

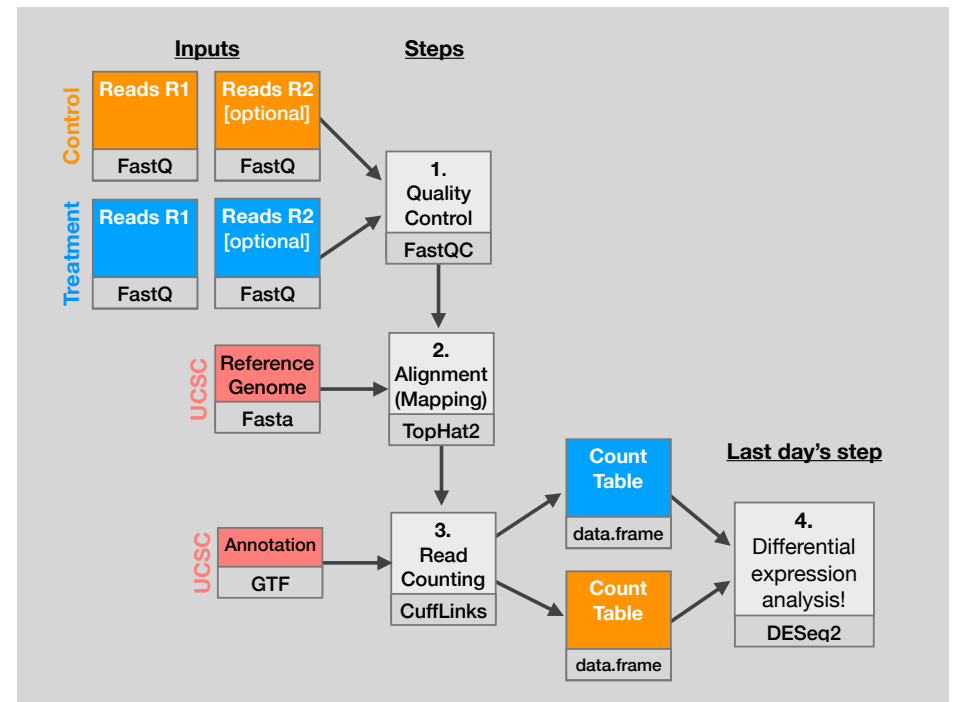
BGGN 213

Pathway Analysis and the Interpretation of Gene Lists

Lecture 15

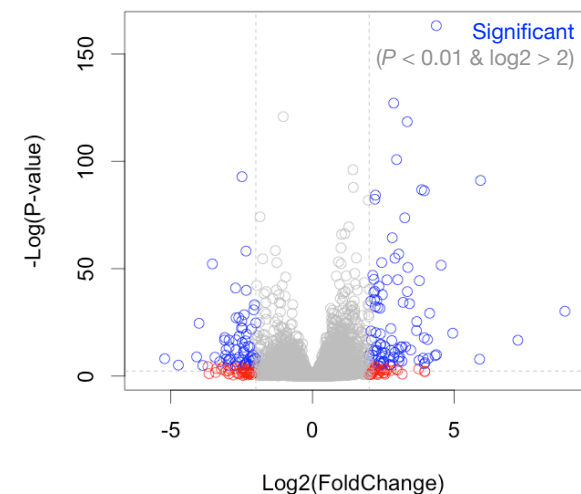
Barry Grant
UC San Diego

<http://thegrantlab.org/bgg213>



| X | baseMean | log2FoldChange | lfcSE | stat | pvalue | padj | symbol |
|-----------------|-------------|----------------|------------|------------|--------------|--------------|---------|
| ENSG00000152583 | 954.77093 | 4.3683590 | 0.23713648 | 18.421286 | 8.867079e-76 | 1.342919e-71 | SPARCL1 |
| ENSG00000179094 | 743.25269 | 2.8638885 | 0.17555825 | 16.313039 | 7.972621e-60 | 6.037267e-56 | PER1 |
| ENSG00000116584 | 2277.91345 | -1.0347000 | 0.06505273 | -15.905557 | 5.798513e-57 | 2.927283e-53 | ARHGEF2 |
| ENSG00000189221 | 2383.75371 | 3.3415441 | 0.21241508 | 15.731200 | 9.244206e-56 | 3.500088e-52 | MAOA |
| ENSG00000120129 | 3440.70375 | 2.9652108 | 0.20370277 | 14.556557 | 5.306416e-48 | 1.607313e-44 | DUSP1 |
| ENSG00000148175 | 13493.92037 | 1.4271683 | 0.10036663 | 14.219550 | 6.929711e-46 | 1.749175e-42 | STOM |
| ENSG00000178695 | 2685.40974 | -2.4890689 | 0.17806407 | -13.978501 | 2.108817e-44 | 4.562576e-41 | KCTD12 |
| ENSG00000109906 | 439.54152 | 5.9275950 | 0.42819442 | 13.843233 | 1.397758e-43 | 2.646131e-40 | ZBTB16 |
| ENSG00000134686 | 2933.64246 | 1.4394898 | 0.10582729 | 13.602255 | 3.882769e-42 | 6.533838e-39 | PHC2 |
| ENSG00000101347 | 14134.99177 | 3.8504143 | 0.28490701 | 13.514635 | 1.281894e-41 | 1.941428e-38 | SAMHD1 |
| ENSG00000096060 | 2630.23049 | 3.9450524 | 0.29291821 | 13.468102 | 2.409807e-41 | 3.317866e-38 | FKBP5 |
| ENSG00000166741 | 7542.25287 | 2.2195906 | 0.16673544 | 13.312050 | 1.970000e-40 | 2.486304e-37 | NNMT |
| ENSG00000125148 | 3695.87946 | 2.1985636 | 0.16700546 | 13.164621 | 1.402400e-39 | 1.633797e-36 | MT2A |
| ENSG00000162614 | 5646.18314 | 1.9711402 | 0.15020631 | 13.122885 | 2.434854e-39 | 2.633990e-36 | NEXN |
| ENSG00000106976 | 989.04683 | -1.8501713 | 0.14778657 | -12.519211 | 5.861471e-36 | 5.918132e-33 | DNM1 |
| ENSG00000187193 | 199.07694 | 3.2551424 | 0.26090711 | 12.476250 | 1.006146e-35 | 9.523804e-33 | MT1X |
| ENSG00000256235 | 1123.47954 | 1.2801193 | 0.10547438 | 12.136779 | 6.742862e-34 | 6.007096e-31 | SMIM3 |
| ENSG00000177666 | 2639.57020 | 1.1399947 | 0.09606884 | 11.866436 | 1.768422e-32 | 1.487930e-29 | PNPLA2 |
| ENSG00000164125 | 7257.00808 | 1.0248523 | 0.08657600 | 11.837603 | 2.494830e-32 | 1.988642e-29 | FAM198B |
| ENSG00000198624 | 2020.04495 | 2.8141014 | 0.24063429 | 11.694515 | 1.359615e-31 | 1.029569e-28 | CCDC69 |
| ENSG00000123562 | 5008.55294 | 1.0045453 | 0.08901501 | 11.285123 | 1.554241e-29 | 1.120904e-26 | MORF4L2 |
| ENSG00000144369 | 1283.77980 | -1.3090041 | 0.11714863 | -11.173875 | 5.473974e-29 | 3.768333e-26 | FAM171B |
| ENSG00000196517 | 241.91536 | -2.3456877 | 0.21047366 | -11.144804 | 7.591120e-29 | 4.998588e-26 | SLC6A9 |
| ENSG00000135821 | 19973.40000 | 3.0413943 | 0.27601796 | 11.018828 | 3.100706e-28 | 1.956675e-25 | GLUL |

Volcano Plot
Fold change vs P-value



My high-throughput
experiment generated a
long list of genes/proteins...

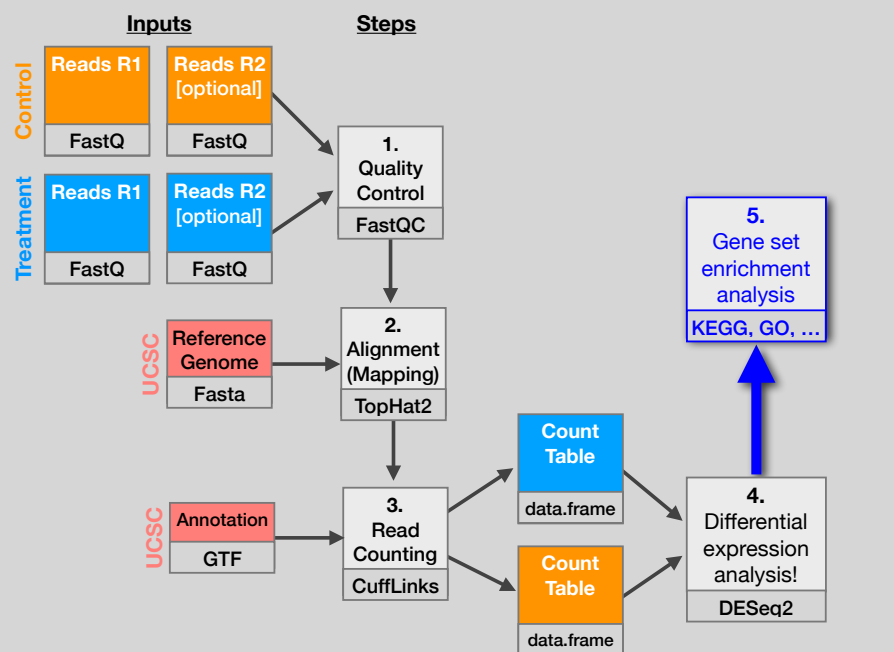
What do I do now?



Pathway analysis!

(a.k.a. geneset enrichment)

Use bioinformatics methods to help extract biological meaning from such lists...



Basic idea

Differentially Expressed Genes (DEGs)

[illegible]

Gene-sets (Pathways, annotations, etc...)

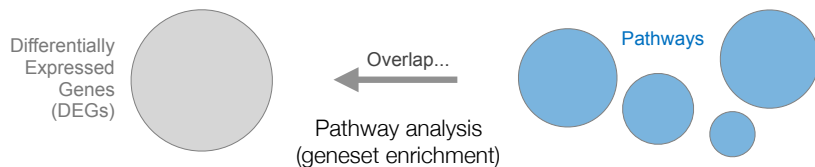
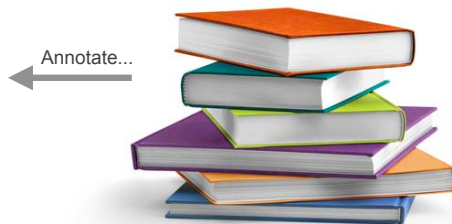


Basic idea

Differentially Expressed Genes (DEGs)

[illegible]

Gene-sets (Pathways, annotations, etc...)



Pathway analysis (a.k.a. geneset enrichment)

Limitations

- **Geneset annotation bias:** can only discover what is already known
- **Non-model organisms:** no high-quality genesets available
- **Post-transcriptional regulation** is neglected
- **Tissue-specific** variations of pathways are not annotated
 - e.g. NF- κ B regulates metabolism, not inflammation, in adipocytes
- **Size bias:** stats are influenced by the size of the pathway
 - Many pathways/receptors **converge** to few regulators
e.g. Tens of innate immune receptors activate four TFs:
NF- κ B, AP-1, IRF3/7, NFAT

Side-note:

Pathway analysis (a.k.a. geneset enrichment)

Principle



- DEGs come from your experiment ➤ *Critical, needs to be as clean as possible*
- Pathway genes ("geneset") come from annotations ➤ *Important, but typically not a competitive advantage*
- Variations of the math: overlap, ranking, networks... ➤ *Not critical, different algorithms show similar performances*

Starting point for pathway analysis: **Your gene list**

- You have a list of genes/proteins of interest
- You have quantitative data for each gene/protein

- Fold change
- p-value
- Spectral counts
- Presence/absence

| | | | |
|--------|--------------------|-----------|----------|
| 228018 | ENSGG000000090339 | NP_000192 | C20orf58 |
| 226 | ENSGG000000003030 | NP_057219 | |
| 207 | ENSGG00000003383 | 055029 | |
| 225 | ENSGG0000000151513 | 000585 | |
| 221 | ENSGG00000003613 | 006125 | 01 |
| 1553 | ENSGG00000007124 | 89495 | 03 |
| 2184 | ENSGG0000000757 | 01032249 | 08 |
| 2049 | ENSGG000000092370 | 78870 | 11 |
| 2026 | ENSGG000000079646 | 4515 | orf112 |
| 23095 | ENSGG000000056892 | 8339 | orf58 |
| 22801 | ENSGG000000125450 | 412 | NMB |
| 15540 | ENSGG000000253982 | 069 | PA2 |
| 20312 | ENSGG000000140688 | 83 | MEM50B |
| 22518 | ENSGG00000010457 | 01 | MP2 |
| 225079 | ENSGG00000009518 | 05340 | MSI2 |
| 243010 | ENSGG00000002013 | 5 | C20orf58 |
| 230668 | ENSGG00000004050 | 033666 | C8orf4 |
| 128541 | ENSGG000000033666 | 002332 | ETV7 |
| 242225 | ENSGG000000033666 | 002332 | LTB |
| 207339 | ENSGG000000033666 | 002332 | ITC1 |
| 202637 | s at W03F8.6 | | |

Translating between identifiers

- Many different identifiers exist for genes and proteins, e.g. UniProt, Entrez, etc.
- Often you will have to translate one set of ids into another
 - A program might only accept certain types of ids
 - You might have a list of genes with one type of id and info for genes with another type of id

Translating between identifiers

- Many different identifiers exist for genes and proteins, e.g. UniProt, Entrez, etc.
- Often you will have to translate one set of ids into another
 - A program might only accept certain types of ids
 - You might have a list of genes with one type of id and info for genes with another type of id
- **Various web sites translate ids -> *best for small lists***
 - UniProt < www.uniprot.org>; IDConverter < idconverter.biinfo.cnio.es >

Translating between identifiers: UniProt < www.uniprot.org >

The screenshot shows the UniProt website's ID Mapping tool. At the top, there is a navigation bar with links for Downloads, Contact, Documentation, and Help. Below this is a search bar with a dropdown menu set to 'Protein Knowledgebase (UniProtKB)'. To the right of the search bar are buttons for Search, Clear, and Fields. Below the search bar are buttons for Search, Blast, Align, Retrieve, and ID Mapping, with the ID Mapping button highlighted by a red rectangle. Below the search bar, there are sections for WELCOME and NEWS. In the center, there is a box titled 'Identifiers' with a large empty text area. To the right of this box are two dropdown menus labeled 'From' and 'To'. The 'From' dropdown is set to 'EMBL/GenBank/DBJ' and the 'To' dropdown is set to 'UniProtKB AC'. Below these dropdowns are buttons for Map, Swap, and Clear. At the bottom, there is a section labeled 'or' with a 'Choose File' button and the text 'no file selected'.

Translating between identifiers

- Many different identifiers exist for genes and proteins, e.g. UniProt, Entrez, etc.
- Often you will have to translate one set of ids into another
 - A program might only accept certain types of ids
 - You might have a list of genes with one type of id and info for genes with another type of id
- Various web sites translate ids -> *best for small lists*
 - UniProt < www.uniprot.org>; IDConverter < idconverter.biinfo.cnio.es >
- **VLOOKUP in Excel - good if you are an excel whizz - I am not!**
 - Download flat file from Entrez, Uniprot, etc; Open in Excel; Find columns that correspond to the 2 IDs you want to convert between; Sort by ID; Use vlookup to translate your list

Translating between identifiers: Excel VLOOKUP

VLOOKUP(lookup_value, table_array, col_index_num)

The screenshot shows an Excel spreadsheet with a formula bar containing `=VLOOKUP(A3,SG$3:$O$30490,2,FALSE)`. The spreadsheet is divided into two main sections: a 'Data Table' on the left and an 'Annotation Table' on the right.

| Data Table | | | | | Annotation Table | | | | |
|------------|------------|------------|------------|------------|------------------|-----------|-----------|-----------|--------|
| RefSeq | Symbol | Exp1 | Exp2 | Exp3 | RefSeq | Symbol | Entrez ID | Unigene | RefSeq |
| NM_153103 | Kif1c | 2.31975457 | 1.24558927 | 2.78816871 | NM_001001 | Zfp85-rs1 | 22746 | Mm.288396 | NM_001 |
| NM_146017 | Gabrp | 4.15029735 | 3.08055836 | 1.18919962 | NM_001001 | Scap | 235623 | Mm.288741 | NM_001 |
| NM_018883 | Camkk1 | 3.83282512 | 0.0522951 | 0.64684259 | NM_001001 | Scap | 235623 | Mm.288741 | NM_001 |
| NM_145936 | Tspyl2 | 0.45449369 | 1.62761318 | 7.59770627 | NM_001001 | Fbxo41 | 330369 | Mm.38777 | NM_001 |
| NM_026599 | Cgml1 | 4.84541871 | 2.84751796 | 1.61595768 | NM_001001 | Taf9b | 407786 | Mm.19440 | NM_001 |
| NM_013926 | Cbx8 | 1.22903318 | 0.2863077 | 0.02952665 | NM_001001 | Taf9b | 407786 | Mm.19440 | NM_001 |
| NR_015566 | A330023F24 | 1.44699053 | 0.98809479 | 1.59330144 | NM_001001 | BC051142 | 407788 | Mm.73205 | NM_001 |
| NM_008623 | Mpz | 0.50749263 | 0.94350028 | 6.10581569 | NM_001001 | BC051142 | 407788 | Mm.73205 | NM_001 |
| NM_183127 | Fate1 | 2.45672795 | 4.87960794 | 3.60759511 | NM_001001 | BC048546 | 232400 | Mm.259234 | NM_001 |
| NM_008943 | | 4.78701069 | 4.15302647 | 0.85432314 | NM_001001 | Zfp941 | 407812 | Mm.359154 | NM_001 |
| NM_025382 | | 0.66397344 | 1.40664187 | 3.09539802 | NM_001001 | BC031181 | 407819 | Mm.29866 | NM_001 |
| NM_182841 | | 1.25528938 | 0.20505996 | 2.76879488 | NM_001001 | Baz2b | 407823 | Mm.486364 | NM_001 |
| NM_030061 | | 0.17670108 | 2.75415469 | 2.98900691 | NM_001001 | Tmem204 | 407831 | Mm.34379 | NM_001 |
| NM_133216 | | 6.572343 | 0.59671282 | 3.84650536 | NM_001001 | Ccdc111 | 408022 | Mm.217385 | NM_001 |
| NM_030063 | | 7.05132762 | 0.65043627 | 1.68111836 | NM_001001 | BC048507 | 408058 | Mm.177840 | NM_001 |

Translating between identifiers

- Many different identifiers exist for genes and proteins, e.g. UniProt, Entrez, etc.
- Often you will have to translate one set of ids into another
 - A program might only accept certain types of ids
 - You might have a list of genes with one type of id and info for genes with another type of id
- Various web sites translate ids -> *best for small lists*
 - UniProt < www.uniprot.org >; IDConverter < idconverter.bioinfo.cnio.es >
- VLOOKUP in Excel -> *good if you are an excel whizz - I am not!*
 - Download flat file from Entrez, Uniprot, etc; Open in Excel; Find columns that correspond to the two ids you want to convert between; Use vlookup to translate your list

- Use the `merge()` or `mapIds()` functions in R - *fast, versatile & reproducible!*
 - Also `clusterProfiler::bitr()` function and many others... [\[Link to clusterProfiler vignette\]](#)

The screenshot shows an R console window titled "2. class-material (bash)". It displays the following code:

```
# Using the merge() function
> anno <- read.csv("data/annotables_grch38.csv")
> merge(mygenes, anno, by.x="row.names", by.y="ensgene")
```

Annotations:

- A green box highlights the file path: "This is an annotation file".
- A red box highlights the `merge()` function call: "This is our differential expressed genes".

A red banner in the top right corner says "Reminder".

The screenshot shows an R console window titled "2. class-material (bash)". It displays the following code:

```
# Using the mapIds() function from bioconductor
> library("AnnotationDbi")
> library("org.Hs.eg.db")
> mygenes$symbol <- mapIds(org.Hs.eg.db,
  column="SYMBOL",
  keys=row.names(mygenes),
  keytype="ENSEMBL")
```

Annotations:

- A green box highlights the library loading: "Load the required Bioconductor packages".
- A red box highlights the `mapIds()` function call: "Annotation we want to add".
- A red box highlights the `row.names(mygenes)` argument: "Our vector of gene names & their format".

A red banner in the top right corner says "Reminder".

bitr: Biological Id Translator

clusterProfiler provides `bitr` and `bitr_kegg` for converting ID types. Both `bitr` and `bitr_kegg` support many species including model and many non-model organisms.

```
x <- c("GPX3", "GLRX", "LBP", "CRYAB", "DEFB1", "HCLS1", "SOD2", "HSPA2",
      "ORM1", "IGFBP1", "PTHLH", "GPC3", "IGFBP3", "TOB1", "MITF", "NDRG1",
      "NR1H4", "FGFR3", "PVR", "IL6", "PTPRM", "ERBB2", "NID2", "LAMB1",
      "COMP", "PLS3", "MCAM", "SPP1", "LAMC1", "COL4A2", "COL4A1", "MYOC",
      "ANXA4", "TFPI2", "CST6", "SLPI", "TIMP2", "CPM", "GGT1", "NNMT",
      "MAL", "EEF1A2", "HGD", "TCN2", "CDA", "PCCA", "CRYM", "PDXK",
      "STC1", "WARS", "HMOX1", "FXRD", "RBP4", "SLC6A12", "KDEL3", "ITM2B")
eg = bitr(x, fromType="SYMBOL", toType="ENTREZID", OrgDb="org.Hs.eg.db")
head(eg)
```

```
## SYMBOL ENTREZID
## 1 GPX3 2878
## 2 GLRX 2745
## 3 LBP 3929
## 4 CRYAB 1410
## 5 DEFB1 1672
## 6 HCLS1 3059
```

See package vignette:

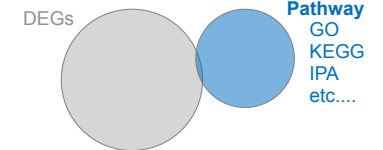
<https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html>

Alternative...

What functional set databases do you want?

• Most commonly used:

- **Gene Ontology (GO)**
- **KEGG Pathways** (mostly metabolic)
- **GeneGO MetaBase**
- **Ingenuity Pathway Analysis (IPA)**



• Many others...

- **Enzyme Classification, PFAM, Reactome,**
- Disease Ontology, MSigDB, Chemical Entities of Biological Interest, Network of Cancer Genes etc...
- See: Open Biomedical Ontologies (www.obofoundry.org)

GO < www.geneontology.org >

• What function does HSF1 perform?

- *response to heat; sequence-specific DNA binding; transcription; etc*

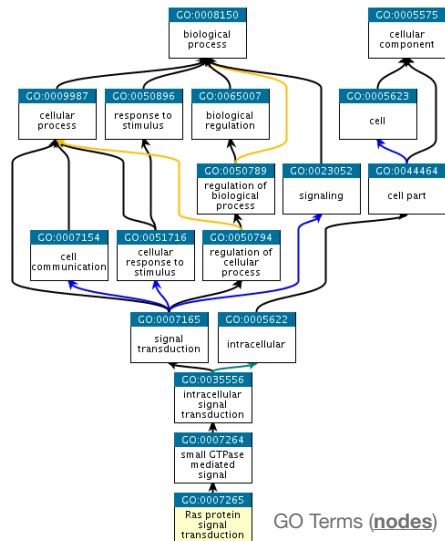
- **Ontology** => a structured and controlled vocabulary that allows us to annotate gene products consistently, interpret the relationships among annotations, and can easily be *handled by a computer*

- GO database consists of 3 ontologies that describe gene products in terms of their associated **biological processes**, **cellular components** and **molecular functions**

GO Annotations

- GO is not a stand-alone database of genes/proteins or sequences
- Rather gene products get annotated with **GO terms** by UniProt and other organism specific databases, such as Flybase, Wormbase, MGI, ZFIN, etc.
- Annotations are available through AmiGO < amigo.geneontology.org >

GO is structured as a “directed graph”



Relationships (edges)

Parent terms are more general & child terms more specific

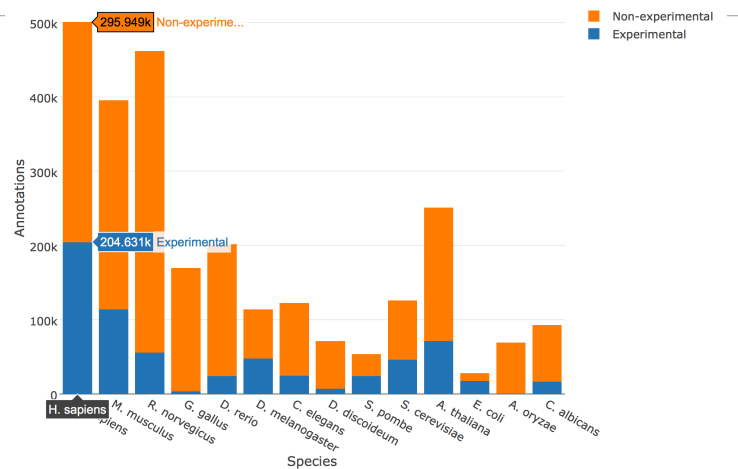
GO evidence codes

| 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

Use and misuse of the gene ontology annotations
Seung Yon Rhee, Valerie Wood, Kara Dolinski & Sorin Draghici
Nature Reviews Genetics 9, 509-515 (2008)

Experimental annotations by species



• See AmiGO for details: http://amigo.geneontology.org/amigo/base_statistics

Can now do gene list analysis with GeneGO online!

pantherdb.org/webservices/go/overrep.jsp

GENE ONTOLOGY
Uniting biology

PANTHER
Classification System

Home About PANTHER Data PANTHER Tools Workspace Downloads Help/Tutorial

NEW! PANTHER13.1 released.

Search

Quick links

Whole genome function views

Genome statistics

Data Version

How to cite PANTHER

NEWS

PANTHER13.1 Released

Click for additional info.

Newsletter subscription

Enter your Email:

Subscribe

Print/Save/SQL

Gene List Analysis

Browse Sequence Search cSNP Scoring Keyword Search

Please refer to our article in Nature Protocols for detailed instructions on how to use this page.

Error parsing request, no input specified

Help Tips

Steps:

1. Select list and list type to analyze
2. Select Organism
3. Select operation

1. Enter IDs and or select file for batch upload. Else enter ids or select file or list from workspace for comparing to a reference list.

Enter IDs: Supported IDs

Upload IDs: File format

Select List Type:

Choose File no file selected

Please login to be able to select lists from your workspace.

ID List

Previously exported text search results

Workspace list

PANTHER Generic Mapping File

VCF File Flanking region 20 Kb

2. Select organism.

Homo sapiens

Mus musculus

Rattus norvegicus

Gallus gallus

Danio rerio

3. Select Analysis.

Functional classification viewed in gene list

Another popular online tool: **DAVID** at NIAID < david.abcc.ncifcrf.gov >

DAVID

• Functional Annotation Chart

Functional Annotation Chart [Help and Manual](#)

Current Gene List: Uploaded List_1
Current Background: Homo sapiens
2316 DAVID IDs

☐ Options [Download File](#)

[Rerun Using Options](#) [Create Sublist](#)

| Sublist | Category | Term | RT | Gene | Count | % | P-Value | Benjamini |
|--------------------------|-------------|--|----|------|-------|-----|---------|-----------|
| <input type="checkbox"/> | GOTERM_BP_5 | regulation of progression through cell cycle | RT | | 98 | 4.2 | 3.3E-7 | 8.6E-4 |
| <input type="checkbox"/> | GOTERM_BP_5 | apoptosis | RT | | 131 | 5.7 | 1.6E-6 | 2.1E-3 |
| <input type="checkbox"/> | GOTERM_BP_5 | cell death | RT | | 136 | 5.9 | 3.8E-6 | 3.3E-3 |
| <input type="checkbox"/> | GOTERM_BP_5 | regulation of transcription from RNA polymerase II promoter | RT | | 83 | 3.6 | 3.7E-5 | 2.4E-2 |
| <input type="checkbox"/> | GOTERM_BP_5 | protein kinase cascade | RT | | 71 | 3.1 | 4.7E-5 | 2.4E-2 |
| <input type="checkbox"/> | GOTERM_BP_5 | regulation of kinase activity | RT | | 48 | 2.1 | 5.4E-5 | 2.3E-2 |
| <input type="checkbox"/> | GOTERM_BP_5 | negative regulation of cell proliferation | RT | | 48 | 2.1 | 1.0E-4 | 3.7E-2 |
| <input type="checkbox"/> | GOTERM_BP_5 | regulation of cell size | RT | | 41 | 1.8 | 1.2E-4 | 3.9E-2 |
| <input type="checkbox"/> | GOTERM_BP_5 | monocarboxylic acid metabolic process | RT | | 48 | 2.1 | 1.3E-4 | 3.6E-2 |
| <input type="checkbox"/> | GOTERM_BP_5 | positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process | RT | | 61 | 2.6 | 1.5E-4 | 3.8E-2 |
| <input type="checkbox"/> | GOTERM_BP_5 | positive regulation of cellular metabolic process | RT | | 72 | 3.1 | 1.7E-4 | 3.8E-2 |

Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources
Da Wei Huang, Brad T Sherman & Richard A Lempicki
Nature Protocols **4**, 44 - 57 (2009)

Overlapping functional sets

• Many functional sets overlap

- In particular those from databases that are hierarchical in nature (e.g. GO)

• Hierarchy enables:

- Annotation flexibility (e.g. allow different degrees of annotation completeness based on what is known)
- Computational methods to “understand” function relationships (e.g. ATPase function is a subset of enzyme function)

• Unfortunately, this also makes functional profiling trickier

- Clustering of functional sets can be helpful in these cases

DAVID

• DAVID now offers functional annotation clustering:

Annotation Summary Results [Help and Tool Manual](#)

Current Gene List: Uploaded List_3
Current Background: HOMO SAPIENS
2320 DAVID IDs
Check Defaults ☒ [Clear All](#)

☐ Main Accessions (0 selected)
☐ Other Accessions (0 selected)
☐ Gene Ontology (4 selected)
☐ Protein Domains (3 selected)
☐ Pathways (3 selected)
☐ General Annotations (0 selected)
☐ Functional Categories (3 selected)
☐ Protein Interactions (0 selected)
☐ Literature (0 selected)
☐ Disease (1 selected)
☐ Tissue Expression

Combined View for Selected Annotation

[Functional Annotation Clustering^{new!}](#) [Functional Annotation Chart](#) [Functional Annotation Table](#)

DAVID Functional Annotation Clustering

- Based on shared genes between functional sets

Functional Annotation Clustering

Current Gene List: Uploaded List_3
2320 DAVID IDs

Options: Classification Stringency: Medium

Download File

| Annotation Cluster | Enrichment Score | Count | P-Value | Benjamini |
|-----------------------------|--|-------|---------|-----------|
| Annotation Cluster 1 | Enrichment Score: 3.72 | | | |
| GOTERM_BP_5 | regulation of transcription from RNA polymerase II promoter | 83 | 3.7E-5 | 2.4E-2 |
| GOTERM_BP_5 | positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process | 61 | 1.5E-4 | 3.8E-2 |
| GOTERM_BP_5 | positive regulation of cellular metabolic process | 72 | 1.7E-4 | 3.8E-2 |
| GOTERM_BP_5 | positive regulation of transcription | 58 | 3.8E-4 | 5.0E-2 |
| GOTERM_BP_5 | positive regulation of transcription, DNA-dependent | 48 | 7.4E-4 | 7.6E-2 |
| Annotation Cluster 2 | Enrichment Score: 3.54 | | | |
| GOTERM_BP_5 | regulation of cell size | 41 | 1.2E-4 | 3.9E-2 |
| GOTERM_BP_5 | regulation of cell growth | 33 | 3.7E-4 | 5.1E-2 |
| GOTERM_BP_5 | cell morphogenesis | 81 | 5.2E-4 | 5.7E-2 |
| Annotation Cluster 3 | Enrichment Score: 3.37 | | | |
| GOTERM_BP_5 | apoptosis | 131 | 1.6E-6 | 2.1E-3 |
| GOTERM_BP_5 | cell death | 136 | 3.8E-6 | 3.3E-3 |
| GOTERM_BP_5 | regulation of programmed cell death | 88 | 3.2E-4 | 5.8E-2 |
| GOTERM_BP_5 | positive regulation of apoptosis | 48 | 3.5E-4 | 5.6E-2 |
| GOTERM_BP_5 | regulation of apoptosis | 87 | 3.5E-4 | 5.2E-2 |
| GOTERM_BP_5 | positive regulation of programmed cell death | 48 | 4.0E-4 | 5.0E-2 |

Want more?

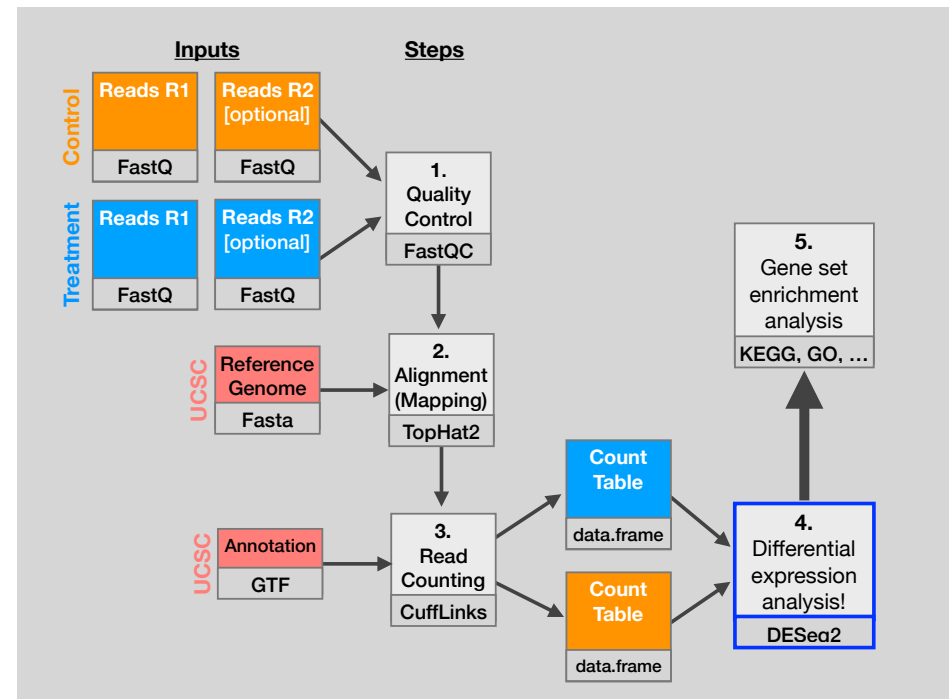


- GeneGO** < portal.genego.com >
 - MD/PhD curated annotations, great for certain domains (eg, Cystic Fibrosis)
 - Nice network analysis tools
 - Email us for access
- Oncomine** < www.oncomine.org >
 - Extensive cancer related expression datasets
 - Nice concept analysis tools
 - Research edition is free for academics, Premium edition \$\$\$
- Lots and lots other R/Bioconductor packages in this area!!!**

Hands-on time!

<http://thegrantlab.org/bgg213>

Do it Yourself!



counts + metadata

1 countData

| gene | ctrl_1 | ctrl_2 | exp_1 | exp_2 |
|-------|--------|--------|-------|-------|
| geneA | 10 | 11 | 56 | 45 |
| geneB | 0 | 0 | 128 | 54 |
| geneC | 42 | 41 | 59 | 41 |
| geneD | 103 | 122 | 1 | 23 |
| geneE | 10 | 23 | 14 | 56 |
| geneF | 0 | 1 | 2 | 0 |
| ... | ... | ... | ... | ... |

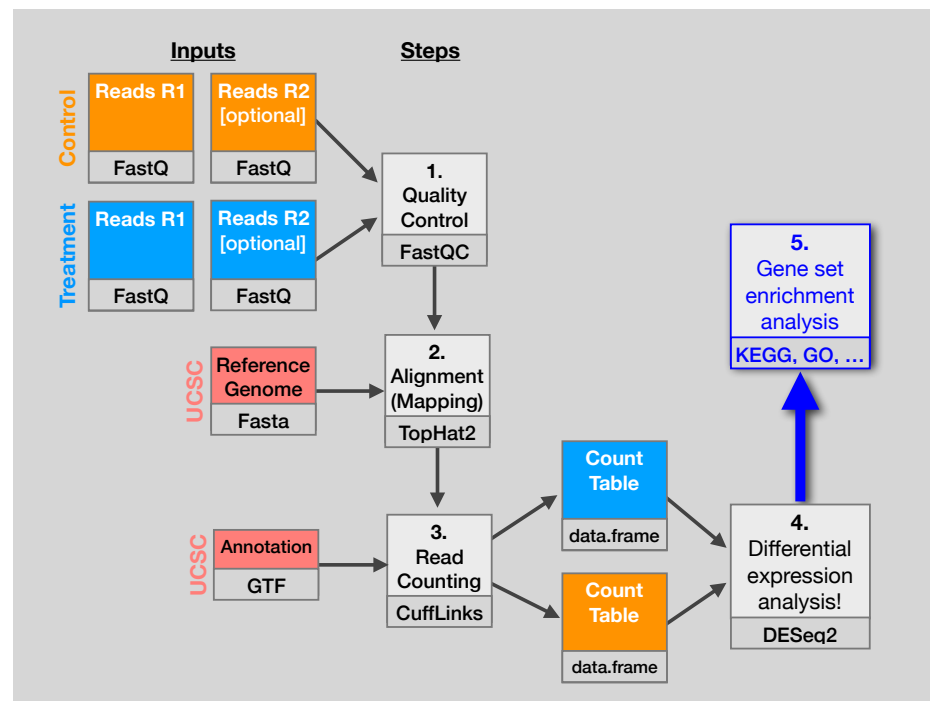
countData is the count matrix
(Number of reads coming from each gene for each sample)

2 colData

| id | treatment | sex | ... |
|--------|-----------|--------|-----|
| ctrl_1 | control | male | ... |
| ctrl_2 | control | female | ... |
| exp_1 | treated | male | ... |
| exp_2 | treated | female | ... |

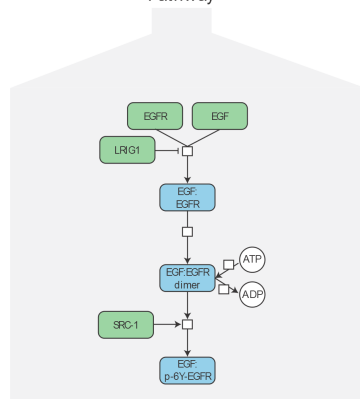
colData describes metadata about the columns of countData

N.B. First column of **colData** must match column names (i.e. **sample names**) of countData (-1st)



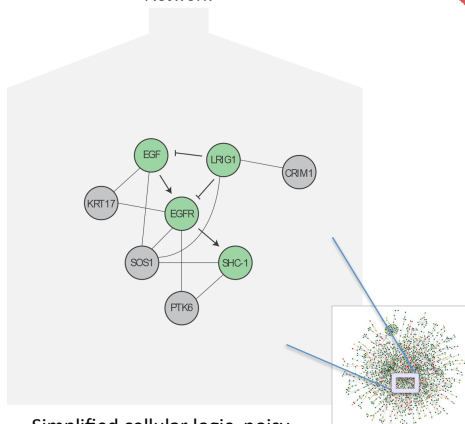
Pathways vs Networks

EGFR-centered Pathway



- Detailed, high-confidence consensus
- Biochemical reactions
- Small-scale, fewer genes
- Concentrated from decades of literature

EGFR-centered Network



- Simplified cellular logic, noisy
- Abstractions: directed, undirected
- Large-scale, genome-wide
- Constructed from *omics* data integration

Next Class

Goal

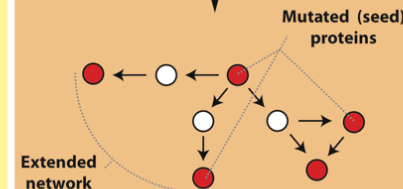
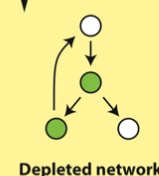
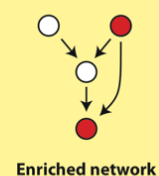
1 Enrichment of fixed gene sets

Identification of pre-built pathways or networks that are enriched in a set of mutated or differentially expressed genes

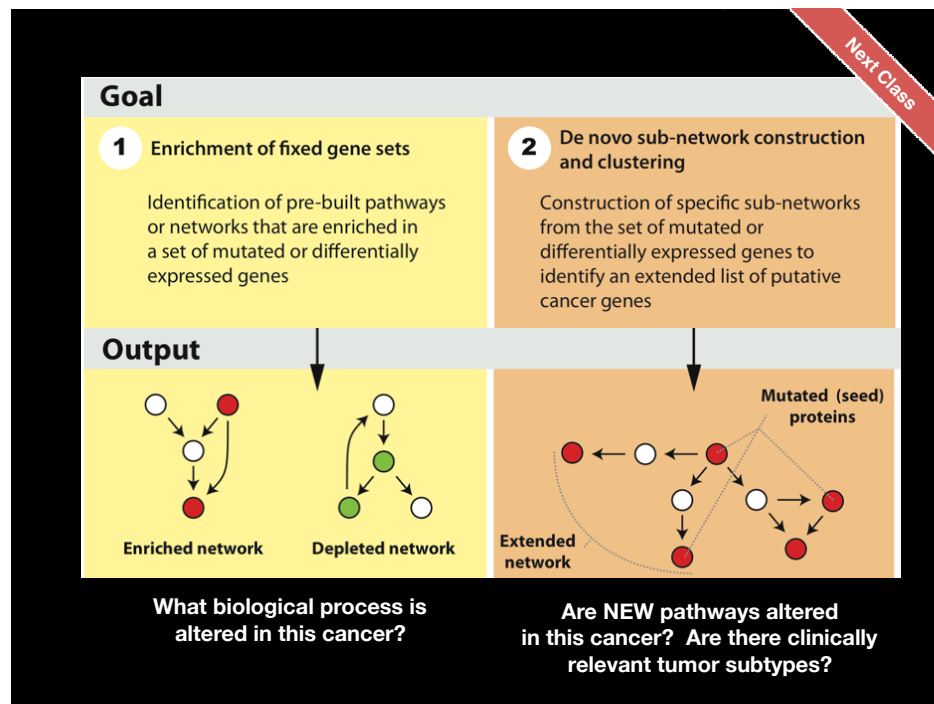
2 De novo sub-network construction and clustering

Construction of specific sub-networks from the set of mutated or differentially expressed genes to identify an extended list of putative cancer genes

Output



Next Class



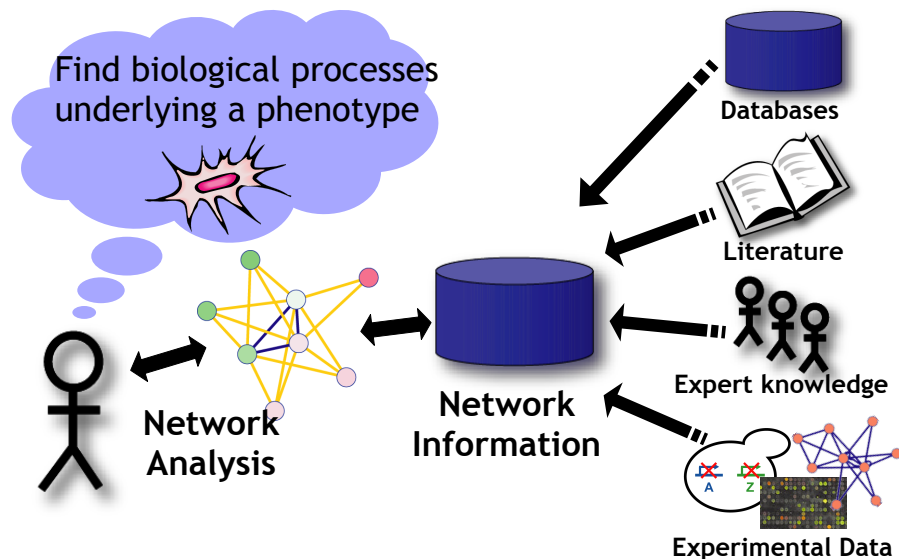
Pathway analysis (a.k.a. geneset enrichment)

Limitations

Side-note:

- **Geneset annotation bias:** can only discover what is already known
- **Non-model organisms:** no high-quality genesets available
- **Post-transcriptional regulation** is neglected
- **Tissue-specific** variations of pathways are not annotated
 - e.g. NF- κ B regulates metabolism, not inflammation, in adipocytes
- **Size bias:** stats are influenced by the size of the pathway
 - Many pathways/receptors **converge** to few regulators
e.g. Tens of innate immune receptors activate four TFs: NF- κ B, AP-1, IRF3/7, NFAT

Pathway & Network Analysis Overview



R Knowledge Check For BGGN-213 Quiz

This will be marked but not graded
(i.e. will not factor into your course grade)

Time Limit: 30 mins

Do it Yourself!

Your results will be e-mailed to you upon quiz close

Thank you!