# Explaining Confounding Factors in eQTL studies using a Dictionary of Latent Variables

Nicolò Fusi[1]

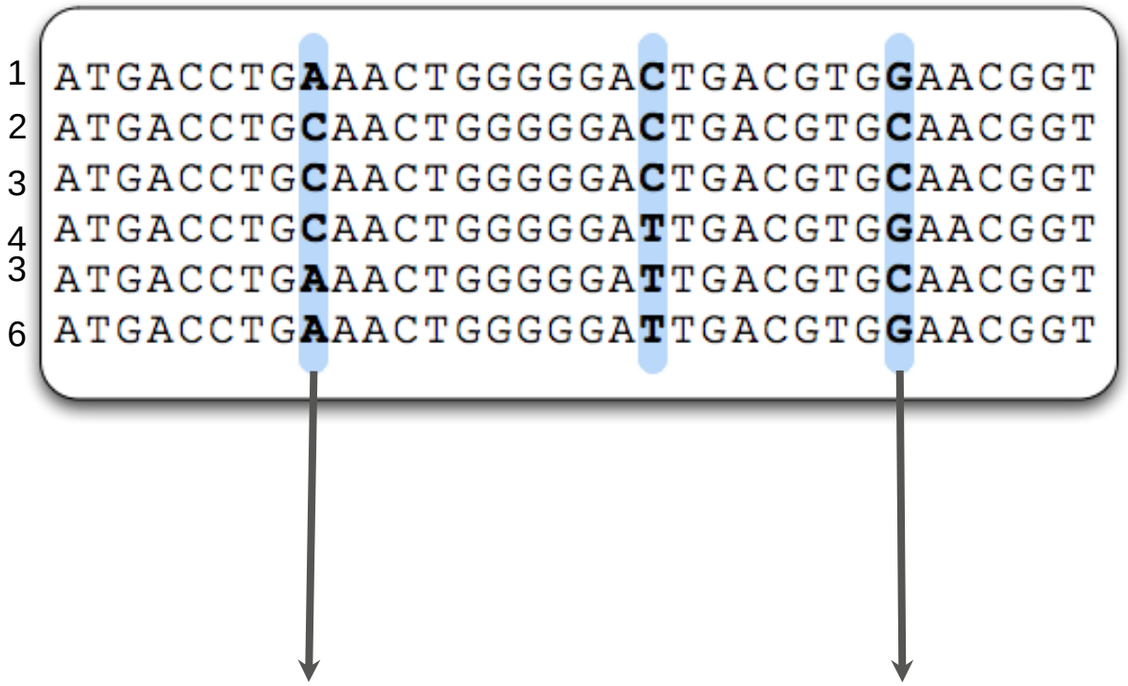
Joint work with Oliver Stegle[2] and Neil Lawrence[1]

*1. Department of Neuro- and Computer Science, University of Sheffield, UK*
*2. Department of Empirical Inference, Max Planck Institutes Tubingen, Germany*

# eQTL mapping

statistical technique with the goal to identify causal associations between variable genetic loci and the expression levels of individual genes.

SNPs

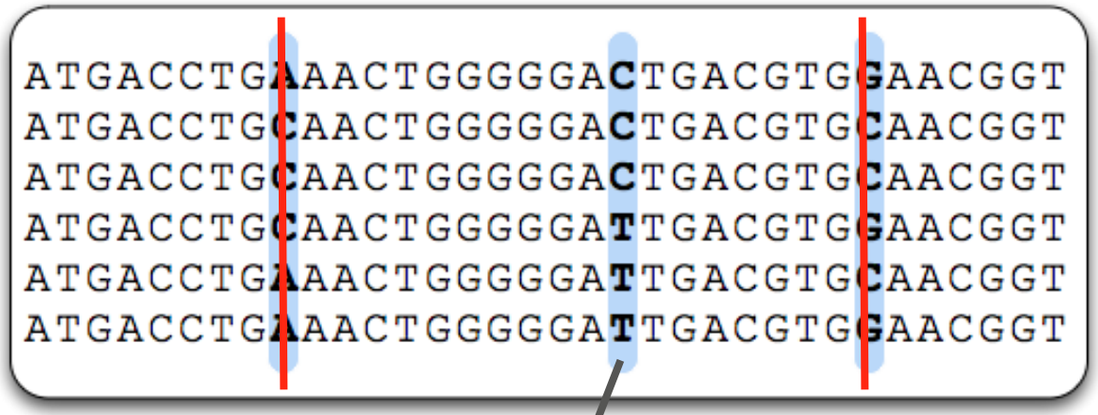



Gene expression

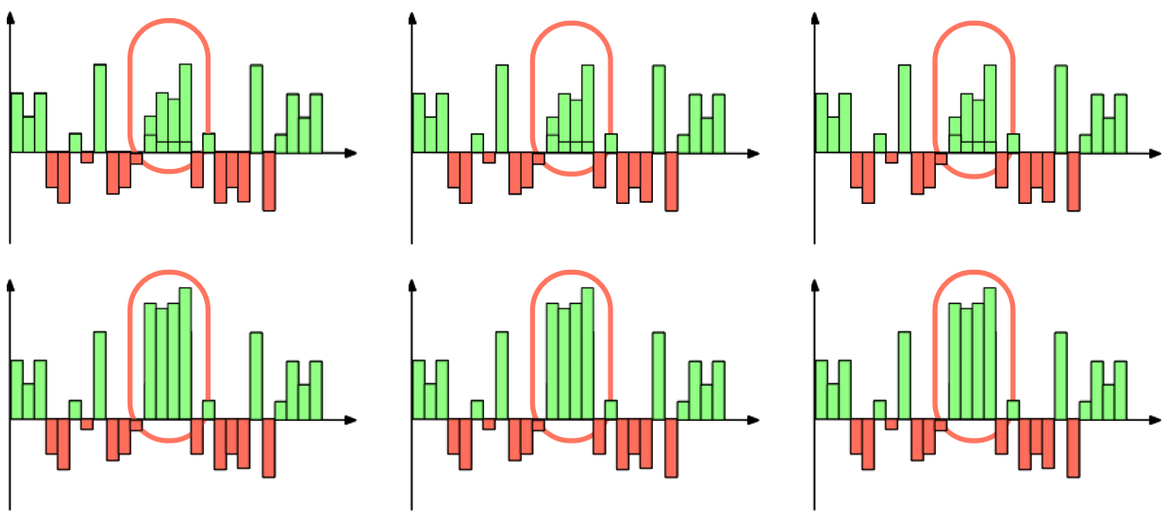
# eQTL mapping

Confounders introduce artifactual correlation in the expression levels of set of genes

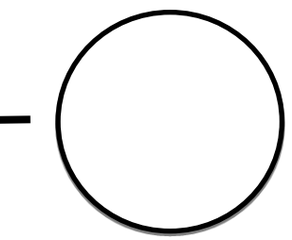
SNPs




Gene expression



Confounder



# The good news

Some of these confounders are known

- gender
- age
- ethnicity

# The bad news

Most of them are completely **unknown** or **unmeasurable**

- optical effects
- laboratory conditions
- in humans, exposure to diesel fumes
- in humans, the stress of taking exams

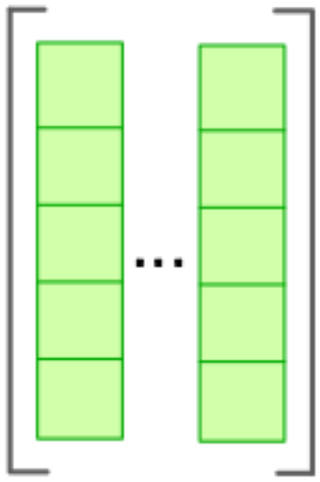
# PANAMA

## Probabilistic ANalysis of MicroArray data

A non-parametric probabilistic model, that:

- can account for both known and unknown confounders
- is based on a linear additive model
- greatly improves the quality of the results
- fast

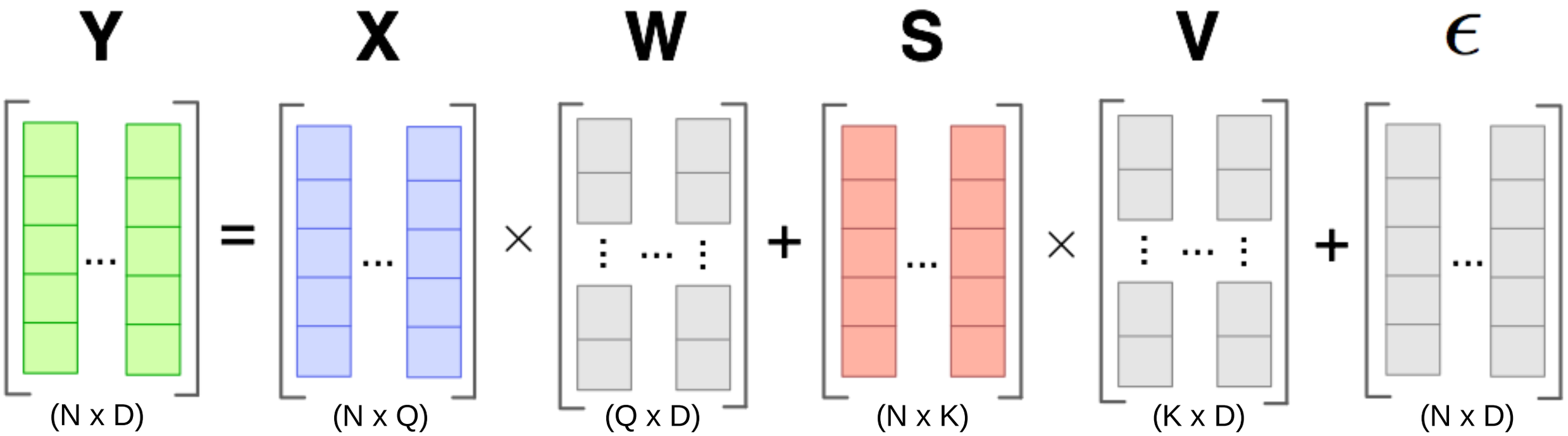
**Y**



(N x D)

Gene expression

# PANAMA


Y
X
W
S
V
ϵ
=
×
+
×
+
(N x D)
(N x Q)
(Q x D)
(N x K)
(K x D)
(N x D)
Gene expression
Latent Confounders
Genotype
Noise

The likelihood is:

$$P(\mathbf{Y} \mid \mathbf{W}, \mathbf{X}, \mathbf{S}) = \prod_{j=1}^{D} N(\mathbf{y}_j \mid \mathbf{W}\mathbf{x}_j + \mathbf{V}\mathbf{s}_j, \sigma^2 \mathbf{I}).$$

If we put spherical Gaussian priors over **V** and **W**,

$$P(\mathbf{W}) = \prod_{i=1}^{D} N(\mathbf{w_i} | \mathbf{0}, \alpha_w \mathbf{I})$$

$$P(\mathbf{V}) = \prod_{i=1}^{D} N(\mathbf{v_i} | \mathbf{0}, \alpha_v \mathbf{I})$$

We can obtain the marginal likelihood

$$P(\mathbf{Y} \mid \mathbf{X}) = \prod_{j=1}^{D} N(\mathbf{y}_j \mid \mathbf{0}, \mathbf{C}),$$

Where

$$\mathbf{C} = \alpha_w \mathbf{X}\mathbf{X}^\top + \alpha_v \mathbf{S}\mathbf{S}^\top + \sigma^2 \mathbf{I}.$$

We can determine the parameters  and the latent variables X from the data by maximum likelihood

$$\{\hat{\boldsymbol{\theta}}, \hat{\mathbf{X}}\} = \arg\max_{\boldsymbol{\theta}, \mathbf{X}} P(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta}).$$

# PANAMA's action is divided into two phases:

1. it learns a **dictionary** of latent variables that capture the main components of confounding variation
2. for each pair gene-SNPs it refits the weight parameters

# Association testing

We determine the presence or the absence of an association by comparing two models

$$\mathbf{C} = \alpha_w \mathbf{X}\mathbf{X}^\top + \alpha_v \mathbf{S}\mathbf{S}^\top + \sigma^2 \mathbf{I}.$$

The significance cutoff is determined by computing the positive False Discovery Rate (Storey, 2003)

$$\mathrm{pFDR}_{k,d} = \frac{\pi_0 \cdot p(\mathbf{y}_d \,|\, \mathcal{H}_0)}{\pi_0 \cdot p(\mathbf{y}_d \,|\, \mathcal{H}_0) + \pi_1 \cdot p(\mathbf{y}_d \,|\, \mathcal{H}_1)}.$$

# Experimental results

# Simulated dataset

**80** diploid individuals, **100** SNPs with a minor allele frequency of 0.4, **400** genes

| | FDR 0.01 | | FDR 0.05 | | FDR 0.1 | |
|---|---|---|---|---|---|---|
| | *cis* | *trans* | *cis* | *trans* | *cis* | *trans* |
| linear | 0.16 | 0.11 | 0.16 | 0.15 | 0.16 | 0.17 |
| SVA | 0.74 | 0.53 | 0.74 | 0.53 | 0.74 | 0.54 |
| ICE | 0.85 | 0.30 | 0.85 | 0.35 | 0.85 | 0.36 |
| PANAMA | 0.83 | 0.72 | 0.85 | 0.73 | 0.85 | 0.74 |

# PANAMA-ARD

We want to allow an individual weighting of the latent variables.

$$\mathbf{C} = \alpha_w \mathbf{X}\mathbf{X}^\top + \alpha_v \mathbf{S}\mathbf{S}^\top + \sigma^2 \mathbf{I}.$$

We constrain **X** to be orthonormal ($\mathbf{X}^\top \mathbf{X} = \mathbf{I}$)
and modify the structure of the covariance

$$\mathbf{C} = \mathbf{X}\mathbf{M}\mathbf{X}^\top + \alpha_v \mathbf{S}\mathbf{S}^\top + \sigma^2 \mathbf{I}.$$

Where M is a matrix

$$\mathbf{M} = \begin{pmatrix} \alpha_{w_1} & 0 & \cdots & 0 \\ 0 & \alpha_{w_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_{w_n} \end{pmatrix}$$

# Experimental results (again)

# Simulated dataset

**80** diploid individuals, **100** SNPs with a minor allele frequency of 0.4, **400** genes

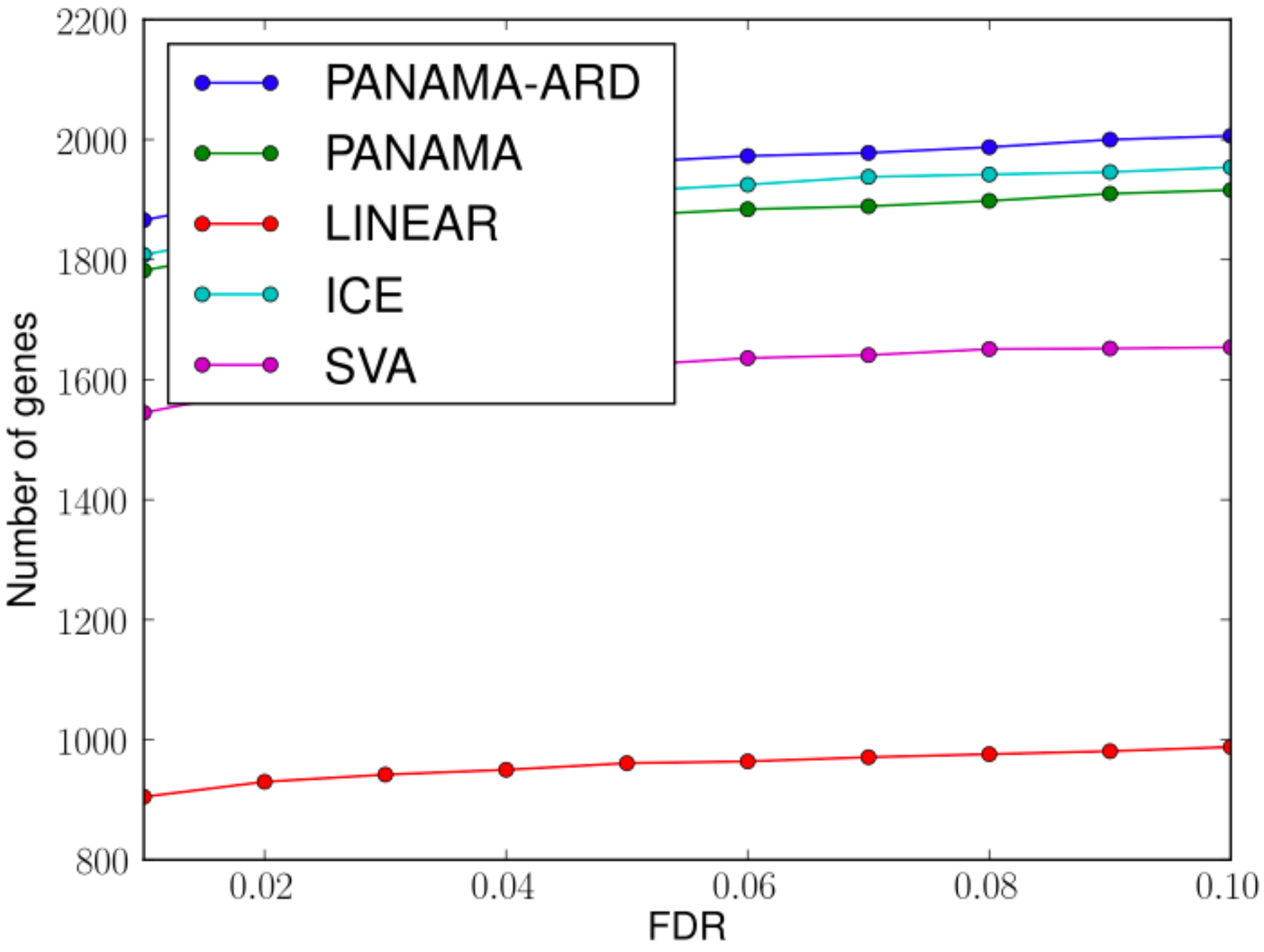
| | FDR 0.01 | | FDR 0.05 | | FDR 0.1 | |
|---|---|---|---|---|---|---|
| | *cis* | *trans* | *cis* | *trans* | *cis* | *trans* |
| linear | 0.16 | 0.11 | 0.16 | 0.15 | 0.16 | 0.17 |
| SVA | 0.74 | 0.53 | 0.74 | 0.53 | 0.74 | 0.54 |
| ICE | 0.85 | 0.30 | 0.85 | 0.35 | 0.85 | 0.36 |
| PANAMA | 0.83 | 0.72 | 0.85 | 0.73 | 0.85 | 0.74 |
| PANAMA-ARD | **0.87** | **0.80** | **0.89** | **0.81** | **0.89** | **0.82** |

# Yeast dataset

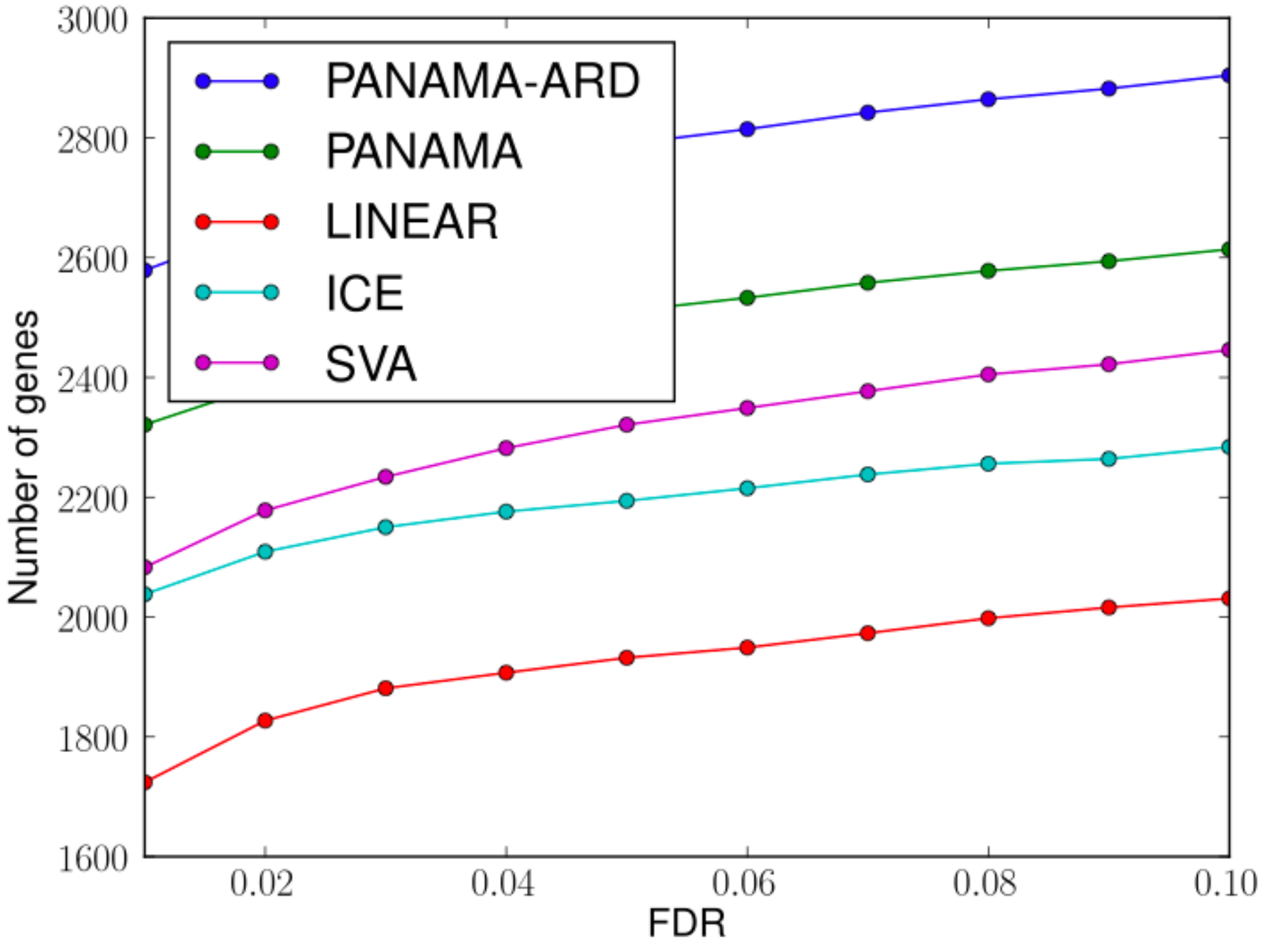
*Smith and Kruglyak,*
*Gene-environment  interaction in yeast gene expression.*
*PLoS biology, 2008*

- genotypes and expression profiles of 108 yeast segregants
- grown in two environmental conditions: sugar and ethanol
- very strong environmental influence
- **known**, but **not included** in the model

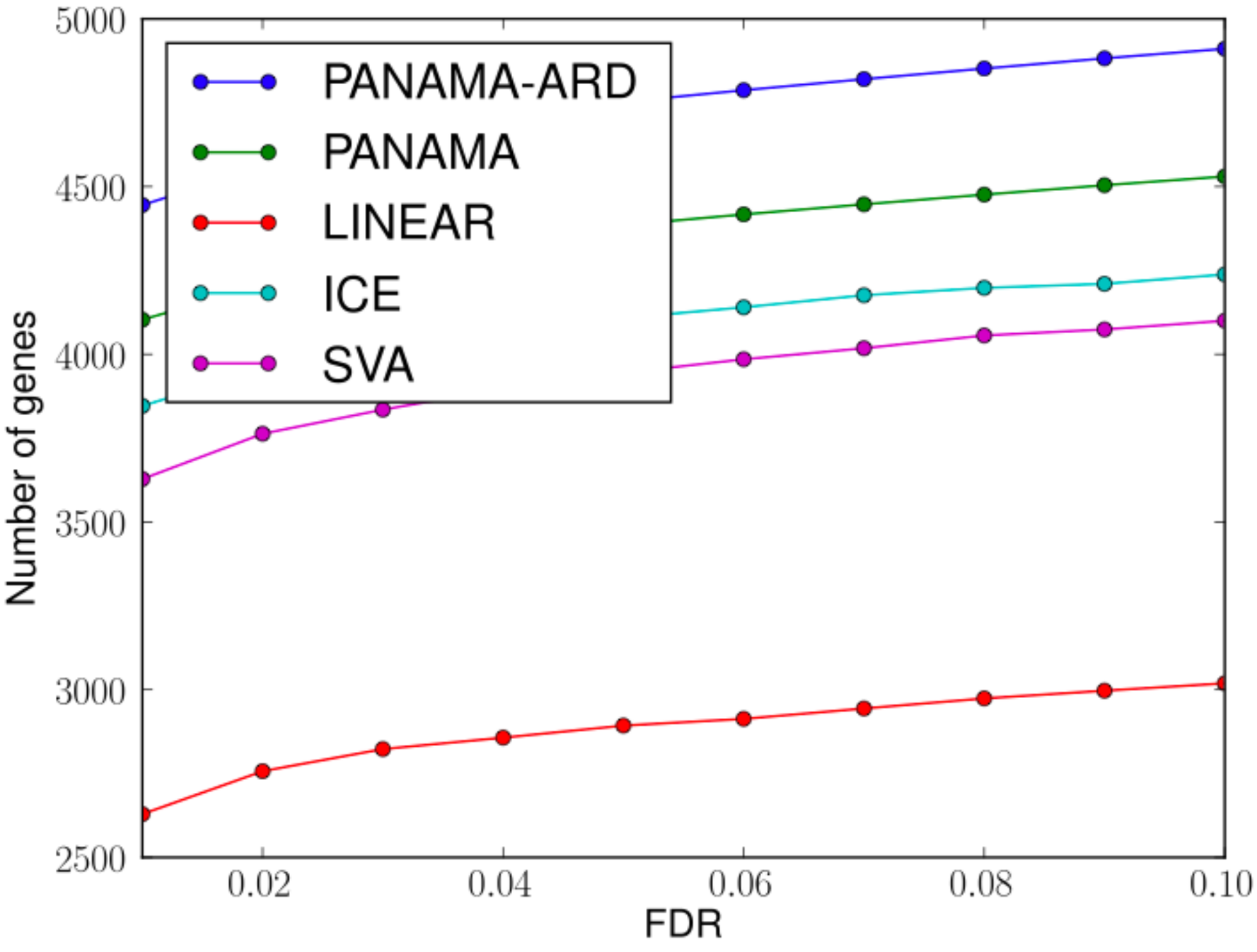
# Cis associations



# Trans associations

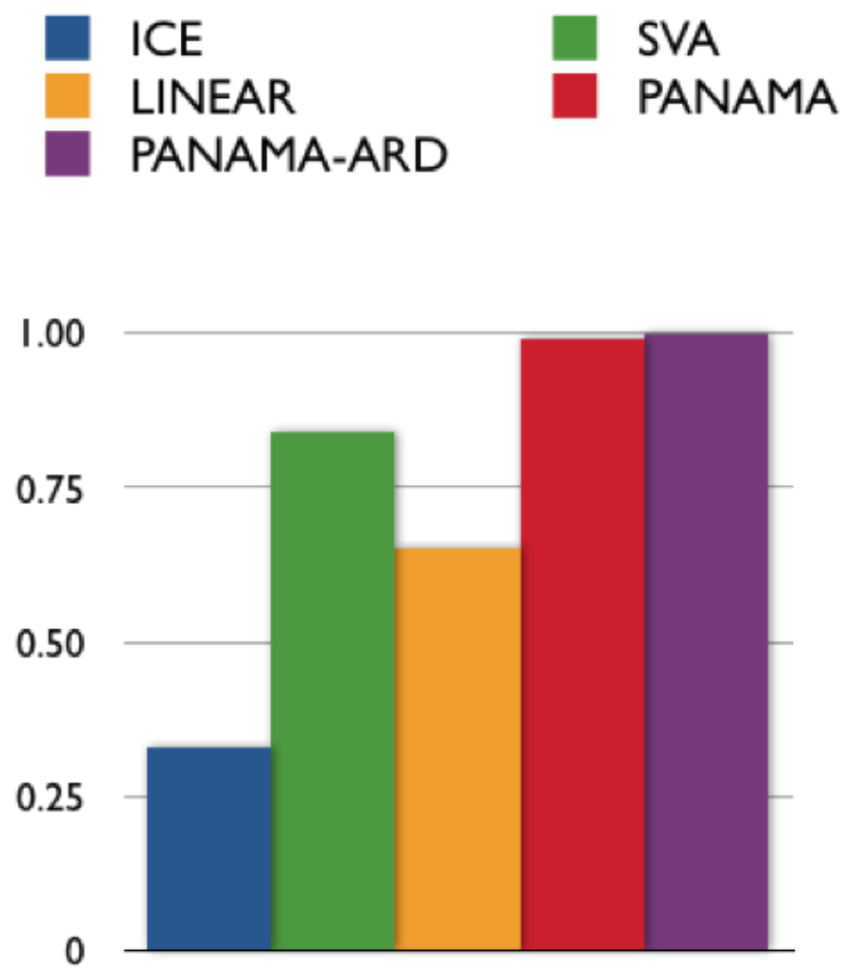


# Overall associations



# Validation

We include the environmental condition as a known covariate and measure the “overlap” between the calls made in the two settings



# Conclusions

- Confounding factors are a serious threat to the significance of eQTL studies and an accurate modeling is necessary
- the predominant assumption of a global set of confounders is clearly suboptimal
- PANAMA-ARD, trough the individual reweighting of the inferred confounding factors is able to greatly improve the results of eQTL association studies
- Additional work is needed to improve the performance