leidos

Leidos Biomedical Research, Inc.

CCR-SF Data Storage, Archive and Meta Data

Yongmei Zhao

CCR-SF Bioinformatics Group/ABCC

April 13, 2015

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute

Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

CCR-Sequencing Facility (SF)

- The **Center for Cancer Research (CCR) Sequencing Facility (SF)** provided investigators with access to Illumina and PacBio® sequencers
- Currently operate 2 HiSeq 2500, 1 Hiseq 2000, 2 NextSeq500 and I Miseq illumina platforms and 1 PacBio RSII.

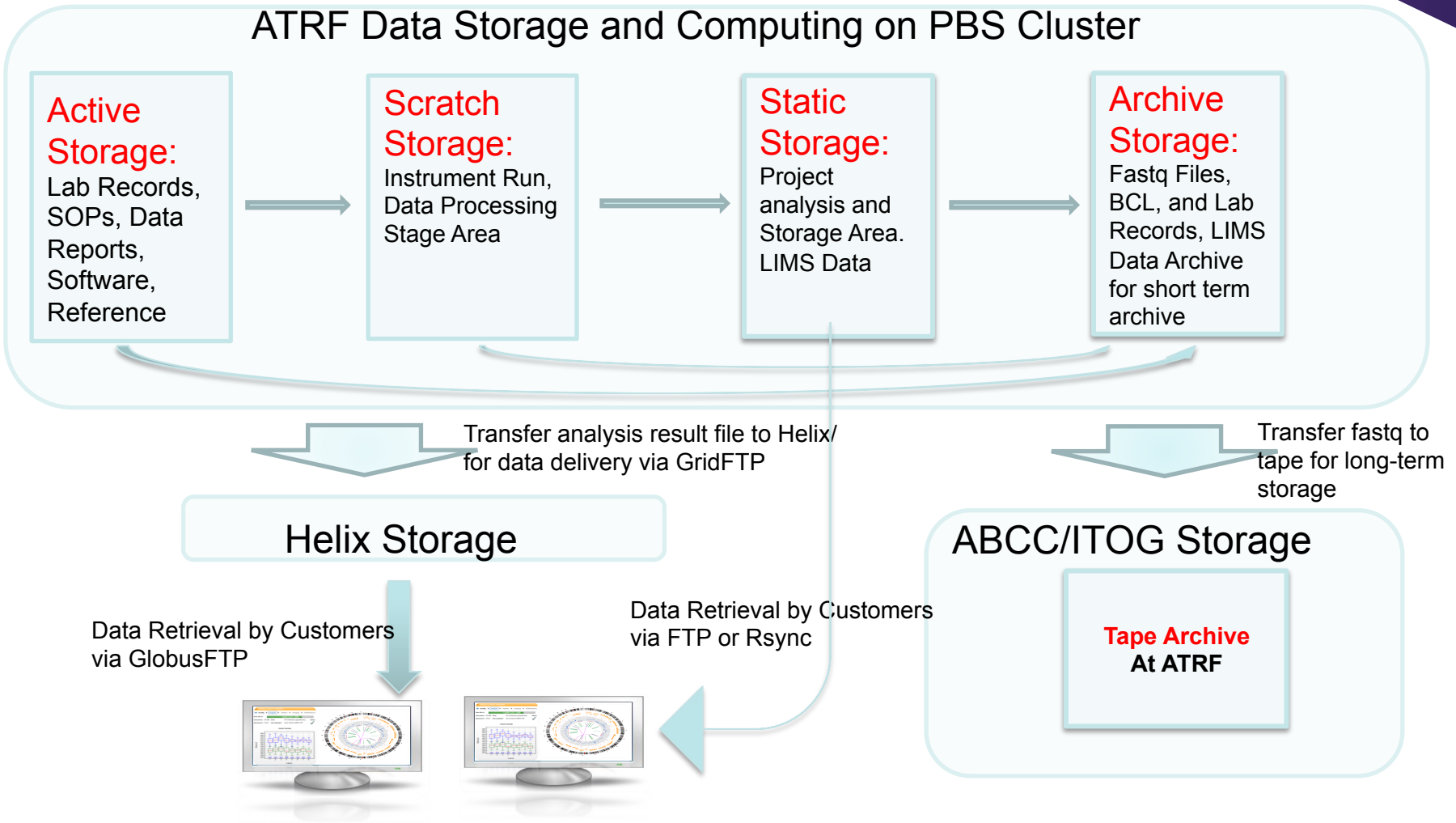
Sequencing Throughput By Instrument Types

Instrument Type	Run Protocol	Sequence Run Time	Monthly Maximum Runs	Estimated Yield in Trillion bases (Dual Flowcell for Hiseq)	Monthly Yield (Trillion bases)
HiSeq 2500	2x100 (SBS V4)	5 days	5	700 - 800Gb	3.5 - 4Tb
(high output mode)	2x125 (SBS V4)	6 days	4	900 -1000Gb	3.6 - 4Tb
	2x100 (SBS V4)	27hr	16	100 -120Gb	1.6 - 1.9Tb
	2x125 (SBS V4)	40hr	9	150 -180Gb	1.3 - 1.8Tb
NextSeq500	2x150	29hr	14	100-120Gb	1.4 -1.6Tb
(high output mode)	2 × 150 bp	26 hrs	18	32.5-39 Gb	0.5-0.7Tb
	2 × 75 bp	15 hrs	28	16.25-19.5 Gb	0.4 - 0.6Tb
HiSeq2000	2x100 bp	8 days	3	550 -660Gb	1.5 - 1.8Tb
	1x50 bp	2 days	14	35 -50Gb	0.5 -0.7Tb
GAIIx	1x36 bp	2 days	9	20-30Gb	0.2 - 0.3Tb
Miseq	1x36, 2x75, 2x150, 2x300	4 hrs - 39 hrs	15	3 - 15Gb	~0.1Tb

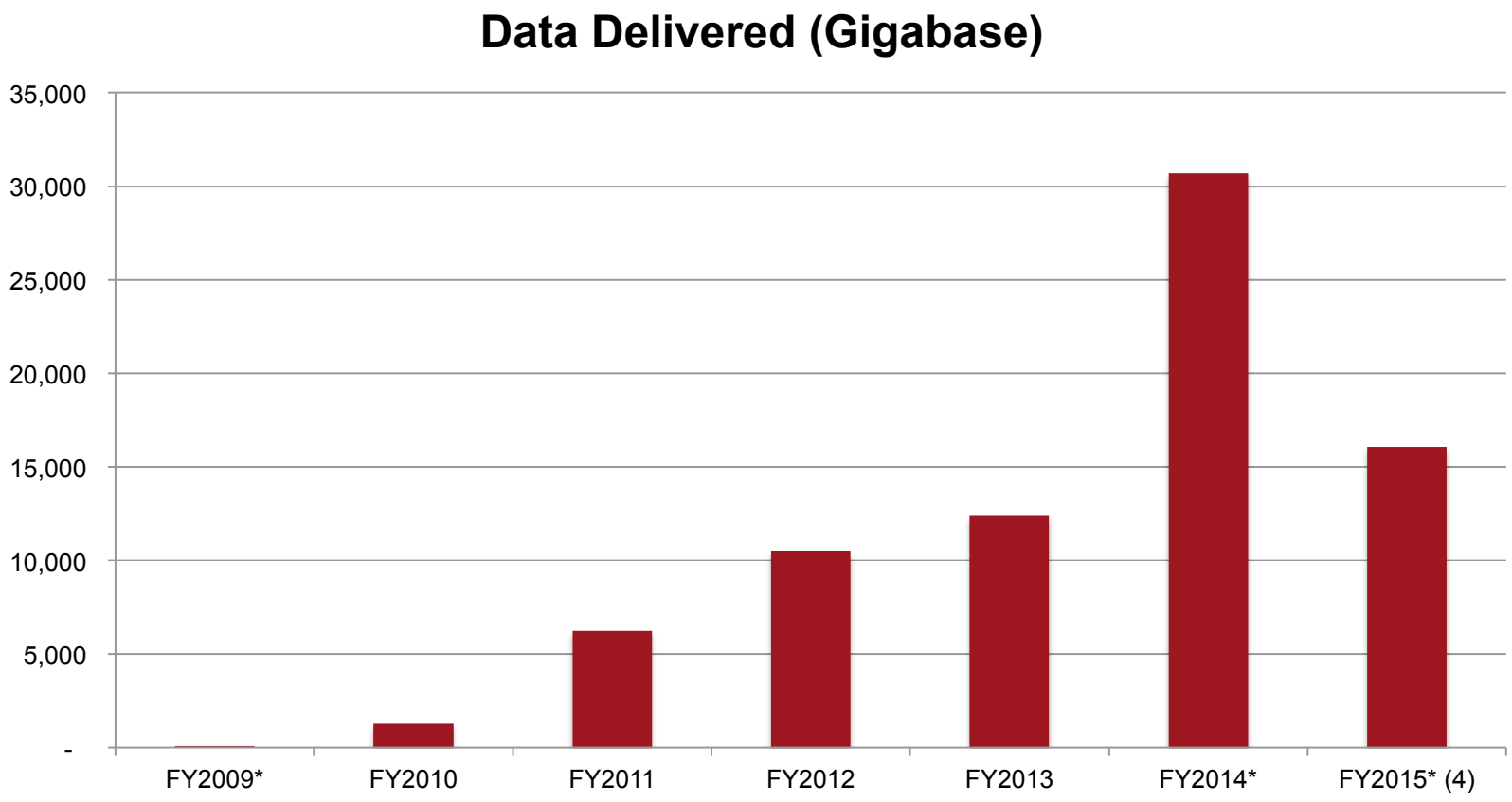
CCR-SF Current Storage Categories

- Active – tier 2 Isilon storage
  - data is regularly changing, but must keep for long term. Need change protection.
- Scratch – tier 3 Isilon storage
  - data is highly volatile and regularly changing, only keep for short term storage. Need low I/O latency for instrument output data and computing.
- Static – tier 3 Isilon storage
  - Data is regularly changing, but keep for short – medium term storage. Need low I/O latency computing on PBS clusters.
- Archive –tier 3 Isilon storage
  - Data is static and stored for medium term. Can be relative high I/O latency. Data periodically transfer to tape archive
- Tape Archive
  - Large scale long term storage for backup archives from instrument run and analysis results.

Current CCR-SF NGS Data Flow

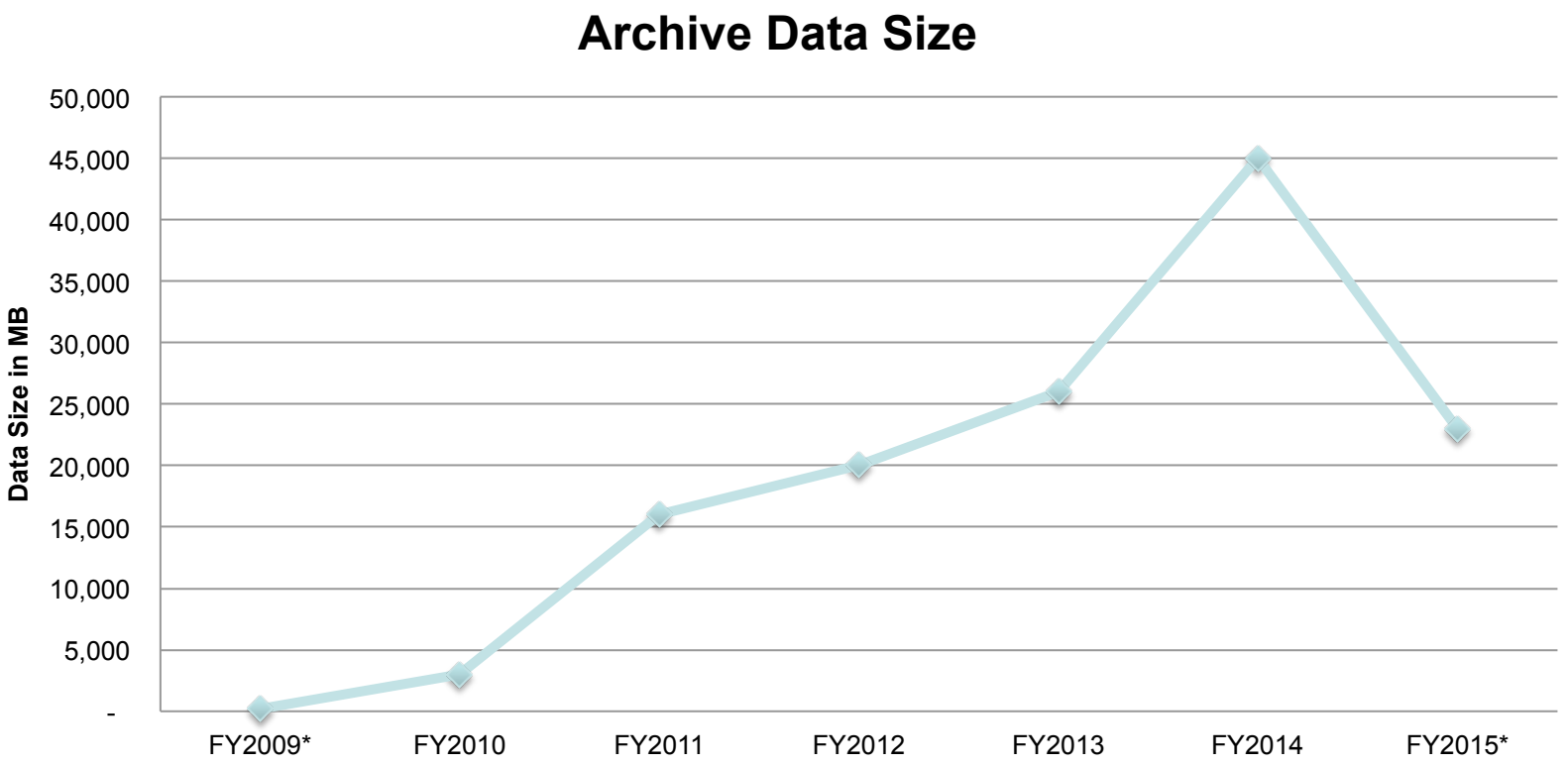


Illumina Platform Production Statistics



FY15 only includes 6 month data

FASTQ Archive Data Size



FY15 only includes 6 month data

Current and Future Archive Storage Estimate

Operation Status	Instrument	Average Monthly Archive Storage Usage (fastq only, 2014) (TB)	Average Monthly Storage Usage for FASTQ and BAM Files (TB)	Total Archive Storage Usage for FY2014 (fastq only) (TB)	Predicted Total Archive Storage Usage for FY2015 (fastq only) (TB)
Current Operation	1 HiSeq2000 2 HiSeq2500 2 NextSeq500	4 TB	8TB	30TB – 40TB	40TB – 50TB
Future	2 HiSeq2500 1 HiSeq2000 2 NextSeq500 Other	4-6 TB	8-12TB	N/A	40TB – 60TB

Archive Data Types

- Raw Fastq files
- Alignment BAM files
- MD5SUM of fastq and BAM files
- Run meta data

Meta Data for SF Archive

- Lab information (Investigator username, lab contact)
- Project information (Project ID – CSAS, Project description)
- Study information
- Platform
- Sample Name
- Sequencing Flowcell ID (unique ID for illumina run)
- Run Date
- Sequencing application type – mRNA pair-end run, chip-seq single-end run, etc.
- Ref genome and annotation (if store BAM files)
- Software version and parameters
- Sequencing center

Data Submission to GEO and SRA

- GEO repository accepted NGS data types:
  - mRNA profiling (NOT transcriptome or transcript assemblies)
  - small RNA profiling
  - ChIP-Seq
  - methyl-Seq
  - bisulfite sequencing
  - digital gene expression tag profiling
  - traditional SAGE
- SRA accepts the following NGS data types:
  - Raw sequence data and alignment BAM files from NGS technologies including 454, IonTorrent, Illumina, SOLiD, Helicos, PacBio and Complete Genomics.
  - Sequencing application include Whole genome sequencing, resequencing, whole exome sequencing, Metegenomics, transcriptome or transcriptome assemblies, etc.

Meta Data for Support Data Submission

- GEO Data Submission
  - Platform and application specific template
  - Meta data includes overall experiment, samples and protocol information
- SRA Data Submission:
  - BioProject – project meta data
  - BioSample – sample meta data
  - SRA meta data – seq protocol, platform, library protocol, data analysis pipeline information, etc.
  - Datafiles:
    - BAM, SFF, PacBio Hdf5, fastq