The Assembly Database

Tracking and accessing versions of genomic assemblies available for different organisms

https://www.ncbi.nlm.nih.gov/assembly/

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Scope

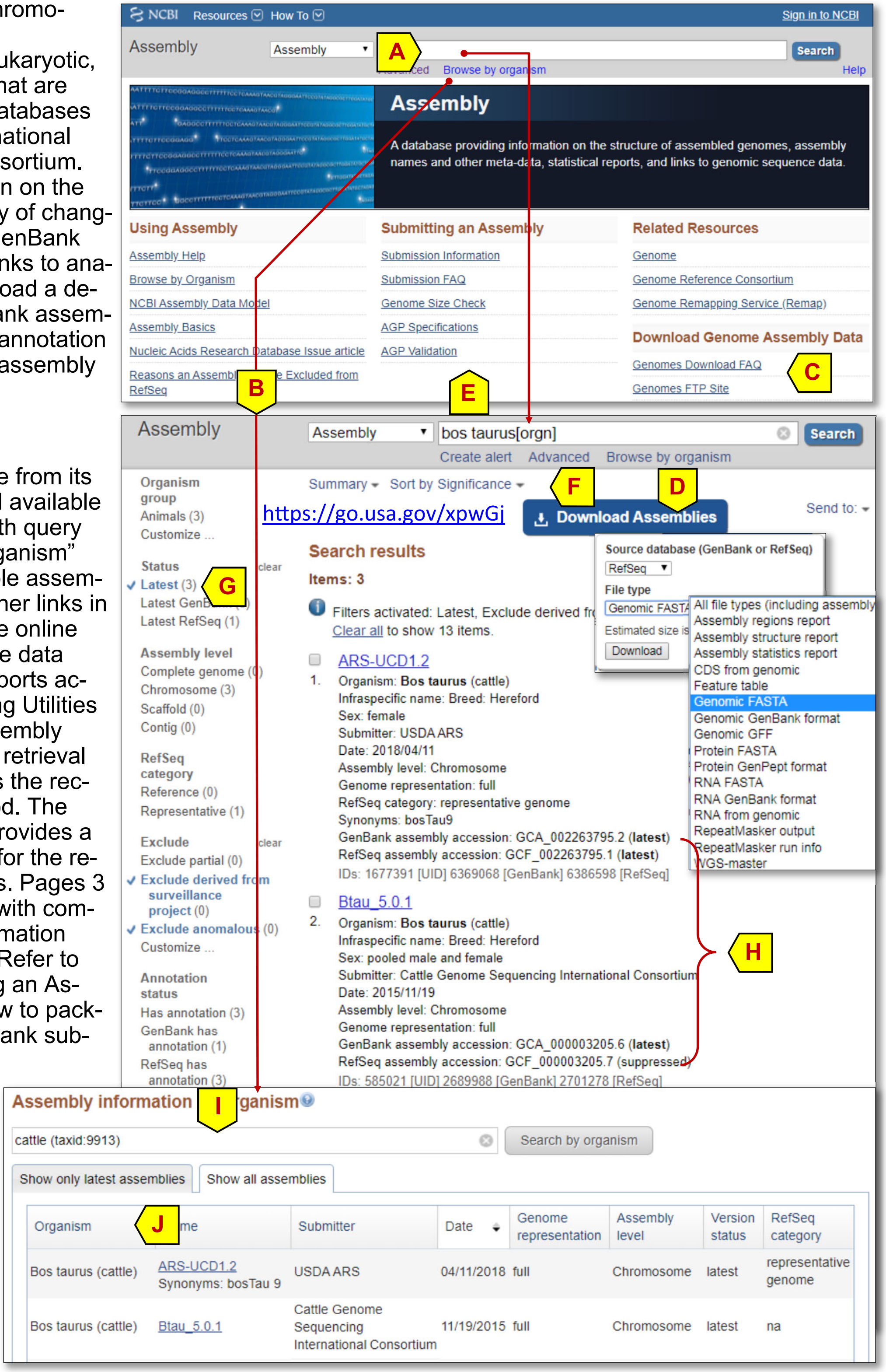
Advances in sequencing technology have led to genome assemblies being available for an increasing number of organisms. NCBI's Assembly resource tracks the set of sequences that comprise a genome assembly and the structure of the assembly (e.g., contigs, scaffolds, chromosomes, gaps). A stable assembly "accession.version" is provided for eukaryotic, bacterial, and archaeal assemblies that are submitted to GenBank and related databases maintained by members of the International Nucleotide Sequence Database Consortium. Assembly records provide information on the assembly structure, submitter, history of changes, statistics, relationship between GenBank and RefSeq genomes, and include links to analyze the assembly via BLAST, download a detailed report, or download the GenBank assembly sequence and annotation (when annotation has been submitted), or the RefSeq assembly sequence and annotation data.

Data Access

The assembly database is accessible from its homepage (right). Here, you can find available genome assemblies by searching with query terms (**A**), or use the "Browse by Organism" page (**B**) to browse the list of available assemblies and filter them by organism. Other links in the "Using Assembly" column provide online help and additional information on the data model. Assembly database also supports access through the Entrez Programming Utilities API, when the database is set to assembly (db=assembly). However, full record retrieval through efetch is pending. FTP (**C**) is the recommended bulk data retrieval method. The "Download Assemblies" button (**D**) provides a convenient venue to download data for the retrieved assemblies in various formats. Pages 3 and 4 of this present two use cases with command line examples. Assembly information comes from GenBank submissions. Refer to documentation under the "Submitting an Assembly" (**E**) column for details on how to package assembly information to a GenBank submission.

Finding Assemblies

The system displays assemblies retrieved with your query terms in the summary format (**F**). You can see historic assembly versions by unchecking "Latest" filter (**G**). The assembly accession.version (**H**) is the key for downloading sequences, assembly summary, and other data from the FTP site. You can use "Browse by Organism" to see what genome assemblies are available for an organism, with the results presented in an easy to read tabular format (**I**). Click the name of the assembly (**J**) in either format of the search result retrieve the detailed report page.



Assembly | Assembly ▾ | Search

Advanced   Browse by organism   Help

Assembly

A database providing information on the structure of assembled genomes, assembly names and other meta-data, statistical reports, and links to genomic sequence data.

Using Assembly
- Assembly Help
- Browse by Organism
- NCBI Assembly Data Model
- Assembly Basics
- Nucleic Acids Research Database Issue article
- Reasons an Assembly may be Excluded from RefSeq

Submitting an Assembly
- Submission Information
- Submission FAQ
- Genome Size Check
- AGP Specifications
- AGP Validation

Related Resources
- Genome
- Genome Reference Consortium
- Genome Remapping Service (Remap)

Download Genome Assembly Data
- Genomes Download FAQ
- Genomes FTP Site

Assembly | Assembly ▾ | bos taurus[orgn] | Search

Create alert   Advanced   Browse by organism

Summary ▾   Sort by Significance ▾

https://go.usa.gov/xpwGj

Download Assemblies   Send to: ▾

Organism group
- Animals (3)
- Customize ...

Status (clear)
- ✓ Latest (3)
- Latest GenBank (3)
- Latest RefSeq (1)

Assembly level
- Complete genome (0)
- Chromosome (3)
- Scaffold (0)
- Contig (0)

RefSeq category
- Reference (0)
- Representative (1)

Exclude (clear)
- Exclude partial (0)
- ✓ Exclude derived from surveillance project (0)
- ✓ Exclude anomalous (0)
- Customize ...

Annotation status
- Has annotation (3)
- GenBank has annotation (1)
- RefSeq has annotation (3)

Search results

Items: 3

Filters activated: Latest, Exclude derived fr[om surveillance project, Exclude anomalous.] Clear all to show 13 items.

Source database (GenBank or RefSeq): RefSeq ▾

File type: Genomic FASTA

Estimated size is

Download

- All file types (including assembly
- Assembly regions report
- Assembly structure report
- Assembly statistics report
- CDS from genomic
- 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
- WGS-master

□ 1. ARS-UCD1.2
Organism: **Bos taurus** (cattle)
Infraspecific name: Breed: Hereford
Sex: female
Submitter: USDA ARS
Date: 2018/04/11
Assembly level: Chromosome
Genome representation: full
RefSeq category: representative genome
Synonyms: bosTau9
GenBank assembly accession: GCA_002263795.2 (**latest**)
RefSeq assembly accession: GCF_002263795.1 (**latest**)
IDs: 1677391 [UID] 6369068 [GenBank] 6386598 [RefSeq]

□ 2. Btau_5.0.1
Organism: **Bos taurus** (cattle)
Infraspecific name: Breed: Hereford
Sex: pooled male and female
Submitter: Cattle Genome Sequencing International Consortium
Date: 2015/11/19
Assembly level: Chromosome
Genome representation: full
GenBank assembly accession: GCA_000003205.6 (**latest**)
RefSeq assembly accession: GCF_000003205.7 (suppressed)
IDs: 585021 [UID] 2689988 [GenBank] 2701278 [RefSeq]

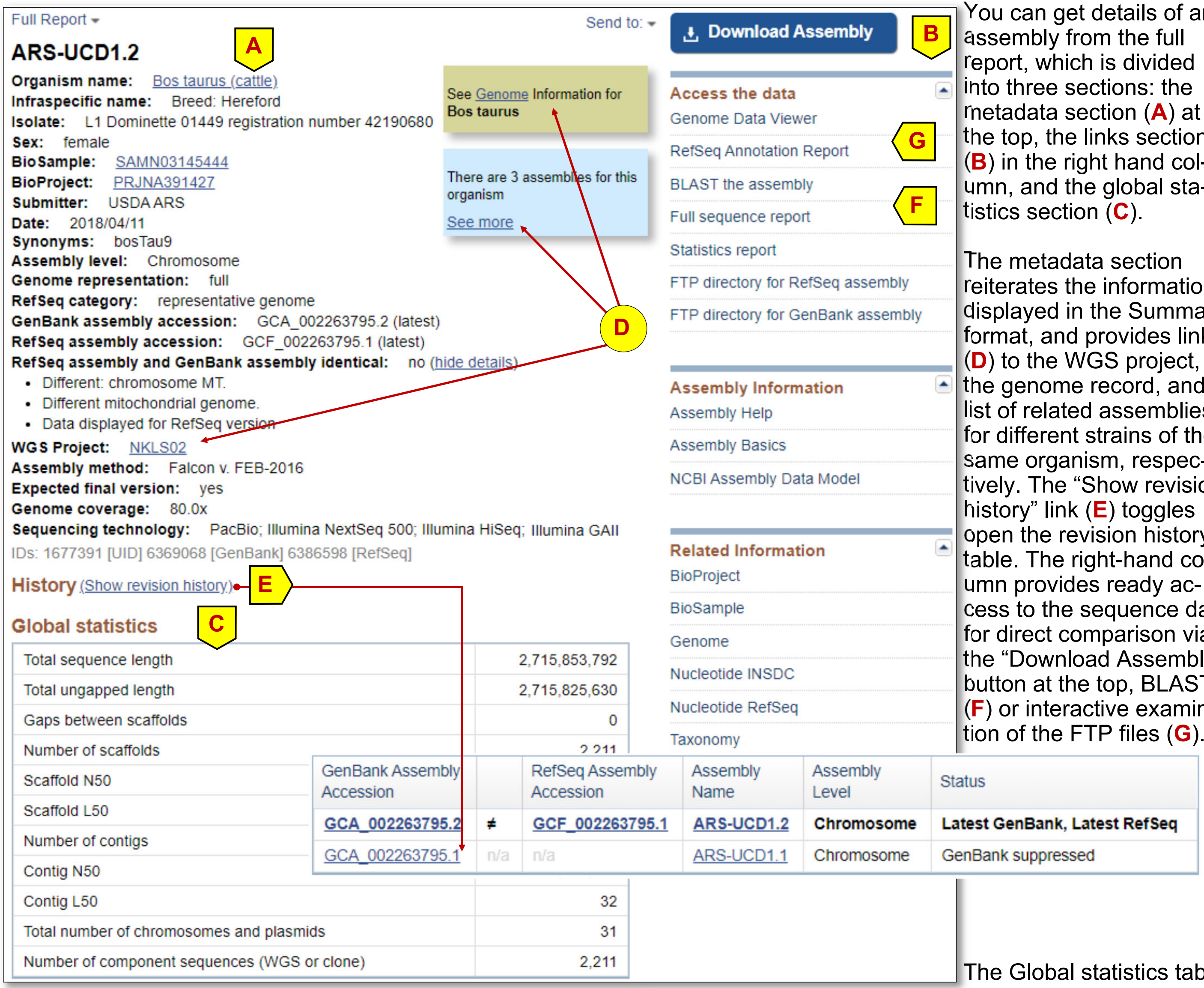
Assembly information for organism

cattle (taxid:9913) | Search by organism

Show only latest assemblies | Show all assemblies

Organism	Name	Submitter	Date	Genome representation	Assembly level	Version status	RefSeq category
Bos taurus (cattle)	ARS-UCD1.2 Synonyms: bosTau 9	USDA ARS	04/11/2018	full	Chromosome	latest	representative genome
Bos taurus (cattle)	Btau_5.0.1	Cattle Genome Sequencing International Consortium	11/19/2015	full	Chromosome	latest	na

The full report of an Assembly record



ARS-UCD1.2

Organism name: Bos taurus (cattle)
Infraspecific name: Breed: Hereford
Isolate: L1 Dominette 01449 registration number 42190680
Sex: female
BioSample: SAMN03145444
BioProject: PRJNA391427
Submitter: USDA ARS
Date: 2018/04/11
Synonyms: bosTau9
Assembly level: Chromosome
Genome representation: full
RefSeq category: representative genome
GenBank assembly accession: GCA_002263795.2 (latest)
RefSeq assembly accession: GCF_002263795.1 (latest)
RefSeq assembly and GenBank assembly identical: no (hide details)

- Different: chromosome MT.
- Different mitochondrial genome.
- Data displayed for RefSeq version

WGS Project: NKLS02
Assembly method: Falcon v. FEB-2016
Expected final version: yes
Genome coverage: 80.0x
Sequencing technology: PacBio; Illumina NextSeq 500; Illumina HiSeq; Illumina GAII
IDs: 1677391 [UID] 6369068 [GenBank] 6386598 [RefSeq]

See Genome Information for **Bos taurus**

There are 3 assemblies for this organism
See more

History (Show revision history)

GenBank Assembly Accession		RefSeq Assembly Accession	Assembly Name	Assembly Level	Status
GCA_002263795.2	**≠**	**GCF_002263795.1**	**ARS-UCD1.2**	**Chromosome**	**Latest GenBank, Latest RefSeq**
GCA_002263795.1	n/a	n/a	ARS-UCD1.1	Chromosome	GenBank suppressed

Global statistics

Total sequence length	2,715,853,792
Total ungapped length	2,715,825,630
Gaps between scaffolds	0
Number of scaffolds	2,211
Scaffold N50	
Scaffold L50	
Number of contigs	
Contig N50	
Contig L50	32
Total number of chromosomes and plasmids	31
Number of component sequences (WGS or clone)	2,211

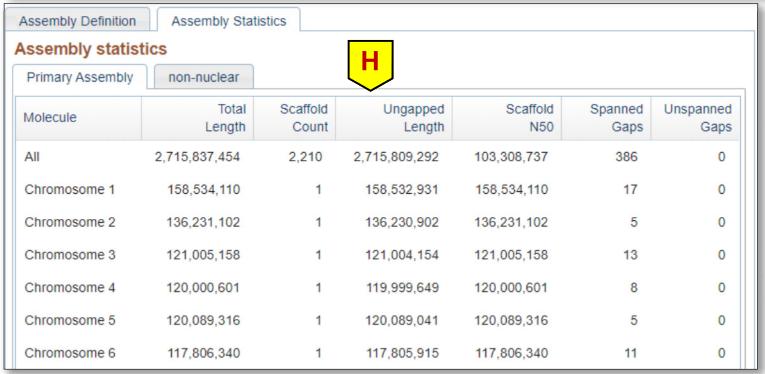
Access the data
Genome Data Viewer
RefSeq Annotation Report
BLAST the assembly
Full sequence report
Statistics report
FTP directory for RefSeq assembly
FTP directory for GenBank assembly

Assembly Information
Assembly Help
Assembly Basics
NCBI Assembly Data Model

Related Information
BioProject
BioSample
Genome
Nucleotide INSDC
Nucleotide RefSeq
Taxonomy

You can get details of an assembly from the full report, which is divided into three sections: the metadata section (**A**) at the top, the links section (**B**) in the right hand column, and the global statistics section (**C**).

The metadata section reiterates the information displayed in the Summary format, and provides links (**D**) to the WGS project, the genome record, and list of related assemblies for different strains of the same organism, respectively. The "Show revision history" link (**E**) toggles open the revision history table. The right-hand column provides ready access to the sequence data for direct comparison via the "Download Assembly" button at the top, BLAST (**F**) or interactive examination of the FTP files (**G**).



Assembly statistics

Primary Assembly | non-nuclear

Molecule	Total Length	Scaffold Count	Ungapped Length	Scaffold N50	Spanned Gaps	Unspanned Gaps
All	2,715,837,454	2,210	2,715,809,292	103,308,737	386	0
Chromosome 1	158,534,110	1	158,532,931	158,534,110	17	0
Chromosome 2	136,231,102	1	136,230,902	136,231,102	5	0
Chromosome 3	121,005,158	1	121,004,154	121,005,158	13	0
Chromosome 4	120,000,601	1	119,999,649	120,000,601	8	0
Chromosome 5	120,089,316	1	120,089,041	120,089,316	5	0
Chromosome 6	117,806,340	1	117,805,915	117,806,340	11	0

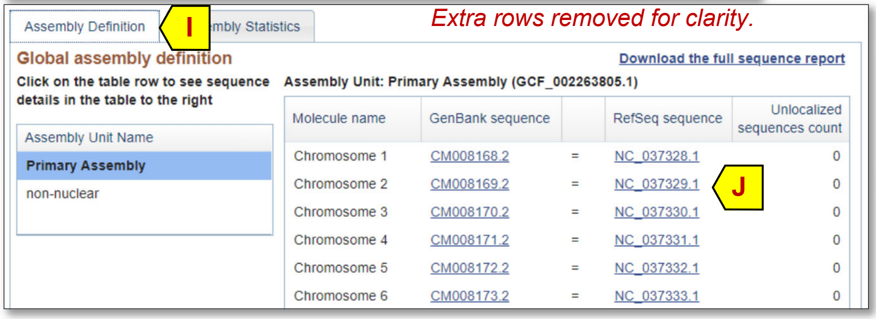
The Global statistics table (**C**) is an aggregated report for the assembly as a whole. The table under" Assembly statistics" (**H**) provides chromosome by chromosome statistics. Individual chromosomal sequences and their feature annotations reside in the Nucleotide database. The "Assembly Definition" tab (**I**) provides a tabular list of these accessions. This table cross references the GenBank sequence accession with its RefSeq counterpart, which always contains annotated features. Sequence accessions (**J**) link to corresponding records in the Nucleotide database to allow detailed interactive examination through the Graphical Sequence Viewer (SV) when displayed in the graphical format. An example like below is to the graphical display of RefSeq Chromosome X for *Bos Taurus*:

- https://www.ncbi.nlm.nih.gov/nuccore/NC_037351.1?report=graph

Refer to the SV handout for additional information on how to use this graphical sequence display tool:

- https://ftp.ncbi.nih.gov/pub/factsheets/Factsheet_Graphical_SV.pdf

Extra rows removed for clarity.



Global assembly definition

Click on the table row to see sequence details in the table to the right

Assembly Unit Name
Primary Assembly
non-nuclear

Download the full sequence report

Assembly Unit: Primary Assembly (GCF_002263805.1)

Molecule name	GenBank sequence		RefSeq sequence	Unlocalized sequences count
Chromosome 1	CM008168.2	=	NC_037328.1	0
Chromosome 2	CM008169.2	=	NC_037329.1	0
Chromosome 3	CM008170.2	=	NC_037330.1	0
Chromosome 4	CM008171.2	=	NC_037331.1	0
Chromosome 5	CM008172.2	=	NC_037332.1	0
Chromosome 6	CM008173.2	=	NC_037333.1	0

Use Cases: Batch Retrieval of Assembly Sequence Data from the FTP Site

Submitted assemblies selected for NCBI Reference Sequence project (RefSeq) are annotated by NCBI's genome annotation pipeline and made available through the NCBI FTP site (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/). The ***/genomes/refseq*** and ***/genomes/genbank*** directories organize available data by large taxonomic groups, i.e., archaea, bacteria, fungi, invertebrate, plant, protozoa, vertebrate_mammalian, vertebrate_other, and viral (last is RefSeq only, details at: https://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/). Each group level directory contains an assembly_summary.txt file with details on the latest versions of assemblies available for that group. This file also contains many fields of metadata, useful in identifying genome assemblies of interest, as well as the URLs for the subdirectories from which the data files can be downloaded. For a detailed description of the file structure, see https://ftp.ncbi.nlm.nih.gov/genomes/README_assembly_summary.txt.

NCBI organizes the genomes data files with a consistent directory hierarchy. For example, the RefSeq entries have the following naming convention (GenBank entries have GCA instead of GCF initial): /all/GCF/aaa/bbb/ccc/GCF_aaabbbccc.V_NAME/GCF_aaabbbccc.V_NAME_X_Y.gz, where GCF_aaabbbccc.V is the assembly's accession plus version, NAME is the assembly name, and _X_Y are sequence and file type. Workflows below use Linux shell utilities to process the assembly_summary.txt for a selected taxonomic group into FTP URLs for genomic sequences or full subdirectory content download.

This directory structure is not well suited for downloading data for all assemblies from a broad taxonomic group. Here, we describe a Linux shell command-based workflow that takes advantage of the assembly_summary.txt file for representative taxonomic groups, to extract URLs for files or directories of interest, and use them to download selected sequences or all data files.

Case 1: Get all the genomic sequence files for the fungal RefSeq assemblies

Under the ***/genomes/refseq*** directory of the NCBI FTP site, available data are grouped by large taxonomic groups, i.e., archaea, bacteria, fungi, invertebrate, plant, protozoa, vertebrate_mammalian, vertebrate_other, and viral, each with its own assembly_summary.txt file that provides detailed information of available assemblies along with the URLs for those subdirectories in the 20th column. The workflow consists of two steps, collecting and modifying the FTP URLs for the desired file format (genomic FASTA sequences), and downloading the relevant files using the collected URLs as input.

Step 1. Collect and modify the FTP URLs to point to the _genomic.fna.gz files

The command line is a pipe symbol– linked set:

The "\" is a Linux shell command to indicate that the command line continues in the next line. We use it to break the linked commands into distinctive steps so we can clearly see and discuss each sub-step:

- The first **curl** command simply gets the specified assembly_summary.txt file and passes its content to the next step with a pipe ("**|**", instead of displaying it in console).
- The second command uses the **awk** utility to separate each line's content by tab (FS="\t"), skip header line (!/^#/) and print out the value of the 20th column (print $20), and passes the output to the next step with pipe ("|").
- The third command uses **sed** to modify the extracted URL string that points to an assembly directory to point to the "_genomic.fna.gz" file instead. Specifically, with the pipe ("|") as delimiter, it first matches the URL into substrings using regular expression patterns and captures them using parentheses (**s|(ftp://ftp.ncbi.nlm.nih.gov/genomes/all/.+/)(GCF_.+)|**), then reconstructs the string (**\1\2**) for the path and add another directory level (**/**) and a specific file name (**\2_genomic.fna.gz**). This modifies the existing URL to point to the **_genomic.fna.gz** file for that assembly. The last part (**>genomic_file**) redirects the output into a file named genomic_file (partially shown below).

```
curl 'ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/fungi/assembly_summary.txt' | \

awk '{FS="\t"}  !/^#/ {print $20} ' | \

sed -r 's|(ftp://ftp.ncbi.nlm.nih.gov/genomes/all/.+/)(GCF_.+)|\1\2/\2_genomic.fna.gz|' > genomic_file
```

Use Cases (cont.)

Step 2. Use step 1's output as input to download through wget utility

The second step is very simple. The command (below left) calls the wget utility and passes the output from step 1 as an argument to the "-- input-file" switch. Given a file with a list of full FTP URLs (**A**), running the wget command with the input file (**B**) will retrieve the files (**C**).

A

```
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/002/945/GCF_000002945.1_ASM294v2/
GCF_000002945.1_ASM294v2_genomic.fna.gz

ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/149/845/GCF_000149845.2_SJ5/
GCF_000149845.2_SJ5_genomic.fna.gz
```

B

```
wget --input genomic_file
```

C

```
-rw-r--r-- 1 samd sdesk  3989616 Dec 31 10:52
GCF_000002945.1_ASM294v2_genomic.fna.gz

-rw-r--r-- 1 samd sdesk  3492573 Dec 31 10:52 GCF_000149845.2_SJ5_genomic.fna.gz
```

The command gunzip *.gz will unpack them all to regenerate the FASTA files for further downstream need. For better file management, first move *.gz files to a new directory so they are isolated from other files.

Case 2: Get the directories and their contents for all the fungal RefSeq assemblies

To completely mirror and archive the files for this group or organisms, you can modify the above commands to download the directories along with all their contents (excluding subdirectories).

Step 1. Collect and modify the FTP URLs to get only the directory name

We can modify the command from Case 1 by dropping the sed command, and modify the **awk** slightly to get the directory URLs (**D**). Last two lines are example output.

D

```
curl 'ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/fungi/assembly_summary.txt' | \
awk '{FS="\t"} !/^#/ {print $20"/"}'  > genomic_directory
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/149/845/GCF_000149845.2_SJ5/
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/150/505/GCF_000150505.1_SO6/
```

Step 2. Pass the above output to wget to pull down the content

We use the wget command (**E**) to get all the directories and their files. The command is more complex than in Case 1, so we explain the meaning of each command line arguments separately below the command. Make sure you have enough disk space since this will pull all the directories and their contents down to your Linux box. Users on *PCs (without Linux or Cygwin)* can do Step 1. above in a different way using the example inline Perl commands below. It generates the same outputs as on Linux for use with the same wget command for the second steps for content download. The PC port of wget is available from: https://eternallybored.org/misc/wget/

E

```
wget -r --no-parent --no-host-directories --cut-dirs=2 --level=1 \
--input-file=genomic_directory
```

-r: recursively works through the directory

--no-parent: ignores the parent directory

--no-host-directories: saves the files without prepending the NCBI FTP URL

--cut-dirs=2: saves the files without creating intermediate directories

--level=1: works only at that level of directory

--input-file=value: sets directory input to the file specified by value

```
perl -e "use LWP::Simple; $file=get(\"https://ftp.ncbi.nlm.nih.gov/genomes/refseq/fungi/assembly_summary.txt\");
while ($file =~ /(ftp.+)(GCF_.+?)\s/g){print $1, $2, \"\/\", $2, \"_genomic.fna.gz\n\";}"  > fungi_genomic_files

perl -e "use LWP::Simple; $file=get(\"https://ftp.ncbi.nlm.nih.gov/genomes/refseq/fungi/assembly_summary.txt\");
while ($file =~ /(ftp:.+GCF.+?)\s/g){print $1, \"\/\n\";}"  > fungi_directory
```