main

August 21, 2021

1 GO analysis using GOATOOLS

[1]:
```
import functools
import pandas as pd
import collections as cx
from pybiomart import Dataset
# GO analysis
from goatools.base import download_go_basic_obo
from goatools.base import download_ncbi_associations
from goatools.obo_parser import GODag
from goatools.anno.genetogo_reader import Gene2GoReader
from goatools.goea.go_enrichment_ns import GOEnrichmentStudyNS
```

[2]:
```
@functools.lru_cache()
def get_database():
    dataset = Dataset(name="hsapiens_gene_ensembl",
                      host="http://www.ensembl.org",
                      use_cache=True)
    db = dataset.query(attributes=["ensembl_gene_id",
                                   "external_gene_name",
                                   "entrezgene_id"],
                       use_attr_names=True).dropna(subset=['entrezgene_id'])
    return db


@functools.lru_cache()
def get_deg():
    fn = '../../_m/genes/diffExpr_EAvsAA_FDR05.txt'
    return pd.read_csv(fn, sep='\t')


@functools.lru_cache()
def convert2entrez():
    df = get_deg()
    if 'EntrezID' in df.columns:
        return df.rename(columns={'EntrezID': 'entrezgene_id'})
    else:
```

```
        return df.merge(get_database(), left_on='ensemblID',
                        right_on='ensembl_gene_id')


@functools.lru_cache()
def get_upregulated():
    df = convert2entrez()
    return df.loc[(df['t'] > 0)]


@functools.lru_cache()
def get_downregulated():
    df = convert2entrez()
    return df.loc[(df['t'] < 0)]
```

[3]:
```
def obo_annotation(alpha=0.05):
    # database annotation
    fn_obo = download_go_basic_obo()
    fn_gene2go = download_ncbi_associations() # must be gunzip to work
    obodag = GODag(fn_obo) # downloads most up-to-date
    anno_hs = Gene2GoReader(fn_gene2go, taxids=[9606])
    # get associations
    ns2assoc = anno_hs.get_ns2assc()
    for nspc, id2gos in ns2assoc.items():
        print("{NS} {N:,} annotated human genes".format(NS=nspc, N=len(id2gos)))
    goeaobj = GOEnrichmentStudyNS(
        get_database()['entrezgene_id'], # List of human genes with entrez IDs
        ns2assoc, # geneid/GO associations
        obodag, # Ontologies
        propagate_counts = False,
        alpha = alpha, # default significance cut-off
        methods = ['fdr_bh'])
    return goeaobj


def run_goea(direction):
    if direction == "Up":
        df = get_upregulated()
    elif direction == "Down":
        df = get_downregulated()
    else:
        df = convert2entrez()
    geneids_study = {z[0]:z[1] for z in zip(df['entrezgene_id'], df['Symbol'])}
    goeaobj = obo_annotation()
    goea_results_all = goeaobj.run_study(geneids_study)
    goea_results_sig = [r for r in goea_results_all if r.p_fdr_bh < 0.05]
```

```
    ctr = cx.Counter([r.NS for r in goea_results_sig])
    print('Significant results[{TOTAL}] = {BP} BP + {MF} MF + {CC} CC'.format(
        TOTAL=len(goea_results_sig),
        BP=ctr['BP'],  # biological_process
        MF=ctr['MF'],  # molecular_function
        CC=ctr['CC'])) # cellular_component

    if direction == "Up":
        label = "upregulated"
    elif direction == "Down":
        label = "downregulated"
    else:
        label = "allDEG"
    goeaobj.wr_xlsx("GO_analysis_%s.xlsx" % label, goea_results_sig)
    goeaobj.wr_txt("GO_analysis_%s.txt" % label, goea_results_sig)
```

1.1 Gene ontology

[4]:
```
for direction in ["All", "Up", "Down"]:
    run_goea(direction)
```

```
requests.get(http://purl.obolibrary.org/obo/go/go-basic.obo, stream=True)
  WROTE: go-basic.obo

FTP RETR ftp.ncbi.nlm.nih.gov gene/DATA gene2go.gz -> gene2go.gz
  gunzip gene2go.gz
go-basic.obo: fmt(1.2) rel(2021-08-18) 47,217 GO Terms
HMS:0:00:04.387035 330,313 annotations, 20,685 genes, 18,684 GOs, 1 taxids READ:
gene2go
MF 18,190 annotated human genes
BP 18,505 annotated human genes
CC 19,422 annotated human genes

Load BP Gene Ontology Analysis …
 70% 20,236 of 29,107 population items found in association

Load CC Gene Ontology Analysis …
 74% 21,428 of 29,107 population items found in association

Load MF Gene Ontology Analysis …
 70% 20,354 of 29,107 population items found in association

Run BP Gene Ontology Analysis: current study set of 2955 IDs … 82%  1,852 of
2,260 study items found in association
 76%  2,260 of  2,955 study items found in population(29107)
Calculating 12,429 uncorrected p-values using fisher
  12,429 GO terms are associated with 17,848 of 29,107 population items
   4,987 GO terms are associated with  1,852 of  2,955 study items
```

```
  METHOD fdr_bh:
      19 GO terms found significant (< 0.05=alpha) ( 15 enriched +   4
purified): statsmodels fdr_bh
     261 study items associated with significant GO IDs (enriched)
      52 study items associated with significant GO IDs (purified)

Run CC Gene Ontology Analysis: current study set of 2955 IDs … 86%  1,945 of
2,260 study items found in association
 76%  2,260 of  2,955 study items found in population(29107)
Calculating 1,753 uncorrected p-values using fisher
   1,753 GO terms are associated with 18,711 of 29,107 population items
     876 GO terms are associated with  1,945 of  2,955 study items
  METHOD fdr_bh:
      55 GO terms found significant (< 0.05=alpha) ( 55 enriched +   0
purified): statsmodels fdr_bh
   1,820 study items associated with significant GO IDs (enriched)
       0 study items associated with significant GO IDs (purified)

Run MF Gene Ontology Analysis: current study set of 2955 IDs … 84%  1,894 of
2,260 study items found in association
 76%  2,260 of  2,955 study items found in population(29107)
Calculating 4,420 uncorrected p-values using fisher
   4,420 GO terms are associated with 17,838 of 29,107 population items
   1,624 GO terms are associated with  1,894 of  2,955 study items
  METHOD fdr_bh:
      13 GO terms found significant (< 0.05=alpha) ( 10 enriched +   3
purified): statsmodels fdr_bh
   1,589 study items associated with significant GO IDs (enriched)
      29 study items associated with significant GO IDs (purified)
Significant results[87] = 19 BP + 13 MF + 55 CC
     87 items WROTE: GO_analysis_allDEG.xlsx
     87 GOEA results for  1952 study items. WROTE: GO_analysis_allDEG.txt
  EXISTS: go-basic.obo
  EXISTS: gene2go
go-basic.obo: fmt(1.2) rel(2021-08-18) 47,217 GO Terms
HMS:0:00:04.604500 330,313 annotations, 20,685 genes, 18,684 GOs, 1 taxids READ:
gene2go
MF 18,190 annotated human genes
BP 18,505 annotated human genes
CC 19,422 annotated human genes

Load BP Gene Ontology Analysis …
 70% 20,236 of 29,107 population items found in association

Load CC Gene Ontology Analysis …
 74% 21,428 of 29,107 population items found in association

Load MF Gene Ontology Analysis …
```

```
 70% 20,354 of 29,107 population items found in association

Run BP Gene Ontology Analysis: current study set of 1546 IDs … 78%    857 of
1,097 study items found in association
 71%  1,097 of  1,546 study items found in population(29107)
Calculating 12,429 uncorrected p-values using fisher
  12,429 GO terms are associated with 17,848 of 29,107 population items
   3,085 GO terms are associated with    857 of  1,546 study items
  METHOD fdr_bh:
      25 GO terms found significant (< 0.05=alpha) ( 23 enriched +   2
purified): statsmodels fdr_bh
     168 study items associated with significant GO IDs (enriched)
       6 study items associated with significant GO IDs (purified)

Run CC Gene Ontology Analysis: current study set of 1546 IDs … 82%    902 of
1,097 study items found in association
 71%  1,097 of  1,546 study items found in population(29107)
Calculating 1,753 uncorrected p-values using fisher
   1,753 GO terms are associated with 18,711 of 29,107 population items
     559 GO terms are associated with    902 of  1,546 study items
  METHOD fdr_bh:
      16 GO terms found significant (< 0.05=alpha) ( 16 enriched +   0
purified): statsmodels fdr_bh
     627 study items associated with significant GO IDs (enriched)
       0 study items associated with significant GO IDs (purified)

Run MF Gene Ontology Analysis: current study set of 1546 IDs … 80%    880 of
1,097 study items found in association
 71%  1,097 of  1,546 study items found in population(29107)
Calculating 4,420 uncorrected p-values using fisher
   4,420 GO terms are associated with 17,838 of 29,107 population items
     980 GO terms are associated with    880 of  1,546 study items
  METHOD fdr_bh:
      10 GO terms found significant (< 0.05=alpha) (  9 enriched +   1
purified): statsmodels fdr_bh
     680 study items associated with significant GO IDs (enriched)
       1 study items associated with significant GO IDs (purified)
Significant results[51] = 25 BP + 10 MF + 16 CC
     51 items WROTE: GO_analysis_upregulated.xlsx
     51 GOEA results for   816 study items. WROTE: GO_analysis_upregulated.txt
  EXISTS: go-basic.obo
  EXISTS: gene2go
go-basic.obo: fmt(1.2) rel(2021-08-18) 47,217 GO Terms
HMS:0:00:04.492485 330,313 annotations, 20,685 genes, 18,684 GOs, 1 taxids READ:
gene2go
MF 18,190 annotated human genes
BP 18,505 annotated human genes
CC 19,422 annotated human genes
```

```
Load BP Gene Ontology Analysis …
 70% 20,236 of 29,107 population items found in association

Load CC Gene Ontology Analysis …
 74% 21,428 of 29,107 population items found in association

Load MF Gene Ontology Analysis …
 70% 20,354 of 29,107 population items found in association

Run BP Gene Ontology Analysis: current study set of 1409 IDs … 86%    995 of
1,163 study items found in association
 83%  1,163 of  1,409 study items found in population(29107)
Calculating 12,429 uncorrected p-values using fisher
  12,429 GO terms are associated with 17,848 of 29,107 population items
   3,358 GO terms are associated with    995 of  1,409 study items
  METHOD fdr_bh:
       9 GO terms found significant (< 0.05=alpha) (  6 enriched +   3
purified): statsmodels fdr_bh
     135 study items associated with significant GO IDs (enriched)
      13 study items associated with significant GO IDs (purified)

Run CC Gene Ontology Analysis: current study set of 1409 IDs … 90%  1,043 of
1,163 study items found in association
 83%  1,163 of  1,409 study items found in population(29107)
Calculating 1,753 uncorrected p-values using fisher
   1,753 GO terms are associated with 18,711 of 29,107 population items
     695 GO terms are associated with  1,043 of  1,409 study items
  METHOD fdr_bh:
      41 GO terms found significant (< 0.05=alpha) ( 41 enriched +   0
purified): statsmodels fdr_bh
     966 study items associated with significant GO IDs (enriched)
       0 study items associated with significant GO IDs (purified)

Run MF Gene Ontology Analysis: current study set of 1409 IDs … 87%  1,014 of
1,163 study items found in association
 83%  1,163 of  1,409 study items found in population(29107)
Calculating 4,420 uncorrected p-values using fisher
   4,420 GO terms are associated with 17,838 of 29,107 population items
   1,101 GO terms are associated with  1,014 of  1,409 study items
  METHOD fdr_bh:
      12 GO terms found significant (< 0.05=alpha) ( 11 enriched +   1
purified): statsmodels fdr_bh
     876 study items associated with significant GO IDs (enriched)
       0 study items associated with significant GO IDs (purified)
Significant results[62] = 9 BP + 12 MF + 41 CC
     62 items WROTE: GO_analysis_downregulated.xlsx
     62 GOEA results for  1034 study items. WROTE: GO_analysis_downregulated.txt
```

[ ]: