Big Data for Cancer Research

Patipark Kueanjinda, Ph.D.

Department of Microbiology
Faculty of Medicine, Chulalongkorn University

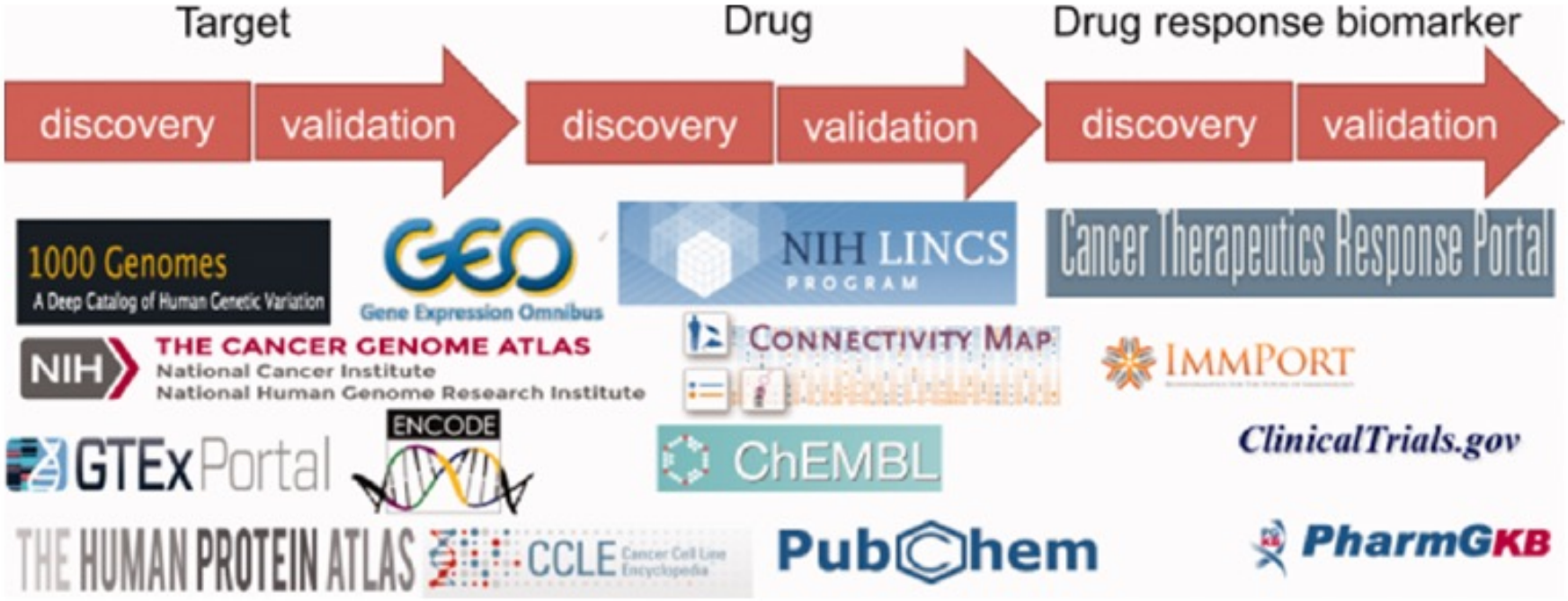
Objectives

- Understanding the Cancer Genome Atlas (TCGA) and Genomic Data Commons (GDC)
- Install and configure TCGABiolink and GenomicDataCommons R packages
- Access and download TCGA cancer genomics data using TCGABiolink
- Explore and visualize gene expression, mutation, and clinical data using R
- Utilize Galaxy for data analysis.

Course materials

- R program
- Rstudio
- Galaxy subscription

Public Databases

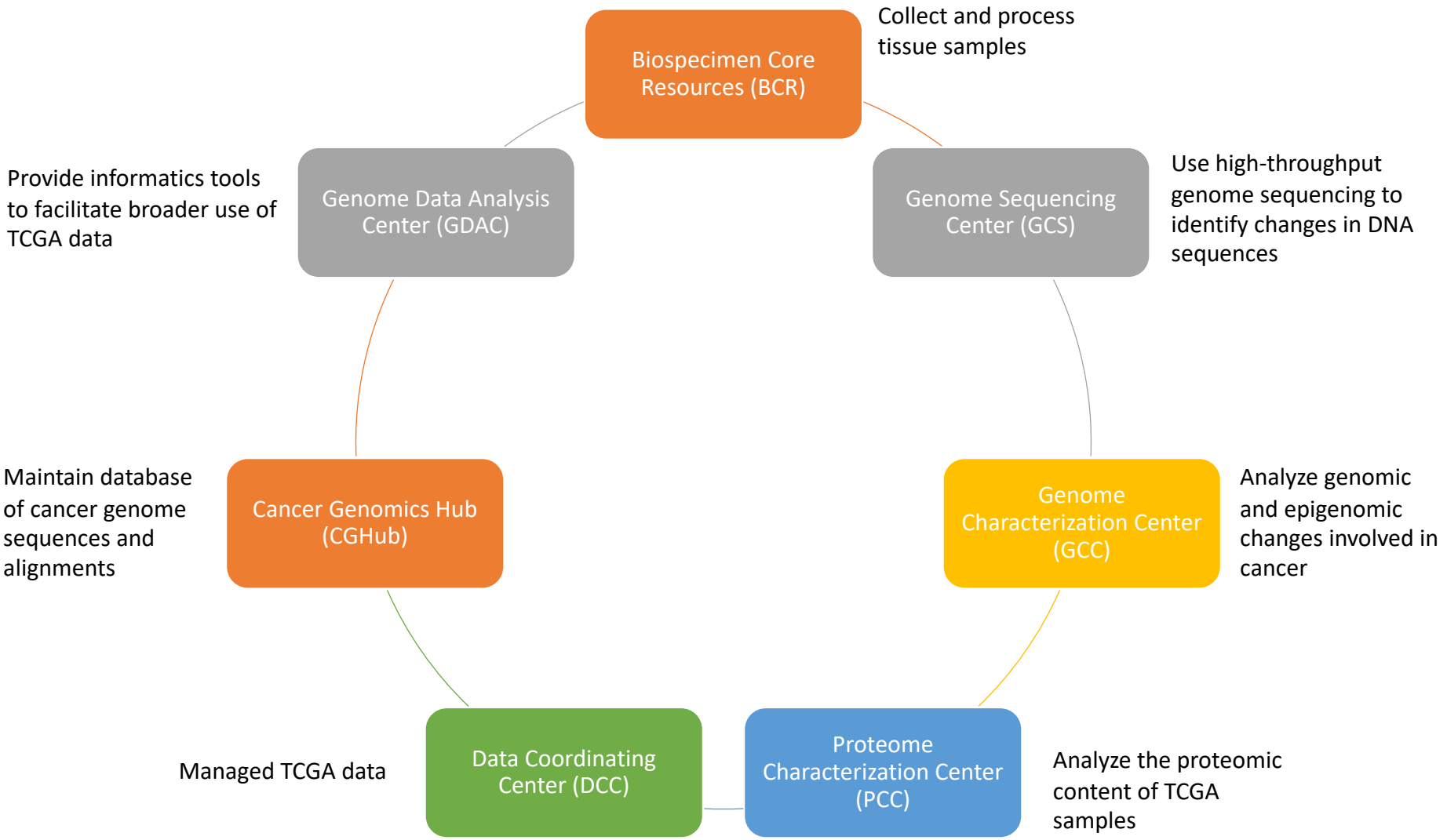


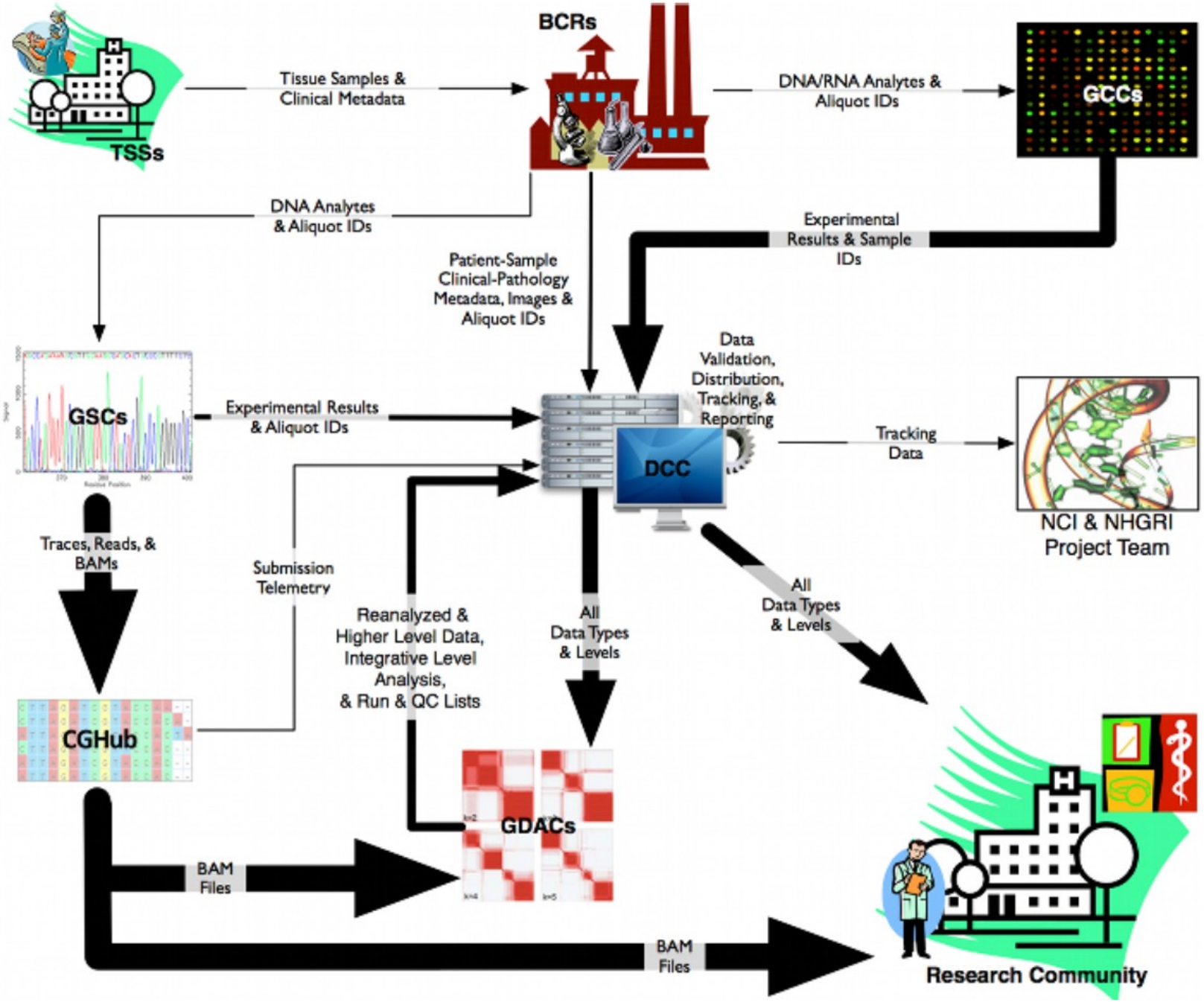
The Cancer Genome Atlas

- National Cancer Institute + National Human Genome Research Institute.
- Increase scientific understanding of the molecular basis of cancer and apply this information to improve the ability to dianose, treat, and prevent cancer.
- Develop a complete atlas of all genomic alterations involved in cancer.

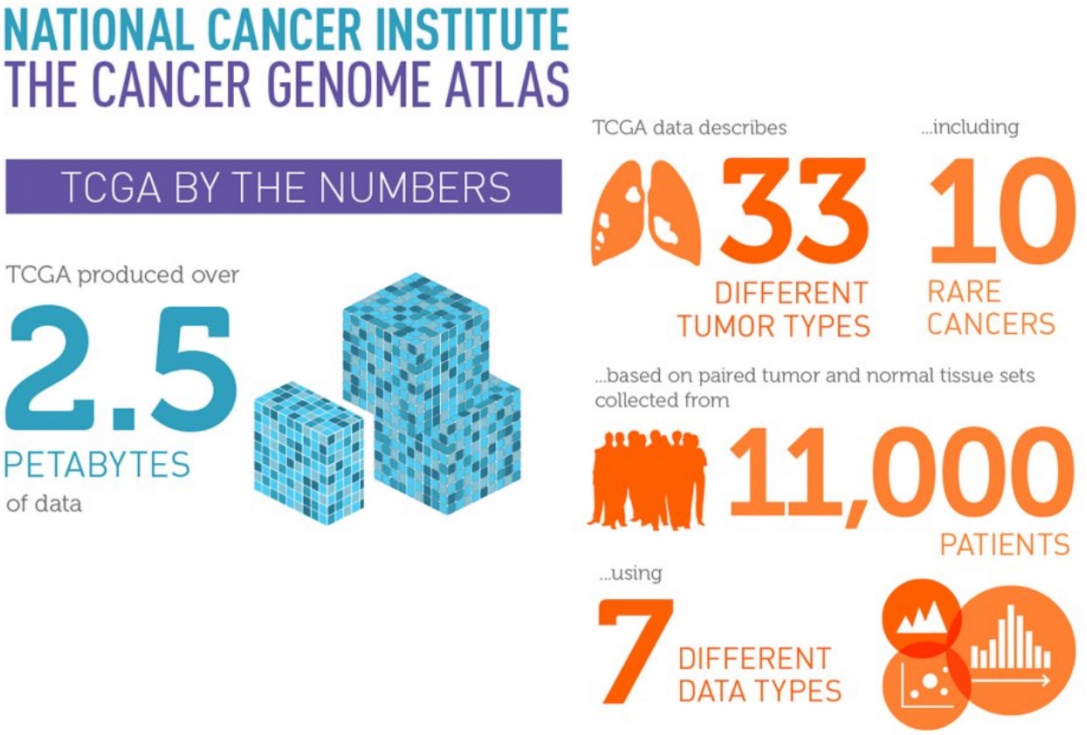
Tumor and adjacent normal tissues from ~15,000 patients coverying 33 cancer types.

Major TCGA Research Components




TSSs
Tissue Samples & Clinical Metadata
BCRs
DNA/RNA Analytes & Aliquot IDs
GCCs
DNA Analytes & Aliquot IDs
Patient-Sample Clinical-Pathology Metadata, Images & Aliquot IDs
Experimental Results & Sample IDs
GSCs
Experimental Results & Aliquot IDs
Data Validation, Distribution, Tracking, & Reporting
DCC
Tracking Data
NCI & NHGRI Project Team
Traces, Reads, & BAMs
Submission Telemetry
Reanalyzed & Higher Level Data, Integrative Level Analysis, & Run & QC Lists
All Data Types & Levels
All Data Types & Levels
CGHub
GDACs
BAM Files
BAM Files
Research Community

TCGA Data Outputs



1. Clinical data
2. DNA sequencing
3. miRNA sequencing
4. Protein expression
5. mRNA sequencing
6. DNA methylation
7. Copy number

TCGA Data Levels

Data level	Level type	Description	Example
1	Raw	-Low-level data for single sample -Not normalized	-Sequence trace file -Affymetrix CEL file -BAM file
2	Processed	-Normalizd single sample data -Interpreted for presence or absence of specific molecular abnormalities	-Putative mutation call for a single sample -Probed locus amplification/deletion/loss of heterozygosity calls in a sample -Signal of a probe or probe set for a sample
3	Segmented/Interpreted	-Aggregate of processed data from single sample -Grouped by probed loci to form larger contiguous regions	-Validated mutation call for a single sample -Amplification/deletion/loss of heterzygosity calls for a sample region -Expression signal of a gene for a sample -Genomic copy-number data
4	Summary/Region of Interest	-Quantified assiciation across classes of samples -Associations based on two or more molecular abnormalities, sample characteristics, clinical variables	-Discovery that a genomic region is amplified in 10% of TCGA glioma samples

TCGA Data Access

- Controlled access: Tier 1 data
  - Fastq files, aligned BAM files, VCF files, etc.
  - Need dbGAP permission to access these files.
  - Permission needs to be renewed every years.
- Open access: Tier 3 data
  - De-identified clinical and demographic data
  - Gene expression data
  - Copy number alterations in regions of the genome
  - Epigenetic data
  - Etc.

Genomic Data Commons

- Part of the NIH Big Data to Knowledge (BD2K) initiative which was launched on 2016.
- Unified knowledge base that promotes sharing of genomic and clinical data between researchers and facilitates precision medicine in oncology.
- Contains standardized data from apprx. 14,500 patients in 42 cancer types derived from NCI programs, including:
  - The Cancer Genome Atlas (TCGA)
  - Therapeutically Applicable Research to Generate Effective Treatment (TARGET)
  - Cancer Genome Characterization Initiative (CGCI)
  - The Cancer Cell Line Encyclopedia (CCLE)

Where to download data?

- TCGA Data Portal
  - https://tcga-data.nci.nih.gov/tcga/
- GDAC at Broad Institute
  - http://gdac.broadinstitute.org
- cBioPortal
  - http://www.cbioportal.org/public-portal/
- The Cancer Genomics Hub (CGHub)
  - https://cghub.ucsc.edu

Access to Tier 3 data

Access to Tier 1 data

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home | Projects | Exploration | Analysis | Repository | Quick Search | Manage Sets | Login | Cart 0 | GDC Apps

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Projects | Exploration | Analysis | Repository

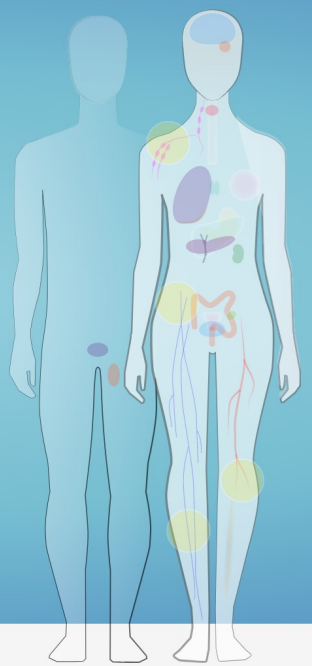
e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

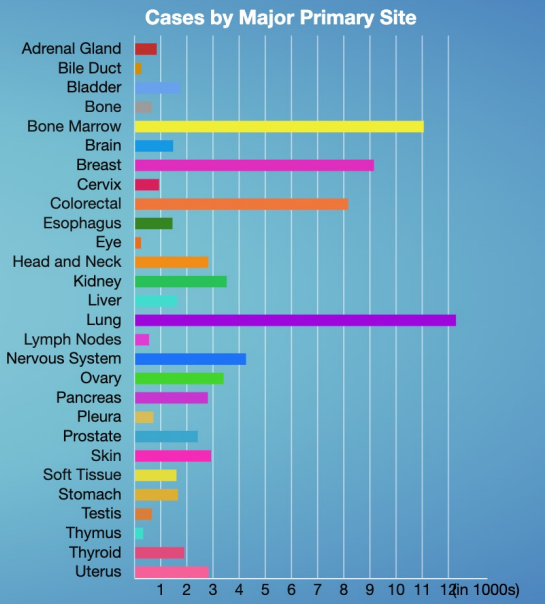
Data Portal Summary

Data Release 38.0 - August 31, 2023

PROJECTS	PRIMARY SITES	CASES
82	68	88,991

FILES	GENES	MUTATIONS
1,003,747	22,588	2,903,037

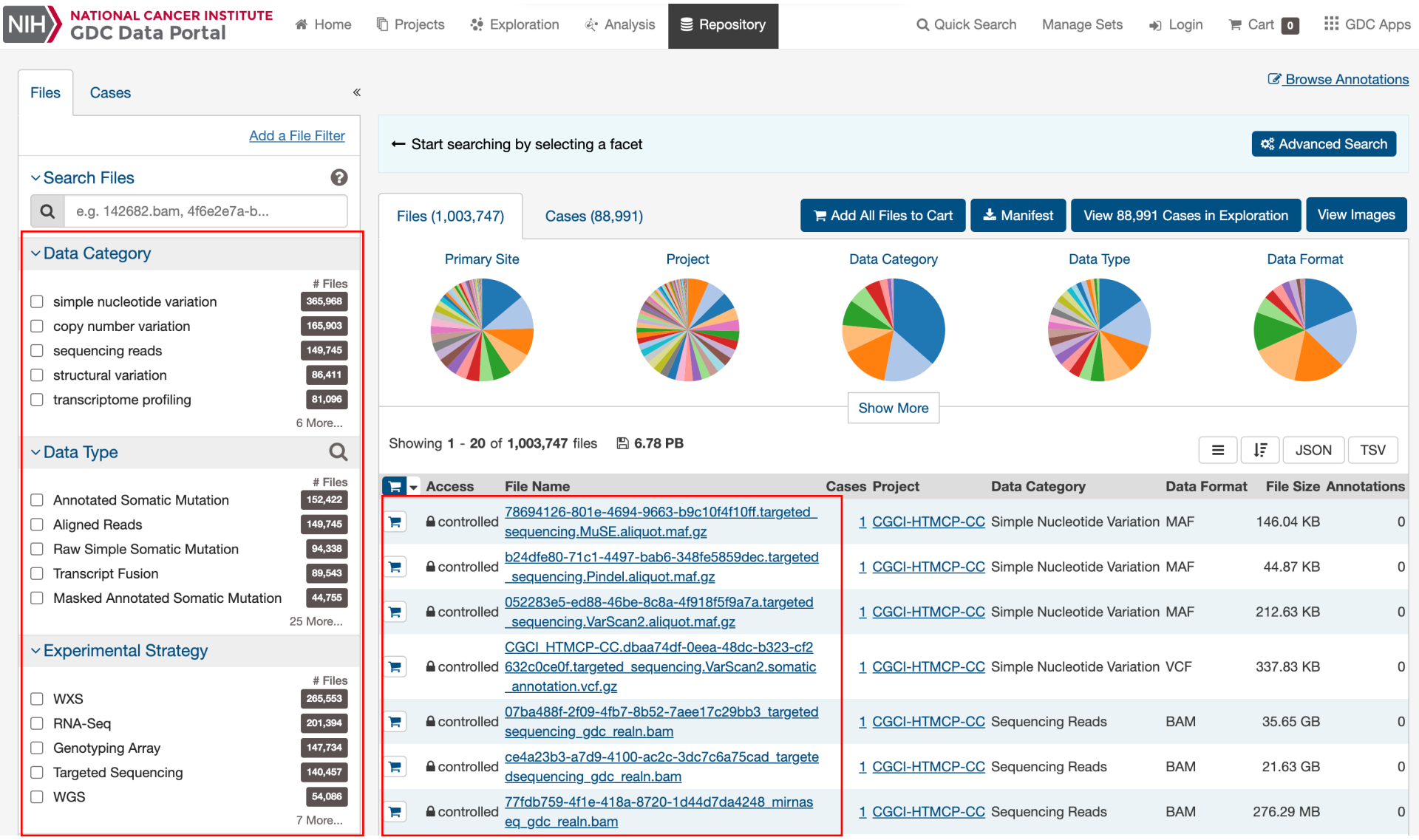




GDC Applications

The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

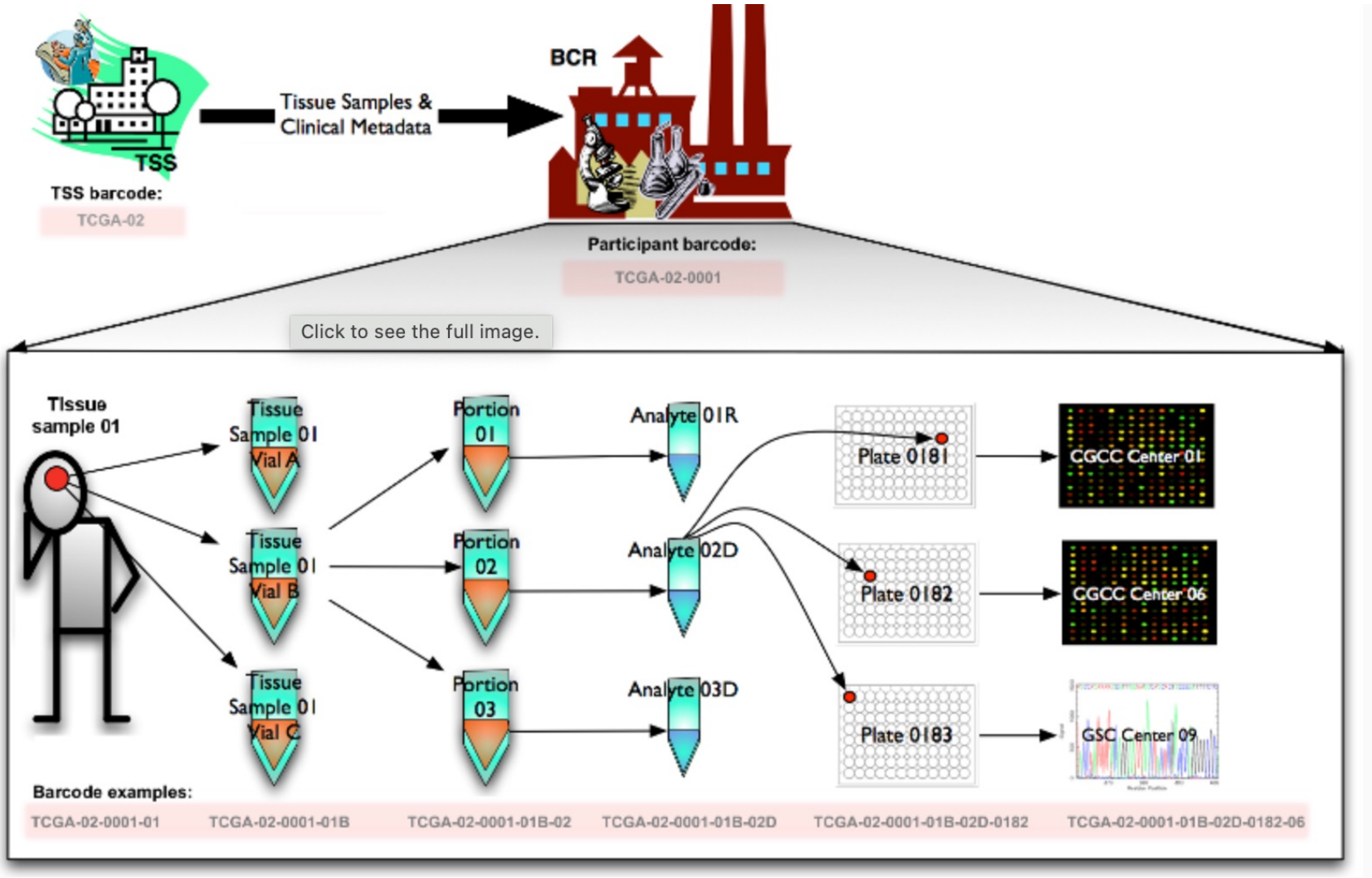
Data Portal | Website | API | Data Transfer Tool | Documentation | Data Submission Portal | Publications

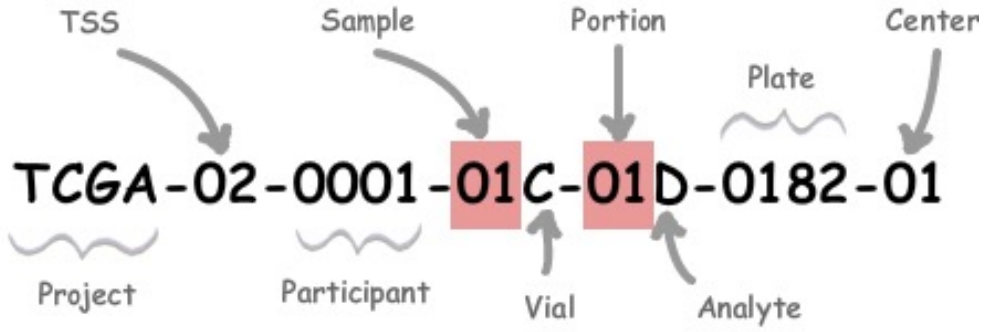


Criteria to filter data

Controlled data (Tier 1)

TCGA Barcode





Label	Identifier for	Value	Value Description	Possible Values
Analyte	Molecular type of analyte for analysis	D	The analyte is a DNA sample	See Code Tables Report
Plate	Order of plate in a sequence of 96-well plates	182	The 182nd plate	4-digit alphanumeric value
Portion	Order of portion in a sequence of 100 - 120 mg sample portions	1	The first portion of the sample	01-99
Vial	Order of sample in a sequence of samples	C	The third vial	A to Z
Project	Project name	TCGA	TCGA project	TCGA
Sample	Sample type	1	A solid tumor	Tumor types range from 01 - 09, normal types from 10 - 19 and control samples from 20 - 29. See Code Tables Report for a complete list of sample codes
Center	Sequencing or characterization center that will receive the aliquot for analysis	1	The Broad Institute GCC	See Code Tables Report
Participant	Study participant	1	The first participant from MD Anderson for GBM study	Any alpha-numeric value
TSS	Tissue source site	2	GBM (brain tumor) sample from MD Anderson	See Code Tables Report

Clinical Data

- Collected by Biospeciment Core Resource (BCR)
- Biotab format
  - Tab-delimited, convenient, easy to sort and manipulate
- Include a variety of clinical information
  - Demographic, drug treatment, radiation treatment, survival

DNA Sequencing

- Sequences of genomic DNA for identification of structural variants
- Whole genome
  - BAM files contain sequences of the entire genome (WGS)
- Exome sequencing
  - BAM files contain sequences of coding regions (WES)
- Mutation
  - MAF files contain mutation annotations
  - VCF files contain variant call and mutation annotations

Copy Number

- Copy number variations (CNVs) are a type of structural variants involving alterations in the number of copies of specific regions of DNA, which can either be deleted or duplicated.
- SNP array
  - CEL files with raw data
  - TXT files with normalized copy number data
- CN array
  - TXT files with raw signals per probe
  - TXT files with copy number alterations for aggregated regions per sample

RNA Sequencing

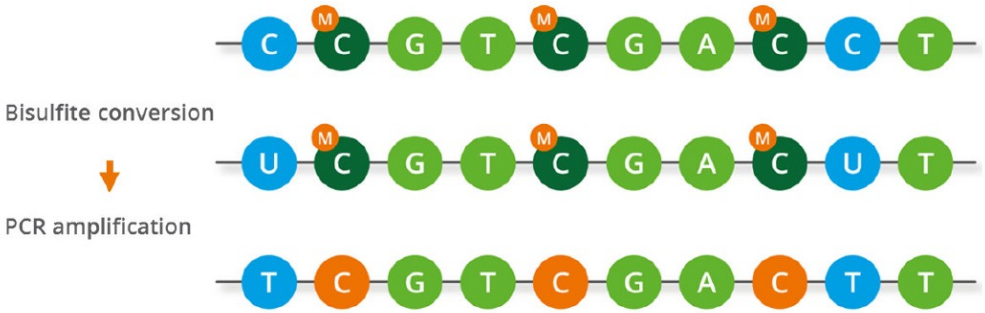
- miRNA-seq contains small non-coding RNA
  - BAM files with raw sequences
  - TXT files with signals at miRNA regions, or isoforms
- mRNA-seq contains transcriptome or expressed genes
  - BAM or FASTQ files contain raw sequences
  - TXT files with expression signals in exon regions, gene regions, splice junction regions, or isoforms
- Total RNA-seq includes both coding (mRNA) and non-coding RNA (miRNA)

Protein Expression

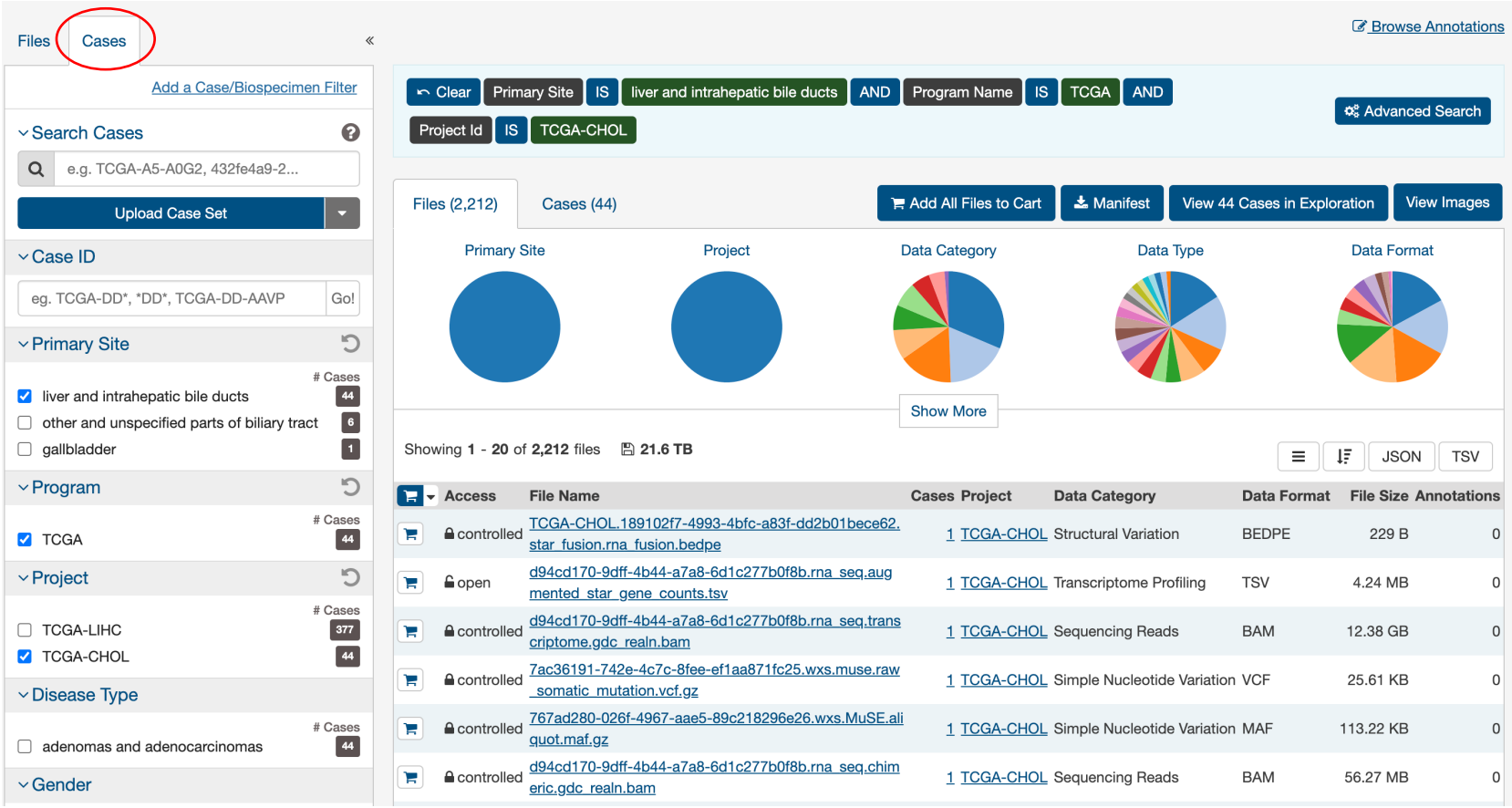
- Generated at MD Anderson Reverse Phase Protein Array (RPPA) core facility
  - High-throughput assay
  - Antibodies printed across slides (150-500 proteins)
  - Quantify the amount of protein in multiple samples simultaneously
- TXT files contain normalized protein expression for each gene per sample

DNA Methylation

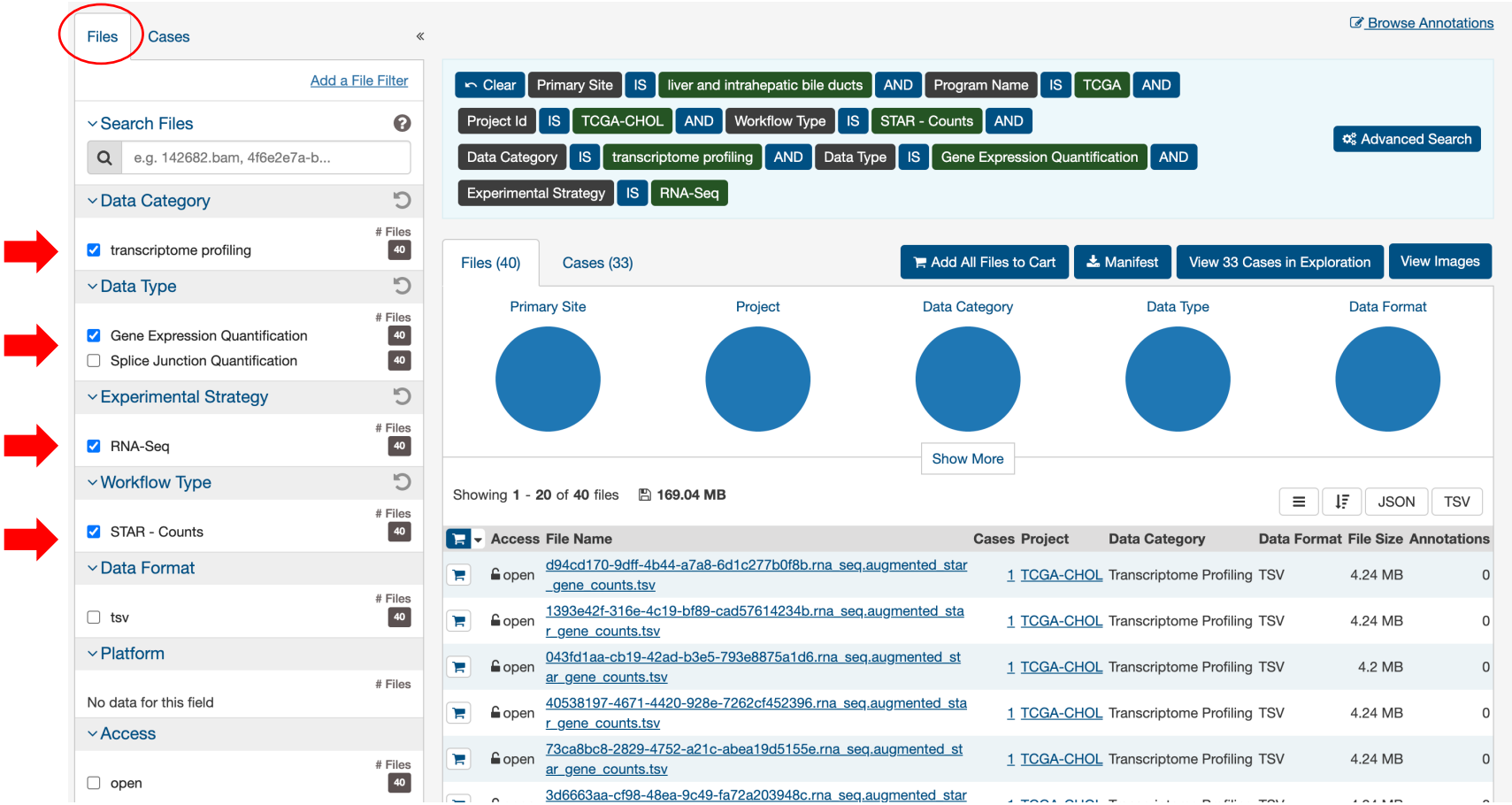
- Sequencing of methylome or epigenome revals DNA methylation profile
- Bisulfite sequencing
  - Use of bisulfite treatment of DNA to determine the pattern of methylation
  - BAM (raw seq), VCF (methylation and mutation calls), or BED (methylation calls) files



Data Download from GDC

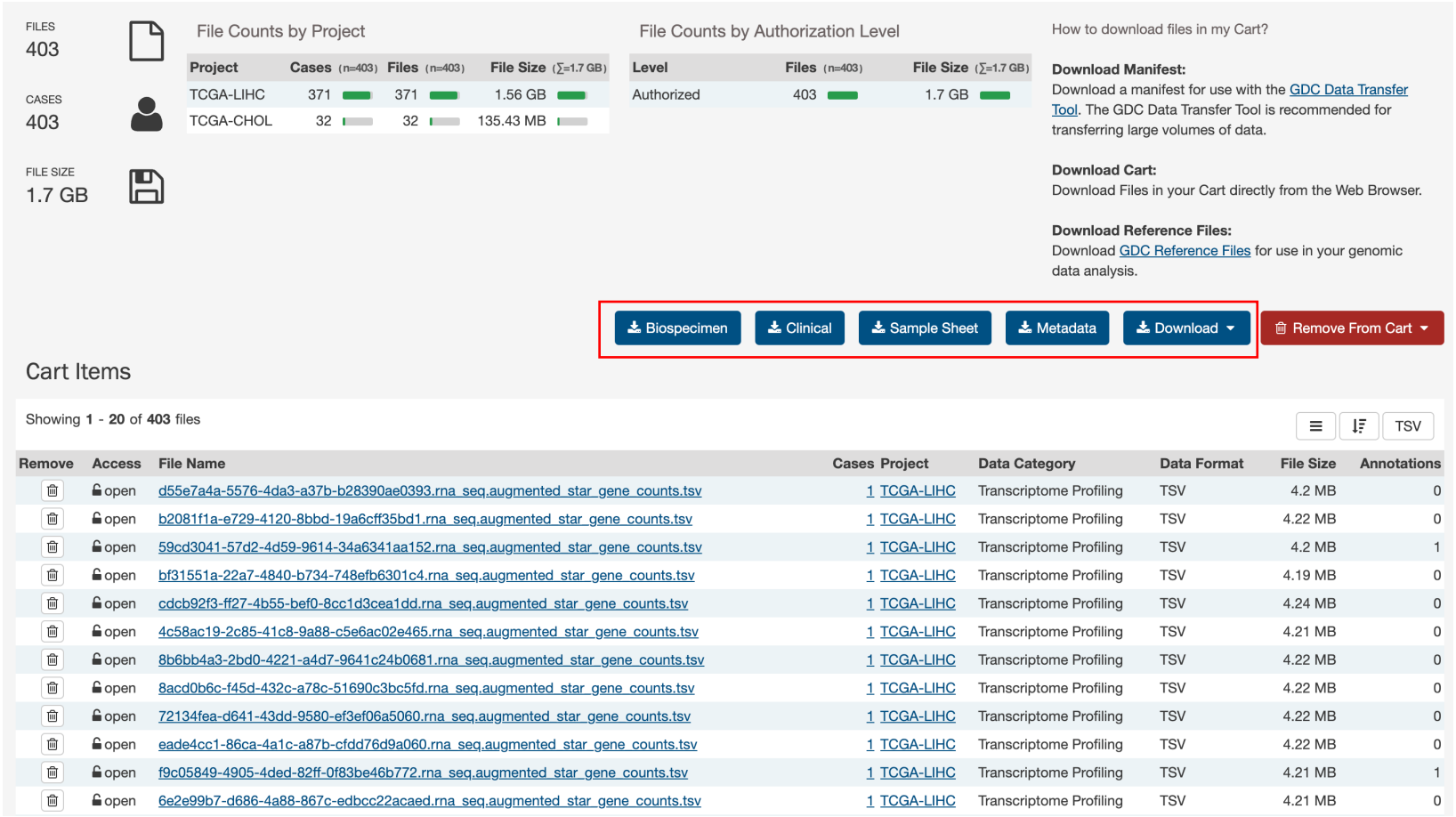


Data Download from GDC

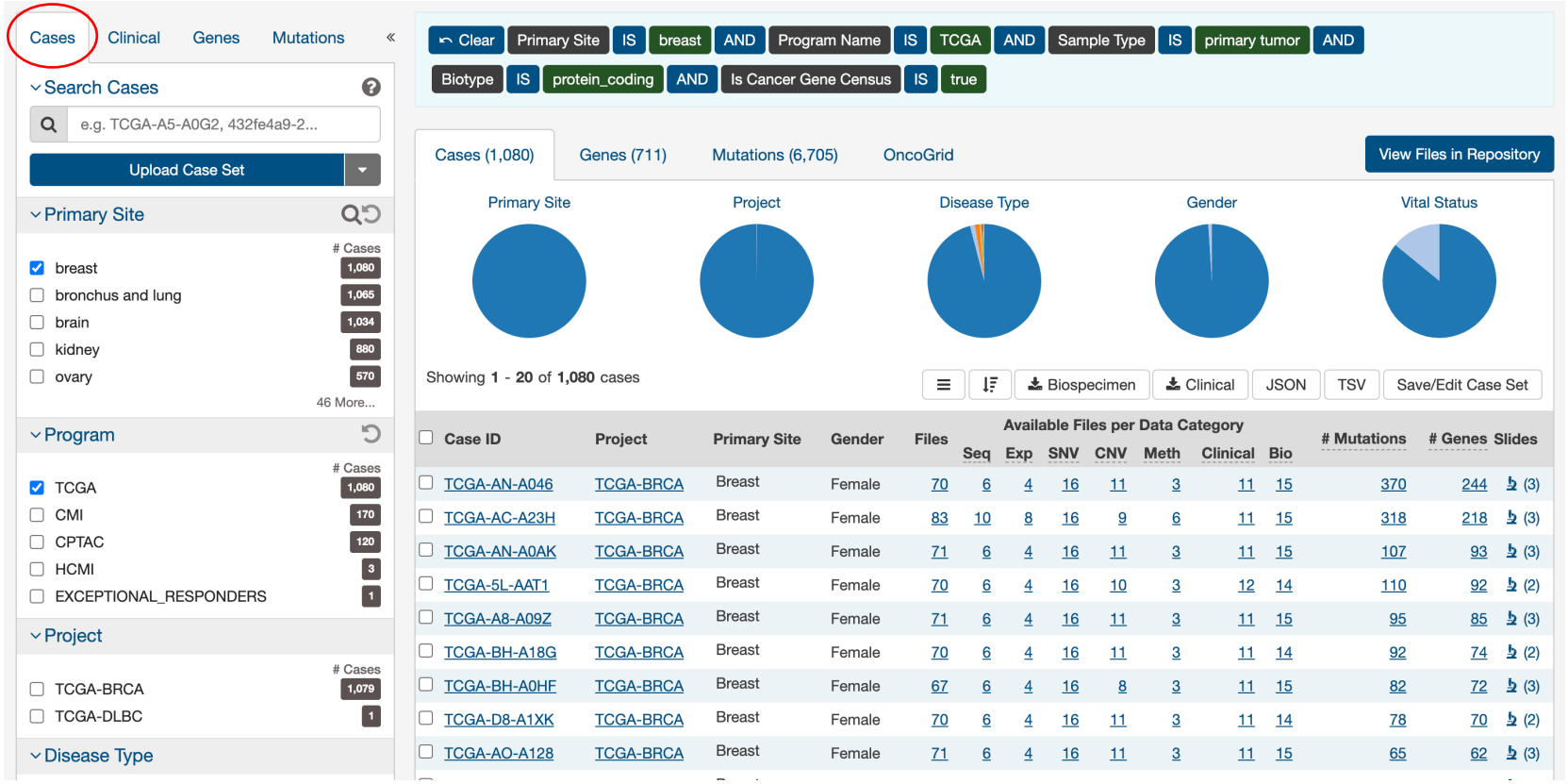


Data Download from GDC

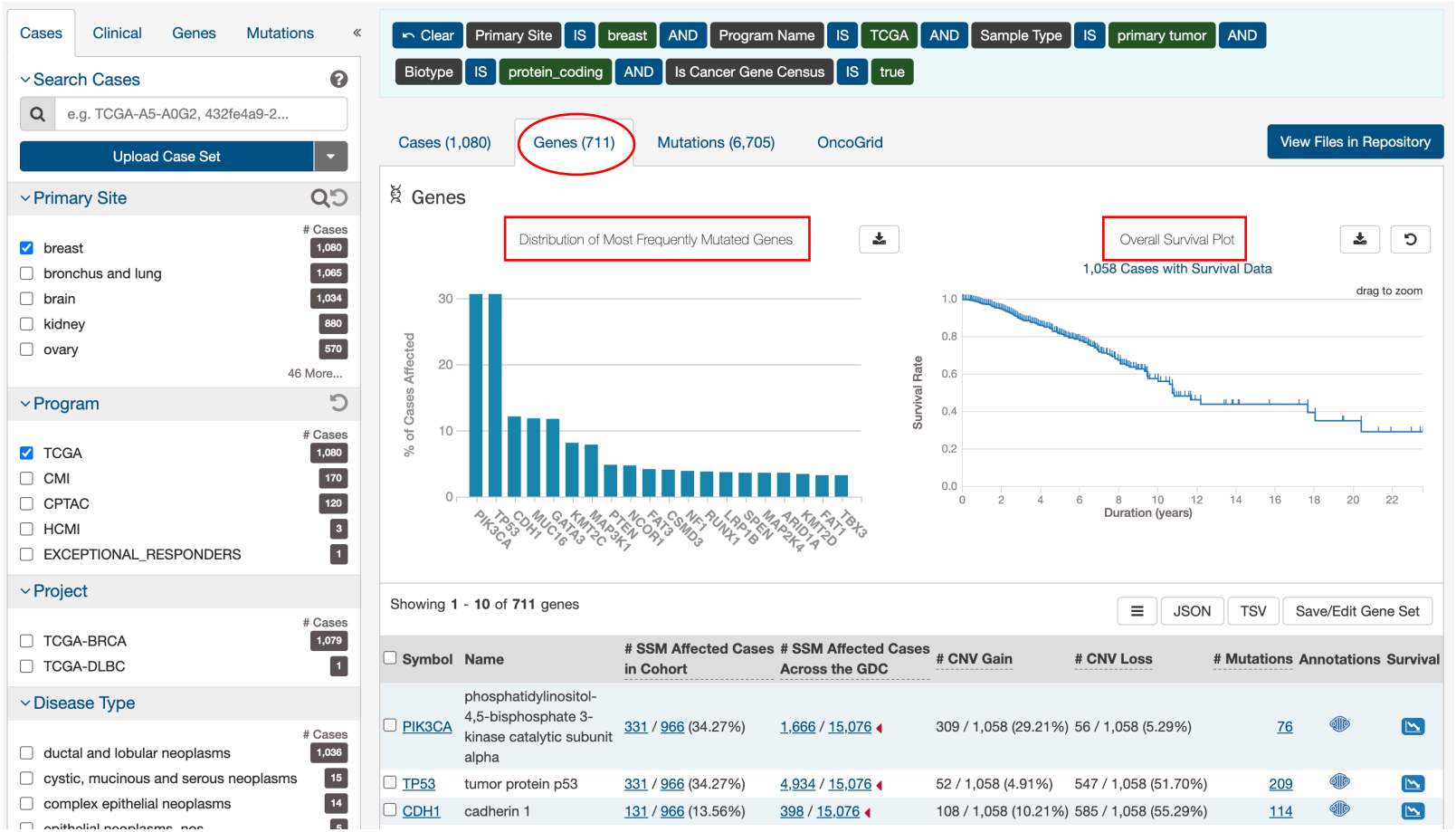
Data Download from GDC



Gene Mining in GDC



Gene Mining in GDC



Gene Mining in GDC

Cases (1,080) | Genes (711) | Mutations (6,705) | OncoGrid

View Files in Repository

OncoGrid

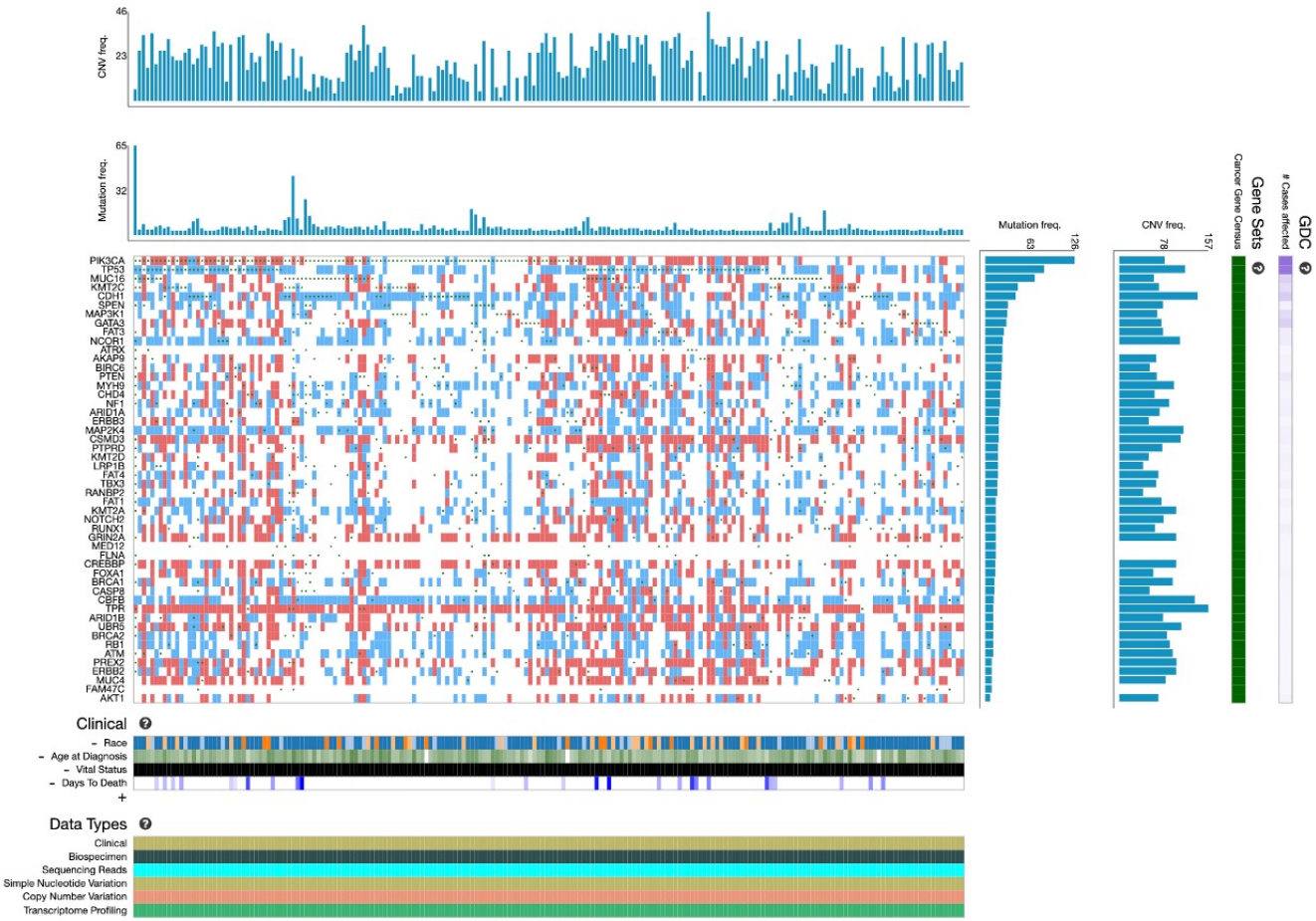
200 Most Mutated Cases and Top 50 Mutated Genes By SSM

Mutations

- [x] Show Mutations
  - [x] Missense
  - [x] Start Lost
  - [x] Stop Gained
  - [x] Frameshift
  - [x] Stop Lost

CNV Changes

- [x] Show Copy Number Variations
  - [x] Loss
  - [x] Gain



Break

Tutorial:
Galaxy project

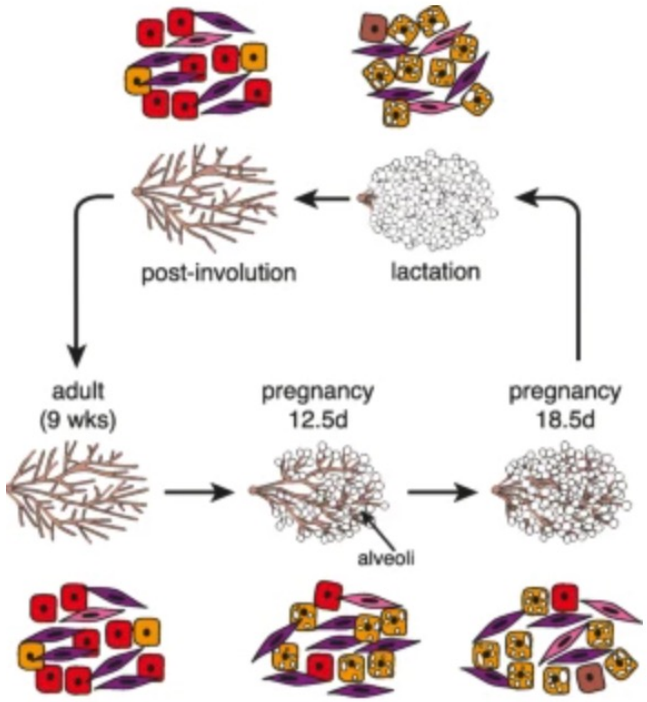
Download data

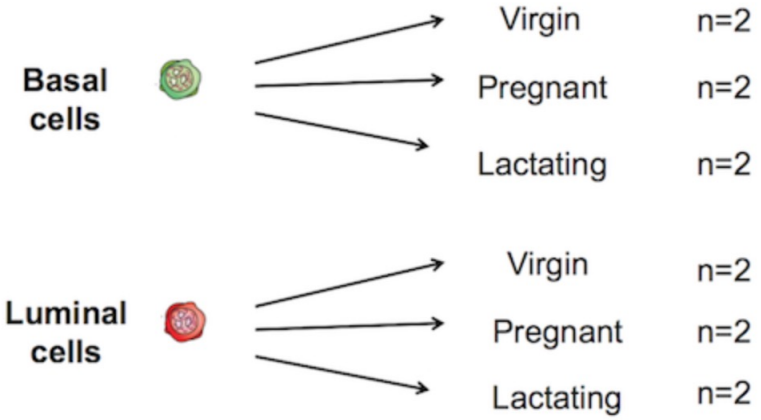
Mus musculus RNA-Seq data

- Two types, three conditions.
  - Basal, luminal cells.
  - Virgin, pregnant, lactate.
- Two replicates.
- Download at https://bit.ly/3OnLTCq

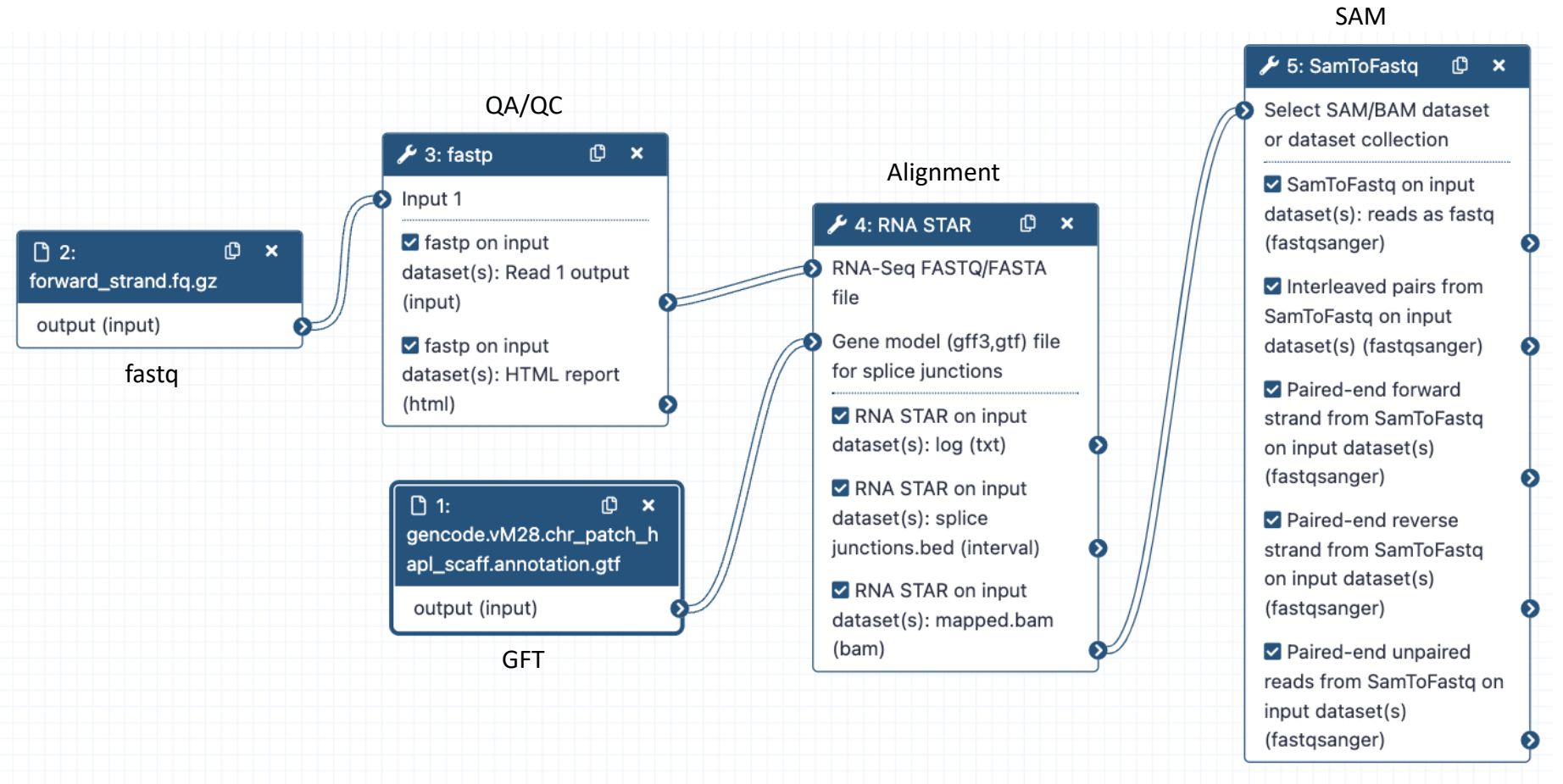
Reference database

- Mouse genome.
- Mouse transcriptome.
- Hallmark gene sets.
- Download at https://bit.ly/3UUCKUr

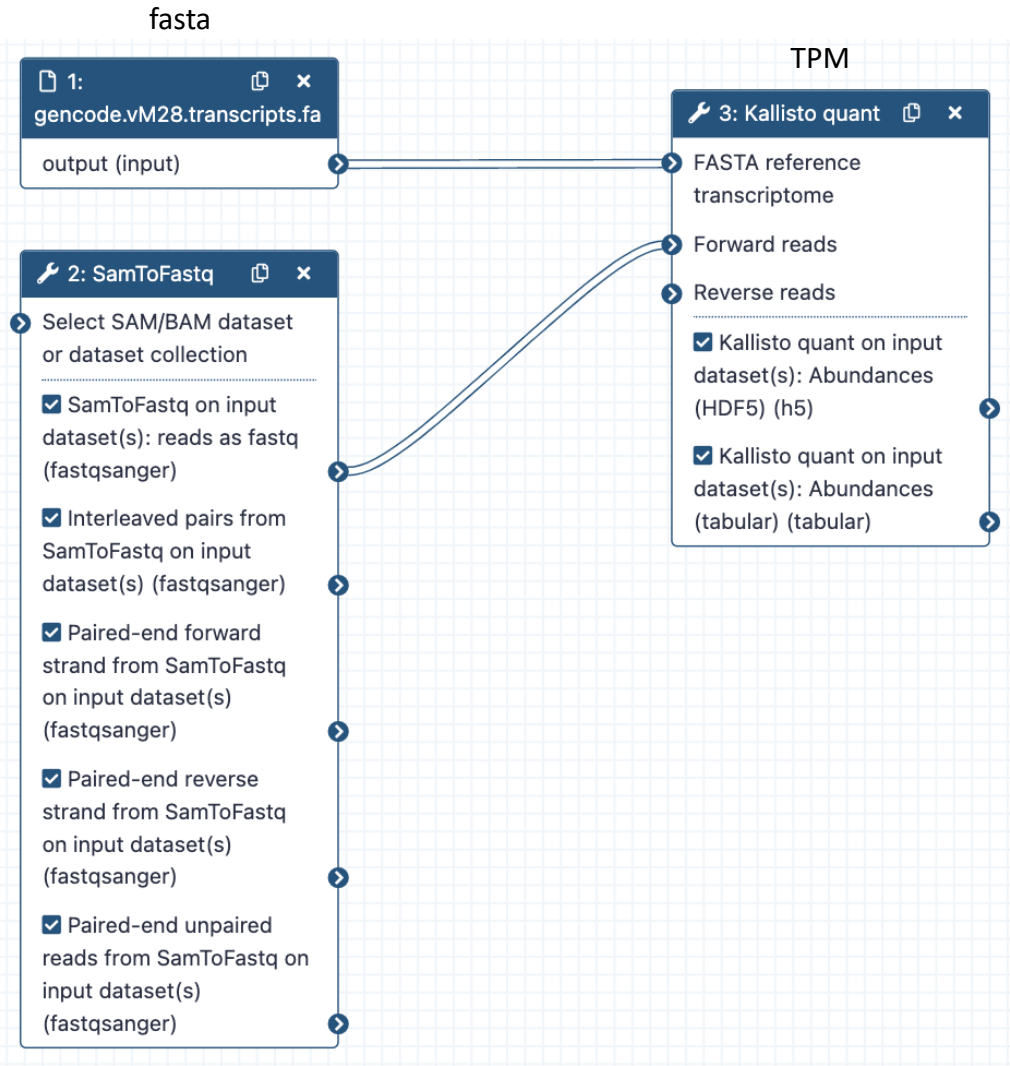




STAR alignment workflow



Kallisto quantification workflow

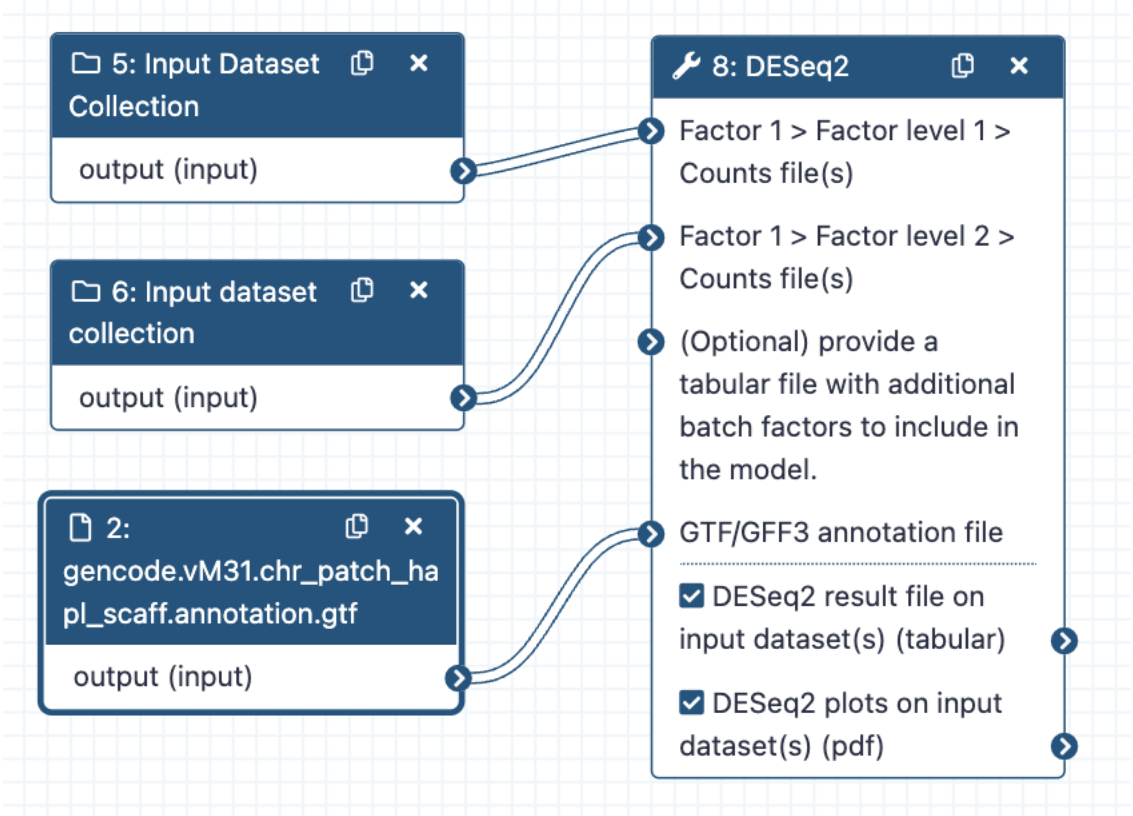


DEG using DESeq2 workflow

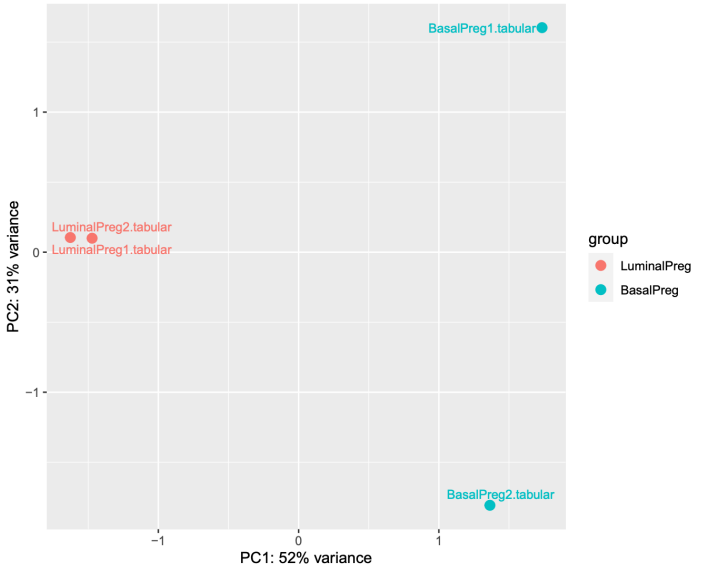
Data collection A
(output from Kallisto)

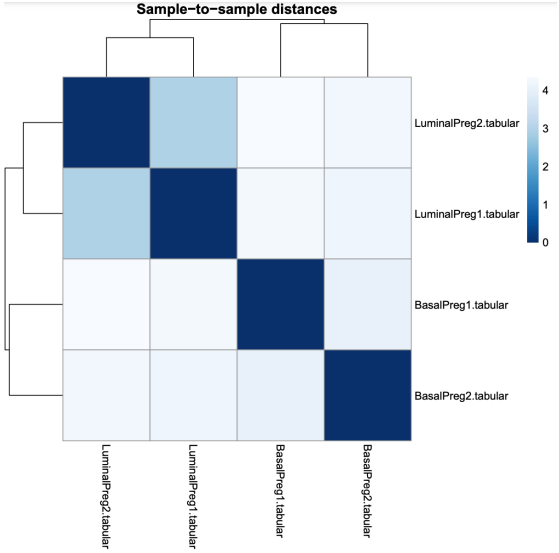
Data collection B
(output from Kallisto)

Reference genome
(GFT format)



DEG results: basal vs. luminal mammary cells in pregnant mice





GeneID	Base mean	log2(FC)	StdErr	Wald-Stats	P-value
ENSMUSG00000063157.10	22.7053166147197	-4.82006060543848	0.870801517979073	-5.53520004951842	3.10873323980544e-08
ENSMUSG00000061937.8	19.0770163172454	-4.53054003503599	0.879112516114901	-5.15353831504755	2.55616816979082e-07
ENSMUSG00000070702.10	12.5885026345625	-3.93134050235266	0.914268536648597	-4.29998446273088	1.70810084454306e-05
ENSMUSG00000032554.16	8.28597383794684	-5.37544553758729	1.32438171526543	-4.05883400203088	4.9318355241298e-05
ENSMUSG00000035783.10	6.09107459376949	5.15837631229398	1.41015102798059	3.65803109733646	0.000254160215059699
ENSMUSG00000018830.11	5.28307026982656	4.88146247165951	1.46694178029199	3.32764567567765	0.000875831842578717
ENSMUSG00000000001.5	0.275301258287342	0.856584944313586	1.85637240555842	0.461429474898877	0.644490508794769
ENSMUSG00000000058.7	0.291953014585935	0.829860348660042	1.85273328853974	0.447911393287541	0.65421714549303
ENSMUSG00000000085.17	0.291953014585935	0.829860348660042	1.85273328853974	0.447911393287541	0.65421714549303
ENSMUSG00000000303.13	0.441691506920243	-0.913681174625269	1.83380113775239	-0.498244414737974	0.618311780919906

Q&A

Patipark.k@chula.ac.th