Traineeship Part 1 NCBI-DPP9_continued_adapted

April 29, 2020

Traineeship Part 1: Data collection using NCBI eUtils and esummary CONTINUED/ADAPTED - DPP9

Author: Iris Raes

The University of Antwerp, Medical Biochemistry, Campus Drie Eiken

Loading required packages

[1]:
```
# pip3 install --user eutils
import MySQLdb as my
from eutils import Client
from Bio import Entrez
import csv
```

UCSC connection for ncbiRefSeq search (mRNA/RNA transcripts with hg19 coordinates)

[2]:
```
### Connection UCSC
db = my.connect(host="genome-euro-mysql.soe.ucsc.edu",
   user="genomep",
   passwd="password",
   db="hg19")
c = db.cursor()
### ncbiRefSeq search
no_rows = c.execute("""SELECT * FROM ncbiRefSeq WHERE name2 LIKE 'DPP9%'""")
result = c.fetchall()
### Close database
db.close()
```

[3]:
```
print("\nLoading currently available accession numbers from NCBI RefSeq table...
 ↪")
print("="*50)
print("\nTranscript variant accession numbers: ")
accList = []
### Save data to csv file
with open('results-transcripts-UCSC.csv', mode='w') as result_transcripts:
    result_writer = csv.writer(result_transcripts,delimiter=';')
```

```
    result_writer.
 ↪writerow(["chromosome","start","end","strand","gene","exonCount","accession"])
    for row in result:
        transcript = row[1]
        print(transcript)
        accList.append(transcript)
        starts = str(row[9])[2:-2]
        ends = str(row[10])[2:-2]
        starts1 = starts.split(",")
        ends1 = ends.split(",")
        j = 0
        for i in starts1:
            result_writer.
 ↪writerow([row[2],i,ends1[j],row[3],row[12],row[8],row[1]])
            j += 1
print("\nSearch results: {}\n".format(no_rows))
### Close csv file
result_transcripts.close()
```

```
Loading currently available accession numbers from NCBI RefSeq table…
==================================================

Transcript variant accession numbers:
NM_139159.5
NR_158699.2
NM_001365987.2
NR_164163.1

Search results: 4
```

Personal API-key for NCBI search

[4]:
```
eclient = Client(api_key="8ecce891e7fa036ff84bccc7c74e5138dc09")
```

---

1) Entrez Nucleotide Search - mRNA Transcript Variants

[5]:
```
### Creating query
mRNAtranscripts = []
transcriptmRNA_esearch = eclient.esearch(db='nucleotide',
            term='(DPP9[gene] AND "Homo sapiens"[Primary Organism] AND␣
 ↪refseq[filter]) NOT biomol_genomic[PROP]')
print("\nLoading currently available ids from Entrez nucleotide...")
print("="*50)
```

```
print("\nTranscript variant ids: ")
print(transcriptmRNA_esearch.ids)
for item in transcriptmRNA_esearch.ids:
    mRNAtranscripts.append(item)
print("\nSearch results: {}\n".format(transcriptmRNA_esearch.count))
```

```
Loading currently available ids from Entrez nucleotide…
=======================================================

Transcript variant ids:
[1370476185, 1034610004, 1034610002, 768004630, 768004626, 768004622, 768004618,
768004616, 578833714, 1677498370, 1677499978, 1700660497]

Search results: 12
```

[ ]:
```
### Esummary for retrieving information
Entrez.email = "iris.raes@hotmail.com"
### For each id in mRNAtranscripts
### Save data to csv file
with open('results-nucleotide.csv', mode='w') as result_nucleotide:
    result_writer = csv.writer(result_nucleotide,delimiter=';')
    result_writer.
 ↪writerow(["transcript_id","description","transcript_variant","accession","length_in_bp"])
    counter = 1
    for ids in mRNAtranscripts:
        handle = Entrez.esummary(db="nucleotide", id=ids)
        record = Entrez.read(handle)
        handle.close()
        ### Write info to csv file, row by row
        splittedtitle = record[0]["Title"].split(",")
        print(splittedtitle)
        result_writer.
 ↪writerow([record[0]["Id"],splittedtitle[0],splittedtitle[1],record[0]["AccessionVersion"],r
        ###
        counter += 1
### Close csv file
result_nucleotide.close()
```

2) dbVar Search - Pathogenic Copy Number Variation in Human

[6]:
```
### Creating query
CNV = []
CNV_esearch = eclient.esearch(db='dbVar',
            term='DPP9[All Fields] AND ("Homo sapiens"[Organism] AND "copy␣
 ↪number variation"[Variant Type] AND "Pathogenic"[clinical_interpretation])')
```

```
print("\nLoading currently available ids from dbVar...")
print("="*50)
print("dbVar ids: ")
print(CNV_esearch.ids)
for item in CNV_esearch.ids:
    CNV.append(item)
print("\nSearch results: {}\n".format(CNV_esearch.count))
```

```
Loading currently available ids from dbVar…
==================================================
dbVar ids:
[49623411, 49353191, 49353005, 49350830, 49349701, 49349293, 49345450, 49344315,
48468240, 48466558, 48466447, 48453939, 45807136, 17813982, 17813734, 3740775,
3739972, 3738955, 3738954, 3738649, 1212838, 1137112]

Search results: 22
```

[ ]:
```
### Esummary for retrieving information
Entrez.email = "iris.raes@hotmail.com"
### For each id in CNV
### Save data in csv file
with open('results-CNV-dbVar.csv', mode='w') as result_CNV:
    result_writer = csv.writer(result_CNV,delimiter=';')
    result_writer.
↪writerow(["CNV_variant_id","variant_region_id","type","study_ID","clinical_assertion","Chr_
    counter = 1
    for ids in CNV:
        handle = Entrez.esummary(db="dbVar", id=ids)
        record = Entrez.read(handle)
        handle.close()
        varregid = record['DocumentSummarySet']['DocumentSummary'][0].get('SV')
        types = record['DocumentSummarySet']['DocumentSummary'][0].
↪get('dbVarVariantTypeList')
        studyid = record['DocumentSummarySet']['DocumentSummary'][0].get('ST')
        clinicalassertion = record['DocumentSummarySet']['DocumentSummary'][0].
↪get('dbVarClinicalSignificanceList')
        if␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'] !=␣
↪[]:
            Chr_1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Chr')
            assembly1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Assembly')
```

```
            start1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Chr_start')
            end1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Chr_end')
            Chr_2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Chr')
            assembly2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Assembly')
            start2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Chr_start')
            end2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Chr_end')
        ### Write info to csv file, row by row
        result_writer.
↪writerow([ids,varregid,types,studyid,clinicalassertion,Chr_1,assembly1+":
↪"+start1+"-"+end1,Chr_2,assembly2+":"+start2+"-"+end2])
        ###
        counter += 1
### Close csv file
result_CNV.close()
```

3) dbVar Search - Insertions in Human

[7]:
```
### Creating query
insertion = []
insertion_esearch = eclient.esearch(db='dbVar',
            term='DPP9[All Fields] AND ("Homo sapiens"[Organism] AND␣
↪"insertion"[Variant Type])')
print("\nLoading currently available ids from dbVar...")
print("="*50)
print("dbVar ids: ")
print(insertion_esearch.ids)
for item in insertion_esearch.ids:
    insertion.append(item)
print("\nSearch results: {}\n".format(insertion_esearch.count))
```

```
Loading currently available ids from dbVar…
==================================================
dbVar ids:
```

```
[49597698, 49580472, 48530760, 48377645, 48377627, 47753859, 47564069, 47178696,
46791711, 45897195, 45896455, 45807279, 36885535, 24618684, 24516168, 24501143,
17814018, 17813982, 14212055, 14211117, 14209696, 13414404, 11399938, 8023314,
7738722, 7694891, 7590450, 7474153, 6477950, 6451851, 6354196, 5661470, 5431842,
5195919, 1297001, 1028299, 286824, 285317, 284926, 40396]

Search results: 40
```

[ ]:
```
### Esummary for retrieving information
Entrez.email = "iris.raes@hotmail.com"
### For each id in insertion
### Save data to csv file
with open('results-insertion-dbVar.csv', mode='w') as result_insertion:
    result_writer = csv.writer(result_insertion,delimiter=';')
    result_writer.
↪writerow(["insertion_variant_id","variant_region_id","type","study_ID","Chr_1","assembly1",
    counter = 1
    for ids in insertion:
        handle = Entrez.esummary(db="dbVar", id=ids)
        record = Entrez.read(handle)
        handle.close()
        varregid = record['DocumentSummarySet']['DocumentSummary'][0].get('SV')
        types = record['DocumentSummarySet']['DocumentSummary'][0].
↪get('dbVarVariantTypeList')
        studyid = record['DocumentSummarySet']['DocumentSummary'][0].get('ST')
        if␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'] !=␣
↪[]:
            Chr_1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Chr')
            assembly1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Assembly')
            start1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Chr_start')
            end1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Chr_end')
            Chr_2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Chr')
            assembly2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Assembly')
```

```
            start2 =␣
 ↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
 ↪get('Chr_start')
            end2 =␣
 ↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
 ↪get('Chr_end')
        ### Write info to csv file, row by row
        result_writer.writerow([ids,varregid,types,studyid,Chr_1,assembly1+":
 ↪"+start1+"-"+end1,Chr_2,assembly2+":"+start2+"-"+end2])
        ###
        counter += 1
### Close csv file
result_insertion.close()
```

4) dbVar Search - Inversions in Human

[8]:
```
### Creating query
inversion = []
inversion_esearch = eclient.esearch(db='dbVar',
            term='DPP9[All Fields] AND ("Homo sapiens"[Organism] AND␣
 ↪"inversion"[Variant Type])')
print("\nLoading currently available ids from dbVar...")
print("="*50)
print("dbVar ids: ")
print(inversion_esearch.ids)
for item in inversion_esearch.ids:
    inversion.append(item)
print("\nSearch results: {}\n".format(inversion_esearch.count))
```

```
Loading currently available ids from dbVar…
==================================================
dbVar ids:
[48377627, 47178696, 46791711, 45807289, 45807279, 36885535, 24618684, 24516168,
24501143, 17814018, 17813982, 5195919, 1297001, 1028299]

Search results: 14
```

[ ]:
```
### Esummary for retrieving information
Entrez.email = "iris.raes@hotmail.com"
### For each id in inversion
### Save data to csv file
with open('results-inversion-dbVar.csv', mode='w') as result_inversion:
    result_writer = csv.writer(result_inversion,delimiter=';')
    result_writer.
 ↪writerow(["inversion_variant_id","variant_region_id","type","study_ID","Chr_1","assembly1",
```

```
    counter = 1
    for ids in inversion:
        handle = Entrez.esummary(db="dbVar", id=ids)
        record = Entrez.read(handle)
        handle.close()
        varregid = record['DocumentSummarySet']['DocumentSummary'][0].get('SV')
        types = record['DocumentSummarySet']['DocumentSummary'][0].
 ↪get('dbVarVariantTypeList')
        studyid = record['DocumentSummarySet']['DocumentSummary'][0].get('ST')
        if␣
 ↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarRemappedAssemblyList']␣
 ↪!= []:
            assembly1 =␣
 ↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarRemappedAssemblyList'][0]
            assembly2 =␣
 ↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarRemappedAssemblyList'][0]
        ### Write info to csv file, row by row
        result_writer.
 ↪writerow([ids,varregid,types,studyid,"Chr19",assembly1,"Chr19",assembly2])
        ###
        counter += 1
### Close csv file
result_inversion.close()
```

5) dbVar Search - Short Tandem Repeats in Human (seems to be less important)

[9]:
```
### Creating query
STR = []
STR_esearch = eclient.esearch(db='dbVar',
            term='DPP9[All Fields] AND ("Homo sapiens"[Organism] AND "short␣
 ↪tandem repeat"[Variant Type])')
print("\nLoading currently available ids from dbVar...")
print("="*50)
print("dbVar ids: ")
print(STR_esearch.ids)
for item in STR_esearch.ids:
    STR.append(item)
print("\nSearch results: {}\n".format(STR_esearch.count))
```

```
Loading currently available ids from dbVar…
==================================================
dbVar ids:
[35728959, 35728956, 35728945, 35728942, 35728939, 35728922, 35728913, 35728902,
35728888, 35728883, 35728872, 35728679, 35728652, 35728650, 35728640, 35728610,
35728601, 35728076, 35727391, 35727380, 35727364, 35727355, 35727352, 35727332,
35727324, 35726686, 35726677, 35726669, 35726663, 35726639, 30349921]
```

```
Search results: 31
```

[ ]:
```
### Esummary for retrieving information
Entrez.email = "iris.raes@hotmail.com"
### For each id in STR
### Save data to csv file
with open('results-STR-dbVar.csv', mode='w') as result_STR:
    result_writer = csv.writer(result_STR,delimiter=';')
    result_writer.
↪writerow(["STR_variant_id","variant_region_id","type","study_ID","Chr_1","assembly1","Chr_2
    counter = 1
    for ids in STR:
        handle = Entrez.esummary(db="dbVar", id=ids)
        record = Entrez.read(handle)
        handle.close()
        varregid = record['DocumentSummarySet']['DocumentSummary'][0].get('SV')
        types = record['DocumentSummarySet']['DocumentSummary'][0].
↪get('dbVarVariantTypeList')
        studyid = record['DocumentSummarySet']['DocumentSummary'][0].get('ST')
        if␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'] !=␣
↪[]:
            Chr_1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Chr')
            assembly1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Assembly')
            start1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Chr_start')
            end1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Chr_end')
            Chr_2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Chr')
            assembly2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Assembly')
            start2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Chr_start')
```

```
            end2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Chr_end')
        ### Write info to csv file, row by row
        result_writer.writerow([ids,varregid,types,studyid,Chr_1,assembly1+":
↪"+start1+"-"+end1,Chr_2,assembly2+":"+start2+"-"+end2])
        ###
        counter += 1
### Close csv file
result_STR.close()
```

6) ClinVar Search - Genetic Variations in Human

[10]:
```
### Creating query
ClinVar = []
ClinVar_esearch = eclient.esearch(db='ClinVar',
            term='DPP9[gene] AND "Single gene"')
print("\nLoading currently available ids from ClinVar...")
print("="*50)
print("\nClinVar ids: ")
print(ClinVar_esearch.ids)
for item in ClinVar_esearch.ids:
    ClinVar.append(item)
print("\nSearch results: {}\n".format(ClinVar_esearch.count))
```

```
Loading currently available ids from ClinVar…
==================================================

ClinVar ids:
[788833, 779179, 778595, 769947, 717743, 713315, 615908]

Search results: 7
```

[ ]:
```
### Esummary for retrieving information
Entrez.email = "iris.raes@hotmail.com"
### For each id in ClinVar
### Save data to csv file
with open('results-ClinVar.csv', mode='w') as result_ClinVar:
    result_writer = csv.writer(result_ClinVar,delimiter=';')
    result_writer.
↪writerow(["ClinVar_variant_id","title","accession","type","description","protein_change","C
    counter = 1
    for ids in ClinVar:
        handle = Entrez.esummary(db="ClinVar", id=ids)
        record = Entrez.read(handle)
```

```
        handle.close()
        title = record['DocumentSummarySet']['DocumentSummary'][0].get('title')
        accession = record['DocumentSummarySet']['DocumentSummary'][0].
↪get('accession_version')
        types = record['DocumentSummarySet']['DocumentSummary'][0].
↪get('obj_type')
        description =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['clinical_significance'].
↪get('description')
        protein_change = record['DocumentSummarySet']['DocumentSummary'][0].
↪get('protein_change')
        if␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['variation_set'][0]['variation_loc']␣
↪!= []:
            Chr_1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['variation_set'][0]['variation_loc'][0].
↪get('chr')
            assembly1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['variation_set'][0]['variation_loc'][0].
↪get('assembly_name')
            start1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['variation_set'][0]['variation_loc'][0].
↪get('start')
            end1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['variation_set'][0]['variation_loc'][0].
↪get('stop')
            if␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['variation_set'][0].
↪get('variation_loc') != []:
                try:
                    Chr_2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['variation_set'][0]['variation_loc'][1].
↪get('chr')
                    assembly2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['variation_set'][0]['variation_loc'][1].
↪get('assembly_name')
                    start2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['variation_set'][0]['variation_loc'][1].
↪get('start')
                    end2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['variation_set'][0]['variation_loc'][1].
↪get('stop')
                except:
                    assembly2 = "not applicable"
                    start2 = "X"
                    end2 = "X"
```

```
        if record['DocumentSummarySet']['DocumentSummary'][0]['trait_set'] !=␣
↪[]:
            dbsource =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['trait_set'][0]['trait_xrefs'][0]['db_so
            dbid =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['trait_set'][0]['trait_xrefs'][0]['db_id
        ### Write info to csv file, row by row
        result_writer.
↪writerow([ids,title,accession,types,description,protein_change,Chr_1,assembly1+":
↪"+start1+"-"+end1,Chr_2,assembly2+":"+start2+"-"+end2,dbsource+"␣
↪("+dbid+")"])
        ###
        counter += 1
### Close csv file
result_ClinVar.close()
```