

Lecture 4

Calculating Expression Measures with Affymetrix Data

Stat 697K, CS 691K,
Microbio 690K

Expression Measure Summaries

- Summarize 20 PM,MM pairs (probe level data) into one number for each probeset (gene)
- We call this number an expression measure
- Affymetrix GeneChip Software has defaults.
- Do they work? Can they be improved?

2

From Spot Intensity to Expression Measure

- For Affymetrix arrays, there have been several approaches to summarizing probe-level data.
 - 3 of them have become standard, and are implemented in the R BioConductor package:
- 1) **Affymetrix Average Approach**: Affymetrix MicroArray Suite 5.0 (MAS 5.1) Software
 - 2) **Model Based Expression Index Approach (MBEI)**: Li & Wong (2001) PNAS **98**: 31-36
 - 3) **Robust Multi-Array Approach (RMA)**: Irizarry/Bolstad/Speed (2003) NAR **31**: e15

3

Expression Measure

- **Affymetrix average approach**
- Model Based Expression Index (MBEI) approach (Li & Wong)
- Robust Multi-Array approach (Irizarry, Bolstad & Speed)

4

OLIGONUCLEOTIDE MICROARRAYS (GeneChips)

Gene Sequence: 3' ————— 5'
Probe Sequences: — — — — —

Perfect match: A-C-T-G-T-T-T-A-C-G-C-T-C-A-G-T-C-G-G-G-T-C-A-A-T
Mismatch: A-C-T-G-T-T-T-A-C-G-C-T-A-A-G-T-C-G-G-G-T-C-A-A-T

Probe set: 11 to 20 probe pairs (PM & MM)
to interrogate each gene

There may be 5,000-20,000 probe sets per chip

Data and Notation

PM_{ijn} , MM_{ijn} = Intensity for perfect/mismatch
in chip i , probe j , gene n

$i = 1, \dots, I$ (chips, ranging from 1 to hundreds)

$j = 1, \dots, J$ (probes, usually 11)

$n = 1, \dots, N$ (genes, between 8,000 and 12,000)

6

Affymetrix Average Approach MAS 4.0

- Takes average of (PM - MM) for the 20 probes as the intensity measure for each gene

7

Affymetrix Average Approach (single chip method)

- Affymetrix MicroArray Suite 4.0 software (MAS 4.0) uses Average Differences: *Avg.diff.*

$$Avg.diff. = \frac{1}{|A|} \sum_{j \in A} (PM_j - MM_j)$$

for probe pair j , and A a set of suitable probe pairs chosen by the software. ($|A|$ is number of elements of a set)

- probe pair outliers are removed: > 3 SD from mean (PM-MM) value
- using PM/MM differences (PM-MM) eliminates most cross-hybridization signals

8

Limitations of MAS 4.0 Average Differences Method

- Can result in negative intensities if $MM > PM$
- Approximately 1/3 of MMs are greater than PMs
- Reasons for negative intensity
 - cross-hybridization
 - changing middle base does not change hybridization for some probes
 - MM for one gene is a PM for another gene
- Implemented in BioConductor package “affy”, `summary.method=“avgdiff”`

9

Affymetrix New Approach

Improved 2 things:

- 1) Uses a weighted average of probes
 - weighs outlier probes less
- 2) Fixed problem of $MM > PM$

10

Affymetrix's New Approach: called MAS 5.0 (single chip method)

- Affymetrix new software, MAS 5.0 uses a **weighted** average of (PM-MM), using Tukey's Biweight function.

$$\text{Signal} = \text{TukeyBiweight}\{\log(PM_j - MM_j^*)\}$$

with MM^* a version of MM that is never larger than PM.

http://www.stat.berkeley.edu/users/terry/zarray/Affy/GL_Workshop/genelomic2001.html
(see Hubbell 2001, see also AffyStatGuide.pdf)

11

Tukey's BiWeight Function

- Weights probes that are outliers from median, less

12

MM > PM

Affymetrix new rules:

- If PM > MM, the probe is used
- If only a few probes for a gene have MM > PM, these MMs are called uninformative
 - replaced with values from good probes (based on PM to MM ratios)
- If most MMs are uninformative for a gene, the gene is flagged and removed

13

Affymetrix's MAS 5.0 Approach: Tukey's Biweight Function

- New approach to avoid negative signals
 - negative values do not make physiologic sense
 - negative signals make log-transformations difficult
- Robust weighted mean that is insensitive to outliers
 - weights values closest to median the highest
- Uses (PM - **adjusted** MM),
where “**adjusted** MM” is set so the difference is not negative

14

Affymetrix MAS 5.0 Summarization Method

- Implemented in BioConductor package “affy”, `summary.method=“mas”`
- see “Background Notes” slides at end of this lecture for further notes on Tukey's Biweight function

15

Some Possible Problems with Affymetrix MAS 5.0 Summarization Method

What if

- A small number of the probe pairs hybridize much better than the rest?
- Changing the middle base in mismatch does not make a difference for some probes?
- Some MMs are PMs for some other gene?

16

Expression Measure

- Affymetrix average approach
- **Model Based Expression Index (MBEI) approach (Li & Wong)**
- Robust Multi-Array approach (Irizarry, Bolstad & Speed)

17

Model based approach to quantify gene expression

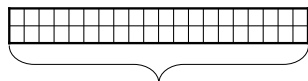
- Li and Wong proposed a model-based method for summarizing probe-level Affymetrix data
- See Cheng Li and Wing Wong, (2001) PNAS

18

Li & Wong Model (multiple chip approach)

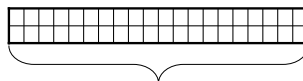
- Consider 10 chips for an experiment, and one gene.

$i = \text{chip } 1$



2x20 probe cells: PM and MM

$i = \text{chip } 10$



2x20 probe cells: PM and MM

Goal: Estimate the expression of this one gene in the 10 chips

Data: There are 2x10x20 measurements used to obtain estimates (10x20 PMs and 10x20 MM).

θ_i : Denotes "expression value" (signal) for this gene on the i^{th} chip,
will estimate $\theta_1, \theta_2, \dots, \theta_{10}$

19

Probe Intensity vs. Gene Expression

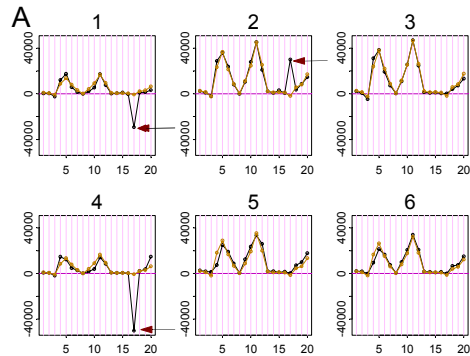
- Gene expression increases \Rightarrow probe intensities increase
- This relation is assumed linear, but rate is different for each probe.
- **Different probes measure gene expression differently**
 - some probes are higher quality than others
- Look at same probes on multiple chips

20

Probe 17 is not concordant with other probes:

could be due to cross hybridization

Future array design: remove probe 17



Probe outlier: large standard errors of ϕ_{17}

21

Li & Wong MBEI Model

Measures 2 things:

- 1) Probe quality measure for each probe for each gene, ϕ_j 's, i.e. 20 of these for 20 probes
- 2) Expression value for each gene on each chip, θ_i 's, i.e. 10 of these for 10 chips
 - Each θ_i is a weighted average of the probe level expression values multiplied by the probe quality value
 - higher quality probes get higher weight

22

Li & Wong MBEI Model

For ONE gene:

$$y_{ij} = PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij}$$

y_{ij} : probe - level expression for probe j on array i

θ_i : Expression signal for array i

ϕ_j : probe sensitivity (probe response) for probe j

$\varepsilon_{ij} \sim N(0, \sigma^2)$, $\sum \phi_j^2 = J$, J = total # of probes

23

Gene Expression is a Weighted Average

$$\theta_i = \frac{\sum_j y_{ij} \phi_j}{J}$$

y_{ij} = probe expression value

ϕ_j 's = probe quality measure

J = total number of probes

24

Li & Wong Model for Probe Level PM-MM differences

- Least square estimates (regression technique) for the parameters are carried out:
 - iteratively fit the set of θ 's, regarding ϕ 's as known,
 - then the set of ϕ 's, regarding θ 's as known.
- NOTE: Recommended to have at least 10 chips for this method

25

Why Model?

- Automatic handling of outliers
- Model produces standard error estimates of both probes (ϕ 's) and overall expression (θ 's) (see Li & Wong 2001)
- Standard errors can be used to detect:
 - Probes that are poor quality, i.e. outlier probes (large SE)
 - Due to: cross-hybridization, image contamination, other reasons
 - Outlier chips (large SE)
- Meta-analysis: pooled data from different experiments

26

Other Advantages of Model-Based Analysis

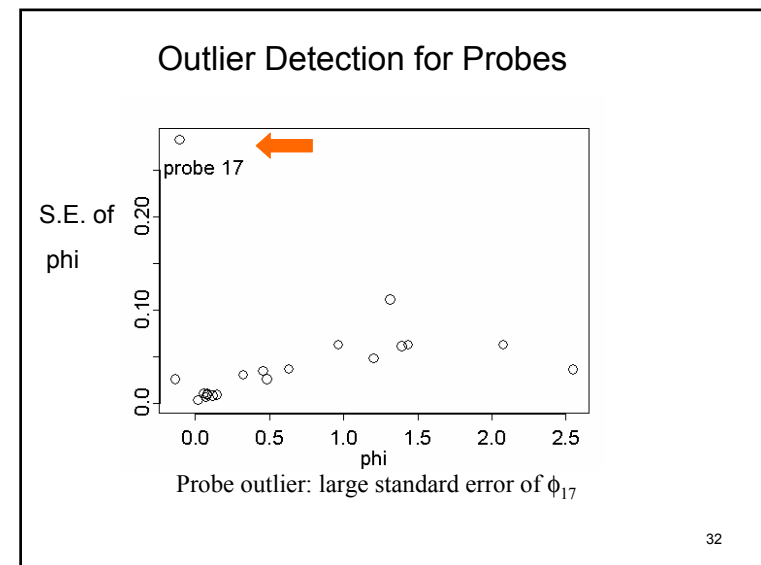
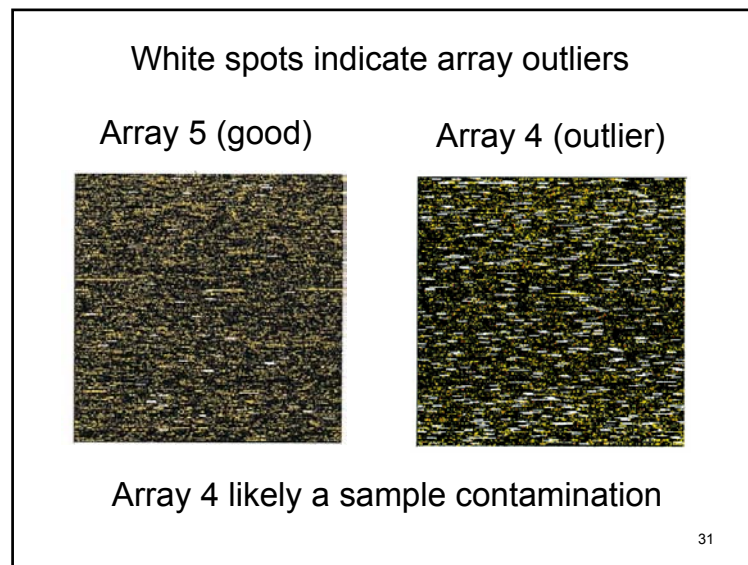
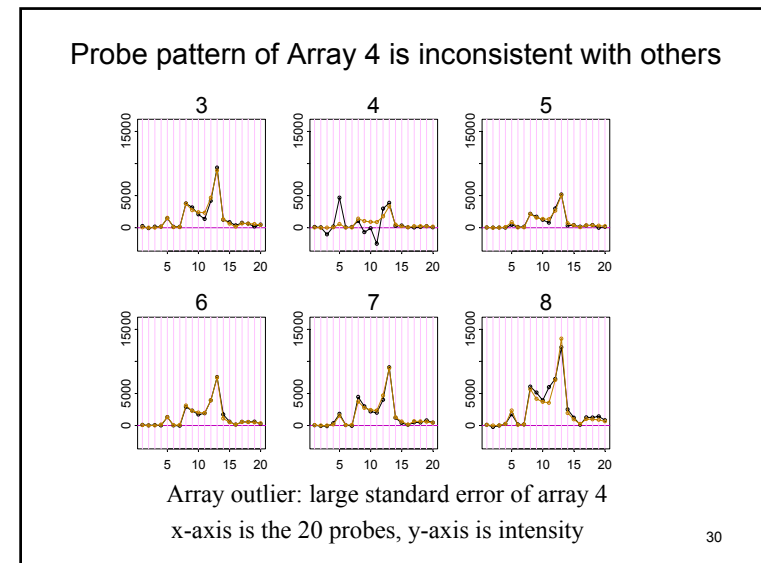
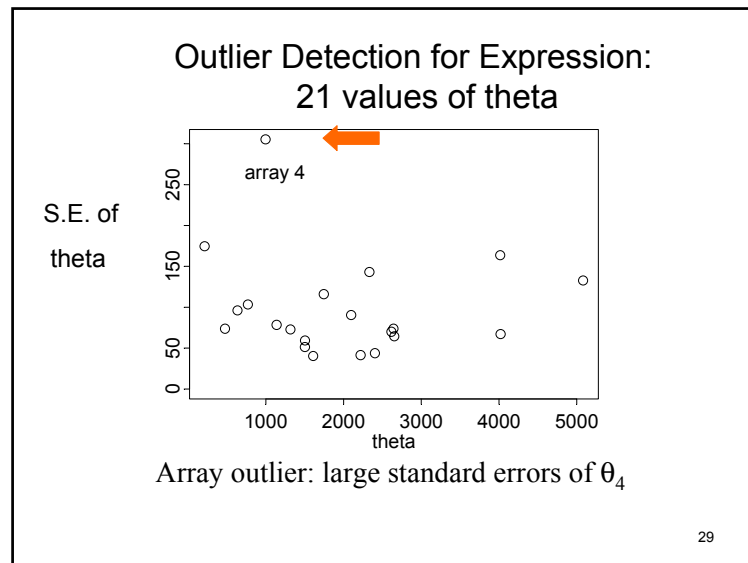
- Identifies good-quality probes
 - helps improve array design by removing low quality probes
- Saves time
 - humans don't need to look at large-scale studies by hand for bad quality chips, probes

27

Li & Wong use 21 HuGeneFL chips to illustrate their method

- The Affymetrix GeneChip HuGeneFL Array is a single array with 5,600 full-length human genes (initially released by Affymetrix in November, 1998).
- For each gene, there are 420 data points: 20 (PM-MM) probe values x 21 chips
- Model has 41 parameters for each gene: 21 θ 's, 20 ϕ 's

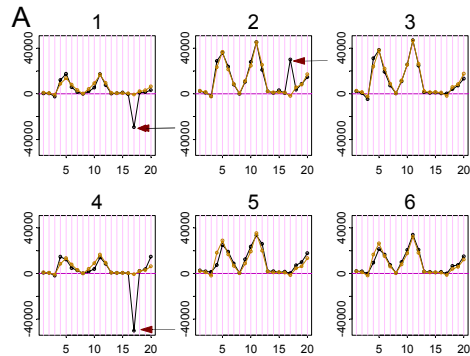
28



Probe 17 is not concordant with other probes:

could be due to **cross hybridization**

Future array design: remove probe 17



Probe outlier: large standard errors of ϕ_{17}

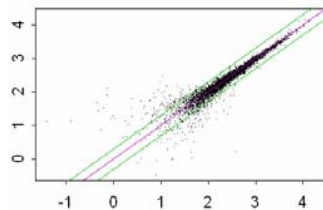
33

Comparison of Li & Wong MBEI with MAS 4.0 Average Difference

- Examined in Li & Wong 2001 Genome Biology
- Examined replicate arrays
- Good expression index should have ratios = 1 between genes on replicate arrays
- Found MBEI can detect low expression better than MAS 4.0
- Much lower variance across replicates for MBEI than MAS 4.0

34

(A) MEI method



(B) AD method

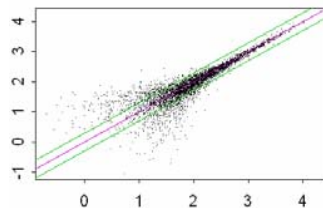


Figure 2.5. Log (base 10) expression indexes of a pair of replicate arrays (array 1 and 2 of array set 5, **brain tissue samples**) for MBEI method (A) and AD method (B). The center line is $y=x$, and the flanking lines indicate the difference of a factor of two.

- For replicate arrays, MBEI is closer to line $y=x$ than AD

35

Model-based Expression Index (MBEI) implementation

- Li & Wong model is applied to probe-level data **already normalized**
- Uses either (PM-MM) or PM only (default PM only)
- Implemented in publicly available software **dChip**
- Implemented in BioConductor package "affy", `summary.method="liwong"`

36

Expression Measure

- Affymetrix average approach
- Model Based Expression Index (MBEI) approach (Li & Wong)
- Robust Multi-Array approach (Irizarry, Bolstad & Speed)

37

Robust Multi-Array Approach (RMA)

Previous studies showed:

- Don't subtract or divide by MM
- PM-MM introduced too much noise, especially at the very low end
- Many probe pairs with MM much larger than PM
- About 1/3 of probes MM>PM
- Take logs

38

Robust Multi-Array Approach (RMA)

- RMA uses PM only
- Steps:
 - 1) Background correct the PM intensities
 - 2) Take \log_2 of background adjusted PM
 - 3) Normalize $\log_2(\text{PM}_{\text{bg-corrected}})$ using quantile normalization, with chips in suitable sets
 - 4) Conduct a robust multi-array analysis (RMA) of the quantiles

39

RMA Steps

1) RMA background correction

- Model based correction. We observe intensities:

$$O = S + N$$

where S = signal

N = noise

- RMA background model estimates S for each PM probe (see Irizarry et al. 2003, Bolstad et al. 2003 and BioConductor documentation for details)
- Does not correct MM probes

40

RMA Steps

- 2) Take \log_2 of background-corrected PM
- 3) Normalize $\log_2(\text{PM}_{\text{bg-corrected}})$ using quantile normalization, with chips in suitable sets
 - suitable sets include replicate chips or chips from similar experiments

41

RMA Steps

4) Robust Multiple-Array (RMA) Analysis of Quantiles

Assume additive model for each gene k

$$T(PM_{ij}^{(k)}) = e_i^{(k)} + a_j^{(k)} + \varepsilon_{ij}^{(k)}$$

$e_i = \log_2$ gene k 's expression on i^{th} array

$a_j = \log_2$ effect for j^{th} probe

ε_{ij} = error

T is the transformation that background corrects, logs and normalizes the PM intensities

Irizarry et al, NAR 2003

42

RMA Analysis

- The parameter of interest is:

$$e_i^{(k)}$$

this is the expression value for gene k on array i

- The parameters $a_j^{(k)}$ are adjustments to overall expression for each probe

43

RMA Analysis

- The RMA model is fit using a median polish algorithm (see *Exploratory Data Analysis*, Tukey 1977)
- Median polish is similar to an ANOVA model, but is robust to outliers
- see “Background Notes” slides at end of this lecture for notes on the median polish algorithm

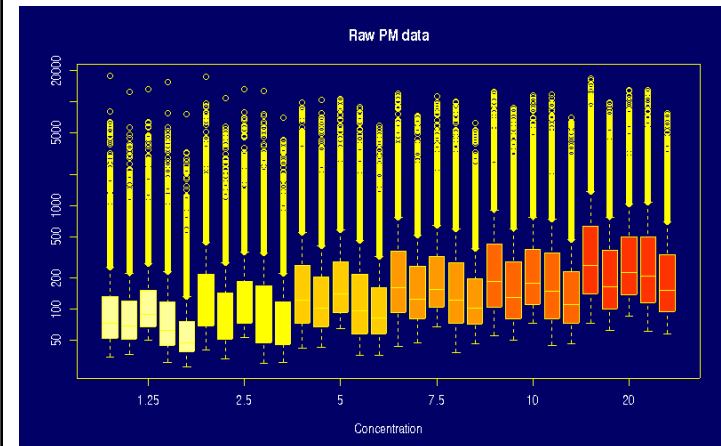
44

Example: Dilution Experiments

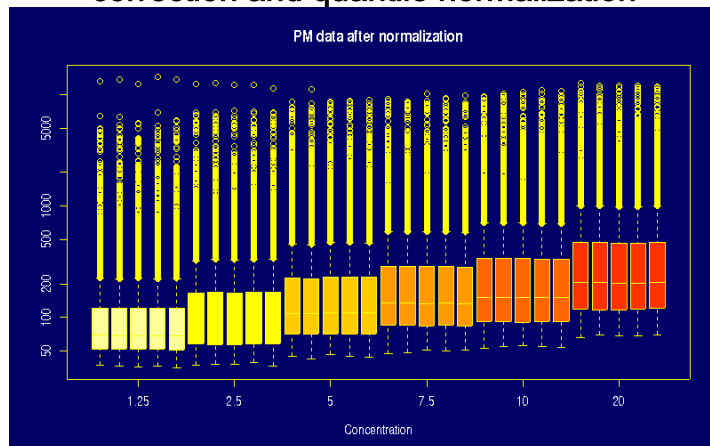
- cRNA hybridized to human chips (HG-U95A) in a range of proportions and dilutions
- Dilution series began at 1.25 μg cRNA per GeneChip array, and rose through 2.5, 5.0, 7.5, 10.0, to 20.0 μg per array.
 - 5 replicate chips were used at each of the 6 dilutions
- Normalization was performed within each set of 5 replicates
- Compute expression values for each gene at each dilution value

Irizarry et al, NAR 2003 45

Dilution experiment data



Dilution experiment data, after background correction and quantile normalization

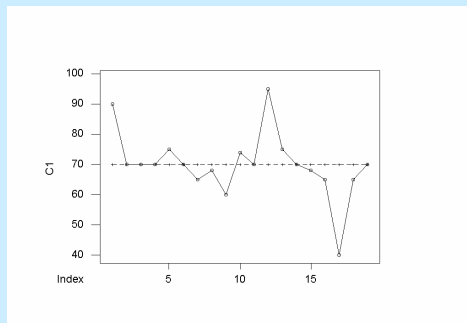


Comparing Methods by Standard Deviation

48

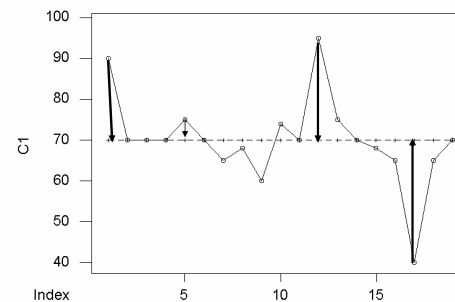
Measures of Variability

- The mean provides an idea of the baseline sample value.
- How representative of the sample is this value?



49

Variance



50

Computations

- Each sample point differs from the mean by some quantity:
 - Compute the differences, square them and sum all of them. This number gives an idea of how distant the sample points are from the mean.
 - We square the values to keep all values positive
 - Divide by (n-1) to “average the variability”. It is called the **sample variance**
 - Tells how variable the sample is with respect to the mean.

sample values x_1, x_2, \dots, x_n

$$\text{mean is } \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$$

$$\text{Sample variance } \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = s^2$$

51

Sample Standard Deviation

- The sample variance, s^2 , does not have the same dimension as the sample data:
 - Eg if data is gene expression, the sample variance is measured in (gene expression)x(gene expression)
- By taking the square root we have a number which has the same measurement unit as the sample.
- Sample standard deviation: $s = \sqrt{s^2}$

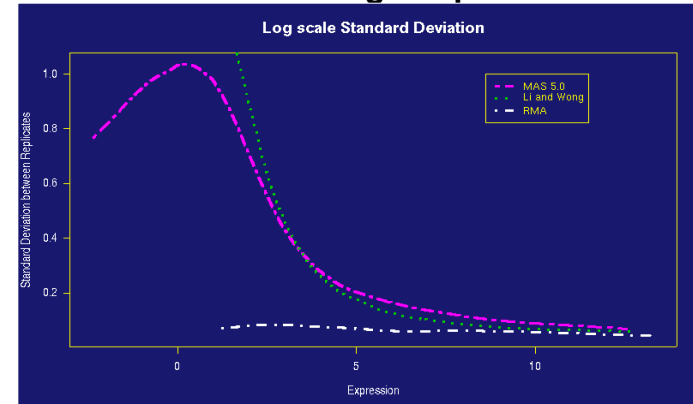
52

Assessing Precision of RMA vs. MBEI and MAS 5.0

- Computed standard deviation of expression for each gene across the 5 replicates
 - computed for RMA, MBEI, MAS 5.0
- Plotted S.D. vs. Average Expression for each gene
 - fitted a smooth curve through the points (all points for all genes)

53

SD vs. Average Expression



RMA has much lower standard deviation than MBEI, MAS 5.0

Source: Terry Speed

Comparison

- AvgDiff and MBEI (Li & Wong) sometimes underestimate expression; may be caused by subtracting MM
- RMA has less variance than all other measures at lower RNA concentrations
- RMA gives better estimates of standard errors of expression level than Li & Wong model
- **RMA is fine for 2 chips; 10 for Li & Wong**

55

RMA conclusions/suggestions

- MM could be created by changing more than one base in PM sequence
- Place mismatched bases in different positions than the middle position (Nimblegen chips)
- Use only PM
 - Allows space currently used for MM to be used for other PM
 - Allows twice as many genes to be printed on arrays

56

RMA Implementation

- Implemented in BioConductor package “affy”, summary.method=“**medianpolish**”
- Or use BioConductor package “rma”

57

Background Notes

Tukey's biweight function

- The **Tukey biweight function** is a robust location measure for the center of the data.
- It is even more robust than the median.
- Let $\mathbf{x}=(x_1, \dots, x_n)$ be a real-valued vector (e.g. expression values of probes)
- The Tukey biweight of \mathbf{x} is calculated as follows:
 - 1) Calculate the median M of x
 - 2) Calculate the **median of the absolute differences** of each datapoint to M ,
 $MAD(\mathbf{x}) = \text{median} (|x_1-M|, |x_2-M|, \dots, |x_n-M|)$.
 (this is a measure for the variability of the data)

Background Notes

3) Standardize the data:

$$y_j = (x_j - M) / (c * MAD(x) + \epsilon)$$

where ϵ is a very small constant which is introduced merely to avoid division by zero.

The constant c is 5 by default; it determines the robustness of the Tukey biweight.

4) Define the biweight function

$$w(t) = \begin{cases} (1-t^2)^2 & \text{for } |t| < 1 \\ 0 & \text{else} \end{cases}$$

here, values farther from median (outliers) will have smaller weight. Replace t with y 's.

Background Notes

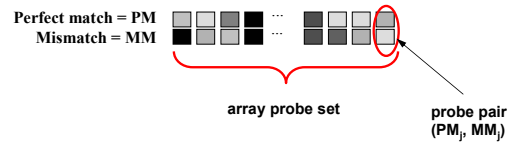
5) The Tukey biweight function $Tukey(x)$ is then:

$$Tukey(x) = \frac{\sum_{j=1}^n w(y_j) x_j}{\sum_{j=1}^n w(y_j)}$$

60

Background Notes

From spot intensity to expression value



Calculation of the overall signal intensity of probe set
 $P = (PM_1, \dots, PM_n, MM_1, \dots, MM_n)$:

$$v_j = \max(PM_j - MM_j, \delta), \delta = 2^{-20}$$

$$x_j = \log(v_j) \quad j=1, \dots, n, \quad x = (x_1, \dots, x_n)$$

$$Signal(P) = Tukey(x)$$

Remark: The formula as stated here is not correct for probes with $MM_j > PM_j$. In these cases, v_j is replaced by some non-negative value (for details see the Affymetrix technical manual).



Background Notes

Median Polish algorithm

1. Take the median of each row and record the value to the side of the row – subtract the row median from each value in that row
2. Compute the median of the row medians, and record the value as the overall effect. Subtract the overall effect from each of the row medians
3. Take the median of each column and record the value beneath the column. Subtract the column median from each value in that particular column
4. Compute the median of the column medians, and add the values to the current overall effect. Subtract this addition to the overall effect from each of the column medians.

www.spatial.maine.edu/~beard/

62

Credits

These slides are based in large part on lectures by Steve Qin, University of Michigan, with generous permission.

- Steve Qin
- Cheng Li
- Wing Wong
- Sandrine Dudoit
- Robert Gentleman
- Terry Speed
- Rafael Irizarry
- Ben Bolstad
- Yee Hwa Yang
- Rebecca Fry
- Leona Samson
- Fraunhofer Institute Algorithms and Scientific Computing
- Christina Kendziorski
- Kate Beard-Tisdale
- Paola Sebastiani

63