CBIO Computational Biology @ UCT

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Practical Course on "Gene/Protein Functional Networks & Interactomes"

23 and 24 November 2015 - UCT, Cape Town, South Africa

(thanks to Prof. Nicola Mulder)

Dr. Javier De Las Rivas
Cancer Research Center (CiC-IBMCC, CSIC/USAL), Salamanca, Spain

CBIO Computational Biology @ UCT

CiC

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

DAY 1

Session 1 (9:30 - 12:30, 3h)

Bioinformatic tools for Functional Enrichment Analysis (FEA)

Session 2 (13:30 - 16:30, 3h)

Construction of gene functional networks

DAY2

Session 3 (9:30 - 12:30, 3h)

Protein interaction networks

Session 4 (13:30 - 16:30, 3h)

Construction and analysis of gene/protein networks

Dr. Javier De Las Rivas
Cancer Research Center (CiC-IBMCC, CSIC/USAL), Salamanca, Spain

CBIO Computational Biology @ UCT

CiC

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Session 1 (9:30 - 12:30, 3h)
Bioinformatic tools for Functional Enrichment Analysis (FEA)
Session 2 (13:30 - 16:30, 3h)
Construction of gene functional networks

- Introduction to biological information and annotation spaces: GO, KEGG, Interpro

- Functional Enrichment Analysis (EA): from single to modular methods

- Using EA tools to annotate gene lists: DAVID (single), GSEA (gene sets), GeneCodis (modular)

- Sort out problems after EA: post-enrichment tool GeneTermLinker (postEA)

- From co-annotation and enrichment to functional networks: networks construction using a R tool

Session 1 (9:30 - 12:30, 3h)
Bioinformatic tools for Functional Enrichment Analysis (FEA)
Session 2 (13:30 - 16:30, 3h)
Construction of gene functional networks

- Introduction to biological information and annotation spaces: GO, KEGG, Interpro

- Functional Enrichment Analysis (EA): from single to modular methods

- Using EA tools to annotate gene lists: DAVID (single), GSEA (gene sets), GeneCodis (modular)

- Sort out problems after EA: post-enrichment tool GeneTermLinker (postEA)

- From co-annotation and enrichment to functional networks: networks construction using a R tool

Many biological data resources (= annotation spaces)

Genome Annotation

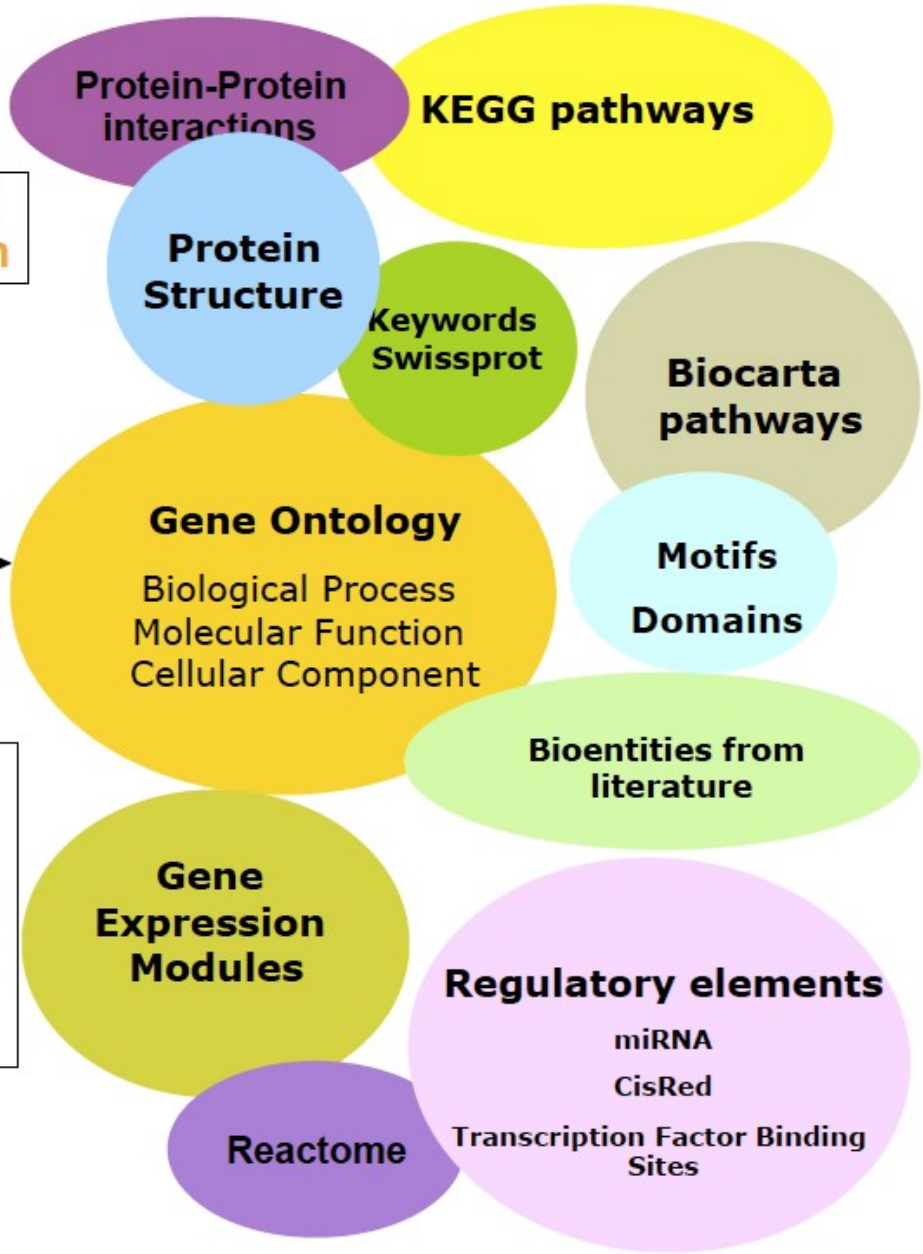
Structural Annotation

Functional Annotation

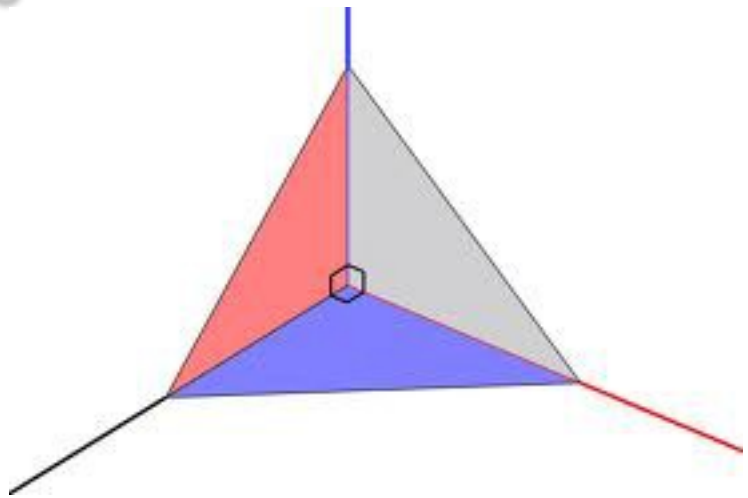
Biological Databases

There are:
- A LOT OF biomolecular databases/sources
- A LOT OF overlap of information/redundancy
- A LOT OF TOOLS

large increase in the "omics" era !!!



Biological information and annotation spaces: GO, KEGG, Interpro



3 orthogonal annotation spaces

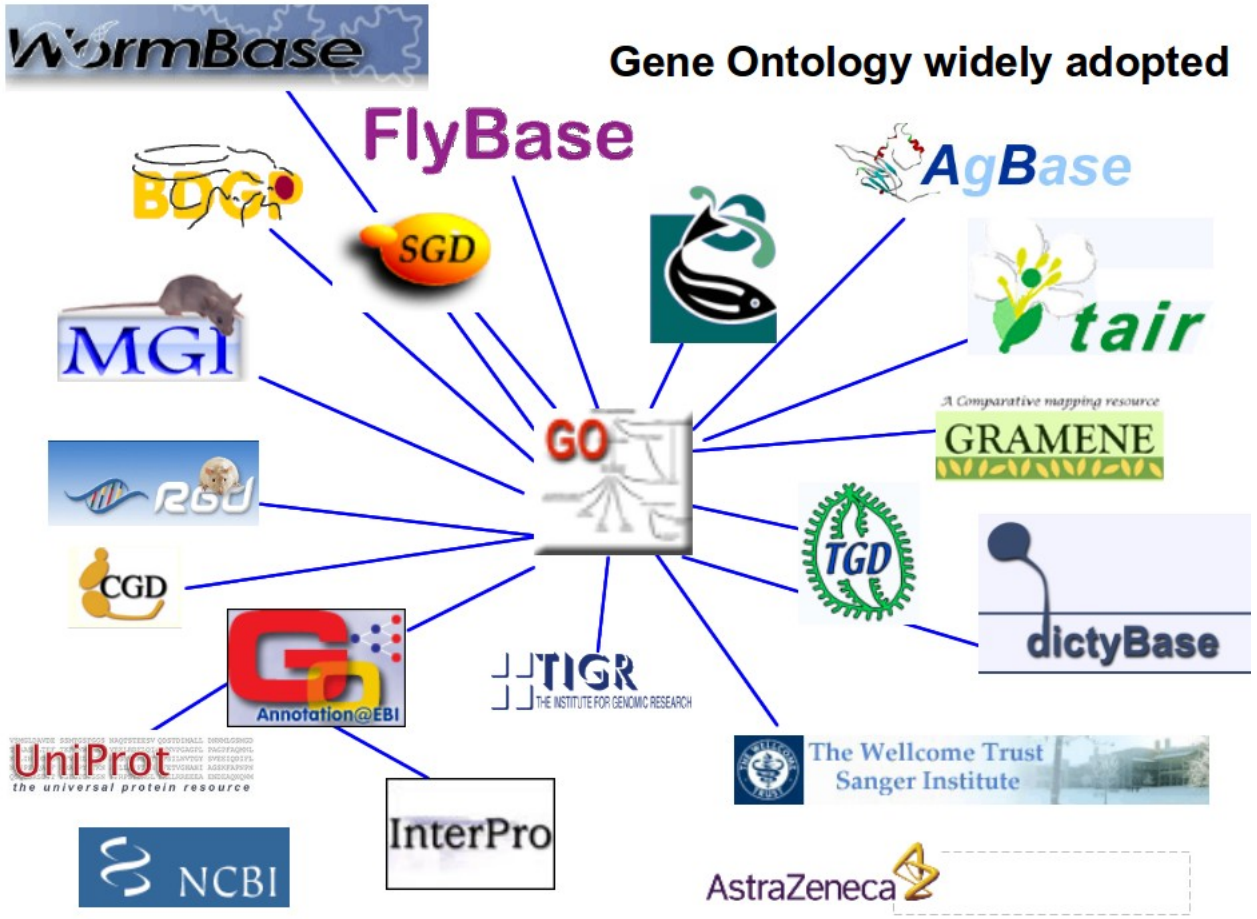
Biological terms and definitions
Gene Ontology (GO) http://www.geneontology.org/

Biomolecular pathways and reactions
KEGG (pathways) http://www.genome.jp/kegg/

Biomolecular sequences and structure (domains, motifs)
InterPro (sequences & structure) http://www.ebi.ac.uk/interpro/

Biological information and annotation spaces: GO, KEGG, Interpro

Gene Ontology (GO) http://www.geneontology.org/

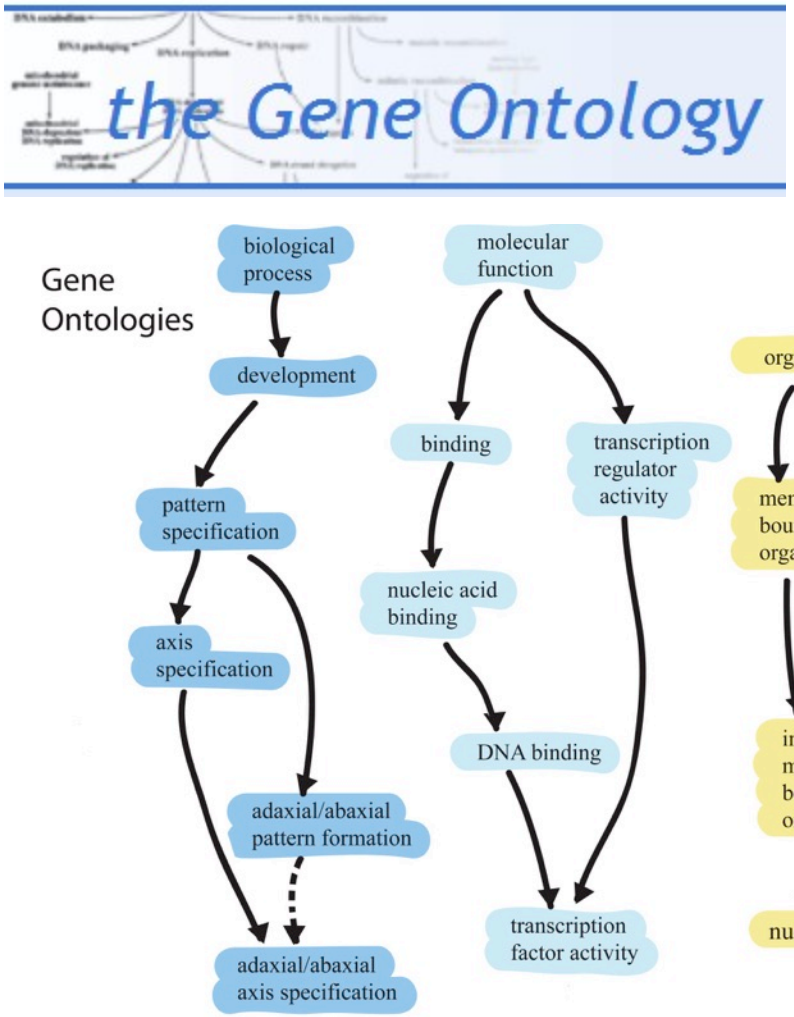


The **Gene Ontology** project is a major bioinformatics initiative with the aim of **standardizing the representation of gene and gene product attributes** across species and databases.

The project provides a **controlled vocabulary of biological terms** for describing gene product **characteristics** and gene product **annotation** data from GO Consortium members, as well as **tools** to access and process this data.

Biological information and annotation spaces: GO, KEGG, Interpro

Gene Ontology (GO) http://www.geneontology.org/



*Modified from **Clark et al. (2005)***

GO is organized in 3 independent hierarchies:

- **Biological Process Ontology (BP)**

Biological and cellular processes where a given gene or gene product is involved.

- **Molecular Function Ontology (MF)**

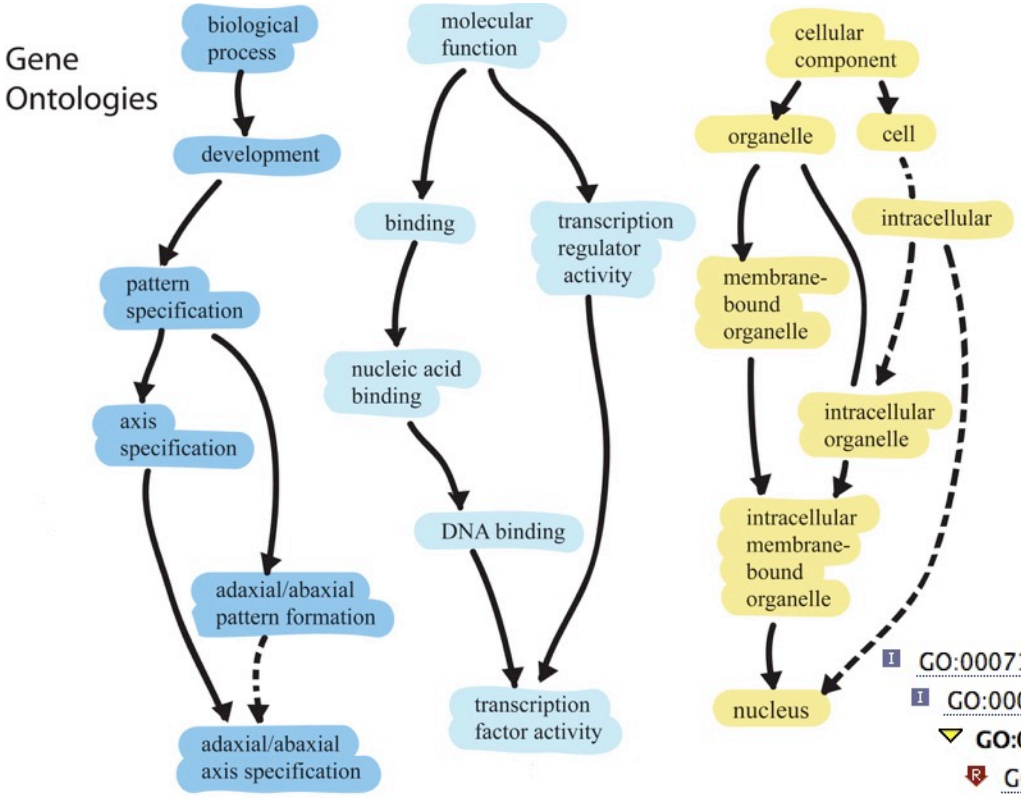
Molecular functions and activities that a given gene or gene product has.

- **Cellular Component Ontology (CC)**

Place or part of the cell where a given gene or gene product works most of its time.

Biological information and annotation spaces: GO, KEGG, Interpro

Gene Ontology (GO) http://www.geneontology.org/



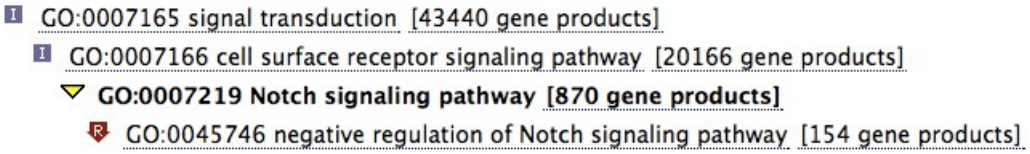
GO is organized in 3 independent hierarchies:

- **Biological Process Ontology (BP)**

Example, search for:
"NOTCH" or "NOTCH signaling"

Find:
GO:0007219 Notch signaling pathway

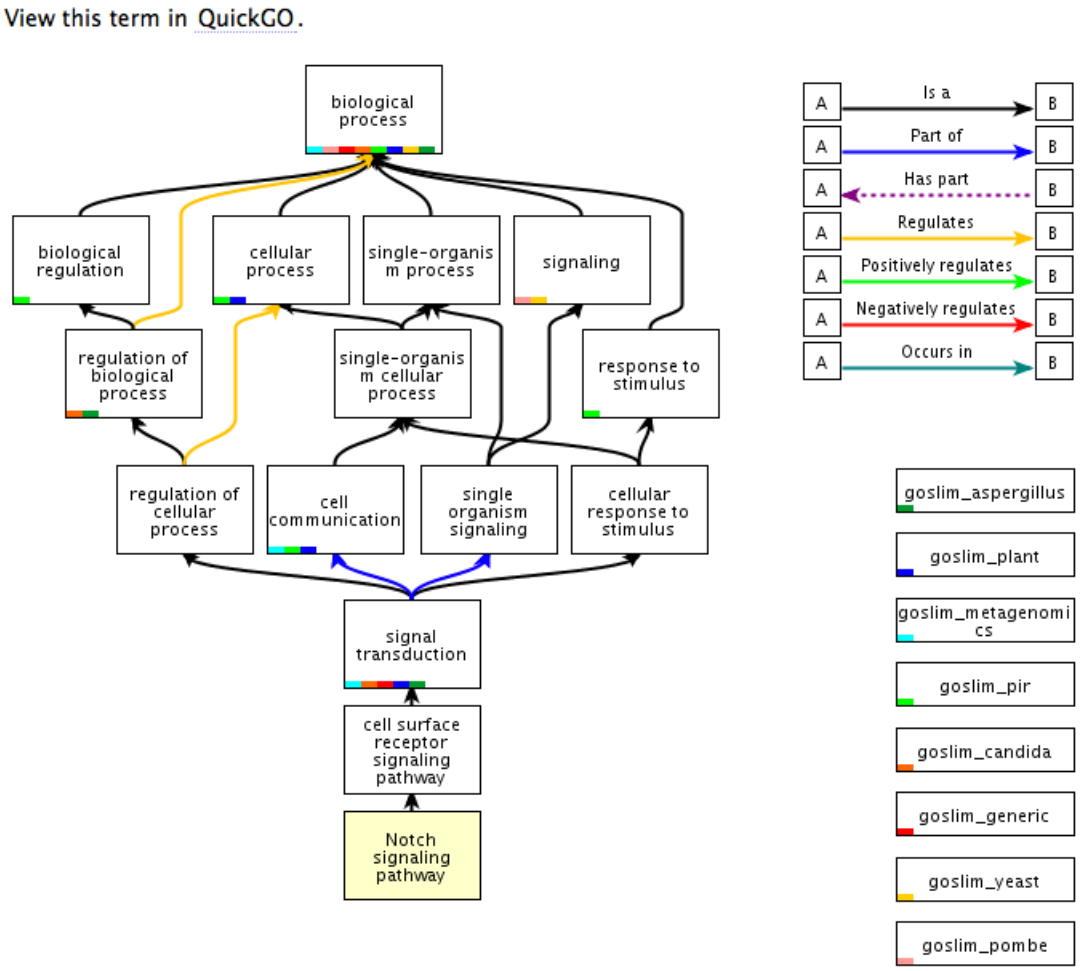
GO includes
ancestors & **children** terms



*Modified from **Clark et al. (2005)***

Biological information and annotation spaces: GO, KEGG, Interpro

Gene Ontology (GO) http://www.geneontology.org/



GO is organized in 3 independent hierarchies:

- **Biological Process Ontology (BP)**

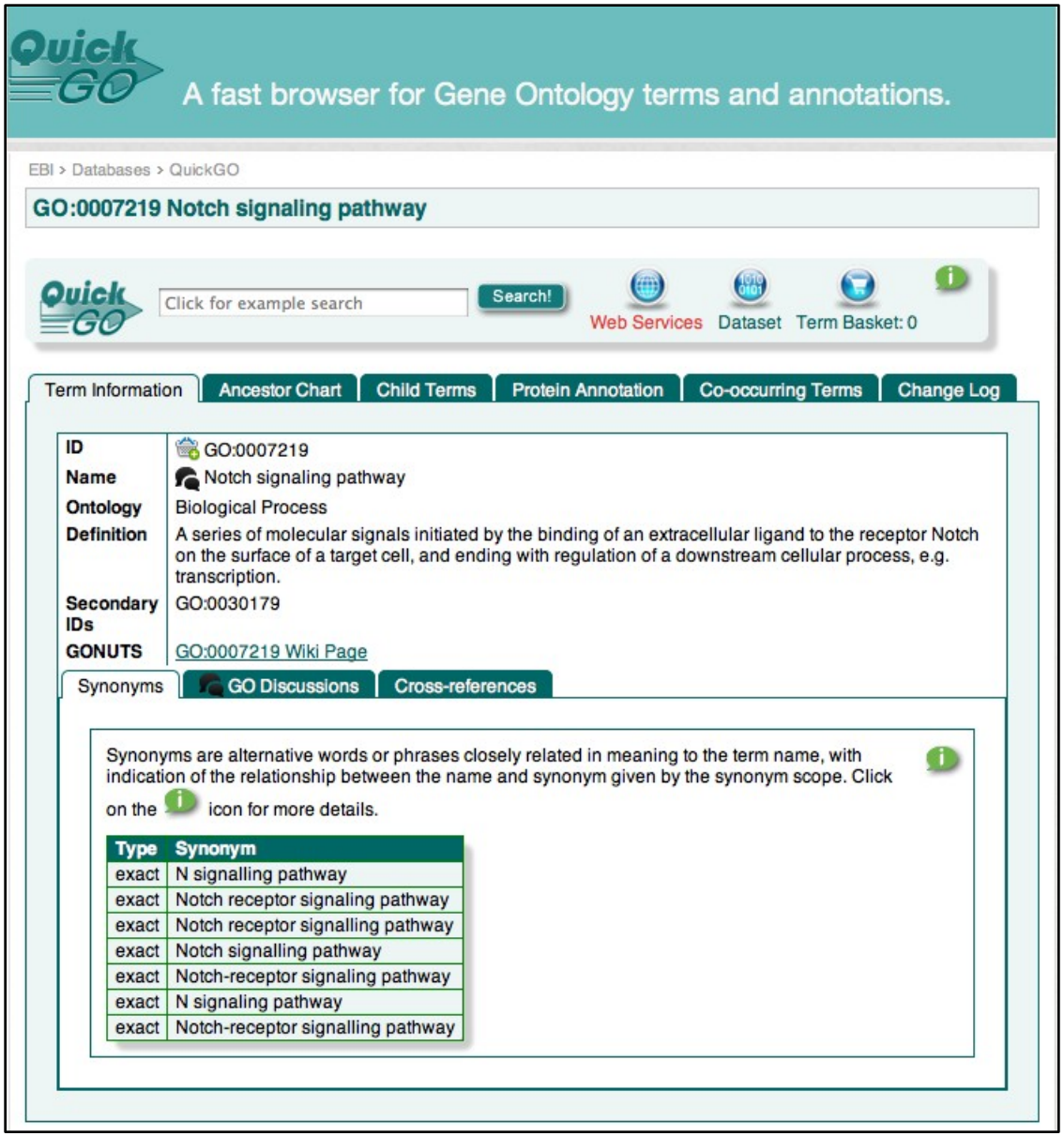
Example, search for:
"NOTCH" or "NOTCH signaling"

Find:
GO:0007219 Notch signaling pathway

GO includes
ancestors & **children** terms

2187 gene products
assigned in all species (Oct.2014)
1661 gene products
assigned in mammalia (Oct.2014)

Biological information and annotation spaces: GO, KEGG, Interpro



http://www.ebi.ac.uk/QuickGO/

GO is organized in 3 independent hierarchies:

- **Biological Process Ontology (BP)**

Example, search for:
"NOTCH" or "NOTCH signaling"

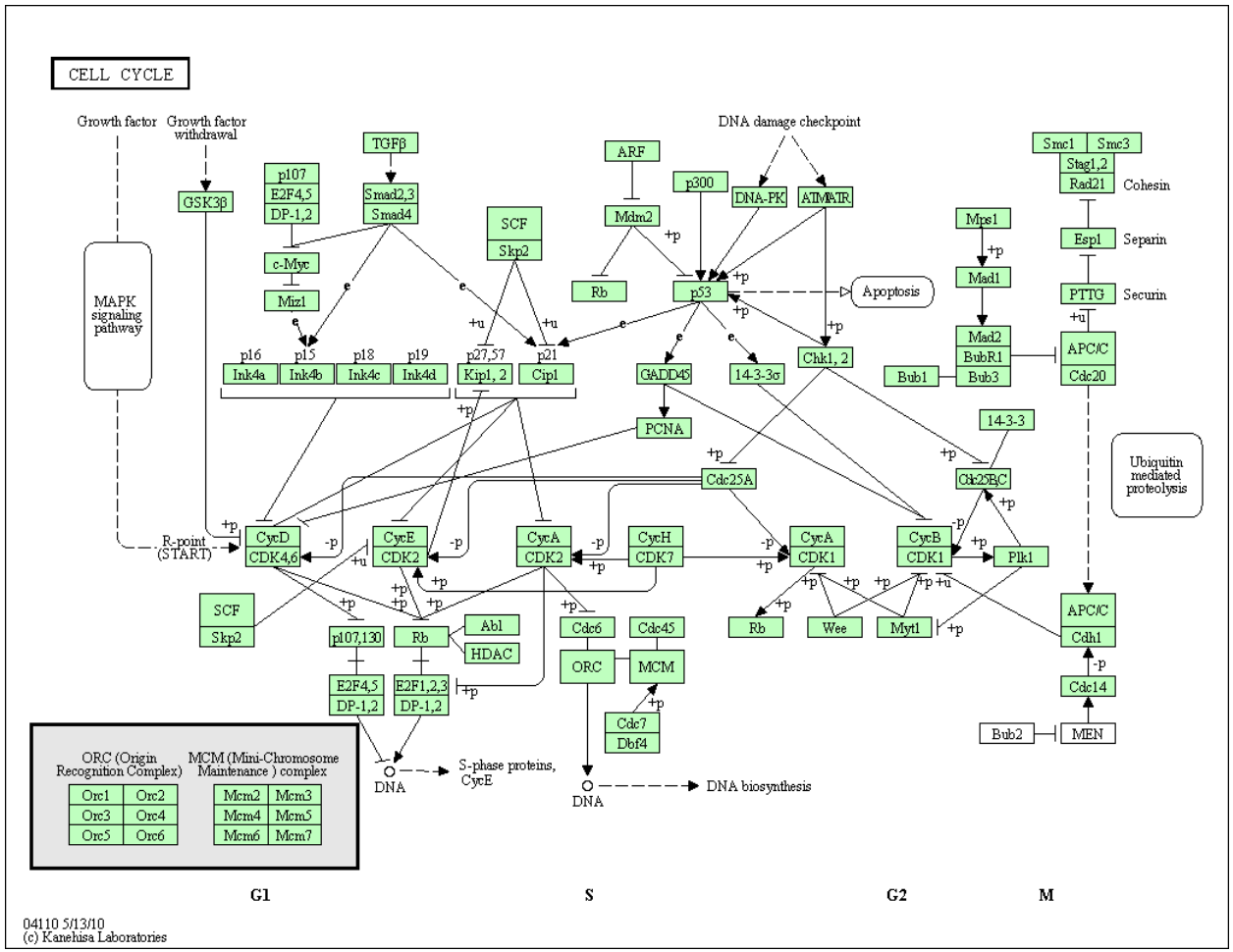
Find:
GO:0007219 Notch signaling pathway

GO includes
ancestors & **children** terms

2187 gene products
assigned in all species (Oct.2014)
1661 gene products
assigned in mammalia (Oct.2014)

Biological information and annotation spaces: GO, KEGG, Interpro

KEGG (pathways) http://www.genome.jp/kegg/



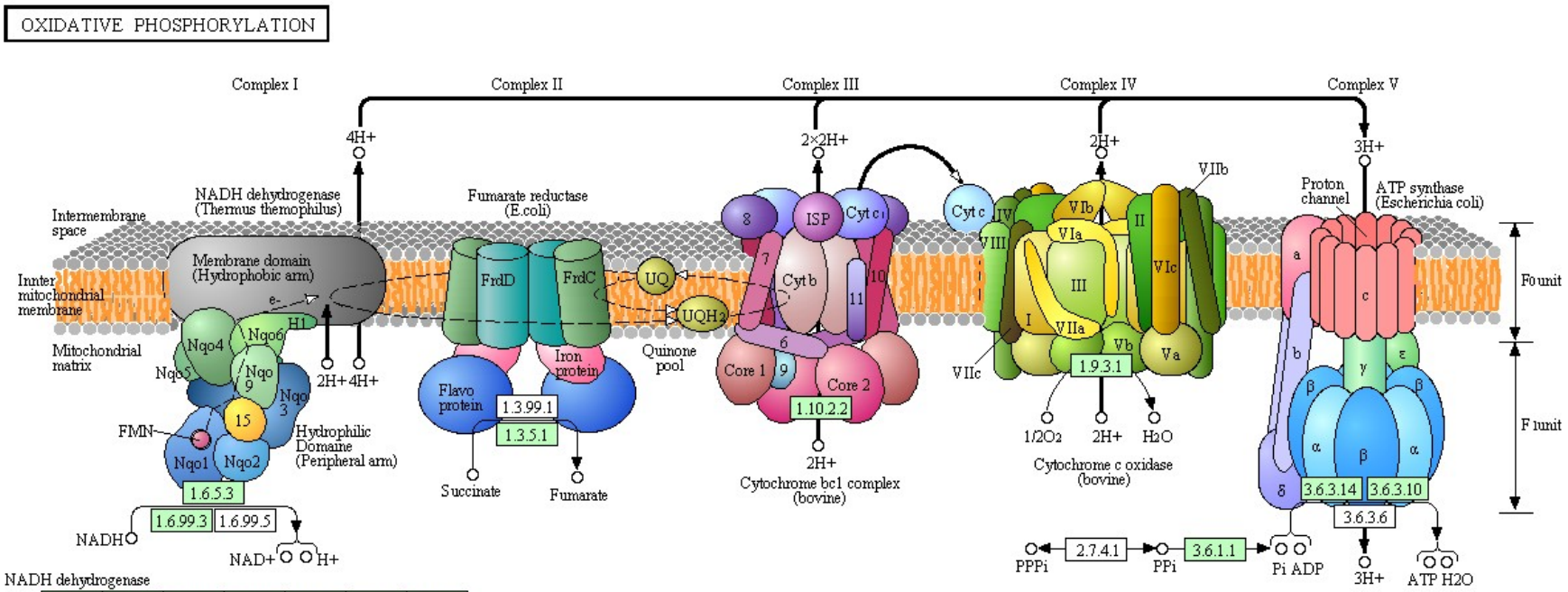
KEGG PATHWAY is a collection of **manually** drawn **pathway maps** representing our **knowledge** on the molecular **interaction** and **reaction** networks

1. Metabolism
2. Genetic Information Processing
3. Environmental Informat. Processing
4. Cellular Processes

5. Organismal Systems
6. Human Diseases

7. Drug Development

Biological information and annotation spaces: GO, KEGG, Interpro



NADH dehydrogenase

E	ND1	ND2	ND3	ND4	ND4L	ND5	ND6
E	Ndufs1	Ndufs2	Ndufs3	Ndufs4	Ndufs5	Ndufs6	Ndufs7	Ndufs8	Ndufv1	Ndufv2	Ndufv3
B/A	NuoA	NuoB	NuoC	NuoD	NuoE	NuoF	NuoG	NuoH	NuoI	NuoJ	NuoK	NuoL	NuoM	NuoN
B/A	NdhC	NdhK	NdhJ	NdhH	NdhA	NdhI	NdhG	NdhE	NdhF	NdhD	NdhB	NdhL	NdhM	NdhN	HoxE	HoxF	HoxU
E	Ndufa1	Ndufa2	Ndufa3	Ndufa4	Ndufa5	Ndufa6	Ndufa7	Ndufa8	Ndufa9	Ndufa10	Ndufab1	Ndufa11	Ndufa12	Ndufa13
E	Ndufb1	Ndufb2	Ndufb3	Ndufb4	Ndufb5	Ndufb6	Ndufb7	Ndufb8	Ndufb9	Ndufb10	Ndufb11	Ndufc1	Ndufc2

Succinate dehydrogenase / Fumarate reductase

E	SDHC	SDHD	SDHA	SDHB
B/A	SdhC	SdhD	SdhA	SdhB
			FrdA	FrdB	FrdC	FrdD

Cytochrome c reductase

E/B/A	ISP	Cyt b	Cyt 1							
E				COR1	QCR2	QCR6	QCR7	QCR8	QCR9	QCR10

Cytochrome c oxidase

E	COX10		COX3	COX1	COX2	COX4	COX5A	COX5B	COX6A	COX6B	COX6C	COX7A	COX7B	COX7C	COX8	E/B/A: COX11	COX15	COX17
B/A	CyoE	CyoD	CyoC	CyoB	CyoA													
		CoxD	CoxC	CoxA	CoxB													
		QoxD	QoxC	QoxB	QoxA													

Cytochrome c oxidase , cbb3-type

B	I	II	IV	III

Cytochrome bd complex

B/A	CydA	CydB

F-type ATPase (Bacteria)

beta	alpha	gamma	delta	epsilon	c	a	b

F-type ATPase (Eukaryotes)

beta	alpha	gamma	OSCP	delta	epsilon	c	a
b	e	f6	f	8			
d	f	h	j	k	g		

V-type ATPase (Prokaryotes)

A	B	C	D	E	F	I	K

V-type ATPase (Eukaryotes)

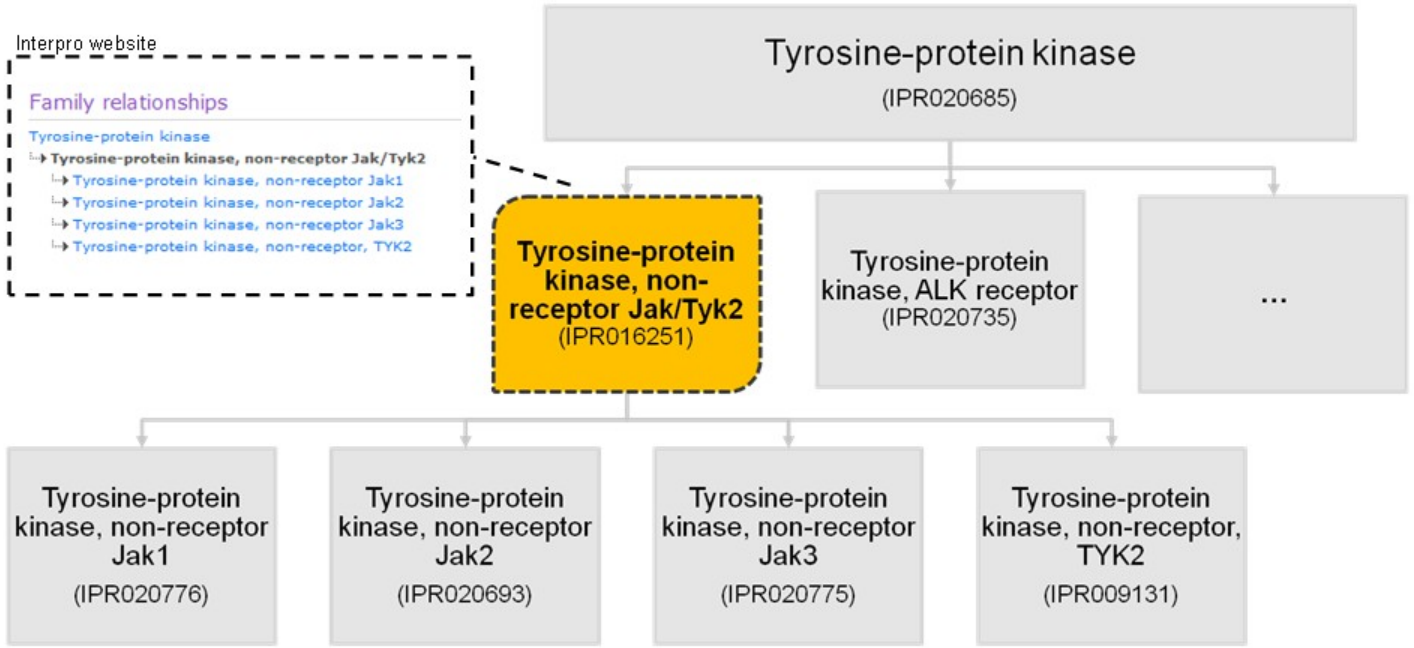
A	B	C	D	E	F	G	H
I	AC39	54kD	S1	lipid			

00190 6/1/12
(c) Kanehisa Laboratories

Biological information and annotation spaces: GO, KEGG, Interpro

InterPro (sequences & structure) http://www.ebi.ac.uk/interpro/

InterPro provides **sequence/structure analysis** and classification of proteins into families using found protein **motifs**, protein **domains** and important **sites**. They combine protein signatures from a number of member databases into a single searchable resource to produce a integrated database and diagnostic tool.



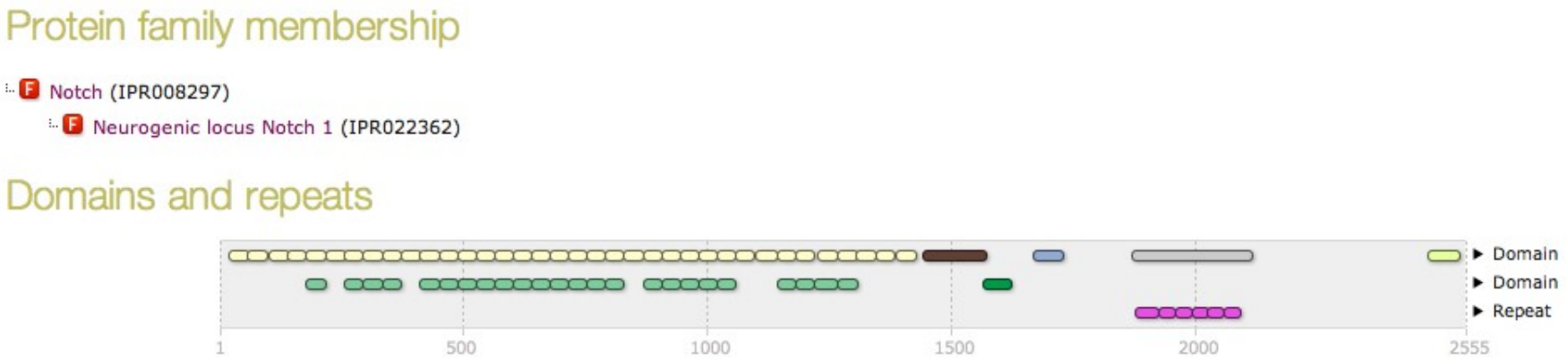
Biological information and annotation spaces: GO, KEGG, Interpro

InterPro (sequences & structure) http://www.ebi.ac.uk/interpro/

InterPro provides **sequence/structure analysis** and classification of proteins into families using found protein **motifs**, protein **domains** and important **sites**. They combine protein signatures from a number of member databases into a single searchable resource to produce a integrated database and diagnostic tool.

InterPro
Protein sequence analysis & classification

Search for NOTCH1 sequence: **NOTC1_HUMAN** (2555 aa)



CBO Computational Biology @ UCT

CiC

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Hands-on: Practical Examples

Explore web sites: GO, KEGG, InterPro (& Pfam)

(NOTCH and NOTCH signaling)

Session 1 (9:30 - 12:30, 3h)
Bioinformatic tools for Functional Enrichment Analysis (FEA)
Session 2 (13:30 - 16:30, 3h)
Construction of gene functional networks

- Introduction to biological information and annotation spaces: GO, KEGG, Interpro

- Functional Enrichment Analysis (EA): from single to modular methods

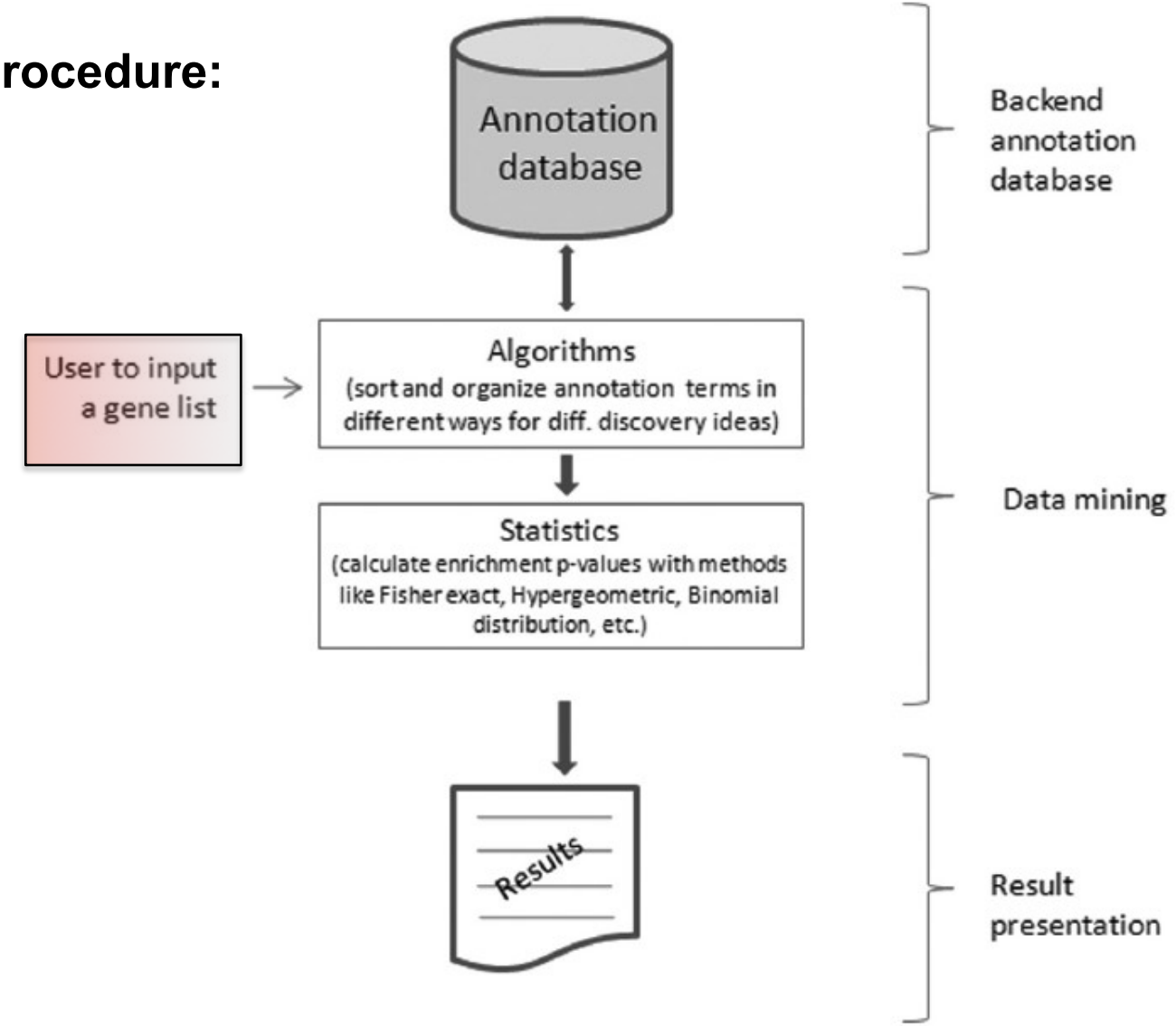
- Using EA tools to annotate gene lists: DAVID (single), GSEA (gene sets), GeneCodis (modular)

- Sort out problems after EA: post-enrichment tool GeneTermLinker (postEA)

- From co-annotation and enrichment to functional networks: networks construction using a R tool

Functional Enrichment Analysis (EA)

general procedure:



Functional Enrichment Analysis (EA)

Enrichment tool name	Year of release	Key statistical method	Category
FunSpec	2002	Hypergeometric	Class I
Onto-express	2002	Various chi-square, Fisher's exact	Class I
EASE	2003	Modified Fisher's exact (EASE score)	Class I
FatiGO/FatiWise/FatiGO +	2003	Fisher's exact; family-wise error rate	Class I
FuncAssociate	2003	Hypergeometric	Class I
GARBAN	2003	Binomial	Class I
GeneMerge	2003	Hypergeometric	Class I
GoMiner	2003	Hypergeometric	Class I
MAPPFinder	2003	Hypergeometric	Class I
CLENCH	2004	Unclear	Class I
GO::TermFinder	2004	Hypergeometric	Class I
GOAL	2004	Hypergeometric	Class I
GOArray	2004	Hypergeometric	Class I
GOStat	2004	Fisher's exact; chi-square	Class I
GoSurfer	2004	Hypergeometric	Class I
OntologyTraverser	2004	Hypergeometric	Class I
THEA	2004	Binomial; hypergeometric	Class I
BiNGO	2005	Hypergeometric	Class I
FACT	2005	Hypergeometric	Class I
gfinder	2005	Hypergeometric	Class I
Gobar	2005	Hypergeometric	Class I
GOCluster	2005	Fisher's exact	Class I
GOSSIP	2005	Hypergeometric	Class I
L2L	2005	Hypergeometric	Class I
WebGestalt	2005	Hypergeometric	Class I
BayGO	2006	Bayesian	Class I
eGOn/GeneTools	2006	Hypergeometric	Class I
Gene Class Expression	2006	Hypergeometric	Class I
GOALIE	2006	Hypergeometric	Class I
g:Profiler	2007	Hypergeometric	Class I
ProbCD	2007	Yule's Q; Goodman-Kruskal's gamma; Cramer's T	Class I
GOEAST	2008	Hypergeometric	Class I
GOHyperGAll	2008	Hypergeometric	Class I
CatMap	2004	Permutations	Class II
Godist	2004	Kolmogorov–Smirnov test	Class II
GO-Mapper	2004	Gaussian distribution; EQ-score	Class II
iGA	2004	Permutations; hypergeometric; *t*-test; *Z*-score	Class II
GSEA	2005	Kolmogorov–Smirnov-like statistic	Class II
MEGO	2005	*Z*-score	Class II
PAGE	2005	*Z*-score	Class II
T-profiler	2005	*t*-Test	Class II
FuncCluster	2006	Fisher's exact	Class II
FatiScan	2007	Fisher's Exact	Class II
FINA	2007	Fisher's exact	Class II
GAzer	2007	*Z*-statistics; permutation	Class II
GeneTrail	2007	Hypergeometric; Kolmogorov–Smirnov	Class II
MetaGP	2007	*Z*-score	Class II
Ontologizer	2004	Fisher's exact	Class III
POSOC	2004	POSET (a discrete math: finite partially ordered set)	Class III
topGO	2006	Fisher's exact	Class III
GO-2D	2007	Hypergeometric; binomial	Class III
GENECODIS	2007	Hypergeometric; chi-square	Class III
GOSim	2007	Resnik's similarity	Class III
PalS	2008	Percent	Class III
ProfCom	2008	Greedy heuristics	Class III
GOTM	2004	Hypergeometric	Class I,II
ermineJ	2005	Permutations; Wilcoxon rank-sum test	Class I,II
DAVID	2003	Fisher's Exact (modified as EASE score)	Class I,III
GOToolBox	2004	Hypergeometric; Fisher's exact; Binomial	Class I,III
ADGO	2006	*Z*-statistic	Class II,III
FunNet	2008	Unclear	Unclear

Huang et al. (2009) NAR

Functional Enrichment Analysis (EA)

3 major types:

singular EA = SEA
selected gene list used to query different annotation spaces (one by one)

gene set EA = GSEA
not a list, but the entire genes ranked used to query different annotation spaces (one by one)

modular EA = MEA
selected gene list used to query multiple annotation spaces (at once)

Huang et al. (2009) NAR

Tool category	Description
Class I: singular enrichment analysis (SEA)	Enrichment *P*-value is calculated on each term from the pre-selected interesting gene list. Then, enriched terms are listed in a simple linear text format. This strategy is the most traditional algorithm. It is still dominantly used by most of the enrichment analysis tools.
Class II: gene set enrichment analysis (GSEA)	Entire genes (without pre-selection) and associated experimental values are considered in the enrichment analysis. The unique features of this strategy are: (i) No need to pre-select interesting genes, as opposed to Classes I and II; (ii) Experimental values integrated into *P*-value calculation.
Class III: modular enrichment analysis (MEA)	This strategy inherits key spirit of SEA. However, the term–term/gene–gene relationships are considered into enrichment *P*-value calculation. The advantage of this strategy is that term–term/gene–gene relationship might contain unique biological meaning that is not held by a single term or gene. Such network/modular analysis is closer to the nature of biological data structure.

Session 1 (9:30 - 12:30, 3h)
Bioinformatic tools for Functional Enrichment Analysis (FEA)
Session 2 (13:30 - 16:30, 3h)
Construction of gene functional networks

- Introduction to biological information and annotation spaces: GO, KEGG, Interpro

- Functional Enrichment Analysis (EA): from single to modular methods

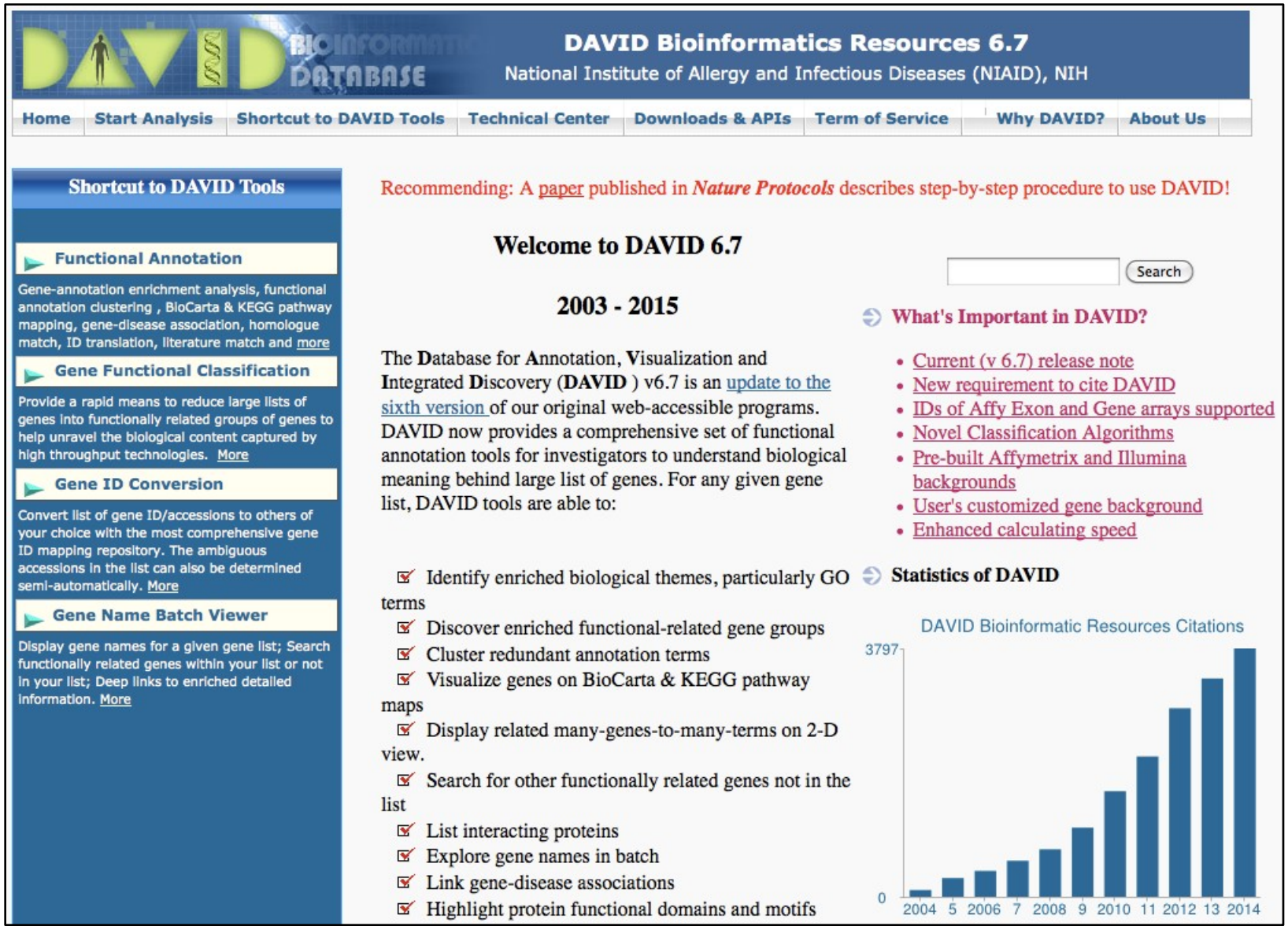
- Using EA tools to annotate gene lists: DAVID (single), GSEA (gene sets), GeneCodis (modular)

- Sort out problems after EA: post-enrichment tool GeneTermLinker (postEA)

- From co-annotation and enrichment to functional networks: networks construction using a R tool

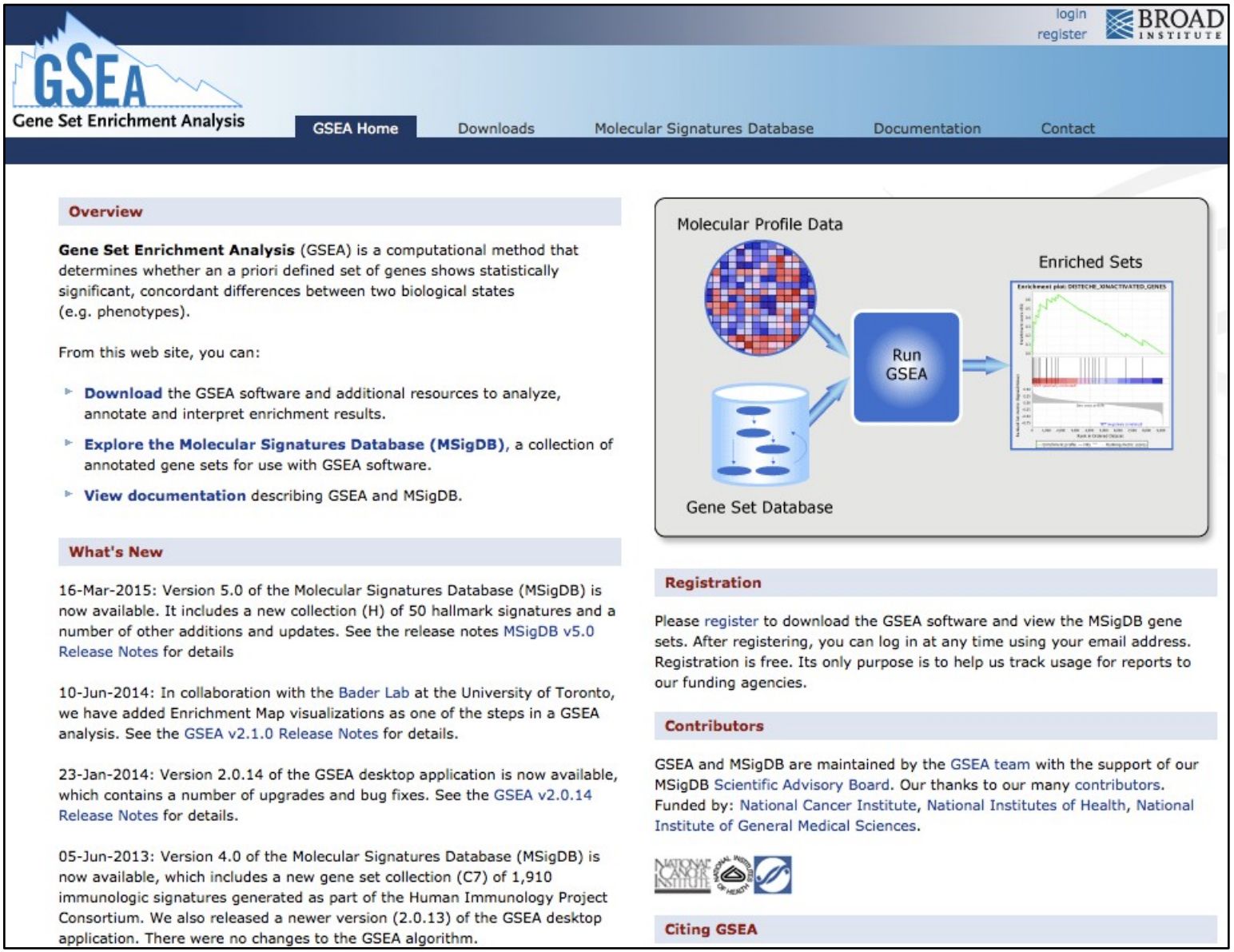
Functional Enrichment Analysis (EA)

Type 1: singular EA = SEA (DAVID)
selected gene list used to query different annotation spaces (one by one)



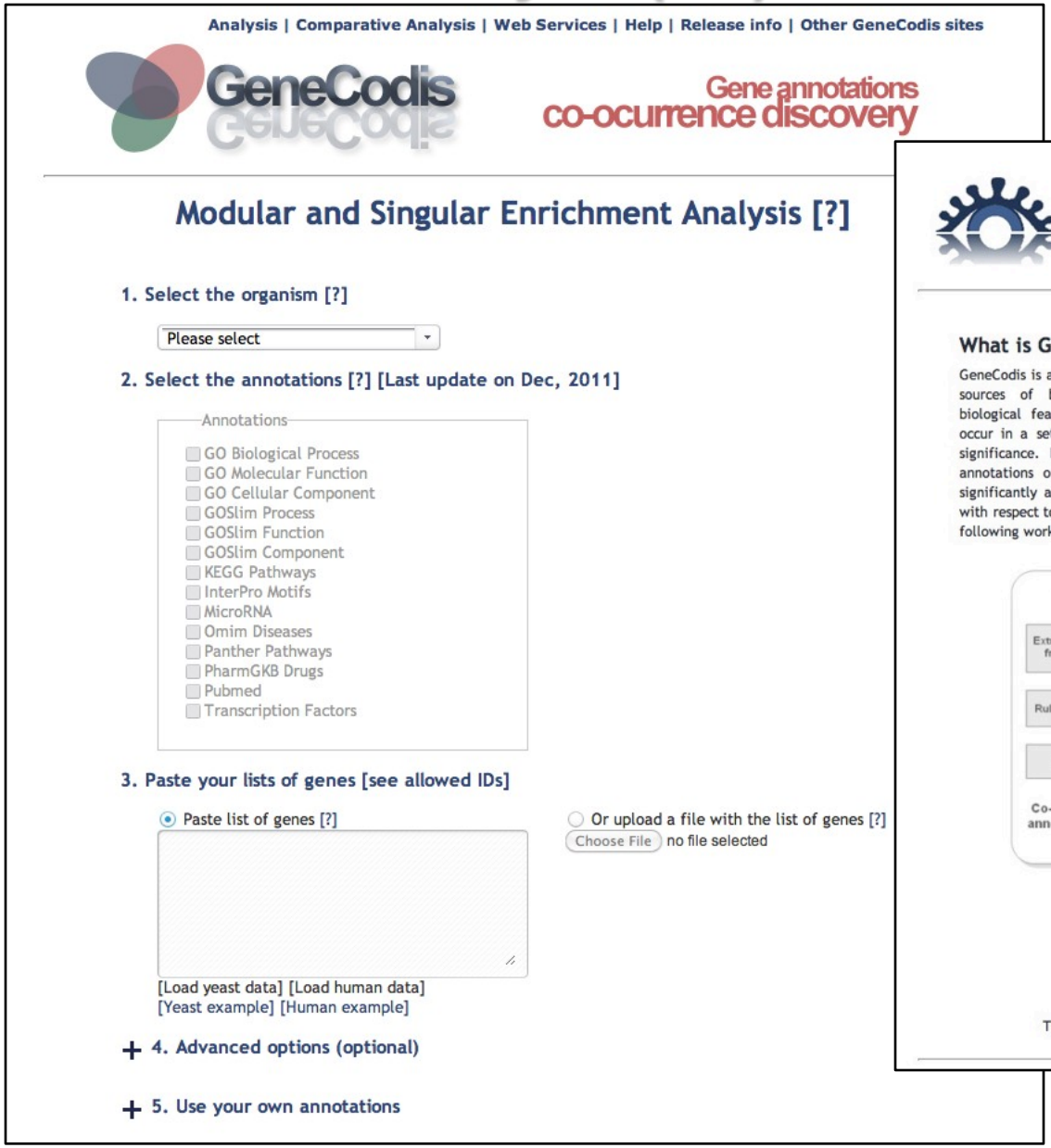
Functional Enrichment Analysis (EA)

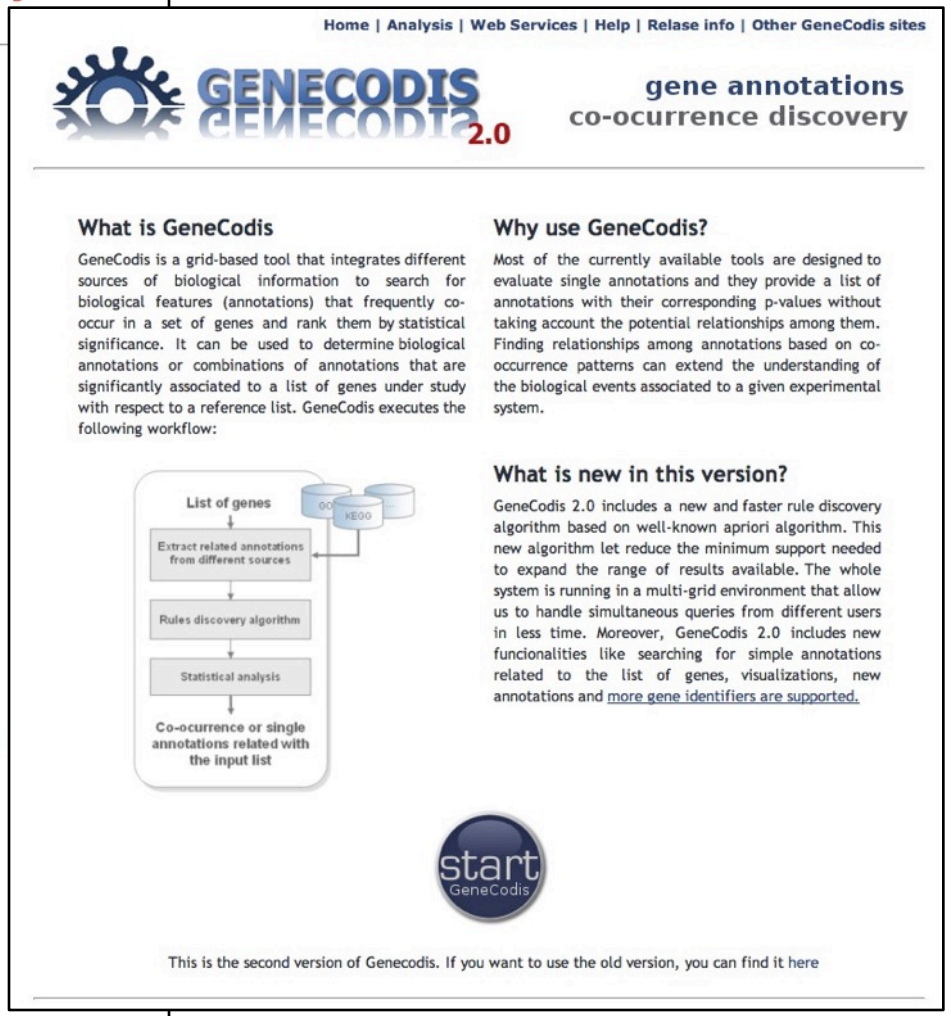
Type 2: gene set EA = GSEA (GSEA)
entire genes ranked used to query different annotation spaces (one by one)



Functional Enrichment Analysis (EA)

Type 3: modular EA = MEA (GeneCodis)
selected gene list used to query
multiple annotation spaces (at once)





Functional Enrichment Analysis (EA)

3 major types:

singular EA = SEA (DAVID *ncbi*)
selected gene list used to query different annotation spaces (one by one)

gene set EA = GSEA (GSEA *Broad*)
entire genes ranked used to query different annotation spaces (one by one)

modular EA = MEA (GeneCodis *CNB*)
selected gene list used to query multiple annotation spaces (at once)

Huang et al. (2009) NAR

Tool category	Description
Class I: singular enrichment analysis (SEA)	Enrichment *P*-value is calculated on each term from the pre-selected interesting gene list. Then, enriched terms are listed in a simple linear text format. This strategy is the most traditional algorithm. It is still dominantly used by most of the enrichment analysis tools.
Class II: gene set enrichment analysis (GSEA)	Entire genes (without pre-selection) and associated experimental values are considered in the enrichment analysis. The unique features of this strategy are: (i) No need to pre-select interesting genes, as opposed to Classes I and II; (ii) Experimental values integrated into *P*-value calculation.
Class III: modular enrichment analysis (MEA)	This strategy inherits key spirit of SEA. However, the term–term/gene–gene relationships are considered into enrichment *P*-value calculation. The advantage of this strategy is that term–term/gene–gene relationship might contain unique biological meaning that is not held by a single term or gene. Such network/modular analysis is closer to the nature of biological data structure.

Functional Enrichment Analysis (EA)

3 major types:

singular EA = SEA
selected gene list used to query different annotation spaces (one by one)

gene set EA = GSEA
entire genes ranked used to query different annotation spaces (one by one)

modular EA = MEA
selected gene list used to query multiple annotation spaces (at once)

statistical and theoretical basis of these methods

Functional Enrichment Analysis (EA)

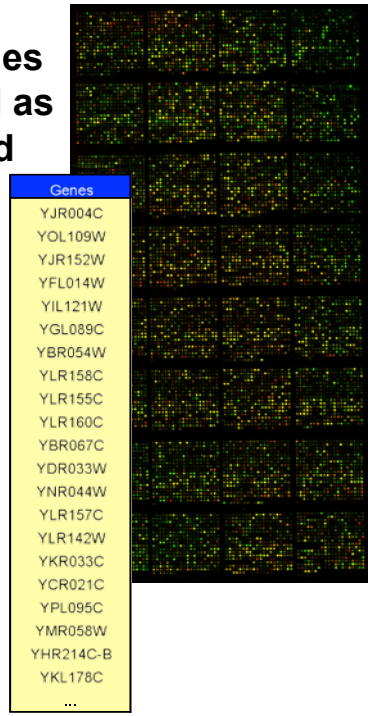
Functional **Enrcihment Analysis** (EA, *Análisis de Enriquecimiento Funcional*): is the identification of **biological functions** and **processes** that are **over-represented** in a given **gene list** or in a **gene subset** selected in a given study.

Find all the **genes** that are **annotated** to a given biological term and evaluate how **relevant** is such functional annotation

¿Which is the **probability** that such **annotation** is **by chance** (i.e. random)? The **p-value** allows to calculate the **statistical probability** of a given selection and allows to fix confidence thresholds (for example, set the test to allow less than 5% expected selection by chance)

10,000 genes total analysed in a 'omic' study

500 genes selected as altered



If in 10,000 genes total => 100 genes are red

In 500 genes selected in one study => 5 genes are expected red
so any ratio found ≤ 1% (0.01) is normal and equal to *by chance*

Functional Enrichment Analysis (EA)

AATF	cell adhesion, ribosome biogenesis, **apoptosis**,...
BAG1	neuron differentiation, **apoptosis**, anti-apoptosis,...
CLIC4	cytoplasm, actin cytoskeleton, membrane ...
ATRIP	nucleus, DNA repair, protein binding, ...
NFKB1	signal trasduction, **apoptosis**, response to...
MDM1	nucleus
OFD1	centrosome, oxidoreductase activity, ...
MAD1L1	mitotic metaphase, cytosol, spindle, ...
PDCD1	**apoptosis**, humoral immune response, protein ...
SNCG	axon, cell soma, adult locomotory behavior

QUERY gene list studied: 4 genes of 10 (in the gene list) annotated to '*apoptosis*'
REFERENCE gene list (i.e. whole proteome): 452 genes of 29,905 (universe) annotated to 'apoptosis'

↓

Statistical test for significant enrichment in the gene list studied

↓

'*apoptosis*' is an ENRICHED TERM with p-value = 1.12×10^{-5}

Functional Enrichment Analysis (EA)

3 major types:

singular EA = SEA
selected gene list used to query different annotation spaces (one by one)

gene set EA = GSEA
entire genes ranked used to query different annotation spaces (one by one)

modular EA = MEA
selected gene list used to query multiple annotation spaces (at once)

These methods alocate groups of **genes** to **terms** in a **contingency table** (with **YES** or **NO** categories) and then they apply the:

Fisher Exact test
or
Hypergeometric test

that calculate the probability using the **contingency table** of frequency data cross-classified according the two categorical variables of assignment or not:
YES = if the **gene** in the list is assigned to the **term**
NO = if the **gene** in the list is not-assigned to the **term**

Real Event	Event observed		
	Yes	No	Marginal total
Yes	Hit	False alarm	Real Yes
No	Miss	Correct non-event	Real No
Marginal total	Obs Yes	Obs No	Sum total

Real Event	Event observed		
	Yes	No	Marginal total
Yes	a	b	a + b
No	c	d	c + d
Marginal total	a + c	b + d	a + b + c + d = n

Functional Enrichment Analysis (EA)

3 major types:

singular EA = SEA
selected gene list used to query different annotation spaces (one by one)

gene set EA = GSEA
entire genes ranked used to query different annotation spaces (one by one)

modular EA = MEA
selected gene list used to query multiple annotation spaces (at once)

These methods alocate groups of **genes** to **terms** in a **contingency table** (with **YES** or **NO** categories) and then they apply the:

Fisher Exact test
or
Hypergeometric test

that calculate the probability using the **contingency table** of frequency data cross-classified according the two categorical variables of assignment or not:
YES = if the **gene** in the list is assigned to the **term**
NO = if the **gene** in the list is not-assigned to the **term**

Gene Annotated White	Gene Selected in Query List YES / NO		
	Yes	No	Marginal total
Yes	Hit	False alarm	W Yes
No	Miss	Correct non-event	W No
Marginal total	Obs Yes	Obs No	Sum total

Gene Annotated White	Gene Selected in Query List YES / NO		
	Yes	No	Marginal total
Yes	a	b	a + b
No	c	d	c + d
Marginal total	a + c	b + d	a + b + c + d =n

Functional Enrichment Analysis (EA)

Hypergeometric test

The classical application of the hypergeometric distribution is **sampling without replacement**. Think of an urn with two types of marbles, black ones and white ones. Define drawing a white marble as a success and drawing a black marble as a failure (analogous to the binomial distribution). If the variable N describes the number of **all marbles in the urn** (see contingency table below) and K describes the number of **white marbles**, then $N - K$ corresponds to the number of **black marbles**. In this example X is the random variable whose outcome is k, the number of white marbles actually drawn in the experiment. This situation is illustrated by the following contingency table:

	drawn	not drawn	total
white marbles	k	$K - k$	K
black marbles	$n - k$	$N + k - n - K$	$N - K$
total	n	$N - n$	N

Now, assume (for example) that there are 5 white and 45 black marbles in the urn. Standing next to the urn, you close your eyes and draw 10 marbles without replacement. What is the probability that exactly 4 of the 10 are white? *Note that although we are looking at success/failure, the data are not accurately modeled by the binomial distribution, because the probability of success on each trial is not the same, as the size of the remaining population changes as we remove each marble.*

This problem is summarized by the following contingency table:

	drawn	not drawn	total
white marbles	$k = 4$	$K - k = 1$	$K = 5$
black marbles	$n - k = 6$	$N + k - n - K = 39$	$N - K = 45$
total	$n = 10$	$N - n = 40$	$N = 50$

Universe

5 red and 45 black marbles in the urn of 50

&

Selection

4 red and 6 black of 10 select marbles

The probability of drawing exactly k white marbles can be calculated by the formula

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

Hence, in this example calculate

$$P(X = 4) = f(4; 50, 5, 10) = \frac{\binom{5}{4}\binom{45}{6}}{\binom{50}{10}} = \frac{5 \cdot 8145060}{10272278170} = 0.003964583\ldots.$$

CBO Computational Biology @ UCT

CIC

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Hands-on: Practical Examples

Enrichment statistical tests
(Hypergeometric test and Fisher's Exact test)

run R Protocol: Protocol_1_EnrichmentTests.R

Functional Enrichment Analysis (EA)

3 major types:

singular EA = SEA (DAVID *ncbi*)
selected gene list used to query different annotation spaces (one by one)

gene set EA = GSEA (GSEA *Broad*)
entire genes ranked used to query different annotation spaces (one by one)

modular EA = MEA (GeneCodis *CNB*)
selected gene list used to query multiple annotation spaces (at once)

Huang et al. (2009) NAR

Tool category	Description
Class I: singular enrichment analysis (SEA)	Enrichment *P*-value is calculated on each term from the pre-selected interesting gene list. Then, enriched terms are listed in a simple linear text format. This strategy is the most traditional algorithm. It is still dominantly used by most of the enrichment analysis tools.
Class II: gene set enrichment analysis (GSEA)	Entire genes (without pre-selection) and associated experimental values are considered in the enrichment analysis. The unique features of this strategy are: (i) No need to pre-select interesting genes, as opposed to Classes I and II; (ii) Experimental values integrated into *P*-value calculation.
Class III: modular enrichment analysis (MEA)	This strategy inherits key spirit of SEA. However, the term–term/gene–gene relationships are considered into enrichment *P*-value calculation. The advantage of this strategy is that term–term/gene–gene relationship might contain unique biological meaning that is not held by a single term or gene. Such network/modular analysis is closer to the nature of biological data structure.

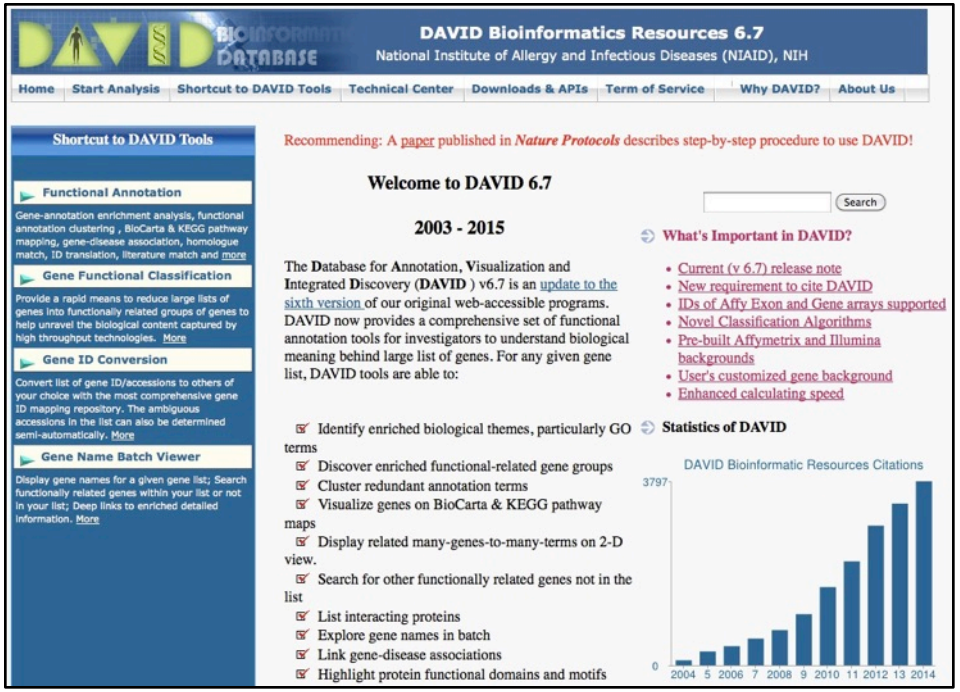
Functional Enrichment Analysis (EA)

3 major types:

singular EA = SEA
selected gene list used to query different annotation spaces (one by one)

gene set EA = GSEA
entire genes ranked used to query different annotation spaces (one by one)

modular EA = MEA
selected gene list used to query multiple annotation spaces (at once)



Type 1: singular EA = SEA (DAVID)
selected gene list used to query different annotation spaces (one by one)

Functional Enrichment Analysis (EA)

Type 1: singular EA = SEA (DAVID)
selected gene list used to query different annotation spaces (one by one)

94 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	GOTERM_BP_FAT	defense response	RT		22	22.7	6.2E-11	5.1E-8
☐	GOTERM_BP_FAT	immune response	RT		20	20.6	2.1E-8	8.5E-6
☐	GOTERM_CC_FAT	extracellular region	RT		30	30.9	3.2E-6	5.6E-4
☐	GOTERM_CC_FAT	anchored to membrane	RT		10	10.3	7.3E-6	6.5E-4
☐	GOTERM_BP_FAT	response to bacterium	RT		9	9.3	2.1E-5	5.9E-3
☐	GOTERM_CC_FAT	extracellular region part	RT		18	18.6	5.3E-5	3.1E-3
☐	GOTERM_BP_FAT	defense response to bacterium	RT		7	7.2	5.9E-5	1.2E-2
☐	GOTERM_MF_FAT	carbohydrate binding	RT		10	10.3	7.1E-5	1.6E-2
☐	GOTERM_BP_FAT	response to wounding	RT		13	13.4	7.8E-5	1.3E-2
☐	GOTERM_CC_FAT	secretory granule	RT		8	8.2	1.1E-4	4.9E-3
☐	GOTERM_BP_FAT	inflammatory response	RT		10	10.3	1.5E-4	2.0E-2
☐	GOTERM_CC_FAT	extracellular space	RT		14	14.4	2.3E-4	8.3E-3
☐	GOTERM_CC_FAT	cytoplasmic vesicle	RT		13	13.4	4.8E-4	1.4E-2
☐	GOTERM_CC_FAT	cytoplasmic membrane-bounded vesicle	RT		12	12.4	4.8E-4	1.2E-2
☐	GOTERM_CC_FAT	membrane-bounded vesicle	RT		12	12.4	6.3E-4	1.4E-2

annotation space — terms — genes — p-values

CBO Computational Biology @ UCT

CiC

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Hands-on: Practical Examples

Protein_ SETs_ 2015.xls

(106g hs, 175g hs, yeast 11pathways, yeast 59g5pc)

run DAVID-FAC

Functional Enrichment Analysis (EA)

3 major types:

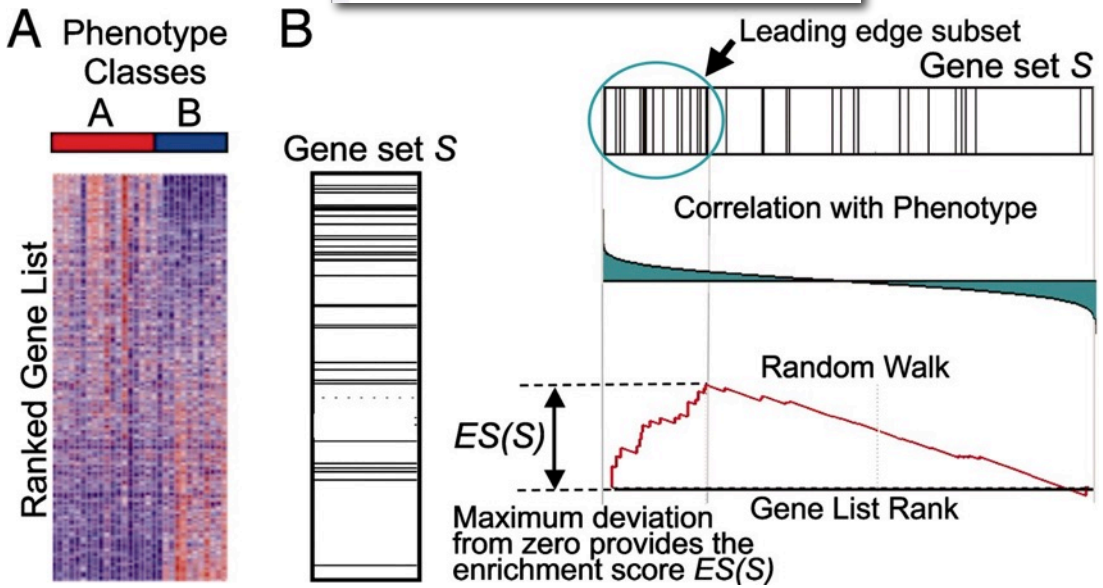
singular EA = SEA
selected gene list used to query different annotation spaces (one by one)

gene set EA = GSEA
entire genes ranked used to query different annotation spaces (one by one)

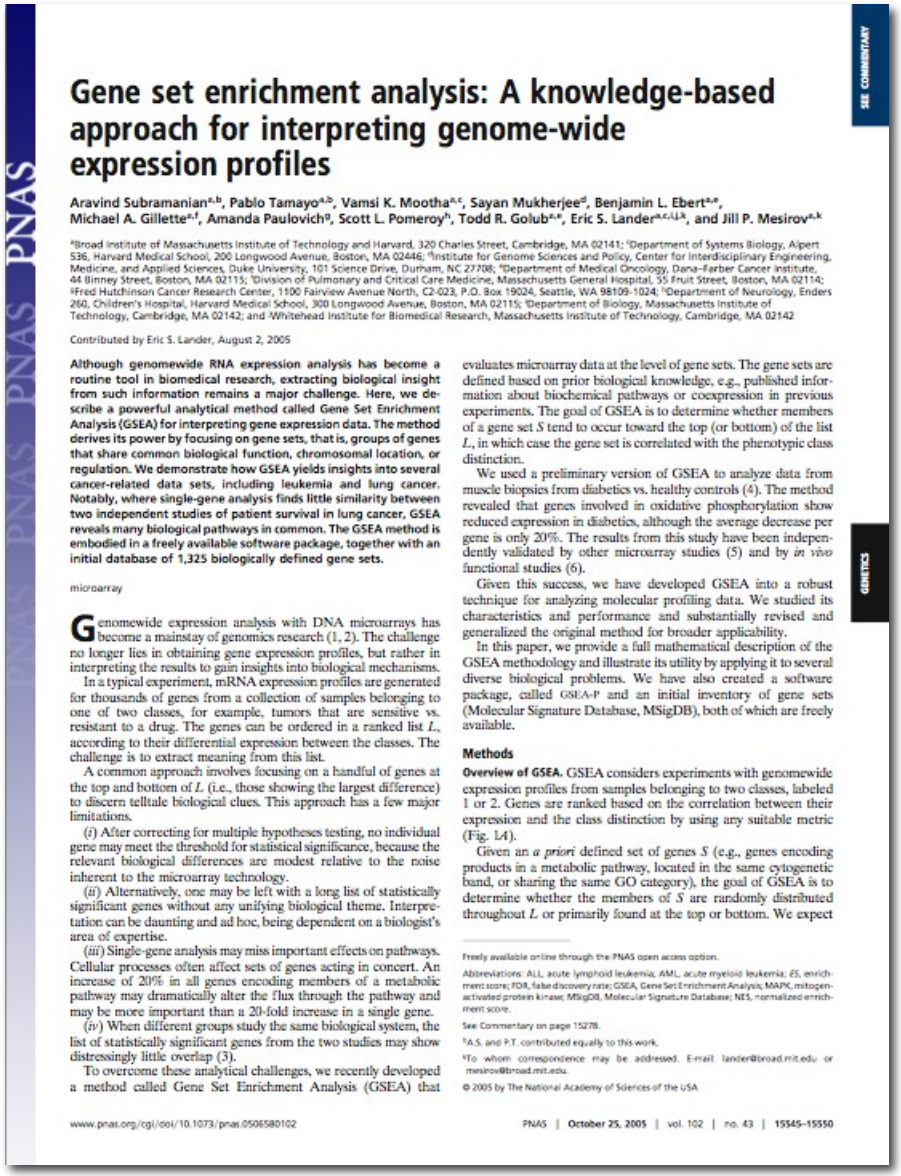
modular EA = MEA
selected gene list used to query multiple annotation spaces (at once)

Subramanian et al. (2005) PNAS

Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles



Functional Enrichment Analysis (EA): GSEA



Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles

Aravind Subramanian[a,b], Pablo Tamayo[a,b], Vamsi K. Mootha[a,c], Sayan Mukherjee[d], Benjamin L. Ebert[a,e], Michael A. Gillette[a,f], Amanda Paulovich[g], Scott L. Pomeroy[h], Todd R. Golub[a,e], Eric S. Lander[a,c,i,k], and Jill P. Mesirov[a,k]

[a]Broad Institute of Massachusetts Institute of Technology and Harvard, 320 Charles Street, Cambridge, MA 02141; [b]Department of Systems Biology, Alpert 536, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02446; [c]Institute for Genome Sciences and Policy, Center for Interdisciplinary Engineering, Medicine, and Applied Sciences, Duke University, 101 Science Drive, Durham, NC 27708; [d]Department of Medical Oncology, Dana–Farber Cancer Institute, 44 Binney Street, Boston, MA 02115; [e]Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114; [f]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, C2-023, P.O. Box 19024, Seattle, WA 98109-1024; [h]Department of Neurology, Enders 260, Children's Hospital, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115; [i]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142; and [k]Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology, Cambridge, MA 02142

Contributed by Eric S. Lander, August 2, 2005

Although genomewide RNA expression analysis has become a routine tool in biomedical research, extracting biological insight from such information remains a major challenge. Here, we describe a powerful analytical method called Gene Set Enrichment Analysis (GSEA) for interpreting gene expression data. The method derives its power by focusing on gene sets, that is, groups of genes that share common biological function, chromosomal location, or regulation. We demonstrate how GSEA yields insights into several cancer-related data sets, including leukemia and lung cancer. Notably, where single-gene analysis finds little similarity between two independent studies of patient survival in lung cancer, GSEA reveals many biological pathways in common. The GSEA method is embodied in a freely available software package, together with an initial database of 1,325 biologically defined gene sets.

microarray

Genomewide expression analysis with DNA microarrays has become a mainstay of genomics research (1, 2). The challenge no longer lies in obtaining gene expression profiles, but rather in interpreting the results to gain insights into biological mechanisms.

In a typical experiment, mRNA expression profiles are generated for thousands of genes from a collection of samples belonging to one of two classes, for example, tumors that are sensitive vs. resistant to a drug. The genes can be ordered in a ranked list L, according to their differential expression between the classes. The challenge is to extract meaning from this list.

A common approach involves focusing on a handful of genes at the top and bottom of L (i.e., those showing the largest difference) to discern telltale biological clues. This approach has a few major limitations.

(*i*) After correcting for multiple hypotheses testing, no individual gene may meet the threshold for statistical significance, because the relevant biological differences are modest relative to the noise inherent to the microarray technology.

(*ii*) Alternatively, one may be left with a long list of statistically significant genes without any unifying biological theme. Interpretation can be daunting and ad hoc, being dependent on a biologist's area of expertise.

(*iii*) Single-gene analysis may miss important effects on pathways. Cellular processes often affect sets of genes acting in concert. An increase of 20% in all genes encoding members of a metabolic pathway may dramatically alter the flux through the pathway and may be more important than a 20-fold increase in a single gene.

(*iv*) When different groups study the same biological system, the list of statistically significant genes from the two studies may show distressingly little overlap (3).

To overcome these analytical challenges, we recently developed a method called Gene Set Enrichment Analysis (GSEA) that evaluates microarray data at the level of gene sets. The gene sets are defined based on prior biological knowledge, e.g., published information about biochemical pathways or coexpression in previous experiments. The goal of GSEA is to determine whether members of a gene set S tend to occur toward the top (or bottom) of the list L, in which case the gene set is correlated with the phenotypic class distinction.

We used a preliminary version of GSEA to analyze data from muscle biopsies from diabetics vs. healthy controls (4). The method revealed that genes involved in oxidative phosphorylation show reduced expression in diabetics, although the average decrease per gene is only 20%. The results from this study have been independently validated by other microarray studies (5) and by *in vivo* functional studies (6).

Given this success, we have developed GSEA into a robust technique for analyzing molecular profiling data. We studied its characteristics and performance and substantially revised and generalized the original method for broader applicability.

In this paper, we provide a full mathematical description of the GSEA methodology and illustrate its utility by applying it to several diverse biological problems. We have also created a software package, called GSEA-P and an initial inventory of gene sets (Molecular Signature Database, MSigDB), both of which are freely available.

Methods

Overview of GSEA. GSEA considers experiments with genomewide expression profiles from samples belonging to two classes, labeled 1 or 2. Genes are ranked based on the correlation between their expression and the class distinction by using any suitable metric (Fig. 1*A*).

Given an *a priori* defined set of genes S (e.g., genes encoding products in a metabolic pathway, located in the same cytogenetic band, or sharing the same GO category), the goal of GSEA is to determine whether the members of S are randomly distributed throughout L or primarily found at the top or bottom. We expect

Freely available online through the PNAS open access option.

Abbreviations: ALL, acute lymphoid leukemia; AML, acute myeloid leukemia; *ES*, enrichment score; FDR, false discovery rate; GSEA, Gene Set Enrichment Analysis; MAPK, mitogen-activated protein kinase; MSigDB, Molecular Signature Database; NES, normalized enrichment score.

See Commentary on page 15278.

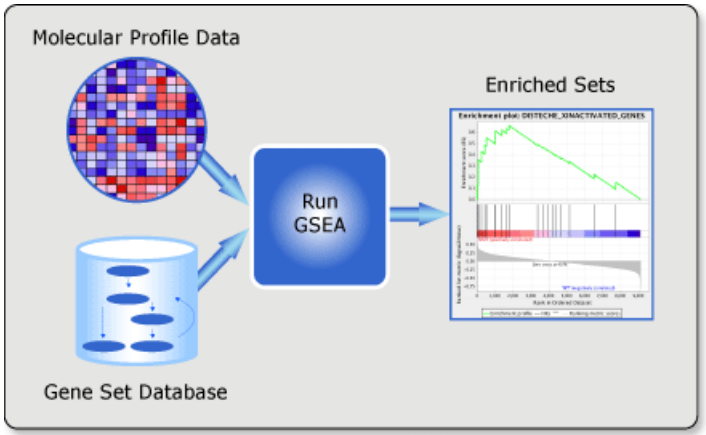
[†]A.S. and P.T. contributed equally to this work.

[‡]To whom correspondence may be addressed. E-mail: lander@broad.mit.edu or mesirov@broad.mit.edu.

Subramanian et al. (2005) PNAS

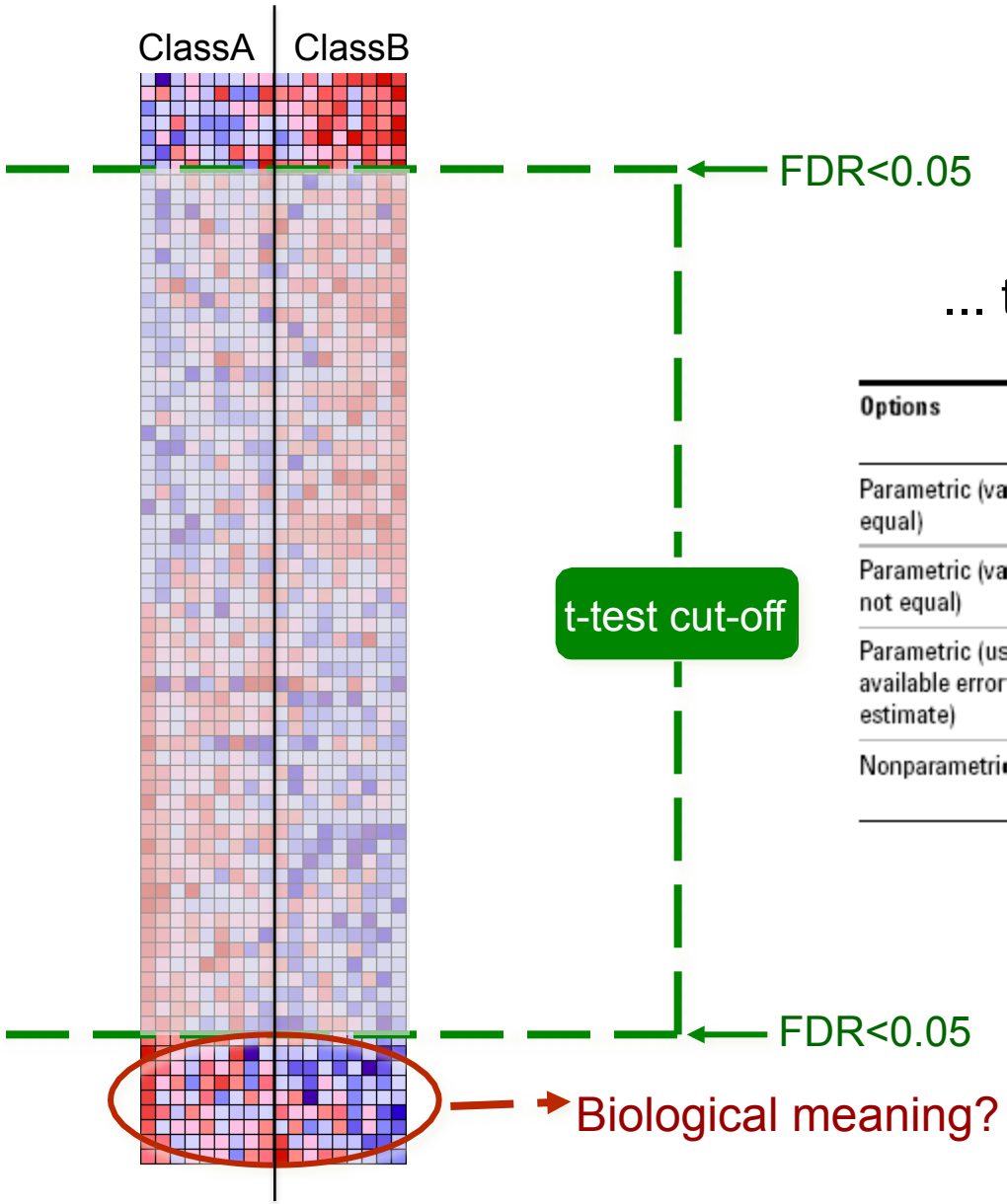
GSEA

MIT - Broad Institute



Functional Enrichment Analysis (EA): GSEA

How GSEA works?



... testing genes independently ...

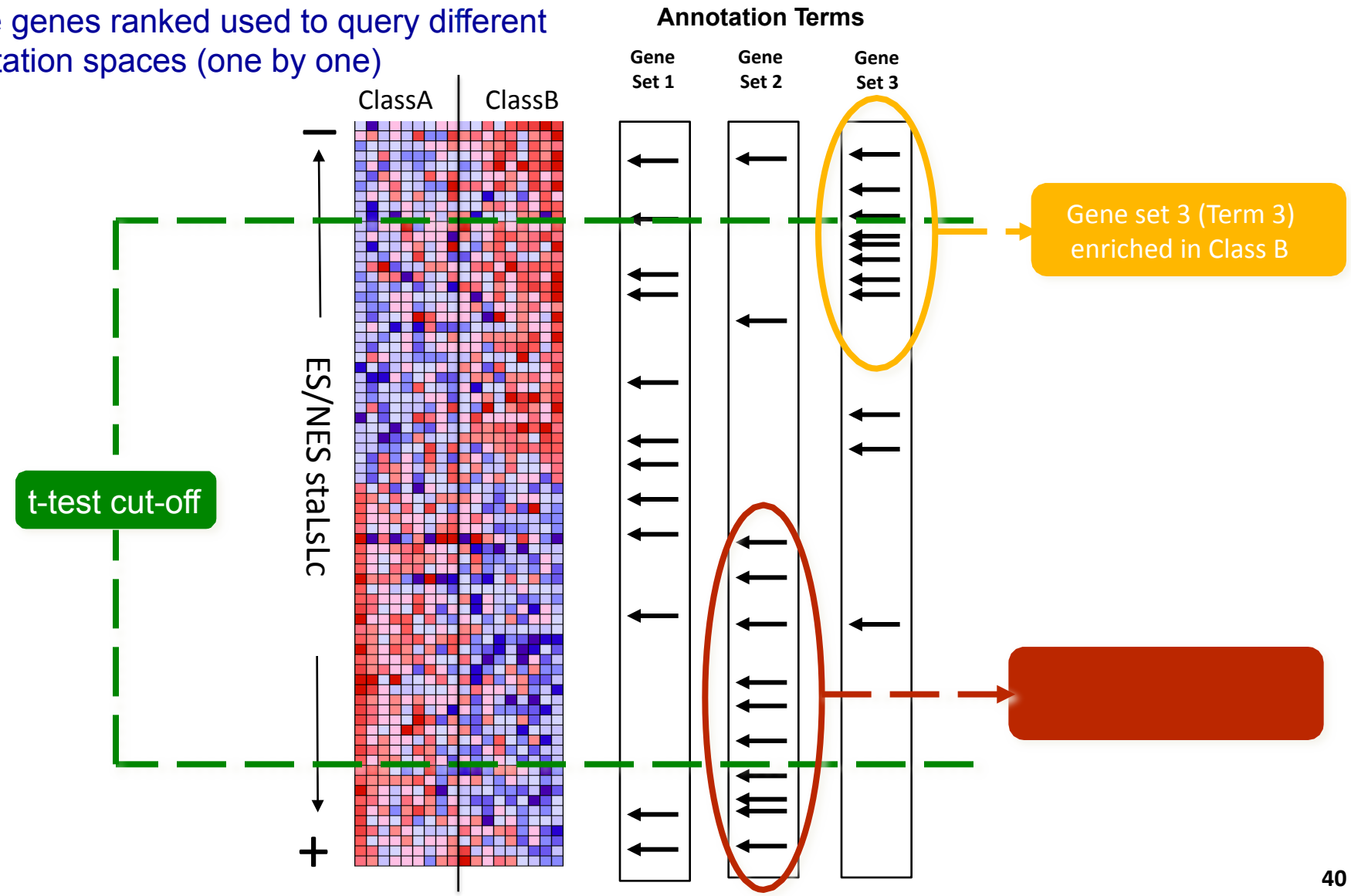
Options	Specific test name: Analyzing 2 groups	Specific test name : Analyzing more than 2 groups
Parametric (variances equal)	Student's T-test	ANOVA
Parametric (variances not equal)	Welch t-test	Welch ANOVA
Parametric (use all available error estimate)	Welch t-test using error model variances	Welch ANOVA using error model variances
Nonparametric	Wilcoxon-Mann-Whitney test	Kruskal-Wallis test

Functional Enrichment Analysis (EA): GSEA

How GSEA works?

gene set EA = GSEA

entire genes ranked used to query different annotation spaces (one by one)



Functional Enrichment Analysis (EA): GSEA

How GSEA works?

gene set EA = GSEA

entire genes ranked used to query different annotation spaces (one by one)

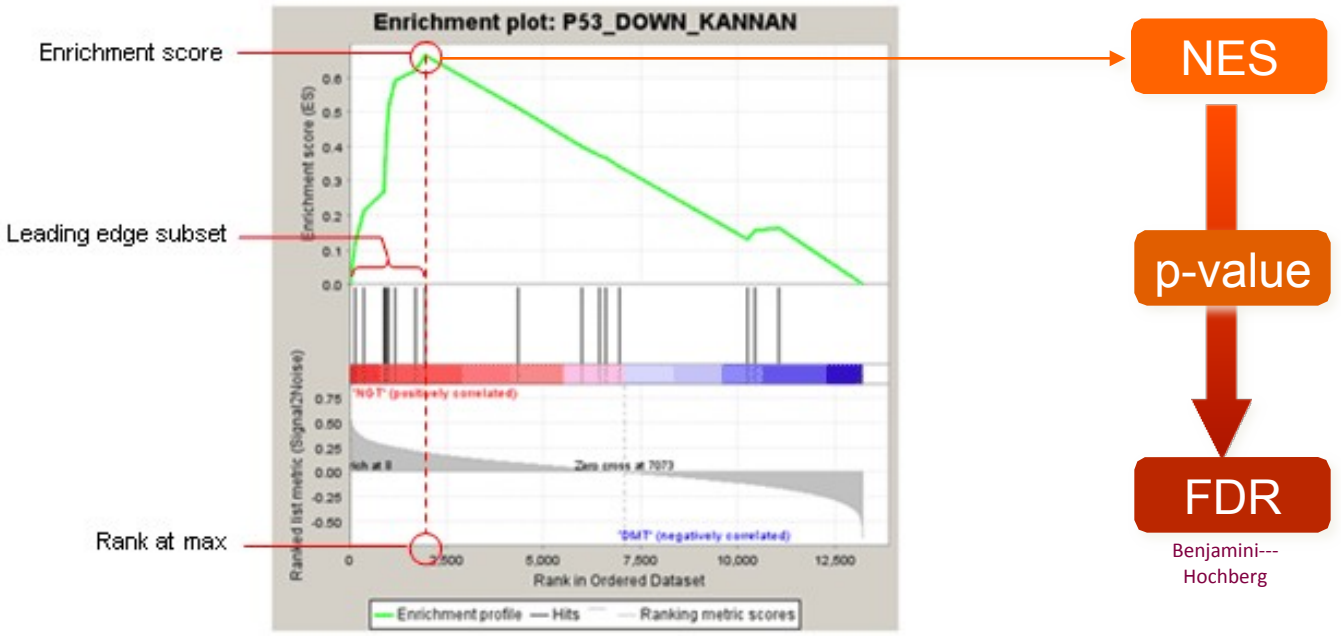


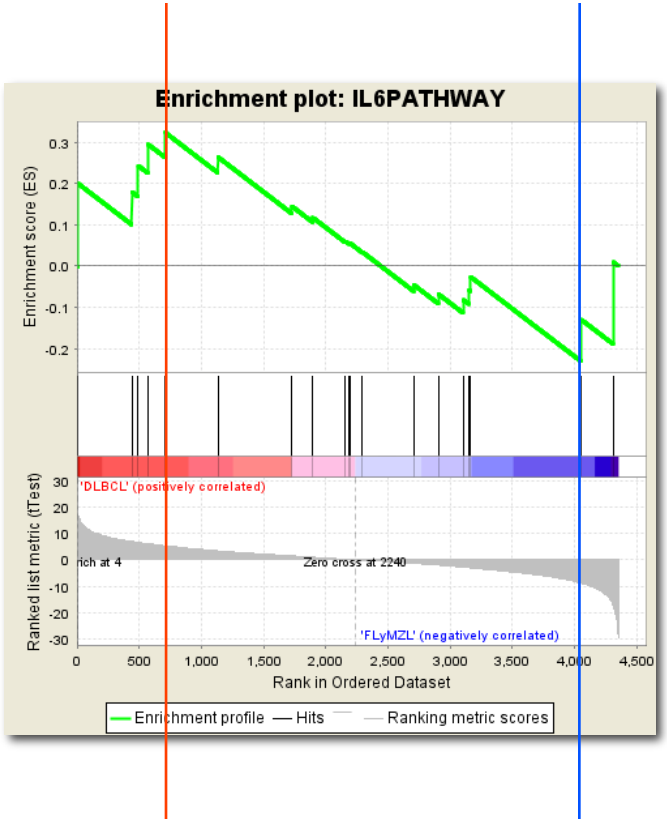
Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

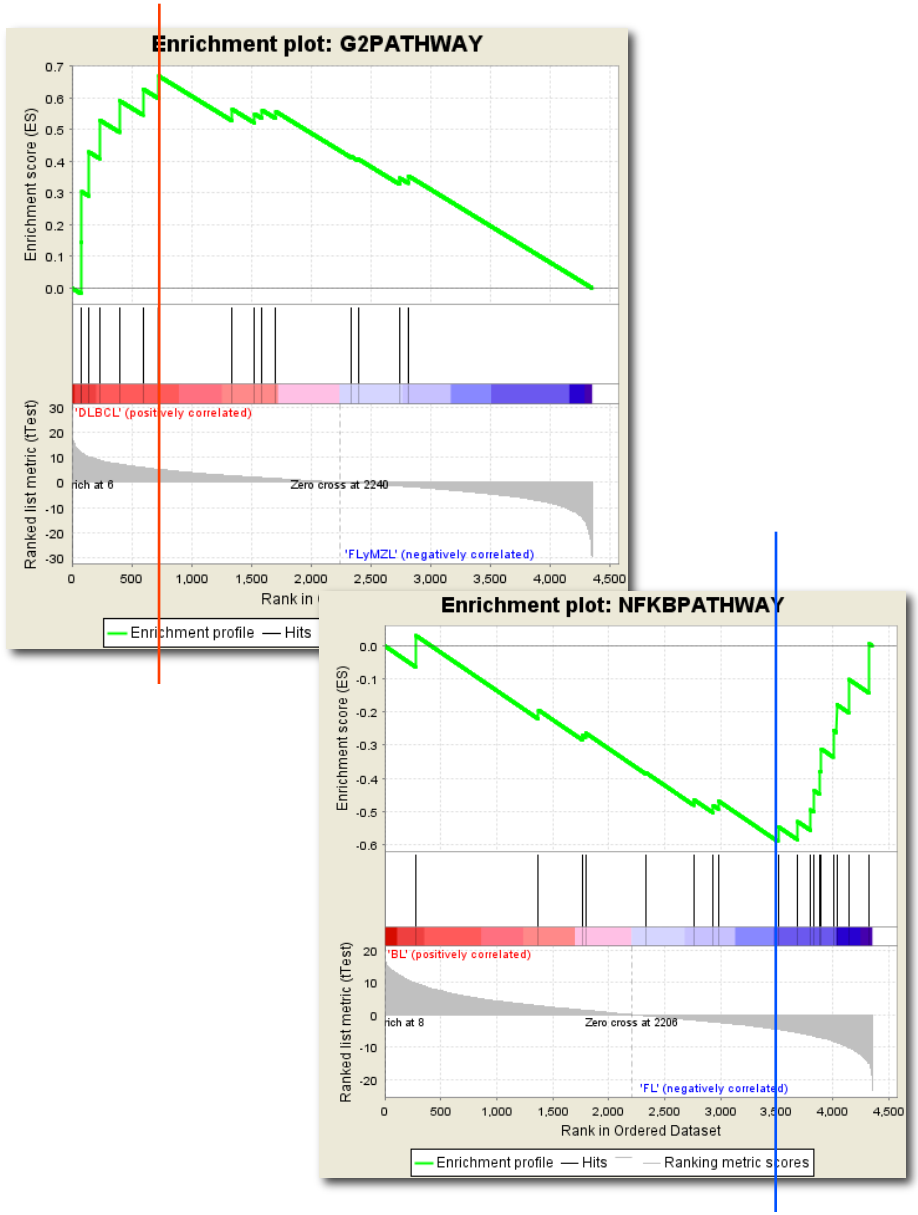
Functional Enrichment Analysis (EA): GSEA

How GSEA works?

gene set EA = GSEA

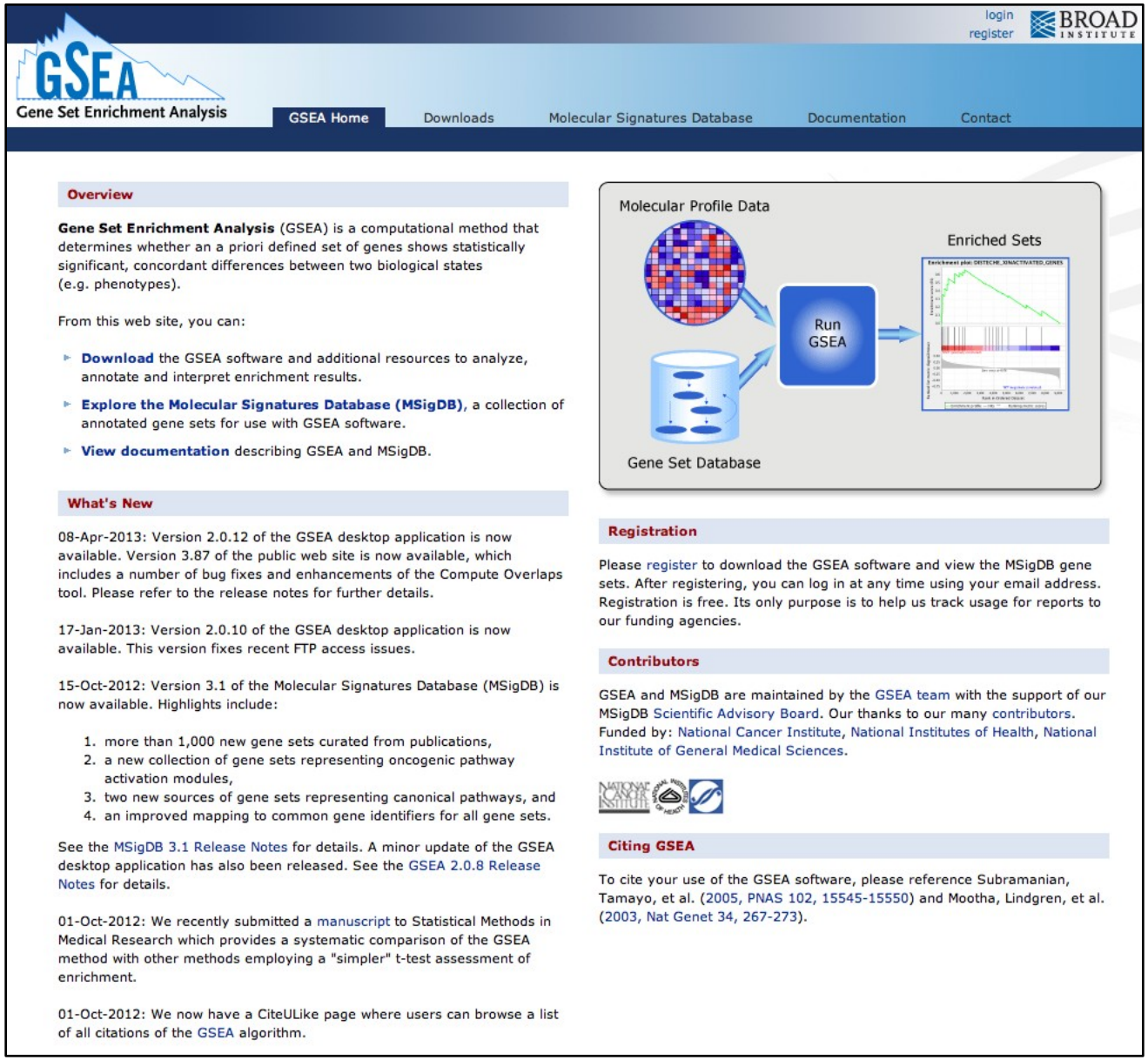
entire genes ranked used to query different annotation spaces (one by one)





Functional Enrichment Analysis (EA): GSEA

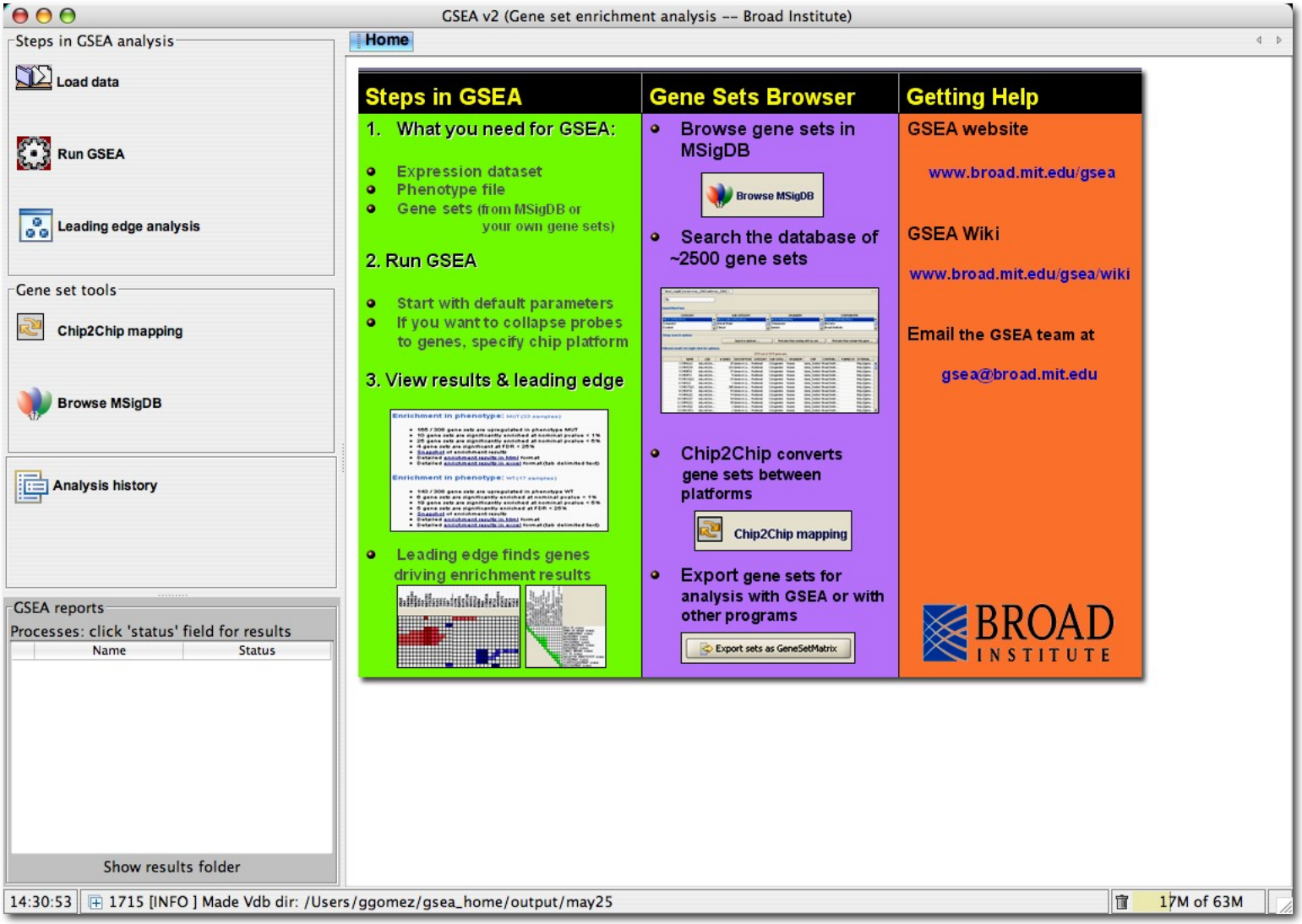
http://www.broad.mit.edu/gsea/



Functional Enrichment Analysis (EA): GSEA

GSEA software

Main Window



Functional Enrichment Analysis (EA): GSEA

GSEA software

Downloads

Downloads

The GSEA software and source code and the Molecular Signatures Database (MSigDB) are freely available to individuals in both academia and industry for internal research purposes. Please see the GSEA/MSigDB license for more details.

Software

There are several options for GSEA software. All options implement exactly the same algorithm. Usage recommendations and installation instructions are listed below. Current Java implementations of GSEA require Java 1.6 or higher. If your computer has Java 1.5 and cannot upgrade to Java 1.6, please see the FAQ.

javaGSEA Desktop Application	▸ Easy-to-use graphical user interface ▸ Runs on any desktop computer (Windows, Mac OS X, Linux etc.) that supports Java1.6+ ▸ Produces richly annotated reports of enrichment results ▸ Integrated gene sets browser to view gene set annotations, search for gene sets and map gene sets between platforms ▸ The GSEA team suggests always starting GSEA by using these Launch buttons, or by clicking the icon that the application installs on your desktop, in order to ensure optimal memory allocation	Launch with 512Mb memory Launch Launch with 1Gb memory Launch
javaGSEA Java Jar file	▸ Command line usage ▸ Runs on any platform that supports Java1.6+ ▸ We recommend using the 'Launch' buttons above instead of this mode for most users	download gsea2-2.07.jar
GSEA Java Source Code Java source files	▸ 100% Java implementation of GSEA ▸ Incorporate GSEA into your own data analysis pipeline ▸ Programmatically call the open source GSEA java API	download gsea2_distrib-2.04.zip
R-GSEA R Script	▸ Usage from within the R programming environment ▸ Easily inspect, learn and tweak the algorithm ▸ Incorporate GSEA into your own data analysis pipeline ▸ Programmatically call the open source GSEA R API ▸ Click here to learn more about the R-GSEA script	download GSEA-P-R.1.0.zip
GenePattern GSEA Module	▸ Use GSEA from within GenePattern ▸ Use GSEA in concert with a large suite of other analytics found in GenePattern (a powerful and flexible analysis platform developed at the Broad Institute)	GenePattern site

For details on the GSEA algorithm and software refer to the Documentation.
For details on the latest release refer to the Release Notes.

CBIO Computational Biology @ UCT

CIC

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Session 1 (9:30 - 12:30, 3h)
Bioinformatic tools for Functional Enrichment Analysis (FEA)
Session 2 (13:30 - 16:30, 3h)
Construction of gene functional networks

- Introduction to biological information and annotation spaces:
GO, KEGG, Interpro

- Functional Enrichment Analysis (EA): from single to modular methods

- Using EA tools to annotate gene lists:
DAVID (single), GSEA (gene sets), GeneCodis (modular)

- Sort out problems after EA: post-enrichment tool
GeneTermLinker (postEA)

- From co-annotation and enrichment to functional networks:
networks construction using a R tool

Functional Enrichment Analysis (EA)

3 major types: (examples)

singular EA = SEA
selected gene list used to query different annotation spaces (one by one)

gene set EA = GSEA
entire genes ranked used to query different annotation spaces (one by one)

modular EA = MEA
selected gene list used to query multiple annotation spaces (at once)

Tool category	Description
Class I: singular enrichment analysis (SEA)	Enrichment *P*-value is calculated on each term from the pre-selected interesting gene list. Then, enriched terms are listed in a simple linear text format. This strategy is the most traditional algorithm. It is still dominantly used by most of the enrichment analysis tools.
Class II: gene set enrichment analysis (GSEA)	Entire genes (without pre-selection) and associated experimental values are considered in the enrichment analysis. The unique features of this strategy are: (i) No need to pre-select interesting genes, as opposed to Classes I and II; (ii) Experimental values integrated into *P*-value calculation.
Class III: modular enrichment analysis (MEA)	This strategy inherits key spirit of SEA. However, the term–term/gene–gene relationships are considered into enrichment *P*-value calculation. The advantage of this strategy is that term–term/gene–gene relationship might contain unique biological meaning that is not held by a single term or gene. Such network/modular analysis is closer to the nature of biological data structure.

Huang et al. (2009) NAR

Functional Enrichment Analysis (EA)

3 major types:

singular EA = SEA (DAVID *ncbi*)
selected gene list used to query different annotation spaces (one by one)

gene set EA = GSEA (GSEA *Broad*)
entire genes ranked used to query different annotation spaces (one by one)

modular EA = MEA (GeneCodis *CNB*)
selected gene list used to query multiple annotation spaces (at once)

Huang et al. (2009) NAR

Tool category	Description
Class I: singular enrichment analysis (SEA)	Enrichment *P*-value is calculated on each term from the pre-selected interesting gene list. Then, enriched terms are listed in a simple linear text format. This strategy is the most traditional algorithm. It is still dominantly used by most of the enrichment analysis tools.
Class II: gene set enrichment analysis (GSEA)	Entire genes (without pre-selection) and associated experimental values are considered in the enrichment analysis. The unique features of this strategy are: (i) No need to pre-select interesting genes, as opposed to Classes I and II; (ii) Experimental values integrated into *P*-value calculation.
Class III: modular enrichment analysis (MEA)	This strategy inherits key spirit of SEA. However, the term–term/gene–gene relationships are considered into enrichment *P*-value calculation. The advantage of this strategy is that term–term/gene–gene relationship might contain unique biological meaning that is not held by a single term or gene. Such network/modular analysis is closer to the nature of biological data structure.

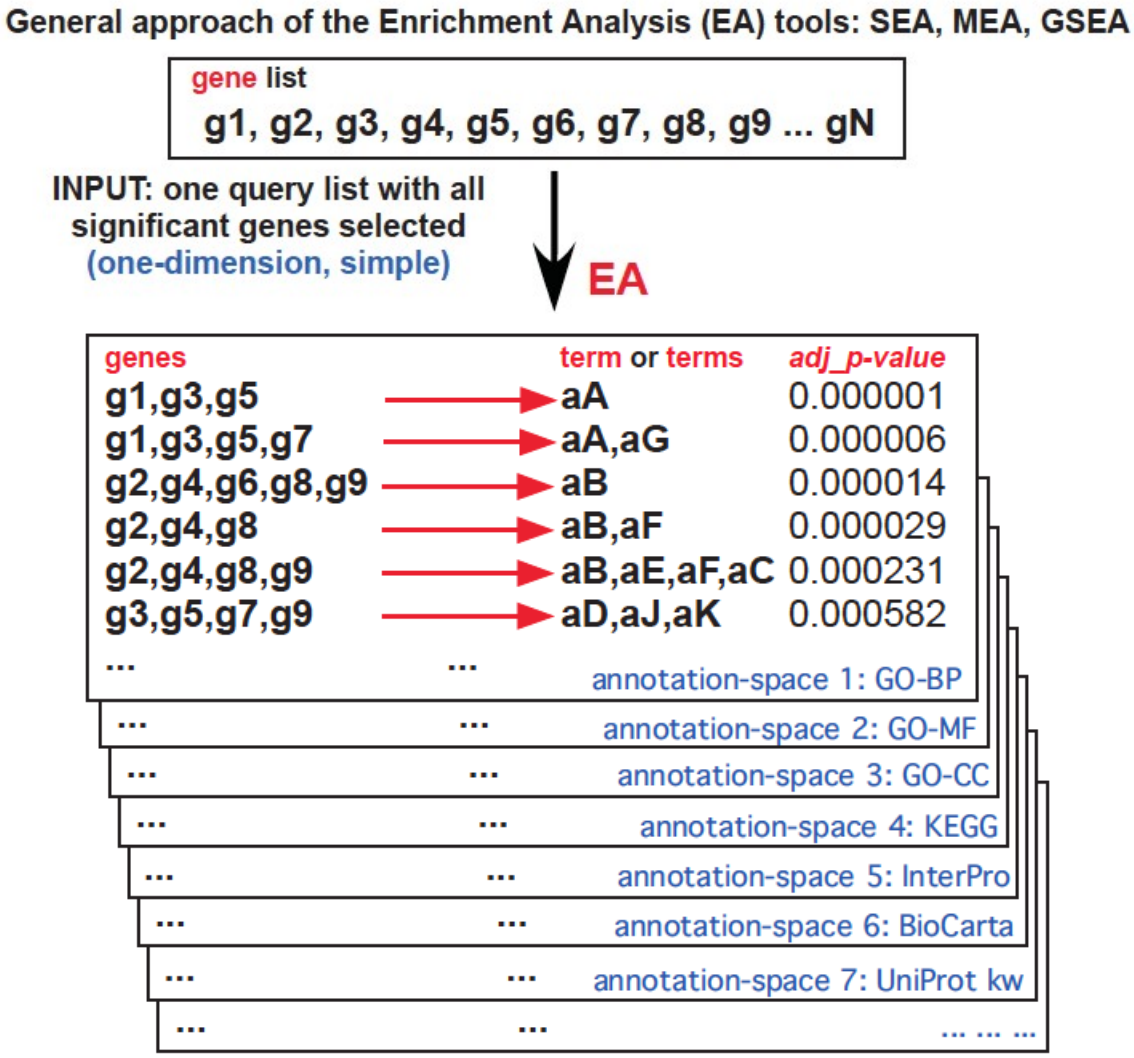
Beyond Enrichment Analysis (EA)

problems:

1.there are many annotation spaces that **overlap**

2.this functional overlapping produces **repetitive** results

3.for a **single** query **gene list** SEA and GSEA tools produced many **result lists** that are highly **redundant**



Beyond
Enrichment Analysis (EA)

Gene List

g1
g2
g3
g4
g5
g6
g7
g8
g9
g10
g11

Functional Enrichment

Result Lists

genes		term or terms	adj_p-value
g1,g3,g5	→	aA	0.000001
g1,g3,g5,g7	→	aA,aG	0.000006
g2,g4,g6,g8,g9	→	aB	0.000014
g2,g4,g8	→	aB,aF	0.000029
g2,g4,g8,g9	→	aB,aE,aF,aC	0.000231
g3,g5,g7,g9	→	aD,aJ,aK	0.000582
...	...	annotation-space 1: GO-BP	
...	...	annotation-space 2: GO-MF	
...	...	annotation-space 3: GO-CC	
...	...	annotation-space 4: KEGG	
...	...	annotation-space 5: InterPro	
...	...	annotation-space 6: BioCarta	
...	...	annotation-space 7: UniProt kw	
...	...	... ...	

Problems:

•Many different result lists, one for each annotation space (in singular EA)

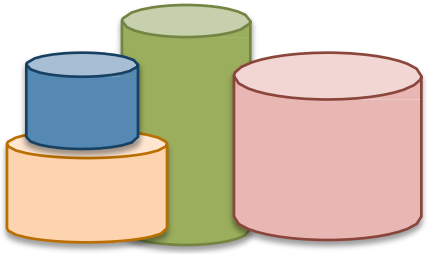
• One big results list (in concurrent modular EA)

• In both cases many **repited** and **redundant** terms

redundant, repetitive, complex, ofuscated results

Beyond
Enrichment Analysis (EA)

Examples of: **redundant, repetitive**
... inherent problems to the annotation spaces:



- Equivalent terms, with the same biological meaning

  GO:0007049 *cell cycle* === GO:0022402 *cell cycle process*

- Redundancy of terms, repeated in different annotation spaces

  GO:0007049: *cell cycle* === KEGG hsa04110: *cell cycle*

- Bias due to highly frequent promiscuous terms that are unspecific

  GO:0005515: *regulation of biological process* includes
  ≈ 45% of all annotated human genes in GO-BP

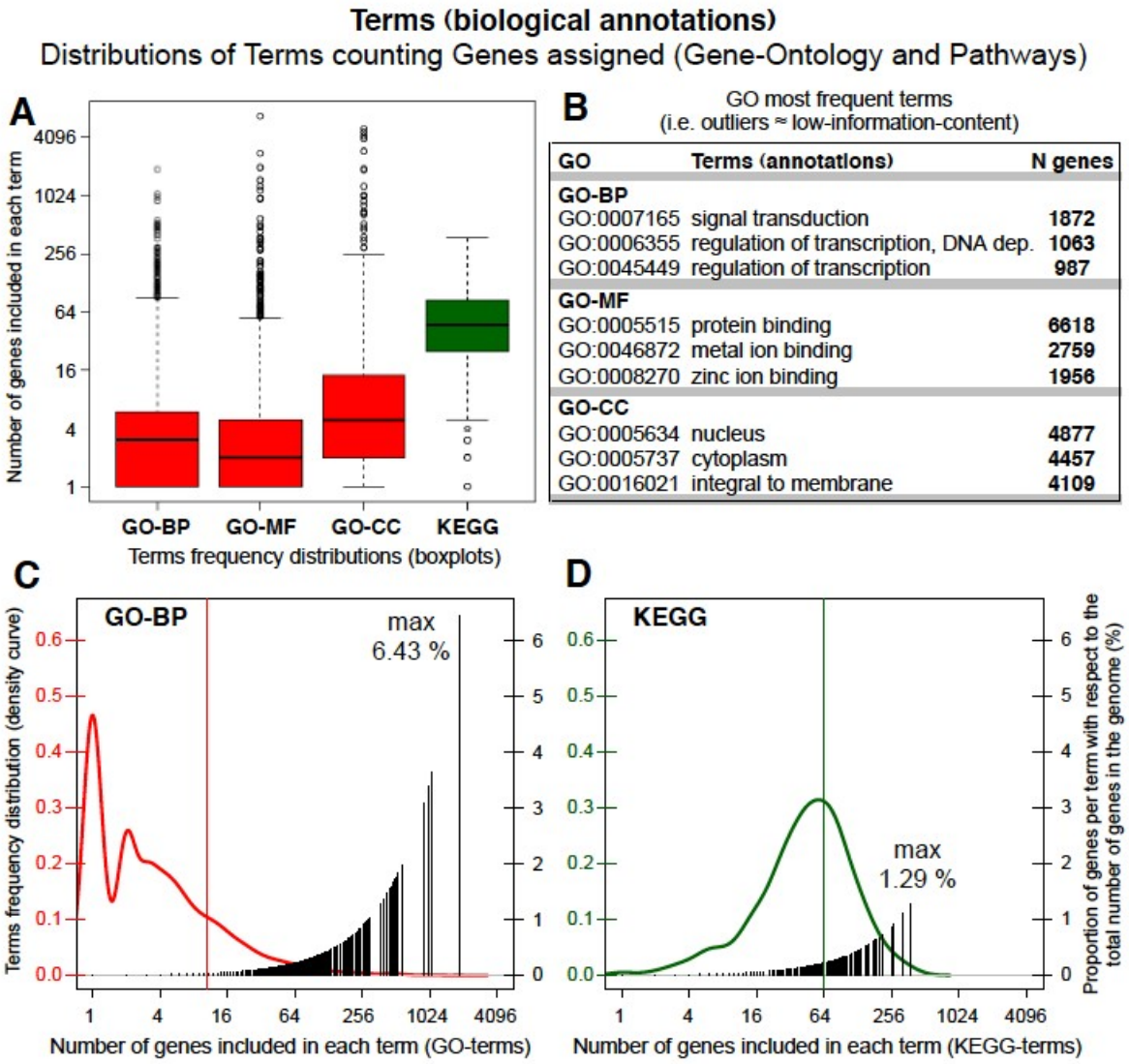
Beyond Enrichment Analysis (EA)

problems:

1.there are many annotation spaces that **overlap**

2.this functional overlapping produces **repetitive** results

3.for a **single** query **gene list** SEA and GSEA tools produced many **result lists** that are highly **redundant**

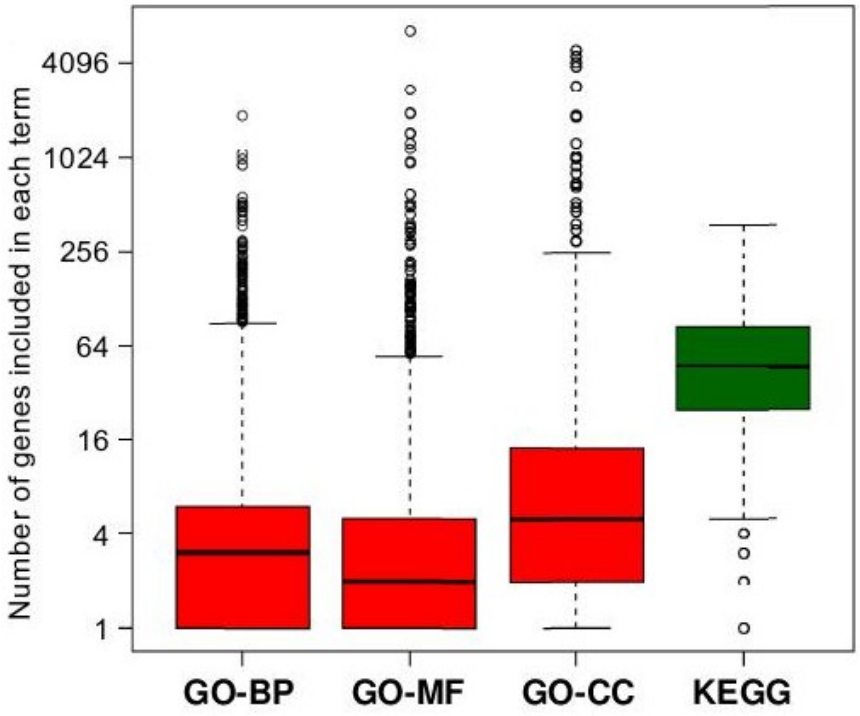


showing most **frequent terms** in GO and KEGG

Beyond Enrichment Analysis (EA)

Analysis of the distribution of **genes per term** in different annotation spaces

Distributions of number genes per term in each annotation space



GO terms with a very high number of genes annotated that have **low-information-content** and easily become **over-represented**

GO	Terms (annotations)	Genes (N)
GO-BP		
GO:0007165	signal transduction	**1872**
GO:0006355	regulation of transcription, DNA dep.	**1063**
GO:0045449	regulation of transcription	**987**
GO-MF		
GO:0005515	protein binding	**6618**
GO:0046872	metal ion binding	**2759**
GO:0008270	zinc ion binding	**1956**
GO-CC		
GO:0005634	nucleus	**4877**
GO:0005737	cytoplasm	**4457**
GO:0016021	integral to membrane	**4109**

Beyond
Enrichment Analysis (EA)

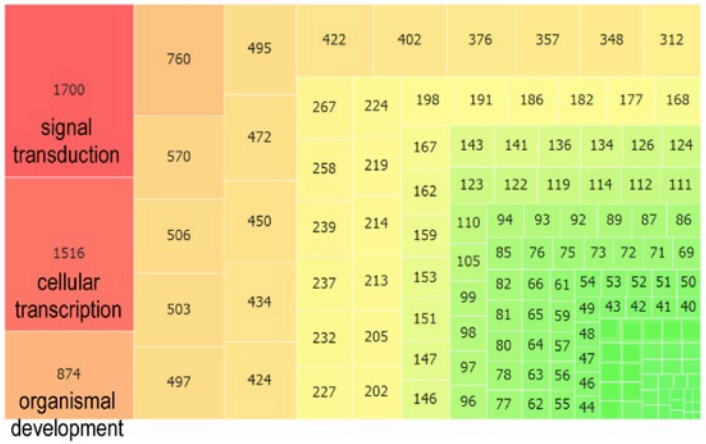
problems:

Distribution of genes per term (GeneOntology terms and Pathway terms)

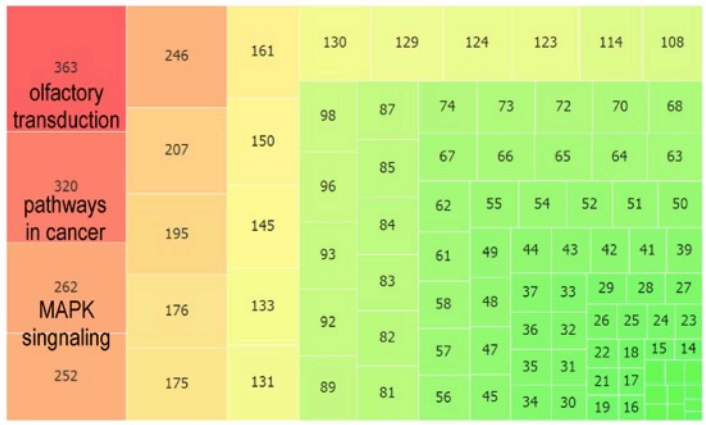
Most Frequent Terms
(highly used ≈ low-information-content)

GO-BP	N genes	Terms
GO:0006355	**1869**	regulat. cellular transcription
GO:0007165	**1700**	signal transduction
GO:0006350	**1516**	cellular transcription
GO-MF		
GO:0005515	**5306**	protein binding
GO:0046872	**2370**	metal ion binding
GO:0008270	**2251**	zinc ion binding
GO-CC		
GO:0005634	**4930**	nucleus
GO:0005737	**4264**	cytoplasm
GO:0016021	**4013**	integral to membrane

Comparative gene-size of the annotation-terms that include at least 3 genes (GO-BP *versus* KEGG)



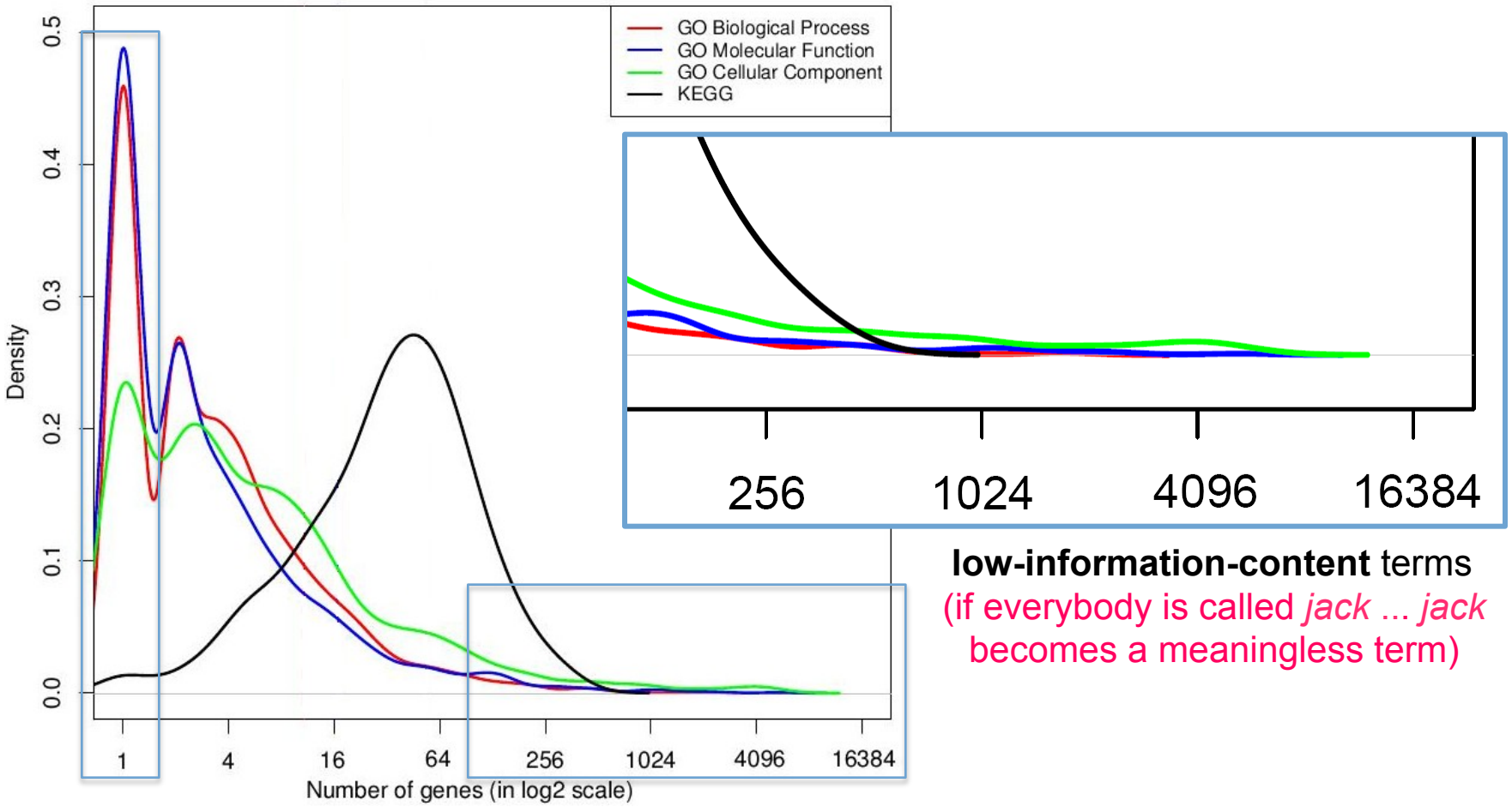
GO-BP (*human*) aprox.1800 terms
terms with **nº genes > [Med+2.5*SD]**
Highly used terms (red) ≈ low-information-content
33 terms that include **36.8** % of genes



KEGG (*human*) aprox. 200 terms
terms with **nº genes > [Med+2.5*SD]**
Highly used terms (red) ≈ low-information-content
9 terms that include **20.8** % of genes

Beyond Enrichment Analysis (EA)

Analysis of the distribution of **genes per term** in different annotation spaces



low-information-content terms
(if everybody is called *jack* ... *jack* becomes a meaningless term)

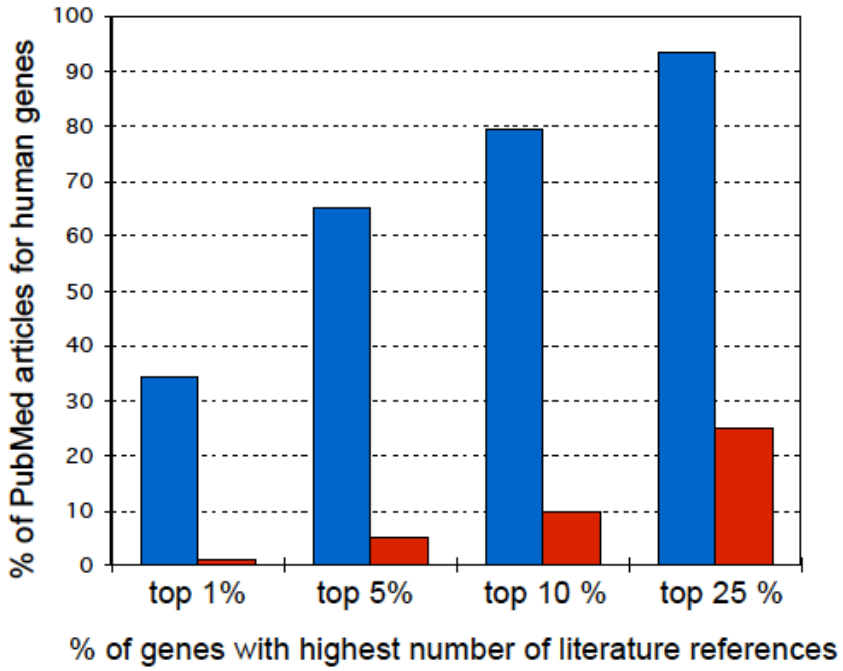
Distribution of **terms and genes** in different annotation spaces

Beyond
Enrichment Analysis (EA)

problems: there is also a clear **knowledge bias** in the scientific **literature**

Genes (biological entities):
Distribution of Genes counting number of articles (PubMed)

A

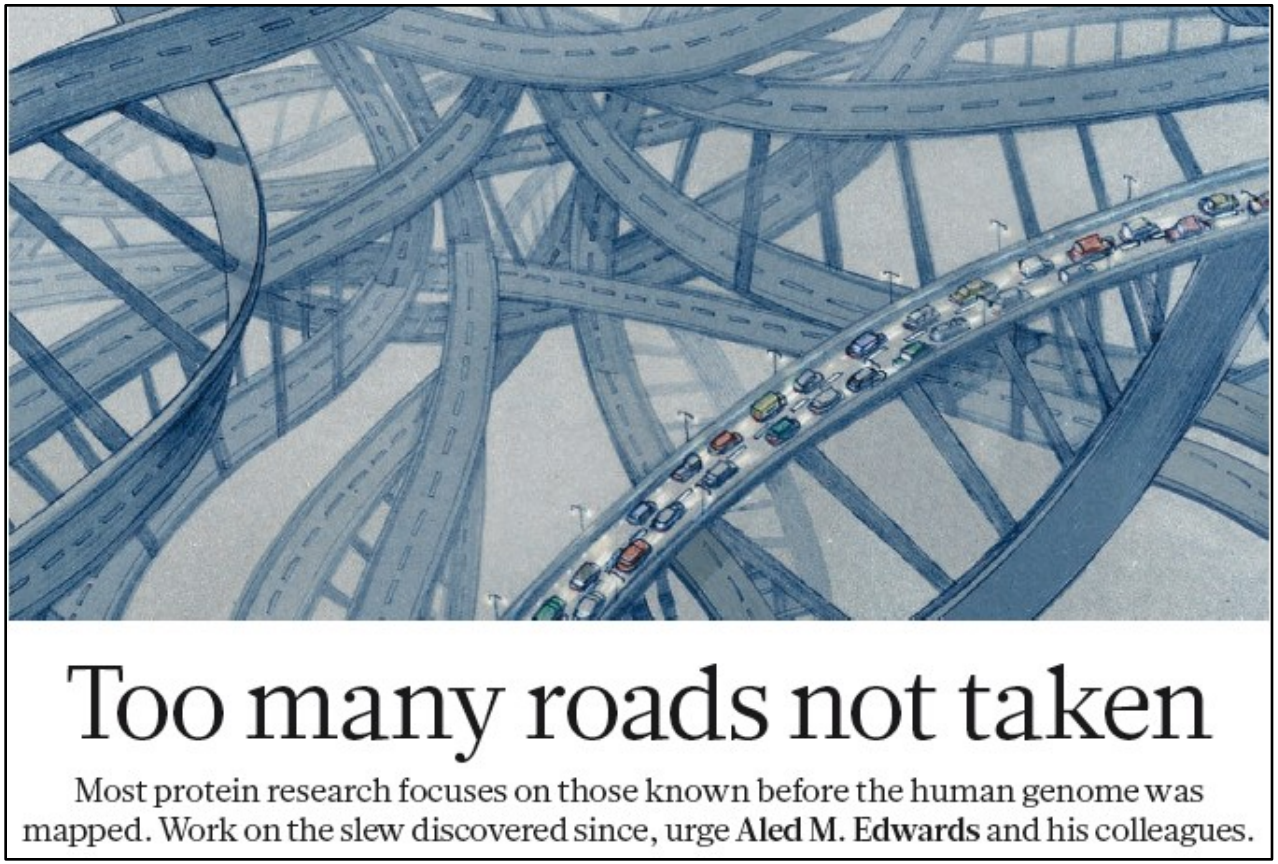


B

	Genes	N articles (PubMed)
1	TP53	**26891**
2	TNF (TNF-a)	**21616**
3	INS	**20298**
4	NFKB1	**16441**
5	FOS (c-FOS)	**16164**
6	MYC (c-MYC)	**12839**
7	CD4	**12427**
8	TGFB1	**12207**
9	IFNG	**12185**
10	IL1B	**11767**
11	EGF	**10493**
12	BCL2	**10123**
13	CALCA	**10006**
14	IL6	**9321**
15	IL2	**9186**
16	VEGFA	**8651**
17	ESR1	**8564**
18	CAT	**8274**
19	FOSB	**7227**
20	IL4	**7013**

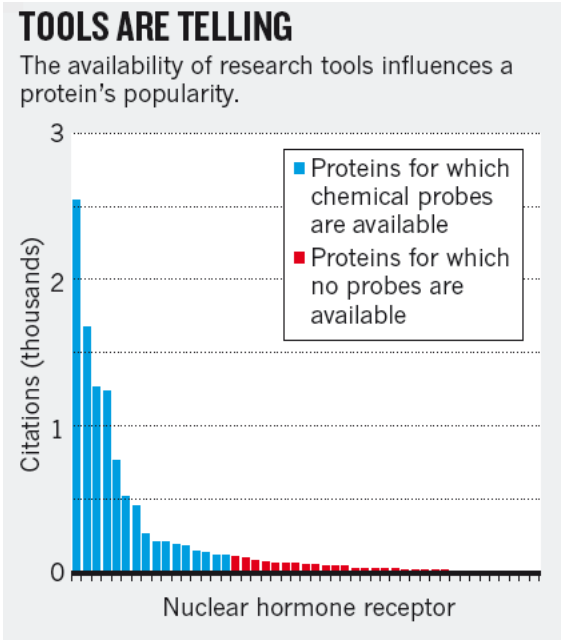
Beyond
Enrichment Analysis (EA)

problems: there is also a clear **knowledge bias** in the scientific **literature**



Too many roads not taken

Most protein research focuses on those known before the human genome was mapped. Work on the slew discovered since, urge **Aled M. Edwards** and his colleagues.

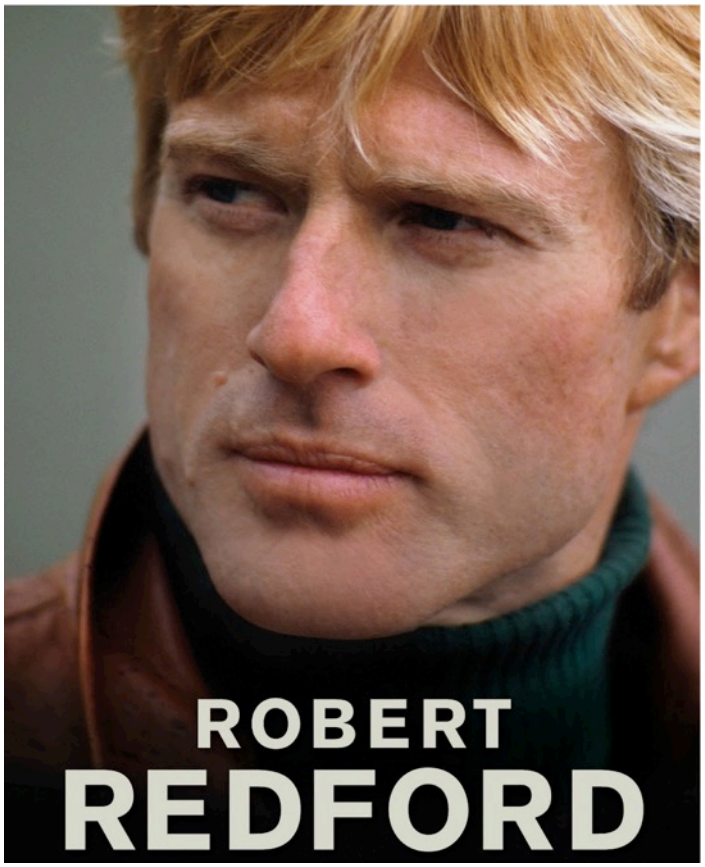


Edwards et al. (2011) SCIENCE

Beyond
Enrichment Analysis (EA)

problems: knowledge bias in the scientific **literature** === everybody wants to publish about **celebrity** proteins ('**robert redford**' et al.)

Genes (biological entities):
Distribution of Genes counting number of articles (PubMed)



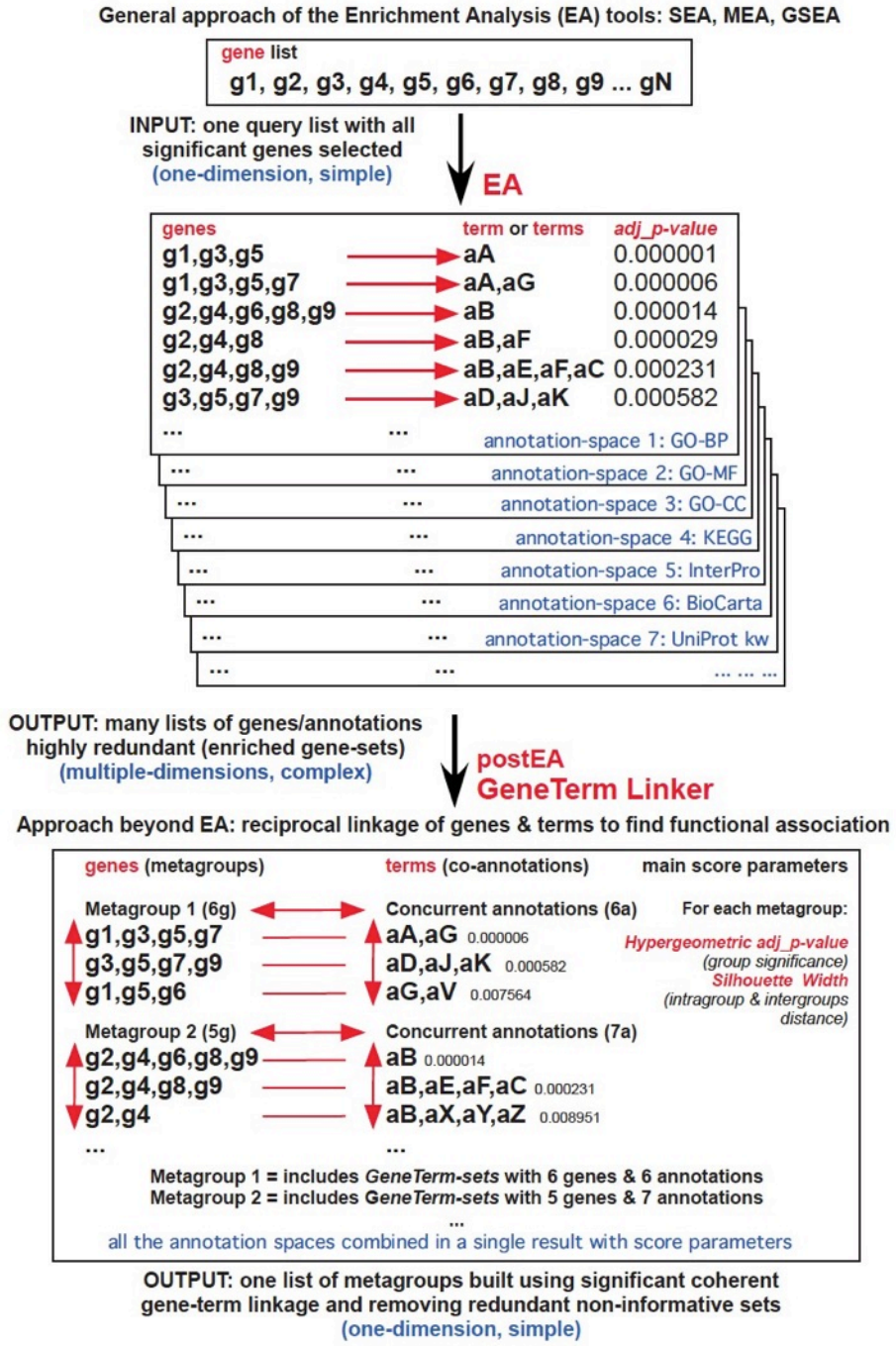
	Genes	N articles (PubMed)
1	TP53	**26891**
2	TNF (TNF-a)	**21616**
3	INS	**20298**
4	NFKB1	**16441**
5	FOS (c-FOS)	**16164**
6	MYC (c-MYC)	**12839**
7	CD4	**12427**
8	TGFB1	**12207**
9	IFNG	**12185**
10	IL1B	**11767**
11	EGF	**10493**
12	BCL2	**10123**
13	CALCA	**10006**
14	IL6	**9321**
15	IL2	**9186**
16	VEGFA	**8651**
17	ESR1	**8564**
18	CAT	**8274**
19	FOSB	**7227**
20	IL4	**7013**

Beyond Enrichment Analysis (EA)

solution:

design a new system of multiple EA that avoids redundant terms and provides a **unified result** based in functional convergence of **genes** & **terms** =

GeneTerm Linker *tool*

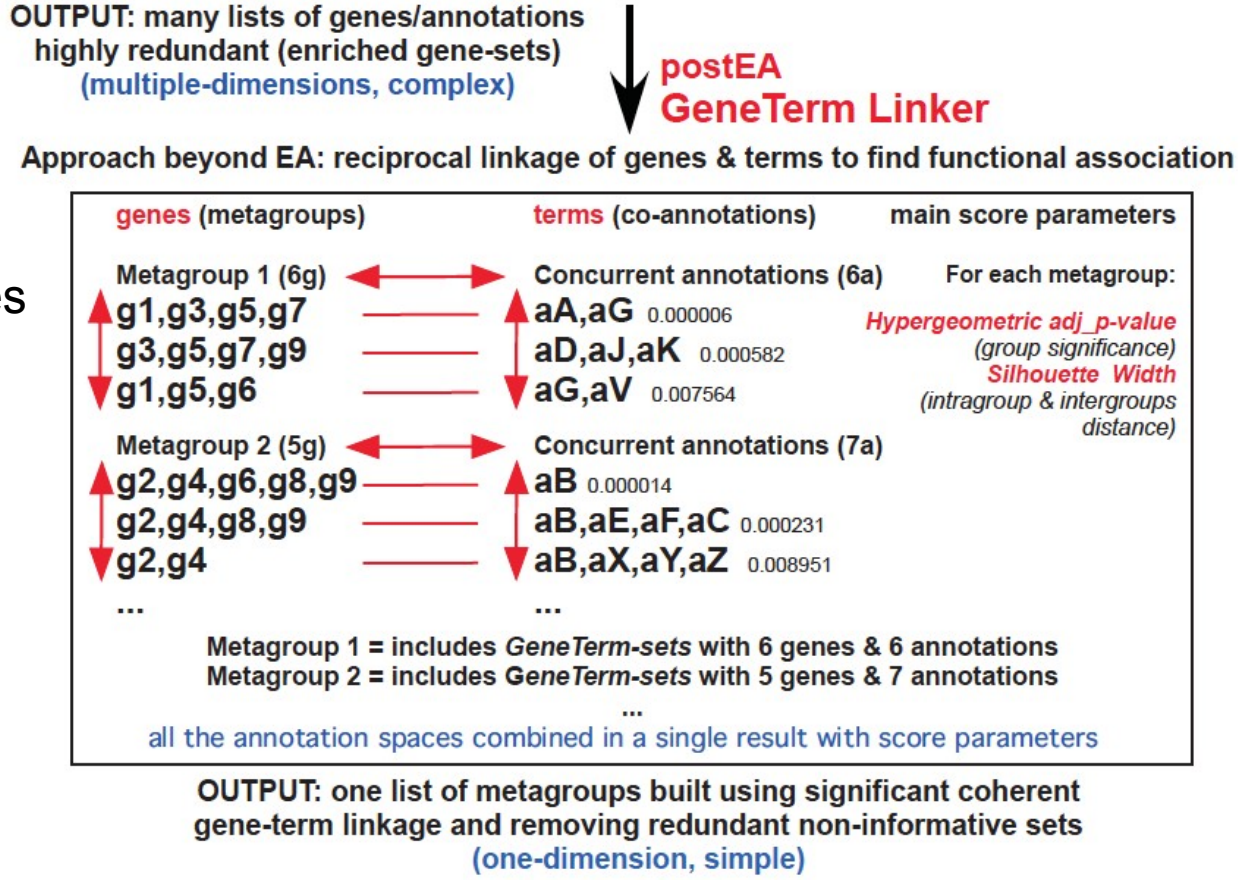


Beyond
Enrichment Analysis (EA)

solution:

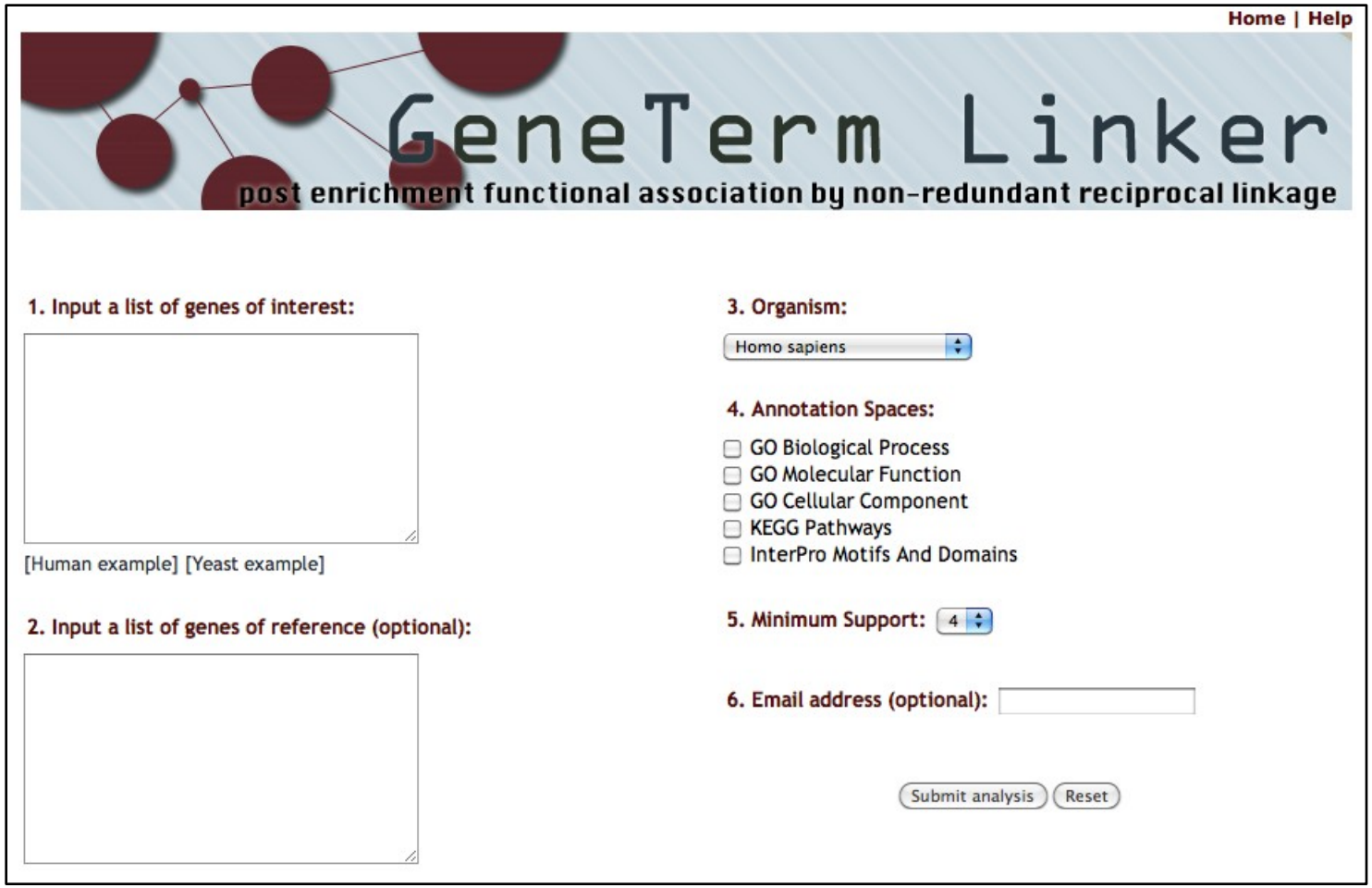
design a new system of multiple EA that avoids redundant terms and provides a **unified result** based in functional convergence of **genes** & **terms** =

GeneTerm Linker *tool*



Beyond
Enrichment Analysis (EA)

solution: ***GeneTerm Linker*** *tool*



Functional Enrichment Analysis (EA)

singular EA = SEA :: DAVID http://david.abcc.ncifcrf.gov/
selected gene list used to query different annotation spaces (one by one)

gene set EA = GSEA :: GSEA http://www.broadinstitute.org/gsea/
entire genes ranked used to query different annotation spaces (one by one)

modular EA = MEA :: GeneCodis http://genecodis.dacya.ucm.es/
selected gene list used to query multiple annotation spaces (at once)

***Beyond* EA :: GeneTerm Linker** http://gtlinker.dacya.ucm.es/
selected gene list used to query multiple annotation spaces (at once) avoiding repetition and redundancy and linking genes and terms

CBO Computational Biology @ UCT

CiC

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Hands-on: Practical Examples

Protein_ SETs_ 2014.xls

(106g hs, 175g hs, yeast 11pathways, yeast 59g5pc)

run *GeneTerm Linker*

CBIO Computational Biology @ UCT

CiC

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Session 1 (9:30 - 12:30, 3h)
Bioinformatic tools for Functional Enrichment Analysis (FEA)
Session 2 (13:30 - 16:30, 3h)
Construction of gene functional networks

- Introduction to biological information and annotation spaces:
GO, KEGG, Interpro

- Functional Enrichment Analysis (EA): from single to modular methods

- Using EA tools to annotate gene lists:
DAVID (single), GSEA (gene sets), GeneCodis (modular)

- Sort out problems after EA: post-enrichment tool
GeneTermLinker (postEA)

- From co-annotation and enrichment to functional networks:
networks construction using a R tool

GeneTerm Linker
a post enrichment tool

OPEN ACCESS Freely available online

Fontanillo et al. (2011) PLoS ONE

PLoS one

Functional Analysis beyond Enrichment: Non-Redundant Reciprocal Linkage of Genes and Biological Terms

Celia Fontanillo[1☯], Ruben Nogales-Cadenas[2☯], Alberto Pascual-Montano[2], Javier De Las Rivas[1*]

1 Cancer Research Center (CiC-IBMCC, CSIC/USAL), Campus Miguel de Unamuno, Salamanca, Spain, **2** National Center of Biotechnology (CNB, CSIC), Campus de Cantoblanco UAM, Madrid, Spain

Abstract

Functional analysis of large sets of genes and proteins is becoming more and more necessary with the increase of experimental biomolecular data at *omic*-scale. Enrichment analysis is by far the most popular available methodology to derive functional implications of sets of cooperating genes. The problem with these techniques relies in the redundancy of resulting information, that in most cases generate lots of trivial results with high risk to mask the reality of key biological events. We present and describe a computational method, called ***GeneTerm Linker***, that filters and links enriched output data identifying sets of associated genes and terms, producing metagroups of coherent biological significance. The method uses fuzzy reciprocal linkage between genes and terms to unravel their functional convergence and associations. The algorithm is tested with a small set of well known interacting proteins from yeast and with a large collection of reference sets from three heterogeneous resources: multiprotein complexes (CORUM), cellular pathways (SGD) and human diseases (OMIM). Statistical *Precision*, *Recall* and balanced *F-score* are calculated showing robust results, even when different levels of random noise are included in the test sets. Although we could not find an equivalent method, we present a comparative analysis with a widely used method that combines enrichment and functional annotation clustering. A web application to use the method here proposed is provided at http://gtlinker.cnb.csic.es.

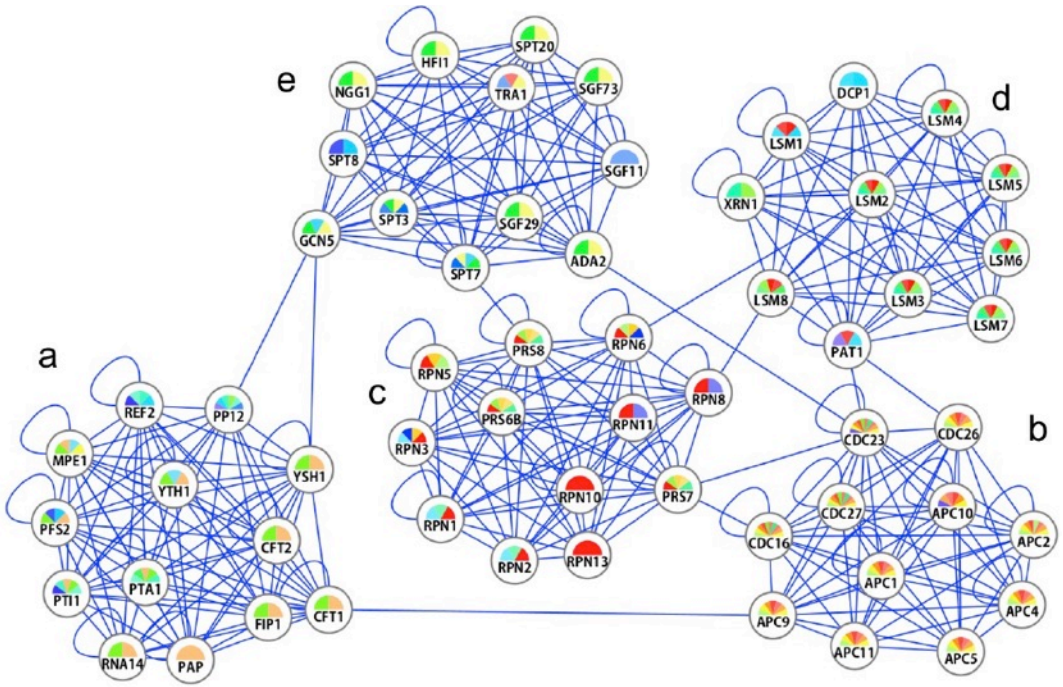
Citation: Fontanillo C, Nogales-Cadenas R, Pascual-Montano A, De Las Rivas J (2011) Functional Analysis beyond Enrichment: Non-Redundant Reciprocal Linkage of Genes and Biological Terms. PLoS ONE 6(9): e24289. doi:10.1371/journal.pone.0024289

Validation example

Protein Complexes (yeast)	Number of proteins
a. mRNA cleavage and polyadenylation specificity factor complex CFT1,CFT2,FIP1,GLC7,MPE1,PAP1,PFS2,PTA1,PTI1,REF2,RNA14,YSH1,YTH1	13
b. anaphase-promoting complex APC1,APC2,APC4,APC5,APC9,APC11,CDC16,CDC23,CDC26,CDC27,DOC1	11
c. proteasome, 19/22S regulator complex RPN1,RPN2,RPN3,RPN5,RPN6,RPN8,RPN10,RPN11,RPN13,RPT1,RPT3,RPT6	12
d. U6 snRNP complex DCP1,KEM1,LSM1,LSM2,LSM3,LSM4,LSM5,LSM6,LSM7,LSM8,PAT1	11
e. SAGA complex ADA2,GCN5,HFI1,NGG1,SGF11,SGF29,SGF73,SPT3,SPT7,SPT8,SPT20,TRA1	12

Testing the method with a set of **59** yeast nuclear **proteins** working in **5** known **complexes**

Bader et al (2003)



Reconstruction of the complexes using experimental protein interaction data (from APID and APID2NET)

Prieto et al (2006)
Hernández-Toro et al (2007)

Validation example

Query list of 59 yeast proteins

GeneTerm Linker results: **5 metagroups**

Gene Metagroups found (include all the related groups)	GENES found	p---value	Co---AnnotaGons (GO---BP, GO---MF, GO---CC, KEGG, InterPro)
Metagroup 1; SilhouePe: 0.529 GLC7, REF2, YTH1, FIP1, PAP1, PFS2, CFT1, RNA14, PTI1, PTA1, MPE1, CFT2, YSH1	**13**(59)	2.26E---26	GO:0005847:mRNA cleavage and polyadenylaLon specificity factor complex (CC), GO:0006378:mRNA polyadenylaLon (BP), GO:0006379:mRNA cleavage (BP), GO:0031123:RNA 3'---end processing (BP), GO:0006353:transcripLon terminaLon (BP), …
Metagroup 2; SilhouePe: 0.823 CDC23, APC5, CDC16, APC2, APC1, DOC1, APC9, APC11, APC4, CDC27, CDC26, GLC7, TRA1	**11**(59)	4.85E---24	04111:Cell cycle – yeast, GO:0004842:ubiquiLn---protein ligase acLvity (MF),GO:0000070:mitoLc sister chromaLd segregaLon (BP),GO:0016567:protein ubiquiLnaLon (BP),GO:0000022:mitoLc spindle elongaLon (BP),GO:0005680:anaphase---promoLng complex (CC), …
Metagroup 3; SilhouePe: 0.609 RPN13, RPN8, RPN1, RPT1, RPN3, RPN10, RPN11, RPN2, RPN5, RPT3, RPT6, RPN6, APC2, DOC1	**14**(59)	8.11E---15	GO:0000502:proteasome complex (CC), 03050:Proteasome, GO:0034515:proteasome storage granule (CC), GO:0008541:proteasome regulatory parLcle, lid subcomplex (CC), GO:0004175:endopepLdase acLvity (MF), GO:0030234:enzyme regulator acLvity (MF)
Metagroup 4; SilhouePe: 0.750 LSM5, DCP1, PAP1, LSM3, LSM8, LSM6, LSM1, LSM4, LSM7, LSM2, PTA1, KEM1, PAT1, YSH1	**14**(59)	1.31E---14	03018:RNA degradaLon, GO:0008033:tRNA processing (BP), GO:0030529:ribonucleoprotein complex (CC), GO:0046540:U4/U6 x U5 tri---snRNP complex (CC), GO:0000398:nuclear mRNA splicing, via spliceosome (BP), IPR001163:Like---Sm ribonucleoprotein (LSM), GO:0005688:U6 snRNP (CC), …
Metagroup 5; SilhouePe: 0.570 NGG1, HFI1, TRA1, SPT20, SGF29, SPT7, SGF73, SPT8, SGF11, GCN5, SPT3, ADA2, RPN6, CFT2	**14**(59)	5.26E---12	GO:0000124:SAGA complex (CC), GO:0016568:chromaLn modificaLon (BP), GO:0016573:histone acetylaLon (BP), GO:0046695:SLIK (SAGA---like) complex (CC), GO:0003702:RNA polymerase II transcripLon factor acLvity (MF), GO: 0004402:histone acetyltransferase acLvity (MF),…

Validation example

a. mRNA cleavage and polyadenilation specificity factor complex: **13 genes**

Gene Metagroups found (include all the related groups)	GENES found	p---value	Co---AnnotaGons (GO---BP, GO---MF, GO---CC, KEGG, InterPro)
Metagroup 1; SilhouePe: 0.529 GLC7, REF2, YTH1, FIP1, PAP1, PFS2, CFT1, RNA14, PTI1, PTA1, MPE1, CFT2, YSH1	**13**(59) (BP),	2.26E---26	GO:0005847:**mRNA cleavage and polyadenylaGon specificity factor complex** (CC), GO:0006378:**mRNA polyadenylaGon** (BP), GO:0006379:mRNA cleavage GO:0031123:RNA 3'---end processing (BP), GO: 0006353:transcripLon terminaLon (BP), …
Metagroup 2; SilhouePe: 0.823 CDC23, APC5, CDC16, APC2, APC1, DOC1, APC9, APC11, APC4, CDC27, CDC26, GLC7, TRA1	**11**(59)	4.85E---24	00[illegible]:Cell cycle – yeast, GO:0004842:ubiquiLn---protein ligase acLvity (MF),GO:0000070:mitoLc sister chromaLd segregaLon (BP),GO:0016567:protein ubiquiLnaLon (BP),GO:0000022:mitoLc spindle (BP),GO:0005680:anaphase---promoLng complex (CC), …
Metagroup 3; SilhouePe: 0.609 RPN13, RPN8, RPN1, RPT1, RPN3, RPN10, RPN11, RPN2, RPN5, RPT3, RPT6, RPN6, APC2, DOC1	**14**(59)	8.11E---15	GO:0000502:proteasome complex (CC), 03050:Proteasome, GO: 0034515:proteasome storage granule (CC), GO:0008541:proteasome regulatory parLcle, lid subcomplex (CC), GO:0004175:endopepLdase acLvity (MF), GO:0030234:enzyme regulator acLvity (MF)
Metagroup 4; SilhouePe: 0.750 LSM5, DCP1, PAP1, LSM3, LSM8, LSM6, LSM1, LSM4, LSM7, LSM2, PTA1, KEM1, PAT1, YSH1	**14**(59)	1.31E---14	03018:RNA degradaLon, GO:0008033:tRNA processing (BP), GO: 0030529:ribonucleoprotein complex (CC), GO:0046540:U4/U6 x U5 tri---snRNP complex (CC), GO:0000398:nuclear mRNA splicing, via spliceosome (BP), IPR001163:Like---Sm ribonucleoprotein (LSM), GO:0005688:U6 snRNP (CC), …
Metagroup 5; SilhouePe: 0.570 NGG1, HFI1, TRA1, SPT20, SGF29, SPT7, SGF73, SPT8, SGF11, GCN5, SPT3, ADA2, RPN6, CFT2	**14**(59)	5.26E---12	GO:0000124:SAGA complex (CC), GO:0016568:chromaLn modificaLon (BP), GO:0016573:histone acetylaLon (BP), GO:0046695:SLIK (SAGA---like) complex (CC), GO:0003702:RNA polymerase II transcripLon factor acLvity (MF), GO: 0004402:histone acetyltransferase acLvity (MF),…

Validation example

b. Anaphase-promoting complex: **11 genes**

Gene Metagroups found (include all the related groups)	GENES found	p---value	Co---AnnotaGons (GO---BP, GO---MF, GO---CC, KEGG, InterPro)
Metagroup 1; SilhouePe: 0.529 GLC7, REF2, YTH1, FIP1, PAP1, PFS2, CFT1, RNA14, PTI1, PTA1, MPE1, CFT2, YSH1	**13**(59) 0006353	2.26E---26 transcripLon	GO:0005847:mRNA cleavage and polyadenylaLon specificity factor complex (CC), GO:0006378:mRNA polyadenylaLon (BP), GO:0006379:mRNA cleavage (BP), GO:0031123:RNA 3'---end processing (BP), GO: terminaLon (BP), …
Metagroup 2; SilhouePe: 0.823 CDC23, APC5, CDC16, APC2, APC1, DOC1, APC9, APC11, APC4, CDC27, CDC26, GLC7, TRA1	**11**(59) elongaLon	4.85E---24	04111:Cell cycle – yeast, GO:0004842:ubiquiLn--- protein ligase acLvity (MF), GO:0000070: mitoGc sister chromaGd segregaGon (BP), GO:0000022:mitoLc spindle (BP),GO:0005680:anaphase---promoGng complex (CC), …
Metagroup 3; SilhouePe: 0.609 RPN13, RPN8, RPN1, RPT1, RPN3, RPN10, RPN11, RPN2, RPN5, RPT3, RPT6, RPN6, APC2, DOC1	**14**(59) 15	8.11E---	GO:0000502:proteasome complex (CC), 03050:Proteasome, GO: 0034515:proteasome storage granule (CC), GO:0008541:proteasome regulatory parLcle, lid subcomplex (CC), GO:0004175:endopepLdase acLvity (MF), GO:0030234:enzyme regulator acLvity (MF)
Metagroup 4; SilhouePe: 0.750 LSM5, DCP1, PAP1, LSM3, LSM8, LSM6, LSM1, LSM4, LSM7, LSM2, PTA1, KEM1, PAT1, YSH1	**14**(59) 14	1.31E---	03018:RNA degradaLon, GO:0008033:tRNA processing (BP), GO: 0030529:ribonucleoprotein complex (CC), GO:0046540:U4/U6 x U5 tri--- snRNP complex (CC), GO:0000398:nuclear mRNA splicing, via spliceosome (BP), IPR001163:Like---Sm ribonucleoprotein (LSM), GO:0005688:U6 snRNP (CC), …
Metagroup 5; SilhouePe: 0.570 NGG1, HFI1, TRA1, SPT20, SGF29, SPT7, SGF73, SPT8, SGF11, GCN5, SPT3, ADA2, RPN6, CFT2	**14**(59) 12	5.26E---	GO:0000124:SAGA complex (CC), GO:0016568:chromaLn modificaLon (BP), GO:0016573:histone acetylaLon (BP), GO:0046695:SLIK (SAGA---like) complex (CC), GO:0003702:RNA polymerase II transcripLon factor acLvity (MF), GO: 0004402:histone acetyltransferase acLvity (MF),…

Validation example

Gene Metagroups found (include all the related groups)	GENES found	p---value	Co---AnnotaGons (GO---BP, GO---MF, GO---CC, KEGG, InterPro)
Metagroup 1; SilhouePe: 0.529 GLC7, REF2, YTH1, FIP1, PAP1, PFS2, CFT1, RNA14, PTI1, PTA1, MPE1, CFT2, YSH1	**13**(59) 0006353	2.26E---26	GO:0005847:mRNA cleavage and polyadenylaLon specificity factor complex (CC), GO:0006378:mRNA polyadenylaLon (BP), GO:0006379:mRNA cleavage (BP), GO:0031123:RNA 3'---end processing (BP), GO: terminaLon (BP), …
Metagroup 2; SilhouePe: 0.823 CDC23, APC5, CDC16, APC2, APC1, DOC1, APC9, APC11, APC4, CDC27, CDC26, GLC7, TRA1	**11**(59)	4.85E---24	04111:Cell cycle – yeast, GO:0004842:ubiquiLn---protein ligase acLvity (MF),GO:0000070:mitoLc sister chromaLd segregaLon (BP),GO:0016567:protein ubiquiLnaLon (BP),GO:0000022:mitoLc spindle (BP),GO:0005680:anaphase---promoLng complex (CC), …
Metagroup 3; SilhouePe: 0.609 RPN13, RPN8, RPN1, RPT1, RPN3, RPN10, RPN11, RPN2, RPN5, RPT3, RPT6, RPN6, APC2, DOC1	**14**(59) 15	8.11E---	GO:0000502:**proteasome complex** (CC), 03050:**Proteasome**, GO:0034515:**proteasome** storage granule (CC), GO:0008541:proteasome regulatory parLcle, lid subcomplex (CC), GO: 0004175:endopeptidase acLvity (MF), GO:0030234:enzyme regulator acLvity (MF)
Metagroup 4; SilhouePe: 0.750 LSM5, DCP1, PAP1, LSM3, LSM8, LSM6, LSM1, LSM4, LSM7, LSM2, PTA1, KEM1, PAT1, YSH1	**14**(59) 14	1.31E---	03019:RNA degradaLon, GO:0008033:tRNA processing (BP), GO: 0030529:ribonucleoprotein complex (CC), GO:0046540:U4/U6 x U5 tri---snRNP complex (CC), GO:0000398:nuclear mRNA splicing, via spliceosome (BP), IPR001163:Like---Sm ribonucleoprotein (LSM), GO:0005688:U6 snRNP (CC), …
Metagroup 5; SilhouePe: 0.570 NGG1, HFI1, TRA1, SPT20, SGF29, SPT7, SGF73, SPT8, SGF11, GCN5, SPT3, ADA2, RPN6, CFT2	**14**(59) 12	5.26E---	GO:0000124:SAGA complex (CC), GO:0016568:chromaLn modificaLon (BP), GO:0016573:histone acetylaLon (BP), GO:0046695:SLIK (SAGA---like) complex (CC), GO:0003702:RNA polymerase II transcripLon factor acLvity (MF), GO: 0004402:histone acetyltransferase acLvity (MF),…

c. proteasome, regulator complex: **12 genes**

Validation example

Gene Metagroups found (include all the related groups)	GENES found	p---value	Co---AnnotaGons (GO---BP, GO---MF, GO---CC, KEGG, InterPro)
Metagroup 1; SilhouePe: 0.529 GLC7, REF2, YTH1, FIP1, PAP1, PFS2, CFT1, RNA14, PTI1, PTA1, MPE1, CFT2, YSH1	**13**(59) 0006353	2.26E---26 transcripLon	GO:0005847:mRNA cleavage and polyadenylaLon specificity factor complex (CC), GO:0006378:mRNA polyadenylaLon (BP), GO:0006379:mRNA cleavage (BP), GO:0031123:RNA 3'---end processing (BP), GO: terminaLon (BP), …
Metagroup 2; SilhouePe: 0.823 CDC23, APC5, CDC16, APC2, APC1, DOC1, APC9, APC11, APC4, CDC27, CDC26, GLC7, TRA1	**11**(59)	4.85E---24	04111:Cell cycle – yeast, GO:0004842:ubiquiLn---protein ligase acLvity (MF),GO:0000070:mitoLc sister chromaLd segregaLon (BP),GO:0016567:protein ubiquiLnaLon (BP),GO:0000022:mitoLc spindle (BP),GO:0005680:anaphase---promoLng complex (CC), …
Metagroup 3; SilhouePe: 0.609 RPN13, RPN8, RPN1, RPT1, RPN3, RPN10, RPN11, RPN2, RPN5, RPT3, RPT6, RPN6, APC2, DOC1	**14**(59) 15	8.11E---	GO:0000502:proteasome complex (CC), 03050:Proteasome, GO: 0034515:proteasome storage granule (CC), GO:0008541:proteasome regulatory parLcle, lid subcomplex (CC), GO:0004175:endopepLdase acLvity (MF), GO:0030234:enzyme regulator acLvity (MF)
Metagroup 4; SilhouePe: 0.750 LSM5, DCP1, PAP1, LSM3, LSM8, LSM6, LSM1, LSM4, LSM7, LSM2, PTA1, KEM1, PAT1, YSH1	**14**(59) 14	1.31E---	03018:RNA degradaLon, GO:0008033:tRNA processing (BP), GO:0030529:**ribonucleoprotein complex** (CC), GO: 0046540:**U4/U6 x U5 tri---snRNP complex** (CC), GO: 0000398:nuclear mRNA splicing, via spliceosome (BP), IPR001163:Like---Sm ribonucleoprotein (LSM), GO: 0005688:U6 snRNP (CC), …
Metagroup 5; SilhouePe: 0.570 NGG1, HFI1, TRA1, SPT20, SGF29, SPT7, SGF73, SPT8, SGF11, GCN5, SPT3, ADA2, RPN6, CFT2	**14**(59) 12	5.26E---	GO:0000124:SAGA complex (CC), GO:0016568:chromaLn modificaLon (BP), GO:0016573:histone acetylaLon (BP), GO:0046695:SLIK (SAGA---like) complex (CC), GO:0003702:RNA polymerase II transcripLon factor acLvity (MF), GO: 0004402:histone acetyltransferase acLvity (MF),…

d. U6 snRNP complex: **11 genes**

Validation example

Gene Metagroups found (include all the related groups)	GENES found	p---value	Co---AnnotaGons (GO---BP, GO---MF, GO---CC, KEGG, InterPro)
Metagroup 1; SilhouePe: 0.529 GLC7, REF2, YTH1, FIP1, PAP1, PFS2, CFT1, RNA14, PTI1, PTA1, MPE1, CFT2, YSH1	**13**(59) 0006353	2.26E---26 transcripLon	GO:0005847:mRNA cleavage and polyadenylaLon specificity factor complex (CC), GO:0006378:mRNA polyadenylaLon (BP), GO:0006379:mRNA cleavage (BP), GO:0031123:RNA 3'---end processing (BP), GO: terminaLon (BP), …
Metagroup 2; SilhouePe: 0.823 CDC23, APC5, CDC16, APC2, APC1, DOC1, APC9, APC11, APC4, CDC27, CDC26, GLC7, TRA1	**11**(59)	4.85E---24	04111:Cell cycle – yeast, GO:0004842:ubiquiLn---protein ligase acLvity (MF),GO:0000070:mitoLc sister chromaLd segregaLon (BP),GO:0016567:protein ubiquiLnaLon (BP),GO:0000022:mitoLc spindle (BP),GO:0005680:anaphase---promoLng complex (CC), …
Metagroup 3; SilhouePe: 0.609 RPN13, RPN8, RPN1, RPT1, RPN3, RPN10, RPN11, RPN2, RPN5, RPT3, RPT6, RPN6, APC2, DOC1	**14**(59) 15	8.11E---	GO:0000502:proteasome complex (CC), 03050:Proteasome, GO: 0034515:proteasome storage granule (CC), GO:0008541:proteasome regulatory parLcle, lid subcomplex (CC), GO:0004175:endopepLdase acLvity (MF), GO:0030234:enzyme regulator acLvity (MF)
Metagroup 4; SilhouePe: 0.750 LSM5, DCP1, PAP1, LSM3, LSM8, LSM6, LSM1, LSM4, LSM7, LSM2, PTA1, KEM1, PAT1, YSH1	**14**(59) 14	1.31E---	03018:RNA degradaLon, GO:0008033:tRNA processing (BP), GO: 0030529:ribonucleoprotein complex (CC), GO:0046540:U4/U6 x U5 tri---snRNP complex (CC), GO:0000398:nuclear mRNA splicing, via spliceosome (BP), IPR001163:Like---Sm ribonucleoprotein (LSM), GO:0005688:U6 snRNP (CC), …
Metagroup 5; SilhouePe: 0.570 NGG1, HFI1, TRA1, SPT20, SGF29, SPT7, SGF73, SPT8, SGF11, GCN5, SPT3, ADA2, RPN6, CFT2	**14**(59) 12	5.26E---	GO:0000124:**SAGA complex** (CC), GO:0016568:chromaLn modificaLon (BP), GO:0016573:histone acetylaLon (BP), GO:0046695:**SLIK (SAGA---like) complex** (CC), GO: 0003702:RNA polymerase II transcripLon factor acLvity (MF), GO:0004402:histone acetyltransferase acLvity (MF),…

e. SAGA complex: **12 genes**

Validation with other gene/protein sets from multiple biological sources

Use 3 sets of genes/proteins grouped by different biological criteria:

- Comprehensive Resource of Mammalian protein complexes (**CORUM**) genes in 10 mammal **multi-protein complexes**

- *Saccaromyces* Genome Database (**SGD**) genes in 10 known yeast **pathways**

- Online Mendelian Inheritance in Man (**OMIM**) genes in 10 human **diseases**

Validation with other gene/protein sets from multiple biological sources

Use 3 sets of genes/proteins grouped by different biological criteria:

- Comprehensive Resource of Mammalian protein complexes (**CORUM**) genes in 10 mammal **multi-protein complexes**
- *Saccaromyces* Genome Database (**SGD**) genes in 10 known yeast **pathways**
- Online Mendelian Inheritance in Man (**OMIM**) genes in 10 human **diseases**

We evaluate the signal considering noise tolerance (20% noise added):

- *Precision*
- *Recall*

$$Precission = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Validation with other gene/protein sets from multiple biological sources

	COMPLEXES (from CORUM db, human)	GENES in Ref.	GENES Tested	GENES Found	Common GENES	Precision (%)	Recall (%)	F-score (%)	adjusted p.value	TERMS Found	TERMS Found (only first shown)
1c	C complex spliceosome	80	96	68	68	100.00	85.00	91.89	5.25E-138	6	GO:0005681:spliceosomal …
2c	Mediator (transcriptional coactivator) complex	32	39	28	28	100.00	87.50	93.33	3.96E-064	10	GO:0016592:mediator complex …
3c	Proteasome (20S/26S)	22	27	22	22	100.00	100.00	100.00	5.34E-063	12	03050:Proteasome …
4c	RNA polymerase II (RNAPII)	26	32	24	24	100.00	92.31	96.00	1.21E-059	12	GO:0006350:transcription …
5c	F1F0-ATP synthase (EC 3.6.3.14), mitochondrial	16	20	14	14	100.00	87.50	93.33	2.63E-045	12	GO:0005753:mitochondrial ATPase ...
6c	DAB complex, transcription preinitiation complex	16	20	16	16	100.00	100.00	100.00	3.20E-042	6	03022:Basal transcription factors …
7c	Exosome	11	14	11	11	100.00	100.00	100.00	2.47E-040	4	GO:0006364:rRNA processing …
8c	eIF3 complex, eukaryotic initiation of translation factor-3	13	16	11	11	100.00	84.62	91.67	8.98E-038	4	GO:0005852:eukar. translation initiation factor …
9c	Nup 107-160 nuclear pore subcomplex	9	11	9	9	100.00	100.00	100.00	1.34E-033	6	GO:0005635:nuclear envelope …
10c	CENP-A NAC-CAD kinetochore complex	13	16	13	13	100.00	100.00	100.00	2.32E-028	2	GO:0005694:chromosome …
	with 20% noise (% of random genes included)		+20% noise		average values=	100.00	93.69	96.62			
	…										
	DISEASES (from OMIM db, human)	**GENES in Ref.**	**GENES Tested**	**GENES Found**	**Common GENES**	**Precision (%)**	**Recall (%)**	**F-score (%)**	**adjusted p.value**	**TERMS Found**	**TERMS Found (only first shown)**
1d	Retinitis pigmentosa	51	62	39	38	97.44	74.51	84.44	1.15E-066	1	GO:0007601:visual perception (BP)
2d	Deafness (autosomal dominant and recesive)	84	101	31	31	100.00	36.90	53.91	4.11E-053	6	GO:0007605:sensory perception of sound …
3d	Cardiomyopathy (dilated, familial and hypertrophic)	44	53	19	19	100.00	43.18	60.32	1.48E-031	5	GO:0008307:structural constituent of muscle …
4d	Epidermolysis bullosa (dystrophica, simplex, junctional)	11	14	11	11	100.00	100.00	100.00	2.77E-024	2	GO:0031581:hemidesmosome assembly …
5d	Congenital disorder of glycosylation (type I and II)	23	28	11	11	100.00	47.83	64.71	1.18E-023	1	00510:N-Glycan biosynthesis
6d	Muscular dystrophy (congenital, limb-girdle, rigid spine)	25	30	11	11	100.00	44.00	61.11	4.45E-019	4	GO:0007517:muscle organ development …
7d	Glycogen storage disease	19	23	9	9	100.00	47.37	64.29	2.53E-018	3	00500:Starch and sucrose metabolism …
8d	Leigh syndrome	8	10	7	7	100.00	87.50	93.33	1.16E-017	10	GO:0006120:mitochondrial electron transport …
9d	Acute Leukemia (lymphoblastic ALLs & myeloid AMLs)	37	45	9	9	100.00	24.32	39.13	7.71E-016	2	05221:Acute myeloid leukemia …
10d	Diabetes mellitus (type 1 or 2, gestational, neonatal)	13	16	5	5	100.00	38.46	55.56	1.01E-010	2	GO:0005975:carbohydrate metabolic process …
	with 20% noise (% of random genes included)		+20% noise		average values=	99.74	54.41	67.68			
	…										
	PATHWAYS (from SGD db, yeast)	**GENES in Ref.**	**GENES Tested**	**GENES Found**	**Common GENES**	**Precision (%)**	**Recall (%)**	**F-score (%)**	**adjusted p.value**	**TERMS Found**	**TERMS Found (only first shown)**
1p	gluconeogenesis	22	27	22	22	100.00	100.00	100.00	7.05E-047	7	GO:0006094:gluconeogenesis …
2p	TCA cycle, aerobic respiration	22	27	23	22	95.65	100.00	97.78	2.36E-037	10	00020:Citrate cycle (TCA cycle) …
3p	sphingolipid metabolism	19	23	16	16	100.00	84.21	91.43	7.64E-033	4	00600:Sphingolipid metabolism …
4p	de novo biosynthesis of purine nucleotides	35	42	21	21	100.00	60.00	75.00	2.02E-030	4	00230:Purine metabolism …
5p	lipid-linked oligosaccharide biosynthesis	12	15	12	12	100.00	100.00	100.00	1.38E-029	5	GO:0005783:endoplasmic reticulum …
6p	ergosterol biosynthesis	11	14	11	11	100.00	100.00	100.00	1.02E-028	7	GO:0006696:ergosterol biosynthetic process …
7p	superpathway of glucose fermentation	14	17	13	13	100.00	92.86	96.30	9.18E-027	5	00010:Glycolysis / Gluconeogenesis …
8p	fatty acid biosynthesis, initial steps	17	21	13	12	92.31	70.59	80.00	9.93E-027	6	GO:0006631:fatty acid metabolic process …
9p	inositol phosphate biosynthesis	19	23	11	11	100.00	57.89	73.33	1.02E-025	3	00562:Inositol phosphate metabolism …
10p	folate biosynthesis	18	22	10	9	90.00	50.00	64.29	6.98E-019	6	GO:0006730:one-carbon metabolic process …
	with 20% noise (% of random genes included)		+20% noise		average values=	97.80	81.56	87.81			
	…										

Session 1 (9:30 - 12:30, 3h)

Bioinformatic tools for Functional Enrichment Analysis (FEA)

Session 2 (13:30 - 16:30, 3h)

Construction of gene functional networks

- Introduction to biological information and annotation spaces: GO, KEGG, Interpro

- Functional Enrichment Analysis (EA): from single to modular methods

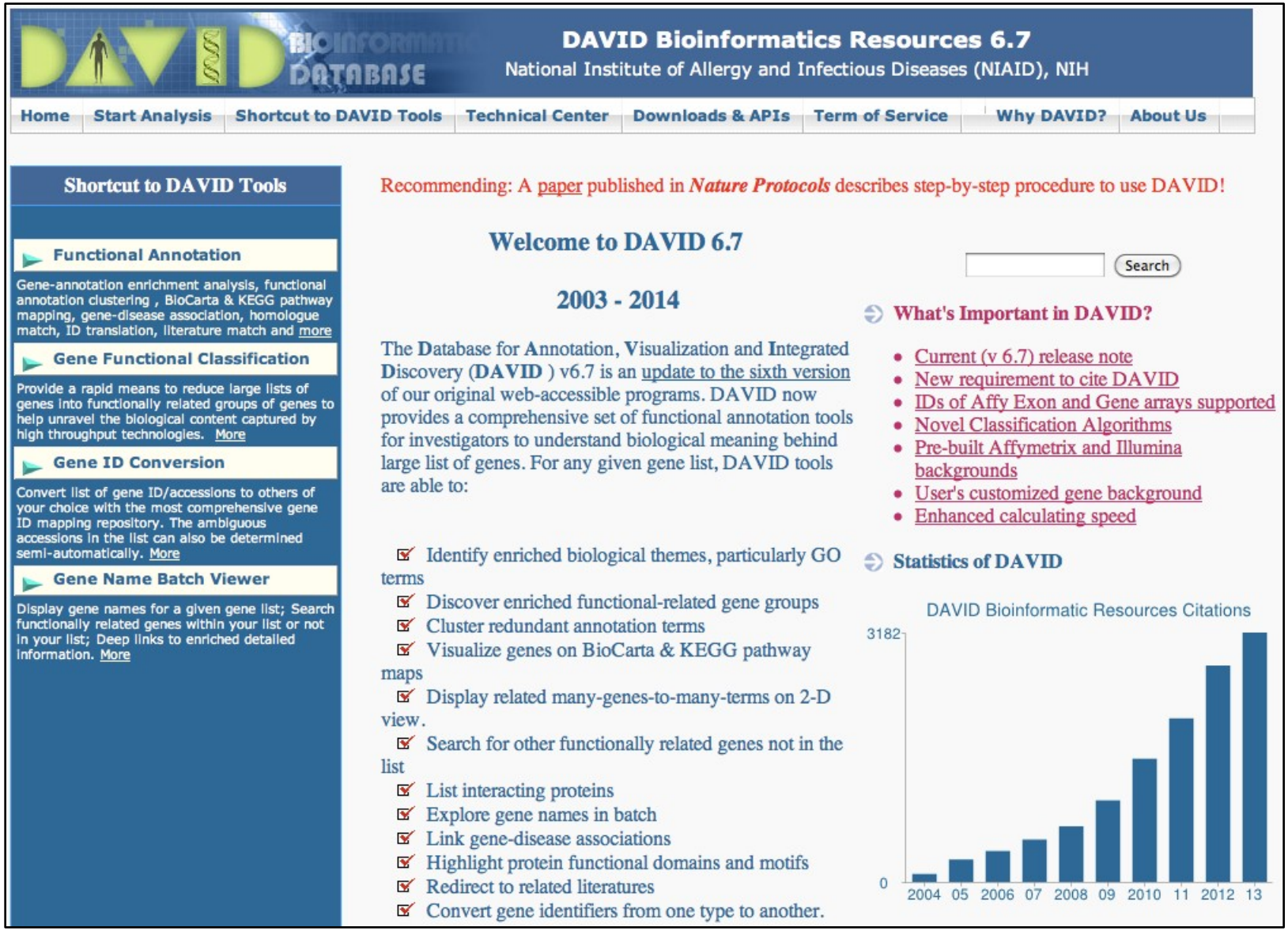
- Using EA tools to annotate gene lists: DAVID (single), GSEA (gene sets), GeneCodis (modular)

- Sort out problems after EA: post-enrichment tool GeneTermLinker (postEA) vs DAVID FAC (single+clustering)

- From co-annotation and enrichment to functional networks: networks construction using a R tool

Comparison
GeneTerm Linker vs *DAVID FAC*

http://david.abcc.ncifcrf.gov/



Comparison
GeneTerm Linker vs *DAVID FAC*

http://david.abcc.ncifcrf.gov/

Huang et al (2007) Genome Biol.

Software

Open Access

The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists

Da Wei Huang[¤*], Brad T Sherman[¤*], Qina Tan[*], Jack R Collins[†], W Gregory Alvord[‡], Jean Roayaei[‡], Robert Stephens[†], Michael W Baseler[§], H Clifford Lane[¶] and Richard A Lempicki[*]

Addresses: [*]Laboratory of Immunopathogenesis and Bioinformatics, Clinical Services Program, SAIC-Frederick, Inc., National Cancer Institute at Frederick, Frederick, Frederick, Frederick, M Frederick, MD 21702, U USA. [¶]Laboratory of Im 20892, USA.

Huang et al (2009) Nature Protoc.

PROTOCOL

Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources

Da Wei Huang[1,2], Brad T Sherman[1,2] & Richard A Lempicki[1]

[1]Laboratory of Immunopathogenesis and Bioinformatics, Clinical Services Program, SAIC-Frederick Inc., National Cancer Institute at Frederick, Frederick, Maryland 21702, USA. [2]These authors contributed equally to this work. Correspondence should be addressed to R.A.L. (rlempicki@mail.nih.gov) or D.W.H. (huangdawei@mail.nih.gov)

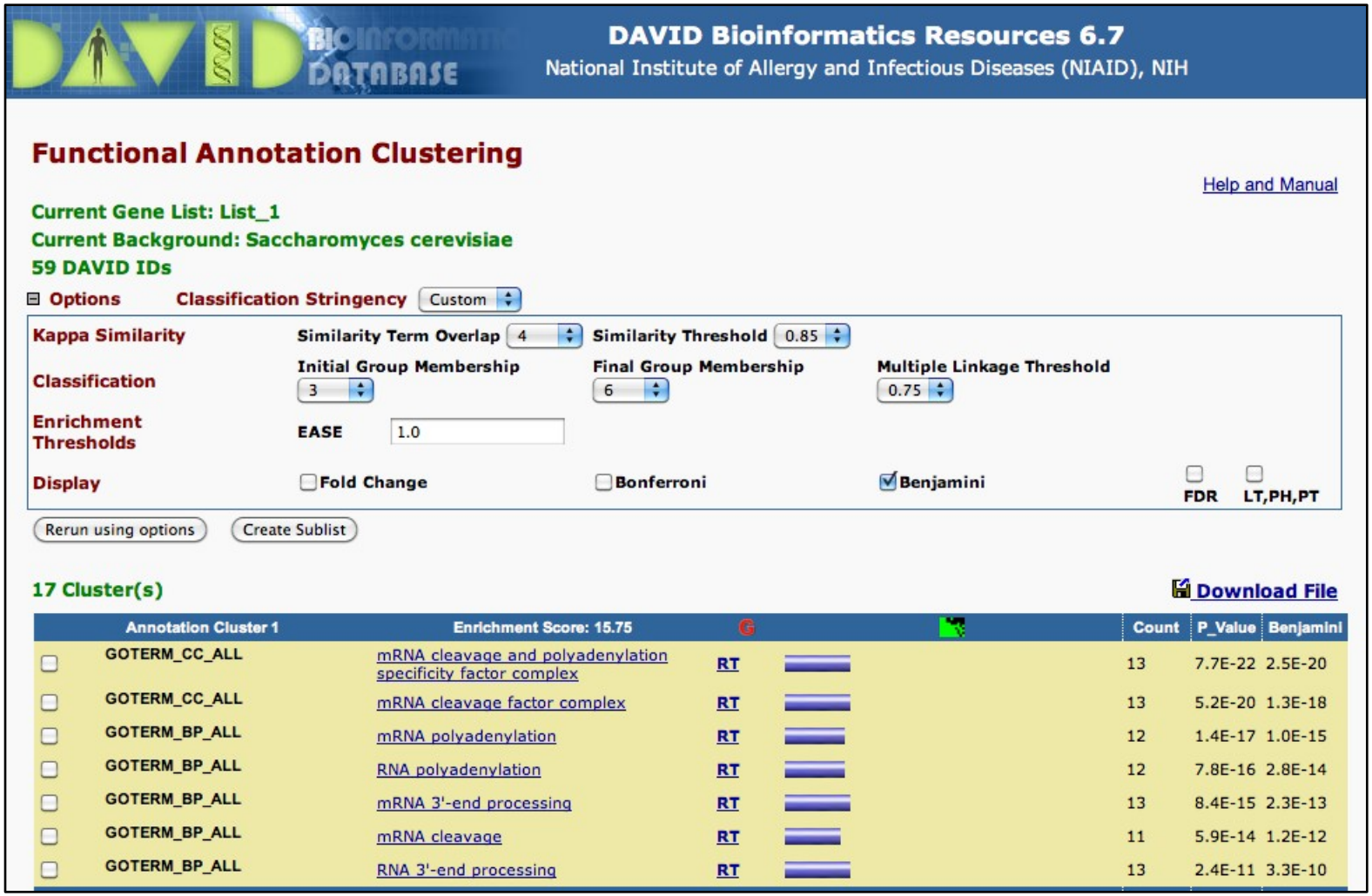
Published online 18 December 2008; doi:10.1038/nprot.2008.211

DAVID bioinformatics resources consists of an integrated biological knowledgebase and analytic tools aimed at systematically extracting biological meaning from large gene/protein lists. This protocol explains how to use DAVID, a high-throughput and integrated data-mining environment, to analyze gene lists derived from high-throughput genomic experiments. The procedure first requires uploading a gene list containing any number of common gene identifiers followed by analysis using one or more text and pathway-mining tools such as gene functional classification, functional annotation chart or clustering and functional annotation table. By following this protocol, investigators are able to gain an in-depth understanding of the biological themes in lists of genes that are enriched in genome-scale studies.

Comparison
GeneTerm Linker vs *DAVID FAC*

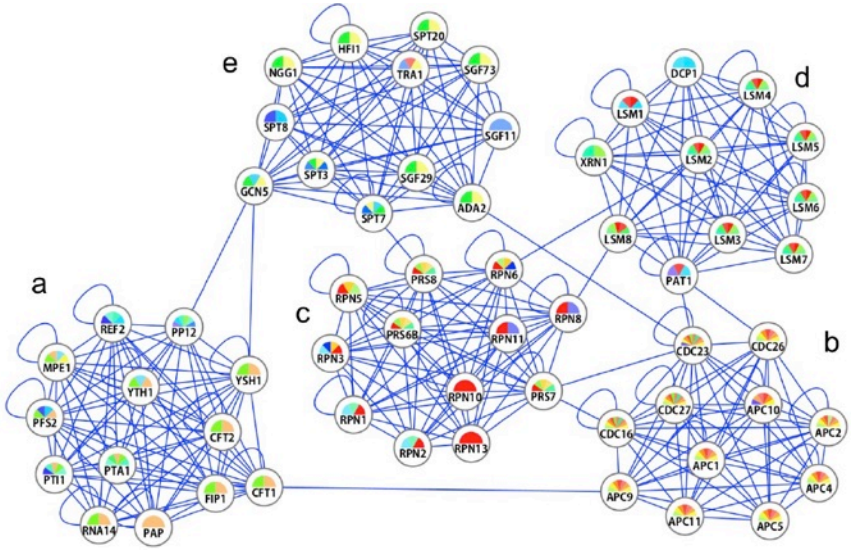
DAVID bioinformatics suite provides a **post-enrichment** tool:

DAVID Functional Annotation Clustering (DAVID - FAC), after single enrichment analysis it clusters the terms based in the genes that they share.



Comparison
GeneTerm Linker vs *DAVID FAC*

Comparative analysis of two approaches that provide groups of genes and terms



Testing a set of **59** yeast nuclear **proteins** working in **5** known **complexes**

	GeneTerm Linker	DAVID FAC (by default)	DAVID FAC (tuned to find 5 groups)
Total groups reference	5	5	5
Total groups found	5	15	5
Accuracy	**0.952**	**0.311**	**0.884**
Jaccard coefficient	**0.769**	**0.213**	**0.562**

Comparison
GeneTerm Linker vs *DAVID FAC*

	GeneTerm Linker	DAVID FAC (by default)	DAVID FAC (tuned to find 5 groups)
Total groups reference	5	5	5
Total groups found	5	15	5
Accuracy	**0.952**	**0.311**	**0.884**
Jaccard coefficient	**0.769**	**0.213**	**0.562**

Accuracy : percentage of genes grouped correctly.

$$Accuracy = \frac{(TP + TN)}{TP + TN + FP + FN}$$

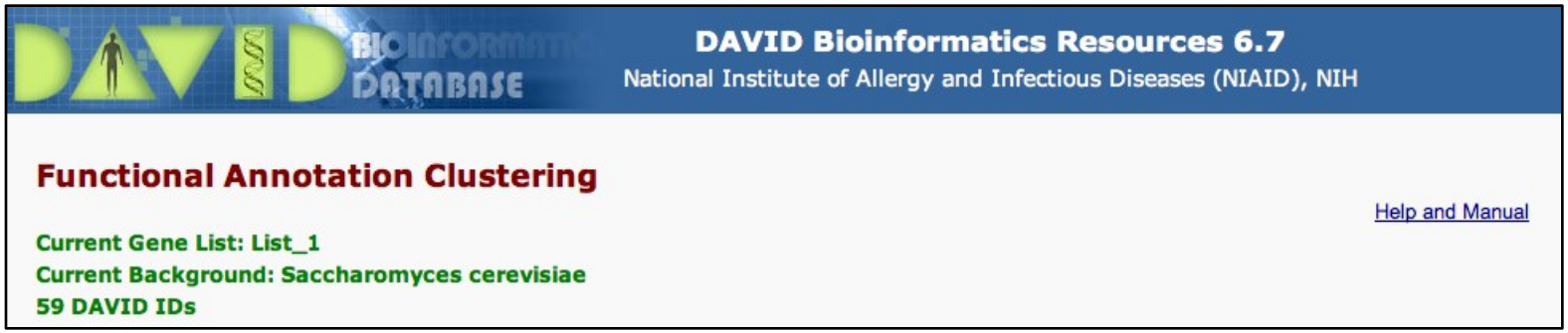
Jaccard coefficient : (measures similarity between sample sets) proportion of genes that belong to the same complex with respect to the total number of genes that are known to belong such complex or to any other.

$$Coeficiente\ de\ Jaccard = \frac{TP}{TP + FP + FN}$$

Comparison
GeneTerm Linker vs *DAVID FAC*

DAVID bioinformatics suite provides a **post-enrichment** tool:

DAVID Functional Annotation Clustering (DAVID - FAC), after single enrichment analysis it clusters the terms based in the genes that they share.



DAVID Functional Annotation Clustering (DAVID - FAC)

1. It groups/clusters **genes or terms** in an independent way
2. It does not considers the enrichment ***p-value*** of each gene-term set to rank them during the clustering
3. It does not eliminates **redundancies**

CBO Computational Biology @ UCT

CiC

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Hands-on: Practical Examples

Protein_ SETs_ 2014.xls

(106g hs, 175g hs, yeast 11pathways, yeast 59g5pc)

run DAVID-FAC & *GeneTerm Linker*

CBIO Computational Biology @ UCT

CIC

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Session 1 (9:30 - 12:30, 3h)
Bioinformatic tools for Functional Enrichment Analysis (FEA)
Session 2 (13:30 - 16:30, 3h)
Construction of gene functional networks

- Introduction to biological information and annotation spaces: GO, KEGG, Interpro

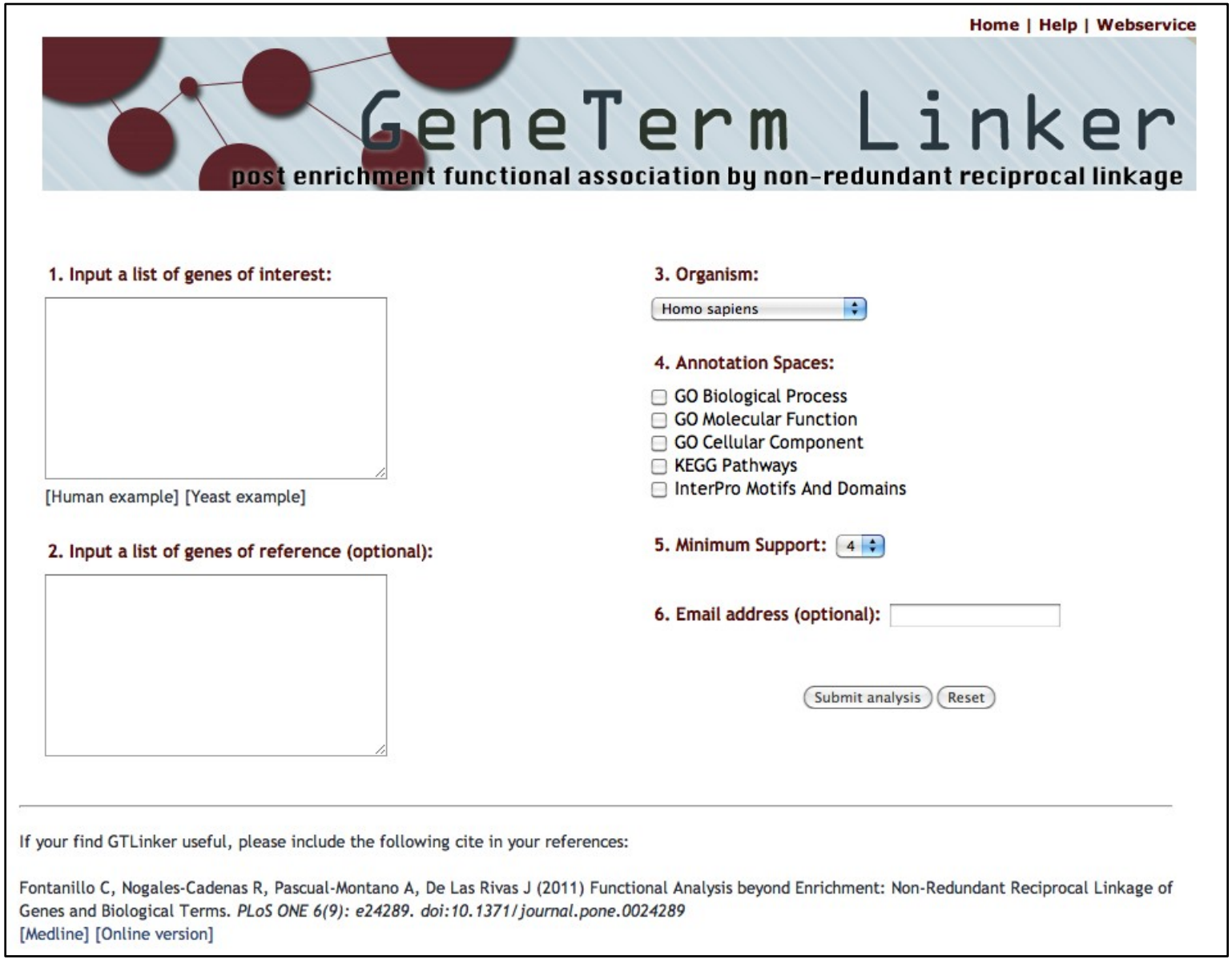
- Functional Enrichment Analysis (EA): from single to modular methods

- Using EA tools to annotate gene lists: DAVID (single), GSEA (gene sets), GeneCodis (modular)

- Sort out problems after EA: post-enrichment tool GeneTermLinker (postEA)

- From co-annotation and enrichment to functional networks: networks construction using a R tool

GeneTerm Linker
web tool



GeneTerm Linker results

FuncLonal analysis results for a **100 genes** signature of **Acute LinfoblasGc Leukemia (AML)**

	Genes	#list	#ref_list	adjusted pValue	Silhouette Width	Terms
Metagroup 3 Show details	BMP2 CTGF CTSH PF4 TESC	5(94)	104(34208)	1.04382e-05	0.8589	GO:0010628 positive regulation of gene expression (BP)
Metagroup 2 Show details	BMP2 CMTM8 CXCL2 CXCL3 GNAI1 PF4 PPBP TNFSF4	8(94)	310(34208)	2.35437e-06	0.7652	GO:0006935 chemotaxis (BP) GO:0008009 chemokine activity (MF) 04060 Cytokine-cytokine receptor interaction 04062 Chemokine signaling pathway IPR001811 Chemokine interleukin-8-like domain IPR002473 CXC chemokine, interleukin 8 IPR001089 CXC chemokine
Metagroup 4 Show details	GNAI1 LRRK2 RAB31 RAB32 RASD1	5(94)	165(34208)	9.53831e-05	0.7143	GO:0003924 GTPase activity (MF) IPR005225 Small GTP-binding protein domain IPR001806 Ras GTPase IPR003579 Ras small GTPase, Rab type
Metagroup 5 Show details	COL5A1 CTGF ECM1 LGALS3 PXDN	5(94)	272(34208)	0.000944947	0.7112	GO:0031012 extracellular matrix (CC) GO:0005578 proteinaceous extracellular matrix (CC)
Metagroup 1 Show details	LRRK2 MME PSD3 SHANK3 STXBP5 SYT1	6(94)	77(34208)	7.41975e-08	0.4167	GO:0008021 synaptic vesicle (CC) GO:0045202 synapse (CC)

GeneTerm Linker results

FuncLonal analysis results for a **100 genes** signature of **Acute LinfoblasGc Leukemia (AML)**

Genes	#list	#ref_list	adjusted pValue	Terms
CXCL2 CXCL3 PF4 PPBP	4(94)	9(34208)	8.59786e-07	GO:0006955 immune response (BP) GO:0006935 chemotaxis (BP) GO:0008009 chemokine activity (MF) GO:0005615 extracellular space (CC) GO:0005576 extracellular region (CC) 04060 Cytokine-cytokine receptor interaction 04062 Chemokine signaling pathway IPR001811 Chemokine interleukin-8-like domain IPR002473 CXC chemokine, interleukin 8 IPR001089 CXC chemokine
CMTM8 CXCL2 CXCL3 PF4 PPBP	5(94)	65(34208)	1.8799e-05	GO:0006935 chemotaxis (BP) GO:0005615 extracellular space (CC)
CXCL2 CXCL3 PF4 PPBP TNFSF4	5(94)	76(34208)	3.19339e-05	GO:0006955 immune response (BP) GO:0005615 extracellular space (CC) 04060 Cytokine-cytokine receptor interaction
BMP2 CXCL2 CXCL3 PF4 PPBP TNFSF4	6(94)	146(34208)	4.20382e-05	GO:0005615 extracellular space (CC) 04060 Cytokine-cytokine receptor interaction
CXCL2 CXCL3 GNAI1 PF4 PPBP	5(94)	186(34208)	0.000567261	04062 Chemokine signaling pathway

GeneTerm Linker results

Functional analysis results for a **100 genes** signature of **Acute Linfoblastic Leukemia (AML)**

Statistics of the analysis

Total number of **genes** sent to the analysis: 100
Number of **genes** recognized by the enrichment tool: 94
Total number of **genes** not-included in the enrichment analysis: 13 (show/hide)
Genes not present in the metagroups, only annotated to generic terms: 56 (show/hide)
Total number of **genes** included in the metagroups: 31 (show/hide)

Total number of significant **GeneTerm-sets** provided by the enrichment analysis: 98
Download the initial Enrichment Analysis (text format)
Number of filtered **GeneTerm-sets** that include generic terms in the enrichment: 80
Download the GeneTerm-sets filtered as generic (text format)
Total number of **GeneTerm-sets** of the enrichment used: 18
NOTE: Maximum number of **GeneTerm-sets** analysed in the web = 1000
Number of redundant **GeneTerm-sets** of the enrichment filtered: 2
Final number of **GeneTerm-sets** of the enrichment included in the metagroups: 16

Functional Network
derived from *GeneTerm Linker*

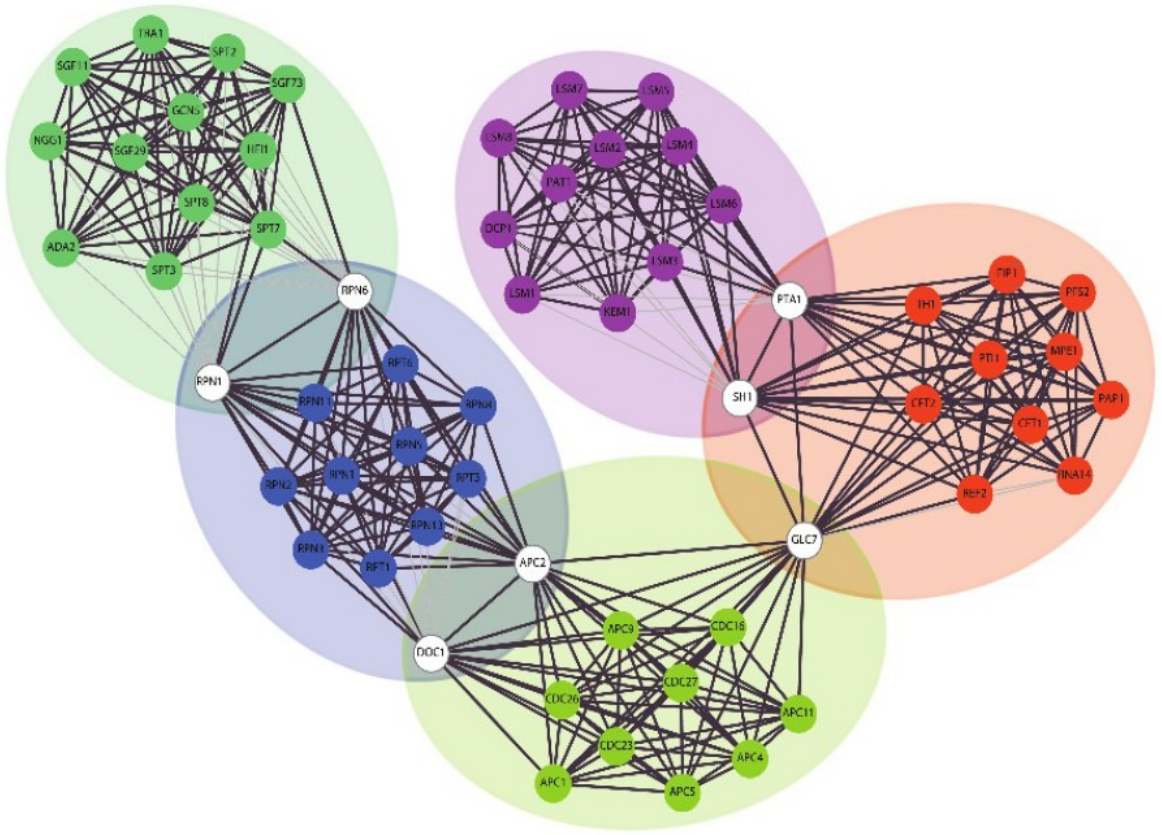
The results of the gene---term **metagroups** generated by *GeneTerm Linker* can be used in a further analysis to build **funcGonal networks**

	gts 1	gts 2	gts 3	gts 4
g1	0	0	1	1
g2	0	0	1	0
g3	1	1	0	1
g4	1	0	1	0

Common **gene term sets** adjacency matrix (g :: gts)

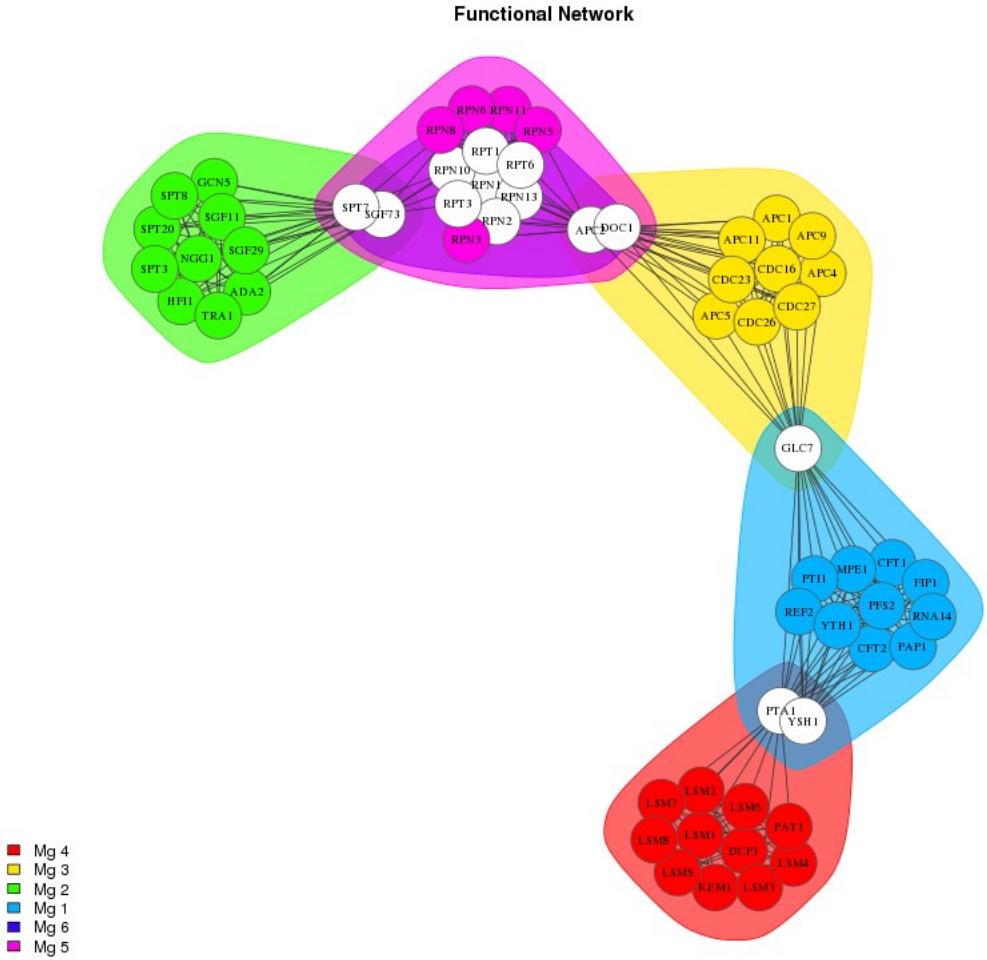
	Mg 1	Mg 2	Mg 3	Mg 4
g1	0	1	0	1
g2	1	0	1	0
g3	0	1	0	1
g4	1	0	1	0

Common **metagroups** adjacency matrix (g :: Mg)



Functional Network
derived from *GeneTerm Linker*

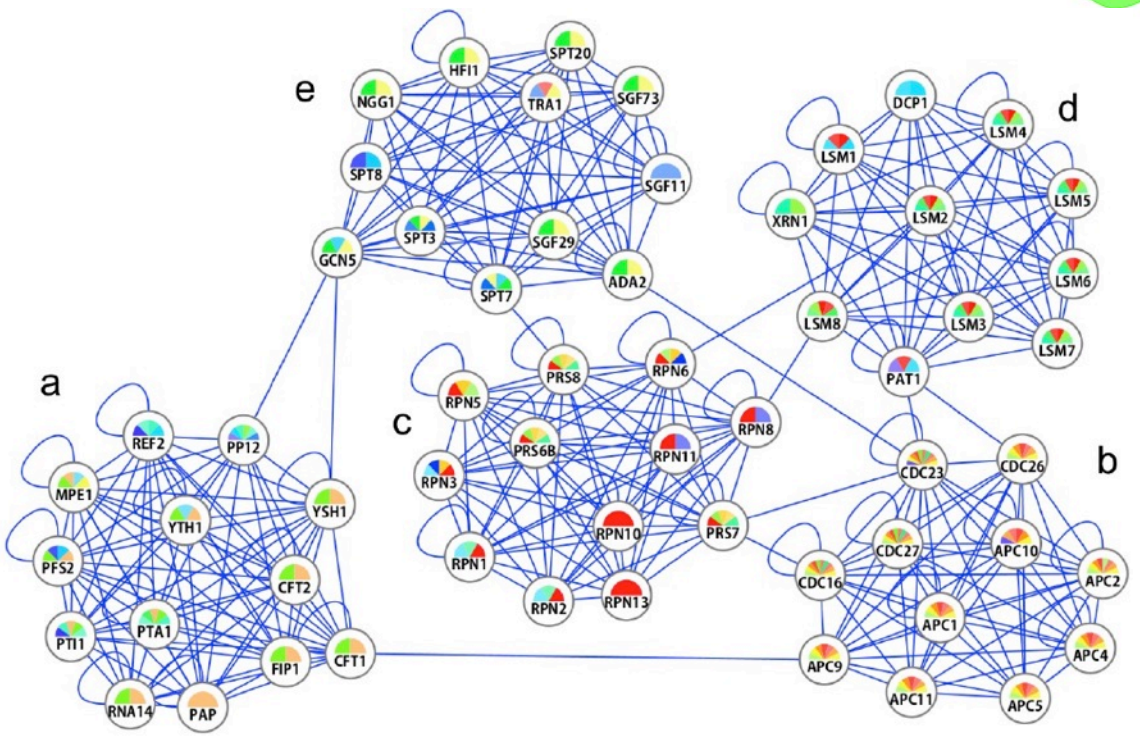
The results of the gene---term **metagroups** generated by *GeneTerm Linker* can be used in a further analysis to build **funcGonal networks**

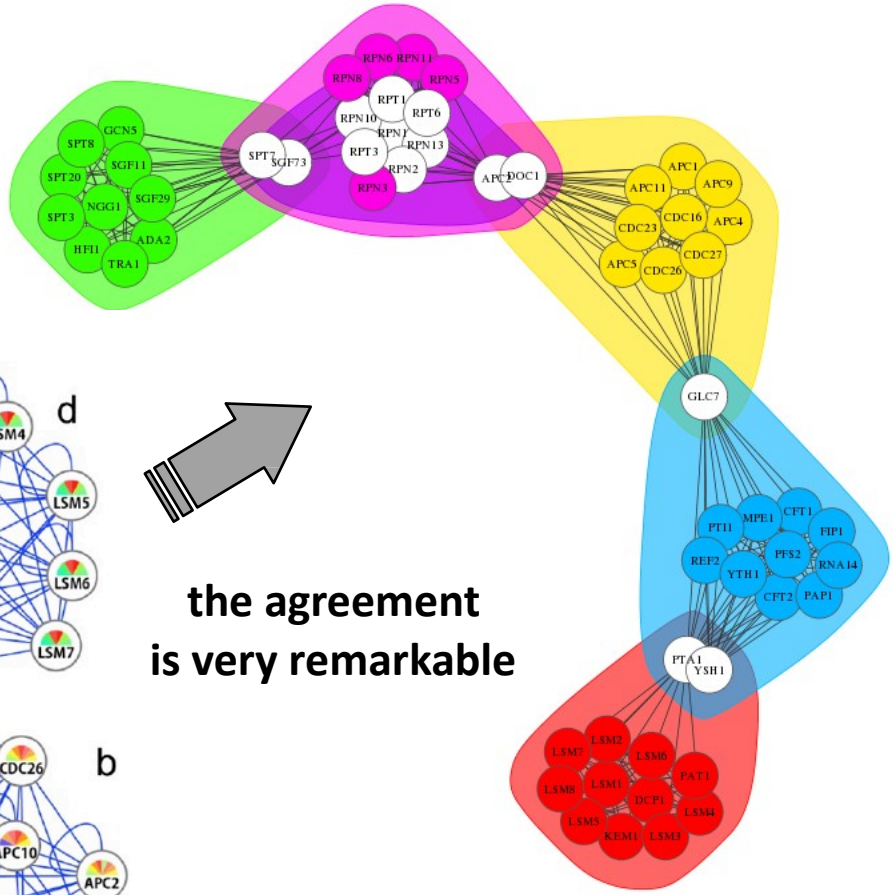


Functional Network
derived from *GeneTerm Linker*

Build ***Func&onalNetworks*** :: from an experiment---based **protein interacGon network** to a knowledge---based **protein/gene funcGonal network**

1 = d = U6 snRNP complex (LSM...)
2 = b = Anaphase-promoting complex (APC...)
3 = e = SAGA complex (SPT...)
4 = a = mRNA cleavage & polyA spec. fact. complex
5 = c = 19/22S regulator, proteasome (RPN...)





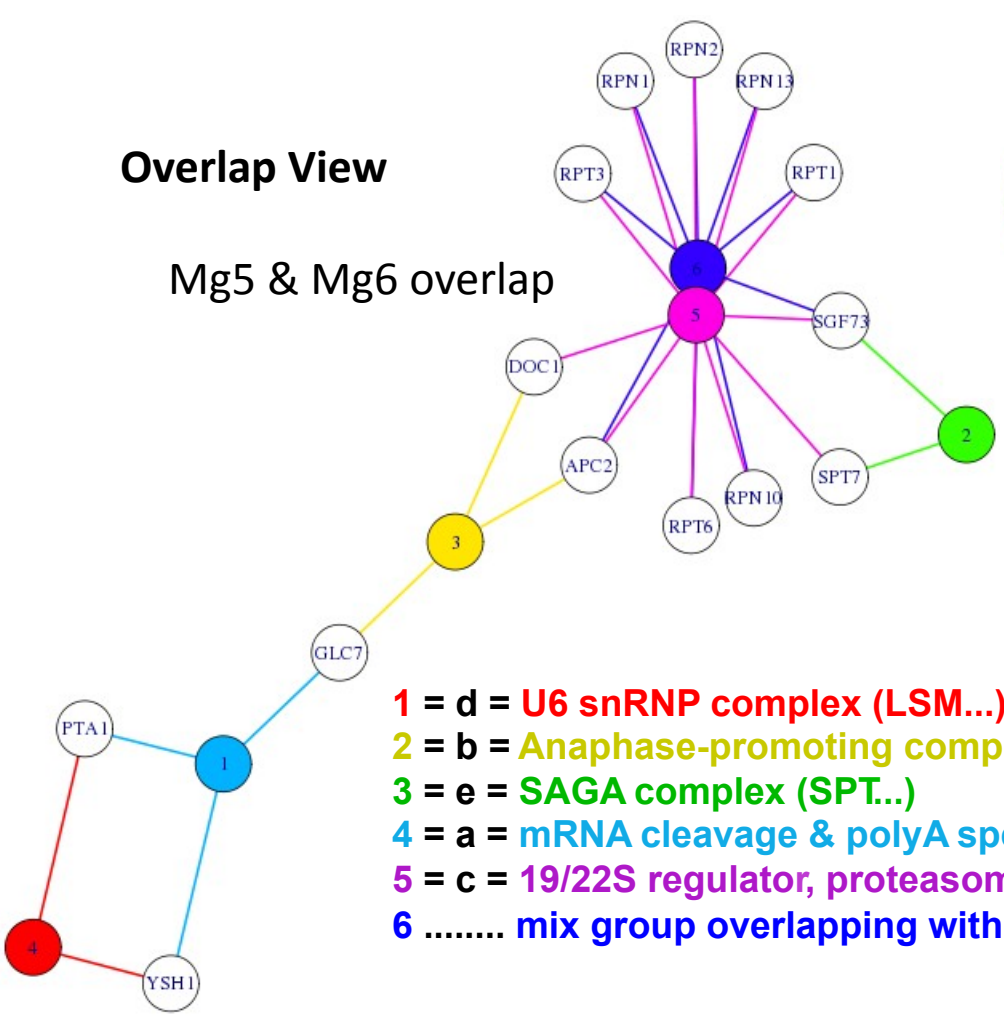
the agreement is very remarkable

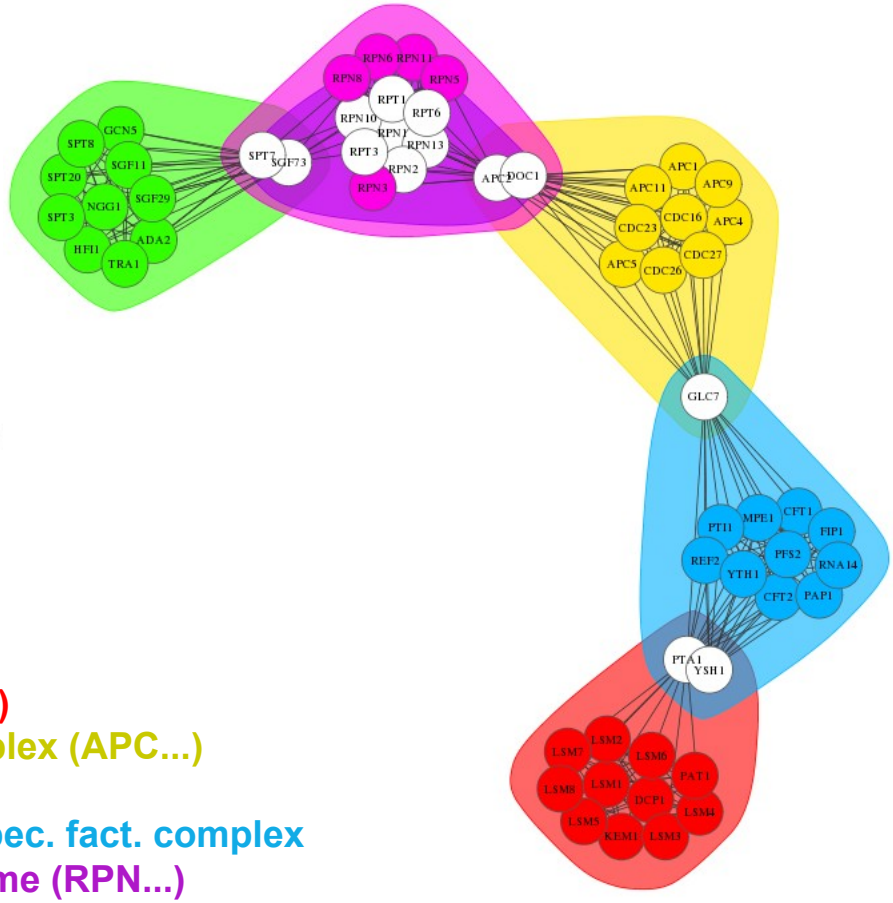
Functional Network
derived from *GeneTerm Linker*

Build ***Func&onalNetworks*** :: from an experiment---based **protein interacGon network** to a knowledge---based **protein/gene funcGonal network**

Overlap View

Mg5 & Mg6 overlap

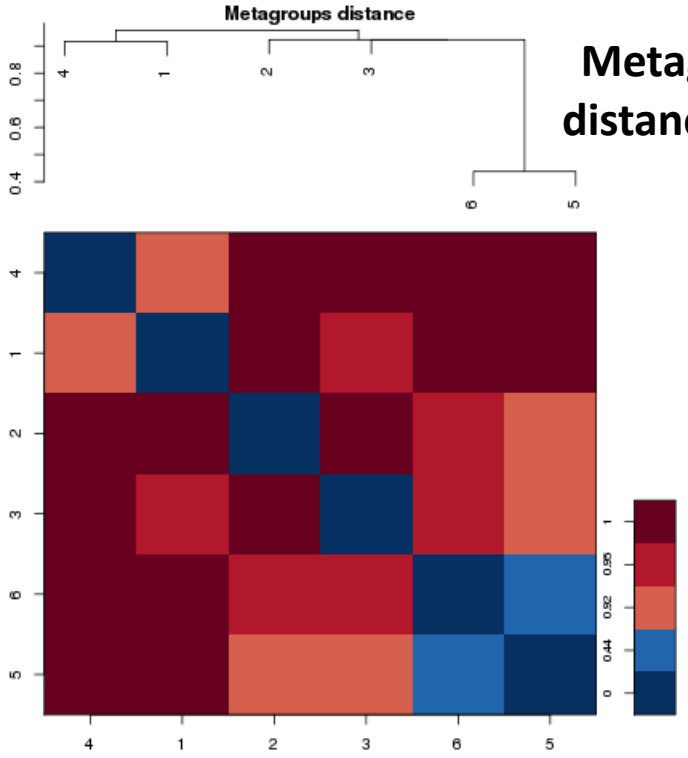




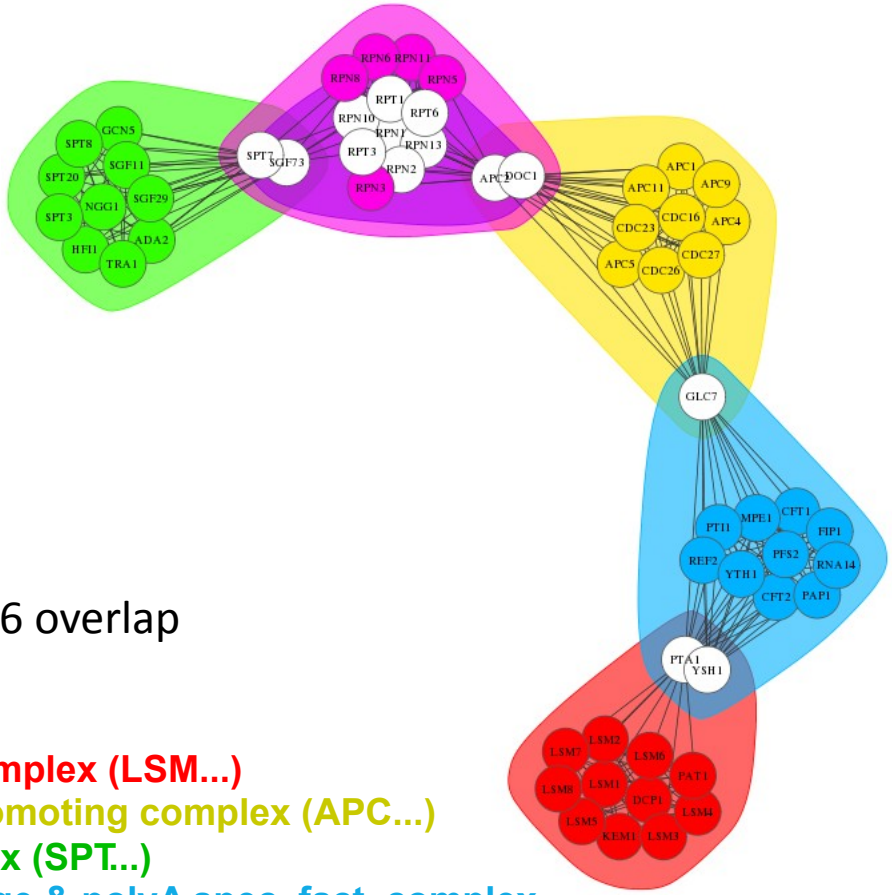
1 = d = U6 snRNP complex (LSM...)
2 = b = Anaphase-promoting complex (APC...)
3 = e = SAGA complex (SPT...)
4 = a = mRNA cleavage & polyA spec. fact. complex
5 = c = 19/22S regulator, proteasome (RPN...)
6 ..... mix group overlapping with 3 other: 5, 3, 2

Functional Network
derived from *GeneTerm Linker*

Build ***Func&onalNetworks*** :: from an experiment---based **protein interacGon network** to a knowledge---based **protein/gene funcGonal network**



Metagroups distances plot



Mg5 & Mg6 overlap

1 = d = U6 snRNP complex (LSM...)
2 = b = Anaphase-promoting complex (APC...)
3 = e = SAGA complex (SPT...)
4 = a = mRNA cleavage & polyA spec. fact. complex
5 = c = 19/22S regulator, proteasome (RPN...)
6 ..... mix group overlapping with 3 other: 5, 3, 2