Sparsity and information borrowing in Bayesian hierarchical models for genetic association problems

Hélène Ruffieux
MRC Biostatistics Unit, University of Cambridge

CUSO Summer School
7–10 September 2025

“In theory there is no difference between theory and practice, while in practice there is.”

—— Yale student Benjamin Brewster, class of 1882

or its variant:

“The difference between practice and theory is greater in practice than in theory.”

This slide set (and any updated content) is available at https://tinyurl.com/cuso-summer-2025-bhm

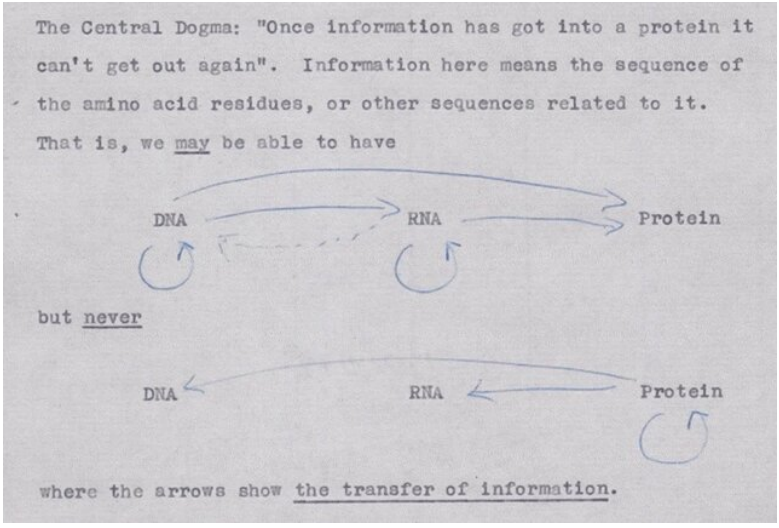
Outline

Cambridge region
Europe's largest biotechnology cluster

Image: Cambridge University Health Partners.

Introduction

The Central Dogma of molecular biology (Crick, 1956)

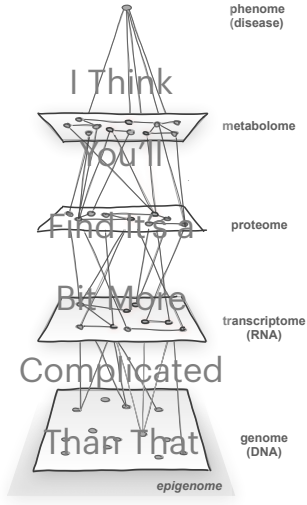


Crick's first outline of the Central Dogma, from an unpublished note made in 1956. Image: Wellcome Library, London.

Molecular pathways from genotype to phenotype

Today, we know that biology involves more than sequence-level information flow.

- **DNA carries** the **instructions** that shape how our bodies function and how diseases may develop.
- These instructions are **read through a complex network of molecular processes** inside each cell.
- This network operates **across multiple layers** of biological activity – including genes, proteins and biochemical entities – which **together influence physical traits and disease risk**.
- The relationships between layers are subtle and depend on context – for example, they **can vary across cell types, tissues or developmental stages**.
- External factors such as **lifestyle or environmental exposures can also influence** how these molecular processes unfold.



Ben Goldacre

Guest-Editorial

Talkin' Omics

The 1990s will be remembered as the decade when advances in biomedical research launched the genomics era. While new information and technologies are clearly important products of the genomics revolution, perhaps most important is a change in mindset of how we pursue scientific discovery. We are no longer satisfied to study a gene or gene product in isolation, but rather we strive to view each gene within the complex circuitry of a cell. Understanding how genes and their products interact will open many exciting avenues in biological and biomedical research. In rapid succession, this new mindset has invigorated the analysis of all molecular entities, from the genome, to transcripts (transcriptome) and proteins (proteome). And it is clear that this is just the beginning of the omics revolution.

While the understanding and treatment of many diseases will be impacted by omics, arguably the greatest biomedical opportunity for discovery is cancer. As a family of diseases, all cancer results from changes in the genome. The genomic changes take many forms, from point mutations, to amplifications and deletions, of disease that might be most amenable to intervention.

The complexity of molecular events within the genome are reflected and amplified by the diversity of transcripts and proteins within a cell. A variety of transcripts can be derived from the same gene, and the encoded proteins can be modified extensively to fulfill specific biological functions within a cell. The omics revolution has challenged researchers to integrate the study of the genome, transcriptome, and proteome, for this is the most promising approach to attaining a comprehensive omic view of the molecular circuitry within a cell.

In this issue of *Disease Markers*, we are fortunate to have contributions from some of the leaders of the omics approach. While many of the articles feature cancer research, we hope that the more general applicability of the described approaches is apparent. We have tried to make this inaugural omics special issue of *Disease Markers* provocative and informative, and we hope that it captures the excitement that has led to the description of omics research as revolutionary.

Robert L. Strausberg, *Disease Markers*, 2001.

From isolated genes to systems biology

Systems biology:

- In the early 2000s, advances such as the **first sequencing of the human genome** (14 April 2003) and the development of **high-throughput technologies** (e.g., microarrays, next-generation sequencing, mass spectrometry) transformed molecular biology.
- The concept of “systems biology” emerged, promoting a **holistic view of biological function**.
- Biological questions shifted from “Which gene is involved?” to “How do genes interact in networks and pathways?”
- These changes opened new opportunities for understanding disease mechanisms and therapeutic targeting.

Large and complex datasets:

- High-throughput techniques generate **multi-layered molecular data** across tissues, individuals and time, shaped by structured, context-specific biological processes.
- Historically limited in genomics by computational constraints, **Bayesian methods have become increasingly tractable and impactful** over the past 20 years.

Why Bayesian modelling in genomics?

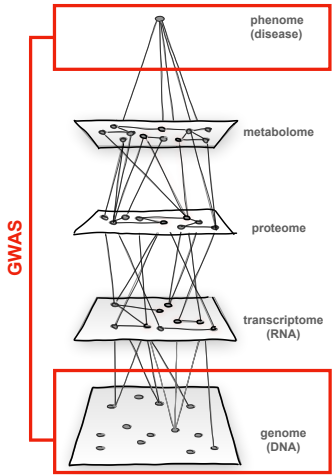
In the context of genomics studies, **Bayesian hierarchical models offer:**

- principled **quantification of uncertainty**, critical for decision-making and interpretation;
- flexible **modelling of complex dependencies**, with information borrowing (when supported by the data);
- **integration of domain knowledge** via custom priors: external (e.g., pathways, annotations, networks) or structural (e.g., sparsity, smoothness, modularity);
- natural **handling of multi-level data** (e.g., samples nested within individuals);
- a **generative view of the genotype-to-phenotype map**, aligned with the systems biology perspective.

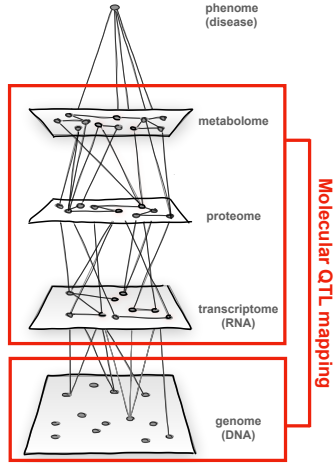
Genetic association studies

Two types of genetic association studies

Part I

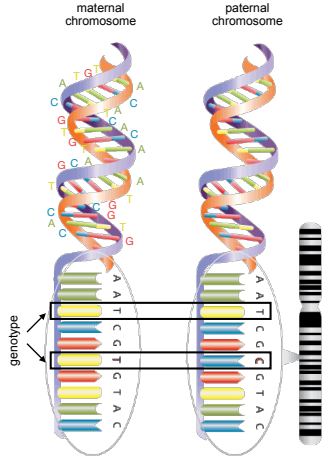


Part II



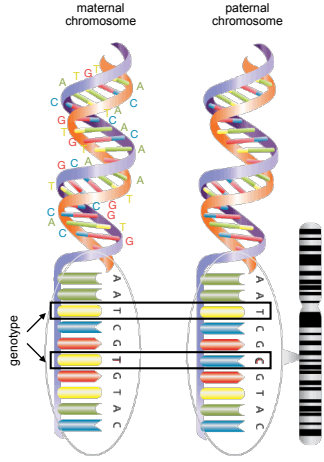
DNA and genotypes

- **DNA** (deoxyribonucleic acid) consists of two complementary strands forming a double helix. Each strand is made up of nucleotides: **A (adenine), T (thymine), C (cytosine) and G (guanine)**.
- Nucleotides pair specifically across the two strands: A pairs with T, and C pairs with G. These pairings form the **base pairs** that encode genetic information.
- Human DNA is organised into **23 pairs of chromosomes** (22 autosomes + 1 sex chromosome), with one set inherited from each parent.
- An individual's **genotype** at a given position (**"locus"**) refers to the combination of nucleotides inherited from both parents.
- Specific regions of DNA called **genes** contain instructions for producing proteins, but much of the genome is non-coding and still functionally important.



SNPs, alleles and minor allele frequency

- A **genetic variant** is a change at a specific location in the DNA sequence across individuals in a population. It often involves a change in a single nucleotide but can also include insertions, deletions or structural alterations.
- An **allele** refers to one of the possible nucleotides observed at a given genomic position where variation occurs across individuals.
- The **minor allele frequency (MAF)** is the frequency at which the less common allele occurs in a population.
- A **single nucleotide polymorphism (SNP)** is a single base substitution (e.g., C$\rightarrow$T) that is present in at least 1% of the population (i.e., MAF > 0.01; typical threshold to distinguish common SNPs from rare variants).
- **Example:** Suppose in a population, the genotypes at a SNP are distributed as:
  - 60% CC, 30% CT, 10% TT $\Rightarrow$ allele C has frequency 0.75, T has frequency 0.25.
  - Here, **C is the major allele**, and **T is the minor allele** with MAF = 0.25.



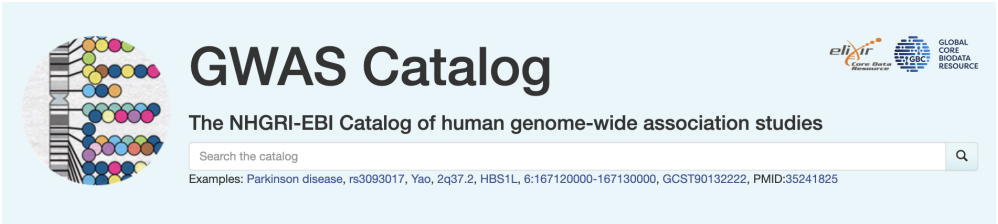
Genome-wide association studies (GWAS)

- **Genome-wide association studies (GWAS)** aim to identify statistical associations between **genetic variants (usually SNPs)** and a **trait** of interest (or *phenotype*), such as height, disease status or a clinical measurement.
- In a GWAS, hundreds of thousands to millions of SNPs across the genome are tested for association with the phenotype, typically in large populations.
- To model the effect of a SNP (say with minor allele T) on the phenotype, different **genetic models** can be used:
  - The **additive model** (most used) assumes a linear and cumulative effect of the number of minor alleles. Genotypes are coded as:
    0: homozygous major (CC) 1: heterozygous (CT or TC) 2: homozygous minor (TT);
  - The **dominant model** assumes the presence of at least one minor allele confers the full effect. Genotypes are coded as:
    0: homozygous major (CC) 1: heterozygous or homozygous minor (CT, TC, or TT);
  - The **recessive model** assumes an effect only when two copies of the minor allele are present. Genotypes are coded as:
    0: homozygous major or heterozygous (CC, CT or TC) 1: homozygous minor (TT).

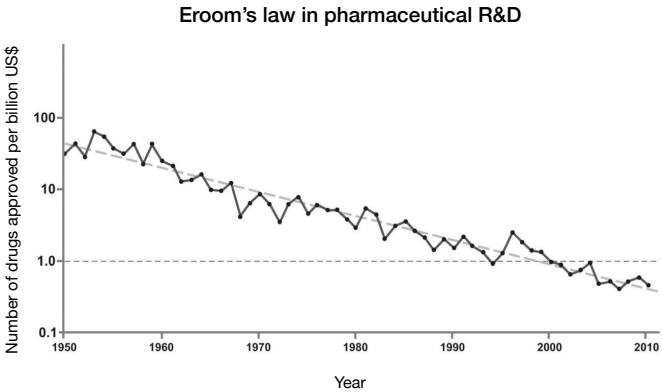
Why study genetic associations?

- Many important health outcomes – such as body mass index (BMI), blood pressure or type-2 diabetes – are **complex traits**, meaning that they are **influenced by multiple genetic variants** together with environmental factors, **rather than variation at a single gene.**
- With the sequencing of the human genome, GWAS have started to reveal the genetic architecture of a tapestry of complex traits through identification of genetic variants associated with such traits across diverse populations.

Last data release 2025-08-24 (`https://www.ebi.ac.uk/gwas/search`):
7 369 publications, 106 320 traits (EFO) and 955 930 top associations.



The costly decline of drug discovery efficiency...



Logarithmic decline in the number of new drugs approved per US$ billion in R&D spending from 1950 to 2010. Adapted from Scannell et al. (2012).

GWAS support drug discovery

GWAS have advanced our understanding of human biology and disease. In some cases, they have also **contributed directly to drug discovery** and therapeutic development.

- *>90% of drug candidates fail in clinical trials (Sun et al., 2022);*
- *Drug targets with genetic support are* **2.6× more likely to reach approval** *(Minikel et al., 2024);*
- *Genetic support underlies 8.2% of all approved drugs (Nelson et al., 2015), . . . but 66% of those approved in 2021 (Ochoa et al., 2022).*

News & analysis

Biobusiness briefs

https://doi.org/10.1038/d41573-023-00158-x

Genetic support for FDA-approved drugs over the past decade

Human genetic evidence has been reported as a proxy for the success of drugs in clinical trials owing to its ability to pinpoint causal mechanistic links between drug targets and diseases. Two-thirds of the new drugs approved by the FDA in 2021 were supported by human genetic evidence associating the intended pharmacological target or its physical interactors with the indication or a surrogate trait. Here, we expand this retrospective analysis

REGULATORY WATCH

Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs

In 2021, 50 drugs were approved by the FDA's Center for Drug Evaluation and Research, continuing a spell of improved productivity.

often requires the integration of mu layers of genetic and functional gen combined with clinical knowledge a

Analysis

Refining the impact of genetic evidence on clinical success

https://doi.org/10.1038/s41586-024-07316-0

Received: 5 July 2023

Accepted: 14 March 2024

Published online: 17 April 2024

Open access

Check for updates

Eric Vallabh Minikel[1], Jeffery L. Painter[2,3], Coco Chengliang Dong[2] & Matthew R. Nelson[2,4]

The cost of drug discovery and development is driven primarily by failure[1], with only about 10% of clinical programmes eventually receiving approval[2–5]. We previously estimated that human genetic evidence doubles the success rate from clinical development to approval[2]. In this study we leverage the growth in genetic evidence over the past decade to better understand the characteristics that distinguish clinical success and failure. We estimate the probability of success for drug mechanisms with genetic support is 2.6 times greater than those without. This relative success varies among therapy areas and development phases, and improves with increasing

From genetic associations to biomedical translation

***PCSK9* and cholesterol-lowering therapy** (Chaudhary et al., 2017):

- GWAS identified variants in the ***PCSK9* gene associated with LDL cholesterol** levels.
- Functional studies confirmed *PCSK9*'s role in lipid metabolism.
- This **led to the development of *PCSK9* inhibitors**, such as alirocumab and evolocumab, now approved to treat high cholesterol and reduce cardiovascular events.

***FTO* and obesity** (Loos and Yeo, 2014):

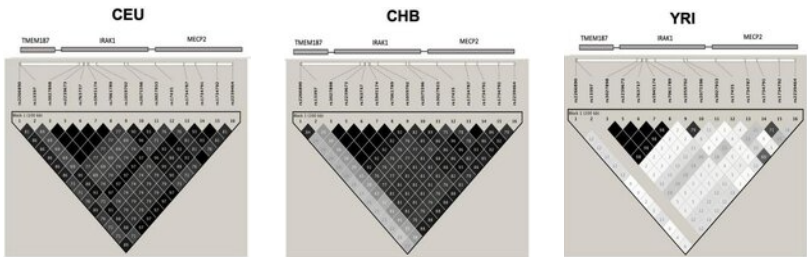
- A landmark **GWAS linked common variants in the *FTO* gene to higher BMI** and obesity risk.
- These variants are frequent in the European population (MAF > 0.4), with substantial public health relevance.
- Follow-up work revealed **a role in appetite regulation**, uncovering new pathways for potential intervention.

Should we consider SNP associations as causal in GWAS?

- Causal interpretations in observational studies are rarely justified without explicit causal modelling.
- In GWAS settings:
  - **Genotypes precede phenotypes temporally**, so reverse causation is implausible.
  - Genotypes **are less susceptible to** many common **confounders**.
  - Therefore, **an association is often interpreted as a causal effect of the SNP** (or a linked SNP) on the phenotype.
- Well-documented exceptions are confounding due to **population structure** (typically corrected using principal components or mixed-model approaches) or **indirect environmental effects** (typically avoided using family-based designs):
  - Regional ancestry differences can confound genetic analysis, creating patterns that resemble genetic effects but are actually due to cultural variation (e.g., cheese type preference).
  - Parental DNA can shape both the child's environment (e.g., TV watching habits) and the child's own genotype.
- This causal interpretation underpins methods like **Mendelian randomisation**, which use genetic variants as instruments to infer causal effects of exposures on outcomes under assumptions of relevance, independence and exclusion restriction (see, e.g., Davey Smith and Ebrahim, 2003; Burgess et al., 2015).

Linkage disequilibrium

- **Linkage disequilibrium (LD)** refers to the non-random association of alleles at different loci.
- It arises because **nearby genetic variants tend to be inherited together**.
- LD is crucial in association studies, as **nearby SNPs often serve as proxies for a causal variant**.
- **Population-specific LD patterns**: e.g., shorter blocks in YRI vs. CEU/CHB, due to greater genetic diversity and older population history in African ancestries.



LD plots of the SNPs in the Xq28 region. CEU: Utah Residents with Northern and Western European ancestry, CHB: Han Chinese in Beijing, YRI: Yoruba in Ibadan, Nigeria (HapMap project).

Univariate screening and the "omitted variable misspecification"

- Most GWAS for continuous traits rely on a **series of marginal regressions**: for n i.i.d. samples and an $n \times 1$ centred response y,

$$\boldsymbol{y} = \boldsymbol{X}_s \beta_s + \boldsymbol{\varepsilon}, \quad s = 1, \ldots, p,$$

  where $\boldsymbol{X}_s$ is an $n \times 1$ centred predictor, β_s is its regression coefficient and $\boldsymbol{\varepsilon}$ is an $n \times 1$ Gaussian error term.

- Suppose the **correct model** is:

$$\boldsymbol{y} = \boldsymbol{X}_1 \beta_1 + \boldsymbol{X}_2 \beta_2 + \boldsymbol{\varepsilon}, \quad \beta_1, \beta_2 \neq 0.$$

- Then, for any SNP $s \in \{1, \ldots, p\}$:

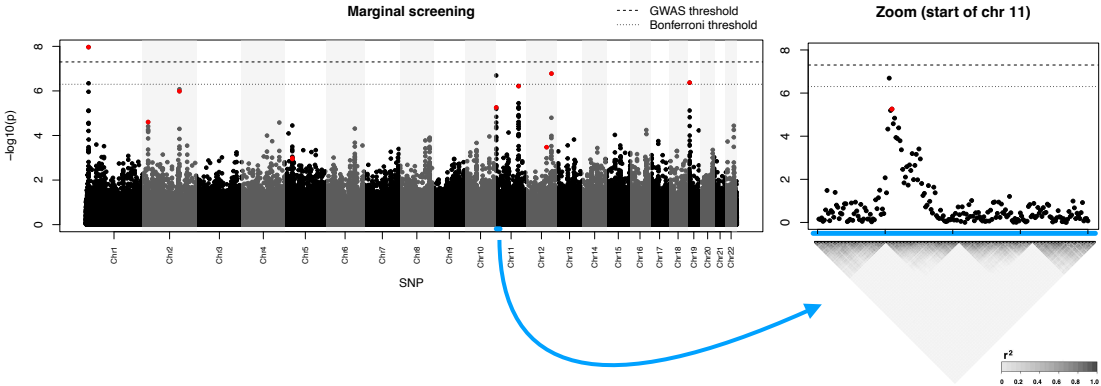
$$\mathbb{E}(\hat{\beta}_s) = (\boldsymbol{X}_s^T \boldsymbol{X}_s)^{-1} \boldsymbol{X}_s^T (\boldsymbol{X}_1 \beta_1 + \boldsymbol{X}_2 \beta_2),$$

  which shows that the **marginal estimator is biased** if $\boldsymbol{X}_s$ is correlated (i.e., "in LD") with $\boldsymbol{X}_1$ or $\boldsymbol{X}_2$.

This bias may:

- lead to spurious associations – for SNPs $s \notin \{1, 2\}$ in LD with $\boldsymbol{X}_1$ or $\boldsymbol{X}_2$.
- hinder the identification of truly associated SNPs – e.g., for SNP $s = 1$, if in LD with $\boldsymbol{X}_2$, the bias may shrink $\hat{\beta}_1$ towards zero.
- distort the estimated effect sizes – e.g., even if $s = 1$ is correctly identified as associated, $\hat{\beta}_1$ may be biased.

Manhattan plot and regional plot



Example of GWAS using simulated data; true non-zero effects are in red. Left: genome-wide (Manhattan) plot, right: regional plot. Most strong marginal associations are solely due to local correlation among predictors.

Frequentist and Bayesian regularisation

High-dimensional regression

Consider the **linear regression model**

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma_{\varepsilon}^2 \boldsymbol{I}_n), \tag{1}$$

where:

- $\boldsymbol{y} \in \mathbb{R}^n$ is a response vector;
- $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is a fixed design matrix of predictors;
- $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of unknown regression coefficients;
- $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a vector of i.i.d. Gaussian errors with mean zero and variance σ_{ε}^2.

- For simplicity, we assume that the **response and predictors are centred**, and set the intercept to zero (assumed throughout this course);
- In GWAS, **scaling the predictors is not necessary** as SNPs are measured on the same scale (allele counts); scaling implies priors favouring larger effects for rare variants, in line with natural selection (Park et al., 2011);
- We focus on the **high-dimensional regime** $p \gg n$.

Main objectives: (i) identify relevant predictors (variable selection) and (ii) estimate their effects.

Sparsity in GWAS

- High-dimensional regression: $\boldsymbol{X}^T\boldsymbol{X}$ is singular when $p > n$;
- Standard estimators are ill-posed without regularisation or structural assumptions.

Sparsity: assume that most entries of β are zero or negligible; this encourages parsimony and enables stable, identifiable inference.

- For complex traits, strong-effect SNPs are the exception rather than the rule.
- Recent theories like the omnigenic model (Boyle et al., 2017) suggest a pervasive polygenic architecture: a small subset of “core genes” have large effects, yet many SNPs, in nearly all expressed genes, contribute weak effects.

$\longrightarrow$ this tension motivates **flexible approaches that capture both a few strong and many weak signals**.

- GWAS emphasises selection over prediction, yet many true effects may be too small to detect confidently;
- This places value on interpretable, variable-level measure of confidence (posterior probabilities) – which we argue is a strength of Bayesian variable selection.

Bayesian model selection formalism

Model selection:

- Model space: 2^p possible models, each defined by an inclusion vector $\boldsymbol{\gamma} \in \{0,1\}^p$.
- For each model $\mathcal{M}_\gamma$, the data distribution is:

$$y_i \mid \mathcal{M}_\gamma, \boldsymbol{\beta}_\gamma, \sigma_\varepsilon^2 \sim \mathcal{N}(\boldsymbol{x}_i^\gamma \boldsymbol{\beta}_\gamma, \sigma_\varepsilon^2), \quad i = 1, \ldots, n,$$

  where $\boldsymbol{x}_i^\gamma$ is the p_γ-vector of predictors included in model $\mathcal{M}_\gamma$, and $\boldsymbol{\beta}_\gamma$ the corresponding vector of regression coefficients.

- Given a prior $p(\boldsymbol{\beta}_\gamma, \sigma_\varepsilon^2)$, the posterior probability of model $\mathcal{M}_\gamma$ is

$$p(\mathcal{M}_\gamma \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \mathcal{M}_\gamma) p(\mathcal{M}_\gamma)}{\sum_\gamma p(\boldsymbol{y} \mid \mathcal{M}_\gamma) p(\mathcal{M}_\gamma)},$$

  where

$$p(\boldsymbol{y} \mid \mathcal{M}_\gamma) = \int p(\boldsymbol{y} \mid \mathcal{M}_\gamma, \boldsymbol{\beta}_\gamma, \sigma_\varepsilon^2) p(\boldsymbol{\beta}_\gamma, \sigma_\varepsilon^2) \, d\boldsymbol{\beta}_\gamma \, d\sigma_\varepsilon^2.$$

Bayesian model averaging (BMA)

- Posterior distribution of a quantity of interest Δ:

$$p(\Delta \mid \boldsymbol{y}) = \sum_{\gamma} p(\Delta \mid \mathcal{M}_\gamma, \boldsymbol{y})\, p(\mathcal{M}_\gamma \mid \boldsymbol{y});$$

- Predictive distribution for new observation $\tilde{\boldsymbol{y}}$:

$$p(\tilde{\boldsymbol{y}} \mid \boldsymbol{y}) = \sum_{\gamma} p(\tilde{\boldsymbol{y}} \mid \mathcal{M}_\gamma, \boldsymbol{y})\, p(\mathcal{M}_\gamma \mid \boldsymbol{y}).$$

Challenges:

- Specification of priors over models and parameters;
- Computation of marginal likelihoods;
- Exploration of model space of size 2^p. When $p \gg n$ and in presence of small effects, the posterior probability on any single model will be very small, making identification of a single best model extremely difficult.

Frequentist approaches to regularisation

Frequentist high-dimensional regression is typically framed as the minimisation of

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \text{pen}_\lambda(\boldsymbol{\beta}), \tag{2}$$

where $\text{pen}_\lambda(\boldsymbol{\beta})$ is a penalty function indexed by $\lambda > 0$, a tuning parameter controlling the strength of the penalty.

- ***Separable*** **penalty functions** are typically employed, i.e., $\text{pen}_\lambda(\boldsymbol{\beta}) = \sum_{s=1}^{p} \rho_\lambda(\beta_s)$, e.g., $\rho_\lambda(\beta_s) = \lambda|\beta_s|$ for the lasso (Tibshirani, 1996).
- Optimal estimation properties may be obtained by imposing sparsity conditions like

$$\|\boldsymbol{\beta}\|_q^q \log p \ll n, \quad 0 \leq q < \infty,$$

  where $\|\cdot\|_q$ is the ℓ_q norm and $\|\boldsymbol{\beta}\|_0^0 = \|\boldsymbol{\beta}\|_0 = \#\{1 \leq s \leq p : \beta_s \neq 0\}$ (Bühlmann and van de Geer, 2011).
- In the $p \gg n$ regime, asymptotic normality no longer holds in general and bootstrapping becomes challenging.

Bayesian regularisation via priors

Bayesian regularisation replaces penalty functions with priors:

- Many **penalised estimators** are equivalent to the **mode of a posterior distribution** with prior

$$p(\beta_s) \propto \exp\{-\rho_\lambda(\beta_s)\}, \quad s = 1, \ldots, p. \tag{3}$$

- Example: the lasso is equivalent to a maximum a posteriori (MAP) estimator with *independent* Laplace priors on the entries of $\boldsymbol{\beta}$:

$$p(\beta_s) = \frac{\lambda}{2} \exp\{-\lambda|\beta_s|\}, \quad s = 1, \ldots, p.$$

  This prior, known as the Bayesian lasso, was first introduced by Park and Casella (2008).

- Instead of just the mode, **Bayesian inference provides the full posterior distribution** and thereby quantifies uncertainty.

- Opportunity to treat λ as a parameter to be estimated. The fully Bayes approach of **placing a prior on λ renders the penalty *non-separable***, which allows sharing information across different coordinates.

Scale mixtures of normals

- A wide range of sparsity priors can be expressed as **scale mixtures of normal densities**:

$$p(\beta_s) = \int \mathcal{N}\Big(\beta_s \mid 0, \omega_s\Big)\, dG(\omega_s), \quad s = 1, \ldots, p, \tag{4}$$

  where G is a distribution on the variance parameter.

- This representation **unifies many popular shrinkage priors** (e.g., Laplace, Student-t, Horseshoe) and simplifies:
  - **Computation**: It induces *conditional conjugacy* for each coefficient, which simplifies posterior sampling. For example, the Laplace prior can be written as:

$$\beta_s \mid \omega_s \sim \mathcal{N}(0, \omega_s), \quad \omega_s \sim \text{Exp}(\lambda^2/2), \quad s = 1, \ldots, p.$$

    This turns a non-conjugate shrinkage prior into a conditionally conjugate model for β_s.
  - **Theory**: The mixing distribution G clarifies how the prior concentrates mass near zero (for shrinkage) and in the tails (for robustness). This simplifies the analysis of adaptivity to sparsity and posterior behaviour.

Posterior consistency and asymptotics

- **Posterior consistency:** Under the assumption that the true parameter $\boldsymbol{\beta}_0$ is in the support of the prior, the posterior converges to a Dirac measure at $\boldsymbol{\beta}_0$ (see Doob, 1949; Schwartz, 1965).
- **Concentration rates:** Establishing the rate at which the posterior contracts provides insight into the required sample size n for a desired accuracy. These rates can be compared to minimax risk bounds to assess adaptivity.
- **Bernstein–von Mises theorem:** In low-dimensional settings, the posterior approximates a Gaussian distribution centred at the maximum likelihood estimator. However, in the $p \gg n$ regime, the required conditions (e.g., flat priors around $\boldsymbol{\beta}_0$) may fail, making standard asymptotic normality results inapplicable.

> *"For a Bayesian, the problem with the 'bias' concept is that is conditional on the true parameter value. But you don't know the true parameter value. There's no particular virtue in unbiasedness."*
>
> —— *Andrew Gelman*

Normal means model

To ease the presentation and unless stated otherwise, we describe these priors in the context of the **normal means problem** (Stein, 1981),

$$y_i = \beta_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}\left(0, 1\right), \qquad i = 1, \ldots, n, \tag{5}$$

where the dimensionality n is large.

Theory for model (5) is frequently examined under a *nearly-black* sparsity assumption, i.e., assuming that the unknown true parameter $\boldsymbol{\beta}_0$ belongs to (Donoho et al., 1992; Johnstone, 1994)

$$l_0[p_n; n] = \left\{\boldsymbol{\beta} \in \mathbb{R}^n : \|\boldsymbol{\beta}\|_0 \leq p_n\right\}, \quad p_n = o(n), \quad n \to \infty. \tag{6}$$

Two-group priors

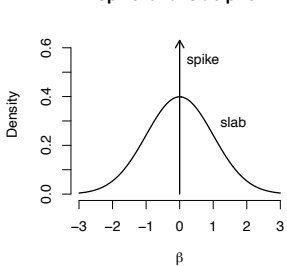
Two-group priors

Formulation (spike-and-slab prior; Mitchell and Beauchamp, 1988; George and McCulloch, 1993):
Split the prior into two components:

$$\beta_i \mid \pi \sim \pi\, g_\beta + (1-\pi)\,\delta_0, \quad i = 1, \ldots, n, \tag{7}$$

where:

- g_β is an absolutely continuous density (often taken as a centred normal with variance σ^2_β);
- δ_0 is the Dirac distribution at zero;
- π is the prior probability that β_i is nonzero.



Spike-and-slab prior (with Gaussian slab).

Interpretation:
The ***spike*** at zero models **noise**; the ***slab*** models **signal.** This formulation allows **separate modelling of the magnitude of effects and the sparsity level.**

Posterior shrinkage under the spike-and-slab prior

It is possible to express model selection priors as shrinkage priors.

Lemma (Adapted from Bhadra et al., 2017)

Assume the normal means model (5) and prior (7) with $g_\beta = \mathcal{N}(0, \sigma_\beta^2)$ for β_i. Then the posterior mean of β_i can be expressed as

$$\mathbb{E}(\beta_i \mid y_i) = \pi(y_i) \frac{\sigma_\beta^2}{1 + \sigma_\beta^2} y_i, \tag{8}$$

where $\pi(y_i) = p(\beta_i \neq 0 \mid y_i)$. As $\sigma_\beta^2 \to \infty$,

$$\mathbb{E}(\beta_i \mid y_i) = \big(1 + o(1)\big)\, \pi(y_i) y_i.$$

- Global shrinkage is controlled by σ_β^2.
- The term $\pi(y_i)$ provides adaptive, local shrinkage.

Hierarchical representation of the spike-and-slab prior

A reparameterisation introduces binary latent indicators γ_i:

$$\begin{aligned} \beta_i \mid \gamma_i &\sim \gamma_i \, g_\beta + (1 - \gamma_i) \, \delta_0, \\ \gamma_i \mid \pi &\sim \text{Bernoulli}(\pi), \quad i = 1, \ldots, n. \end{aligned} \tag{9}$$

Posterior interpretation for variable selection:

- For regression models (1), with spike-and-slab prior on the regression coefficients (replacing the index *i* with the predictor index *s*), the ***marginal posterior probability of inclusion (PPI)***,

  $$\mathbb{E}(\gamma_s \mid \boldsymbol{y}) = p(\gamma_s = 1 \mid \boldsymbol{y}),$$

  directly quantifies the evidence for including predictor *s* in the model.
- This hierarchical form facilitates both variable selection and uncertainty quantification.
- We use marginal summaries of variable inclusion – indeed, remember that we are not attempting to identify a single "best" model $\mathcal{M}_\gamma$ (i.e., combination of predictors) or estimate posterior probabilities for specific models.

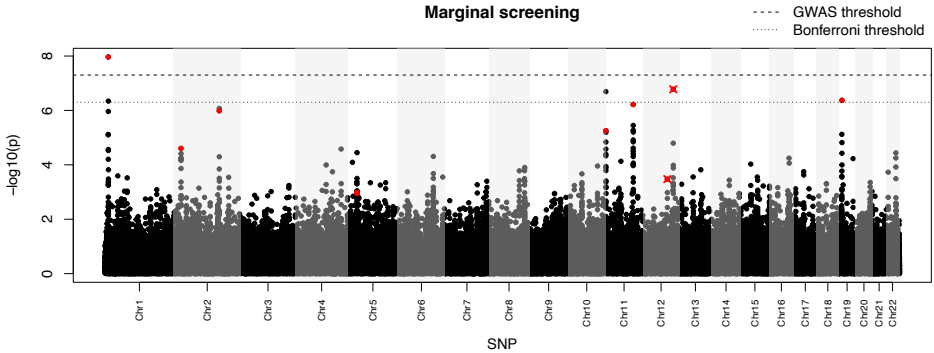
Example GWAS for systolic blood pressure (SBP)

Data simulation settings:

- 22 chromosomes with lengths reflecting their relative proportions in the human genome (GRCh38);
- 100 000 SNPs with minor allele frequencies (MAF) uniformly drawn between 0.05 and 0.5, for $n = 500$ individuals;
- Simulated correlation patterns (linkage disequilibrium, LD) using block-wise autocorrelated structures with realistic block counts per chromosome;
- Selected 10 risk SNPs among the top associations from a large systolic blood pressure GWAS[1];
- Assigned effect sizes such that the total proportion of variance explained (PVE, i.e., narrow-sense heritability, see later) equals 30%;
- Per-SNP PVE sampled across chromosomes from a half-Cauchy distribution;
- Simulated the phenotype as a linear combination of the risk SNPs, with additive noise reflecting the specified PVE.

[1] **Reference GWAS:** *Genome-wide analysis in over 1 million individuals of European ancestry yields improved polygenic risk scores for blood pressure traits* (Keaton et al., 2024).

Marginal GWAS

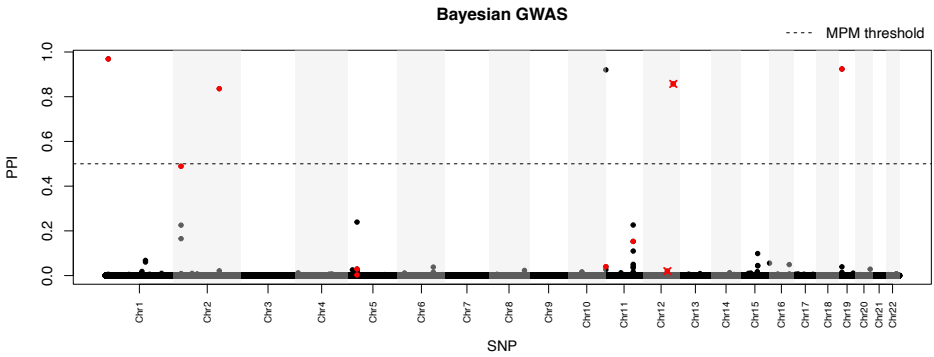


Manhattan plot, using frequentist marginal screening. Red: simulated signals.

SNPs marked with a red cross, on chromosome 12:

- rs7137828: regulator of cytokine signalling, known associations with blood pressure, hypertension and autoimmunity;
- rs11105354: regulator of calcium levels in blood vessels, known association with blood pressure.

Spike-and-slab posterior inclusion probabilities



Bayesian spike-and-slab regression. Shown: marginal posterior probabilities of inclusion (PPI). Red simulated signals.

- Nearly half of the “active” SNPs are correctly assigned high PPIs.
- Others receive very low PPIs despite being truly associated – weak signal and/or difficulty resolving signals when SNPs are highly correlated (LD).

Continuous two-group shrinkage priors

Formulation (continuous spike-and-slab prior):
Use a mixture of two *continuous* densities:

$$\beta_i \mid \gamma_i \sim \gamma_i\, g_\beta + (1-\gamma_i)\, g_0, \quad \gamma_i \sim \text{Bernoulli}(\pi), \tag{10}$$

where:

- g_0 is a density with strong concentration near zero (continuous “spike”);
- g_β is a diffuse distribution allowing large signals (continuous “slab”).

Typically, both are from **location-scale families**, and conjugate choices aid computation.

Characteristics:

- Avoids discontinuity from point-mass at zero (improves MCMC mixing);
- Does not produce exact zeros (shrinkage rather than formal variable selection).

Examples of spike-and-slab densities

Common choices for g_β and g_0:

- **Normal–Normal:**
$$g_\beta = \mathcal{N}(0, \sigma_\beta^2), \quad g_0 = \mathcal{N}(0, \sigma_0^2), \qquad 0 < \sigma_0^2 \ll \sigma_\beta^2$$
(e.g., Ishwaran and Rao, 2005);
- **Laplace–Laplace:**
$$g_\beta = \text{Laplace}(0, \lambda_\beta), \quad g_0 = \text{Laplace}(0, \lambda_0), \qquad 0 < \lambda_0 \ll \lambda_\beta$$
(e.g., Ročková and George, 2018);
- **Cauchy–Cauchy or Student's t–Student's t:** to allow for heavier tails and robustness to large signals.

Remarks:

- The choice of g_0 controls shrinkage near zero; g_β governs signal adaptivity.
- Gaussian pairs yield conjugacy; Laplace induces sparsity via ℓ_1 penalty analogues.

Additional priors for the “slab” component

- Formulations introduced so far placed independent priors on the nonzero coefficients, $\boldsymbol{\beta}_\gamma$, conditional on $\boldsymbol{\gamma}$, e.g.,

$$\boldsymbol{\beta}_\gamma \mid \boldsymbol{\gamma}, \sigma_\beta^2, \sigma_\varepsilon^2 \sim \mathcal{N}_{p_\gamma}\left(0, \sigma_\beta^2 \sigma_\varepsilon^2 \boldsymbol{I}_{p_\gamma}\right);$$

- Instead, Zellner (1986) introduced the ***g*-prior**, which assumes correlations among the regression **coefficients mimicking the correlations among predictors**:

$$\boldsymbol{\beta}_\gamma \mid \boldsymbol{\gamma}, g, \sigma_\varepsilon^2 \sim \mathcal{N}_{p_\gamma}\left(0, g\sigma_\varepsilon^2 (\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma)^{-1}\right);$$

- Its covariance is proportional to the **inverse Fisher information**, $\mathcal{I}(\boldsymbol{\beta}_\gamma) = \sigma_\varepsilon^{-2} \boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma$, and hence **mirrors the uncertainty of the MLE**, giving more prior variance where the data is less informative, and vice versa.
- This ties it conceptually to the Jeffreys[2] prior, $p(\boldsymbol{\beta}_\gamma) \propto |\mathcal{I}(\boldsymbol{\beta}_\gamma)|^{1/2}$ (same idea but improper/non-conjugate form).
- g controls prior strength relative to data – common choices include $g = n$ (unit information), g large (flat prior) or a hyperprior on g for flexibility.
- In GWAS, independent priors are often preferred as the effects $\boldsymbol{\beta}_\gamma$ need not mirror SNP correlations.

[2] Harold Jeffreys (1891-1989) also introduced Bayes factors in *Theory of Probability* (1935, expanded 1948/1961) as a general framework for Bayesian model comparison.

Prior specification for the “slab” variance

The slab variance, σ^2_β, controls the **amount of shrinkage** applied to non-zero coefficients. Choosing a hyperprior for σ^2_β has **strong implications** for inference.

See:

Andrew Gelman. “Prior distributions for variance parameters in hierarchical models”.
Bayesian Analysis, 1, 515 - 534, 2006.

Some specifications:

- Inverse-Gamma prior, which is conjugate, often used with very small shape and scale, e.g., 0.001 – this specification is not truly non-informative despite being commonly used as such (overly concentrated near zero);
- Less common: heavy-tailed priors (e.g., Half-Cauchy) or improper priors (e.g., flat on log-scale, $p(\sigma^2_\beta) \propto 1/\sigma^2_\beta$) often used to reflect vague prior knowledge;
- Alternative: fix σ^2_β to a constant (e.g., 1 or 10) – risk of miscalibrated shrinkage.

An alternative: prior based on PVE

Guan and Stephens (2011) argue that independence from model size (i.e., $\boldsymbol{\gamma}$) implicitly assumes that **complex models explain more variance** than simpler ones, which **can be unrealistic**, as biologically, one may expect many small effects (strong polygenicity of Boyle et al. (2017)'s omnigenic model) or few large ones.

They consider the following discrete spike-and-slab specification:

$$\beta_s \mid \gamma_s \sim \gamma_s \mathcal{N}\left(0, \sigma_\beta^2 \sigma_\varepsilon^2\right) + \left(1 - \gamma_s\right) \delta_0,$$

where

- σ_ε^2 is the residual variance;
- σ_β^2 is the slab variance, representing the typical size of nonzero effects.

Main idea: Define the prior on σ_β^2 given $\boldsymbol{\gamma}$, such that the prior on proportion of variance explained (PVE),

$$\mathrm{PVE}\left(\boldsymbol{\beta}, \sigma_\varepsilon^2\right) = \frac{\mathbb{V}\mathrm{ar}\left(\boldsymbol{X}\boldsymbol{\beta}\right)}{\mathbb{V}\mathrm{ar}\left(\boldsymbol{X}\boldsymbol{\beta}\right) + \sigma_\varepsilon^2},$$

is approximately uniform on $\left(0, 1\right)$, *independently* of $\boldsymbol{\gamma}$.

A structured prior via the PVE

- Assuming centred predictors, we have

$$\mathrm{PVE}(\boldsymbol{\beta}, \sigma_\varepsilon^2) \approx \frac{V(\boldsymbol{\beta}, \sigma_\varepsilon^2)}{V(\boldsymbol{\beta}, \sigma_\varepsilon^2) + 1}, \quad \text{where} \quad V(\boldsymbol{\beta}, \sigma_\varepsilon^2) = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{X}\boldsymbol{\beta})_i^2 / \sigma_\varepsilon^2$$

  is the empirical variance of $\boldsymbol{X}\boldsymbol{\beta}$ relative to error variance σ_ε^2.

- Next,

$$h^2(\boldsymbol{\gamma}, \sigma_\beta^2) := \frac{v(\boldsymbol{\gamma}, \sigma_\beta^2)}{v(\boldsymbol{\gamma}, \sigma_\beta^2) + 1}$$

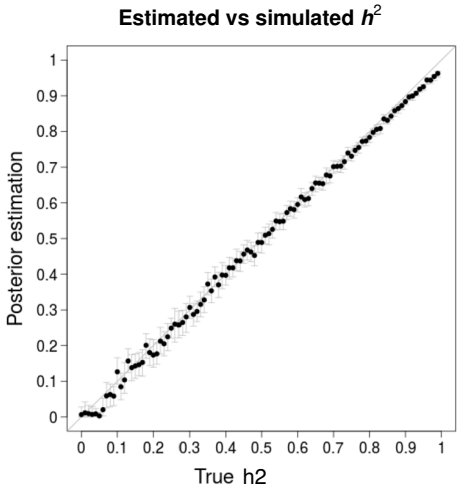
  may be considered a **proxy for the expected PVE**, where

$$v(\boldsymbol{\gamma}, \sigma_\beta^2) = \mathbb{E}[V(\boldsymbol{\beta}, \sigma_\varepsilon^2) | \boldsymbol{\gamma}, \sigma_\beta^2] = \sigma_\beta^2 \sum_{j:\gamma_j=1} s_j,$$

  with $s_j := \frac{1}{n} \sum_{i=1}^{n} X_{ij}^2$ the empirical variance of variable j.

- $h^2(\boldsymbol{\gamma}, \sigma_\beta^2)$ is only a rough guide to the expected PVE (ratio of expectations, not the expectation of the ratio).
- Impose $h^2 \sim \mathrm{Unif}(0, 1)$, *independently* of $\boldsymbol{\gamma}$; this induces a prior on $\sigma_\beta^2 \mid \boldsymbol{\gamma}$, such that more **complex models receive stronger shrinkage**, counteracting the tendency of common priors to inflate PVE with model size.

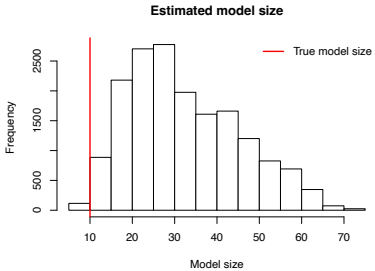
Posterior PVE estimates using simulated data

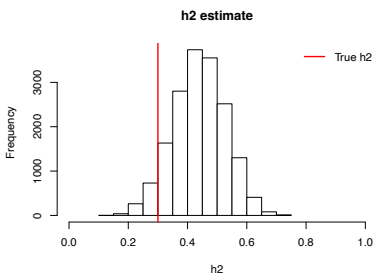


Posterior mean estimates with 95% credible intervals.

- 10 000 real SNPs sampled across the genome [height GWAS dataset from Yang et al. (2010), $n = 3\,925$]; local LD means sampled SNPs are nearly uncorrelated.
- Traits simulated with varying true h^2, using 200 randomly chosen “active” SNPs.
- Posterior intervals are tight and h^2 estimates remain **well-calibrated overall**.
- Some **downward bias** appears **at high** h^2 since the model must explain more variance with the same number of SNPs, so true effects are larger and the slab shrinks them inward; missed SNPs also contribute.
- A slab with heavier tails could help better preserve large effects.

Back to our simulated SBP example: model size and PVE





Individual models often include too many transient, correlated SNPs, reflecting instability in SNP selection under LD.

Estimated h^2 is variable and inflated, with wide posterior uncertainty.

- h^2 is evaluated at each MCMC iteration from the corresponding sampled model.
- Large models with redundant SNPs can reduce residual variance due to overfitting, creating the illusion of higher genetic signal.
- h^2 appears high in some iterations and lower in others, depending on the degree of redundancy and tagging.

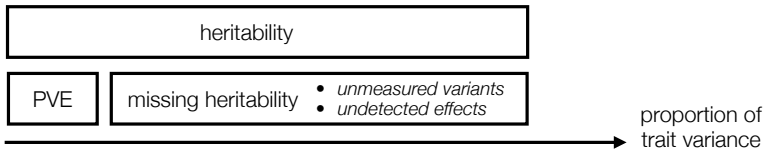
Later in the course, we will see how “fine-mapping” strategies help stabilise inference and disentangle overlapping signals.

A note on the terminology: PVE, heritability and missing heritability

- **PVE:** proportion of phenotypic variance explained by a linear predictor from *measured* genetic variants.
- **Heritability:** proportion of variance explained by *all* genetic variants (measured or unmeasured).
- **Polygenic risk score (PRS):** individual-level score estimating genetic susceptibility to a trait or disease, computed as a linear prediction where an individual's genotypes are weighted by effect sizes from GWAS.

"Missing heritability": GWAS often find much smaller PVE than heritability estimates from family studies (Maher, 2008), because of

- *unmeasured variants:* rare variants not well tagged by common SNPs;
- *undetected effects:* small-effect common variants, gene-gene and gene-environment interactions not captured by standard GWAS.



One-group priors

One-group shrinkage priors

- Unlike the two-group approach, one-group priors **do not explicitly partition coefficients into signal and noise**.
- All coefficients are modelled using **a single, continuous shrinkage component**.
- The classical **James–Stein estimator** (Stein, 1956; James and Stein, 1961) illustrates the benefits of **global shrinkage**, and can be viewed as an **early example** of empirical Bayes estimation based on a common normal prior:

$$\beta_i \mid \sigma_0^2 \sim \mathcal{N}(0, \sigma_0^2).$$

- **Modern one-group priors** extend this idea with hierarchical structures that allow **adaptive coefficient-specific shrinkage**, suitable for sparse, high-dimensional settings.

Global-local scale mixture priors

The general form is given by:

$$\beta_i \mid \lambda_i^2, \sigma_0^2 \sim \mathcal{N}(0, \sigma_0^2 \lambda_i^2), \qquad \lambda_i \sim f, \qquad \sigma_0 \sim g, \tag{11}$$

where f and g are densities on $\mathbb{R}^+$.

- This leads to **non-normal marginal distributions** for β_i that can accommodate both sparsity and heavy-tailed signals.
- σ_0 is the ***global* scale** (affecting all coefficients): it **adapts to** the overall **sparsity level.**
- λ_i are the ***local* scales** (providing coefficient-specific adaptation): they allow **individual signals** to **escape shrinkage** when supported by the data.

Posterior shrinkage under one-group priors

Lemma (Adapted from Carvalho et al., 2009)

Assume the normal means model (5) and prior (11) for β_i. Let

$$\kappa_i = \frac{1}{1 + \sigma_0^2 \lambda_i^2}, \qquad \kappa_i \in (0, 1). \tag{12}$$

Then, the conditional posterior mean of β_i can be expressed as

$$\mathbb{E}\left(\beta_i \mid y_i, \sigma_0^2, \lambda_i^2\right) = (1 - \kappa_i)\, y_i + \kappa_i \times 0,$$

so,

$$\mathbb{E}\left(\beta_i \mid y_i, \sigma_0^2\right) = \Big(1 - \mathbb{E}\left(\kappa_i \mid y_i, \sigma_0^2\right)\Big) y_i.$$

Parameter κ_i is called **shrinkage factor**, as it represents the weight placed on zero by the posterior mean of β_i. In this one-group framework, the quantity $1 - \mathbb{E}\left(\kappa_i \mid y_i, \sigma_0^2\right)$ plays a role analogous to the posterior inclusion probability in two-group models.

Examples of global-local priors

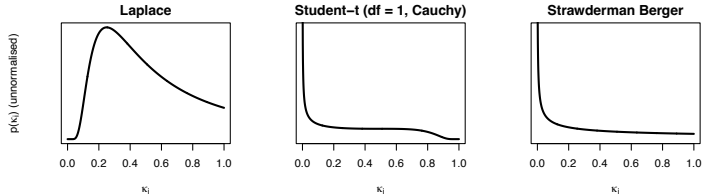
- **Laplace and Student-t prior:** local variances follow exponential and inverse-Gamma distributions, respectively.
- **Strawderman–Berger prior** (Strawderman, 1971; Berger, 1980):

$$\beta_i \mid \kappa_i \sim \mathcal{N}\left(0, \frac{1}{\kappa_i} - 1\right), \qquad \kappa_i \sim \text{Beta}\left(\frac{1}{2}, 1\right).$$

- **Normal/inverted-beta prior**: local variances follow an inverted-Beta density,

$$p\left(\lambda_i^2\right) = \frac{\left(\lambda_i^2\right)^{\alpha-1}\left(1+\lambda_i^2\right)^{-\alpha-\beta}}{\mathrm{B}(\alpha,\beta)}, \qquad \alpha, \beta > 0, \tag{13}$$

where $\mathrm{B}(\cdot,\cdot)$ is the beta function.



Examples of global-local priors

- **Horseshoe prior** (Carvalho et al., 2009, 2010):

$$\beta_i \mid \sigma_0^2, \lambda_i^2 \sim \mathcal{N}(0, \sigma_0^2 \lambda_i^2), \qquad \lambda_i \sim \mathrm{C}^+(0, 1), \tag{14}$$

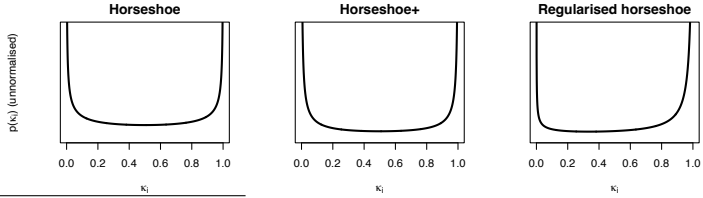
  where $\mathrm{C}^+(\cdot, \cdot)$ denotes the half-Cauchy distribution.

- **Horseshoe+ prior** (Bhadra et al., 2017): more aggressive noise reduction without sacrificing tail robustness:

$$\beta_i \mid \sigma_0^2, \lambda_i^2 \sim \mathcal{N}(0, \sigma_0^2 \lambda_i^2), \quad \lambda_i \mid \eta_i \sim \mathrm{C}^+(0, \eta_i), \quad \eta_i \sim \mathrm{C}^+(0, 1).$$

- **Regularised horseshoe prior** (Piironen and Vehtari, 2017): prevents overly large signals:

$$\beta_i \mid \lambda_i, \tau, c \sim \mathcal{N}\left(0, \tau^2 \tilde{\lambda}_i^2\right), \quad \tilde{\lambda}_i^2 = \frac{c^2 \lambda_i^2}{c^2 + \tau^2 \lambda_i^2}, \quad \lambda_i \sim \mathrm{C}^+(0, 1), \quad c > 0.$$[3]

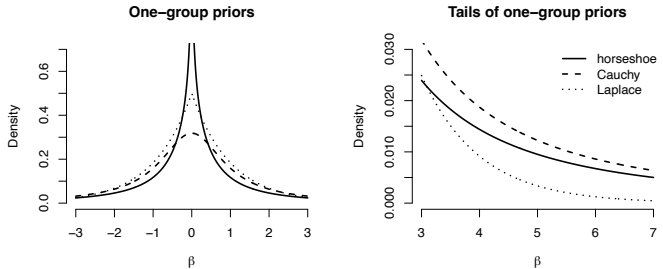


[3] In practice the authors recommend placing a prior on c.

Interpretation of shrinkage profiles

The marginal prior density of κ_i (shrinkage factor) reveals how the prior treats small versus large signals.

- **Horseshoe prior:** $\kappa_i \sim \text{Beta}(1/2, 1/2)$ places mass near 0 (no shrinkage for large signals) and near 1 (strong shrinkage for noise).
- **Student-t and Strawderman–Berger priors:** The density of κ_i exhibits a pole at zero, reflecting fat tails, but does not enforce full shrinkage.
- **Laplace prior:** Lighter tails lead to little mass near zero and may shrink genuine coefficients too much.



Densities of different one-group priors (Cauchy is t with df = 1).

Two-group priors as local-scale mixtures

A general prior formulation for $\boldsymbol{\beta}$ assumes a Gaussian prior:

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{\Lambda}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{0},\ \boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda}),$$

where

- $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, with $\lambda_i > 0$, $i = 1, \ldots, n$;
- $\boldsymbol{\Sigma}$ is positive semidefinite which may depend on σ^2_ε.

Placing a mixture prior over λ_i,

$$\lambda_i \mid \pi \sim \pi g_\lambda + (1 - \pi)\delta_0, \quad i = 1, \ldots, n,$$

where $\pi \in (0, 1)$ controls sparsity and g_λ is an absolutely continuous density on $\mathbb{R}^+$, gives rise to a two-group prior for $\boldsymbol{\beta}$. This representation bridges model selection and shrinkage:

- **model selection** via δ_0 on λ_i (or, equivalently, via introduction of latent binary variables γ_i) ...
- ... **embedded within the class of local-scale mixtures** of normal distributions.

Remarks and literature

> *"Because posterior sampling is computation-intensive and because variable selection is most desirable in contexts with many predictor variables,* ***computational considerations are important*** *in motivating and evaluating the approaches above.* ***The discrete model selection approach and the continuous shrinkage prior approach are both quite challenging in terms of posterior sampling.****"*
>
> *— Hahn and Carvalho*

Related work:

- George and McCulloch (1993)
- Ishwaran and Rao (2005)
- Park and Casella (2008), Hans (2009)
- Carvalho et al. (2009)
- Clyde et al. (2011)
- Polson and Scott (2010), Griffin and Brown (2017)
- etc.

Book: Gelman, A., Carlin, J. B., Stern, H., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.), available at `https://sites.stat.columbia.edu/gelman/book/BDA3.pdf`

Selection and multiplicity

Weak vs. strong sparsity

- **Strong sparsity:** most coefficients are exactly zero (model selection / two-group models);
- **Weak sparsity:** most coefficients are small, but not exactly zero (one-group models);
- In genomics, strong sparsity has been a common assumption, yet recent biological models challenge this view.
- *Omnigenic hypothesis* (Boyle et al., 2017): complex traits are influenced by thousands of variants, including many with tiny effects. While individual effects are weak, their combined contribution to phenotypic variance can be substantial.
- The distinction between weak and strong sparsity reflects a broader debate in Bayesian analysis: whether inference should be framed in terms of parameter estimation or formal hypothesis testing.

Variable selection: two-group vs. one-group

Recall that: in **two-group models**, variable selection is achieved via the binary latent indicators γ_s.

- The **highest posterior probability model** (HPM) is given by

$$\arg \max_{\boldsymbol{\gamma} \in \{0,1\}^p} p(\mathcal{M}_\gamma \mid \boldsymbol{y}). \tag{15}$$

- ... but with 2^p candidate models, enumeration or efficient sampling is very hard. Instead, marginal PPIs, $p(\gamma_s = 1 \mid \boldsymbol{y})$, are used.
- The **median probability model** (MPM; Barbieri and Berger, 2004) includes variables with PPI > 0.5 and is shown to outperform the HPM for prediction when predictors are orthogonal.

In **one-group models**, no exact zeros exist, so selection is done by thresholding posterior summaries.

- For instance, Carvalho et al. (2010) propose thresholding $1 - \mathbb{E}(\kappa_s \mid \boldsymbol{y})$ at 0.5 based on its **analogy with two-group posterior inclusion probabilities**.
- Threshold choices depend on prior structure and inferential goals – **no universally optimal rule** (for instance, stricter thresholds for sparse, interpretable models; looser thresholds for predictive performance).

Multiplicity control in two-group variable selection

- In two-group models, with prior slab g_β in

$$\beta_s \mid \pi \sim \pi\, g_\beta + (1-\pi)\,\delta_0, \quad s = 1, \ldots, p,$$

  π can be interpreted as a **prior proportion of included variables**.

- Given π, the **prior probability of model** $\mathcal{M}_\gamma$ is:

$$p(\mathcal{M}_\gamma \mid \pi) = \pi^{p_\gamma}(1-\pi)^{p-p_\gamma},$$

  where p_γ is the number of variables included in $\mathcal{M}_\gamma$ and $\gamma_s \mid \pi \sim \text{Bernoulli}(\pi)$, independently.

- Letting $\pi = 1/2$: each variable has **equal prior probability of being included or excluded** $\rightarrow$ **no sparsity** enforced.

- Other choices that imply **low inclusion probabilities** for individual predictors can effectively **enforce sparsity.**

- But **no fixed choice** of π that is **independent of** p **can adjust for multiplicity.**

Multiplicity control in two-group variable selection

Setting $\pi = 1/p$ is a common choice that favours sparse models as $p(\mathcal{M}_\gamma \mid \pi)$ decays rapidly with p_γ, however:

- Define **prior odds (PO)** penalties as:

$$\text{PO}(p_\gamma - 1 : p_\gamma) = \frac{p(\mathcal{M}_{\gamma_{p_\gamma - 1}} \mid \pi)}{p(\mathcal{M}_{\gamma_{p_\gamma}} \mid \pi)},$$

  where $\mathcal{M}_{\gamma_{p_\gamma - 1}}$ and $\mathcal{M}_{\gamma_{p_\gamma}}$ are models with $p_\gamma - 1$ and p_γ included variables, respectively.

- For fixed inclusion probability $\pi = 1/p$, we have

$$\text{PO}(p_\gamma - 1 : p_\gamma) = \frac{1 - \pi}{\pi} = p - 1.$$

$\rightarrow$ **Same penalty regardless of the number of included variables** (p_γ): no increasing preference for simpler models (*still favours sparse models, but* ***every new variable "costs" the same*** *regardless of how many are already in!*).

- In contrast, Scott and Berger (2010) have shown that **a fully Bayesian treatment** (e.g., Beta prior on π) **adjusts to the actual sparsity level** and **induces adaptive penalty** for model complexity **via non-separability**.

Multiplicity control in *fully Bayesian* two-group variable selection

- Assume $\pi \sim \text{Beta}(\alpha, \beta)$. Then prior model probability is:

$$p(\mathcal{M}_\gamma) = \int_0^1 p(\mathcal{M}_\gamma \mid \pi) p(\pi)\, d\pi = \frac{\text{B}(\alpha + p_\gamma, \beta + p - p_\gamma)}{\text{B}(\alpha, \beta)},$$

  where $\text{B}(\cdot, \cdot)$ is the Beta function.

- If $\alpha = \beta = 1$ (uniform prior on π), then:

$$p(\mathcal{M}_\gamma) = \frac{p_\gamma!\,(p - p_\gamma)!}{(p+1)\,p!} = \frac{1}{p+1}\binom{p}{p_\gamma}^{-1}.$$

- **Prior probability inversely proportional to the number of models of size p_γ**
  $\rightarrow$ built-in preference for simpler models!

Multiplicity control in *fully Bayesian* two-group variable selection

- The posterior probability for model $\mathcal{M}_\gamma$ therefore is:

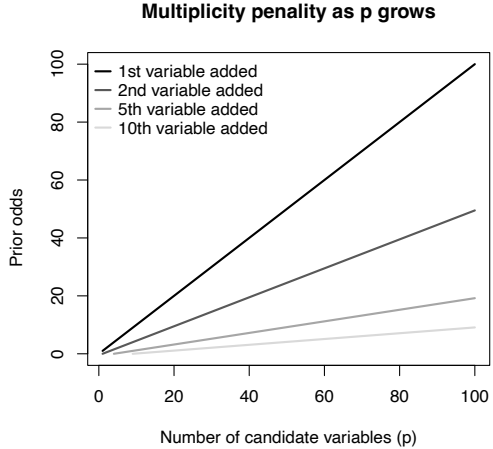
$$p(\mathcal{M}_\gamma \mid \boldsymbol{y}) \propto \frac{1}{p+1}\binom{p}{p_\gamma}^{-1} p(\boldsymbol{y} \mid \mathcal{M}_\gamma);$$

- Prior odds penalise more complex models:

$$\text{PO}(p_\gamma - 1 : p_\gamma) = \frac{p(\mathcal{M}_{\gamma_{p_\gamma - 1}})}{p(\mathcal{M}_{\gamma_{p_\gamma}})} = \frac{p - p_\gamma + 1}{p_\gamma}.$$

  - The **penalty depends** dynamically on **how many variables** are already **included.**
  - Adding the first variable is heavily penalised. It is easier to add variables once some are already included.

- Penalty also **grows with the number of candidate variables** *p*.



Typical hyperparameter choices for Beta prior on π

With $\pi \sim \text{Beta}(\alpha, \beta)$, the hyperparameter specification $\alpha = 1, \beta = 1$ (uniform prior) may be inadequate in GWAS and other high-dimensional settings. Instead, common choices are:

- $\alpha = \beta = 0.5$ (Jeffreys' prior):
  - U-shaped prior density places more mass near $\pi = 0$ and $\pi = 1$;
  - Favours either very sparse or very dense models;
- $\alpha = 1, \beta = p$ (more common):
  - encourages sparsity;
  - implies, a priori:

$$\mathbb{E}(\pi) = \frac{\alpha}{\alpha + \beta} \approx \frac{1}{p}, \qquad \mathbb{V}\text{ar}(\pi) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \approx \frac{1}{p^2},$$

    for p large.

Alternative prior on π

In high-dimensional problems, where sparsity can span orders of magnitude, standard $\text{Beta}(1, p)$ can concentrate too much mass near extreme sparsity levels.

Guan and Stephens (2011) propose a prior on $\log(\pi)$:

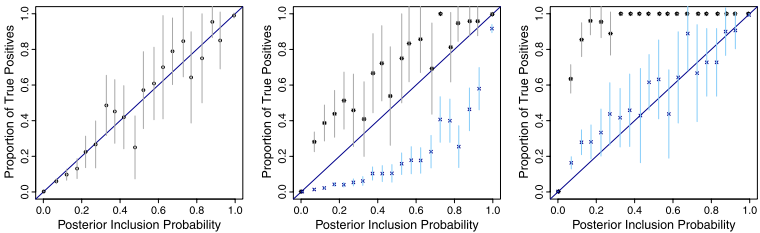
$$\log(\pi) \sim \mathcal{U}(a, b),$$

where $a = \log(1/p)$ and $b = \log(M/p)$.

- This gives **roughly equal prior weight across orders of magnitude of π.**
- Lower and upper bounds correspond to expectations of 1 and M variables included (the choice of M often motivated by computational cost).
- **Prior variance** under this prior is $\mathbb{V}\text{ar}(\pi) \approx M^2/p^2$ (times log factors), so for $M \gg 1$ (e.g., 500), it is much larger than the variance of $\pi \sim \text{Beta}(1, p)$, **reflecting genuine uncertainty about the degree of sparsity.**

Impact on calibration of PPIs

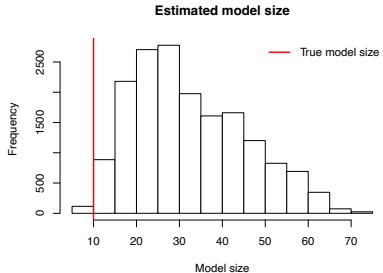
Estimating π and σ_β adapts inference to signal strength and multiplicity (important in polygenic or high LD cases).

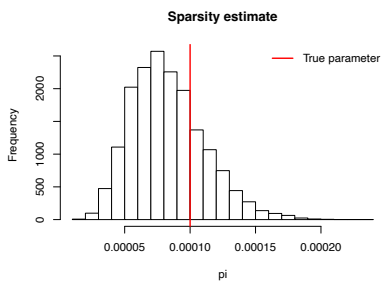


PPIs grouped into bins of width 0.05. *x*-axis: average PPI, *y*-axis: corresp. proportion of true effects within that bin, with $\pm$ SE.

- Left: Spike-and-slab regression with hyperpriors on π and $\sigma_\beta \rightarrow$ PPIs are well calibrated.
- Middle and right: Fixing π (middle) or σ_β (right) to values five-fold too small (black) or large (blue) leads to poorly calibrated PPIs. Note: σ_β five-fold too large has limited impact on calibration $\rightarrow$ **heavy-tailed priors are safer.**
- The **rankings remain largely robust to misspecification** of π and σ_β, even when PPI calibration is affected.

Back to the systolic blood pressure GWAS





When applied to the systolic blood pressure example:

- π is estimated as low (strong shrinkage), in line with the low PPIs for true signals in the presence of LD and modest effect sizes.
- This happens even though individual models tend to include too many SNPs: there is instability in selection, with many **correlated SNPs** in LD regions **appearing only sporadically, without accumulating support.**

Multiplicity control in one-group variable selection

Consider the example of a linear model (1) with a global-local prior on the regression coefficients:

$$\beta_s \mid \lambda_s^2, \sigma_0^2 \sim \mathcal{N}\big(0, \lambda_s^2 \sigma_0^2\big), \qquad \lambda_s \sim f, \qquad s = 1, \ldots, p, \tag{16}$$

where f is a density on $\mathbb{R}^+$.

- The global scale σ_0 controls the **overall sparsity level**.
- Typical choices for σ_0:
  - **Half-Cauchy prior:** $\sigma_0 \sim \mathrm{C}^+\big(0, 1\big)$ (default suggestion in early works; Carvalho et al., 2009);
  - **Fixed value:** Set σ_0 to a small constant $\sigma_0 = c > 0$.
- $\mathrm{C}^+\big(0, 1\big)$ is often too vague: can lead to insufficient shrinkage, especially when the data are weakly informative.
- Fixing σ_0 results in a lack of adaptation to the actual sparsity level.

Multiplicity control in one-group variable selection

- Piironen and Vehtari (2017) propose a **prior** specification **based on the expected number of relevant variables.**
- The conditional posterior for $\boldsymbol{\beta}$ given the hyperparameters and data can be written as

$$p(\boldsymbol{\beta} \mid \boldsymbol{\Lambda}, \sigma_0^2, \sigma_\varepsilon^2, \boldsymbol{y}) = \mathcal{N}(\bar{\boldsymbol{\beta}}, \boldsymbol{\Sigma}),$$

  where

$$\bar{\boldsymbol{\beta}} = \sigma_0^2 \boldsymbol{\Lambda} \left(\sigma_0^2 \boldsymbol{\Lambda} + \sigma_\varepsilon^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1} \right)^{-1} \hat{\boldsymbol{\beta}}, \qquad \boldsymbol{\Sigma} = \left(\sigma_0^{-2} \boldsymbol{\Lambda}^{-1} + \sigma_\varepsilon^{-2} \boldsymbol{X}^T \boldsymbol{X} \right)^{-1},$$

  with $\boldsymbol{\Lambda} = \text{diag}(\lambda_1^2, \ldots, \lambda_p^2)$ and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$ is the OLS estimate (assuming the inverse exists).
- If the predictors are uncorrelated, with $\mathbb{E}(\boldsymbol{X}_s) = 0$ and $\mathbb{V}\text{ar}(\boldsymbol{X}_s) = s_s^2$, then $\boldsymbol{X}^T \boldsymbol{X} \approx n \, \text{diag}(s_1^2, \ldots, s_p^2)$, and

$$\bar{\beta}_s \approx (1 - \kappa_s) \hat{\beta}_s, \qquad s = 1, \ldots, p,$$

  where

$$\kappa_s = \frac{1}{1 + n \sigma_\varepsilon^{-2} \sigma_0^2 s_s^2 \lambda_s^2}$$

  is the shrinkage factor (with same interpretation as seen for the normal means model!).
- Note that $\bar{\boldsymbol{\beta}} \to 0$ as $\sigma_0 \to 0$ and $\bar{\boldsymbol{\beta}} \to \hat{\boldsymbol{\beta}}$ as $\sigma_0 \to \infty$.

Multiplicity control in one-group variable selection

- **In the case of the horseshoe prior**, i.e., where f is $\mathrm{C}^{+}(0, 1)$ in (16), independently for all λ_s, the implied prior on κ_s is:

$$p(\kappa_s \mid \sigma_0, \sigma_\varepsilon) = \frac{1}{\pi} \frac{a_s}{\sqrt{\kappa_s(1-\kappa_s)}\,[(a_s^2 - 1)\kappa_s + 1]}, \tag{17}$$

  where $a_s = \sqrt{n}\sigma_\varepsilon^{-1}\sigma_0 s_s$.

- When $a_s = 1$, the distribution reduces to a Beta$(1/2, 1/2)$, with horseshoe shape.

- For fixed σ_0, though, the prior does not adapt to the dimension p (the effective sparsity depends p).

Multiplicity control in one-group variable selection

- To control sparsity across different p, Piironen and Vehtari (2017) propose choosing σ_0 based on the **"effective number of nonzero coefficients"** defined as:

$$m_{\text{eff}} = \sum_{s=1}^{p} \big(1 - \kappa_s\big). \tag{18}$$

- When κ_s are close to 0 and 1 (as they typically are for the horseshoe prior), $\big(18\big)$ describes the number of variables included in the model, therefore serving as an indicator of the effective model size.
- Using (17), it can be shown that

$$\mathbb{E}\big(m_{\text{eff}} \mid \sigma_0, \sigma_\varepsilon\big) = \sum_{s=1}^{p} \frac{a_s}{1 + a_s}, \qquad \mathbb{V}\text{ar}\big(m_{\text{eff}} \mid \sigma_0, \sigma_\varepsilon\big) = \sum_{s=1}^{p} \frac{a_s}{2\big(1 + a_s\big)^2},$$

where $a_s = \sqrt{n}\sigma_\varepsilon^{-1}\sigma_0 s_s$.

Multiplicity control in one-group variable selection

- Assuming that the variables are also standardised ($s_s = 1$), this simplifies to

$$\mathbb{E}(m_{\text{eff}} \mid \sigma_0, \sigma_\varepsilon) = \frac{\sqrt{n}\sigma_\varepsilon^{-1}\sigma_0}{1 + \sqrt{n}\sigma_\varepsilon^{-1}\sigma_0} p, \qquad \mathbb{V}\text{ar}(m_{\text{eff}} \mid \sigma_0, \sigma_\varepsilon) = \frac{\sqrt{n}\sigma_\varepsilon^{-1}\sigma_0}{2(1 + \sqrt{n}\sigma_\varepsilon^{-1}\sigma_0)^2} p.$$

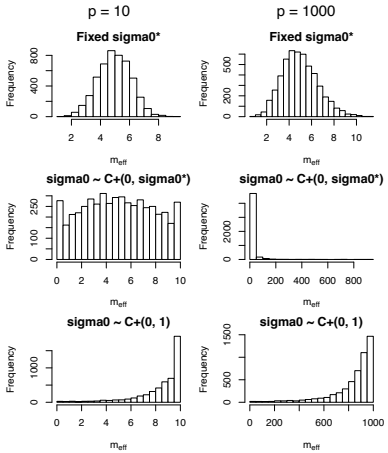
- Note that σ_0 should scale as $\sigma_0 \propto \sigma_\varepsilon / \sqrt{n}$ to avoid a prior specification favouring models of varying sizes depending on the noise level and sample size.
- Piironen and Vehtari (2017) propose setting σ_0 by solving

$$\mathbb{E}(m_{\text{eff}} \mid \sigma_0, \sigma_\varepsilon) = p_0,$$

for **a prior guess p_0 of the number of relevant variables**, giving

$$\sigma_0^* = \frac{p_0}{p - p_0} \frac{\sigma_\varepsilon}{\sqrt{n}}.$$

Effect of prior choices for σ_0 on prior distribution of m_{eff}



Adapted from Piironen and Vehtari (2017). Prior draws for m_{eff} under different priors (rows) for σ_0, with $n = 100$, $\sigma_\varepsilon = 1$, $p \in \{10, 1000\}$ (columns). The first two priors use $p_0 = 5$ as prior guess for the nb of active variables.

- Fixing $\sigma_0 = \sigma_0^*$ leads to a symmetric prior for m_{eff} centred around p_0.
- **Half-Cauchy prior with scale** σ_0^*: heavier tail, placing **more mass on large** m_{eff}, especially when p is large.
- Standard half-Cauchy $\sigma_0 \sim \text{C}^+(0, 1)$: favours solutions with most coefficients unshrunk, causing weak shrinkage; problematic for large p unless σ_0 is strongly identified by data.
- Note that changing σ_ε or n would alter the induced prior for m_{eff} for this standard half-Cauchy, unlike for the other priors.

$\rightarrow$ Based on further numerical experiments, the authors recommend:

$$\sigma_0 \mid \sigma_\varepsilon \sim \text{C}^+(0, \sigma_0^*).$$

as a **weakly informative default choice** instead of fixing the global scale to σ_0^*.

Summary

- **Principled multiplicity control** is crucial in problems with a large number of candidate predictors, like GWAS.
- **Univariate analysis** ignores correlations and **provides no global calibration** $\rightarrow$ **choices** about the sparsity and typical size of the nonzero coefficients **implicitly made** when specifying significance thresholds.
- Estimation of the global parameters π and σ_β (two-group models) or σ_0 (one-group models) via appropriate **hierarchical prior specifications renders the entries of $\boldsymbol{\beta}$ dependent** in the marginal prior $p(\boldsymbol{\beta})$:
  - yields ***self-adaptivity to sparsity*** via a ***non-separable*** penalty that borrows strength across coefficients and adapts to varying sparsity levels;
  - enables **direct estimation of interpretable global quantities** – such as the proportion of variance explained (PVE) quantifying the total genetic contribution to complex traits.
  - provides a ***built-in correction for multiplicity***, that discourages over-selection unless justified by strong evidence.
- Different strategies have been proposed for specifying such hyperpriors and their hyperparameters, based on expected numbers of non-zero coefficients or upper bounds for them.

Bayesian false discovery rate (FDR): local vs. tail-area

We test hypotheses H_{0s} versus H_{1s}, for $s = 1, \ldots, p$, based on test statistics z_s. Efron et al. (2001) introduces:

- **The local FDR:**

$$\text{fdr}(z_s) = p(H_{0s} \mid z_s) = \frac{\pi_0 f_0(z_s)}{f(z_s)},$$

  where $f_0(z)$ is the density under the null, $f(z)$ the overall (mixture) density and π_0 the prior null probability. Interpreted as the posterior probability that H_{0s} is true given z_s. Local (pointwise), useful for ranking discoveries.

- **The tail-area FDR:**

$$\text{Fdr}(z_s) = p(H_0 \mid Z \leq z_s) = \frac{\pi_0 F_0(z_s)}{F(z_s)},$$

  where $F_0(z)$ and $F(z)$ are the CDFs corresponding to $f_0(z)$ and $f(z)$, respectively.
  Estimates the expected proportion of nulls among all test statistics z_r such that $z_r \leq z_s$. Cumulative, similar in spirit to classical frequentist FDR methods.

Note: Although these quantities are defined using posterior-like expressions, they are derived under an empirical Bayes framework and rely on large-scale testing assumptions. They consider the distribution of test statistics, rather than model parameters.

Bayesian FDR for spike-and-slab models

In two-group models, posterior probabilities of inclusion offer a natural basis for model-based FDRs.

One can define the Bayesian FDR (Newton et al., 2004) as

$$\text{FDR}(\tau) = \frac{\sum_{s=1}^{p} (1 - \text{PPI}_s)\, \mathbf{1}\{\text{PPI}_s > \tau\}}{\sum_{s=1}^{p} \mathbf{1}\{\text{PPI}_s > \tau\}}, \tag{19}$$

where $\text{PPI}_s = p(\gamma_s = 1 \mid \boldsymbol{y})$, for a given threshold $\tau \in [0, 1]$.

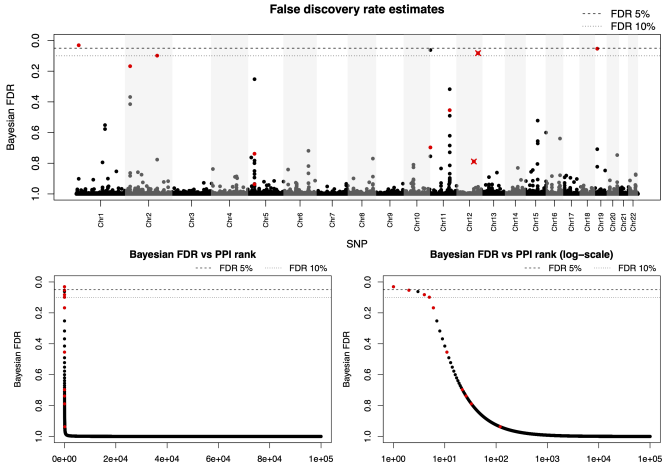
- Gives a *global* empirical Bayes FDR estimate over all variables selected at threshold τ.
- Model-based control: uses PPIs from full Bayesian model.
- Varying τ over a fine grid yields an estimated $\text{FDR}(\tau)$ curve.
- Selecting the smallest $\tau^\star$ such that $\text{FDR}(\tau^\star) \leq \alpha$ enables declaring all variables with $\text{PPI}_s > \tau^\star$ as discoveries at level α.

Bayesian FDR: step-up assignment

Define ordered PPIs: $\mathrm{PPI}_{(p)} \geq \cdots \geq \mathrm{PPI}_{(1)}$, and for each $k = 1, \ldots, p$, compute:

$$\mathrm{FDR}_{(k)} = \frac{1}{k} \sum_{s=0}^{k-1} \big(1 - \mathrm{PPI}_{(p-s)}\big).$$

- This is the expected false discovery rate among the top-k variables.
- Each variable is assigned the minimum FDR level at which it would be selected.
- Closely approximates threshold-based FDR(τ) curves (19) in large-scale settings.



Estimated Bayesian FDR for the systolic blood pressure GWAS example.

Red: simulated signals.

Correlated tests and Bayesian FDR

- **Correlated predictors or test statistics** can substantially distort the null distribution of marginal test statistics, which complicates the estimation of $f_0(z)$ and $f(z)$ in empirical Bayes FDR procedures (Efron, 2007).
- As a result, **local FDR** and **tail-area FDR** estimates may be biased.
- As we have seen, **fully Bayesian models**, such as two-group spike-and-slab priors, can in principle account for correlations via the likelihood and prior, but PPIs may still be sensitive to local dependencies.
- **Permutation-based approaches** help empirically preserve dependence in the null, but they are computationally expensive which limits their applicability within hierarchical Bayesian models.
- Developing **reliable Bayesian FDR procedures under dependence** remains an active and important area of research.

Structured priors

Local LD structure and region-level inference

- Common models assume that SNPs are assigned the same prior inclusion probability, which ignores spatial correlation where **SNPs in linkage disequilibrium (LD) form natural *groups* (*loci*** [4]**)**;
- This LD **correlation is mostly *local***: the genotype correlation matrix $\boldsymbol{X}^T\boldsymbol{X}$ **is approximately *banded***;
- Correlated SNPs from a same locus which display associations with the trait **often tag the same underlying mechanisms.**

$\rightarrow$ motivates a shift in focus from pinpointing risk SNPs to **identifying *associated loci*.**

- This mitigates method-specific differences in how signal is estimated (e.g., marginal methods flag any SNP correlated with a functional SNP; sparse penalised methods, such as the lasso, tend to select one or few representatives; Bayesian approaches may spread posterior inclusion probabilities across correlated SNPs).
- **SNPs within the identified loci** can **subsequently** be **prioritised** through dedicated follow-up analyses (see later).

[4] *Defining loci is itself non-trivial and involves a series of choices (e.g., LD thresholds, physical distance, gene boundaries).*

Encoding biological group structure

Can we exploit group structures to improve inference and interpretability?

- Bayesian hierarchical models can encode such structure:
  - **Group-level parameters** capture shared information (e.g., **group-level activation**);
  - **Predictor-level parameters** remain flexible (e.g., **within-group adaptivity**).
- **Multilevel modelling** permits:
  - **information borrowing** within and across groups;
  - **selective shrinkage** that respects group relevance;
  - **mitigation of** correlation-induced **redundancy.**
- Note: beyond LD, **group structure may arise from functional annotations** (e.g., coding, regulatory, conserved), gene membership, biological pathways, cell-type or tissue-specific effects.

Frequentist group lasso

Performs joint selection of predefined groups $g = 1, \ldots G$.

Optimisation problem (group lasso, Yuan and Lin, 2006):

$$\min_{\boldsymbol{\beta}} \left\{ \left\| \boldsymbol{y} - \sum_{g=1}^{G} \boldsymbol{X}_g \boldsymbol{\beta}_g \right\|_2^2 + \lambda \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_{\boldsymbol{K}_g} \right\}, \quad \|\boldsymbol{z}\|_{\boldsymbol{K}_g} = (\boldsymbol{z}^\mathsf{T} \boldsymbol{K}_g \boldsymbol{z})^{1/2},$$

where:

- $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector;
- $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_G]$, with $\boldsymbol{X}_g \in \mathbb{R}^{n \times |g|}$, is the design matrix;
- $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \ldots, \boldsymbol{\beta}_G^\top]^\top$, with $\boldsymbol{\beta}_g \in \mathbb{R}^{|g|}$, is the coefficient vector;
- $\boldsymbol{K}_g$, $g = 1, \ldots, G$, are positive definite matrices.

- Encourages **sparsity at the group level**, through penalty for $\lambda > 0$;
- Reduces to **standard lasso when each group contains a single variable** and $\boldsymbol{K}_g = \boldsymbol{I}_{|g|}$.

Bayesian group lasso and spike-and-slab extension

Kyung et al. (2010) show that the **multivariate Laplace prior**,

$$p(\boldsymbol{\beta}_g) \propto \exp\left(-\frac{\lambda}{\sigma}\|\boldsymbol{\beta}_g\|_2\right), \quad g = 1, \dots G,$$

corresponding to a group lasso penalty, can be written as a scale mixture of normals (easier posterior inference):

$$\boldsymbol{\beta}_g \mid \sigma^2, \tau_g^2 \sim \mathcal{N}_{|g|}\left(\mathbf{0}, \sigma^2\tau_g^2\boldsymbol{I}_{|g|}\right), \quad \tau_g^2 \sim \text{Gamma}\left(\frac{|g|+1}{2}, \frac{\lambda^2}{2}\right).$$

- Encourages group-wise shrinkage of coefficients (like the frequentist group lasso).
- Estimation based on posterior means or medians does not produce exact zero estimates (unlike the frequentist group lasso).

Extension (spike-and-slab prior for group selection): introduce binary indicators $\gamma_g \in \{0, 1\}$ for group inclusion:

$$\boldsymbol{\beta}_g \mid \gamma_g, \sigma^2, \tau_g^2 \sim \gamma_g\mathcal{N}_{|g|}(\mathbf{0}, \sigma^2\tau_g^2\boldsymbol{I}_{|g|}) + (1-\gamma_g)\delta_0(\boldsymbol{\beta}_g), \quad \gamma_g \sim \text{Bernoulli}(\pi).$$

This prior formulation formally encodes group *selection* by assigning exact zeros to $\boldsymbol{\beta}_g$ of inactive groups.

Structured grouping in spike-and-slab priors

The latent indicator vector $\boldsymbol{\gamma} \in \{0, 1\}^p$ is reduced to $\boldsymbol{\gamma} \in \{0, 1\}^G$, for $G \ll p$ groups.

Chipman (1996) was probably the first to use **structural grouping information** in Bayesian variable selection, using a continuous spike-and-slab prior:

> *"Not only does the grouping principle reduce the size of the total model space, but it makes headway in dealing with the pitfalls of multiple comparisons."*
>
> *— Hugh Chipman (1996)*

In practice:

- Identify groups $g \in \{1, \ldots, G\}$ with large posterior probability of inclusion: $p(\gamma_g = 1 \mid \boldsymbol{y})$;
- Within selected groups, inspect entries of $\boldsymbol{\beta}_g$ with large posterior means.

Grouped horseshoe prior: different formulations

BGHS (Bayesian grouped horseshoe, Xu et al., 2016; He and Wand, 2024):

$$\boldsymbol{\beta}_g \mid \sigma_0^2, \nu_g^2 \sim \mathcal{N}_{|g|}\left(\mathbf{0}, \sigma_0^2 \nu_g^2 \boldsymbol{I}_{|g|}\right), \quad \nu_g \sim \mathrm{C}^+(0,1), \quad \sigma_0 \sim \mathrm{C}^+(0,A), \quad A > 0, \quad g = 1, \ldots, G.$$

- Controls overall sparsity via the global scale σ_0 and group-level sparsity via the local scale ν_g;
- Shrinks all entries of $\boldsymbol{\beta}_g$ in a group specific way, via ν_g, which has heavy tail.

HBGHS (Hierarchical BGHS, Xu et al., 2016, Alt. formulation):

$$\beta_s \mid \sigma_0^2, \nu_{g(s)}^2, \lambda_s^2 \sim \mathcal{N}\left(0, \sigma_0^2 \nu_{g(s)}^2 \lambda_s^2\right), \quad \lambda_s \sim \mathrm{C}^+(0,1), \quad \nu_g \sim \mathrm{C}^+(0,1), \quad \sigma_0 \sim \mathrm{C}^+(0,A), \quad s = 1, \ldots, p.$$

- Activates/deactivates groups (SNP loci) via the group-level scale ν_g;
- Enables predictor-level adaptivity via the local scale λ_s (allows a few strong effects to escape from “inactive groups”).

Beyond exchangeability: exploiting SNP-level information

- So far we saw how to encode **known group structures**, such as LD blocks via Bayesian group shrinkage;
- But **individual SNPs** may also **differ in biological plausibility**, which breaks exchangeability in a different way;
- External information, from prior studies or genomic annotations, may help **prioritise** likely functional variants, especially within high-LD loci where true signals harder to isolate.

Definition (De Finetti, 1937)

Exchangeability means the joint distribution of random variables is invariant under permutation.
For $\beta_1, \beta_2, \ldots, \beta_p$, they are exchangeable if:

$$(\beta_1, \beta_2, \ldots, \beta_p) \overset{d}{=} (\beta_{\sigma(1)}, \beta_{\sigma(2)}, \ldots, \beta_{\sigma(p)}),$$

for any permutation σ.

What kind of information?

Definition (Morgensztern et al., 2018)

The epigenome is the complete description of all the chemical modifications to DNA and histone proteins that regulate the expression of genes within the genome.

- These regulatory mechanisms include DNA methylation, histone modifications and small noncoding RNAs.
- They underpin tissue- and context-specific gene regulation.

("Epigenome" comes from the Greek prefix epi-, meaning 'on top of', highlighting its role in regulating gene activity "above" the DNA sequence itself.)

- Other types of **SNP-level annotations**:
  - **Genomic location:** exonic, intronic, intergenic, UTRs, regulatory regions.
  - **Functional features:** enhancer/promoter overlap, TF binding sites, chromatin states.
  - **Quantitative indicators:** prior GWAS hits, conservation scores, allele frequency.

How to encode such information?

- **Most GWAS use this only *post hoc***: practitioners inspect peaks for known genomic marks, which is *ad hoc*, **subjective** and **difficult to scale**.
- The Bayesian framework lets us **encode such information *a priori* via the model hierarchy.**
- Simplest example: In a spike-and-slab model, a higher prior inclusion probability π_s can be assigned to a SNP $\boldsymbol{X}_s$ based on existing evidence about a likely higher functional relevance.

> *"Exchangeability is a function not just of reality, but of the information you have."*
>
> *— Andrew Gelman*

Can we *learn* which annotations matter?

SNP effects may be **exchangeable, conditionally, based on shared annotations or biological profiles** ("conditional exchangeability").

- Many sources of annotation data exist, but not all are equally relevant for a given trait or context.
- Pre-specifying prior inclusion probabilities based on functional annotations may be too rigid or subjective: relevance of annotations will be tissue, condition and region-specific.
- **Can we learn which annotations may matter** for the genetic association problem at hand ?

This motivates introducing the concept of **co-data**.

Definition (te Beest et al., 2017):

Co-data are any type of information that is available on the variables of the primary data, but does not use its response labels.

Co-data can be used as predictor-level information in the model hierarchy to guide variable selection:
In GWAS, SNP-level annotations can serve as co-data capturing the potential biological relevance of each SNP in controlling the trait of interest, which in turn could improve association estimates.

Top-level co-data submodel

Examples of hierarchical co-data priors for:

- spike-and-slab regression (van de Wiel et al., 2018):

$$\beta_s \mid \gamma_s \sim \gamma_s\, g_\beta + (1-\gamma_s)\, g_0, \quad \gamma_s \mid \pi_s \sim \text{Bernoulli}(\pi_s),$$

where g_0 and g_β are the "spike" and "slab" distributions, respectively. Predictor-specific inclusion probabilities are modulated by co-data:

$$\pi_s = h^{-1}(\boldsymbol{V}_s^\top \boldsymbol{\xi}), \quad s = 1, \ldots, p,$$

where $\boldsymbol{V}_s$ is a $L \times 1$ vector of co-data for predictor s, $\boldsymbol{\xi}$ is the corresponding vector of effects (assigned its own prior) and h is a link function (e.g., logit or probit).
$\rightarrow$ **adaptive selection:** Predictors with supportive annotations will be **prioritised for inclusion**.

- horseshoe regression (Busatto and van de Wiel, 2023):

$$\beta_s \mid \lambda_s^2, \sigma_0^2 \sim \mathcal{N}(0, \lambda_s^2 \sigma_0^2), \qquad \lambda_s \sim \text{C}(\mathbf{v}_s^\top \boldsymbol{\xi}, 1)\, \mathbb{1}(\lambda_s > 0), \quad s = 1, \ldots, p,$$

where $\text{C}(\cdot, \cdot)$ denotes the Cauchy distribution; the local scales λ_s are influenced by co-data $\boldsymbol{V}_s$ via the effects $\boldsymbol{\xi}$.
$\rightarrow$ **adaptive shrinkage:** Effects of predictors with supportive annotations will be **less shrunk**.

Inferring the relevance of annotations

- When many annotations are available, **some (or even all) may be irrelevant** to specific GWAS of interest.
- We may want to ***select* the annotations** relevant to the specific GWAS considered, **out of a potentially large number of candidates** (hundreds or thousands).
- Place a **sparse prior** on $\boldsymbol{\xi}$ (e.g., spike-and-slab or continuous shrinkage) to:
  - pinpoint relevant annotations from the data, for the association problem at hand;
  - quantify the extent of their relevance for the effects of the predictors they concern.
- Example:

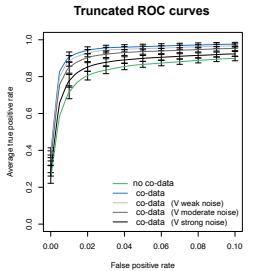
$$\xi_l \mid \zeta_l \sim \zeta_l g_\beta + (1 - \zeta_l) g_0, \quad \zeta_l \sim \text{Bernoulli}(\rho), \quad l = 1, \ldots, L,$$

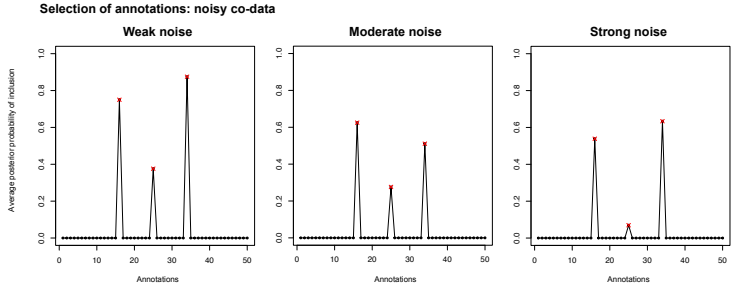
  where $p(\zeta_l = 1 \mid \boldsymbol{y})$ can be used to select the relevant annotations.
  $\rightarrow$ **Improves accuracy** of association estimates ***and* generates mechanistic hypotheses** by highlighting specific **annotations that might underlie the mechanisms of interest.**
- **Caveat:** The top-level model hierarchy can only be reliably inferred if the data are sufficiently informative.

Simulated example

Vanilla vs. annotation-informed (co-data) spike-and-slab models (simulated data, 100 replicates):

- $p = 250$ candidate SNPs for $n = 400$ individuals.
- $L = 50$ candidate annotations supplied to the co-data model, with different levels of noise added to the annotation data.
- Annotations 12, 37, 49 are simulated as triggering the genetic signal (red).





Bayesian fine-mapping

Bayesian fine-mapping

Fine-mapping aims to **prioritise SNPs** that are most likely to be functional in a pre-identified GWAS-implicated genomic region ("risk locus").

- In given risk loci, association signals can span multiple SNPs in LD but, typically, only a few are functional; others tag them through LD. Note: the functional SNP(s) may or may not be genotyped!
- Goals: (i) help understand how many distinct causal associations may underlie the association results; (ii) infer **which subset** of SNPs in the risk locus may be causal (or best tag) causal variants; (iii) quantify the strength of evidence.
- This guides: (i) the selection of variants for follow-up in downstream functional validation experiments; (ii) the identification of therapeutic targets; (iii) the discovery of new biological mechanisms behind diseases.

Bayesian approach:

- Define a prior over all possible *models* $\mathcal{M}_\gamma$ (i.e., subsets of all SNPs in the risk locus);
- Compute posterior probabilities for models and SNPs;
- Use prior constraints (e.g., ≤ 4 functional SNPs) to reduce search space.

Bayesian model for fine-mapping

- Assume that $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma, \sigma_\varepsilon^2 \boldsymbol{I}_n)$;
- Let $\mathcal{M}_\gamma$ be a model with candidate predictors indexed by $\boldsymbol{\gamma} \subset \{1, \ldots, p\}$, with $p_\gamma \leq M$ included predictors, for M small;
- Prior on models (sparser models favoured & uniform over all models of a given size):

$$p(\mathcal{M}_\gamma) \propto \binom{p}{p_\gamma}^{-1} \mathbb{1}(p_\gamma \leq M);$$

- Prior on effects:

$$\boldsymbol{\beta}_\gamma \mid \sigma_\beta^2 \sim \mathcal{N}(0, \sigma_\beta^2 \boldsymbol{I}_{p_\gamma});$$

- **MCMC over model space:** propose local moves (add, remove, swap predictors);
- Provides a posterior over models,

$$p(\mathcal{M}_\gamma \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \mathcal{M}_\gamma) p(\mathcal{M}_\gamma),$$

from which **credible sets** of predictors can be derived and ranked for fine mapping.

Typical posterior summaries of interest

- **PPI for each predictor:** sum of posterior probabilities over models containing that predictor;
- **Top models:** highest posterior probability models (HPM);
- **Model size distribution:** posterior on the number of active predictors;
- **Credible sets:** minimal sets of predictors designed to capture the active predictors $\rightarrow$ smaller credible sets reflect greater certainty (assuming correct coverage).

Three definitions of credible sets in fine-mapping

Let $\mathcal{S} = \{1, \ldots, p\}$ be the indices of all candidate predictors. Definitions in use for **a ρ-level credible set** $\mathcal{C}$:

1. Marginal PPI-based

$$\text{Find a minimal set } \mathcal{C} \subseteq \mathcal{S} \text{ such that } \sum_{s \in \mathcal{C}} \text{PPI}_s \ \geq \ \rho,$$

where PPI_s denote the posterior probability of inclusion for predictor $\boldsymbol{X}_s$. *Used in FINEMAP (Benner et al. 2016).*

2. Model–based (at least one active predictor)

$$\text{Find a minimal set } \mathcal{C} \subseteq \mathcal{S} \text{ such that } \sum_{\mathcal{M}_\gamma \cap \mathcal{C} \neq \emptyset} p(\mathcal{M}_\gamma \mid \boldsymbol{y}) \ \geq \ \rho,$$

i.e., $\mathcal{C}$ has probability $\geq \rho$ of containing *at least one active predictor. Used in SuSiE (Wang et al. 2020).*

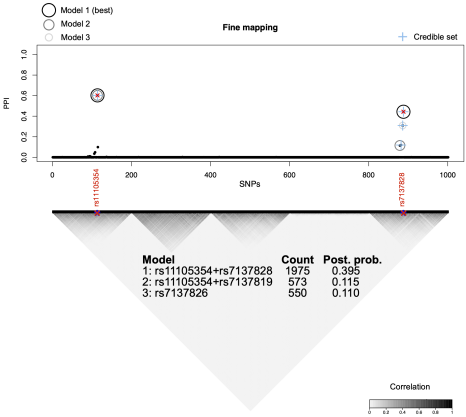
3. Model–based (all active predictors)

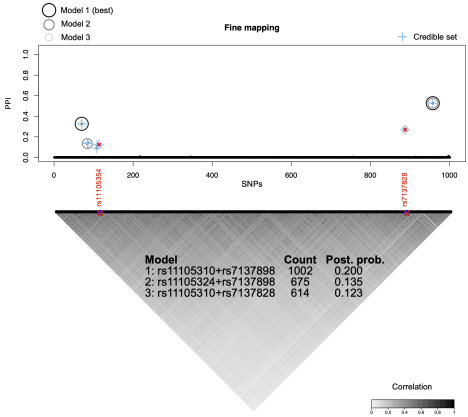
$$\text{Find a minimal set } \mathcal{C} \subseteq \mathcal{S} \text{ such that } \sum_{\mathcal{M}_\gamma \subseteq \mathcal{C}} p(\mathcal{M}_\gamma \mid \boldsymbol{y}) \ \geq \ \rho,$$

i.e., $\mathcal{C}$ has probability $\geq \rho$ of containing *all active predictors. Used in CAVIAR (Hormozdiari et al. 2014).*

Simulated example

Yiran Li





Bayesian fine-mapping on a systolic blood pressure locus (involving the active SNPs rs11105354 and rs7137828, on chr12). Metropolis–Hastings algorithm described on slide 97 with $M = 4$. Scenarios with two simulated LD structures: five SNP blocks with varied autocorrelation each (left) and single block with highly correlated SNPs (right).

From GWAS to causal variant identification

Given that GWAS estimates tend to be unstable due to LD structure, the possibility of subsequently refining signals with fine-mapping suggests the following pipeline:

- Apply *strong* LD pruning before GWAS for a more stable detection of signals;
- Identify regions of interest (loci) based on GWAS hits;
- Narrow down likely functional SNPs using Bayesian fine-mapping applied to the full SNP set within each locus.

In practice:

- Fine-mapping is ideally conducted in an **independent dataset**, often with larger sample size and denser genotyping or imputation, to avoid data reuse and maximise resolution;
- Functional annotations can inform the prior over model space within risk loci (similar use as for GWAS co-data models);
- Many fine-mapping approaches rely on GWAS summary statistics (effect estimates and standard errors), combined with LD information from a suitable reference panel to derive marginal likelihoods or Bayes factors.

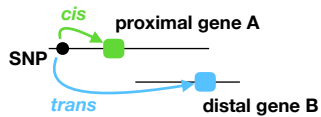
Hierarchical regression for multiple responses

Molecular mechanisms: from the genotype to the phenotype

- A variety of **molecular mechanisms mediate the action of the genotype on the phenotype** (trait of interest).
- Genetic variants can regulate gene expression (the transcriptome).
- Changes at the transcript level can have downstream effects on proteins (the proteome) or metabolites (the metabolome).
- These regulatory effects are often subtle, involve pathway-level interactions and can be specific to particular tissues or cell types.
- Molecular readouts (e.g., gene, protein or metabolite levels) are often referred to as **endophenotypes**, because they serve as **intermediate molecular proxies for the phenotype**.
- We aim to understand **not just whether** a genetic variant is associated with disease, **but how it operates biologically** by studying its effect on molecular traits.

Molecular QTL studies

- Molecular QTL (quantitative trait locus) studies aim to identify genetic variants that influence molecular traits, e.g.:
  - **eQTLs**: expression QTLs – SNPs affecting gene expression.
  - **pQTLs**: protein QTLs – SNPs affecting protein abundance.
  - **mQTLs**: methylation QTLs – SNPs influencing methylation levels.
- Genetic effects can be:
  - ***cis***: the SNP regulates a nearby gene product (e.g., within 1 Mb).
  - ***trans***: the SNP regulates a distant gene product, possibly even on another chromosome.

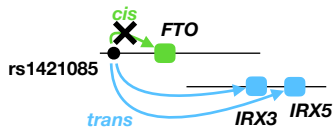


- A single SNP may influence multiple traits, a phenomenon known as **pleiotropy**.
- **Hotspot SNPs** are pleiotropic SNPs associated with large numbers of gene products in *cis* and/or *trans*.
- Univariate screening methods (one pair of SNP/trait at a time) do not account for shared genetic architecture.

From GWAS hits to mechanisms: revisiting *FTO*

Recall the *FTO* example from earlier:

- GWAS linked SNPs in the *FTO* locus to BMI and obesity risk;
- Initially thought to act "in *cis*", through the nearby *FTO* gene;
- But functional studies revealed a **distal regulatory mechanism**: SNP **rs1421085** alters a regulatory element "in *trans*" that modulates expression of *IRX3* and *IRX5* via long-range chromatin looping (Smemo et al., 2014);
- *IRX3* and *IRX5* might be regarded as potential targets for obesity treatment, not *FTO*.



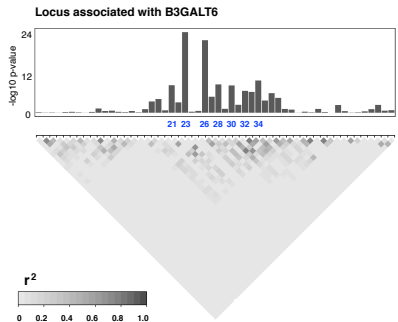
Studies have found that **drugs whose target is supported by eQTL mapping are 2-4 times more likely to be successful** (Sadler et al., 2023).

Pairwise eQTL screening

- Example dataset: CD14$^+$ monocytes from 432 healthy European individuals, with $>$ 24 400 gene transcripts (traits) and $>$ 380 000 SNPs.
- A marginal pairwise screening on 29 607 SNPs from chromosome 1 shows:
  - About 2.5 times more *cis* associations than *trans* associations;
  - Many *cis* associations are probably redundant due to LD;
  - *Trans* effects tend to be weaker than *cis* effects (general fact).

	Number	After LD pruning	Effect magnitude
Cis associations	1 611	1 049	0.11 (0.10)
Trans associations	655	641	0.04 (0.03)

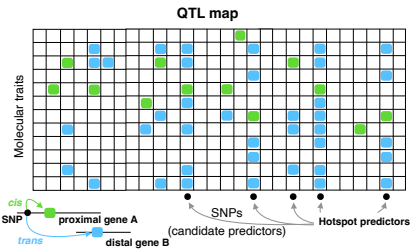
Table: Summary of *cis* and *trans* associations at FDR 20% (LD pruning genetic $r^2 > 0.5$ and window size 2 Mb).

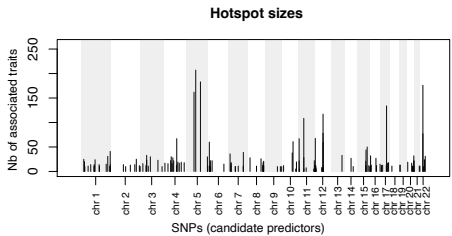


LD and Manhattan plots for *cis* associations with gene *B3GALT6* (involved in the transfer galactose).

Beyond marginal screening: the need for joint models

- **Most *trans* associations** identified by marginal screening **fail to survive multiplicity correction**: they are largely masked by the large number of redundant (and stronger) *cis* signals due to LD.
- Detecting weaker but biologically important *trans* effects (including hotspots) requires models that account for all SNPs and all traits jointly.
- **Borrowing information across traits** under shared genetic control can substantially **improve power**, especially for *trans*-acting effects.
- This (again!) highlights the need for **interpretable** multivariate methods for high-dimensional QTL analysis.





Sparse multivariate regression model

Consider the canonical sparse multivariate regression model for $\mathcal{M}_\gamma$:

$$\boldsymbol{Y} = \boldsymbol{X}_\gamma \boldsymbol{B}_\gamma + \boldsymbol{E}, \qquad \boldsymbol{E} \sim \mathcal{MN}_{n \times q}\left(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{\Sigma}\right), \tag{20}$$

where

- $\boldsymbol{Y}$ is a $n \times q$ response matrix (molecular traits);
- $\boldsymbol{X}_\gamma$ is a $n \times p_\gamma$ matrix of selected predictors (SNPs), where $p_\gamma = \sum_{s=1}^{p} \gamma_s$;
- $\boldsymbol{B}_\gamma$ is a $p_\gamma \times q$ matrix of regression coefficients;
- $\boldsymbol{E}$ is a $n \times q$ matrix of error terms, assigned a matrix-variate normal distribution (Dawid, 1981) with independent rows and covariance $\boldsymbol{\Sigma}$ across traits.

Interpretable multivariate inference

Need for dual selection:

- Model (20) represents pairwise SNP-trait associations via $\boldsymbol{B}$, but selection is still row-wise via the $p \times 1$ binary indicator vector $\boldsymbol{\gamma}$.
- This assumes that a SNP is **either associated with all traits or none**, which may be a reasonable simplification in specific multivariate GWAS settings (e.g., shared genetic basis for cholesterol, lipid traits and blood pressure).
- In molecular QTL studies however, each SNP typically regulates a few traits, requiring *within-row* sparsity in $\boldsymbol{B}_{\boldsymbol{\gamma}}$.

Need for scalability in high-dimensional response settings:

- Molecular QTL studies: q (traits) in tens of thousands, p (SNPs) in millions.
- Estimating an (unstructured) trait residual covariance $\boldsymbol{\Sigma}$ in such settings is **prohibitive due to high memory and runtime cost.**
- Existing multivariate models typically restrict q and assume conjugacy to integrate $\boldsymbol{\Sigma}$ out (e.g., Petretto et al., 2010; Lewin et al., 2015).

$\longrightarrow$ Flexible multivariate modelling of molecular QTL data is out of reach without structural simplifications.

A hierarchical model for molecular QTL mapping

One alternative: **borrow strength via** the spike-and-slab **model hierarchy only.**

For traits $t = 1, \ldots, q$, consider a **series of conditionally independent regressions:**

$$\boldsymbol{y}_t = \boldsymbol{X}\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}_n\left(\mathbf{0}, \tau_t^{-1}\boldsymbol{I}_n\right), \qquad \tau_t \sim \text{Gamma}\big(\eta_t, \kappa_t\big), \qquad t = 1, \ldots, q, \tag{21}$$

where

- $\boldsymbol{y}_t$ is a $n \times 1$ response vector (molecular traits, typically $q = 10^2$–10^4),
- $\boldsymbol{X}$ is a $n \times p$ matrix of candidate predictors (SNPs, typically, $p = 10^5$–10^6),
- $\boldsymbol{\beta}_t$ is a $p \times 1$ vector of regression coefficient,

and place a pairwise spike-and-slab prior on the regression coefficients:

$$\beta_{st} \mid \gamma_{st}, \sigma^2, \tau_t \sim \gamma_{st}\,\mathcal{N}\big(0, \sigma^2\tau_t^{-1}\big) + \big(1 - \gamma_{st}\big)\,\delta_0, \qquad \gamma_{st} \sim \text{Bernoulli}\big(\omega_s\big), \tag{22}$$

with

$$\omega_s \sim \text{Beta}\big(a_s, b_s\big), \qquad \sigma^{-2} \sim \text{Gamma}\big(\lambda, \nu\big). \tag{23}$$

Why this hierarchical model?

Model (21)–(23) takes a different path from canonical multivariate approaches:

- It replaces the $p \times 1$ binary indicator vector $\boldsymbol{\gamma}$ with a matrix $\boldsymbol{\gamma} = \{\gamma_{st}\}$, which enables **direct selection of predictor-response pairs**, via the marginal posterior inclusion probabilities $p(\gamma_{st} = 1 \mid \boldsymbol{y})$:
- It **does not model residual covariance** explicitly, which avoids the curse of dimensionality in $q \gg n$ settings;
- It introduces **dependence across responses** *through the prior* on effect inclusion, via shared ω_s and σ; for molecular QTL mapping this allows **information sharing across traits *under shared genetic control***.
- In particular, the model directly parametrises **pleiotropic effects** via ω_s, so that $\mathbb{E}(\omega_s \mid \boldsymbol{y})$ can be used to select **hotspot predictors** controlling multiple responses.

Prior-induced sparsity and multiplicity control

Sparsity in the association pattern is controlled via the prior on ω_s.

- The prior odds penalty representing the support for a model to have an additional response associated with a given predictor $\boldsymbol{X}_s$ is

$$\text{PO}(q_s - 1 : q_s) = \frac{\text{pr}(\mathcal{M}_{q_s-1})}{\text{pr}(\mathcal{M}_{q_s})} = \frac{b_s + q - q_s}{a_s + q_s - 1},$$

  where $\mathcal{M}_{q_s}$ now denotes a model in which $\boldsymbol{X}_s$ is associated with $1 \leq q_s \leq q$ responses.

- Penalty increases with the number of responses q but **no inherent correction for the predictor dimension** p.

- This is because ω_s is predictor-specific so we have separability across predictors, unlike in the example discussed in the context of single-response spike-and-slab regression.

p	50	250	500	1,000	2,500
Mean # FP					
Uncorrected	1.06	6.70	16.52	35.22	73.55
Mean # TP					
Uncorrected	9.67	9.69	9.72	9.77	9.80

For instance, choosing $a_s \equiv 1$, $b_s \equiv 2q - 1$, so $\mathbb{E}(\omega_s) \equiv (2q)^{-1}$ (prior mean number of responses associated with $\boldsymbol{X}_s$ is 0.5 independently of p) leads to a linear increase of false positives as p grows (using the MPM selection rule, PPI > 0.5).

Prior-induced sparsity and multiplicity control

- To address this, we control the probability of any association between $\boldsymbol{X}_s$ and responses:

$$\text{pr}\left(\cup_{t=1}^{q}\{\gamma_{st}=1\}\right) = 1 - \prod_{t=1}^{q} \frac{b_s + q - t}{a_s + b_s + q - t},$$

  setting it to p_0/p, where p_0 is a prior guess of the number of predictors included in the model.

- This can be achieved by setting (assuming exchangeability):

$$a_s \equiv 1\,, \qquad b_s \equiv \frac{q(p-p_0)}{p_0}\,, \qquad 0 < p_0 < p.$$

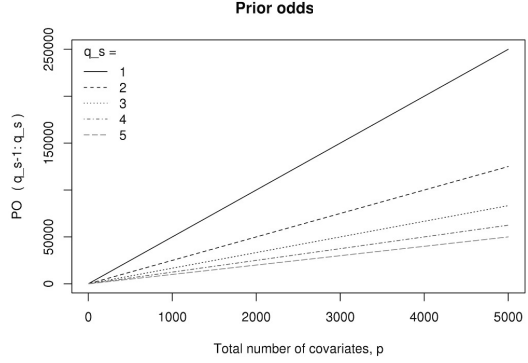
- The model now adjusts for multiplicity in terms of numbers of both candidate predictors p and responses q.

p	50	250	500	1,000	2,500
Mean # FP					
Uncorrected	1.06	6.70	16.52	35.22	73.55
Corrected	0.39	0.30	0.36	0.34	0.31
Mean # TP					
Uncorrected	9.67	9.69	9.72	9.77	9.80
Corrected	9.69	9.42	9.20	9.25	8.98

The number of false positives remains roughly constant and close to zero with correction, small cost on the number of true positives.

Prior-induced sparsity and multiplicity control

- Similarly to the single-response spike-and-slab model, there is also a relation between the **penalty on the model complexity** and the current model size.
- Here the penalty is not uniform depending on the number of *responses* associated with a given predictor $\boldsymbol{X}_s$: moving from zero to one response associated with $\boldsymbol{X}_s$ is harder than moving from nine to ten.
- **Ensures a sufficient regularisation on the "hotspot sizes"**, especially important in weakly informative, large-q settings.



Alternative priors for inclusion probabilities

Recall: The predictor-specific prior on the spike-and-slab inclusion probability in model (21)–(23): $\omega_s \sim \text{Beta}(a_s, b_s)$;

- Controls the extent to which predictor s is associated with lots of traits ("hotspot propensity");
- Allows sparsity control through a prior on the total number of predictors entering the model.

Alternative formulations have been proposed for the inclusion probabilities:

Top-level spike-and-slab prior (Scott-Boyer et al., 2012):

$$\omega_s \mid \pi_s \sim \pi_s \text{Beta}(a_s, b_s) + (1-\pi_s)\delta_0, \quad \pi_s \sim \text{Beta}(a_0, b_0);$$

- Allows some predictors to have zero probability of inclusion across all traits.

"Independent" formulation:

$$\omega_t \sim \text{Beta}(a_t, b_t);$$

- Implements a trait-specific probability of inclusion, direct extension of single-trait models (no sharing across traits).

Alternative priors for inclusion probabilities

Multiplicative formulation (Richardson et al., 2010):

$$\omega_{st} = \omega_s \rho_t, \quad \rho_t \sim \text{Beta}(a_t, b_t), \quad \omega_s \sim \text{Gamma}(c_s, d_s);$$

- ω_s: predictor-specific hotspot propensity (similarly as before) and ρ_t: trait-specific modulation.

Hotspot-tailored formulation (Ruffieux et al., 2020):

$$\omega_{st} = \Phi(\theta_s + \zeta_t), \quad \zeta_t \sim \mathcal{N}(n_0, t_0^2), \quad \theta_s \mid \lambda_s, \sigma_0 \sim \mathcal{N}(0, \sigma_0^2 \lambda_s^2), \quad \lambda_s \sim \text{C}^+(0,1), \quad \sigma_0 \sim \text{C}^+(0, q^{-1/2}),$$

where $\Phi(\cdot)$ is the standard normal CDF and $\text{C}^+(\cdot,\cdot)$ is a half-Cauchy distribution.

- θ_s: predictor-specific hotspot propensity and ζ_t: trait-specific modulation;
- Horseshoe prior on θ_s provides local and global shrinkage on the propensity of predictors to be involved in associations: global scale σ_0 adapts to overall sparsity while local scales λ_s allow flexible deviations (heavy tails).

Alternative formulations

- These models allow increased flexibility, but beware: top-level parameters may not be well-informed if data are weakly informative.
- Similar shrinkage structures could be considered using one-group priors (e.g., the horseshoe) directly on β_{st}.
- **Important caveat:** By design, the above formulations capture dependence solely via the spike-and-slab model hierarchy: they assume a diagonal residual covariance $\boldsymbol{\Sigma}$ (independent errors).
- This can be deleterious when residual correlations reflect meaningful biological or technical structure that the model fails to account for: can bias effect estimates, miscalibrate uncertainty, and either reduce power or inflate false positives.
- Multivariate SSL (mSSL; Deshpande et al. 2019) estimates trait residual covariance jointly with sparse β_{st}, but at higher computational cost (not applicable at the scales typically encountered in molecular QTL studies).

GTEx Portal

Home Downloads Expression Single Cell QTL IGV Browser Tissues & Histology Documentation About

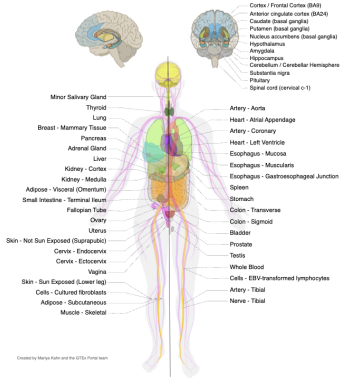
Adipose Subcutaneous

Data Source: GTEx Analysis Release V10 (dbGaP Accession phs000424.v10.p2)

UBERON UBERON:0002190

Main Sampling Site

Subcutaneous tissue beneath the leg's skin sample.



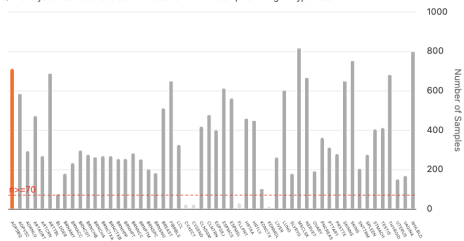
Bulk Tissue RNA-Seq Sample Info

Table: Sample count in Adipose - Subcutaneous

Samples	All	Female	Male
Total	714	233	481
With Genotype	711	232	479

Chart: GTEx samples by tissue

QTL analysis results are available for tissues with >70 samples with genotype data



A flexible hierarchical model for cross-tissue eQTL effects

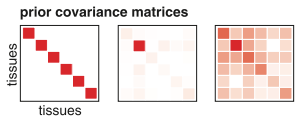
Understand how a given gene is regulated by SNPs across multiple tissues:

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E}, \qquad \boldsymbol{E} \sim \mathcal{MN}_{n \times r}(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{\Sigma}), \tag{24}$$

- $\boldsymbol{Y}$: $n \times r$ matrix of expression for a given gene across r tissues;
- $\boldsymbol{X}$: $n \times p$ matrix of *cis*-SNP genotypes;
- $\boldsymbol{B}$: $p \times r$ matrix of SNP effects across tissues (typically $r = 1 - 100$);
- $\boldsymbol{\Sigma}$: residual covariance across tissues.

Prior for effects of SNP s across the r tissues (mixture of normals; Morgante et al., 2023):

$$\boldsymbol{b}_s \sim \sum_{k=1}^{K} w_k \, \mathcal{N}_r(\boldsymbol{0}, \boldsymbol{S}_k)$$



- $\{\boldsymbol{S}_k\}_{k=1}^{K}$: covariance matrices encoding plausible patterns of tissue sharing;
- $\boldsymbol{w} = (w_1, \ldots, w_K)$: mixture weights learned via empirical Bayes.

Goal: Learn which patterns of effect sharing are supported by the data and exploit them to improve estimation.

Structured priors for cross-tissue effects

Specifically, the prior distribution over SNP effects is defined as a combination of scaled, covariance matrices:

$$\boldsymbol{b}_s \sim w_{0,0}\boldsymbol{\delta}_0 + \sum_{l=1}^{L}\sum_{t=1}^{T} w_{0,l,t}\,\mathcal{N}_r\left(\mathbf{0}, \omega_l^2 \boldsymbol{U}_{0,t}\right),$$

- $\boldsymbol{\delta}_0$: point mass at zero (spike) to induce sparsity;
- $\boldsymbol{U}_{0,t}$: fixed normalised covariance matrices (largest diagonal entry is 1);
- ω_l: scaling factors, log-spaced to cover a wide effect size range;
- $w_{0,l,t}$: mixture weights estimated from the data (empirical Bayes).

Effect sharing patterns – specification of covariance matrices $\boldsymbol{U}_{0,t}$:

- **Canonical**: e.g., identity (tissue-specific), equal effects (shared), rank-1 (single-tissue) or sparse block structures (subset sharing).
- **Data-driven**: estimated once across genes from summary statistics (marginal screening).

GTEx illustration: Bayesian vs. elastic net

Predict gene expression across 48 tissues using GTEx data (for 1 000 randomly chosen genes).

- Bayesian model **naturally handles missingness** (60%) by imputing $\boldsymbol{Y}_{\text{miss}}$ at each step using current estimates of effects and residual covariance;
- Benchmark: elastic net (frequentist, $\ell_1 + \ell_2$ penalised regression) with 5-fold cross-validation;
- Accuracy in each tissue quantified using root mean squared errors (RMSE):

$$\text{RMSE}(m) = \sqrt{\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left(y_{im} - \hat{y}_{im}\right)^2},$$

where y_{im} is the observed expression in tissue m for test sample i, $\hat{y}_{im}$ is the predicted expression and n_{test} is the test set size. To improve comparability across tissues, RMSEs are standardised by the standard deviation of y_{im} in the test set.

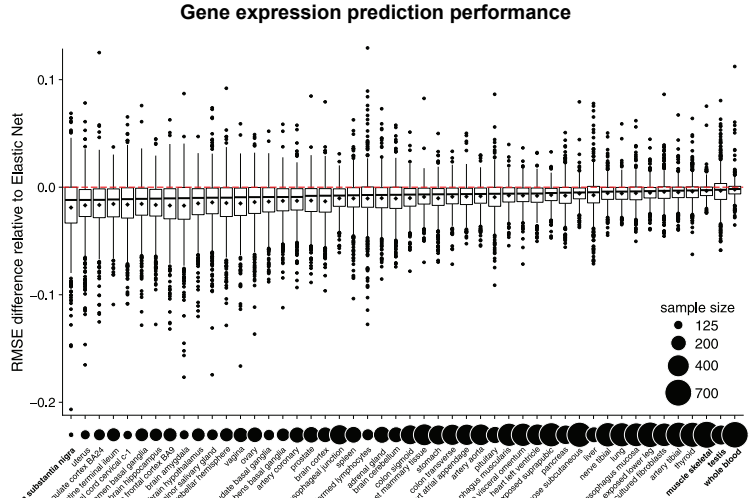
GTEx illustration: Bayesian vs. elastic net

Bayesian model outperforms elastic net in most tissues.

- **Strongest gains are in low-sample tissues** or those **with shared eQTL effects** (e.g., among the brain tissue);
- Smaller gains are in large-sample, tissue-specific settings (e.g., muscle, testis).

Relative RMSE difference:

$$\frac{\text{RMSE}_{\text{Bayesian}} - \text{RMSE}_{\text{elastic net}}}{\text{RMSE}_{\text{elastic net}}}$$



Further directions

Integrating QTL mapping and GWAS signals

Broad motivation: Understand the molecular mechanisms linking SNPs to disease phenotypes.

- **Colocalisation** (Giambartolomei et al., 2014):
  Goal: assess whether a QTL and a GWAS signal in the same region are likely to share a causal SNP.
  - Enables functional interpretation of GWAS loci and prioritisation of putative target genes.
- **Mendelian randomisation** (MR; Davey Smith and Ebrahim, 2003):
  Goal: use SNPs as instruments to assess whether a biomarker (such as a gene or protein) causally affects an outcome (trait or disease).
  - Supports causal inference in observational studies by using SNPs as natural experiments.
- **Phenome-wide association studies** (PheWAS; Bush et al., 2016):
  Goal: scan many phenotypes (from EHRs, registries) for association with a given SNP or risk score.
  - Identifies pleiotropic effects across diverse phenotypes.
  - Detects opportunities for drug repurposing through shared genetic architecture.
  - Supports hypothesis generation about gene function or comorbidities.

New directions in biomedical genomics

- **Emerging modalities:** Single-cell genomics, spatial transcriptomics, diagnostic imaging, longitudinal designs – all require richer datasets with matched genotypes.
- **Beyond traditional phenotypes:** Leverage ML/AI to derive biologically relevant latent traits (e.g., brain connectivity, immune profiles, mental health signatures).
- **Knowledge-aware AI:** Incorporate biological constraints into predictive models – e.g., structured priors, Bayesian neural networks, constraint-based loss functions.

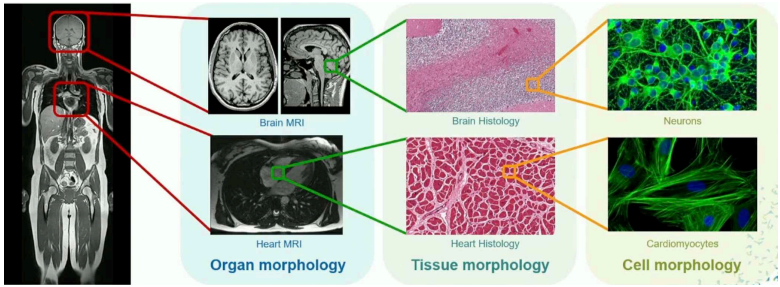
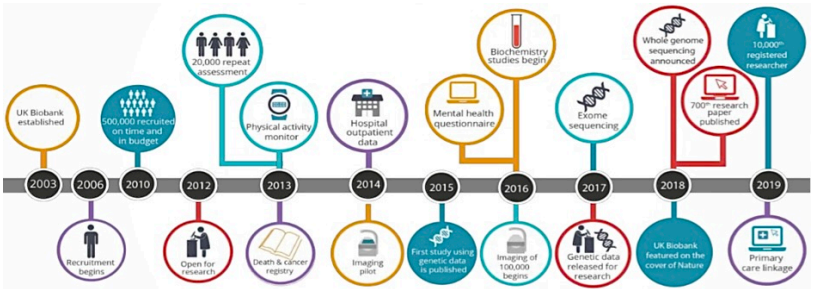


Image: Human Technopole.

New directions in biomedical genomics (2)

- **Big resources and biases:** Biobanks (UKBB, FinnGen, etc.) enable broad studies, but raise concerns about ancestry representation, selection bias and portability of findings.
- **Scalability and privacy:** Widespread use of summary statistics, emerging use federated learning and synthetic data.
- **From association to function:** CRISPR screens and perturbation assays enable functional validation of genetic findings at scale.



UK Biobank overview from its creation to 2019. Image: UKBB.

Summary: Bayesian hierarchical models – beyond genomics

- **Interpretability and generative insights:**
  - Reflect nested variation and generative processes $\rightarrow$ conceptual clarity and alignment with domain knowledge.
  - Especially suited for scientific inference and mechanism-based reasoning.
- **Information borrowing and adaptive shrinkage:**
  - Partial pooling enables principled information sharing across units (non-separability).
  - Regularisation arises naturally via priors $\rightarrow$ stabilises inference under limited data or high dimensionality.
- **Flexibility and modularity:**
  - Models can often be seamlessly extended to multi-level, spatio-temporal or latent structures.
  - Structured priors and co-data can be integrated.
- **Uncertainty propagation:**
  - Posteriors at all levels enable coherent interval estimation.
  - Averaging over plausible configurations helps in collinear and missing-data settings.

Thank you for your attention

Upcoming: 22nd Armitage Workshop and Lecture, 23 October 2025, Cambridge.

Topic: “Integration of data from multiple domains”. Keynote: Prof. Matthew Stephens, University of Chicago.

Registration (in-person/online):

https://www.mrc-bsu.cam.ac.uk/events/22nd-armitage-workshop-and-lecture

References I

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, 32:870–897.

Berger, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *The Annals of Statistics*, 8:716–761.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, 12:1105–1131.

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169:1177–1186.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer, Berlin, Germany.

Burgess, S., Timpson, N. J., Ebrahim, S., and Davey Smith, G. (2015). Mendelian randomization: where are we now and where are we going?

Busatto, C. and van de Wiel, M. (2023). Informative co-data learning for high-dimensional horseshoe regression. *arXiv preprint arXiv:2303.05898*.

Bush, W. S., Oetjens, M. T., and Crawford, D. C. (2016). Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nature Reviews Genetics*, 17:129–145.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In van Dyk, D. and Welling, M., editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, pages 73–80, Clearwater Beach, United States. Proceedings of Machine Learning Research.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97:465–480.

Chaudhary, R., Garg, J., Shah, N., and Sumner, A. (2017). PCSK9 inhibitors: a new era of lipid lowering therapy. *World Journal of Cardiology*, 9:76.

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24:17–36.

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, 20:80–101.

Davey Smith, G. and Ebrahim, S. (2003). Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32:1–22.

Dawid, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68:265–274.

De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68.

Deshpande, S. K., Ročková, V., and George, E. I. (2019). Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *Journal of Computational and Graphical Statistics*, 28:921–931.

Donoho, D. L., Johnstone, I. M., Hoch, J. C., and Stern, A. S. (1992). Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 41–81.

Doob, J. L. (1949). Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications*, pages 23–27, Paris. Colloques Internationaux du Centre National de la Recherche Scientifique.

References II

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102:93–103.

Efron, B., Tibshirani, R. J., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881–889.

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, 10:e1004383.

Griffin, J. and Brown, P. (2017). Hierarchical shrinkage priors for regression models. *Bayesian Analysis*, 12:135–159.

Guan, Y. and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Annals of Applied Statistics*, 5:1780–1815.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96:835–845.

He, V. X. and Wand, M. P. (2024). Bayesian generalized additive model selection including a fast variational option. *AStA Advances in Statistical Analysis*, 108:639–668.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, 33:730–773.

James, W. and Stein, C. M. (1961). Estimation with quadratic loss. In Neyman, J., editor, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1*, pages 361–379, Berkeley, United States. University of California Press.

Johnstone, I. M. (1994). On minimax estimation of a sparse normal mean vector. *The Annals of Statistics*, pages 271–289.

Keaton, J. M., Kamali, Z., Xie, T., Vaez, A., Williams, A., Goleva, S. B., Ani, A., Evangelou, E., Hellwege, J. N., Yengo, L., et al. (2024). Genome-wide analysis in over 1 million individuals of European ancestry yields improved polygenic risk scores for blood pressure traits. *Nature Genetics*, 56:778–791.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5:369–411.

Lewin, A., Saadi, H., Peters, J. E., Moreno-Moral, A., Lee, J. C., Smith, K. G. C., Petretto, E., Bottolo, L., and Richardson, S. (2015). MT-HESS: an efficient Bayesian approach for simultaneous association detection in OMICS datasets, with application to eQTL mapping in multiple tissues. *Bioinformatics*, 32:523–532.

Loos, R. J. F. and Yeo, G. S. H. (2014). The bigger picture of FTO—the first GWAS-identified obesity gene. *Nature Reviews Endocrinology*, 10:51–61.

Maher, B. (2008). Personal genomes: The case of the missing heritability.

Minikel, E. V., Painter, J. L., Dong, C. C., and Nelson, M. R. (2024). Refining the impact of genetic evidence on clinical success. *Nature*, 629:624–629.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–1032.

References III

Morgante, F., Carbonetto, P., Wang, G., Zou, Y., Sarkar, A., and Stephens, M. (2023). A flexible empirical bayes approach to multivariate multiple regression, and its improved accuracy in predicting multi-tissue gene expression from genotypes. *PLoS Genetics*, 19:e1010539.

Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P. C., Li, M. J., Wang, J., et al. (2015). The support of human genetic evidence for approved drug indications. *Nature Genetics*, 47:856–860.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5:155–176.

Ochoa, D., Karim, M., Ghoussaini, M., Hulcoop, D. G., McDonagh, E. M., and Dunham, I. (2022). Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. *Nature Reviews Drug Discovery*, 21:551.

Park, J.-H., Gail, M. H., Weinberg, C. R., Carroll, R. J., Chung, C. C., Wang, Z., Chanock, S. J., Fraumeni, J. F., and Chatterjee, N. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences*, 108:18026–18031.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103:681–686.

Petretto, E., Bottolo, L., Langley, S. R., Heinig, M., McDermott-Roe, C., Sarwar, R., Pravenec, M., Hübner, N., Aitman, T. J., and Cook, S. A. (2010). New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS Computational Biology*, 6:e1000737.

Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11:5018–5051.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., editors, *Bayesian Statistics*, volume 9, pages 501–538. Oxford University Press, New York, United States.

Richardson, S., Bottolo, L., and Rosenthal, J. S. (2010). Bayesian models for sparse regression analysis of high-dimensional data. In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., editors, *Bayesian Statistics*, volume 9, pages 539–569. Oxford University Press, New York, United States.

Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, 113:431–444.

Ruffieux, H., Davison, A. C., Hager, J., Inshaw, J., Fairfax, B., Richardson, S., and Bottolo, L. (2020). A global-local approach for detecting hotspots in multiple response regression. *The Annals of Applied Statistics*, 14:905–928.

Sadler, M. C., Auwerx, C., Deelen, P., and Kutalik, Z. (2023). Multi-layered genetic approaches to identify approved drug targets. *Cell Genomics*, 3.

Scannell, J. W., Blanckley, A., Boldon, H., and Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*, 11:191–200.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4:10–26.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38:2587–2619.

References IV

Scott-Boyer, M. P., Imholte, G. C., Tayeb, A., Labbe, A., Deschepper, C. F., and Gottardo, R. (2012). An integrated hierarchical Bayesian model for multivariate eQTL mapping. *Statistical Applications in Genetics and Molecular Biology*, 11:1515–1544.

Smemo, S., Tena, J. J., Kim, K.-H., Gamazon, E. R., Sakabe, N. J., Gómez-Marín, C., Aneas, I., Credidio, F. L., Sobreira, D. R., and Wasserman, N. F. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, 507:371.

Stein, C. M. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Technical report, Stanford University, United States.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 1135–1151.

Strausberg, R. L. (2001). Talkin' Omics. *Disease Markers*, 17:39.

Strawderman, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics*, 42:385–388.

Sun, D., Gao, W., Hu, H., and Zhou, S. (2022). Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B*, 12:3049–3062.

te Beest, D. E., Mes, S. W., Wilting, S. M., Brakenhoff, R. H., and van de Wiel, M. A. (2017). Improved high-dimensional prediction with Random Forests by the use of co-data. *BMC Bioinformatics*, 18:584.

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

van de Wiel, M. A., Te Beest, D. E., and Münch, M. M. (2018). Learning from a lot: Empirical Bayes for high-dimensional model-based prediction. *Scandinavian Journal of Statistics*.

Xu, Z., Schmidt, D. F., Makalic, E., Qian, G., and Hopper, J. L. (2016). Bayesian grouped horseshoe regression with application to additive models. In *Australasian Joint Conference on Artificial Intelligence*, pages 229–240. Springer.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42:565–569.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68:49–67.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Studies in Bayesian Econometrics*, volume 6, pages 233–243. Elsevier, New York, United States. P. K. Goel and A. Zellner, editors.