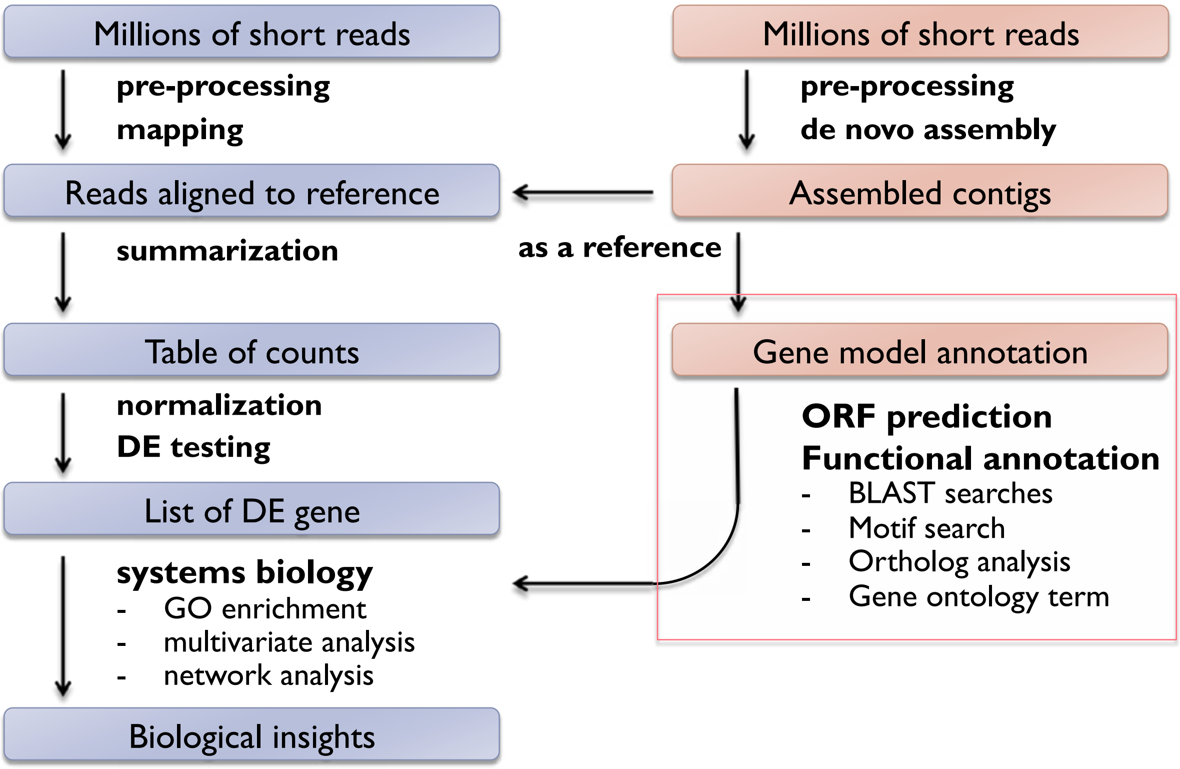
基礎生物学研究所　ゲノムインフォマティクストレーニングコース
RNA-seq入門：実践編　2017年 3月

機能アノテーションと
Gene Ontology解析

Shuji Shigenobu
重信　秀治

基礎生物学研究所
生物機能解析センター

RNA-seq analysis pipeline (*de novo* strategy)



Millions of short reads

pre-processing
mapping

Reads aligned to reference

summarization

Table of counts

normalization
DE testing

List of DE gene

systems biology
- GO enrichment
- multivariate analysis
- network analysis

Biological insights

Millions of short reads

pre-processing
de novo assembly

Assembled contigs

as a reference

Gene model annotation

ORF prediction
Functional annotation
- BLAST searches
- Motif search
- Ortholog analysis
- Gene ontology term

ORF prediction

- Special consideration in ORF prediction after de novo RNA-seq assembly
  - Sometimes partial: Start Met or terminal codon may be missing.
  - Ideally one ORF is present per contig, but erroneously joined contigs may include multiple ORFs.
  - Possible frame shifts.
    - Frame shifts do not occur so often in Illumina, while it happens very frequently in 454 and IonProton.

- Recommended software: TransDecoder

Functional Annotation of Predicted ORFs

- BLAST
  - NCBI NR (or UniProt)
  - species of interest (model organisms, close relatives etc)
  - specific DB (SwissProt, rRNA DB, CEGMA etc)
  - self (assembly v.s. assembly)
- Motif search
  - Pfam, SignalP etc.
- Ortholog analysis
  - vs model organism
  - ortholog database (OrthoDB, eggNOG, OrthoMCL etc)
  - close relatives
- Gene Ontology term assignment

Quick annotation by BLASTX

- Query: assembled contigs

  (nucleotide sequences in multi-fasta format)
- DB: Protein sequences of a model organism

Format DB

```
$ makeblastdb –in protein.fa -dbtype prot
```

Search

```
$ blastx -query trinity_contigs –db protein.fa \
  -num_threads 8 -evalue 1.0e-8 –outfmt 0 > blastxout.txt
```

Protein motif search using InterProScan

- Query: Translated ORF sequences
- Software: InterProScan
  - https://github.com/ebi-pf-team/interproscan/wiki

Search

```
$ interproscan.sh  -I proteins.fasta -f XML,TSV --goterms
--pathways
```

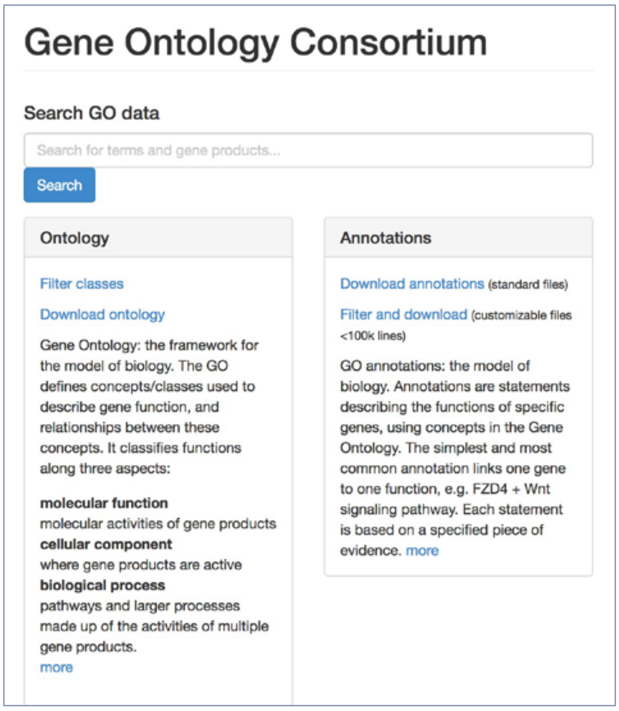
What is Gene Ontology (GO)?

- GO project describes gene products from all organisms using a consistent and computable language.
- GO produces sets of explicitly defined, structured vocabularies in both a computer- and human-readable manner.
- 3 categories
  - Biological processes
  - Molecular functions
  - Cellular components
- 2 components
  - Ontology: term definition terms and the structured relationships between them
  - Associations between gene products and the GO terms.

`http://www.geneontology.org/`

Two components of GO

- Ontology
- Gene associations



Gene Ontology Consortium

Search GO data

Ontology

Filter classes

Download ontology

Gene Ontology: the framework for the model of biology. The GO defines concepts/classes used to describe gene function, and relationships between these concepts. It classifies functions along three aspects:

molecular function
molecular activities of gene products
cellular component
where gene products are active
biological process
pathways and larger processes made up of the activities of multiple gene products.
more

Annotations

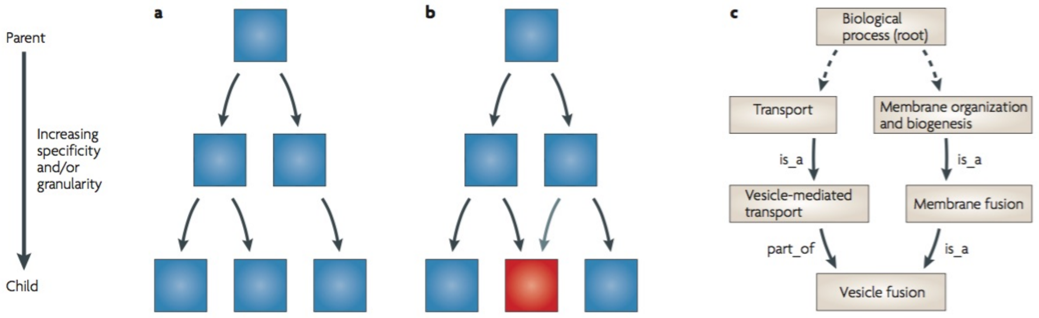
Download annotations (standard files)

Filter and download (customizable files <100k lines)

GO annotations: the model of biology. Annotations are statements describing the functions of specific genes, using concepts in the Gene Ontology. The simplest and most common annotation links one gene to one function, e.g. FZD4 + Wnt signaling pathway. Each statement is based on a specified piece of evidence. more

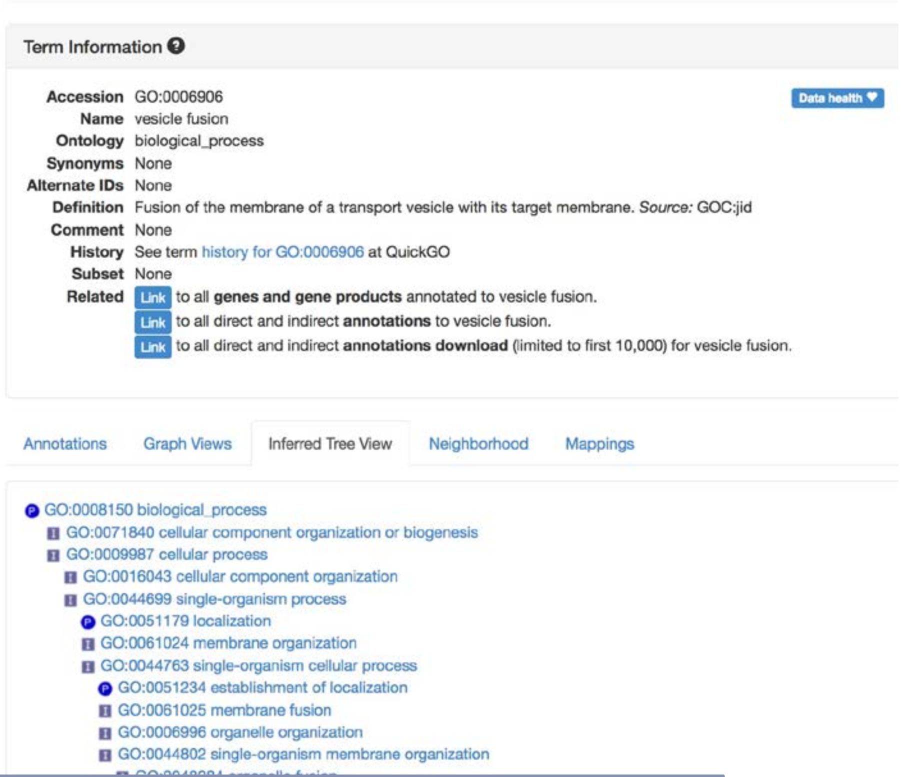
Ontology structure

- Ontologies are represented as a directed acyclic graph (DAG).
- Parent-child relationship
  - is_a
  - part_of
- Ontology can be changed / updated


Parent
Increasing specificity and/or granularity
Child
a
b
c
Biological process (root)
Transport
Membrane organization and biogenesis
is_a
is_a
Vesicle-mediated transport
Membrane fusion
part_of
is_a
Vesicle fusion

Rhee et al., 2008

vesicle fusion



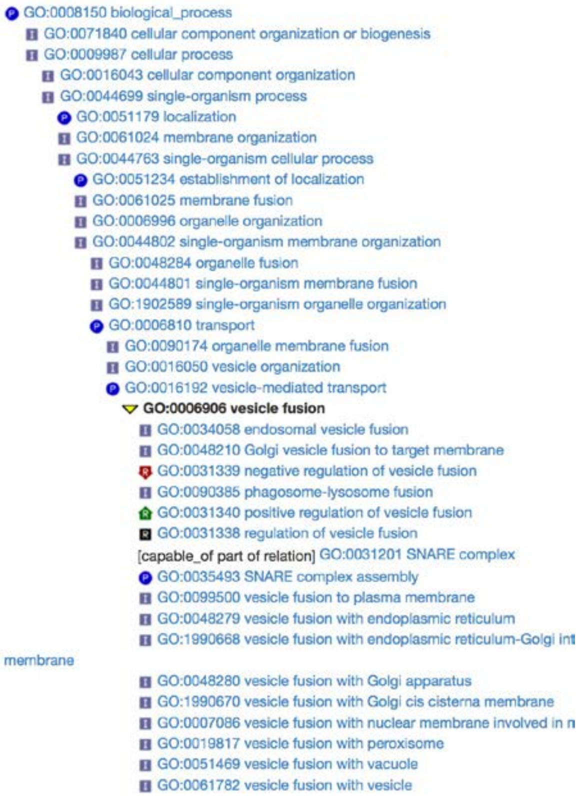
Term Information

Accession GO:0006906
Name vesicle fusion
Ontology biological_process
Synonyms None
Alternate IDs None
Definition Fusion of the membrane of a transport vesicle with its target membrane. *Source:* GOC:jid
Comment None
History See term history for GO:0006906 at QuickGO
Subset None
Related Link to all **genes and gene products** annotated to vesicle fusion.
Link to all direct and indirect **annotations** to vesicle fusion.
Link to all direct and indirect **annotations download** (limited to first 10,000) for vesicle fusion.

Annotations | Graph Views | Inferred Tree View | Neighborhood | Mappings

- P GO:0008150 biological_process
  - I GO:0071840 cellular component organization or biogenesis
  - I GO:0009987 cellular process
    - I GO:0016043 cellular component organization
    - I GO:0044699 single-organism process
      - P GO:0051179 localization
      - I GO:0061024 membrane organization
      - I GO:0044763 single-organism cellular process
        - P GO:0051234 establishment of localization
        - I GO:0061025 membrane fusion
        - I GO:0006996 organelle organization
        - I GO:0044802 single-organism membrane organization

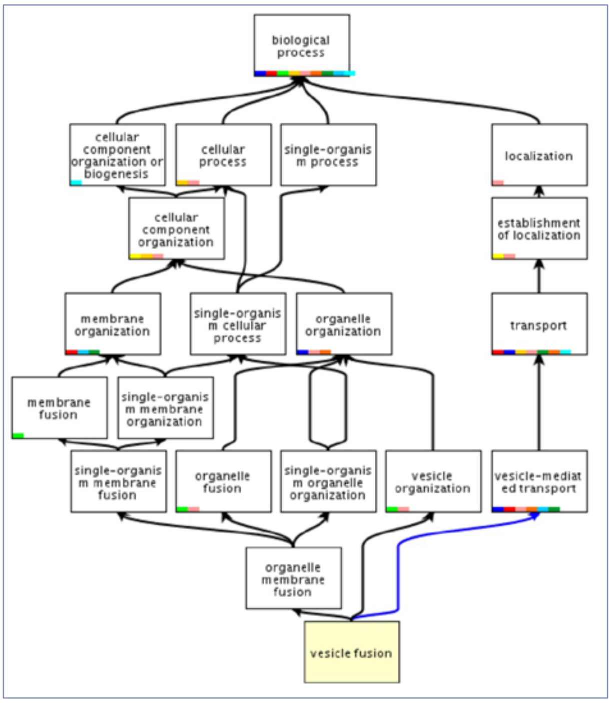
http://amigo.geneontology.org/amigo/term/GO:0006906



- P GO:0008150 biological_process
  - I GO:0071840 cellular component organization or biogenesis
  - I GO:0009987 cellular process
    - I GO:0016043 cellular component organization
    - I GO:0044699 single-organism process
      - P GO:0051179 localization
      - I GO:0061024 membrane organization
      - I GO:0044763 single-organism cellular process
        - P GO:0051234 establishment of localization
        - I GO:0061025 membrane fusion
        - I GO:0006996 organelle organization
        - I GO:0044802 single-organism membrane organization
          - I GO:0048284 organelle fusion
          - I GO:0044801 single-organism membrane fusion
          - I GO:1902589 single-organism organelle organization
          - P GO:0006810 transport
            - I GO:0090174 organelle membrane fusion
            - I GO:0016050 vesicle organization
            - P GO:0016192 vesicle-mediated transport
              - ▽ **GO:0006906 vesicle fusion**
                - I GO:0034058 endosomal vesicle fusion
                - I GO:0048210 Golgi vesicle fusion to target membrane
                - R GO:0031339 negative regulation of vesicle fusion
                - I GO:0090385 phagosome-lysosome fusion
                - R GO:0031340 positive regulation of vesicle fusion
                - R GO:0031338 regulation of vesicle fusion
                - [capable_of part of relation] GO:0031201 SNARE complex
                - P GO:0035493 SNARE complex assembly
                - I GO:0099500 vesicle fusion to plasma membrane
                - I GO:0048279 vesicle fusion with endoplasmic reticulum
                - I GO:1990668 vesicle fusion with endoplasmic reticulum-Golgi int

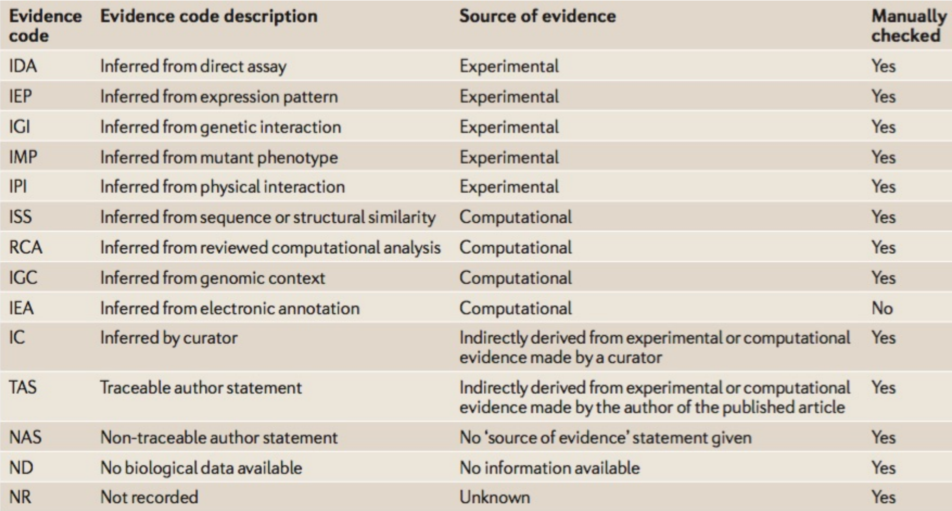
membrane

- I GO:0048280 vesicle fusion with Golgi apparatus
- I GO:1990670 vesicle fusion with Golgi cis cisterna membrane
- I GO:0007086 vesicle fusion with nuclear membrane involved in n
- I GO:0019817 vesicle fusion with peroxisome
- I GO:0051469 vesicle fusion with vacuole
- I GO:0061782 vesicle fusion with vesicle


biological process
cellular component organization or biogenesis
cellular process
single-organism process
localization
cellular component organization
establishment of localization
membrane organization
single-organism cellular process
organelle organization
transport
membrane fusion
single-organism membrane organization
single-organism membrane fusion
organelle fusion
single-organism organelle organization
vesicle organization
vesicle-mediated transport
organelle membrane fusion
vesicle fusion

Gene association

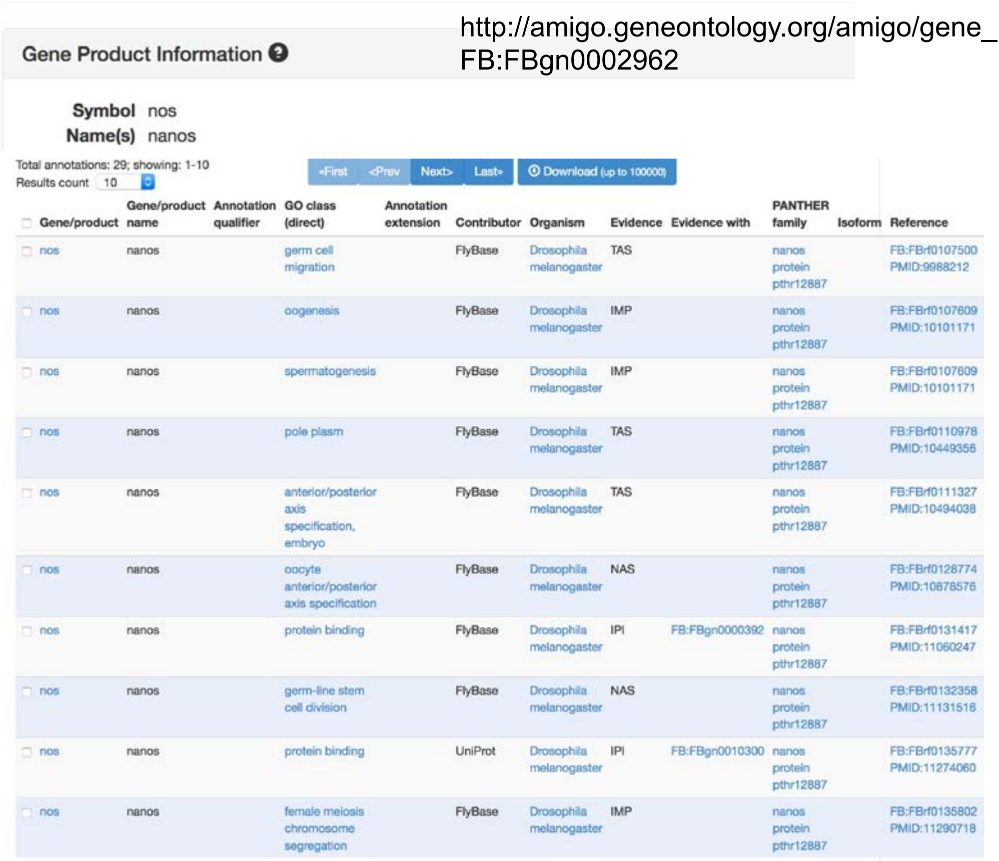
- Gene <=> GO
- A gene may associate with multiple GO terms.
- Evidence codes.



Evidence code	Evidence code description	Source of evidence	Manually checked
IDA	Inferred from direct assay	Experimental	Yes
IEP	Inferred from expression pattern	Experimental	Yes
IGI	Inferred from genetic interaction	Experimental	Yes
IMP	Inferred from mutant phenotype	Experimental	Yes
IPI	Inferred from physical interaction	Experimental	Yes
ISS	Inferred from sequence or structural similarity	Computational	Yes
RCA	Inferred from reviewed computational analysis	Computational	Yes
IGC	Inferred from genomic context	Computational	Yes
IEA	Inferred from electronic annotation	Computational	No
IC	Inferred by curator	Indirectly derived from experimental or computational evidence made by a curator	Yes
TAS	Traceable author statement	Indirectly derived from experimental or computational evidence made by the author of the published article	Yes
NAS	Non-traceable author statement	No 'source of evidence' statement given	Yes
ND	No biological data available	No information available	Yes
NR	Not recorded	Unknown	Yes

nanos

http://amigo.geneontology.org/amigo/gene_product/FB:FBgn0002962



Gene Product Information

Symbol nos
Name(s) nanos

Total annotations: 29; showing: 1-10
Results count 10

Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence	Evidence with	PANTHER family	Isoform	Reference
□ nos	nanos		germ cell migration		FlyBase	Drosophila melanogaster	TAS		nanos protein pthr12887		FB:FBrf0107500 PMID:9988212
□ nos	nanos		oogenesis		FlyBase	Drosophila melanogaster	IMP		nanos protein pthr12887		FB:FBrf0107609 PMID:10101171
□ nos	nanos		spermatogenesis		FlyBase	Drosophila melanogaster	IMP		nanos protein pthr12887		FB:FBrf0107609 PMID:10101171
□ nos	nanos		pole plasm		FlyBase	Drosophila melanogaster	TAS		nanos protein pthr12887		FB:FBrf0110978 PMID:10449356
□ nos	nanos		anterior/posterior axis specification, embryo		FlyBase	Drosophila melanogaster	TAS		nanos protein pthr12887		FB:FBrf0111327 PMID:10494038
□ nos	nanos		oocyte anterior/posterior axis specification		FlyBase	Drosophila melanogaster	NAS		nanos protein pthr12887		FB:FBrf0128774 PMID:10878576
□ nos	nanos		protein binding		FlyBase	Drosophila melanogaster	IPI	FB:FBgn0000392	nanos protein pthr12887		FB:FBrf0131417 PMID:11060247
□ nos	nanos		germ-line stem cell division		FlyBase	Drosophila melanogaster	NAS		nanos protein pthr12887		FB:FBrf0132358 PMID:11131516
□ nos	nanos		protein binding		UniProt	Drosophila melanogaster	IPI	FB:FBgn0010300	nanos protein pthr12887		FB:FBrf0135777 PMID:11274060
□ nos	nanos		female meiosis chromosome segregation		FlyBase	Drosophila melanogaster	IMP		nanos protein pthr12887		FB:FBrf0135802 PMID:11290718

How to annotate GO for non-model organisms?

- Ortholog grouping with a model organism and then transfer the GO terms from the reference organism to your target organism.
- BLAST2GO

Gene Ontology enrichment analysis

- What is GO enrichment analysis?
- Why GO enrichment analysis is required in DEG studies?
- Type of GO enrichment analysis.
  - gene set
  - gene score
- Software
  - gene set type: DAVID (web), metascape (web), goseq (R), GOstat (R)
  - gene score: GSEA, roast, camera
  - both: ErmineJ

Basic over-representation test: 2 x 2 table and Fisher's exact test

- Suppose we perform a test of DE and find a list of 200 significant genes out of 10,000
- Consider a specific GO term, apoptosis. Among the 200 DE genes, 100 genes are annotated as apoptosis related, while 300 / 10,000 are associated with apoptosis in the whole gene set.
- Question: Is the gene set "apoptosis" over-represented among "significant" genes?

	apoptosis	non-apoptosis	total
DE	**20**	180	**200**
non-DE	280	9,520	9,800
	300	9,700	**10,000**

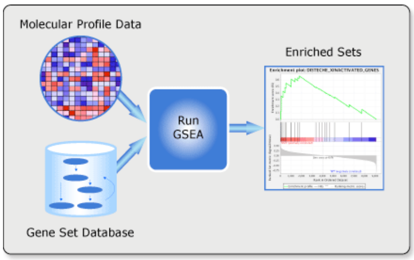
```
> mat <- matrix(c(20,200-20,300-20, 10000-300-(200-20)),
nrow=2, byrow=T)
> fisher.test(mat, alternative="greater")

   Fisher's Exact Test for Count Data

data:  mat
p-value = 2.269e-06
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 2.418508      Inf
sample estimates:
odds ratio
  3.777069
```

Gene score type enrichment analysis

- Drawback of basic 2x2 table method
  - Threshold value is arbitral
  - Magnitude of significance is ignored
- GSEA
  - http://software.broadinstitute.org/gsea/index.jsp
- ROAST, CAMERA
  - implemented within edgeR



Tutorial: ErmineJ

- http://erminej.chibi.ubc.ca/

- Easy to use Java software with both GUI and CUI
- Three enrich methods supported
  - ORA: overrepresentation analysis
  - GSR: gene score resampling
  - ROC: rank-based gene score in receiver-operator curves