We have seen that microarray expression data suffers from “unwanted variation”.

We have discussed a number of preprocessing algorithms intended to increase the signal-to-noise in such data.

But sometimes (unwanted) variation remains.
Batch effects

8 normal samples
color: processing date
Batch effects

Although specific other technical effects violate the assumptions of normalization methods, there may even exacerbate technical artefacts in some cases. Normalization procedures affect different genes in different ways. In many cases we have found that these can lead to erroneous biological conclusions, supporting the conclusions of differential expression is nearly always across biological groups, as real biological variables is to examine the gene expression patterns of variation, such as ozone levels, laboratory-effects have shown strong laboratory-specific downstream consequences of batch effects. In high-throughput studies it is commonly used to account for batch effects. In a typical experiment these are associated with DNA preparation groups. For example, multiple laboratory-expressions of technical artefacts. For example, the presence/absence of CIS was strongly confounded with the presence or absence of CIS was strongly associated with the processing date, as reported in microarray studies focusing on copy the processing date was confounded with the processing date. Hierarchical cluster analysis separated the sTCC samples according to the presence/absence of CIS. Hierarchical cluster analysis separated the sTCC samples according to the presence/absence of CIS. Hierarchical cluster analysis separated the sTCC samples according to the presence/absence of CIS.

One gene

Clustering

© 2010 Genetics Express
Batch effects

Batch effects: unwanted variation remaining after normalization.

Often associated with processing date or processing batch.

Often thought to be technical, but can be biological (later).

Two extreme situations:
(1) Batch effect is orthogonal to comparison of interest. This will increase noise and dilute signal, but otherwise ok.
(2) Batch effect is confounded with comparison of interest. This will make any conclusion highly circumspect (and likely wrong).
Sources of Heterogeneity

External Factors (like environment)  Genetics/Epigenetics  Technical Factors
The Effect of Heterogeneity

Color = Environment
(Idaghdour et al. 2008)

Color = Processing Year
(Cheung et al. 2008)

Color = Allele
(Brem et al. 2005)
A Simple Simulated Example

Independent E

Dependent E
Gene by Gene Model

expression = $b_0 + b_1 \times \text{group} + \text{noise}$

Test whether $b_1 = 0 \iff \text{T-test for gene I}$

Calculate a P-value
Null P-Value Distributions

Independent E

Dependent E
Null P-Value Distributions

**Correlation**

\[ |\rho| = 0.40 \quad |\rho| = 0.31 \quad |\rho| = 0.10 \quad |\rho| = 0.00 \]

**Independent E**

**Dependent E**
False Discovery Rate Estimates

Independent E

Dependent E
Ranking Estimates

Independent $\mathbf{E}$

Dependent $\mathbf{E}$
Principal Components Analysis / Singular Value Decomposition

• A method to identify patterns in the data that explain a large percentage of the variation

• PCA and SVD have different mathematical goals but end up estimating the same thing

• First proposed for genomics by Alter et al. (2000) PNAS
Singular Value Decomposition

\[ \text{samples} \times \text{genes} = U \times D \times V^T \]

- Data: eigenarrays / left singular vectors / loadings
- U: singular vectors / principal components
- D: singular values
- V^T: eigengenes / right singular vectors / principal components
Properties of SVD

Columns of $V^T$/rows of $U$ are orthogonal and calculated one at a time.

Columns of $V^T$ describe patterns across genes.

Columns of $U$ describe patterns across arrays.

\[
\frac{d_i^2}{\sum_{i=1}^{n} d_i^2} \text{ is the percent of variation explained by the } \text{ith column of } V
\]
1 Pattern $1^{st}$ SV

$1^{st}$ Column of $U$ $1^{st}$ Column of $V^T$
2 Patterns, 1\textsuperscript{st} SV

\[ \text{1\textsuperscript{st} Column of } U \times \text{1\textsuperscript{st} Column of } V^T \]
Surrogate Variable Analysis

The Data

Pr(\text{Group & Batch})

Estimate of Batch

True Batch
Surrogate Variable Analysis

The Data

True Batch

Pr(!Group & Batch)

Estimate of Batch

True Batch
Surrogate Variable Analysis

The Data

Pr(!Group & Batch)

Estimate of Batch

True Batch
Surrogate Variable Analysis

The Data $\Pr(\text{Group} \& \text{Batch})$

Estimate of Batch

True Batch
Surrogate Variable Analysis

The Data  Pr(!Group & Batch)

Estimate of Batch

True Batch
Surrogate Variable Analysis

The Data

Pr(\text{Group} \& \text{Batch})

Estimate of Batch

True Batch
SVA Adjusted Gene by Gene Model

expression = $b_0 + b_1 \times \text{group} + \text{surrogates} + \text{noise}$

Test whether $b_1 = 0$

Calculate a P-value
Ranking Estimates

Independent E

Dependent E

Dependent E + IRW-SVA
Methods

(IRW-) SVA by Leek and Storey.

RUV by Gagnon-Bartsch and Speed

Many methods are small modifications of SVA.

The idea behind RUV is to use control (genes / probes) to estimate batch effects. Controls could be probes unaffected by biology (negative controls) or genes known to be differentially expressed.
An example of “biological” batch effect

Changes in cell type composition.

Expression changes associated with Age or Cell cycle.