Taking HEP data management outside of HEP

Ian Fisk, Robert Illingworth, **Bo Jayatilaka**

CHEP 2016, San Francisco

11 October 2016

Challenges of large datasets

- The rapid growth in data acquisition rates has presented the community with new challenges
- Datasets are growing rapidly to petabyte scale and beyond
- Traditional data management methods (lists of paths on a filesystem) don’t scale well
- Large datasets require organized processing workflows
  - Use the cloud as well as/instead of traditional resources
- These issues are familiar ones in high energy physics

Data Management Requirements

- There needs to be an efficient method of classifying, storing and retrieving large amounts of data

- Needs to be:
  - Capable of handling ANY type of data
  - Compatible with distributed analysis and analytics gathering computing models
  - Compatible with arbitrary data stores
  - Scalable (many millions of data files)
  - Optimize access to the data

These problems are not specific to any particular field

Fermilab Storage System and the Active Archive Facility

- Fermilab currently has 100PB of HEP data in active tape storage
  - Uses Enstore, a Fermilab-developed system
  - Access is via ~30PB of disk-based caching
    - Uses dCache (DESY+international collaboration)

- **Active Archive Facility** (AAF)
  - Leverage existing Enstore/dCache based system to provide archival storage for non-HEP scientific use

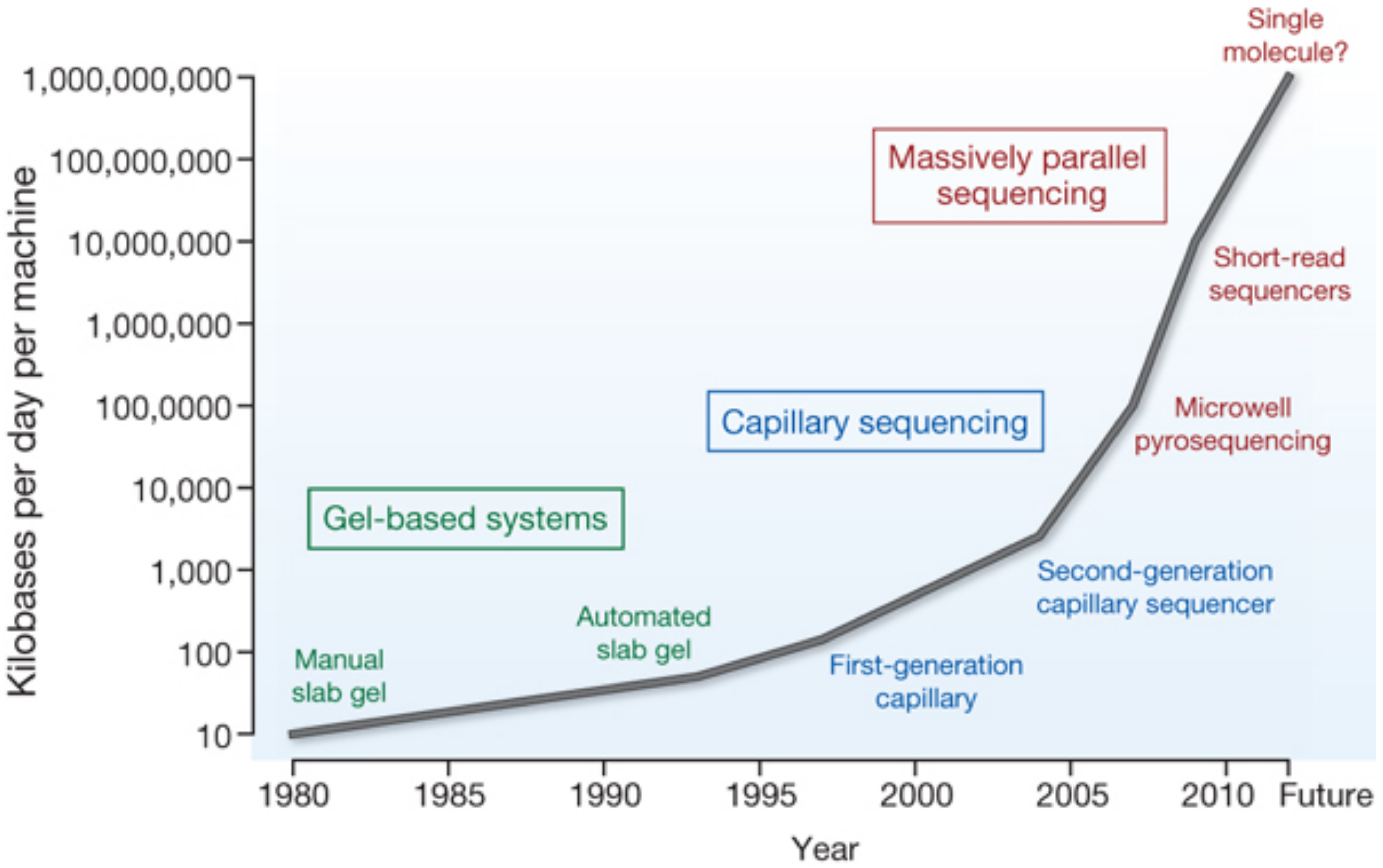
- SFARI is storing data in the AAF
  - Currently 1.2 PB of genome data archived

SFARI

- Simons Foundation Autism Research Initiative (SFARI)
  - https://sfari.org/
  - Improve the understanding, diagnosis and treatment of autism spectrum disorders by funding innovative research of the highest quality and relevance.
  - Understand genetic and environmental factors that lead to autism spectrum disorder (ASD)

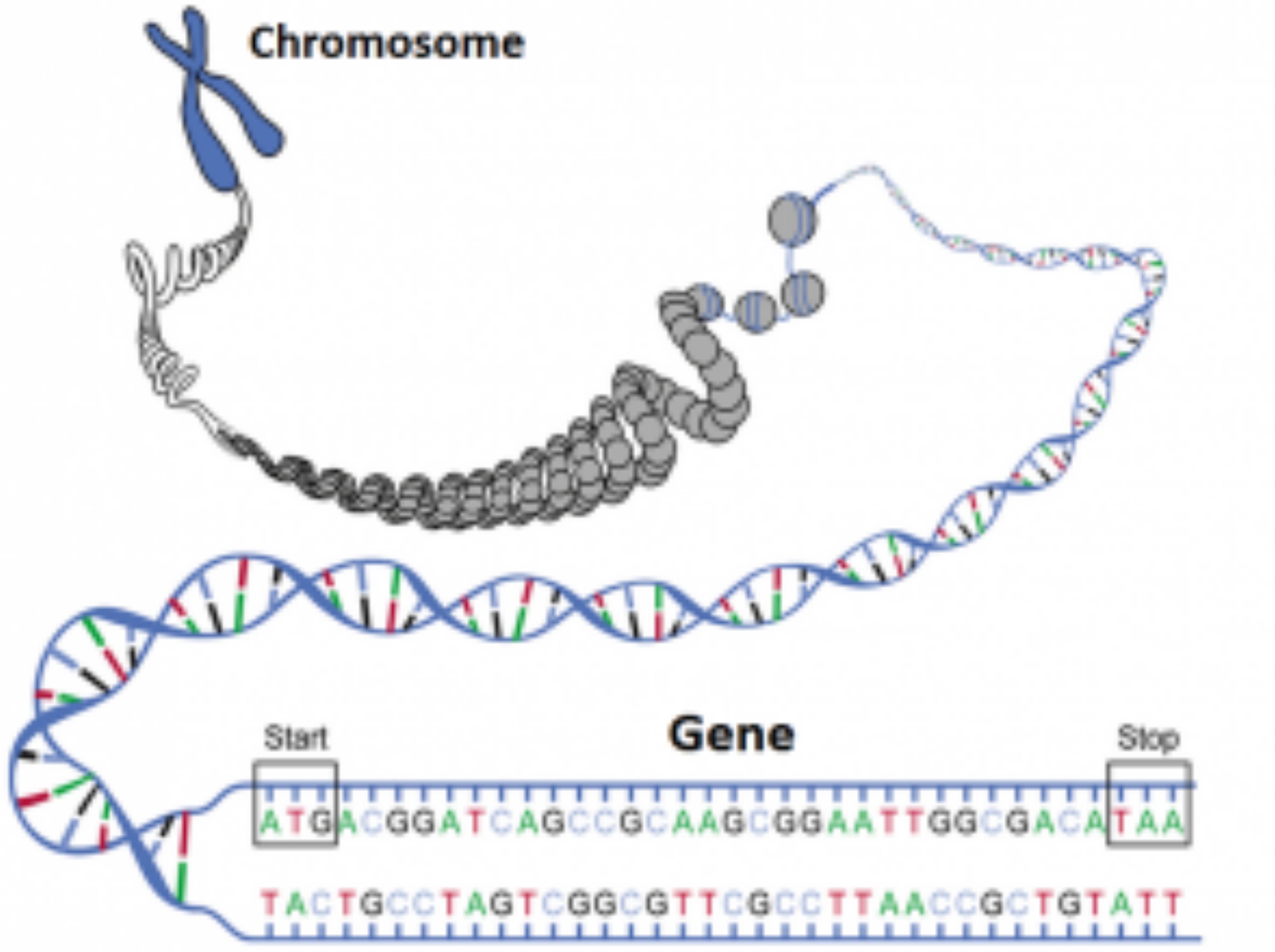
SFARI SIMONS FOUNDATION
AUTISM RESEARCH INITIATIVE

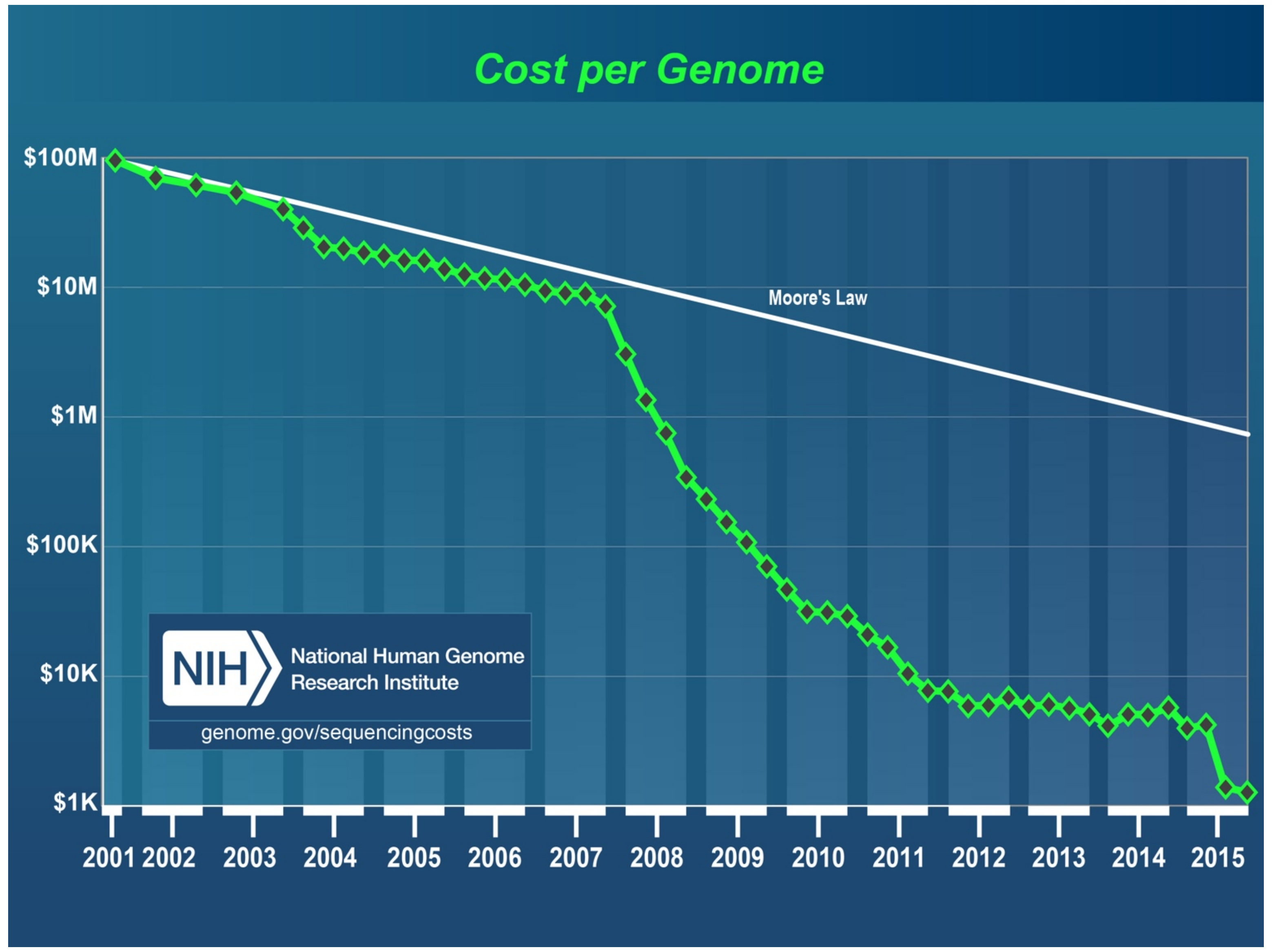
Genomics data



Improvements in the rate of DNA sequencing over the past 30 years and into the future.

MR Stratton *et al. Nature* **458**, 719-724 (2009) doi:10.1038/nature07943


Chromosome
Gene
Start
Stop
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAA
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT

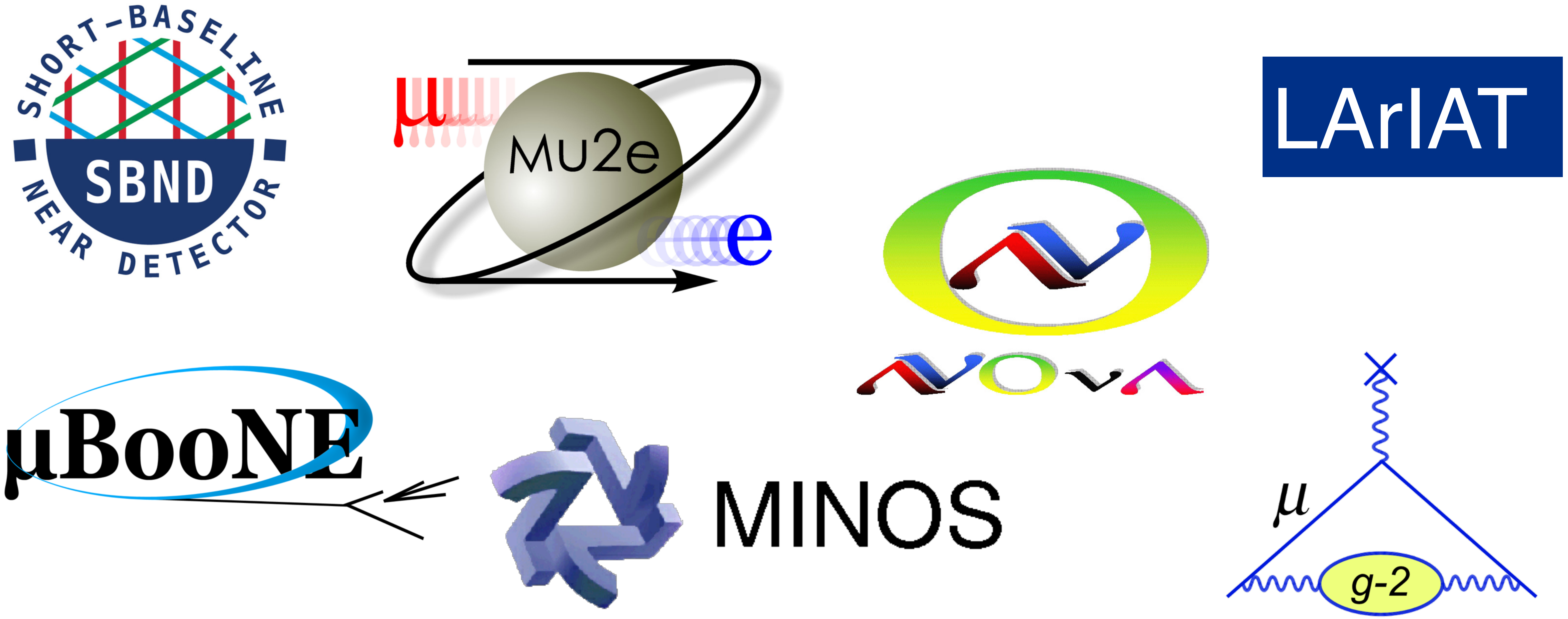


SFARI datasets

- SFARI has created and supports several resources for autism scientists
  - Simons Simplex Collection (SSC), which contains extensive genetic and phenotypic data from nearly 3,000 families with a child affected by autism
  - Simons Variation in Individuals Project (Simons VIP), which aims to identify and study large numbers of individuals sharing recurrent genetic variants known to increase the risk of developing autism spectrum and other neurodevelopmental disorders.
  - SPARK, an ongoing online research initiative that aims to recruit, engage and retain a community of 50,000 individuals with autism and their family members living in the U.S.
    - Total 2 PB over 5 years (assuming current technology)

Generic HEP data management at Fermilab

- Many Fermilab experiments are using the SAM data management system
  - Originally developed for Run II of the Tevatron
  - Now used by a dozen Intensity Frontier experiments

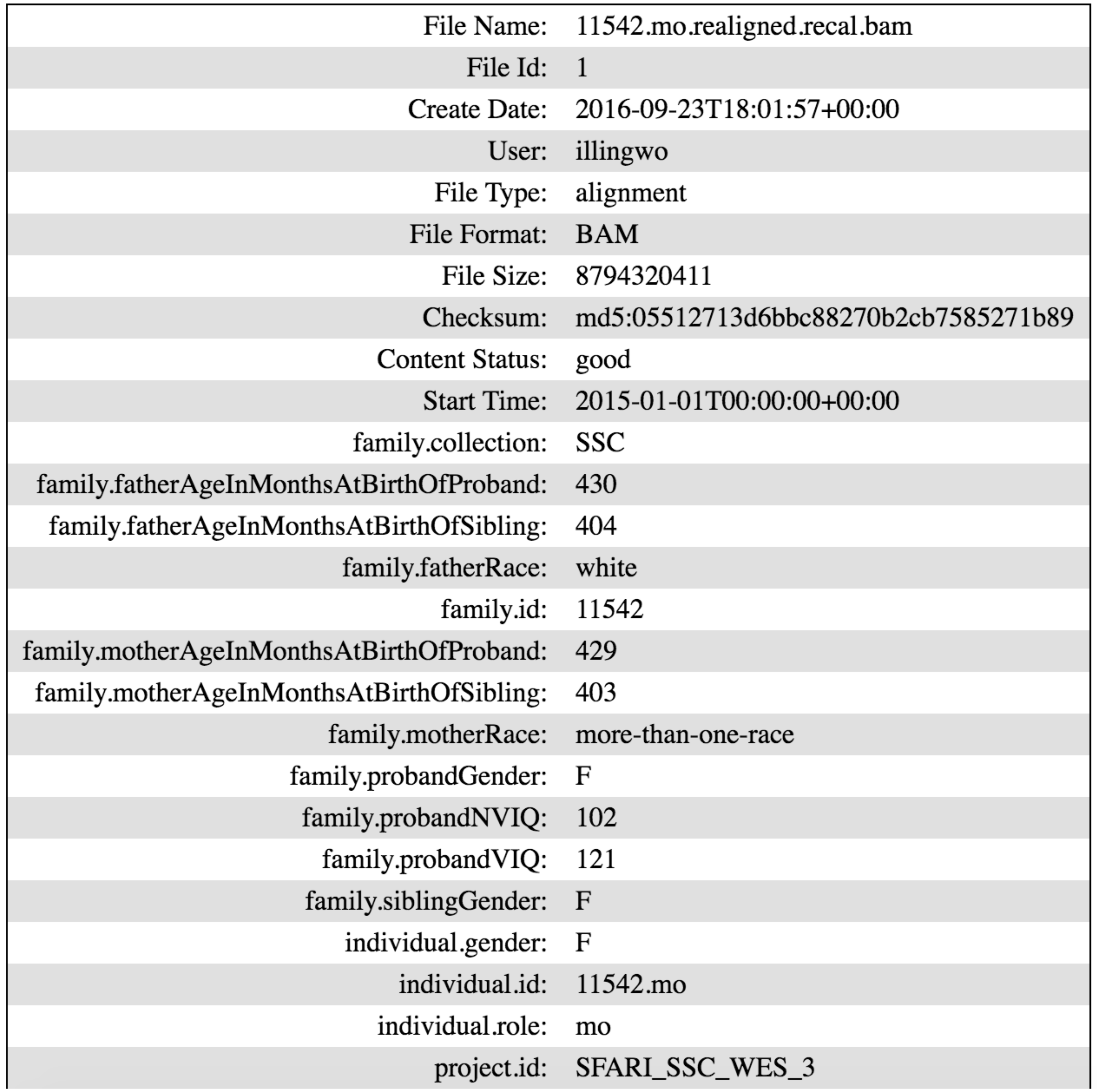


SAM Concepts

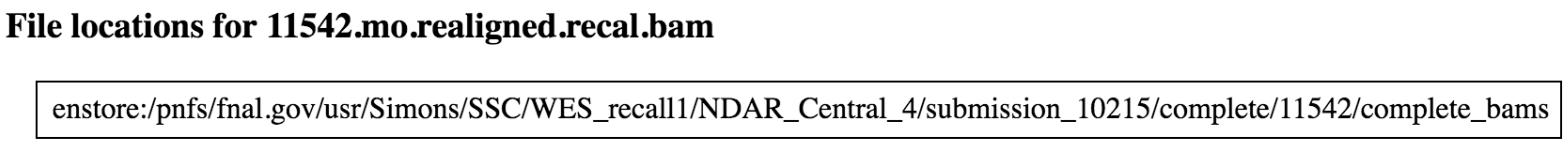
- Metadata based catalog system
  - Describe data files
  - User defined fields allow attaching arbitrary values
- Replica Catalog
  - Track the location of data files in an abstract, non-storage, specific way
  - Abstract identifiers can be mapped to access URLs to retrieve data
- Framework for organizing analysis dataflow
  - Provides a generic method for organizing access to data as part of a processing workflow

Application of SAM to genomics data

- Metadata about files
  - Includes information about individuals and families
- File locations in AAF



File Name:	11542.mo.realigned.recal.bam
File Id:	1
Create Date:	2016-09-23T18:01:57+00:00
User:	illingwo
File Type:	alignment
File Format:	BAM
File Size:	8794320411
Checksum:	md5:05512713d6bbc88270b2cb7585271b89
Content Status:	good
Start Time:	2015-01-01T00:00:00+00:00
family.collection:	SSC
family.fatherAgeInMonthsAtBirthOfProband:	430
family.fatherAgeInMonthsAtBirthOfSibling:	404
family.fatherRace:	white
family.id:	11542
family.motherAgeInMonthsAtBirthOfProband:	429
family.motherAgeInMonthsAtBirthOfSibling:	403
family.motherRace:	more-than-one-race
family.probandGender:	F
family.probandNVIQ:	102
family.probandVIQ:	121
family.siblingGender:	F
individual.gender:	F
individual.id:	11542.mo
individual.role:	mo
project.id:	SFARI_SSC_WES_3



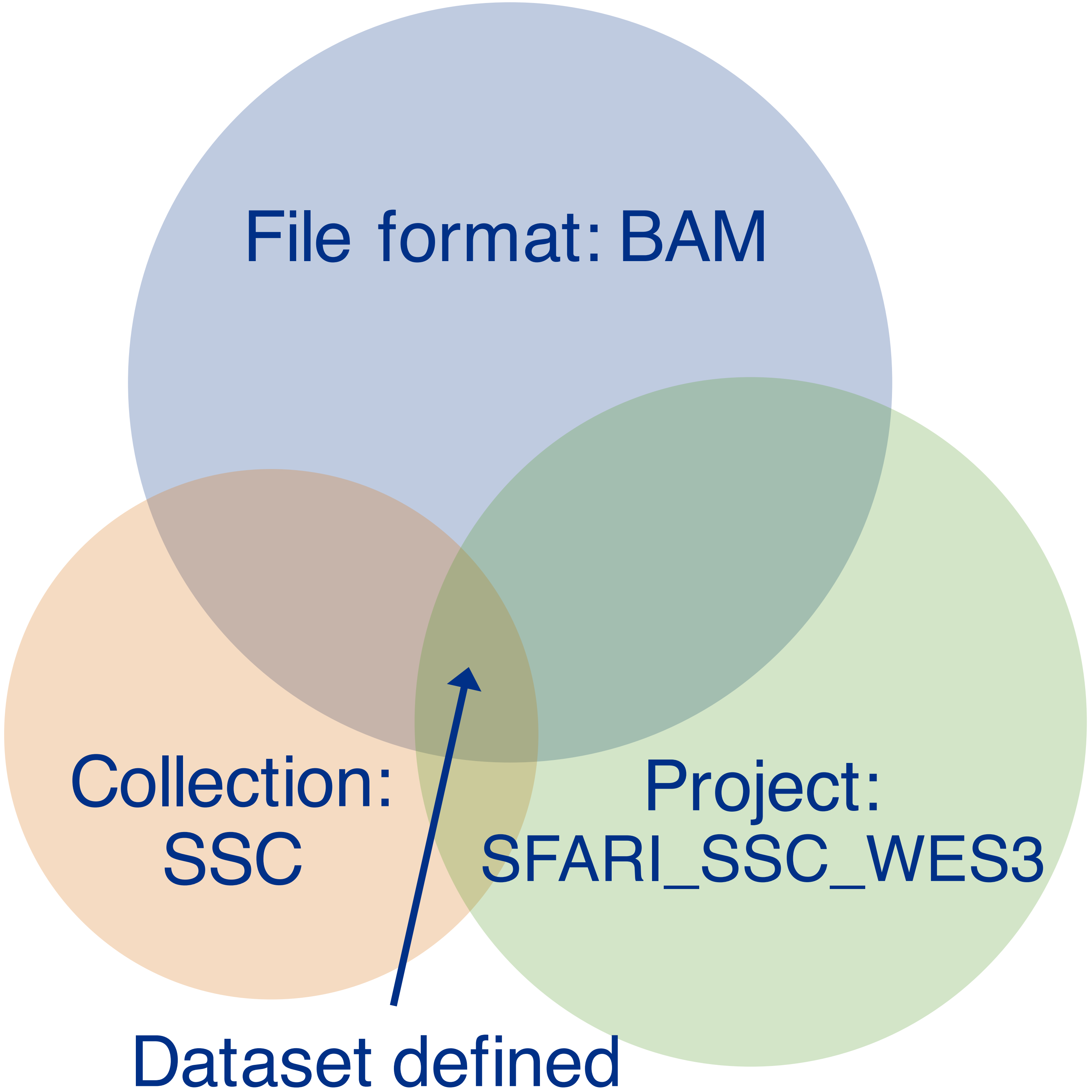
File locations for 11542.mo.realigned.recal.bam

enstore:/pnfs/fnal.gov/usr/Simons/SSC/WES_recall1/NDAR_Central_4/submission_10215/complete/11542/complete_bams

Making datasets

- SAM defines datasets in terms of metadata queries
  - For example “all BAM files from collection SSC and project SFARI_SSC_WES3”

- This gives great flexibility in creating datasets for different purposes
  - Not forced into a particular concept of a dataset



Dataset defined as all files matching combination of criteria

Future plans

- Expand replica location catalogue to data stores beyond Fermilab Active Archive
- Implement automatic movement of datasets between data stores
- Integrate with processing workflows

Summary

- Genomics data production rate is comparable to HEP experiments
- Many of the data management requirements – metadata cataloging, locating files in multiple data stores, defining datasets and processing the files within them – are similar
- An HEP data management system that is generic enough to handle multiple experiments is generic enough to handle data from outside of HEP
- We have started to apply this to data from SFARI