

# Tutorial 7: Normalization Methods



# Normalization methods for Affymetrix data

- MAS 5

Can be applied on single array

- RMA

- DChip

- Plier

- Plier +16

Can be applied on multiple arrays

# Normalization methods for Affymetrix data - continued

Database Contents

- ANONYMOUS
  - Affy\_Rat\_MAS5\_only
    - Gene Lists
      - D0\_T12\_B\_a D0\_T12\_B[Biotin]
        - MAS5 {D0\_T12\_B\_a} [file: ...]
          - Mean/Median Scaling, ...
      - D0\_T12\_B\_b D0\_T12\_B[Bio]
        - MAS5 {D0\_T12\_B\_b} [file: ...]
          - Mean/Median Scaling, ...
      - D0\_T12\_C\_a D0\_T12\_C[Bio]
        - MAS5 {D0\_T12\_C\_a} [file: ...]
          - Mean/Median Scaling, ...
      - D0\_T12\_C\_b D0\_T12\_C[Bio]
        - MAS5 {D0\_T12\_C\_b} [file: ...]
          - Mean/Median Scaling, ...
      - D0\_T12\_D\_a D0\_T12\_D[Bio]
        - MAS5 {D0\_T12\_D\_a} [file: ...]
          - Mean/Median Scaling, ...
      - D0\_T12\_D\_b D0\_T12\_D[Bio]
        - MAS5 {D0\_T12\_D\_b} [file: ...]
          - Mean/Median Scaling, ...
      - D2\_T12\_B\_a D2\_T12\_B[Bio]
        - MAS5 {D2\_T12\_B\_a} [file: ...]
          - Mean/Median Scaling, ...
      - D2\_T12\_B\_b D2\_T12\_B[Bio]

Right-click the selected data, choose "Convert affy cel files to probe sets."

View data set(s) as wide spreadsheet - datasets side by side

Export

Convert affy cel files to probe sets

Mixed scatterplot

Virtual array images for data

Actual array images for data <<Dev. Only>>

Rank intensity plots for data

BarChart

Create gene list by data filtering...

Analysis

Quality Control

Normalize...

Duplicate data sets

Copy data sets for pasting elsewhere

Studies

Tree options...

Selection of methods

Select methods:

- MAS5
- RMA
- DChip
- Plier
- qPlier 16

Quantile normalization for Plier

OK Cancel

# Normalization methods for Affymetrix data - continued

The converted probe sets files will be shown under the experiment.

# Other normalization methods

The following methods are for one and/or two channels

- Lowess
  - Total intensity norm
  - Linear Lowess
  - GenePix Mean Log ratio
  - Mean/Median scaling
  - Ref Avg Comp
  - Quantile
- For 2 channel only
- For either 1 or 2 channel
- For 1 channel only



# Normalize Data

Select the data, right-click, choose “Normalize...”

The screenshot shows a software interface with a file tree on the left and a context menu on the right. The file tree is organized as follows:

- Strain\_mice\_two\_Channel
  - Gene Lists
  - Strain A Q380 Strain A - mice 1 [Cy5] | Reference [Cy3]
    - raw data {Strain A Q380} [Cy5] | Reference [Cy3]
    - LOWESS, ri=3
  - Strain A Q381 Strain A - mice 1 [Cy5] | Reference [Cy3]
    - no-name {Strain A Q381} [Cy5] | Reference [Cy3]
    - LOWESS, ri=3
  - Strain A Q382 Strain A - mice 1 [Cy5] | Reference [Cy3]
    - no-name {Strain A Q382} [Cy5] | Reference [Cy3]
    - LOWESS, ri=3
  - Strain B Q385 Strain B - mice 1 [Cy5] | Reference [Cy3]
    - no-name {Strain B Q385} [Cy5] | Reference [Cy3]
    - LOWESS, ri=3
  - Strain B Q386 Strain B - mice 1 [Cy5] | Reference [Cy3]
    - no-name {Strain B Q386} [Cy5] | Reference [Cy3]
    - LOWESS, ri=3
  - Strain B Q387 Strain B - mice 1 [Cy5] | Reference [Cy3]
    - no-name {Strain B Q387} [Cy5] | Reference [Cy3]
    - LOWESS, ri=3

The context menu is open over the 'raw data {Strain A Q380} [Cy5] | Reference [Cy3]' item. The menu items are:

- View data set(s) as wide spreadsheet - datasets side by side
- Export
  - Scatter plots for data
  - Mixed scatterplot
  - MA plots for data
  - Virtual array images for data
  - Actual array images for data <<Dev. Only>>
  - Rank intensity plots for data
  - BarChart
- Create gene list by data filtering...
- Analysis
- Quality Control
- Normalize...** (highlighted by the mouse cursor)

# Lowess

Lowess is for two channel data only.

**Select Normalization Method**

LOWESS

For 2 channel data only

The Lowess normalization method performs a robust locally weighted regression on the log ratio (M) vs. log geometric average (A) spot data, using each spot's locally estimated M value for spot by spot correction of log ratio values. Thus the Lowess method differs from many other normalization techniques because it is able to correct intensity (A) dependent ratio biases in an intensity-specific way. Visually this amounts to "straightening out" a curved A-M plot.

When the final local regression estimate Mfit of the log ratio has been calculated for a spot having channel values c1 and c2, the corrected ratio c1'/c2' is determined by

$$\log(c1'/c2') = \log(c1/c2) - Mfit$$

*(fitted M value become new zero point of M values at this avg intensity)*

which we can rewrite as

$$c1'/c2' = (c1/c2) * 1/b^{Mfit}$$

*(where b is the logarithm base for computed M-values)*

To present this correction in two channel format, we spread the ratio correction factor reciprocally to both channels, so that the same ratio correction is achieved:

$$c1' = c1 * \sqrt{1/b^{Mfit}}$$
$$c2' = c2 / \sqrt{1/b^{Mfit}}$$

Filters

Gene List: Include only spots with genes in gene list <all genes>

OK Cancel

subtract backgrounds Yes  
smoothing factor 0.2  
robustness iterations 3  
delta

# Total Intensity Ratio Normalization

Select Normalization Method

Total Intensity Norm.

Total Intensity Ratio Normalization

subtract backgrounds Yes

Total intensity ratio normalization is for two channel only.

Method (2 ch only)

- 1) Compute sum of each channels' intensities, optionally subtracting backgrounds
- 2) Let  $r$  be the ratio of these sums, ie  $(\text{sum ch1 vals})/(\text{sum ch2 vals})$
- 3) scale factor for first channel is  $1/\sqrt{r}$ , for channel 2 is  $\sqrt{r}$

Properties

- Ratio of sums computed for the normalized data should be 1.0
- Each spot is adjusted so that the ratio of channel values is  $1/r$  times it's unnormalized value.

Filters

Gene List: Include only spots with genes in gene list <all genes>

OK Cancel

# Mean/Median Scaling Normalization

Select Normalization Method

Mean/Median Scaling

## Mean/Median Scaling Normalization

Method (for either 1 or 2 channel data)

- Multiply each value by  $T/m$  where  $m$  is the mean or median of the channel data and  $T$  is the target mean/median value option (default is 1000). Channels are scaled separately in the case of two channel data.

Properties

- Normalized channel data will have a mean/median matching the target value option.

subtract backgrounds Yes

scaling Median

target value 1000.0

include flagged spots No

Filters

Gene List: Include only spots with genes in gene list <all genes>

OK Cancel

For one or two channel data

# GenePix Mean Log Ratio Normalization

This method is for two channel data only

**Select Normalization Method**

GenePix Mean Log Ratio Norm.

**Method (2 ch only)**

- 1) Compute channel ratios after respective background subtraction if specified (which channel is the numerator vs which is denominator doesn't matter.)
- 2) If the **exclude ratio limit** parameter M has been specified non-zero, then spots are ignored whose ratios don't lie between  $1/M$  and  $M$
- 3) take log of remaining ratios (base doesn't matter, will cancel out)
- 4) Apply anti-log to the avg of these log ratios, to get r
- 5) scale factor for numerator channel is  $1/\sqrt{r}$ , for denominator channel is  $\sqrt{r}$  (applied after background subtraction, if specified)

**Properties**

- After normalization, the average of the log of channel ratios is 0, corresponding to  $r = 1.0$
- Each spot value is adjusted such that the ratio of channels is  $1/r$  times it's unnormalized value

**Filters**

Gene List: Include only spots with genes in gene list <all genes>

subtract backgrounds Yes

exclude ratio limit 10.0

OK Cancel

# Linear & Lowess Normalization

The screenshot shows a software dialog box titled "Select Normalization Method". The selected method is "Linear&Lowess". The dialog is divided into a main description area on the left and a configuration area on the right.

**Linear&Lowess Normalization**

**Method (for 2 channel data only)**

- First, values for each channel are multiplied by  $T/m$  where  $m$  is the mean or median of the channel data and  $T$  is the target mean/median value option (default is 1000).
- Then a Lowess normalization is performed on the resulting scaled channel data.

**Configuration Parameters:**

- subtract backgrounds: Yes
- smoothing factor: 0.2
- robustness iterations: 3
- delta: (empty field)
- scaling: Geometric Mean
- target value: 1000.0
- include flagged spots for scaling: No

**Filters**

Gene List: Include only spots with genes in gene list <all genes>

Buttons: OK, Cancel

**Annotations:** A red oval highlights the text "Method (for 2 channel data only)". A red arrow points from a red-bordered box containing the text "For two channel data only" to the circled text.

# Quantile Normalization

Select Normalization Method

Quantile

## Quantile Normalization

**Method (1 ch only)**

All datasets are first sorted by ascending intensity value, after removing flagged values if the option is set to exclude them. All datasets are then trimmed to the length of the shortest dataset involved, by trimming the smallest (front) values from each. Then the least (front of list) value is taken from each dataset, and the arithmetic or geometric mean of these least values is written as the normalized intensity value for each of the front spots. Thus the same normalized intensity value is written for each dataset, however the spot (gene) that this intensity value is written for depends on the dataset. This process is repeated until the end of the datasets are reached.

**Properties**

- Each dataset normalized together via quantile normalization will have the same set of output intensity values.

subtract backgrounds Yes

include flagged spots Yes

mean type arithmetic

**Filters**

Gene List: Include only spots with genes in gene list <all genes>

OK Cancel

For one channel only

# Reference Average Comparison Normalization

For one channel only

**Select Normalization Method**

Ref Avg Comp

## Reference Average Comparison Normalization

**Method (1 ch only)**

The normalized intensity  $I'$  is given in terms of the original intensity  $I$  and the reference intensities  $R(k)$  by

$$I' = I / g$$

where

$$g = (R(1) * R(2) * \dots * R(N))^{(1/N)}$$

is the geometric mean of the reference intensities.

**Properties**

- $\log(I') = \log(I) - \text{Avg}(\log(R(k)))$ , where Avg is the usual arithmetic average.

subtract backgrounds Yes

include flagged spots No

**Highlight Reference Data Sets (Use Shift and/or Control keys for multiple selection)**

- Affy\_Rat\_MAS5\_only/D0\_T12\_B\_a[r10649]
- Affy\_Rat\_MAS5\_only/D0\_T12\_B\_b[r10650]
- Affy\_Rat\_MAS5\_only/D0\_T12\_C\_a[r10651]
- Affy\_Rat\_MAS5\_only/D0\_T12\_C\_b[r10652]
- Affy\_Rat\_MAS5\_only/D0\_T12\_D\_a[r10653]
- Affy\_Rat\_MAS5\_only/D0\_T12\_D\_b[r10654]
- Affy\_Rat\_MAS5\_only/D2\_T12\_B\_a[r10655]

Exclude reference datasets from being normalized

**Filters**

Gene List: Include only spots with genes in gene list <all genes>

OK Cancel