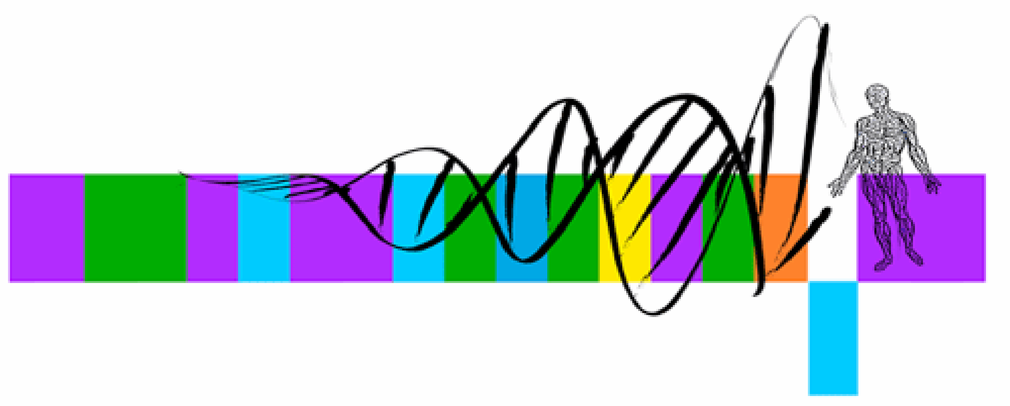


Analysis of Microarray Data

Lecture 2:

Differential Expression, Filtering and Clustering

George Bell, Ph.D.
Senior Bioinformatics Scientist
Bioinformatics and Research Computing
Whitehead Institute

Outline

- Review
- Measuring differential expression
- Multiple hypothesis testing
- Gene filtering
- Measuring distance between profiles
- Clustering methods

Review

- Assumption: Expression microarrays measure specific mRNA levels
- Why perform the experiment?
- What best design addresses your goals?
- Normalize to increase power of comparisons.
- Precision doesn’t necessarily indicate analysis success.
- Does your analysis pipeline make sense biologically and statistically?

Caveats and limitations

- Are the probes on the chip for a specific transcript? gene?
- Are mRNA levels correlated with transcription activity?
- Is transcriptional regulation important?
- Are mRNA levels correlated with protein activity?
- Is this the best technology to answer your question(s)?

Measuring differential expression

- One common goal is to rank all the genes on a chip in order of evidence for differential expression
- Ways to score genes:
  - Fold change
  - T-statistic p-value
  - Another statistic (nonparametric, etc.)
  - A combination of several scores

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

Hypothesis testing

- We may want to test …
  - Is the expression of my gene different in a set in one condition compared to another condition?
  - How big is the difference?
  - Is the mean of one set of values different from the mean of another set of values?
  - If we say “yes”, how much confidence do we have that the means are truly different?
- Assumptions:
  - Data are normally distributed
  - Samples are randomly chosen

Hypothesis testing with the t-test

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

$$t = \frac{\text{mean}_r - \text{mean}_g}{\sqrt{\frac{s_r^2}{n_r} + \frac{s_g^2}{n_g}}}$$

… 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)
- Moderated t-tests: pool data for many genes
  - Significance Analysis of Microarrays (SAM)
  - Limma (Bioconductor)

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

ANOVA

- Analysis of variance – like a multidimensional t-test
- Measure 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

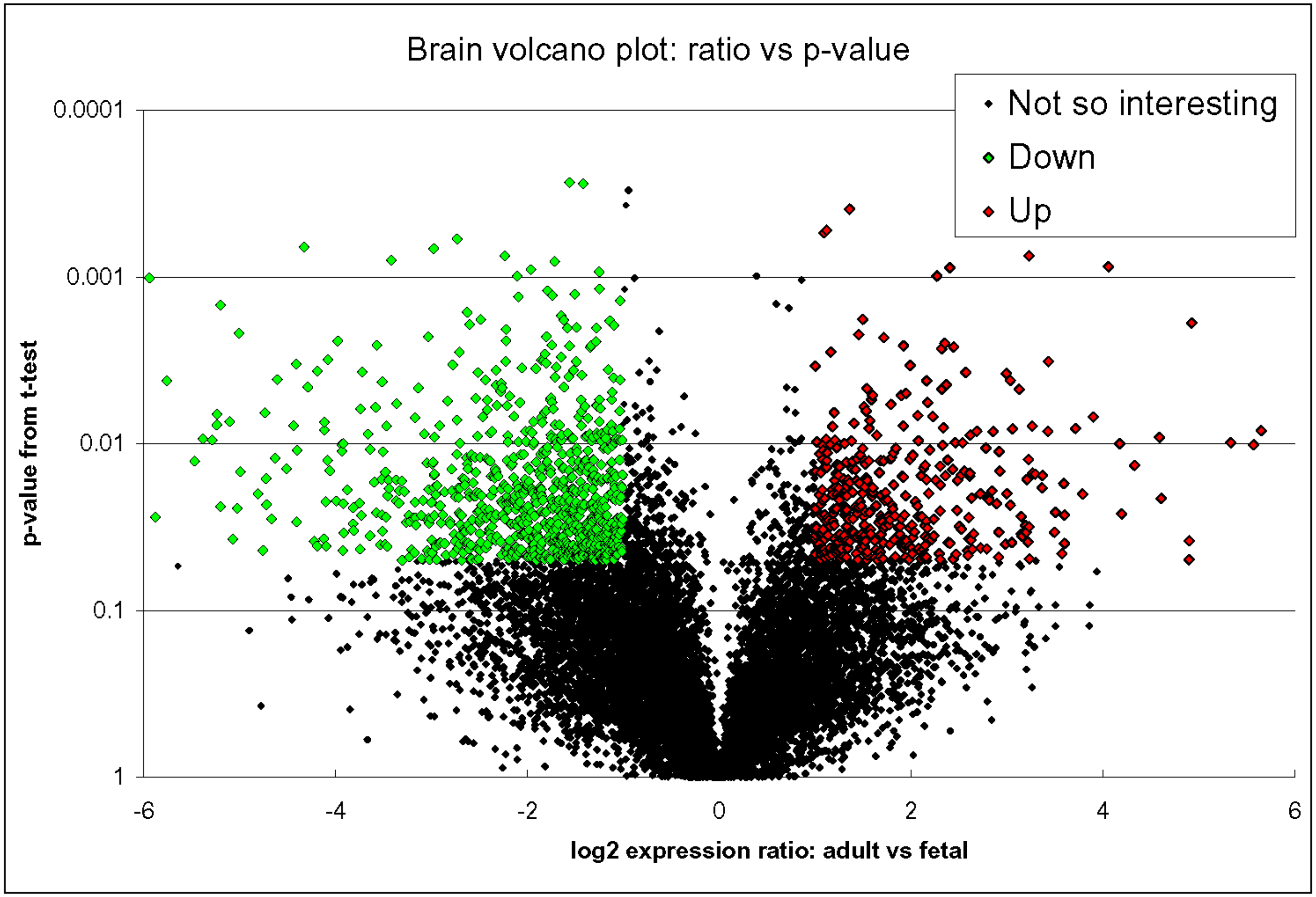
Bootstrap analysis

- Powerful non-parametric statistical tests
- Do not assume a normal distribution but do require a lot of computer time
- Example: Compare means of two sets of data while creating a custom distribution
  - Shuffle data and calculate t statistic
  - Repeat at least 1000 times
  - How often is the result more extreme that the real data?
- Calculate the p-value from your distribution

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 do something like this.

Volcano plots



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
- Looking at all the data: volcano plots
- Where do your positive controls fit in?
- There may be no single best way

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?

Why correct for multiple hypothesis testing?

Number of genes tested (N)	FP incidence $(p < 0.05)$	Probability of >= 1 FPs $100(1 – 0.95^{N})$
1	1 / 20	5%
10	10 / 20	40.1%
100	100 / 20	99.4%

FP = false positive

Correcting for multiple hypothesis testing

- If false positives are not tolerated
  - Perform Bonferroni correction
  - If you perform 100 t-tests, multiply each p-value by 100 to get corrected (adjusted) values

    $p = 0.0005 \quad => \quad p = 0.05$
- If false positives can be tolerated
  - Use False Discovery Rate (FDR)
  - If you can tolerate 15% false positives, calculate FDR p-values and then select 0.15 as your threshold
- FDR method is less conservative than Bonferroni and usually more appropriate for microarrays.

Performing a FDR correction

- Sort list of p-values in increasing order
- Starting at the bottom row,

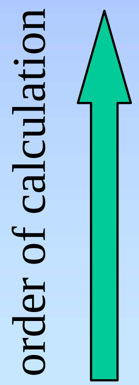
  corrected p-value = the minimum between

  1: raw p-value * (n/rank)

  2: corrected p-value below

  – n is the number of tests

  – rank is the position in the sorted list
- Example: a microarray assays 5 genes for differential expression



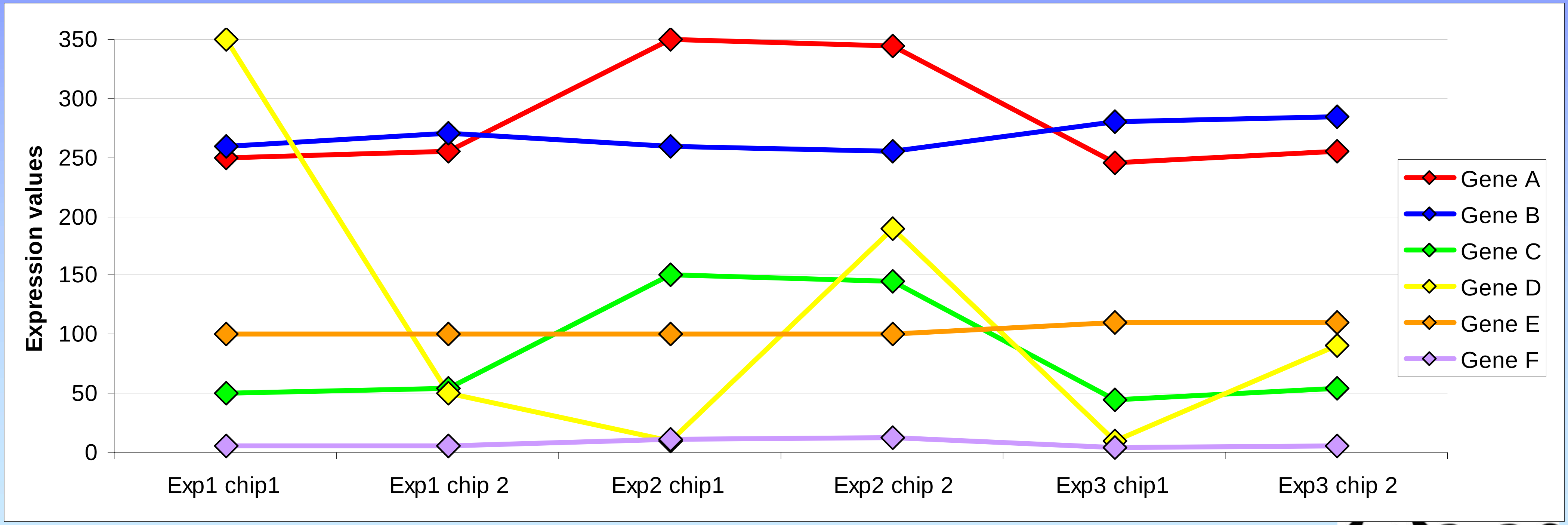
Gene	Rank	Raw p-value	Formula	Corrected p-value
C	1	0.001	min (0.001 * (5/1), 0.0125)	0.005
A	2	0.005	min (0.005 * (5/2), 0.017)	0.0125
B	3	0.01	min (0.01 * (5/3), 0.063)	0.017
E	4	0.05	min (0.05 * (5/4), 0.1)	0.063
D	5	0.1	0.1 * (5/5)	0.1

Gene filtering

- An infinite number of methods can select “interesting” genes
- Not all genes on the chip need consideration: any meaningful selection is possible
- Filtering by function: using GO or other annotations
- Often the major question: How many genes to choose for further analysis?

Measuring distance between profiles

- Distance metric is most important choice when comparing genes and/or experiments
- What are you trying to do?



Common distance metrics

- Pearson correlation
  - Measures the difference in the shape of two curves
  - Modification: absolute correlation
- Euclidean distance: multidimensional Pythagorean Theorem
  - Measures the distance between two curves
- Nonparametric or Rank Correlation
  - Similar to the Pearson correlation but data values are replaced with their ranks
  - Ex: Spearman Rank, Kendall’s Tau
  - More robust (against outliers) than other methods

Clustering and segmenting

- Goal: organize a set of data to show relationships between data elements
- With microarray analysis: genes and/or chips
- Most data does not inherently exist in clusters
- Clustering vs segmenting
- Most effective with optimal quantity of data
- Interpretation of data in obvious clusters: is it filtered?

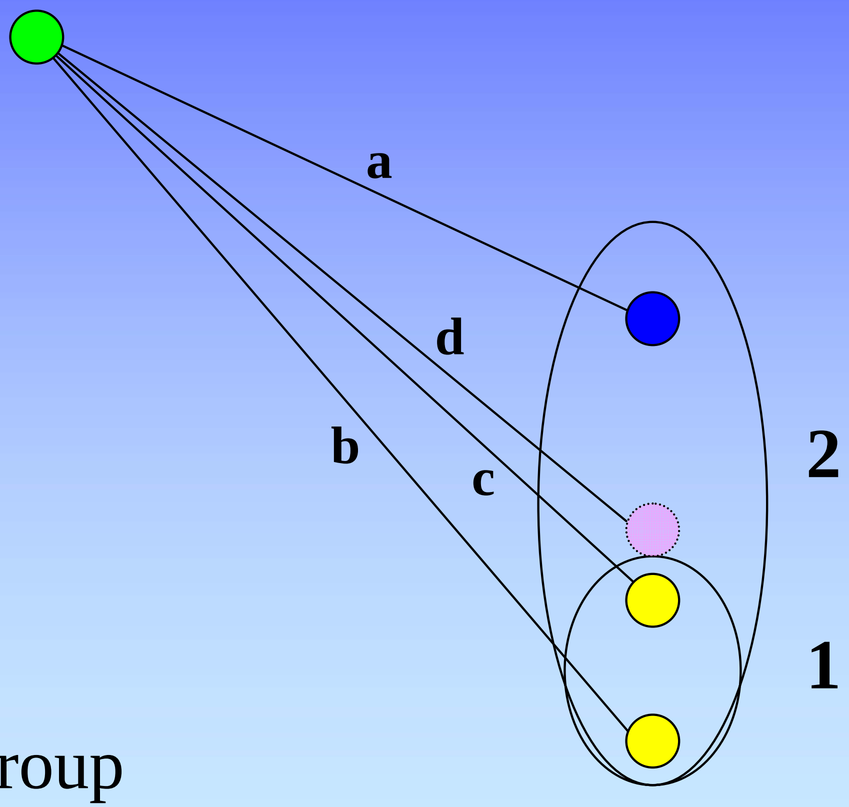
Clustering basics

- How to start:
  - One big cluster (divisive)
  - n clusters for n objects (agglomerative)
  - K clusters, where k is some pre-defined number
- Hierarchical agglomerative clustering
  - Popular method producing a tree showing relationships between objects (genes or chips)
  - Start by creating an all vs. all distance matrix
  - Fuse closest objects, then…

Representing groups of objects during clustering

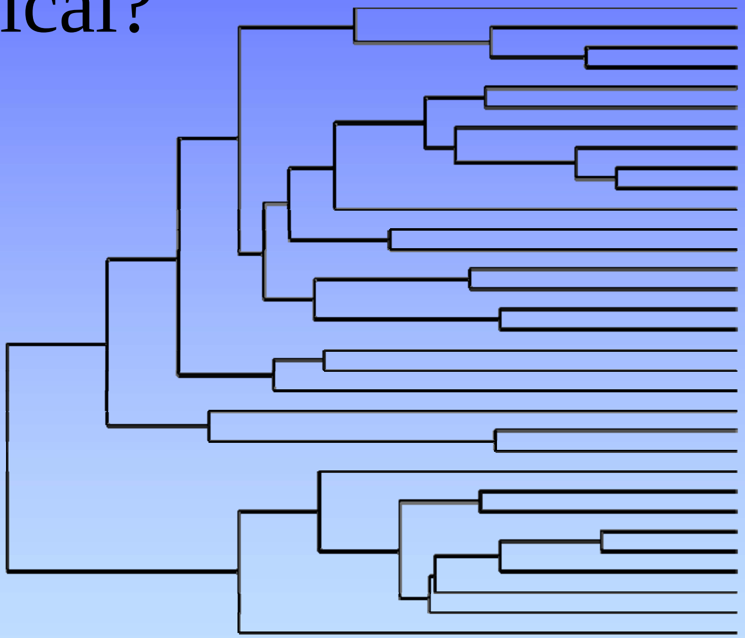
How is distance measured to a cluster of objects?

- Single linkage (a)
  - – minimum distance
- Complete linkage (b)
  - – maximum distance
- Average linkage (c)
  - – average distance
- Centroid linkage (d)
  - – distance to "centroid" of group



Representing clustered data

- Hierarchical clustering produces a dendrogram showing relationships between objects
- Are the data really hierarchical?
- Order of leaves 2^{N-2}
- How can objects be partitioned into groups?
  - k-means clustering
  - self-organizing maps
  - How many clusters (k)?
- Original distance matrix may be more informative



Summary

- Determining differential expression:
  - t-test, fold change, etc.
  - methods may be used in combination
- Correcting for multiple hypothesis testing
  - Bonferroni, False Discovery Rate, etc.
- Distance metrics: select carefully
- Clustering/segmentation types and methods
  - hierarchical, k-means, etc.; linkage types
  - Which protocol is best for your experiment?

References

- Dov Stekel. Microarray Bioinformatics. Cambridge, 2003.
- Speed, T. (ed.) Statistical Analysis of Microarray Data. Chapman & Hall, 2003
- Smyth GK et al. Statistical issues in cDNA microarray data analysis. Methods Mol Biol. 224:111-36, 2003.
- Pavlidis P. Using ANOVA for gene selection from microarray studies of the nervous system. Methods. 31(4):282-9, 2003.
- Quackenbush J. Computational analysis of microarray data. Nature Reviews Genetics 2:418-427, 2001.

Microarray tools

- Course page:
  - http://jura.wi.mit.edu/bio/education/bioinfo2007/arrays/
- BaRC analysis tools:
  - http://jura.wi.mit.edu/bioc/tools/
- Bioconductor (R statistics package)
  - http://www.bioconductor.org/
- Excel
- Many commercial and open source packages
- Cluster 3.0 and Java TreeView

Selecting a large matrix in Excel

1	Select the bottom right cell of the desired matrix	
2	Control - Shift - Up arrow	Select everything above the original cell
3	Control - Shift - Left arrow	Select everything to the left of the original cell
4	Shift - Down arrow	Move down one row
5	Shift - Right arrow	Move to the right one column

Exercise 2: Excel functions

- LOG
- IF
- TTEST
- CONCATENATE
- VLOOKUP
- MIN
- RANK

Exercise 2 - To do

- Use t-test to identify differentially expressed genes
- Use the "Absent/Present" calls from the Affymetrix algorithm to filter out genes with questionable expression levels
- List all the gene IDs for those that meet your significance threshold (such as $p < 0.05$) and are present in at least one sample.
- Gather expression data for these genes
- Cluster this selected data (multiple methods)
- Visualize clustered data as a heatmap