From CEL files to lists of interesting genes

Rafael A. Irizarry
Department of Biostatistics
Johns Hopkins University
<table>
<thead>
<tr>
<th>e-mail</th>
<th><a href="mailto:rafa@jhu.edu">rafa@jhu.edu</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal webpage</td>
<td><a href="http://www.biostat.jhsph.edu/~ririzarr">http://www.biostat.jhsph.edu/~ririzarr</a></td>
</tr>
<tr>
<td>Department webpage</td>
<td><a href="http://www.biostat.jhsph.edu/">http://www.biostat.jhsph.edu/</a></td>
</tr>
<tr>
<td>Bioinformatics Program</td>
<td><a href="http://www.biostat.jhsph.edu/bioinfo">http://www.biostat.jhsph.edu/bioinfo</a></td>
</tr>
</tbody>
</table>

http://www.bioconductor.org
Acknowledgements

- Ben Bolstad
- Leslie Cope
- Laurent Gautier
- Robert Gentleman
- Wolfgang Huber
- Christina Kendziorski
- James MacDonald
- Francisco Martínez-Murillo
- Felix Naef
- Forrest Spencer
- Terry Speed
- Zhijin Wu
Outline

• Introduction to Technology
• Probe Level Data Analysis
• Differential Expression
• More Probe Level Data Analysis  (Time permitting)
Affymetrix GeneChip Design

Reference sequence

...TGTGATGATGGGGAATGGGTCAGAAGGCCTCCGATGCGCCGATTGAGAAT...

Perfect match

Mismatch

NSB & SB

NSB
Work flow

Raw data (.DAT files)

Probe intensities (.CEL files)

Expression measures (tables)

Rank (list)

Candidate genes (short list)

Image analysis

Pre-processing normalization

Statistical test

Choose filter

Significance level
• 10-20K genes represented by 11-20 pairs of probe intensities (PM & MM)
• Obtain expression measure for each gene on each array by summarizing these pairs
• Background adjustment and normalization are important issues
• There are many methods
Previous Work

• MAS 4.0
  – Negative expression
  – Very noisy for low expressed genes
  – Averages without taking log
• dChip (Li and Wong PNAS 2001)
  – Account for the strong probe effect
  – Need for non-linear normalization
  – Multi-chip analysis reveals outliers
• MAS 5.0 (Current default)
  – No more negatives
  – Average taking log
Throughout the rest of this presentation we will be using Data from Affymetrix’s spike-in experiment. Replicate RNA was hybridized to various arrays. Some probe-sets were spiked in at different concentrations across the different arrays. This gives us a way to assess precision and accuracy.
### Spikein Experiment

<table>
<thead>
<tr>
<th>A</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>0.25</td>
<td>1024</td>
<td>0</td>
<td>512</td>
<td>64</td>
</tr>
<tr>
<td>C</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>0.5</td>
<td>0</td>
<td>0.25</td>
<td>1024</td>
<td>128</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>1</td>
<td>0.25</td>
<td>0.5</td>
<td>0</td>
<td>256</td>
</tr>
<tr>
<td>E</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>2</td>
<td>0.5</td>
<td>1</td>
<td>0.25</td>
<td>512</td>
</tr>
<tr>
<td>F</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>0.5</td>
<td>1024</td>
</tr>
<tr>
<td>G</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>8</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>16</td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>0.25</td>
</tr>
<tr>
<td>I</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>32</td>
<td>8</td>
<td>16</td>
<td>4</td>
<td>0.5</td>
</tr>
<tr>
<td>J</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>64</td>
<td>16</td>
<td>32</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>K</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>128</td>
<td>32</td>
<td>64</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>L</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>256</td>
<td>64</td>
<td>128</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>M</td>
<td>512</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>0</td>
<td>512</td>
</tr>
<tr>
<td>N</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>128</td>
<td>512</td>
<td>0</td>
</tr>
<tr>
<td>O</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>128</td>
<td>512</td>
<td>512</td>
</tr>
<tr>
<td>P</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>128</td>
<td>512</td>
<td>512</td>
</tr>
<tr>
<td>Q</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>128</td>
<td>512</td>
<td>512</td>
</tr>
<tr>
<td>R</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>128</td>
<td>512</td>
<td>512</td>
</tr>
<tr>
<td>S</td>
<td>1024</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>128</td>
<td>512</td>
<td>512</td>
</tr>
</tbody>
</table>
MvA Plot

\[ \log_2(\text{expression 1}) \]

\[ \log_2(\text{expression 2}) \]
MvA Plot

\[ M = \log_2 \left( \frac{\text{expression 2}}{\text{expression 1}} \right) \]

\[ A = \frac{\log_2(\text{expression 2}) + \log_2(\text{expression 1})}{2} \]
Can MAS 5.0 be improved?
MAS 5.0

• Current version, MAS 5.0, uses **Signal**

\[ \text{signal} = \text{Tukey Biweight}\{\log(PM_j - MM_j^*)}\]  

• MAS 4.0 did not take log, now MAS 5.0 does  
• But what about negative PM-MM ?  

• MM* is a new version of MM that is never larger than PM.  
• If MM < PM, MM* = MM.  
• If MM >= PM,  
  – SB = Tukey Biweight (log(PM)-log(MM))  
    (log-ratio).  
  – log(MM*) = log(PM)-log(max(SB, +ve)).  
• Tukey Biweight: \( B(x) = (1 - (x/c)^2)^2 \) if \( |x| < c \), 0 ow.
Before we continue

• We will demonstrate:
  – There is a large multiplicative probe effect
  – The need for non-linear normalization
  – The need for background adjustment
  – Subtracting MM introduces variance

• Then we will show how these facts are used to improve MAS 5.0
Probe effect

This strong probe-effect will result in very high correlation between replicates. Do not get too exited. Look at correlation or variance of relative expression (log FC) instead.
Probe effect makes correlation deceiving

Correlation for absolute expression of replicates looks great! But…
Probe effect makes correlation deceiving

- It is better to look at relative expression because probe effect is somewhat cancelled out.
- Later we will see that we can take advantage of probe effect to find outlier probes.
Why non-linear normalization?

Shifting medians is not enough
Why so much noise?
Why so much noise?

Default algorithm seems to be inspired by the following deterministic model for background:

\[ PM = O + N + S \]
\[ MM = O + N \]

And a multiplicative error model for signal (they take the log before averaging):

\[ PM - MM = S \]
Deterministic model is wrong

- Do MM measure non-specific binding?
- Look at Yeast DNA hybridized to Human Chip
- Look at PM, MM log-scale scatter-plot
- $R^2$ is only 0.5
**Stochastic Model**

*(Additive background/multiplicative error)*

\[ PM = O_{PM} + N_{PM} + S, \]
\[ MM = O_{MM} + N_{MM} \]

\[ \log (N_{PM}), \log (N_{MM}) \sim \text{Bivariate Normal} \ (\rho \approx 0.7) \]

\[ S = \exp (s + a + \epsilon) \]

*s is the quantity of interest* *(log scale expression)*

\[ E[PM - MM] = S, \ 	ext{but} \]
\[ \text{Var}[\log( PM - MM )] \sim 1/S \ (\text{can be very large}) \]
Why we can not ignore BG?

- The data shown here comes from spike-in experiments
- NSB causes fold-change attenuation at low expression level
  - $(E_1+K)/(E_2+K) \approx \frac{E_1}{E_2}$ if $E_1$, $E_2$ are large
  - $(E_1+K)/(E_2+K) \approx 1$ if $E_1$, $E_2$ are small
Alternative background adjustment

• Model observed PM as the sum of a signal intensity $S$ and a background intensity $B$
  \[ PM = B + S, \]
  where it is assumed that $S$ is \textit{Exponential} ($\alpha$), $B$ is \textit{Normal} ($\mu$, $\sigma^2$), and $S$ and $B$ are independent

• Background adjusted PM are then $E[S \mid PM]$

• These parametric distributions were chosen to provide a \textit{closed form} solution
Can MAS 5.0 be improved?

Rank of Spikeins (out of 12626)

141 250 364 368 480 586 686 838 945 1153 1567 NA NA NA NA
Rank of Spikeins (out of 12626)

1
2
3
4
7
11
15
21
35
122
1182
230
450
1380
11700

RMA

Irizarry et al. (2003) NAR 31:e15
More about RMA

- Background adjust as mentioned
- Use quantile normalization
- Assume log-scale probe level additive model

$$\log_2(\text{PM}_{ij}^*) = a_i + b_j + \varepsilon_{ij}$$

- Estimate RMA = $a_i$ for chip $i$ using robust method, such as median polish
Quality Control
QC from probe level models

- RMA fits a probe level model
- From these fits we can obtain residuals
- We can also get weights if we use formal robust regression procedures instead of median polish
- These probe-level residuals and summaries of their size can be used for quality control
- Software available: affyPLM Bioconductor package (Ben Bolstad)
Images of probe level data

This is the raw data
Images of probe level data

Log scale version much more informative
Images of probe level data

Residuals (or weights) from probe level model fits show problem clearly
Images of probe level data

Here is a more subtle artifact. Can you see it? The strong probe effect does not let you.
Images of probe level data

Probe level fit residuals really show it
Other pseudo-chip images

Weights

Positive Residuals

Negative Residuals

Residuals
Differential Expression in Two Populations
Work flow

- Raw data (.DAT files)
- Probe intensities (.CEL files)
- Expression measures (tables)
- Rank (list)
- Candidate genes (short list)

Image analysis

Pre-processing normalization

Statistical test

Choose filter
  Significance level
Introduction

• Many microarray experiments are carried out to find genes which are differentially expressed between two (or more) samples of cells. Examples abound!
• Initially, comparative microarray experiments were done with few, if any replicates, and statistical criteria were not used for identifying differentially expressed genes. Instead, simple criteria were used such as fold-change, with 2-fold being a popular cut-off.
• It did not take long for people to want to assign statistical significance to their findings concerning differentially expressed genes. Could $p$-values be attached, confidence statements be made, and so on?
• All my examples are for 2 population

Compliments of Terry Speed
Different Experiments

• The simplest Affymetrix experiment involves two arrays. Efforts are many times successful but the problem is probably beyond formal statistical inference (valid p-values, etc)

• The second simplest experiment uses various replicate hybridizations. In this case there are some relatively simple procedures for statistical inference
  – Technical replicates will permit inference about the sample at hand
  – Biological replicates will permit inference about a population
  – What should we look at? Average fold-change, t-tests, Wilcoxon tests?
Let us start with EDA

• If we are interested in finding a few genes that are differentially expressed it seems obvious we want to plot differences of average (usually of log intensities)

• However, some use heat maps as default
Clustering is not a good tool for this
• If we are interested in genes with over-all large fold changes why not look at average (log) fold changes?

• Experience has shown that one usually wants to stratify by over-all expression

• We can make MA plots:
  – \( M = \) difference in *average* log intensities and
  – \( A = \) average of log intensities
MA plot of average log ratios
Should we consider gene-specific variance?

Left panels are 3 replicate measurements from pooled RNA (12 mice). These are 3 gene. Right panels have measurements for each of the 12. The colors represent two populations.
• If the number of replicates is very large, the t-statistic is normally distributed with mean 0 and SD of 1.
• If the observed data is normally distributed, then the t-statistic follows a t distribution regardless of sample size.
• Regardless, the square of the t-test is proportional to the ratio of across-group variance to within-group variance.
$t$ - statistic squared if sample sizes the same:

$$N \times \frac{(\bar{Y} - \bar{X})^2}{S_Y^2 + S_X^2}$$
Two Useful Plots

• The MA plot shows
• The volcano plot shows, for a particular test, negative log p-value against the effect size (M)
With $N=3$ t-test is not powerful
If you insist on using MAS 5.0, t-test really helps.
Estimating the variance

- If different genes (or probes) have different variation then it is not a good idea to use average log ratios even if we do not care about significance
- Under a random model we need to estimate the SE
- The t-test divides by SE
- But with few replicates, estimates of SE are not stable
- This explains why t-test is not powerful
- There are many proposals for estimating variation
- Many borrow strength across genes
- Empirical Bayesian Approaches are popular
- SAM, an ad-hoc procedure, is even more popular
- Many are what some call “moderated” t-tests
Some Examples of Tests

Notation:

- $T$ is average log expression in Tx
- $C$ is average log expression in Control
- $S$ is SE

- Note taking log before average is important
- Tests:
  - Average log fold-change: $(T-C)$
  - t-statistic: $(T-C) / S$
  - SAM shrunken t-statistic: $(T-C) / (S + S_0)$
  - Bayesian posteriors: $(T-C) / \sqrt{(S^2 + K^2)}$
  - Wilcoxon Rank test
  - Ad-hoc pairwise comparison No formula

Many of these are in the `limma` package
SAM is in the `siggenes` package
Also `EBarrays` package has some tests
ROC curves

• ROC = Receiver Operator Characteristic
• To compare tests it is important to look at both specificity and sensitivity
• For every cut-off value there will be some true positives and some false positives
• We can make a curve that plots true positives versus false positives as we move the cut-off
Does it make a difference (N=3)?
Does it make a difference (N=12)?
Work flow

- Raw data (.DAT files)
  - Probe intensities (.CEL files)
    - Expression measures (tables)
      - Rank (list)
        - Candidate genes (short list)
  - Image analysis
    - Pre-processing normalization
      - Statistical test
        - Choose filter
          - Significance level
Hypothesis testing

- Once you have a score for each gene, how do you decide on a cut-off? p-values are popular. Are they appropriate?
- Test for each gene null hypothesis: no differential expression.
- Notice that if you have look at 10,000 genes for which the null is true you expect to see 500 attain p-values of 0.05
- This is called the multiple comparison problem. Statisticians fight about it. But not about the above.
- Main message: p-values can’t be interpreted in the usual way
Conclusions

• All steps are important (including preprocessing)
• EDA is powerful tool
• Bioconductor provides flexible software for preprocessing, analysis, annotation, and report creation
More on Probe Level Analysis
Can RMA be improved?

RMA attenuates signal slightly to achieve gains in precision

<table>
<thead>
<tr>
<th>method</th>
<th>slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAS 5.0</td>
<td>0.69</td>
</tr>
<tr>
<td>RMA</td>
<td>0.61</td>
</tr>
</tbody>
</table>
Current work on preprocessing

- We need better estimates of means and variances of bivariate normal background noise
- Use observed MM intensities along with sequence information
- We also have a solution that does not use the MM
Problems with MM

- MM detect signal
Problems with MM

- MM detect signal
- The estimate of non-specific binding is sequence-dependent

Plot is courtesy of Felix Naef
Problems with MM

- MM detect signal
- The estimate of non-specific binding is sequence-dependent
Problems with MM

- MM detect signal
- The estimate of non-specific binding is sequence-dependent
- Take up half the chip ($250)
Predict NSB with sequence

- Fit simple linear model to yeast on human data to obtain base/position effects
Predict NSB with sequence

- Fit simple linear model to yeast on human data to obtain base/position effects
- Call these *affinities* and use them to obtain parameters for background model
Does it help?

- We can predict empirical results with model
- Accuracy of expression measures improves...
Does it help?

- We can predict empirical results with model
- Accuracy of expression measures improves...
- Without adding too much variance
- A correspondence in Nature Biotech and a paper in JASA (forthcoming) describes the new expression measure: GCRMA
Supplemental Slides
RMA references

- Current work: GCRMA (uses sequence info)
  To appear in JASA 2004
Terminology

• Each gene or portion of a gene is represented by 1q to 20 oligonucleotides of 25 base-pairs.

• **Probe**: an oligonucleotide of 25 base-pairs, i.e., a 25-mer.
• **Perfect match (PM)**: A 25-mer complementary to a reference sequence of interest (e.g., part of a gene).
• **Mismatch (MM)**: same as PM but with a single homomeric base change for the middle (13th) base (transversion purine <-> pyrimidine, G <->C, A <->T).
• **Probe-pair**: a (PM,MM) pair.
• **Probeset**: a collection of probe-pairs (1q to 20) related to a common gene or fraction of a gene.
• **Affy ID**: an identifier for a probe-pair set.
• The purpose of the MM probe design is to measure non-specific binding and background noise.
Affymetrix files

• Main software from Affymetrix company *MicroArray Suite - MAS*, now version 5.
• **DAT** file: Image file, \( \sim 10^7 \) pixels, \( \sim 50 \) MB.
• **CEL** file: Cell intensity file, probe level PM and MM values.
• **CDF** file: Chip Description File. Describes which probes go in which probe sets and the location of probe-pair sets (genes, gene fragments, ESTs).
Image analysis

- Raw data, **DAT image files** → **CEL files**
- Each probe cell: 10x10 pixels.
- **Gridding**: estimate location of probe cell centers.
- **Signal**:
  - Remove outer 36 pixels → 8x8 pixels.
  - The probe cell signal, PM or MM, is the 75\textsuperscript{th} percentile of the 8x8 pixel values.
- **Background**: Average of the lowest 2% probe cell values is taken as the background value and subtracted.
- Compute also quality measures.
GeneChip® MAS 4.0 software used AvDiff up until 2001

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

where A is a set of “suitable” pairs, e.g., pairs with \( d_j = PM_j - MM_j \) within 3 SDs of the average of \( d_{(2)}, \ldots, d_{(J-1)} \)

• Obvious problems:
  – Negative values
  – No log scale
Why use log?

Original scale vs. Log scale
Li and Wong’s observations

- There is a large probe effect
- There are outliers that are only noticed when looking across arrays
- Non-linear normalization needed

PNAS vol. 98. no. 1, 31-36
Expression from 2 replicate arrays

Correlation is higher than 0.99
Expression from probesets divided into 2 (at random)

Correlation drops to 0.55
Expression measures
Li & Wong

• Li & Wong (2001) fit a model for each probe set, i.e., gene

\[ \begin{align*}
PM_{ij} - MM_{ij} &= \theta_i \phi_j + \varepsilon_{ij}, \\
\varepsilon_{ij} &\sim N(0, \sigma^2)
\end{align*} \]

where
– \( \theta_i \): model based expression index (MBEI),
– \( \phi_j \): probe sensitivity index.

• Maximum likelihood estimate of MBEI is used as expression measure for the gene in chip \( i \).
• Non-linear normalization used
• Ad-hoc procedure used to remove outliers
• Need at least 10 or 20 chips.
MBEI not much better
Why background correct?
Quantile normalization

- Many non-linear normalization methods perform similarly
- Quantiles is my favorite because it's fast
- Basic idea:
  - order value in each array
  - take average across probes
  - Substitute probe intensity with average
  - Put in original order
<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>4</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>3</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>5</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>5</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>
Example: QC
QC

raw values

log2-transformed values
T-test and fold change tell slightly different stories

In red are the top 100 genes ranked by fold change. Notice some have very small t-tests i.e. large (marginal) p-values
Why Keep Probe Level Data?

• Quality control
  – Spatial Effects
  – RNA degradation
• Detection of defective probes
• Transcript sequence “estimates” change
• Ways to reduce to expression measure keep improving
Affymetrix GeneChip Design

GeneChip Probe Array

Hybridized Probe Cell
- Single stranded, labeled RNA target
- Oligonucleotide probe

Image of Hybridized Probe Array

- 1.28 cm
- 24 μm

Millions of copies of a specific oligonucleotide probe

>200,000 different complementary probes

Compliments of D. Gerhold
Should we use variance?

This has all 15K+ genes. The y-axis has average fold change for the 3 replicates. The x-axis has percentiles of variances obtained from the 12 individuals.