Modelling molecular heterogeneity between individuals and single cells

GCC 2015

Oliver Stegle
EMBL, European Bioinformatics Institute

Heterogeneity between individuals and single cells

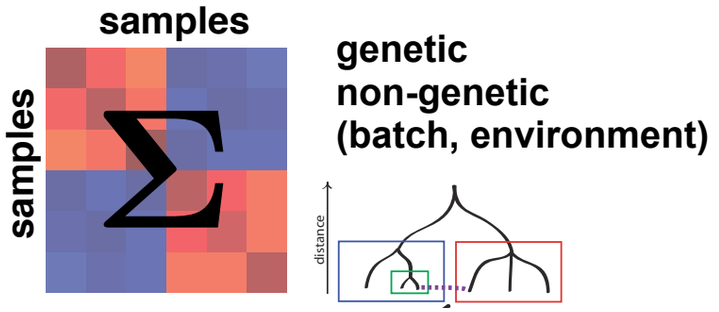
variation of interest

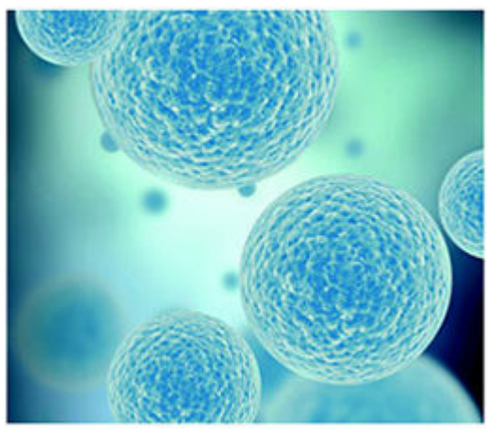
confounding

population variation

genetic associations with phenotype

sample covariance

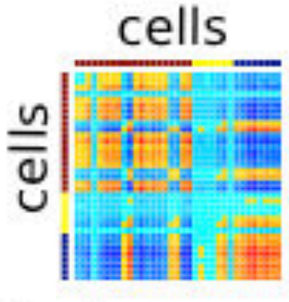




single-cell variation

differentiation processes

cell covariance



cell cycle
stress

UK human iPS cell consortium: genotype to phenotype

iPS genomics

stem cell proteomics

- Discover how
  - genetic variation affects cellular function
  - genetic lesions lead to disease phenotypes

- In so doing, create an open access resource for the wider biomedical community:
  - HapMap-like consent for ~500 healthy normals
  - Controlled access for ~500 disease samples

www.hipsci.org

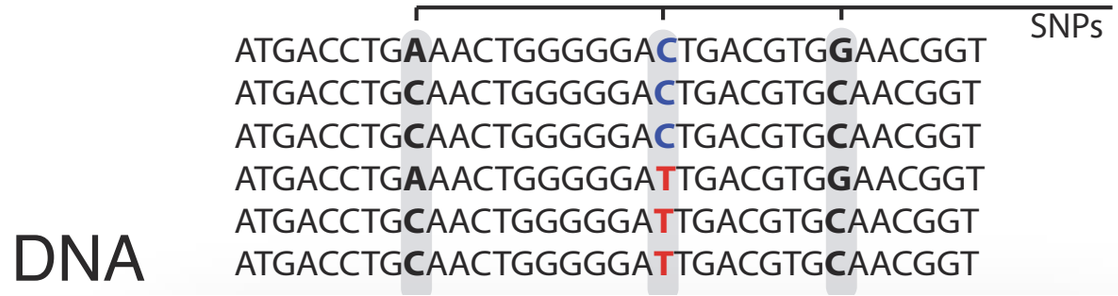
Phil Beales, Ewan Birney,Laura Clarke,
Daniel Gaffney,Angus Lamond, Richard Durbin, Fiona Watt

wellcome trust Strategic Award

MRC Medical Research Council

Multi-omics association genetics

HipSci



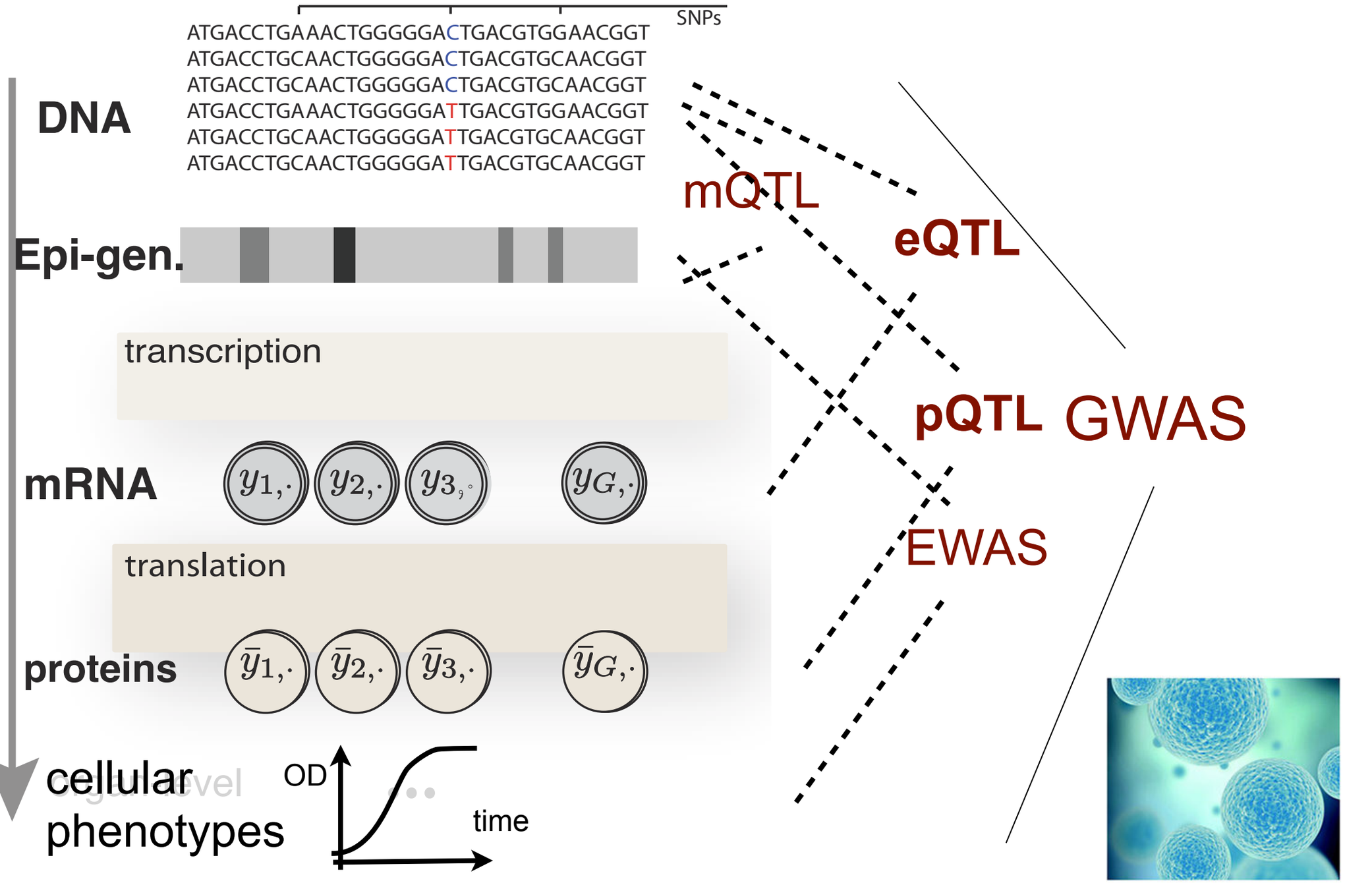
DNA

GWAS

cellular phenotypes

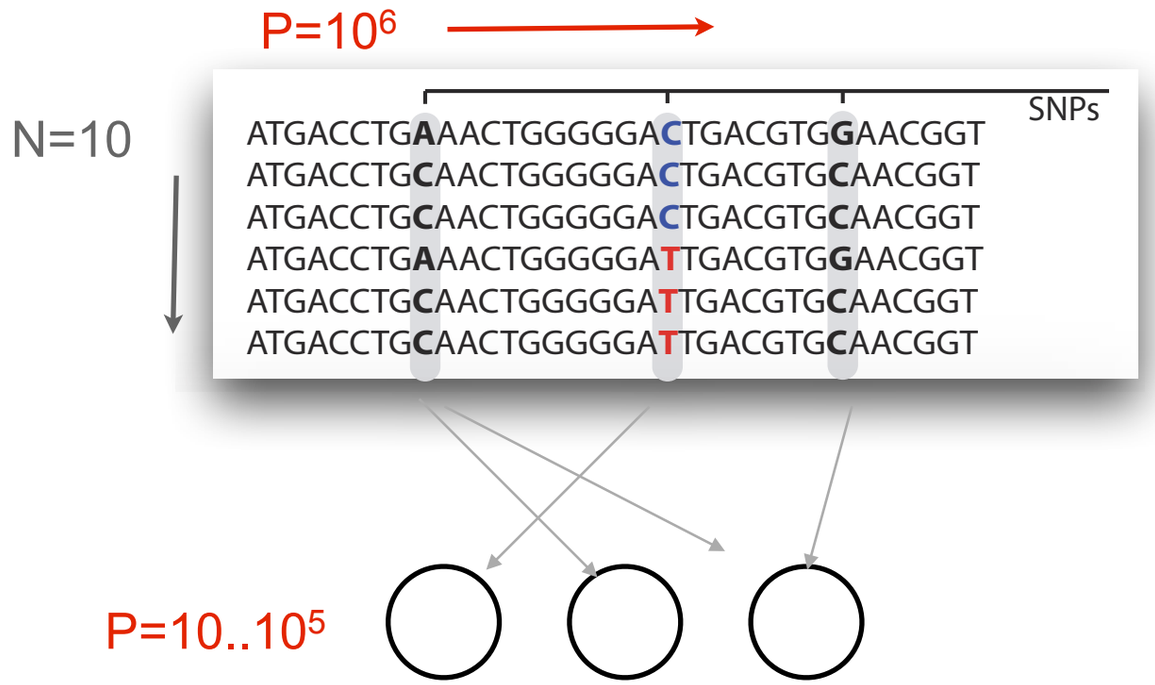
Multi-modal data integration across molecular layers

HipSci

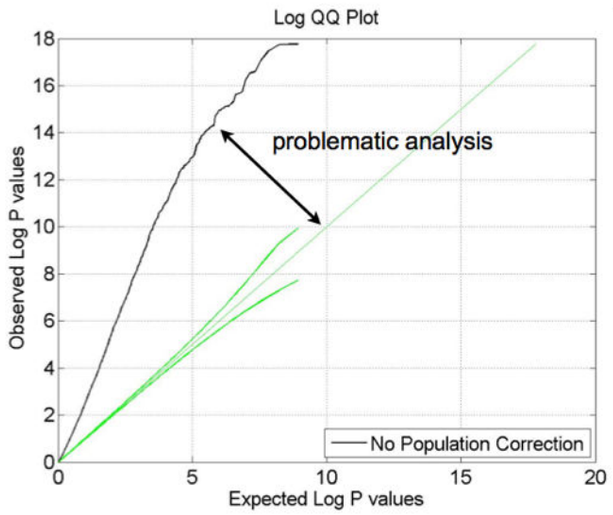


Multi-model association genetics: statistical challenges and opportunities

- **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

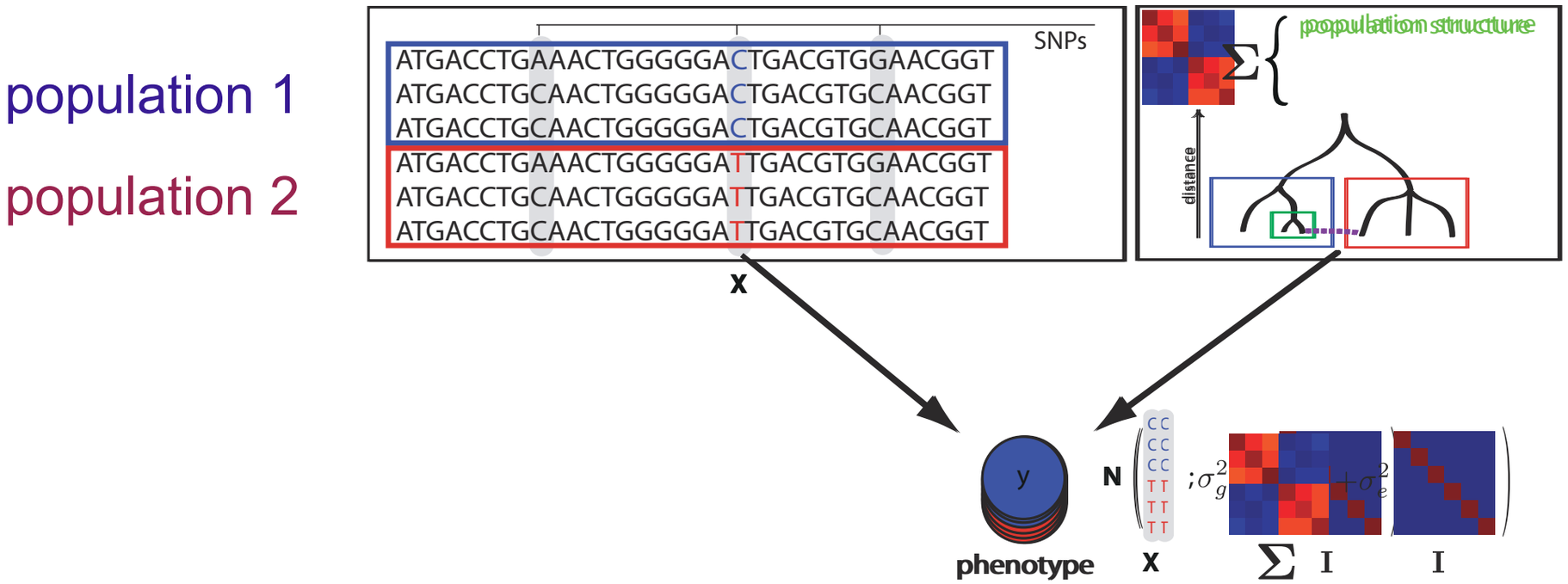


- **Win:** Large dataset allow to test modeling assumptions / fit better models
  - **Inference of confounding structures**
  - Not possible before large-scale hypothesis testing/large datasets
  - More **power due to large datasets**
  - Gain in power by joint analysis of **multiple traits**



Hidden structure: population structure

- Population structure (genetic)



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

- Estimate Σ
  - Population structure: genotype data

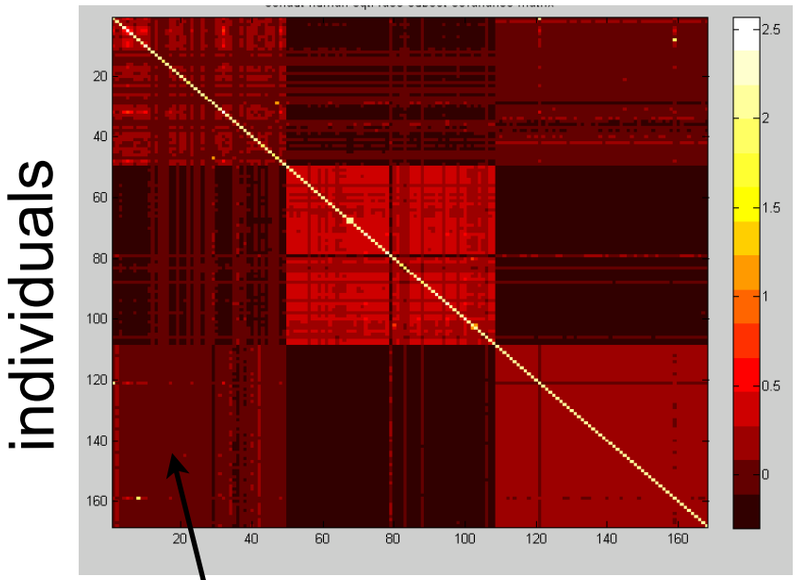
Kang et al., 2008, 2010
Lippert et al. 2011
Zhou et al. 2012

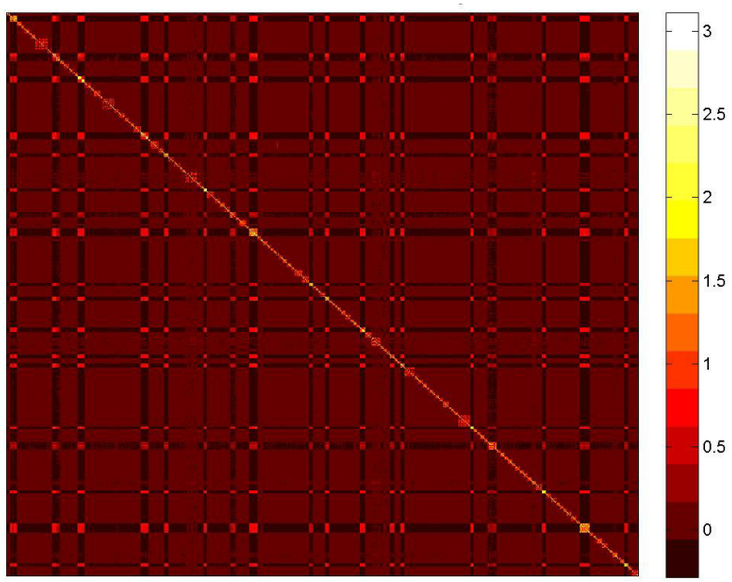
Hidden structure: population structure

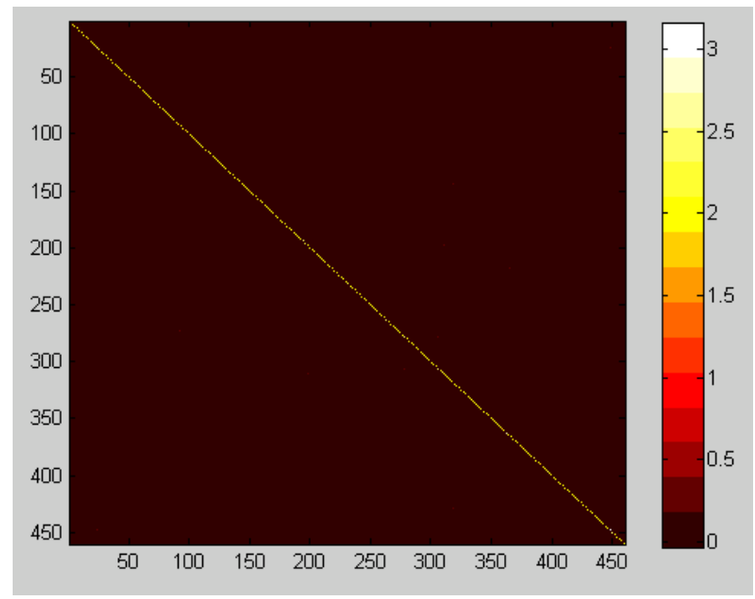
3 populations

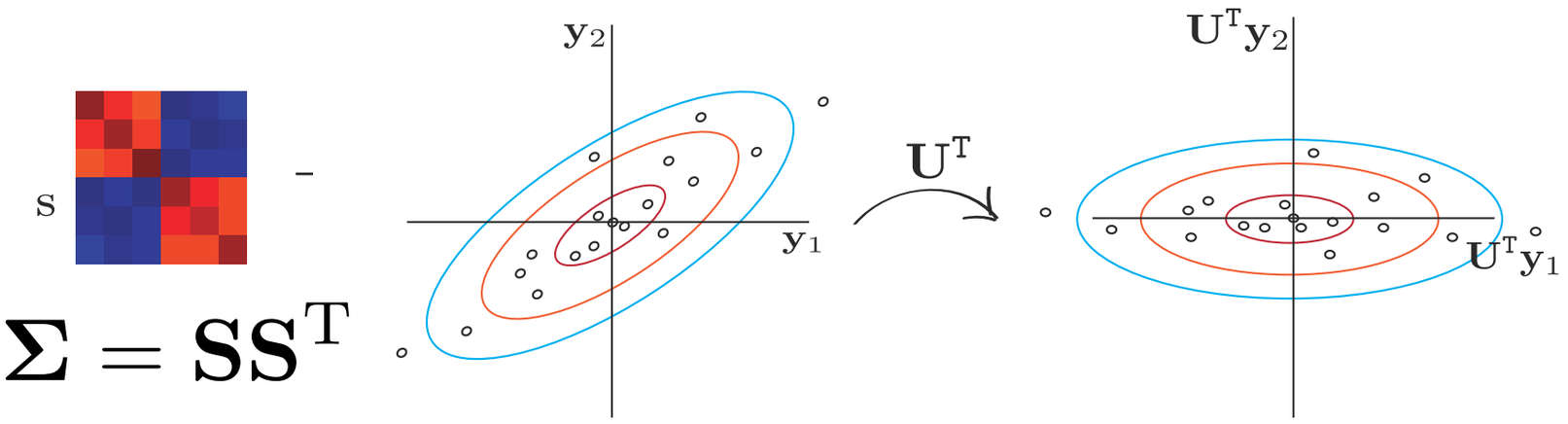
families

no structure









Hidden structure: population structure

LINEAR MODEL

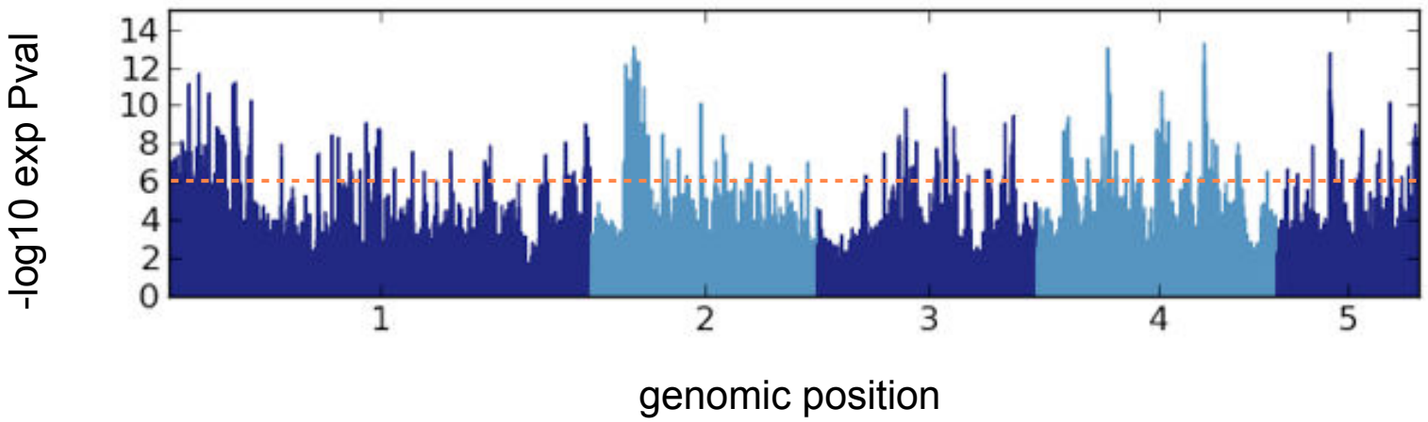
$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\psi}$$

pheno | **X SNP** | **noise**

NOISE

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

N > 1,000



flowering time
A. Thaliana
(Atwell et al. 2010)

Flowering in A. thaliana

Hidden structure: population structure

LINEAR MIXED MODEL

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} + \boldsymbol{\psi}$$

pheno | **SNP** | **confounding** | **noise**

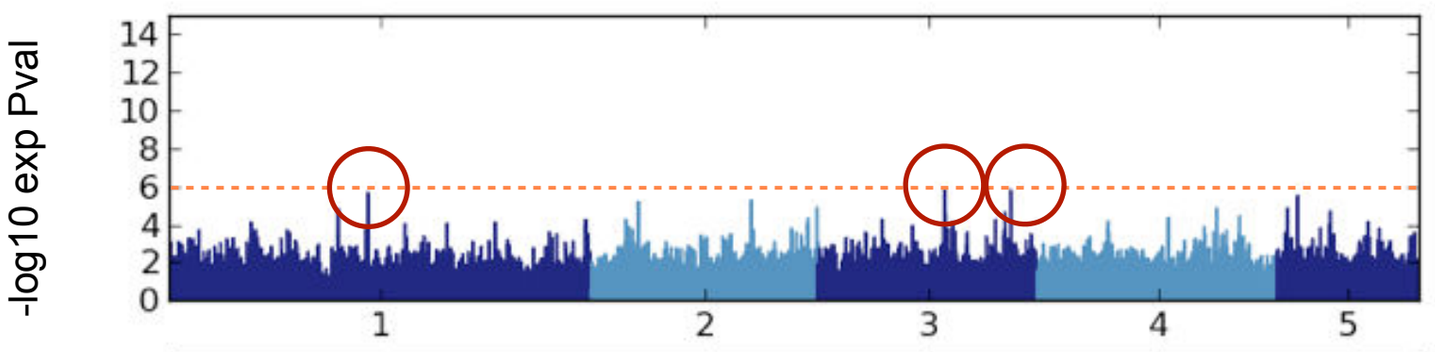
GW GENETIC TERM

$$\mathbf{u} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2_{g\,\mathrm{s}} \right)$$

NOISE

$$\boldsymbol{\psi} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2_e \right)$$

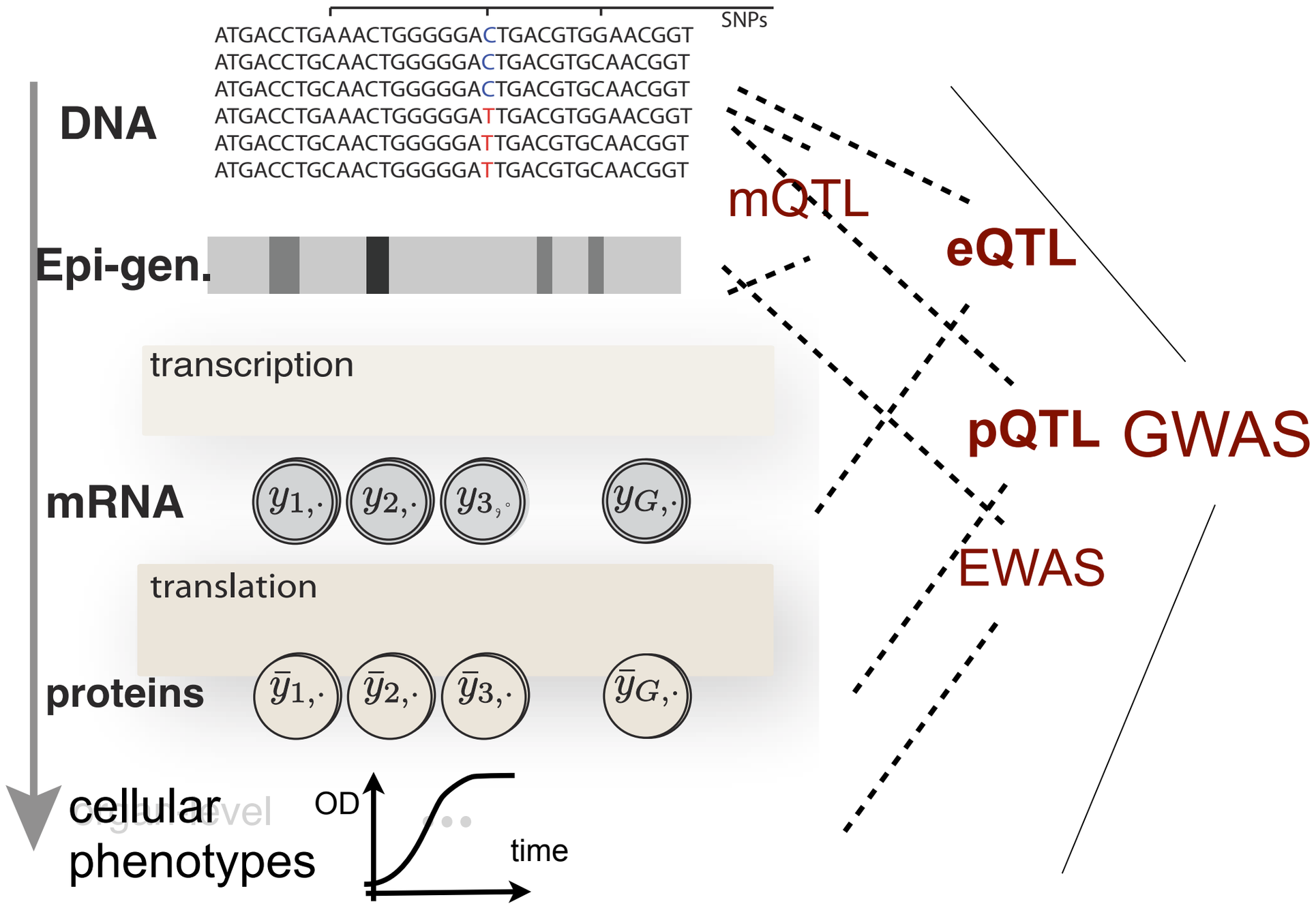
N ~ 5



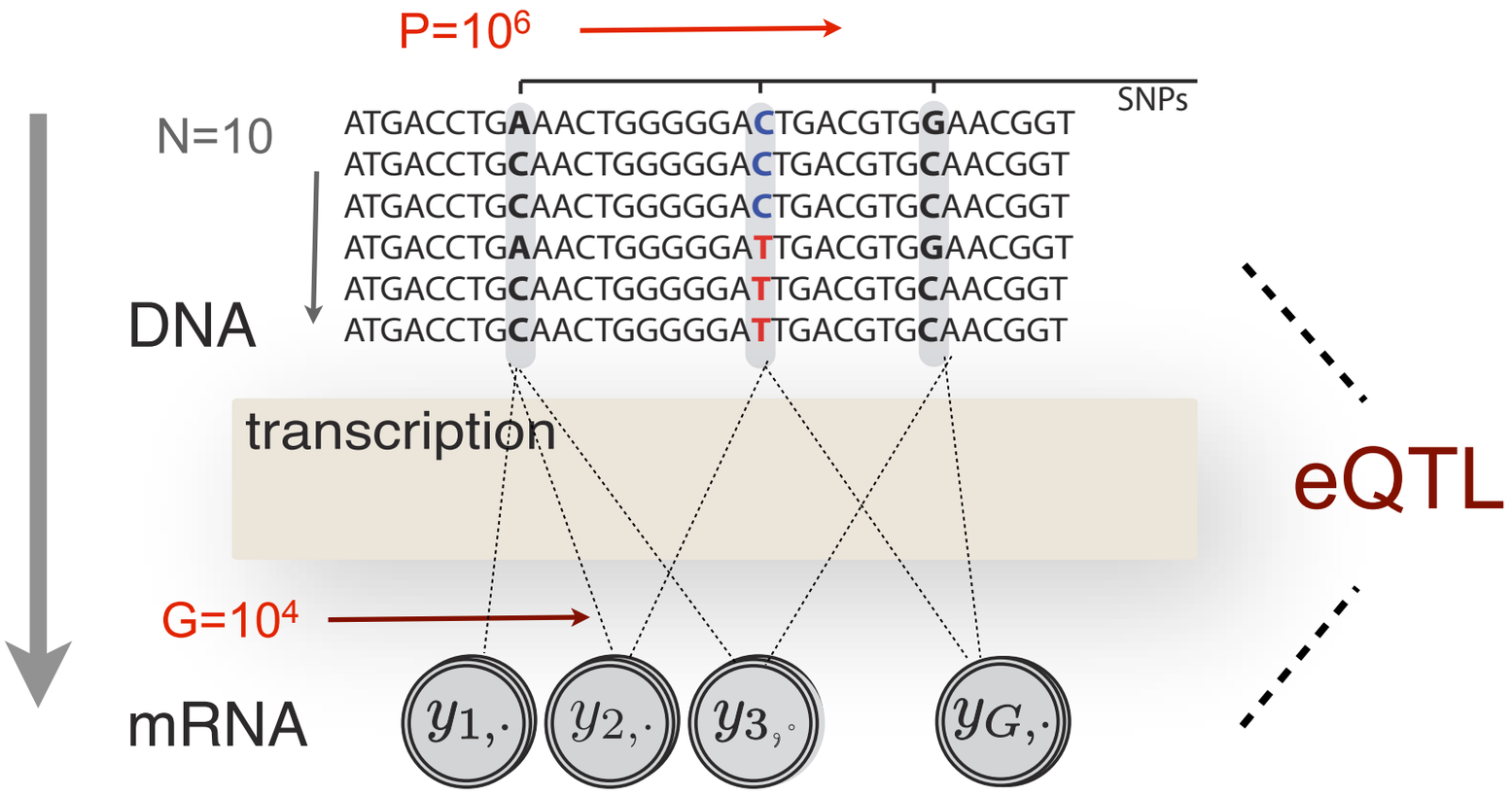
flowering time
A. Thaliana
(Atwell et al. 2010)

Flowering in A. thaliana

Multi-modal data integration across molecular layers



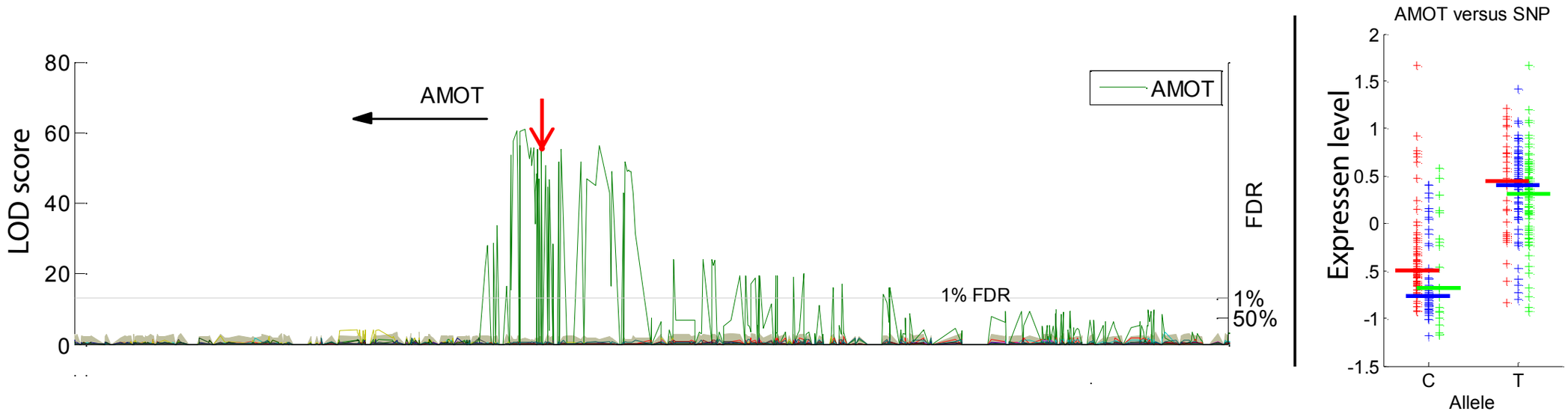
Association genetics with high-dimensional molecular phenotypes



challenge:

- multiple testing
- non-genetic sources of variation

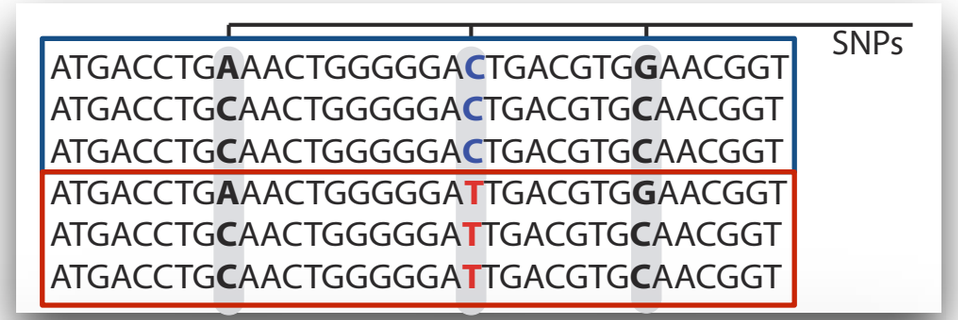
Association genetics with high-dimensional phenotypes



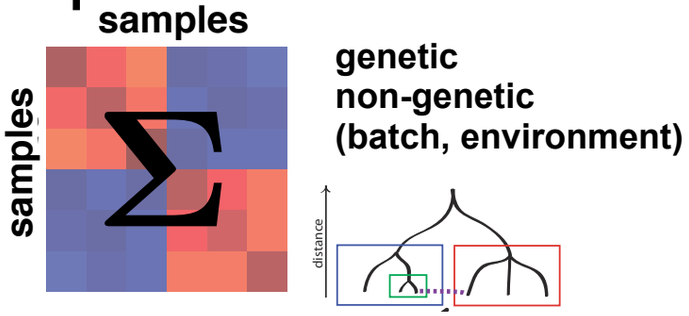
Linear mixed models to account for sample-to-sample covariance

population 1

population 2



sample covariance



transcriptome

- 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

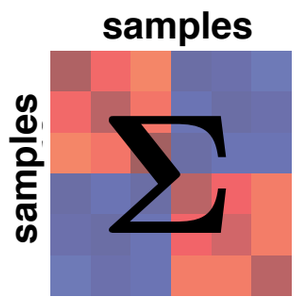
Confounding factors: genetic and non-genetic structure

- genetic

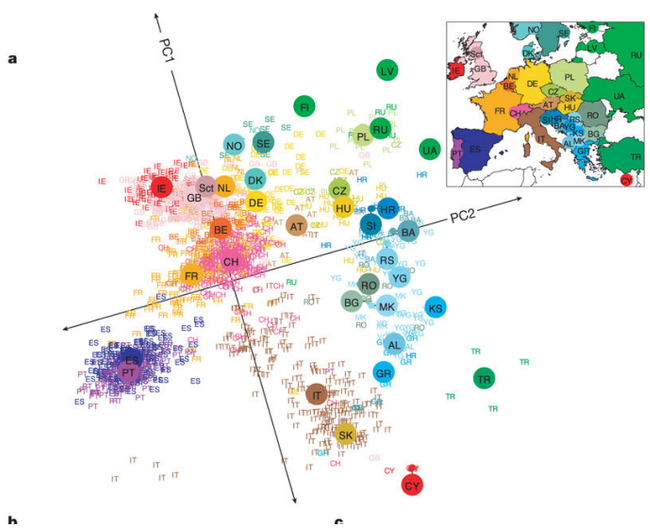
$$\boldsymbol{\Sigma} = \mathbf{S}\mathbf{S}^{\mathrm{T}}$$

- non-genetic

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



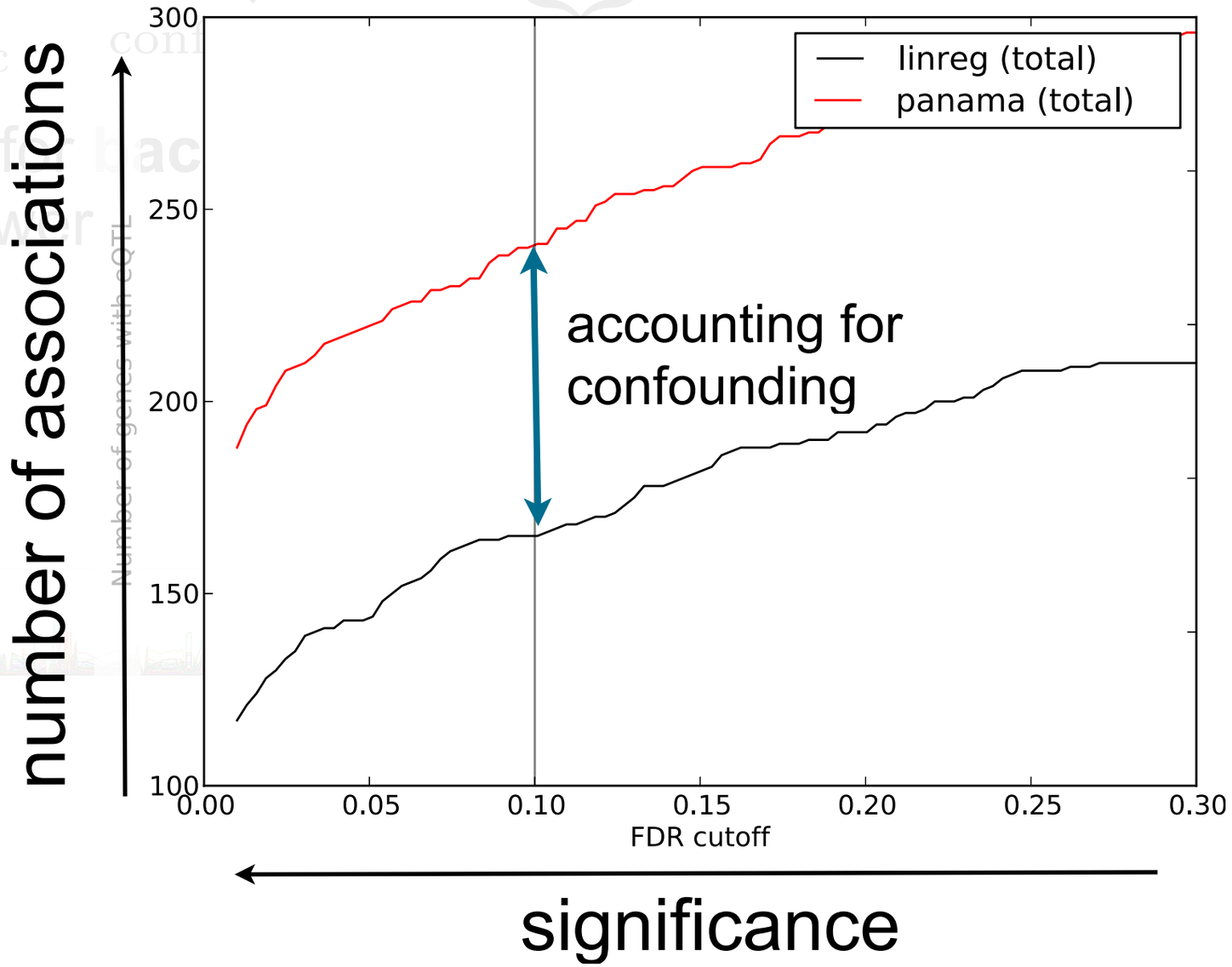
- Empirical gene expression covariance



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

genetic non-genetic

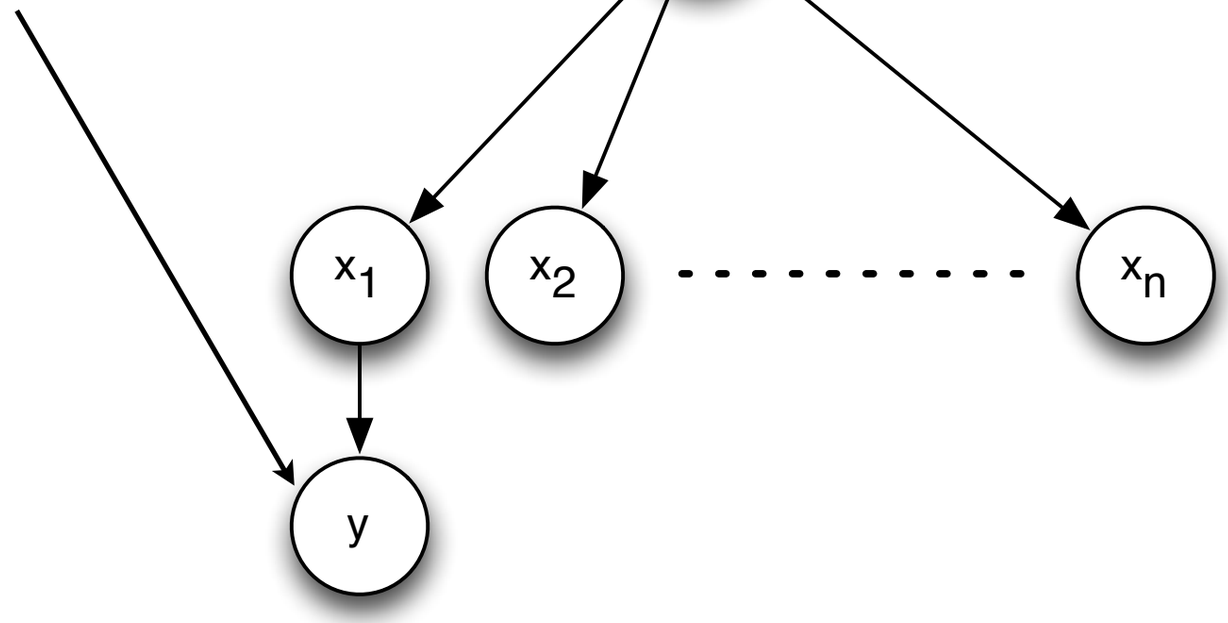
Confounding factors: genetic and non-genetic structure



Nature, Lappalainen et al. 2013

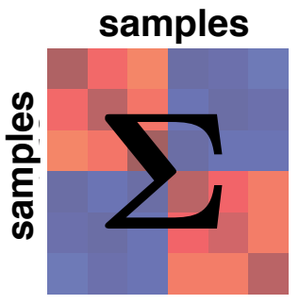
Confounding factors: genetic and non-genetic structure

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



SNPs

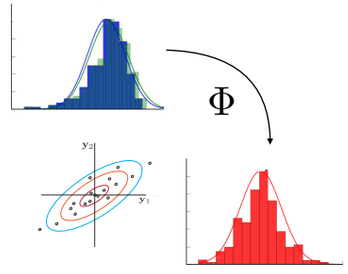
phenotype



Extending linear mixed models: beyond single-SNP single phenotype analyses

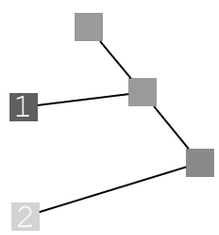
- Statistical challenges in high-dimensional association genetics
  - Normalization and scaling of quantitative traits

    Fusi et al., Nat Comm (2014)



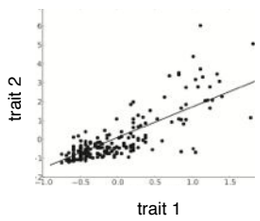
  - Accounting for epistasis and non-linear genetic interactions

    Stephan et al., Nat Comm (2015)

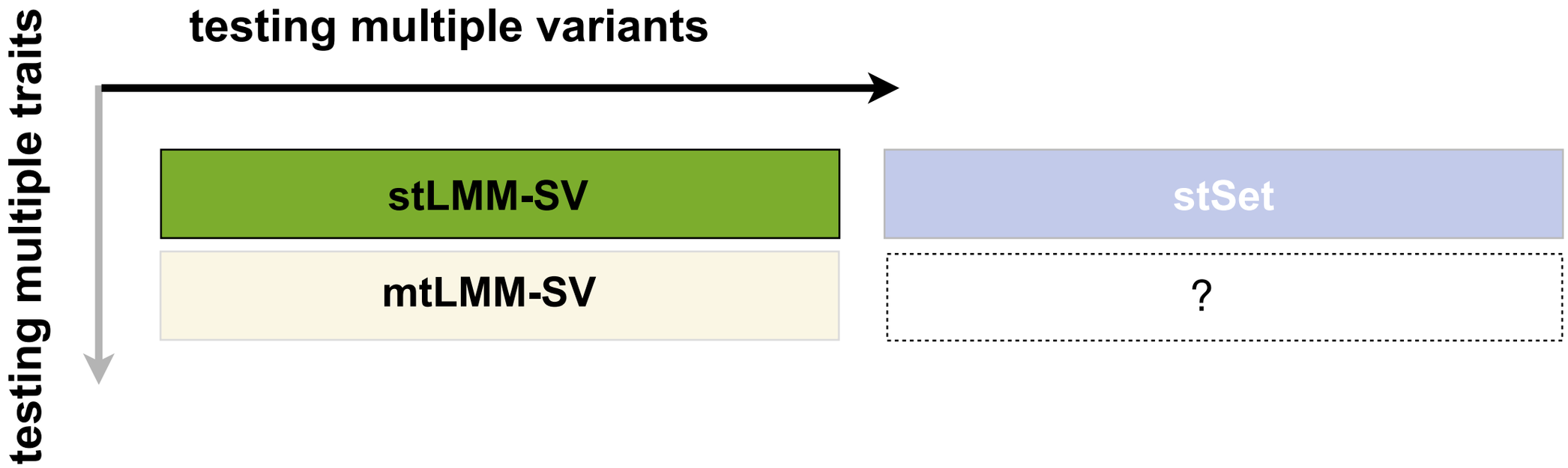


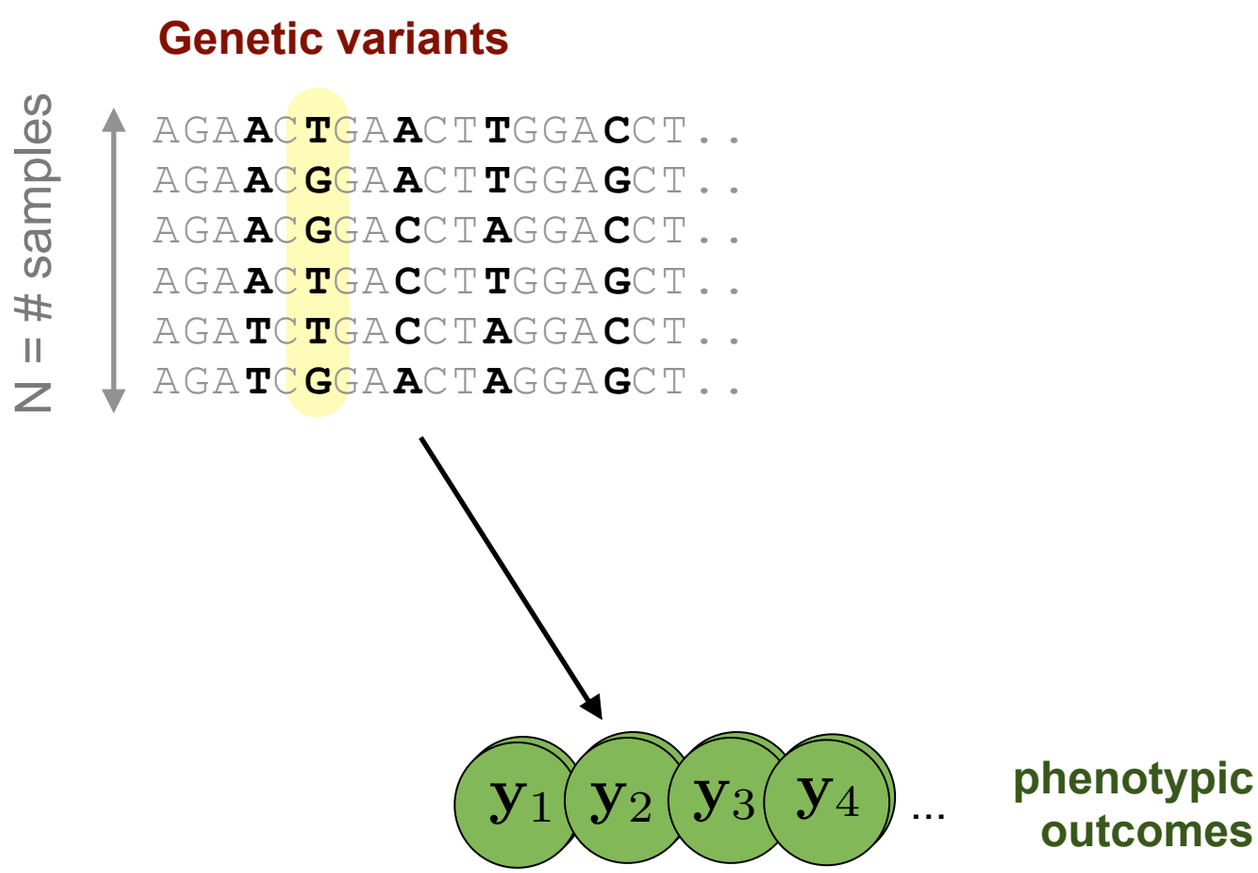
  - **Joint modeling of multiple traits and variants**

    Casale et al., Nat Meth (2015)

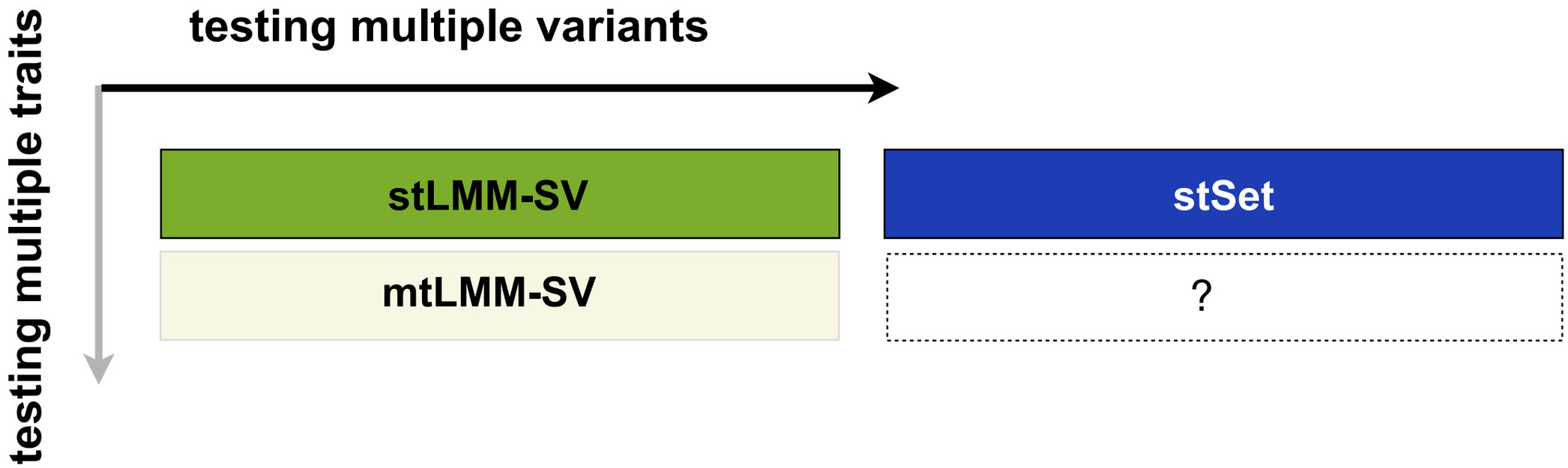


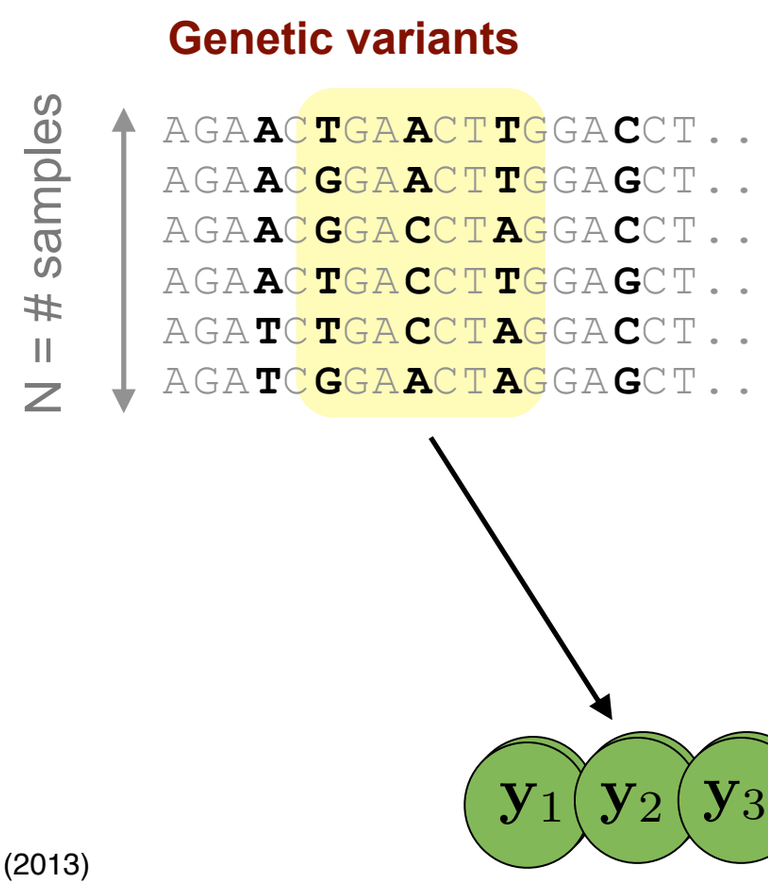
Joint modelling of traits and variants





Joint modelling of traits and variants



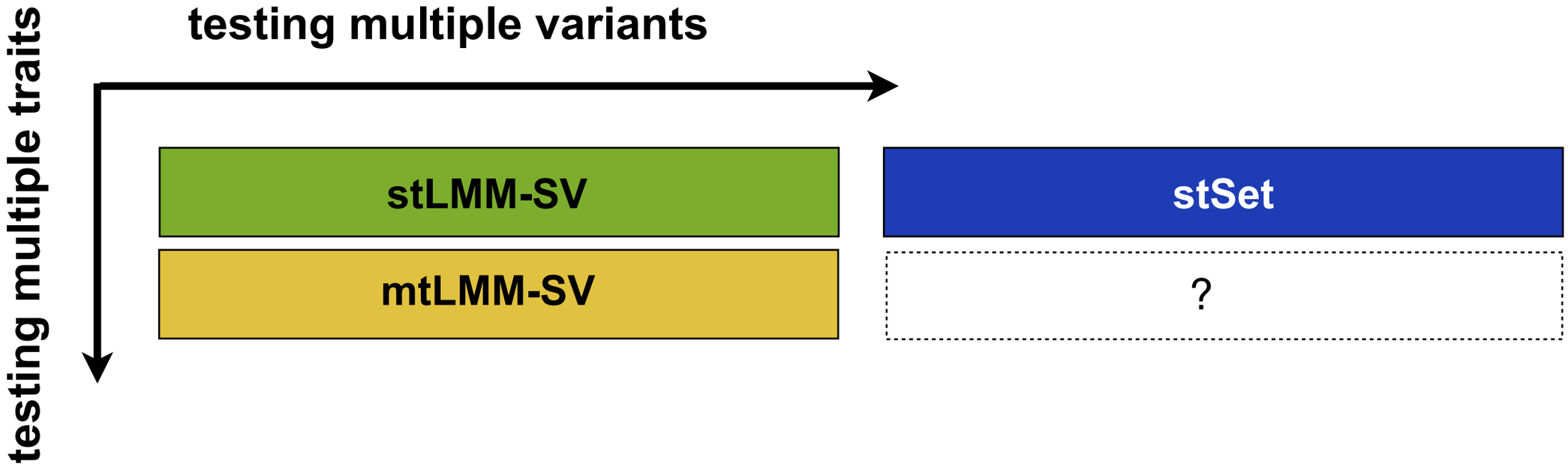


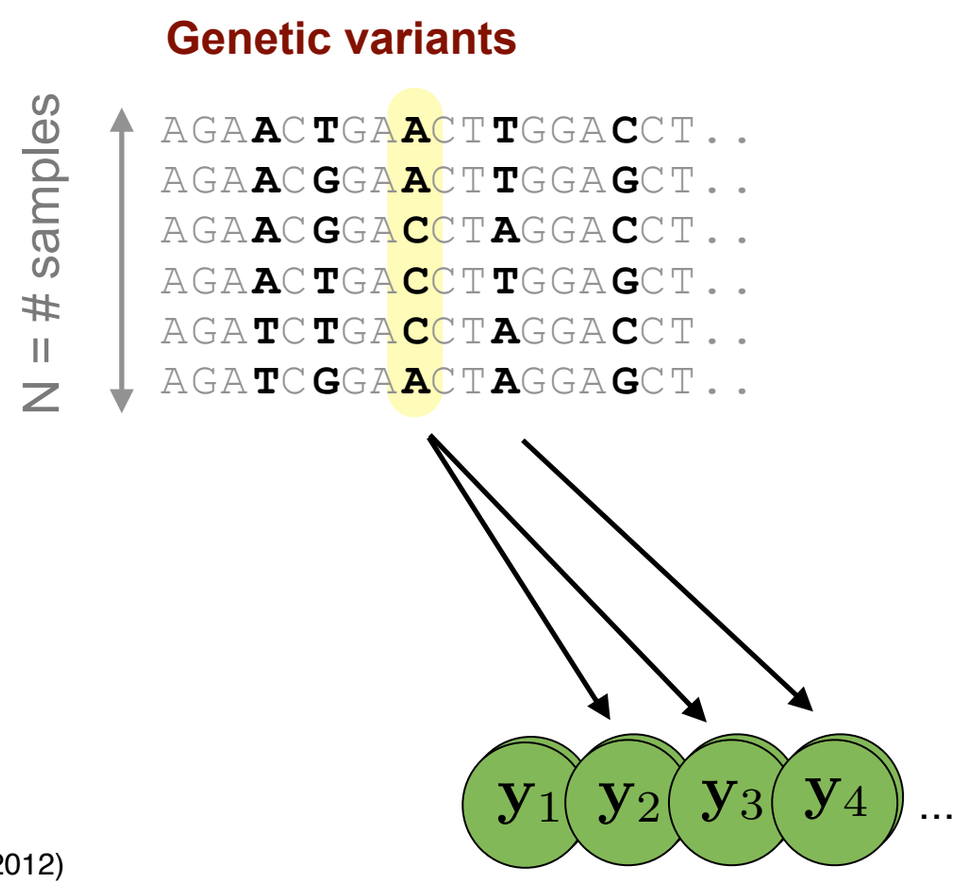
- multiple causal variants
- allelic heterogeneity

phenotypic outcomes

Listgarten et al. Bioinformatics (2013)

Joint modelling of traits and variants

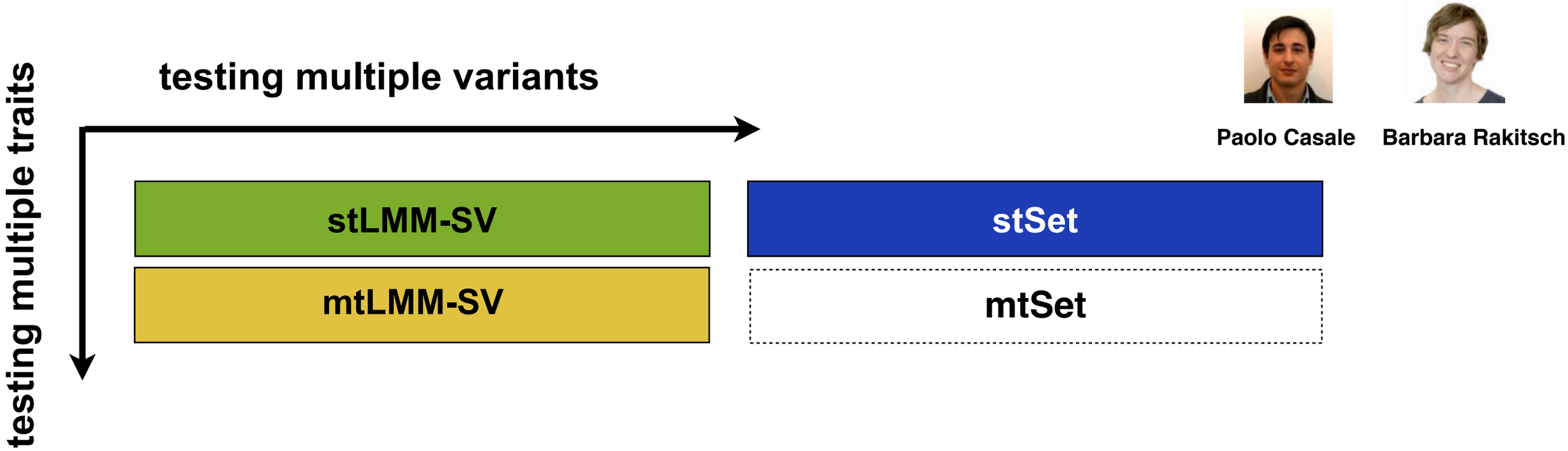




- leverage correlated traits

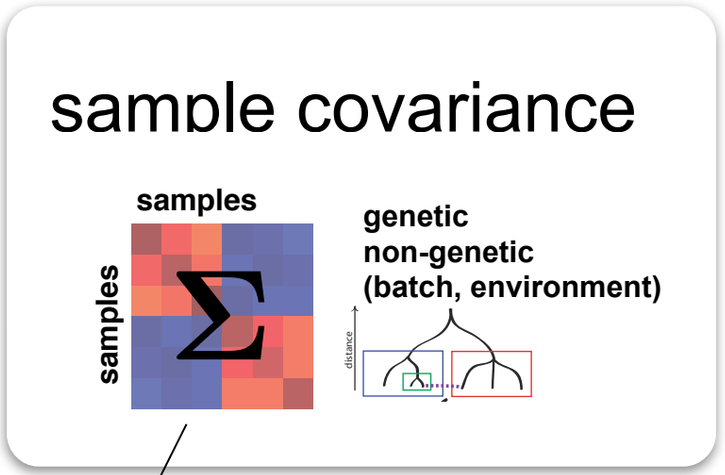
Korte et al., Nature Genetics (2012)

Joint modelling of traits and variants



Casale et al., Nat Meth (2015)

mtSet: aggregation across traits and causal variants



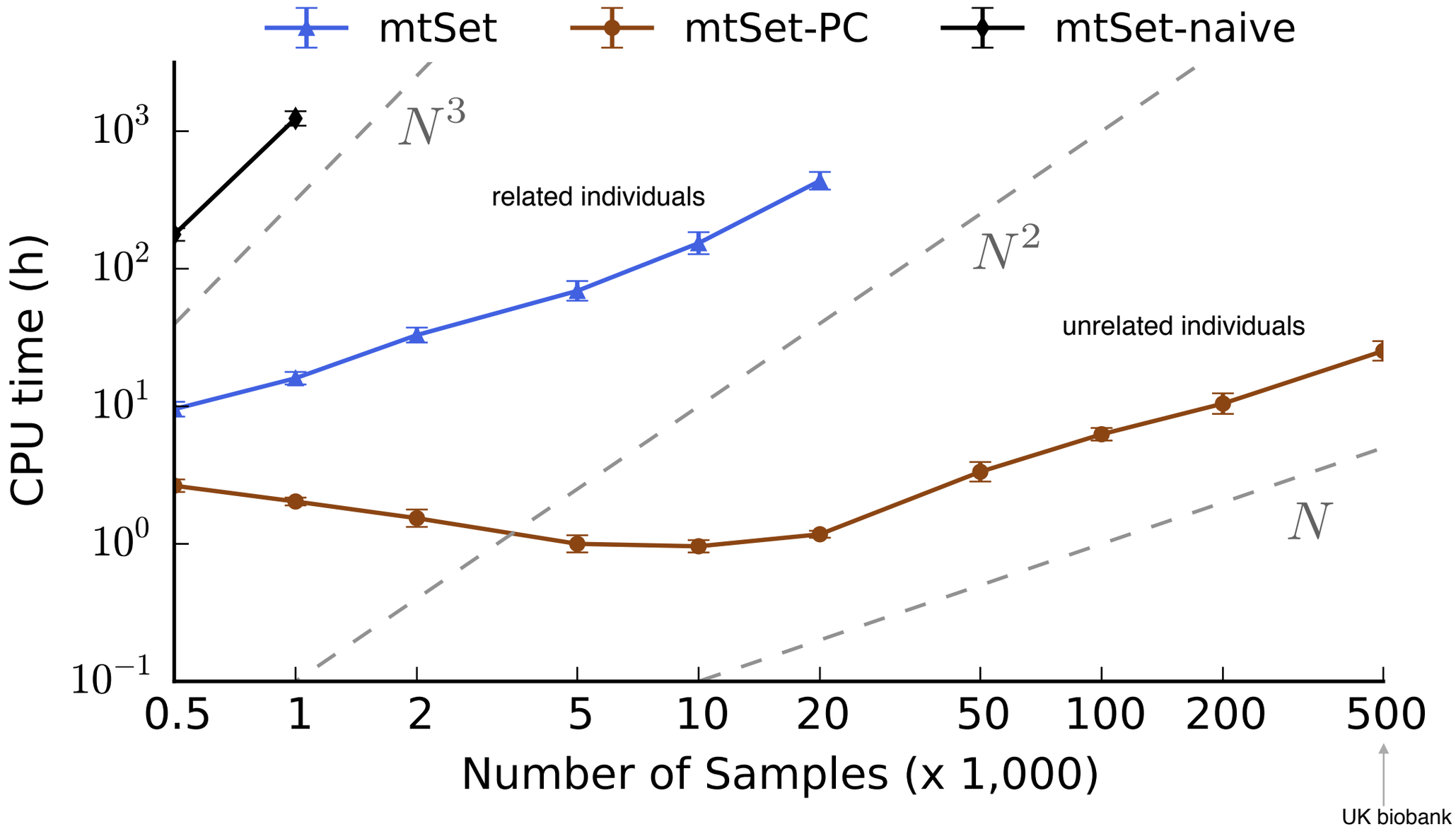
$$\underbrace{\mathbf{Y}}_{\text{phenotypes}} = \underbrace{\mathbf{FW}}_{\text{covariates}} + \underbrace{\mathbf{R}}_{\text{SNPs}} + \underbrace{\mathbf{U}}_{\text{relatedness}} + \underbrace{\mathbf{\Psi}}_{\text{noise}}$$

variance components
(random effects)

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

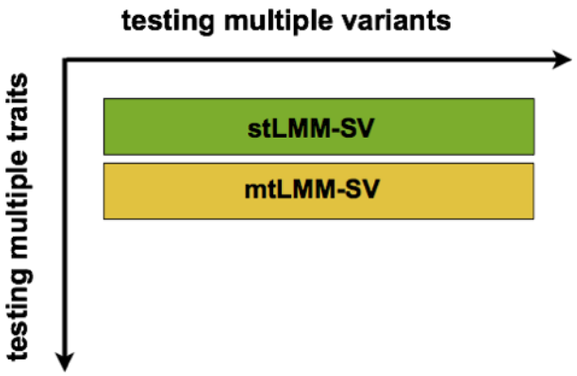
Challenge: Cubical scaling means such an algorithm is impractical for even moderately-size datasets!

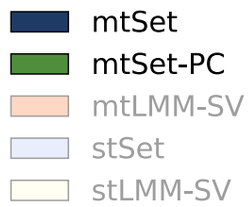
Efficient inference for large-scale GWAS

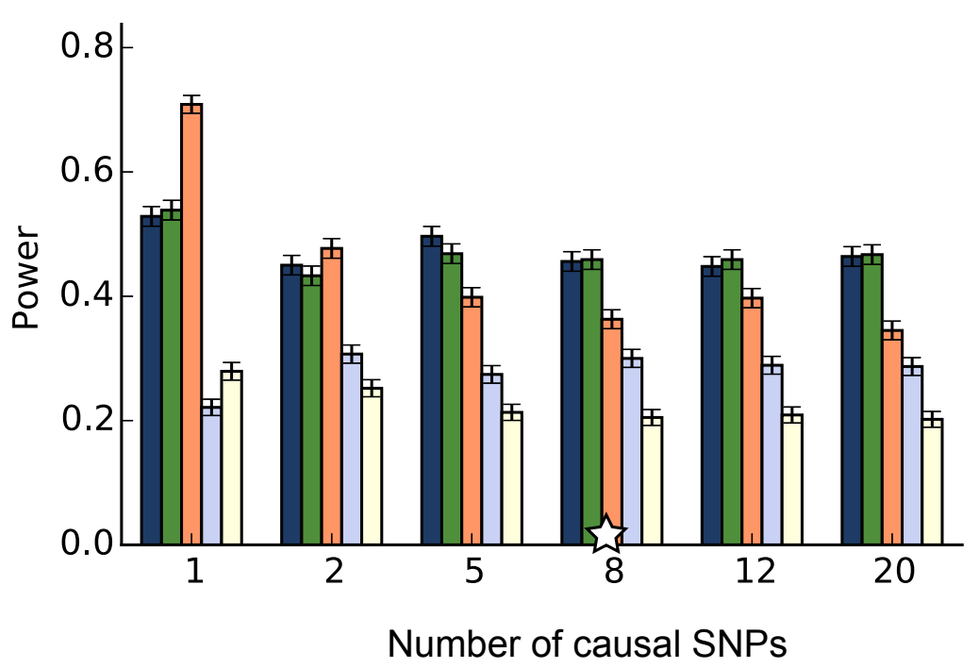


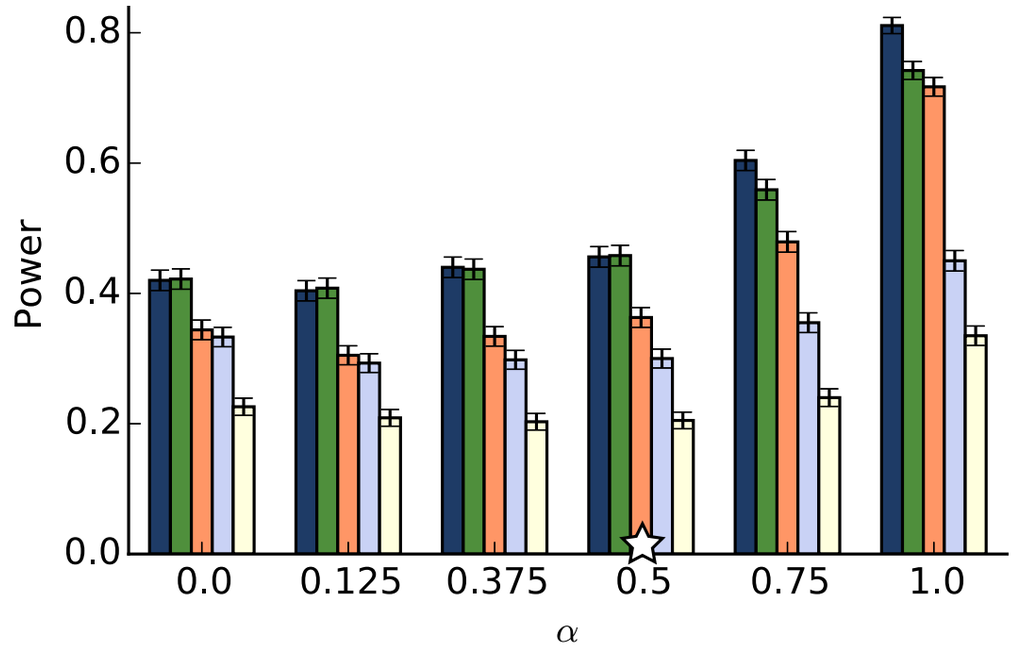
(human chrom20, 3,975 set tests for 4 traits)

Simulation study: aggregating across multiple causal variants and correlated traits









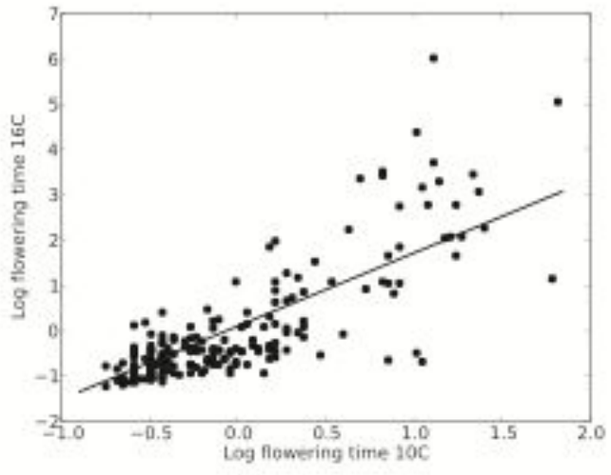
multiple causal variants →

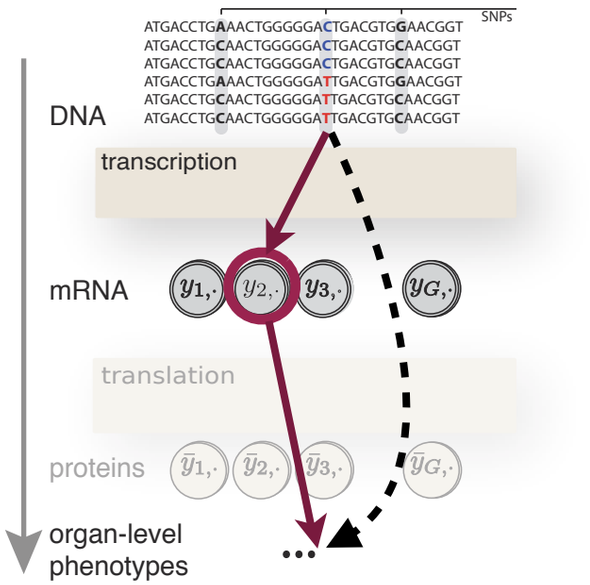
correlation between traits →

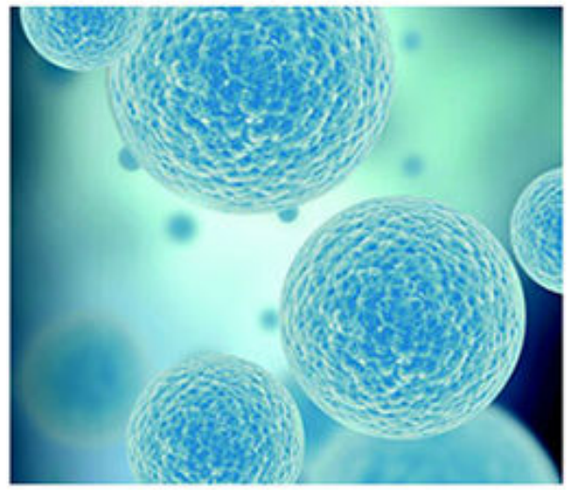
Summary so far

- Linear mixed models enable accounting for sample heterogeneity in genetic analyses
  - genetic (population structure): reduces false positives
  - non-genetic (batch, environment): increases power
- Scalability to large datasets
- Joint modeling of multiple (correlated) traits and multiple causal variants

Accounting for heterogeneity between individuals and single-cells







Statistical genetics

- Software (LIMIX)[1]
- Phenotype normalisation[2]
- Modelling epistatic relationships[3]
- Joint modelling of correlated traits[4]

Molecular heterogeneity

- Genetics of gene expression
- Causality[5]

- Human iPS biology
- Drosophila/yeast/plant genetics
- Cancer
- Drug susceptibility screens

single-cell heterogeneity

- Modelling heterogeneity in scRNA-Seq[6]
- Single cell DNA methylation profiling

1.Lippert et al., BioArxiv
2. Fusi et al., Nat Comm (2014)
3. Stephan et al., Nat Comm (2015)
4.Casale & Rakitsch et al., Nat Meth (2015)
5.Gagneur et al. PLoS Genet (2013)
6.Buettner et al., Nat BioTech (2015)

EMBL-EBI

Single-cell RNA-Seq

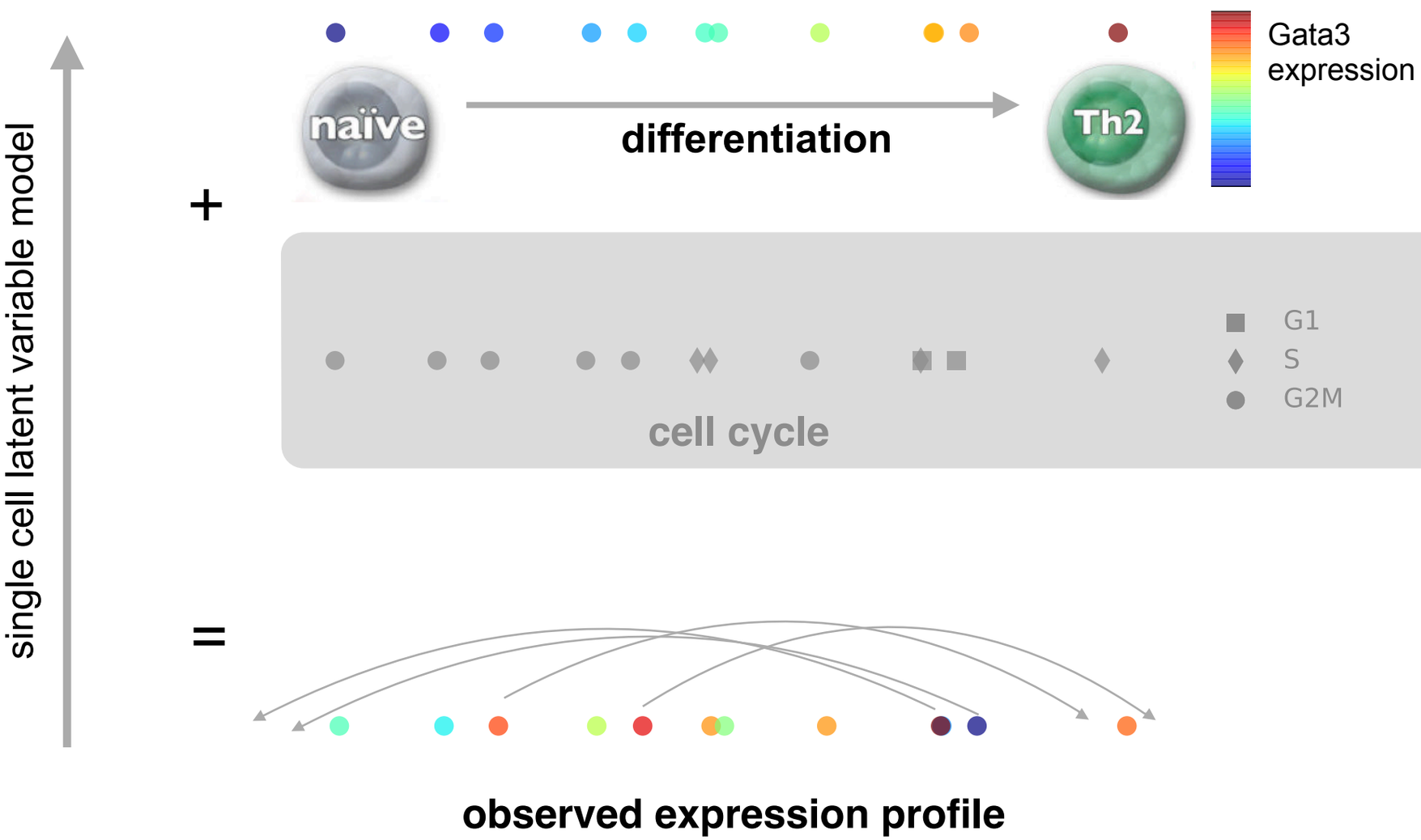
- Conventional RNA-Seq profiles are obtained from a pool of typically ~100,000+ cells.
- Using single-cell RNA-sequencing technologies, we can now assay RNA abundance in single cells.

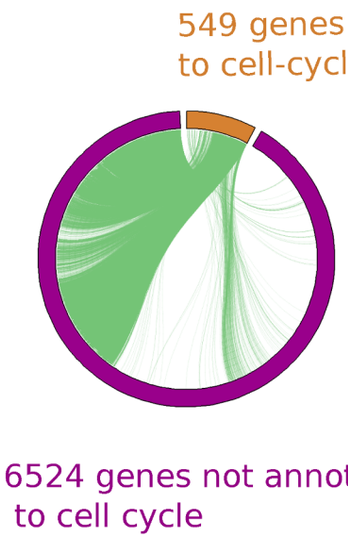
- novel variation between cells:
  cell type composition, **differentiation**

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

Fluidigm C1®

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





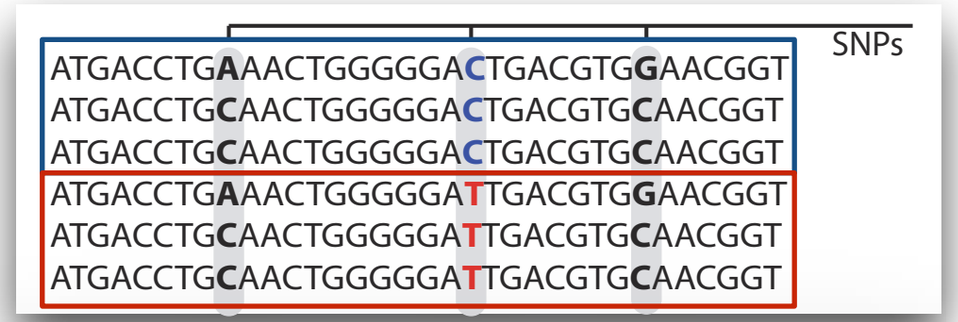
wide-spread correlation between cell cycle genes and non-cycle genes

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

Gene expression heterogeneity is not new…

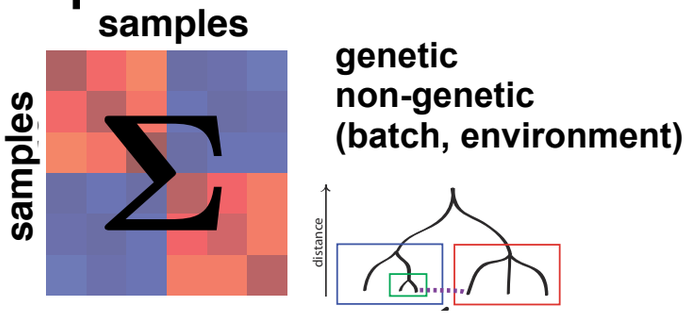
population 1

population 2

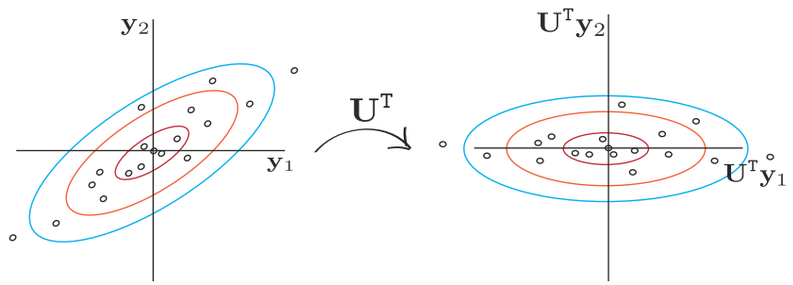


transcriptome

sample covariance



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

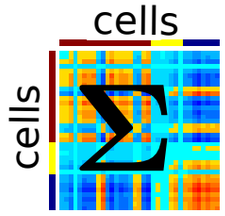


Leek & Storey, 2007
Kang et al. 2008
Listgarten et al, 2010
Lipper et al. 2011
Stegle* & Parts* et al. 2010, 2012
Fusi* & Stegle* et al. 2012

Single-cell latent variable model (scLVM)

Florian Buettner **Kedar Natarajan**

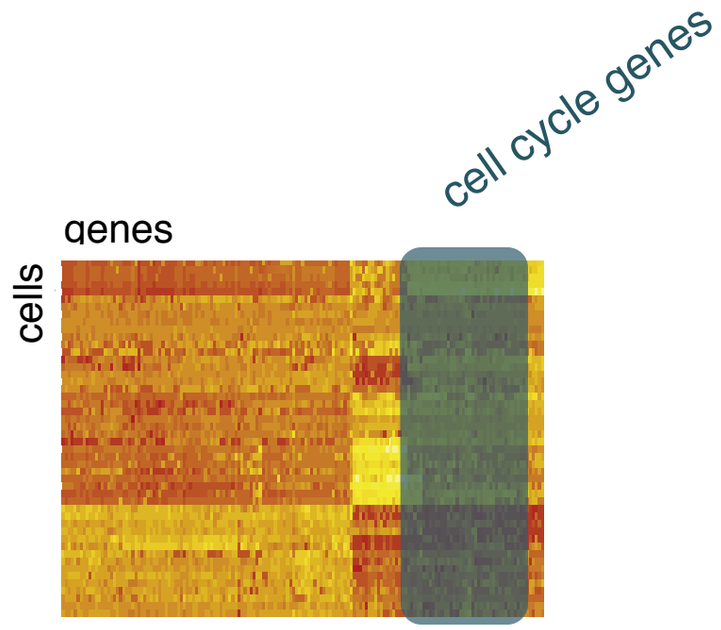
- Random effect model for cell cycle effects. Two-stage approach:
    1. Estimate a cell-cell covariance that captures cell cycle



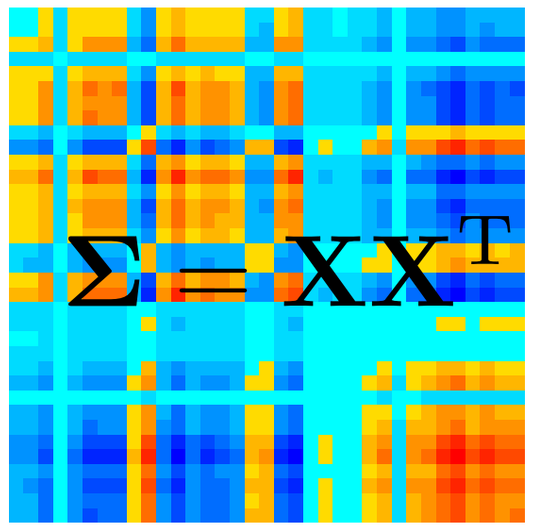
Estimation of cell-cycle induced

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**)

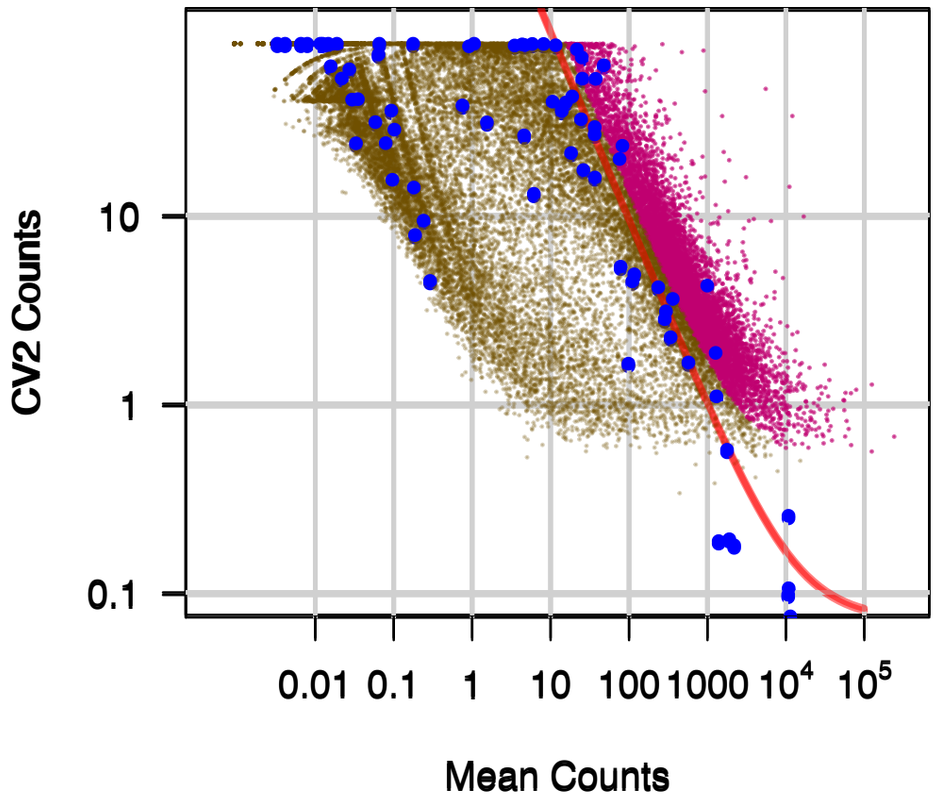


$$\mathbf{Y}_{\mathrm{cc}} \sim \prod_{g} \mathcal{N}(\mathbf{0} \mid \underbrace{\mathbf{X}\mathbf{X}^{\mathrm{T}}}_{\text{cell cycle covariance}} + \underbrace{\delta_b \mathbf{I}}_{\text{residual variance}})$$



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
  - Extrapolation to genome-wide genes
  - 7,073 highly variable genes



Brennecke et al. 2013

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

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)

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

$$N(0,\ \ ) \quad N(0,\ \ ) \quad N(0,\ \ )$$

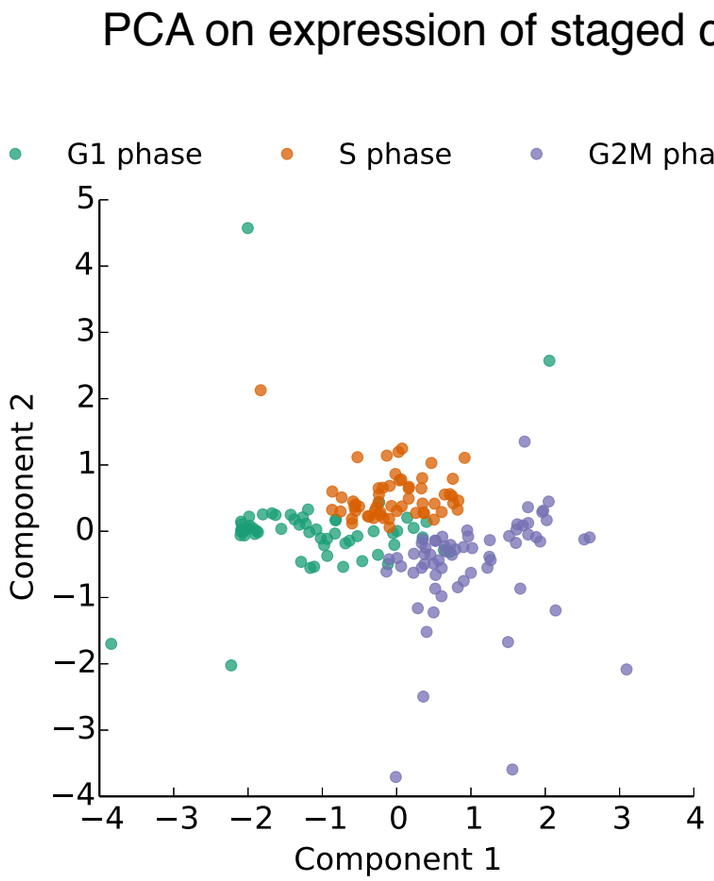
cell cycle

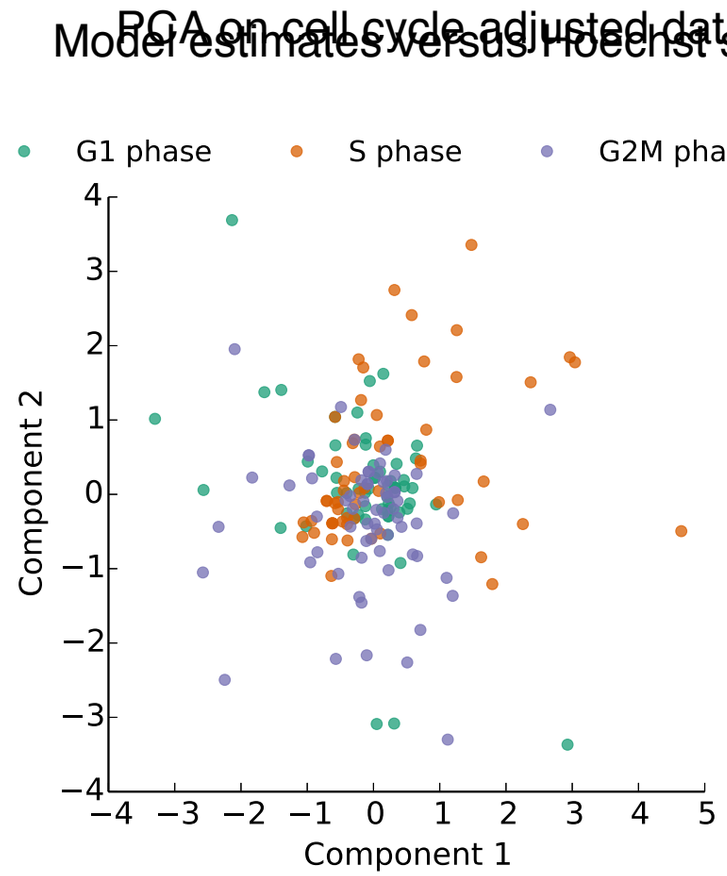
res. biological variability

technical noise

Model validation on mouse ESCs

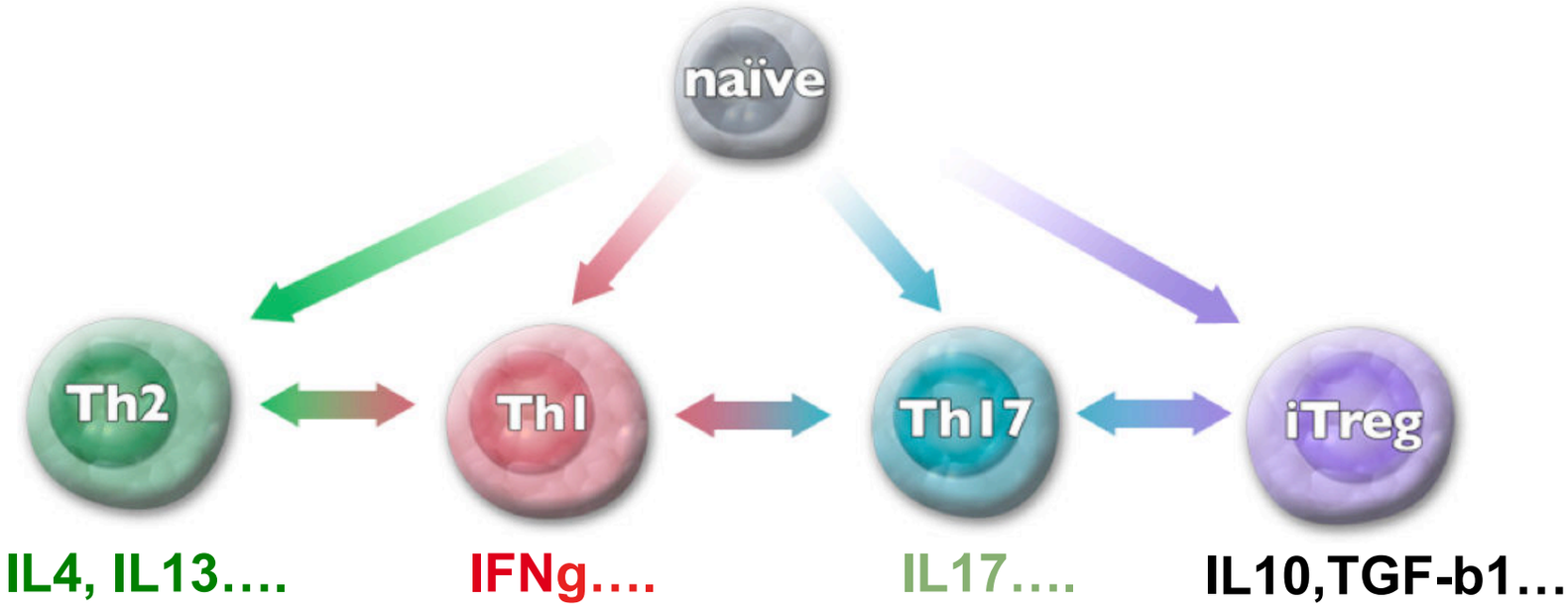
- To test our model, we used single-cell RNA-Seq data generated from ~300 ES cells collected at different stages of the cell cycle





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

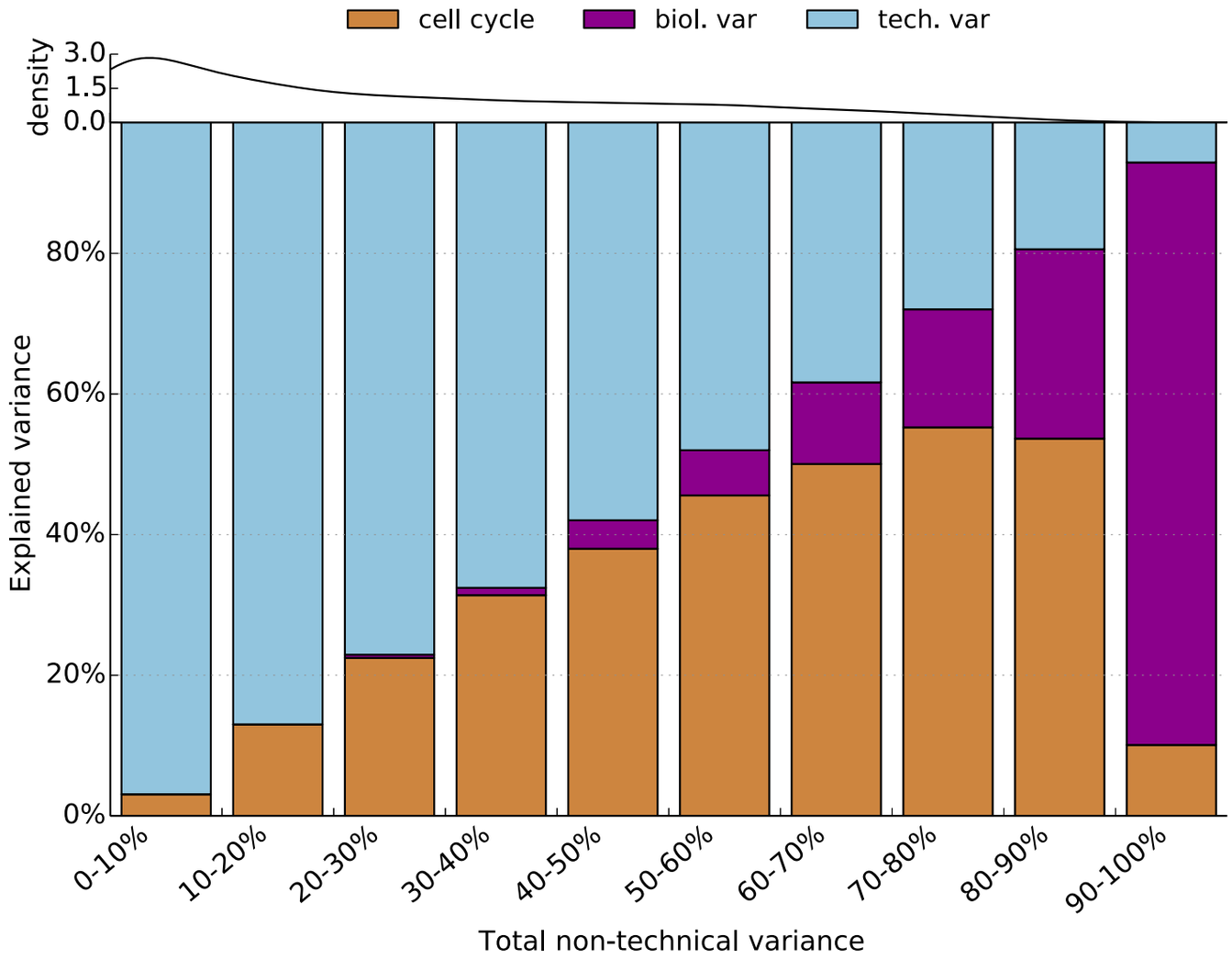
Application to T-cell differentiation



- 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

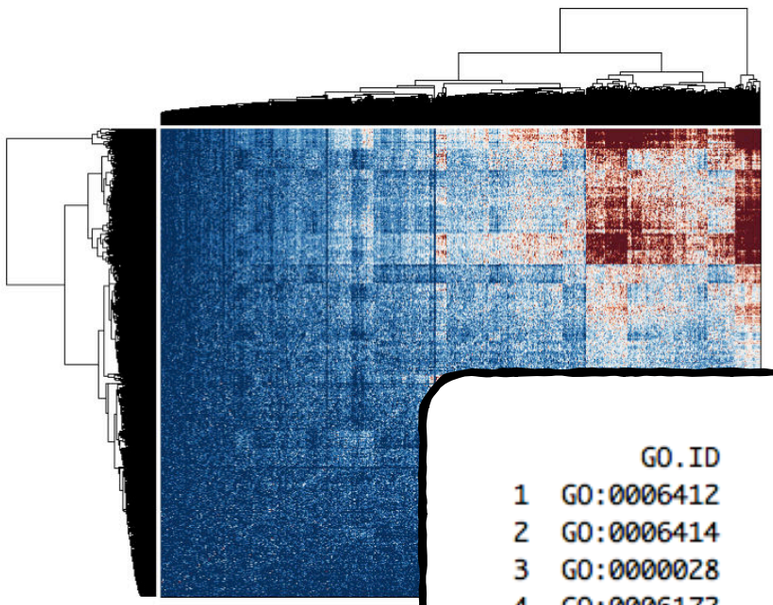
- **Technical noise**
  For 27% of the genes, variation of expression can be entirely explained by the (technical) null variability.
- **Cell-cycle**
  For 42% of the genes, >30% of the observed variance is explained by the cell cycle state.

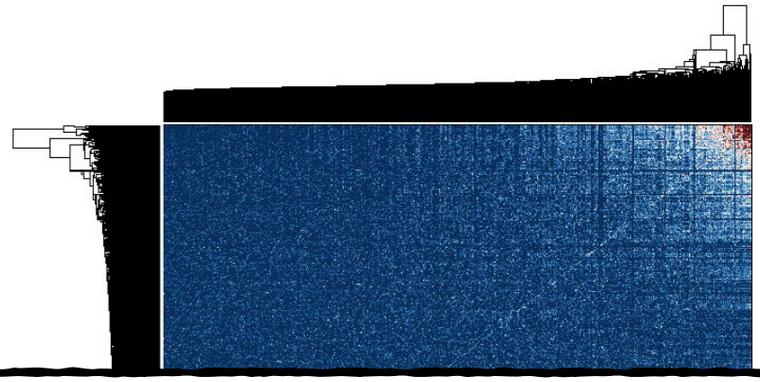


The impact of cell cycle on gene-gene correlations

Gene-gene correlations (unadjusted)

Gene-gene correlations (adjusted for cell cycle)

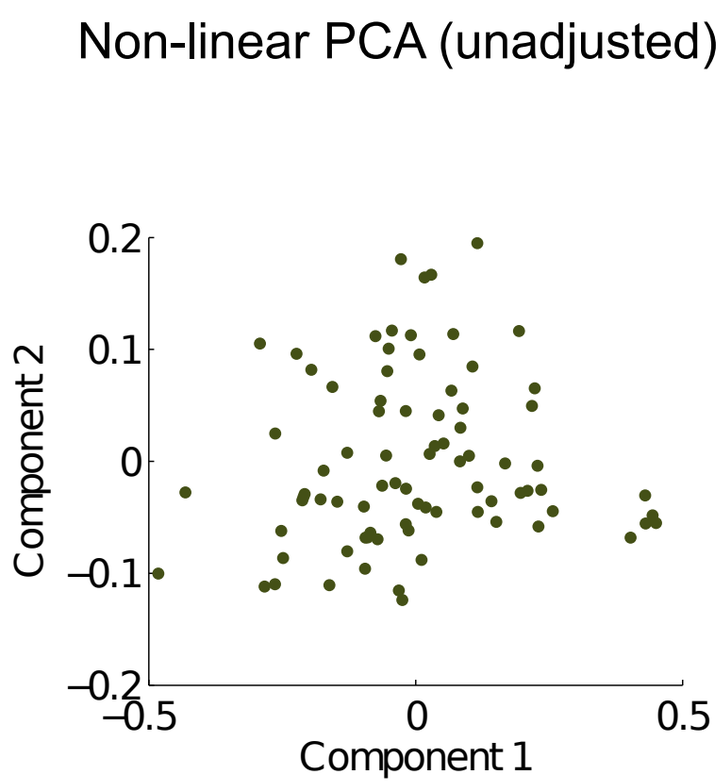


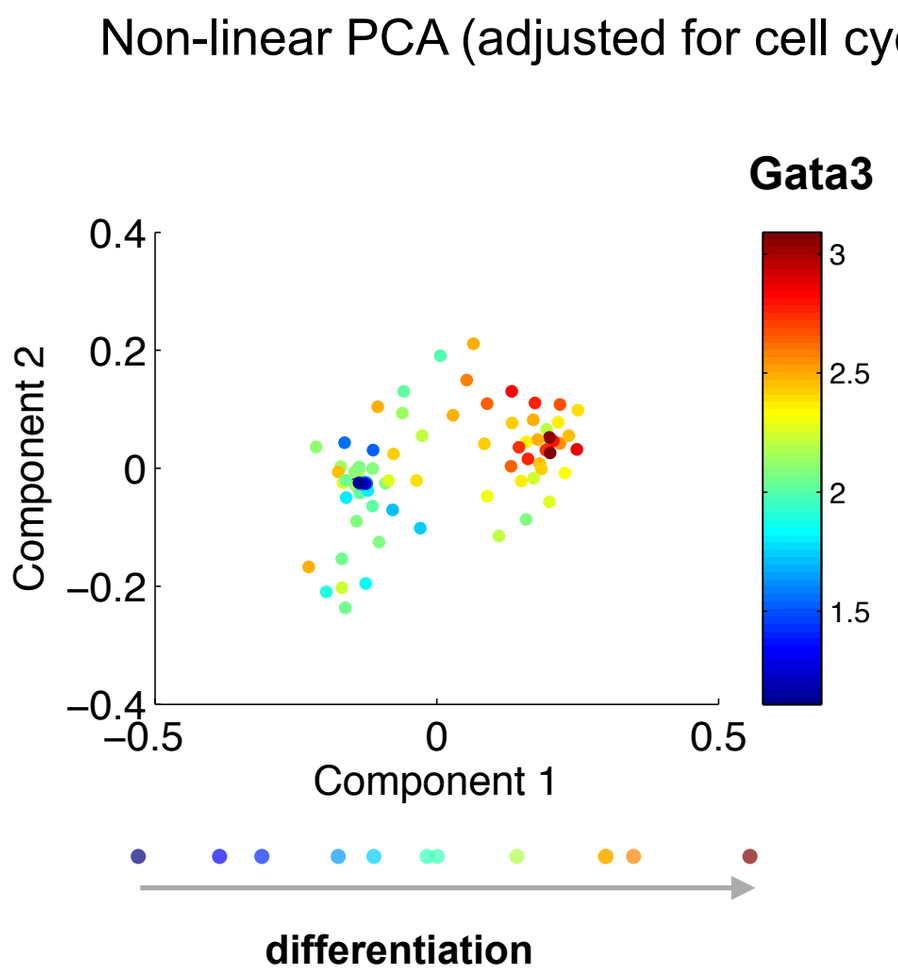


> 500,

	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

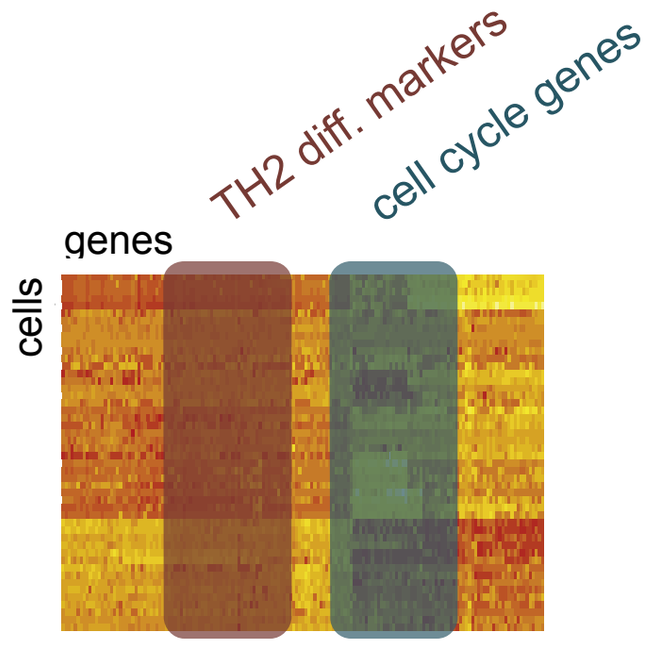




- After correction for cell heterogeneity, cells appear to separate better into two groups than without correction.

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

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

$N(0,$) $N(0,$) $N(0,$) $N(0,$)

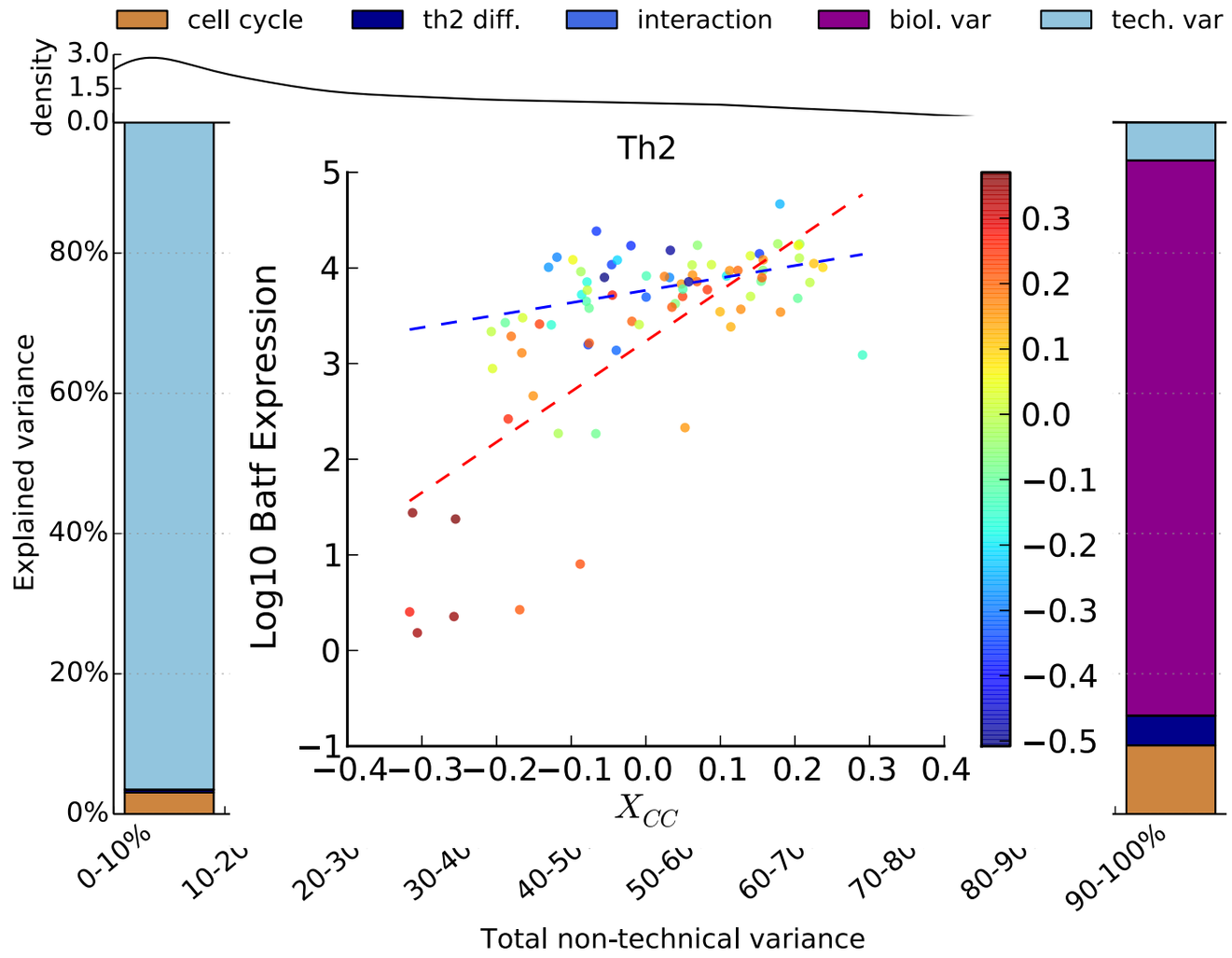
cell cycle
th2 differentiation
res. biological variability
technical noise

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



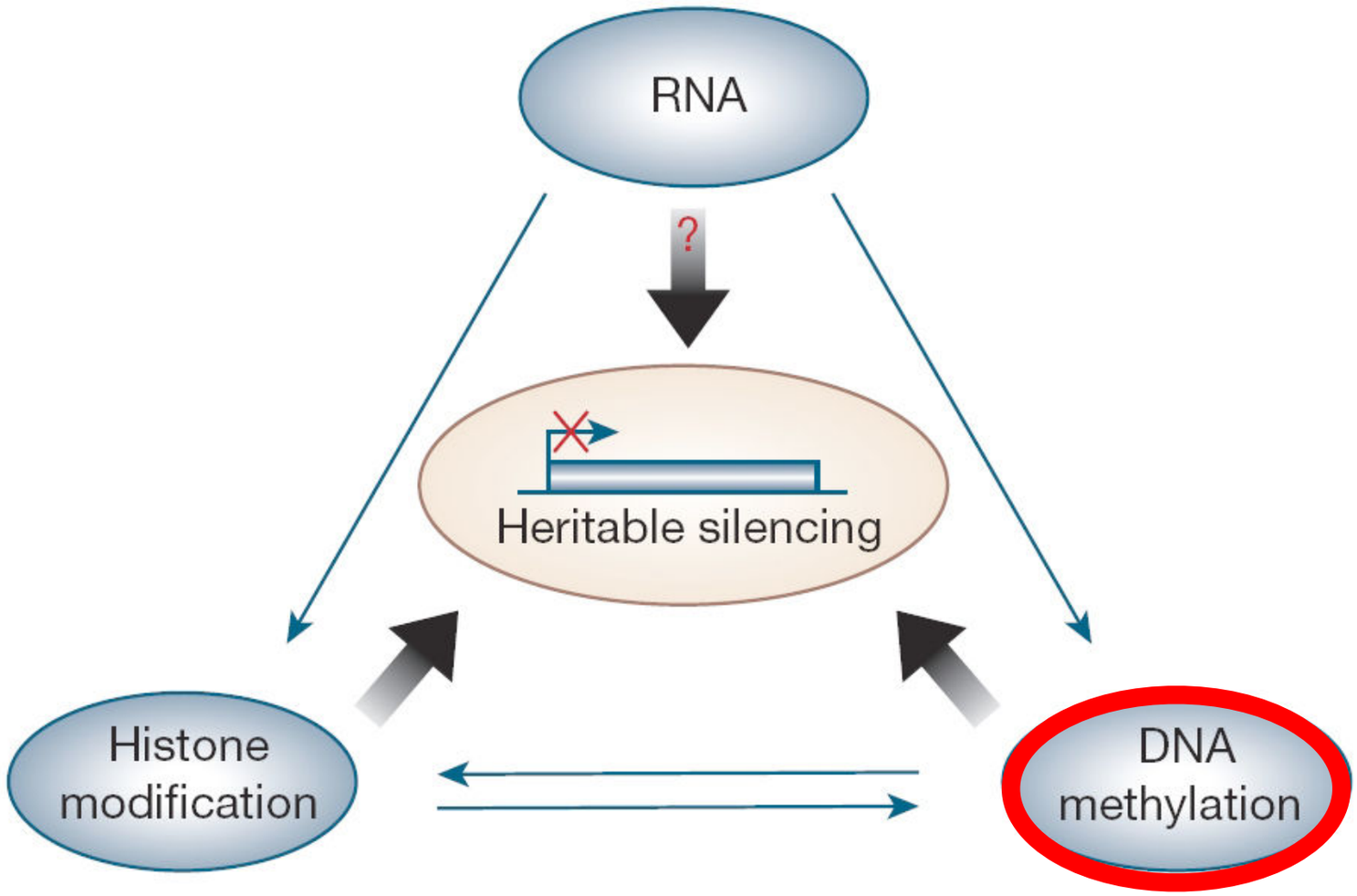
The origin of transcriptome diversity?

Christof Angermueller

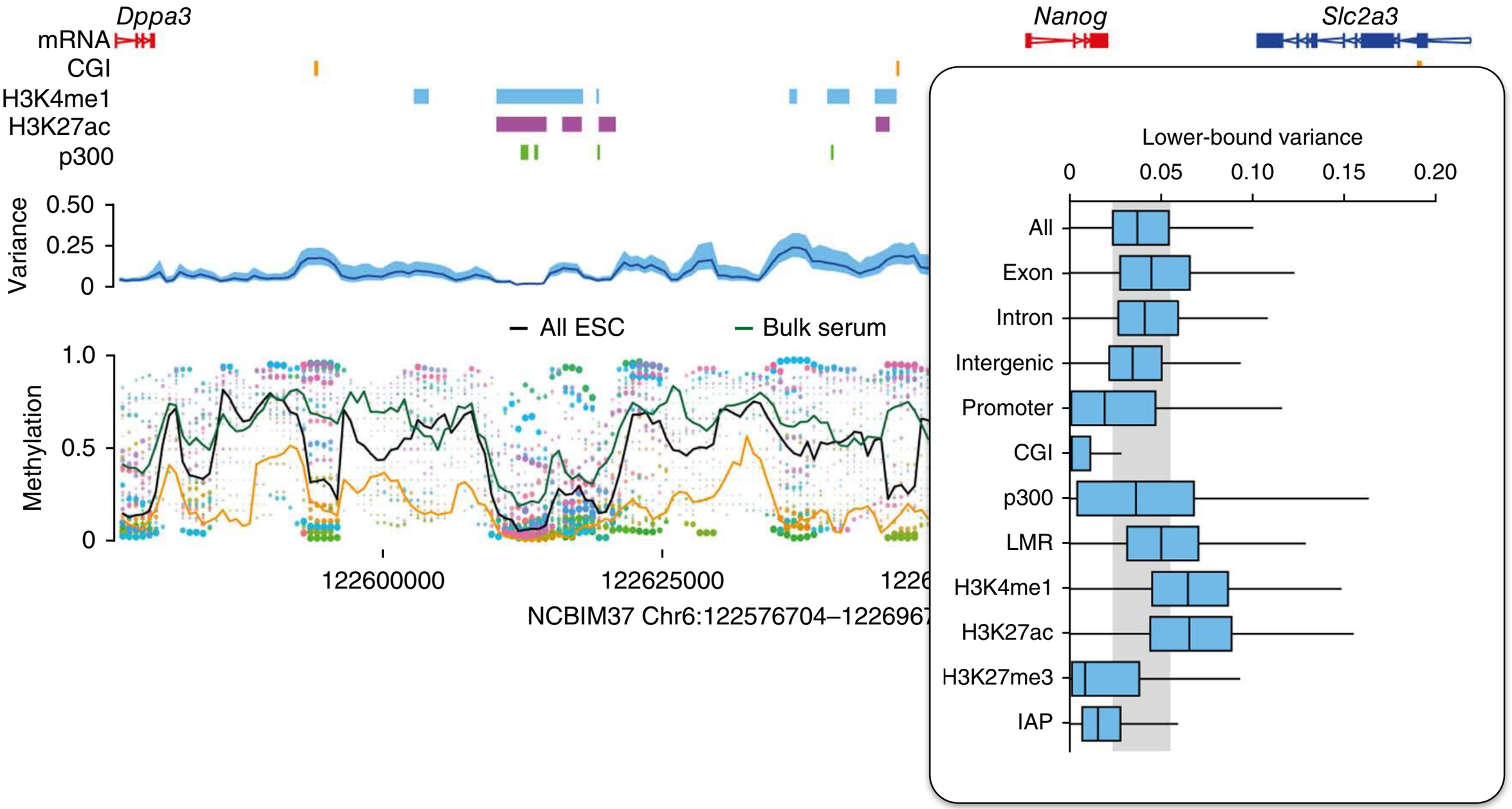
Heather Lee

Wolf Reik

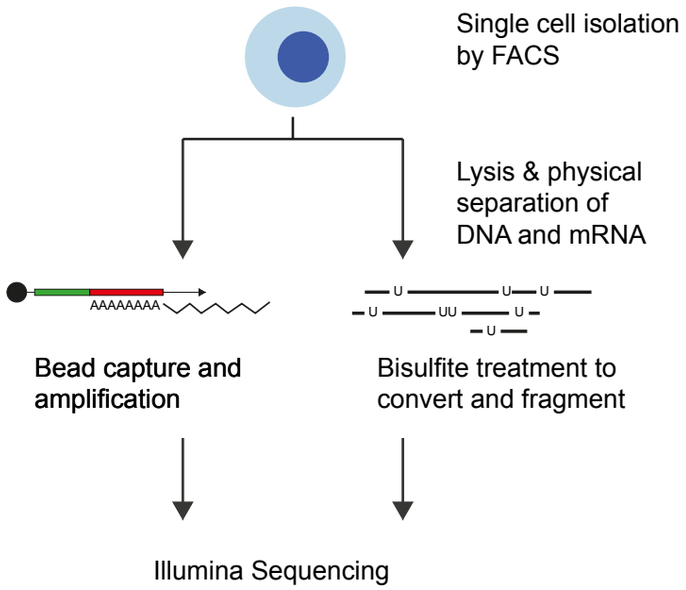
Thierry Voet



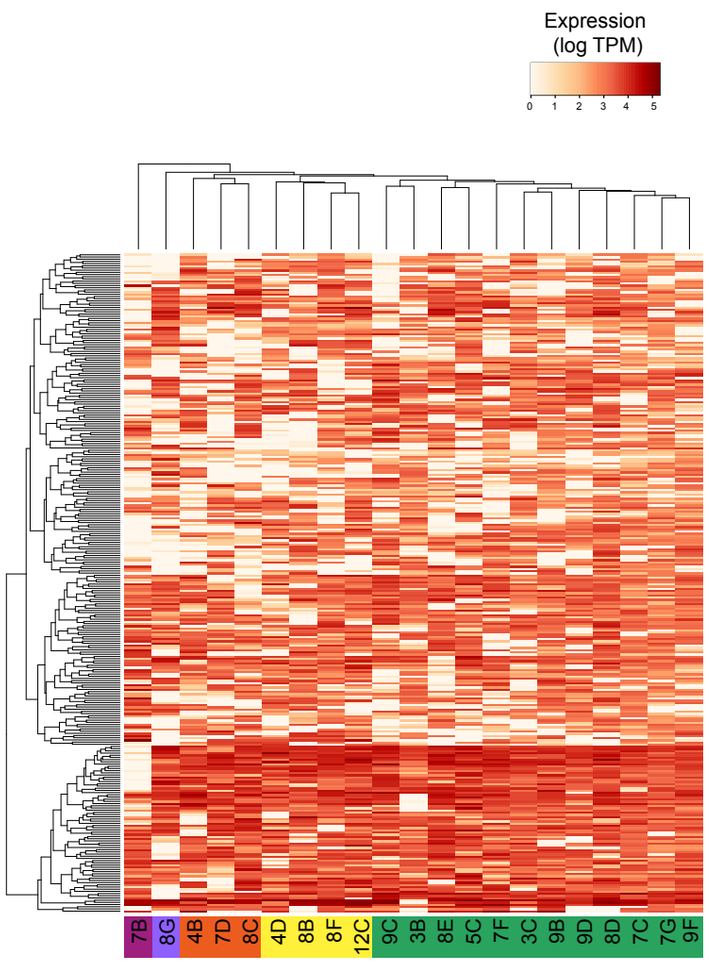
Single-cell bisulfite sequencing (20 ESC cells)



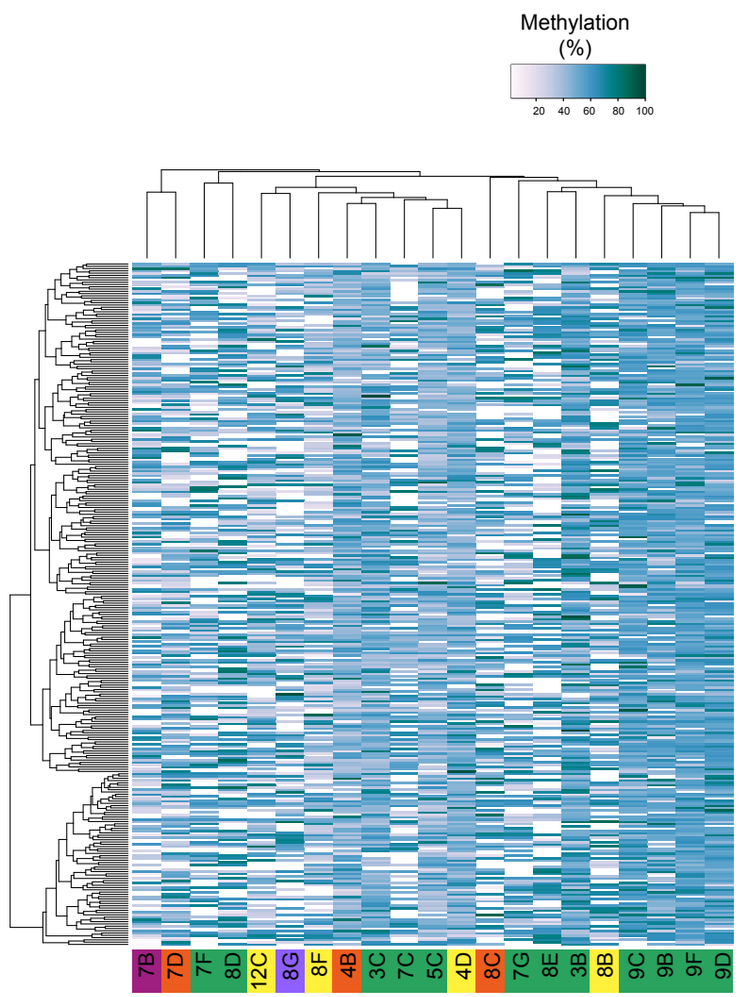
Parallel bs-seq & RNA-seq profiling in 21 serum ES cells



Gene expression

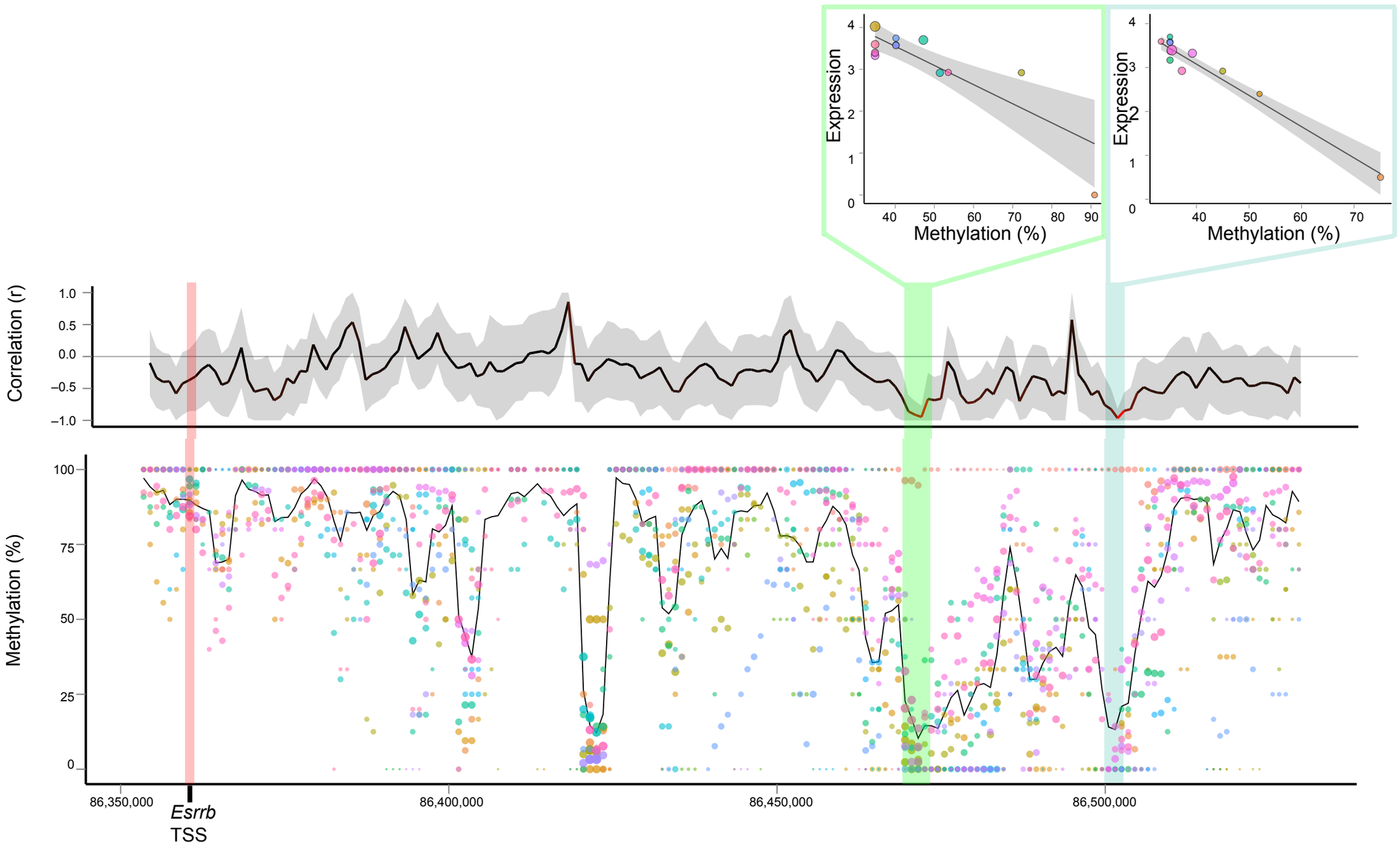


DNA methyatlion



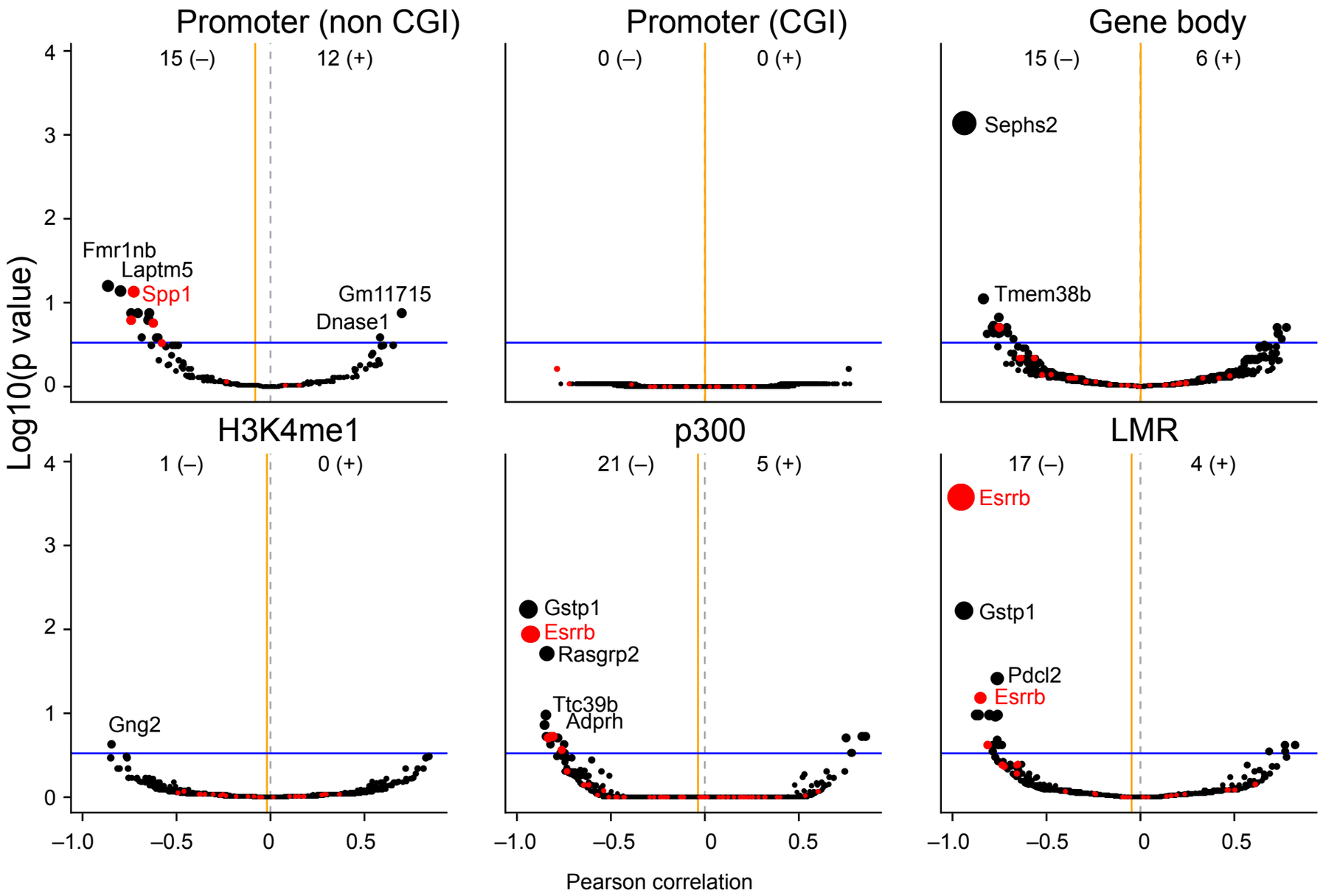
EMBL

Parallel bs-seq & RNA-seq profiling in 21 serum ES cells



EMBL

Parallel bs-seq & RNA-seq profiling in 21 serum ES cells



Conclusions

- Latent variable models can effectively account for gene expression heterogeneity & confounding
- (e)QTL analysis
  - population structure & env. /technical confounding to improve power
- 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
  - Parallel bs-seq/RNA-seq profiling in the same cells reveals associations between methylation and transcriptome variation

Acknowledgments

group
Barbara Rakitsch
Florian Buettner
Christof Angermüller
Paolo Casale
Helena Kilpinen
Amelie Baud
Danilo Horta
Johannes Stephan
Bogdan Mirauta
Kate Howell
Fatemeh Ghavidel

EBI/Sanger
John Marioni
Sarah Teichmann
Thierry Voet

Kedar Natarajan
Valentina Proserpio
Antonio Scialdone
Iain Macaulay

Babraham Inst.
Wolf Reik
Gavin Kelsey
Heather Lee
Stephen Clark

Helmholtz Munich
Fabian Theis

Microsoft Research
Nicolo Fusi
Christoph Lippert

University of Sheffield
Neil Lawrence

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

scLVM: https://github.com/PMBio/scLVM

Postdoc opportunities:
Single-cell genomics
Statistical genetics