Generation and Management of Large Sequence Files:

Perspectives from the DNA Sequencing Core

Alvaro G. Hernandez, Ph.D.
Director of DNA Services
University of Illinois at Urbana-Champaign
Roy J. Carver Biotechnology Center

Genomic Information Representation
April 18, 2018

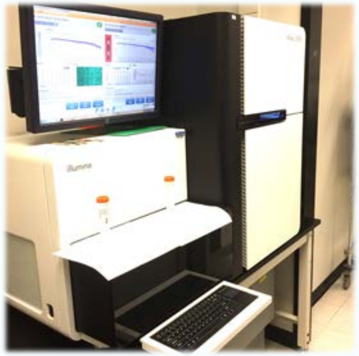
MPEG

Outline

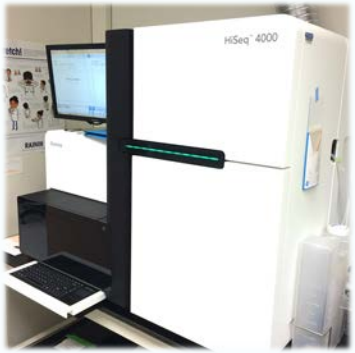
- Overview of the lab and of our portion of the NGS market
- Workflow of genomics data generation
- Size of Illumina run folders and fastq files
- Size and number of files from Oxford Nanopore
- Final considerations

www.biotech.illinois.edu/htdna

HiSeq 2500



HiSeq 4000



NovaSeq 6000

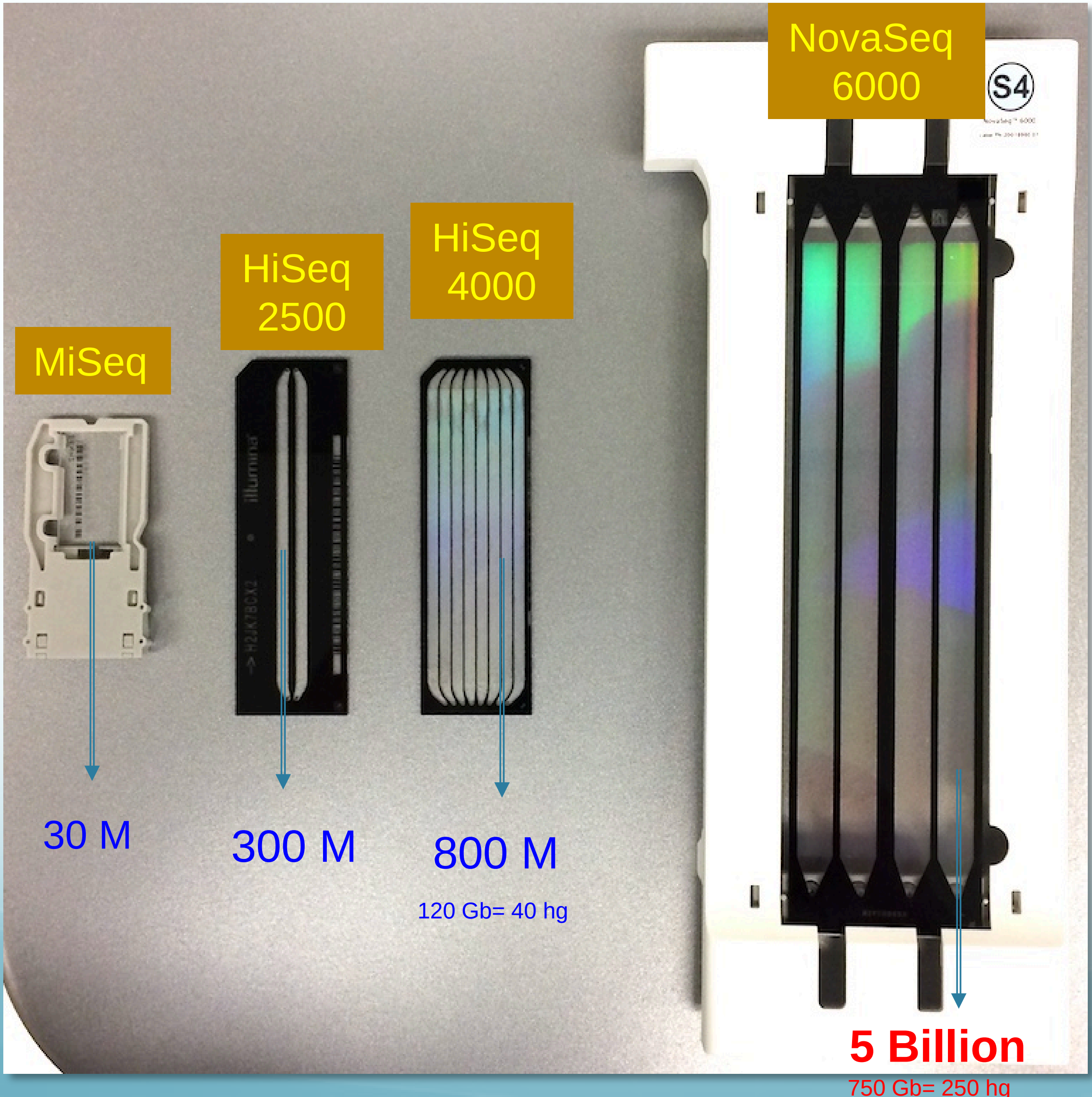
3 MiSeq

10x Genomics
Single-cell RNAseq
de novo assembly

Oxford Nanopore
GridION x5

1.5 PB tape
active archive

Output (# of reads) per lane



With every new instrument, researchers can do more for comparatively less ➔ projects keep increasing in size and amount of data generated in the core facilities (size or number of files) increases

HT DNA Sequencing Laboratory

Unlimited market ➔ massive amount of data being generated

Illumina: 9k sequencing instruments worldwide

Biologists, veterinarians, agronomists, engineers, all fields of science

The amount of data

- In 2017 we produced:

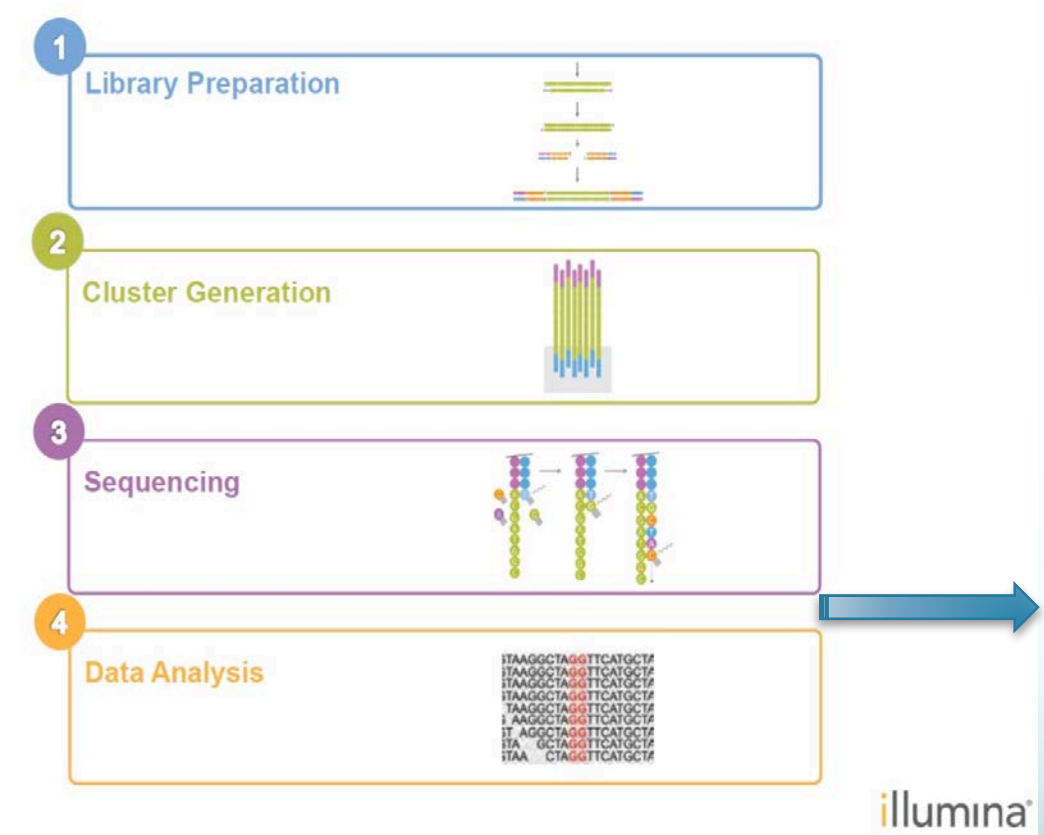
  580 TB of raw data (bcl + fastq files)
  posted 65 TB of compressed (bz2 files)
  archived 68 TB of compressed files

- NovaSeq: each run produces ~ 5 TB to 12 TB of raw data
- Oxford Nanopore produces ~ 100 GB to 1+ TB per run, many flowcells needed per genome

This is RAW data from one of hundreds of facilities in the US and the world

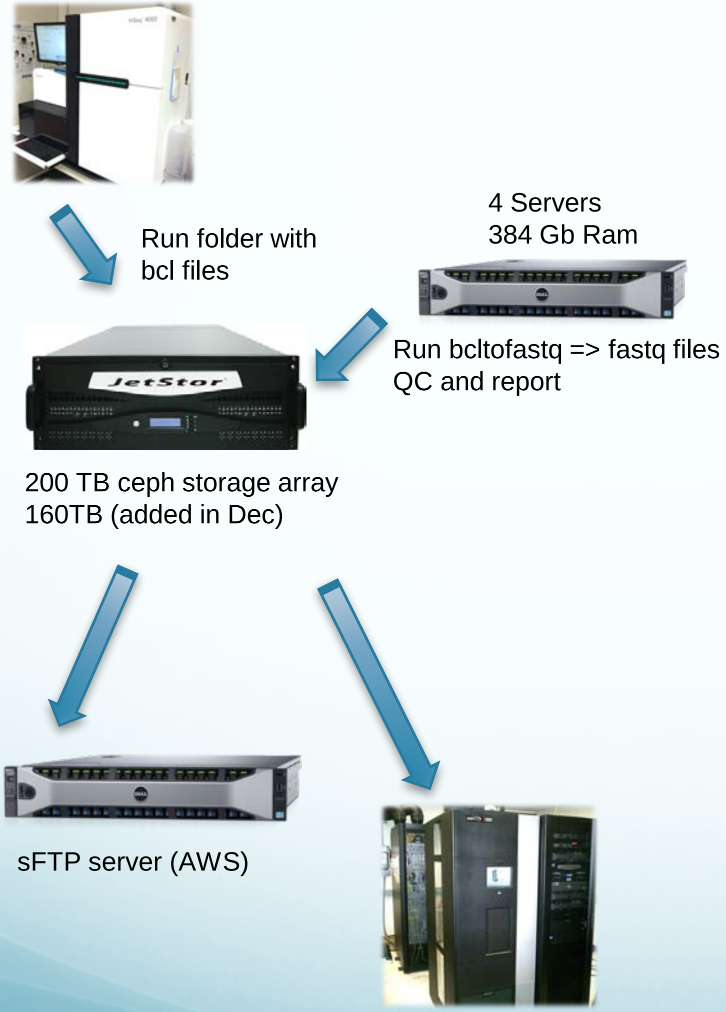
From a core perspective, the problem is the massive efforts and expenses dedicated to network, storage and other hardware infrastructure to handle the ever expanding datasets as well as the duplication of the data

Illumina Sequencing Workflow



- Transfer bcl files to server
- Demutliplex and generate fastq files with bcltofastq
- QC fastq files with FastQC
- Prepare reports for users
- Post files to sFTP server or S3 AWS
- Store runs for a few months (4-6)
- Archive runs (5 years)

Our Workflow



Run folder with
bcl files

4 Servers
384 Gb Ram

Run bcltofastq => fastq files
QC and report

200 TB ceph storage array
160TB (added in Dec)

sFTP server (AWS)

archive

Our Workflow

Run folder with bcl files

4 Servers
384 Gb Ram

Run bcltofastq => fastq files
QC and report

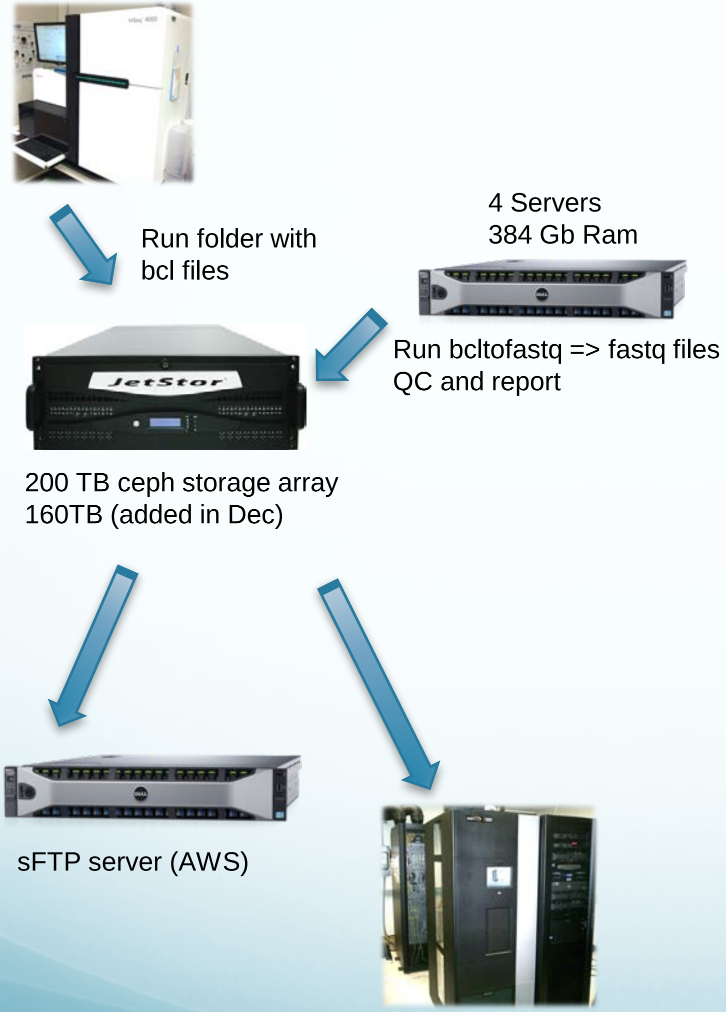
200 TB ceph storage array
160TB (added in Dec)

sFTP server (AWS)

archive

	HiSeq 4000 PE run (151x8x8x151)	
Number of bcl files	288,000	

Our Workflow



Run folder with bcl files

4 Servers
384 Gb Ram

Run bcltofastq => fastq files
QC and report

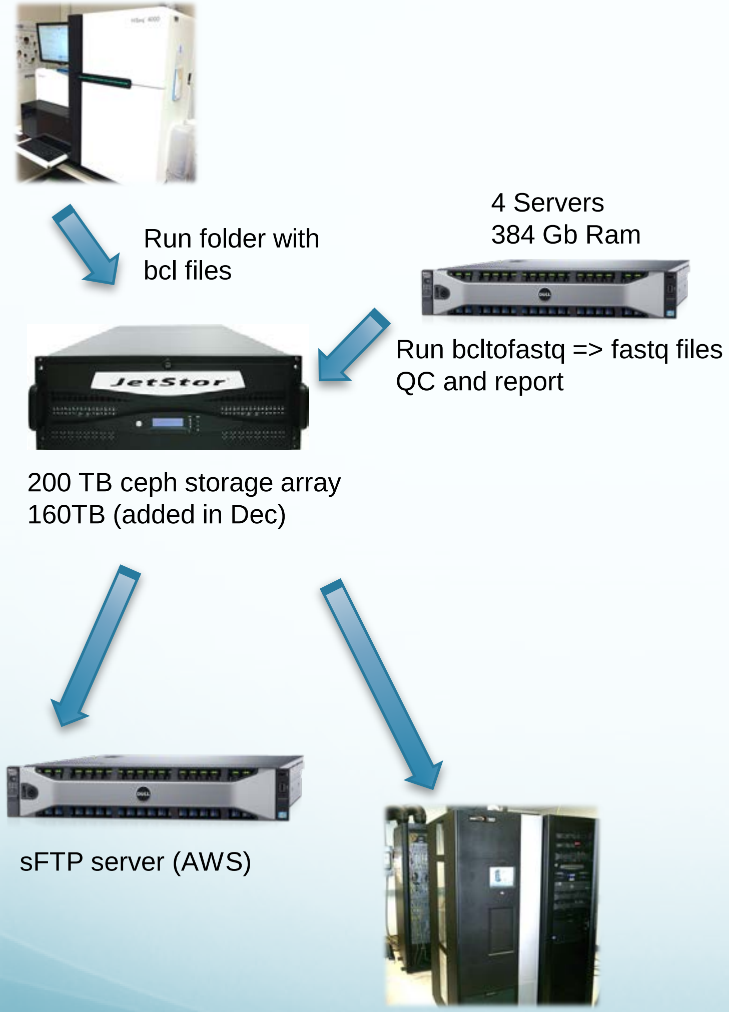
200 TB ceph storage array
160TB (added in Dec)

sFTP server (AWS)

archive

	HiSeq 4000 PE run (151x8x8x151)	
Number of bcl files	285,000	
bcl files total size	560 GB	

Our Workflow



Run folder with
bcl files

4 Servers
384 Gb Ram

Run bcltofastq => fastq files
QC and report

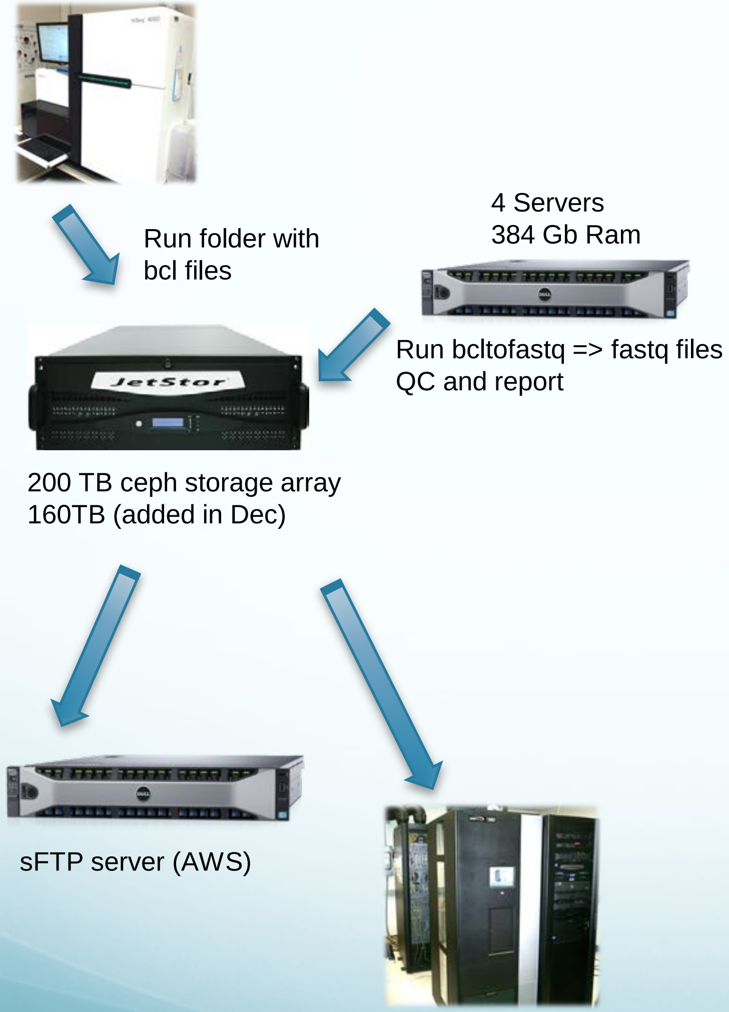
200 TB ceph storage array
160TB (added in Dec)

sFTP server (AWS)

archive

	HiSeq 4000 PE run (151x8x8x151)	
Number of bcl files	288,000	
bcl files total size	560 GB	
fastq files total size	4 TB	

Our Workflow



Run folder with
bcl files

4 Servers
384 Gb Ram

Run bcltofastq => fastq files
QC and report

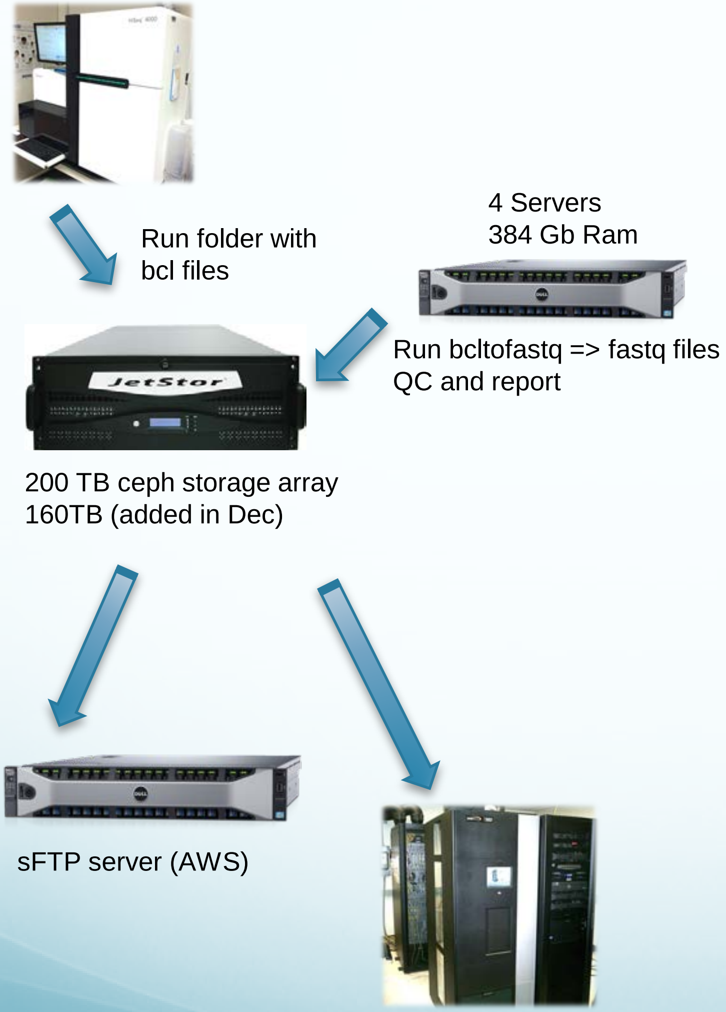
200 TB ceph storage array
160TB (added in Dec)

sFTP server (AWS)

archive

	HiSeq 4000 PE run (151x8x8x151)	
Number of bcl files	288,000	
bcl files total size	560 GB	
fastq files total size	4 TB	
Compressed files delivered to user	384 GB	

Our Workflow



Run folder with
bcl files

4 Servers
384 Gb Ram

Run bcltofastq => fastq files
QC and report

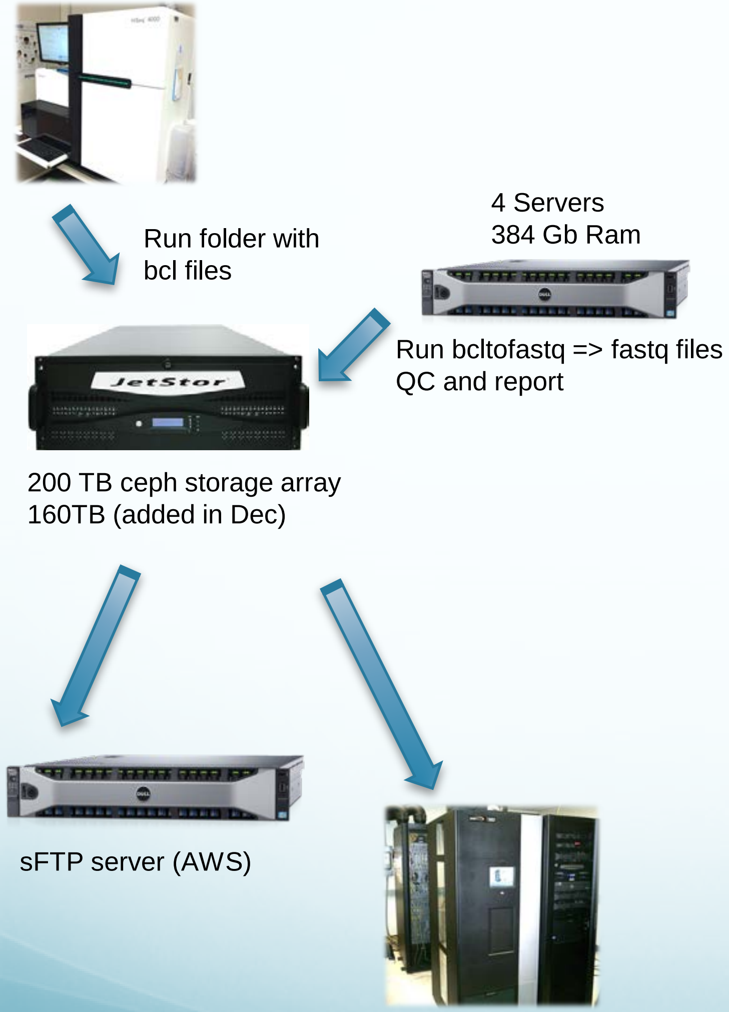
200 TB ceph storage array
160TB (added in Dec)

sFTP server (AWS)

archive

	HiSeq 4000 PE run (151x8x8x151)	
Number of bcl files	288,000	
bcl files total size	560 GB	
fastq files total size	4 TB	
Compressed files delivered to user	384 GB	
Compressed run folder to archive	400 GB	

Our Workflow



Run folder with
bcl files

4 Servers
384 Gb Ram

Run bcltofastq => fastq files
QC and report

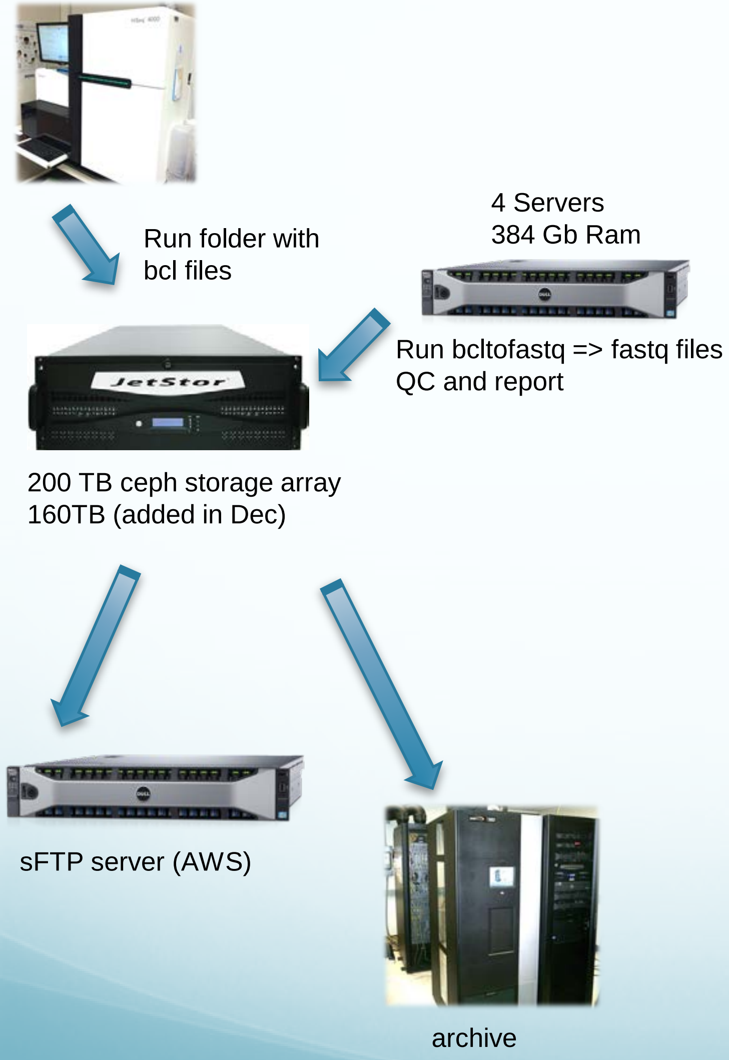
200 TB ceph storage array
160TB (added in Dec)

sFTP server (AWS)

archive

	HiSeq 4000 PE run (151x8x8x151)	
Number of bcl files	288,000	
bcl files total size	560 GB	
fastq files total size	4 TB	
Compressed files delivered to user	384 GB	
Compressed run folder to archive	400 GB	
Total 2017 **~ 20 runs a month**	580 TB raw 65 TB posted 68 TB archived	

Our Workflow



Run folder with bcl files

4 Servers
384 Gb Ram

Run bcltofastq => fastq files
QC and report

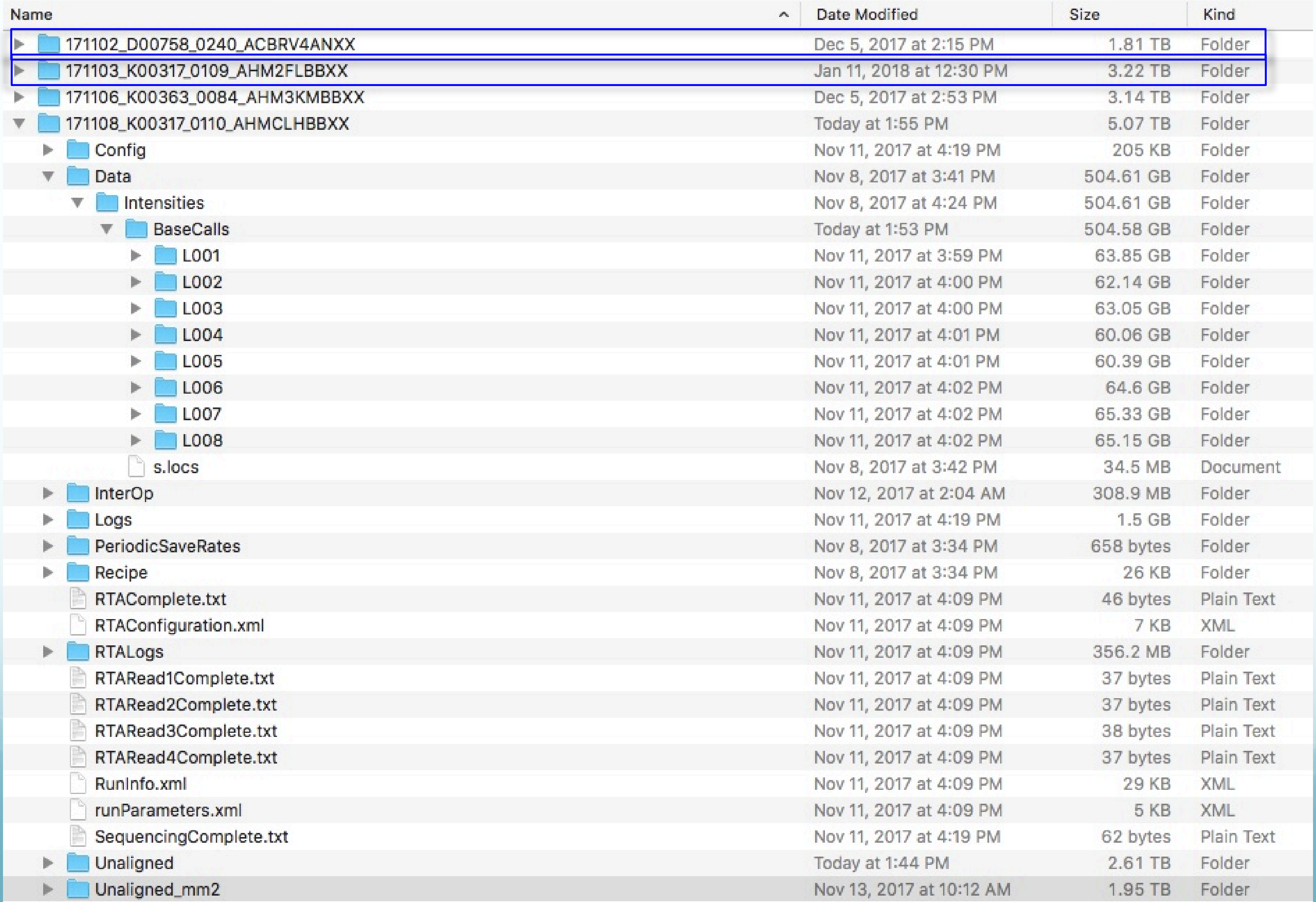
200 TB ceph storage array
160TB (added in Dec)

sFTP server (AWS)

archive

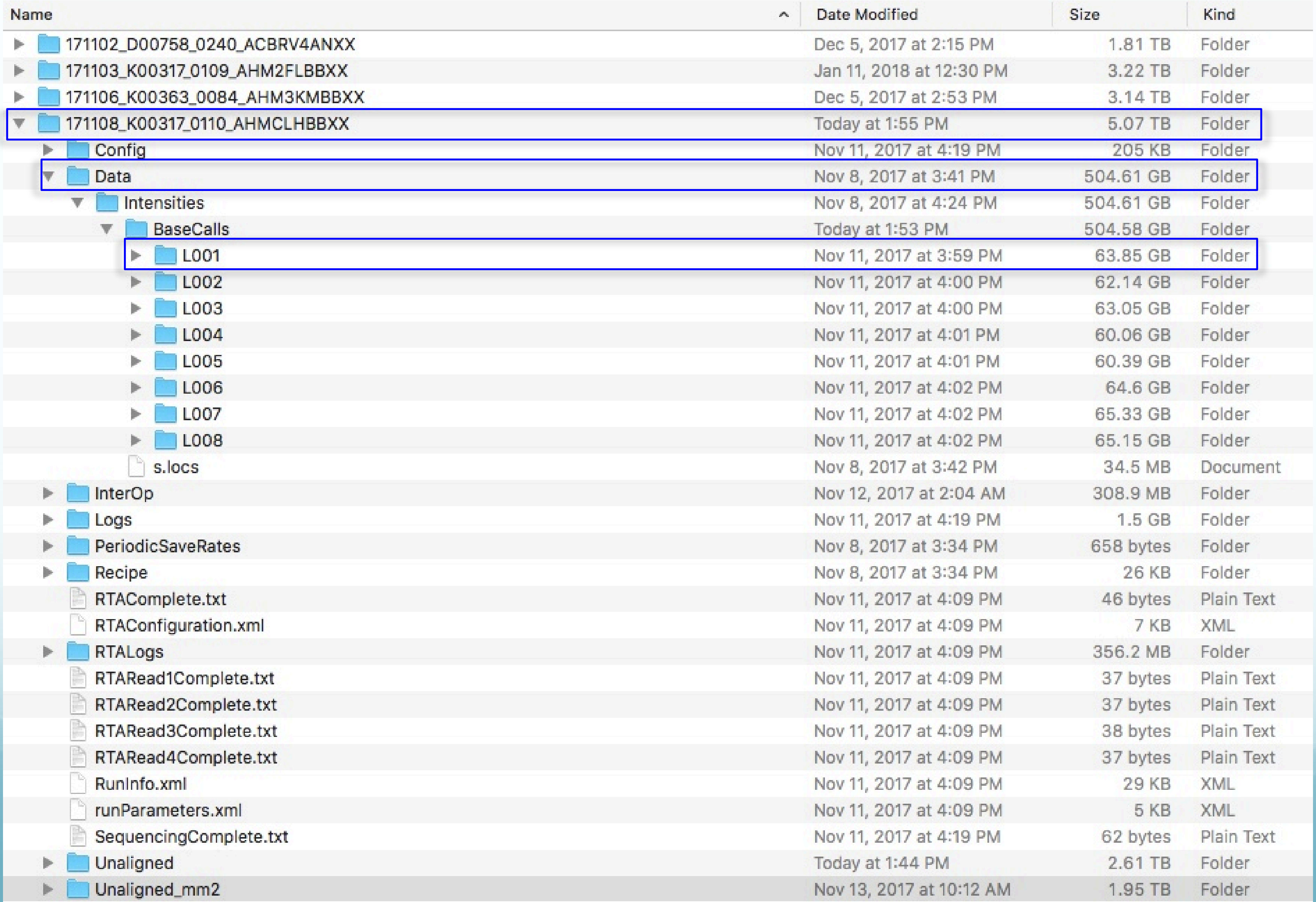
	HiSeq 4000 PE run (151x8x8x151)	NovaSeq PE run 151x8x8x151
Number of bcl files	288,000	1,272
bcl files total size	560 GB	2.2TB
fastq files total size	4 TB	10 TB
Compressed files delivered to user	384 GB	1.2 TB
Compressed run folder to archive	400 GB	---
Total 2017 **~ 20 runs a month**	580 TB raw 65 TB posted 68 TB archived	2018: much more

The Run Folders



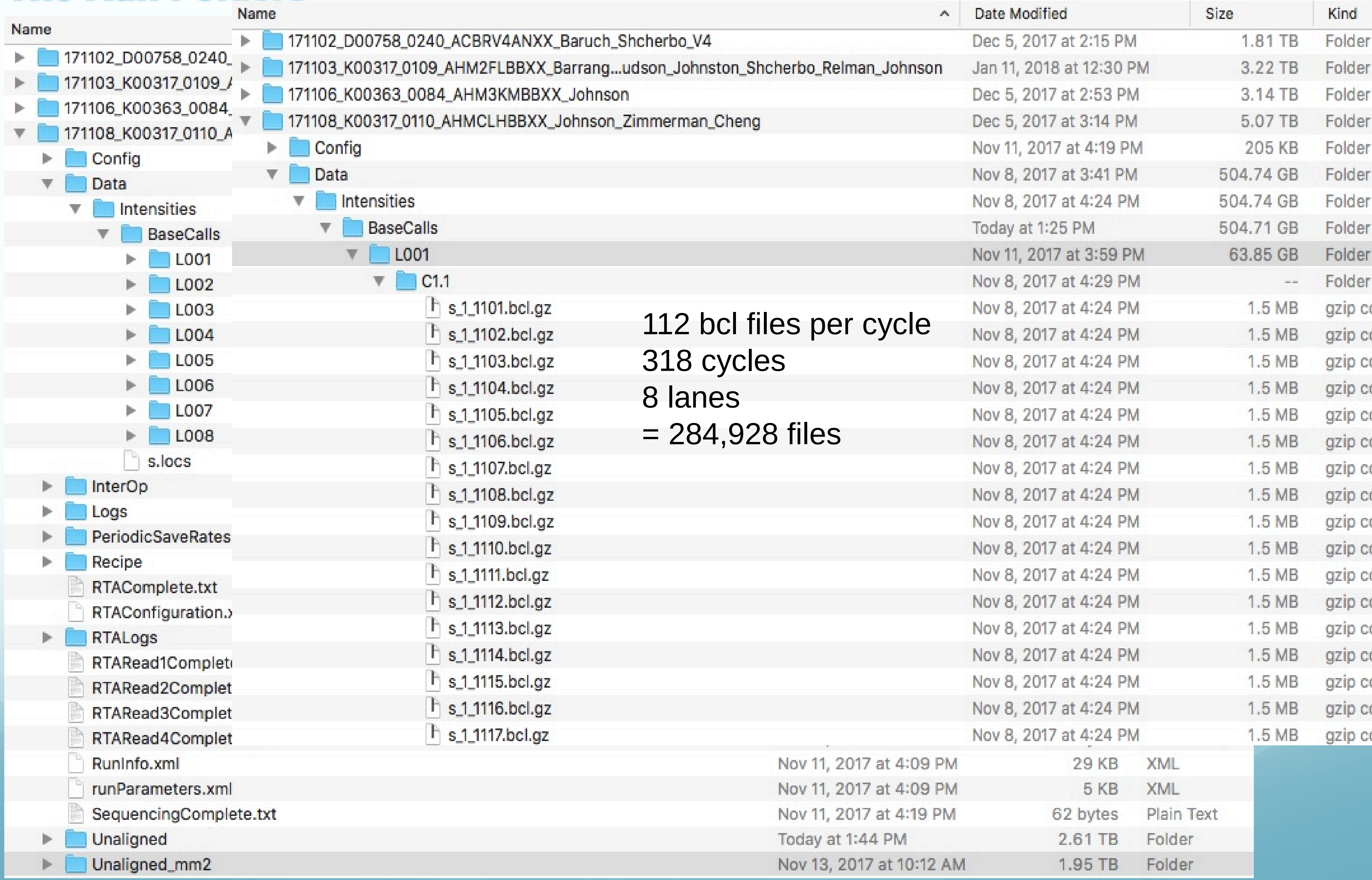
Name	Date Modified	Size	Kind
▶ 171102_D00758_0240_ACBRV4ANXX	Dec 5, 2017 at 2:15 PM	1.81 TB	Folder
▶ 171103_K00317_0109_AHM2FLBBXX	Jan 11, 2018 at 12:30 PM	3.22 TB	Folder
▶ 171106_K00363_0084_AHM3KMBBXX	Dec 5, 2017 at 2:53 PM	3.14 TB	Folder
▼ 171108_K00317_0110_AHMCLHBBXX	Today at 1:55 PM	5.07 TB	Folder
▶ Config	Nov 11, 2017 at 4:19 PM	205 KB	Folder
▼ Data	Nov 8, 2017 at 3:41 PM	504.61 GB	Folder
▼ Intensities	Nov 8, 2017 at 4:24 PM	504.61 GB	Folder
▼ BaseCalls	Today at 1:53 PM	504.58 GB	Folder
▶ L001	Nov 11, 2017 at 3:59 PM	63.85 GB	Folder
▶ L002	Nov 11, 2017 at 4:00 PM	62.14 GB	Folder
▶ L003	Nov 11, 2017 at 4:00 PM	63.05 GB	Folder
▶ L004	Nov 11, 2017 at 4:01 PM	60.06 GB	Folder
▶ L005	Nov 11, 2017 at 4:01 PM	60.39 GB	Folder
▶ L006	Nov 11, 2017 at 4:02 PM	64.6 GB	Folder
▶ L007	Nov 11, 2017 at 4:02 PM	65.33 GB	Folder
▶ L008	Nov 11, 2017 at 4:02 PM	65.15 GB	Folder
s.locs	Nov 8, 2017 at 3:42 PM	34.5 MB	Document
▶ InterOp	Nov 12, 2017 at 2:04 AM	308.9 MB	Folder
▶ Logs	Nov 11, 2017 at 4:19 PM	1.5 GB	Folder
▶ PeriodicSaveRates	Nov 8, 2017 at 3:34 PM	658 bytes	Folder
▶ Recipe	Nov 8, 2017 at 3:34 PM	26 KB	Folder
RTAComplete.txt	Nov 11, 2017 at 4:09 PM	46 bytes	Plain Text
RTAConfiguration.xml	Nov 11, 2017 at 4:09 PM	7 KB	XML
▶ RTALogs	Nov 11, 2017 at 4:09 PM	356.2 MB	Folder
RTARead1Complete.txt	Nov 11, 2017 at 4:09 PM	37 bytes	Plain Text
RTARead2Complete.txt	Nov 11, 2017 at 4:09 PM	37 bytes	Plain Text
RTARead3Complete.txt	Nov 11, 2017 at 4:09 PM	38 bytes	Plain Text
RTARead4Complete.txt	Nov 11, 2017 at 4:09 PM	37 bytes	Plain Text
RunInfo.xml	Nov 11, 2017 at 4:09 PM	29 KB	XML
runParameters.xml	Nov 11, 2017 at 4:09 PM	5 KB	XML
SequencingComplete.txt	Nov 11, 2017 at 4:19 PM	62 bytes	Plain Text
▶ Unaligned	Today at 1:44 PM	2.61 TB	Folder
▶ Unaligned_mm2	Nov 13, 2017 at 10:12 AM	1.95 TB	Folder

The Run Folders



Name	Date Modified	Size	Kind
171102_D00758_0240_ACBRV4ANXX	Dec 5, 2017 at 2:15 PM	1.81 TB	Folder
171103_K00317_0109_AHM2FLBBXX	Jan 11, 2018 at 12:30 PM	3.22 TB	Folder
171106_K00363_0084_AHM3KMBBXX	Dec 5, 2017 at 2:53 PM	3.14 TB	Folder
171108_K00317_0110_AHMCLHBBXX	Today at 1:55 PM	5.07 TB	Folder
Config	Nov 11, 2017 at 4:19 PM	205 KB	Folder
Data	Nov 8, 2017 at 3:41 PM	504.61 GB	Folder
Intensities	Nov 8, 2017 at 4:24 PM	504.61 GB	Folder
BaseCalls	Today at 1:53 PM	504.58 GB	Folder
L001	Nov 11, 2017 at 3:59 PM	63.85 GB	Folder
L002	Nov 11, 2017 at 4:00 PM	62.14 GB	Folder
L003	Nov 11, 2017 at 4:00 PM	63.05 GB	Folder
L004	Nov 11, 2017 at 4:01 PM	60.06 GB	Folder
L005	Nov 11, 2017 at 4:01 PM	60.39 GB	Folder
L006	Nov 11, 2017 at 4:02 PM	64.6 GB	Folder
L007	Nov 11, 2017 at 4:02 PM	65.33 GB	Folder
L008	Nov 11, 2017 at 4:02 PM	65.15 GB	Folder
s.locs	Nov 8, 2017 at 3:42 PM	34.5 MB	Document
InterOp	Nov 12, 2017 at 2:04 AM	308.9 MB	Folder
Logs	Nov 11, 2017 at 4:19 PM	1.5 GB	Folder
PeriodicSaveRates	Nov 8, 2017 at 3:34 PM	658 bytes	Folder
Recipe	Nov 8, 2017 at 3:34 PM	26 KB	Folder
RTAComplete.txt	Nov 11, 2017 at 4:09 PM	46 bytes	Plain Text
RTAConfiguration.xml	Nov 11, 2017 at 4:09 PM	7 KB	XML
RTALogs	Nov 11, 2017 at 4:09 PM	356.2 MB	Folder
RTARead1Complete.txt	Nov 11, 2017 at 4:09 PM	37 bytes	Plain Text
RTARead2Complete.txt	Nov 11, 2017 at 4:09 PM	37 bytes	Plain Text
RTARead3Complete.txt	Nov 11, 2017 at 4:09 PM	38 bytes	Plain Text
RTARead4Complete.txt	Nov 11, 2017 at 4:09 PM	37 bytes	Plain Text
RunInfo.xml	Nov 11, 2017 at 4:09 PM	29 KB	XML
runParameters.xml	Nov 11, 2017 at 4:09 PM	5 KB	XML
SequencingComplete.txt	Nov 11, 2017 at 4:19 PM	62 bytes	Plain Text
Unaligned	Today at 1:44 PM	2.61 TB	Folder
Unaligned_mm2	Nov 13, 2017 at 10:12 AM	1.95 TB	Folder

The Run Folders

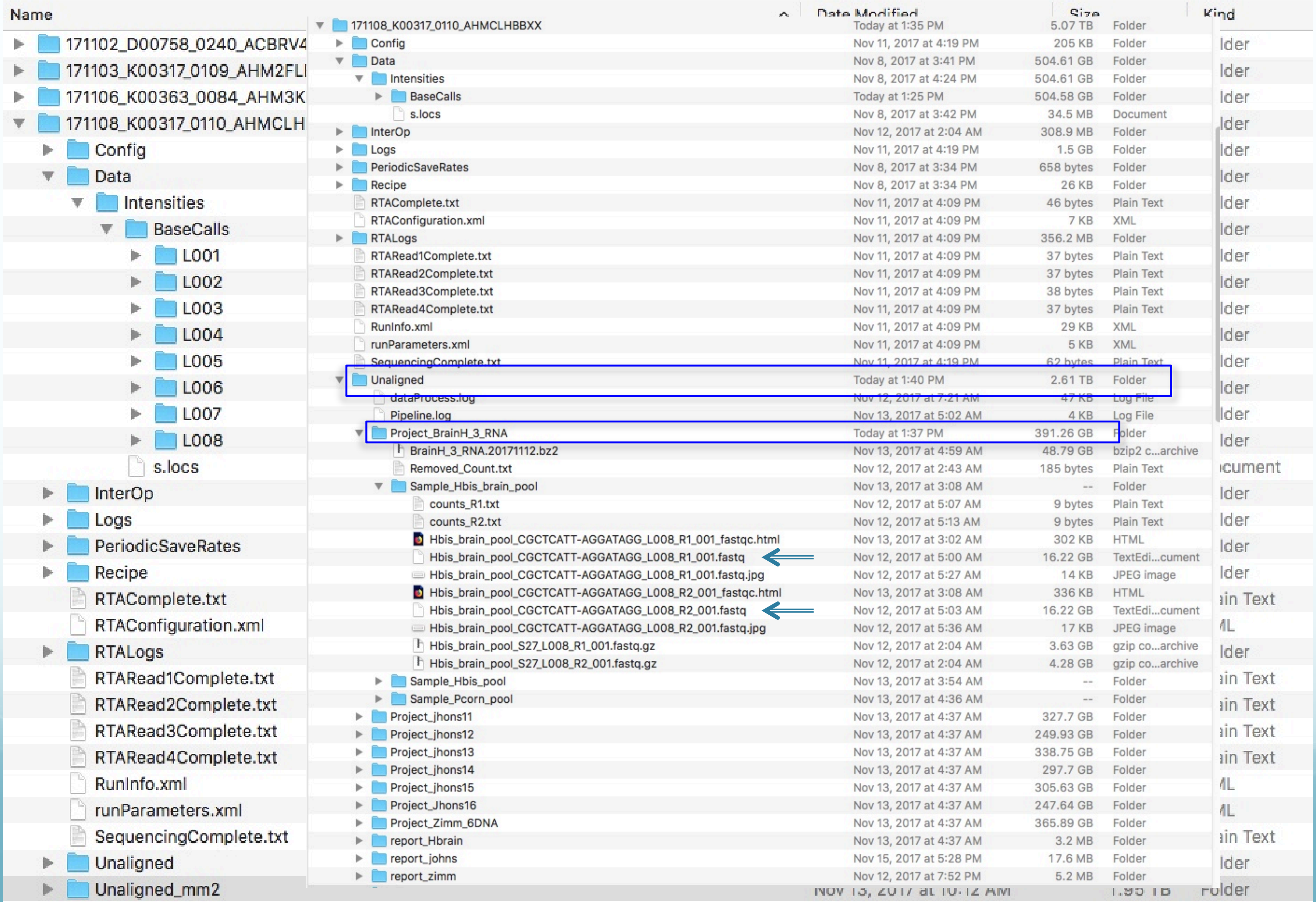


Name	Date Modified	Size	Kind
171102_D00758_0240_ACBRV4ANXX_Baruch_Shcherbo_V4	Dec 5, 2017 at 2:15 PM	1.81 TB	Folder
171103_K00317_0109_AHM2FLBBXX_Barrang...udson_Johnston_Shcherbo_Relman_Johnson	Jan 11, 2018 at 12:30 PM	3.22 TB	Folder
171106_K00363_0084_AHM3KMBBXX_Johnson	Dec 5, 2017 at 2:53 PM	3.14 TB	Folder
171108_K00317_0110_AHMCLHBBXX_Johnson_Zimmerman_Cheng	Dec 5, 2017 at 3:14 PM	5.07 TB	Folder
Config	Nov 11, 2017 at 4:19 PM	205 KB	Folder
Data	Nov 8, 2017 at 3:41 PM	504.74 GB	Folder
Intensities	Nov 8, 2017 at 4:24 PM	504.74 GB	Folder
BaseCalls	Today at 1:25 PM	504.71 GB	Folder
L001	Nov 11, 2017 at 3:59 PM	63.85 GB	Folder
C1.1	Nov 8, 2017 at 4:29 PM	--	Folder
s_1_1101.bcl.gz	Nov 8, 2017 at 4:24 PM	1.5 MB	gzip c
s_1_1102.bcl.gz	Nov 8, 2017 at 4:24 PM	1.5 MB	gzip c
s_1_1103.bcl.gz	Nov 8, 2017 at 4:24 PM	1.5 MB	gzip c
s_1_1104.bcl.gz	Nov 8, 2017 at 4:24 PM	1.5 MB	gzip c
s_1_1105.bcl.gz	Nov 8, 2017 at 4:24 PM	1.5 MB	gzip c
s_1_1106.bcl.gz	Nov 8, 2017 at 4:24 PM	1.5 MB	gzip c
s_1_1107.bcl.gz	Nov 8, 2017 at 4:24 PM	1.5 MB	gzip c
s_1_1108.bcl.gz	Nov 8, 2017 at 4:24 PM	1.5 MB	gzip c
s_1_1109.bcl.gz	Nov 8, 2017 at 4:24 PM	1.5 MB	gzip c
s_1_1110.bcl.gz	Nov 8, 2017 at 4:24 PM	1.5 MB	gzip c
s_1_1111.bcl.gz	Nov 8, 2017 at 4:24 PM	1.5 MB	gzip c
s_1_1112.bcl.gz	Nov 8, 2017 at 4:24 PM	1.5 MB	gzip c
s_1_1113.bcl.gz	Nov 8, 2017 at 4:24 PM	1.5 MB	gzip c
s_1_1114.bcl.gz	Nov 8, 2017 at 4:24 PM	1.5 MB	gzip c
s_1_1115.bcl.gz	Nov 8, 2017 at 4:24 PM	1.5 MB	gzip c
s_1_1116.bcl.gz	Nov 8, 2017 at 4:24 PM	1.5 MB	gzip c
s_1_1117.bcl.gz	Nov 8, 2017 at 4:24 PM	1.5 MB	gzip c

Name	Date Modified	Size	Kind
171102_D00758_0240_			
171103_K00317_0109_			
171106_K00363_0084_			
171108_K00317_0110_A			
Config			
Data			
Intensities			
BaseCalls			
L001			
L002			
L003			
L004			
L005			
L006			
L007			
L008			
s.locs			
InterOp			
Logs			
PeriodicSaveRates			
Recipe			
RTAComplete.txt			
RTAConfiguration.			
RTALogs			
RTARead1Complet			
RTARead2Complet			
RTARead3Complet			
RTARead4Complet			
RunInfo.xml	Nov 11, 2017 at 4:09 PM	29 KB	XML
runParameters.xml	Nov 11, 2017 at 4:09 PM	5 KB	XML
SequencingComplete.txt	Nov 11, 2017 at 4:19 PM	62 bytes	Plain Text
Unaligned	Today at 1:44 PM	2.61 TB	Folder
Unaligned_mm2	Nov 13, 2017 at 10:12 AM	1.95 TB	Folder

112 bcl files per cycle
318 cycles
8 lanes
= 284,928 files

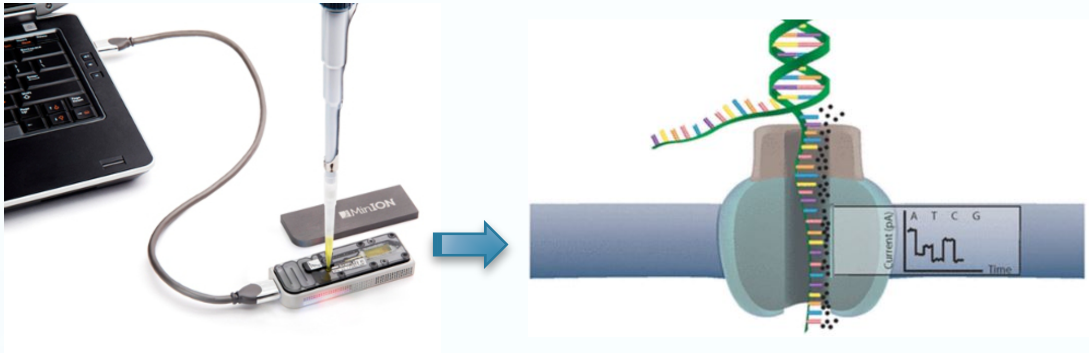
The Run Folders



- 171102_D00758_0240_ACBRV4
- 171103_K00317_0109_AHM2FLI
- 171106_K00363_0084_AHM3K
- 171108_K00317_0110_AHMCLH
  - Config
  - Data
    - Intensities
      - BaseCalls
        - L001
        - L002
        - L003
        - L004
        - L005
        - L006
        - L007
        - L008
      - s.locs
  - InterOp
  - Logs
  - PeriodicSaveRates
  - Recipe
  - RTAComplete.txt
  - RTAConfiguration.xml
  - RTALogs
  - RTARead1Complete.txt
  - RTARead2Complete.txt
  - RTARead3Complete.txt
  - RTARead4Complete.txt
  - RunInfo.xml
  - runParameters.xml
  - SequencingComplete.txt
  - Unaligned
  - Unaligned_mm2 — Nov 13, 2017 at 10:12 AM — 1.95 TB — Folder

Name	Date Modified	Size	Kind
171108_K00317_0110_AHMCLHBBXX	Today at 1:35 PM	5.07 TB	Folder
Config	Nov 11, 2017 at 4:19 PM	205 KB	Folder
Data	Nov 8, 2017 at 3:41 PM	504.61 GB	Folder
Intensities	Nov 8, 2017 at 4:24 PM	504.61 GB	Folder
BaseCalls	Today at 1:25 PM	504.58 GB	Folder
s.locs	Nov 8, 2017 at 3:42 PM	34.5 MB	Document
InterOp	Nov 12, 2017 at 2:04 AM	308.9 MB	Folder
Logs	Nov 11, 2017 at 4:19 PM	1.5 GB	Folder
PeriodicSaveRates	Nov 8, 2017 at 3:34 PM	658 bytes	Folder
Recipe	Nov 8, 2017 at 3:34 PM	26 KB	Folder
RTAComplete.txt	Nov 11, 2017 at 4:09 PM	46 bytes	Plain Text
RTAConfiguration.xml	Nov 11, 2017 at 4:09 PM	7 KB	XML
RTALogs	Nov 11, 2017 at 4:09 PM	356.2 MB	Folder
RTARead1Complete.txt	Nov 11, 2017 at 4:09 PM	37 bytes	Plain Text
RTARead2Complete.txt	Nov 11, 2017 at 4:09 PM	37 bytes	Plain Text
RTARead3Complete.txt	Nov 11, 2017 at 4:09 PM	38 bytes	Plain Text
RTARead4Complete.txt	Nov 11, 2017 at 4:09 PM	37 bytes	Plain Text
RunInfo.xml	Nov 11, 2017 at 4:09 PM	29 KB	XML
runParameters.xml	Nov 11, 2017 at 4:09 PM	5 KB	XML
SequencingComplete.txt	Nov 11, 2017 at 4:19 PM	62 bytes	Plain Text
Unaligned	Today at 1:40 PM	2.61 TB	Folder
dataProcess.log	Nov 12, 2017 at 7:21 AM	47 KB	Log File
Pipeline.log	Nov 13, 2017 at 5:02 AM	4 KB	Log File
Project_BrainH_3_RNA	Today at 1:37 PM	391.26 GB	Folder
BrainH_3_RNA.20171112.bz2	Nov 13, 2017 at 4:59 AM	48.79 GB	bzip2 c...archive
Removed_Count.txt	Nov 12, 2017 at 2:43 AM	185 bytes	Plain Text
Sample_Hbis_brain_pool	Nov 13, 2017 at 3:08 AM	--	Folder
counts_R1.txt	Nov 12, 2017 at 5:07 AM	9 bytes	Plain Text
counts_R2.txt	Nov 12, 2017 at 5:13 AM	9 bytes	Plain Text
Hbis_brain_pool_CGCTCATT-AGGATAGG_L008_R1_001_fastqc.html	Nov 13, 2017 at 3:02 AM	302 KB	HTML
Hbis_brain_pool_CGCTCATT-AGGATAGG_L008_R1_001.fastq ⟸	Nov 12, 2017 at 5:00 AM	16.22 GB	TextEdi...cument
Hbis_brain_pool_CGCTCATT-AGGATAGG_L008_R1_001.fastq.jpg	Nov 12, 2017 at 5:27 AM	14 KB	JPEG image
Hbis_brain_pool_CGCTCATT-AGGATAGG_L008_R2_001_fastqc.html	Nov 13, 2017 at 3:08 AM	336 KB	HTML
Hbis_brain_pool_CGCTCATT-AGGATAGG_L008_R2_001.fastq ⟸	Nov 12, 2017 at 5:03 AM	16.22 GB	TextEdi...cument
Hbis_brain_pool_CGCTCATT-AGGATAGG_L008_R2_001.fastq.jpg	Nov 12, 2017 at 5:36 AM	17 KB	JPEG image
Hbis_brain_pool_S27_L008_R1_001.fastq.gz	Nov 12, 2017 at 2:04 AM	3.63 GB	gzip co...archive
Hbis_brain_pool_S27_L008_R2_001.fastq.gz	Nov 12, 2017 at 2:04 AM	4.28 GB	gzip co...archive
Sample_Hbis_pool	Nov 13, 2017 at 3:54 AM	--	Folder
Sample_Pcorn_pool	Nov 13, 2017 at 4:36 AM	--	Folder
Project_jhons11	Nov 13, 2017 at 4:37 AM	327.7 GB	Folder
Project_jhons12	Nov 13, 2017 at 4:37 AM	249.93 GB	Folder
Project_jhons13	Nov 13, 2017 at 4:37 AM	338.75 GB	Folder
Project_jhons14	Nov 13, 2017 at 4:37 AM	297.7 GB	Folder
Project_jhons15	Nov 13, 2017 at 4:37 AM	305.63 GB	Folder
Project_Jhons16	Nov 13, 2017 at 4:37 AM	247.64 GB	Folder
Project_Zimm_6DNA	Nov 13, 2017 at 4:37 AM	365.89 GB	Folder
report_Hbrain	Nov 13, 2017 at 4:37 AM	3.2 MB	Folder
report_johns	Nov 15, 2017 at 5:28 PM	17.6 MB	Folder
report_zimm	Nov 12, 2017 at 7:52 PM	5.2 MB	Folder

Oxford Nanopore GridION Runs



- Each read has its own fast5 file (HDF5 format)
- Each run can have from a few hundreds to several million reads ➔ millions of files
- Each fast5 file can be from 80KB to 10 MB
- Runs with > 1 million long reads ➔ > 1 TB of fast5 files, sometimes > 5 TB
  ~ 50 GB of fastq files
- For an eukaryotic genome assembly= may need > 10 runs

Final Considerations

- DNA sequencing market is unlimited, the amount of data generated is a tiny portion of what will be generated

- Large investment in resources to produce data, store it, make it available

- The majority of these files are often multiplied over 4 times:

  in production storage
  posted
  archived
  downloaded by end user, once, twice?
  downloaded by the bioinformaticians

- Analysis files (BAM, SAM, intermediary files), are at least 5x, generally 10x larger than fastq data ➔ BI facility has a bigger storage problem + end user wants the data so now it grows even more.

- x 100’s of sequencing facilities across the world

Thank you for listening!

And thank you to my group: