# Analysis of Microarray Data

Lecture 3:
Visualization and Functional Analysis

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

---

# Outline

- Review
- Visualizing all the data
- What to do with a set of interesting genes?
  - Basic annotation
  - Comparing lists
  - Genome mapping
  - Obtaining and analyzing promoters
  - Gene Ontology and pathway analysis
  - Other expression data

---

# Generic Microarray Pipeline

- Design experiment
- Prepare samples and perform hybridizations ⎫ not covered in course
- Quantify scanned slide image ⎭
- Calculate expression values
- Normalize
- Handle low-level expression values
- Merge data for replicates
- Determine differentially expressed genes
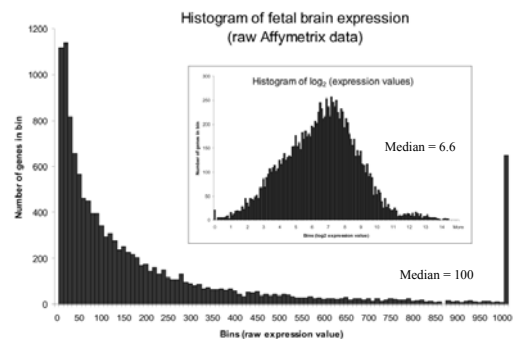- Cluster interesting data

---

# Review

- Preliminary filtering?
- Measuring differential expression:
- Correcting for multiple hypothesis testing
  - Fold change, t-test, ANOVA
  - Bonferroni, False Discovery Rate, etc.
- Filtering; identifying "interesting" genes
- Distance measures for clustering
- Clustering/segmentation types and methods
- What is the best analysis pipeline?
  - Why are you doing the experiment?
  - Are you being reasonable with the statistics?

---

# Why draw figures?

- Get a global perspective of the experiments
- Quality control: check for low-quality data and errors
- Compare raw and normalized data
- Compare controls: are they homogeneous?
- Help decide how to filter data
- Look at a subset of data in detail

---

# Intensity histogram



Histogram of fetal brain expression
(raw Affymetrix data)

Histogram of $\log_2$ (expression values)

Median = 6.6

Median = 100

## Intensity histogram

- Most genes have low expression levels
- Using $\log_2$ scale to transform data
  - more normal distribution
  - more helpful interpretation
- One way to observe overall intensity of chip
- How to choose genes with "no" expression?

## Intensity scatterplot



Expression in brain: fetal vs adult
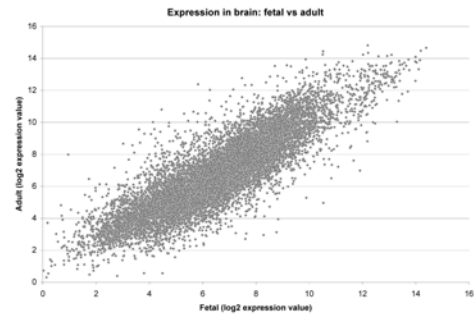
## Intensity scatterplot

- Compares intensity on two colors or chips
- Genes with similar expression are on the diagonal
- Use log-transformed expression values
- Genes with lower expression
  - noisier expression
  - harder to call significant

## R-I and M-A plots



R-I plot for brain: fetal vs adult

## R-I and M-A plots

- Compares intensity on two colors or chips
- Like an intensity scatterplot rotated 45º

$$R \text{ (ratio)} = \log(chip1 / chip2)$$
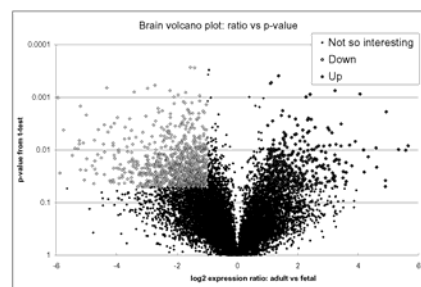$$I \text{ (intensity)} = \log(chip1 * chip2)$$

$$M = \log_2(chip1 / chip2)$$
$$A = \tfrac{1}{2}(\log_2(chip1*chip2))$$

- Popularized with lowess normalization
- Easier to intrepret than an intensity scatterplot
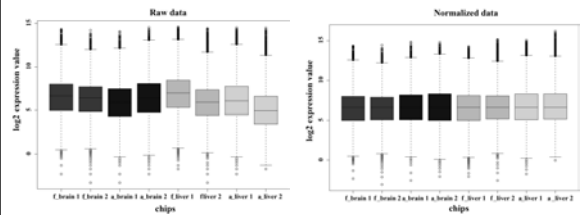
## Volcano plot



Brain volcano plot: ratio vs p-value

# Volcano plot

- Scatterplot showing differential expression statistics and fold change
- Visualize effects of filtering genes by both measures
- Using fold change vs. statistical measures for differential expression produce very different results

13

# Boxplots



Raw and median-normalized $\log_2$ (expression values)

14

# Boxplots

- Display summary statistics about the distribution of each chip:
  - Median
  - Quartiles (25% and 75% percentiles)
  - Extreme values (>3 quartiles from median)
  - Note that mean-normalized chips wouldn't have the same median
  - Easy in R; much harder to do in Excel

15

# Chip images
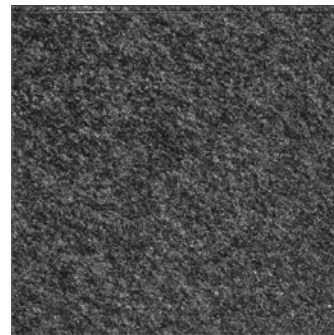


- Affymetrix U95A chip hybridized with fetal brain

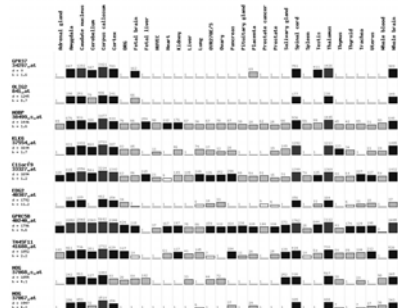- Image generated from .cel file

- Helpful for quality control

16

# Heatmaps
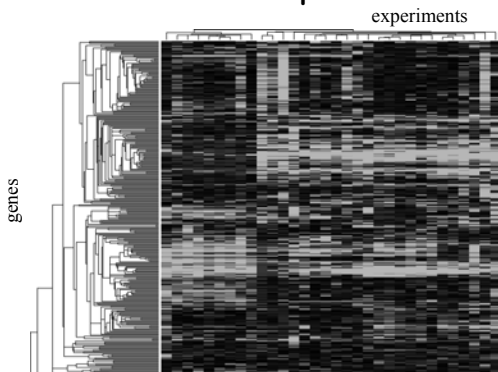


# Using distance measurements

Genes with most similar profiles to GPR37



18

# Functional Analysis: intro

- After data is normalized, compared, filtered, clustered, and differentially expressed genes are found, what happens next?
- Driven by experimental questions
- Specificity of hypothesis testing increases power of statistical tests
- One general question: what's special about the differentially expressed genes?

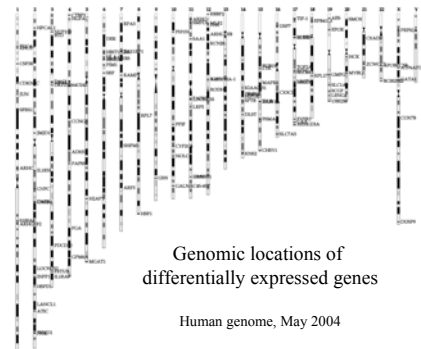# Annotation using sequence databases

- Gene data can be "translated" into IDs from a wide variety of sequence databases:
  - LocusLink, Ensembl, UniGene, RefSeq, genome databases
  - Each database in turn links to a lot of different types of data
  - Use Excel or programming tools to do this quickly
- Web links, instead of actual data, can also be used.
- What's the difference between these databases?
- How can all this data be integrated?

# Venn diagrams

- Show intersection(s) between at least 2 sets



Typical figure    More informative figure
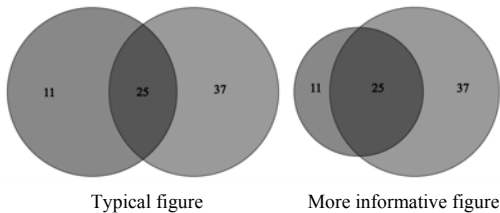
# Mapping genes to the genome



Genomic locations of differentially expressed genes

Human genome, May 2004

22

# Promoter extraction

- Prerequisite of any promoter analysis
- Requires a sequenced genome and a complete, mapped cDNA sequence
- "Promoter" is defined in this context as upstream regulatory sequence
- Extract genomic DNA using a genome browser: UCSC, Ensembl, NCBI, GBrowse, etc.
- Functional promoter needs to be determined experimentally

# Promoter analysis

- TRANSFAC contains curated binding data
- Transcription factor binding sites can be predicted
  - matrix (probabilities of each nt at each site)
  - pattern (fuzzy consensus of binding site)
- Functional sites tend to be evolutionarily conserved
- Functional promoter activity needs to be verified experimentally

## Gene Ontology

- GO is a systematic way to describe protein (gene) function
- GO comprises ontologies and annotations
- The ontologies:
  - Molecular function
  - Biological process
  - Cellular component

  Gene_Ontology ( GO:0003673 )
  molecular_function ( GO:0003674 )
  binding ( GO:0005488 )
  nucleic acid binding ( GO:0003676 )
  DNA binding ( GO:0003677 )
  transcription factor activity ( GO:0003700 )
  RNA polymerase II transcription factor activity, enhancer binding ( GO:0003705 )
  transcription regulator activity ( GO:0030528 )
  transcription factor activity ( GO:0003700 )
  RNA polymerase II transcription factor activity, enhancer binding ( GO:0003705 )

- Ontologies are like hierarchies except that a "child" can have more than one "parent".
- Annotation sources: publications (TAS), bioinformatics (IEA), genetics (IGI), assays (IDA), phenotypes (IMP), etc.

---

## Gene Ontology enrichment analysis

- Unbiased method to ask question, "What's so special about my set of genes?"
- Obtain GO annotation (most specific term(s)) for genes in your set
- Climb an ontology to get all "parents" (more general, "induced" terms)
- Look at occurrence of each term in your set compared to terms in population (all genes or all genes on your chip)
- Are some terms over-represented?
  Ex: sample:10/100  pop1: 600/6000  pop2: 15/6000

---

## Pathway enrichment analysis

- Unbiased method to ask question, "Is my set of genes especially involved in specific pathways?"
- First step: Link genes to pathways
- Are some pathways over-represented?
- Caveats
  - What is meant by "pathway"?
  - Multiple DBs with varied annotations
  - Annotations are very incomplete

---

## Enrichment analysis on sorted expression data

- Input 1: complete sorted gene list
  - no threshold value or definition of significance
- Input 2: set of biologically meaningful gene sets
  - pathway, genome location, function, ...
- Is the rank of genes from any gene set in your sorted list non-random?
- Example: GSEA



Broad Institute

---
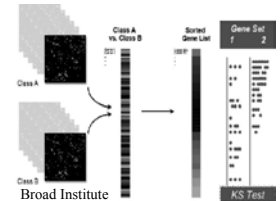
## Comparisons with other expression studies

- Array repositories: GEO (NCBI), ArrayExpress (EBI), WADE (WIBR)
- Search for genes, chips, types of experiments, species
- View or download data
- Normalize but still expect noise
  - Check medians and distribution of data
- It's much easier to make comparisons within an experiment than between experiments

---

## Summary

- Plots: histogram, scatter, R-I, volcano, box
- Other visualizations: whole chip, heatmaps, bar graphs, Venn diagrams
- Annotation to sequence DBs
- Genome mapping
- Promoter extraction and analysis
- GO and pathway enrichment analysis
- Comparison with published studies

# More information

- Course page:
  - http://jura.wi.mit.edu/bio/education/bioinfo2006/arrays/
- Bioconductor short courses: http://www.bioconductor.org/
- BaRC analysis tools:
  - http://jura.wi.mit.edu/bioc/tools/
- Gene Ontology Consortium website:
  - http://www.geneontology.org/
- Causton HC et al. <u>Microarray Gene Expression Data Analysis: A Beginner's Guide</u>. Blackwell, 2003.
- Parmigiana G et al. <u>The Analysis of Gene Expression Data: Methods and Software</u>. Springer, 2003.

# Exercises

- Graphing all data
  - Scatterplot
  - R-I (M-A) plot
  - Volcano plot
- Functional analysis
  - Annotation
  - Comparisons
  - Genome mapping
  - Promoter extraction and analysis
  - GO and pathway analysis
  - Using other expression studies