Gene Set Enrichment Analysis

Robert Gentleman

Outline

- Description of the experimental setting
- Defining gene sets
- Description of the original GSEA algorithm
  - proposed by Mootha et al (2003)
- Our approach + some extensions

Experiments/Data

- there are n samples
- for each sample G different genes are measured
- the resultant data are stored in a matrix **X** (G x n)
- a univariate, per gene, statistic can be computed, **x**, (G x 1)
  - often a t-test comparing two groups, but we can pretty much deal with anything

Differential Expression

- Usual approach is to
  1. find the set of differentially expressed genes [those with extreme values of the univariate statistic, **x**]
  2. use a Hypergeometric calculation to identify those gene sets with too many (sometimes too few) differentially expressed genes

Differential Expression

- dividing genes into two groups
  - differentially expressed
  - not differentially expressed

  is somewhat artificial
- *p*-value correction methods don't really do what we want
  - they seldom change the ranking (and shouldn't) so they might change the location of the cut
  - but the artificial distinction remains
- favors finding groups enriched for some genes whose expression changes a lot

A Different Approach

- a different approach is to make use of all of the genes not just the DE ones
- we recommend only using the non-specific filtering methods
- we will attempt to find gene sets where there are potentially small but coordinated changes in gene expression
- an obvious situation is one where genes in a gene set all show small but consistent change in a particular direction

Gene Sets

- can be obtained from biological motiviations: GO, KEGG etc
- from experimental observations: DE genes reported in some paper
- predefined sets from the published literature etc
- regions of synteny; cytochrome bands

Gene Sets

- the GSEABase package in BioC provides substantial infrastructure for holding and manipulating Gene Sets
- they can have values associated with the genes
  - weights
  - +/- 1 to indicate positive or negative regulation
- a collection of gene sets does not need to be exhaustive or disjoint

Gene Sets

- the mapping from a set of entities (genes) to a collection of gene sets can be represented as a bipartite graph
  - one set of nodes are the genes
  - the other are the gene sets
- this mapping can be represented by an incidence matrix, **A** (C x G)

Gene Sets

- the elements of **A**, **A**[*i,j*]=1 if gene *j* is in gene set *I*, it is 0 otherwise
- the row sums represent the number of genes in each gene set
- the column sums represent the number of gene sets a gene is in
- if two rows are identical (for a given set of genes) then the two gene sets are aliased (in the usual statistical sense)
- other patterns can cause problems and need some study

Gene Sets

- the simplest transformation is to use

  $\mathbf{z} = \mathbf{A}\mathbf{x}$

- $\mathbf{x}$ is the vector of t-statistics (or alternatives)
- so that $\mathbf{z}$ is a C-vector, and in this case represents the per gene set sums of the selected test statistics
- we are interested in large or small $\mathbf{z}$'s
- potentially adjusted for the number of entities in the gene set (size)
  - often division by the square root of the number of genes in the gene set

Other Properties

- there is a certain amount of robustness to being correct about the mapping
- a strong signal may be detected even if not all genes in a gene set are identified
- there is also tolerance to some genes being incorrectly associated with the gene set
- this is in contrast to the usual method of differential expression - there we identify particular genes and hence are more subject to errors in annotation

Gene Set Enrichment (Original)

- 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.

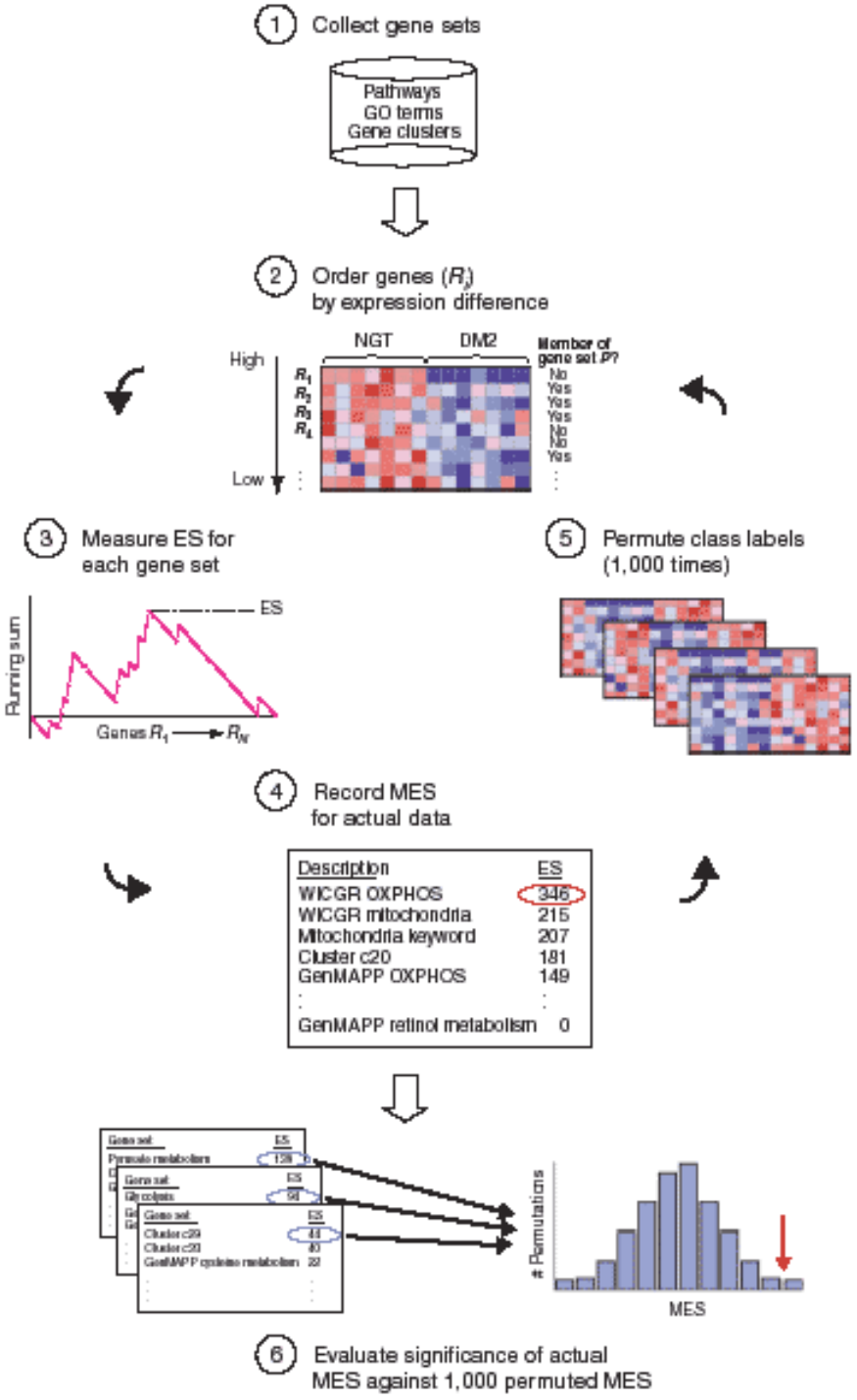
Gene Set Enrichment(Original)

- 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.



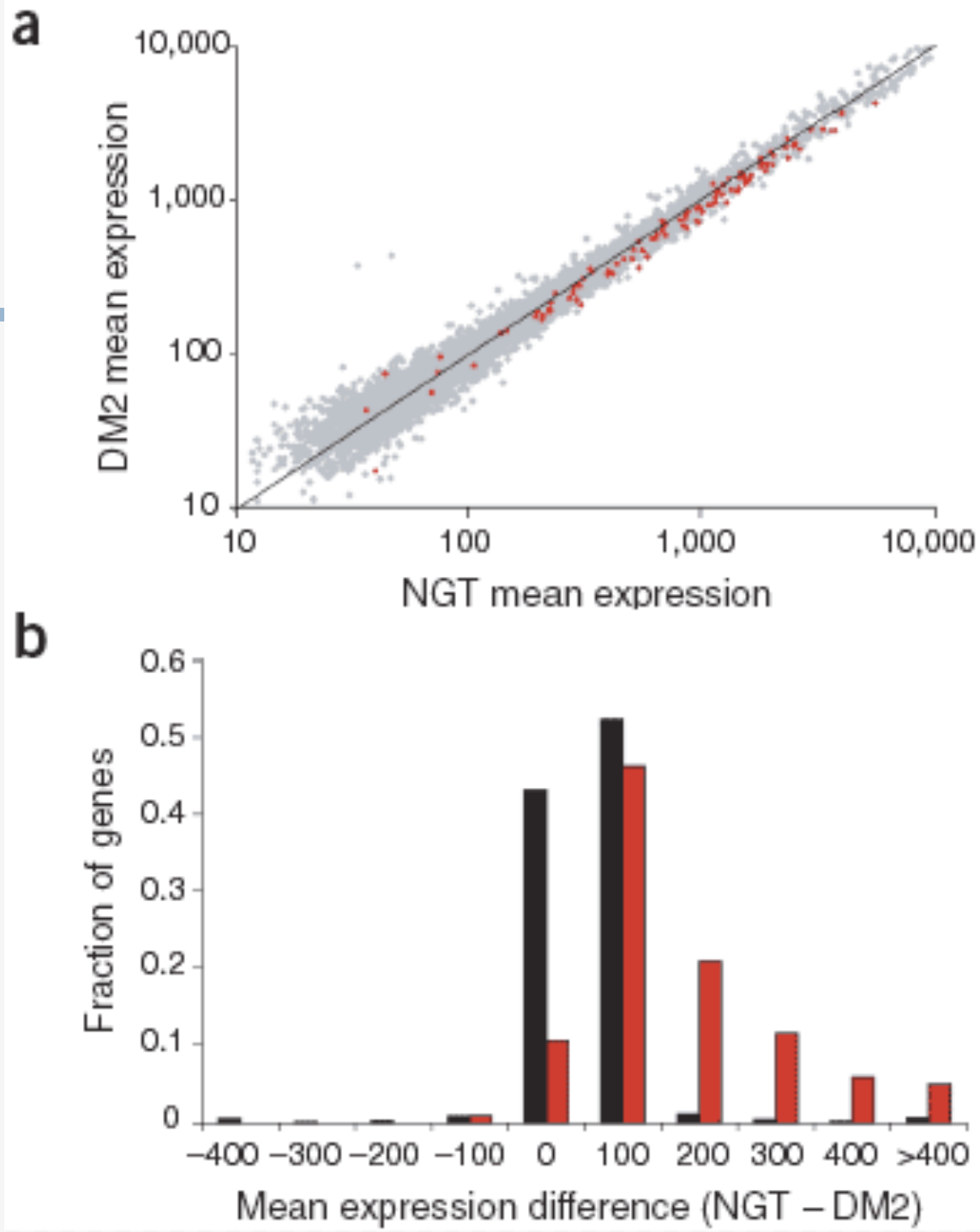
OXPHOS

Other

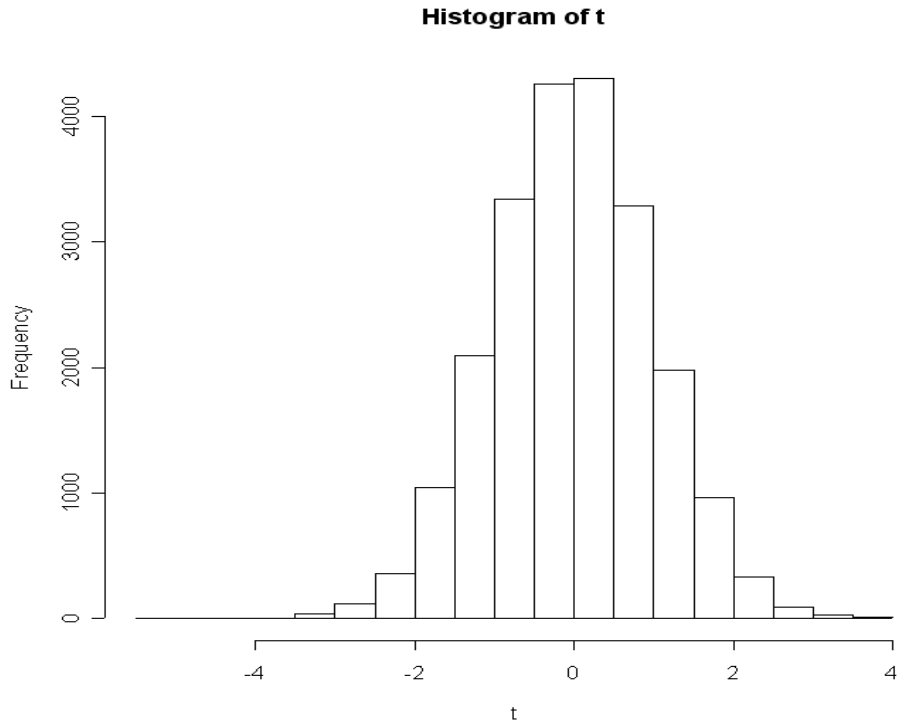
(A small difference for many genes)

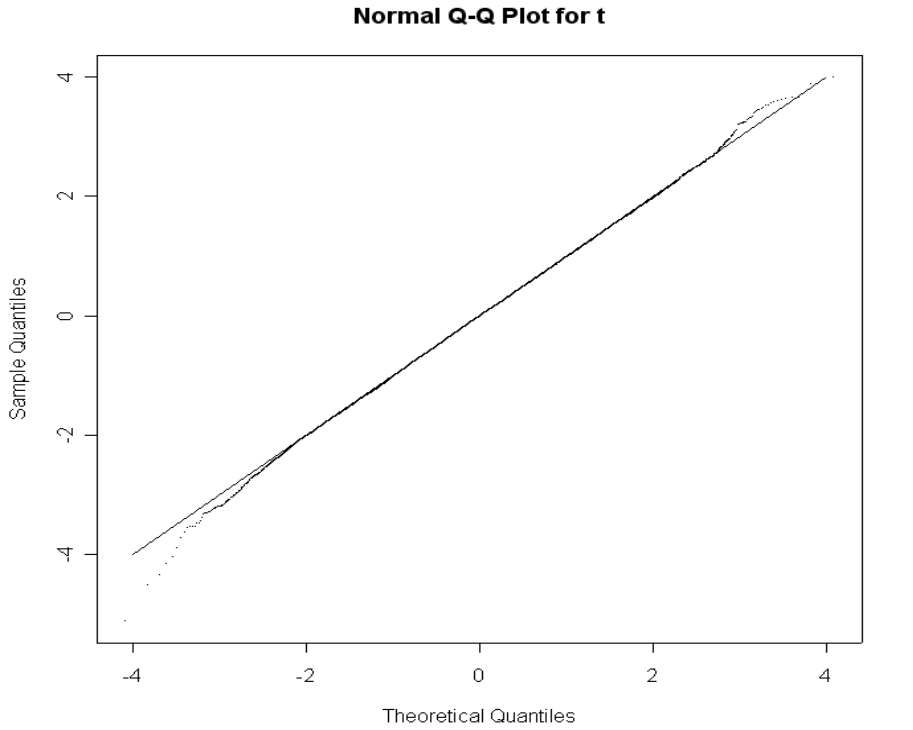
All genes

OXPHOS

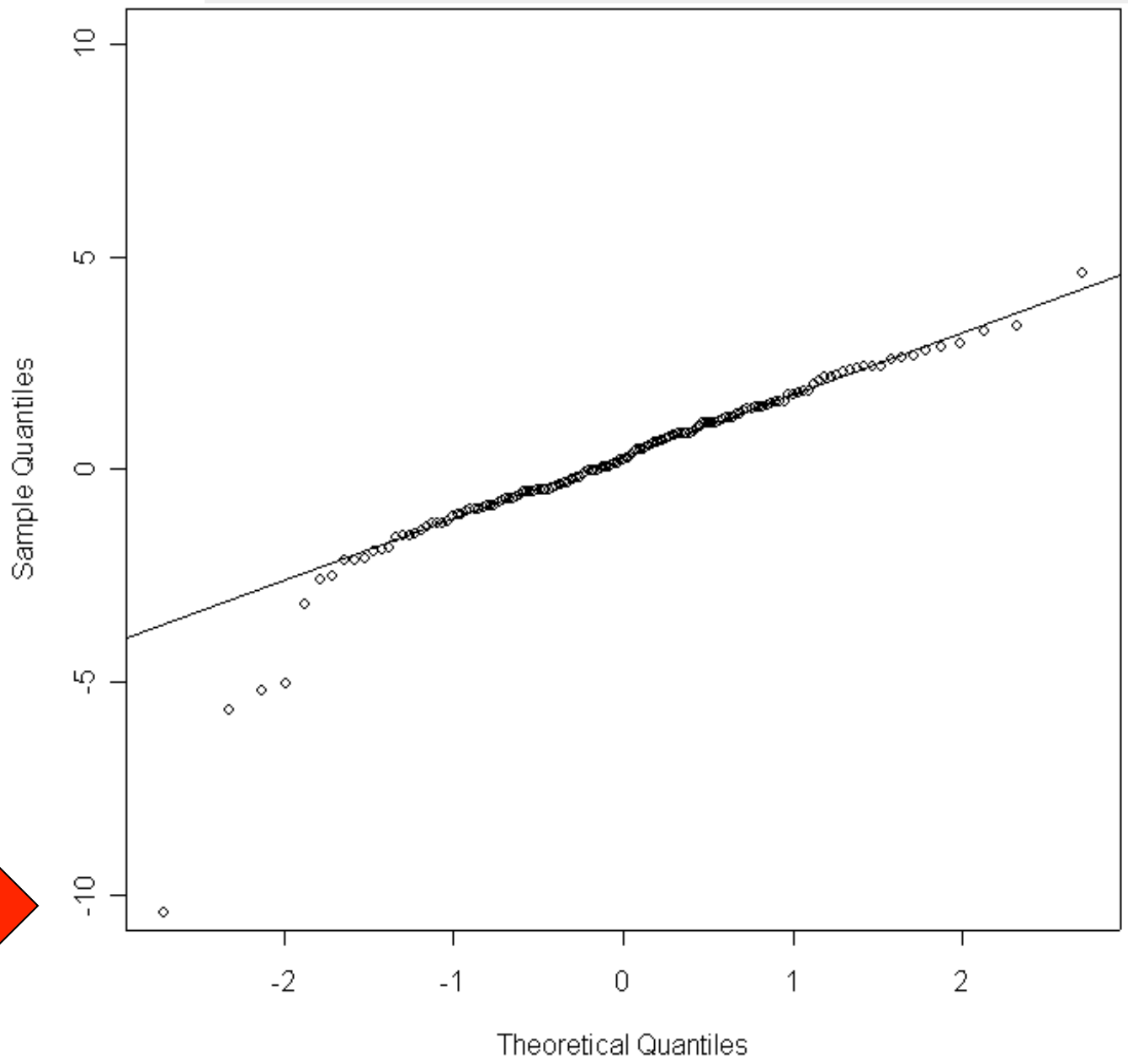


Mootha's ts are approx normal





Normal qq-plot of $\Sigma t/\sqrt{n}$



OXPHOS

Gene Sets: Distribution

- so what might be sensible
- if n (the number of samples) is large-ish and we use a *t*-test to compare two groups
- and if H_0*: no difference between the group means* is true, for all genes
- then the elements of **x** are approximately *t* with n-1 df (for large n this is approximately N(0,1))
- so that the elements of **z** are sums of N(0,1) and if we divide by the square root of the row sums of **A** we are back at N(0,1) [sort of]

Gene Sets: Distribution

- the problem is that that relies on the assumption of independence between the elements of **x**, which does not hold
- but it does give some guidance and a qq-plot of the **z**'s can be quite useful (as we saw above)

Summary Statistic

- one choice is to use:

$$T = \frac{\sum X}{\sqrt{n}}$$

- a second is to use the regression:

$$Y_i = \alpha + \beta 1_{i \in GS} + \varepsilon_i$$

Gene Sets: Reference Distribution

- an alternative is to generate many **x**'s from a reference distribution
- one distribution of interest is to go back to the original expression data and either permuting the sample labels or bootstrapping can be used to provide a reference distribution

Comparisons

- you can test whether for a given gene set is the observed test statistic unusual
- or test whether any of the observed gene set statistics are unusually large with respect to the entire reference distribution

Extensions

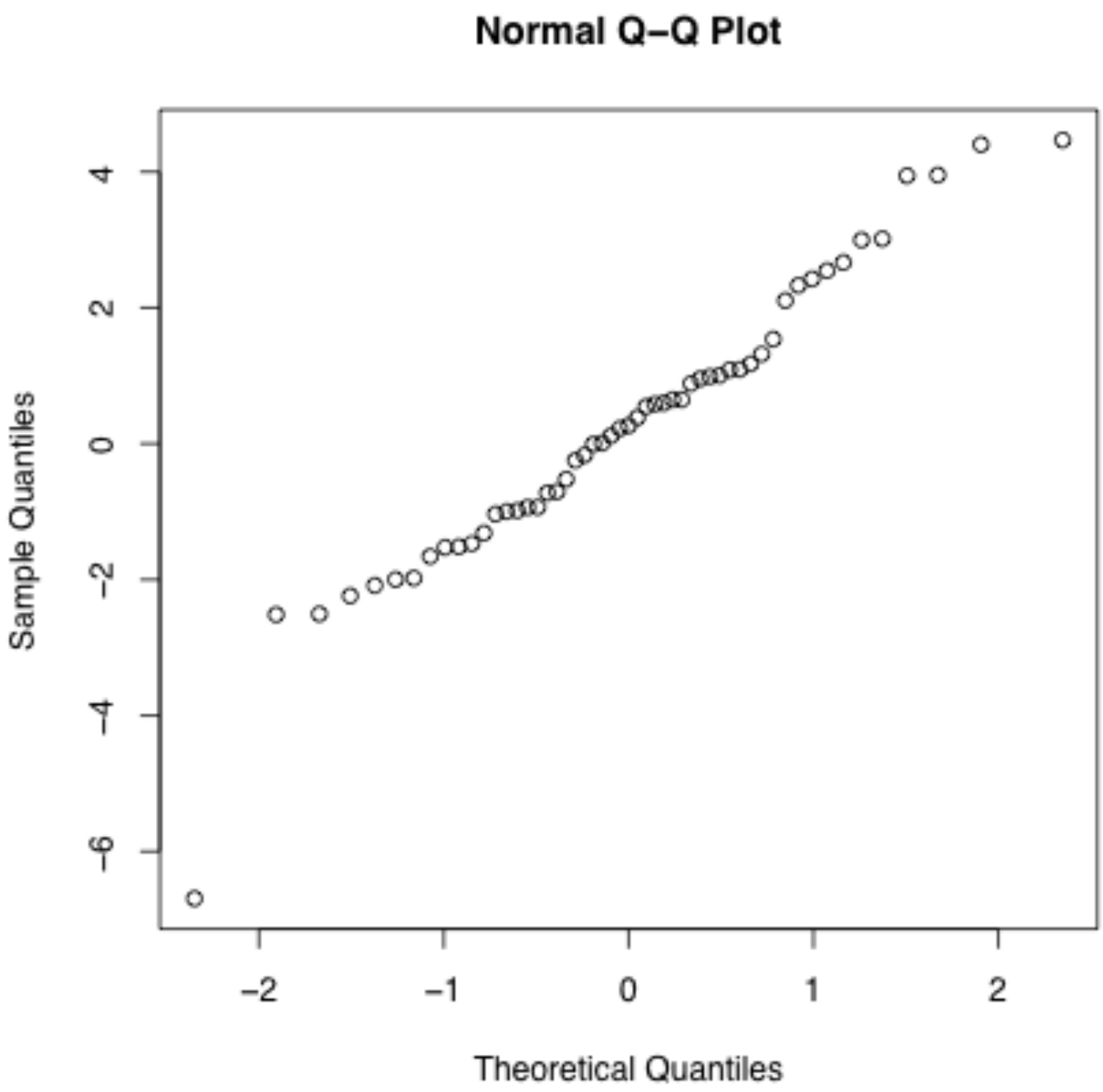
- there is no need to compute sums over gene sets
  - you could use medians, any other statistic, such as a sign test
- the regression approach can be extended to
  - include covariates/multiple gene sets
  - use residuals (both for gene sets and for samples)

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

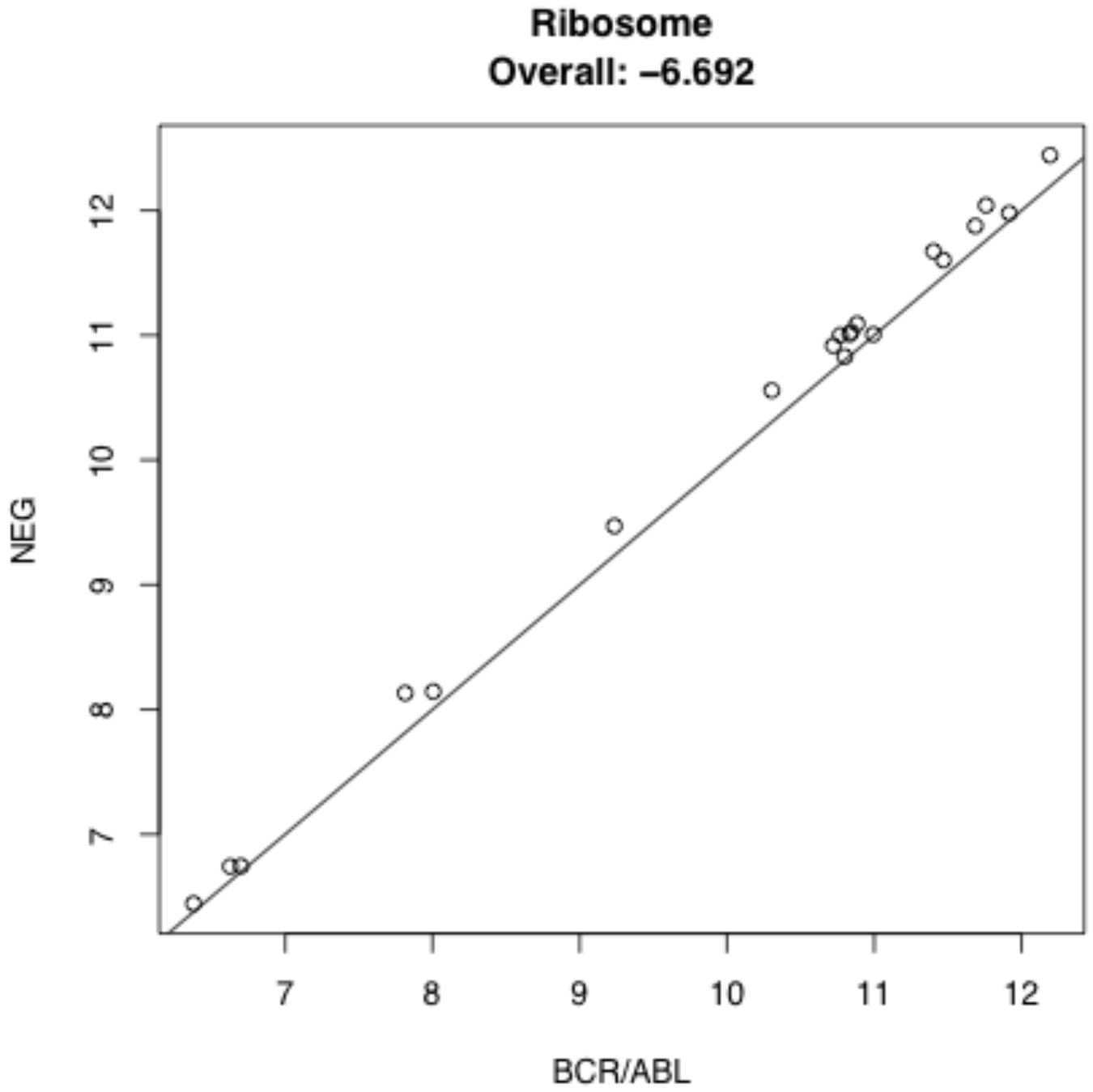
Example: ALL Data

- 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 gene sets 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 Gene Sets

- so the qq-plot looks interesting and identifies at least one gene set that is different
- we identify it (Ribosome), 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
- unfortunately the result is spurious - sex needs to be accounted for
  - the groups are not balanced by sex
  - and there is a ribosomal gene encoded on the Y chromosome

Alternative: 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

Recap

- basic idea is to make use of all genes
- summarize per gene data **X** (G x n) to **x** (G x 1)
  - $\mathbf{x} = f_1(\mathbf{X})$
- use predefined gene sets
  - these define a bipartite graph **A** (C x G)
- summarize the relationship between the gene sets and the per gene summary stats
  - $\mathbf{z} = f_2(\mathbf{A}, \mathbf{x})$

Recap

- the summaries of the data, **X**, f_1, can be any test statistic
  - doesn't really need to be 1 dimensional
- the transformations **(A, x)**, f_2, can be sums, or many other things (medians, sign tests etc)

Some other extensions

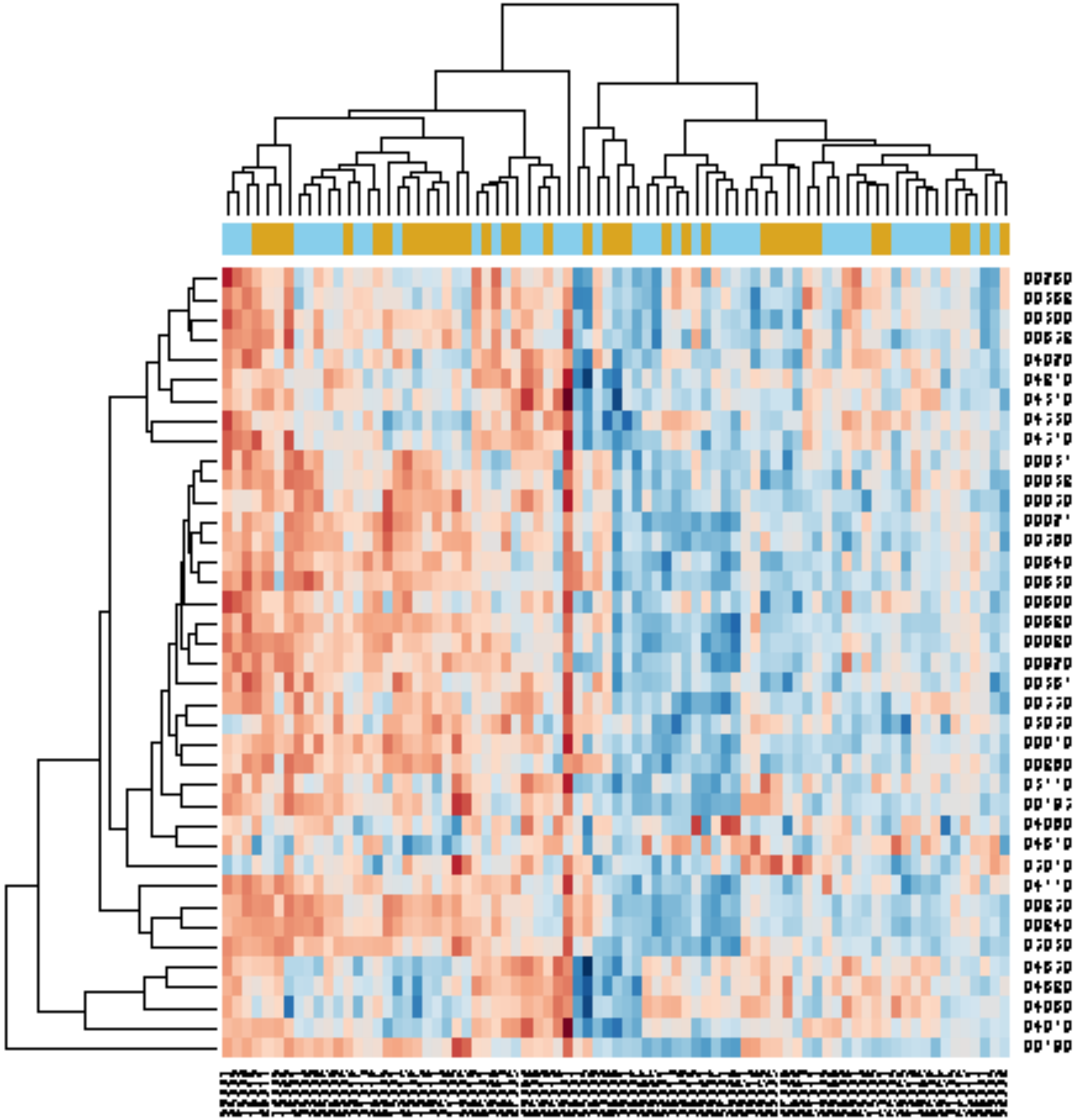
- gene sets might be a better way to do meta-analysis
- one of the fundamental problems with meta-analysis on gene expression data is the gene matching problem
- even technical replicates on the same array do not show similar expression patterns

Extensions: Meta-analysis

- if instead we compute per gene set 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 gene sets 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 gene set for each sample



References

- there is a rich body of literature
- my two main contributions are

Gene set enrichment analysis using linear models and diagnostics. Oron AP, Jiang Z, Gentleman R.Bioinformatics. 2008 Nov 15;24(22):2586-91. Epub 2008 Sep 11.

Extensions to gene set enrichment.Jiang Z, Gentleman R.Bioinformatics. 2007 Feb 1;23(3):306-13. Epub 2006 Nov 24.

Acknowledgements

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