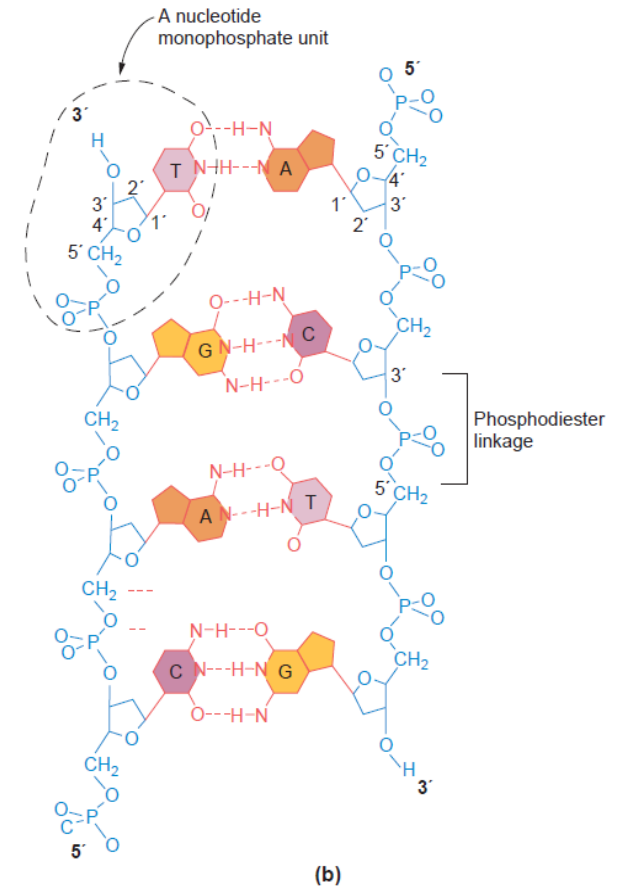
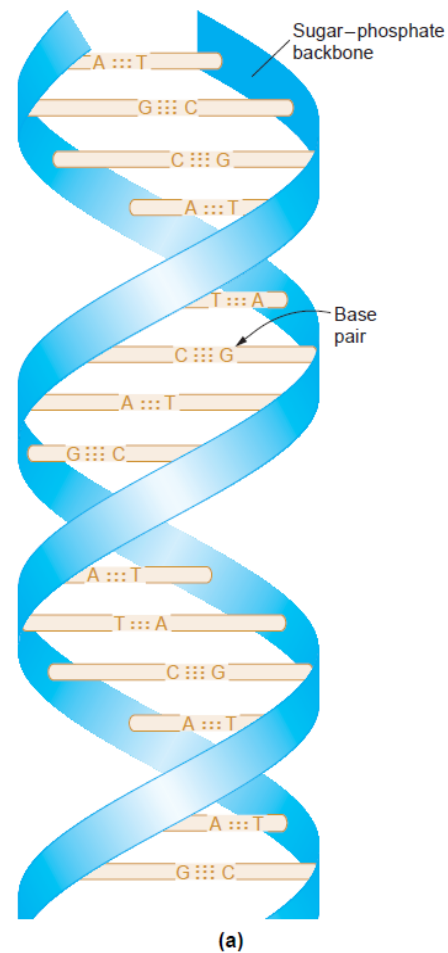
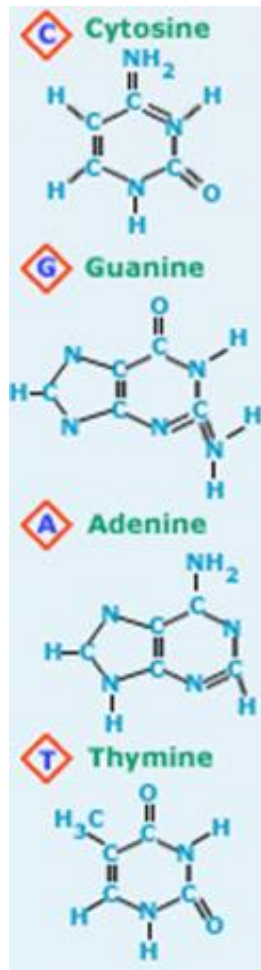


# **Project : TCGA data analysis**

# DNA

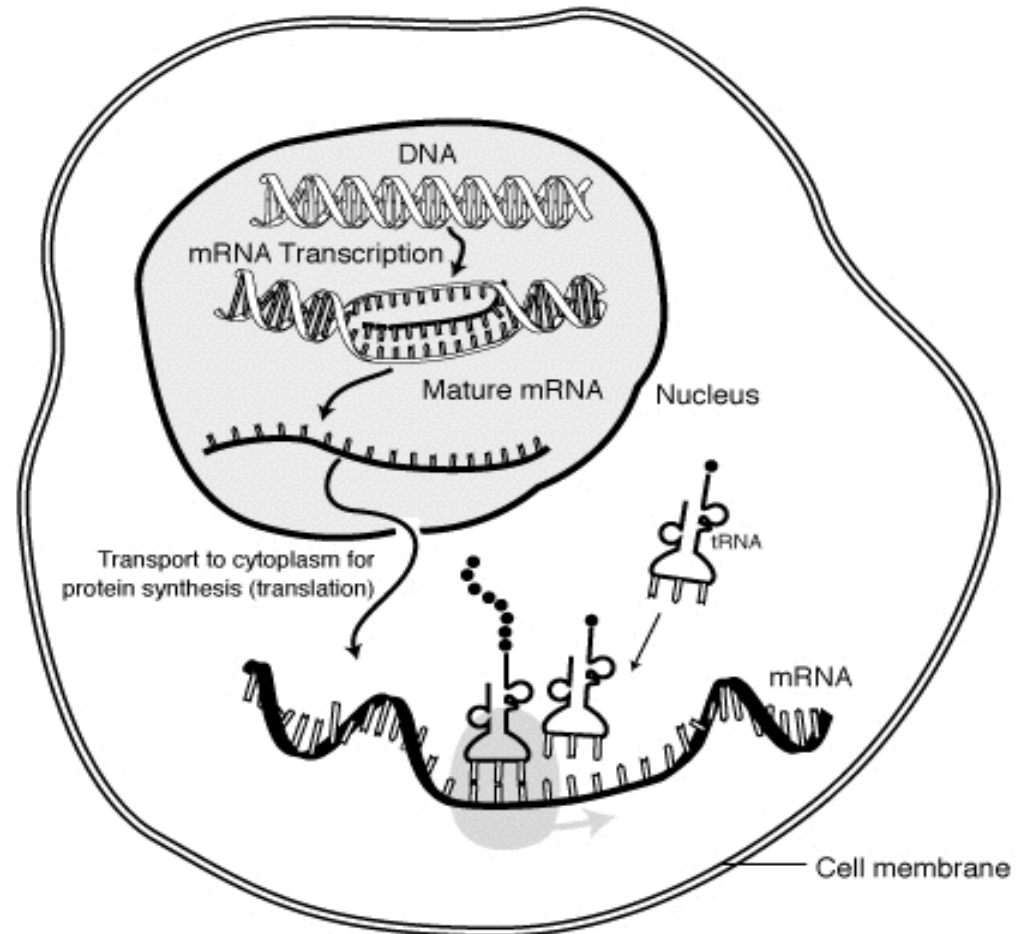


Source: Introduction to Genetic Analysis, Griths et al., 2008.

<http://esheets.biz/dexterity-dna-and-rna-worksheet/step-nucleic-acids-dna-and-rna/>

# DNA to Protein

- DNA  
(map of human body)
- mRNA  
(transcription)
- Amino acid  
(translation)
- Protein
- miRNA



Source: <http://en.wikipedia.org/wiki/File:MRNA-interaction.png>  
(genome.gov, 2004)

# DNA to Protein

<https://www.youtube.com/watch?v=gG7uCskUOrA>

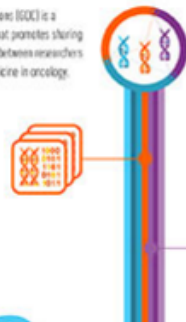
# TCGA data

[Home](#)[About Cancer Genomics](#)[Cancers Selected for Study](#)[Research Highlights](#)[Publications](#)[News and Events](#)[About TCGA](#)

## NATIONAL CANCER INSTITUTE GENOMIC DATA COMMONS

The NCI Genomic Data Commons (GDC) is a knowledge base for cancer that promotes sharing of genomic and clinical data between researchers and facilitates precision medicine in oncology.

The GDC brings together harmonized genomic datasets, starting with 3 petabytes of NCI genomic data from TCGA and other initiatives.



Researchers are encouraged to

## The NCI Genomic Data Commons Launches

The Genomic Data Commons (GDC) is a data sharing platform that promotes precision medicine in oncology, and it will host all of the TCGA data.

[Learn More](#)

## Launch Data Portal

The Genomic Data Commons (GDC) Data Portal is an interactive data system for researchers to search, download, upload, and analyze harmonized cancer genomic data sets, including TCGA.

## Questions About Cancer

Visit [www.cancer.gov](http://www.cancer.gov)

Call 1-800-4-CANCER

Use [LiveHelp Online Chat](#)

## Multimedia Library

Images

Videos and Animations

Podcasts

Interactive

Genomic Data Commons Launches

Analysis of Adrenocortical Carcinoma

Cancers Selected for Study

About TCGA

## Research Briefs

July 2016

**Longitudinal Study Charts Brain Tumor Evolution**  
Scientists analyzed the genomes of 114 glioblastoma tumors at the time of diagnosis and at the time of recurrence, including recurrent tumors from The Cancer Genome Atlas. The study found that the relapsed tumors likely arose years before the initial

## News and Announcements

June 06, 2016

**Newly launched Genomic Data Commons to facilitate data and clinical information sharing**  
The Genomic Data Commons (GDC) is a data sharing platform that promotes precision medicine in oncology, and it will host all of the TCGA data.

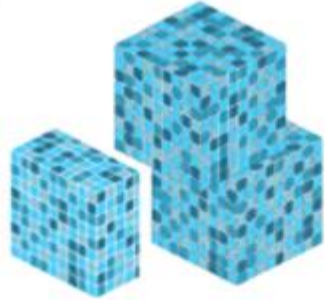
# TCGA data

## TCGA BY THE NUMBERS

TCGA produced over

**2.5**

PETABYTES  
of data



To put this into perspective, 1 petabyte of data  
is equal to

**212,000**

DVDs



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**

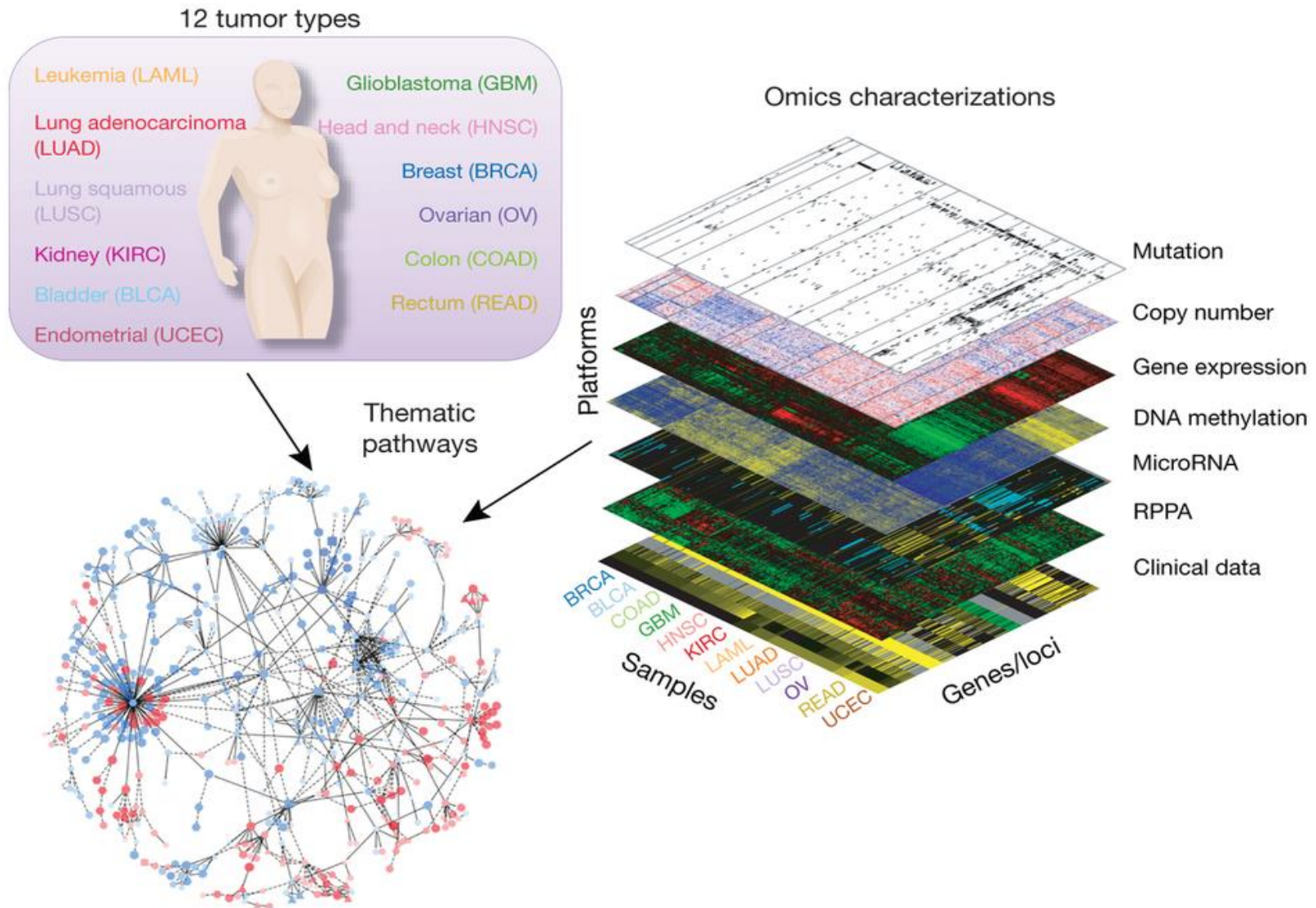
DIFFERENT  
DATA TYPES



<http://cancergenome.nih.gov/abouttcga/overview>



# TCGA data : the TCGA Pan-Cancer project

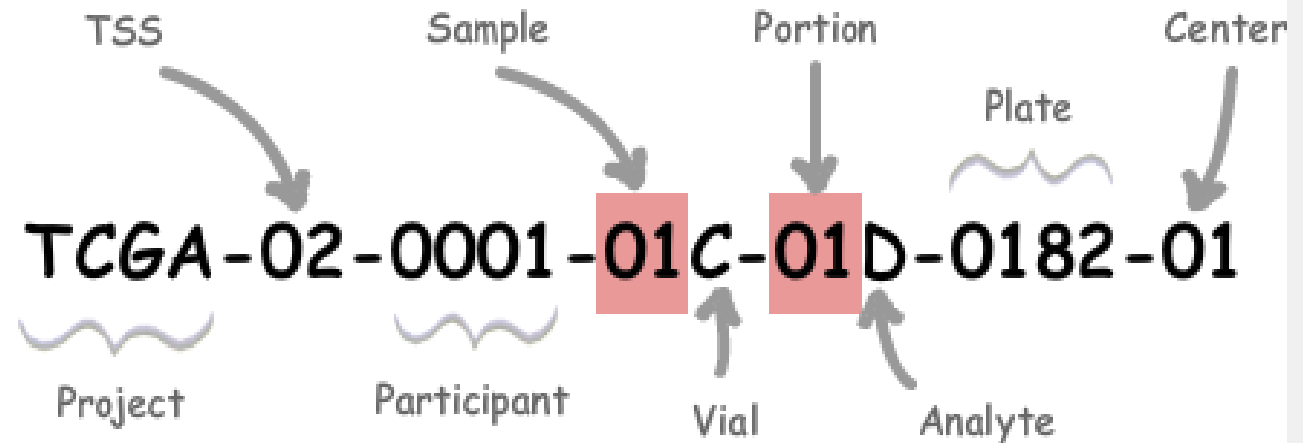


[http://www.nature.com/ng/journal/v45/n10/fig\\_tab/ng.2764\\_F1.html](http://www.nature.com/ng/journal/v45/n10/fig_tab/ng.2764_F1.html) (Weinstein et al., 2013)

# TCGA data : sample ID

sample types 0  
1 (solid tumor n  
ormal)

11 (solid tissue  
normal)



This figure of an aliquot barcode shows how it can be broken down into its components and translated into its metadata. The barcode metadata are further described in the following table.

<https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode>



# TCGA data : sample ID

Label	Identifier for	Value	Value description	Possible values
Project	Project name	TCGA	TCGA project	TCGA
<a href="#">TSS</a>	Tissue source site	02	GBM (brain tumor) sample from M D Anderson	See <a href="#">Code Tables Report</a>
<a href="#">Participant</a>	Study participant	0001	The first participant from MD Anderson for GBM study	Any alpha-numeric value
<a href="#">Sample</a>	Sample type	01	A solid tumor	Tumor types range from 01 - 09, normal types from 10 – 19, and control samples from 20 - 29. See <a href="#">Code Tables Report</a> for a complete list of sample codes
<a href="#">Vial</a>	Order of sample in a sequence of samples	C	The third vial	A to Z
<a href="#">Portion</a>	Order of portion in a sequence of 100 - 120 mg sample portions	01	The first portion of the sample	01-99
<a href="#">Analyte</a>	Molecular type of analyte for analysis	D	The analyte is a DNA sample	See <a href="#">Code Tables Report</a>
<a href="#">Plate</a>	Order of plate in a sequence of 96-well plates	0182	The 182nd plate	4-digit alphanumeric value
<a href="#">Center</a>	<a href="#">Sequencing</a> or <a href="#">characterization</a> center that will receive the aliquot for analysis	01	The Broad Institute <a href="#">GCC</a>	See <a href="#">Code Tables Report</a>

# https://gdac.broadinstitute.org/



*Liberating scientific knowledge  
from data, at scale*

[About](#)
[FireBrowse](#)
[Dashboards](#)
[Data](#)
[Analyses](#)
[Software](#)
[Documentation](#)
[FAQ](#)
[Download](#)
[Contact Us](#)
[What's New?](#)

Disease Name	Cohort	Cases	Analyses	Data
Adrenocortical carcinoma	ACC	<a href="#">92</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Bladder urothelial carcinoma	BLCA	<a href="#">412</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Breast invasive carcinoma	BRCA	<a href="#">1098</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Cervical and endocervical cancers	CESC	<a href="#">307</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Cholangiocarcinoma	CHOL	<a href="#">51</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Colon adenocarcinoma	COAD	<a href="#">460</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Colorectal adenocarcinoma	COADREAD	<a href="#">631</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	<a href="#">58</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Esophageal carcinoma	ESCA	<a href="#">185</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
FFPE Pilot Phase II	FPPP	<a href="#">38</a>	None	<a href="#">Browse</a>
Glioblastoma multiforme	GBM	<a href="#">613</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Glioma	GBMLGG	<a href="#">1129</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Head and Neck squamous cell carcinoma	HNSC	<a href="#">528</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Kidney Chromophobe	KICH	<a href="#">113</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Pan-kidney cohort (KICH+KIRC+KIRP)	KIPAN	<a href="#">973</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Kidney renal clear cell carcinoma	KIRC	<a href="#">537</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Kidney renal papillary cell carcinoma	KIRP	<a href="#">323</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Acute Myeloid Leukemia	LAML	<a href="#">200</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Brain Lower Grade Glioma	LGG	<a href="#">516</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Liver hepatocellular carcinoma	LIHC	<a href="#">377</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Lung adenocarcinoma	LUAD	<a href="#">585</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Lung squamous cell carcinoma	LUSC	<a href="#">504</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Mesothelioma	MESO	<a href="#">87</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Ovarian serous cystadenocarcinoma	OV	<a href="#">602</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Pancreatic adenocarcinoma	PAAD	<a href="#">185</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Pheochromocytoma and Paraganglioma	PCPG	<a href="#">179</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Prostate adenocarcinoma	PRAD	<a href="#">499</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Rectum adenocarcinoma	READ	<a href="#">171</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Sarcoma	SARC	<a href="#">261</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Skin Cutaneous Melanoma	SKCM	<a href="#">470</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Stomach adenocarcinoma	STAD	<a href="#">443</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Stomach and Esophageal carcinoma	STES	<a href="#">628</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Testicular Germ Cell Tumors	TGCT	<a href="#">150</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Thyroid carcinoma	THCA	<a href="#">503</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Thymoma	THYM	<a href="#">124</a>	<a href="#">Browse</a>	<a href="#">Browse</a>
Uterine Corpus Endometrial Carcinoma	UCEC	<a href="#">560</a>	<a href="#">Browse</a>	<a href="#">Browse</a>

- Colon adenocarcinoma / COAD / [460](#)
- Brain Lower Grade Glioma / LGG / [516](#)
- Liver hepatocellular carcinoma / LIHC / [377](#)
- Stomach adenocarcinoma / STAD / [443](#)
- Lung adenocarcinoma / LUAD / [585](#)
- Thyroid carcinoma / THCA / [503](#)
- Skin Cutaneous Melanoma / SKCM / [470](#)

# Download data (Example: BRCA)

## BRCA Archives

PrimaryAuxiliarySDRF/MageSend To

Files may also be downloaded [here](#), or with [firehose\\_get](#), or exported to [GenomeSpace](#) with the SendTo tab.  
Downloading data constitutes agreement to [TCGA data usage policy](#)

### Clinical

Clinical\_Pick\_Tier1 (MD5)  
Merge\_Clinical (MD5)

### SNP6 CopyNum

genome\_wide\_snp\_6-FFPE-segmented\_scna\_hg19 (MD5)  
genome\_wide\_snp\_6-segmented\_scna\_minus\_germline\_cnv\_hg18 (MD5)  
genome\_wide\_snp\_6-segmented\_scna\_hg18 (MD5)  
genome\_wide\_snp\_6-FFPE-segmented\_scna\_minus\_germline\_cnv\_hg18 (MD5)  
genome\_wide\_snp\_6-FFPE-segmented\_scna\_hg18 (MD5)  
genome\_wide\_snp\_6-FFPE-segmented\_scna\_minus\_germline\_cnv\_hg19 (MD5)  
genome\_wide\_snp\_6-segmented\_scna\_hg19 (MD5)  
genome\_wide\_snp\_6-segmented\_scna\_minus\_germline\_cnv\_hg19 (MD5)

### LowPass DNaseq CopyNum

illuminahtseq\_dnaseqc-segmentation (MD5)

### Mutation Annotation File

Mutation\_Packager\_Oncotated\_Calls (MD5)  
Mutation\_Packager\_Calls (MD5)

### methylation

Methylation\_Preprocess (MD5)  
humanmethylation27-within\_bioassay\_data\_set\_function (MD5)  
humanmethylation450-FFPE-within\_bioassay\_data\_set\_function (MD5)  
Methylation\_Preprocess-FFPE (MD5)  
humanmethylation450-within\_bioassay\_data\_set\_function (MD5)

### miRSeq

illuminaga\_mimaseq-miR\_isoform\_expression (MD5)  
miRseq\_Preprocess-FFPE (MD5)  
illuminahtseq\_mimaseq-FFPE-miR\_isoform\_expression (MD5)  
miRseq\_Mature\_Preprocess-FFPE (MD5)  
miRseq\_Mature\_Preprocess (MD5)  
miRseq\_Preprocess (MD5)  
illuminaga\_mimaseq-miR\_gene\_expression (MD5)  
illuminahtseq\_mimaseq-miR\_gene\_expression (MD5)  
illuminahtseq\_mimaseq-miR\_isoform\_expression (MD5)  
illuminahtseq\_mimaseq-FFPE-miR\_gene\_expression (MD5)

# Download data : Next Generation Sequencing Technology



## MiniSeq System

Power and simplicity for targeted sequencing.



## MiSeq Series

Small genome and targeted sequencing.



## NextSeq Series

Everyday genome, exome transcriptome sequencing, and more.



## HiSeq Series

Production-scale genome, exome, transcriptome sequencing and more.



## HiSeq X Series

Population- and production-scale human whole-genome sequencing.

[www.illumina.com/documents/products/techspotlights/techspotlight\\_sequencing.pdf](http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf)

# Download data

## Clinical

Clinical\_Pick\_Tier1 (MD5)

## mRNASeq

illuminahiseq\_rnaseqv2-RSEM\_genes\_normalized (MD5)

illuminaga\_rnaseqv2-RSEM\_genes\_normalized (MD5)

## miRSeq

illuminahiseq\_mirnaseq-miR\_gene\_expression (MD5)

illuminaga\_mirnaseq-miR\_gene\_expression (MD5)



# Download data (Example: BRCA)

## Clinical

Clinical\_Pick\_Tier1 (MD5)

Merge\_Clinical (MD5)

## mRNASeq

illuminahiseq\_rnaseqv2-RSEM\_isoforms (MD5)

illuminahiseq\_rnaseqv2-exon\_quantification (MD5)

illuminahiseq\_rnaseqv2-RSEM\_genes\_normalized (MD5)

illuminahiseq\_rnaseq-exon\_expression (MD5)

illuminahiseq\_rnaseqv2-junction\_quantification (MD5)

illuminahiseq\_rnaseqv2-RSEM\_isoforms\_normalized (MD5)

mRNAseq\_Preprocess (MD5)

illuminahiseq\_rnaseqv2-RSEM\_genes (MD5)

illuminahiseq\_rnaseq-splice\_junction\_expression (MD5)

illuminahiseq\_rnaseq-gene\_expression (MD5)

## miRSeq

illuminaga\_mimaseq-miR\_isoform\_expression (MD5)

miRseq\_Preprocess-FFPE (MD5)

illuminahiseq\_mimaseq-FFPE-miR\_isoform\_expression (MD5)

miRseq\_Mature\_Preprocess-FFPE (MD5)

miRseq\_Mature\_Preprocess (MD5)

miRseq\_Preprocess (MD5)

illuminaga\_mimaseq-miR\_gene\_expression (MD5)

illuminahiseq\_mimaseq-miR\_gene\_expression (MD5)

illuminahiseq\_mimaseq-miR\_isoform\_expression (MD5)

illuminahiseq\_mimaseq-FFPE-miR\_gene\_expression (MD5)

# Download data : after extracting the compressed files

## Clinical

Name	Date modified	Type	Size
All_CDEs.csv	9/21/2016 12:15 AM	Microsoft Excel C...	952 KB
All_CDEs.txt	9/21/2016 12:12 AM	TXT File	946 KB
BRCA.clin.merged.picked.txt	2/13/2016 8:19 AM	TXT File	148 KB
MANIFEST.txt	2/13/2016 8:26 AM	TXT File	1 KB
stage3_params_clin_selection_BRCA.tsv	2/13/2016 8:19 AM	TSV File	1 KB

## mRNASeq

Name	Date modified	Type	Size
BRCA.rnaseqv2_illuminaiseq_rnaseqv2_unc_edu_Level_3_RSEM_genes_normalized_data.csv	9/21/2016 12:27 AM	Microsoft Excel C...	187,277 KB
BRCA.rnaseqv2_illuminaiseq_rnaseqv2_unc_edu_Level_3_RSEM_genes_normalized_data.data.txt	9/21/2016 12:25 AM	TXT File	187,277 KB
MANIFEST.txt	2/14/2016 2:20 AM	TXT File	1 KB

## miRSeq

Name	Date modified	Type	Size
BRCA.mirnaiseq_illuminaiseq_mirnaiseq_bcgsc_ca_Level_3_miR_gene_expression_data.csv	9/21/2016 1:09 AM	Microsoft Excel...	8,758 KB
BRCA.mirnaiseq_illuminaiseq_mirnaiseq_bcgsc_ca_Level_3_miR_gene_expression_data.data.txt	9/21/2016 1:09 AM	TXT File	8,758 KB
MANIFEST.txt	2/14/2016 5:34 AM	TXT File	1 KB

# Q & A