ADVANCED STATISTICAL METHODS FOR GENE EXPRESSION DATA

Veera Baladandayuthapani & Kim-Anh Do

University of Texas M.D. Anderson Cancer Center
Houston, Texas, USA
veera@mdanderson.org

Course Website:
http://odin.mdacc.tmc.edu/~kim/TeachBioinf/AdvStatGE-Prot.htm

BIG PICTURE

- Genomics/Bioinformatics (general)
  - As the generation, organization, and analysis of biological data (initially genomic data)
  - Attracted lot of interest in different fields: Computer Science, Physics, Engineering and of course Statistics
- Statistical Genomics
  - Class of statistical methods for dealing with large biological data sets
  - Goal: statistically identify significant changes in biological processes to answer relevant biological questions.
  - High-throughput studies; get data matrix; mine the matrix for information

EXAMPLES

- Changes in DNA sequence
- Quantitative trait locus identification
- Differential expression of genes (microarrays)
- Changes in protein abundance (proteomics)
- Cell and molecular based studies
- And many many more.....

ROAD MAP FOR TODAY

- How to get data?
- How to clean data?
- In the context of Microarrays
  - Get data: Image Processing
  - Clean data: Pre-processing data

ROAD MAP FOR TODAY

- How to get data?
- How to clean data?
- In the context of Microarrays
  - Get data: Image Processing
  - Clean data: Pre-processing data
- The literature is huge!

Microarray Technology

- High-throughput assays for understanding molecular biology
- Simultaneously measure expression levels for thousands of genes

- By understanding how "gene expression" changes across multiple conditions
  - Researches gain clues about gene functions
  - How genes work together to carry out biological functions
- Many applications in a variety of studies; attracted considerable statistical literature
- Other techniques to measure gene expression
  - Serial analysis of gene expression (SAGE); cDNA library sequencing; differential display; cDNA subtraction; multiplex quantitative RT-PCR

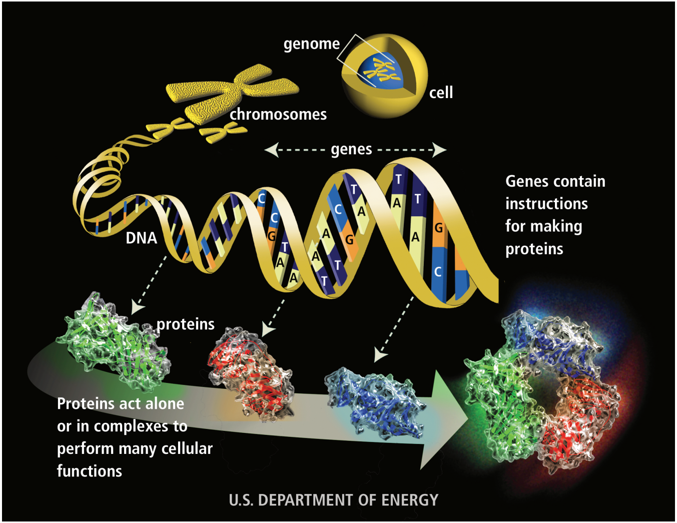
BASIC PRINCIPLES OF MICROARRAYS

- Central dogma of molecular biology: "information transfer" (Nguyen et al., 2003)

$$\text{DNA} \Rightarrow \text{mRNA} \longrightarrow \text{amino acid} \longrightarrow \text{protein} \longrightarrow \text{cell phenotype} \longrightarrow \text{organism phenotype}$$

- Different levels of gene expression
  - Transcription level: DNA → RNA (microarrays)
  - Protein level: mRNA → proteins (protein arrays)
- Three primary information processes in functioning organisms
  - Replication (duplication) - DNA
  - Transcription (copying) - RNA
  - Translation - Protein production

Central Dogma: DNA and Protein


genome
cell
chromosomes
genes
Genes contain instructions for making proteins
DNA
C
C
G
T
A
A
C
A
A
T
T
T
C
A
G
T
A
T
A
G
C
proteins
Proteins act alone or in complexes to perform many cellular functions
U.S. DEPARTMENT OF ENERGY

http://www.ornl.gov/sci/techresources/HumanGenome/graphics/slides/images1.shtml

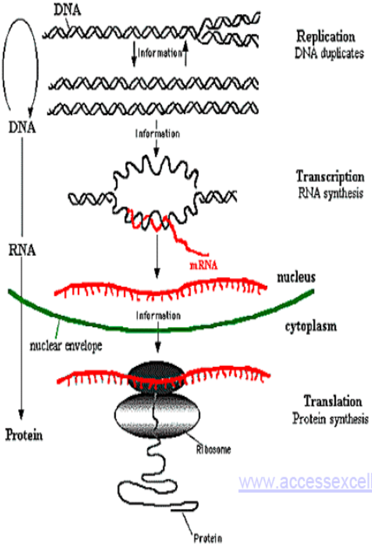
DNA, Genes and DNA Transcription

- DNA
  - DNA in native state is double stranded
  - Complementary base pairing: A-T, G-C

...AAAAGCTAGTCGATGCTAG...
...TTTTCGATCAGCTACGATC...

- RNA
  - Single stranded
  - Base pairing: A-U, G-C (same as DNA with T replaced with U)
- DNA Transcription
  - Inside the nucleus. DNA strand encoding the gene copied (mRNA)
  - Section of one strand of DNA corresponding to the gene is copied using base complementarity.

Central Dogma: RNA

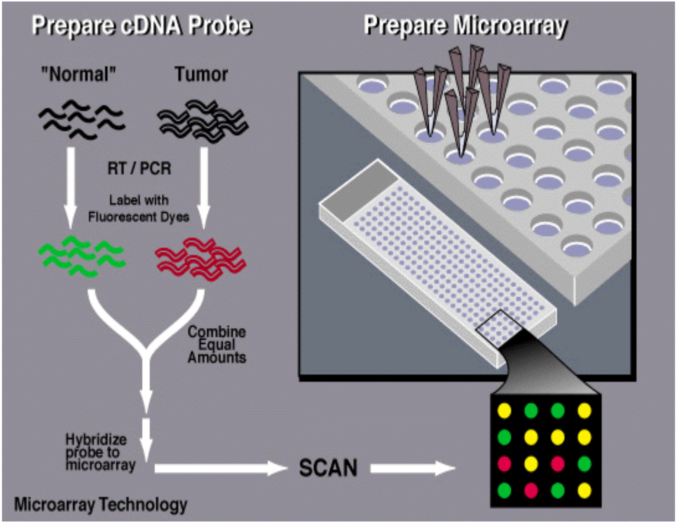


www.accessexcellence.com/AB/GG/

MICROARRAY TECHNOLOGIES

- Basic Premise: if we know target mRNA sequence we can build a probe for it using the complementary sequence. Probe location tells us the identity of the gene.
  Two variants:
- Reverse transcription from mRNA to cDNA
  - cDNA Microarray technology; Duggan et al. (1999)
- Synthesis of short subsequences (oligos)
  - Affymetrix (www.affymetrix.com); Genechips

The cDNA Microarray Technology: Overview



http://www.accessexcellence.org/RC/VL/GG/microArray.html

THE cDNA MICROARRAY TECHNOLOGY

- Array fabrication: preparing the glass slide, obtaining the DNA sequences and depositing ("printing") the cDNA sequences
- Sample preparation: processing and preparing the biological sample of interest
  - Isolating total RNA (mRNA and other RNA's) from tissue samples
  - Much variability comes from this step
- cDNA synthesis and labeling: making and labeling cDNA's from experimental and reference samples.
- Hybridization: applying the experimental and reference cDNA mixture solution to the array.
- Many sources of variation come from these processes.

Data Collection: Microarray Quantification

- Expression levels measured by spot intensities of Cy5 and Cy3 dyes
- Intensities obtained by scanning array with confocal laser microscope
- Array scanned at two wavelengths: Cy5 and Cy3 tagged sample
- Result: Two 16-bit TIFF images containing fluorescence intensities of pixels
- This is the raw data!

Oligonucleotide (Oligo) Arrays

- Affymetrix GeneChip most popular commercially produced high-density arrays; Genechips
- Oligonucleotide: short sequence of nucleotides
- Each gene (more accurately sequence of interest or feature) is represented by multiple short (25-nucleotide) oligo probes.
- Probes sequences are chosen to have good and relatively uniform hybridization characteristics
- A probe is chosen to match a portion of its target mRNA transcript that is unique to that sequence.
- Oligo probes can distinguish among multiple mRNA transcripts with similar sequences

Genechips and Probes

- A probe set is used to measure mRNA levels of a single gene.
- Each probe set consists of multiple probe cells.
- Each probe cell contains millions of copies of one oligo.
- Each oligo is intended to be 25 nucleotides in length.
- Probe cells in a probe set are arranged in probe pairs.
- Each probe pair contains a perfect match (PM) probe cell and a mismatch (MM) probe cell.
- A PM oligo perfectly matches part of a gene sequence.
- A MM oligo is identical to a PM oligo except that the middle nucleotide (13th of 25) is intentionally replaced by its complementary nucleotide.

Array design

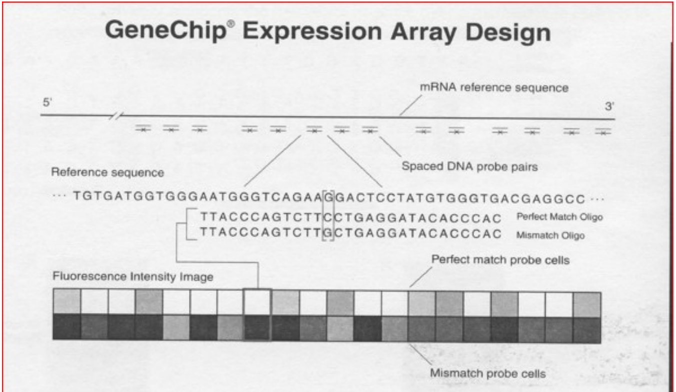
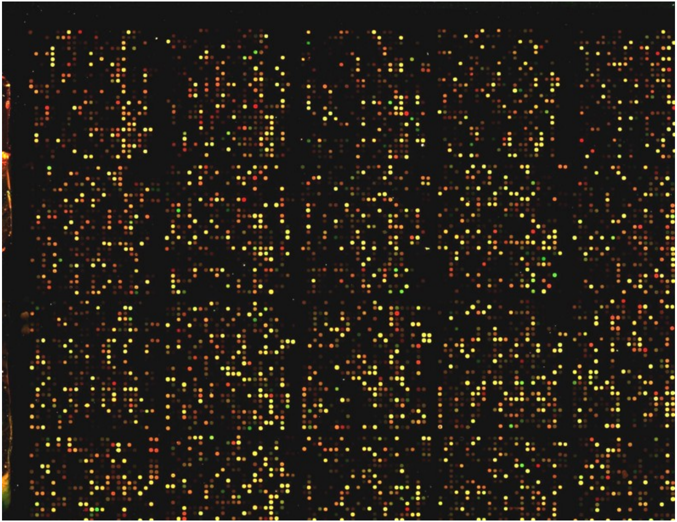

GeneChip® Expression Array Design
mRNA reference sequence
5'
3'
Spaced DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC ···
TTACCCAGTCTTCCTGAGGATACACCCAC
Perfect Match Oligo
TTACCCAGTCTTGCTGAGGATACACCCAC
Mismatch Oligo
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells

Figure 1-3 Expression tiling strategy

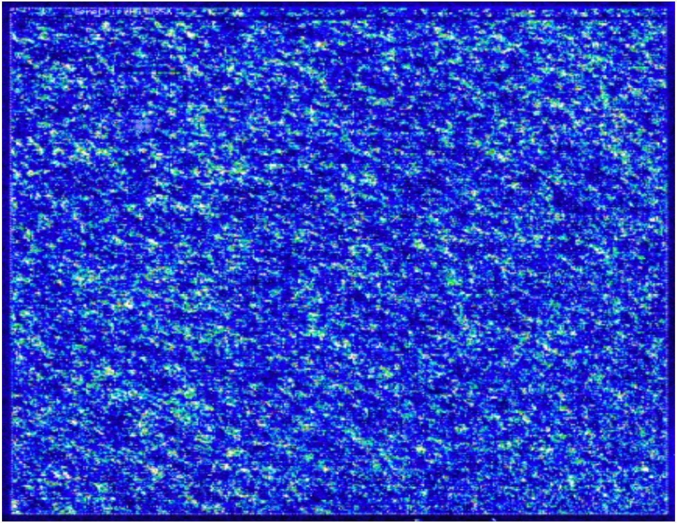
Shown are probe pairs, PM and MM
There are tens of thousands of probe sets per chip

http://www.affymetrix.com

WHAT DO WE FINALLY GET: cDNA ARRAYS



What do we finally get: Affymetrix arrays

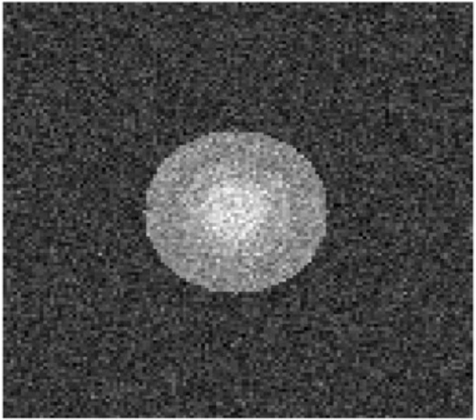


http://www.affymetrix.com

WHAT DO WE FINALLY GET

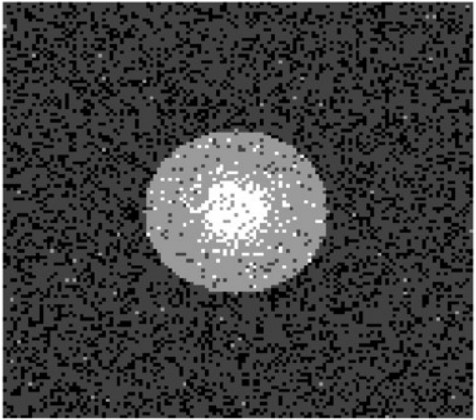
- *Digital image*: rectangular array of intensity values
- Each intensity value corresponds to a *pixel*
- *Color Depth*: is the number of bits used to store the intensity value of one pixel
- Color depth of 16 bits/pixel (common for microarray scanners) means the intensity values of each pixel is an integer between 0 and 65,535 (= $2^{16} - 1$)
- The number of pixels contained in a digital image is called *resolution*

Color Depth = 6; Resolution = 128 x 128



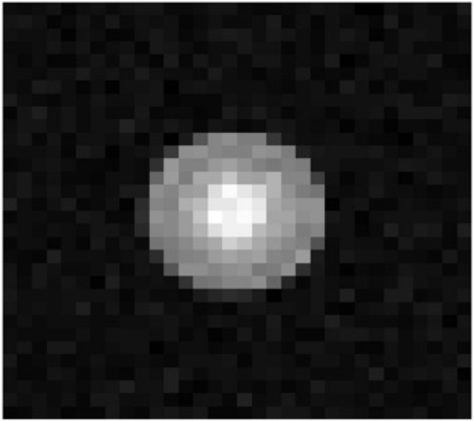
(Adapted from Dan Nettleton's JSM short course slides)

Color Depth = 2; Resolution = 128 x 128



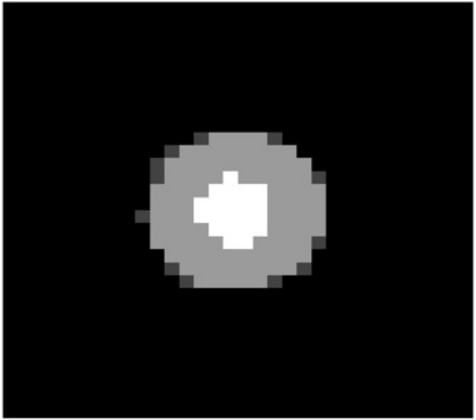
(Adapted from Dan Nettleton's JSM short course slides)

Color Depth = 6; Resolution = 32 x 32



(Adapted from Dan Nettleton's JSM short course slides)

Color Depth = 2; Resolution = 32 x 32



(Adapted from Dan Nettleton's JSM short course slides)

Image Processing for cDNA arrays

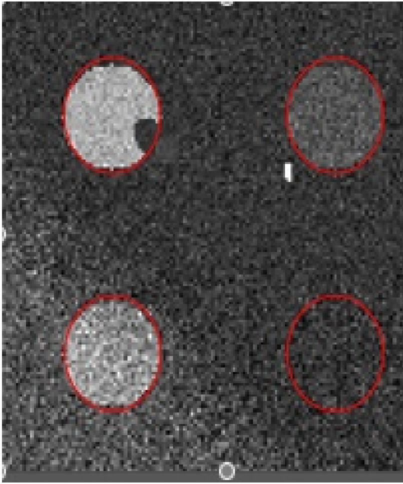
4 basic steps:

- Array localization - locate the spots
- Image segmentation - categorize each spot as foreground (signal), background or other
- Quantification - assign signal and background values to each spot
- Spot quality assessment - compute measures of spot quality for each spot

These steps use specialized software and can involve varying degrees of human intervention.

Array Localization

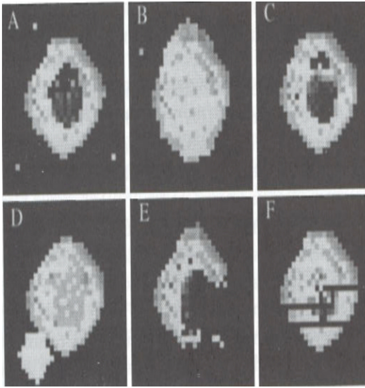
- Mostly software driven
- Ideally every spot should have shape of a circle and all spots should have consistent diameters
- Users may
  - Aid software by outlining grids, providing information about spot size and the number of rows and columns spotted on slide
  - Make manual adjustments to improve upon automated spot adjustments



(Adapted from Dan Nettleton's JSM short course slid

Array Localization

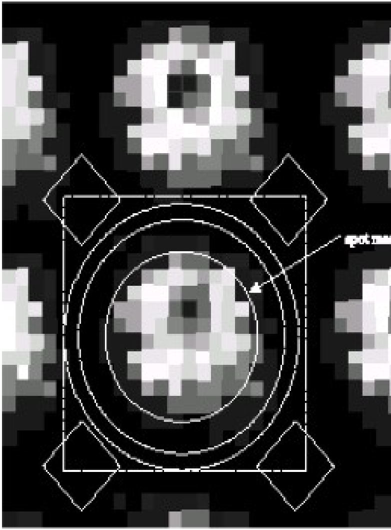
- Observed spots are hardly circular
- Donut shape, sickle shape, oval or pear shape, black holes inside spot
- Image analysis software rectify the spatial problems by capturing true shape of spots
- Other image analysis techniques use distribution of pixels e.g. histogram
- Hybrid approaches: combine both spatial and distributional approaches



(Adapted from McLachlan, Do and Ambroise, 2004

Image Segmentation

- Segmentation technique to classify each pixel in target area as foreground (spot signal)/background
- Spot signal is flourescence intensity due to target molecules hybridized to probe sequences contained in the spot (which is what we want to measure) plus background flourescence (which we would rather not measure)
- Background is fluorescence that may contribute to spot pixel intensities but is not due to target molecules
- Dust particles, stray fluorescent molecules, fluorescence in the slide itself etc.
- Background varies across slide so most softwares attempt to measure local background by quantifying pixel intensities around each spot.



Adapted from Yang et al., 2002

Quantification

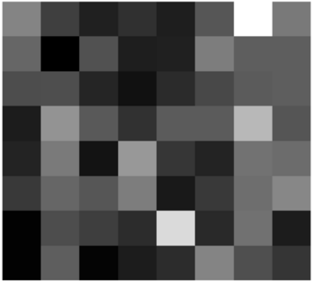
- Pixels in spot and background area used to compute intensities
- Spot intensity: Some statistics representing intensities for all pixels in spot area; similarly for background intensity
  - Mean: mean of pixel intensities
  - Median: median of pixel intensities
  - Mode: location of peak in histrogram of intensities
  - Area: number of pixels
  - Total: sum of pixel intensities
- Many open questions still remain
- Imaging software also output some spot quality statistics.
- Different image analysis programs: GenePix, SPOT, ScanAlyze, UCSF Spot and Imagene

IMAGE PROCESSING FOR OLIGO ARRAYS

- Affymetrix Genechips use propriety Affymetrix software
- Genechip surface covered with square shaped cells containing probes
- Probes are synthesized on the chip in precise locations
- Thus spot finding and image segmentation are not major issues

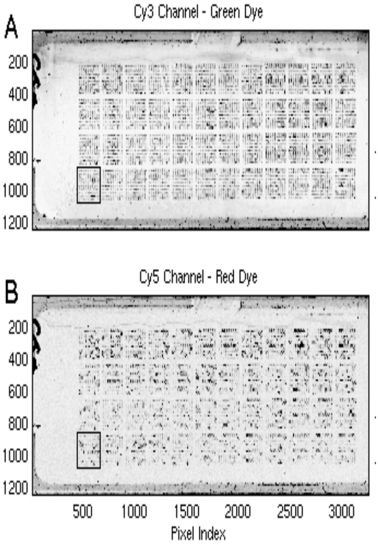
Array Localization

- 8 x 8 = 64 pixels
- Border pixels excluded
- 75th percentile of the 36 pixel intensities corresponding to the center 36 pixels is used to quantify fluorescence intensity for each probe cell
- These values are called PM values for perfect-match probe cells and MM values for mismatch probe cells
- The PM and MM values are used to compute expression measures for each probe set



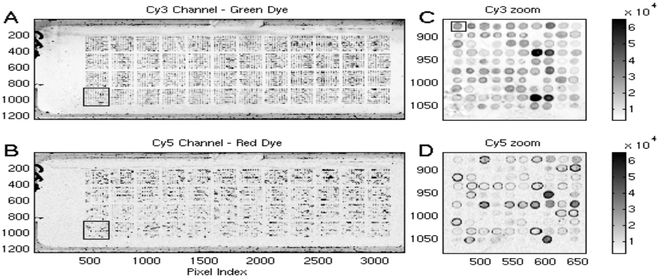
An Example

- Early experiment at MD Anderson; 4800 dots; 4 x 12 grid of 10 x 10 patches
- Each pixel a 16-bit intensity measurement; values between 0 and 65535
- Each image 8 M in size
- Nowadays more genes; more resolution; greater size
- Some image analysis software assume 8-bit (0 to 255); lose some gradation information; damaging for analysis



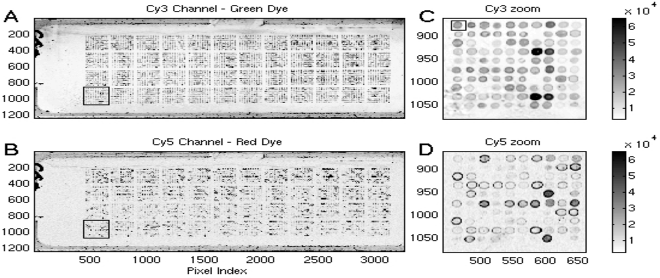
Adapted from Baggerly et al., 2006

An Example



- Patch zoomed out
- Replicate spotting of same genes; top half of patch replicated in the bottom
- Cy3 spots in rows 4 and 9 in column 7 are a replicate pair; some confidence in assay

An Example: Analysis Caveats

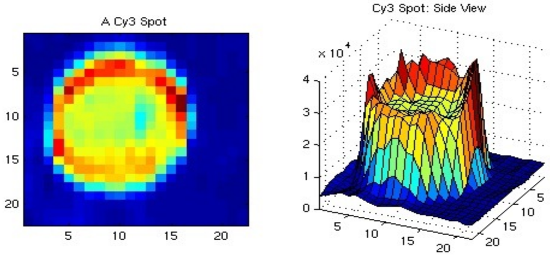


- “Dots” really not “dot-like”; rather rings of high intensity about lower-level centers; true for both channels; surface tension dries cause clumping at edges

- Dots not equal size; automated procedure not possible; human intervention

- Some smearing at the lower left hand corner (green channel)!; affects assessment of spot intensity

An Example: Analysis Caveats



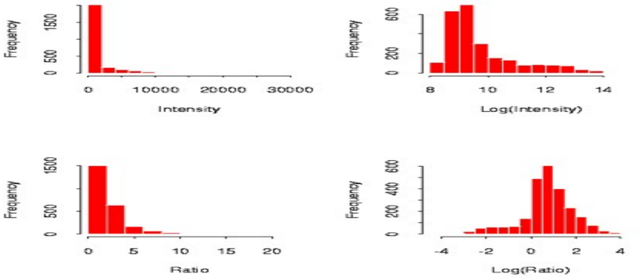
Baggerly et al., 2006

- One single spot zoomed out and side view
- Ring shape visible ⇒ uneven hybridization
- Measurements outside the spot not at 0 intensity ⇒ need some type of background correction
- Conclusion: need good image quantification algorithms

Processed Data

- $\mathbf{I}_R = (m_{ij}^R)$ and $\mathbf{B}_R = (b_{ij}^R)$ be $n \times p$ matrices containing spot and background intensities of genes $j = 1, \ldots, p$ in samples (arrays) $i = 1, \ldots, n$ from Cy5-channel(red) image
- $\mathbf{I}_G = (m_{ij}^G)$ and $\mathbf{B}_G = (b_{ij}^G)$ corresponding matrices from Cy3-channel(green) image
- Many analysis based on:
  - Background corrected intensities:
    $\mathbf{R} = (r_{ij}) = (m_{ij}^R - b_{ij}^R)$ and $\mathbf{G} = (g_{ij}) = (m_{ij}^G - b_{ij}^G)$
  - Intensity ratios: $X = (x_{ij})$ where $x_{ij} = r_{ij}/g_{ij}$
  - Most common: Log(X) log ratio of intensities

Why log?



- Makes variation of intensities and ratios of intensities more independent of absolute magnitude
- Evens out highly skewed distributions; gives more realistic sense of variation
- Approximates normal distribution; treats up- and down- regulated genes symmetrically; helps visualize variation in both directions

Normalization

- Describes the process of removing (or minimizing) non-biological (techincal) variation in the measured expression levels
- Aim: Biological differences can be more easily detected
- Typically, normalization attempts to remove global effects i.e. effects shown by exploratory plots for a slide or multiple slides
- Nothing to do with normal distribution
- Not a panacea for bad data!

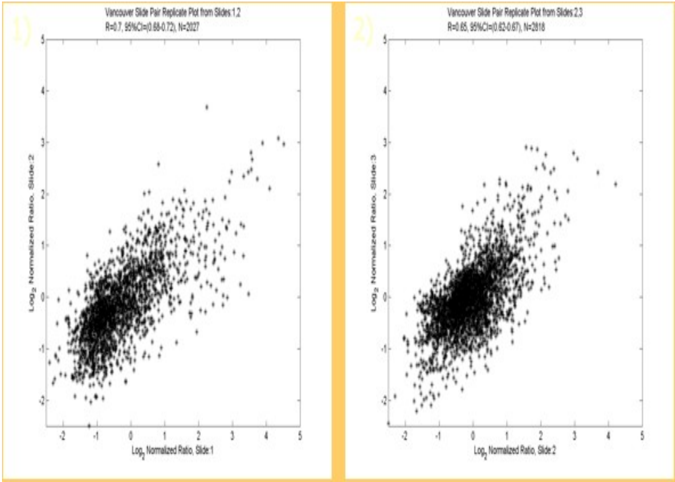
Sources of Non-biological variation

- Dye bias: differences in heat and light sensitivity; efficiency of dye incorporation
- Differences in amount of labeled cDNA hybridized
- Different amounts of mRNA
- Different scanning parameters
- Different technicians producing the arrays
- Any process that induces systematic error

Normalization methods

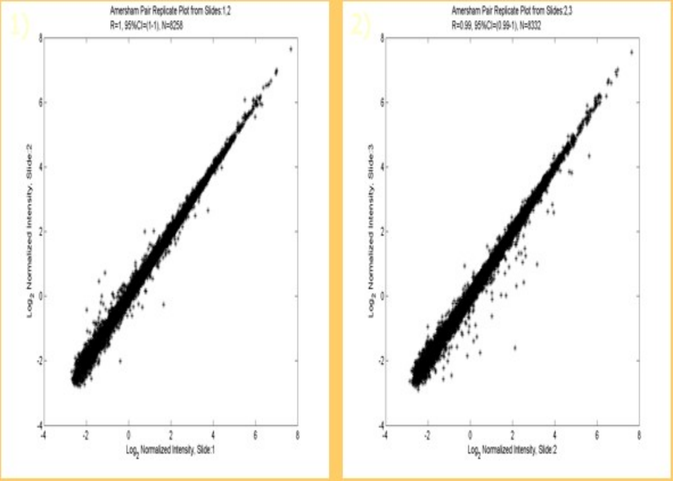
- Several methods
  - Global mean methods
  - (Iterative) linear regression method
  - Curvilinear methods (e.g. Lowess)
  - Variance model methods
- Basic idea: Try to get slope of $\sim$1 and a correlation of $\sim$1

Example: normalization needed



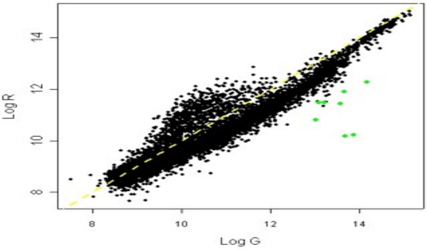
Courtesy Jonathen Wren

Example: normalization not needed

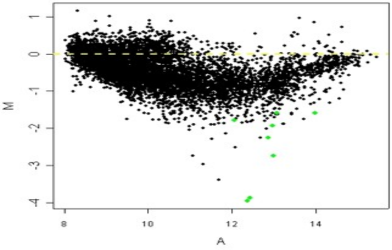


Courtesy Jonathen Wren

First Step: M-A Plots



log_2R vs log_2G



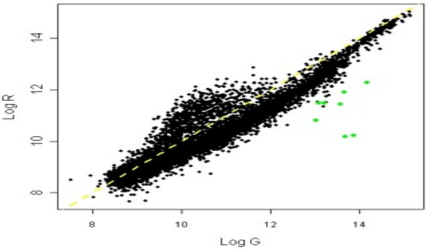
$M = log_2(R/G)$ vs $A = log_2(RG)/2$

- Essentially a rotation of the log plot so that the 45 degres line is now the horizontal axis; Originally proposed for microarrays by Dudoit et. al. (2002, Statistica Sinica); M = *Minus*, A = *Average*

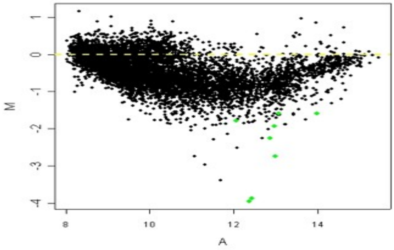
- Shows eventual non-linear and unwanted dependence between ratios and fluorescence intensities; M is units of 2-fold change (if log base is 2) and A is in units of 2-fold increase in brightness

- Shows that using only ratios is a naive way to identify differentially expressed genes.

M-A Plots



log_2R vs log_2G



$M = log_2(R/G)$ vs $A = log_2(RG)/2$ ‘

- Using just the ratios or log ratios to visuzalize the data does not enable us to see the systematic dependence of the ratio on intesity values
- Ideally, if few genes expressed the cloud is centered around 0; if not then lowess
- For two-channel arrays shown above; for one channel usual to plot pair of slides

NORMALIZATION: GLOBAL

- Normalization based on

$$log_2 R/G \longrightarrow log_2 R/G - c = log_2 R/(kG)$$

- Common choices for k or c = $log_2 k$ are
    - $c =$ mean/median of log rations for a particular gene set (e.g. all genes, or control or housekeeping genes)
    - Alternative: $k = \sum R_i / \sum G_i$; total intensity normalization
- Changes roughly symmetric at all intensities
- Not intensity/spatial dependent

NORMALIZATION: INTENSITY-DEPENDENT

- Run a smoother through the MA plot, shifting the M value of the pair (A,M) by c=C(A), i.e.

$$log_2 R/G \longrightarrow log_2 R/G - c(A) = log_2 R/\{k(A)G\}$$

- One estimate of c(A) is made using the LOWESS function of Cleveland (1979)
  - LOcally WEighted Scatterplot Smoothing
  - First proposed for microarrays by Yang et al. (2002)

- Global LOWESS use implicit assumptions that, when stratified by mRNA abundance,
  - Only a minority of genes are expected to be differentially expressed or,
  - any differential expression is as likely to be up-regulation as well as down-regulation

Normalization: print-tip

- Both intensity-dependent variation and spatial bias can be significant sources of systematic error
- Global methods do not correct for spatial effects produced by hybridization artifacts or print-tip or plate effects during microarray construction
- Can correct for both print-tip and intensity dependent bias by performing LOWESS fits to the data within print-tip groups, i.e.

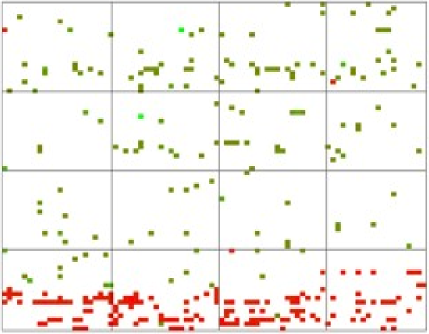
  $$log_2 R/G \longrightarrow log_2 R/G - c_i(A) = log_2 R/\{k_i(A)G\}$$

  where $c_i(A)$ is the lowess fit to the MA plot for the ith grid only (ith print group), $i = 1, \ldots, I$ (= number of print tips)
- Also called sub-array normalization

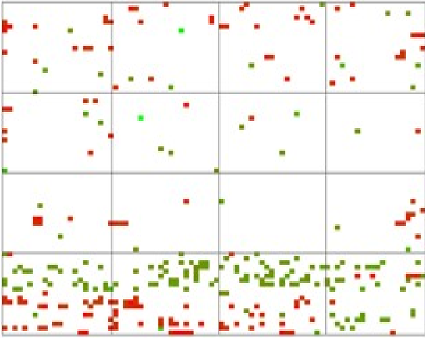
Local smoothing and regression

- LOWESS is a form of a local smoother
- Classical (global) regression: draws a single line to the entire set of points
- Local regression: draws a curve through noisy data by smoothing.
- Linear (or polynomial) function of the predictor(s) is created in a local neighborhood, points are weighted
- As you move through values of the predictor, the neighborhood moves as well
- Lot of active research in the general area of smoothing

Effects of Location normalization



Before normalization



After print-tip group normalization

(Courtesy Yee Hwa Yang)

NORMALIZATION CONTINUED...

- The LOWESS lines can be run through many different sets of points; each strategy has its own implicit set of assumptions, justifying its applicability
- What genes to use
  - All genes on the array
  - Housekeeping genes: genes whose expression does not change over a variety of conditions.
  - Controls: Spiked controls (e.g. plant genes) or genomic DNA titration series; regulate amount of spike-in relative to the amount of control.
- Different arrays often do not show identical signal distribution of M values: various technical reasons (e.g. labeling efficiency, amount of labelled RNA, scanner settings, etc...)
- Need to normalize the signal between chips: multiple possibilities, one often used: ”scale normalization”

Scale normalization

Assume: All slides have the same spread in $\boldsymbol{M}$

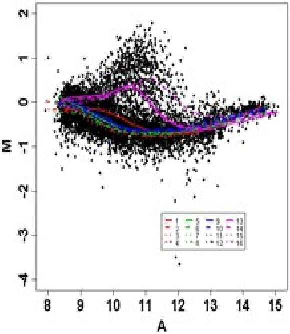
- True log ratio is μ_{ij} where $\boldsymbol{i}$ represents different slides and $\boldsymbol{j}$ represents different spots
- Observed is $\boldsymbol{M_{ij}}$, where $\boldsymbol{M_{ij} = a_i \mu_{ij}}$
- Robust estimate of $\boldsymbol{a_i}$ is

$$\frac{MAD_i}{I\sqrt{\prod_{k=1}^{J} MAD_i}}$$

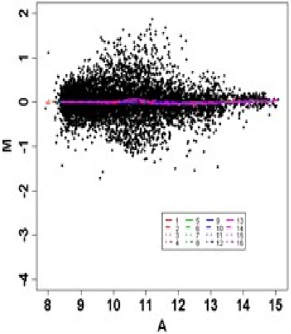
where $\boldsymbol{MAD_i = median_j\{|y_{ij} - median(y_{ij})|\}}$

- Could instead make same assumption for print tip groups (rather than slides)
- Scale normalization changes scale of data; affects fold change calculations

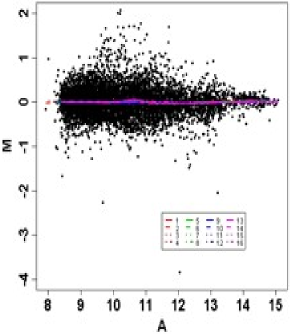
An example



Un-normalized



Print-tip normalization



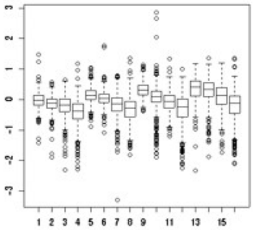
Print tip & scale norm.

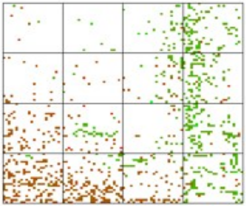
Point: location normalization takes out non-linear effects

(Courtesy: Darlene Goldstein)

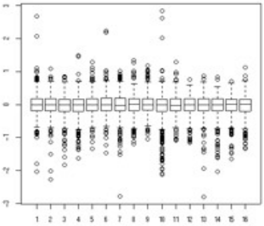
Another example

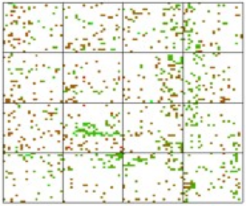
Before





After





(Courtesy: Darlene Goldstein)

Quantile Normalization

- Bolstad et. al. (2003; Bioinformatics) propose quantile normalization for microarray data; most commonly used in normalization of Affy data
- Goal: to give same empirical distribution of intensities to each array i.e. after quantile normalization the histogram of intensities on each array will be identical
- Target distribution is found by averaging the quantiles for each of the arrays in the dataset
- An intensity is transformed in the following manner:

  $$x_{ij}^* = F^{-1}\{G_j(x_{ij})\}$$

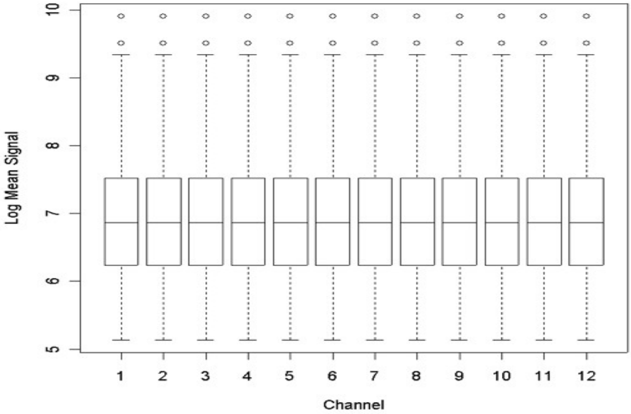
  where x_{ij} is measurement i on array j, G_j is the distribution function for array j, and F^{-1} is the inverse of the distribution function to be normalized.

  In practice, G_j estimated using the empirical distribution function and F is the average distribution across all arrays in the data set.

Details of Quantile Normalization

- Very easy to implement
  - Find the smallest log signal on each channel
  - Average the values from step 1
  - Replace each value in step 1 with the average computed in step 2
  - Repeat steps 1 through 3 for the second smallest values, third smallest values,..., largest values
- Quantile normalization changes expression over many slides i.e. changes the correlation structure of the data, may effect subsequent analysis.

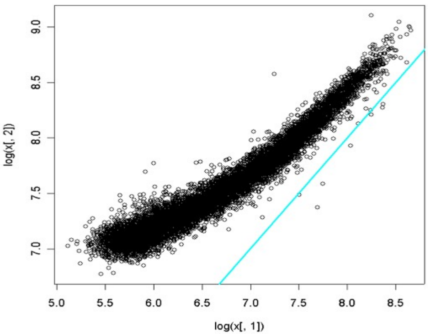
AN EXAMPLE



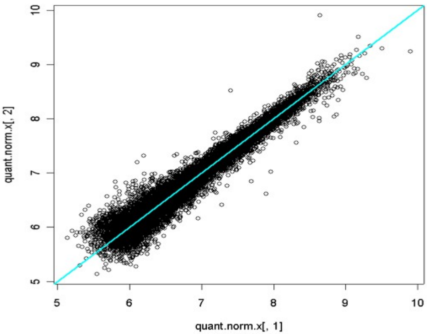
Boxplot of log signal means after quantile normalization

(Courtesy Dan Nettleton)

Another example



Original log means



After quantile normalization

(Courtesy Dan Nettleton)

Oligo arrays pre-processing: RMA

- RMA = Robust Multichip Analysis (Irizarry et. al., Bolstad et. al.)
- Implemented in R package: affy
- Other alternatives:
  - MAS 5.0: Affymetrix
  - Model Based Expression Index (MBEI): Li-Wong method, implemented in dChip
  - vsn (Huber et al., Rocke)
  - plier, plier+16 (Hubbell, new Affymetrix)
  - gcrma (Irizarry et al.)
- For a comprehensive list go to:

  `http://affycomp.biostat.jhsph.edu/`

RMA - I

- Use only PM, ignore MM (variant: gcrma)
- Background correct PM on raw intensity scale
- Quantile Normalization of $y_{ij} = \log_2(\mathrm{PM} - \mathrm{BG})$
- Assume additive model (on $\log_2$ scale):

  $$y_{ij} = \beta_i + \alpha_j + \epsilon_{ij}$$

  where (i, j) indexes array and probe respectively
  β_i = gene expression of the probe set on array i
  α_j = probe affinity affect for the jth probe in the probe set
  ϵ_{ij} = residual for the jth probe on the ith GeneChip
- Estimate (β_j, α_j) = chip and probe effect using a robust method
  - Median polish: quick
  - Robust linear model: yields quality diagnostics

RMA - II

- The parameters in the above equations are unidentifiable. Need constraint $\sum \alpha_j = 0$
- Perform Tukey's Median Polish on the matrix of y_{ij} values with y_{ij} in the ith row and jth column. Basically, entails iteratively normalizing row and column medians to 0 until convergence.
- Let $\hat{y}_{ij}$ denote the fitted value for y_{ij} that results from the median polish procedure
- Let $\hat{\alpha}_j = \hat{y}_{.j} - \hat{y}_{..}$ where $\hat{y}_{.j} = \sum_i y_{ij}/I$ and $\hat{y}_{..} = \sum_i \sum_j y_{ij}/IJ$ where (I, J) = number of arrays and probes
- Let $\hat{\beta}_i = \sum_j y_{ij}/J$
- Then, $\hat{\beta}_i$'s are the RMA measure of expression for array i

RMA - III

- RMA ignores MM values
  - MM values have information about both signal and noise; Typically 30% higher than MM; Subtracting them might lead to negative expression values; log?
  - Using it without adding more noise is challenging and is a topic of current research (gcrma)
  - Hope: possible to improve the BG correction using MM, without having the noise level increase greatly
- Multi-chip: to put each chip's measurements in the context of similar values
- Robust: to provide summaries that really improve over the standard ones by down-weighting outliers
- Conclusions of Irizarry et al: RMA was arguably the best summary in terms of bias, variance and model fit

Final Comments: Normalization

- Reduces systematic (not random) effects; makes it possible to compare several arrays
- There are many variations and extensions of the methods covered today. Its a still emerging field.
- Normalization affects the final analysis but not often clear which strategy is the best; normalization introduces more variability
- Two-step procedure vs. integrated normalization and analysis?
- Preprocessing can improve the quality of analysis, remove technical effects
- But bad data IS bad data!

FINAL DATA FOR ANALYSIS

- What statisticians work with: Gene Expression Matrix

Samples	Gene $\mathbf{1}$	Gene $\mathbf{2}$	$\cdots$	Gene $\boldsymbol{p}$
$\mathbf{1}$	$\boldsymbol{X}$	$\boldsymbol{X}$	$\cdots$	$\boldsymbol{X}$
$\mathbf{2}$	$\boldsymbol{X}$	$\boldsymbol{X}$	$\cdots$	$\boldsymbol{X}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{n}$	$\boldsymbol{X}$	$\boldsymbol{X}$	$\cdots$	$\boldsymbol{X}$

- $\boldsymbol{X}$ = Gene expression intensities (some form)
- $\boldsymbol{p}$ = Number of genes (usually in thousands)
- $\boldsymbol{n}$ = Number of samples (micorarrays) ($\boldsymbol{n} \ll \boldsymbol{p}$)
- $\boldsymbol{Y}$ (tissue type/phenotype) = 0 if Normal; 1 if Cancer (binary)
- $\boldsymbol{Z}$ = Design variables for controlled experiments (e.g. Drug A/B) OR Covariates

FINAL THOUGHTS

- Before complex statistical analysis some other preprocessing issues need to be addressed
- Few basic issues affecting quality of data to be analyzed (not covered today)
  - Variation: within and between arrays; identify areas of experimentation that require improvements
  - Design of experiments: understand “biology”

    Careful design on experiments; Kerr and Churchill (2001) examined variation due to array, dye, treatment (variety), gene and labeling design; ANOVA models for cDNA normalization

Gene Expression Variation

- “Biological” versus “Technological”
- Need replication!; three kinds
- Spot to Spot
  - Depositing probes for same genes multiple times on the array
  - Assesses within array variation
- Array to Array
  - Multiple hybridizations using same mix of RNA source
  - Assesses between array variation
- Subject to Subject
  - Sample multiple individuals
  - Assesses biological variation

Microarray Data: Statistician's View

- Experimental design
  - Choice of sample size; assignment of experimental conditions to arrays
- Signal extraction
  - Image analysis; gene filtering; probe level analysis for oligo arrays; normalization
- Data analysis
  - Gene selection; clustering and classification of biological samples and genes; dimension reduction
- Validation and interpretation
  - Comparisons across platforms; use of multiple datasets
- A last two points will be covered throughout the course

Moral of the Story

- Microarray data: powerful tools to understand basic biological processes
- Opened a plethora of interesting methodological statistical problems
  - Small n large p problems
- Careful review of procedures generating data; errors propagated
- Still evolving: new biology and new data analysis

LIST OF RESOURCES

- `http://www.bioconductor.org`: Open source software for the analysis of genomic data sets based upon R.
- `http://www.affymetrix.com; www.dchip.org`: Information about Affymetrix arrays and technology; alternate expression measures for Genechip data.
- `http://affycomp.biostat.jhsph.edu; rmaexpress.bmbolstad.com`: Benchmarking tool for comparing the performance of alternate expression measures for Genechip data; also windows GUI for RMA procedure.
- `http://www.stat.berkeley.edu/∼terry/zarray/`: cDNA arrays
- Of course our very own: `http://bioinformatics.mdanderson.org`