Identifying and displaying differentially expressed genes from microarrays

George Bell, Ph.D.
Bioinformatics and Research Computing

Hot Topics – October 2008

Outline

- Rationale
- Identifying differentially expressed genes
- Sample microarray study
- Displaying differentially expressed genes

Why?

- Most microarray experiments are performed to identify targets of transcriptional regulation.
- What transcripts have differential abundance between cell types and/or treatments?
- How confident are we that they really are transcriptionally regulated?
- How much change do they exhibit?
- How can we display what's going on?
- How can we make sense of the results in the context of our biological interests?

Measuring differential expression

- Magnitude of fold change
- Magnitude of variation between samples
- Traditional statistical measures of confidence
  - T-test
  - Moderated t-test
  - ANOVA
  - Paired t-test
  - Non-parametric test (Wilcoxon rank-sum test)
- Other methods

Fold change

- Advantage: Fold change makes sense to biologists

$$\text{Fold change} = \frac{\text{expression value in sample 1}}{\text{expression value in sample 2}}$$

- What cutoff should be used?
- Should it be the same for all genes?
- Disadvantages:
  - Only mean values – not variability – are considered
  - Genes with large variances are more likely to make the cutoff just because of noise
- A log-transformation may be useful to compare up and down genes.

Statistical testing with the t-test

- Considers mean values and variability
- Equation for the t-statistic in the Welch test:

$$t = \frac{mean_r - mean_g}{\sqrt{\frac{s_r^2}{n_r} + \frac{s_g^2}{n_g}}}$$

… and then a p-value is calculated
r ; g = data sets to compare
s = standard deviation
n = no. of measurements

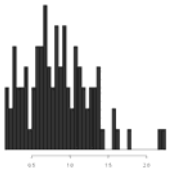
- Disadvantages:
  - Genes with small variances are more likely to make the cutoff
  - Works best with larger data sets than one usually has

Flavors of the t-test

- Are we only considering up-regulated or down-regulated genes, or both?
  - If both, perform a 2-tailed test
- Can we assume that the variance of the gene is similar in both samples?
  - Yes => Homoscedastic (the standard t-test)
  - No => Heteroscedastic (Welch's test)
- For specific experimental designs: paired t-tests
- Moderated t-tests: pool variation data for many genes

Moderated t-tests

- If a standard t-test is performed on each set of data (for each gene) separately, some genes will appear to be less or more variable just by chance.
- Can we use data from the whole array to better estimate the variation for each gene?
- Perhaps: Shrink each gene's sd towards that of a pooled sd for all genes
- A moderated t-test is available in MeV, SAM, and Bioconductor



$$t = \frac{\bar{x}_1 - \bar{x}_2}{s + s_0}$$

ANOVA

- Compares multiple groups at once (instead of just 2)
- Measures effect of multiple treatments and their interactions
- A thoughtful ANOVA design can help answer several questions with one analysis
- ANOVA can also analyze factors that should be controlled – just to confirm absence of confounding effects
- ANOVA generally identifies genes that are influenced by some factor – but then post-hoc tests must be run to identify the specific nature of the influence
  - Ex: t-tests between all pairs of data

Bootstrap and permutation testing

- Powerful parametric and non-parametric statistical tests
- Does not assume a normal distribution but does require a lot of computer time
- Example: Compare means of two sets of data while creating a custom distribution
  - Sample and/or shuffle data and calculate t or other statistic
  - Repeat at least 1000 times
  - Calculate the p-value
- Implemented in software like MeV and SAM

Multiple hypothesis testing

- We need both sensitivity and specificity:
  - Sensitivity: probability of successfully identifying a real effect
  - Specificity: probability of successfully rejecting a nonexistent effect
  - These are inversely related.
- The problem
  - The number of false positives greatly increases as one performs more and more t-tests
  - How seriously do you want to limit false positives?
- Correcting p-values for False Discovery Rate addresses this problem.

Combining p-values and fold changes

- What's important biologically?
  - How significant is the difference?
  - How large is the difference?
- Both amounts can be used to identify genes.
- What cutoffs to use?
- How many genes should be selected?
- Where are your positive controls?
- Moderated t-tests are somewhere between fold change (ignoring sd) and t-tests (gene-specific sd).

Statistical vs. biological significance

- Statistical significance sets out to find how often a given result could be produced by chance.
- Biological significance sets out to find results that are interesting relative to the problem under investigation.
- An assumption is that statistical significance can help lead us to biological significance.
- Convention, not theory, supports 95% percent confidence ($p < 0.05$) as the only choice.

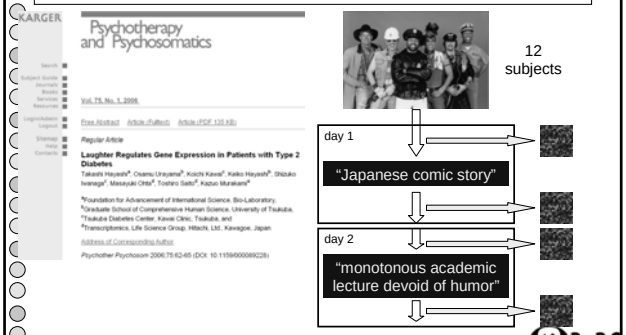
Differential expression - summary

- Multiple methods can produce lists of differentially expressed genes
- Which ways make most sense biologically and statistically?
- Be aware of multiple hypothesis testing
- Where do your positive controls fit in?
- There may be no single best way
- Is what you find biologically meaningful?

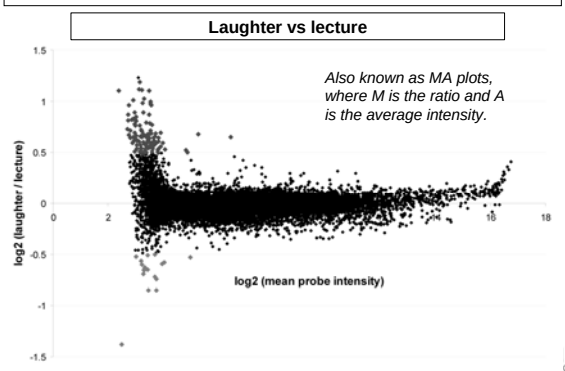
Displaying some or all genes

- What does all of your data look like?
- Are the arrays of good quality? consistent?
- Did you preprocess (normalize) correctly?
- How could/should you choose differentially expressed genes?
- What patterns do you see in differentially expressed genes?
- What could/should you do next?

Sample dataset: GSE1322

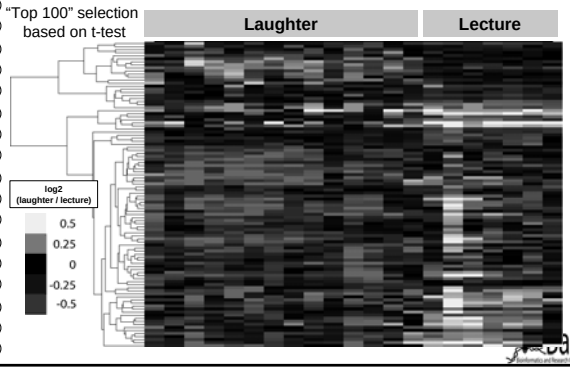


Ratio-intensity plots



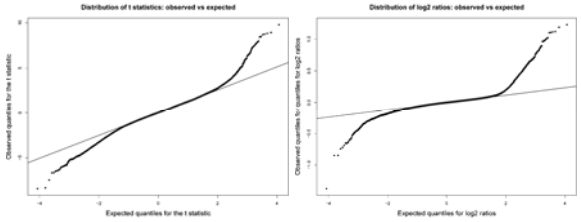
Volcano plots

Heatmaps



Q-Q (quantile-quantile) plots

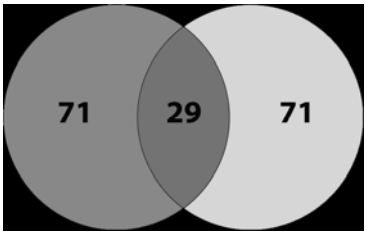
- How does our distribution of data differ from a normal distribution?



- Can this help us select an appropriate threshold?

Effect of moderated vs standard t-test

- How do the top 100 (by p-value) differentially expressed genes compare?



Display - summary

- Does you data look correct?
- Can you find evidence of normalization issues?
- Is the distribution what you'd expect?
- Does your definition of differential expression make sense?
- Do your figures lead you to subsequent analysis steps?

Microarray tools

- Excel
- Bioconductor (modules from R statistics package)
  - http://www.bioconductor.org/
- MeV: MultiExperimentViewer (part of TM4 suite)
- SAM (Significance Analysis of Microarrays; Stanford)
- Cluster 3.0 and Java TreeView
- BaRC analysis tools:
  - http://jura.wi.mit.edu/bioc/tools/