

Testing groups of genes

Jelle Goeman

Department of Medical Statistics
Leiden University Medical Center

MGC array course

22 June 2011

Beyond differentially expressed genes

The typical result of a microarray experiment

A list of differentially expressed genes

Biological theory is not about isolated genes

Typical biological research questions and hypotheses

- About pathways
- About biological processes
- About areas of the genome

About **sets** of related genes

Question

How to analyze microarray data from a gene set perspective?

A gene list

	P-value	FDR.adjusted
ZYX	1.6248e-11	1.1583e-07
SRGN	4.9409e-10	1.2723e-06
CD33	5.3541e-10	1.2723e-06
CFD	1.2903e-09	2.1064e-06
APLP2	1.7072e-09	2.1064e-06
MGST1	2.0128e-09	2.1064e-06
CST3	2.1556e-09	2.1064e-06
LYN	2.3637e-09	2.1064e-06
CTSA	3.8935e-09	2.9511e-06
CTSD	4.1395e-09	2.9511e-06
FAH	6.1026e-09	3.9550e-06
ATP6VOC	8.8919e-09	5.2825e-06
LTC4S	1.4490e-08	7.9459e-06
RHOG	1.7997e-08	9.1644e-06
RNASE2	1.9911e-08	9.4629e-06
FTL	2.3560e-08	1.0497e-05
TIMP2	2.8443e-08	1.1928e-05
SPI1	4.8150e-08	1.9070e-05
IL18	5.0842e-08	1.9076e-05

A gene list: Gene Ontology Terms involved

ZYX	cell adhesion, signal transduction, cell-cell signaling , interspecies interaction between organisms, p
SRGN	ossification, apoptosis, negative regulation of bone mineralization, mast cell secretory granule organ
CD33	cell adhesion, signal transduction, cell-cell signaling , negative regulation of cell proliferation, pl
CFD	proteolysis , complement activation, alternative pathway, complement activation, alternative pathway, e
APLP2	G-protein coupled receptor protein signaling pathway, plasma membrane, integral to membrane, nucleus,
MGST1	glutathione metabolic process, membrane, mitochondrion, endoplasmic reticulum, mitochondrial outer mem
CST3	extracellular region, cysteine protease inhibitor activity, protein homodimerization activity
LYN	protein amino acid phosphorylation, intracellular signaling cascade, positive regulation of cell proli
CTSA	proteolysis , intracellular protein transport, lysosome, endoplasmic reticulum, protein binding, serine
CTSD	proteolysis , lysosome, mitochondrion, extracellular region, melanosome, peptidase activity
FAH	metabolic process, arginine catabolic process, L-phenylalanine catabolic process, tyrosine catabolic p
ATP6VOC	ATP biosynthetic process, ion transport, proton transport, interspecies interaction between organisms,
LTC4S	leukotriene biosynthetic process, membrane, integral to membrane, membrane fraction, microsome, glutat
RHOG	small GTPase mediated signal transduction, Rho protein signal transduction, positive regulation of cel
RNASE2	RNA catabolic process, chemotaxis, lysosome, extracellular region, nucleic acid binding, endonuclease
FTL	iron ion transport, cellular iron ion homeostasis, cellular iron ion homeostasis, ferritin complex, bi
TIMP2	negative regulation of cell proliferation, regulation of cAMP metabolic process, regulation of MAPKKK
SPI1	negative regulation of transcription from RNA polymerase II promoter, transcription, regulation of tra
IL18	angiogenesis, response to hypoxia, immune response, cell-cell signaling , response to cold, regulation

Gene set testing: aims

More directly interpretable results

Study relevant processes directly, not via single genes

Reduce multiple testing load

Fewer gene sets than genes

Introduce biological knowledge in the analysis

Make use of knowledge in pathway databases to improve power

Find gene sets with consistent but small effects

Of which no single gene shows up in a single gene analysis

More reproducible results

Also across platforms

What gene sets?

Any externally defined set that has something in common

Pathways

KEGG, Biocarta

Gene Ontology terms

Biological process, Molecular function, Cellular component

Chromosomal regions

Chromosome arms, cytobands, linkage peaks, genes

Published gene sets

Predictive signatures, gene lists

Two types of gene set testing methods

Enrichment methods

- Test gene sets as a secondary step after single gene testing
- Are gene sets enriched with diff. expr. genes?
- Example: 2×2 table methods

Aggregate score methods

- Test gene sets in a single step
- Is there any differential expression in the gene set?
- Example: global test

Enrichment: general idea

Primary analysis

- First do a single gene analysis
- Gather the results (gene list, p-values, fold change. . .)

Secondary (post hoc) analysis

- Input: the results of the primary analysis
- Do genes in the gene set tend to have low p-values?
- Do the low p-values tend to belong to the same gene set?

Compare

Searching for “themes” in the gene list

Fisher's exact test: practicalities

A 2×2 table

Make a 2×2 table:

	<i>diff. expr. gene</i>	<i>non-diff. expr. gene</i>
<i>in gene set</i>		
<i>not in gene set</i>		

Test row/column association with Fisher's exact test

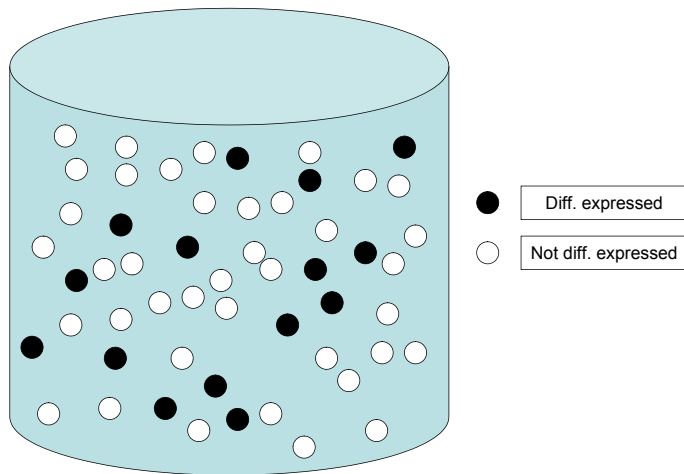
Equivalent variants

χ^2 test; hypergeometric test; binomial z-test

Overview

Khatri and Dhragici, *Bioinformatics*, 2005

Urn model: randomly draw genes from an urn



Fisher's exact test: example

Microarray: 20000 genes; gene set: 100 genes; diff. expr: 200

Could arise due to chance: $p = 0.26$

	<i>diff. expr. gene</i>	<i>non-diff. expr. gene</i>	<i>total</i>
<i>in gene set</i>	2	98	100
<i>not in gene set</i>	198	19702	19900
<i>total</i>	200	19800	20000

Could not arise due to chance: $p = 0.0005$

	<i>diff. expr. gene</i>	<i>non-diff. expr. gene</i>	<i>total</i>
<i>in gene set</i>	6	94	100
<i>not in gene set</i>	194	21706	19900
<i>total</i>	200	19800	20000

Why the urn model is problematic

Null hypothesis

Genes in the gene set are just a random draw from the urn of genes

Meaning of a significant result

Genes in the gene set are more alike than randomly drawn genes

Genes in a gene set are always similar

Because they are selected to have something in common

Consequence: false positives

P-values are only correct if all genes are independent

Remedy the problems of the urn model

A permutation

Randomly reassign group labels to array data

Permutation p-value

- Permute many times
- Calculate enrichment score for each permutation
- Permutation P-value
= percentage of permuted scores better than true score

Why?

- Under permutations no gene is differentially expressed
- But correlations between genes are retained

Permutation testing

True data

	0	0	0	1	1	1
$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	$X_{1,5}$	$X_{1,6}$	
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	
$X_{p,1}$	$X_{p,2}$	$X_{p,3}$	$X_{p,4}$	$X_{p,5}$	$X_{p,6}$	

Random permutation of sample labels

	1	0	1	1	0	0
$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	$X_{1,5}$	$X_{1,6}$	
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	
$X_{p,1}$	$X_{p,2}$	$X_{p,3}$	$X_{p,4}$	$X_{p,5}$	$X_{p,6}$	

Urn model methods

Many gene set methods are based on the urn model

How to recognize urn model methods?

- They only require a single value per gene as input
- They do not require the complete data set

Avoid using urn model methods

- Use only when you have no access to the full data
- Beware: many false positives
- Do not trust the p-values

Globaltest: general idea

Direct gene set testing

No need for first testing single genes

Basic idea

Based on a regression model: predict group label (response) from expressions

Null hypothesis

Group label cannot be predicted from the gene expressions

Alternative

Group label can (partly) be predicted from the gene expressions

The Global Test model

Question

Can the genes in the gene set predict the response?

In a nutshell

- A response y to be predicted
- Gene expressions = a set of covariates (x_1, \dots, x_p)
- Clinical covariates: (z_1, \dots, z_m)
- Generalized linear model or Cox model
- Linear predictor: $\alpha + \sum_{j=1}^m z_j \gamma_j + \sum_{i=1}^p x_i \beta_i$
- Null hypothesis of no predictive ability:

$$H_0 : \beta_1 = \dots = \beta_p = 0$$

Power properties

Score test

Directed at alternatives close to the null hypothesis

Consequence

Good at detecting gene sets with many small effects

Application: gene expression

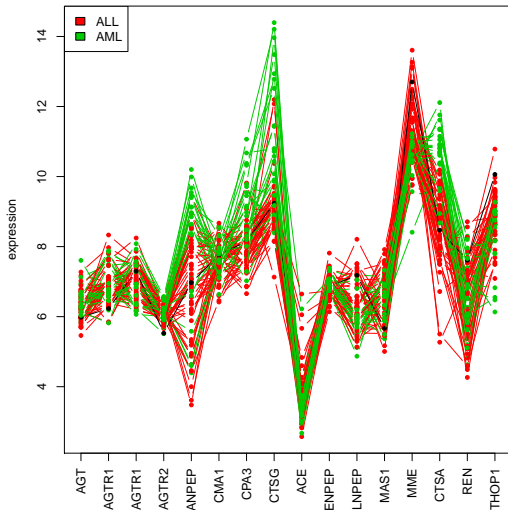
AML/ALL data

Famous old data set of Golub et al. (Science 1999)

Details

- Response: AML vs. ALL
- $n = 72$ patients (25 AML; 47 ALL)
- 7129 genes
- Affymetrix hu6800 chips
- Pathway of interest (KEGG):
Renin-angiotensin system (16 probe sets)

Pathway Profile: prediction



GlobalTest: Genes interpretation

Pathway is associated with the response

Many genes in the pathway are associated with the response

Maybe only slightly

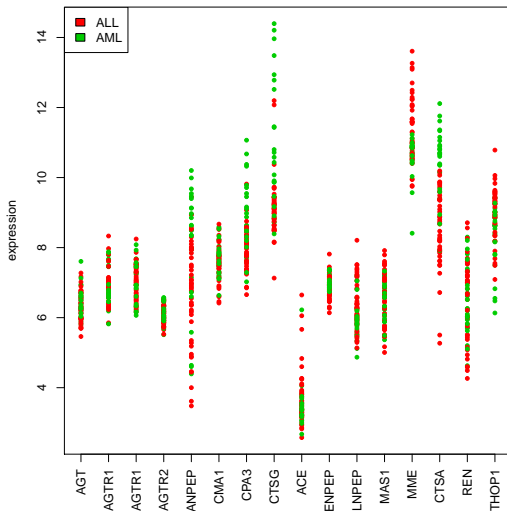
Not enough to be significant individually

But together they may show a consistent pattern

Significant Global Test result:

- 'On average' the genes are associated with the response
- Pathway usually a mix of up- and downregulated genes
- Not every gene needs to be associated

Pathway Profile: prediction



Features plot

Decomposition of the test: contribution per gene

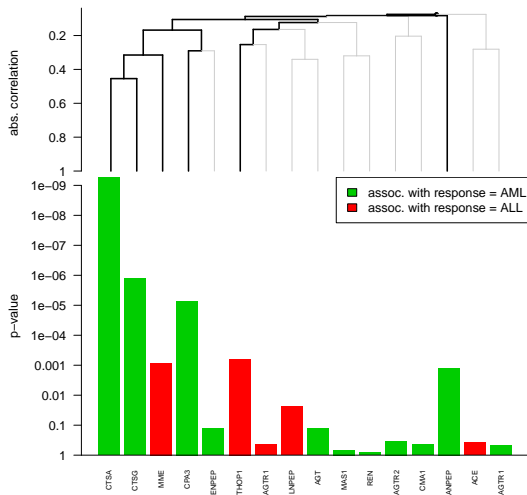
Shows three things at once

- P-values: association of the gene with the response
- Direction of association of gene with response
- Correlations among genes

Multiple testing method based on clustering graph

Where to attribute the significant result of global test?

Features Plot: look at influential genes



GlobalTest: Samples interpretation

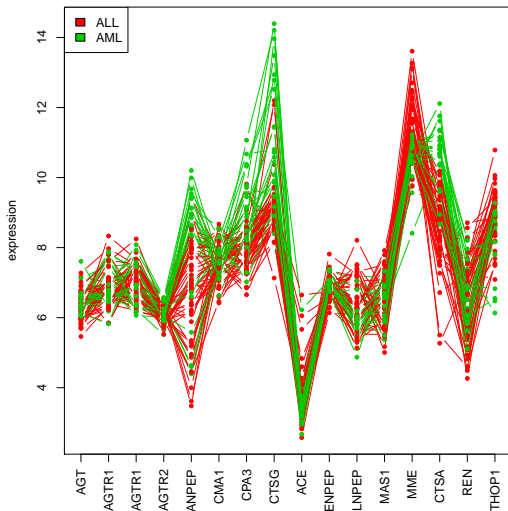
Pathway is associated with the response

- The pathway expression profile differs for different values of the response
- Samples with similar response have relatively similar expression profiles
- Samples with different response have relatively dissimilar expression profiles

Consequence

Samples with the same condition tend to cluster together in a cluster analysis based on only the genes in the pathway

Pathway profiles



Subjects plot

Decomposition of the test: contribution per subject

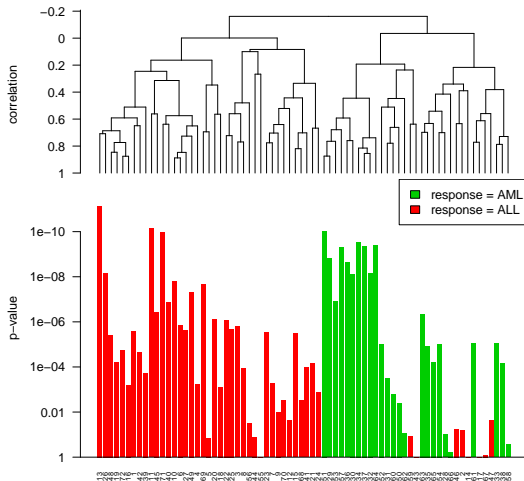
Shows two things at once

- P-values: see below
- Correlations between subjects

P-value

Null hypothesis: subjects with the same label are as similar to this subject as subjects with opposite labels

Subjects plot



Is the test significant?

Null hypothesis

Expression profile of KEGG *Renin-angiotensin system* is not associated with the response AML/ALL

Result

- P-value: 2×10^{-15}
- Null hypothesis rejected
- KEGG Renin-angiotensin system clearly associated with AML/ALL
- As could be seen from the graphs

Dutch Cancer Institute—breast cancer data

Breast cancer data of the Netherlands Cancer Institute

- Paper by Van 't Veer *et al.* (*Nature*, 2002)
- Followed up by Van de Vijver *et al.* (*NEJM*, 2002)
- 295 breast cancer patients
- Response: survival (up to 16 years follow-up)
- Microarray: 4,919 genes preselected (Rosetta technology)

Covariates

- Tumor size, grade, lymph node status, estrogen receptor status
- Age
- Treatment: mastectomy, chemotherapy, hormonal therapy

Global association of gene expression with survival

Association of gene expression with survival

Without looking at the covariates:

	genes	Statistic Z	p-value
All genes	4919	4.64	0.000002

Conclusions

- Patient survival associated with gene expression profile
- Potential in predicting survival from gene expression
- Patients with similar gene expression profiles tend to have similar survival times
- Many individual genes associated with survival

Mining Gene Ontology

Use Global Testing to mine Gene Ontology

- Which GO terms can most clearly be used to predict survival?
- Which GO terms' expression profile is associated with survival?

Disadvantage of mining: highly undirected search

Consequence: heavy multiple testing penalty

Better: a small collection of well-chosen candidate gene sets

Mining the breast cancer data

We tested all 4045 biological process GO terms

Association with survival

	genes	FDR-adj. p-value
establishment of organelle localization	8	0.0000007
mitotic chromosome condensation	7	0.0000007
cytokinesis	13	0.0000007
pos. reg. of progr. thr. cell cycle	9	0.0000007
chromosome segregation	21	0.0000007
regulation of caspase activity	12	0.0000007
chromosome condensation	8	0.0000007
pos. reg. of progr. thr. mitotic cell cycle	3	0.0000007
sister chromatid segregation	15	0.0000007
mitotic sister chromatid segregation	15	0.0000007
organelle localization	10	0.0000007
regulation of hydrolase activity	19	0.0000007
pos. reg. of exit from mitosis	2	0.0000008
G2/M transition of mitotic cell cycle	6	0.0000008
DNA replication	58	0.0000010
regulation of exit from mitosis	3	0.0000010

The importance of covariates

Multiple risk factors

- Look at the expression signature as a risk factor
- One among many others
Most clinical risk factors are much cheaper to measure
- Imagine: the microarray as an expensive way of measuring lymph node status

Confounders

- Many microarray studies for prediction are observational
- Population variables may be different between study groups
Gene expression relates strongly to population variables
- Imagine: the microarray signature turns out to predict age

Global association with covariates

Association of gene expression with survival

Adjusting for Age, tumor covariates, treatment covariates

	genes	Statistic Z	p-value
All genes	4919	1.84	0.033

Conclusions (much less clear)

- Some additional potential for predicting survival
- Patients with similar gene expression profiles have slightly similar residuals from the clinical model
- Some individual genes associated with survival after correcting for covariates

Association with survival (with covariates)

	genes	FDR-adj. p-value
cytokinesis	13	0.034
pos. reg. of progr. thr. mitotic cell cycle	3	0.034
establishment of organelle localization	8	0.034
pos. reg. of progr. thr. cell cycle	9	0.039
spindle localization	3	0.043
establ. of spindle localization	3	0.043
establ. of mitotic spindle localization	3	0.043
organelle localization	10	0.051
chromosome localization	2	0.051
establ. of chromosome localization	2	0.051
protein complex localization	2	0.057
pos. reg. of exit from mitosis	2	0.057
regulation of exit from mitosis	3	0.067
cytokinesis after mitosis	2	0.073
regulation of caspase activity	12	0.078
chloride ion homeostasis	1	0.078

Association with covariates (e.g.: tumor diameter)

	genes	FDR-adj. p-value
mitotic spindle elongation	2	0.0007
spindle elongation	2	0.0007
mitotic spindle organization and biogenesis	7	0.0007
anterior/posterior pattern formation	16	0.0007
lipid catabolism	25	0.0007
positive regulation of glycolysis	1	0.0007
pentose-phosphate shunt	3	0.0007
NADPH regeneration	3	0.0007
mitotic checkpoint	8	0.0008
DNA damage checkpoint	9	0.0008
G2/M transition DNA damage checkpoint	4	0.0008
G2/M transition checkpoint	4	0.0008
mitotic G2 checkpoint	4	0.0008
spindle organization and biogenesis	14	0.0008
DNA damage response, signal transduction	11	0.0009
cell division	78	0.0009

Recommendations

Gene set testing

- More power than single gene analysis
- Direct statements about relevant pathways

Methodological issues

- Avoid urn-model based methods (enrichment)
- Think about confounders
- Avoid testing too many pathways