Mapping of high-dimensional traits: eQTL and beyond

Christoph Lippert¹ Oliver Stegle²

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



Basel 09. September 2012



C. Lippert & O. Stegle

Mapping high dimensional traits

September 2012 1

イロト イポト イヨト イヨト

Models of molecular variation Motivation

Goal:

- Dissect genetic GWAS signals
- Improve predictive models of quantitative traits
- Understand genetic architecture

DNA	ATGACCTGAAACTGGGGGATGACGTGACGGT ATGACCTGCACTGGGGGGGGATGACGTGCAACGGT ATGACCTGCACTGGGGGGTGACGTGCAACGGT ATGACCTGCACTGGGGGGTGACGGGATGACGTGCAACGGT ATGACCTGCAACTGGGGGATGACGTGCAACGGT ATGACCTGCAACTGGGGGATGACGTGCAACGGT	SNPs		
trar	nscription			
mRNA	$(\mathbf{y}_{1,\cdot})$ $(y_{2,\cdot})$ $(\mathbf{y}_{3,\cdot})$ $(\mathbf{y}_{G,\cdot})$			
tra	Inslation			
proteins	s $(\overline{y}_{1,\cdot})(\overline{y}_{2,\cdot})(\overline{y}_{3,\cdot})$ $(\overline{y}_{G,\cdot})$			
organ-level •••				

イロト イポト イヨト イヨト

Models of molecular variation Motivation

Goal:

- Dissect genetic GWAS signals
- Improve predictive models of quantitative traits
- Understand genetic architecture

DNA	ATGACCTGAAACTGGGGGACTGACG ATGACCTGCAACTGGGGGACTGACG ATGACCTGCAACTGGGGGACTGACG ATGACCTGCAAACTGGGGGATTGACG ATGACCTGCAACTGGGGGATTGACG ATGACCTGCAACTGGGGGATTGACG	SNPs TGGAACGGT TGCAACGGT TGGAACGGT TGGAACGGT TGCAACGGT TGCAACGGT	\backslash	
tra	anscription			
mRNA	y _{1,} , y _{2,} y _{3,}	yg,)		GWAS
proteir	anslation ans $(\overline{y}_{1,\cdot})(\overline{y}_{2,\cdot})(\overline{y}_{3,\cdot})$	(yG,·)	/	/
organ- pheno	level types •••			

<ロ> (日) (日) (日) (日) (日)

Models of molecular variation Motivation

Goal:

- Dissect genetic GWAS signals
- Improve predictive models of quantitative traits
- Understand genetic architecture



Models of molecular variation Motivation

Goal:

- Dissect genetic GWAS signals
- Improve predictive models of quantitative traits
- Understand genetic architecture



Models of molecular variation eQTL



C. Lippert & O. Stegle

Mapping high dimensional traits

・ < ≣ ▶ ≣ ∽ < @ September 2012 3

(日) (周) (日) (日)

Models of molecular variation eQTL

Statistical challenges:

- Large-scale
 (N << p regime)
- Millions of tests
- Limited power

DNA	ATGACCTGAAAC ATGACCTGCAAC ATGACCTGCAAC ATGACCTGCAAC ATGACCTGCAAC ATGACCTGCAAC	TGGGGGACTGACC TGGGGGACTGACC TGGGGGACTGACC TGGGGGATTGACC TGGGGGATTGACC TGGGGGATTGACC	GGGAACGGT GGCAACGGT GGGAACGGT GGGAACGGT GGCAACGGT GGCAACGGT	SNPs	
trar	nscription				eQT
mRNA			¥G,	/	
tra		¥			
proteins	s (<u>ÿ</u> 1,·)(<u>ÿ</u>	$\overline{y}_{3,\cdot}$	$\overline{y}_{G,\cdot}$		
organ-le	evel ypes	0 0 0			

イロト イポト イヨト イヨト

Models of molecular variation eQTL

Statistical challenges:

- Large-scale
 (N << p regime)
- Millions of tests
- Limited power

DNA	ATGACCTGAAAC ATGACCTGCAAC ATGACCTGCAAC ATGACCTGCAAC ATGACCTGCAAC ATGACCTGCAAC	TGGGGGACTGACG TGGGGGACTGACG TGGGGGACTGACG TGGGGGATTGACG TGGGGGATTGACG TGGGGGATTGACG	TGGAACGGT TGCAACGGT TGCAACGGT TGGAACGGT TGCAACGGT TGCAACGGT	SNPs
trai	nscription		7 /	eQT
mRNA	$\begin{array}{c} & \\ & \\ & \\ & \\ & \\ & \\ & \\ & \\ & \\ & $	y 2,.)(y _{3,.})	¥G,•	1
tra		¥		
protein	s (y 1,·) (y	$\overline{y}_{3,\cdot}$	$\overline{y}_{G,\cdot}$	
organ-l phenot				

イロト イポト イヨト イヨト

Models of molecular variation eQTL

Statistical challenges:

- Large-scale
 (N << p regime)
- Millions of tests
- Limited power

Bioinformatics challenges:

- Phenotyping
- Using NGS technologies



(日) (同) (三) (三)

September 2012 3

Outline

Outline

C. Lippert & O. Stegle

Mapping high dimensional traits

September 2012 4

◆□ → < @ → < 差 → < 差 → < 差 → の < ??</p>

Outline

Motivation

Accounting for background variation in eQTL studies

Mechanistic models: Genetic analyses with learnt cellular features

The role of GxE in the A. thaliana transcriptional landscape

Summary

C. Lippert & O. Stegle

Mapping high dimensional traits

イロト イポト イヨト イヨト

Models of molecular variation Univariate phenotypes, examples

Single-marker mapping using linear models:



SNPs n can either be proximal or distal to gene g

▲□▶ ▲□▶ ▲□▶ ▲□▶ □ ののの

Models of molecular variation

Univariate phenotypes, examples

Single-marker mapping using linear models:



• SNPs n can either be proximal or distal to gene g



C. Lippert & O. Stegle

Mapping high dimensional traits

Models of molecular variation

Univariate phenotypes, examples

Single-marker mapping using linear models:



C. Lippert & O. Stegle

Mapping high dimensional traits

- promoter, RNA stability. chromatin structure
- typically cis mechanisms



- regulatory protein, pathway
- hotspots
- typically trans mechanisms

Known and unknown confounding in genomic analyses

- Standard to model known factors
 - Population background
 - Gender
- It is key to account for unknown hidden factors as well
 - Sample preparation
 - Sample history



Known and unknown confounding in genomic analyses

- Standard to model known factors
 - Population background
 - Gender
- It is key to account for unknown hidden factors as well
 - Sample preparation
 - Sample history

ATGACCTGAAACTGGGGGATGACGGGT ATGACCTGGAACGGT ATGACCTGGAACTGGGGGATGACGGCA ATGACCTGGAACTGGGGATGACGGCATGAACGGT ATGACCTGAAACTGGGGATTGACGTGGAACGGT ATGACCTGAAACTGGGGATTGACGTGGAACGGT	$ \begin{array}{c} $
transcription	experimental procedures environment sample history
mRNA (1, (1, (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1	
translation 🖌	
proteins $(\overline{y}_1, \overline{y}_2, \overline{y}_3, $	
♥ organ-level	

イロト イポト イヨト イヨト

[Leek and Storey, 2007]

September 2012 7

Example, Human



イロト イヨト イヨト イヨト

Association model

PEER: accounting for hidden factors

- Start with standard association model.
- Include (few) global hidden factors (confounders) in the model.
- Factors $\mathbf{H} = {\mathbf{h}_1, \dots, \mathbf{h}_K}$ need to be learnt from the expression data.
- Controlling model complexity using hierarchical Bayesian modeling. $p(w_{g,k}, \alpha_k) = \mathcal{N}\left(w_{g,k} \middle| 0, \frac{1}{\alpha_k}\right) \Gamma(\alpha_k | a_k, b_k)$



[Stegle et al., 2010]

Mapping high dimensional traits

September 2012 9

▲□▶ ▲□▶ ▲□▶ ▲□▶ □ ののの

Association model

PEER: accounting for hidden factors

- Start with standard association model.
- Include (few) global hidden factors (confounders) in the model.



Mapping high dimensional traits

September 2012 9

[Stegle et al., 2010]

Association model

PEER: accounting for hidden factors

- Start with standard association model.
- Include (few) global hidden factors (confounders) in the model.
- Factors $\mathbf{H} = {\mathbf{h}_1, \dots, \mathbf{h}_K}$ need to be learnt from the expression data.



C. Lippert & O. Stegle

Mapping high dimensional traits

September 2012 9

[Stegle et al., 2010]

Association model

PEER: accounting for hidden factors

- Start with standard association model.
- Include (few) global hidden factors (confounders) in the model.
- Factors $\mathbf{H} = {\mathbf{h}_1, \dots, \mathbf{h}_K}$ need to be learnt from the expression data.
- Controlling model complexity using hierarchical Bayesian modeling. $p(w_{g,k}, \alpha_k) = \mathcal{N}\left(w_{g,k} \mid 0, \frac{1}{\alpha_k}\right) \Gamma(\alpha_k \mid a_k, b_k)$



C. Lippert & O. Stegle

Mapping high dimensional traits

September 2012 9

[Stegle et al., 2010]

Use cases



[Stegle et al., 2010]

Increased power

Similarly on yeast, mouse, A. thaliana

C. Lippert & O. Stegle

Mapping high dimensional traits

September 2012 10

- E - N

Image: A match a ma

Use cases



[Stegle et al., 2010]

- Increased power
- Similarly on yeast, mouse, A. thaliana

Mixed model implementation



 Exploit large number of expression traits to estimate the empirical covariance structure.

- Iterative learning on the covariance structure induced by all traits.
 L. Learn confounders, explaining broad covariance within expression profiles.
 - 2. Test for genetic (SNP) control of learnt confounders.
 - 3. Add relevant SNPs to the covariance structure.
- Add known confounding, e.g. population structure.
- Derive confounding covariance structure Σ for association testing

C. Lippert & O. Stegle

< ロト < 同ト < ヨト < ヨト

Mixed model implementation



 $+\psi$

noise

- Exploit large number of expression traits to estimate the empirical covariance structure.
- Iterative learning on the covariance structure induced by all traits.
 - 1. Learn confounders, explaining broad covariance within expression profiles.
 - 2. Test for genetic (SNP) control of learnt confounders.
 - 3. Add relevant SNPs to the covariance structure.
- Add known confounding, e.g. population structure.
- Derive confounding covariance structure Σ for association testing.

September 2012 11

[Fusi et al., 2012]

Mixed model implementation



- Exploit large number of expression traits to estimate the empirical covariance structure.
- Iterative learning on the covariance structure induced by all traits.
 - 1. Learn confounders, explaining broad covariance within expression profiles.
 - 2. Test for genetic (SNP) control of learnt confounders.
 - 3. Add relevant SNPs to the covariance structure.
- Add known confounding, e.g. population structure.
- Derive confounding covariance structure Σ for association testing.

Mixed model implementation



- Exploit large number of expression traits to estimate the empirical covariance structure.
- Iterative learning on the covariance structure induced by all traits.
 - 1. Learn confounders, explaining broad covariance within expression profiles.
 - 2. Test for genetic (SNP) control of learnt confounders.
 - 3. Add relevant SNPs to the covariance structure.
- Add known confounding, e.g. population structure.

• Derive confounding covariance structure Σ for association testing.

4 D N 4 A N 4 F N

[Fusi et al., 2012]

Mixed model implementation



- Exploit large number of expression traits to estimate the empirical covariance structure.
- Iterative learning on the covariance structure induced by all traits.
 - 1. Learn confounders, explaining broad covariance within expression profiles.
 - 2. Test for genetic (SNP) control of learnt confounders.
 - 3. Add relevant SNPs to the covariance structure.
- Add known confounding, e.g. population structure.
- \blacktriangleright Derive confounding covariance structure Σ for association testing.

September 2012 11

4 D N 4 B N 4 B N

[Fusi et al., 2012]

Mixed model implementation



Formally, the expression levels are independent given genotype, hidden factors and population structure.

$$P(\mathbf{Y} | \mathbf{X}, \mathbf{H},) = \prod_{g=1}^{G} \mathcal{N}\left(\mathbf{y}_{:,g} \left| \mathbf{0}, \sigma_{g}^{2} \sum_{s=1}^{S} \mathbf{x}_{s} \mathbf{x}_{s}^{\mathrm{T}} + \sigma_{h}^{2} \sum_{k=1}^{K} \mathbf{h}_{k} \mathbf{h}_{k}^{\mathrm{T}} + \mathbf{K}_{\mathsf{pop}} + \sigma_{e}^{2} \mathbf{I} \right)$$

C. Lippert & O. Stegle

(日) (同) (日) (日)

Mixed model implementation Testing strategies



Mixed model likelihood ratio

$$\mathsf{LOD}_{g,s} = \log \frac{\mathcal{N}\left(\mathbf{y}_{g} \mid \mathbf{x}_{s}\beta_{s,g}, \sigma_{g}^{2}\boldsymbol{\Sigma} + \sigma_{e}^{2}\mathbf{I}\right)}{\mathcal{N}\left(\mathbf{y}_{g} \mid \mathbf{0}, \sigma_{g}^{2}\boldsymbol{\Sigma} + \sigma_{e}^{2}\mathbf{I}\right)}$$

C. Lippert & O. Stegle

Mapping high dimensional traits

Use cases



[Fusi et al., 2012]

Increased power

Improved consistency between studies

Better calibrated test statistics



C. Lippert & O. Stegle

Mapping high dimensional traits

Use cases



[Fusi et al., 2012]

- Increased power
- Improved consistency between studies
- Better calibrated test statistics



Mapping high dimensional traits

Use cases



[Fusi et al., 2012]

- Increased power
- Improved consistency between studies
- Better calibrated test statistics



Summary

- Accounting for hidden factors can greatly increase the power and meaningfulness of analysis results.
- ▶ Open source **PEER** software package (Python, R, C++) [Stegle et al., 2012]

[Stegle et al., 2012]

1000 Genomes project

A map of human genome variation from population-scale sequencing Nature (Nature, 1000 genomes consortium 2010)

HapMap III expression analysis

Patterns of Cis Regulatory Variation in Diverse Human Populations, PLoS Genet 2012

Genome and transcriptome variation in Arabidopsis

Multiple reference genomes and transcriptomes for Arabidopsis (Nature, Gan* & Stegle* et al. 2011)

C. Lippert & O. Stegle

Mapping high dimensional traits

소리가 소문가 소문가 소문가 ...

Summary

- Accounting for hidden factors can greatly increase the power and meaningfulness of analysis results.
- ▶ Open source **PEER** software package (Python, R, C++) [Stegle et al., 2012]

[Stegle et al., 2012]

Applied use in a number of studies

1000 Genomes project

A map of human genome variation from population-scale sequencing Nature (Nature, 1000 genomes consortium 2010)

HapMap III expression analysis

Patterns of Cis Regulatory Variation in Diverse Human Populations, PLoS Genet 2012

• Genome and transcriptome variation in Arabidopsis

Multiple reference genomes and transcriptomes for Arabidopsis (Nature, Gan* & Stegle* et al. 2011)

C. Lippert & O. Stegle

Mapping high dimensional traits

September 2012 15

イロト 不得下 イヨト イヨト 二日

Summary

- Accounting for hidden factors can greatly increase the power and meaningfulness of analysis results.
- ▶ Open source **PEER** software package (Python, R, C++) [Stegle et al., 2012]

[Stegle et al., 2012]

Applied use in a number of studies

1000 Genomes project

A map of human genome variation from population-scale sequencing Nature (Nature, 1000 genomes consortium 2010)

HapMap III expression analysis

Patterns of Cis Regulatory Variation in Diverse Human Populations, PLoS Genet 2012

• Genome and transcriptome variation in Arabidopsis

Multiple reference genomes and transcriptomes for Arabidopsis (Nature, Gan* & Stegle* et al. 2011)



C. Lippert & O. Stegle

Mapping high dimensional traits
Outline

Motivation

Accounting for background variation in eQTL studies

Mechanistic models: Genetic analyses with learnt cellular features

The role of GxE in the A. thaliana transcriptional landscape

Summary

C. Lippert & O. Stegle

Mapping high dimensional traits

イロト イポト イヨト イヨト

Regulatory and external factors





Account for regulatory factors in transcription?



Regulatory and external factors





Account for regulatory factors in transcription?



Regulatory factors

- Regulatory factors:
 - Transcription factors
 - Pathway components
- Mechanistic hypothesis: regulatory factors mediate the association signals to target genes.
- Measuring T?
 - Difficult and expensive
- Learn the unobserved factors T



[Parts et al., 2011, Lee and Bussemaker, 2010]

(日) (同) (日) (日)

September 2012 18

Regulatory factors

- Regulatory factors:
 - Transcription factors
 - Pathway components
- Mechanistic hypothesis: regulatory factors mediate the association signals to target genes.
- Measuring T?
 - Difficult and expensive
- Learn the unobserved factors T



[Parts et al., 2011, Lee and Bussemaker, 2010]

(日) (同) (日) (日)

September 2012 18

Regulatory factors

- Regulatory factors:
 - Transcription factors
 - Pathway components
- Mechanistic hypothesis: regulatory factors mediate the association signals to target genes.
- Measuring T?
 - Difficult and expensive
- \blacktriangleright Learn the unobserved factors ${\bf T}$



[Parts et al., 2011, Lee and Bussemaker, 2010]

(日) (同) (三) (三)

September 2012 18

Regulatory factors A linear model of gene regulation

Inference of regulatory factors:

$$\underbrace{\mathbf{Y}_{J\cdot G}}_{\text{Expr.}} = \underbrace{\mathbf{T}_{J\cdot K}}_{\text{Factors}} \cdot \underbrace{\mathbf{W}_{K\cdot G}}_{\text{Weights}} + \underbrace{\mathbf{\Psi}_{J\cdot G}}_{\text{Noise}}.$$

- W is sparse; each factor regulates a specific subset of all genes.
- Incorporation of prior knowledge to render factors interpretable:
 - Transcription factor binding.



(日) (同) (三) (三)

Regulatory factors A linear model of gene regulation

Inference of regulatory factors:



- W is sparse; each factor regulates a specific subset of all genes.
- Incorporation of prior knowledge to render factors interpretable:
 - Transcription factor binding.



ATGACCTGAAACTGGGCGGATGACGTGGAACGGTATGACCTGAAACTGGGGGGACTGACGTGGAACGGTATGACCTGAAACT

< □ > < □ > < □ > < □ > < □ >

Sparse factor analysis Probabilistic model

• Graphical model $\mathbf{Y} = \mathbf{T} \cdot \mathbf{W} + \boldsymbol{\Psi}$.

Indicators z_{g,k} determine the sparsity pattern:

$$P(w_{g,k} | z_{g,k} = 0) = \mathcal{N}(w_{g,k} | 0, \sigma_0^2)$$

$$P(w_{g,k} | z_{g,k} = 1) = \mathcal{N}(w_{g,k} | 0, \sigma_1^2)$$



イロト イポト イヨト イヨト

Sparse factor analysis Probabilistic model

- Graphical model $\mathbf{Y} = \mathbf{T} \cdot \mathbf{W} + \boldsymbol{\Psi}$.
- Indicators z_{g,k} determine the sparsity pattern:

$$P(w_{g,k} | z_{g,k} = 0) = \mathcal{N}(w_{g,k} | 0, \sigma_0^2)$$
$$P(w_{g,k} | z_{g,k} = 1) = \mathcal{N}(w_{g,k} | 0, \sigma_1^2).$$



Sparse factor analysis Probabilistic model

- Graphical model $\mathbf{Y} = \mathbf{T} \cdot \mathbf{W} + \boldsymbol{\Psi}$.
- Indicators z_{g,k} determine the sparsity pattern:

$$P(w_{g,k} \mid z_{g,k} = 0) = \mathcal{N}\left(w_{g,k} \mid 0, \sigma_0^2\right)$$
$$P(w_{g,k} \mid z_{g,k} = 1) = \mathcal{N}\left(w_{g,k} \mid 0, \sigma_1^2\right)$$



Application to yeast Factor associations

Application to 108 yeast strains.

- Genotyped and expression profiled in 2 conditions.
- Prior knowledge: TF binding affinities.

Biological hypotheses

- 1. Genetic variation (SNPs) may regulate factor activations.
- Genotype-specific regulation of target genes.



(日) (同) (三) (三)

[Smith and Kruglyak, 2008]

Application to yeast Factor associations

Application to 108 yeast strains.

- Genotyped and expression profiled in 2 conditions.
- Prior knowledge: TF binding affinities.
- Biological hypotheses
 - 1. Genetic variation (SNPs) may regulate factor activations.
 - 2. Genotype-specific regulation of target genes.



(日) (同) (三) (三)

[Smith and Kruglyak, 2008]

Application to yeast Factor associations

- Application to 108 yeast strains.
 - Genotyped and expression profiled in 2 conditions.
 - Prior knowledge: TF binding affinities.
- Biological hypotheses
 - 1. Genetic variation (SNPs) may regulate factor activations.
 - 2. Genotype-specific regulation of target genes.



[Smith and Kruglyak, 2008]

Application to yeast Factor interactions

Example of genotype-specific factor regulation.



[Parts et al., 2011]

Image: A math a math

Application to yeast Factor interactions

Genome-wide interaction density.



Image: A math a math

[[]Parts et al., 2011]

Summary

- Accounting for hidden factors can greatly increase the power and meaningfulness of analysis results.
- Joint genetic analysis of cellular features and gene expression for improved interpretability.
- Open source PEER software package (Python, R, C++) http://github.com/PMBio/peer

[Stegle et al., 2012]

イロト 不得 トイヨト イヨト

Outline

Motivation

Accounting for background variation in eQTL studies

Mechanistic models: Genetic analyses with learnt cellular features

The role of GxE in the A. thaliana transcriptional landscape

Summary

C. Lippert & O. Stegle

Mapping high dimensional traits

イロト イポト イヨト イヨト

Swedish lines

- 160 Lines, extensive population structure
- Genome sequencing
- Transcriptome sequencing
- Bisulfite sequencing
- Two environments; 10C and 16C



Swedish lines

- 160 Lines, extensive population structure
- Genome sequencing
- Transcriptome sequencing
- Bisulfite sequencing
- Two environments; 10C and 16C



Image: A match a ma

Swedish lines

- 160 Lines, extensive population structure
- Genome sequencing
- Transcriptome sequencing
- Bisulfite sequencing
- Two environments; 10C and 16C



Image: A math a math

Swedish lines

- 160 Lines, extensive population structure
- Genome sequencing
- Transcriptome sequencing
- Bisulfite sequencing

Two environments; 10C and 16C



Image: A math a math

Swedish lines

- 160 Lines, extensive population structure
- Genome sequencing
- Transcriptome sequencing
- Bisulfite sequencing
- Two environments; 10C and 16C



Variance component analysis

A random effect variance estimation model

 \blacktriangleright Variance dissection of expression levels of gene g in environment $e=\{0,1\}$



Standard multi trait correlation model

cis & trans genotype prior

 $\mathbf{b}_{n} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \beta_{0}^{2} & \beta_{0,1} \\ \beta_{0,1} & \beta_{1}^{2} \end{bmatrix}\right) \quad \mathbf{d}_{n} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \delta_{0}^{2} & \delta_{0,1} \\ \delta_{0,1} & \delta_{1}^{2} \end{bmatrix}\right)$ correlated noise covariance $\mathcal{L}\left(-\begin{bmatrix} \sigma_{1}^{2} & 0 \end{bmatrix} - \mathbf{1}\right)$

Variance component analysis

A random effect variance estimation model

 \blacktriangleright Variance dissection of expression levels of gene g in environment $e=\{0,1\}$



- Standard multi trait correlation model
 - cis & trans genotype prior:

$$\mathbf{b}_{n} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \beta_{0}^{2} & \beta_{0,1} \\ \beta_{0,1} & \beta_{1}^{2} \end{bmatrix}\right) \quad \mathbf{d}_{n} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \delta_{0}^{2} & \delta_{0,1} \\ \delta_{0,1} & \delta_{1}^{2} \end{bmatrix}\right)$$

Uncorrelated noise covariance

$$oldsymbol{\psi}_g \sim \mathcal{N}\left(\mathbf{0}, egin{bmatrix} \sigma_0^2 & 0 \ 0 & \sigma_1^2 \end{bmatrix} \otimes \mathbf{I}
ight).$$

C. Lippert & O. Stegle

Mapping high dimensional traits

イロト 不得下 イヨト イヨト

Variance component analysis

A random effect variance estimation model

 \blacktriangleright Variance dissection of expression levels of gene g in environment $e=\{0,1\}$



- Standard multi trait correlation model
 - cis & trans genotype prior:

$$\mathbf{b}_n \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \beta_0^2 & \beta_{0,1} \\ \beta_{0,1} & \beta_1^2 \end{bmatrix}\right) \quad \mathbf{d}_n \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \delta_0^2 & \delta_{0,1} \\ \delta_{0,1} & \delta_1^2 \end{bmatrix}\right)$$

Uncorrelated noise covariance

$$\boldsymbol{\psi}_g \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix} \otimes \mathbf{I} \right).$$

C. Lippert & O. Stegle

Mapping high dimensional traits

イロト 不得下 イヨト イヨト

Variance component analysis

A random effect variance estimation model

 \blacktriangleright Variance dissection of expression levels of gene g in environment $e=\{0,1\}$



- Standard multi trait correlation model
 - cis & trans genotype prior:

$$\mathbf{b}_n \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \beta_0^2 & \beta_{0,1} \\ \beta_{0,1} & \beta_1^2 \end{bmatrix}\right) \quad \mathbf{d}_n \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \delta_0^2 & \delta_{0,1} \\ \delta_{0,1} & \delta_1^2 \end{bmatrix}\right)$$

Uncorrelated noise covariance

$$\boldsymbol{\psi}_{g} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_{0}^{2} & 0\\ 0 & \sigma_{1}^{2} \end{bmatrix} \otimes \mathbf{I} \right).$$

C. Lippert & O. Stegle

Mapping high dimensional traits

イロト 不得下 イヨト イヨト

Variance component analysis

Observational model needs to acknowledge count statistics

- \blacktriangleright The observed quantities are read counts $\mathbf{c}_{g,e}$ and not true expression levels
- Log-normal model on the Poisson rates
 - MCMC inference
 - Approximate Bayesian inferences
 - Variance stabilizing transform (Anscombe trans



Image: A image: A

Variance component analysis

Observational model needs to acknowledge count statistics

- \blacktriangleright The observed quantities are read counts $\mathbf{c}_{g,e}$ and not true expression levels
- Log-normal model on the Poisson rates
 - MCMC inference
 - Approximate Bayesian inference
 - ▶ Variance stabilizing transform (Anscombe transform



$$p(\mathbf{c}_{g,e} \,|\, \mathbf{C}(\boldsymbol{\theta})) = \mathcal{N}\left(\mathbf{y}_{g,e} \,|\, \mathbf{0}, \mathbf{C}(\boldsymbol{\theta})\right) \prod_{j=1}^{J} \underbrace{\operatorname{Poisson}(c_{g,e,j} \,|\, e^{y_{g,e,j}})}_{\operatorname{Poisson observation model}}$$

C. Lippert & O. Stegle

A B A B A
 A
 B
 A
 A
 B
 A
 A
 B
 A
 A
 B
 A
 A
 B
 A
 A
 B
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A

Variance component analysis

Observational model needs to acknowledge count statistics

- \blacktriangleright The observed quantities are read counts $\mathbf{c}_{g,e}$ and not true expression levels
- Log-normal model on the Poisson rates
 - MCMC inference
 - Approximate Bayesian inference
 - Variance stabilizing transform (Anscombe transform)



$$p(\mathbf{c}_{g,e} \,|\, \mathbf{C}(\boldsymbol{\theta})) = \mathcal{N}\left(\mathbf{y}_{g,e} \,|\, \mathbf{0}, \mathbf{C}(\boldsymbol{\theta})\right) \prod_{j=1}^{J} \underbrace{\operatorname{Poisson}(c_{g,e,j} \,|\, e^{y_{g,e,j}})}_{\operatorname{Poisson observation model}}$$

Image: A image: A

Variance component analysis

Observational model needs to acknowledge count statistics

- \blacktriangleright The observed quantities are read counts $\mathbf{c}_{g,e}$ and not true expression levels
- Log-normal model on the Poisson rates
 - MCMC inference
 - Approximate Bayesian inference
 - Variance stabilizing transform (Anscombe transform)



$$p(\mathbf{c}_{g,e} \mid \mathbf{C}(\boldsymbol{\theta})) = \mathcal{N}\left(\mathbf{y}_{g,e} \mid \mathbf{0}, \mathbf{C}(\boldsymbol{\theta})\right) \prod_{j=1}^{J} \underbrace{\frac{\mathsf{Poisson}(c_{g,e,j} \mid e^{y_{g,e,j}})}{\mathsf{Poisson observation model}}}_{\mathsf{Poisson} \left(\mathbf{y}_{g,e} \mid \mathbf{0}, \mathbf{C}(\boldsymbol{\theta})\right) \prod_{j=1}^{J} \underbrace{\mathcal{N}\left(y_{g,e,j} \mid \mu_{g,e,j}, \sigma_{g,e,j}^{2}\right)}_{\mathsf{Gaussian approximation}}$$

C. Lippert & O. Stegle

Mapping high dimensional traits

Variance component analysis

Observational model needs to acknowledge count statistics

- \blacktriangleright The observed quantities are read counts $\mathbf{c}_{g,e}$ and not true expression levels
- Log-normal model on the Poisson rates
 - MCMC inference
 - Approximate Bayesian inference
 - Variance stabilizing transform (Anscombe transform)

$$p(\mathbf{c}_{g,e} \mid \mathbf{C}(\boldsymbol{\theta})) = \mathcal{N}\left(\mathbf{y}_{g,e} \mid \mathbf{0}, \mathbf{C}(\boldsymbol{\theta})\right) \prod_{j=1}^{J} \underbrace{\frac{\mathsf{Poisson}(c_{g,e,j} \mid e^{y_{g,e,j}})}{\mathsf{Poisson observation model}}}_{\mathsf{Poisson} \left(\mathbf{y}_{g,e} \mid \mathbf{0}, \mathbf{C}(\boldsymbol{\theta})\right) \prod_{j=1}^{J} \underbrace{\mathcal{N}\left(y_{g,e,j} \mid \mu_{g,e,j}, \sigma_{g,e,j}^{2}\right)}_{\mathsf{Gaussian approximation}}$$

C. Lippert & O. Stegle

Mapping high dimensional traits

 $\sigma^2 \sim$

Variance component analysis Impact of environment

 Environment greatly affects heritability

 Absolute environment contribution small



Variance component analysis Impact of environment



(Average across upper 50% quantile of heritable genes)

イロト イポト イヨト イヨト

C. Lippert & O. Stegle

Mapping high dimensional traits

▶ ৰ ≣ ▶ ≣ ৩ ৭ ৫ September 2012 28

GWAS main effects

	Expressed gene position										
	Chromasame 3 (Mb)			Chromesone 2 (Mo) Chromesone 3 (Mo)			Ovamosome & Mol		Oveneseme 5 (#b)		
		0 0 0			1 5 8		<u>.</u>			0 8 0	
										and the state of the	
										and the second second	
										Section Provide	
										and the second second product	
					distance and the second					and the second second second second	_
2										1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	
÷.										1. A 2 1 1 1 1 1 1 1 1	_
- 11										1.	
3	ALC: NOT A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTIONO								100 B 100	5	
- 1	Contract of the second second								 Comparison (1997) 	and the state of the state	
- 6 "	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1				1				and the second sec	5.1.5 Arrows 10.5	
									the second of the		- I
									1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1		
									-702 - 1		
									Sec. Sec.		
									1		110
									and a second second second		
											111
								1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 -			
								1			
								and the second second			
- 8											
Ξ								1.			
1								and the second se			
3								The state of the second second			
- E								and the second second			
0.1								A COMPANY OF A COMPANY OF			
							1244				
							20				
	and the second second							A CONTRACTOR OF			
						1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1				1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	
						1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1					
						1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1					
						1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1					
3						1000000					
- 51						1					
2						1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1			1.5 6 11 12		
					5 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	And the second		A CONTRACTOR OF			
- 3 *					and the second s						
ΕÅ.					10 A. 200 A.			Contraction and the			
£7.					and the second second						
1.1					1						
Ł					1						
÷ .	100 B 100 B 100 B				1						
1	a second second second				and the second second second						
ξ.,											
£				1							
•				1				and the second second			
2											
ň.				1.2							
- t.,				20 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1							
- 8				1 Mar 1997							
- 8				A CONTRACTOR OF THE OWNER							
δ,											
			1.1	1.1							
						18 11 11					
			6. 1								
		1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1									
		and the second			solution and the second						
		- 2 Content of the state of the									
		1									
- 2		and the second second second									
8		Sector Standards									
а.	and a second	which was the latest							S 40 22 4		
- 11	The stand of the stand of the								10000		
- 5	and a state of the second										
- S.,	and the second s										
	Service March										
	Subsection Statistics and										
	Server March 199	Contraction of the second									
	Bear Section 1										
	a state of the second								Sec. 35		
	No. 1 Sec.										

C. Lippert & O. Stegle

Mapping high dimensional traits

◆□→ ◆□→ ◆注→ ◆注→ □注

GWAS GxE effects

		Expressed gene position								
Chromesene 3 (HM			Charamanana 2 (Mil)	Owners	Chromesene 3 (MH)		Dramowing & Mr.)	Overeseme 5 (#b)		
	6 B B			1 1 1	·····		· · · · · · · ·	1 1 2 2 2 2	0	
A										
									- 1 - 1 - 1	
a second second									A	
and the second se									1111 (111))))))))	
									2 C 2 C 2 C 2 C 2 C 2 C 2 C 2 C 2 C 2 C	
a standard start									1	
and the second second					and the same			the mark that is the strength	2021	
0										
									S. 125	
									- 52 C	
									the second s	
									1.00	
									ba	
The second second second										
A COMPANY OF A COM										
and the second										
C to a state of the state of th										
and the second second										
and the second se										
8										
and the second second										
and the second s										
S manual + they have										
S managers.										
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1										
A series and series and										
a)										
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1										
P										
8 2432 States 19 1										
2										
8										
a start and a start										
the second second second										
Participation of the second se										
· · · · · · · · · · · · · · · · · · ·										
a second of the second of										
1 N N N N N N N N N N N N N N N N N N N				A set i star						
2 strates and a second										
8 10 KK KL 1 K K										
Sal an anna a success										
and the state of the										
1										
D										
1. 1. 1. 1. 1. 1.										
A CONTRACTOR OF A CONTRACT										

C. Lippert & O. Stegle

Mapping high dimensional traits

◆□> ◆圖> ◆臣> ◆臣>
The role of GxE in the A. thaliana transcriptional landscape

GWAS GxE effects



	and a subsection for	
		Contraction of the second s
and the second se		the second se
		and the second se
		and the second se
a contraction of the second		A REAL PROPERTY OF A READ REAL PROPERTY OF A READ R
		Contraction of the second second second
		A REAL PROPERTY OF A READ REAL PROPERTY OF A REAL P
A CONTRACTOR OF		
NUTCHOUSE A DESCRIPTION		
the second		
A 1		
the second s		
and a second sec		
A CONTRACTOR OF A CONTRACTOR O		
a spinishen and the second second		

◆□→ ◆圖→ ◆臣→ ◆臣→

The role of GxE in the A. thaliana transcriptional landscape

GWAS GxE effects



Hanny states (trade they are a state and state
and a state of the state of a state with the state
sectors constant a manufacture sectors and a sector and a sector and a sector and a sector a
searchest of a state of the second second of the second second second second second second second second second
I second contracted to monteners association to reconcitive second to
internet internet in the second second second in the second s
Planting (Decempes of the lease density 1968 b) Strangers Strain Algebrash
interest interferent of contribution comparis, contribution of proceeding statistics and the process
Transmiss measures in received statement maps in particular entered contraction

・ロト ・ 日 ・ ・ ヨ ・ ・ ヨ ・

- Lack of power to detect GxE?
- GxE largely aligned with population structure?

C. Lippert & O. Stegle

Mapping high dimensional traits

E ▶ 4 E ▶ E ∽ Q (September 2012 30

Outline

Motivation

Accounting for background variation in eQTL studies

Mechanistic models: Genetic analyses with learnt cellular features

The role of GxE in the A. thaliana transcriptional landscape

Summary

C. Lippert & O. Stegle

Mapping high dimensional traits

イロト イポト イヨト イヨト

1. eQTL mapping is sensitive to background signals

- co-factors
- population structure
- "hidden confounding"
- Leverage on high dimensionality of gene expression
- 2. Mechanistic models in eQTL
 - ▶ Intermediate molecular traits can be measured or learnt from data.
- 3. Analysis of variance is well applicable for NGS data
 - ▶ GxE explains substantial variance, however is difficult to map

イロト イポト イヨト イヨト

1. eQTL mapping is sensitive to background signals

- co-factors
- population structure
- "hidden confounding"
- Leverage on high dimensionality of gene expression

2. Mechanistic models in eQTL

- ▶ Intermediate molecular traits can be measured or learnt from data.
- 3. Analysis of variance is well applicable for NGS data
 - ▶ GxE explains substantial variance, however is difficult to map

(日) (周) (三) (三)

Summary

1. eQTL mapping is sensitive to background signals

- co-factors
- population structure
- "hidden confounding"
- Leverage on high dimensionality of gene expression
- 2. Mechanistic models in eQTL
 - ► Intermediate molecular traits can be measured or learnt from data.
- 3. Analysis of variance is well applicable for NGS data
 - GxE explains substantial variance, however is difficult to map

< ロ > < 同 > < 回 > < 回 > < 回 > <

Summary

1. eQTL mapping is sensitive to background signals

- co-factors
- population structure
- "hidden confounding"
- Leverage on high dimensionality of gene expression
- 2. Mechanistic models in eQTL
 - ► Intermediate molecular traits can be measured or learnt from data.
- 3. Analysis of variance is well applicable for NGS data
 - GxE explains substantial variance, however is difficult to map

< ロ > < 同 > < 回 > < 回 > < 回 > <

- 1. eQTL mapping is sensitive to background signals
 - co-factors
 - population structure
 - "hidden confounding"
 - Leverage on high dimensionality of gene expression
- 2. Mechanistic models in eQTL
 - Intermediate molecular traits can be measured or learnt from data.
- 3. Analysis of variance is well applicable for NGS data
 - ► GxE explains substantial variance, however is difficult to map

イロト イポト イヨト イヨト

Acknowledgements

- Accounting for background variation in eQTL studies
 L. Parts, J. Winn, Nicolo Fusi, N. Lawrence, Richard Durbin
- Mechanistic models: Genetic analyses of cellular features
 L. Parts, J. Winn, R. Durbin
- The role of GxE in the A. thaliana transcriptional landscape E. J. Osborne, M. Remigereau, P. Zhang,, Oliver Stegle, Philipp Drewe, Quan Long, Ümit Seren, Andre Kahles, Qiang Song, Arthur Korte, Andrew D. Smith, Gunnar Rätsch, Richard M Clark, Magnus Nordborg, Bjarni Vilhjalmsson

イロト 不得下 イヨト イヨト 二日

References I

- N. Fusi, O. Stegle, and N. Lawrence. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS computational biology*, 8(1):e1002330, 2012.
- E. Lee and H. Bussemaker. Identifying the genetic determinants of transcription factor activity. *Molecular systems biology*, 6(1), 2010.
- J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):e161+, September 2007. doi: 10.1371/journal.pgen.0030161.
- L. Parts, O. Stegle, J. Winn, and R. Durbin. Joint genetic analysis of gene expression data with inferred cellular phenotypes. *PLoS genetics*, 7(1):e1001276, 2011.
- E. Smith and L. Kruglyak. Gene–environment interaction in yeast gene expression. PLoS biology, 6(4):e83, 2008.
- O. Stegle, L. Parts, R. Durbin, and J. Winn. A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies. *PLoS Comput Biol*, 6(5):e1000770, May 2010.
- O. Stegle, L. Parts, M. Piipari, J. Winn, and R. Durbin. Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7(3):500–507, 2012.

▲□▶ ▲□▶ ▲□▶ ▲□▶ □ のの⊙