

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
- Quantify scanned slide image
- Calculate expression values
- Normalize
- Handle low-level expression values
- Merge data for replicates
- Determine differentially expressed genes
- Cluster interesting data

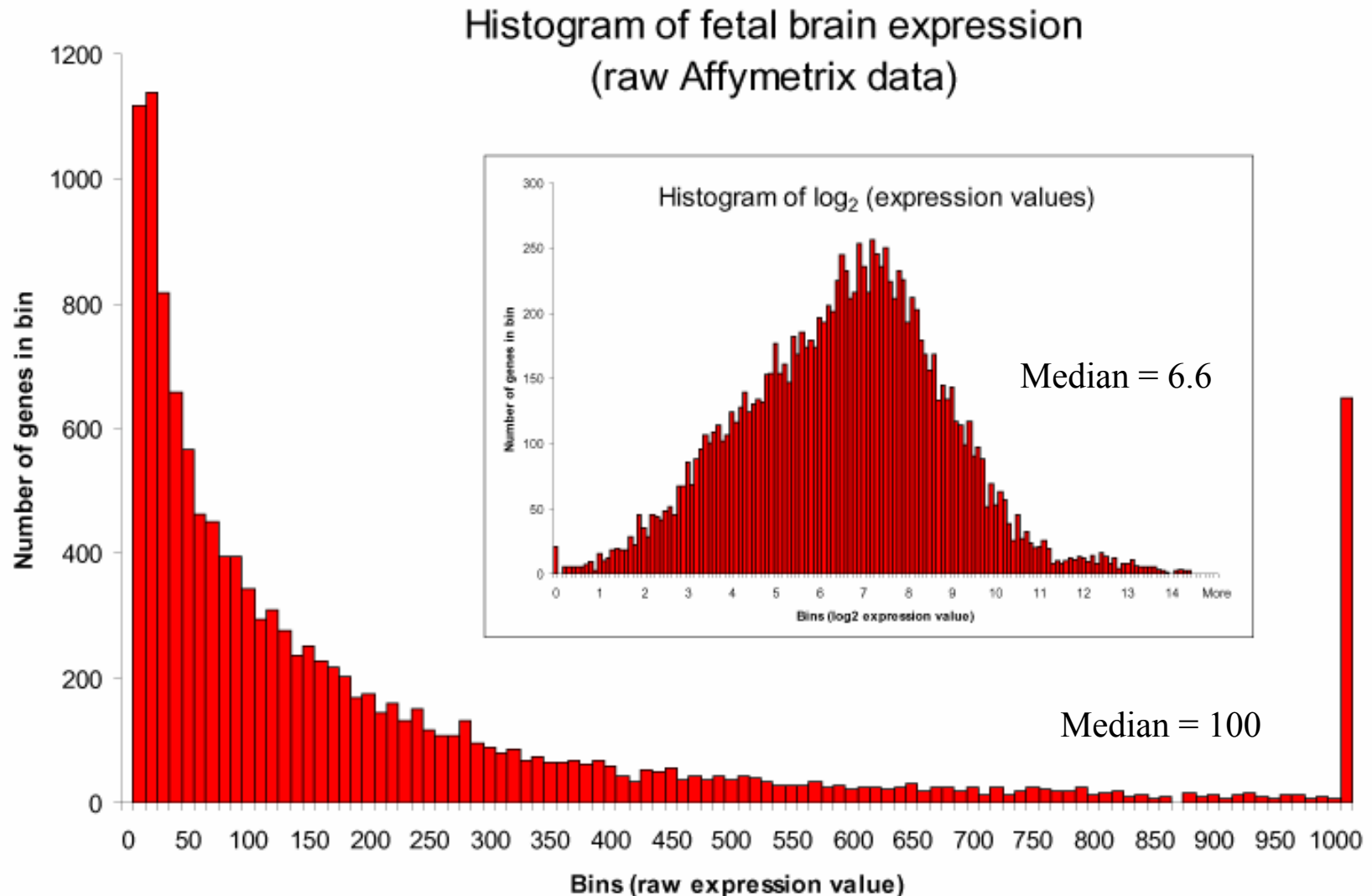
Review

- 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

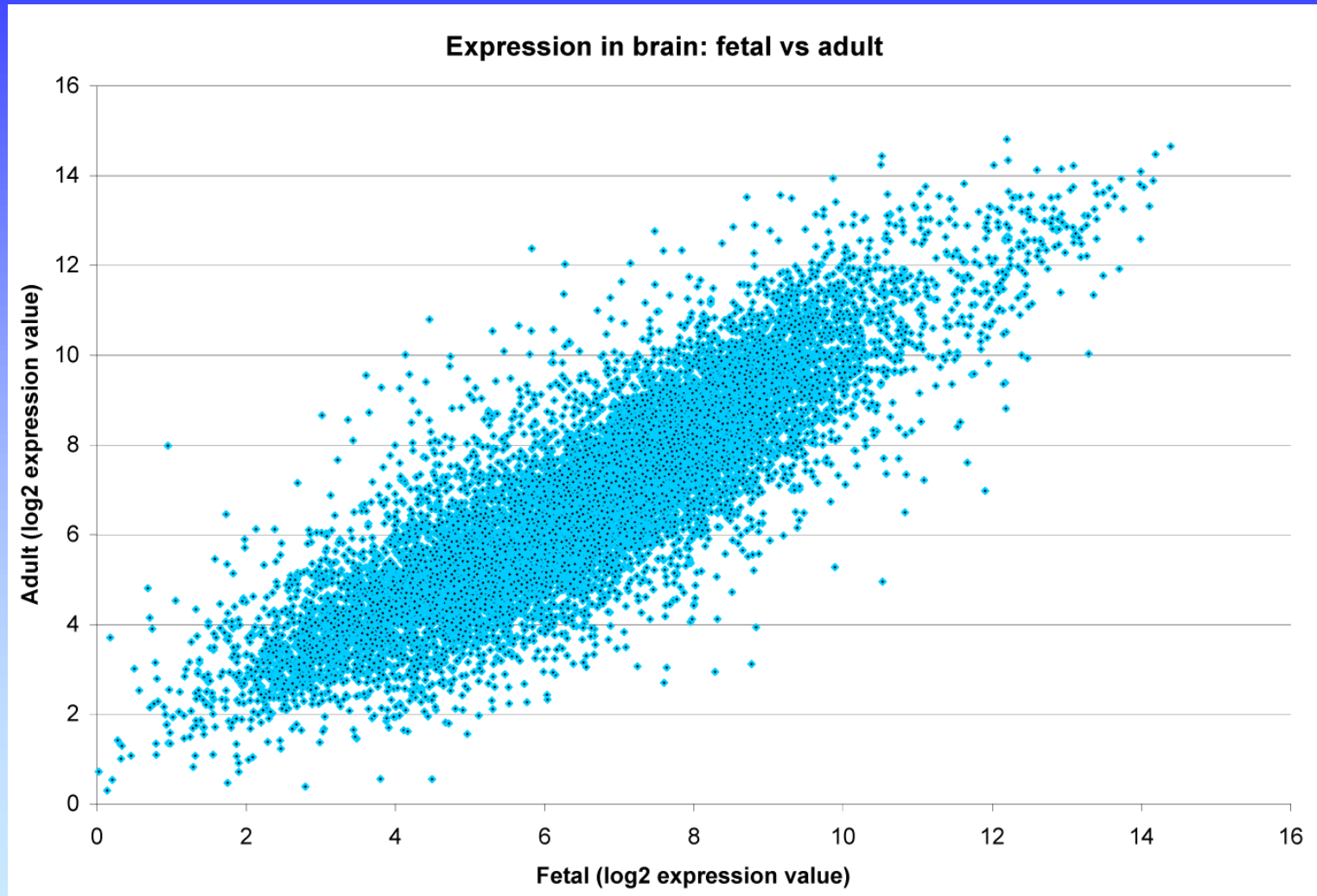
Intensity histogram



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



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 and M-A plots

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

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

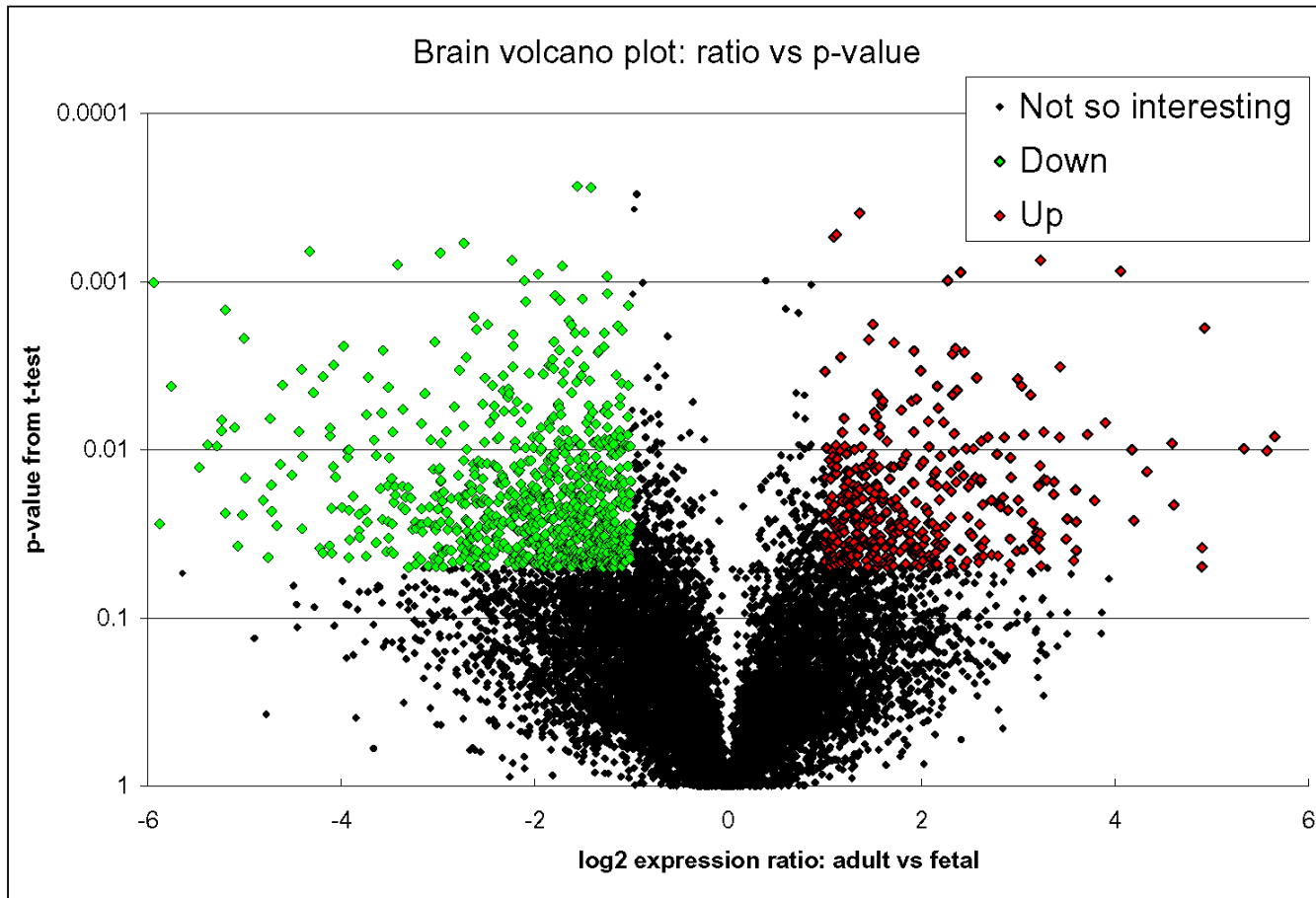
$$I \text{ (intensity)} = \log(\text{chip1} * \text{chip2})$$

$$M = \log_2(\text{chip1} / \text{chip2})$$

$$A = \frac{1}{2}(\log_2(\text{chip1} * \text{chip2}))$$

- Popularized with lowess normalization
- Easier to interpret than an intensity scatterplot

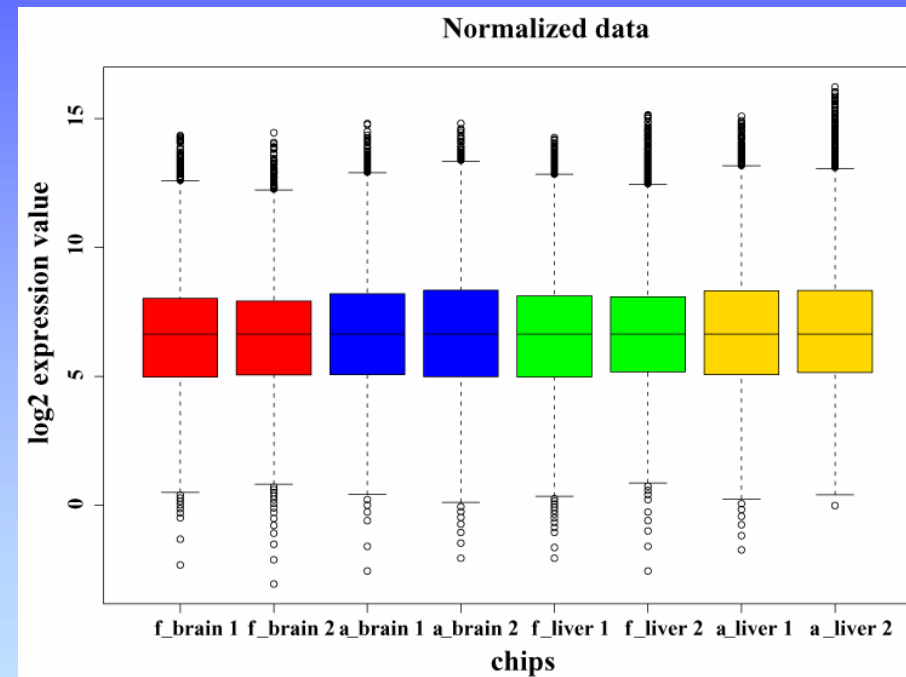
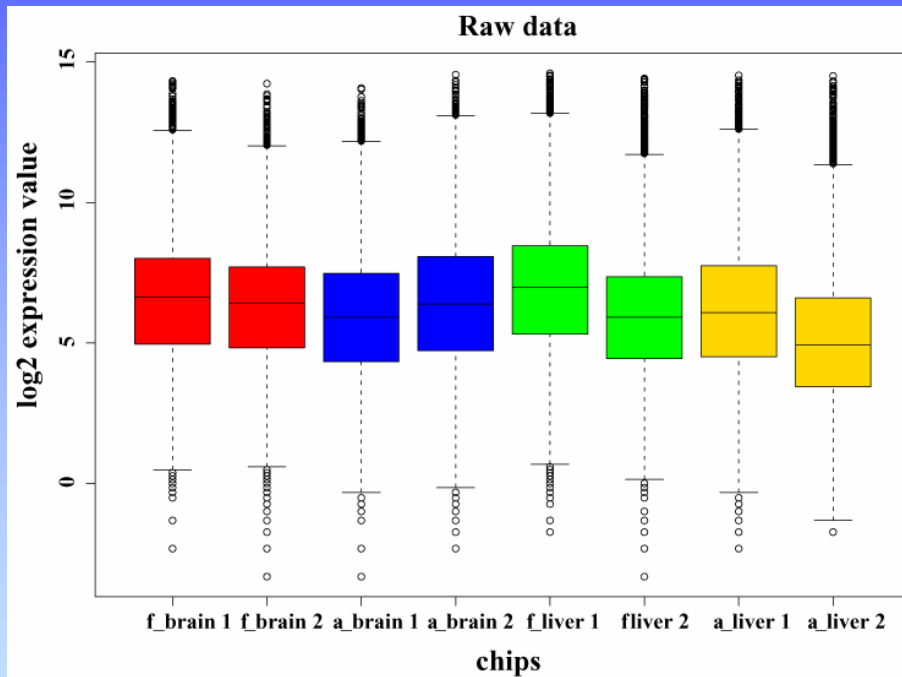
Volcano plot



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

Boxplots

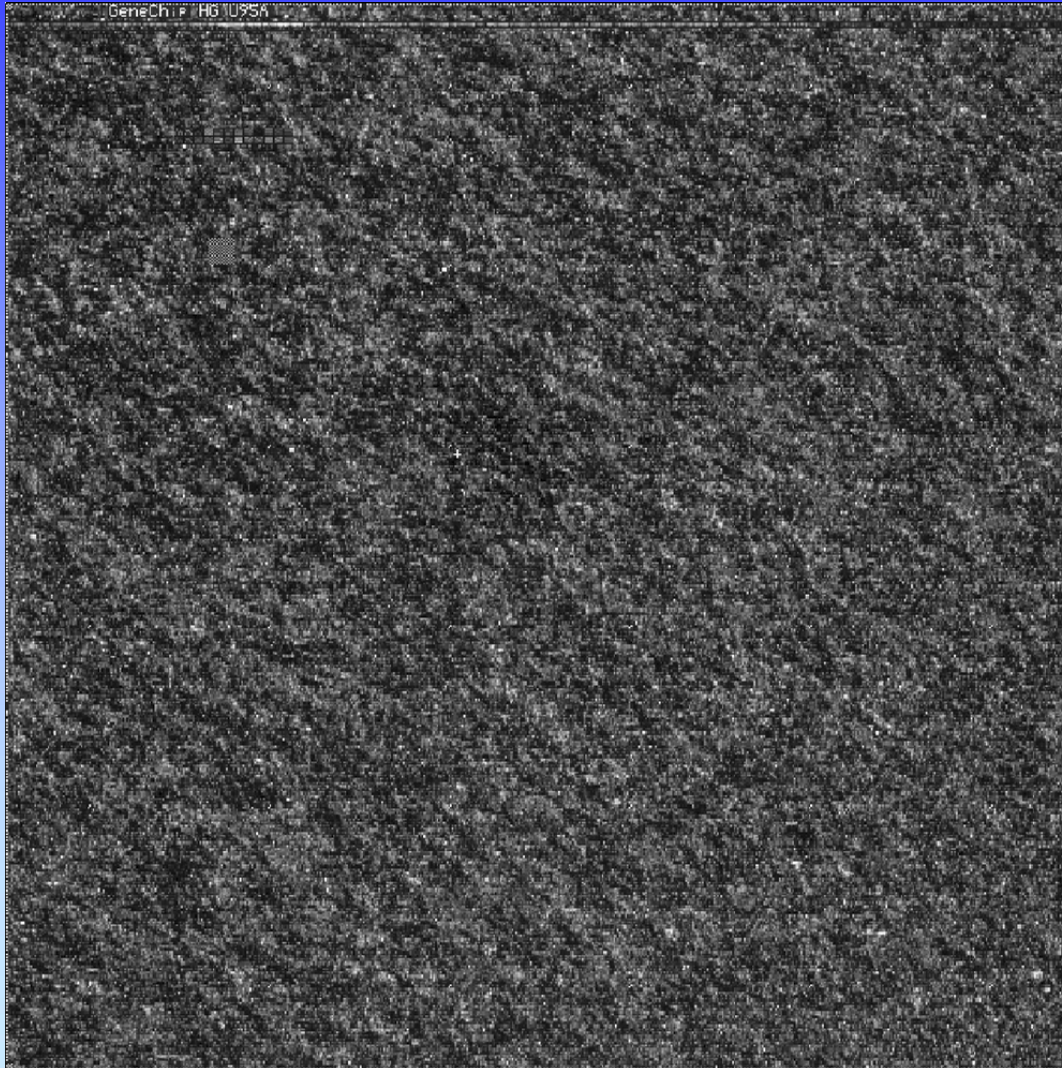


Raw and median-normalized \log_2 (expression values)

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

Chip images

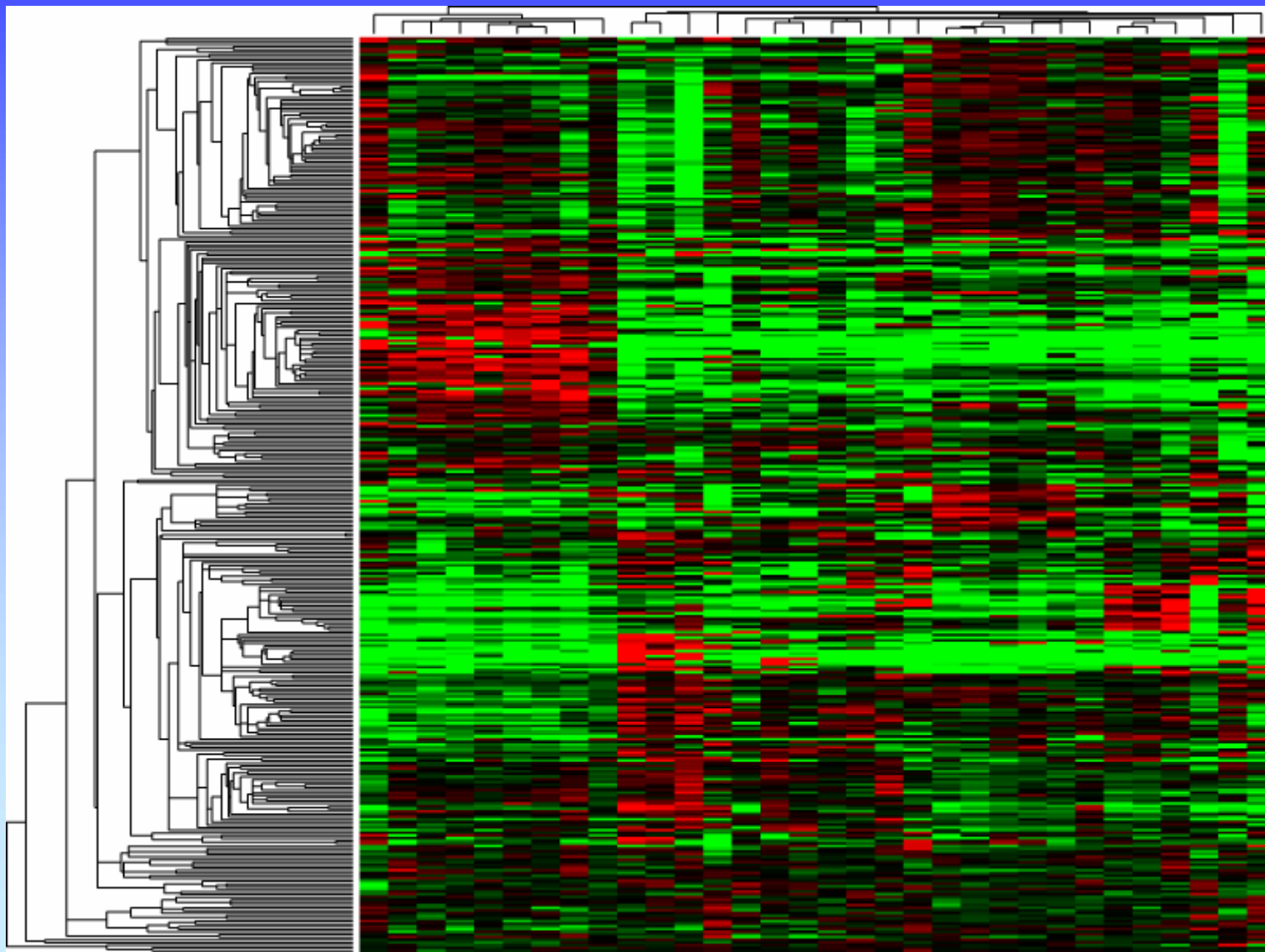


- Affymetrix U95A chip hybridized with fetal brain
- Image generated from .cel file
- Helpful for quality control

Heatmaps

experiments

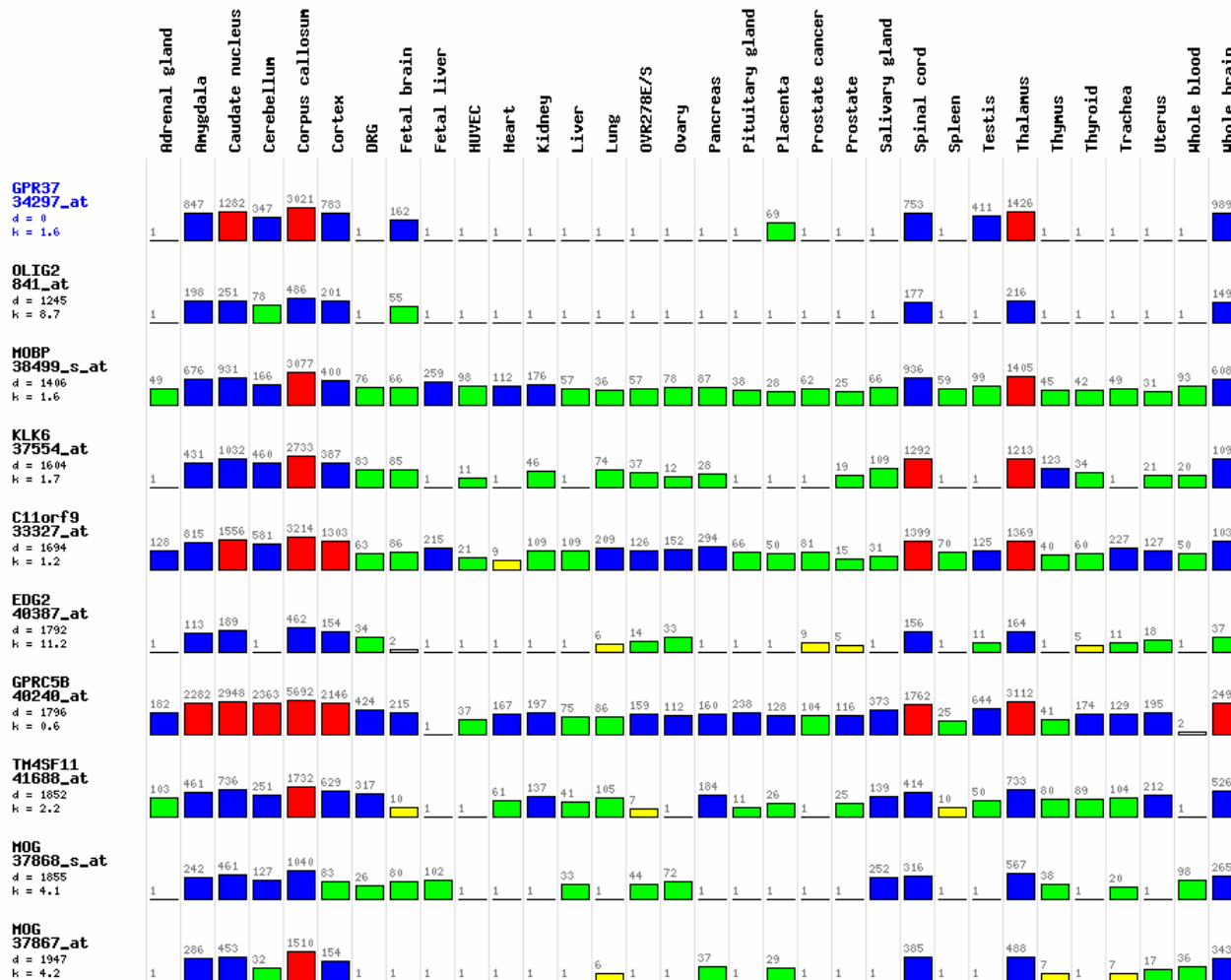
genes



Using distance measurements

Genes with most similar profiles to GPR37

Expression profiles in Novartis human set (U95) closest to 34297_at



Functional Analysis: intro

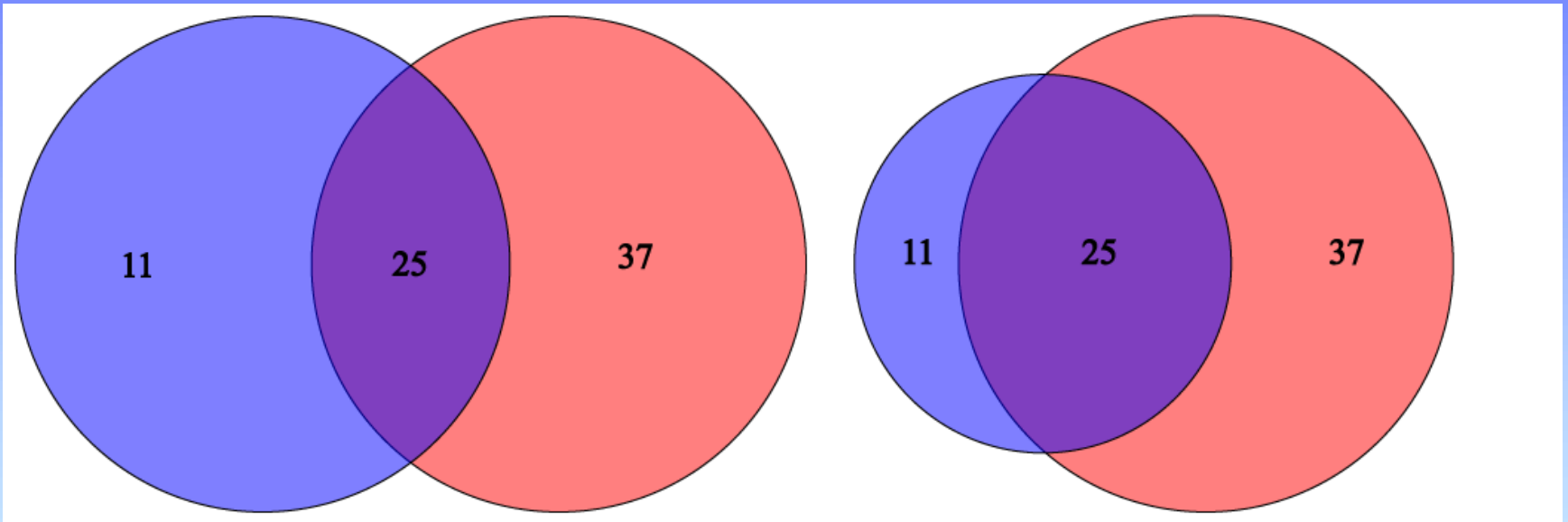
- After data is normalized, compared, filtered, clustered, and differentially expressed genes are found, what happens next?
- Driven by experimental questions
- Specificity of hypothesis testing increases power of statistical tests
- One general question: what's special about the differentially expressed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

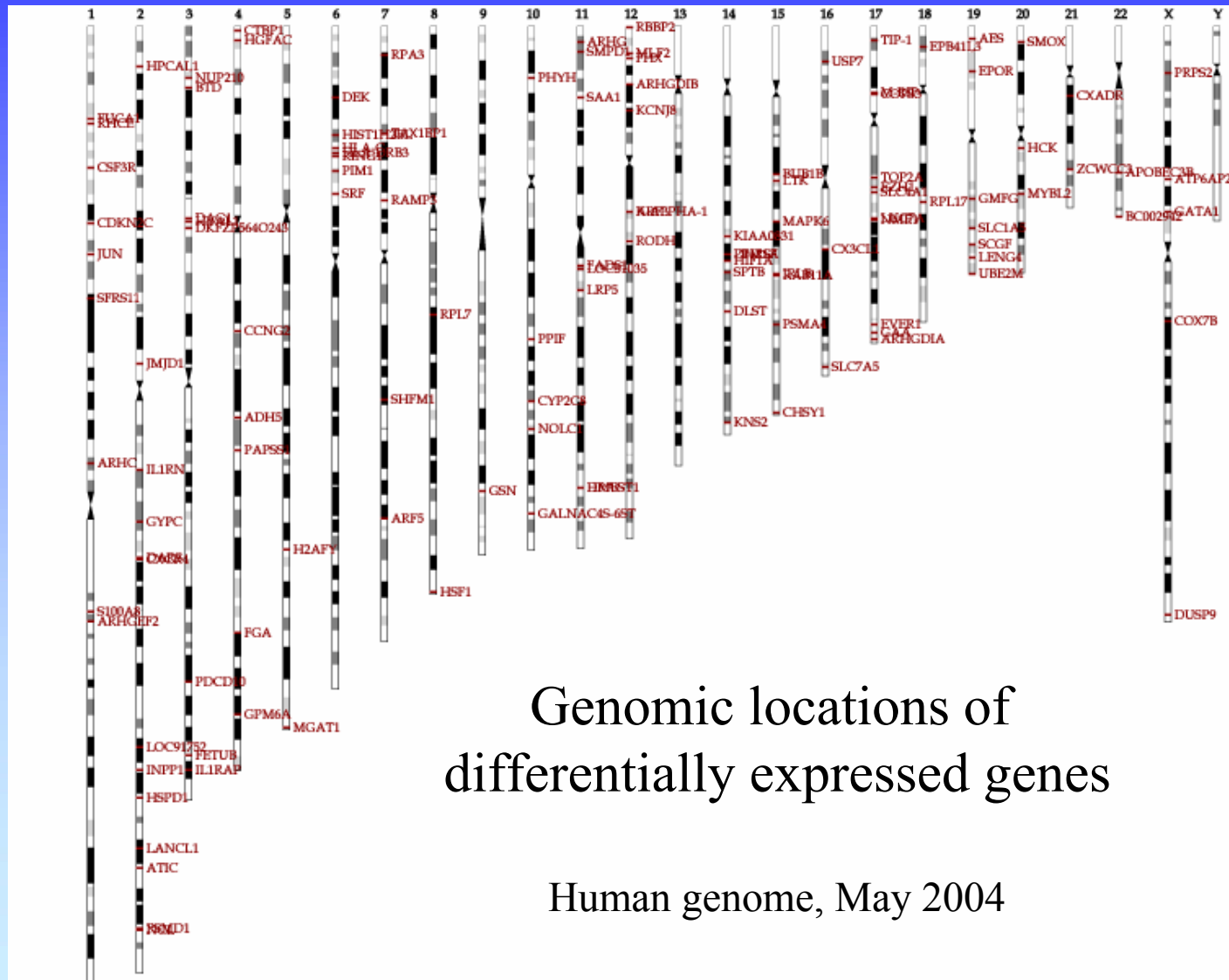
- Show intersection(s) between at least 2 sets



Typical figure

More informative figure

Mapping genes to the genome



Genomic locations of
differentially expressed genes

Human genome, May 2004

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
- 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 [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]

Gene Ontology enrichment analysis

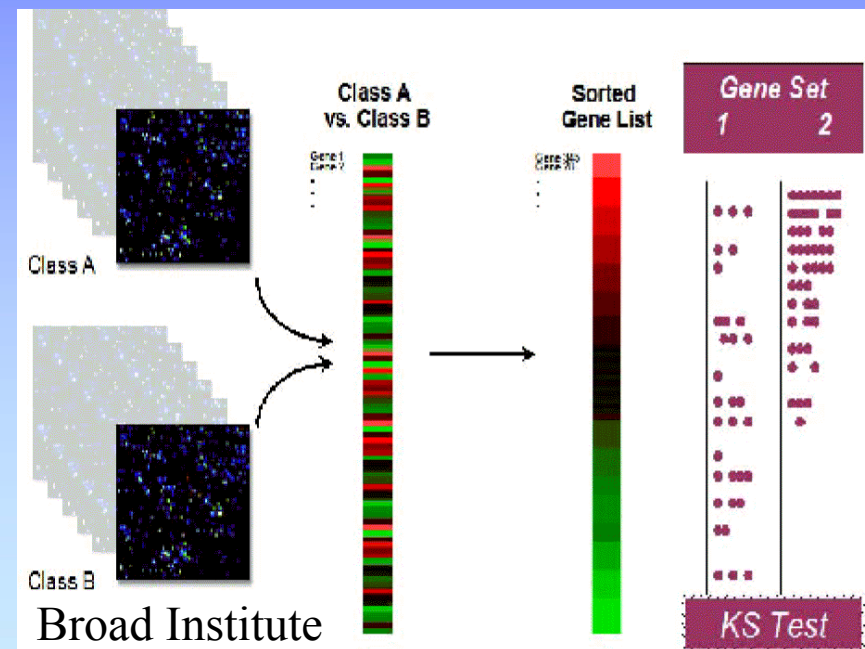
- Unbiased method to ask question, “What’s so special about my set of genes?”
 - Obtain GO annotation (most specific term(s)) for genes in your set
 - Climb an ontology to get all “parents” (more general, “induced” terms)
 - Look at occurrence of each term in your set compared to terms in population (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



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/bioinfo2005/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. Microarray Gene Expression Data Analysis: A Beginner's Guide. Blackwell, 2003.
- Parmigiana G et al. The Analysis of Gene Expression Data: Methods and Software. 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