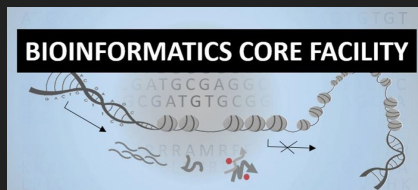


Welcome to ABC.6

3. October 2024

<https://abc.au.dk>



Health
Data Science
Sandbox

Agenda



- What's new
- Topic presentation
- Tutorial and/or open coding

What's new

DataViz material

A curated list of learning resources and examples for data visualization from AU

<https://visualization.info/>

 Collected and Curated by Hans-Jörg Schulz 

Data Visualization Teaching and Learning Materials

Start typing or click labels to search everything

 All Categories **86**

 Textbooks and Lecture Notes **30**

 Lecture Videos and Slides **29**

 Tutorials and Notebooks **17**

 Exercises and Other Materials **10**

TEXTBOOKS AND LECTURE NOTES 2024



Rhetorical DataVis Course (Pilot)

Enrico Bertini

The course is, above all, about thinking effectively with data visualization. When exposed to new data visualizations, many elements play a role in deciding


TUTORIALS AND NOTEBOOKS 2024



Information Visualization Tutorials

FH Potsdam

EXERCISES AND OTHER MATERIALS 2024



GIS&T Body of Knowledge: Cartography and Visualization

UCGIS

The Cartography & Visualization section encapsulates competencies related to the design and use of maps and mapping technology. This section covers

What's new

GenomeDK basics Workshop

- 10.October
- 11-12 am + 13-15 pm in 1540-K26
- Please get a functioning account (with 2FA already secured) at <https://console.genome.au.dk/user-requests/create/>
- We will introduce the very basics of GDK usage and bash
- Of course come with your questions and problems
- Outlook invitation or sign up from our website <https://abc.au.dk>

Tutorial topic

Online databases

- Lots of databases, bigger ones (GEO+SRA) and smaller ones (project-specific), databases of databases, open source data/data on request
- Different omics. <https://www.omicsdi.org/database> is a good resource for main omics databases
- Many features
 - Data download
 - programmatic download
 - specific download tools
 - different stages of data-format-chaos
 - Query for finding items
 - Exploratory/interactive utilities

Tutorial topic

Online databases



Tutorial topic

First: look for an ftp



Project Files

Name	Type	Size (M)	Download
checksum.txt	OTHER	1707 bit	FTP
Ga13HJH_bjhb1_9.raw	RAW	230	FTP
Ga13HJH_bjhb1_8.raw	RAW	229	FTP
Ga13HJH_bjhb1_7.raw	RAW	243	FTP
Ga13HJH_bjhb1_6.raw	RAW	273	FTP
Ga13HJH_bjhb1_5.raw	RAW	264	FTP
Ga13HJH_bjhb1_4.raw	RAW	270	FTP

Total 17 items [<](#) [1](#) [>](#) [20 /page](#)

Tutorial topic

First: look for an ftp

FTP (file transfer protocol) is a folder structure like in your pc, but hosted remotely and explorable...

...through your browser...

...or your command line!



The screenshot shows a web browser window with the address bar displaying `https://ftp.pride.ebi.ac.uk/pride/data/archive/2024/09/PXD056312/`. The browser's tab bar shows several tabs, including 'Papers 1', 'Learning', 'Sandbox', 'Blogging', 'Articles', 'Online-Tools', 'Software', and 'Hus'. The main content area displays the title 'Index of /pride/data/archive/2024/09/PXD056312' and a table with columns for 'Name', 'Last modified', 'Size', and 'Description'. The table lists various files, including XML, RAW, and checksum files, with their respective modification dates and sizes.

Name	Last modified	Size	Description
Parent Directory	-	-	-
Ga13HJH_bjhb1_.pep.xml	2024-09-27 12:35	27M	
Ga13HJH_bjhb1_.prot.xml	2024-09-27 12:33	14M	
Ga13HJH_bjhb1_1.raw	2024-09-27 12:33	268M	
Ga13HJH_bjhb1_2.raw	2024-09-27 12:35	255M	
Ga13HJH_bjhb1_3.raw	2024-09-27 12:35	254M	
Ga13HJH_bjhb1_4.raw	2024-09-27 12:35	270M	
Ga13HJH_bjhb1_5.raw	2024-09-27 12:34	264M	
Ga13HJH_bjhb1_6.raw	2024-09-27 12:35	273M	
Ga13HJH_bjhb1_7.raw	2024-09-27 12:33	243M	
Ga13HJH_bjhb1_8.raw	2024-09-27 12:34	229M	
Ga13HJH_bjhb1_9.raw	2024-09-27 12:33	230M	
Ga13HJH_bjhb1_10.raw	2024-09-27 12:34	242M	
Ga13HJH_bjhb1_11.raw	2024-09-27 12:34	246M	
Ga13HJH_bjhb1_12.raw	2024-09-27 12:34	241M	
Ga13HJH_bjhb1_13.raw	2024-09-27 12:33	270M	
Ga13HJH_bjhb1_14.raw	2024-09-27 12:34	257M	
checksum.txt	2024-09-27 12:33	1.7K	



The screenshot shows a terminal window with a dark background. The prompt is `(geofetch) samuele@D55749:~$`. The command entered is `curl -L ftp.pride.ebi.ac.uk/pride/data/archive/2024/09/PXD056312/Ga13HJH_bjhb1_.pep.xml`. The output shows the file being fetched and its size (14M).

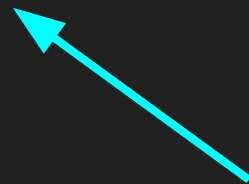
```
(geofetch) samuele@D55749:~$ curl -L ftp.pride.ebi.ac.uk/pride/data/archive/2024/09/PXD056312/Ga13HJH_bjhb1_.pep.xml
Ga13HJH_bjhb1_.prot.xml
Ga13HJH_bjhb1_1.raw
Ga13HJH_bjhb1_10.raw
Ga13HJH_bjhb1_11.raw
Ga13HJH_bjhb1_12.raw
Ga13HJH_bjhb1_13.raw
Ga13HJH_bjhb1_14.raw
Ga13HJH_bjhb1_2.raw
Ga13HJH_bjhb1_3.raw
Ga13HJH_bjhb1_4.raw
Ga13HJH_bjhb1_5.raw
Ga13HJH_bjhb1_6.raw
Ga13HJH_bjhb1_7.raw
Ga13HJH_bjhb1_8.raw
Ga13HJH_bjhb1_9.raw
checksum.txt
(geofetch) samuele@D55749:~$ curl -L ftp.pride.ebi.ac.uk/pride/data/archive/2024/09/PXD056312/Ga13HJH_bjhb1_.pep.xml
Ga13HJH_bjhb1_.prot.xml
Ga13HJH_bjhb1_1.raw
Ga13HJH_bjhb1_10.raw
Ga13HJH_bjhb1_11.raw
Ga13HJH_bjhb1_12.raw
Ga13HJH_bjhb1_13.raw
Ga13HJH_bjhb1_14.raw
Ga13HJH_bjhb1_2.raw
Ga13HJH_bjhb1_3.raw
Ga13HJH_bjhb1_4.raw
Ga13HJH_bjhb1_5.raw
Ga13HJH_bjhb1_6.raw
Ga13HJH_bjhb1_7.raw
Ga13HJH_bjhb1_8.raw
Ga13HJH_bjhb1_9.raw
checksum.txt
```


Tutorial topic

First: look for an ftp

Download is simple by saving the objects on the browser, or downloading from the command line

```
wget --random-wait \  
      -r -p -e robots=off \  
      -U mozilla \  
      ftp.pride.ebi.ac.uk/pride/data/archive/2024/09/PXD0  
      56312/
```



Downloads the whole PXD056312 folder content.

Tutorial topic

Second: look for programmatic download guidelines as in GEO

E-Util programs

FTP directory structure

All GEO data are available for download from the FTP site. Directory structure is organized by type, GEO accession range, GEO accession number, and format. Range subdirectory name is created by replacing the three last digits of the accession with letters "nnn". For example,

GSM575: /samples/GSMnnn/GSM575/
GSM1234: /samples/GSM1nnn/GSM1234/
GSM12345: /samples/GSM12nnn/GSM12345/

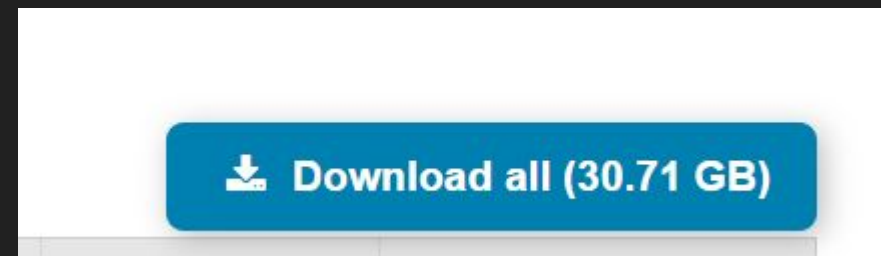
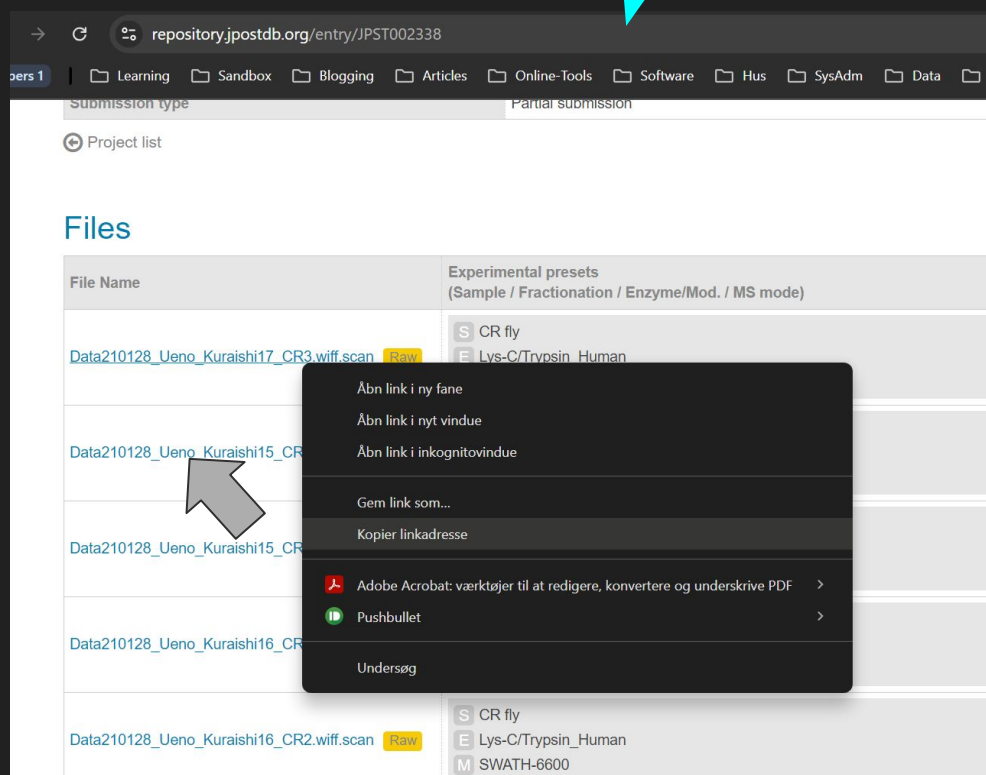
For more information, please see [README](#).

Format	Example
SOFT, by DataSet	ftp://ftp.ncbi.nlm.nih.gov/geo/datasets/GDS1nnn/GDS1001/soft/GDS1001.soft.gz
SOFT full, by DataSet	ftp://ftp.ncbi.nlm.nih.gov/geo/datasets/GDS1nnn/GDS1001/soft/GDS1001_full.soft.gz
SOFT, by Platform	ftp://ftp.ncbi.nlm.nih.gov/geo/platforms/GPLnnn/GPL10/soft/GPL10_family.soft.gz
SOFT, by Series	ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSEnnn/GSE1/soft/GSE1_family.soft.gz
MINiML, by Platform	ftp://ftp.ncbi.nlm.nih.gov/geo/platforms/GPLnnn/GPL10/miniml/GPL10_family.xml.tgz
MINiML, by Series	ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSEnnn/GSE1/miniml/GSE1_family.xml.tgz
SeriesMatrix	ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSEnnn/GSE1/matrix/GSE1_series_matrix.txt.gz
Supplementary files, by Platform	ftp://ftp.ncbi.nlm.nih.gov/geo/platforms/GPL1nnn/GPL1073/suppl/
Supplementary files, by Series	ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE1nnn/GSE1000/suppl/GSE1000_RAW.tar
Supplementary files, by Sample	ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM1nnn/GSM1137/suppl/GSM1137.CEL.gz

Tutorial topic

Third: brute-force right-click+download-link

Right-click on a **download link to a file**, then copy the link address. See if you can do the same on a **bulk download button** to get all data at once if you need it



Tutorial topic

Third: brute-force right-click+download-link

Paste the link to the WGET command on the command line, and see if it works or if it instead gets only a weird small file.

```
(geofetch) samuele@D55749:~$ wget https://storage.jpostdb.org/JPST002338/Data210128_Ueno_Kuraishi16_CR2.wiff.scan
--2024-10-02 15:31:29-- https://storage.jpostdb.org/JPST002338/Data210128_Ueno_Kuraishi16_CR2.wiff.scan
Resolving storage.jpostdb.org (storage.jpostdb.org)... 133.39.78.111
Connecting to storage.jpostdb.org (storage.jpostdb.org)|133.39.78.111|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 4373527172 (4.1G) [application/octet-stream]
Saving to: 'Data210128_Ueno_Kuraishi16_CR2.wiff.scan'

Data210128_Ueno_Kur  0%[                               ] 50.23K  30.2KB/s
```

Tutorial topic

Fourts: specific software or access procedure

- Some databases have specific ways of downloading
 - GEO+SRA uses sra-tools to download in sra format, which must be converted to fastq
 - you can get direct link to ftp download in GEO, but still you need sra→ fastq conversion
- Some useful tools are developed to make it easier to download from those databases
 - geofetch (tutorial at abc.au.dk/Documentation)
 - ffq (<https://github.com/pachterlab/ffq>)

Come with your input

<https://tinyurl.com/makeyourabc>

or from our home

<https://abc.au.dk/#suggestions>



Tutorial and open coding

At our home <https://abc.au.dk/Documentation> you can find

- 5 tutorials
- 1 conference workshop

Or you can code and ask for help or generic coding/bioinf/data science questions

