Course on Microarray Gene Expression Analysis

::: Normalization methods and data preprocessing

The probe-level data must be cleaned and processed to obtain biologically meaningful measurement:

✓ **Background correction**: eliminate signals due to non-specific binding;

• **Normalization**: make multiple arrays comparable;

• **Preprocessing**: Scale linearization, summarization (calculating ratios, log scale transformation), missing values imputation, etc.
Normalization techniques are needed to accurately interpret the results:

- **Remove aberrations** and outliers due to non-biological factors.
- **Reveal the patterns** and outliers of actual biological factors.
Why normalize?

Sources of variation between multiple high-density oligonucleotide arrays:

- Biological (e.g., diseased vs. normal)
- Non-biological:
  - Total RNA preparation, transcription, amplification, abundances;
  - Sample labeling differences;
  - Hybridization parameters;
  - Scanner differences;
  - Image analysis;
  etc.

Goal: make multiple arrays comparable.
Changes in expression are independent of abundance.
Rare transcripts are as likely to change in response to a given stress as common ones.

Most transcripts are not differentially expressed in response to a given stress.
Expression ratio of typical spot: tumor/control = 1

Range of abundance begins at 0.
“Negative transcripts” = error measurement

Outliers are biologically relevant.
Average cases are less interesting.
::: Normalization plots

**Boxplots**

IQR: InterQuartile Range

"the difference between the 75th percentile and 25th percentile"

Whiskers: In most cases represents 1.5 times the box width. Can be customized.

Outliers

- <1.5 IQR below 0-25 quartile
- >1.5 IQR over 75-100 quartile
MA plots (ratio vs intensity)

The MA-plots show the relationship between $A$, the "average signal" $[(\log R + \log G)/2]$, where $R$ is the background subtracted red [mean of F635 - median of B635] and $G$ the background subtracted green [mean of F532 - median of B532], and $M$, the log [base 2] differential ratio: $\log(R/G)$.

$$M = \log R - \log G = \log \left(\frac{R}{G}\right)$$

$$A = \left(\frac{\log R + \log G}{2}\right) = \log \sqrt{R \cdot G}$$
Normalization Methods and Data Preprocessing

::: Normalization plots

Two main categories for study:

- Intra-slide normalization (*within array*).
  Normalizes expression values to make intensities consistent within each array.

- Inter-slide normalization (*between array*).
  Normalizes expression values to achieve consistency between arrays.

  Normalization between arrays is usually, but not necessarily, applied after normalization within arrays (an exception in VSN method).
Normalization methods

Within-array normalization.

- Median: Subtracts the weighted median from the M-values for each array.
- **Lowess**: LОcally WEighted Scatterplot Smoothing, 2-channel microarrays. Utilize a locally weighted polynomial regression of the intensity scatterplot in order to obtain the calibration factor.
- Print-tip Lowess: Regional Lowess, 2-channel spotted microarrays.
- Control: Fits a global loess curve through a set of control spots and applies that curve to all the other spots.
- Robust spline: Normalize the M-values for a single microarray using robustly fitted regression splines and empirical Bayes shrinkage.

Between-array normalization.

- Scale: Simply to scale the log-ratios to have the same median-absolute-deviation across arrays
- MAS5: Affymetrix
- RMA: Affymetrix
- Cyclic Loess: Affymetrix, CodeLink
- **Quantiles**: Ensures that the intensities have the same empirical distribution across arrays and across channels. Affymetrix, Agilent
- Invariant set normalization: Affymetrix
- VSN: Variance Stabilization Normalization
Global Lowess Normalization

• Lowess is a technique for fitting a smoothing curve to a dataset. An intensity-dependent normalization.

• Predicted loess value is subtracted from the data to decrease the standard deviation and place the mean log ratio at 0.

• Lowess normalization may be applied to a two-color array expression dataset.

• All samples in the dataset are corrected independently.

• Lowess normalization can be applied to complete or incomplete datasets.

Global normalized data \{ (M,A)_{n=1..5184} \}

\[ M_{\text{norm}} = M - c(A) \]

where \( c(A) \) is an intensity dependent function.
Within-array methods

Print-tip Lowess normalization

- One loess line for each block (print-tip block).

\[
M_{\text{print-tip normalized}} = M_p - c_p(A);
\]

where \(c_p(A)\) is an intensity dependent function for print tip \(p\).

Print-tip normalized data \(\{M(A)\}_{k=1,16}\):
• Significant variation in the distribution of intensity values across arrays
• Transforms the distribution of probe intensities to be same across arrays
• Final distribution is the average of each quantile across chips
Quantiles normalization

Sort each column of original matrix

Take average across rows

Substitute each value to corresponding row average

Unsort columns of matrix to original order

\[
\begin{pmatrix}
1 & 5 & 3 & 5 \\
2 & 1 & 6 & 7 \\
3 & 2 & 2 & 6 \\
4 & 6 & 1 & 8
\end{pmatrix}
\longrightarrow
\begin{pmatrix}
1 & 1 & 1 & 5 \\
2 & 2 & 2 & 6 \\
3 & 5 & 3 & 7 \\
4 & 6 & 6 & 8
\end{pmatrix}
\longrightarrow
\begin{pmatrix}
2 & 3 & 4.5 & 6 \\
\end{pmatrix}
\longrightarrow
\begin{pmatrix}
2 & 2 & 2 & 2 \\
3 & 3 & 3 & 3 \\
4.5 & 4.5 & 4.5 & 4.5 \\
6 & 6 & 6 & 6
\end{pmatrix}
\longrightarrow
\begin{pmatrix}
2 & 4.5 & 4.5 & 2 \\
3 & 2 & 6 & 4.5 \\
4.5 & 3 & 3 & 3 \\
6 & 6 & 2 & 6
\end{pmatrix}
\]
Variance Stabilization Normalization (VSN)


- VSN transformation = \( \text{asinh}(ai + bi \times y) \)
- Well-defined and meaningful close to 0.
- Original intensities may be negative.
Log Transformation

- Common technique used for two-color arrays (one-color as well)
- Log ratio transformations convert data to a linear scale
- $M = \log_2(Cy5/Cy3)$
- $A = \log_2(Cy5 \times Cy3) \times 0.5$
Imputing Missing Values

- Fill with zeros: Replace missing values by zeros.
- Fill with row average: Replace missing values by the row average.
- Fill with row median: Replace missing values by the row median.
- K-Nearest Neighbors (KNN) impute: Replace missing values by the average value of the K nearest patterns.
KNN Imputing

• Machine learning algorithm.
• Usually euclidean (or Manhattan) metric, confined to the columns for which that gene is NOT missing
• $K < \text{number of samples in smallest class.}$

<table>
<thead>
<tr>
<th>Class A</th>
<th>Class B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4</td>
<td>1 2 3 4</td>
</tr>
<tr>
<td>5 6 7 8</td>
<td>5 6 7 8</td>
</tr>
<tr>
<td>9 10 11 12</td>
<td>9 10</td>
</tr>
<tr>
<td>13 14 15 16</td>
<td>13 14 15 16</td>
</tr>
</tbody>
</table>
Other Parameters

• Filter Flat Patterns (no relevant differences between classes):
  - By number of peaks
  - By root mean square
  - By standard deviation

• Standardize Patterns: Subtracts the mean of the pattern and divide it by the standard deviation (z-score).
  Allows comparison of observations from different distributions (recommended)

\[ z = \frac{x - \mu}{\sigma} \]

Indicates how many standard deviations an observation is above or below the mean.
Thanks!

Visit us:
http://ubio.bioinfo.cnio.es/ubio/

Gonzalo Gómez
ggomez@cnio.es