Genome Annotation - 2

Qi Sun

Bioinformatics Facility

Cornell University

Topics in this lecture

BLAST & HMM

Gene Ontology

Function Annotation

Function Enrichment

Given a protein, how to predict its function?

>unknown_protein

MEPSSETGMDPPLSQETFEDLWSLLPDPLQTVTCRLDNLSEFPDYPLAADMSVLQEGLMGNAVPTVTSCA
PSTDDYAGKYGLQLDFQQNGTAKSVTCTYSPELNKLFCQLAKTCPLLVRVESPPPRGSILRATAVYKKSE
HVAEVVKRCPHHERSVEPGEDAAPPSHLMRVEGNLQAYYMEDVNSGRHSVCVPYEGPQVGTECTTVLYNY
MCNSSCMGGMNRRPILTIITLETPQGLLLGRRCFEVRVCACPGRDRRTEEDNYTKKRGLKPSGKRELAHP
PSSEPPLPKKRLVVDDDEEIFTLRIKGRSRYEMIKKLNDALELQESLDQQKVTIKCRKCRDEIKPKKGKK
LLVKDEQPDSE

BLAST

mutant tumor protein p53 [Homo sapiens]

Sequence ID: AYE20623.1 **Length:** 393 **Number of Matches:** 1

Range 1: 3 to 393 GenPept Graphics ▼ Next Match ▲ Previou

Score	Expect	Method	Identities	Positives	Gaps
345 bits(884)	4e-116	Compositional matrix adjust.	203/399(51%)	252/399(63%)	47/399(11%)

```
Query  2    EPSSETGMDPPLSQETFEDLWSLLPD-----PLQTVTCRLDNLSEFPD---YPLAADMSV  53
            EP S+  ++PPLSQETF DLW LLP+     PL +    +D+L   PD        D
Sbjct  3    EPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQA--MDDLMLSPDDIEQWFTEDPGP  60

Query  54   LQEGLMGNAVPT--------------------VTSCAPSTDDYAGKYGLQLDFQQNGTAK  93
             +   M  A P                    ++S  PS   Y G YG +L F  +GTAK
Sbjct  61   DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK  120

Query  94   SVTCTYSPELNKLFCQLAKTCPLLVRVESPPPRGSILRATAVYKKSEHVAEVVKRCPHHE  153
            SVTCTYSP LNK+FCQLAKTCP+ + V+S PP G+ +RA A+YK+S+H+ EVV+RCPHHE
Sbjct  121  SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE  180

Query  154  RSVEPGEDAAPPSHLMRVEGNLQAYYMEDVNSGRHSVCVPYEGPQVGTECTTVLYNYMCN  213
            R  +  +  APP HL+RVEGNL+  Y++D N+ RHSV VPYE P+VG++CTT+ YNYMCN
Sbjct  181  RCSD-SDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCN  239

Query  214  SSCMGGMNRRPILTIITLETPQGLLLGRRCFEVRVCACPGRDRRTEEDNYTKK----RGL  269
            SSCMGGMNRRPILTIITLE   G LLGR  FEVRVCACPGRDRRTEE+N+ KK      L
Sbjct  240  SSCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENFRKKGEPHHEL  299

Query  270  KP-SGKRELAHPPSSEPPLPKKRLVVDDDEEIFTLRIKGRSRYEMIKKLNDALELQESLD  328
             P S KR L +  SS P   KK L    D E FTL+I+GR R+EM ++LN+ALEL+++
Sbjct  300  PPGSTKRALPNNTSSSPQPKKKPL----DGEYFTLQIRGRERFEMFRELNEALELKDA-Q  354

Query  329  QQKVTIKCRKCRDEIKPKKG------KKLLVKDEQPDSE  361
              K     R     +K KKG      KKL+ K E PDS+
Sbjct  355  AGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD  393
```

HMM

Domain	Description	Evalue
P53_TAD	P53 transactivation motif	3.80E-10
P53	P53 DNA-binding domain	2.70E-59
P53_tetramer	P53 tetramerisation motif	1.30E-17

How does BLAST work

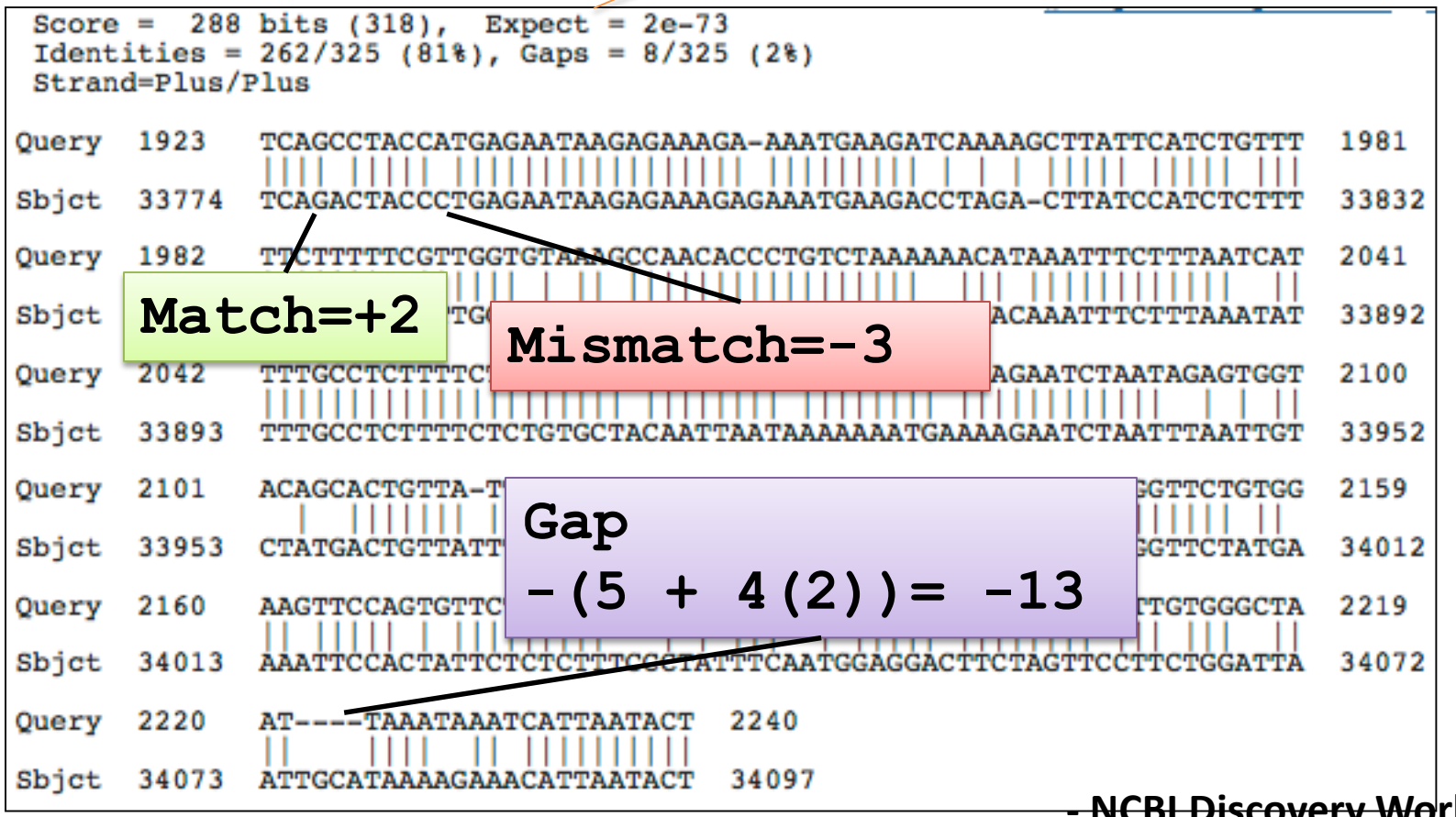
Step 1. find alignments

ACCAGAGGACGATA **ACG** GGACTAAGCAGCTAGA

AACCGAGAGATCGGACGATA **ACG** GGACTAAGCAACGAAAGACGA

Step 2. scoring alignments

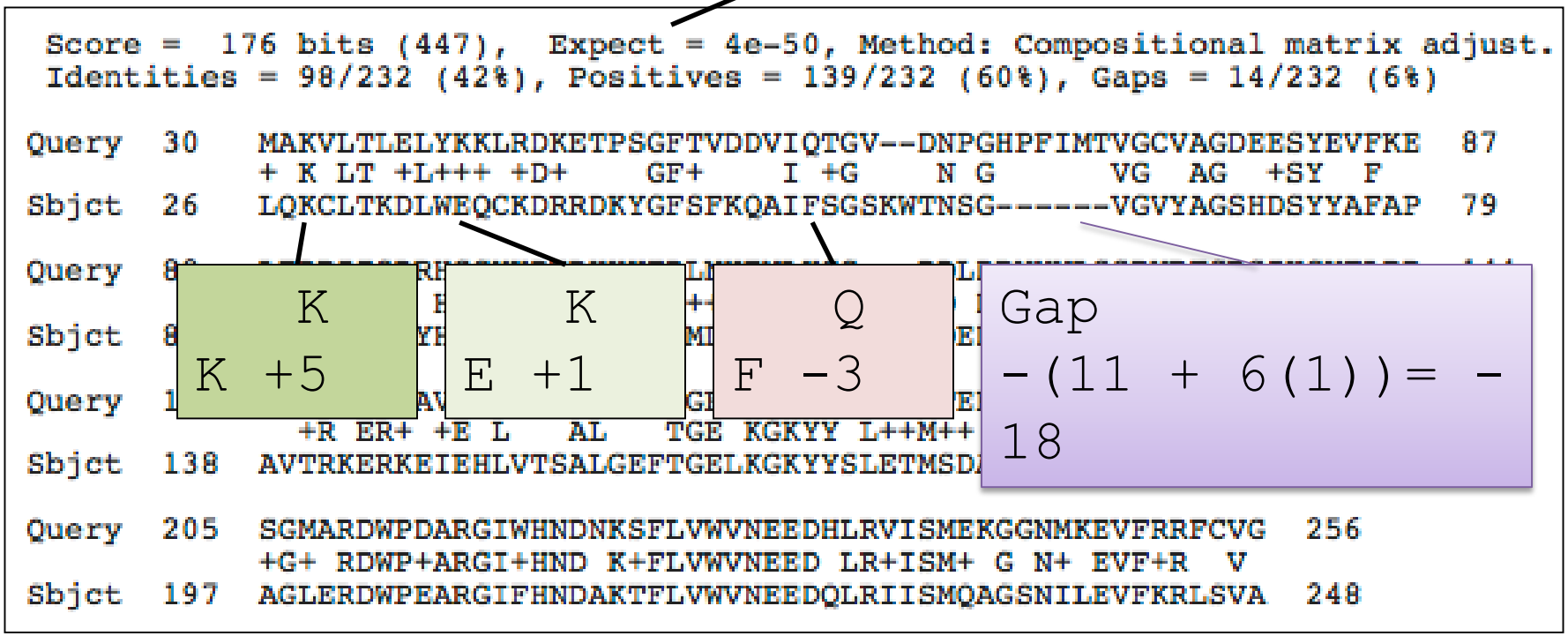
Number of Chance Alignments = 2 X 10^{-73}



How does BLAST work

Step 2. Score each alignment – protein alignment

Number of Chance Alignments = 4×10^{-50}



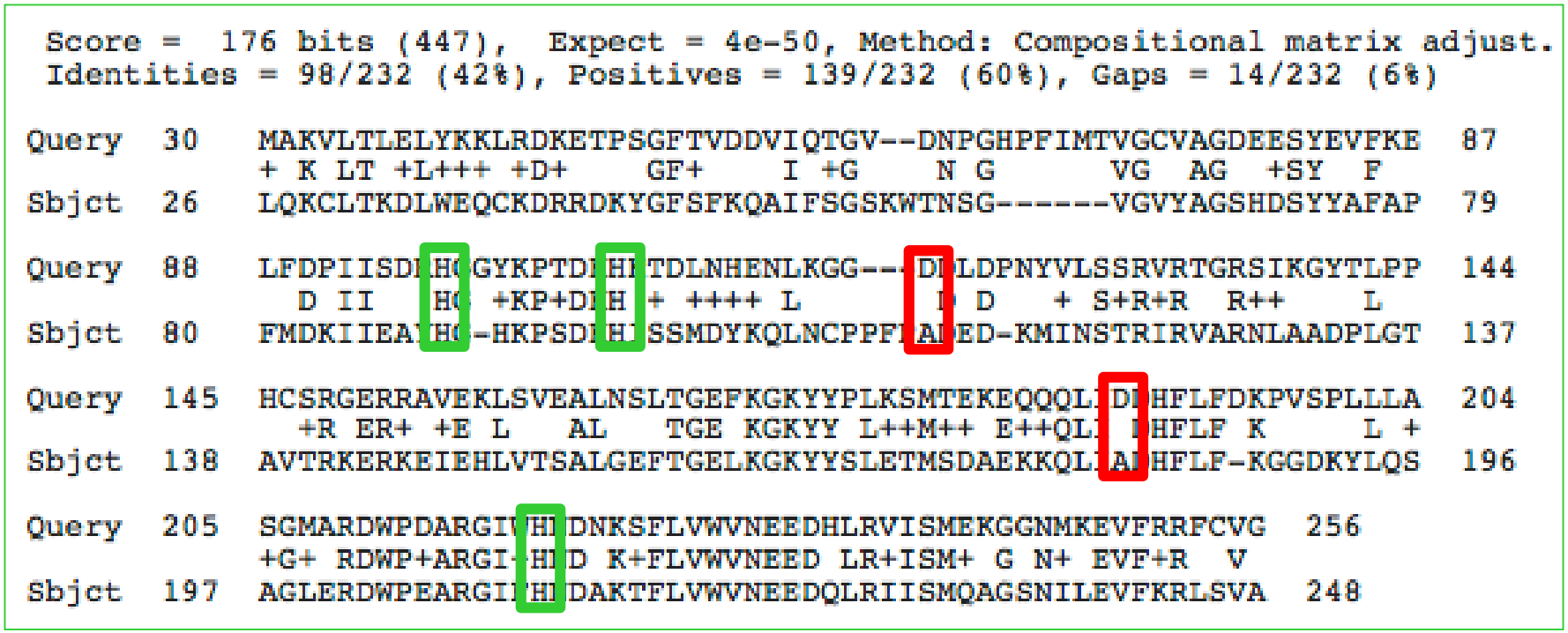
Scores from BLOSUM62, a position independent matrix

- NCBI Discovery Workshops

BLOSUM62, a position independent matrix

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

BLOSUM62 substitution score is position independent

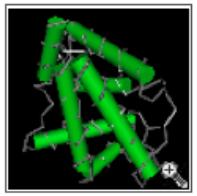


Scores from BLOSUM62, a position independent matrix

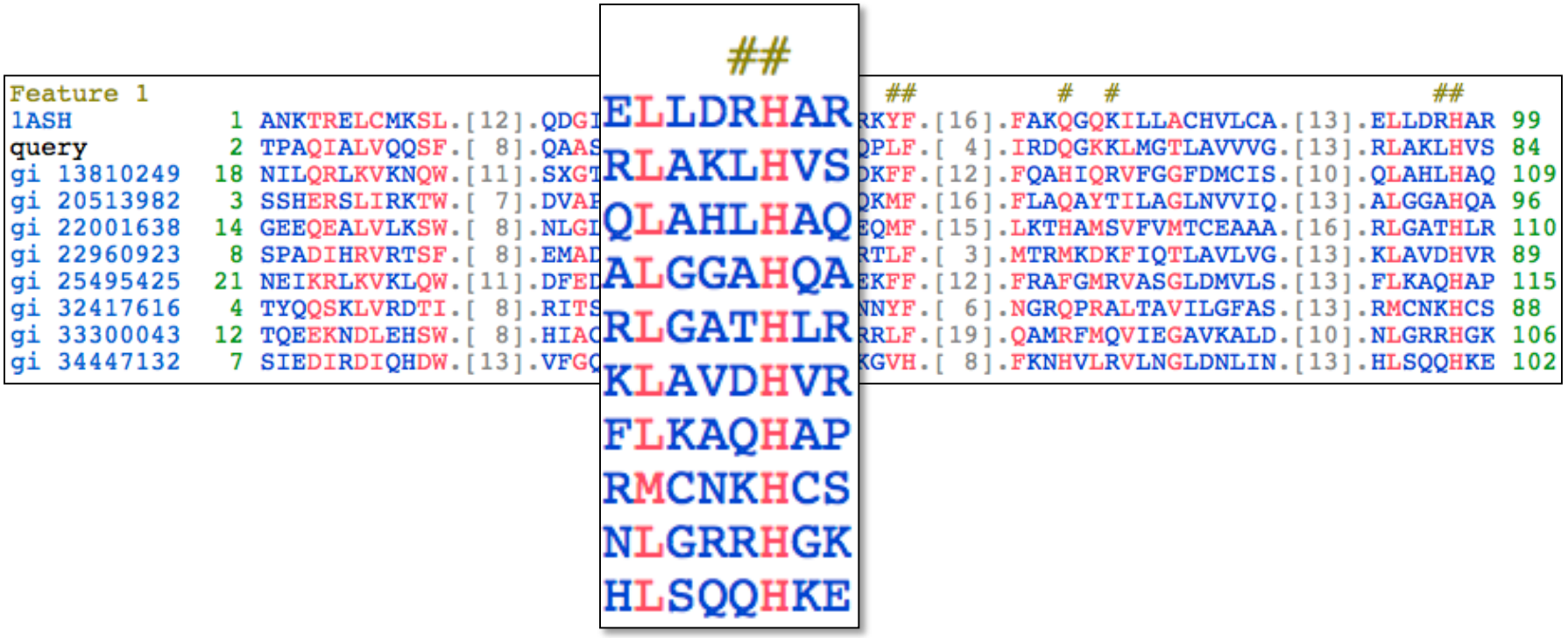
- NCBI Discovery Workshops

PSSM Alignment: Globins

cd01040: globin, with user query added ?



Globins are heme proteins, which bind and transport oxygen. This family summarizes a diverse set of homologous protein domains, including: (1) tetrameric vertebrate hemoglobins, which are the major protein component of erythrocytes and transport oxygen in the bloodstream, (2) microorganismal flavohemoglobins, which are linked to C-terminal FAD-dependend reductase domains, (3) homodimeric bacterial hemoglobins, such as from Vitreoscilla, (4) plant leghemoglobins (symbiotic hemoglobins, involved in nitrogen metabolism in plant rhizomes), (5) plant non-symbiotic hexacoordinate globins and hexacoordinate globins from bacteria and animals, such as neuroglobin, (6) invertebrate hemoglobins, which may occur in tandem-repeat arrangements, and (7) monomeric myoglobins found in animal muscle tissue.



Conserved Histidine

- NCBI Discovery Workshops

Heme Binding Site

Conserved Histidine

blastp

```
TFATLSELHCDKLHVD----PENFRLLG
      S L   KLHV      P ++  +G
ILPAASRLA--KLHVSYGVQPTHYAPVG
```

DELTA-BLAST

```
TF---ATLSELHCDKLHVDPENFRLLG
     + L++LH      V P ++  +G
ILPAASRLAKLHVS-YGVQPTHYAPVG
```

- NCBI Discovery Workshops

Heme Binding Site

Conserved Histidine

bla

DELTA-BLAST

ILPAASRLAKLHVS-YGVQPTHYAPVG

BLAST is not reliable for alignment of homologous genes between distantly related species.

- NCBI Discovery Workshops

Search PSSM with DELTA-BLAST

DELTA-BLAST employs a subset of NCBI's Conserved Domain Database (CDD) to construct PSSM

Enter Query Sequence

BLASTP programs search protein databases using a protein query. more...

Reset page | Bookmark

Enter accession number(s), gi(s), or FASTA sequence(s)

NP_001265090

Clear

Query subrange

From

To

Or, upload file

Choose File | No file chosen

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database: Reference proteins (refseq_protein)

Organism (Optional): Drosophila melanogaster | Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): Models (XM/XP) | Uncultured/environmental sample sequences

Entrez Query (Optional)

Enter an Entrez query to limit search

Program Selection

Algorithm

- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST

Search database Reference proteins (refseq_protein) using DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

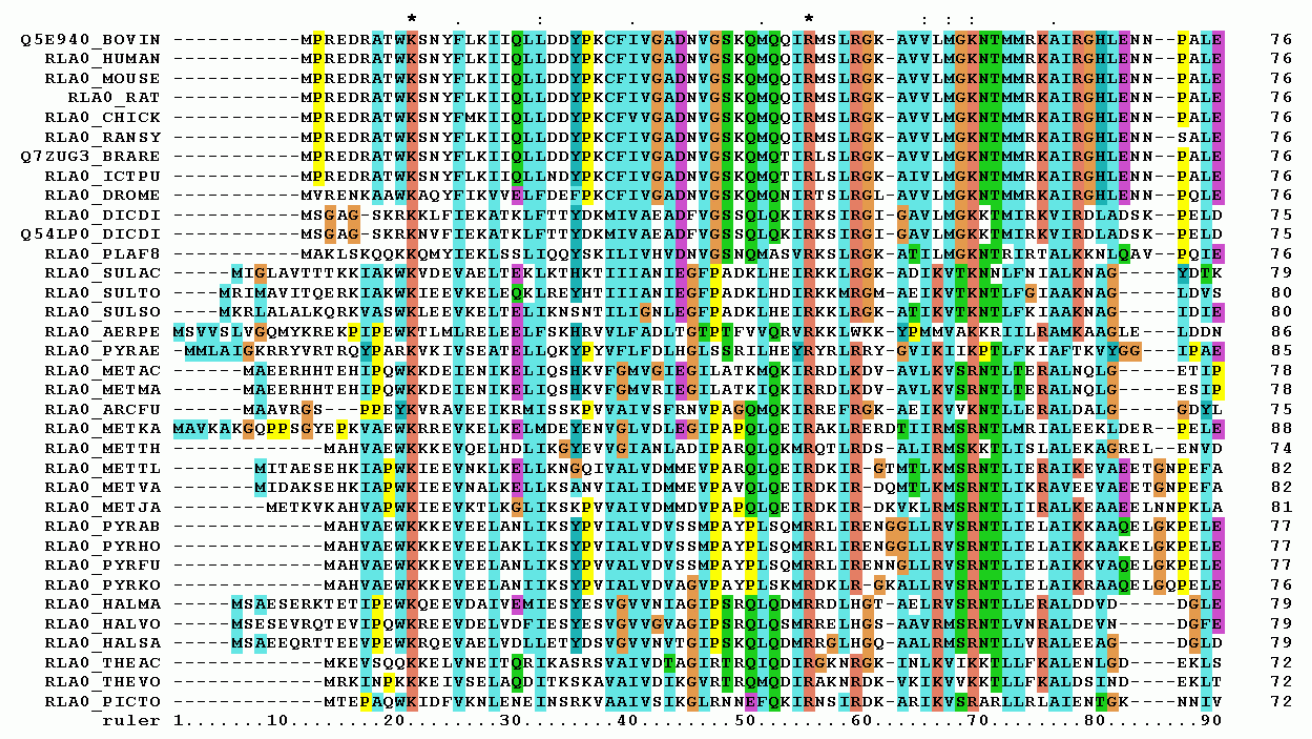
Show results in a new window

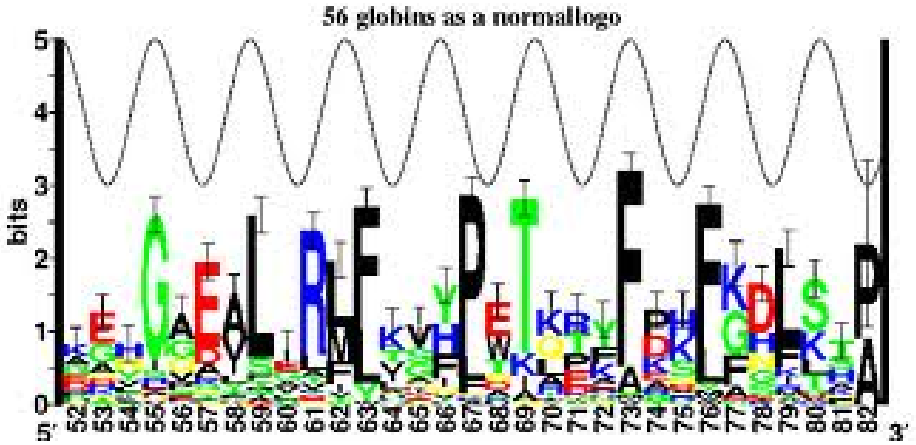
Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

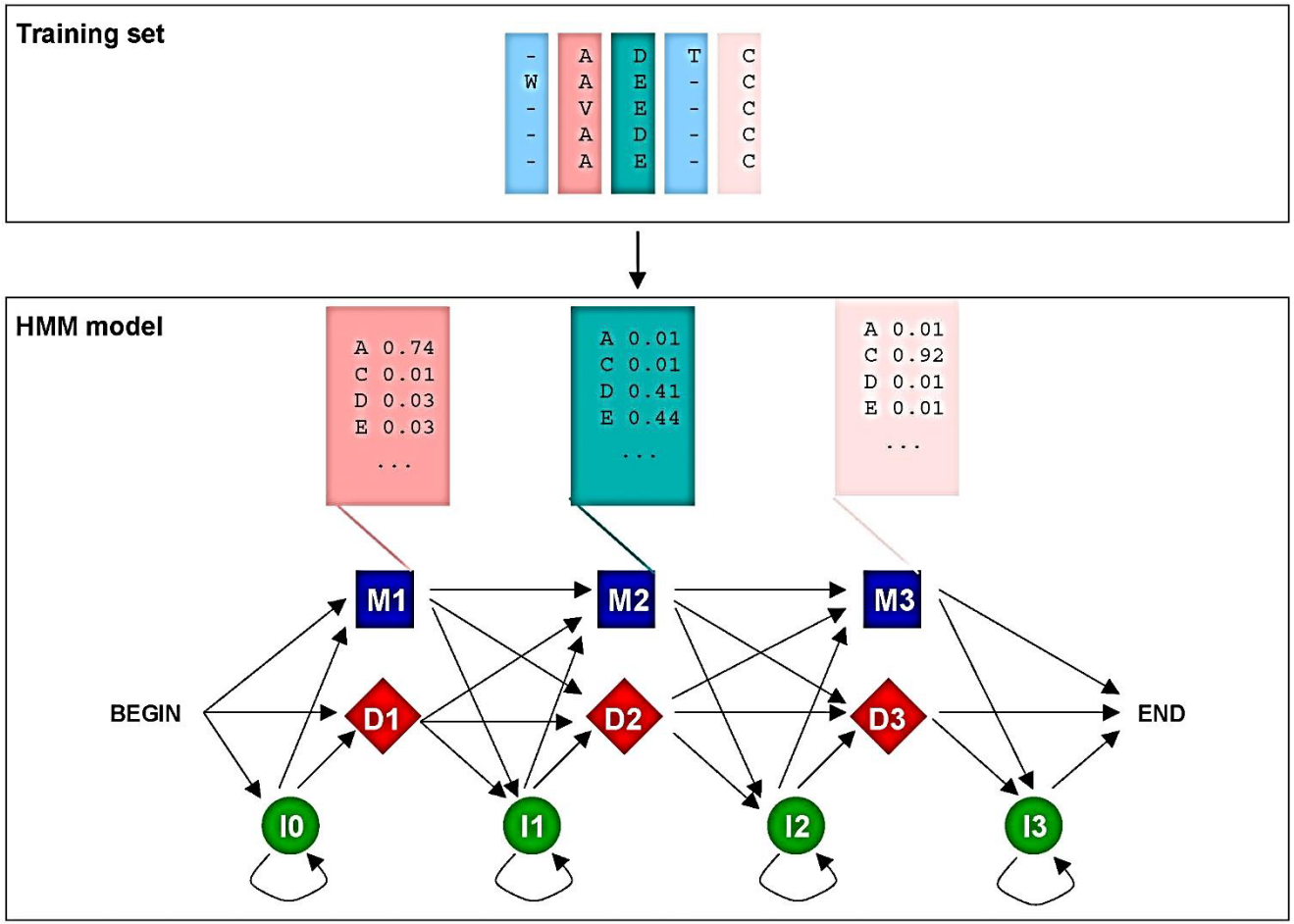
Hidden Markov Model

HMMs are trained from a multiple sequence alignment



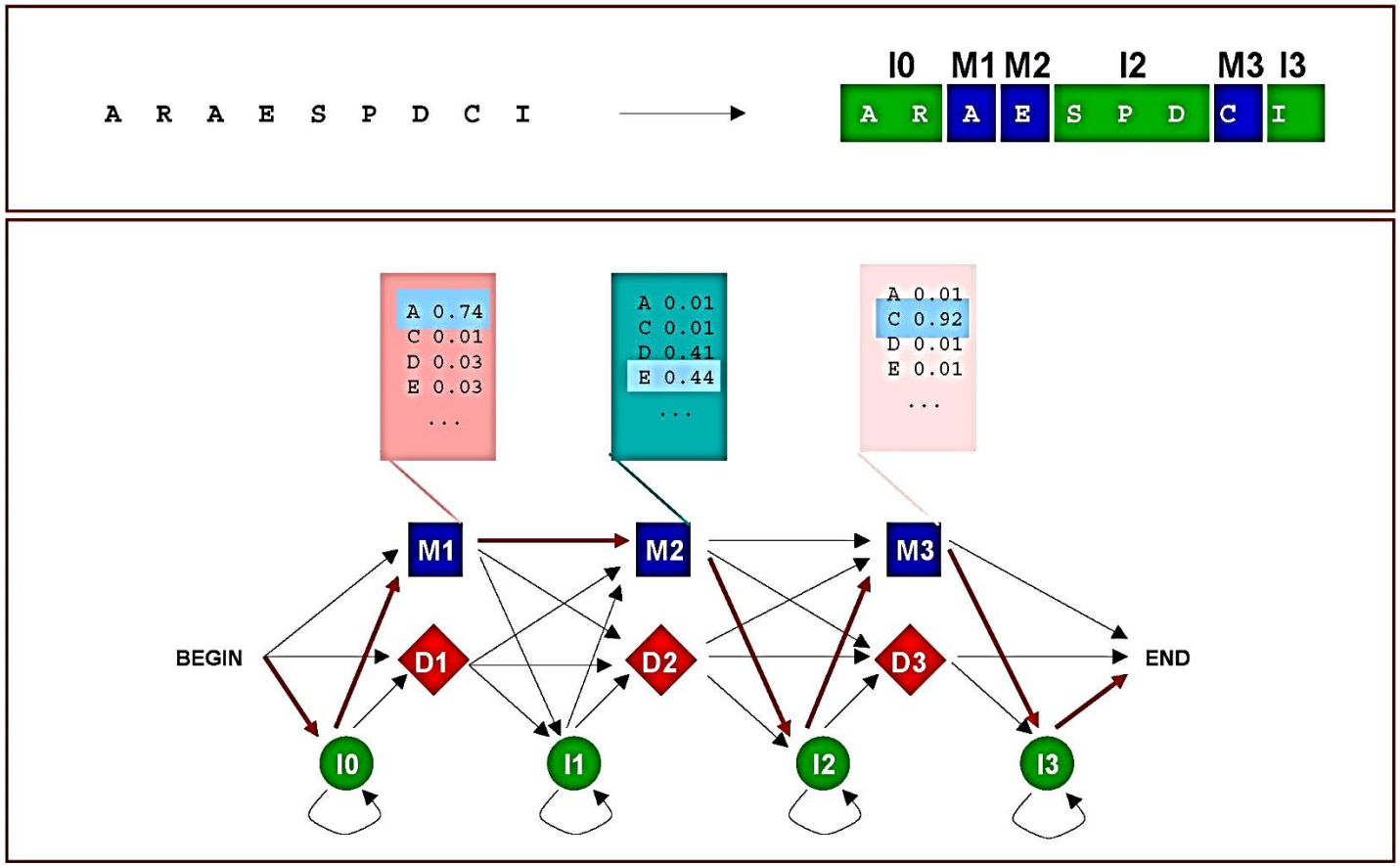


Hidden Markov Model (HMM) is more general than PSSM



Pagni, et al. EMBnet Course 2003

Match a sequence to a model
Application: Function Prediction



Pagni, et al. EMBnet Course 2003

```
>unknown_protein
MALLYRRMSMLLNIILAYIFLCAICVQGSVKQEWAEIGKNVSLECASENEAVAWKLGNQTINKNHTRYKI
RTEPLKSNDDGSENNDSQDFIKYKNVLALLDVNIKDSGNYTCTAQTGQNHSTEFQVRPYLPSKVLQSTPD
RIKRKIKQDVMLYCLIEMYPQNETTNRNLKWLKDGSQFEFLDTFSSISKLNDTHLNFTLEFTEVYKKENG
TYKCTVFDDTGLEITSKEITLFVMEVPQVSIDFAKAVGANKIYLNWTVNDGNDPIQKFFITLQEAGTPTF
TYHKDFINGSHTSYILDHFKPNTTYFLRIVGKNSIGNGQPTQYPQGITTLSYDPIFIPKVETTGSTASTI
TIGWNPPPPDLIDYIQYYELIVSESGEVPKVIEEAIYQQNSRNLPYMFDKLKTATDYEFRVRACSDLTKT
CGPWSENVNGTTMDGVATKPTNLSIQCHHDNVTRGNSIAINWDVPKTPNGKVVSYLIHLLGNPMSTVDRE
MWGPKIRRIDEPHHKTLYESVSPNTNYTVTVSAITRHKKNGEPATGSCLMPVSTPDAIGRTMWSKVNLDS
KYVLKLYLPKISERNGPICCYRLYLVRINNDNKELPDPEKLNIATYQEVHSDNVTRSSAYIAEMISSKYF
RPEIFLGDEKRFSENNDIIRDNDEICRKCLEGTPFLRKPEIIHIPPQGSLSNSDSELPILSEKDNLIKGA
NLTEHALKILESKLRDKRNAVTSDENPILSAVNPNVPLHDSSRDVFDGEIDINSNYTGFLEIIVRDRNNA
LMAYSKYFDIITPATEAEPIQSLNNMDYYLSIGVKAGAVLLGVILVFIVLWVFHHKKTKNELQGEDTLTL
RDSLRALFGRRNHNHGHEITSGNHKGEDAGDIHRLDLENAYKNPHKDTDYGFLDEYEMLDNPESDRTTKN
SDLKENACKNRYP
EQHLEIIVMLTNL
RRQITQYHYLTWK
SVSIYNTVCDLRH
EKLLATADEISKS
QDPLENTIGDFWR
TNCKIDDTLKVTQ
VAMCILVQHLRLE
```

PFAM

a pre-constructed HMM model database

for protein function domain prediction

wellcome trust sanger institute

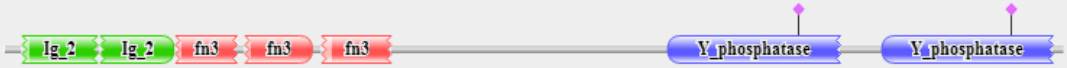
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

Sequence search results

Show the detailed description of this results page.

We found **7** Pfam-A matches to your search sequence (**all** significant). You did not choose to search for Pfam-B matches.



Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
Ig_2	Immunoglobulin domain	Domain	CL0011	24	127	35	126	**11**	**78**	80	27.0	3.5e-06	n/a	Show
Ig_2	Immunoglobulin domain	Domain	CL0011	132	233	135	233	**4**	80	80	19.8	0.00063	n/a	Show
fn3	Fibronectin type III domain	Domain	CL0159	237	321	244	320	**8**	**84**	85	39.3	5.2e-10	n/a	Show
fn3	Fibronectin type III domain	Domain	CL0159	333	425	340	425	**6**	85	85	40.9	1.6e-10	n/a	Show
fn3	Fibronectin type III domain	Domain	CL0159	439	534	452	532	**11**	**83**	85	27.3	2.8e-06	n/a	Show
Y_phosphatase	Protein-tyrosine phosphatase	Domain	CL0031	916	1154	916	1153	1	**234**	235	283.6	9.6e-85	1096,1096	Show
Y_phosphatase	Protein-tyrosine phosphatase	Domain	CL0031	1212	1448	1212	1447	1	**234**	235	211.8	8.5e-63	1390,1390	Show

Comments or questions on the site? Send a mail to **pfam-help@sanger.ac.uk**. Our **cookie policy**.

The Wellcome Trust

http://pfam.sanger.ac.uk/

What is Gene Ontology (GO)
How to describe the function of a gene?

- **Free text description**

Gene ID	Gene description
GRMZM2G002950	Putative leucine-rich repeat receptor-like protein kinase family
GRMZM2G006470	Uncharacterized protein
GRMZM2G014376	Shikimate dehydrogenase; Uncharacterized protein
GRMZM2G015238	Prolyl endopeptidase
GRMZM2G022283	Uncharacterized protein

- **Controlled vocabulary (Gene Ontology)**

What is Gene Ontology (GO)

How to describe the function of a gene?

- **Gene description line**

- **Controlled vocabulary (Gene Ontology)**

Gene ID	GO
GRMZM5G888620	GO:0003674
GRMZM5G888620	GO:0008150
GRMZM5G888620	GO:0008152
GRMZM5G888620	GO:0016757
GRMZM5G888620	GO:0016758
GRMZM2G133073	GO:0003674
GRMZM2G133073	GO:0016746

Three Groups of GO Terms

Molecular Function

id: GO:0004396
name: hexokinase activity

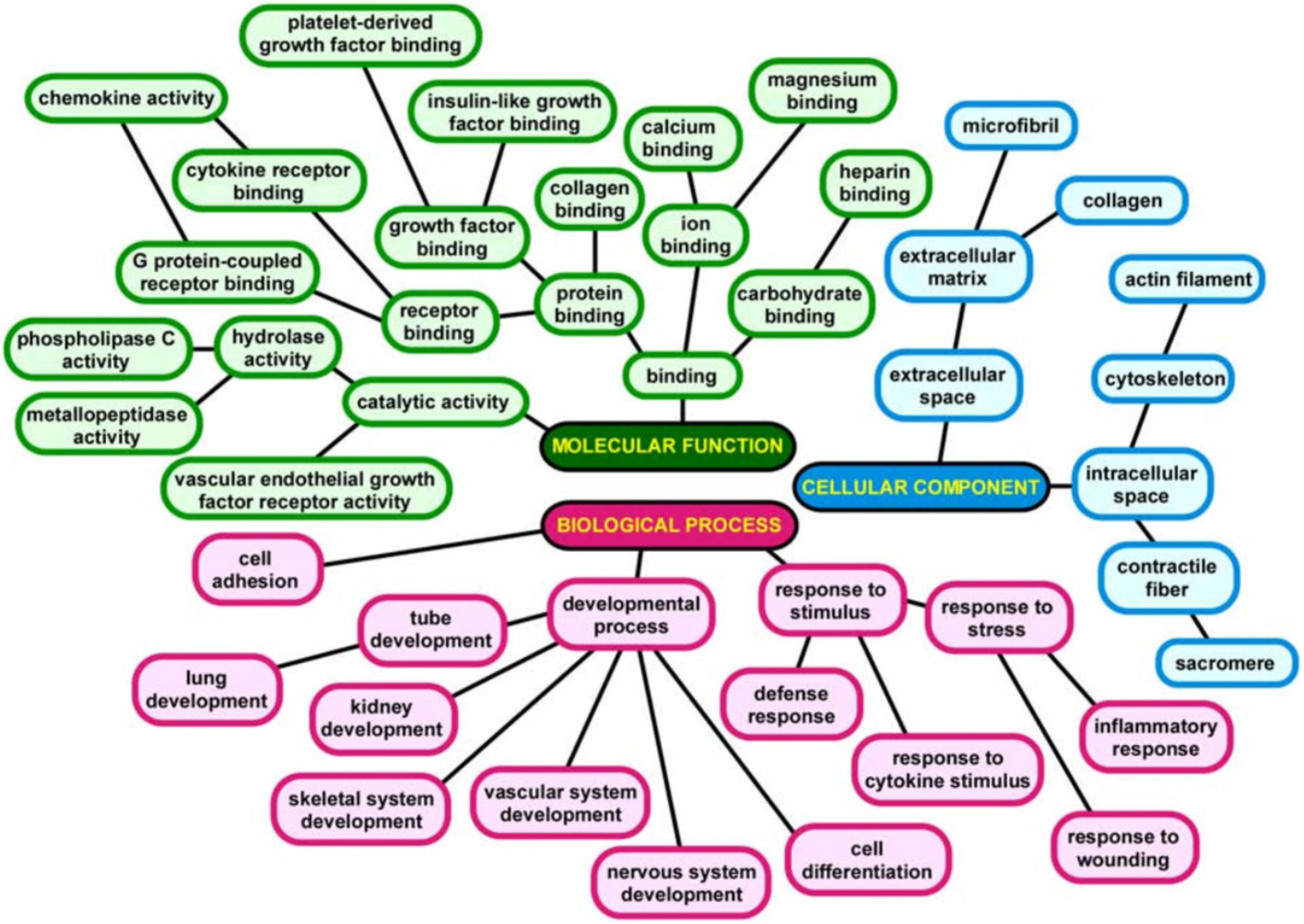
Biological Process

id: GO:0000018
name: regulation of DNA recombination

Cellular Component

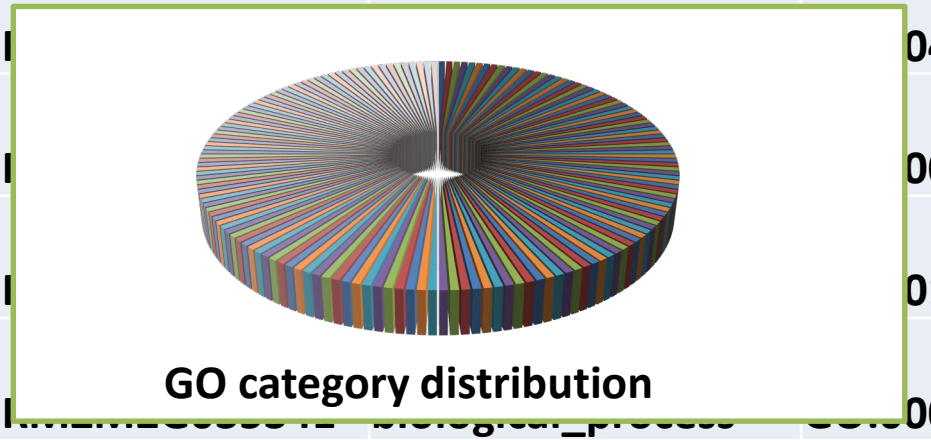
id: GO:0032590
name: dendrite membrane

Hierarchical structure of gene ontology?

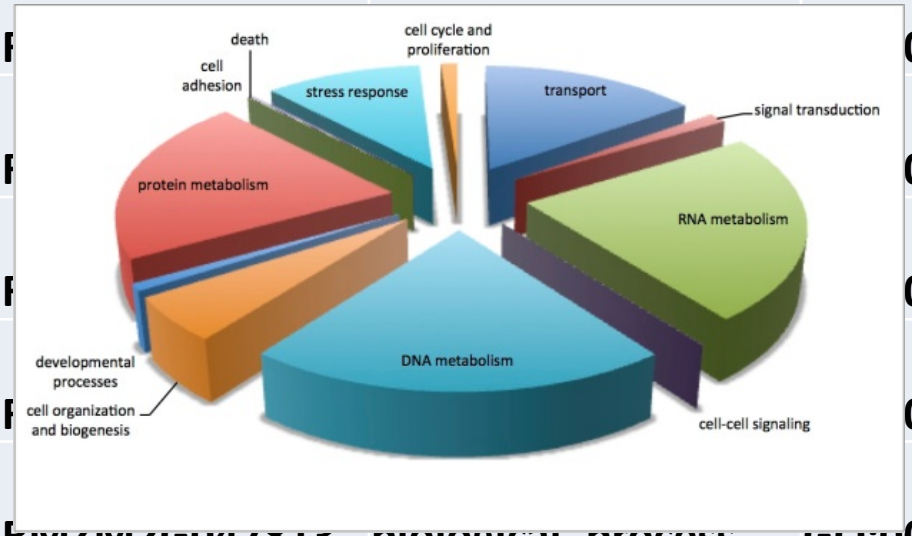


GO SLIM

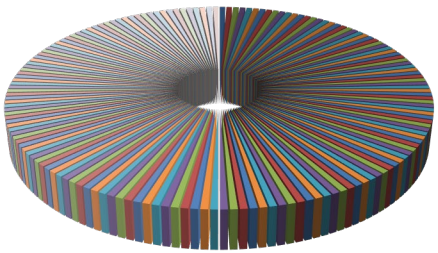
GRMZM2G035341	molecular_function	GO:0008270	zinc ion binding
GI		046872	metal ion binding
GI		005622	intracellular
GI		019005	SCF ubiquitin ligase complex
GI		009733	response to auxin
GI		003677	DNA binding
GI		005634	nucleus
GI		005694	chromosome
GI		006259	DNA metabolic process
GRMZM2G047813	biological_process	GO:0034641	cellular nitrogen compound metabolic process

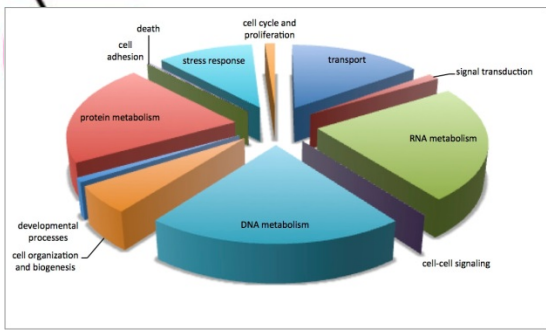


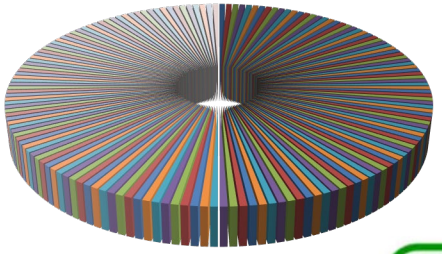
GO category distribution



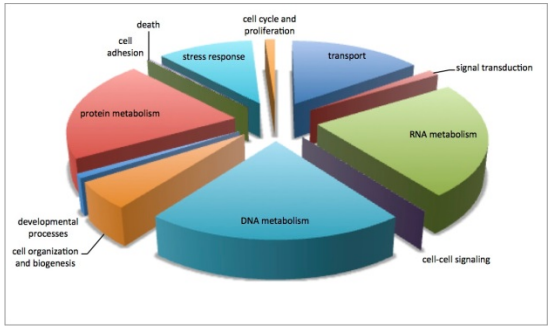
The Necessity for GO Slim

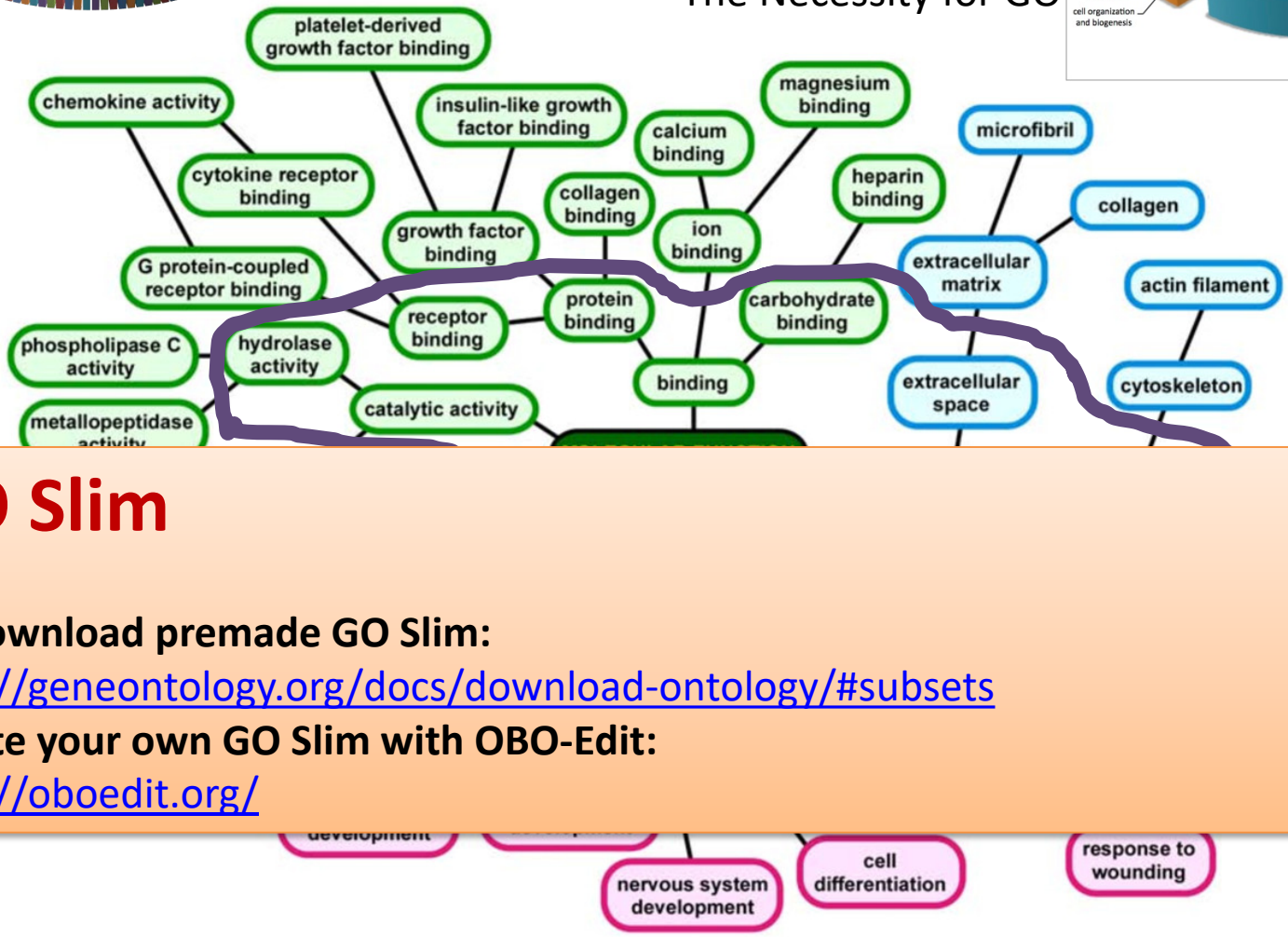






The Necessity for GO





GO Slim

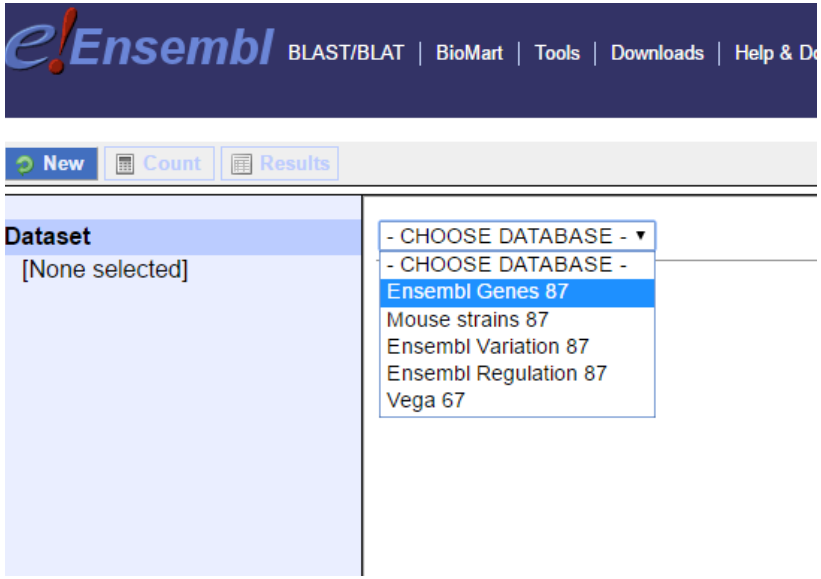
To download premade GO Slim:
http://geneontology.org/docs/download-ontology/#subsets
Create your own GO Slim with OBO-Edit:
http://oboedit.org/

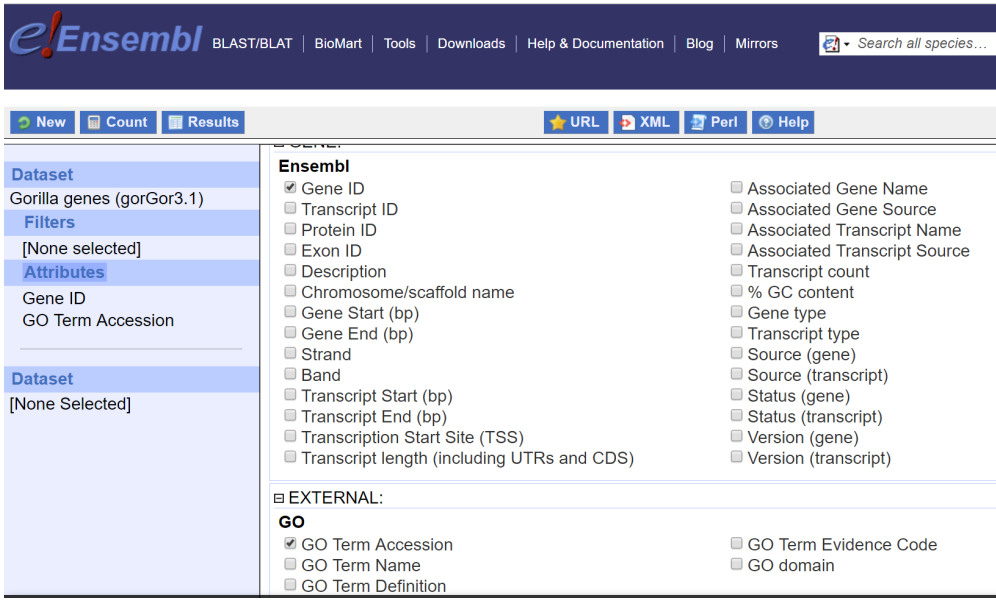
How to get Gene Ontology ?

GRMZM2G035341	molecular_function	GO:0008270	zinc ion binding
GRMZM2G035341	molecular_function	GO:0046872	metal ion binding
GRMZM2G035341	cellular_component	GO:0005622	intracellular
GRMZM2G035341	cellular_component	GO:0019005	SCF ubiquitin ligase complex
GRMZM2G035341	biological_process	GO:0009733	response to auxin
GRMZM2G047813	molecular_function	GO:0003677	DNA binding
GRMZM2G047813	cellular_component	GO:0005634	nucleus
GRMZM2G047813	cellular_component	GO:0005694	chromosome
GRMZM2G047813	biological_process	GO:0006259	DNA metabolic process
GRMZM2G047813	biological_process	GO:0034641	cellular nitrogen compound metabolic process

Model organisms: Ensembl BioMart:

Animal genomes: http://www.ensembl.org
Plant genomes: http://plants.ensembl.org





Non model organism

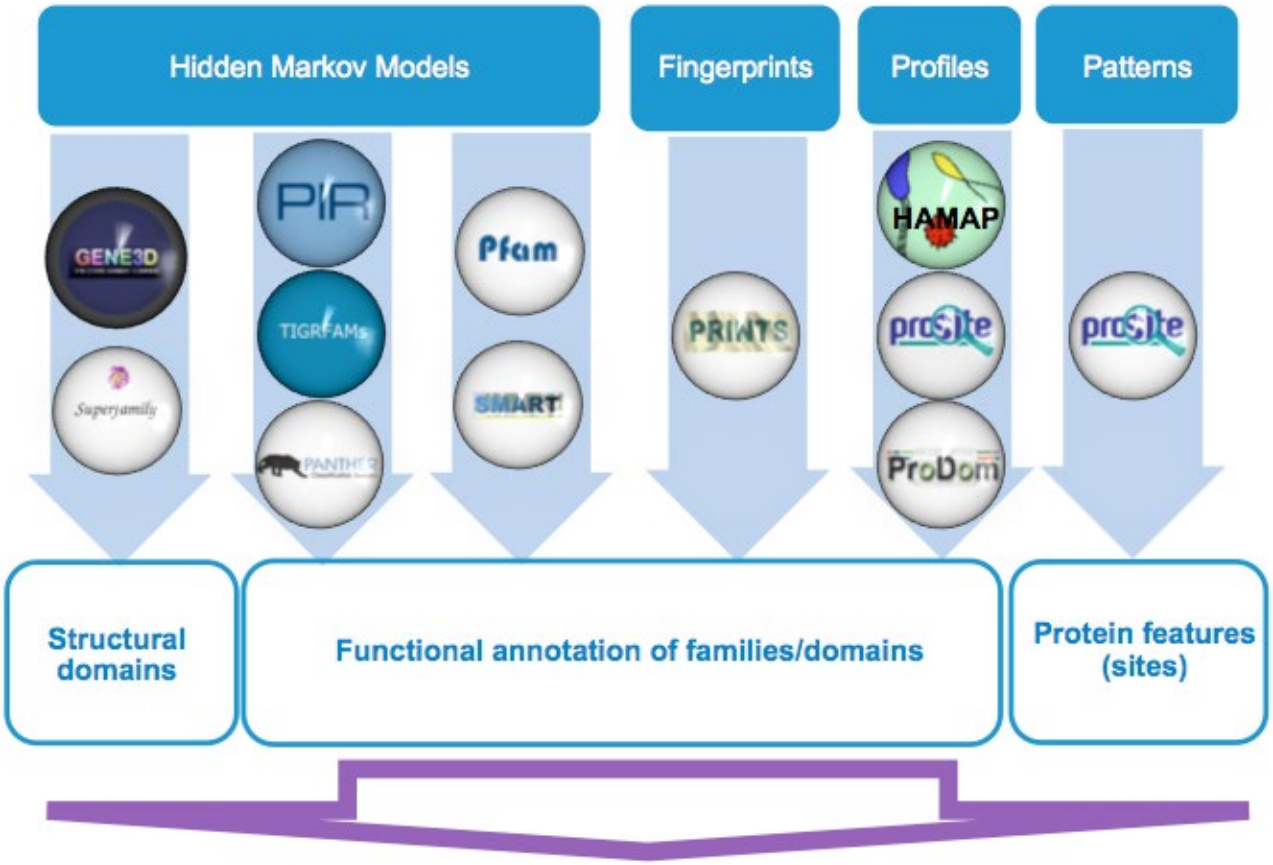
Public tool: InterProScan

Commercial software: BLAST2GO

- **Command line license (on cbsumm10)**
- **GUI license**

Open source gene function annotation software:

InterProScan



Run InterProScan on multiple BioHPC computers

(General or intermediate memory computer)

- Each gene would takes a few minutes. Split the large FASTA into multiple files and run on different computers. Merging the result files.
- Even though it can accept nucleotide, it is strongly recommended to use protein sequences. The BLAST2GO software cannot accept nucleotide sequence based interproscan.

- Version on BioHPC 2016/3. Contact us if you need newer version

```
tar -xf /shared_data/genome_db/interproscan.tar


interproscan/interproscan.sh -b ipsout -f XML -i
annot_exercise.fasta  --goterms --pathways --
iprlookup -t p
```

Specify input data type:
n: DNA; p: protein

Commercial gene function annotation software

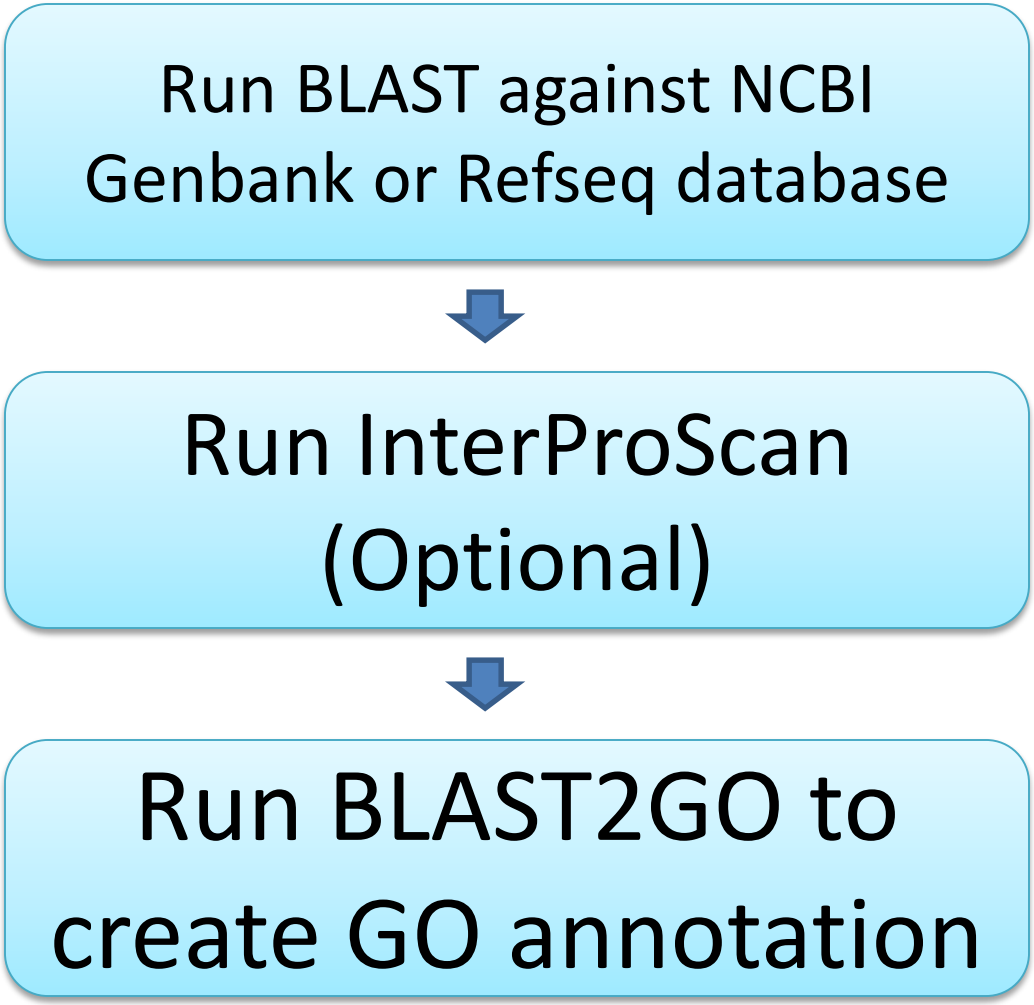
BLAST2GO on BioHPC Lab

Details for **blast2go** (hide)

Name:	blast2go
Version:	DB: Mar.2016; Software: v1.2.1
OS:	Linux
About:	Gene Ontology annotation and function enrichment analysis.
Added:	4/15/2013 5:20:07 PM
Updated:	4/25/2016 12:13:57 PM
Link:	https://www.blast2go.com/
Manual:	https://www.blast2go.com/images/b2g_pdfs/blast2go_cli_manual.pdf
Download:	https://www.blast2go.com/blast2go-pro/b2g-register-basic
Notes:	

```
################################################

####   Run BLAST on any BioHPC computer   ######
################################################
#you can run blast on any of the biohpc computers, adjust the num_threads based on computer you are
using:general machine: 8; medium memory:24; large memory: 64
# you have an option to use swissprot, refseq or nr for blast database. In most cases swissprot is fast and good
enough. However, if a large percentage of your genes have no blast hits to swissprot, you can try refseq. The nr
database is too big, the blast run would take very long time.
#replace test.fa with your own fasta file. Make sure you are using the right blast software (blastx or blastp). To
save time, it is preferrable to use blastp on protein queries. We recommend to to use TransDecoder software to
identify protein coding sequences from cDNA sequences.
#replace swissprot with nr if you want to blast against nr database
#adjust the blast parameters in blast command
# BLAST might take hours to finish. With nr, it might take days

#commands (use swissprot as an example. To use refseq, replate swissprot with refseq_protein)

cd /workdir/myUserName
cp /shared_data/genome_db/BLAST_NCBI/swissprot* ./

blastp -num_threads 24 -query test.fa -db swissprot -out blastresults.xml  -max_target_seqs 20 -evalue 1e-5 -
outfmt 5  -culling_limit 10 >& blastlogfile &

After this step, the blast result file blastresults.xml will be created. Copy this file to your home directory.



################################################
####   Optional: Run Interproscan on any BioHPC computer   ######
################################################
#you can run interproscan on any of the biohpc computers,

Follow the instruction to run interproscan on BioHPC lab
```

BLAST2GO, a pipeline for function annotation



Which BLAST Database to Use
* use Protein Database

- Swissprot: fast
- NCBI NR: could take weeks
- NCBI Refseq Protein: a good compromise

Run BLAST on any BioHPC computer

- **Use protein queries if possible**

***** set -num_threads according to the computer you are using.**

```
cp /shared_data/genome_db/BLAST_NCBI/refseq_protein* ./


blastx -num_threads 8 \
-query annot_exercise.fasta  \
-db swissprot \
-out blastresults.xml  \
-max_target_seqs 20 \
-evalue 1e-5 -outfmt 5 \
-culling_limit 10 >& logfile &
```

Blast database is available on BioHPC lab

Use blastx if query is DNA sequence
Use blastp if query is proteins

Specify output format 5 (XML format)

Culling_limit restrict maximum target for each site of the query

Run BLAST2GO on cbsumm10

```
/usr/local/blast2go/blast2go_cli.run \
-properties annotation.prop \
-useobo go.obo \
-loadblast blastresults.xml \
-loadips50 ipsout.xml \
-mapping -annotation -annex -statistics all \
-saveb2g myresult -saveannot myresult -
savereport myresult -tempfolder ./ \
>& annotatelogfile &
```

Default works for most cases. Modify the property file if needed.

Output from BLAST2GO

myresult.b2g: A binary project file that can be opened in BLAST2GO software

myresult.annot: a tab-delimited text file with GO annotation for each gene

myresult.pdf: statistic report of the annotation

Function enrichment analysis

ORA

Over Representation Analysis

- Identify DE genes;
- Assess GO terms over-represented in the DE gene list;

GSEA

Gene Set Enrichment Analysis

- Rank genes on DE level;
- Evaluate gene sets over-represented at either the top or bottom of the list

Software

- **Free:**
  - DAVID (online tool http://david.abcc.ncifcrf.gov/)
  - topGO (command line tool)
  - GSEA (Win/Mac/Linux software)

- **Commercial:**
  - IPA (Ingenuity Pathway Analysis)

(Cornell license information http://www.biotech.cornell.edu/node/137)

ORA - Over Representation Analysis

	Total Genes	DE genes
In P53 Pathway	40	3 -1
Not in P53 Pathway	29960	297

Using Fisher's Exact Test to identify over represented genes in a pathway or function category

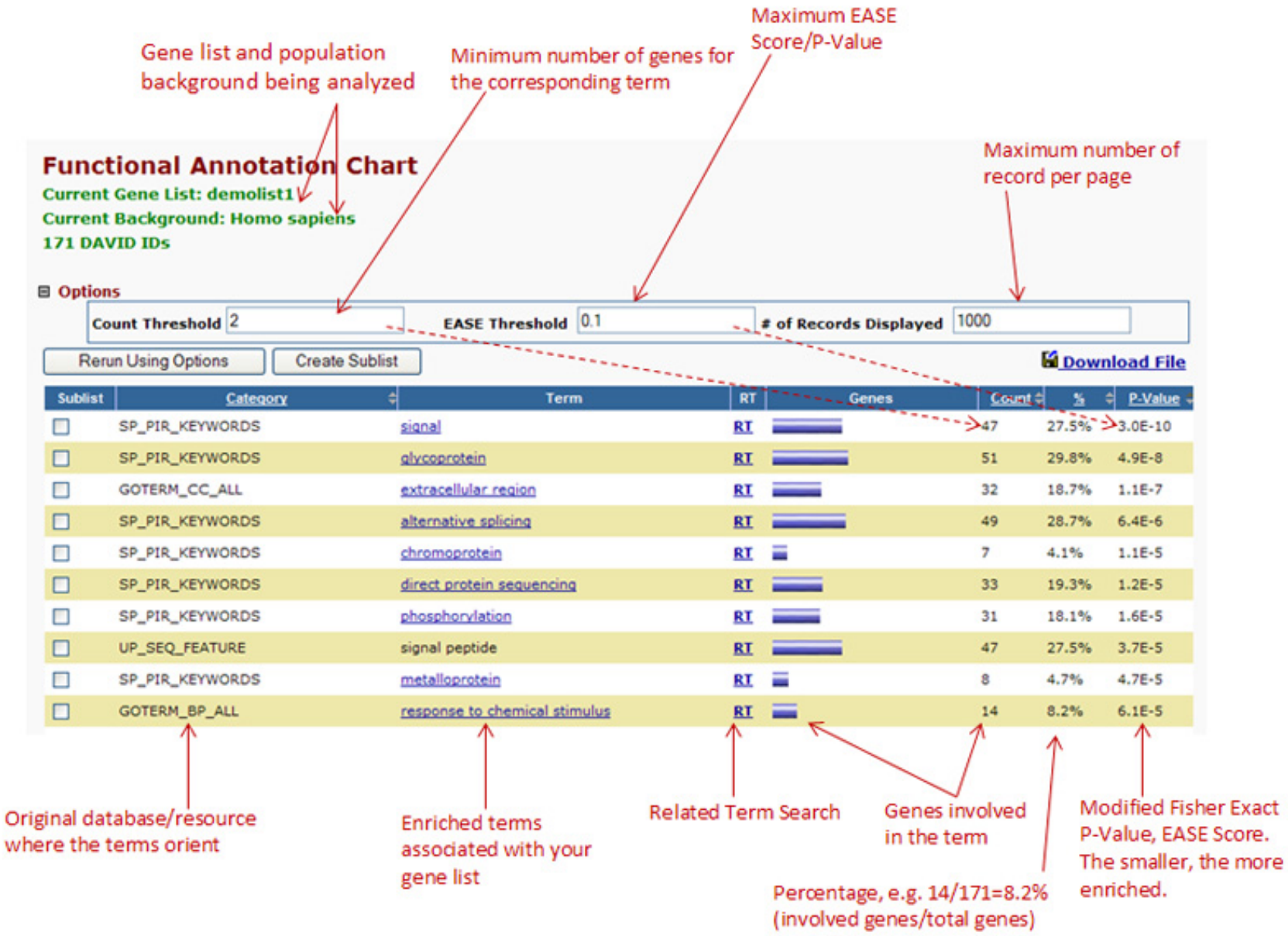
Standard Fisher's exact test: P value= 0.008

EASE Score (in red): P value=0.06

http://david.abcc.ncifcrf.gov/content.jsp?file=functional_annotation.html

Online tools

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



If you work on a non-model organism.

- **Option 1: "Humanized" your gene list**

  Convert your gene list to human orthologs using Ensembl BioMart.

- **Option 2: Use custom GO annotation file with topGO**

```
gene1    GO:0005488, GO:0003774, GO:0001539, GO:0006935, GO:0009288
gene2    GO:0005634, GO:0030528, GO:0006355,
gene3    GO:0016787, GO:0017057, GO:0005975, GO:0005783, GO:0005792
gene4    GO:0043565, GO:0000122, GO:0003700, GO:0005634
gene5    GO:0004803, GO:0005634, GO:0008270, GO:0003677
gene6    GO:0015031, GO:0005794, GO:0016020, GO:0017119, GO:0000139
```

gene | tab | List of GO ids

Run topGo

```
Rscript topGO.r  go.annot refset testset 0.05 BP myBP
```

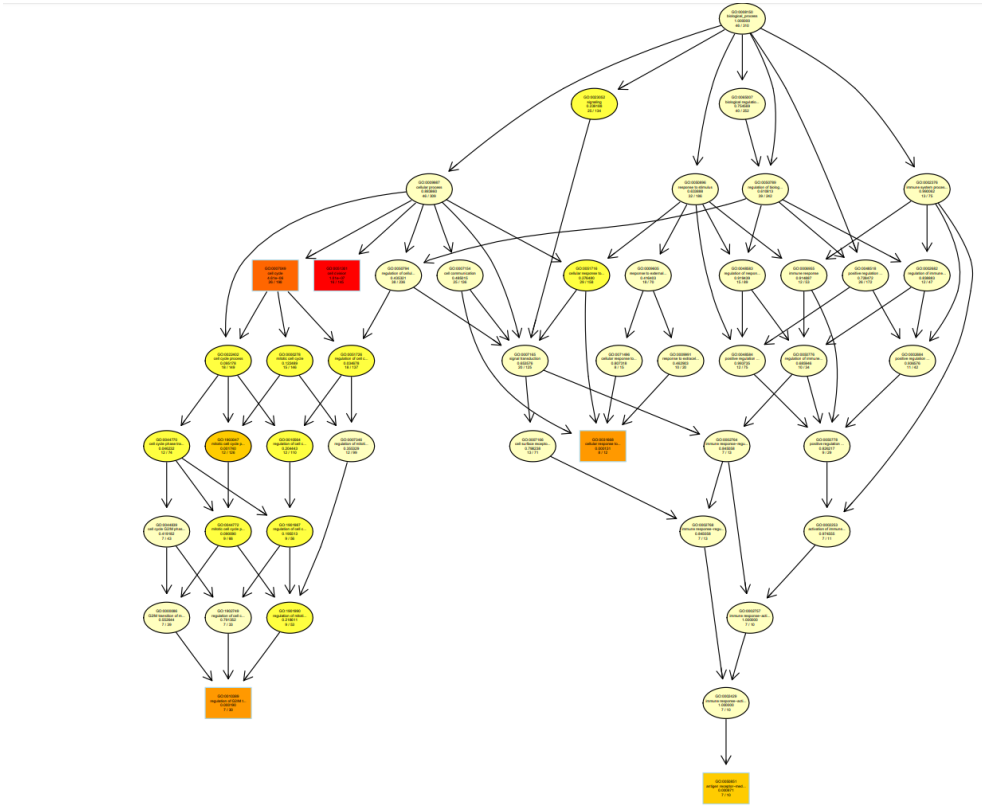
Required input files

- **go.annot**: Go annotation file
- **Refset**: Reference gene sets (all expressed gene list)
- **Testset**: Test gene set (e.g. DE gene list)

- 0.05: P-value cutoff
- BP: test Biology Process GO terms
- myBP: output file

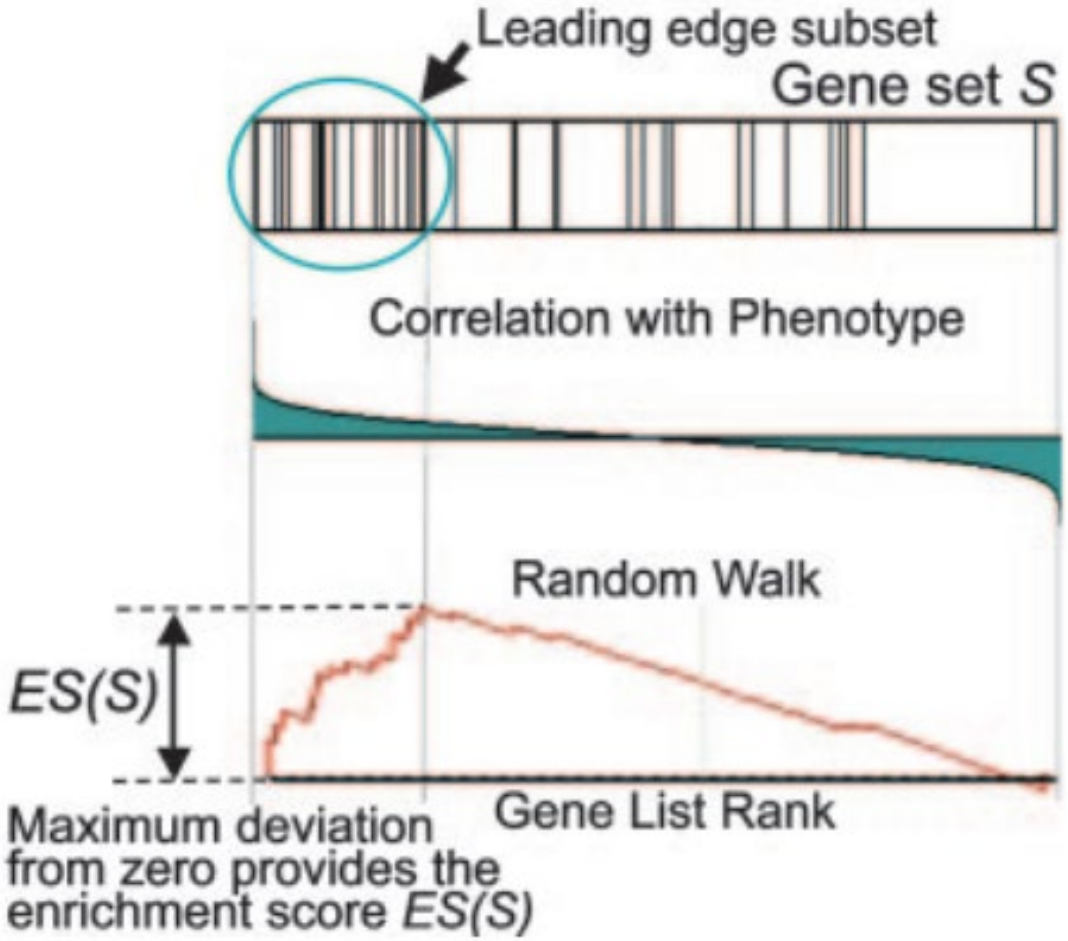
FISHER & Kolmogorov Smirnov (KS) Test

	GO.ID	Term	Annotated	Significant	Expected	Rank in classicFisher	classicFisher	classicKS	elimKS
1	GO:0051301	cell division	145	16	21.52	942	0.97	1.0e-07	1.0e-07
2	GO:0007049	cell cycle	198	26	29.38	857	0.90	3.8e-11	4.6e-06
3	GO:0031668	cellular response to extracellular stimu...	12	8	1.78	1	4.2e-05	0.00013	0.00013
4	GO:0010389	regulation of G2/M transition of mitotic...	30	7	4.45	246	0.14	0.00019	0.00019
5	GO:0050851	antigen receptor-mediated signaling path...	10	7	1.48	2	8.8e-05	0.00087	0.00087
6	GO:0051054	positive regulation of DNA metabolic pro...	24	6	3.56	233	0.13	0.00147	0.00147
7	GO:1903047	mitotic cell cycle process	126	12	18.70	958	0.99	2.5e-05	0.00174
8	GO:0051276	chromosome organization	87	7	12.91	957	0.99	0.00245	0.00245
9	GO:0000226	microtubule cytoskeleton organization	66	8	9.79	739	0.81	0.00377	0.00377
10	GO:0007292	female gamete generation	13	2	1.93	557	0.60	0.00422	0.00422



GSEA - Gene Set Enrichment Analysis

- Rank genes based on shrunken Log_2(Fold_Change) *
- ES score of each gene set (e.g. diabetes related genes)



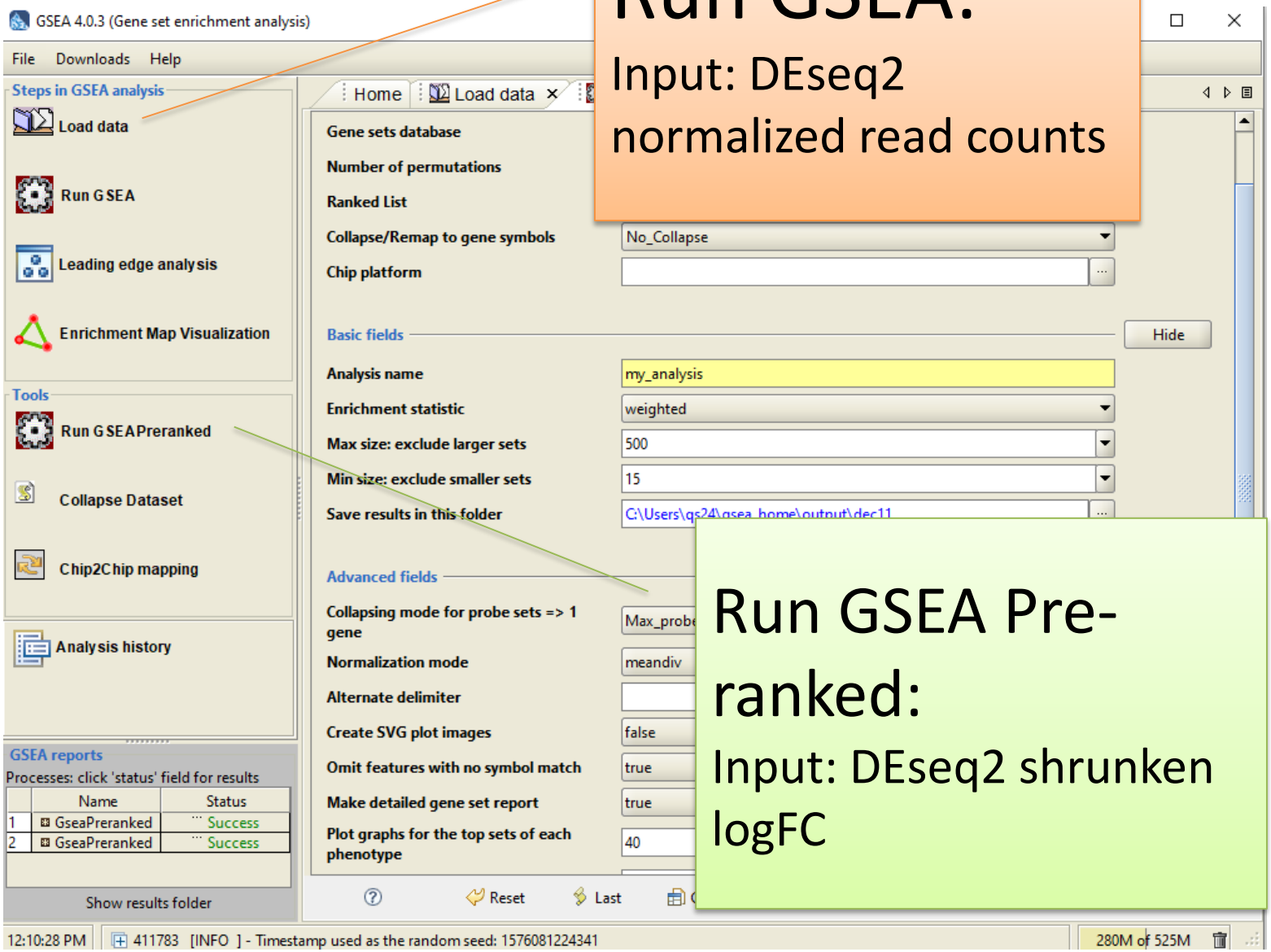
Two alternative ways to analyze RNA-seq data with GSEA

Run GSEA:

Input: DEseq2 normalized read counts

Run GSEA Pre-ranked:

Input: DEseq2 shrunken logFC



GSEA
Input files

.rnk file
- ranked gene list

Gene	log2(ratio)
YDL248W	0.446508
YDL243C	0.285379
YDL241W	2.006822
YDL240W	-0.87753
YDL239C	-0.00886
YDL238C	0.837298
YDL237W	-0.14496
YDL236W	0.417735
YDL235C	-0.31365
YDL234C	0.832606

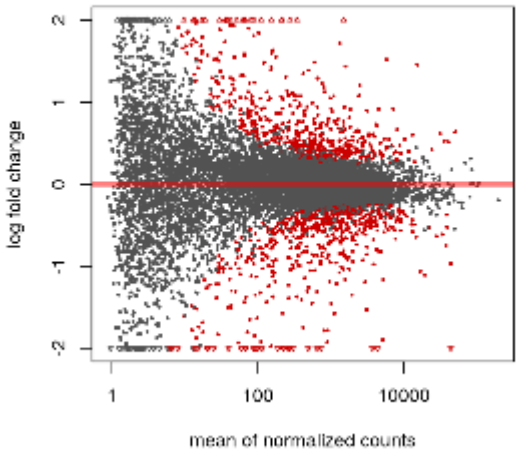
.gmt file
- gene sets

90S_preribosome	http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0030686	YBL004W	YBR247C	YCL031C	YCR057C	YDL148C	YDL213C
AP_type_membrane_coat_adaptor	http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0030119	YBL037W	YBR288C	YDR358W	YGR261C	YHL019C	YHR108W
ATPase_complex	http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:1904949	YAL011W	YAR007C	YBL006C	YBL035C	YBR087W	
COPII_coated_ER_to_Golgi	http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0030134	YAL007C	YAL042W	YAR002C-A	YAR033W	YBR210W	YCL001W
COPII_coated_vesicle_budding	http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0090114	YCR067C	YDL195W	YFL038C	YGR058W	YHR098C	YIL109C
COPI_coated_vesicle	http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0030137	YAR033W	YCL001W	YDL145C	YDR238C		
DASH_complex	http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0042729	YBR233W-A	YDR016C	YDR201W	YDR320C-A	YGL061C	YGR113W
RNA_polymerase_II_specific	http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0001228	YAL051W	YBL005W	YBR033W	YBR083W	YBR297W	YCR018C

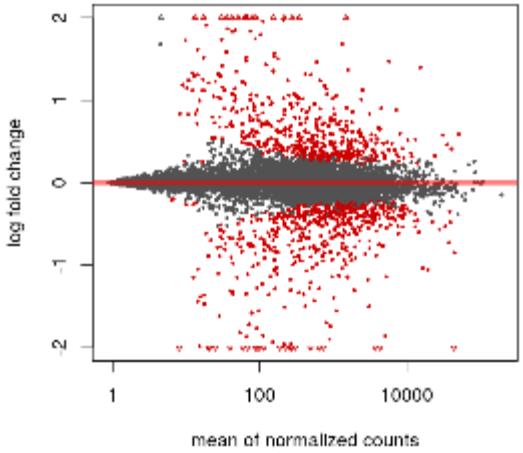
Use shrunken logFC from DESeq2

To shrink the log(Fold-Change) of genes with high noise

MA Plots



Before shrinkage

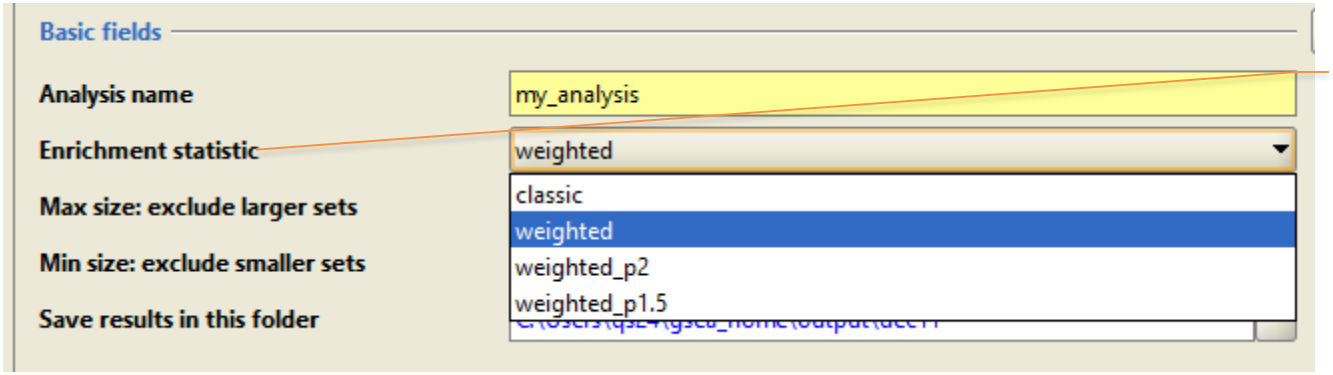


After shrinkage

DESeq2 command for shrink logFC

```
resLFC <- lfcShrink(dds,
coef="condition_treated_vs_untreated", type="apeglm")
```

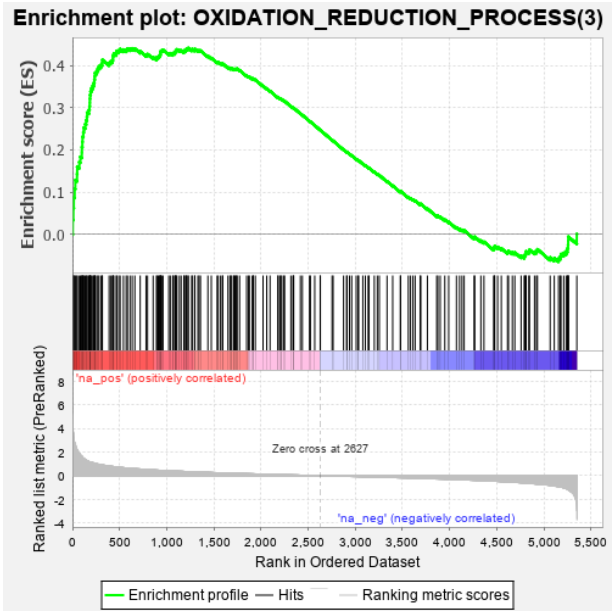
Enrichment statistics



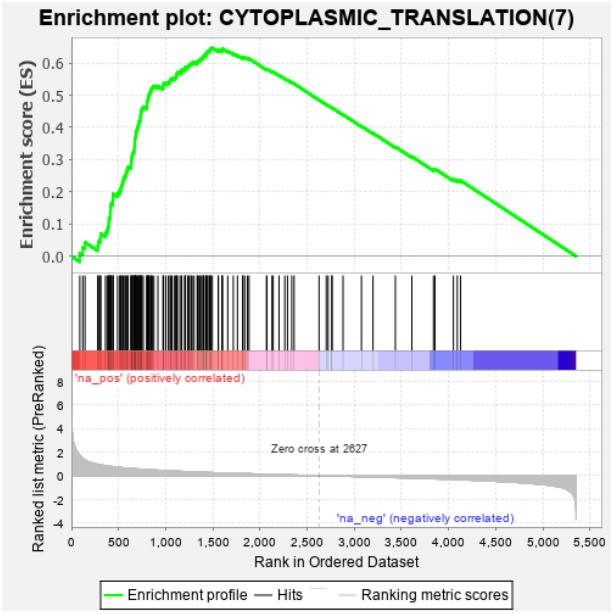
Weighted P-value:
Default: 1

Higher value would enhance the weight of fold change in ES calculation.

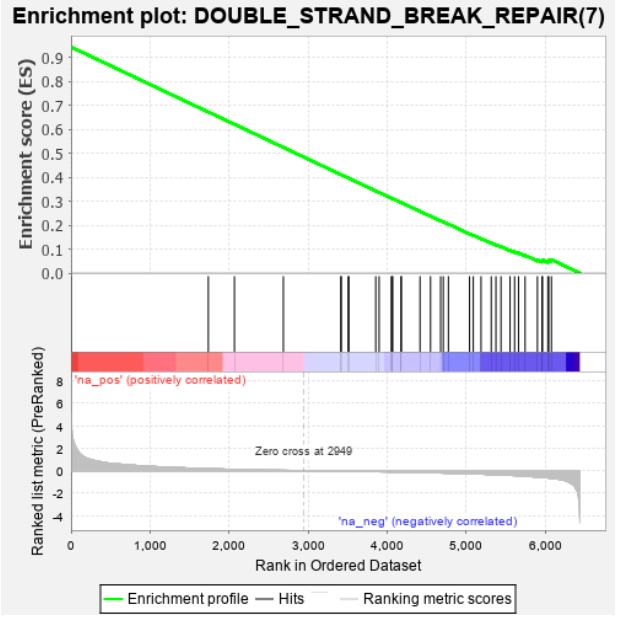
Top hit in ORA



Top hit in GSEA (p=1)



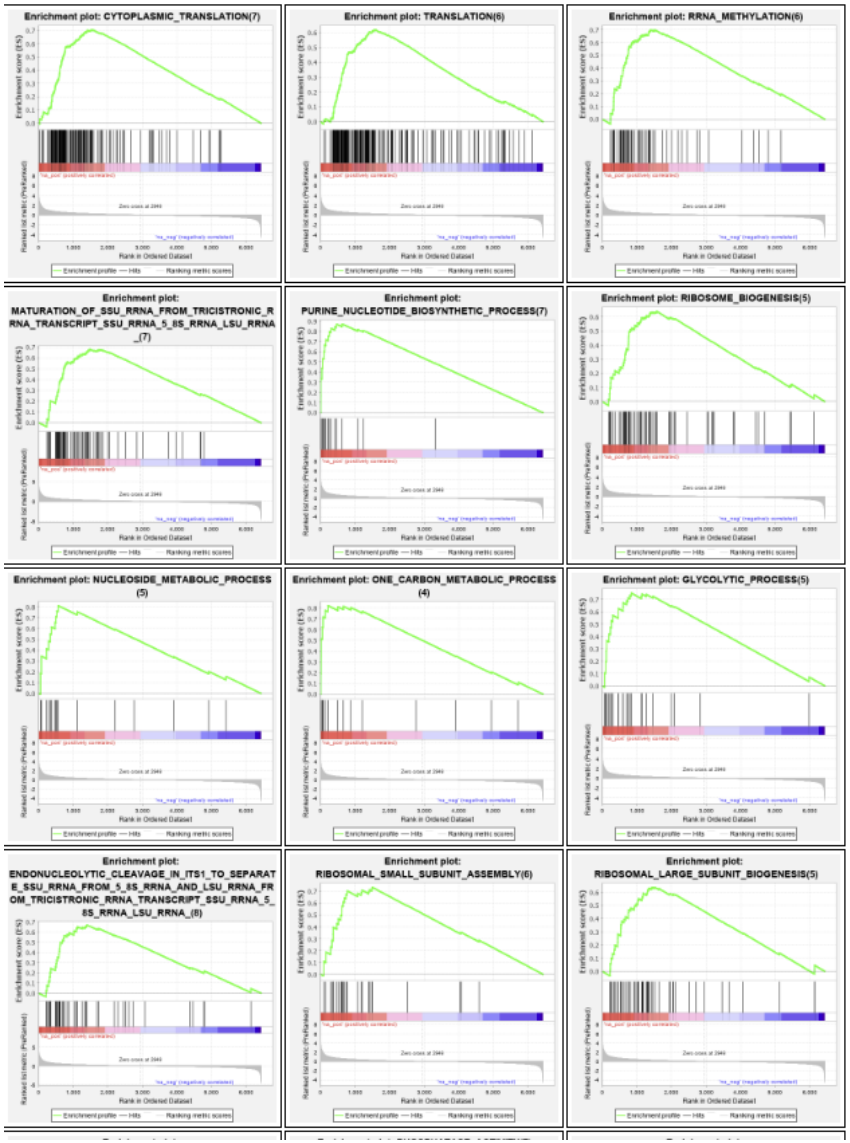
Top hit in GSEA (p=2)



Snap shots of top 12 gene sets with p=1 and p=2

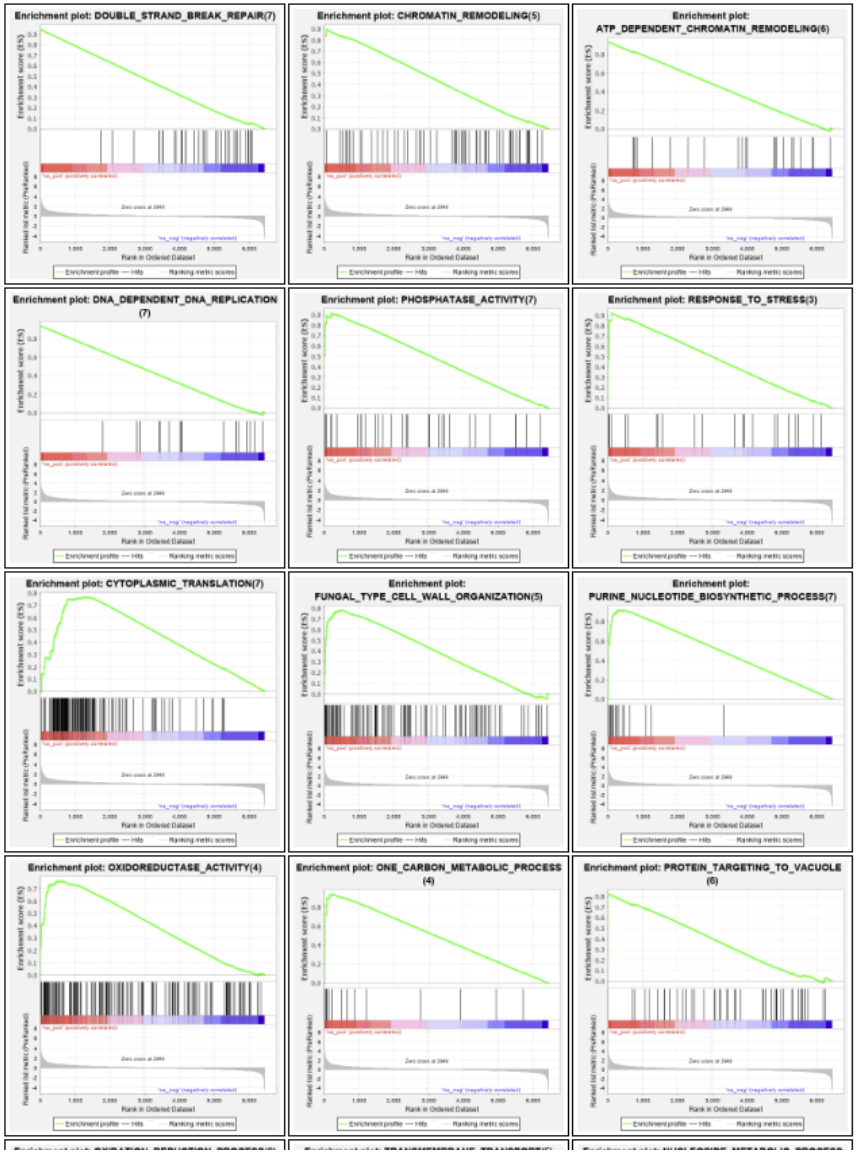
Weighted

Table: Snapshot of enrichment results



Weighted p=2

Table: Snapshot of enrichment results



GSEA Output

Enrichment in phenotype: na

- 199 / 486 gene sets are upregulated in phenotype **na_pos**
- 41 gene sets are significant at FDR < 25%
- 33 gene sets are significantly enriched at nominal pvalue < 1%
- 41 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

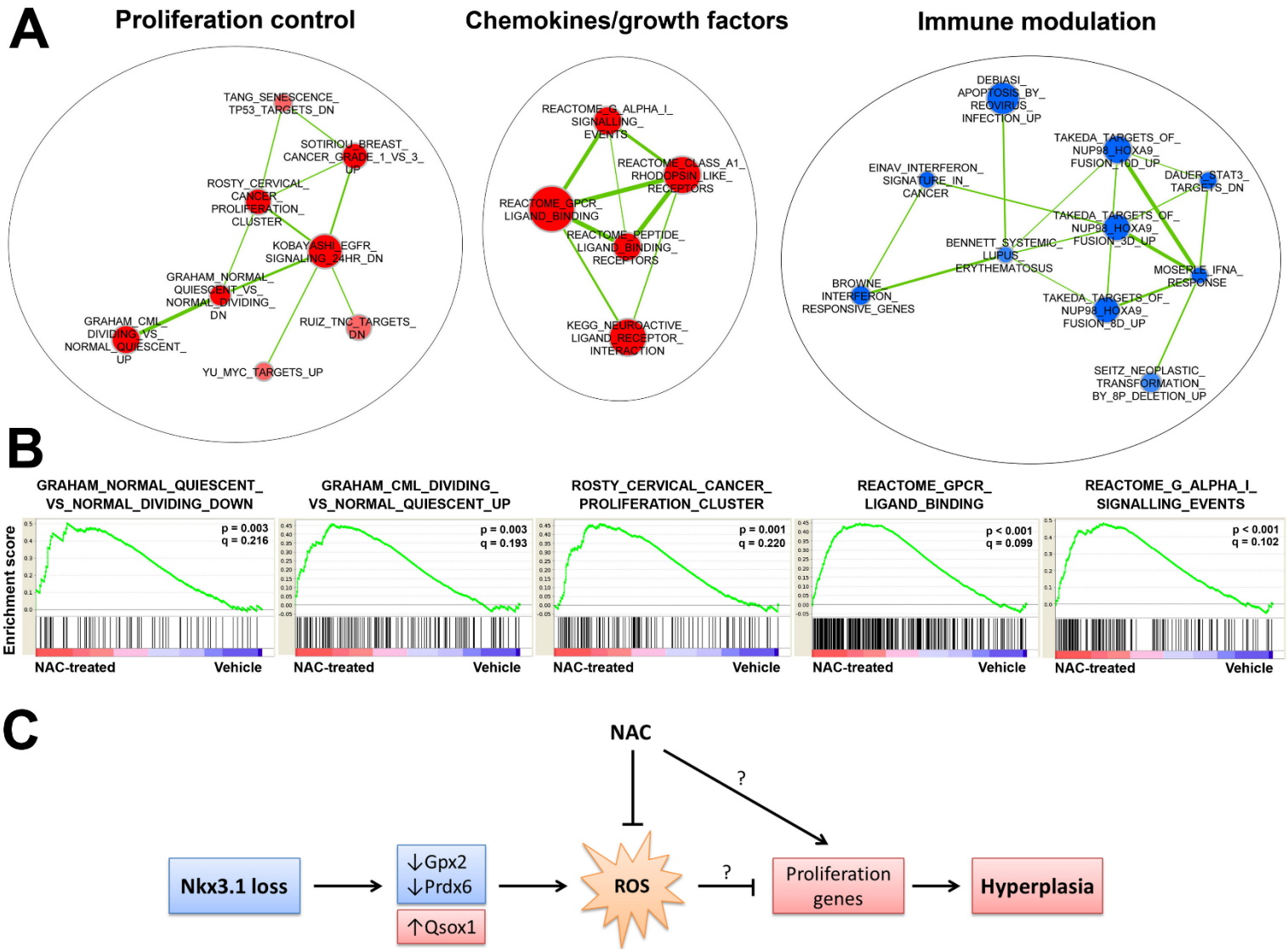
Enrichment in phenotype: na

- 287 / 486 gene sets are upregulated in phenotype **na_neg**
- 70 gene sets are significantly enriched at FDR < 25%
- 55 gene sets are significantly enriched at nominal pvalue < 1%
- 76 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Enriched gene sets from GSEA

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val
1	CYTOPLASMIC_TRANSLATION(7)	Details ...	151	0.70	2.48	0.000	0.000	0.000
2	TRANSLATION(6)	Details ...	185	0.62	2.22	0.000	0.000	0.000
3	RRNA_METHYLATION(6)	Details ...	57	0.70	2.14	0.000	0.000	0.000
4	MATURATION_OF_SSU_RRNA_FROM_TRICISTRONIC_RRNA_TRANSCRIPT_SSU_RRNA_5_8S_RRNA_LSU_RRNA_(7)	Details ...	70	0.68	2.13	0.000	0.000	0.000
5	PURINE_NUCLEOTIDE_BIOSYNTHETIC_PROCESS(7)	Details ...	16	0.87	2.09	0.000	0.000	0.000
6	RIBOSOME_BIOGENESIS(5)	Details ...	64	0.64	2.00	0.000	0.002	0.017
7	ENDONUCLEOLYTIC_CLEAVAGE_IN_ITS1_TO_SEPARATE_SSU_RRNA_FROM_5_8S_RRNA_AND_LSU_RRNA_FROM_TRICISTRONIC_RRNA_TRANSCRIPT_SSU_RRNA_5_8S_RRNA_LSU_RRNA_(8)	Details ...	43	0.67	1.94	0.000	0.005	0.043
8	RIBOSOMAL_LARGE_SUBUNIT_ASSEMBLY(6)	Details ...	40	0.67	1.94	0.000	0.005	0.044
9	RIBOSOMAL_SMALL_SUBUNIT_ASSEMBLY(6)	Details ...	26	0.74	1.94	0.000	0.004	0.045
10	GLYCOLYTIC_PROCESS(5)	Details ...	24	0.75	1.93	0.002	0.005	0.057
11	NUCLEOSIDE_METABOLIC_PROCESS(5)	Details ...	16	0.81	1.92	0.002	0.006	0.073
12	ONE_CARBON_METABOLIC_PROCESS(4)	Details ...	15	0.82	1.91	0.000	0.006	0.079
13	PHOSPHATASE_ACTIVITY(7)	Details ...	31	0.70	1.91	0.000	0.006	0.086
14	RIBOSOMAL_LARGE_SUBUNIT_BIOGENESIS(5)	Details ...	52	0.64	1.90	0.000	0.006	0.088

Get network representation of enriched gene sets



https://journals.plos.org/plosone/article/figure?id=10.1371/journal.pone.0046792.g007