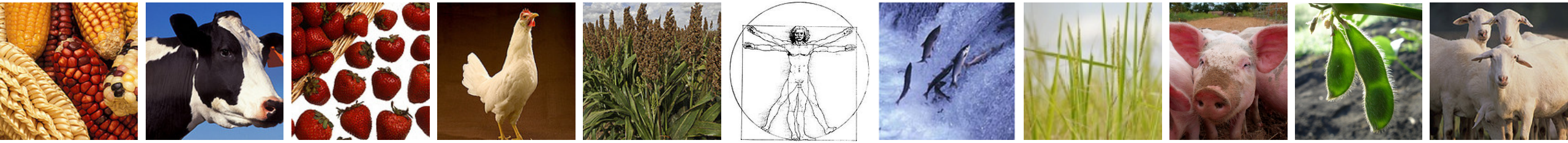

Programmatic Access to NCBI Resources

Vamsi K. Kodali (kodalivk@nih.gov)

NCBI Genome Resources Workshop

PAG XXVII January 14, 2019

U.S. National Library of Medicine
National Center for Biotechnology Information

ncbi.nlm.nih.gov/home/coursesandwebinars

bit.ly/entrez-direct

bit.ly/ncbi_factsheets

support.nlm.nih.gov

$ esearch -help

NCBI booth #223

EDirect

`bit.ly/entrez-direct`

- access NCBI databases from the UNIX terminal
- build multi-step pipelines
- compatible with other UNIX utilities
- extract specific data in tabular format

Installation

bit.ly/install-edirect-windows
bit.ly/install-edirect-mac_linux

Installation

EDirect will run on UNIX and Macintosh computers that have the Perl language installed, and under the Cygwin UNIX-emulation environment on Windows PCs. To install the EDirect software, copy the following commands and paste them into a terminal window:

```
cd ~
/bin/bash
perl -MNet::FTP -e \
  '$ftp = new Net::FTP("ftp.ncbi.nlm.nih.gov", Passive => 1);
   $ftp->login; $ftp->binary;
   $ftp->get("/entrez/entrezdirect/edirect.tar.gz");'
gunzip -c edirect.tar.gz | tar xf -
rm edirect.tar.gz
builtin exit
export PATH=${PATH}:$HOME/edirect >& /dev/null || setenv PATH "${PATH}:$HOME/edirect"
./edirect/setup.sh
```

This downloads several scripts into an "edirect" folder in the user's home directory. The setup.sh script then downloads any missing Perl modules, and may print an additional command for updating the PATH environment variable in the user's configuration file. Copy that command, if present, and paste it into the terminal window to complete the installation process. The editing instructions will look something like:

```
echo "export PATH=\$PATH:\$HOME/edirect" >> $HOME/.bash_profile
```


Handout available

EDirect commands

Tool	Description
`esearch`	Entrez searches
`efilter`	Filter results
`elink`	Related data within the same or another database
`esummary`	Document summary downloads
`efetch`	Data record downloads
`einfo`	Information about databases and indexed fields
`epost`	Upload UIDs or Accessions
`nquire`	Send a URL request to a web page or CGI service
`xtract`	Extract data from XML into a table

EDirect commands

Tool	Description
`esearch`	Entrez searches
`efilter`	Filter results
`elink`	Related data within the same or another database
`esummary`	Document summary downloads
`efetch`	Data record downloads
`einfo`	Information about databases and indexed fields
`epost`	Upload UIDs or Accessions
`nquire`	Send a URL request to a web page or CGI service
`xtract`	Extract data from XML into a table

esearch

- Performs a text search and returns a list of UIDs
- Search for genes with a symbol beginning with QSOX in cow

esearch

- Performs a text search and returns a list of UIDs
- Search for genes with a symbol beginning with QSOX in cow

```
$ esearch -db gene -q "QSOX*[Gene Symbol] AND bos taurus[Organism] AND
alive[Properties]"

<ENTREZ_DIRECT>
  <Db>gene</Db>
  <WebEnv>NCID_1_22473879_130.14.18.34_..._0MetA0_S_MegaStore</WebEnv>
  <QueryKey>1</QueryKey>
  <Count>2</Count>
  <Step>1</Step>
</ENTREZ_DIRECT>
```

esearch

- Performs a text search and returns a list of UIDs
- Search for genes with a symbol beginning with QSOX in cow

```
$ esearch -db gene -q "QSOX*[Gene Symbol] AND bos taurus[Organism] AND
alive[Properties]"

<ENTREZ_DIRECT>
  <Db>gene</Db>
  <WebEnv>NCID_1_22473879_130.14.18.34_..._0MetA0_S_MegaStore</WebEnv>
  <QueryKey>1</QueryKey>
  <Count>2</Count>
  <Step>1</Step>
</ENTREZ_DIRECT>
```

esummary

- retrieves a brief summary called ‘Document Summary’ for a list of UIDs

```
$ esearch -db gene -q "QSOX*[Gene Symbol]
      AND bos taurus[Organism]
      AND alive[Properties]" \
  | esummary
```

esummary

- For a given list of UIDs, retrieves a brief summary called ‘Document Summary’

```
$ esearch -db gene -q "QSOX*[Gene Symbol]
      AND bos taurus[Organism]
      AND alive[Properties]" \
  | esummary
```


```
▼<eSummaryResult>
  ▼<DocumentSummarySet status="OK">
    <DbBuild>Build190102-2115m.1</DbBuild>
    ▼<DocumentSummary uid="522986">
      <Name>QSOX1</Name>
      <Description>quiescin sulfhydryl oxidase 1</Description>
      <Status>0</Status>
      <CurrentID>0</CurrentID>
      <Chromosome>16</Chromosome>
      <GeneticSource>genomic</GeneticSource>
      <MapLocation/>
      <OtherAliases/>
      ▼<OtherDesignations>
        sulfhydryl oxidase 1|quiescin Q6 sulfhydryl oxidase 1
      </OtherDesignations>
      <NomenclatureSymbol>QSOX1</NomenclatureSymbol>
      <NomenclatureName>quiescin sulfhydryl oxidase 1</NomenclatureName>
      <NomenclatureStatus>Official</NomenclatureStatus>
      <Mim> </Mim>
      ▼<GenomicInfo>
        ▼<GenomicInfoType>
          <ChrLoc>16</ChrLoc>
          <ChrAccVer>NC_037343.1</ChrAccVer>
          <ChrStart>61363054</ChrStart>
          <ChrStop>61404621</ChrStop>
          <ExonCount>12</ExonCount>
        </GenomicInfoType>
      </GenomicInfo>
      <GeneWeight>445</GeneWeight>
      <Summary/>
      <ChrSort>16</ChrSort>
      <ChrStart>61363054</ChrStart>
      ▼<Organism>
        <ScientificName>Bos taurus</ScientificName>
        <CommonName>cattle</CommonName>
        <TaxID>9913</TaxID>
      </Organism>
      ▼<LocationHist>
        ▼<LocationHistType>
          <AnnotationRelease>106</AnnotationRelease>
          <AssemblyAccVer>GCF_002263795.1</AssemblyAccVer>
          <ChrAccVer>NC_037343.1</ChrAccVer>
          <ChrStart>61363054</ChrStart>
          <ChrStop>61404621</ChrStop>
```

esummary

```
$ esearch -db gene -q "QSOX*[Gene Symbol]
      AND bos taurus[Organism]
      AND alive[Properties]" \
  | esummary
```

```
▼<DocumentSummary uid="522986">
    <Name>QSOX1</Name>
    <Description>quiescin sulfhydryl oxidase 1</Description>
    <Status>0</Status>
    <CurrentID>0</CurrentID>
    <Chromosome>16</Chromosome>
    <GeneticSource>genomic</GeneticSource>
    <MapLocation/>
    <OtherAliases/>
```

```
▼<eSummaryResult>
  ▼<DocumentSummarySet status="OK">
    <DbBuild>Build190102-2115m.1</DbBuild>
    ▼<DocumentSummary uid="522986">
      <Name>QSOX1</Name>
      <Description>quiescin sulfhydryl oxidase 1</Description>
      <Status>0</Status>
      <CurrentID>0</CurrentID>
      <Chromosome>16</Chromosome>
      <GeneticSource>genomic</GeneticSource>
      <MapLocation/>
      <OtherAliases/>
      ▼<OtherDesignations>
        sulfhydryl oxidase 1|quiescin Q6 sulfhydryl oxidase 1
      </OtherDesignations>
      <NomenclatureSymbol>QSOX1</NomenclatureSymbol>
      <NomenclatureName>quiescin sulfhydryl oxidase 1</NomenclatureName>
      <NomenclatureStatus>Official</NomenclatureStatus>
      <Mim> </Mim>
      ▼<GenomicInfo>
```

```
rAccVer>
rt>
>
/ScientificName>
ame>
notationRelease>
<AssemblyAccVer>GCF_002263795.1</AssemblyAccVer>
<ChrAccVer>NC_037343.1</ChrAccVer>
<ChrStart>61363054</ChrStart>
<ChrStop>61404621</ChrStop>
```

xtract

- parses XML and extracts specified elements into a table

```
▼<DocumentSummary uid="522986"
    <Name>QSOX1</Name>
    <Description>quiescin sulfhy
    <Status>0</Status>
    <CurrentID>0</CurrentID>
    <Chromosome>16</Chromosome>
    <GeneticSource>genomic</Gene
    <MapLocation/>
    <OtherAliases/>
```

```
$ esearch -db gene -q "QSOX*[Gene Symbol]
    AND bos taurus[Organism]
    AND alive[Properties]" \
  | esummary \
  | xtract -pattern DocumentSummary \
     -element Name,Description,Chromosome

QSOX1  quiescin sulfhydryl oxidase 1  16
QSOX2  quiescin sulfhydryl oxidase 2  11
```

xtract

- structured data extraction
- nested exploration
- conditionals
- INSDC feature extraction
- compatible with other UNIX utilities

bit.ly/entrez-direct

```
$ xtract -help
```

Handout available

efetch

- retrieves data records for a given list of UIDs
- pipe output of esearch to efetch to get information about the two QSOX genes

```
$ esearch -db gene -q "QSOX*[Gene Symbol] AND bos taurus[Organism] AND
alive[Properties]" | efetch

1. QSOX1
Official Symbol: QSOX1 and Name: quiescin sulfhydryl oxidase 1 [Bos taurus (cattle)]
Other Designations: sulfhydryl oxidase 1; quiescin Q6 sulfhydryl oxidase 1
Chromosome: 16
Annotation: Chromosome 16 NC_037343.1 (61363055..61404622)
ID: 522986
…
```

efetch

```
$ efetch -help
```

```
gene
                full_report                     Detailed Report
                gene_table                      Gene Table
                native                          Gene Report
                native              asn.1       Entrezgene ASN.1
                native              xml         Entrezgene-Set XML
                tabular                         Tabular Report
```

elink

- explore links between UIDs
- can be within the same or another database
- Example:
  - find all RefSeq RNAs annotated on cow QSOX genes

elink

NCBI Resources How To vkkodali My NCBI Sign Out

Gene

Gene | QSOX*[Gene Symbol] AND bos taurus[Organism] | Search

Create RSS Create alert Advanced Help

Gene sources
Genomic

Categories
Alternatively spliced
Annotated genes
Protein-coding

Sequence content
Ensembl
RefSeq

Status clear
✓ Current

Clear all

Show additional filters

Tabular Sort by Relevance Send to:

Hide sidebar >>

Filters: Manage Filters

Search results

Items: 2

Showing Current items.

Name/Gene ID	Description	Location	Aliases
QSOX1 ID: 522986	quiescin sulfhydryl oxidase 1 [*Bos taurus* (cattle)]	Chromosome 16, NC_037343.1 (61363055..61404622)	
QSOX2 ID: 788063	quiescin sulfhydryl oxidase 2 [*Bos taurus* (cattle)]	Chromosome 11, NC_037338.1 (103691351..103716243, complement)	

Find related data

Database: Select

GEO Profiles
GSS
GTR
HomoloGene
MedGen
Nucleotide
OMIM
PMC
Probe
Protein

Search d

Recent a

elink

```
$ esearch -db gene -q "QSOX*[Gene Symbol]
    AND bos taurus[Organism] AND alive[Properties]" \
  | elink -db gene -target nuccore -name gene_nuccore_refseqrna \
  | efetch -db nuccore -format acc

NM_001102074.2
XM_024976246.1
XM_002691616.4
```

bit.ly/elink-descriptions

elink

```
$ esearch -db gene -q "QSOX*[Gene Symbol]
    AND bos taurus[Organism] AND alive[Properties]" \
  | elink -db gene -target nuccore -name gene_nuccore_refseqrna \
  | efetch -db nuccore -format fasta

>NM_001102074.2 Bos taurus quiescin sulfhydryl oxidase 1 (QSOX1), mRNA
AAGTCTTCCGAAAACTTGTGCGCAGCGGTGGCCTCGGCGCGGCGGCCGAGGATGGGGTGGTGC...

>XM_024976246.1 PREDICTED: Bos taurus quiescin sulfhydryl oxidase 1
(QSOX1), transcript variant X1, mRNA
GGGCGGGGCCGGACCATGGCGGGCGGGGTCTGCGGGCTCCACCTCCCGACGGAGGCAGGAGGT...

>XM_002691616.4 PREDICTED: Bos taurus quiescin sulfhydryl oxidase 2
(QSOX2), mRNA
CGCGGCACTCCCAACATGGCGGCGGCCAGGACGGCGGCGTCGGCAGCGCGCAGCCCGGGAGCT...
```

Example: Gene ranges – task description

- Get gene range sequences in FASTA format for cow QSOX genes
  - esearch – search for QSOX genes in cow
  - esummary – to get the gene docsum
  - xtract – parse docsum XML to extract chr, start and stop
  - efetch – to fetch nucleotide sequences in FASTA format

Example: Gene ranges – command

```
$ esearch -db gene -q "QSOX*[Gene Symbol]
    AND bos taurus[Organism]
    AND alive[Properties]" \
  | esummary \
  | xtract -pattern DocumentSummary \
    -if GenomicInfoType -element Id \
    -block GenomicInfoType \
    -element ChrAccVer ChrStart ChrStop \
  | while read -r gene_id chr_acc chr_start chr_stop ;
    do
      efetch -db nuccore \
             -id $chr_acc \
             -chr_start $chr_start \
             -chr_stop $chr_stop \
             -format fasta ;
    done
```

```
▼<GenomicInfo>
  ▼<GenomicInfoType>
      <ChrLoc>16</ChrLoc>
      <ChrAccVer>NC_037343.1</ChrAccVer>
      <ChrStart>61363054</ChrStart>
      <ChrStop>61404621</ChrStop>
      <ExonCount>12</ExonCount>
    </GenomicInfoType>
  </GenomicInfo>
```

Example: Gene ranges – output

```
>NC_037343.1:61363055-61404622 Bos taurus isolate L1 Dominette 01449 registration
number 42190680 breed Hereford chromosome 16, ARS-UCD1.2, whole genome shotgun
sequence
GGGCGGGGCCGGACCATGGCGGGCGGGGTCTGCGGGCTCCACCTCCCGACGGAGGCAGGAGGTGCCGCGG...

>NC_037338.1:c103716243-103691351 Bos taurus isolate L1 Dominette 01449
registration number 42190680 breed Hereford chromosome 11, ARS-UCD1.2, whole
genome shotgun sequence
CGCGGCACTCCCAACATGGCGGCGGCCAGGACGGCGGCGTCGGCAGCGCGCAGCCCGGGAGCTCCGGCGG...
```

Example: CDD domains

```
$ esearch -db gene -q "QSOX*[Gene Symbol]
    AND bos taurus[Organism]
    AND alive[Properties]" \
  | elink -db gene \
      -target protein \
      -name gene_protein_refseq \
  | efetch -db protein \
      -format gpc \
  | xtract -insd Region \
      INSDInterval_from \
      INSDInterval_to \
      region_name \
      db_xref \
  | grep 'CDD:'
```

NP_001095544.1	41	153	PDI_a_QSOX	CDD:239290
NP_001095544.1	407	502	Evr1_Alr	CDD:282612
XP_024832014.1	82	194	PDI_a_QSOX	CDD:239290
XP_024832014.1	448	540	Evr1_Alr	CDD:309769
XP_002691662.2	63	175	PDI_a_QSOX	CDD:239290
XP_002691662.2	433	530	Evr1_Alr	CDD:309769

Downloading genomic data

- For whole genomes, FTP is the best source
- NCBI Assembly provides stable, accessioned data
  - GCA for GenBank, GCF for RefSeq
- RefSeq Assemblies have annotation
- Need to know the FTP path
  - EDirect, NCBI Assembly Portal, Assembly Summary Files

Handout available

1. EDirect for Assembly FTP paths

```
$ esearch -db assembly -q 'bos taurus[Organism] AND latest[Filter]’ \
  | esummary \
  | xtract -pattern DocumentSummary
     -element AssemblyAccession,FtpPath_RefSeq

GCF_002263795.1 …/genomes/all/GCF/002/263/795/GCF_002263795.1_ARS-UCD1.2
GCF_000003205.7 …/genomes/all/GCF/000/003/205/GCF_000003205.7_Btau_5.0.1
GCF_000003055.6 …/genomes/all/GCF/000/003/055/GCF_000003055.6_Bos_taurus_UMD_3.1.1
```

2. NCBI Assembly

NCBI Resources How To

Assembly

Assembly | bos [organism]

Create alert Advanced Browse by organism

Organism group
Animals (9)
Customize ...

Status clear
Latest (9)
Latest GenBank (9)
Latest RefSeq (4)

Assembly level
Complete genome (0)
Chromosome (8)
Scaffold (1)
Contig (0)

RefSeq category
Reference (0)
Representative (4)

Exclude clear

Summary 20 per page Sort by Significance

Download Assemblies

Search results

Items: 9

Filters activated: Latest, Exclude anomalous. Clear all to show 22 items.

1. Bos indicus 1.0
Organism: Bos indicus (zebu cattle)
Infraspecific name: Breed: Nelore
Sex: male
Submitter: Genoa Biotecnologia SA
Date: 2014/11/25
Assembly level: Chromosome
Genome representation: full
RefSeq category: representative genome

2. NCBI Assembly

NCBI Resources How To

Assembly

Assembly | bos [organism]

Create alert Advanced Browse by organism

Organism group
Animals (9)
Customize ...

Status clear
Latest (9)
Latest GenBank (9)
Latest RefSeq (4)

Assembly level
Complete genome (0)
Chromosome (8)
Scaffold (1)
Contig (0)

RefSeq category
Reference (0)
Representative (4)

Exclude clear

Summary 20 per page Sort by Significance

Download Assemblies

Search results

Items: 9

Filters activated: Latest, Exclude anomalous. Clear all to show 22 items.

1. Bos indicus 1.0
Organism: Bos indicus (zebu cattle)
Infraspecific name: Breed: Nelore
Sex: male
Submitter: Genoa Biotecnologia SA
Date: 2014/11/25
Assembly level: Chromosome
Genome representation: full
RefSeq category: representative genome

Source database (GenBank or RefSeq)
RefSeq

File type
Genomic GFF

Estimated size is 91.1 MB

Download

2. NCBI Assembly

NCBI Resources How To

Assembly

Assembly | bos [organism]

Create alert Advanced Browse by organism

Organism group
Animals (9)
Customize ...

Status clear
✓ Latest (9)
Latest GenBank (9)
Latest RefSeq (4)

Assembly level
Complete genome (0)
Chromosome (8)
Scaffold (1)
Contig (0)

RefSeq category
Reference (0)
Representative (4)

Exclude clear

Summary 20 per page Sort by Significance

Download Assemblies

Search results

Items: 9

Filters activated: Latest, Exclude anomalous. Clear all to show 22 items.

1. Bos indicus 1.0
Organism: Bos indicus (zebu cattle)
Infraspecific name: Breed: Nelore
Sex: male
Submitter: Genoa Biotecnologia SA
Date: 2014/11/25
Assembly level: Chromosome
Genome representation: full
RefSeq category: representative genome

Source database (GenBank or RefSeq)
RefSeq

File type
Genomic GFF

Estimated size is 91.1 MB

Download

All file types (including assembly-structure directory)
Assembly regions report
Assembly structure report
Assembly statistics report
CDS from genomic
Feature count
Feature table
Genomic FASTA
Genomic GenBank format
Genomic GFF
Protein FASTA
Protein GenPept format
RNA FASTA
RNA GenBank format
RNA from genomic
RepeatMasker output
RepeatMasker run info
Translated CDS
WGS-master

Feature Counts

	A	B	C	D	E	F	G
1	# Feature	Class	Full Assembly	Assembly-unit accession	Assembly-unit name	Unique Ids	Placements
2	CDS	with_protein	GCF_002263795.1	GCF_002263805.1	Primary Assembly	63674	63683
3	CDS	without_protein	GCF_002263795.1	GCF_002263805.1	Primary Assembly	na	176
4	C_region		GCF_002263795.1	GCF_002263805.1	Primary Assembly	na	18
5	V_segment		GCF_002263795.1	GCF_002263805.1	Primary Assembly	na	159
6	gene	C_region	GCF_002263795.1	GCF_002263805.1	Primary Assembly	18	18
7	gene	RNase_MRP_RNA	GCF_002263795.1	GCF_002263805.1	Primary Assembly	1	1
8	gene	SRP_RNA	GCF_002263795.1	GCF_002263805.1	Primary Assembly	1	1
9	gene	V_segment	GCF_002263795.1	GCF_002263805.1	Primary Assembly	159	159
10	gene	antisense_RNA	GCF_002263795.1	GCF_002263805.1	Primary Assembly	1	1
11	gene	guide_RNA	GCF_002263795.1	GCF_002263805.1	Primary Assembly	28	28
12	gene	lncRNA	GCF_002263795.1	GCF_002263805.1	Primary Assembly	5179	5179
13	gene	miRNA	GCF_002263795.1	GCF_002263805.1	Primary Assembly	797	797
14	gene	misc_RNA	GCF_002263795.1	GCF_002263805.1	Primary Assembly	82	82
15	gene	other	GCF_002263795.1	GCF_002263805.1	Primary Assembly	1	7
16	gene	protein_coding	GCF_002263795.1	GCF_002263805.1	Primary Assembly	21026	21035

Feature Table

	A	B	C	D	E	F	G	H	I	J	K
1	# feature	class	assembly	assembly_unit	seq_type	chromosome	genomic_accession	start	end	strand	product_acc
2	gene	pseudogene	GCF_002263795.1	Primary Assembly	chromosome	1	NC_037328.1	207958	209108	-	
3	gene	lncRNA	GCF_002263795.1	Primary Assembly	chromosome	1	NC_037328.1	210759	214966	-	
4	ncRNA	lncRNA	GCF_002263795.1	Primary Assembly	chromosome	1	NC_037328.1	210759	214966	-	XR_003035
5	gene	lncRNA	GCF_002263795.1	Primary Assembly	chromosome	1	NC_037328.1	217517	257046	-	
6	ncRNA	lncRNA	GCF_002263795.1	Primary Assembly	chromosome	1	NC_037328.1	217517	257046	-	XR_003035
7	ncRNA	lncRNA	GCF_002263795.1	Primary Assembly	chromosome	1	NC_037328.1	217517	257046	-	XR_003035
8	ncRNA	lncRNA	GCF_002263795.1	Primary Assembly	chromosome	1	NC_037328.1	217517	257046	-	XR_003035
9	ncRNA	lncRNA	GCF_002263795.1	Primary Assembly	chromosome	1	NC_037328.1	217517	257046	-	XR_003035
10	gene	protein_coding	GCF_002263795.1	Primary Assembly	chromosome	1	NC_037328.1	260033	269063	-	
11	mRNA		GCF_002263795.1	Primary Assembly	chromosome	1	NC_037328.1	260033	269063	-	XM_024993
12	CDS	with_protein	GCF_002263795.1	Primary Assembly	chromosome	1	NC_037328.1	260033	268757	-	XP_024849
13	gene	lncRNA	GCF_002263795.1	Primary Assembly	chromosome	1	NC_037328.1	[illegible]	[illegible]		

3. FTP – Assembly Summary Files

ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS

```
     # assembly_accession [  1]: GCF_000004155.1
               bioproject [  2]: PRJNA264109
                biosample [  3]: SAMN02953762
               wgs_master [  4]: ACQJ00000000.2
          refseq_category [  5]: representative genome
                    taxid [  6]: 653667
             species_taxid [  7]: 866546
             organism_name [  8]: Schizosaccharomyces cryophilus OY26
       infraspecific_name [  9]: strain=OY26
                  isolate [ 10]:
            version_status [ 11]: latest
            assembly_level [ 12]: Scaffold
              release_type [ 13]: Major
                genome_rep [ 14]: Full
              seq_rel_date [ 15]: 2013/07/31
                  asm_name [ 16]: SCY4
                 submitter [ 17]: Broad Institute
          gbrs_paired_asm [ 18]: GCA_000004155.2
          paired_asm_comp [ 19]: identical
                 ftp_path [ 20]: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/004/155/GCF_000004155.1_SCY4
     excluded_from_refseq [ 21]:
relation_to_type_material [ 22]: assembly from type material
```

3. FTP – Assembly Summary Files

ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS

```
# assembly_accession [  1]: GCF_000001215.4
      refseq_category [  5]: reference genome
        species_taxid [  7]: 7227
        organism_name [  8]: Drosophila melanogaster
       version_status [ 11]: latest
           genome_rep [ 14]: Full
         seq_rel_date [ 15]: 2014/08/01
             asm_name [ 16]: Release 6 plus ISO1 MT
     gbrs_paired_asm [ 18]: GCA_000001215.4
     paired_asm_comp [ 19]: identical
             ftp_path [ 20]: ftp://ftp.ncbi.nlm.nih.gov…
```

3. FTP – Assembly Summary Files

`ftp.ncbi.nlm.nih.gov/genomes/refseq`

```
├── archaea
├── bacteria
├── fungi
├── invertebrate
├── plant
├── protozoa
├── vertebrate_mammalian
├── vertebrate_other
└── viral
```

Example: Download all *E. coli* genomes

```
## download the assembly_summary.txt file
wget ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/assembly_summary.txt

## parse assembly_summary.txt to create a url list
cat assembly_summary.txt \
  | awk 'BEGIN{FS="\t";OFS="\t"} \
    ($7 == "562" && $5 == "reference genome") \
    {print $20}' \
  | sed -r 's/(GC[AF]_[0-9.]*_.*$)/\1\/\1_genomic.fna.gz/g' \
  > url_list.txt

## download data
wget -i url_list.txt
```

Summary

- EDirect
  - Access NCBI databases from UNIX command line
  - Extract data in tabular format
- Use FTP for whole genome data
  - Multiple formats
  - Batch downloads

Thank you!

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

RefSeq/Gene

Terence Murphy

Eric Cox
Catherine Farrell
Tamara Goldfarb
Diana Haddad
John Jackson
Vinita Joardar
Kelly McGarvey
Michael Murphy
Nuala O'Leary
Shashi Pujar
Bhanu Rajput
Sanjida Rangwala
Lillian Riddick
Barbara Robbertse
Brian Smith-White
Pooja Strope
Anjana Vatsan

David Webb
Alex Astashyn
Olga Ermolaeva
Craig Wallin

Annotation Pipeline

Francoise Thibaud-Nissen
Paul Kitts
Mike Dicuccio
Wratko Hlavina
Avi Kimchi
Jinna Choi
Boris Kiryutin
Patrick Masterson
Eyal Mozes
Anton Perkov
Robert Smith
Alexandre Souvorov

E-utilities

Colleen Bollin
Jonathan Kans

Leadership

Kim Pruitt
Jim Ostell

ncbi.nlm.nih.gov/home/coursesandwebinars

bit.ly/entrez-direct

bit.ly/ncbi_factsheets

support.nlm.nih.gov

$ esearch -help

NCBI booth #223