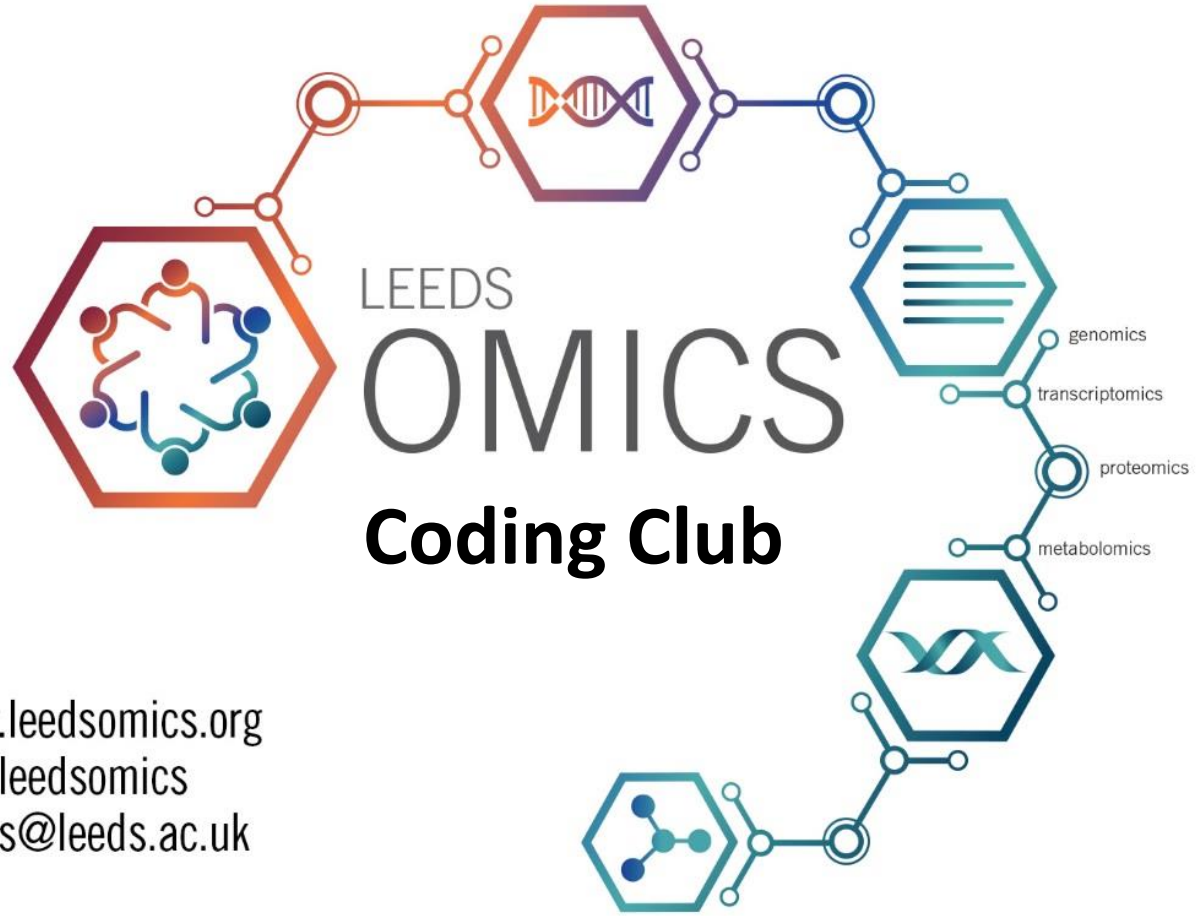


Coding Club

www.leedsomics.org
@leedsomics
omics@leeds.ac.uk

Accessing and Downloading Data from Online Databases

Club Moderators: Elton Vasconcelos, Euan McDonell, Chew Cheng, and Dapeng Wang

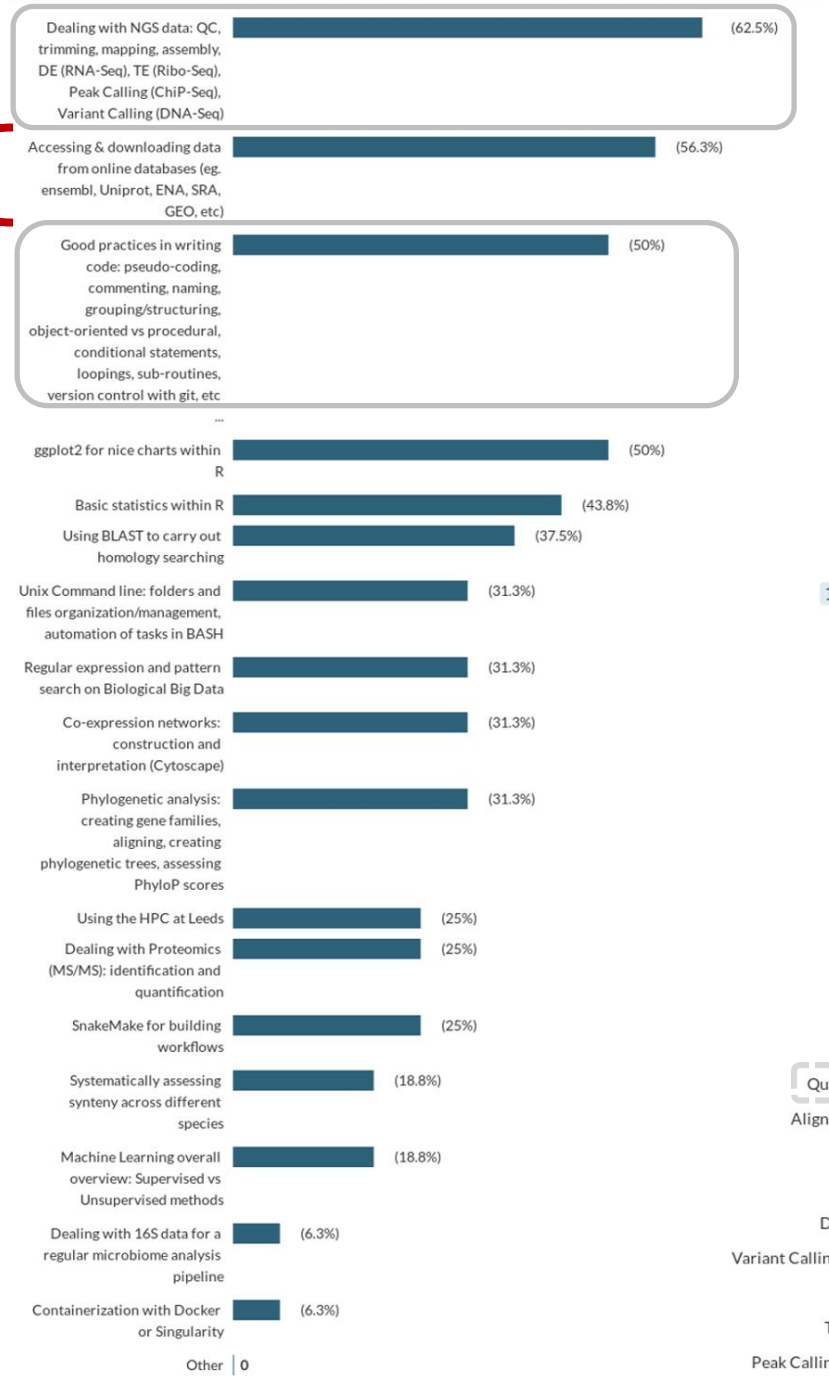
Topics to be addressed on the **2020-21** season - **Survey Result**

1st, 4th, 6th, 8th, ... sessions

2nd session (today)

3rd, 5th, 7th, 9th, ... sessions

1 Which of the following topics would you like to attend in our Coding Club sessions?



LEEDS OMICS

genomics
transcriptomics
proteomics
metabolomics

Coding Club

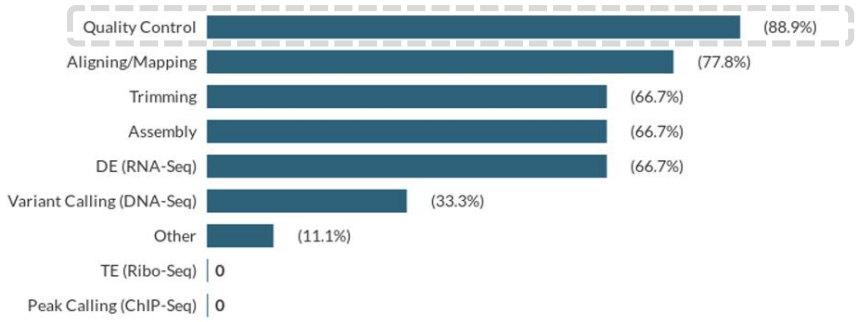
Survey Result (2020-2021)

www.leedsomics.org
@leedsomics
omics@leeds.ac.uk

1.b Do you think we should address "Good practices in writing code" topic in more sessions?

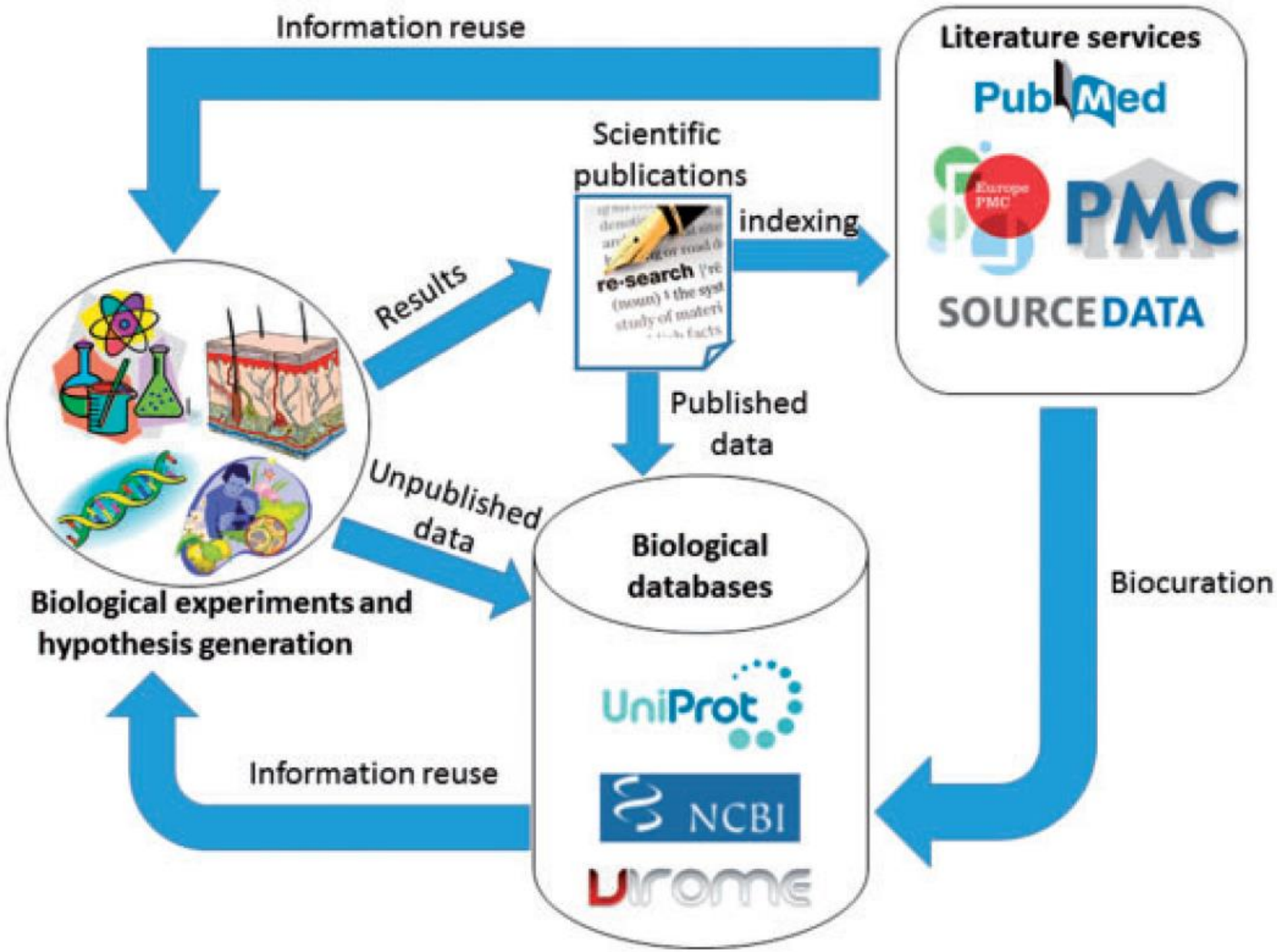
1.c Do you think we should address "Dealing with NGS data" topic in more sessions?

1.c.i Which sub-topics would be of most interest to you?

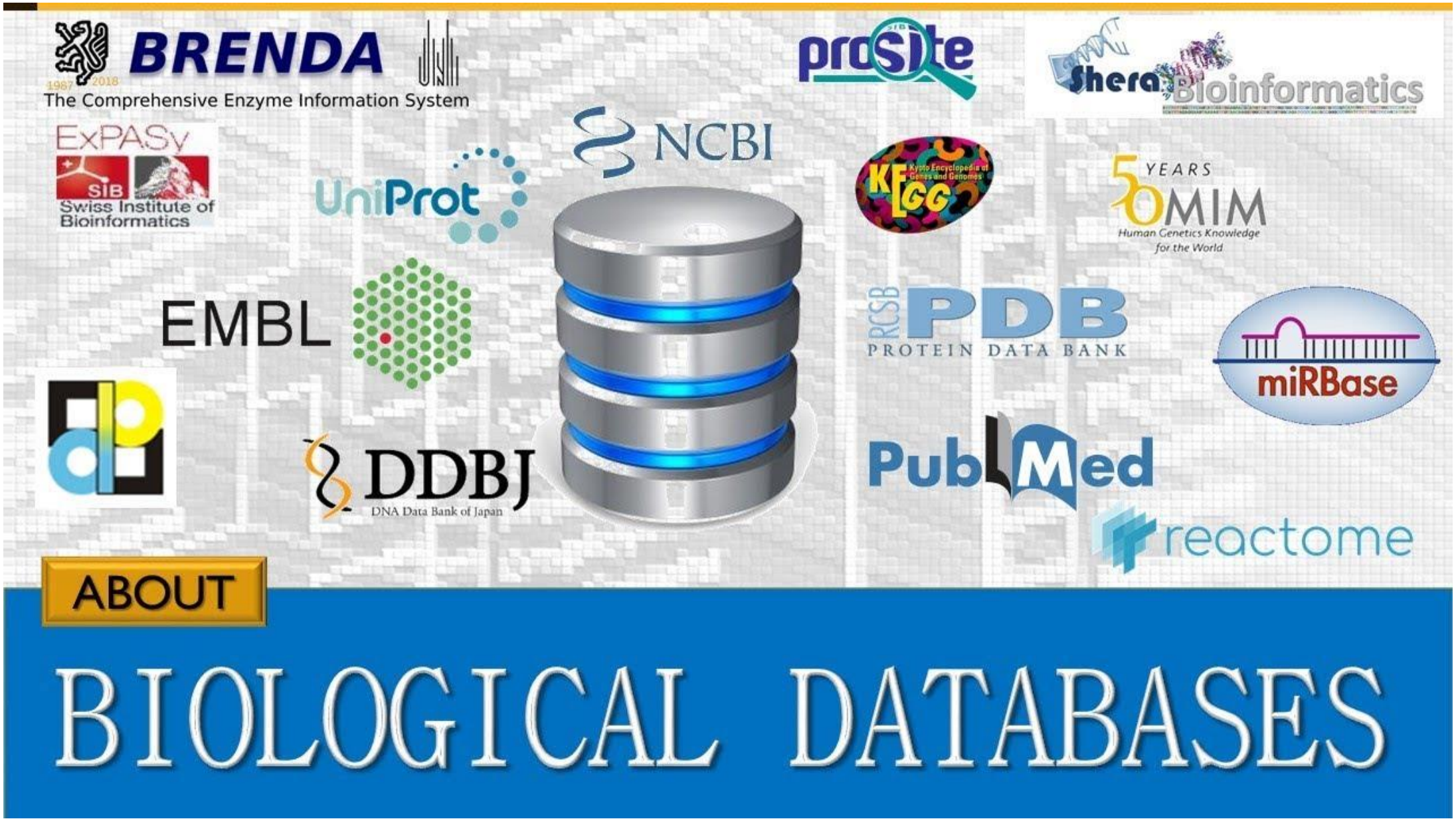


1st session (15/06/20)

Interconnection between literature services and biological databases



Singhal et al., 2016 (*Database*)



https://www.youtube.com/watch?v=0-u54qTYoy0 (Shera Bioinformatics, 2018)

- **Tips for downloading large amount of heavy data from any biological database**

A. Via ftp repositories (batch mode)

1. Find out whether your target web resource stores its data on an ftp repository, and localize the ftp address where your desired files are placed in;

2. Create a plain text file ("file.txt") where each line is a different target ftp address pointing to a specific file to be downloaded;

3. Under the directory (folder) where you saved "file.txt", run the following command on any Unix Shell terminal (Mac or Linux) :

```
$ for i in `cat file.txt`; do wget -c $i; done
```

B. Via http web pages

1. Right-clicking on your desired file link shall display something like "Copy Web Link address";

2. Go to your Unix Shell terminal and under your desired working directory type **wget -c** and paste the copied address from B1 step;

   → If doing for several ("n") files, repeat B1 step "n" times combined with A2, and then run A3.

- **Dealing specifically with NGS raw data**

→ e.g. Fastq files from NCBI-SRA database

1. Create a plain text file (“file.txt”) where each line points to a target SRA “run” accession number (SRR or ERR);

```
SRR7691569
SRR7691583
SRR7691584
SRR7691591
SRR7691595
```

2. Download and install the NCBI sratoolkit software;

https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software → OS-specific download links

https://www.ncbi.nlm.nih.gov/books/NBK158900/ → HOW-TO tutorial

3. Under the directory (folder) where you saved “file.txt”, run the following command on any Unix Shell:

```
$ for i in `cat file.txt`; do /path/to/your/sratoolkit/bin/fastq-dump --split-3 $i; done
```

NOTE: Most of SRA-deposited fastq files are quite heavy, consequently fastq-dump executions can be quite time-consuming. In that case, one may consider running on the system’s background with both nohup and xargs commands.

```
$ nohup cat file.txt | xargs –i /path/to/your/sratoolkit/bin/fastq-dump --split-3 {} &
```

Both command lines work the same, but the latter will print download progress log messages on an output file called nohup.out, rather than on the terminal screen like the former. “Nohup” therefore releases your terminal prompt so you can keep working on other tasks in the same terminal screen while your download goes on.

Bring your issues on!