main

July 9, 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_szVctl_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 Enrichment analysis

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

```
All
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(2021-07-02) 47,229 GO Terms
HMS:0:00:05.312360 341,880 annotations, 20,685 genes, 18,610 GOs, 1 taxids READ:
gene2go
CC 19,421 annotated human genes
BP 18,709 annotated human genes
MF 18,179 annotated human genes

Load BP Gene Ontology Analysis …
 71% 20,538 of 29,107 population items found in association

Load CC Gene Ontology Analysis …
 74% 21,427 of 29,107 population items found in association

Load MF Gene Ontology Analysis …
 70% 20,342 of 29,107 population items found in association

Run BP Gene Ontology Analysis: current study set of 2700 IDs … 89%  2,154 of
2,432 study items found in association
 90%  2,432 of  2,700 study items found in population(29107)
Calculating 12,433 uncorrected p-values using fisher
```

```
  12,433 GO terms are associated with 18,051 of 29,107 population items
   5,738 GO terms are associated with  2,154 of  2,700 study items
  METHOD fdr_bh:
     165 GO terms found significant (< 0.05=alpha) (162 enriched +   3
purified): statsmodels fdr_bh
   1,367 study items associated with significant GO IDs (enriched)
      17 study items associated with significant GO IDs (purified)

Run CC Gene Ontology Analysis: current study set of 2700 IDs … 92%  2,249 of
2,432 study items found in association
 90%  2,432 of  2,700 study items found in population(29107)
Calculating 1,755 uncorrected p-values using fisher
   1,755 GO terms are associated with 18,710 of 29,107 population items
     924 GO terms are associated with  2,249 of  2,700 study items
  METHOD fdr_bh:
     113 GO terms found significant (< 0.05=alpha) (113 enriched +   0
purified): statsmodels fdr_bh
   2,195 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 2700 IDs … 90%  2,186 of
2,432 study items found in association
 90%  2,432 of  2,700 study items found in population(29107)
Calculating 4,361 uncorrected p-values using fisher
   4,361 GO terms are associated with 17,827 of 29,107 population items
   1,757 GO terms are associated with  2,186 of  2,700 study items
  METHOD fdr_bh:
      79 GO terms found significant (< 0.05=alpha) ( 75 enriched +   4
purified): statsmodels fdr_bh
   1,966 study items associated with significant GO IDs (enriched)
      33 study items associated with significant GO IDs (purified)
Significant results[357] = 165 BP + 79 MF + 113 CC
    357 items WROTE: GO_analysis_allDEG.xlsx
    357 GOEA results for  2305 study items. WROTE: GO_analysis_allDEG.txt
Up
  EXISTS: go-basic.obo
  EXISTS: gene2go
go-basic.obo: fmt(1.2) rel(2021-07-02) 47,229 GO Terms
HMS:0:00:05.045684 341,880 annotations, 20,685 genes, 18,610 GOs, 1 taxids READ:
gene2go
CC 19,421 annotated human genes
BP 18,709 annotated human genes
MF 18,179 annotated human genes

Load BP Gene Ontology Analysis …
 71% 20,538 of 29,107 population items found in association

Load CC Gene Ontology Analysis …
```

```
 74% 21,427 of 29,107 population items found in association

Load MF Gene Ontology Analysis …
 70% 20,342 of 29,107 population items found in association

Run BP Gene Ontology Analysis: current study set of 1301 IDs … 88%  1,028 of
1,165 study items found in association
 90%  1,165 of  1,301 study items found in population(29107)
Calculating 12,433 uncorrected p-values using fisher
  12,433 GO terms are associated with 18,051 of 29,107 population items
   3,740 GO terms are associated with  1,028 of  1,301 study items
  METHOD fdr_bh:
      48 GO terms found significant (< 0.05=alpha) ( 46 enriched +   2
purified): statsmodels fdr_bh
     389 study items associated with significant GO IDs (enriched)
       3 study items associated with significant GO IDs (purified)

Run CC Gene Ontology Analysis: current study set of 1301 IDs … 93%  1,078 of
1,165 study items found in association
 90%  1,165 of  1,301 study items found in population(29107)
Calculating 1,755 uncorrected p-values using fisher
   1,755 GO terms are associated with 18,710 of 29,107 population items
     634 GO terms are associated with  1,078 of  1,301 study items
  METHOD fdr_bh:
      46 GO terms found significant (< 0.05=alpha) ( 46 enriched +   0
purified): statsmodels fdr_bh
     991 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 1301 IDs … 89%  1,036 of
1,165 study items found in association
 90%  1,165 of  1,301 study items found in population(29107)
Calculating 4,361 uncorrected p-values using fisher
   4,361 GO terms are associated with 17,827 of 29,107 population items
   1,094 GO terms are associated with  1,036 of  1,301 study items
  METHOD fdr_bh:
      23 GO terms found significant (< 0.05=alpha) ( 22 enriched +   1
purified): statsmodels fdr_bh
     867 study items associated with significant GO IDs (enriched)
       1 study items associated with significant GO IDs (purified)
Significant results[117] = 48 BP + 23 MF + 46 CC
    117 items WROTE: GO_analysis_upregulated.xlsx
    117 GOEA results for  1078 study items. WROTE: GO_analysis_upregulated.txt
Down
  EXISTS: go-basic.obo
  EXISTS: gene2go
go-basic.obo: fmt(1.2) rel(2021-07-02) 47,229 GO Terms
HMS:0:00:05.020085 341,880 annotations, 20,685 genes, 18,610 GOs, 1 taxids READ:
```

```
gene2go
CC 19,421 annotated human genes
BP 18,709 annotated human genes
MF 18,179 annotated human genes

Load BP Gene Ontology Analysis …
 71% 20,538 of 29,107 population items found in association

Load CC Gene Ontology Analysis …
 74% 21,427 of 29,107 population items found in association

Load MF Gene Ontology Analysis …
 70% 20,342 of 29,107 population items found in association

Run BP Gene Ontology Analysis: current study set of 1399 IDs … 89%  1,126 of
1,267 study items found in association
 91%  1,267 of  1,399 study items found in population(29107)
Calculating 12,433 uncorrected p-values using fisher
  12,433 GO terms are associated with 18,051 of 29,107 population items
   4,076 GO terms are associated with  1,126 of  1,399 study items
  METHOD fdr_bh:
      48 GO terms found significant (< 0.05=alpha) ( 45 enriched +   3
purified): statsmodels fdr_bh
     520 study items associated with significant GO IDs (enriched)
       4 study items associated with significant GO IDs (purified)

Run CC Gene Ontology Analysis: current study set of 1399 IDs … 92%  1,171 of
1,267 study items found in association
 91%  1,267 of  1,399 study items found in population(29107)
Calculating 1,755 uncorrected p-values using fisher
   1,755 GO terms are associated with 18,710 of 29,107 population items
     710 GO terms are associated with  1,171 of  1,399 study items
  METHOD fdr_bh:
      70 GO terms found significant (< 0.05=alpha) ( 70 enriched +   0
purified): statsmodels fdr_bh
   1,116 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 1399 IDs … 91%  1,150 of
1,267 study items found in association
 91%  1,267 of  1,399 study items found in population(29107)
Calculating 4,361 uncorrected p-values using fisher
   4,361 GO terms are associated with 17,827 of 29,107 population items
   1,213 GO terms are associated with  1,150 of  1,399 study items
  METHOD fdr_bh:
      40 GO terms found significant (< 0.05=alpha) ( 37 enriched +   3
purified): statsmodels fdr_bh
   1,024 study items associated with significant GO IDs (enriched)
```

```
      11 study items associated with significant GO IDs (purified)
Significant results[158] = 48 BP + 40 MF + 70 CC
    158 items WROTE: GO_analysis_downregulated.xlsx
    158 GOEA results for  1194 study items. WROTE: GO_analysis_downregulated.txt
```

[ ]: