Using NCBI edirect to access data at NCBI

NCBI has developed the edirect command line application for interfacing with their system. We can use that to query NCBI from the command line.

For many of these examples, check out the EDirect Cookbook and the NCBI EDirect book

For all of the `eutils`, you can add the option `-help` to the command to get more information about that command. For example, `efetch -help` will give you lots of information about the `efetch` command.

For each of these commands you need to tell NCBI what your email address is. The AWS instances have a generic address of `ec2-user@ip-xxx-xxx-xxx-xxx.us-east-2.compute.internal` but obviously that is not meaningful. NCBI requires an email address so they can contact you if you write a script that goes off the rails and starts DOS attacking them (also, as I know from personal experience, they will simply block your IP so you can no longer access their services!)

NCBI Databases

Before we start, we need to know which databases are available. You can find a list of databases using `einfo`:

```
einfo -email xxx@sdsu.edu -dbs | sort
```

Note: you need to specify your own email address here!

annotinfo
assembly
biocollections
bioproject
biosample
biosystems
blastdbinfo
books
cdd
clinvar
clone
dbvar
gap
gapplus
gds
gencoll
gene
genome
geoprofiles
grasp
gtr
homologene
medgen
mesh
ncbisearch
nlmcatalog
nuccore
nucest
nucgss
nucleotide
omim
orgtrack
pcassay
pccompound
pcsubstance
pmc
popset
probe
protein
proteinclusters
pubmed
pubmedhealth
seqannot
snp
sparcle
sra
structure
taxonomy
unigene

We can find out more about any of these databases using the einfo command again:

```
einfo -email xxx@sdsu.edu -db assembly | less
```

However, this prints out XML, which is not very readable. Instead, eutils from NCBI comes with a program called xtract that converts XML into text. We can use that to extract the name and description associated with the fields in a database:

```
einfo -email xxx@sdsu.edu -db assembly -fields
```

or

```
einfo -email xxx@sdsu.edu -db assembly | xtract -pattern Field -element Name Description
```

You can do this with any of the databases listed above to see what fields those databases have. This is the output from the assembly database.

Name	Description
ALL	All terms from all searchable fields
UID	Unique number assigned to publication
FILT	Limits the records
ACCN	Chromosome accessions
ASAC	Space delimited assembly accessions w/ & w/o versions
ASLV	How assembled is this assembly. ‘Contig’ to ‘Chromosome’
TXID	Taxonomy ID
ORGN	Exploded organism names
RUID	Id of RefSeq Assembly.
GUID	Id of GenBank synonym of this Assembly.
UIDS	Pair-id, GB-id, and RS-id of this Assembly.
PROJ	Uid and accessions of this assembly’s projects
SAMP	BioSample Accession and Id
NAME	Assembly name
ALLN	All names, space separated
DESC	Assembly description
COV	Sequencing coverage
TYPE	Type of the assembly
SRDT	Date the most recent sequence went live in ID
UPDT	Date the assembly was last updated
LEN	Total length of chromosome/genome including bases and gaps divided by 1,000,000.
REPL	Number of chromosomes in assembly
PLAC	Number of placed scaffolds
UNLO	Number of unordered(unlocalized) scaffolds belonging to chromosomes
UNPL	Number of unplaced scaffolds which do not belong to any chromosome, ie ChrUn
CN50	Contig length at which 50% of total bases in assembly are in contigs of that length or greater
SN50	Scaffold length at which 50% of total bases in assembly are in contigs of that length or greater
CL50	Number of contigs that are greater than or equal to the N50 length.
SL50	Number of scaffolds that are greater than or equal to the N50 length.
CNTG	Number of contigs
UNGL	Total length excluding gaps in chromosome/genome divided by 1,000,000
PROP	Properties

Name	Description
SUBO	Organization that submitted this assembly
INFR	Infraspecific Name: breed, cultivar, strain, ecotype
ISOL	Isolate name
SEX	Sex
ASMM	Assembly Method
GCOV	Genome Coverage
TECH	Sequencing Technology
EXFV	Expected Final Version
RGAS	Reference Guided Assembly
SCAM	Single Cell Amplification
RCAT	RefSeq Category
FTYP	From Type Material
NFRS	Reasons assembly was excluded from RefSeq
GRLS	Date the GenBankassembly was first released
RRLS	Date the RefSeq assembly was first released
RTYP	Release Type
RLEN	Total length of chromosome/genome including bases and gaps

Downloading Genomes

Let's start with a simple search. The command esearch allows us to search the databases. For example to search through the assembly database, we can use:

```
esearch -db assembly -query "Faecalibacterium prausnitzii[ORGN]"
```

This should give you an answer like:

```
<ENTREZ_DIRECT>
  <Db>assembly</Db>
  <WebEnv>NCID_1_27641186_130.14.18.34_9001_1536599806_648492441_0MetA0_S_MegaStore</WebEnv>
  <QueryKey>1</QueryKey>
  <Count>49</Count>
  <Step>1</Step>
</ENTREZ_DIRECT>
```

The key field in this response is the **`<Count>49</Count>`** field – this shows you how many things match your query,

We can use `efetch` to get those matches to the query, and we use the document summary format of `efetch` to summarize the document.

```
esearch -db assembly -query "Faecalibacterium prausnitzii[ORGN]" | efetch -format docsum |
```

This is still XML format, and so to extract specific elements from that output, we can use xtract again. Here, we are looking for elements that contain a link to the NCBI ftp site:

```
<FtpPath_GenBank>ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/003/467/805/GCA_003467805.1_ASM3
<FtpPath_RefSeq>ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/467/805/GCF_003467805.1_ASM34
<FtpPath_Assembly_rpt>ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/467/805/GCF_003467805.1
<FtpPath_Stats_rpt>ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/467/805/GCF_003467805.1_AS
```

This is the path to different formats of the file We want the annotated sequences in [RefSeq](, and you can view that path in your browser by looking at this link: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/467/805/GCF_003467805.1_ASM346780v1/.

For these entries, the RefSeq file will be called the same as the name of the directory, with "_genomic.fna.gz" appended.

Thus, for the above entry, the RefSeq nucleotide sequence file is: ‘ GCF_003467805.1_ASM346780v1_genomic.fna.gz

This xtract syntax will get just the URLs for all the entries:

```
xtract -pattern DocumentSummary -element FtpPath_RefSeq
```

Which gives us a list like:

```
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/467/805/GCF_003467805.1_ASM346780v1
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/465/525/GCF_003465525.1_ASM346552v1
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/434/165/GCF_003434165.1_ASM343416v1
...
```

Note: I just trimmed this to the first three elements in the list

We can write a simple program using awk to append the directory name and `_genomic.fna.gz` on the end of these URLs:

```
esearch -db assembly -query "Faecalibacterium prausnitzii[ORGN]" | efetch -format docsum | \
xtract -pattern DocumentSummary -element FtpPath_RefSeq | \
awk -F"/" '{print $0"/"$NF"_genomic.fna.gz"}'
```

In this awk command, `$0` is the thing that was piped to the command. `$NF` is the last element in the record and we have said to split with a `/` using the option `-F"/"`.

So that command gives us this list (again trimmed to the first three entries):

```
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/467/805/GCF_003467805.1_ASM346780v1/GCF_0034
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/465/525/GCF_003465525.1_ASM346552v1/GCF_0034
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/434/165/GCF_003434165.1_ASM343416v1/GCF_0034
```

There are several programs that you can use to download these files. One I recommend is `wget` as it is very straightforward. However, I also recommend using `curl` as it has a lot more options!

For example, we can modify the above command to use curl to get the files and save them with their original names:

```
esearch -db assembly -query "Faecalibacterium prausnitzii[ORGN]" | efetch -format docsum | \
xtract -pattern DocumentSummary -element FtpPath_RefSeq | \
awk -F"/" '{print "curl -o "$NF"_genomic.fna.gz " $0"/"$NF"_genomic.fna.gz"}'
```

Downloading Reference Databases

If you are working with shotgun metagenomic data you will eventually have to make sense of that data in a taxonomic and/or functional context. Therefore knowing how to access large amounts of reference databases quickly and efficiently is a must. Below are some examples of how you can accomplish this.

1. Downloading the NCBI Virus RefSeq Project Database

```
esearch -db bioproject -query "PRJNA485481" | elink -target nuccore | efetch -format fasta
```

If you click on the NCBI virus link above you will see that accession number will be the query name you use to download the viral database which in this case is *PRJNA485481.

You could also download this same database using the database id:

```
esearch -db bioproject -query "485481[id]" | elink -target nuccore | efetch -format fasta >
```

Also note that I wanted to get the nucleotide database so my link target is *nuccore* but you can easily change that to *protein* if that is your desired database like the example below.

2. Downloading the Bacterial Antimicrobial Resistance Reference Gene Database from NCBI

```
esearch -db bioproject -query "PRJNA313047" | elink -target protein | efetch -format fasta >
```

3. Downloading all Bacterial genomes in RefSeq Your query can also be a list of filters allowing you to download only certain datasets from NCBI - in this case it is all bacterial nucleotide genomes found in the RefSeq database.

```
esearch  -db "nucleotide" -query "Bacteria[Organism] OR bacteria[All Fields] AND Refseq[Filt
```