Gene Set Analyses using Bayesian Linear Regression Models

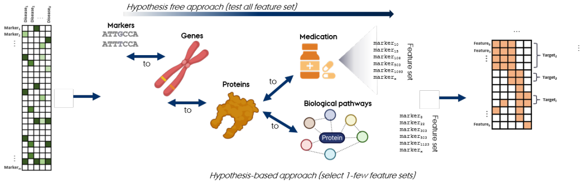
Peter Sørensen

From GWAS to Biological Discovery

BLR Adjusted
GWAS summary statistics

Link genetic markers to
information in biological
databases

Functional marker sets
enriched for disease
association



“*Utilising cleaner and stronger marker signals linked to functional marker information and jointly analysing multiple functional marker sets and diseases may help better understand disease biology and subsequently be used to identify novel drug targets.*”

Gene Set Analysis

Gene set analysis evaluates the **coordinated action of genes or sets of variants** within predefined biological pathways or functional groups.

GWAS identify single genetic variants (SNPs) associated with traits or diseases.

- **Many variants have small individual effects**
  → Use larger datasets or make better use of existing data.
- **Some effects are clustered within functionally related genes or pathways**
  → Use prior information on functional marker groups to improve detection power and interpretation.
- **Some effects are shared across multiple traits**
  → Leverage correlated trait information to enhance detection power and prediction accuracy.

Gene Set Analyses

Many different gene set analysis approaches have been proposed.

- **MAGMA**: Multi-marker Analysis of GenoMic Annotation (Leuww et al 2015)
  - generalized gene set analysis of GWAS data
  - based on a multiple regression model
  - rank gene sets
- **PoPS**: Polygenic Prioritisation Scoring (Weeks et al 2023)
  - leveraging polygenic enrichment of gene features (e.g. gene sets) to predict genes underlying complex diseases
  - based on a multiple regression model
  - rank genes

MAGMA: Linear Model Approach

MAGMA fits a **linear regression model** to test associations between **gene sets** and traits.

1. Aggregate **SNP-level GWAS statistics** into **gene-level statistics**, accounting for LD.
2. Use gene-level statistics as the **response variable**.
3. Represent gene sets as a **predictor matrix**, typically indicating gene membership, but not necessarily limited to binary values.
4. Estimate regression coefficients to assess the **strength of association** between each gene set and the trait.
5. Evaluate significance using **permutation** or **model-based null distributions**.

MAGMA: Limitations

When analyzing thousands of gene sets, several issues arise:
- **Overfitting** – the model may capture noise rather than true signals.
- **Multicollinearity** – many gene sets are correlated due to biological overlap.
- **Multiple testing** – increases false-positive risk.
- **Interpretation difficulty** – hard to disentangle contributions of overlapping sets.

$\rightarrow$ Use of **regularization** and **variable selection** to improve model robustness and interpretability.

Additionally, many complex traits are **genetically correlated**, sharing overlapping biological pathways.

$\rightarrow$ Incorporating **multi-trait information** in MAGMA can increase detection power and reveal shared genetic mechanisms across traits.

Bayesian MAGMA: Idea

Develop and evaluate a **Bayesian gene-set prioritization approach** using BLR within the MAGMA framework.

- Advantages:
  - Incorporates **regularization** and **variable selection** via **spike-and-slab priors**.
  - Controls false positives and **handles correlated gene sets**.
  - Provides **posterior inclusion probabilities (PIP)** as evidence of gene set association.
- Flexible framework supporting:
  - **Single- and multi-trait models**
  - **Integration of diverse genomic features**
  - **Modeling of correlated traits** to uncover shared genetic factors.

Bayesian MAGMA: Regression Model

The **Bayesian MAGMA** framework builds on the standard regression model:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$$

- $\mathbf{Y}$ — vector of **observed outcomes** or **gene-/set-level association measures**
- $\mathbf{X}$ — matrix of **genomic predictors** (e.g., gene membership, functional annotations, or pathway indicators)
- β — vector of **effect sizes** describing how predictors in $\mathbf{X}$ explain variation in $\mathbf{Y}$
- ε — **residual noise** capturing unexplained variation

In the **Bayesian formulation**, each β_j is assigned a **prior distribution** that encodes assumptions about **effect size magnitude, sparsity, or functional grouping**.
These priors enable **regularization**, **variable selection**, and **information sharing** across correlated features or biological layers.

Bayesian MAGMA: Multivariate Motivation

- Traditional single-trait analyses may miss associations that are **weak individually but consistent across traits**.
- **Multi-trait Bayesian MAGMA** leverages these correlations by jointly modeling multiple traits to:
  - Increase **power** for detecting gene sets and pathways,
  - Improve **accuracy** of effect estimation, and
  - Reveal **shared biological mechanisms** underlying complex diseases.

Bayesian MAGMA: Multivariate Regression Model

In the **multivariate BLR** model, we model **multiple correlated outcomes** jointly:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

- $\mathbf{Y}$: $(n \times T)$ matrix of outcomes
  (e.g., association measures for T traits or omic layers)
- $\mathbf{X}$: $(n \times p)$ feature matrix
- $\mathbf{B}$: $(p \times T)$ matrix of effect sizes
- $\mathbf{E}$: $(n \times T)$ residual matrix

Each row of $\mathbf{Y}$ corresponds to an observation or gene, and each column to a trait, phenotype, or molecular layer.

Bayesian MAGMA: Error and Effect Priors

We extend the univariate priors to the multivariate setting:

$$\mathbf{e}_{i\cdot} \sim \mathcal{N}_T(\mathbf{0}, \Sigma_e)$$
$$\mathbf{b}_j \sim \mathcal{N}_T(\mathbf{0}, \Sigma_b)$$

- Σ_e: residual covariance among traits
- Σ_b: covariance of effect sizes across traits
- When Σ_e and Σ_b are diagonal, the model reduces to T independent univariate BLR models.

Allows **information sharing across correlated traits or omic layers** and can be used to identify **pleiotropic effects** and **cross-trait genetic architectures**.

Bayesian MAGMA: Estimation of Effects

Ordinary multiple regression

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

With regularization

$$\mathbf{b} = \left(\mathbf{X}'\mathbf{X} + \mathbf{I}\frac{\sigma_e^2}{\sigma_b^2}\right)^{-1}\mathbf{X}'\mathbf{y}$$

Using information from multiple responses

$$\mathbf{b} = \left(\mathbf{X}'\mathbf{X} + \mathbf{I} \otimes \Sigma_B^{-1}\Sigma_E\right)^{-1}\mathbf{X}'\mathbf{y}$$

Bayesian MAGMA: Indicator Variables and PIPs

In the **multivariate BLR**, each feature j may affect multiple outcomes (traits).
We extend the indicator variable to capture **cross-trait activity patterns**:

$$\delta_j = \begin{bmatrix} \delta_{j1} \\ \delta_{j2} \\ \vdots \\ \delta_{jT} \end{bmatrix}, \qquad \delta_{jt} = \begin{cases} 1, & \text{if feature } j \text{ affects trait } t \\ 0, & \text{otherwise.} \end{cases}$$

After inference, we estimate $\text{PIP}_{jt} = P(\delta_{jt} = 1 \mid \text{data})$ — the **posterior inclusion probability** that feature j affects trait t.

Bayesian MAGMA: Posterior Parameters

In the multivariate setting, we generalize each posterior quantity:

Parameter	Interpretation
$\mathbf{B} = [\beta_{jt}]$	Effect matrix across traits (j: feature, t: trait)
PIP$_j$	Probability that feature j affects 1 trait
Σ_b	Covariance of effects across traits
Σ_e	Residual covariance among traits

These allow us to identify:

- **Shared genetic effects** (pleiotropy)
- **Trait-specific vs. shared signals**
- **Cross-trait enrichment** of biological sets

Bayesian MAGMA. Study Aim and Design

Evaluate a **Bayesian gene-set prioritization approach** using BLR within the MAGMA framework.

Simulation study:

- Assessed model performance under varying gene set characteristics and genetic architectures.
- Used **UK Biobank genetic data** for realistic evaluation.

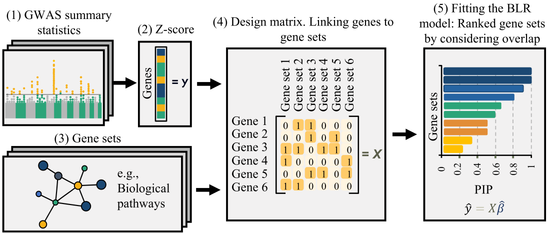
Comparative analysis:

- Benchmarked Bayesian MAGMA against the standard MAGMA approach.

Applications:

- Applied to **nine complex traits** using publicly available GWAS data.
- Developed a **multi-trait BLR model** to integrate GWAS results across traits and uncover shared genetic architecture.

Bayesian MAGMA: Overview



(1) GWAS summary statistics

(2) Z-score

(3) Gene sets

e.g., Biological pathways

(4) Design matrix. Linking genes to gene sets

$$\begin{array}{c|cccccc} & \text{Gene set 1} & \text{Gene set 2} & \text{Gene set 3} & \text{Gene set 4} & \text{Gene set 5} & \text{Gene set 6} \\ \hline \text{Gene 1} & 0 & 1 & 1 & 0 & 0 & 0 \\ \text{Gene 2} & 0 & 0 & 1 & 0 & 1 & 0 \\ \text{Gene 3} & 1 & 1 & 0 & 1 & 1 & 0 \\ \text{Gene 4} & 1 & 0 & 0 & 0 & 0 & 1 \\ \text{Gene 5} & 0 & 0 & 1 & 1 & 0 & 1 \\ \text{Gene 6} & 1 & 1 & 0 & 0 & 0 & 0 \end{array} = X$$

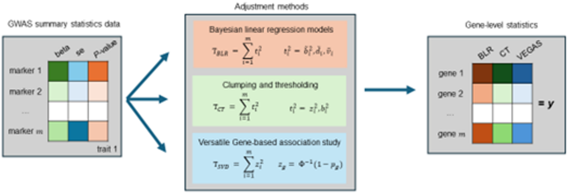
(5) Fitting the BLR model: Ranked gene sets by considering overlap

$\hat{y} = X\hat{\beta}$

- Fits a Bayesian regression model that allows **regularization** and **variable selection**
- Supports **single- or multi-trait** a nalyses
- Identifies associated features based on **posterior inclusion probabilities** (PIPs) for the regression effects

Gholipourshahraki et al., 2024

Bayesian MAGMA: Gene-level Statistics


GWAS summary statistics data
beta
se
p-value
marker 1
marker 2
...
marker m
trait 1
Adjustment methods
Bayesian linear regression models
$T_{BLR} = \sum_{i=1}^{m} t_i^2$
$t_i^2 = \hat{b}_i^2, \hat{d}_i, \hat{v}_i$
Clumping and thresholding
$T_{CT} = \sum_{i=1}^{m} t_i^2$
$t_i^2 = z_i^2, b_i^2$
Versatile Gene-based association study
$T_{SVD} = \sum_{i=1}^{m} z_i^2$
$z_g = \Phi^{-1}(1 - p_g)$
Gene-level statistics
BLR
CT
VEGAS
gene 1
gene 2
...
gene m
= $\mathbf{y}$

Compute gene-level (or other feature-level) association statistics:

- Account for correlation among marker statistics (i.e., linkage disequilibrium, LD)
- Different LD-adjustment methods (e.g., SVD, clumping and thresholding, BLR)
- The choice of method depends on the quality of the available GWAS summary statistics and LD reference panel

Bai et al., 2025

Bayesian MAGMA – KEGG Pathway

- GWAS summary statistics from nine studies (**T2D**, **CAD**, **CKD**, **HTN**, **BMI**, **WHR**, **Hb1Ac**, **TG**, **SBP**)
- **Gene sets** are defined by genes linked to **KEGG pathways**.
- Pathways relevant to **diabetes** are associated with **Type 2 Diabetes (T2D)** and correlated traits
- Enables identification of **cross-disease patterns** to better understand **comorbidities**

Gholipourshahraki et al., 2024

Bayesian MAGMA – DGIdb

- **Gene sets** are defined by genes linked to the **Anatomical Therapeutic Chemical (ATC)** classification system using the **Drug–Gene Interaction Database (DGIdb)**
- **Drug gene sets relevant to diabetes** show associations with **Type 2 Diabetes (T2D)** and related traits
- **Novel drug–gene set associations** may reveal opportunities for **drug repurposing**

Hjelholt et al., 2025

Bayesian MAGMA – Across Ancestries

- **Gene sets** are defined by genes linked to **KEGG pathways**.
- Joint analysis of **T2D** across three ancestries (**EUR**, **EAS**, **SAS**).
- Pathways relevant to **diabetes** show associations with **Type 2 Diabetes (T2D)** across two of the ancestries (**EUR** and **EAS**).
- Comparing these associations helps reveal **ancestry-specific biological mechanisms**.

Overview of BLR Models used in Gene Set Analyses

Model Type	Feature Integration	Grouping Basis	Prior Structure	What It Captures
Single-component BLR	Combines all biological features in one model	None	One global variance (τ^2)	All features contribute equally; uniform shrinkage
Multiple-component BLR	Integrates all layers but allows heterogeneous contributions	Learned from data	Mixture of variances ($\{\tau_k^2\}$)	Large, small, and null effect classes
Hierarchical BLR	Groups features by biological structure (e.g., genes, pathways)	Defined *a priori*	Group-specific **mixture of variances** ($\{\tau_{gk}^2\}$)	Within-group heterogeneity; enrichment and

Learning at Different Levels

Model Level	Key Parameters Learned	What They Represent	How They Are Learned	What We Learn Biologically
Effect sizes	β	Strength and direction of association for each feature	Posterior mean/median given priors and data	Which features drive the outcome
Indicator variables	δ_j (single trait), δ_j (multi-trait)	Whether feature j is active (and for which traits)	Estimated as posterior inclusion probabilities (PIPs)	Which features are relevant, and whether effects are shared or trait-specific
Variance	τ^2, $\{\tau_k^2\}$, $\{\tau_{ck}^2\}$	Magnitude of expected	Inferred hierarchically	How strongly different

Bayesian MAGMA: Summary

Advantages

- Incorporates **regularization** and **variable selection** via *spike-and-slab* priors.
- Controls **false positives** and handles **correlated gene sets**.
- Provides **posterior inclusion probabilities (PIP)** as evidence of association strength.
- Flexible framework supporting:
  - **Single- and multi-trait models**
  - **Integration of diverse genomic features**
  - **Modeling correlated traits** to uncover shared genetic architecture.

Limitations and Future Work

- Dependent on the **quality and granularity** of GWAS summary data and annotations.
- **Computationally more demanding** than standard MAGMA, especially for multi-trait analyses.
- Opportunities for **hierarchical model extensions** (e.g., grouping by pathways/tissues).
- **Multi-omics integration** to capture regulatory complexity across data layers.
- Systematic **comparisons with machine learning and deep learning** methods to assess performance and scalability.

References

Sørensen P, Rohde PD. *A Versatile Data Repository for GWAS Summary Statistics-Based Downstream Genomic Analysis of Human Complex Traits.*
medRxiv (2025). https://doi.org/10.1101/2025.10.01.25337099

Sørensen IF, Sørensen P. *Privacy-Preserving Multivariate Bayesian Regression Models for Overcoming Data Sharing Barriers in Health and Genomics.*
medRxiv (2025). https://doi.org/10.1101/2025.07.30.25332448

Hjelholt AJ, Gholipourshahraki T, Bai Z, Shrestha M, Kjølby M, Sørensen P, Rohde P. *Leveraging Genetic Correlations to Prioritize Drug Groups for Repurposing in Type 2 Diabetes.*
medRxiv (2025). https://doi.org/10.1101/2025.06.13.25329590

Gholipourshahraki T, Bai Z, Shrestha M, Hjelholt A, Rohde P, Fuglsang MK, Sørensen P. *Evaluation of Bayesian Linear Regression Models for Gene Set Prioritization in Complex Diseases.*
PLOS Genetics 20(11): e1011463 (2025).
https://doi.org/10.1371/journal.pgen.1011463