Comparing normalization methods based on splice array experiments

Jean Yee Hwa Yang

http://www.biostat.ucsf.edu/jean/
University of California, San Francisco

Acknowledgements

- UCSF
  - Yuanyuan Xiao
  - Mark Segal

- UCSC
  - Grant Hartzog
  - Todd Burcin

- UCB
  - Terry Speed
  - Sandrine Dudoit

- WEHI (Melbourne)
  - Natalie Thorne
  - Gordon Smyth
Outline

• Background
• Preprocessing
  – Stepwise normalization
• Experimental design of the splice arrays
• Comparison study
• Summary

Background
Nuclear RNA processing events

- 5’ capping
- 3’ cleavage and polyadenylation
- Intron removal – splicing
- mRNA transport to cytoplasm for translation

For Yeast
- ~ 6000 genes
- ~ 250 contain introns

Biological background

- Mutants
  - Spt4-5 -- chromatin specific elongation factors. Spt4d, Spt5.4, Spt5.194 and Spt5.242
  - Ceg1 – Capping enzyme mutant

- Long term question
  - How does the Spt4-Spt5 complex affect transcription elongation?
  - Investigate the role of Spt-Spt5 complex in splicing.

- Specific question
  - Identification of genes with splicing defects in mutant strains. i.e. Identify DE genes in the splice array.
Expression profiling with DNA microarrays

- cDNA “A” Cy5 labeled
- cDNA “B” Cy3 labeled

Hybridization

Laser 1
Laser 2

Scanning

Analysis

Image Capture

Array Layout
Splicing-specific microarrays

Intron-containing genes

exon 1  |  exon 2

Splicing

exon 1  |  exon 2

Intronless genes

Print Layout:
4 X 4 Print tips
15X24 Probes / Print tip
5760 Probes total

These mutants are defective for transcription elongation. 22 arrays were hybridized, scanned and quantified using GenePix.
Normalization

- This is known as the process of identifying and removing systematic variation not due to real differences between RNA treatments i.e. differential gene expression.

- These systematic variation can be observed from the dependence of ratios on
  - Fluorescent intensity (A)
  - Spatial (S) heterogeneity.
    - Print-tip.
    - 384-well plate.
    - Time order of print.

- Often, these dependencies are correlated with each other.

Preprocessing steps and options

- Which genes to use
  - All
  - Intronless
  - Exon

- Normalization methods
  - Ratios [two channels]
    - Median
    - Loess
    - Print-tip / pins
  - Intensities [single channel]
    - ANOVA
    - Quantile normalization
    - VSN
Within-slide normalization: adjusting A

• To correct for any dye-biases that commonly occur in cDNA microarrays.
  – Global normalization, median shift.
  – Robust linear normalization (local regression model) [Kelper et al Genome Biology 2003.]
  – An Intensity (A) dependent loess fit to log-ratios.
Adjusting A

Before

After

Within-slide normalization: adjusting S

• To correct for any spatial imbalance that commonly occur in cDNA microarrays.

  – Adjustment to print-tip-groups.
  – 2D-loess: Local spatial smoothing.
    [These are implemented in Bioconductor.]
  – ANOVA adjusting for rows and columns effect.
  – Use median filter to estimate and adjust for the spatial trend. Size of smoothing element is a 3 by 3 block of spots. [Ref Wilson et al Bioinformatics, 2003 and is implemented in a Rpackage “tRMA” which is available at http://www.pi.csiro.au/gena/tRMA]
Between-slide normalization: adjusting scale

- Here, we are concerned with making the single-channels between slides comparable.

- Quantile normalisation is based on the idea of normalising for equivalent medians or quartiles, requiring that every quantile across channels be equal and forcing the channels to have the same distribution.

- This distribution is estimated by the average of each quantile across all channels.

- [Ref: Natalie Thorne and Gordon Smyth have implemented this method in the Bioconductor package “limma”.]
Stepwise normalization

• Motivation:
  – Different slides within an experiment are similar but distinct from each other, therefore, we propose a data-specific normalization.
  
  – Avoid over fitting and introducing too much noise.
At each step, select the best model based on
\[ \text{BIC} = -2\log(L) + K\log(N) \]

This is an example of a print-tip median normalization
Experimental design of splice arrays
+
Comparison

Criteria for comparison

- It's often hard to use DE genes as the comparisons criteria, unless we have a set of spike-ins.
- Splice arrays are constructed arrays that can be used to compare different normalisation methods.
Experimental design – Target Samples

These mutants are defective for transcription elongation. 22 arrays were hybridized, scanned and quantified using GenePix.

Array Layout
Splicing-specific microarrays

Array layout

- **Probes:**
  - ~ Examine 260 genes
  - 40mer oligonucleotides from SJ, Int, Exon and Intronless and 4 replicates for each gene.
  - ~ 1100 SJ
  - ~ 1100 Int
  - ~ 1100 Exon
  - ~ 800 Intronless

- **Print Layout:**
  - 4 X 4 Print tips
  - 15X24 Probes / Print tip
  - 5760 Probes total

Probe of interest

Use for self-normalization

Without using exon information

\[
M = \log_2 \left( \frac{S_{J_{mut}}}{S_{J_{wt}}} \right)
\]

Assumption:
We assume that the probes are Close to each other on the slides

Self normalization and Index forming

\[ M_{mt} = SJ \text{ index} = \log_2 \frac{SJ_{mut}}{SJ_{wt}} \cdot \frac{Ex_{mut}}{Ex_{wt}} \]

Assumption:
We assume that the probes are close to each other on the slides.


Criteria for comparison

• We use SJ-Index as the standard and compared the various normalization to SJ-Index based on Euclidean distance.

• For each gene,
  – Observed \( \log_2 SJ_{obs} = \log_2 SJ - C1 \).
  – Observed \( \log_2 Ex_{obs} = \log_2 Ex - C2 \).
  – We assume.
    • \( C1 = C2 \) and
    • \( E (\log_2 (Ex_{MT} / Ex_{WT})) = 0 \).
Normalization methods

- Assume there are no Exons (or gene-specific controls) on the arrays. This is the case for most experiment, only the probe of interest are spotted (i.e. SJ probes).
  - No Normalization
  - Median = Global median.
  - Loess = Global loess fit.
  - PrintTip = Print-tip loess.
  - CSIRO = Spatial method proposed by Wilson et al
  - VSN = VSN method proposed by Huber et al
  - Quantile = Quantile normalization (this method adjust for between arrays).
  - Step = Stepwise normalization.
Summary

- Controls spots are essential to validate assumptions before individual normalization.

- We could include comparisons where we use a subset of genes for normalization rather than “all genes”.

- The assumption $E(\log_2(\frac{E_{MT}}{E_{WT}})) = 0$ may not hold as it may contain biological variation.