Mapping of high-dimensional traits: eQTL and beyond

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

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

Basel
09. September 2012

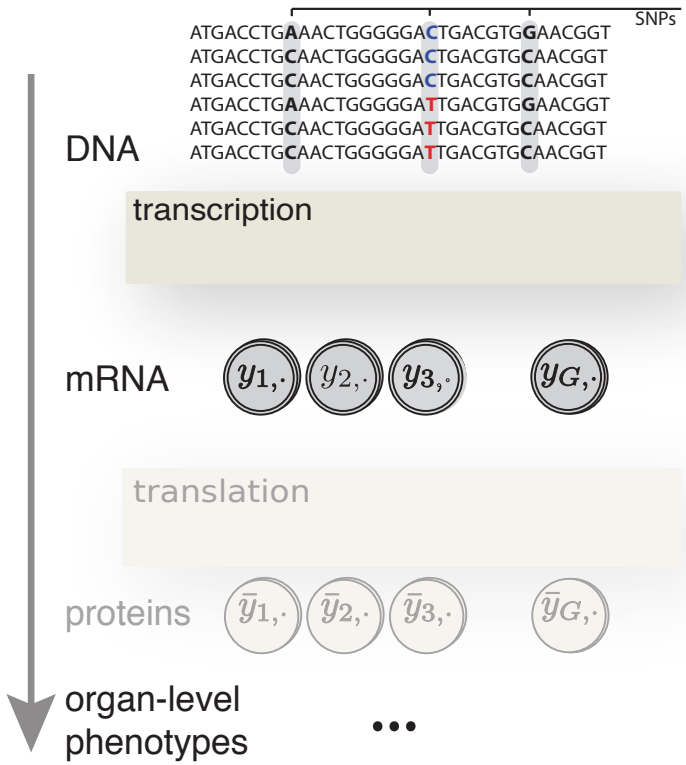
Microsoft Research

Models of molecular variation

Motivation

Goal:

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

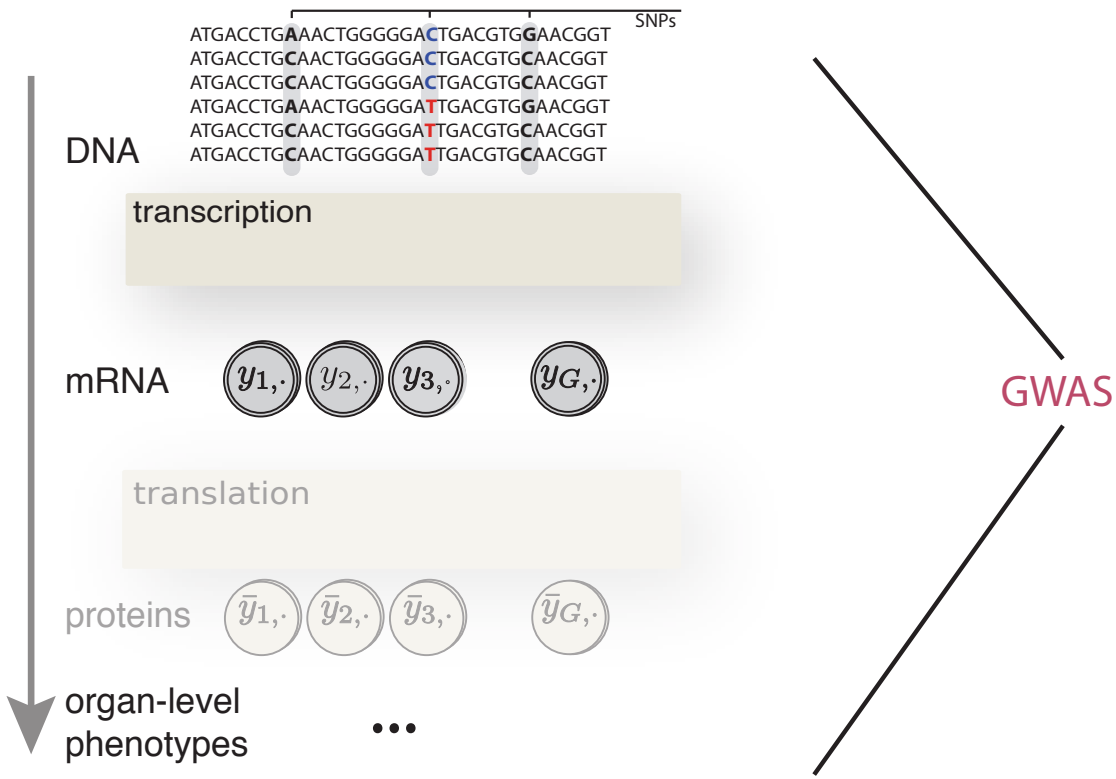


Models of molecular variation

Motivation

Goal:

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

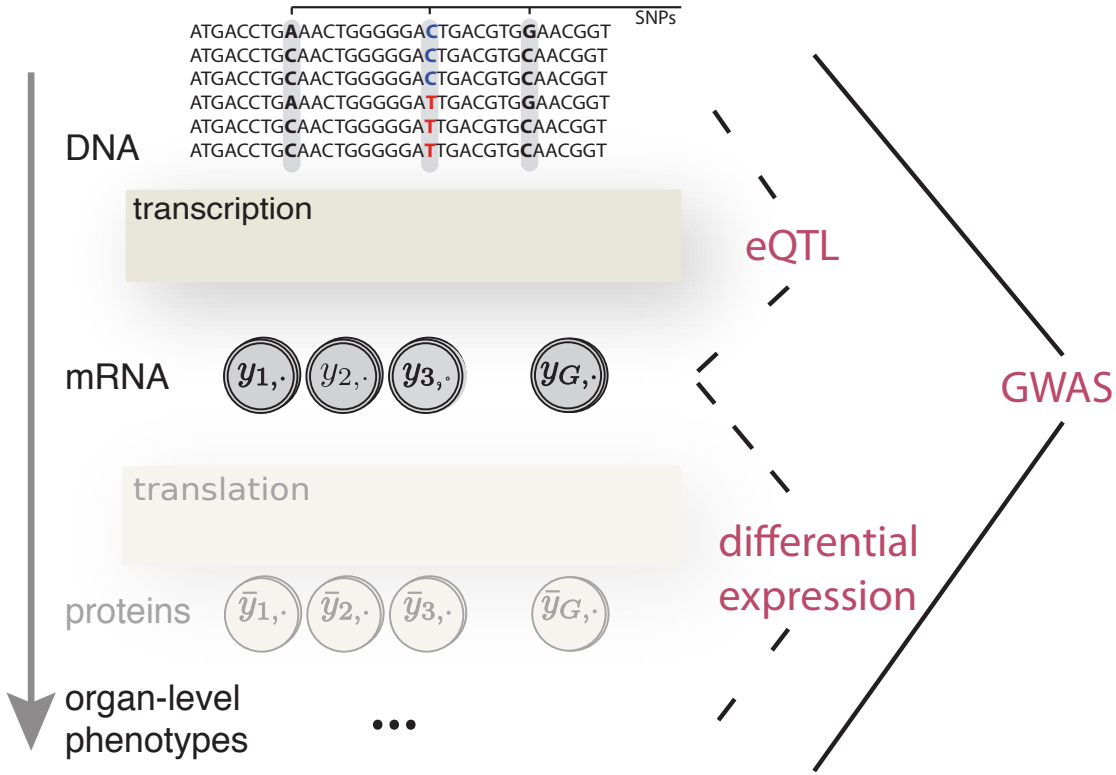


Models of molecular variation

Motivation

Goal:

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

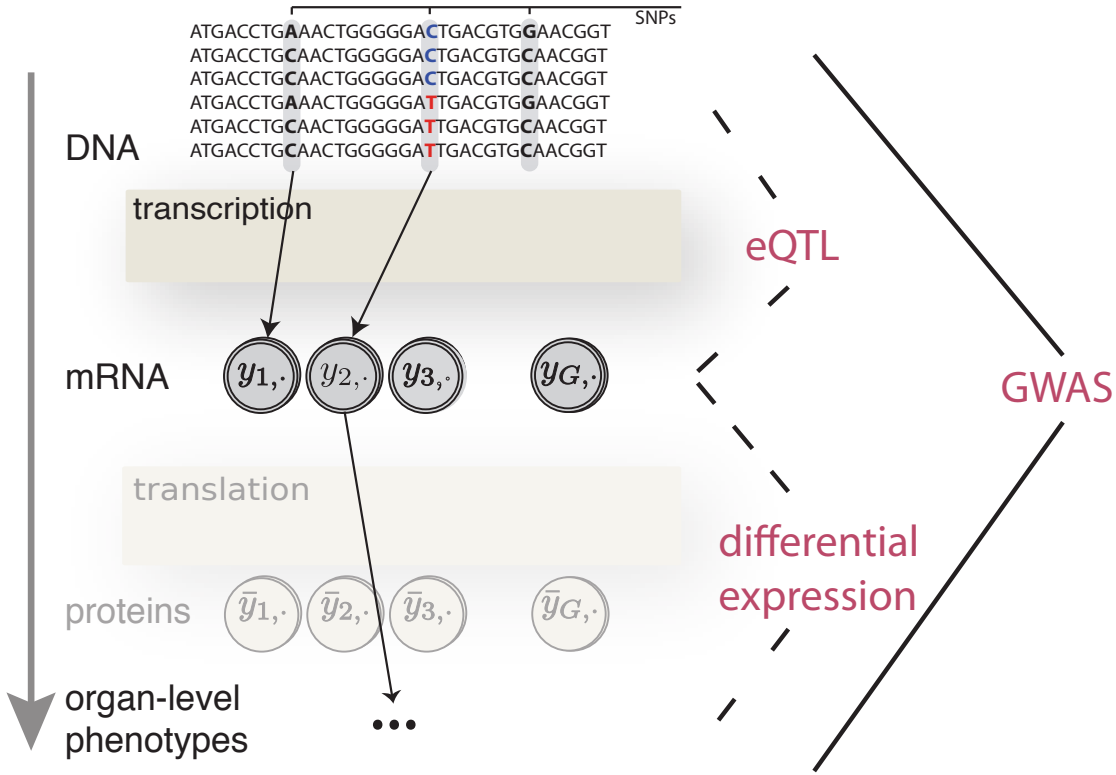


Models of molecular variation

Motivation

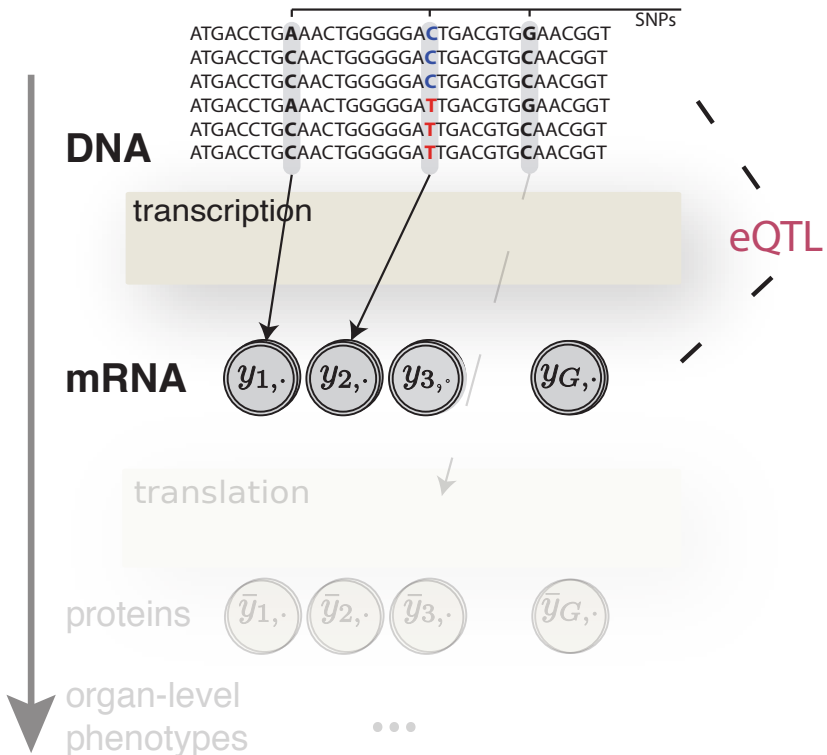
Goal:

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



Models of molecular variation

eQTL

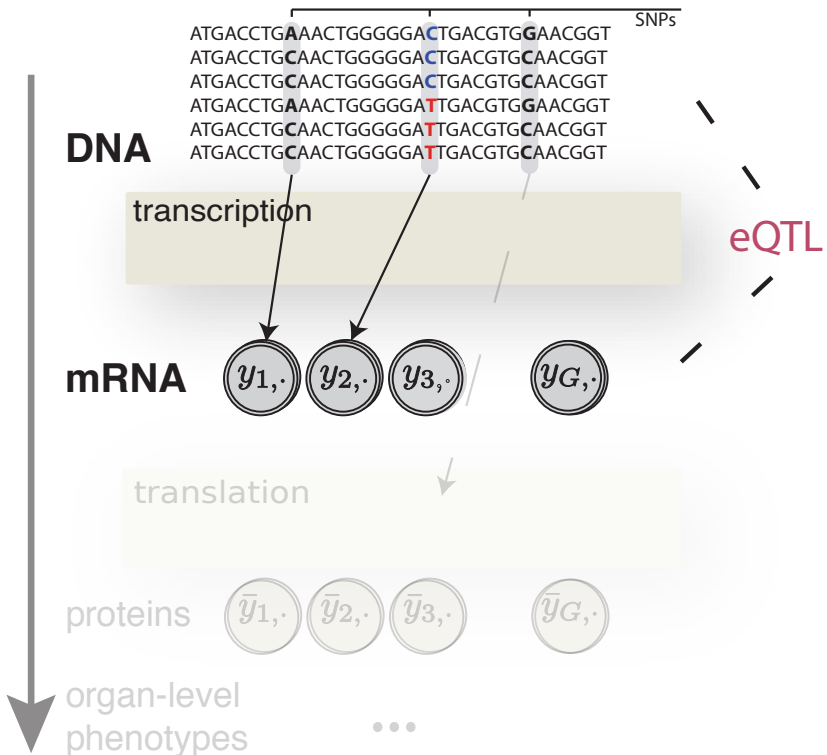


Models of molecular variation

eQTL

Statistical challenges:

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

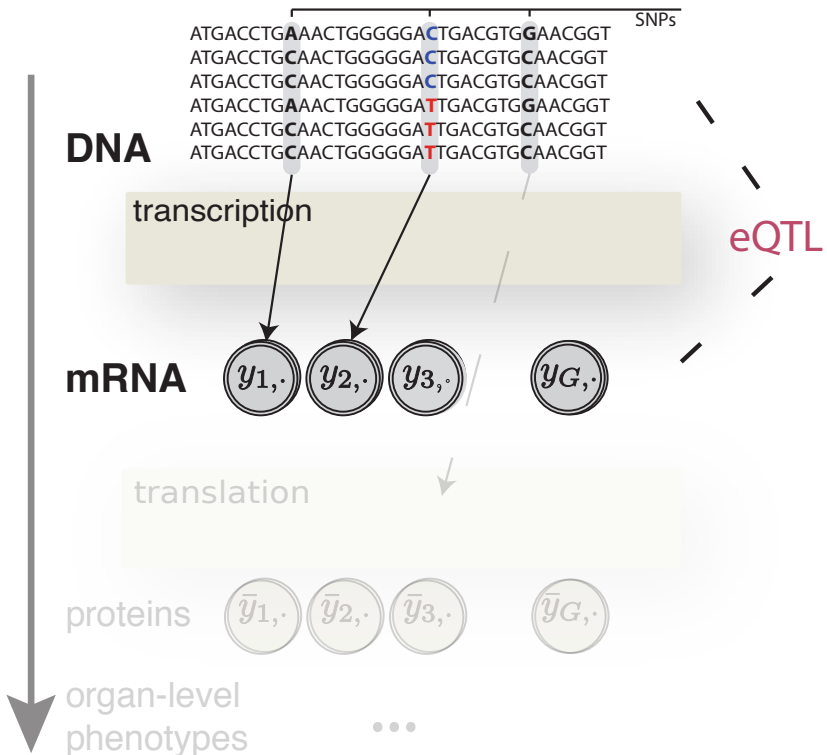


Models of molecular variation

eQTL

Statistical challenges:

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



Models of molecular variation

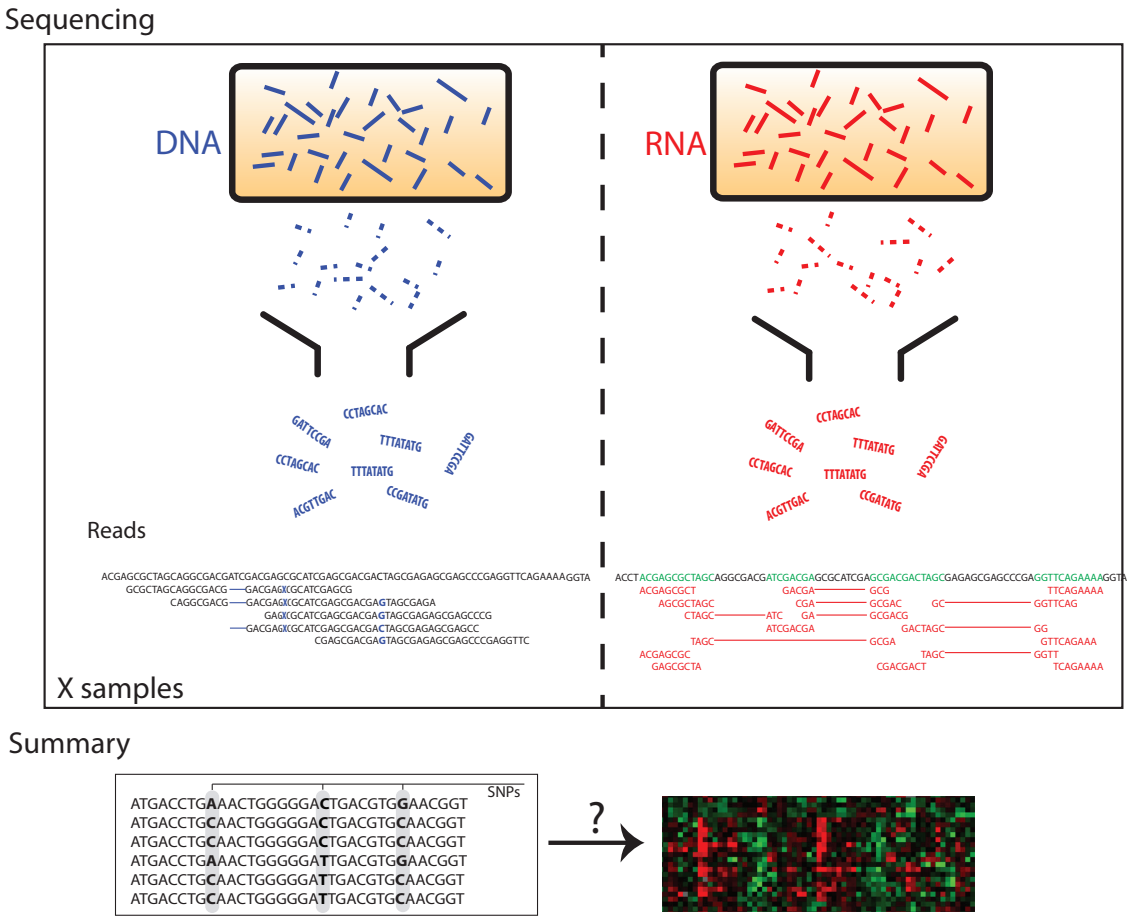
eQTL

Statistical challenges:

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

Bioinformatics challenges:

- Phenotyping
- Using NGS technologies



Outline

Outline

Models of molecular variation

Univariate phenotypes, examples

- Single-marker mapping using linear models:

$$\mathbf{y}_g = \underbrace{\mathbf{x}_s \beta_{s,g}}_{\text{genetic}} + \underbrace{\mathbf{u}}_{\text{background}} + \underbrace{\boldsymbol{\epsilon}_g}_{\text{noise}}$$

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

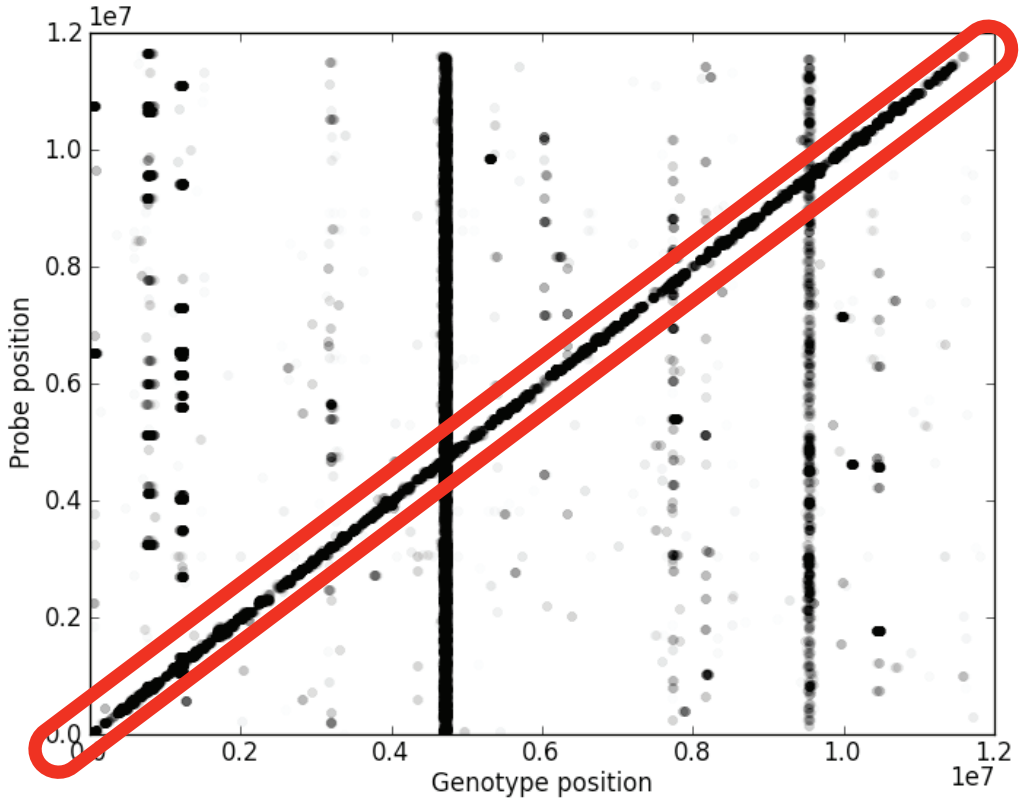
Models of molecular variation

Univariate phenotypes, examples

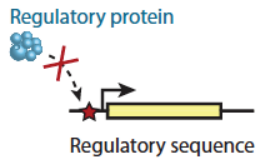
- Single-marker mapping using linear models:

$$\mathbf{y}_g = \underbrace{\mathbf{x}_s \beta_{s,g}}_{\text{genetic}} + \underbrace{\mathbf{u}}_{\text{background}} + \underbrace{\boldsymbol{\epsilon}_g}_{\text{noise}}$$

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



Local regulation



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

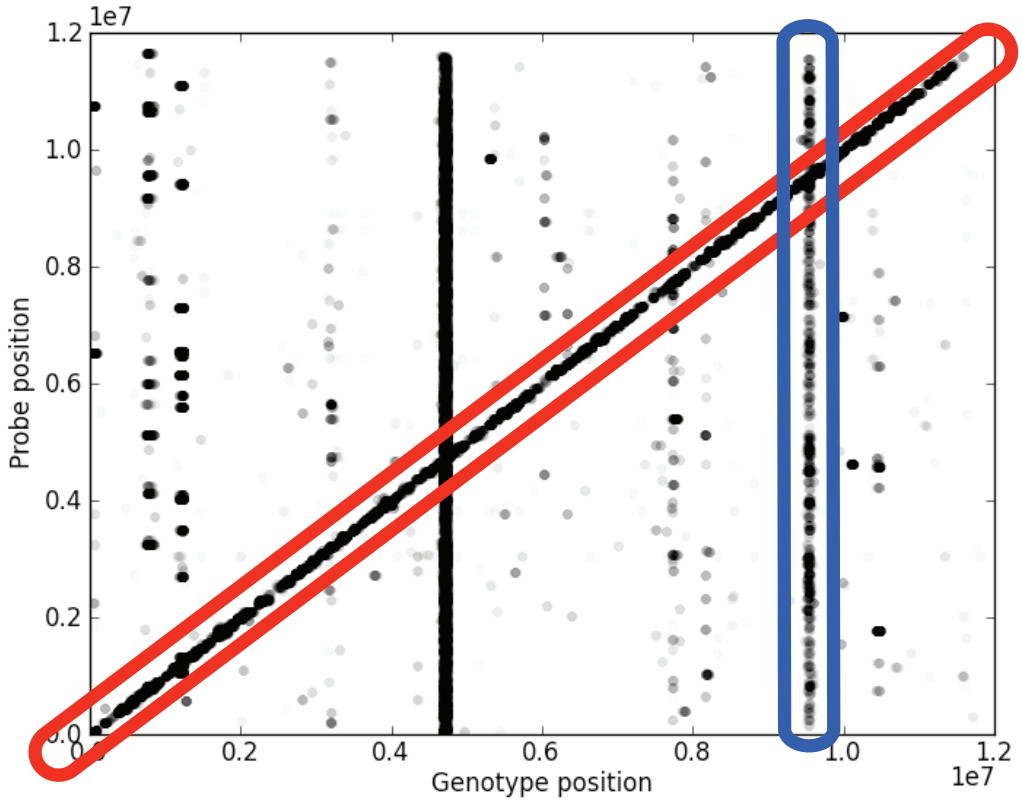
Models of molecular variation

Univariate phenotypes, examples

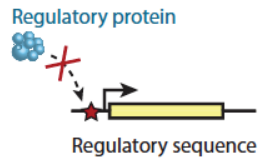
- Single-marker mapping using linear models:

$$\mathbf{y}_g = \underbrace{\mathbf{x}_s \beta_{s,g}}_{\text{genetic}} + \underbrace{\mathbf{u}}_{\text{background}} + \underbrace{\boldsymbol{\epsilon}_g}_{\text{noise}}$$

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

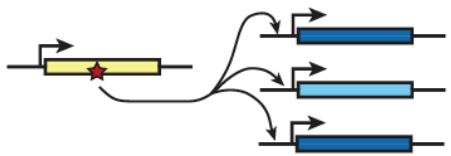


Local regulation



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

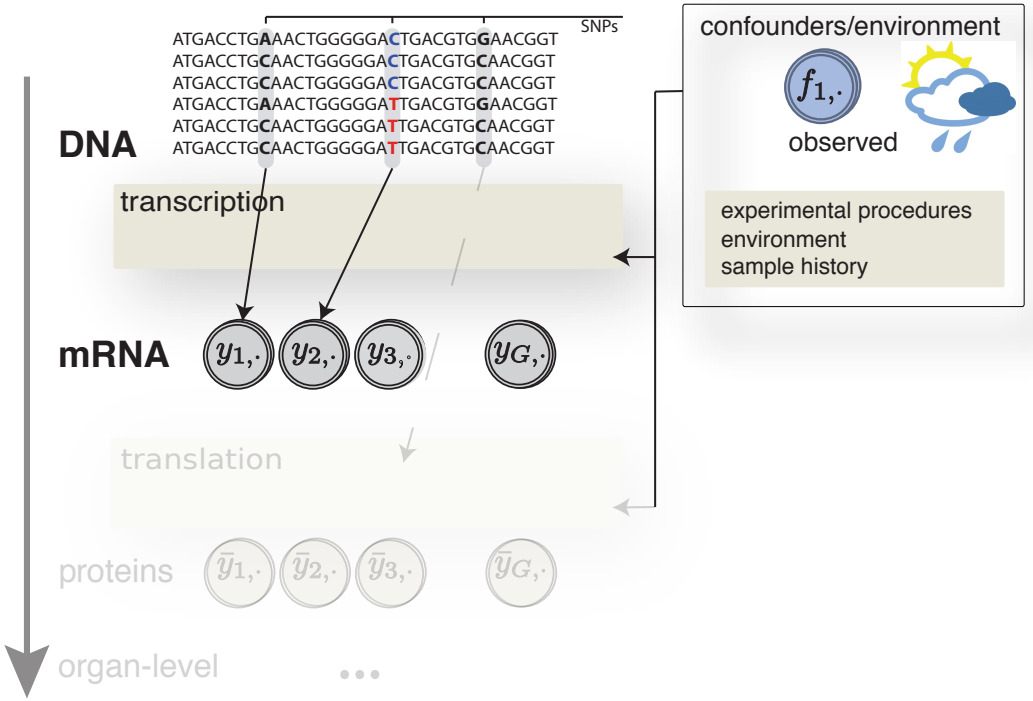
Distal regulation



- sequence of a TF,
- regulatory protein, pathway
- hotspots
- typically *trans* mechanisms

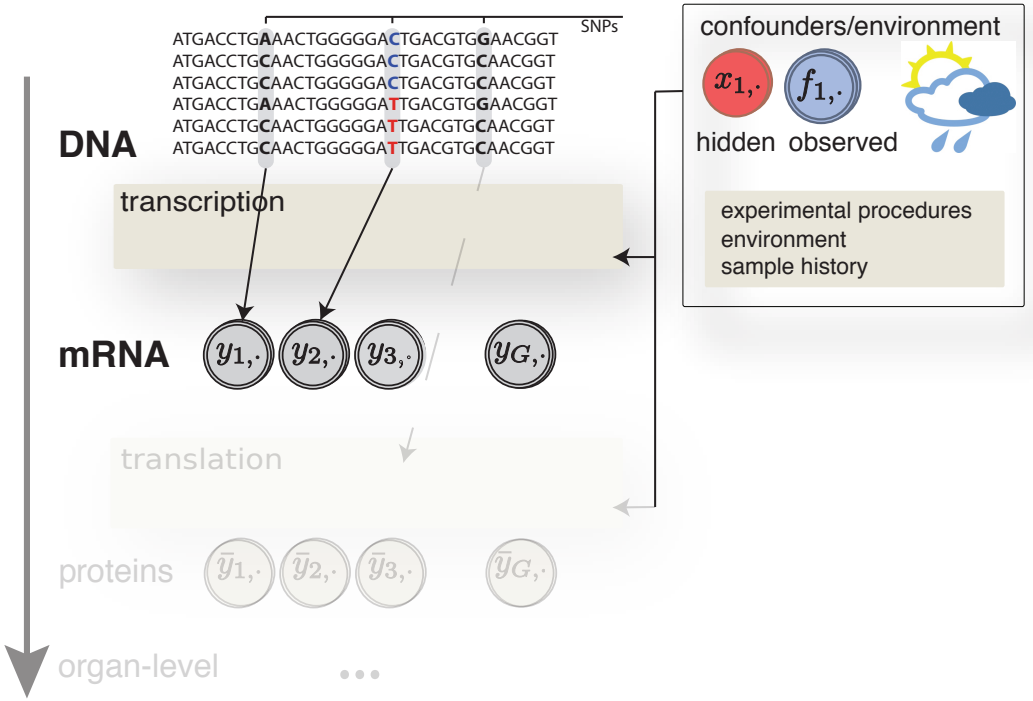
Known and unknown confounding in genomic analyses

- 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

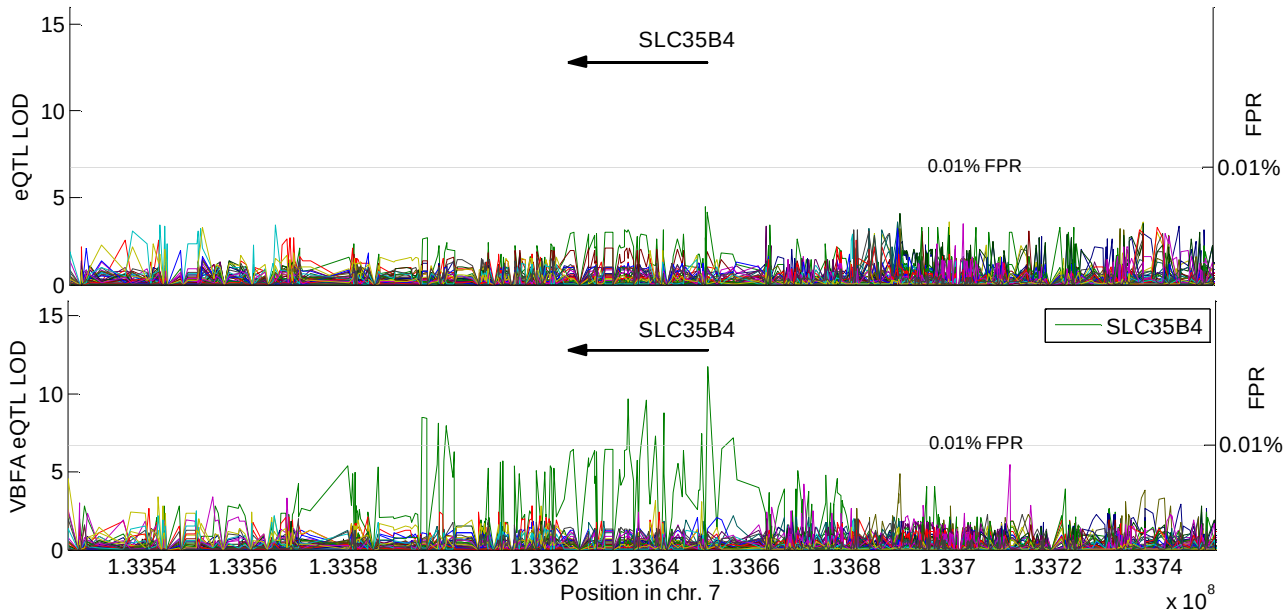


[Leek and Storey, 2007]

Example, Human

(a) Standard

(b) VBFA accounting for hidden factors



Association model

PEER: accounting for hidden factors

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

$$p(w_{g,k}, \alpha_k) = \mathcal{N}\left(w_{g,k} \middle| 0, \frac{1}{\alpha_k}\right) \Gamma(\alpha_k \,|\, a_k, b_k)$$

$$y_g^n = \overbrace{(x_s^n \beta_{s,g})}^{\text{genetic}} + \underbrace{\mathbf{f}^n \mathbf{v}_g}_{\text{known factors}} + \underbrace{\epsilon_g^n}_{\text{noise}}$$

[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 $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_K\}$ need to be learnt from the expression data.
- Controlling model complexity using hierarchical Bayesian modeling.

$$p(w_{g,k}, \alpha_k) = \mathcal{N}\left(w_{g,k} \,\middle|\, 0, \frac{1}{\alpha_k}\right) \Gamma(\alpha_k \,|\, a_k, b_k)$$

$$y_g^n = \overbrace{(x_s^n \beta_{s,g})}^{\text{genetic}} + \underbrace{\mathbf{f}^n \mathbf{v}_g}_{\text{known factors}} + \underbrace{\mathbf{h}^n \mathbf{w}_g}_{\text{hidden factors}} + \underbrace{\epsilon_g^n}_{\text{noise}}$$

[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 $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_K\}$ need to be learnt from the expression data.
- Controlling model complexity using hierarchical Bayesian modeling.

$$p(w_{g,k}, \alpha_k) = \mathcal{N}\left(w_{g,k} \,\middle|\, 0, \frac{1}{\alpha_k}\right) \Gamma(\alpha_k \,|\, a_k, b_k)$$

$$y_g^n = \overbrace{(x_s^n \beta_{s,g})}^{\text{genetic}} + \underbrace{\mathbf{f}^n \mathbf{v}_g}_{\text{known factors}} + \underbrace{\mathbf{h}^n \mathbf{w}_g}_{\text{hidden factors}} + \underbrace{\epsilon_g^n}_{\text{noise}}$$

[Stegle et al., 2010]

Association model

PEER: accounting for hidden factors

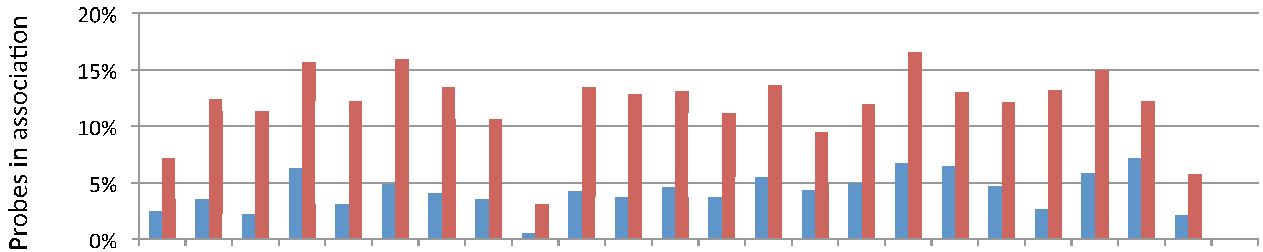
- Start with standard association model.
- Include (few) global hidden factors (confounders) in the model.
- Factors $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_K\}$ need to be learnt from the expression data.
- Controlling model complexity using hierarchical Bayesian modeling.

$$p(w_{g,k}, \alpha_k) = \mathcal{N}\left(w_{g,k} \,\middle|\, 0, \frac{1}{\alpha_k}\right) \Gamma(\alpha_k \,|\, a_k, b_k)$$

$$y_g^n = \overbrace{(x_s^n \beta_{s,g})}^{\text{genetic}} + \underbrace{\mathbf{f}^n \mathbf{v}_g}_{\text{known factors}} + \underbrace{\mathbf{h}^n \mathbf{w}_g}_{\text{hidden factors}} + \underbrace{\epsilon_g^n}_{\text{noise}}$$

[Stegle et al., 2010]

Use cases

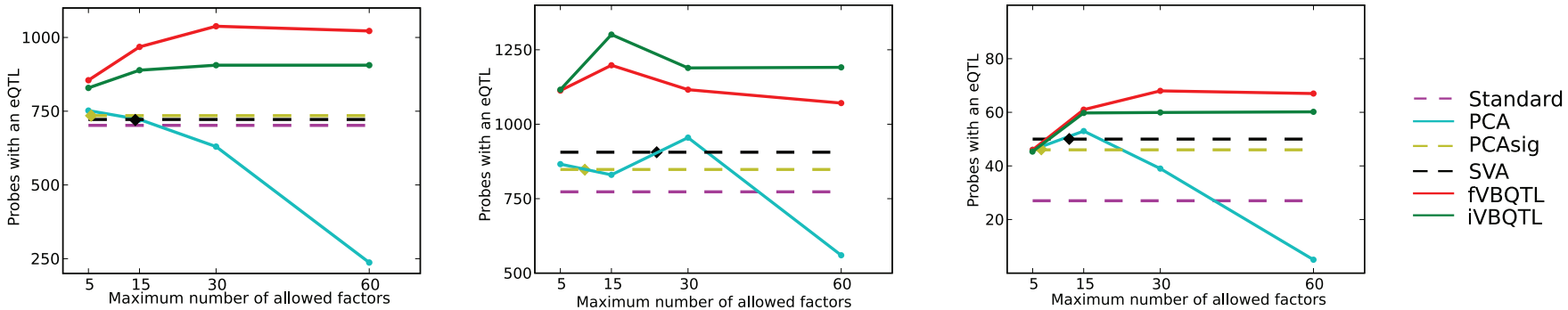


Chr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y
Probes	1009	644	540	384	449	571	468	338	387	380	545	520	189	330	348	426	549	154	618	266	120	238	328	15
Standard	25	23	12	24	14	28	19	12	2	16	20	24	7	18	15	21	37	10	29	7	7	17	7	0
VBQTL	72	80	61	60	55	91	63	36	12	51	70	68	21	45	33	51	91	20	75	35	18	29	19	0

[Stegle et al., 2010]

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

Use cases



(a) Yeast *cis* eQTLs (b) Mouse *cis* eQTLs (c) Human *cis* eQTLs

[Stegle et al., 2010]

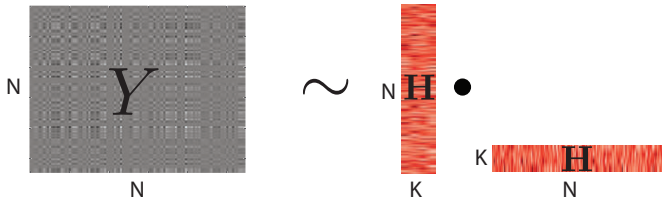
- Increased power
- Similarly on yeast, mouse, *A. thaliana*

Mixed model implementation

$$\mathbf{y}_g = \underbrace{\mathbf{x}_s \beta_{s,g}}_{\text{SNP effect}} + \underbrace{\mathbf{vU}}_{\text{covariates}} + \underbrace{\mathbf{Hw}_g}_{\text{unknown confounding}} + \boldsymbol{\epsilon}_g$$

- 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



Gene expression hidden confounders

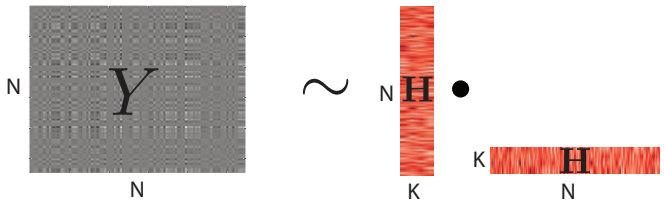
$+\psi$

noise

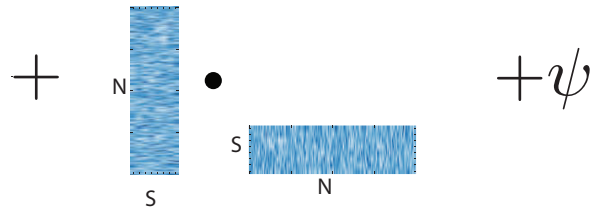
- Exploit large number of expression traits to estimate the empirical covariance structure.
- Iterative learning on the covariance structure induced by all traits.
  1. Learn confounders, explaining broad covariance within expression profiles.
  2. Test for genetic (SNP) control of learnt confounders.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

[Fusi et al., 2012]

Mixed model implementation



Gene expression

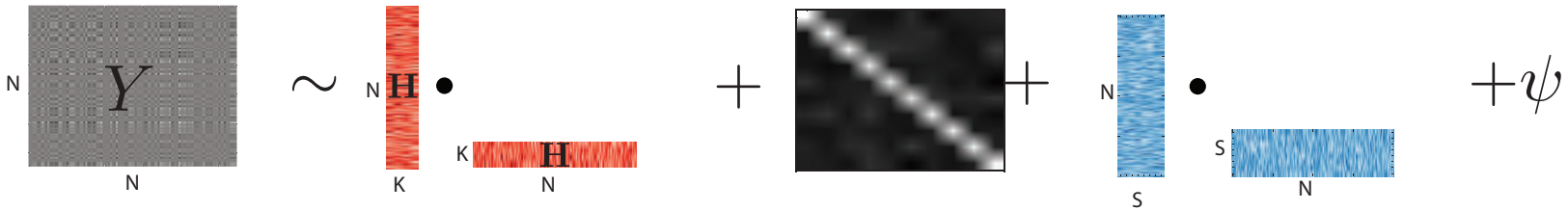


master regulators noise

- 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.

[Fusi et al., 2012]

Mixed model implementation

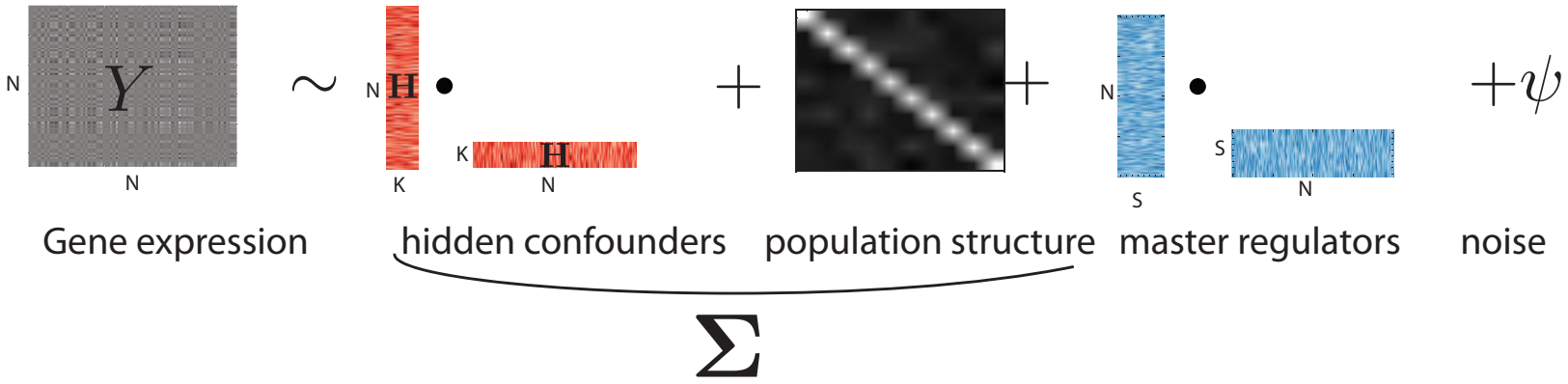


Gene expression hidden confounders population structure master regulators noise

- 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.

[Fusi et al., 2012]

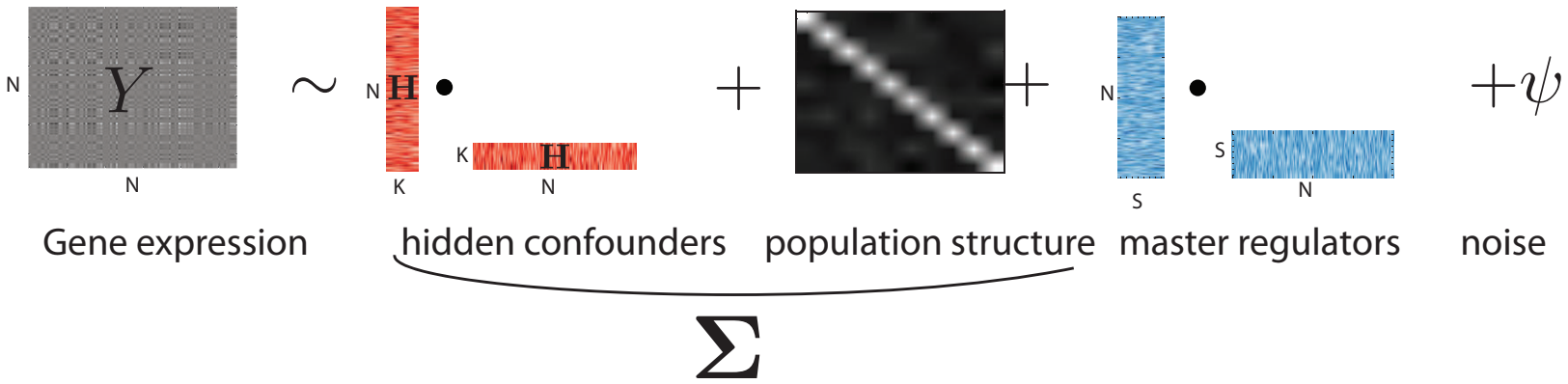
Mixed model implementation



- 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.

[Fusi et al., 2012]

Mixed model implementation

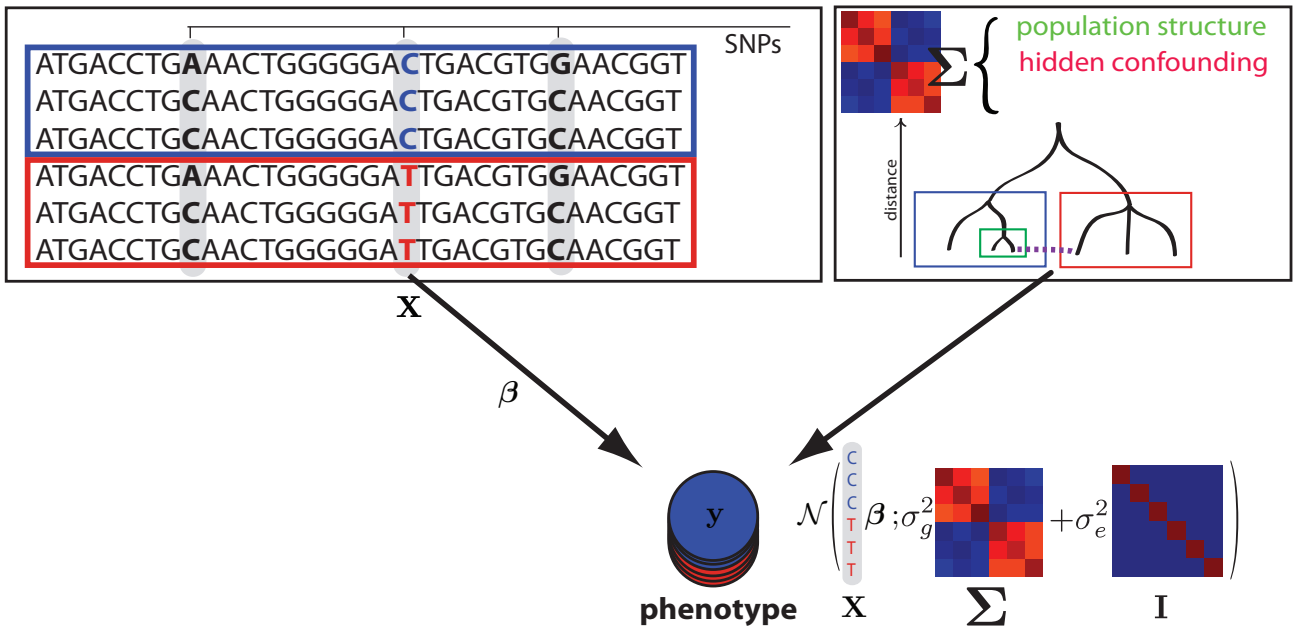


- Formally, the expression levels are independent given genotype, hidden factors and population structure.

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

Mixed model implementation

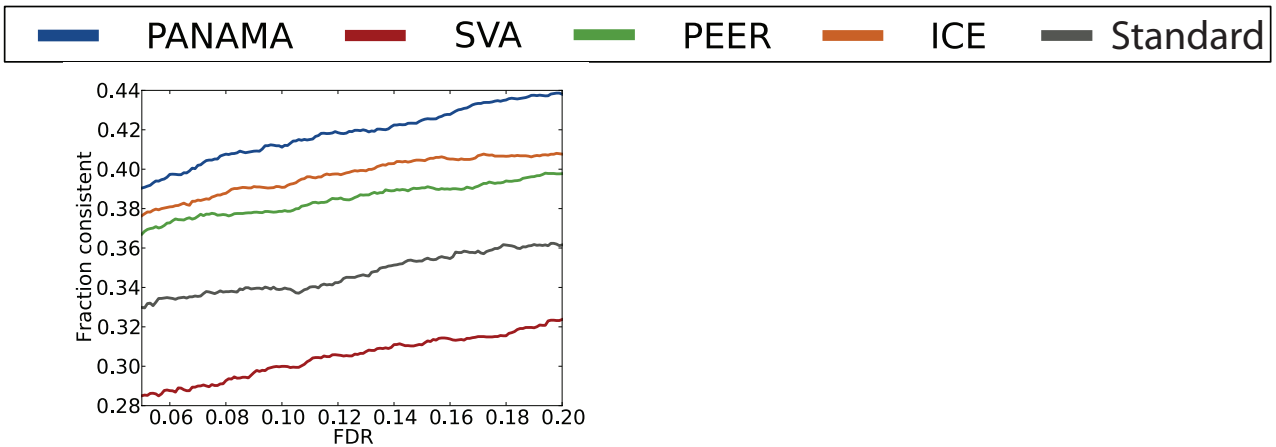
Testing strategies



- Mixed model likelihood ratio

$$\mathrm{LOD}_{g,s} = \log \frac{\mathcal{N}\left(\mathbf{y}_g \,\middle|\, \mathbf{x}_s\beta_{s,g}, \sigma_g^2\boldsymbol{\Sigma} + \sigma_e^2\mathbf{I}\right)}{\mathcal{N}\left(\mathbf{y}_g \,\middle|\, \mathbf{0} \quad , \sigma_g^2\boldsymbol{\Sigma} + \sigma_e^2\mathbf{I}\right)}$$

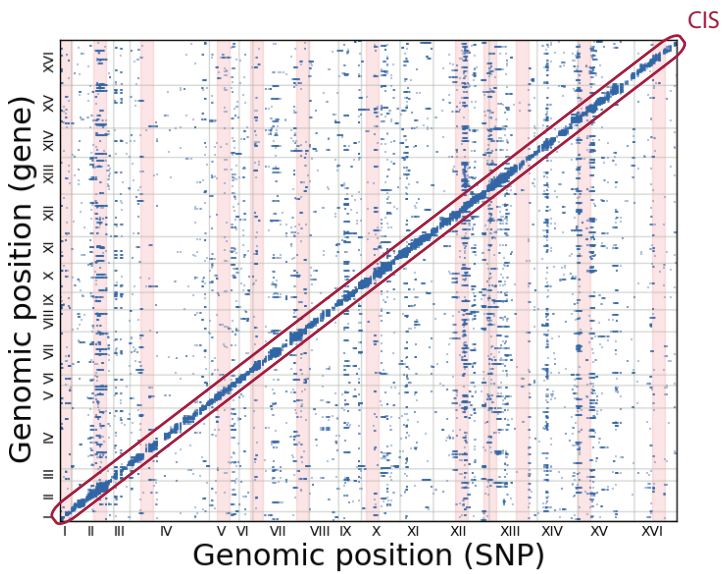
Use cases



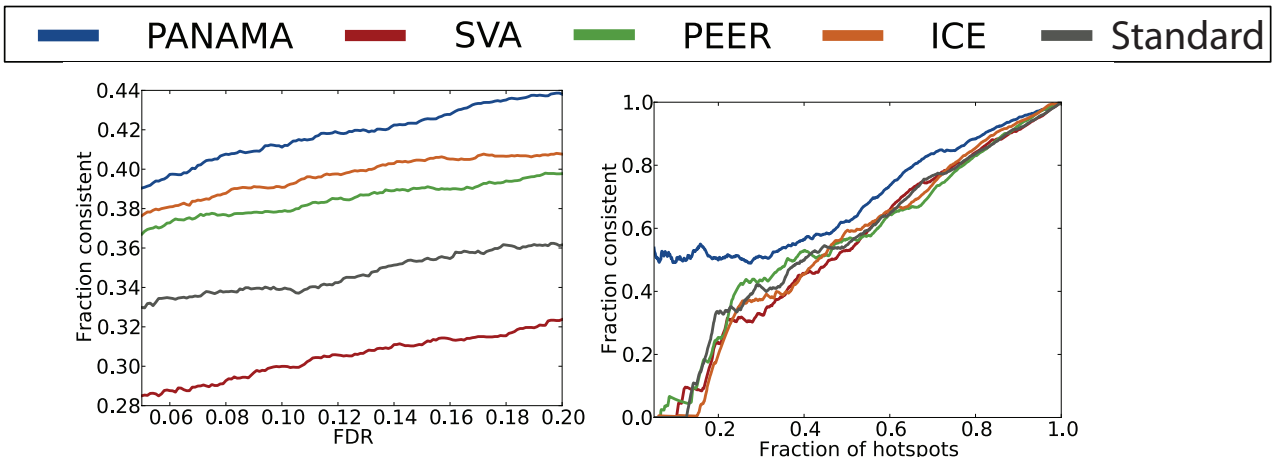
(d) Cis calling consistency

[Fusi et al., 2012]

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



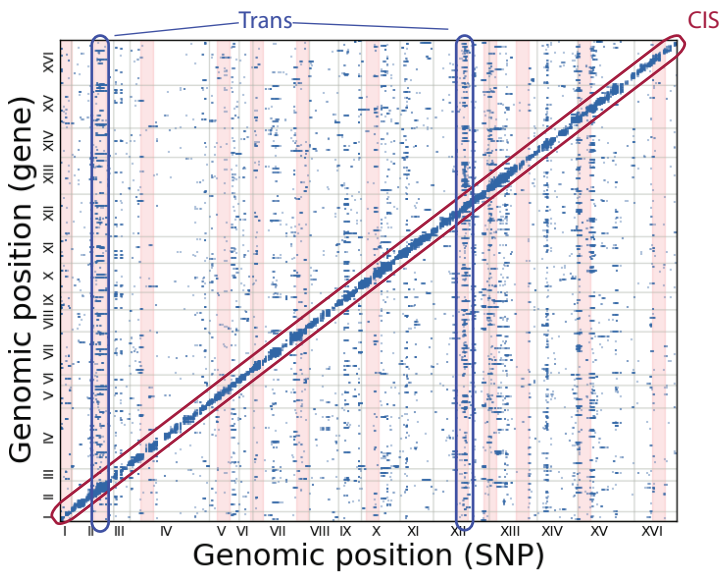
Use cases



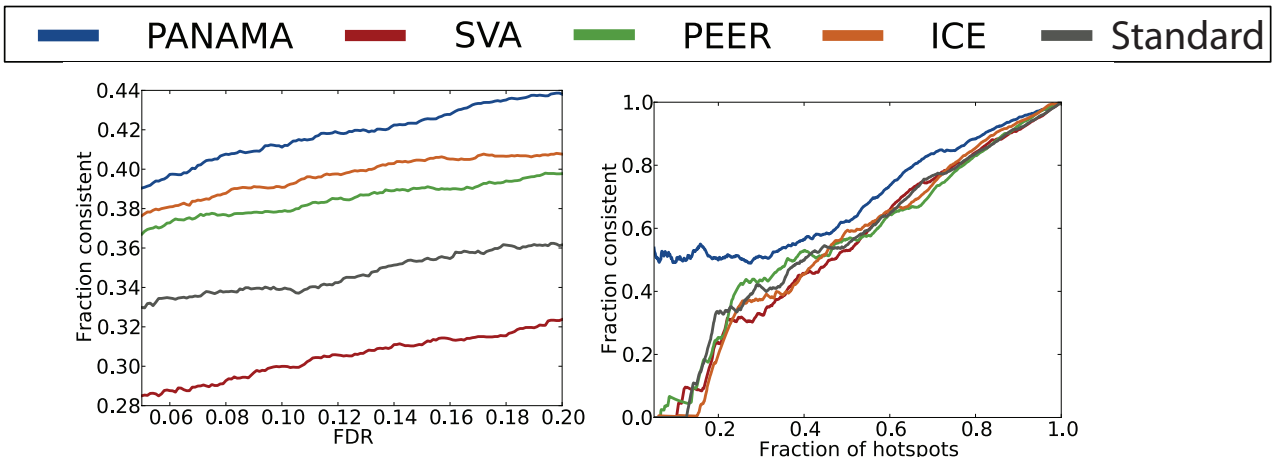
(d) Cis calling consistency (e) Trans calling consistency

[Fusi et al., 2012]

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



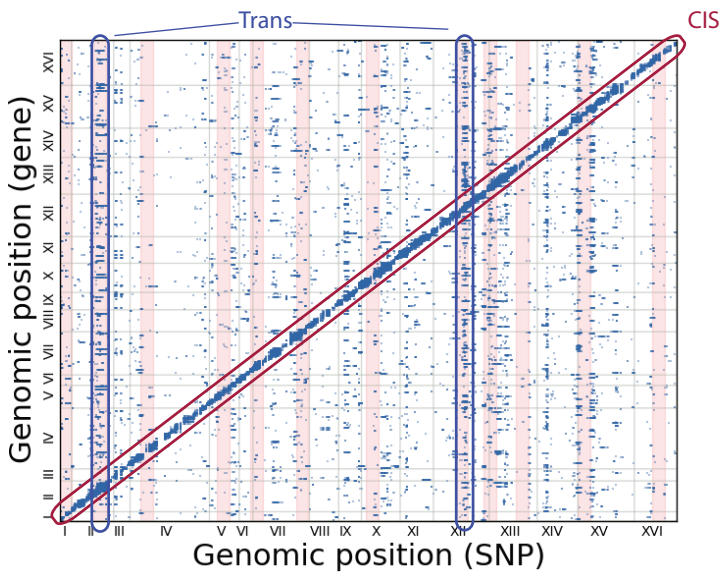
Use cases



(d) Cis calling consistency (e) Trans calling consistency

[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)

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)

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)

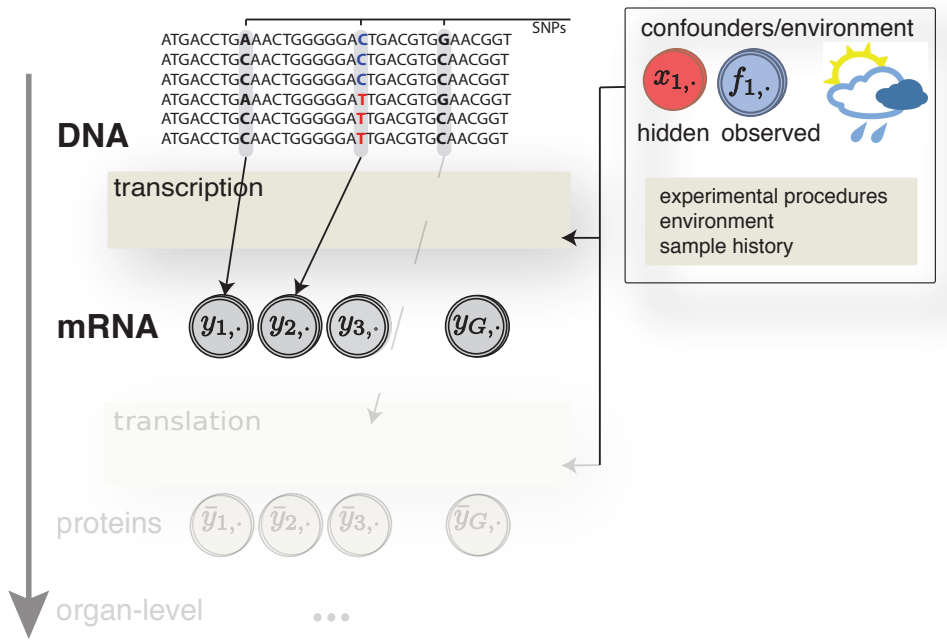
Outline

Regulatory and external factors

- Confounding factors

$$y_g^n = \overbrace{b_{n,g}\left(\mathbf{x}_s^n \theta_{n,g}\right)}^{\text{genetic}} + \underbrace{\mathbf{f}^n \mathbf{v}_g}_{\text{known factors}} + \underbrace{\mathbf{h}^n \mathbf{w}_g}_{\text{hidden factors}} + \underbrace{\epsilon_g^n}_{\text{noise}} .$$

- Account for regulatory factors in transcription?

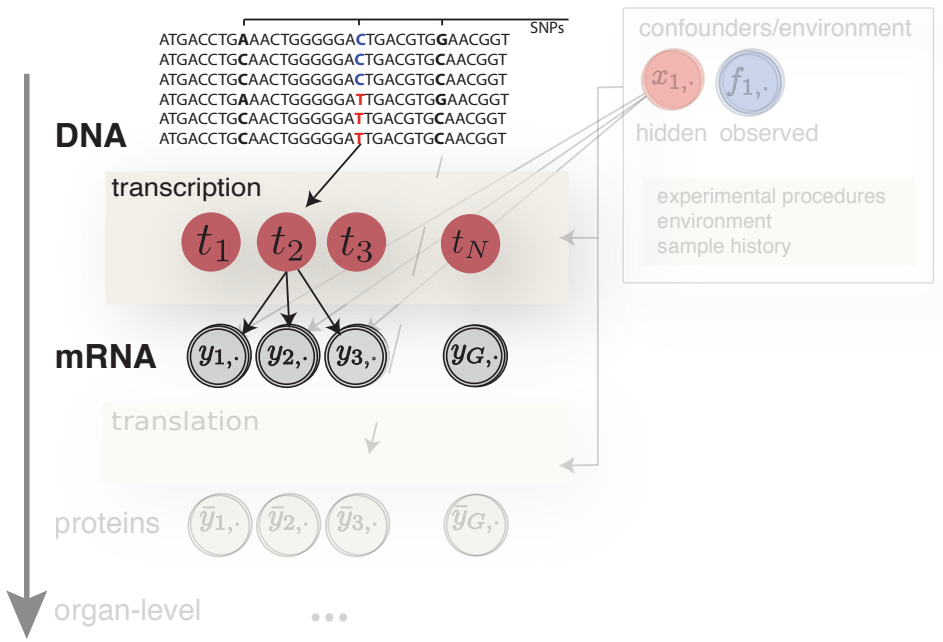


Regulatory and external factors

- Confounding factors

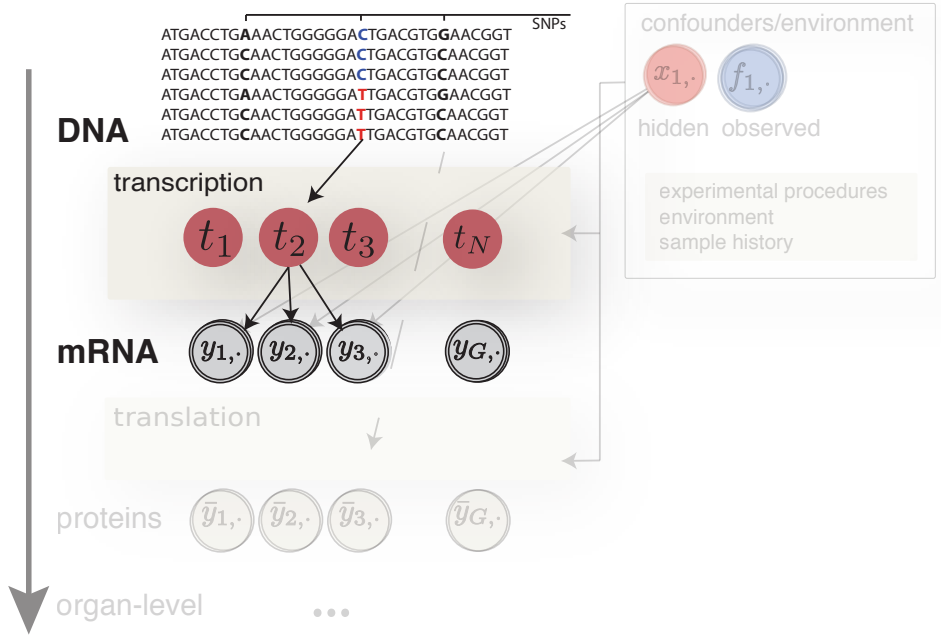
$$y_g^n = \overbrace{b_{n,g}\left(\mathbf{x}_s^n \theta_{n,g}\right)}^{\text{genetic}} + \underbrace{\mathbf{f}^n \mathbf{v}_g}_{\text{known factors}} + \underbrace{\mathbf{h}^n \mathbf{w}_g}_{\text{hidden factors}} + \underbrace{\epsilon_g^n}_{\text{noise}} .$$

- Account for regulatory factors in transcription?



Regulatory factors

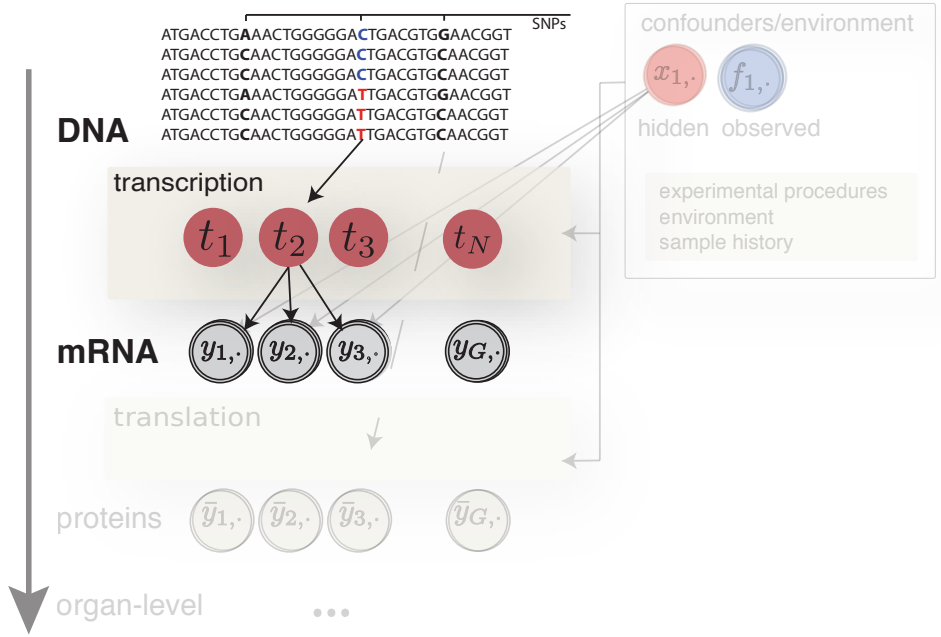
- Regulatory factors:
  - Transcription factors
  - Pathway components
- Mechanistic hypothesis: regulatory factors mediate the association signals to target genes.
- Measuring $\mathbf{T}$?
  - Difficult and expensive
- Learn the unobserved factors $\mathbf{T}$



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

Regulatory factors

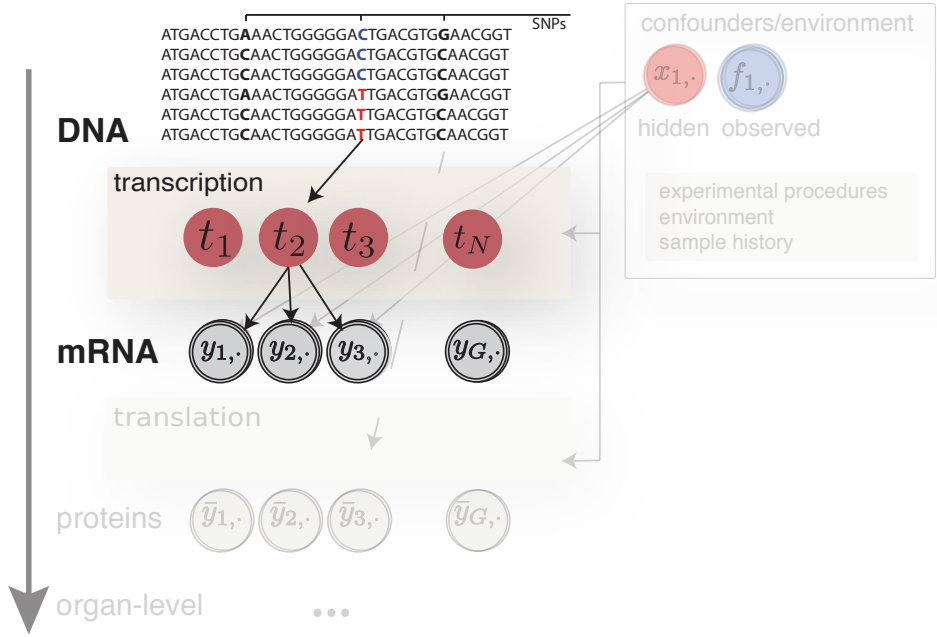
- Regulatory factors:
  - Transcription factors
  - Pathway components
- Mechanistic hypothesis: regulatory factors mediate the association signals to target genes.
- Measuring $\mathbf{T}$?
  - Difficult and expensive
- Learn the unobserved factors $\mathbf{T}$



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

Regulatory factors

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



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

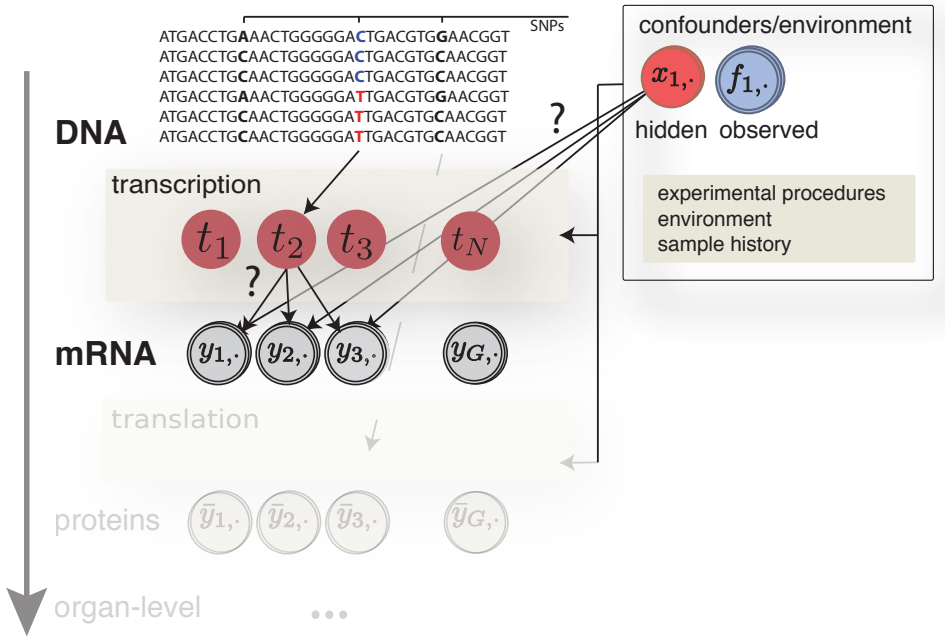
Regulatory factors

A linear model of gene regulation

- Inference of regulatory factors:

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

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



Regulatory factors

A linear model of gene regulation

- Inference of regulatory factors:

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

  - $\mathbf{W}$ is sparse; each factor regulates a specific subset of all genes.

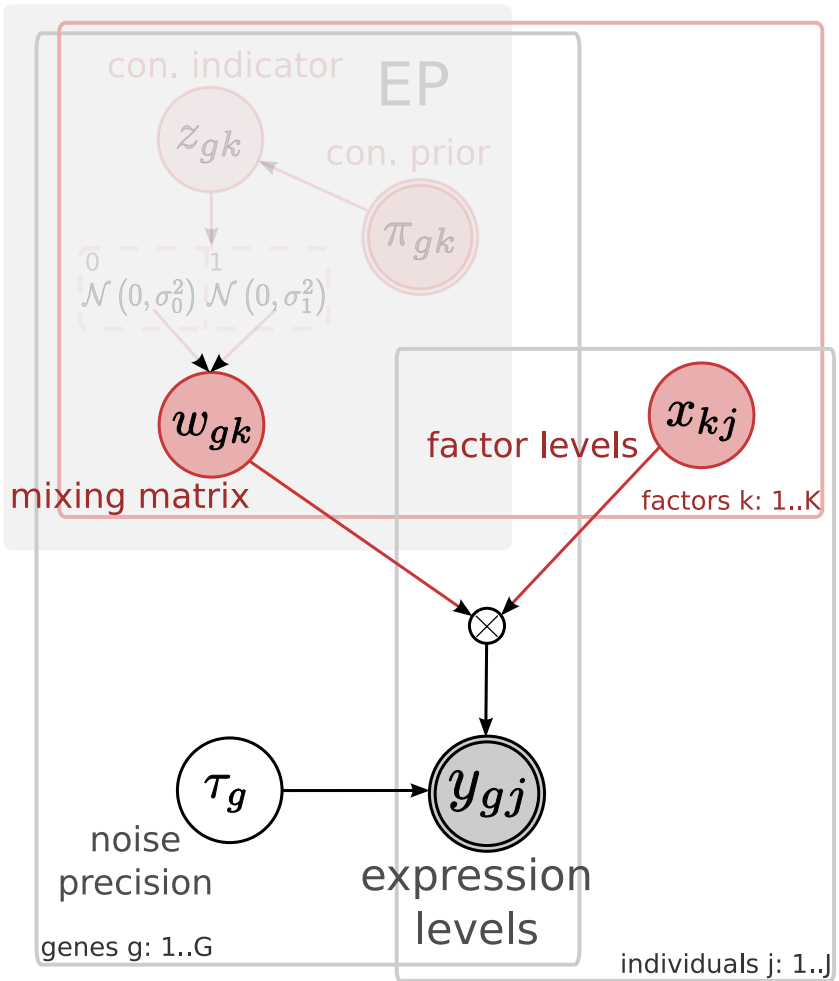
- Incorporation of prior knowledge to render factors interpretable:
  - Transcription factor binding.

Sparse factor analysis

Probabilistic model

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

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

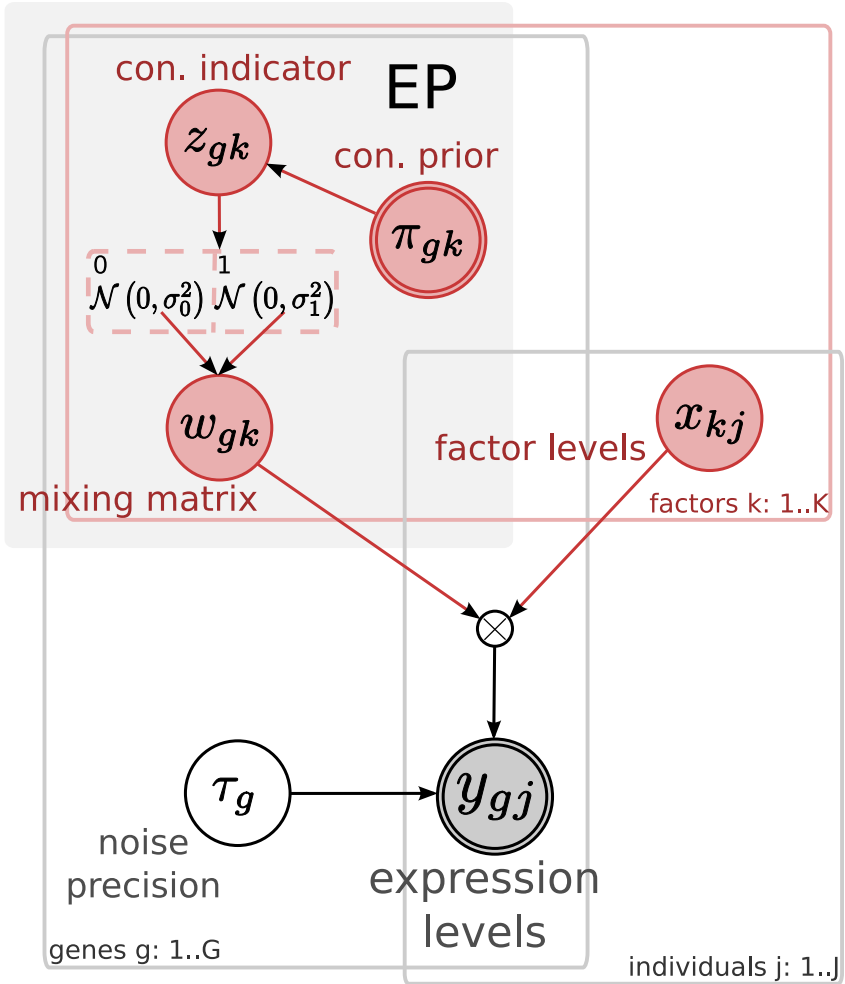


Sparse factor analysis

Probabilistic model

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

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

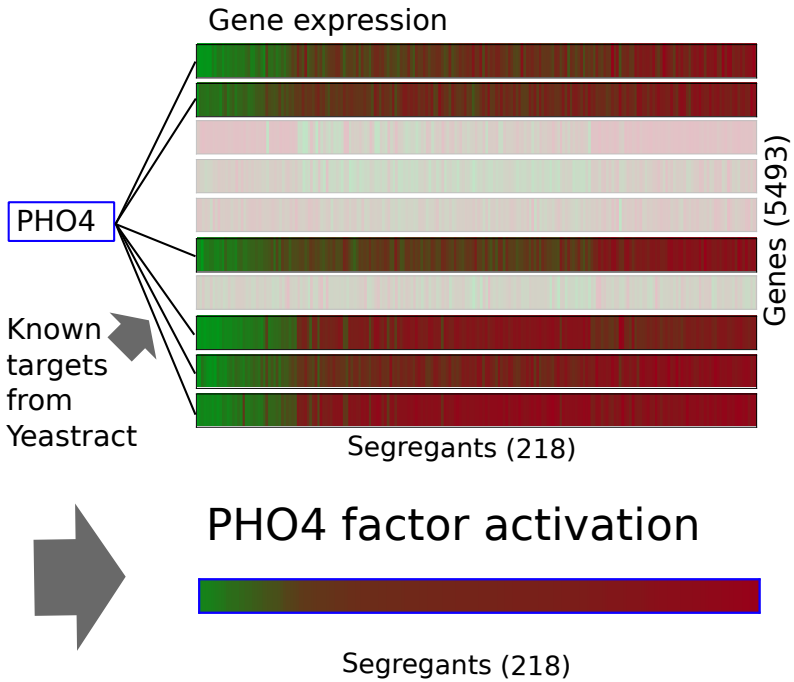


Sparse factor analysis

Probabilistic model

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

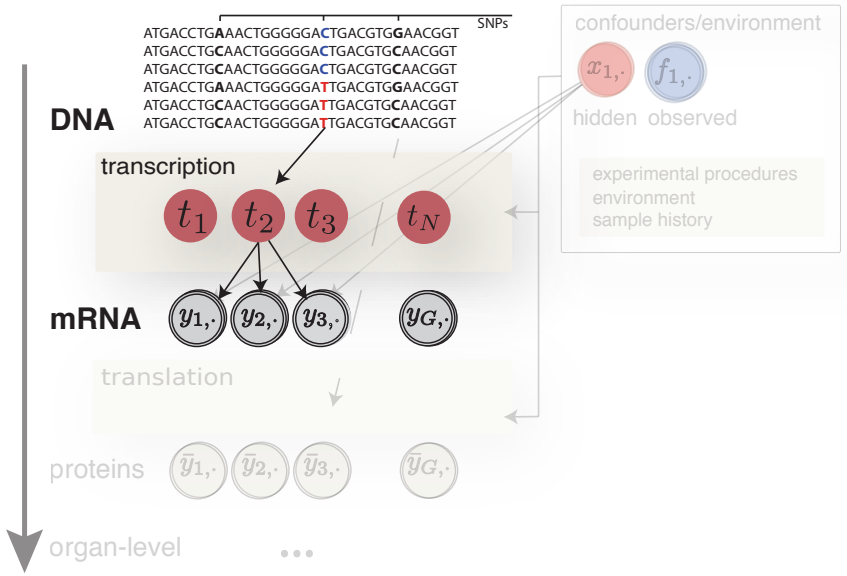
$$P(w_{g,k} \,|\, z_{g,k} = 0) = \mathcal{N}\left(w_{g,k} \,\middle|\, 0, \sigma_0^2\right)$$
$$P(w_{g,k} \,|\, z_{g,k} = 1) = \mathcal{N}\left(w_{g,k} \,\middle|\, 0, \sigma_1^2\right).$$



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.
  2. Genotype-specific regulation of target genes.

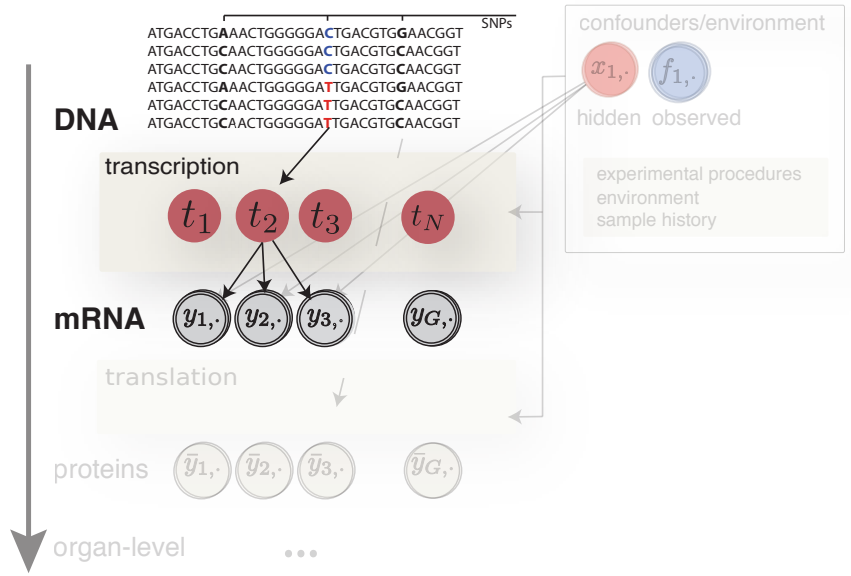


[Smith and Kruglyak, 2008]

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 associations

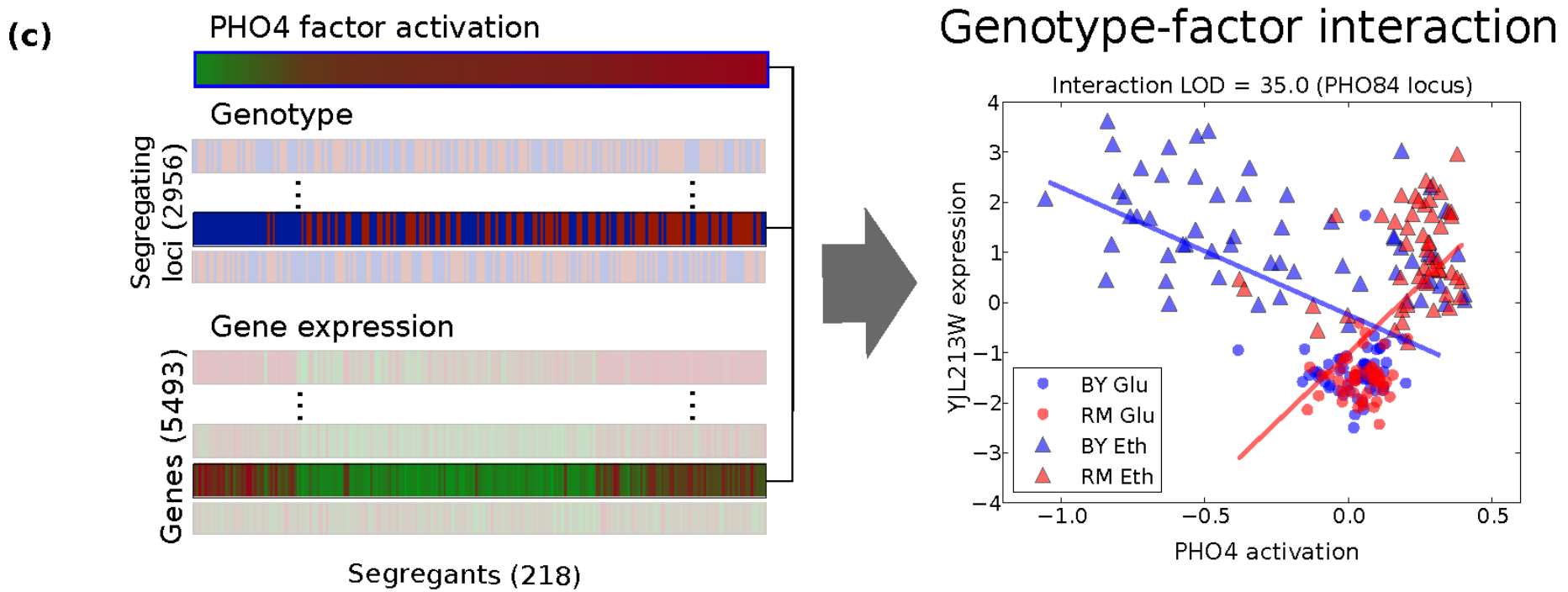
- Application to 108 yeast strains.
  - Genotyped and expression profiled in 2 conditions.
  - Prior knowledge: TF binding affinities.
- Biological hypotheses
  1. Genetic variation (SNPs) may regulate factor activations.
  2. Genotype-specific regulation of target genes.

[Smith and Kruglyak, 2008]

Application to yeast

Factor interactions

- Example of genotype-specific factor regulation.

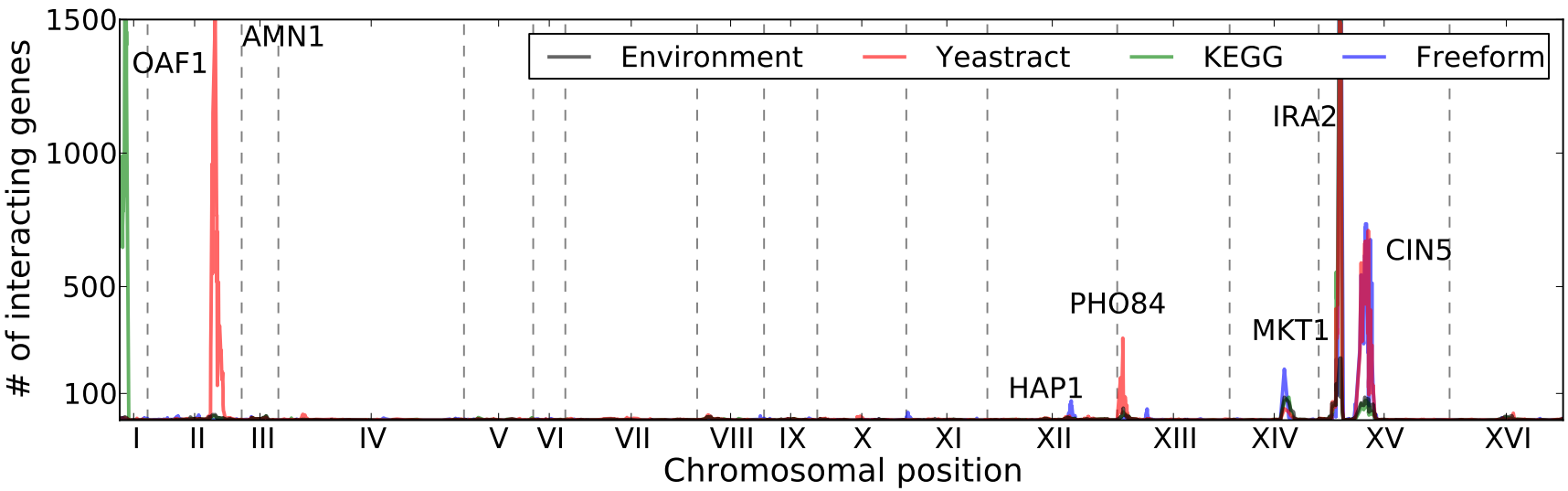


[Parts et al., 2011]

Application to yeast

Factor interactions

- Genome-wide interaction density.



[Parts et al., 2011]

Summary

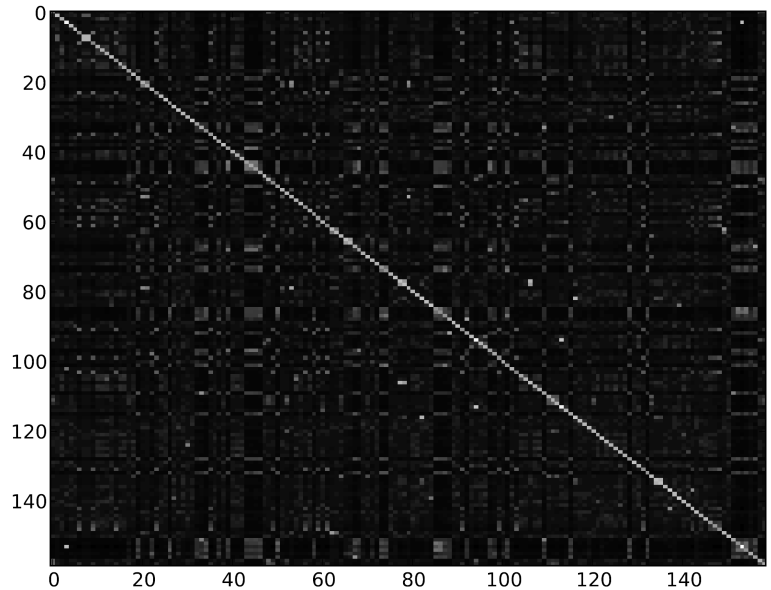
- Accounting for hidden factors can greatly increase the power and meaningfulness of analysis results.
- Joint genetic analysis of cellular features and gene expression for improved interpretability.
- Open source **PEER** software package (Python, R, C++)
  `http://github.com/PMBio/peer`

[Stegle et al., 2012]

Outline

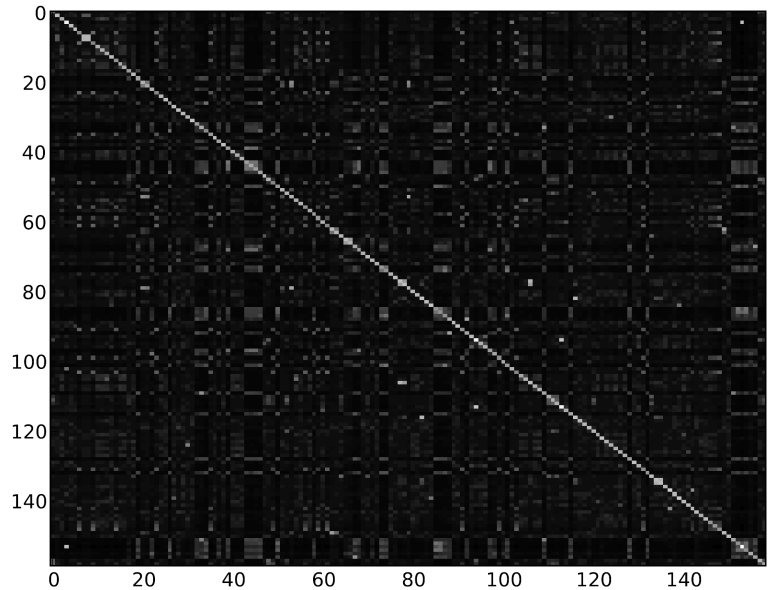
Swedish lines

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



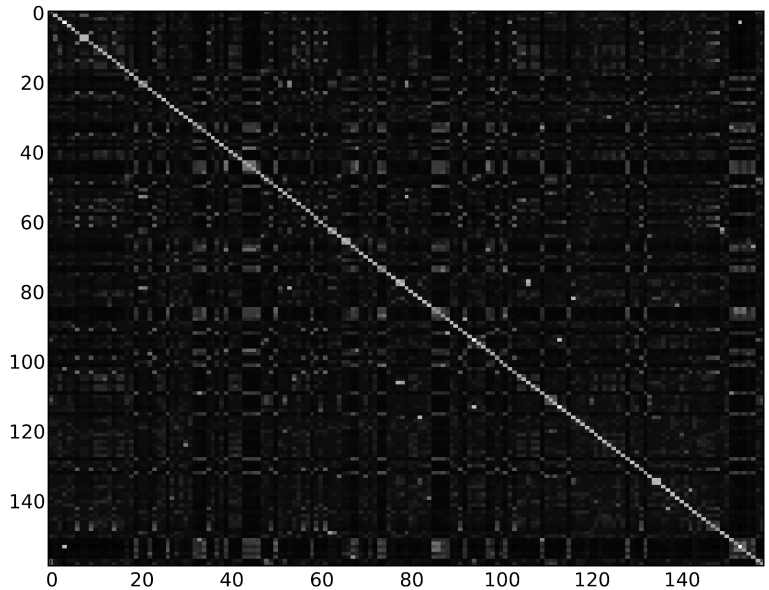
Swedish lines

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



Swedish lines

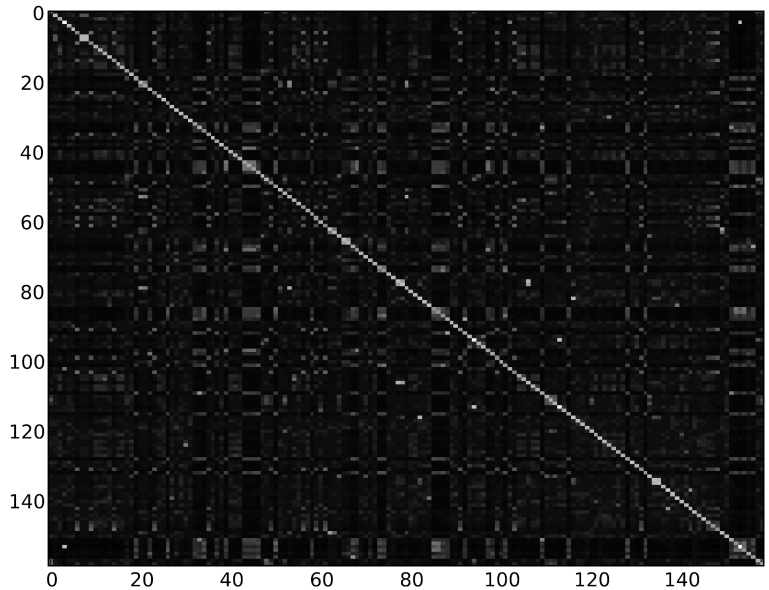
- 160 Lines, extensive population structure
- Genome sequencing
- Transcriptome sequencing
- Bisulfite sequencing
- Two environments; 10C and 16C



Swedish lines

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

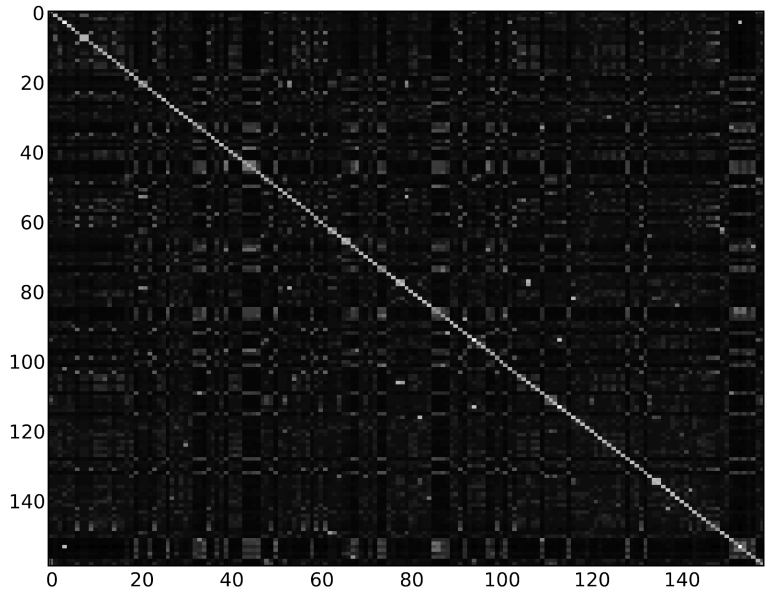
- Two environments; 10C and 16C



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

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

$$\mathbf{y}_{g,e} = \underbrace{\mu_e}_{\text{env effect}} + \underbrace{\sum_{n=1}^{N_{\text{cis}}} b_{n,e}\mathbf{s}_n}_{\textit{cis} \text{ genetics}} + \underbrace{\sum_{l=1}^{L_{\text{trans}}} d_{l,e}\mathbf{s}_l}_{\textit{trans} \text{ genetics}} + \boldsymbol{\psi}_{g,e}.$$

- 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

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

Variance component analysis

A random effect variance estimation model

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

$$\mathbf{y}_{g,e} = \underbrace{\mu_e}_{\text{env effect}} + \underbrace{\sum_{n=1}^{N_{\text{cis}}} b_{n,e}\mathbf{s}_n}_{\textit{cis} \text{ genetics}} + \underbrace{\sum_{l=1}^{L_{\text{trans}}} d_{l,e}\mathbf{s}_l}_{\textit{trans} \text{ genetics}} + \boldsymbol{\psi}_{g,e}.$$

- 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

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

Variance component analysis

A random effect variance estimation model

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

$$\mathbf{y}_{g,e} = \underbrace{\mu_e}_{\text{env effect}} + \underbrace{\sum_{n=1}^{N_{\text{cis}}} b_{n,e}\mathbf{s}_n}_{\textit{cis}\text{ genetics}} + \underbrace{\sum_{l=1}^{L_{\text{trans}}} d_{l,e}\mathbf{s}_l}_{\textit{trans}\text{ genetics}} + \boldsymbol{\psi}_{g,e}.$$

- 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

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

Variance component analysis

A random effect variance estimation model

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

$$\mathbf{y}_{g,e} = \underbrace{\mu_e}_{\text{env effect}} + \underbrace{\sum_{n=1}^{N_{\text{cis}}} b_{n,e}\mathbf{s}_n}_{\textit{cis}\text{ genetics}} + \underbrace{\sum_{l=1}^{L_{\text{trans}}} d_{l,e}\mathbf{s}_l}_{\textit{trans}\text{ genetics}} + \boldsymbol{\psi}_{g,e}.$$

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

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

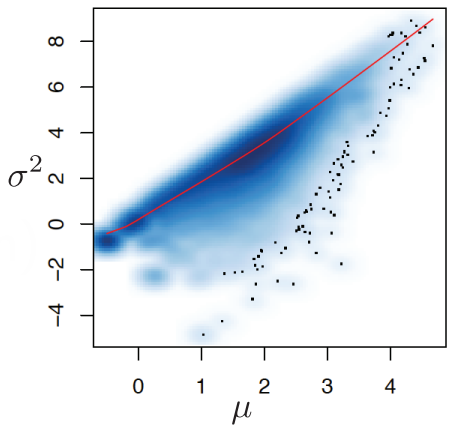
  - Uncorrelated noise covariance

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

Variance component analysis

Observational model needs to acknowledge count statistics

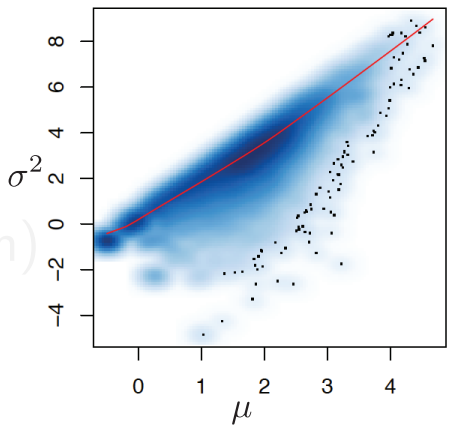
- The observed quantities are read counts $\mathbf{c}_{g,e}$ and not true expression levels
- Log-normal model on the Poisson rates
  - MCMC inference
  - Approximate Bayesian inference
  - Variance stabilizing transform (Anscombe transform)



Variance component analysis

Observational model needs to acknowledge count statistics

- 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)

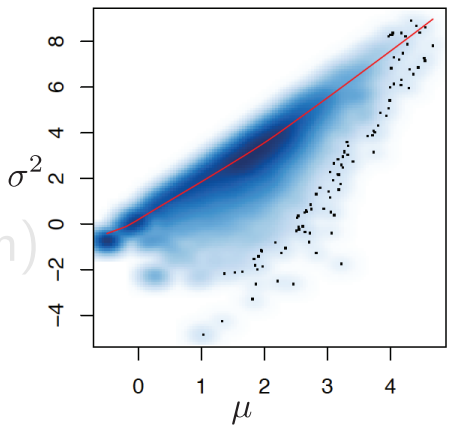


$$p(\mathbf{c}_{g,e} \,|\, \mathbf{C}(\boldsymbol{\theta})) = \mathcal{N}\left(\mathbf{y}_{g,e} \,|\, \mathbf{0}, \mathbf{C}(\boldsymbol{\theta})\right) \prod_{j=1}^{J} \underbrace{\text{Poisson}(c_{g,e,j} \,|\, e^{y_{g,e,j}})}_{\text{Poisson observation model}}$$

Variance component analysis

Observational model needs to acknowledge count statistics

- 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)

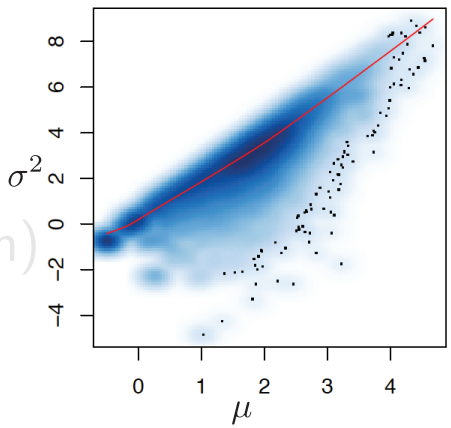


$$p(\mathbf{c}_{g,e} \,|\, \mathbf{C}(\boldsymbol{\theta})) = \mathcal{N}\left(\mathbf{y}_{g,e} \,|\, \mathbf{0}, \mathbf{C}(\boldsymbol{\theta})\right) \prod_{j=1}^{J} \underbrace{\text{Poisson}(c_{g,e,j} \,|\, e^{y_{g,e,j}})}_{\text{Poisson observation model}}$$

Variance component analysis

Observational model needs to acknowledge count statistics

- 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)



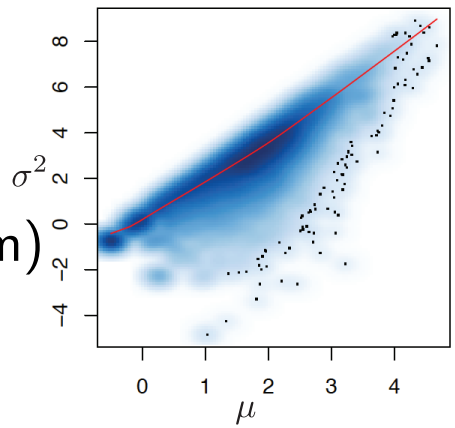
$$p(\mathbf{c}_{g,e} \,|\, \mathbf{C}(\boldsymbol{\theta})) = \mathcal{N}\left(\mathbf{y}_{g,e} \,|\, \mathbf{0}, \mathbf{C}(\boldsymbol{\theta})\right) \prod_{j=1}^{J} \underbrace{\text{Poisson}(c_{g,e,j} \,|\, e^{y_{g,e,j}})}_{\text{Poisson observation model}}$$

$$\approx \mathcal{N}\left(\mathbf{y}_{g,e} \,|\, \mathbf{0}, \mathbf{C}(\boldsymbol{\theta})\right) \prod_{j=1}^{J} \underbrace{\mathcal{N}\left(y_{g,e,j} \,|\, \mu_{g,e,j}, \sigma^2_{g,e,j}\right)}_{\text{Gaussian approximation}}$$

Variance component analysis

Observational model needs to acknowledge count statistics

- 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)



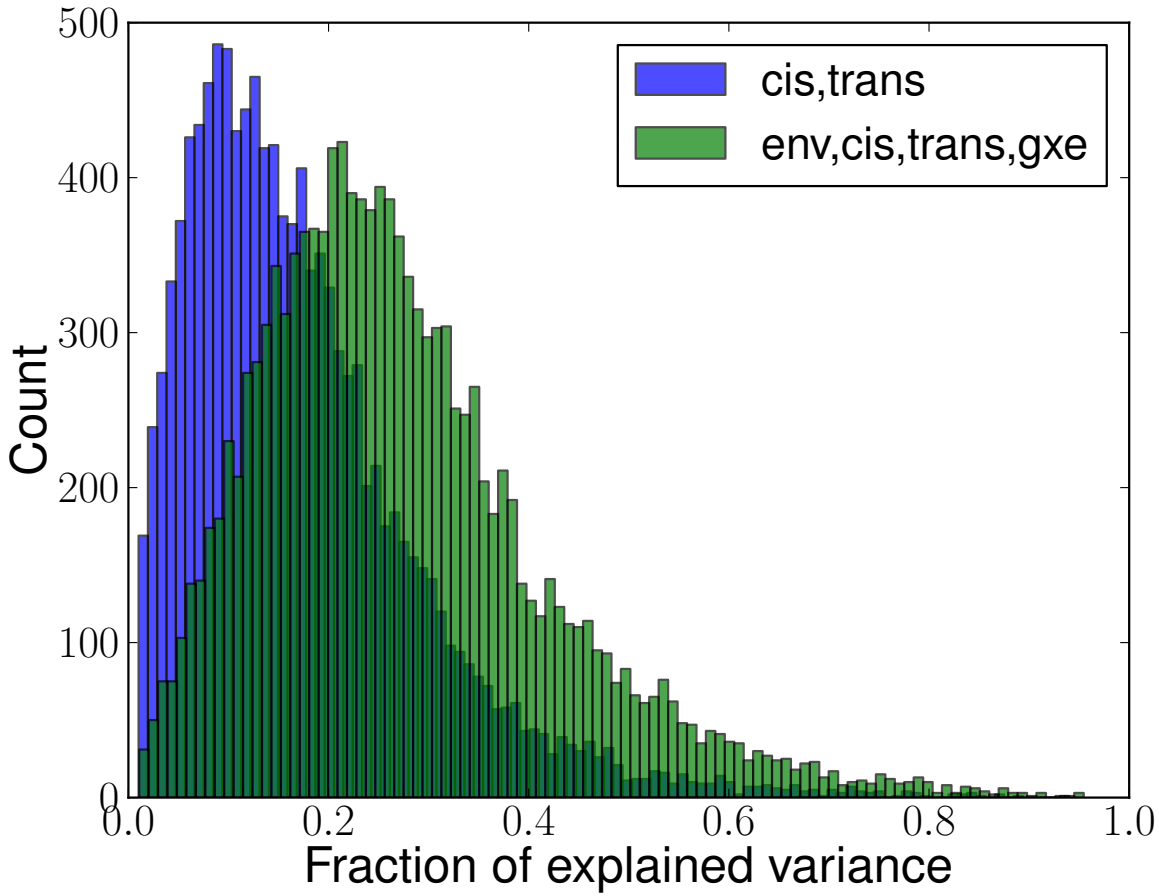
$$p(\mathbf{c}_{g,e} \,|\, \mathbf{C}(\boldsymbol{\theta})) = \mathcal{N}\left(\mathbf{y}_{g,e} \,|\, \mathbf{0}, \mathbf{C}(\boldsymbol{\theta})\right) \prod_{j=1}^{J} \underbrace{\text{Poisson}(c_{g,e,j} \,|\, e^{y_{g,e,j}})}_{\text{Poisson observation model}}$$

$$\approx \mathcal{N}\left(\mathbf{y}_{g,e} \,|\, \mathbf{0}, \mathbf{C}(\boldsymbol{\theta})\right) \prod_{j=1}^{J} \underbrace{\mathcal{N}\left(y_{g,e,j} \,|\, \mu_{g,e,j}, \sigma^2_{g,e,j}\right)}_{\text{Gaussian approximation}}$$

Variance component analysis

Impact of environment

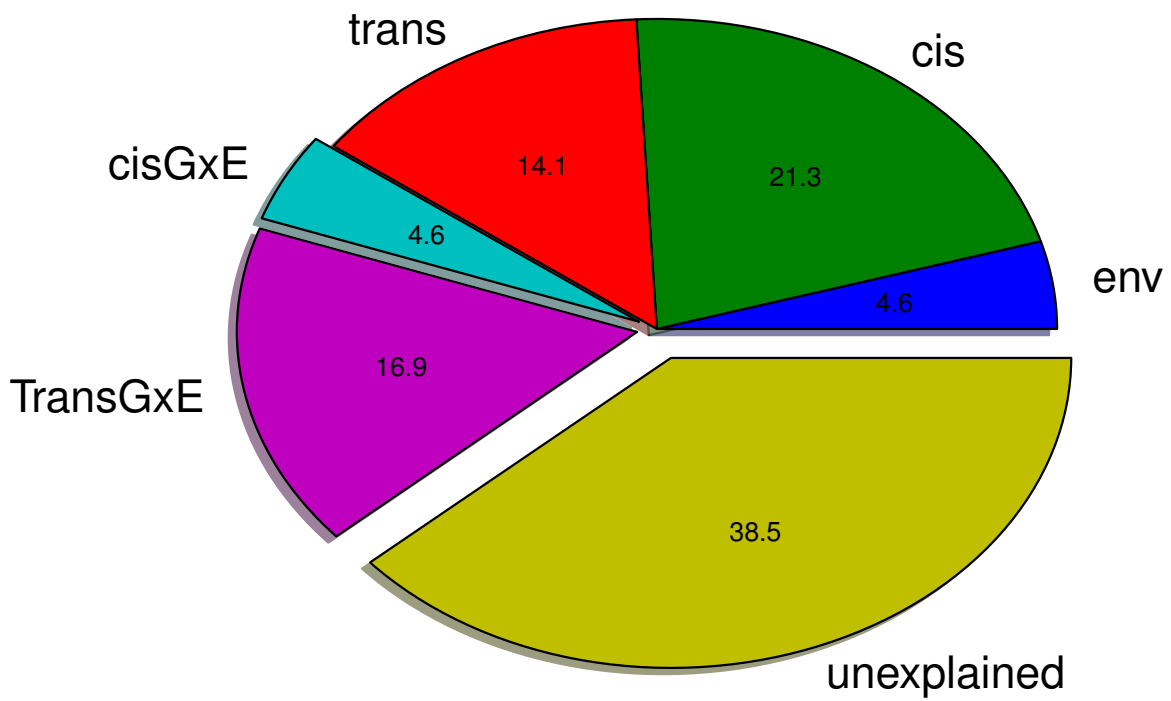
- Environment greatly affects heritability
- Absolute environment contribution small



Variance component analysis

Impact of environment

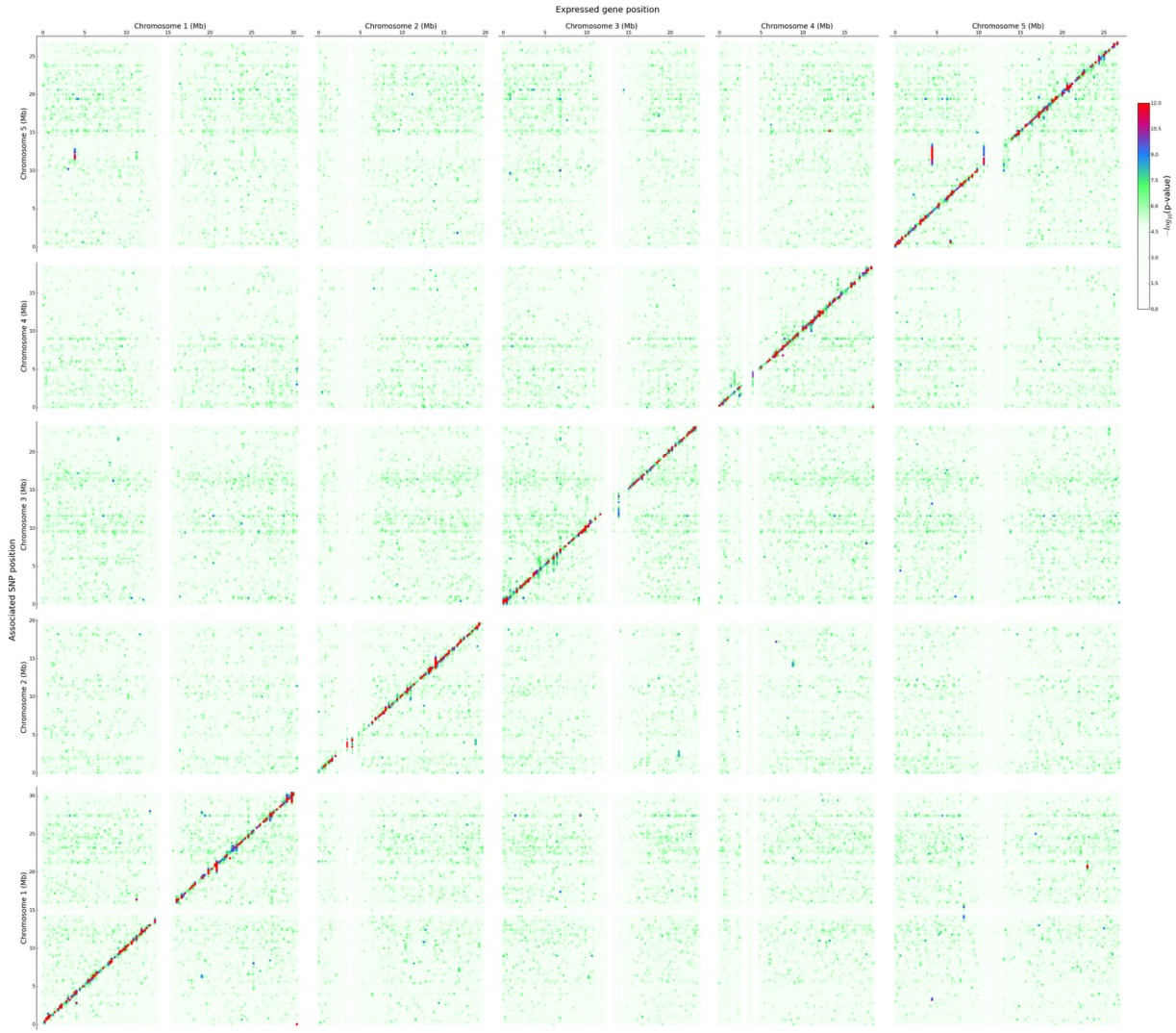
- Environment greatly affects heritability
- Absolute environment contribution small



(Average across upper 50% quantile of heritable genes)

GWAS

main effects

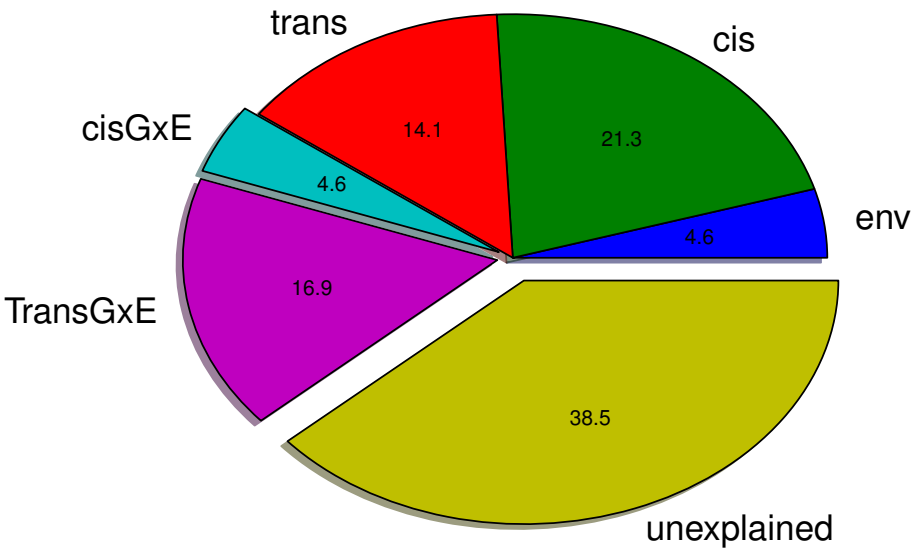


GWAS

GxE effects

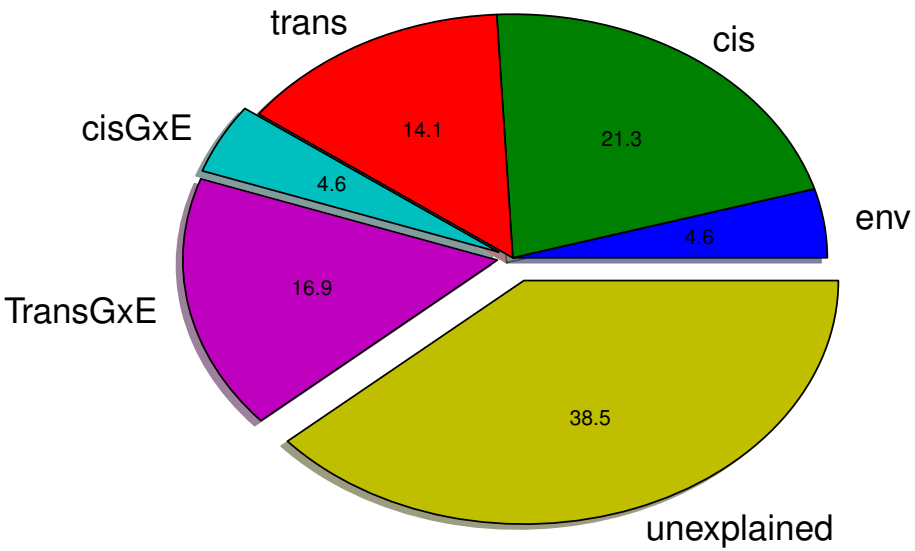
GWAS

GxE effects



GWAS

GxE effects



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

Outline

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

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

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

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

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

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

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.