Linear models II: multi locus models and variance decomposition

Christoph Lippert[1] **Oliver Stegle**[2]

[1] Microsoft Research, Los Angeles, USA
[2] Max-Planck-Institutes Tübingen, Germany

Basel
09. September 2012

Microsoft Research

Overview

- Single marker association model with random effect term

$$\mathbf{y} = \underbrace{\mathbf{x}_s \beta_s}_{\text{genetic effect}} + \underbrace{\mathbf{u}}_{\text{random effect covariates}} + \underbrace{\boldsymbol{\epsilon}}_{\text{noise}}$$

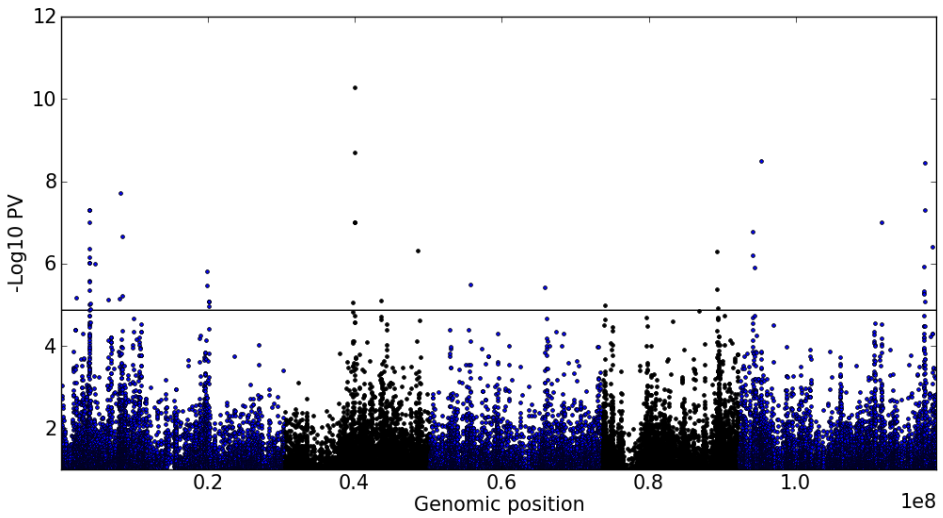
- Shortcomings
  - Weak effects are not captured by single-marker analysis.
  - Complex traits are controlled by more than a single SNP.

Overview

- Single marker association model with random effect term

$$\mathbf{y} = \underbrace{\mathbf{x}_s \beta_s}_{\text{genetic effect}} + \underbrace{\mathbf{u}}_{\text{random effect covariates}} + \underbrace{\boldsymbol{\epsilon}}_{\text{noise}}$$

- Shortcomings
  - Weak effects are not captured by single-marker analysis.
  - Complex traits are controlled by more than a single SNP.

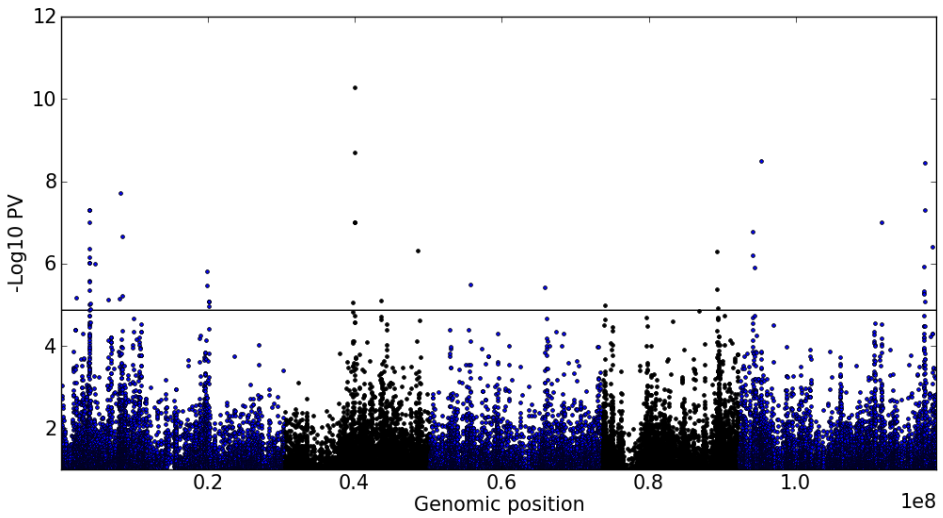


Overview

- Single marker association model with random effect term

$$\mathbf{y} = \underbrace{\mathbf{x}_s \beta_s}_{\text{genetic effect}} + \underbrace{\mathbf{u}}_{\text{random effect covariates}} + \underbrace{\boldsymbol{\epsilon}}_{\text{noise}}$$

- Shortcomings
  - Weak effects are not captured by single-marker analysis.
  - Complex traits are controlled by more than a single SNP.



Multi locus models

- Generalization to multiple genetic factors

$$\mathbf{y} = \underbrace{\sum_{s=1}^{S} \mathbf{x}_s \beta_s}_{\text{genetic effect}} + \underbrace{\mathbf{u}}_{\text{random effect covariates}} + \underbrace{\boldsymbol{\epsilon}}_{\text{noise}}$$

- Challenge: $N << S$: explicit estimation of all β_s is not feasible.
- Solutions
  - Regularize β_s (Ridge regression, LASSO)
  - Variance component models

[Wu et al., 2011]

Multi locus models

- Generalization to multiple genetic factors

$$\mathbf{y} = \underbrace{\sum_{s=1}^{S} \mathbf{x}_s \beta_s}_{\text{genetic effect}} + \underbrace{\mathbf{u}}_{\text{random effect covariates}} + \underbrace{\boldsymbol{\epsilon}}_{\text{noise}}$$

- Challenge: $N << S$: explicit estimation of all β_s is not feasible.
- Solutions
  - Regularize β_s (Ridge regression, LASSO)
  - Variance component modeling

[Wu et al., 2011]

Multi locus models

- Generalization to multiple genetic factors

$$\mathbf{y} = \underbrace{\sum_{s=1}^{S} \mathbf{x}_s \beta_s}_{\text{genetic effect}} + \underbrace{\mathbf{u}}_{\text{random effect covariates}} + \underbrace{\boldsymbol{\epsilon}}_{\text{noise}}$$

- Challenge: $N << S$: explicit estimation of all β_s is not feasible.
- Solutions
  - Regularize β_s (Ridge regression, LASSO)
  - Variance component modeling

[Wu et al., 2011]

Multi locus models

- Generalization to multiple genetic factors

$$\mathbf{y} = \underbrace{\sum_{s=1}^{S} \mathbf{x}_s \beta_s}_{\text{genetic effect}} + \underbrace{\mathbf{u}}_{\text{random effect covariates}} + \underbrace{\boldsymbol{\epsilon}}_{\text{noise}}$$

- Challenge: $N << S$: explicit estimation of all β_s is not feasible.
- Solutions
  - Regularize β_s (Ridge regression, LASSO)
  - Variance component modeling

[Wu et al., 2011]

Outline

Outline

Multi locus models

Random effect models

- For now, let's drop the random effect term

$$\mathbf{y} = \sum_{s=1}^{S} \mathbf{x}_s \beta_s + \boldsymbol{\epsilon}.$$

- For mathematical convenience, we choose a shared Gaussian prior on the weights and Gaussian noise

$$p(\beta_1, \ldots, \beta_S) = \prod_{s=1}^{S} \mathcal{N}\left(\beta_s \,|\, 0, \sigma_g^2\right) \quad p(\boldsymbol{\epsilon}) = \mathcal{N}\left(\boldsymbol{\epsilon} \,|\, \mathbf{0}, \sigma_e^2 \mathbf{I}\right)$$

- Marginalize out the weights $\beta_1, \ldots, \beta_S$

$$p(\mathbf{y} \,|\, \mathbf{X}, \sigma_{\mathrm{g}}^2, \sigma_{\mathrm{e}}^2) = \int_{\boldsymbol{\beta}} \mathcal{N}\left(\mathbf{y} \,\middle|\, \sum_{s=1}^{S} \mathbf{x}_s \beta_s, \sigma_{\mathrm{e}}^2 \mathbf{I}\right) \prod_{s=1}^{S} \mathcal{N}\left(\beta_s \,|\, 0, \sigma_{\mathrm{g}}^2\right) \mathrm{d}\boldsymbol{\beta}$$

$$= \mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{0}, \sigma_{\mathrm{g}}^2 \sum_{s=1}^{S} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}} + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)$$

Multi locus models

Random effect models

- For now, let's drop the random effect term

$$\mathbf{y} = \sum_{s=1}^{S} \mathbf{x}_s \beta_s + \boldsymbol{\epsilon}.$$

- For mathematical convenience, we choose a shared Gaussian prior on the weights and Gaussian noise

$$p(\beta_1, \ldots, \beta_S) = \prod_{s=1}^{S} \mathcal{N}\left(\beta_s \,|\, 0, \sigma_g^2\right) \quad p(\boldsymbol{\epsilon}) = \mathcal{N}\left(\boldsymbol{\epsilon} \,|\, \mathbf{0}, \sigma_e^2 \mathbf{I}\right)$$

- Marginalize out the weights $\beta_1, \ldots, \beta_S$

$$p(\mathbf{y} \,|\, \mathbf{X}, \sigma_{\mathrm{g}}^2, \sigma_{\mathrm{e}}^2) = \int_{\boldsymbol{\beta}} \mathcal{N}\left(\mathbf{y} \,\middle|\, \sum_{s=1}^{S} \mathbf{x}_s \beta_s, \sigma_{\mathrm{e}}^2 \mathbf{I}\right) \prod_{s=1}^{S} \mathcal{N}\left(\beta_s \,|\, 0, \sigma_{\mathrm{g}}^2\right) \mathrm{d}\boldsymbol{\beta}$$

$$= \mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{0}, \sigma_{\mathrm{g}}^2 \sum_{s=1}^{S} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}} + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)$$

Multi locus models

Random effect models

- For now, let's drop the random effect term

$$\mathbf{y} = \sum_{s=1}^{S} \mathbf{x}_s \beta_s + \boldsymbol{\epsilon}.$$

- For mathematical convenience, we choose a shared Gaussian prior on the weights and Gaussian noise

$$p(\beta_1, \ldots, \beta_S) = \prod_{s=1}^{S} \mathcal{N}\left(\beta_s \,|\, 0, \sigma_g^2\right) \quad p(\boldsymbol{\epsilon}) = \mathcal{N}\left(\boldsymbol{\epsilon} \,|\, \mathbf{0}, \sigma_e^2 \mathbf{I}\right)$$

- Marginalize out the weights $\beta_1, \ldots, \beta_S$

$$\begin{aligned} p(\mathbf{y} \,|\, \mathbf{X}, \sigma_{\mathrm{g}}^2, \sigma_{\mathrm{e}}^2) &= \int_{\boldsymbol{\beta}} \mathcal{N}\left(\mathbf{y} \,\middle|\, \sum_{s=1}^{S} \mathbf{x}_s \beta_s, \sigma_{\mathrm{e}}^2 \mathbf{I}\right) \prod_{s=1}^{S} \mathcal{N}\left(\beta_s \,|\, 0, \sigma_{\mathrm{g}}^2\right) \mathrm{d}\boldsymbol{\beta} \\ &= \mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{0}, \sigma_{\mathrm{g}}^2 \sum_{s=1}^{S} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}} + \sigma_{\mathrm{e}}^2 \mathbf{I}\right) \end{aligned}$$

Multi locus models

Remarks

$$p(\mathbf{y} \mid \mathbf{X}, \sigma_{\mathrm{g}}^2, \sigma_{\mathrm{e}}^2) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{0}, \sigma_{\mathrm{g}}^2 \underbrace{\sum_{s=1}^{S} \mathbf{s}_s \mathbf{x}_s^{\mathrm{T}}}_{\mathbf{K}_g} + \sigma_{\mathrm{e}}^2 \mathbf{I}\right) \tag{1}$$

- $\mathbf{K}_g$ genotype covariance matrix.
  - Closely related to Kinship explaining population structure.
- Inference can be done my maximum likelihood.
- The ratio of σ_{g}^2 and σ_{e}^2 defines the narrow sense heritability of the trait

$$h = \frac{\sigma_{\mathrm{g}}^2}{\sigma_{\mathrm{g}}^2 + \sigma_{\mathrm{e}}^2}.$$

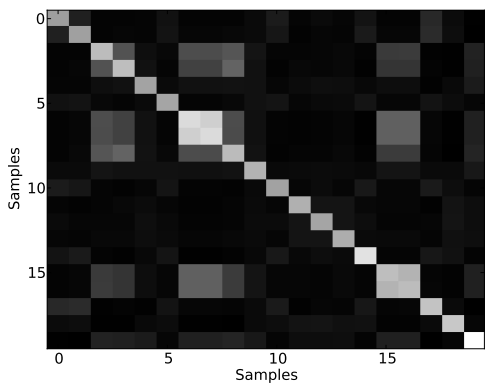
Multi locus models

Remarks

$$p(\mathbf{y} \mid \mathbf{X}, \sigma_{\mathrm{g}}^2, \sigma_{\mathrm{e}}^2) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{0}, \sigma_{\mathrm{g}}^2 \underbrace{\sum_{s=1}^{S} \mathbf{s}_s \mathbf{x}_s^{\mathrm{T}}}_{\mathbf{K}_g} + \sigma_{\mathrm{e}}^2 \mathbf{I}\right) \tag{1}$$

- $\mathbf{K}_g$ genotype covariance matrix.
  - Closely related to Kinship explaining population structure.
- Inference can be done my maximum likelihood.
- The ratio of σ_{g}^2 and σ_{e}^2 defines the narrow sense heritability of the trait

$$h = \frac{\sigma_{\mathrm{g}}^2}{\sigma_{\mathrm{g}}^2 + \sigma_{\mathrm{e}}^2}.$$



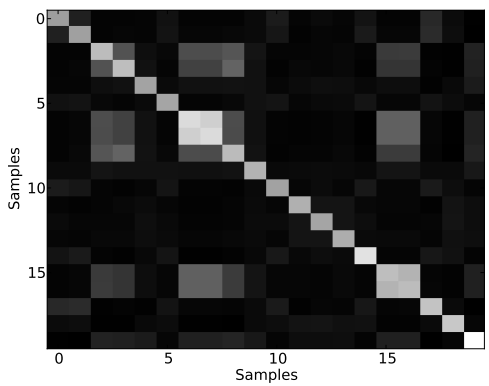
Multi locus models

Remarks

$$p(\mathbf{y} \mid \mathbf{X}, \sigma_{\mathrm{g}}^2, \sigma_{\mathrm{e}}^2) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{0}, \sigma_{\mathrm{g}}^2 \underbrace{\sum_{s=1}^{S} \mathbf{s}_s \mathbf{x}_s^{\mathrm{T}}}_{\mathbf{K}_g} + \sigma_{\mathrm{e}}^2 \mathbf{I}\right) \tag{1}$$

- $\mathbf{K}_g$ genotype covariance matrix.
  - Closely related to Kinship explaining population structure.
- Inference can be done my maximum likelihood.
- The ratio of σ_{g}^2 and σ_{e}^2 defines the narrow sense heritability of the trait

$$h = \frac{\sigma_{\mathrm{g}}^2}{\sigma_{\mathrm{g}}^2 + \sigma_{\mathrm{e}}^2}.$$

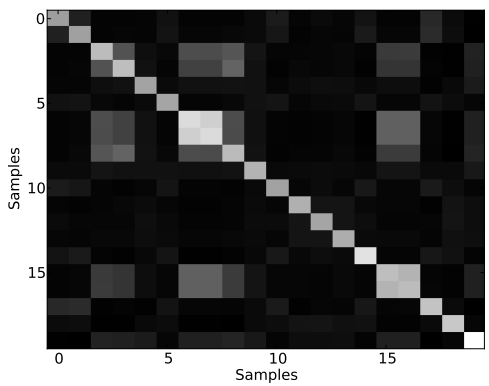


Multi locus models

Remarks

$$p(\mathbf{y} \mid \mathbf{X}, \sigma_{\mathrm{g}}^2, \sigma_{\mathrm{e}}^2) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{0}, \sigma_{\mathrm{g}}^2 \underbrace{\sum_{s=1}^{S} \mathbf{s}_s \mathbf{x}_s^{\mathrm{T}}}_{\mathbf{K}_g} + \sigma_{\mathrm{e}}^2 \mathbf{I}\right) \tag{1}$$

- $\mathbf{K}_g$ genotype covariance matrix.
  - Closely related to Kinship explaining population structure.
- Inference can be done my maximum likelihood.
- The ratio of σ_{g}^2 and σ_{e}^2 defines the narrow sense heritability of the trait

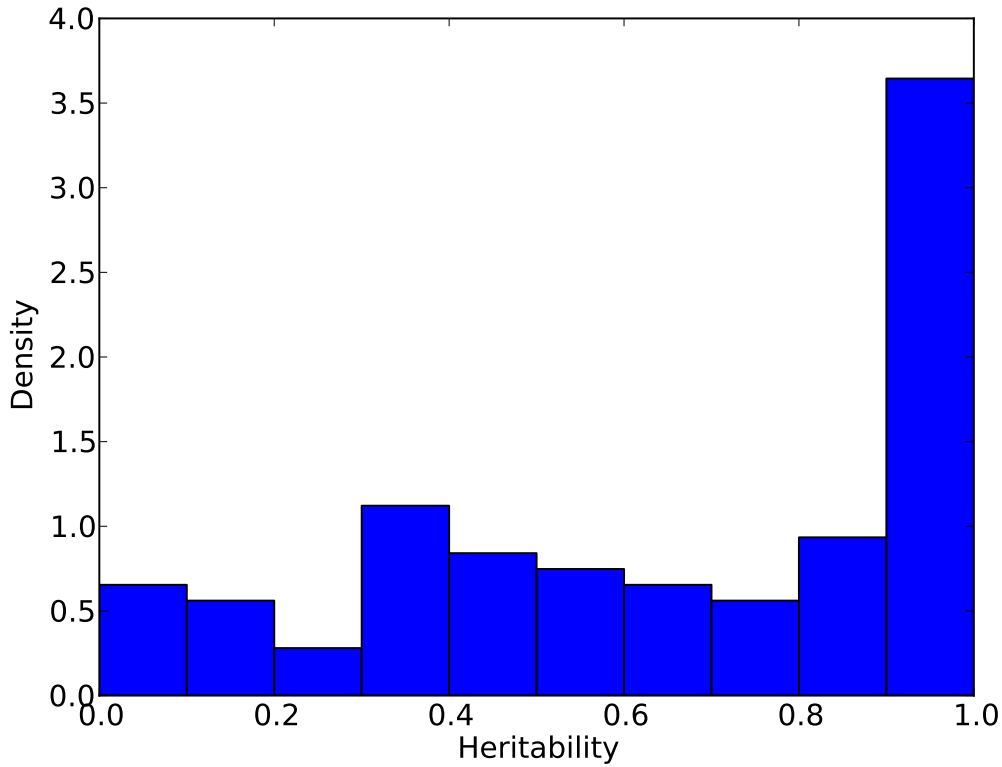


$$h = \frac{\sigma_{\mathrm{g}}^2}{\sigma_{\mathrm{g}}^2 + \sigma_{\mathrm{e}}^2}.$$

Heritability

Heritability estimated on 107 *A. thaliana* phenotypes

Global genetic heritability

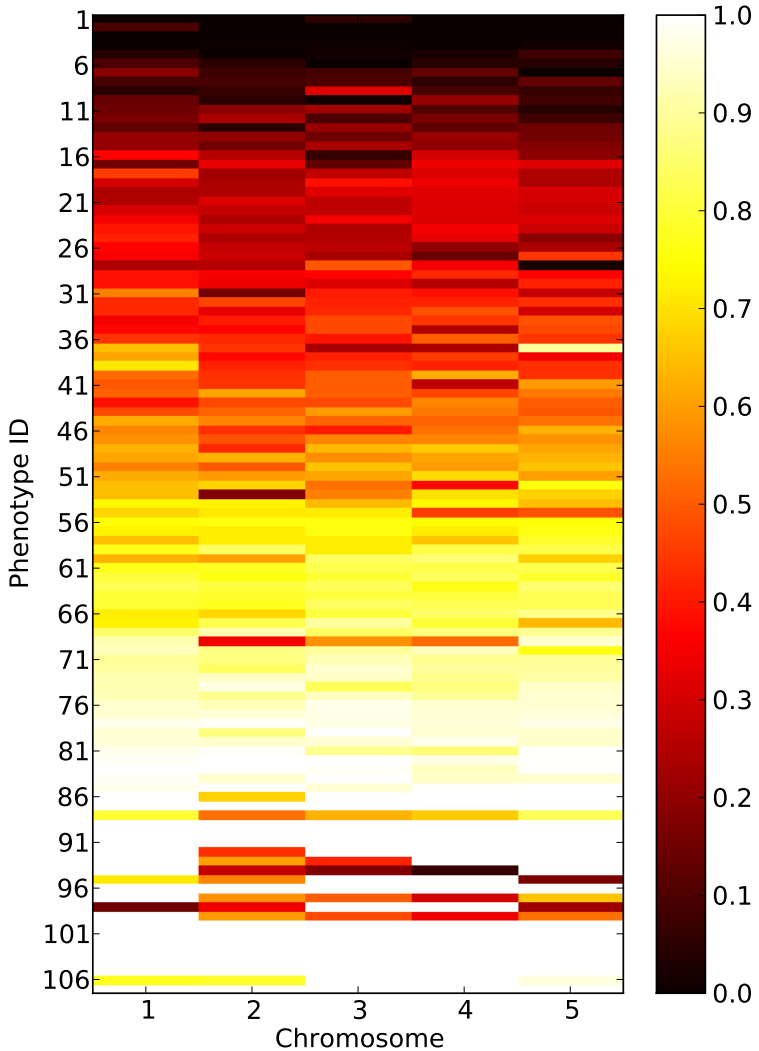


Heritability

Heritability estimated on 107 *A. thaliana* phenotypes

- Estimate can be restricted to a genomic region such as a single chromosome, etc.

$$\mathcal{N}\left(\mathbf{y} \,|\, \mathbf{0}, \sigma_{\mathrm{g}}^2 \sum_{s \in \mathrm{Chrom}} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}} + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)$$



Window-based composite variance analysis

Region-based testing

- Just fitting a particular region ignores the genome-wide context
- Variance dissection with region-based separation

$$p(\mathbf{y} \mid W) = \mathcal{N}(\mathbf{y} \mid \mathbf{0}, \sigma_w^2 \underbrace{\sum_{s \in W} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}}}_{\mathbf{K}_w} + \sigma_g^2 \underbrace{\sum_{s \notin W} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}}}_{\mathbf{K}_g} + \sigma_{\mathrm{e}}^2 \mathbf{I})$$

- Explained variance components can be read off subject to suitable normalization of the covariances $\mathbf{K}_w$ and $\mathbf{K}_g$.
- "Local" heritability

$$h(W) = \frac{\sigma_w^2}{\sigma_w^2 + \sigma_g^2 + \sigma_e^2}$$

Window-based composite variance analysis

Region-based testing

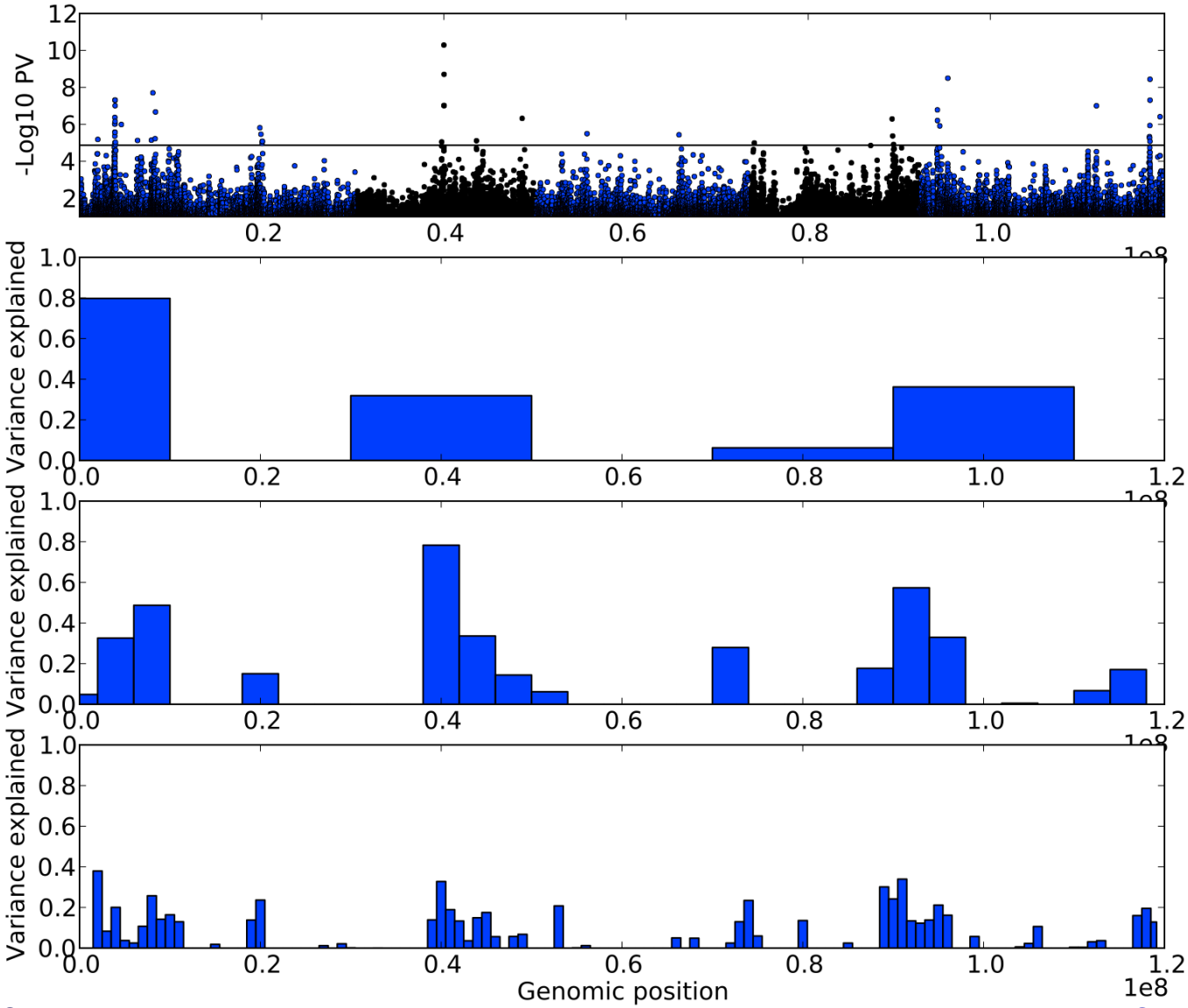
- Just fitting a particular region ignores the genome-wide context
- Variance dissection with region-based separation

$$p(\mathbf{y} \mid W) = \mathcal{N}(\mathbf{y} \mid \mathbf{0}, \sigma_w^2 \underbrace{\sum_{s \in W} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}}}_{\mathbf{K}_w} + \sigma_g^2 \underbrace{\sum_{s \notin W} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}}}_{\mathbf{K}_g} + \sigma_{\mathrm{e}}^2 \mathbf{I})$$

- Explained variance components can be read off subject to suitable normalization of the covariances $\mathbf{K}_w$ and $\mathbf{K}_g$.
- "Local" heritability

$$h(W) = \frac{\sigma_w^2}{\sigma_w^2 + \sigma_g^2 + \sigma_e^2}$$

Window-based composite variance analysis



Window-based composite variance analysis

Significance testing

- Analogously to fixed effect testing, the significance of a specific window can be tested.
- Likelihood-ratio statistics to score the relevance of a particular genomic region W

$$\mathrm{LOD}(W) = \frac{\mathcal{N}\left(\mathbf{y} \middle| \mathbf{0}, \sigma_w^2 \sum_{s \in W} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}} + \sigma_g^2 \sum_{s \notin W} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}} + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)}{\mathcal{N}\left(\mathbf{y} \middle| \mathbf{0}, \sigma_g^2 \sum_{s \notin W} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}} + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)}$$

- P-values can be obtained from permutation statistics or analytical approximation (variants of score tests or likelihood ratio tests).

Window-based composite variance analysis

Significance testing

- Analogously to fixed effect testing, the significance of a specific window can be tested.
- Likelihood-ratio statistics to score the relevance of a particular genomic region W

$$\mathrm{LOD}(W) = \frac{\mathcal{N}\left(\mathbf{y} \middle| \mathbf{0}, \sigma_w^2 \sum_{s \in W} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}} + \sigma_g^2 \sum_{s \notin W} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}} + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)}{\mathcal{N}\left(\mathbf{y} \middle| \mathbf{0}, \sigma_g^2 \sum_{s \notin W} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}} + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)}$$

- P-values can be obtained from permutation statistics or analytical approximation (variants of score tests or likelihood ratio tests).

Window-based composite variance analysis

Significance testing

- Analogously to fixed effect testing, the significance of a specific window can be tested.
- Likelihood-ratio statistics to score the relevance of a particular genomic region W

$$\text{LOD}(W) = \frac{\mathcal{N}\left(\mathbf{y} \middle| \mathbf{0}, \sigma_w^2 \sum_{s \in W} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}} + \sigma_g^2 \sum_{s \notin W} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}} + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)}{\mathcal{N}\left(\mathbf{y} \middle| \mathbf{0}, \sigma_g^2 \sum_{s \notin W} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}} + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)}$$

- P-values can be obtained from permutation statistics or analytical approximation (variants of score tests or likelihood ratio tests).

Phenotype prediction

Best linear unbiased prediction

- Given the phenotype values of a set of individuals and the genetic relatedness, we can predict the genetic component of the phenotype of a new individual.
- $P(\mathbf{y}^\star \mid \mathbf{y}) = \mathcal{N}\left(\mathbf{y}^\star \mid \boldsymbol{\mu}^\star, \sigma_{\mathrm{g}}^2 \mathbf{V}_g^\star + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)$
- Predicitve mean: $\boldsymbol{\mu}^\star = \underbrace{\sigma_{\mathrm{g}}^2 \mathbf{K}_g^{\star,:} \left(\sigma_{\mathrm{g}}^2 \mathbf{K}_g + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)^{-1}}_{\text{BLUP}} \mathbf{y}$
- Predictive Variance: $\mathbf{V}_g^\star = \mathbf{K}_g^{\star,\star} - \sigma_{\mathrm{g}}^2 \mathbf{K}_g^{\star,:} \left(\sigma_{\mathrm{g}}^2 \mathbf{K}_g + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)^{-1} \mathbf{K}_g^{:,\star}$

Phenotype prediction

Best linear unbiased prediction

- Given the phenotype values of a set of individuals and the genetic relatedness, we can predict the genetic component of the phenotype of a new individual.
- $P(\mathbf{y}^\star \,|\, \mathbf{y}) = \mathcal{N}\left(\mathbf{y}^\star \,\middle|\, \boldsymbol{\mu}^\star, \sigma_{\mathrm{g}}^2 \mathbf{V}_g^\star + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)$
- Predicitve mean: $\boldsymbol{\mu}^\star = \underbrace{\sigma_{\mathrm{g}}^2 \mathbf{K}_g^{\star,:} \left(\sigma_{\mathrm{g}}^2 \mathbf{K}_g + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)^{-1}}_{\text{BLUP}} \mathbf{y}$
- Predictive Variance: $\mathbf{V}_g^\star = \mathbf{K}_g^{\star,\star} - \sigma_{\mathrm{g}}^2 \mathbf{K}_g^{\star,:} \left(\sigma_{\mathrm{g}}^2 \mathbf{K}_g + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)^{-1} \mathbf{K}_g^{:,\star}$

Phenotype prediction

Best linear unbiased prediction

- Given the phenotype values of a set of individuals and the genetic relatedness, we can predict the genetic component of the phenotype of a new individual.
- $P(\mathbf{y}^\star \,|\, \mathbf{y}) = \mathcal{N}\left(\mathbf{y}^\star \,\middle|\, \boldsymbol{\mu}^\star, \sigma_{\mathrm{g}}^2 \mathbf{V}_g^\star + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)$
- Predicitve mean: $\boldsymbol{\mu}^\star = \underbrace{\sigma_{\mathrm{g}}^2 \mathbf{K}_g^{\star,:} \left(\sigma_{\mathrm{g}}^2 \mathbf{K}_g + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)^{-1} \mathbf{y}}_{\text{BLUP}}$
- Predictive Variance: $\mathbf{V}_g^\star = \mathbf{K}_g^{\star,\star} - \sigma_{\mathrm{g}}^2 \mathbf{K}_g^{\star,:} \left(\sigma_{\mathrm{g}}^2 \mathbf{K}_g + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)^{-1} \mathbf{K}_g^{:,\star}$

Phenotype prediction

Best linear unbiased prediction

- Given the phenotype values of a set of individuals and the genetic relatedness, we can predict the genetic component of the phenotype of a new individual.
- $P(\mathbf{y}^\star \,|\, \mathbf{y}) = \mathcal{N}\left(\mathbf{y}^\star \,\middle|\, \boldsymbol{\mu}^\star, \sigma_{\mathrm{g}}^2 \mathbf{V}_g^\star + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)$
- Predicitve mean: $\boldsymbol{\mu}^\star = \underbrace{\sigma_{\mathrm{g}}^2 \mathbf{K}_g^{\star,:} \left(\sigma_{\mathrm{g}}^2 \mathbf{K}_g + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)^{-1} \mathbf{y}}_{\text{BLUP}}$
- Predictive Variance: $\mathbf{V}_g^\star = \mathbf{K}_g^{\star,\star} - \sigma_{\mathrm{g}}^2 \mathbf{K}_g^{\star,:} \left(\sigma_{\mathrm{g}}^2 \mathbf{K}_g + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)^{-1} \mathbf{K}_g^{:,\star}$

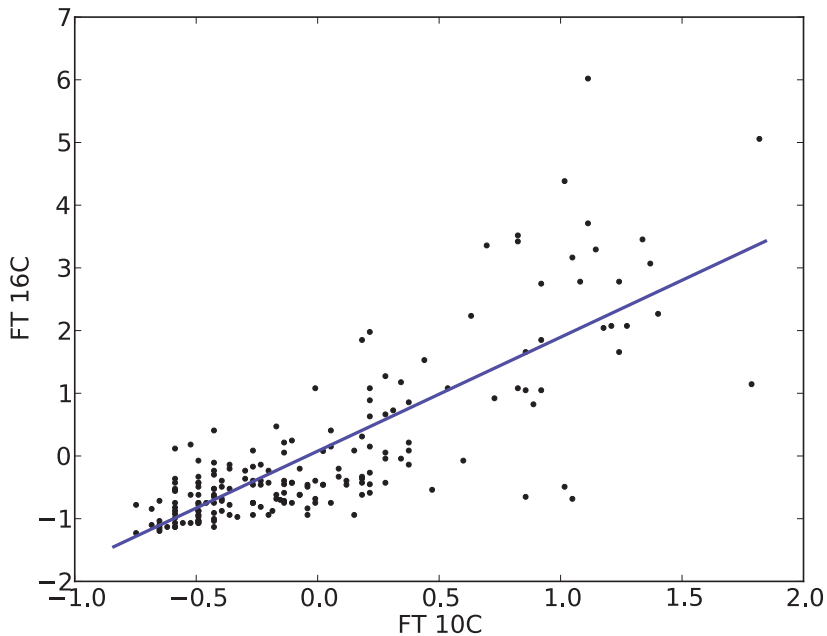
Outline

Overview

- Frequently, we are interested in the genetic architecture of related traits.
- Example: flowering time in 10C and 16C

Overview

- Frequently, we are interested in the genetic architecture of related traits.
- Example: flowering time in 10C and 16C



Multi-trait mixed models

- Extend variance component models to pairs of traits for environments $e \in \{0, 1\}$

$$\mathbf{y}_e = \underbrace{\mu_e}_{\text{env effect}} + \underbrace{\sum_{s=1}^{S} \mathbf{x}_s \beta_{s,e}}_{\text{genetic effects}} + \boldsymbol{\epsilon}_e.$$

- Again, prior is *shared* across SNPs but environment specific

$$p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_S) = \prod_{s=1}^{S} \mathcal{N}\left(\boldsymbol{\beta}_s \middle| \mathbf{0}, \begin{bmatrix} \sigma^2_{g,0} & \rho_{0,1} \\ \rho_{0,1} & \sigma^2_{g,1} \end{bmatrix}\right)$$

genetic variance and correlation.

- One noise level per environment

$$\boldsymbol{\epsilon}_0 \sim \mathcal{N}(0, \sigma^2_{e,0}\mathbf{I}) \;\; \boldsymbol{\epsilon}_1 \sim \mathcal{N}(0, \sigma^2_{e,1}\mathbf{I}).$$

Multi-trait mixed models

- Extend variance component models to pairs of traits for environments $e \in \{0, 1\}$

$$\mathbf{y}_e = \underbrace{\mu_e}_{\text{env effect}} + \underbrace{\sum_{s=1}^{S} \mathbf{x}_s \beta_{s,e}}_{\text{genetic effects}} + \boldsymbol{\epsilon}_e.$$

- Again, prior is *shared* across SNPs but environment specific

$$p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_S) = \prod_{s=1}^{S} \mathcal{N}\left(\boldsymbol{\beta}_s \middle| \mathbf{0}, \begin{bmatrix} \sigma^2_{g,0} & \rho_{0,1} \\ \rho_{0,1} & \sigma^2_{g,1} \end{bmatrix}\right)$$

  genetic variance and correlation.

- One noise level per environment

$$\boldsymbol{\epsilon}_0 \sim \mathcal{N}(0, \sigma^2_{e,0}\mathbf{I}) \quad \boldsymbol{\epsilon}_1 \sim \mathcal{N}(0, \sigma^2_{e,1}\mathbf{I}).$$

Multi-trait mixed models

- Extend variance component models to pairs of traits for environments $e \in \{0, 1\}$

$$\mathbf{y}_e = \underbrace{\mu_e}_{\text{env effect}} + \underbrace{\sum_{s=1}^{S} \mathbf{x}_s \beta_{s,e}}_{\text{genetic effects}} + \boldsymbol{\epsilon}_e.$$

- Again, prior is *shared* across SNPs but environment specific

$$p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_S) = \prod_{s=1}^{S} \mathcal{N}\left(\boldsymbol{\beta}_s \middle| \mathbf{0}, \begin{bmatrix} \sigma^2_{g,0} & \rho_{0,1} \\ \rho_{0,1} & \sigma^2_{g,1} \end{bmatrix}\right)$$

  genetic variance and correlation.

- One noise level per environment

$$\boldsymbol{\epsilon}_0 \sim \mathcal{N}(0, \sigma^2_{e,0}\mathbf{I}) \;\; \boldsymbol{\epsilon}_1 \sim \mathcal{N}(0, \sigma^2_{e,1}\mathbf{I}).$$

Multi-trait mixed models

Marginalized multi trait variance component model

- Because of the Gaussian assumption, again noise and weights can be marginalized out analytically

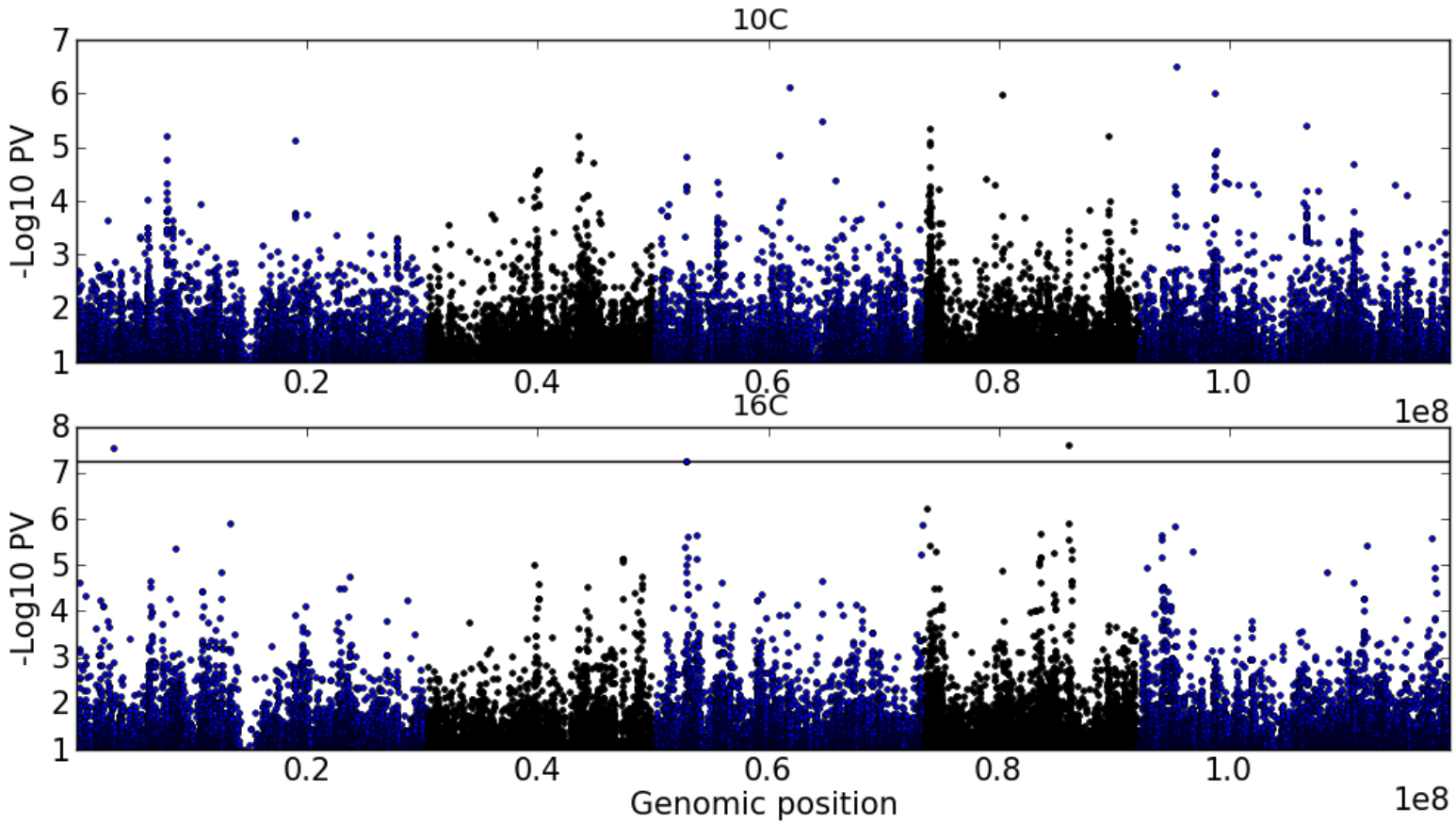
$$p\left(\mathbf{y}_0, \mathbf{y}_1 \,|\, \mathbf{X}, \sigma^2_{g,0}, \sigma^2_{g,1}, \dots\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{y}_0 \\ \mathbf{y}_1\end{bmatrix} \middle| \begin{matrix}\mu_0\mathbf{I} \\ \mu_1\mathbf{I}\end{matrix}, \begin{bmatrix}\sigma^2_{g,0}\mathbf{K}_g & \rho_{0,1}\mathbf{K}_g \\ \rho_{0,1}\mathbf{K}_g & \sigma^2_{g,1}\mathbf{K}_g\end{bmatrix} + \begin{bmatrix}\sigma^2_{e,0}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma^2_{e,1}\mathbf{I}\end{bmatrix}\right)$$

As before, $\mathbf{K}_g = \sum_{s=1}^{S} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}}$ denotes the genotype covariance.

Multi-trait mixed models

Illustration on two related *A. thaliana* traits

Independent single-marker GWAs on flowering time (10C and 16C)



Multi-trait mixed models

Illustration on two related *A. thaliana* traits

Mixed model inference

- Genotype covariance

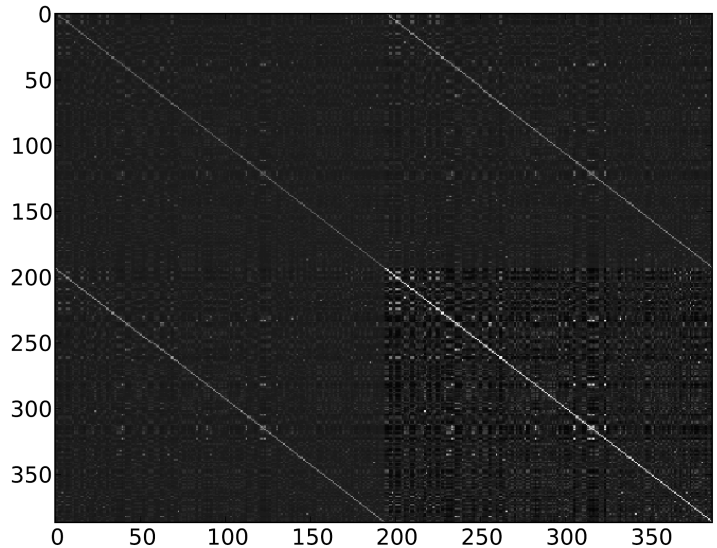
$$\begin{pmatrix} \sigma^2_{g,0} = 0.2 & \sigma_{0,1} = 0.33 \\ \sigma_{0,1} = 0.33 & \sigma^2_{g,1} = 0.92 \end{pmatrix}$$

- Noise covariance

$$\begin{pmatrix} \sigma^2_{e,0} = 0.03 & 0 \\ 0 = 0.33 & \sigma^2_{e,1} = 0.02 \end{pmatrix}$$

- Marginal heritabilities from joint analysis

$$h_0 = 0.88 \quad h_1 = 0.95$$



Multi-trait mixed models

Illustration on two related *A. thaliana* traits

Mixed model inference

- Genotype covariance

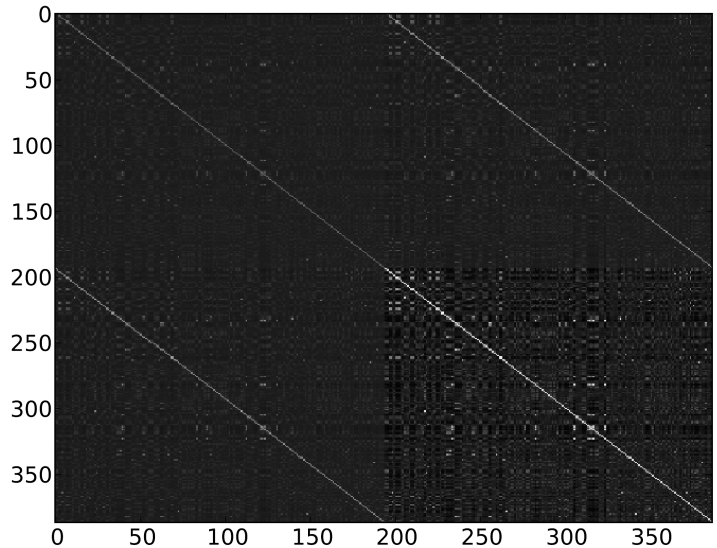
$$\begin{pmatrix} \sigma^2_{g,0} = 0.2 & \sigma_{0,1} = 0.33 \\ \sigma_{0,1} = 0.33 & \sigma^2_{g,1} = 0.92 \end{pmatrix}$$

- Noise covariance

$$\begin{pmatrix} \sigma^2_{e,0} = 0.03 & 0 \\ 0 = 0.33 & \sigma^2_{e,1} = 0.02 \end{pmatrix}$$

- Marginal heritabilities from joint analysis

$$h_0 = 0.88 \quad h_1 = 0.95$$



Multi-trait mixed models

Illustration on two related *A. thaliana* traits

Mixed model inference

- Genotype covariance

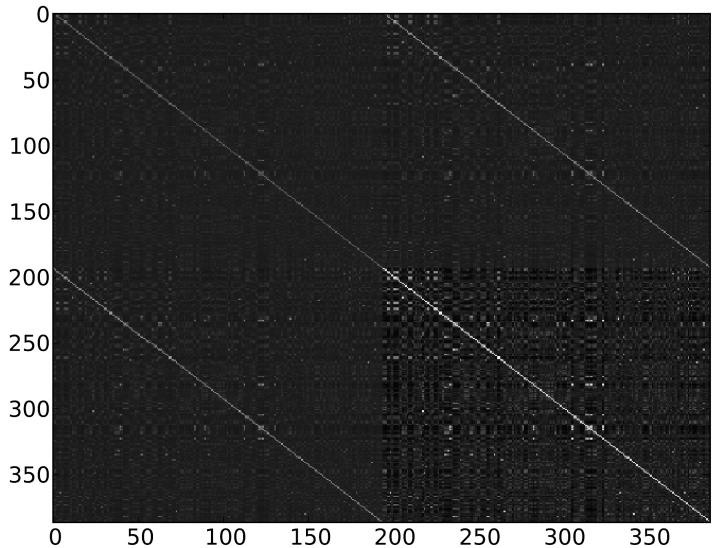
$$\begin{pmatrix} \sigma^2_{g,0} = 0.2 & \sigma_{0,1} = 0.33 \\ \sigma_{0,1} = 0.33 & \sigma^2_{g,1} = 0.92 \end{pmatrix}$$

- Noise covariance

$$\begin{pmatrix} \sigma^2_{e,0} = 0.03 & 0 \\ 0 = 0.33 & \sigma^2_{e,1} = 0.02 \end{pmatrix}$$

- Marginal heritabilities from joint analysis

$$h_0 = 0.88 \quad h_1 = 0.95$$



Multi-trait mixed models

- Multi-trait models can be used to control for population structure in a multi-environment setting
- Common association test affecting both traits

$$\mathcal{N}\left(\begin{bmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \end{bmatrix} \mid \underbrace{\begin{bmatrix} \mu_0 \mathbf{I} \\ \mu_1 \mathbf{I} \end{bmatrix}}_{\text{env effect}} + \underbrace{\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_s \end{bmatrix} \beta_s}_{\text{SNP effect}}, \mathbf{K}_{GxE} \right).$$

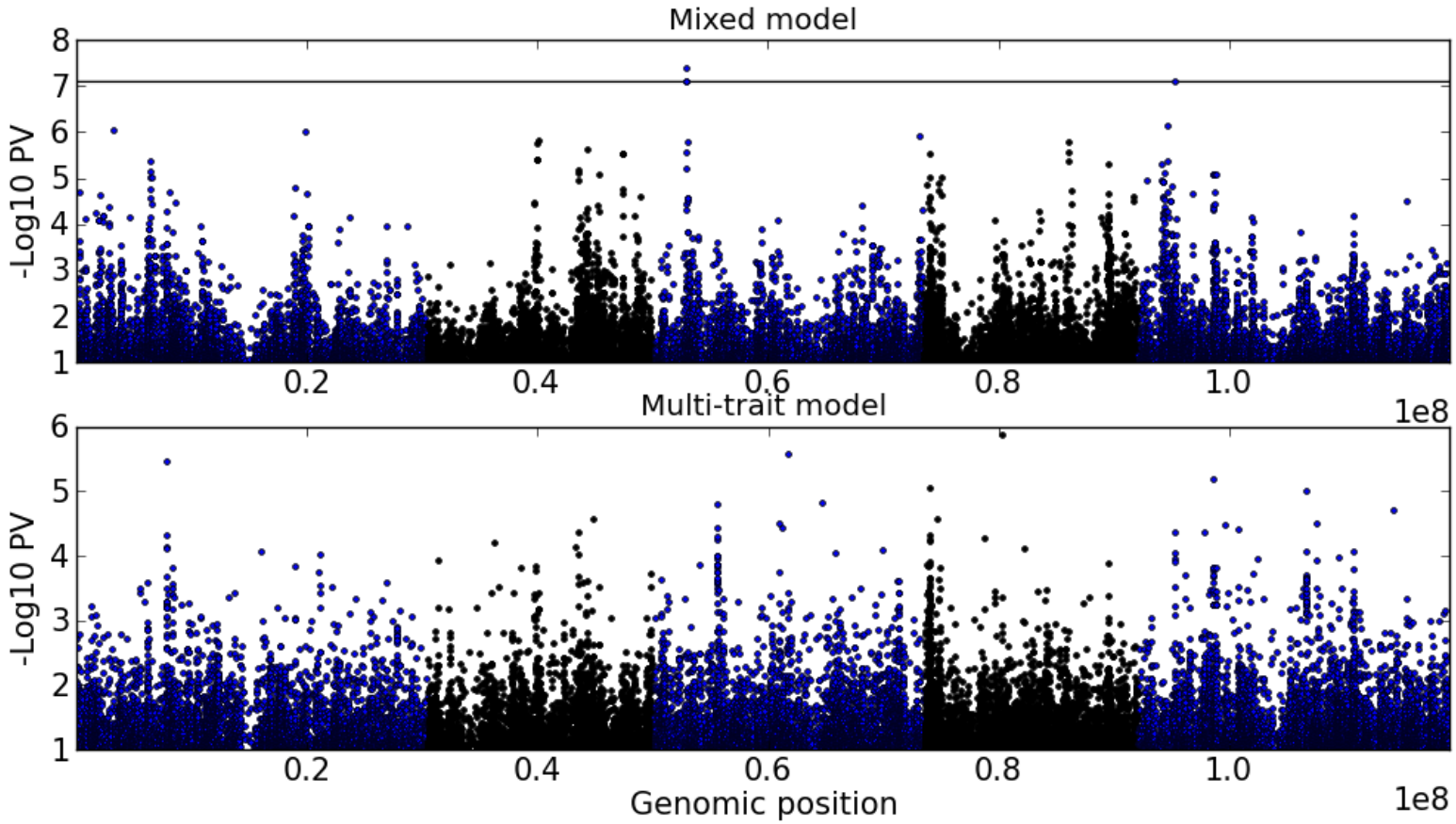
- Interaction test, specific for one environment

$$\mathcal{N}\left(\begin{bmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \end{bmatrix} \mid \underbrace{\begin{bmatrix} \mu_0 \mathbf{I} \\ \mu_1 \mathbf{I} \end{bmatrix}}_{\text{env effect}} + \underbrace{\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_s \end{bmatrix} \beta_s}_{\text{SNP effect}} + \underbrace{\begin{bmatrix} \mathbf{x}_s \\ \mathbf{0} \end{bmatrix} \beta_s^I}_{\text{GxE effect}}, \mathbf{K}_{GxE} \right).$$

where $\mathbf{K}_{GxE} = \begin{bmatrix} \sigma^2_{g,0}\mathbf{K}_g & \rho_{0,1}\mathbf{K}_g \\ \rho_{0,1}\mathbf{K}_g & \sigma^2_{g,1}\mathbf{K}_g \end{bmatrix} + \begin{bmatrix} \sigma^2_{e,0}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma^2_{e,1}\mathbf{I} \end{bmatrix}$

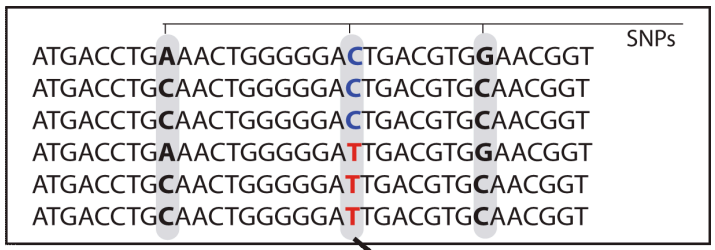
[Korte et al., 2012]

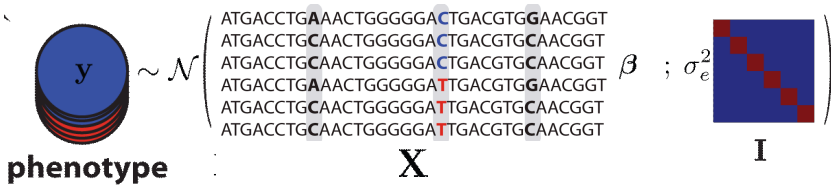
Common effect mapping on two traits

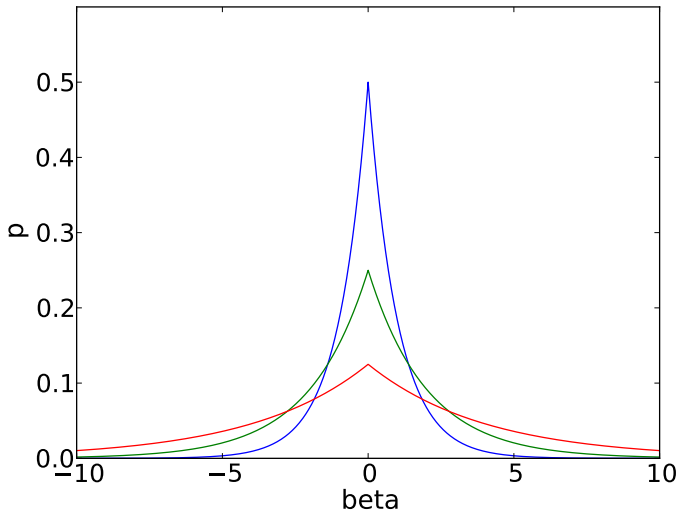


Outline

Simultaneous Analysis of all SNPs: LassoTibshirani [1996]



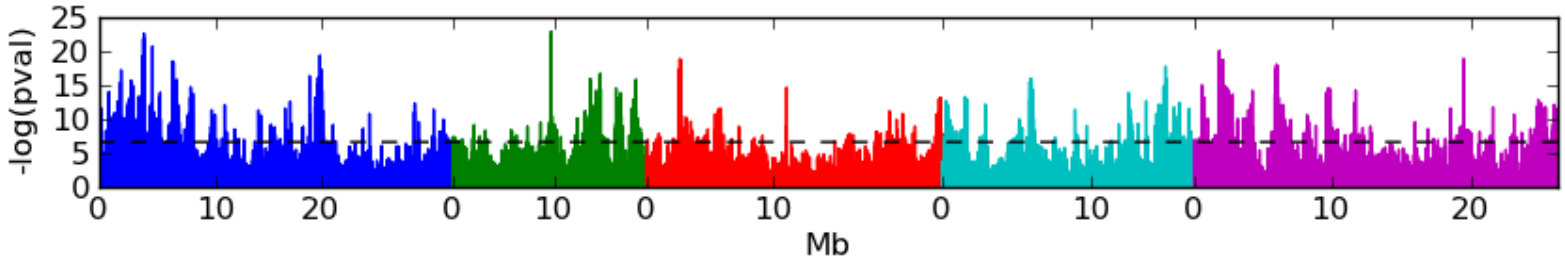




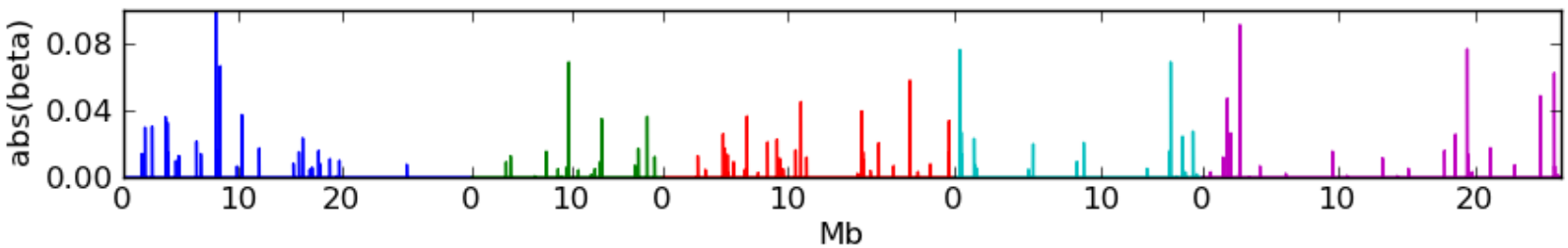
$$
\begin{aligned}
\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \\
\boldsymbol{\beta} &\sim p(\boldsymbol{\beta}|\lambda), \quad p(\boldsymbol{\beta}|\lambda) \propto \prod e^{-\lambda|\boldsymbol{\beta}_i|} \\
\boldsymbol{\epsilon} &\sim \mathcal{N}(0, \sigma_{\mathrm{e}}^2\mathbf{I})
\end{aligned}
$$

GWAS for Flowering Time in *Arabidopsis thaliana*

- Linear Model

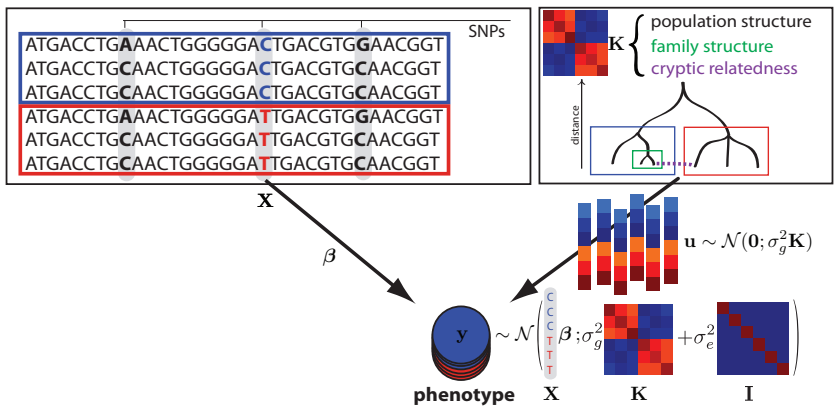


- Lasso



Linear Mixed Models

$$
\begin{aligned}
\mathbf{y} &= \mathbf{x}\beta + \mathbf{u} + \epsilon, \\
\mathbf{u} &\sim \mathcal{N}(0, \sigma_{\mathrm{g}}^2 \mathbf{K}) \\
\epsilon &\sim \mathcal{N}(0, \sigma_{\mathrm{e}}^2 \mathbf{I})
\end{aligned}
$$



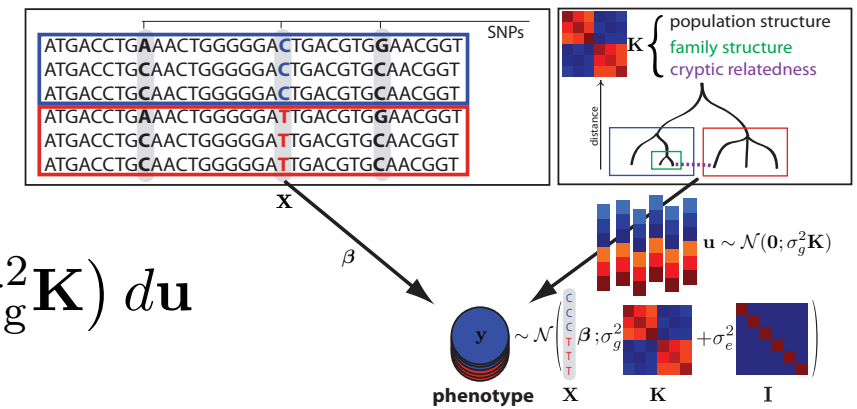
- Kinship Matrix $\mathbf{K}$: measures genetic similarity between pairs of samples

Linear Mixed Models

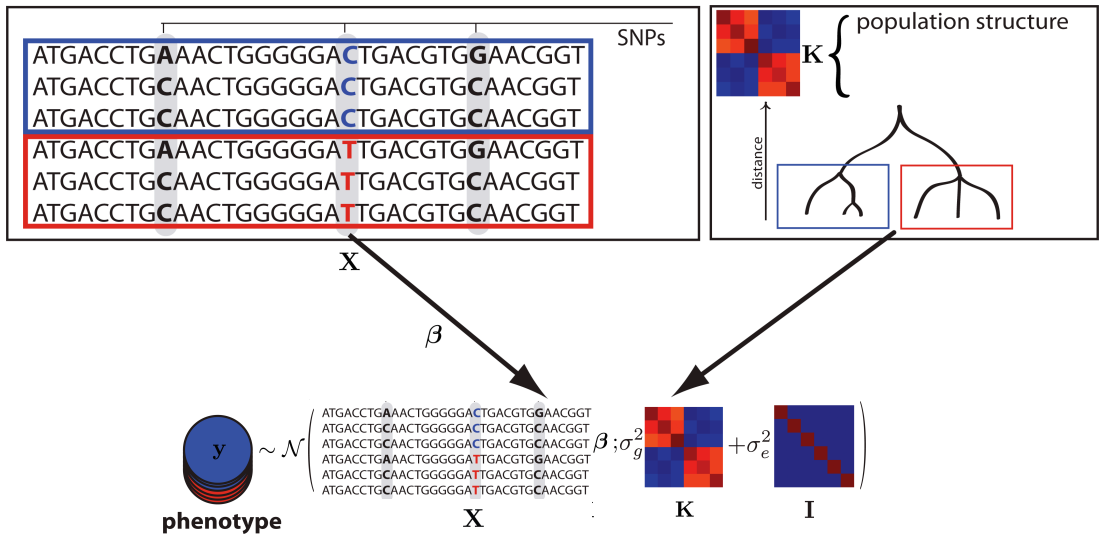
$$\mathbf{y} = \mathbf{x}\beta + \mathbf{u} + \boldsymbol{\epsilon},$$

Integrating over the random effects yields:

$$\begin{aligned} & \int_{\mathbf{u}} \mathcal{N}\left(\mathbf{y}|\mathbf{x}_s\beta_s + \mathbf{u}; \sigma_{\mathrm{e}}^2\mathbf{I}\right) \mathcal{N}\left(\mathbf{u}|\mathbf{0}; \sigma_{\mathrm{g}}^2\mathbf{K}\right) d\mathbf{u} \\ = \quad & \mathcal{N}\left(\mathbf{y}|\mathbf{x}_s\beta_s; \sigma_{\mathrm{g}}^2\mathbf{K} + \sigma_{\mathrm{e}}^2\mathbf{I}\right) \end{aligned}$$



LMM-Lasso



$$
\begin{aligned}
\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\epsilon}, \\
\boldsymbol{\beta} &\sim p(\boldsymbol{\beta}|\lambda), \quad p(\boldsymbol{\beta}|\lambda) \propto \prod e^{-\lambda|\boldsymbol{\beta}_i|} \\
\mathbf{u} &\sim \mathcal{N}(0, \sigma_{\mathrm{g}}^2 \mathbf{K}) \\
\boldsymbol{\epsilon} &\sim \mathcal{N}(0, \sigma_{\mathrm{e}}^2 \mathbf{I})
\end{aligned}
$$

[Rakitsch et al., 2012]

Inference

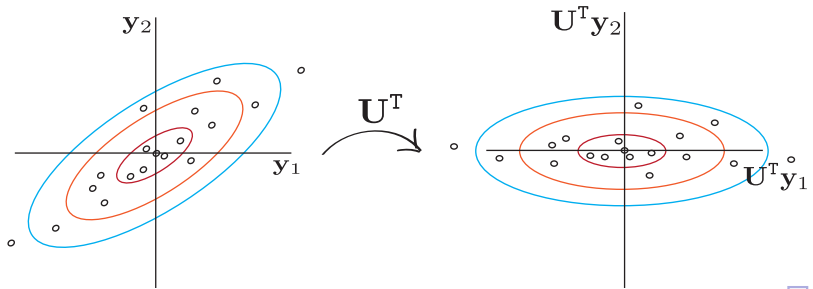
Minimize the objective:

$$-\log \mathcal{N}\left(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}; \sigma_{\mathrm{g}}^2\mathbf{K} + \sigma_{\mathrm{e}}^2\mathbf{I}\right) + \lambda\|\boldsymbol{\beta}\|_1$$

Algorithm:

1. Fix $\delta = \sigma_{\mathrm{e}}^2/\sigma_{\mathrm{g}}^2$ on the null model:

$$\begin{aligned} -\log \mathcal{N}\left(\mathbf{y}|\mathbf{0}; \sigma_{\mathrm{g}}^{\mathbf{2}}(\mathbf{K} + \delta\mathbf{I})\right) &= -\log \mathcal{N}\left(\mathbf{y}|\mathbf{0}; \sigma_{\mathrm{g}}^{\mathbf{2}}(\mathbf{U}\mathbf{S}\mathbf{U}^{\mathbf{T}} + \delta\mathbf{I})\right) \\ &= -\log \mathcal{N}\left(\mathbf{U}^{\mathbf{T}}\mathbf{y}|\mathbf{0}; \sigma_{\mathrm{g}}^{\mathbf{2}}(\mathbf{S} + \delta\mathbf{I})\right) \\ &= \frac{1}{\sigma_{\mathrm{g}}^2}(\mathbf{U}^{\mathbf{T}}\mathbf{y})^{\mathbf{T}}(\mathbf{S} + \delta\mathbf{I})^{-\mathbf{1}}(\mathbf{U}^{\mathbf{T}}\mathbf{y}) \end{aligned}$$



Inference (continued)

Minimize the objective:

$$-\log \mathcal{N}\left(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}; \sigma_{\mathrm{g}}^2(\mathbf{K}+\delta\mathbf{I})\right) + \lambda\|\boldsymbol{\beta}\|_1$$

Algorithm:

2. Train $\boldsymbol{\beta}$:

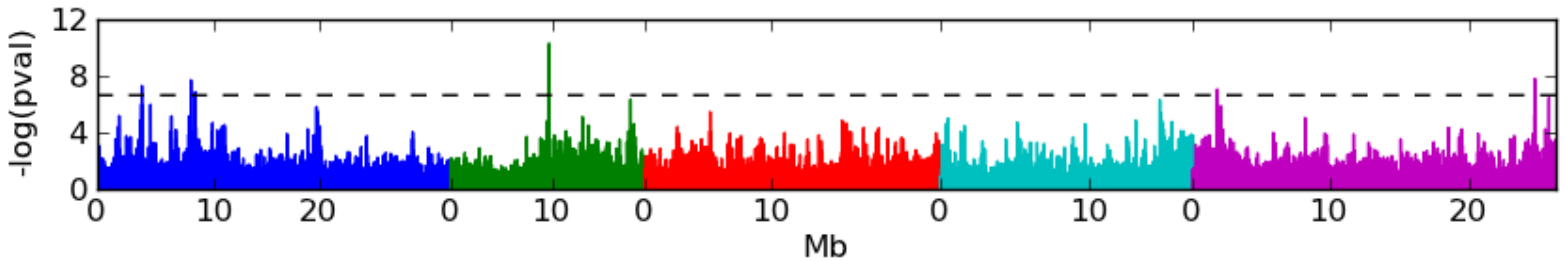
$$\begin{aligned} & \mathrm{argmin}_{\boldsymbol{\beta}} \quad (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\sigma_{\mathrm{g}}^2(\mathbf{K}+\delta\mathbf{I}))^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1 \\ = \quad & \mathrm{argmin}_{\boldsymbol{\beta}} \quad \frac{1}{\sigma_{\mathrm{g}}^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{U}\mathbf{S}\mathbf{U}^T+\delta\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1 \\ = \quad & \mathrm{argmin}_{\boldsymbol{\beta}} \quad \frac{1}{\sigma_{\mathrm{g}}^2}(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta})^T(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1, \end{aligned}$$

where

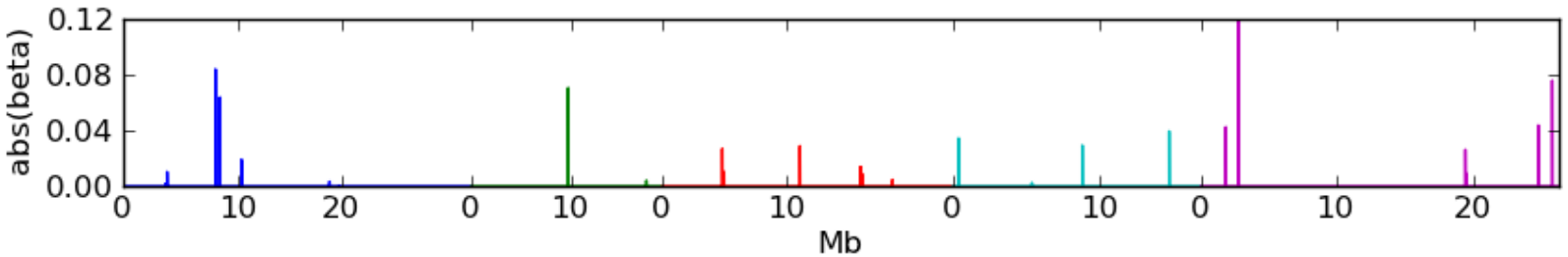
$$\begin{aligned} \tilde{\mathbf{X}} &= (\mathbf{S}+\delta\mathbf{I})^{-\frac{1}{2}}\mathbf{U}^{\mathbf{T}}\mathbf{X} \\ \tilde{\mathbf{y}} &= (\mathbf{S}+\delta\mathbf{I})^{-\frac{1}{2}}\mathbf{U}^{\mathbf{T}}\mathbf{y} \end{aligned}$$

GWAS for Flowering Time in *Arabidopsis thaliana*

- Linear Mixed Model

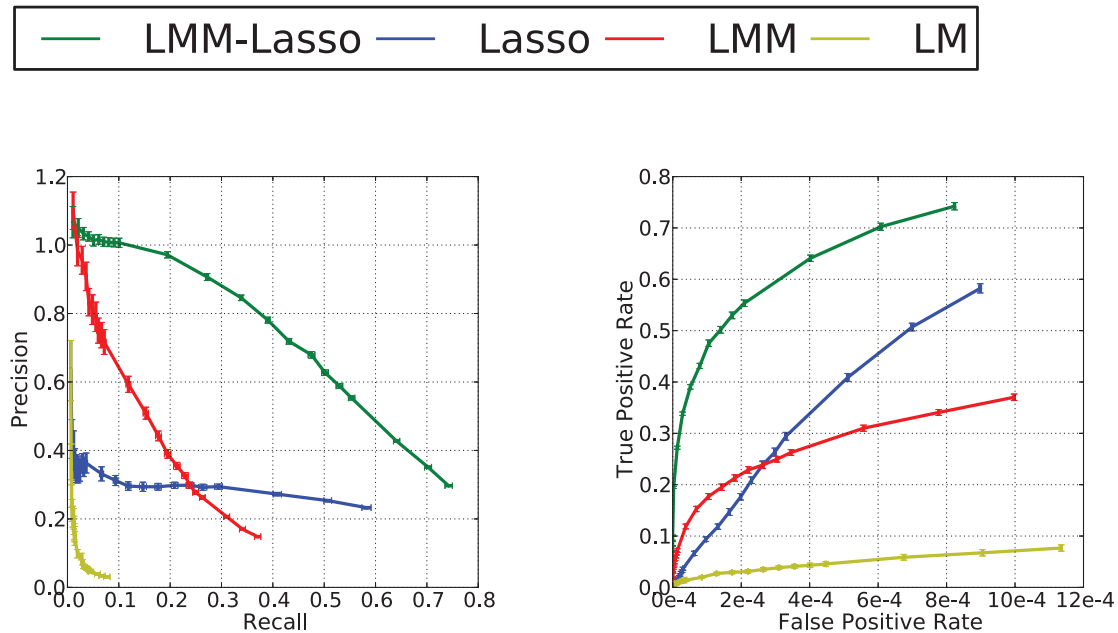


- LMM - Lasso



Power comparison

- Power comparison on semi-empirical data
  - LASSO models perform better than univariate testing.
  - Combining mixed models with LASSO outperforms other models.

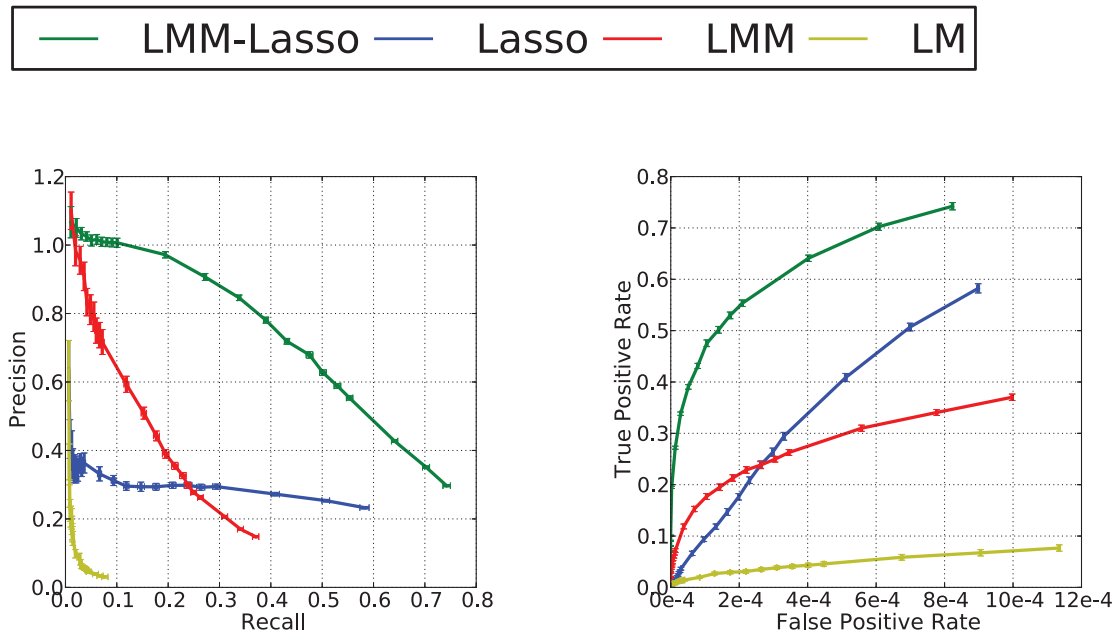


(a) Precision/Recall

(b) ROC

Power comparison

- Power comparison on semi-empirical data
  - LASSO models perform better than univariate testing.
  - Combining mixed models with LASSO outperforms other models.

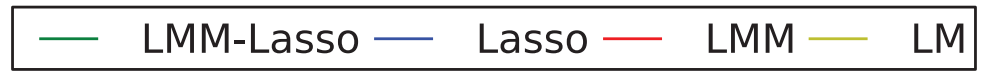


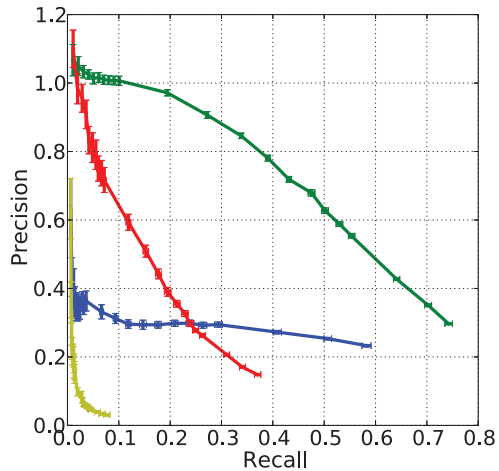
(a) Precision/Recall

(b) ROC

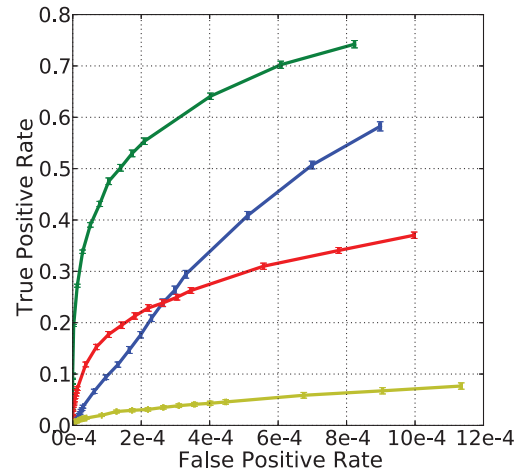
Power comparison

- Power comparison on semi-empirical data
  - LASSO models perform better than univariate testing.
  - Combining mixed models with LASSO outperforms other models.





(a) Precision/Recall

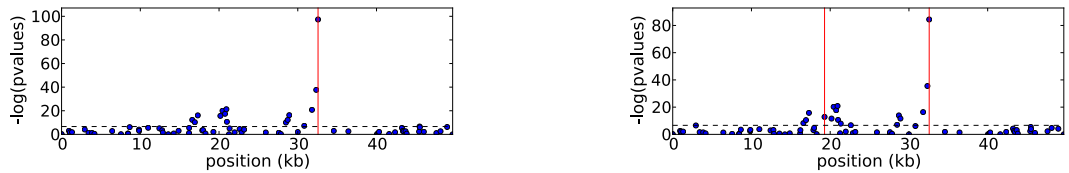


(b) ROC

Detection of multiple linked causal variants.pdf

- Two causal variants in linkage
  - LMM-Lasso better differentiates LD from genetic heterogeneity.

(a) Linear Model



(b) Linear Mixed Model

(c) Lasso

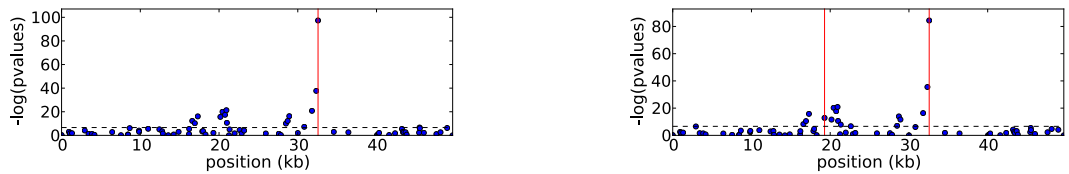
(d) LMM-Lasso

Figure 3

Detection of multiple linked causal variants.pdf

- Two causal variants in linkage
  - LMM-Lasso better differentiates LD from genetic heterogeneity.

(a) Linear Model



(b) Linear Mixed Model

(c) Lasso

(d) LMM-Lasso

Figure 3

Phenotype prediction with the LMM-Lasso

- Given the phenotype values of a set of individuals and the genetic relatedness, we can predict the genetic component of the phenotype of a new individual.
- $P(\mathbf{y}^\star \mid \mathbf{y}) = \mathcal{N}\left(\mathbf{y}^\star \mid \boldsymbol{\mu}^\star, \sigma_{\mathrm{g}}^2 \mathbf{V}^\star + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)$
- Predicitve mean: $\boldsymbol{\mu}^\star = \underbrace{\mathbf{X}^\star \boldsymbol{\beta}}_{\text{LASSO component}} + \underbrace{\mathbf{K}^{\star,:}\left(\mathbf{K} + \delta \mathbf{I}\right)^{-1}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)}_{\text{BLUP component}}$
- Predictive Variance: $\mathbf{V}^\star = \mathbf{K}^{\star,\star} - \mathbf{K}^{\star,:}\left(\mathbf{K} + \delta \mathbf{I}\right)^{-1} \mathbf{K}^{:,\star}$

Phenotype prediction with the LMM-Lasso

- Given the phenotype values of a set of individuals and the genetic relatedness, we can predict the genetic component of the phenotype of a new individual.
- $P(\mathbf{y}^\star \,|\, \mathbf{y}) = \mathcal{N}\left(\mathbf{y}^\star \,\middle|\, \boldsymbol{\mu}^\star, \sigma_{\mathrm{g}}^2 \mathbf{V}^\star + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)$
- Predicitve mean: $\mu^\star = \underbrace{\mathbf{X}^\star \beta}_{\text{LASSO component}} + \underbrace{\mathbf{K}^{\star,:} \left(\mathbf{K} + \delta \mathbf{I}\right)^{-1} \left(\mathbf{y} - \mathbf{X}\beta\right)}_{\text{BLUP component}}$
- Predictive Variance: $\mathbf{V}^\star = \mathbf{K}^{\star,\star} - \mathbf{K}^{\star,:} \left(\mathbf{K} + \delta \mathbf{I}\right)^{-1} \mathbf{K}^{:,\star}$

Phenotype prediction with the LMM-Lasso

- Given the phenotype values of a set of individuals and the genetic relatedness, we can predict the genetic component of the phenotype of a new individual.
- $P(\mathbf{y}^\star \,|\, \mathbf{y}) = \mathcal{N}\left(\mathbf{y}^\star \,\middle|\, \boldsymbol{\mu}^\star, \sigma_{\mathrm{g}}^2 \mathbf{V}^\star + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)$
- Predicitve mean: $\boldsymbol{\mu}^\star = \underbrace{\mathbf{X}^\star \boldsymbol{\beta}}_{\text{LASSO component}} + \underbrace{\mathbf{K}^{\star,:} \left(\mathbf{K} + \delta \mathbf{I}\right)^{-1} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)}_{\text{BLUP component}}$
- Predictive Variance: $\mathbf{V}^\star = \mathbf{K}^{\star,\star} - \mathbf{K}^{\star,:} \left(\mathbf{K} + \delta \mathbf{I}\right)^{-1} \mathbf{K}^{:,\star}$

Phenotype prediction with the LMM-Lasso

- Given the phenotype values of a set of individuals and the genetic relatedness, we can predict the genetic component of the phenotype of a new individual.
- $P(\mathbf{y}^\star \,|\, \mathbf{y}) = \mathcal{N}\left(\mathbf{y}^\star \,\middle|\, \boldsymbol{\mu}^\star, \sigma_{\mathrm{g}}^2 \mathbf{V}^\star + \sigma_{\mathrm{e}}^2 \mathbf{I}\right)$
- Predicitve mean: $\boldsymbol{\mu}^\star = \underbrace{\mathbf{X}^\star \boldsymbol{\beta}}_{\text{LASSO component}} + \underbrace{\mathbf{K}^{\star,:} \left(\mathbf{K} + \delta \mathbf{I}\right)^{-1} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)}_{\text{BLUP component}}$
- Predictive Variance: $\mathbf{V}^\star = \mathbf{K}^{\star,\star} - \mathbf{K}^{\star,:} \left(\mathbf{K} + \delta \mathbf{I}\right)^{-1} \mathbf{K}^{:,\star}$

Phenotype prediction with the LMM-Lasso

- Out of sample prediction on FT in *A. thaliana*

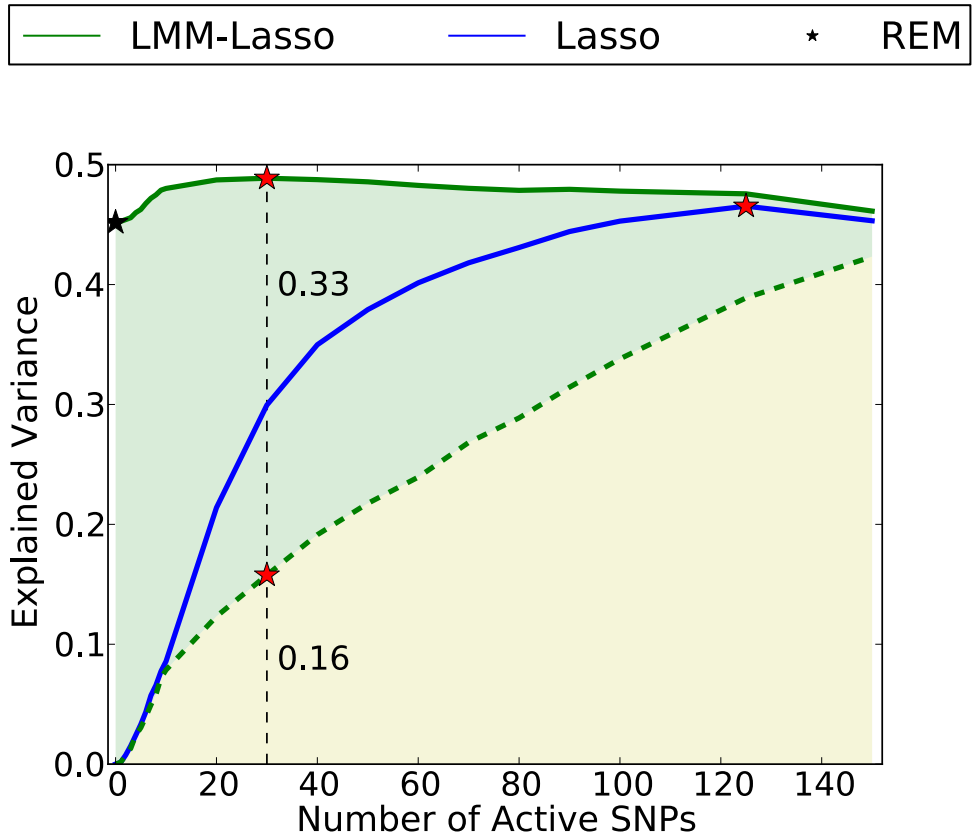
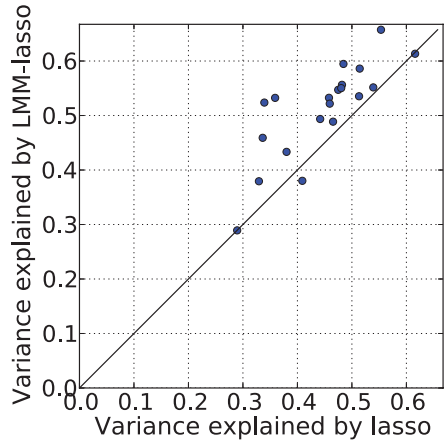


Figure 5

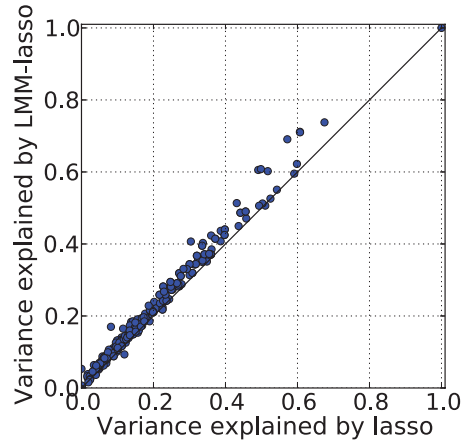
Phenotype prediction with the LMM-Lasso

Summary on a compendium of phenotypes

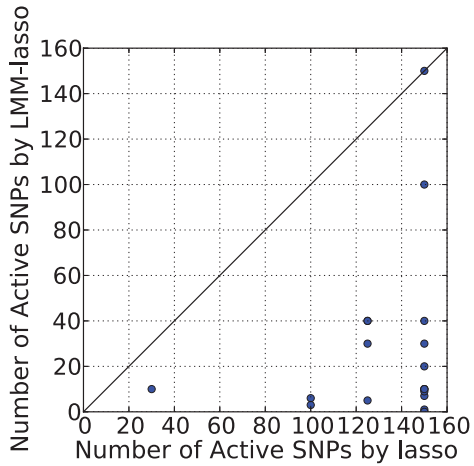
- Out of sample predictions on *A. thaliana* and mouse.



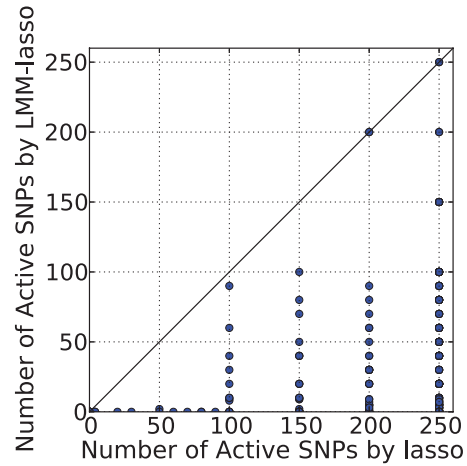
(a) Arabidopsis test variance



(b) Mouse test variance



(c) Arabidopsis number of SNPs



(d) Mouse number of SNPs

Summary

- Joint modeling of multiple SNPs is compromised by sample size.
- Solutions: shrinkage (LASSO) or marginalization.
- Variance component models allow for estimating the proportion of genetic variance.
- Genetic co-regulation causes phenotype correlation which can be incorporated into variance component models.
- Shrinkage-based LASSO models and variance component modeling can be combined, significantly improving phenotype prediction.

Summary

- Joint modeling of multiple SNPs is compromised by sample size.
- Solutions: shrinkage (LASSO) or marginalization.
- Variance component models allow for estimating the proportion of genetic variance.
- Genetic co-regulation causes phenotype correlation which can be incorporated into variance component models.
- Shrinkage-based LASSO models and variance component modeling can be combined, significantly improving phenotype prediction.

Summary

- Joint modeling of multiple SNPs is compromised by sample size.
- Solutions: shrinkage (LASSO) or marginalization.
- Variance component models allow for estimating the proportion of genetic variance.
- Genetic co-regulation causes phenotype correlation which can be incorporated into variance component models.
- Shrinkage-based LASSO models and variance component modeling can be combined, significantly improving phenotype prediction.

Summary

- Joint modeling of multiple SNPs is compromised by sample size.
- Solutions: shrinkage (LASSO) or marginalization.
- Variance component models allow for estimating the proportion of genetic variance.
- Genetic co-regulation causes phenotype correlation which can be incorporated into variance component models.
- Shrinkage-based LASSO models and variance component modeling can be combined, significantly improving phenotype prediction.

Summary

- Joint modeling of multiple SNPs is compromised by sample size.
- Solutions: shrinkage (LASSO) or marginalization.
- Variance component models allow for estimating the proportion of genetic variance.
- Genetic co-regulation causes phenotype correlation which can be incorporated into variance component models.
- Shrinkage-based LASSO models and variance component modeling can be combined, significantly improving phenotype prediction.

Acknowledgements

- **Mixed model Lasso**
  Barbara Rakitsch, Karsten Borgwardt

References I

A. Korte, B. Vilhjálmsson, V. Segura, A. Platt, Q. Long, and M. Nordborg. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*, 44(9):1066–1071, 2012.

B. Rakitsch, C. Lippert, O. Stegle, and K. Borgwardt. Lmm-lasso: A lasso multi-marker mixed model for association mapping with population structure correction. *Arxiv preprint arXiv:1205.6986*, 2012.

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

M. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 2011.