CSE 5559 – Translational Bioinformatics
Week 5
Two major technologies

- Affymetrix-like arrays
  - *single channel* (background-green, foreground-red)

- cDNA arrays
  - *two channel* (red, green, yellow)
Both
Reading Material – Chapter 3, Text
Two-color CDNA
Two-channel microarray

- Printed microarrays
- Long probe oligonucleotides (80-100) long are “printed” on the glass chip
- Comparative hybridization experiment

http://www.accesssexcellence.org/AB/GG/microArray.html
Affymetrix GeneChip

- silicon chip
- oligonucleotide probes lithographically synthesized on the array
Matching

RNA fragments with fluorescent tags from sample to be tested

RNA fragment hybridizes with DNA on GeneChip® array
Photolithography

**FIGURE 3.9:** Photolithographic fabrication of microarrays. Synthetic linkers modified with photochemical removable protecting groups are attached to a glass surface. Light is shed through a photolithographic mask to a specific area on the surface to produce a localized photodeprotection. The first of a series of hydroxyl-protected deoxynucleosides is incubated on the surface. In this example, it is the protected deoxynucleoside C. The surface of the array is protected again, and the array is ready for the next mask.
Photolithography

**FIGURE 3.10:** Photolithographic fabrications of microarrays. The second mask is applied and light is used to deprotect the areas that are designed to receive the next nucleoside (A). The fabrication process would generally require 4 masking steps for each element of the probes. Several steps later, each area has its own sequence as designed.
Fluorescence

Shining a laser light at GeneChip® array causes tagged DNA fragments that hybridized to glow

Non-hybridized DNA

Hybridized DNA
Recap

The diagram illustrates the process of gene expression:

1. **Transcription**: DNA is transcribed into pre-mRNA, which then undergoes RNA processing.
2. **RNA processing**: Intron sequences are removed, resulting in a mature mRNA.
3. **Translation**: The mature mRNA is translated into a polypeptide.
Probes

First generation expression arrays

Exon tiling arrays

Whole genome tiling arrays

Splicing arrays

Probe type legend
- Exon
- Intron
- Exon-Exon junction
- Exon-Intron boundary

Chromosome

Example gene

5’
Start codon

STOP codon
poly(A)

Tissue A
50 copies per cell
E1 E2 E3

Tissue B
1000 copies per cell
E1 E2 E3

Tissue A
E1 E2 E3

Tissue B
E1 E3

Identify novel transcribed regions in a single tissue sample

Novel exon

Detect exon skipping events:

E1 E2 E3

Exon junction probe

Detect alternate exon boundaries and intron retentions:

Alternate 5’ boundary
Intron retention
Exon boundary probe

Tissue A
E1 E2 I1 E2-I2 E3

Tissue B
E1 E2 I2 E2-I3 E3

Tissue A
E1 I1 E2 E2-I3 E3

Tissue B
E1 I2 E2 E3

Wexner Medical Center
Probe Selection

Chromosome

Example gene

5' Start codon

3' Stop codon Poly(A)

Probe type legend
Exon
Intron
Exon-Exon junction
Exon-Intron boundary

First generation expression arrays

Tissue A

50 copies per cell

Tissue B

1000 copies per cell

Wexner Medical Center
Probe selection

✓ 3’ bias – why? Think degradation
✓ Multiple probes for one region
✓ G-C content - Think binding

http://www.bcgsc.ca/people/malachig/htdocs/alexa_platform/alexa_arrays/intro.htm
Worth Visiting

http://www.arrayanalysis.org/
Probe Design

- Fabrication expense and frequency of error increases with probe length - **25 oligonucleotide probes are employed**

- Problem: cross hybridization

- Solution: introduce mismatched probe with one position (central) different with the matched probe. The difference gives a more accurate reading.
The principle of the Affymetrix technology. The probes correspond to short oligonucleotide sequences thought to be representative for a given gene. Each oligonucleotide sequence is represented by two probes: one with the exact sequence of the chosen fragment of the gene (perfect match or PM), and one with a mismatch nucleotide in the middle of the fragment (mismatch or MM). The hybridization occurs at a critical temperature – a single mismatch can totally inhibit hybridization. For each gene, the value that is usually taken as representative for the expression level of the gene is the average difference between PM and MM.
Affymetrix Genechips

- Each gene represented by 11-20 ‘probe pairs’
- Probe pairs are 3’ biased.
- ‘Probe Pair’ consists of Perfect Match (PM) and MisMatch (MM) probes.
- MM has altered middle (13th) base. Designed to measure non-specific binding (NSB).
Actually

Problem: cross hybridization

FIGURE 20.10: Affymetrix terminology and background zone weighting computation. A gene is represented by a number of perfect match (PM) and mismatch (MM) probe pairs. Each pair is composed of two cells. The background value is computed as a weighted sum of the nearby zone background values. The weight of a zone is inversely proportional to the distance to it. The background value for a given zone is calculated from the cells with the lowest 2% intensities in that zone.
Reading Data

cDNA array – ratio, log ratio

\[ T_i = \frac{R_i}{G_i} \quad \text{OR} \quad \log \text{ratio} = \log_2 \frac{R_i}{G_i} \]

Affymetrix array

Difference_{probe \, pair} = PM - MM

Average Difference_{probe \, set} = \frac{\sum_{i=1}^{n} (PM_i - MM_i)}{n}
Problems with Mis-Match data

![Graph showing intensity of probe pairs with different categories: PM, MM, and CT. The graph illustrates variations in intensity across different probe pairs.](image-url)
Problems with Mis-Match data

- MM intensity levels are greater than PM intensity levels in ~1/3 of all probes.
- Suggests that MM probes measure actual signal, and not just NSB (non-specific binding).
- Subtracting MM data will result in loss of interesting signal in many probes.
- Several methods have been proposed using only PM data.
Signal Calculation

• Self-reading – Section 20.4.2, 20.4.3
• Ideal Mismatch: Using MM
• Probe Values
• Detection calls

\[ R_i = \frac{PM_i - MM_i}{PM_i + MM_i} \]

Present (P): \( p - \text{value} < \alpha_1 \)
Marginal (M): \( \alpha_1 \leq p - \text{value} < \alpha_2 \)
Absent (A): \( p - \text{value} \geq \alpha_2 \)
The Meaning of A/M/P Calls

<table>
<thead>
<tr>
<th>Detection calls</th>
<th>Magnitude of exp/base ratio</th>
<th>Biological meaning</th>
<th>Interesting gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>M/A M/A</td>
<td>meaningless; can be anything</td>
<td>the gene is probably not expressed in either condition</td>
<td>no</td>
</tr>
<tr>
<td>M/A P</td>
<td>meaningless; close to zero</td>
<td>gene expressed on the baseline array, not expressed in the experiment</td>
<td>yes</td>
</tr>
<tr>
<td>P M/A</td>
<td>meaningless; very large</td>
<td>gene expressed in the experiment, not expressed in the baseline array</td>
<td>yes</td>
</tr>
<tr>
<td>P P</td>
<td>meaningful; can be anything</td>
<td>gene expressed on both channels, useful ratio</td>
<td>depends on the ratio</td>
</tr>
</tbody>
</table>

**TABLE 20.1:** The interplay between the A/M/P calls and the expression values when an experiment array is compared to a baseline array. The only genes that can be safely ignored are those that are absent (A) or only marginally detected (M) in both arrays. All other combinations are or may be interesting from a biological point of view.
Tiling arrays cover the entire length of the genome after the repetitive elements have been removed. Probes of 25 oligonucleotides are tiled at an average resolution of 35 bps, with an average gap of 10 bps. Some tiling array sets contain both perfect match and mismatch sequences. Others contain only perfect matches.
Tiling

Exon tiling arrays

Whole genome tiling arrays

Identify novel transcribed regions in a single tissue sample
Recap - Image
Pipeline
Genechip Scanning

- RNA sample prepared, labelled and hybridised to chip.
- Chip fluorescently scanned.
- Gives a raw pixelated image - .DAT file.
- Grid used to separate pixels related to individual probes.
- Pixel intensities averaged to give single intensity for each probe - .CEL file.
- Probe level intensities combined for each probe set to give single intensity value for each gene.
Affymetrix Data Files

- Image file (.dat file)
- **Probe results file (.cel file)**
- Library file (.cdf, .gin files)
- Results file (.chp file)
Microarray Data Analysis
Problems?

Poorly defined borders
Large holes
Saturated spot
Dust specs
Fiber or scratch?
Bubble
Edge effect
Background haze

(McShane, NCI)
### Spatial Images of Microarrays

- Data for the same brain voxel but for the untreated control mouse.
- Background levels are much higher than those for the Parkinson’s disease model mouse.
- There appears to be something non-random affecting the background of the green channel of this slide.

<table>
<thead>
<tr>
<th>Channel</th>
<th>Median Value</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>F635</td>
<td>10,000</td>
<td>x 10^4</td>
</tr>
<tr>
<td>F532</td>
<td>2,500</td>
<td>x 10^4</td>
</tr>
<tr>
<td>B635</td>
<td>5,000</td>
<td></td>
</tr>
<tr>
<td>B532</td>
<td>1,000</td>
<td></td>
</tr>
</tbody>
</table>
Microarray Data Analysis

- Raw data (image)
- QC (artifact)
- Probe level data
- PM vs MM probes
- Reading for probes

Normalization
- QC (MA plot, box plot, batch effect, PCA)

Comparative analysis
- Clustering analysis
- Classification (feature selection)

Enrichment analysis
- Pathway and network analysis

QC (artifact)
- Artifact

Feature selection
Preprocessing/Data Analysis

Chapter 8, 9, 10 – Background in Statistics
Chapter 20 – Data Preprocessing and Normalization
Chapter 11, 12, 13, 14, 16–Classical/Linear Approaches
Visualization

• Chapter 17
Why Normalize?
Array Normalization
Goal of Normalization

- Arbitrary comparison of gene expressions in various condition and/or tissues
- Goal of normalization – data is independent of experiment and technology
- Ideal measurement – number of mRNA copies per cell
Confounding Reasons

+ Intensity imbalance between RNA samples
+ Affects all genes
+ Not due to biology of samples
+ Reasons include
  - difference in the settings of the photodetector voltage
  - imbalance in total amount of RNA in each sample
  - difference in uptaking of the dyes, etc.
  - difference in sites!
Array Normalization
Consider This

+ A measure of the actual expression levels

+ Differences between median foreground and median background for the red and green channels:

"F635 Median - B635"
"F532 Median - B532"

Slope = 1
The Task

Adjust expression values of all genes so that the ones that are not really differentially expressed have similar values across the array(s)
The MV Box Plot

Median

Upper quartile

Median

Low quartile
General Preprocessing Techniques
Log Transform – Fold Changes

- $\log_{10}(100) = 2$
- $\log_{10}(10,000) = 4$
- $\log_{10}(1,000) = 3$

Scale:
- 0
- 100
- 1,000
- 900
- 9,000
- 10,000
- -1
- 1
Log Transform - Distribution
Log Transform - Distribution
Platforms and Slides

+ Do not mix platforms

+ Affymetrix - Intra-slide normalization
Normalization in R
Use

- library(affy)
- CELfiles <- list.celfiles()

- Rawdata <- ReadAffy(filenames=CELfiles)
- Intensity(Rawdata) # extract intensity

#Visualize
M-A Plot

- `par(mfrow=c(2,3))`
- `Image(rawdata,transfo=log)`
- `Maplot(rawdata, pairs=T, plot.method="smoothScatter", cex = 0.9,...)`
**FIGURE 20.14:** MA-plots of the Affymetrix arrays in this sample data set. The smoothed blue color density representations are derived from kernel density estimations for each array. Data points located in low-density area are depicted as black dots. The red line is the fitted Loess curve. The two summary statistics, median and IQR (interquantile range), for M-values are also shown.
Background Adjustment

```r
> data.bg <- bg.correct(rawdata, method="rma")
> data.qtl <- normalize(data.bg, method="quantiles")
> data.scl <- normalize(data.bg, method="constant")
> data.dChip <- normalize(data.bg, method="invariantset")
> usr.col <- rep(c("red", "blue"), each=3)
> usr.line <- rep(1:3, 2)

> hist(data.bg, lty=usr.line, col=usr.col, main="background + corrected")
> legend("topright", sampleNames(rawdata), lty=usr.line, + col=usr.col)

> hist(data.scl, lty=usr.line, col=usr.col, main="scaling")

> legend("topright", sampleNames(rawdata), lty=usr.line, + col=usr.col)

> hist(data.qtl, lty=usr.line, col=usr.col, main="quantiles")
> legend("topright", sampleNames(rawdata), lty=usr.line, + col=usr.col)

> hist(data.dChip, lty=usr.line, col=usr.col, main="invariantset")
> legend("topright", sampleNames(rawdata), lty=usr.line, + col=usr.col)
```
FIGURE 20.15: A comparison of three normalization methods. The figure shows the probe intensity distributions of the background corrected data (upper left), and the same data after three different normalization methods: scaling (upper right), quantiles (bottom left), and invariant set (bottom right).
Background Adjustment

```r
> boxplot(data.bg, col=usr.col, las=3, names=
+ sampleNames(rawdata), ylim=c(4,8), cex.axis=0.6, main=
+ "background corrected")

> boxplot(data.scl, col=usr.col, las=3, names=sampleNames
+ (rawdata), ylim=c(4,8), cex.axis=0.6, main="scaling")

> boxplot(data.qml, col=usr.col, las=3, names=sampleNames
+ (rawdata), ylim=c(4,8), cex.axis=0.6, main="quantiles")

> boxplot(data.dChip, col=usr.col, las=3, names=sampleNames
+ (rawdata), ylim=c(4,8), cex.axis=0.6, main="invariantset")
```
FIGURE 20.16: A comparison of three normalization methods. The figure shows the box plot of probe intensities of the background corrected data (upper left) and after three different normalization methods: scaling (upper right), quantiles (bottom left), and invariant set (bottom right). Note the almost perfect alignment of the distributions provided by the quantile normalization, and the very good alignment provided by the invariant set method used by dChip. The red and blue colors show the class membership of that particular sample.
Array Normalization

✓ Which genes to use for normalization

✓ Which methods to use
Which Genes?

+ Housekeeping genes

+ Genes involved in activities of cell maintenance and survival

+ Not in cell function & proliferation

+ These genes will be similarly expressed in all samples
Affymetrix & House Genes

+ Difficult to identify – need to be confirmed

+ Affymetrix GeneChip provides a set of house keeping genes based on a large set of tests on different tissues

+ Found to have low variability in these samples (but still no guarantee).
Spiked Controls

✓ Genes that are not usually found in the samples – both control and test sample

✓ E.g., yeast gene in human tissue samples.
The Simple Route

+ Use all adequately expressed genes for normalization

+ Assumption - majority of genes on array are housekeeping genes and proportion of over expressed genes is similar to under expressed genes

+ If the genes on chip are specially selected, then this method will not work.
Methods
Array Normalization

- Scaling/Non-linear
  - Using baseline array or reference
  - Baseline – Median of median intensities
  - Linear Regression
  - Standard Affymetrix method

- Complete - Combine information from all arrays
  - Cyclic Loess
  - Quantile Normalization
First Examine
Scaling - Dividing by Array Mean

- z-transform

- Substitute mean by median/mode/percentile

- Remove upper/lower 10% and divide by mean

- cDNA arrays – adjusts overall intensity problems
Subtracting Mean

\[ x_{in} = \log \left( \frac{x_i}{\bar{x}_i} \right) \]

\[ x_{in} = \log(x_i) - \frac{\sum \log(x_i)}{n} = \log \left( \frac{x_i}{\bar{x}_{ig}} \right) \]
Iterative linear Regression

Repeat regression by discarding points: error > 2\sigma

\[ X_{i}^{\text{norm}} = k \times X_i \]
\[ c = \log(k) \]
\[ M_{i}^{\text{norm}} = \log(X_{i}^{\text{norm}}) = c + M_i \]
Line (global) normalization

- Simplest but most consistent
- Move the median to zero –
  - slope 1 in scatter plot
  - changes the intersection
Using Control Spots & Genes

[Image of Whac-a-Mole game box]

Ohio State Wexner Medical Center
Affymetrix MicroArray Suite (MAS) v4.0

- Uses MM probes to correct for NSB.
- MAS4.0 used simple Average Difference method:

\[
AvDiff = \frac{1}{|A|} \sum_{j \in A} (PM_j - MM_j)
\]

- A is the subset of probes where \(d_j = (PM_j - MM_j)\) is within 3 SDs of the average of \(d_{(2)}, \ldots, d_{(J-1)}\)
- Excludes outliers, but not a robust averaging method.
Affymetrix MicroArray Suite (MAS) v5.0

- Current method employed by Affymetrix.
- Weighted mean using one-step Tukey Biweight Estimate:

\[ \text{signal} = \log^{-1}(\text{Tukey Biweight}(\log(\text{PM}_j - \text{CT}_j))) \]

- \(\text{CT}_j\) (contrast tau) is a quantity derived from \(\text{MM}_j\) never larger than \(\text{PM}_j\).
- Weights each probe intensity based on it’s distance from the mean.
- Robust average (insensitive to small changes from any assumptions made).
Tukey Biweight

\[ \psi(x) = \begin{cases} 
  x \left(1 - \frac{x^2}{c^2}\right)^2 & \text{for } |x| < c \\
  0 & \text{for } |x| > c 
\end{cases} \]
Problems with MAS5.0

- Loss of probe-level information
- Background estimate may cause noise at low intensity levels due to subtraction of MM data.
LOWESS/LOESS

• Locally weighted Polynomial Regression
• Over-fitting vs. under-fitting
• LOWESS-1 – degree 1, LOWESS-2 - degree 2
• Divide domain into windows of size $w$
• And fit

$$w(x) = \begin{cases} 
(l - |x|^3), & |x| < 1 \\
0, & |x| \geq 1 
\end{cases}$$
Effects of LOWESS
Effects of LOWESS

**FIGURE 20.7**: LOWESS normalization. A ratio-intensity plot before (left panel) and after (right panel) the LOWESS correction. The horizontal axis represents $\log_2(cy3) + \log_2(cy5)$ which is a quantity proportional to the overall signal intensity. The vertical axis represents the log-ratio of the two samples $\log_2(cy3/cy5)$. The left panel also shows the nonlinear regression curve.
FIGURE 20.8: LOWESS normalization. A scatter plot before (left panel) and after (right panel) the LOWESS correction. The horizontal axis represents $\log_2(cy3)$ and the vertical axis represents $\log_2(cy5)$. 

Effects of LOWESS
Operation

• For n points: \( \frac{d+1}{n} < q < 1 \)

• In each window, \( n \times q \) points will be chosen
• Small q – less smooth but close to function
• Large q – more smooth
Use for LOWESS

Overall magnitude of spot intensity has impact on relative intensity between channels.

(McShane, NCI)
Use for LOWESS

- Nonlinear
- Gene-by-gene, could introduce bias
- Use only when there is a compelling reason

(McShane, NCI)
Advantages

• No need to specify type of function
• Specify degree of polynomial and smoothing factor q
• Least square is well studied
Disadvantages

• No model or function is generated
• LOWESS will need to be applied for every data; cannot transfer models
• Very expensive
• Suspetible to noise and outliers
Cyclic LOESS – M&A

- Arrays $i, j$
- Probe $k$
- Apply to pairs of arrays
- M – fold change, A log average
- Do LOESS on M-A plot
- Repeat till threshold adjustment reached

$$M_k = \log_2 \left( \frac{x_{ki}}{x_{kj}} \right)$$

$$A_k = \frac{1}{2} \log_2 \left( x_{ki} \cdot x_{kj} \right)$$

$$M_k' = M_k - \hat{M}_k$$

$$x_{ki} = 2^{A_k + \frac{M_k}{2}}$$

$$x_{ki} = 2^{A_k - \frac{M_k}{2}}$$
FIGURE 20.14: MA-plots of the Affymetrix arrays in this sample data set. The smoothed blue color density representations are derived from kernel density estimations for each array. Data points located in low-density area are depicted as black dots. The red line is the fitted Loess curve. The two summary statistics, median and IQR (interquantile range), for M-values are also shown.
Application

“Straighten” Lowess fit line (MA plot) to horizontal & move to zero

Before

After

\[ X_i^{\text{norm}} = k(A) \times X_i \]

\[ c(A) = \log(k(A)) \]

\[ M_i^{\text{norm}} = \log(X_i^{\text{norm}}) = c(A) + M_i \]
Robust Multiarray Average (RMA)
The Need

- Subtraction of MM data corrects for NSB, but introduces noise.
- Want a method that gives positive intensity values.
- Normalising at probe level avoids the loss of information.
Background Correction

- PM data is combination of background and signal.

\[ B(PM_{ijn}) \equiv E(S_{ijn} \mid PM_{ijn}) \]

- Assume strictly positive distribution for signal. Then background corrected signal is also positively distributed.
- Background correction performed on each array separately.
Normalizing

• ‘Quantile Normalization’ used to correct for array biases.
• Compares expression levels between arrays for various quantiles.
• Can view this on quantile-quantile plot.
• Protects against outliers.
These 2 batches do not appear to have come from populations with a common distribution.

The batch 1 values are significantly higher than the corresponding batch 2 values.

The differences are increasing middle of range. Then the values for the 2 batches get closer again.
Robust Multiarray Average (RMA)

- Linear model.
- Uses background corrected, normalised, log transformed probe intensities: \( Y_{ijn} = T(P_{m_{ij}}) \)

\[
Y_{ijn} = \mu_{in} + \alpha_{jn} + \varepsilon_{ijn}
\]

\( \mu_{in} \) = Log scale expression level (RMA measure).

\( \alpha_{jn} \) = Probe affinity affect.

\( \varepsilon_{ijn} \) = Independent identically distributed error term (with mean 0).

- Not all Probes are used
Comparison

**FIGURE 20.11:** A comparison of the three most popular Affymetrix normalization approaches. Each panel shows a scatter plot of the normalized values obtained after hybridizing two different amounts from the same sample. Since these are different amounts from the same biological sample, there should be no differentially expressed (DE) genes. The results obtained with the Affymetrix MAS 5 package (section 20.4), dChip (section 20.5.2), and RMA (section 20.5.3) are shown in panels A, B, and C, respectively. Genes reported as having two- to three-fold changes are shown in yellow. Genes reported as having fold changes larger than three fold are shown in red. Figure reprinted with permission from [235].
Summarization

- Combine intensity values from the probes in the probe set to get a single intensity value for each gene.
- Uses ‘Median Polishing’.
- Each chip normalised to its median.
- Each gene normalised to its median.
- Repeated until medians converge.
- Maximum of 5 iterations to prevent infinite loops.
Effect

Pre-Normalisation
Effect

Post-Normalisation
GC-RMA

- Corrects for background noise as well as NSB.
- Probe affinity calculated using position dependant base effects:

\[ \alpha = \sum_{k=1}^{25} \sum_{j \in \{A,T,G,C\}} \mu_{j,k} 1_{b_k = j} \quad \text{with} \quad \mu_{j,k} = \sum_{l=0}^{3} \beta_{j,l} k^l, \]

- MM data adjusted based on probe affinity, then subtracted from PM.
- Does not lose MM data.
Advantages of RMA/GC-RMA

- Gives less false positives than MAS5.0.
- See less variance at lower expression levels than MAS5.0.
- Provides more consistent fold change estimates.
- Exclusion of MM data in RMA reduces noise, but loses information.
- Inclusion of adjusted MM data in GC-RMA reduces noise, and retains MM data.
Disadvantages

• May hide real changes, especially at low expression levels (false negatives).
• Makes quality control after normalisation difficult.
• Normalisation assumes equal distribution which may hide biological changes.
Conclusions

• RMA is more precise than MAS5.0, but may result in false negatives at low expression levels.
• Useful for fold change analysis, but not for studying statistical significance. Makes quality control difficult.
• Ideal solution – Use standard MAS5.0 techniques for quality control. Then go back and perform probe level normalisation on quality controlled genes.
The Best for The Last
Quantile Normalization

Diagram showing sample array and reference array with cumulative and density plots.
Steps

1. Given a set of $n$ arrays with $p$ probes, construct a matrix $X$ with $p$ rows and $n$ columns, in which each column is an array of raw intensities.

2. Sort each column to construct the matrix $X_{sort}$.

3. Calculate the mean across each row of $X_{sort}$ and assign this mean to each element in the row to get $X'_{sort}$.

4. Construct $X_{normalized}$ by rearranging each column of $X'_{sort}$ in the ordering of the original $X$.

The idea of the quantile normalization is illustrated in Fig. 20.13. In
Quantile Normalization

After Lowess normalization

After quantile normalization
FIGURE 20.15: A comparison of three normalization methods. The figure shows the probe intensity distributions of the background corrected data (upper left), and the same data after three different normalization methods: scaling (upper right), quantiles (bottom left), and invariant set (bottom right).
FIGURE 20.16: A comparison of three normalization methods. The figure shows the box plot of probe intensities of the background corrected data (upper left) and after three different normalization methods: scaling (upper right), quantiles (bottom left), and invariant set (bottom right). Note the almost perfect alignment of the distributions provided by the quantile normalization, and the very good alignment provided by the invariant set method used by dChip. The red and blue colors show the class membership of that particular sample.
In Closing
Overall Comparison

- Linear (global) – the chips have equal median (or mean) intensity
- Intensity-based (Lowess) – the chips have equal medians (means) at all intensity values
- Quantile – the chips have identical intensity distribution
- Quantile is the “best” in term of normalizing the data to desired distribution, however it also changes the gene expression level individually
The End
Two-color CDNA
Background Correction

- Local correction
- Sub-grid correction
- Using blank spots (no DNA deposited)
- Using control spots (exogenous DNA)
Color Normalization
Color Normalization

- Cy3 dye is more efficient
- mRNA for Cy3 dye is abundant
The Curse of the Bannana
Curve Fitting and Correction
## cDNA Variation (Table 3.3)

<table>
<thead>
<tr>
<th>Factor</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>mRNA preparation transcription</td>
<td>Tissues, kits, and procedures vary</td>
</tr>
<tr>
<td>Labeling</td>
<td>Inherent variation in the reactions, type of enzymes used</td>
</tr>
<tr>
<td>Amplification (PCR protocol)</td>
<td>Depends on the type of labeling and procedures as well as age of labels</td>
</tr>
<tr>
<td>Pin geometry variations</td>
<td>PCR is difficult to quantify</td>
</tr>
<tr>
<td>Target volume</td>
<td>Different surfaces and properties due to production random errors</td>
</tr>
<tr>
<td>Target fixation</td>
<td>Fluxuates stochastically even for the same pin</td>
</tr>
<tr>
<td>Hybridization parameters</td>
<td>The fraction of target cDNA that is chemically linked to the slide surface from the droplet is unknown</td>
</tr>
<tr>
<td>Slide inhomogeneities</td>
<td>Influenced by many factors, such as temperature of the laboratory, time, buffering conditions, and others</td>
</tr>
<tr>
<td>Nonspecific hybridization</td>
<td>Slide production parameters, batch-to-batch variations</td>
</tr>
<tr>
<td>Gain setting (PMT)</td>
<td>cDNA hybridizes to background or to sequences that are not their exact complement</td>
</tr>
<tr>
<td>Dynamic range limitations</td>
<td>Shifts the distribution of the pixel intensities</td>
</tr>
<tr>
<td>Image alignment</td>
<td>Variability at low end or saturation at the high end</td>
</tr>
<tr>
<td>Grid placement</td>
<td>Images of the same array at various wavelengths corresponding to different channels are not aligned; different pixels are considered for the same spot</td>
</tr>
<tr>
<td>Nonspecific background</td>
<td>Center of the spot is not located properly</td>
</tr>
<tr>
<td>Spot shape</td>
<td>Erroneous elevation of the average intensity of the background</td>
</tr>
<tr>
<td>Segmentation</td>
<td>Irregular spots are hard to segment from background</td>
</tr>
<tr>
<td>Spot quantification</td>
<td>Bright contaminants can seem like signal (e.g., dust)</td>
</tr>
<tr>
<td></td>
<td>Pixel mean, median, area, etc.</td>
</tr>
</tbody>
</table>

**TABLE 3.3:** Sources of fluctuations in a typical cDNA microarray experiment.
LOWESS/LOESS

- Locally weighted Polynomial Regression
- Over-fitting vs. under-fitting
- LOWESS-1 – degree 1, LOWESS-2 - degree 2
- Divide domain into windows of size $w$
- And fit

$$w(x) = \begin{cases} 
(l - |x|^3), & x < 1 \\
0, & x \geq 1 
\end{cases}$$
Effects of LOWESS
FIGURE 20.7: LOWESS normalization. A ratio-intensity plot before (left panel) and after (right panel) the LOWESS correction. The horizontal axis represents $\log_2(cy3) + \log_2(cy5)$ which is a quantity proportional to the overall signal intensity. The vertical axis represents the log-ratio of the two samples $\log_2(cy3/cy5)$. The left panel also shows the nonlinear regression curve.
Effects of LOWESS

**FIGURE 20.8**: LOWESS normalization. A scatter plot before (left panel) and after (right panel) the LOWESS correction. The horizontal axis represents $\log_2(cy3)$ and the vertical axis represents $\log_2(cy5)$. 
Operation

- For n points \( \frac{d+1}{n} < q < 1 \)

- In each window, \( n \times q \) points will be chosen
- Small q – less smooth but close to function
- Large q – more smooth
Overall magnitude of spot intensity has impact on relative intensity between channels.

(McShane, NCI)

Use for LOWESS

(McShane, NCI)
Use for LOWESS

- Nonlinear
- Gene-by-gene, could introduce bias
- Use only when there is a compelling reason

(McShane, NCI)
Advantages

• No need to specify type of function
• Specify degree of polynomial and smoothing factor $q$
• Least square is well studied
Disadvantages

• No model or function is generated
• LOWESS will need to be applied for every data; cannot transfer models
• Very expensive
• Suspetible to noise and outliers