1. Quality Control
   - FastQC

2. Alignment (Mapping)
   - TopHat2

3. Read Counting
   - CuffLinks
   - Count Table (data.frame)

4. Differential expression analysis!
   - DESeq2

**Inputs**
- **Control**
  - Reads R1: FastQ
  - Reads R2 [optional]: FastQ

- **Treatment**
  - Reads R1: FastQ
  - Reads R2 [optional]: FastQ

**Reference Genome**
- UCSC Fasta

**Annotation**
- UCSC GTF

**Last day’s step**
- Differential expression analysis!
  - DESeq2
<table>
<thead>
<tr>
<th>Gene ID</th>
<th>baseMean</th>
<th>log2FoldChange</th>
<th>lfcSE</th>
<th>stat</th>
<th>pvalue</th>
<th>padj</th>
<th>symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000152583</td>
<td>954.77093</td>
<td>4.3683590</td>
<td>0.23713648</td>
<td>18.421286</td>
<td>8.867079e-76</td>
<td>1.342919e-71</td>
<td>SPARCL1</td>
</tr>
<tr>
<td>ENSG00000179094</td>
<td>743.25269</td>
<td>2.8638885</td>
<td>0.17555825</td>
<td>16.313039</td>
<td>7.972621e-60</td>
<td>6.037267e-56</td>
<td>PER1</td>
</tr>
<tr>
<td>ENSG00000116584</td>
<td>2277.91345</td>
<td>-1.0347000</td>
<td>0.06505273</td>
<td>-15.905557</td>
<td>5.798513e-57</td>
<td>2.927283e-53</td>
<td>ARHGEF2</td>
</tr>
<tr>
<td>ENSG00000189221</td>
<td>2383.75371</td>
<td>3.3415441</td>
<td>0.21241508</td>
<td>15.731200</td>
<td>9.244206e-56</td>
<td>3.500088e-52</td>
<td>MAOA</td>
</tr>
<tr>
<td>ENSG00000120129</td>
<td>3440.70375</td>
<td>2.9652108</td>
<td>0.20370277</td>
<td>14.556557</td>
<td>5.306416e-48</td>
<td>1.607313e-44</td>
<td>DUSP1</td>
</tr>
<tr>
<td>ENSG00000148175</td>
<td>13493.92037</td>
<td>1.4271683</td>
<td>0.10036663</td>
<td>14.219550</td>
<td>6.929711e-46</td>
<td>1.749175e-42</td>
<td>STOM</td>
</tr>
<tr>
<td>ENSG00000178695</td>
<td>2685.40974</td>
<td>-2.4890689</td>
<td>0.17806407</td>
<td>-13.978501</td>
<td>2.108817e-44</td>
<td>4.562576e-41</td>
<td>KCTD12</td>
</tr>
<tr>
<td>ENSG00000109906</td>
<td>439.54152</td>
<td>5.9275950</td>
<td>0.42819442</td>
<td>13.843233</td>
<td>1.397758e-43</td>
<td>2.646131e-40</td>
<td>ZBTB16</td>
</tr>
<tr>
<td>ENSG00000134686</td>
<td>2933.64246</td>
<td>1.4394898</td>
<td>0.10582729</td>
<td>13.602255</td>
<td>3.882769e-42</td>
<td>6.533883e-39</td>
<td>PHC2</td>
</tr>
<tr>
<td>ENSG00000101347</td>
<td>14134.99177</td>
<td>3.8504143</td>
<td>0.28490701</td>
<td>13.514635</td>
<td>1.281894e-41</td>
<td>1.941428e-38</td>
<td>SAMHD1</td>
</tr>
<tr>
<td>ENSG00000096060</td>
<td>2630.23049</td>
<td>3.9450524</td>
<td>0.29291821</td>
<td>13.468102</td>
<td>2.409807e-41</td>
<td>3.317866e-38</td>
<td>FKBP5</td>
</tr>
<tr>
<td>ENSG00000166741</td>
<td>7542.25287</td>
<td>2.2195906</td>
<td>0.16673544</td>
<td>13.312050</td>
<td>1.970000e-40</td>
<td>2.486304e-37</td>
<td>NNMT</td>
</tr>
<tr>
<td>ENSG00000125148</td>
<td>3695.87946</td>
<td>2.1985636</td>
<td>0.16700546</td>
<td>13.164621</td>
<td>1.402400e-39</td>
<td>1.633797e-36</td>
<td>MT2A</td>
</tr>
<tr>
<td>ENSG00000162614</td>
<td>5646.18314</td>
<td>1.9711402</td>
<td>0.15020631</td>
<td>13.122885</td>
<td>2.434854e-39</td>
<td>2.633990e-36</td>
<td>NEXN</td>
</tr>
<tr>
<td>ENSG00000106976</td>
<td>989.04683</td>
<td>-1.8501713</td>
<td>0.14778657</td>
<td>-12.519211</td>
<td>5.861471e-36</td>
<td>5.918132e-33</td>
<td>DNMT1</td>
</tr>
<tr>
<td>ENSG00000187193</td>
<td>199.07694</td>
<td>3.2551424</td>
<td>0.26090711</td>
<td>12.476250</td>
<td>1.006146e-35</td>
<td>9.523804e-33</td>
<td>MT1X</td>
</tr>
<tr>
<td>ENSG00000256235</td>
<td>1123.47954</td>
<td>1.2801193</td>
<td>0.10547438</td>
<td>12.136779</td>
<td>6.742862e-34</td>
<td>6.007096e-31</td>
<td>SMIM3</td>
</tr>
<tr>
<td>ENSG00000177666</td>
<td>2639.57020</td>
<td>1.1399947</td>
<td>0.09606884</td>
<td>11.866436</td>
<td>1.768422e-32</td>
<td>1.487930e-29</td>
<td>PNPLA2</td>
</tr>
<tr>
<td>ENSG00000164125</td>
<td>7257.00808</td>
<td>1.0248523</td>
<td>0.08657600</td>
<td>11.837603</td>
<td>2.494830e-32</td>
<td>1.988642e-29</td>
<td>FAM198B</td>
</tr>
<tr>
<td>ENSG00000198624</td>
<td>2020.04495</td>
<td>2.8141014</td>
<td>0.24063429</td>
<td>11.694515</td>
<td>1.359615e-31</td>
<td>1.029569e-28</td>
<td>CCDC69</td>
</tr>
<tr>
<td>ENSG00000123562</td>
<td>5008.55294</td>
<td>1.0045453</td>
<td>0.08901501</td>
<td>11.285123</td>
<td>1.554241e-29</td>
<td>1.120904e-26</td>
<td>MORF4L2</td>
</tr>
<tr>
<td>ENSG00000144369</td>
<td>1283.77980</td>
<td>-1.3090041</td>
<td>0.11714863</td>
<td>-11.173875</td>
<td>5.473974e-29</td>
<td>3.768333e-26</td>
<td>FAM171B</td>
</tr>
<tr>
<td>ENSG00000196517</td>
<td>241.91536</td>
<td>-2.3456877</td>
<td>0.21047366</td>
<td>-11.144804</td>
<td>7.591120e-29</td>
<td>4.998588e-26</td>
<td>SLC6A9</td>
</tr>
<tr>
<td>ENSG00000135821</td>
<td>19973.40000</td>
<td>3.0413943</td>
<td>0.27601796</td>
<td>11.018828</td>
<td>3.100706e-28</td>
<td>1.956675e-25</td>
<td>GLUL</td>
</tr>
</tbody>
</table>
Volcano Plot
Fold change vs P-value

Significant
(P < 0.01 & log2 > 2)
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…
**Steps**

1. Quality Control
   - FastQC

2. Alignment (Mapping)
   - TopHat2

3. Read Counting
   - CuffLinks

4. Differential expression analysis!
   - DESeq2

5. Gene set enrichment analysis
   - KEGG, GO, …

**Inputs**

- **Control**
  - Reads R1
    - FastQ
  - Reads R2 [optional]
    - FastQ

- **Treatment**
  - Reads R1
    - FastQ
  - Reads R2 [optional]
    - FastQ

- UCSC
  - Reference Genome
    - Fasta

- UCSC
  - Annotation
    - GTF
Basic idea

Differentially Expressed Genes (DEGs)

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Log2FoldChange</th>
<th>P.Value</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG0000152183</td>
<td>4.3683590</td>
<td>0.2371648</td>
<td>SPARC1</td>
</tr>
<tr>
<td>ENSG00001770094</td>
<td>7.4325699</td>
<td>2.8638885</td>
<td>PER1</td>
</tr>
<tr>
<td>ENSG00001181584</td>
<td>2.277.91345</td>
<td>-1.0347000</td>
<td>ABKG2F</td>
</tr>
<tr>
<td>ENSG00001849221</td>
<td>2.383.75371</td>
<td>3.3415441</td>
<td>MAGA</td>
</tr>
<tr>
<td>ENSG00001512020</td>
<td>2.6492108</td>
<td>2.0397071</td>
<td>DIUSF1</td>
</tr>
<tr>
<td>ENSG00001484175</td>
<td>3.449.26297</td>
<td>1.0423681</td>
<td>STOM</td>
</tr>
<tr>
<td>ENSG00001786905</td>
<td>2.685.40974</td>
<td>-1.2890689</td>
<td>KCTD12</td>
</tr>
<tr>
<td>ENSG00001090906</td>
<td>4.393.54132</td>
<td>5.0279950</td>
<td>ZBTB16</td>
</tr>
<tr>
<td>ENSG00001468568</td>
<td>2.933.64246</td>
<td>1.0364768</td>
<td>PHC2</td>
</tr>
<tr>
<td>ENSG00001013147</td>
<td>1.614.99177</td>
<td>3.8504143</td>
<td>SAMHD1</td>
</tr>
<tr>
<td>ENSG00001092050</td>
<td>2.682.20349</td>
<td>3.9450524</td>
<td>PEP5</td>
</tr>
<tr>
<td>ENSG00001468474</td>
<td>7.541.25327</td>
<td>2.2339590</td>
<td>NMI</td>
</tr>
<tr>
<td>ENSG00001211485</td>
<td>3.695.87946</td>
<td>2.1985636</td>
<td>MTAD</td>
</tr>
<tr>
<td>ENSG00001602197</td>
<td>5.646.18314</td>
<td>1.9711402</td>
<td>NNK</td>
</tr>
<tr>
<td>ENSG00001068976</td>
<td>9.898.04683</td>
<td>1.8501713</td>
<td>DNM1</td>
</tr>
<tr>
<td>ENSG00001871939</td>
<td>199.70694</td>
<td>3.2551424</td>
<td>MTIX</td>
</tr>
<tr>
<td>ENSG00002562335</td>
<td>1.123.47954</td>
<td>1.2801139</td>
<td>SMW3</td>
</tr>
<tr>
<td>ENSG00001776686</td>
<td>2.693.75070</td>
<td>1.1930947</td>
<td>PNPX2</td>
</tr>
<tr>
<td>ENSG00001641428</td>
<td>7.577.08090</td>
<td>1.0248523</td>
<td>FAM18B</td>
</tr>
<tr>
<td>ENSG00001908064</td>
<td>2.202.04491</td>
<td>2.8140141</td>
<td>CCDC99</td>
</tr>
<tr>
<td>ENSG00001235652</td>
<td>5.008.55294</td>
<td>1.0045453</td>
<td>MIF4G1L2</td>
</tr>
<tr>
<td>ENSG00001416199</td>
<td>1283.77980</td>
<td>1.3090041</td>
<td>SIKCA9</td>
</tr>
<tr>
<td>ENSG00001969117</td>
<td>241.91356</td>
<td>2.3456877</td>
<td>FLJ</td>
</tr>
</tbody>
</table>

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

Annotation...
Basic idea

Differentially Expressed Genes (DEGs)

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>baseMean</th>
<th>log2FoldChange</th>
<th>adjPvalue</th>
<th>symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN000000152183</td>
<td>954.7709</td>
<td>4.368359</td>
<td>0.23713648</td>
<td>SPARC1</td>
</tr>
<tr>
<td>EN000000178904</td>
<td>743.2526</td>
<td>2.863885</td>
<td>0.17555825</td>
<td>PER1</td>
</tr>
<tr>
<td>EN000000185814</td>
<td>2277.9134</td>
<td>0.10347006</td>
<td>0.06505273</td>
<td>ABK</td>
</tr>
<tr>
<td>EN000000189221</td>
<td>2384.7551</td>
<td>3.341544</td>
<td>0.21410518</td>
<td>MAGA</td>
</tr>
<tr>
<td>EN000000191229</td>
<td>3440.7037</td>
<td>2.963108</td>
<td>0.03701377</td>
<td>DUSP1</td>
</tr>
<tr>
<td>EN000000144175</td>
<td>13493.8293</td>
<td>1.472168</td>
<td>0.00106663</td>
<td>STAM</td>
</tr>
<tr>
<td>EN000000178695</td>
<td>2685.4097</td>
<td>2.489068</td>
<td>0.17806407</td>
<td>CDK2</td>
</tr>
<tr>
<td>EN000000109906</td>
<td>439.5412</td>
<td>5.927950</td>
<td>0.28191443</td>
<td>CDK12</td>
</tr>
<tr>
<td>EN000000134686</td>
<td>2933.6426</td>
<td>1.439489</td>
<td>0.10582729</td>
<td>CDK12</td>
</tr>
<tr>
<td>EN000000101134</td>
<td>4114.9517</td>
<td>3.850413</td>
<td>0.28405710</td>
<td>SAMHD1</td>
</tr>
<tr>
<td>EN000000090605</td>
<td>2630.2304</td>
<td>3.945052</td>
<td>0.29218212</td>
<td>FER1BP3</td>
</tr>
<tr>
<td>EN000000164743</td>
<td>7541.2337</td>
<td>2.139500</td>
<td>0.06371444</td>
<td>HTR2C</td>
</tr>
<tr>
<td>EN000000125148</td>
<td>3895.8794</td>
<td>2.198536</td>
<td>0.01700546</td>
<td>HTR2C</td>
</tr>
<tr>
<td>EN000000163014</td>
<td>5646.1834</td>
<td>1.971140</td>
<td>0.15020631</td>
<td>ND11</td>
</tr>
<tr>
<td>EN000000106976</td>
<td>989.0463</td>
<td>1.850173</td>
<td>0.14778687</td>
<td>ND11</td>
</tr>
<tr>
<td>EN000000187193</td>
<td>199.0769</td>
<td>3.255142</td>
<td>0.26090711</td>
<td>MTLX</td>
</tr>
<tr>
<td>EN000000256235</td>
<td>1123.4795</td>
<td>1.280113</td>
<td>0.10547438</td>
<td>MTLX</td>
</tr>
<tr>
<td>EN000000177686</td>
<td>2630.1700</td>
<td>1.119946</td>
<td>0.09668484</td>
<td>PDK2</td>
</tr>
<tr>
<td>EN000000144312</td>
<td>737.0080</td>
<td>1.024653</td>
<td>0.06371444</td>
<td>PFKFB2</td>
</tr>
<tr>
<td>EN000000192624</td>
<td>2020.0495</td>
<td>2.814314</td>
<td>0.24063429</td>
<td>PFKFB2</td>
</tr>
<tr>
<td>EN000000123562</td>
<td>5008.5259</td>
<td>1.034543</td>
<td>0.09017152</td>
<td>STK31</td>
</tr>
<tr>
<td>EN000000144369</td>
<td>1283.7790</td>
<td>1.309094</td>
<td>0.11748683</td>
<td>STK31</td>
</tr>
<tr>
<td>EN000000196517</td>
<td>241.9513</td>
<td>2.345687</td>
<td>0.21047366</td>
<td>STK31</td>
</tr>
<tr>
<td>DEG000000111831</td>
<td>1093.4000</td>
<td>3.041324</td>
<td>0.27017286</td>
<td>STK31</td>
</tr>
</tbody>
</table>

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

Annotate...

Pathway analysis (geneset enrichment)

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

**Principle**

- **DEGs** come from your experiment
- **Pathway genes (“geneset”)** come from annotations
- Variations of the math: overlap, ranking, networks...
  - Critical, needs to be as clean as possible
  - Important, but typically not a competitive advantage
  - Not critical, different algorithms show similar performances
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-κB 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-kB, AP-1, IRF3/7, NFAT
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
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.bioinfo.cnio.es >
Translating between identifiers:
UniProt < www.uniprot.org >
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 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)
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]
# Using the merge() function

```r
> anno <- read.csv("data/annotables_grch38.csv")

> merge(mygenes, anno, by.x="row.names", by.y="ensgene")
```

This is an annotation file

This is our differential expressed genes
# Using the merge() function
> anno <- read.csv("data/annotables_grch38.csv")

> merge(mygenes, anno, by.x="row.names", by.y= "ensgene")

# Using 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" )
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.

```r

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:
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](http://www.obofoundry.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”

Parent terms are more general & child terms more specific
## GO evidence codes

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

*October 2007 release

---

**Use and misuse of the gene ontology annotations**
Seung Yon Rhee, Valerie Wood, Kara Dolinski & Sorin Draghici
• See AmiGO for details: http://amigo.geneontology.org/amigo/base_statistics
Can now do gene list analysis with GeneGO online!

Gene List Analysis

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

Error parsing request, no input specified

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

   Choose File □ no file selected

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

   Select List Type:
   □ 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
   □ Enrichment analysis if functional classification available

Help Tips
Steps:
  ▸ 1. Select list and list type to analyze
  ▸ 2. Select Organism
  ▸ 3. Select operation
Another popular online tool:
DAVID at NIAID < david.abcc.ncifcrf.gov >
DAVID

• Functional Annotation Chart

Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources
Da Wei Huang, Brad T Sherman & Richard A Lempicki
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:
DAVID Functional Annotation Clustering

- Based on shared genes between functional sets
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!

https://bioboot.github.io/bggn213_S19/lectures/#15
Reads R1
FastQ

Reads R2
[optional]
FastQ

Quality Control
FastQC

1. Alignment (Mapping)
TopHat2

2. Read Counting
CuffLinks

3. Count Table
data.frame

4. Differential expression analysis
DESeq2

5. Gene set enrichment analysis
KEGG, GO, ...

Control
FastQ

Treatment
FastQ

Reference Genome
UCSC Fasta

Annotation
UCSC GTF
**counts + metadata**

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

<table>
<thead>
<tr>
<th>gene</th>
<th>ctrl_1</th>
<th>ctrl_2</th>
<th>exp_1</th>
<th>exp_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>geneA</td>
<td>10</td>
<td>11</td>
<td>56</td>
<td>45</td>
</tr>
<tr>
<td>geneB</td>
<td>0</td>
<td>0</td>
<td>128</td>
<td>54</td>
</tr>
<tr>
<td>geneC</td>
<td>42</td>
<td>41</td>
<td>59</td>
<td>41</td>
</tr>
<tr>
<td>geneD</td>
<td>103</td>
<td>122</td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>geneE</td>
<td>10</td>
<td>23</td>
<td>14</td>
<td>56</td>
</tr>
<tr>
<td>geneF</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**colData** describes metadata about the columns of countData

<table>
<thead>
<tr>
<th>id</th>
<th>treatment</th>
<th>sex</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>ctrl_1</td>
<td>control</td>
<td>male</td>
<td></td>
</tr>
<tr>
<td>ctrl_2</td>
<td>control</td>
<td>female</td>
<td></td>
</tr>
<tr>
<td>exp_1</td>
<td>treated</td>
<td>male</td>
<td></td>
</tr>
<tr>
<td>exp_2</td>
<td>treated</td>
<td>female</td>
<td></td>
</tr>
</tbody>
</table>

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

2. Alignment (Mapping)
   TopHat2

3. Read Counting
   CuffLinks

4. Differential expression analysis!
   DESeq2

5. Gene set enrichment analysis
   KEGG, GO, …
Pathways vs Networks

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

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

- **Enriched network**
- **Depleted network**
- **Extended network**

**Mutated (seed) proteins**
Module 12: Introduction to Pathway and Network Analysis

Types of Pathway/Network Analysis

What biological processes are altered in this cancer? Are new pathways altered in this cancer? Are there clinically-relevant tumor subtypes? How are pathway activities altered in a particular patient? Are there targetable pathways in this patient?

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

Enriched network
Depleted network

Extended network
Mutated (seed) proteins

What biological process is altered in this cancer? Are NEW pathways altered in this cancer? Are there clinically relevant tumor subtypes?
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-κB 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-κB, AP-1, IRF3/7, NFAT
Find biological processes underlying a phenotype.

- Network Analysis
- Network Information
- Databases
- Literature
- Expert knowledge
- Experimental Data
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: 40 mins