

Inference of Genetic Effects via Approximate Message Passing

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

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



Institute of
Science and
Technology
Austria



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

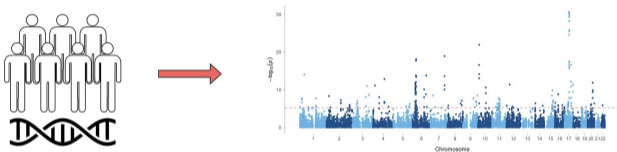
Agenda

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)

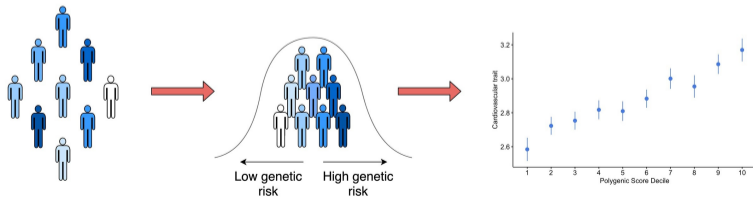
1. Genome-Wide Association Studies

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):

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

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):

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

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):

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

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

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

Modelling genetic effects on a trait

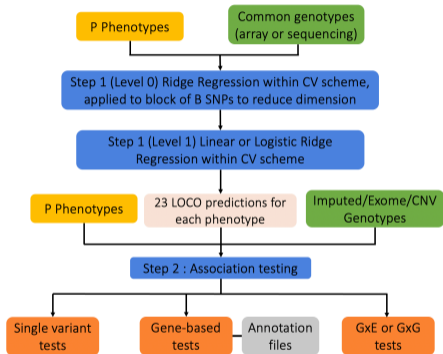
- almost no limit to the amount of measured genetic variants (hundreds of millions; more genetic variants \implies better generalization), but limited sample size
- Data format (genotype matrices normalized column-wise):

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

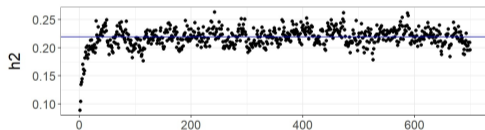
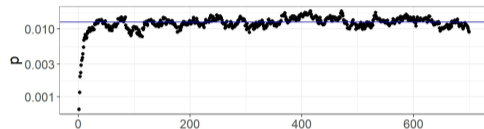
- Bayesian Linear Regression for the individual-level model:

$$y_i = \langle \mathbf{X}(i, \cdot), \beta \rangle + \epsilon_i \text{ for } i \in [N] = \{1, \dots, N\} \quad \text{and}$$
$$\beta_j \sim (1 - \lambda) \cdot \delta_0(\cdot) + \lambda \cdot \sum_{i=1}^L \pi_i \cdot \mathcal{N}(\cdot, 0, \sigma_i^2), \quad \epsilon_i \sim \mathcal{N}(0, \gamma_\epsilon^{-1})$$

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

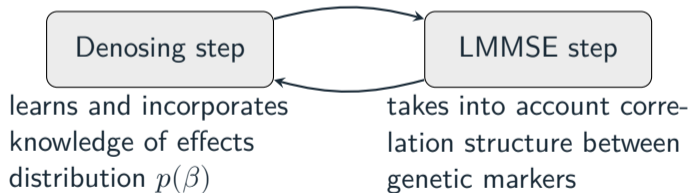
- Problem: correlation structure between columns of \mathbf{X} ?

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



genomicVAMP

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)

genomicVAMP

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 probabilities and variances)

genomicVAMP

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 probabilities and variances)
3. *Auto-tuning* of denoising signal error precision [FSR+17] combined with EM steps [VS12, FS17] that updates estimate of $p(\beta)$

genomicVAMP

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 probabilities 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)

genomicVAMP

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 probabilities 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]

genomicVAMP

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 probabilities 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

genomicVAMP

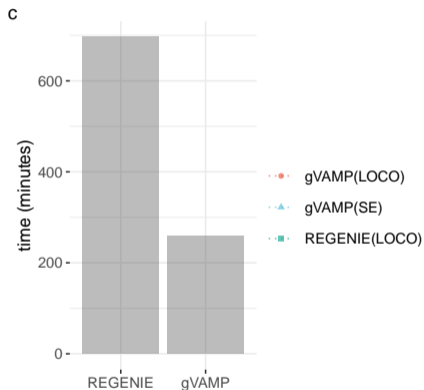
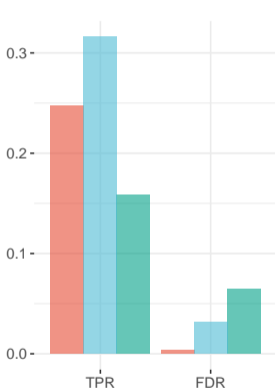
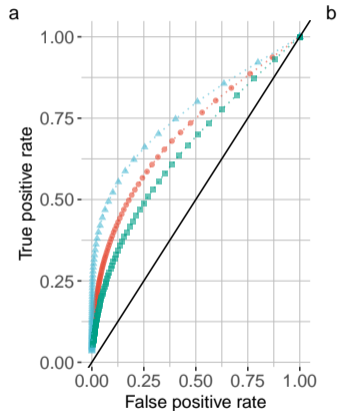
1. *Filtering* the normalized genotype matrix for first-degree relatives to reduce the correlation between rows ($\sim 400,000$ out of 460,000 participants from UK Biobank study)
2. Initialization matters (sparsity $\sim 50k$ genetic positions, geometric sequence for prior mixture probabilities and variances)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

genomicVAMP

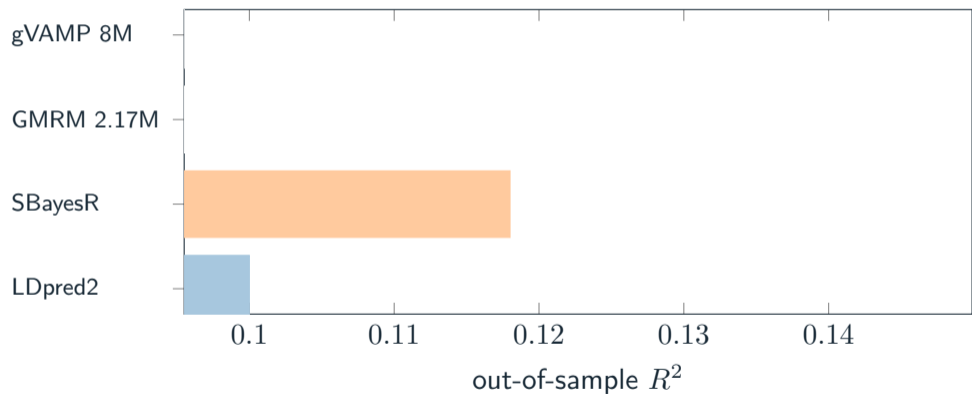
1. *Filtering* the normalized genotype matrix for first-degree relatives to reduce the correlation between rows ($\sim 400,000$ out of 460,000 participants from UK Biobank study)
2. Initialization matters (sparsity $\sim 50k$ genetic positions, geometric sequence for prior mixture probabilities and variances)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

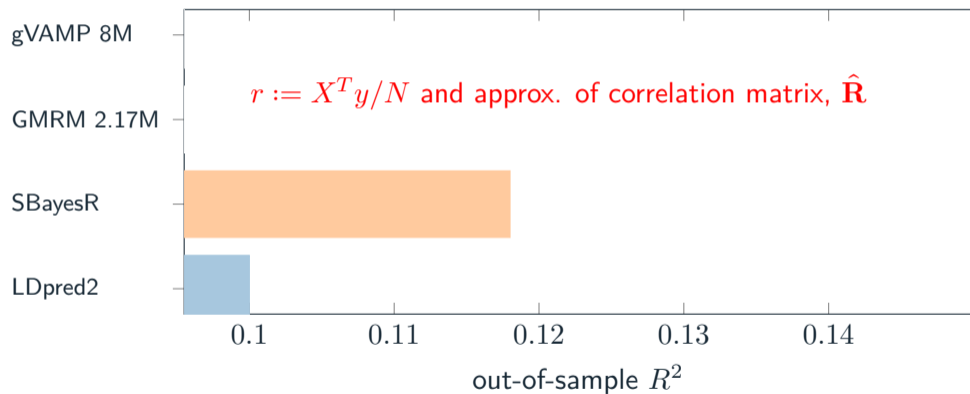
$$y^{(i)} := y - \mathbf{X}_{\setminus \text{chr}(i)} \hat{\beta}_{\setminus \text{chr}(i)} \sim \mathbf{X}(:, i)$$



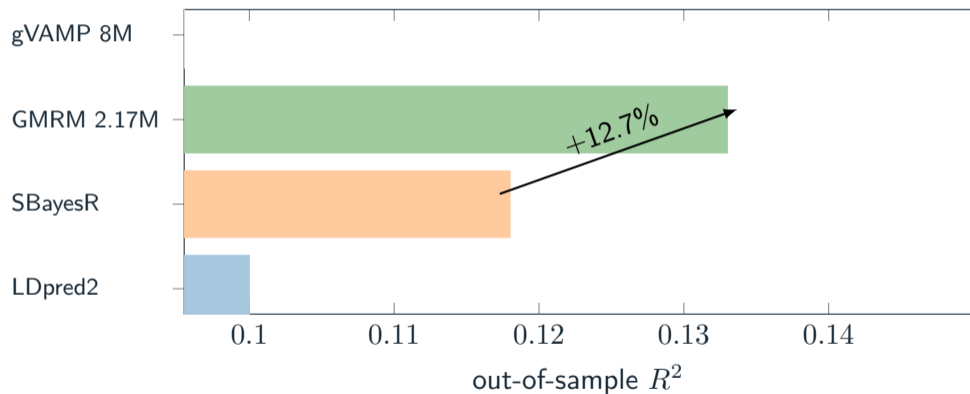
Prediction accuracy for BMI (Body Mass Index)



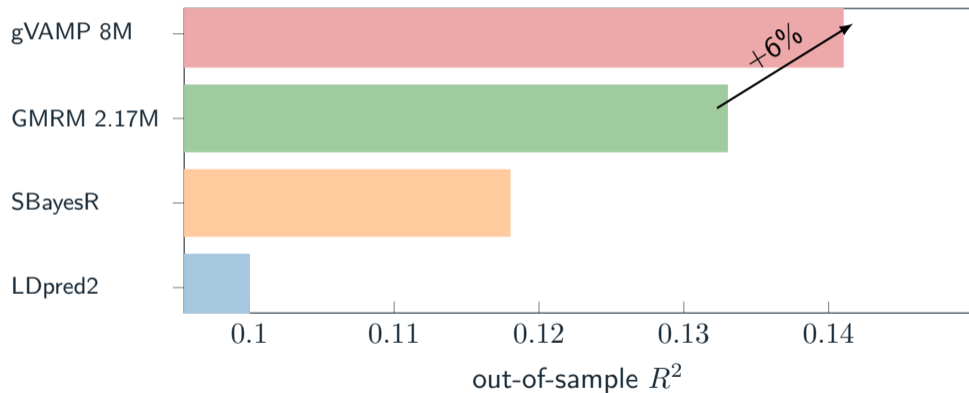
Prediction accuracy for BMI (Body Mass Index)



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

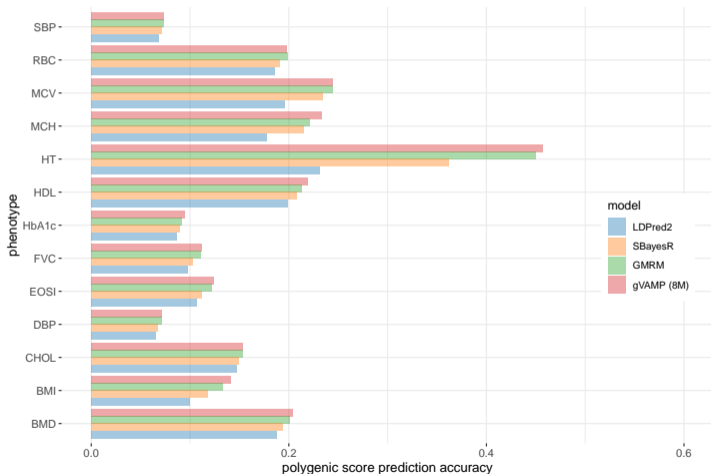
EOSI: Eosinophill count

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 – 12M 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 – 12M genetic variants)

4. **low-complexity alternatives to VAMP?**

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

The screenshot displays the GitHub repository interface for `medical-genomics-group/gVAMP`. The repository is a C++ project with 4 branches and 0 tags. The commit history table shows the following entries:

File	Commit Message	Time Ago
<code>ctggroup</code>	Merge pull request #1 from medical-genomics-group/ma...	4 months ago
<code>README.md</code>	added seed, default Z2mixtures, R2 storing, default...	5 months ago
<code>data.cpp</code>	typo in data.cpp LOCD calculation	4 months ago
<code>data.hpp</code>	multitrait LOCD and LOO testing added	4 months ago
<code>denoiseRXT.cpp</code>	latest updated 10/23	5 months ago
<code>drop_het.hpp</code>	latest updated 10/23	5 months ago
<code>main_real.cpp</code>	latest updated 10/23	5 months ago
<code>main_real.cpp</code>	multitrait LOCD and LOO testing added	4 months ago
<code>main_real_probit.cpp</code>	latest updated 10/23	5 months ago
<code>na_het.hpp</code>	latest updated 10/23	5 months ago
<code>options.cpp</code>	added seed, default Z2mixtures, R2 storing, default...	5 months ago
<code>options.hpp</code>	realistic sims added, debugged prev changes	5 months ago
<code>sim.cpp</code>	multitrait LOCD and LOO testing added	4 months ago
<code>sim_heavy_tails.cpp</code>	multitrait LOCD and LOO testing added	4 months ago
<code>sim_probit.cpp</code>	latest updated 10/23	5 months ago
<code>sim_realistic.cpp</code>	multitrait LOCD and LOO testing added	4 months ago
<code>utilReal.cpp</code>	prior init problems solved	5 months ago
<code>utilReal.hpp</code>	prior init problems solved	5 months ago
<code>vamp.cpp</code>	multitrait LOCD and LOO testing added	4 months ago
<code>vamp.hpp</code>	realistic sims added, debugged prev changes	5 months ago
<code>vamp_MuBer.cpp</code>	latest updated 10/23	5 months ago
<code>vamp_probit.cpp</code>	latest updated 10/23	5 months ago

Repository metadata:

- About:** Vector Approximate Message Passing inference framework for GWAS
- Releases:** No releases published
- Packages:** No packages published
- Contributors:** 2 (ADePope, ctggroup)
- Languages:** C++ 100.0%

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

The screenshot shows the GitHub repository page for `medical-genomics-group/gVAMP`. The repository is a C++ project with 4 branches and 0 tags. It contains a list of files, including `README.md`, `data.cpp`, `data.hpp`, `denoiseRXT.cpp`, `diag_hct.hpp`, `main_nest_en.cpp`, `main_nest.cpp`, `main_nest_probit.cpp`, `na_hct.hpp`, `options.cpp`, `options.hpp`, `sim.cpp`, `sim_heavy_tails.cpp`, `sim_probit.cpp`, `sim_realistic.cpp`, `utils.cpp`, `utils.hpp`, `vamp.cpp`, `vamp.hpp`, `vamp_Muber.cpp`, and `vamp_probit.cpp`. The repository has 17 commits and 1 contributor (ADeSope). The languages section shows C++ at 100%.

File Name	Commit Message	Time Ago
<code>README.md</code>	added seed, default ZDistributions, R2 storing, default...	5 months ago
<code>data.cpp</code>	typo in data.cpp LOOC calculation	4 months ago
<code>data.hpp</code>	multitrait LOOC and LOO testing added	4 months ago
<code>denoiseRXT.cpp</code>	latest updated 10/23	5 months ago
<code>diag_hct.hpp</code>	latest updated 10/23	5 months ago
<code>main_nest_en.cpp</code>	latest updated 10/23	5 months ago
<code>main_nest.cpp</code>	multitrait LOOC and LOO testing added	4 months ago
<code>main_nest_probit.cpp</code>	latest updated 10/23	5 months ago
<code>na_hct.hpp</code>	latest updated 10/23	5 months ago
<code>options.cpp</code>	added seed, default ZDistributions, R2 storing, default...	5 months ago
<code>options.hpp</code>	realistic sims added, debugged prev changes	5 months ago
<code>sim.cpp</code>	multitrait LOOC and LOO testing added	4 months ago
<code>sim_heavy_tails.cpp</code>	multitrait LOOC and LOO testing added	4 months ago
<code>sim_probit.cpp</code>	latest updated 10/23	5 months ago
<code>sim_realistic.cpp</code>	multitrait LOOC and LOO testing added	4 months ago
<code>utils.cpp</code>	prior int problems solved	5 months ago
<code>utils.hpp</code>	prior int problems solved	5 months ago
<code>vamp.cpp</code>	multitrait LOOC and LOO testing added	4 months ago
<code>vamp.hpp</code>	realistic sims added, debugged prev changes	5 months ago
<code>vamp_Muber.cpp</code>	latest updated 10/23	5 months ago
<code>vamp_probit.cpp</code>	latest updated 10/23	5 months ago

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

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

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

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

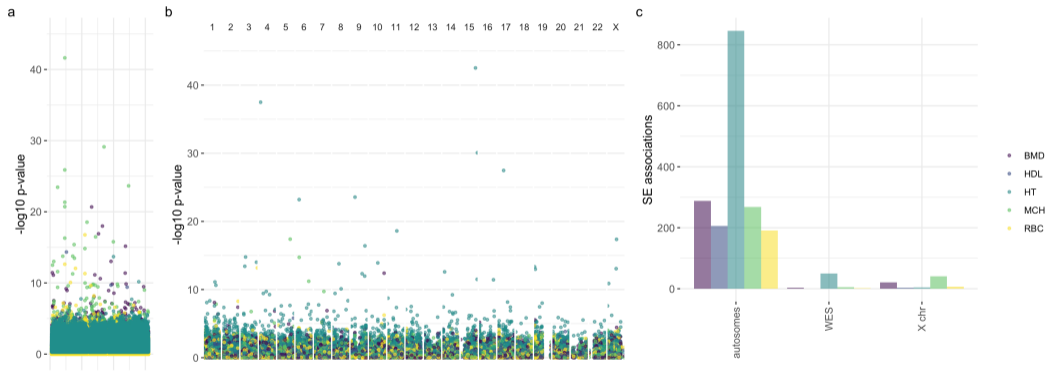
- each MPI worker sees approximately equal number of consecutive columns (\mathbf{X} is stored in a column-major format)
- $v \mapsto \mathbf{X}^T v$ operation is brought down to the level of single markers and combined with OpenMP reduction

- $u \mapsto \mathbf{X}u = \sum_{w=1}^W \mathbf{X}_w u_w \rightarrow 2 \cdot (W - 1) \cdot N$ 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):

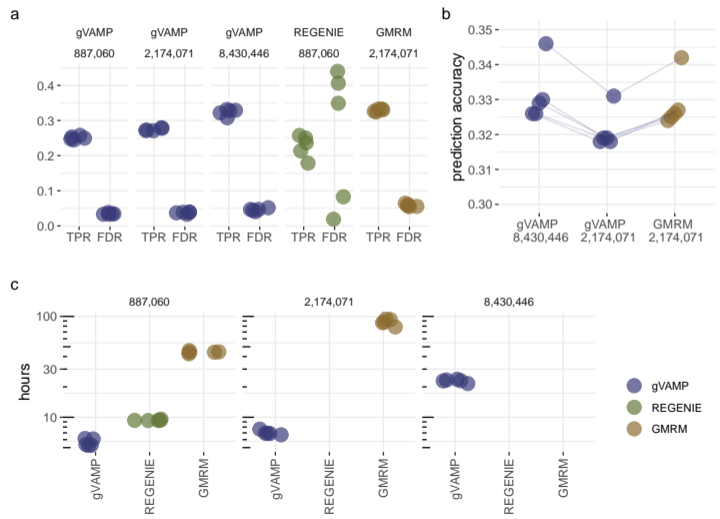
$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix} \mapsto \begin{pmatrix} \text{NaN} & 2 & 0 & 1 \end{pmatrix}$$

Autosomal imputed data + X + WES analysis



3. Association testing

3. Association testing



Association testing: gVAMP vs GMRM

