main

August 23, 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
```

1.1 Functions

1.1.1 Cached functions

[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_bs_specific(tissue):
    df = pd.read_csv("../../_m/genes/brainseq_ancestry_4tissues_mashr.tsv",␣
 ↪sep='\t')
    df = df[(df["N_Regions_Shared"] == 1) & (df[tissue] == 1)].copy()
    df["ensemblID"] = df.Feature.str.replace("\\..*", "", regex=True)
    return df
```

```
@functools.lru_cache()
def convert2entrez(tissue):
    df = get_bs_specific(tissue)
    return df.merge(get_database(), left_on='ensemblID',␣
 ↪right_on='ensembl_gene_id')
```

1.1.2 Simple functions

[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(tissue):
    df = convert2entrez(tissue)
    geneids_study = {z[0]:z[1] for z in zip(df['entrezgene_id'],␣
 ↪df['external_gene_name'])}
    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
    # Save data
    label = tissue.lower().replace(" ", "_")
    goeaobj.wr_xlsx("GO_analysis_mashr_%s.xlsx" % label, goea_results_sig)
    goeaobj.wr_txt("GO_analysis_mashr_%s.txt" % label, goea_results_sig)
```

1.2 Gene ontology

[4]:
```
for tissue in ["Caudate", "Dentate Gyrus", "DLPFC", "Hippocampus"]:
    print(tissue)
    run_goea(tissue)
```

```
Caudate
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:0:00:04.429761 330,313 annotations, 20,685 genes, 18,684 GOs, 1 taxids READ:
gene2go
BP 18,505 annotated human genes
MF 18,190 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 683 IDs … 83%    564 of
683 study items found in association
100%    683 of    683 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
   2,530 GO terms are associated with    564 of    683 study items
  METHOD fdr_bh:
       6 GO terms found significant (< 0.05=alpha) (  6 enriched +   0
purified): statsmodels fdr_bh
      72 study items associated with significant GO IDs (enriched)
       0 study items associated with significant GO IDs (purified)

Run CC Gene Ontology Analysis: current study set of 683 IDs … 87%    597 of
683 study items found in association
100%    683 of    683 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
     493 GO terms are associated with    597 of    683 study items
  METHOD fdr_bh:
      16 GO terms found significant (< 0.05=alpha) ( 16 enriched +   0
```

```
purified): statsmodels fdr_bh
     521 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 683 IDs … 83%    570 of
683 study items found in association
100%    683 of    683 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
     815 GO terms are associated with    570 of    683 study items
  METHOD fdr_bh:
       4 GO terms found significant (< 0.05=alpha) (  4 enriched +   0
purified): statsmodels fdr_bh
     426 study items associated with significant GO IDs (enriched)
       0 study items associated with significant GO IDs (purified)
Significant results[26] = 6 BP + 4 MF + 16 CC
     26 items WROTE: GO_analysis_mashr_caudate.xlsx
     26 GOEA results for   574 study items. WROTE: GO_analysis_mashr_caudate.txt
Dentate Gyrus
  EXISTS: go-basic.obo
  EXISTS: gene2go
go-basic.obo: fmt(1.2) rel(2021-08-18) 47,217 GO Terms
HMS:0:00:04.463966 330,313 annotations, 20,685 genes, 18,684 GOs, 1 taxids READ:
gene2go
BP 18,505 annotated human genes
MF 18,190 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 710 IDs … 81%    578 of
710 study items found in association
100%    710 of    710 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
   2,760 GO terms are associated with    578 of    710 study items
  METHOD fdr_bh:
       5 GO terms found significant (< 0.05=alpha) (  5 enriched +   0
purified): statsmodels fdr_bh
     112 study items associated with significant GO IDs (enriched)
       0 study items associated with significant GO IDs (purified)
```

```
Run CC Gene Ontology Analysis: current study set of 710 IDs … 87%    617 of
710 study items found in association
100%    710 of    710 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
     475 GO terms are associated with    617 of    710 study items
  METHOD fdr_bh:
      15 GO terms found significant (< 0.05=alpha) ( 15 enriched +   0
purified): statsmodels fdr_bh
     401 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 710 IDs … 84%    596 of
710 study items found in association
100%    710 of    710 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
     796 GO terms are associated with    596 of    710 study items
  METHOD fdr_bh:
       3 GO terms found significant (< 0.05=alpha) (  3 enriched +   0
purified): statsmodels fdr_bh
     443 study items associated with significant GO IDs (enriched)
       0 study items associated with significant GO IDs (purified)
Significant results[23] = 5 BP + 3 MF + 15 CC
     23 items WROTE: GO_analysis_mashr_dentate_gyrus.xlsx
     23 GOEA results for   582 study items. WROTE:
GO_analysis_mashr_dentate_gyrus.txt
DLPFC
  EXISTS: go-basic.obo
  EXISTS: gene2go
go-basic.obo: fmt(1.2) rel(2021-08-18) 47,217 GO Terms
HMS:0:00:04.435358 330,313 annotations, 20,685 genes, 18,684 GOs, 1 taxids READ:
gene2go
BP 18,505 annotated human genes
MF 18,190 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 602 IDs … 82%    495 of
```

```
602 study items found in association
100%    602 of    602 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
   2,374 GO terms are associated with    495 of    602 study items
  METHOD fdr_bh:
       0 GO terms found significant (< 0.05=alpha) (  0 enriched +   0
purified): statsmodels fdr_bh
       0 study items associated with significant GO IDs (enriched)
       0 study items associated with significant GO IDs (purified)

Run CC Gene Ontology Analysis: current study set of 602 IDs … 89%    535 of
602 study items found in association
100%    602 of    602 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
     494 GO terms are associated with    535 of    602 study items
  METHOD fdr_bh:
      17 GO terms found significant (< 0.05=alpha) ( 17 enriched +   0
purified): statsmodels fdr_bh
     473 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 602 IDs … 86%    519 of
602 study items found in association
100%    602 of    602 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
     746 GO terms are associated with    519 of    602 study items
  METHOD fdr_bh:
       5 GO terms found significant (< 0.05=alpha) (  5 enriched +   0
purified): statsmodels fdr_bh
     420 study items associated with significant GO IDs (enriched)
       0 study items associated with significant GO IDs (purified)
Significant results[22] = 0 BP + 5 MF + 17 CC
     22 items WROTE: GO_analysis_mashr_dlpfc.xlsx
     22 GOEA results for   518 study items. WROTE: GO_analysis_mashr_dlpfc.txt
Hippocampus
  EXISTS: go-basic.obo
  EXISTS: gene2go
go-basic.obo: fmt(1.2) rel(2021-08-18) 47,217 GO Terms
HMS:0:00:04.468244 330,313 annotations, 20,685 genes, 18,684 GOs, 1 taxids READ:
gene2go
BP 18,505 annotated human genes
MF 18,190 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 599 IDs … 79%    474 of
599 study items found in association
100%    599 of    599 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
   2,185 GO terms are associated with    474 of    599 study items
  METHOD fdr_bh:
       0 GO terms found significant (< 0.05=alpha) (  0 enriched +   0
purified): statsmodels fdr_bh
       0 study items associated with significant GO IDs (enriched)
       0 study items associated with significant GO IDs (purified)

Run CC Gene Ontology Analysis: current study set of 599 IDs … 85%    509 of
599 study items found in association
100%    599 of    599 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
     453 GO terms are associated with    509 of    599 study items
  METHOD fdr_bh:
       4 GO terms found significant (< 0.05=alpha) (  4 enriched +   0
purified): statsmodels fdr_bh
     310 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 599 IDs … 81%    486 of
599 study items found in association
100%    599 of    599 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
     639 GO terms are associated with    486 of    599 study items
  METHOD fdr_bh:
       1 GO terms found significant (< 0.05=alpha) (  1 enriched +   0
purified): statsmodels fdr_bh
     359 study items associated with significant GO IDs (enriched)
       0 study items associated with significant GO IDs (purified)
Significant results[5] = 0 BP + 1 MF + 4 CC
      5 items WROTE: GO_analysis_mashr_hippocampus.xlsx
      5 GOEA results for   431 study items. WROTE:
GO_analysis_mashr_hippocampus.txt
```

[ ]: