

Bulk Downloads

Class Summary:

- Why?
- Data Formats
- Data Sources
- Exercises

What Are Bulk Downloads?

Sometimes you will need to gather many records, and rather than collecting everything manually, one piece at a time, you can save a great deal of time by downloading a large set of data and sort out what you need.

We will cover the manual 'Bulk Download' process only. Almost all of the databases, genome browsers, and other tools we have seen so far will offer access to their raw data in some format. This access is usually provided via 2 routes: manual downloads, and programmatic downloads.

The programmatic downloads require a little bit of programming experience. We will cover programmatic access to bulk data during some of the upcoming courses.

Data Formats:

As Alex outlined on the first day, bioinformatics data comes in many formats.

Flat File:

- Genbank and GenPept (*gb, *gp)
- FASTA and QUAL (*fasta, *fa, *fna, *qu, *qual)
- Alignment
- Multiple Alignment
- Tab-Delimited (*tsv, *txt)
- Comma Separated (*csv)

NOTE: csv and txt file data can be formatted however the authors wish. Always check the file format specifications before working with these files.

Hierarchical:

- ASN.1
- XML
- HTML

Binary:

- Sequence Chromatogram or Trace Files (*ab1 or *scf)

Depending on the data source, you may end up with any number of these data formats, or new specialized data formats.

Always check to know what format you are dealing with- Especially with text files.

Data Sources:

NCBI:

<http://eutils.ncbi.nlm.nih.gov/Ftp/>

NOTE: for programmatic access, NCBI offers the utilities software:

http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

-and-

“Building Customized Data Pipelines Using the Entrez Programming Utilities”

<http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=coursework.chapter.eutils>

Download some NCBI data:

- Open this link: <http://eutils.ncbi.nlm.nih.gov/Ftp/>
- Click Gene
- Open the README
- Read about the data files and formats in this directory tree.
- Open the DATA link
- Download and save the files (save the files in a new dir named ‘class4Data’):
 - `gene2go.gz`
 - `gene_info.gz`
- Open a terminal session (Black square in dock)
- Move to the directory you saved your data (ex: `cd ./Desktop/class4Data`)
- Unzip the files (`gunzip ./ gene2go.gz`)
- View the contents (more `./gene2go`) Hit q to exit.
- Open the files in excel.
- Notice that you get a warning “File Not Completely Loaded”. This is telling you the file is > 50,000 lines. Excel will not display more than 50,000 records. For most cases, you’ll want to use this bulk data with software w/out these limitations. This file is so large b/c it contains gene info for all NCBI organisms. We’ll now use a quick UNIX trick to get just the human records.
- On the terminal command line, type:
 - `grep ‘^9606’ gene_info > gene_info_human`
 - `grep ‘^9606’ gene2go > gene2go_human`
- This will grab all lines from the file which start with 9606 (the code for human), and write them to the file `gene_info_human`
- Open `gene_info_human` in excel.
- Search for EGFR.

- Open the README file and find the entry describing the format of the gene_info file.
- Note that the second column is the ENTREZgeneID. This id is used to key in to other tables which annotate genes.
- Note the synonyms for EGFR.
- Open the gene2go_human file in excel (its ok that the file gives an error on opening.... The records we want are at the top)
- Find the EGFR entries (entrezID = 1956)
- Find the description of the gene2go file in the README
- So, if you wanted to get all of the go terms for a set of genes, you could quickly use these tables to get what you need.

ENSEMBL:

<http://www.ensembl.org/info/data/download.html>

- Open the link :
<http://www.ensembl.org/info/data/download.html>
- Scroll down to the H.sapiens entry.
- There are several links to the right of each species entry. These allow you to download sequence for various genomic features, and grab annotation files in Genbank and EMBL format.
- The annotation files are not easy to work with manually, so we will leave them for the programming class.

BIOMART (programmatic interface) : <http://www.biomart.org/>
Open this link. Run through a quick search.

UCSC:

<http://genome.ucsc.edu/cgi-bin/hgTables?org=Human&db=hg18>

Programmatic interface :: nothing provided.. but you can recreate their table structure and use SQL to run any queries.

Get Gene Records in a Region:

- Open the genome browser.
- Change the coordinates to: chr21:26,074,733–26,965,003
- Click tables
- Note the region we were viewing is now preset in the position field.
- For group, select the ‘Gene and Prediction Tracks’.
- For track, select ‘known genes’
- For output format select ‘selected fields from primary and related tables’
- Select all boxes in the first section, in the second section select the ucsc_kgxfref box.

- Select ‘Allow Selection From Checked Tables’ at bottom.
- Hit OK
- Select the gene symbol, mrna, protid, and description fields in ucsc_kgxfref
- -get output-

Repeat the first UCSC exercise.... But add a filter

Get SNP records in a region

- Run through steps 1-4 above.
- For group, select “Variation and Repeats”
- For track select “SNPs”
- For output format select ‘All Fields... “
- Change file name
- -get output-
- Unzip the file.
- Open in excel

COSMIC:

<ftp://ftp.sanger.ac.uk/pub/CGP/cosmic>

- Open the link to the cosmic data.
- Read the GUIDELINES* file
- Follow the ‘data_export’ link
- Open EGFR.csv entry

Tip: To save a linked file, hold the –control- button and click on the link, then choose ‘save link as’.

iHOP:

<http://www.ihop-net.org/UniPub/iHOP/>

From the help page:

Bulk downloads.

Please note that iHOP is a freely available tool from the academic domain. Not a company! Thus, it is necessary to limit server load and to give preference to individual users. Bulk downloads may lead to the banning of IP addresses for specific servers or institutions!

Please contact directly with Robert Hoffmann if there should be a scientific reason for bulk downloads. Thank you for your cooperation!

Others: If you are working with a tool or database, and do not see a way to bulk access the supporting data, contact the authors.

Exercises:

Ex1: UCSC (Biomart does not have the same depth of info for cytoband)::

For all of Chromosome Y in human, grab all of the cytoband/ideogram entries.
(hint- Mapping and Sequencing* group)

Ex 2: UCSC or Biomart::

For all genes on chrm X in human, grab the name, chrm, strand, start, end, HGNCgeneSymbol, geneDescription, and EnsemblID.
(hint adding in one more table to the report from the 2nd UCSC walkthrough above)

Ex 3: Biomart would be easiest ::

Retrieve the HGNC gene symbols, chrm, start&end positions, and affy U95 expression annotations for all genes on Chromosome 1.