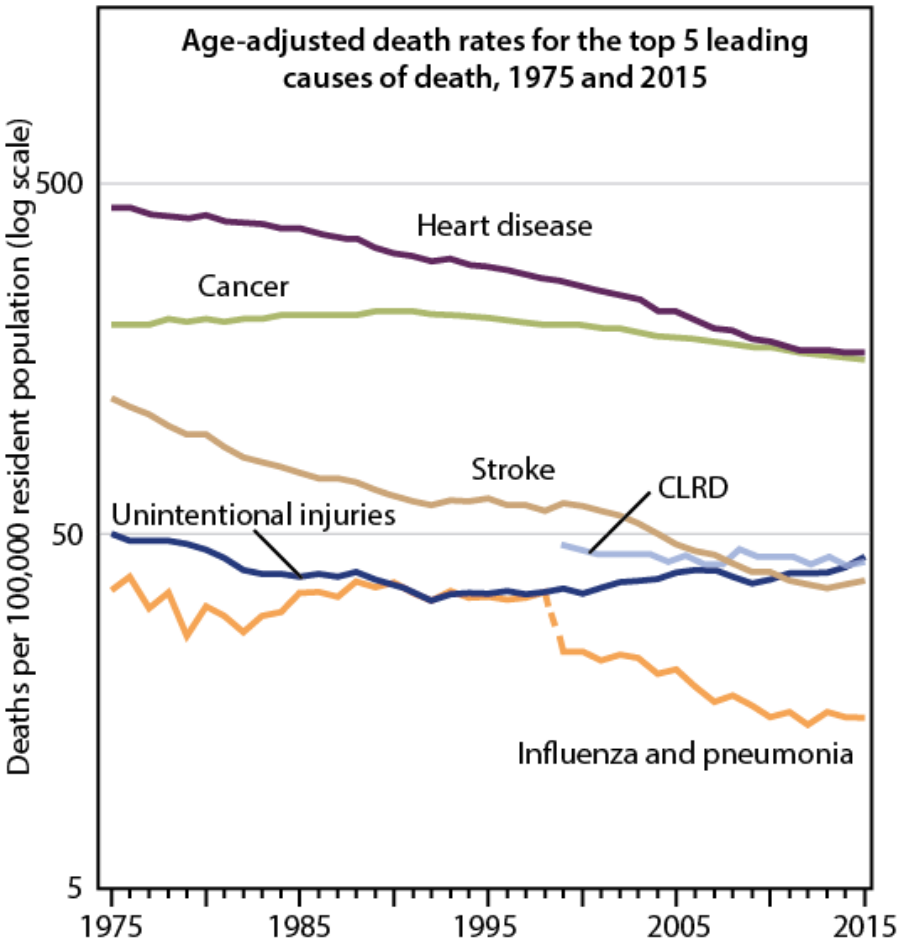
# Advancing the Big Data Genomics Analysis using Cloud

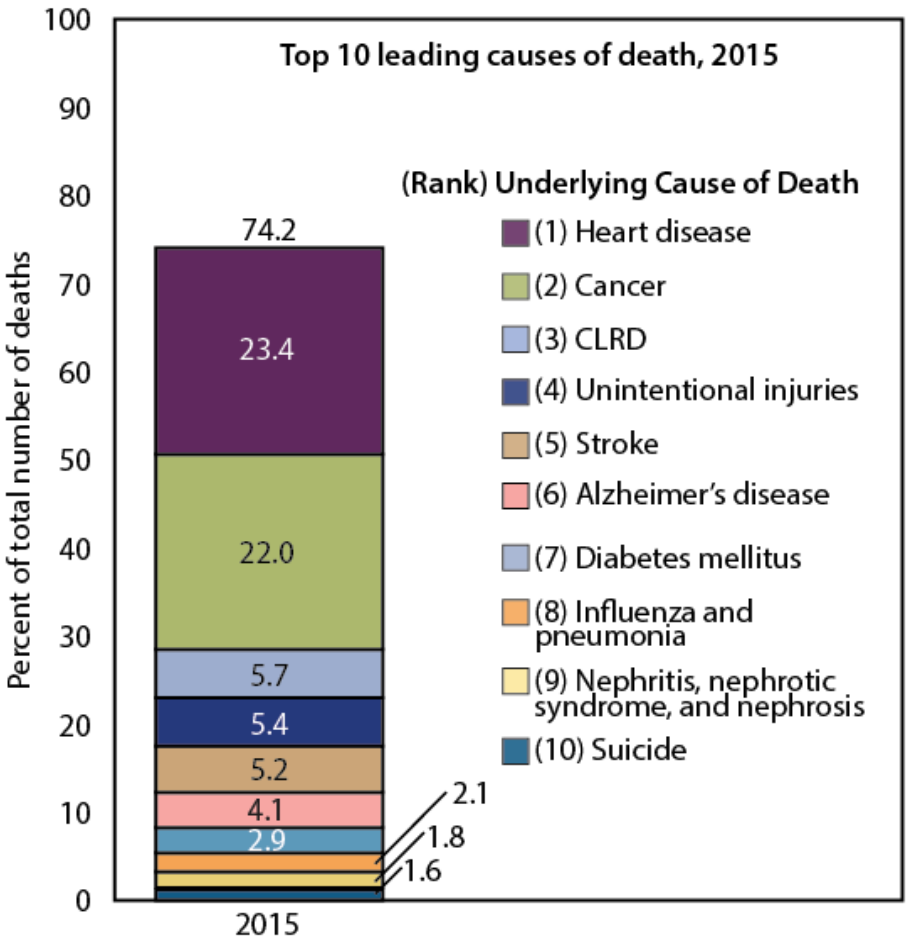
Segun C. Jung

Biomedical Informatics Consultant

University of Chicago & Argonne National Laboratory

# Leading causes of death is the US



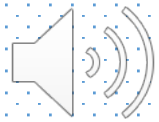




NOTE: Due to coding changes for chronic lower respiratory diseases (CLRD) between ICD–9 and ICD–10, which prevent the direct comparison of trends prior to 1998 and after 1999, rates for CLRD are only shown for 1999 onwards.

SOURCE: NCHS, *Health, United States, 2016*, Figure 8. Data from the National Vital Statistics System (NVSS).

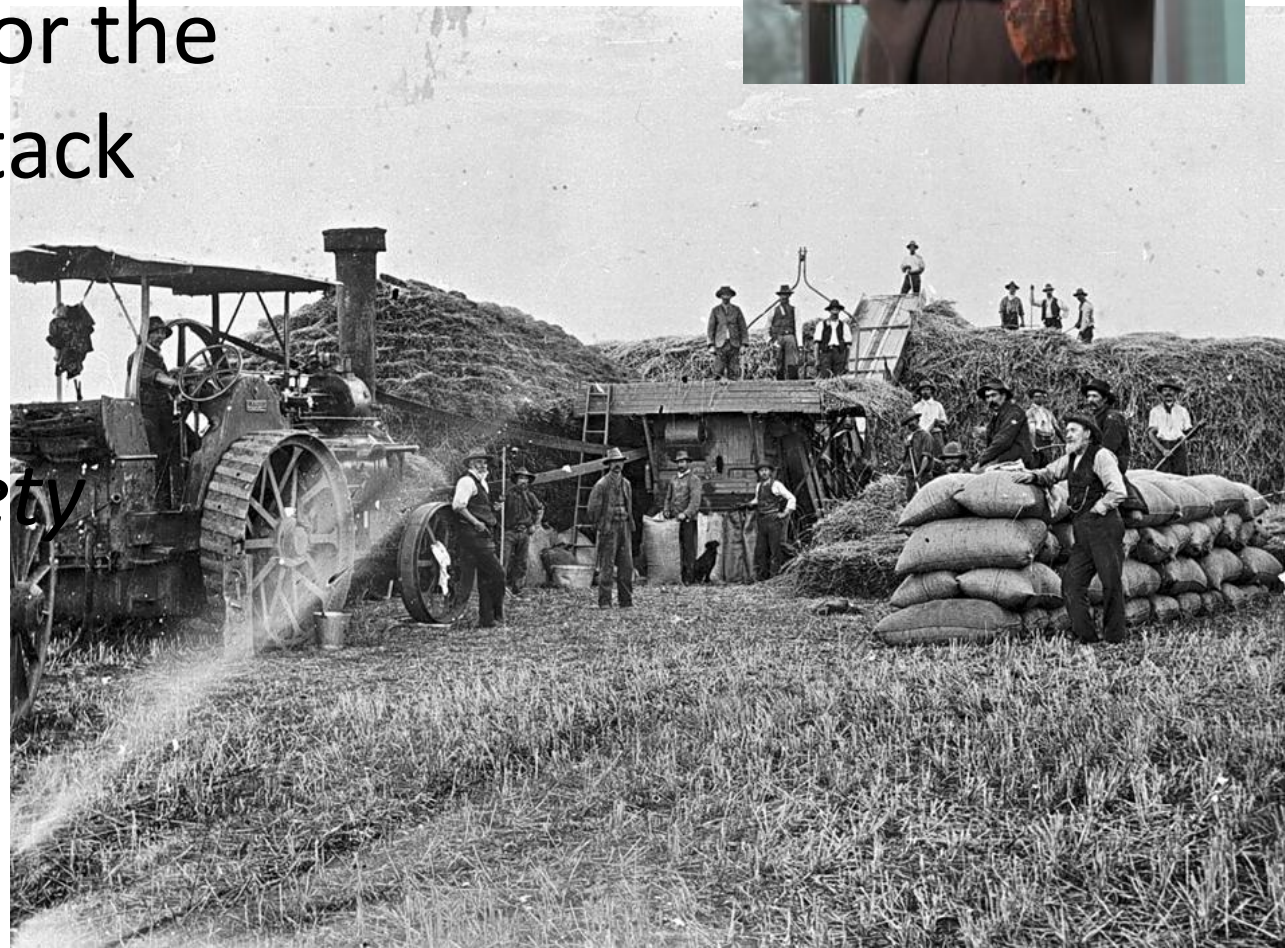
More than 90% of cancer patients carry a mutation that may be responsive to a known drug

Mark Rubin, Weill Cornell Medical College and NewYork-Presbyterian Hospital in New York in *Nature, April, 2015*



- Trying to find a single causative gene for diseases with a complex genetic background is like looking for the proverbial needle in a haystack

- Dr. Nancy Cox
  *President of American Society of Human Genetics*

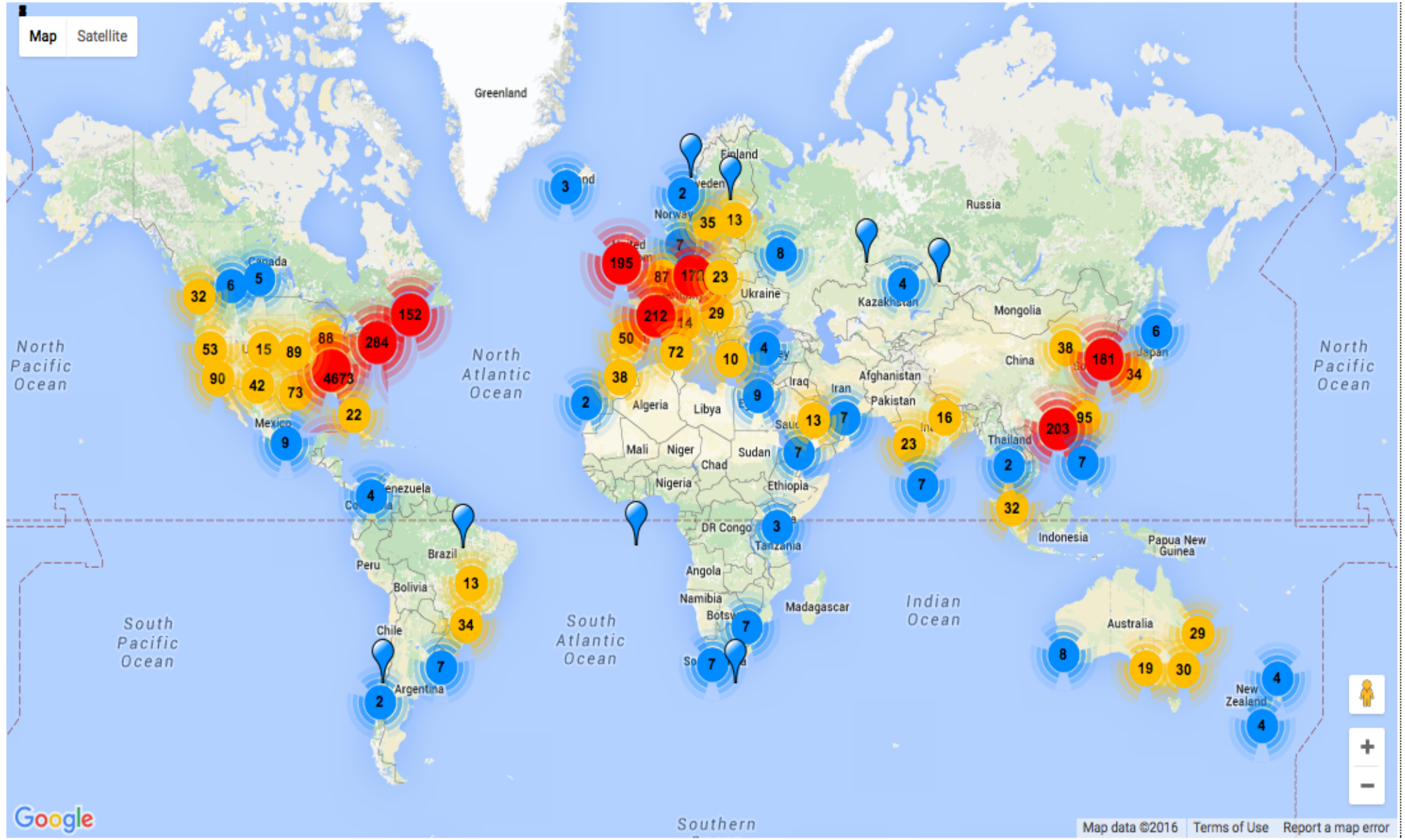


# Next Generation Genomics: World Map of High-throughput Sequencers

Machine Statis

Show all platforms | 454 | HiSeq | HiSeq X Ten | Illumina GA2 | Ion Torrent | MiSeq | MinION | NextSeq | PacBio | Polonator | Proton | SOLiD | Service Provider

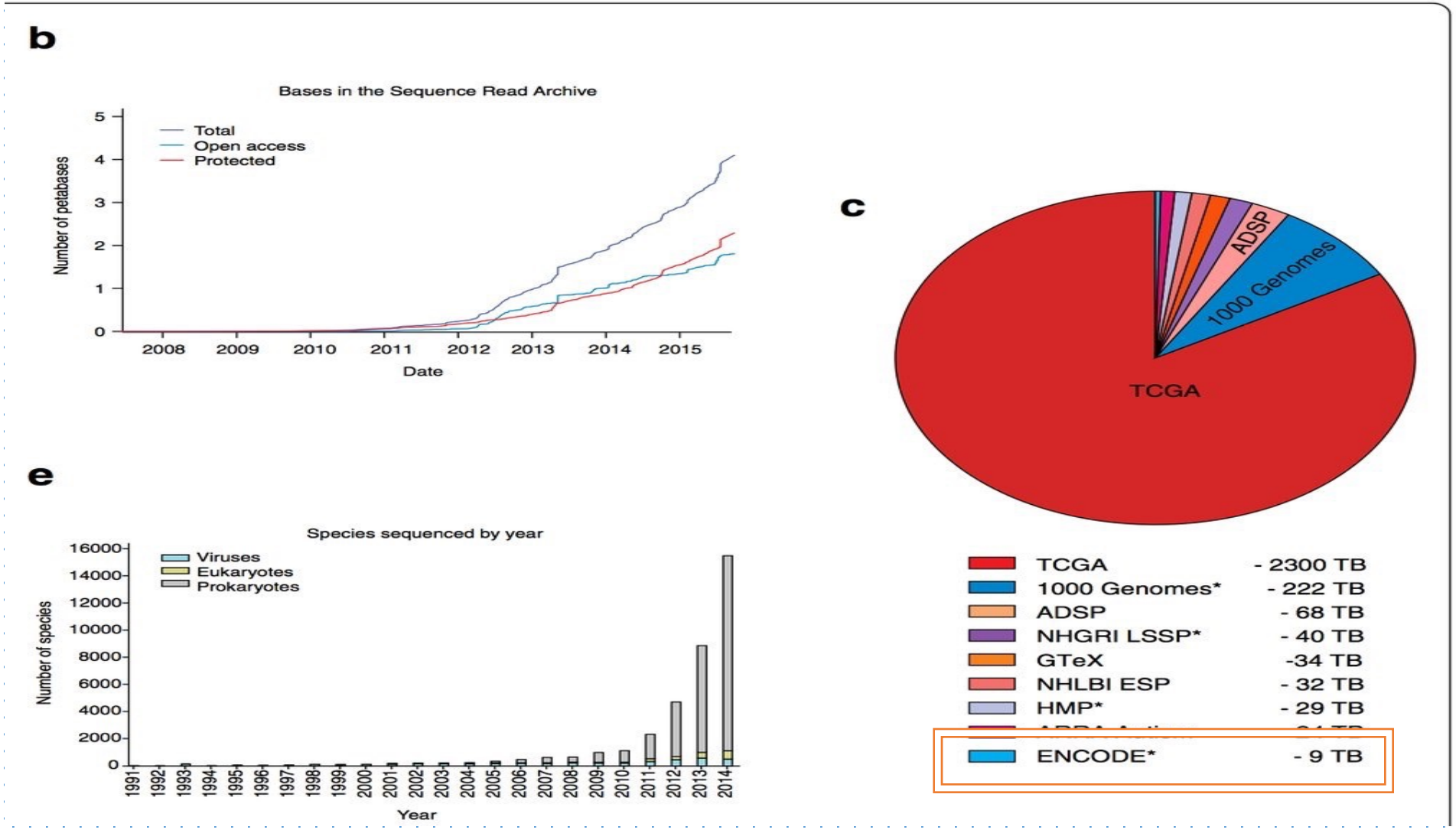
Jump to country --



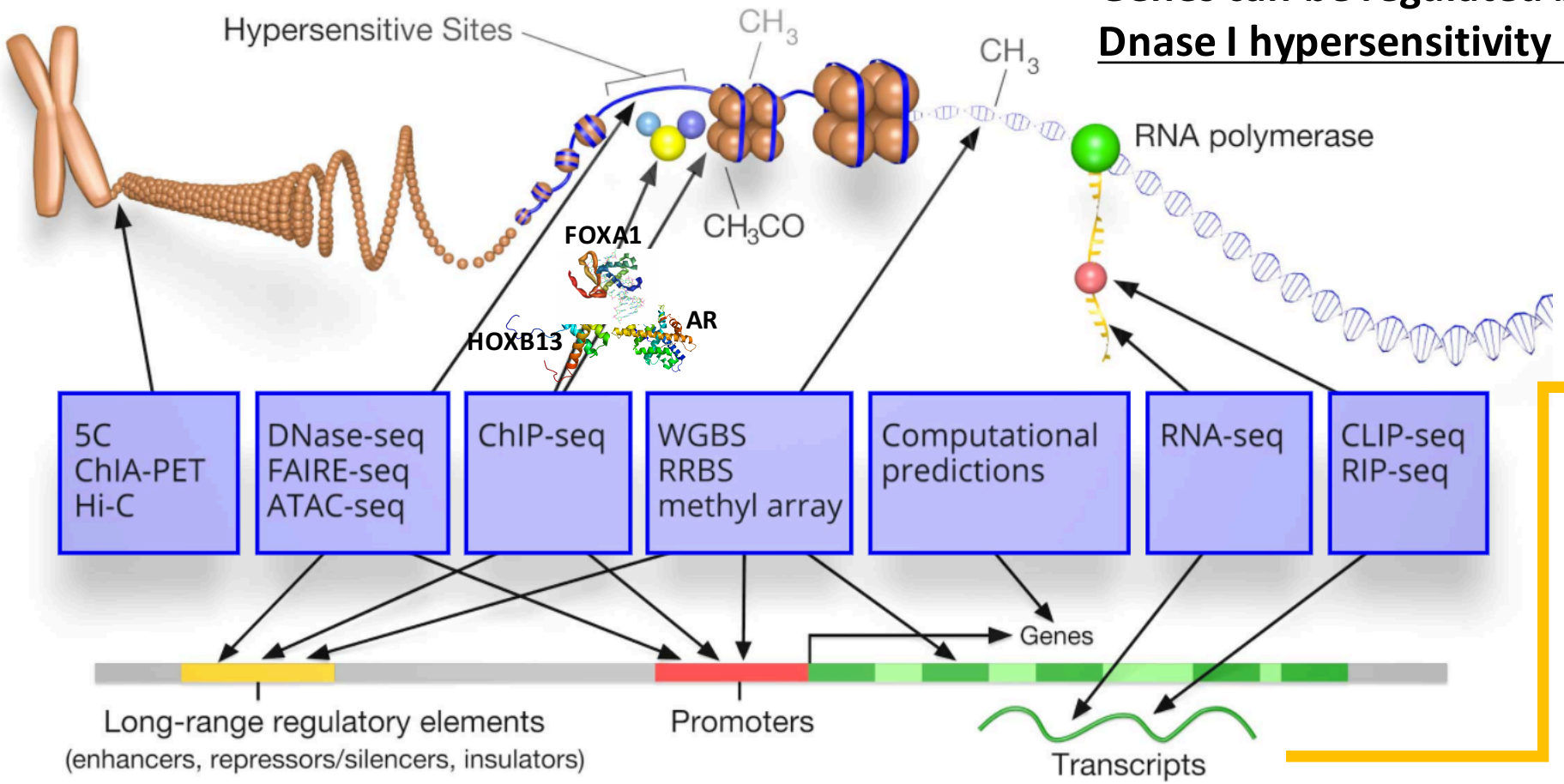

Add New Facility

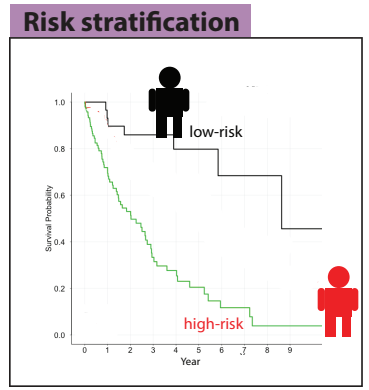
Bioinformatic Center for Health Research,S
Key Laboratory of Cardiovascular Disease
China
Beijing, China
"G.Prodi" Cancer Research Center
Emilia-Romagna, Italy
Aalborg University
Nordjylland, Denmark
ABi SOLiD 5500xl
Csongrád, Hungary
AC-Gen Reading Life
Valladolid , Spain
Academic Medical Center
Noord-Holland, Netherlands
ACGT Sdn Bhd
WP Kuala Lumpur, Malaysia
ACHRI Genomics and Bioinformatics
AB, Canada
Advanced Center for Treatment, Research
Education in Cancer (ACTREC)
Maharashtra, India
Agricultural Genetic Engineering Research
Institute
Giza, Egypt
Agricultural Genetic Engineering Research
Institute
Giza, Egypt
AgriLife Genomics and Bioinformatics Serv
Facility
TX, United States
AITbiotech Pte Ltd
Singapore, Singapore
Akershus University Hospital
Lørenskog, Norway
Alacris Theranostics GmbH
Berlin, Germany
Albert Einstein College of Medicine,
Epigenomics Shared Facility
NY, United States
Albert Einstein College of Medicine, Genor
Core

# Big Biomedical Data Sources



**Genes can be regulated by TFs, histon modification,**
**Dnase I hypersensitivity (DNAse-seq)**




**Disease association**

**Clinical data**

Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

https://www.encodeproject.org/

# Big Data Challenges in Transcriptional Regulatory Network Analysis (TReNA)

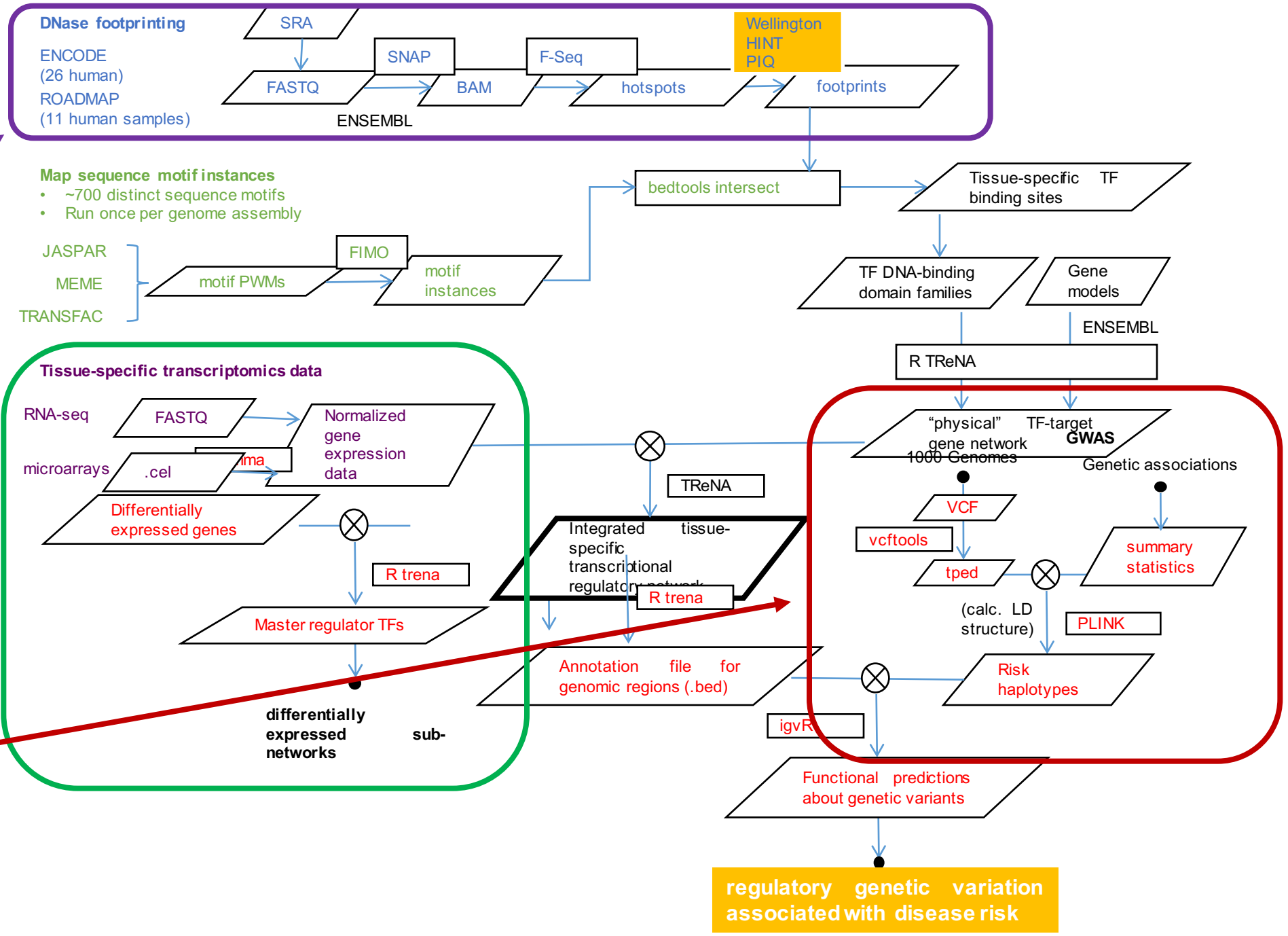
- Identifying and transferring raw data objects
  - BDBag and minid
- Building and optimizing workflows
  - Shell scripting to drag/drop tool box, File I/O
- Scalable data analysis on cloud
  - Batch submit, Monitor jobs

# What is TReNA?

Identifying sites for DNA-Protein interactions

Gene expression

Disease association



# DNA-binding data in ENCODE

- DNAse-seq, FAIRE-seq, and ATAC-seq for the hypersensitive site
- Total number of tissues: 27 (lymphoblast, brain, skin, etc)
- Total number of patient samples: 206
- Total number of fastq files: **1379**
  - Each patient sample has a few to many replicates
- Total size of the raw data: **2.5 TB**


T.SCUDIERO
WWW.TSCUDIERO.COM

# BDDS Solutions:

## Enabling TReNA – BDBag

BDDS BIG DATA *for* DISCOVERY SCIENCE

Create a BDBag from an ENCODE search.

For example enter the following search:

https://www.encodeproject.org/search/?type=Experiment&assay_title=RNA-seq&replicates.library.biosample.biosample_type=stem+cell

Or paste in an Encode metadata file.

Encode Search Query | Encode Metadata File

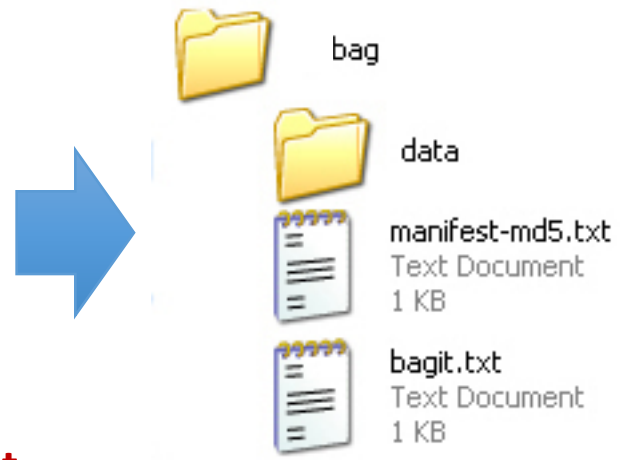
File accession File format Output type Experiment accession Assay Biosample term id Biosample term name Biosample type Biosample life stage
Biosample sex Biosample organism Biosample treatments Biosample subcellular fraction term name Biosample phase Biosample synchronization stage
Experiment target Antibody accession Library made from Library depleted in Library extraction method Library lysis method Library crosslinking method
Experiment date released Project RBNS protein concentration Library fragmentation method Library size range Biosample Age Biological replicate(s)
Technical replicate Read length Run type Paired end Paired with Derived from Size Lab md5sum File download URL Assembly Platform

Create BDBag

BDBag created: ark:/99999/fk40294z6m

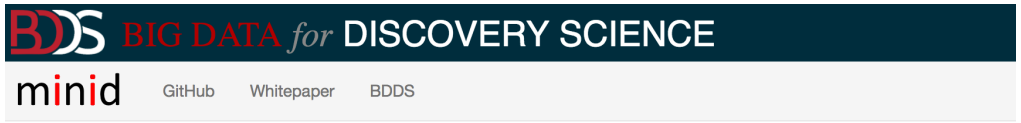
or you can access, transfer, and share the complete, materialized BDBag with Globus

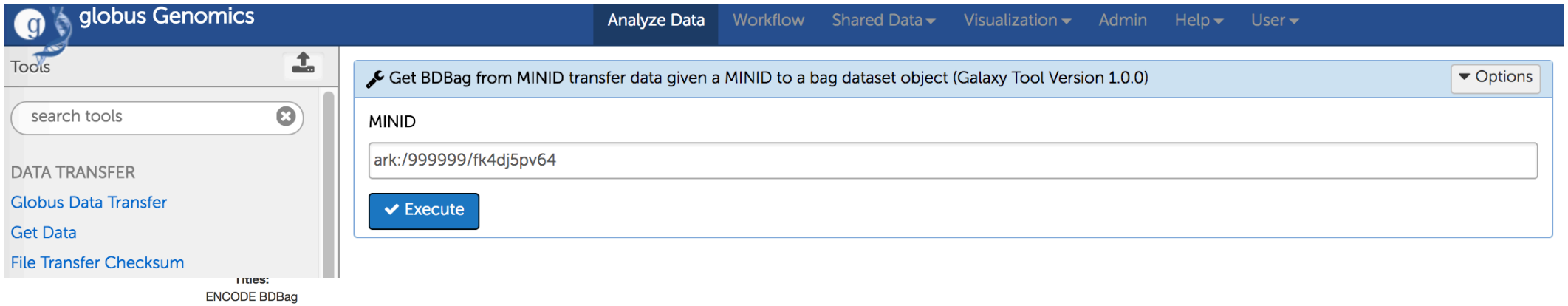
**Lymphoblast metadata**




https://github.com/ini-bdds/bdbag

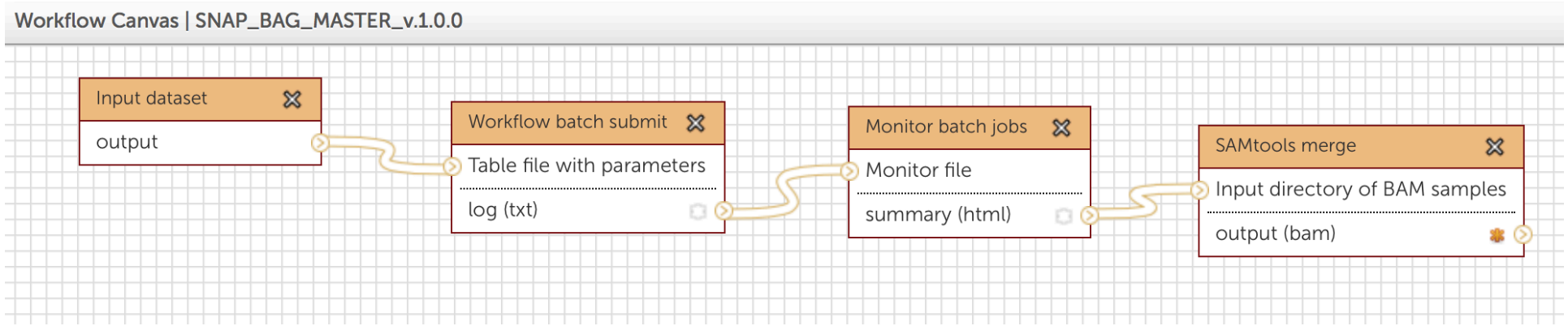
# Enabling TReNA – BDBag



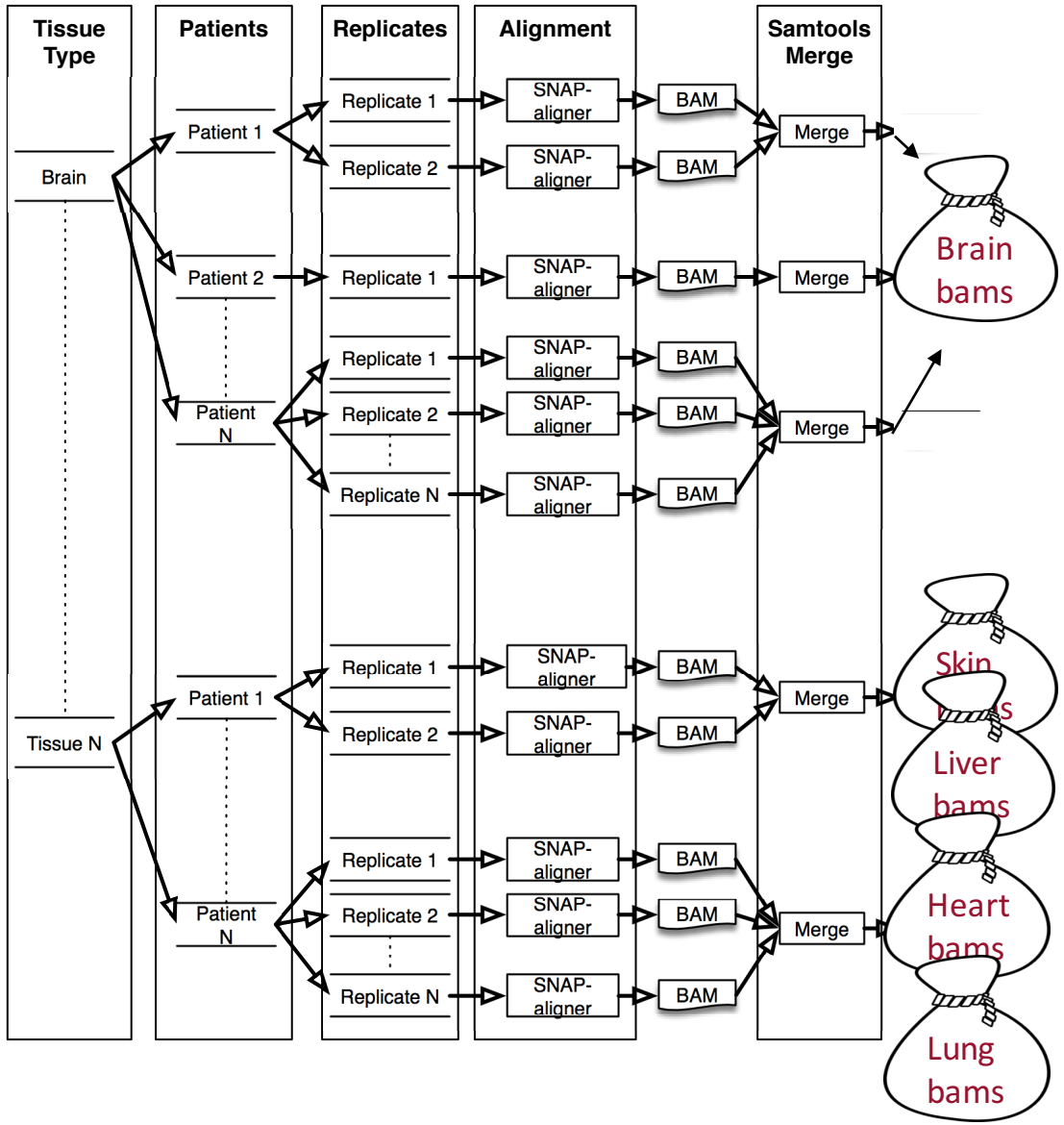

https://github.com/ini-bdds/bdbag

# BDDS Solutions:
# Enabling TReNA – Analysis pipelines

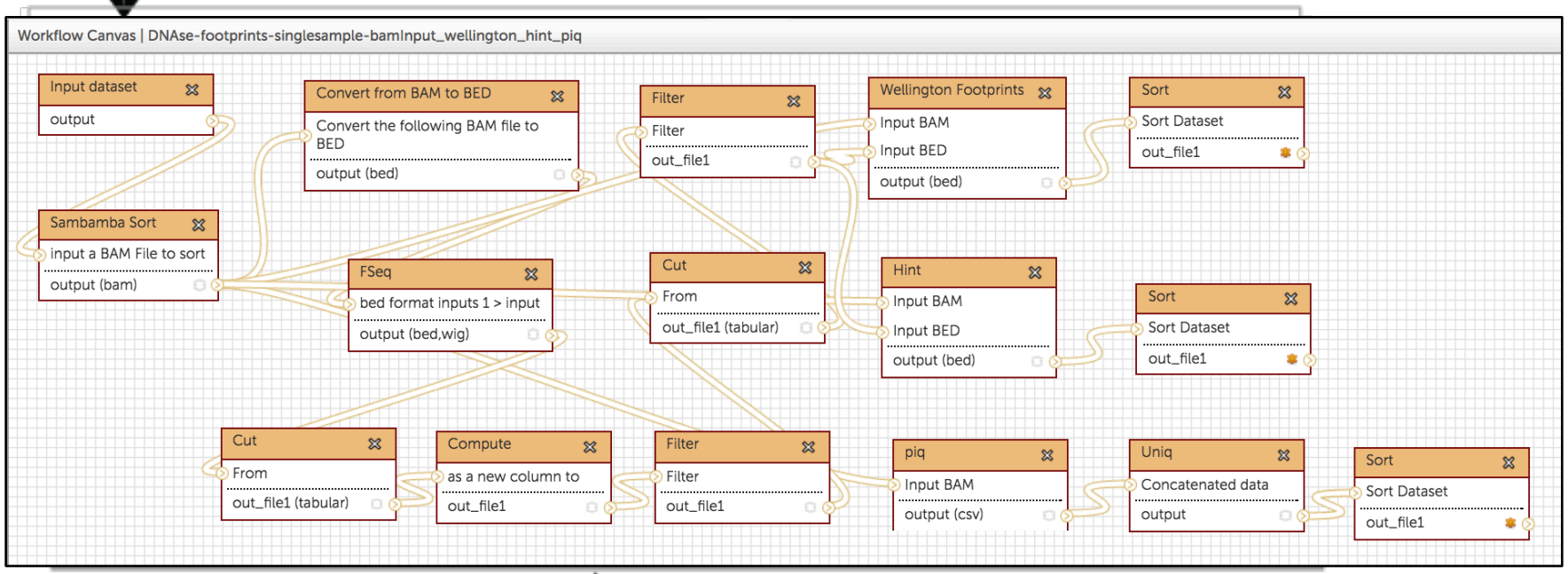



- Take a tissue sample bag (i.e. brain, skin, etc)
- Submit each bag to the alignment with the latest human reference genome GRCh38
- Merges samples that are from the same patient group

https://bdds.globusgenomics.org/

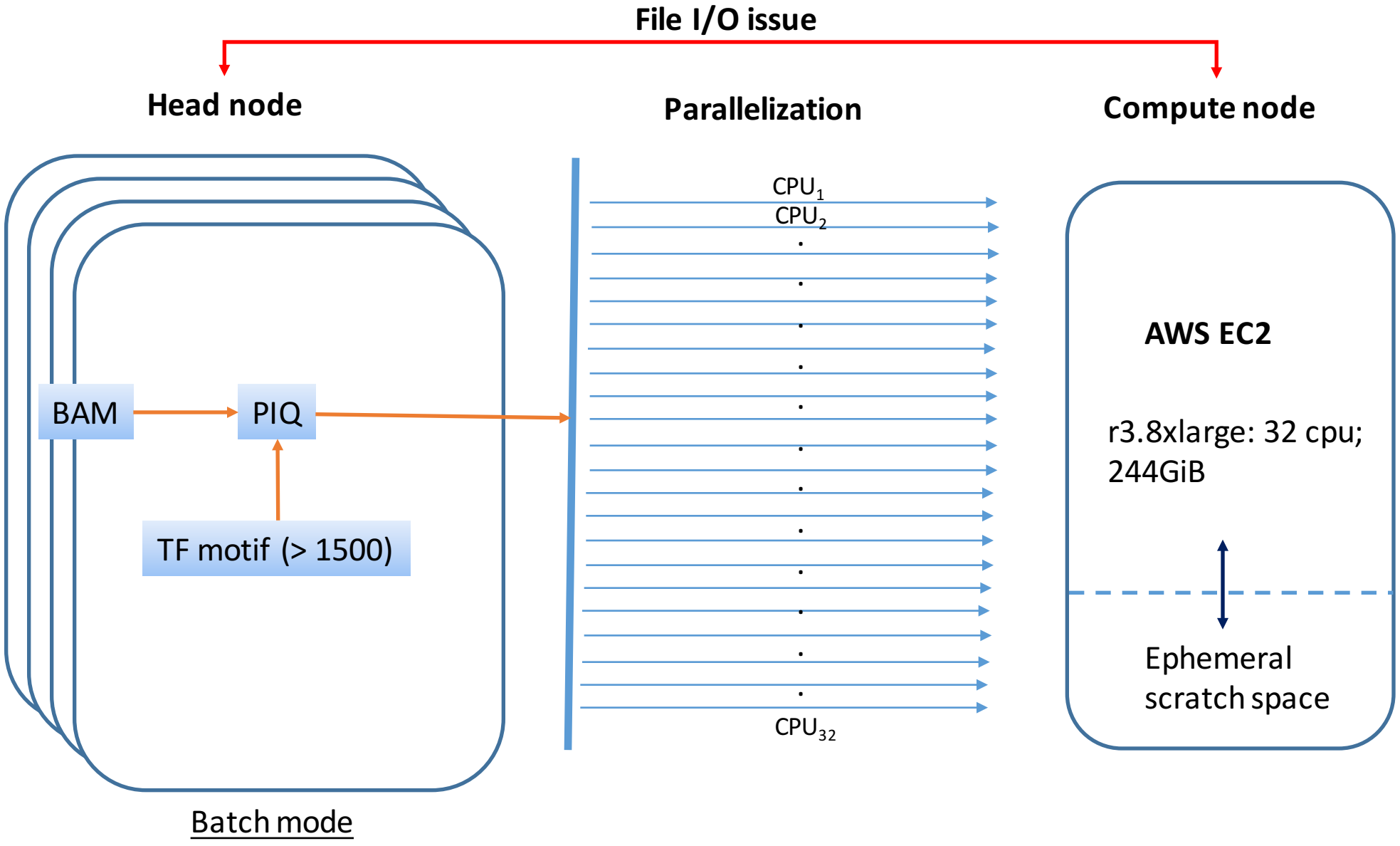

Tissue Type
Patients
Replicates
Alignment
Samtools Merge
Brain
Tissue N
Patient 1
Patient 2
Patient N
Patient 1
Patient N
Replicate 1
Replicate 2
Replicate N
SNAP-aligner
BAM
Merge
Brain bams
Skin
Liver bams
Heart bams
Lung bams


27 bags




**Translation of command lines to a galaxy workflow**

**Running at scale with batch submission**


File I/O issue
Head node
Parallelization
Compute node
BAM
PIQ
TF motif (> 1500)
CPU1
CPU2
CPU32
AWS EC2
r3.8xlarge: 32 cpu;
244GiB
Ephemeral
scratch space
Batch mode

# BDDS Solutions: Enabling TReNA – Analysis pipelines

```
### METADATA
#######################################
Workflow Name  DNAse-footprints-singlesample-bamInput_wellington_hint_piq
Workflow id    6f2c19b5325b338f
Project Name   <lymphoblast>
#######################################


###TABLE DATA
#######################################
SampleName      ##SourceType::SourceName::Input Bam
ENCSR000DBY     library::DNAse-lymphoblast::ENCSR000DBY.batch_submit.txt.batch.log.results.bam
ENCSR000DBZ     library::DNAse-lymphoblast::ENCSR000DBZ.batch_submit.txt.batch.log.results.bam
ENCSR000DCA     library::DNAse-lymphoblast::ENCSR000DCA.batch_submit.txt.batch.log.results.bam
ENCSR000DCB     library::DNAse-lymphoblast::ENCSR000DCB.batch_submit.txt.batch.log.results.bam
ENCSR000DCC     library::DNAse-lymphoblast::ENCSR000DCC.batch_submit.txt.batch.log.results.bam
ENCSR000EJB     library::DNAse-lymphoblast::ENCSR000EJB.batch_submit.txt.batch.log.results.bam
ENCSR000EJC     library::DNAse-lymphoblast::ENCSR000EJC.batch_submit.txt.batch.log.results.bam
ENCSR000EJD     library::DNAse-lymphoblast::ENCSR000EJD.batch_submit.txt.batch.log.results.bam
ENCSR000EJE     library::DNAse-lymphoblast::ENCSR000EJE.batch_submit.txt.batch.log.results.bam
ENCSR000EJF     library::DNAse-lymphoblast::ENCSR000EJF.batch_submit.txt.batch.log.results.bam
ENCSR000E                                                                   h.log.results.bam
ENCSR000E                                                                   h.log.results.bam
ENCSR000E                                                                   h.log.results.bam
ENCSR000E                                                                   h.log.results.bam
ENCSR000E                                                                   h.log.results.bam
ENCSR000E                                                                   h.log.results.bam
ENCSR000E                                                                   h.log.results.bam
ENCSR000E                                                                   h.log.results.bam
ENCSR000E                                                                   h.log.results.bam
ENCSR000E                                                                   h.log.results.bam
```

Monitor batch jobs (version 1.0.0)

Monitor file:

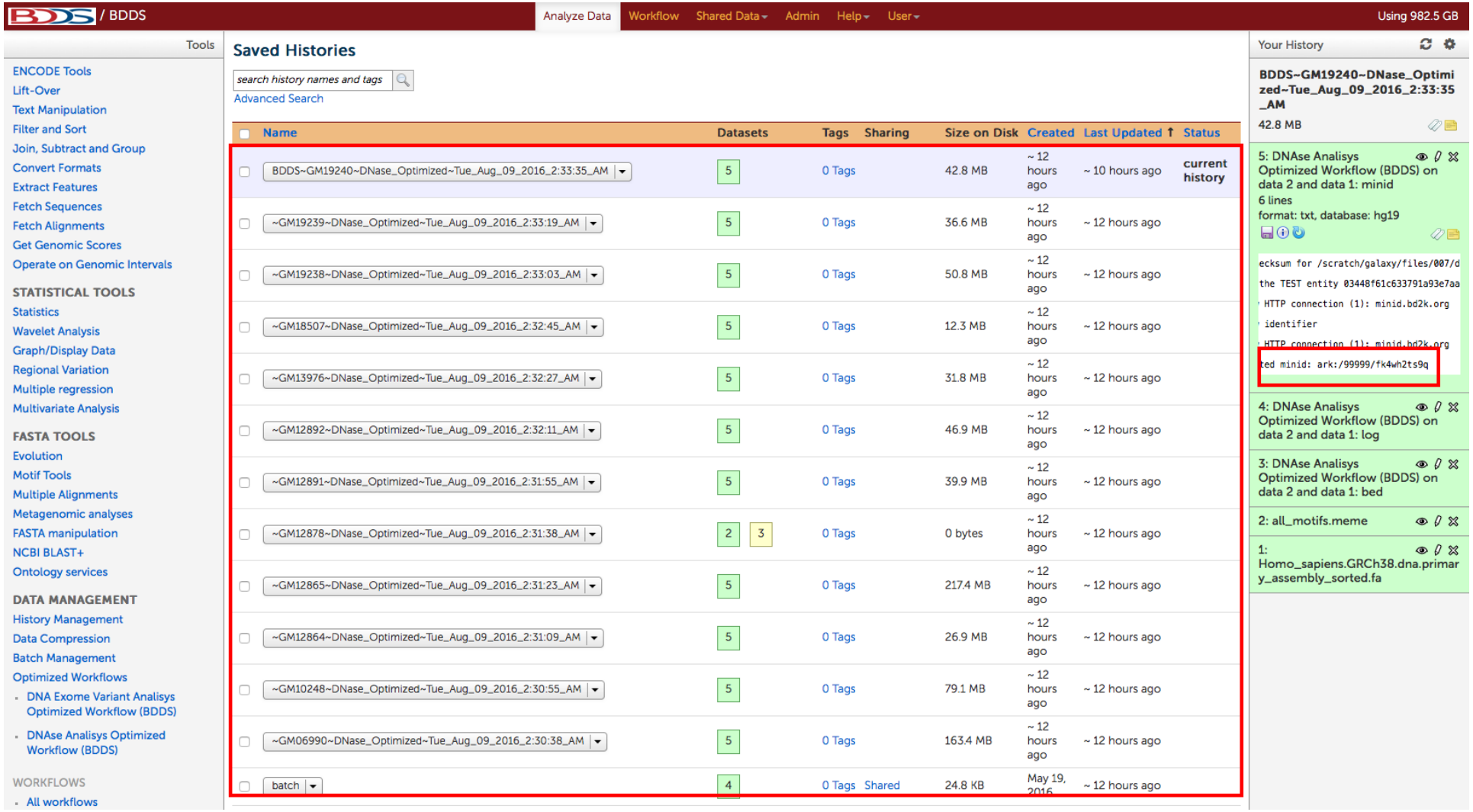
68: Monitor for batch submission data 64

This should be the output from the batch submit job

Execute

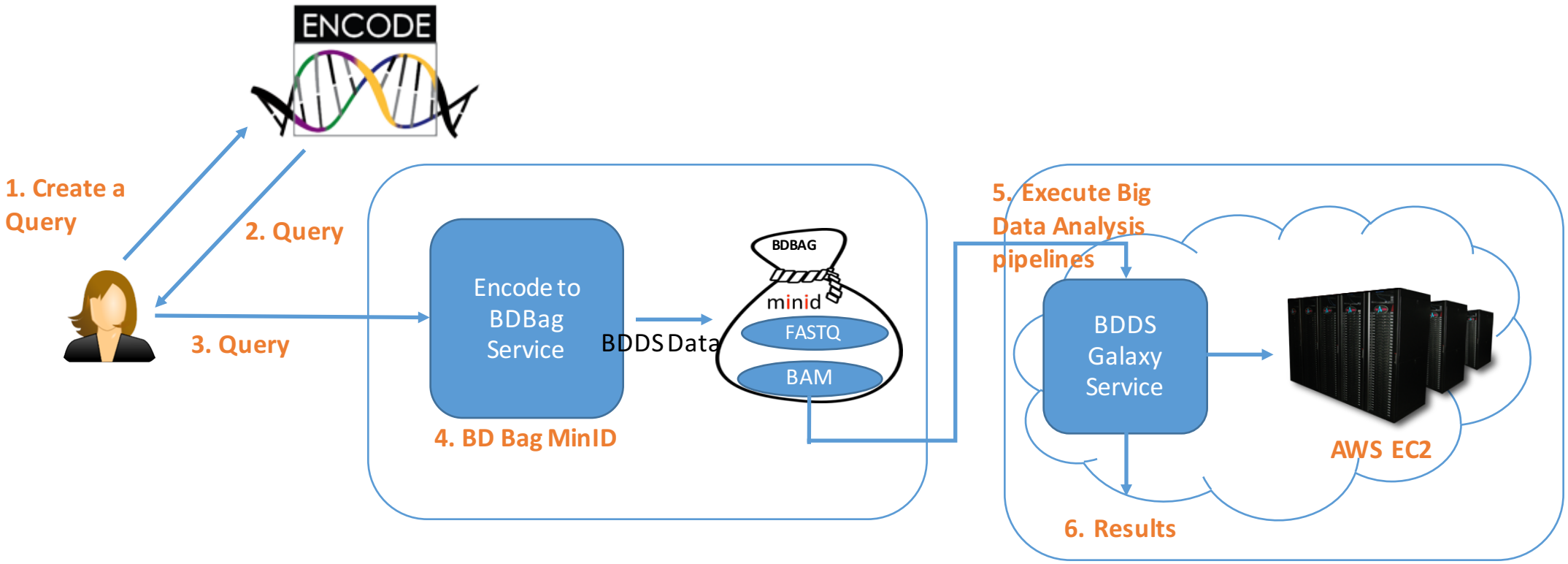
- Run at scale with batch submit to accelerate the performance
- Upon job completion, the monitor batch jobs tool collects all the final output

https://bdds.globusgenomics.org/




bdds.globusgenomics.org

# Dnase Hypersensitivity Analysis



# Results from TReNA analysis

- Total number of tissues: 27
- Total number of patient samples: 206
- Total number of fastq files: 1379
- Total size of the raw data: 2.5 TB
- Number of new tools added: 20
- Number of HPC workflows created: 11
- Number of compute hours for alignment: 24,000 CPU hours
- Number of compute hours for footprinting: 150,000 CPU hours
- Number of databases created: 106

- Notably, all the work was completed within **two weeks** which was originally expected for **several months**