Accounting for heterogeneity between individuals and single cells
(using linear mixed models)

Machine Learning in Personalized Medicine Summer school 2015

Oliver Stegle
European Bioinformatics Institute

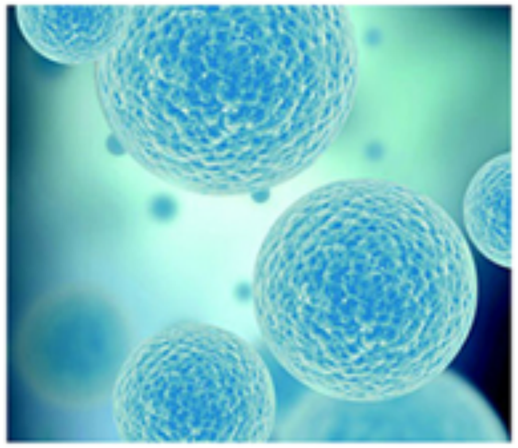
EMBL-EBI

Gene expression heterogeneity between individuals and single cells

variation of interest

population variation

genetic associations with phenotype



single-cell variation

differentiation processes
Correlations between genes

Gene expression heterogeneity between individuals and single cells

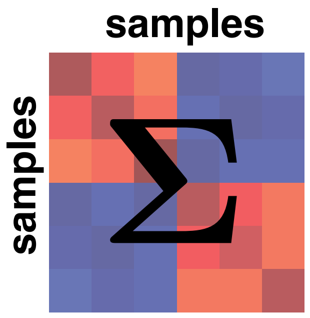
variation of interest

confounding

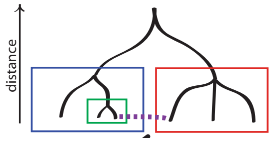
population variation

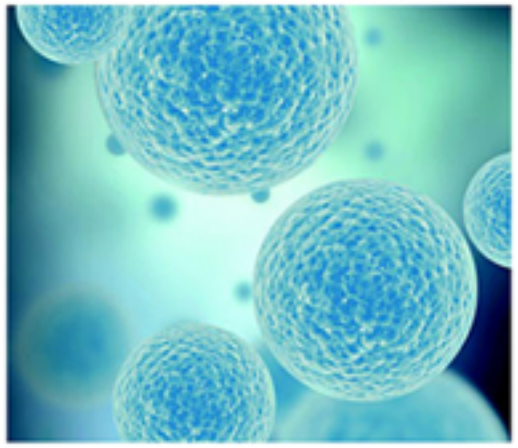
genetic associations with phenotype

sample covariance



genetic
non-genetic
(batch, environment)

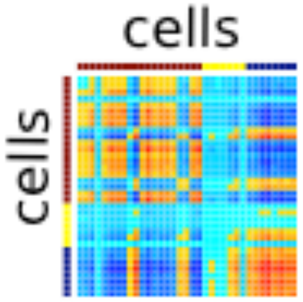




single-cell variation

differentiation processes
Correlations between genes

cell covariance

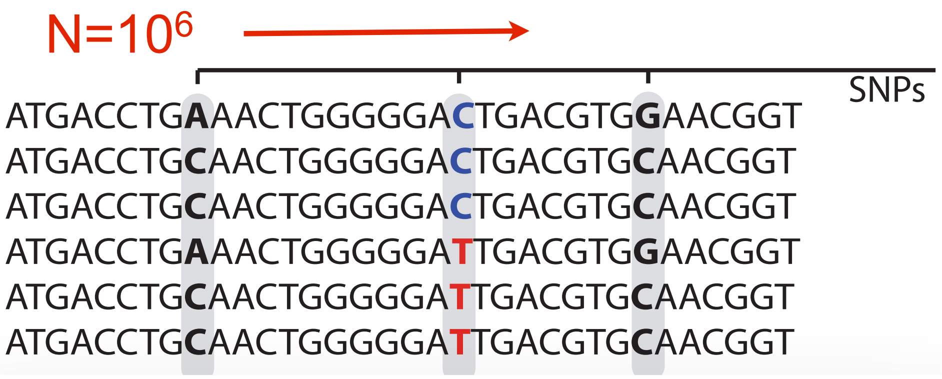


cell cycle
stress

Multi-omics association genetics

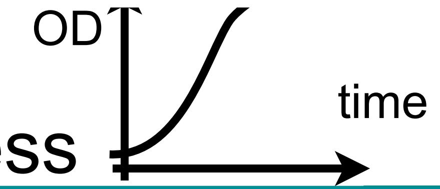
natural
randomized
perturbation!

DNA



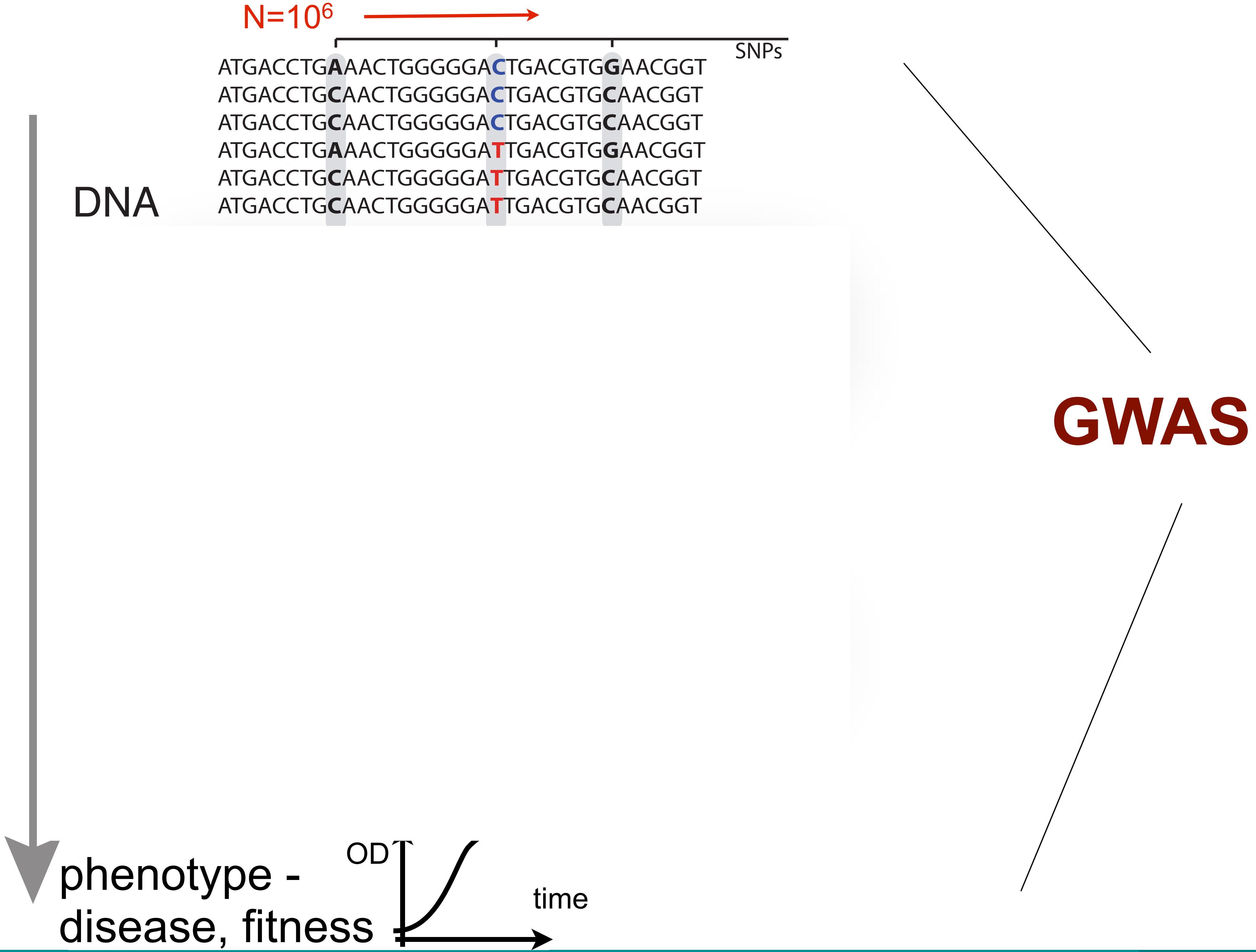
ATGACCTG**A**AACTGGGGGA**C**TGACGTG**G**AACGGT
ATGACCTG**C**AACTGGGGGA**C**TGACGTG**C**AACGGT
ATGACCTG**C**AACTGGGGGA**C**TGACGTG**C**AACGGT
ATGACCTG**A**AACTGGGGGA**T**TGACGTG**G**AACGGT
ATGACCTG**C**AACTGGGGGA**T**TGACGTG**C**AACGGT
ATGACCTG**C**AACTGGGGGA**T**TGACGTG**C**AACGGT

phenotype -
disease, fitness



Multi-omics association genetics

natural
randomized
perturbation!



N=10^6

SNPs

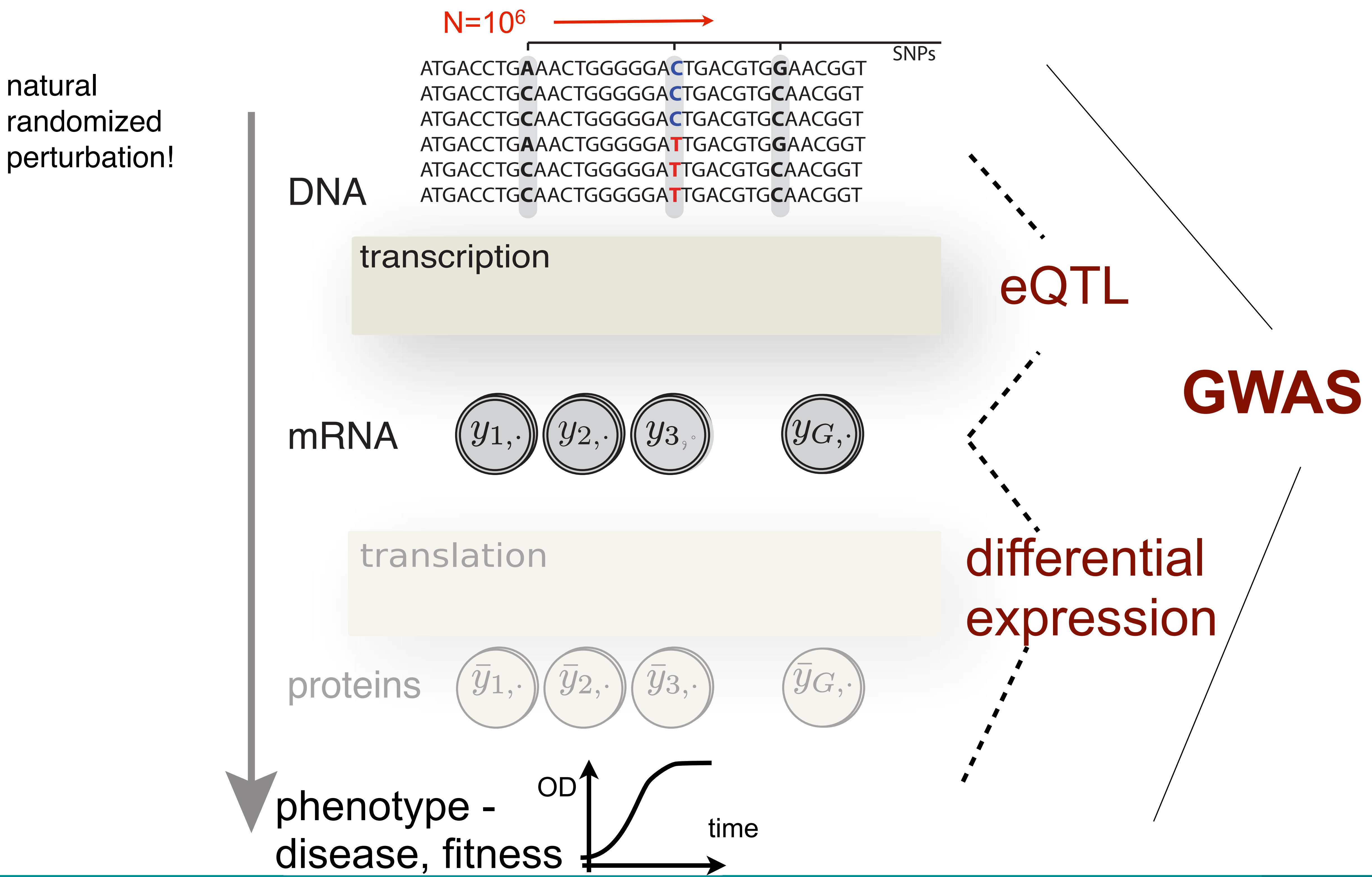
```
ATGACCTGAAACTGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT
```

DNA

phenotype -
disease, fitness

GWAS

Multi-omics association genetics



natural randomized perturbation!

N=10^6

SNPs

ATGACCTG**A**AACTGGGGA**C**TGACGTG**G**AACGGT
ATGACCTG**C**AACTGGGGA**C**TGACGTG**C**AACGGT
ATGACCTG**C**AACTGGGGA**C**TGACGTG**C**AACGGT
ATGACCTG**A**AACTGGGGA**T**TGACGTG**G**AACGGT
ATGACCTG**C**AACTGGGGA**T**TGACGTG**C**AACGGT
ATGACCTG**C**AACTGGGGA**T**TGACGTG**C**AACGGT

DNA

transcription

mRNA $y_{1,\cdot}$ $y_{2,\cdot}$ $y_{3,\cdot}$ $y_{G,\cdot}$

translation

proteins $\bar{y}_{1,\cdot}$ $\bar{y}_{2,\cdot}$ $\bar{y}_{3,\cdot}$ $\bar{y}_{G,\cdot}$

phenotype -
disease, fitness

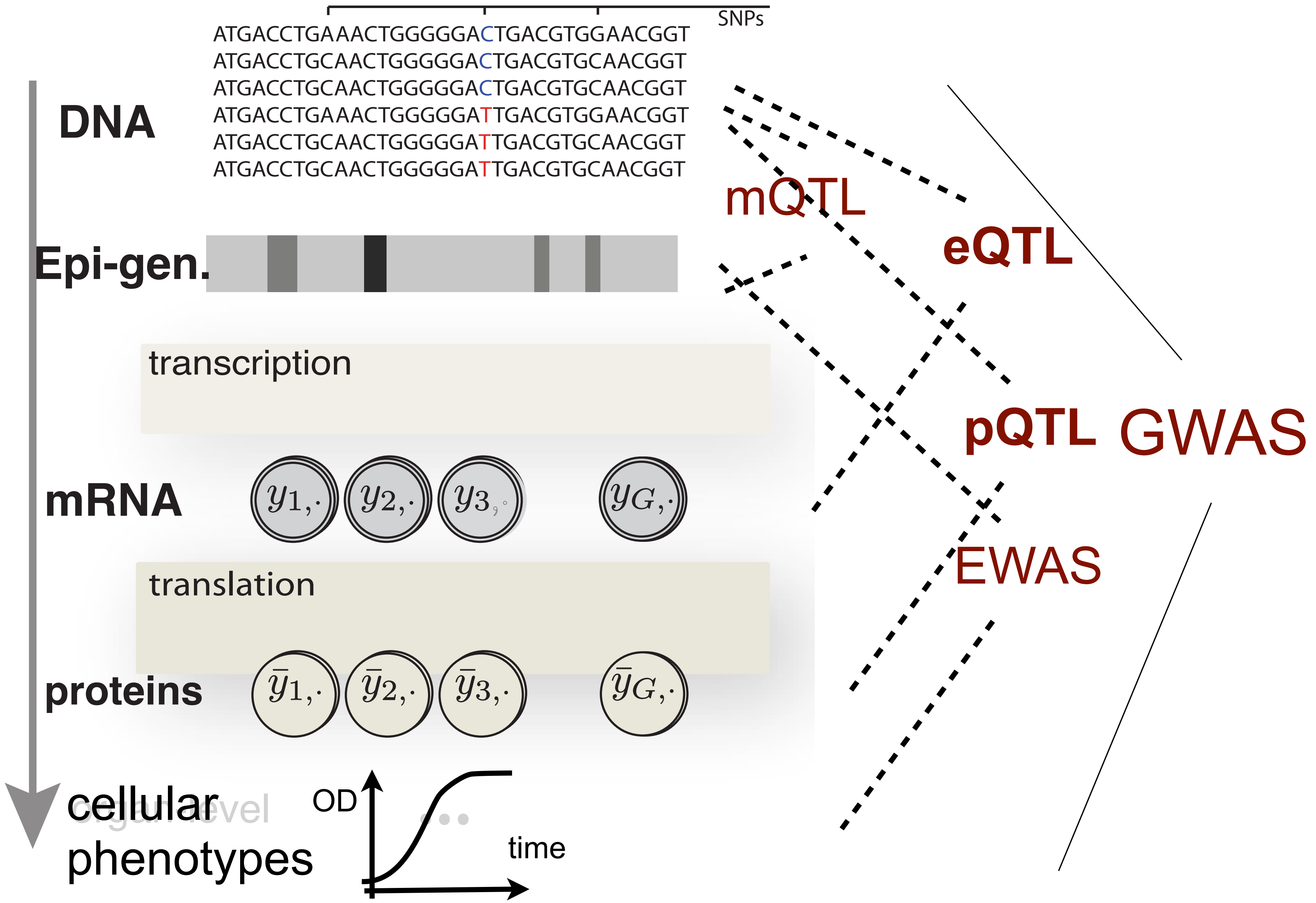
eQTL

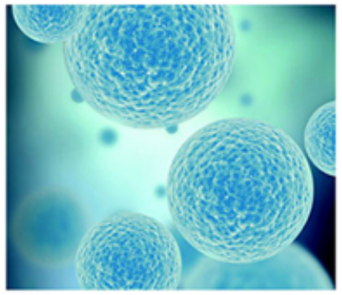
differential
expression

GWAS

Multi-omics association genetics

- Open access iPSC resource for the wider biomedical community
- Aims to discover how genetic variation affects cellular function in iPSC and leads to disease phenotypes

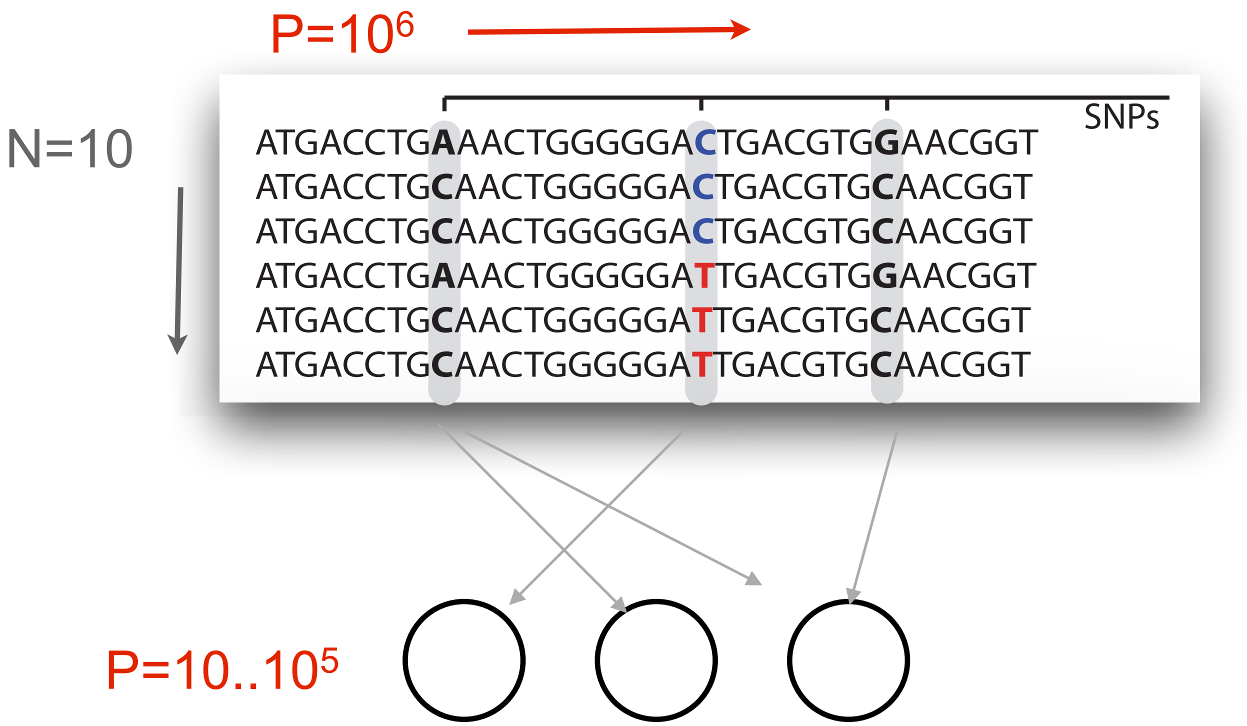




single-cell

Big data in molecular genetics: statistical challenges and opportunities

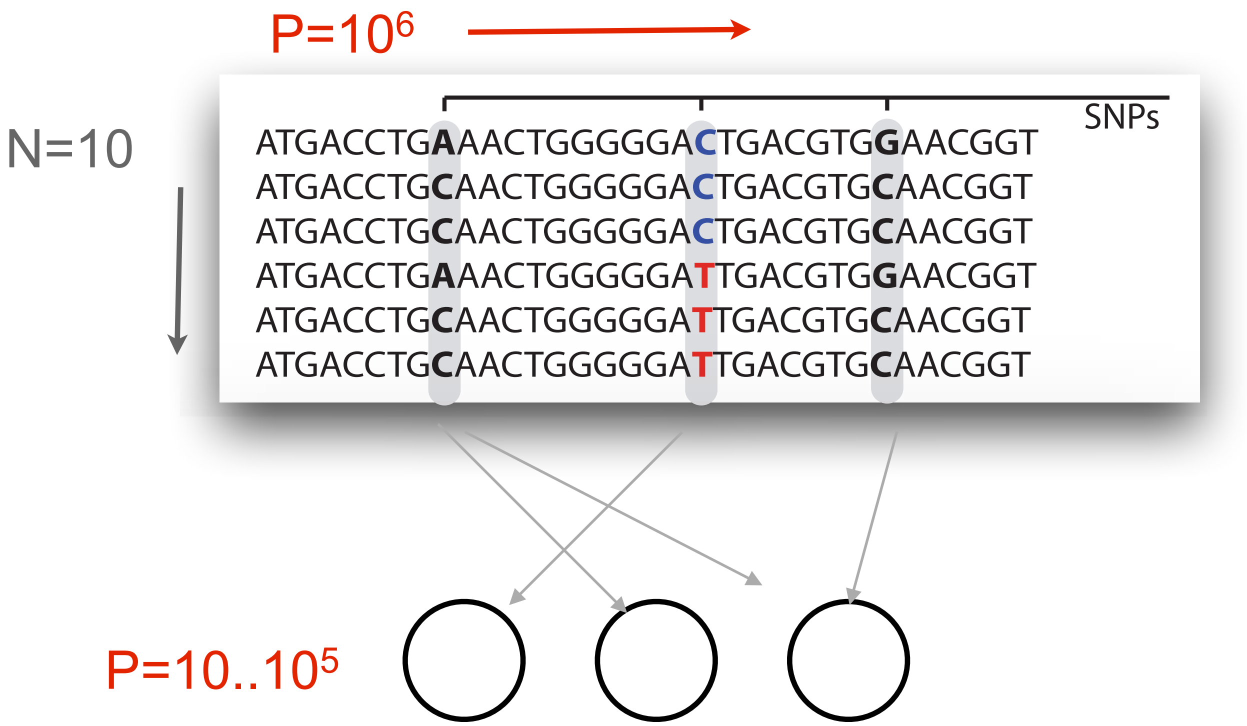
- **Challenge**: Large-scale multiple testing problem:
  - Need to consider potentially millions of loci and adjust for multiple testing.
  - Account for **confounding**
  - Need appropriate corrections (e.g. False Discovery Rate)
  - Scalability to **large cohorts (computation, not storage)**



P=10^6
N=10
SNPs
ATGACCTGAAACTGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT
P=10..10^5

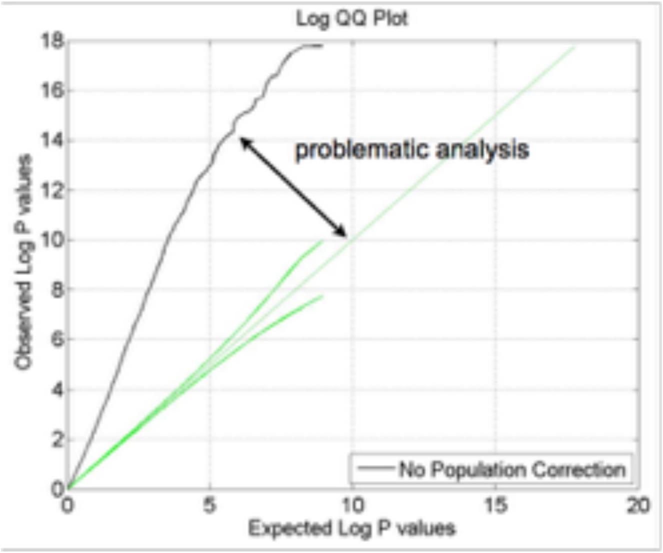
Big data in molecular genetics: statistical challenges and opportunities

- **Challenge**: Large-scale multiple testing problem:
  - Need to consider potentially millions of loci and adjust for multiple testing.
  - Account for **confounding**
  - Need appropriate corrections (e.g. False Discovery Rate)
  - Scalability to **large cohorts (computation, not storage)**



$P=10^6$
N=10
SNPs
ATGACCTGAAACTGGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
$P=10..10^5$

- **Win:** Large dataset allow to test modeling assumptions / fit better models
  - **Inference of confounding structures**
  - Not possible before large-scale hypothesis testing/large datasets
  - More power due to large datasets
  - Gain in power by joint analysis of **multiple traits**



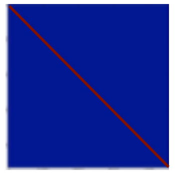
Hidden structure: population structure

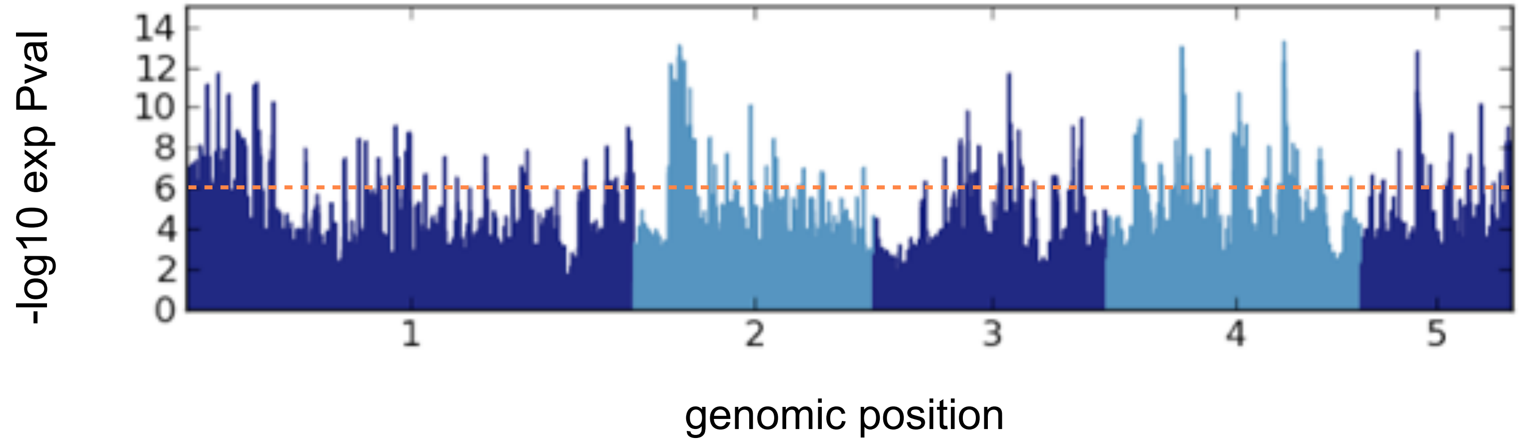
LINEAR MODEL

$$\mathbf{y} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \beta + \psi$$

pheno $\quad$ $\mathbf{x}$ **SNP** $\quad$ **noise**

NOISE

$\psi \sim \mathcal{N}\left(\mathbf{0}, \sigma_e^2\right.$  $\left.\right)$

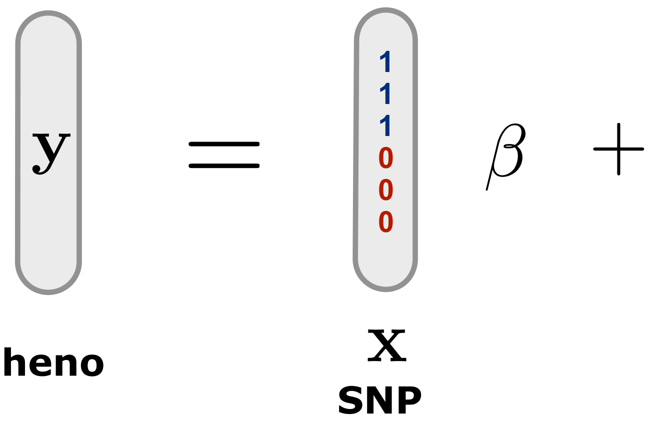


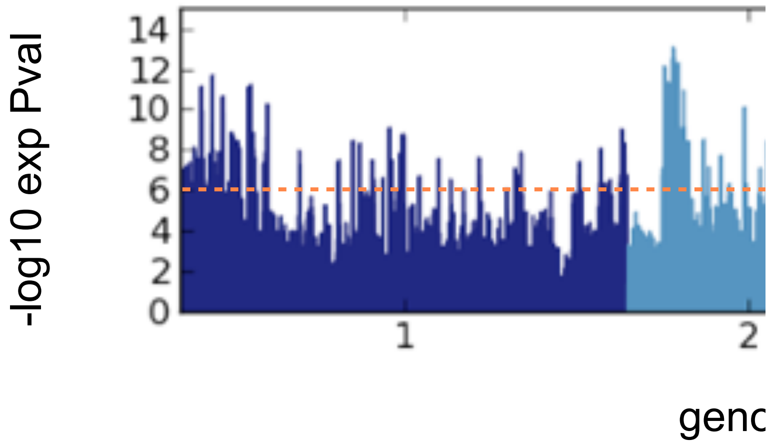
flowering time
A. Thaliana

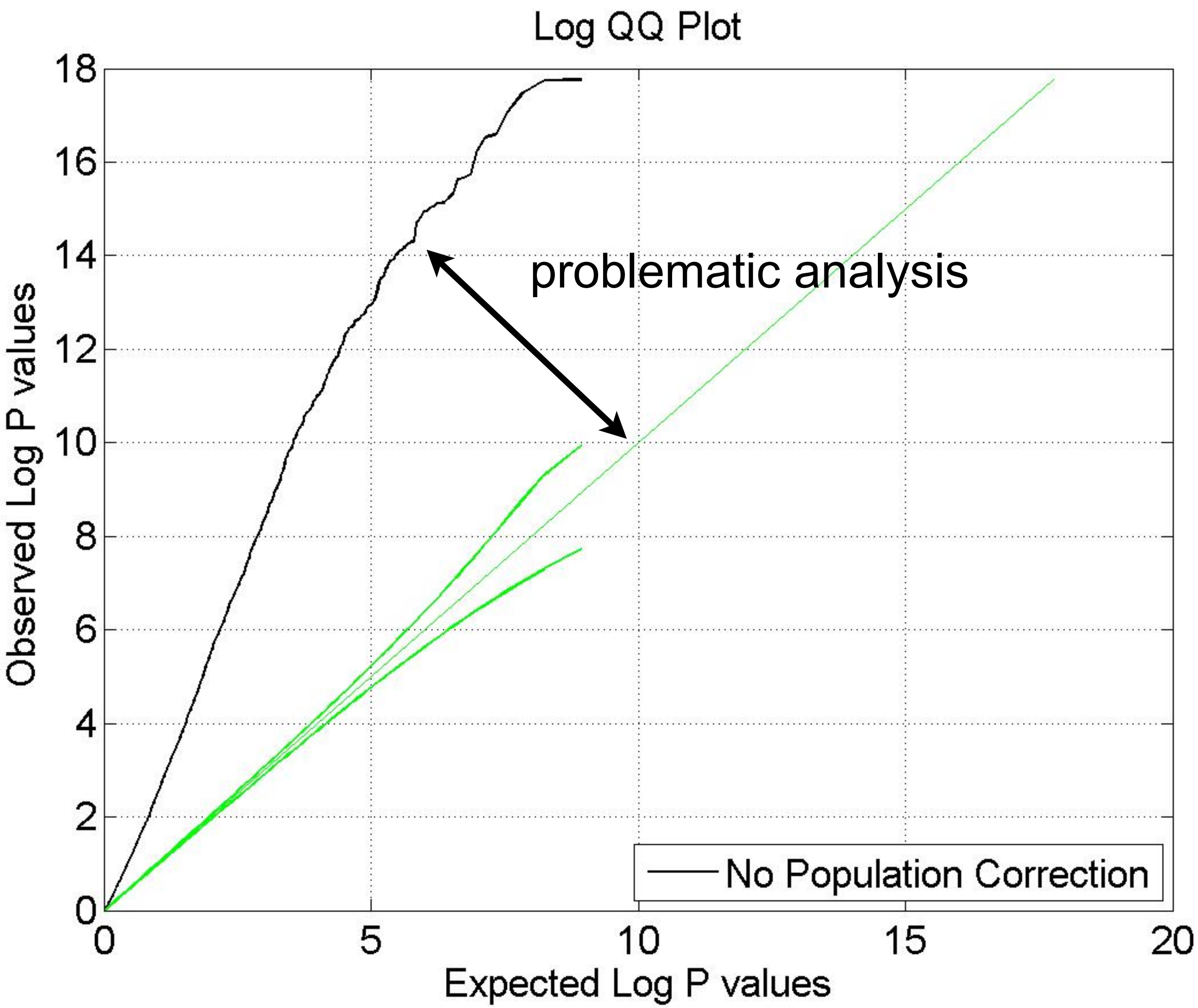
Flowering in A. thaliana

Hidden structure: population structure

LINEAR MODEL







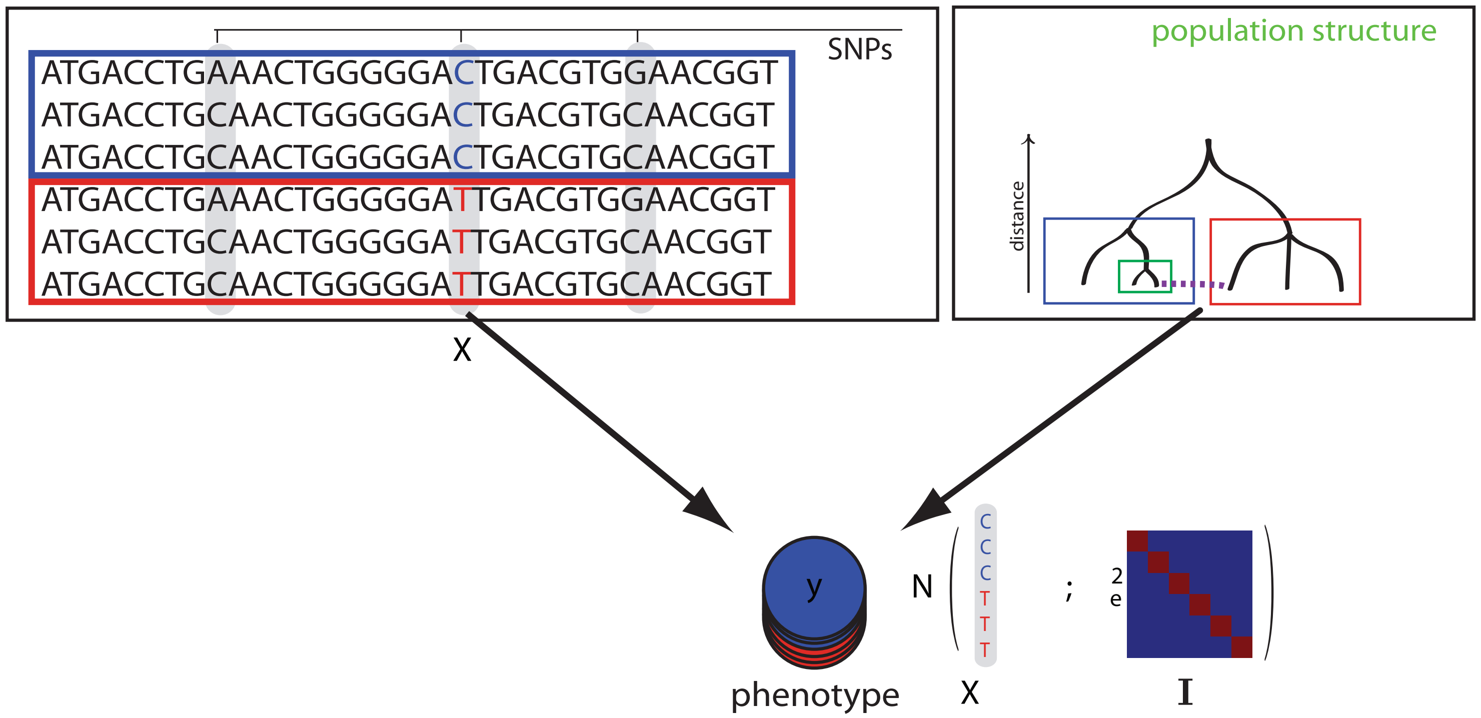
problematic analysis

Flowering in A. thaliana

Hidden structure: population structure

population 1

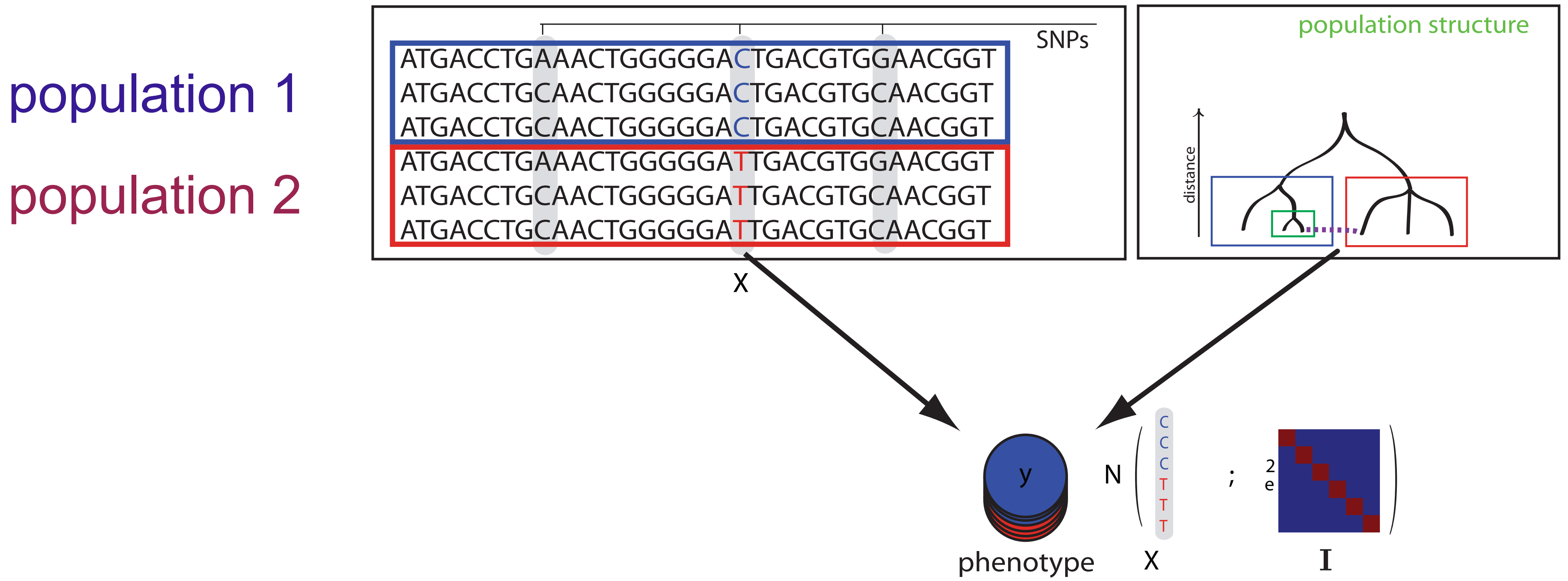
population 2



ATGACCTGAAACTGGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT

Hidden structure: population structure

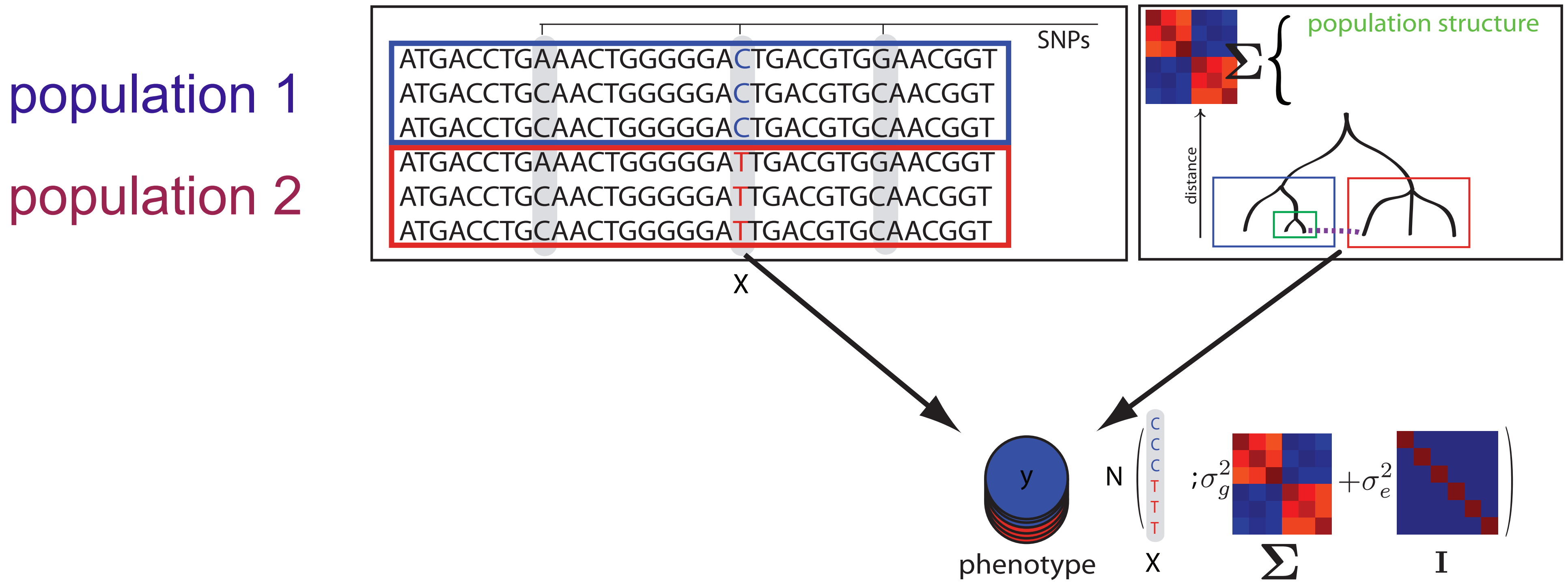
- Population structure (genetic)



ATGACCTGAAACTGGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT

Hidden structure: population structure

- Population structure (genetic)

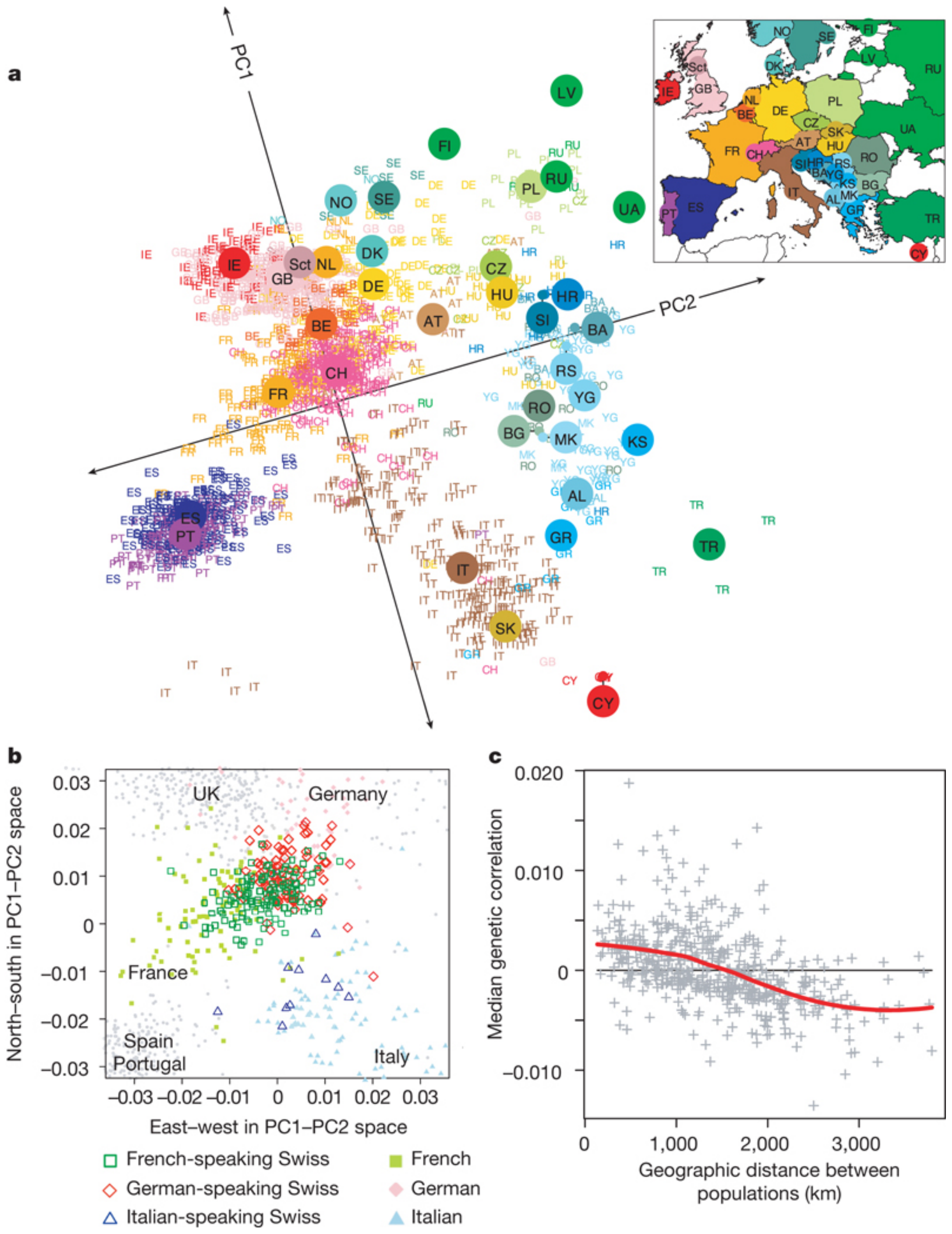

population 1
population 2
SNPs
ATGACCTGAAACTGGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
X
Σ
population structure
distance
y
phenotype
N
C C C T T T
;σ_g^2
+σ_e^2
X
Σ
I

$$\mathbf{y}_g = \underbrace{\mathbf{s}_i\beta_{i,g}}_{\text{genetic}} + \underbrace{\mathbf{u}}_{\text{confounding}} + \underbrace{\boldsymbol{\epsilon}_g}_{\text{noise}} \qquad \mathbf{u} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$$

- Estimate $\boldsymbol{\Sigma}$
  - Population structure: genotype data

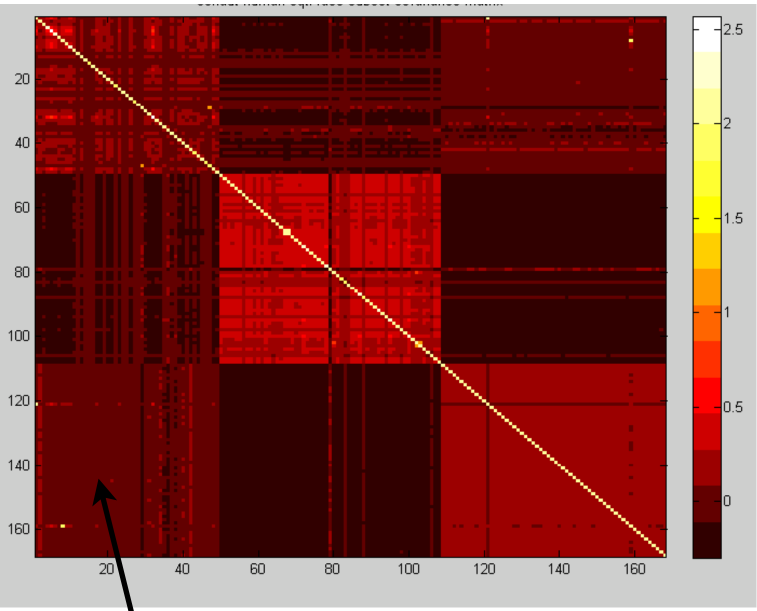
Genetics, Kang et al. 2008
Lippert et al. 2011
Zhou & Stepens, 2012

Hidden structure: population structure



J Novembre *et al. Nature* **000**, 1-4 (2008) doi:10.1038/nature07331

Hidden structure: population structure



individuals

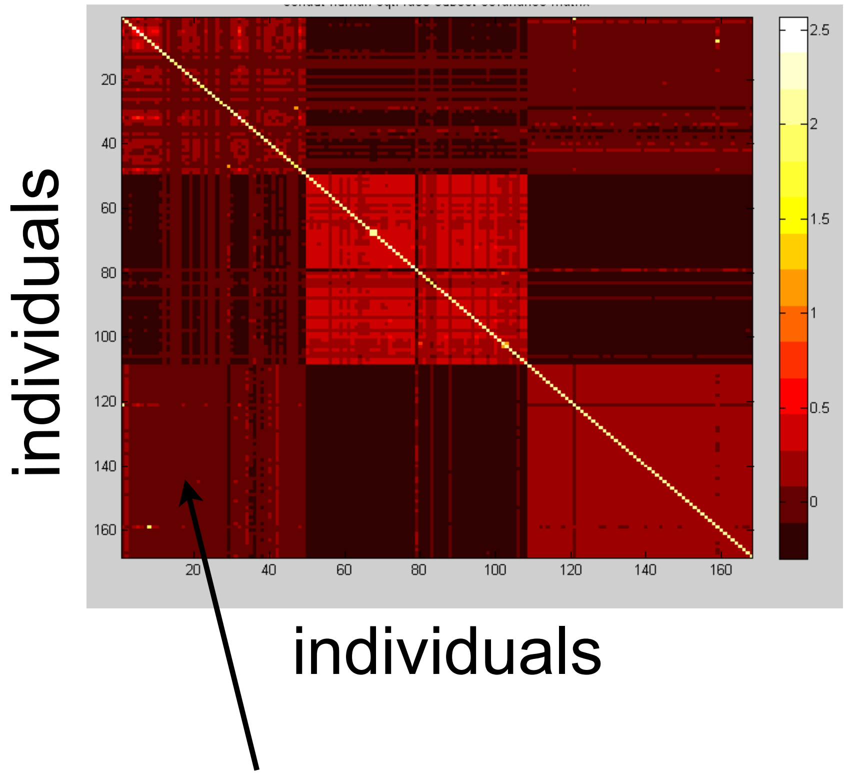
individuals

genetic concordance

adapted from Christoph Lippert

Hidden structure: population structure

3 populations

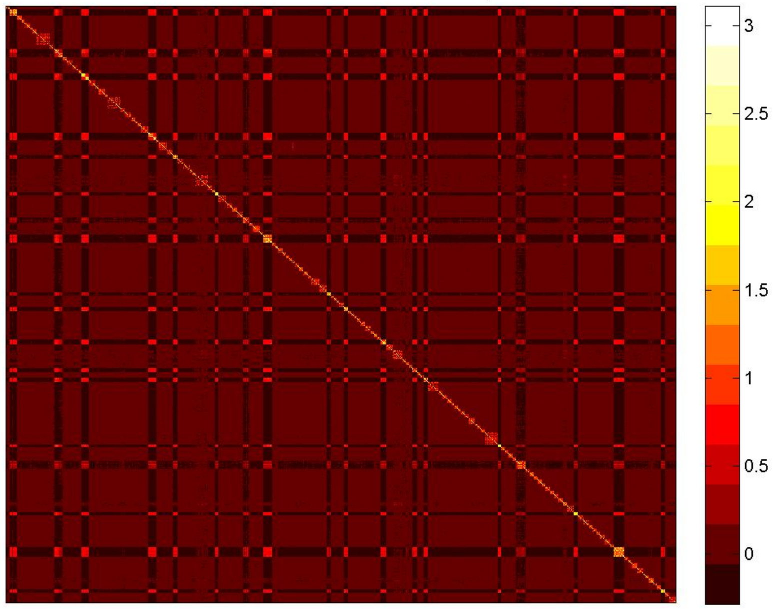


individuals

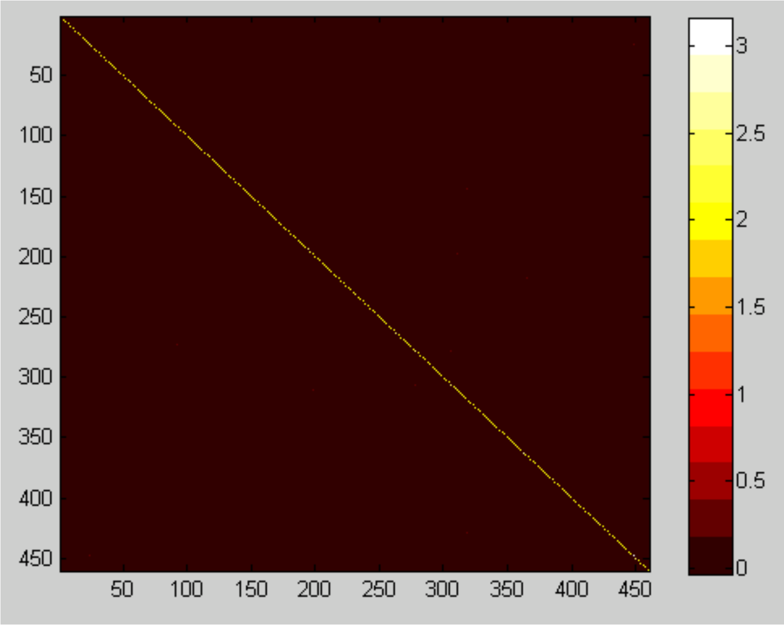
individuals

genetic concordance

families



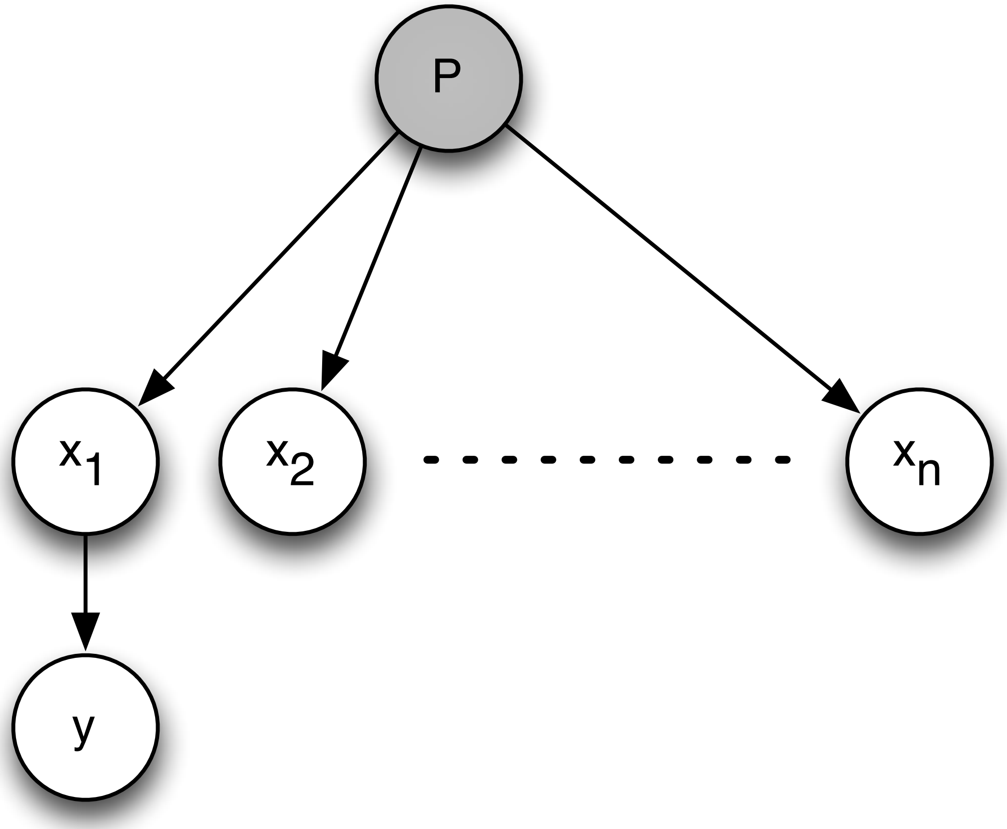
no structure



adapted from Christoph Lippert

Hidden structure: population structure

- genetic confounding (population structure)



SNPs

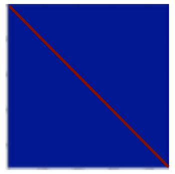
phenotype

Hidden structure: population structure

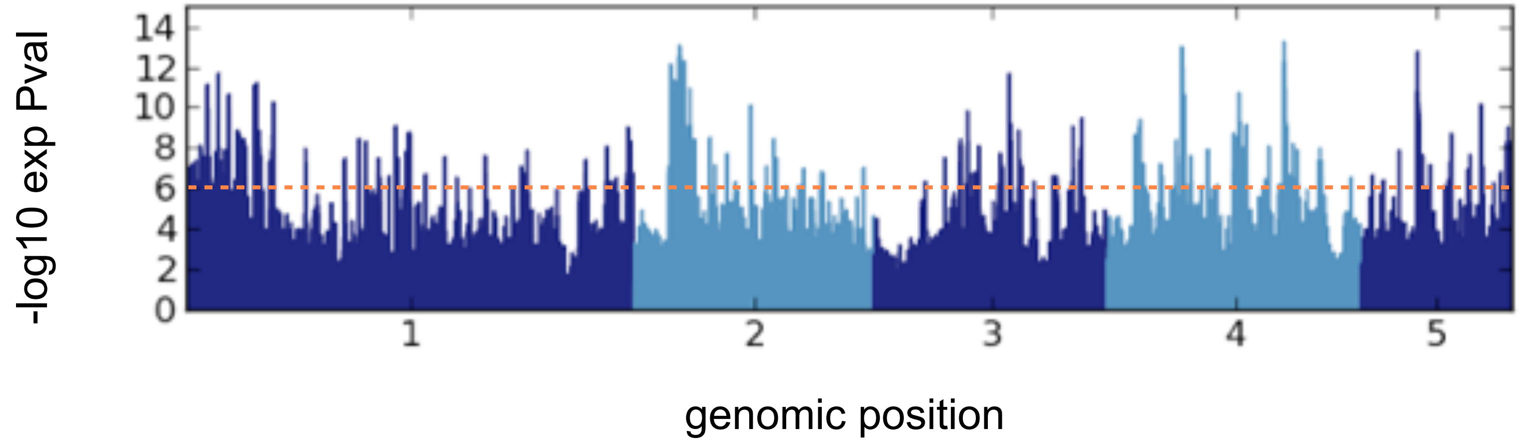
LINEAR MODEL

$$\underbrace{\mathbf{y}}_{\text{pheno}} = \underbrace{\begin{pmatrix}1\\1\\1\\0\\0\\0\end{pmatrix}}_{\substack{\mathbf{x}\\ \text{SNP}}} \beta + \underbrace{\psi}_{\text{noise}}$$

NOISE

$\psi \sim \mathcal{N}\Big(\mathbf{0}, \sigma_e^2$  $\Big)$

N > 1,000



flowering time
A. Thaliana

Flowering in A. thaliana

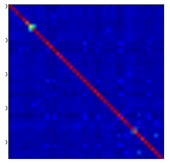
Hidden structure: population structure

LINEAR MIXED MODEL

$$\underset{\text{pheno}}{\mathbf{y}} = \underset{\text{SNP}}{\underset{\mathbf{x}}{\begin{pmatrix}1\\1\\1\\0\\0\\0\end{pmatrix}}}\beta + \underset{\text{genetic term}}{\mathbf{g}} + \underset{\text{noise}}{\psi}$$

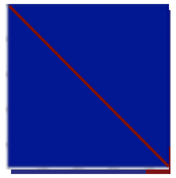
GW GENETIC TERM

$$\mathbf{g} \sim \mathcal{N}\left(\mathbf{0}, \sigma_g^2 \quad \right)$$

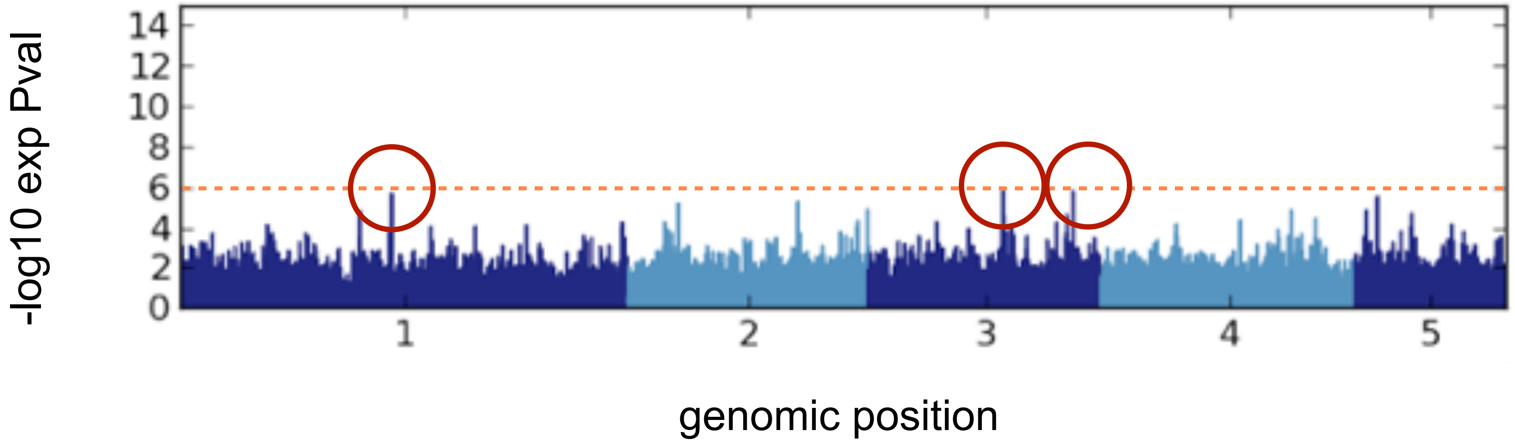


NOISE

$$\psi \sim \mathcal{N}\left(\mathbf{0}, \sigma_e^2 \quad \right)$$



N ~ 5



flowering time
A. Thaliana

Flowering in A. thaliana

Applications of LMMs in genetics

$$\mathbf{y} \sim \mathcal{N}(\beta \mathbf{x}_i, \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I}) \qquad \mathbf{K} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$$

Association testing

$$\mathrm{LLR} = 2 \log \frac{\mathcal{N}\left(\mathbf{y} \,\middle|\, \beta \mathbf{s}_i, \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I}\right)}{\mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{0}, \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I}\right)}$$

Heritability estimation

$$h = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

phenotype prediction

$$\hat{y^{\star}} = \mathbf{K}_{\star,\cdot}(\mathbf{K}_{\cdot,\cdot} + \delta \mathbf{I})^{-1} \mathbf{y}$$

Applications of LMMs in genetics

$$\mathbf{y} \sim \mathcal{N}(\beta \mathbf{x}_i, \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I}) \qquad \mathbf{K} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$$

Association testing

$$\mathrm{LLR} = 2 \log \frac{\mathcal{N}\left(\mathbf{y} \,\middle|\, \beta \mathbf{s}_i, \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I}\right)}{\mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{0}, \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I}\right)}$$

Heritability estimation

$$h = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

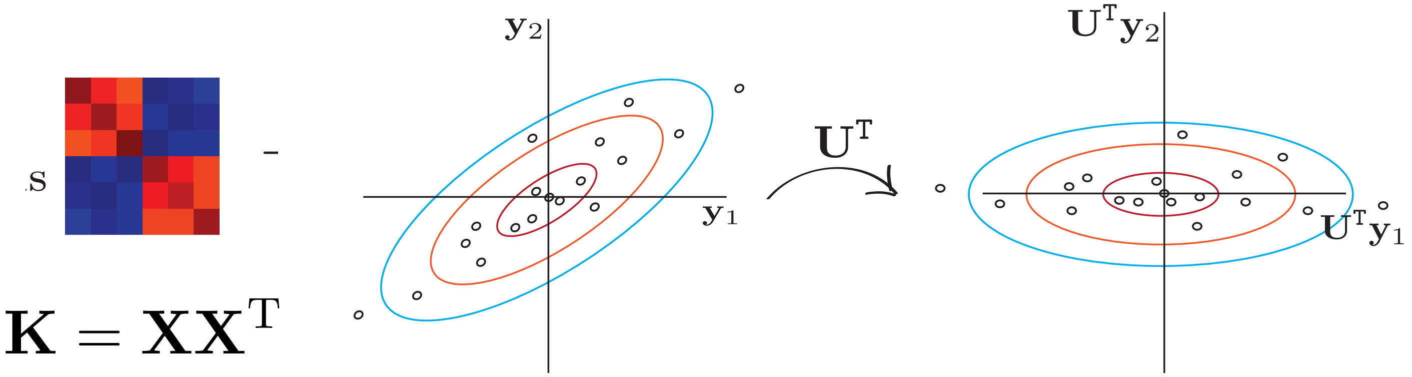
phenotype prediction

$$\hat{y^\star} = \mathbf{K}_{\star,\cdot}(\mathbf{K}_{\cdot,\cdot} + \delta \mathbf{I})^{-1} \mathbf{y}$$

- Efficient inference methods to scale analysis to large cohorts

Lippert et al. *Nature Methods* 8.10 (2011): 833-835.

Zhou & Stephens. *Nature genetics* 44.7 (2012): 821-824.

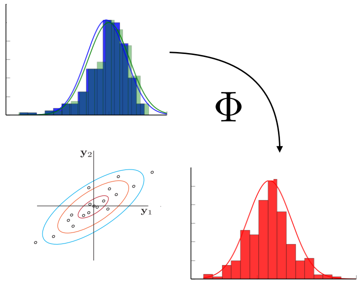

S
$\mathbf{y}_2$
$\mathbf{y}_1$
$\mathbf{U}^{\mathrm{T}}$
$\mathbf{U}^{\mathrm{T}}\mathbf{y}_2$
$\mathbf{U}^{\mathrm{T}}\mathbf{y}_1$

$$\mathbf{K} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$$

Extending linear mixed models

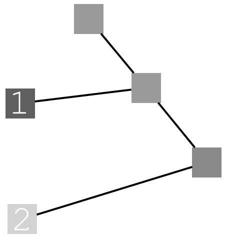
- Statistical challenges in high-dimensional association genetics
    - Normalization and scaling of quantitative traits

      Fusi et al., Nat Comm (2014)



    - Accounting for epistasis and non-linear genetic interactions

      Stephan et al., Nat Comm (2015)

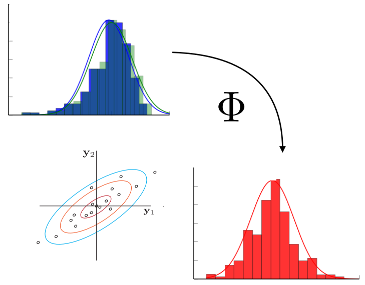


Extending linear mixed models

- Statistical challenges in high-dimensional association genetics

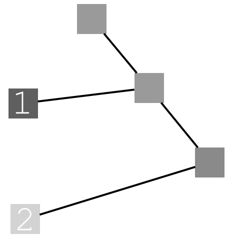
  - Normalization and scaling of quantitative traits

    Fusi et al., Nat Comm (2014)

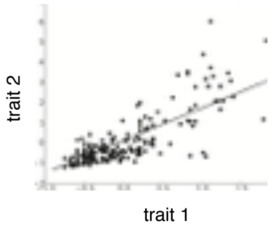


  - Accounting for epistasis and non-linear genetic interactions

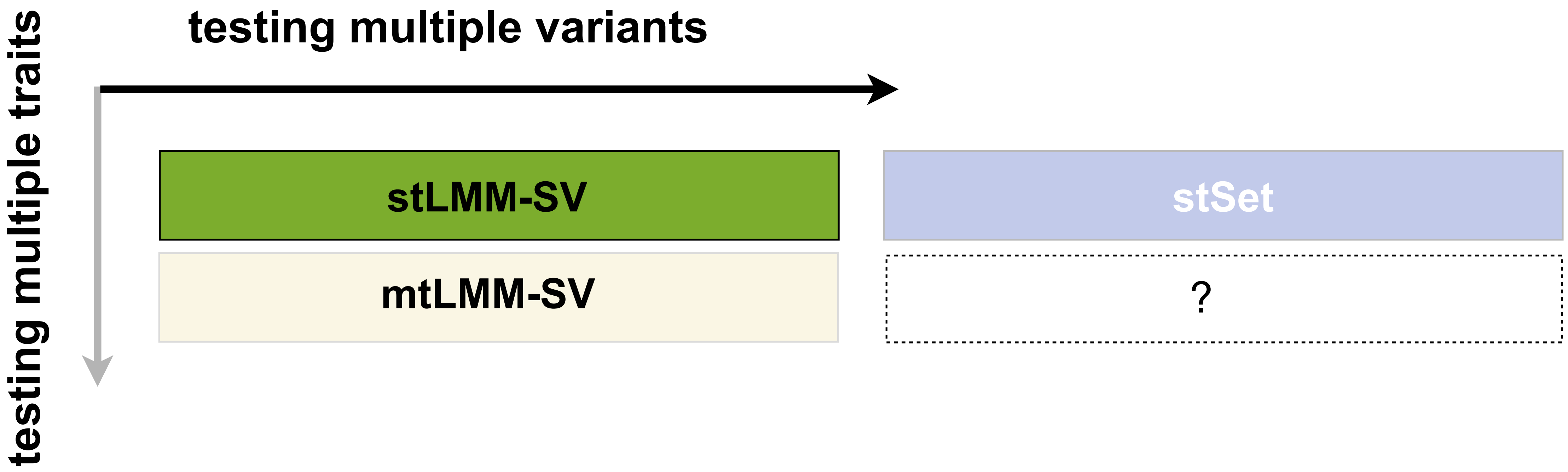
    Stephan et al., Nat Comm (2015)



  - **Joint modeling of multiple (correlated) traits**



Joint modelling of traits and variants



testing multiple variants

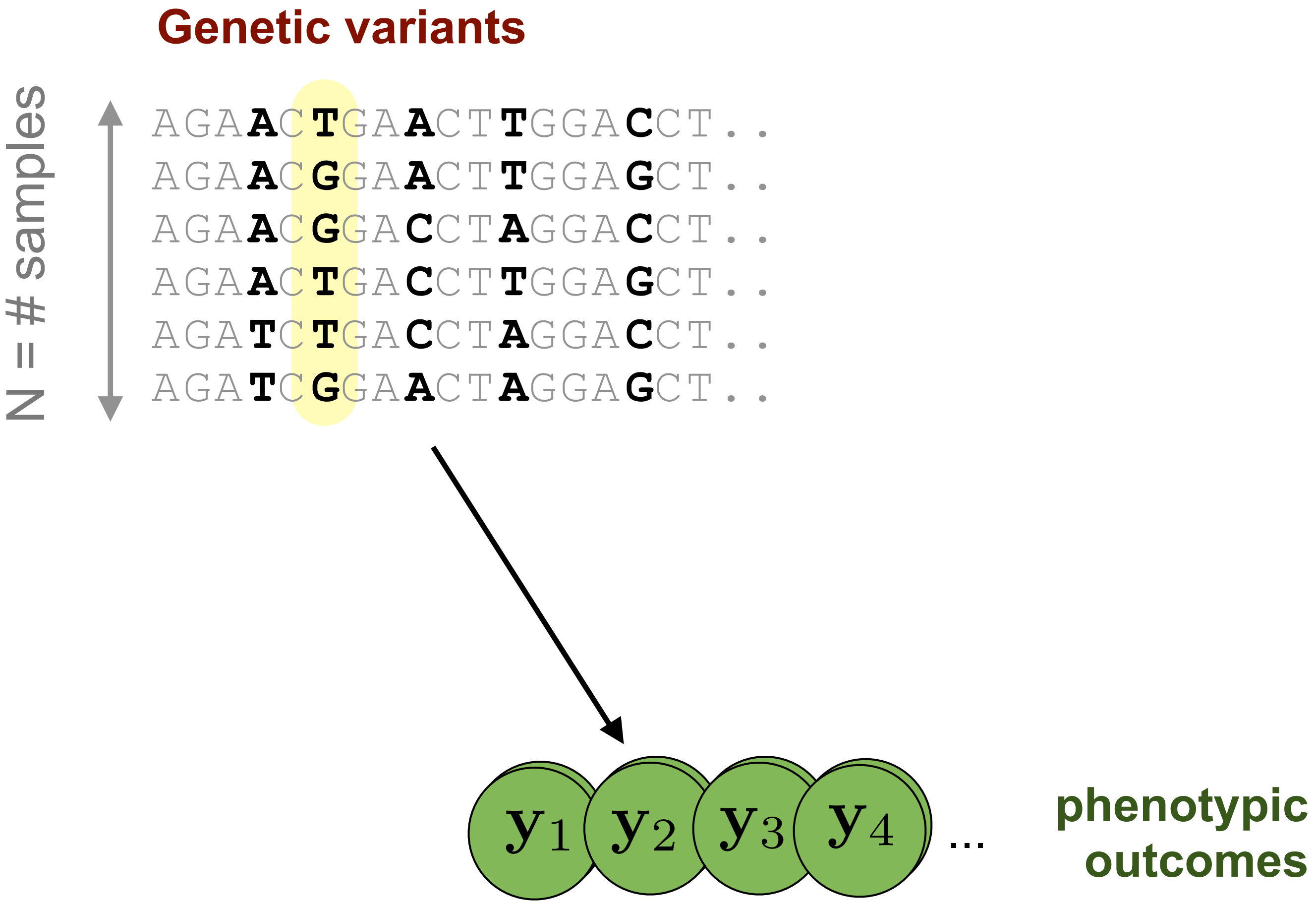
testing multiple traits

stLMM-SV

stSet

mtLMM-SV

?



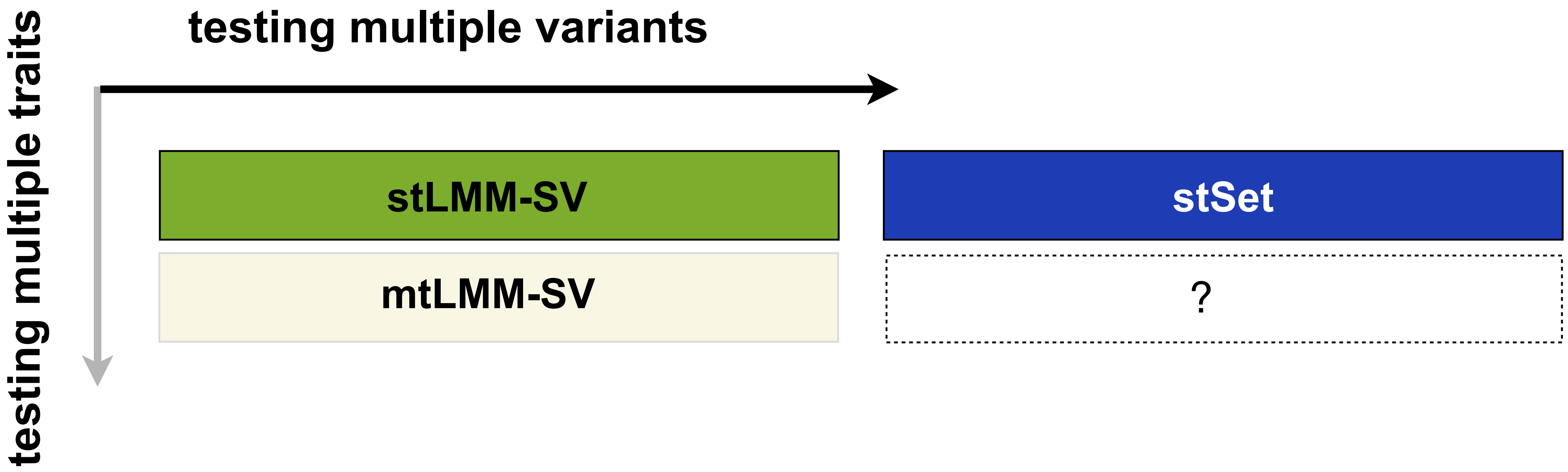
Genetic variants

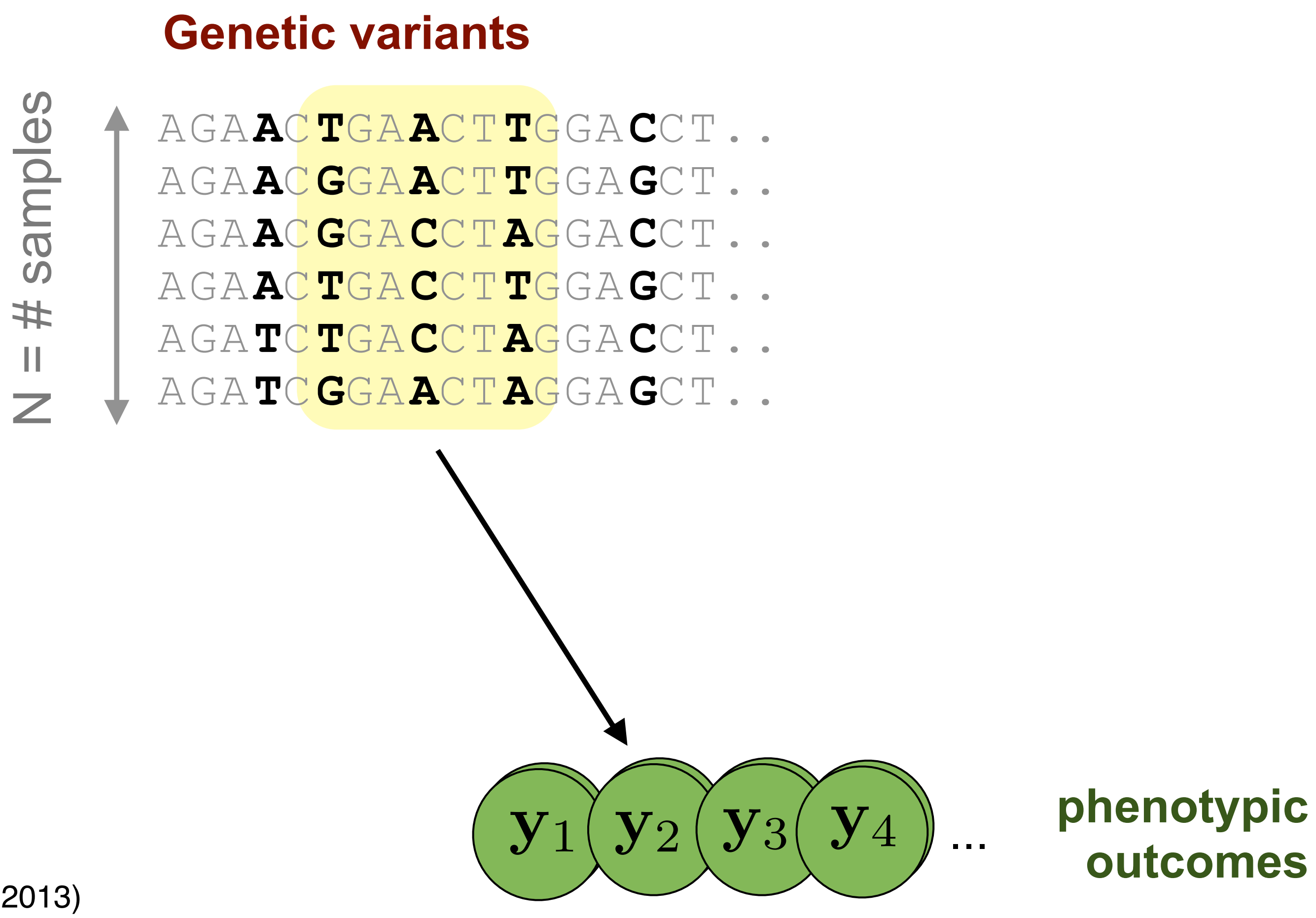
N = # samples

```
AGAACTGAACTTGGACCT..
AGAACGGAACTTGGAGCT..
AGAACGGACCTAGGACCT..
AGAACTGACCTTGGAGCT..
AGATCTGACCTAGGACCT..
AGATCGGAACTAGGAGCT..
```

phenotypic outcomes

Joint modelling of traits and variants

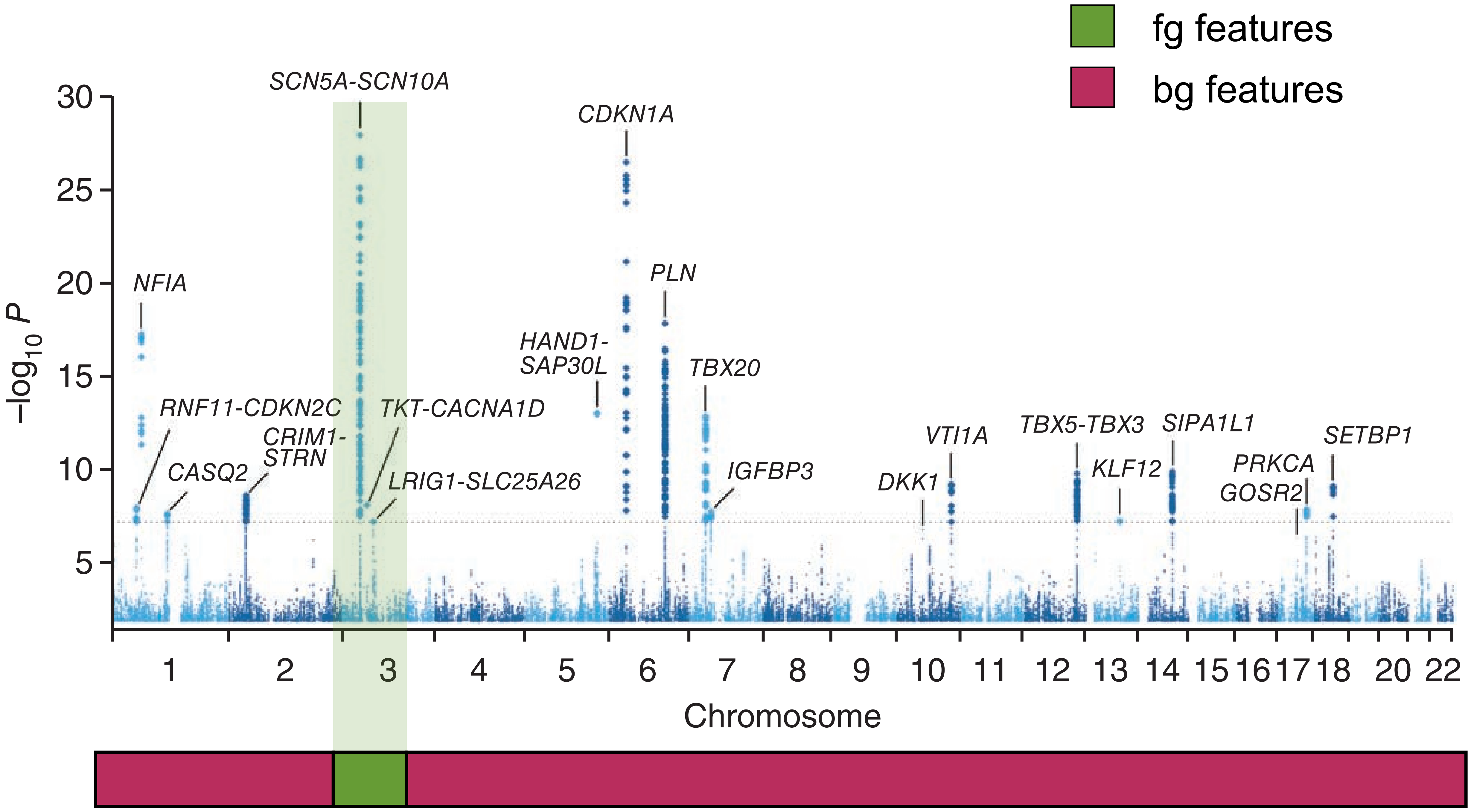




```
AGAACTGAACTTGGACCT..
AGAACGGAACTTGGAGCT..
AGAACGGACCTAGGACCT..
AGAACTGACCTTGGAGCT..
AGATCTGACCTAGGACCT..
AGATCGGAACTAGGAGCT..
```

Listgarten et al. Bioinformatics (2013)

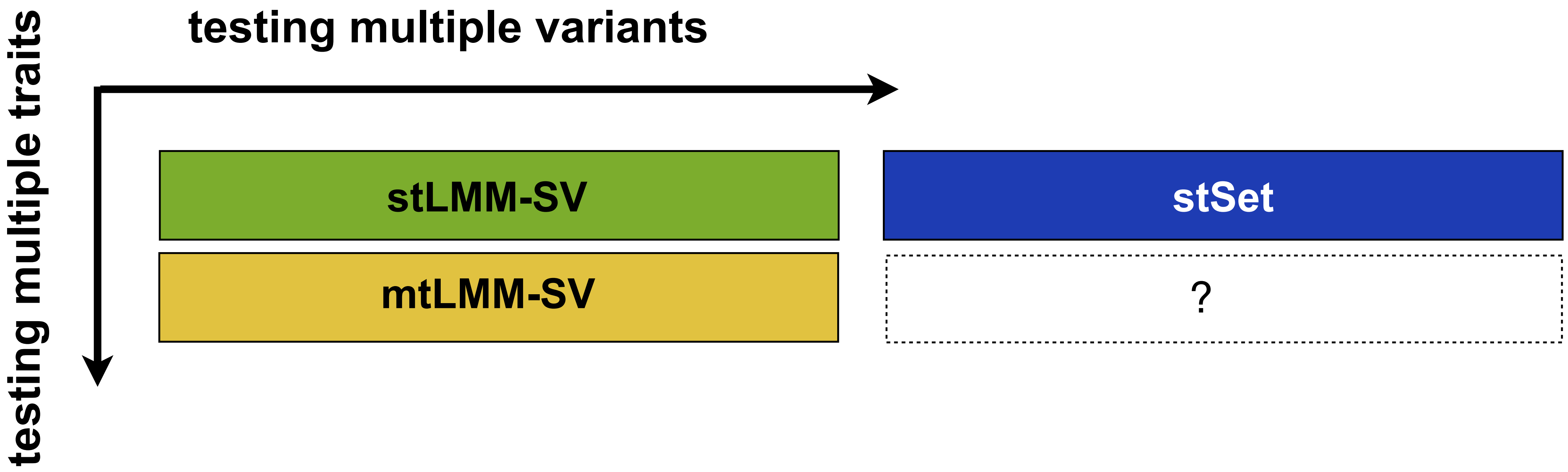
Region-based testing



- rare variant associations
- accounting for allelic heterogeneity

Sotoodehnia et al, Nature Genetics (2010)

Joint modelling of traits and variants



testing multiple variants

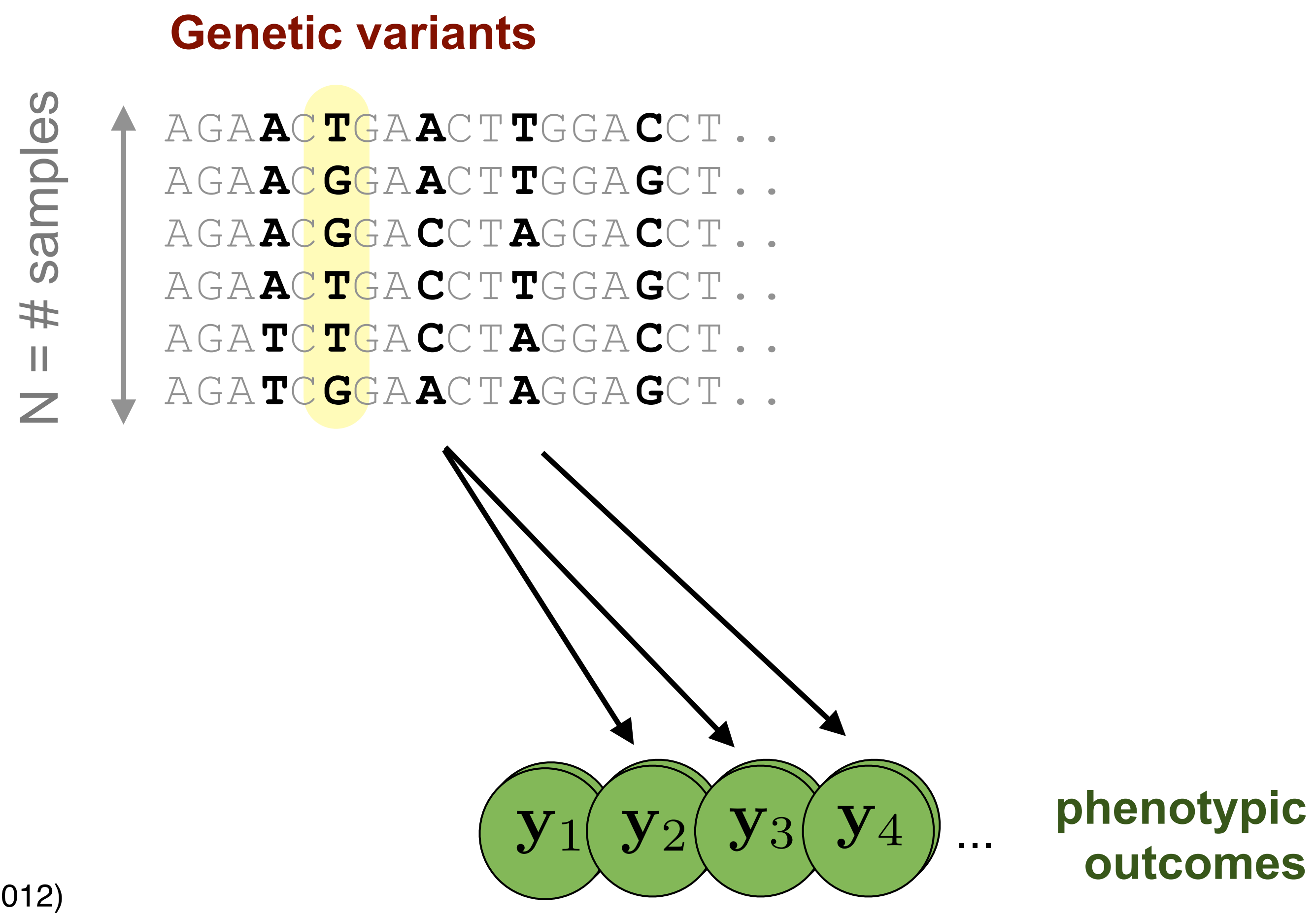
testing multiple traits

stLMM-SV

stSet

mtLMM-SV

?



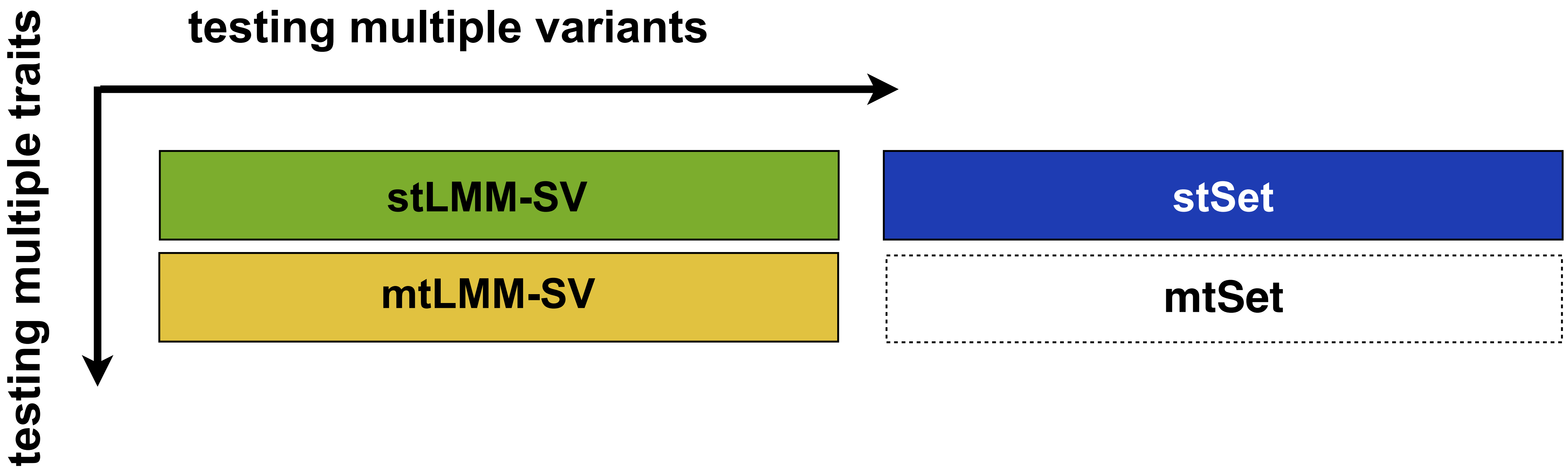
Genetic variants

N = # samples

phenotypic outcomes

Korte et al., Nature Genetics (2012)

Joint modelling of traits and variants



testing multiple variants

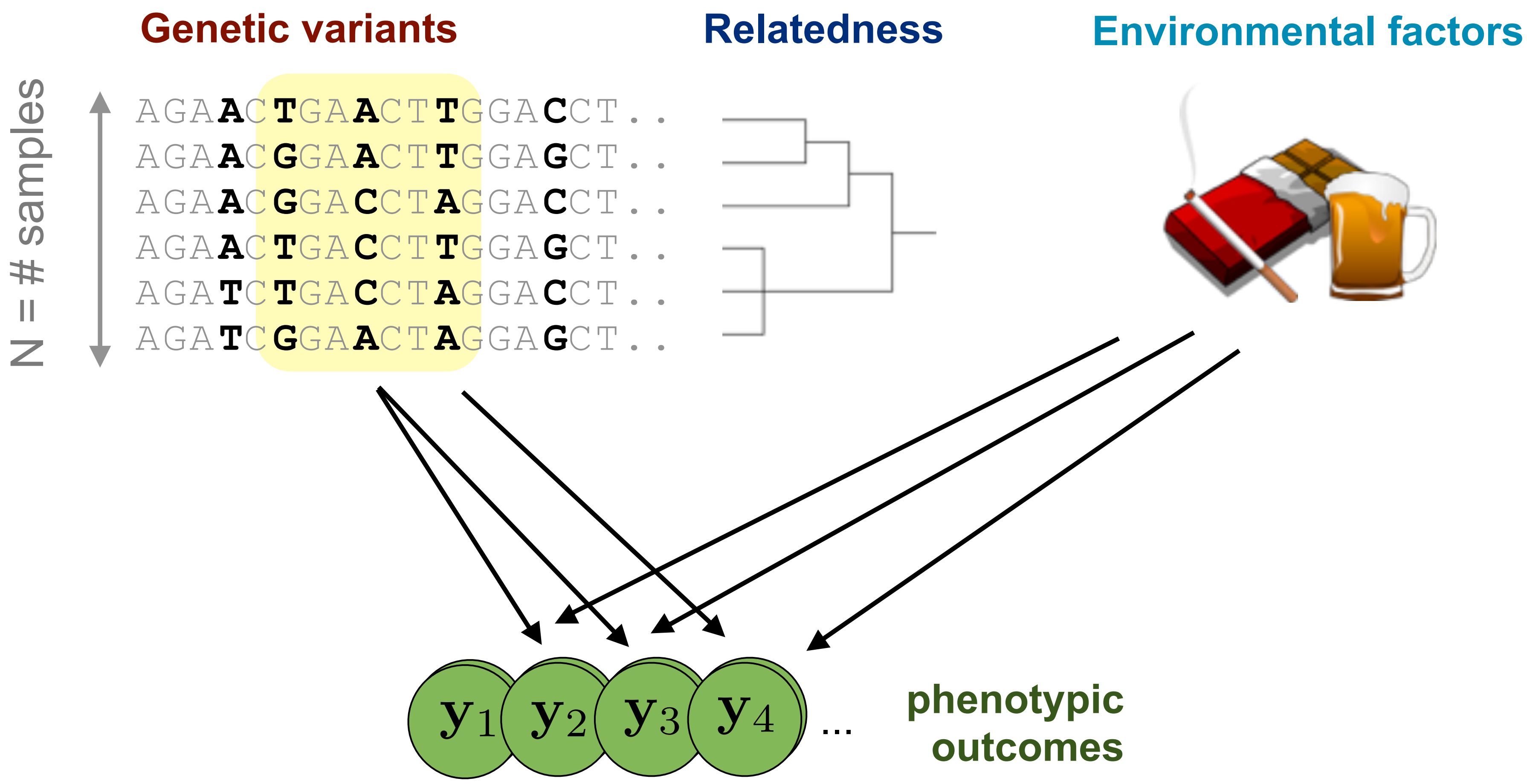
testing multiple traits

stLMM-SV

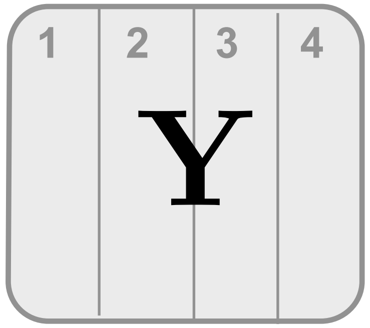
mtLMM-SV

stSet

mtSet



mtSet: aggregation across traits and causal variants

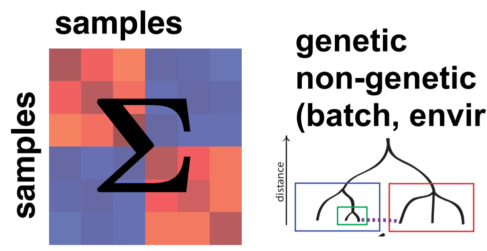


$$\mathbf{Y} = \mathbf{FW} + \mathbf{R} + \mathbf{U} + \mathbf{\Psi}$$

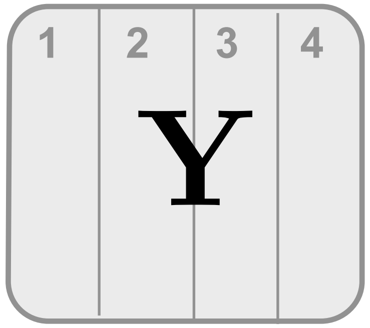
phenotypes covariates SNPs relatedness noise

mtSet: aggregation across traits and causal variants

sample covariance



genetic
non-genetic
(batch, environment)

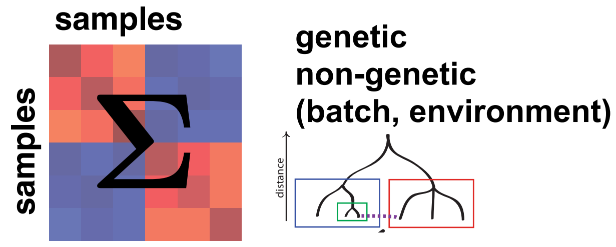


$$\mathbf{Y} = \mathbf{FW} + \mathbf{R} + \mathbf{U} + \mathbf{\Psi}$$

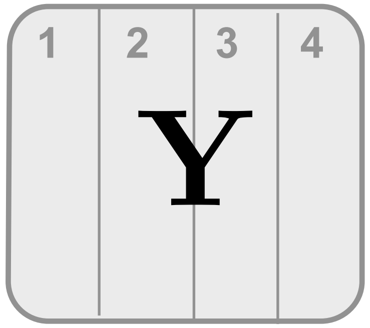
phenotypes covariates SNPs relatedness noise

mtSet: aggregation across traits and causal variants

sample covariance



genetic
non-genetic
(batch, environment)



$$\mathbf{Y} = \mathbf{FW} + \mathbf{R} + \mathbf{U} + \boldsymbol{\Psi}$$

phenotypes covariates SNPs relatedness noise

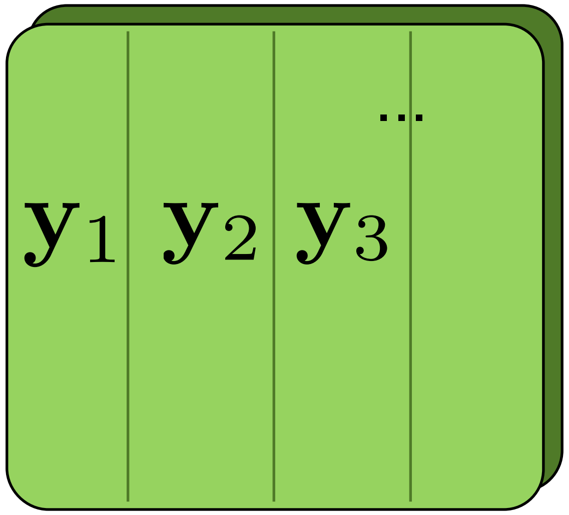
variance components
(random effects)

mtSet: aggregation across traits and causal variants

genetic variants

AGA**A**C**T**GAACT**T**GGA**C**CT..
AGA**A**C**G**GAACT**T**GGA**G**CT..
AGA**A**C**G**GAACT**A**GGA**C**CT..
AGA**A**C**T**GAACT**T**GGA**G**CT..
AGA**T**C**G**GAACT**A**GGA**C**CT..
AGA**T**C**G**GAACT**A**GGA**G**CT..

phenotypes



$$\begin{aligned}\mathbf{X} &= [\mathbf{x}_{:,1}, \ldots, \mathbf{x}_{:,F}] \\ &= [\mathbf{x}_{1,:}, \ldots, \mathbf{x}_{N,:}]^\top\end{aligned}$$

$$\begin{aligned}\mathbf{Y} &= [\mathbf{y}_{:,1}, \ldots, \mathbf{y}_{:,T}] \\ &= [\mathbf{y}_{1,:}, \ldots, \mathbf{y}_{N,:}]^\top\end{aligned}$$

$N =$ # samples

$T =$ # traits

$F =$ # snps

mtSet: aggregation across traits and causal variants

Linear model for trait *t*

$$\mathbf{y}_{:,t} = \sum_k \mathbf{g}_{:,k} w_{k,t} + \sum_f \mathbf{x}_{:,f} v_{f,t} + \boldsymbol{\psi}_{:,t}$$

Introducing MVN priors on weights and residuals and marginalizing out

mtSet: aggregation across traits and causal variants

Linear model for trait *t*

$$\mathbf{y}_{:,t} = \sum_k \mathbf{g}_{:,k} w_{k,t} + \sum_f \mathbf{x}_{:,f} v_{f,t} + \boldsymbol{\psi}_{:,t}$$

Introducing MVN priors on weights and residuals and marginalizing out

$$p(\mathbf{W}^T) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}_{:,k} \mid \mathbf{0}, \mathbf{C}_r\right)$$

$$p(\mathbf{V}^T) = \prod_f \mathcal{N}\left(\mathbf{v}_{f,:} \mid \mathbf{0}, \mathbf{C}_g\right)$$

$$p(\boldsymbol{\Psi}^T) = \prod_n \mathcal{N}\left(\boldsymbol{\psi}_{n,:} \mid \mathbf{0}, \boldsymbol{\Sigma}\right)$$

mtSet: aggregation across traits and causal variants

Linear model for trait *t*

$$\mathbf{y}_{:,t} = \sum_{k} \mathbf{g}_{:,k} w_{k,t} + \sum_{f} \mathbf{x}_{:,f} v_{f,t} + \boldsymbol{\psi}_{:,t}$$

Introducing MVN priors on weights and residuals and marginalizing out

$$p(\mathbf{W}^T) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}_{:,k} \mid \mathbf{0}, \mathbf{C}_r\right) \qquad p(\mathbf{V}^T) = \prod\nolimits_{f} \mathcal{N}\left(\mathbf{v}_{f,:} \mid \mathbf{0}, \mathbf{C}_g\right)$$

$$p(\boldsymbol{\Psi}^T) = \prod\nolimits_{n} \mathcal{N}\left(\boldsymbol{\psi}_{n,:} \mid \mathbf{0}, \boldsymbol{\Sigma}\right)$$

Marginal likelihood

$$p(\mathbf{Y} \mid \mathbf{C}_r, \mathbf{R}_r, \mathbf{C}_g, \mathbf{R}_g, \boldsymbol{\Sigma}) = \mathcal{N}\left(\mathrm{vec}\left(\mathbf{Y}\right) \,\middle|\, \mathbf{0}, \boxed{\underbrace{\mathbf{C}_r \otimes \overset{\mathbf{GG}^T}{\overset{\downarrow}{\mathbf{R}_r}}}_{\text{fg signal}} + \underbrace{\mathbf{C}_g \otimes \overset{\mathbf{XX}^T}{\overset{\downarrow}{\mathbf{R}_g}}}_{\text{bg signal}} + \underbrace{\boldsymbol{\Sigma} \otimes \mathbf{I}}_{\text{struct. noise}}}\right)$$

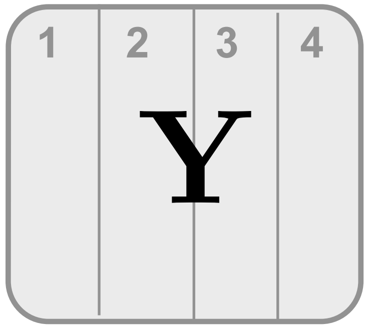
$$\mathbf{R} \qquad\qquad \mathbf{U} \qquad\qquad \boldsymbol{\Psi}$$

Closely related to multi-task kernel models in ML
Rakitsch et al., NIPS 2013
Bonilla et al., NIPS 2008

mtSet: aggregation across traits and causal variants

upfront computation

$$O(\boxed{N^3 + N^2R} + NR^2P^2 + NRP^4)$$



$$= \mathbf{FW} + \mathbf{R} + \mathbf{U} + \mathbf{\Psi}$$

phenotype covariates SNPs relatedness noise

variance components
(random effects)

$$O(N^3P^3)$$

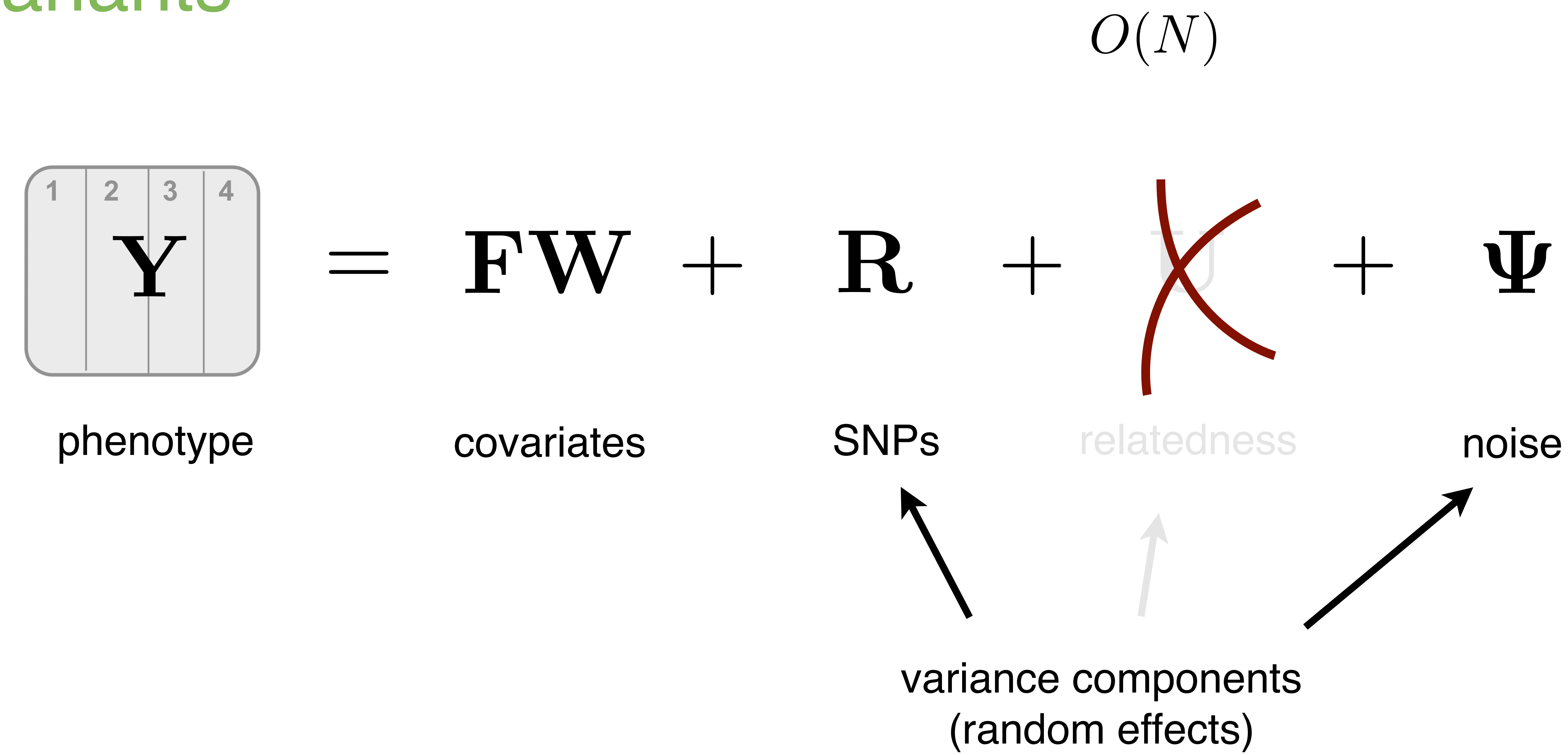
Challenge: Cubical scaling means such an algorithm is impractical for even moderately-size datasets!

# tested SNPs << # samples

mtSet

1.Casale, P. & Rakitsch, B. et a., Nature Methods (2015)

mtSet: aggregation across traits and causal variants



$O(N)$

$$\mathbf{Y} = \mathbf{FW} + \mathbf{R} + \cancel{\mathbf{U}} + \mathbf{\Psi}$$

phenotype

covariates

SNPs

relatedness

noise

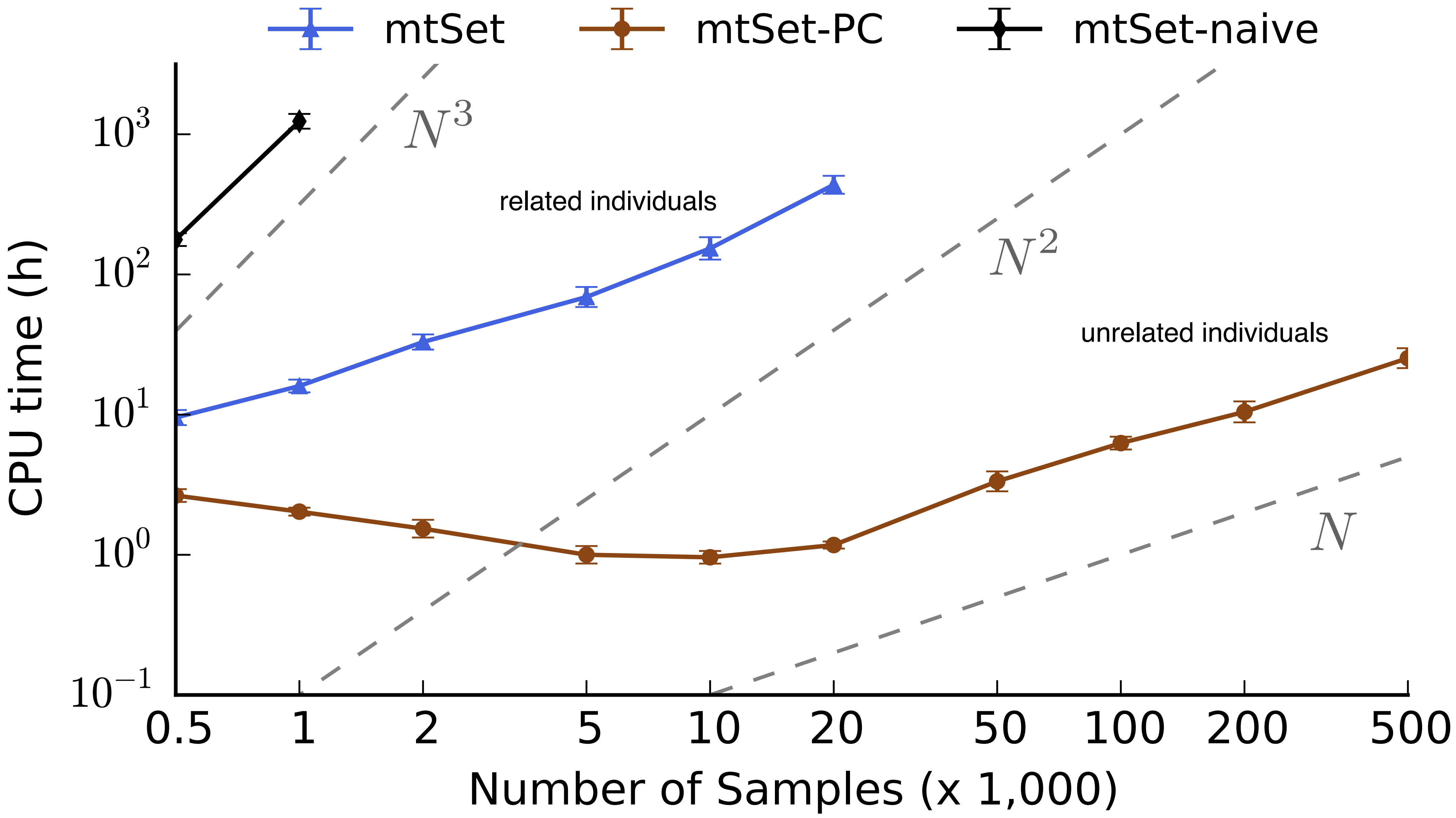
variance components
(random effects)

mtSet-PC

# tested SNPs **<<** # samples

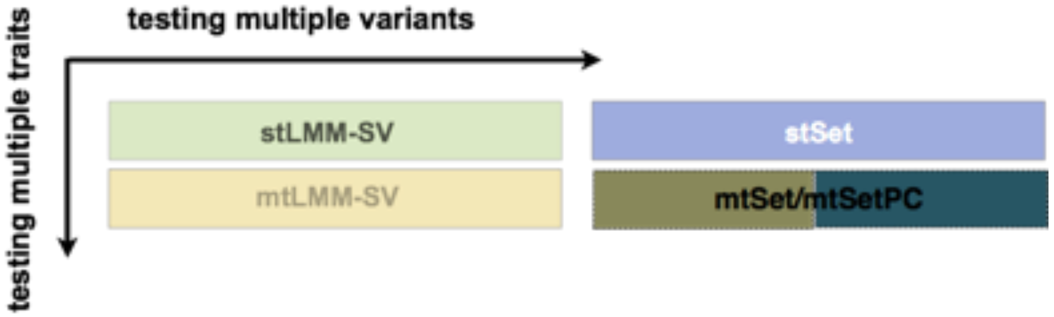
1.Casale, P. & Rakitsch, B. et a., Nature Methods (2015)

Efficient inference for large-scale GWAS



(human chrom20, 3,975 set tests for 4 traits)

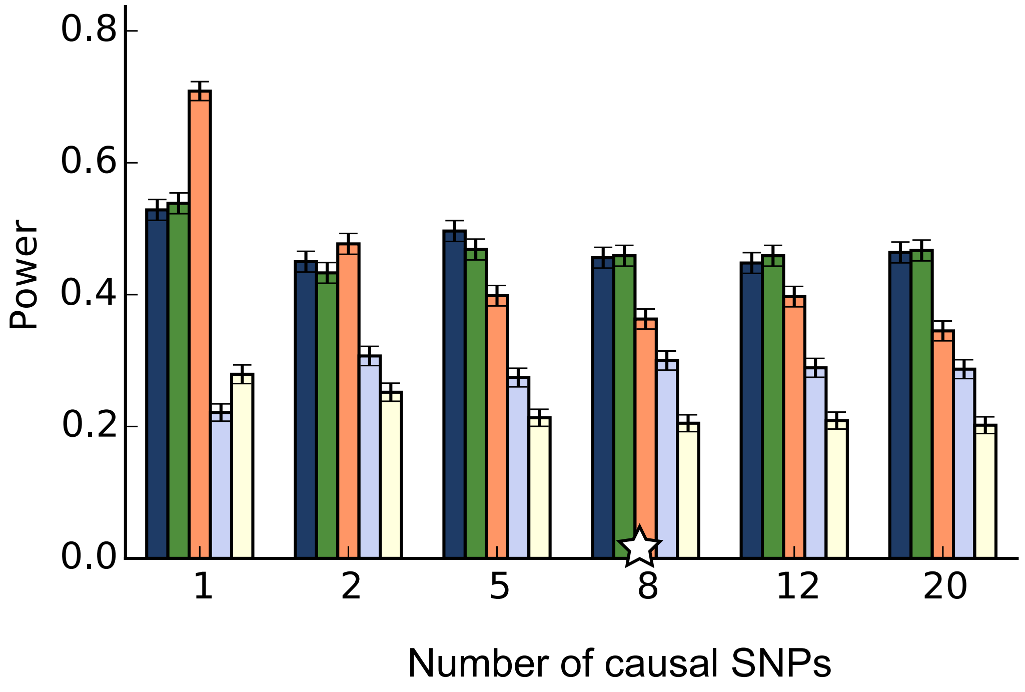
Simulation study: aggregating across multiple causal variants and correlated traits



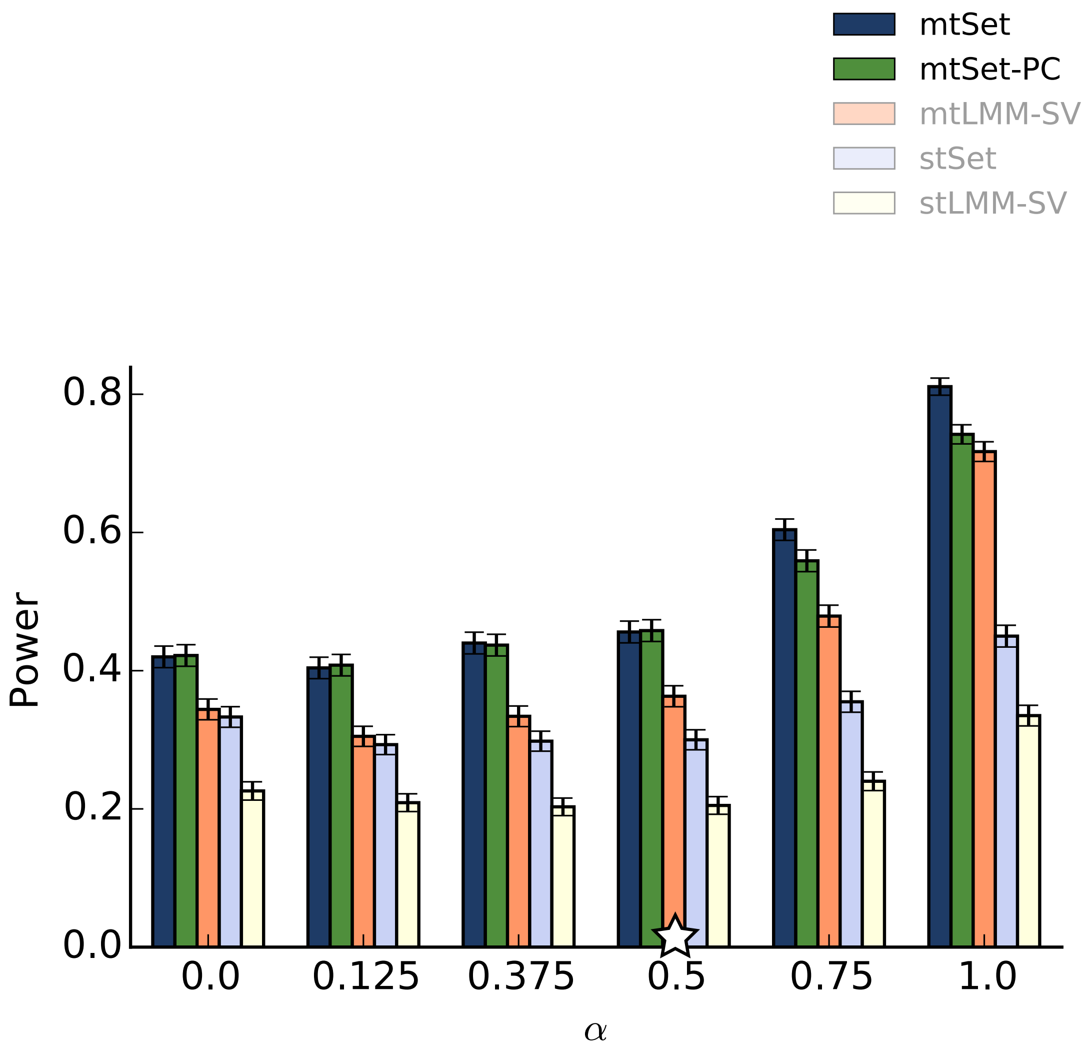
testing multiple variants

testing multiple traits

stLMM-SV	stSet
mtLMM-SV	mtSet/mtSetPC



multiple causal variants

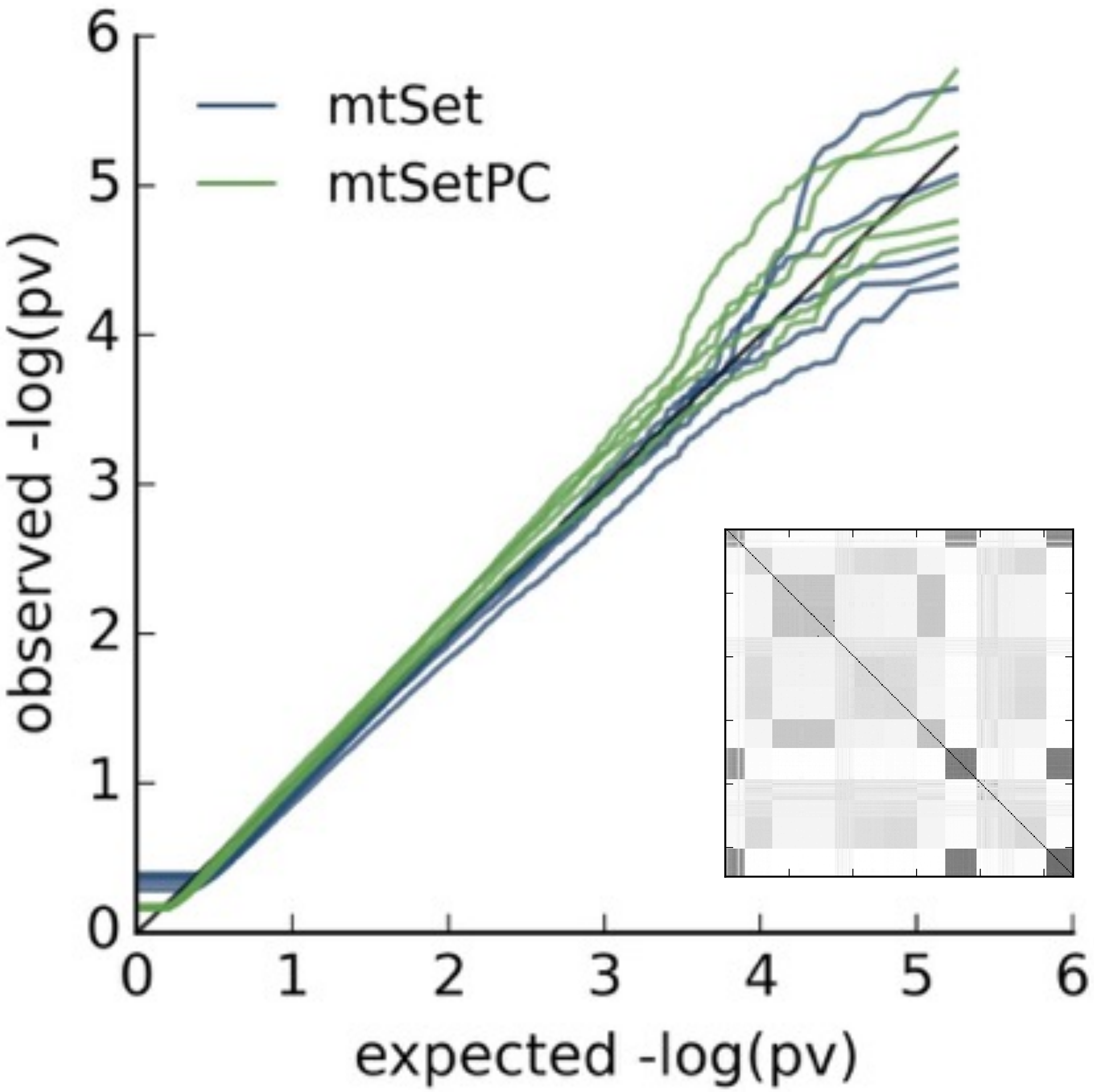


correlation between traits

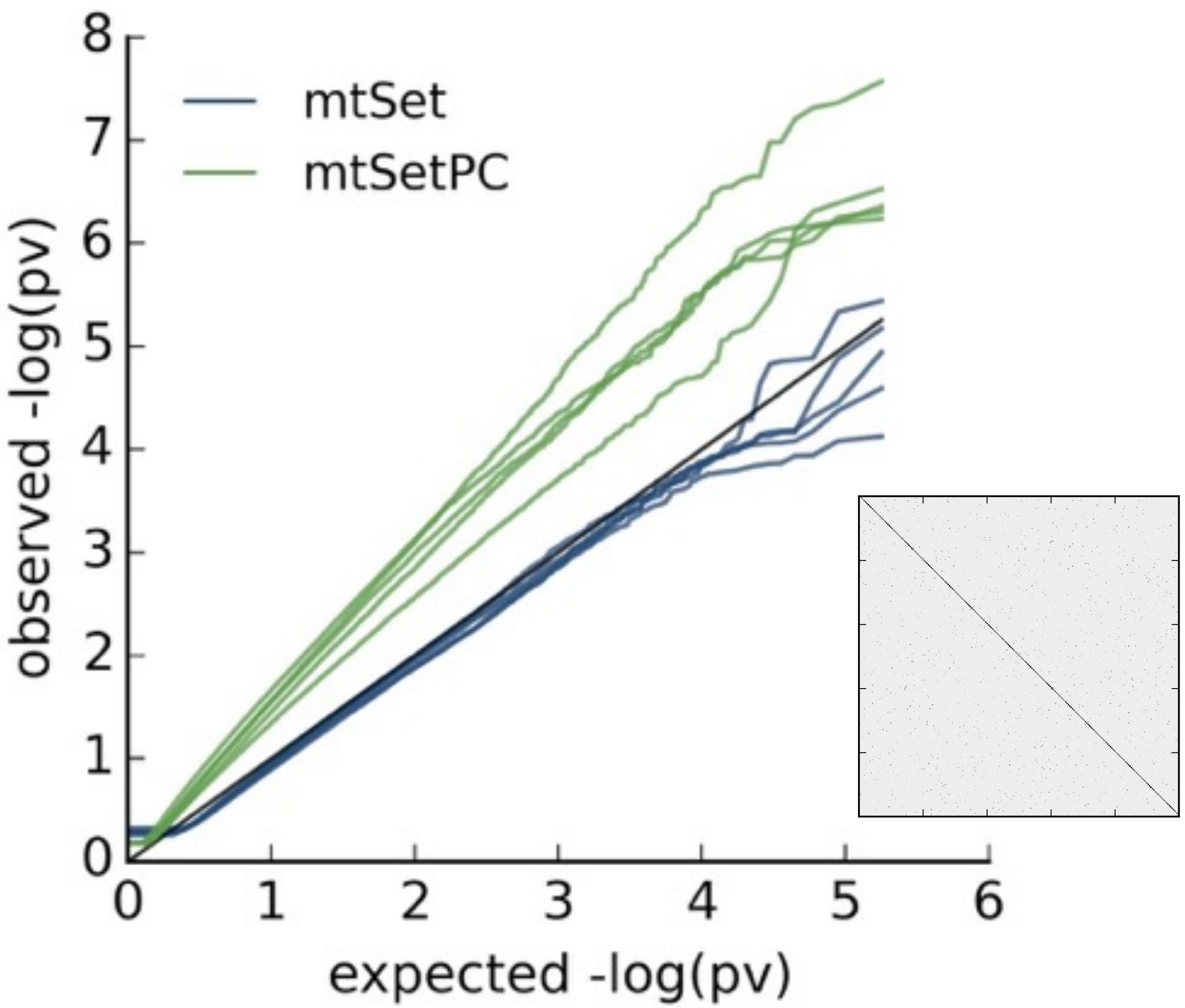
Accounting for relatedness

$$\boxed{\mathbf{Y}} = \underbrace{\mathbf{U}}_{\text{relatedness}} + \underbrace{\mathbf{E}}_{\text{env. fact.}} + \underbrace{\mathbf{\Psi}}_{\text{iid noise}}$$

Population Structure (1000G)

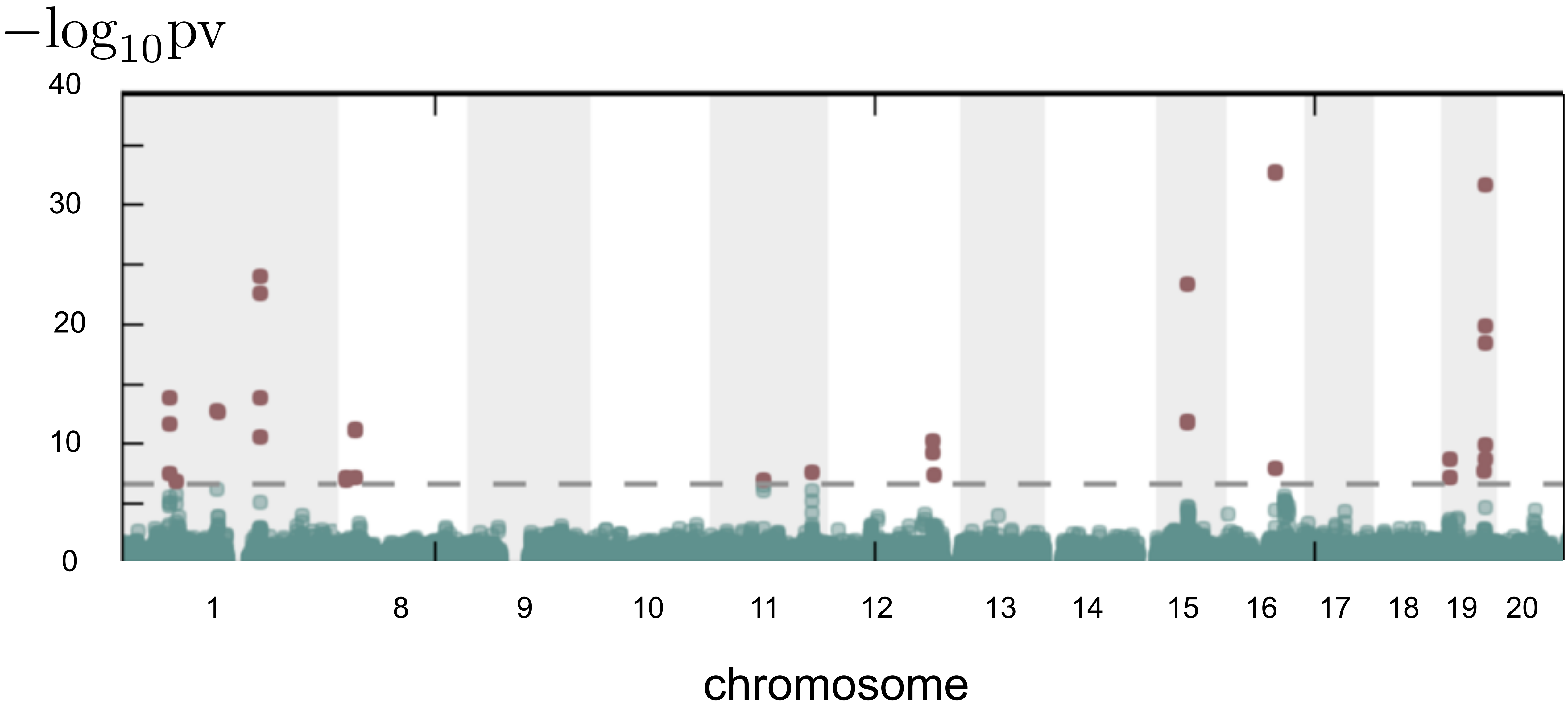


Family Structure (sim from 1000G)



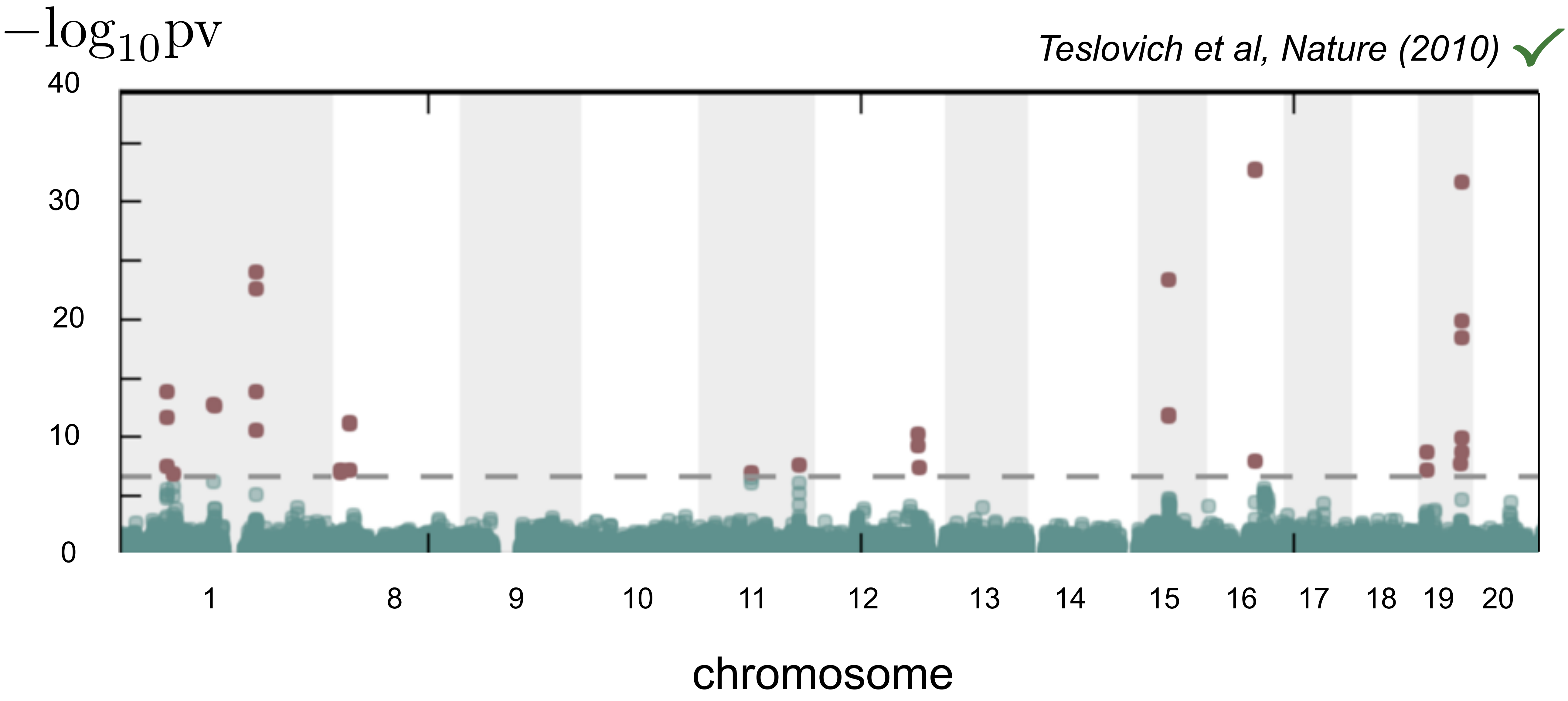
Analysis of lipid-related traits in Human

- $N =$ 5,246
- 4 lipid traits: LDL, HDL, CRP, Trig



Analysis of lipid-related traits in Human

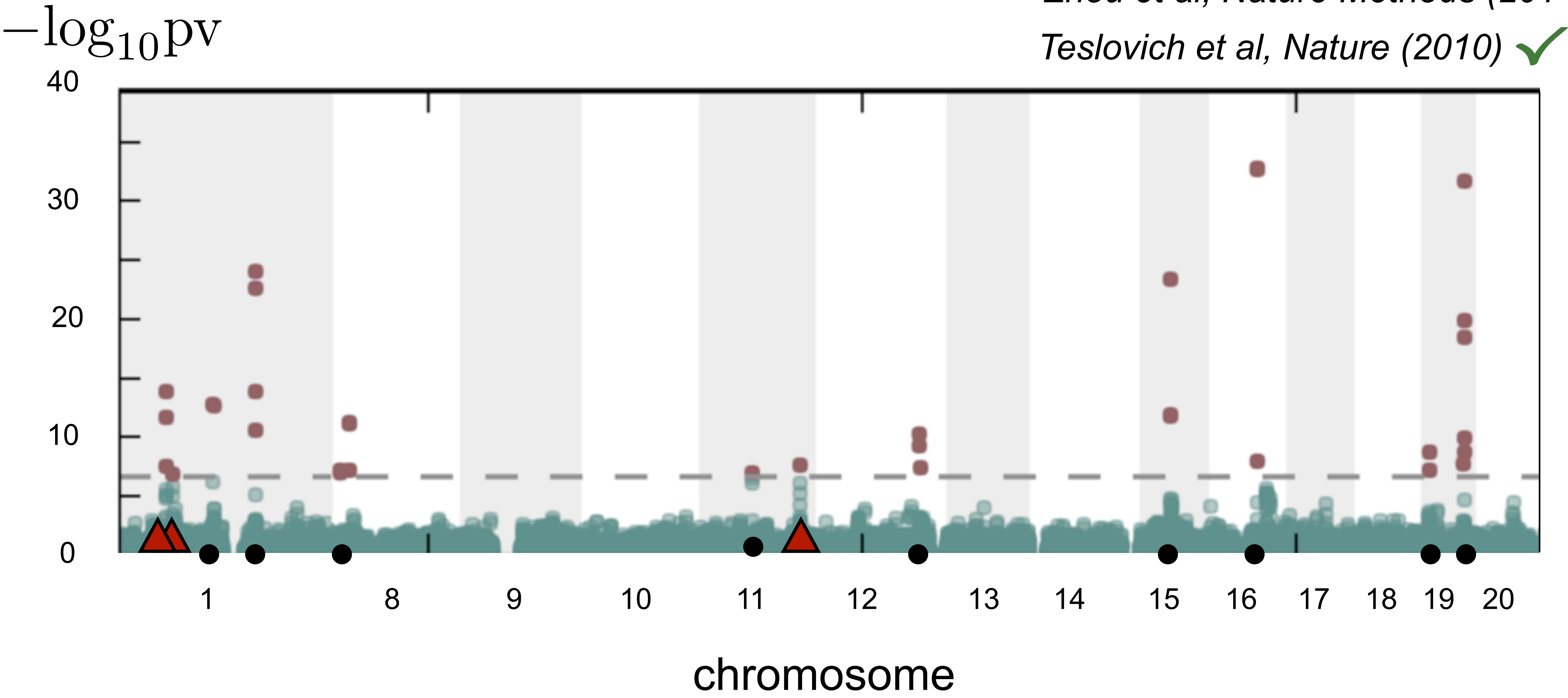
- $N =$ 5,246
- 4 lipid traits: LDL, HDL, CRP, Trig



Teslovich et al, Nature (2010) ✓

Analysis of lipid-related traits in Human

- $N =$ 5,246
- 4 lipid traits: LDL, HDL, CRP, Trig

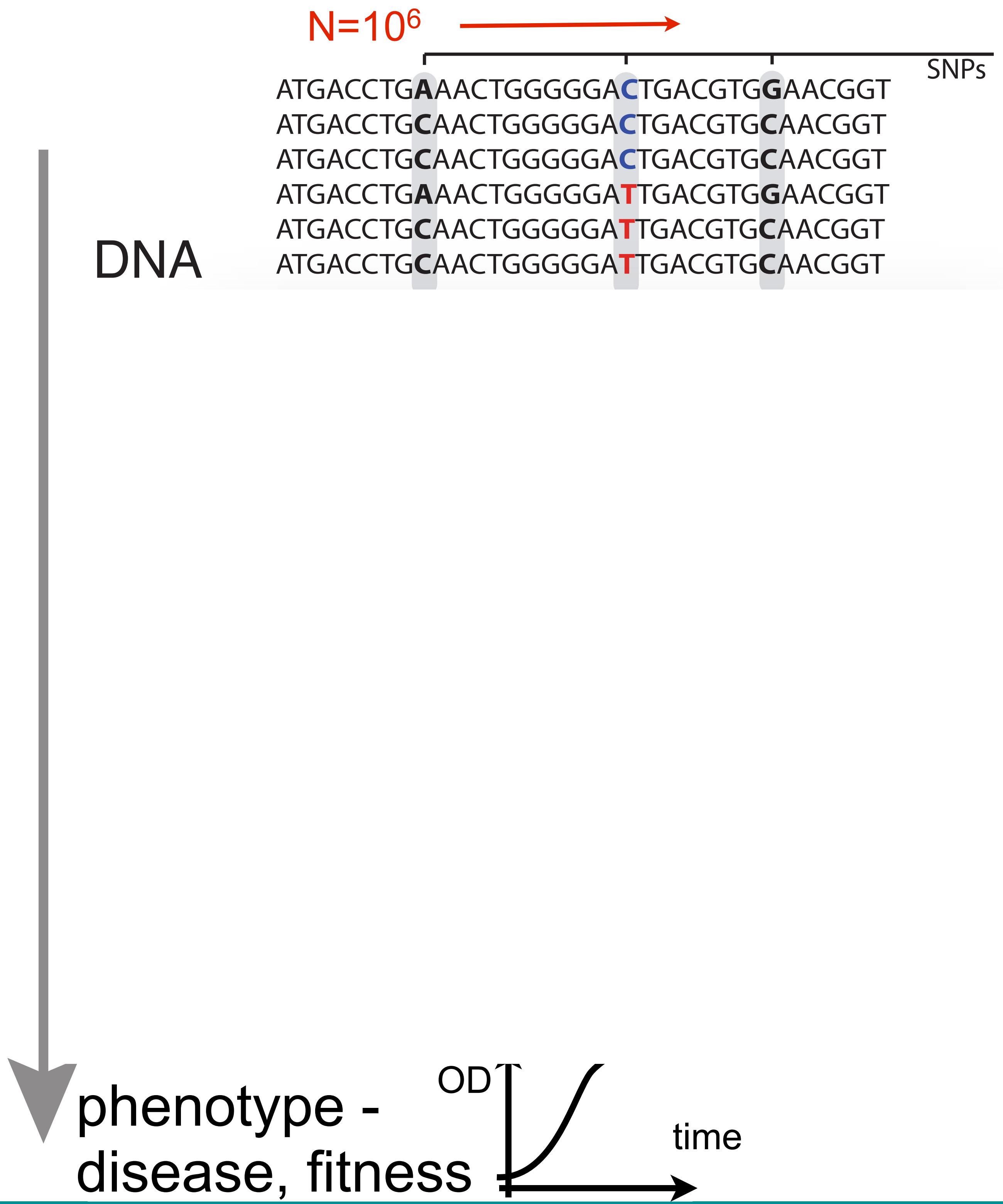


● multi-trait single-SNP model

Zhou et al, Nature Methods (2014)

Teslovich et al, Nature (2010) ✓

Multi-omics association genetics



N=10^6

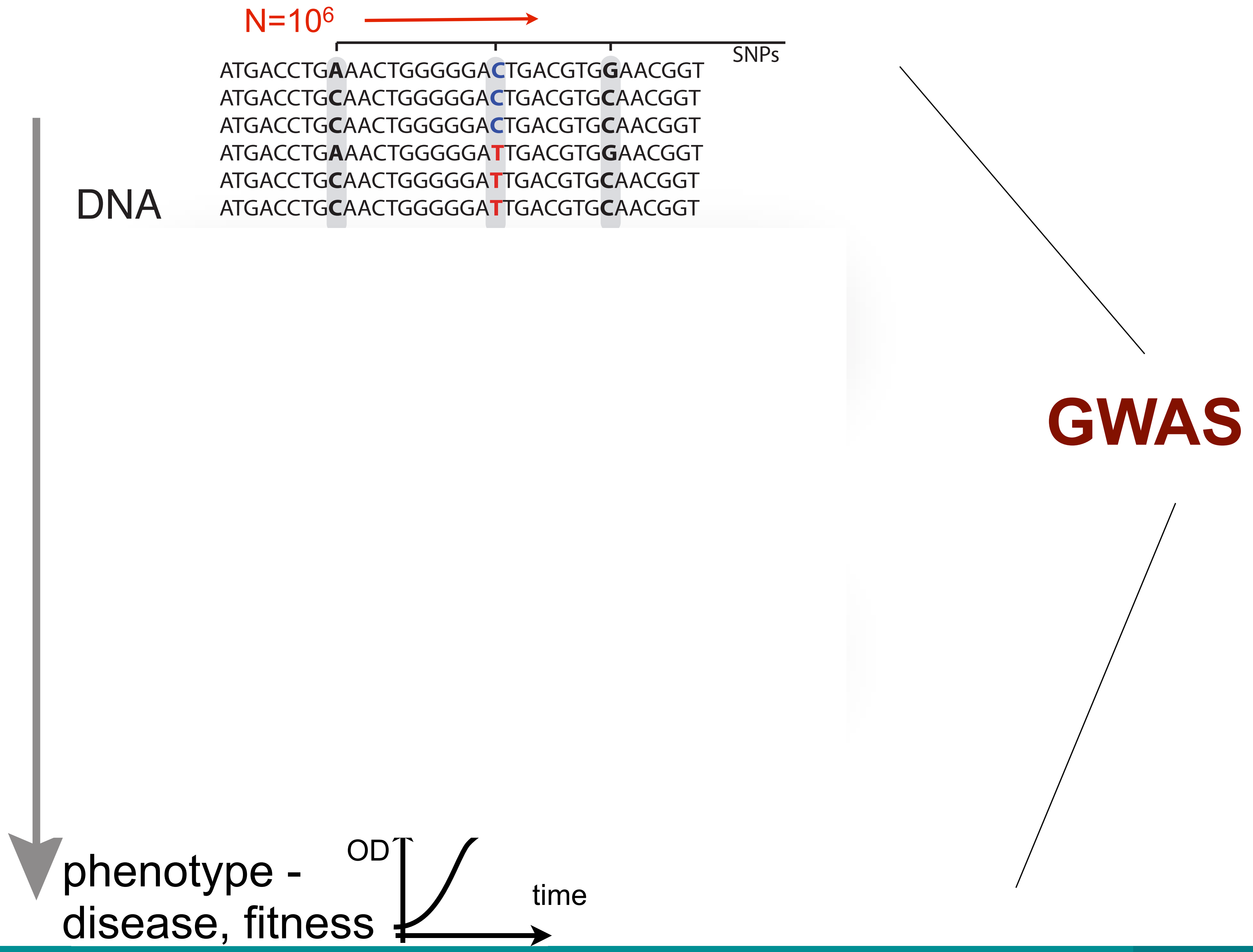
SNPs

```
ATGACCTGAAACTGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT
```

DNA

phenotype -
disease, fitness

Multi-omics association genetics



N=10^6

SNPs

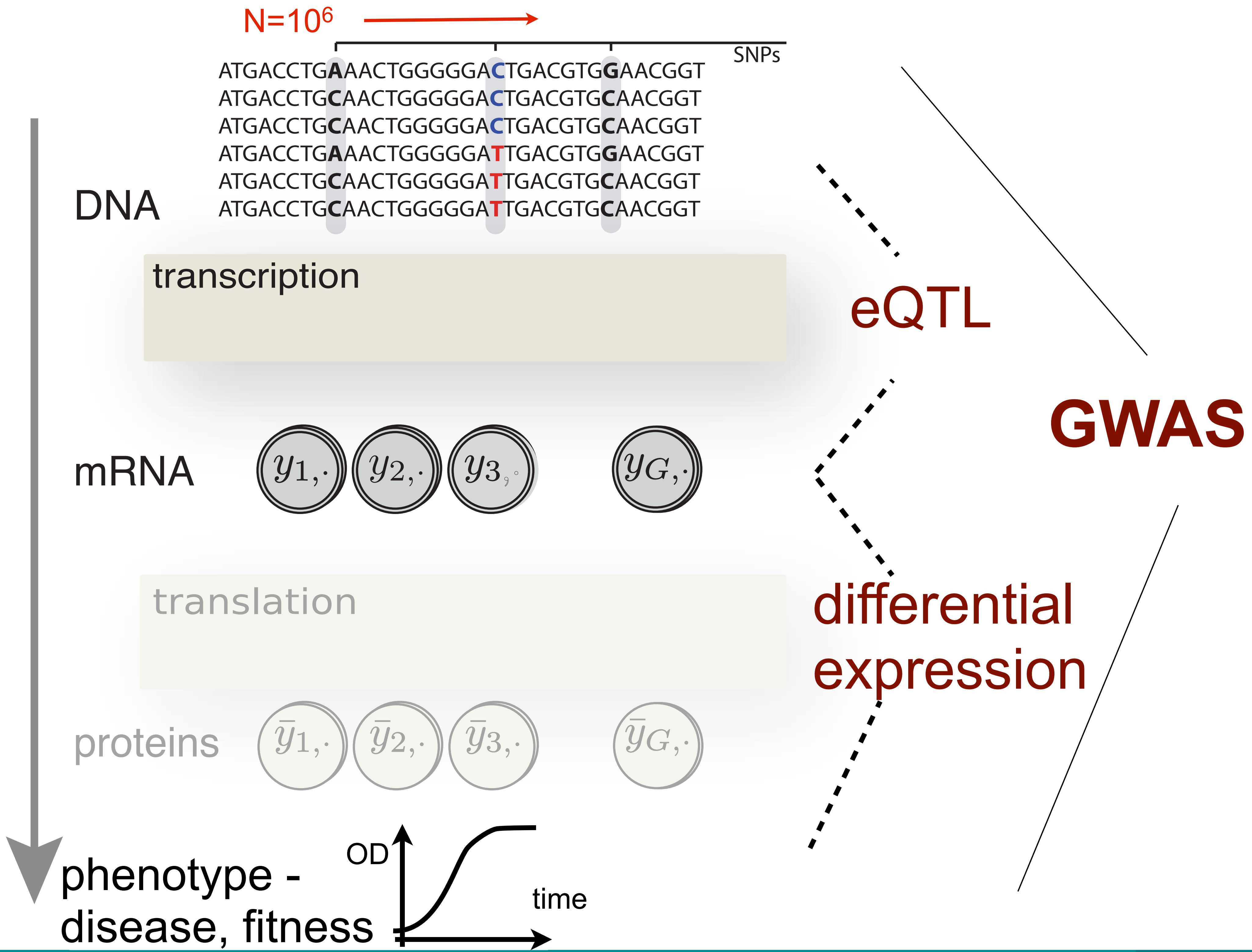
ATGACCTGAAACTGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT

DNA

phenotype -
disease, fitness

GWAS

Multi-omics association genetics



N=10^6

SNPs

ATGACCTGAAACTGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT

DNA

transcription

mRNA $y_{1,\cdot}$ $y_{2,\cdot}$ $y_{3,\cdot}$ $y_{G,\cdot}$

translation

proteins $\bar{y}_{1,\cdot}$ $\bar{y}_{2,\cdot}$ $\bar{y}_{3,\cdot}$ $\bar{y}_{G,\cdot}$

phenotype -
disease, fitness

OD

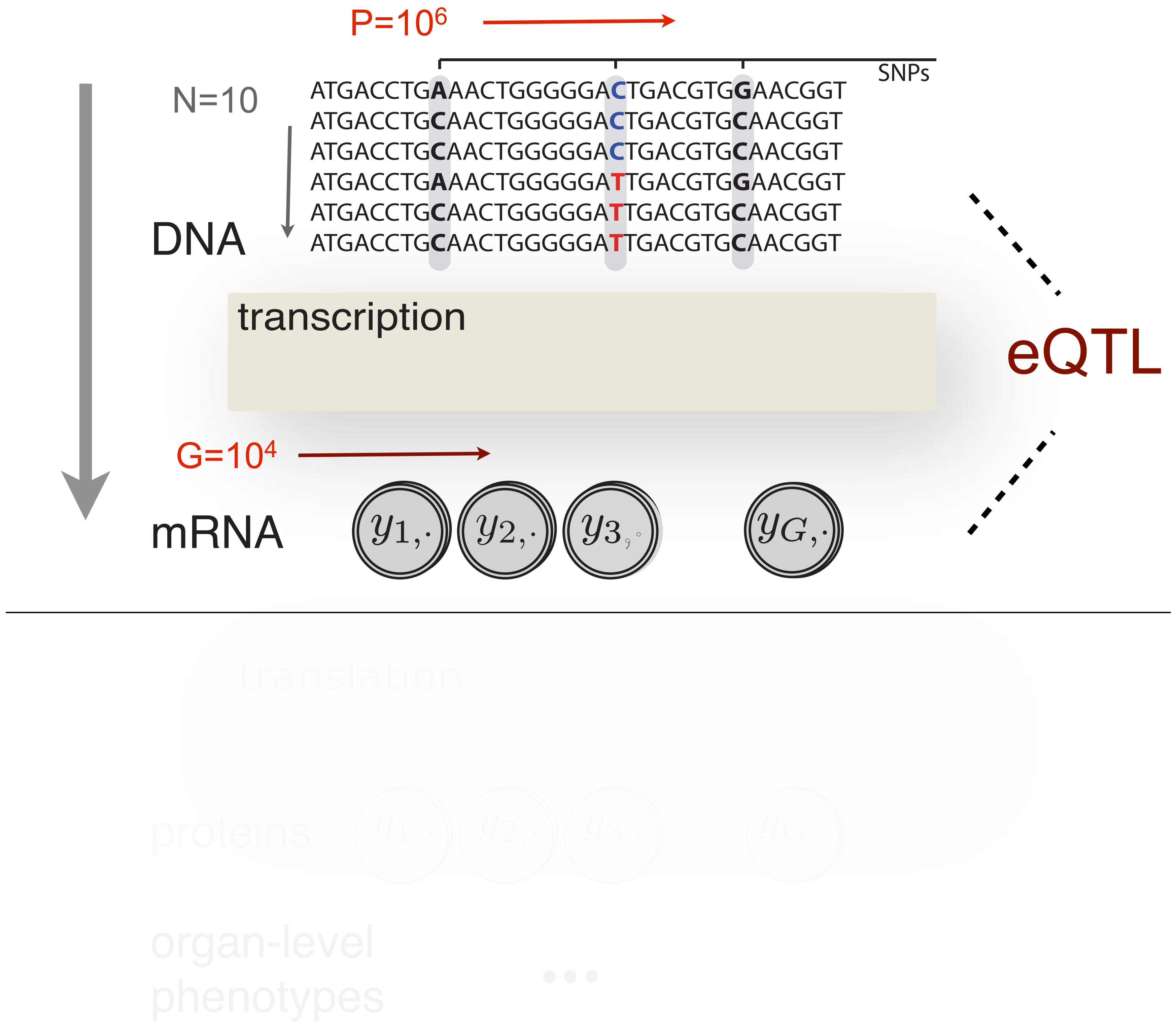
time

eQTL

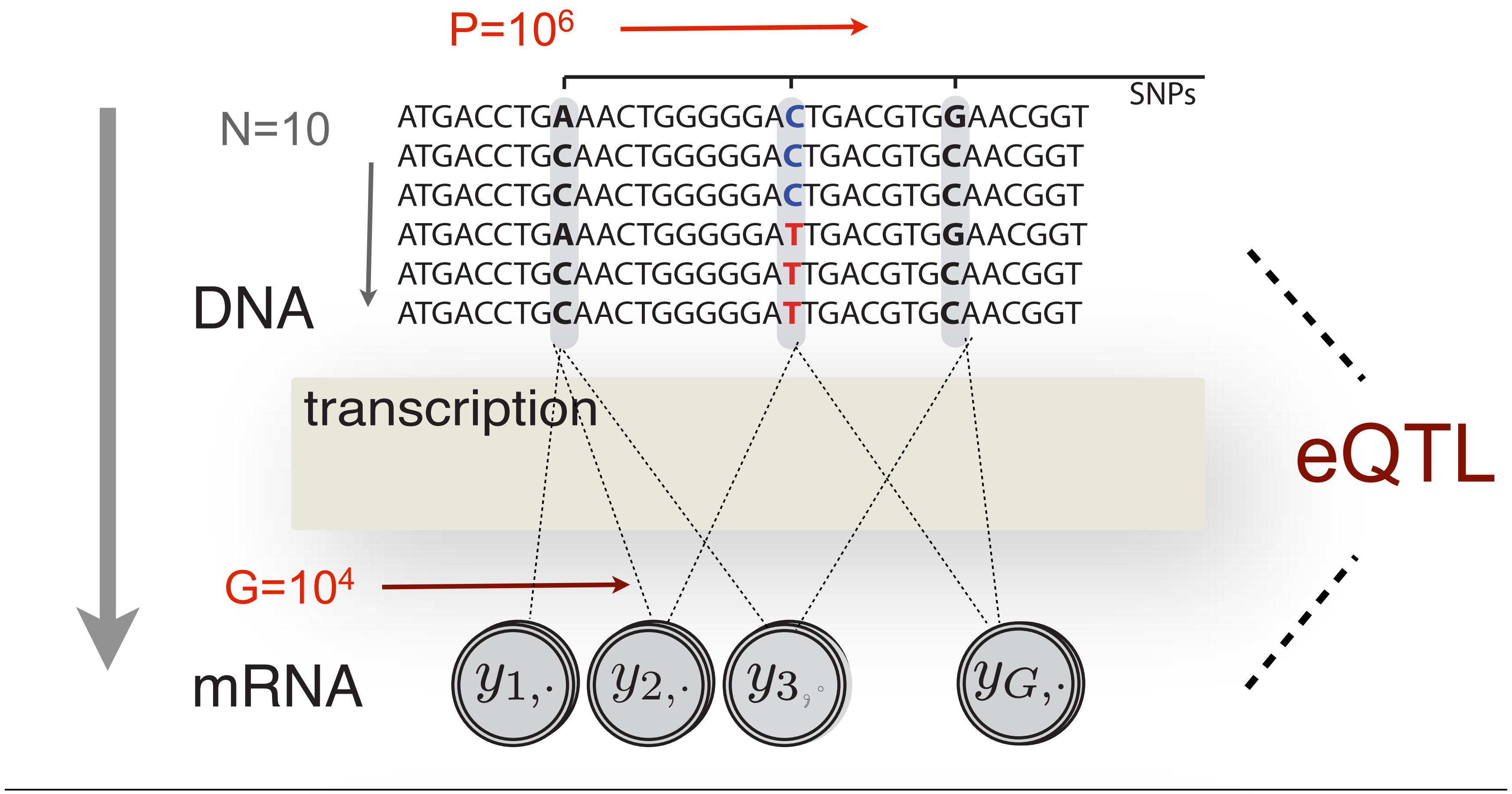
differential
expression

GWAS

Association genetics with high-dimensional phenotypes



Association genetics with high-dimensional phenotypes

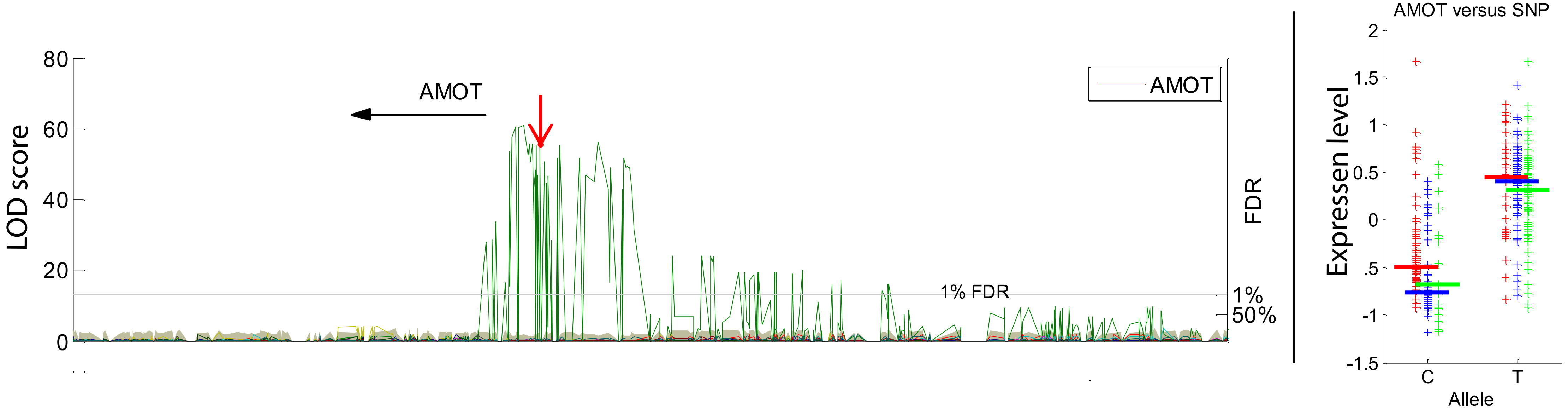


- statistical power
- false positives

Expression quantitative trait loci

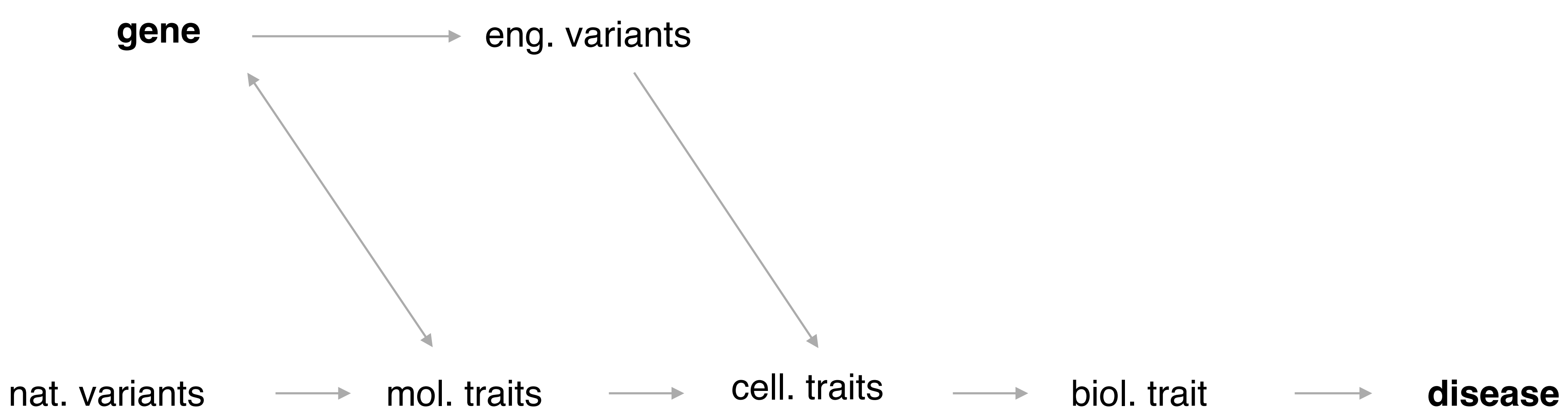
- Single marker genetic mapping

$$\mathbf{y}_g = \underbrace{\mathbf{s}_i \beta_{i,g}}_{\text{genetic}} + \underbrace{\boldsymbol{\epsilon}_g}_{\text{noise}}$$



Stegle et. al PLoS Comp. Biol. 2010
Fusi et. al PLoS Comp. Biol. 2012
Stegle et. al Nat. Protoc. 2012

Why should we care about eQTLs?



gene

eng. variants

nat. variants

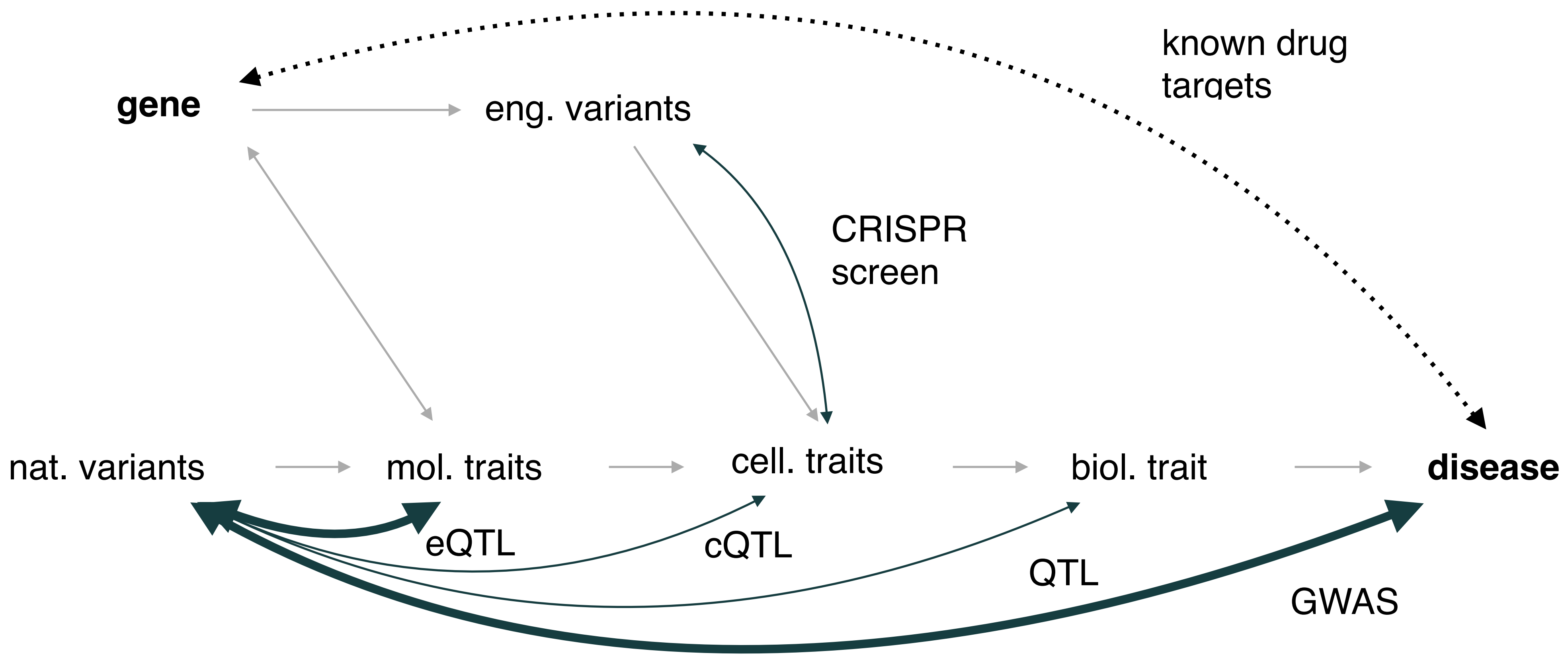
mol. traits

cell. traits

biol. trait

disease

Why should we care about eQTLs?



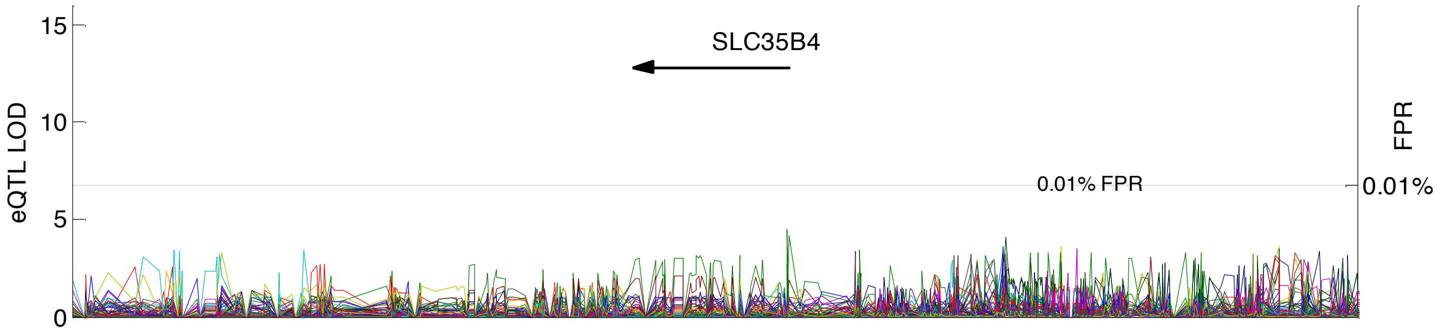
- Challenges:
  - Almost no direct evidence of gene->disease relationships
  - Overlaying eQTLs and GWAS is one of the key evidences
- Wins:
  - Even weak associations (genetic is) are useful.

Expression quantitative trait loci - accounting for row covariances

- Single marker genetic mapping

$$\mathbf{y}_g = \underbrace{\mathbf{s}_i \beta_{i,g}}_{\text{genetic}} + \underbrace{\mathbf{u}}_{\text{confounding}} + \underbrace{\boldsymbol{\epsilon}_g}_{\text{noise}}$$

(a) Standard

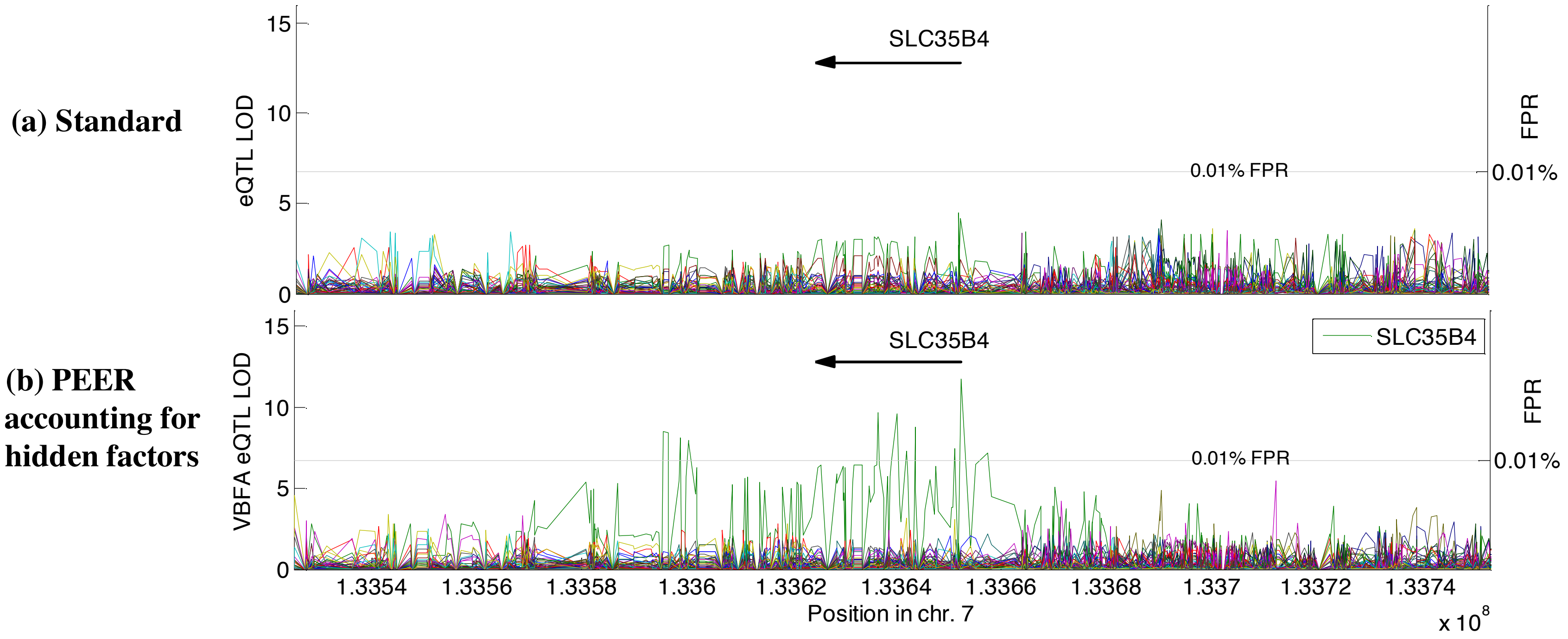


Expression quantitative trait loci - accounting for row covariances

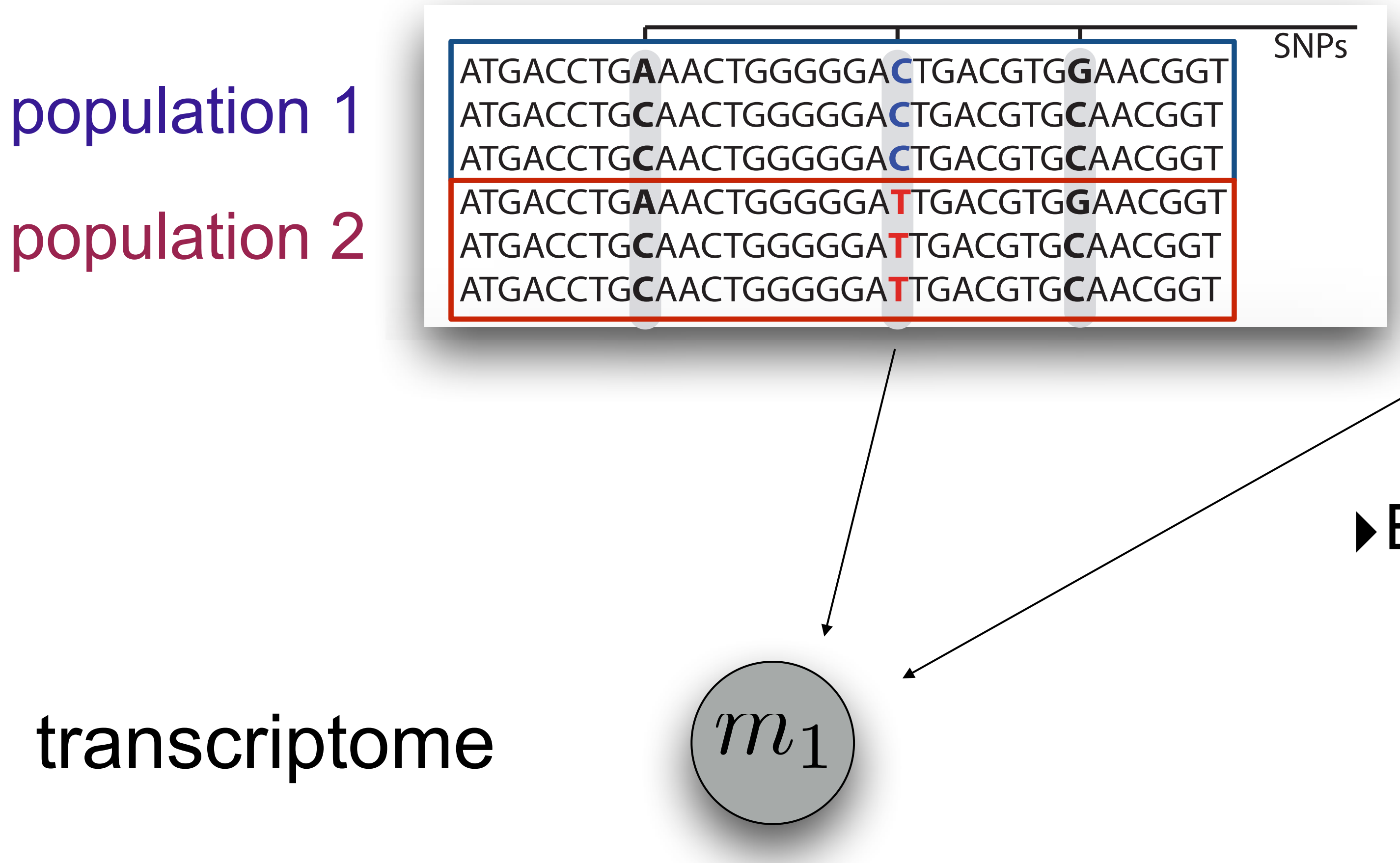
- Single marker genetic mapping

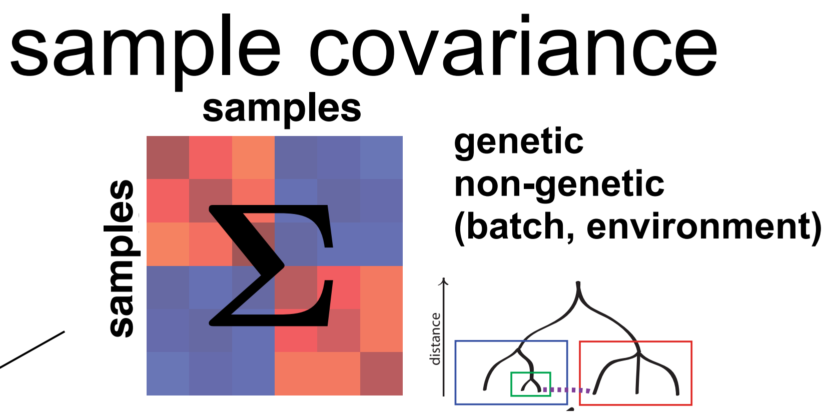
$$\mathbf{y}_g = \underbrace{\mathbf{s}_i \beta_{i,g}}_{\text{genetic}} + \underbrace{\mathbf{u}}_{\text{confounding}} + \underbrace{\boldsymbol{\epsilon}_g}_{\text{noise}}$$

- Accounting for **non-genetic sample heterogeneity** increases power



Accounting for genetic and non-genetic sample covariance


population 1
population 2
SNPs
ATGACCTGAAACTGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT
transcriptome
m_1



- Estimate Σ
  - Population structure: genotype data
  - **Environment/batch: gene expression levels**

$$\mathbf{y}_g = \underbrace{\mathbf{s}_i \beta_{i,g}}_{\text{genetic}} + \underbrace{\mathbf{u}}_{\text{confounding}} + \underbrace{\boldsymbol{\epsilon}_g}_{\text{noise}}$$

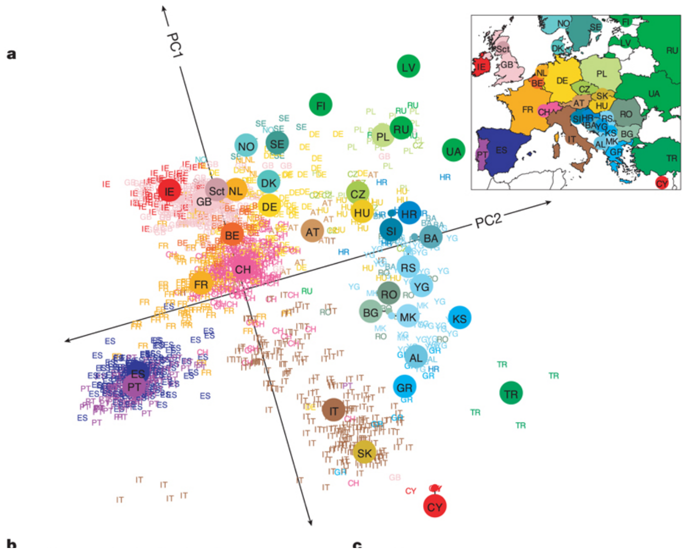
$$\mathbf{u} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$$

Leek & Storey, 2007
Kang et al. 2008
Listgarten et al, 2010
Lipper et al. 2011
Stegle* & Parts* et al. 2010, 2012
Fusi* & Stegle* et al. 2012

Accounting for genetic and non-genetic sample covariance

- genetic

$$\mathbf{\Sigma} = \mathbf{S}\mathbf{S}^{\mathrm{T}}$$

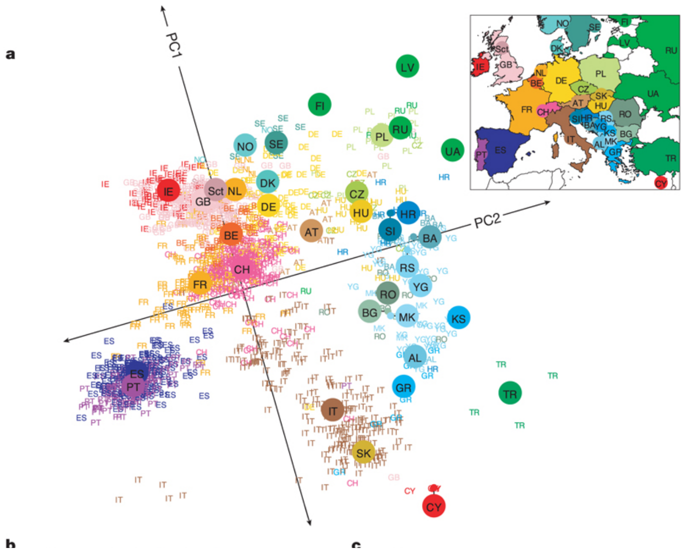


Fusi, N. & Stegle, O. et al. PLoS Computational Biology, 2012

Accounting for genetic and non-genetic sample covariance

- genetic

$$\mathbf{\Sigma} = \mathbf{S}\mathbf{S}^{\mathrm{T}}$$



- non-genetic

$$\mathbf{\Sigma} = \mathbf{Y}\mathbf{Y}^{\mathrm{T}}$$

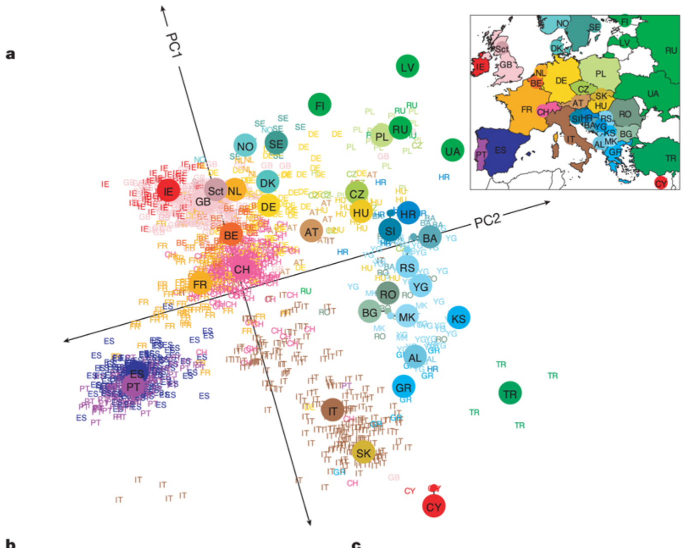
- Empirical gene expression covariance

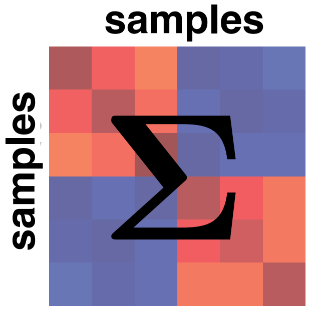
Fusi, N. & Stegle, O. et al. PLoS Computational Biology, 2012

Accounting for genetic and non-genetic sample covariance

- genetic

$$\mathbf{\Sigma} = \mathbf{S}\mathbf{S}^{\mathrm{T}}$$





- non-genetic

$$\mathbf{\Sigma} = \mathbf{Y}\mathbf{Y}^{\mathrm{T}}$$

- Empirical gene expression covariance

$$p(\mathbf{Y} \mid \sigma_g^2, \sigma_k^2, \sigma_e^2, \mathbf{S}, \mathbf{X}) = \prod_{g=1}^{G} \mathcal{N}\left(\mathbf{y}_g \mid \mathbf{0}, \underset{\text{genetic}}{\sigma_g^2 \mathbf{S}\mathbf{S}^{\mathrm{T}}} + \underset{\text{non-genetic}}{\sigma_k^2 \mathbf{X}\mathbf{X}^{\mathrm{T}}} + \sigma_e^2 \mathbf{I}\right)$$

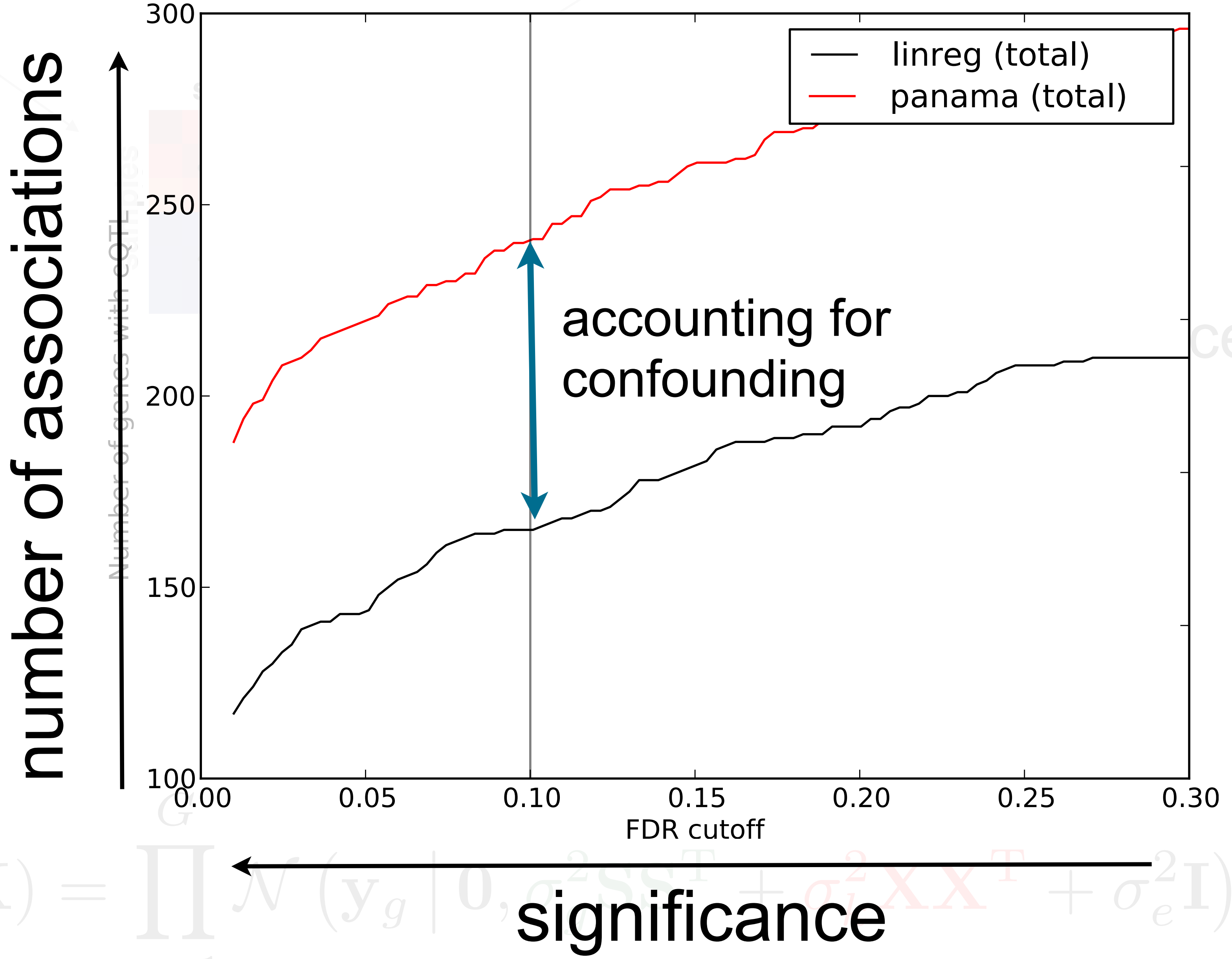
Fusi, N. & Stegle, O. et al. PLoS Computational Biology, 2012

Accounting for genetic and non-genetic sample covariance

- genetic

- non-genetic

$\Sigma = \mathbf{S}\mathbf{S}^{\mathrm{T}}$



$$p(\mathbf{Y} \,|\, \sigma_g^2, \sigma_k^2, \sigma_e^2, \mathbf{S}, \mathbf{X}) = \prod_{g=1}^{G} \mathcal{N}\left(\mathbf{y}_g \,|\, \mathbf{0}, \sigma_g^2 \mathbf{S}\mathbf{S}^{\mathrm{T}} + \sigma_k^2 \mathbf{X}\mathbf{X}^{\mathrm{T}} + \sigma_e^2 \mathbf{I}\right)$$

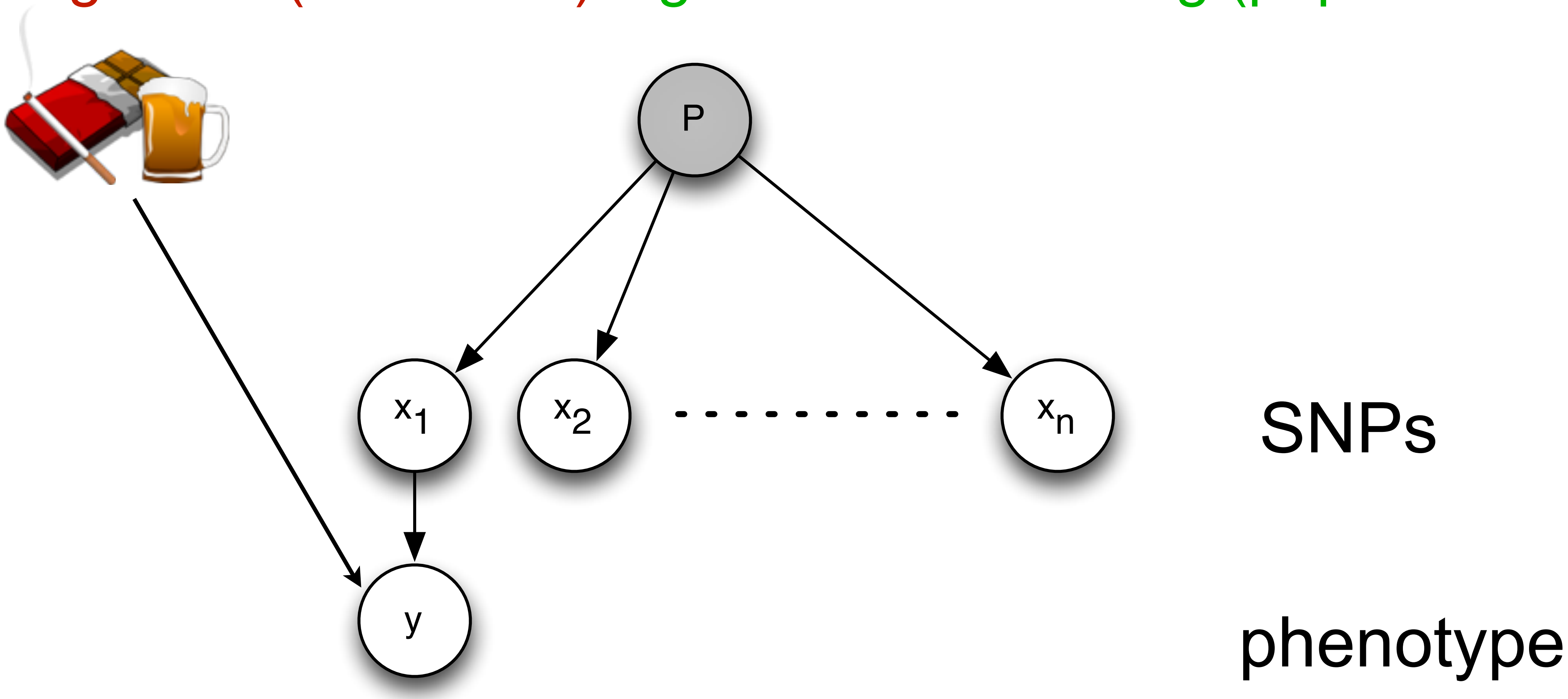
genetic non-genetic

Nature, Lappalainen et al. 2013

Fusi, N. & Stegle, O. et al. PLoS Computational Biology, 2012

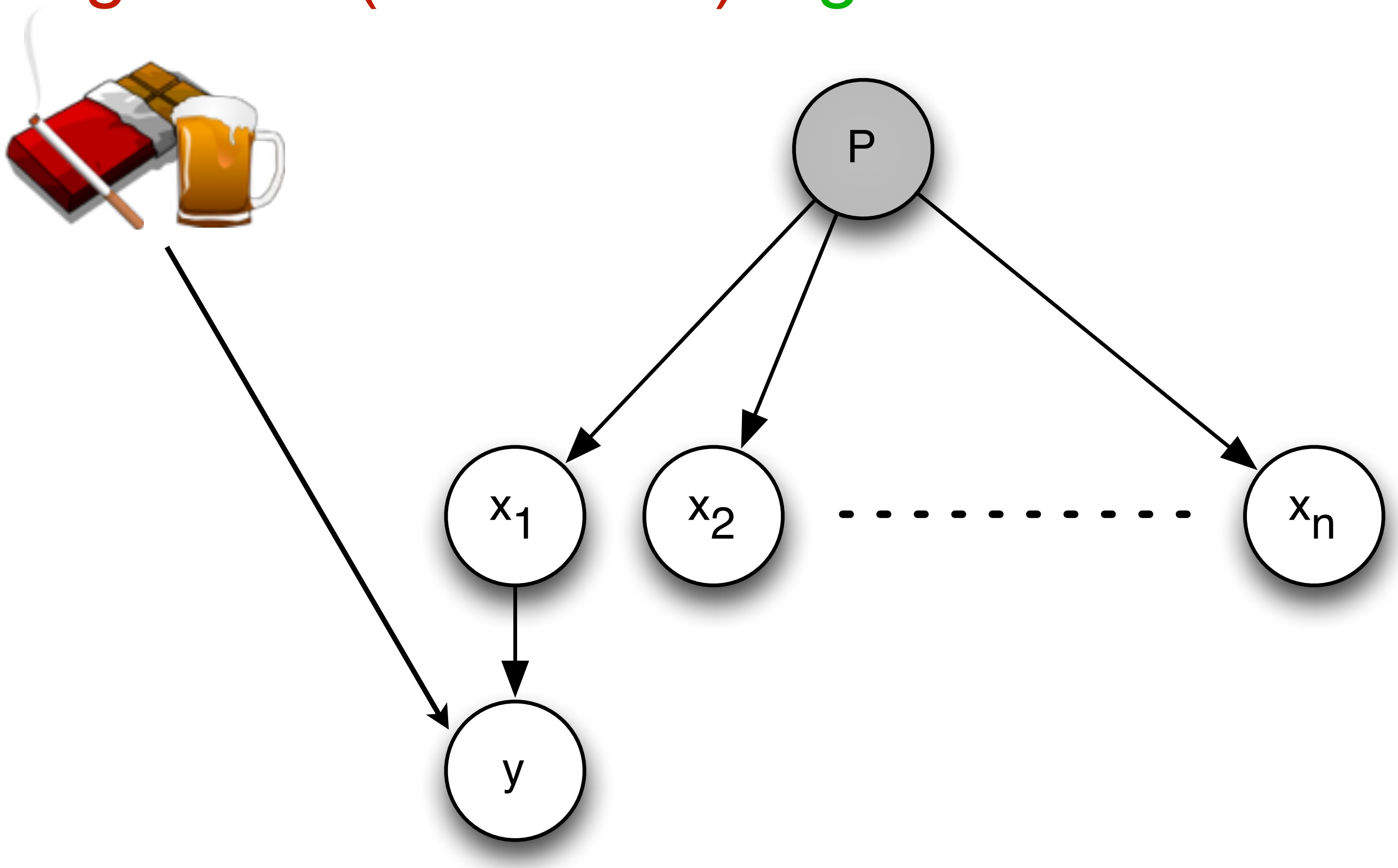
Confounding factors: genetic and non-genetic structure

- non-genetic (batch/env)
- genetic confounding (population structure)



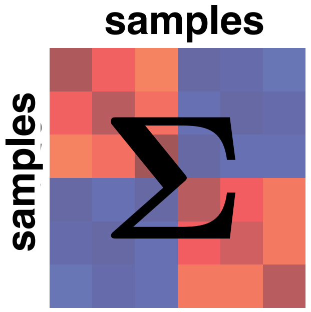
Confounding factors: genetic and non-genetic structure

- non-genetic (batch/env)
- genetic confounding (population structure)



SNPs

phenotype



Summary so far

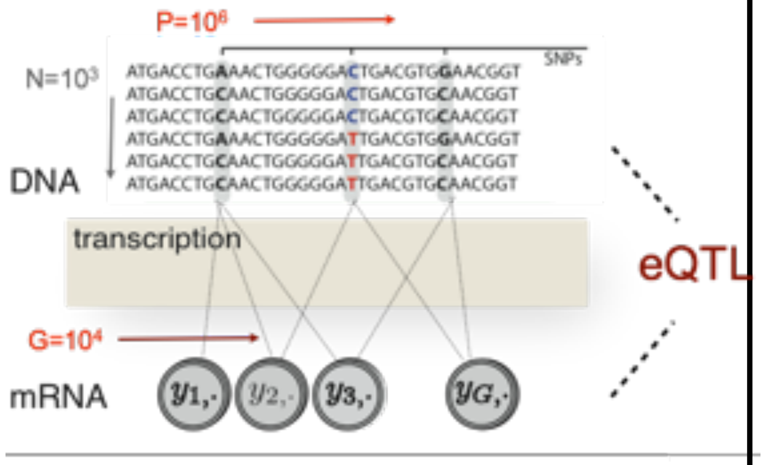
- Linear mixed models help to adjust for non-IID sample structure such as relatedness and population structure.
- Both local and global **genetic structure** can be estimated from the genotype data itself.
- Multivariate modeling allows to exploit genetic covariances in different ways, including to test for the effect of local regions.

- If phenotypes are high-dimensional, **non-genetic sample structure** can be estimated from the phenotype data itself, allowing to account for environment factors or batch.

Accounting for heterogeneity is key…

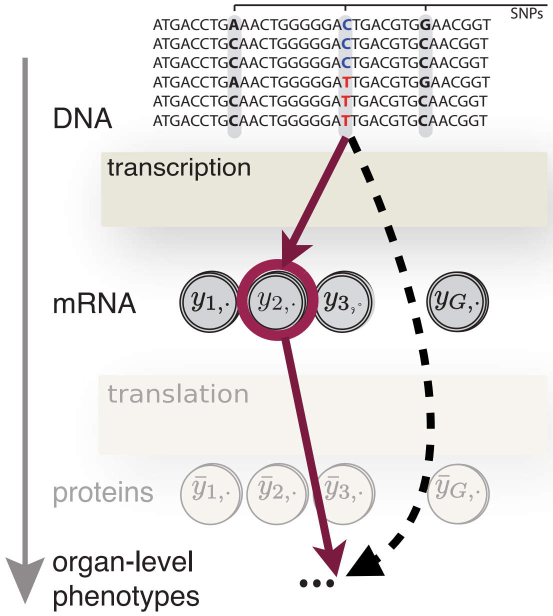
(e)QTL mapping

- multiple phenotype models
- variance components


P=10^6
N=10^3
SNPs
DNA
ATGACCTGAAACTGGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
transcription
eQTL
G=10^4
mRNA
$y_{1,\cdot}$
$y_{2,\cdot}$
$y_{3,\cdot}$
$y_{G,\cdot}$

Causality in molecular systems

- prediction of causal mediators
- ordering of pathways

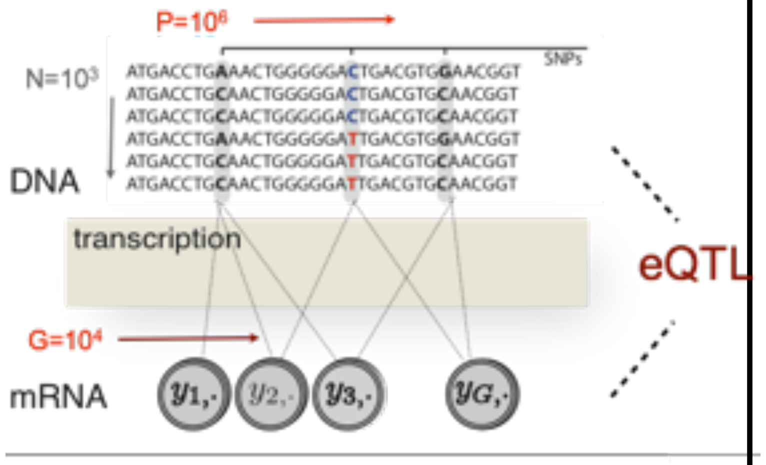

SNPs
DNA
ATGACCTGAAACTGGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
transcription
mRNA
$y_{1,\cdot}$
$y_{2,\cdot}$
$y_{3,\cdot}$
$y_{G,\cdot}$
translation
proteins
$\bar{y}_{1,\cdot}$
$\bar{y}_{2,\cdot}$
$\bar{y}_{3,\cdot}$
$\bar{y}_{G,\cdot}$
organ-level phenotypes
...

- PLoS Genet, Gagneur et al. 2013

Accounting for heterogeneity is key…

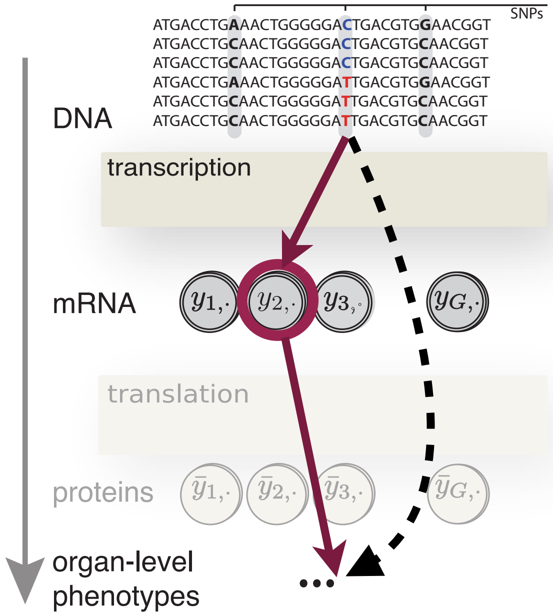
(e)QTL mapping

- multiple phenotype models
- variance components


P=10⁶
N=10³
SNPs
ATGACCTGAAACTGGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
DNA
transcription
G=10⁴
mRNA
$y_{1,\cdot}$
$y_{2,\cdot}$
$y_{3,\cdot}$
$y_{G,\cdot}$
eQTL

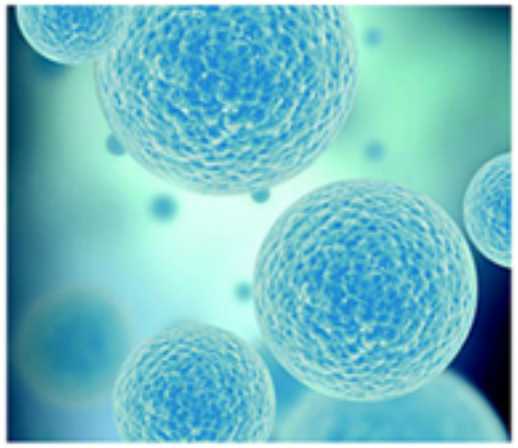
Causality in molecular systems

- prediction of causal mediators
- ordering of pathways


SNPs
ATGACCTGAAACTGGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
DNA
transcription
mRNA
$y_{1,\cdot}$
$y_{2,\cdot}$
$y_{3,\cdot}$
$y_{G,\cdot}$
translation
proteins
$\tilde{y}_{1,\cdot}$
$\tilde{y}_{2,\cdot}$
$\tilde{y}_{3,\cdot}$
$\tilde{y}_{G,\cdot}$
organ-level phenotypes
...

- PLoS Genet, Gagneur et al. 2013

Single-cell transcriptomics



Gene expression heterogeneity between individuals and single cells

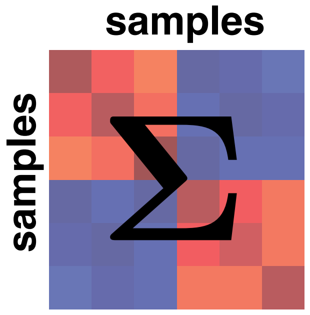
variation of interest

confounding

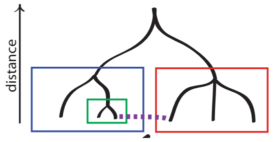
population variation

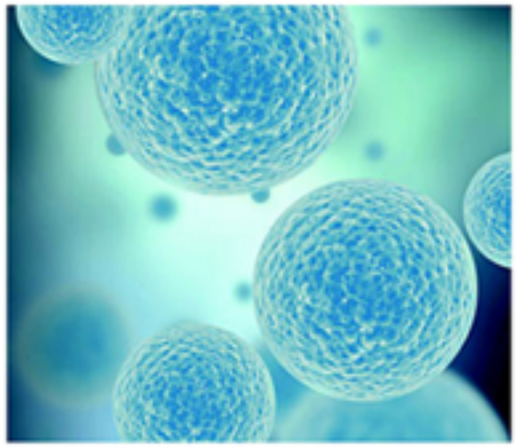
genetic associations with phenotype

sample covariance



genetic
non-genetic
(batch, environment)

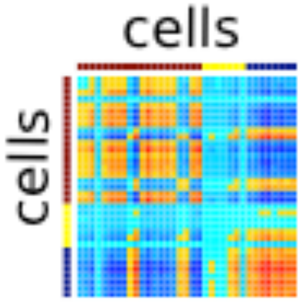




single-cell variation

differentiation processes
Correlations between genes

cell covariance



cell cycle
stress

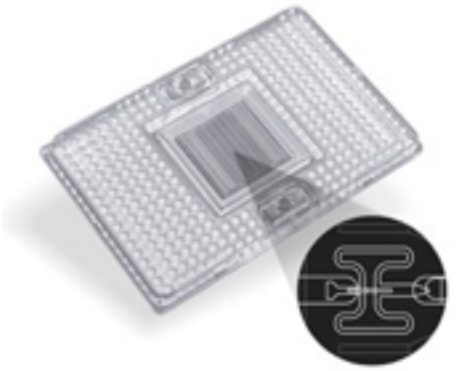
Single-cell RNA-Seq

- Conventional RNA-Seq profiles are obtained from a pool of typically ~100,000+ cells.
- Using single-cell RNA-sequencing technologies, we can now assay RNA abundance in single cells.

- novel variation between cells:

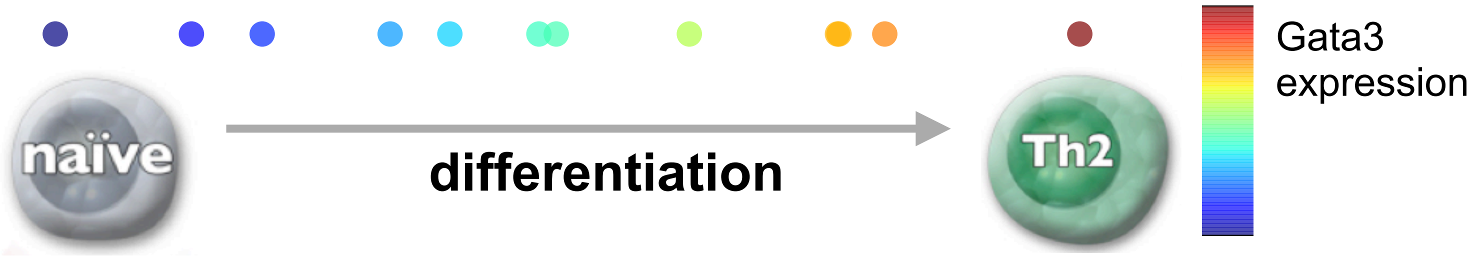
  cell type composition, **differentiation**

- additional (confounding) expression heterogeneity: **cell cycle**, apoptosis, …

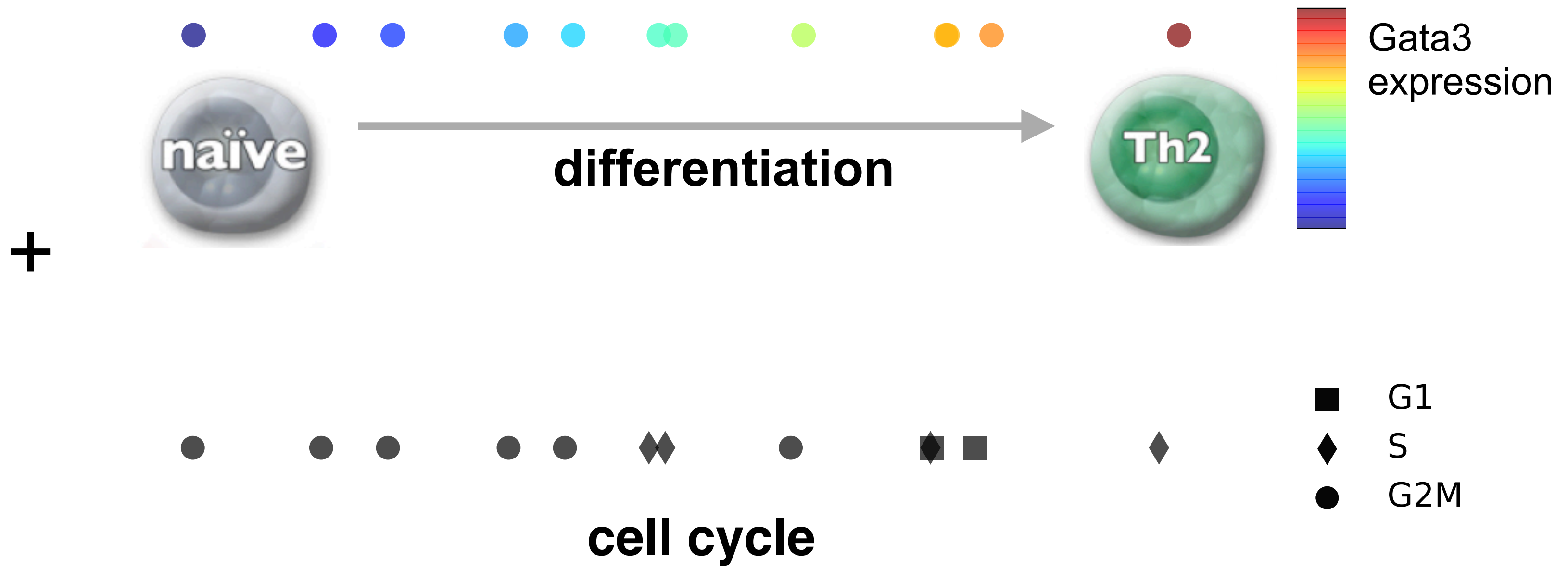


Fluidigm C1®

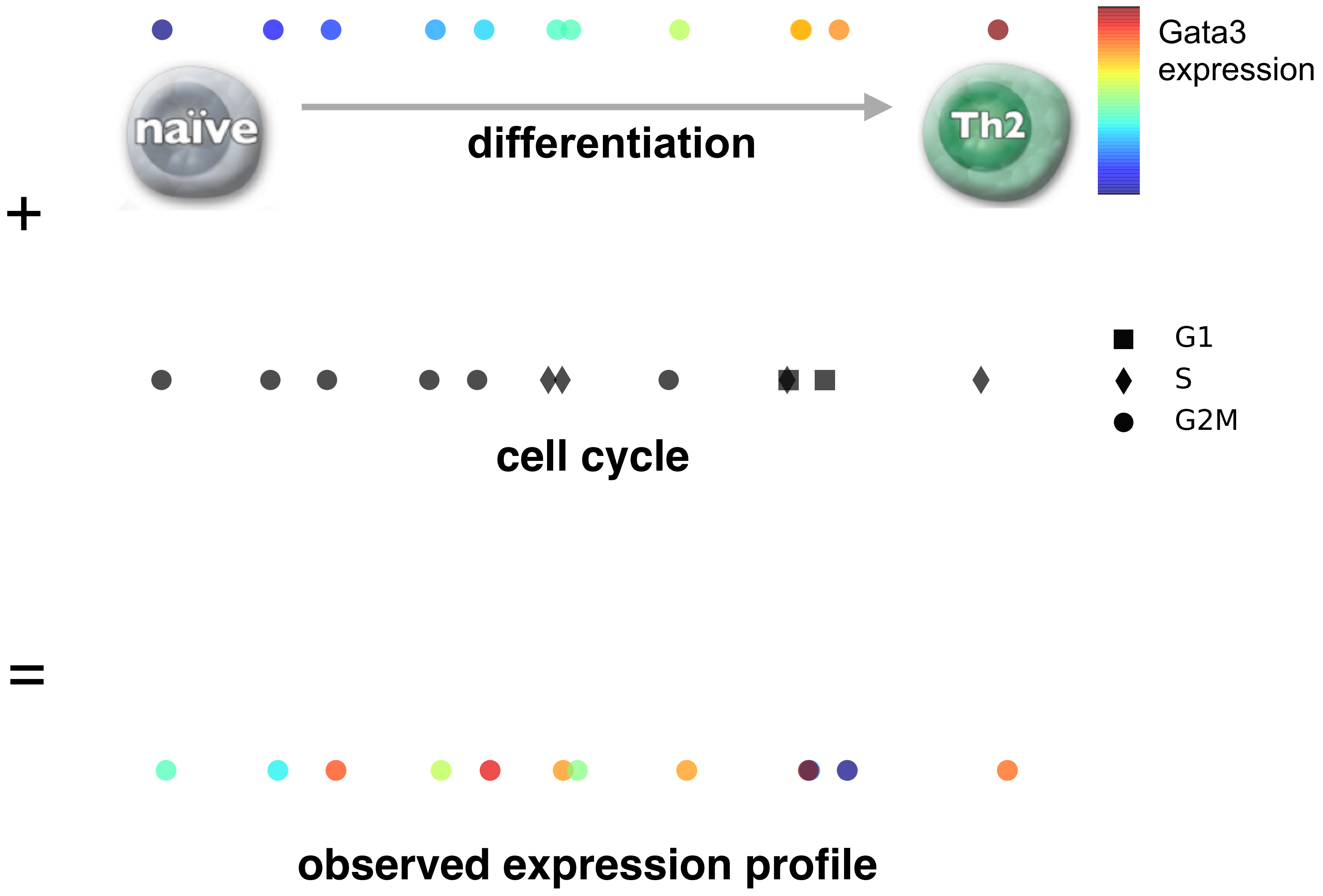
Cell cycle masks differentiation processes in single-cell RNA-Seq



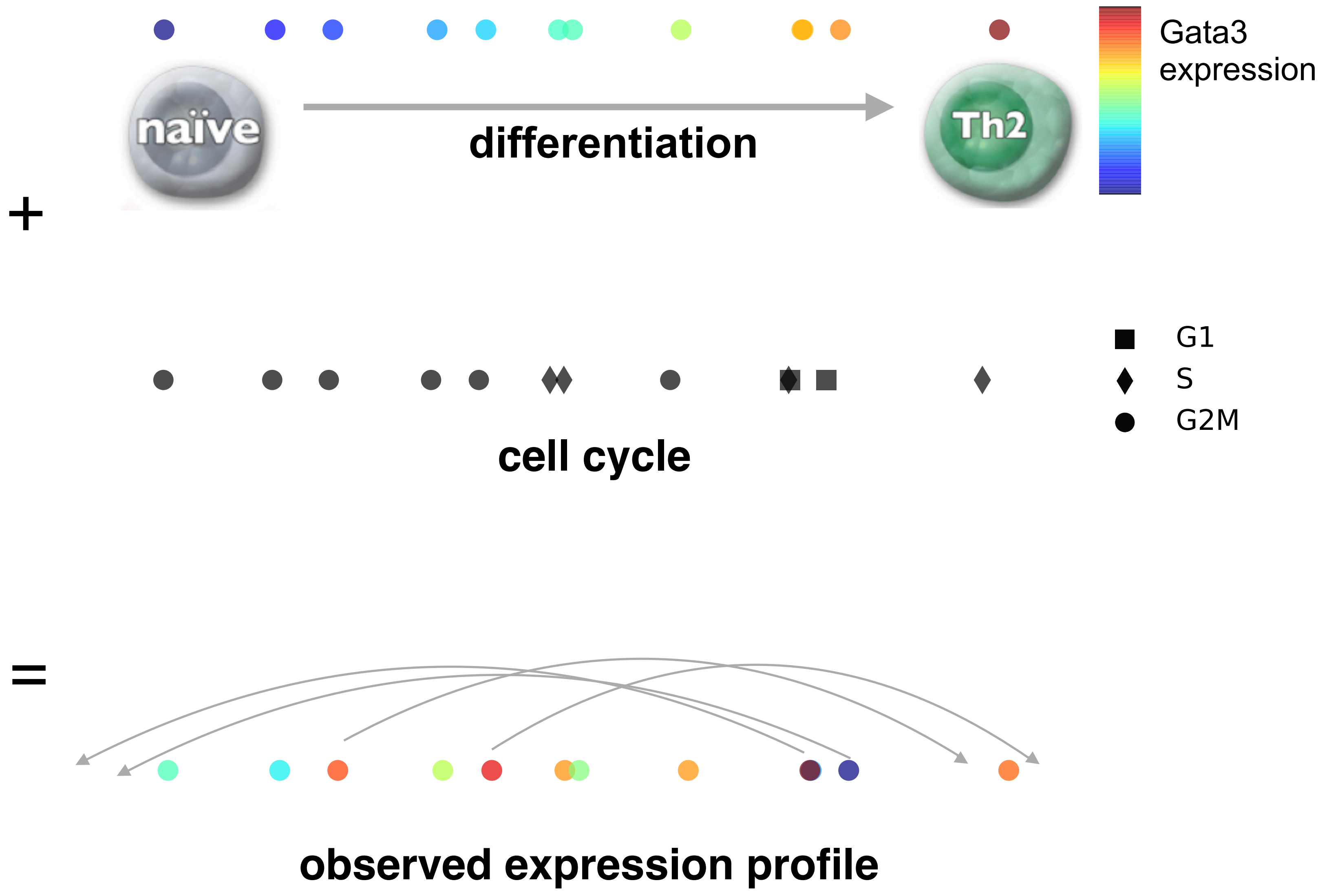
Cell cycle masks differentiation processes in single-cell RNA-Seq

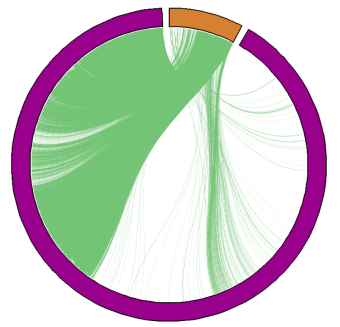


Cell cycle masks differentiation processes in single-cell RNA-Seq



Cell cycle masks differentiation processes in single-cell RNA-Seq





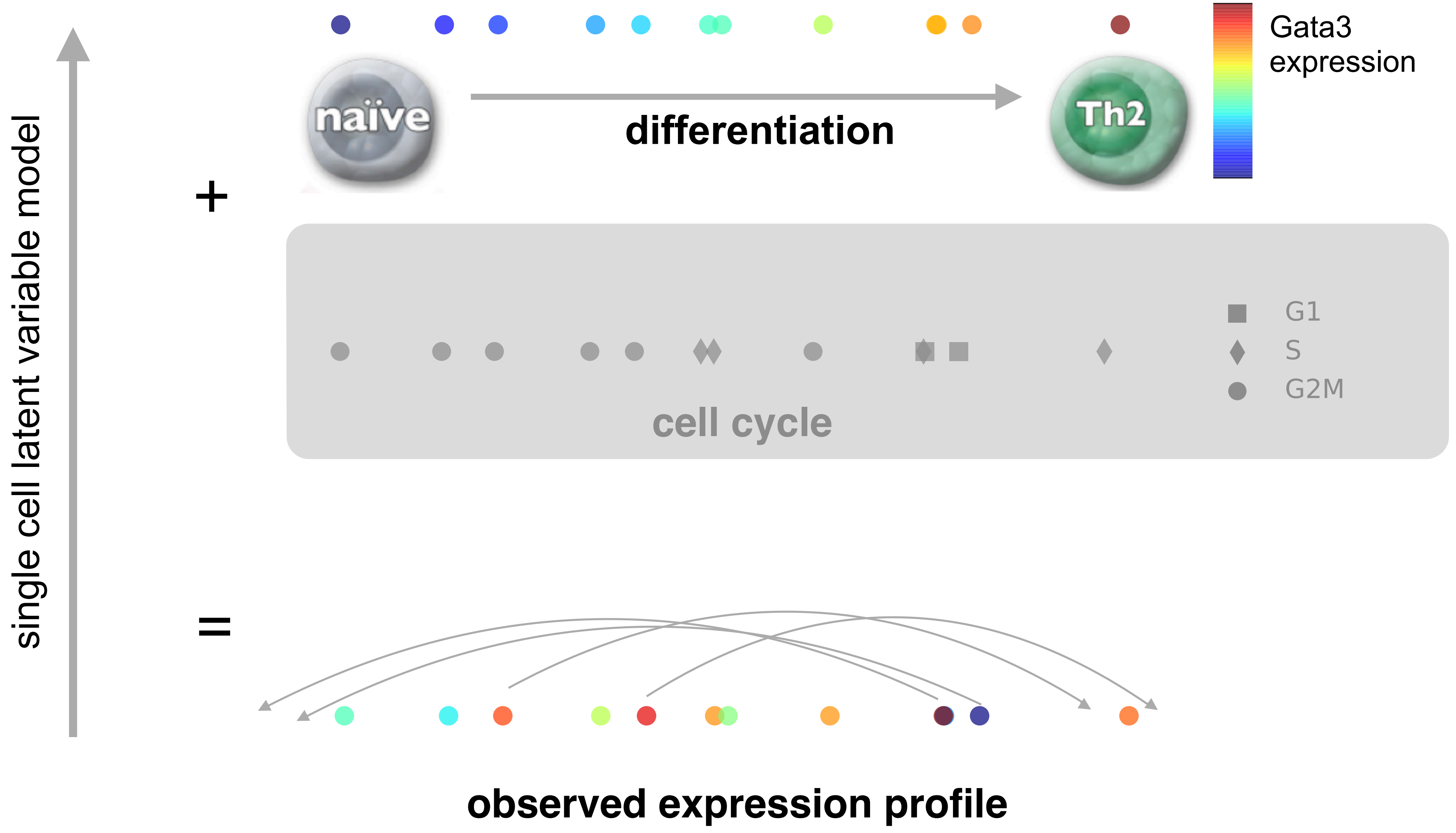
549 genes annotated to cell-cycle

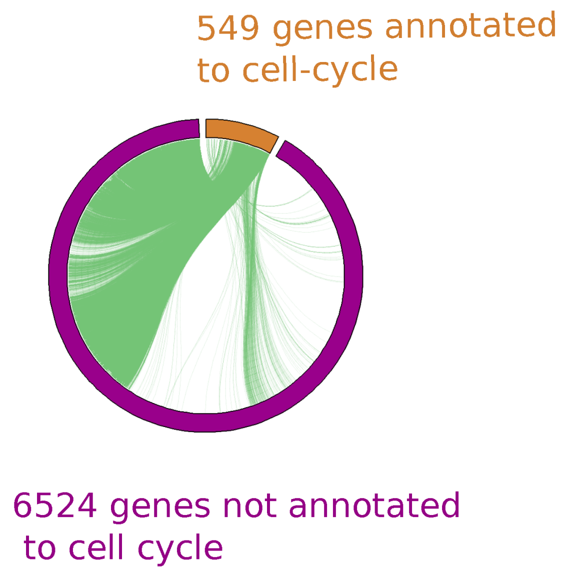
6524 genes not annotated to cell cycle

wide-spread correlation between cell cycle genes and non-cycle genes

- Observed expression profiles do not enable recovering of the differentiation process.

Cell cycle masks differentiation processes in single-cell RNA-Seq

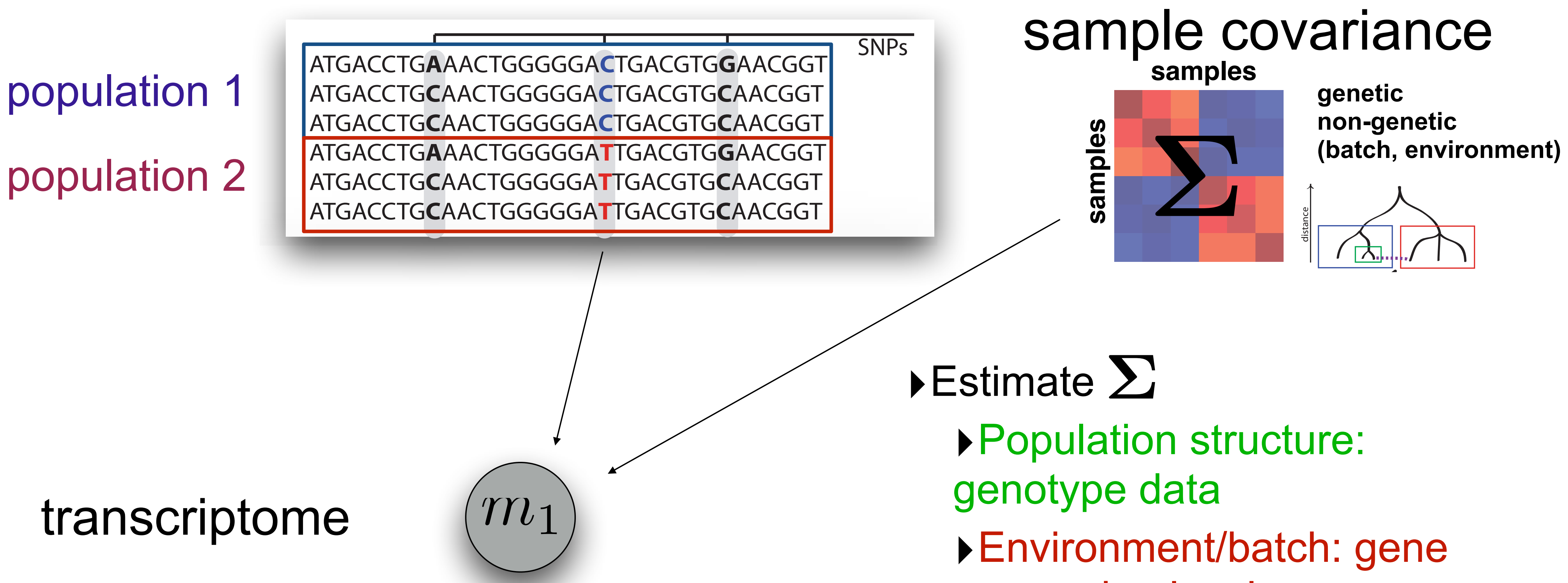

single cell latent variable model
naive
differentiation
Th2
Gata3 expression
+
G1
S
G2M
cell cycle
=
observed expression profile



wide-spread correlation between cell cycle genes and non-cycle genes

- Observed expression profiles do not enable recovering of the differentiation process.

Gene expression heterogeneity is not new…



population 1

population 2

SNPs

ATGACCTGAAACTGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT

sample covariance

samples

samples

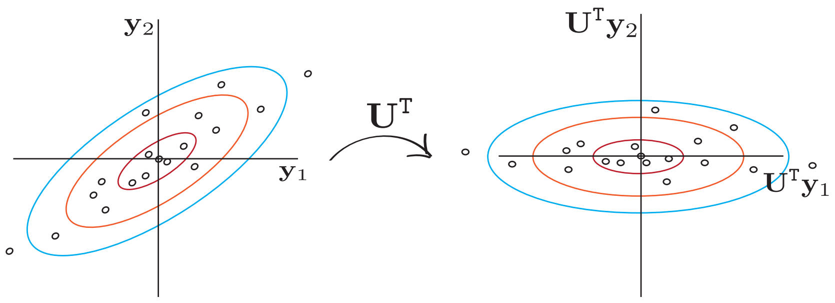
Σ

genetic
non-genetic
(batch, environment)

transcriptome

m_1

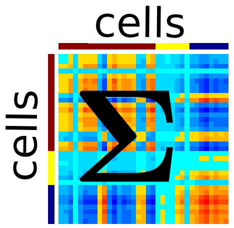
- Estimate Σ
  - Population structure: genotype data
  - Environment/batch: gene expression levels
- Correct for Σ using mixed models
  - “Rotate” phenotypes & genotypes



Leek & Storey, 2007
Kang et al. 2008
Listgarten et al, 2010
Lipper et al. 2011
Stegle* & Parts* et al. 2010, 2012
Fusi* & Stegle* et al. 2012

Single-cell latent variable model (scLVM)

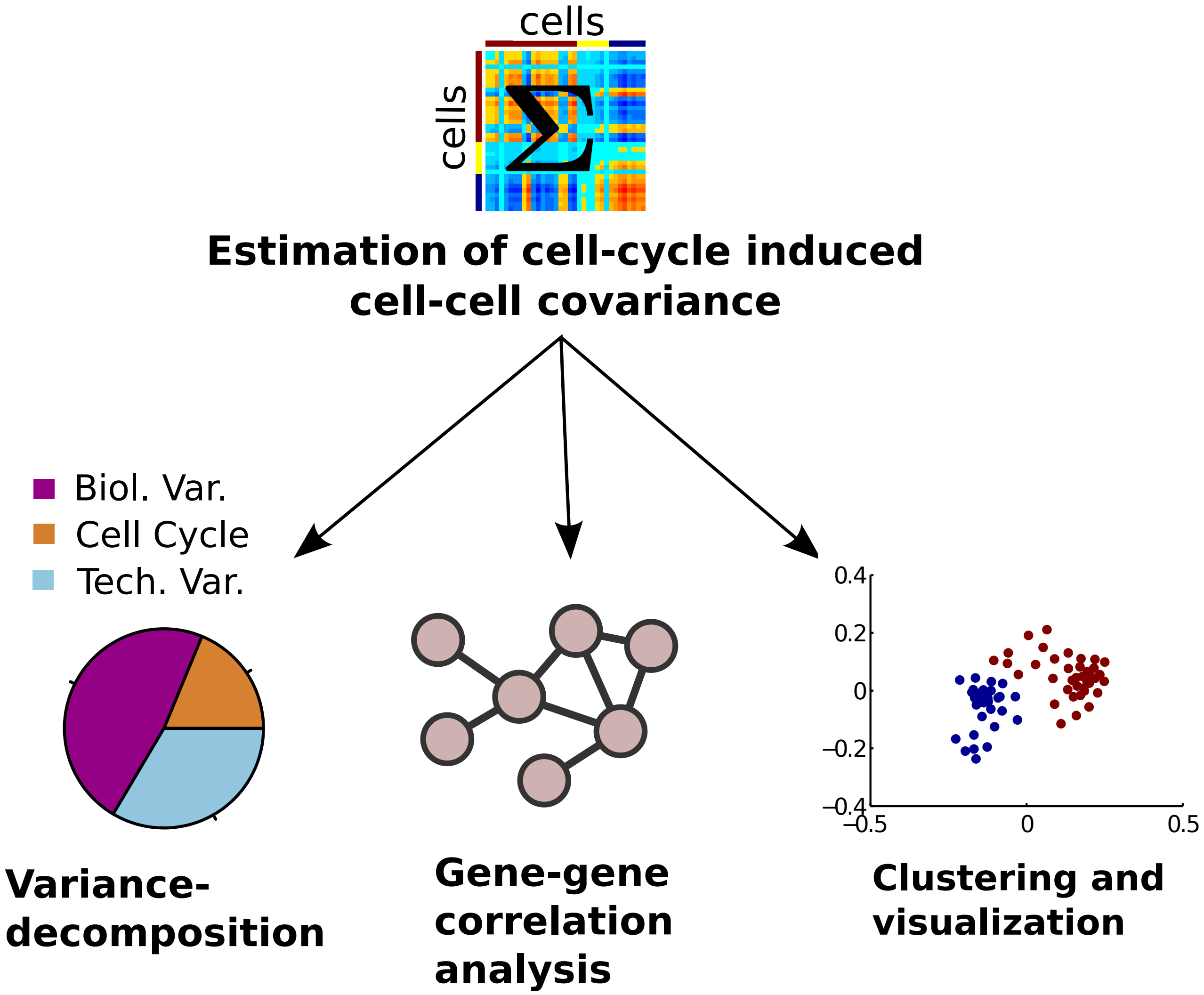
- Random effect model for cell cycle effects. Two-stage approach:
    1. Estimate a cell-cell covariance that captures cell cycle



Estimation of cell-cycle induced

Single-cell latent variable model (scLVM)

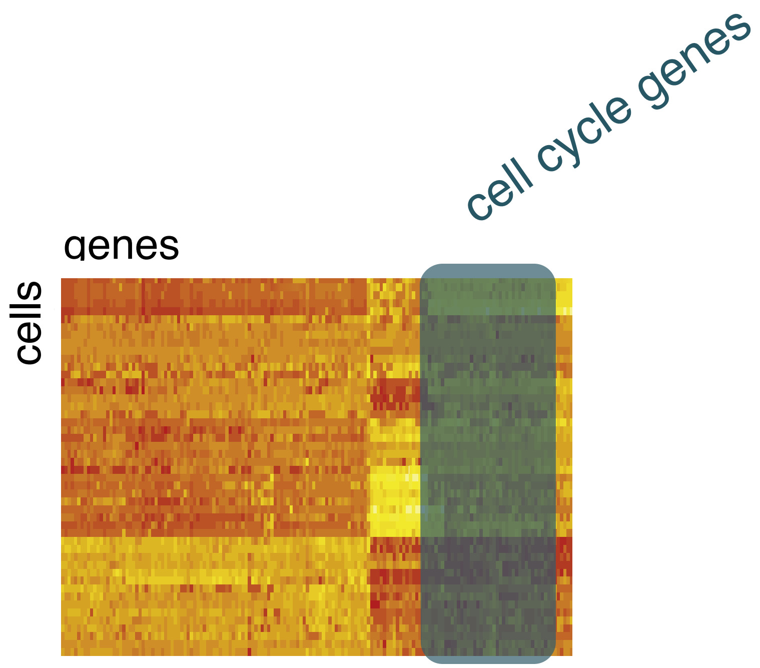
- Random effect model for cell cycle effects. Two-stage approach:
  1. Estimate a cell-cell covariance that captures cell cycle
  2. Account for cell cycle in
     - Variance decomposition
     - Gene-gene correlation analysis
     - Cell clustering



cells
cells
Σ
Estimation of cell-cycle induced cell-cell covariance
Biol. Var.
Cell Cycle
Tech. Var.
Variance-decomposition
Gene-gene correlation analysis
0.4
0.2
0
−0.2
−0.4
−0.5
0
0.5
Clustering and visualization

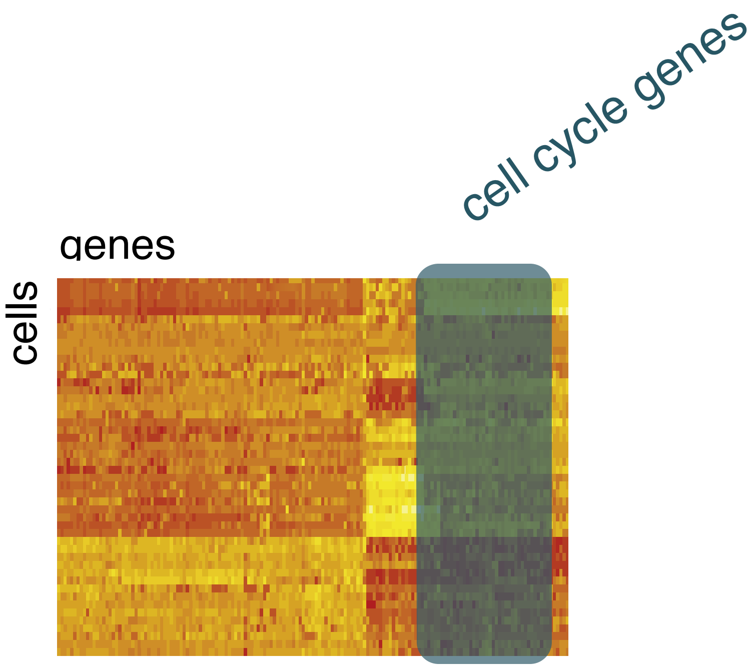
Estimating the cell cycle covariance

- Reconstruct cell cycle from the observed expression data
- Use known annotated cell cycle gene set



Estimating the cell cycle covariance

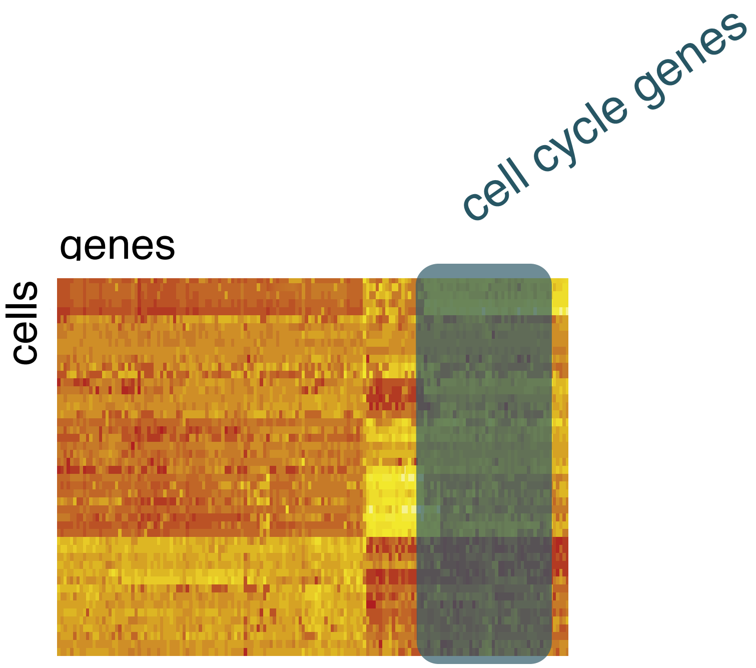
- Reconstruct cell cycle from the observed expression data
- Use known annotated cell cycle gene set
- Employ latent variable modeling to reconstruct a cell cycle factor (**X**)



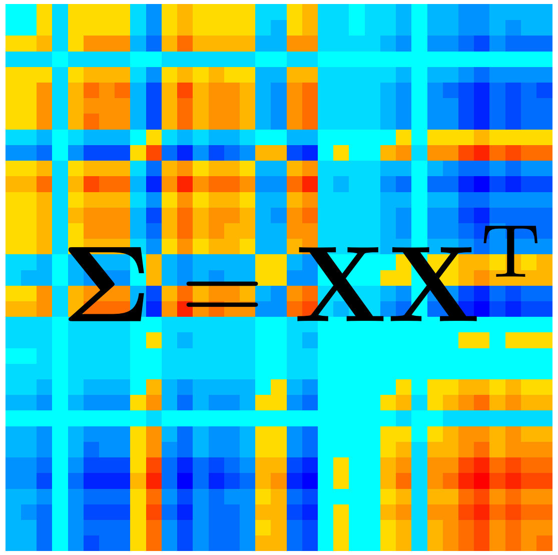
$$\mathbf{Y}_{\mathrm{cc}} \sim \prod_g \mathcal{N}(\mathbf{0} \,|\, \underbrace{\mathbf{X}\mathbf{X}^{\mathrm{T}}}_{\text{cell cycle covariance}} + \underbrace{\delta_b \mathbf{I}}_{\text{residual variance}})$$

Estimating the cell cycle covariance

- Reconstruct cell cycle from the observed expression data
- Use known annotated cell cycle gene set
- Employ latent variable modeling to reconstruct a cell cycle factor (**X**)

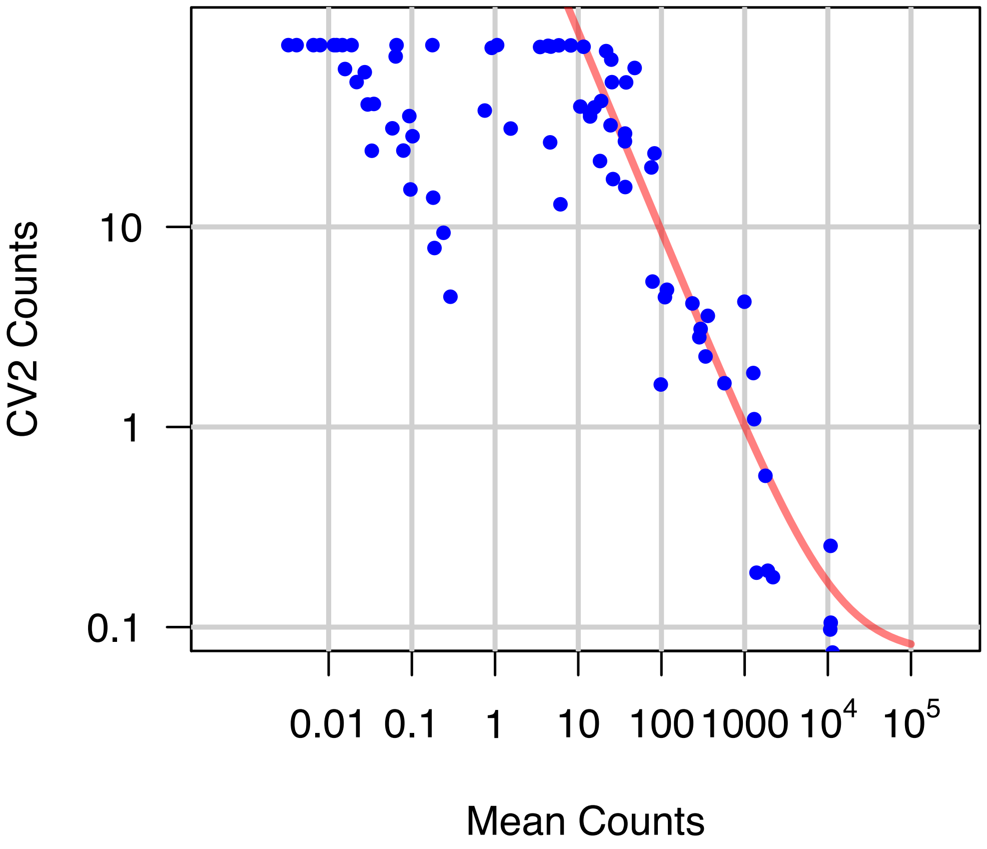


$$\mathbf{Y}_{\mathrm{cc}} \sim \prod_{g} \mathcal{N}(\mathbf{0} \mid \underbrace{\mathbf{X}\mathbf{X}^{\mathrm{T}}}_{\text{cell cycle covariance}} + \underbrace{\delta_b \mathbf{I}}_{\text{residual variance}})$$



Technical noise requires special attention

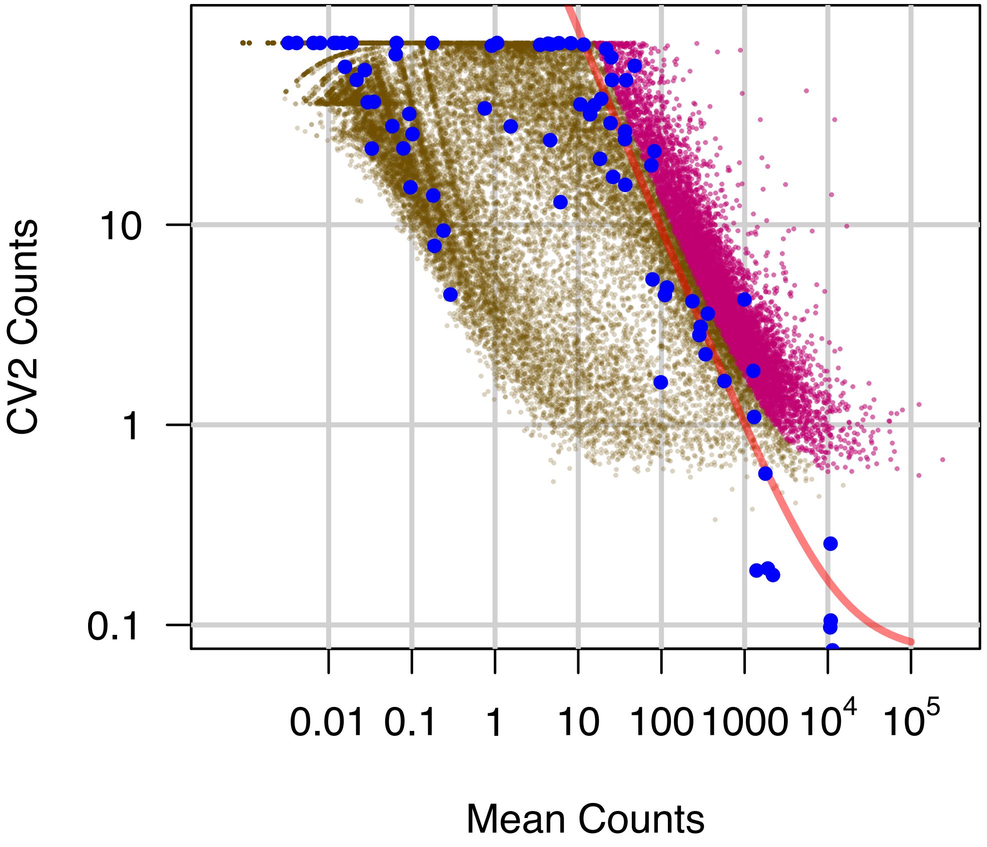
- Large proportions of technical variability due to low quantities of starting material
- Estimation of technical noise
  - Mean/variance fit from ERCC spike ins



Brennecke et al. 2013

Technical noise requires special attention

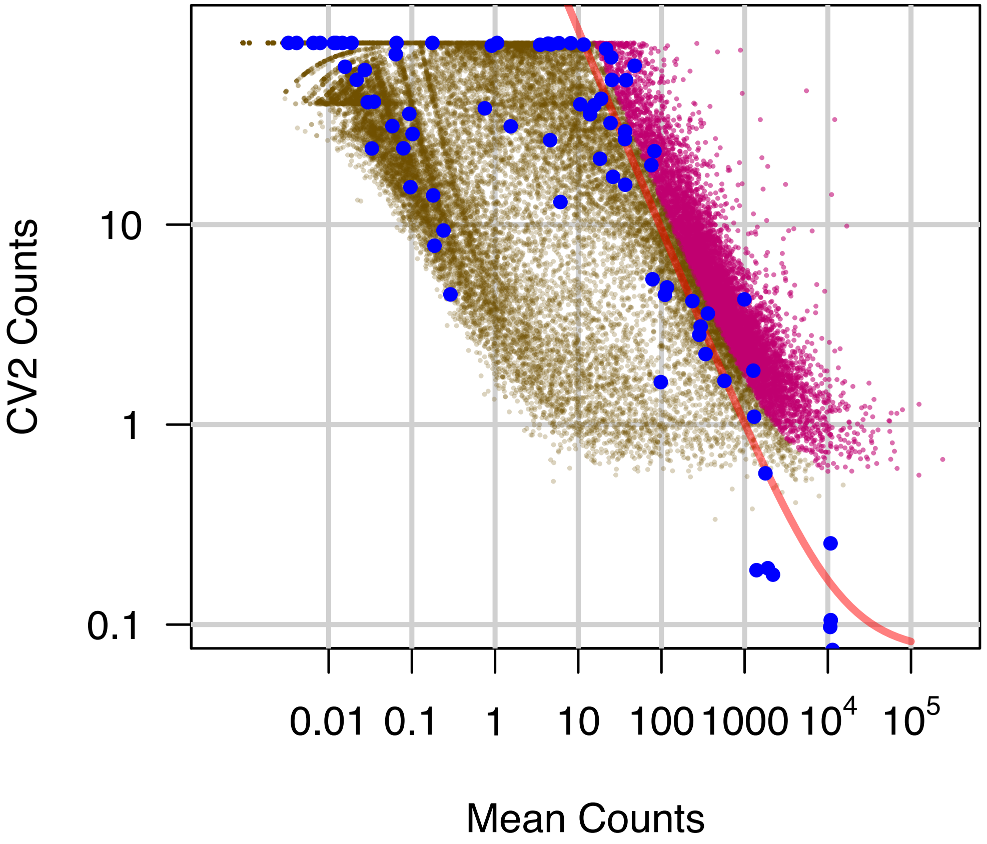
- Large proportions of technical variability due to low quantities of starting material
- Estimation of technical noise
  - Mean/variance fit from ERCC spike ins
  - Extrapolation to genome-wide genes
  - 7,073 highly variable genes



Brennecke et al. 2013

Technical noise requires special attention

- Large proportions of technical variability due to low quantities of starting material
- Estimation of technical noise
  - Mean/variance fit from ERCC spike ins
  - Extrapolation to genome-wide genes
  - 7,073 highly variable genes



Brennecke et al. 2013

$$\mathbf{Y}_{\text{cc}} \sim \prod_g \mathcal{N}(\mathbf{0} \,|\, \underbrace{\mathbf{X}\mathbf{X}^{\mathrm{T}}}_{\text{cell cycle covariance}} + \underbrace{\delta_b \mathbf{I}}_{\text{residual variance}} + \underbrace{\mathrm{diag}(\boldsymbol{\sigma}_g^2)}_{\text{technical variance}})$$

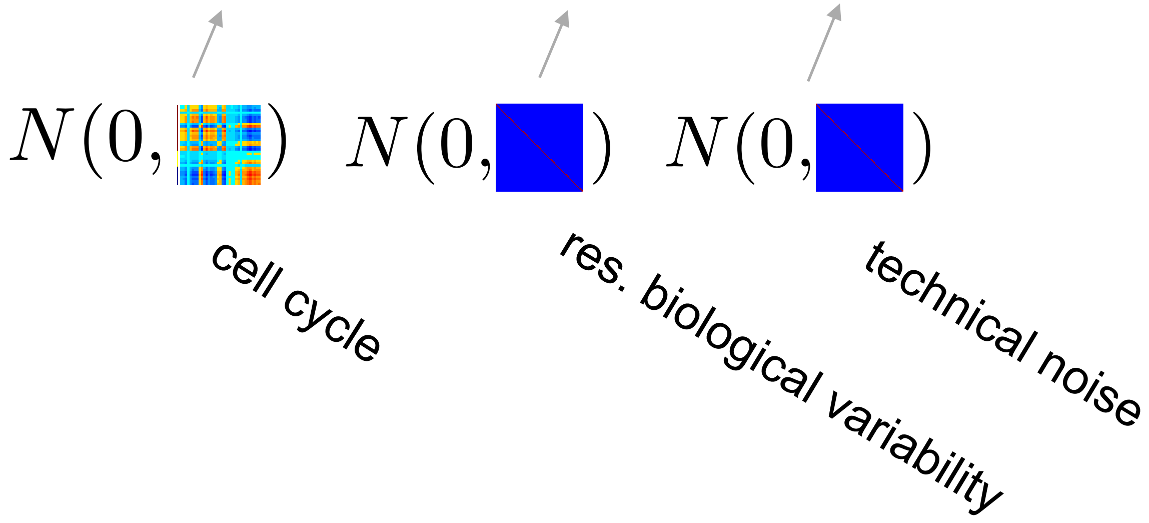
Decomposing sources of gene expression variation

- Variance decomposition of gene expression, considering
  - cell cycle (using estimated covariance)
  - residual biological variability
  - technical noise (estimated via spike-ins)

Decomposing sources of gene expression variation

- Variance decomposition of gene expression, considering
  - cell cycle (using estimated covariance)
  - residual biological variability
  - technical noise (estimated via spike-ins)

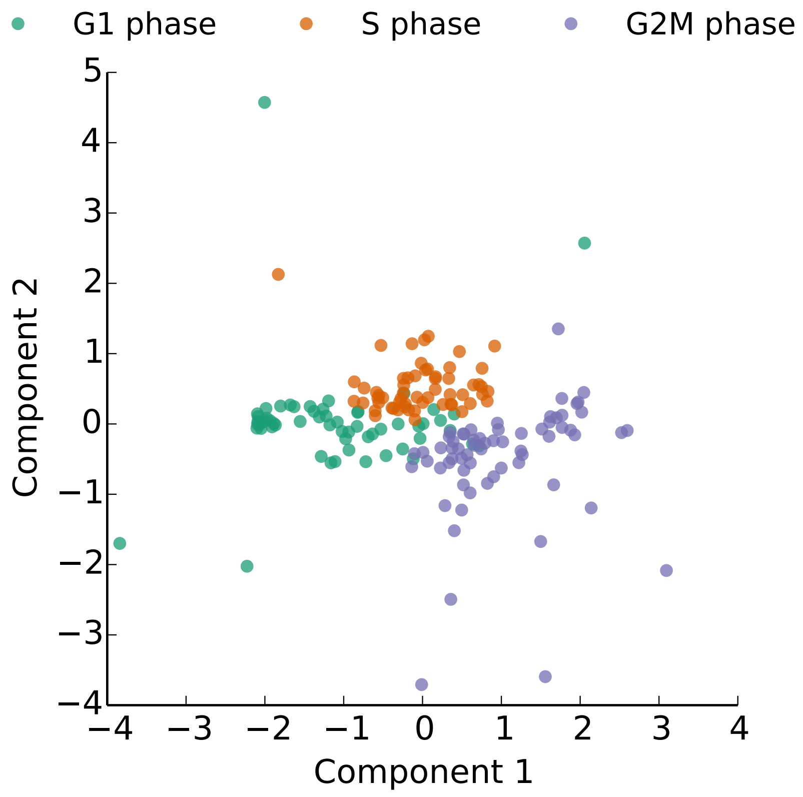
$$\mathbf{Y}_g = \mu\mathbf{I} + \alpha\mathbf{u}_{\mathrm{cc}} + \delta_b\mathbf{u}_{\mathrm{b}} + \mathbf{u}_{\mathrm{n}}$$



Model validation on mouse ESCs

- To test our model, we used single-cell RNA-Seq data generated from ~300 ES cells collected at different stages of the cell cycle

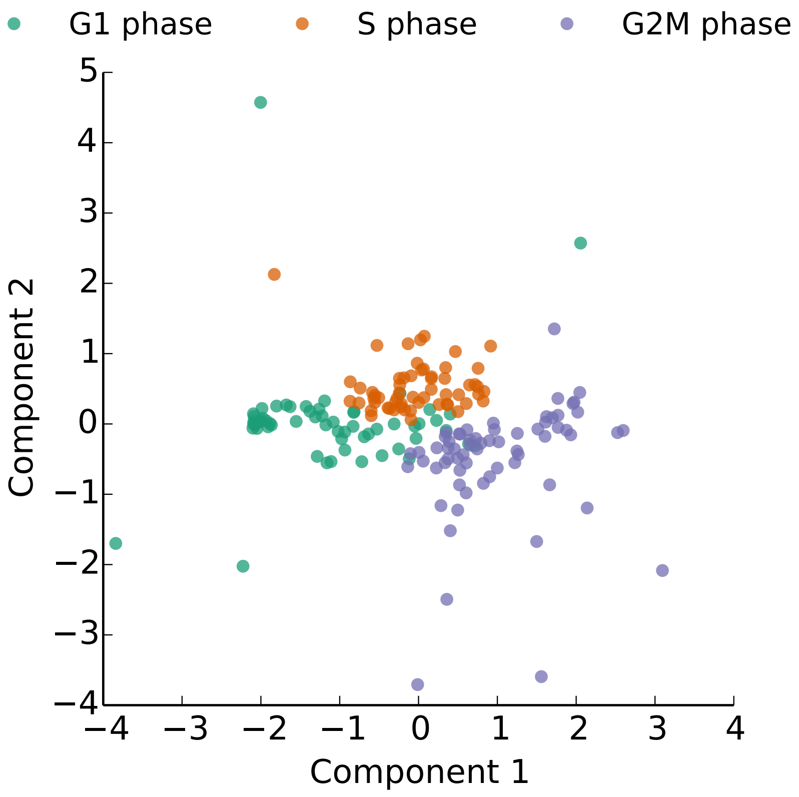
PCA on expression of staged cells



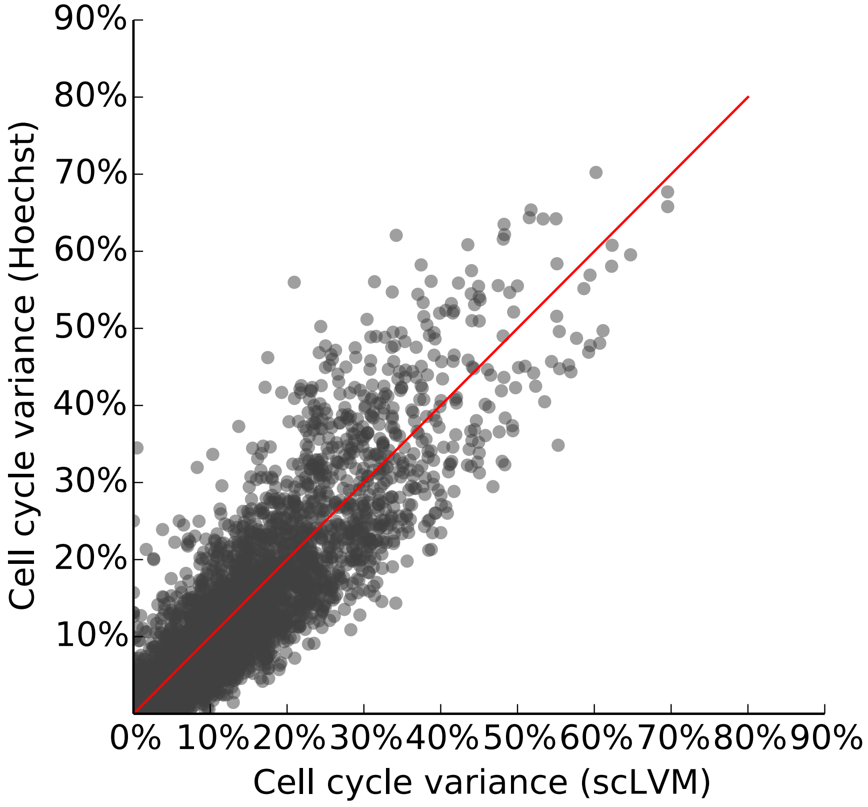
Model validation on mouse ESCs

- To test our model, we used single-cell RNA-Seq data generated from ~300 ES cells collected at different stages of the cell cycle

PCA on expression of staged cells



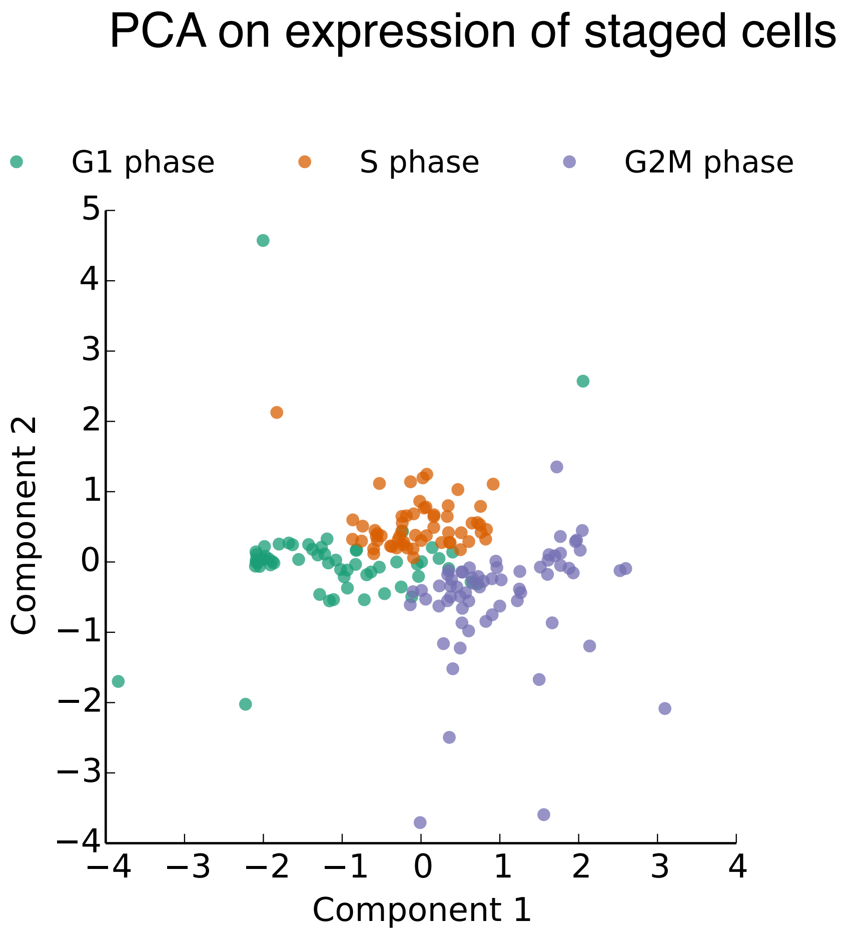
Model estimates versus Hoechst staining

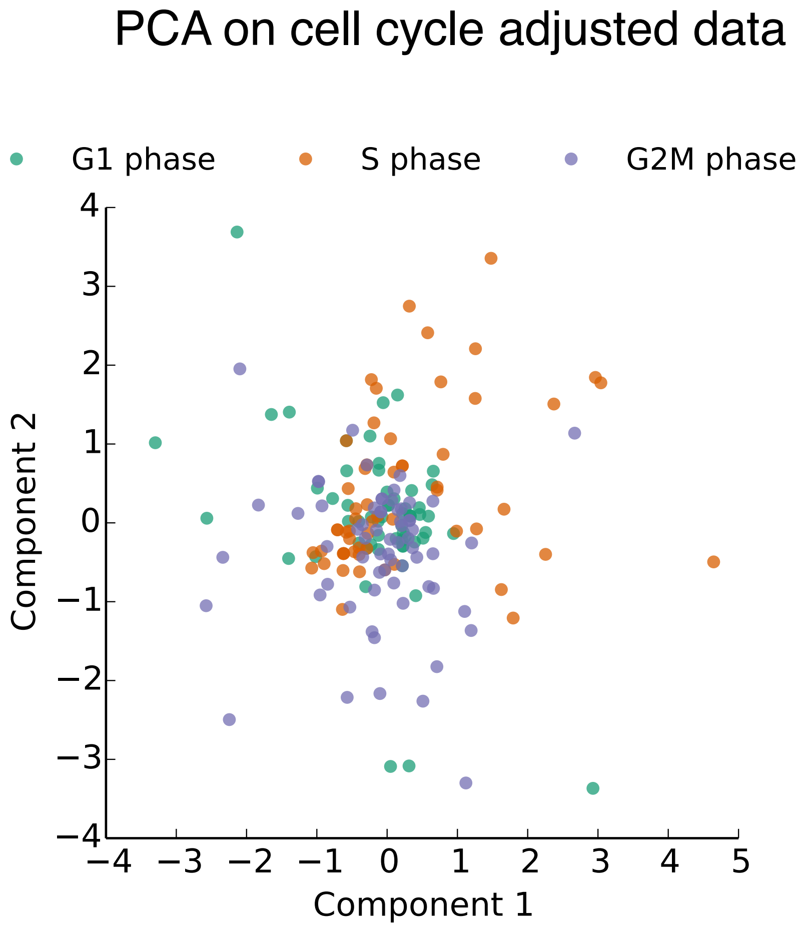


- scLVM accurately estimates variability due to the cell cycle.

Model validation on mouse ESCs

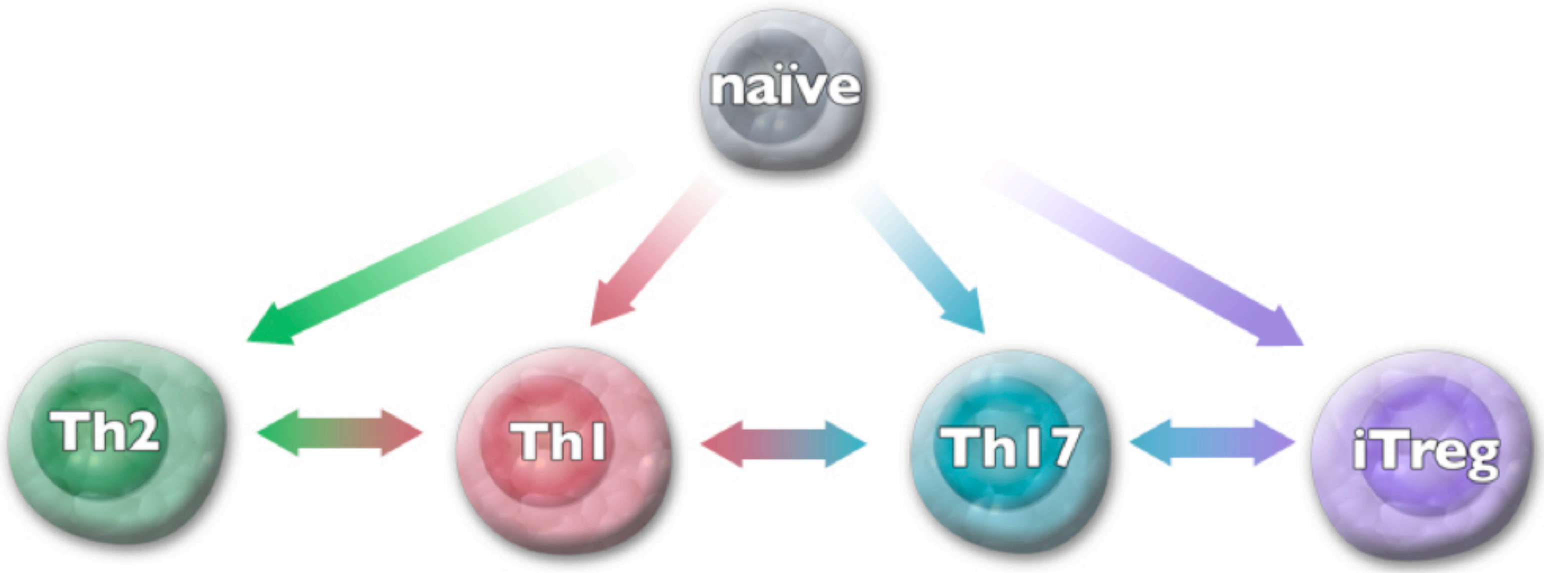
- To test our model, we used single-cell RNA-Seq data generated from ~300 ES cells collected at different stages of the cell cycle





- scLVM accurately estimates variability due to the cell cycle.
- Cell cycle effects are not visible on the model residuals.

Application to T-cell differentiation

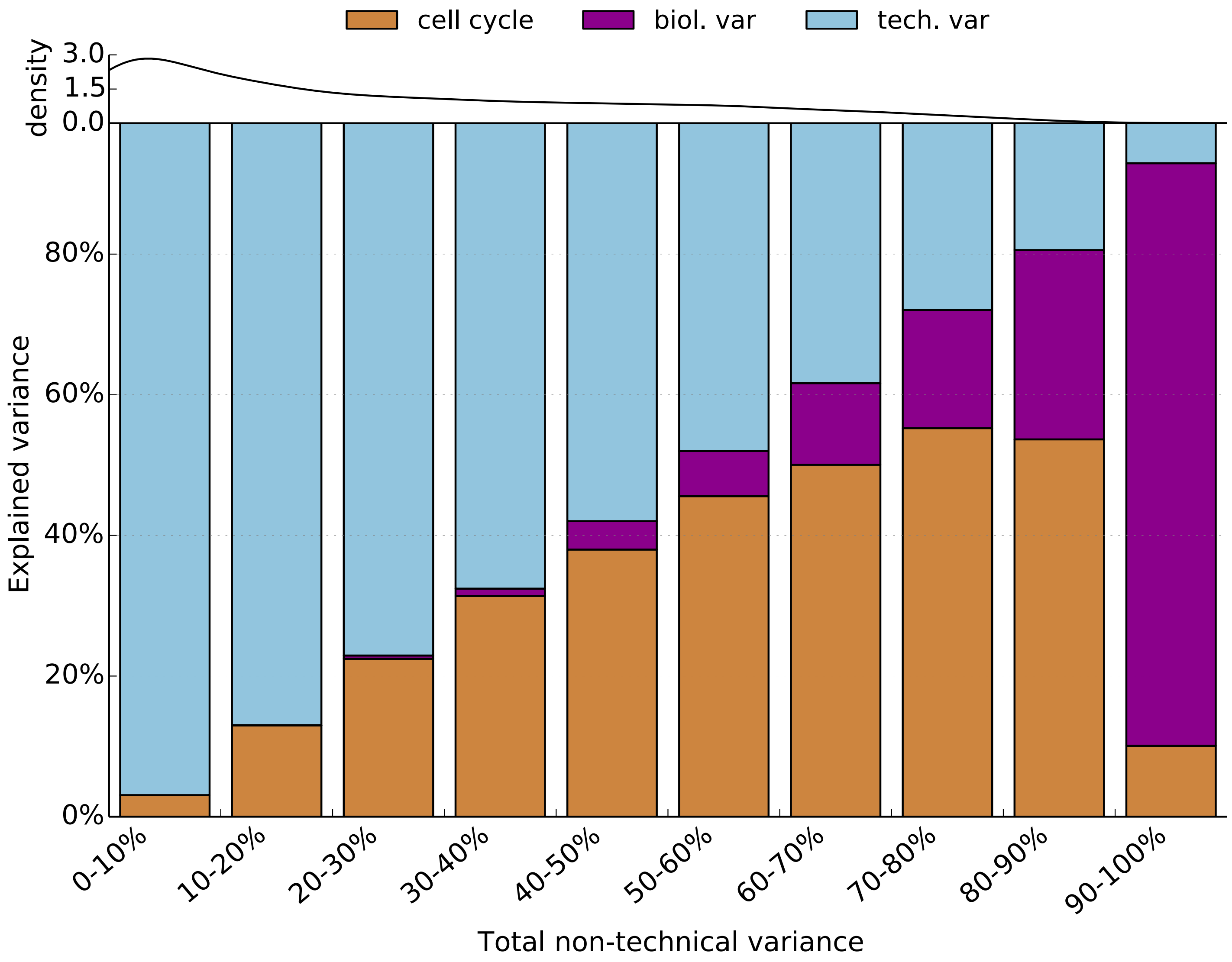


IL4, IL13….	**IFNg….**	**IL17….**	**IL10,TGF-b1…**
Helminth, Allergy	Bacteria, Viruses,	Bacteria, Fungi	Suppression, Tolerance

- Focus on cells being differentiated in vitro from the naïve state towards the Th2 cell type
- 96 cells transcription profiled using the Fluidigm C1 system

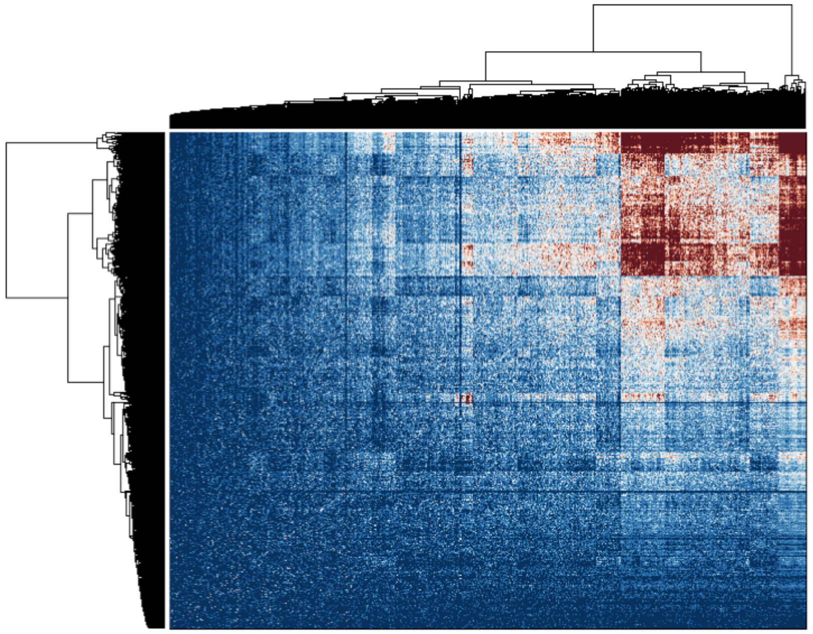
Dissecting the sources of transcriptional variation

- **Technical noise**
  For 27% of the genes, variation of expression can be entirely explained by the (technical) null variability.
- **Cell-cycle**
  For 42% of the genes, >30% of the observed variance is explained by the cell cycle state.



The impact of cell cycle on gene-gene correlations

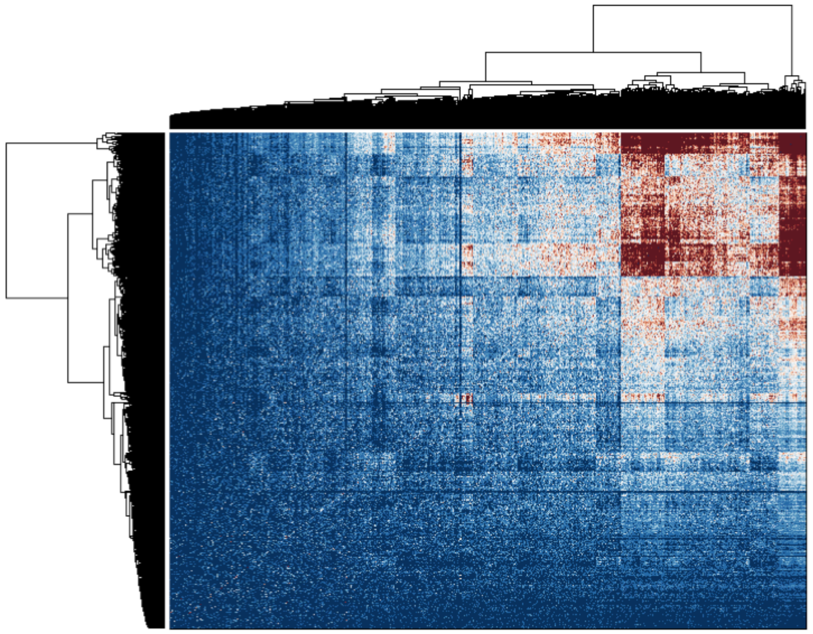
Gene-gene correlations (unadjusted)



> 500,000 edges

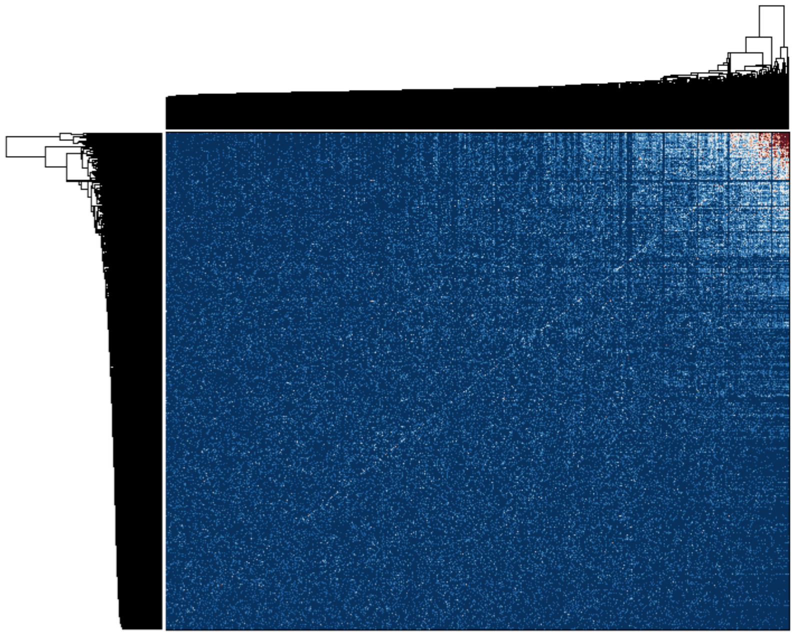
The impact of cell cycle on gene-gene correlations

Gene-gene correlations (unadjusted)



> 500,000 edges

Gene-gene correlations (adjusted for cell cycle)

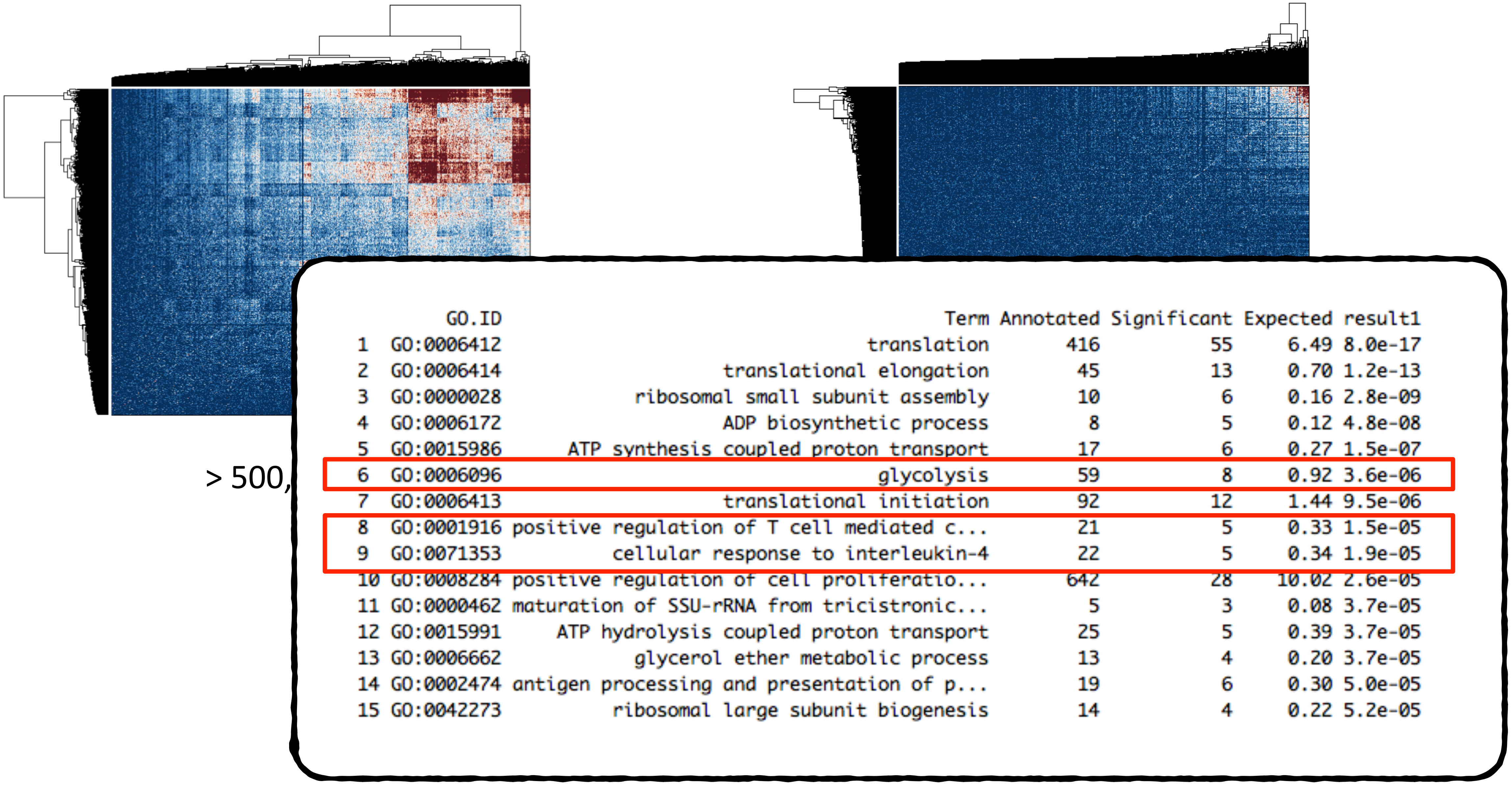


~ 20,000 edges

The impact of cell cycle on gene-gene correlations

Gene-gene correlations (unadjusted)

Gene-gene correlations (adjusted for cell cycle)

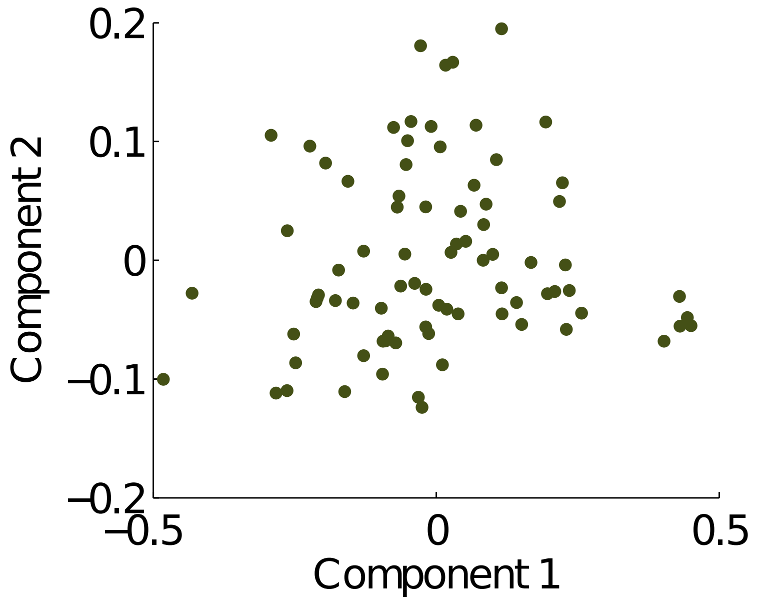


> 500,

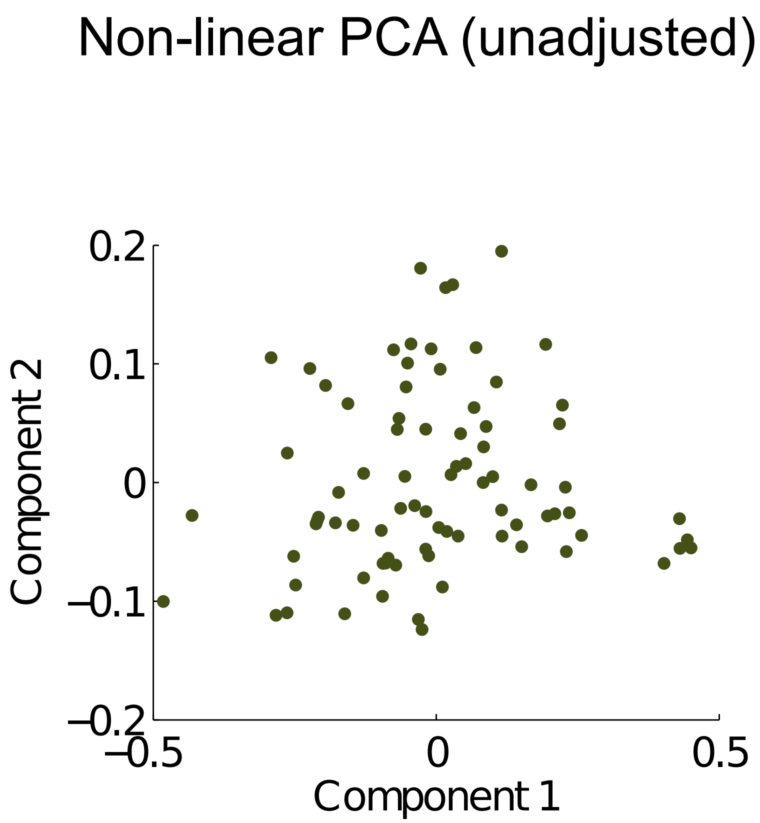
```
        GO.ID                                        Term Annotated Significant Expected result1
1  GO:0006412                                 translation       416          55     6.49 8.0e-17
2  GO:0006414                     translational elongation        45          13     0.70 1.2e-13
3  GO:0000028                 ribosomal small subunit assembly     10           6     0.16 2.8e-09
4  GO:0006172                     ADP biosynthetic process         8           5     0.12 4.8e-08
5  GO:0015986        ATP synthesis coupled proton transport       17           6     0.27 1.5e-07
6  GO:0006096                                  glycolysis        59           8     0.92 3.6e-06
7  GO:0006413                     translational initiation        92          12     1.44 9.5e-06
8  GO:0001916 positive regulation of T cell mediated c...        21           5     0.33 1.5e-05
9  GO:0071353            cellular response to interleukin-4       22           5     0.34 1.9e-05
10 GO:0008284 positive regulation of cell proliferatio...       642          28    10.02 2.6e-05
11 GO:0000462 maturation of SSU-rRNA from tricistronic...         5           3     0.08 3.7e-05
12 GO:0015991         ATP hydrolysis coupled proton transport     25           5     0.39 3.7e-05
13 GO:0006662               glycerol ether metabolic process      13           4     0.20 3.7e-05
14 GO:0002474 antigen processing and presentation of p...        19           6     0.30 5.0e-05
15 GO:0042273             ribosomal large subunit biogenesis      14           4     0.22 5.2e-05
```

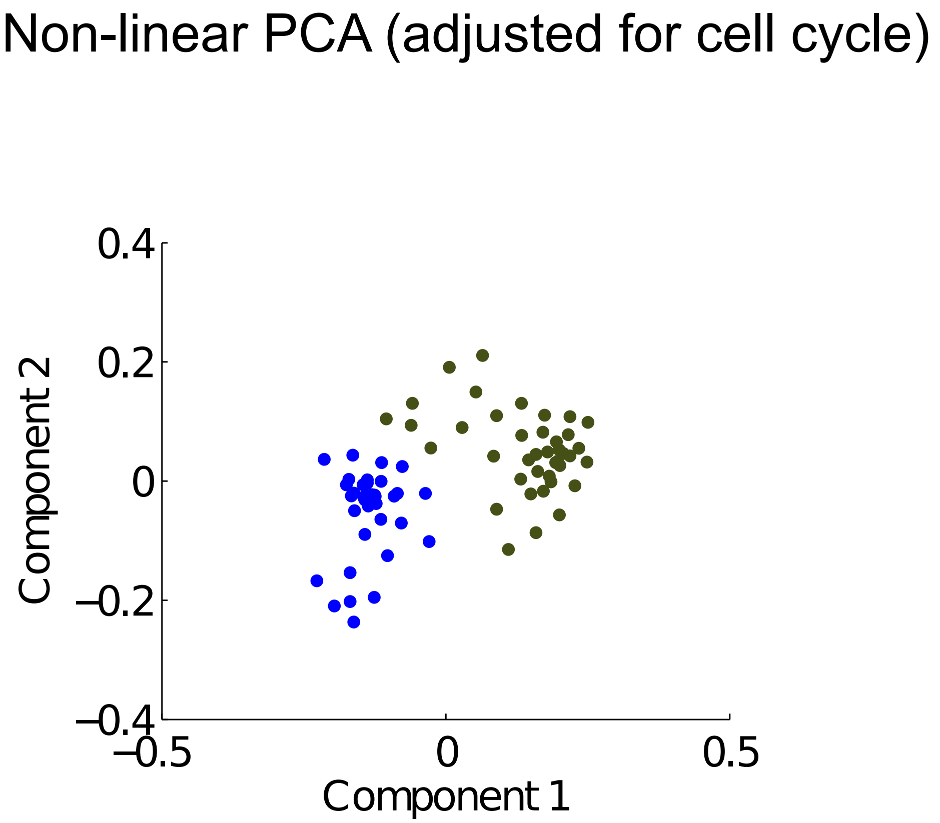
Discrimination between differentiated and undifferentiated cells

Non-linear PCA (unadjusted)

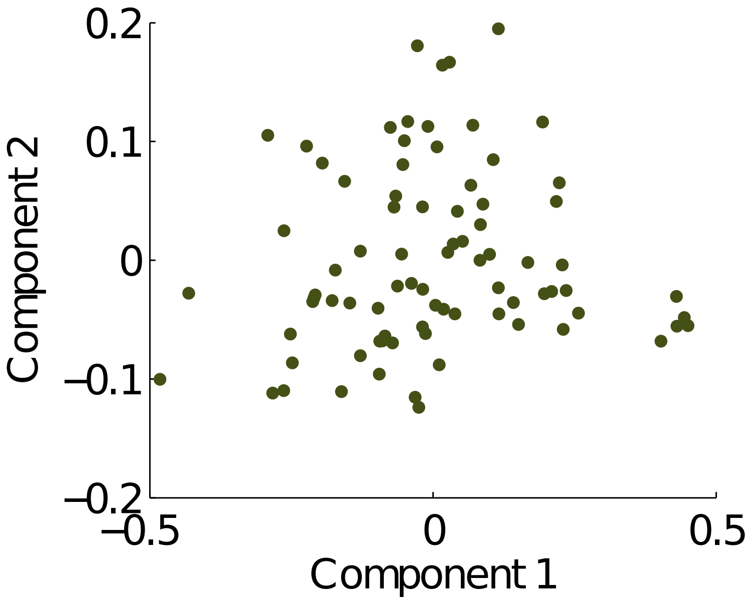


Discrimination between differentiated and undifferentiated cells

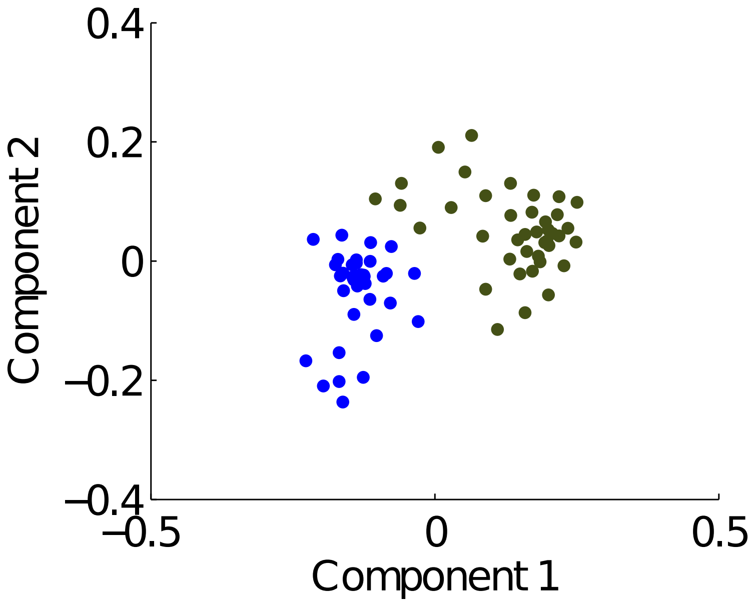




Discrimination between differentiated and undifferentiated cells



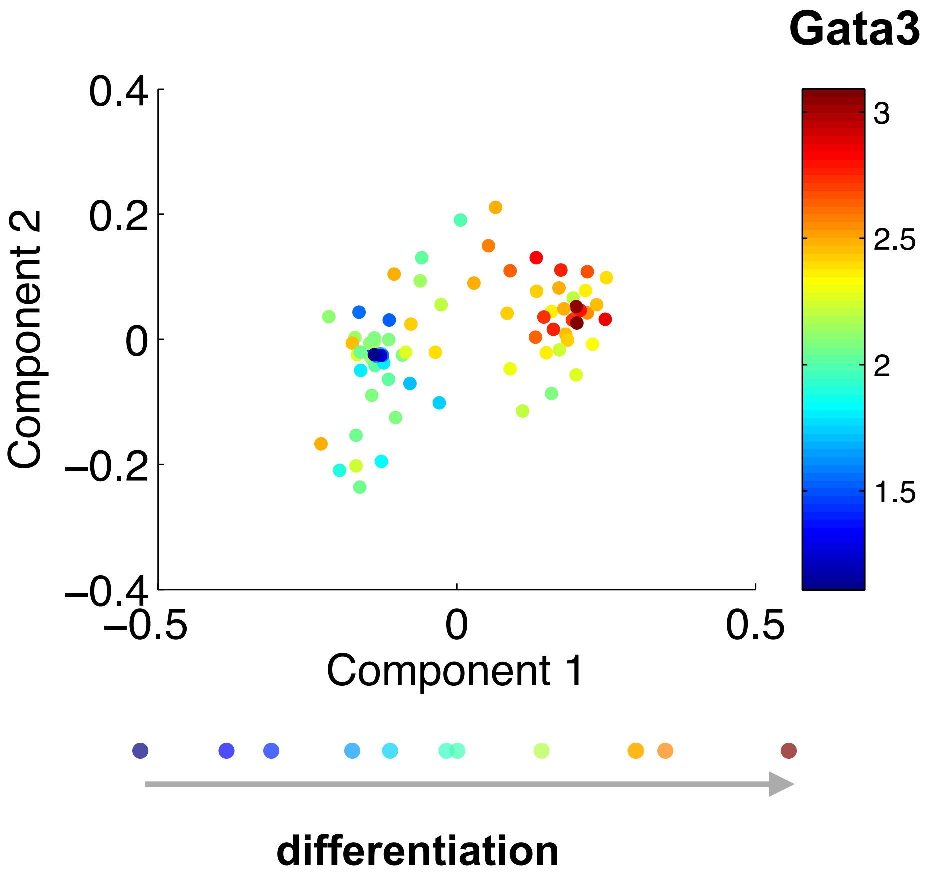
Non-linear PCA (unadjusted)



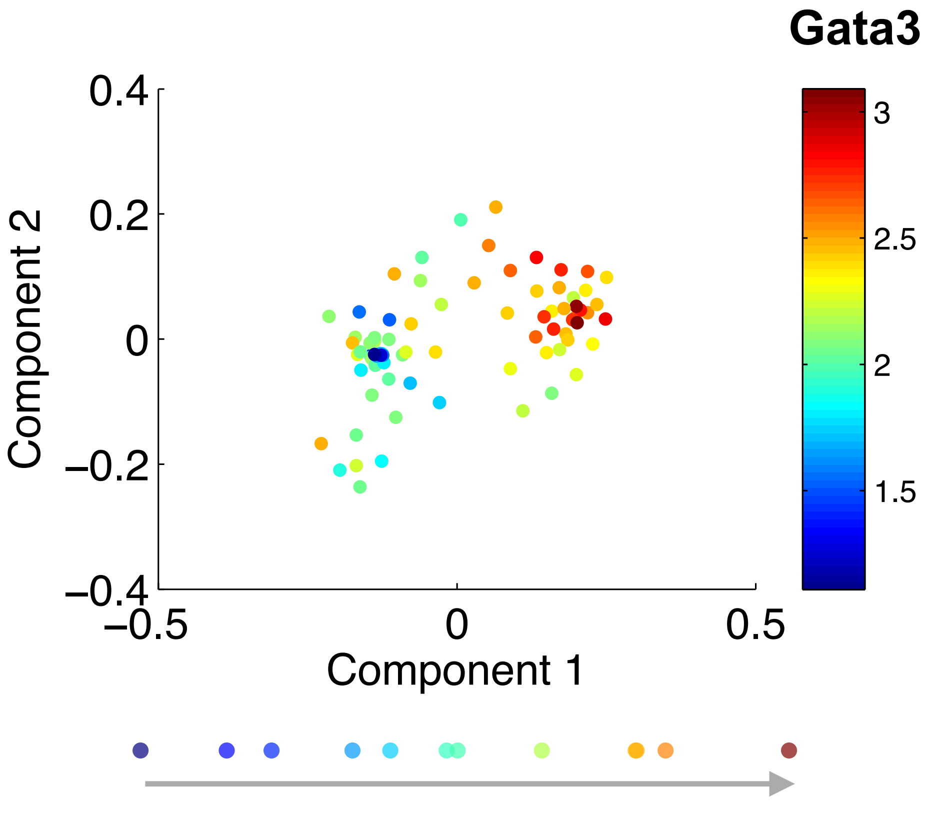
Non-linear PCA (adjusted for cell cycle)

- After cell cycle correction, cells appear to separate better into two groups than without correction.

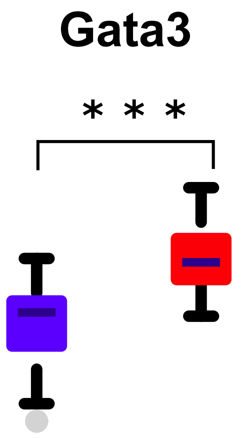
Are the identified subpopulations meaningful?



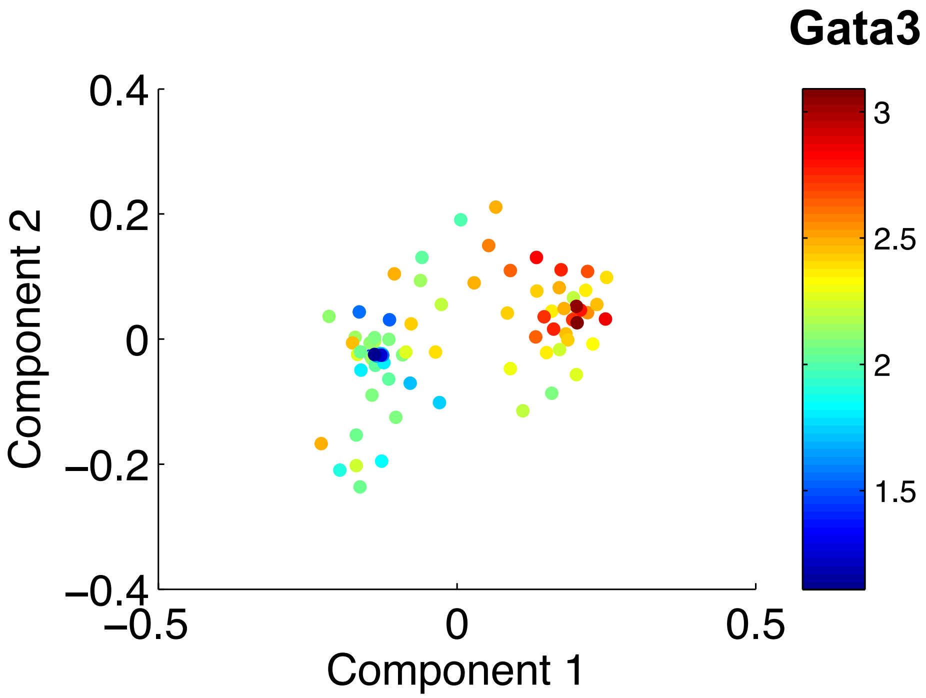
Are the identified subpopulations meaningful?



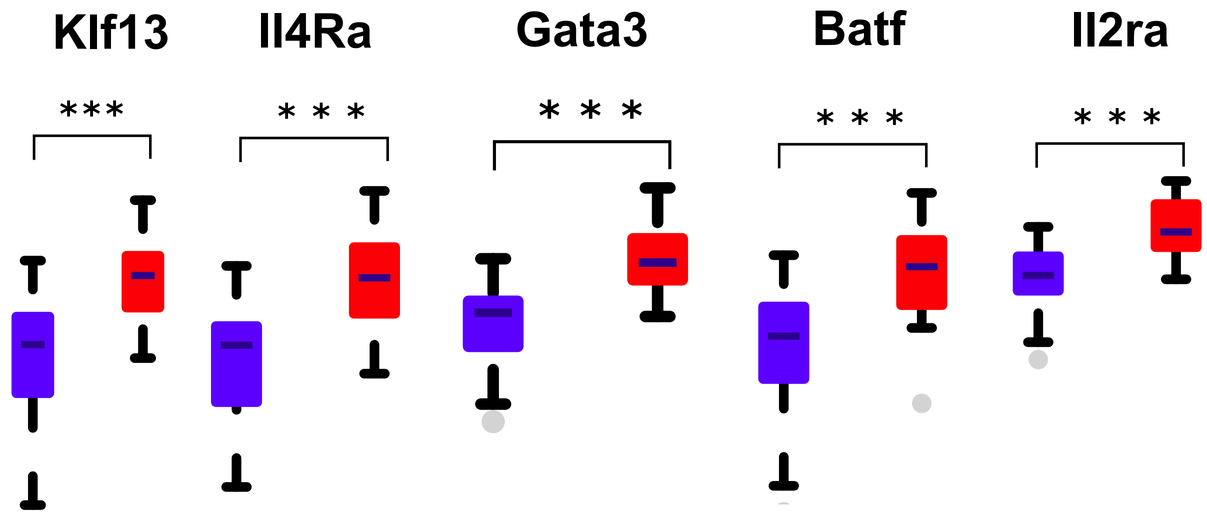
differentiation



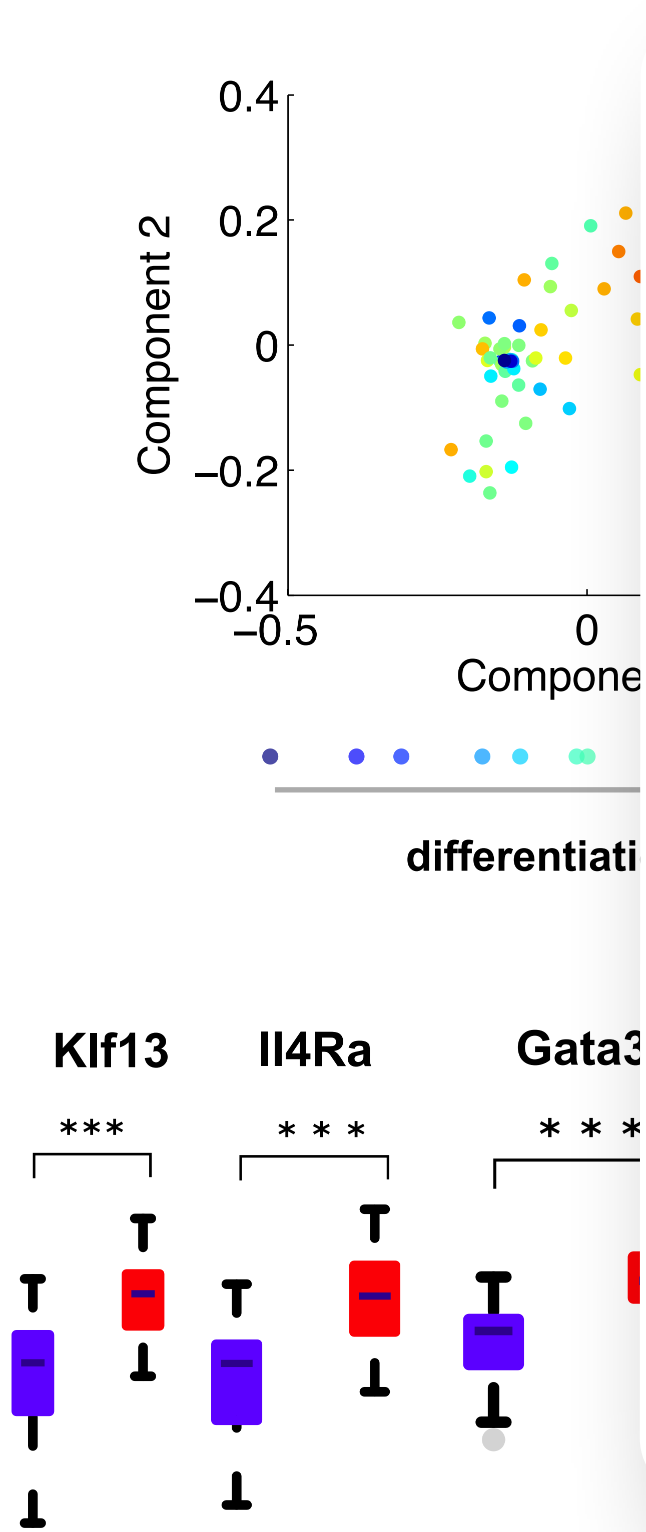
Are the identified subpopulations meaningful?



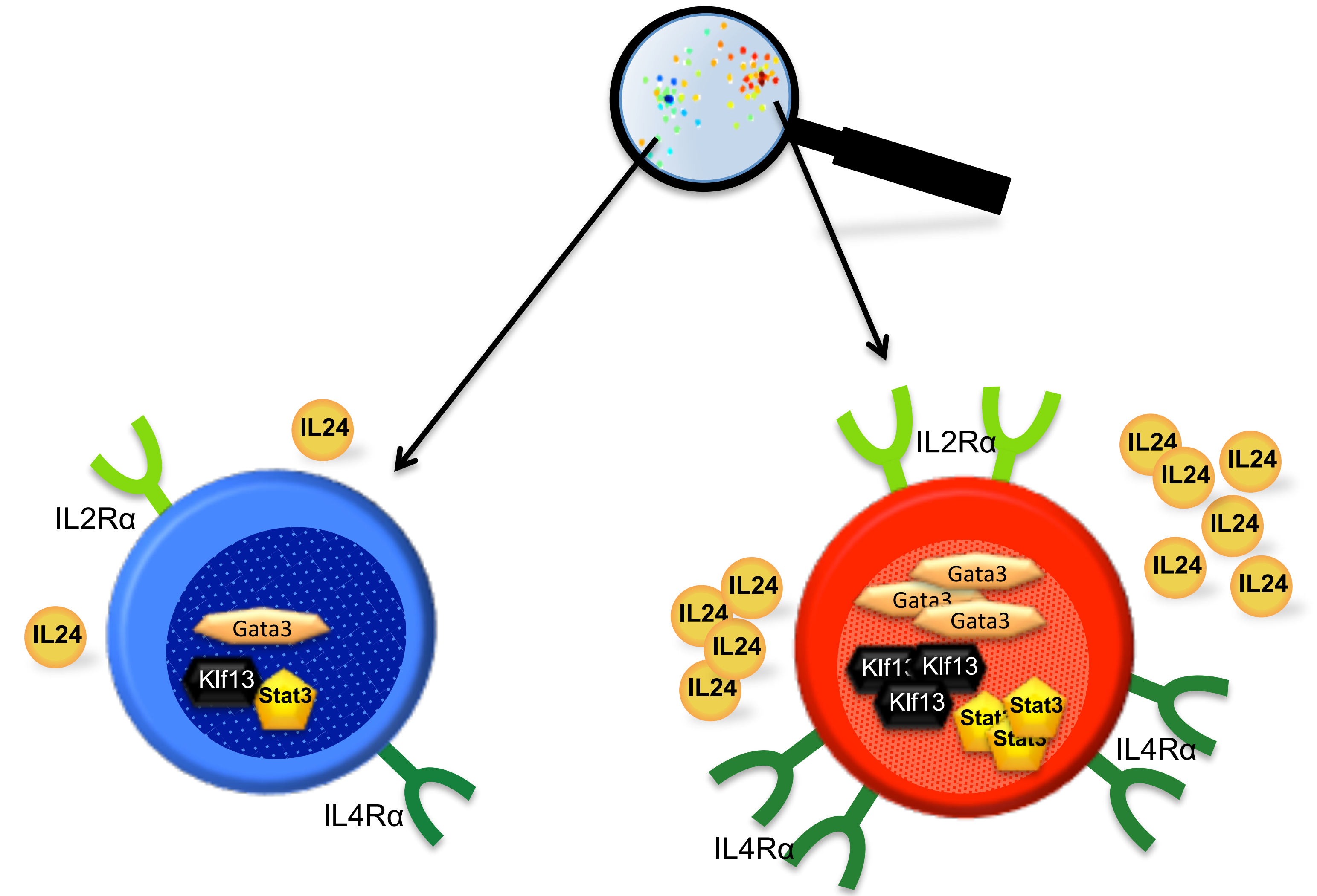
differentiation



Are the identified subpopulations meaningful?

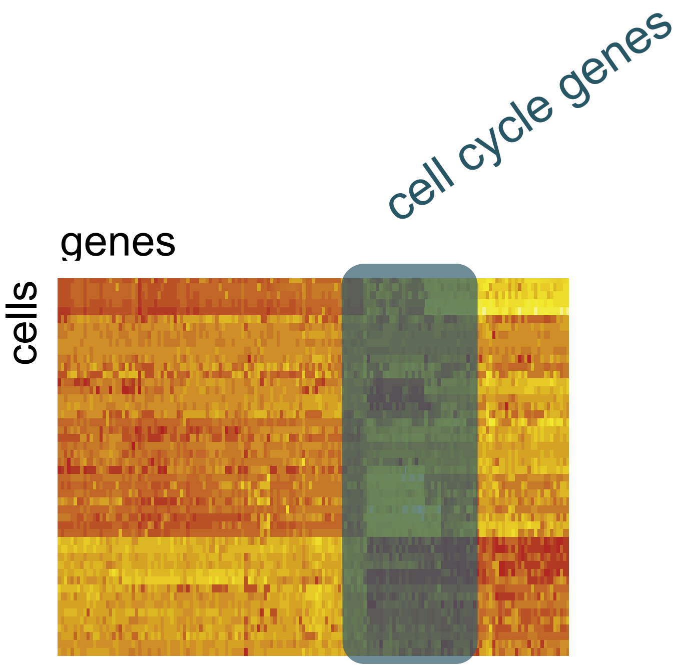


- 401 Genes differentially expressed
- Strikingly enriched in Th2 markers



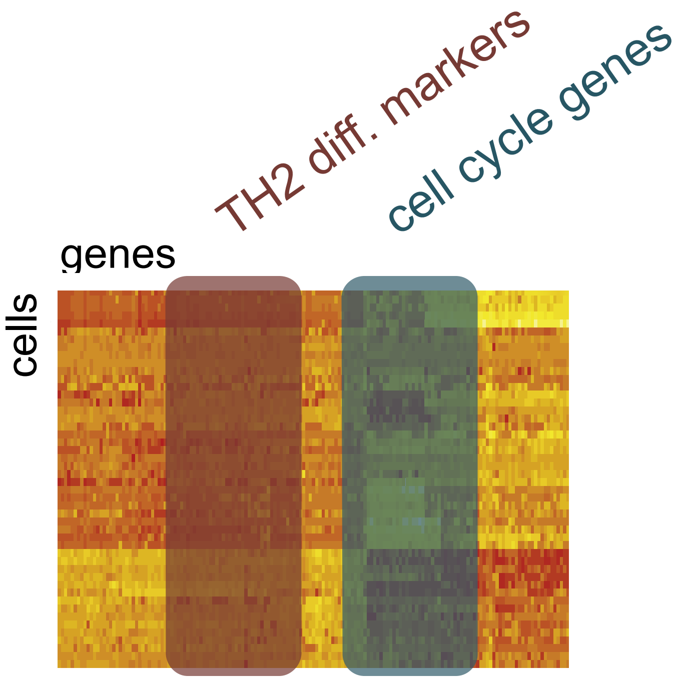
Can we better tease apart the effect of cell cycle and differentiation ?

- scLVM also enables learning multiple latent factors
  - Genes annotated for cell cycle



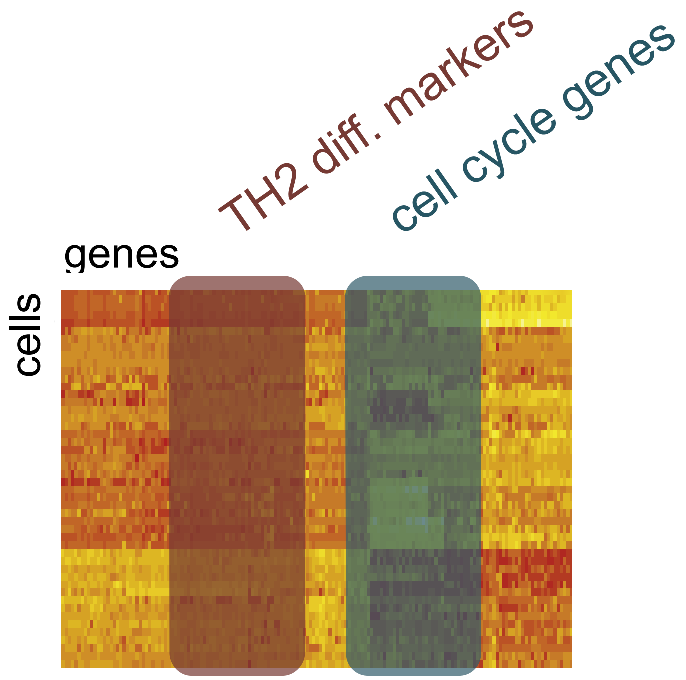
Can we better tease apart the effect of cell cycle and differentiation ?

- scLVM also enables learning multiple latent factors
  - Genes annotated for cell cycle
  - Th2 differentiation marker genes



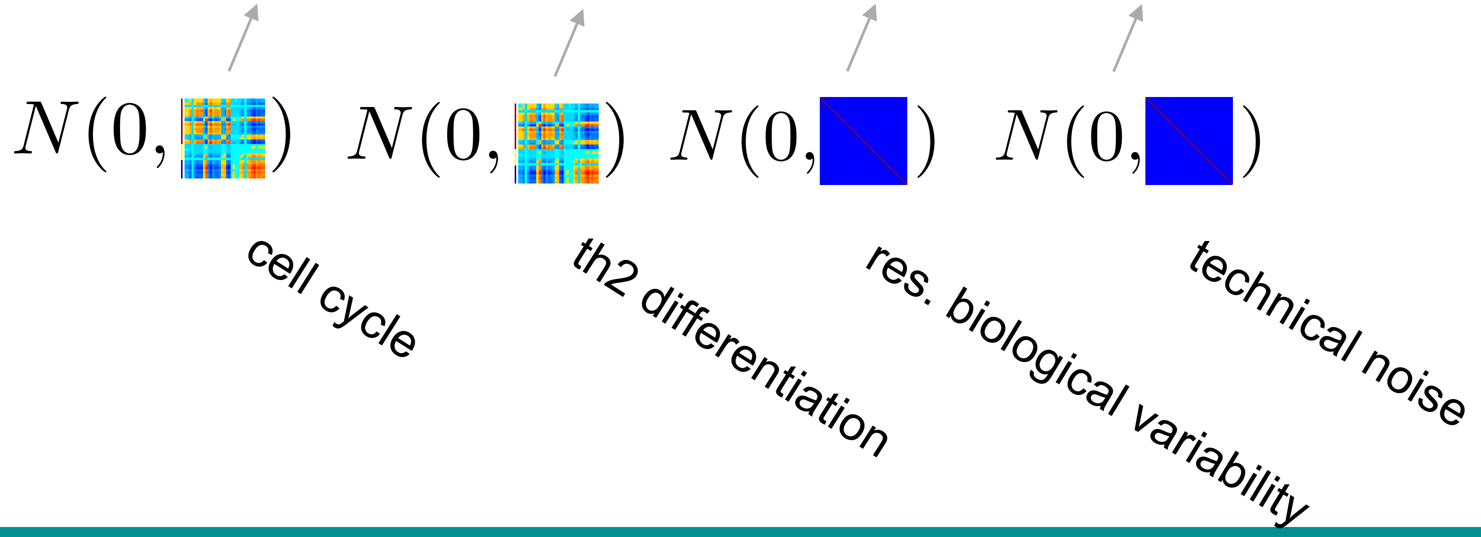
Can we better tease apart the effect of cell cycle and differentiation ?

- scLVM also enables learning multiple latent factors
  - Genes annotated for cell cycle
  - Th2 differentiation marker genes



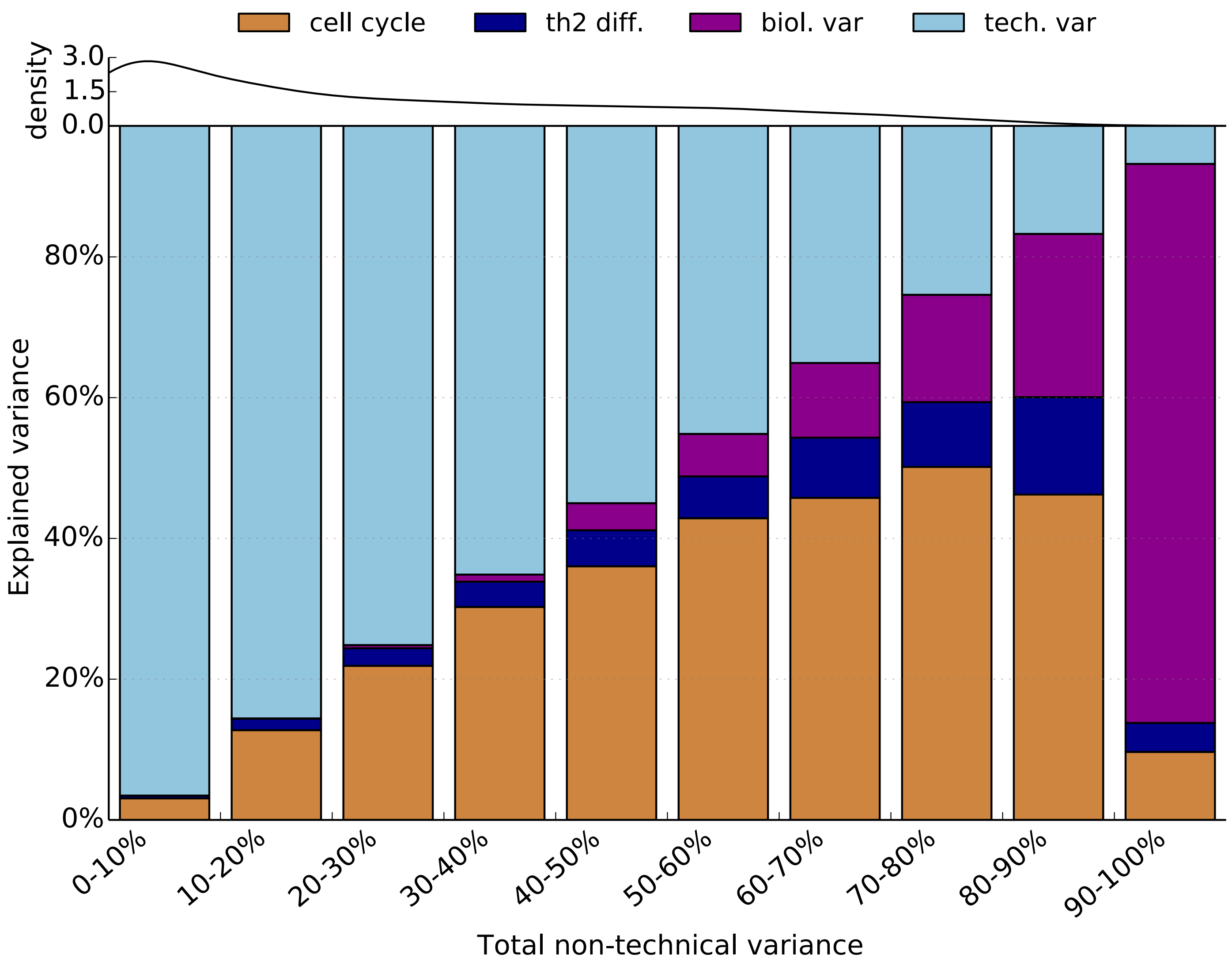
- Extended variance component analysis

$$\mathbf{Y}_g = \mu\mathbf{I} + \alpha\mathbf{u}_{\mathrm{cc}} + \beta\mathbf{u}_{\mathrm{th2}} + \delta_b\mathbf{u}_{\mathrm{b}} + \mathbf{u}_{\mathrm{n}}$$



Can we better tease apart the effect of cell cycle and differentiation ?

- **Th2 differentiation**
  928 genes with affected by the Th2 differentiation factor

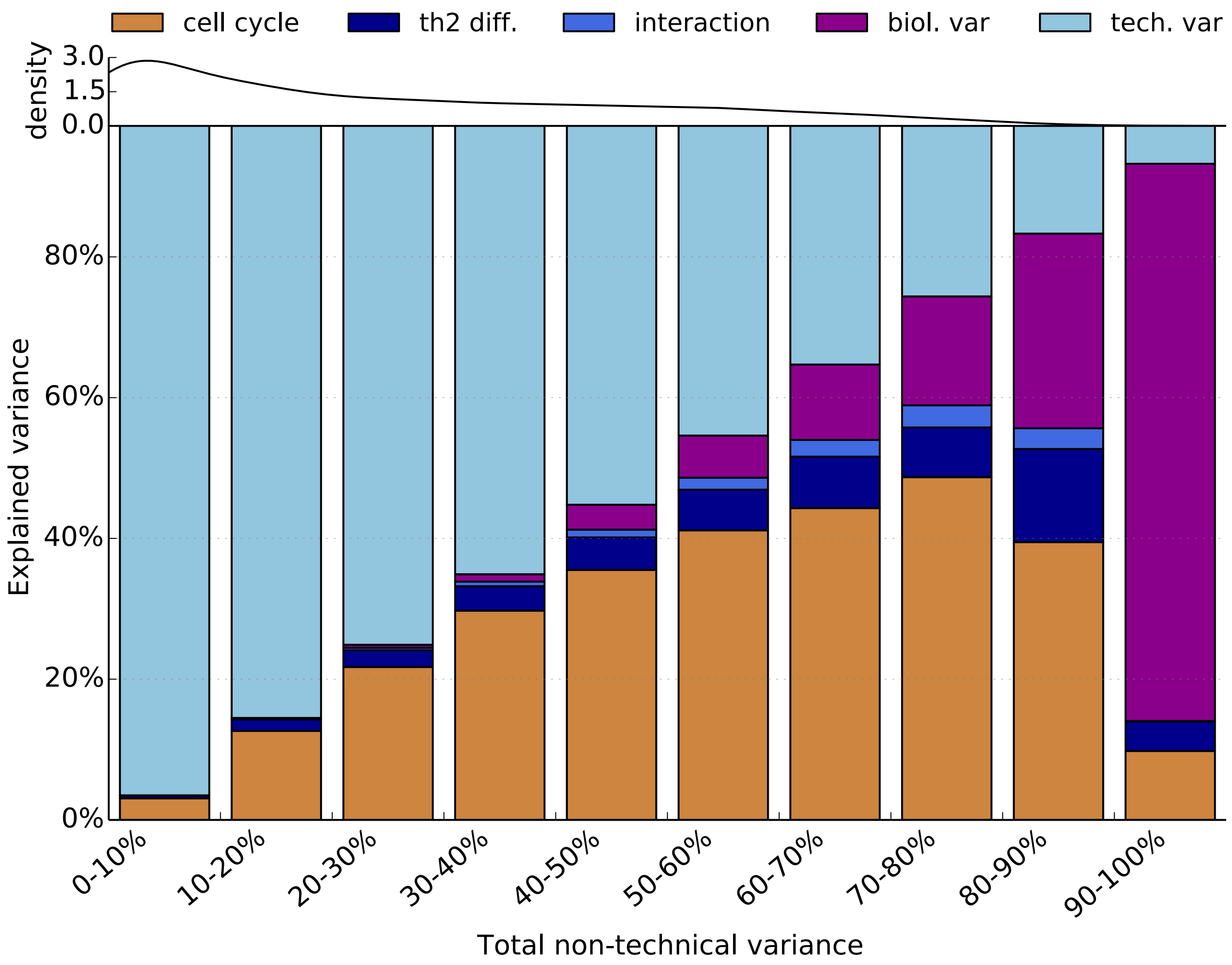


Can we better tease apart the effect of cell cycle and differentiation ?

- **Th2 differentiation**
  928 genes with affected by the Th2 differentiation factor

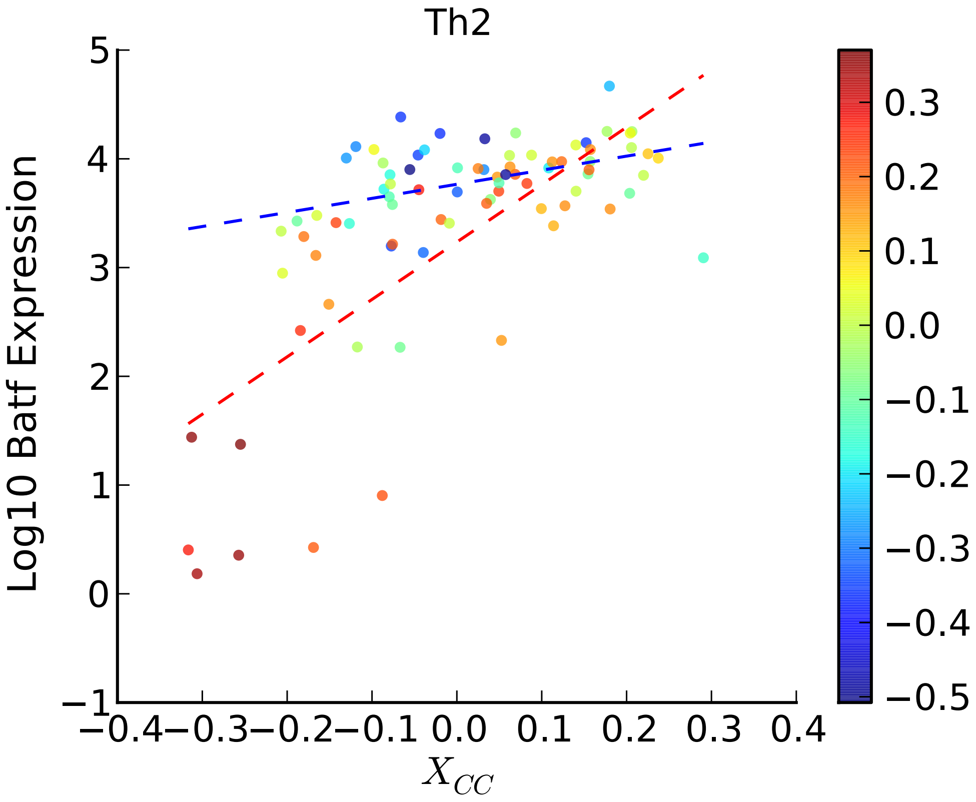
- **Th2/cell-cycle interaction**
  200 genes with interaction effects

- Enriched for
  positive cell proliferation
  negative regulation of apoptosis



Can we better tease apart the effect of cell cycle and differentiation ?

- **Th2 differentiation**
  928 genes with affected by the Th2 differentiation factor
- **Th2/cell-cycle interaction**
  200 genes with interaction effects
- Enriched for
  positive cell proliferation
  negative regulation of apoptosis

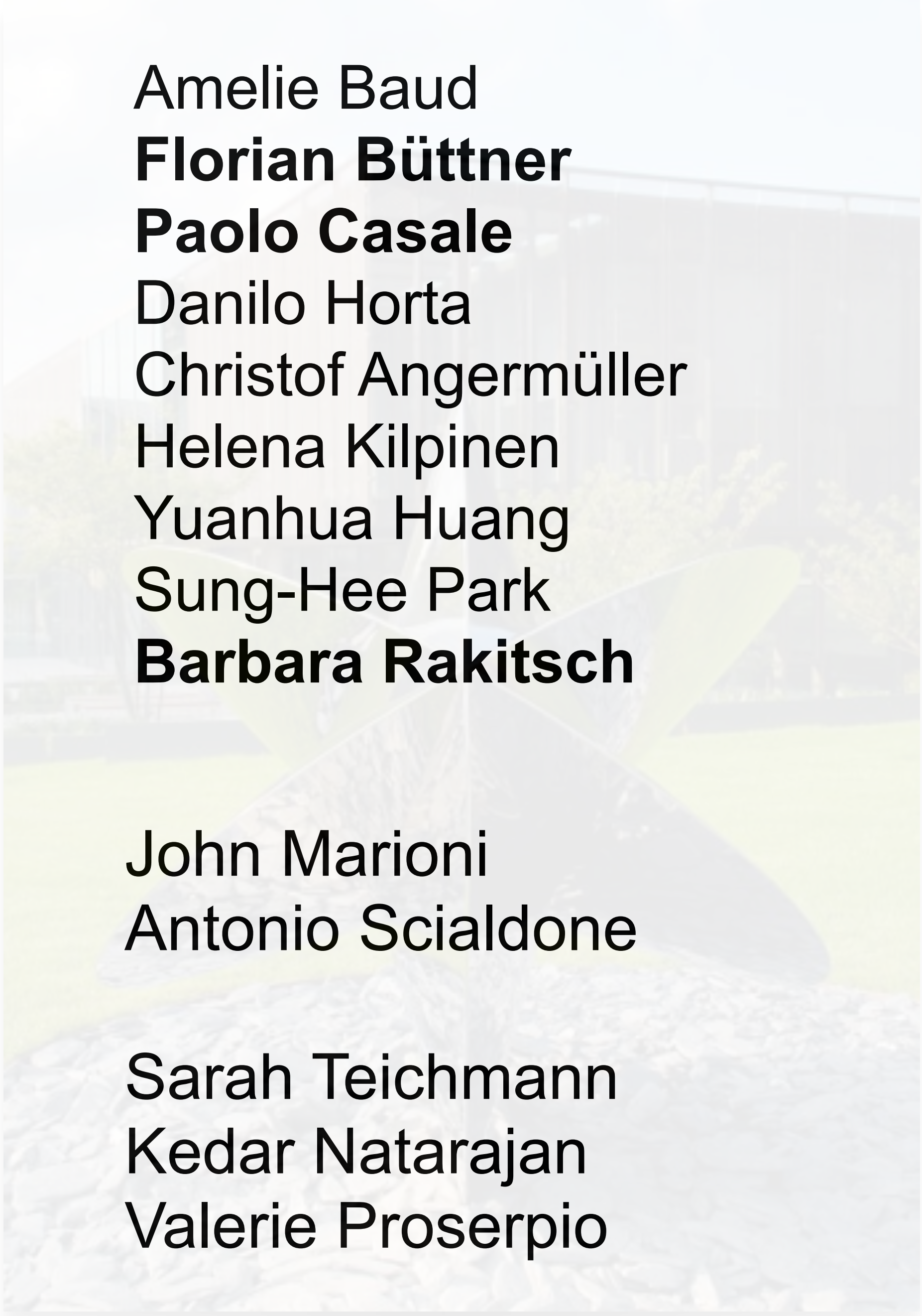


Closing comments

- Random effect covariance models can be flexibly applied to account for different levels of sample heterogeneity
- (e)QTL analysis
  - population structure & env. /technical confounding to improve power and accuracy

- Single-cell RNA-seq analysis
  - a small number of genes with known cell cycle annotation is sufficient to estimate a cell covariance due to cell cycle
  - more compact gene-gene correlations
  - detection of genes with interactions involving multiple biological processes

Acknowledgments



Amelie Baud
Florian Büttner
Paolo Casale
Danilo Horta
Christof Angermüller
Helena Kilpinen
Yuanhua Huang
Sung-Hee Park
Barbara Rakitsch

John Marioni
Antonio Scialdone

Sarah Teichmann
Kedar Natarajan
Valerie Proserpio

<u>Helmholtz Munich</u>
Fabian Theis

<u>University of Shefifeld</u>
Neil Lawrence
Nicolo Fusi

<u>Microsoft Resarch</u>
Nicolo Fusi

<u>Human Longevity INC</u>
Christoph Lippert

<u>Sanger</u>
Thierry Voet
Iain Macaulay

<u>Babraham Inst.</u>
Heather Lee
Stephen Clark
Wolf Reik
Gavin Kelsey

Mixed model inference: https://github.com/PMBio/limix

Single cell latent variable model: https://github.com/PMBio/scLVM

Tutorial pointers

https://github.com/PMBio/limix-tutorials

https://github.com/PMBio/scLVM/tree/master/R/tutorials

Tutorial pointers

- Practical to use LIMIX for genetic analyses:

  https://github.com/PMBio/limix-tutorials

- R version of scLVM, recommended for R users:

  https://github.com/PMBio/scLVM/tree/master/R/tutorials

Tutorial pointers

- Practical to use LIMIX for genetic analyses:

  https://github.com/PMBio/limix-tutorials

- R version of scLVM, recommended for R users:

  https://github.com/PMBio/scLVM/tree/master/R/tutorials

- scLVM python module & ipython notebooks with an example that interfaces to GPy.

  https://github.com/PMBio/scLVM/blob/master/tutorials/tcell_demo.ipynb

```
> git clone git@github.com:PMBio/scLVM.git
> cd scLVM/tutorials
> ipython notebook ./tcell_demo.ipynb
```

- GPy example on non-linear dimensionality reduction applied to single-cell RNA-Seq

  https://github.com/SheffieldML/notebook/blob/master/compbio/SingleCellDataWithGPyTutorial.ipynb