V9 funktionelle Annotation

- **Analyse von Gen-Expression**
- **Funktionelle Annotation: Gene Ontology (GO)**
- **Signifikanz der Annotation: Hypergeometrischer Test**
- **Annotationsanalysen z.B. mit NIH-Tool DAVID**
- **Ähnlichkeit von GO-Termen automatisch bestimmen**
- **OMIM-Datenbank**

Ausgangslage

Daten aus Microarray-Analyse wurden ursprünglich als sehr „verrauscht“ angesehen.

Mittlerweile wurden jedoch sowohl die experimentellen Schritte wie auch die Datenauswertung gründlich verfeinert.

Microarray-Analyse ist daher heute eine (zwar teure, aber zuverlässige) Routine-Methode, die in allen großen Firmen verwendet wird.

Heute wird die MA-Analyse zunehmend durch RNA-seq ersetzt.

Die Datenaufbereitung kann in beiden Fällen folgende Schritte enthalten: Normalisierung, Logarithmierung, Clustering, evtl. Ko-Expressionsanalyse, **Annotation der Genfunktion (Inhalt von V9).**

Sehr wichtig ist es immer, die Signifikanz der Ergebnisse zu bewerten.

Gentleman et al. Genome Biology 5, R80 (2004)

Beispiel: differentielle Gen-Expression für ALL-Patienten

Input:

Genexpressionsdaten für 128 Patienten mit akuter lymphatischer Leukämie (ALL).

Alle ALL-Patienten haben chromosomale Veränderungen.

Der Therapieerfolg ist jedoch sehr unterschiedlich.

Hintergrundinformation:

- Eine Gruppe von Patienten (ALL1/AF4) hat eine genetische Translokation zwischen den Chromosomen 4 und 11.
- Eine zweite Gruppe von Patienten (BCR/ABL) hat eine genetische Translokation zwischen den Chromosomen 9 und 22.
- Die Krankheitsursachen + optimale Therapie können für die beiden Gruppen verschieden sein.

Ziel:

Identifiziere Gene, die zwischen den beiden Gruppen differentiell exprimiert werden.

Beispiel für die Anwendung der `Bioconductor`-Software (siehe Ref unten, bisher rund 10000 mal zitiert).

Gentleman et al. Genome Biology 5, R80 (2004)

Auswahl der differentiell exprimierten Gene

```
> f <- factor(as.character(eset$mol))
> design <- model.matrix(~f)
> fit <- lmFit(eset, design)
> fit <- eBayes(fit)
> topTable(fit, coef = 2)
```

Bioconductor Kommandos

	ID	M	A	t	*p*-value	B
1016	1914_at	−3.1	4.6	−27	5.9e-27	56
7884	37809_at	−4.0	4.9	−20	1.3e-20	44
6939	36873_at	−3.4	4.3	−20	1.8e-20	44
10865	40763_at	−3.1	3.5	−17	7.2e-18	39
4250	34210_at	3.6	8.4	15	3.5e-16	35
11556	41448_at	−2.5	3.7	−15	1.8e-15	34
3389	33358_at	−2.3	5.2	−13	3.3e-13	29
8054	37978_at	−1.0	6.9	−10	6.5e-10	22
10579	40480_s_at	1.8	7.8	10	9.1e-10	21
330	1307_at	1.6	4.6	10	1.4e-09	21

Figure 1
Limma analysis of the ALL data. The leftmost numbers are row indices, ID is the Affymetrix HGU95av2 accession number, M is the log ratio of expression, A is the log average expression, and B is the log odds of differential expression.

Differential expression (D.E.) = log(R) / log(G)

Log ratio M : 2^M = log(R) / log(G); M = 1 -> zweifach D.E.

Wie signifikant ist dies? -> bewerte mit statistischem Test.

Vergleiche Gen-Expression in den beiden Gruppen.

Fokussiere auf Gene mit stark unterschiedlicher Expression.

Wähle z.B. alle Gene mit p-Wert < 0.05 aus.

Es bleiben 165 Gene übrig.

Gentleman et al. Genome Biology 5, R80 (2004)

Differentielle Gen-Expression als Heatmap visualisieren

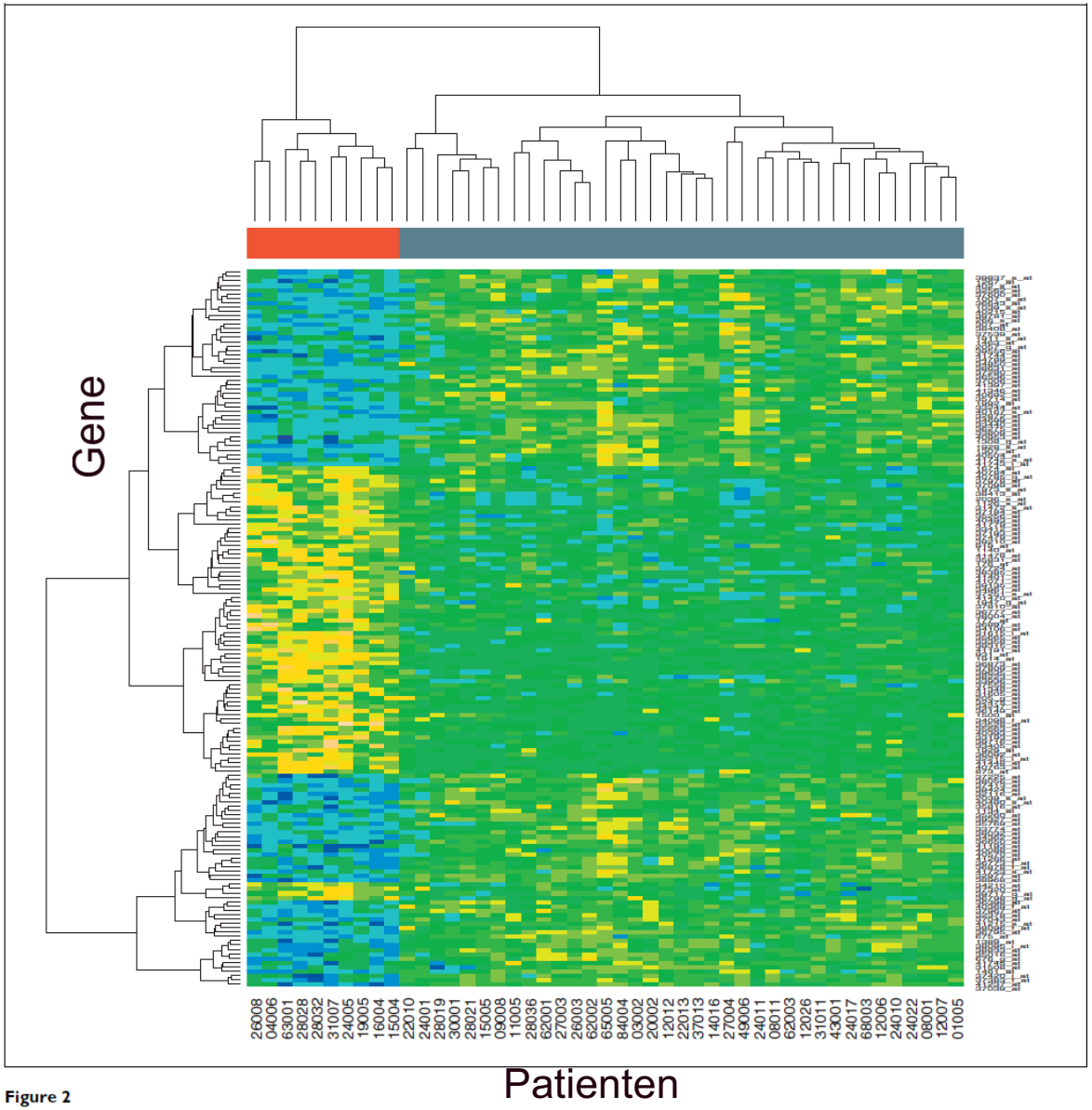


Figure 2
Heat map (produced by the Bioconductor function heatmap()) of the ALL leukemia data.

Mit einem Abstandsmaß und einem Cluster-Algorithmus werden die Ähnlichkeiten zwischen den Patienten (x-Achse) und den einzelnen Gene (y-Achse) erfasst.

Die beiden Patienten-Gruppen haben deutlich unterschiedliche Expressionsprofile (rot/grau).

Gelb: stark hochreguliert

Blau: stark runterreguliert

Gentleman et al. Genome Biology 5, R80 (2004)

Zuordnung von Gen-Funktion

```
> f <- factor(as.character(eset$mol))
> design <- model.matrix(~f)
> fit <- lmFit(eset, design)
> fit <- eBayes(fit)
> topTable(fit, coef = 2)
```

Bioconductor Kommandos

	ID	M	A	t	*p*-value	B
1016	1914_at	–3.1	4.6	–27	5.9e-27	56
7884	37809_at	–4.0	4.9	–20	1.3e-20	44
6939	36873_at	–3.4	4.3	–20	1.8e-20	44
10865	40763_at	–3.1	3.5	–17	7.2e-18	39
4250	34210_at	3.6	8.4	15	3.5e-16	35
11556	41448_at	–2.5	3.7	–15	1.8e-15	34
3389	33358_at	–2.3	5.2	–13	3.3e-13	29
8054	37978_at	–1.0	6.9	–10	6.5e-10	22
10579	40480_s_at	1.8	7.8	10	9.1e-10	21
330	1307_at	1.6	4.6	10	1.4e-09	21

Figure 1
Limma analysis of the ALL data. The leftmost numbers are row indices, ID is the Affymetrix HGU95av2 accession number, M is the log ratio of expression, A is the log average expression, and B is the log odds of differential expression.

Links gezeigt ist dieselbe Tabelle wie zwei Folien zuvor.

Nun interessiert uns, welche Funktionen diese Gene in der Zelle ausüben.

Verwende dazu Informationen aus der `Gene Ontology` über diese Gene.

Gentleman et al. Genome Biology 5, R80 (2004)

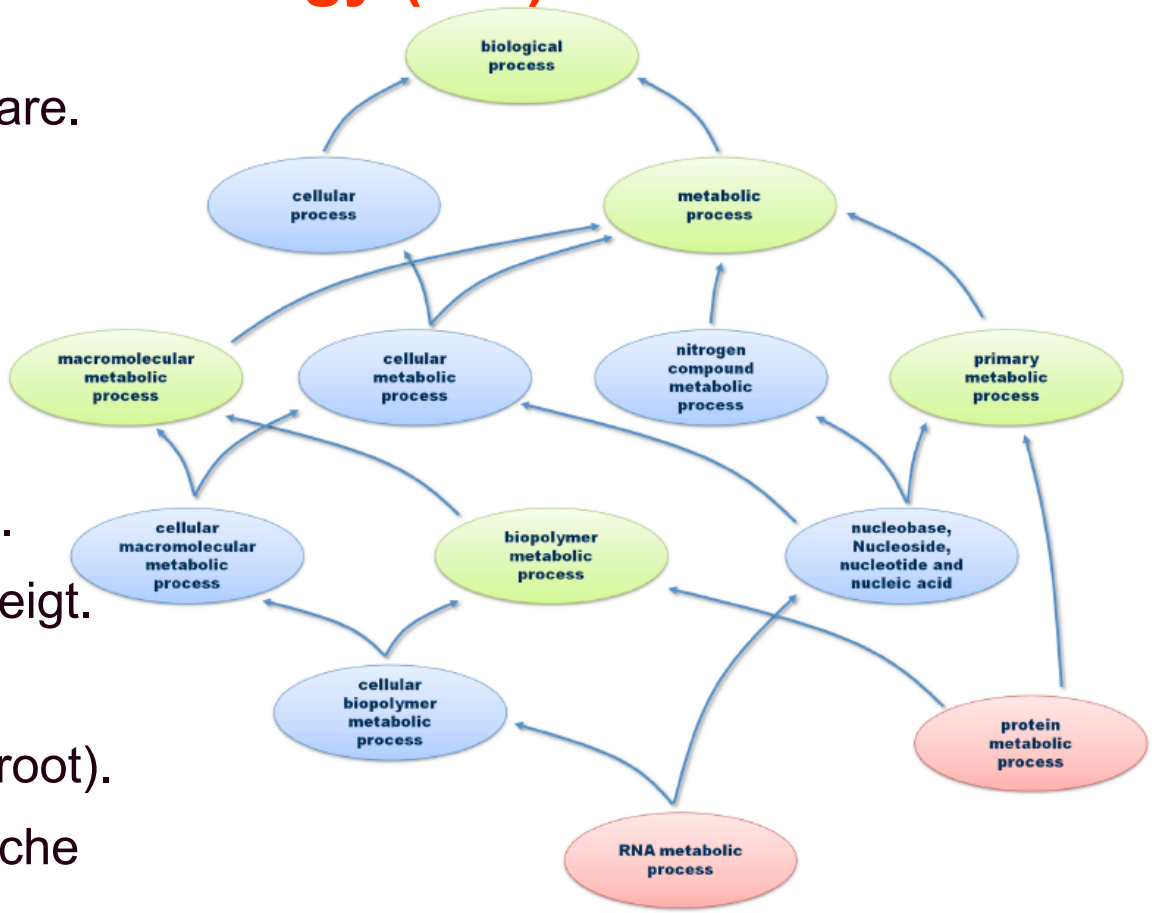
Die Gene Ontology (GO)

Ontologien sind strukturierte Vokabulare.

Die Gene Ontology hat 3 Bereiche:

- biologischer Prozess (BP)
- molekulare Funktion (MF)
- zelluläre Komponente (Lokalisation).

Hier ist ein Teil der BP-Ontologie gezeigt.



Oben ist der allgemeinste Ausdruck (root).

Rot: Blätter des Baums (sehr spezifische GO-Terme)

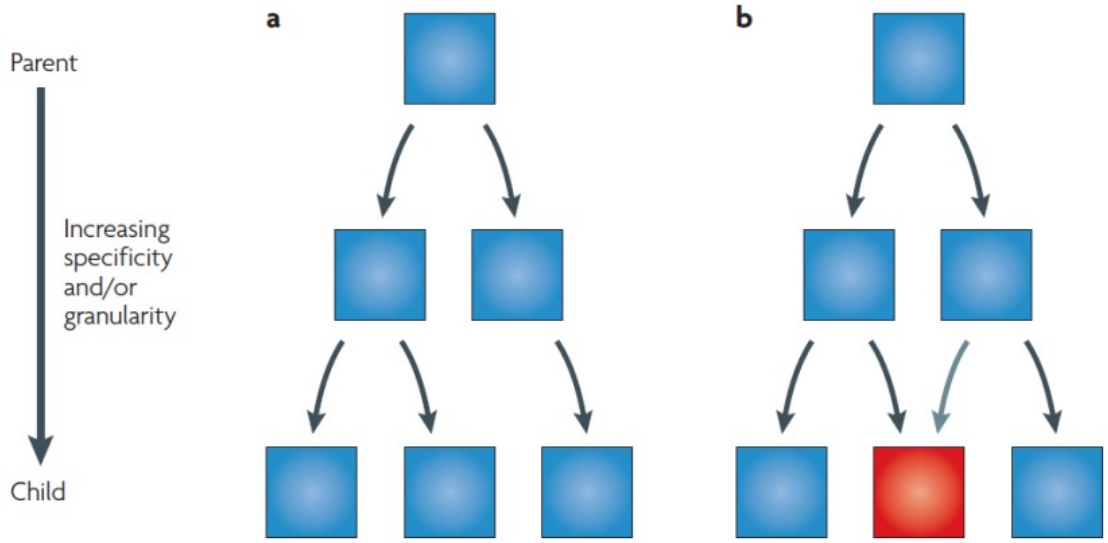
Grün: gemeinsame Vorgänger.

Blau: andere Knoten.

Linien: „Y ist in X enthalten“-Beziehungen

Dissertation Andreas Schlicker (UdS, 2010)

Baum vs. azyklische Graphen

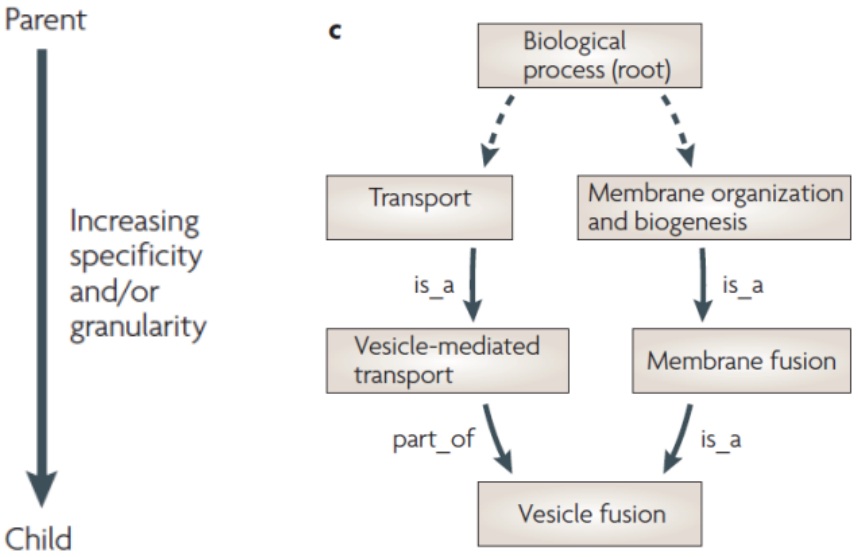


a | Einfacher **Baum**, in dem jedes Kind genau ein Elternteil hat.
Die Kanten sind vom Elternteil zum Kind gerichtet.

b | In einem **gerichteten azyklischen Graph** (DAG) kann jedes Kind ein oder mehrere Eltern haben. Hier besitzt z.B. der rote Knoten 2 Eltern.
Azklisch heisst, dass der Graph keine gerichteten Zyklen enthält.

Rhee et al. (2008) Nature
Rev. Genet. 9: *509*

Die Gene Ontology ist ein directed acyclic graph



Der Knoten *vesicle fusion*
besitzt in der BP Ontologie mehrere Eltern.

Gestrichelte Kanten: andere dazwischen liegende Knoten sind nicht gezeigt.

Root : keine Kanten zeigen zu diesem Knoten hin. Er hat mindestens ein Kind.

Leaf node : ein Endknoten ohne Kinder.

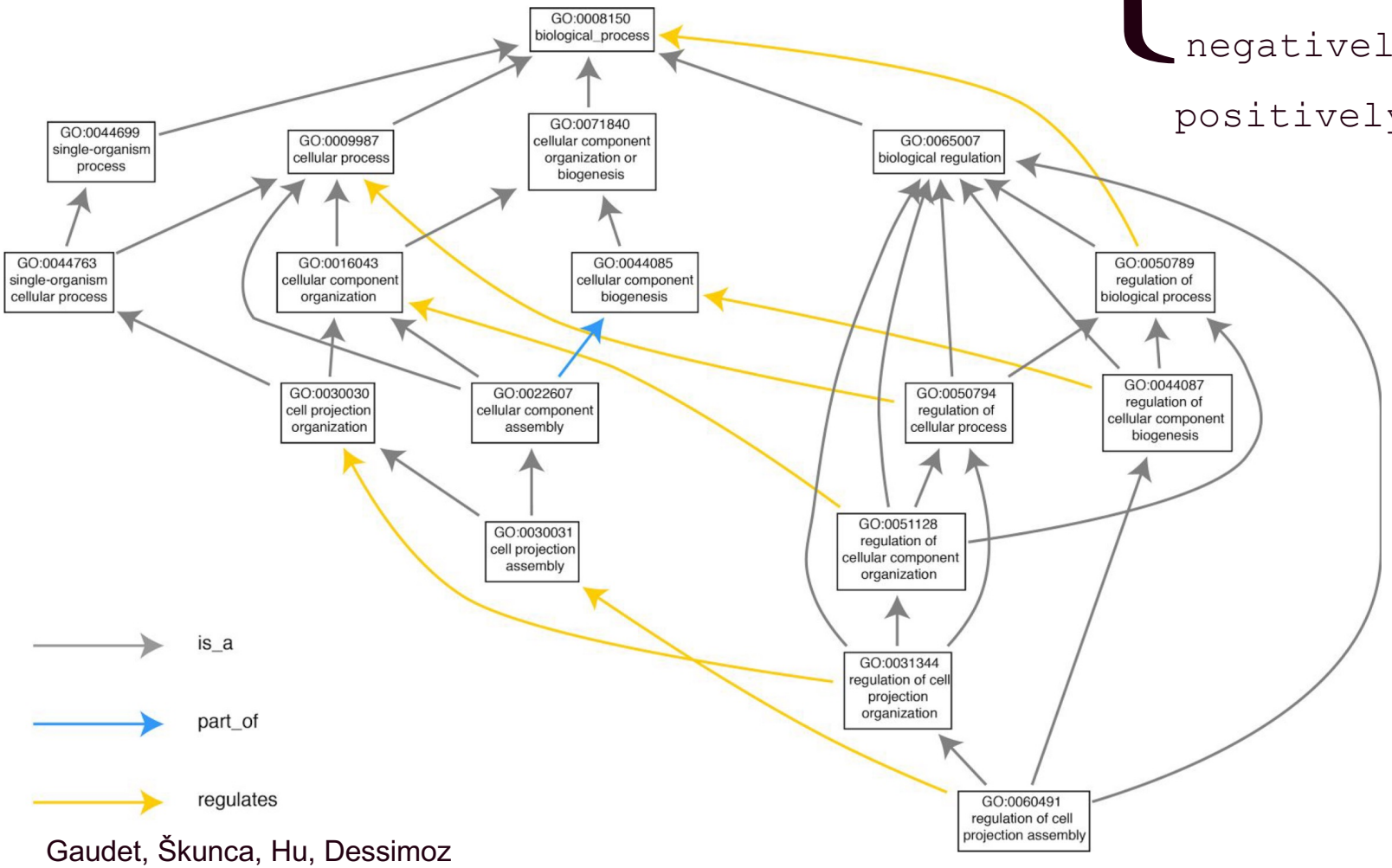
Tiefe eines Knotens: Länge des längsten Pfades von root zu diesem Knoten.

Höhe eines Knotens: Länge des längsten Pfades von diesem Knoten zu einem Endknoten.

Rhee et al. (2008) Nature
Rev. Genet. 9: *509*

Beziehungen in der GO

Gen X { is_a, is a part_of, regulates, negatively_regulates, positively_regulates } Beziehung



Gaudet, Škunca, Hu, Dessimoz
Primer on the Gene Ontology,
https://arxiv.org/abs/1602.01876

Gene Ontology (GO) - Konsortium

Berkeley Bioinformatics Open-source Project (BBOP)

British Heart Foundation - University College London (BHF-UCL)

dictyBase

EcoliWiki

FlyBase

GeneDB

UniProtKB-Gene Ontology Annotation @ EBI (UniProtKB-GOA)

GO Editorial Office at the European Bioinformatics Institute

Gramene

Institute of Genome Sciences, Univ. of Maryland

J Craig Venter Institute

Mouse Genome Informatics (MGI)

Rat Genome Database (RGD)

Reactome

Saccharomyces Genome Database (SGD)

The Arabidopsis Information Resource (TAIR)

WormBase

The Zebrafish Information Network (ZFIN)

Woher stammen die Gene Ontology Annotationen?

Evidence code	Evidence code description	Source of evidence	Manually checked	Current number of annotations*
IDA	Inferred from direct assay	Experimental	Yes	71,050
IEP	Inferred from expression pattern	Experimental	Yes	4,598
IGI	Inferred from genetic interaction	Experimental	Yes	8,311
IMP	Inferred from mutant phenotype	Experimental	Yes	61,549
IPI	Inferred from physical interaction	Experimental	Yes	17,043
ISS	Inferred from sequence or structural similarity	Computational	Yes	196,643
RCA	Inferred from reviewed computational analysis	Computational	Yes	103,792
IGC	Inferred from genomic context	Computational	Yes	4
IEA	Inferred from electronic annotation	Computational	No	15,687,382
IC	Inferred by curator	Indirectly derived from experimental or computational evidence made by a curator	Yes	5,167
TAS	Traceable author statement	Indirectly derived from experimental or computational evidence made by the author of the published article	Yes	44,564
NAS	Non-traceable author statement	No 'source of evidence' statement given	Yes	25,656
ND	No biological data available	No information available	Yes	132,192
NR	Not recorded	Unknown	Yes	1,185

*October 2007 release

Rhee et al. Nature Reviews Genetics 9, 509-515 (2008)

Woher stammen die Gene Ontology Annotationen?

Species (NCBI taxon ID)	Genes* with experimental annotations‡	Total annotated genes*	Percentage of genes* with at least one experimental annotation	Total genes*	Percentage annotated§	Percentage known in genome‖
Schizosaccharomyces pombe (4896)	4,482	4,930	90.9%	4,930	100%	90.9%
Saccharomyces cerevisiae (4932)	4,947	5,794	85.4%	5,794	100%	85.4%
Mouse (10090)	10,621	18,386	57.8%	27,289	67.4%	38.9%
Caenorhabditis elegans (6239)	4,614	14,154	32.6%	20,163	70.2%	22.9%
Human¶ (9606)	4,780	17,021	28.1%	20,887	81.5%	22.9%
Arabidopsis thaliana# (3702)	5,530	26,637	20.8%	27,029	98.5%	20.5%
Rat (10116)	3,566	17,243	20.7%	17,993	95.8%	19.8%
Fruitfly (7227)**	2,790	9,563	29.2%	14,141	67.6%	19.7%
Candida albicans (5476)	806	3,756	21.4%	6,166	60.9%	13.0%
Pseudomonas aeruginosa PAO1 (208964)	491	2,506	19.6%	5,568	45.0%	8.82%
Slime mold (44689)	797	6,892	11.6%	13,625	50.6%	5.9%
Trypanosoma brucei (5691)	449	3,914	11.5%	9,154	42.8%	4.92%
Zebrafish (7955)	1,235	13,574	5.8%	21,322	63.7%	3.7%
Plasmodium falciparum (5833)	188	3,243	5.8%	5,420	59.8%	3.47%
Rice (39947)	654	29,877	2.2%	41,908	71.3%	1.57%
Chicken¶ (9031)	75	6,063	1.2%	16,737	36.2%	0.4%
Cow¶ (9913)	96	8,536	1.1%	21,756	39.2%	0.4%

*Total genes in genomes include only those that encode proteins. These numbers were obtained from the databases that contribute annotations to GO and are listed on the GO annotations download page (http://www.geneontology.org/GO.current.annotations.shtml). ‡Experimental annotations include those only with the following evidence codes: IDA (inferred from direct assay), IEP (inferred from expression pattern), IGI (inferred from genetic interaction), IMP (inferred from mutant phenotype) and IPI (inferred from physical interaction). §Percentage annotated is determined by dividing the number of genes annotated by total genes. ‖Percentage known in genome is determined by multiplying the percentage of experimentally derived annotations by the percentage of the genome annotated. This is an approximation of the extent of knowledge about the portion of the genome that encodes proteins in an organism with a complete genome sequence that is captured by annotation. ¶Numbers are from the GO annotation project at the European Bioinformatics Institute, human data last updated 14 September 2007, cow data last updated 17 January 2007, chicken data last updated 10 July 2007. #Numbers are from The *Arabidopsis* Information Resource (TAIR), last updated 14 December 2007. **Numbers are based on release 5.4 of the *Drosophila melanogaster* genome and GO annotations from FlyBase release FB2007_03 (dated 11 January 2007). NCBI, National Center for Biotechnology Information.

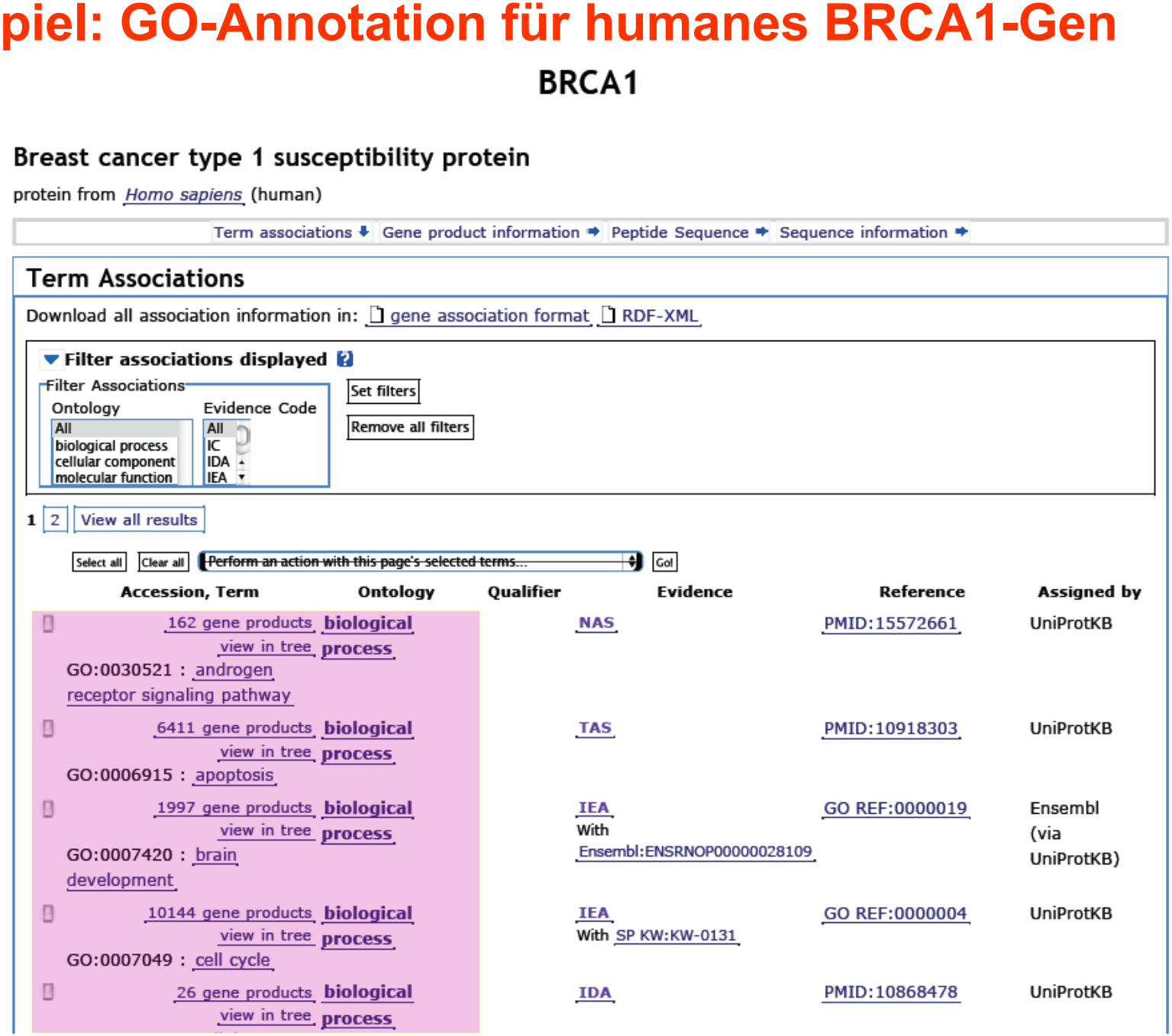
Rhee et al. Nature Reviews Genetics 9, 509-515 (2008)

Format des GO flat files

Column	Content	Required?	Cardinality	Example
1	DB	required	1	UniProtKB
2	DB Object ID	required	1	P12345
3	DB Object Symbol	required	1	PHO3
4	Qualifier	optional	0 or greater	NOT
5	GO ID	required	1	GO:0003993
6	DB:Reference (|DB:Reference)	required	1 or greater	PMID:2676709
7	Evidence Code	required	1	IMP
8	With (or) From	optional	0 or greater	GO:0000346
9	Aspect	required	1	F
10	DB Object Name	optional	0 or 1	Toll-like receptor 4
11	DB Object Synonym (|Synonym)	optional	0 or greater	hToll|Tollbooth
12	DB Object Type	required	1	protein
13	Taxon(|taxon)	required	1 or 2	taxon:9606
14	Date	required	1	20090118
15	Assigned By	required	1	SGD
16	Annotation Extension	optional	0 or greater	part_of(CL:0000576)
17	Gene Product Form ID	optional	0 or 1	UniProtKB:P12345-2

Beispiel: GO-Annotation für humanes BRCA1-Gen

Einzelne GO-Terme, mit denen das Brustkrebs-Gen BRCA1 annotiert ist.



BRCA1

Breast cancer type 1 susceptibility protein

protein from *Homo sapiens* (human)

Term associations | Gene product information | Peptide Sequence | Sequence information

Term Associations

Download all association information in: gene association format | RDF-XML

Filter associations displayed

Filter Associations

Ontology: All; biological process; cellular component; molecular function

Evidence Code: All; IC; IDA; IEA

Set filters | Remove all filters

1 2 View all results

Select all | Clear all | Perform an action with this page's selected terms... | Go!

Accession, Term	Ontology	Qualifier	Evidence	Reference	Assigned by
162 gene products, view in tree; GO:0030521 : androgen receptor signaling pathway	biological process		NAS	PMID:15572661	UniProtKB
6411 gene products, view in tree; GO:0006915 : apoptosis	biological process		TAS	PMID:10918303	UniProtKB
1997 gene products, view in tree; GO:0007420 : brain development	biological process		IEA With Ensembl:ENSRNOP00000028109	GO REF:0000019	Ensembl (via UniProtKB)
10144 gene products, view in tree; GO:0007049 : cell cycle	biological process		IEA With SP KW:KW-0131	GO REF:0000004	UniProtKB
26 gene products, view in tree	biological process		IDA	PMID:10868478	UniProtKB

Signifkanz von GO-Annotationen

Sehr **allgemeine GO-Terme** wie z.B. “cellular metabolic process“ werden vielen Genen im Genom zugeordnet.

Sehr **spezielle Terme** gehören jeweils nur zu wenigen Genen.

Man muss also vergleichen, wie **signifikant** das Auftreten jedes GO-Terms in einer Testmenge an Genen im Vergleich zu einer zufällig ausgewählten Menge an Genen derselben Größe ist.

Dazu verwendet man meist den **hypergeometrischen Test**.

Dissertation Andreas Schlicker (UdS, 2010)

Vorbemerkung

Zieht man aus einer Urne mit *n* Kugeln insgesamt *k* Kugeln **ohne Beachtung der Reihenfolge**, so gibt es hierfür genau

$$\frac{n\cdot(n-1)\cdot(n-2)\cdot\ldots\cdot(n-k+1)}{k!}=\frac{n!}{k!\cdot(n-k)!}=\binom{n}{k}$$ Möglichkeiten

http://www.schule-bw.de/

Hypergeometrischer Test

$$\text{p-Wert} = \sum_{i=k_\pi}^{min(n,K_\pi)} \frac{\binom{K_\pi}{i}\binom{N-K_\pi}{n-i}}{\binom{N}{n}}$$

Der hypergeometrische Test ist ein statistischer Test, der z.B. überprüft, ob in einer vorgegebenen Testmenge an Genen eine biologische Annotation π gegenüber dem gesamten Genom statistisch signifikant angereichert ist.

- Sei *N* die Anzahl an Genen im Genom.
- Sei *n* die Anzahl an Genen in der Testmenge.
- Sie K_π die Anzahl an Genen im Genom mit der Annotation π.
- Sei k_π die Anzahl an Genen in der Testmenge mit der Annotation π.

Der hypergeometrische p-Wert drückt die Wahrscheinlichkeit aus, dass k_π oder mehr **zufällig** aus dem Genom ausgewählte Gene auch die Annotation π haben.

http://great.stanford.edu/

Hypergeometrischer Test

Wähle $i = k_\pi$ Gene mit Annotation π aus dem Genom.

Davon gibt es genau K_π.

Die anderen $n - i$ Gene in der Testmenge haben dann nicht die Annotation π. Davon gibt es im Genom genau N - K_π.

$$\text{p-Wert} = \sum_{i=k_\pi}^{min(n,K_\pi)} \frac{\binom{K_\pi}{i}\binom{N-K_\pi}{n-i}}{\binom{N}{n}}$$

Korrigiert für die kombinatorische Vielfalt an Möglichkeiten um n Elemente aus einer Menge mit N Elementen auszuwählen.

N.B. dies gilt für den Fall, dass die Reihenfolge der Elemente egal ist.

Die Summe läuft von mindestens k_π Elementen bis zur maximal möglichen Anzahl an Elementen.

Eine Obergrenze ist durch die Anzahl an Genen mit Annotation π im Genom gegeben (K_π).

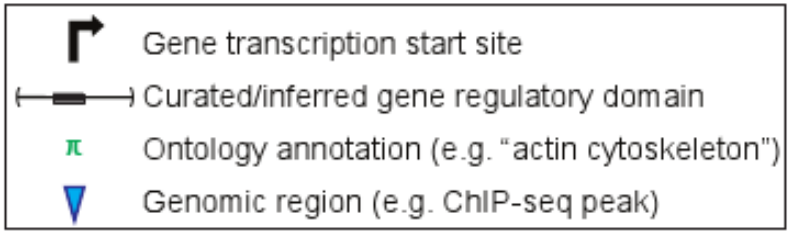
Die andere Obergrenze ist die Zahl der Gene in der Testmenge (n).

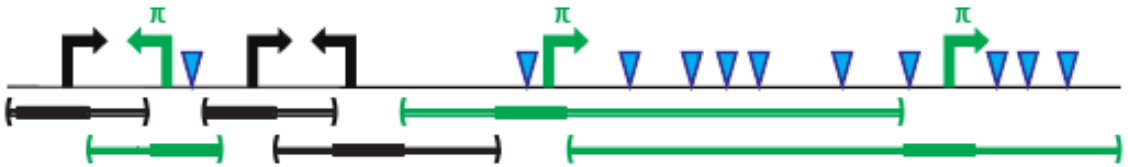
http://great.stanford.edu/

http://www.schule-bw.de/

Beispiel

p-Wert = $$\sum_{i=k_\pi}^{min(n,K_\pi)} \frac{\binom{K_\pi}{i}\binom{N-K_\pi}{n-i}}{\binom{N}{n}}$$





Hypergeometric test over genes
N = 6 total genes
K_π = 3 genes annotated with π
n = 3 genes with an associated genomic region
k_π = 3 genes annotated and with a genomic region
P-value = 0.05

Frage: ist die Annotation π in der Testmenge von 3 Genen signifikant angereichert?

Ja! p = 0.05 ist (knapp) signifikant.

http://great.stanford.edu/

Anwendung auf ALL-Beispiel

```
> ll <- mget(geneNames(esetSel),
    hgu95av2LOCUSID)
> ll <- unique(unlist(ll))
> mf <- as.data.frame(GOHyperG(ll))[, 1:3]
> mf <- mf[mf$pvalue < 0.01, ]
> mf
```

Bioconductor Kommandos

	p-values	goCounts	intCounts
GO:0045012	1.1e-08	12	6
GO:0030106	6.4e-03	8	2
GO:0004888	7.0e-03	523	16
GO:0004872	9.7e-03	789	21

```
> as.character(unlist(mget(row.names(mf),
    GOTERM)))
[1] "MHC class II receptor activity"
[2] "MHC class I receptor activity"
[3] "transmembrane receptor activity"
[4] "receptor activity"
```

Figure 3
Hypergeometric analysis of molecular function enrichment of genes selected in the analysis described in Figure 1.

Die signifikanteste Anreicherung ergibt sich für MHC Klasse 2 Rezeptoraktivität.

6 von 12 Genen im Genom mit dieser Annotation sind in den 2 ALL-Klassen differentiell exprimiert.

Gentleman et al. Genome Biology 5, R80 (2004)

NIH Tool David: Tool für Annotation der Genfunktion

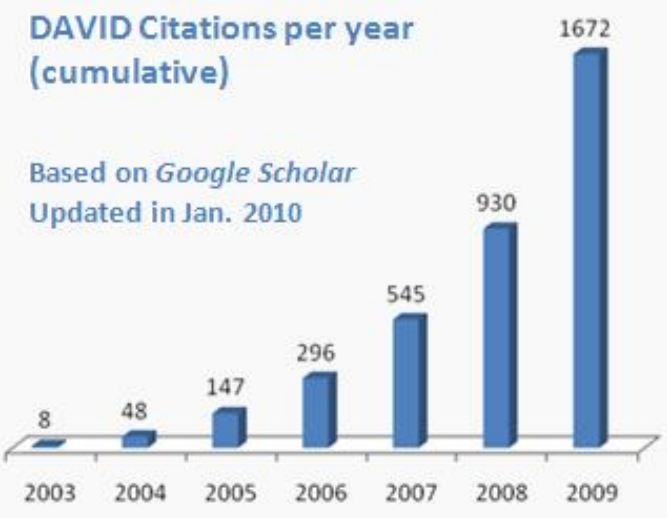
PROTOCOL

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)

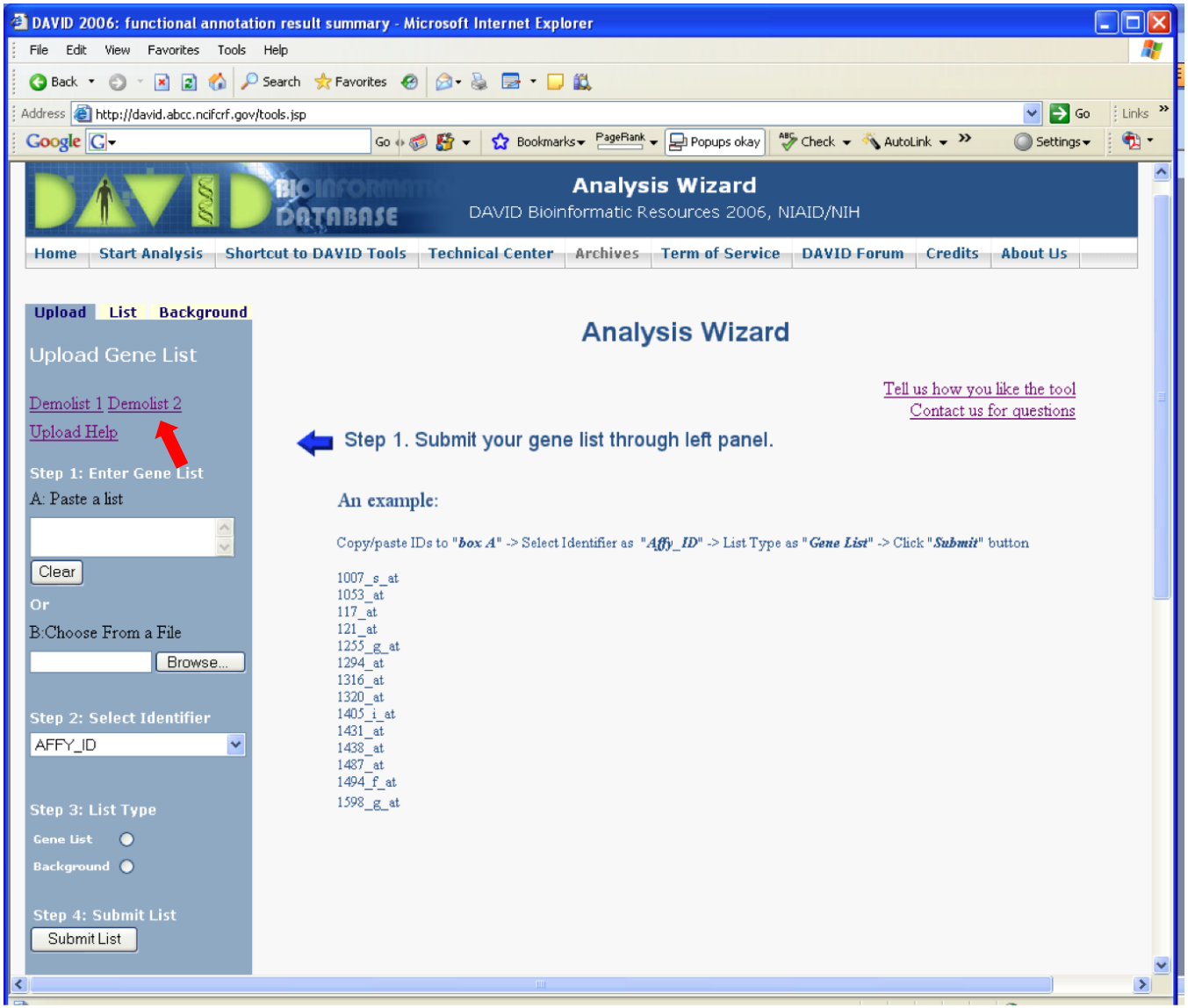
Published online 18 December 2008; doi:10.1038/nprot.2008.211



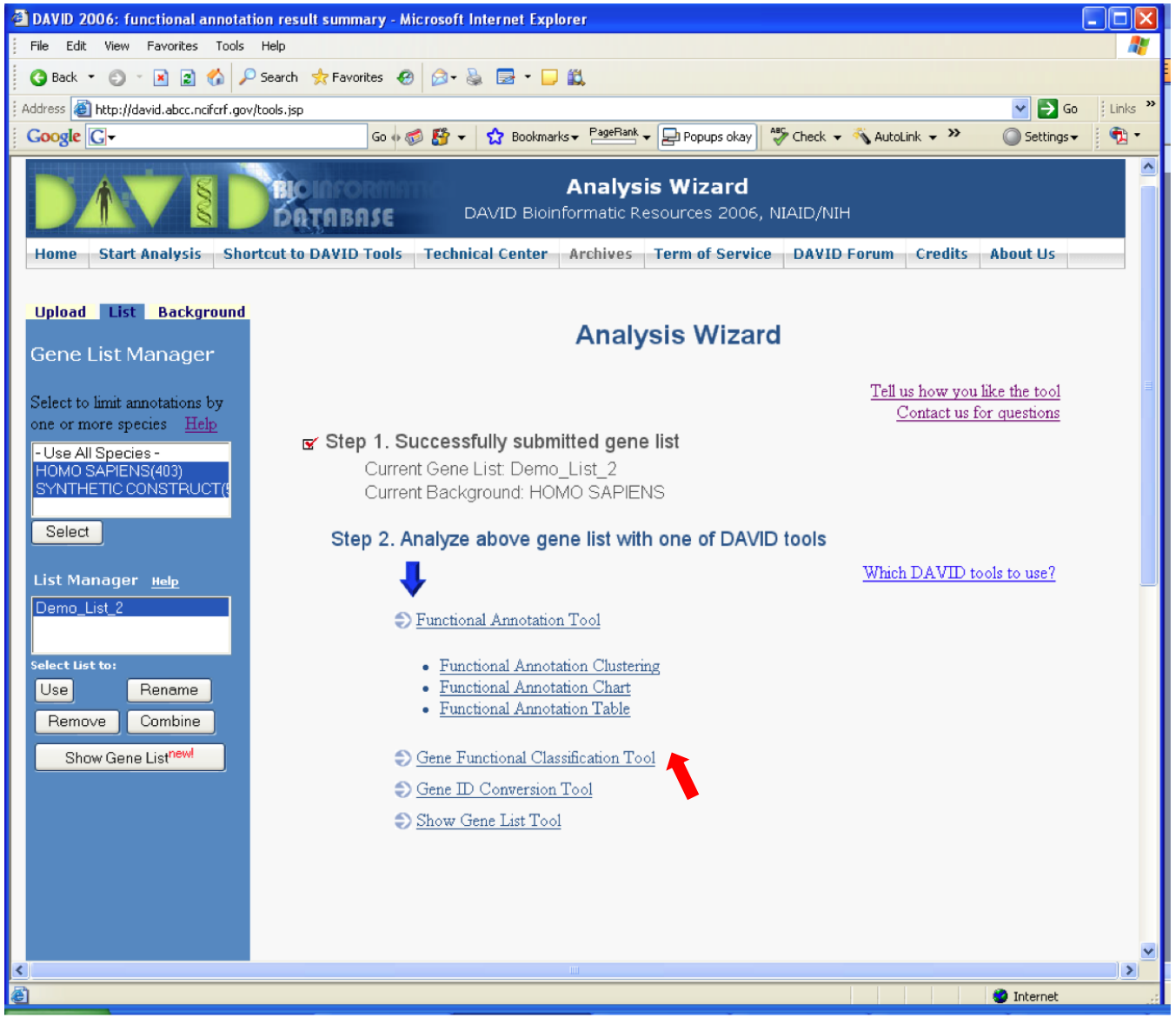
Huang et al. Nat. Protocol 4, 44 (2009)

NIH Tool David

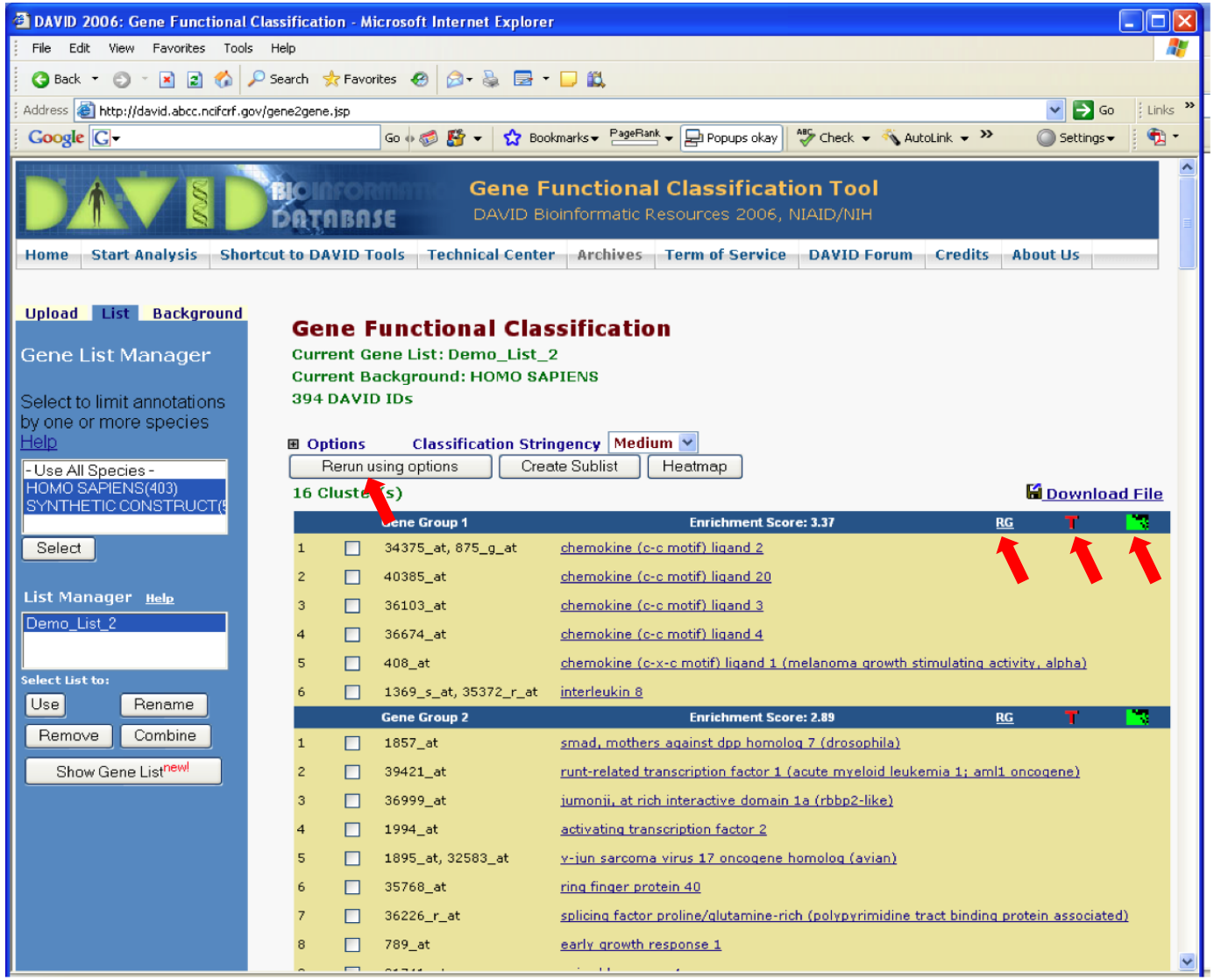
Submit gene list or use built-in demo_lists



Select the DAVID Gene Functional Classification Tool

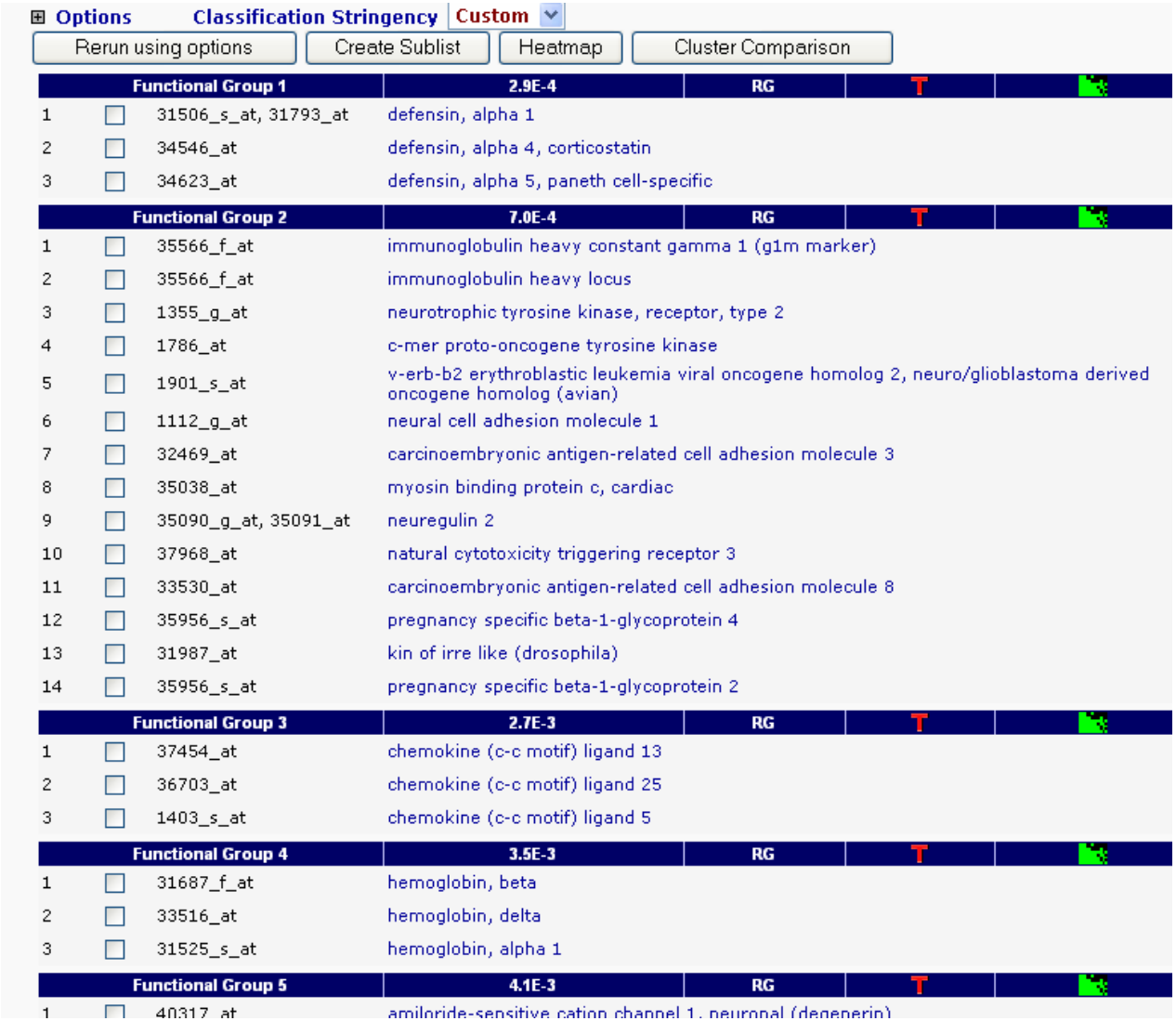


Select the DAVID Gene Functional Classification Tool



Select the DAVID Gene Functional Annotation Tool

Funktionelles Clustering von angereicherten GO-Termen



Options Classification Stringency Custom

Rerun using options | Create Sublist | Heatmap | Cluster Comparison

	Functional Group 1	2.9E-4	RG
1	31506_s_at, 31793_at	defensin, alpha 1	
2	34546_at	defensin, alpha 4, corticostatin	
3	34623_at	defensin, alpha 5, paneth cell-specific	

	Functional Group 2	7.0E-4	RG
1	35566_f_at	immunoglobulin heavy constant gamma 1 (g1m marker)	
2	35566_f_at	immunoglobulin heavy locus	
3	1355_g_at	neurotrophic tyrosine kinase, receptor, type 2	
4	1786_at	c-mer proto-oncogene tyrosine kinase	
5	1901_s_at	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	
6	1112_g_at	neural cell adhesion molecule 1	
7	32469_at	carcinoembryonic antigen-related cell adhesion molecule 3	
8	35038_at	myosin binding protein c, cardiac	
9	35090_g_at, 35091_at	neuregulin 2	
10	37968_at	natural cytotoxicity triggering receptor 3	
11	33530_at	carcinoembryonic antigen-related cell adhesion molecule 8	
12	35956_s_at	pregnancy specific beta-1-glycoprotein 4	
13	31987_at	kin of irre like (drosophila)	
14	35956_s_at	pregnancy specific beta-1-glycoprotein 2	

	Functional Group 3	2.7E-3	RG
1	37454_at	chemokine (c-c motif) ligand 13	
2	36703_at	chemokine (c-c motif) ligand 25	
3	1403_s_at	chemokine (c-c motif) ligand 5	

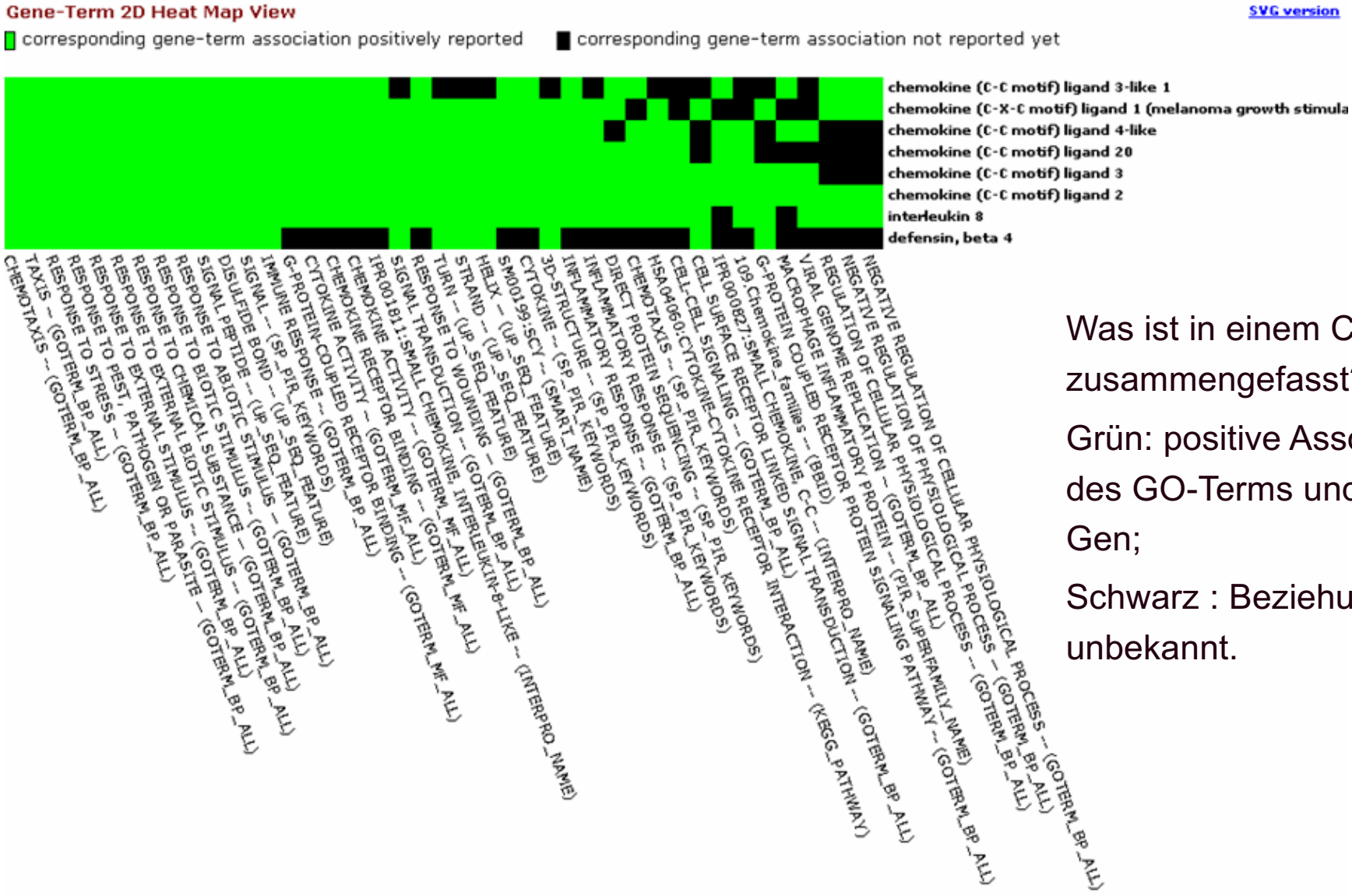
	Functional Group 4	3.5E-3	RG
1	31687_f_at	hemoglobin, beta	
2	33516_at	hemoglobin, delta	
3	31525_s_at	hemoglobin, alpha 1	

	Functional Group 5	4.1E-3	RG
1	40317_at	amiloride-sensitive cation channel 1, neuronal (degenerin)	

XXXX_at sind die Kürzel für einzelne Proben auf Affymetrix-Microarray-Chip

Huang *et al. Genome Biology 2007* ***8:R183***

David: Genes-to-terms 2D view



Was ist in einem Cluster zusammengefasst?

Grün: positive Assoziation des GO-Terms und einem Gen;

Schwarz : Beziehung ist unbekannt.

Huang *et al. Genome Biology 2007* ***8:R183***

Vergleich von GO-Termen

Die hierarchische Struktur der GO-Ontologie ermöglicht es, Proteine miteinander zu vergleichen, die mit verschiedenen GO-Termen annotiert sind.

Dies geht so lange, wie die Terme Beziehungen zueinander haben.

Nahe beieinander liegende Terme im GO-Graphen (d.h. mit wenigen dazwischen liegenden Termen) sind tendentiell **semantisch ähnlicher** zueinander als solche, die weiter voneinander entfernt sind.

Man könnte einfach die **Anzahl an Kanten** zwischen 2 Knoten als Maß für ihre Ähnlichkeit nehmen.

Dies ist jedoch problematisch, da verschiedene Regionen der GO-Ontologie unterschiedlich dicht mit Termen abgedeckt sind.

Gaudet, Škunca, Hu, Dessimoz
Primer on the Gene Ontology,
https://arxiv.org/abs/1602.01876

Messe funktionelle Ähnlichkeit von GO-Termen

Die **Wahrscheinlichkeit eines Knoten** *t* kann man auf 2 Arten ausdrücken:

Wieviele Gene besitzen die Annotation *t* relativ zur Häufigkeit der Wurzel?

Anzahl an GO-Termen im bei *t* startenden Unterbaum relativ zu der Anzahl an GO-Termen im Gesamtbaum.

$$p_{anno}(t) = \frac{occur(t)}{occur(root)} \qquad p_{graph}(t) = \frac{D(t)}{D(root)}$$

Die Wahrscheinlichkeit hat Werte zwischen 0 und 1 und nimmt zwischen den Blättern bis zur Wurzel monoton zu.

Aus der Wahrscheinlichkeit *p* berechnet man den **Informationsgehalt** jedes Knotens:

$$IC(t) = -\log p(t)$$

Je seltener ein Knoten ist, desto höher sein Informationsgehalt.

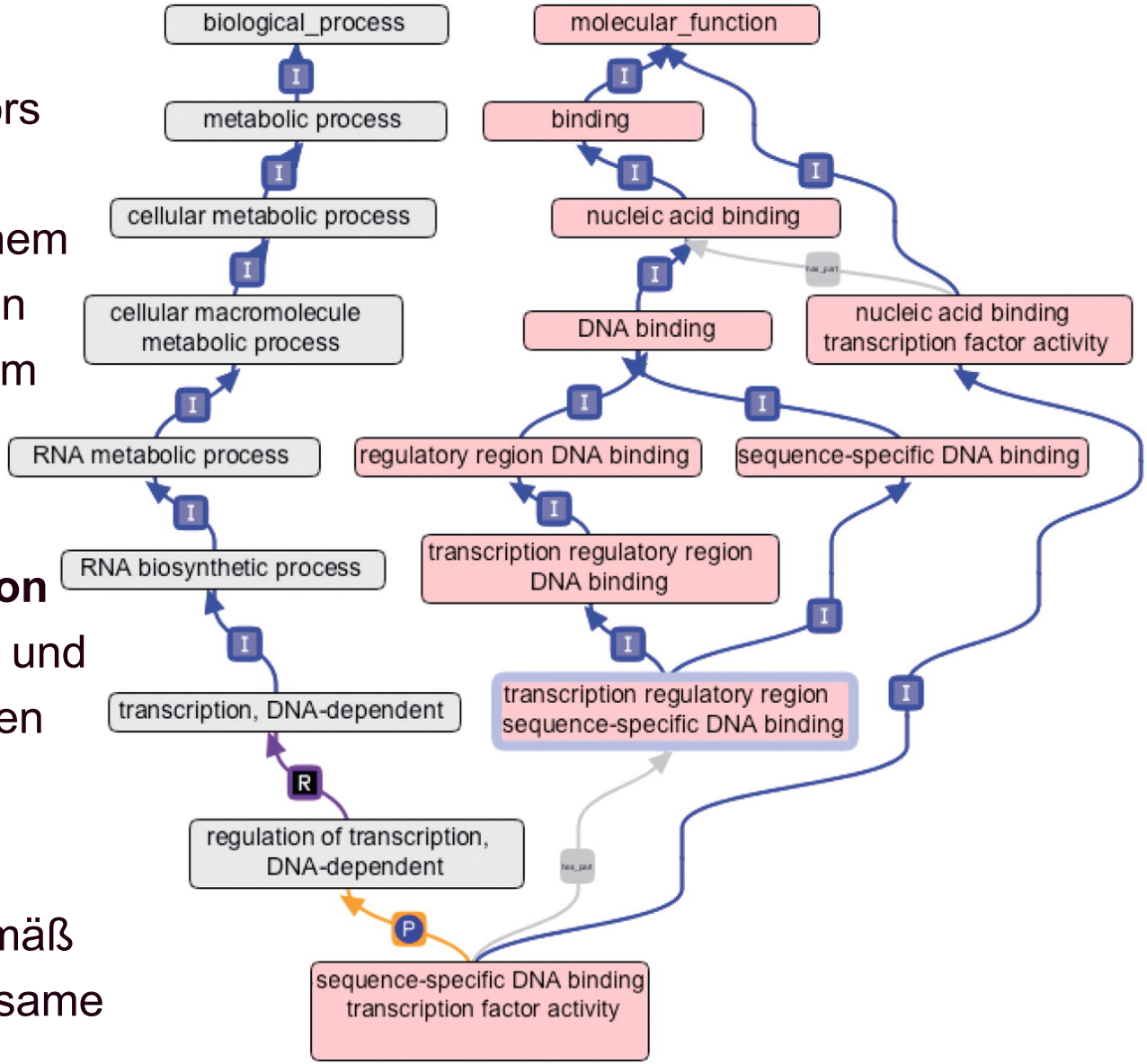
Schlicker PhD dissertation (2010)

Messe funktionelle Ähnlichkeit von GO-Termen

Die Menge an gemeinsamen Vorgängern (common ancestors (CA)) zweier Knoten t_1 und t_2 enthält alle Knoten, die auf einem Pfad von t_1 zum Wurzel-Knoten **UND** auf einem Pfad von t_2 zum Wurzelknoten liegen.

Der **most informative common ancestor** (MICA) der Terme t_1 und t_2 ist der Term mit dem höchsten Informationsgehalt in CA.

Normalerweise ist das der gemäß dem Abstand nächste gemeinsame Vorgänger.



Nucl. Acids Res. (2012) 40 (D1): D559-D564

Messe funktionelle Ähnlichkeit von GO-Termen

Schlicker et al. definierten aus dem Abstand zum **most informative common ancestor** (MICA) die Ähnlichkeit der Terme t_1 und t_2

$$sim_{Rel}(t_1,t_2) = \frac{2 \cdot IC(MICA)}{IC(t_1)+IC(t_2)} \cdot (1 - p(MICA))$$

Der hintere Faktor gewichtet die Ähnlichkeit mit der Häufigkeit *p(MICA)*. Dies ergab Vorteile in der Praxis.

Schlicker PhD dissertation (2010)

Messe funktionelle Ähnlichkeit von GO-Termen

Zwei Gene oder zwei Mengen an Genen *A* und *B* haben jedoch meist jeweils mehr als eine GO-Annotation. Betrachte daher die Ähnlichkeit aller Terme *i* und *j*:

$$s_{ij} = sim(GO_i^A, GO_j^B), \forall i \in 1,...,N, \forall j \in 1,...,M.$$

und wähle daraus in den Reihen und Spalten jeweils die Maxima

$$rowScore(A,B) = \frac{1}{N}\sum_{i=1}^{N} \max_{1 \le j \le M} s_{ij}, \qquad GOscore_{avg}^{BMA}(A,B) = \frac{1}{2} \cdot (rowScore(A,B) + columnScore(A,B))$$

$$columnScore(A,B) = \frac{1}{M}\sum_{j=1}^{M} \max_{1 \le i \le N} s_{ij}. \qquad GOscore_{max}^{BMA}(A,B) = max(rowScore(A,B), columnScore(A,B))$$

Aus den Scores für den BP-Baum und den MF-Baum wird der *funsim*-Score berechnet.

$$funsim(A,B) = \frac{1}{2} \cdot \left[\left(\frac{BPscore}{\max(BPscore)} \right)^2 + \left(\frac{MFscore}{\max(MFscore)} \right)^2 \right]$$

Schlicker PhD dissertation (2010)

GO ist unvollständig

Die Gen-Ontologie repräsentiert eine Auswahl des aktuell verfügbaren **Wissens**. Daher ist sie sehr **dynamisch**.

Die Ontologie wird ständig verbessert um die Biologie aller Organismen möglichst genau darzustellen.

Sobald neue Entdeckungen gemacht warden, werden diese in GO aufgenommen.

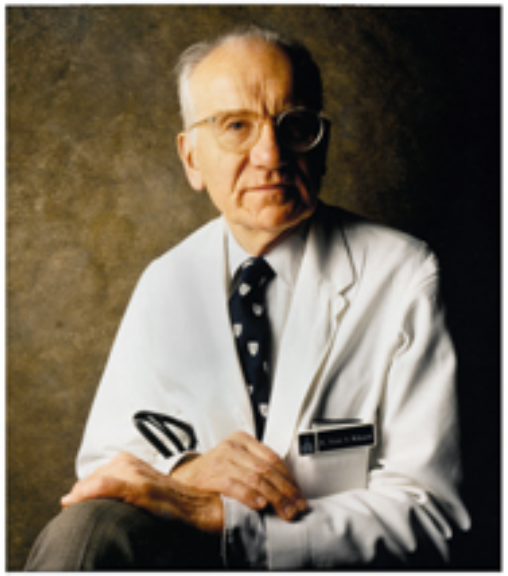
Allerdings stellt die Geschwindigkeit der aktuellen Forschung das GO-Konsortium vor die hohe Herausforderung, damit Schritt zu halten.

Auf jeden Fall ist die Information in GO notwendigerweise **unvollständig.**

Daher bedeutet fehlende Evidenz über (eine bestimmte) Funktion NICHT, dass diese Funktion nicht vorliegt.

Gaudet, Dessimoz,
Gene Ontology: Pitfalls, Biases, Remedies
https://arxiv.org/abs/1602.01876

OMIM-Datenbank



Victor McKusick (1921-2008),
Johns Hopkins Universität,
- begründete das Gebiet
Medical genetics
- gründete die Datenbank
Mendelian Inheritance in Man

OMIM®, Online Mendelian Inheritance in Man®.

OMIM is a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes.

NCBI OMIM Online Mendelian Inheritance in Man

All Databases PubMed Nucleotide Protein Genome Structure PMC OMIM

My NCBI [Sign In] [Register]

Search OMIM for Go Clear

Limits Preview/Index History Clipboard Details

Display Detailed Show 20 Send to

All: 1 OMIM UniSTS: 0 OMIM dbSNP: 0

MIM ID #211980

GeneTests, Links

LUNG CANCER

Other entities represented by this entry

ALVEOLAR CELL CARCINOMA, INCLUDED
ADENOCARCINOMA OF LUNG, INCLUDED
NONSMALL CELL LUNG CANCER, INCLUDED
LUNG CANCER, PROTECTION AGAINST, INCLUDED

Gene map locus: 17q21.1, 12p12.1, etc.

Clinical Synopsis

Text Back to Top

A number sign (#) is used with this entry because mutations in several different genes are associated with lung cancer. Both germline and somatic mutations have been identified in the EGFR (131550) and p53 (TP53; 191170) genes, and somatic mutations have been identified in the KRAS (190070), BRAF (164757), ERBB2 (164870), MET (164860), STK11 (602216), PIK3CA (171834), and PARK2 (602544) genes. Amplification of several genes,

Table of Contents

MIM #211980
Text
Description
Clinical Features
Inheritance
Population Genetics
Pathogenesis
Clinical Management
Mapping
Molecular Genetics
Cytogenetics
Clinical Synopsis
See Also
References
Contributors
Creation Date
Edit History

Links

BIOINFORMATICS

Vol. 26 ECCB 2010, pages i561–i567
doi:10.1093/bioinformatics/btq384

Improving disease gene prioritization using the semantic similarity of Gene Ontology terms

Andreas Schlicker†, Thomas Lengauer and Mario Albrecht*

Max Planck Institute for Informatics, Department of Computational Biology and Applied Algorithmics, Campus E1.4, 66123 Saarbrücken, Germany

ONIM-Datenbank &

UniProt Datenbank:

GO-Annotationen für bekannte Krankheits-gene.

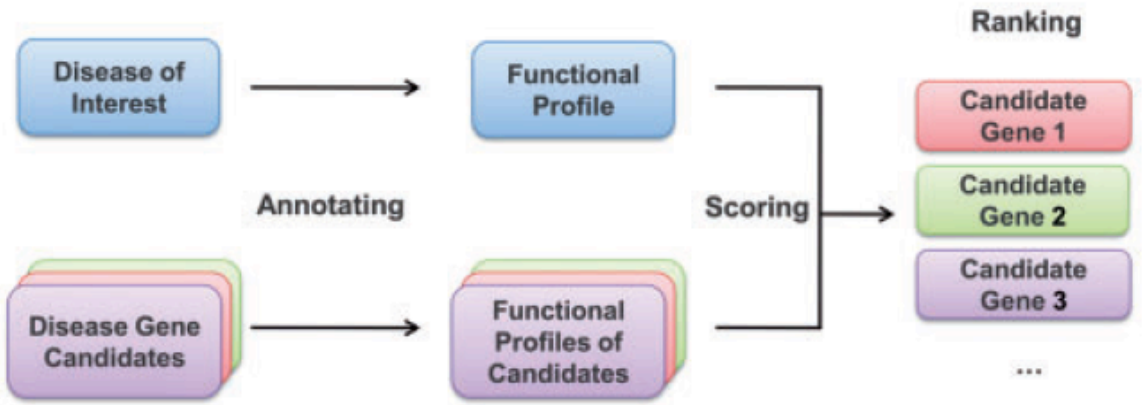
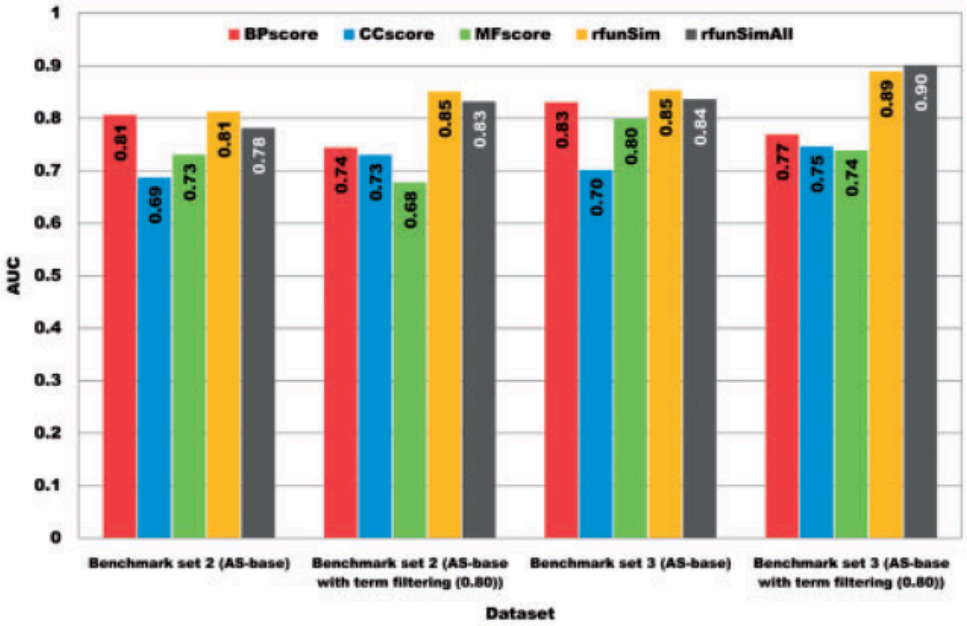


Fig. 1. Flow chart of the MedSim approach. First, the functional profiles of the disease of interest and the disease gene candidates are created using one of the annotation strategies. Afterwards, the functional profile of the disease is scored against each functional profile of a candidate, and the candidates are ranked according to this functional similarity score.

Schlicker et al. Bioinformatics 26, i561 (2010)

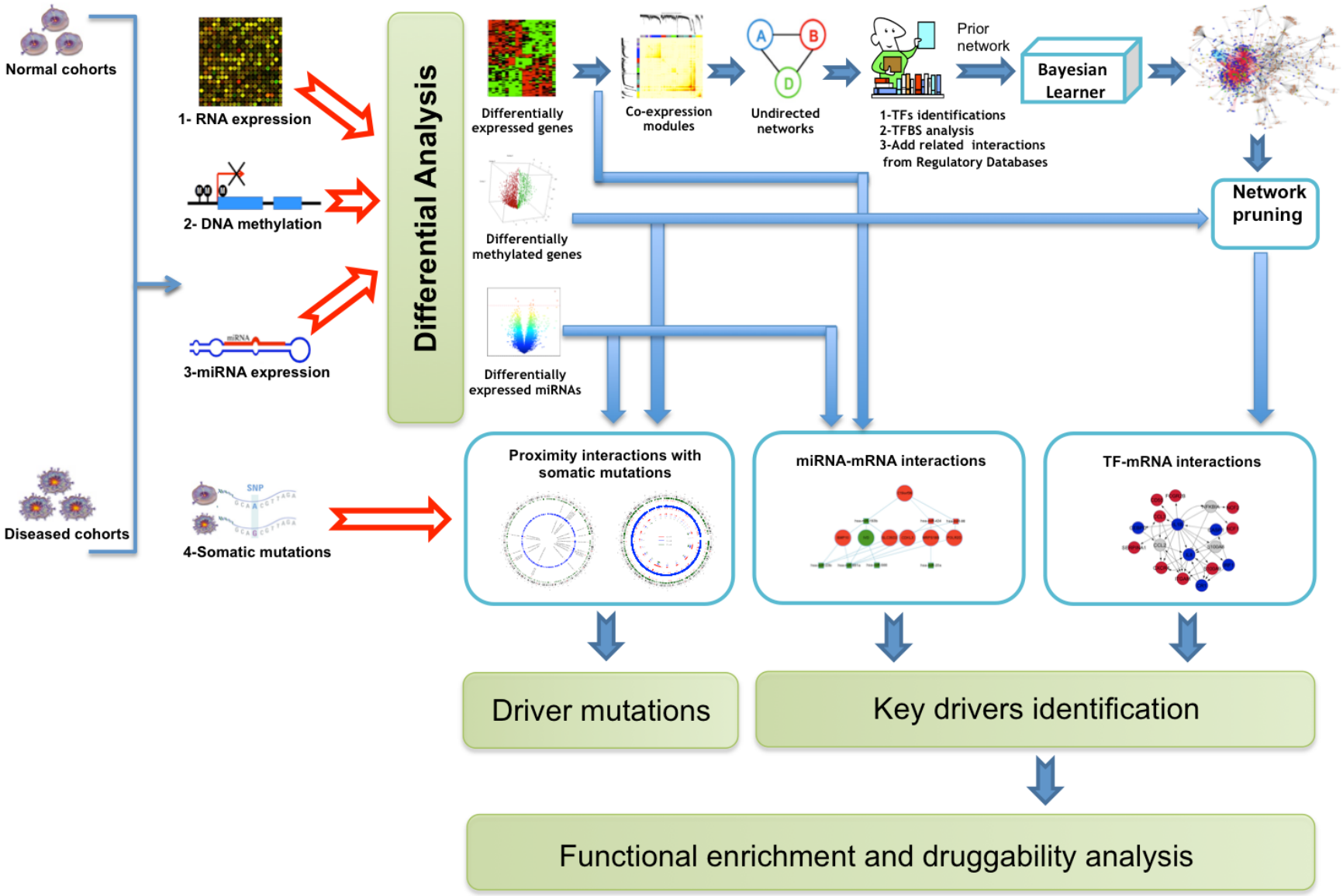
Die Methode liefert recht genaue Vorhersagen, mit welchen Krankheiten Gene in Verbindung stehen könnten.

Die Sensitivität, d.h. die Anzahl der korrekten Vorhersagen relativ zur Anzahl aller Vorhersagen, beträgt 73%.



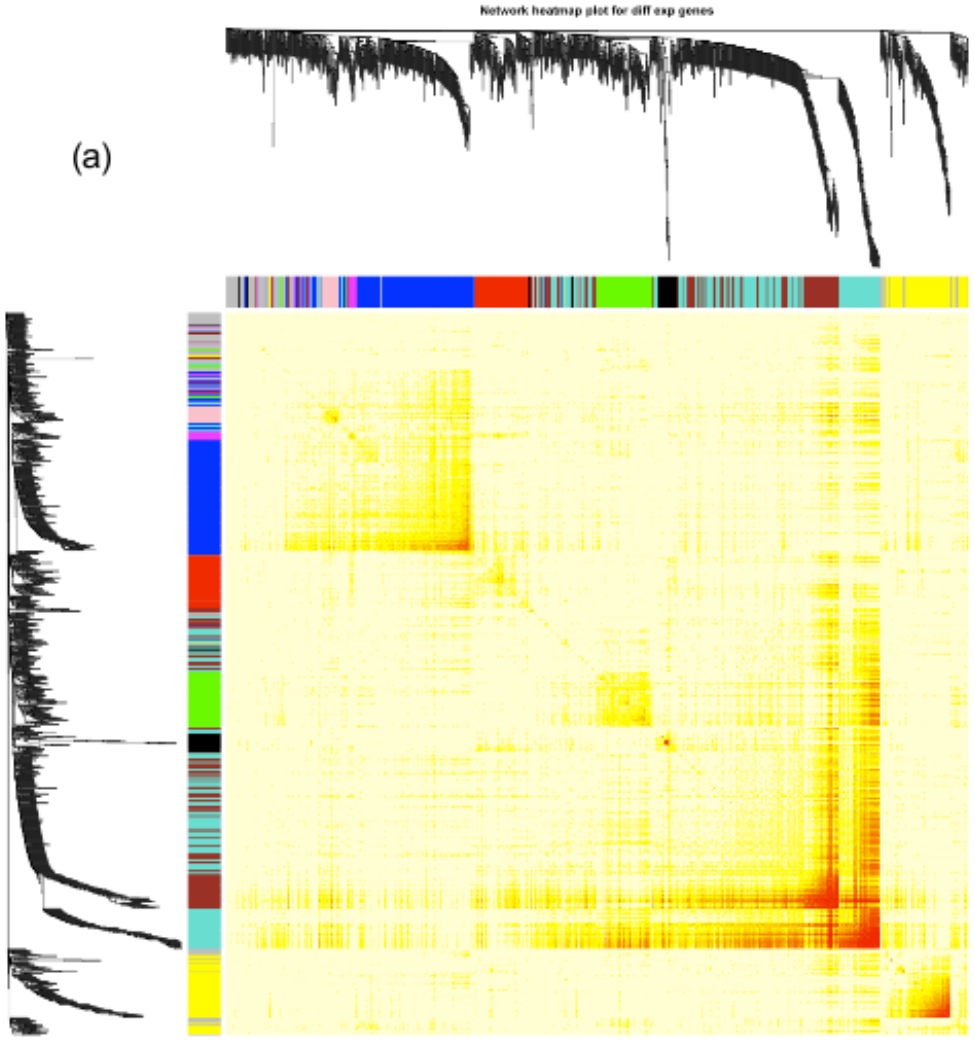
Schlicker et al. Bioinformatics 26, i561 (2010)

funktionelle Annotation von OMICS-Daten für Brustkrebs



Hamed et al. BMC Genomics (2015)

Analyse von Ko-Expression



Ko-Expression der 1317 differenziell exprimierten Gene (Krebs vs. Normal)

Hierarchisches Clustern

-> 10 Module mit 26 – 295 Genen

Hamed et al. BMC Genomics (2015)

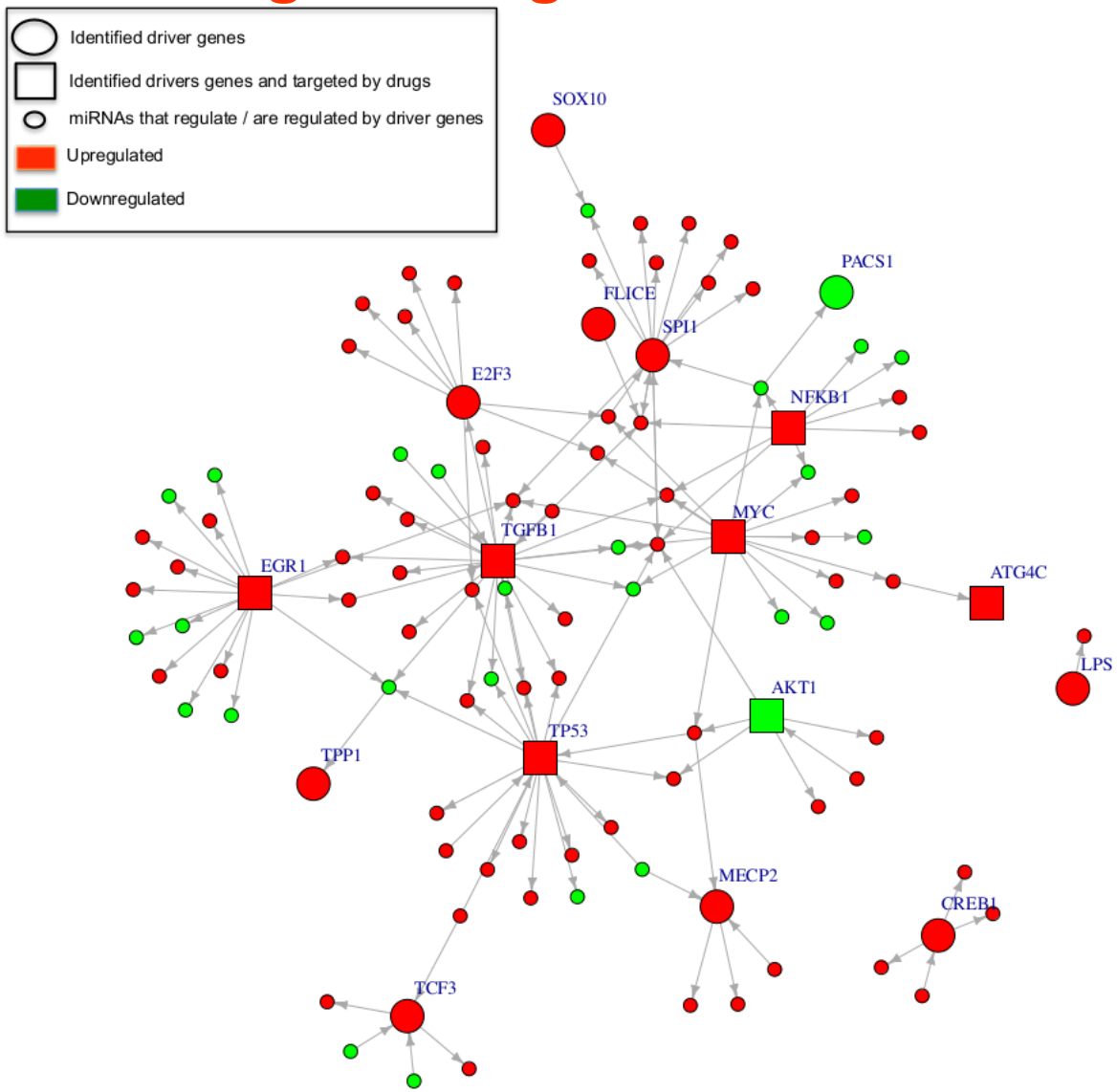
Gibt es angereicherte Genfunktionen in diesen Modulen?

	Module	Gene count	Top GO category	Top KEGG categories	Key driver count	Key drivers
TF- mRNA interactions	black	41	Regulation of transcription	Pathways in cancer, Renal cell carcinoma	5	SORBS3, ZNF43, ZNF681, RBMX, POU2F1
	blue	247	Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	Cell cycle, Prostate cancer, Melanoma	9	AR, BRCA1, ESR1, JUN, MYB, RPN1, E2F1, E2F2, PPARD
	brown	195	Anatomical structure morphogenesis	Leukocyte transendothelial migration	5	TMOD3, CREB1, POU5F1, SP3, TERT
	green	110	Cellular macromolecule metabolic process	Endometrial cancer, Insulin signaling pathway	15	B4GALT7, OS9, CDC34, MAN2C1, MYO1C, SH3GLB2, INPP5E, PLXNB1, USF2, PPP1R12C, CDK9, DAP, E4F1, E2F4, USF1
	grey	148	Anatomical structure development	Sulfur metabolism	18	AHCTF1, NQO2, FGFR2, CCDC130, ABCG4, BIRC6, CA6, SP4, RNF2, SPRR1B, C16orf65, DNAJC5G, SNCAIP, GRIK5, SLC6A4, SMAD1, DAD1, POU4F2
	magenta	26	Regulation of metabolic process	p53 signaling pathway, Alzheimer's disease	3	ATF6, NGEF, POGK
	pink	30	Transcription initiation from RNA polymerase II promoter	Basal transcription factors	4	CCDC92, TMEM70, RNF139, E2F5
	red	93	Regulation of cellular process	Endometrial cancer, Neurotrophin signaling pathway	14	ATP1B1, STAT3, ABCB8, MYC, TGFB1, SP1, TP53, PCGF1, SUMF2, GTF3A, IPO13, GMPPA, HTR6, TGIF1
	turquoise	295	Regulation of cellular metabolic process	p53 signaling pathway, Pancreatic cancer, Apoptosis	2	UBL5, RNF111
	yellow	132	Immune system process	Chemokine signaling pathway, Natural killer cell mediated cytotoxicity	19	APOC1, CD2, CD79B, LRRC28, DAPK1, FAM124B, EML2, LAP3, TSPAN2, FCRL3, ELMO1, SLC7A7, RASSF5, SLC31A2, TRAF3IP3, GALNT12, ITGA4, SPI1, TFAP2A
	Total	1317				

Module hängen mit Prozessen zusammen, die bereits mit Brustkrebs in Verbindung gebracht werden (endometrical cancer, p53, Prostatakrebs ...)

Hamed et al. BMC Genomics (2015)

Ergänze regulatorische Information + driver genes



Differenziell experimierte Gene eines Moduls

-> extrahiere regulatorische Interaktionen (TF -> Gen) aus den öffentlichen Datenbanken JASPAR, TRED, MSigDB

Driver genes sind Transkriptions-faktoren, die möglichst viele Gene des Moduls regulieren.

31% der Driver genes kodieren für Proteine, die Targets für bekannte Krebs-Medikamente sind!

Hamed et al. BMC Genomics (2015)

Ausblick auf den 3. Teil der Vorlesung

- Protein-Protein-Interaktionsnetzwerke – Analyse mit `Cytoscape`

- metabolische Netzwerke – Simulation mit `Copasi`

- Ko-Expression / Go-Annotation – Prozessierung mit `Bioconductor`