Gene enrichment analysis

Most of today's material courtesy of Terry Speed

Are sets of genes differentially expressed?

The sets we refer to here are all the *outcomes* of analyses. Later we discuss sets specified *a priori*.

Examples of sets. They could be the list of all genes whose differential expression (e.g. average *M*-value) exceeds a given threshold, typically a liberal one, which would not correspond to any real "significance", e.g. *1.5*-fold. They might be clusters.

What do we mean by a set being differentially expressed. Here it is a convenient shorthand for being unusual in relation to all the genes represented on the array, for example, by being functionally enriched, in the sense of having more genes of a given category than one would expect, *by chance*.

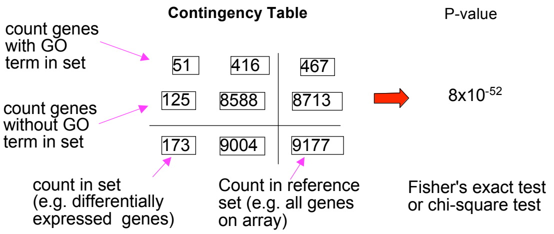
GO and microarray gene sets

Hypothesis: Functionally related, differentially expressed genes should accumulate in the corresponding GO-group.

Problem: to find a method which scores accumulation of differential gene expression in a node of the GO.

We describe the calculation from the program Gostat. For all the genes analysed, it determines the annotated GO terms and all splits. It then counts the # of appearances of each GO term for the genes in the set, as well as the # in the reference set, which is typically all genes on the array. Then a 2×2 table is formed, see over page, and a *p*-value calculated.

Is a GO term is specific for a set?



Contingency Table

51	416	467
125	8588	8713
173	9004	9177

count genes with GO term in set

count genes without GO term in set

count in set (e.g. differentially expressed genes)

Count in reference set (e.g. all genes on array)

P-value

$8x10^{-52}$

Fisher's exact test or chi-square test

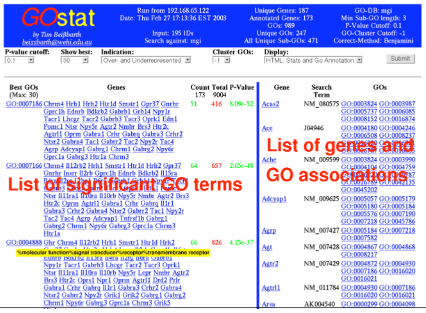
The multiple testing problem

Naturally one doesn't test a single GO term or split, but many, perhaps 1000s. As with testing of single genes, we need to deal with the multiple testing problem. Many of the solutions from there carry over: Bonferroni, Holm, step-down minP, FDR, and so on. But there are also special problems here, deriving from the nesting relationships between splits. In my view, these are not easily dealt with, and require more research.

Related questions. How can we compare the results of different lists being compared? And, rather than select a set of genes using a cut-off, can we make use the gene abundances or p-values for differential expression?

GOstat: Tool for finding significant GO terms in a list of genes

http://gostat.wehi.edu.au



Run from 192.168.65.122
Date: Thu Feb 27 17:13:36 EST 2003

Input: 195 IDs
Search against: mgi

Unique Genes: 187
Annotated Genes: 173
GOs: 989
Unique GOs: 247
All Unique Sub-GOs: 471

GO-DB: mgi
Min Sub-GO length: 3
P-Value Cutoff: 0.1
GO-Cluster Cutoff: -1
Correct-Method: Benjamini

P-value cutoff: 0.1
Show best: 30
Indication: Over- and Underrepresented
Cluster GOs: -1
Display: HTML, Stats and Go Annotation

Best GOs (Max: 30)	Genes	Count 173	Total 9004	P-Value
GO:0007186	Chrm4 Hrh1 Hrh2 Htr1d Smstr1 Gpr37 Gnrhr Gprc1h Ednrb Bdkrb2 Gabrb1 Grb14 Npy1r Tacr1 Lhcgr Tacr2 Gabrb3 Tacr3 Oprk1 Edn1 Pomc1 Ntsr Npy5r Agtr2 Nmbr Brs3 Htr2c Agtrl1 Oprm Gabra1 Crhr Gabrq Gabra3 Crhr2 Ntsr2 Gabra4 Tac1 Gabrr2 Tac2 Npy2r Tac4 Agrp Adcyap1 Gabrg1 Chrm1 Gabrg2 Npy6r Gprc1a Gabrg3 Htr1a Chrm3	51	416	8.09e-52
GO:0007166	Chrm4 Il12rb2 Hrh1 Smstr1 Htr1d Hrh2 Gpr37 Gnrhr Insrr Il2rb Gprc1h Ednrb Bdkrb2 Il15ra [illegible] Ntsr Il11ra1 Il10ra Il10rb Npy5r Nmbr Agtr2 Brs3 Htr2c Oprm Agtrl1 Gabra1 Crhr Gabrq Il1r1 Gabra3 Crhr2 Gabra4 Ntsr2 Gabrr2 Tac1 Npy2r Tac2 Tac4 Agrp Adcyap1 Tnfrsf1b Gabrg1 Gabrg2 Chrm1 Npy6r Gabrg3 Gprc1a Chrm3 Htr1a	64	657	2.65e-48
GO:0004888	Ghr Chrm4 Il12rb2 Hrh1 Smstr1 Htr1d Hrh2 [illegible] Ednrb Bdkrb2 Il13ra Il4ra Il2rg Il6ra Gabrb1 Npy1r Tacr1 Gabrb3 Lhcgr Tacr2 Tacr3 Oprk1 Ntsr Il11ra1 Il10ra Il10rb Npy5r Lepr Nmbr Agtr2 Brs3 Htr2c Oprs1 Npr1 Oprm Agtrl1 Drd2 Prlr Gabra1 Crhr Gabrq Il1r1 Gabra3 Crhr2 Gabra4 Ntsr2 Gabrr2 Npy2r Grik1 Grik2 Gabrg1 Gabrg2 Chrm1 Npy6r Gabrg3 Gprc1a Chrm3 Grik5	66	826	4.15e-37

%molecular function%signal trancducer%receptor%transmembrane receptor

Gene	Search Term	GOs
Acas2	NM_080575	GO:0003824 GO:0003987 GO:0005737 GO:0006085 GO:0008152 GO:0016874
Ace	J04946	GO:0004180 GO:0004246 GO:0006508 GO:0008237 [illegible]
Ache	NM_009599	GO:0003824 GO:0003990 GO:0004104 GO:0004759 [illegible] GO:0016789 GO:0042135 GO:0045202
Adcyap1	NM_009625	GO:0005057 GO:0005179 GO:0005180 GO:0005184 GO:0005576 GO:0007190 GO:0007218 GO:0045786
Agrp	NM_007427	GO:0005184 GO:0007218 GO:0007582
Agt	NM_007428	GO:0004867 GO:0004868 GO:0008217
Agtr2	NM_007429	GO:0004872 GO:0004930 GO:0007186 GO:0016020 GO:0016021
Agtrl1	NM_011784	GO:0004930 GO:0007186 GO:0016020 GO:0016021
Arsa	AK004540	GO:0000299 GO:0004098

List of significant GO terms

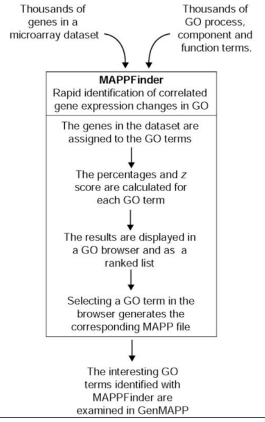
List of genes and GO associations

There are many similar tools

Here are a few.

GenMAPP, and MAPPFinder
EASE (DAVID)
FunSpec
FatiGO
…..

Outline of MAPPfinder:
MAPP = MicroArray Pathway Profiler



Thousands of genes in a microarray dataset

Thousands of GO process, component and function terms.

MAPPFinder
Rapid identification of correlated gene expression changes in GO

The genes in the dataset are assigned to the GO terms

The percentages and *z* score are calculated for each GO term

The results are displayed in a GO browser and as a ranked list

Selecting a GO term in the browser generates the corresponding MAPP file

The interesting GO terms identified with MAPPFinder are examined in GenMAPP

Analyzing microarray data by functional gene sets defined *a priori*

Analysis at the level of single gene:

- Identifying differentially expressed genes becomes a challenge when the magnitude of differential expression is small.
- For some differences, many genes are involved.

Analysis at the level of functional group: why?

By incorporating biological knowledge, we can hope to detect modest but coordinate expression changes of sets of functionally related genes.

PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes

Mootha *et al*, Nature Genetics July 2003

Data: Affymetrix microarray data on 22,000 genes in skeletal muscle biopsy samples from 43 males, 17 with normal glucose tolerance (NGT), 8 with impaired glucose tolerance and 18 with Type 2 diabetes (DM2).

In their single gene analysis, a *t*-statistic was calculated for each gene. No significant difference found between NTG and DM2 after adjusting for multiple testing.

Their idea: test 149 *a priori* defined gene sets for association with disease phenotypes.

149 gene sets

Sets of metabolic pathways:

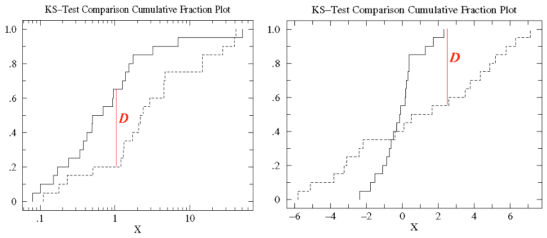
- manually curated pathways (standard textbook literature reviews, and LocusLink)
- Netaffx annotations using GenMAPP metabolic pathways

Sets of coregulated genes:

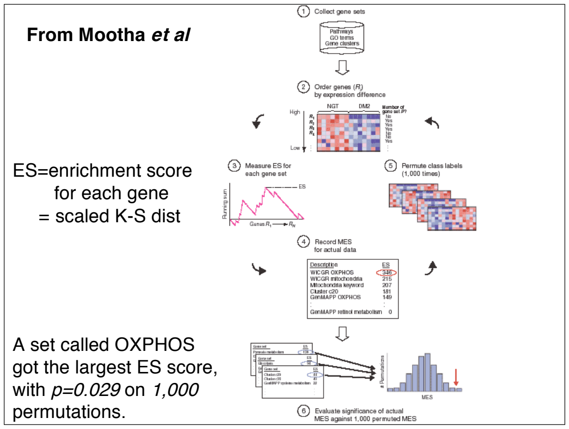
- SOM clustering of the mouse expression atlas

How do you compare group of interest to rest?

Two sample Kolmogorov-Smirnov test



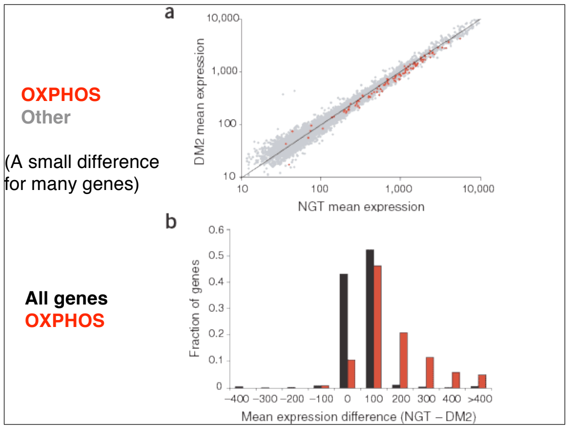
To compare two empirical cdfs, $S_M(x)$ and $S_N(x)$ based on samples of size M and N, resp, the Kolmogorov-Smirnov (K-S) test uses the K-S distance $D_{MN} = \max_x |S_M(x) - S_N(x)|$. This is normalized by multiplying by $\sqrt{(M^{-1} + N^{-1})}$. It has a complicated null distribution, which can be approximated by permuting.



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.



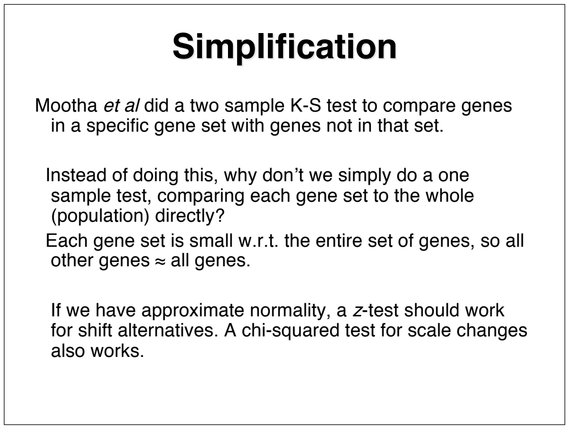
OXPHOS

Other

(A small difference for many genes)

All genes

OXPHOS



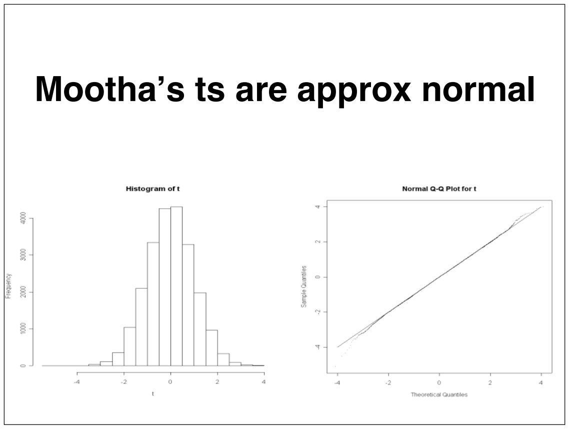
Simplification

Mootha *et al* did a two sample K-S test to compare genes in a specific gene set with genes not in that set.

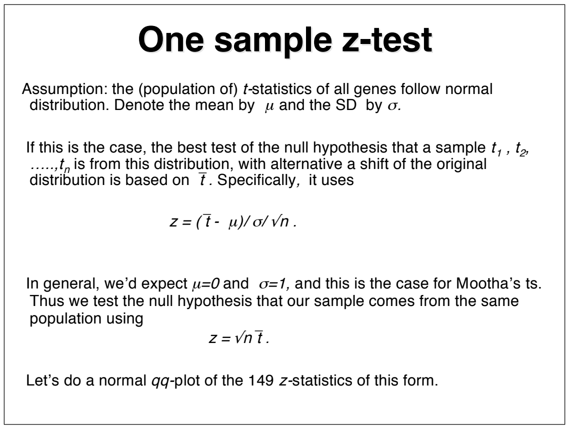
Instead of doing this, why don't we simply do a one sample test, comparing each gene set to the whole (population) directly?

Each gene set is small w.r.t. the entire set of genes, so all other genes ≈ all genes.

If we have approximate normality, a z-test should work for shift alternatives. A chi-squared test for scale changes also works.



Mootha's ts are approx normal



One sample z-test

Assumption: the (population of) *t*-statistics of all genes follow normal distribution. Denote the mean by μ and the SD by σ.

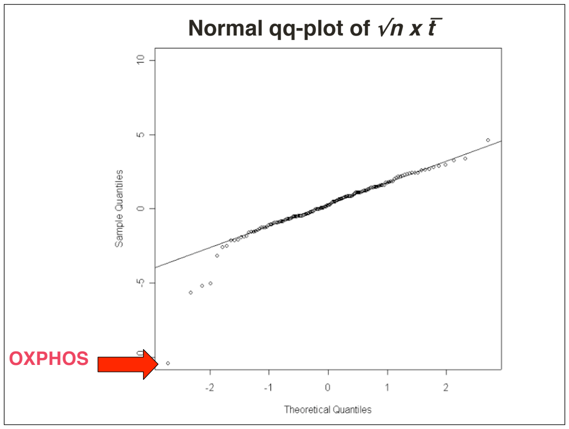
If this is the case, the best test of the null hypothesis that a sample t_1 , t_2, …..,t_n is from this distribution, with alternative a shift of the original distribution is based on $\bar{t}$. Specifically, it uses

$$z = (\bar{t} - \mu)/\sigma/\sqrt{n} .$$

In general, we'd expect $\mu=0$ and $\sigma=1$, and this is the case for Mootha's ts. Thus we test the null hypothesis that our sample comes from the same population using

$$z = \sqrt{n}\,\bar{t} .$$

Let's do a normal *qq*-plot of the 149 *z*-statistics of this form.



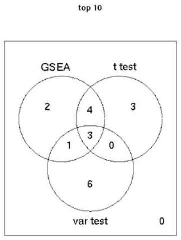
Normal qq-plot of $\sqrt{n}$ x $\bar{t}$

Result from one sample z-test

- OXPHOS is easily identified as ≈ -10.
- The next three sets on the top ranking list are all related to oxidative phosphorylation.

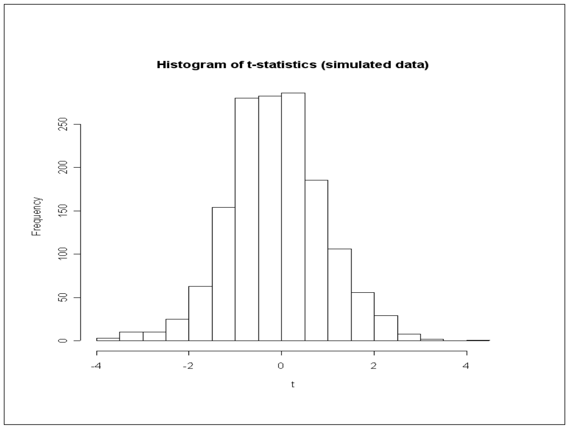
	z	n	# overlapping w/ OXPHOS
OXPHOS_HG-U133A_Probes	-10.4	114	114
Human_mitoDB_6_2002_HGU133A_probes	-5.6	594	106
Mitochondr_HG-U133A_probes	-5.2	615	103
MAP00190_Oxidative_phosphorylation	-5.0	75	29

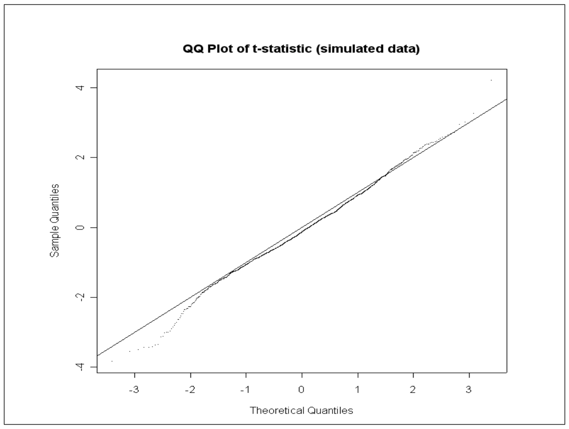
Similar Results

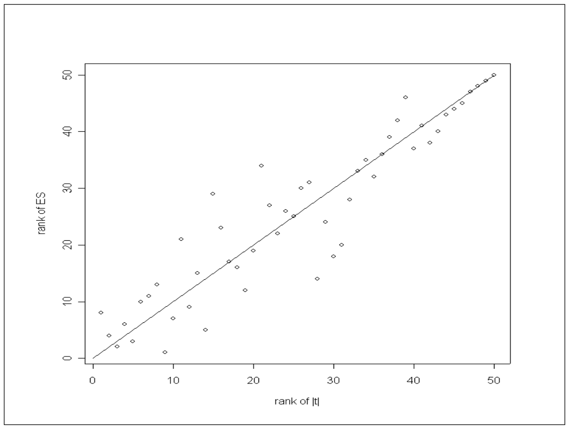


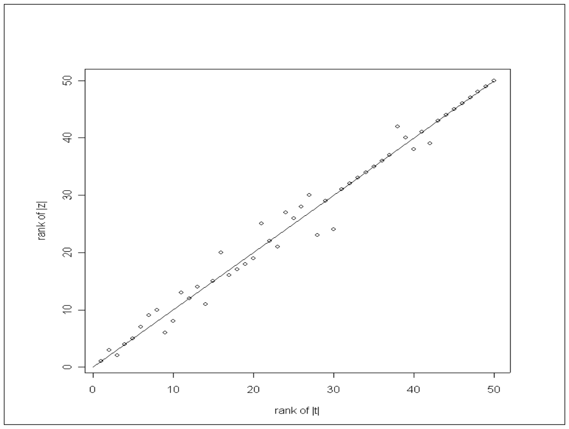
Simulation 1

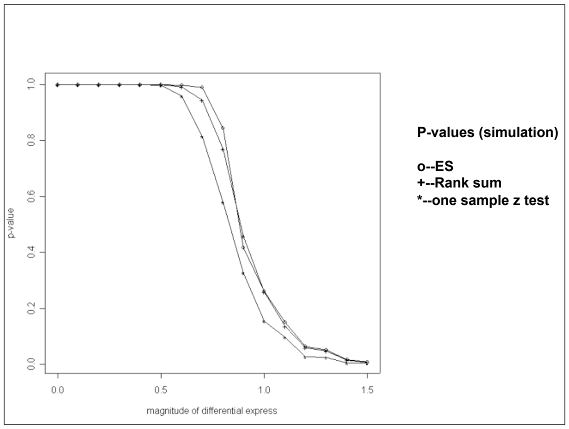
- 1500 × 29 gene expression values are generated from N(0,1), representing 1500 genes for 9 cases and 20 controls.
- The 1500 genes are divided into 50 gene sets, each with 30 genes. The genes are correlated within each gene set.
- We manipulate the gene expression level of the cases of the first gene set so that the magnitude of difference is known.


Histogram of t-statistics (simulated data)
Frequency
0
50
100
150
200
250
-4
-2
0
2
4
t


QQ Plot of t-statistic (simulated data)
Sample Quantiles
-4
-2
0
2
4
-3
-2
-1
0
1
2
3
Theoretical Quantiles


rank of ES
0
10
20
30
40
50
0
10
20
30
40
50
rank of |t|





P-values (simulation)

o--ES
+--Rank sum
***--one sample z test**

Conclusion

- When the population follows a normal distribution, the one-sample *z*-test is most powerful for shift alternatives (no surprise: theory says it has to be).
- From the simulation study, the one sample z-test is seen to be more powerful than the two sample *K-S* test for shift alternatives (even less of a surprise).
- The new method is not as compute intensive as the K-S test**.**
- Similar results can be given for the following test statistic, for scale change alternatives: for a set of n genes

$$z' = \sum_{i=1,..n} [(t_i - \bar{t})^2 - (n-1)] / \sqrt{(2(n-1))}.$$

(A test of no scale change might locate a set of genes that was split, with some having larger and others having smaller ts than average.)