MANCHESTER
1824
The University of Manchester

UoM UK Biobank User Community Meet-up

June 2019

Data Download Tools

George Leaver, Research IT

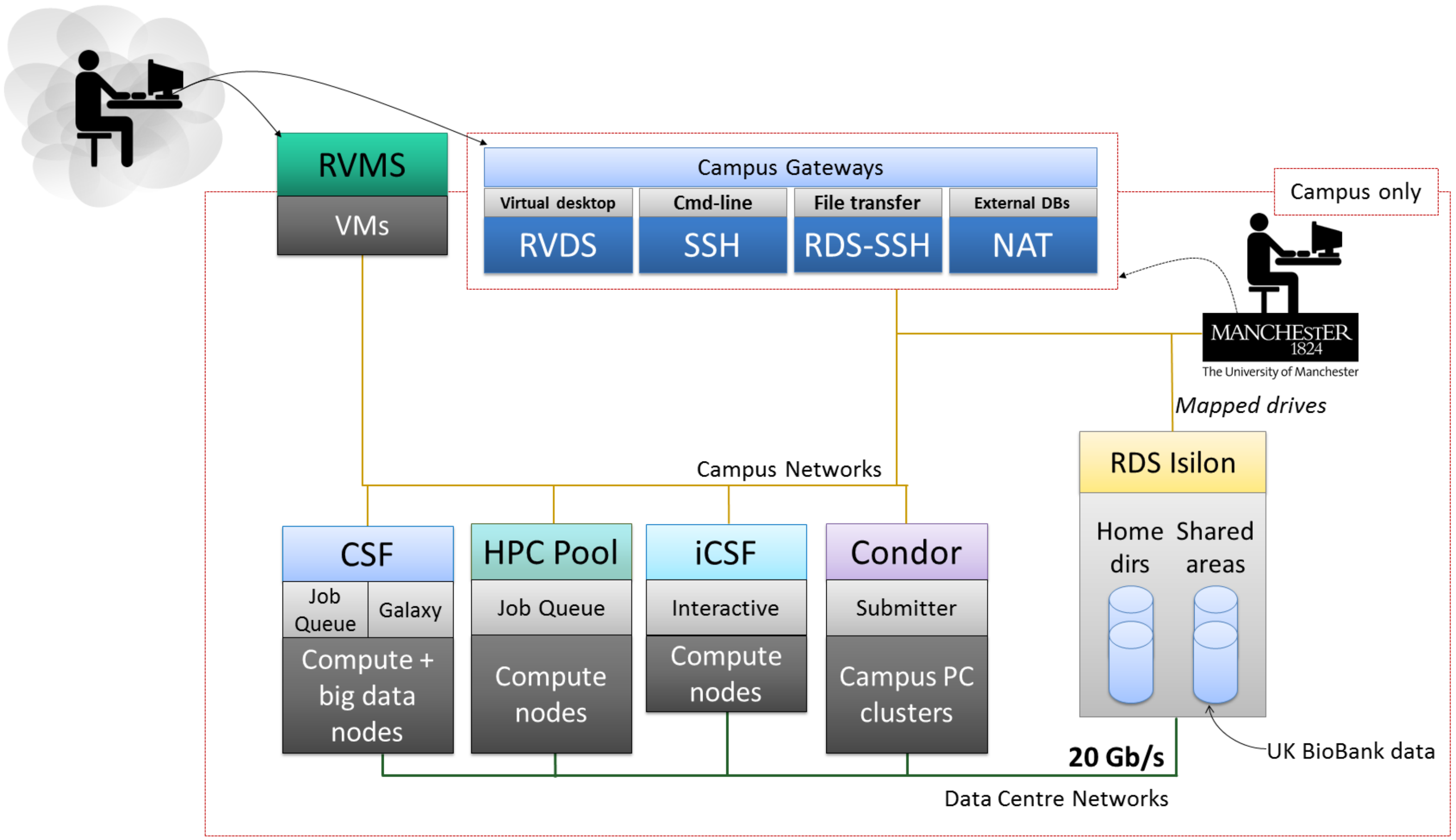
Downloading Datasets

- Basics: What are the *data-provider*'s requirements?
  - Register a username/email address?
  - Obtain a password / key-file?
  - Can you do a web / ftp download? (unlikely)
  - Do you need to use their special / custom download tool? (probably)

- *How* you download determines *which* RI system you can use

- Storage: we will consider hosting a central copy of a dataset if generic enough to be useful to other research groups, *and* there is demand.

Computationally Intensive Research *Ecosystem*



Downloading while on CSF, iCSF

- HTTP/HTTPS/FTP-over-HTTP web downloads (and 'git clone') ONLY
  - Internal 10.99 IP addresses (for security)
  - Access to outside via University web proxy: `proxy.man.ac.uk`

```
module load tools/env/proxy
wget http://example.com/data/data1.tgz
```

  - Submit long-running / large / numerous downloads as batch jobs

```
#!/bin/bash --login
#$ -cwd
module load tools/env/proxy
wget http://example.com/data/data1.tgz
# or 'curl -L http://example.com/data/data1.tgz -o data1.tgz'
```

  - Can save to CSF *home directory*, Research Data Storage area, or *scratch* area

Downloading while on rds-ssh

- rds-ssh.itservices.manchester.ac.uk
  - 130.88 IP address
  - Direct (but restricted) connections *to* outside world
  - Two servers available: "gorg" & "zola"
  - Same *home* directory and any additional RDS areas as CSF, iCSF. NOT *scratch* dir.
  - No *modulefiles* needed to access download tools
  - Specified IP addresses permitted for various download tools (anything missing?):

    biota.ndph.ox.ac.uk, chest.ndph.ox.ac.uk
    ega.ebi.ac.uk, pg-www.ebi.ac.uk, fasp.ebi.ac.uk, ftp-private.ebi.ac.uk,
    Some from: .ed.ac.uk
- HTTP/HTTPS/FTP-over-HTTP web downloads
  - Same as CSF but no need for `module load tools/env/proxy`
  - Web downloads still go through proxy

UKBioBank tools rds-ssh

- Available by default - just run them, no need to install them yourself:

Command	Description	Links
ukbmd5	Calculate size and MD5 (checksum) of a file to verify it	download and basic docs
ukbconv	Convert unpacked UKB data to other formats	download and basic docs
ukbunpack	Unpack (decrypt and decompress) UKB data	download and basic docs
ukbfetch	Download approved **bulk** data files	download (RHEL6) Detailed docs
ukblink	Download Returned-datasets and link between Applications	download (RHEL6) Detailed docs
ukbgene	Download approved **genetic** data (supersedes `gfetch`)	download (RHEL6) Detailed docs

- All rds-ssh installs work (it uses RHEL6 versions)
  - Default downloads are for RHEL7. Email its-ri-team@manchester.ac.uk if having problems on your desktop.
- See http://biobank.ctsu.ox.ac.uk/crystal/download.cgi

The University of Manchester

UKBioBank tools rds-ssh

- Try running an app without any flags / files to get basic help

```
ssh username@rds-ssh.itservices.manchester.ac.uk
(enter password)

ukbgene

ukbgene on unx - ver Mar 14 2018 14:21:52 - using Glibc2.12(stable)
Run start : 2019-06-27T12:11:25
Missing compulsory parameters
Usage: ukbgene datatype [flags]
 -a    authentication file (application_id + 24-char key)
 -c    chromosome (1-26, X, Y, XY or MT)
 -d    name of output datafile
 -h    show this usage message then exit
 -i    show program version information only then exit
 -m    fetch mapping/family file associated with datatype
 -v    verbose mode on

Compiled : Mar 14 2018 14:21:52
```

MANCHESTER 1824
The University of Manchester

UKBioBank tools rds-ssh

- Try running an app without any flags / files to get basic help

```
ssh username@rds-ssh.itservices.manchester.ac.uk
(enter password)

ukbfetch

ukbfetch on unx - ver Jan 30 2019 15:39:51 - using Glibc2.12(stable)
Run start : 2019-06-26T17:35:13
Must specify encoded_id for participant (-e flag)
Usage: ukbfetch parameters...
 -a    authentication file (application_id + 24-char key)
 -b    batch file containing list of participants and datafiles
 -d    datafile name
 -e    encoded id for participant
 -h    show this usage message then exit
 -i    show program version information only then exit
 -m    maximum datafiles to fetch (batch mode only, capped at 50000)
 -o    name of output file recording successful fetches
 -s    starting line (batch mode only)
 -v    verbose mode on
Compiled : Jan 30 2019 15:39:51
```

UKBioBank tools rds-ssh

- UKB repositories need *proof of identity*:
  - The downloader apps accept an *Authentication Keyfile* (a plain text file), containing:
    - Application ID (first line)
    - 64-character *decryption password* (second line)
    - For example (for application 5137 and 64-char *decryption password* sent by UKBioBank):
      ```
      5137
      a1b2c3d4a1b2c3d4a1b2c3d4e5f6a44b343d334eef232ce3d3298ba847abcde
      ```
  - Put this is a text file named `.ukbkey` (notice the `.` at the start)
    (Linux `ls` command won't show the file, use `ls -a` instead)
  - Place it in the folder where you will be running the downloader app from
  - Some apps also accept a flag instead of `.ukbkey` (eg: `-a`*`mykeyfile`*)
- NB: Some apps used a 24-character password (e.g., ukblink)
  - If your downloader complains about a 64-char password, try shortening it in your `.ukbfile` to 24-characters (delete from the end of the line)!!

Download example

- ukbgene (simpler tool than ukbfetch)

```
ukbgene typename -cchrom [flags]
ukbgene cal -c22       # Anonymous genotype calls for Chromosome 17
ukbgene cal -c17 -m    # -m fetches Link file in addition to dataset
```

- Please follow the "Detailed docs" link (also in earlier table) for details of which file types and groups can be downloaded

- *typename*:

Typename	type of data to be retrieved	format	link format
cal	genotype calls	bed	fam
con	genotype confidences	txt	fam
int	genotype intensities	bin	fam
baf	genotype CNV b-allele frequencies	txt	fam
l2r	genotype CNV log2ratios	txt	fam
imp	imputation	bgen	sample
hap	haplotypes	bgen	sample

UKBioBank tools rds-ssh

- Downloading a huge dataset (many 1000s of files) may *time out*:
  - Server may be rate-limiting in some way to reduce the load
- Try doing it in batches if the downloader app supports it.
- For example: `ukbfetch`
  - Can read a text file with participant-ID and data-file-ID pairs on each line

    ```
    529523 2323_0_0
    529585 2348_0_1
    529585 2348_1_1
    …
    ```

  - According to the ukbfetch docs can download up to 50,000 data files in one go
  - but 1000 seems to be the actual limit

Bulk Download example

```
# Go to you data dir
cd ~/my/data/area
# Create a .ukbkey from the file sent to you by UKB for use with ukbfetch
cp ~/k9876.key .ukbkey

# Verify checksum of downloaded encrypted file [Q: how is this downloaded - via the UKB showcase?]
ukbmd5 ukb12345.enc

# Decrypt using auth keyfile for our Application 9876 (contains App ID and 64-char password)
ukbunpack ukb12345.enc ~/k9876.key
# Generate list of bulk downloads for field 456, say. Data-type to convert to is 'bulk'. Generates  ukb12345.bulk
ukbconv ukb12345.enc_ukb bulk -s456

# Now download in batches of 1000 (eg ukb12345.bulk has 10000s of lines)
# Start reading ukb12345.bulk from line one, then line 1001, and so on. Batch size (-m) is 1000.
# We log successful downloads to a file named downloaded.1-1000.lis (ukbfetch will add .lis to filename)
# Auth keyfile .ukbkey is used (see *). Alternatively add to the ukbfetch command: -ak9876.key
ukbfetch -v -bukb12345.bulk -s1 -m1000 -odownloaded.1-1000
ukbfetch -v -bukb12345.bulk -s1001 -m1000 -odownloaded.1001-2000
…
# If you want to count how many downloads were logged in each .lis file:
wc -l downloaded.1-1000.lis
```

Script it

- Write a simple bash script you can leave running on rds-ssh

```
#!/bin/bash
i=0
BATCHSIZE=1000
NUMBATCHES=130     # Num lines in ukb12345.bulk <= BATCHSIZE x NUMBATCHES
while [ $i -lt $NUMBATCHES ]; do
  START=$((i*BATCHSIZE+1))
  ukbfetch -v -bukb12345.bulk -s$START -m1000 -odownloaded.$START
  # Report how many lines are in the .lis
  # (assuming 1 line output per downloaded file)
  wc -l downloaded.$START.lis
  ((i++))
done
```

EGA Tools on rds-ssh

- Available by default, just run them

Command	Description	Links
`egaclient`	v2.2.2 Java batch and *interactive* downloader for gentyping and imputation datasets (e.g., EGAD00010001497)	download and basic docs
`egacryptor`	Encrypts files and generates md5sum for submission to EGA	download and basic docs

- Username (your email addr) and the password of your EGA account needed
- Decryption Key-file `ega.key` containing 64-char password required (plain text)
  - When using commands that require the key, don't give it the name of the file.
  - Do give it the actual password in the file!
- For convenience the alias `egaclient` actually runs:

```
java -jar <install_path>/EgaDemoClient.jar
```

egaclient on rds-ssh

- You first make a *request* for a dataset and possibly specific files within it

```
egaclient -p demo@test.org 123pass -rfd EGAD00010000498 -re abc \
          -label request_EGAD00010000498
```

- Then you *download the request* to get the files (using eg 7 parallel streams)

```
egaclient -p demo@test.org 123pass -dr request_EGAD00010000498 -nt 7
```

- Then you *decrypt* the downloaded files (getting the 64-char p/w from a file)

```
egaclient -p demo@test.org 123pass filename -dck `cat ega.key`
```

- Full reference and examples at
  https://ega-archive.org/download/downloader-quickguide-v2
  https://ega-archive.org/download/using-ega-download-client

MANCHESTER
1824
The University of Manchester

Aspera on rds-ssh

- Available by default, just run them

Command	Description	Links
ascp	v3.3.3 (previously used by EGA/UKBioBank before ukbfetch?)	EGA ascp basic docs
ascp_noid	ascp but *without* the default aspera_web id key	EGA ascp basic docs

- Performs parallel downloads of large datasets
- The `ascp` command uses a default ID file which some remote servers use
- If a different auth method is used by data provider, run ascp_noid

  `ascp_noid` (will ask for password)

  ```
  ascp_noid -i <path>/file_id_dsa.openssh
  ```

- Example - download all data from ega-box-800 (ports 33001-33010)

```
# -Q: Use fair download (shares bandwidth), -T turn of encryption for speed, -L- log to screen, -l 1000M limit bandwidth
ascp -QT -L- -l 1000M ega-box-800@fasp.ega.ebi.ac.uk:. .
```

Illumina Basemount on rds-ssh

- Only one user from CRUK has requested / used this. Installed, just run it.

  "I have just received via BaseSpace the results from a sequencing run from the CRUK-MI"

- basemount is a method of remote-mounting a *basespace* filesystem
- Uses an encrypted (port 443) connection

  ```
  basemount basespace
  cd basespace
  ls
  cd MySharedData
  ```

  (perform file-commands as usual)

- For help: `man basemount`
- https://basemount.basespace.illumina.com/

Our Hosted Datasets

- Currently the *500,000 participant* datasets (v3)
  - Imputation data: EGAD00010001474
  - Genotyping data: EGAD00010001497
  - Activity data …. in progress
- How to obtain access to our copy
  - Principal Investigator (PI) informs UK BioBank of their intent to use an institute-held dataset *during the application review process (preferably)*
  - PI *must* name people who will access the data on the *Material Transfer Agreement* (MTA)
  - Research IT will provide access following confirmation a user is on the MTA
  - http://ri.itservices.manchester.ac.uk/hosted-data-sets/ukbiobank/

access@ukbiobank.ac.uk

Contact Us

- Any other tools needed?

- Please Contact the Research Infrastructure Team

  its-ri-team@manchester.ac.uk