

Big Data for Cancer Research

Patipark Kueanjinda, Ph.D.

Department of Microbiology
Faculty of Medicine, Chulalongkorn University

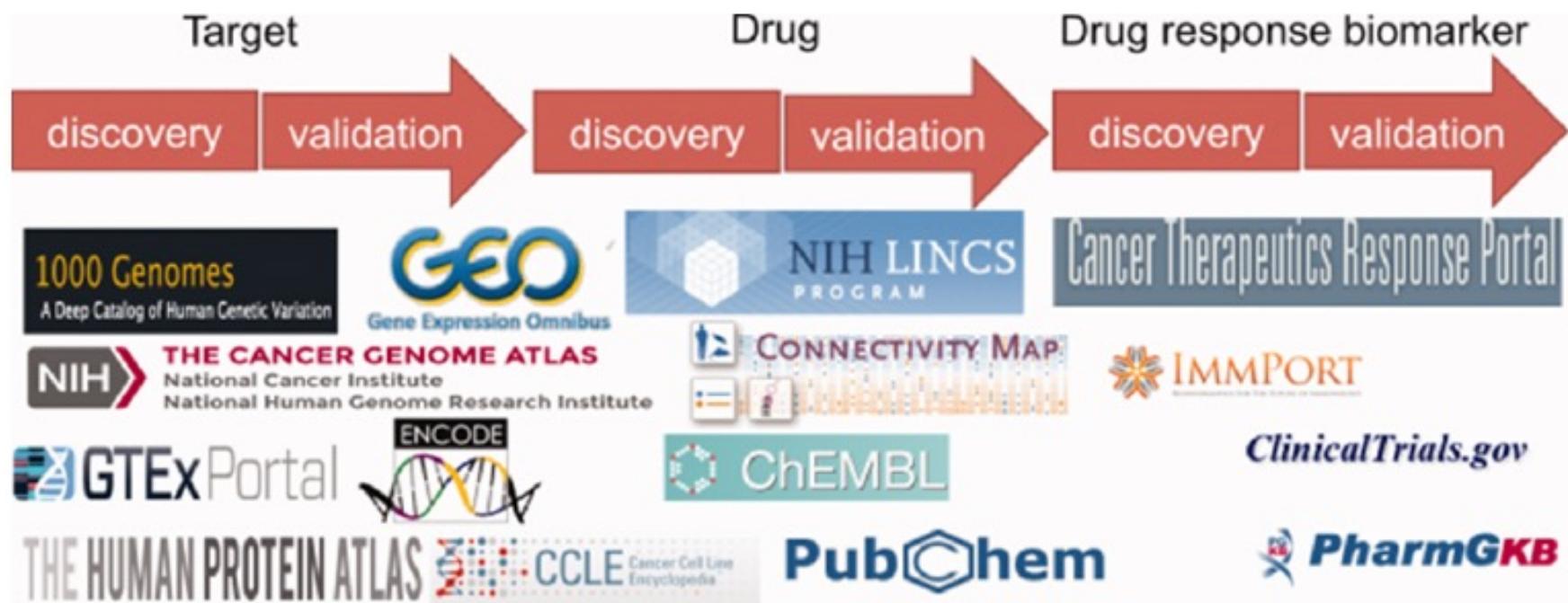
Objectives

- 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

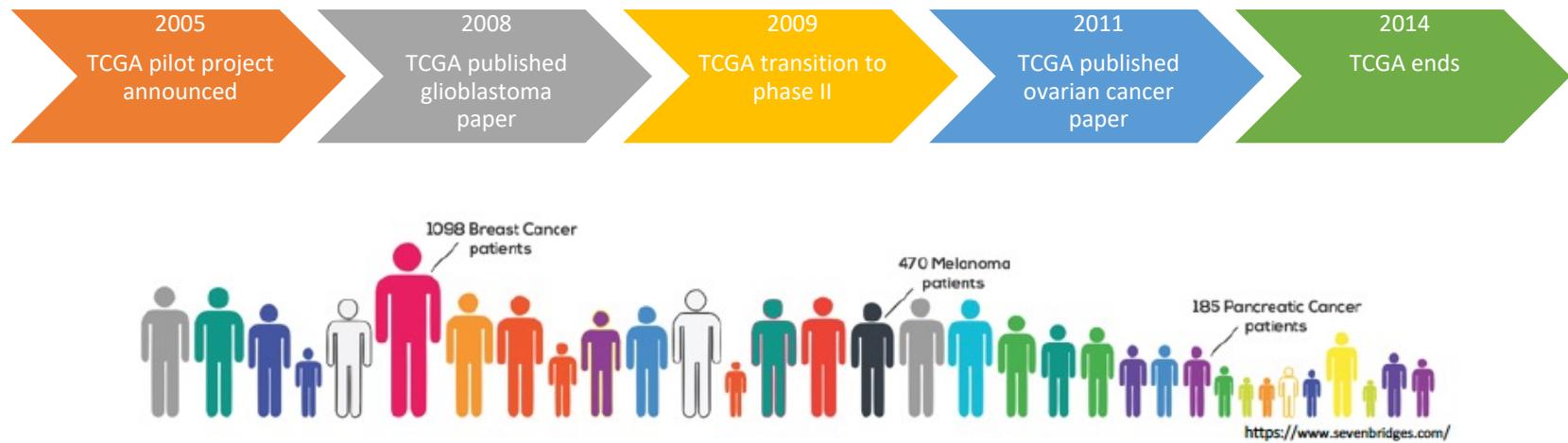
- R program
- Rstudio
- Galaxy subscription

Public Databases



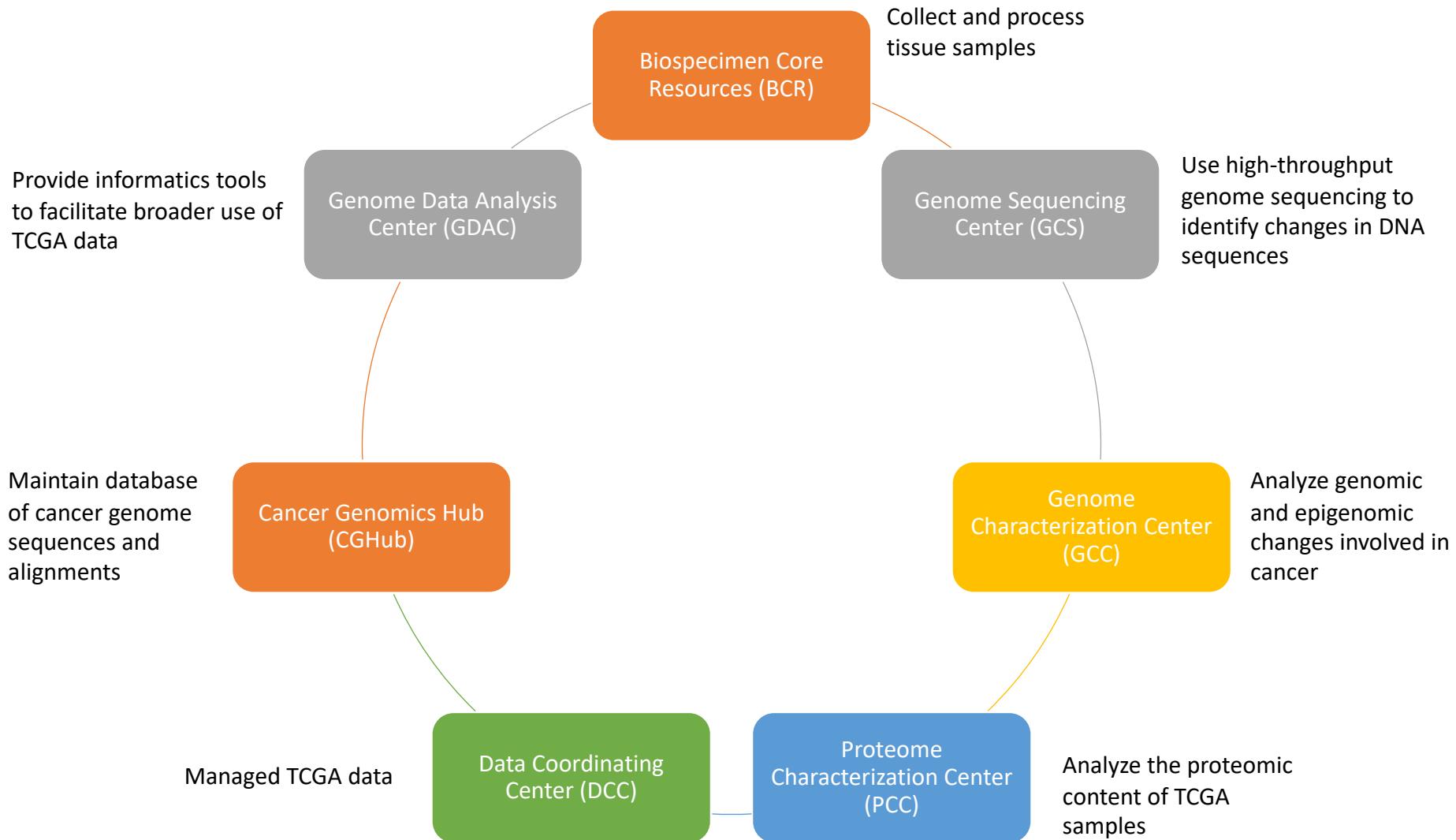
The Cancer Genome Atlas

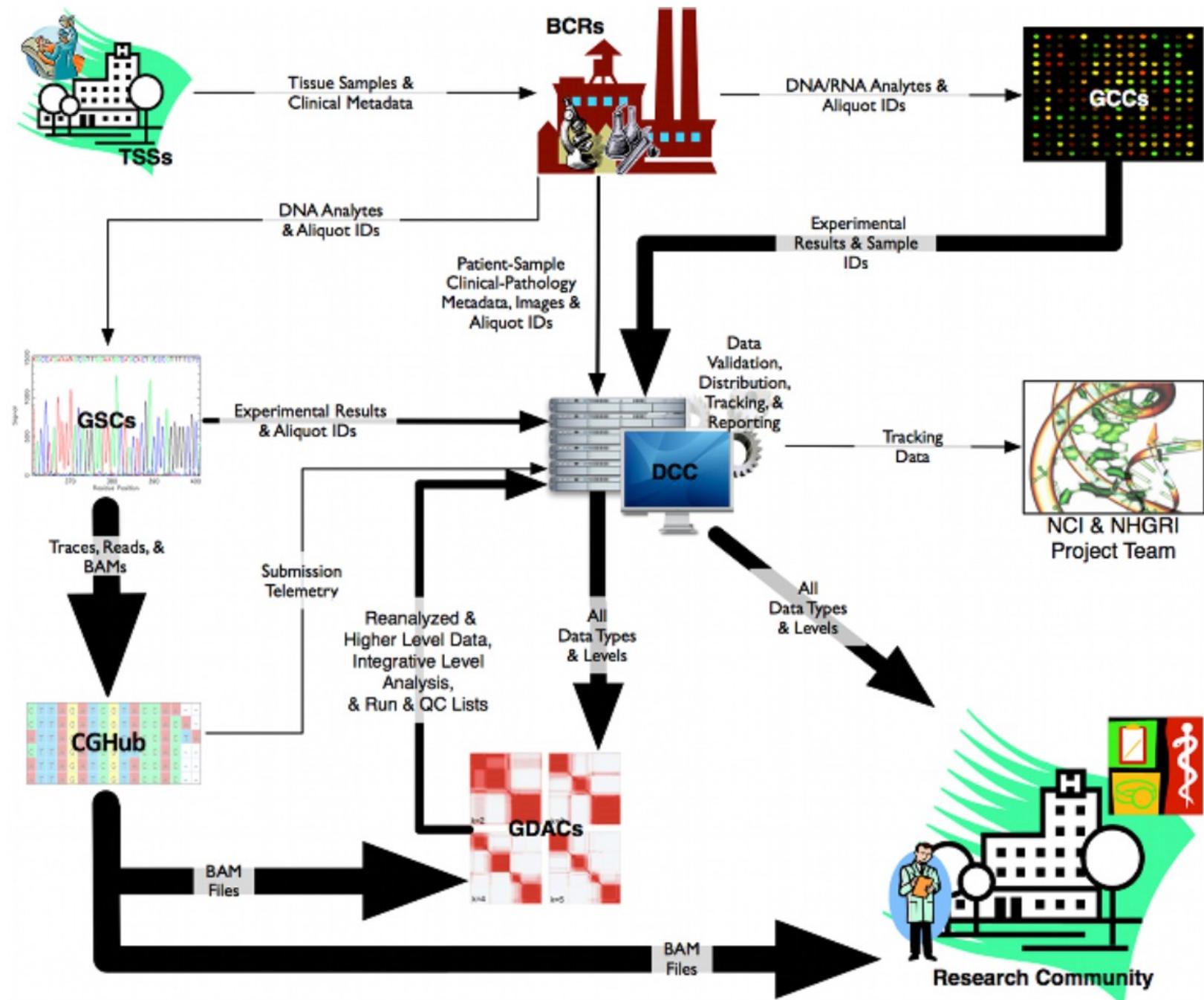
- National Cancer Institute + National Human Genome Research Institute.
- Increase scientific understanding of the molecular basis of cancer and apply this information to improve the ability to diagnose, treat, and prevent cancer.
- Develop a complete atlas of all genomic alterations involved in cancer.



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

Major TCGA Research Components





TCGA Data Outputs

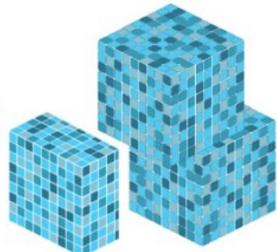
NATIONAL CANCER INSTITUTE THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

TCGA produced over

2.5
PETABYTES

of data



TCGA data describes



33
DIFFERENT
TUMOR TYPES

...including

10
RARE
CANCERS

...based on paired tumor and normal tissue sets collected from



11,000
PATIENTS

...using

7
DIFFERENT
DATA TYPES



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	-Normalized 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 heterozygosity calls for a sample region -Expression signal of a gene for a sample -Genomic copy-number data
4	Summary/Region of Interest	-Quantified association 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)

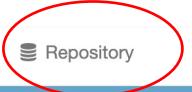
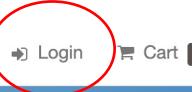
Where to download data?

- 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

NATIONAL CANCER INSTITUTE
GDC Data Portal

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

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

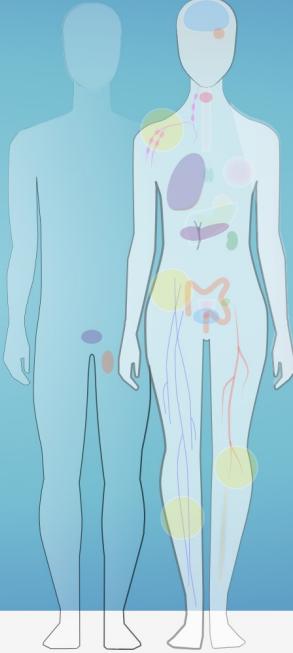
Get Started by Exploring:

Projects Exploration Analysis Repository 

Q 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



Cases by Major Primary Site

Primary Site	Cases (in 1000s)
Adrenal Gland	1
Bile Duct	1
Bladder	2
Bone	1
Bone Marrow	11
Brain	2
Breast	9
Cervix	1
Colorectal	8
Esophagus	1
Eye	1
Head and Neck	3
Kidney	3
Liver	1
Lung	12
Lymph Nodes	1
Nervous System	4
Ovary	3
Pancreas	2
Pleura	1
Prostate	2
Skin	3
Soft Tissue	1
Stomach	1
Testis	1
Thymus	1
Thyroid	1
Uterus	2

1 2 3 4 5 6 7 8 9 10 11 12 (in 1000s)

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:



Files Cases < Add a File Filter

Search Files e.g. 142682.bam, 4f6e2e7a-b...

Start searching by selecting a facet Advanced Search

Files (1,003,747) Cases (88,991) Add All Files to Cart Manifest View 88,991 Cases in Exploration View Images

Primary Site Project Data Category Data Type Data Format

Show More

Showing 1 - 20 of 1,003,747 files 6.78 PB

Access File Name Cases Project Data Category Data Format File Size Annotations

	controlled 78694126-801e-4694-9663-b9c10f4f10ff.targeted_sequencing.MuSE.aliquot.maf.gz	1 CGCI-HTMCP-CC	Simple Nucleotide Variation MAF	146.04 KB	0	
	controlled b24dfe80-71c1-4497-bab6-348fe5859dec.targeted_sequencing.Pindel.aliquot.maf.gz	1 CGCI-HTMCP-CC	Simple Nucleotide Variation MAF	44.87 KB	0	
	controlled 052283e5-ed88-46be-8c8a-4f918f5f9a7a.targeted_sequencing.VarScan2.aliquot.maf.gz	1 CGCI-HTMCP-CC	Simple Nucleotide Variation MAF	212.63 KB	0	
	controlled CGCI HTMCP-CC.dbaa74df-0eea-48dc-b323-cf2632c0ce0f.targeted_sequencing.VarScan2.somatic_annotation.vcf.gz	1 CGCI-HTMCP-CC	Simple Nucleotide Variation VCF	337.83 KB	0	
	controlled 07ba488f-2f09-4fb7-8b52-7aee17c29bb3_targeted_sequencing_gdc_realm.bam	1 CGCI-HTMCP-CC	Sequencing Reads	BAM	35.65 GB	0
	controlled ce4a23b3-a7d9-4100-ac2c-3dc7c6a75cad_targeted_sequencing_gdc_realm.bam	1 CGCI-HTMCP-CC	Sequencing Reads	BAM	21.63 GB	0
	controlled 77fdb759-4f1e-418a-8720-1d44d7da4248_mirnas_eq_gdc_realm.bam	1 CGCI-HTMCP-CC	Sequencing Reads	BAM	276.29 MB	0

Data Category

- simple nucleotide variation # Files 365,968
- copy number variation # Files 165,903
- sequencing reads # Files 149,745
- structural variation # Files 86,411
- transcriptome profiling # Files 81,096

6 More...

Data Type

- Annotated Somatic Mutation # Files 152,422
- Aligned Reads # Files 149,745
- Raw Simple Somatic Mutation # Files 94,338
- Transcript Fusion # Files 89,543
- Masked Annotated Somatic Mutation # Files 44,755

25 More...

Experimental Strategy

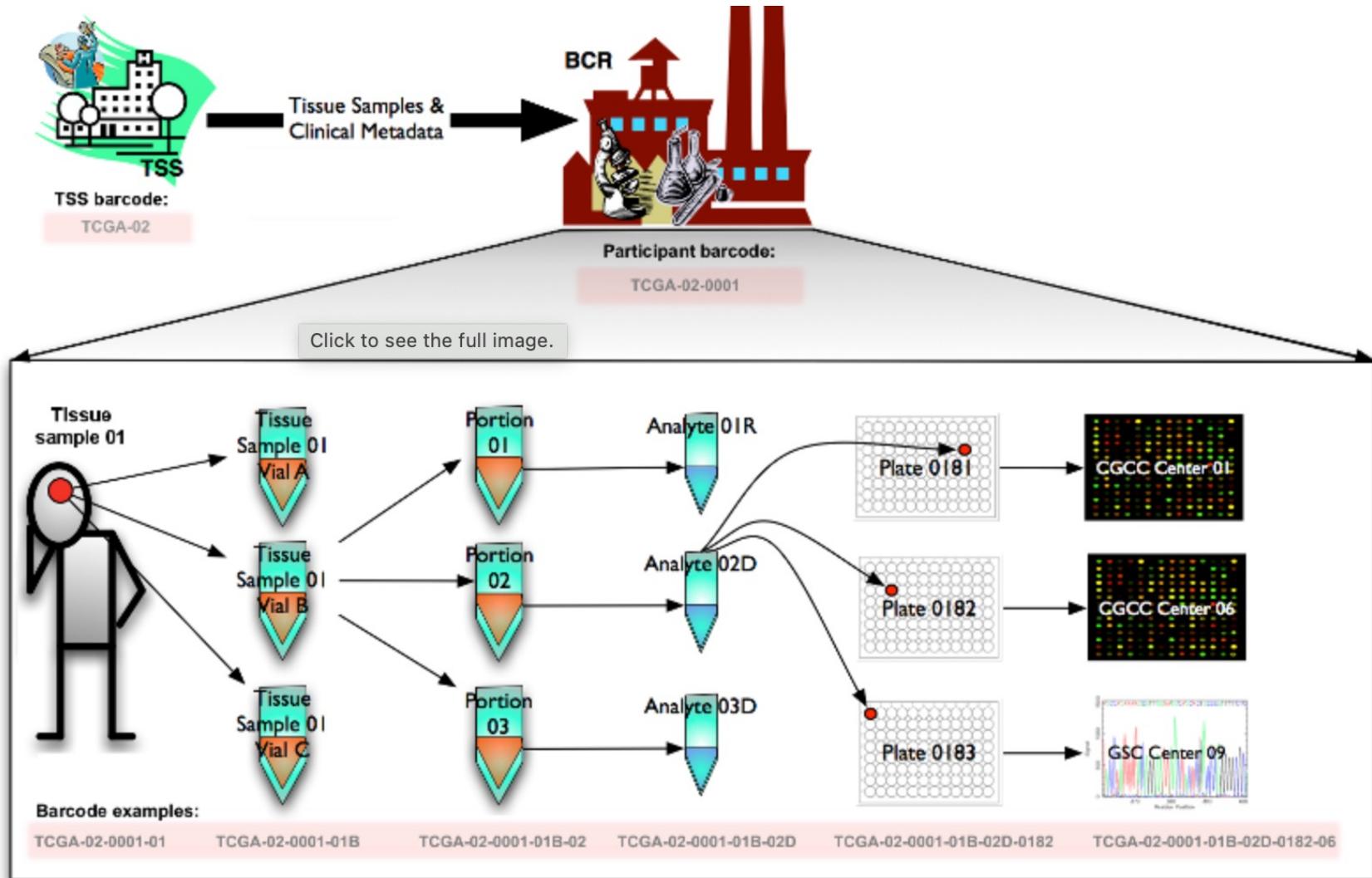
- WXS # Files 265,553
- RNA-Seq # Files 201,394
- Genotyping Array # Files 147,734
- Targeted Sequencing # Files 140,457
- WGS # Files 54,086

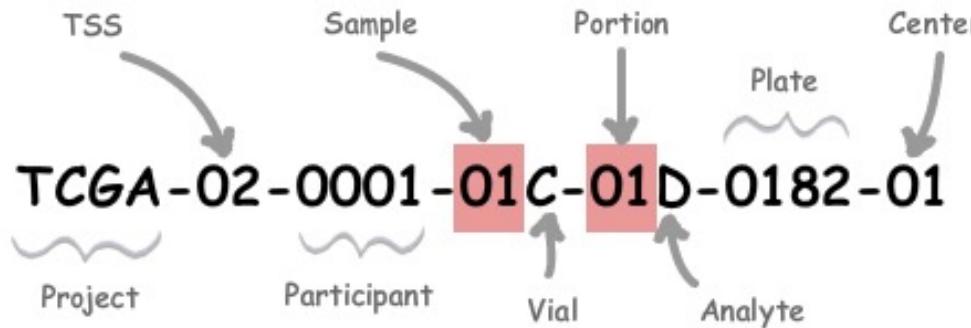
7 More...

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 Biospecimen 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

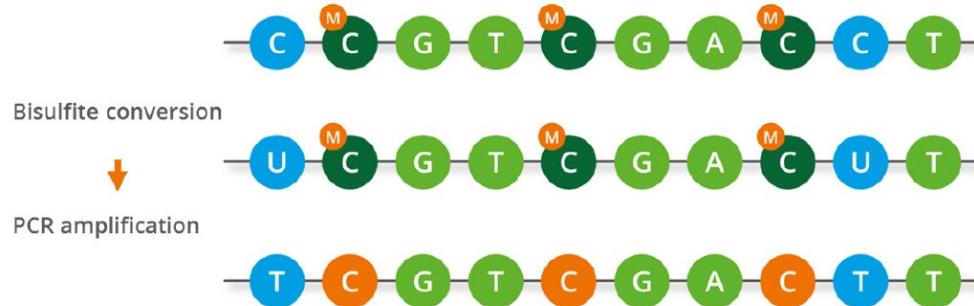
- miRNA-seq contains small non-coding RNA
 - BAM files with raw sequences
 - TXT files with signals at miRNA regions, or isoforms
- mRNA-seq contains transcriptome or expressed genes
 - BAM or FASTQ files contain raw sequences
 - TXT files with expression signals in exon regions, gene regions, splice 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 reveals 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

The screenshot shows the GDC (Genomic Data Commons) search interface for a dataset related to liver and intrahepatic bile ducts (TCGA-CHOL). The interface is divided into two main sections: a left sidebar for filtering and a right main area for results.

Left Sidebar (Filtering):

- Top Navigation:** Files (selected) and Cases (highlighted with a red circle).
- Search Bar:** Add a Case/Biospecimen Filter (e.g., TCGA-A5-A0G2, 432fe4a9-2...).
- Search Cases:** Search bar (e.g., TCGA-A5-A0G2, 432fe4a9-2...) and Upload Case Set button.
- Case ID:** Search bar (e.g., TCGA-DD*, *DD*, TCGA-DD-AAVP) and Go! button.
- Primary Site:** Filtered to liver and intrahepatic bile ducts (44 cases). Other options: other and unspecified parts of biliary tract (6), gallbladder (1).
- Program:** Filtered to TCGA (44 cases). Other option: TCGA-LIHC (377).
- Project:** Filtered to TCGA-CHOL (44 cases). Other options: TCGA-LIHC (377).
- Disease Type:** Filtered to adenomas and adenocarcinomas (44 cases).
- Gender:** No filter applied.

Right Main Area (Results):

- Search Bar:** Clear, Primary Site IS liver and intrahepatic bile ducts AND Program Name IS TCGA AND Project Id IS TCGA-CHOL. Advanced Search button.
- Summary Metrics:** Files (2,212) and Cases (44).
- Visualizations:** Primary Site (blue circle), Project (blue circle), Data Category (pie chart), Data Type (pie chart), and Data Format (pie chart).
- File Listing:** Showing 1 - 20 of 2,212 files (21.6 TB).

Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
controlled	TCGA-CHOL_189102f7-4993-4bfc-a83f-dd2b01bece62_star_fusion.rna_fusion.bedpe	1	TCGA-CHOL	Structural Variation	BEDPE	229 B	0
open	d94cd170-9dff-4b44-a7a8-6d1c277b0f8b.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-CHOL	Transcriptome Profiling	TSV	4.24 MB	0
controlled	d94cd170-9dff-4b44-a7a8-6d1c277b0f8b.rna_seq.transcriptome.gdc_realm.bam	1	TCGA-CHOL	Sequencing Reads	BAM	12.38 GB	0
controlled	7ac36191-742e-4c7c-8fee-e1aa871fc25.wxs.muse.raw_somatic_mutation.vcf.gz	1	TCGA-CHOL	Simple Nucleotide Variation VCF	VCF	25.61 KB	0
controlled	767ad280-026f-4967-aae5-89c218296e26.wxs.MuSE.aliquot.maf.gz	1	TCGA-CHOL	Simple Nucleotide Variation MAF	MAF	113.22 KB	0
controlled	d94cd170-9dff-4b44-a7a8-6d1c277b0f8b.rna_seq.chimeric.gdc_realm.bam	1	TCGA-CHOL	Sequencing Reads	BAM	56.27 MB	0

Data Download from GDC

The screenshot shows the GDC Data Explorer interface. The top navigation bar has two tabs: 'Files' (circled in red) and 'Cases'. Below the tabs is a search bar with placeholder text 'e.g. 142682.bam, 4f6e2e7a-b...'. To the right of the search bar is a complex search query consisting of multiple AND clauses: 'Primary Site IS liver and intrahepatic bile ducts AND Program Name IS TCGA AND Project Id IS TCGA-CHOL AND Workflow Type IS STAR - Counts AND Data Category IS transcriptome profiling AND Data Type IS Gene Expression Quantification AND Experimental Strategy IS RNA-Seq'. There is also a link to 'Advanced Search'. On the left side, there are several filter panels: 'Data Category' (transcriptome profiling, 40 files), 'Data Type' (Gene Expression Quantification, 40 files; Splice Junction Quantification, 40 files), 'Experimental Strategy' (RNA-Seq, 40 files), 'Workflow Type' (STAR - Counts, 40 files), 'Data Format' (tsv, 40 files), 'Platform' (No data for this field), and 'Access' (open, 40 files). In the center, there are five large circular summaries: Primary Site, Project, Data Category, Data Type, and Data Format. Below these are buttons for 'Add All Files to Cart', 'Manifest', 'View 33 Cases in Exploration', and 'View Images'. At the bottom, it says 'Showing 1 - 20 of 40 files' and '169.04 MB'. A table lists the first few files: d94cd170-9dff-4b44-a7a8-6d1c277b0f8b.rna_seq.augmented_star_gene_counts.tsv, 1393e42f-316e-4c19-bf89-cad57614234b.rna_seq.augmented_star_gene_counts.tsv, 043fd1aa-cb19-42ad-b3e5-793e8875a1d6.rna_seq.augmented_star_gene_counts.tsv, 40538197-4671-4420-928e-7262cf452396.rna_seq.augmented_star_gene_counts.tsv, 73ca8bc8-2829-4752-a21c-abea19d5155e.rna_seq.augmented_star_gene_counts.tsv, and 3d6663aa-cf98-48ea-9c49-fa72a203948c.rna_seq.augmented_star_gene_counts.tsv. The table includes columns for Access, File Name, Cases, Project, Data Category, Data Format, File Size, and Annotations.

Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
open	d94cd170-9dff-4b44-a7a8-6d1c277b0f8b.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-CHOL	Transcriptome Profiling	TSV	4.24 MB	0
open	1393e42f-316e-4c19-bf89-cad57614234b.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-CHOL	Transcriptome Profiling	TSV	4.24 MB	0
open	043fd1aa-cb19-42ad-b3e5-793e8875a1d6.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-CHOL	Transcriptome Profiling	TSV	4.2 MB	0
open	40538197-4671-4420-928e-7262cf452396.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-CHOL	Transcriptome Profiling	TSV	4.24 MB	0
open	73ca8bc8-2829-4752-a21c-abea19d5155e.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-CHOL	Transcriptome Profiling	TSV	4.24 MB	0
open	3d6663aa-cf98-48ea-9c49-fa72a203948c.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-CHOL	Transcriptome Profiling	TSV	4.24 MB	0

Data Download from GDC

Files **Cases** Browse Annotations

[Reset](#) | [Add a Case/Biospecimen Filter](#)

Samples Sample Type

- primary tumor # Cases 404
- blood derived normal 355
- solid tissue normal 100
- recurrent tumor 2

Search Cases

e.g. TCGA-A5-A0G2, 432fe4a9-2...

[Upload Case Set](#)

Case ID

eg. TCGA-DD*, *DD*, TCGA-DD-AAVP [Go!](#)

Primary Site

- liver and intrahepatic bile ducts # Cases 404
- other and unspecified parts of biliary tract 2
- gallbladder 1

Program

- TCGA # Cases 404

Project

- TCGA-LIHC # Cases 371
- TCGA-CHOL # Cases 33

Advanced Search

Clear Primary Site IS liver and intrahepatic bile ducts AND Program Name IS TCGA AND
Project Id IN (TCGA-CHOL, TCGA-LIHC) AND Sample Type IS primary tumor AND
Workflow Type IS STAR - Counts AND Data Category IS transcriptome profiling AND
Data Type IS Gene Expression Quantification AND Experimental Strategy IS RNA-Seq

Files (403) **Cases (404)** [Add All Files to Cart](#) [Manifest](#) [View 404 Cases in Exploration](#) [View Images](#)

Primary Site Project Data Category Data Type Data Format

Show More

Showing 1 - 20 of 403 files **1.7 GB**

[Access File Name](#) [Cases](#) [Project](#) [Data Category](#) [Data Format](#) [File Size](#) [Annotations](#)

Access File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
d55e7a4a-5576-4da3-a37b-b28390ae0393.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-LIHC	Transcriptome Profiling	TSV	4.2 MB	0
b2081f1a-e729-4120-8bbd-19a6cff35bd1.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-LIHC	Transcriptome Profiling	TSV	4.22 MB	0
59cd3041-57d2-4d59-9614-34a6341aa152.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-LIHC	Transcriptome Profiling	TSV	4.2 MB	1
bf31551a-22a7-4840-b734-748efb6301c4.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-LIHC	Transcriptome Profiling	TSV	4.19 MB	0
ccdb92f3-ff27-4b55-bef0-8cc1d3cea1dd.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-LIHC	Transcriptome Profiling	TSV	4.24 MB	0
4c58ac19-2c85-41c8-9a88-c5e6ac02e465.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-LIHC	Transcriptome Profiling	TSV	4.21 MB	0

Data Download from GDC

FILES
403

CASES
403

FILE SIZE
1.7 GB

File Counts by Project

Project	Cases (n=403)	Files (n=403)	File Size (Σ=1.7 GB)
TCGA-LIHC	371	371	1.56 GB
TCGA-CHOL	32	32	135.43 MB

File Counts by Authorization Level

Level	Files (n=403)	File Size (Σ=1.7 GB)
Authorized	403	1.7 GB

How to download files in my Cart?

Download Manifest:
Download a manifest for use with the [GDC Data Transfer Tool](#). The GDC Data Transfer Tool is recommended for transferring large volumes of data.

Download Cart:
Download Files in your Cart directly from the Web Browser.

Download Reference Files:
Download [GDC Reference Files](#) for use in your genomic data analysis.

Download Buttons: Biospecimen, Clinical, Sample Sheet, Metadata, Download ▾, Remove From Cart ▾

Cart Items

Showing 1 - 20 of 403 files

Remove	Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
	🔓 open	d55e7a4a-5576-4da3-a37b-b28390ae0393.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-LIHC	Transcriptome Profiling	TSV	4.2 MB	0
	🔓 open	b2081f1a-e729-4120-8bbd-19a6cff35bd1.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-LIHC	Transcriptome Profiling	TSV	4.22 MB	0
	🔓 open	59cd3041-57d2-4d59-9614-34a6341aa152.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-LIHC	Transcriptome Profiling	TSV	4.2 MB	1
	🔓 open	b31551a-22a7-4840-b734-748efb6301c4.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-LIHC	Transcriptome Profiling	TSV	4.19 MB	0
	🔓 open	cddcb92f3-ff27-4b55-bef0-8cc1d3cea1dd.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-LIHC	Transcriptome Profiling	TSV	4.24 MB	0
	🔓 open	4c58ac19-2c85-41c8-9a88-c5e6ac02e465.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-LIHC	Transcriptome Profiling	TSV	4.21 MB	0
	🔓 open	8b6bb4a3-2bd0-4221-a4d7-9641c24b0681.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-LIHC	Transcriptome Profiling	TSV	4.22 MB	0
	🔓 open	8acd0b6c-f45d-432c-a78c-51690c3bc5fd.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-LIHC	Transcriptome Profiling	TSV	4.22 MB	0
	🔓 open	72134fea-d641-43dd-9580-ef3ef06a5060.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-LIHC	Transcriptome Profiling	TSV	4.22 MB	0
	🔓 open	eade4cc1-86ca-4a1c-a87b-cfdd76d9a060.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-LIHC	Transcriptome Profiling	TSV	4.22 MB	0
	🔓 open	f9c05849-4905-4ded-82ff-0f83be46b772.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-LIHC	Transcriptome Profiling	TSV	4.21 MB	1
	🔓 open	6e2e99b7-d686-4a88-867c-edbcc22acaed.rna_seq.augmented_star_gene_counts.tsv	1	TCGA-LIHC	Transcriptome Profiling	TSV	4.21 MB	0

Gene Mining in GDC

Cases Clinical Genes Mutations <

Search Cases ?
e.g. TCGA-A5-A0G2, 432fe4a9-2...

Upload Case Set

Primary Site
breast (1,080)
bronchus and lung (1,065)
brain (1,034)
kidney (880)
ovary (570)
46 More...

Program
TCGA (1,080)
CMI (170)
CPTAC (120)
HCMI (3)
EXCEPTIONAL_RESPONDERS (1)

Project
TCGA-BRCA (1,079)
TCGA-DLBC (1)

Disease Type
TCGA-AO-A128 (1)

Clear Primary Site IS breast AND Program Name IS TCGA AND Sample Type IS primary tumor AND Biotype IS protein_coding AND Is Cancer Gene Census IS true

Cases (1,080) Genes (711) Mutations (6,705) OncoGrid View Files in Repository

Primary Site Project Disease Type Gender Vital Status

Showing 1 - 20 of 1,080 cases

Case ID	Project	Primary Site	Gender	Files	Available Files per Data Category								# Mutations	# Genes	Slides
					Seq	Exp	SNV	CNV	Meth	Clinical	Bio				
TCGA-AN-A046	TCGA-BRCA	Breast	Female	70 6 4 16 11 3 11 15	370	244	2 (3)								
TCGA-AC-A23H	TCGA-BRCA	Breast	Female	83 10 8 16 9 6 11 15	318	218	2 (3)								
TCGA-AN-A0AK	TCGA-BRCA	Breast	Female	71 6 4 16 11 3 11 15	107	93	2 (3)								
TCGA-5L-AAT1	TCGA-BRCA	Breast	Female	70 6 4 16 10 3 12 14	110	92	2 (2)								
TCGA-A8-A09Z	TCGA-BRCA	Breast	Female	71 6 4 16 11 3 11 15	95	85	2 (3)								
TCGA-BH-A18G	TCGA-BRCA	Breast	Female	70 6 4 16 11 3 11 14	92	74	2 (2)								
TCGA-BH-A0HF	TCGA-BRCA	Breast	Female	67 6 4 16 8 3 11 15	82	72	2 (3)								
TCGA-D8-A1XK	TCGA-BRCA	Breast	Female	70 6 4 16 11 3 11 14	78	70	2 (2)								
TCGA-AO-A128	TCGA-BRCA	Breast	Female	71 6 4 16 11 3 11 15	65	62	-								

Gene Mining in GDC

Cases Clinical Genes Mutations <

Search Cases e.g. TCGA-A5-A0G2, 432fe4a9-2...

Upload Case Set

Primary Site

- breast # Cases 1,080
- bronchus and lung 1,065
- brain 1,034
- kidney 880
- ovary 570
- 46 More...

Program

- TCGA # Cases 1,080
- CMI 170
- CPTAC 120
- HCMII 3
- EXCEPTIONAL_RESPONDERS 1

Project

- TCGA-BRCA # Cases 1,079
- TCGA-DLBC 1

Disease Type

- ductal and lobular neoplasms # Cases 1,036
- cystic, mucinous and serous neoplasms 15
- complex epithelial neoplasms 14
- epithelial neoplasms, nec 5

Clear Primary Site IS breast AND Program Name IS TCGA AND Sample Type IS primary tumor AND Biotype IS protein_coding AND Is Cancer Gene Census IS true

Cases (1,080) Genes (711) Mutations (6,705) Oncogrid View Files in Repository

Genes Distribution of Most Frequently Mutated Genes

Overall Survival Plot 1,058 Cases with Survival Data drag to zoom

Symbol	Name	# SSM Affected Cases in Cohort	# SSM Affected Cases Across the GDC	# CNV Gain	# CNV Loss	# Mutations	Annotations	Survival
<input type="checkbox"/> PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha	331 / 966 (34.27%)	1,666 / 15,076 ↘	309 / 1,058 (29.21%)	56 / 1,058 (5.29%)	76		
<input type="checkbox"/> TP53	tumor protein p53	331 / 966 (34.27%)	4,934 / 15,076 ↘	52 / 1,058 (4.91%)	547 / 1,058 (51.70%)	209		
<input type="checkbox"/> CDH1	cadherin 1	131 / 966 (13.56%)	398 / 15,076 ↘	108 / 1,058 (10.21%)	585 / 1,058 (55.29%)	114		

Gene Mining in GDC

breast # Cases 1,080

bronchus and lung # Cases 1,065

brain # Cases 1,034

kidney # Cases 880

ovary # Cases 570

46 More...

TCGA # Cases 1,080

CMI # Cases 170

CPTAC # Cases 120

HCMC # Cases 3

EXCEPTIONAL_RESPONDERS # Cases 1

TCGA-BRCA # Cases 1,079

TCGA-DLBC # Cases 1

Disease Type

- ductal and lobular neoplasms # Cases 1,036
- cystic, mucinous and serous neoplasms # Cases 15
- complex epithelial neoplasms # Cases 14
- epithelial neoplasms, nos # Cases 5
- adenomas and adenocarcinomas # Cases 3

5 More...

primary tumor # Cases 1,080

blood derived normal # Cases 994

solid tissue normal # Cases 160

metastatic # Cases 7

Distribution of Most Frequently Mutated Genes

Overall Survival Plot

S₁ (N = 696) - PTEN Not Mutated Cases S₂ (N = 362) - PTEN Mutated Cases

Log-Rank Test P-Value = 2.41e-2

Showing 1 - 10 of 711 genes

Symbol	Name	# SSM Affected Cases in Cohort	# SSM Affected Cases Across the GDC	# CNV Gain	# CNV Loss	# Mutations	Annotations	Survival
<input type="checkbox"/>	PIK3CA	331 / 966 (34.27%)	1,666 / 15,076 ↘	309 / 1,058 (29.21%)	56 / 1,058 (5.29%)	76		
<input type="checkbox"/>	TP53	331 / 966 (34.27%)	4,934 / 15,076 ↘	52 / 1,058 (4.91%)	547 / 1,058 (51.70%)	209		
<input type="checkbox"/>	CDH1	131 / 966 (13.56%)	398 / 15,076 ↘	108 / 1,058 (10.21%)	585 / 1,058 (55.29%)	114		
<input type="checkbox"/>	MUC16	128 / 966 (13.25%)	2,839 / 15,076 ↘	211 / 1,058 (19.94%)	149 / 1,058 (14.08%)	169		
<input type="checkbox"/>	GATA3	127 / 966 (13.15%)	384 / 15,076 ↘	281 / 1,058 (26.56%)	94 / 1,058 (8.88%)	89		
<input type="checkbox"/>	KMT2C	88 / 966 (9.11%)	1,046 / 15,076 ↘	237 / 1,058 (22.40%)	151 / 1,058 (14.27%)	97		
<input type="checkbox"/>	MAP3K1	85 / 966 (8.80%)	383 / 15,076 ↘	157 / 1,058 (14.84%)	224 / 1,058 (21.17%)	115		
<input type="checkbox"/>	PTEN	52 / 966 (5.38%)	1,228 / 15,076 ↘	98 / 1,058 (9.26%)	253 / 1,058 (23.91%)	50		

Cases (1,080)

Genes (711)

Mutations (6,705)

OncoGrid

View Files in Repository

OncoGrid

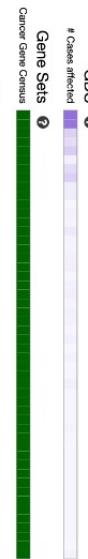
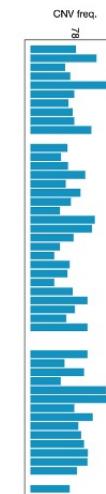
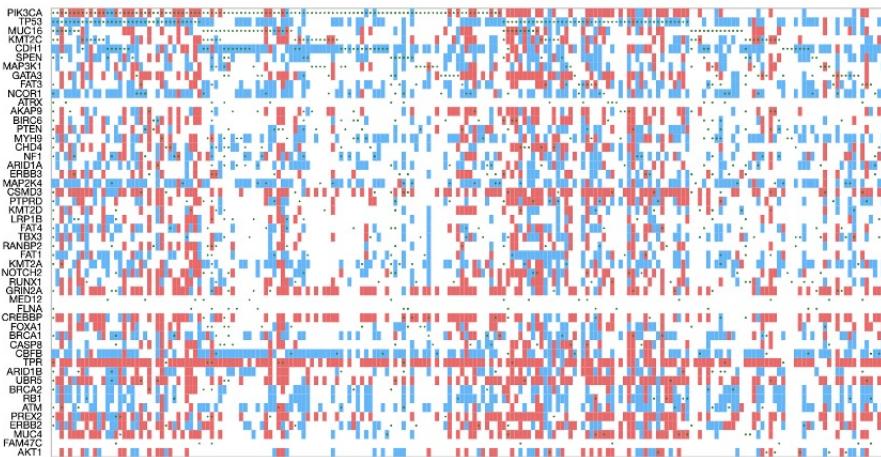
200 Most Mutated Cases and Top 50 Mutated Genes By SSM

Mutations

- Show Mutations
- Missense Start Lost Stop Gained
 Frameshift Stop Lost

CNV Changes

- Show Copy Number Variations
- Loss Gain



Clinical

- Race
- Age at Diagnosis
- Vital Status
- Days To Death
- +

Data Types

- Clinical
- Biospecimen
- Sequencing Reads
- Simple Nucleotide Variation
- Copy Number Variation
- Transcriptome Profiling

Break

Tutorial: Galaxy project

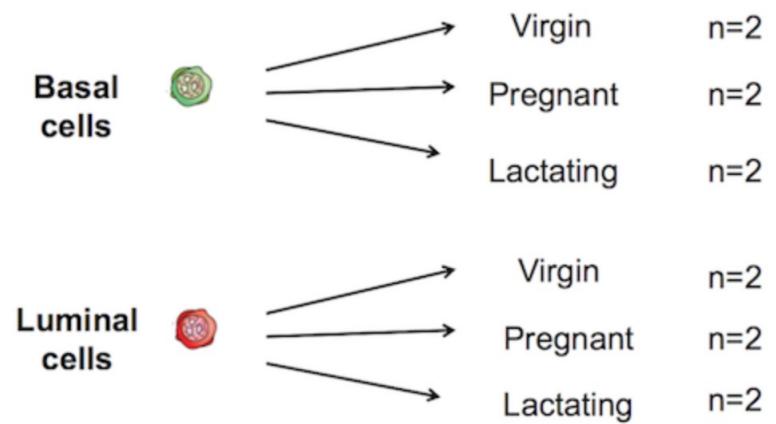
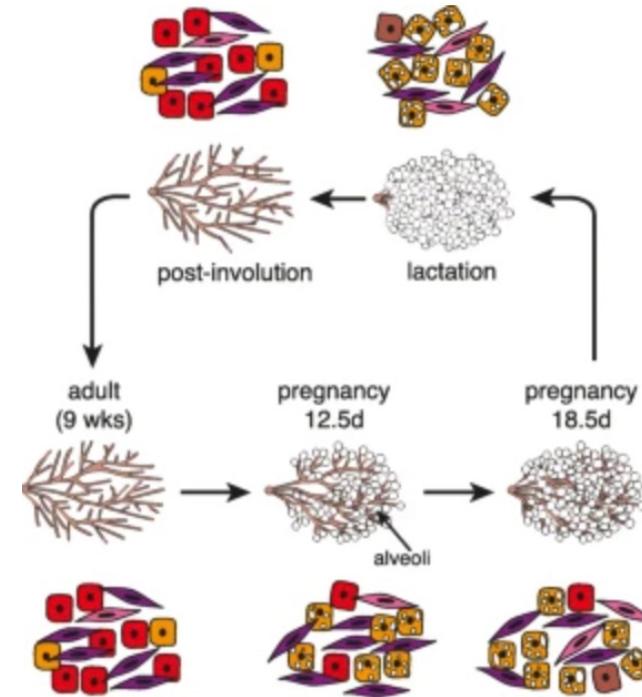
Download data

Mus musculus RNA-Seq data

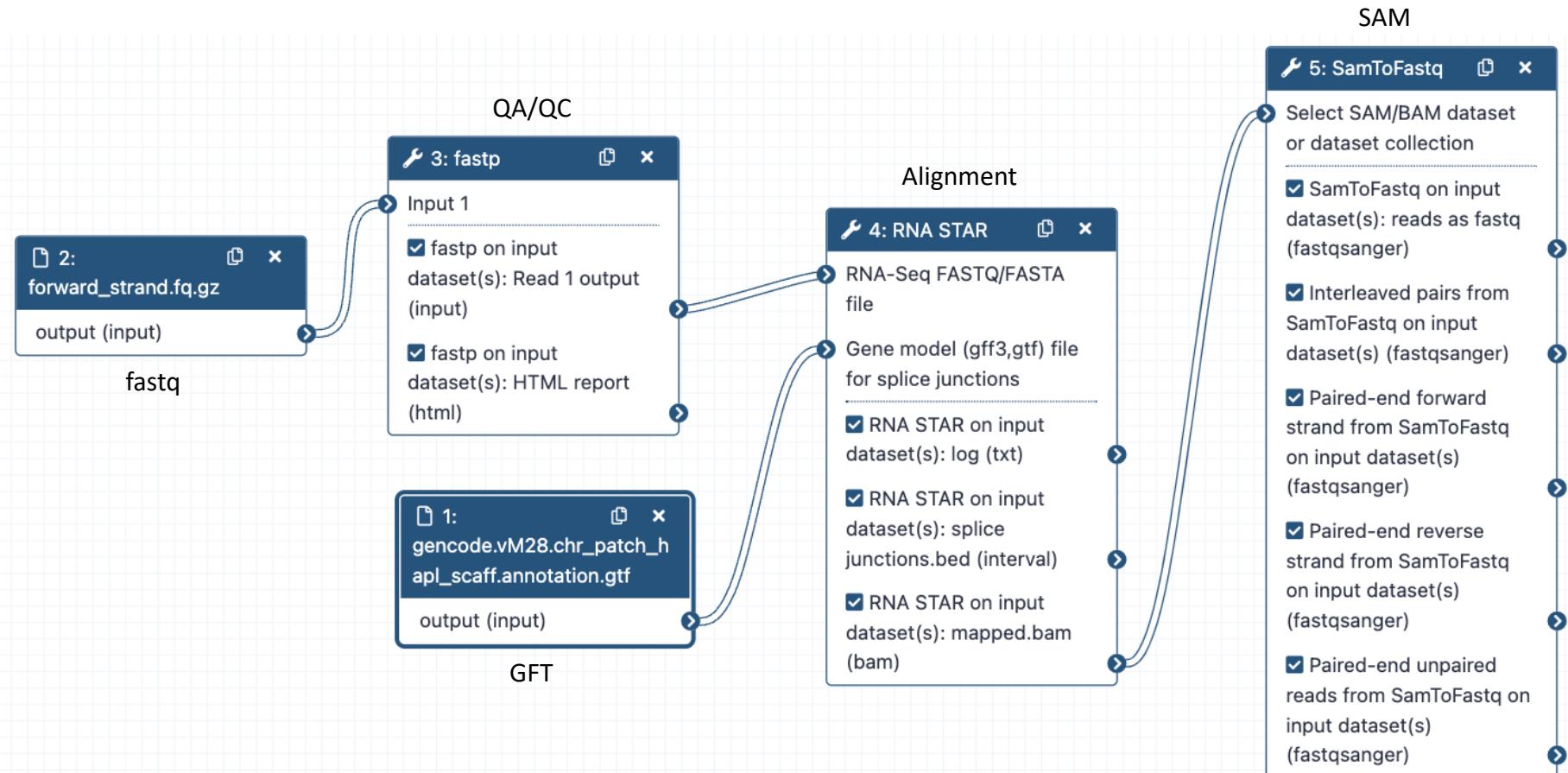
- Two types, three conditions.
 - Basal, luminal cells.
 - Virgin, pregnant, lactate.
- Two replicates.
- Download at
<https://bit.ly/3OnLTCq>

Reference database

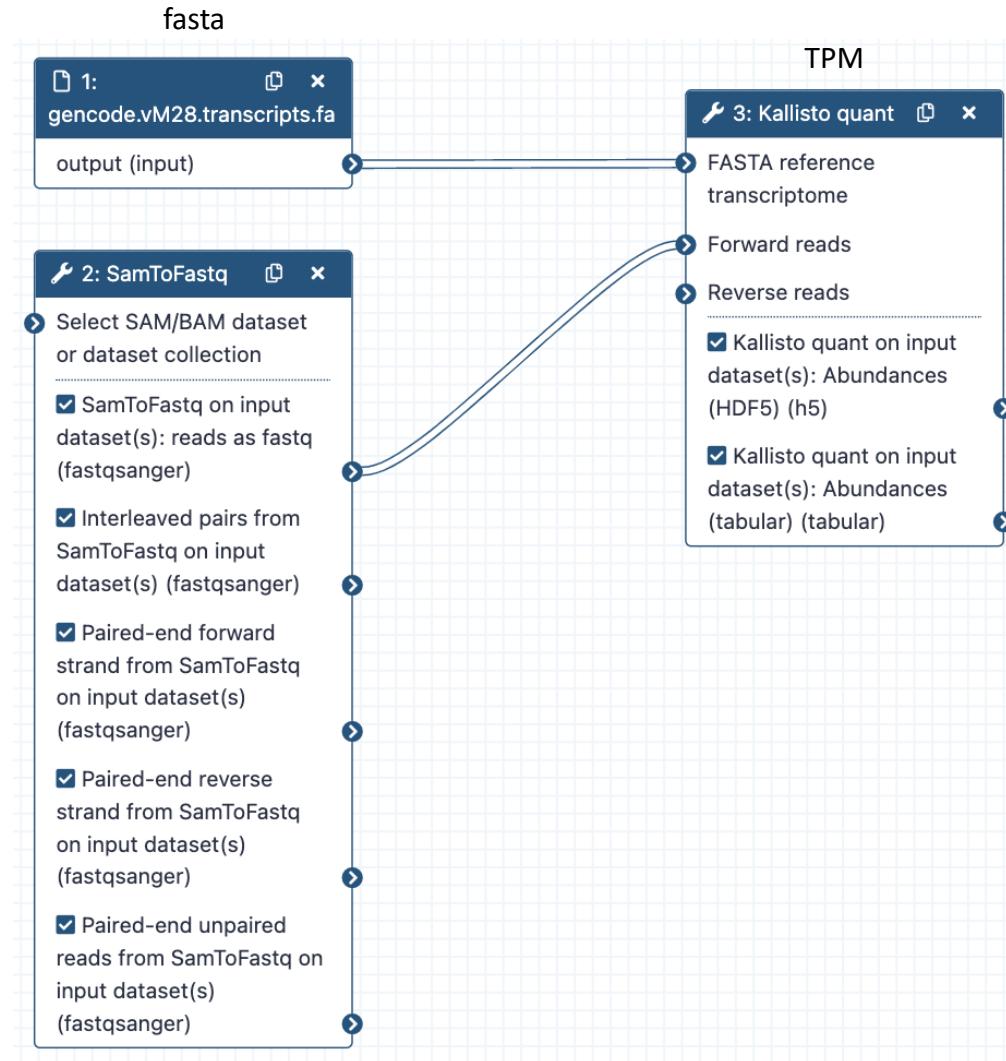
- Mouse genome.
- Mouse transcriptome.
- Hallmark gene sets.
- Download at
<https://bit.ly/3UUCKUr>



STAR alignment workflow



Kallisto quantification workflow

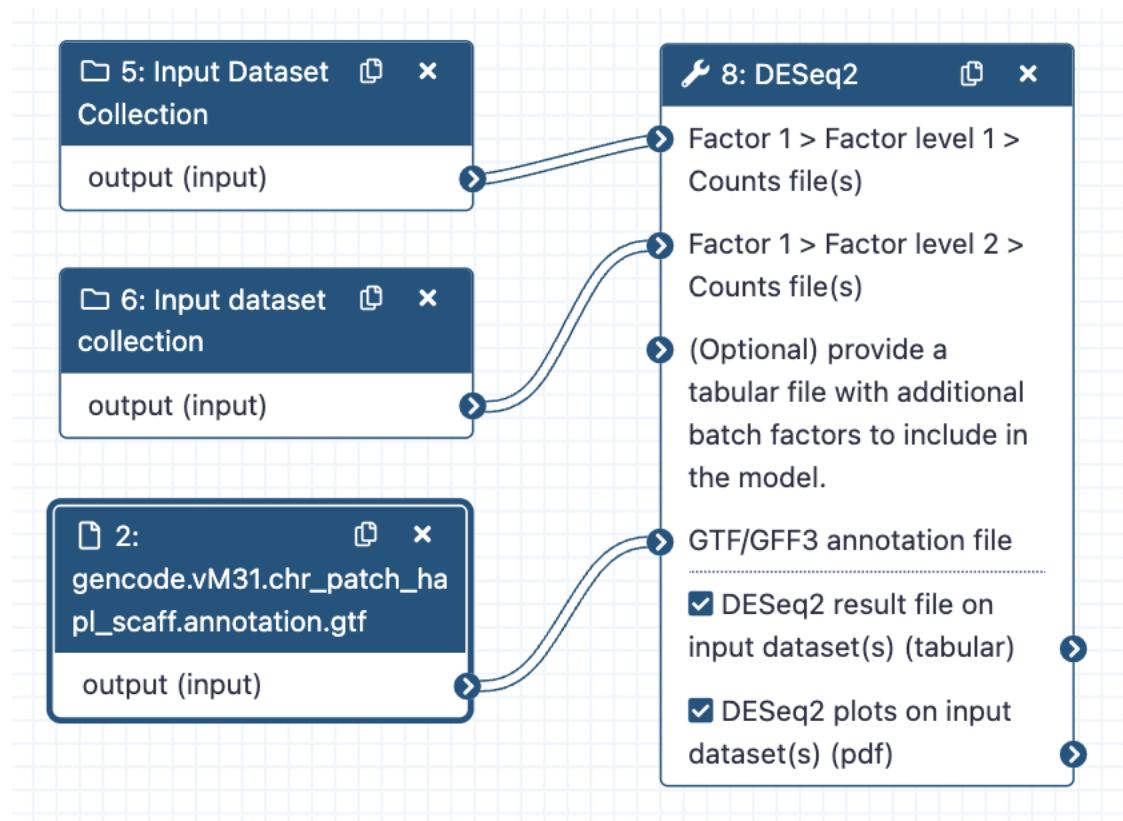


DEG using DESeq2 workflow

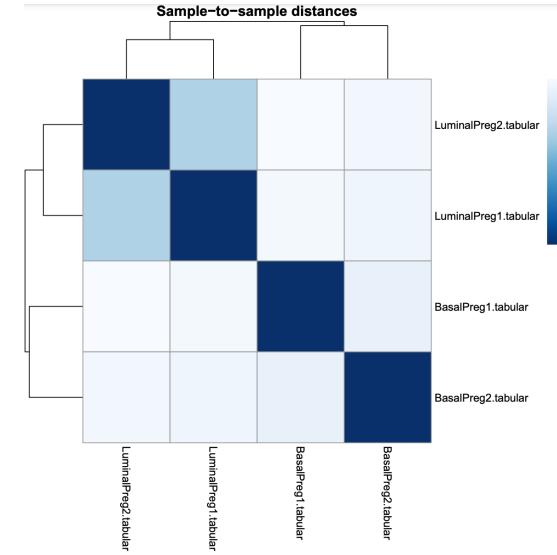
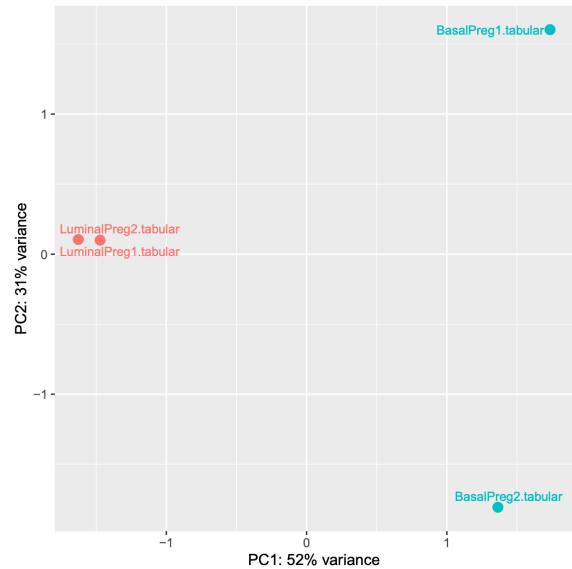
Data collection A
(output from Kallisto)

Data collection B
(output from Kallisto)

Reference genome
(GTF format)



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



GenelD	Base mean	log2(FC)	StdErr	Wald-Stats	P-value	
ENSMUSG00000063157.10	22.7053166147197	-4.82006060543848	0.870801517979073	-5.53520004951842	3.10873323980544e-08	1
ENSMUSG00000061937.8	19.0770163172454	-4.53054003503599	0.879112516114901	-5.15353831504755	2.55616816979082e-07	0
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