Preprocessing and normalization

The path from colored specks to priceless data
Intensities are not just mRNA concentrations

- Tissue contamination
- Clone identification and mapping
- Image segmentation
- RNA degradation
- PCR yield, contamination
- Signal quantification
- Amplification efficiency
- Spatial effects
- Reverse transcription efficiency and specificity
- Other issues related to array manufacturing

Example of spatial effects on microarrays

In theory, the spatial location of a given spot should matter little, since the locations were randomly selected.

But in reality, things like the distribution of solvent over the array surface and the quality of washing, have their say on the matter.

Example data:
- Raw data
- Spatial bias estimate
## Two degrees of variation

<table>
<thead>
<tr>
<th>Array-specific variation</th>
<th>Gene-specific variation:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Amount of RNA in the biopsy</strong></td>
<td><strong>PCR yield / DNA quality</strong></td>
</tr>
<tr>
<td><strong>Efficiencies of:</strong></td>
<td><strong>Spotting efficiency,</strong></td>
</tr>
<tr>
<td>– RNA extraction</td>
<td>– Spot size</td>
</tr>
<tr>
<td>– Reverse transcription</td>
<td><strong>Cross-/unspecific hybridization</strong></td>
</tr>
<tr>
<td>– Labeling</td>
<td></td>
</tr>
<tr>
<td>– Photodetection</td>
<td></td>
</tr>
</tbody>
</table>

**Systematic** | **Stochastic**
Stochastic noise can be dealt with by a t-test...

PCA Plot of 34 patients, 8973 dimensions (genes) reduced to 2
...like we will see later today

PCA for 100 most significant genes reduced to 2 dimensions
Sources of variation

<table>
<thead>
<tr>
<th>Array-specific variation:</th>
<th>Gene-specific variation:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Systematic</strong></td>
<td><strong>Stochastic</strong></td>
</tr>
</tbody>
</table>

- Similar effect on many measurements
- Corrections can be estimated from data
- Too random to be explicitly accounted for
- “noise”

Normalization → Statistical testing
Calibration = Normalization = Scaling
Nonlinear normalization

Unnormalized data

q spline normalized data
The Qspline method

From the empirical distribution, a number of quantiles are calculated for each of the channels to be normalized (one channel shown in red) and for the reference distribution (shown in black).

A QQ-plot is made and a normalization curve is constructed by fitting a cubic spline function.

As reference one can use an artificial “median array” for a set of arrays or use a log-normal distribution, which is a good approximation.
Non-linear normalization

- Raw data
- After intensity normalization
- Spatial bias estimate
- After spatial normalization
The really cool thing about R...

...is all the nice libraries out there

The BioConductor packages encompasses many very useful methods for microarray analysis

– Including the *qspline* method, and other normalization algorithms

Check out the [www.bioconductor.org](http://www.bioconductor.org) website!
Exercise in normalization

• Download the normalization exercise and open the pdf document
  – Please consider that you learn more if you read the commands thoroughly before you copy-and-paste