to the coordination of the SNP genotyping and all the analysis tasks. In this second chapter, I present the manuscript: **Haematological changes from conception to childbirth: an indicator of major pregnancy complications** (Patxot *et al.* (2022) - see Appendix B), in which I sought to delineate the haematological changes during healthy pregnancies and pregnancies affected by hypertensive disorders of pregnancy (HDP), GDM and post-partum hemorrhage (PPH) in our cohort. I then present my work on the genetic data for which SNP genotyping was completed in January 2022. This work is not included in the appendix.

Haematological changes in low and high risk pregnancies

In the manuscript - Appendix B, we used data on 14 cell blood count (CBC) parameters, which are routinely performed and easily accessible in any prenatal clinic, to establish haematological changes during pregnancy. We extensively describe these changes in healthy pregnancies and assess differences in the variation of CBC during pregnancies with either HDP, GDM or PPH using a cubic polynomial regression and a mixed effects model. We additionally apply a survival model to define the association of CBC and pregnancies complicated by HDP or GDM, with birth timing. The manuscript raises the following questions: are high measurements early in pregnancy indicative of higher later measures? Are women more likely to experience complications if measures fluctuate? Can risk groups be predicted in early pregnancy? The short answer is most probably, by setting up large comprehensive and integrative cohorts for maternal and fetal medicine. Results confirmed and refined previous findings in healthy pregnancies where pregnant women present a net decrease in red blood cell parameters followed by a stabilisation in the third trimester. White blood cell counts mainly rises because of the physiologic stress imposed on the body. And finally, platelet count decreases continuously. We demonstrate that routine CBC can

be sufficiently sensitive to identify pathophysiological mechanisms occurring in early pregnancy that will later lead to the development of obstetric or postpartum complications, identifying the 10th to 20th weeks of gestation as the most informative period to do so. However, a larger, more recent and complete dataset from the beginning of pregnancy would allow us to further study these changes and attempt to predict pregnancy-related complications using CBC. I am the main author of this article, I designed the study, performed the statistical analyses, and wrote the manuscript with helpful feedback from the co-authors.

Genetic risk prediction of maternal complex diseases

Because of its rich phenotypic data, the CHUV maternity cohort aims to better understand maternal health. Specifically, it has been set up, as effectively as possible, to study HDP, GDM and PPH. In this section, I describe how we generated the genetic data and then present some preliminary results where I explore the major obstetric complications using predictors from the UK Biobank.

Methods

DNA extraction and SNP genotyping

Blood samples from the mother and from the umbilical cord collected at delivery were stored at -80°C at the CHUV Maternity ward. As samples were collected for a serological study 10 years ago, they consisted largely of circulating free DNA in blood cells, which was extracted by Rosanna Pescini Gobert and Microsynth AG. The 1,628 samples processed only include women whose pregnancy resulted in a single live birth and for whom we have a delivery report in the phenotypic data. Stillbirths, multiple pregnancies, and women with ICD-10 codes for conditions other than those classified as Pregnancy, Childbirth, and Puerperium (O00-O9A) were not included in the cohort. As DNA was extracted, our laboratory technician shipped 48- and 96-well plates to iGE3 Genomics Platform for genotyping. SNP genotyping was conducted using the Illumina Infinium Global Screening Array (GSA) version 2 (v2, from 2019 to 2020) and version 3 (v3, from 2021 to 2022). GSA v2 is based on the GRCh37/hg19 human genome assembly [Church et al., 2011] while v3 is based on the GRCh38/hg38 assembly [Schneider et al., 2017]. Both arrays genotype $\sim 654,027$ genome-wide SNPs and $\sim 100,000$ identified clinical variants. The latter are updated from v2 to v3 and come from the ClinVar [Landrum et al., 2020], CPIC [Relling and Klein, 2011] and PharmGKB [Whirl-Carrillo et al., 2021] projects.

Quality control of the genotype data

To carry out any SNP-trait association study, a quality assessment and control of the SNP genotype data is necessary to identify markers or samples of low quality [Anderson et al., 2010]. For each

Table 2. Steps to format genotype data.

Step	Software
convert each .ped/.map files to .bed/.bim/.fam format merge files from the same array version (v2 or v3)	plink 1.9 [Purcell et al., 2007]
convert v3 plink file from hg38 to hg19 assembly	liftOverPlink [Ritchie, 2014]
remove any non-ACGT alleles	plink 1.9 [Purcell et al., 2007]
strand and ambiguous alleles check	snpflip [Stovner and Cole, 2019]
exclude identified ambiguous alleles flip identified alleles on the reverse strand set reference allele based on the genome assembly hg19 update snp name to chromosome:position remove any multi-allelic snps merge v2 and v3 formatted .bed/.bim/.fam files	plink 1.9 [Purcell et al., 2007]

plate sent together, the iGE3 genomic platform provided us with IDAT intensity data files and SNP genotypes coded in .ped and .map plink format [Purcell et al., 2007]. Table 2 describes the first part of the quality control which consists of combining the data into a single file containing all bi-allelic SNPs genotyped. These are first filtered then flipped on the direct strand and referenced using the hg19 human genome assembly. I then filtered the data according to the following 5 criteria: (1) excluded individuals with a missing genotype rate $\geq 30\%$, which is more lenient than the typical call-rate threshold of 3% [Anderson et al., 2010] because of the lower quality of the extracted DNA, (2) excluded individuals with a heterozygosity rate ± 2 standard deviation from the mean, (3) excluded SNPs with a missing genotype rate $\geq 10\%$ or more, (4) excluded SNPs not passing the Hardy-Weinberg test at $p\text{-value} \leq 0.00001$, (5) excluded SNPs with $\text{MAF} < 0.01$. A total of 137 individuals and 284'929 markers were removed. I performed a principal component analysis (PCA) to ensure that there were no specific biases, i.e. batch effects. One additional individual was excluded from the analysis due to a much higher missing genotyping rate than the others, leaving us with a dataset of 1482 individuals and 389'717 SNPs to impute.

Genotype imputation

The data were imputed in collaboration with the group of Prof. Olivier Delaneau at the Department of Computational Biology, UNIL. PhD student Robin Hofmeister did the imputation of chromosome 1 to 22 in the following way: (1) ensuring SNPs are consistent with the hg19 reference genome assembly using bcftools [Danecek et al., 2021], (2) phasing to infer haplotypes from the genotype data using SHAPEIT4 v4.2.1 [Delaneau et al., 2019] and the HRC reference panel [McCarthy et al., 2016], (3) imputing alleles using IMPUTE5 [Rubinacci et al., 2020] and the HRC reference panel, and (4) filtering of imputed SNPs based on $\text{INFOscore} \geq 0.8$ using bcftools, which

rendered a total of 18,026,053 SNPs. I then applied the same filters as for the genotype quality control excluding 169,257 SNPs that failed the Hardy Weinberg test (p -value < 0.0001) and 9,715,525 SNPs with $MAF < 0.01$. In addition, 49 individuals were excluded for having opposite genetically inferred gender from the X chromosome prior imputation (using plink 1.9 [Purcell et al., 2007]) and reported gender in the phenotypic data. Finally, to validate the sufficient quality of the data, I performed two additional checks. First, I compared MAF of SNPs present in our data and in the UK Biobank which resulted in a correlation of 0.98 (p -value $< 2.2e-16$). Second, I predicted maternal height in our data using GMRM [Orliac et al., 2021] predictors from the UK Biobank. Again, the accuracy was reassuring with a correlation of 0.47 (p -value $< 2.2e-16$). As a comparison, this value is approximately 0.6 when predicting height in the Estonian Genome Centre data (see Appendix A).

The final dataset includes 8,141,271 SNPs and 1470 individuals, of which 814 are babies, 649 are mothers, and 7 could not be linked to phenotypic information. Of the 649 mothers, 5 had incomplete phenotypic information. These were not included, leaving 644 mothers in the following analysis. We also validated 429 maternal-infant pairs from the phenotypic data that had a Kinship coefficient > 0.17 and an $IBS0 < 0.0012$ computed using the KING software [Manichaikul et al., 2010]. Figure 7 shows the ethnic background across maternal samples and different sample sizes according to four major pregnancy-related complications: PTB, PPH, HDP and GDM. These complications are defined using the delivery report and ICD-10 classification, as described in the methods section of the manuscript in Appendix B.

Prediction into the Maternity CHUV cohort

Given the low number of maternal samples (649 women) and as it is becoming increasingly evident that the genetics of complex diseases are characterized by a large number of markers with tiny effects, we would most likely have very limited statistical power to conduct any SNP-trait association study. To compare, a recent paper published a multi-ancestry GWAS of GDM in which five loci were identified as genome-wide significant in a much larger sample: 347,856 controls and 5,458 cases [Pervjakova et al., 2022]. For this reason, I decided to investigate the genetics of GDM, HDP, PTB and PPH cases through different prediction models.

Prediction of GDM, HDP, PTB and PPH - to predict the four complications in the maternal samples, I used posterior mean effect sizes (mean of each SNP beta included in the model) obtained for 30 UK Biobank complex traits and generated using the GMRM model as described in [Orliac et al., 2021]. With Sven Erik Ojavee from, a PhD student in the group, we additionally applied

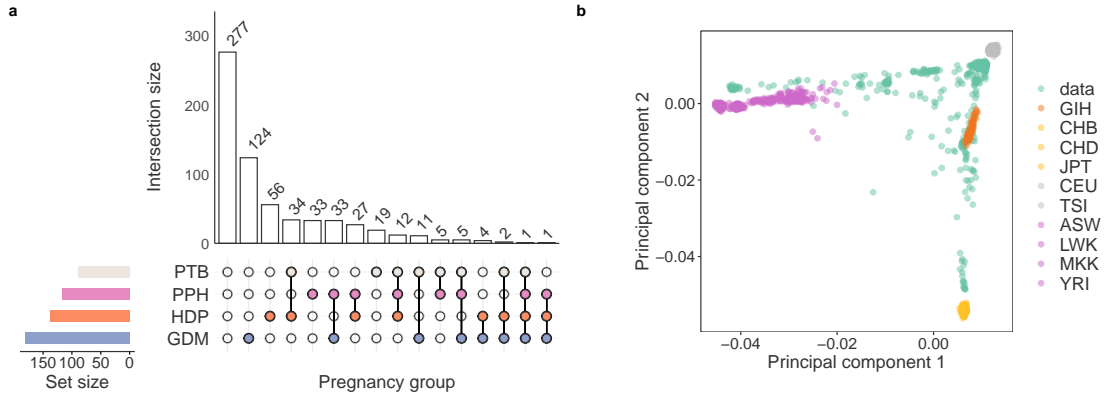


Figure 7. The Maternity CHUV cohort. (a) Upset plot showing the number of women in the data who experienced premature birth (PTB), post-partum hemorrhage (PPH), hypertensive disorders of pregnancy (HDP), gestational diabetes mellitus (GDM), a combination or none of the above. The data includes 277 women who had none of these 4 complications and 367 pregnancies presenting one or more complications. Among these, there are 181 cases of GDM, 137 of HDP, 89 of PTB, and 117 of PPH. (b) I combined the 644 maternal genotypes with the genotypes from the 1000G study and then did a PCA to map our genotyped individuals (data) to individuals with known ancestries in the reference panel. The scatter plot shows the first two PC reflecting ancestry diversity across maternal samples. GIH: Gujarati Indians in Houston, Texas, USA. CHB: Han Chinese in Beijing, China. CHD: Chinese in metropolitan Denver. JPT: Japanese in Tokyo, Japan. CEU: Utah residents with ancestry from northern and western Europe. TSI: Toscani in Italy. ASW: African ancestry in SW USA. LWK: Luhya in Webuye, Kenya. MKK: Maasai in Kinyawa, Kenya. YRI: Yoruba in Ibadan, Nigeria.

GMRM to GDM for 187'299 women in the UK Biobank data genotyped at 2,174,071 pruned SNP markers ($LD R^2 \geq 0.8$ within a 1Mb window, using plink 1.9 [Purcell et al., 2007]). Of the 187,299 women, 794 self-reported that a physician told them they had diabetes and that they had diabetes only during pregnancy. These women were set as GDM cases. The remaining 186,505 controls are women who self-reported having 1, 2 or 3 live births. We adjusted the phenotype for age, age squared, east-west coordinates, UK biobank centre, genotype batch, top 20 genotypic PCs and number of live births. We applied GMRM running one chain for 6000 iterations (burn-in 200) with SNP markers split into 8 groups: 4 MAF quartiles, each split in 2 LD bins. Each group was modelled with a mixture of four normal distribution with variance 0.0001, 0.001, 0.01. 0.1 and a dirac spike at zero. The 31 traits with the number of SNPs overlapping the imputed maternal genotypes, thus included in the prediction analysis are presented in Table 3. I then multiplied the posterior means from each trait to the standardized maternal genotypes creating trait-specific genetic predictors of GDM, PPH, PTB, and PPH for each woman. GDM, HDP, PTB, and PPH were adjusted for maternal age, SNP genotype plate and top 10 PCs. The effects of GMRM markers are estimated from traits that were standardized to a z-score prior to analysis, which is why I also standardized the maternal phenotypes. I predicted all four complications in 644 women and repeated the analysis, restricting the sample size to 484 women of European nationality. In Figure 8, I show the correlation between the genetic predictors and the observed GDM, HDP, PTB and PPH outcomes, for all traits and in both populations.

Table 3. GMRM predictors from the UK Biobank.

Phenotype	Complex trait	SNP count	
Disease	T2D	Type-2-diabetes	1,151,856
	CAD	Coronary artery disease	1,146,095
	BP	High blood pressure	1,153,045
	GDM	Gestational diabetes mellitus	1,154,257
Measure	HT	Height	1,154,056
	BMI	Body mass index	1,152,423
	BMD	Heel bone mineral density T-score	1,152,979
	DBP	Diastolic blood pressure	1,154,144
	SBP	Systolic blood pressure	1,153,217
	FVC	Forced vital capacity	1,152,603
CBC	RBC	Red blood cell count	1,153,266
	HMC	Haemoglobin count	1,148,084
	HAC	Haematocrit count	1,153,408
	MCV	Mean corpuscular volume	1,139,850
	MCH	Mean corpuscular haemoglobin	1,150,445
	MCHC	Mean corpuscular haemoglobin concentration	1,151,248
	RCDW	Red cell distribution width	1,144,813
	WBC	White blood cell count	1,131,373
	NEUT	Neutrophil count	1,154,029
	LYMPH	Lymphocyte count	1,122,028
	MONO	Monocyte count	1,122,956
	EOSI	Eosinophil count	1,151,763
	PLATE	Platelet count	1,151,898
	MPV	Mean platelet volume	1,130,286
BASO	Basophil count	1,138,607	
Biomarker	HbA1c	Glycated haemoglobin	1,152,760
	GLU	Glucose	1,149,618
	CRET	Creatinine	1,148,941
	CHOL	Cholesterol	1,141,994
	HDL	High-density lipoprotein cholesterol	1,153,869
	LDL	Direct low-density lipoprotein cholesterol	1,152,929

Zooming in on GDM - here, I focus on the prediction of GDM only in 484 women of European nationality. I combined the genetic predictors of women previously calculated from GDM, T2D, GLU and HbA1c, to explore whether a combined genomic risk score would improve our prediction of GDM. To do so, I combined these genetic predictors in a linear regression model, in which the beta effects effectively weight the contribution of each genetic predictor to the phenotype [Maier et al., 2018]. The square root of the r-squared then gives us the the correlation between the observed and the predicted phenotype. I constructed a second combined genomic risk score called metaGRS [Inouye et al., 2018], which consists of a weighted average of the normalized genetic predictors. As described in their supplementary information, I calculated the metaGRS from the

correlation between GDM, T2D, GLU and HbA1c genetic predictors and the independent effect of each one on the phenotype, estimated from a linear regression model. In Figure 9a, I compare the correlation between the observed and predicted phenotype, obtained from the two multi-trait models to our previous results. In Figure 9b, I stratified the single- and multi-trait predictors to select women in the top 10%, presenting a higher risk for GDM, and applied a logistic regression of the top 10% on GDM to directly compare individuals at high and lower risk.

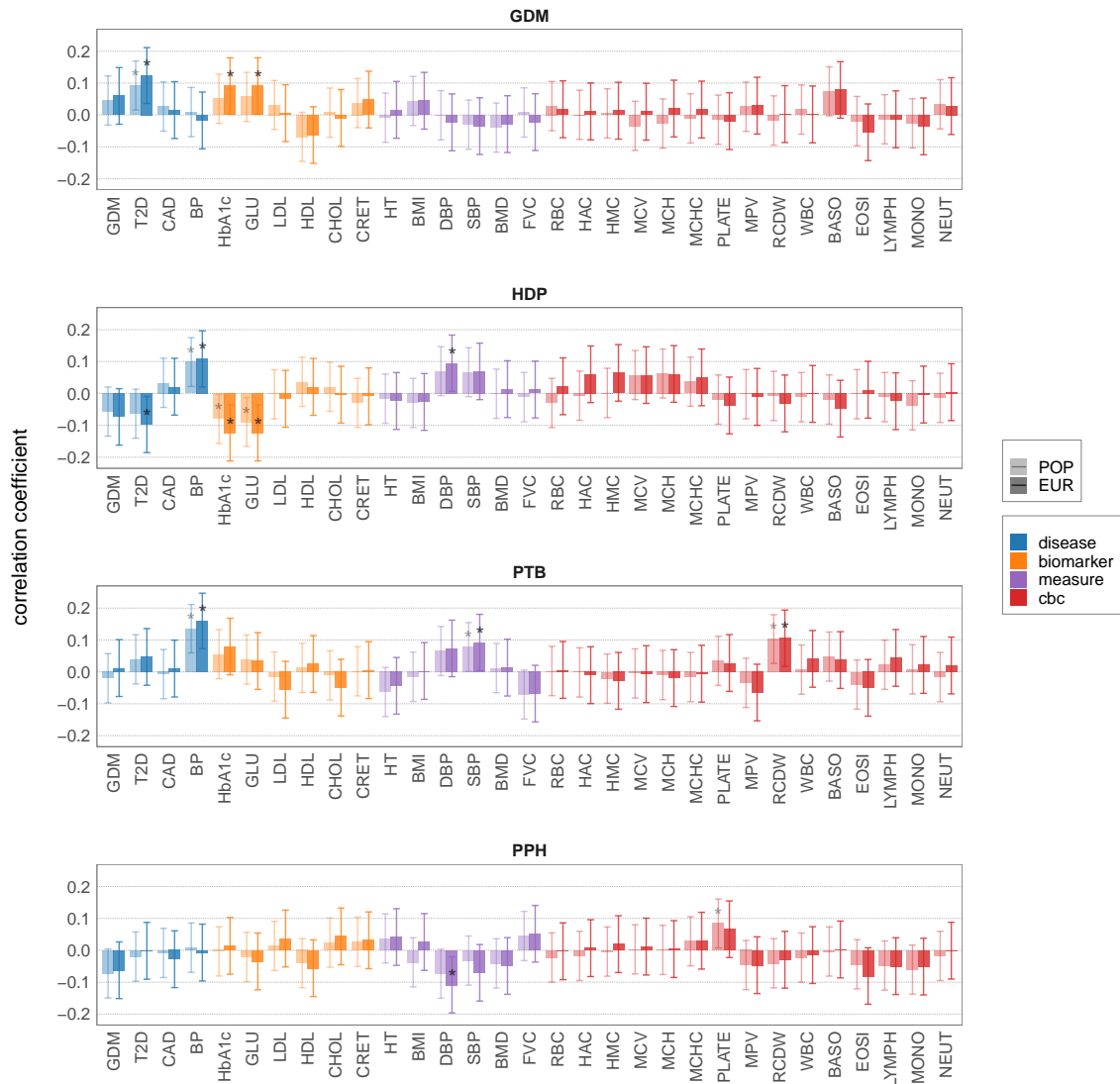


Figure 8. Out-of-sample prediction of major maternal complications. Bar plots with error bars giving 95% confidence intervals of the correlation between the observed and the predicted phenotype when predicting GDM, HDP, PTB and PPH from GMRM predictors of 31 traits in the UK Biobank (on the x-axis). The asterisk (*) indicates a correlation at p-value < 0.05. Results are presented for the prediction of the four complications in 644 women of diverse ancestry and in 484 women of European nationality.

Results

Prediction of GDM, HDP, PTB and PPH

Altogether, results show meaningful correlations, such as high genetic risk for T2D increasing the risk of GDM (Figure 8). Indeed, we find a correlation at p -value < 0.05 of 0.093 (95% CI 0.016, 0.169) for T2D when predicting GDM in women of diverse ancestry. Although not significant, we find a positive correlation for GDM too, which is promising given the lower sample size for the GMRM model of GDM: 794 cases for 186,505 women against 28,230 T2D cases for 428,747 individuals. Similarly, we observe that a high genetic risk for BP increases the risk of HDP with a correlation of 0.099 (95% CI 0.022, 0.175) at p -value < 0.05 . For HDP, we also find two negative correlations: -0.081 (95% CI -0.157, -0.003) for HbA1c and -0.90 (95% CI -0.166, -0.013) for GLU, which indicates that the predictors associated with high HbA1c and GLU levels would decrease the genetic risk of HDP. Moreover, we observe an overall improvement in prediction from the 31 traits when restricting the sample to women of European nationality. This is expected because the GMRM models were applied in the UK Biobank to a sample of European-ancestry individuals. For GDM, predictions from T2D, HbA1c and GLU in women of European nationality, are higher and now all significant at p -value < 0.05 with correlations of 0.125 (95% CI 0.036, 0.211), 0.092 (95% CI 0.002 0.179) and 0.092 (95% CI 0.003 0.179) respectively. We also find stronger positive and negative correlations for HDP with two new signals at p -value < 0.05 . A positive correlation of 0.095 (95% CI 0.006, 0.183) for DBP and a negative one of -0.098 (95% CI -0.186, -0.009) for T2D.

The prediction of PPH cases yielded a negative correlation at p -value < 0.05 of -0.109 (95% CI -0.197, -0.020) for DBP. More interestingly, when using predictors associated with high PLATE count, we see a more pronounced increase in the risk of PPH when including all women of diverse ancestry. We find a correlation of 0.085 (95% CI 0.007, 0.161) with p -value = 0.032 against 0.067 (95% CI -0.023, 0.155) with p -value = 0.143 in women of European nationality only, which could imply a lower genetic risk of PPH in women of European nationality. Finally, we find positive correlations at p -value < 0.05 for BP, SBP, and RCDW traits when predicting PTB in both women of diverse and European nationality. It is important to note that PTB is particularly difficult to explore because it can occur spontaneously or be induced, i.e. in case of an emergency c-section. In addition, PTB may also be associated with other maternal conditions such as HDP, which itself can lead to premature pregnancy induction due to severe hypertension (see Appendix B).

For significance testing I have used a nominal p -value of 0.05 and the correlations found in the results show fairly wide confidence intervals, which is expected given the size of our sample and

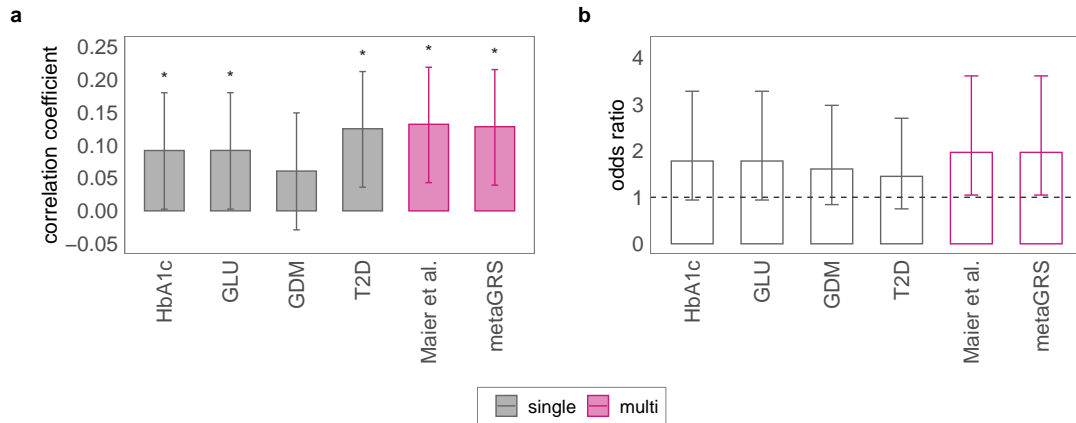


Figure 9. Maximising the prediction of GDM in European maternal samples. (a) Bar plots with error bars giving 95% confidence intervals of the correlation between the observed and the predicted phenotype, when predicting GDM from GMRM predictors of HbA1c, GLU, T2D, GDM and from the multi-trait [Maier et al., 2018] and metaGRS [Inouye et al., 2018] models. (b) Bar plots with error bars giving 95% confidence intervals of odds ratio of developing GDM for women with genetic predictors of HbA1c, GLU, T2D, and GDM in the top 10%. The asterisk (*) indicates a correlation at p-value < 0.05. Results are presented for the prediction of GDM in 484 women of European nationality.

taking into account that maternal phenotypes occur in a very specific setting that is pregnancy (Figure 8). Nevertheless, these preliminary results suggest several leads regarding a woman’s genetic risk of having GDM, HDP, PTB, or PPH. They show (1) that it is possible to predict a woman’s genetic risk of having a pregnancy-related complication, to a small extent, even before she becomes pregnant based on her DNA and (2) that genes influencing the 31 different traits in the general population might be associated with maternal complications.

Zooming in on GDM

I further explored the prediction of GDM by combining previously identified genetic predictors at p-value < 0.05, namely T2D, HbA1c, and GLU, in addition to GDM in the 484 women of European nationality (Figure 9). Results show that we can improve prediction if we combine genetic predictors together in a sensible weighting. Indeed, we find a correlation at p-value < 0.05 of 0.131 (95% CI 0.043, 0.218) for the [Maier et al., 2018] predictor and 0.128 (95% CI 0.039, 0.214) for the metaGRS predictor compared to 0.125 (95% CI 0.036, 0.211) for the T2D. Although the difference is small, results also suggest a better weighting of genetic predictors for the [Maier et al., 2018]. When we stratify the single- and multi-trait predictors to select women at higher risk for GDM (top 10%), we observe that the multi-trait models find a subset of individuals more prone to GDM than the single models. We also observe in Figure 9a that T2D has a higher prediction accuracy across individuals than HbA1c and GLU. However, in Figure 9b, the odds ratio is higher for HbA1c and GLU, although not significant. This may indicate that the genetics of biomarkers

directly involved in the mechanism of GDM may be more informative than a woman’s genetic risk of T2D. Given the size of our 95% confidence intervals, this is obviously a hypothesis to be explored.

Limitations of the prediction analysis

These results show that we can better understand maternal quantitative traits by examining which genetic predictor correlates with pregnancy-related complications (Figure 8). For example, the results obtained from the prediction of GDM are in line with the latest published GWAS, where they report a positive genetic correlation between GDM and other glycaemic traits including T2D, GLU and HbA1c [Pervjakova et al., 2022]. In both analysis, the respective correlation was highest for T2D, confirming once again that GDM and T2D share genetic risk factors. In addition, we improve the genetic prediction of GDM by combining relevant predictors into a single multi-trait predictor (Figure 9).

The negative correlation of -0.098 (95% CI -0.186, -0.009) between the HDP observed phenotype and the predicted phenotype from T2D highlights a limitation of the analysis (Figure 8). A recent study found a genetic correlation of more than 0.3 between pre-eclampsia and BP, SBP, DBP and T2D [Steinthorsdottir et al., 2020]. Our results point towards a negative genetic correlation between pre-eclampsia and T2D rather than a positive one. This is surprising and could simply be due to chance as I use a nominal p-value of 0.05. It could also be due to the selection of controls in the analysis. For each phenotype, I compared the cases to the rest of the women in the cohort. This choice could be a potential issue for HDP because controls then included 277 women who had no complications during their pregnancy and 168 cases of GDM. Moreover, the number of cases overlapping the two complications is only 8 (Figure 7). As GDM is positively correlated with T2D, the prevalence of GDM cases in the controls could explain the resulting negative correlation. Another difference with [Steinthorsdottir et al., 2020] is the phenotype itself. We defined HDP as all women who reported gestational pregnancy-induced hypertension, preeclampsia, HELLP syndrome and unspecified maternal hypertension (ICD10 codes O13,O14 and O16) to increase the number of cases to 137. This highlights a second limitation related to study design, namely the difference in phenotype definition and the limited number of cases available across studies. This is particularly true for maternal health studies as there is a significant lack of data to study pregnancy [Barbitoff et al., 2020].

Moreover, the prediction of complications would most likely increase if pregnancy-specific genetic associations were included. For example, for GDM, we would ideally want to use HbA1c

mean posterior effect sizes estimated in pregnant women rather than in the general population. Two questions we might ask are: are there pregnancy-specific effects? And is there a difference in prediction when using genetic predictors calculated from HbA1c levels in the general population and in pregnant women? We could also compare the prediction accuracy for GDM between potential SNP markers associated with high HbA1c in the first and in the second trimester of pregnancy. In addition, to improve GDM prediction, we could add actual clinical measurements of HbA1c and GLU in women in early pregnancy to the genetic risk scores. Finally, to further study the genetics of pregnancy-related complications, we could apply an alternative and recently published method named pleiotropic decomposition regression (PDR) [Ballard and O'Connor, 2022]. PDR identifies genetic components that are shared across genetically correlated complex traits. This method could help us clarify the relationship between pregnancy-related traits and other complex traits as well as identify shared genetic risk factors involved.

Discussion

The introduction outlines the substantial efforts in quantitative genetics to study the underlying genetic architecture of complex traits. It also introduces quantitative genetics in maternal health and how GWAS can contribute to obstetric medicine. Over the past four years, I have explored how to better model LD in genetic data with the BayesRR-RC method and contributed to maternal research through the CHUV maternity cohort. In the discussion, I present potential extensions to BayesRR-RC and share perspectives on the work I have undertaken in maternal health.

BayesRR-RC: extensions and future directions

Fully described in Appendix A, the resulting BayesRR-RC model assumes additive genetic effects $\beta_\varphi \in \mathbb{R}^{M_G \times 1}$ split into φ groups over a trait $\mathbf{y} \in \mathbb{R}^{N \times 1}$ such that:

$$\mathbf{y} = \mathbf{1}\mu + \sum_{\varphi=1}^g \mathbf{X}_\varphi \beta_\varphi + \epsilon \quad (3)$$

where there is a single intercept term $\mathbf{1}\mu$ and a single error term ϵ . We assume that the genotype matrices \mathbf{X}_φ have been centered and scaled to unit variance. SNPs are allocated into groups $\varphi = (1, \dots, g)$, each of which having its own set of model parameters $\Theta_\varphi = \{\beta_\varphi, \pi_\varphi, \sigma_\varphi^2\}$, where β_{φ_j} is distributed according to:

$$\beta_{\varphi_j} \sim \pi_{0\varphi} \delta_0 + \pi_{1\varphi} \mathcal{N}(0, \sigma_{1\varphi}^2) + \pi_{2\varphi} \mathcal{N}(0, \sigma_{2\varphi}^2) + \dots + \pi_{L\varphi} \mathcal{N}(0, \sigma_{L\varphi}^2) \quad (4)$$

for each marker j from group φ . For each φ group, $\{\pi_{0\varphi}, \pi_{1\varphi}, \dots, \pi_{L\varphi}\}$ are the mixture proportions, $\{\sigma_{1\varphi}^2, \sigma_{2\varphi}^2, \dots, \sigma_{L\varphi}^2\}$ are the mixture specific-variances where for any mixture k , $\sigma_{k\varphi}^2 = C_k^\varphi \sigma_\varphi^2$ and δ_0 is a discrete probability mass at zero. The mixture proportions, the mixture constants C_k^φ and the group variance explained by the SNP markers σ_φ^2 are all unique and independent across SNP marker groups.

With a single command line, BayesRR-RC provides (1) unbiased MAF-LD annotation-specific joint genetic effect size estimates, (2) the total SNP heritability and the SNP-heritability of annotations, (3) the probability that each SNP, genetic region or annotation is associated with a phenotype and (4) a posterior predictive distribution for each individual in genomic prediction. This framework, informs us about the number of SNPs entering the model and their magnitudes. SNPs entering the model can be fully explored with greater confidence because obtained proba-

bilities also convey the uncertainty of the estimates. The BayesRR-RC model is able to answer core questions raised when it comes to the genetic architecture of complex traits, i.e. how many SNPs and genes are involved? what is the contribution of coding and non-coding regions to the phenotypic variance and to the susceptibility of an individual to a disease? [Loos, 2020].

The obvious next question is how best to use BayesRR-RC to explore the complex underlying biology of traits? The model can be used in a general way to compare the architecture across multiple traits, or on a case-by-case basis using the prior biological knowledge specific to each trait. For instance, we could investigate T2D and CAD using trait-specific annotations from Gene Ontology terms [Consortium, 2019] associated with each trait. However, inference on specific annotations will depend on their quality and combination used in the model. As shown in simulation, randomly assigning SNPs to different annotations does not fit any enrichment pattern and breaks the model with the genetic variance being evenly split across groups. Furthermore, SNPs may contribute to multiple annotations and one way to extend our model would be to accommodate for annotation overlap. This could be implemented by allowing markers to swap groups. For instance, let’s consider an overlap with SNP marker j assigned to groups 1 and 2. Then, from the BayesRR-RC Gibbs sampling algorithm (see Appendix A), we would compute the inclusion probability of β_j , where β_j would be distributed according to:

$$\beta_j \sim \sum_{\varphi=1}^2 (\pi_{0\varphi} \delta_0 + \pi_{1\varphi} \mathcal{N}(0, \sigma_{1\varphi}^2) + \pi_{2\varphi} \mathcal{N}(0, \sigma_{2\varphi}^2) + \dots + \pi_{L\varphi} \mathcal{N}(0, \sigma_{L\varphi}^2)) \quad (5)$$

for marker j from group $\varphi = (1, 2)$. Here, we model β_j with a prior that follows a mixture of Gaussian probability densities including a discrete spike at zero, the Gaussian probability densities set on group 1 and the Gaussian probability densities set on group 2. Modeling annotation overlap has already been considered in GWAS summary-based methods, such as s-LDSC [Finucane et al., 2015]. Modeling it in the BayesRR-RC model adds an additional layer of biological information to the individual-level data, possibly improving our inference of genetic architectures.

Additional work may be carried out in different directions. First, computationally, to cope with the increase in the number of markers tested, which with genome sequencing will exceed 150 million variants [Loos, 2020]. This is key because, although we provide LD-unbiased estimates, we use imputed genotype data from a specific population, which may poorly represent rare variants. And findings from 15 years of GWAS seem to suggest that SNPs yet to be discovered are either common, with tiny effect sizes, or rare [Yong et al., 2020]. Second, a major criticism of

GWAS is the diversity in ancestry of the populations studied. European individuals are clearly over-represented and additional efforts are needed to include other ancestries. This is important because health disparities may follow with inequalities in care based on genetic research [Loos, 2020]. Although we have shown that BayesRR-RC accounts for data structure and does a better job than standard GWAS, it would be interesting to test it empirically and understand how best to adjust for multiple ancestries in a sample. Third, with the increase of omics data, we could also set up the model to estimate the contribution of other omics data, i.e. copy number variants or transcriptomic data, as in [Banos et al., 2020] where the contribution of SNPs and methylation probes are jointly estimated in the BayesR framework. Finally, it is important to keep in mind that markers in LD tagging the same effect are correlated and interchangeable in the model (we cannot tell which SNP is causal), which is why we use the PPWV [Fernando et al., 2017] to control for false positives and accurately fine-map regions that contribute to the phenotypic variance of complex traits. Here, additional work could be conducted to fine-map causal SNPs.

Further extended in GMRM, multiple phenotypes can now be analysed simultaneously within the BayesRR-RC modelling framework [Orliac et al., 2021]. GMRM can also use joint estimates to improve marginal SNP effect estimation by adjusting the phenotype using LOCO (leave one chromosome out) and return a marginal summary statistic for GWAS discovery. This extension allows for easier usage of Bayesian model and will hopefully familiarise and expose users to a Bayesian approach in quantitative human genetics.

Maternal complex traits

During pregnancy, women experience physiological changes to facilitate the growing foetus and to prepare for labour [Soma-Pillay and Catherine, 2016]. Understanding these changes is important to improve prevention, early diagnosis and care for women during pregnancy, labour and post-partum. In Chapter 2, we demonstrated that routine CBC can be sufficiently sensitive to identify unusual patterns linked to obstetric complications early in pregnancy and possibly improve the stratification of high-risk pregnancies. Genetic predictors such as PRS have also been proven good enough to stratify individuals in the extremes with high or low genetic risks [Yong et al., 2020]. By further combining PRSs in a multi-trait predictor for GDM, we are able to highlight those that are most informative and improve the stratification of women who are more likely to experience the complication (Figure 9).

Much like other maternal cohorts, the CHUV maternity cohort is limited by its sample size to fully explore the genetic architecture of maternal traits. The small sample sizes available in maternity research are a real challenge, largely due to study design [Barbitoff et al., 2020]. First, women have long been considered vulnerable individuals in research, which complicates ethical considerations regarding maternal data and leads to extremely limited participation of pregnant women in clinical trials [Biggio, 2020]. Second, there are disparities in data collection and analysis. Diagnoses are not standardised and there are no clear guidelines on inclusion and exclusion criteria for samples [Barbitoff et al., 2020]. For example, in my thesis, cases of GDM are defined using the ICD-10 O24 code classification. At the CHUV, pregnant women are screened for GDM by fasting blood glucose test and if necessary by blood glucose at 1 hour and 2 hours after ingestion of 75 g of sugar also known as an oral glucose challenge test. In the [Lamri et al., 2020] study, GDM was diagnosed via an oral glucose challenge test whereas in the UK Biobank, cases are self-reported, which may result in misclassification and limit the number of cases. In [Lamri et al., 2020], they redefine cases of GDM in the UK Biobank by selecting women who have had at least one pregnancy. Similarly, I have included women who have had 1 to 3 pregnancies reducing the number of cases in the prediction analysis from GDM posterior mean effect sizes. These disparities complicate comparisons between cohorts and across maternal studies. Finally, and as discussed earlier, it is important to include populations of different ancestries so that translational research can be useful worldwide and tailored to each individual. Interestingly, I have observed that maternal studies are quite specific to one ethnicity, probably reflecting the considerable efforts around the world to generate maternal data. GWAS design consists of discovery steps followed by replication steps, achieved with strict thresholds for significance testing [Loos, 2020]. Few studies each on a specific ancestry adds to the difficulty of making reproducible findings across maternal health, i.e. in [Kwak et al., 2012], the *CDKAL1* and *MTNR1B* genes are found to be associated with GDM in Korean women while the [Wu et al., 2021] study identifies four other genes in Chinese women.

In addition to large-scale genetic studies, the field needs data specific to pregnancy to understand molecular pathways and improve prediction of complex maternal traits. We show that genetic risk factors influencing the 31 selected traits in the general UK Biobank population can be predictive of maternal complications. However, pregnancy is a complex environment where physiological, physical and metabolic changes occur in women and the prediction of complications would benefit from genetic predictors specific to pregnancy such as SNP markers associated with high GLU levels early in pregnancy. Moreover, both mother and child have an interest in staying alive so they cooperate, but they are also in conflict [Boddy et al., 2015]. The child tries to receive optimal resources through the placenta, while the mother will try to compensate for it to keep her

metabolism functioning efficiently. Not only that, but the maternal and fetal systems are intrinsically linked as cells and tissues interact. For example, the [Rasmussen et al., 2022] study published in January shows that it is possible to explore gene transcripts involved in pregnancy using cell-free RNA taken from blood at different weeks of gestation. Their analysis is sensitive enough to track maternal, fetal and placental changes and confidently predict pre-eclampsia before it occurs based on RNA profiles. In line with our manuscript on haematological changes (see Appendix B), this study also shows that a simple blood test has enormous potential to monitor normal pregnancy and identify complications as early as possible. Maternal and fetal genetics also interact and may, independently or together, influence a maternal trait, which complicates the analysis [Zhang et al., 2018]. In GWAS listed in Table 1, fetal effects have been associated with a maternal phenotype, namely preeclampsia and preterm birth. Lastly, whole genome sequencing data might enhance new discoveries in maternal health as some pregnancy complications leading to adverse outcomes might be disfavored by evolution [Barbitoff et al., 2020].

Despite the limitations presented here, I have shown that we can use prediction to explore the genetic basis of maternal health. Future directions could include addressing specific questions such as: Does genetics tell us more than a blood test? Can genetics predict the changes in blood measurements that we see during pregnancy? We could also explore other traits that may be linked to complications, such as insulin levels for GDM, and potentially improve the multi-trait prediction. A particular challenge in maternal health is that the clinic and research must work in close collaboration, as pregnancy is a whole system where women undergo rapid changes that vary between individuals. For instance, it would be interesting to integrate clinical factors into the multi-trait predictor, i.e. HbA1c and GLU measurements taken at 20 weeks of gestation combined with targeted regions of the genome associated with T2D, high GLU and high HbA1c levels, to predict a woman's risk of GDM. Moreover, the CHUV maternity cohort includes mother-infant pairs that could be used to explore maternal and fetal effects in a prediction or replication analysis. Indeed, this unique cohort is a valuable resource because it can be used as a validation dataset in maternal-fetal health. Because it is phenotypically rich, maternal samples can also be used to validate other outcomes such as the association between Rhesus blood group and haematological traits [Auwerx et al., 2022].

Conclusion

To conclude, BayesRR-RC is a new implementation of BayesR that overcomes the increased time and memory cost of Bayesian inference while improving prediction accuracy. The inferred posterior distributions help the user to directly fine-map and draw conclusions over the genetic architecture of traits including maternal complex traits, and highlight the immense challenge of understanding the molecular underpinning of each association.

GWAS have been carried out for 15 years and their popularity continues to grow with the advances in technology and the increase in omics data [Loos, 2020]. Remarkably, the UK Biobank has just published WGS for 200,000 participants in November 2021 and made it available to researchers. As most of the genome remains to be explored, large-scale WGS data will bring new insights into the genetic basis of complex traits and human health.

In the future, considerable efforts will be required to combine omics data in GWAS. Targeted high quality data will be needed to fully explore the underlying molecular mechanisms of disease. This is critical in the context of maternal health, where pregnancy is a complex, time-limited environment with rapid changes that many women experience. Larger studies on diverse populations will further increase discovery and reduce clinical disparities arising from quantitative genetics. Researchers would also benefit from additional independent cohorts made publicly available. These would give a wider perspective towards personalised medicine because we could evaluate at which group of individuals a predictor is no longer recommended, i.e. populations with different ancestries or different pregnancy status for women.

A key step in personalised medicine will be the inclusion of genetic data in medical records to provide the most appropriate individual PRS for a given trait or disease. To achieve this goal, ethical and legal concerns need to be addressed, and the understanding of genetics by a wider public needs to be improved. Quantitative genetics has made significant contributions to human health encouraging the transition towards a genome-based precision medicine.

References

- Altshuler, D., Donnelly, P., Consortium, I. H., et al. (2005). A haplotype map of the human genome. *Nature*, 437(7063):nature04226.
- Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature protocols*, 5(9):1564–1573.
- Antwi, E., Amoakoh-Coleman, M., Vieira, D. L., Madhavaram, S., Koram, K. A., Grobbee, D. E., Agyepong, I. A., and Klipstein-Grobush, K. (2020). Systematic review of prediction models for gestational hypertension and preeclampsia. *PLoS One*, 15(4):e0230955.
- Ardissino, M., Slob, E. A., Millar, O., Reddy, R. K., Lazzari, L., Patel, K. H. K., Ryan, D., Johnson, M. R., Gill, D., and Ng, F. S. (2022). Maternal hypertension increases risk of preeclampsia and low fetal birthweight: Genetic evidence from a mendelian randomization study. *Hypertension*, 79(3):588–598.
- Ardlie, K. G., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 3(4):299–309.
- Auwerx, C., Lepamets, M., Sadler, M. C., Patxot, M., Stojanov, M., Baud, D., Mägi, R., Esko, T., Metspalu, A., Milani, L., et al. (2022). The individual and global impact of copy-number variants on complex human traits. *The American Journal of Human Genetics*.
- Ballard, J. L. and O’Connor, L. J. (2022). Shared components of heritability across genetically correlated traits. *The American Journal of Human Genetics*.
- Banos, D. T., McCartney, D. L., Patxot, M., Anchieri, L., Battram, T., Christiansen, C., Costeira, R., Walker, R. M., Morris, S. W., Campbell, A., et al. (2020). Bayesian reassessment of the epigenetic architecture of complex traits. *Nature Communications*, 11(1):1–14.
- Barbitoff, Y. A., Tsarev, A. A., Vashukova, E. S., Maksiutenko, E. M., Kovalenko, L. V., Belotserkovtseva, L. D., and Glotov, A. S. (2020). A data-driven review of the genetic factors of pregnancy complications. *International journal of molecular sciences*, 21(9):3384.
- Biggio, J. R. (2020). Research in pregnant subjects: Increasingly important, but challenging. *Ochsner Journal*, 20(1):39–43.
- Boddy, A. M., Fortunato, A., Wilson Sayres, M., and Aktipis, A. (2015). Fetal microchimerism and maternal health: a review and evolutionary analysis of cooperation and conflict beyond the womb. *BioEssays*, 37(10):1106–1118.
- Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, 169(7):1177–1186.
- Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295.
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O’Connell, J., et al. (2018). The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209.
- Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W. M., Ritchie, G. R., et al. (2011). Modernizing reference genome assemblies. *PLoS biology*, 9(7):e1001091.
- Consortium, . G. P. et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68.

- Consortium, G. O. (2019). The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338.
- Consortium, I. H. et al. (2007). A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851.
- Consortium, T. E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2). giab008.
- Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L., and Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature communications*, 10(1):1–10.
- Durbin, R. M. and Altshuler, t. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.
- Dwivedi, O. P., Lehtovirta, M., Hastoy, B., Chandra, V., Krentz, N. A., Kleiner, S., Jain, D., Richard, A.-M., Abaitua, F., Beer, N. L., et al. (2019). Loss of znt8 function protects against diabetes by enhanced insulin secretion. *Nature genetics*, 51(11):1596–1606.
- Erbe, M., Hayes, B. J., Matukumalli, L. K., Goswami, S., Bowman, P. J., Reich, C. M., Mason, B. A., and Goddard, M. E. (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):4114–4129.
- Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of human genetics*, 29(1):51–76.
- Falconer, D. S. and Mackay, T. F. (1983). *Quantitative genetics*. Longman.
- Fejzo, M. S., Sazonova, O. V., Sathirapongsasuti, J. F., Hallgrímsson, I. B., Vacic, V., MacGibbon, K. W., Schoenberg, F. P., Mancuso, N., Slamon, D. J., and Mullin, P. M. (2018). Placenta and appetite genes *gdf15* and *igfbp7* are associated with hyperemesis gravidarum. *Nature communications*, 9(1):1–9.
- Fernando, R., Toosi, A., Wolc, A., Garrick, D., and Dekkers, J. (2017). Application of whole-genome prediction methods for genome-wide association studies: a bayesian approach. *Journal of Agricultural, Biological and Environmental Statistics*, 22(2):172–193.
- Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P. R., Anttila, V., Xu, H., Zang, C., Farh, K., Ripke, S., Day, F. R., Purcell, S., Stahl, E., Lindstrom, S., Perry, J. R., Okada, Y., Raychaudhuri, S., Daly, M. J., Patterson, N., Neale, B. M., and Price, A. L. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11):1228–1235.
- Flannick, J., Thorleifsson, G., Beer, N. L., Jacobs, S. B., Grarup, N., Burt, N. P., Mahajan, A., Fuchsberger, C., Atzmon, G., Benediktsson, R., et al. (2014). Loss-of-function mutations in *slc30a8* protect against type 2 diabetes. *Nature genetics*, 46(4):357–363.
- Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. (2002). The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229.
- GENOMICS, H. (2010). Integrating common and rare genetic variation in diverse human populations. *NATURE REVIEWS/ GENETICS*, 11:20.

- Gong, J., Mei, S., Liu, C., Xiang, Y., Ye, Y., Zhang, Z., Feng, J., Liu, R., Diao, L., Guo, A. Y., Miao, X., and Han, L. (2018). PanCanQTL: Systematic identification of cis -eQTLs and trans -eQTLs in 33 cancer types. *Nucleic Acids Research*, 46(D1):D971–D976.
- GTEx Consortium, G. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multi-tissue gene regulation in humans. *Science (New York, N. Y.)*, 348(6235):648–60.
- Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B. V., Hjartarson, E., et al. (2015). Large-scale whole-genome sequencing of the icelandic population. *Nature genetics*, 47(5):435–444.
- Guennebaud, G., Jacob, B., Lenz, M., et al. (2015). Eigen v3, 2010.
- Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., De Geus, E. J., Boomsma, D. I., Wright, F. A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245.
- Gusev, A. and et al., L. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *American Journal of Human Genetics*, 95(5):535–552.
- Hao, X., Zeng, P., Zhang, S., and Zhou, X. (2018). Identifying and exploiting trait-relevant tissues with multiple functional annotations in genome-wide association studies. *PLoS Genetics*, 14(1).
- Hivert, V., Sidorenko, J., Rohart, F., Goddard, M. E., Yang, J., Wray, N. R., Yengo, L., and Visscher, P. M. (2021). Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *The American Journal of Human Genetics*, 108(5):786–798.
- Inouye, M., Abraham, G., Nelson, C. P., Wood, A. M., Sweeting, M. J., Dudbridge, F., Lai, F. Y., Kaptoge, S., Brozynska, M., Wang, T., et al. (2018). Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *Journal of the American College of Cardiology*, 72(16):1883–1893.
- J Rowe, S. and Tenesa, A. (2012). Human complex trait genetics: lifting the lid of the genomics toolbox-from pathways to prediction. *Current Genomics*, 13(3):213–224.
- Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., and Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. *Nature genetics*, 51(12):1749–1755.
- Johnson, M. P., Brennecke, S. P., East, C. E., Göring, H. H., Kent Jr, J. W., Dyer, T. D., Said, J. M., Roten, L. T., Iversen, A.-C., Abraham, L. J., et al. (2012). Genome-wide association scan identifies a risk locus for preeclampsia on 2q14, near the inhibin, beta b gene. *PloS one*, 7(3):e33666.
- Kundaje, A. and Meuleman, t. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.
- Kwak, S. H., Kim, S.-H., Cho, Y. M., Go, M. J., Cho, Y. S., Choi, S. H., Moon, M. K., Jung, H. S., Shin, H. D., Kang, H. M., et al. (2012). A genome-wide association study of gestational diabetes mellitus in korean women. *Diabetes*, 61(2):531–541.
- Laisk, T., Soares, A. L. G., Ferreira, T., Painter, J. N., Censin, J. C., Laber, S., Bacelis, J., Chen, C.-Y., Lepamets, M., Lin, K., et al. (2020). The genetic architecture of sporadic and multiple consecutive miscarriage. *Nature communications*, 11(1):1–12.
- Lambert, S. A., Abraham, G., and Inouye, M. (2019). Towards clinical utility of polygenic risk scores. *Human molecular genetics*, 28(R2):R133–R142.
- Lamri, A., Mao, S., Desai, D., Gupta, M., Paré, G., and Anand, S. S. (2020). Fine-tuning of genome-wide polygenic risk scores and prediction of gestational diabetes in south asian women. *Scientific reports*, 10(1):1–9.

- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome.
- Landrum, M. J., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., et al. (2020). Clinvar: improvements to accessing data. *Nucleic acids research*, 48(D1):D835–D844.
- Lango Allen, H. and *et al.*, E. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838.
- Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163.
- Lee, S. H., Wray, N. R., Goddard, M. E., and Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 88(3):294–305.
- Liu, X., Helenius, D., Skotte, L., Beaumont, R. N., Wielscher, M., Geller, F., Juodakis, J., Mahajan, A., Bradfield, J. P., Lin, F. T., et al. (2019). Variants in the fetal genome near pro-inflammatory cytokine genes on 2q13 associate with gestational duration. *Nature communications*, 10(1):1–13.
- Locke, A. E. and *et al.*, K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206.
- Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K., Pollack, S. J., de Candia, T. R., Lee, S. H., Wray, N. R., Kendler, K. S., et al. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature genetics*, 47(12):1385.
- Loos, R. J. (2020). 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications*, 11(1):1–3.
- Ma, Y. and Zhou, X. (2021). Genetic prediction of complex traits with polygenic scores: a statistical review. *Trends in Genetics*, 37(11):995–1011.
- 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.
- Maier, R. M., Zhu, Z., Lee, S. H., Trzaskowski, M., Ruderfer, D. M., Stahl, E. A., Ripke, S., Wray, N. R., Yang, J., Visscher, P. M., and Robinson, M. R. (2018). Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nature Communications*, 9(1):989.
- Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873.
- Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., and Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, 27(2):e1608.
- McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews genetics*, 9(5):356.
- McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*, 48(10):1279.
- McGinnis, R., Steinhorsdottir, V., Williams, N. O., Thorleifsson, G., Shooter, S., Hjartardottir, S., Bumpstead, S., Stefansdottir, L., Hildyard, L., Sigurdsson, J. K., et al. (2017). Variants in the fetal genome near *flt1* are associated with risk of preeclampsia. *Nature genetics*, 49(8):1255–1260.

- Meuwissen, T. H., Hayes, B. J., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.
- Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2015a). Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. *PLOS Genetics*, 11(4):e1004969.
- Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2015b). Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLOS Genetics*, 11(4):1–22.
- Ojavee, S. E., Kutalik, Z., and Robinson, M. R. (2022). Liability-scale heritability estimation for biobank studies of low prevalence disease. *medRxiv*.
- Ongen, H., Brown, A. A., Delaneau, O., Panousis, N. I., Nica, A. C., Dermitzakis, E. T., and Dermitzakis, E. T. (2017). Estimating the causal tissues for complex traits and diseases. *Nature Genetics*, 49(12):1676–1683.
- Orliac, E. J., Banos, D. T., Ojavee, S. E., Kristi, L., Reedik, M., Visscher, P. M., Robinson, M. R., et al. (2021). Maximizing gwas discovery and genomic prediction accuracy in biobank data. *bioRxiv*.
- Pazokitoroudi, A., Wu, Y., Burch, K. S., Hou, K., Pasaniuc, B., and Sankararaman, S. (2019). Scalable multi-component linear mixed models with application to SNP heritability estimation. *bioRxiv*, page 522003.
- Pazokitoroudi, A., Wu, Y., Burch, K. S., Hou, K., Zhou, A., Pasaniuc, B., and Sankararaman, S. (2020). Efficient variance components analysis across millions of genomes. *Nature communications*, 11(1):1–10.
- Pers, T. H., Karjalainen, J. M., Chan, Y., Westra, H.-J., Wood, A. R., Yang, J., Lui, J. C., Vedantam, S., Gustafsson, S., Esko, T., Frayling, T., Speliotes, E. K., Boehnke, M., Raychaudhuri, S., Fehrmann, R. S. N., Hirschhorn, J. N., Franke, L., and Franke, L. (2015). Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications*, 6(1):5890.
- Pervjakova, N., Moen, G.-H., Borges, M.-C., Ferreira, T., Cook, J. P., Allard, C., Beaumont, R. N., Canouil, M., Hatem, G., Heiskala, A., et al. (2022). Multi-ancestry genome-wide association study of gestational diabetes mellitus highlights genetic links with type 2 diabetes. *Human Molecular Genetics*.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575.
- Rappoport, N., Toung, J., Hadley, D., Wong, R. J., Fujioka, K., Reuter, J., Abbott, C. W., Oh, S., Hu, D., Eng, C., et al. (2018). A genome-wide association study identifies only two ancestry specific variants associated with spontaneous preterm birth. *Scientific reports*, 8(1):1–11.
- Rasmussen, M., Reddy, M., Nolan, R., Camunas-Soler, J., Khodursky, A., Scheller, N. M., Cantonwine, D. E., Engelbrechtsen, L., Mi, J. D., Dutta, A., et al. (2022). Rna profiles reveal signatures of future health and disease in pregnancy. *Nature*, pages 1–6.
- Relling, M. and Klein, T. (2011). Cplic: clinical pharmacogenetics implementation consortium of the pharmacogenomics research network. *Clinical Pharmacology & Therapeutics*, 89(3):464–467.
- Ripke, S. and Neale, t. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427.
- Ritchie, S. (2014). liftOverPlink. <https://github.com/sritchie73/liftOverPlink.git>, Last accessed on 2022-03-22.











- Rubinacci, S., Delaneau, O., and Marchini, J. (2020). Genotype imputation using the positional burrows wheeler transform. *PLoS genetics*, 16(11):e1009049.
- Sawyer, S. L., Mukherjee, N., Pakstis, A. J., Feuk, L., Kidd, J. R., Brookes, A. J., and Kidd, K. K. (2005). Linkage disequilibrium patterns vary substantially among populations. *European journal of human genetics*, 13(5):677–686.
- Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., et al. (2017). Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome research*, 27(5):849–864.
- Segrè, A. V., Groop, L., Mootha, V. K., Daly, M. J., Altshuler, D., Daly, M. J., and Altshuler, D. (2010). Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits. *PLoS Genetics*, 6(8):e1001058.
- Slatkin, M. (2008). Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485.
- Soma-Pillay, P. and Catherine, N. (2016). P, tolppanen h, mebazaa a, tolppanen h, mebazaa a. *Physiological changes in pregnancy. Cardiovasc J Afr*, 27(2):89–94.
- Speed, D. and Balding, D. J. (2019). SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nature Genetics*, 51(2):277–284.
- Steinthorsdottir, V., McGinnis, R., Williams, N. O., Stefansdottir, L., Thorleifsson, G., Shooter, S., Fadista, J., Sigurdsson, J. K., Auro, K. M., Berezina, G., et al. (2020). Genetic predisposition to hypertension is associated with preeclampsia in european and central asian women. *Nature communications*, 11(1):1–14.
- Stovner, E. B. and Cole, B. S. (2019). snpflip. <https://github.com/biocore-ntnu/snpflip>, Last accessed on 2022-03-22.
- Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779.
- Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., et al. (2021). Sequencing of 53,831 diverse genomes from the nhlni topmed program. *Nature*, 590(7845):290–299.
- Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484.
- Tiensuu, H., Haapalainen, A. M., Karjalainen, M. K., Pasanen, A., Huusko, J. M., Marttila, R., Ojaniemi, M., Muglia, L. J., Hallman, M., and Rämetsä, M. (2019). Risk of spontaneous preterm birth and fetal growth associates with fetal slit2. *PLoS genetics*, 15(6):e1008107.
- Timpson, N. J., Greenwood, C. M., Soranzo, N., Lawson, D. J., and Richards, J. B. (2018). Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics*, 19(2):110–124.
- Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):1–21.
- Vilhjálmsdóttir, B. and et al., Y. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics*, 97(4):576–592.
- Visscher, P. M. (2008). Sizing up human height variation. *Nature genetics*, 40(5):489–490.
- Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, 101(1):5–22.

- Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., et al. (2018). Unraveling the polygenic architecture of complex traits using blood eqtl meta-analysis. *bioRxiv*, page 447367.
- Warrington, N. M. and Beaumont, t. (2019). Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. *Nature Genetics*, 51(5):804–814.
- Whirl-Carrillo, M., Huddart, R., Gong, L., Sangkuhl, K., Thorn, C. F., Whaley, R., and Klein, T. E. (2021). An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 110(3):563–572.
- Workalemahu, T., Enquobahrie, D. A., Gelaye, B., Sanchez, S. E., Garcia, P. J., Tekola-Ayele, F., Hajat, A., Thornton, T. A., Ananth, C. V., and Williams, M. A. (2018). Genetic variations and risk of placental abruption: A genome-wide association study and meta-analysis of genome-wide association studies. *Placenta*, 66:8–16.
- Wray, N. R., Ripke, S., and *et al.*, M. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, 50(5):668–681.
- Wu, N.-N., Zhao, D., Ma, W., Lang, J.-N., Liu, S.-M., Fu, Y., Wang, X., Wang, Z.-W., and Li, Q. (2021). A genome-wide association study of gestational diabetes mellitus in chinese women. *The Journal of Maternal-Fetal & Neonatal Medicine*, 34(10):1557–1564.
- Xu, Z., Wu, C., Wei, P., and Pan, W. (2017). A powerful framework for integrating eqtl and gwas summary data. *Genetics*, 207(3):893–902.
- Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011a). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, 88(1):76–82.
- Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M. G., Hill, W. G., Landi, M. T., Alonso, A., Lettre, G., Lin, P., Ling, H., Lowe, W., Mathias, R. A., Melbye, M., Pugh, E., Cornelis, M. C., Weir, B. S., Goddard, M. E., and Visscher, P. M. (2011b). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature genetics*, 43(6):519–25.
- Yong, S. Y., Raben, T. G., Lello, L., and Hsu, S. D. (2020). Genetic architecture of complex traits and disease risk predictors. *Scientific reports*, 10(1):1–14.
- Zhang, G., Feenstra, B., Bacelis, J., Liu, X., Muglia, L. M., Juodakis, J., Miller, D. E., Litterman, N., Jiang, P.-P., Russell, L., et al. (2017). Genetic associations with gestational duration and spontaneous preterm birth. *New England Journal of Medicine*, 377(12):1156–1167.
- Zhang, G., Srivastava, A., Bacelis, J., Juodakis, J., Jacobsson, B., and Muglia, L. J. (2018). Genetic studies of gestational duration and preterm birth. *Best practice & research Clinical obstetrics & gynaecology*, 52:33–47.
- Zhang, Q., Privé, F., Vilhjálmsson, B., and Speed, D. (2021). Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nature communications*, 12(1):1–9.
- Zwick, M. E., Cutler, D. J., and Chakravarti, A. (2000). Patterns of genetic variation in mendelian and complex traits. *Annual review of genomics and human genetics*, 1(1):387–407.

Appendix A - Probabilistic inference of the genetic architecture underlying functional enrichment of complex traits

This article (Patxot *et al.* 2021) is presented in Chapter 1. Supplementary Data Tables can be downloaded from <https://www.nature.com/articles/s41467-021-27258-9>.

Probabilistic inference of the genetic architecture underlying functional enrichment of complex traits

Marion Patxot ^{1,11}, Daniel Trejo Banos ^{1,11}, Athanasios Kousathanas ^{1,11}, Etienne J. Orlicac², Sven E. Ojavee ¹, Gerhard Moser ³, Alexander Holloway¹, Julia Sidorenko ⁴, Zoltan Kutalik ^{1,5,6}, Reedik Mägi⁷, Peter M. Visscher ⁴, Lars Rönnegård ^{8,9} & Matthew R. Robinson ¹⁰✉

We develop a Bayesian model (BayesRR-RC) that provides robust SNP-heritability estimation, an alternative to marker discovery, and accurate genomic prediction, taking 22 seconds per iteration to estimate 8.4 million SNP-effects and 78 SNP-heritability parameters in the UK Biobank. We find that only $\leq 10\%$ of the genetic variation captured for height, body mass index, cardiovascular disease, and type 2 diabetes is attributable to proximal regulatory regions within 10kb upstream of genes, while 12–25% is attributed to coding regions, 32–44% to introns, and 22–28% to distal 10–500kb upstream regions. Up to 24% of all cis and coding regions of each chromosome are associated with each trait, with over 3,100 independent exonic and intronic regions and over 5,400 independent regulatory regions having $\geq 95\%$ probability of contributing $\geq 0.001\%$ to the genetic variance of these four traits. Our open-source software (GMRM) provides a scalable alternative to current approaches for biobank data.

¹Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. ²Scientific Computing and Research Support Unit, University of Lausanne, Lausanne, Switzerland. ³Australian Agricultural Company Limited, Brisbane, QLD, Australia. ⁴Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD, Australia. ⁵University Center for Primary Care and Public Health, Lausanne, Switzerland. ⁶Swiss Institute of Bioinformatics, Lausanne, Switzerland. ⁷Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia. ⁸School of Technology and Business Studies, Dalarna University, Falun, Sweden. ⁹Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden. ¹⁰Institute of Science and Technology Austria, Klosterneuburg, Austria. ¹¹These authors contributed equally: Marion Patxot, Daniel Trejo Banos, Athanasios Kousathanas. ✉email: matthew.robinson@ist.ac.at

As whole-genomes are collected for hundreds of thousands of individuals, we require regression methods that are not only computationally efficient, but which also provide improved inference. Rather than relying on subsets of the SNPs, methods should fully utilise the data, exploiting computational power to facilitate discovery of additional genomic regions, to improve understanding of the genomic architecture of common disease, and to provide more informative genomic prediction.

For example, when estimating the proportion of phenotypic variance attributable to different categories of genetic markers (the SNP-heritability, h_{SNP}^2 of a genomic region), recent studies^{1–4} highlight the importance of accounting for minor allele frequency (MAF) and LD structure of the genomic data. Generally, assessment of the relative contribution of different genomic regions is currently made assuming that markers within a category all contribute to the variance, with enrichment defined as the estimated share of the variance explained divided by its expected share^{5,6}. However ideally, the estimated distribution of marker effects for each category would be directly obtained, accounting for MAF and LD structure and allowing for some of the marker effects to be zero, as this would yield a better understanding of the polygenicity of genomic effects across different genomic annotation groups.

Furthermore, statistical inference usually follows a multi-step approach. Current mixed-linear association models such as those implemented in the software fastGWA⁷, BoltLMM⁸ and REGENIE⁹, use a two-step approach, first estimating the variance contributed by the SNP markers without the use of MAF-LD-annotation information, and then estimating the marker effect sizes one-by-one as fixed effects in a second step^{7,8,10}. Following this initial mixed-model association step, statistical inference (variance components, fine mapping and risk prediction) is then typically conducted on the summary statistics generated. The advantage of a multi-step approach is that large sample size can be easily obtained through meta-analyses, combining summary statistics from different studies and avoiding the need for individual-level data sharing. However, as large-scale biobank data is increasingly available, methods that provide joint estimates of the marker effects in a single step by estimating the effect sizes as random under flexible prior formulations may become beneficial as they: (i) can account for differences in the variance contributed across MAF, LD or annotation groups providing unbiased MAF-LD annotation-specific genetic effect size estimates and h_{SNP}^2 of different annotations, allowing for a contrasting of the genetic architectures of complex traits; (ii) give the probability that each marker, genomic region, annotation, gene-coding region, or SNP is associated with a phenotype, alongside the proportion of phenotypic variation contributed by each, yielding test statistics that describe the *gene* architecture of complex traits and the uncertainty over the estimates; and (iii) provide improved genomic prediction, whilst providing a posterior predictive distribution for each individual.

Here, we outline the fastest Bayesian penalised regression model to date, with a hybrid-parallel algorithm for analysing large-scale genomic biobank using a single command-line tool implemented in our grouped mixture regressions model (GMRM) software. We validate our approach in large-scale simulation study and provide an empirical example using four traits measured in both the UK Biobank and Estonian Biobank data.

Results

A Bayesian model for large-scale genomic data. We derive a model that we call BayesRR-RC in Supplementary Note 1 and the “Methods” section, which is based on grouped effects with

mixture priors, improving on the formulations of refs. 11–13. Like these former methods, we consider a spike probability at zero (Dirac delta function), and a scale mixture of Gaussian distributions as a slab probability density. Unlike these models, we have genetic markers grouped into MAF-LD-annotation specific sets, with independent hyper-parameters for the phenotypic variance attributable to each group, so that the mixture proportions, the variance explained by the SNP markers, and the mixture constants are all unique and independent across SNP marker groups. This enables estimation of the phenotypic variance attributable to the group-specific effects, and differences in the underlying distribution of the β_ϕ effects among MAF-LD-annotation groups, with different degrees of sparsity. Assuming N individuals and p genetic markers, our model of an observed phenotype vector \mathbf{y} is:

$$\mathbf{y} = \mathbf{1}\mu + \sum_{\phi=1}^{\Phi} \mathbf{X}_\phi \beta_\phi + \boldsymbol{\epsilon}, \quad (1)$$

where there is a single intercept term $\mathbf{1}\mu$ and a single error term, a vector ($N \times 1$) of residuals $\boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} | \sigma_\epsilon^2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_\epsilon^2)$. An N by p matrix of single nucleotide polymorphism (SNP) genetic markers, centred and scaled to unit variance, which we denote as \mathbf{X}_ϕ . The effects are allocated into groups $(1, \dots, \Phi)$. Each group has a set of model parameters $\Theta_\phi = \{\beta_\phi, \pi_\phi, \sigma_{G\phi}^2\}$, with β_ϕ as a $p_\phi \times 1$ vector of partial regression coefficients, where β_{ϕ_j} is the effect of a 1 SD change in the j th covariate within the ϕ th group. The spike and slab prior, contains what is called a Dirac spike^{14,15} for β_ϕ , which induces sparsity in the model through a Dirac-delta at zero, excluding variables from the model by setting their coefficients to zero. A finite scale mixture of normal distributions centred at zero constitute the slab component. The slab shrinks the non-zero coefficients towards zero according to the slab’s width, and by having a scale mixture of Gaussians, the distribution has heavier tails and can accommodate big and small effects¹⁶. Therefore, each β_{ϕ_j} is distributed according to:

$$\beta_{\phi_j} \sim \pi_{0\phi} \delta_0 + \pi_{1\phi} \mathcal{N}(0, \sigma_{1\phi}^2) + \pi_{2\phi} \mathcal{N}(0, \sigma_{2\phi}^2) + \dots + \pi_{L_\phi\phi} \mathcal{N}(0, \sigma_{L_\phi\phi}^2), \quad (2)$$

where for each SNP marker group $\{\pi_{0\phi}, \pi_{1\phi}, \dots, \pi_{L_\phi\phi}\}$ are the mixture proportions and $\{\sigma_{1\phi}^2, \sigma_{2\phi}^2, \dots, \sigma_{L_\phi\phi}^2\}$ are the mixture-specific variances proportional to

$$\begin{bmatrix} \sigma_{1\phi}^2 \\ \vdots \\ \sigma_{L_\phi\phi}^2 \end{bmatrix} = \sigma_{G\phi}^2 \begin{bmatrix} C_{1\phi} \\ \vdots \\ C_{L_\phi\phi} \end{bmatrix},$$

with $\sigma_{G\phi}^2$ the phenotypic variance associated with the SNPs in group ϕ , which, like all the other parameters, is estimated directly from the data. Here, we use 78 MAF-LD-annotation SNP marker groups. SNPs are partitioned into seven location annotations preferentially to coding (exonic) regions first, then to intronic regions, then to 1 kb upstream regions, then to 1–10 kb regions, then to 10–500 kb regions, then to 500–1 Mb regions. Remaining SNPs were grouped in a category labelled “others” and also included in the model so that variance is partitioned relative to these also. Thus, we assigned SNPs to their closest upstream region, for example if a SNP is 1 kb upstream of gene X, but also 10–500 kb upstream of gene Y and 5 kb downstream for gene Z, then it was assigned to be a 1 kb region SNP. This ensures that SNPs 10–500 kb and 500 kb–1 Mb upstream are distal from any known gene. We further partition upstream regions to experimentally validated promoters, transcription factor binding sites (tfbs) and enhancers (enh) using the HACER, snp2tfbs databases

(see “Code availability” section). All SNP markers assigned to 1 kb regions map to promoters; 1–10 kb SNPs, 10–500 kb SNPs, 500 kb–1 Mb SNPs are then split into enh, tfbs and others (unmapped SNPs) extending the model to 13 annotation groups (Supplementary Data 1). Within each of these annotations, we have three minor allele frequency groups ($MAF \leq 0.01$, $0.01 < MAF \leq 0.05$, and $MAF > 0.05$), and then each MAF group is further split into two based on median LD score. This gives 78 non-overlapping groups for which our BayesRR-RC model jointly estimates the phenotypic variation attributable to, and the SNP marker effects within, each group. For each of the 78 groups, SNPs were modelled using five mixture groups with variance equal to the phenotypic variance attributable to the group multiplied by constants (0, 0.0001, 0.001, 0.01, 0.1).

One of the major limitations preventing the application of Bayesian approaches to large-scale genomic data is the view that the computation of a posterior distribution is too expensive. In Supplementary Note 2, we derive a Bulk Synchronous hybrid-parallel (BSP) Gibbs sampling scheme for large-scale genomic data that allows both the data and the compute tasks to be split within and across compute nodes in a series of message-passing interface (MPI) tasks. We extend previous sparse residual updating schemes by deriving sampling steps to utilise whole genome sequence or SNP genetic data stored in mixed binary/sparse-index representation (see Supplementary Note 2), reducing computational complexity of a single Gibbs step from $\mathcal{O}(n)$ to $\mathcal{O}(n_z)$, with n_z the number of non-zero genotypes, as SNP-phenotype covariance estimation (dot product calculation) is conducted as a series of look-up tables. We provide publicly available open source software (GMRM) that requires as little as 22 s per MCMC sample to estimate 78 group-specific h^2_{SNP} parameters, and the inclusion probabilities and effect sizes of 8,433,421 markers in 382,466 individuals on standard Intel Xeon CPU processors (see “Code availability” section, Supplementary Note 2).

Simulation study. We first compare the model performance of BayesRR-RC to existing approaches across 18 different genetic architectures. We randomly selected 40,000 unrelated UK Biobank individuals and used 596,741 imputed SNP markers from chromosomes 19 to 22. We randomly selected either 1000, 10,000 or 100,000 LD independent ($LD R^2 < 0.1$) causal SNP markers. For each SNP marker set, we then simulated effect sizes from a normal distribution with zero mean and variance of 0.1, 0.3 or 0.6 divided by the number of causal variants and $\alpha N(0, [p(1-p)]^{-0.25})$, with p the allele frequency (see “Methods” section). This simulates stronger effect sizes for rare variants in line with recent empirical estimates and we simulated ten replicate phenotypes for each of the nine different genetic architectures. We then additionally repeat each simulation, sampling the SNP marker effects this time from 13 different distributions, one for each of 13 different genomic annotation groups with different proportions of h^2_{SNP} to create nine further different genetic architectures. We compare our BayesRR-RC model to the following statistical models: (i) a restricted maximum likelihood (REML) model implemented in the software BoltREML¹⁷ with the same 78 MAF-LD-annotation groups enabling a direct comparison, (ii) a Haseman–Elston (HE) regression using the same 78 group model implemented in the software RHEmc¹⁸, (iii) summary statistic linkage disequilibrium score regression (LDSC)¹⁹, with LD scores calculated using the same data, and the same 78 non-overlapping annotations in a 78 component LDSC annotation model, and (iv) summary statistic SumHer⁶ (LDAK) with the same 78 non-overlapping annotations.

We find that BayesRR-RC estimates the phenotypic variation attributable to different genomic annotation groups comparable

with the BoltREML model, with similar correlation of the estimated and simulated values within each simulation replicate (Fig. 1a). In comparison, RHEmc, which also uses individual-level data, yields estimates with lower correlation with the simulated value, but higher than both summary statistic approaches implemented in LDSC and SumHer (Fig. 1a). We calculate estimates of enrichment, defined as the proportion of h^2_{SNP} attributable to the annotation divided by the proportion of SNPs mapping to the annotation (for BayesRR-RC, because there is sparsity in the SNP effects, we define enrichment as the proportion of SNPs in the model that map to the annotation, see “Methods” section) and we compare these to the true simulated value. Compared to other approaches, we find that BayesRR-RC gives a lower probability of false enrichment, calculated as the proportion of times within a simulation replicate that an annotation group was incorrectly assigned as having enrichment greater than 2 (Fig. 1b). Thus, BayesRR-RC provides accurate partitioning of genomic enrichment across the genome.

In Supplementary Note 3, we propose a posterior probability window variance (PPWV) approach²⁰, which provides a probabilistic determination of association of a given LD block, genomic window, gene, or upstream region, relative to the amount of phenotypic variation attributable to that window. Our PPWV approach determines the posterior inclusion probability that each region and each gene contributes at least 0.001% to the h^2_{SNP} , with theory outlined in Supplementary Note 3 suggesting well controlled FDR. We determine the ability of our PPWV approach to correctly localise an association to LD blocks (defined as groups of markers with $LD R^2 \geq 0.1$) that contain causal variants, and compare this to using LD to clump mixed-linear model association estimates obtained using the BoltLMM software (Fig. 2a). We find that a PPWV approach identifies associated LD blocks with higher probability as compared to clumped MLMA associations, for all genetic architectures, with the exception of simulated phenotypes with enrichment and low polygenicity, where the small numbers of relatively large effect size regions are better identified with a single-marker regression approach (Fig. 2a). Thus, BayesRR-RC provides an alternative to standard genome-wide association studies to localise SNP-phenotype associations at the regional level, especially for traits with high polygenicity.

We then also compare the prediction accuracy obtained in an independent sample when creating genomic predictors using (i) effect sizes estimated by BayesRR-RC, (ii) fixed-effect SNP effect sizes estimated in the MLMA approach implemented in bolt, and (iii) effect size estimates obtained from four different genomic prediction models proposed in a recent paper²¹, implemented in the LDAK software, which are suggested to outperform all other current approaches. In comparison to the best LDAK predictor, we find that BayesRR-RC obtains similar or improved prediction accuracy across all genetic architectures, with greater prediction accuracy gains observed under genetic architectures where the SNP effect distributions differed across genomic annotations (Fig. 2b). We find that given sufficient power, BayesRR-RC can obtain or even exceed the theoretical expectation of prediction accuracy under ridge regression assumptions (Fig. 2b, see “Methods” section).

We then conduct a number of follow-up simulation studies. Recent work has highlighted differences in statistical model performance depending upon the relationship of SNP marker effect size, LD and MAF^{1,3,4}. We explore the performance of our model in theory, with highly correlated genetic markers in Supplementary Note 4. We also conducted another large-scale, but well-powered, simulation study to explore the model performance of BayesRR-RC as compare to existing approaches

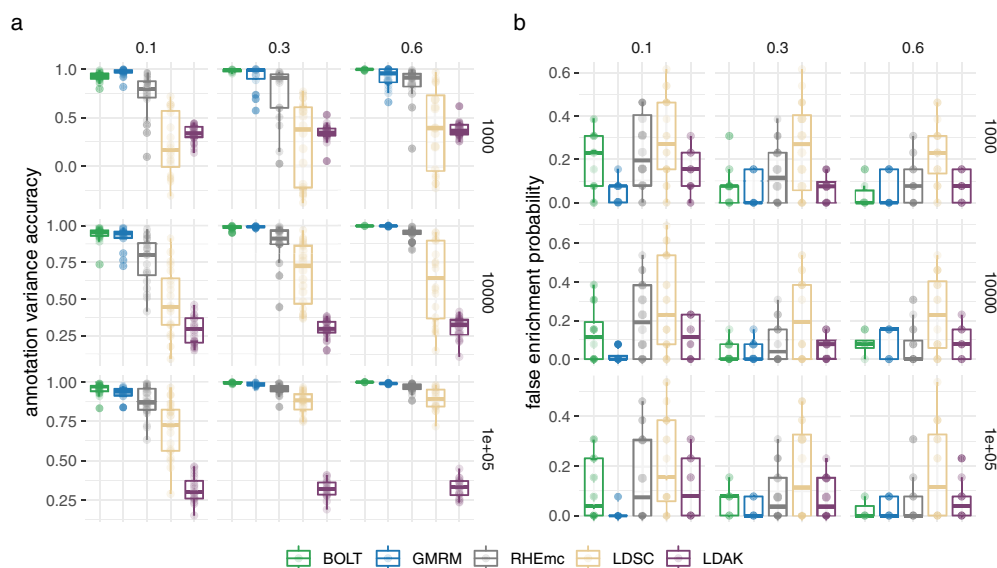


Fig. 1 Simulation study for the performance of our BayesRR-RC model implemented in the GMRM software against existing approaches for variance component and genomic annotation enrichment estimation. **a** Correlation of the simulated and estimated SNP heritability across 13 genomic annotation groups within each of 20 replicates for five different statistical models: a mixture of regression model with multiple group-specific variance components described in this work (GMRM), Haseman-Elston regression with annotation-specific relationship matrices implemented in the RHEmc software (RHEmc), a multiple group-specific variance component REML model implemented in the software bolt (BOLT), and two annotation summary statistic models implemented in the software LDSC and LDAK. The column facets give the simulated heritability and rows give the number of causal variants. **b** Probability of falsely assigning one of the 13 genomic annotation groups as explaining 2 times greater proportion of variance given the proportion of SNPs mapping to the annotation. The column facets give the simulated heritability and rows give the number of causal variants. Boxplots give the median with 25th and 75th percentile and 95% credible intervals for $n = 20$ simulation replicates in both panels.

across a wide range of 20 different effect size, LD, and MAF relationships as described in Supplementary Table 1. For the estimation of h^2_{SNP} and the proportion of h^2_{SNP} attributable to different annotation groups, we find that all statistical models other than BayesRR-RC are sensitive to the underlying generative genetic model, with no other approach providing consistent estimates across the 20 generative genetic models (Supplementary Fig. 1a). As in the previous simulation, BayesRR-RC estimates the variance attributable to different genomic regions on the correct scale, with higher correlation as compared with other approaches (Supplementary Fig. 1b), and this results in the estimated average effect size for each annotation group having high correlation with the simulated value (Supplementary Fig. 1c). Again, summary statistic approaches performed poorly for both variance component estimation (Supplementary Fig. 1b) and quantification of enrichment as compared to individual-level methods, often even incorrectly selecting the group of highest average effect size (Supplementary Fig. 1c).

We confirmed our genomic prediction results, finding that BayesRR-RC outperforms all methods implemented in the LDAK software across all generative models, with BayesRR-RC very marginally outperforming a single variance component BayesR model in the enrichment simulations of each of the 20 generative genetic models (Supplementary Fig. 2).

We further explored the ability of our PPWV approach to localise SNP-phenotype associations in the 20 generative models, by comparing the z -scores of the marker effect estimates from their true simulated value across the minor allele frequency spectrum (Supplementary Fig. 3) and the area under the precision-recall curve (AUPRC, Supplementary Fig. 4) for BayesRR-RC and a series of MLMA methods. We find that the

z -scores of the BayesRR-RC estimates are generally stable across generative genetic models and that the MLMA estimates have higher estimation error, especially when the causal variant is rare, or in high-LD with many other SNPs (Supplementary Fig. 3). We also find that our PPWV approach outperforms MLMA methods in their precision-recall curves across the range of genetic architectures (Supplementary Fig. 4). We confirmed that population stratification and relatedness are well-controlled for using a PPWV approach, as compared to an MLMA model with the leading PCs of the genomic data included (Supplementary Fig. 5). We compared the ability of our approach to identify candidate SNPs and to provide a probabilistic assessment of the most likely associated set of SNP markers. Finally, we show that our PPWV approach is analogous to the approach suggested in a recent paper (SuSiE²²) of selecting credible sets of markers with high probability of association, finding that BayesRR-RC has higher power to localise associations to sets of SNP markers (Supplementary Fig. 6). The advantage of BayesRR-RC is also that assessment of associated regions is done genome-wide, with estimates obtained through simple summary of the posterior distribution instead of running numerous statistical models at different genomic regions. Taken together, these simulation results indicate that BayesRR-RC provides accurate estimates of the underlying effect size distribution for different genomic groups, yielding improved genomic prediction, across a wide range of different underlying generative genetic models.

The genetic architecture of four complex traits in the UK Biobank. We apply BayesRR-RC to cardiovascular disease outcomes (CAD), type-2 diabetes (T2D), body mass index (BMI) and height measured for 382,466 unrelated individuals from the

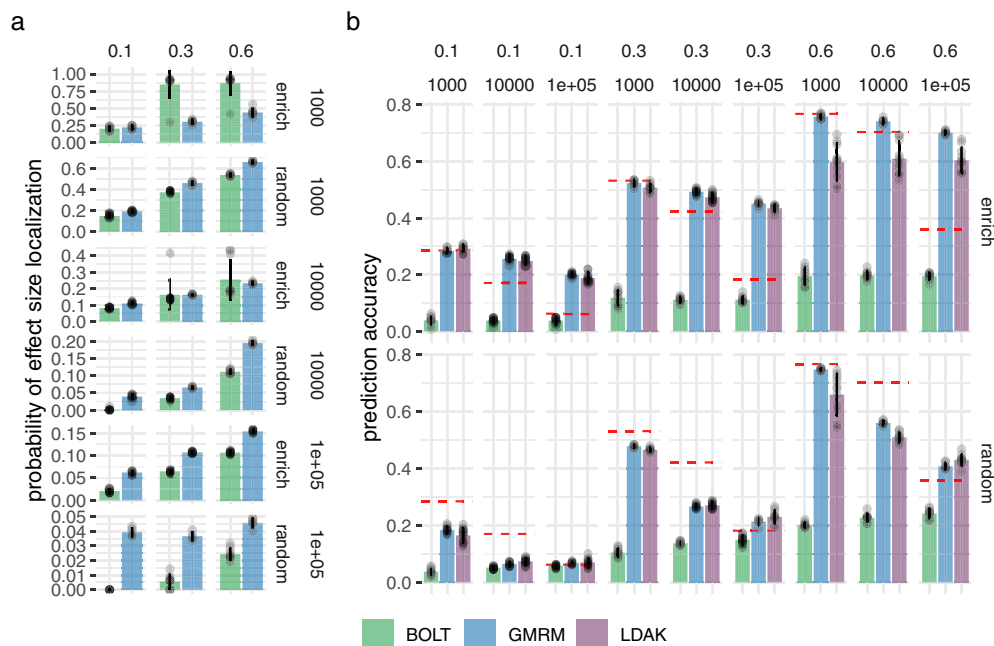


Fig. 2 Simulation study for the performance of our BayesRR-RC model implemented in the GMRM software against existing approaches for localisation of associations and genomic prediction. **a** Probability of detecting genomic regions containing simulated causal variants by a Bayesian regional fine-mapping approach (GMRM: blue) versus standard mixed linear model association (MLMA) testing (BOLT: green). The column facets give the simulated heritability and rows give the number of causal variants and whether the effect sizes differed across genomic annotation groups (enrich) or were randomly assigned (random). **b** Correlation of a genomic predictor and a phenotype in an independent sample when the genomic predictor is created from GMRM effects sizes (blue), MLMA effect sizes using BOLT (green), and the optimal effect sizes obtained from individual-level and summary statistic models implemented in the Mega-PRS LDAK approach (purple). The column facets give the simulated heritability and the number of causal variants. The row facets give whether the effect sizes differed across genomic annotation groups (enrich) or were randomly assigned (random). The red lines give the expected prediction accuracy based on ridge regression theory. Error bars show the SD in both panels.

UK Biobank data genotyped at 8,433,421 imputed SNP markers. These markers were selected as they overlap with the Estonian Genome Centre data (see “Methods” section) and have minor allele frequency >0.0002 . We adjust each phenotype for age, sex, year of birth, genotype batch effects, UK Biobank assessment centre, and the leading 20 principal components of the SNP data. We conducted a series of convergence diagnostic analyses of the posterior distributions to ensure we obtained estimates from a converged set of four Gibbs chains, each run for 6000 iterations with a thin of five for each trait (Supplementary Figs. 7–10).

We find that 32–44% of the h^2_{SNP} is attributable to intronic regions, 12–25% is attributable to exonic regions, 22–28% is attributable to markers 10–500 kb upstream of genes, with proximal (within 10 kb) promoters, enhancers and transcription factor binding sites cumulatively contributing $<10\%$ (Fig. 3b and Supplementary Fig. 11, with estimates summed across MAF and LD groups Table 1, and full results in Supplementary Data 2). The large contribution of exonic and intronic annotations to variation is in-line with the fact that these annotations account for $\sim 40\%$ of the total genome length. All four traits show the same pattern of group-specific variation, with the exception of height, where the proportion of h^2_{SNP} attributable to exons is almost twice as large as the other phenotypes (Fig. 3b; Table 1 and Supplementary Fig. 11 and Supplementary Data 2). For all annotation groups in exons, introns, and within 500 kb of genes across all traits, $\geq 60\%$ of the h^2_{SNP} attributable to these groups is contributed by many thousands of common variants, each of small effect (Fig. 3b and Supplementary Figs. 11 and 12).

Our estimates compare similarly to those obtained by RHEmc and SumHer, but differ to those obtained by LDSC (Table 1 and Supplementary Data 3, 4, and 5 for full results). In addition to providing variance component estimates, our model facilitates assessment of differences in the underlying effect size distribution across annotation groups. For each group, we modelled the SNP effects as coming from a series of five Gaussian mixtures, and we find that at least 45% of the h^2_{SNP} attributable to both introns and 500 kb upstream regions is underlain by many thousands of SNPs that on average each contribute 0.001% (estimates summed across MAF and LD groups in Fig. 3b and Supplementary Figs. 11 and 12). In contrast, the variance is spread more evenly across the mixtures for the other groups, implying that 10–500 kb upstream regions and introns are more polygenic than other groups. This is especially so for BMI where 35% of the h^2_{SNP} is attributable to many thousands of intronic variants (Fig. 3 and Supplementary Fig. 12). Therefore, we find that the polygenicity of the genetic effects varies across different genomic regions, with remarkably consistent patterns across traits in the partitioning of h^2_{SNP} across the genome.

Across traits, posterior mean effect sizes scale to their differences in h^2_{SNP} , and we find that exonic and intronic region effect sizes were higher than the rest of the genome, across all mixture groups, followed by 10–500 kb upstream regions (Fig. 3c). We find little evidence that SNPs located in proximal promoters, enhancers, and transcription factor binding sites within 10 kb of genes showed average effect sizes that were higher than SNPs located 1 MB away from genes, or those that were not mapped to

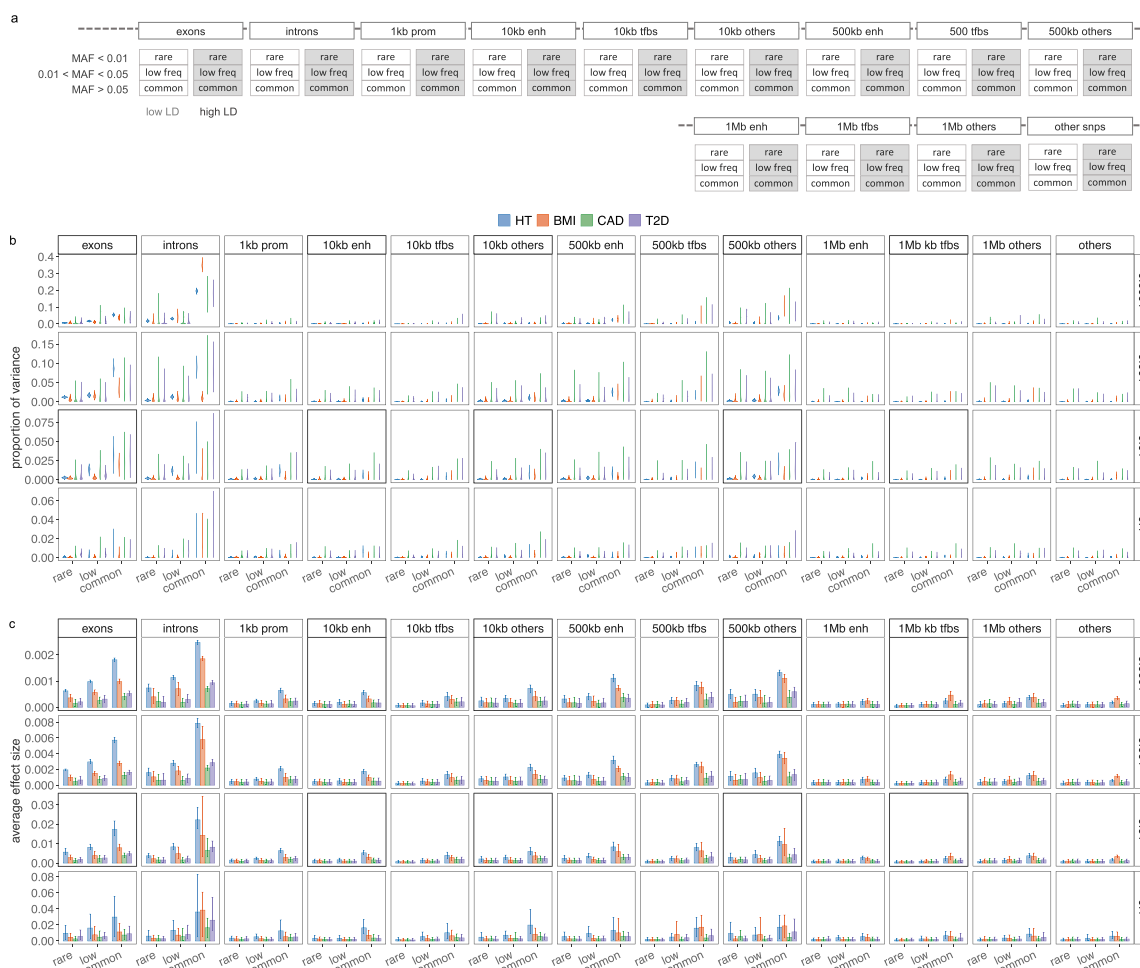


Fig. 3 Genetic architecture of enrichment for height (HT), body mass index (BMI), cardiovascular disease (CAD) and type-2 diabetes (T2D) for 382,466 unrelated European ancestry UK Biobank individuals genotyped at 8,430,446 SNP markers. a We partition SNP markers into seven location annotations (coding regions, intronic regions, and windows 1, 1–10, 10–500 kb and 500 kb–1 Mb upstream of genes, with other SNPs grouped in a category labelled “others”). Windows 1–10 kb, 10–500 kb and 500 kb–1 Mb upstream of genes are further split into SNPs mapped to enhancers (enh), transcription factor binding sites (tfbs) and others. Within each of the 13 annotations, we have three minor allele frequency groups (MAF ≤ 0.01 annotated as rare, $0.01 < \text{MAF} \leq 0.05$ annotated as low, and $\text{MAF} > 0.05$ annotated as common), and then each MAF group is further split into two based on median LD score. This gives 78 groups for which our BayesRR-RC model jointly estimates the phenotypic variation attributable to, and the SNP marker effects within, each group. For each of the 78 groups, SNPs were modelled using five mixture groups with variance equal to the phenotypic variance attributable to the group multiplied by constants (mixture 0 = 0, mixture 1 = 0.0001, 2 = 0.001, 3 = 0.01, 4 = 0.1). **b** Posterior distribution of the proportion of the total phenotypic variance attributable to the SNP markers that is contributed by each of the four non-zero mixtures within each MAF-annotation group for HT, BMI, CAD and T2D. Within these, are boxplots of the posterior mean and 95% credible intervals. Values are summed over LD groups. **c** Bar plots with error bars giving the 95% credible intervals for the average effect size of markers in the model for each MAF-annotation group, split by mixture.

a specific category, with perhaps the exception of high MAF variants (Fig. 3c). Generally, all phenotypes simply appear to be predominantly underlain by very many common variants, with SNPs within distal regulatory regions, coding and intronic regions contributing more to the variance. We also re-scaled the marker effects by the standard deviation of each marker, to give effect sizes on the allele substitution effect size scale, and again we find that rare variants have higher average allele substitution effects than common variants for exonic, intronic, promoter and enhancers (Supplementary Fig. 12b). An exception to these patterns were BMI-associated intronic and 10–500 kb group SNPs, where we find no evidence that the allele substitution effect size differs across frequency groups (Supplementary Fig. 12b). We

also did not find evidence that the allele substitution effect size differed across frequency groups for transcription factor binding sites, distal SNPs 1 Mb upstream of genes, or those not mapping to an annotation group (Supplementary Fig. 12b).

Discovery of associated genomic regions. We then partitioned the variance attributed to SNP markers across 50kb regions of the genome, then across SNPs annotated to genes, and then to LD blocks of the DNA using our PPWV approach. We find 1660 50 kb regions for height with $\geq 95\%$ posterior probability of explaining 0.001% of the h_{SNP}^2 , 520 regions for BMI, 70 regions for CAD and 87 regions for T2D (Fig. 4a and Table 2). We then map

Table 1 Proportion of genetic variance attributable to different genomic regions for height (HT), body mass index (BMI), type-2 diabetes (T2D) and cardiovascular disease (CAD).

Group	Trait	BayesRR-RC Posterior mean (95% CI)	RHE-mc ^a h_{obs}^2 (se) %	sLDSC ^a h_{obs}^2 (se) %	SumHer ^a h_{obs}^2 (se) %
Variance attributable to SNP markers genome-wide	HT	57.66 (56.09, 59.14)	63.28 (3.57)	64.16 (2.86)	98.58 (0.69)
	BMI	28.74 (27.62, 30.0)	26.76 (1.06)	31.03 (0.9)	44.98 (0.53)
	CAD	5.94 (5.30, 6.67)	4.49 (>100)	4.73 (0.28)	7.33 (0.43)
	T2D	8.45 (7.83, 9.18)	6.90 (0.47)	6.53 (0.3)	11.65 (0.44)
Proportion of genetic variance attributable to exonic regions of genes	HT	24.75 (23.39, 26.071)	27.09	3.00	16.74
	BMI	12.98 (10.98, 14.84)	12.62	4.37	7.60
	CAD	13.23 (8.40, 18.84)	18.68	1.69	15.34
	T2D	14.49 (10.74, 18.54)	14.60	2.46	10.12
Proportion of genetic variance attributable to intronic regions of genes	HT	41.54 (39.91, 43.39)	41.60	46.07	43.03
	BMI	44.17 (41.36, 47.25)	47.87	44.61	48.19
	CAD	32.05 (24.98, 39.51)	41.15	47.22	41.94
	T2D	37.28 (32.22, 42.57)	48.66	38.52	48.02
Proportion of genetic variance attributable to snps 1 kb upstream of genes	HT	2.81 (2.24, 3.42)	1.76	1.46	1.74
	BMI	1.62 (0.75, 2.69)	0.36	1.90	1.15
	CAD	4.20 (1.71, 7.55)	2.49	<0.00	1.26
	T2D	3.58 (1.77, 5.86)	3.40	<0.00	1.57
Proportion of genetic variance attributable to snps 10 kb upstream of genes	HT	6.60 (5.84, 7.40)	6.73	4.29	12.87
	BMI	5.28 (3.92, 6.87)	3.19	6.58	4.10
	CAD	13.06 (8.70, 18.16)	5.70	6.02	8.91
	T2D	9.08 (5.90, 13.28)	4.02	20.44	7.56
Proportion of genetic variance attributable to snps 500 kb upstream of genes	HT	22.13 (21.00, 23.40)	21.53	37.23	24.14
	BMI	28.58 (26.41, 31.01)	28.81	35.86	31.17
	CAD	28.02 (21.24, 35.04)	30.23	38.90	29.58
	T2D	27.42 (22.68, 32.36)	24.33	32.49	27.47
Proportion of genetic variance attributable to exonic regions that is explained by common variants	HT	72.09 (69.77, 74.14)	62.62	75.35	51.22
	BMI	69.41 (62.60, 76.42)	59.67	16.43	54.31
	CAD	64.97 (43.08, 83.16)	61.72	>100	49.17
	T2D	68.57 (56.00, 79.82)	66.33	>100	64.11
Proportion of genetic variance attributable to intronic regions that is explained by common variants	HT	81.19 (79.30, 83.02)	79.96	70.88	66.12
	BMI	85.05 (78.28, 91.49)	86.10	70.62	69.68
	CAD	84.68 (65.64, 95.91)	96.55	61.11	78.17
	T2D	87.62 (75.65, 94.85)	87.63	67.93	71.39
Proportion of genetic variance attributable to snps 500 kb upstream of genes that is explained by common variants	HT	81.59 (78.91, 83.96)	80.66	71.86	77.28
	BMI	86.78 (80.56, 91.60)	89.95	67.38	74.81
	CAD	66.49 (49.11, 81.79)	88.51	60.52	79.91
	T2D	72.35 (58.71, 83.75)	94.91	69.48	75.12

^aRHEmc¹⁸, LDSC¹⁹ and SumHer⁶ provide the total SNP heritability observed (%) and single heritability estimates per genetic component (see Supplementary Data 2-5) that we summarised to obtain the proportion of genetic variance attributed to exonic regions, intronic regions and windows 1, 1-10 and 10-500 kb upstream of genes.

SNPs to their closest gene (+/-50 kb from SNP position) and we use our annotations to label them (see “Methods” section). We find 243 independent coding regions for height with $\geq 95\%$ posterior probability of explaining at least 0.001% of the h_{SNP}^2 , 29 independent coding regions for BMI, 5 for CAD and 13 for T2D. We find many more associations in the cis region of genes with 1254 independent cis-regions for height with $\geq 95\%$ posterior probability of explaining 0.001% of the h_{SNP}^2 , 1765 independent cis-regions for BMI, 1166 for CAD and 1221 for T2D. We additionally find 9 independent promoter regions and 1072 independent introns for height with $\geq 95\%$ posterior probability of explaining at least 0.001% of the h_{SNP}^2 , 1162 independent intronic gene regions for BMI, 307 for CAD and 347 for T2D. When we

calculate the number of exons, introns, promoters and cis regions with $\geq 95\%$ posterior probability of explaining 0.001% of the h_{SNP}^2 , as a proportion of the total number within each chromosome, we find that up to 24% of the genes on each chromosome are associated with each of the four traits (Fig. 4b). Generally, we find that only 1% or less of the available exons and promoter regions of genes per chromosome show an association with each of the phenotypes, but up to 14% of the available intronic regions and up to 10% of the cis-regions surrounding genes contribute to the phenotypic variance with $\geq 95\%$ probability (Fig. 4b). The variance contributed by each exonic, intronic, promoter, or cis region is typically only a small fraction of a percent, with largest effect sizes being the exonic region of GDF5 contributing 0.26%



Fig. 4 Contribution of genes and 50kb regions to height (HT), body-mass-index (BMI), cardiovascular disease (CAD) and type-2-diabetes (T2D).

a We grouped SNPs in 50 kb-regions genome-wide and estimated the sum of the squared regression coefficient estimates for each 50 kb-region. We then select the number of 50 kb regions that explain at least 0.001% of the variance attributed to all SNP markers in 80, 90 and 95% of the iterations. This gives a measure called the posterior probability that the window variance (PPVW)²⁰ exceeds 1/10,000 of the phenotypic variation attributed to SNP markers. **b** We mapped SNPs to the closest gene ± 50 kb from the SNP position and labelled them as located in a coding region, an intron, 1 kb upstream of a gene using our functional annotations (Fig. 3a). Remaining snps are labelled as located in a cis-region (up to ± 50 kb from a gene). We then select the number of regions where PPVW is higher than 95% and explains at least 0.001 % of the phenotypic variance attributed to all SNP markers. We then calculate the number of significant coding regions, introns, 1 kb regions and cis regions as a proportion of the total number of genes for each chromosome. Genic associations that explain at least 0.001% of the phenotypic variance attributed to all SNP markers are again spread across chromosomes according to the chromosome length. **c** Shows the mean of the phenotypic variance attributed to intron and cis regions (y-axis) and coding regions (x-axis) that explain at least 0.001% of the phenotypic variance attributable to SNP markers in $\geq 95\%$ of the iterations (PPVW > 0.95). These results provide joint estimates of the proportions of variance contributed by different gene bodies and automatic fine-mapping of gene bodies and their cis-regulatory regions. For example, introns and cis-regulatory regions of FTO respectively contribute 0.48% (95% CI 0.29, 1.12) and 0.01% (95% CI 0, 0.01) to the phenotypic variance of BMI. **d** We calculated the phenotypic variance contributed by exonic, intronic, promoter region and SNPs ± 50 kb outside of the exon and promoter regions (cis) for each gene. Bar plots show the correlation among the variance explained by the groups across genes. Error bars show the SD.

(95% CI 0.21, 0.32) to the phenotypic variance of height, the intronic region of FTO contributing 0.48% (95% CI 0.29, 1.12) to BMI, both the exonic-region and intronic-region of LPA contributing a combined 0.08% (95% CI 0.04, 0.13) to the risk of CAD, and the intronic region of TCF7L2 contributing 0.28% (95% CI 0.23, 0.35) to the risk of T2D (Fig. 4c, full results in Supplementary Data 6–9). Taken together, these results support an infinitesimal contribution of many thousands of genes to common complex trait variation and give joint estimates of the proportions of variance contributed by each gene and their probability of association.

For each gene, we also calculated the phenotypic variance contributed by exonic, intronic, promoter region, and cis SNPs and then calculated the correlation among the variances explained by the groups across genes. Across traits, we find small positive correlations of the variance attributable to exonic and intronic regions of 0.17 (0.09, 0.24 95% CI) for height, 0.02 (0.001, 0.05 95% CI) for BMI, 0.103 (–0.007, 0.71 95% CI) for CAD, and 0.064 (0.01, 0.19 95% CI) for T2D. Similarly, we find small positive correlations between introns and cis regions

(Fig. 4d). With the exception of height, there was no evidence for a relationship among the following groups: (i) SNPs in the exons of each gene and SNPs ± 50 kb outside of the exon and promoter regions; (ii) SNPs in the exons of each gene and SNPs in proximal promoters; and (iii) intronic SNPs and SNPs in promoter regions (Fig. 4d). This implies that trait associated SNPs in proximal and distal regulatory regions are largely independent of the effects of SNPs in their closest exon, as they do not align in terms of the variance they explain (Fig. 4d). For height, small weakly positive correlations across all gene regions in their contribution to variance, implies a degree of alignment across genes in regulatory variants and the closest exon (Fig. 4d). These results suggest a regulatory link between introns and distal cis regions outside of the promoter, or that introns may be correlated with structural variation. They also imply that the variance contributed by regulatory regions and those in the closest coding regions are not strongly coupled for these common complex traits.

Finally, our approach provides automatic fine-mapping of SNP loci, and of these region-level and gene-level associations, 360

Table 2 Summary of findings for height (HT), body mass index (BMI), type-2 diabetes (T2D) and cardiovascular disease (CAD).

Findings	Method	HT	BMI	CAD	T2D
Associated SNPs	COJO-plink2	1673	517	34	85
	COJO-	2131	565	34	84
	BoltLMM				
	COJO-	2134	555	34	82
	Regenie				
50 kb regions (PPWV \geq 95%)	BayesRR-RC	1660	520	70	87
Genic regions (PPWV \geq 95%)	BayesRR-RC	2578	2956	1478	1581
Exons		243	29	5	13
Introns		1072	1162	307	347
cis ^a		1254	1765	1166	1221
SNPs (PIP \geq 95%)	BayesRR-RC	360	20	2	9
Exons		216	16	1	4
Introns		73	2	1	5
10–500 kb		48	1	0	0
LD clumps with $r_2 = 0.1$ (PPWV \geq 95%)	BayesRR-RC	1220	206	16	19

^aSNPs located up to ± 50 kb from the closest gene.

SNPs for height, 20 for BMI, 2 for CAD and 9 for T2D could be mapped to a single SNP with greater than 95% inclusion probability across all four chains (Supplementary Data 10 and Supplementary Fig. 13). Of these fine-mapped SNPs, only 53.45% are top loci with a p -value $< 5 \times 10^{-8}$ from the fastGWAS UK Biobank summary statistic data for standing height, BMI, angina/heart attack and type-2 diabetes (fastGWA, see “Code availability”). This highlights that selecting on the top SNP markers identified by standard association studies would give a different set of variants than those obtained from selecting high PIP SNPs.

Out-of-sample prediction into another European healthcare system. We generated a full posterior predictive distribution for each trait in each of 32,500 individuals from the Estonian Genome Centre data, which allows the transmission of uncertainty in the marker effect estimates from the UK Biobank to the genomic predictors created in Estonia. First, despite this study having almost half the sample size, we show improved genomic prediction as compared to recently proposed summary statistic approaches²³, when taking the mean of the predictor across iterations and correlating this with the phenotype with correlation of 0.62 for height, 0.34 for BMI, 0.16 for T2D, and 0.07 for CAD (Supplementary Fig. 14a). The area under the receiver operator curve (AUC) for T2D was 0.67 and 0.57 for CAD. In comparison, using the 64 BLD-LDAK annotations recommended by a recent study²¹, the highest prediction accuracy obtained from MegaPRS was 0.55 for height, 0.32 for BMI, 0.10 for T2D, and 0.05 for CAD.

We then estimated the distribution of the partial correlations between the trait and genomic predictors created from our different annotation groups and find that exonic, intronic, and 10–500 kb upstream regions contribute proportionally more to the prediction accuracy than other genomic groups, replicating our results from the UK Biobank (Supplementary Fig. 14). We find evidence for zero/low correlations of genomic predictors created from different annotation groups, which supports our results from the UK Biobank (Supplementary Fig. 14e). This suggests that individuals have a different portfolio of risk variants, with different genomic regions contributing for different

individuals to their overall genetic value, as expected under a highly polygenic model.

Additionally, for height and BMI we also determined the proportion of the posterior predictive distribution for each individual that was within ± 1 SD of their true phenotypic value. On average 67.5% of an individual's posterior predictive distribution is within ± 1 SD of their true phenotype for BMI and 75% for height, with similar prediction accuracy across individuals (Supplementary Fig. 14c). For T2D and CAD, we extended the PCF metric, typically defined as the proportion of cases with larger estimated risk than the top p th percentile of the distribution of genetic risk in the general population. For each individual, we calculated the proportion of their posterior predictive distribution that falls above the top 25% of the distribution of genetic risk in the general population. The distribution of these probabilities is shown for confirmed cases and those without diagnosis in the Estonian Biobank (Supplementary Fig. 14d). We find 25 individuals for T2D and 15 individuals for CAD where $\geq 90\%$ of their posterior predictive distribution is within the high risk group of which 40 and 18% are currently defined as cases for T2D and CAD, respectively based on recent medical records. This is compared to 1% and 2% case rate for those with $\leq 10\%$ probability of being in the high risk group for T2D and CAD respectively, giving an odds ratio of 20 and 18 between the $\geq 90\%$ and $\leq 10\%$ groups. However, our results clearly show that the individual-level sensitivity and specificity of genomic prediction for these common complex diseases is very poor, as 75% of T2D cases and 92% of CAD cases have $\leq 50\%$ of their distribution within the high-risk category. These results highlight how variation contained within a posterior predictive distribution that is typically ignored in human genomic prediction can be used. We show that genomic prediction for personalised medicine with patient-specific predictions or stratification of patients is currently extremely limited.

Discussion

There is no single statistical model appropriate for all settings and thus there will always be a situation where a model poorly fits the data. We have provided theoretical and empirical evidence that a grouped Dirac spike-and-slab model (which we term BayesRR-RC), has a prior that is flexible enough to show robust model performance across the data analysed here, improving inference in many settings over commonly applied approaches. We develop a range of computational and statistical approaches which allow this, or any similar Gibbs sampling algorithm, to scale to whole genome sequence data on many hundreds of thousands of individuals. This has enabled us to compare and contrast the inferred underlying genetic distribution for four complex phenotypes under this prior, providing novel insight into the genetic architecture of these traits. We observe that all phenotypes simply appear to be predominantly underlain by very many common variants, with SNPs within distal regulatory regions, coding and intronic regions each contributing more to the phenotypic variance and having higher allele substitution effects.

There has been debate on how to best estimate SNP heritability^{1,3,4} and here we validate that one approach could be to split SNP markers by LD to improve genetic effect size estimates. Our results suggest that the proportion of genomic variation attributable to mutations in regulatory regions and mutations in the closest genic regions are largely independent. Additionally our model tests association within groups in a probabilistic way and we find 290 independent coding, 2888 independent intronic, and 5406 independent cis regions with $\geq 95\%$ probability of contributing at least 0.001% of the SNP heritability. Understand how these coding, intronic and proximal and distal regulatory regions combine to contribute to

phenotypic variance remains a substantial challenge and our results suggest a predominant role for introns and for distal, and thus likely more global enhancers, rather than locally dominant proximal expression QTL. The recent “omnigenic” model²⁴, suggests that trait-associated variants in regulatory regions influence a local gene which is not directly causal to the disease, and also co-regulate other disease causal genes (or “core” gene). Our findings of little correlation of exonic and proximal regulatory variance and a large number of trait-associated intronic and cis regions do not rule this out, but suggest a more complex infinitesimal picture with differences occurring among traits, potentially due to their evolutionary history.

There are important caveats and limitations to consider. Here, we present an approach for analysing large-scale biobank data, which is becoming increasingly available. However, a substantial number of GWAS have already been conducted, with associated published genome-wide summary association statistic estimates. Many methods have been developed to take advantage of these estimates, with downstream analysis models making use of various summary statistics resources in efficient and flexible ways. We show here that two leading summary statistic approaches perform poorly as compared to individual-level models for estimation of enrichment and genomic prediction. Despite this, the sample sizes obtained in consortia study meta-analyses will exceed those from single biobanks, especially for disease, and thus the genomic prediction accuracy of consortia study meta-analysis summary statistic prediction models may exceed those from individual-level analyses. Combining the posterior distribution obtained from BayesRR-RC across different individual-level biobank studies would alleviate this issue.

Additionally, in this work we do not extend past a limited number of functional annotations and thus we do not provide a model capable of further partitioning the variation into specific regulatory functions (eQTL, mQTL, pQTL etc.) or directly modelling the relationships among components. LDSC functional methods take the approach that SNPs can be assigned to different categories (e.g., both coding and conserved), with the categories competing against each other to explain the signal, with the downside that enrichment is relative and that the total variance is not partitioned. Here, the total variance is partitioned but this is based on preferential allocation of SNPs to coding regions, then introns, and then to their nearest upstream gene position. These SNPs are most likely to be allocated accurately, with 1 and 1–10 kb groups being more ambiguous in high gene density regions and likely mislabelled. However, if this was the case then variance would still be partitioned to these mislabelled groups and it would just be evenly split across them, with experimentally validated promoter, enhancer and tfs regions assisting to some degree in alleviating this. Rather, here we see a clear pattern of increasing variance contributed, increasing average effect size, and an increasing pattern of higher rare allele substitution effects by individual markers as distance from the nearest gene increases. 10–500 kb distal regions may contribute more variance as marker density and marker coverage is higher in these regions, with missing variation within 10 kb upstream as causal variants are poorly correlated with SNPs. The posterior distributions for the variance explained by 1 kb, 1–10 kb regions, and 10–500 kb regions are negatively correlated (Supplementary Fig. 8, meaning that these groups are competing with each other, as if variance goes to one then it is being taken away from the other because they are in LD), and thus there is the risk that the model cannot separate these effectively. However, this is true of any enrichment analysis conducted to date and we can only make inference in the data that we have currently available. Resolving this requires the application of this model to whole genome sequence data where the total variance can be partitioned across upstream regions without marker coverage concerns. Irrespective of exactly which

upstream region variance is allocated to, our inference that genic regions are uncorrelated in their contribution to variance with the promoter and upstream regions still holds as does our probabilistic inference on the associations of each gene and their contribution to the phenotypic variation.

Our results provide evidence for an infinitesimal contribution of many thousands of common genomic regions to common complex trait variation and for a predominant role of intronic, exonic, and distal regulatory regions. This highlights the immense challenge of understanding the molecular underpinning of each association and the difficulties in improving the estimation of many tens of thousands of small-effect associations that are required to improve genomic prediction. This work represents a step toward maximising the probabilistic inference that can be obtained from large-scale Biobank studies.

Methods

BayesRR-RC model. We extend the BayesR model to a BayesRR-RC model as follows

$$\mathbf{y} = \mathbf{1}\mu + \sum_{\varphi=1}^{\Phi} \mathbf{X}_{\varphi} \beta_{\varphi} + \epsilon, \quad (3)$$

where there is a single intercept term $\mathbf{1}\mu$ and a single error term ϵ but now SNPs are allocated into groups ($\varphi_1, \dots, \varphi_{\Phi}$), each of which having its own set of model parameters $\Theta_{\varphi} = \{\beta_{\varphi}, \pi_{\beta_{\varphi}}, \sigma_{\epsilon}^2\}$. As such, each β_{φ} is distributed according to:

$$\beta_{\varphi} \sim \pi_0 \delta_0 + \pi_1 \mathcal{N}(0, \sigma_{1\varphi}^2) + \pi_2 \mathcal{N}(0, \sigma_{2\varphi}^2) + \dots + \pi_{L\varphi} \mathcal{N}(0, \sigma_{L\varphi}^2), \quad (4)$$

where for each SNP marker group $\{\pi_0, \pi_1, \dots, \pi_{L\varphi}\}$ are the mixture proportions and $\{\sigma_{1\varphi}^2, \sigma_{2\varphi}^2, \dots, \sigma_{L\varphi}^2\}$ are the mixture-specific variances proportional to

$$\begin{bmatrix} \sigma_{1\varphi}^2 \\ \vdots \\ \sigma_{L\varphi}^2 \end{bmatrix} = \sigma_{\beta_{\varphi}}^2 \begin{bmatrix} C_{1\varphi} \\ \vdots \\ C_{L\varphi} \end{bmatrix}$$

Thus the mixture proportions, variance explained by the SNP markers, and mixture constants are all unique and independent across SNP marker groups. This extends previous models (known as BayesRC²⁵ and BayesRS²⁶), which have used additional mixtures for different SNP groups, but kept a single global variance component. Importantly, a single variance component with more mixtures serves only to change the amount of mass allocated at different sizes of the distribution, but does not alter the sizes of the effects themselves as there is still a single distribution. In contrast, the formulation presented here of having an independent variance parameter $\sigma_{\beta_{\varphi}}^2$ per group of markers, and independent mixture variance components, enables estimation of the amount of phenotypic variance attributable to the group-specific effects and enables differences in the distribution of effects among groups. In this work, we use 78 SNP marker groups, each with five mixture components (including 0).

We can sketch the difference in the models by looking at the respective conditional posteriors, again, assuming a single component for simplification purposes. We have a BayesRC or BayesRS estimator by assuming different groups of effects as described in Supplementary Note 4 Eq. 35, which yields:

$$f(\alpha, \gamma | \pi_{\beta_{\varphi}}, \sigma_{\beta_{\varphi}}^2, \sigma_{\epsilon}^2, \mathbf{y}) \propto \exp \left\{ \frac{1}{2\sigma_{\epsilon}^2} \|\mathbf{y} - \mathbf{X}_{\gamma \neq 0} \alpha_{\gamma \neq 0}\|_2^2 - \frac{1}{2\sigma_{\beta_{\varphi}}^2} \|\alpha\|_2^2 - \log \left(\frac{1 - \pi_{\beta_{\varphi}}}{\pi_{\beta_{\varphi}}} \right) \|\gamma_{\varphi}\|_0 \right\}, \quad (5)$$

where $\pi_{\beta_{\varphi}}$ are the group-specific mixture proportions and $\|\gamma_{\varphi}\|_0$ is the cardinality of the group. The corresponding MAP estimate would amount to adding extra penalisation on sparsity through the $\pi_{\beta_{\varphi}}$ terms, while keeping the same level of shrinkage as the baseline BayesR.

In our model the conditional posterior is:

$$f(\alpha, \gamma | \pi_{\beta_{\varphi}}, \sigma_{\beta_{\varphi}}^2, \sigma_{\epsilon}^2, \mathbf{y}) \propto \exp \left\{ \frac{1}{2\sigma_{\epsilon}^2} \|\mathbf{y} - \mathbf{X}_{\gamma \neq 0} \alpha_{\gamma \neq 0}\|_2^2 - \frac{1}{2\sigma_{\beta_{\varphi}}^2} \|\alpha\|_2^2 - \log \left(\frac{1 - \pi_{\beta_{\varphi}}}{\pi_{\beta_{\varphi}}} \right) \|\gamma_{\varphi}\|_0 \right\} \quad (6)$$

now each marker has a group-specific shrinkage $\sigma_{\beta_{\varphi}}^2$, which translates to a specific λ_{φ} per group in the MAP estimate. This amounts to markers being shrunk according to the scale of their group, instead of the scale of all other markers. So instead of solving a single model selection and regularisation problem we are solving Φ model selection and regularisation problems, with shared information only through the residuals. If we subset by MAF and LD bins, the resulting groups of columns will have a correlation pattern similar to an exponential decay (LD decays with distance). If we take the whole genotype matrix, the pattern would be closer to a block diagonal matrix of correlations, in refs. 16,27

it is shown that the former case requires weaker conditions in order to recover the true vector β consistently than the latter. Although the sampling scheme was different, we have shown that a similar model with only two groups: genetic markers and epigenetic markers, is successful in identifying BMI and smoking epigenetic signatures¹³. The baseline model derivations for this model are outlined in Supplementary Note 1, a BSP Gibbs sampling scheme and an assessment of its performance is outlined in Supplementary Note 2, and an assessment of the model performance with correlated covariates is outlined in Supplementary Note 4.

Simulation study

Genetic architecture. We first compare the model performance of BayesRR-RC to existing approaches across 18 different genetic architectures. We randomly selected 40,000 unrelated UK Biobank individuals and used 596,741 imputed SNP markers from chromosomes 19 to 22. We randomly selected either 1000, 10,000, or 100,000 LD independent ($LD R^2 < 0.1$) causal SNP markers. For each SNP marker set there were two settings.

In the first setting, we simulated effect sizes from a normal distribution with zero mean and variance of 0.1, 0.3, or 0.6 divided by the number of causal variants $\propto N(0, [p(1-p)]^{-0.25})$, with p the allele frequency. We sampled individual-level environmental (residual) variance from a normal distribution with zero mean and variance equal to 1 minus either 0.1, 0.3, or 0.6 to give phenotypes with zero mean and unit variance. This gave $h_{SNP}^2 = 0.1, 0.3, \text{ or } 0.6$ and simulates stronger effect sizes for rare variants in line with recent empirical estimates. We simulated ten replicate phenotypes for each of the nine different genetic architectures. In the second setting, we repeat each simulation, sampling the SNP marker effects from 13 different normal distributions, one for each of 13 different genomic annotation groups described in the main text. The 13 groups were allocated different proportions of the h_{SNP}^2 as follows: for exonic variants $P(h_{SNP}^2) = 0.167$, intronic variants $P(h_{SNP}^2) = 0.334$, 1 kb promoter variants $P(h_{SNP}^2) = 0.0835$, 1–10 kb enhancer variants $P(h_{SNP}^2) = 0.04175$, 1–10 kb transcription factor binding sites $P(h_{SNP}^2) = 0.04175$, 1–10 kb other variants $P(h_{SNP}^2) = 0$, 10–500 kb enhancers $P(h_{SNP}^2) = 0.0835$, 10–500 kb transcription factor binding sites $P(h_{SNP}^2) = 0.0835$, 10–500 kb other variants $P(h_{SNP}^2) = 0$, 500 kb–1 Mb enhancers $P(h_{SNP}^2) = 0.0835$, 500 kb–1 Mb transcription factor binding sites $P(h_{SNP}^2) = 0.0835$, 500 kb–1 Mb other variants $P(h_{SNP}^2) = 0$, and other non-annotated SNPs $P(h_{SNP}^2) = 0$. For each of the 13 groups marker effects were simulated as $\propto N(0, [p(1-p)]^{-0.25})$ to give $h_{SNP}^2 = 0.1, 0.3, \text{ or } 0.6$, with stronger effect sizes for rare variants. Four of these 13 groups had zero variance indicating that no associations were created for these groups.

Thus, in the first setting we simulate variance explained by annotation groups that is on average proportional to the number of SNPs within each annotation (due to the random allocation of SNPs and effect sizes). In the second setting, the variance and average effect size differ across annotation groups. We refer to these as two different enrichment settings: “random”, or “enriched”.

For these 180 phenotypes, we ran the following individual-level models:

- A restricted maximum likelihood model implemented in the software GCTA with a single relationship matrix providing an estimate of the variance attributable to SNPs genome-wide.
- A restricted maximum likelihood model implemented in the software BoltREML¹⁷. Here, we used a 78 MAF-LD-annotation group model using the non-overlapping genomic annotation groups described below in the UK Biobank analysis providing an estimate of the variance attributable to SNPs genome-wide and an estimate of the variance attributable to SNP markers of each annotation group.
- A Haseman-Elston regression using the same 78 group model implemented in the software RHEmc¹⁸, providing an estimate of the variance attributable to SNPs genome-wide and an estimate of the variance attributable to SNP markers of each annotation group.
- Mixed linear association model (MLMA), which is a two-stage approach where the variance attributable to the SNP markers genome-wide is estimated and this estimate is then used in a second generalised least squares step to test for SNP-phenotype associations one marker at a time. There are two forms of this model. In the first, the SNP is fitted twice as it is included in both the fixed and random terms (MLMAi). In the second, the SNP to be tested as fixed is removed from the random term alongside those on the same chromosome (MLMA). We used the software BoltLMM⁸, Regenie⁹, and GCTA to fit these models. These approaches provided estimates of the SNP regression coefficients (marker effect sizes).
- Single marker marginal least squares regression using plink2²⁸, whilst fitting 20 principal components of the marker data as covariates.
- Linkage disequilibrium score regression (LDSC¹⁹), with LD scores calculated using the same data, and the same 78 non-overlapping annotations in a 78 component LDSC annotation model. We included SNPs with MAF > 1% following the software instructions. This model is intended to approximate an individual-level REML analysis with 78 annotations and provides an estimate of the variance attributable to SNPs genome-wide and an estimate of the variance attributable to SNP markers of each annotation group.
- We used the software SumHer⁶. We calculated marker taggings under the same 78 component annotation model. We ignored the LD weights when

calculating the taggings as we found this gave the best estimates we could obtain from the simulated data across all scenarios. We set the relationship of effect size and minor allele frequency to be -0.25 as suggested by the authors and which matches the simulation setting. This model is intended to approximate an individual-level REML analysis with 78 annotations, but using a different scaling of the relationship matrix, and provides an estimate of the variance attributable to SNPs genome-wide and an estimate of the variance attributable to SNP markers of each annotation group.

- Our BayesRR-RC model implemented in GMRM with 78 SNP-marker groups and run for 5000 iterations with a burn-in period of 2000 iterations.
- Our BayesRR-RC model implemented in GMRM with only a single SNP-marker group, which is equivalent to BayesR, run for 5000 iterations with a burn-in period of 2000 iterations.

We then ran the following prediction models, using a testing set of 10,000 UK Biobank unrelated individuals, that were also unrelated to the training data, and focusing on the models proposed in a recent paper²¹. These methods contain two approximations to our BayesRR-RC model and the authors claim to outperform all other existing methods, including individual-level models. The models are:

- An individual-level bayesR model using genomic annotation SNP variance estimates from the SumHer models as implemented in the software MegaPRS²¹. This provides estimates of the SNP marker effects for creating a genetic risk predictor.
- An individual-level boltREML model using genomic annotation SNP variance estimates from the SumHer models as implemented in the software MegaPRS²¹. This provides estimates of the SNP marker effects for creating a genetic risk predictor.
- A summary statistic bayesR model using genomic annotation SNP variance estimates from the SumHer models as implemented in the software MegaPRS²¹. This provides estimates of the SNP marker effects for creating a genetic risk predictor.
- A summary statistic boltREML model using genomic annotation SNP variance estimates from the SumHer models as implemented in the software MegaPRS²¹. This provides estimates of the SNP marker effects for creating a genetic risk predictor.

First, we compared the correlation of the simulated and estimated proportion of phenotypic variance attributable to the 13 genomic annotation groups across all models in Fig. 1. We determined the ability of the approaches to correctly identify enriched regions of the DNA by estimating the probability within each simulation replicate that a SNP marker group would have an estimated enrichment of ≥ 2 (i.e., being described as having average effect sizes that are twice as large as expected) when the simulated value was ≤ 1.1 . As BayesRR-RC induces sparsity in the SNP effect estimates, with some markers always remaining in the variance = 0 spike, we propose a different enrichment definition where the proportion of h_{SNP}^2 is divided by the proportion of markers that are in the model for the SNP group, rather than the proportion of markers mapping to the SNP group.

In Supplementary Note 3, we propose a posterior probability window variance (PPWV) approach²⁰, which provides a probabilistic determination of association of a given LD block, genomic window, gene, or upstream region, relative to the amount of phenotypic variation attributable to that window. Our PPWV approach determines the posterior inclusion probability that each region and each gene contributes at least 0.001% to the h_{SNP}^2 , with theory and small-scale simulations outlined in Supplementary Note 3 suggesting well controlled FDR. We partitioned the 596,741 imputed SNP markers in LD blocks, defined as groups of markers with $LD R^2 \geq 0.1$. Within each simulation replicate, we estimated the probability that LD blocks containing a causal variant were identified by PPWV. We compared this to MLMA estimates obtained using the BoltLMM software, by estimating the probability that LD blocks containing a causal variant were identified as having a SNP with p -value $\leq 5 \times 10^{-8}$, the standard genome-wide significance threshold. We present these results in Fig. 2a.

We then compare the prediction accuracy obtained in a testing set of 10,000 UK Biobank unrelated individuals, that were also unrelated to the training data. We predicted phenotype using SNP marker effect sizes obtained from BayesRR-RC, MLMA effect sizes from BoltLMM, and the four MegaPRS methods outlined above implemented in the LDK software. While we would suggest that fixed-effect MLMA estimates are improper for prediction we include this comparison as polygenic risk scores have often been created from fixed-effect SNP estimates. We calculate the correlation between the simulated phenotype in the testing set and the genomic predictor within each simulation replicate and we compare the mean correlation across the 18 different genomic annotations in Fig. 2. Additionally, to provide a benchmark, we compare to the theoretical expectation under ridge regression approximations²⁹, with the number of markers set to the number of causal variants.

Relationship between effect size, minor allele frequency and LD. We then conducted another large-scale, but this time well-powered simulation study, where we ascertained the causal variant SNP markers in different ways and varied the relationship between effect size, minor allele frequency and LD. We used the same randomly selected 40,000 unrelated individuals and all 596,741 imputed (version 3) genetic markers from chromosomes 19 through 22 from the UK Biobank. We

simulated a wide-range of different possible underlying genetic effect size distributions as follows:

- We chose either 5000 or 10,000 imputed SNP markers for which to assign a genetic effect size, providing two different levels of polygenicity.
- We selected these 5000 or 10,000 markers in two different ways. Either, we selected SNPs at random, or we selected the marker of highest minor allele frequency per LD block of the genome, with an LD block defined as a group of SNP markers with absolute LD of at least 0.05. Randomly allocating markers creates a set of associated variants with the same distribution of LD and MAF as the SNP data, which is composed of predominantly low frequency variants. Selecting only the highest frequency marker per LD block creates a setting where for each set of markers in LD with each other, there is only one causal genetic variant, and where the distribution of associated markers differs to that of the SNP markers as a whole.
- Having created four different ways of selecting associated markers (5000 or 10,000 and high-MAF or random) we then created five different ways of assigning effect sizes to them:
 - We simulated effect sizes from a normal distribution with zero mean and variance 0.6 divided by the number of markers (5000 or 10,000) with no relationship to the LD or MAF of the markers. Thus, effects had variance $\propto N(0, w^0[p(1-p)]^0)$ with w the LD score of the marker and p the allele frequency.
 - We simulated effect sizes from a normal distribution with zero mean and variance 0.6 divided by the number of markers (5000 or 10,000) $\propto N(0, w^{-0.25}[p(1-p)]^{-0.25})$. This simulates stronger effect sizes for rare variants and those in low LD.
 - We simulated effect sizes from a normal distribution with zero mean and variance 0.6 divided by the number of markers (5000 or 10,000) $\propto N(0, w^{0.25}[p(1-p)]^{0.25})$. This simulates stronger effect sizes for rare variants and those in high LD.
 - We simulated effect sizes from a normal distribution with zero mean and variance 0.6 divided by the number of markers (5000 or 10,000) $\propto N(0, w^{-0.25}[p(1-p)]^{0.75})$. This simulates equivalent effect sizes for common and rare variants, and greater effects for markers in low LD.
 - We simulated effect sizes from a normal distribution with zero mean and variance 0.6 divided by the number of markers (5000 or 10,000) $\propto N(0, w^{0.25}[p(1-p)]^{0.75})$. This simulates equivalent effect sizes for common and rare variants, and greater effects for markers in high LD.
- For each of the four different sets of markers, each with five different effect size sampling schemes, we then created two additional settings. In the first setting markers were sampled from the various normal distribution, as described above, for the five different effect size sampling schemes. In the second setting, for each of the five different effect size sampling schemes we simulated effects from 13 different distributions, one for each of 13 different genomic annotation groups with different proportions of total SNP heritability (h_{SNP}^2). For each of the five different effect size sampling schemes the relationship to LD and MAF remained the same, but the total variance attributed to the SNP markers was partitioned across annotation groups as follows for exonic variants ($h_{\text{SNP}}^2 = 0.1$), intronic variants ($h_{\text{SNP}}^2 = 0.2$), 1 kb promoter variants ($h_{\text{SNP}}^2 = 0.05$), 1–10 kb enhancer variants (0.025), 1–10 kb transcription factor binding sites ($h_{\text{SNP}}^2 = 0.025$), 1–10 kb other variants ($h_{\text{SNP}}^2 = 0$), 10–500 kb enhancers ($h_{\text{SNP}}^2 = 0.05$), 10–500 kb transcription factor binding sites ($h_{\text{SNP}}^2 = 0.05$), 10–500 kb other variants ($h_{\text{SNP}}^2 = 0$), 500 kb–1 Mb enhancers ($h_{\text{SNP}}^2 = 0.05$), 500 kb–1 Mb transcription factor binding sites ($h_{\text{SNP}}^2 = 0.05$), 500 kb–1 Mb other variants ($h_{\text{SNP}}^2 = 0$), and other non-annotated SNPs ($h_{\text{SNP}}^2 = 0$). Four of these distributions had zero variance indicating that no associations were created for these groups. In the first setting, this simulates variance explained by annotation groups that is on average proportional to the number of SNPs within each annotation. In the second scheme, the variance and average effect size differs across annotation groups. We refer to these as two different enrichment settings: “random”, or “enriched”.
- This created four different sets of associated markers (5000 or 10,000 and high-MAF or random), each with five different marker effect size sampling schemes, which we refer to in the main text as the 20 different generative genetic models (Table 1), each of which has two enrichment settings. This gave 40 different sampling schemes for the genetic effects and we simulated ten replicates for each setting, giving a total set of 400 simulated phenotypes.
- For each generative model the total genetic variance was 0.6 and we sampled individual-level environmental (residual) variance from a normal distribution with zero mean and variance 0.4 to give phenotypes with zero mean and unit variance.

This range covers generative genetic models discussed in the literature and provides models that both fit and violate the assumptions of the range of variance component statistical models. This includes both individual-level and summary statistic approaches, that are currently applied in the literature for estimation of the

variance attributable to the SNP markers, for testing association of genetic markers with phenotypes genome-wide, and for genomic prediction.

This simulation provides a range of different scenarios for which we can explore the model performance of BayesRR-RC and compare it to existing approaches. In Supplementary Fig. 1, we compare the h_{SNP}^2 estimation, estimation of the annotation genetic variance along with the RMSE of the estimates, and the estimated average effect size.

We then extend our model comparisons in a number of ways. While direct comparisons of frequentist and Bayesian approaches are difficult and often ill advised, we wished to show that BayesRR-RC provides accurate effect size estimation in the presence of LD. We provide three simple comparable metrics to assess model performance of BayesRR-RC against frequentist mixed linear association models (MLMA) applied as two-stage approaches, where either the SNP is fitted twice as it is included in both the fixed and random terms (MLMAi implemented in GCTA), or the SNP to be tested as fixed is removed from the random term alongside those on the same chromosome (MLMA implemented in BoltLMM and Regenie).

First, we calculated z -scores of the marker effect estimates from their true simulated value. As MLMA approaches estimate marker effects one-at-a-time, we calculated the z -score of the estimate from the true simulated value for the causal variants in each simulation replicate, across generative genetic models. For the Bayesian methods, at any one iteration, LD among the markers is controlled for (see Supplementary Note 4). However across iterations as the chain mixes, markers in LD will enter and leave the model, with their posterior inclusion probabilities reflecting their association with the trait. Thus, we summed the squared regression coefficient estimates of SNPs in the model at each iteration for each LD block (markers in LD $R^2 \geq 0.1$ within 1 MB) of each simulation replicate, took the posterior mean across iterations, and then calculated the z -score of the estimate from the simulated value. This metric provides an assessment of the ability of BayesRR-RC to accurately estimate the contribution of a genomic region to the phenotypic variance as compared to MLMA approaches. We present these results in Supplementary Fig. 2, where we find that the z -scores of the estimated BayesRR-RC effects are generally stable across generative genetic models and comparable to those obtained from BayesR but with slightly elevated variance in many cases as the model is less sparse (Supplementary Fig. 2a). We find that SNP effect size estimates from MLMA models have higher estimation error, especially when the causal variant is rare, or in high-LD with many other SNPs (Supplementary Fig. 2a). MLMAi models show lower estimation error than MLMA approaches, likely as they control for both distant and local LD (Supplementary Fig. 2a). We explore this further in Supplementary Note 4.

Second, to further test our PPVW approach we calculated precision-recall curves, where associations are defined as LD blocks with PPVW of $\geq 95\%$ at 0.001% proportion of variance explained. True positives were the number of identified 5000 or 10,000 LD blocks that contained a causal variant. False positives were the number of identified LD blocks that did not contain a causal variant. Precision was defined as the ratio of true positives to the sum of true positives and false positives. Recall was defined as the ratio of true positives to the sum of true positives plus false negatives. The FDR was defined as the proportion of LD blocks with PPVW of $\geq 95\%$ at 0.001% proportion of variance explained that did not contain a causal variant. For the MLMA methods, following standard practice, we clumped the marker effect estimates using Plink, as local LD is not controlled for, selecting LD independent markers ($LD R^2 \leq 0.01$ with other markers) across the genome. True associations were defined as selected SNPs that were in LD with a simulated causal variant ($LD R^2 \geq 0.01$). False associations were defined as selected SNPs that were not in LD ($LD R^2 \leq 0.01$) with a simulated causal variant. Precision and recall were calculated across thresholds of the chi-squared statistics of the selected markers, and the area under the curve was calculated using the trapezoid rule for calculating the integrals, assuming the curve is linear between the points. FDR is then calculated as the proportion of markers with p -value $\leq 5 \times 10^{-8}$ that were not in LD with a causal variant ($LD R^2 \geq 0.01$). This provides a way to directly compare model performance for the discovery of associated genomic regions across Bayesian and frequentist approaches and tests our hypothesis that a PPVW approach controls FDR well in comparison with Bonferroni p -value correction (Supplementary Fig. 2b, c). For both MLMA and Bayesian approaches our definition of FDR is not strictly the FDR. Markers in $LD R^2 \leq 0.01$ with the clumped selected markers may still show a weak correlation with the simulated causal variants, and likewise blocks of SNPs in $LD R^2 \leq 0.1$ may still be in weak LD with the causal variants. Our approach instead captures the ability of MLMA and Bayesian approaches to localise an effect within $R^2 \geq 0.01$ and $R^2 \geq 0.1$ respectively. We present these results in Supplementary Fig. 2.

Third, we wished to determine the out-of-sample phenotypic prediction performance of BayesRR-RC. We used the same randomly selected 10,000 individuals from the UK Biobank that were unrelated to those used in the simulation. Using the same SNP markers and the simulated marker effects we calculated a simulated genetic value for each individual across the replicates. Then, using the effects generated by BayesR and BayesRR-RC, we calculated the predicted genetic value and determined the correlation with the simulated genetic value. We took the marker effect estimates from the MLMA approaches and conducted LD clumping with p -value thresholding using Plink to find the set of markers that gave the highest correlation of the genetic predictor and the simulated genetic value

within the 10,000 UK Biobank individual selected for out-of-sample prediction. We also used the MegaPRS methods implemented in the software LDK running the four different models described above. We compared the correlation of predicted and simulated genetic value across approaches for each of the 400 simulated phenotypes (Supplementary Fig. 2d).

The influence of population structure and relatedness. We then investigated the importance of controlling for multicollinearity for the control of population genetic and data structure effects. In principle, a MLMA approach will control for bias with correlated markers (either local or long-range LD) fitted as random when testing for the effects of a focal SNP. For two markers, X_1 and X_2 in LD correlation ρ_{X_1, X_2} , with $\beta_2 = 0$ we can express the MLMA fixed effect solution as a partial regression coefficient of the phenotype regressed onto the focal SNP after adjusting for X_2 estimated as $\mu_{X_1} = \frac{X_1 y}{X_1^2 X_2 + \lambda}$. Following our derivation above for a shrinkage estimator of a partial regression coefficient the effect size of X_1 is estimated as $\hat{\beta}_{y, X_1 | X_2} = \frac{N}{X_1^2 X_2} \times \rho_{y, X_1} - \frac{\rho_{X_1, X_2} X_2 y}{1 - \rho_{X_1, X_2}^2}$ and in this two-SNP example the bias is accounted for in the term $\frac{\rho_{X_1, X_2} X_2 y}{1 - \rho_{X_1, X_2}^2}$ when the fixed effect is estimated. Multicollinearity acts to increase the σ_G term of λ , reducing the denominator $X_1^2 X_2 + \lambda$ in the estimation of μ_{X_1} , and increasing the variance of the estimates of common markers in high LD, those with the highest average F_{ST} .

We conducted a simulation study using real genomic data from chromosome 22 where 10,000 individuals were selected from two UK Biobank assessment centres (Glasgow and Croydon). First, causal variants were allocated to 5000 high-LD SNPs with effect sizes simulated from a normal distribution with variance proportional to the F_{ST} among the two populations at each SNP. Second, we selected the same high-LD SNPs as the causal variants, but simulated effect sizes to have correlation 0.5 with the allele frequency differences of the SNPs among the two populations, and thus not only is the effect size proportional to the F_{ST} , but there is also directional differentiation (trait increasing loci tend to be those with higher allele frequency in Croydon, trait decreasing alleles have lower frequency in Croydon). For each of these two scenarios, we simulated 50 replicate phenotypes where the phenotypic variance attributable to the causal SNPs is 0.5, there is a phenotypic difference in which Croydon individuals have a phenotype that is 0.5 SD higher than Glasgow individuals (contributing variance 0.05), and residual variance was simulated from a normal with variance 0.45, to give a phenotype with mean of zero and variance of 1. The data were then analysed using a mixed-linear model association (MLMAi implemented in GCTA) and a grouped Bayesian dirac spike and slab models (BayesR implemented in GMRM). In the analysis, we either adjusted the phenotype by the first 20 PCs of the genetic data used in the simulation study, or we did not adjust the phenotype for the PCs, to examine the effects of this common methods of population stratification control. In a two-population scenario the leading eigenvector encapsulates the allele frequency differentiation between the populations and thus the expectation is that this should adjust for these differences when estimating the marker associations. The results are presented in Fig. S5a, where we find that an MLMA approach overestimates the variance attributable to the SNPs under all scenarios, both with and without adjustment for PCs. BayesR returns accurate estimates when the variance of the marker effects is proportional to F_{ST} and underestimates the variance when there is a directional associations, with this underestimation being less severe with PC adjustment.

Finally, we also assess the influence of relatedness on the estimates obtained from a BayesR model using real genomic data from chromosome 21 and 22 (226,662 SNP markers) and 10,000 families randomly selected from the UK Biobank (26,034 individuals). Here, we selected 2000 LD blocks with a single causal SNP per block at random, where an LD block is defined as a group of SNP markers with absolute LD of at least 0.01. We assigned effect sizes to these 2000 selected SNPs, drawing them from a normal distribution with zero mean and variance 0.5/2000. We then multiplied effect sizes by the simulated marker values scaled to zero mean and unit variance to create the genetic values with variance 0.5. In addition to the genetic component, we added a common environment component to simulate effects coming from shared familial environment. We simulated four scenarios where each family was assigned the same common environment effect drawn from a normal distribution with variance 0 (no common environment), 0.1, 0.2, and 0.3. Finally, we added an environmental component simulated from a normal distribution with mean zero and variance 1 minus the genetic variance and minus the common environment variance. We analysed 20 replicates of each of the four scenarios with BayesRR-RC with six MAF-LD groups (terciles of MAF, each split into two groups based on median LD score within each MAF tercile). In Supplementary Fig. 5, we summarise 800 samples of the posterior distribution from 5000 iterations with a thin of five and removing the first 1000 iterations as burn-in. We find that the variance attributable to the SNPs, the regression coefficients and the posterior probability of window variance (PPWV) remain unchanged with relatedness and with increasing family effects.

Localisation and fine-mapping of SNP-phenotype associations. We further validate the use of PPWV in an another simulation study with 500 replicate data sets of 10,000 SNP markers for 5000 individuals for each of two scenarios. In the first scenario, 1000 SNPs are randomly selected to be causal variants and all 10,000 SNP markers are LD independent. In the second, the 1000 causal variants are each in LD

with four other variants with LD = 0.95, with the remaining 5000 variants having zero effect size and LD = 0. For each scenario, we simulate effect sizes as an equally spaced sequence from an effect size of -0.04 SD, to 0.04 SD giving genetic variance of 0.55, and we simulate residual variance from a normal distribution with zero mean and variance 0.45, to give a phenotype with zero mean and unit variance. For the first scenario, we calculate the posterior inclusion probability of each causal SNP. For the second scenario, we calculate the PPWV for each 5-SNP group. Across the 500 replicates of each scenario, we take the mean PPWV and mean PIP for each of the 1000 different effect sizes and compare these in Fig. S6a. Additionally, we grouped SNPs in 50kb regions and selected the number of regions that explain at least 0.1, 0.01 and 0.001% of the variance attributed to all SNP markers in 0.8–100% of the iterations using the simulated data described above for the multiple group enrichment scenario for chromosome 22 in the UK Biobank. We then calculated the false discovery rate (FDR), defined as the proportion of 50 kb regions identified that do not contain a causal variant, at PPWV thresholds ranging from 0.8 to 100%. We compare these in Supplementary Fig. 6b where as we lower the PPWV variance threshold, the number of false discoveries in the model increases but remains at $\leq 5\%$ when the PPWV is $\geq 95\%$. This further demonstrates that our proposed PPWV approach is an appropriate metric of summarising the posterior distribution to identify associated genomic regions, irrespective of the genomic region used.

We also focused on the ability of our approach to fine-map associated regions to identify candidate SNPs and to provide a probabilistic assessment of the most likely associated set of SNP markers. To do this we used our large-scale simulation data and focused on seven focal regions within a blocks of chromosome 22. We allocated effect sizes to the following SNPs: rs131529 with MAF 0.32 which had LD $R^2 \geq 0.15$ with 348 other SNPs, rs2096537 with MAF 0.14 which had LD $R^2 \geq 0.15$ with 295 other SNPs, rs131538 with MAF 0.05 which had LD $R^2 \geq 0.15$ with 82 other SNPs, rs141962840 with MAF 0.007 which had LD $R^2 \geq 0.15$ with 11 other SNPs, rs117873986 with MAF 0.02 which had LD $R^2 \geq 0.15$ with 12 other SNPs, rs9606483 with MAF 0.005 which had LD $R^2 \geq 0.15$ with 1 other SNP, and rs78881648 with MAF 0.009 which had LD $R^2 \geq 0.15$ with 1 other SNP. To these seven SNPs, we assigned the same effect sizes in four different scenarios, either 0.05, 0.025, 0.0125, or 0.01 on the SD scale. On the remainder of chromosomes 19, 20, 21 and 22, we randomly selected 1000 SNPs as causal variants to give a polygenic background, sampling their effects from a normal distribution with zero mean and variance 0.5/1000. We repeated each of the four scenarios 20 times. We selected these regions to compare the performance of BayesRR-RC to the fine-mapping approach SuSiE as outlined in a recent paper²². For BayesRR-RC, we calculate the PPWV of the LD blocks containing the seven focal SNPs, and then prune these blocks based on the LD among the markers in the block (described as “purity” in the SuSiE paper²²) to identify a credible set with LD $R^2 \geq 0.9$. We then count the proportion of times across the simulations that each causal variant was contained with one of the credible sets. For SuSiE, we ran the model from the individual-level data of the whole block of chromosome 22 using the suggested settings and setting $K = 10$. We then calculate the proportion of times that the identified credible sets contained one of the seven causal variants. We present these results in Supplementary Fig. 6c.

UK Biobank data. We restricted our discovery analysis of the UK Biobank to a sample of European-ancestry individuals. To infer ancestry, we used both self-reported ethnic background (UK Biobank data code 21000-0) selecting coding 1 and genetic ethnicity (UK Biobank data code 22006-0) selecting coding 1. We also took the 488,377 genotyped participants and projected them onto the first two genotypic principal components (PC) calculated from 2504 individuals of the 1000 Genomes project with known ancestries. Using the obtained PC loadings, we then assigned each participant to the closest population in the 1000 Genomes data: European, African, East-Asian, South-Asian or Admixed, selecting individuals with PC1 projection < absolute value 4 and PC 2 projection < absolute value 3. This gave a sample size of 456,426 individuals.

To facilitate contrasting the genetic basis of different phenotypes, we then removed closely related individuals as identified in the UK Biobank data release. While the BayesRR model can accommodate relatedness similar to mixed linear models, we wished to simply compare phenotypes at markers that enter the model due to LD with underlying causal variants. Relatedness leads to the addition of markers within the model to capture the phenotypic covariance of closely related individuals, and this will vary across traits in accordance with the genetic and environmental covariance for each phenotype. For these unrelated individuals, we used the imputed autosomal genotype data of the UK Biobank provided as part of the data release. We used the genotype probabilities to hard-call the genotypes for variants with an imputation quality score above 0.3. The hard-call threshold was 0.1, setting the genotypes with probability ≤ 0.9 as missing. From the good quality markers (with missingness less than 5% and p -value for Hardy–Weinberg test larger than 10⁻⁶, as determined in the set of unrelated Europeans) were selected those with minor allele frequency (MAF) > 0.0002 and rs identifier, in the set of European-ancestry participants, providing a data set 9,144,511 SNPs, short indels and large structural variants. From these, we took the overlap with the Estonian Genome centre data to give a final set of 8,430,446 markers. From the UK Biobank European data set, samples were excluded if in the UKB quality control procedures they (i) were identified as extreme heterozygosity or missing genotype outliers; (ii)

had a genetically inferred gender that did not match the self-reported gender; (iii) were identified to have putative sex chromosome aneuploidy; (iv) were excluded from kinship inference. Information on individuals who had withdrawn their consent for their data to be used was also removed. These filters resulted in a data set with 382,466 individuals.

We then selected the recorded measures of BMI (UK Biobank variable identifier 21001-0.0) and height (variable identifier 50-0.0) collected during initial assessment visit (year 2006-2010). BMI and height phenotypes six standard deviations (SD) away from the mean were not included in the analyses. For Type 2 Diabetes (T2D) in UKB, we selected cases very broadly as individuals who have main or secondary diagnosis (UKB fields 41202-0.0–41202-0.379 and 41204-0.0–41204-0.434) of “non-insulin-dependent diabetes mellitus” (ICD 10 code E11) or self-reported non-cancer illness (UKB field 20002-0.0–20002-2.28) “type 2 diabetes” (code 1223). From respondents self-reporting just “diabetes” (code 1220), we selected as cases those who did not self-report “type 1 diabetes” (code 1222) and had no Type 1 Diabetes (ICD code E10) diagnosis. Individuals with self-reported “diabetes” and “type 1 diabetes”/E10 were also left out from controls. We also defined coronary artery disease (CAD) cases broadly as participants with one of the following primary or secondary diagnoses or cause of death: ICD 10 codes I20 to I28; self-reported angina (code 1074) or self-reported heart attack/myocardial infarction (code 1075). Participants with self-reported “heart/cardiac problem” (code 1066) were not included as cases but also excluded from controls. This gave a sample size for each trait of 25,773 T2D cases and 359,730 T2D controls, 39,766 CAD cases and 344,054 CAD controls, 382,402 measures of height and 381,899 measures of BMI.

UK Biobank has approval from the North West Multi-centre Research Ethics Committee (MREC) to obtain and disseminate data and samples from the participants (<http://www.ukbiobank.ac.uk/ethics/>), and these ethical regulations cover the work in this study. Written informed consent was obtained from all participants. Data from this project were held under UK Biobank project ID 35520.

All phenotypes were adjusted for age of attending assessment centre (UKB code 21003-0.0, factor with levels for each age), year of birth (UKB field 34-0.0, factor with levels for each year), UK Biobank recruitment centre (UKB field 54-0.0, factor with levels for each centre), Genotype batch (UKB field 22000, factor with levels for each batch) and final 20 leading principal components of 1.2 million LD clumped markers from the 8,430,446 markers included in the analysis, calculated using flashPCA (see “Code availability” section). The residuals were then converted to z -scores with 0 mean and variance of 1. Similarly as for relatedness, population stratification is also accounted for within the BayesRR model through the addition of a background of marker effects entering the model, however we also wished to account for this in the standard manner by adjusting for the leading 20 PCs of the SNP data to get as close as possible to the inclusion of markers in the model that reflect LD with the causal variants. We note that as with any association model, while we take steps to adjust for known spatial (UKB centre), batch, and ancestry effects, and that the effects of each SNP is estimated jointly (and thus conditionally on the effects of all the other SNPs) environmentally induced covariance between SNP markers and a phenotype is still possible.

We partition SNP markers into seven location annotations using the knownGene table from the UCSC browser data (see “Code availability” section). We preferentially assigned SNPs to coding (exonic) regions first, then in the remaining SNPs, we preferentially assigned them to intronic regions, then to 1 kb upstream regions, then to 1–10 kb regions, then to 10–500 kb regions, then to 500–1 Mb regions. Remaining SNPs were grouped in a category labelled “others” and also included in the model so that variance is partitioned relative to these also. Thus, we assigned SNPs to their closest upstream region, for example if a SNP is 1 kb upstream of gene X, but also 10–500 kb upstream of gene Y and 5 kb downstream for gene Z, then it was assigned to be a 1 kb region SNP. This means that SNPs 10–500 kb and 500 kb–1 Mb upstream are distal from any known nearby genes. We further partition upstream regions to experimentally validated promoters, transcription factor binding sites (TFBS) and enhancers (enh) using the HACER, snp2tfbs databases (see “Code availability” section). All SNP markers assigned to 1 kb regions map to promoters; 1–10 kb SNPs, 10–500 kb SNPs, 500 kb–1 Mb SNPs are split into enh, TFBS and others (un-mapped SNPs) extending the model to 13 annotation groups. Within each of these annotations, we have three minor allele frequency groups ($MAF < 0.01$, $0.01 > MAF > 0.05$, and $MAF > 0.05$), and then each MAF group is further split into two based on median LD score. This gives 78 non-overlapping groups for which our BayesRR-RC model jointly estimates the phenotypic variation attributable to, and the SNP marker effects within, each group. For each of the 78 groups, SNPs were modelled using five mixture groups with variance equal to the phenotypic variance attributable to the group multiplied by constants (mixture 0 = 0, mixture 1 = 0.0001, 2 = 0.001, 3 = 0.01, 4 = 0.1). We conducted a series of convergence diagnostic analyses of the posterior distributions to ensure we obtained estimates from a converged set of four Gibbs chains, each run for 6000 iterations with a thin of five and burn-in of 500 for each trait (Supplementary Figs. 7–10).

We calculate PPWV for LD blocks of the genome, by first calculating the minor allele frequency of each SNP (p) and using $1 - p$ in a Plink clumping procedure to select LD independent ($correlation^2 \leq 0.1$) blocks of SNPs. We then repeat the estimation of the PPWV of 50 kb regions across the genome, then map SNPs to the coding region of genes, and to the closest gene ± 50 kb from the SNP position. These are labelled as located in a coding region, an intron, 1 kb upstream of a gene using our functional annotations. Remaining SNPs are labelled as located in a cis-

region (up to ± 50 kb from a gene, Supplementary Data 6–9). Finally, we mapped SNPs with greater than 50% posterior inclusion probability (PIP) across all four chains labelling them using our seven location annotations (Supplementary Fig. 13). We report SNPs with PIP $> 95\%$ and their corresponding p -value from reported GWAS summary statistics (fastGWA, see “Code availability”) with “body mass index” entry for BMI, “standing height” for HT, “angina/heart attack” for CAD and “diabetes” for T2D (Supplementary Data 10).

We then compared our BayesRR-RC estimates for height, BMI, T2D and CAD to RHEmc¹⁸ which also relies on individual level data. We ran RHEmc with ten independent random vectors and 100 jackknife blocks on the 382,466 individuals and 8,430,446 SNP markers assigned to our 78 non-overlapping groups. SNP heritability estimates, enrichment and standard errors per genetic component are reported in Supplementary Data 3. We intended to include SNP heritability estimates from Bolt-REML¹⁷ in the method comparison but the run time and memory usage exceeded 7 days and 900 GB which is the limiting run-time and memory for our HPC system. Among the summary statistic methods, we ran sLDSC¹⁹ and SumHer⁶. To do so, we created summary statistics containing marginal associations for each of the 8,430,446 markers using plink2²⁸ for height, BMI, T2D and CAD. For sLDSC, we computed univariate LD scores and annotation-specific LD scores for the 78 non-overlapping groups using a window size of 10,000 kb and a subset of 20,000 individuals randomly selected from the full data set. We then partitioned heritability with our annotations and no restriction on MAF. SNP heritability estimates, proportions of heritability, enrichment and standard errors per genetic component are reported in Supplementary Data 4. For SumHer, we computed LDAC weightings and created tagging files separately by chromosomes using the full data set ($M = 8,430,446$ and $N = 382,466$) as reference and a window size of 1000 kb. Because SNPs included in groups *others* and *rare 1Mb tfs* are not present in all chromosomes, tagging files are constructed using 70 non-overlapping annotations only. The remaining SNPs are modelled together in an extra partition. Finally, we merged the tagging files and regressed the summary statistics onto this file assuming the LDAC model. SNP heritability estimates, proportions of heritability, enrichment and standard errors per genetic component are reported in Supplementary Data 5. The proportion of genetic variance estimated genome-wide with RHE-mc, sLDSC, and SumHer are shown in Table 1. We also report the proportion of genetic variance attributed to SNPs located in exons, introns, 1, 1–10 and 10–500 kb regions and the proportion of common SNPs located in exons, introns and 10–500 kb regions computed from the single heritability estimates observed (Table 1).

In addition to plink2²⁸ summary statistics, we also applied Bolt-LMM⁸ and Regenie⁹ to height, BMI, T2D and CAD. In the first step, we pruned SNPs using plink³⁰ with a pairwise r^2 threshold of 0.5 and a window size of 1000 kb, resulting in a subset of 1,362,013 SNPs markers. We restricted the random effects in the mixed model for bolt-LMM and the ridge regression predictors for Regenie to this subset of pruned SNPs. In the second step, all 8,430,446 SNPs from the full genotype data were then tested for association in both methods. Following recommendations, we used the provided hg19 genetic map file and 1000 Genomes LD scores reference for Bolt-LMM and performed the default mixed linear model association test. For Regenie, the 1,362,013 SNP markers are split in blocks of 1000 consecutive SNP markers and ridge regression predictors are computed for a range of five shrinkage parameters within each block. For the association testing, we split the 8,430,446 SNP markers in blocks of 400 consecutive SNP markers and set the Firth correction p -value threshold to 0.01. We then applied an approximate and joint association analysis called GCTA-COJO³¹ to the summary statistics obtained with Bolt-LMM, Regenie and plink2. We ran GCTA-COJO using a subset of 20,000 individuals randomly selected from the 382,466 individuals as reference with a window size of 10,000 kb and a r^2 cutoff value of 0.5 for the LD among the SNPs in the data. Finally, we set a p -value threshold to $5e-8$ to report significant SNPs associated with height, BMI, CAD and T2D in Table 2.

Estonian Genome Centre data. For the Estonian Genome Centre Data, 32,594 individuals were genotyped on Illumina Global Screening (GSA) arrays and we imputed the data set to an Estonian reference, created from the whole genome sequence data of 2244 participants³². From 11,130,313 markers with imputation quality score > 0.3 , we selected SNPs that overlapped with the UK Biobank, resulting in a set of 8,433,421 markers.

We selected height and BMI measures from the Estonian Genome Centre data, in 32,594 individuals genotyped on GSA array and converted them to sex-specific z -scores after applying the same outlier removal procedure as in UKB and adjusting for the age at agreement. Prevalent cases of CAD and T2D in the Estonian Biobank cohort were first identified on the basis of the baseline data collected at recruitment, where the information on prevalent diseases was either retrieved from medical records or self-reported by the participant. The cohort was subsequently linked to the Estonian Health Insurance database that provided additional information on prevalent cases (diagnoses confirmed before the date of recruitment) as well as on incident cases during the follow-up.

All Estonian biobank participants have signed a broad informed consent form and the study was carried out under ethical approval 1.1 12/2856 from the Estonian Committee on Bioethics and Human Research (Estonian Ministry of Social Affairs).

As the UK Biobank marker effects are estimated from traits that were standardised to a z -score prior to analysis, all effect sizes obtained are on the SD

scale. Thus when we create a genomic predictor, for say coding SNPs, by multiplying SNPs mapped to coding regions genotyped in Estonia to the effect sizes obtained in the UK Biobank for each iteration, we obtain a genetic predictor for each iteration, providing a posterior predictive distribution that is also on the SD scale. For each trait, we created 2000 genomic predictors for each individual in the Estonian Biobank, at each of the 13 annotation groups, by selecting effect size estimates obtained every tenth iteration from the last 3000 iterations of each of the four Gibbs chains and combining them together in a single posterior. We calculated prediction accuracy as the proportion of phenotypic variation explained by the genomic predictor, and area under the receiver operator curve (AUC) for T2D and CAD using each individual's mean genetic predictor. For each of the 13 annotation groups, we calculated the partial correlation of the genetic predictor of each of the 2000 iterations and the phenotype. We then used this to estimate the independent proportional contribution of each group to the total prediction accuracy, providing a metric of replication for our UK Biobank enrichment results.

For height and BMI, we determined the probability that each Estonian individual's predictor accurately reflected their phenotypic value. To do this, we calculated the proportion of posterior samples with $|\hat{g} - y|$ of less than 1 for each individual, which gives a measure of the degree to which each posterior predictive distribution overlaps with the phenotype within ± 1 SD.

For T2D and CAD, we extended the PCF metric, typically defined as the proportion of cases with larger estimated risk than the top p^{th} percentile of the distribution of genetic risk in the general population. We calculated the proportion of posterior samples for each individual with values in the top 25% of the distribution of genomic predictors for each trait. Thus for each individual, we calculate the probability that the posterior predictive distribution is in the top 25% of the distribution of genetic risk in the general population.

As a comparison, we also estimated a boltLMM prediction model using MegaPRS²¹ as recommended by the authors and as shown to have the best prediction performance out of the MegaPRS approaches in our simulation study. We clumped SNPs with r^2 threshold of 0.5 resulting in 1,508,624 SNP markers to be included in the analysis and randomly selected 20,000 individuals to compute the LD weights. We then computed the tagging file using the same data set as reference and the 64 BLD-LDAK annotations. Here, weights are models as an extra annotation and we save the heritability matrix. We then regress the plink2²⁸ summary statistics for height, BMI, CAD and T2D onto the tagging file, saving the per-predictor heritabilities. We then created four reference panels with the same 1,508,624 SNP markers but randomly selecting different 5000 related individuals from the UK Biobank and we used these to: (i) calculate predictor-predictor correlations with a window size of 3000 kb to estimate the LD structure; (ii) compute pseudo summaries from the plink2 summary statistics including ambiguous alleles, which creates pseudo training and test summary statistics to be used in the construction of the prediction model; (iii) estimate effect sizes specifying a Bolt-LMM model for height, BMI, CAD and T2D, using the predictor-predictor correlations, the per-predictor heritabilities, the plink2 summary statistics and training pseudo summary statistics, whilst including ambiguous allele and specifying a 1000 kb window; (iv) test prior distributions to determine the most accurate model and obtain the best effect sizes. These steps resulted in 1,397,514 predictors for height, 1,471,586 for BMI, 1,397,514 for CAD and 1,389,364 for T2D and we ensured that at no point was the Estonian genome centre data used, nor was any overlapping individuals in the UK Biobank subsets used to train the models and the data used to generate the summary statistics. Finally, we then calculated genomic predictors for each individual in the Estonian Biobank using the best effect sizes. We report the squared correlations between the genomic predictor and phenotypes.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

This project uses UK Biobank data under project 35520. The Estonian Genome Centre data are protected and are not available due to data privacy laws. The Estonian Genome Centre data can be made available under restricted access upon request from the cohort author R.M. with appropriate research agreements. Summaries of all posterior distributions generated in this study are provided in Supplementary Data tables. Full posterior distributions of the SNP marker effects sizes and estimated variance components for each trait are deposited on Dryad with <https://doi.org/10.5061/dryad.sqv9s4n51>.

Code availability

Our BayesRR-RC model is implemented within the software GMRM, with full open source code available at: <https://github.com/medical-genomics-group/gmrmm>. UCSC Table Browser <https://genome.ucsc.edu/cgi-bin/hgTables>. flashPCA <https://github.com/gabraham/flashpca>. Plink1.90 <https://www.cog-genomics.org/plink2/>. GCTA <https://cns.genomics.com/content/software>. HACER database <http://bioinfo.vanderbilt.edu/AE/HACER/>. snp2tfs database <https://ccg.epfl.ch/snp2tfs/>. fastGWA database <http://fastgwa.info/ukbimp/phenotypes/>. Computing environment <https://www.epfl.ch/research/facilities/scitas/hardware/helvetios/>.

Received: 10 February 2021; Accepted: 5 November 2021;
Published online: 30 November 2021

References

- Evans, L. M. et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat. Genet.* **50**, 737–745 (2018).
- Speed, D. et al. Reevaluation of snp heritability in complex human traits. *Nat. Genet.* **49**, 986 (2017).
- Speed, D., Holmes, J. & Balding, D. J. Evaluating and improving heritability models using summary statistics. *Nat. Genet.* **52**, 458–462 (2020).
- Hou, K. et al. Accurate estimation of snp-heritability from biobank-scale data irrespective of genetic architecture. *Nat. Genet.* **51**, 1244–1251 (2019).
- Gazal, S., Marquez-Luna, C., Finucane, H. K. & Price, A. L. Reconciling s-Ldsc and lDak functional enrichment estimates. *Nat. Genet.* **51**, 1202–1204 (2019).
- Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* **51**, 277–284 (2019).
- Jiang, L. et al. A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.* **51**, 1749–1755 (2019).
- Loh, P.-R. et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284 (2015).
- Mbatchou, J. et al. *Computationally Efficient Whole Genome Regression for Quantitative and Binary Traits* (Nature Publishing Group, 2020).
- Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
- Erbe, M. et al. 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 (2012).
- Moser, G. et al. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genet.* **11**, 1–22 (2015).
- Banos, D. T. et al. Bayesian reassessment of the epigenetic architecture of complex traits. *Nat. Commun.* **11**, 1–14 (2020).
- George, E. I. & McCulloch, R. E. Variable selection via gibbs sampling. *J. Am. Stat. Assoc.* **88**, 881–889 (1993).
- Malsiner-Walli, G. & Wagner, H. Comparing spike and slab priors for bayesian variable selection. *Austrian J. Stat.* **40**, 241–264 (2016).
- Castillo, I. et al. Bayesian linear regression with sparse priors. *Ann. Stat.* **43**, 1986–2018 (2015).
- Loh, P.-R. et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385 (2015).
- Pazokitoroudi, A. et al. Efficient variance components analysis across millions of genomes. *Nat Commun* **11**, 4020 (2020).
- Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
- Fernando, R., Toosi, A., Wolc, A., Garrick, D. & Dekkers, J. Application of whole-genome prediction methods for genome-wide association studies: a bayesian approach. *J. Agric. Biol. Environ. Stat.* **22**, 172–193 (2017).
- Zhang, Q., Privé, F., Vilhjálmsson, B. et al. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat Commun* **12**, 4192 (2021).
- Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc.* **82**, 1273–1300 (2020).
- Lloyd-Jones, L. R. et al. Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019).
- Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: From polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
- MacLeod, I. M. et al. Exploiting biological priors and sequence variants enhances qtl discovery and genomic prediction of complex traits. *BMC Genomics* **17**, 144 (2016).
- Brøndum, R. F. et al. Genome position specific priors for genomic prediction. *BMC Genom.* **13**, 543 (2012).
- Zhao, P. & Yu, B. On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–2563 (2006).
- Chang, C. C. et al. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742–015 (2015).
- Daetwyler, H. D., Villanueva, B. & Woolliams, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE* **3**, 1–8 (2008).
- Purcell, S. et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

31. Yang, J. et al. Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
32. Tasa, T. et al. Genetic variation in the estonian population: pharmacogenomics study of adverse drug effects using electronic health records. *Eur. J. Hum. Genet.* **27**, 442–454 (2019).

Acknowledgements

This project was funded by an SNSF Eccellenza Grant to MRR (PCEGP3-181181), and by core funding from the Institute of Science and Technology Austria. We would like to thank the participants of the cohort studies, and the Ecole Polytechnique Federal Lausanne (EPFL) SCITAS for their excellent compute resources, their generosity with their time and the kindness of their support. P.M.V. acknowledges funding from the Australian National Health and Medical Research Council (1113400) and the Australian Research Council (FL180100072). L.R. acknowledges funding from the Kjell & Märta Beijer Foundation (Stockholm, Sweden). We also would like to acknowledge Simone Rubinacci, Oliver Delanau, Alexander Terenin, Eleonora Porcu, and Mike Goddard for their useful comments and suggestions.

Author contributions

M.R.R. conceived and designed the study. M.P., D.T.B. and A.K. contributed to the study design. M.P. and M.R.R. conducted the experiments and analyses with input from D.T.B., A.K., S.E.O., A.H., J.S., P.M.V., R.M. and L.R. M.R.R., D.T.B., S.E.O. and L.R. derived the equations and the algorithm. EJO and DTB developed the software, with contributions from M.R.R., M.P., S.E.O., A.K. and G.M. M.R.R., M.P. and DTB wrote the paper. RM and ZK provided study oversight and contributed data to the analysis. All authors approved the final manuscript prior to submission.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-27258-9>.

Correspondence and requests for materials should be addressed to Matthew R. Robinson.

Peer review information *Nature Communications* thanks Luke O'Connor and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021

Supplementary Information

Probabilistic inference of the genetic architecture underlying functional enrichment of complex traits

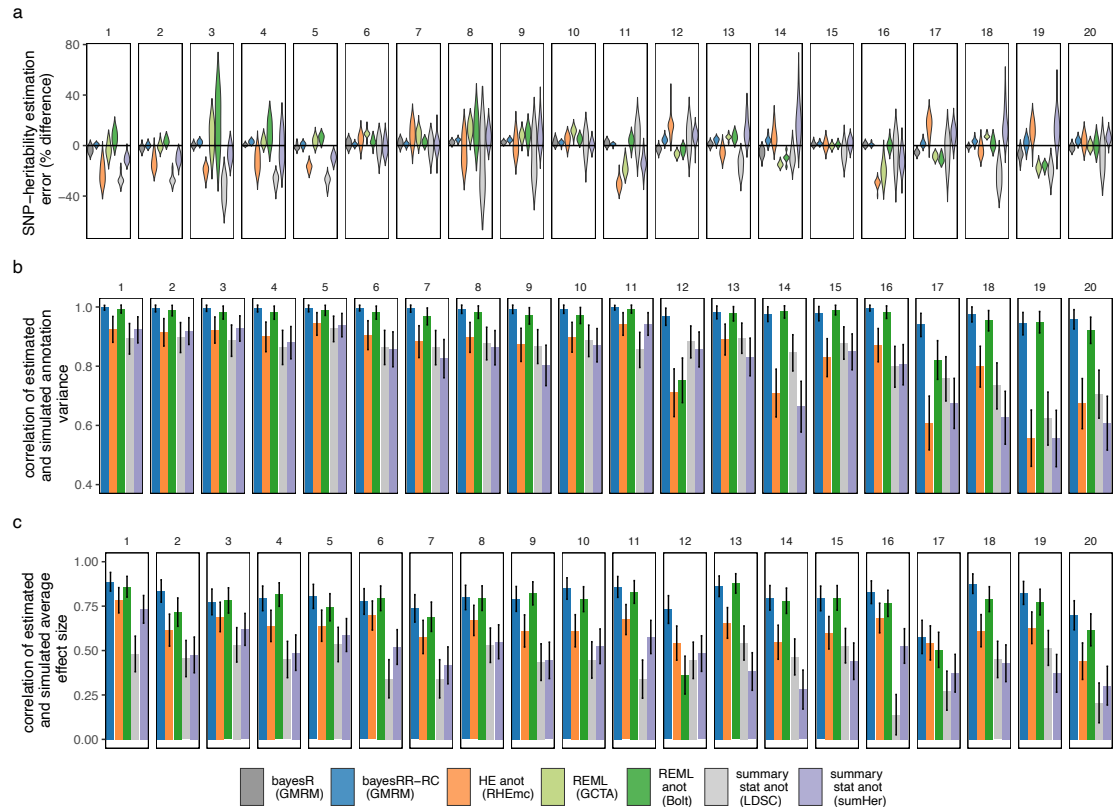
Marion Patxot, Daniel Trejo Banos, Athanasios Kousathanas, Etienne J. Orliac, Sven E. Ojavee, Gerhard Moser, Alexander Holloway, Julia Sidorenko, Zoltan Kutalik, Reedik Mägi, Peter M. Visscher, Lars Rönnegård, Matthew R. Robinson

Supplementary Tables

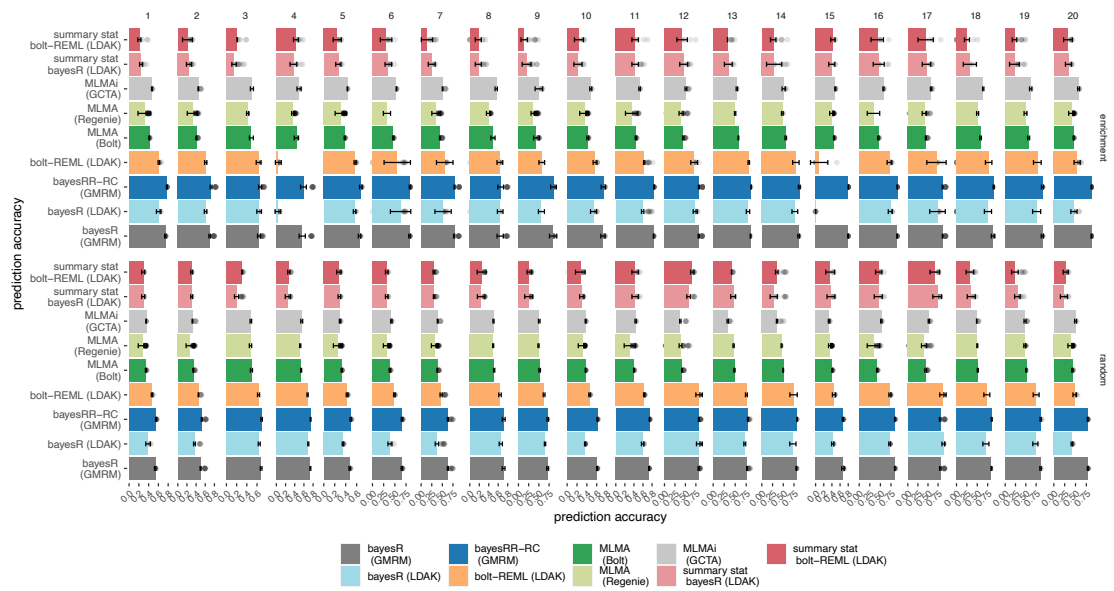
generative model	causal variant allocation	causal variants	effect size (b), LD (w), MAF (p) relationship
1	highest MAF per LD block	10,000	$b \propto N(0, w^{-0.25} [p(1-p)]^{-0.25})$
2	highest MAF per LD block	10,000	$b \propto N(0, w^{0.25} [p(1-p)]^{-0.25})$
3	highest MAF per LD block	10,000	$b \propto N(0, w^{-0.25} [p(1-p)]^{0.75})$
4	highest MAF per LD block	10,000	$b \propto N(0, w^{0.25} [p(1-p)]^{0.75})$
5	highest MAF per LD block	10,000	$b \propto N(0, w^0 [p(1-p)]^0)$
6	highest MAF per LD block	5,000	$b \propto N(0, w^{-0.25} [p(1-p)]^{-0.25})$
7	highest MAF per LD block	5,000	$b \propto N(0, w^{0.25} [p(1-p)]^{-0.25})$
8	highest MAF per LD block	5,000	$b \propto N(0, w^{-0.25} [p(1-p)]^{0.75})$
9	highest MAF per LD block	5,000	$b \propto N(0, w^{0.25} [p(1-p)]^{0.75})$
10	highest MAF per LD block	5,000	$b \propto N(0, w^0 [p(1-p)]^0)$
11	random	10,000	$b \propto N(0, w^{-0.25} [p(1-p)]^{-0.25})$
12	random	10,000	$b \propto N(0, w^{0.25} [p(1-p)]^{-0.25})$
13	random	10,000	$b \propto N(0, w^{-0.25} [p(1-p)]^{0.75})$
14	random	10,000	$b \propto N(0, w^{0.25} [p(1-p)]^{0.75})$
15	random	10,000	$b \propto N(0, w^0 [p(1-p)]^0)$
16	random	5,000	$b \propto N(0, w^{-0.25} [p(1-p)]^{-0.25})$
17	random	5,000	$b \propto N(0, w^{0.25} [p(1-p)]^{-0.25})$
18	random	5,000	$b \propto N(0, w^{-0.25} [p(1-p)]^{0.75})$
19	random	5,000	$b \propto N(0, w^{0.25} [p(1-p)]^{0.75})$
20	random	5,000	$b \propto N(0, w^0 [p(1-p)]^0)$

Supplementary Table 1. The generative genetic models used in the simulation study. Imputed SNP marker data from chromosomes 19, 20, 21 and 22 of 40,000 randomly selected UK Biobank participants were selected, giving 596,741 markers in total. Marker effects were simulated according to the 20 generative models in two ways: (i) a single distribution of marker effects, and (ii) 13 distributions of marker effects for 13 different genomic annotation groups with different proportions of SNP heritability (h_{SNP}^2) explained for exonic variants ($h_{SNP}^2 = 0.1$), intronic variants ($h_{SNP}^2 = 0.2$), 1kb promotor variants ($h_{SNP}^2 = 0.05$), 1-10kb enhancer variants (0.025), 1-10kb transcription factor binding sites ($h_{SNP}^2 = 0.025$), 1-10kb other variants ($h_{SNP}^2 = 0$), 10-500kb enhancers ($h_{SNP}^2 = 0.05$), 10-500kb transcription factor binding sites ($h_{SNP}^2 = 0.05$), 10-500kb other variants ($h_{SNP}^2 = 0$), 500kb-1Mb enhancers ($h_{SNP}^2 = 0.05$), 500kb-1Mb transcription factor binding sites ($h_{SNP}^2 = 0.05$), 500kb-1Mb other variants ($h_{SNP}^2 = 0$), and other non-annotated SNPs ($h_{SNP}^2 = 0$). 10 simulation replicates were created for both (i) and (ii) giving a total set of 400 simulated phenotypes.

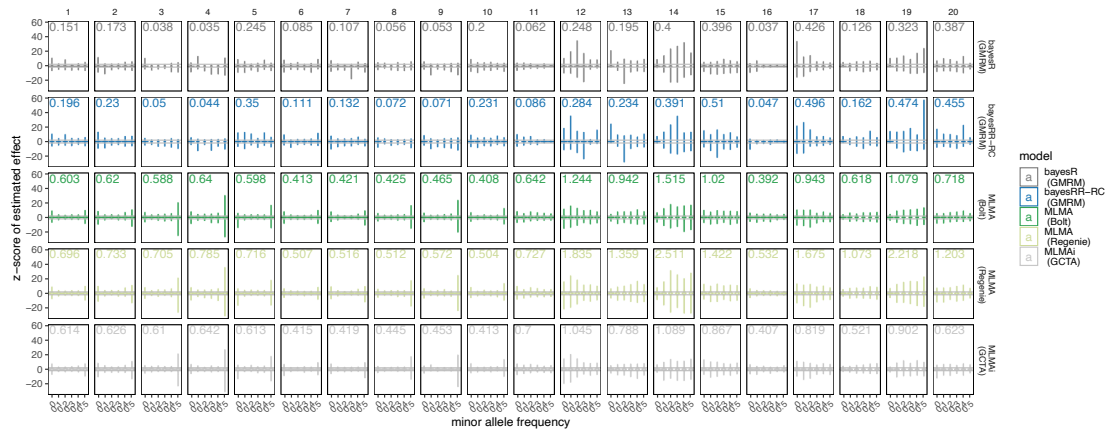
Supplementary Figures



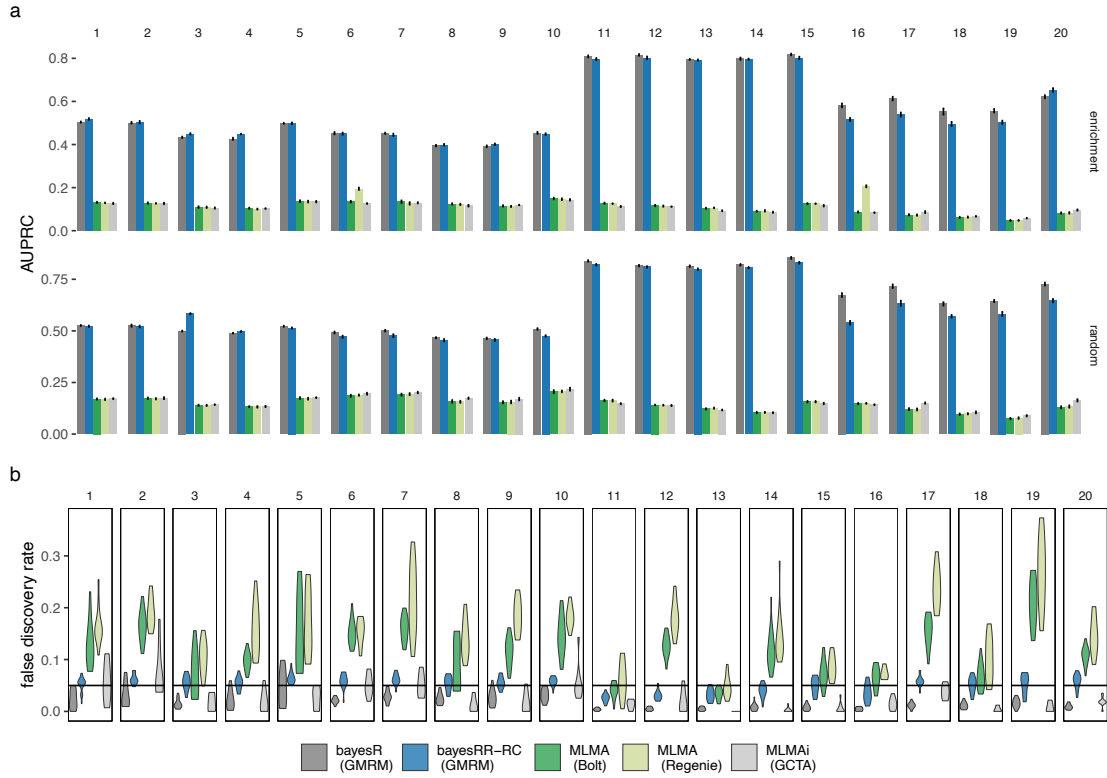
Supplementary Figure 1. Simulation study of the variance component estimation performance of BayesRR-RC implemented in GMRM. (a) Violin-plot of the genome-wide SNP-heritability estimates as a percentage difference from the simulated value for 40 replicates, for each of 20 different generative genetic models described in Table S1. For each generative genetic model we compare seven different statistical models: a mixture of regression model with a single global variance component known as "bayesR" implemented in our GMRM software (bayesR GMRM), the mixture of regression model with multiple group-specific variance components described in this work (bayesRR-RC GMRM), Haseman-Elston regression with annotation-specific relationship matrices implemented in the RHEmc software (HE anot RHEmc), a single component REML model implemented in the software GCTA (REML GCTA), a multiple group-specific variance component REML model implemented in the software bolt (REML anot Bolt), and two annotation summary statistic models implemented in the software LDSC and sumHer. (b) The correlation of the estimated genetic variance for each of 13 genetic annotation groups and the simulated genetic variance across the 40 replicates, for each of the five statistical approaches which enable annotation-specific estimation. (c) Bar-plots of the correlation of the estimated and simulated average effect size of each annotation across simulation replicates. Error bars give the SD.



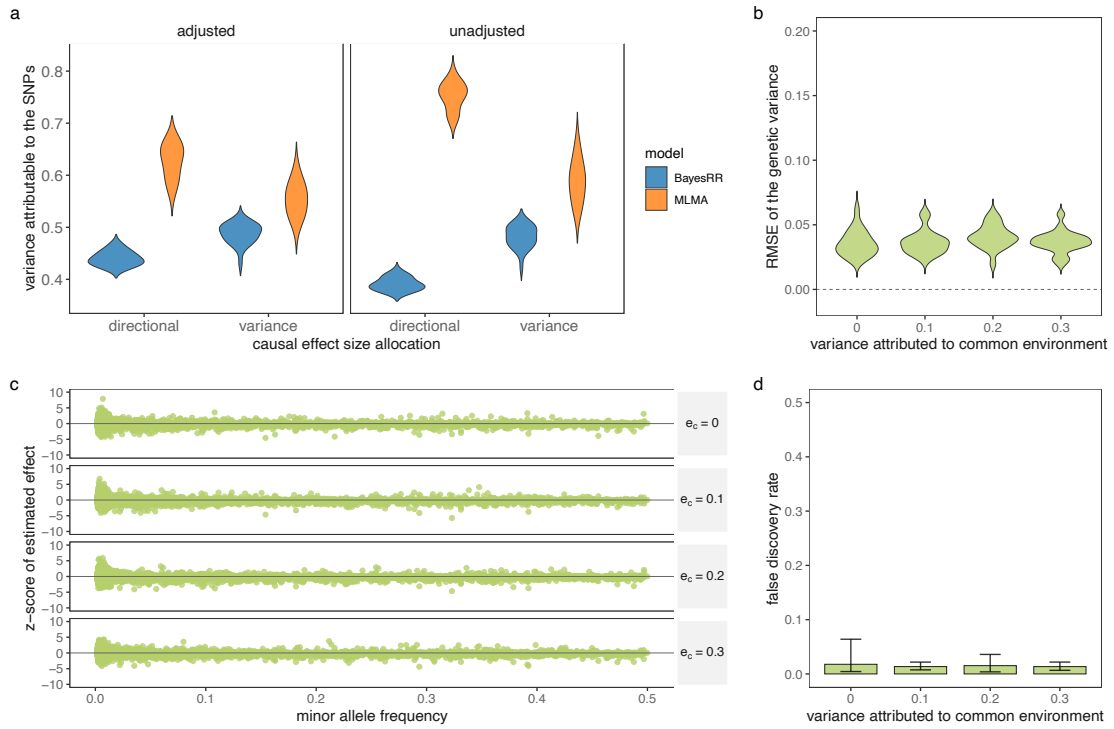
Supplementary Figure 2. Simulation study of the prediction performance of BayesRR-RC implemented in GMRM. Average prediction accuracy in an independent sample, defined as the squared correlation of the predicted and simulated genetic value, with error bars giving the SD. For each of the 20 different generative genetic models described in Supplementary Table 1, we compare the prediction accuracy obtained in a testing set of 10,000 unrelated individuals from the UK Biobank, selected at random and unrelated to the training data. We predicted simulated phenotypes using SNP marker effect sizes obtained from nine different statistical methods: bayesR implemented in our GMRM software (bayesR GMRM); the mixture of regression model with multiple group-specific variance components described in this work (bayesRR-RC GMRM); three frequentist mixed-linear association models (MLMA) where the genetic marker tested for association is removed from the relationship matrix (implemented in software Bolt and Regenie), or fitted both as fixed and random (MLMAi implemented in the software GCTA); and four MegaPRS models using genomic annotation SNP variance estimates from SumHer and implemented in the software LDAK: (i) an individual-level bayesR model (bayesR LDAK), (ii) an individual-level boltREML model (bolt-REML LDAK), (iii) a summary statistic bayesR model (summary stat bayesR LDAK) and (iv) a summary statistic boltREML model (summary stat bolt-REML).



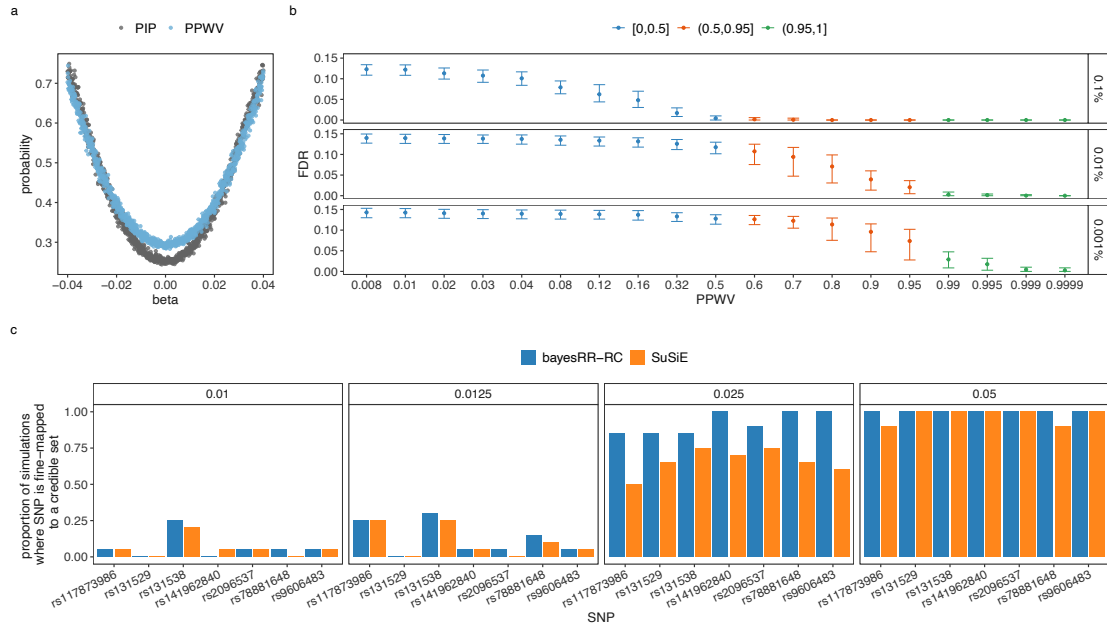
Supplementary Figure 3. Simulation study of the effect size estimation of BayesRR-RC implemented in GMRM. For each of the 20 different generative genetic models described in Supplementary Table 1, we compare model performance of our approach (bayesRR-RC GMRM) to bayesR implemented in our GMRM software (bayesR GMRM) and frequentist mixed-linear association models (MLMA) where the genetic marker tested for association is removed from the relationship matrix (implemented in software Bolt and Regenie), or fitted both as fixed and random (MLMAi implemented in the software GCTA). For bayesR (GMRM) and bayesRR-RC (GMRM), we summed the squared regression coefficient estimates of all SNPs in LD with each causal variant (markers in LD $R^2 \geq 0.1$ within 1MB), took the posterior mean, and calculated the z-score from the simulated value. For the MLMA approaches, we calculated the z-score of the causal marker estimate from the simulated value. Violin plots for groups of minor allele frequency of the causal variant are shown, with values giving the variance in each facet.



Supplementary Figure 4. Simulation study of the effect size localization of BayesRR-RC implemented in GMRM. (a) For each of the 20 different generative genetic models described in Table S1, we compare the area-under the precision-recall curve (AUPRC) for bayesRR-RC (described in this work and implemented in GMRM), bayesR (implemented in GMRM) and mixed-linear association models (MLMA). For Bayesian methods bayesR (GMRM) and bayesRR-RC (GMRM), we use our PPWV metric (see Methods), with true positives defined as LD blocks that contain a causal variant and false positives defined as LD blocks that did not contain a causal variant. For MLMA methods implemented in GCTA (MLMAi GCTA), Bolt (MLMA Bolt) and Regenie (MLMA Regenie), we LD-clumped the results ($LD R^2 \geq 0.01$) using the p-value of the chi-squared statistics. Markers in $R^2 \geq 0.01$ with simulated causal variants were defined as true positives and those not in $LD R^2 \geq 0.01$ as false positives. (b) False discovery rate (FDR), with the line giving the 5% threshold. For the MLMA methods, FDR was calculated as the proportion of LD independent SNPs with p-value $\leq 5 \times 10^{-8}$ that were not in $LD R^2 \geq 0.01$ with causal variants. For the Bayesian methods, we defined FDR as the proportion of LD blocks with posterior probability of window variance (PPWV), of $\geq 95\%$ at 0.001% variance threshold that did not contain a causal variant.

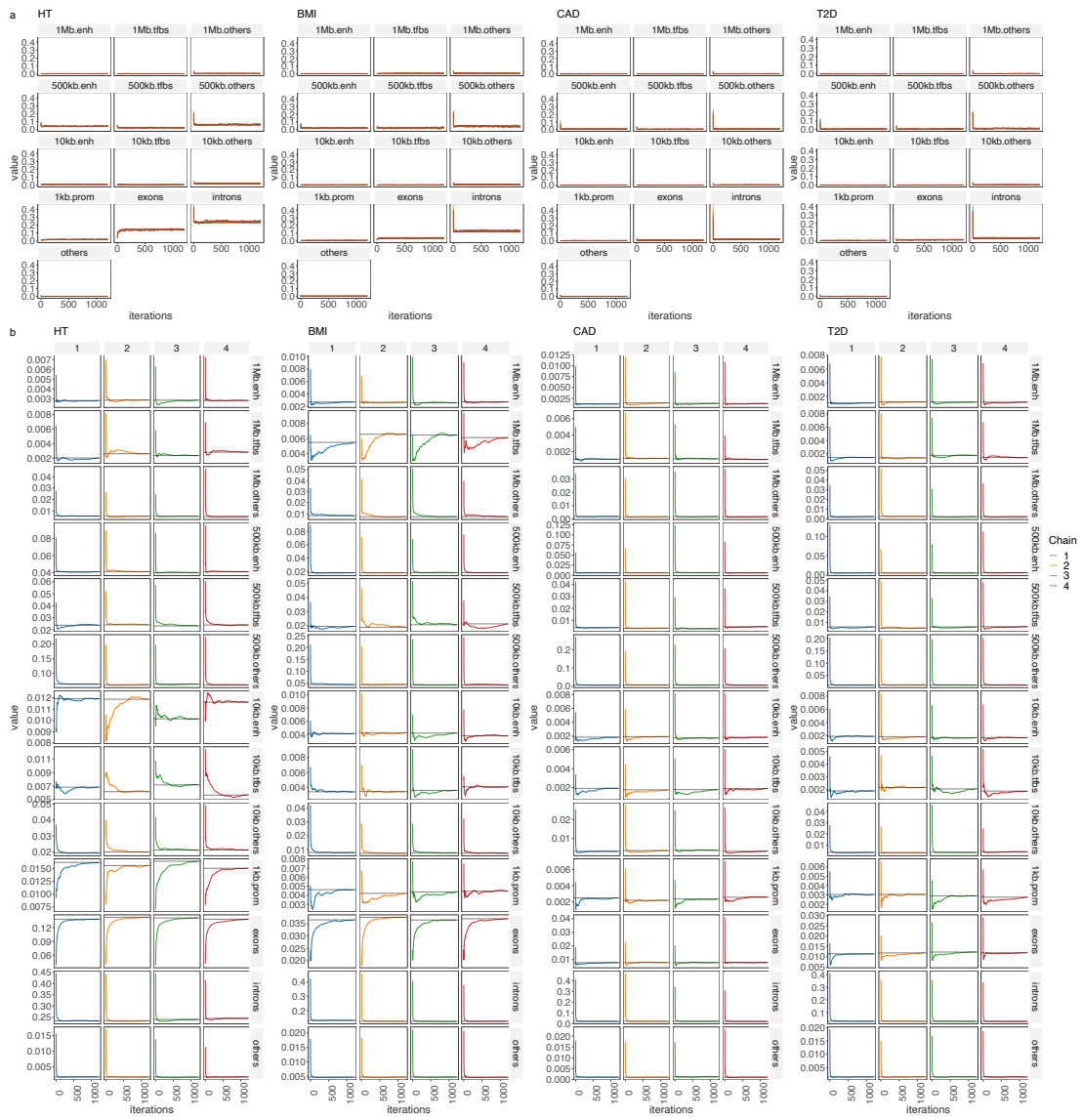


Supplementary Figure 5. Exploring effects of population stratification and relatedness among samples. (a) Simulation study using real genomic data from chromosome 22 where 10,000 individuals were selected from 2 UK Biobank assessment centres (Glasgow and Croydon). First, causal variants were allocated to 5000 high-LD SNPs with effect sizes simulated from a normal distribution with variance proportional to the F_{ST} among the two populations at each SNP (labelled 'variance', see Methods). Second, we selected the same high-LD SNPs as the causal variants, but simulated effect sizes to have correlation 0.5 with the allele frequency differences of the SNPs among the two populations, and thus not only is the effect size proportional to the F_{ST} , but there is also directional differentiation (trait increasing loci tend to be those with higher allele frequency in Croydon, trait decreasing alleles have lower frequency in Croydon). For each of these two scenarios, we simulated 50 replicate phenotypes where the phenotypic variance attributable to the causal SNPs is 0.5, there is a phenotypic difference where Croydon individuals have a phenotype that is on average 0.5 SD higher than Glasgow individuals (contributing variance 0.05), and residual variance was simulated from a normal with variance 0.45, to give a phenotype with mean of zero and variance of 1. The distribution across simulations of the estimated phenotypic variance attributable to the SNP markers is shown for each of the two causal effect size allocation scenarios when the data was analysed using a mixed-linear model association (MLMA, distribution of the point estimates) and a grouped Bayesian dirac spike and slab models (BayesRR, distribution of the posterior means). In the analysis, we either adjusted the phenotype by the first 20 PCs of the genetic data used in the simulation study ("adjusted") or we did not adjust the phenotype for the PCs ("unadjusted"). (b), (c) and (d) show BayesRR-RC simulation results using real genomic data from chromosome 21 and 22 and 10,000 families randomly selected from the UK Biobank. We simulated 20 replicates where we selected 2000 LD blocks at random, with an LD block defined as a group of SNP markers with squared LD correlation of at least 0.15. We assigned a causal SNP per LD block and for each replicate, we simulated 4 phenotypes increasing the variance attributed to family effects from 0 (no common environment) to 0.3 (see Methods). (b) Violin-plot of the root mean square error (RMSE) of the SNP-heritability estimates across simulation replicates. (c) For each LD block of each simulation replicate, we summed the squared regression coefficient estimates of all SNPs in the block and took the posterior mean. We then calculated the z-score of the LD block and plotted it against the minor allele frequency of the causal variant of the block. (d) Shows mean and 95% credible intervals of the false discovery rate defined as the posterior probability of window variance (PPWV), of $\geq 95\%$ at 0.001% variance threshold that did not contain a causal variant.

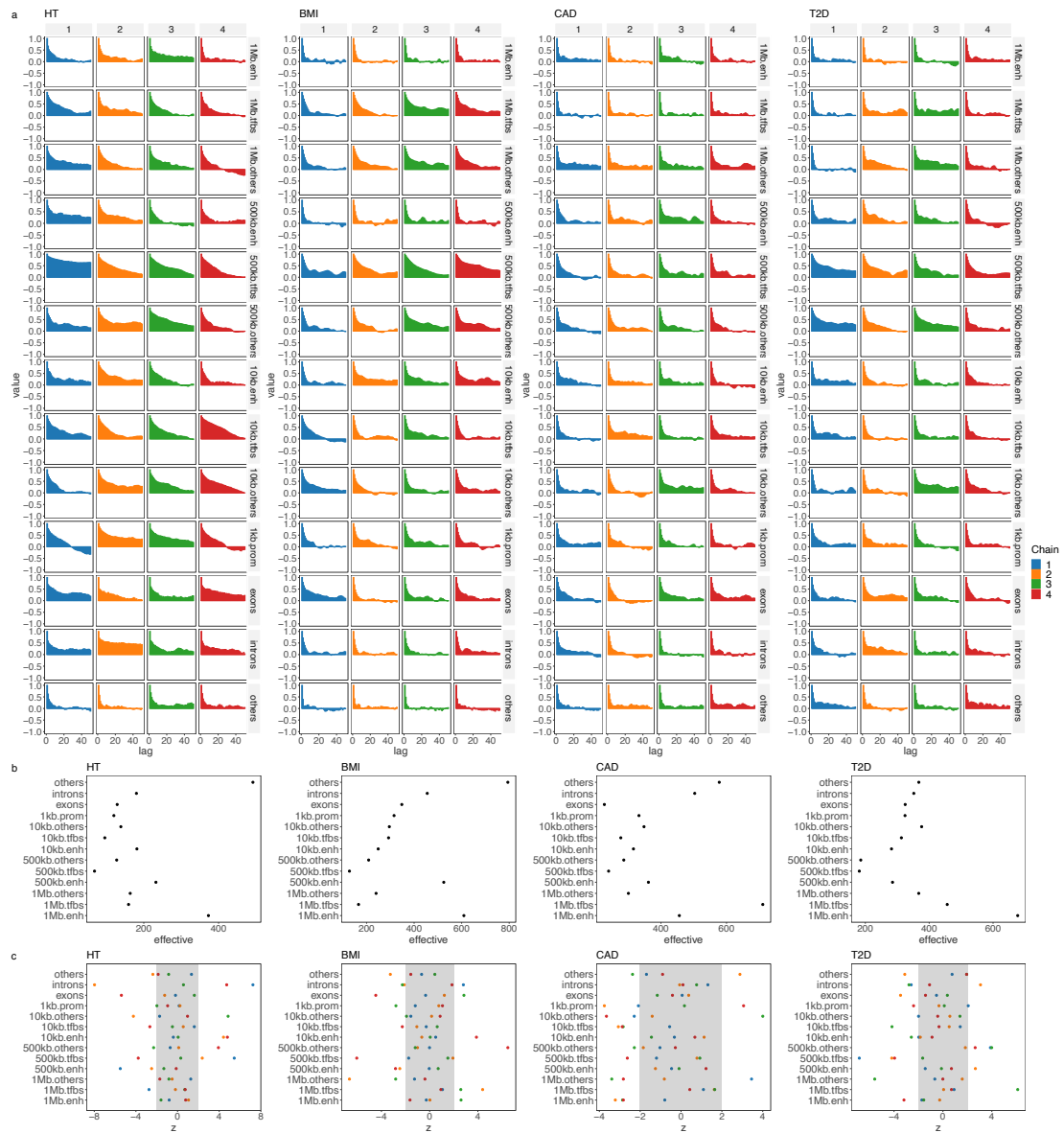


Supplementary Figure 6. Posterior inclusion probability (PIP) and posterior probability of window variance (PPWV).

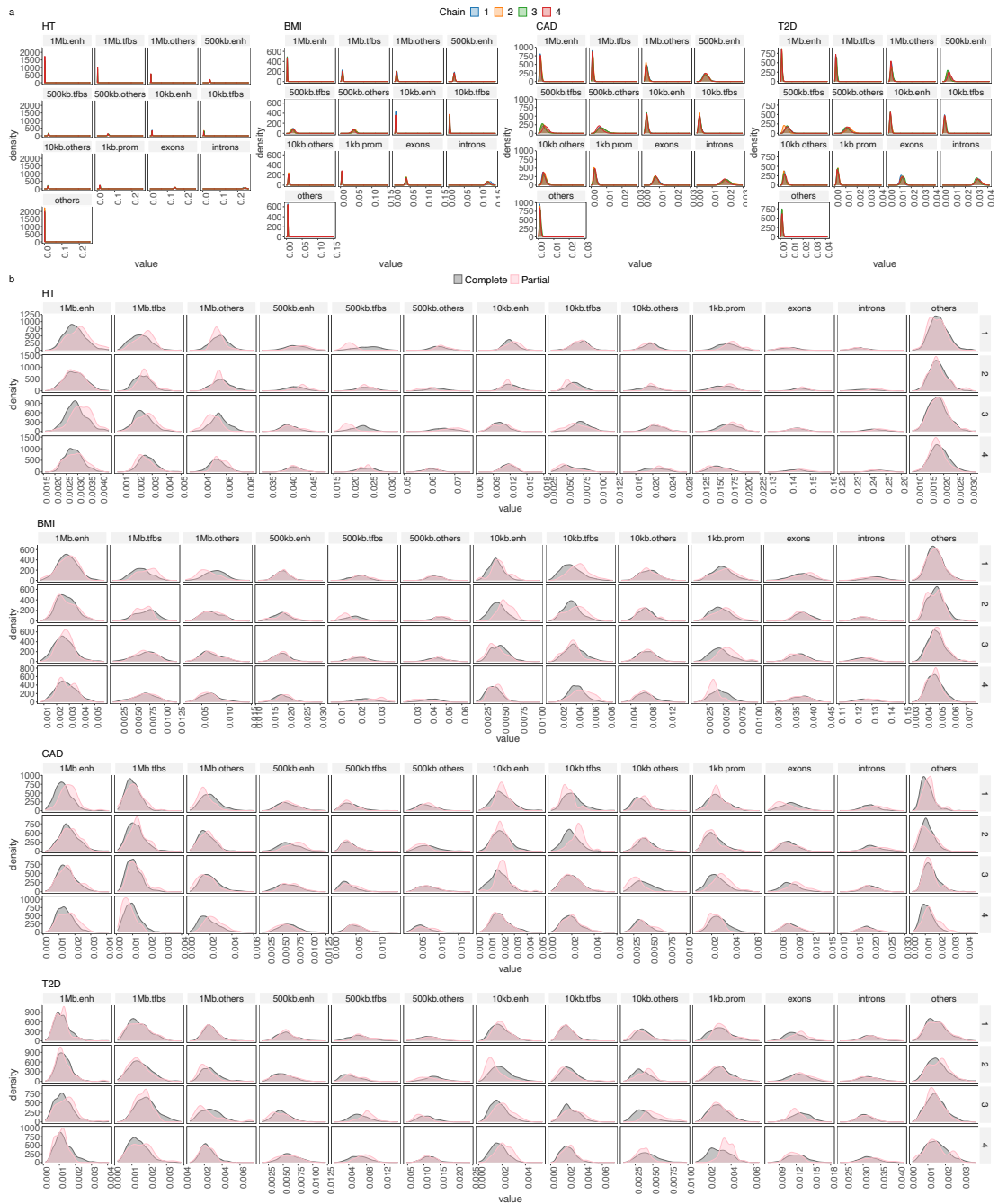
(a) We validate the use of PPWV in simulation study, first simulating 500 replicate data sets of 10,000 SNP markers for 5,000 individuals for each of two scenarios. In the first scenario, 1000 SNPs are randomly selected to be causal variants and all 10,000 SNP markers are LD independent. In the second, the 1000 causal variants are each in LD with four other variants with $LD = 0.95$, with the remaining 5000 variants having zero effect size and $LD = 0$. For each scenario, we simulate effect sizes as an equally spaced sequence from an effect size of -0.04 SD, to 0.04 SD giving genetic variance of 0.55 , and we simulate residual variance from a normal distribution with zero mean and variance 0.45 , to give a phenotype with zero mean and unit variance. For the first scenario, we calculate the posterior inclusion probability of each causal SNP. For the second scenario, we calculate the PPWV for each 5-SNP group. Across the 500 replicates, we take the mean PIP for each SNP of the 1000 different effect sizes for the first scenario and the mean PPWV of each of the 1000 5-SNP windows for the second scenario, and these are the points on the figure. (b) Shows the mean and 95% credible interval of the false discovery rate (FDR), defined as the proportion of regions identified that do not contain a causal variant, at PPWV thresholds ranging from 0.8% to 100%. Here, we grouped SNPs in 50kb regions and selected the number of regions that explain at least 0.1%, 0.01% and 0.001% of the variance attributed to all SNP markers in 0.8% to 100% of the iterations using simulated data for chromosome 22 in the UK Biobank (see Methods). We compare the FDR at these different PPWV thresholds and as we lower the PPWV variance, the number of false discoveries in the model increases, but remains at $\leq 5\%$ at $PPWV \geq 95\%$. (c) A comparison of BayesRR-RC and SuSiE where we assigned effect sizes of either 0.05, 0.025, 0.0125, or 0.01 on the SD scale to seven SNPs. For BayesRR-RC, we calculate the PPWV of the LD blocks containing the seven focal SNPs, and then prune these blocks based on the LD among the markers in the block to identify a credible set with $LD R^2 \geq 0.9$. We then count the proportion of times across 20 simulation replicates that each causal variant was contained with one of the credible sets. For SuSiE, we calculate the proportion of times that the credible sets identified contained one of the seven causal variants.



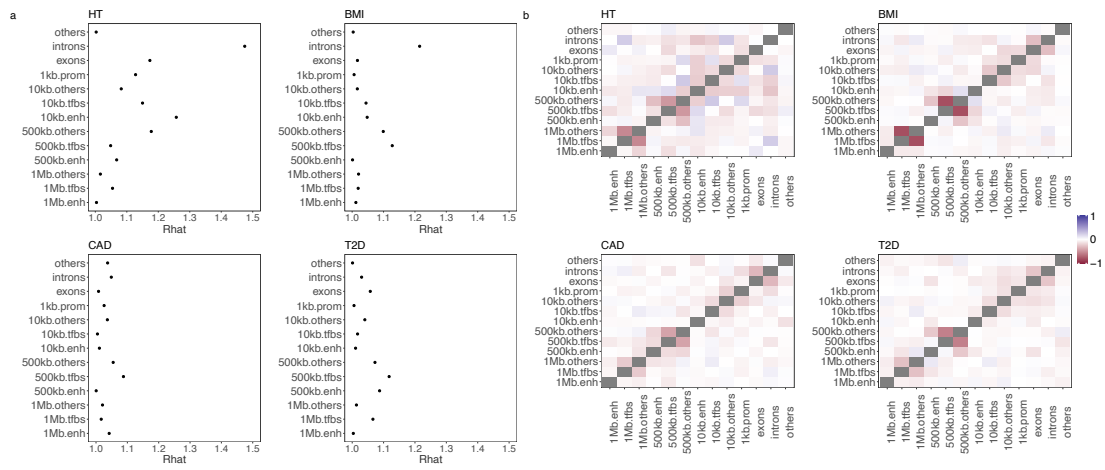
Supplementary Figure 7. Convergence diagnostics of model chains for UK Biobank analysis. (a) Traceplot of the phenotypic variance attributable to SNP markers for each trait across functional annotation of exonic regions, intronic regions, promoters (prom) 1kb upstream of coding regions, enhancers (enh) 1kb to 10kb upstream of coding regions, transcription factor binding sites (tfbs) 1kb to 10kb upstream of coding regions, other snps 1kb to 10kb upstream of coding regions, enh 10kb to 500kb upstream, tfbs 10kb to 500kb upstream, other snps 10kb to 500kb upstream, enh 500kb to 1Mb upstream, tfbs 500kb to 1Mb upstream, other snps 500kb to 1Mb upstream and SNP markers elsewhere in the genome (other), with colours representing the different chains. (b) A time series of the running mean of each chain, for each annotation group and each trait showing all chains approach the same mean value for each parameter.



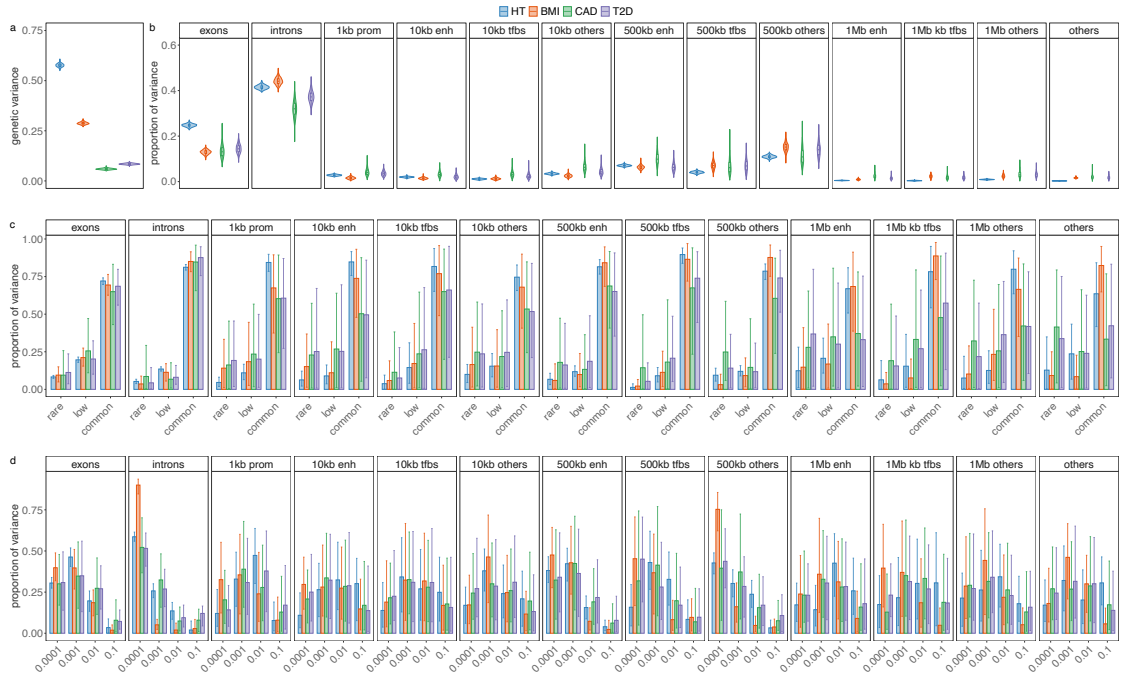
Supplementary Figure 8. Convergence diagnostics of model chains for UK Biobank analysis.(a) Lagged autocorrelation plot of each chain, for each annotation group and each trait and (b) Effective number of uncorrelated samples obtained for each annotation group and each trait. As phenotypic variance is being partitioned it is not expected that posterior estimates obtained are entirely uncorrelated. (c) Geweke z-score statistic comparing the initial part of the chain to the final part, for each annotation group and each trait.

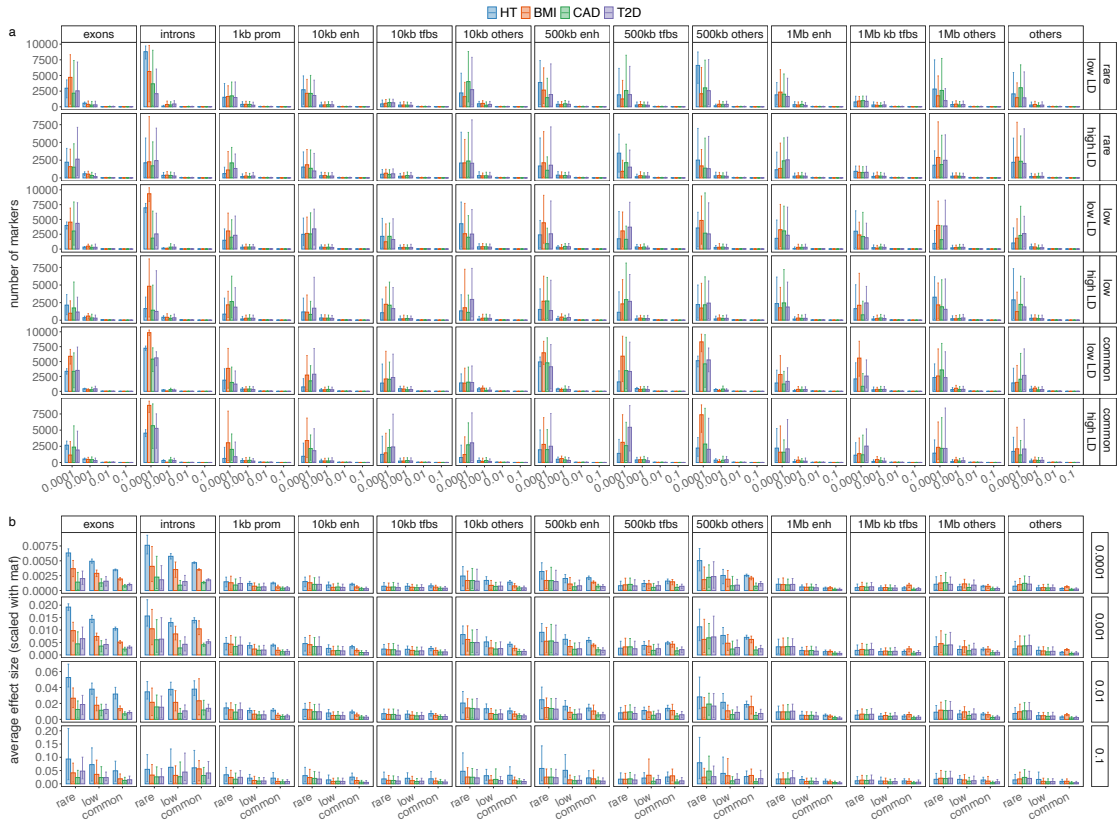


Supplementary Figure 9. Convergence diagnostics of model chains for UK Biobank analysis.(a) Overlapped density plots to compare the target distribution by chain showing each chain has converged in a similar space, for each annotation group and each trait. (b) Overlapped density plots comparing the last 10 percent of the chain (green), with the whole chain (pink), showing that the initial and final parts of the chain are sampling the same target distribution for each annotation group and each trait.

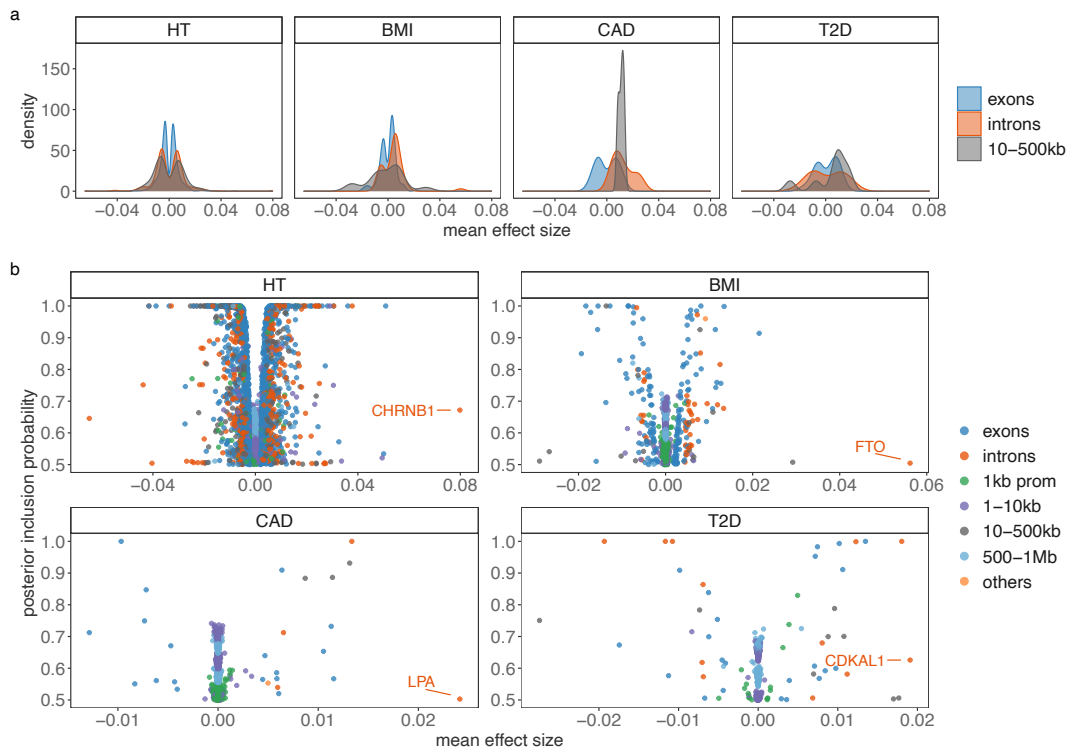


Supplementary Figure 10. Convergence diagnostics of model chains for UK Biobank analysis.(a) The potential scale reduction factor comparing the among- and within-chain variance for each annotation group and each trait. (b) The cross-correlation between all parameters for each annotation group and each trait.

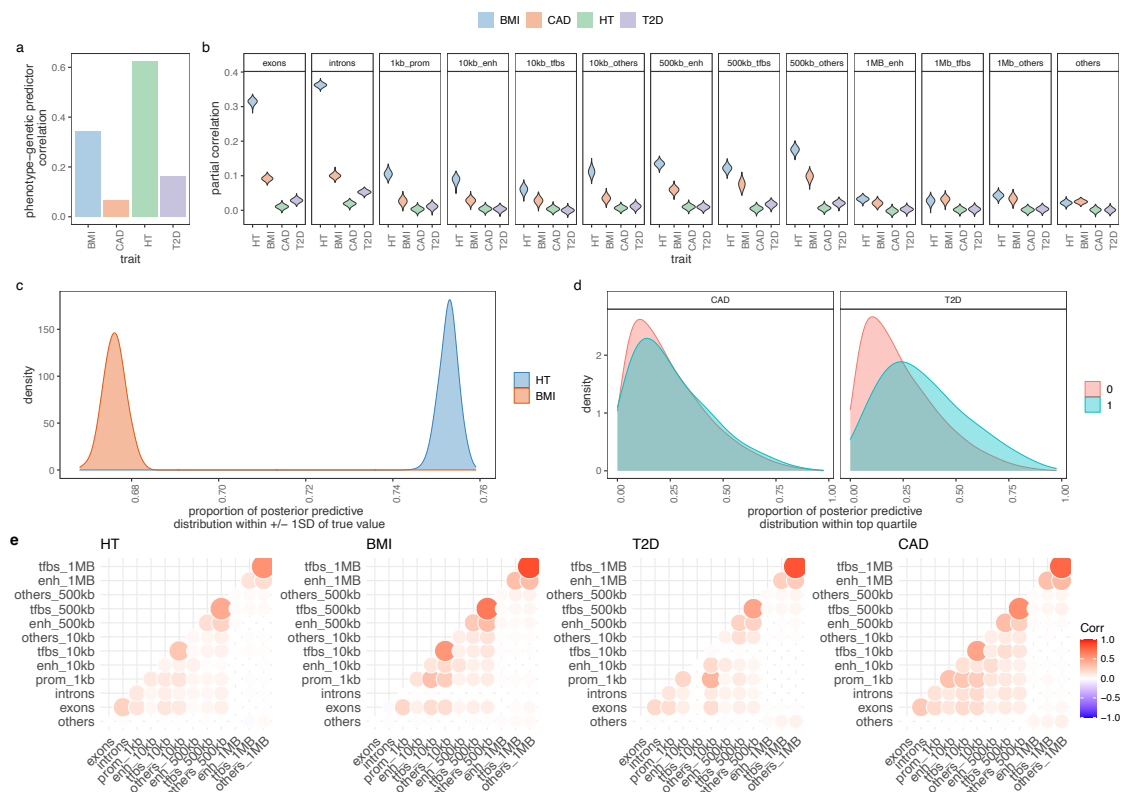




Supplementary Figure 12. Marker inclusion and effect estimate overview. (a) Bar plots of the number of markers entering the model for each mixture group (x-axis), within each MAF-LD group (y-axis facets, with top row MAF and bottom row LD), within each annotation (x-axis facets). Mixture 1 = 0.0001, 2 = 0.001, 3 = 0.01, 4 = 0.1. (b) Bar plots of the average effect size of markers in the model for each annotation group, scaling the effects to their frequency and split by mixture. Posterior summary of $n = 6,000$ iterations with a thin of 5 and burn-in of 500 for each trait in (a) and (b). Error bars give the 95% credible intervals in both panels.



Supplementary Figure 13. Contribution of SNPs with posterior inclusion probability (PIP) > 0.5 to height, body-mass-index (BMI), cardiovascular disease (CAD) and type-2-diabetes (T2D). (a) Shows the distribution of mean effect sizes for SNPs with PIP > 0.5 attributed to exons, introns and 500kb upstream of genes in each trait. (b) We then plot the relationship between mean effect size and posterior inclusion probability for SNPs with PIP > 0.5 attributed to the annotation groups (exons, introns, SNPs located 1kb, 1-10kb, 10-500kb and 500-1Mb upstream of genes and other un-mapped SNPs). We labelled the closest gene to the SNP with the highest mean effect size in each trait.



Supplementary Figure 14. Cross-cohort prediction accuracy and the posterior predictive distribution. (a) Correlation of the posterior mean predictor and height (HT), body mass index (BMI), type-2 diabetes (T2D), and cardiovascular disease (CAD). (b) the partial correlations of the phenotype and genomic predictors specific to different genomic annotations. (c) For height and BMI, we calculate the probability that the distribution of genomic predictors obtained for each individual is within 1 SD of the true phenotypic value. The density of these probabilities is shown. (d) For CAD and T2D, we plot density plots of the proportion of the posterior predictive distribution for each individual that is within the top quartile of the risk distribution. (e) Correlation of genetic predictors obtained across annotation groups.

Supplementary Notes

Supplementary Note 1

Model Specification

We begin by outlining the basic model BayesR, before then presenting our extensions. Consider p single nucleotide polymorphism (SNP) markers. If we gather samples for $i = 1, \dots, N$ subjects in an $N \times p$ matrix, \mathbf{G} , in which the elements are coded as 0 for homozygous individuals at the major allele, 1 for heterozygous individuals and 2 for minor allele homozygotes. Now, we wish to model their linear association with the phenotype $\mathbf{y} = (y_i)$ of subjects $i = 1, \dots, N$ in a standard linear regression model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (1)$$

We assume that the genotypes are standardized so that $\mathbf{X}_j = \frac{(\mathbf{G}_j - \mu_j \mathbf{1})}{\sigma_j}$ is the vector of genotypes for the j^{th} marker ($j = 1, p$) with zero mean and unit variance, i.e. the centered and scaled j^{th} column of \mathbf{G} . The column's mean $\mu_j \approx 2f_j$ and the column's standard deviation $\sigma_j \approx \sqrt{2f_j(1-f_j)}$ being f_j the minor allele frequency (MAF) of the SNP. We define $\boldsymbol{\beta}$ as a $p \times 1$ vector of partial regression coefficients with β_j the effect of a 1 SD change in the j^{th} covariate, and $\boldsymbol{\epsilon}$ is a vector ($N \times 1$) of residuals.

We estimate the model's parameters using Bayesian inference, assuming that the error term $\boldsymbol{\epsilon} | \sigma_\epsilon^2 \sim \mathcal{N}(0, \mathbf{I}\sigma_\epsilon^2)$. The log-likelihood of this model can be written as

$$l(\mu, \boldsymbol{\beta}, \sigma_\epsilon^2) = -\frac{N}{2} \log(2\pi\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2} \left(N(\hat{y} - \mu)^2 + (\mathbf{y}_c - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y}_c - \mathbf{X}\boldsymbol{\beta}) \right) \quad (2)$$

with $\mathbf{y}_c = \frac{(\mathbf{y} - \mu \mathbf{1})}{\sigma_y}$ a vector of centred and scaled responses (SD 1).

As we adopt a Bayesian approach, we place priors over the model parameters. For the covariate effects, $\boldsymbol{\beta}$, we use a mixture prior with Dirac spike and slab components, which have been extensively used for variable selection [1, 2]. The prior induces sparsity in the model through a Dirac-delta at zero, excluding variables from the model by setting their coefficients to zero. A slab component is centered at zero and shrinks the non-zero coefficients towards zero according to the slab's width. In our approach, the slab component is a scale mixtures of normals and thus each $\beta_j \in \boldsymbol{\beta}$ is distributed according to:

$$\beta_j \sim \pi_0 \delta_0 + \pi_1 \mathcal{N}(0, \sigma_1^2) + \dots + \pi_L \mathcal{N}(0, \sigma_L^2)$$

where $\pi_\beta = (\pi_0, \pi_1, \dots, \pi_L)$ are the mixture proportions, $\{\sigma_1^2, \dots, \sigma_L^2\}$ are the mixture-specific variances, and δ_0 is a discrete probability mass at zero. We further constrain the prior by assuming a single parameter representing the total variance explained by the effects σ_G^2 , with the component-specific variances proportional to σ_G^2 multiplied by a constant $\{C_1, \dots, C_L\}$ so that

$$\begin{bmatrix} \sigma_1^2 \\ \vdots \\ \sigma_L^2 \end{bmatrix} = \sigma_G^2 \begin{bmatrix} C_1 \\ \vdots \\ C_L \end{bmatrix}$$

The remaining prior structure for the model is then

$$\begin{aligned} \pi &\sim \text{Dirichlet}(\mathbf{1}) \\ \sigma_G^2 &\sim \text{Inv-Scaled}\chi^2(v_0, s_0^2) \\ \sigma_\epsilon^2 &\sim \text{Inv-Scaled}\chi^2(v_0, s_0^2) \end{aligned} \quad (3)$$

with weakly informative parameters for hyperparameters $v_0 = s_0^2 = 0.001$.

For notational convenience, we will refer to the mixture membership labels as (l_0, l_1, \dots, l_L) and we define a latent indicator of each SNP, j , $\boldsymbol{\gamma} = (\gamma_j, \dots, \gamma_p)^T$ with $\gamma_{j,l} = 0$ or 1, indicating whether or not the effect of SNP j falls into the zeroth mixture $\gamma_{j,l} = 0$, or follows a normal distribution with variance σ_l^2 . We define the "active set of coefficients" as those β_j such that $\beta_j \neq 0$ denoted as $\boldsymbol{\beta}_{\gamma \neq 0}$ with cardinality $|\gamma_\neq|_0$. Thus the objective of our inference scheme is to compute an estimate of the posterior distribution $f(\boldsymbol{\beta}_{\gamma \neq 0}, \sigma_\epsilon^2, \sigma_G^2, \mu | \mathbf{y}_c)$. This model has been termed BayesR [3, 4] and an effective proposed Gibbs sampling scheme [4] follows the following steps:

- (i) sample μ from $\mathcal{N}\left(\frac{\sum_{i=1}^N (\mathbf{y}_{e_i} - \mathbf{X}_j \beta_{\gamma \neq 0})}{N}, \frac{\sigma_\epsilon^2}{N}\right)$
- (ii) sample $\beta_{\gamma \neq 0}$ from its conditional as described below
- (iii) sample σ_G^2 from Inv - Scaled $\chi^2\left(\|\gamma_\varphi\|_0 + v_0, \frac{\|\gamma_\varphi\|_0 \|\beta_{\gamma \neq 0}\|^2 + v_0 S_0^2}{v_0 + \|\gamma_\varphi\|_0}\right)$
- (iv) sample σ_ϵ^2 from Inv - Scaled $\chi^2\left(v_0 + N, \frac{\|\mathbf{y}_e - \mu - \mathbf{X} \beta_{\gamma \neq 0}\|^2 + v_0 S_0^2}{v_0 + N}\right)$

From the former algorithm, steps (i), and (iv) are straight-forward applications of conjugacy and are common to many Gibbs sampling algorithms for linear regression. Step (iii) follows from conjugacy and the assumption that the individual mixtures represent fractions of the total variance explained by the coefficients. Step (ii) is the biggest bottleneck in any linear regression problem, and in the next section we will proceed to detail the derivations of the sampling scheme for this step.

While it is not uncommon to use non-proper priors for the residual's variance σ_ϵ^2 , in our case we chose to keep a proper prior for algorithmic and modeling reasons as: (a) conjugacy is amenable to Gibbs sampling (b) we assume σ_ϵ^2 and σ_G^2 are not nuisance parameters, and in some cases we possess prior information on its distribution. It is also common to specify the distribution of β_j having a variance depending on the residual's variance σ_ϵ^2 , which would make the estimates transformation-invariant. Recent results suggest the estimates for σ_ϵ^2 in this latter transformation-invariant formulation are biased [5]. Another concern may be that the prior's hyperparameters induce biased estimates for small variances [6], we acknowledge that may be an issue, and allow parameters v_0, s_0^2 to be adjusted if deemed necessary. The scale mixture of Gaussians, allows the prior distribution to have heavier tails than a single Gaussian, which allows big effects to be shrunk to a lesser degree than small effects [7]. Finally, the original formulation of [3, 4] assumes $\sigma_G^2 = r^2 \sigma_y$ which for centered and scaled phenotypes and genotypes, with heritability h^2 equal to reliability $r^2 = \frac{\text{Var}[\mathbf{X} \beta_{\gamma \neq 0}]}{\text{Var}[y]}$, would mean $\sigma_G^2 = h^2 = r^2 = \text{Var}[\mathbf{X} \beta_{\gamma \neq 0}] = \sum_{\gamma \neq 0} \beta_{\gamma \neq 0}^2$, but there is no constraint in the model ensuring $\sigma_G^2 + \sigma_\epsilon^2 = \sigma_y^2$. As we will see, further assumptions are necessary for having unbiased estimates of σ_G^2 and h^2 under varying LD and MAF. These estimates will achieve the equivalence $\sigma_G^2 = r^2 = h^2$ without relying in either using a point estimate of r^2 [3], informative priors on σ_G^2 , or normalising the posterior variances by $h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_\epsilon^2}$ [8].

Sampling the effects

For sampling β , the challenge is two-fold: (a) determining if the effect β_j is part of $\beta_{\gamma \neq 0}$, and if so, to which component it belongs; and then (b) sampling the vector $\beta_{\gamma \neq 0}$ from a multivariate Gaussian with covariance matrix $\Sigma = \mathbf{X}_{l \neq 0}^T \mathbf{X}_{l \neq 0} + \Lambda$ where Λ is the diagonal matrix with entries $\lambda_{l,j} = \frac{\sigma_\epsilon^2}{\sigma_{j,l}^2}$, with $\sigma_{j,l}^2$ the variance of the mixture component to which marker β_j was assigned. For (a), marginalization of each effect individually is required to compute the membership probability, which requires solving a determinant of the size of $\|\gamma_\varphi\|_0 - 1$ [2]. For (b), either a system of size $\|\gamma_\varphi\|_0$ must be solved through LU decomposition, or Cholesky decomposition of size $\|\gamma_\varphi\|_0$, and both operations are resource intensive when the size of $\|\gamma_\varphi\|_0$ is large. Instead, we determine the inclusion of a marker in the active set, along with its mixture membership and its partial regression coefficient β_j , in single-site updates. Single-site Gibbs sampling, also known as stochastic relaxation [9], has a long history given its equivalence to iterative Gauss Siedel methods to solve matrix equations [10]. Although we choose to use the BayesR model, many alternative models can easily be placed within the iterative solving and computational framework we outline here.

In this scheme, we sample each element, j , of β from the full conditional posterior $f(\beta_j | \beta_{\setminus j}, \mathbf{y}) \propto f(\beta_j, \beta_{\setminus j}, \mathbf{y})$ which can be written as $f(\beta_j, \beta_{\setminus j}, \mathbf{y}) = f(\mathbf{y} | \beta) f(\beta_j) f(\beta_{\setminus j})$ where $f(\mathbf{y} | \beta)$ is the density function of the conditional distribution of $\mathbf{y} | \beta$ and $f(\beta_j)$ and $f(\beta_{\setminus j})$ are the densities of the prior distributions of β_j and $\beta_{\setminus j}$ respectively, with notation $\setminus j$ representing all other covariates except j . The kernel of the full conditional posterior for β_j is proportional to the product of the likelihood, the prior distribution for β_j and the prior distributions of the variances, and thus ignoring factors that are constant with respect to β_j gives

$$f(\beta_j | l_j, \boldsymbol{\theta}_{\setminus j}, \mathbf{y}) \propto \exp\left[-\frac{(\mathbf{y}_e - \mathbf{X} \beta)^T (\mathbf{y}_e - \mathbf{X} \beta)}{2\sigma_\epsilon^2}\right] \exp\left[-\frac{\beta_j^2}{2\sigma_{j,l}^2}\right] \quad (4)$$

where l_j represents the mixture β_j is assigned, $\boldsymbol{\theta}_{\setminus j} = \{\beta_{\setminus j}, \sigma_\epsilon^2, \sigma_G^2, \pi_\beta, \mu\}$ and $\sigma_{j,l}^2$ the corresponding mixture variance. We can reduce the expanded form and drop terms that are free from β_j as

$$\begin{aligned}
f(\beta_j | l_j, \boldsymbol{\theta}_{\setminus j}, \mathbf{y}) &\propto \exp \left[-\frac{1}{2\sigma_\epsilon^2} (\mathbf{y}_c - \mathbf{X}_j \beta_j - \mathbf{X}_{\setminus j} \boldsymbol{\beta}_{\setminus j})^T (\mathbf{y}_c - \mathbf{X}_j \beta_j - \mathbf{X}_{\setminus j} \boldsymbol{\beta}_{\setminus j}) + \frac{\beta_j^2 \sigma_\epsilon^2}{2\sigma_{j,l}^2} \right] \\
&\propto \exp \left[-\frac{1}{2\sigma_\epsilon^2} \left(\tilde{\mathbf{y}}^T \tilde{\mathbf{y}} - 2\mathbf{X}_j^T \tilde{\mathbf{y}} \beta_j + \mathbf{X}_j^T \mathbf{X}_j \beta_j^2 + \frac{\beta_j^2 \sigma_\epsilon^2}{2\sigma_{j,l}^2} \right) \right] \\
&\propto \exp \left[-\frac{1}{2\sigma_\epsilon^2} (\tilde{\mathbf{y}}^T \tilde{\mathbf{y}} - 2\mathbf{X}_j^T \tilde{\mathbf{y}} \beta_j + \beta_j^2 \Sigma_{j,l}) \right] \\
&\propto \exp \left[-\frac{1}{2\sigma_\epsilon^2} (\tilde{\mathbf{y}}^T \tilde{\mathbf{y}} - 2\hat{\beta}_j \Sigma_{j,l} \beta_j + \beta_j^2 \Sigma_{j,l} + \hat{\beta}_j^2 \Sigma_{j,l} - \hat{\beta}_j^2 \Sigma_{j,l}) \right] \\
&\propto \exp \left[-\frac{1}{2} \frac{(\beta_j - \hat{\beta}_j)^2}{\frac{\sigma_\epsilon^2}{\Sigma_{j,l}}} \right] \tag{5}
\end{aligned}$$

with $\tilde{\mathbf{y}} = \mathbf{y}_c - \mathbf{X}_{\setminus j} \boldsymbol{\beta}_{\setminus j}$, $\Sigma_{j,l} = \mathbf{X}_j^T \mathbf{X}_j + \lambda_{j,l}$ and $\hat{\beta}_{j,l} = \frac{\mathbf{X}_j^T \tilde{\mathbf{y}}}{\Sigma_{j,l}}$. This gives the Gibbs sampling update for β_j as

$$\beta_j \sim \mathcal{N}(\Sigma_{j,l}^{-1} \mathbf{X}_j^T \tilde{\mathbf{y}}, \sigma_\epsilon^2 \Sigma_{j,l}^{-1}) \tag{6}$$

To avoid reducibility of the Markov chain, prior to drawing the effect β_j , we first need to select the mixture K for each covariate j , and as above we can condition on the individual coordinates and to obtain the probability that a coefficient j belongs to a given mixture.

$$\mathbb{P}(l_j = K | \boldsymbol{\theta}_{\setminus j}, \mathbf{y}) = \frac{f(\tilde{\mathbf{y}} | l_j = K, \boldsymbol{\theta}, \mathbf{y}) \mathbb{P}(l_j = K)}{\sum_{k=1}^L f(\tilde{\mathbf{y}} | l_j = k, \boldsymbol{\theta}, \mathbf{y}) \mathbb{P}(l_j = k)} \tag{7}$$

We integrate out the β_j coordinate following the equations above with

$$\begin{aligned}
f(\tilde{\mathbf{y}} | l_j, \boldsymbol{\theta}, \mathbf{y}) &= \int f(\tilde{\mathbf{y}} | \beta_j, \sigma_\epsilon^2) f(\beta_j | l_j, \sigma_{j,l}^2) d\beta_j \\
&= \int (2\pi\sigma_\epsilon^2)^{-n/2} \exp \left[-\frac{(\tilde{\mathbf{y}} - \mathbf{X}_j \beta_j)^T (\tilde{\mathbf{y}} - \mathbf{X}_j \beta_j)}{2\sigma_\epsilon^2} \right] (2\pi\sigma_{j,l}^2)^{-q/2} \exp \left[-\frac{\beta_j^2}{2\sigma_{j,l}^2} \right] d\beta_j
\end{aligned}$$

where $q = 2$. We then expand this equation using the relationship $\Sigma_{j,l} \hat{\beta}_j = \mathbf{X}_j^T \tilde{\mathbf{y}}$ from Eq. 6 and complete the squares

$$\begin{aligned}
f(\tilde{\mathbf{y}} | l_j, \boldsymbol{\theta}, \mathbf{y}) &= \int (2\pi\sigma_{j,l}^2)^{-q/2} (2\pi\sigma_\epsilon^2)^{-n/2} \exp \left[-\frac{1}{2\sigma_\epsilon^2} (\tilde{\mathbf{y}}^T \tilde{\mathbf{y}} - 2\hat{\beta}_j \Sigma_{j,l} \beta_j + \beta_j^2 \Sigma_{j,l} + \hat{\beta}_j^2 \Sigma_{j,l} - \hat{\beta}_j^2 \Sigma_{j,l}) \right] d\beta_j \\
&= (2\pi|\sigma_\epsilon^2 \Sigma_{j,l}^{-1}|)^{1/2} (2\pi\sigma_{j,l}^2)^{-q/2} (2\pi\sigma_\epsilon^2) \exp \left[-\frac{1}{2\sigma_\epsilon^2} (\tilde{\mathbf{y}}^T \tilde{\mathbf{y}} - \hat{\beta}_j^2 \Sigma_{j,l}) \right] \times \\
&\int (2\pi|\sigma_\epsilon^2 \Sigma_{j,l}^{-1}|)^{-1/2} \exp \left[-\frac{1}{2\sigma_\epsilon^2} (\beta_j - \hat{\beta}_j)^2 \Sigma_{j,l} \right] d\beta_j \\
&= (|\lambda_{l,j} \Sigma_{j,l}^{-1}|)^{\frac{1}{2}} (2\pi\sigma_\epsilon^2)^{-\frac{n}{2}} \exp \left[-\frac{1}{2\sigma_\epsilon^2} (\tilde{\mathbf{y}}^T \tilde{\mathbf{y}} - \hat{\beta}_j^2 \Sigma_{j,l}) \right] \tag{8}
\end{aligned}$$

where the final reduction in Eq. 8 occurs as the integral component is now a normal distribution that integrates to 1 and then terms are removed that do not contain, nor depend upon $\Sigma_{j,l}$ nor $\hat{\beta}_{j,l}$. The probability for inclusion in the model in the first mixture, as compared to the spike, then depends upon the ratio

$$\begin{aligned}
\frac{f(\tilde{\mathbf{y}} \mid l_j = 0, \boldsymbol{\theta}, \mathbf{y})}{f(\tilde{\mathbf{y}} \mid l_j = 1, \boldsymbol{\theta}, \mathbf{y})} &= \frac{(2\pi\sigma_\epsilon^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma_\epsilon^2}(\tilde{\mathbf{y}}^T \tilde{\mathbf{y}})\right]}{(|\lambda_{l,j}\Sigma_{j,2}^{-1}|)^{\frac{1}{2}} (2\pi\sigma_\epsilon^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma_\epsilon^2}(\tilde{\mathbf{y}}^T \tilde{\mathbf{y}} - \hat{\beta}_{j,l}^2 \Sigma_{j,2})\right]} \\
&= (|\lambda_{l,j}\Sigma_{j,2}^{-1}|)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_\epsilon^2}(\tilde{\mathbf{y}}^T \tilde{\mathbf{y}}) + \frac{1}{2\sigma_\epsilon^2}(\tilde{\mathbf{y}}^T \tilde{\mathbf{y}}) - \frac{1}{2\sigma_\epsilon^2}(\hat{\beta}_{j,l}^2 \Sigma_{j,2})\right] \\
&= (|\lambda_{l,j}\Sigma_{j,2}^{-1}|)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_\epsilon^2}(\hat{\beta}_{j,l}^2 \Sigma_{j,2})\right] \tag{9}
\end{aligned}$$

Analogous to equation 9, any comparison between mixtures has the same form and allows us to omit the $\tilde{\mathbf{y}}^T \tilde{\mathbf{y}}$ term. Thus placing Eq.9 into Eq.7 and re-arranging to a numerically more stable version [3] gives

$$\mathbb{P}(l_j = K \mid \boldsymbol{\theta}_{\setminus j}, \mathbf{y}) = \frac{1}{1 + \sum_{k=0}^L \exp[\log(LK_K) - \log(LK_k)]} \tag{10}$$

with $\log(LK_0) = \log(\pi_0)$ and $\log(LK_l) = -\frac{1}{2} \left[-\log(|\lambda_{l,j}\Sigma_{j,l}^{-1}|) - \left(\frac{\hat{\beta}_{j,l}^2 \Sigma_{j,l}}{\sigma_\epsilon^2} \right) \right] + \log(\pi_l)$ for l in $(1 \dots L)$.

Having derived the regression coefficients and their inclusion probabilities, fully specifying the BayesR model, we now proceed to: (1) extend this to a BayesRR-RC model in the Methods section; (2) derive a computational implementation that facilitate the application of the model to biobank sized data in Supplementary Note 2; and (3) derive the properties of the model parameters when applied to highly correlated genomic data (under multicollinearity) and compare these to estimates made by other approaches in the field in Supplementary Note 4.

Supplementary Note 2

A Gibbs sampling scheme for biobank size data

For " $p \gg n$ " regimes, such as in genomics, where the number of covariates is greater than the number of individuals, hierarchical models controlling assumptions over the sparsity of the model are typically proposed, with examples of sparsity-inducing priors like the "spike and slab" prior [1, 11], the Bayesian LASSO [12] and the Horseshoe [13] prior. There are efficient tools to perform Bayesian regression analysis "out-of-the-box" using MCMC and variational inference [14–16], but these methods are limited to problems with explanatory variables in the low thousands of observations. Recent results show that Gibbs samplers for the Horseshoe prior [17], or for the Bayesian LASSO [18], offer a competitive advantage when combined with approximation schemes for problems of high dimensionality (over 100,000 covariates). These latter methods exchange the inversion of the coefficient matrix, for a matrix multiplication, thus reducing complexity from cubic to almost quadratic on the number of variables. However, despite these good properties, scaling these approaches up to a factor of millions of variables remains prohibitive.

We now describe an effective algorithmic implementation of our BayesRR-RC model that scales to millions of individuals, each genotyped at millions of genetic markers. We outline a Gibbs sampling algorithm that enables all sampling steps to utilize genetic data stored in mixed binary/sparse-index representation, reducing computational complexity of a single Gibbs step from $\mathcal{O}(n)$ to $\mathcal{O}(n_z)$, with n_z the number of non-zero genotypes. We then outline a Bulk Synchronous Parallel Gibbs sampling scheme implemented based on a hybrid MPI + OpenMP model, distributing data across MPI tasks over as many compute nodes as required to hold all the data in memory. Uniquely, this enables large-scale genomic data to be split up into smaller manageable segments, whilst still conducting the analysis in the same way, estimating the marker effects jointly.

Algorithm 1: Serial Algorithm for sampling over the posterior distribution $p(\mu, \beta, \epsilon, \sigma_\epsilon, \theta)$. \mathbf{X}_{marker_j} represents column of \mathbf{X} corresponding to the column j of the vector $marker$. Given that $marker$ is shuffled before sampling the effects, this is equivalent to permuting the order of the effects to be sampled.

Data: Coefficient matrix \mathbf{X} , measurement vector \mathbf{y} , prior hyperparameters v_0, s_0^2 , iterations I

Result: mean μ , effects vector β , residual vector ϵ , residual variance σ_ϵ^2 and variance contributed by the marker effects, σ_G^2

```

1 Initialize  $\beta, \mu, \sigma_\epsilon^2, \sigma_G^2, \pi_\phi$  ;
2  $effects = 1, \dots, p$ ;
3  $\epsilon = \mathbf{y} - \mu$ ;
4 for  $i \leftarrow 1$  to  $I$  do
5   Sample  $\mu$ ;
6   Shuffle ( $effects$ );
7   for  $j \leftarrow 1$  to  $p$  do
8      $\beta_j^{old} = \beta_j$ ;
9      $\hat{\beta}_{j,l} = \frac{\mathbf{X}_j^T (\epsilon + \mathbf{X}_j \beta_j^{old})}{\Sigma_{j,l}}$ ;
10    Determine mixture component and sample the new value  $\beta_j$ ;
11     $\epsilon^{new} = \epsilon + (\beta_j^{old} - \beta_j) \mathbf{X}_j$ ;
12  Sample  $\sigma_\epsilon^2$ ;
13  Sample  $\sigma_G^2$ ;

```

Algorithm 1 provides a full overview of the sampling scheme of the model as it has been previously implemented. For each marker j , we must compute $\hat{\beta}_{j,l}$ to determine which mixture a marker belongs to, before then sampling β_j given the mixture group assigned. This quantity depends on the dot product $\mathbf{X}_j^T \mathbf{y}_c$, with \mathbf{y}_c the centred phenotype. If we keep in memory the vector of residuals $\epsilon = \mathbf{y}_c - \mathbf{X}\beta_{\gamma \neq 0}$, then we can compute efficiently $\mathbf{y}_c - \mathbf{X}_{\setminus j} \beta_{\gamma \neq 0, \setminus j}$ by the update $\mathbf{y}_c - \mathbf{X}_{\setminus j} \beta_{\gamma \neq 0, \setminus j} = \tilde{\epsilon} + \mathbf{X}_j \beta_j$, thus sampling from the joint distribution with a complexity $\mathcal{O}(p)$. The most expensive operation in Algorithm 1 is computing the numerator in step 9: $\mathbf{X}_j^T (\tilde{\epsilon} + \mathbf{X}_j \beta_j^{old})$. As the column vector \mathbf{X}_j contains the centered and scaled genotypes, step 9 involves one sum of two dense vectors and a dot product of two dense vectors. However, if we store in memory the mean, μ_j , and standard deviation σ_j of each column of the genotype matrix, we can express the numerator in step 9 with these quantities and the j -th column of the original genotype matrix \mathbf{G} as (with

$\sigma_j^2 = (\mathbf{G}_j - \mu_j \mathbf{1})^T (\mathbf{G}_j - \mu_j \mathbf{1}) / (n - 1)$ by definition):

$$\begin{aligned}
num &= \frac{(\mathbf{G}_j - \mu_j \mathbf{1})^T}{\sigma_j} \left(\epsilon + \beta_j^{old} \frac{(\mathbf{G}_j - \mu_j \mathbf{1})}{\sigma_j} \right) \\
&= \frac{(\mathbf{G}_j - \mu_j \mathbf{1})^T}{\sigma_j} \epsilon + \beta_j^{old} \frac{(\mathbf{G}_j - \mu_j \mathbf{1})^T (\mathbf{G}_j - \mu_j \mathbf{1})}{\sigma_j} \\
&= \frac{\mathbf{G}_j^T}{\sigma_j} \epsilon - \frac{\mu_j}{\sigma_j} \sum_{i=1}^n \epsilon + \beta_j^{old} (n - 1)
\end{aligned} \tag{11}$$

and we can do the same for the ϵ update:

$$\epsilon_{new} = \epsilon + (\beta_j^{old} - \beta_j) \frac{(\mathbf{G}_j - \mu_j \mathbf{1})}{\sigma_j} = \epsilon + \frac{(\beta_j^{old} - \beta_j)}{\sigma_j} (\mathbf{G}_j - \mu_j \mathbf{1}) \tag{12}$$

for which we only have to compute the difference of a sparse vector and a dense vector, and the sum of two dense vectors. Finally, to avoid computing $\sum_{i=1}^n \epsilon_{new}$ for each marker, we assign a variable to this quantity and update it after each ϵ update as follows (with $\mu_j = \sum_{i=1}^n \mathbf{G}_{i,j} / n$ by definition):

$$\sum_{i=1}^n \epsilon_{new} = \sum_{i=1}^n \epsilon + \frac{(\beta_j^{old} - \beta_j)}{\sigma_j} \left(\sum_{i=1}^n \mathbf{G}_{i,j} - n\mu_j \right) = \sum_{i=1}^n \epsilon \tag{13}$$

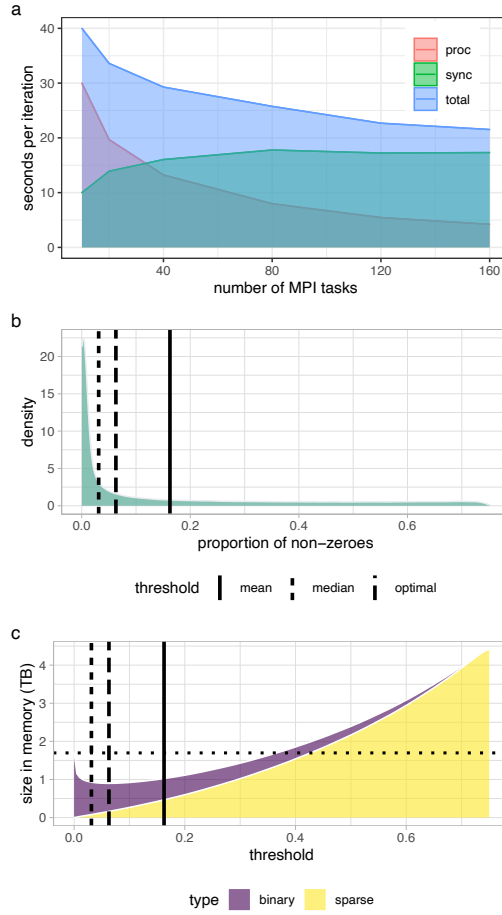
meaning that the sum of ϵ elements is constant during the algorithm execution (as expected as all involved vectors are zero-mean). Therefore, the only quantity to be computed per run (apart from the ϵ update) is the dot product $\frac{\mathbf{G}_j^T}{\sigma_j} \epsilon$ which can also be reduced, as the elements of \mathbf{G}_j can only be either $\{0, 1, 2\}$ with sequence

data or hard-coded genotype. We call \mathcal{I}_1 the indicator function such that $\epsilon \mathcal{I}_1 = \begin{cases} \epsilon_j & x_j = 1 \\ 0 & else \end{cases}$ and similarly

$\epsilon \mathcal{I}_2 = \begin{cases} \epsilon_j & x_j = 2 \\ 0 & else \end{cases}$ which then gives the dot product as $\frac{\mathbf{G}_j^T}{\sigma_j} \epsilon = \frac{\sum \epsilon \mathcal{I}_1 + 2 \sum \epsilon \mathcal{I}_2}{\sigma_j}$ meaning that multiple $\mathcal{O}(n)$

multiplications are now $\mathcal{O}(n_z)$ sums, and also that instead of storing in memory a sparse matrix of elements plus its indexes, we just need to store three ragged arrays of indexes, one for the "1" elements, a second one for the "2" elements, and a third one for the "M"issing elements. Those arrays contain information for all markers processed by a MPI task and are of unsigned integer type (32 bits). They store indices of the 1, 2 and M elements within the marker (i.e. ranging from 0 to $N - 1$). It corresponds to the smallest integer type that allows us to scale to hundreds of thousands or millions individuals. On top of those 3 ragged arrays there are two meta-data arrays for each element type which provide the starts and lengths of the 1, 2 and M elements for each marker in the ragged arrays. They are loaded in memory from reading sparse data files stemming from the conversion of the original Plink .bed file and accessed in parallel by the tasks with MPI I/O.

Even though the sparse representation is optimal in number of operations, performance may vary depending on hardware as a vectorised dot product may be faster than sparse dot product. Spatially, the sparse representation is optimal as long as the columns are sparse. In genotype data, even though the expected number of non-zeros per column is given by the average MAF ($\sim 20\%$ in the UK Biobank data), the distribution is long tailed (Supplementary Figure 15). These columns at the tail of the distribution can dominate the total size of the data structure in memory. Encoding a single column has a constant size of $N \times 2$ bits in plink's .bed file format (referred from now on as binary format), while in sparse representation a column has varying size of $n_z \times 32$ bits. If we encode the columns with less than 6% of non-zeros as sparse and the rest in the original binary format, we can have a total memory occupancy of 60% the size of the original genotype matrix in Plink bed format. In Supplementary Figure 15, we represent on panel (b) the distribution of the proportion non-zeros per column of a genotype matrix for $\sim 4 \times 10^5$ individuals and $\sim 1.5 \times 10^7$ SNPs, solid line representing the mean of the distribution and slashed line the median. In panel (c) we show the total size of the data in memory as a function of the threshold used to split between binary and sparse format, in purple we see how the binary representations dominates the total size up until the mean of the distribution, after which, the size of the sparse data structure starts to dominate and ends up being around four times bigger than the original .bed file size (dotted horizontal line). We found the optimal threshold to be around 0.064 (6.4%, Supplementary Figure 15).



Supplementary Figure 15. A mixed representation bulk synchronous hybrid-parallel Gibbs sampling scheme for genomic data.

(a) The minimum seconds per iteration achieved for 382,466 unrelated individuals from the UK Biobank data genotyped at 8,430,466 markers, with an increasing number of message-passing interface (MPI) tasks used. The total seconds is given in blue and this is subset into (i) the time taken to process the markers and estimate all of the 8,433,421 marker effects and hyper-parameters (proc), and (ii) the time taken to synchronise the estimates as they are being obtained (sync). With increasing data parallelism parameter estimation times drop quickly to less than 5 seconds with 160 MPI tasks, however the time taken to synchronise the estimates increases as the number of tasks increases. The SD was 1 second, with variation in sampling times induced by fluctuations in networking speed that influenced the synchronisation times. Each MPI task was able to used 4 CPUs. (b) the distribution of the proportion non-zeros per column of a genotype matrix for $\sim 4 \times 10^5$ individuals and $\sim 1.5 \times 10^7$ SNPs taken from UKB, with solid line representing the mean of the distribution and dashed line the median. (c) the size in memory in TB of the data as the coding of the SNP markers moves from binary to the sparse indexed format, the optimal threshold is achieved between mean and median of the distribution of non-zeros in the genotype matrix. Above this threshold columns are coded in binary format below in sparse index. Through a combination of a mixed data representation and highly vectorized look-up tables, memory usage is reduced while maintaining fast computational speed.

Finally, we implement a vectorized dot product for genotype data stored in the raw binary format based on a couple of look-up tables, by writing the dot product as:

$$\begin{aligned} \frac{(\mathbf{G}_j - \mu_j \mathbf{1})^T}{\sigma_j} \epsilon &= \sum_i \frac{\psi_{i,j} \epsilon_i}{\sigma_j} \\ &= \frac{1}{\sigma_j} \left(\sum_i a_i \epsilon_i - \mu_j \sum_i b_i \epsilon_i \right) \end{aligned} \quad (14)$$

with coefficients a_i and b_i being 0.0, 1.0 or 2.0 depending on the value of $\mathbf{G}_{i,j}$ and following Table 2.

$\mathbf{G}_{i,j}$	0	1	2	NA
2-bit	11	10	00	01
a_i	0.0	1.0	2.0	0.0
b_i	1.0	1.0	1.0	0.0
$\psi_{i,j}$	$0.0 - 1.0\mu_j$	$1.0 - 1.0\mu_j$	$2.0 - 1.0\mu_j$	$0.0 - 0.0\mu_j$

Supplementary Table 2. a and b coefficient values used for building up the two look-up tables needed for the vectorization of the dot product computation when processing binary data.

As 1 byte of plink’s .bed can contain $4^4 = 256$ different combinations of information for 4 individuals, we can setup two lookup tables with 256×4 entries each that will give for any byte the corresponding 4 a_i and b_i coefficients, hence allowing for vectorisation of Eq. 14 by performing $a_i\epsilon_i$ and $b_i\epsilon_i$ and accumulating them for 4 individuals at once. Additionally, we use OpenMP to parallelize the loop over the marker’s bytes. This greatly extends previously proposed sparse residual updating schemes and also facilitates the synchronous, fully parallel bulk-synchronous Gibbs sampling scheme that we describe in the next section below.

Bulk-synchronous parallel Hogwild Gibbs sampling with sparse data

Bulk-synchronous parallel Hogwild Gibbs sampling [19] assigns block of columns from \mathbf{X} to workers that then sample from $f(\beta_j|\beta_{\setminus j}, \mathbf{y})$ for each of the columns in their block. Workers can communicate between each other exchanging the current values of the variables they are sampling, or the whole state of variables for workers in particular. If we perform global synchronisation steps the algorithm is called Bulk-synchronous parallel Hogwild (BSP), if on the other hand, workers exchange messages without a global synchronisation, the algorithm is called Asynchronous parallel Hogwild (ASP) [20].

Algorithm 2: Hogwild Gibbs with ‘ $\Delta\epsilon$ -exchange’.

components : Define K parallel workers
1 Define global variables $\mu, \beta, \epsilon, \pi, \sigma_g^2, \sigma_e^2$;
2 Initialize variables;
3 **for** $i \leftarrow 1$ **to** I **do**
4 Update μ ;
5 Update β in parallel using $\mathbf{DEpsX}(K)$;
6 Update hyperparameters $\pi, \sigma_g^2, \sigma_e^2$;

We propose Algorithm 2, which is a modification of a BSP algorithm where we sample the individual coefficients in parallel conditioned on the hyperparameters. We assign workers (MPI tasks) subsets of coefficients to sample, and each worker performs local Gibbs steps until a global synchronisation is triggered. This global synchronisation happens many times in each iteration, during the phase in which we sample the individual coefficients β_j . For this algorithm, we developed a synchronisation scheme called ‘ $\Delta\epsilon$ -exchange’ as outlined in Algorithm 3. In this scheme each individual worker is assigned a block of columns from \mathbf{X} and is in charge of sampling from $f(\beta_j|\beta_{\setminus j}, \mathbf{y})$ for each of the columns in its block. We add an additional parameter for the synchronisation rate Ω . After Ω columns have been sampled in all workers (around 5-10 in practice to avoid divergence occurring), a synchronisation move is executed.

The purpose of the synchronisation move is to update all of the workers’ state based on the coefficients sampled from $t = 1$ until $t = \Omega$ in all workers. The sufficient statistic for this state is contained in the residual vector ϵ . Thus from $t = 1$ until $t = \omega$ each worker computes $f(\beta_j|\epsilon_{t=1})$ and keeps track of its local change in ϵ which we denote $\Delta\epsilon = \sum_1^\Omega \mathbf{X}_\omega \beta_\omega$ for ω in the set of indexes for the current batch of variables in the workers list of variables. For the synchronisation step, we use the MPI_Allreduce collective, meaning that each task will receive the sum of locally accumulated $\Delta\epsilon$ from all tasks to update its $\epsilon_{t=1} = \sum^w \Delta\epsilon_w$ for $w = (1..W)$ workers. With the new $\epsilon_{t=1}$, the worker proceeds to sample the next Ω -sized batch of columns from its set of columns. This synchronisation scheme allows workers to exchange state information in compact form, as the total size of memory occupied in total by the messages is $\mathcal{O}(NW)$.

Algorithm 3: ‘ $\Delta\epsilon$ -exchange’ for synchronising changes in backfitted residuals in our BSP Gibbs sampling algorithm.

```

1 DEpsX ( $K$ )
   components: Set of  $K$  workers, each one  $\beta_k$ , Set of  $K$  messages, each one  $\Delta\epsilon_K$ ,  $K$  sets of  $\sim \frac{p}{K}$ 
                 columns, each set of columns assigned to a worker.
2   foreach worker  $\beta_k$  do
3      $\epsilon_k = \epsilon$ ;
4      $\Delta\epsilon_k = 0$ ;
5     foreach column  $i$  in a subset of size  $\Omega$  of the columns assigned to  $\beta_k$  do
6        $\beta_j^{old} = \beta_i$ ;
7       draw  $\beta_i$  from  $f(\beta_i | \epsilon, \sigma_\epsilon^2, \sigma_G^2, \pi)$ ;
8        $\Delta\epsilon_k = \Delta\epsilon_k - X_i(\beta_i - \beta_j^{old})$ ;
9     Wait until all workers are finished processing their  $\Omega$  sets;
10     $\epsilon = \epsilon + \sum_k \Delta\epsilon_k$ ;

```

Previous results point to BSP Gibbs sampling for a multivariate Gaussian converging if the covariance matrix is strictly diagonal-dominant [20] with zero covariance of the markers split across workers. The risk for genomic data, is that two markers in LD get updated at the same time in parallel, double counting their effects, and leading to ϵ being mis-estimated after a synchronization has occurred. Suppose we have one fixed causal marker and two other markers i and j that are assigned to different MPI tasks. Suppose that the Pearson correlation between the causal marker and marker i or j is ρ_i and ρ_j , respectively. Finally, let ρ denote the correlation between the markers i and j . For simplicity in this example suppose that the inclusion probability of the causal marker is q and we make an assumption that the inclusion probability of the marker i is then $P(\beta_i \neq 0) = q\rho_i$ and for marker j it is $P(\beta_j \neq 0) = q\rho_j$, that means that the inclusion probability is proportional to the correlation between causal and other markers. In reality, the effect size estimate is actually proportional to the causal effect: $\hat{\beta}_i = \rho_i\beta_{causal}$ and the function between posterior inclusion probability and causal effect size $q(\beta_{causal})$ is not linear for $\beta_{causal} \geq 0$ as described in Eq.(10) and thus we cannot assume that $P(\beta_i \neq 0) = q\rho_i$ in practice. In the case of parallelising the markers between two tasks we are interested in the probability that two markers from different tasks will absorb the effect of a same causal variant. Thus, we are interested in the probability $P(\beta_i \neq 0, \beta_j \neq 0 | i, j \in U)$, where U is the set of markers that are updated simultaneously in two different tasks. Thus, we can write:

$$P(\beta_i \neq 0, \beta_j \neq 0 | i, j \in U) = P(\beta_i \neq 0)P(\beta_j \neq 0) = q^2\rho_i\rho_j.$$

We see that the probability of making a mistake is dependant on the product $\rho_i\rho_j$. The correlation matrix R of the three markers

$$R = \begin{pmatrix} 1 & \rho_i & \rho_j \\ \rho_i & 1 & \rho \\ \rho_j & \rho & 1 \end{pmatrix}$$

has to be positive semi-definite and thus we can examine what are the possible values for the product $\rho_i\rho_j$ given that we know ρ . Note that the value of ρ can be controlled by providing some blocking mechanism that would assign SNPs to the tasks so that the correlation for the markers from different tasks would be limited to ρ and this is what we advocate here, placing contiguous blocks of markers into different tasks, so as to maximise the LD within a block (MPI task), but minimise the LD across blocks. The maximum possible values for the product follow a linear function that depends on ρ as

$$\max_{\rho_i, \rho_j, \rho = \tilde{\rho}} = 0.5 + 0.5\tilde{\rho}.$$

To get better estimates for the constraints for the product $\rho_i\rho_j$ then we need to make further assumptions about the distribution of ρ_i or ρ_j . Therefore, we can say that $P(\beta_i \neq 0, \beta_j \neq 0 | i, j \in U) \leq q^2(0.5 + 0.5\rho)$. This result and inequality only holds per sampled pair (i, j) . We then multiply this result with the probability of sampling the pair (i, j) that both have correlations $\rho_i, \rho_j > 0$. Denoting a set of markers that have a positive correlation with one specific causal marker as the causal radius C , The probability of sampling any pair (i, j) is

$$P(i, j \in U) = \frac{1}{T^2},$$

where T is the number of markers per one task. The probability of pair (i, j) belonging to C is $P(i, j \in C) = c(\ll 1)$, some reasonable values could be proposed or estimated for this (for example, $c = (\frac{\#(\text{markers-in-LD})}{2T})^2$). Combining the results together we get that the probability of making a mistake at one update of a pair (i, j) :

$$P(\beta_i \neq 0, \beta_j \neq 0) = P(\beta_i \neq 0, \beta_j \neq 0 | (i, j) \in U; (i, j) \in C)P((i, j) \in U)P((i, j) \in C) = \\ P(\beta_i \neq 0, \beta_j \neq 0 | (i, j) \in U) \frac{c}{T^2} \leq q^2(0.5 + 0.5\rho) \frac{c}{T^2}.$$

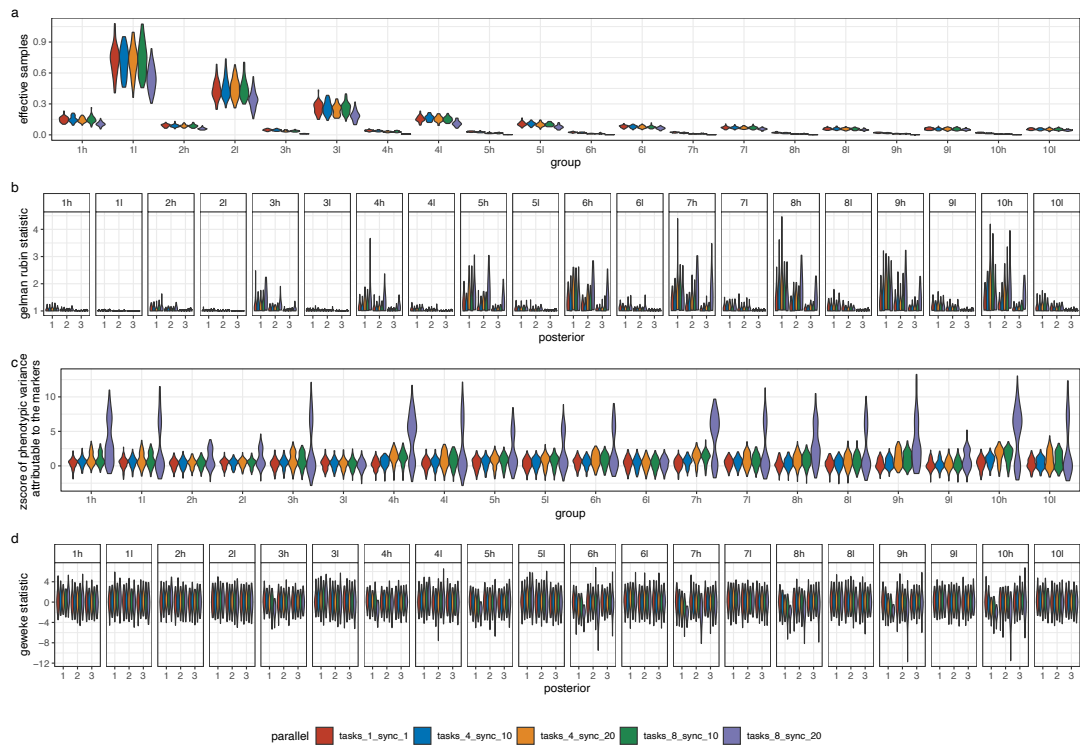
This result goes for one fixed causal marker and it also represents the expected number of mistakes per sampled pair (i, j) for one causal marker. If we want to find the expected number of mistakes per sampled pair, we should sum across the P causal markers:

$$Errors \leq \sum_{i=1}^P q_i^2(0.5 + 0.5\rho) \frac{c}{T^2} = (0.5 + 0.5\rho) \frac{c}{T^2} \sum_{i=1}^P q_i^2 \leq (0.5 + 0.5\rho) \frac{cP}{T^2}$$

To provide some intuition, we can think of an extreme scenario and assume that there are 100,000 variants in the SNP marker data that would enter the model as they are in LD with underlying causal variants, that each of these variants has posterior inclusion probability of 1, and that for each variant there are two blocks with 30,000 markers in total of which 100 markers have $LD = 1$ with the causal variant, and that both blocks contain 30,000 markers. Placing these values into what we derive above and sampling over 10,000 iterations leads to probability of an error ~ 0.1 throughout the sampling for this extreme example. Having derived a stable highly parallel Gibbs sampling algorithm for large-scale genomics data, we then performed exhaustive empirical validation of our algorithm in simulation study as described below.

Testing algorithm performance and parallelism in simulation

We explored the influence of increasing parallelism in our algorithm. We used the simulated data described above for the randomly sampled 50,000 UK Biobank individuals with imputed genotype data for chromosome 22, where we sampled randomly 4988 evenly spaced markers as causal variants and randomly assigned the effect sizes from a normal distribution with zero mean and variance $0.6/4988$ (the fourth scenario). For each of the 50 simulation replicates, we compared the three chains obtained by running the BayesRR-RC model (with 20 MAF-LD groups) in serial, with a single MPI task and synchronisation rate of 1 (residual updating after sampling each SNP), to three chains obtained by increasing the number of MPI tasks to 4 and then to 8, with synchronisation rates of 10 and 20 sampling steps before residual updating. For each simulation, we ran three chains of our BayesRR-RC model with different starting values for 3000 iterations. Like with all MCMC chains of regression models, convergence and sampling properties will be problem specific and dependent upon the LD of the markers, LD among the causal variants, the phenotypic variation attributable to the SNP markers across the MAF and LD spectrum, the study sample size, the degree of data parallelism per total marker number, and the synchronisation rate. Thus, the aim here is to simply show a series of diagnostic tests that can be utilized to explore the properties of the posterior to highlight how the different metrics can be used to identify convergence issues. We use the distribution, across simulations, of the proportion of effective samples obtained for the hyperparameter estimate of the proportion of phenotypic variance attributable to the markers of each group. This shows that for all ranges of parallelism, we achieve more effective samples for low MAF and low LD variants. As high MAF SNPs are interchangeable in the model to a large degree, their entry and exit from the model is correlated across iterations, and thus this is entirely expected and is actually a consequence of the model mixing. With high synchronisation rates, where many marker updates occur before residual updating by message passing a reduction in effective sample sizes occurs. We also use the distribution of the Gelman-Rubin test statistic for the three chains, a general metric to monitor convergence that compares within- and among-chain variance, as the number of iterations increases. Finally, a Geweke statistic value can be used to test the equality of the means of the first and last part of the Markov chains. We present the results of this simulation in Supplementary Figure 16 also including the distribution of z-scores of the posterior distribution of the phenotypic variance attributable to the markers for each MAF-LD group from the simulated values, which show stability of the estimates obtained with increasing data parallelism (tasks), but that a very high synchronisation rate with high parallelism can lead to poor convergence rates, meaning that the chains would have to be run for longer (Supplementary Figure 16).



Supplementary Figure 16. Simulation study of increasing task parallelism and increasing message passing rate for our hybrid-parallel sampling scheme.

We aimed to compare (a) the effective samples obtained, (b) the convergence rate of the algorithm, (c) the accuracy of the estimation, and (d) the stability of the estimates obtained as data parallelism increases within a burn-in period of the initial 3000 iterations. For 50,000 randomly selected UK Biobank individuals, and 111,425 imputed SNP markers of chromosome 22, we simulated 50 replicate phenotypes by randomly selecting 4,988 SNPs as causal variants and randomly allocating effect sizes from a normal distribution, with SNP heritability of 0.5. For each simulation, we ran three chains of our BayesRR model with different starting values for 3000 iterations. The SNP marker data was grouped into deciles of the distribution of linkage disequilibrium (LD), giving twenty groups in total (11 = MAF decile 1, low LD; 1h = MAF decile 1, high LD; ...; 10l = MAF decile 10, low LD; 10h = MAF decile 10, high LD). We repeated the three chains, but with increasing data parallelism: (1) in serial where one MPI task is used and the residual is updated after each marker is sampled (tasks_1_sync_1); (2) where the markers were split across four MPI processes with synchronisation occurring by message passing after 10 markers have been updated (tasks_4_sync_10); (3) where the markers were split across four MPI processes with synchronisation occurring after 20 markers have been updated (tasks_4_sync_20); (4) with 8 MPI processes and synchronisation of 10 (tasks_8_sync_10); and (5) with 8 MPI processes and synchronisation of 20 (tasks_8_sync_20). (a) shows the distribution across simulations of the proportion of effective samples obtained for the hyperparameter estimate of the proportion of phenotypic variance attributable to the markers of each group. For all ranges of parallelism, we achieve more effective samples for low MAF and low LD variants. With high synchronisation rates, where many marker updates occur before residual updating by message passing a reduction in effective sample sizes occurs. (b) gives the distribution of the Gelman-Rubin test statistic for the three chains, a general metric to monitor convergence that compares within- and among-chain variance, as the number of iterations increases. On the x-axis, 1 gives the distribution of the statistic across chains and MAF-LD groups for the first 500 iterations showing divergence of the chains (y-axis value $\gg 1$) across all MAF-LD groups, 2 gives the distribution for the first 1000 iterations, and 3 gives the distribution for the whole chain showing convergence of the chains by the end of this initial 3000 iteration sampling period irrespective of the data parallelism, with the exception of a few groups with infrequent synchronisation and high data parallelism which have yet to converge within this burn-in phase. (c) gives the distribution of z-scores of the posterior distribution of the phenotypic variance attributable to the markers for each MAF-LD group from the simulated values, showing stability of the estimates with increasing data parallelism (tasks), but not with infrequent synchronisation within the 3000 iterations run here. (d) shows the distribution of the Geweke statistic value which is a test of the equality of the means of the first and last part of the Markov chains. On the x-axis, 1 gives the distribution of the statistic calculated using all iterations across all MAF-LD groups, 2 gives the distribution discarding the first 500 iterations, and 3 gives the distribution discarding the first 1000 iterations. (a) - (d) suggest that our hybrid-parallelism sampling scheme achieves the same accuracy and convergence rates as a serial sampling scheme, provided that frequent synchronisation occurs and data parallelism is kept moderate. At high data parallelism and infrequent synchronisation, our theory shows that we are more likely to make a sampling mistake, preventing chains from converging and requiring longer sampling times. Convergence and accuracy of the MCMC Gibbs sampling chain will be problem specific and dependent upon the LD of the markers, LD among the causal variants, the phenotypic variation attributable to the SNP markers across the MAF and LD spectrum, the study sample size, the degree of data parallelism per total marker number, and the synchronisation rate. Therefore, like with all MCMC chains, a series of diagnostic tests can be utilized to explore the properties of the posterior and here we show how different metrics can be used to identify convergence issues.

Implementation and processing setup

We implement algorithms 2 and 3 in C++ as a pure CPU MPI + OpenMP hybrid solution. All data structures were properly aligned in memory to assist vectorization and assembly code was examined to ensure that the code was properly vectorized where expected. We utilize the scientific library boost (see Code Availability) and we profiled and benchmarked the code with Intel performance analysis tools such as Advisor and Ampflifier. Current implementation requires to be compiled with Intel compiler on an architecture supporting at least AVX2 although support for AVX512 is recommended for performance. UK Biobank results were generated on the cluster Helvetios from EPFL (see Code Availability) using 10 compute nodes and setting 8 MPI tasks per node and dedicating 4 (physical) cores to each task. 10 is the minimal number of nodes that was required to hold all the data in memory in its mixed-representation. An overview of the run times and memory use are provided in Supplementary Figure 15.

Supplementary Note 3

Posterior summaries and discovery

The ability of the additive regression model outlined and applied here to infer the underlying distribution of genomic effects is limited unless an additive model with many 0 coefficients holds as approximately true and the true number of underlying nonzero coefficients is $\ll n$. Various ad hoc penalty functions in machine learning, and the range of proper priors employed by members of the Bayesian alphabet and beyond, all impose a restriction on the size of the regression coefficients, and while these restrictions differ, they all provide shrinkage estimators that by their definition are biased as they are shrunk toward zero (this is true of mixed-linear association models also). In other words, the penalty function (prior) will be important and will influence the inference made here. Thus, the inference we obtain can only be made with respect to our *a priori* assumption that many marker effects are zero, and that the effects of those that are not zero can be reflected by a mixture of zero centred Gaussian distributions. Given this, we focused on comparing the posterior distributions of different traits obtained under the same model, focusing on the hyper-parameter estimates obtained for MAF-LD-annotation groups, and comparing these across traits. It has been shown in Bayesian penalized regression models that what is learned about β is a function of what is learned about $\mathbf{X}\beta$ and thus by placing separate hyper-parameters over different genomic groups we can obtain inference as to the variance contributed by each group [21]. As we show through theory and simulation study described below, MAF-LD-annotation specific hyper-parameters likely results in improved inference as to the distribution of genetic effects. However, with the exception of very rare variants with $LD \sim 0$, we cannot treat each β_j as independent and thus here we outline a strategy to identify associated genes, or genomic regions within a probabilistic framework.

For a simple example, consider two markers in LD that are correlated with a single causal variant, where either or both markers may be in the model at any one iteration and the expected posterior inclusion probability of each SNP is 0.5. In this scenario, we cannot use the posterior inclusion probability of each marker to assess association and thus instead, we take an approach of assessing the contribution of different genomic regions to trait variation whilst controlling the posterior type I error rate (PER), which is more suitable controlling for false positives, than controlling the genome-wide error rate (GER). Many papers have discussed the advantages of controlling the false discovery rate (FDR), and related measures rather than controlling GER [22] and here we follow [23] where the posterior probability that β_j is nonzero for at least one SNP j in a window or genomic segment is used to make inferences on the presence of an association in that segment.

Briefly, following [23], we will refer to this probability as the window posterior probability of association (WPPA). The underlying assumption is that if a genomic window contains a marker in LD with a causal variant, one or more SNPs in that window will have nonzero β_j . Thus, WPPA, which is estimated by counting the number of MCMC samples in which β_j is nonzero for at least one SNP j in the window, can be used as a proxy for the posterior probability that the genomic region contains a causal variant. Because WPPA for a given window is a partial association conditional on all other SNPs in the model, including those flanking the region, the influence of flanking markers on the WPPA signal for any given window will be inversely related to the distance k of the flanking markers. Thus, as the number of markers between a causal variant and the focal window increases, the influence of the causal variant on the WPPA signal will decrease and so WPPA computed for a given window can be used to locate associations for that given window [23].

This measure can be shown to control the PER, which in frequentist statistics would be associated with the test of a hypothesis. The null hypothesis in this case is that the genomic region does not contain any SNPs associated with the trait. Using this notation, WPPA is the conditional probability that the null is false given the observed data, while PER is the conditional probability that the null hypothesis is true given that it has been rejected based on some statistical test. Suppose the test is based on WPPA and the null is rejected whenever WPPA is larger than some value t . Then, PER is the probability that the null hypothesis is true given WPPA is larger than t , and it can be written as:

$$\text{PER} = \Pr(H_0 \text{ is true} | \text{WPPA} > t) = E[(1 - \text{WPPA}) | \text{WPPA} > t] \quad (15)$$

Thus, for any interval with $\text{WPPA} > t$ the proportion of false positives among significant results will be $\leq (1 - t)$. Here, we are interested in detecting genes and genomic regions that explain more than some proportion v of the total phenotypic variance attributable to the SNP markers (genetic variance). The genomic segment variance is defined as the sum of the squared partial regression coefficient estimates at each iteration and these are divided by the sum of all the squared partial regression coefficient estimates genome-wide to give a proportion for each genomic region at each iteration. Then we simply count the

proportion of MCMC samples where the proportion of genetic variance is greater than a thresholds of 0.001% and we denote this metric as the posterior probability of window variance (PPWV).

We extend this PPWV approach to develop an association metric for LD blocks of the genome. Currently, association studies predominantly estimate SNP effects and test for association one marker at a time, which does not control for local LD among SNP markers. Thus, the level of association determined is at the regional level as results are reduced, using LD patterns, to a subset of the strongest associated LD-independent variables. We can solve the problem of having a correlated posterior distribution by applying our PPWV approach to the LD blocks of the genome providing a Bayesian probabilistic metric of association that is equivalent to selecting the number of independent associated SNP markers. We define LD blocks as a group of SNPs that have squared correlation greater than 0.15 and then for each iteration, we sum the squared partial regression coefficient estimates for all the SNPs within the block, divide this by the sum of all the squared partial regression coefficient estimates genome-wide to give a proportion for each genomic region at each iteration. Then we simply count the proportion of MCMC samples where the proportion of genetic variance is greater than a thresholds of 0.001% providing a probabilistic association metric for each LD block that controls the FDR genome-wide. Within each associated region the individual SNP posterior inclusion probabilities can then also be used to "fine-map" the associations, in order to select the base-pair position that is most likely to be closest to the true underlying causal variant in imputed SNP data.

Supplementary Note 4

Comparison to other approaches under collinearity

Genome-wide association studies have predominantly been conducted using single marker regression via ordinary least squares (OLS). Recently, it has been proposed that if aggregation due to familial or molecular similarity (e.g. population stratification) exists in the data, a better estimation approach is generalized least squares (GLS), as it poses a more general covariance structure than OLS. GLS estimates can be obtained within mixed-linear association models, which first declare all marker effects as random variables, for example, assuming that $u_j \sim N(0, \sigma_u^2)$, or from a mixture of distributions, with all markers in the set taken as independently and identically distributed random variables. Second, when the markers are evaluated for association, they are then treated as a fixed effect. The resulting model can be written as

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_1u_1 + \mathbf{X}_{\setminus 1}\mathbf{u}_{\setminus 1} + \boldsymbol{\epsilon} \quad (16)$$

where a focal genetic marker, here \mathbf{X}_1 is fitted twice, first as a fixed effect to estimate the regression coefficient β_1 , and also as part of all of the other markers with their effects, u , estimated as random (note here $\setminus 1$ indicates all markers other than marker 1). Under this model the phenotypic covariance structure is

$$\mathbf{V} = \mathbf{X}_1\mathbf{X}_1^T\sigma_G^2 + \mathbf{X}_{\setminus 1}\mathbf{X}_{\setminus 1}^T\sigma_G^2 + \mathbf{I}\sigma_\epsilon^2 \quad (17)$$

With orthogonal covariates, the estimated variance components that compose \mathbf{V} can remain constant when testing each marker in turn. However, with collinearity among markers the situation becomes more complex. Below, we first describe the impact of multicollinearity on ridge regression estimates. We then outline the equivalence of a ridge regression and a mixed linear model, before then demonstrating increased variance of the estimates obtained from Eq. (16) under multicollinearity. Finally, we then go on to show that estimates from BayesR are less subject to inflated variance, except under extensive multicollinearity, before then describing how extending the model to provide minor allele frequency and LD specific hyperparameters provides estimates with improved properties across a range of underlying generative data models.

In Eq. (16) if markers were all simply estimated as random, following a single distribution, then a ridge regression estimator of Hoerl and Kennard 1970 [24] would be obtained, which was proposed to replace $\mathbf{X}^T\mathbf{X}$ in the OLS solutions by $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$, with $\lambda \in [0, \infty]$ a tuning or penalty parameter. This gives the ridge regression estimator

$$\hat{\boldsymbol{\beta}}(\lambda) = [\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}]^{-1}\mathbf{X}^T\mathbf{Y} \quad (18)$$

where λ is strictly positive and the solution or regularization path of the ridge estimate $\hat{\boldsymbol{\beta}}(\lambda) : \lambda \in [0, \infty]$ is the set of ridge estimates across the values of λ . The expectation of the ridge estimator

$$\begin{aligned} \mathbb{E}[\hat{\boldsymbol{\beta}}(\lambda)] &= \mathbb{E}[(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}] \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbb{E}(\mathbf{Y}) \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} \end{aligned} \quad (19)$$

with $\hat{\boldsymbol{\beta}}$ the maximum likelihood OLS estimator. If we consider an orthonormal design matrix \mathbf{X} , with $\mathbf{X}^T\mathbf{X} = \mathbf{I} = (\mathbf{X}^T\mathbf{X})^{-1}$ then we can express the relationship between $\hat{\boldsymbol{\beta}}$, and the ridge estimator, $\hat{\boldsymbol{\beta}}(\lambda)$, as

$$\begin{aligned} \hat{\boldsymbol{\beta}}(\lambda) &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (\mathbf{I} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (1 + \lambda\mathbf{I})^{-1}\mathbf{I}\mathbf{X}^T\mathbf{Y} \\ &= (1 + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (1 + \lambda\mathbf{I})^{-1}\hat{\boldsymbol{\beta}} \end{aligned} \quad (20)$$

If we define $\mathbf{W}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X})$ then the ridge estimator $\hat{\boldsymbol{\beta}}(\lambda)$ can be expressed as $\mathbf{W}_\lambda\hat{\boldsymbol{\beta}}$ for

$$\begin{aligned} \mathbf{W}_\lambda\hat{\boldsymbol{\beta}} &= \mathbf{W}_\lambda(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= [(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})]^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \hat{\boldsymbol{\beta}}(\lambda) \end{aligned} \quad (21)$$

The variance of the ridge estimator is then

$$\begin{aligned}
\text{Var}[\hat{\beta}(\lambda)] &= \text{Var}[\mathbf{W}_\lambda \hat{\beta}] \\
&= \mathbf{W}_\lambda \text{Var}[\hat{\beta}] \mathbf{W}_\lambda^T \\
&= \sigma_\epsilon^2 \mathbf{W}_\lambda (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{W}_\lambda^T \\
&= \sigma_\epsilon^2 (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1}
\end{aligned} \tag{22}$$

and the mean square error of $\hat{\beta}(\lambda)$ is

$$\begin{aligned}
\text{MSE}[\hat{\beta}(\lambda)] &= \mathbb{E}[(\mathbf{W}_\lambda \hat{\beta})^T (\mathbf{W}_\lambda \hat{\beta})] \\
&= \mathbb{E}[\hat{\beta}^T \mathbf{W}_\lambda^T \mathbf{W}_\lambda \hat{\beta}] - \mathbb{E}[\beta^T \mathbf{W}_\lambda \hat{\beta}] - \mathbb{E}[\hat{\beta}^T \mathbf{W}_\lambda^T \beta] + \mathbb{E}[\beta^T \beta] \\
&= \mathbb{E}[\hat{\beta}^T \mathbf{W}_\lambda^T \mathbf{W}_\lambda \hat{\beta}] - \mathbb{E}[\beta^T \mathbf{W}_\lambda^T \mathbf{W}_\lambda \hat{\beta}] - \mathbb{E}[\hat{\beta}^T \mathbf{W}_\lambda^T \mathbf{W}_\lambda \beta] + \mathbb{E}[\beta^T \mathbf{W}_\lambda^T \mathbf{W}_\lambda \beta] \\
&\quad - \mathbb{E}[\beta^T \mathbf{W}_\lambda^T \mathbf{W}_\lambda \beta] + \mathbb{E}[\beta^T \mathbf{W}_\lambda^T \mathbf{W}_\lambda \hat{\beta}] + \mathbb{E}[\hat{\beta}^T \mathbf{W}_\lambda^T \mathbf{W}_\lambda \beta] \\
&\quad - \mathbb{E}[\beta^T \mathbf{W}_\lambda \beta] - \mathbb{E}[\hat{\beta}^T \mathbf{W}_\lambda^T \beta] - \mathbb{E}[\beta^T \beta] \\
&= \mathbb{E}[(\hat{\beta} - \beta)^T \mathbf{W}_\lambda^T \mathbf{W}_\lambda (\hat{\beta} - \beta)] \\
&\quad - \beta^T \mathbf{W}_\lambda^T \mathbf{W}_\lambda \beta + \beta^T \mathbf{W}_\lambda^T \mathbf{W}_\lambda \beta + \beta^T \mathbf{W}_\lambda^T \mathbf{W}_\lambda \beta - \beta^T \mathbf{W}_\lambda \beta - \beta^T \mathbf{W}_\lambda \beta + \beta^T \beta \\
&= \mathbb{E}[(\hat{\beta} - \beta)^T \mathbf{W}_\lambda^T \mathbf{W}_\lambda (\hat{\beta} - \beta)] + \beta^T (\mathbf{W}_\lambda - \mathbf{I})^T (\mathbf{W}_\lambda - \mathbf{I}) \beta \\
&= \sigma_\epsilon^2 \text{tr}[\mathbf{W}_\lambda (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{W}_\lambda^T] + \beta^T (\mathbf{W}_\lambda - \mathbf{I})^T (\mathbf{W}_\lambda - \mathbf{I}) \beta
\end{aligned} \tag{23}$$

The first summand is the sum of the variances of the ridge estimator, while the second summand is the squared bias of the ridge estimator. With an orthonormal design matrix, \mathbf{X} , Theorem 2 of Theobald 1974 [25] shows:

$$\text{MSE}[\hat{\beta}(\lambda)] = \frac{p\sigma_\epsilon^2}{(1+\lambda)^2} + \frac{\lambda^2}{(1+\lambda)^2} \beta^T \beta \tag{24}$$

which achieves a minimum at $\lambda = p\sigma_\epsilon^2/\beta^T \beta = \sigma_\epsilon^2/\sigma_\beta^2$, with σ_β^2 the variance of the β coefficients. This has been stated in the genetics literature as the optimal shrinkage parameter [26] for a ridge regression. However, this is derived under the assumption of uncorrelated covariates within the design matrix \mathbf{X} .

To explore the effects of correlated covariates we use the ridge loss function, defined as

$$\mathcal{L}_{\text{ridge}}(\beta; \lambda) = \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 = \sum_{i=1}^n (Y_i - \mathbf{X}_i \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2 \tag{25}$$

which is the sums-of-squares with a penalty, $\lambda \sum_{j=1}^p \beta_j^2$, referred to as the ridge penalty, which shrinks the regression coefficients towards zero. The radius of the ridge constraint, the squared Euclidean norm of β , $\|\beta\|_2^2$, depends upon λ , \mathbf{X} and \mathbf{Y} , and taking its expectation

$$\begin{aligned}
\mathbb{E}[\|\hat{\beta}(\lambda)\|_2^2] &= \mathbb{E}[(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} (\mathbf{X}^T \mathbf{X}) \hat{\beta}]^T (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} (\mathbf{X}^T \mathbf{X}) \hat{\beta}] \\
&= \mathbb{E}[\mathbf{Y}^T \mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-2} \mathbf{X}^T \mathbf{Y}] \\
&= \sigma_\epsilon^2 \text{tr}[\mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-2} \mathbf{X}^T] + \beta \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-2} \mathbf{X}^T \mathbf{X} \beta
\end{aligned} \tag{26}$$

provides a measure that can be evaluated given different properties of the design matrix \mathbf{X} . With the same λ and the same β , Eq. (26) shows that the degree of collinearity among the covariates alters the variance of the estimated effects. Thus, in a ridge regression penalization does not remove collinearity but simply reduces its effects on the variance of the ridge estimator provided that the λ value is sufficiently large (and thus the σ_β^2 is small). We explore Eq. (26) in a simulation study described below and presented in Figure S1, Figure 1 and Figure 2. This theory is an extension of previous work [27] which showed that the inflation of the SNP heritability is proportional to a ratio of the average LD among causal variants and the markers and the average LD among all the markers, with inflation expected when causal variants are in higher LD with the markers than on average. Eq. (26) is a function of $\mathbf{X}^T \mathbf{X}$, with the LD values the off-diagonal elements in $\mathbf{X}^T \mathbf{X}$, but it suggests that inflation would be irrespective of the average LD across the genome, simply being expected if high-LD markers had strong effects and showing that inflation would occur only for the estimates of markers that are in LD with those causal variants. Thus, if SNP heritability is allocated across SNPs at random then estimation will on average be correct, irrespective of the LD among SNPs. If the effects of SNPs

vary according to the MAF or LD of the SNP, and assumptions are made that all SNP effects are sampled from the same distribution, then this will lead to bias as the estimates at high-LD markers in strong LD with underlying causal variants will be inflated and this inflation will be sufficiently large and occur at a sufficient number of genomic locations so as to impact upon the global estimate of SNP heritability.

This issue has been detected, and demonstrated in simulation, in a number of recent papers [28–31]. However, to date it has remained little understood from a theoretical perspective. The LD-MAF corrections proposed in the literature all serve to alter the lambda value for SNPs, or sets of SNPs, so that it becomes proportional to the LD and MAF of the marker, in essence reducing the σ_G^2 , or making it more specific to the markers in question, and increasing the λ value for common, highly correlated covariates. The equivalence of ridge regression and mixed-linear models has been shown many times, using well-established results from prediction of random variables dating back to Henderson [32]. The model $\mathbf{Y} = \mathbf{g} + \boldsymbol{\epsilon}$, with \mathbf{g} the genetic value of the individuals, and the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\mathbf{g} = \mathbf{X}\boldsymbol{\beta}$, $\mathbf{g} \sim N(0, \mathbf{X}\mathbf{X}^T\sigma_G^2)$ with marker effects thus $\boldsymbol{\beta} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{g}$, are equivalent. Following Henderson [32], assuming σ_ϵ^2 and σ_G^2 are known, with no fixed effect component, the log-likelihood can be shown to be proportional to:

$$\sigma_\epsilon^{-2}\|\mathbf{Y} - \mathbf{g}\|_2^2 + \mathbf{g}^T\mathbf{I}\sigma_G^2\mathbf{g} \quad (27)$$

equating the partial derivatives of this mixed model loss function with respect to \mathbf{g} to zero, yields the estimating equations known as Henderson’s mixed model equations. Returning to the mixed linear association model described in Eq.(16), using \mathbf{u} to denote the marker effects estimated as random, β for the focal marker effect estimated as fixed, and assuming independent marker effects, Henderson’s mixed model equations (MME) take the form:

$$\begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_{\setminus 1} \\ \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_1 + \mathbf{I}\lambda & \mathbf{X}_1^T\mathbf{X}_{\setminus 1} \\ \mathbf{X}_{\setminus 1}^T\mathbf{X}_1 & \mathbf{X}_{\setminus 1}^T\mathbf{X}_1 & \mathbf{X}_{\setminus 1}^T\mathbf{X}_{\setminus 1} + \mathbf{I}\lambda \end{bmatrix} \begin{bmatrix} \beta_1 \\ u_1 \\ \mathbf{u}_{\setminus 1} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^T\mathbf{y} \\ \mathbf{X}_1^T\mathbf{y} \\ \mathbf{X}_{\setminus 1}^T\mathbf{y} \end{bmatrix} \quad (28)$$

where $\lambda = \frac{\sigma_\epsilon^2}{\sigma_\beta^2}$. Subtracting the u_1 from the β equations gives $u_1 = 0$ and thus the MME reduce to:

$$\begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_{\setminus 1} \\ \mathbf{X}_{\setminus 1}^T\mathbf{X}_1 & \mathbf{X}_{\setminus 1}^T\mathbf{X}_{\setminus 1} + \mathbf{I}\lambda \end{bmatrix} \begin{bmatrix} \beta_1 \\ \mathbf{u}_{\setminus 1} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^T\mathbf{y} \\ \mathbf{X}_{\setminus 1}^T\mathbf{y} \end{bmatrix} \quad (29)$$

This has been derived previously [33], however there is an explicit assumption that any estimation error of the random marker effect estimates go into the residual and does not influence the fixed estimate of the marker. For the random effect component, the equivalence with the ridge regression estimator of Eq.(18) is evident, as is the equivalence of Eq. (27) with Eq. (25) above. Thus an MLMAi model returns “ridge regression” estimate of the marker effects, and as we show above ridge regression estimates are inflated when effect sizes are higher for high LD markers. It then follows that mixed model effect size estimates could be biased when effect sizes are higher for high LD markers.

Seen in this light, we can now explore the influence of multicollinearity on the BayesR dirac spike and slab model described above and compare it to that of a ridge regression. If we denote a measure of fit, such as the ridge loss function described above, being composed of $l(\beta)$ and a penalty function $pen_\lambda(\beta)$, then from a Bayesian perspective these correspond to the negative logarithms of the likelihood and the prior distribution, respectively. We can parameterize the BayesR dirac spike and slab model described above using the latent indicator of each SNP, j , $\boldsymbol{\gamma} = (\gamma_j, \dots, \gamma_p)^T$ with $\gamma_{j,l} = 0$ or 1, indicating whether or not the effect of SNP j follows a normal distribution with variance σ_l^2 ($l = 1, 2, 3, 4$). Then $p(\gamma_{j,l} = 1|\pi_l) = \pi_l$ and the prior distribution of each SNP effect β_j conditional on the indicator $\gamma_{j,l}$ is

$$f(\beta_j|\gamma_{j,l}) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma_l^2}}\exp(-\frac{\beta_j^2}{2\sigma_l^2}), & \text{if } \gamma_{j,l} = 1 \quad (l = 2, 3, 4) \\ \delta_0(\beta_j), & \text{if } \gamma_{j,l} = 0 \end{cases} \quad (30)$$

The joint distribution $p(\beta_j, \boldsymbol{\gamma}_j)$ conditional on π_β is

$$\begin{aligned} f(\beta_j, \boldsymbol{\gamma}_j|\pi_\beta, \sigma_\beta^2) &= \prod_{l=1}^4 f(\beta_j|\gamma_{j,l}) f(\gamma_{j,l} = 1|\pi_l) \\ &= (\delta_0(\beta_j)\pi_1)^{\gamma_{j,1}} \prod_{l=2}^4 \left(\frac{1}{\sqrt{2\pi\sigma_l^2}}\exp(-\frac{\beta_j^2}{2\sigma_l^2})\pi_l \right)^{\gamma_{j,l}} \end{aligned} \quad (31)$$

to simplify the following, we assume only a single normal distribution with $\pi_1 + \pi_2 = 1$ and we redefine the regression coefficient as $\beta_j = \gamma_j \alpha_j$ with $\alpha_j | \sigma_\beta^2 \sim N(0, \sigma_\beta^2)$. then:

$$\begin{aligned} f\left(\alpha_j, \gamma_j | \pi_\beta, \sigma_\beta^2\right) &= (\delta_0(\alpha_j) \pi_1)^{\gamma_{j,1}} \left(\frac{1}{\sqrt{2\pi\sigma_\beta^2}} \exp\left(-\frac{\alpha_j^2}{2\sigma_\beta^2}\right) \pi_1 \right)^{\gamma_{j,2}} \\ &= \pi_1^{\gamma_{j,1}} (1 - \pi_1)^{\gamma_{j,2}} \frac{1}{\sqrt{2\pi\sigma_\beta^2}} \exp\left(-\frac{\alpha_j^2}{2\sigma_\beta^2}\right) \end{aligned} \quad (32)$$

Now as above, if we define an active set of markers, $\mathbf{X}_{\gamma \neq 0}$, as those columns of \mathbf{X} where $\beta_{\gamma \neq 0}$, with an active set of γ , and $\|\gamma\|_0 = \sum_{j=1}^p \gamma_j$ be its cardinality. The joint prior on the vector γ, α then factorizes across all the markers as

$$\begin{aligned} f\left(\alpha, \gamma | \pi_\beta, \sigma_\beta^2\right) &= \prod_{j=1}^p f\left(\alpha_j, \gamma_j | \pi_\beta, \sigma_\beta^2\right) \\ &= \pi_1^{\|\gamma\|_0} (1 - \pi_1)^{p - \|\gamma\|_0} (2\pi\sigma_\beta^2)^{-\frac{p}{2}} \exp\left\{-\frac{1}{2\sigma_\beta^2} \sum_{j=1}^p \alpha_j^2\right\} \end{aligned} \quad (33)$$

as above we can express the likelihood in terms of γ, α as

$$f(y | \gamma, \alpha, \pi_\beta, \sigma_\epsilon) = (2\pi\sigma_\epsilon^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma_\epsilon^2} \|y - \mathbf{X}_{\gamma \neq 0} \alpha_{\gamma \neq 0}\|_2^2\right\} \quad (34)$$

and then under this reparamterisation the posterior is given as

$$\begin{aligned} f(\alpha, \gamma | \pi_\beta, \sigma_\beta^2, \sigma_\epsilon^2, y) &\propto f(\alpha, \gamma | \pi_\beta, \sigma_\beta^2) f(y | \gamma, \alpha, \pi_\beta, \sigma_\epsilon) \\ &\propto \exp\left\{\frac{1}{2\sigma_\epsilon^2} \|y - \mathbf{X}_{\gamma \neq 0} \alpha_{\gamma \neq 0}\|_2^2 - \frac{1}{2\sigma_\beta^2} \|\alpha\|_2^2 - \log\left(\frac{1 - \pi_1}{\pi_1}\right) \|\gamma\|_0\right\} \end{aligned} \quad (35)$$

The regularized maximum a posterior estimator is equivalent to minimising over γ, α the least squares objective function as

$$\min_{\gamma, \alpha} \|y - \mathbf{X}_{\gamma \neq 0} \alpha_{\gamma \neq 0}\|_2^2 + \lambda \|\alpha\|_2^2 + 2\sigma_\epsilon^2 \log\left(\frac{1 - \pi_1}{\pi_1}\right) \|\gamma\|_0 \quad (36)$$

In comparison to the ridge loss function described above, the first two terms are very similar and the third term imposes a sparsity constraint on the model. The term $\lambda \|\alpha\|_2^2$ has the same expectation as in Eq. (26) but with \mathbf{X} replaced with $\mathbf{X}_{\gamma \neq 0}$. To give some insight into the influence of collinearity on $\mathbb{E}[\|\gamma\|_0]$ and on the active set, we explore a two SNP scenario.

In a single site updating scheme, the probability that the first marker enters the model is given by Eq. 10. We seek to derive the probability that the second marker enters the model conditional on the first marker being in the model. We consider a scenario where we observe our standardised outcome $\tilde{\mathbf{y}}_c$ and two correlated predictors \mathbf{X}_1 and \mathbf{X}_2 . We assume that $\tilde{\mathbf{y}}_c, \mathbf{X}_1$ and \mathbf{X}_2 are scaled with zero mean and unit variance. We can then derive the partial least squares regression for $\tilde{\mathbf{y}}_c$ regressed on \mathbf{X}_2 , adjusting for \mathbf{X}_1 . If $\beta_{x_1, \tilde{y}_c} = \frac{\mathbf{X}_1^T \tilde{\mathbf{y}}_c}{\Sigma_{1,1}}$, with $\Sigma_{1,1} = \mathbf{X}_1^T \mathbf{X}_1 + \lambda_1 \mathbf{I}$, then a residual vector $\epsilon_{y_c, X_1} = \tilde{\mathbf{y}}_c - \beta_{x_1, \tilde{y}_c} \mathbf{X}_1$ is the vector left after backfitting β_{x_1, \tilde{y}_c} and we define $\epsilon_{X_1, X_2} = \mathbf{X}_2 - \rho_{X_1, X_2} \mathbf{X}_1$ as the additional information in X_2 left to fit $\beta_{x_2, \epsilon_{y_c, X_1}}$, with ρ_{X_1, X_2} the correlation of X_1 and X_2 . The correlation between the two residuals ϵ_{y_c, X_1} and ϵ_{X_1, X_2} can be used to estimate $\beta_{x_2, \epsilon_{y_c, X_1}}$, since $\beta_{x_2, \epsilon_{y_c, X_1}} = \frac{N}{\Sigma_{1,1}} \rho_{\epsilon_{y_c, X_1}, \epsilon_{X_1, X_2}}$. The correlation is a ratio between a covariance and a variance as

$$\begin{aligned} Cov_{\epsilon_{y_c, X_1}, \epsilon_{X_1, X_2}} &= \frac{1}{N} \sum (\tilde{\mathbf{y}}_c - \beta_{x_1, \tilde{y}_c} \mathbf{X}_1) (\mathbf{X}_2 - \rho_{X_1, X_2} \mathbf{X}_1) \\ &= \frac{1}{N} \sum (\tilde{\mathbf{y}}_c X_2 - \rho_{x_1, x_2} X_1 \tilde{y}_c - \beta_{x_1, \tilde{y}_c} X_1 X_2 + N \beta_{x_1, \tilde{y}_c} \rho_{X_1, X_2}) \\ &= \rho_{\epsilon_{y_c, X_2}} - \rho_{X_1, X_2} \beta_{x_1, \tilde{y}_c} \frac{\Sigma_{1,1}}{N} - \beta_{x_1, \tilde{y}_c} \rho_{X_1, X_2} + \beta_{x_1, \tilde{y}_c} \rho_{X_1, X_2} \\ &= \rho_{\epsilon_{y_c, X_2}} - \rho_{X_1, X_2} \beta_{x_1, \tilde{y}_c} \frac{\Sigma_{1,1}}{N} \\ &= \rho_{\epsilon_{y_c, X_2}} - \rho_{X_1, X_2} \frac{1}{N} X_1 \tilde{y}_c \end{aligned} \quad (37)$$

The variance in the correlation denominator is $S_{\epsilon_{X_1, X_2}}^2 = 1 - \rho_{X_1, X_2}^2$ which gives

$$\beta_{y_c, X_2 | X_1} = \frac{N}{\Sigma_{2,l}} \times \frac{\rho_{\epsilon_{y_c, X_2}} - \rho_{X_1, X_2} \frac{1}{N} X_1 \tilde{y}_c}{1 - \rho_{X_1, X_2}^2} \quad (38)$$

Eq. 38 can then be used in Eq. 9 and Eq. 10 to determine the posterior inclusion probability of the second covariate conditional on the first covariate being in the model. From this, the expectation, $\mathbb{E}[|\gamma|_0]$ for a two SNP scenario is then

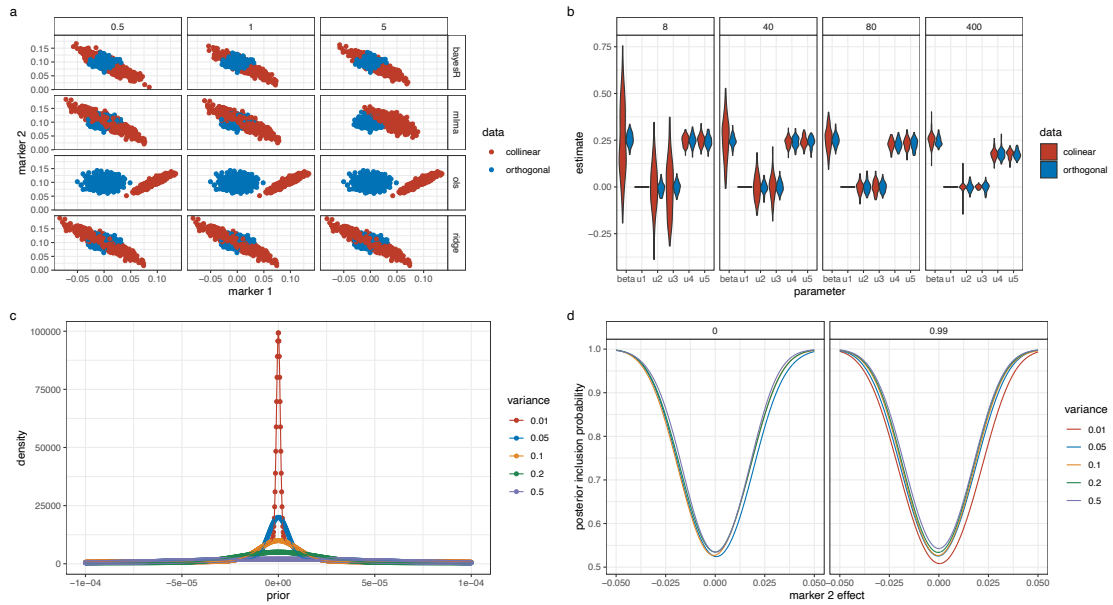
$$\begin{aligned} \mathbb{E}[|\gamma|_0] &= p(l_1 = 1 | \boldsymbol{\theta}, \mathbf{y}) + p(l_2 = 1 | \boldsymbol{\theta}, \mathbf{y}) \\ &= \frac{1}{1 + \exp \left[\log(\pi_0) - \left(-\frac{1}{2} [-\log(|\lambda \Sigma_{1,1}^{-1}|) - \left(\frac{\beta_{y_c, X_1}^2 \Sigma_{1,1}}{\sigma_\beta^2} \right)] \right) \right]} \\ &\quad + \frac{1}{1 + \exp \left[\log(\pi_0) - \left(-\frac{1}{2} [-\log(|\lambda \Sigma_{2,1}^{-1}|) - \left(\frac{\beta_{y_c, X_2 | X_1}^2 \Sigma_{2,1}}{\sigma_\beta^2} \right)] \right) \right]} \end{aligned} \quad (39)$$

With the dirac spike and slab and ridge regression estimators minimizing the same sum-of-squares, the key difference with the constrained estimation formulation of ridge regression is not in the explicit form of λ but in what is bounded the domain of acceptable values for α . For the BayesR estimator the domain is specified by a bound on the ℓ_0 norm of the regression parameter, while for its ridge counterpart the bound is applied to the squared ℓ_2 norm of β . Multicollinearity will reduce the likelihood of the second covariate entering the model as it's inclusion is dependent upon ρ_{X_1, X_2} the correlation among covariates and $\rho_{\epsilon_{y_c, X_2}}$ the correlation of the second marker and the residual vector after backfitting the first marker. This will limit the range of possible estimates to be lower than those obtained from ridge regression, reducing inflation of $\lambda \|\alpha\|_2^2$ under high collinearity, but not entirely removing it. Due to the sampling of markers from a series of normal distributions, collinearity will still inflate $\lambda \|\alpha\|_2^2$, however, the degree to which this occurs will depend upon the number of correlated markers, the degree of correlation among them and the strength of the effects. Therefore, our aim here is not to derive a general solution predictive of all situations, merely it is to highlight that in order to make some inference as to the underlying distribution of genetic effects, it is required to extend the model as outlined in the following section.

Small-scale simulation example

While we assess the performance of our model in the large-scale simulation work, smaller-scale focused simulation work was also conducted to support and test the inference made. Our theory suggests that there will be increased variance of the regression coefficient estimates and, as a result, an inflated estimate of the phenotypic variance attributable to SNP markers under high multicollinearity for both mixed linear model approaches and a Dirac spike and slab mixture model. To create a toy example of this, we conducted a simulation study where for each of 50 replicates, we simulated 50 independent genomic regions, each containing two SNP markers. In each simulation replicate, we simulated values for 5,000 individuals at each of the 50 SNP marker pairs, by first simulating from a standard multivariate normal distribution with correlation set to either 0 or 0.99. From this, we obtained the integral from $-\infty$ to q of the probability density function, where q is the z-score of the values obtained for each individual from the multivariate normal. From these integrals, we then made two draws from the inverse of the cumulative density function of the binomial distribution to obtain the marker value for each individual, with frequency 0.3. This gave marker values (0, 1, or 2), with the pairs of SNPs having either all LD = 0, or all LD = 0.99. For each of the 50 pairs of SNPs, we assigned effect size 0 to the first marker and 0.1 to the second marker. We then scaled the SNP markers to zero mean and unit variance and multiplied the markers by the effect sizes to obtain the genetic values for the 5,000 individuals, with variance 0.5. We then simulated the environmental component of the phenotype from a normal distribution with zero mean and variance 0.5 and then created a phenotype as the sum of the genetic values and the environmental values, with zero mean and unit variance.

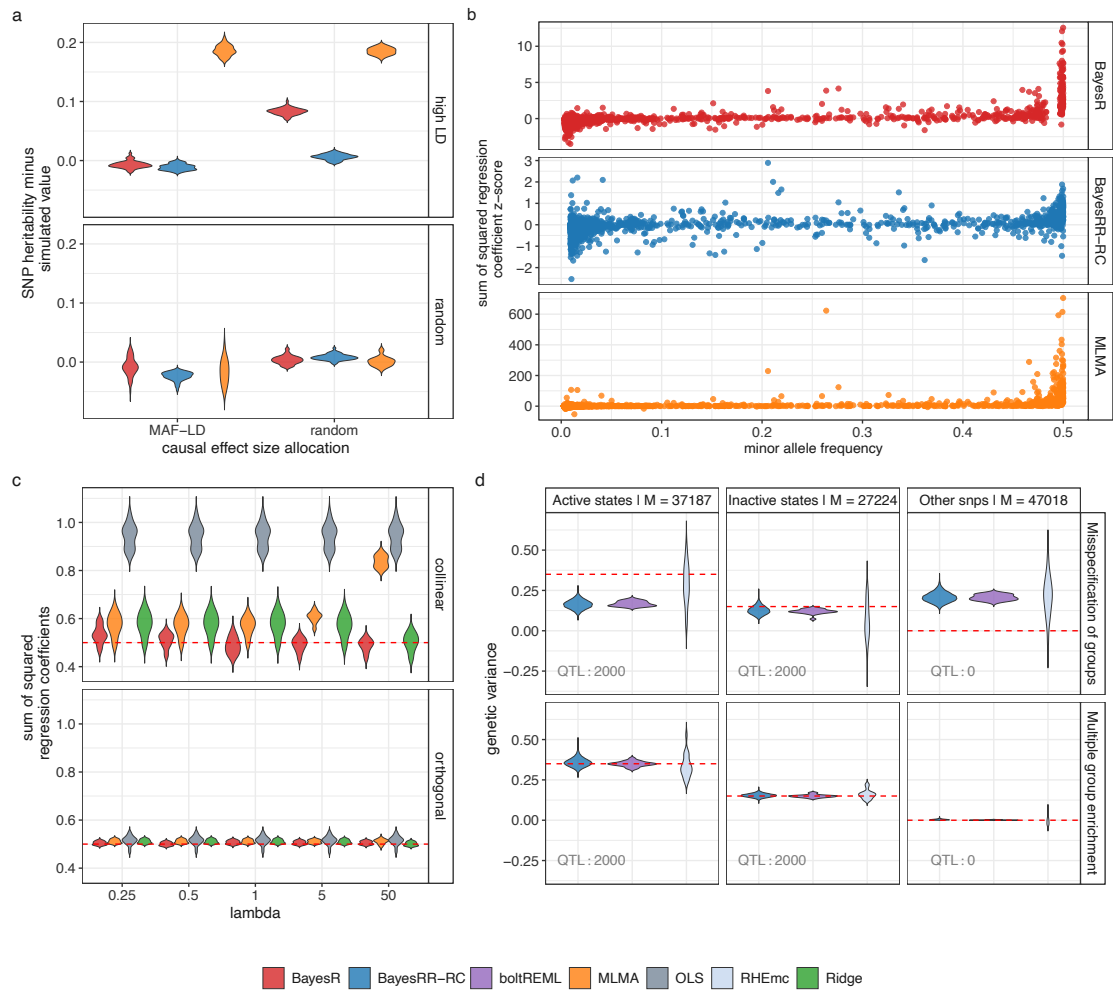
We then analysed these 50 data sets using different methods of single-marker OLS regression (OLS), mixed-linear model association (MLMA), ridge regression (Ridge), and a Dirac spike and slab mixture of regressions model (BayesR), all of which are described above. For the frequentist approaches, we directly solved the estimation equations, scaling the SNP markers to have zero mean and unit variance. For BayesR we sampled the effects for 5000 iterations, with burn-in period of 2000 iterations to obtain the posterior mean effect sizes, again scaling the SNP markers to zero mean and unit variance. We repeated these analyses many times, each time fixing the estimated phenotypic variance attributable to the markers σ_G^2 to be a



Supplementary Figure 17. Theory and simulation study of SNP marker model parameters. (a) accompanies Eq. (26) and shows the distribution of the point estimates of the effect sizes of two correlated markers of effect size (0,0.1) under orthogonality (LD = 0) and collinearity (LD = 0.99) across 2500 replicates (50 independent genomic regions for 5,000 individuals within each of 50 replicates) for a range of different models: a dirac spike and slab mixture of regressions model (bayesR), a mixed linear association model (MLMA), single-marker ordinary least squares (OLS), and ridge regression (Ridge). Panels give the lambda shrinkage parameter of the model, the error variance divided by the phenotypic variance attributable to the SNP markers, showing that as lambda decreases the variation of the estimates increases under multicollinearity. (b) accompanies Eq.(29) and shows the marker estimates obtained from Henderson’s mixed model equations for a MLMA with the focal marker as fixed (beta) and random (u1), with four other markers in the model. Markers were either uncorrelated (orthogonal, LD=0) or the focal marker was correlated with the first two out of the four other markers (collinear, LD=0.99). Panels give the lambda shrinkage parameter, showing that as lambda decreases the variation of the estimates increases under multicollinearity. (c) shows the prior density of the BayesR model for different hyperparameter values of the phenotypic variance attributable to genetic effects (variance), showing that as the variance attributable to the markers decreases, the prior has higher mass around zero. Thus, with a grouped mixture of regressions model (BayesRR-RC), each hyperparameter estimate will be smaller and thus there will be higher prior density around zero. This then has consequences for marker inclusion in the BayesRR model. Higher prior mass around zero makes little difference for the inclusion of uncorrelated markers, but it results in reduced posterior inclusion probability for correlated markers as shown in (d). For (d), we calculated the inclusion probability (PIP) of two markers with LD = 0 and LD = 0.99, as the variance attributable to the SNP markers, and thus the prior distribution, changes assuming a background inclusion probability of 0.1, a sample size of 5000, and an effect size of 0.01 SD for marker 1 (see Methods). (d) shows that the PIP of the second marker is reduced across a range of possible effect size values (the average of 1000 replicated simulations for 1000 marker 2 effect values for each line) as the hyperparameter estimate decreases, and thus the smaller hyperparameter estimates in a BayesRR model means that correlated markers are less likely to enter the model, controlling better for the effects of multicollinearity.

different value. We selected (2, 1, 0.5, 0.1, and 0.01) and fixed the residual variance σ_ϵ^2 to be 0.5, to give different lambda values $\lambda = \frac{\sigma_\epsilon^2}{\sigma_G^2}$, giving $\lambda = 0.25, 0.5, 1, 5, \text{ and } 50$. Our aim here was to explore the pattern of effect sizes that we obtain under these λ values. So first, we plotted the effect sizes obtained for each of the 50 SNP pairs obtained across the 50 simulation replicates in Supplementary Figure 17a, to show the differences in the variance of the estimates obtained across approaches when the pairs of SNP markers were orthogonal (LD=0), or collinear (LD=0.99), under different lambda values. Second, we then plot the distribution of the sum of the squared regression coefficients in Supplementary Figure 18c across approaches, when the pairs of SNP markers were orthogonal (LD=0), or collinear (LD=0.99), under different lambda values, where the expectation is 0.5 (sum of the 50 squared 0.1 SD effect sizes). This simulation confirmed, that regression coefficient estimates have higher variance under multicollinearity, resulting in inflation of the sum of the squared coefficient estimates for all approaches when the variation attributable to SNP markers is overestimated, resulting in a reduction in the lambda values.

We then further explored the performance of the MLMA and BayesR models under multicollinearity to (i)



Supplementary Figure 18. Theory and simulation study for genetic penalized regression models under multicollinearity. (a) Smaller-scale simulation study than that presented in the main text using real genomic data from chromosome 22 where 50 replicate phenotypes were generated by either allocating 5000 LD-independent causal variants to high LD SNPs (y-axis panel: high LD), or randomly allocating 5000 SNPs as causal variants (y-axis panel: random), and then either randomly allocating effect sizes to those SNPs (x-axis: random), or allocating effect sizes proportional to their LD and MAF (x-axis: MAF-LD, see Methods). In this simulation every LD block of chromosome 22 contributes to the trait variance. SNP heritability estimation error is plotted as the difference of the estimate and the true simulated value across the 50 replicates. (b) We then investigated this further for the scenario where causal variants are allocated to high-LD SNPs. While the 5000 causal variants are LD-independent, they are each correlated with a large number of SNPs of simulated effect size 0. For each causal variant, we took all the markers in $LD \geq 0.05$ and summed the squared estimated regression coefficients of these markers. The true simulated value is simply the square of the effect size allocated to the causal variant, and we subtracted this from the sum of the squared regression coefficients divided by the SD of the simulated genetic effects, to give a z-score for each causal variant and this is plotted on the y-axis for MLMA, BayesR, and BayesRR-RC. (c) Our theory outlines how this overestimation is the result of the effect of multicollinearity (see Methods) and an example is shown here, where 50 pairs of SNP markers with $LD = 0.9$ were simulated for each of 50 simulation replicates, where only one marker of each pair has an effect (0,0.1 SD), giving the sum of the squared regression coefficients as 0.5 for each simulation (dotted red line). lambda is the shrinkage parameter, the ratio of the error variance and the variance attributable to the SNP markers, used for MLMA, ridge regression (Ridge) and the BayesR model to estimate the effects. (d) Simulation of a genetic architecture (dotted red line) using real annotations from the Epigenome Roadmap Project [34] (active states, inactive states, other snps). We compared BayesRR-RC to other recent approaches providing annotation-specific variance component estimates in individual-level data when SNPs are randomly assigned to an annotation (labelled: misspecification of groups), or when specifying enrichment using prior knowledge (labelled: multiple group enrichment)

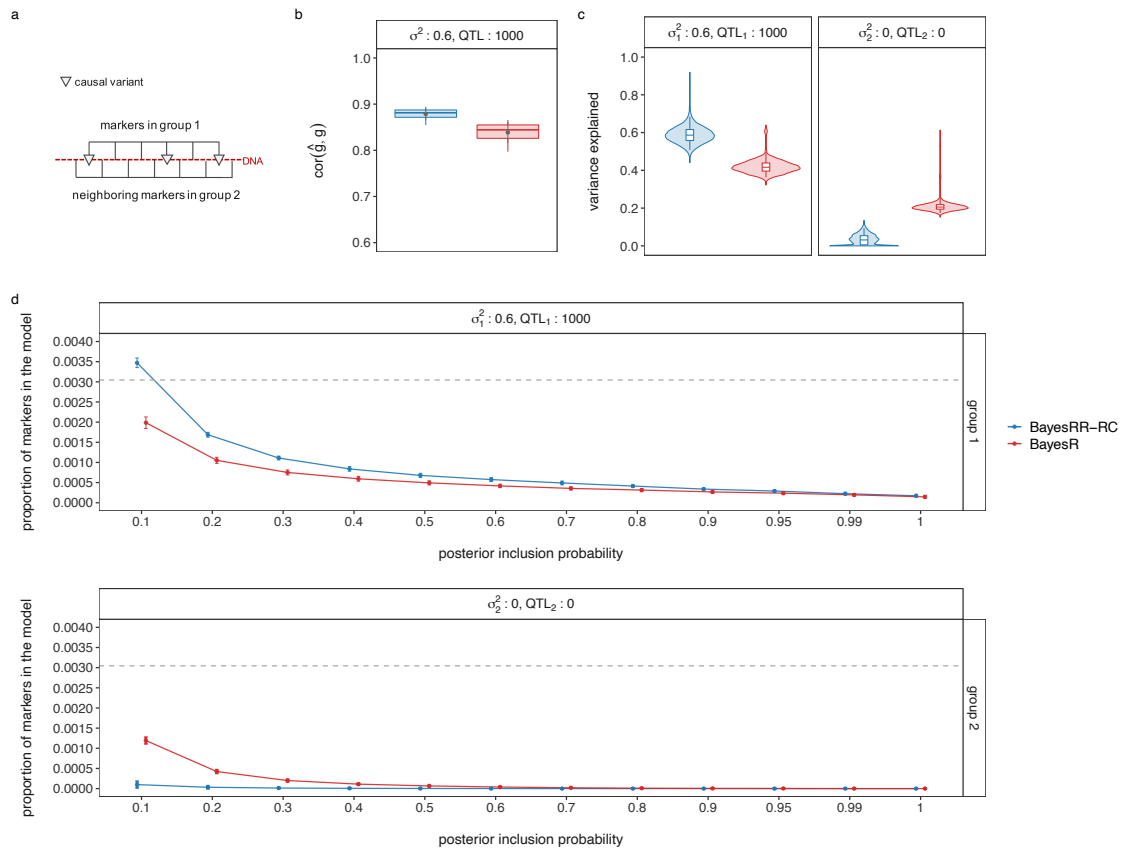
better understand the interplay between the fixed GLS estimate obtained and the random marker effects, and (ii) to better understand how the prior of the BayesR model changes with lambda and how this constrains the

inclusion probabilities of correlated markers. We first examined the influence of varying lambda and varying the collinearity of markers on the variation of the effect size estimates obtained from the Henderson’s mixed model equations, where one focal marker is estimated as fixed, and a further five markers are estimated as random, with LD between the markers estimated as fixed and random. To do this, we simulated five markers in the same manner as described above that were either (i) entirely orthogonal with $LD = 0$, or (ii) had $LD = 0.99$ among the first three markers, with the final two markers having $LD = 0$ with all others. We assigned effect sizes to the five markers as $\beta = (0.25, 0, 0, 0.25, 0.25)$, multiplied these effect sizes by the simulated marker values scaled to zero mean and unit variance to create the genetic values, and then added an environmental component simulated from a normal distribution with mean zero and variance 1 minus the variance of the genetic values (0.1875) to give a phenotype with zero mean and unit variance. We directly solved the Henderson’s mixed model equations, fixing the lambda value at different levels (the appropriate lambda from theory assuming orthogonal covariate would be $(1 - 0.1875)/0.1875 = 4.333$). We find that even with high shrinkage, a lambda value of almost 20 times greater than the theoretical orthogonal expectation is required to produce effect sizes under collinearity, with similar variance to those obtained under orthogonality (Supplementary Figure 17b).

For BayesR, we first explored the density of the posterior distribution by simulating draws from the prior as we change the variance attributable to the SNP markers. Supplementary Figure 17c shows these densities, revealing how the prior becomes strongly centred on zero and almost exponentially distributed as the variance becomes small. This is in contrast to the almost flat prior observed with high variance, which will do little to constrain effect size estimates toward zero. We then conducted 1000 simulation replicates of paired SNP markers for 10 different scenarios of variance attributable to the SNP markers of 0.01, 0.05, 0.1, 0.2, and 0.5, for pairs of SNPs with correlation of either 0 or 0.99. For each of these 10,000 data sets, we simulate a pair of SNPs for 5000 individuals, assuming error variance of 0.5, effect size for the first marker of 0.01 SD and then we simulated a sequence of 1000 different effect sizes from -0.05 to 0.05. Of these 10 million phenotypes and pairs of SNPs obtained, we then determine the posterior inclusion probability of the second marker, given that the first marker is in the model, with the effect size correctly estimated as 0.01, from the BayesR model derivations presented above. The lines presented in Supplementary Figure 17d go through the mean posterior inclusion probability of the second SNP marker across the 1000 simulation replicates, for each of the 1000 different effect sizes from -0.05 to 0.05 for marker 2, with a different colour for each scenario of the variance attributable to the SNP markers. The plot shows a reduction in the posterior inclusion probability of the second SNP marker as the variance attributable to the SNP markers decreases under multicollinearity. Thus, if the hyperparameter estimates of the variance contributed by markers is kept small, by having different hyperparameters for different groups of markers, then the BayesR model acts to constrain the inclusion of any additional correlated markers in the model.

Having confirmed our theory, we then conducted a further simulation study to replicate these observations using real genomic data. We randomly selected 50,000 individuals from the UK Biobank study data and used the imputed SNP data from chromosome 22 as supplied in the data release. We simulated phenotypes under contrasting generative models:

- We chose markers of high LD with other SNPs to be the causal variants and we assigned effects proportional to the LD score of those markers and their minor allele frequency. To do this, we first grouped the SNPs using the clumping procedure in Plink (see Code Availability) based on 1 - MAF, selecting the highest frequency variants and removing any variants with $LD < 0.01$, to obtain 4988 independent SNPs. For these 4988 SNPs we calculated the LD score of the markers. We then assigned effect sizes to these selected SNPs, drawing them from a single normal distribution with variance $\sim LD_score^1 MAF^{-1}$. We multiplied these effect sizes by the simulated marker values scaled to zero mean and unit variance to create the genetic values with variance 0.5, and then added an environmental component simulated from a normal distribution with mean zero and variance 1 minus the variance of the genetic values to give a phenotype with zero mean and unit variance.
- We then took the same 4988 SNPs but assigned effect sizes to the markers at random from a normal distribution with zero mean and variance 0.5/4988. We multiplied these effect sizes by the simulated marker values scaled to zero mean and unit variance to create the genetic values with variance 0.5, and then added an environmental component simulated from a normal distribution with mean zero and variance 1 minus the variance of the genetic values to give a phenotype with zero mean and unit variance.
- We then sampled randomly 4988 evenly spaced markers as causal variants, but assigned effect sizes proportional to the LD score and minor allele frequency of the markers as described above. We



Supplementary Figure 19. Classification power of BayesRR-RC. Grouping effects in a BayesRR-RC model improves the power of BayesR to estimate effect sizes and infer the genetic architecture of common complex traits and diseases. This setting compares 10 simulations of 5 chains with different starting values (chain length : 2500, burn-in : 500, thin : 5) executed using BayesRR-RC. (a) Each simulation has two groups in high LD with an interdigitated structure where one in two SNPs is assigned to group 1 and all genetic variance is assigned to group 1 with 1000 QTL. Annotation-specific estimates for BayesR are calculated post-analysis for each group. (b) Estimation of markers effects in an independent data set. BayesRR-RC improves on correlation between predicted and simulated genetic values. This increase in prediction implies that adding functional information to BayesR better fits the data and improves prediction accuracy. (c) Genetic variance and (d) proportion of markers entering the model at posterior inclusion probability (pip) thresholds summarized across 10 simulations for group 1 and group 2. The proportion of markers included in the model is closer to the truth (dotted grey line) when using BayesRR-RC compared to BayesR. Effects are thus more likely attributed to the correct group using our approach, which also explains why we estimate more accurately the group genetic variance compared to the baseline. Simulation setting: $N = 20,000$ unrelated European individuals from the UK Biobank, $M = 328,385$ markers (chromosome 2). Dots in box plots show the mean of the correlation between predicted and simulated genetic values.

multiplied these effect sizes by the simulated marker values scaled to zero mean and unit variance to create the genetic values with variance 0.6, and then added an environmental component simulated from a normal distribution with mean zero and variance 1 minus the variance of the genetic values to give a phenotype with zero mean and unit variance.

- Finally, we then sampled randomly 4988 evenly spaced markers as causal variants and randomly assigned the effect sizes from a normal distribution with zero mean and variance $0.5/4988$. We multiplied these effect sizes by the simulated marker values scaled to zero mean and unit variance to create the genetic values with variance 0.5, and then added an environmental component simulated from a normal distribution with mean zero and variance 1 minus the variance of the genetic values to give a phenotype with zero mean and unit variance.

This replicates our main simulation study, but creates a situation where there is an association at every LD block on chromosome 22 and thus the results seen in the main simulation study should be magnified

here. We analysed 50 simulation replicates of each of the four scenarios with BayesR, BayesRR-RC with 20 MAF-LD groups (deciles of MAF, each split into two groups based on median LD score within each MAF decile), and a MLMAi model implemented in software GCTA. For the Bayesian methods we ran three chains with different starting values for each of the 200 simulation replicates for 3000 iterations, removing the first 1500 iterations as burn-in and taking the posterior mean across the three chains. In Supplementary Figure 18a we plot the distribution of the posterior mean for BayesR and BayesRR-RC, and the MLMA point estimates, of the proportion of variance attributable to the SNP markers minus the true simulated value obtained across the 50 simulation replicates for each of the four scenarios, showing inflation of the MLMA estimates when selecting high LD variants, and inflation of the BayesR estimates with high LD and random effect size estimates. In contrast, estimates obtained from BayesRR-RC were unbiased across all scenarios. By simulating an effect size MAF relationship $\sim \text{LDscore}^1 \text{MAF}^{-1}$, we are assigning the smallest absolute effect size values to the most common SNPs, which appears to limit the inflation of the estimates for BayesR, when selecting high LD SNPs as causal variants (Supplementary Figure 18a). We then examined the effect size estimates obtained from these three approaches across the MAF spectrum under the second scenario of high LD causal variant selection, but random effect size allocation, to show using z-scores calculated as the estimated effects minus the simulated effects, divided by the SD of the simulated effects. We find overestimation of common variant effect sizes under BayesR, and dramatic inflation of effect size estimates under MLMA showing poor recovery of the underlying effect size distribution (Supplementary Figure 18b). Grouping effects by MAF and LD in a BayesRR-RC model resolved this overestimation issue (Supplementary Figure 18b) as seen in our original large-scale simulation study.

We then explore the ability of the model to recover a different set of annotation-specific effect sizes using the same set of 50,000 randomly selected UK Biobank individuals and imputed genotype data for chromosome 22 grouped by chromatin state annotations (15-state ChromHMM model) from the epigenome of primary mononuclear cells from peripheral blood (E062) of the Epigenome Roadmap Project [34]. We simulated the genetic architecture as follows :

- We first mapped SNPs to active and inactive chromatin states from the mnemonic bed files for E062 (see Code availability). 37,187 SNPs mapped to active chromatin states including transcription start site (TSS) and their flanking regions, genic and other enhancers, untranslated transcribed regions (UTR) and actively transcribed regions and zinc finger genes states. 27,224 SNPs mapped to inactive states including heterochromatin, bivalent/poised TSS and their flanking regions, bivalent enhancers and repressed polycomb states. The remaining 47,018 SNPs were grouped and labelled as Other SNPs.
- To simulate enrichment in both chromatin states, we randomly sampled 2000 SNPs as causal variants from variants mapped to active chromatin states and another 2000 SNPs from variants mapped to inactive chromatin states. We then assigned effect sizes to these 4000 selected SNPs, drawing them from a normal distribution with zero mean and variance $0.35/2000$ for active states and $0.15/2000$ for inactive states.
- We multiplied annotation-specific effect sizes by the simulated marker values scaled to zero mean and unit variance to create the annotation-specific genetic values with variance 0.35 for active states, 0.15 for inactive states and 0 for other SNPs. We finally added an environmental component simulated from a normal distribution with mean zero and variance 1 minus 0.5 (the sum of the genetic values) to give a phenotype with zero mean and unit variance.

We analyzed 20 simulation replicates with our BayesRR-RC software specifying annotations (active states, inactive states and other SNPs) with 2 LD groups based on median LD score within each annotation. We compared our software to BoltREML [35] and RHEmc [36] both multi-variance component methods that also use individual-level data but provide single heritability estimates per genetic component. For BayesRR-RC we ran three chains with different starting values for each of the 20 simulation replicates for 3000 iterations, removing the first 1000 iterations as burn-in and taking the posterior mean across the three chains. We then performed the same analysis but randomly assigning SNPs to each annotation resulting in mis-specification of the underlying genetic architecture. In Supplementary Figure 18d, we plot the estimated sum of the squared regression coefficients that is evenly split across the three annotations when misspecifying the underlying genetic architecture (labelled : Misspecification of groups) and shows enrichment when we properly assign SNPs to annotation (labelled : Multiple group enrichment). We find that BayesRR-RC performs as BoltREML and RHEmc, with RHEmc estimates showing higher variability, supporting our main simulation results.

We also further examined the ability of BayesRR-RC to recover effect sizes compared to BayesR by comparing 10 simulations of 5 chains with different starting values where each simulation has two groups in high LD with an interdigitated structure where one in two SNPs is assigned to group 1 (Supplementary Figure 19a). We then simulated phenotypes as previously described, randomly selecting 1000 causal variants in group 1 only, using 20,000 randomly selected UK Biobank individuals and imputed genotype data for chromosome 2 (with $\text{MAF} > 0.05$). In Supplementary Figure 19, we compare the proportion of markers entering the model in group 1 and group 2 at different posterior inclusion probability thresholds. Annotation-specific estimates for BayesR are calculated post-analysis for each group. We also compare the correlation of estimated genetic values with the truth when using BayesRR-RC and BayesR. For this, we conducted estimation of marker effects in an independent data set to compare prediction accuracy. We simulated 10 new phenotypes and computed the genetic value $\hat{g} = X\hat{\beta}$ where X is the genotype matrix and $\hat{\beta}$ is a vector of estimated marker effects for each individual. Supplementary Figure 19 shows BayesRR-RC has improved model performance over BayesR to recover effect sizes and infer underlying genetic architectures.

Supplementary References

1. Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.
2. Gertraud Malsiner-Walli and Helga Wagner. Comparing spike and slab priors for bayesian variable selection. *Austrian Journal of Statistics*, 40(4):241–264, Feb. 2016.
3. M. Erbe, B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. 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, 2020/05/10 2012.
4. Gerhard Moser, Sang Hong Lee, Ben J. Hayes, Michael E. Goddard, Naomi R. Wray, and Peter M. Visscher. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLOS Genetics*, 11(4):1–22, 04 2015.
5. Gemma E. Moran, Veronika Ročková, and Edward I. George. Variance prior forms for high-dimensional bayesian variable selection. *Bayesian Anal.*, 14(4):1091–1119, 12 2019.
6. Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):515–534, 2006.
7. Ismaël Castillo, Johannes Schmidt-Hieber, Aad Van der Vaart, et al. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.
8. Daniel Trejo Banos, Daniel L McCartney, Marion Patxot, Lucas Anchieri, Thomas Battram, Colette Christiansen, Ricardo Costeira, Rosie M Walker, Stewart W Morris, Archie Campbell, et al. Bayesian reassessment of the epigenetic architecture of complex traits. *Nature Communications*, 11(1):1–14, 2020.
9. Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. In *Readings in Computer Vision*, pages 564–584. Elsevier, 1987.
10. Yali Amit and Ulf Grenander. Comparing sweep strategies for stochastic relaxation. *Journal of Multivariate Analysis*, 37(2):197–222, 1991.
11. M. Goddard. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, 136:245 EP –, 08 2009.
12. Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.
13. Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.
14. Andrew Gelman, Daniel Lee, and Jiqiang Guo. Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40(5):530–543, 2015.
15. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
16. John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, 2:e55, apr 2016.
17. James Johndrow, Paulo Orenstein, and Anirban Bhattacharya. Scalable approximate mcmc algorithms for the horseshoe prior. *Journal of Machine Learning Research*, 21(73):1–61, 2020.

18. Bala Rajaratnam, Doug Sparks, Kshitij Khare, and Liyuan Zhang. Uncertainty quantification for modern high-dimensional regression via scalable bayesian methods. *Journal of Computational and Graphical Statistics*, 28(1):174–184, 2019.
19. Matthew Johnson, James Saunderson, and Alan Willsky. Analyzing hogwild parallel gaussian gibbs sampling. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2715–2723. Curran Associates, Inc., 2013.
20. Elaine Angelino, Matthew James Johnson, Ryan P Adams, et al. Patterns of scalable bayesian inference. *Foundations and Trends® in Machine Learning*, 9(2-3):119–247, 2016.
21. Daniel Gianola. Priors in whole-genome regression: The bayesian alphabet returns. *Genetics*, 194(3):573–596, 2013.
22. Matthew Stephens and David J. Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690, 2009.
23. Rohan Fernando, Ali Toosi, Anna Wolc, Dorian Garrick, and Jack Dekkers. Application of whole-genome prediction methods for genome-wide association studies: a bayesian approach. *Journal of Agricultural, Biological and Environmental Statistics*, 22(2):172–193, 2017.
24. Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.
25. C. M. Theobald. Generalizations of mean square error applied to ridge regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1):103–106, 1974.
26. Robert M. Maier, Zhihong Zhu, Sang Hong Lee, Maciej Trzaskowski, Douglas M. Ruderfer, Eli A. Stahl, Stephan Ripke, Naomi R. Wray, Jian Yang, Peter M. Visscher, and Matthew R. Robinson. Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nature Communications*, 9(1):989, 2018.
27. Jian Yang, Andrew Bakshi, Zhihong Zhu, Gibran Hemani, Anna AE Vinkhuyzen, Sang Hong Lee, Matthew R Robinson, John RB Perry, Ilja M Nolte, Jana V van Vliet-Ostaptchouk, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics*, 47(10):1114, 2015.
28. Luke M Evans, Rasool Tahmasbi, Scott I Vrieze, Gonçalo R Abecasis, Sayantan Das, Steven Gazal, Douglas W Bjelland, Teresa R De Candia, Michael E Goddard, Benjamin M Neale, et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature Genetics*, 50(5):737–745, 2018.
29. Doug Speed, Na Cai, Michael R Johnson, Sergey Nejentsev, David J Balding, UCLEB Consortium, et al. Reevaluation of snp heritability in complex human traits. *Nature Genetics*, 49(7):986, 2017.
30. Doug Speed, John Holmes, and David J Balding. Evaluating and improving heritability models using summary statistics. *Nature Genetics*, 52(4):458–462, 2020.
31. Kangcheng Hou, Kathryn S Burch, Arunabha Majumdar, Huwenbo Shi, Nicholas Mancuso, Yue Wu, Sriram Sankararaman, and Bogdan Pasaniuc. Accurate estimation of snp-heritability from biobank-scale data irrespective of genetic architecture. *Nature Genetics*, page 1, 2019.
32. C.R. Henderson. Best linear unbiased prediction of breeding values not in the model for records. *Journal of Dairy Science*, 60(5):783 – 787, 1977.
33. Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2):100–106, 2014.
34. Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317, 2015.

35. Po-Ru Loh, Gaurav Bhatia, Alexander Gusev, Hilary K Finucane, Brendan K Bulik-Sullivan, Samuela J Pollack, Teresa R de Candia, Sang Hong Lee, Naomi R Wray, Kenneth S Kendler, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature Genetics*, 47(12):1385, 2015.
36. Ali Pazokitoroudi, Yue Wu, Kathryn S Burch, Kangcheng Hou, Bogdan Pasaniuc, and Sriram Sankararaman. Scalable multi-component linear mixed models with application to snp heritability estimation. *bioRxiv*, page 522003, 2019.

Description of Additional Supplementary Files

File Name: Supplementary Data 1

Description: SNP partitioned into 13 annotation groups.

File Name: Supplementary Data 2

Description: SNP heritability attributable to each genomic annotation and phenotype.

File Name: Supplementary Data 3

Description: SNP heritability estimates from RHE-mc.

File Name: Supplementary Data 4

Description: SNP heritability estimates from stratified-LDSC.

File Name: Supplementary Data 5

Description: SNP heritability estimates from SumHer.

File Name: Supplementary Data 6

Description: Mean effect sizes of gene components for exons contributing to the phenotypic variance with > 95% probability.

File Name: Supplementary Data 7

Description: Mean effect sizes of gene components for introns contributing to the phenotypic variance with > 95% probability.

File Name: Supplementary Data 8

Description: Mean effect sizes of gene components for 1kb regions contributing to the phenotypic variance with > 95% probability.

File Name: Supplementary Data 9

Description: Mean effect sizes of gene components for cis regions contributing to the phenotypic variance with > 95% probability.

File Name: Supplementary Data 10

Description: Contribution of SNPs with posterior inclusion probability (PIP) > 0.95 to each phenotype and corresponding p-value from UKB GWAS summary statistics (see Code Availability).

Appendix B - Haematological changes from conception to childbirth: an indicator of major pregnancy complications.

This article (Patxot *et al.* 2022) is presented in Chapter 2.

Haematological changes from conception to childbirth: an indicator of major pregnancy complications.

Marion Patxot^{1,5,*}, Miloš Stojanov², Sven Erik Ojavee^{1,5}, Rosanna Pescini Gobert¹, Zoltán Kutalik^{1,4,5}, Mathilde Gavillet⁶, David Baud^{2,†}, Matthew R. Robinson^{3,*;†}

1 Department of Computational Biology, University of Lausanne, Lausanne, Switzerland.

2 Materno-fetal and Obstetrics Research Unit, Department of Obstetrics and Gynecology, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland.

3 Institute of Science and Technology Austria, Klosterneuburg, Austria.

4 University Center for Primary Care and Public Health, Lausanne, Switzerland.

5 Swiss Institute of Bioinformatics, Lausanne, Switzerland.

6 Service and Central Laboratory of Haematology, Department of Oncology and Department of Laboratories and Pathology, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland.

† denotes equal contribution.

*corresponding authors: marion.patxotbertran@unil.ch, matthew.robinson@ist.ac.at

Abstract

About 800 women die every day worldwide from pregnancy-related complications, including excessive blood loss, infections and high blood pressure (World Health Organization, 2019). To improve screening for high-risk pregnancies, we set out to identify patterns of maternal haematological changes associated with future pregnancy complications. Using mixed effects models, we established changes in 14 complete blood count (CBC) parameters for 1,710 healthy pregnancies and compared them to measurements from 98 pregnancy-induced hypertension, 106 gestational diabetes and 339 postpartum hemorrhage cases. Results show inter-individual variations, but good individual repeatability in CBC values during physiological pregnancies, allowing the identification of specific alterations in women with obstetric complications. For example, in women with uncomplicated pregnancies, haemoglobin count decreases of 0.12 g/L (95% CI -0.16, -0.09) significantly per gestation week (p-value < 0.001). Interestingly, this decrease is 3 times more pronounced in women who will develop pregnancy-induced hypertension, with an additional decrease of 0.39 g/L (95% CI -0.51, -0.26). We also confirm that obstetric complications and white CBC predict the likelihood of giving birth earlier during pregnancy. We provide a comprehensive description of the associations between haematological changes through pregnancy and three major obstetric complications to support strategies for prevention, early-diagnosis and maternal care.

Introduction

During pregnancy, women experience physiological changes to facilitate the growing foetus and to prepare for labour [1]. Understanding these changes as well as improving prevention, early-diagnosis and care for women during pregnancy, labour and post-partum, requires increased efforts to generate data and large reference samples. While reference values for maternal health are established to avoid unnecessary interventions [2, 3], very few studies focus on blood cell count changes from conception to childbirth. Physiological changes, including haematological changes, that may be perceived as pathological outside of pregnancy are poorly understood as the participation of pregnant women is extremely limited in clinical trials [4-6] and large-scale cohort data are lacking. It is therefore important to fully characterise what makes a healthy pregnancy as the pregnancy progresses, to facilitate identification of unusual patterns of obstetric complications and improve the stratification of high-risk pregnancies.

In this study, we used data on 14 complete blood count (CBC) parameters, collected from 2003 to 2009 at the Lausanne University Hospital (CHUV), to establish haematological changes during pregnancy. CBC is routinely performed to assess any abnormal fluctuations in blood values that help to screen for clinical risk factors associated with pregnancy [7]. For example, gestational thrombocytopenia in pregnant women, a common haematologic complication of pregnancy, is identified by a platelet count below 150,000/ μ L no earlier than 100 days before delivery [8]. Intra- and inter-individual CBC may thus fluctuate during pregnancy and

be influenced by maternal life-style factors and pregnancy-related complications (Supplementary Figure S1). To characterise haematological changes, we first set up a reference for the 14 CBC parameters from healthy pregnancies. We then assessed differences in the variation of CBC during pregnancies with three major complications: (i) hypertensive disorders of pregnancy (HDP) including pregnancy-induced hypertension, preeclampsia, HELLP (Hemolysis, Elevated Liver enzymes and Low Platelets) syndrome and unspecified maternal hypertension, (ii) gestational diabetes mellitus (GDM) and (iii) post-partum haemorrhage (PPH). Finally, we estimate the association of CBC parameters and pregnancy-related complications, namely HDP and GDM, on birth timing using a Cox proportional hazards model.

Methods

1 Study design and participants

The Lausanne University Hospital (CHUV) maternity cohort aims to study maternal health, and maternal and fetal outcomes. We obtained local canton ethical approval from CER-VD [9] under the project ID 2019-00280 for re-use of data initially collected for a serological surveillance study to investigate the prevalence of maternal and fetal toxoplasmosis infections between 2009 and 2014. The data consists of maternal medical record information for 4,347 pregnancies including haematological measures of which CBC taken at prenatal visits and described in Table 1. Our central laboratory uses an automated blood counter (Sysmex XN®), which rely on complementary techniques to determine CBC; (i) photometry after total red blood cell lysis for HB; (ii) impedance for ERY and PLATE counts (cell type being discriminated by size cut-offs), MCV, and MPV; and (iii) flow-cytometry for LEUC counts and differentiation. Other CBC parameters (HT, MCH, MCHC, RDW) are derived from these measures. The data also includes (i) intrapartum measures (reason for admission, age, maternal weight, maternal height, blood pressure, gestational age, fetal position, fundal height, method of delivery, contraction number, interventions to assist with birth, delivery date and delivery time), (ii) newborn measures (sex, birth weight, birth height, pH of umbilical cord, Apgar score), (iii) ICD-10 classification for diagnosis and medical procedures, and (iv) blood samples from the mother and from the umbilical cord collected at delivery.

In this study, we focus on the CBC measures taken throughout pregnancy (Table 1). We considered a longitudinal cohort study of 2,253 pregnancies with a single live birth (ICD10 Z370). From these, 42 pregnant women have data for two consecutive pregnancies. The study includes 1,710 control pregnancies with single spontaneous full-term uncomplicated delivery live births (ICD10 O80), 98 cases of HDP including pregnancy-induced hypertension, preeclampsia, HELLP syndrome and unspecified maternal hypertension (ICD10 codes O13, O14 and O16), 106 cases of GDM (ICD10 O24) and 339 cases of PPH (ICD10 O72). Still births (ICD10 Z371), liveborn twins (ICD10 Z372), multiple pregnancies (ICD10 O30) and complications specific to multiple gestation (ICD10 O31) were excluded from the analysis. Women with ICD-10 classifications unrelated to pregnancy, childbirth and the puerperium were filtered out excluding pre-pregnancy diseases of which pre-existing hypertension (ICD10 O100). To better compare cases of HDP, GDM and PPH in our analysis, pregnancies with two or all three complications studied and women with additional ICD-10 classifications for edema, proteinuria, and hypertensive disorders of pregnancy, delivery and puerperium (O10-O16), other maternal disorders primarily related to pregnancy (O20-O29), polyhydramnios (O40), other amniotic fluid and membrane disorders (O41), placental disorders (O43-O44), and maternal care for fetal abnormality (O35-O36), were excluded. In addition to the ICD10 codes, we also used reports at delivery. Women with an indication of hypertension were added to HDP cases and those with blood loss greater than 500 ml following delivery to PPH cases. The distributions for maternal age at birth, maternal weight at birth, gestational age, parity and nationality of the selected participants are shown in Supplementary Figures S2-S3. Distributions for each CBC parameter included in our study are shown in Supplementary Figure S4 and reported in Supplementary Tables S1-S2.

2 Statistical analysis

82

Haematological changes throughout low and high-risk pregnancies. We applied a cubic polynomial regression model with a random intercept for each woman to (i) define a reference for the evolution of CBC measures taken throughout control pregnancies and (ii) to assess CBC changes in women with major obstetric complications. For each CBC measure, we have:

83
84
85
86

$$\begin{aligned}
 CBC_{ij} = & \gamma_0 + \gamma_1 \cdot I(\text{group}_i = GDM) + \gamma_2 \cdot I(\text{group}_i = HDP) + \gamma_3 \cdot I(\text{group}_i = PPH) \\
 & + (\beta_1 \cdot \text{week}_{ij} + \beta_2 \cdot \text{week}_{ij}^2 + \beta_3 \cdot \text{week}_{ij}^3) \\
 & + (\beta_4 \cdot \text{week}_{ij} \cdot I(\text{group}_i = GDM) + \beta_5 \cdot \text{week}_{ij}^2 \cdot I(\text{group}_i = GDM) + \beta_6 \cdot \text{week}_{ij}^3 \cdot I(\text{group}_i = GDM)) \\
 & + (\beta_7 \cdot \text{week}_{ij} \cdot I(\text{group}_i = HDP) + \beta_8 \cdot \text{week}_{ij}^2 \cdot I(\text{group}_i = HDP) + \beta_9 \cdot \text{week}_{ij}^3 \cdot I(\text{group}_i = HDP)) \\
 & + (\beta_{10} \cdot \text{week}_{ij} \cdot I(\text{group}_i = PPH) + \beta_{11} \cdot \text{week}_{ij}^2 \cdot I(\text{group}_i = PPH) + \beta_{12} \cdot \text{week}_{ij}^3 \cdot I(\text{group}_i = PPH)) \\
 & + Z_i \xi + u_i + \varepsilon_{ij}
 \end{aligned} \tag{1}$$

where CBC is the outcome variable, $i = 1, \dots, N$ and $j = 1, \dots, n_i$, with N the number of women and n_i the number of measurements done for women i . Variable $u_i \sim N(0, \sigma_u^2)$ is the random intercept for individual i and $\varepsilon_{ij} \sim N(0, \sigma_e^2)$ is the random error for individual's i j th measurement. Z are the covariate values and ξ is a vector with each of the covariate parameters including maternal age at birth, maternal weight at birth, parity, gestational age at birth and nationality. Of these, 0.09% of women had missing nationality, 0.04% had missing values for gestational age and 12.2% of pregnancies had no maternal weight reported. Nationality is categorized into European coded as 0 and non-European coded as 1. Variable week is the timing of the CBC measure in gestation week (GW). γ_0 is the intercept parameter, describing the woman's initial blood count value of CBC measure at the start of a control pregnancy. γ_1 , γ_2 and γ_3 are the fixed-effect regression coefficients describing the difference from the intercept in GDM, HDP and PPH pregnancy groups. β_1 to β_{12} are the fixed-effects regression coefficients of the polynomial terms in each pregnancy group.

87
88
89
90
91
92
93
94
95
96
97

When dividing CBC measurements by trimester, 581 measurements were taken up to GW 14, 480 between GW 15 and 28 and 3,196 after GW 28 in control pregnancies. We included cases of HDP with 35, 44, and 677 CBC measurements collected in trimesters 1, 2, and 3, respectively; cases of GDM with 47, 31, and 289 CBC measurements collected in trimesters 1, 2, and 3, respectively; and cases of PPH with 112, 102, and 1,229 CBC measurements collected in trimesters 1, 2, and 3, respectively (Supplementary Figure S4b). To assess CBC changes in women with the latter obstetric complications, we specified an interaction term between the timing of CBC measurements and a categorical variable specifying the pregnancy group. CBC measures, parity, gestational age, maternal age and weight at birth were centered and scaled with respect to their standard deviation prior analysis. Missing values were imputed using multiple imputation with predictive mean matching in the R package *mice* [10] and including case-control groups, maternal age at birth, the newborn's weight and height and a binary variable for premature birth (gestational age < 37 weeks) in the imputation model. We analysed 5 sets of complete data and pooled the results. Differences in the evolution of CBC between control and high-risk pregnancies are shown in Figure 1. Complete results from each polynomial regression model are reported in Supplementary Table S3 and S4. We primarily chose the cubic polynomial as it had the lowest Akaike information criterion (AIC) and Bayesian information criterion (BIC) for most CBC, reflecting a flexible but parsimonious number of parameters (Supplementary Table S5). We additionally applied a linear mixed effect model with random intercepts for each pregnancy describing woman's initial blood count value, similarly to the cubic polynomial model, as follows:

98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116

$$\begin{aligned}
CBC_{ij} = & \gamma_0 + \gamma_1 \cdot I(\text{group}_i = GDM) + \gamma_2 \cdot I(\text{group}_i = HDP) + \gamma_3 \cdot I(\text{group}_i = PPH) \\
& + \beta_1 \cdot \text{week}_{ij} \\
& + \beta_2 \cdot \text{week}_{ij} \cdot I(\text{group}_i = GDM) \\
& + \beta_3 \cdot \text{week}_{ij} \cdot I(\text{group}_i = HDP) \\
& + \beta_4 \cdot \text{week}_{ij} \cdot I(\text{group}_i = PPH) \\
& + Z_i \xi + u_i + \varepsilon_{ij}
\end{aligned} \tag{2}$$

where the coefficients β_1 to β_4 give us an overview of the direction of change during pregnancy and the additional effect of having one of the three obstetric complications (Figure 2 and Supplementary Figure S5). Regression coefficients from each linear mixed effect model are reported in Supplementary Table S6. Linear and polynomial models are fit using the R package *lme4* [11].

Effect of CBC, HDP and GDM on birth timing. We fit a time-dependent covariate Cox proportional-hazards model [12] to describe how CBC parameters and obstetric complications jointly influence the hazard rate of birth at a particular point in time. The model is expressed as follows:

$$h_i(t) = h_0(t) \cdot \exp(\beta_1 \cdot X_1^i + \beta_2 \cdot X_2^i + \dots + \beta_p \cdot X_p^i) \tag{3}$$

where t represents the pregnancy time, $h_i(t)$ is the hazard function for individual i which can vary overtime and can be interpreted as the risk of labour at time t . The regression coefficients $\beta_1, \beta_2, \dots, \beta_p$ measure the effect size of p covariates. h_0 is the baseline hazard describing how the risk of birth changes over time at baseline levels of covariates. To select the most adequate predictors to include in the model, we applied a backward stepwise selection model using the BIC. Backward stepwise selection performs model comparison by removing predictors included in the model and evaluating the BIC of models of decreasing complexity until the most optimal model is reached. Maternal age at birth, maternal weight at birth, parity, nationality, the 14 CBC parameters in Table 1 and a categorical variable for controls, GDM and HDP pregnancies, were considered in the stepwise analysis. Maternal weight and parity were selected and included in the multivariate analysis as time-constant covariates. From the CBC parameters, HB, RDW, LEUC, ALYMPH, ANEUT and PLATE levels were retained. The pregnancy groups were also added to the multivariate analysis setting the control group as the reference. Complete results from the multivariate Cox proportional hazards model are reported in Supplementary Table S7 and the hazard ratios (HR) are shown in Figure 3. The HR measures the likelihood of women whose pregnancy is complicated by GDM or HDP, to give birth at time t compared to controls. For the continuous variables, the HR reflects the hazard of birth at time t if the variable in question increases by one unit. Cox proportional-hazards models are fit using the R package *survival* [13].

Finally, we further investigated significant associations found between time to birth, pregnancy complications and CBC measurements by exploring the number of C-sections and oxytocin administration in pregnancy complications. We also explored the proportion of obstetric infections and premature rupture of membranes (PROM) using ICD-10 codes reported in each pregnancy group (Supplementary Figure S6).

Results

Haematological changes throughout low and high-risk pregnancies. Firstly, we investigated haematological changes during pregnancies with single spontaneous full-term uncomplicated delivery live birth (ICD10 code O80) and compared them to values in pregnancies complicated by either GDM, HDP or PPH. For this purpose, a cubic polynomial regression was applied with (i) an interaction between time and a categorical variable indicating the pregnancy group, (ii) a number of covariates and (iii) random intercepts for each pregnancy to model individual-level differences in repeated blood count values. We identify the

following changes at a P-value < 0.0036 adjusting the significance level with the Bonferroni correction for the 14 phenotypes tested (Supplementary Table S1-S3).

Figure 1 depicts how CBC parameters change from the first to the third trimester, while Figure 2 shows CBC predictions at GW 0, 10, 20 and 30 using estimates from the linear mixed effect model. In control pregnancies, we observe a decrease in the second trimester followed by an increase in the third trimester for erythrocyte parameters (HB, HT and ERY), a decrease in platelets (PLATE), and an increase in leucocyte (LEUC). The same patterns are delineated when the linear mixed effect model is applied to estimate an overall direction of haematological changes during pregnancy (Supplementary Figure S5). First, we found that HB levels significantly decrease by 0.008 (95% CI -0.011, -0.006), HT by 0.005 (95% CI -0.007, -0.003) and ERY by 0.01 (95% CI -0.012, -0.008) per GW. We also observe that the mean corpuscular haemoglobin (MCH) is not significantly altered, while the mean red cell volume (MCV) shows a marginal increase, explaining a marginal decrease in the mean corpuscular haemoglobin concentration (MCHC) across pregnancy. Lastly, red blood cell distribution width (RDW), which measures the change in red blood cell size, increases continuously as pregnancy progresses. Second, PLATE count also decreases by 0.012 (95% CI -0.014, -0.010), whereas the mean platelet volume (MPV) decreases and then increases in the second trimester. And third, LEUC counts increase by 0.015 (95% CI 0.012, 0.017). Among leucocytes, the polynomial slope for absolute neutrophil (ANEUT) count increases early in pregnancy while absolute monocyte (AMONO) count increases from the second trimester onward. Absolute lymphocyte (ALYMPH) count decreases marginally early in pregnancy. The linear estimation confirms these trends with a significant increase by 0.015 (95% CI 0.013, 0.018) in ANEUT, 0.017 (95% CI 0.015, 0.019) in AMONO, and a decrease by 0.005 (95% CI -0.008, -0.003) in both ALYMPH and absolute eosinophil count (AEOSI) per GW.

In pregnancies complicated by either GDM, HDP or PPH, we find statistically significant differences at P-value < 0.0036 , in the polynomial terms for all CBC except ALYMPH and RDW (Supplementary Table S1). We observe group specific differences in the polynomial slopes, especially early in pregnancy, between GW 10 and 20 (Figure 1 and Supplementary Figure S7). As with the polynomial curves, CBC values, predicted from the linear mixed-effect model, vary overtime and are subject to changes specific to women with at-risk pregnancies (Figure 2). For pregnant women who will develop HDP, we find that erythrocyte parameters are approximately 2-fold higher compared to controls at GW 0, and that they decrease significantly faster during pregnancy with the following interaction effect sizes: -0.026 (95% CI -0.035, -0.018) for HB, -0.026 (95% CI -0.035, -0.17) for HT and -0.021 (95% CI -0.029, -0.012). PLATE count is significantly lower by -0.42 (95% CI -0.70, -0.14) at the start of pregnancy but although not significant, the decrease is slower in women who will develop HDP compared to controls (Figure 2, Supplementary Figure S5). We find similar trends in pregnancies leading to PPH with approximately 3-fold higher values in erythrocyte parameters at GW 0, which also decrease significantly faster: -0.041 (95% CI -0.046, -0.036) for HB, -0.044 (95% CI -0.049, -0.039) for HT, -0.043 (95% CI -0.048, -0.038) for ERY. PPH pregnancies also have 2-fold lower values LEUC and ANEUT counts at GW 0. As pregnancy progresses, these counts and AMONO increase significantly faster with interaction effect sizes: 0.026 (95% CI 0.020, 0.031) for LEUC, 0.028 (95% CI 0.023, 0.033) for ANEUT and 0.013 (95% CI 0.007, 0.018) for AMONO (Figure 2, Supplementary Figure S5). Finally, women diagnosed with GDM are distinguished by approximately 2-fold higher LEUC and ANEUT counts at GW 0. Interestingly, LEUC count remains constant throughout the pregnancy as it significantly decreases by 0.014 (95% CI: -0.022, -0.005) counteracting the increase of 0.015 found in the control pregnancies. We also note that their number of HB and HT decreases by -0.015 (95% CI -0.023, -0.007) and -0.013 (95% CI -0.021, -0.004) with each passing week of pregnancy (Figure 2, Supplementary Figure S5).

Random intercepts for individuals fitted within each polynomial model revealed that 30.9% to 84.6% of the variance is attributed to inter-individual differences indicating that the pregnant women differ in their initial blood cell count (Supplementary Table S5). The proportion of total variance attributable to this term also informs us about individual repeatability, i.e., how similar observations of an individual are as compared to the rest of the population, as the pregnancy progresses [14]. For instance, 84.6% of the variance in MCV measurements would be explained by the random intercept and thus the rate of change

for those measurements are likely to be the same for all women. We also find variation in blood measure with maternal age and weight at delivery, parity, nationality and gestational age at birth (Supplementary Figure S8). Estimates of the effect of nationality on the 14 CBC have a wider confidence interval than the other covariates, which could imply a broader genetic background in individuals than that described by their nationality. Compared to other CBC, RDW is more affected by higher parity with an increase of 0.16 (95% CI 0.13,0.20). We also observe a negative effect with a 1-SD increase in gestational age on LEUC count which appears to be driven by ANEUT and ALYMPH with a decrease of -0.21 (95% CI -0.25,-0.17) and -0.18 (95% CI -0.22,-0.14) respectively. Finally, maternal age and weight show significant opposite effects on ERY, MCV, MCH, MCHC, RDW and ALYMPH counts.

Effect of CBC, HDP and GDM on birth timing. Using a time-dependent covariate Cox-proportional hazards model, we assess how CBC measurements taken throughout pregnancy, GDM and HDP obstetric complications jointly influence birth timing measured as gestational age at birth. To select the most adequate multivariate model, we used backward step-wise selection and the final model used for the analysis included maternal weight, parity, HB, RDW, LEUC, ALYMPH, ANEUT and PLATE CBC parameters, and the case-control categorical variable for HDP, GDM and control pregnancies.

Kaplan Meier curves in Figure 3a describe the probability of birth from GW 24 to 42 in control pregnancies and in cases of GDM and HDP. The median gestational age at delivery is 39.55 (95% CI 39.50, 39.60) in controls and 39.10 (95% CI 39.00, 39.30) in GDM-complicated pregnancies, both being within the limits of what is considered to be full-term delivery; as opposed to 37.60 GW (95% CI 37.10, 38.20) in HDP-complicated pregnancies, which lies very close to the cut-off for preterm delivery. Indeed, 39.8% of HDP-complicated pregnancies resulted in birth ≤ 37 GW and, although ICD10 code O80 indicates single spontaneous full-term uncomplicated delivery in controls, only 6.6% of healthy and 5.7% of GDM-complicated pregnancies delivered prematurely. Furthermore, at GW 41, 16.1% of controls, 7.5% of GDM and 5.1% of HDP pregnancies are still ongoing and can be classified as late deliveries. In agreement with the Kaplan Meier curves, we find that the hazard of birth is 2.49 (95% CI 1.95, 3.18; P-value = 2.82e-13) times higher in pregnancies complicated by HDP and, although not significant, the hazard of birth was found to be 1.35 (95% CI 0.98, 1.85) times higher in pregnancies complicated by GDM (Figure 3b). Moreover, the hazard of giving birth at a given time rises by 2.94 for 1 G/l increase in ALYMPH (95% CI 2.07, 4.17; P-value = 1.83e-09) and by 2.06 for 1 G/l increase in ANEUT (95% CI 1.54, 2.74; P-value = 8.63e-07). We also find that for one unit increase in HB, RDW and LEUC cell counts, the hazard of giving birth falls by 0.99 (95% CI 0.983, 0.996; P-value = 1.53e-03), 0.90 (95% CI 0.85, 0.95; P-value = 4.98e-05) and 0.56 (95% CI 0.42, 0.73; P-value = 2.92e-05) respectively. Finally, the hazard of birth increases by 1.15 (95% CI 1.07, 1.23; P-value = 1.93e-04) for one unit increase in parity and decreases by 0.99 (95% CI 0.98, 0.99; P-value = 5.41e-06) for one unit increase in maternal weight.

We further investigate the association found between time to birth and HDP complication by exploring the number of C-sections and oxytocin administration reported. Of the 98 HDP cases in our study, 67.3% resulted in a C-section, 22.6% received oxytocin to induce or to accelerate labour and 10.2% gave birth without either intervention. Out of 39 preterm deliveries in the HDP group, only 3 infants were born by natural labour. We also investigated ANEUT and ALYMPH associations with time to birth, looking at the proportion of obstetric infections and PROM in each pregnancy group (Supplementary Figure S6). Using the reported ICD-10 codes in our sample, 0.7% of patients had an infection during pregnancy and we find $\leq 2\%$ of infections per pregnancy group. On the other hand, we observe 15.3% of PROM, of which 7.9% are preterm PROM.

Discussion

In this study, we extensively describe the haematological changes that occur during healthy pregnancies and compare them with pregnancies involving three major obstetric complications, namely GDM, HDP and PPH. Maternal data are also analysed in a novel manner, by exploring the effect of obstetric complications and CBC measurements throughout pregnancy, on gestational age at delivery.

Our results are in line with previous studies on haematological changes in healthy pregnancy. Blood volume is known to increase by about 1.5L throughout pregnancy. Erythrocyte count increases due to a greater erythropoietin production [15]; however, the volume of plasma increases proportionally more, thus pregnant women present a net decrease in red blood cell parameters, notably haemoglobin and haematocrit in the second trimester, followed by a stabilisation in the third [1, 15]. Leucocyte counts have been shown to rise mainly due to neutrophilia and monocytosis caused by the physiologic stress imposed during pregnancy, alongside a decrease in lymphocytes [15–17]. Finally, previous studies have also shown a decrease in platelet count [1, 15, 18]. Here, we confirm and refine the description of these physiological changes. More importantly, we explore how repeated monitoring of CBC can help identify complications. Compared to healthy pregnancies, we found significant differences in the variation of CBC parameters throughout pregnancies complicated by GDM, HDP or PPH, each having a unique alteration pattern. We thus demonstrate that assessing haematological changes has the potential to timely identify women who will develop obstetric complications. Recent findings show that CBC can help predict the risk of PPH [14] and GDM [19]. Regarding HDP, few studies have used platelet count to predict preeclampsia alone [20–22]. An increase in hematocrit count between the first and second trimester has also been shown to be predictive of preeclampsia as well as other pregnancy outcomes such as fetal growth restriction [23, 24]. In our analysis, we identify the 10th to 20th week of gestation as the most informative period for identifying these complications. This information brings novelty to the field as it indicates a specific time frame, early in pregnancy, that we can focus on to predict and identify maternal complications occurring weeks later and improve prenatal care. For instance, we demonstrate that routine red blood cell and platelet count are sensitive enough to identify pathophysiological mechanism occurring early in the course of the pregnancy that will later lead to the development of HDP. The observed pattern of significantly accelerated decrease in red blood cell and platelet counts, is consistent with low grade thrombotic microangiopathy, which takes many forms during pregnancy including preeclampsia and the HELLP syndrome [25, 26]; however further research is needed to establish this hypothesis. In current clinical practice, it is not usual to continuously measure CBC values during pregnancy. However, our results indicate a pattern of change in CBC values that differs in women with negative pregnancy outcomes, as opposed to healthy pregnancies. This suggests that more routine CBC testing throughout pregnancy may contribute to earlier diagnosis. Finally, as in [14], we observe that post-labour complications can also be related to what happens during pregnancy.

When exploring associations between time to birth, pregnancy complications and CBC measurements, we find that HDP significantly increases the hazard of birth and thus may shorten pregnancy duration compared to controls regardless of the GW. Among the HDP cases, we observe that few pregnant women, three of whom gave birth before GW 37, delivered without C-section or use of oxytocin. It is thus likely that our results reflect medically indicated births due to obstetric guidelines, recommending induction of labour in women who develop hypertension; and that onset of HDP before GW 37 increases the likelihood of preterm delivery. With regards to the CBC results, we show that the hazard of giving birth rises with higher values of ALYMPH and ANEUT. As both cell types are mobilised by the immune system in the presence of pathogens [27, 28], results may reflect an underlying subclinical infection. Using reported ICD-10 codes, we find less than 1% of obstetric infections. However, we observe 15.3% of PROM, a complication associated with infections in pregnancy [29, 30]. Following PROM, if labour does not begin spontaneously within 24 hours, obstetric guidelines recommend induction of labour. Of the reported cases, 52% are preterm PROM thus resulting in preterm birth. Additionally, physiological changes during labour that have been described as inflammatory reactions [31, 32] and treatments such as steroids in preparation of preterm birth, are additional factors that may lead to significant changes in white blood cell counts. We believe, further research is needed to investigate haematological changes specifically during labour and in complications that play a key role in the timing of birth. With regards to parity, a recent study showed that women in their first pregnancy are at greater risk of spontaneous preterm birth compared to women in their second pregnancy; and that the risk increases steadily in multiparous women [33]. We find that for every new pregnancy, the hazard of birth increases by 14.7% and so, it would be interesting to deconstruct our analysis and compare different parity status in a larger sample size.

309
310 Although the CHUV maternity cohort is phenotypically rich, the main limitation in our study is the
311 heterogeneity in the data collection. First, although we have repeated measures throughout pregnancy, these
312 have not been collected at similar time points for each pregnancy and 75.1% to 89.6% of measures are taken
313 in the third trimester (Supplementary Figure S4b). Our results are thus more reliable in the end of pregnancy
314 where we can confidently establish specific ranges for each cell blood count (Supplementary Table S1) and
315 we believe that a more complete dataset from the beginning of pregnancy would show a more pronounced
316 difference between the groups. Second, the time of onset of the various pregnancy complications was not
317 reported in the CHUV maternity cohort. With a more complete data set and an increase in sample size, we
318 would be able to further investigate whether hematologic changes in early pregnancy are also dependent on the
319 timing of pregnancy complications. A larger sample size is also required to determine the prediction accuracy
320 when predicting these complications from longitudinal CBC measures taken between the 10th and 20th week
321 of gestation. Third, the data was collected 10 years ago and guidelines in maternal health have evolved since
322 then. These three limitations emphasise the importance of data collection in a rapidly changing field. Finally,
323 as obstetric complications have a multi-factorial etiology including a number of medical interventions, bigger
324 sample sizes of diverse ancestries and genetic data are required to (i) investigate maternal risk factors, (ii)
325 better predict and understand obstetric complications, (iii) stratify women that are at risk and (iv) develop
326 risk specific guidelines. Fortunately, maternal health is increasingly becoming part of the research agenda.
327 Available data and collaborations to improve maternal care are increasing and we are currently filling the
328 gaps in the field.
329

Author contributions

330

MP and MRR conceived and designed the study. MP conducted the analysis with contributions from MRR, DB, MS, SEO and ZK. MRR and DB provided study oversight, and DB contributed data. MP drafted the paper and all authors reviewed and approved the final manuscript prior to submission.

331

332

333

Author competing interests

334

The authors declare no competing interests.

335

Acknowledgements

336

This project was funded by an SNSF Eccellenza Grant to MRR (PCEGP3-181181), and by core funding from the Institute of Science and Technology Austria. We would like to thank the participants of the study and all the midwives and doctors involved for the computerized obstetrical data from the CHUV Maternity Hospital.

337

338

339

Tables

340

Table 1. Cell blood counts (unit) included in the study.

Type	Name	Description
Red blood cells	ERY (T/L)	Red blood cell count
	HB (g/L)	Total amount of the oxygen-carrying protein in the blood
	HT (%)	Volume percentage of red blood cells in blood
	MCV (fl)	Mean cell volume
	MCH (pg)	Mean corpuscular hemoglobin
	MCHC (g/L)	Mean corpuscular hemoglobin concentration
	RDW (%)	Red blood cell distribution width
White blood cells	LEUC (G/L)	White blood cell count
	ANEUT (G/L)	Absolute number of neutrophils
	ALYMPH (G/L)	Absolute number of lymphocytes
	AMONO (G/L)	Absolute number of monocytes
	AEOSI (G/L)	Absolute number of eosinophils
Platelets	PLATE (G/L)	Platelet count
	MPV (fl)	Mean platelet volume

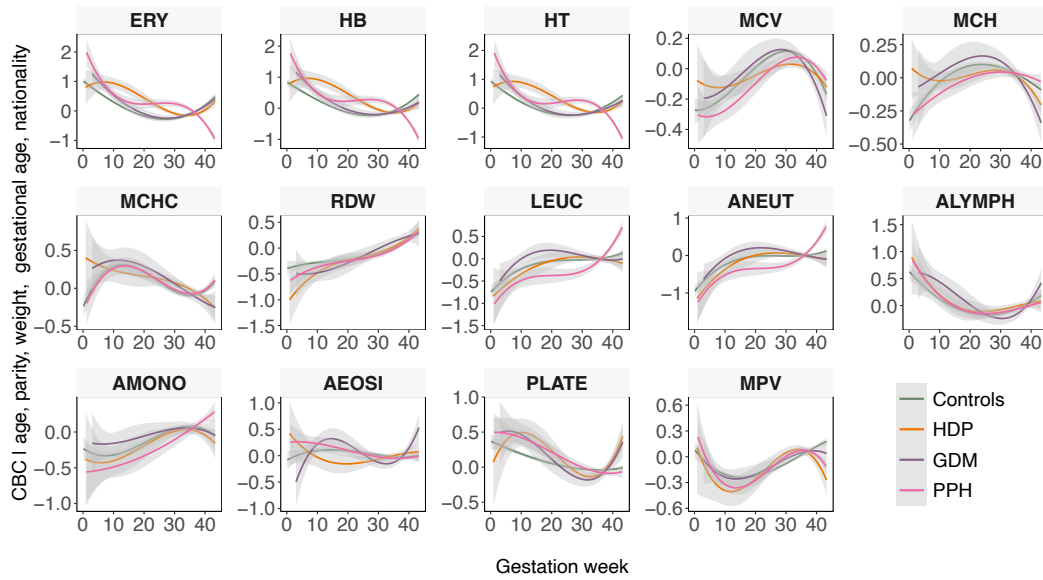


Figure 1. Haematological changes in major pregnancy complications. Cubic polynomial slopes showing the course of 14 cell blood counts (CBC) in control pregnancies, pregnancies with gestational diabetes mellitus (GDM), pregnancies with induced hypertension (HDP) and pregnancies resulting in post-partum haemorrhage (PPH). Slopes are adjusted for maternal age and weight at delivery, parity, gestational age and nationality. 95% CI are displayed in grey around the slopes. CBC parameters and covariates are centered and scaled with respect to their standard deviation. In control pregnancies, red CBC (ERY) and platelets (PLATE) decrease compared to white CBC (LEUC) which gradually increases overtime. Results show different patterns for HDP, GDM and PPH with a greater difference between the course of CBC in control and in HDP and PPH pregnancies. For instance, we observe non-overlapping 95% CI in ERY, HB and HT mean counts between control pregnancies and pregnancies complicated by HDP or by PPH.

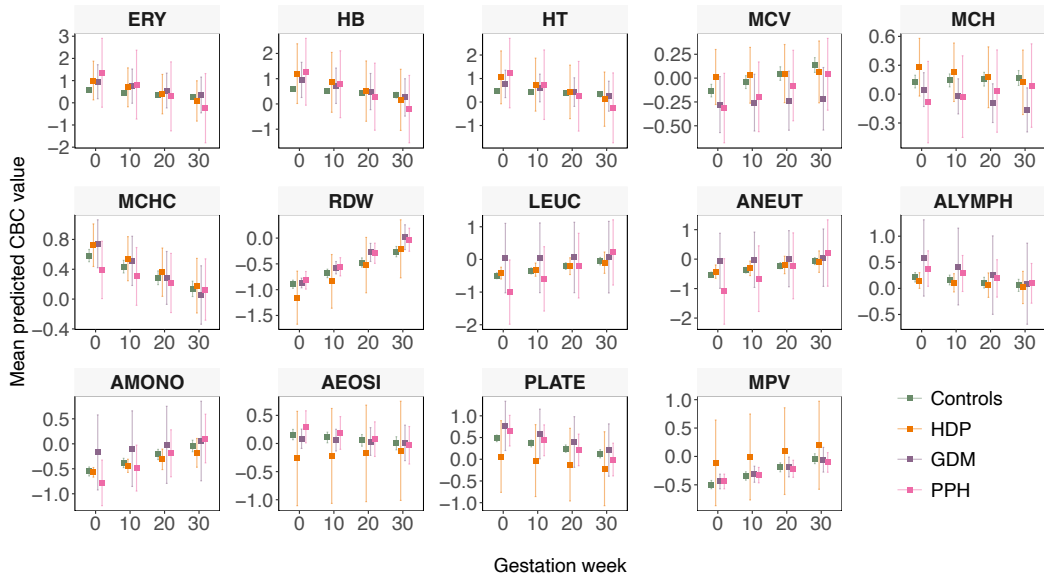


Figure 2. CBC prediction of each group across pregnancy. Mean predicted CBC values of control, GDM, HDP and PPH pregnancy groups at week 0, 10, 20 and 30. The error bars show the variance of the prediction error, except at week 0, where we show (i) the estimated mean intercept and 95% CI for the control group and (ii) the predicted mean values and the prediction error variance for the case groups by adding the estimated effect of pregnancies complicated by either GDM, HDP or PPH to the estimated control intercept. At week 10, 20 and 30, the predicted CBC value in control groups is computed by adding the estimated intercept and the estimated effect of time in weeks times gestation week = (10 or 20 or 30). To predict values in a pregnancy affected by one of the three complications, we also sum the effect of time in weeks for the group of interest times gestation week = (10 or 20 or 30). We observe clear differences in the groups, i.e., in the count of AMONO in case pregnancies compared to the reference mean predicted values in green, as pregnancy progresses.

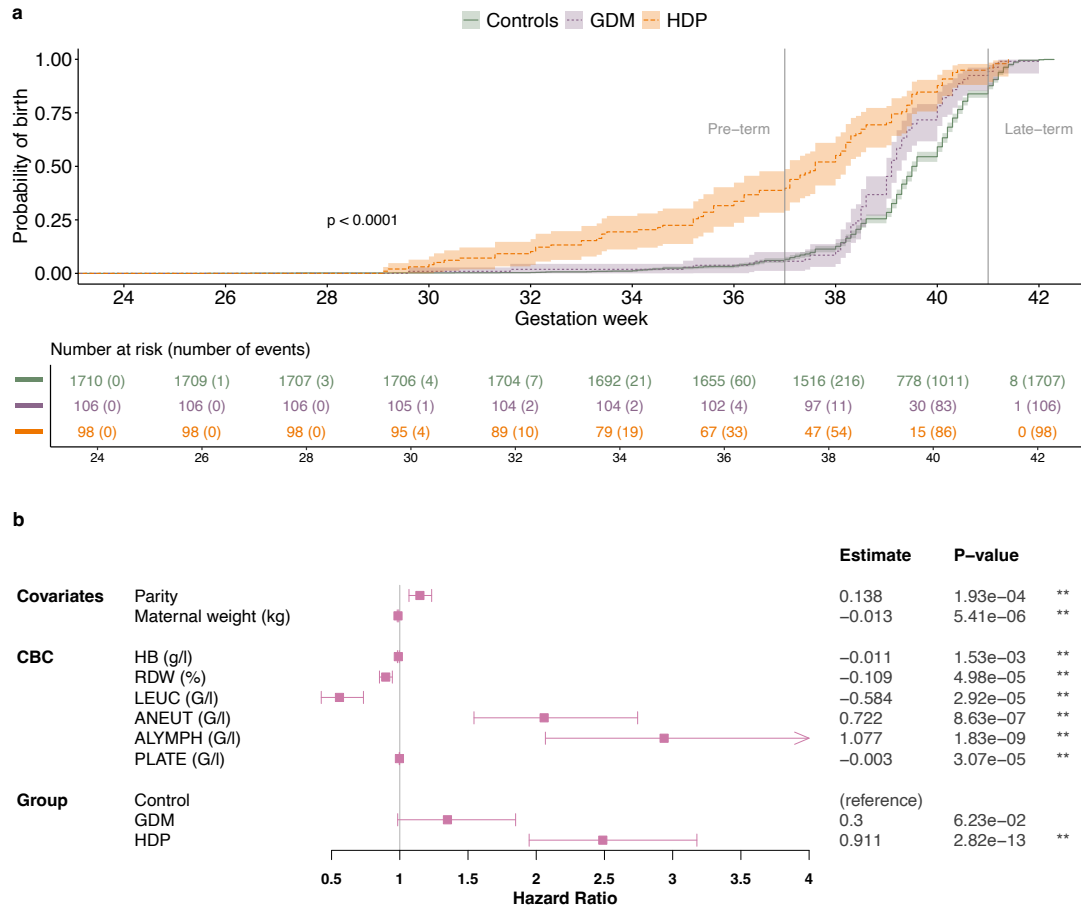


Figure 3. Time-to-birth analysis. (a) Cumulative incidence plot showing the probability of giving birth throughout controls and high risk pregnancies including gestational diabetes mellitus (GDM) and hypertensive disorders in pregnancy (HDP). The cut-off point for a pre-term and late-term delivery is indicated by the grey lines. The plot also includes a risk table with the number of women susceptible to give births and the cumulative number of events at a given gestation week. The log-rank test p-value < 0.0001 indicates that birth timing is significantly different between our groups. Pregnancies with HDP are found to have a higher probability of preterm birth. (b) Forest plot showing the hazard ratio with 95% confidence intervals, estimates and p-values associated with variables included in the cox proportional hazards model. HDP pregnancies give birth 2.5x the rate per unit time compared to control pregnancies (P-value = $6.33e-13$). We also observe that lymphocyte (ALYMPH) and neutrophil (ANEUT) counts increase the probability of giving birth, with a hazard by a factor of 2.9x (P-value = $2.97e-09$) and 2x (P-value = $1.09e-06$) respectively. This might reflect the presence of an infection, which is one of the main causes of spontaneous preterm delivery, or a reactive neutrophilia and lymphocytosis accompanying the causal event for birth.

References

1. P. Soma-Pillay and N. Catherine, "P, tolppanen h, mebazaa a, tolppanen h, mebazaa a," *Physiological changes in pregnancy. Cardiovasc J Afr*, vol. 27, no. 2, pp. 89–94, 2016.
2. G. Edelstam, C. Lowbeer, G. Kral, S. Gustafsson, and P. Venge, "New reference values for routine blood samples and human neutrophilic lipocalin during third-trimester pregnancy," *Scandinavian journal of clinical and laboratory investigation*, vol. 61, no. 8, pp. 583–591, 2001.
3. R. Sivasankar, R. A. Kumar, R. Baraz, and R. E. Collis, "The white cell count in pregnancy and labour: a reference range," *The Journal of Maternal-Fetal & Neonatal Medicine*, vol. 28, no. 7, pp. 790–792, 2015.
4. P. Dashraath, K. Nielsen-Saines, S. A. Madhi, and D. Baud, "Covid-19 vaccines and neglected pregnancy," *The Lancet*, vol. 396, no. 10252, p. e22, 2020.
5. N. Chakhtoura, J. J. Chinn, K. L. Grantz, E. Eisenberg, S. A. Dickerson, C. Lamar, and D. W. Bianchi, "Importance of research in reducing maternal morbidity and mortality rates," *American journal of obstetrics and gynecology*, vol. 221, no. 3, p. 179, 2019.
6. J. R. Biggio, "Research in pregnant subjects: Increasingly important, but challenging," *Ochsner Journal*, vol. 20, no. 1, pp. 39–43, 2020.
7. M. H. Gandhi and V. Gupta, "Physiology, maternal blood," *StatPearls [Internet]*, 2020.
8. E. Habas Sr, A. Rayani, G. Alfitori, G. E. Ahmed, and A.-N. Y. Elzouki, "Gestational thrombocytopenia: A review on recent updates," *Cureus*, vol. 14, no. 3, 2022.
9. "Commission cantonale d'éthique de la recherche sur l'être humain." [Online]. Available: <https://www.cer-vd.ch/>
10. S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," pp. 1–67, 2011. [Online]. Available: <https://www.jstatsoft.org/v45/i03/>
11. D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.
12. D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.
13. T. M. Therneau, *A Package for Survival Analysis in R*, 2021, r package version 3.2-13. [Online]. Available: <https://CRAN.R-project.org/package=survival>
14. M. R. Robinson, M. Patxot, M. Stojanov, S. Blum, and D. Baud, "Postpartum hemorrhage risk is driven by changes in blood composition through pregnancy," *Scientific Reports*, vol. 11, no. 1, p. 19238, Sep 2021. [Online]. Available: <https://doi.org/10.1038/s41598-021-98411-z>
15. S. Chandra, A. K. Tripathi, S. Mishra, M. Amzarul, and A. K. Vaish, "Physiological changes in hematological parameters during pregnancy," *Indian journal of hematology and blood transfusion*, vol. 28, no. 3, pp. 144–146, 2012.
16. O. Pughikumo, D. Pughikumo, and H. Omunakwe, "White blood cell counts in pregnant women in port harcourt, nigeria," *IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)*, vol. 14, no. 3, pp. 01–03, 2015.
17. I. Siegel and N. Gleicher, "Peripheral white blood cell alterations in early labor." *Diagnostic gynecology and obstetrics*, vol. 3, no. 2, pp. 123–126, 1981.

18. J. A. Reese, J. D. Peck, D. R. Deschamps, J. J. McIntosh, E. J. Knudtson, D. R. Terrell, S. K. Vesely, and J. N. George, "Platelet counts during pregnancy," *New England Journal of Medicine*, vol. 379, no. 1, pp. 32–43, 2018.
19. P. Pattanathaiyanon, C. Phaloprakarn, and S. Tangjitgamol, "Comparison of gestational diabetes mellitus rates in women with increased and normal white blood cell counts in early pregnancy," *Journal of Obstetrics and Gynaecology Research*, vol. 40, no. 4, pp. 976–982, 2014.
20. A. Kirbas, A. O. Ersoy, K. Daglar, T. Dikici, E. H. Biberoglu, O. Kirbas, and N. Danisman, "Prediction of preeclampsia by first trimester combined test and simple complete blood count parameters," *Journal of clinical and diagnostic research: JCDR*, vol. 9, no. 11, p. QC20, 2015.
21. M. A. AlSheeha, R. S. Alaboudi, M. A. Alghasham, J. Iqbal, and I. Adam, "Platelet count and platelet indices in women with preeclampsia," *Vascular health and risk management*, vol. 12, p. 477, 2016.
22. F. Tesfay, M. Negash, J. Alemu, M. Yahya, G. Teklu, M. Yibrah, T. Asfaw, and A. Tsegaye, "Role of platelet parameters in early detection and prediction of severity of preeclampsia: A comparative cross-sectional study at ayder comprehensive specialized and mekelle general hospitals, mekelle, tigray, ethiopia," *Plos one*, vol. 14, no. 11, p. e0225536, 2019.
23. Z. M. Lu, R. L. Goldenberg, S. P. Cliver, G. Cutter, and M. Blankson, "The relationship between maternal hematocrit and pregnancy outcome." *Obstetrics and gynecology*, vol. 77, no. 2, pp. 190–194, 1991.
24. M. G. Khoigani, S. Goli, and A. HasanZadeh, "The relationship of hemoglobin and hematocrit in the first and second half of pregnancy with pregnancy outcome," *Iranian journal of nursing and midwifery research*, vol. 17, no. 2 Suppl1, p. S165, 2012.
25. F. Fakhouri, M. Scully, F. Provôt, M. Blasco, P. Coppo, M. Noris, K. Paizis, D. Kavanagh, F. Pène, S. Quezada *et al.*, "Management of thrombotic microangiopathy in pregnancy and postpartum: report from an international working group," *blood*, vol. 136, no. 19, pp. 2103–2117, 2020.
26. S. Z. Vahed, Y. R. Saadat, and M. Ardalan, "Thrombotic microangiopathy during pregnancy," *Microvascular Research*, p. 104226, 2021.
27. R. L. Goldenberg, J. F. Culhane, J. D. Iams, and R. Romero, "Epidemiology and causes of preterm birth," *The lancet*, vol. 371, no. 9606, pp. 75–84, 2008.
28. H. L. Malech, F. R. DeLeo, and M. T. Quinn, "The role of neutrophils in the immune system: an overview," *Neutrophil*, pp. 3–10, 2020.
29. L. L. Klein and R. S. Gibbs, "Infection and preterm birth," *Obstetrics and Gynecology Clinics*, vol. 32, no. 3, pp. 397–410, 2005.
30. R. Menon and S. J. Fortunato, "Infection and the role of inflammation in preterm premature rupture of the membranes," *Best practice & research Clinical obstetrics & gynaecology*, vol. 21, no. 3, pp. 467–478, 2007.
31. M. Yuan, F. Jordan, I. McInnes, M. Harnett, and J. Norman, "Leukocytes are primed in peripheral blood for activation during term and preterm labour," *Molecular human reproduction*, vol. 15, no. 11, pp. 713–724, 2009.
32. N. Gomez-Lopez, D. StLouis, M. A. Lehr, E. N. Sanchez-Rodriguez, and M. Arenas-Hernandez, "Immune cells in term and preterm labor," *Cellular & molecular immunology*, vol. 11, no. 6, pp. 571–581, 2014.
33. B. Koullali, M. D. Van Zijl, B. M. Kazemier, M. A. Oudijk, B. W. Mol, E. Pajkrt, and A. C. Ravelli, "The association between parity and spontaneous preterm birth: a population based study," *BMC pregnancy and childbirth*, vol. 20, no. 1, pp. 1–8, 2020.

Supplementary Online Material

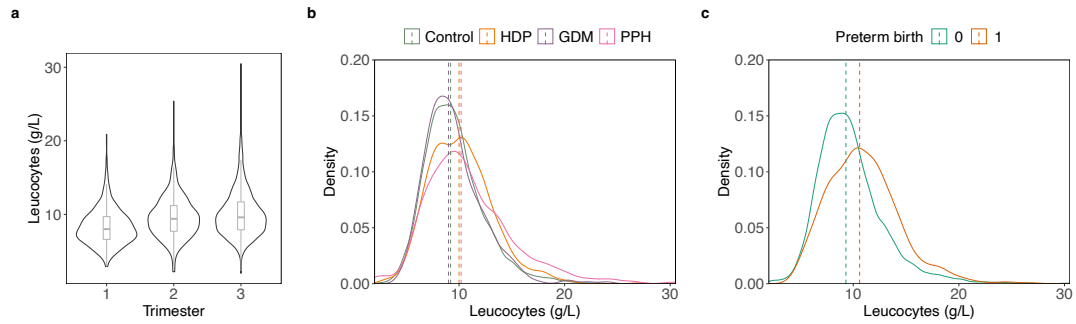


Figure S1. Example of CBC changes in maternal leucocyte cell count. The distribution of leucocytes (G/L) in blood varies (a) over time (b) between low and high risk pregnancies including hypertensive disorders of pregnancy (HDP), gestational diabetes mellitus (GDM) and post-partum haemorrhage (PPH), and (c) between preterm and term births. Violin plots show the distribution of leucocytes in trimester 1, 2 and 3 and boxplots with the median, the 25th and 75th percentiles. Coloured dashed lines show the median leucocyte count in blood for each group. We observe an increase in leucocyte count from trimester 1 to 3. The leucocyte distribution for HDP and PPH is slightly shifted towards higher values. The same is true for pregnancies with preterm delivery demonstrating the interest in monitoring blood values during pregnancy.

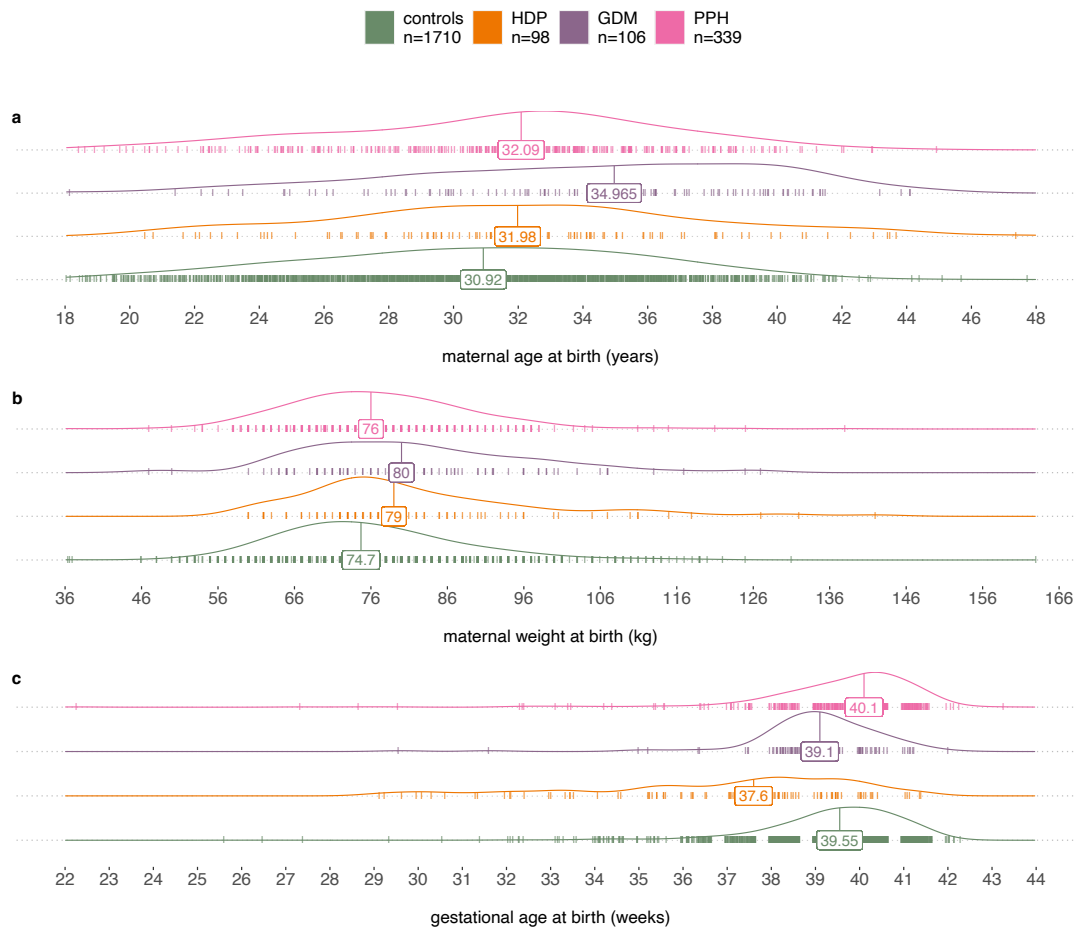


Figure S2. Distribution of maternal measures at birth. Distribution of (a) maternal age, (b) maternal weight and (c) gestational age at birth in pregnancies with spontaneous uncomplicated delivery (controls), gestational diabetes mellitus (GDM), hypertensive disorders of pregnancy (HDP) and post-partum hemorrhage (PPH). The median for each group is shown in the labelled box. HDP pregnancies have a lower median for gestational age compared to the other groups. We also observe that the age distribution of women with GDM is shifted to the right.

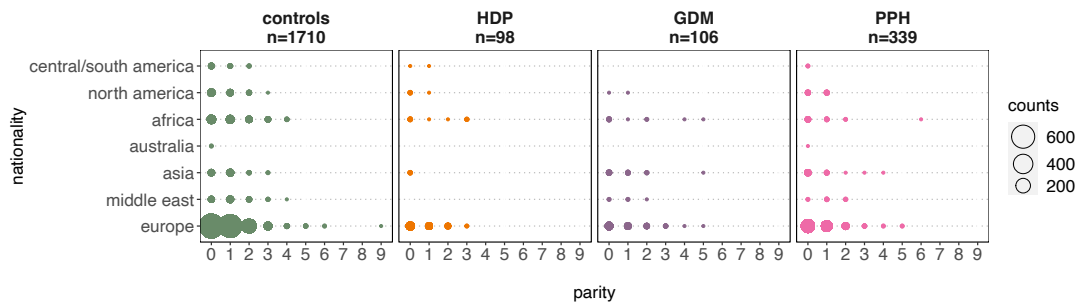


Figure S3. Nationality and parity in the CHUV maternity cohort. Distribution of nationality group by parity in pregnancies with spontaneous uncomplicated delivery (controls), gestational diabetes mellitus (GDM), hypertensive disorders of pregnancy (HDP) and post-partum hemorrhage (PPH). Solid dots give an order of magnitude of sample size. The majority of women in the CHUV maternity cohort are of European nationality. The parity ranges from 0 to 9 and there are mainly first (parity=0) or second (parity=1) pregnancies for women in the cohort.

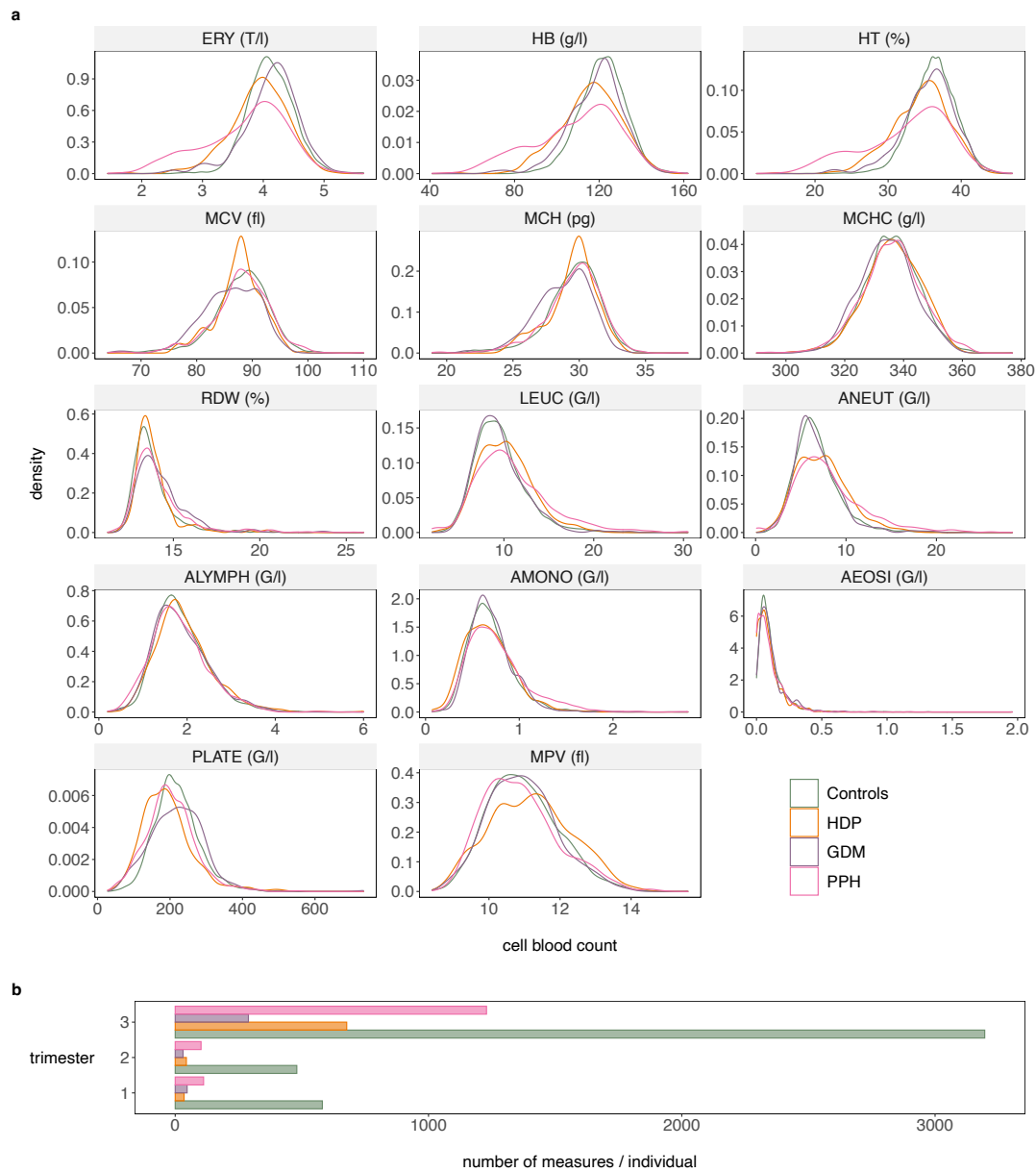


Figure S4. Distribution of the 14 cell blood count measures. (a) Distribution of erythrocyte count (ERY), hemoglobin level (HB), hematocrit count (HT), mean red cell volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), red blood cell distribution width (RDW), leucocyte count (LEUC), absolute number of neutrophils (ANEUT), absolute number of lymphocytes (ALYMPH), absolute number of monocytes (AMONO), absolute number of eosinophils (AEOSI), platelet count (PLATE) and mean platelet volume (MPV) for the 1,710 control pregnancies, 98 cases of gestational hypertensive disorders of pregnancy, 106 cases of gestational diabetes and 339 cases of post-partum hemorrhage included in this study. (b) Number of complete blood count (CBC) tests performed by case-control pregnancies and by trimester 1 to 3. Each CBC test includes the 14 CBC parameters described in Table 1.

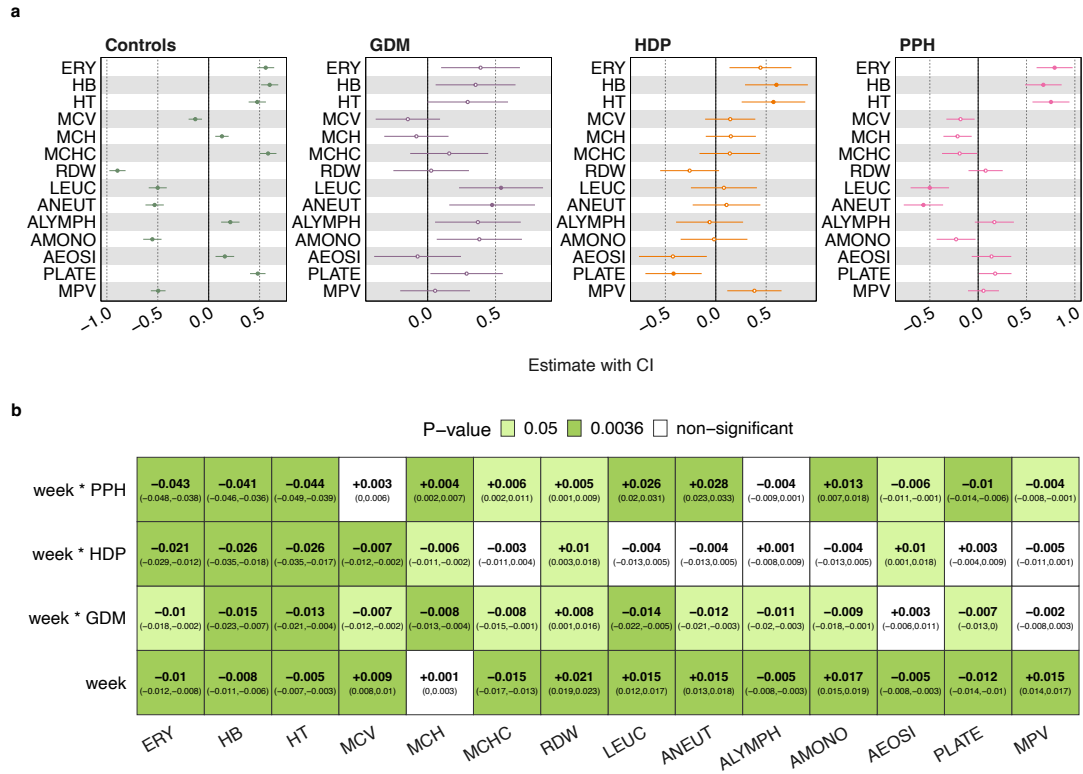


Figure S5. Interaction between gestation week and pregnancy groups. (a) Forest plots with estimates and 95% CI from the mixed effect model describing the estimated mean intercept of CBC values in the control pregnancy group and the estimated difference from this intercept in the GDM, HDP and PPH pregnancy groups at week=0. Solid points show significant effects at P-value $< 0.05/14 = 0.0036$ with Bonferroni correction for multiple testing. (b) Estimates (in bold) and (95% CI) from the linear mixed effect model, of gestation time in weeks and of the incremental effect of gestation time in weeks for women with GDM, HDP or PPH on CBC. For example, at a P-value level < 0.00036 , results show a significant decrease of -0.01 (95% CI -0.012, -0.008) in erythrocyte count (ERY) during pregnancy. ERY counts in women with HDP are -0.021 (95% CI -0.029, -0.012) lower. This effect is in addition to the reference established in controls and thus HDP pregnancies have a total ERY decrease of $0.01 + 0.021 = 0.031$ per gestation week. Effects are adjusted for maternal age and maternal weight at birth, parity, preterm birth and nationality, and CBC phenotypes are scaled with respect to their standard deviation.

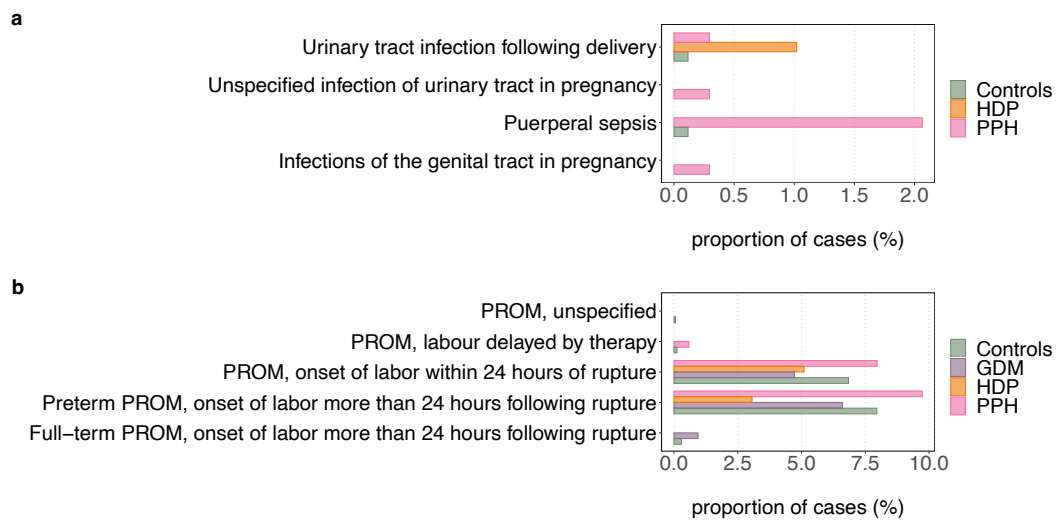


Figure S6. Proportion of infections and premature rupture of membranes reported in Controls, HDP, GDM and PPH. (a) Proportion of infections reported in each pregnancy group based on the following ICD 10 codes : O234 for unspecified infection of urinary tract in pregnancy, O235 for infections of the genital tract in pregnancy, O862 for urinary tract infection following delivery and O85 for Puerperal sepsis. No infections were reported in GDM cases. (b) Proportion of premature rupture of membranes (PROM) reported in each pregnancy group based on the following ICD 10 codes: O4212 for full-term PROM with onset of labor more than 24 hours following rupture, O4211 for preterm PROM with onset of labor more than 24 hours following rupture, O420 for PROM with onset of labor within 24 hours of rupture, O422 for PROM with labour delayed by therapy and O429 for unspecified PROM. The study includes 1,710 control, 98 HDP, 106 GDM and 339 PPH pregnancies.

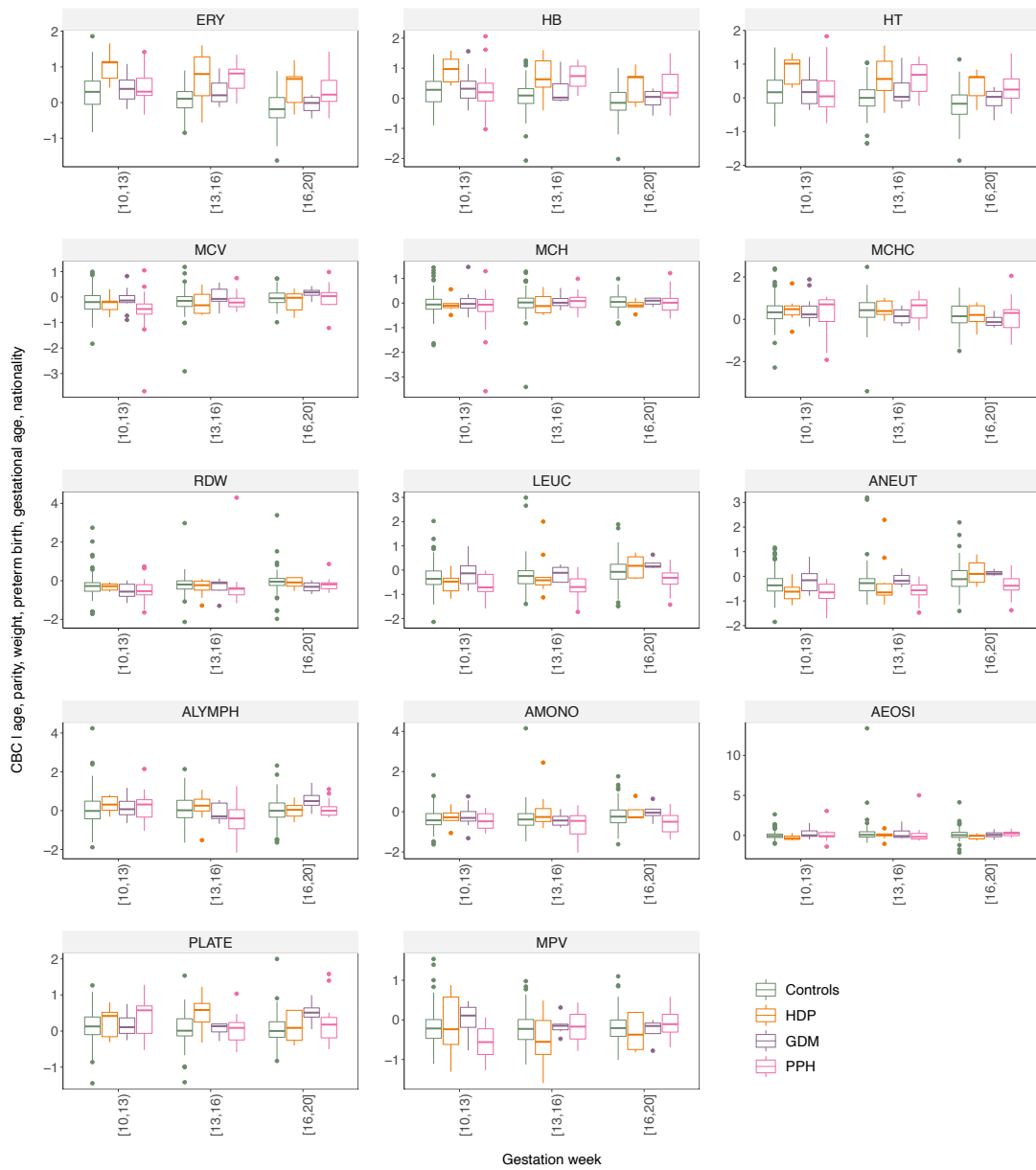


Figure S7. Course of the 14 complete blood counts from 10 to 20 weeks of gestation. Boxplots showing the distribution (median, 25% and 75% quartiles) of each blood measure described in Table 1 between cases of pregnancy induced hypertension (HDP), gestational diabetes mellitus (GDM), post-partum hemorrhage (PPH) and control pregnancies, from gestation week 10 to 20. This zoom shows that the polynomial curves (Figure 1) represent a real and visible difference between the groups in the average cell blood count. The first trimester lasts until the 13th week of gestation and the second trimester until the 26th week.

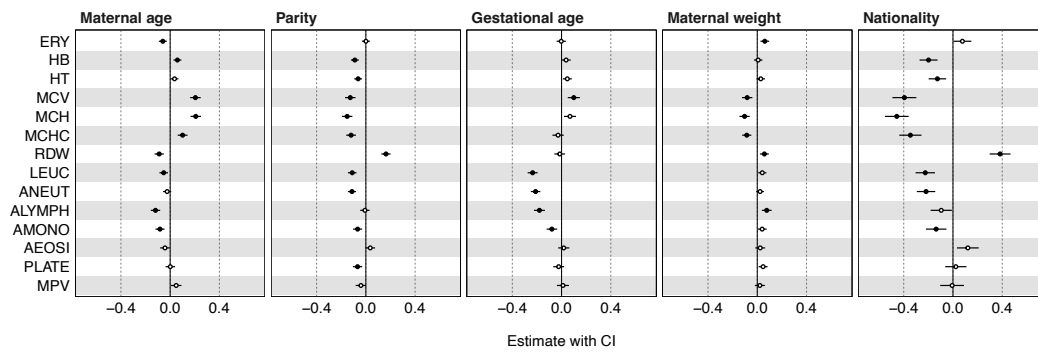


Figure S8. Covariates included in the cubic polynomial regression. Forest plots display estimates with 95% CI for covariates included in the cubic polynomial regression. Solid points show significant effects at P-value $\leq 0.05/14 = 0.0036$ adjusting the significance level with the Bonferroni correction for the 14 phenotypes tested.

Table S1. Summary distribution of CBC values per trimester and pregnancy group.

trimester	group	test	min	1stQ	median	mean	3rdQ	max
3	control	AEOSI	0	0.05	0.08	0.105	0.13	1.31
3	control	ERY	2.7	3.9	4.1	4.1	4.3	5.6
3	control	HT	24.0	34.0	36.0	36.3	38.0	47.0
3	control	HB	81.0	115.0	122.0	121.5	129.0	162.0
3	control	LEUC	2.4	7.8	9.4	9.7	11.2	28.5
3	control	ALYMPH	0.4	1.5	1.8	1.9	2.2	4.9
3	control	MCH	20.3	28.5	29.9	29.7	31.0	35.4
3	control	MCHC	299.0	329.0	335.0	335.3	341.0	364.0
3	control	MCV	65.0	86.0	89.0	88.4	92.0	103.0
3	control	AMONO	0.1	0.6	0.7	0.7	0.8	2.6
3	control	MPV	8.5	10.3	11.0	11.1	11.7	15.6
3	control	ANEUT	0.3	5.4	6.6	7.0	8.2	23.1
3	control	PLAQ	46.0	173.0	208.0	213.9	249.0	542.0
3	control	RDW	11.6	13.1	13.6	13.9	14.2	26.0
3	HDP	AEOSI	0.0	0.0	0.1	0.1	0.1	0.6
3	HDP	ERY	1.7	3.6	4.0	3.9	4.2	5.2
3	HDP	HT	15.0	32.0	35.0	34.2	37.0	44.0
3	HDP	HB	52.0	106.0	116.0	114.7	125.0	151.0
3	HDP	LEUC	3.0	8.1	10.1	10.3	12.0	27.0
3	HDP	ALYMPH	0.4	1.5	1.8	1.9	2.3	6.0
3	HDP	MCH	24.1	28.6	29.8	29.5	30.6	33.3
3	HDP	MCHC	302.0	330.0	336.0	336.1	343.0	358.0
3	HDP	MCV	76.0	86.0	88.0	87.7	90.0	99.0
3	HDP	AMONO	0.1	0.5	0.6	0.7	0.8	1.7
3	HDP	MPV	8.7	10.4	11.2	11.3	12.0	14.3
3	HDP	ANEUT	1.3	5.3	7.3	7.5	9.1	23.5
3	HDP	PLAQ	36.0	141.0	179.0	188.5	221.0	518.0
3	HDP	RDW	11.9	13.2	13.6	13.9	14.2	21.0
3	GDM	AEOSI	0.0	0.0	0.1	0.1	0.1	0.7
3	GDM	ERY	2.4	3.9	4.2	4.1	4.4	5.3
3	GDM	HT	22.0	33.0	36.0	35.2	38.0	44.0
3	GDM	HB	70.0	109.0	119.0	117.2	125.0	156.0
3	GDM	LEUC	4.9	7.6	9.0	9.5	10.7	24.0
3	GDM	ALYMPH	0.4	1.4	1.7	1.8	2.2	3.8
3	GDM	MCH	21.5	27.4	29.1	28.8	30.4	33.1
3	GDM	MCHC	307.0	328.0	333.0	333.4	339.0	358.0
3	GDM	MCV	66.0	83.0	87.0	86.4	91.0	96.0
3	GDM	AMONO	0.1	0.6	0.7	0.7	0.8	1.9
3	GDM	MPV	8.8	10.4	11.0	11.1	11.6	13.8
3	GDM	ANEUT	1.7	5.1	6.3	6.8	7.9	19.7
3	GDM	PLAQ	71.0	165.0	212.0	211.4	255.0	734.0
3	GDM	RDW	12.1	13.5	14.1	14.5	14.9	23.8
3	PPH	AEOSI	0.0	0.0	0.1	0.1	0.1	1.2
3	PPH	ERY	1.4	3.1	3.8	3.6	4.1	5.3
3	PPH	HT	12.0	28.0	33.0	31.9	37.0	46.0
3	PPH	HB	41.0	92.0	110.0	106.8	122.0	157.0

3	PPH	LEUC	2.0	8.4	10.6	11.3	13.6	30.5
3	PPH	ALYMPH	0.2	1.4	1.7	1.8	2.2	5.2
3	PPH	MCH	20.0	28.4	29.9	29.7	31.1	38.2
3	PPH	MCHC	302.0	329.0	336.0	335.7	342.0	364.0
3	PPH	MCV	66.0	86.0	89.0	88.3	92.0	110.0
3	PPH	AMONO	0.1	0.6	0.7	0.8	0.9	2.8
3	PPH	MPV	8.6	10.1	10.8	10.9	11.6	14.6
3	PPH	ANEUT	0.1	5.8	7.8	8.5	10.5	28.4
3	PPH	PLAQ	39.0	154.0	191.0	196.9	234.0	557.0
3	PPH	RDW	11.5	13.2	13.8	14.2	14.7	24.7
2	control	AEOSI	0.0	0.1	0.1	0.1	0.2	2.0
2	control	ERY	2.1	3.7	3.9	3.9	4.2	4.9
2	control	HT	19.0	33.0	35.0	34.3	36.0	41.0
2	control	HB	65.0	109.0	116.5	115.2	122.0	142.0
2	control	LEUC	2.3	7.6	9.2	9.6	11.1	21.3
2	control	ALYMPH	0.4	1.4	1.7	1.8	2.1	4.2
2	control	MCH	19.7	28.7	29.9	29.6	30.9	33.7
2	control	MCHC	295.0	330.0	337.0	336.6	343.0	364.0
2	control	MCV	64.0	86.0	89.0	87.9	91.0	98.0
2	control	AMONO	0.1	0.5	0.6	0.7	0.8	2.0
2	control	MPV	8.4	10.1	10.5	10.6	11.2	14.5
2	control	ANEUT	1.5	5.2	6.6	6.9	8.1	18.3
2	control	PLAQ	49.0	188.0	226.0	227.4	262.0	549.0
2	control	RDW	12.1	13.0	13.5	13.8	14.1	19.8
2	HDP	AEOSI	0.0	0.1	0.1	0.1	0.1	0.3
2	HDP	ERY	3.5	3.8	3.9	4.0	4.1	4.8
2	HDP	HT	31.0	33.8	35.0	35.2	37.0	40.0
2	HDP	HB	108.0	114.0	118.0	119.2	124.0	134.0
2	HDP	LEUC	3.0	9.4	10.8	10.5	12.3	15.8
2	HDP	ALYMPH	1.0	1.5	1.8	1.8	2.1	3.1
2	HDP	MCH	26.0	29.5	30.3	30.0	30.8	32.7
2	HDP	MCHC	322.0	332.8	339.0	338.6	344.5	358.0
2	HDP	MCV	80.0	86.0	88.5	88.7	92.0	97.0
2	HDP	AMONO	0.3	0.5	0.7	0.7	0.9	1.3
2	HDP	MPV	9.1	9.5	10.3	10.4	11.1	12.6
2	HDP	ANEUT	1.7	6.3	7.6	7.8	9.5	14.1
2	HDP	PLAQ	122.0	187.3	221.5	221.4	244.0	460.0
2	HDP	RDW	12.6	13.0	13.4	13.4	13.8	14.7
2	GDM	AEOSI	0.0	0.1	0.1	0.2	0.2	0.3
2	GDM	ERY	3.4	3.8	4.0	4.0	4.3	4.5
2	GDM	HT	29.0	33.5	35.0	34.6	36.0	39.0
2	GDM	HB	98.0	110.5	117.0	116.4	121.0	134.0
2	GDM	LEUC	5.9	8.3	9.6	10.2	12.1	15.9
2	GDM	ALYMPH	0.9	1.4	1.9	1.9	2.3	3.8
2	GDM	MCH	24.6	27.9	29.6	29.0	30.2	31.8
2	GDM	MCHC	315.0	330.5	337.0	335.7	343.0	351.0
2	GDM	MCV	78.0	84.0	88.0	86.4	89.0	91.0
2	GDM	AMONO	0.3	0.5	0.7	0.7	0.8	1.1

2	GDM	MPV	9.2	10.1	10.6	10.7	11.1	13.0
2	GDM	ANEUT	4.1	5.9	7.2	7.3	8.9	12.4
2	GDM	PLAQ	136.0	175.0	226.0	226.1	275.5	318.0
2	GDM	RDW	12.5	13.5	13.6	14.0	14.3	16.5
2	PPH	AEOSI	0.0	0.1	0.1	0.1	0.2	0.5
2	PPH	ERY	2.8	3.7	3.9	3.9	4.2	4.8
2	PPH	HT	25.0	33.0	34.0	34.3	36.0	41.0
2	PPH	HB	86.0	109.3	116.0	115.6	123.0	138.0
2	PPH	LEUC	2.2	7.2	9.2	9.1	10.6	25.4
2	PPH	ALYMPH	0.5	1.4	1.6	1.7	2.0	3.5
2	PPH	MCH	22.9	28.8	29.7	29.4	30.5	34.6
2	PPH	MCHC	307.0	330.3	337.0	337.2	344.8	361.0
2	PPH	MCV	73.0	85.0	88.0	87.4	90.0	100.0
2	PPH	AMONO	0.2	0.5	0.6	0.6	0.7	1.5
2	PPH	MPV	9.1	9.9	10.5	10.6	11.1	14.0
2	PPH	ANEUT	0.2	5.0	6.6	6.6	7.8	23.6
2	PPH	PLAQ	98.0	187.0	215.0	214.5	250.8	324.0
2	PPH	RDW	11.6	13.0	13.5	13.9	14.1	25.6
1	control	AEOSI	0.0	0.0	0.1	0.1	0.2	1.3
1	control	ERY	3.1	4.1	4.3	4.4	4.6	5.7
1	control	HT	27.0	36.0	37.0	37.4	39.0	47.0
1	control	HB	88.0	120.0	127.0	126.3	133.0	161.0
1	control	LEUC	2.9	6.7	7.9	8.3	9.7	20.9
1	control	ALYMPH	0.4	1.6	2.0	2.0	2.4	5.7
1	control	MCH	22.2	28.1	29.3	29.0	30.3	34.6
1	control	MCHC	300.0	332.0	338.0	338.2	344.0	377.0
1	control	MCV	67.0	84.0	86.0	85.8	89.0	96.0
1	control	AMONO	0.2	0.5	0.6	0.6	0.7	1.3
1	control	MPV	8.4	10.1	10.7	10.7	11.3	13.6
1	control	ANEUT	1.4	4.0	5.2	5.5	6.6	18.9
1	control	PLAQ	28.0	198.0	235.0	237.3	274.0	425.0
1	control	RDW	11.2	12.7	13.1	13.3	13.6	19.1
1	HDP	AEOSI	0.0	0.1	0.1	0.1	0.1	0.4
1	HDP	ERY	3.9	4.3	4.4	4.5	4.7	5.0
1	HDP	HT	35.0	37.0	39.0	38.6	40.0	42.0
1	HDP	HB	118.0	127.0	131.0	131.1	135.0	141.0
1	HDP	LEUC	4.0	6.6	7.7	8.1	9.4	15.7
1	HDP	ALYMPH	1.1	1.6	2.0	2.0	2.4	3.0
1	HDP	MCH	27.1	28.1	29.3	29.3	30.5	32.2
1	HDP	MCHC	330.0	335.0	340.0	340.0	343.5	353.0
1	HDP	MCV	80.0	83.5	86.0	86.3	89.5	94.0
1	HDP	AMONO	0.3	0.5	0.6	0.6	0.6	1.5
1	HDP	MPV	9.0	10.2	10.7	10.7	11.3	13.2
1	HDP	ANEUT	1.9	3.8	4.6	5.3	6.4	13.1
1	HDP	PLAQ	159.0	203.0	230.0	238.7	259.0	502.0
1	HDP	RDW	11.8	12.8	13.2	13.1	13.6	14.2
1	GDM	AEOSI	0.0	0.1	0.1	0.1	0.2	0.3
1	GDM	ERY	3.8	4.3	4.5	4.5	4.7	5.2

1	GDM	HT	34.0	36.0	38.0	37.9	39.0	41.0
1	GDM	HB	112.0	122.5	128.0	128.4	134.0	142.0
1	GDM	LEUC	4.5	6.8	9.7	9.2	10.9	15.6
1	GDM	ALYMPH	1.2	1.7	2.1	2.2	2.7	3.8
1	GDM	MCH	25.2	27.8	28.8	28.7	29.6	33.1
1	GDM	MCHC	321.0	333.5	340.0	339.6	346.5	355.0
1	GDM	MCV	77.0	82.0	84.0	84.5	87.0	95.0
1	GDM	AMONO	0.3	0.5	0.6	0.6	0.8	1.1
1	GDM	MPV	8.8	9.9	10.3	10.4	10.7	13.3
1	GDM	ANEUT	2.4	4.3	6.1	6.2	7.7	11.8
1	GDM	PLAQ	150.0	229.0	272.0	264.3	291.0	383.0
1	GDM	RDW	12.2	12.8	13.1	13.3	13.7	15.2
1	PPH	AEOSI	0.0	0.1	0.1	0.1	0.2	0.8
1	PPH	ERY	3.5	4.2	4.4	4.4	4.6	5.5
1	PPH	HT	28.0	36.0	37.0	37.6	40.0	45.0
1	PPH	HB	86.0	122.0	127.0	126.5	134.0	154.0
1	PPH	LEUC	4.2	6.4	7.9	7.9	9.0	12.8
1	PPH	ALYMPH	1.0	1.6	2.0	2.1	2.4	3.8
1	PPH	MCH	18.9	28.2	29.1	28.7	30.4	33.1
1	PPH	MCHC	290.0	333.0	338.0	336.7	343.0	359.0
1	PPH	MCV	64.0	84.0	86.0	85.3	89.0	96.0
1	PPH	AMONO	0.3	0.5	0.5	0.6	0.7	1.2
1	PPH	MPV	9.2	10.0	10.6	10.7	11.1	14.9
1	PPH	ANEUT	1.4	3.9	4.9	5.1	6.0	9.8
1	PPH	PLAQ	98.0	208.0	241.0	242.6	276.3	414.0
1	PPH	RDW	11.7	12.6	12.9	13.5	13.7	20.8

Table S2. CBC standard deviation.

test	sd
AEOSI	0.107
ERY	0.503
HT	4.215
HB	14.715
LEUC	3.223
ALYMPH	0.613
MCH	2.057
MCHC	9.567
MCV	5.005
AMONO	0.264
MPV	1.028
ANEUT	2.946
PLAQ	65.835
RDW	1.404

Table S3. Cubic polynomial regression summary.

< 0.0036

test	term	estimate	std.error	statistic	df	p.value	2.5%	97.5%
AEOSI	(Intercept)	-0.003	0.023	-0.145	6791.1	8.84E-01	-0.048	0.041
AEOSI	weeks1	-4.320	1.048	-4.127	6797.2	3.67E-05	-6.380	-2.270
AEOSI	weeks2	-1.330	1.096	-1.212	6797.2	2.26E-01	-3.480	0.821
AEOSI	weeks3	1.940	1.025	1.895	6797.2	5.81E-02	-0.067	3.949
AEOSI	GDM	0.037	0.084	0.441	6743.1	6.59E-01	-0.128	0.202
AEOSI	PPH	-0.060	0.049	-1.223	6796.7	2.21E-01	-0.156	0.036
AEOSI	HDP	-0.097	0.085	-1.146	6678.8	2.52E-01	-0.264	0.069
AEOSI	age_diam	-0.042	0.020	-2.104	6796.8	3.54E-02	-0.081	-0.003
AEOSI	parity	0.036	0.020	1.794	6758	7.29E-02	-0.003	0.075
AEOSI	age_gest	0.018	0.023	0.765	6589.9	4.44E-01	-0.028	0.063
AEOSI	weight_gest	0.024	0.020	1.181	127.2	2.37E-01	-0.016	0.064
AEOSI	EUR	0.121	0.045	2.681	6781.8	7.34E-03	0.033	0.209
AEOSI	weeks1:GDM	4.930	3.658	1.348	6797.2	1.78E-01	-2.240	12.101
AEOSI	weeks2:GDM	5.890	3.749	1.572	6797.2	1.16E-01	-1.460	13.242
AEOSI	weeks3:GDM	13.300	3.873	3.439	6797.2	5.83E-04	5.730	20.910
AEOSI	weeks1:PPH	-5.140	2.269	-2.265	6797.2	2.35E-02	-9.590	-0.691
AEOSI	weeks2:PPH	3.260	2.200	1.482	6797.2	1.38E-01	-1.050	7.572
AEOSI	weeks3:PPH	-0.306	2.213	-0.138	6797.2	8.90E-01	-4.640	4.032
AEOSI	weeks1:HDP	7.530	3.718	2.025	6797.2	4.29E-02	0.240	14.815
AEOSI	weeks2:HDP	11.300	3.510	3.206	6797.2	1.35E-03	4.370	18.136
AEOSI	weeks3:HDP	-3.790	3.374	-1.123	6797.2	2.61E-01	-10.400	2.825
ERY	(Intercept)	0.234	0.019	12.439	6793.1	1.60E-35	0.197	0.271
ERY	weeks1	-4.760	0.913	-5.208	6796.3	1.91E-07	-6.550	-2.966
ERY	weeks2	22.400	0.958	23.360	6795.4	1.10E-120	20.500	24.245
ERY	weeks3	2.900	0.896	3.242	6797.2	1.19E-03	1.150	4.660
ERY	GDM	0.108	0.069	1.562	6664.4	1.18E-01	-0.028	0.244
ERY	PPH	-0.522	0.040	-12.946	6793	2.48E-38	-0.601	-0.443
ERY	HDP	-0.162	0.069	-2.342	6749.3	1.92E-02	-0.298	-0.026
ERY	age_diam	-0.059	0.016	-3.613	6781.2	3.02E-04	-0.091	-0.027
ERY	parity	0.000	0.016	0.027	6686.2	9.79E-01	-0.032	0.033
ERY	age_gest	-0.003	0.019	-0.146	5963.1	8.84E-01	-0.040	0.035
ERY	weight_gest	0.061	0.018	3.415	49.9	6.37E-04	0.025	0.098
ERY	EUR	0.077	0.037	2.062	6790.5	3.92E-02	0.004	0.149
ERY	weeks1:GDM	-9.700	3.202	-3.031	6797.2	2.44E-03	-16.000	-3.428
ERY	weeks2:GDM	0.516	3.285	0.157	6797.2	8.75E-01	-5.920	6.956
ERY	weeks3:GDM	-1.600	3.394	-0.471	6797.1	6.38E-01	-8.250	5.056
ERY	weeks1:PPH	-34.700	1.984	-17.473	6796.7	2.28E-68	-38.600	-30.780
ERY	weeks2:PPH	-33.500	1.924	-17.438	6796.7	4.22E-68	-37.300	-29.777
ERY	weeks3:PPH	-22.900	1.937	-11.823	6797.2	2.98E-32	-26.700	-19.103
ERY	weeks1:HDP	-21.700	3.261	-6.667	6797.2	2.61E-11	-28.100	-15.348
ERY	weeks2:HDP	-4.660	3.056	-1.525	6794.6	1.27E-01	-10.600	1.331
ERY	weeks3:HDP	11.000	2.955	3.721	6796	1.98E-04	5.200	16.786
HT	(Intercept)	0.314	0.019	16.976	6788.5	1.24E-64	0.278	0.350
HT	weeks1	-1.340	0.959	-1.399	6797.2	1.62E-01	-3.220	0.538
HT	weeks2	19.500	1.008	19.321	6797.2	3.60E-83	17.500	21.458
HT	weeks3	1.250	0.944	1.319	6797.2	1.87E-01	-0.606	3.097

HT	GDM	-0.080	0.068	-1.178	6644.4	2.39E-01	-0.212	0.053
HT	PPH	-0.591	0.039	-15.032	6793.9	4.53E-51	-0.668	-0.514
HT	HDP	-0.206	0.067	-3.070	6619.4	2.14E-03	-0.337	-0.074
HT	age_diam	0.035	0.016	2.207	6795.9	2.73E-02	0.004	0.067
HT	parity	-0.063	0.016	-3.935	6714.1	8.31E-05	-0.094	-0.032
HT	age_gest	0.047	0.019	2.493	6337.9	1.27E-02	0.010	0.083
HT	weight_gest	0.028	0.018	1.624	53.4	1.04E-01	-0.007	0.063
HT	EUR	-0.126	0.036	-3.487	6744.8	4.89E-04	-0.198	-0.055
HT	weeks1:GDM	-11.900	3.380	-3.520	6797.2	4.32E-04	-18.500	-5.271
HT	weeks2:GDM	-0.210	3.475	-0.061	6797.2	9.52E-01	-7.020	6.601
HT	weeks3:GDM	-2.300	3.592	-0.640	6797.2	5.22E-01	-9.340	4.742
HT	weeks1:PPH	-34.300	2.092	-16.392	6797.2	2.19E-60	-38.400	-30.192
HT	weeks2:PPH	-33.200	2.028	-16.380	6797.1	2.66E-60	-37.200	-29.249
HT	weeks3:PPH	-23.700	2.045	-11.581	6797.2	5.17E-31	-27.700	-19.675
HT	weeks1:HDP	-25.300	3.454	-7.340	6797.2	2.14E-13	-32.100	-18.578
HT	weeks2:HDP	-5.890	3.199	-1.840	6797.2	6.58E-02	-12.200	0.385
HT	weeks3:HDP	10.800	3.121	3.465	6797.2	5.31E-04	4.690	16.930
HB	(Intercept)	0.329	0.019	17.381	6786.4	1.16E-67	0.292	0.366
HB	weeks1	-3.990	0.946	-4.218	6797.2	2.46E-05	-5.840	-2.136
HB	weeks2	18.900	0.993	19.063	6797.2	5.16E-81	17.000	20.876
HB	weeks3	2.910	0.929	3.135	6797.2	1.72E-03	1.090	4.736
HB	GDM	-0.098	0.069	-1.419	6696.1	1.56E-01	-0.234	0.038
HB	PPH	-0.567	0.040	-14.024	6795.1	1.11E-44	-0.646	-0.488
HB	HDP	-0.203	0.069	-2.936	6574.3	3.33E-03	-0.339	-0.068
HB	age_diam	0.059	0.017	3.576	6797.2	3.48E-04	0.027	0.091
HB	parity	-0.089	0.016	-5.438	6739.3	5.37E-08	-0.121	-0.057
HB	age_gest	0.037	0.019	1.912	6474.8	5.59E-02	-0.001	0.074
HB	weight_gest	0.008	0.018	0.432	68.3	6.65E-01	-0.028	0.043
HB	EUR	-0.198	0.037	-5.316	6719.5	1.06E-07	-0.271	-0.125
HB	weeks1:GDM	-13.900	3.324	-4.183	6797.2	2.88E-05	-20.400	-7.388
HB	weeks2:GDM	-1.500	3.414	-0.440	6797.2	6.60E-01	-8.190	5.189
HB	weeks3:GDM	-3.100	3.527	-0.878	6797.2	3.80E-01	-10.000	3.817
HB	weeks1:PPH	-32.400	2.059	-15.729	6797.2	9.57E-56	-36.400	-28.348
HB	weeks2:PPH	-32.500	1.996	-16.304	6797.2	9.30E-60	-36.500	-28.633
HB	weeks3:PPH	-22.700	2.011	-11.310	6797.2	1.17E-29	-26.700	-18.804
HB	weeks1:HDP	-25.500	3.390	-7.516	6797.2	5.66E-14	-32.100	-18.832
HB	weeks2:HDP	-7.770	3.161	-2.459	6797.2	1.39E-02	-14.000	-1.577
HB	weeks3:HDP	8.420	3.068	2.745	6797.2	6.06E-03	2.410	14.434
LEUC	(Intercept)	-0.013	0.020	-0.640	6783.8	5.22E-01	-0.053	0.027
LEUC	weeks1	12.700	1.030	12.291	6797.2	1.02E-34	10.600	14.673
LEUC	weeks2	-0.957	1.081	-0.885	6797.2	3.76E-01	-3.080	1.163
LEUC	weeks3	3.860	1.012	3.812	6797.2	1.38E-04	1.870	5.843
LEUC	GDM	0.082	0.075	1.105	6777.4	2.69E-01	-0.064	0.229
LEUC	PPH	0.256	0.044	5.874	6797.2	4.25E-09	0.170	0.341
LEUC	HDP	-0.124	0.074	-1.673	6793.3	9.44E-02	-0.270	0.021
LEUC	age_diam	-0.053	0.018	-2.962	6792	3.06E-03	-0.087	-0.018
LEUC	parity	-0.111	0.018	-6.267	6789	3.69E-10	-0.145	-0.076
LEUC	age_gest	-0.235	0.021	-11.383	6723.6	5.11E-30	-0.275	-0.194

LEUC	weight_gest	0.041	0.017	2.393	830	1.67E-02	0.007	0.074
LEUC	EUR	-0.225	0.040	-5.606	6711.6	2.08E-08	-0.303	-0.146
LEUC	weeks1:GDM	-11.700	3.621	-3.230	6797.2	1.24E-03	-18.800	-4.597
LEUC	weeks2:GDM	-6.060	3.720	-1.629	6797.2	1.03E-01	-13.400	1.231
LEUC	weeks3:GDM	1.220	3.844	0.318	6797.2	7.51E-01	-6.310	8.758
LEUC	weeks1:PPH	18.500	2.243	8.233	6797.2	1.83E-16	14.100	22.859
LEUC	weeks2:PPH	16.600	2.174	7.627	6797.2	2.41E-14	12.300	20.844
LEUC	weeks3:PPH	7.610	2.191	3.474	6797.2	5.13E-04	3.320	11.906
LEUC	weeks1:HDP	-2.810	3.695	-0.760	6797.2	4.47E-01	-10.100	4.434
LEUC	weeks2:HDP	-8.750	3.440	-2.544	6797	1.10E-02	-15.500	-2.006
LEUC	weeks3:HDP	-3.840	3.343	-1.148	6797.2	2.51E-01	-10.400	2.714
ALYMPH	(Intercept)	0.045	0.022	2.045	6788.9	4.09E-02	0.002	0.089
ALYMPH	weeks1	-2.210	1.022	-2.165	6796.4	3.04E-02	-4.220	-0.209
ALYMPH	weeks2	12.300	1.069	11.481	6796.1	1.64E-30	10.200	14.368
ALYMPH	weeks3	0.132	0.999	0.132	6797.2	8.95E-01	-1.830	2.091
ALYMPH	GDM	0.035	0.082	0.427	6655.9	6.69E-01	-0.126	0.196
ALYMPH	PPH	0.000	0.048	0.003	6789.8	9.98E-01	-0.094	0.094
ALYMPH	HDP	-0.041	0.083	-0.491	6672.7	6.23E-01	-0.203	0.122
ALYMPH	age_diam	-0.119	0.019	-6.131	6776.3	8.74E-10	-0.157	-0.081
ALYMPH	parity	-0.008	0.019	-0.393	6713.8	6.95E-01	-0.046	0.030
ALYMPH	age_gest	-0.180	0.023	-7.950	6448.2	1.86E-15	-0.224	-0.135
ALYMPH	weight_gest	0.077	0.020	3.824	103.7	1.31E-04	0.037	0.117
ALYMPH	EUR	-0.095	0.044	-2.167	6782.6	3.02E-02	-0.182	-0.009
ALYMPH	weeks1:GDM	-8.900	3.568	-2.495	6797.2	1.26E-02	-15.900	-1.909
ALYMPH	weeks2:GDM	4.310	3.656	1.179	6797.2	2.38E-01	-2.860	11.480
ALYMPH	weeks3:GDM	6.710	3.777	1.776	6797.2	7.58E-02	-0.697	14.110
ALYMPH	weeks1:PPH	-5.120	2.213	-2.312	6796.6	2.08E-02	-9.450	-0.779
ALYMPH	weeks2:PPH	-1.090	2.146	-0.507	6797.2	6.12E-01	-5.290	3.118
ALYMPH	weeks3:PPH	-1.920	2.158	-0.888	6797.2	3.74E-01	-6.150	2.314
ALYMPH	weeks1:HDP	-2.150	3.626	-0.592	6797.2	5.54E-01	-9.250	4.962
ALYMPH	weeks2:HDP	2.630	3.424	0.769	6793.9	4.42E-01	-4.080	9.346
ALYMPH	weeks3:HDP	-1.470	3.291	-0.447	6796.7	6.55E-01	-7.920	4.978
MCH	(Intercept)	0.173	0.024	7.201	6789.8	5.97E-13	0.126	0.221
MCH	weeks1	-0.292	0.585	-0.500	6797.2	6.17E-01	-1.440	0.854
MCH	weeks2	-6.930	0.604	-11.477	6797.2	1.73E-30	-8.120	-5.750
MCH	weeks3	0.432	0.562	0.770	6797.2	4.41E-01	-0.669	1.534
MCH	GDM	-0.368	0.092	-4.002	6794.2	6.28E-05	-0.548	-0.188
MCH	PPH	-0.062	0.054	-1.151	6773.4	2.50E-01	-0.167	0.044
MCH	HDP	-0.081	0.097	-0.833	6672.8	4.05E-01	-0.271	0.109
MCH	age_diam	0.208	0.022	9.613	6770.1	7.02E-22	0.166	0.251
MCH	parity	-0.152	0.022	-7.044	6784.4	1.87E-12	-0.194	-0.110
MCH	age_gest	0.069	0.025	2.748	6554.9	5.99E-03	0.020	0.117
MCH	weight_gest	-0.103	0.021	-4.877	341.5	1.08E-06	-0.145	-0.061
MCH	EUR	-0.456	0.049	-9.275	6636	1.78E-20	-0.553	-0.360
MCH	weeks1:GDM	-7.060	1.992	-3.543	6797.2	3.96E-04	-11.000	-3.152
MCH	weeks2:GDM	-3.110	2.026	-1.533	6797.2	1.25E-01	-7.080	0.866
MCH	weeks3:GDM	-3.000	2.090	-1.434	6797.2	1.52E-01	-7.090	1.101
MCH	weeks1:PPH	5.040	1.240	4.064	6797.2	4.83E-05	2.610	7.467

MCH	weeks2:PPH	2.970	1.203	2.471	6797.2	1.35E-02	0.614	5.330
MCH	weeks3:PPH	-0.862	1.203	-0.716	6797.2	4.74E-01	-3.220	1.496
MCH	weeks1:HDP	-3.990	1.998	-1.997	6797.2	4.59E-02	-7.900	-0.073
MCH	weeks2:HDP	-2.570	1.966	-1.306	6796.8	1.92E-01	-6.420	1.286
MCH	weeks3:HDP	-4.990	1.828	-2.727	6797.2	6.38E-03	-8.570	-1.402
MCHC	(Intercept)	0.113	0.023	4.929	6791.6	8.25E-07	0.068	0.158
MCHC	weeks1	-11.800	0.891	-13.278	6796.5	3.09E-40	-13.600	-10.078
MCHC	weeks2	-0.303	0.927	-0.327	6796.7	7.44E-01	-2.120	1.513
MCHC	weeks3	7.300	0.864	8.450	6797.2	2.91E-17	5.610	8.998
MCHC	GDM	-0.114	0.086	-1.327	6791.1	1.84E-01	-0.282	0.054
MCHC	PPH	-0.007	0.050	-0.142	6789.3	8.87E-01	-0.105	0.091
MCHC	HDP	-0.032	0.088	-0.366	6534.9	7.15E-01	-0.205	0.141
MCHC	age_diam	0.101	0.020	5.004	6779.8	5.61E-07	0.062	0.141
MCHC	parity	-0.120	0.020	-5.947	6793	2.74E-09	-0.160	-0.080
MCHC	age_gest	-0.028	0.024	-1.208	6587.3	2.27E-01	-0.074	0.018
MCHC	weight_gest	-0.085	0.019	-4.501	4201.7	6.76E-06	-0.122	-0.048
MCHC	EUR	-0.346	0.046	-7.535	6780.8	4.88E-14	-0.436	-0.256
MCHC	weeks1:GDM	-7.560	3.078	-2.455	6797.2	1.41E-02	-13.600	-1.522
MCHC	weeks2:GDM	-4.870	3.145	-1.549	6797.2	1.21E-01	-11.000	1.293
MCHC	weeks3:GDM	-5.660	3.246	-1.745	6797.2	8.10E-02	-12.000	0.700
MCHC	weeks1:PPH	2.980	1.912	1.558	6797	1.19E-01	-0.770	6.726
MCHC	weeks2:PPH	0.268	1.854	0.145	6797.2	8.85E-01	-3.370	3.903
MCHC	weeks3:PPH	-0.208	1.861	-0.112	6797.2	9.11E-01	-3.860	3.440
MCHC	weeks1:HDP	-2.920	3.111	-0.938	6797.2	3.48E-01	-9.010	3.181
MCHC	weeks2:HDP	-9.160	2.991	-3.062	6793.9	2.20E-03	-15.000	-3.294
MCHC	weeks3:HDP	-11.000	2.834	-3.872	6796.3	1.08E-04	-16.500	-5.417
MCV	(Intercept)	0.156	0.024	6.443	6793.2	1.17E-10	0.109	0.204
MCV	weeks1	5.500	0.577	9.538	6797.2	1.46E-21	4.370	6.630
MCV	weeks2	-8.420	0.596	-14.140	6797.2	2.17E-45	-9.590	-7.256
MCV	weeks3	-3.250	0.554	-5.875	6797.2	4.22E-09	-4.340	-2.169
MCV	GDM	-0.385	0.093	-4.154	6787.3	3.26E-05	-0.566	-0.203
MCV	PPH	-0.076	0.054	-1.402	6782.3	1.61E-01	-0.182	0.030
MCV	HDP	-0.077	0.098	-0.787	6743.4	4.31E-01	-0.268	0.115
MCV	age_diam	0.205	0.022	9.411	6784.2	4.91E-21	0.162	0.248
MCV	parity	-0.127	0.022	-5.831	6782.2	5.52E-09	-0.169	-0.084
MCV	age_gest	0.099	0.025	3.960	6651.5	7.49E-05	0.050	0.149
MCV	weight_gest	-0.082	0.022	-3.803	246.3	1.43E-04	-0.124	-0.039
MCV	EUR	-0.394	0.050	-7.958	6562.1	1.74E-15	-0.492	-0.297
MCV	weeks1:GDM	-5.490	1.963	-2.798	6797.2	5.14E-03	-9.340	-1.645
MCV	weeks2:GDM	-1.360	1.997	-0.683	6797.2	4.95E-01	-5.280	2.551
MCV	weeks3:GDM	-0.831	2.060	-0.403	6797.2	6.87E-01	-4.870	3.207
MCV	weeks1:PPH	4.930	1.222	4.033	6797.2	5.50E-05	2.530	7.324
MCV	weeks2:PPH	3.320	1.186	2.798	6797.2	5.14E-03	0.994	5.643
MCV	weeks3:PPH	-0.249	1.186	-0.210	6797.2	8.33E-01	-2.570	2.075
MCV	weeks1:HDP	-4.010	1.969	-2.039	6797.2	4.14E-02	-7.870	-0.155
MCV	weeks2:HDP	2.380	1.938	1.229	6797.2	2.19E-01	-1.420	6.181
MCV	weeks3:HDP	-0.157	1.802	-0.087	6797.2	9.31E-01	-3.690	3.376
AMONO	(Intercept)	-0.008	0.022	-0.349	6790.1	7.27E-01	-0.050	0.035

AMONO	weeks1	13.500	1.042	13.000	6797.2	1.23E-38	11.500	15.584
AMONO	weeks2	-2.280	1.092	-2.089	6797.2	3.67E-02	-4.420	-0.141
AMONO	weeks3	-3.480	1.022	-3.406	6797.2	6.58E-04	-5.480	-1.477
AMONO	GDM	0.074	0.079	0.937	6777.3	3.49E-01	-0.081	0.229
AMONO	PPH	0.155	0.046	3.359	6793	7.82E-04	0.065	0.245
AMONO	HDP	-0.172	0.079	-2.164	6757.3	3.04E-02	-0.327	-0.016
AMONO	age_diam	-0.082	0.019	-4.392	6793.8	1.13E-05	-0.119	-0.046
AMONO	parity	-0.067	0.019	-3.575	6782.9	3.50E-04	-0.103	-0.030
AMONO	age_gest	-0.079	0.022	-3.606	6751.8	3.11E-04	-0.121	-0.036
AMONO	weight_gest	0.040	0.019	2.143	281	3.21E-02	0.003	0.076
AMONO	EUR	-0.137	0.043	-3.218	6734	1.29E-03	-0.220	-0.053
AMONO	weeks1:GDM	-7.700	3.651	-2.110	6797.2	3.49E-02	-14.900	-0.547
AMONO	weeks2:GDM	-0.140	3.746	-0.037	6797.2	9.70E-01	-7.480	7.203
AMONO	weeks3:GDM	1.530	3.871	0.395	6797.2	6.93E-01	-6.060	9.116
AMONO	weeks1:PPH	10.700	2.263	4.745	6797.2	2.09E-06	6.300	15.171
AMONO	weeks2:PPH	8.340	2.194	3.802	6797.2	1.44E-04	4.040	12.641
AMONO	weeks3:PPH	4.900	2.209	2.219	6797.2	2.65E-02	0.571	9.231
AMONO	weeks1:HDP	-2.430	3.718	-0.655	6797.2	5.13E-01	-9.720	4.854
AMONO	weeks2:HDP	-3.780	3.485	-1.084	6797.2	2.78E-01	-10.600	3.054
AMONO	weeks3:HDP	-1.470	3.369	-0.437	6797.2	6.62E-01	-8.080	5.131
MPV	(Intercept)	-0.014	0.024	-0.566	6795.4	5.72E-01	-0.061	0.034
MPV	weeks1	14.100	0.693	20.357	6797.2	4.04E-92	12.700	15.462
MPV	weeks2	8.140	0.717	11.347	6797.2	7.71E-30	6.730	9.542
MPV	weeks3	-1.260	0.667	-1.884	6797.2	5.95E-02	-2.570	0.051
MPV	GDM	-0.019	0.092	-0.202	6792.7	8.40E-01	-0.199	0.162
MPV	PPH	-0.063	0.054	-1.165	6797.2	2.44E-01	-0.168	0.043
MPV	HDP	0.161	0.097	1.664	6775.3	9.62E-02	-0.029	0.350
MPV	age_diam	0.049	0.022	2.253	6797.2	2.42E-02	0.006	0.092
MPV	parity	-0.040	0.022	-1.855	6794.5	6.36E-02	-0.083	0.002
MPV	age_gest	0.011	0.025	0.455	6776.6	6.49E-01	-0.038	0.060
MPV	weight_gest	0.021	0.021	1.023	1517.2	3.07E-01	-0.019	0.061
MPV	EUR	-0.007	0.049	-0.141	6778.8	8.88E-01	-0.104	0.090
MPV	weeks1:GDM	-3.780	2.368	-1.595	6797.2	1.11E-01	-8.420	0.865
MPV	weeks2:GDM	-5.600	2.412	-2.321	6797.2	2.03E-02	-10.300	-0.870
MPV	weeks3:GDM	-2.560	2.488	-1.028	6797.2	3.04E-01	-7.440	2.319
MPV	weeks1:PPH	-2.960	1.473	-2.012	6797.2	4.42E-02	-5.850	-0.077
MPV	weeks2:PPH	-7.530	1.430	-5.266	6797.2	1.39E-07	-10.300	-4.726
MPV	weeks3:PPH	-7.250	1.431	-5.067	6797.2	4.04E-07	-10.100	-4.445
MPV	weeks1:HDP	-5.610	2.380	-2.359	6797.2	1.83E-02	-10.300	-0.950
MPV	weeks2:HDP	-16.800	2.329	-7.234	6797.2	4.68E-13	-21.400	-12.283
MPV	weeks3:HDP	-10.200	2.175	-4.694	6797.2	2.68E-06	-14.500	-5.947
ANEUT	(Intercept)	-0.022	0.020	-1.147	6788.5	2.51E-01	-0.061	0.016
ANEUT	weeks1	13.000	1.066	12.173	6797.2	4.33E-34	10.900	15.061
ANEUT	weeks2	-2.890	1.124	-2.568	6797.2	1.02E-02	-5.090	-0.683
ANEUT	weeks3	4.730	1.054	4.494	6797.2	6.98E-06	2.670	6.800
ANEUT	GDM	0.072	0.071	1.013	6792.4	3.11E-01	-0.067	0.211
ANEUT	PPH	0.269	0.041	6.509	6797.2	7.58E-11	0.188	0.350
ANEUT	HDP	-0.099	0.070	-1.424	6797.1	1.55E-01	-0.236	0.037

ANEUT	age_diam	-0.025	0.017	-1.454	6796	1.46E-01	-0.058	0.009
ANEUT	parity	-0.113	0.017	-6.720	6795.3	1.81E-11	-0.146	-0.080
ANEUT	age_gest	-0.211	0.020	-10.720	6775.8	8.20E-27	-0.250	-0.173
ANEUT	weight_gest	0.023	0.016	1.443	2446.3	1.49E-01	-0.008	0.054
ANEUT	EUR	-0.219	0.038	-5.726	6687.4	1.03E-08	-0.293	-0.144
ANEUT	weeks1:GDM	-10.300	3.774	-2.732	6797.2	6.29E-03	-17.700	-2.913
ANEUT	weeks2:GDM	-7.540	3.887	-1.940	6797.2	5.24E-02	-15.200	0.079
ANEUT	weeks3:GDM	-0.774	4.018	-0.193	6797.2	8.47E-01	-8.650	7.103
ANEUT	weeks1:PPH	20.200	2.334	8.658	6797.2	4.79E-18	15.600	24.780
ANEUT	weeks2:PPH	16.900	2.262	7.472	6797.2	7.87E-14	12.500	21.340
ANEUT	weeks3:PPH	7.950	2.284	3.482	6797.2	4.98E-04	3.470	12.429
ANEUT	weeks1:HDP	-2.170	3.867	-0.561	6797.2	5.75E-01	-9.750	5.413
ANEUT	weeks2:HDP	-9.550	3.545	-2.695	6797.2	7.04E-03	-16.500	-2.605
ANEUT	weeks3:HDP	-2.890	3.486	-0.828	6797.2	4.07E-01	-9.720	3.945
PLAQ	(Intercept)	0.102	0.022	4.678	6797.1	2.89E-06	0.059	0.145
PLAQ	weeks1	-9.360	0.811	-11.547	6797.2	7.62E-31	-10.900	-7.771
PLAQ	weeks2	2.550	0.843	3.026	6797.2	2.48E-03	0.898	4.201
PLAQ	weeks3	0.209	0.786	0.266	6797.2	7.90E-01	-1.330	1.749
PLAQ	GDM	0.094	0.082	1.144	6794.7	2.53E-01	-0.067	0.255
PLAQ	PPH	-0.159	0.048	-3.322	6793.8	8.94E-04	-0.253	-0.065
PLAQ	HDP	-0.253	0.085	-2.987	6795.3	2.81E-03	-0.419	-0.087
PLAQ	age_diam	0.001	0.019	0.057	6793.2	9.55E-01	-0.037	0.039
PLAQ	parity	-0.068	0.019	-3.496	6792.6	4.73E-04	-0.105	-0.030
PLAQ	age_gest	-0.024	0.022	-1.061	6772.6	2.89E-01	-0.068	0.020
PLAQ	weight_gest	0.047	0.019	2.563	966.7	1.04E-02	0.011	0.084
PLAQ	EUR	0.023	0.044	0.523	6758.6	6.01E-01	-0.063	0.109
PLAQ	weeks1:GDM	-3.650	2.796	-1.305	6797.2	1.92E-01	-9.130	1.832
PLAQ	weeks2:GDM	10.200	2.854	3.568	6797.2	3.60E-04	4.590	15.777
PLAQ	weeks3:GDM	7.790	2.946	2.646	6797.2	8.15E-03	2.020	13.569
PLAQ	weeks1:PPH	-9.420	1.737	-5.424	6797.2	5.82E-08	-12.800	-6.017
PLAQ	weeks2:PPH	-1.200	1.685	-0.711	6797.2	4.77E-01	-4.500	2.105
PLAQ	weeks3:PPH	1.600	1.690	0.947	6797.2	3.44E-01	-1.710	4.912
PLAQ	weeks1:HDP	1.280	2.822	0.453	6797.2	6.50E-01	-4.250	6.811
PLAQ	weeks2:HDP	16.400	2.723	6.015	6797.2	1.80E-09	11.000	21.720
PLAQ	weeks3:HDP	14.500	2.573	5.644	6797.2	1.66E-08	9.480	19.564
RDW	(Intercept)	-0.211	0.022	-9.804	6794	1.08E-22	-0.253	-0.169
RDW	weeks1	19.400	0.873	22.225	6797	1.99E-109	17.700	21.114
RDW	weeks2	8.370	0.909	9.203	6797.2	3.48E-20	6.590	10.152
RDW	weeks3	2.890	0.849	3.400	6797.2	6.74E-04	1.220	4.549
RDW	GDM	0.297	0.080	3.693	6796.3	2.22E-04	0.139	0.454
RDW	PPH	0.217	0.047	4.613	6791.3	3.97E-06	0.125	0.309
RDW	HDP	0.085	0.082	1.027	6728.1	3.04E-01	-0.077	0.246
RDW	age_diam	-0.089	0.019	-4.687	6791.8	2.78E-06	-0.126	-0.052
RDW	parity	0.163	0.019	8.619	6796.4	6.75E-18	0.126	0.201
RDW	age_gest	-0.015	0.022	-0.676	6748.2	4.99E-01	-0.058	0.028
RDW	weight_gest	0.058	0.018	3.166	846.9	1.55E-03	0.022	0.093
RDW	EUR	0.383	0.043	8.871	6731.1	7.24E-19	0.298	0.467
RDW	weeks1:GDM	5.400	3.024	1.786	6797.2	7.41E-02	-0.527	11.329

RDW	weeks2:GDM	-5.910	3.091	-1.912	6797.2	5.59E-02	-12.000	0.149
RDW	weeks3:GDM	-4.160	3.192	-1.303	6797.2	1.93E-01	-10.400	2.098
RDW	weeks1:PPH	1.590	1.878	0.848	6797.2	3.96E-01	-2.090	5.274
RDW	weeks2:PPH	-0.953	1.821	-0.523	6797.2	6.01E-01	-4.520	2.617
RDW	weeks3:PPH	2.180	1.829	1.192	6797.2	2.33E-01	-1.400	5.765
RDW	weeks1:HDP	6.890	3.060	2.251	6797.2	2.44E-02	0.890	12.885
RDW	weeks2:HDP	1.270	2.931	0.435	6796	6.64E-01	-4.470	7.020
RDW	weeks3:HDP	3.490	2.785	1.253	6796.8	2.10E-01	-1.970	8.950

Table S4. Random effects from the cubic polynomial regression.

test	stddev	Intercept~ IID	Residual~ IID	ICC IID
AEOSI	0.680	0.462	0.515	0.473
ERY	0.543	0.295	0.400	0.424
HT	0.505	0.255	0.454	0.360
HB	0.534	0.285	0.434	0.396
LEUC	0.570	0.325	0.517	0.386
ALYMPH	0.663	0.440	0.490	0.473
MCH	0.863	0.745	0.142	0.840
MCHC	0.741	0.550	0.354	0.609
MCV	0.871	0.759	0.138	0.846
AMONO	0.621	0.386	0.519	0.426
MPV	0.850	0.723	0.203	0.781
ANEUT	0.507	0.258	0.576	0.309
PLAQ	0.720	0.518	0.290	0.641
RDW	0.687	0.472	0.343	0.579

Table S5. Polynomial model comparison.

test	terms	df	AIC	BIC	R2m	R2c
AEOSI	1	15	17556	17659	0.0134	0.479
AEOSI	2	19	17534	17664	0.0164	0.481
AEOSI	3	23	17508	17665	0.0181	0.483
ERY	1	15	16312	16414	0.1803	0.489
ERY	2	19	15676	15805	0.2329	0.548
ERY	3	23	15504	15661	0.2507	0.569
HT	1	15	16666	16769	0.1754	0.442
HT	2	19	16188	16317	0.2192	0.489
HT	3	23	16015	16172	0.2384	0.512
HB	1	15	16510	16612	0.1767	0.475
HB	2	19	16060	16190	0.2153	0.516
HB	3	23	15910	16067	0.2316	0.536
LEUC	1	15	17187	17290	0.1344	0.459
LEUC	2	19	17096	17226	0.1468	0.473
LEUC	3	23	17040	17197	0.1538	0.481
ALYMPH	1	15	17396	17498	0.0387	0.483
ALYMPH	2	19	17180	17310	0.0575	0.504
ALYMPH	3	23	17170	17327	0.0579	0.504
MCH	1	15	12141	12243	0.0971	0.849
MCH	2	19	11933	12063	0.1027	0.856
MCH	3	23	11923	12080	0.103	0.856
MCHC	1	15	15938	16040	0.0632	0.626
MCHC	2	19	15915	16045	0.0659	0.629
MCHC	3	23	15820	15977	0.0733	0.637
MCV	1	15	12111	12213	0.0933	0.852
MCV	2	19	11874	12004	0.0988	0.86
MCV	3	23	11820	11977	0.1001	0.861
AMONO	1	15	17339	17442	0.0711	0.464
AMONO	2	19	17316	17446	0.0735	0.467
AMONO	3	23	17295	17452	0.0751	0.47
MPV	1	15	13813	13915	0.0333	0.779
MPV	2	19	13653	13782	0.0444	0.788
MPV	3	23	13569	13727	0.0478	0.791
ANEUT	1	15	17519	17621	0.1358	0.394
ANEUT	2	19	17427	17557	0.1483	0.408
ANEUT	3	23	17362	17519	0.1567	0.417
PLAQ	1	15	14860	14963	0.0498	0.651
PLAQ	2	19	14758	14887	0.0585	0.66
PLAQ	3	23	14709	14866	0.0623	0.664
RDW	1	15	15558	15661	0.1276	0.624
RDW	2	19	15445	15575	0.1383	0.636
RDW	3	23	15414	15571	0.14	0.638

Table S6. Linear mixed effect model summary.

<0.0036

test	term	estimate	std.error	statistic	df	p.value	2.5%	97.5%
AEOSI	(Intercept)	0.159	0.047	3.396	6805.2	6.88E-04	0.067	0.250
AEOSI	weeks	-0.005	0.001	-4.197	6805.2	2.73E-05	-0.008	-0.003
AEOSI	GDM	-0.074	0.163	-0.456	6801.1	6.48E-01	-0.394	0.245
AEOSI	PPH	0.138	0.104	1.321	6805.2	1.87E-01	-0.067	0.343
AEOSI	HDP	-0.426	0.172	-2.477	6793.2	1.33E-02	-0.762	-0.089
AEOSI	age_diam	-0.042	0.020	-2.086	6804.9	3.70E-02	-0.081	-0.002
AEOSI	parity	0.036	0.020	1.811	6768.8	7.01E-02	-0.003	0.075
AEOSI	age_gest	0.031	0.023	1.384	6595.1	1.66E-01	-0.013	0.075
AEOSI	weight_gest	0.023	0.020	1.124	135.4	2.63E-01	-0.017	0.063
AEOSI	EUR	0.120	0.045	2.665	6789.9	7.72E-03	0.032	0.209
AEOSI	weeks:GDM	0.003	0.004	0.627	6805.2	5.31E-01	-0.006	0.011
AEOSI	weeks:PPH	-0.006	0.003	-2.174	6805.2	2.97E-02	-0.011	-0.001
AEOSI	weeks:HDP	0.010	0.004	2.183	6805.2	2.91E-02	0.001	0.018
ERY	(Intercept)	0.560	0.042	13.220	6803.5	0.00E+00	0.477	0.643
ERY	weeks	-0.010	0.001	-8.766	6803.5	0.00E+00	-0.012	-0.008
ERY	GDM	0.389	0.148	2.629	6792.1	8.59E-03	0.099	0.680
ERY	PPH	0.791	0.095	8.309	6805.0	0.00E+00	0.604	0.977
ERY	HDP	0.443	0.157	2.830	6800.7	4.67E-03	0.136	0.750
ERY	age_diam	-0.058	0.017	-3.483	6786.4	4.99E-04	-0.090	-0.025
ERY	parity	0.002	0.017	0.125	6675.6	9.01E-01	-0.030	0.035
ERY	age_gest	0.057	0.019	3.025	5538.3	2.50E-03	0.020	0.093
ERY	weight_gest	0.065	0.019	3.479	41.5	1.20E-03	0.027	0.102
ERY	EUR	0.064	0.038	1.698	6786.9	8.95E-02	-0.010	0.137
ERY	weeks:GDM	-0.010	0.004	-2.424	6805.0	1.54E-02	-0.018	-0.002
ERY	weeks:PPH	-0.043	0.002	-17.297	6804.5	0.00E+00	-0.048	-0.038
ERY	weeks:HDP	-0.021	0.004	-4.927	6805.2	8.54E-07	-0.029	-0.012
HT	(Intercept)	0.477	0.043	11.022	6805.2	0.00E+00	0.392	0.562
HT	weeks	-0.005	0.001	-4.190	6804.9	2.82E-05	-0.007	-0.003
HT	GDM	0.295	0.152	1.945	6797.4	5.18E-02	-0.002	0.592
HT	PPH	0.753	0.098	7.716	6805.2	1.38E-14	0.562	0.944
HT	HDP	0.572	0.161	3.556	6794.6	3.79E-04	0.257	0.888
HT	age_diam	0.036	0.016	2.209	6803.1	2.72E-02	0.004	0.068
HT	parity	-0.060	0.016	-3.740	6713.1	1.86E-04	-0.092	-0.029
HT	age_gest	0.085	0.018	4.667	6160.0	3.11E-06	0.049	0.121
HT	weight_gest	0.032	0.018	1.810	47.8	7.66E-02	-0.004	0.068
HT	EUR	-0.137	0.037	-3.743	6770.4	1.84E-04	-0.209	-0.065
HT	weeks:GDM	-0.013	0.004	-2.967	6805.2	3.02E-03	-0.021	-0.004
HT	weeks:PPH	-0.044	0.003	-17.144	6805.2	0.00E+00	-0.049	-0.039
HT	weeks:HDP	-0.026	0.004	-5.957	6805.2	2.70E-09	-0.035	-0.017
HB	(Intercept)	0.599	0.043	13.948	6805.2	0.00E+00	0.515	0.683
HB	weeks	-0.008	0.001	-7.132	6805.2	1.09E-12	-0.011	-0.006
HB	GDM	0.353	0.150	2.349	6800.4	1.88E-02	0.058	0.647
HB	PPH	0.674	0.097	6.981	6805.2	3.21E-12	0.485	0.863
HB	HDP	0.602	0.159	3.784	6789.0	1.56E-04	0.290	0.914
HB	age_diam	0.059	0.017	3.547	6805.2	3.93E-04	0.026	0.091
HB	parity	-0.086	0.017	-5.214	6740.6	1.90E-07	-0.119	-0.054

HB	age_gest	0.077	0.019	4.109	6400.8	4.02E-05	0.040	0.114
HB	weight_gest	0.011	0.018	0.640	62.0	5.25E-01	-0.024	0.047
HB	EUR	-0.208	0.038	-5.535	6743.4	3.23E-08	-0.282	-0.134
HB	weeks:GDM	-0.015	0.004	-3.581	6805.2	3.45E-04	-0.023	-0.007
HB	weeks:PPH	-0.041	0.003	-16.161	6805.2	0.00E+00	-0.046	-0.036
HB	weeks:HDP	-0.026	0.004	-6.171	6805.2	7.16E-10	-0.035	-0.018
LEUC	(Intercept)	-0.500	0.045	-11.071	6802.3	0.00E+00	-0.589	-0.412
LEUC	weeks	0.015	0.001	12.145	6805.2	0.00E+00	0.012	0.017
LEUC	GDM	0.541	0.158	3.423	6802.7	6.23E-04	0.231	0.850
LEUC	PPH	-0.500	0.102	-4.927	6805.2	8.56E-07	-0.699	-0.301
LEUC	HDP	0.081	0.167	0.483	6803.9	6.29E-01	-0.247	0.408
LEUC	age_diam	-0.053	0.018	-3.014	6800.7	2.59E-03	-0.088	-0.019
LEUC	parity	-0.113	0.018	-6.415	6797.0	1.51E-10	-0.148	-0.078
LEUC	age_gest	-0.212	0.020	-10.665	6704.8	0.00E+00	-0.251	-0.173
LEUC	weight_gest	0.039	0.017	2.273	740.1	2.33E-02	0.005	0.072
LEUC	EUR	-0.226	0.040	-5.637	6748.5	1.80E-08	-0.304	-0.147
LEUC	weeks:GDM	-0.014	0.004	-3.206	6805.2	1.35E-03	-0.022	-0.005
LEUC	weeks:PPH	0.026	0.003	9.640	6805.2	0.00E+00	0.020	0.031
LEUC	weeks:HDP	-0.004	0.004	-0.937	6805.2	3.49E-01	-0.013	0.005
ALYMPH	(Intercept)	0.213	0.046	4.604	6802.4	4.21E-06	0.122	0.303
ALYMPH	weeks	-0.005	0.001	-4.232	6804.8	2.34E-05	-0.008	-0.003
ALYMPH	GDM	0.370	0.161	2.296	6786.3	2.17E-02	0.054	0.686
ALYMPH	PPH	0.167	0.103	1.622	6805.0	1.05E-01	-0.035	0.370
ALYMPH	HDP	-0.061	0.170	-0.362	6786.2	7.17E-01	-0.394	0.271
ALYMPH	age_diam	-0.116	0.019	-5.954	6784.4	2.74E-09	-0.154	-0.078
ALYMPH	parity	-0.012	0.019	-0.638	6704.9	5.24E-01	-0.051	0.026
ALYMPH	age_gest	-0.123	0.022	-5.563	6292.4	2.76E-08	-0.166	-0.079
ALYMPH	weight_gest	0.076	0.020	3.740	85.9	3.31E-04	0.036	0.117
ALYMPH	EUR	-0.105	0.044	-2.369	6789.4	1.79E-02	-0.191	-0.018
ALYMPH	weeks:GDM	-0.011	0.004	-2.626	6805.2	8.66E-03	-0.020	-0.003
ALYMPH	weeks:PPH	-0.004	0.003	-1.621	6804.8	1.05E-01	-0.009	0.001
ALYMPH	weeks:HDP	0.001	0.004	0.161	6805.2	8.72E-01	-0.008	0.009
MCH	(Intercept)	0.130	0.034	3.880	6803.9	1.05E-04	0.064	0.196
MCH	weeks	0.001	0.001	1.782	6805.2	7.47E-02	0.000	0.003
MCH	GDM	-0.083	0.121	-0.689	6804.1	4.91E-01	-0.320	0.154
MCH	PPH	-0.213	0.075	-2.860	6800.3	4.25E-03	-0.359	-0.067
MCH	HDP	0.149	0.127	1.177	6750.8	2.39E-01	-0.099	0.398
MCH	age_diam	0.206	0.022	9.547	6777.4	0.00E+00	0.164	0.249
MCH	parity	-0.150	0.022	-6.946	6791.3	4.11E-12	-0.192	-0.107
MCH	age_gest	0.036	0.025	1.460	6543.6	1.44E-01	-0.012	0.085
MCH	weight_gest	-0.103	0.021	-4.869	351.1	1.70E-06	-0.144	-0.061
MCH	EUR	-0.451	0.049	-9.174	6638.9	0.00E+00	-0.547	-0.354
MCH	weeks:GDM	-0.008	0.002	-3.446	6805.2	5.72E-04	-0.013	-0.004
MCH	weeks:PPH	0.004	0.001	3.021	6805.2	2.53E-03	0.002	0.007
MCH	weeks:HDP	-0.006	0.002	-2.644	6805.2	8.20E-03	-0.011	-0.002
MCHC	(Intercept)	0.583	0.042	13.929	6804.3	0.00E+00	0.501	0.665
MCHC	weeks	-0.015	0.001	-14.137	6804.9	0.00E+00	-0.017	-0.013
MCHC	GDM	0.159	0.147	1.081	6802.0	2.80E-01	-0.129	0.446

MCHC	PPH	-0.192	0.093	-2.060	6804.4	3.95E-02	-0.374	-0.009
MCHC	HDP	0.140	0.154	0.908	6755.5	3.64E-01	-0.162	0.441
MCHC	age_diam	0.100	0.020	4.924	6787.7	8.69E-07	0.060	0.139
MCHC	parity	-0.118	0.020	-5.847	6801.5	5.23E-09	-0.157	-0.078
MCHC	age_gest	-0.018	0.023	-0.796	6564.8	4.26E-01	-0.063	0.027
MCHC	weight_gest	-0.085	0.019	-4.508	3760.8	6.73E-06	-0.123	-0.048
MCHC	EUR	-0.343	0.046	-7.490	6789.3	7.75E-14	-0.433	-0.254
MCHC	weeks:GDM	-0.008	0.004	-2.189	6805.2	2.87E-02	-0.015	-0.001
MCHC	weeks:PPH	0.006	0.002	2.695	6805.1	7.06E-03	0.002	0.011
MCHC	weeks:HDP	-0.003	0.004	-0.861	6805.2	3.89E-01	-0.011	0.004
MCV	(Intercept)	-0.131	0.034	-3.912	6804.8	9.26E-05	-0.197	-0.066
MCV	weeks	0.009	0.001	12.940	6805.2	0.00E+00	0.008	0.010
MCV	GDM	-0.146	0.121	-1.208	6802.8	2.27E-01	-0.383	0.091
MCV	PPH	-0.183	0.075	-2.451	6802.2	1.43E-02	-0.329	-0.037
MCV	HDP	0.143	0.127	1.128	6781.4	2.59E-01	-0.106	0.392
MCV	age_diam	0.204	0.022	9.370	6792.1	0.00E+00	0.161	0.247
MCV	parity	-0.125	0.022	-5.764	6790.3	8.58E-09	-0.168	-0.082
MCV	age_gest	0.054	0.025	2.176	6660.8	2.96E-02	0.005	0.103
MCV	weight_gest	-0.081	0.021	-3.795	280.0	1.81E-04	-0.123	-0.039
MCV	EUR	-0.389	0.049	-7.855	6582.6	4.66E-15	-0.486	-0.292
MCV	weeks:GDM	-0.007	0.002	-2.904	6805.2	3.70E-03	-0.012	-0.002
MCV	weeks:PPH	0.003	0.001	1.934	6805.2	5.32E-02	0.000	0.006
MCV	weeks:HDP	-0.007	0.002	-2.962	6805.2	3.07E-03	-0.012	-0.002
AMONO	(Intercept)	-0.553	0.046	-12.062	6804.7	0.00E+00	-0.643	-0.463
AMONO	weeks	0.017	0.001	13.878	6805.2	0.00E+00	0.015	0.019
AMONO	GDM	0.381	0.160	2.377	6804.2	1.75E-02	0.067	0.695
AMONO	PPH	-0.228	0.103	-2.223	6805.2	2.62E-02	-0.430	-0.027
AMONO	HDP	-0.017	0.169	-0.103	6802.0	9.18E-01	-0.349	0.314
AMONO	age_diam	-0.083	0.019	-4.416	6802.1	1.02E-05	-0.119	-0.046
AMONO	parity	-0.068	0.019	-3.664	6790.7	2.50E-04	-0.105	-0.032
AMONO	age_gest	-0.091	0.021	-4.315	6755.6	1.62E-05	-0.132	-0.050
AMONO	weight_gest	0.039	0.019	2.128	281.8	3.42E-02	0.003	0.076
AMONO	EUR	-0.137	0.042	-3.234	6752.6	1.23E-03	-0.220	-0.054
AMONO	weeks:GDM	-0.009	0.004	-2.174	6805.2	2.97E-02	-0.018	-0.001
AMONO	weeks:PPH	0.013	0.003	4.734	6805.2	2.24E-06	0.007	0.018
AMONO	weeks:HDP	-0.004	0.004	-0.903	6805.2	3.66E-01	-0.013	0.005
MPV	(Intercept)	-0.496	0.037	-13.443	6805.2	0.00E+00	-0.568	-0.424
MPV	weeks	0.015	0.001	18.394	6805.2	0.00E+00	0.014	0.017
MPV	GDM	0.055	0.131	0.417	6803.8	6.76E-01	-0.202	0.312
MPV	PPH	0.055	0.082	0.672	6805.2	5.01E-01	-0.105	0.215
MPV	HDP	0.383	0.137	2.789	6797.4	5.29E-03	0.114	0.652
MPV	age_diam	0.049	0.022	2.263	6805.2	2.37E-02	0.007	0.092
MPV	parity	-0.040	0.022	-1.828	6802.6	6.75E-02	-0.082	0.003
MPV	age_gest	0.016	0.025	0.631	6787.7	5.28E-01	-0.033	0.064
MPV	weight_gest	0.023	0.020	1.099	1880.0	2.72E-01	-0.018	0.063
MPV	EUR	-0.010	0.049	-0.200	6776.7	8.42E-01	-0.106	0.087
MPV	weeks:GDM	-0.002	0.003	-0.856	6805.2	3.92E-01	-0.008	0.003
MPV	weeks:PPH	-0.004	0.002	-2.287	6805.2	2.23E-02	-0.008	-0.001

MPV	weeks:HDP	-0.005	0.003	-1.679	6805.2	9.32E-02	-0.011	0.001
ANEUT	(Intercept)	-0.531	0.046	-11.542	6804.0	0.00E+00	-0.621	-0.441
ANEUT	weeks	0.015	0.001	12.216	6805.2	0.00E+00	0.013	0.018
ANEUT	GDM	0.475	0.161	2.947	6805.2	3.22E-03	0.159	0.791
ANEUT	PPH	-0.566	0.104	-5.452	6805.2	5.15E-08	-0.770	-0.363
ANEUT	HDP	0.107	0.171	0.622	6805.2	5.34E-01	-0.229	0.442
ANEUT	age_diam	-0.026	0.017	-1.530	6804.4	1.26E-01	-0.059	0.007
ANEUT	parity	-0.115	0.017	-6.823	6803.8	9.69E-12	-0.147	-0.082
ANEUT	age_gest	-0.195	0.019	-10.318	6777.3	0.00E+00	-0.232	-0.158
ANEUT	weight_gest	0.021	0.016	1.312	2426.7	1.90E-01	-0.010	0.052
ANEUT	EUR	-0.218	0.038	-5.708	6734.2	1.19E-08	-0.292	-0.143
ANEUT	weeks:GDM	-0.012	0.005	-2.652	6805.2	8.01E-03	-0.021	-0.003
ANEUT	weeks:PPH	0.028	0.003	10.123	6805.2	0.00E+00	0.023	0.033
ANEUT	weeks:HDP	-0.004	0.005	-0.896	6805.2	3.70E-01	-0.013	0.005
PLAQ	(Intercept)	0.482	0.039	12.422	6805.2	0.00E+00	0.406	0.558
PLAQ	weeks	-0.012	0.001	-12.343	6805.2	0.00E+00	-0.014	-0.010
PLAQ	GDM	0.287	0.136	2.107	6805.2	3.52E-02	0.020	0.554
PLAQ	PPH	0.176	0.086	2.041	6805.2	4.13E-02	0.007	0.345
PLAQ	HDP	-0.419	0.142	-2.942	6805.2	3.27E-03	-0.699	-0.140
PLAQ	age_diam	0.003	0.019	0.160	6801.1	8.73E-01	-0.035	0.041
PLAQ	parity	-0.070	0.019	-3.627	6799.4	2.89E-04	-0.108	-0.032
PLAQ	age_gest	0.009	0.022	0.387	6765.6	6.98E-01	-0.035	0.052
PLAQ	weight_gest	0.046	0.019	2.492	756.5	1.29E-02	0.010	0.083
PLAQ	EUR	0.018	0.044	0.410	6732.7	6.82E-01	-0.068	0.104
PLAQ	weeks:GDM	-0.007	0.003	-2.010	6805.2	4.45E-02	-0.013	0.000
PLAQ	weeks:PPH	-0.010	0.002	-4.989	6805.2	6.23E-07	-0.014	-0.006
PLAQ	weeks:HDP	0.003	0.003	0.733	6805.2	4.63E-01	-0.004	0.009
RDW	(Intercept)	-0.896	0.041	-22.051	6805.2	0.00E+00	-0.975	-0.816
RDW	weeks	0.021	0.001	20.396	6805.2	0.00E+00	0.019	0.023
RDW	GDM	0.026	0.142	0.181	6805.2	8.56E-01	-0.253	0.304
RDW	PPH	0.077	0.090	0.852	6805.1	3.94E-01	-0.100	0.254
RDW	HDP	-0.260	0.149	-1.742	6792.7	8.15E-02	-0.552	0.033
RDW	age_diam	-0.087	0.019	-4.563	6799.7	5.14E-06	-0.124	-0.049
RDW	parity	0.160	0.019	8.468	6804.5	0.00E+00	0.123	0.197
RDW	age_gest	0.034	0.022	1.596	6751.2	1.11E-01	-0.008	0.077
RDW	weight_gest	0.057	0.018	3.146	1099.6	1.70E-03	0.021	0.093
RDW	EUR	0.379	0.043	8.794	6742.8	0.00E+00	0.294	0.463
RDW	weeks:GDM	0.008	0.004	2.349	6805.2	1.89E-02	0.001	0.016
RDW	weeks:PPH	0.005	0.002	2.297	6805.2	2.16E-02	0.001	0.009
RDW	weeks:HDP	0.010	0.004	2.827	6805.2	4.72E-03	0.003	0.018

Table S7. Cox proportional hazards model summary.

term	estimate	std.error	statistic	df	p.value	2.5%	97.5%	HR	HR2.5%	HR97.5%
parity	0.13756	0.0369	3.7278	852	0.00019	0.065	0.21	1.147	1.0674	1.23354
weight	-0.01318	0.0029	-4.5484	211	5.4E-06	-0.019	-0.008	0.987	0.9813	0.99253
HB	-0.01084	0.00342	-3.1695	845	0.00153	-0.018	-0.004	0.989	0.9826	0.99587
RDW	-0.10871	0.0268	-4.0565	849	5E-05	-0.161	-0.056	0.897	0.8511	0.94536
LEUC	-0.5835	0.13962	-4.1793	854	2.9E-05	-0.857	-0.31	0.558	0.4244	0.73356
ALYMPH	0.72169	0.14666	4.92065	855	8.6E-07	0.434	1.009	2.058	1.5438	2.74327
ANEUT	1.07747	0.17922	6.01208	851	1.8E-09	0.726	1.429	2.937	2.0672	4.1734
PLAQ	-0.00265	0.00063	-4.1686	845	3.1E-05	-0.004	-0.001	0.997	0.9961	0.9986
GDM	0.29959	0.16073	1.86398	854	0.06232	-0.015	0.615	1.349	0.9847	1.84894
HDP	0.91148	0.12482	7.30261	821	2.8E-13	0.667	1.156	2.488	1.9481	3.17758