Batch Effects

Solutions to avoid batch effects and remove bias

Batch effects

The problem: Batch effects and other technical artifacts can **seriously** bias and obscure signal in high-throughput experiments.

Potential solutions:

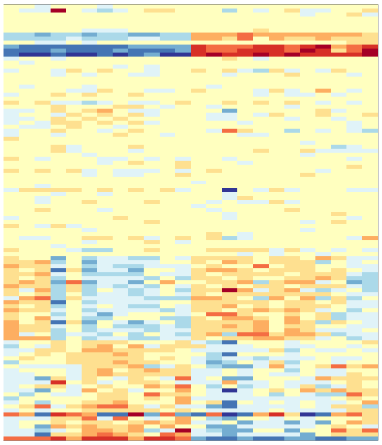
Record information:
Report date, reagent changes, personnel changes, etc.

Good experimental design:
Balance with respect to batches, etc.

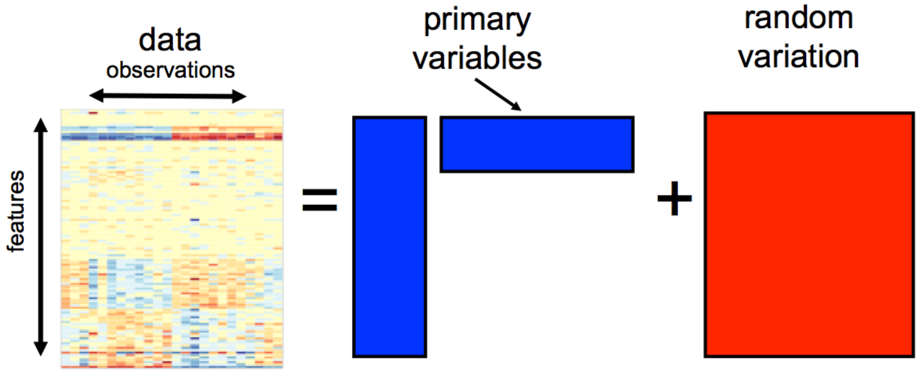
Statistical correction:
Simple regression when variables are known, and surrogate variable analysis (for example) when they unknown or uncertain.

12 males and females, 2 months, 109 genes

	Female	Male
June 2005	3	9
October 2005	9	3



Decomposing variability



data

observations

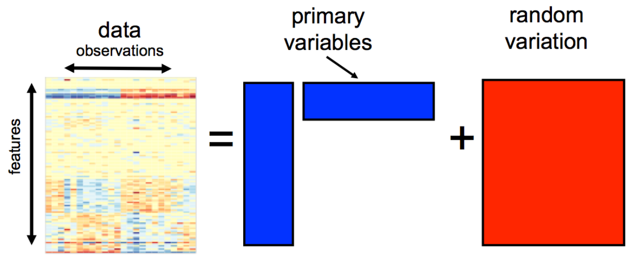
features

primary variables

random variation

$$Y_{m,n} = \beta_{m,p} X_{p,n} + \varepsilon_{m,n}$$

We model sex using this parametrization

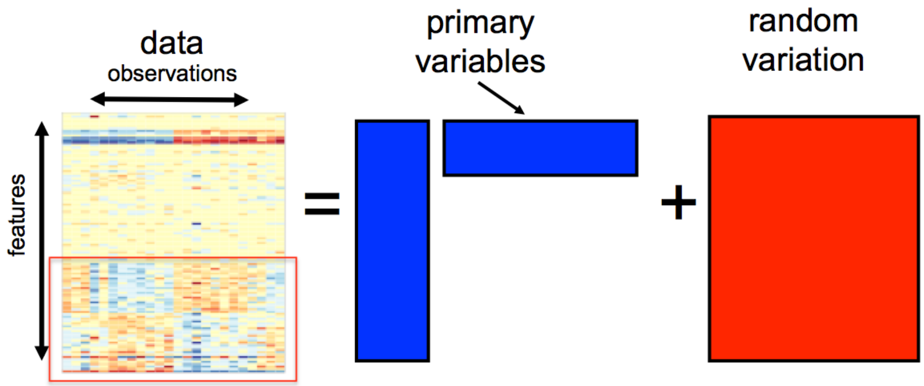


$$Y_{m,n} = \beta_{m,p} X_{p,n} + \varepsilon_{m,n}$$

$$X_{p \times n} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ & \vdots \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ & \vdots \\ 1 & 1 \end{pmatrix}^T$$

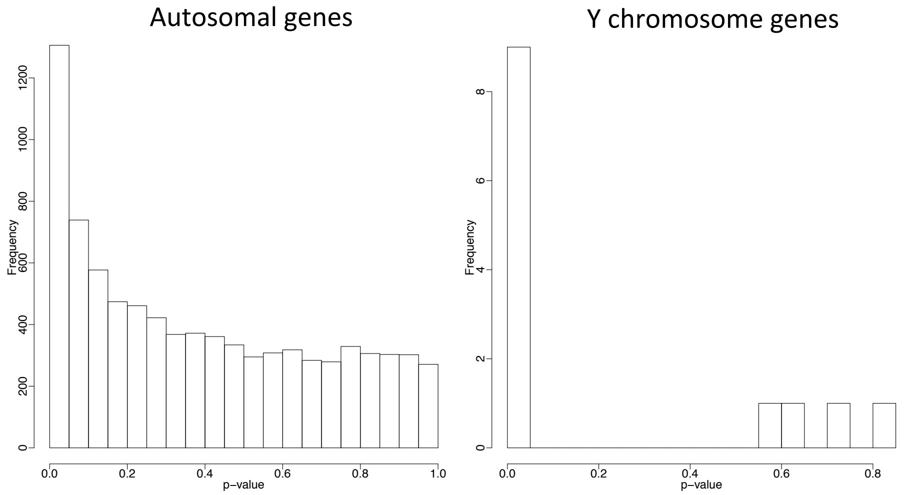
$$\beta_{m \times p} = \begin{pmatrix} \beta_{1,0} & \beta_{1,1} \\ & \vdots & \\ \beta_{m,0} & \beta_{m,1} \end{pmatrix}$$

This model does not account for batch

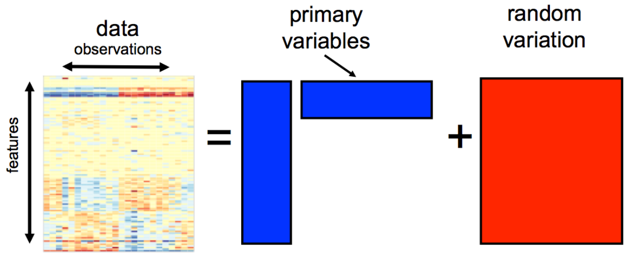


$$Y_{m,n} = \beta_{m,p} X_{p,n} + \varepsilon_{m,n}$$

P-values



We can include a term for batch

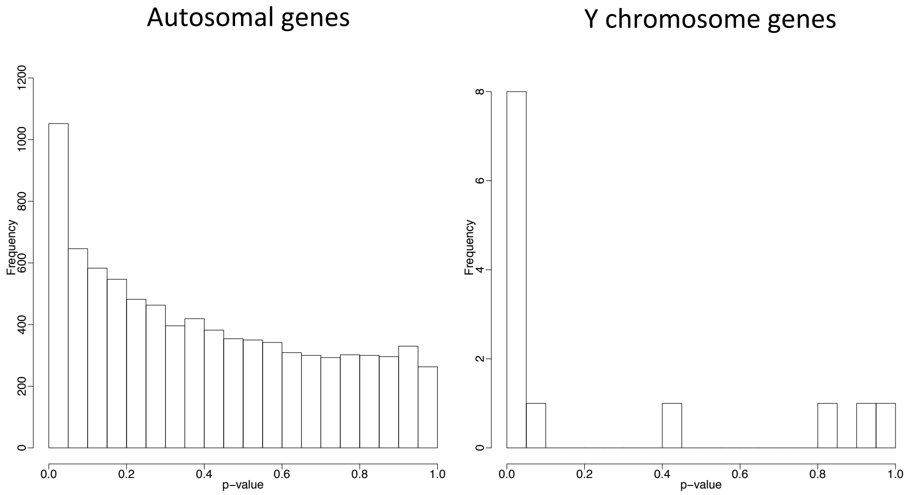


$$Y_{m,n} = \beta_{m,p} X_{p,n} + \varepsilon_{m,n}$$

$$\beta_{m \times p} = \begin{pmatrix} \beta_{1,0} & \beta_{1,1} & \beta_{1,2} \\ & \vdots & \\ \beta_{m,0} & \beta_{m,1} & \beta_{m,2} \end{pmatrix}$$

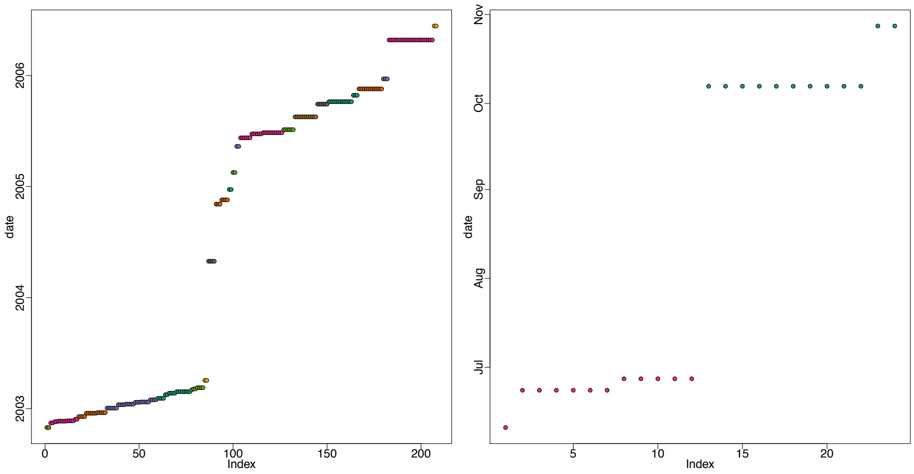
$$X_{p \times n} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ & \vdots & \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ & \vdots & \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}^T$$

P-values after adjusting for batch

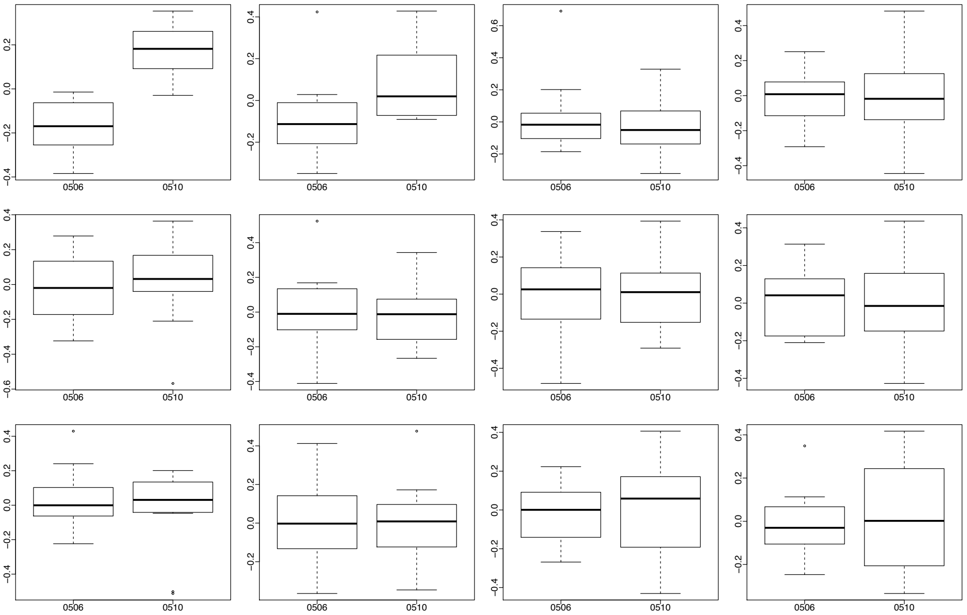

Autosomal genes

Y chromosome genes

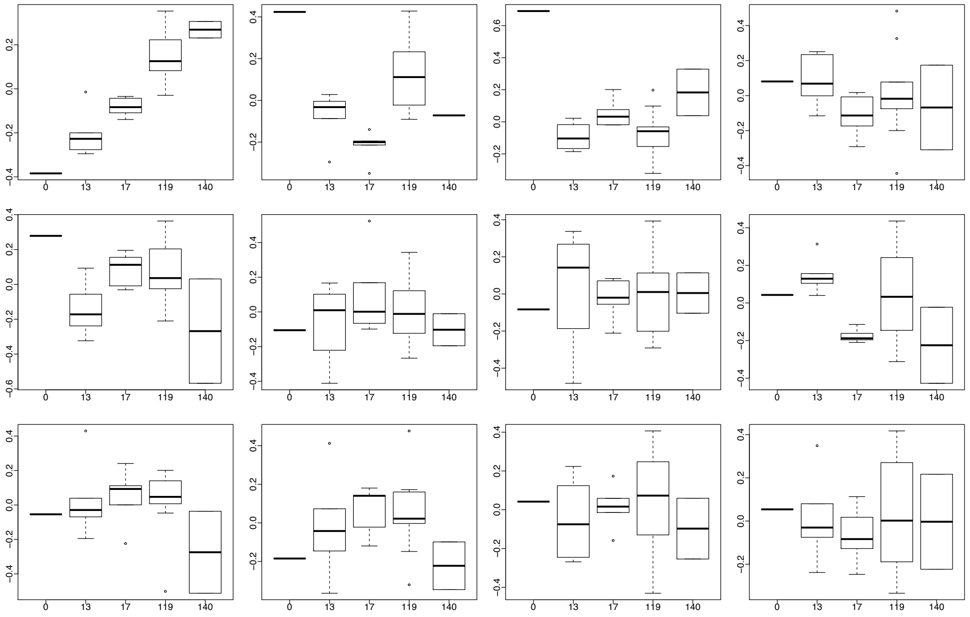
Do we know what the batches are?



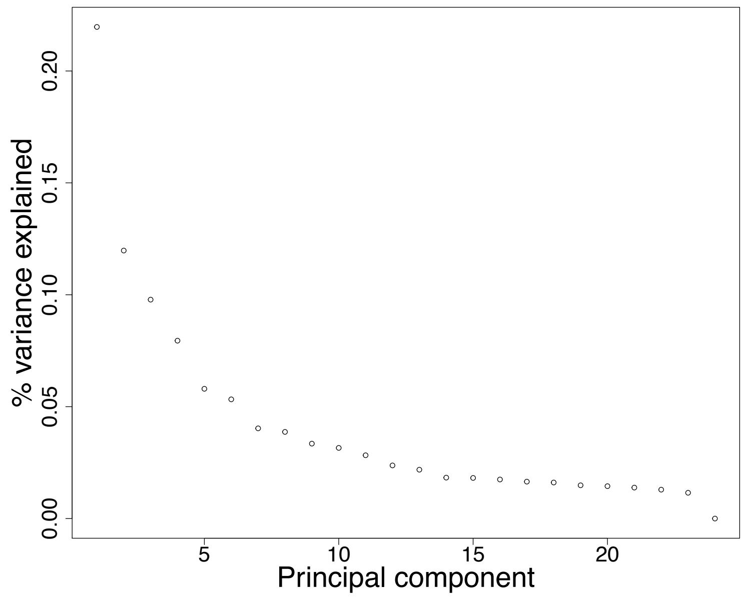
How many eigenvectors explain batch?



How many eigenvectors explain batch?



Variability explained



Batch surrogates

The first k columns give us estimates of surrogates

$$Y_{m\times n} = U_{m\times k} D_{k\times k} V'_{k\times n} + \varepsilon_{m\times n}$$

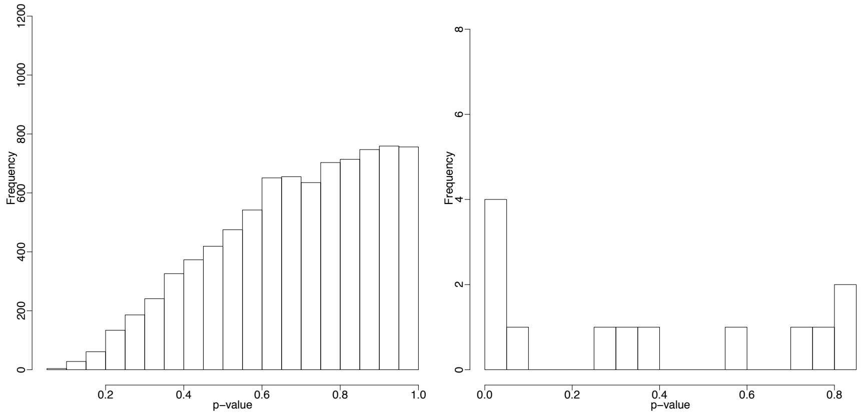
which can also be written like this:

$$Y_{m\times n} = \gamma_1 V_1' + \cdots + \gamma_k V_k' + \varepsilon_{m\times n}$$

P-values after regressing out 6 PCs

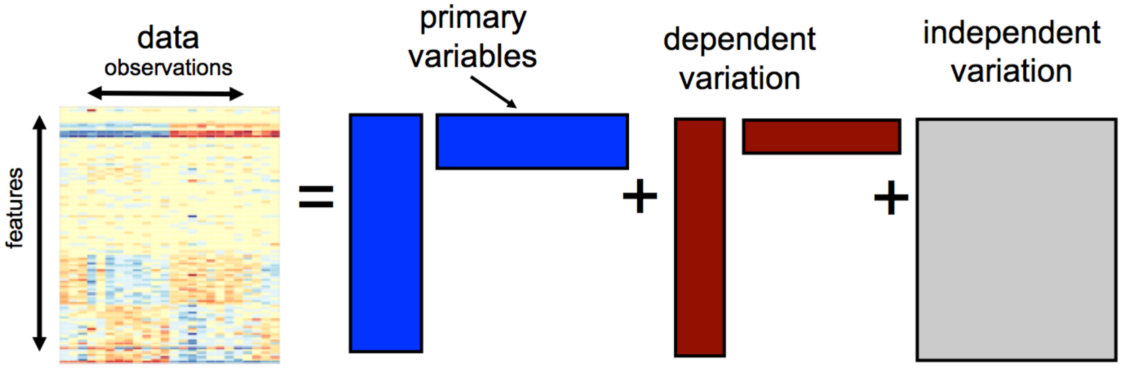
Autosomal genes

Y chromosome genes



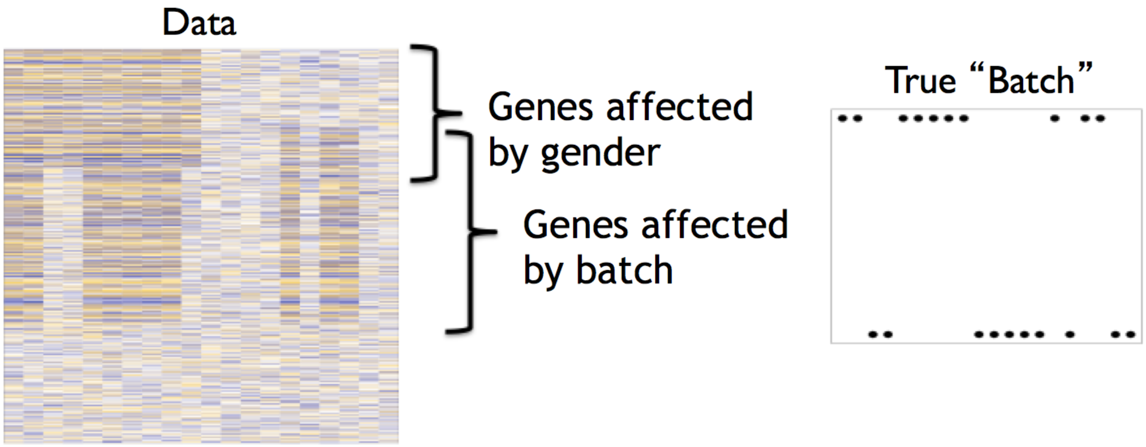
We washed away the signal!

SVA fits this model



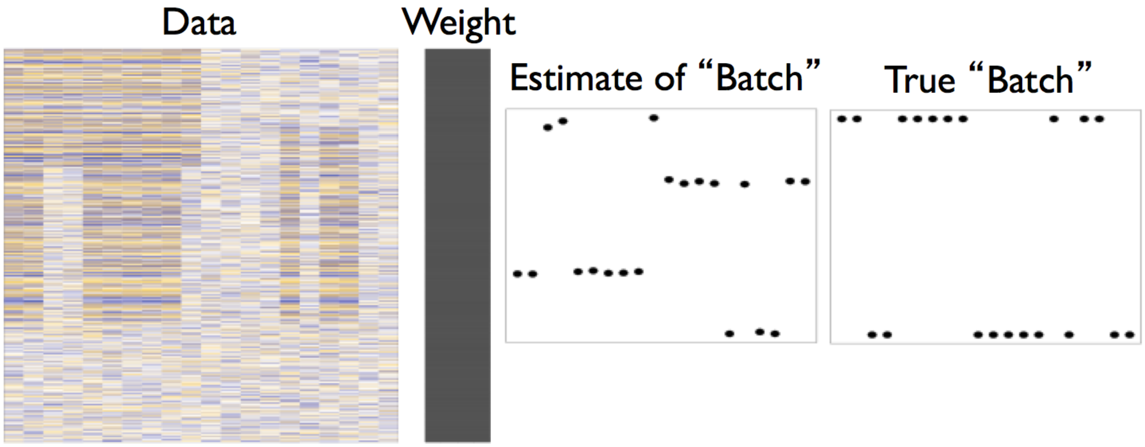
$$Y_{m\times n} = \beta_{m\times p}X_{p\times n} + \alpha_{m\times k}W_{k\times n} + \varepsilon_{m\times n}$$

Surrogate variable analysis



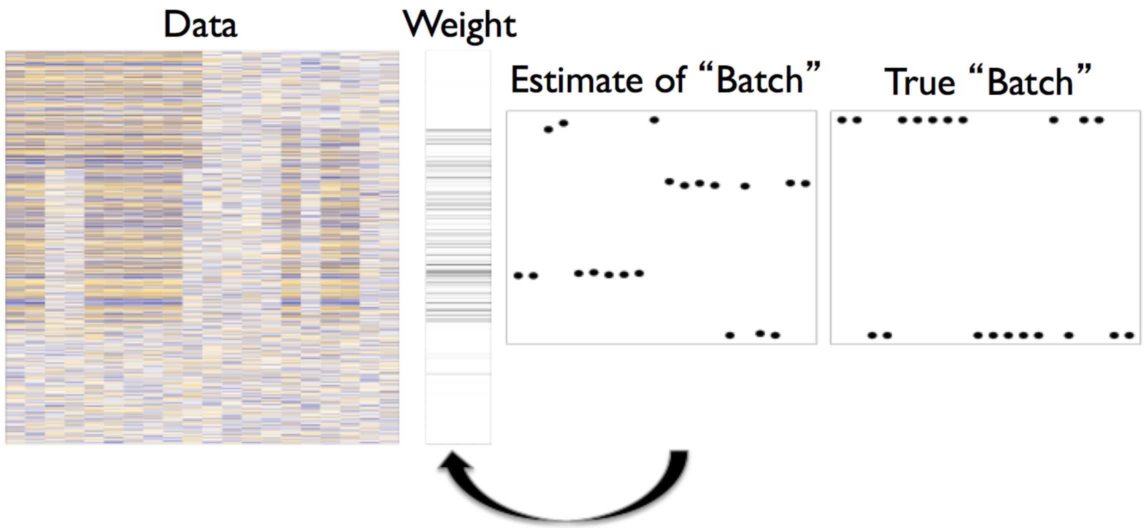
[JL]

Surrogate variable analysis



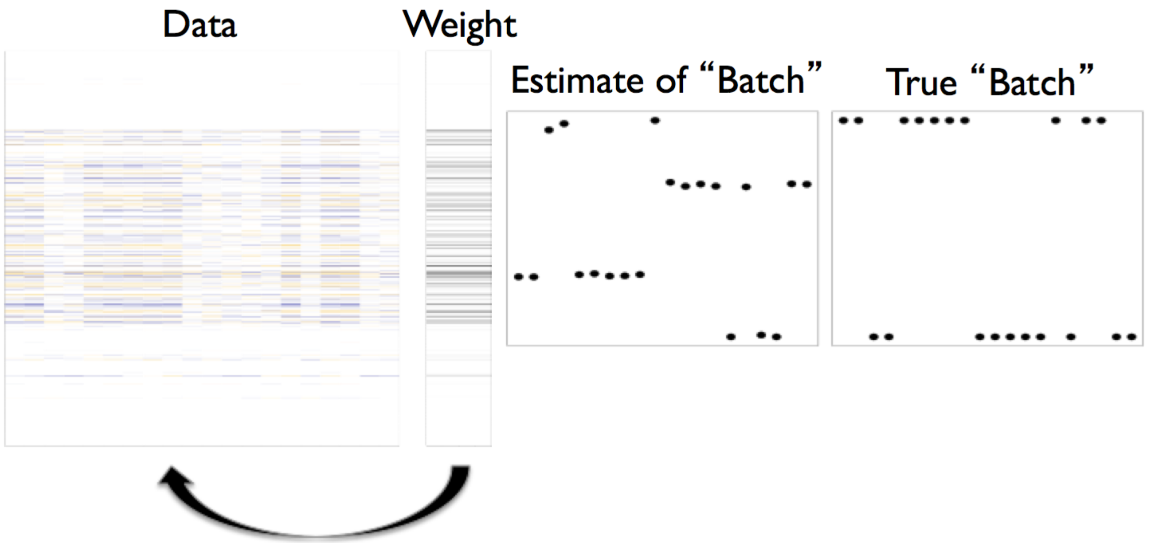
[JL]

Surrogate variable analysis



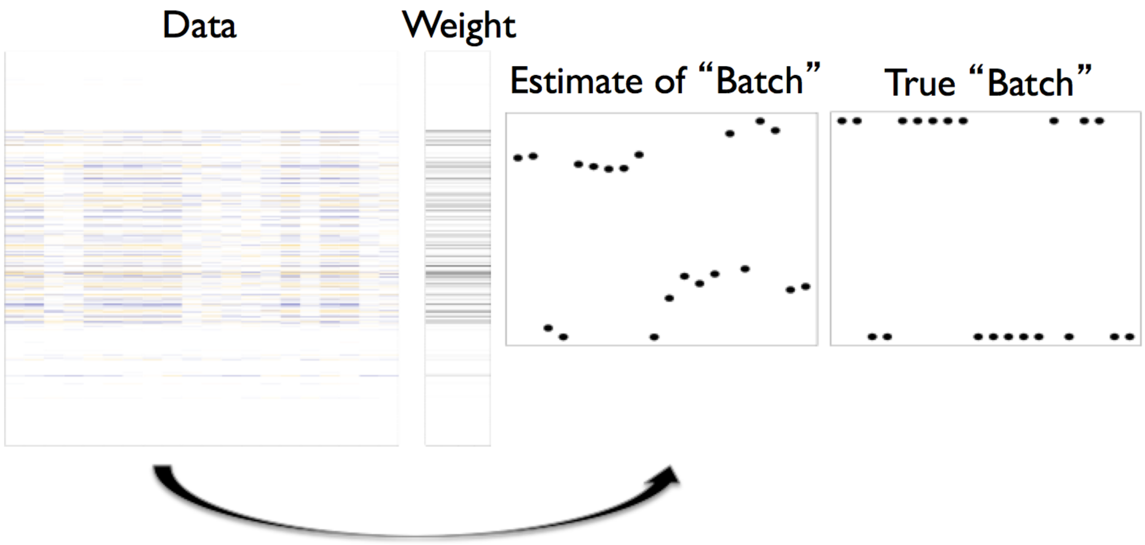
[JL]

Surrogate variable analysis



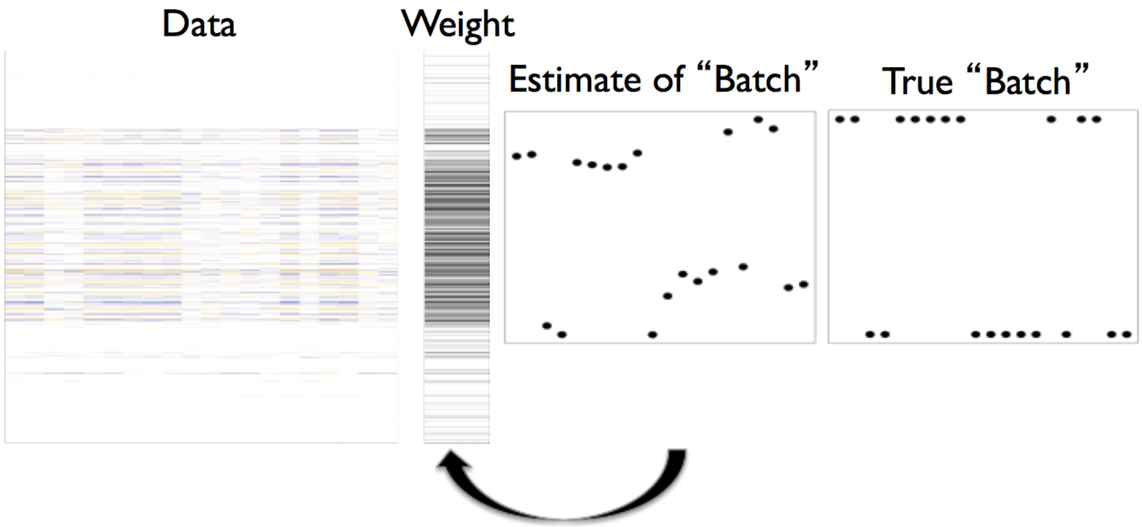
[JL]

Surrogate variable analysis



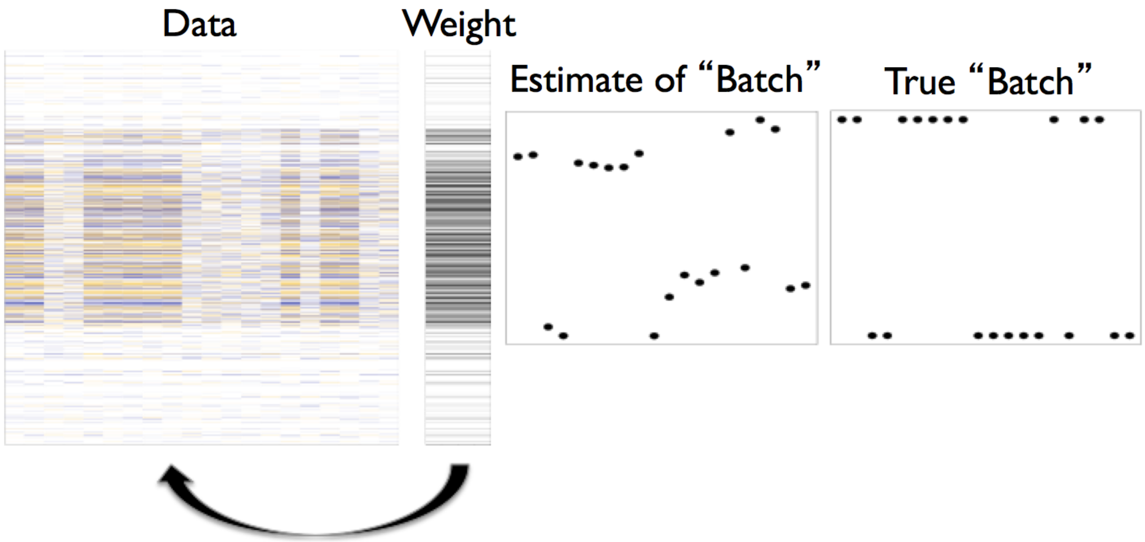
[JL]

Surrogate variable analysis



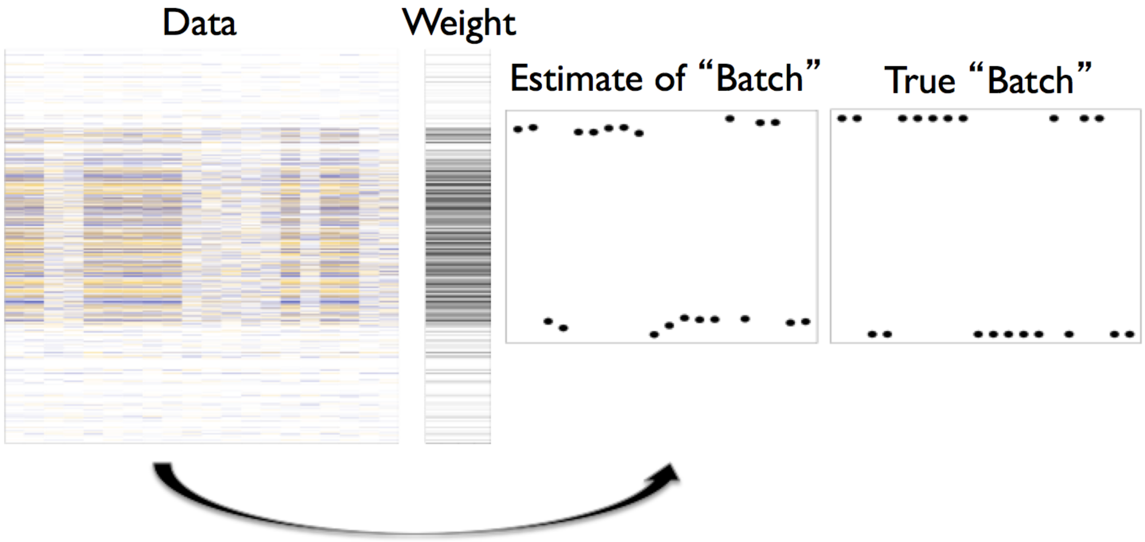
[JL]

Surrogate variable analysis



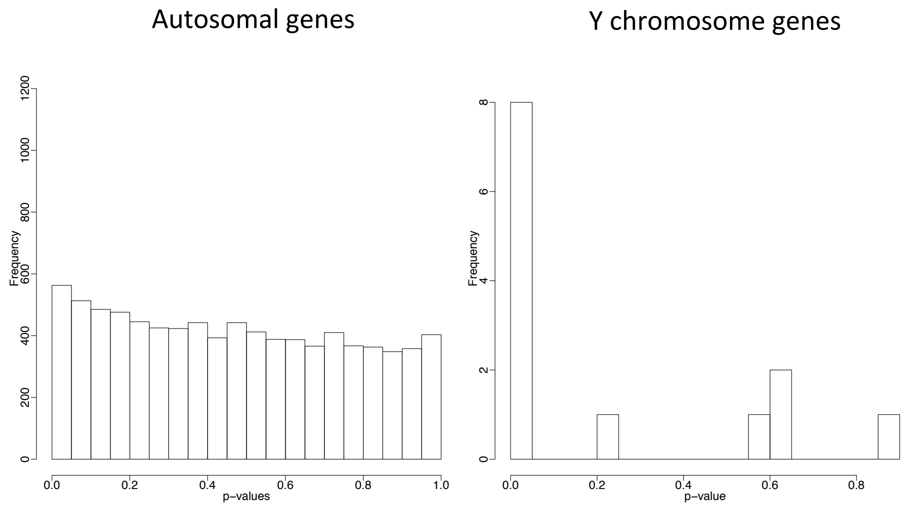
[JL]

Surrogate variable analysis

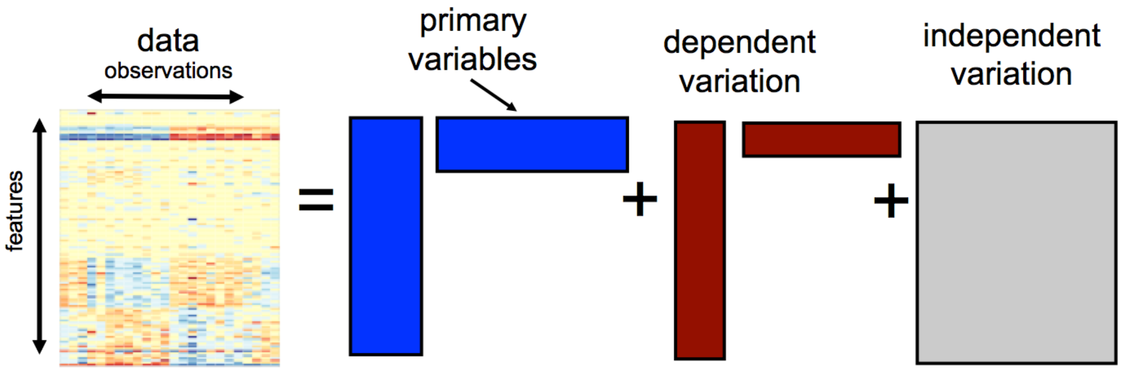


[JL]

P-values after SVA



SVA fits this model



$$Y_{m\times n} = \beta_{m\times p}X_{p\times n} + \alpha_{m\times k}W_{k\times n} + \varepsilon_{m\times n}$$

Decomposed data