# 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 artificial distinction - differentially expressed
  - it focuses attention on only a few genes that change by a large amount

# Differential Expression

- *p*-value correction methods don't really do what we want
- *p-values* are not signed, so the effects may be in different directions
- to see if too many genes of a particular type have been selected 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, or sets, of genes where there are potentially small but coordinated changes in gene expression
- for example, if all genes are expressed at slightly higher (or all at slightly lower) values for one  phenotype versus the other

# 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 for this work) but is more complex and computationally expensive
- they discuss gene sets, S, which are the same as categories
- they sidestep multiple testing issues by testing a single hypothesis (the maximal observed per set statistic)
- I will sidestep multiple testing issues by simply reporting unadjusted *p-values*

# 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.

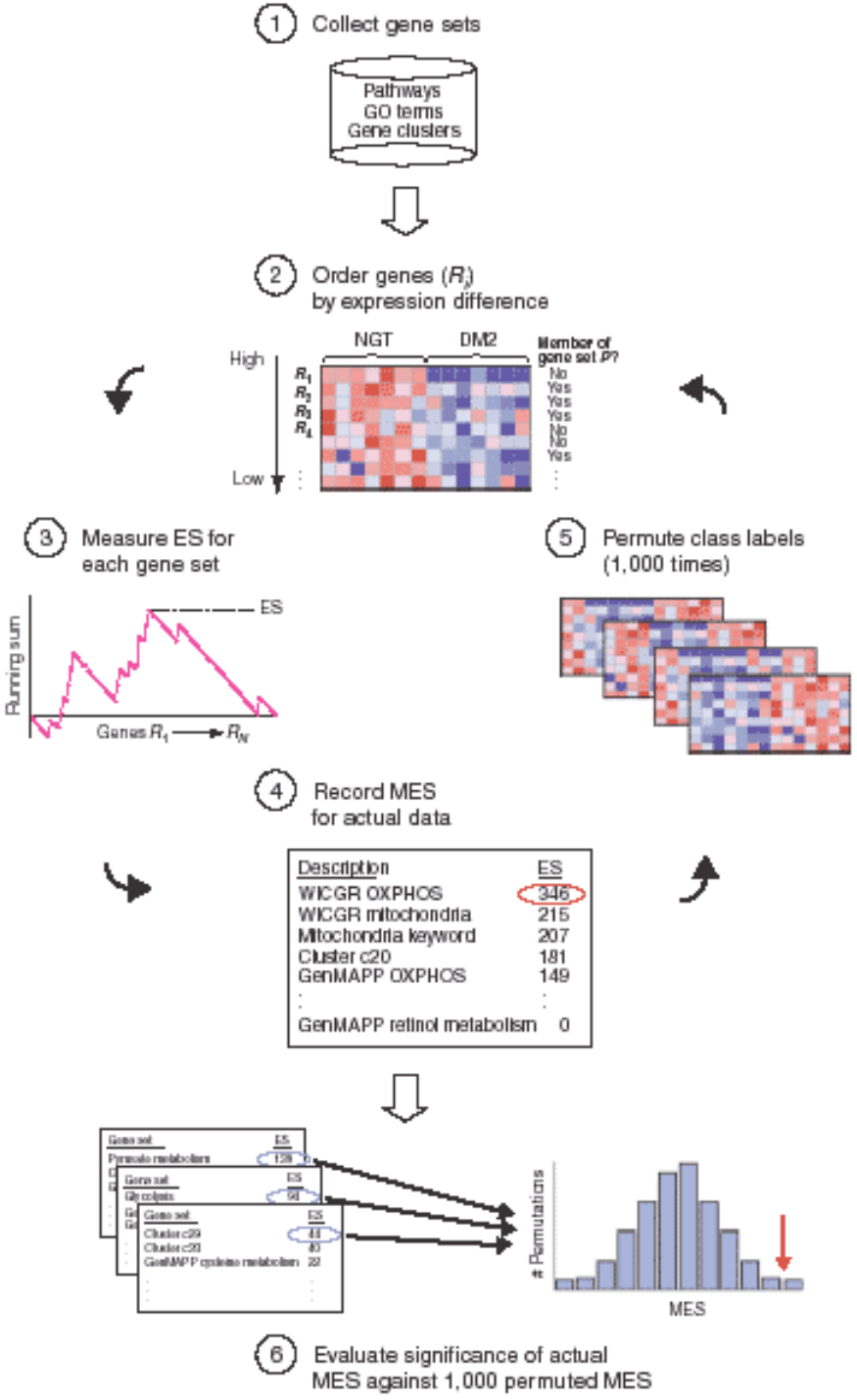
# Gene Set Enrichment

- 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

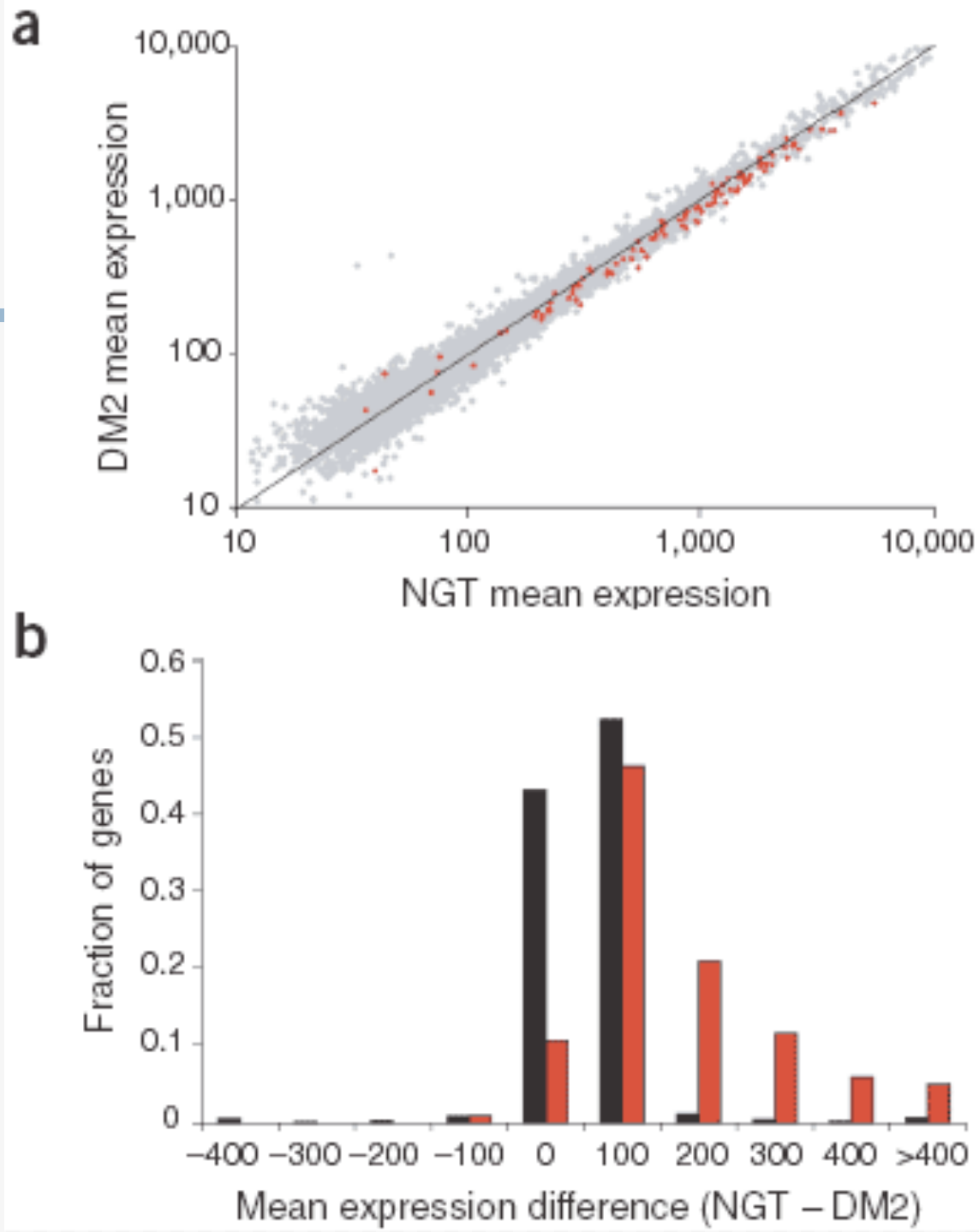
A set called OXPHOS got the largest ES score, with *p=0.029* on *1,000* permutations.



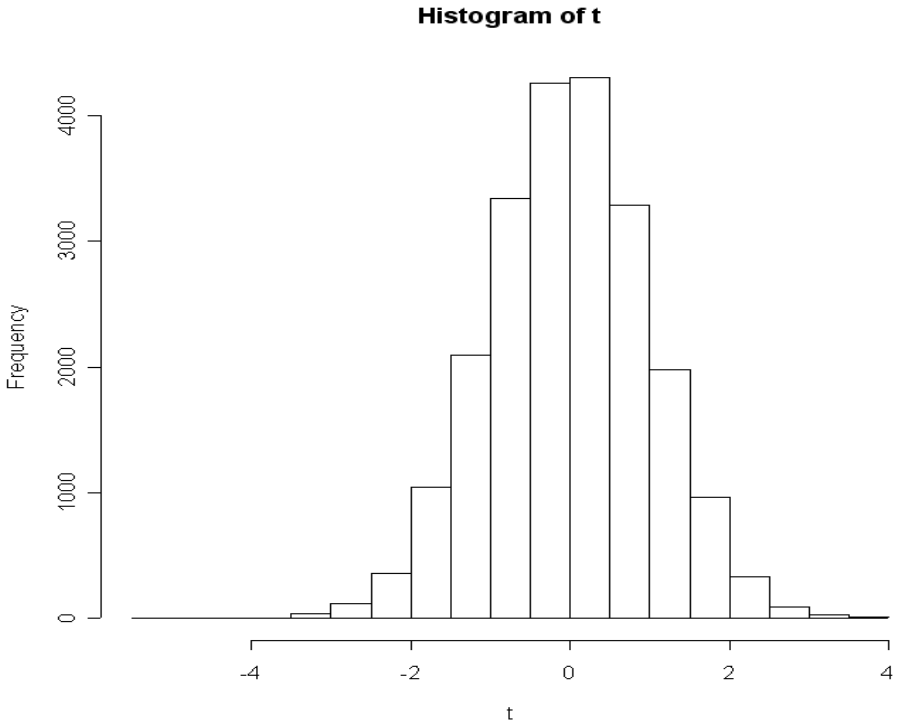
## OXPHOS

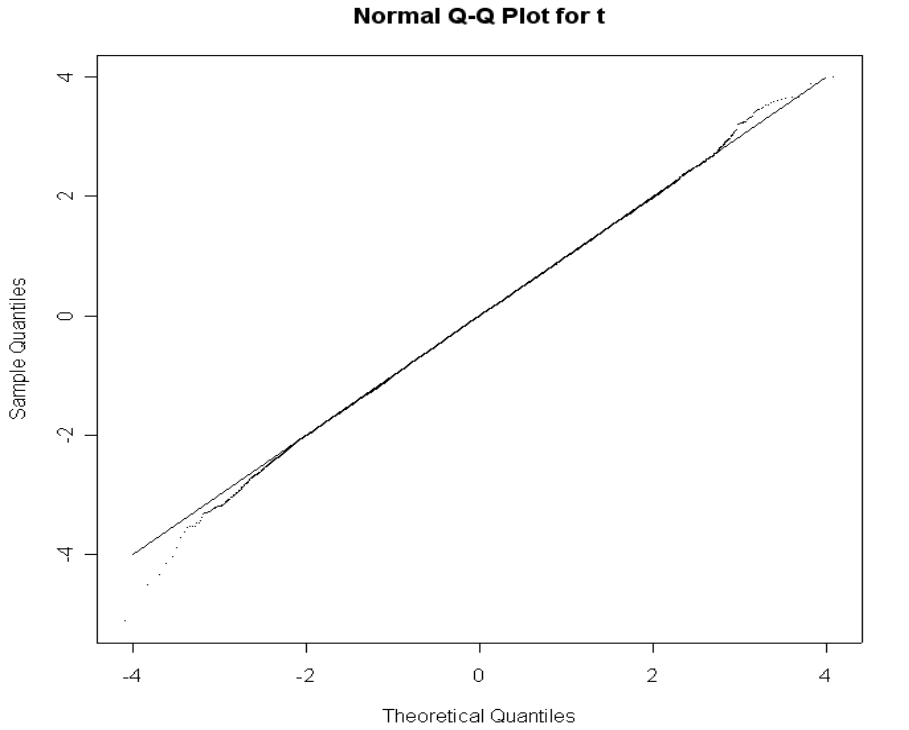
(A small difference for many genes)

**All genes**
**OXPHOS**

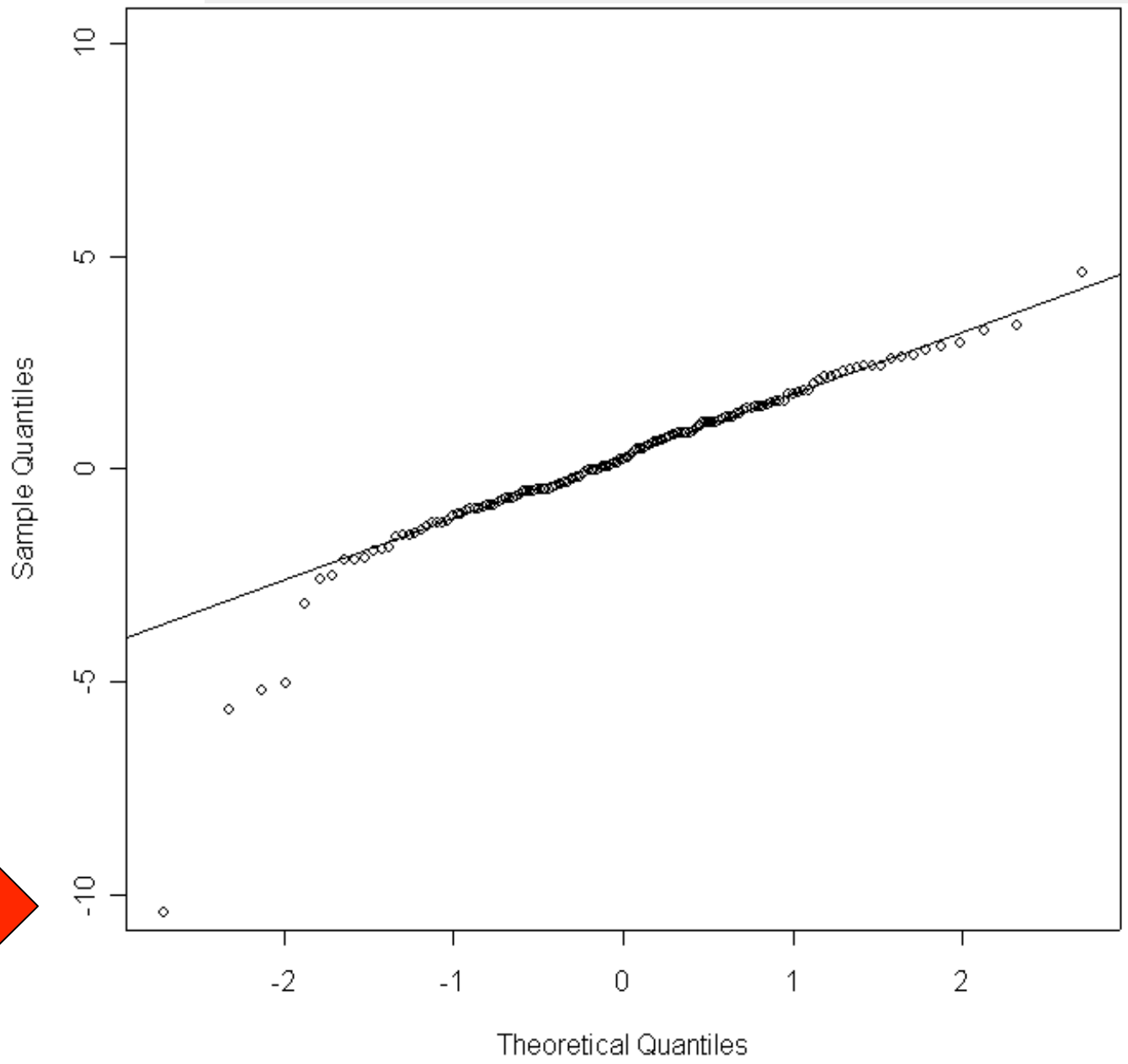


# Mootha's ts are approx normal



# 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; chromosome 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 reasonable

# Simple Statistical Approach

- the data matrix has G rows (one for each gene) and N columns (one for each sample)
- let's assume that there are two phenotypes of interest, so we have a two-sample comparison
- we can compute univariate test statistics, **x**, a G-vector
- select some set of categories, or gene sets, and let C denote the number of such sets
- you should address the problem that very commonly some genes are represented by a single probe and others by many (same for Hypergeometric testing)

# Categories

- define **A**, a C by G matrix, such that **A**[*i,j*]=1 if gene *j* is in category *i,* and **A**[*i,j*]=0, otherwise
- the row sums of **A** represent the number of genes in each category
- the column sums of **A** 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,

  $$\mathbf{z = Ax}$$

- we divide the sum by the square root of the number of genes per category (this is right if genes are independent - very unrealistic)
- then the resultant statistics, under the null hypothesis, have approximately a N(0,1) distribution
- we could also use other, per category, test statistics such as the median, or sign-test

# Categories: Reference Distribution

- an alternative is to generate many versions of **x**, the per category test statistic, from some reference distribution
- e.g. 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

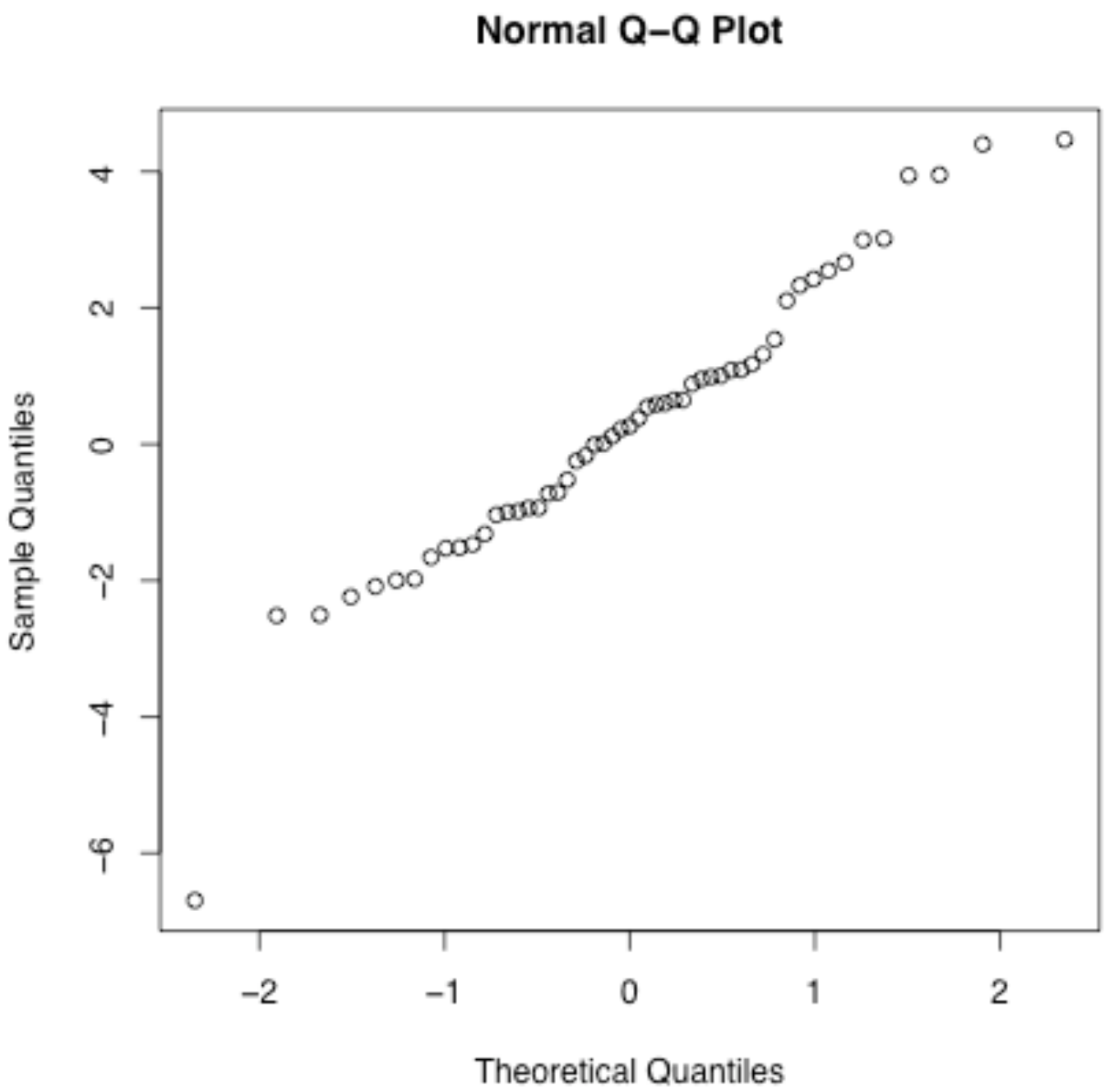
- adjustment for category size is needed
- an expected number per category can be obtained by using p*=mean of the posterior probabilities and the category size
- categories that deviate substantially from that expected number are of interest

# 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

# Example: ALL Data

- we then mapped the probes to KEGG pathways
- the mapping to pathways is via EntrezGene ID
  - we have a many-to-one problem and solve it by taking the probe set with the most extreme *t*-statistic
- we chose to only consider pathways with at least 10 genes
- this leaves us with 79 samples, 1036 genes and 70 pathways


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

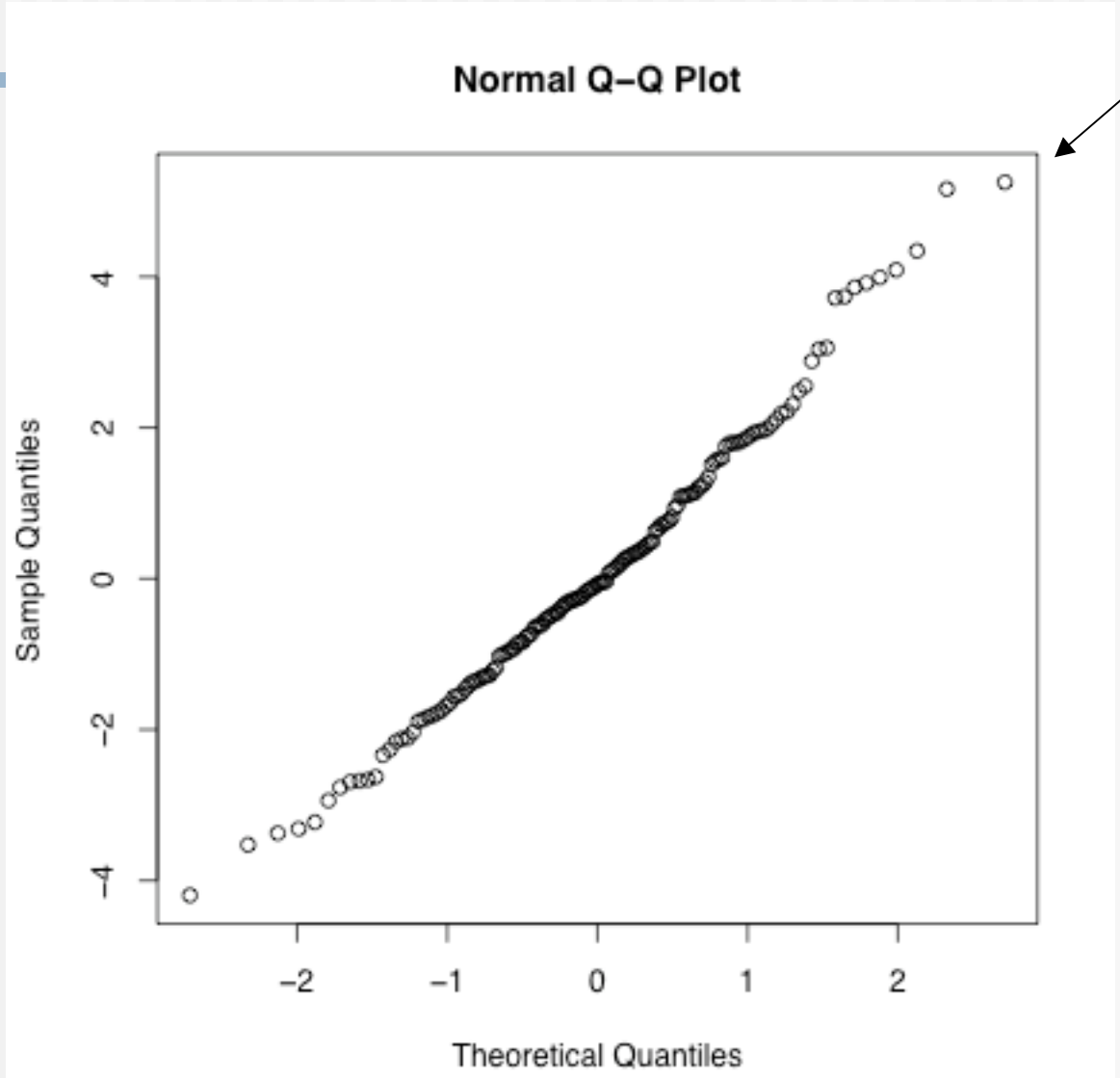
# Which Categories

- 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

# Different Univariate test statistics

| | ID | PW Name | P.v$^{Mn}$ | P.v$^{Md}$ | P.v$^{ST}$ | Size |
|---|---|---|---|---|---|---|
| 1 | 04514 | Cell adhesio... | 0.0000 | 0.0000 | 0.0008 | 38 |
| 2 | 04940 | Type I diabe... | 0.0018 | 0.0020 | 0.0013 | 20 |
| 3 | 04060 | Cytokine-cyt... | 0.0030 | 0.0050 | 0.0001 | 54 |
| 4 | 04610 | Complement a... | 0.0000 | 0.0004 | | 14 |
| 5 | 04512 | ECM-receptor... | 0.0000 | 0.0004 | | 15 |
| 6 | 04530 | Tight juncti... | 0.0000 | 0.0020 | | 40 |
| 7 | 04520 | Adherens jun... | 0.0000 | 0.0034 | | 34 |
| 8 | 04670 | Leukocyte tr... | 0.0002 | 0.0010 | | 49 |
| 9 | 04080 | Neuroactive ... | 0.0002 | 0.0012 | | 20 |
| 10 | 04510 | Focal adhesi... | 0.0006 | 0.0028 | | 73 |
| 11 | 01430 | Cell Communi... | 0.0014 | 0.0004 | | 12 |
| 12 | 03010 | Ribosome | | 0.0080 | 0.0000 | 23 |
| 13 | 04360 | Axon guidanc... | 0.0004 | | | 38 |
| 14 | 04810 | Regulation o... | 0.0066 | | | 79 |
| 15 | 04210 | Apoptosis | 0.0096 | | | 46 |
| 16 | 04640 | Hematopoieti... | | 0.0008 | | 38 |
| 17 | 00190 | Oxidative ph... | | | 0.0001 | 59 |
| 18 | 00620 | Pyruvate met... | | | 0.0003 | 16 |
| 19 | 00230 | Purine metab... | | | 0.0027 | 58 |
| 20 | 04110 | Cell cycle | | | 0.0046 | 66 |
| 21 | 00071 | Fatty acid m... | | | 0.0065 | 14 |
| 22 | 00010 | Glycolysis /... | | | 0.0085 | 22 |

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

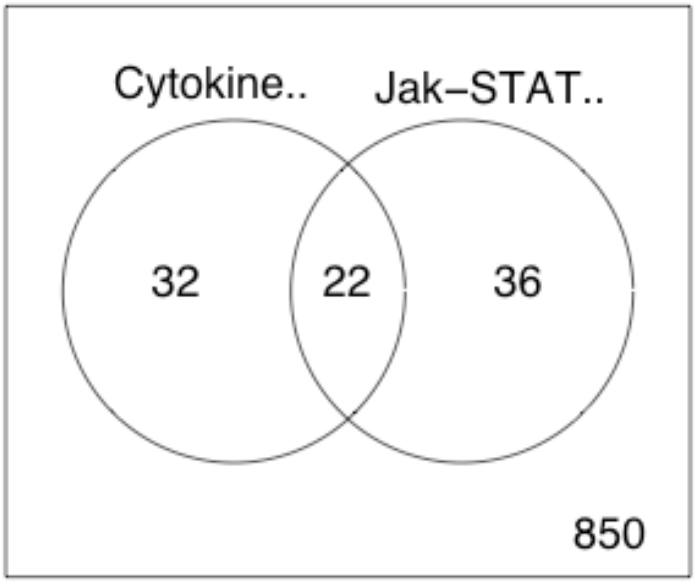



Two largest are 9q34 and 1p36 - both already implicated

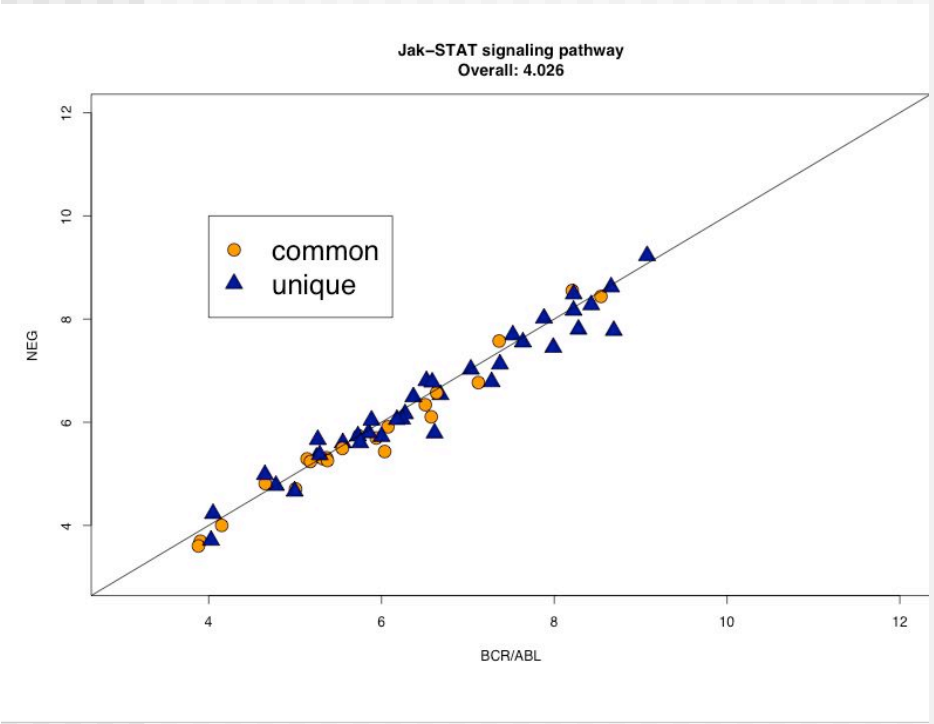
# Aliasing

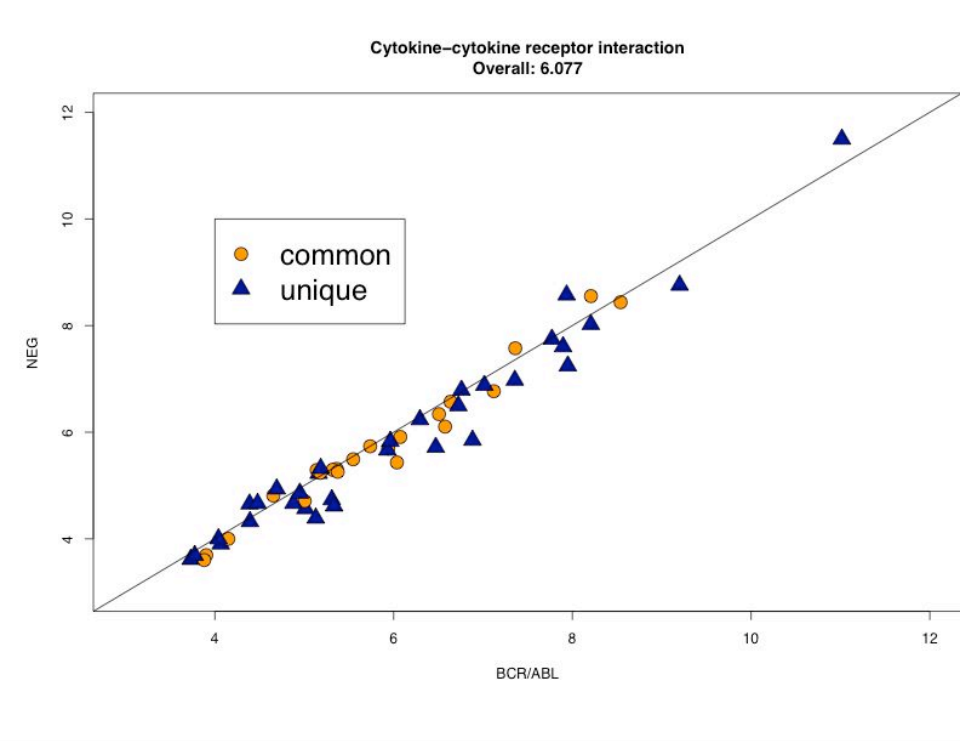
- all others have ignored this - but it does matter
- when we use categories, two categories can have substantial overlap
- if they are both significant, we might ask why

# For cytokine-cytokine and Jak-Stat we have



# Comparison of Gene Expression



# The Analysis

- and when the genes involved, are separated into three groups
  - those in Cytokine-Cytokine only
  - those in Jak-Stat only
  - those common
- then we find that the first and third are significant, but the second (Jak-Stat alone) is not

# Some other extensions

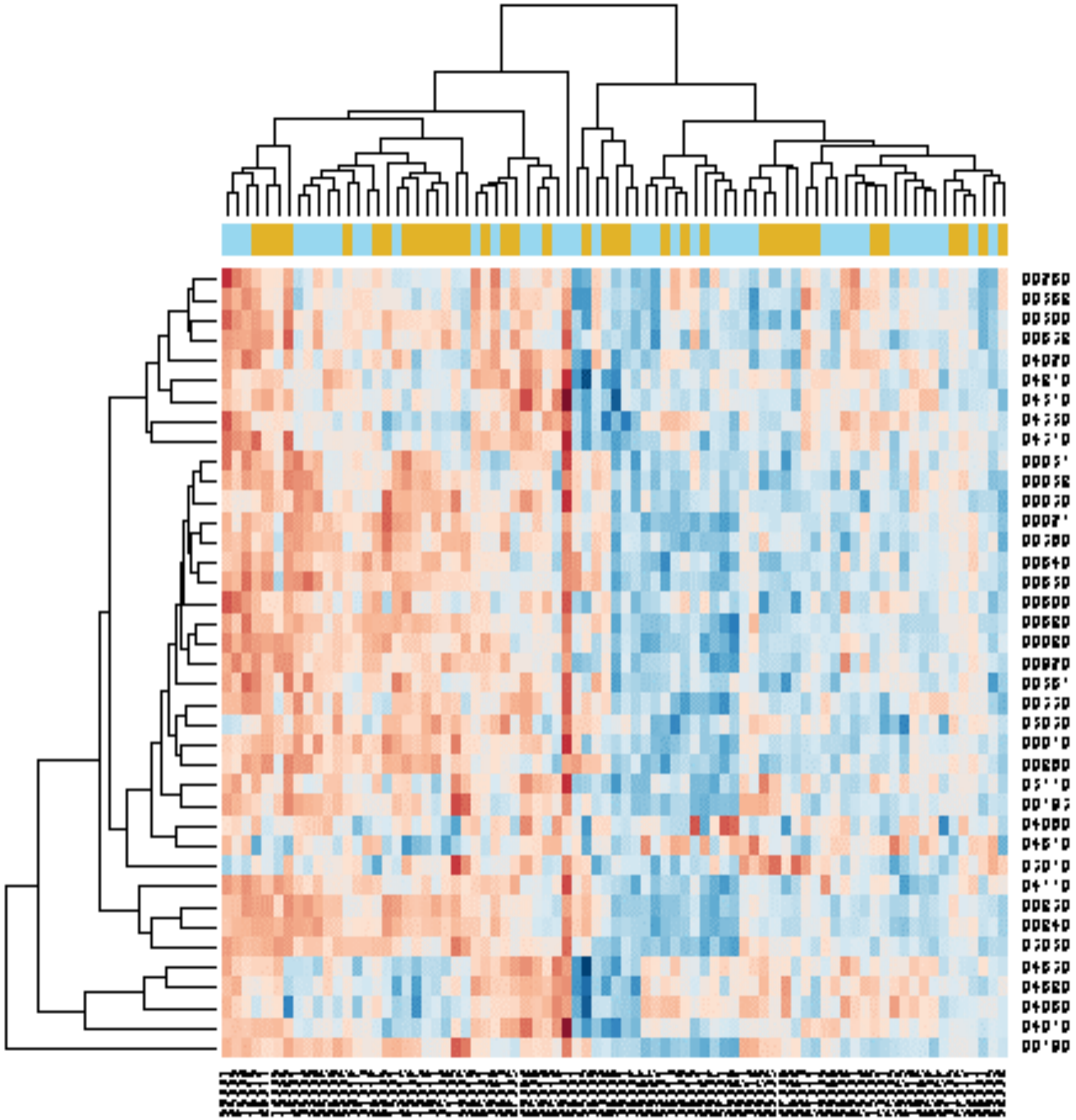
- categories might be a better way to do meta-analysis
- one of the fundamental problems with meta-analysis on gene expression data is the gene matching problem
- even technical replicates on the same array do not show similar expression patterns

# 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 divide by MAD
- now X*= AX, is a Cxn array, one entry for each category for each sample



# 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
- Seth Falcon