Genome Annotation - 2

Qi Sun

Bioinformatics Facility

Cornell University

Output from Maker

GFF file:

- Annotated gene, transcripts, and CDS

FASTA file:

- Predicted transcript sequences
- Predicted protein sequences

Use gffread to convert gff3 file to transcript or protein sequence fasta file:

```
gffread -g genome.fa -y protein.fa -w transcript.fa annotation.gff3
```

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

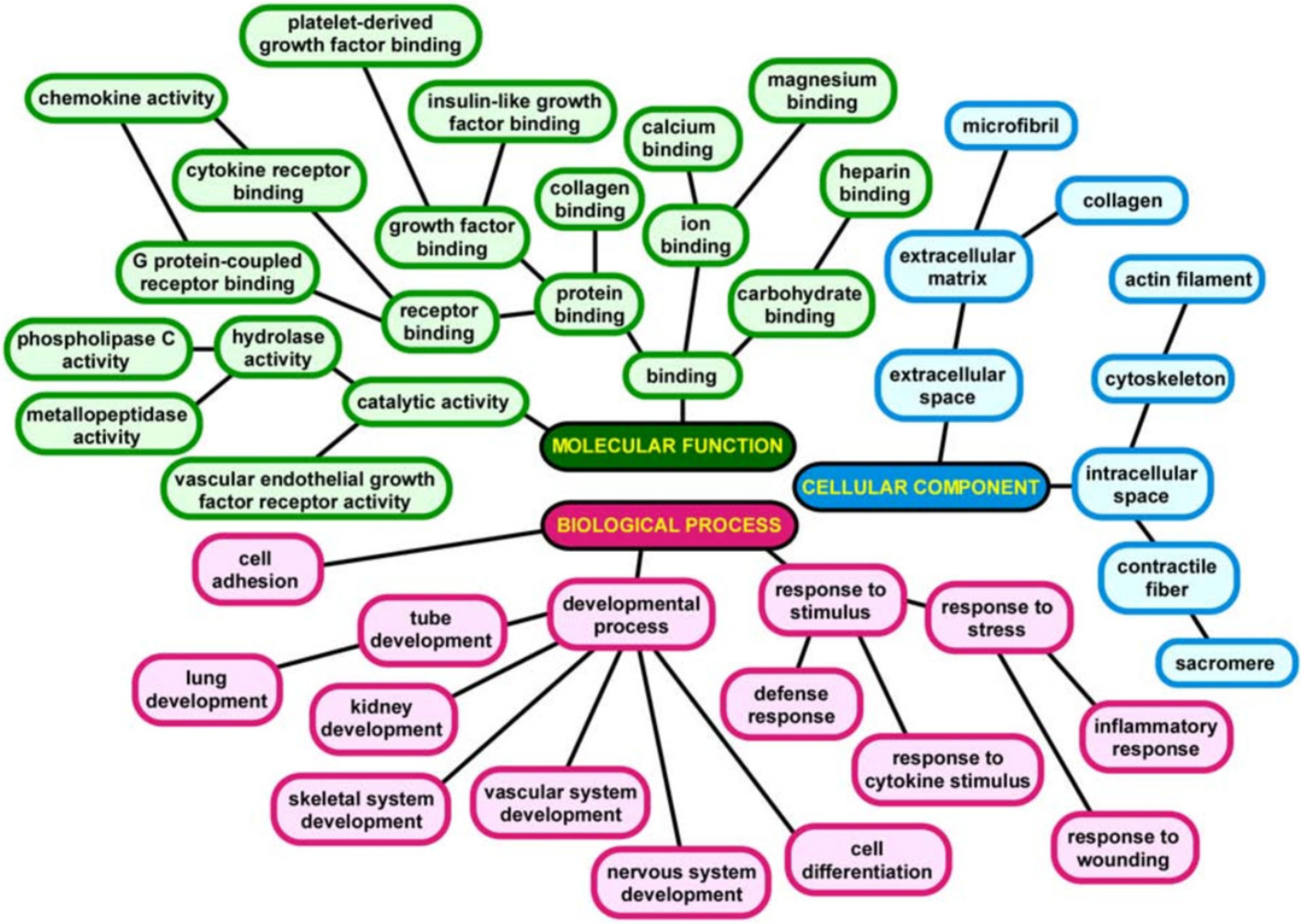
Biological Process

id: GO:0000018
name: regulation of DNA recombination

Cellular Component

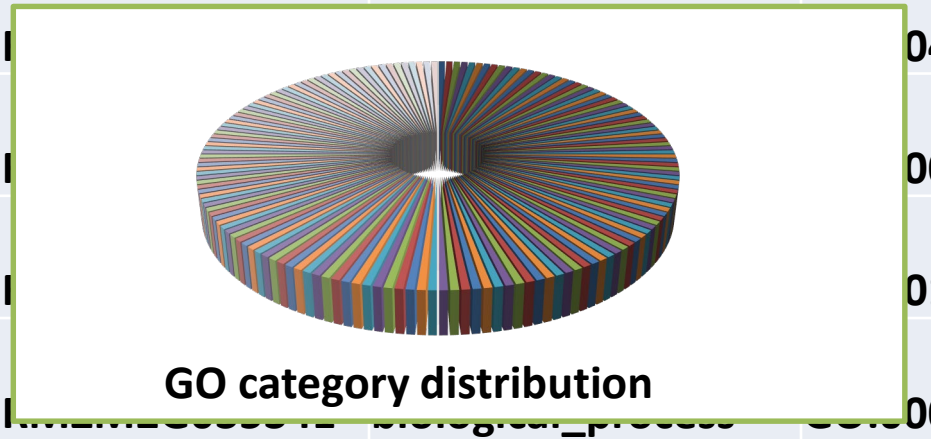
id: GO:0032590
name: dendrite membrane

Hierarchical structure of gene ontology?

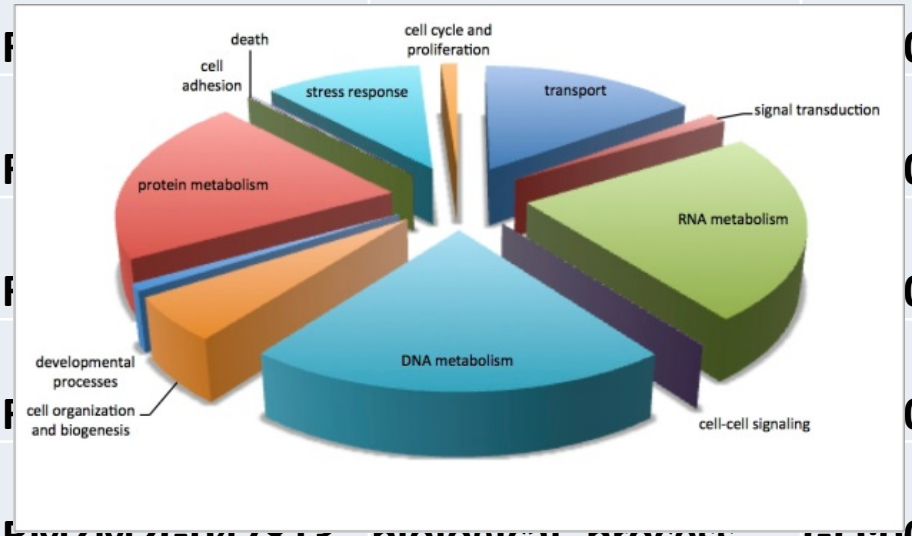


GO SLIM

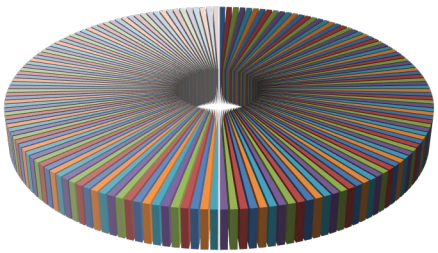
GRMZM2G035341	molecular_function	GO:0008270	zinc ion binding
GI		046872	metal ion binding
GI		005622	intracellular
GI		019005	SCF ubiquitin ligase complex
GI		009733	response to auxin
GI		003677	DNA binding
GI		005634	nucleus
GI		005694	chromosome
GI		006259	DNA metabolic process
GRMZM2G047813	biological_process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

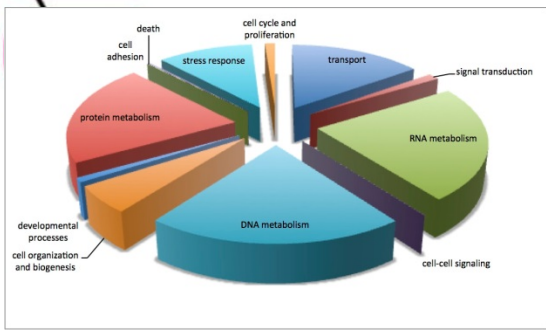


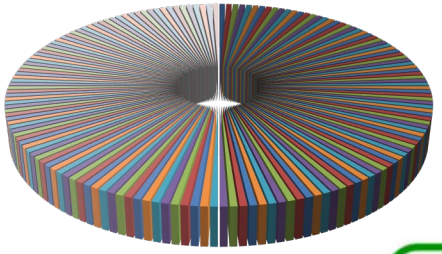
GO category distribution



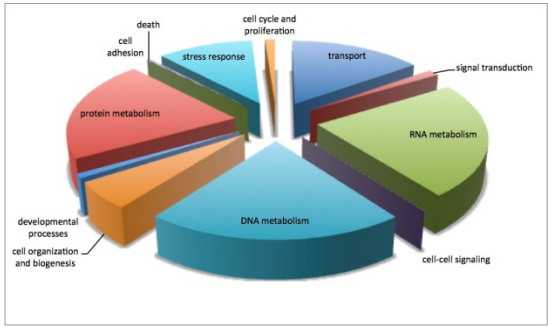
The Necessity for GO Slim

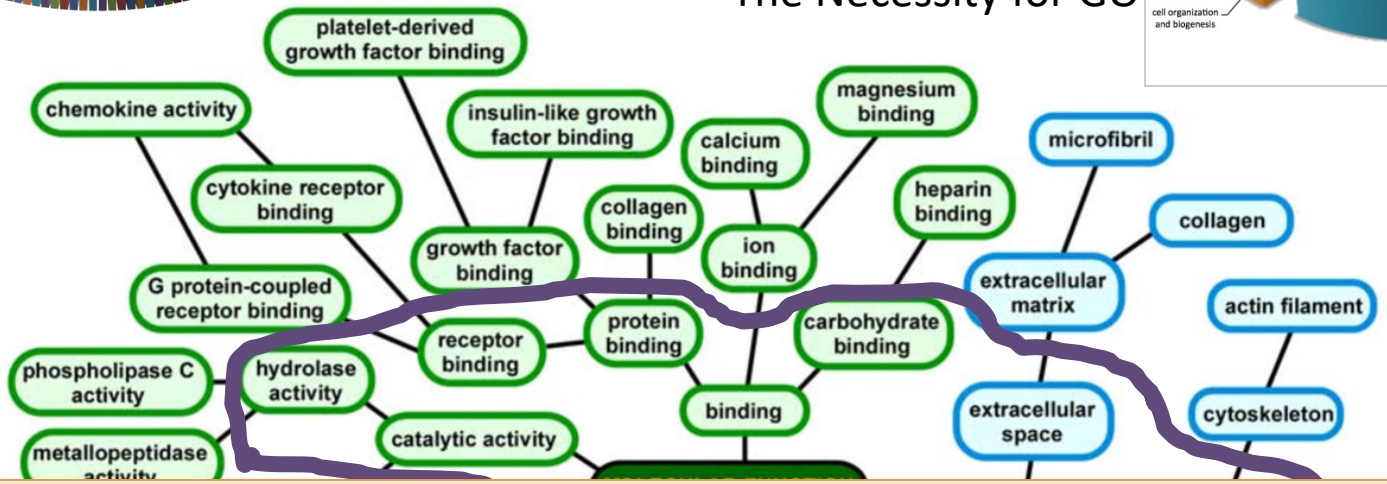






The Necessity for GO





GO Slim

To download premade GO Slim:
http://www.geneontology.org/GO.slims.shtml

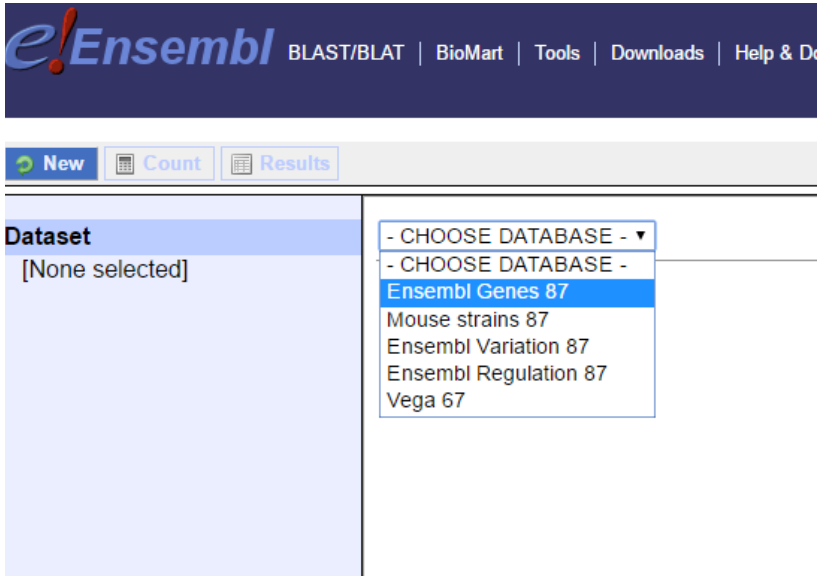
Create your own GO Slim:
http://oboedit.org/docs/html/Creating_Your_Own_GO_Slim_in_OBO_Edit.htm

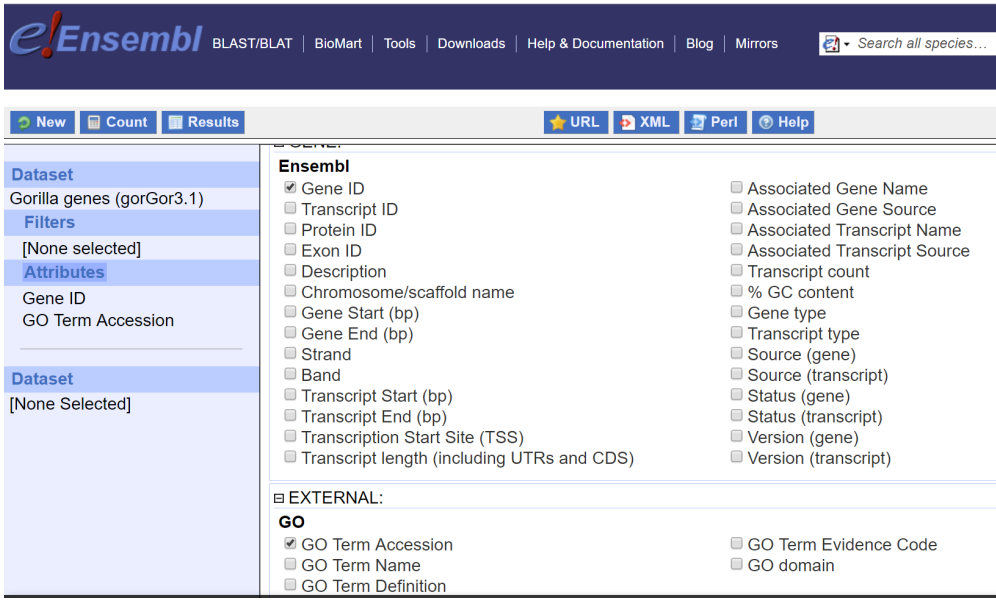
How to get Gene Ontology ?

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





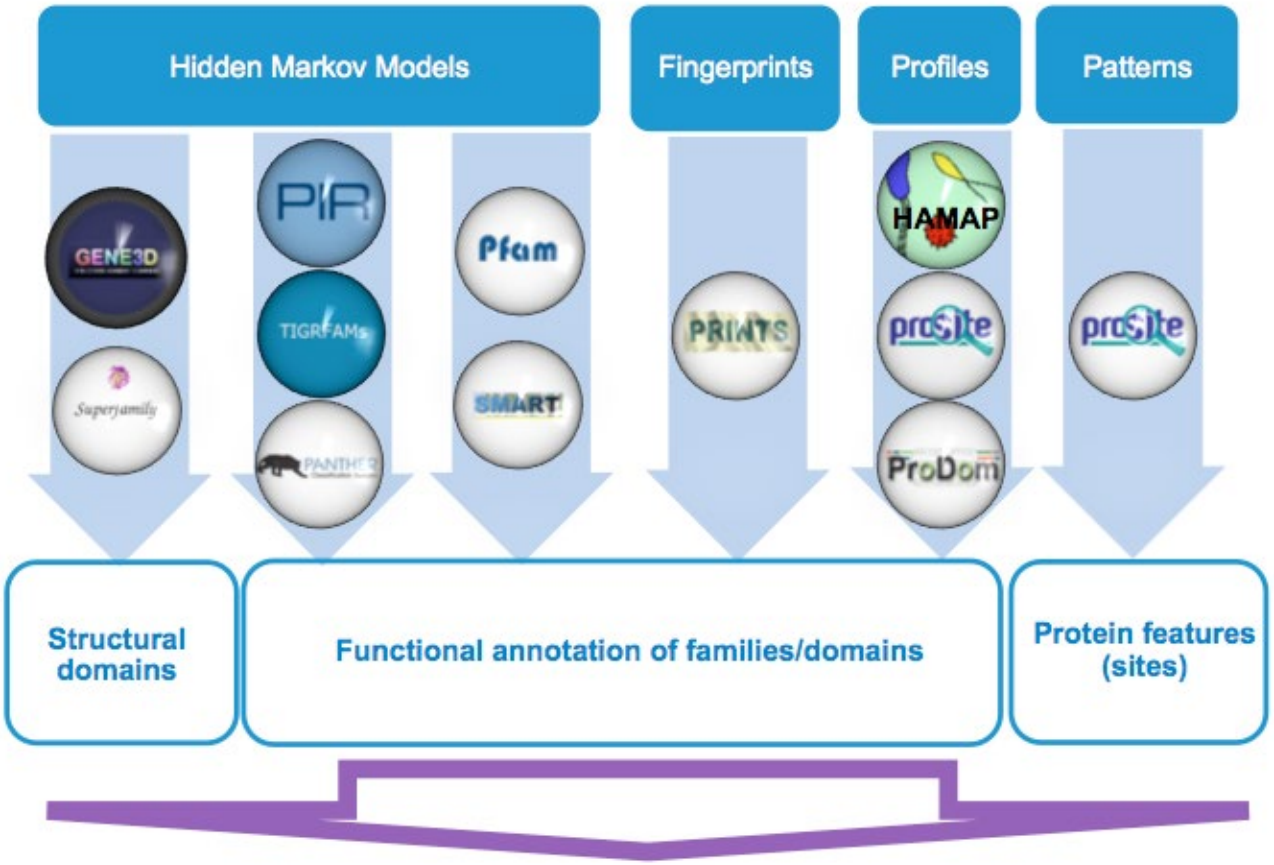
Non model organism

Public tool: InterProScan

Commercial software: BLAST2GO

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

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

How to Get Gene Ontology Data (2)

Your own reference genome

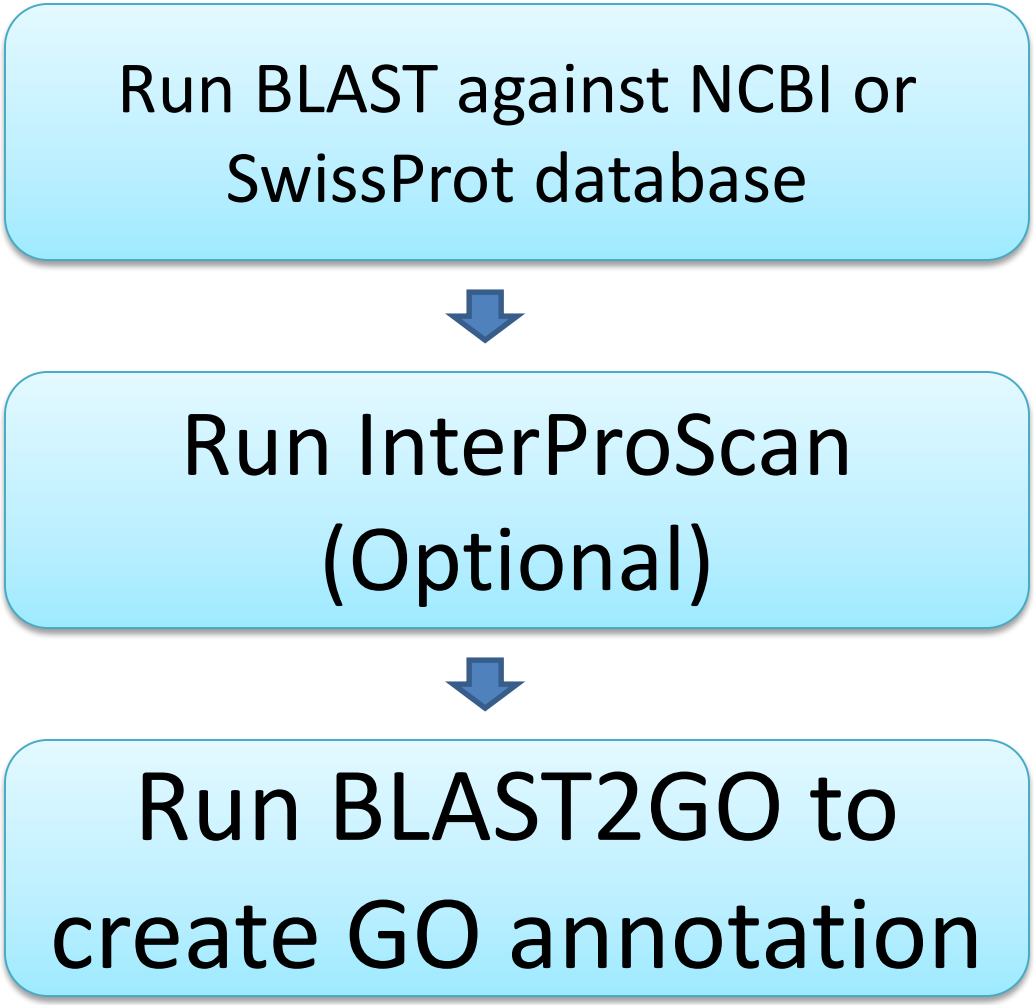
BLAST2GO on BioHPC Lab

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.

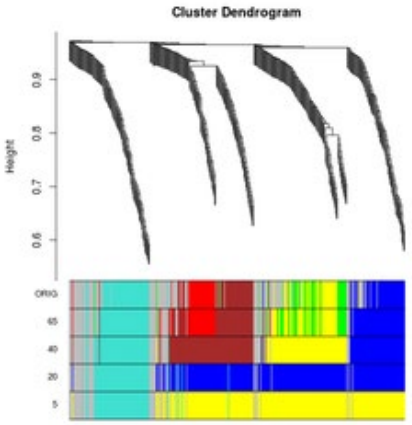
Output from BLAST2GO

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

RNA-seq data



Differentially expressed genes

Co-expression network modules

Function Enrichment Analysis

Public and Commercial Resources of Pathway/Function analysis

- **Public resource:**
  - DAVID Bioinformatics Resources (http://david.abcc.ncifcrf.gov/)

  - Bioconductor: TopGO

- **Commercial Resource:**
  - Ingenuity:

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

How to do GO analysis?

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

	Genes in the genome	DE genes in a experiment
P53 Pathway	40	3 -1
Not P53 Pathway	29960	297

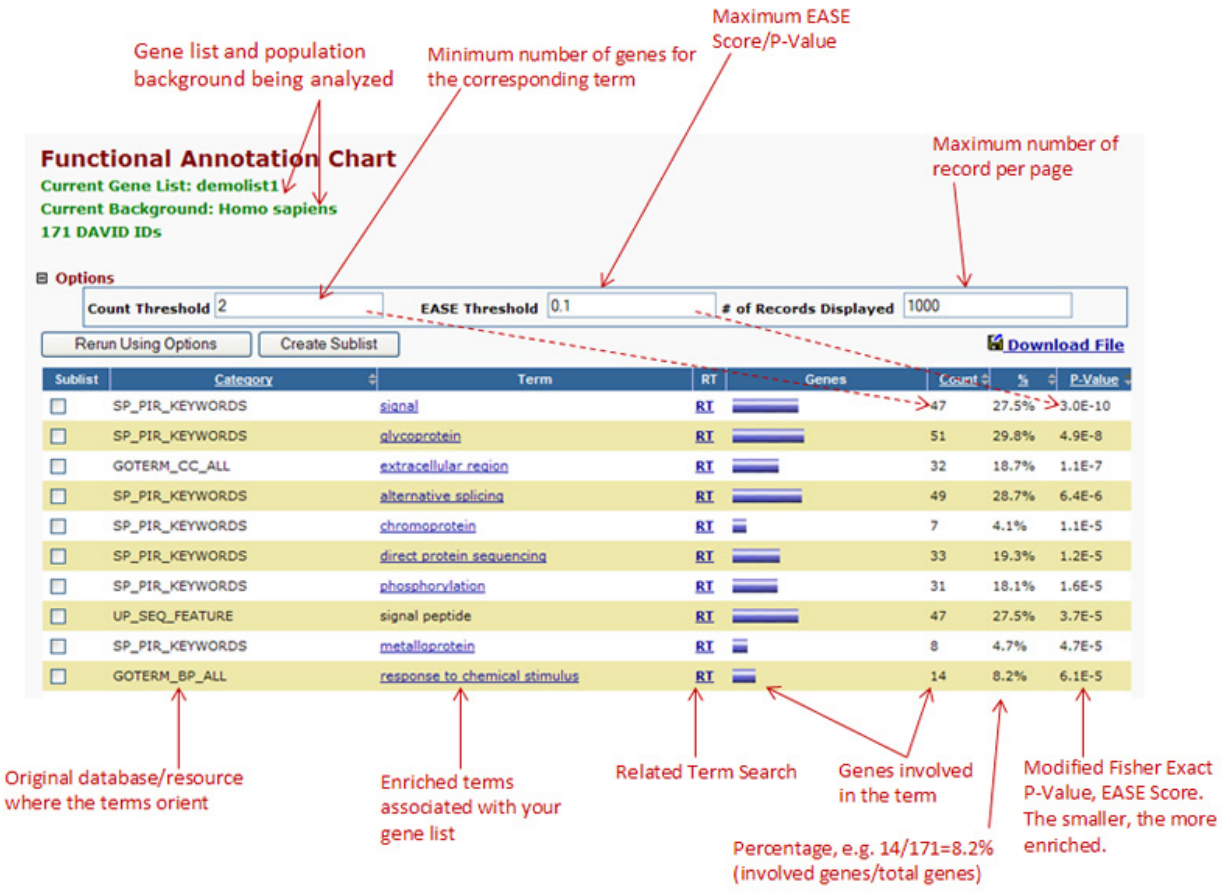
Standard Fisher's exact test: P value= 0.008

EASE Score (in red): P value=0.06

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

Tools for function Enrichment analysis

- DAVID
  - Web based (http://david.abcc.ncifcrf.gov/)
  - Recognized Gene IDs are limited



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

- go.annot: Go annotation file with two columns (gene ID, GO ID)
- Refset: Reference gene sets (all expressed genes)
- Testset: Test gene set (e.g. differentially expressed genes)
- 0.05: P-value cutoff
- BP: only test GO in the Biology Process component (BP, CC or MF)
- 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