Understanding protein lists from proteomics studies

Bing Zhang
Department of Biomedical Informatics
Vanderbilt University
bing.zhang@vanderbilt.edu
A typical comparative shotgun proteomics study

Li et al. JPR, 2010
Omics technologies generate gene/protein lists

- Genomics
  - Genome Wide Association Study (GWAS)
  - Next generation sequencing (NGS)
- Transcriptomics
  - mRNA profiling
    - Microarrays
    - Serial analysis of gene expression (SAGE)
    - RNA-Seq
  - Protein-DNA interaction
    - Chromatin immunoprecipitation
- Proteomics
  - Protein profiling
    - LC-MS/MS
  - Protein-protein interaction
    - Yeast two hybrid
    - Affinity pull-down/LC-MS/MS
Sample files

- Samples files can be downloaded from
  - http://bioinfo.vanderbilt.edu/zhanglab/?q=node/410

- Significant proteins
  - hnscc_sig_proteins.txt

- Significant proteins with log fold change
  - hnscc_sig_withLogRatio.txt

- All proteins identified in the study
  - hnscc_all_proteins.txt
Understanding a protein list

■ Level I

□ What are the proteins/genes behind the IDs and what do we know about the functions of the proteins/genes?
Level one: information retrieval

- One-protein-at-a-time
- Time consuming
- Information is local and isolated
- Hard to automate the information retrieval process

Query interface (http://www.ebi.ac.uk/IPI)  Output
I’ve attached a spreadsheet of our proteomics results comparing 5 Vehicle and 5 Aldosterone treated patients. We’ve included only those proteins whose summed spectral counts are >30 in one treatment group. Would it be possible to get the GO annotations for these? The Uniprot name is listed in column A and the gene name is listed in column R. If this is a time consuming task (and I imagine that it is), can you tell me how to do it?”
Biomart: a batch information retrieval system

http://central.biomart.org/
Biomart analysis

Choose dataset

- Choose database: Ensembl 75 Genes
- Choose dataset: *Homo sapiens* genes (GRCh37.p13)

Set filters

- Gene: a list of genes identified by various database IDs (e.g. IPI IDs)
- Gene Ontology: filter for genes with specific GO terms (e.g. cell cycle)
- Protein domains: filter for genes with specific protein domains (e.g. SH2 domain, signal domains)
- Region: filter for genes in a specific chromosome region (e.g. chr1 1:1000000 or 11q13)
- Others

Select output attributes

- Gene annotation information in the Ensembl database, e.g. gene description, chromosome name, gene start, gene end, strand, band, gene name, etc.
- External data: Gene Ontology, IDs in other databases
- Expression: anatomical system, development stage, cell type, pathology
- Protein domains: SMART, PFAM, Interpro, etc.
**Biomart: sample output**

Biomart Central Portal

<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
<th>HGNC Symbol</th>
<th>GO Term Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP00004924</td>
<td>differentially expressed in FDCP 6 homolog [mouse] [Source:HGNC Symbol:Acc:2760]</td>
<td>DEP6</td>
<td>nucleus</td>
</tr>
<tr>
<td>IP00004924</td>
<td>differentially expressed in FDCP 6 homolog [mouse] [Source:HGNC Symbol:Acc:2760]</td>
<td>DEP6</td>
<td>cytoplasm</td>
</tr>
<tr>
<td>IP00004924</td>
<td>differentially expressed in FDCP 6 homolog [mouse] [Source:HGNC Symbol:Acc:2760]</td>
<td>DEP6</td>
<td>plasma membrane</td>
</tr>
<tr>
<td>IP00004924</td>
<td>differentially expressed in FDCP 6 homolog [mouse] [Source:HGNC Symbol:Acc:2760]</td>
<td>DEP6</td>
<td>phospholipid binding</td>
</tr>
<tr>
<td>IP00004924</td>
<td>differentially expressed in FDCP 6 homolog [mouse] [Source:HGNC Symbol:Acc:2760]</td>
<td>DEP6</td>
<td>protein binding</td>
</tr>
<tr>
<td>IP00006304</td>
<td>cysteine-rich protein 2 [Source:HGNC Symbol:Acc:2361]</td>
<td>CRIP2</td>
<td>positive regulation of cell proliferation</td>
</tr>
<tr>
<td>IP00006304</td>
<td>cysteine-rich protein 2 [Source:HGNC Symbol:Acc:2361]</td>
<td>CRIP2</td>
<td>hemorrhage</td>
</tr>
<tr>
<td>IP00006304</td>
<td>cysteine-rich protein 2 [Source:HGNC Symbol:Acc:2361]</td>
<td>CRIP2</td>
<td>extracellular matrix</td>
</tr>
<tr>
<td>IP00006304</td>
<td>cysteine-rich protein 2 [Source:HGNC Symbol:Acc:2361]</td>
<td>CRIP2</td>
<td>cell cortex</td>
</tr>
<tr>
<td>IP00006304</td>
<td>cysteine-rich protein 2 [Source:HGNC Symbol:Acc:2361]</td>
<td>CRIP2</td>
<td>zinc ion binding</td>
</tr>
<tr>
<td>IP00006304</td>
<td>cysteine-rich protein 2 [Source:HGNC Symbol:Acc:2361]</td>
<td>CRIP2</td>
<td>protein binding</td>
</tr>
<tr>
<td>IP00009991</td>
<td>cyclin-dependent kinase inhibitor 18 [p27, Kip1] [Source:HGNC Symbol:Acc:1766]</td>
<td>CKIN1B</td>
<td>potassium ion transport</td>
</tr>
<tr>
<td>IP00009991</td>
<td>cyclin-dependent kinase inhibitor 18 [p27, Kip1] [Source:HGNC Symbol:Acc:1766]</td>
<td>CKIN1B</td>
<td>negative regulation of cell proliferation</td>
</tr>
<tr>
<td>IP00009991</td>
<td>cyclin-dependent kinase inhibitor 18 [p27, Kip1] [Source:HGNC Symbol:Acc:1766]</td>
<td>CKIN1B</td>
<td>negative regulation of cell proliferation</td>
</tr>
<tr>
<td>IP00009991</td>
<td>cyclin-dependent kinase inhibitor 18 [p27, Kip1] [Source:HGNC Symbol:Acc:1766]</td>
<td>CKIN1B</td>
<td>negative regulation of cell proliferation</td>
</tr>
<tr>
<td>IP00009991</td>
<td>cyclin-dependent kinase inhibitor 18 [p27, Kip1] [Source:HGNC Symbol:Acc:1766]</td>
<td>CKIN1B</td>
<td>cell cycle process</td>
</tr>
<tr>
<td>IP00009991</td>
<td>cyclin-dependent kinase inhibitor 18 [p27, Kip1] [Source:HGNC Symbol:Acc:1766]</td>
<td>CKIN1B</td>
<td>negative regulation of cell proliferation</td>
</tr>
<tr>
<td>IP00009991</td>
<td>cyclin-dependent kinase inhibitor 18 [p27, Kip1] [Source:HGNC Symbol:Acc:1766]</td>
<td>CKIN1B</td>
<td>negative regulation of cell proliferation</td>
</tr>
<tr>
<td>IP00009991</td>
<td>cyclin-dependent kinase inhibitor 18 [p27, Kip1] [Source:HGNC Symbol:Acc:1766]</td>
<td>CKIN1B</td>
<td>negative regulation of cell proliferation</td>
</tr>
<tr>
<td>IP00009991</td>
<td>cyclin-dependent kinase inhibitor 18 [p27, Kip1] [Source:HGNC Symbol:Acc:1766]</td>
<td>CKIN1B</td>
<td>positive regulation of cell proliferation</td>
</tr>
</tbody>
</table>

Next ▶
Understanding a protein list

- Level I
  - What are the proteins/genes behind the IDs and what do we know about the functions of the proteins/genes?

- Level II
  - Which biological processes and pathways are the most interesting in terms of the experimental question?
Enrichment analysis:

- Enrichment analysis: is a functional group (e.g. cell cycle) significantly associated with the experimental question?

Filter for significant proteins

Differentially expressed protein list (260 proteins)

Extracellular space (83 proteins)

All identified proteins (1733)

180

83

1305

1305

9.2

22

Random

Observed

180

83

180

83

Filter for significant proteins

Differentially expressed protein list (260 proteins)

Extracellular space (83 proteins)

All identified proteins (1733)

IPI00375843
IPI00171798
IPI00299485
IPI00009542
IPI0019568
IPI00060627
IPI00168262
IPI0082931
IPI0025084
IPI00412546
IPI00165528
IPI0043992
IPI00384992
IPI0006991
IPI00021885
......

MMP9
SERPINF1
A2ML1
F2
FN1
LYZ
TNXB
FGG
MPO
FBLN1
THBS1
HDLBP
GSN
FBN1
CA2
P11
CCL21
FGB
......

180

83

1305

1305

Observed

Random

All identified proteins (1733)
Enrichment analysis: hypergeometric test

<table>
<thead>
<tr>
<th></th>
<th>Significant proteins</th>
<th>Non-significant proteins</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteins in the group</td>
<td>$k$</td>
<td>$j-k$</td>
<td>$j$</td>
</tr>
<tr>
<td>Other proteins</td>
<td>$n-k$</td>
<td>$m-n-j+k$</td>
<td>$m-j$</td>
</tr>
<tr>
<td>Total</td>
<td>$n$</td>
<td>$m-n$</td>
<td>$m$</td>
</tr>
</tbody>
</table>

Hypergeometric test: given a total of $m$ proteins where $j$ proteins are in the functional group, if we pick $n$ proteins randomly, what is the probability of having $k$ or more proteins from the group?

$$p = \sum_{i=k}^{\min(n,j)} \frac{\binom{m-j}{j} \binom{j}{i}}{\binom{n}{i} \binom{m}{n}}$$

Zhang et.al. Nucleic Acids Res. 33:W741, 2005
Gene Ontology (http://www.geneontology.org)
- Structured, precisely defined, controlled vocabulary for describing the roles of genes and gene products
- Three organizing principles: molecular function, biological process, and cellular component

Pathways
- KEGG (http://www.genome.jp/kegg/pathway.html)
- Pathway commons (http://www.pathwaycommons.org)
- WikiPathways (http://www.wikipathways.org)

Cytogenetic bands

Targets of transcription factors/miRNAs
WebGestalt: Web-based Gene Set Analysis Toolkit

- 8 organisms: Human, Mouse, Rat, Dog, Fruitfly, Worm, Zebrafish, Yeast
- Microarray Probe IDs: Affymetrix, Agilent, Codelink, Illumina
- Gene IDs: Gene Symbol, GenBank, Ensembl Gene, RefSeq Gene, UniGene, Entrez Gene, SGD, MGI, Flybase ID, Wormbase ID, ZFIN
- Protein IDs: UniProt, IPI, RefSeq Peptide, Ensembl Peptide
- Genetic Variation IDs: dbSNP
- 196 ID types with mapping to Entrez Gene ID
- 59,278 functional categories with genes identified by Entrez Gene IDs

Gene Ontology:
- Biological Process
- Molecular Function
- Cellular Component

Pathway:
- KEGG
- Pathway Commons
- WikiPathways

Network module:
- Transcription factor targets
- microRNA targets
- Protein interaction modules

Disease and Drug:
- Disease association genes
- Drug association genes

Chromosomal location:
- Cytogenetic bands

http://www.webgestalt.org

Zhang et al. Nucleic Acids Res. 33:W741, 2005
Wang et al. Nucleic Acids Res. 41:W77, 2013
WebGestalt analysis

- **Select** the organism of interest.

- **Upload** a gene/protein list in the **txt format**, one ID per row. Optionally, a value can be provided for each ID. In this case, put the ID and value in the same row and separate them by a tab. Then pick the ID type that corresponds to the list of IDs.

- **Categorize** the uploaded ID list based upon GO Slim (a simplified version of Gene Ontology that focuses on high level classifications).

- **Analyze** the uploaded ID list for enrichment in various biological contexts. You will need to select an appropriate predefined reference set or upload a reference set. If a customized reference set is uploaded, ID type also needs to be selected. After this, select the analysis parameters (e.g., significance level, multiple test adjustment method, etc.).

- **Retrieve** enrichment results by opening the respective results files. You may also open and/or download a TSV file, or download the zipped results to a directory on your desktop.
Input list
- 260 significant proteins identified in the HNSCC study (hnscc_sig_withLogRatio.txt)

Mapping result
- Total number of User IDs: 260. Unambiguously mapped User IDs to Entrez IDs: 229. Unique User Entrez IDs: 224. The Enrichment Analysis will be based upon the unique IDs.

### Mapped User IDs

<table>
<thead>
<tr>
<th>User IDs</th>
<th>Mapped IDs</th>
<th>Protein Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPI0021948</td>
<td>ENSG00000044113</td>
<td>CTNNAL</td>
</tr>
<tr>
<td>IPI00215948</td>
<td>1495</td>
<td></td>
</tr>
<tr>
<td>IPI0025094</td>
<td>84176</td>
<td>MYH16</td>
</tr>
<tr>
<td>IPI00218414</td>
<td>760</td>
<td>CA2</td>
</tr>
<tr>
<td>IPI00783625</td>
<td>5268</td>
<td>SERPINB5</td>
</tr>
<tr>
<td>IPI0024254</td>
<td>3437</td>
<td>IFIT3</td>
</tr>
<tr>
<td>IPI0025427</td>
<td>6037</td>
<td>RNASE3</td>
</tr>
<tr>
<td>IPI00307545</td>
<td>7145</td>
<td>TNS1</td>
</tr>
<tr>
<td>IPI00216470</td>
<td>8396</td>
<td>PIP4K2B</td>
</tr>
<tr>
<td>IPI0029048</td>
<td>23170</td>
<td>TTLL12</td>
</tr>
<tr>
<td>IPI00419433</td>
<td>80223</td>
<td>RAB11FIP1</td>
</tr>
<tr>
<td>IPI00295400</td>
<td>7453</td>
<td>WARS</td>
</tr>
<tr>
<td>IPI00330546</td>
<td>51486</td>
<td>CTDSB1</td>
</tr>
</tbody>
</table>

### User IDs Not Mapped

<table>
<thead>
<tr>
<th>User IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPI00247583</td>
</tr>
<tr>
<td>IPI00016910</td>
</tr>
<tr>
<td>Cntm_P19013</td>
</tr>
<tr>
<td>IPI00240503</td>
</tr>
<tr>
<td>IPI00478600</td>
</tr>
<tr>
<td>Cntm_P04264</td>
</tr>
<tr>
<td>IPI00658109</td>
</tr>
<tr>
<td>Cntm_P02538</td>
</tr>
<tr>
<td>Cntm_P13646</td>
</tr>
<tr>
<td>IPI00787452</td>
</tr>
<tr>
<td>Cntm_P08779</td>
</tr>
<tr>
<td>Cntm_P02533</td>
</tr>
</tbody>
</table>
WebGestalt: GOSlim classification

Bar chart of Biological Process categories

- Biological process

Bar chart of Molecular Function categories

- Molecular function

Bar chart of Cellular Component categories

- Cellular component
### WebGestalt: top 10 enriched GO biological processes

<table>
<thead>
<tr>
<th>Biological Process</th>
<th>Enriched GO Terms</th>
<th>Gene Symbols</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epiidermis development</td>
<td>GO:0008544</td>
<td>Sprr3, Lamc2, AnxA1, LAMB3, Ppl, Bmp4, Sema4b, Fabr5, Sprr3, Eptdc1</td>
</tr>
</tbody>
</table>

**Reference list:**
CSHL2010_hnscc_all_proteins.txt
WebGestalt: top 10 enriched WikiPathways

- Blood Clotting Cascade
- Focal Adhesion
- Alpha6-Beta4 Integrin Signaling Pathway
- Inflammatory Response Pathway
- Endochondral Ossification
- Senescence and Autophagy
- FAS pathway and Stress induction of HSP regulation
- Oxidative Stress
- Myometrial Relaxation and Contraction Pathways
- Complement and Coagulation Cascades KEGG

BCHM352
Limitation of the over-representation analysis

- Does not account for the order of genes in the significant gene list
- Arbitrary thresholding leads to the lose of information
Gene Set Enrichment Analysis (GSEA)

- Test whether the members of a predefined gene set are randomly distributed throughout the ranked gene list
  - Calculation of an Enrichment Score, modified Kolmogorov Smirnov test
  - Estimation of Significance Level of ES, permutation test
  - Adjustment for Multiple Hypothesis Testing, control False Discovery Rate
  - Leading edge subset: genes contribute to the significance

http://www.broad.mit.edu/gsea/

Subramanian et al. PNAS 102:15545, 2005
Understanding a protein list

- **Level I**
  - What are the proteins/genes behind the IDs and what do we know about the functions of the proteins/genes?

- **Level II**
  - Which biological processes and pathways are the most interesting in terms of the experimental question?

- **Level III**
  - How do the proteins work together to form a network?
Resources

- GeneMANIA
  - [http://genemania.org](http://genemania.org)
- STRING
  - [http://string-db.org/](http://string-db.org/)
- Genes2Networks
  - [http://actin.pharm.mssm.edu/genes2networks/](http://actin.pharm.mssm.edu/genes2networks/)
Understanding a protein list: summary

- **Level I**
  - What are the proteins/genes behind the IDs and what do we know about the functions of the proteins/genes?
  - Biomart (http://www.biomart.org/)

- **Level II**
  - Which biological processes and pathways are the most interesting in terms of the experimental question?
  - WebGestalt (http://bioinfo.vanderbilt.edu/webgestalt)

- **Level III**
  - How do the proteins work together to form a network?
  - GeneMANIA (http://genemania.org)