Pathway and Gene Set Analyses

Self contained and competitive tests

What is a pathway?

No clear definition!

Wikipedia: "In biochemistry, **metabolic pathways** are series of chemical reactions occurring within a cell. In each pathway, a principal chemical is modified by chemical reactions." These pathways describe enzymes and metabolites.

But often the word "pathway" is also used to describe gene regulatory networks or protein interaction networks.

In all cases a pathway describes a biological function very specifically.

[LM]

What is a gene set?

Just what it says: a set of genes!

All genes involved in a pathway are an example of a gene set.
All genes corresponding to a Gene Ontology term are a gene set.
All genes mentioned in a paper of Smith et al might form a gene set.

A gene set is a much more general and less specific concept than a pathway.

Still: we will sometimes use two words interchangeably, as the analysis methods are mainly the same.

[LM]

Gene set / pathway analysis

The aim is to give one number (e.g. a p-value, a score) to a gene set / pathway:

- Are many genes in the pathway differentially expressed (up-regulated / down-regulated)?
- Can we give a number (a p-value) as the probability of observing changes of this magnitude (or larger) just by chance?

[LM]

Resources for sets of genes or proteins

www.geneontology.org

www.genome.jp/kegg

www.bioconductor.org/packages/release/data/annotation

www.broadinstitute.org/gsea/msigdb/collections.jsp

www.pantherdb.org/panther/ontologies.jsp

Some warnings

Pathway / gene set definitions in a database might differ from another databases.

The activity of the genes in a pathway is not necessarily a good measurement of the activity of the pathway.

Genes in a gene set are usually not given by a Probe Set ID, but refer to some gene data base. Entrez IDs, Unigene IDs, …

- Conversion can lead to errors!

There are **many** more resources out there. BioCarta, BioPax, NetPath, …

Commercial packages often use their own pathway / gene set definitions. Ingenuity, Metacore, Genomatix, …

Self contained versus competitive tests

The distinction between “self-contained” and “competitive” methods goes back to Goeman and Buehlman (2007).
PMID 17303618.

A **self-contained** method only uses the values for the genes of a gene set.

A **competitive method** compares the genes within the gene set with the other genes interrogated (array or sequencing).

Gene sets

Entrez IDs

```
> gsids[131:135]
$chryq11
 [1] 2260 3798 3963 3849 4862 4629 4826 4827 4629 4826 4827 4629 4826 4827 4629 4826 4827 4856 4862 4855 2611 4842 3963 4856 5968 3742 4842
[28] 4842 4842 4842 4842 4470

$chr15q13
 [1] 5667 7986 5667 1355 6917 6464 7436 1952 6623  881

$chr3q22
 [1] 5145 7328 7840 1042 5016 6193 1235 7016 5465 3678 3752 3160 1368 5574 7077

$chr6q15
[1] 8416 1607 8212 8082

$chr4q31
 [1] 2635 3295 4107 7827 6266 4581 5543 6509 5949 4313 2138 5751 2295 3258  118  481 2930 4522  812  118  481 1908 2863 7970 7084 5880 2785
[28] 2799 2907
```

Enrichment

Are genes on chromosome yq11 enriched?

	Not in gene set	In gene set
Not differentially expressed	8796	9
Differentially expressed	11	4

Self contained: Under the null, we would expect 13*0.05 < 1 gene to be significant. Using a Binomial(13,0.05) distribution, the probability that 4 or more genes are differentially expressed is 0.003.

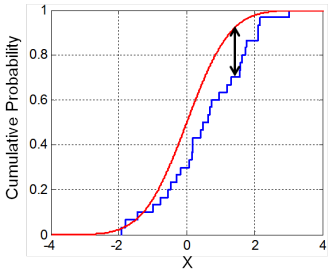
Competitive: Under the null, we would expect that virtually all differentially expressed genes are not in the gene set. Fisher's exact test gives a p-value of less than 1e-8.

Issues with significance cut-off

- A problem with both types of tests discussed so far is that they rely on an arbitrary cut-off for "significance".
- If we call a gene significant at a 1% or 10% p-value threshold, the results and the conclusions will change.
- We also lose information by reducing a p-value to a binary ("significant", "non-significant") variable.
- It should make a difference whether the non-significant genes in the set are "nearly significant" or "far from being significant".

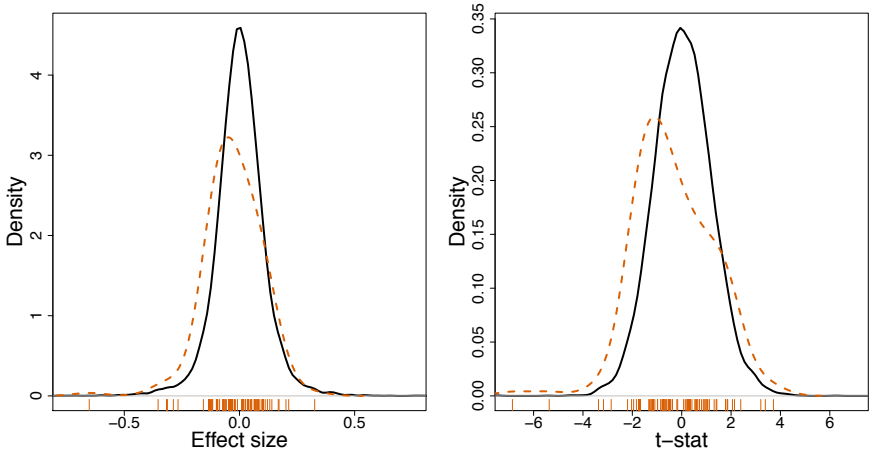
Possible approaches

- Use Fisher's inverse χ^2 test statistic: -2*sum(log (p_i)) [S] .
- Use the Kolmogorov-Smirnov test to check if the set of p-values departs from a Uniform distribution [S].



- Alternatively you could look at the distribution of the **ranks** of the p-values in the gene set [C]. Again one can use the Kolmogorov-Smirnov test to test for uniformity.
- Many more approaches exist, for example based on test statistics, fold changes, etc.

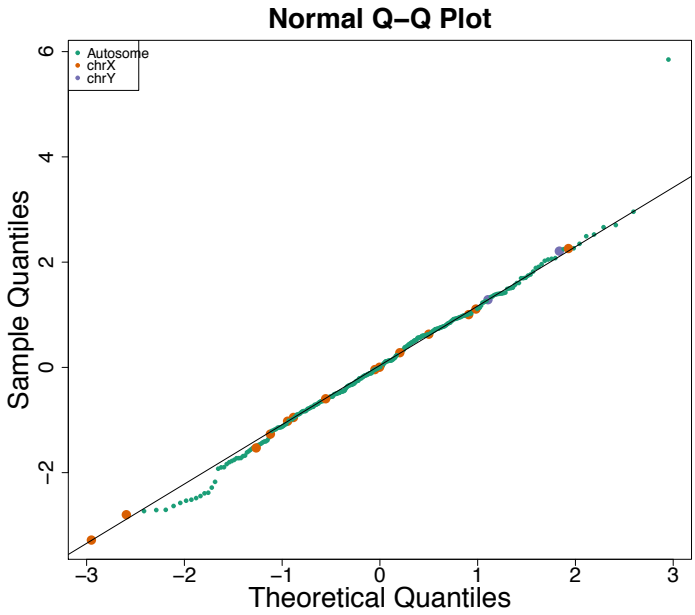
Distribution of the chrxp11 gene set



Few genes are significant, but a mean shift is observed.

[RI]

Wilcoxon test for differential expression



A simple test for a mean shift

The simplest statistic to test for a mean shift is the average difference in mean. One way to summarize this difference is to average the t-statistics.

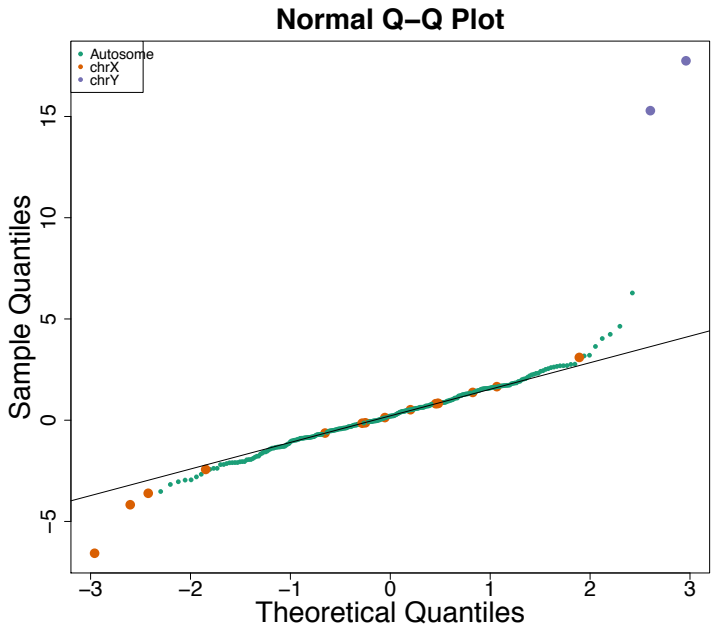
$$\bar{t} = \frac{1}{N} \sum_{i \in G} t_i$$ (with N being the size of gene set G).

Under the null, the t statistics have mean 0 and SD 1.

If they are independent, then:

$$\sqrt{N}\ \bar{t} \sim N(0,1)$$

Average t-statistics



[RI]

A big issue

The **correlation** of expression between genes.

- All tests we discussed so far assumed that the statistics within the gene set(s) are independent. That is highly unlikely!
- If genes are correlated, the p-values of the gene set tests (e.g. from Fisher's exact test) will be incorrect.
- This can be addressed for example by resampling methods:
  - Shuffle the group labels.
  - Repeat the analysis.
  - Compare the re-shuffled with the observed data.
- Note: shuffle the labels, not the genes!

Variance of average of correlated variables

$$\mathrm{var}(\bar{t}) = \frac{1}{N^2}\mathrm{var}\{(1\ldots1)(t_1\ldots t_N)'\}$$

$$= \frac{1}{N^2}(1\ldots1)\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \ldots & \rho & 1 \end{pmatrix}(1\ldots1)'$$

$$= \frac{1}{N}\{1 + (N-1)\rho\}$$

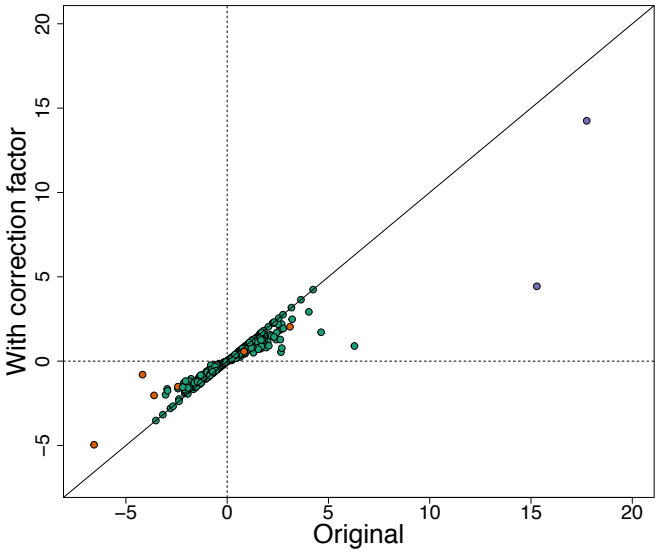
Correction factor

$$\frac{\sqrt{N}}{\sqrt{1 + (N-1)\bar{r}}}\,\bar{t}$$

Here, $\bar{r}$ is the average pairwise correlation within gene set.

Note that there are also plenty of other approaches for dealing with correlation!

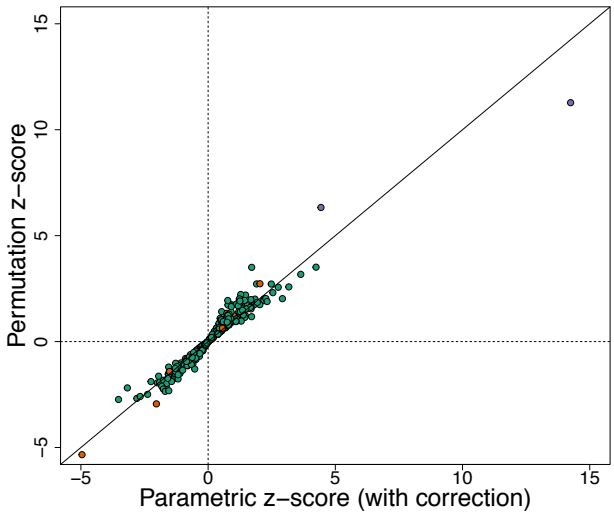
After parametric correction



[RI]

Permutation z-scores

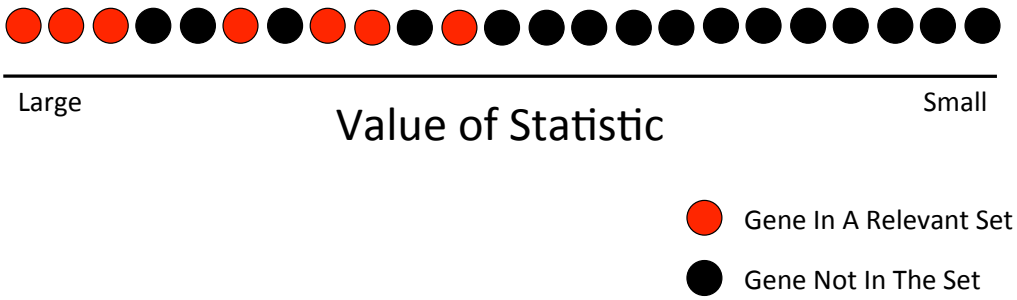
Even better: use permutations to construct gene set specific null distributions!



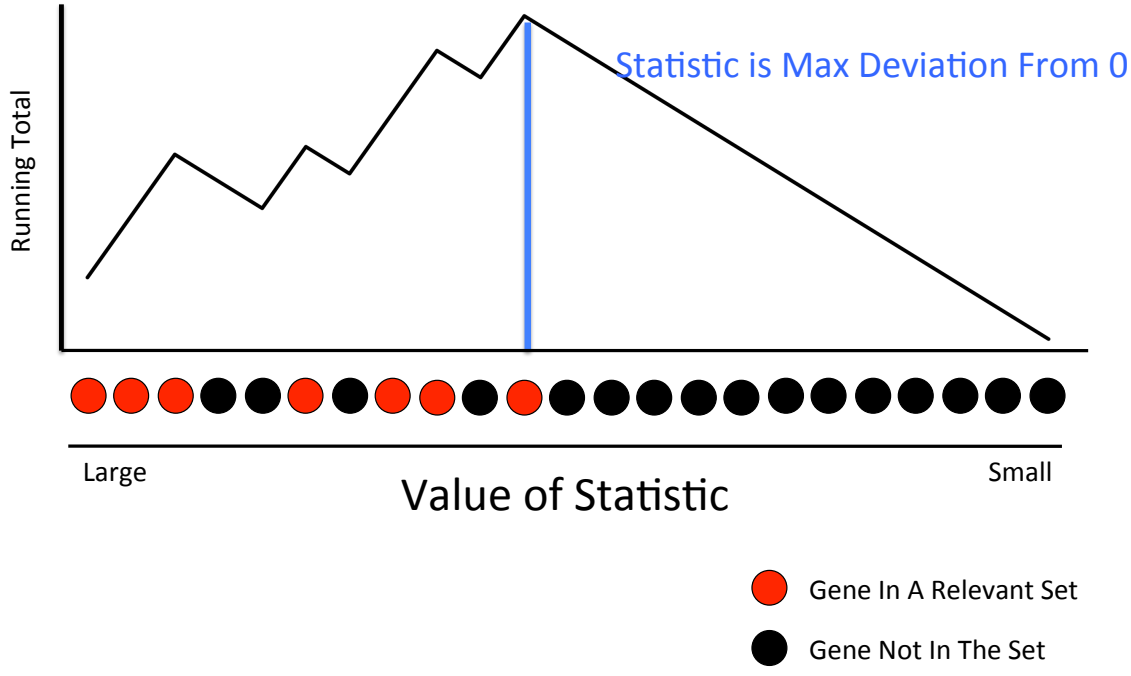
[RI]

Gene Set Enrichment Analysis

GSEA – a weighted version of Kolmogorov-Smirnov tests.



Gene Set Enrichment Analysis



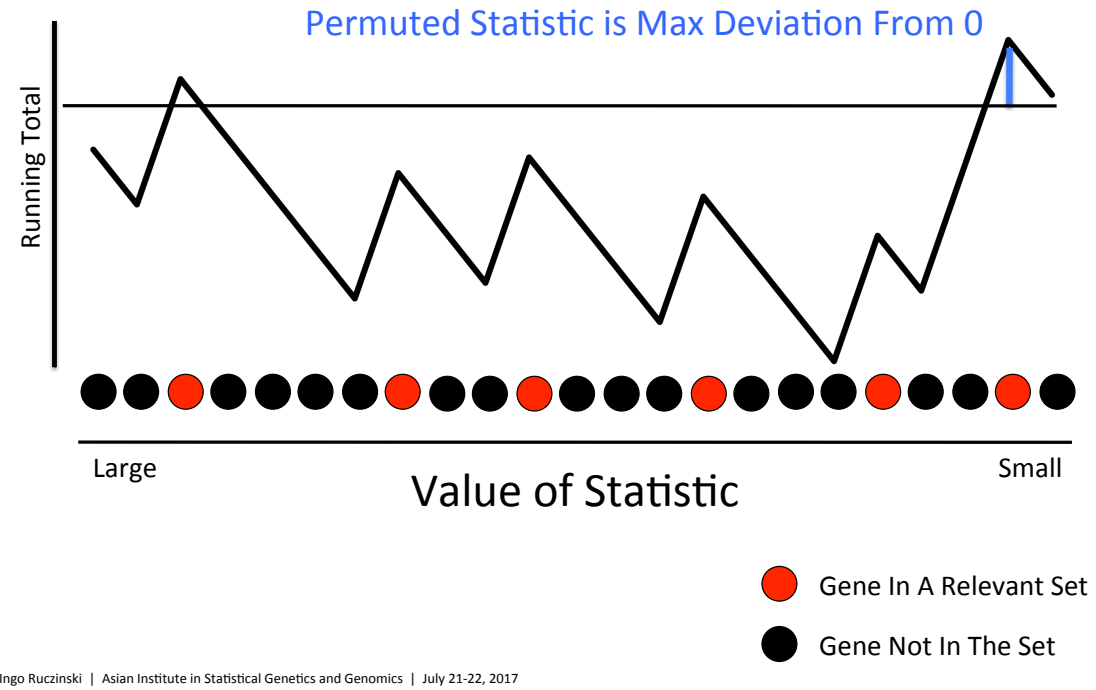
Permutation tests

Response	R	R	•••	NR	NR
	Patient 1	Patient 2	•••	Patient n-1	Patient n
Gene 1	-1.64	-0.42	•••	-1.39	-0.38
Gene 2	-3.12	-3.60	•••	-3.80	-2.82
:	:	:	•••	:	:
:	:	:		:	:
:	:	:	•••	:	:
:	:	:		:	:
Gene m-1	-2.34	-0.22	•••	-1.22	-2.76
Gene m	4.53	3.23	•••	0.29	3.11

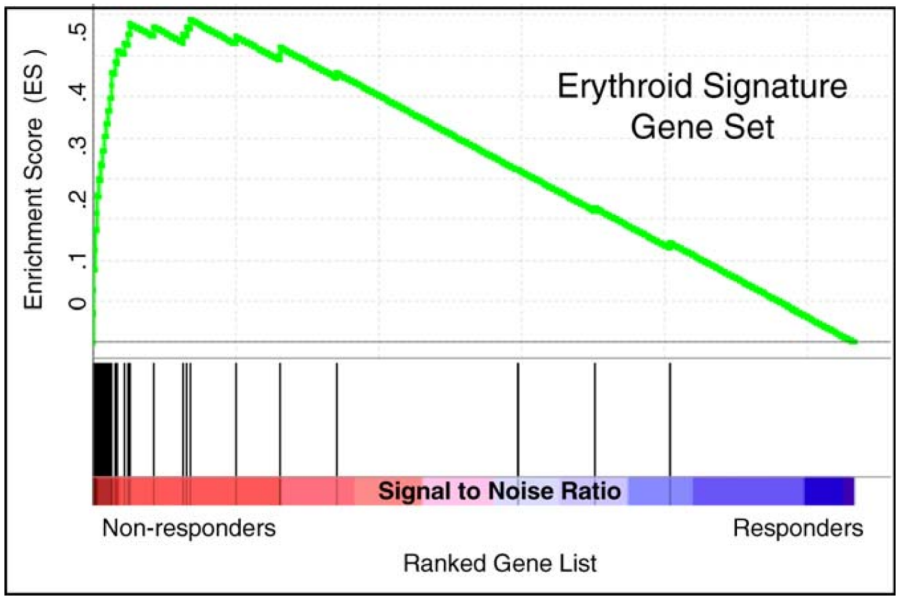
Response	NR	R	•••	NR	R
	Patient 1	Patient 2	•••	Patient n-1	Patient n
Gene 1	-1.64	-0.42	•••	-1.39	-0.38
Gene 2	-3.12	-3.60	•••	-3.80	-2.82
:	:	:	•••	:	:
:	:	:		:	:
:	:	:	•••	:	:
:	:	:		:	:
Gene m-1	-2.34	-0.22	•••	-1.22	-2.76
Gene m	4.53	3.23	•••	0.29	3.11

This leaves the relationship between the genes unchanged.

Permuted GSEA Profile



Gene set enrichment results



PMID 18271621

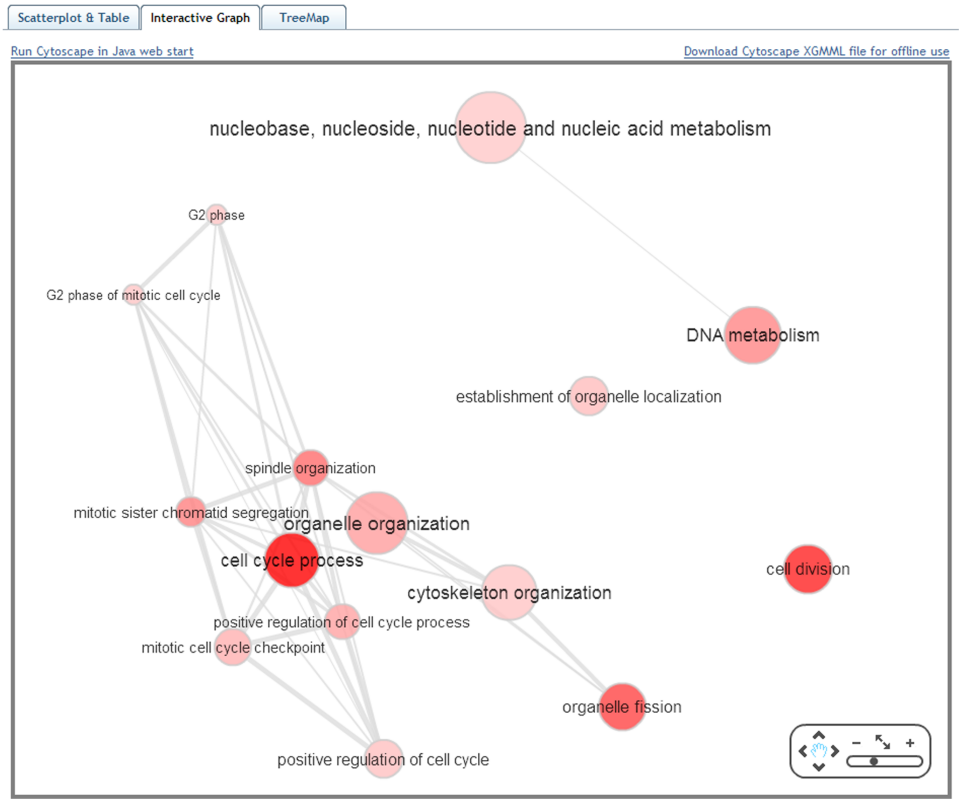
ANALYSIS

The statistical properties of gene-set analysis

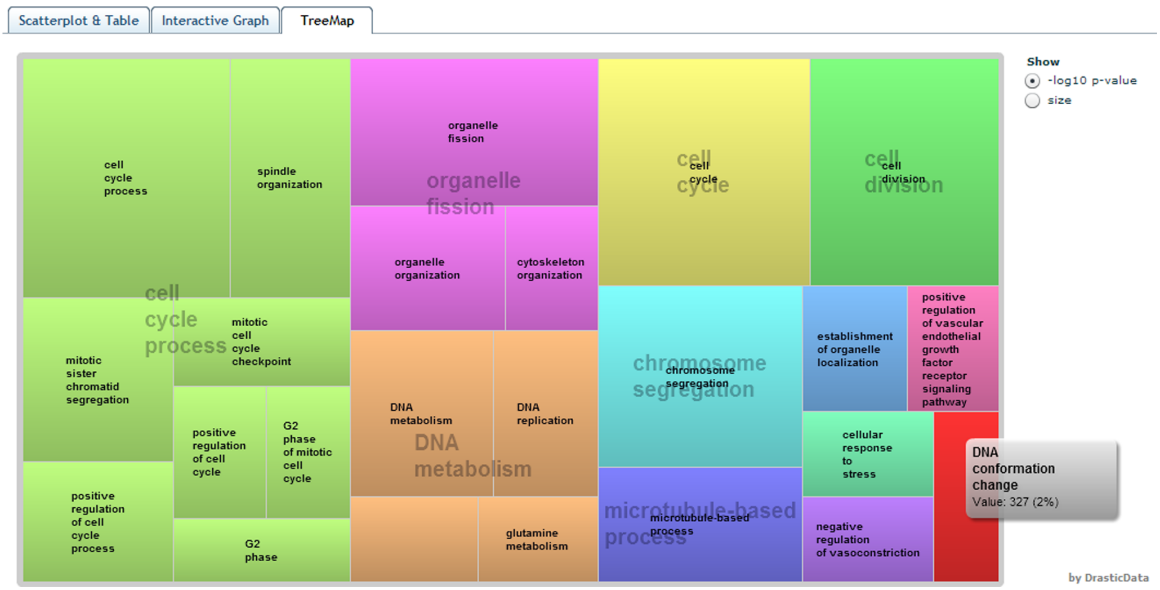
Christiaan A. de Leeuw[1,2], Benjamin M. Neale[3,4], Tom Heskes[2] and Danielle Posthuma[1,5]

Abstract | The rapid increase in loci discovered in genome-wide association studies has created a need to understand the biological implications of these results. Gene-set analysis provides a means of gaining such understanding, but the statistical properties of gene-set analysis are not well understood, which compromises our ability to interpret its results. In this Analysis article, we provide an extensive statistical evaluation of the core structure that is inherent to all gene-set analyses and we examine current implementations in available tools. We show which factors affect valid and successful detection of gene sets and which provide a solid foundation for performing and interpreting gene-set analysis.

PMID 27070863



PMID 21789182



PMID 21789182