Gergely Csaba
LFE Bioinformatik
Institut für Informatik
Ludwig-Maximilians-Universität München

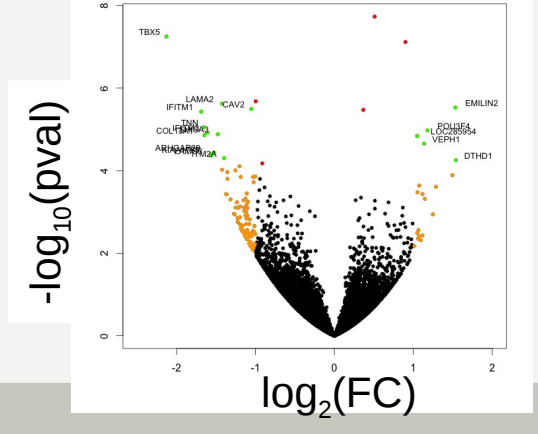
Gene Set Enrichment

GoBi WS 2019/20

Result of gene-based DE tests

RNAseq → mapping → feature counting → differential expression / splicing →

	condition 1 (phenotype 1)	condition 2 (phenotype 2)	log(FC)	pval,FDR
	rep_1 rep_2 ... rep_{n1}	rep_1 rep_2 ... rep_{n2}		
gene 1	counts (or other signals if not RNA-seq)	counts (or other signals if not RNA-seq)	FC 1	FDR 1
gene 2			FC 2	FDR 2
...			...	...
gene N			FC N	FDR N

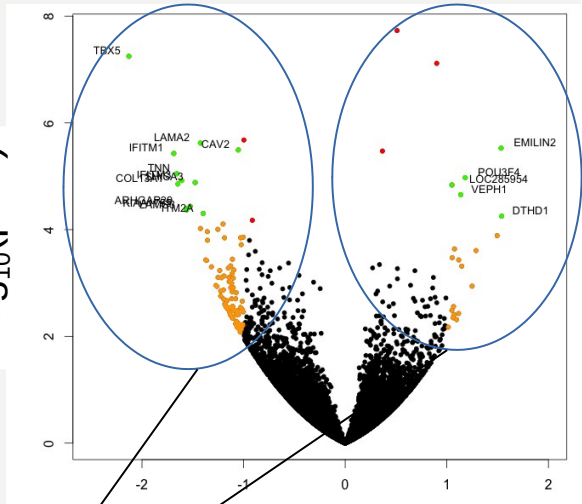


Result of gene-based DE tests

RNAseq → mapping → feature counting → differential expression / splicing →

condition 1 (phenotype 1) condition 2 (phenotype 2) log(FC) pval,FDR

	rep_1 rep_2 ... rep_{n1}	rep_1 rep_2 ... rep_{n2}
gene 1	counts	counts
gene 2	(or other signals if not RNA-seq)	(or other signals if not RNA-seq)
...		
gene N		



typical definition:
differentially expressed (DE) genes (**DEG-s**) :=
high abs(log(FC)) and small p-value

DEG-s

Result of gene-based DE tests

Beyond single gene analysis: what is the "drive" of the changes?

Input: long list of DEG-s

OFFICIAL_GENE_SYMBOL	Gene Name
AANAT	aralkylamine N-acetyltransferase(AANAT)
ABCB4	ATP binding cassette subfamily B member 4(ABCB4)
ABCC2	ATP binding cassette subfamily C member 2(ABCC2)
ABHD2	abhydrolase domain containing 2(ABHD2)
ABR	active BCR-related(ABR)
ACVR1C	activin A receptor type 1C(ACVR1C)
ACVRL1	activin A receptor like type 1(ACVRL1)
ADA	adenosine deaminase(ADA)
ADAM15	ADAM metallopeptidase domain 15(ADAM15)
ADAMTS1	ADAM metallopeptidase with thrombospondin type 1 motif 1(ADAMTS1)
ADCYAP1	adenylate cyclase activating polypeptide 1(ADCYAP1)
ADGRG1	adhesion G protein-coupled receptor G1(ADGRG1)
ADGRL1	adhesion G protein-coupled receptor L1(ADGRL1)
ADIPOQ	adiponectin, C1Q and collagen domain containing(ADIPOQ)
ADIPOR1	adiponectin receptor 1(ADIPOR1)
ADNP	activity dependent neuroprotector homeobox(ADNP)
ADORA1	adenosine A1 receptor(ADORA1)
ADORA2A	adenosine A2a receptor(ADORA2A)
AFP	alpha fetoprotein(AFP)
AGER	advanced glycosylation end-product specific receptor(AGER)
AGO4	argonaute 4, RISC catalytic component(AGO4)
AGRP	agouti related neuropeptide(AGRP)
AGTR2	angiotensin II receptor type 2(AGTR2)
AHCY	adenosylhomocysteinase(AHCY)
AHR	aryl hydrocarbon receptor(AHR)
AIF1	allograft inflammatory factor 1(AIF1)
ALB	albumin(ALB)
ALOX15B	arachidonate 15-lipoxygenase, type B(ALOX15B)
ALOX5AP	arachidonate 5-lipoxygenase activating protein(ALOX5AP)

Result of gene-based DE tests

Beyond single gene analysis: what is the "drive" of the changes?

Input: long list of DEG-s

OFFICIAL_GENE_SYMBOL	Gene Name
AANAT	aralkylamine N-acetyltransferase(AANAT)
ABCB4	ATP binding cassette subfamily B member 4(ABCB4)
ABCC2	ATP binding cassette subfamily C member 2(ABCC2)
ABHD2	abhydrolase domain containing 2(ABHD2)
ABR	active BCR-related(ABR)
ACVR1C	activin A receptor type 1C(ACVR1C)
ACVRL1	activin A receptor like type 1(ACVRL1)
ADA	adenosine deaminase(ADA)
ADAM15	ADAM metallopeptidase domain 15(ADAM15)
ADAMTS1	ADAM metallopeptidase with thrombospondin type 1 motif 1(ADAMTS1)
ADCYAP1	adenylate cyclase activating polypeptide 1(ADCYAP1)
ADGRG1	adhesion G protein-coupled receptor G1(ADGRG1)
ADGRL1	adhesion G protein-coupled receptor L1(ADGRL1)
ADIPOQ	adiponectin, C1Q and collagen domain containing(ADIPOQ)
ADIPOR1	adiponectin receptor 1(ADIPOR1)
ADNP	activity dependent neuroprotector homeobox(ADNP)
ADORA1	adenosine A1 receptor(ADORA1)
ADORA2A	adenosine A2a receptor(ADORA2A)
AFP	alpha fetoprotein(AFP)
AGER	advanced glycosylation end-product specific receptor(AGER)
AGO4	argonaute 4, RISC catalytic component(AGO4)
AGRP	agouti related neuropeptide(AGRP)
AGTR2	angiotensin II receptor type 2(AGTR2)
AHCY	adenosylhomocysteinase(AHCY)
AHR	aryl hydrocarbon receptor(AHR)

molecular function?

e.g.:

GTPase activator activity

Binds to and increases the activity of a GTPase, an enzyme that catalyzes the hydrolysis of GTP.

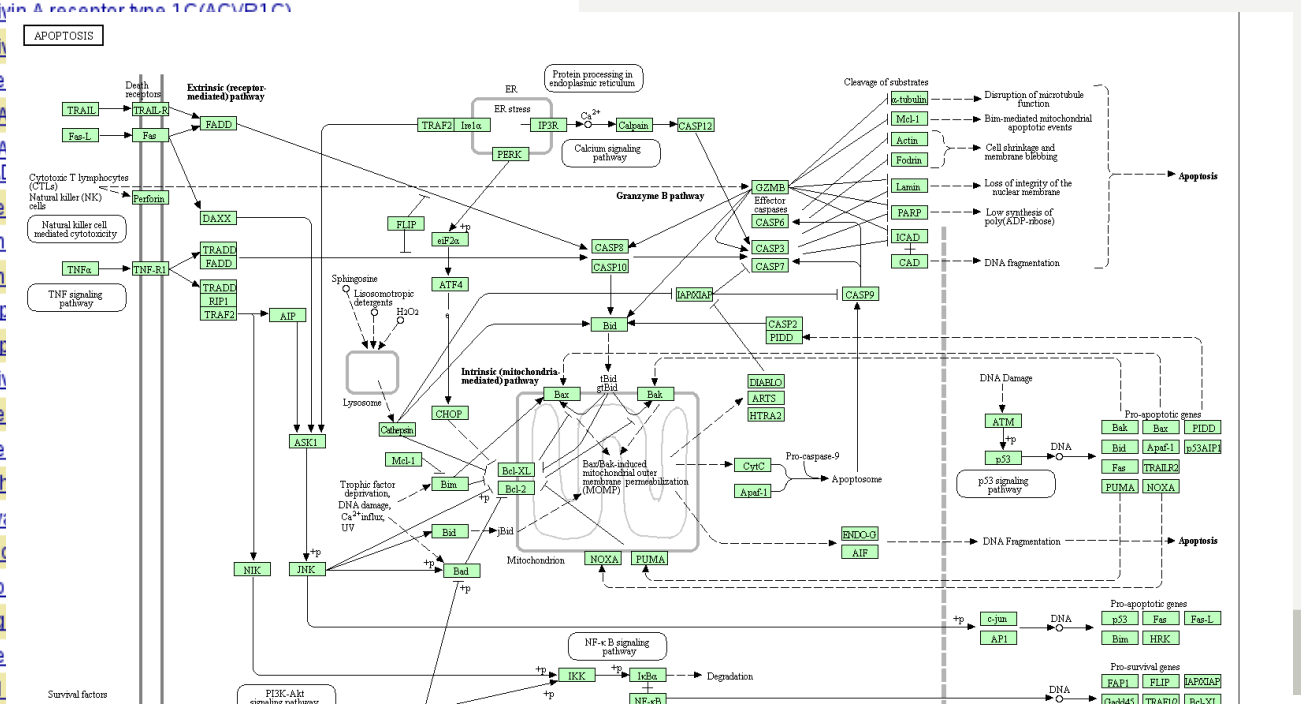
Result of gene-based DE tests

Beyond single gene analysis: what is the "drive" of the changes?

Input: long list of DEG-s

OFFICIAL_GENE_SYMBOL	Gene Name
AANAT	aralkylamine N-acetyltransferase(AANAT)
ABCB4	ATP binding cassette subfamily B member 4(ABCB4)
ABCC2	ATP binding cassette subfamily C member 2(ABCC2)
ABHD2	abhydrolase domain containing 2(ABHD2)
ABR	active BCR-related(ABR)
ACVR1C	activin A receptor type 1C(ACVR1C)
ACVRL1	acti
ADA	ade
ADAM15	ADA
ADAMTS1	ADA 1(AD
ADCYAP1	ade
ADGRG1	adh
ADGRL1	adh
ADIPOQ	adip
ADIPOR1	adip
ADNP	acti
ADORA1	ade
ADORA2A	ade
AFP	alph
AGER	adv
AGO4	argo
AGRP	ago
AGTR2	ang
AHCY	ade
AHR	aryl

molecular function?
biological process?
pathway?



Result of gene-based DE tests

Beyond single gene analysis: what is the "drive" of the changes?

Input: long list of DEG-s

OFFICIAL_GENE_SYMBOL	Gene Name
AANAT	aralkylamine N-acetyltransferase(AANAT)
ABCB4	ATP binding cassette subfamily B member 4(ABCB4)
ABCC2	ATP binding cassette subfamily C member 2(ABCC2)
ABHD2	abhydrolase domain containing 2(ABHD2)
ABR	active BCR-related(ABR)
ACVR1C	activin A receptor type 1C(ACVR1C)
ACVRL1	activin A receptor like type 1(ACVRL1)
ADA	adenosine deaminase(ADA)
ADAM15	ADAM metallopeptidase domain 15(ADAM15)
ADAMTS1	ADAM metallopeptidase with thrombospondin type 1 motif 1(ADAMTS1)
ADCYAP1	adenylate cyclase activating polypeptide 1(ADCYAP1)
ADGRG1	adhesion G protein-coupled receptor G1(ADGRG1)
ADGRL1	adhesion G protein-coupled receptor L1(ADGRL1)
ADIPOQ	adiponectin, C1Q and collagen domain containing(ADIPOQ)
ADIPOR1	adiponectin receptor 1(ADIPOR1)
ADNP	activity dependent neuroprotector homeobox(ADNP)
ADORA1	adenosine A1 receptor(ADORA1)
ADORA2A	adenosine A2a receptor(ADORA2A)
AFP	alpha fetoprotein(AFP)
AGER	advanced glycosylation end-product specific receptor(AGER)
AGO4	argonaute 4, RISC catalytic component(AGO4)
AGRP	agouti related neuropeptide(AGRP)
AGTR2	angiotensin II receptor type 2(AGTR2)
AHCY	adenosylhomocysteinase(AHCY)
AHR	aryl hydrocarbon receptor(AHR)

molecular function?
biological process?
pathway?
targets of TF / microRNA / lincRNA?
disease?

...

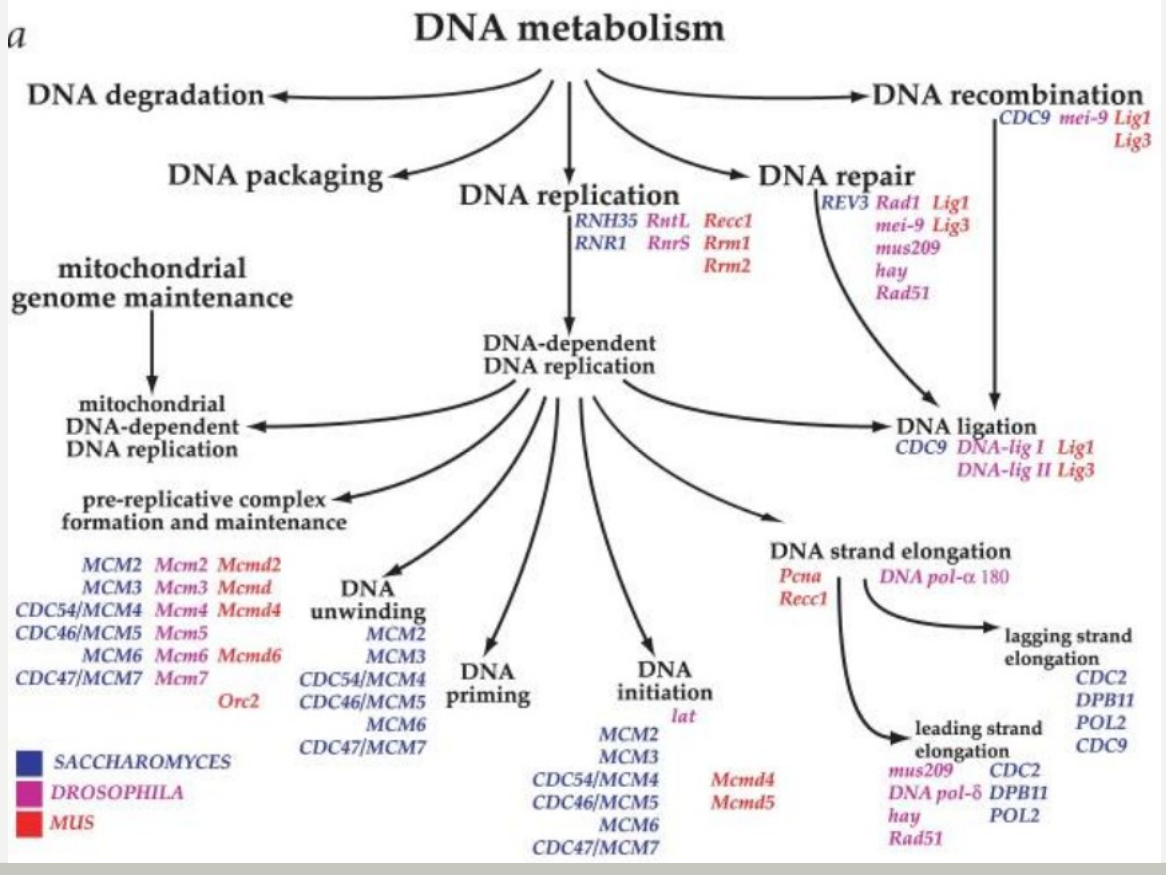
most common definition:

knowledge associated with a **`set of`** genes

Gene Ontology started in 2007

Joint effort of model organism annotators with the goal:

to produce a **structured**, **precisely defined**, **common**, **controlled vocabulary** for describing the roles of **genes** and **gene products** in **any organism**



GO is organized as directed acyclic graph (**DAG**)

annotations associated to a DAG node → but imply association to all parent nodes as well (upward propagation)

GO: constantly updated, extended

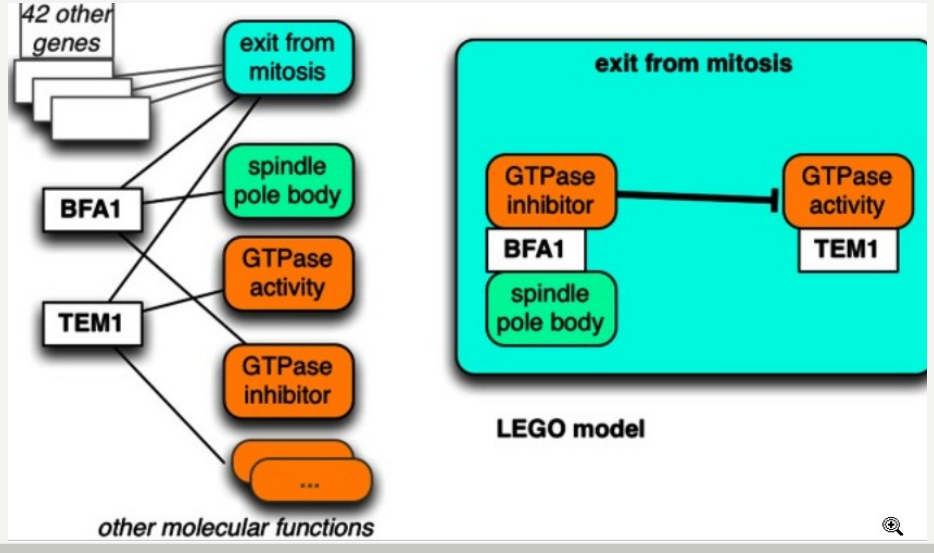
report from 2016

Aspect	Terms (classes)	Relationships
Molecular function (MF)	10 417	14 039
Cellular component (CC)	4022	7854
Biological process (BP)	29 146	71 372

Organism	Biological process EXP	Biological process IBA
Human	38 819	14 596
Mouse	59 517	18 128
Rat	27 591	16 810
Zebrafish	18 004	17 001
Fruit fly	30 560	5913
Nematode (*C. elegans*)	11 679	7683
Slime mold (*D. discoideum*)	3630	4637
Budding yeast	17 646	3608

novel extensions:

- additional links to other databases
- traceable curations (pmids)
- negative (NOT) annotations
- towards LEGO (Linked Expressions using the Gene Ontology)



Gene set enrichment general input

set to gene associations

x sets where set i is associated with p_i genes

set 1 → $(g_1, g_2, g_3, \ldots, g_{p1})$

set 2 → $(g_1, g_2, g_3, \ldots, g_{p2})$

set 3 → $(g_1, g_2, g_3, \ldots, g_{p3})$

set 4 → $(g_1, g_2, g_3, \ldots, g_{p4})$

set 5 → $(g_1, g_2, g_3, \ldots, g_{p5})$

set 6 → $(g_1, g_2, g_3, \ldots, g_{p6})$

set 7 → $(g_1, g_2, g_3, \ldots, g_{p7})$

experimental outcome

g_j: a subset of following information

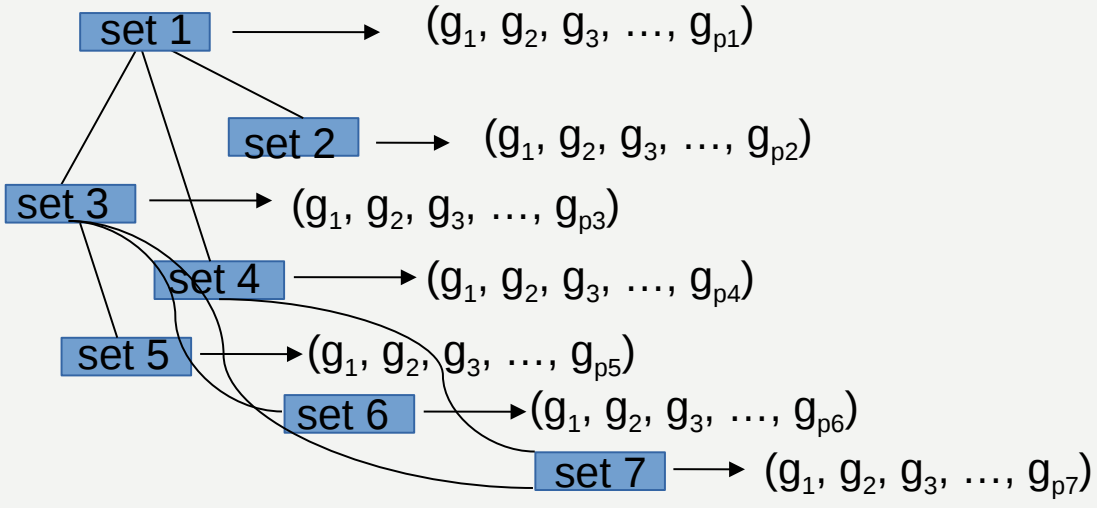
- is target of the enrichment
- (e.g. significantly differentially expressed)
-
- phenotype information **X**, **Y** with $\mathbf{n_1}$, $\mathbf{n_2}$ replicates
- $\mathbf{X = (X_1, \ldots, X_{n1})}$
- $\mathbf{Y = (Y_1, \ldots, Y_{n2})}$

- some derived measure(s) for the gene – phenotype association (e.g. p-value, $\log_2$(FC))

Gene set enrichment general input

set to gene associations
+
ontology

x sets where set i is associated with p_i genes



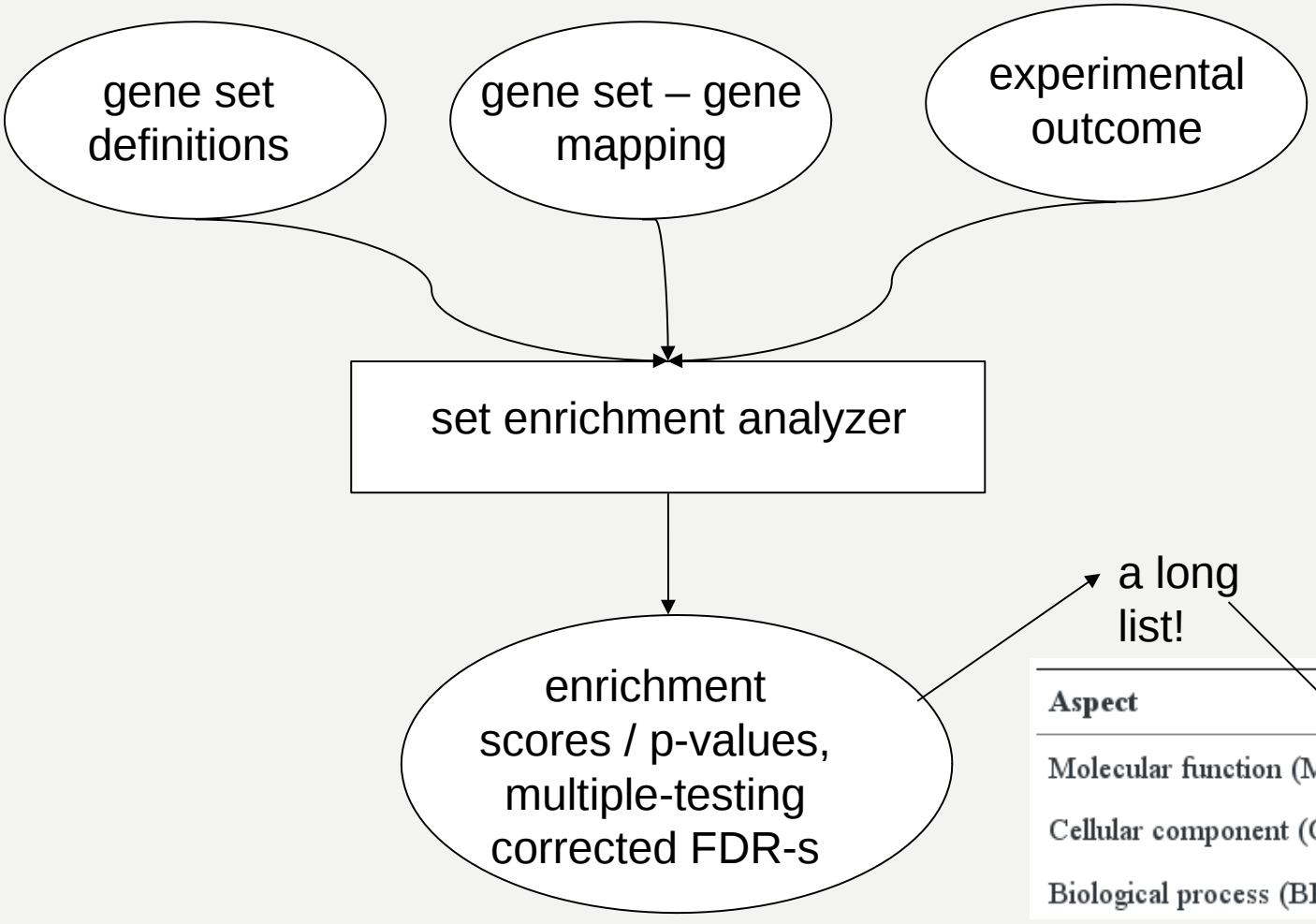
experimental outcome

g_j: a subset of following information

- is target of the enrichment
- (e.g. significantly differentially expressed)
-
- phenotype information **X**, **Y** with $\mathbf{n_1}$, $\mathbf{n_2}$ replicates
- $\mathbf{X = (X_1, \ldots, X_{n1})}$
- $\mathbf{Y = (Y_1, \ldots, Y_{n2})}$

- some derived measure for the gene – phenotype association (e.g. p-value, $\log_2$(FC))

Gene set enrichment general workflow



e.g. GO

Aspect	Terms (classes)	Relationshi
Molecular function (MF)	10 417	14 039
Cellular component (CC)	4022	7854
Biological process (BP)	29 146	71 372

Gene set enrichment strategies

Depends on **what** and **how** we use from the experimental data

Combine p-values of genes (Gamma, Fischer, Stouffer) → one p-value per set

set i → $(g_1, g_2, g_3, \ldots, g_{pi})$

p-value of g_1, p-value of g_2, p-value of g_3, p-value of g_{pi}

e.g. Fischer:

$$X^2_{2k} \sim -2 \sum_{i=1}^{k} \ln(p_i)$$

problem: genes are not independent

Gene set enrichment strategies

Depends on **what** and **how** we use from the experimental data

- **Multivariate tests** on gene set with p genes:
- phenotype information **X**, **Y** with $\mathbf{n_1}$, $\mathbf{n_2}$ replicates
- $X = (X_1, \ldots, X_{n1}) \sim$ distrib F mean μ_x
- $Y = (Y_1, \ldots, Y_{n2}) \sim$ distrib G mean μ_y

Hypothesis:

General:

$H_0: F = G \qquad H_1: F \neq G$

Restricted:

$H_0: \mu_x = \mu_y \qquad H_1: \mu_x \neq \mu_y$

example: N-statistics

$$N_{n_1 n_2} = \frac{n_1 n_2}{n_1 + n_2}\left[\frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} L(X_i, Y_j) - \frac{1}{2n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} L(X_i, X_j) - \frac{1}{2n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2} L(Y_i, Y_j)\right]^{1/2}$$

L: e.g. euclidean distance on some normalized value (e.g. RPKM)

$\mathbf{X_i}$, $\mathbf{Y_j}$ have length p

$$L(X, Y) = \|X - Y\|$$

Gene set enrichment strategies

Depends on **what** and **how** we use from the experimental data

List-based:

reduce the experimental outcome to a simple list of targets:

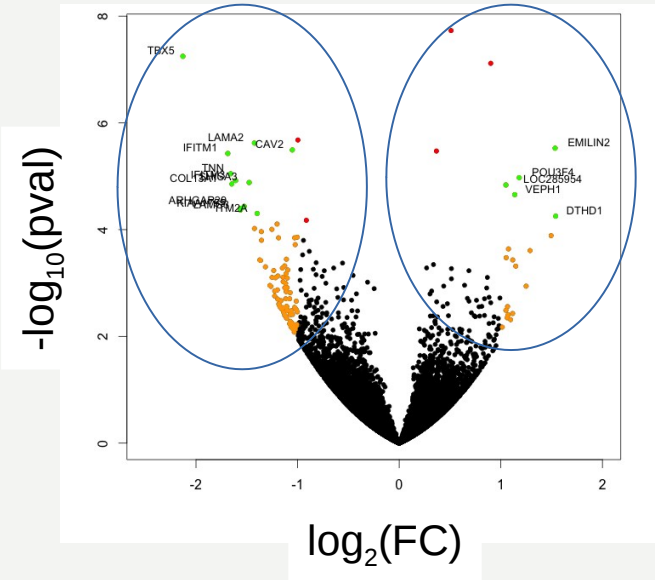
„is target of the enrichment (e.g. DEG-s) „

typical: DE-genes with:

FDR < 0.05,

abs($\log_2$(FC)) >= 1.0

→ **input: a list of gene-ids of interest**



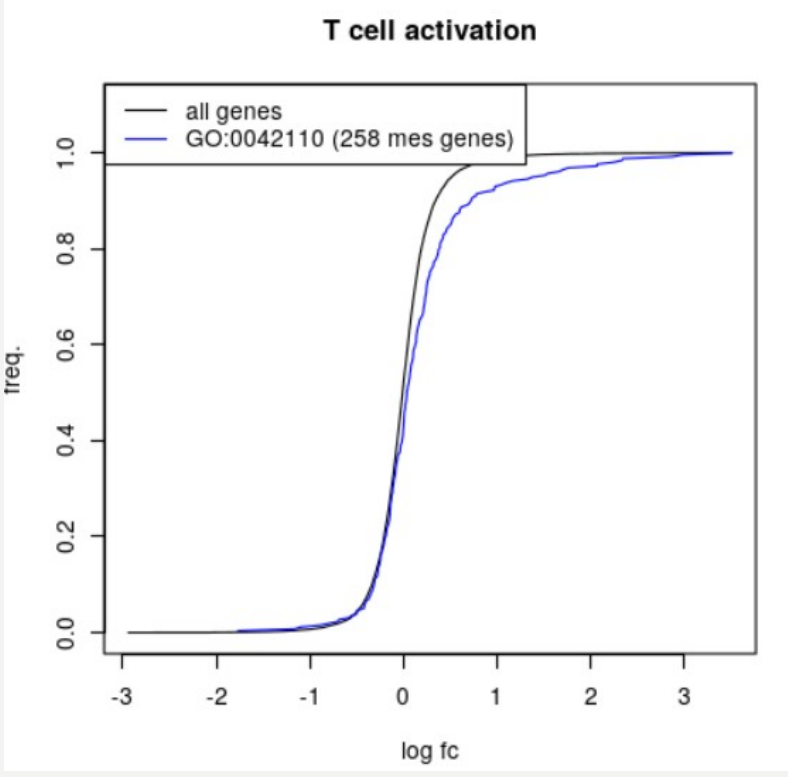
Gene set enrichment strategies

Depends on **what** and **how** we use from the experimental data

Distribution-based:

„some derived measure for the gene – phenotype association (e.g. p-value, $\log_2$(FC))“

example: compare the **fold change** distribution of the **genes within the set against** the fold change distribution of **all other** measured genes



List based Set Enrichment: Over-representation analysis (ORA)

input:

- list of gene-ids of interest (e.g. **DEG-s**)
- a pre-defined **set** of genes to test (e.g. a biological process)

- → contingency table. hypothesis: $\mathbf{H_0}$**: a/b = c/d**

	in set	not in set
significant DE	a	b
non-significant DE	c	d

Over-representation analysis (ORA)

Fischer‘s Exact

	in set	not in set	row total
significant DE	a	b	a+b
non-significant DE	c	d	c+d
column total	a+c	b+d	a + b + c + d (=n)

Hypergeometric

N: total genes
K: DEG-s
n: set size
k: overlap (DE genes, set genes)

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

Fischer‘s exact = Hypergeometric with:

N = a+b+c+d
K = a+b
n = a+c
K = a

Hypergeometric = Fischer‘s exact with:

a = k
b = K – k
c = N – k
d = N – n – K + k

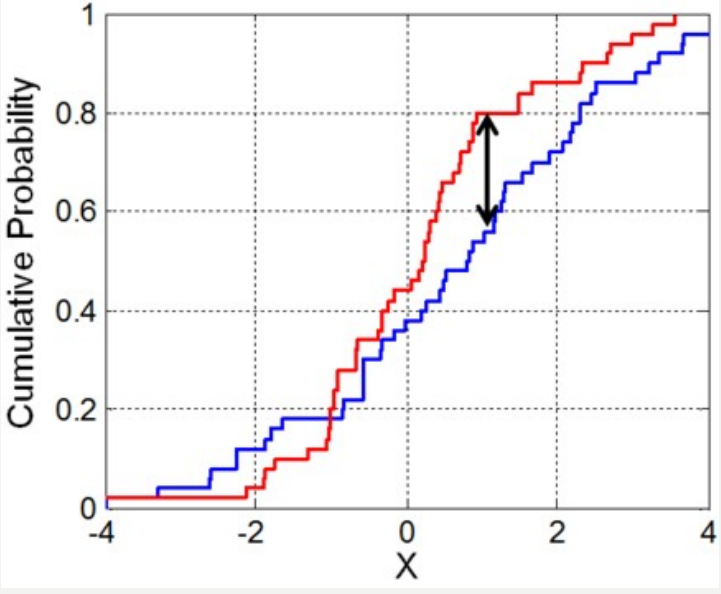
Distribution based set enrichment

input: „some derived measure for the gene – phenotype association“

example: compare the fold change distribution of the genes within the set against the fold change distribution of all other measured genes

Kolmogorov-Smirnov statistics

compares the cumulative distributions:



- statistic value: **Dm,n**: maximum distance between the empirical distribution
- basically a running-sum difference test between two lists of length **n** and **m**,
- no need to know the type and parameter of the underlying distributions

$$\lim_{m,n\to\infty} \Pr\left\{ \sqrt{mn/(m+n)}\, D_{m,n} \le z \right\} = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} \exp\left(-2i^2 z^2\right)$$

DAVID

Widely used web-based ORA since 2003 (Lempicki group (for ref see last slides))

Easy to use as:

- provides a solution for the gene name-mapping problem (many possible input gene identifier types, results depend on how one handles ambiguities, version of mappings etc…)
- integrates many resources (many organisms, but also multiple annotation repositories)
- web-based (no installation needed – but if version changes on the web implications might be also affected)
- provides a convenient clustering of the long output lists (see next slides)

DAVID

Widely used web-based ORA since 2003 (Lempicki group)

Enrichment score: Fischer‘s exact test with **jackknifing**

Jackknifing: a single data point is removed and the statistic is recalculated many times → a distribution of probabilities that is broad if the result is highly variable and tight if the result is robust.

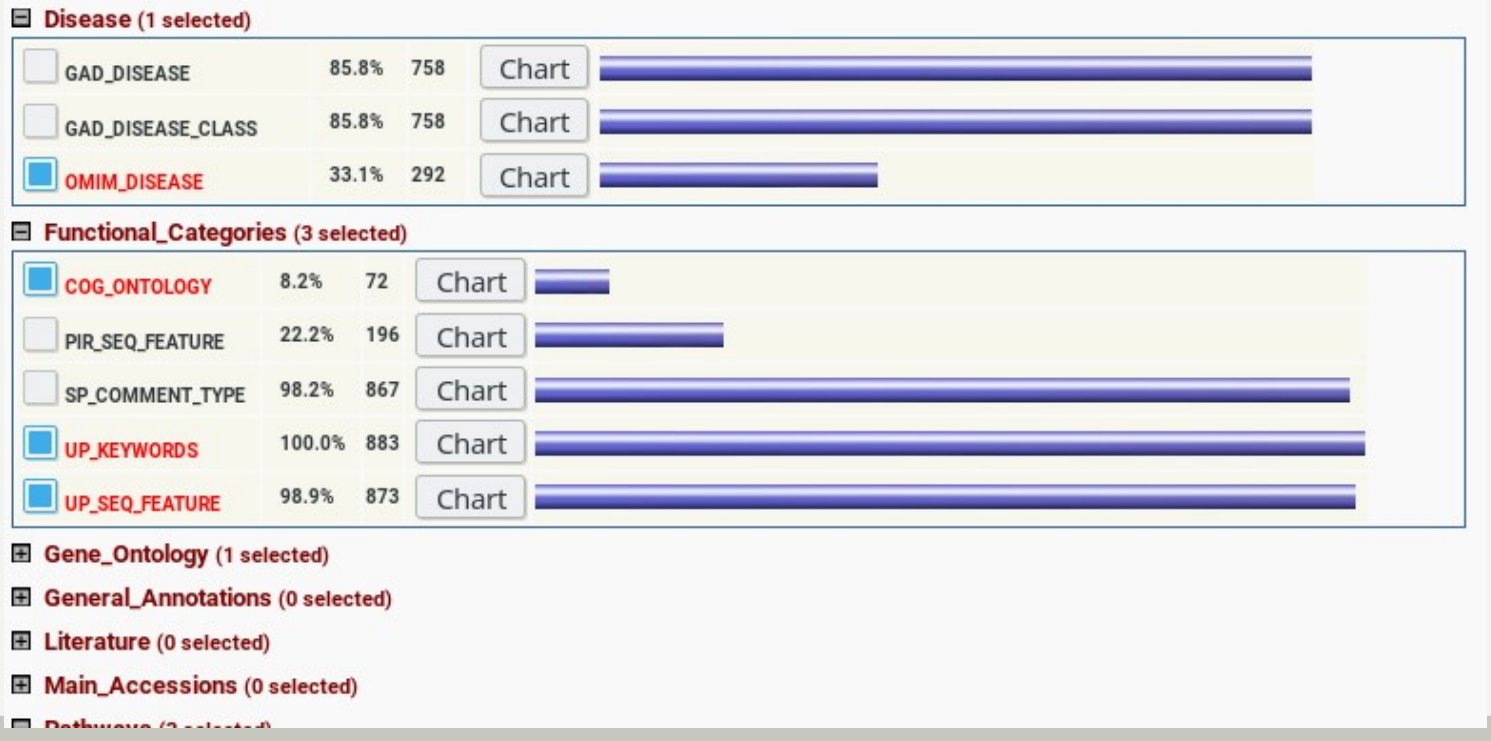
In case of Fischer‘s exact: simply use a-1 instead of a.

DAVID

Widely used web-based ORA since 2003 (Lempicki group)

Integrated resources:

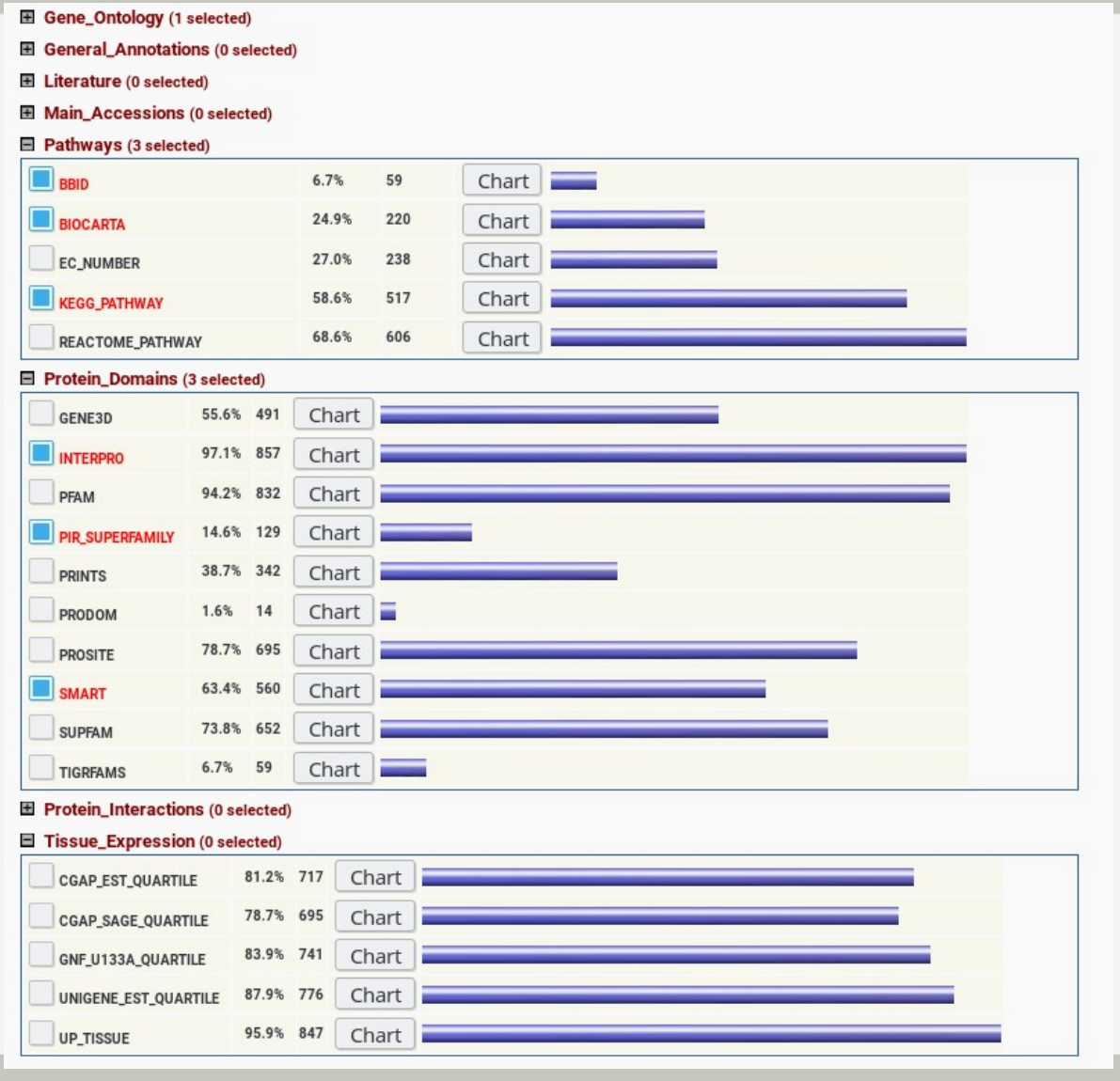
GO, KEGG, Biocarta Pathways, Swiss-Prot keywords, UniProt Sequence features … (many more)



DAVID

Widely used web-based ORA since 2003 (Lempicki group)

Integrate resources: GO, KEGG, Biocarta Pathways, Swiss-Prot keywords, UniProt Sequence Features … (many more)



DAVID

Widely used web-based ORA since 2003 (Lempicki group)
Enrichment score: Fischer's exact test with **jackknifing**

long list

many categories with similar description

4156 chart records

Download File

1, 2, 3, 4, 5

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_BP_ALL	rhythmic process	RT		273	30.9	4.0E-318	3.7E-314
	GOTERM_BP_ALL	negative regulation of locomotion	RT		229	25.9	3.5E-267	1.6E-263
	GOTERM_BP_ALL	negative regulation of cellular component movement	RT		223	25.3	4.4E-253	1.4E-249
	GOTERM_BP_ALL	negative regulation of cell motility	RT		201	22.8	2.2E-232	5.1E-229
	GOTERM_BP_ALL	negative regulation of cell migration	RT		192	21.7	4.3E-224	8.0E-221
	GOTERM_BP_ALL	regulation of localization	RT		479	54.2	1.9E-177	3.0E-174
	GOTERM_BP_ALL	regulation of locomotion	RT		281	31.8	2.2E-170	2.9E-167
	GOTERM_BP_ALL	circadian rhythm	RT		147	16.6	1.2E-162	1.4E-159
	GOTERM_BP_ALL	regulation of cell motility	RT		267	30.2	7.1E-160	7.3E-157
	GOTERM_BP_ALL	regulation of cell migration	RT		258	29.2	1.0E-158	9.7E-156
	GOTERM_BP_ALL	regulation of cellular component movement	RT		272	30.8	2.5E-154	2.1E-151
	GOTERM_BP_ALL	presynaptic process involved in chemical synaptic transmission	RT		127	14.4	9.0E-143	7.0E-140
	GOTERM_BP_ALL	neurotransmitter transport	RT		143	16.2	5.2E-139	3.7E-136
	GOTERM_BP_ALL	regulation of neurotransmitter levels	RT		142	16.1	2.8E-137	1.9E-134
	GOTERM_BP_ALL	neurotransmitter secretion	RT		121	13.7	4.3E-135	2.6E-132
	GOTERM_BP_ALL	signal release from synapse	RT		121	13.7	4.3E-135	2.6E-132
	GOTERM_BP_ALL	cell migration	RT		296	33.5	1.3E-128	7.8E-126
	GOTERM_BP_ALL	regulation of biological quality	RT		502	56.9	4.0E-126	2.2E-123
	GOTERM_BP_ALL	locomotion	RT		329	37.3	2.3E-124	1.2E-121
	GOTERM_BP_ALL	regulation of multicellular organismal process	RT		430	48.7	4.9E-123	2.4E-120
	GOTERM_BP_ALL	cell motility	RT		304	34.4	5.1E-121	2.4E-118
	GOTERM_BP_ALL	localization of cell	RT		304	34.4	5.1E-121	2.4E-118

DAVID

Widely used web-based ORA since 2003 (Lempicki group)

Strategy to make the long output list more interpretable:
cluster genes by their annotations

each gene is a binary vector of length of all available annotation categories:

	Cell death	Apoptosis	Ph domain	Sh2 domain	Apoptosis pathway	Membrane
Gene a	1	1	0	0	1	0
Gene b	1	1	0	1	1	0
Gene c	1	0	0	1	1	1
Gene d	1	1	0	0	1	1
Gene e	0	1	1	1	1	1
Gene f	0	0	1	1	0	1
Gene g	0	0	1	1	0	1

DAVID

Widely used web-based ORA since 2003 (Lempicki group)

Strategy to make the long output list more interpretable: cluster genes by their annotations

each gene is a binary vector of length of all available annotation categories:

	Cell death	Apoptosis	Ph domain	Sh2 domain	Apoptosis pathway	Membrane
Gene a	1	1	0	0	1	0
Gene b	1	1	0	1	1	0
Gene c	1	0	0	1	1	1
Gene d	1	1	0	0	1	1
Gene e	0	1	1	1	1	1
Gene f	0	0	1	1	0	1
Gene g	0	0	1	1	0	1

- similarity is measured by the correlation of the annotation vectors
- as vector binary use the Kappa (K) statistic

$$K_{ab} = \frac{O_{ab} - A_{ab}}{1 - A_{ab}}$$

O_{mn} num co-occurrence

A_{mn} chance co-occurrence

DAVID

- similarity of genes is measured by the Kappa statistic on the co-occurrences of the binary annotation vectors

	Cell death	Apoptosis	Ph domain	Sh2 domain	Apoptosis pathway	Membrane
Gene a	1	1	0	0	1	0
Gene b	1	1	0	1	1	0
Gene c	1	0	0	1	1	1
Gene d	1	1	0	0	1	1
Gene e	0	1	1	1	1	1
Gene f	0	0	1	1	0	1
Gene g	0	0	1	1	0	1

		Gene a		
		1	0	Row total
Gene b	1	3 ($C_{1,1}$)	**1 ($C_{1,0}$)**	4 ($C_{1,*}$)
	0	0 ($C_{0,1}$)	2 ($C_{0,0}$)	2 ($C_{0,*}$)
Column total		3 ($C_{*,1}$)	3 ($C_{*,0}$)	6 (T_{ab})

$$O_{ab} = \frac{C_{1,1} + C_{0,0}}{T_{ab}} = \frac{3+2}{6} = 0.83$$

$$A_{ab} = \frac{C_{*,1} \bullet C_{1,*} + C_{*,0} \bullet C_{0,*}}{T_{ab} \bullet T_{ab}} = \frac{3 \bullet 4 + 3 \bullet 2}{6 \bullet 6} = 0.5$$

$$K_{ab} = \frac{O_{ab} - A_{ab}}{1 - A_{ab}} = \frac{0.83 - 0.5}{1 - 0.5} = 0.66$$

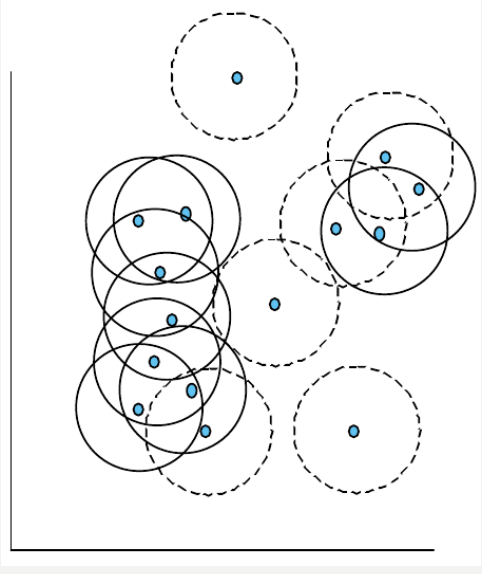
DAVID

- strategy to make the long output list more interpretable: **cluster genes by their annotations**
- similarity of genes is measured by the **Kappa statistic** on the co-occurrences of the binary annotation vectors
- clustering: how? normal strategies (hierarchical clustering, k-means, SOM) would assign exactly one cluster per gene – does not reflect the situation:
  - gene to cluster relationship may be weak – better not to assign
  - gene may belong to multiple clusters
  - → need fuzzy multi-linkage clustering

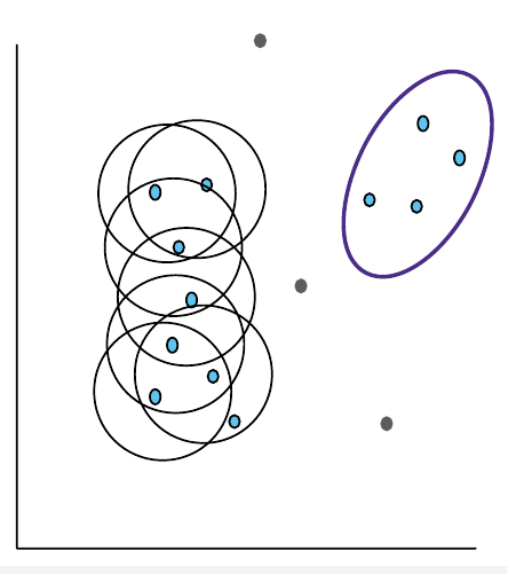
DAVID

- strategy to make the long output list more interpretable: **cluster genes by their annotations:** fuzzy multi-linkage: 3 steps:

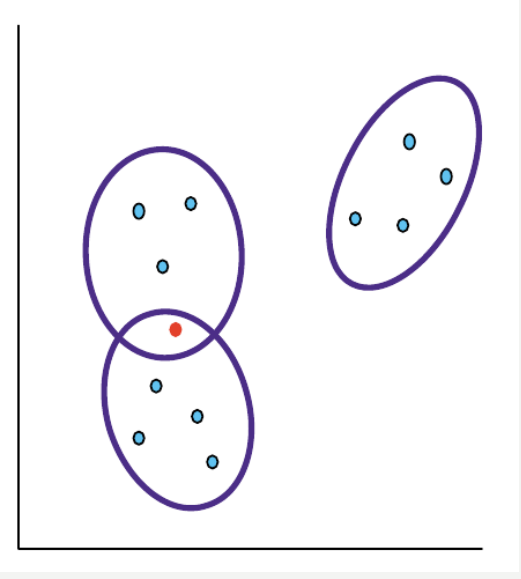
points: genes, distance: kappa



initializing multiple seeds



multi-linkage merging (based on fraction shared group members)



iterate until no more merging

DAVID

- strategy to make the long output list more interpretable: **cluster genes by their annotations → build an own flat DAG**

2354 long

537 Cluster(s)

Download File

Annotation Cluster 1	Enrichment Score: 165.72			Count	P_Value	Benjamini
GOTERM_BP_ALL	negative regulation of locomotion	RT		229	3.5E-267	1.6E-263
GOTERM_BP_ALL	negative regulation of cellular component movement	RT		223	4.4E-253	1.4E-249
GOTERM_BP_ALL	negative regulation of cell motility	RT		201	2.2E-232	5.1E-229
GOTERM_BP_ALL	negative regulation of cell migration	RT		192	4.3E-224	8.0E-221
GOTERM_BP_ALL	regulation of localization	RT		479	1.9E-177	3.0E-174
GOTERM_BP_ALL	regulation of locomotion	RT		281	2.2E-170	2.9E-167
GOTERM_BP_ALL	regulation of cell motility	RT		267	7.1E-160	7.3E-157
GOTERM_BP_ALL	regulation of cell migration	RT		258	1.0E-158	9.7E-156
GOTERM_BP_ALL	regulation of cellular component movement	RT		272	2.5E-154	2.1E-151
GOTERM_BP_ALL	cell migration	RT		296	1.3E-128	7.8E-126
GOTERM_BP_ALL	locomotion	RT		329	2.3E-124	1.2E-121
GOTERM_BP_ALL	cell motility	RT		304	5.1E-121	2.4E-118
GOTERM_BP_ALL	localization of cell	RT		304	5.1E-121	2.4E-118
GOTERM_BP_ALL	movement of cell or subcellular component	RT		329	2.7E-103	8.9E-101
GOTERM_BP_ALL	negative regulation of multicellular organismal process	RT		246	6.3E-101	1.9E-98
Annotation Cluster 2	**Enrichment Score: 86.52**			**Count**	**P_Value**	**Benjamini**
GOTERM_BP_ALL	regulation of multicellular organismal process	RT		430	4.9E-123	2.4E-120
GOTERM_BP_ALL	single-multicellular organism process	RT		616	5.6E-101	1.7E-98
GOTERM_BP_ALL	regulation of developmental process	RT		346	6.9E-94	1.7E-91
GOTERM_BP_ALL	regulation of multicellular organismal development	RT		307	6.2E-92	1.3E-89

-log of geometric mean of member group FDR-s

DAVID

- strategy to make the long output list more interpretable: **cluster genes by their annotations**

- **allows for combining multiple resources for enrichment**

- take care what you select: by default
- BP_direct is used, not BP_all → less (and other) clusters found

182 Cluster(s) Download File

Annotation Cluster 1	**Enrichment Score: 25.11**			**Count**	**P_Value**	**Benjami**
GOTERM_BP_DIRECT	cellular oxidant detoxification	RT		60	6.0E-65	1.1E-61
UP_KEYWORDS	Oxidoreductase	RT		52	1.1E-6	1.4E-5
GOTERM_BP_DIRECT	oxidation-reduction process	RT		58	7.1E-6	2.1E-4
Annotation Cluster 2	**Enrichment Score: 21.86**			**Count**	**P_Value**	**Benjami**
GOTERM_CC_DIRECT	extracellular space	RT		179	4.8E-37	1.5E-34
UP_SEQ_FEATURE	signal peptide	RT		260	5.7E-23	6.9E-20
UP_KEYWORDS	Secreted	RT		174	5.8E-21	5.1E-19
GOTERM_CC_DIRECT	extracellular region	RT		152	4.8E-16	2.3E-14
UP_KEYWORDS	Signal	RT		278	6.5E-16	2.3E-14
Annotation Cluster 3	**Enrichment Score: 21.48**			**Count**	**P_Value**	**Benjami**
GOTERM_BP_DIRECT	calcium ion-regulated exocytosis of neurotransmitter	RT		33	7.0E-35	4.1E-32
GOTERM_MF_DIRECT	syntaxin binding	RT		41	1.3E-31	1.4E-28
GOTERM_BP_DIRECT	regulation of calcium ion-dependent exocytosis	RT		29	4.6E-30	2.4E-27
GOTERM_BP_DIRECT	vesicle fusion	RT		32	1.7E-24	5.8E-22
GOTERM_MF_DIRECT	clathrin binding	RT		29	4.5E-23	1.2E-20
UP_SEQ_FEATURE	domain:C2 1	RT		29	7.1E-23	5.6E-20
UP_SEQ_FEATURE	domain:C2 2	RT		29	7.1E-23	5.6E-20
GOTERM_MF_DIRECT	calcium-dependent phospholipid binding	RT		30	1.2E-22	2.6E-20
INTERPRO	C2 calcium-dependent membrane targeting	RT		44	4.7E-21	3.2E-18
INTERPRO	Synaptotagmin	RT		18	1.6E-20	7.4E-18
SMART	C2	RT		40	2.2E-19	6.2E-17
UP_SEQ_FEATURE	topological domain:Vesicular	RT		14	1.0E-8	4.1E-6
GOTERM_MF_DIRECT	calcium ion binding	RT		67	1.8E-6	4.8E-5
Annotation Cluster 4	**Enrichment Score: 19**			**Count**	**P_Value**	**Benjami**
UP_KEYWORDS	Disulfide bond	RT		287	4.0E-32	8.8E-30
UP_SEQ_FEATURE	disulfide bond	RT		253	2.5E-29	6.0E-26

GSEA (Subramanian et al. 2005)

GSEA – (gene set enrichment analysis) – widely used distribution based method - based on the idea of Kolmogorov-Smirnov

input: “some derived measure for the gene – phenotype association (e.g. p-value, $\log_2$(FC))“ → in GSEA correlation with phenotype

	condition 1 (phenotype 1)	condition 2 (phenotype 2)	log(FC)	pval,FDR
	$\mathbf{rep_1}$ $\mathbf{rep_2}$ ... $\mathbf{rep_{n1}}$	$\mathbf{rep_1}$ $\mathbf{rep_2}$ ... $\mathbf{rep_{n2}}$		
gene 1	(normalized) counts (or other signals if not RNA-seq)	(normalized) counts (or other signals if not RNA-seq)	FC 1	FDR 1
gene 2			FC 2	FDR 2
...			...	...
gene N			FC N	FDR N
encoded phenotype	condition 1 (phenotype 1)	condition 2 (phenotype 2)		
	0 0 ... 0	1 1 ... 1		

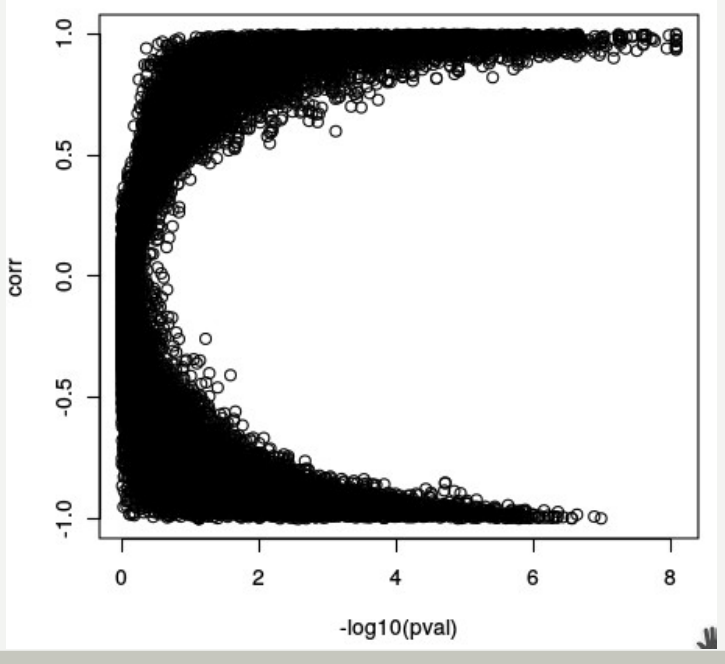
GSEA (Subramanian et al. 2005)

GSEA – derived value: correlation with phenotype

$$\rho_{X,Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

	condition 1 (phenotype 1)	condition 2 (phenotype 2)
encoded phenotype	0 0 ... 0	1 1 ... 1
gene counts	$\mathbf{rep_1}$ $\mathbf{rep_2}$ ... $\mathbf{rep_{n1}}$	$\mathbf{rep_1}$ $\mathbf{rep_2}$ ... $\mathbf{rep_{n2}}$

DE-seq p-value
vs
Phenotype correlation



GSEA

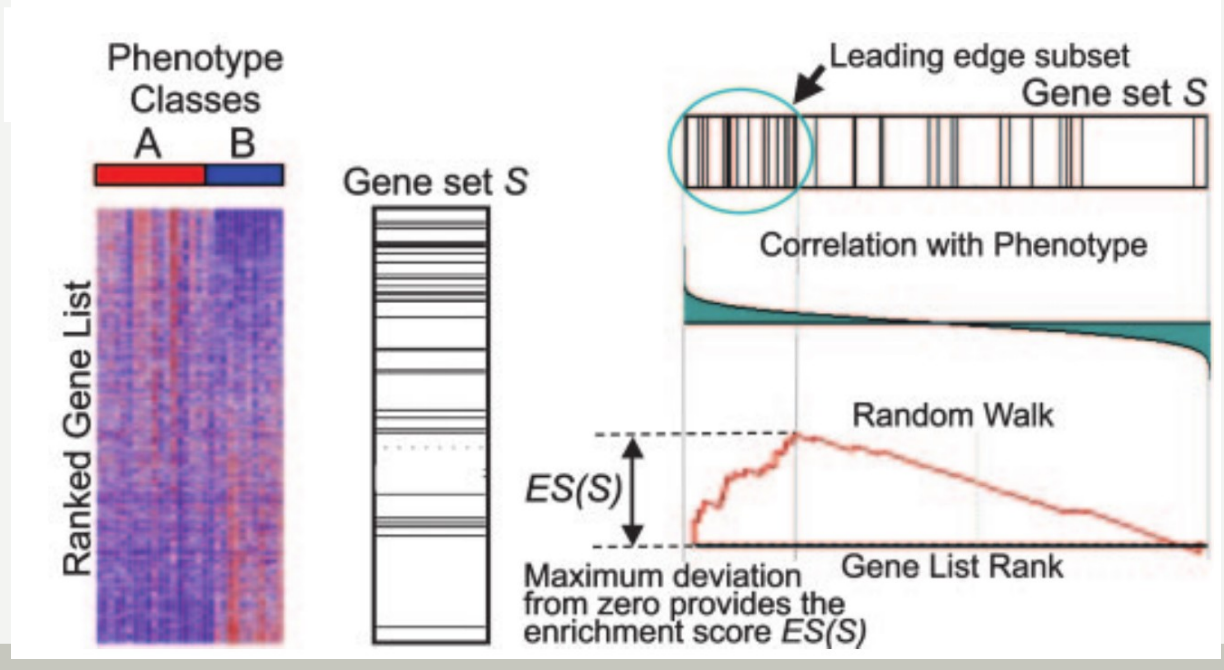
GSEA: weighted KS:

- r_j: correlation of gene j
- statistic value: $P_{\text{hit}} - P_{\text{miss}}$.
- p:
  - if 0 standard KS
  - if 1 weighted KS by correlation

$$P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^p}{N_R}, \quad \text{where } N_R = \sum_{g_j \in S} |r_j|^p$$

$$P_{\text{miss}}(S, i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(N - N_H)}.$$

N = |genes|
N_h = |genes in S|



GSEA

GSEA: weighted KS → standard KS distribution does not apply
→ how to estimate significance?

General solution for such cases (unknown distribution):
calculate the background distribution for the score / statistic value

How?

- **calculate score s for query gene set**
- **permute phenotype labels x times**
- **calculate for each permutation the score ps_i → BG = $[ps_1, ps_2, \ldots, ps_x]$**

significance estimation: p-value = $(1+(|\{ps_i \mid ps_i<s\}|)) / (x+1)$

GSEA

GSEA: weighted KS → standard KS distribution does not apply
→ how to estimate significance?

permute phenotype labels x times

encoded phenotype	0	0	...	0	1	1	...	1	
	1	0	...	0	1	0	...	1	← permutations
				...					
	0	1	...	1	0	0	...	1	
gene counts	rep_1	rep_2	...	rep_{n1}	rep_1	rep_2	...	rep_{n2}	

for a minimal p-value of 0.001 we need ~ 1000 permutations

how many replicates are required to make 1000 different permutations possible?

GSEA

GSEA: weighted KS → standard KS distribution does not apply
→ how to estimate significance?

permute phenotype labels x times

for a minimal p-value of 0.001 we need ~ 1000 permutations

how many replicates are required to make 1000 different permutations possible?

Example: 3 replicates = 10 permutations

cond1	cond2
(x_1, x_2, x_3)	(y_1, y_2, y_3)

(x_1, x_2, y_1)	(x_3, y_2, y_3)	(x_1, x_2, y_2)	(x_3, y_1, y_3)	(x_1, x_2, y_3)	(x_3, y_1, y_2)
(x_1, x_3, y_1)	(x_2, y_2, y_3)	(x_1, x_3, y_2)	(x_2, y_1, y_3)	(x_1, x_3, y_3)	(x_2, y_1, y_2)
(x_2, x_3, y_1)	(x_1, y_2, y_3)	(x_2, x_3, y_2)	(x_1, y_1, y_3)	(x_2, x_3, y_3)	(x_1, y_1, y_2)

GSEA

GSEA: weighted KS → standard KS distribution does not apply → how to estimate significance?

permute phenotype labels x times

for a minimal p-value of 0.001 we need ~ 1000 permutations

how many replicates are required to make 1000 different permutations possible?

For: $n = n_1 = n_2$ $\binom{2n}{n}/2$

replicates	num perm.
3	10
4	35
5	126
6	462
7	**1716**

GSEA

GSEA: weighted KS → standard KS distribution does not apply → how to estimate significance?

permute phenotype labels `1000` times → need >= 7 replicates

in most cases not available… →

fallback solution: permute gene sets:

for significance estimation of the gene set with p:

- calculate score s for query gene set
- select randomly **p genes** x times
- calculate for each the score ps_i→ `BG = [ps1,ps2,…,`ps_x`]`
- significance estimation: `(1+(|`ps_i`<s|)) / (x+1)`

GSEA: weighted KS → standard KS distribution does not apply

permute phenotype labels or genes makes a big difference!

permute phenotype labels (subject sampling):

p-value gives confidence that the associations found between DE-genes and the outcome will be found for a new sample of subjects

→ **self-contained hypothesis** (tests if gene set diff. exp. between **phenotypes**)

permute genes (gene sampling) :

p-value gives confidence that a for a new set of genes from the same subjects (replicates), there will be a similar association being in the set and being called "significant"

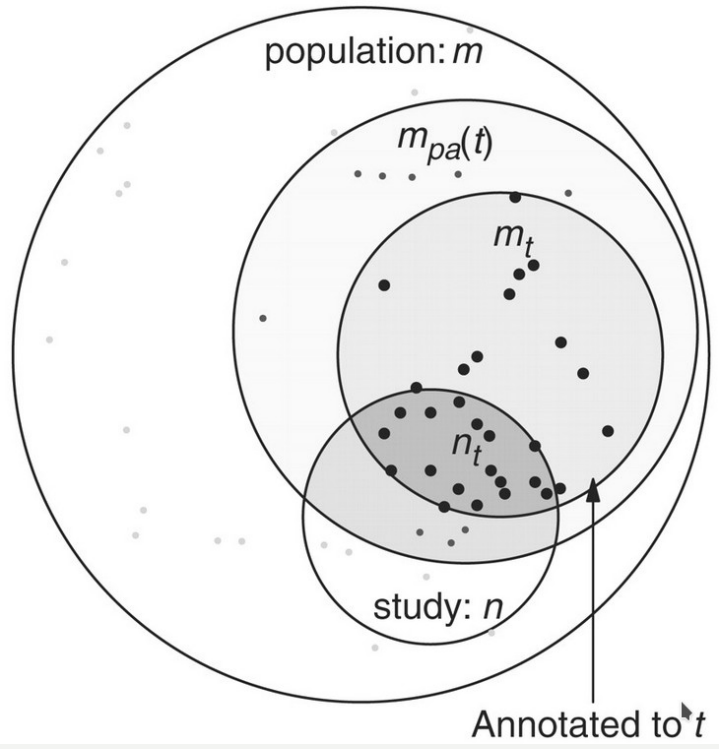
→ **competitive hypothesis** → compares **set versus background**
(standard KS is also competitive)

Dealing with the overlaps /DAG structure

some **overlaps inherently implicated** by the DAG structure: **parent** categories contain all genes from the **child** categories

one can address this with different strategies:

- **eliminate** categories on same path (Grossmann 2007)
- down-weight genes contained in significantly enriched terms while evaluating related terms (Alexa 2006)



Dealing with the overlaps /DAG structure

some **overlaps inherently implicated** by the DAG structure: **parent** categories contain all genes from the **child** categories

there are **many other overlaps** as well…

→ try to address the initial goal: find a **small set of categories explaining the observed changes**

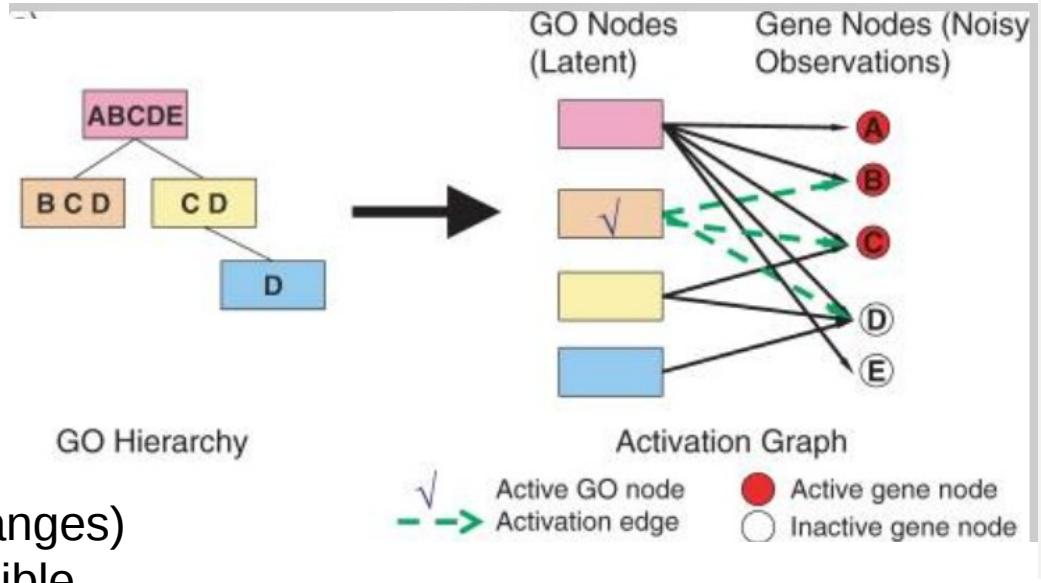
GenGo (Lu 2008) → generatives model maximum likelihood

MGSA (Bauer 2010) → bayesian network

Dealing with the overlaps /DAG structure

GenGo (Lu 2008) → generatives model, maximum likelihood

$$L(C|p,q,G) = |A_g|\log p + |A_n|\log q + |S_g|\log(1-p) + |S_n|\log(1-q) - \alpha|C|$$



p: false negative rate
q: false positive rate
C: set of selected GO terms
(generating the observed gene changes)
α: penalty to use as few terms as possible

A_g—active gene nodes connected to at least one active GO node
A_n—active gene nodes not connected to any active GO nodes
I—inactive gene nodes
S_g—edges connecting nodes in I with active GO nodes
S_n—edges connecting nodes in I with inactive GO nodes

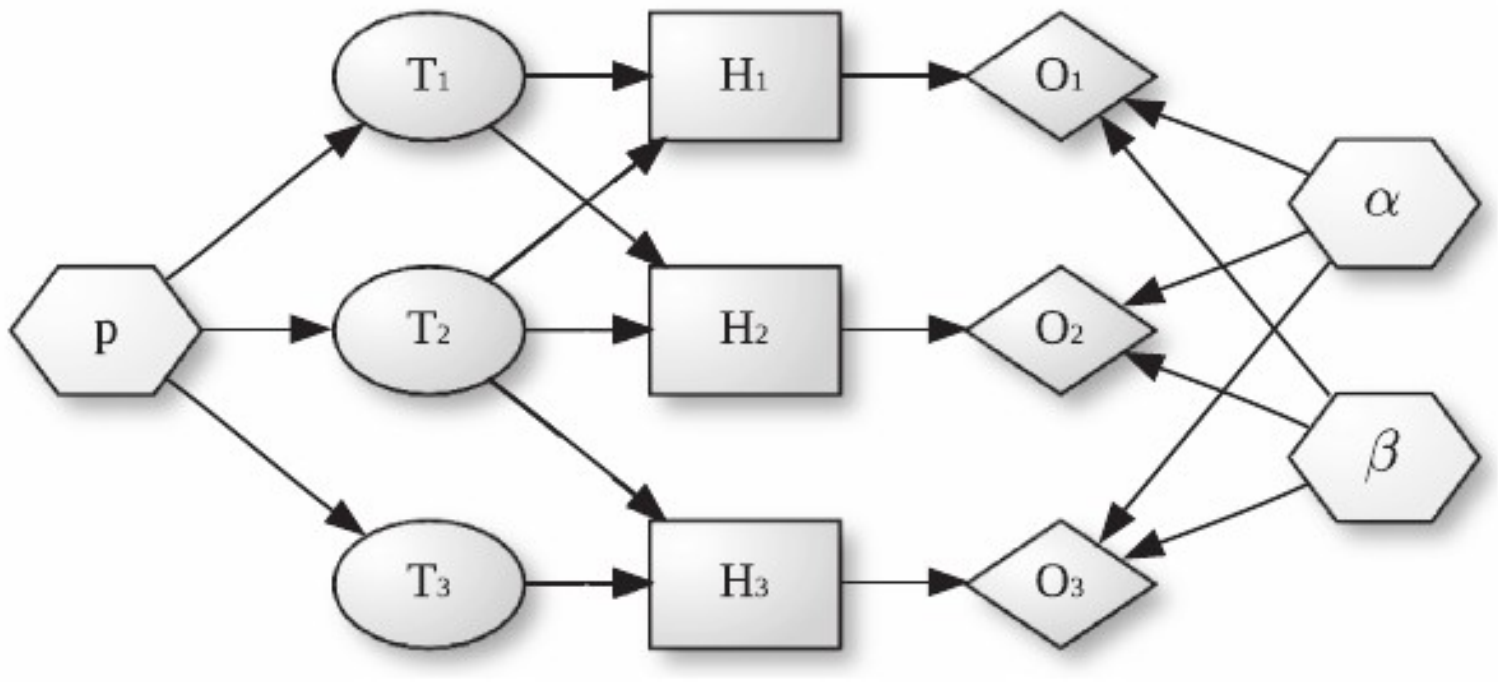
optimization (p, q, C) is NP-hard
→ use greedy heuristic

Dealing with the overlaps /DAG structure

MGSA (Bauer 2010) → bayesian network

p: prior for T (~ penalty to use few)
T_i: terms
H_i: hidden (gene-GO associations)

O_i: observed (genes)
α: false positive rate
β: false negative rate



distributions: Bernoulli, optimization via MCMC (Markov Chain Monte Carlo)

How to evaluate / benchmark?

For real data: no gold standard available:

- **simulation:** (problem: how to simulate realistically?)
  - (+ where to start: from counts / expression signals or sets?)
- **check FP-s**: compare replicates of same condition
- **„cherry picking“:** show that some enriched categories might make sense (i.e. show that the method is usable, but it does not allow for comparing methods or benchmark)
- annotate **„quasi gold standard“** trues and falses based on external knowledge (e.g. experiment was on cancer versus healthy, categories should link to cancer) – limited by how appropriate the trues and falses are selected
- evaluate **robustness** (different aspects) example: take a data-set with many replicates, create subsets from the replicates, run enrichment, define found ones from the full as „true“-s and evaluate how well the subsets agree with these

How to evaluate / benchmark?

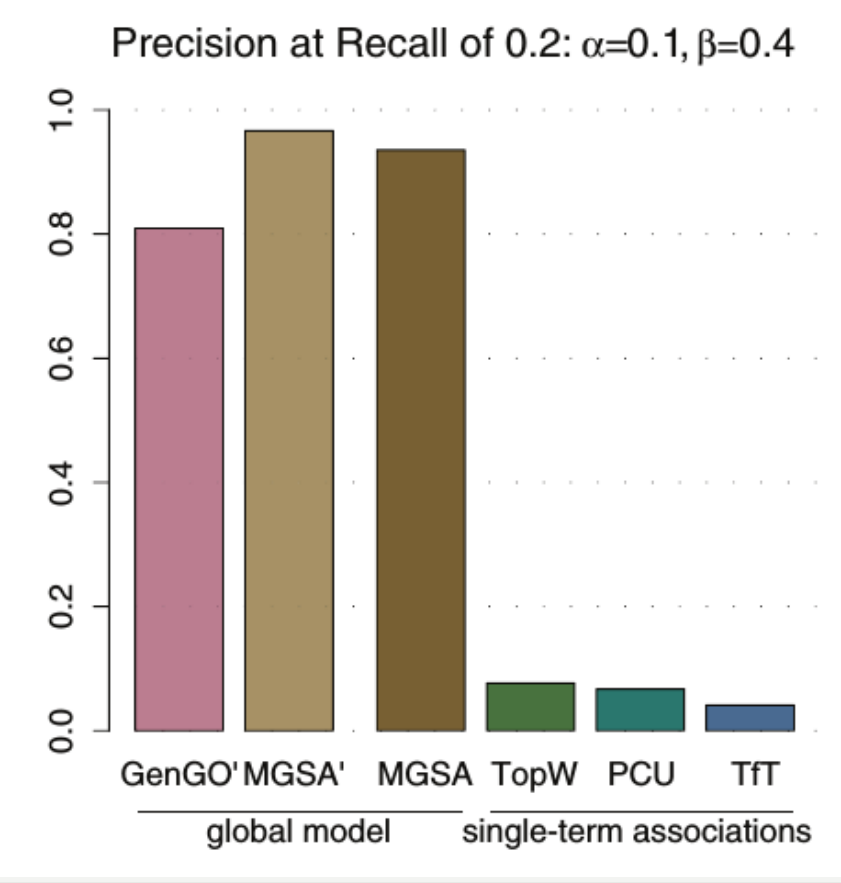
What measures (if applicable):

- FP-rate
- AUROC but if number of trues / falses imbalanced, not appropriate
- → area under precision recall curve: AUPRC (see example next slides)
- robustness: dependence on chosen parameter
- tailored robustness measure e.g.:
  - commonly detecting gene sets in subsets
  - cumulative common detection per subset (CCDS)

simulation results

α: false positive rate β: false negative rate
MGSA‘, GenGo‘ true α, β is given,
MGSA: α, β estimated from data

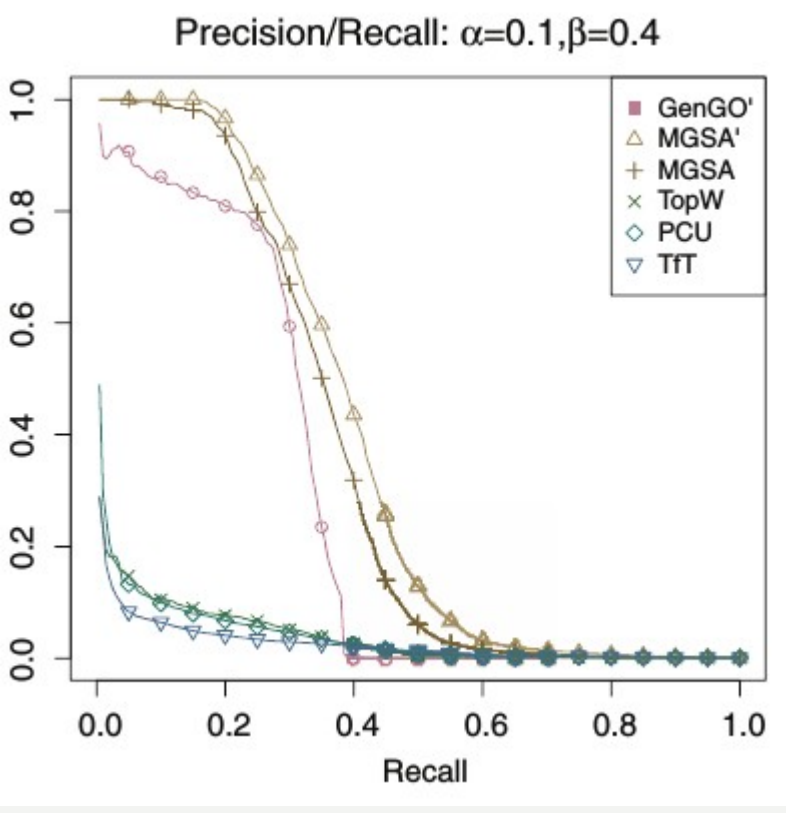
TfT: term form term Fischer‘s Exact
PCU: parent child union
TopW: topological weight

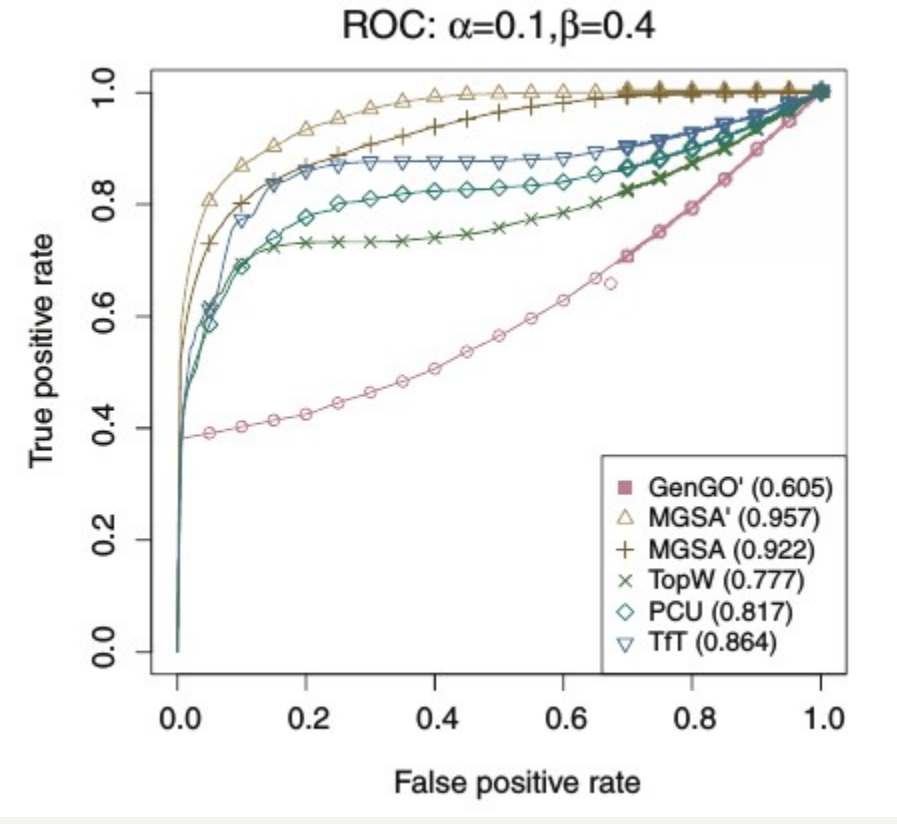


simulation results

α: false positive rate β: false negative rate
MGSA‘, GenGo‘ true α, β is given,
MGSA: α, β estimated from data

TfT: term form term Fischer‘s Exact
PCU: parent child union
TopW: topological weight





Real data example

‚Nigerian data-set‘ (Pickrell 2010) tests: (Rahmatallah 2014, 2016)

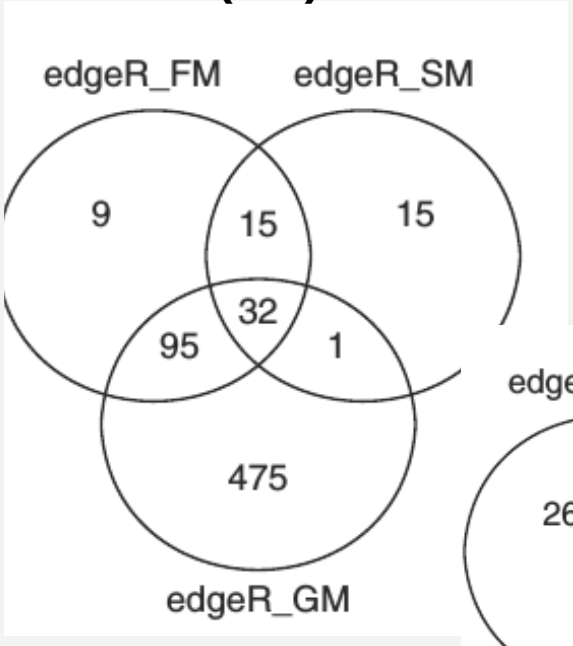
- lymphoblastoid cell lines (LCL)
- 69 Nigerian individuals (use 58 unrelated, **29 male, 29 female**)

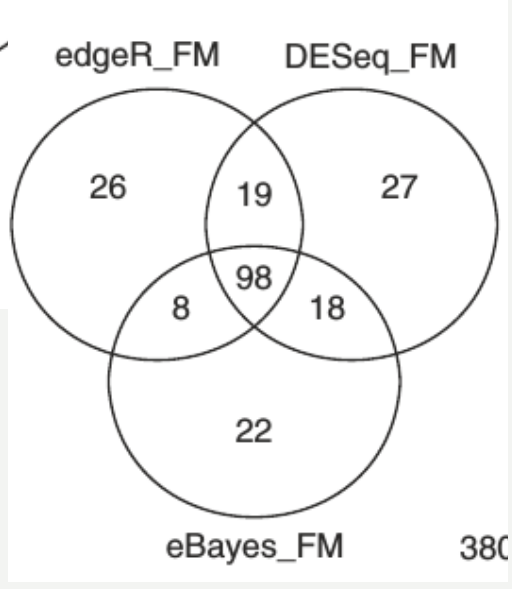
- within same gender allows for FP-check
- „natural“-FP-s: X-linked genes that are not escaping inactivation (**Xi**)
- „natural“-TP-s: genes that are **escaping X-chromosome** inactivation and are therefore over-expressed **in females (XiE: 387 genes)**, and genes that are located on male-specific region of **Y** chromosome and are therefore over-expressed **in males** (**msY**).
- TP: MsigDB: C2 pathway: DISTECHE_ESCAPED_FROM_X_INACTIVATION (**DEX**)

- three TP gene-sets: **XiE, msY, DEX,** one FP**: (Xi)**
- (task is too easy, all tested methods find all TP-s and miss the one FP)

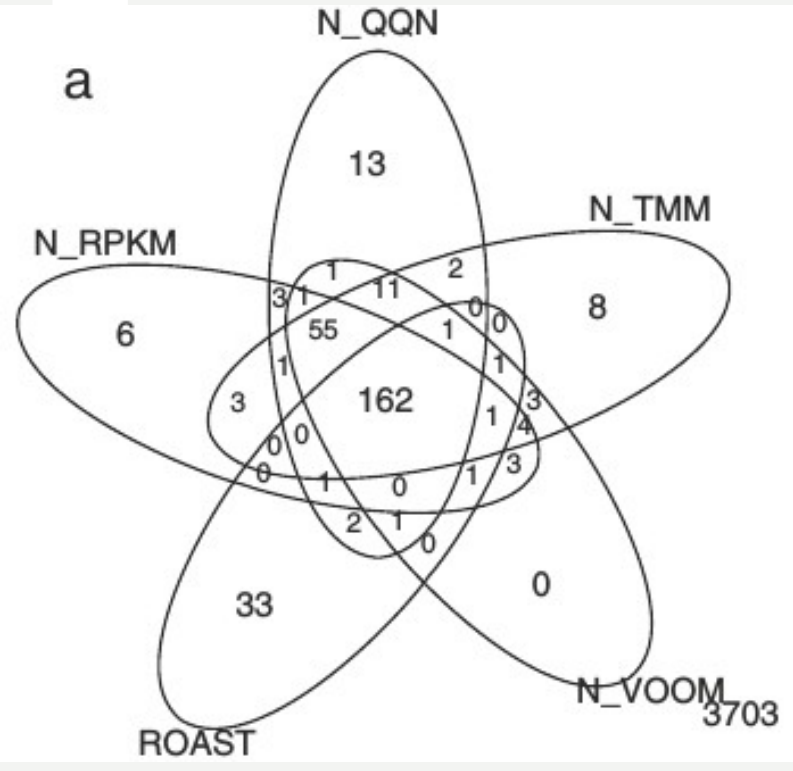
robustness (Nigerian data-set)

p-value combination methods: Gamma (GM), Fischer (FM), Stouffer (SM)





N-statistics with different normalizations



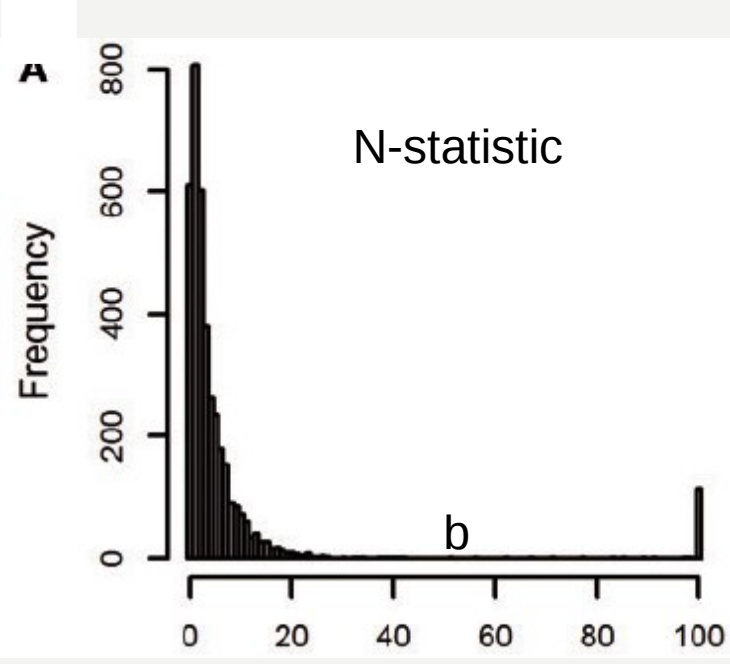
Sample-size robustness (Nigerian data-set)

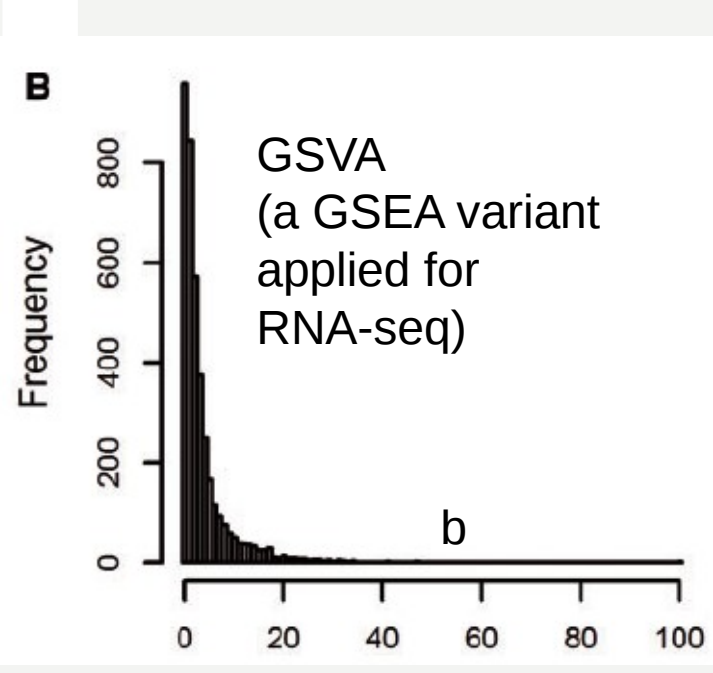
Gene set definitions: C2 from MsigDB (various sources, curated, 3890)

define TP/TN gene-sets from full data-set (58 = 29 + 29 individuals)

create random sub-data-sets for sample sizes [48, 38, 28, 18] (100 data-set each)

- **commonly detecting gene sets in subsets (b).**





sample size = 18

Sample-size robustness (Nigerian data-set)

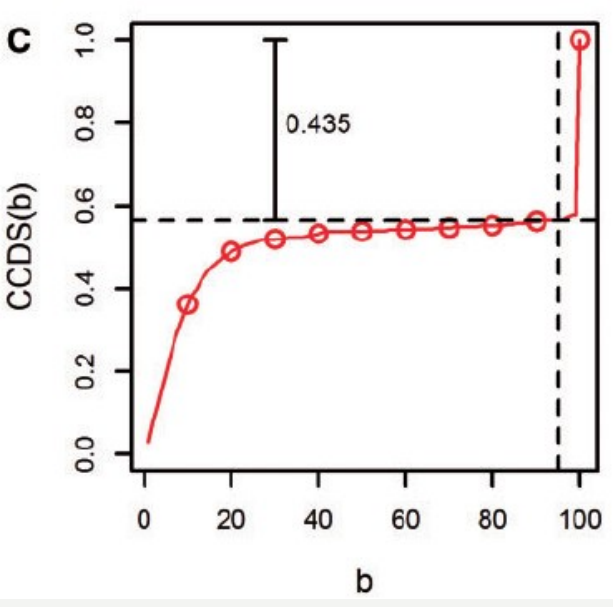
Gene set definitions: C2 from MsigDB (various sources, curated, 3890)

define TP/TN gene-sets from full data-set (58 = 29 + 29 individuals)

create random sub-data-sets for sample sizes [48, 38, 28, 18] (100 data-set each)

- **CCDS: cumulative common detection per subset**

N - statistic



Q is the sum of the numbers of detected gene sets in all b (100)

b between 1 and 100

$$CCDS(b) = \frac{1}{Q} \sum_{k=1}^{b} k \; s_k$$

among the 26.844 gene sets detected by N-statistic in all 100 subsets, 43.5% of them were commonly detected in at least 95% of all subsets

sample size = 18

Sample-size robustness (Nigerian data-set)

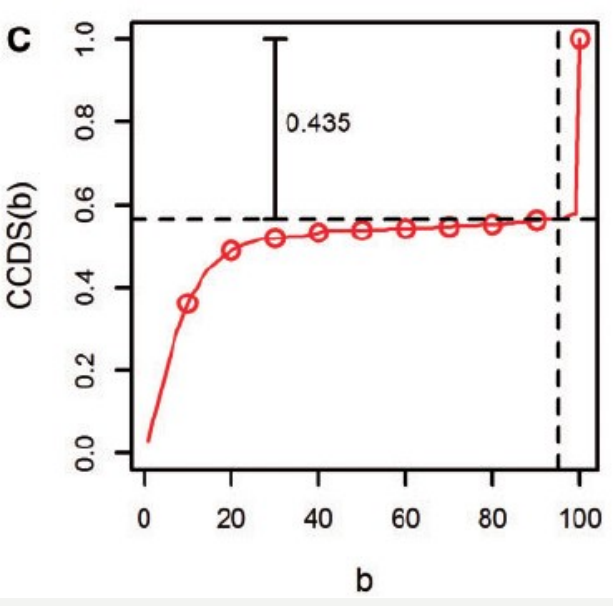
Gene set definitions: C2 from MsigDB (various sources, curated, 3890)

define TP/TN gene-sets from full data-set (58 = 29 + 29 individuals)

create random sub-data-sets for sample sizes [48, 38, 28, 18] (100 data-set each)

- **CCDS: cumulative common detection per subset**

N - statistic



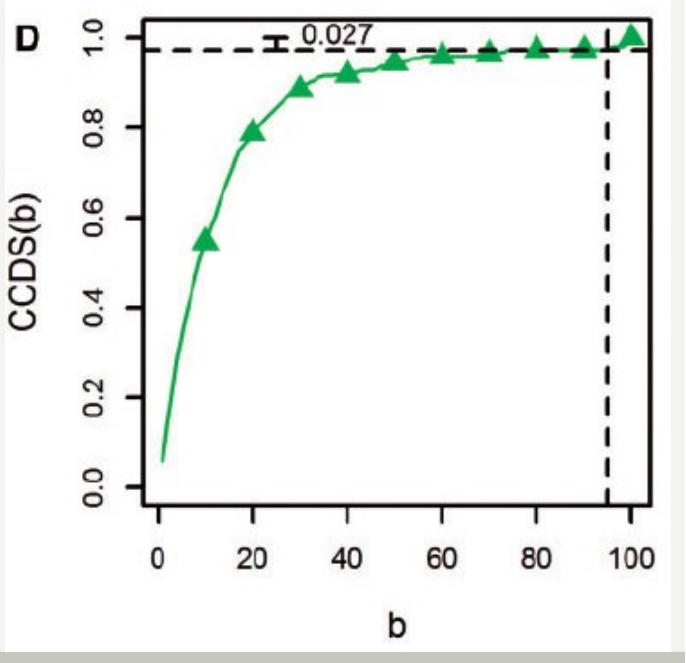
Q is the sum of the numbers of detected gene sets in all b (100)

b is between 1 and 100

$$CCDS(b) = \frac{1}{Q} \sum_{k=1}^{b} k \; s_k$$

sample size = 28

GSVA



Sample-size robustness (Nigerian data-set)

Gene set definitions: C2 from MsigDB (various sources, curated, 3890)

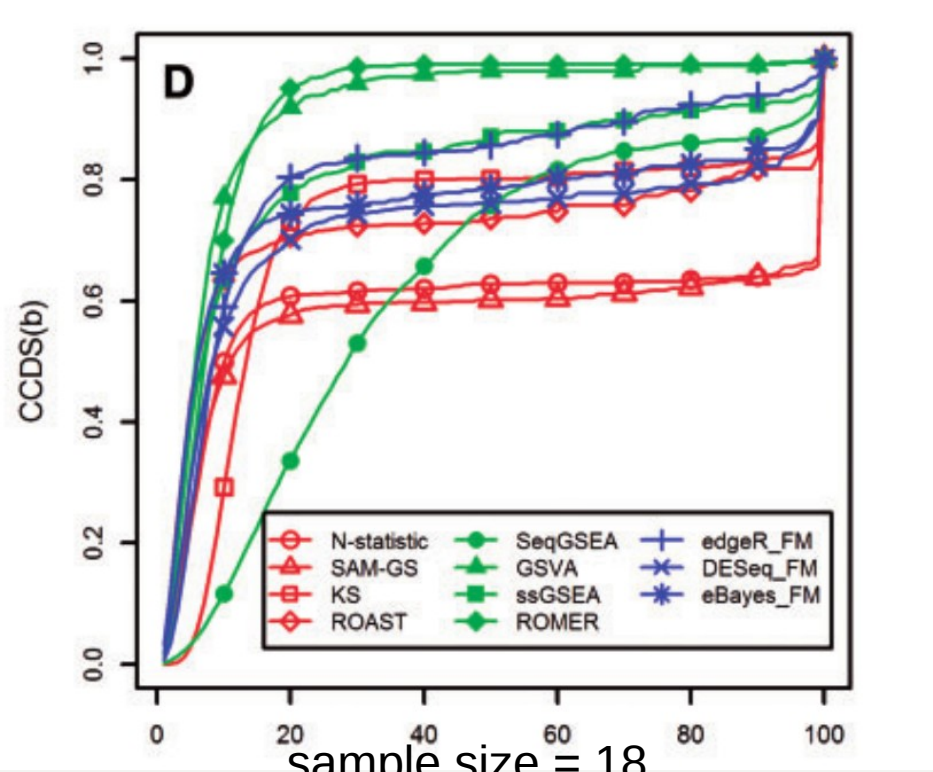
define TP/TN gene-sets from full data-set (58 = 29 + 29 individuals)

create random sub-data-sets for sample sizes [48, 38, 28, 18] (100 data-set each)

- **CCDS: cumulative common detection per subset**

Q is the sum of the numbers of detected gene sets in all b (100)
b is between 1 and 100

$$CCDS(b) = \frac{1}{Q} \sum_{k=1}^{b} k \; s_k$$



sample size = 18

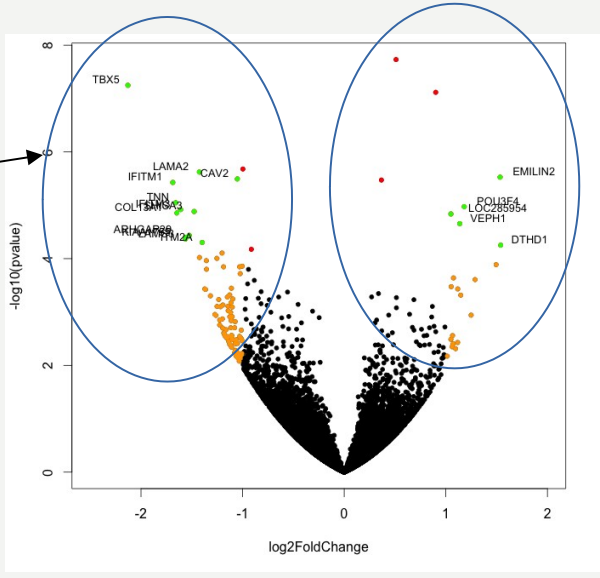
issues

Things to have in mind if performing / reading about set enrichment

what is the background?

	in set	not in set
significant DE	**a**	**b**
non-significant DE	**c**	**d**

a is defined but what is **b, c, d**?



what about genes **without any signal** (read-count)?

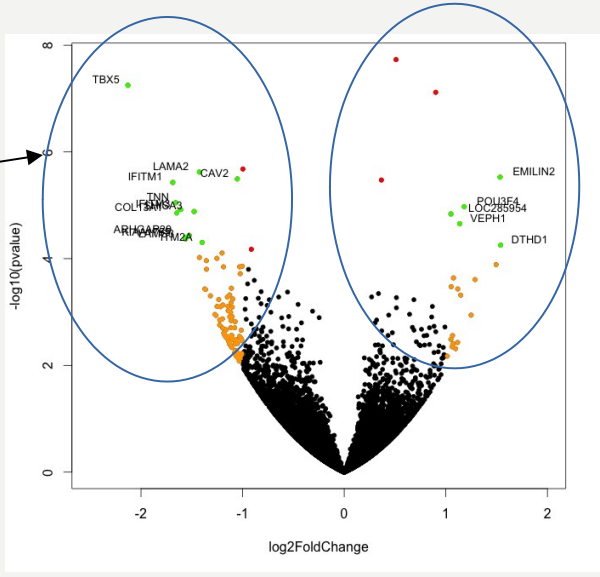
issues

Things to have in mind if performing / reading about set enrichment

what is the background?

	in set	not in set
significant DE	**a**	**b**
non-significant DE	**c**	**d**

a is defined but what is **`b, c, d`**?



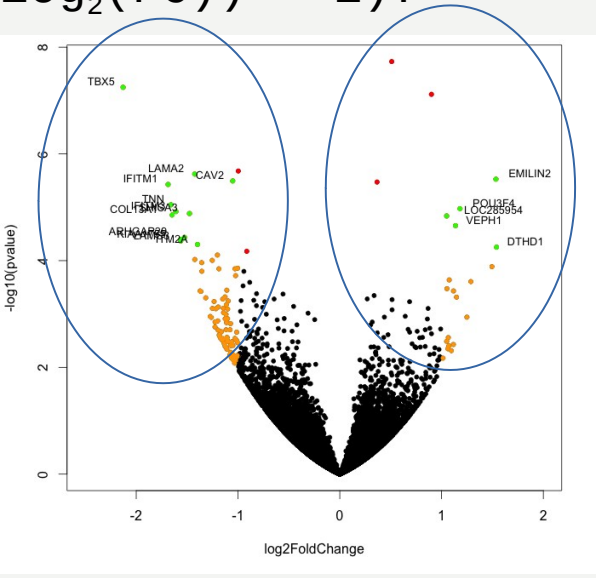
what about genes **without any signal** (count)?

not-measured → might be DE or not DE → one should not consider them

what about genes **not DE** (e.g. `pval >= 0.05 || abs(log`$_2$`(FC)) < 1`)?

issues

what about genes **not DE** (e.g. `pval >= 0.05 || abs(log2(FC)) < 1`)?



the p-val only says something about the H_0 (gene is unchanged between phenotypes) if rejected, if not it might be for example:

- **clearly not changing** (would be ok as non-signif-DE, FC for KS-test would be appropriate (note: this is not tested by default at all!)
- **too low signals** to make a clear difference → not appropriate for **`b, c, d`** (FC for KS would be also invalid)
- **high variance within replicates**, unclear what happens → not appropriate for **`b, c, d`** (FC for KS would be also invalid)

issues

Things to have in mind if performing / reading about set enrichment

especially problem with the **users** of DAVID:

- DAVID enables loading background but works with defaults - in many cases background inappropriate
- ID-mapping affects results
- many possible enrichment/clustering possibilities, default is appropriate?

pro & contra

List based enrichment (ORA):

- (+) easy to use, even for complex set of „genes of interest“ (e.g. up-regulated in expression, targets of TF X, and in chromatin state Y…) → but what is the appropriate background?
- (+) quick (simple calculations) (but only if no permutation tests)
- (-) cutoff thresholds to define significants ad-hoc, can be very unstable
- (-) misses systematic but small effects

Distribution based (e.g. KS):

- (+) uses the experimental values – more information
- (+) no cutoff
- (-) slower
- (-) not possible for non genome-wide data (no background distribution to compare) or in a complex integrated study
- (-) big changes driving the statistic → assumption more changing more important (not necessarily true)

many methods

- in a review from 2009 from the DAVID authors **68** methods are discussed
  - (see literature slides)

- there are methods exploiting the gene networks for enrichment e.g.:
  - SPIA (Tarca, 2009)
  - GGEA (2011)
  - NEA (Alexeyenko, 2012)
  - RelExplain (Berchtold, 2017)

Questions?

references

Gene ontology

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock.
Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.
Nat. Genet., 25(1):25-29, May 2000.

No authors listed.
Expansion of the Gene Ontology knowledgebase and resources.
Nucleic Acids Res., 45(D1):D331-D338, Jan 2017.

DAVID

B. T. Sherman, d. a. W. Huang, Q. Tan, Y. Guo, S. Bour, D. Liu, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki.
DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis.
BMC Bioinformatics, 8:426, Nov 2007.

references

DAVID

D. W. Huang, B. T. Sherman, Q. Tan, J. R. Collins, W. G. Alvord, J. Roayaei, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki.
The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists.
Genome Biol., 8(9):R183, 2007.

Review of (68!) methods from the DAVID authors:

d. a. W. Huang, B. T. Sherman, and R. A. Lempicki.
Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.
Nucleic Acids Res., 37(1):1-13, Jan 2009.

GSEA

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov.
Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.
Proc. Natl. Acad. Sci. U.S.A., 102(43):15545-15550, Oct 2005.

Methods using the GO-structure / overlaps

A. Alexa, J. Rahnenfuhrer, and T. Lengauer.
Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.
Bioinformatics, 22(13):1600-1607, Jul 2006.

S. Grossmann, S. Bauer, P. N. Robinson, and M. Vingron.
Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis.
Bioinformatics, 23(22):3024-3031, Nov 2007.

Y. Lu, R. Rosenfeld, I. Simon, G. J. Nau, and Z. Bar-Joseph.
A probabilistic generative model for GO enrichment analysis.
Nucleic Acids Res., 36(17):e109, Oct 2008.

S. Bauer, J. Gagneur, and P. N. Robinson.
GOing Bayesian: model-based gene set analysis of genome-scale data.
Nucleic Acids Res., 38(11):3523-3532, Jun 2010.

references

Comparative Analysis / Evaluation of Methods for RNA-seq

Y. Rahmatallah, F. Emmert-Streib, and G. Glazko.
Comparative evaluation of gene set analysis approaches for RNA-Seq data.
BMC Bioinformatics, 15:397, 2014.

Y. Rahmatallah, F. Emmert-Streib, and G. Glazko.
Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline.
Brief. Bioinformatics, Sep 2015.

Network based enrichment methods

A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J. S. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero.
A novel signaling pathway impact analysis.
Bioinformatics, 25(1):75-82, Jan 2009.

L. Geistlinger, G. Csaba, R. Kuffner, N. Mulder, and R. Zimmer.
From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems.
Bioinformatics, 27(13):i366-373, Jul 2011.

A. Alexeyenko, W. Lee, M. Pernemalm, J. Guegan, P. Dessen, V. Lazar, J. Lehtio, and Y. Pawitan.
Network enrichment analysis: extension of gene-set enrichment analysis to gene networks.
BMC Bioinformatics, 13:226, Sep 2012.

E. Berchtold, G. Csaba, and R. Zimmer.
RelExplain-integrating data and networks to explain biological processes.
Bioinformatics, 33(12):1837-1844, Jun 2017.