Analyzing Gene Expression Data using Categories (work in progress)

Robert Gentleman

FRED HUTCHINSON
CANCER RESEARCH CENTER
A LIFE OF SCIENCE

Outline

- Description of the experimental setting
- A brief description of differential gene selection
- Categories and how to use them
- Related ideas
- Example: ALL data set from the Ritz Lab
- Concluding Remarks

Experiments/Data

- There are n samples
- for each sample we measure mRNA expression levels on G genes
- we consider the case where there are two phenotypes (e.g. BCR/ABL vs NEG)
- A t-test can be computed, for each gene comparing the two samples (other test statistics can be handled easily)

Differential Expression

- Usual approach is to try and find the set of differentially expressed genes [those with extreme values of the univariate statistic, **x**]
- Often adjusting in some way for multiple comparisons
- This can be criticized on many grounds
  - it introduces an aritical distinction - differentially expresssed
  - it focuses attention on only a few genes that change alot

Differential Expression

- *p*-value correction methods don't really do what we want
- to see if too many genes of a particular type have been seleceted a Hypergeometric calculation is made, but it relies on the articial distinction between expressed and not expressed
- we (and others) propose a different approach: find sets of genes whose expression changes in concert, possibly not by a large amount

Holistic Approach

- we will attempt to find categories of genes where there are potentially small but coordinated changes in gene expression
- an obvious situation is one where genes in a category all show small but consistent change in a particular direction

Related Work

- PGC-1 alpha-responsive genes involed in oxidative phosphorylation are coordinately downregulated in human diabetes. Mootha et al, Nature Genetics, 2003
- mTOR inhibition reverses Akt-dependent prostate intraepithelial neoplasia through regulation of apoptotic and HIF-1 dependent pathways, Majumder et al, Nature Medicine, 2004
- Discovering statistically significant pathways in expression profiling studies. Tian et al, PNAS, 2005,

Gene Set Enrichment

- proposed by Mootha et al (2003)
- very similar (and was one of the motivations) but more complex and computationally expensive
- they discuss gene sets, S, which are the same as categories

Gene Set Enrichment

- For each gene set S, a Kolmogorov-Smirnov running sum is computed
- The assayed genes are ordered according to some criterion (say a two sample *t*-test; or signal-to-noise ratio SNR).
- Beginning with the top ranking gene the running sum increases when a gene in set S is encountered and decreases otherwise
- The enrichment score (ES) for a set S is defined to be the largest value of the running sum.

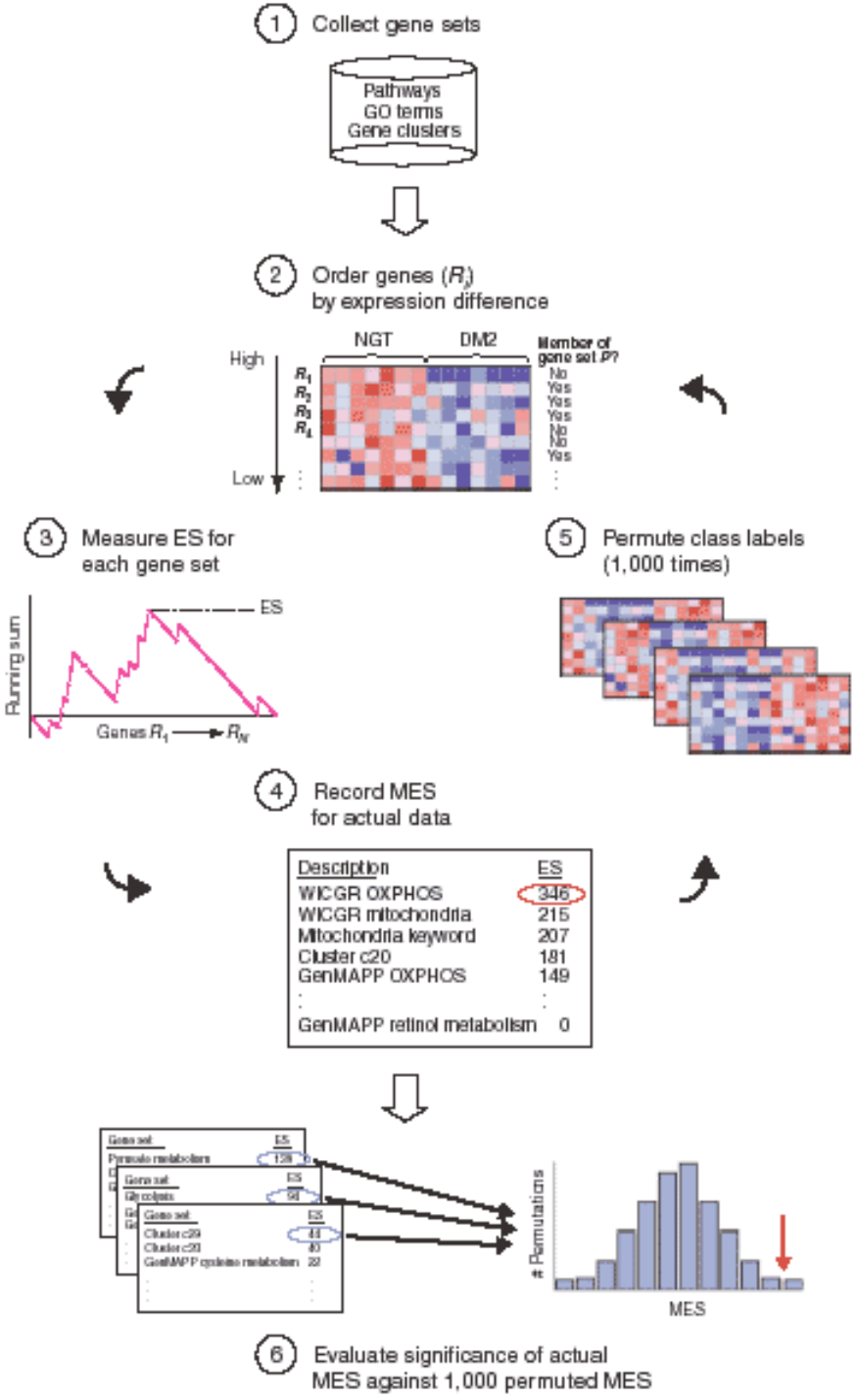
Gene Set Enrichment

- The maximal ES (MES), over all sets S under consideration is recorded.
- For each of B permutations of the class label, ES and MES values are computed.
- The observed MES is then compared to the B values of MES that have been computed, via permutation.
- This is a single *p*-value for all tests and hence needs no correction (on the other hand you are testing only one thing).

From Mootha *et al*

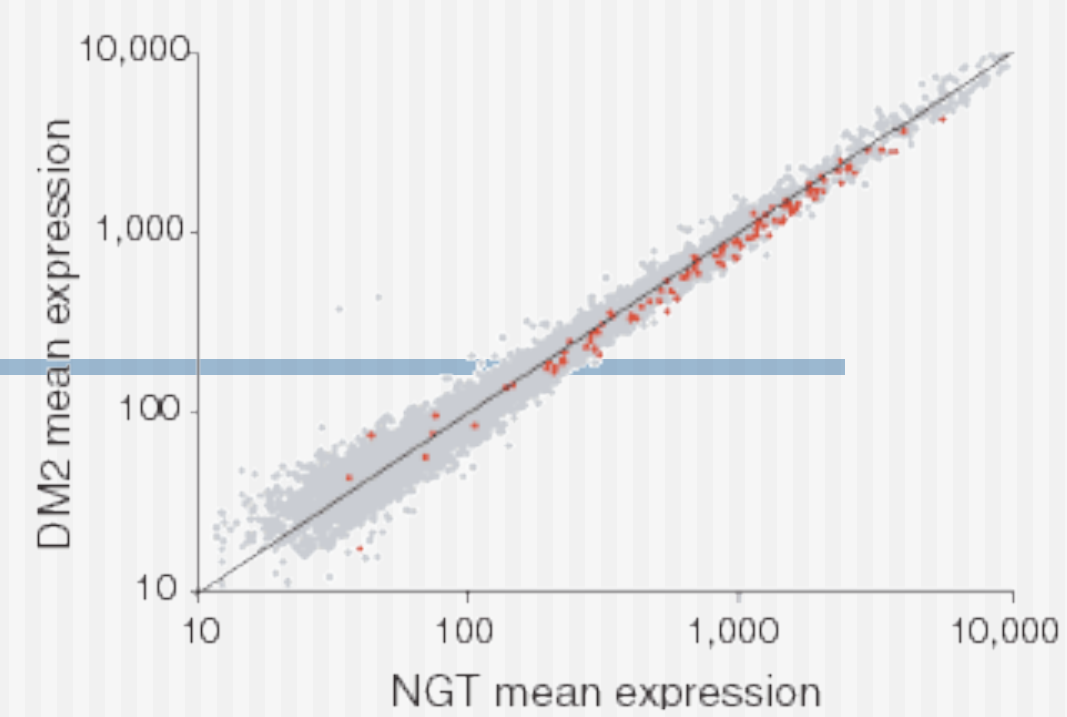
ES=enrichment score
for each gene
= scaled K-S dist

A set called OXPHOS got the largest ES score, with *p=0.029* on *1,000* permutations.

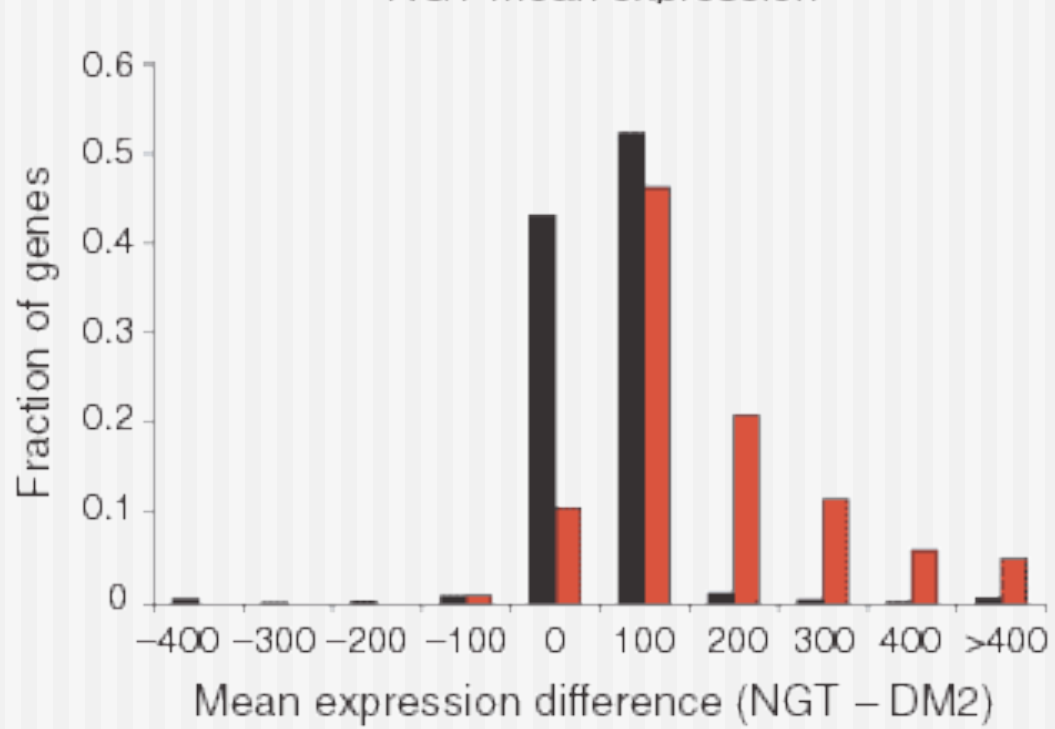


OXPHOS

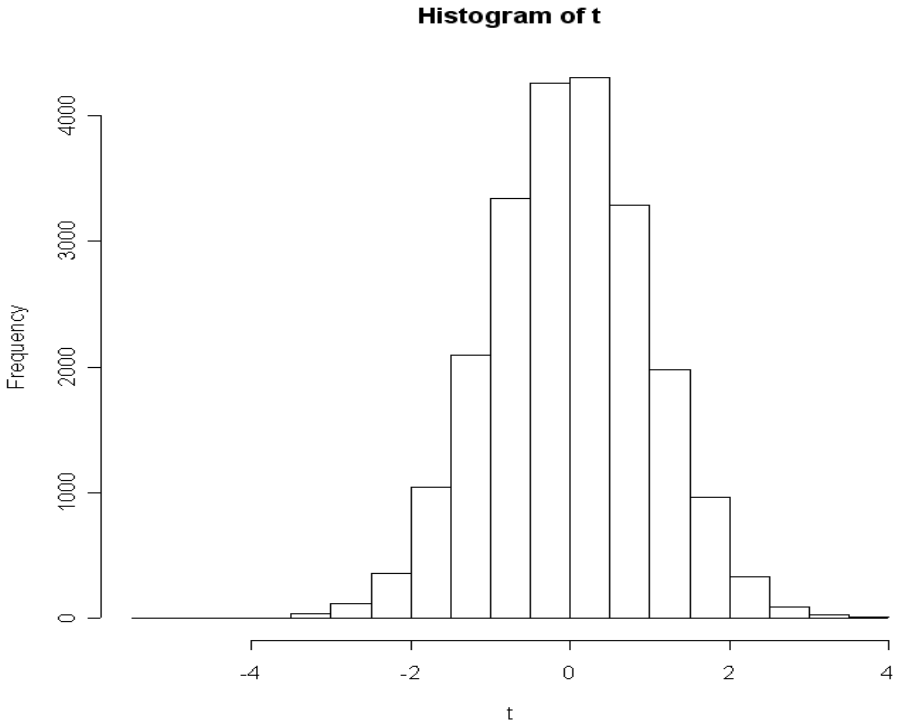
(A small difference for many genes)

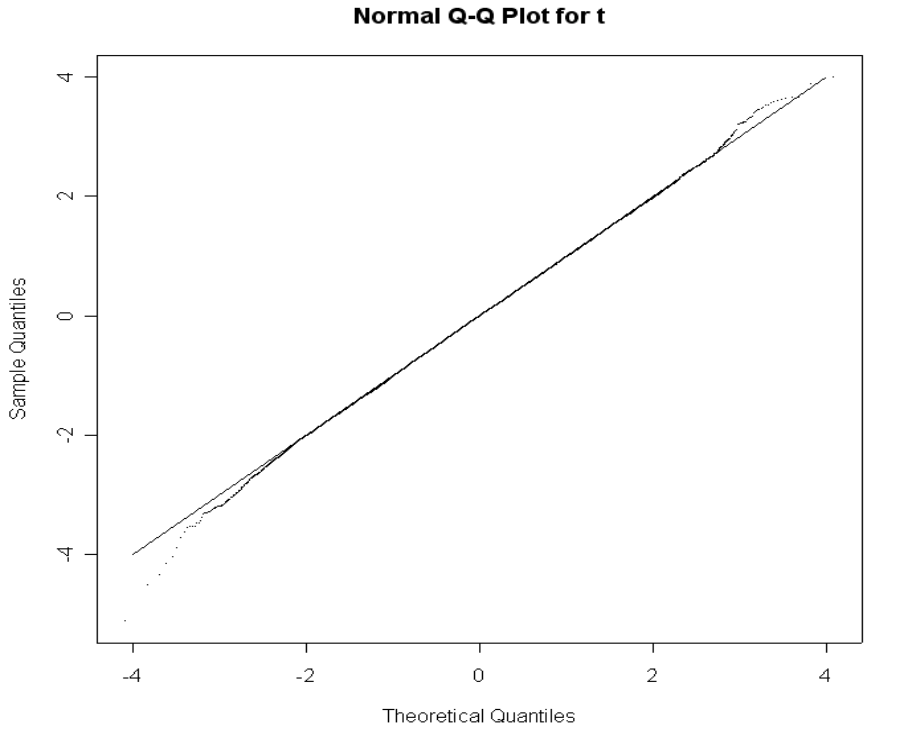


All genes
OXPHOS

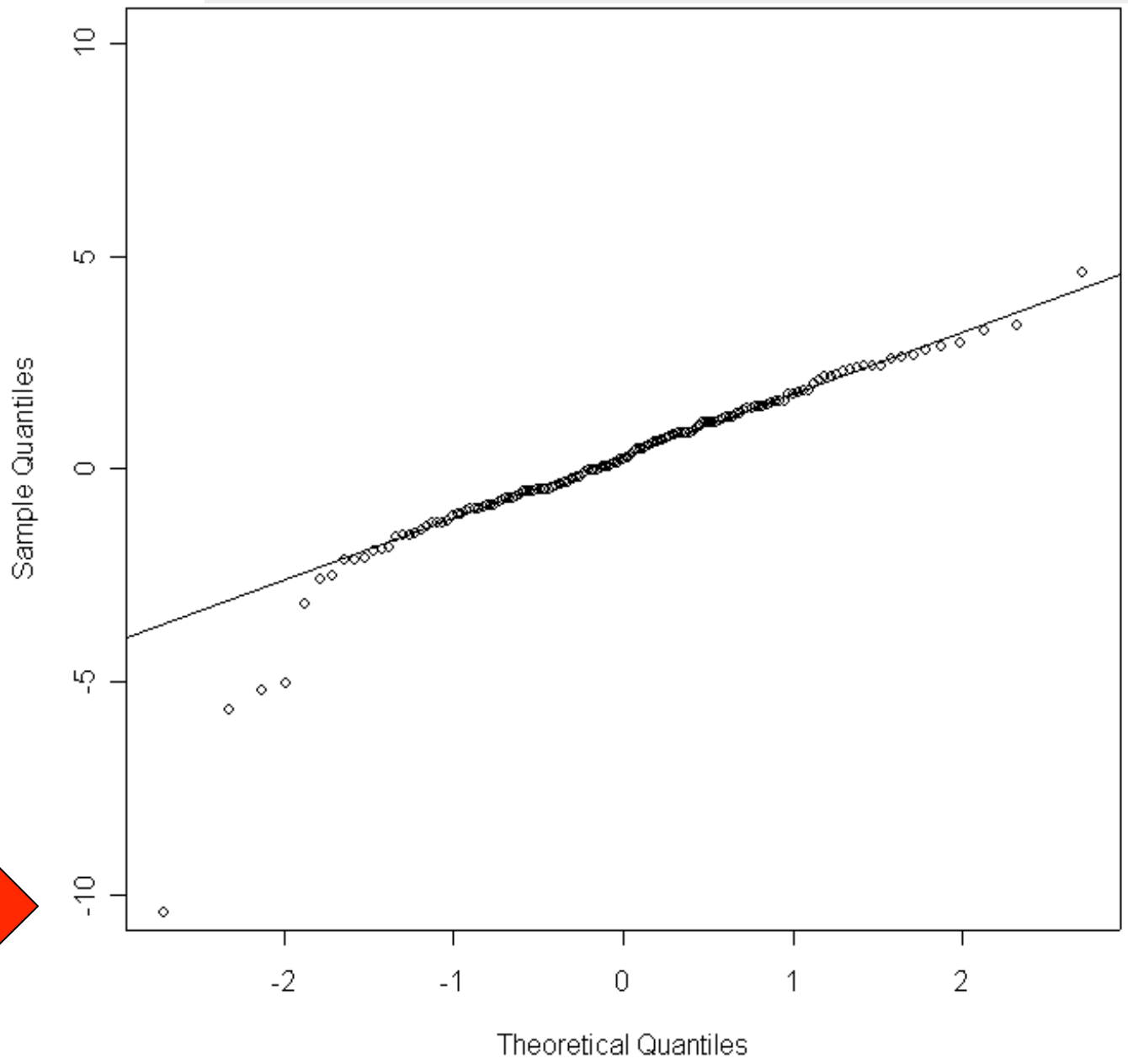


Mootha's ts are approx normal





Normal qq-plot of $\sqrt{n} \times \bar{t}$



OXPHOS

Selection of Categories

- pathways (KEGG, cMAP, BioCarta)
- molecular function, biological process cellular location (GO)
- predefined sets from the published literature etc
- regions of synteny; cytochrome bands
- some care should be exercised to select categories that are of interest *a priori*
  - there are more categories than genes so you will simply end up back in the multiple comparison problem

Categories

- a set of categories is merely a grouping of genes (entities)
- the groups do not need to be exhaustive or disjoint
- we do not need to be completely right, we can have some genes that are not in the category, and we can miss some, but not too many
- we are relying on averaging to help adjust for mistakes
- given the state of genomic knowledge this seems preferable

Categories

- the elements of **A**, **A**[*i,j*]=1 if gene *j* is in category *i*
- the row sums represent the number of genes in each category
- the column sums represent the number of categories a gene is in
- if two rows are identical (for a given set of genes) then the two categories are aliased (in the usual statistical sense)
- other patterns can can cause problems and need some study

Categories

- the simplest transformation is to simply sum up the *t*-statistics for all genes in each category
- we divide the sum by the square root of the number of genes per category (this is right if genes are independent - an unrealistic assumption)
- we could take the median, or use a sign-test within categories
- then the resultant statisics, under the null hypothesis, have approximately a N(0,1) distribution
- we can plot them and look for big/small values

Categories: Reference Distribution

- an alternative is to generate many *t*-tests from a reference distribution
- one distribution of interest is to go back to the original expression data and either permute the sample labels or bootstrap to provide a reference distribution
- you should not (as Tian et al do) permute the gene labels [what is your null hypothesis?]

Comparisons

- you can do either within category comparisons
  - for a given category is the observed test statistic unusual
- or overall comparisons
  - are any of the observed category statistics unusually large with respect to the entire reference distribution
- the former requires some consideration of multiple testing issues
- note that the approach is inherently multivariate, one data set gives G test statistics (one per gene) and these are transformed to yield one per category

Bayesian Approach

- following Newton et al, we could compute the posterior probability that a gene is differentially expressed
- then **x**, our G vector is a set of probabilities
- **z** = **Ax**, is then a C vector of the expected number of differentially expressed genes in each category

Bayesian Approach

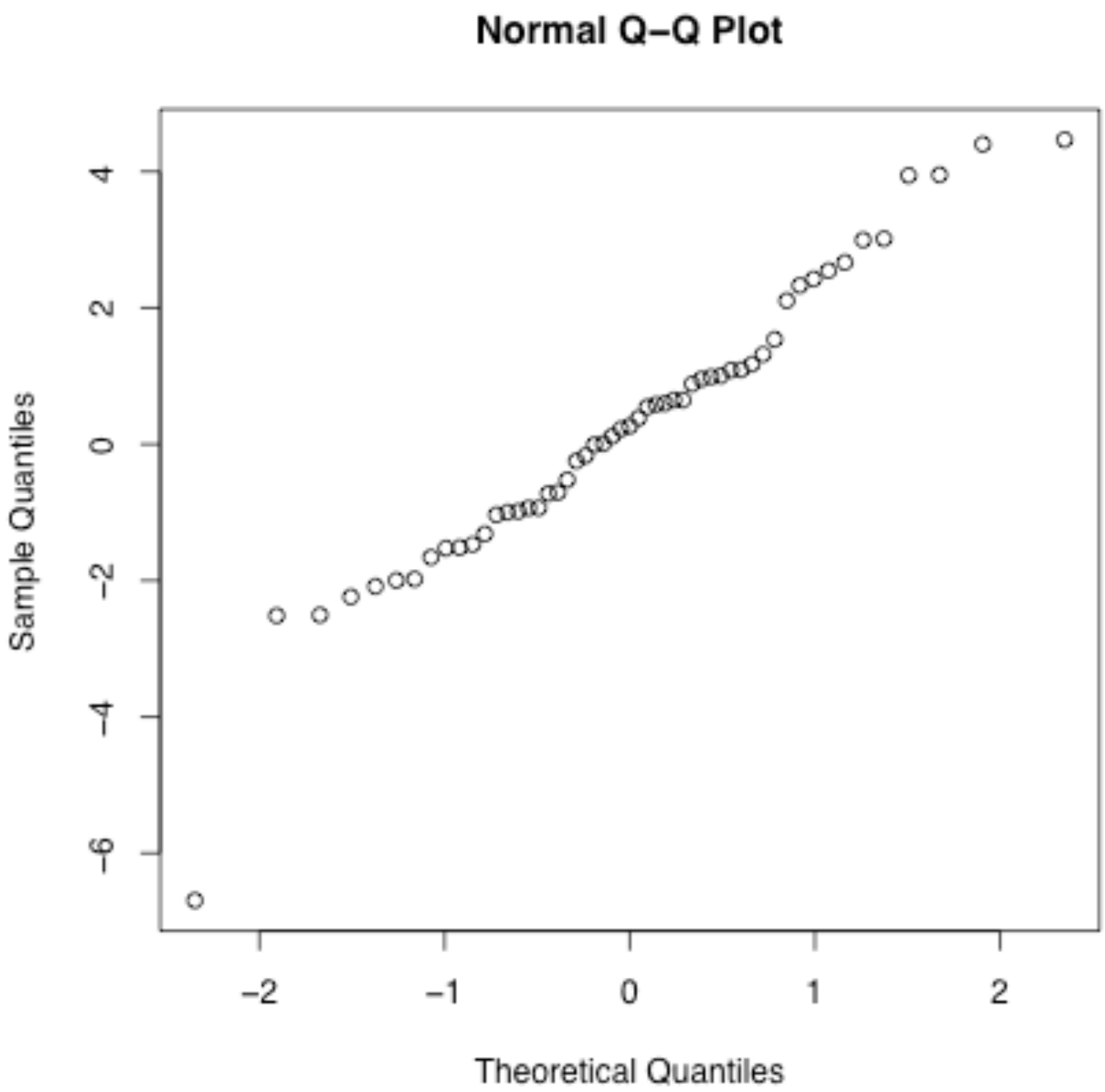
- adjustment for category size is needed
- an expected number per category can be obtained by using p*=mean of the posterior probabilities and the category size
- categories that deviate substantially from that expected number are of interest

Example: ALL Data

- samples on patients with ALL were assayed using HGu95Av2 GeneChips
- we were interested in comparing those with BCR/ABL (basically a 9;22 translocation) with those that had no cytogenetic abnormalities (NEG)
- 37 BCR/ABL and 42 NEG
- non-specific filter left us with 2526 probe sets

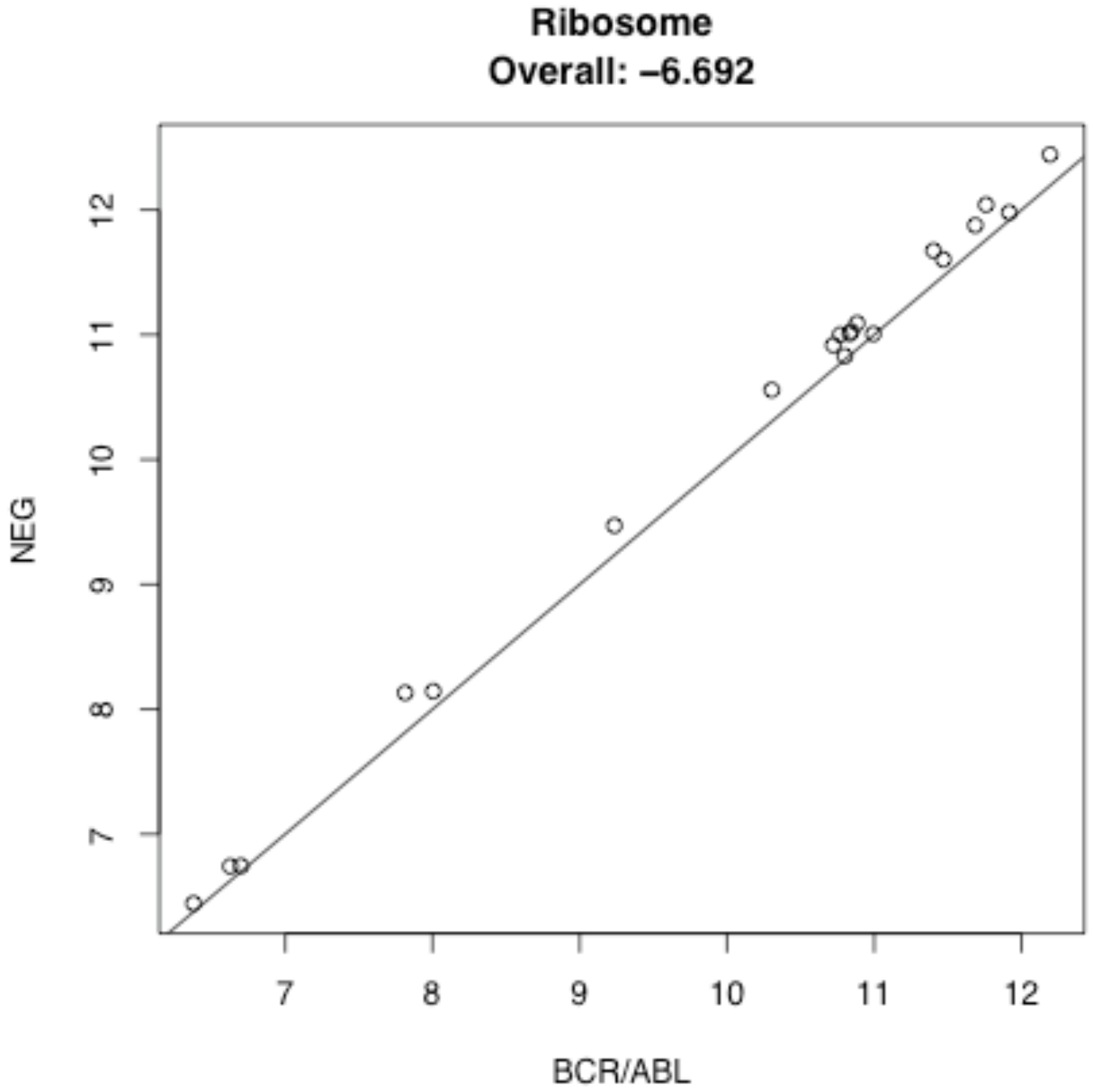
Example: ALL Data

- we then mapped the probes to KEGG pathways
- the mapping to pathways is via LocusLink ID
  - we have a many-to-one problem and solve it by taking the probe set with the most extreme *t*-statistic
- this left 556 genes
- much of the reduction is due to the lack of pathway information (but there is also substantial redundancy on the chip)
- then I decided to ignore categories with fewer than 5 members


Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
4
2
0
−2
−4
−6
−2
−1
0
1
2

Which Categories

- so the qq-plot looks interesting and identifies at least one category that looks interesting
- we identify it, and create a plot that shows the two group means (BCR/ABL and NEG)
- if all points are below or above the 45 degree line that should be interesting


Ribosome
Overall: −6.692
NEG
BCR/ABL
7
8
9
10
11
12

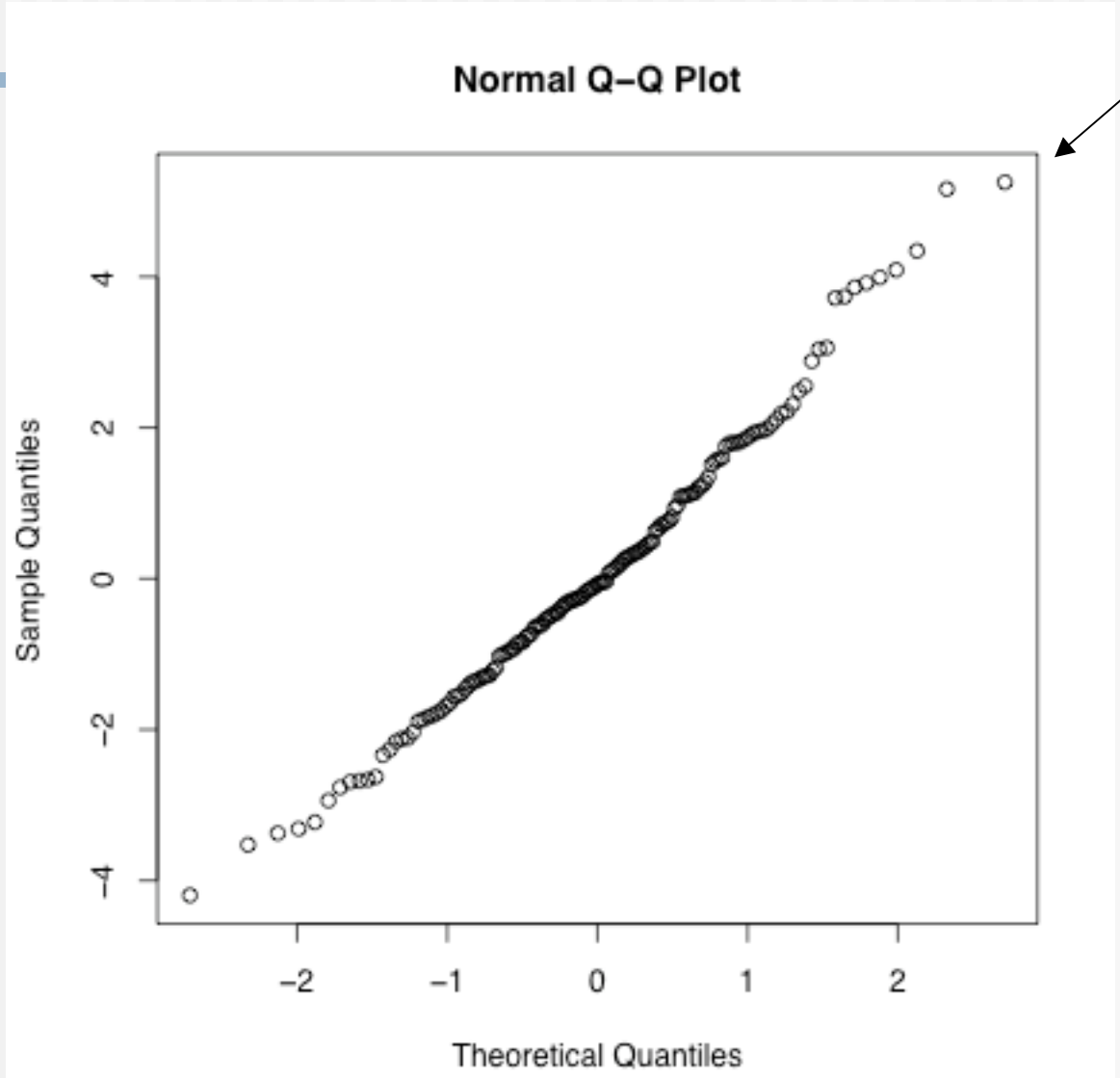
Ribosome

- the mean expression of genes in this pathway seem to be higher in the NEG group
- might be better to say suppressed in BCR/ABL (since they are relatively more homogeneous)

Permutation Test

- B=5000, p=0.05
- NEG> BCR/ABL
  - Ribosome
- BCR/ABL > NEG
  - Cytokine-cytokine receptor interaction
  - MAPK signaling pathway
  - Complement and coagulation cascades
  - TGF-beta signaling pathway
  - Apoptosis
  - Neuroactive ligand-receptor interaction
  - Huntington's disease
  - Prostaglandin and leukotriene metabolism

BCR/ABL vs NEG - Categories are cytochrome band (only those with more than 10 genes per band)

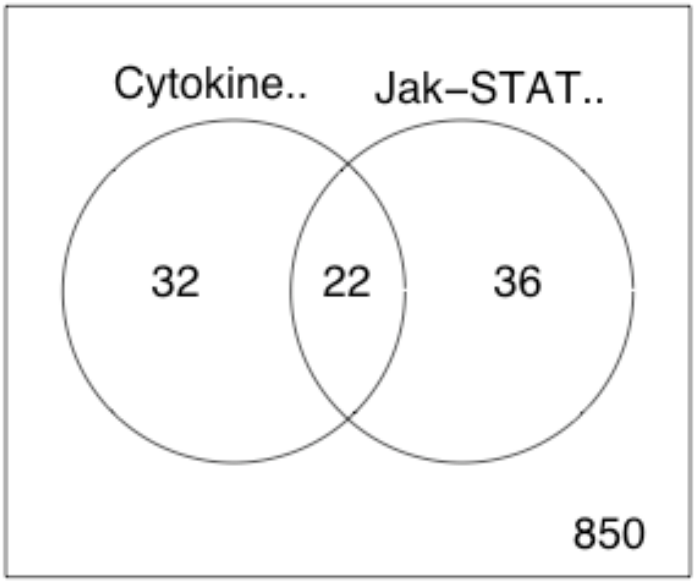


Two largest are 9q34 and 1p36 - both already implicated

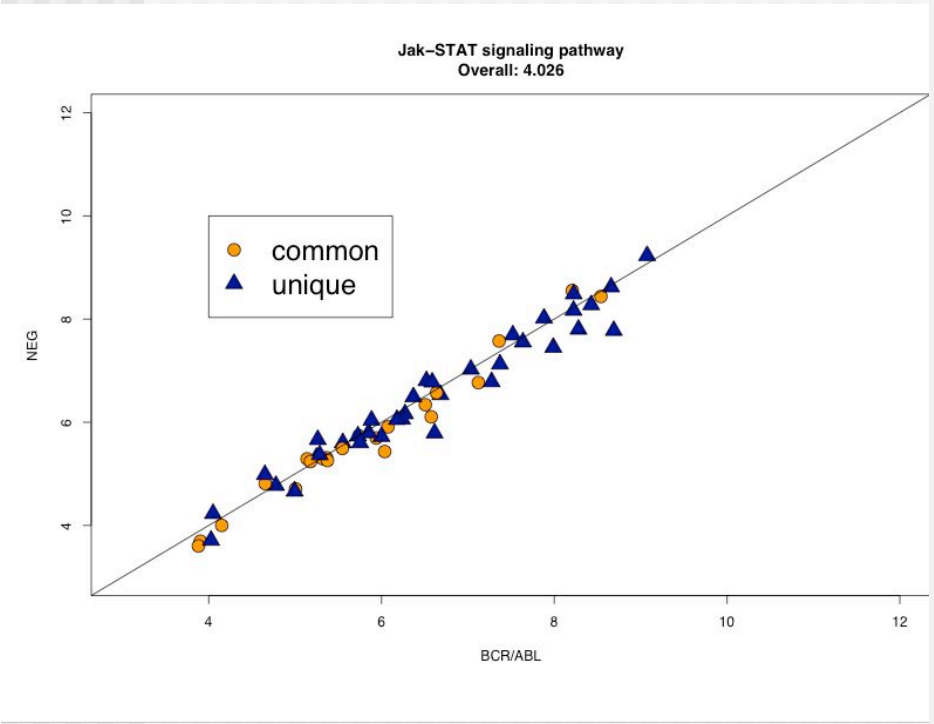
Aliasing

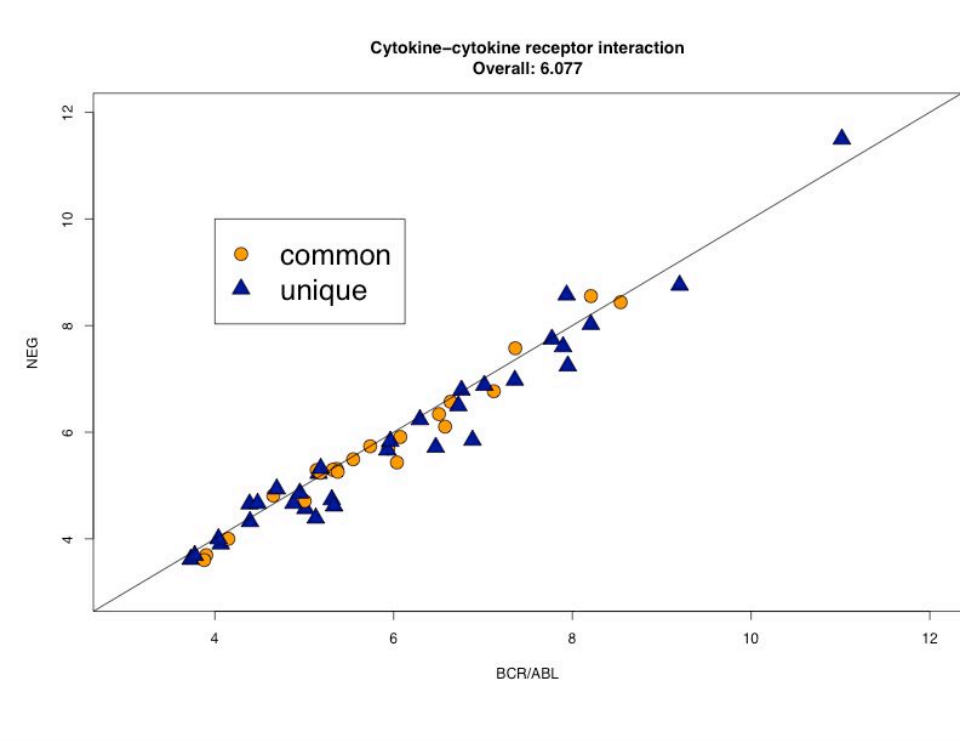
- all others have ignored this - but it does matter
- when we use categories, two categories can have substantial overlap
- if they are both significant, we might ask why

For cytokine-cytokine and Jak-Stat we have



Comparison of Gene Expression





Some other extensions

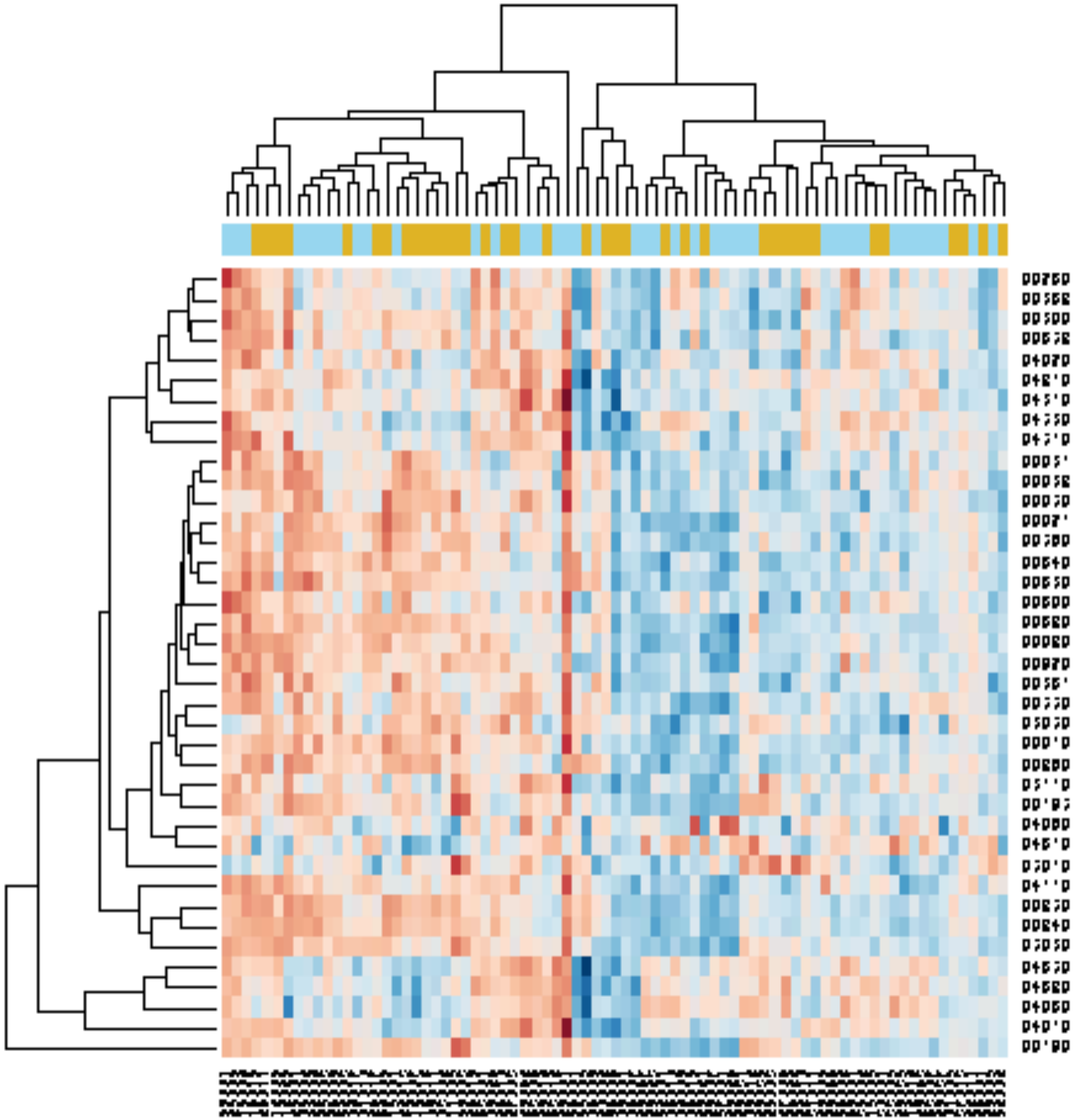
- categories might be a better way to do meta-analysis
- one of the fundamental problems with meta-analysis on gene expression data is the gene matching problem
- even technical replicates on the same array do not show similar expression patterns

Extensions

- if instead we compute per category effects these are sort of independent of the probes that were used
- matching is easier and potentially more biologically relevant
- the problem of adjustment still exists; how do we make two categories with different numbers of expression estimates comparable

Extensions

- you can do per array computations
- residuals are one of the most underused tools for analyzing microarrays
- we first filter genes for variability
- next standardize on a per gene basis - subtract the median divde by MAD
- now X*= AX, is a Cxn array, one entry for each category for each sample



Discovering Categories

- everything I have said up to now requires that categories be predefined
- how do we find new categories?
- use some form of feature selection (BMA, machine learning) and take the resulting features (genes)
- use those as *seeds* to find other genes whose expression is close to the seed gene
- those sufficiently close would form a category

Concluding Remarks

- the analysis of gene expression data still requires more research
- we should be looking at mechanisms for coordinated expression
  - transcription factors
  - amplifications
  - deletions
  - change in chromatin structure

Concluding Remarks

- *p*-value corrections are not really the right approach here
- bringing more biology to bear seems to be more likely to bear fruit
- we need some results to indicate how to deal with the coordinated gene expression (lack of independence within a category)

Acknowledgements

- Terry Speed (also some slides are his)
- Arden Miller
- Vincent Carey
- Michael Newton
- Kasper Hansen
- Jerry Ritz
- Sabina Chiaretti
- Sandrine Dudoit
- Zhen Jiang
- Adrian Raftery