Downloading data

SEQAFRICA Module 2

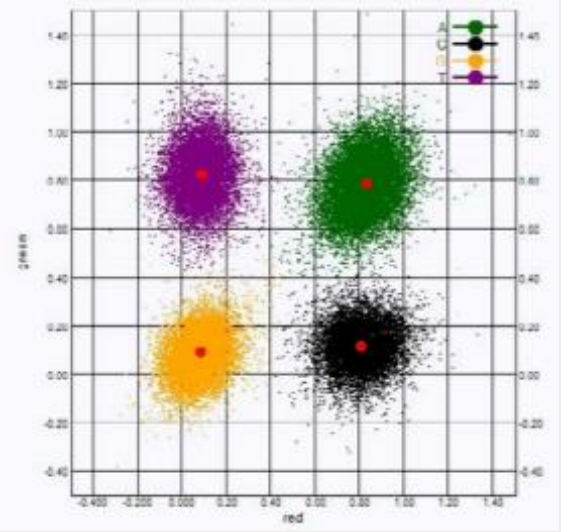
19 February 2021

Stanford Kwenda
Lead Data Analyst

National Institute for Communicable Diseases
South Africa

Illumina Output Files and Folders

- Illumina instruments generate raw data files in binary base call (BCL) format
- Some systems (e.g MiniSeq, MiSeq and NextSeq) provide the option to automatically convert data from BCL to FASTQ format
  - Base calling determines a base (A, C, G, or T) for every cluster of a given tile at a specific cycle
  - The base calling process determines to which population each cluster belongs
  - bcl2fastq2 conversion software for FASTQ conversion
- Quality scoring is a prediction of the probability of an incorrect base call
  - Assigns a quality score to every base call
  - After the Q-score is determined results for each tile are recorded in the base call files
- Base call files
  - Aggregated in one file for each lane for each cycle
  - Contains the base call and encoded quality score for every cluster for that lane



Illumina Output Files and Folders

The Fleming Fund Regional Grants

File structure on the instrument

- Data
  - Intensities
    - BaseCalls
      - L001 — Base call files for lane 1, aggregated in one file per cycle.
      - L002 — Base call files for lane 2, aggregated in one file per cycle.
      - L003 — Base call files for lane 3, aggregated in one file per cycle.
      - L004 — Base call files for lane 4, aggregated in one file per cycle.
    - L001 — An aggregated *.locs file for lane 1.
    - L002 — An aggregated *.locs file for lane 2.
    - L003 — An aggregated *.locs file for lane 3.
    - L004 — An aggregated *.locs file for lane 4.
  - Images
    - Focus
      - L001 — Focus images for lane 1.
      - L002 — Focus images for lane 2.
      - L003 — Focus images for lane 3.
      - L004 — Focus images for lane 4.
  - InterOp — Binary files used by Sequencing Analysis Viewer (SAV).
  - Logs — Log files describing operational steps.
  - Recipe — Run-specific recipe file named with reagent cartridge ID.
  - RTALogs — Log files describing analysis steps.
  - Thumbnail_Images — Thumbnail images for tiles 1, 6, and 12 in each swath at every cycle.

bcl2fastq

Fastq files Copied to

```
HJH37AFX2/Alignment_1/20210111_182955/Fastq$ ls -l N00447*
 N00447_S51_L001_R1_001.fastq.gz
 N00447_S51_L001_R2_001.fastq.gz
 N00447_S51_L002_R1_001.fastq.gz
 N00447_S51_L002_R2_001.fastq.gz
 N00447_S51_L003_R1_001.fastq.gz
 N00447_S51_L003_R2_001.fastq.gz
 N00447_S51_L004_R1_001.fastq.gz
 N00447_S51_L004_R2_001.fastq.gz
```

Merged Fastq ready for downstream analysis

```
H37AFX2/Alignment_1/20210111_182955/Fastq$ ls -l merged/N00447*
merged/N00447_S51_L001_R1_001.fastq.gz
merged/N00447_S51_L001_R2_001.fastq.gz
```

SampleName_S1_L001_R1_001.fastq.gz

- N00447
- Sample number in sample sheet
- Lane number
- The read R1==> Read 1

Setting output folder location

1. BaseSpace Sequence Hub Configuration

1 From the Home screen, select **Manage Instrument**.

2 Select **System Configuration**.

3 Select **BaseSpace Sequence Hub Configuration**.

4 Select from the following options to specify a location where data are transferred for analysis.

- From the Hosting Location list, select the location of the server where data are uploaded to.
- If you have an Enterprise subscription, select the **Private Domain** checkbox and enter the domain name (URL) used for single sign-on to BaseSpace Sequence Hub.

For example: https://**yourlab**.basespace.illumina.com.

5 Select **Save**.

2. Output Folder Location

- For Local Run Manager run mode set the output folder location in the Local Run Manager software

```
stan@LT-StanfordK:~$ ls /mnt/y/201222_M02143_0218_000000000-J45T2/Data/Intensities/BaseCalls/
13440_S7_L001_R1_001.fastq.gz   13814_S13_L001_R1_001.fastq.gz   14000_S19_L001_R1_001.fastq.gz   Alignment
13440_S7_L001_R2_001.fastq.gz   13814_S13_L001_R2_001.fastq.gz   14000_S19_L001_R2_001.fastq.gz   config.xml
13499_S8_L001_R1_001.fastq.gz   13815_S14_L001_R1_001.fastq.gz   14118_S20_L001_R1_001.fastq.gz   gBlock-SetD_S25_
13499_S8_L001_R2_001.fastq.gz   13815_S14_L001_R2_001.fastq.gz   14118_S20_L001_R2_001.fastq.gz   gBlock-SetD_S25_
13506_S9_L001_R1_001.fastq.gz   13871_S15_L001_R1_001.fastq.gz   14128_S21_L001_R1_001.fastq.gz   L001
13506_S9_L001_R2_001.fastq.gz   13871_S15_L001_R2_001.fastq.gz   14128_S21_L001_R2_001.fastq.gz   Matrix
13718_S10_L001_R1_001.fastq.gz  13874_S16_L001_R1_001.fastq.gz   15064_S22_L001_R1_001.fastq.gz   Phasing
13718_S10_L001_R2_001.fastq.gz  13874_S16_L001_R2_001.fastq.gz   15064_S22_L001_R2_001.fastq.gz   SampleSheet.csv
```

Data in the instrument output folder

Illumina data transfer

- BaseSpace Sequence Hub
  - Cloud-based data management powered by Amazon Web Services (AWS)
  - Encrypted data streamed from instrument to BaseSpace Sequence Hub
  - Requires registration and is licensed

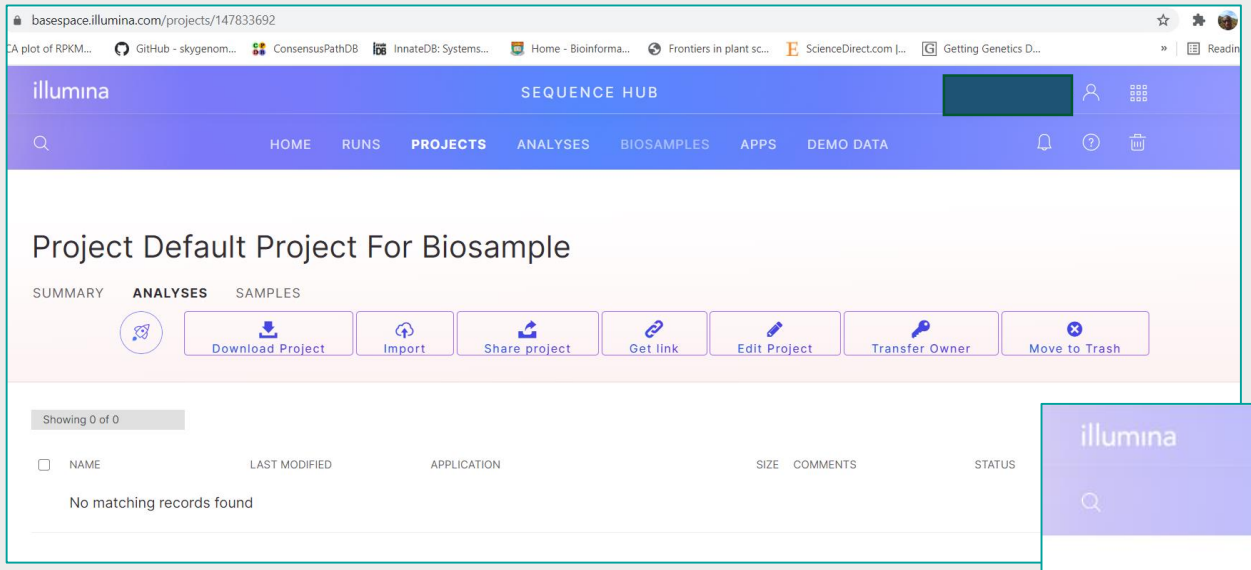
Management of high volume sequencing etc.

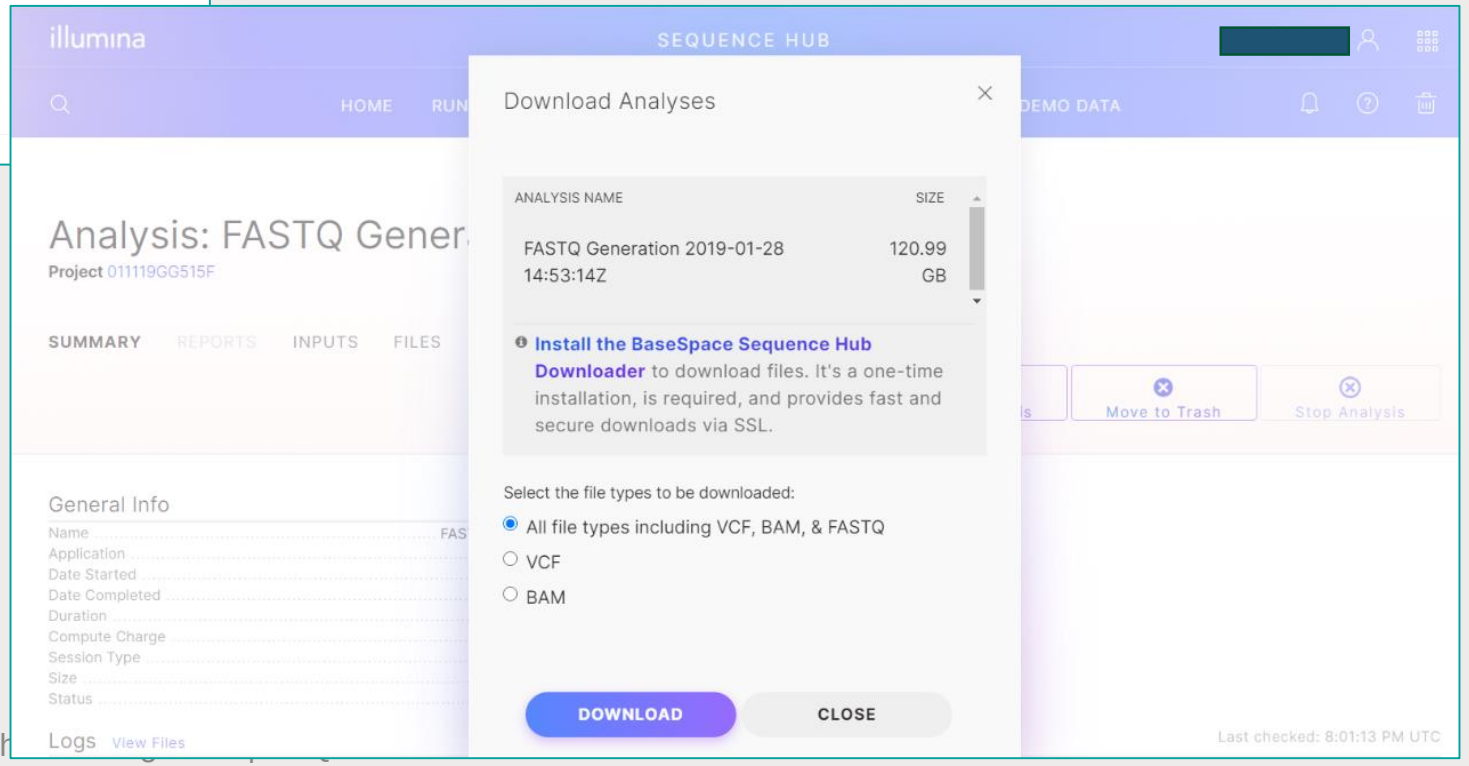
Small to medium sized labs – access to BaseSpace Bioinformatics Apps

Free for Individual users – 1TB storage

The Fleming Fund Regional Grants

BaseSpace Sequence Hub





Data transfer Using a Local Server

Set Output Folder Location

The NextSeq requires an output folder for all runs. Use the full Universal Naming Convention (UNC) path to the output folder. The UNC path includes two backslashes, the server name and directory name, but *not* a letter for a mapped network drive.

- Paths to the output folder that are one level require a trailing backslash.
  - Example UNC path: \\servername\directory1\
- Paths to the output folder that are two or more levels do not require a trailing backslash.
  - Example UNC path: \\servername\directory1\directory2
- Paths to a mapped network drive cause errors. *Do not use.*
  - Example of a mapped network drive path: T:\sbsfiles

Ask your IT support team to set-up your instrument on the network

Data transfer directly from instrument

1. Mount output folder via CLI

```
# mount output folder on e.g. MiSeq Instrument to your Linux computer/server
# example using the WSL terminal
sudo mount -t drvfs '\\10.10.10.10\MiSeqOutput' /mnt/y
```

Mount Instrument's Output folder

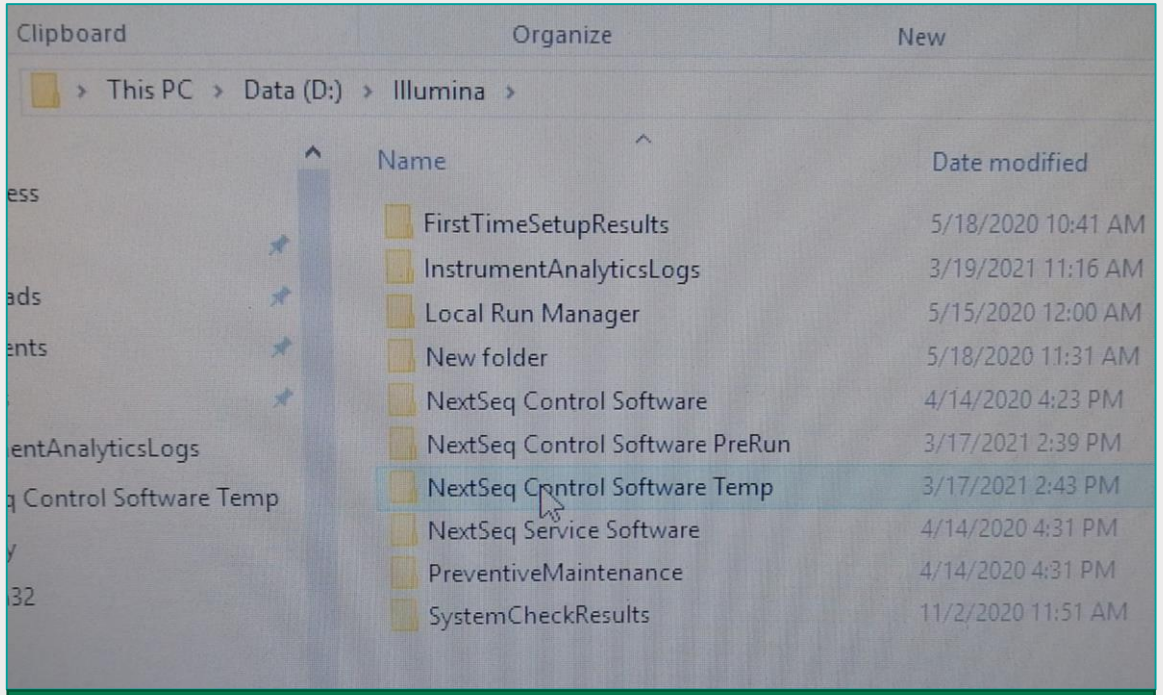
```
stan@LT-StanfordK:~$ copy-illumina-data ~/MiSeq-Run-ID-Lists/run_201222_M02143_0218_000000000-J45T2 /mnt/c/Users/stanfordk/Documents/Data-Delivery
sending incremental file list
rsync: change_dir "/mnt/x/MiSeqOutput/201222_M02143_0218_000000000-J45T2/" failed: No such file or directory (2)
rsync error: errors selecting input/output files, dirs (code 3) at flist.c(2118) [sender=3.1.1]
sending incremental file list
rsync: change_dir "/mnt/x/MiSeqOutput/201222_M02143_0218_000000000-J45T2/" failed: No such file or directory (2)
rsync error: errors selecting input/output files, dirs (code 3) at flist.c(2118) [sender=3.1.1]
find: ‘/mnt/x/MiSeqOutput/201222_M02143_0218_000000000-J45T2/Data/Intensities/BaseCalls’: No such file or directory
sending incremental file list
created directory /mnt/c/Users/stanfordk/Documents/Data-Delivery/201222_M02143_0218_000000000-J45T2
RunInfo.xml
            685 100%    0.00kB/s    0:00:00 (xfr#1, to-chk=11/12)
SampleSheet.csv
          4,258 100%   64.97kB/s    0:00:00 (xfr#2, to-chk=10/12)
config.xml
          7,081 100%   30.73kB/s    0:00:00 (xfr#3, to-chk=9/12)
runParameters.xml
          3,455 100%   10.88kB/s    0:00:00 (xfr#4, to-chk=8/12)
InterOp/
InterOp/ControlMetricsOut.bin
      1,322,401 100%  886.96kB/s    0:00:01 (xfr#5, to-chk=6/12)
InterOp/CorrectedIntMetricsOut.bin
      1,127,234 100%  787.98kB/s    0:00:01 (xfr#6, to-chk=5/12)
InterOp/ErrorMetricsOut.bin
```

Copy data e.g. using rsync to external hard disk or server

Data transfer directly from instrument

2. Manually download data via external hard disk

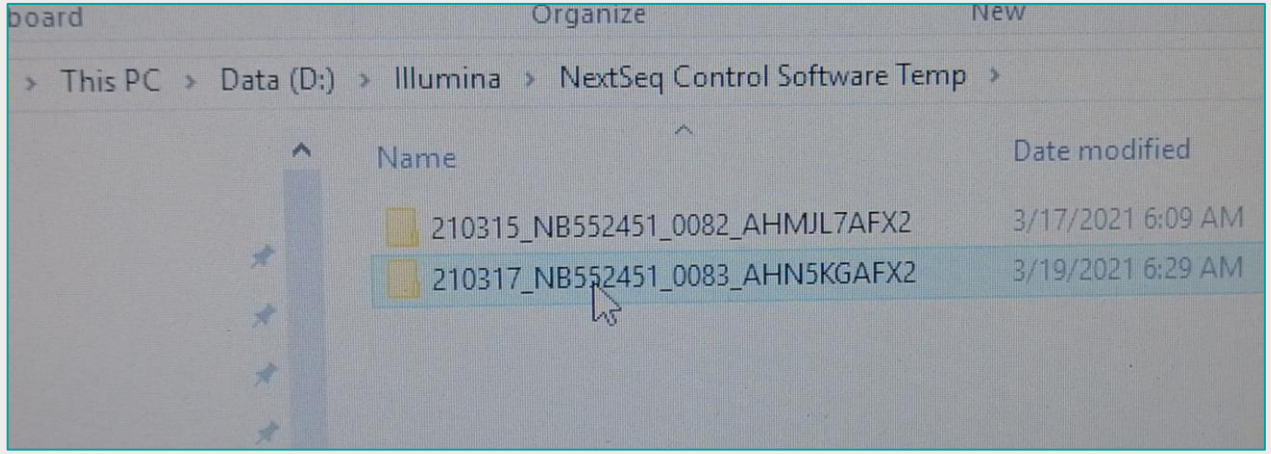
Connect external hard disk

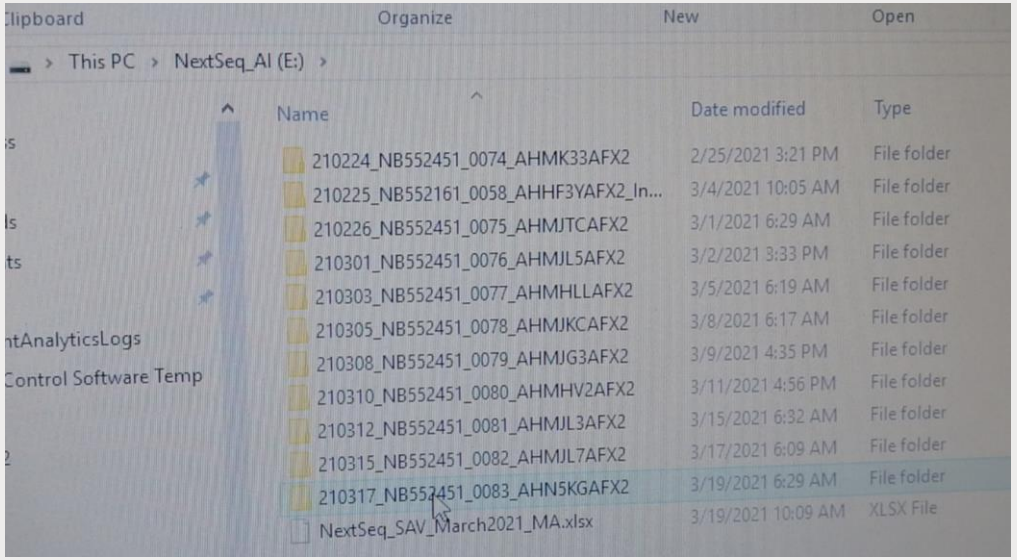


- NextSeq uses Windows 10 OS
- Navigate to the **D-drive ==> Illumina folder ==> NextSeq Control Software Temp**

Data transfer directly from instrument

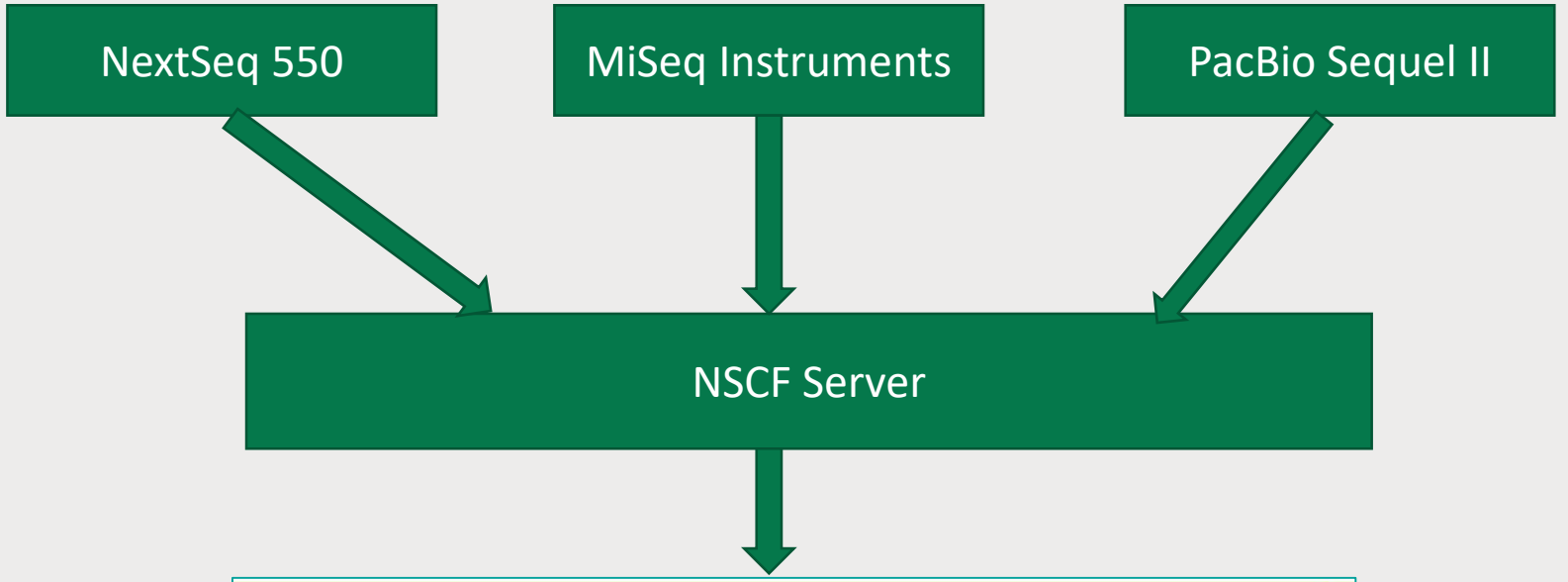
2. Manually download data via external hard disk





- Copy data from the NextSeq instrument to the external hard disk

Data transfer at NICD Sequencing Core Facility



https://nxtshare.nicd.ac.za/index.php/apps/files/?dir=/&fileid=97764

The Fleming Fund
Regional Grants

Thank you

NOGUCHI MEMORIAL INSTITUTE FOR MEDICAL RESEARCH
UNIVERSITY OF GHANA , LEGON

UNIVERSITY OF IBADAN
RECTE SAPERE FONS

NATIONAL INSTITUTE FOR COMMUNICABLE DISEASES
Division of the National Health Laboratory Service

KCRI Kilimanjaro Clinical Research Institute
Academic Centre for Evidence Based Health Interventions
An Institution of the Good Samaritan Foundation

UKaid
from the British people

MOTT MACDONALD

SEQAFRICA

DTU

This programme is being funded by the UK Department of Health and Social Care.
The views expressed do not necessarily reflect the UK Government's official policies.