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

File structure on the instrument	bcl2fastq2	→ Fastq	files Copied to	C
The Data and the D	HJH37AFX2/A	Alignment_1/202101	L11_182955/Fas	tq\$ ls -l N00447
🧰 Intensities	N00447_S53	1_L001_R1_001.fast	tq.gz	
🛅 BaseCalls	N00447_S53	1_L001_R2_001.fast	tq.gz	
💼 L001 — Base call files for lane 1, aggregated in one file per cycle.	N00447_S53	1_L002_R1_001.fast	tq.gz	
🛅 L002 – Base call files for lane 2, aggregated in one file per cycle.	N00447_S53	1_L002_R2_001.fast	tq.gz	
💼 L003 – Base call files for lane 3, aggregated in one file per cycle.	N00447_S53	1_L003_R1_001.fast	tq.gz	
💼 L004 – Base call files for lane 4, aggregated in one file per cycle.	N00447_S5	1_L003_R2_001.fast	tq.gz	
L001—An aggregated *.locs file for lane 1.	N00447_S5:	1_L004_R1_001.fast	tq.gz	
L002—An aggregated *.locs file for lane 2.	N00447_S53	1_L004_R2_001.fast	tq.gz	
L003—An aggregated *.locs file for lane 3.				
L004—An aggregated *.locs file for lane 4.				
🧰 Images				
Focus				
L001 – Focus images for lane 1.	N/I	argad Easta raadu	for downstre	am analysis
L002 – Focus images for lane 2.		eigeu i asty i eau		
L003 – Focus images for lane 3.	H37AFX2/Ali	gnment_1/20210111_1	82955/Fastq\$ ls	-1 merged/N00447
L004 – Focus images for lane 4.	merged/N004	4/_S51_L001_R1_001.	fastq.gz	
InterOp—Binary files used by Sequencing Analysis Viewer (SAV).	menged/1004	47_551_L001_K2_001.	rastq.gz	
Logs — Log files describing operational steps.				
Recipe —Run-specific recipe file named with reagent cartridge ID.				
RTALogs — Log files describing analysis steps.	Samp	leName_S1_L001	R1_001.fasto	q.gz
Thumbnail_Images - Thumbnail images for tiles 1, 6, and 12 in each swath at every c	ycle.			_
	N00447 Samp	ole number in	Lane	The read
19 February 2021	The Fleming Fund SEQAL Sa	mple sheet	number	R1==> Read 1

19 February 2021

The Fleming Fund | SEQA



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	// 01222_M02143_0218_000000000-J45	5T2/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_	Data in the
13499_S8_L001_R2_001.fastq.gz	13815_S14_L001_R2_001.fastq.gz	14118_S20_L001_R2_001.fastq.gz	gBlock-SetD_S25_	instrument output
13506_S9_L001_R1_001.fastq.gz	13871_S15_L001_R1_001.fastq.gz	14128_S21_L001_R1_001.fastq.gz	LOOI	instrument output
13506_S9_L001_R2_001.fastq.gz	13871_S15_L001_R2_001.fastq.gz	14128_S21_L001_R2_001.fastq.gz	Matrix	folder
13718_S10_L001_R1_001.fastq.g	13874_S16_L001_R1_001.fastq.gz	15064_S22_L001_R1_001.fastq.gz	Phasing	
13718_S10_L001_R2_001.fastq.g	13874_S16_L001_R2_001.fastq.gz	15064_S22_L001_R2_001.fastq.gz	SampleSheet.csv	



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





BaseSpace Sequence Hub

basespace:illumina.com/projects/147833692	🖈 😕 🌚		
plot of RPKM 🎧 GitHub - skygenom 😫 ConsensusPathDB 📓 InnateDB: Systems 🧔 Home - Bioinforma 🔇 Frontiers in plant sc 📙 ScienceDirect.com I 🜀 Getting Genetic	ss D » 🗏 Readin		
illumina sequence нив	A III		
Q HOME RUNS PROJECTS ANALYSES BIOSAMPLES APPS DEMO DATA	Д [©] ф		
Droiget Default Droiget For Piecemple			
SUMMARY ANALYSES SAMPLES			
Download Project Import Share project Get link Edit Project Transfer Owner	Move to Trash		
Chausian 0 of 0			
	illumina	SEQUENCE HUB	A
No matching records found	O HOME	Download Analyses	
	_		
	Analysis: FASTO Ge		
	Project 011119GG515F	FASTQ Generation 2019-01-28 120.99 14:53:14Z GB	
		•	
	SUMMARY REPORTS INPUTS F	FILES	
		installation, is required, and provides fast and	e Move to Trach Ston Analysis
		secure downloads via SSL.	Stop Analysis
		Select the file types to be downloaded:	
	Name	All file types including VCF, BAM, & FASTQ	
	Application Date Started	○ VCF	
	Date Completed Duration	○ BAM	
	Compute Charge Session Type		
	Size Status		
10 February 2021		DOWNLOAD CLOSE	Last checked: 8:01:13 PM U



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

The Fleming Fund 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

<pre>stan@LT-StanfordK:~\$ copy-illumina-data ~/MiSeq-Run-ID-Lists/run_201222_M02143_0218_000000000-J45T2 /mnt/c/Users/stanfordk/Documents/Data-Del</pre>	ivery
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_00000000-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/MiSegOutput/201222 M02143 0218 00000000-J45T2/Data/Intensities/BaseCalls': No such file or directory	Convidate a guilling
sending incremental file list	Copy data e.g. using
created directory /mnt/c/Users/stanfordk/Documents/Data-Delivery/201222 M02143 0218 000000000-J45T2	
RunInfo.xml	rsvnc to external hard
685 100% 0.00kB/s 0:00:00 (xfr#1, to-chk=11/12)	
SampleSheet.csv	disk or server
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	



2. Manually download data via external hard disk



Connect external hard disk

Clipboard		Organize	New
> This PC	> Data (D:)) > Illumina >	
	^	Name	Date modified
ess		FirstTimeSetupResults	5/18/2020 10:41 AM
	×	InstrumentAnalyticsLogs	3/19/2021 11:16 AM
ads	. *	Local Run Manager	5/15/2020 12:00 AM
ents	*	New folder	5/18/2020 11:31 AM
	*	NextSeq Control Software	4/14/2020 4:23 PM
entAnalyticsLogs		NextSeq Control Software PreRun	3/17/2021 2:39 PM
a Control Software Temp		NextSeq Control Software Temp	3/17/2021 2:43 PM
		NextSeq Service Software	4/14/2020 4:31 PM
y		PreventiveMaintenance	4/14/2020 4:31 PM
132		SystemCheckResults	11/2/2020 11:51 AM

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



2. Manually download data via external hard disk



lipboard		Organize	Vew	Open
-> This PC >	NextSeq_	AI (E:) >		
	^	Name	Date modified	Туре
s		210224 NB552451 0074 AHMK33AFX2	2/25/2021 3:21 PM	File folder
	*	210225 NB552161 0058 AHHF3YAFX2_In	3/4/2021 10:05 AM	File folder
s	. *	210226 NB552451 0075 AHMJTCAFX2	3/1/2021 6:29 AM	File folder
ts	*	210301 NB552451 0076 AHMJL5AFX2	3/2/2021 3:33 PM	File folder
	-	210303 NB552451 0077 AHMHLLAFX2	3/5/2021 6:19 AM	File folder
		210305 NB552451 0078 AHMJKCAFX2	3/8/2021 6:17 AM	File folder
tAnalyticsLogs		210308 NB552451 0079 AHMJG3AFX2	3/9/2021 4:35 PM	File folder
Control Software Temp		210310 NB552451 0080 AHMHV2AFX2	3/11/2021 4:56 PM	File folder
		210312 NB552451 0081 AHMJL3AFX2	3/15/2021 6:32 AM	File folder
		210315 NB552451 0082 AHMJL7AFX2	3/17/2021 6:09 AM	File folder
		210317 NB557451 0083 AHN5KGAFX2	3/19/2021 6:29 AM	File folder
		NextSeq_SAV_March2021_MA.xlsx	3/19/2021 10:09 AM	

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



Data transfer at NICD Sequencing Core Facility





Thank you



NOGUCHI MEMORIAL INSTITUTE FOR MEDICAL RESEARCH UNIVERSITY OF GHANA, LEGON







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.

