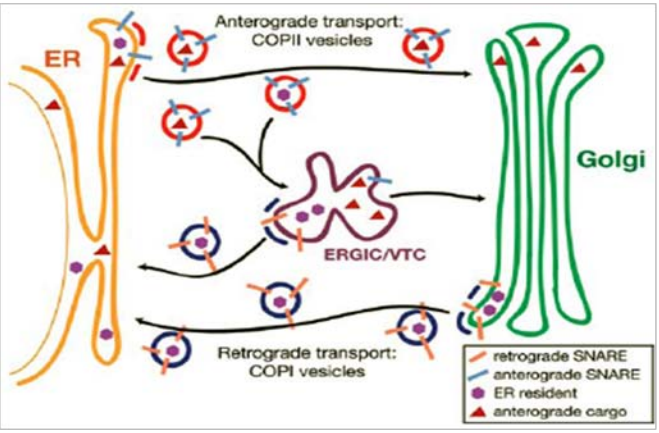
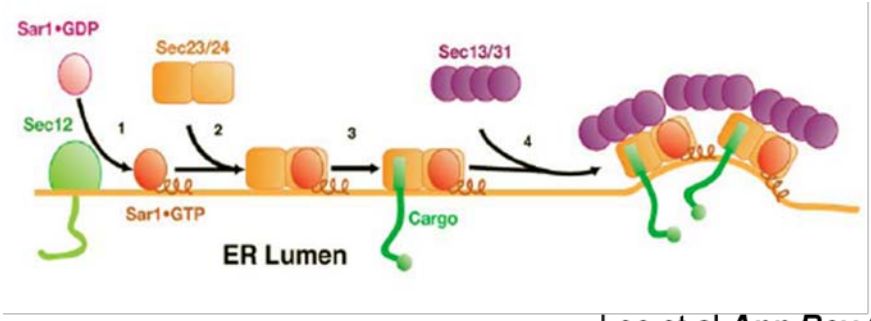
Overview

The bi-directional transport between the endoplasmic reticulum (ER) and the Golgi apparatus is a fundamental process in eukaryotic cells. The process is mediated by vesicles sculpted by coat protein complexes, COPI and COPII.



Lee et al ***Ann Rev Cell Dev Biol 2004***

COPII coated vesicles transport newly synthesized or ER resident proteins to Golgi, a process initiated upon the activation of Sar1 GTPase by it GEF Sec12. Active Sar1 in turn recruits the inner coat heterodimer Sec23/Sec24, the latter of which mediates cargo recognition. The Sar1-Sec23/Sec24 complex then recruits the outer coat heterotetramer Sec13/Sec31, which can further polymerize the coat complex and drive membrane deformation for vesicle budding. However, factors that determine the specificity of COPII cargo recognition remain to be fully elucidated. Interestingly, mammals express four isoforms of Sec24, Sec24A/B/C/D, of which the cargo specificity remains largely unknown.



Lee et al ***Ann Rev Cell Dev Biol 2004***

We generated mice models deficient of Sec24A using genetrap technology, in an effect to characterize the physiological role of specific COPII subunits, particularly, the cargo recognition subunits Sec24. The mice are viable and develop to adulthood with no gross abnormalities. To characterize the gene expression profile in Sec24A deficient mice, we performed a RNA-seq experiment with liver mRNA from Sec24A KO and their wild type littermates.

To test our hypothesis, we integrated the ERANGE RNA-Seq package with our custom differentially expressed gene analysis scripts and used this pipeline to process the RNA-Seq data. We analyzed top differentially expressed genes with FDR<0.10 using ConceptGen – a web based gene set enrichment test tool developed here at the NCIBI-CCMB. ConceptGen results for down-regulated genes show the enrichment of biomedical terms related to fatty acid and lipid metabolic processes. Thus our results report that there is a decreased lipogenesis in the Sec24A KO mice, likely indicating that genes involved in this process might also be affected by Sec24A. However, from current experimental results, we cannot simply deduce that genes involved in fatty acid or cholesterol synthesis are also directly targeted by Sec24A.

Part I. Command Summary for ERANGE RNA-Seq Steps

Step 1. Preparation and path set up

1) First, you will need to download ERANGE package from ERANGE's home page (http://woldlab.caltech.edu/rnaseq/). The chromosome files, knownGene.txt, and repeatMask annotations for each chromosome of mm9 should also be downloaded locally from the UCSC.

2) Set correct path:

```
# set PYTHONPATH to point to the parent directory of the Cistematic
export PYTHONPATH=/ccmb/home/ybai/data11/seq
# set CISTEMATIC_ROOT to the directory that contains the genome directories
export CISTEMATIC_ROOT=/ccmb/home/ybai/data11/seq
# set ERANGEPATH
export ERANGEPATH=/usr/local/bioinf/ERANGE3.1/commoncode
# set CISTEMATIC_TEMP to a local directory with ample space
export CISTEMATIC_TEMP= /data/ybai/tmp
```

3) Under your downloaded mouse genome directory (i.e. /ccmb/CoreBA/Projects/Data/mouse/), perform the following tasks:

```
# create splice file using getsplicefa.py with maxBorder set to 4 bp shorter than the read length
python2.6 $ERANGEPATH/getsplicefa.py mouse /ccmb/CoreBA/Projects/Data/mouse/knownGene.txt mm9splice37.fa 33
cat /ccmb/CoreBA/Projects/Data/mouse/mm9.fa > /ccmb/CoreBA/Projects/Data/mouse/mm9sp37.fa
cat /ccmb/CoreBA/Projects/Data/mouse/mm9splice37.fa >> /ccmb/CoreBA/Projects/Data/mouse/mm9sp37.fa
# build expanded genome using Bowtie's bowtie-build
bowtie-build /ccmb/CoreBA/Projects/Data/mouse/mm9sp37.fa mm9sp37_indexes
# build repeatmask database using buildrmaskdb.py
python2.6 $ERANGEPATH/buildrmaskdb.py /ccmb/CoreBA/Projects/Data/mouse/mm9repeats /ccmb/CoreBA/Projects/Data/mouse/mm9repeats/rmask.db
```

Step 2. Bowtie alignment step

Input: path ($path) to "*_sequence.txt" sequence read file, read file itself ($seq_file), and output path ($bowtie_path) to store result files. A typical read sequence file generated from Illumina pipeline should look like:

```
[ybai@ccmb-comp1 private]$ head -12 s_6_24A-WT_sequence.txt
@unknown_0001:6:1:1156:2319#0/1
GATAATCCATCACNCGTTAAAAATTTGCNTACTACCA
+unknown_0001:6:1:1156:2319#0/1
Za^``aaaaa^Z^ZEZ[[[[aaaaaaaa^E^`^``aaa
@unknown_0001:6:1:1156:3931#0/1
GGAAATACCTCACNCTTTCCCTTCTCCCNCCCCAAAC
+unknown_0001:6:1:1156:3931#0/1
``\`BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@unknown_0001:6:1:1156:20179#0/1
ACATAACCTTAGANGAGAGCCACGGGAANGTGCTAGA
+unknown_0001:6:1:1156:20179#0/1
abbbbbbbb`^`EWVWZXbbbbbY[YZE^Z]^^bbb
```

Command line:

bowtie /ccmb/CoreBA/Projects/Data/mouse/mm9sp37_indexes -v 2 -k 11 -m 10 --best -q $path/$seq_file --un $bowtie_path/$seq_file.unmapped.fa --max $bowtie_path/$seq_file.repeat.fa $bowtie_path/$seq_file.bowtie.txt"

Output: bowtie alignment file ($seq_file.bowtie.txt). i.e.

```
[ybai@ccmb-comp1 private]$ head -7 s_6_24A-WT_sequence.txt.bowtie.txt
unknown_0001:6:1:1156:3931#0/1 +     chr8   63391667     GGAAATACCTCACNCTTTCCCTTCTCCCNCCCCAAAC
``\`BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB  0     13:T>N,28:T>N
unknown_0001:6:1:1156:2319#0/1 +     chr3   14872152     GATAATCCATCACNCGTTAAAAATTTGCNTACTACCA
Za^`aaaaa^Z^ZEZ[[[[aaaaaaaa^E^`^``aaa  0     13:T>N,28:C>N
unknown_0001:6:1:1156:20179#0/1 +     chr13   42253205     ACATAACCTTAGANGAGAGCCACGGGAANGTGCTAGA
abbbbbbbb``^`EWVWZXbbbbbY[YZE^Z]^^bbb  0     13:G>N,28:A>N
unknown_0001:6:1:1157:12474#0/1 -     chr7   25673303     CCACAGCANGGCACCTGTACCTCNGAGGTGGTGCTGG
BBBBBBBBBBBBBBBBBBBBBBBBBBBBB`W````_  2     13:C>N,28:C>N
unknown_0001:6:1:1157:12474#0/1 -     chr18   42412226     CCACAGCANGGCACCTGTACCTCNGAGGTGGTGCTGG
BBBBBBBBBBBBBBBBBBBBBBBBBBBBB`W````_  2     13:C>N,28:C>N
unknown_0001:6:1:1157:12474#0/1 -     chr12   112136498     CCACAGCANGGCACCTGTACCTCNGAGGTGGTGCTGG
BBBBBBBBBBBBBBBBBBBBBBBBBBBBB`W````_  2     13:C>N,28:C>N
unknown_0001:6:1:1157:2991#0/1 +     chr4   133525096     GAAGCTAGTTTGTNATAGCCATGGCAGCNGAAAGCAG
a_aaaaa`BBBBBBBBBBBBBBBBBBBBBBBBBBBB  0     13:A>N,28:A>N
```

Step 3. Counting reads falling on gene models, identifying new regions, filtering out new regions that overlap repeats, and mapping candidate regions within a certain radius of genes

Input: bowtie alignment file ($seq_file.bowtie.txt). See the above example.

1) Making rds file
$label = substr($seq_file,0,-4);
python2.6 /usr/local/bioinf/ERANGE3.1/commoncode/makerdsfrombowtie.py $label $bowtie_path/$seq_file.bowtie.txt /data/ybai/tmp/$seq_file.rds -RNA /ccmb/CoreBA/Projects/Data/mouse/knownGene.txt –index

2) Changing database cache size
python2.6 /usr/local/bioinf/ERANGE3.1/commoncode/rdsmetadata.py /data/ybai/tmp/$seq_file.rds -defaultcache 6000000

3) Copying rds file to the memory
cp /data/ybai/tmp/$seq_file.rds /dev/shm

4) Counting unique reads falling on gene model
python2.6 /usr/local/bioinf/ERANGE3.1/commoncode/geneMrnaCounts.py mouse /dev/shm/$seq_file.rds /data/ybai/tmp/$seq_file.uniqs.count –markNM

5) Counting splice reads
python2.6 /usr/local/bioinf/ERANGE3.1/commoncode/geneMrnaCounts.py mouse /dev/shm/$seq_file.rds /data/ybai/tmp/$seq_file.splices.count -splices –noUniqs

6) Calculating a first-pass RPKM
python2.6 /usr/local/bioinf/ERANGE3.1/commoncode/normalizeExpandedExonic.py mouse /dev/shm/$seq_file.rds /data/ybai/tmp/$seq_file.uniqs.count none /data/ybai/tmp/$seq_file.firstpass.rpkm

7) Recounting the unique reads

python2.6
/usr/local/bioinf/ERANGE3.1/commoncode/geneMrnaCountsWeighted.py mouse
/dev/shm/$seq_file.rds /data/ybai/tmp/$seq_file.firstpass.rpkm
/data/ybai/tmp/$seq_file.uniqs.recount –uniq

8) Finding new regions outside of gene models
python2.6 /usr/local/bioinf/ERANGE3.1/commoncode/findall.py $label
/dev/shm/$seq_file.rds /data/ybai/tmp/$seq_file.newregions.txt -raw -minimum
25 -spacing 40 -nodirectionality -notrim -noshift -flag NM

9) Filtering out new regions that overlap repeats more than a certain fraction
python2.6 /usr/local/bioinf/ERANGE3.1/commoncode/checkrmask.py
/ccmb/CoreBA/Projects/Data/mouse/mm9repeats/rmask.db
/data/ybai/tmp/$seq_file.newregions.txt
/data/ybai/tmp/$seq_file.newregions.repstatus
/data/ybai/tmp/$seq_file.newregions.good -startField 1

10) Mapping all candidate regions that are within a 20kb radius of a gene
python2.6 /usr/local/bioinf/ERANGE3.1/commoncode/getallgenes.py mouse
/data/ybai/tmp/$seq_file.newregions.good
/data/ybai/tmp/$seq_file.candidates.txt 20001 -trackfar –cache

11) Calculating expanded exonic read density
python2.6
/usr/local/bioinf/ERANGE3.1/commoncode/normalizeExpandedExonic.py mouse
/dev/shm/$seq_file.rds /data/ybai/tmp/$seq_file.uniqs.recount
/data/ybai/tmp/$seq_file.splices.count /data/ybai/tmp/$seq_file.expanded.rpkm
/data/ybai/tmp/$seq_file.candidates.txt /data/ybai/tmp/$seq_file.accepted.rpkm

Step 4. Weighing multi-reads, creating bed file, and calculating final reads per KB per million reads (RPKM)

1) Creating bed file of accepted candidate regions

python2.6 /usr/local/bioinf/ERANGE3.1/commoncode/regiontobed.py $label /data/ybai/tmp/$seq_file.accepted.rpkm /data/ybai/tmp/$seq_file.bed -color 255,0,0

2) Weighing multi-reads
python2.6 /usr/local/bioinf/ERANGE3.1/commoncode/geneMrnaCountsWeighted.py mouse /dev/shm/$seq_file.rds /data/ybai/tmp/$seq_file.expanded.rpkm /data/ybai/tmp/$seq_file.multi.count -accept /data/ybai/tmp/$seq_file.accepted.rpkm –multi

3) Calculating final exonic read density
python2.6 /usr/local/bioinf/ERANGE3.1/commoncode/normalizeFinalExonic.py /dev/shm/$seq_file.rds /data/ybai/tmp/$seq_file.expanded.rpkm /data/ybai/tmp/$seq_file.multi.count /data/ybai/tmp/$seq_file.final.rpkm

4) Moving all files to destination directory and clear up memory
mv /data/ybai/tmp/* $bowtie_path/
rm /dev/shm/*

Output: *.accepted.rpkm, *.final.rpkm, *splices.count…

```
[ybai@ccmb-comp1 private]$ head -10 s_6_24A-WT_sequence.txt.accepted.rpkm
FAR1    chr1    79700110      79700577      2.77    467    FAR1
FAR2    chr1    89470263      89470444      9.61    181    FAR2
FAR3    chr1    127574366     127574769     3.65    403    FAR3
FAR4    chr1    152983143     152983540     7.75    397    FAR4
FAR4    chr1    152985298     152986404     4.72    1106   FAR4
FAR4    chr1    152988242     152988675     4.32    433    FAR4
FAR4    chr1    152988680     152988844     8.43    164    FAR4
FAR4    chr1    152989333     152989554     6.45    221    FAR4
FAR4    chr1    152989563     152989743     8.42    180    FAR4
FAR5    chr1    187086555     187086719     7.88    164    FAR5

[ybai@ccmb-comp1 private]$ head -5 s_6_24A-WT_sequence.txt.final.rpkm
#gene   len_kb  RPKM
Alb1    2.028   31420.21
Apoe    1.089   13931.63
Apoa2   0.477   11242.40
Serpina1e       1.374   10543.13
```

Part II. Command Summary for Identifying Differentially Expression Gene Analysis Step under R environment

Input files: $seq_file1_prefix.final.rpkm and $seq_file2_prefix.final.rpkm (*.final.rpkm); $seq_file1_prefix.accepted.rpkm and $seq_file2_prefix.accepted.rpkm (*.accepted.rpkm) for both WT and KO lanes from ERANGE

Command line:

finalFile1=$seq_file1_prefix.final.rpkm finalFile2=$seq_file2_prefix.final.rpkm acceptedFile1=$seq_file1_prefix.accepted.rpkm acceptedFile2=$seq_file2_prefix.accepted.rpkm pathToInput=$input_path pathToOutput=$output_path R CMD BATCH RNA-Seq_Testing_ERANGE_mm9.R

Output files: See Table below.

File Name	Description
ConceptGen-Input-Genes.txt	**A list of down-regulated genes under a certain FDR cutoff criteria**
ERANGE-FAR_Prediction_Results.txt	Predicted candidate exon regions
ERANGE-Results-Genes.txt	Table containing genes along with their normalized RPKM values, fold changes, logfc, P-value, and FDR
ERANGE-Raw-Genes	Genes and their RPKM values in WT and KO samples along with their exon length upon merging by identical gene symbols
AbsM-A plot.jpeg	Estimate variance as function of RPKM
Rplots.pdf	Estimate variance as function of normalized RPKM
M-A plot.jpeg	Plot logarithmic normalized RPKM ratios of KO/WT by "average $\log_2$(normalized RPKM)"

Below is an example for “ConceptGen-Input-Genes.txt”.

```
Entrez Gene ID
77371
13119
30939
76574
13117
14373
22187
68801
217166
19734
232174
227731
194655
13087
20249
…
```

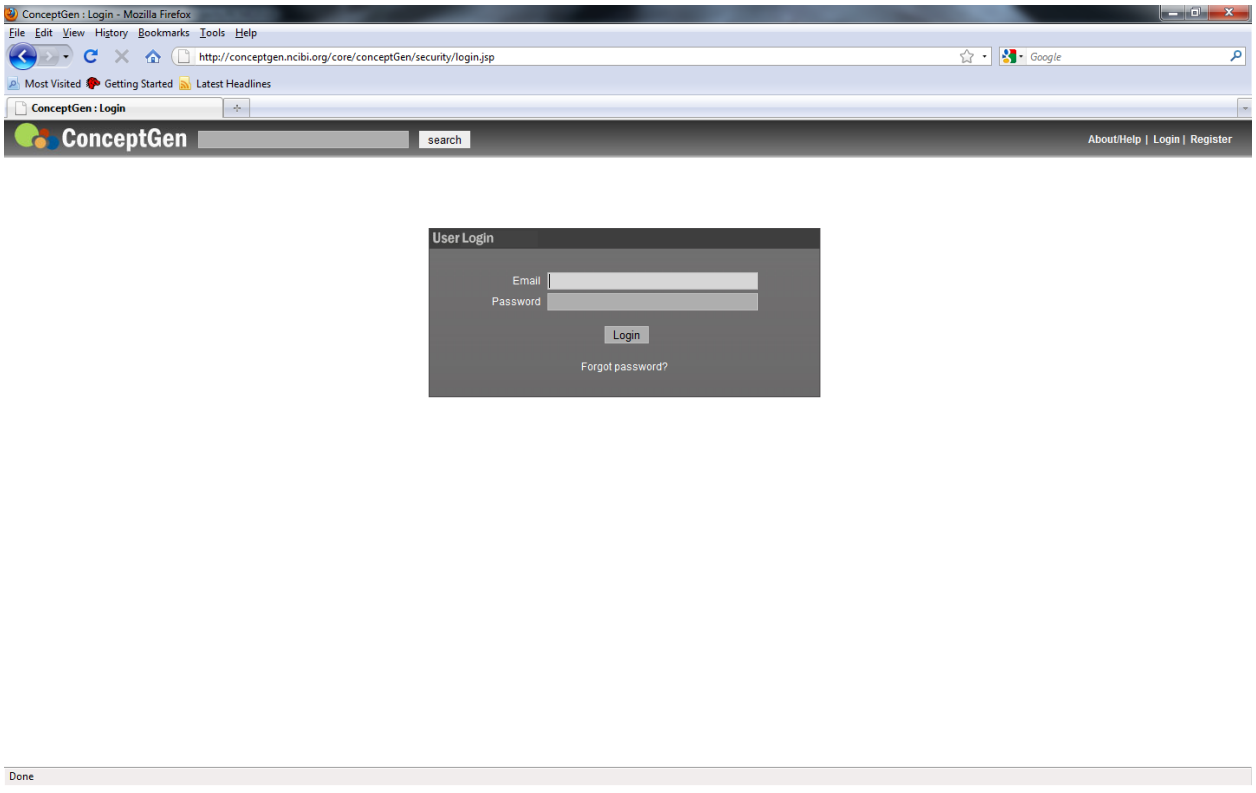
Part III: Biological concept enrichment test for down-regulated genes using web-based ConceptGen tool developed here at the NCIBI at the University of Michigan

Input: A list of down-regulated genes under a certain FDR cutoff criteria from ERANGE-Results-Genes.txt

Output: ConceptGen output

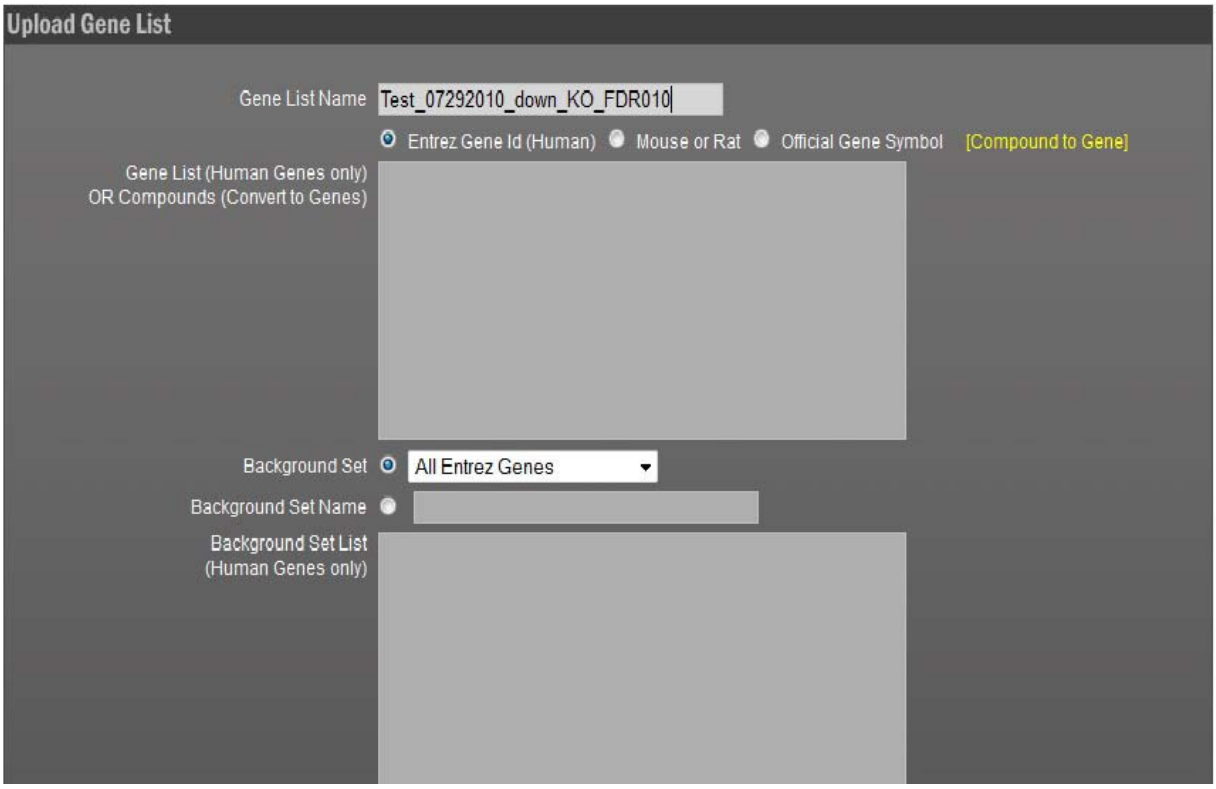
Steps:

1. Load ConceptGen page (http://conceptgen.ncibi.org), register, and log in.

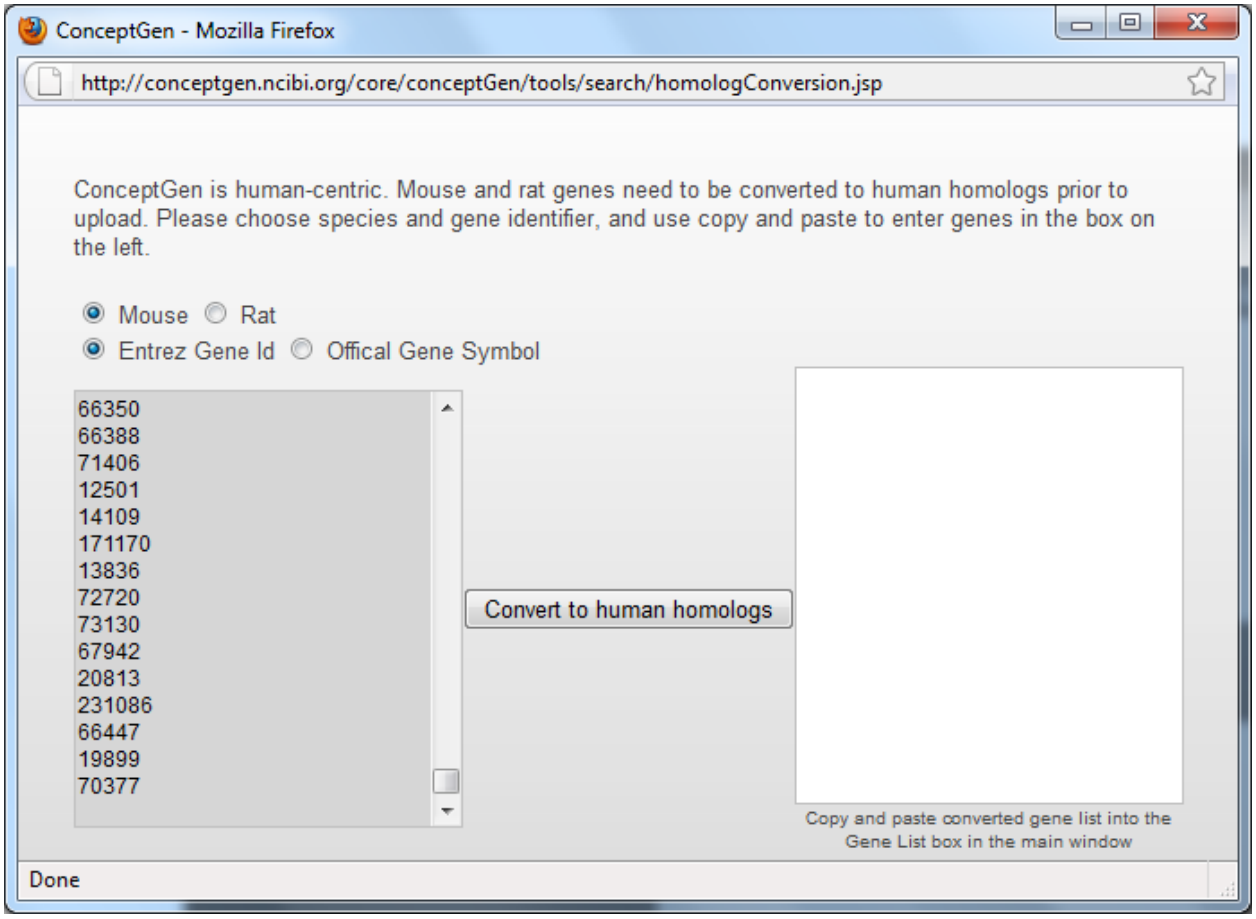


2. Click “Upload Concept” tab, or alternatively click “Upload new gene list” icon under “My Concepts”.

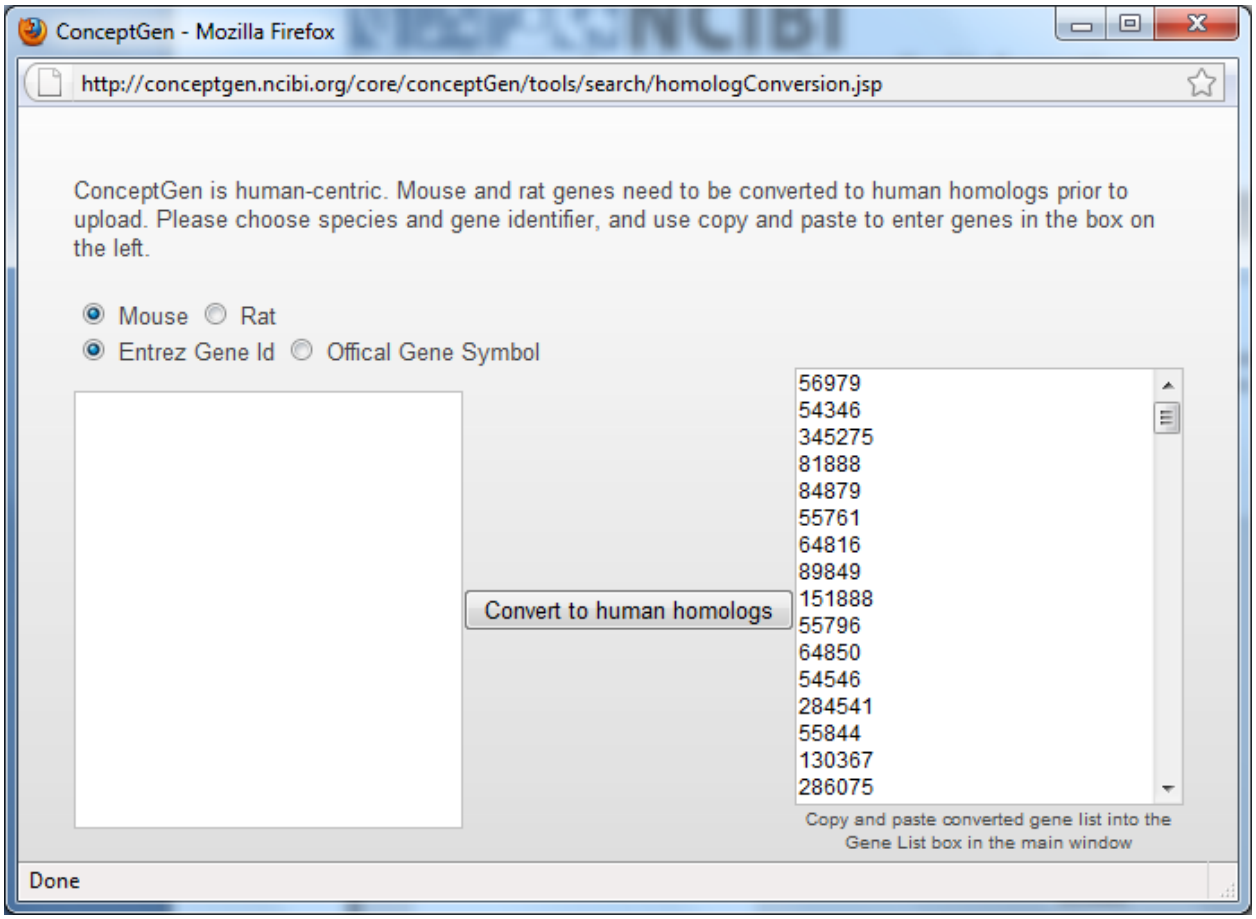
3. Enter a name for your gene list.



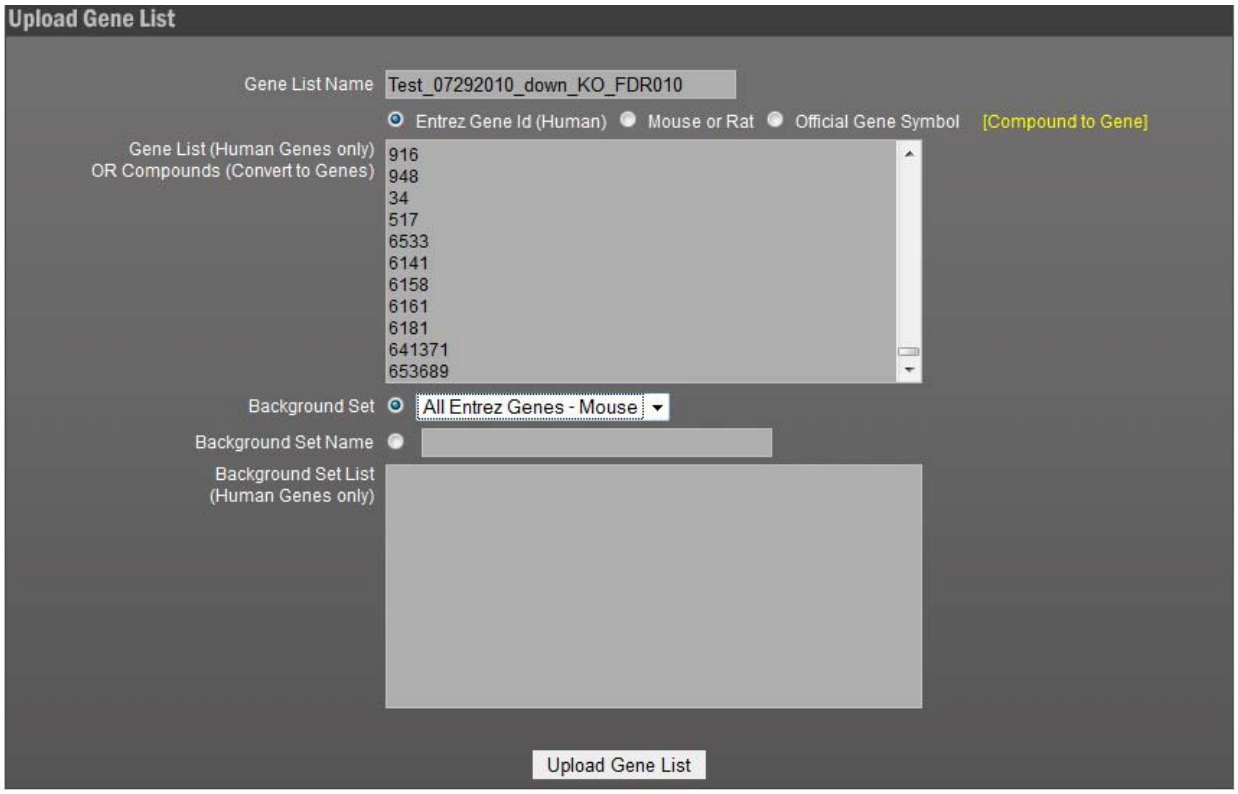
4. When select “Mouse or Rat” option, a window is popped up. Since ConceptGen is human-centric, Mouse or rat genes need to be converted to human homologs prior to upload. In the conversion dialog, select “Mouse” species and “Entrez Gene Id”. Paste 201 mouse genes in the left pane as shown below.



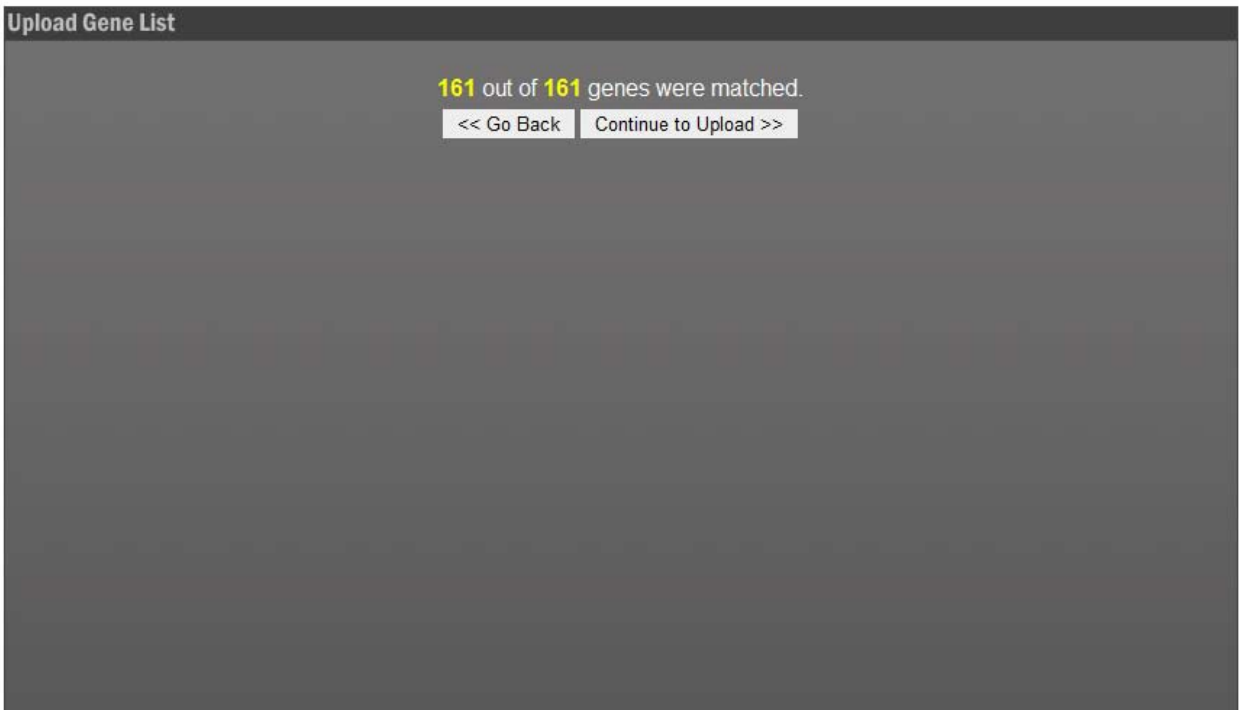
5. Click “convert to human homologs” button.



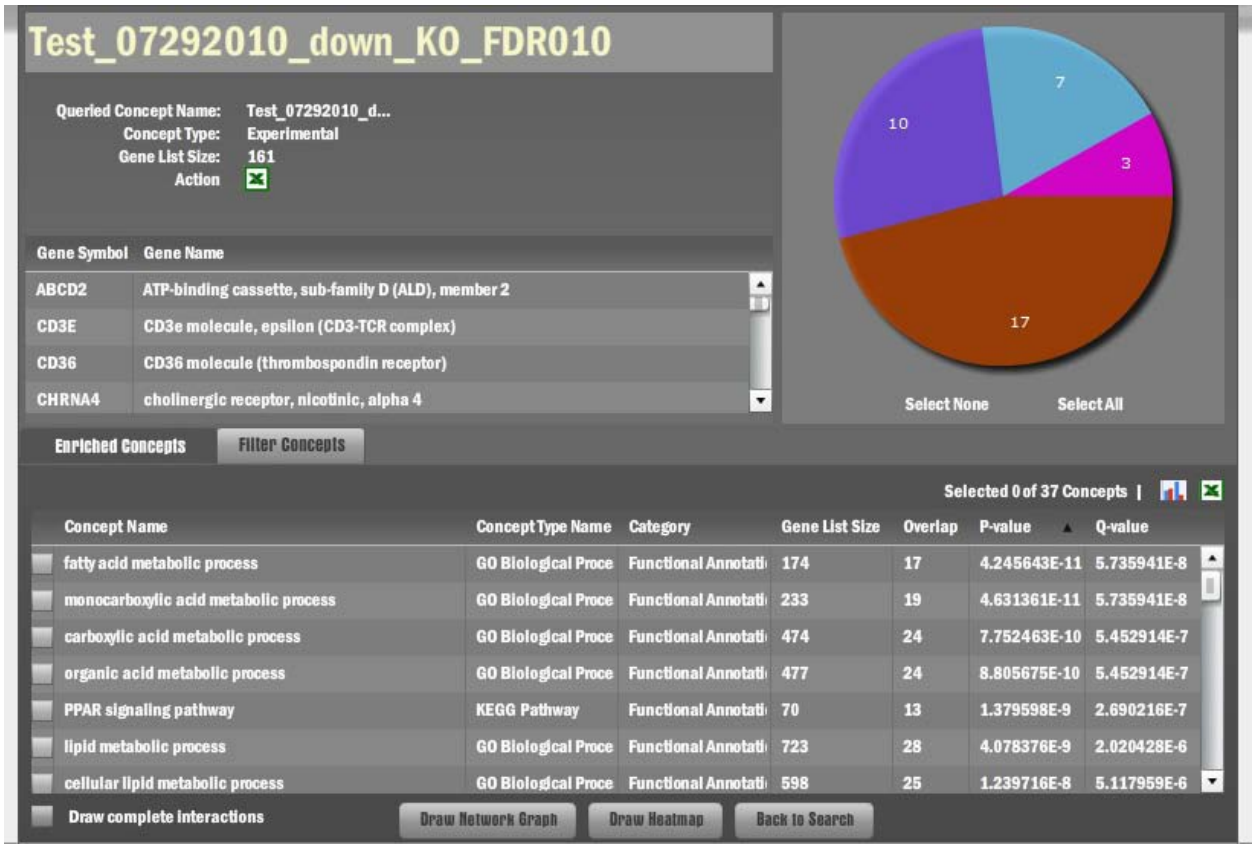
6. Copy and paste converted gene list into the “Gene List (Human Genes only)” box in the main window and select “Background set” as “All Entrez Genes - Mouse”.



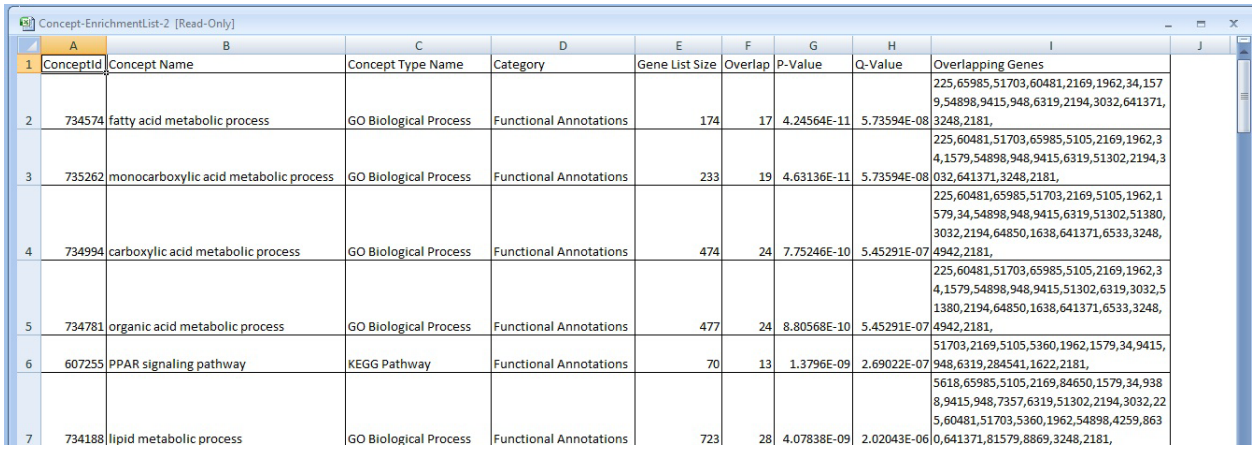
7. Click “Upload Gene List” button. 161 out of 161 genes were matched.



8. Click “continue to upload”. After the “My Concepts ” dialog reappears, you can click your concept name’s "Click to launch Concept Explorer" play button to view the results.



9. Click "Export selected concepts to Excel", open the file to view and/or to save it locally.



Concept-EnrichmentList-2 [Read-Only]

	ConceptId	Concept Name	Concept Type Name	Category	Gene List Size	Overlap	P-Value	Q-Value	Overlapping Genes
2	734574	fatty acid metabolic process	GO Biological Process	Functional Annotations	174	17	4.24564E-11	5.73594E-08	225,65985,51703,60481,2169,1962,34,1579,54898,9415,948,6319,2194,3032,641371,3248,2181,
3	735262	monocarboxylic acid metabolic process	GO Biological Process	Functional Annotations	233	19	4.63136E-11	5.73594E-08	225,60481,51703,65985,5105,2169,1962,34,1579,54898,948,9415,6319,51302,2194,3032,641371,3248,2181,
4	734994	carboxylic acid metabolic process	GO Biological Process	Functional Annotations	474	24	7.75246E-10	5.45291E-07	225,60481,65985,51703,2169,5105,1962,1579,34,54898,948,9415,6319,51302,51380,3032,2194,64850,1638,641371,6533,3248,4942,2181,
5	734781	organic acid metabolic process	GO Biological Process	Functional Annotations	477	24	8.80568E-10	5.45291E-07	225,60481,51703,65985,5105,2169,1962,34,1579,54898,948,9415,51302,6319,3032,51380,2194,64850,1638,641371,6533,3248,4942,2181,
6	607255	PPAR signaling pathway	KEGG Pathway	Functional Annotations	70	13	1.3796E-09	2.69022E-07	51703,2169,5105,5360,1962,1579,34,9415,948,6319,284541,1622,2181,
7	734188	lipid metabolic process	GO Biological Process	Functional Annotations	723	28	4.07838E-09	2.02043E-06	5618,65985,5105,2169,84650,1579,34,9388,9415,948,7357,6319,51302,2194,3032,225,60481,51703,5360,1962,54898,4259,8630,641371,81579,8869,3248,2181,

References:

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 2008*,* 5, 621-628.

Sartor M, Mahavisno V, Keshamouni V, Cavalcoli J, Wright Z, Karnovsky A, Kuick R, Jagadish HV, Mirel B, Weymouth T, Athey B, Omenn G: **ConceptGen**: a gene set enrichment and gene set relation mapping tool. *Bioinformatics* 2010*,* 26(4):456-463.