



BeeSpace



Statistical Analysis for Expression Experiments

Heather Adams

BeeSpace Doctoral Forum
Thursday May 21, 2009

Analytical Workflow

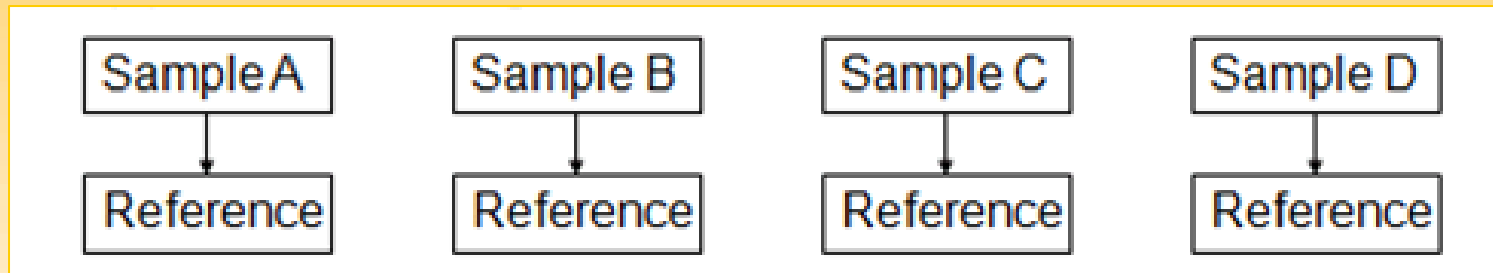
- Analysis of gene expression experiments:
 1. Experimental design
 2. Data preprocessing
 3. Inference (detection of D.E. genes)
 4. Classification
 5. Validation of results

Experimental Design

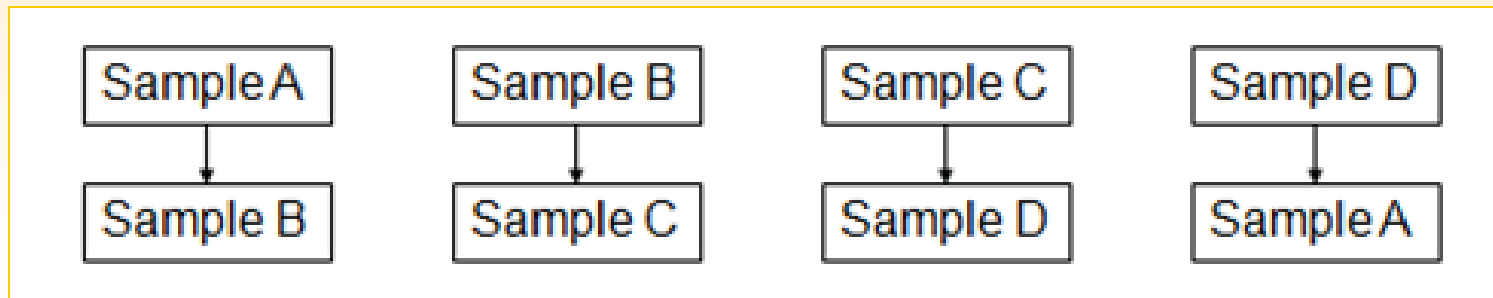
- What is the plan?
 - Assign samples
 - Technical and biological sources of variation and confounding effects are considered
 - Microarray design
 - Two-dye spotted microarrays: Reference and Loop

Experimental Design

- Reference Design

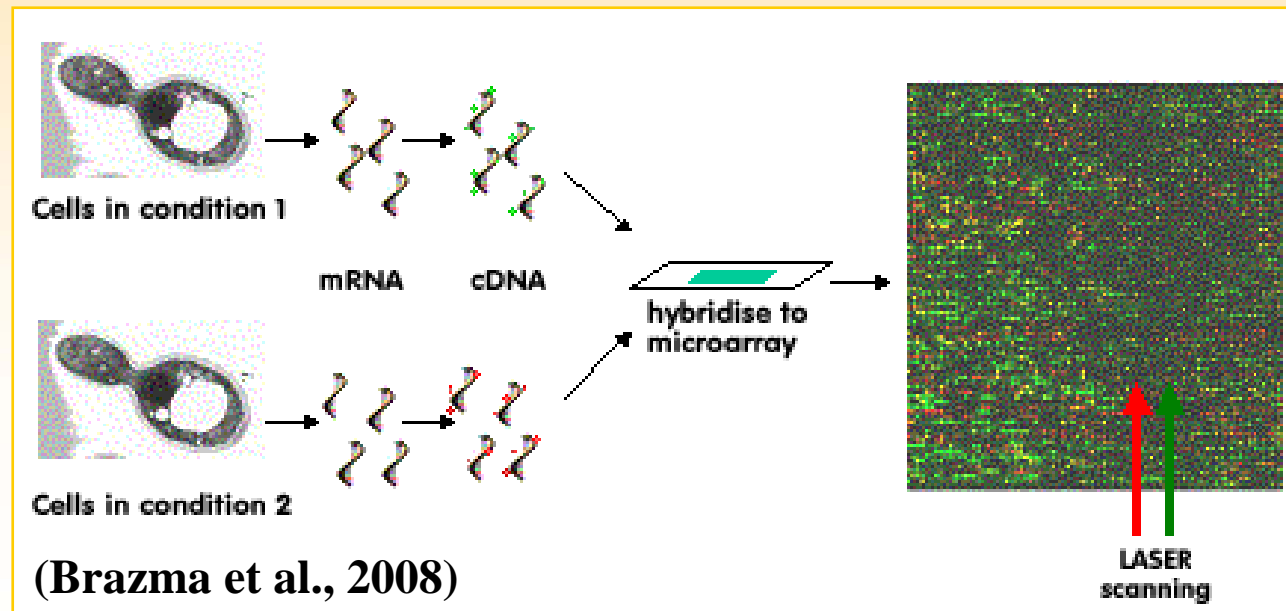


- Loop Design



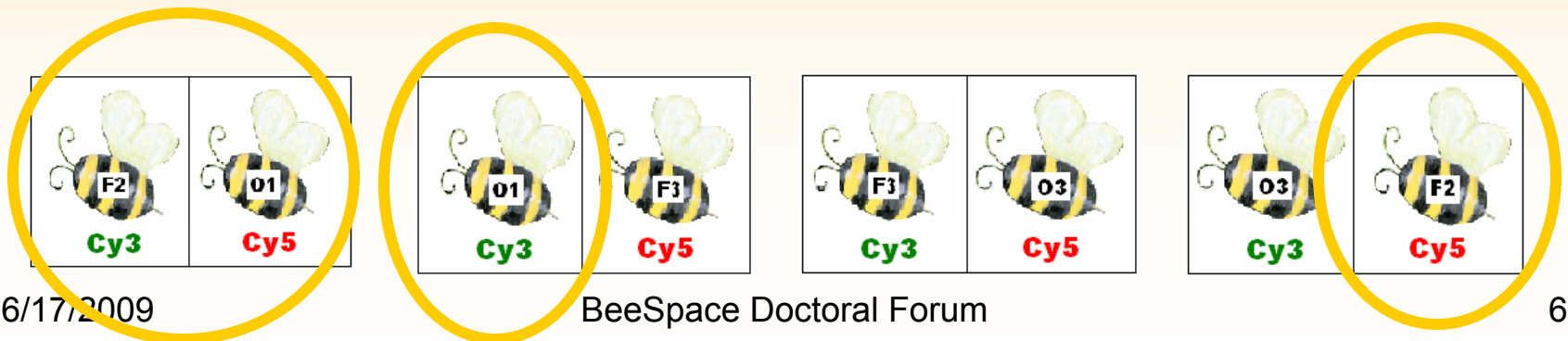
Review of microarray...

- Spotted 2-dye microarrays:
 - mRNA samples are reverse-transcribed into complementary DNA (cDNA) and labeled with red (Cy5) and green (Cy3) fluorescent dyes
 - Samples then hybridized with arrayed probes/DNA sequences
 - Relative abundance of the spotted sequences can be measured (high intensity = high expression, low intensity = low expression)



Review of microarray...

- Two measurements per spot (Cy5, Cy3)
 - Ratio of intensities for each spot represents the relative abundance of corresponding DNA sequence
- Comparison of stages with reverse dye labeling
- Example:



Data Preprocessing

- Remove unreliable observations and systematic sources of variation
 - Feature extraction
 - Data filtering and transformation
 - Normalization
 - Collapsing



Data Preprocessing

- Feature extraction
 - Identify pixels on the image that are part of the feature
 - Identify nearby pixels that are used to calculate the background
 - Calculate numerical information:
 - Signal and background means, medians, standard deviations, pixels
 - Identify flagged spots

Data Preprocessing



- Data Filtering
 - Remove flagged spots
 - Flag – indication of the quality of the feature spot
 - Bad feature: pixel standard deviation is very high relative to the pixel mean
 - Negative feature: signal of the feature is less than the signal of the background
 - Dark feature: signal feature is very low
 - Manually flagged feature: user-flagged feature

Data Preprocessing

– Interpretation of flag values

- If value of flag is zero, feature is good
- Non-zero values suggest problems with the feature
- Common flag threshold values
 - -100, -75, -50



Data Preprocessing

– Background subtraction

- Subtract the background signal from the feature intensity
- What if background is higher than feature?
 - Use lowest available signal-intensity measurement, which is typically 1
 - If background > foreground, gene has little or no expression, and will be filtered out regardless

– Intensity threshold criterion

- Pixel intensity 1 to 60,000 (saturated)
- Remove cDNAs with average expression intensity < 200 (on average)

Data Preprocessing

- Transformation of raw intensities
 - Logarithmic transformation makes the data exhibit a more Normal distribution
 - Log base 2 (\log_2)
 - 2-fold up-regulated genes \rightarrow log ratio of +1
 - 2-fold down-regulated genes \rightarrow log ratio of -1
 - Non-differentially expressed genes \rightarrow log ratio of 0

Data Preprocessing

- Normalization
 - Allows for joint analysis of gene expression measurements from different dyes and microarrays
 - Used to minimize technical variation in gene expression profiles
 - Green dye typically has higher intensity than red

Data Preprocessing

- Normalization

- Global

- On entire dataset
 - Assume red (Cy5) and green (Cy3) intensities are related by a constant
 - Center of the distribution of log ratios is shifted to equal zero

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

(Yang et al., 2002)

OR

$$Z_{rk} = \log_2 (Y_{rk} + C)$$

$$Z_{gk} = \log_2 (Y_{gk} + C)$$

(Kerr et al., 2002) 14

Data Preprocessing

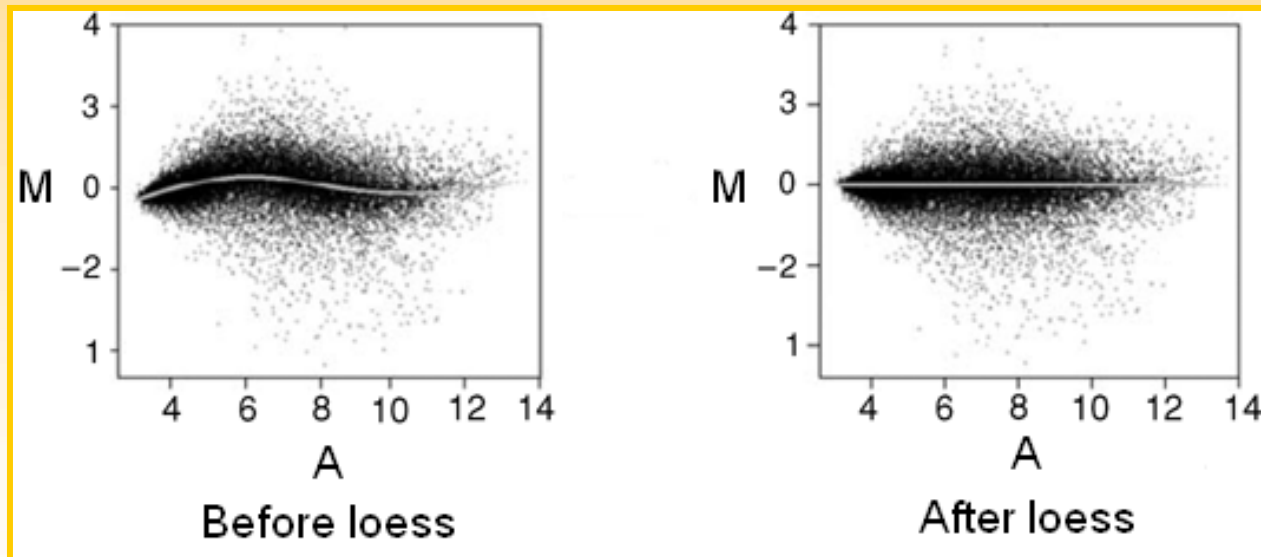
- Normalization
 - Local
 - On a physical subset of the dataset (apply to single microarray or by block)
 - Remove systematic variation within and across microarrays
 - Sources of variation: dye, pin, microarray area effects
 - Remove systematic noise on a more specific level (compared to global)

Data Preprocessing

- Normalization
 - Local
 - LOESS/LOWESS
 - curve-fitting local dye normalization method where a local regression line is fitted to the ratio R/G (**M**) versus the average of the R and G intensities (**A**), and the data is re-centered
 - LOWESS – linear function is used to fit the regression
 - LOESS – quadratic function is used to fit the regression

Data Preprocessing

- Normalization
 - Loess



Data Preprocessing

- Collapsing
 - Platforms containing multiple spots of the same probe, replicate spots next to each other, etc, then these observations tend to be highly correlated
 - Solution? COLLAPSE THE DATA
 - Average or median of the normalized fluorescence intensities from replicate spots
 - One observation per gene becomes available for analysis

Inference

- Analysis of preprocessed measurements to identify D.E. genes
 - Model fitting
 - Hypothesis testing
 - Characterizing functions

Inference

- Model fitting
 - Fixed effect
 - The levels of a factor have been predetermined for the experiment (dye, treatment)
 - Random effect
 - The levels of a factor have been randomly selected from a population of possible levels (microarray, sample)
 - Covariate
 - An independent factor that is not manipulated by the experimenter, but still affects the response

Inference

– Mixed effects model:

$$y_{ijklm} = \mu + D_i + T_j + S_k + A_l + e_{ijklm}$$

$$S \sim (0, \Sigma)$$

$$A \sim (0, \Delta)$$

- y_{ijklm} is the gene expression measurement for dye i ($i = \text{red or green}$), treatment j , sample k , and microarray l
- μ is the overall mean response
- D is the effect of dye l
- T is the effect of treatment j
- S is the effect of sample k
- A is the effect of microarray l
- ε is the error term

Inference

- Hypothesis testing
 - T-test
 - Experiments including two levels of a factor
 - Test statistic compared to t-distribution, and p-value is generated
 - F-test
 - Experiments including more than two levels of a factor, or more than two groups (ANOVA)
 - Test statistic compared to F-distribution, and p-value is generated



Inference

- Hypothesis testing
 - P-value
 - If less than threshold (confidence level), then measurement/gene is significant
 - $< 0.05, 0.01, 0.001, 0.0001$
 - Different thresholds lead to varying numbers of genes found to be important to the question at hand

Inference

- Hypothesis testing
 - Multiple test adjustment
 - Microarray experiments contain hundreds or thousands of genes/probes, which increases the probability of the false positive rate
 - Issue is overcome by multiple test adjustment techniques such as:
 - False Discovery Rate (FDR)
 - Bonferonni adjustment
 - Permutation tests

Inference

- Characterizing functions
 - Gene Ontology (GO) analysis
 - Characterize molecular functions, biological processes or cellular components of differentially expressed genes
 - Interpretation of data from genomic standpoint

Classification

- Classification approaches used to identify distinctions between groups of observations (genes, samples, treatments)
 - Hierarchical clustering, K-nearest neighbors (KNN), etc.

Validation

- Additional verification of differentially expressed genes to clear false positives
 - Real-time PCR
 - Informal application – literature review of similar studies

Thank you!

