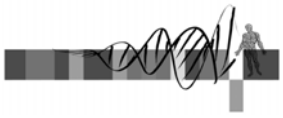


Microarray Analysis

Visualization and Functional Analysis

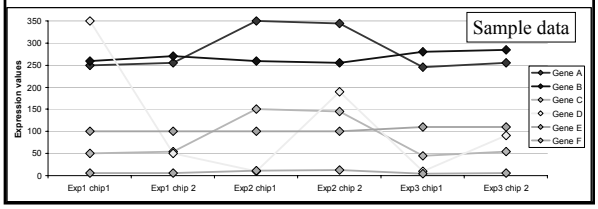
George Bell, Ph.D.
Bioinformatics Scientist
Bioinformatics and Research Computing
Whitehead Institute

Microarray pipeline so far

- 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

Some issues to consider - review

- Quality control – lab work and analysis
- The "best" analysis pipeline
- Filtering; identifying "interesting" genes
- Distance measures for clustering



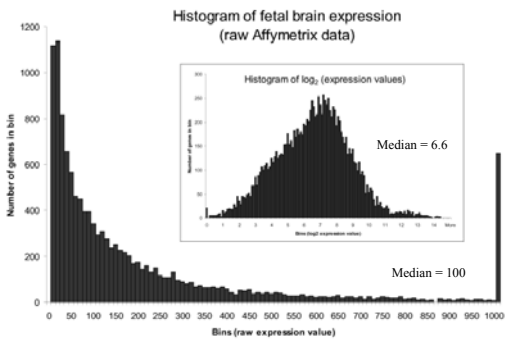
Outline

- 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

Why graphs?

- 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

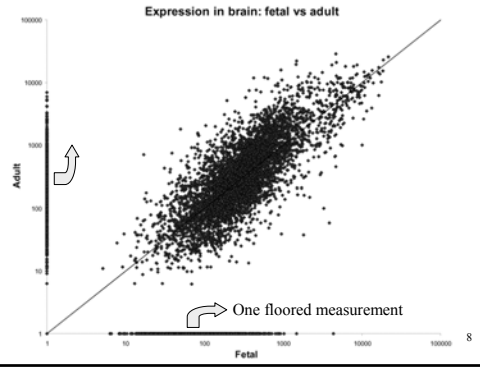
Intensity histogram



Intensity histogram

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

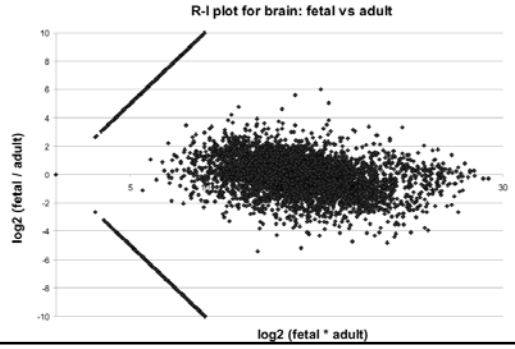
Intensity scatterplot



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

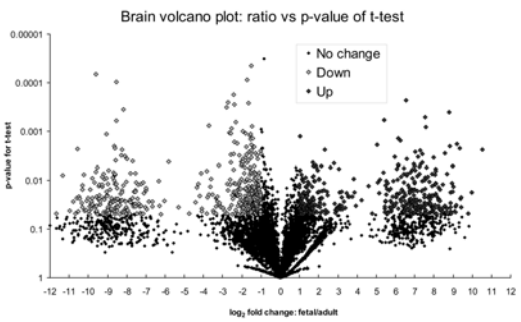
- Compares intensity on two colors or chips
- Like an intensity scatterplot rotated 45º

R (ratio) = log(chip1 / chip2)
I (intensity) = log(chip1 * chip2)

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

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

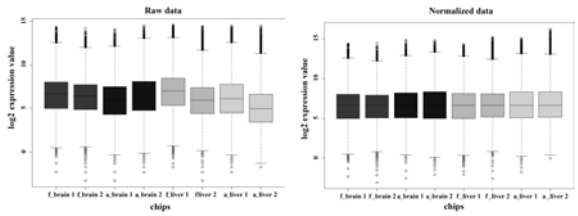
Volcano plot



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)

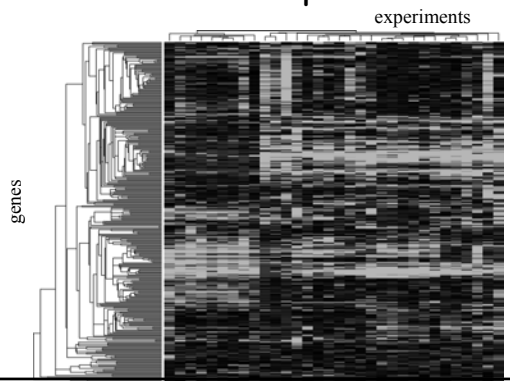
Boxplots

- 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

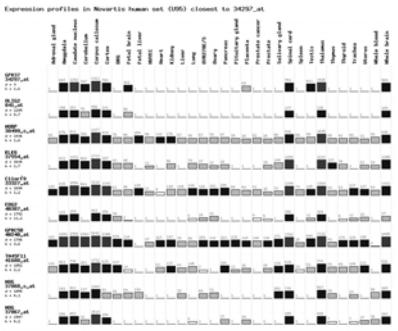
- Affymetrix U95A chip hybridized with fetal brain
- Image generated from .cel file
- Helpful for quality control

Heatmaps



Using distance measurements

Genes with most similar profiles to GPR37



Functional Analysis: intro

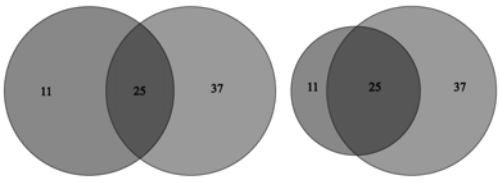
- After data is normalized, compared, filtered, clustered, and differentially expressed genes are found, what happens next?
- Driven by experimental questions
- Specificity of hypothesis testing increases power of statistical tests
- One general question: what's special about the differentially expressed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

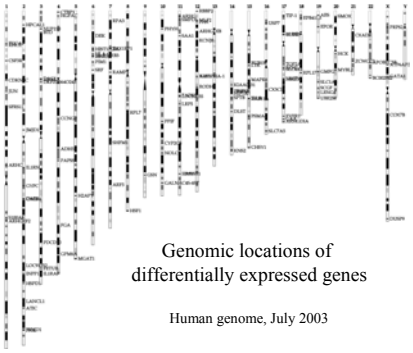
Venn diagrams

- Show intersection(s) between at least 2 sets



Typical figure Informative figure

Mapping genes to the genome



Genomic locations of differentially expressed genes

Human genome, July 2003

Promoter extraction

- 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 gene and protein 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 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 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 analysis

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

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
- 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 analysis
- Comparison with published studies

Tools for array analysis

- Excel; OpenOffice
- R / Bioconductor
- Matlab
- JMP
- GCOS (Affymetrix)
- GeneSpring
- GenePattern; GeneCluster
- Lots more on the web and for download

More information

- Bioconductor short courses: http://www.bioconductor.org/
- BaRC analysis tools: http://iona.wi.mit.edu/bio/tools/bioc_tools.html
- Causton et al., 2003. *Microarray Gene Expression Data Analysis*.
- Gene Ontology Consortium
- *Nature Genetics* (Dec 2002) The Chipping Forecast II (supplement)

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