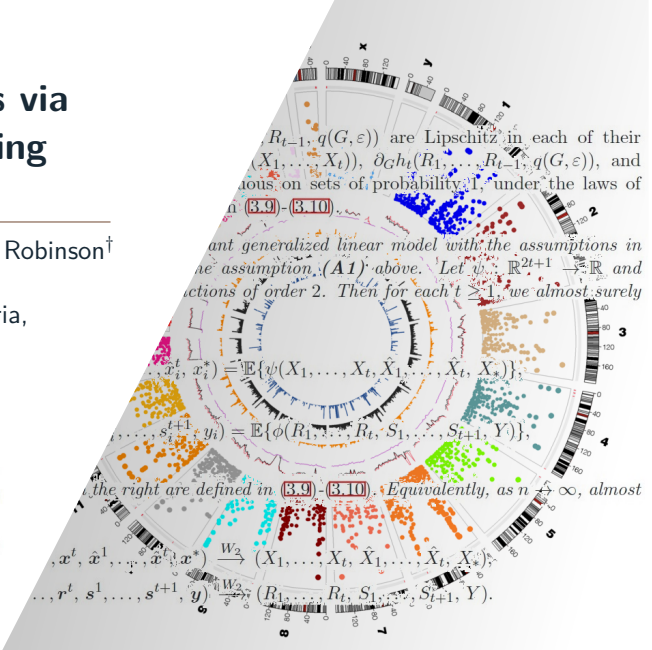
Inference of Genetic Effects via Approximate Message Passing

AI Depope[†], Marco Mondelli[†], Matthew R. Robinson[†]

[†] Institute of Science and Technology Austria, Klosterneuburg, Austria.

Institute of Science and Technology Austria

ICASSP 2024 KOREA



Agenda

1. What is a genome-wide association study (GWAS)?

Agenda

1. What is a genome-wide association study (GWAS)?
2. AMP overview. Making AMP approach scalable and stable for the GWAS inference task

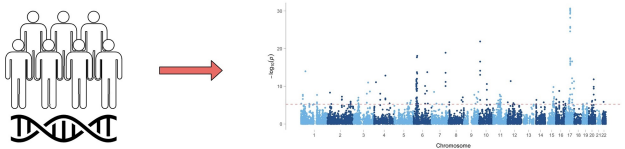
Agenda

1. What is a genome-wide association study (GWAS)?
2. AMP overview. Making AMP approach scalable and stable for the GWAS inference task
3. Comparison to the state-of-the-art methods (regenie, GMRM)

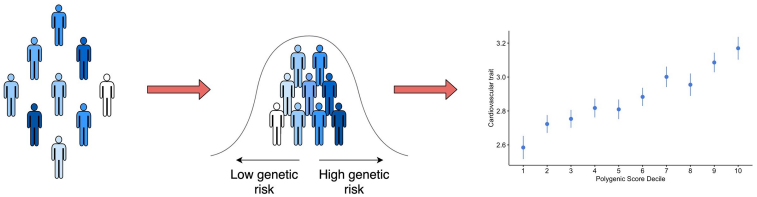
1. Genome-Wide Association Studies

1. Genome-Wide Association Studies

Step 1: Genome-wide association studies in adult populations from the UK Biobank



Step 2: Whole genome polygenic risk scores



Modelling genetic effects on a trait

Modelling genetic effects on a trait

- almost no limit to the amount of measured genetic variants (hundreds of millions; more genetic variants $\implies$ better generalization), but limited sample size

Modelling genetic effects on a trait

- almost no limit to the amount of measured genetic variants (hundreds of millions; more genetic variants $\implies$ better generalization), but limited sample size
- Data format (genotype matrices normalized column-wise):

$$\mathbf{X}_{ij} = \begin{cases} 2, & aa \\ 1, & Aa \\ 0, & AA \end{cases} \qquad \Longrightarrow \qquad \{0,1,2\}^{N\times P} \ni \mathbf{X} = \underbrace{\left.\begin{bmatrix} 1 & 2 & \dots & 0 \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 2 & \dots & 2 \end{bmatrix}\right\} \sim 10^5}_{\sim 10^6}$$

Modelling genetic effects on a trait

- almost no limit to the amount of measured genetic variants (hundreds of millions; more genetic variants $\implies$ better generalization), but limited sample size
- Data format (genotype matrices normalized column-wise):

$$\mathbf{X}_{ij} = \begin{cases} 2, & aa \\ 1, & Aa \\ 0, & AA \end{cases} \qquad \implies \qquad \mathbf{X} = \underbrace{\left.\begin{bmatrix} 1.886 & 4.242 & \dots & -0.472 \\ -0.472 & -1.414 & \dots & 1.886 \\ \vdots & \vdots & \ddots & \vdots \\ -0.472 & 4.242 & \dots & 4.243 \end{bmatrix}\right\} \sim 10^5}_{\sim 10^6}$$

Modelling genetic effects on a trait

- almost no limit to the amount of measured genetic variants (hundreds of millions; more genetic variants $\implies$ better generalization), but limited sample size
- Data format (genotype matrices normalized column-wise):

$$\mathbf{X}_{ij} = \begin{cases} 2, & aa \\ 1, & Aa \\ 0, & AA \end{cases} \qquad \implies \qquad \mathbf{X} = \underbrace{\left.\begin{bmatrix} 1.886 & 4.242 & \dots & -0.472 \\ -0.472 & -1.414 & \dots & 1.886 \\ \vdots & \vdots & \ddots & \vdots \\ -0.472 & 4.242 & \dots & 4.243 \end{bmatrix}\right\} \sim 10^5}_{\sim 10^6}$$

- Bayesian Linear Regression for the **individual-level** model:

$$y_i = \langle \mathbf{X}(i,:), \beta \rangle + \epsilon_i \text{ for } i \in [N] = \{1, \dots, N\}$$

Modelling genetic effects on a trait

- almost no limit to the amount of measured genetic variants (hundreds of millions; more genetic variants $\implies$ better generalization), but limited sample size
- Data format (genotype matrices normalized column-wise):

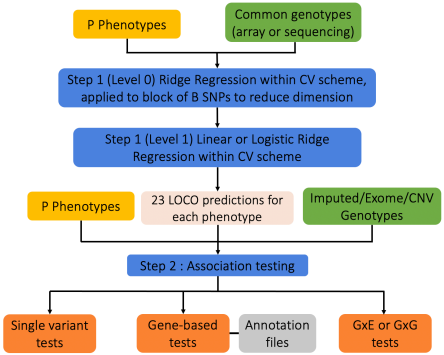
$$\mathbf{X}_{ij} = \begin{cases} 2, & aa \\ 1, & Aa \\ 0, & AA \end{cases} \qquad \implies \qquad \mathbf{X} = \underbrace{\left.\begin{bmatrix} 1.886 & 4.242 & \dots & -0.472 \\ -0.472 & -1.414 & \dots & 1.886 \\ \vdots & \vdots & \ddots & \vdots \\ -0.472 & 4.242 & \dots & 4.243 \end{bmatrix}\right\}}_{\sim 10^6} \sim 10^5$$

- Bayesian Linear Regression for the **individual-level** model:

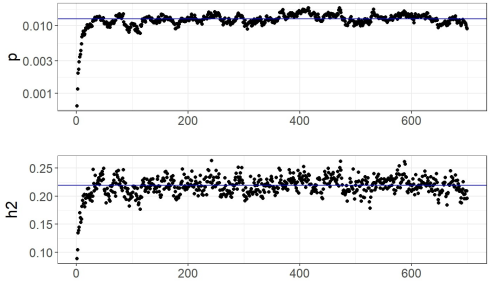
$$y_i = \langle \mathbf{X}(i,:), \beta \rangle + \epsilon_i \text{ for } i \in [N] = \{1, \dots, N\} \quad \text{and}$$

$$\beta_j \sim (1-\lambda) \cdot \delta_0(\cdot) + \lambda \cdot \sum_{i=1}^{L} \pi_i \cdot \mathcal{N}(\cdot, 0, \sigma_i^2), \quad \epsilon_i \sim \mathcal{N}(0, {\gamma_\epsilon}^{-1})$$

Prior Work



[regenie, PLINK]



[LDpred2, SBayesR, SBayesRC, GMRM]

2. Approximate Message Passing

- family of iterative algorithms that incorporate structural information about genetic signal

2. Approximate Message Passing

- family of iterative algorithms that incorporate structural information about genetic signal
- linear models [Kab03, BM12, BM11, DMM09, KMS+12], generalized linear models [BKM+19, MLKZ20, Ran11, SR14, SC19] and low-rank matrix estimation

2. Approximate Message Passing

- family of iterative algorithms that incorporate structural information about genetic signal
- linear models [Kab03, BM12, BM11, DMM09, KMS+12], generalized linear models [BKM+19, MLKZ20, Ran11, SR14, SC19] and low-rank matrix estimation
- achieves Bayes-optimal performance for some models [DM14, DJM13, BKM+19]

2. (EM) Vector AMP

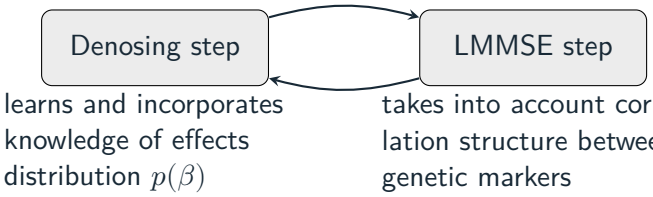
- Problem: correlation structure between columns of $\mathbf{X}$?

2. (EM) Vector AMP

- Problem: correlation structure between columns of $\mathbf{X}$?
- $\mathbf{X}$ right-orthogonally invariant [RSF16, T17]: distributions of objects in the high-dimensional limit precisely characterized by a *state evolution* recursion

2. (EM) Vector AMP

- Problem: correlation structure between columns of $\mathbf{X}$?
- $\mathbf{X}$ right-orthogonally invariant [RSF16, T17]: distributions of objects in the high-dimensional limit precisely characterized by a *state evolution* recursion



genomic**VAMP**

1. *Filtering* the normalized genotype matrix for first-degree relatives to reduce the correlation between rows ($\sim 400,000$ out of $460,000$ participants from UK Biobank study)

genomic**VAMP**

1. *Filtering* the normalized genotype matrix for first-degree relatives to reduce the correlation between rows ($\sim 400,000$ out of $460,000$ participants from UK Biobank study)
2. Initialization matters (sparsity $\sim$ 50k genetic positions, geometric sequence for prior mixture probabilites and variances)

genomic**VAMP**

1. *Filtering* the normalized genotype matrix for first-degree relatives to reduce the correlation between rows ($\sim 400,000$ out of $460,000$ participants from UK Biobank study)
2. Initialization matters (sparsity $\sim$ 50k genetic positions, geometric sequence for prior mixture probabilites and variances)
3. *Auto-tuning* of denoising signal error precision [FSR+17] combined with EM steps [VS12, FS17] that updates estimate of $p(\beta)$

genomic**VAMP**

1. *Filtering* the normalized genotype matrix for first-degree relatives to reduce the correlation between rows ($\sim 400,000$ out of $460,000$ participants from UK Biobank study)
2. Initialization matters (sparsity $\sim$ 50k genetic positions, geometric sequence for prior mixture probabilites and variances)
3. *Auto-tuning* of denoising signal error precision [FSR+17] combined with EM steps [VS12, FS17] that updates estimate of $p(\beta)$
4. Damping of denoised marker effects (momentum)

genomic**VAMP**

1. *Filtering* the normalized genotype matrix for first-degree relatives to reduce the correlation between rows ($\sim 400,000$ out of $460,000$ participants from UK Biobank study)
2. Initialization matters (sparsity $\sim$ 50k genetic positions, geometric sequence for prior mixture probabilites and variances)
3. *Auto-tuning* of denoising signal error precision [FSR+17] combined with EM steps [VS12, FS17] that updates estimate of $p(\beta)$
4. Damping of denoised marker effects (momentum)
5. Warm-start of conjugate gradients for LMMSE calculation [SD20]

genomic**VAMP**

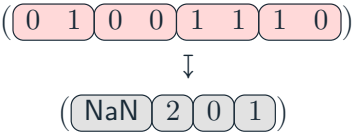
1. *Filtering* the normalized genotype matrix for first-degree relatives to reduce the correlation between rows ($\sim 400,000$ out of $460,000$ participants from UK Biobank study)
2. Initialization matters (sparsity $\sim$ 50k genetic positions, geometric sequence for prior mixture probabilites and variances)
3. *Auto-tuning* of denoising signal error precision [FSR+17] combined with EM steps [VS12, FS17] that updates estimate of $p(\beta)$
4. Damping of denoised marker effects (momentum)
5. Warm-start of conjugate gradients for LMMSE calculation [SD20]
6. Re-using Hutchinson estimator

genomic**VAMP**

1. *Filtering* the normalized genotype matrix for first-degree relatives to reduce the correlation between rows ($\sim 400,000$ out of $460,000$ participants from UK Biobank study)
2. Initialization matters (sparsity $\sim$ 50k genetic positions, geometric sequence for prior mixture probabilites and variances)
3. *Auto-tuning* of denoising signal error precision [FSR+17] combined with EM steps [VS12, FS17] that updates estimate of $p(\beta)$
4. Damping of denoised marker effects (momentum)
5. Warm-start of conjugate gradients for LMMSE calculation [SD20]
6. Re-using Hutchinson estimator
7. MPI + OpenMP

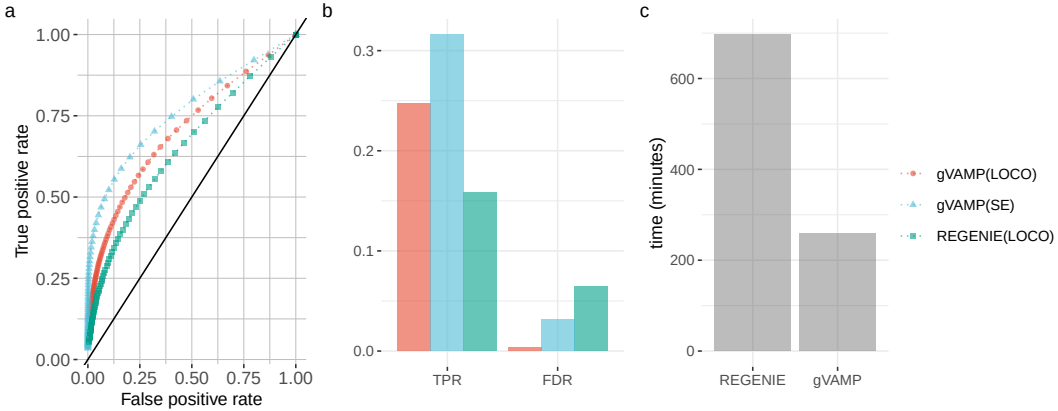
genomic**VAMP**

1. *Filtering* the normalized genotype matrix for first-degree relatives to reduce the correlation between rows ($\sim 400,000$ out of $460,000$ participants from UK Biobank study)
2. Initialization matters (sparsity $\sim$ 50k genetic positions, geometric sequence for prior mixture probabilites and variances)
3. *Auto-tuning* of denoising signal error precision [FSR+17] combined with EM steps [VS12, FS17] that updates estimate of $p(\beta)$
4. Damping of denoised marker effects (momentum)
5. Warm-start of conjugate gradients for LMMSE calculation [SD20]
6. Re-using Hutchinson estimator
7. MPI + OpenMP
8. data processing by using a lookup table + SIMD:

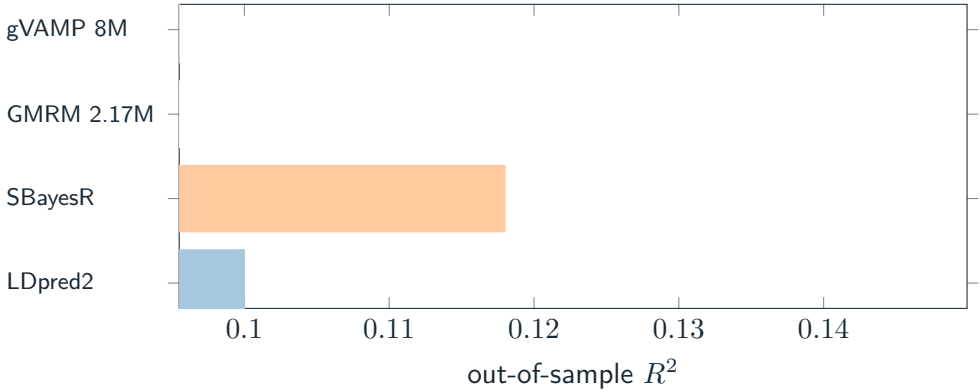


3. Association testing

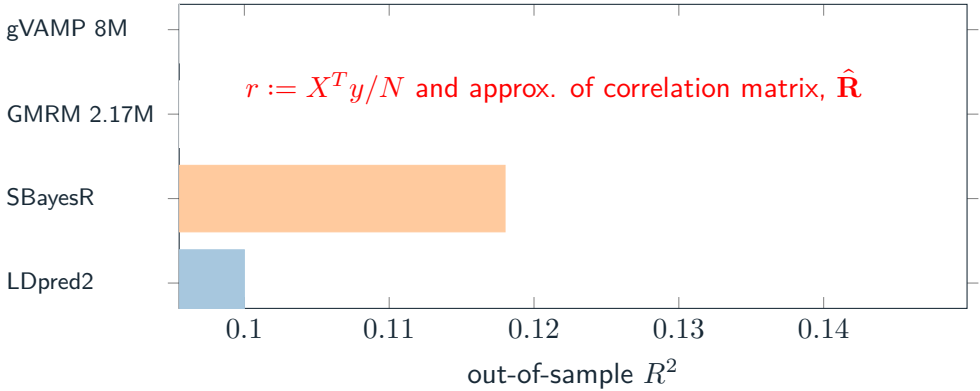
$$y^{(i)} := y - \mathbf{X}_{\setminus \mathsf{chr}(i)} \hat{\beta}_{\setminus \mathsf{chr}(i)} \sim \mathbf{X}(:, i)$$



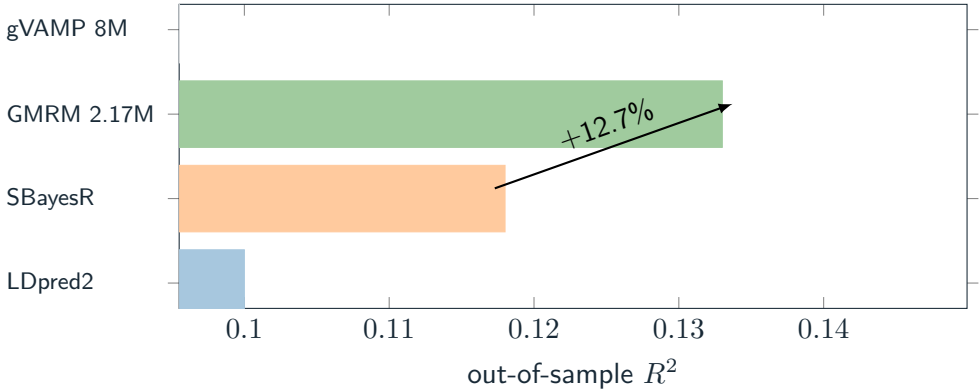
Prediction accuracy for BMI (Body Mass Index)



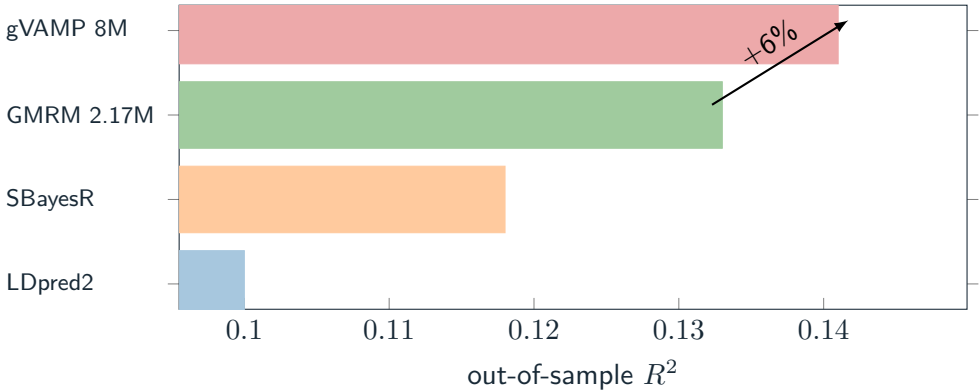
Prediction accuracy for BMI (Body Mass Index)



Prediction accuracy for BMI (Body Mass Index)

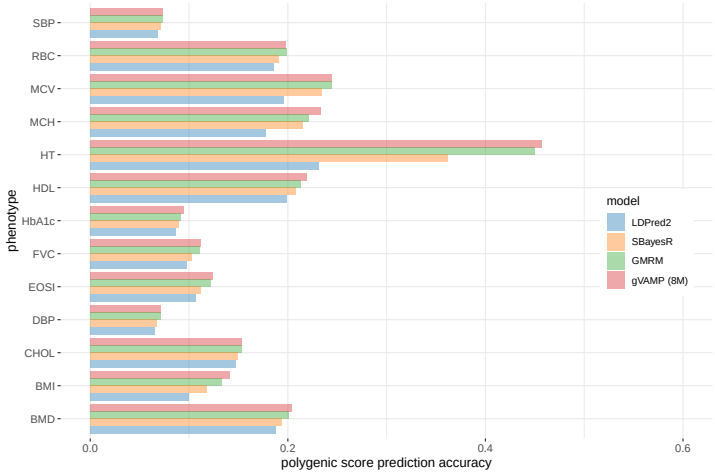


Prediction accuracy for BMI (Body Mass Index)



Prediction accuracy

SBP: Systolic blood pressure
RBC: Red blood cell count
MCV: Mean corpuscular volume
MCH: Mean corpuscular haemoglobin
HT: Standing height
HDL: High density lipoprotein
HbA1c: Glycated haemoglobin
FVC: Forced vital capacity
EOSI: Eosinophill count
DBP: Diastolic blood pressure
CHOL: Cholesterol
BMI: Body mass index
BMD: Heel bone mineral density



Summary & Future Directions

Summary & Future Directions

- gVAMP requires less than a day to model 8.4 million imputed genetic variants jointly in over $400,000$ UK Biobank participants. Other methods such as regenie, GMRM can not do this

Summary & Future Directions

- gVAMP requires less than a day to model 8.4 million imputed genetic variants jointly in over $400,000$ UK Biobank participants. Other methods such as regenie, GMRM can not do this
- exhibits lower FDR, greater TPR than regenie

Summary & Future Directions

- gVAMP requires less than a day to model 8.4 million imputed genetic variants jointly in over $400,000$ UK Biobank participants. Other methods such as regenie, GMRM can not do this
- exhibits lower FDR, greater TPR than regenie
- capable of analysing heterogeneous data (WES, X chromosome data)

Summary & Future Directions

- gVAMP requires less than a day to model 8.4 million imputed genetic variants jointly in over $400,000$ UK Biobank participants. Other methods such as regenie, GMRM can not do this
- exhibits lower FDR, greater TPR than regenie
- capable of analysing heterogeneous data (WES, X chromosome data)

1. **summary statistics & meta analysis models**

Summary & Future Directions

- gVAMP requires less than a day to model 8.4 million imputed genetic variants jointly in over $400,000$ UK Biobank participants. Other methods such as regenie, GMRM can not do this
- exhibits lower FDR, greater TPR than regenie
- capable of analysing heterogeneous data (WES, X chromosome data)

1. **summary statistics & meta analysis models**
   - access only to $r := X^T y/N$ and an approximation of a correlation matrix, called $\hat{\mathbf{R}}$
   - merging information from different databases/cohorts

Summary & Future Directions

- gVAMP requires less than a day to model 8.4 million imputed genetic variants jointly in over $400,000$ UK Biobank participants. Other methods such as regenie, GMRM can not do this
- exhibits lower FDR, greater TPR than regenie
- capable of analysing heterogeneous data (WES, X chromosome data)

1. **summary statistics & meta analysis models**
2. **time-to-event models**

Summary & Future Directions

- gVAMP requires less than a day to model 8.4 million imputed genetic variants jointly in over $400,000$ UK Biobank participants. Other methods such as regenie, GMRM can not do this
- exhibits lower FDR, greater TPR than regenie
- capable of analysing heterogeneous data (WES, X chromosome data)

1. **summary statistics & meta analysis models**
2. **time-to-event models**

$$\log y_i = \mu + \langle x_i, \beta \rangle + \frac{w_i}{\alpha} + \frac{K}{\alpha}$$

Summary & Future Directions

- gVAMP requires less than a day to model 8.4 million imputed genetic variants jointly in over $400,000$ UK Biobank participants. Other methods such as regenie, GMRM can not do this
- exhibits lower FDR, greater TPR than regenie
- capable of analysing heterogeneous data (WES, X chromosome data)

1. **summary statistics & meta analysis models**
2. **time-to-event models**

$$\log y_i = \mu + \langle x_i, \beta \rangle + \frac{w_i}{\alpha} + \frac{K}{\alpha}$$

3. **using gVAMP on WGS data** (between $10 - 12$M genetic variants)

Summary & Future Directions

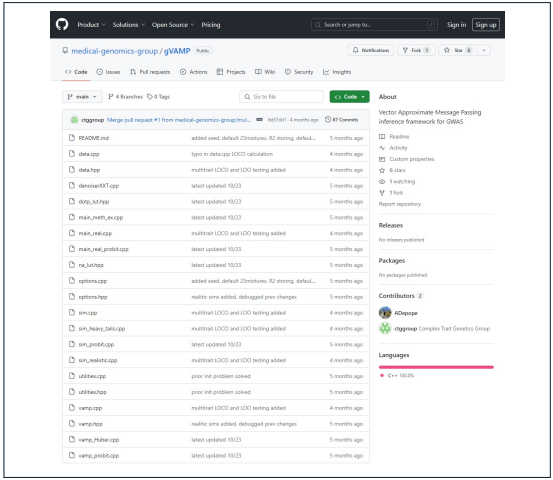
- gVAMP requires less than a day to model 8.4 million imputed genetic variants jointly in over $400,000$ UK Biobank participants. Other methods such as regenie, GMRM can not do this
- exhibits lower FDR, greater TPR than regenie
- capable of analysing heterogeneous data (WES, X chromosome data)

1. **summary statistics & meta analysis models**
2. **time-to-event models**

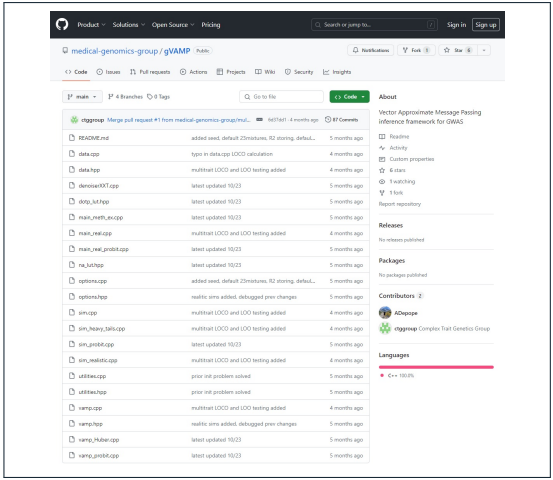
$$\log y_i = \mu + \langle x_i, \beta \rangle + \frac{w_i}{\alpha} + \frac{K}{\alpha}$$

3. **using gVAMP on WGS data** (between $10 - 12$M genetic variants)
4. **low-complexity alternatives to VAMP?**

gVAMP git repo: https://github.com/medical-genomics-group/gVAMP



gVAMP git repo: https://github.com/medical-genomics-group/gVAMP



The End

Thanks for your attention!

Extra Slides

REGENIE overview

- Step 1: (Inference)
  - (Ridge regression): reads P markers in blocks of $B = 1000$ consecutive markers and

$$\mathbf{X} = \begin{array}{c} \begin{matrix} B & B & \dots & B \end{matrix} \\ \begin{pmatrix} 0 & 4.242 & \dots & -1.414 \\ -1.414 & -1.414 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -1.414 & 4.242 & \dots & 1.414 \end{pmatrix} \end{array}$$

    for $\tau \in \{\tau_1, \dots, \tau_J\}$ and block index b calculate $\hat{\beta}_{\tau,b} = (\mathbf{X}_b^T \mathbf{X}_b + \tau I)^{-1} \mathbf{X}_b^T y$
  - (Cross-validation): fitting model $y = W\alpha + \varepsilon$ using ridge with cross-validation, where W contains JM/B predictors stacked
- Step 2: Single-variant association testing using Leave-One-Chromosome-Out (LOCO) approach

Leave-One-Out (LOO) testing approach

- using VAMP we obtain estimators $\hat{\beta}$ for the effect sizes in a linear model

$$y = \mathbf{X}\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I_N).$$

- Leave-One-Out (LOO) p-values for the statistical test $H_0 : \beta_i = 0$ are calculated as a p-value from t-test for testing whether the slope of a regression line is zero when regressing

$$y^{(i)} := y - \mathbf{X}_{\setminus i}\hat{\beta}_{\setminus i} \quad \text{on} \quad \mathbf{X}_i$$

($\mathbf{X}_{\setminus i}$ = all columns of $\mathbf{X}$ except the i-th one)

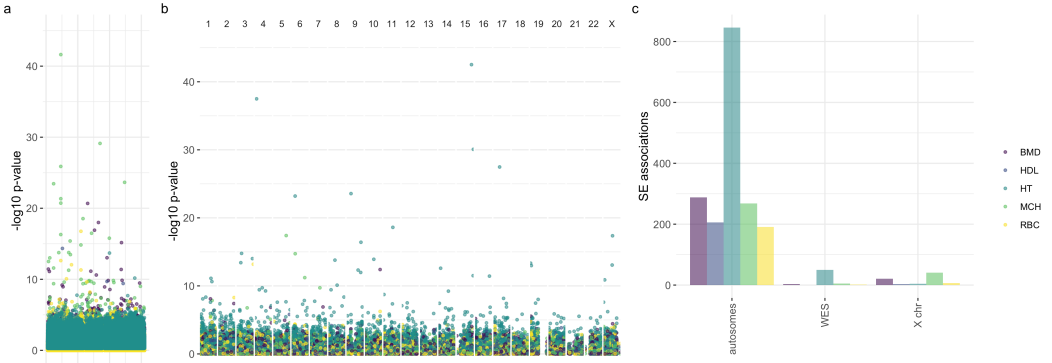
Parallelization of the code

$$\mathbf{X} = \begin{pmatrix} 0 & 4.242 & \dots & -1.414 \\ -1.414 & -1.414 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -1.414 & 4.242 & \dots & 1.414 \end{pmatrix}$$

- each MPI worker sees approximately equal number of consecutive columns ($\mathbf{X}$ is stored in a column-major format)
- $v \mapsto \mathbf{X}^T v$ operation is brought down to the level of single markers and combined with OpenMP reduction

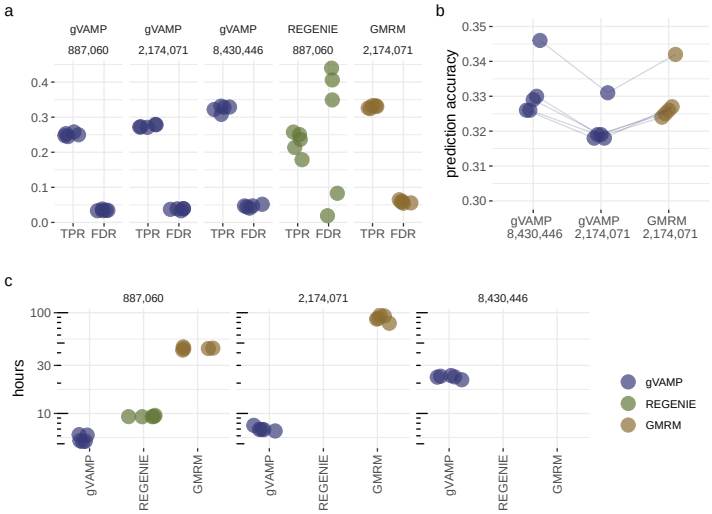
- $u \mapsto \mathbf{X}u = \sum_{w=1}^{W} \mathbf{X}_w u_w \rightarrow$ $2 \cdot (W-1) \cdot N$ doubles sent for communication
- $\mathbf{X}$ is being streamed-in using a lookup table (no additional memory is required, performing 4 basic operations at once):
  $(\boxed{0\ \ 1}\boxed{0\ \ 0}\boxed{1\ \ 1}\boxed{1\ \ 0}) \mapsto$
  $(\boxed{\text{NaN}}\boxed{2}\boxed{0}\boxed{1})$

Autosomal imputed data + X + WES analysis



3. Association testing

3. Association testing



Association testing: gVAMP vs GMRM