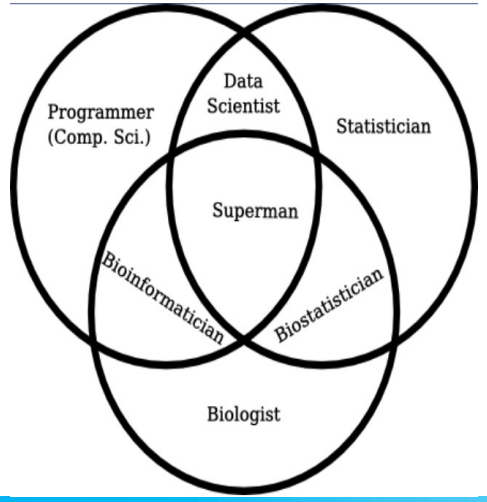
암 유전체 빅데이터



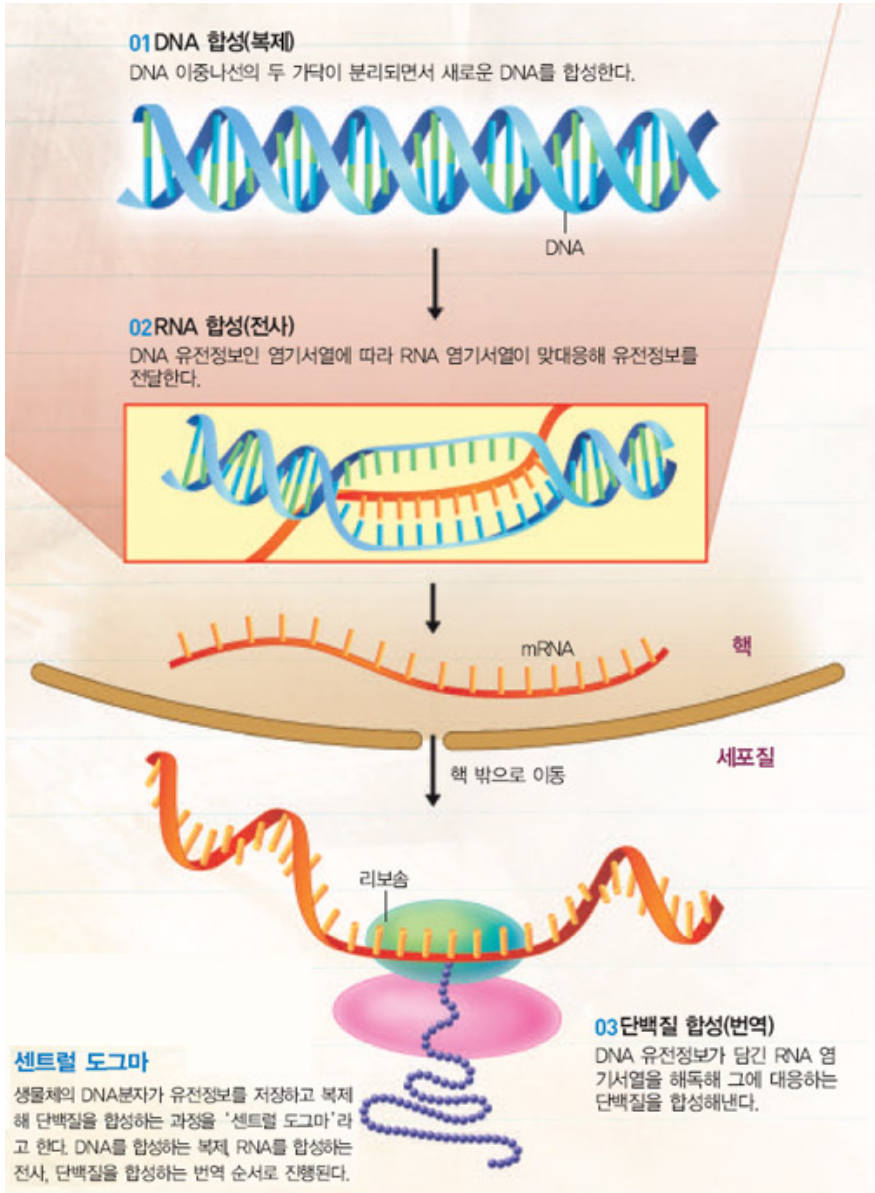
Dongwan Hong, Ph.D.
(dwhong@catholic.ac.kr)
Dept. of Medical Informatics, College of Medicine
Catholic University of Korea

Contents

- **cBioPortal**
  - http://www.cbioportal.org
  - Korean prostate cancers vs TCGA prostate cancers
- **UCSC Xena Browser**
  - http://xena.ucsc.edu
  - Differentially Expressed Gene (DEG)s selection
- **Genomic Data Commons (GDC) Data Portal**
  - http://portal.gdc.cancer.gov
  - Intron retention is a widespread mechanism of tumor-suppressor inactivation, Nature Genetics, 2015

cBioPortal
(http://cbioportal.org)

CENTRAL DOGMA



Multi-OMICS data

Big Data

Fourth Industrial Revolution

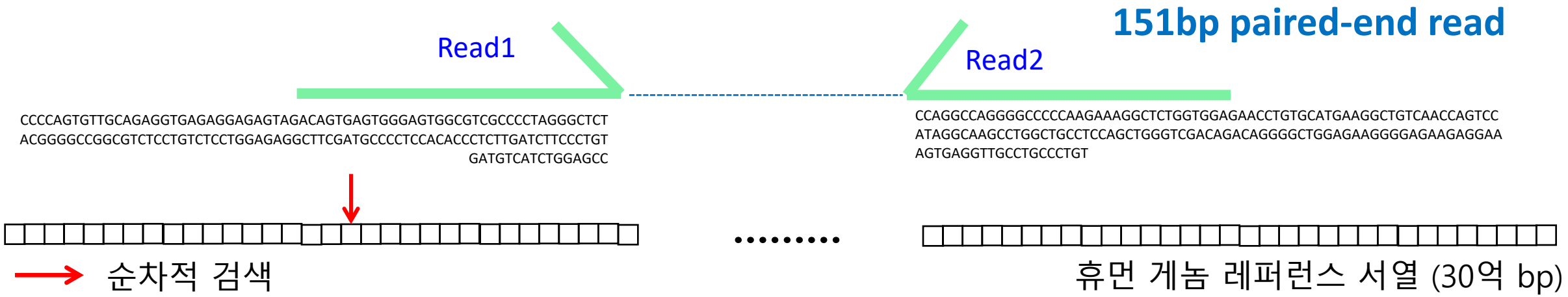
출처: 사이언스 올

Next generation sequencing instruments

	 NovaSeq6000	 HiSeq X	 HiSeq 2500 V4	 PacBio RS II
 Running time	~2 days	~3 days	~6 days	~1 days
 Throughput	6TB (~25명)	1,8TB (~8명)	1TB (~4명)	~2TB
 Cost	GB 당 $7 이하 일 것으로 기대	GB 당 $7	GB 당 $29	GB 당 $400-800
 연간 분석 가능한 유전체수 [30x 기준]	>4,500	>900	>250	>6
 리드길이	2*150bp	2*150bp	2*125bp	10-15kp

* Hiseq X 1기 (연간 생산량): 1주 생산량 (≒35샘플: 30x)×4주 ×12월/2(normal, tumor pair) = 약 1,000 명 한

Alignment: short read mapping (Sequential Search)



소요 시간 = ~ 283초 (4 분 43 초) / 1 리드

리드 개수: 366,489,310

Paired-end read: 732,978,620

151 base pairs: 110,679,771,620 bp

3G 게놈: ~36 x 커버리지, Read depth

총 매핑 시간: 103,716,474,730 초

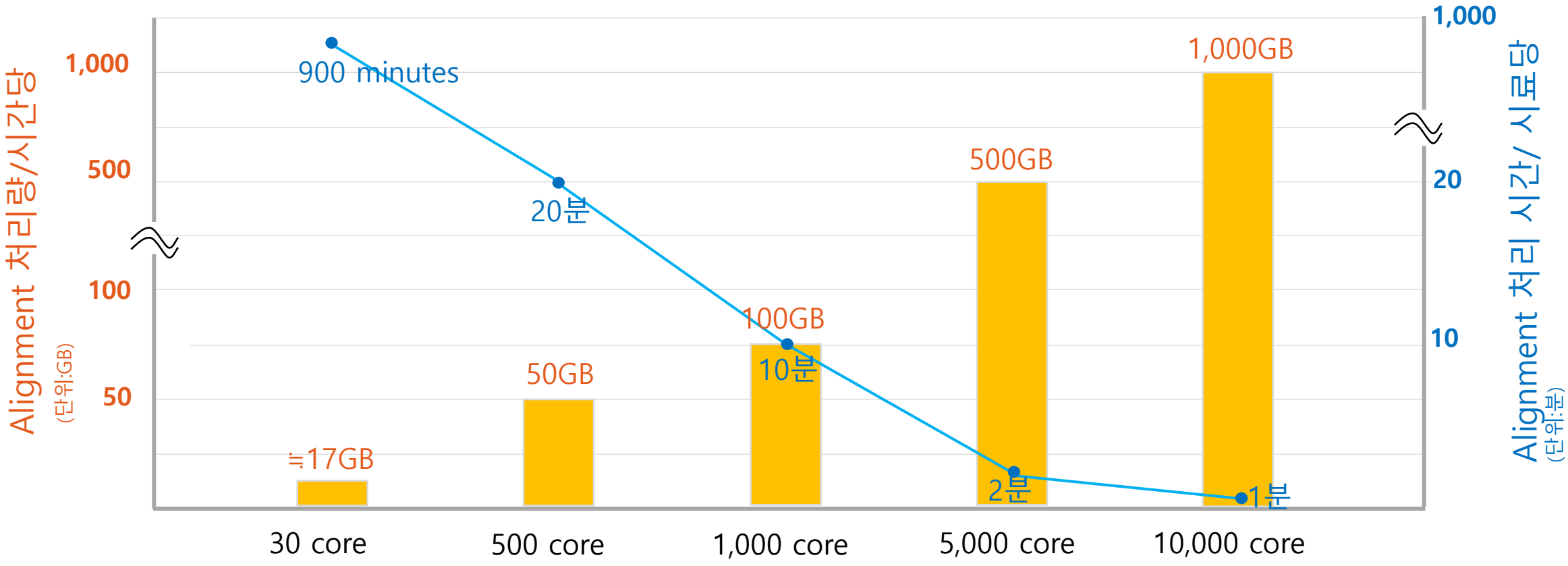
총 매핑 시간: 1,728,607,912 분

총 매핑 시간: 28,810,131 시간

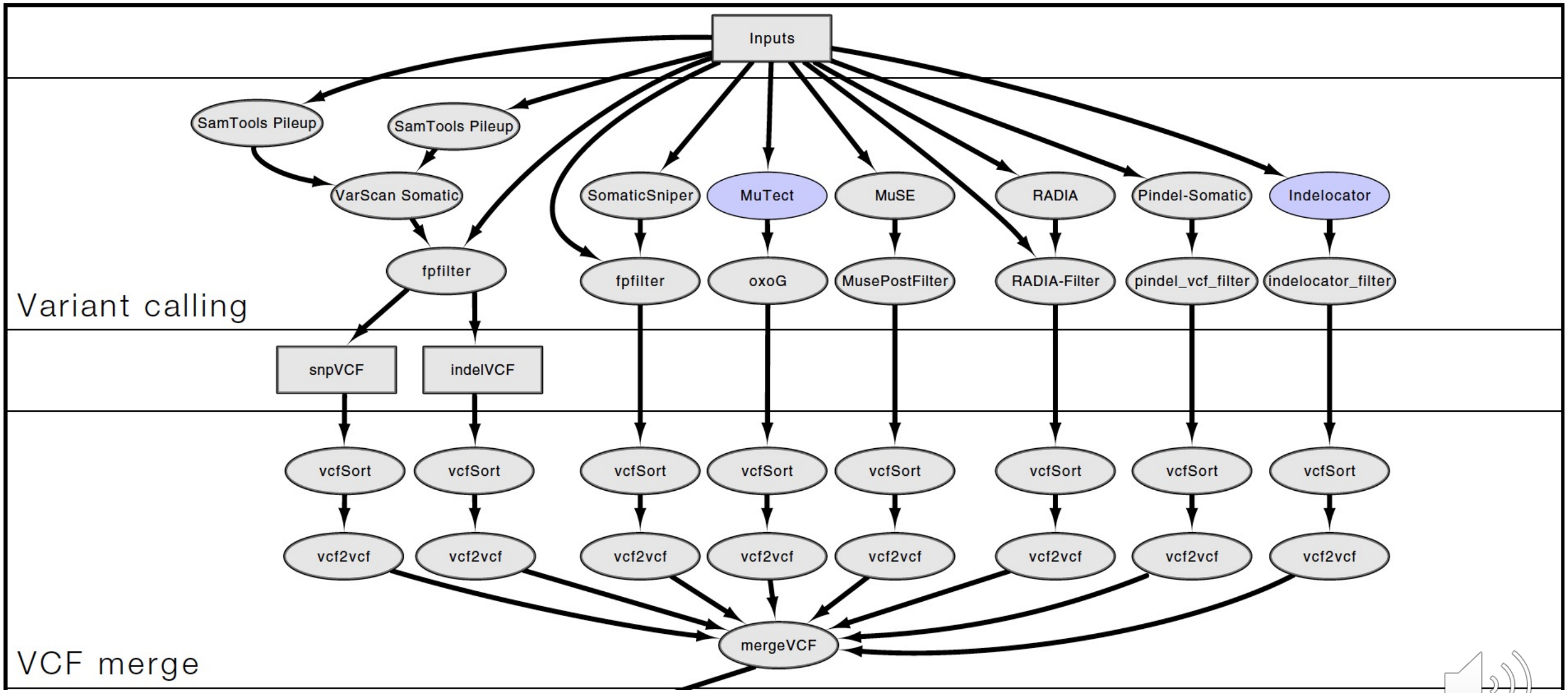
총 매핑 시간: 1,200,422 일

총 매핑 시간: 3,288 년

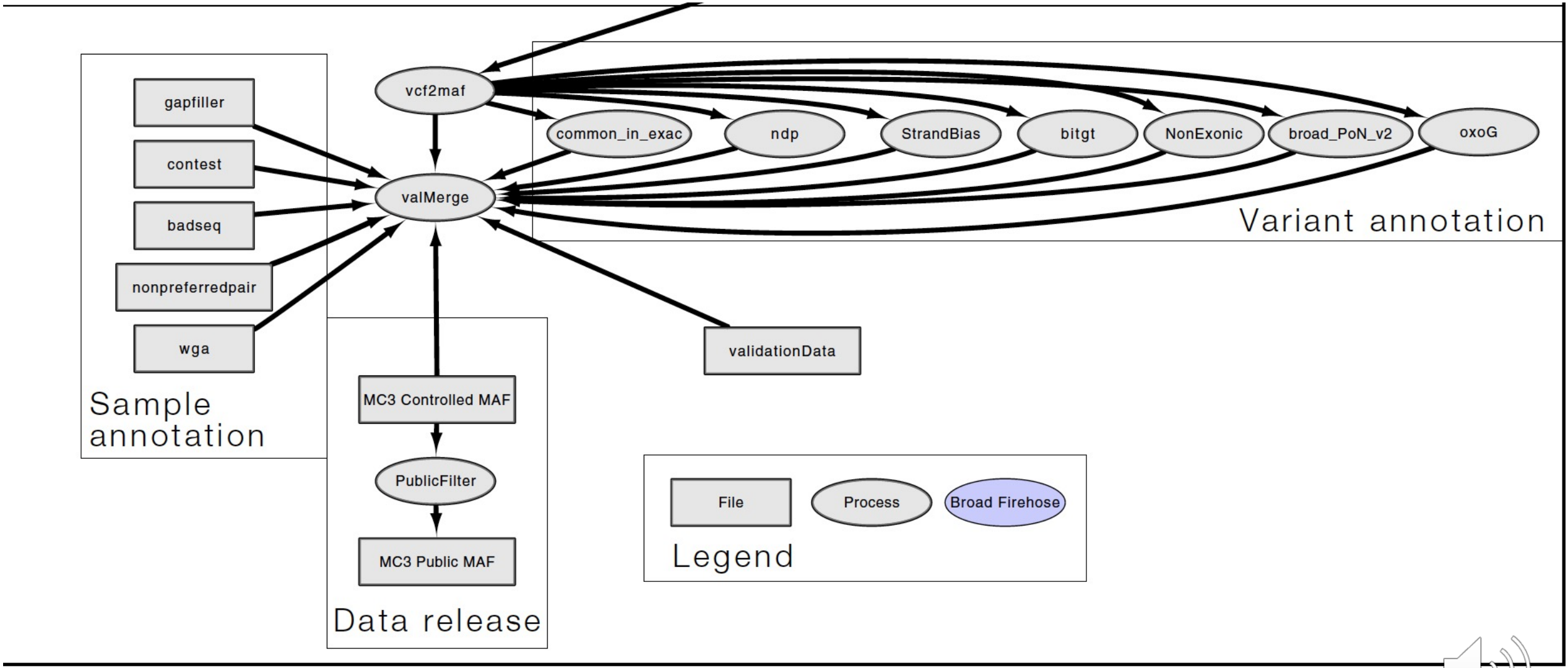
Running time by computing spec.



Workflow for Mutation Detection and Filtering



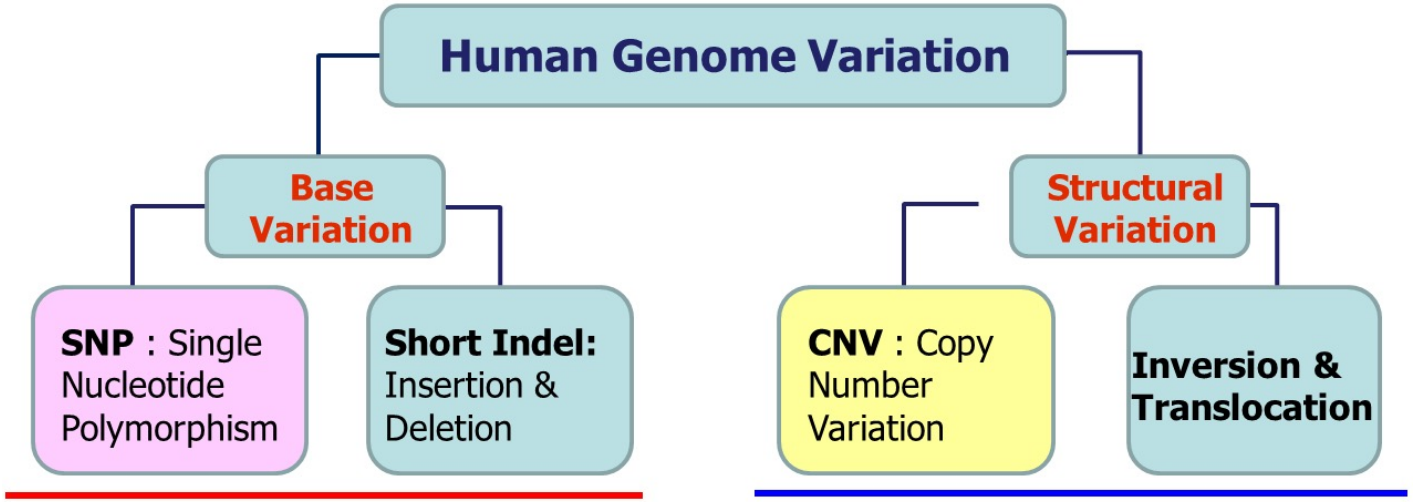
Workflow for Mutation Detection and Filtering



Sequencing Data Size

Nxx-xxxxx (Sample)		< 암 패널 데이터 >	< 전장유전체시퀀싱 > ~80x	< 엑솜 시퀀싱 > ~100x	< 전사체 시퀀싱 > ~200x
1.FASTQ	1.RAW	300 MB	70 GB	4 GB	5 GB
2.BAM	2.BAM	300 MB	310 GB	15 GB	7 GB
3.VCF	3.VCF	55 KB	10 MB	100 KB	10 MB
4.EXCEL	4.EXCEL	51 KB			
5.PDF	5.PDF	150 KB			

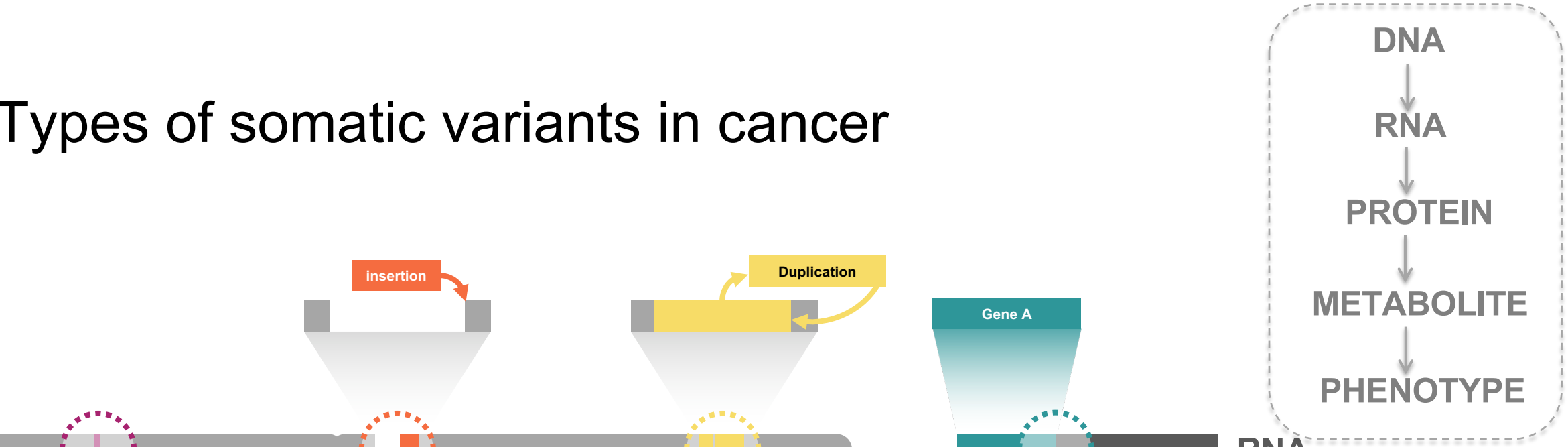
- Cancer genes
  - Oncogene
    - An oncogene is a gene that has the potential to cause cancer. In tumor cells, these genes are often mutated, or expressed at high levels. Most normal cells will undergo a programmed form of rapid cell death (apoptosis) when critical functions are altered and malfunctioning. Activated oncogenes can cause those cells designated for apoptosis to survive and proliferate instead.
  - Tumor suppressor gene
    - A tumor suppressor gene, or antioncogene, is a gene that regulates a cell from advancing to cancer. When this gene is mutated it results in a loss or reduction in its function; in combination with other genetic mutations this could allow the cell to grow abnormally



Cancer Genome Data

- ~1,000-100,000 somatic substitution
- ~300 chromosomal rearrangements
- ~1,000 copy number alterations

Types of somatic variants in cancer



Single
Nucleotide
Variations

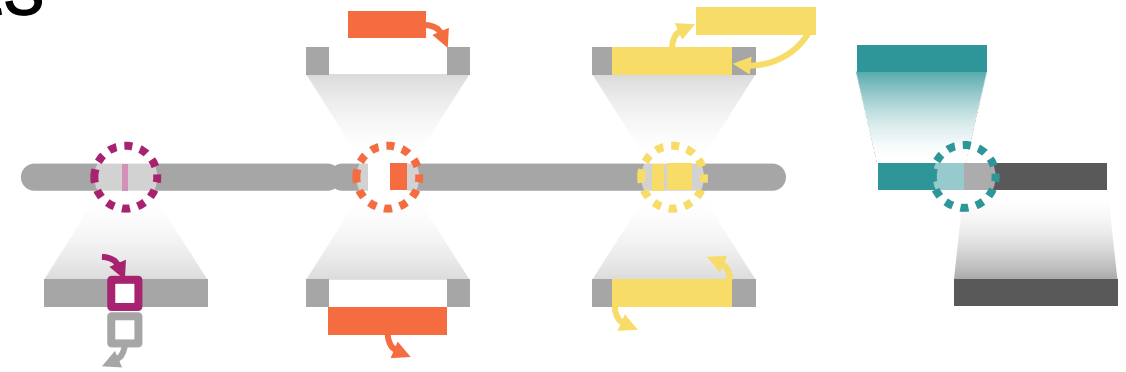
Copy
Number
Variations

Structural
Variations

Gene
Fusion

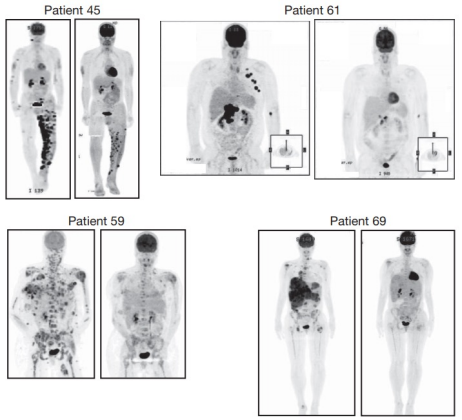
C>A

These variants are used as drug targets



● SNV ● SV
● CNV ● Fusion

Vemurafenib → BRAF V600E



Bollag et al., Nature 2010

Table 1. Genes Used to Guide FDA-Approved Therapies

Mutations Used to Select Targeted Therapy	
ABL1	CML, ALL
EGFR	lung cancer
ALK	lung cancer
ROS1	lung cancer
BRAF	melanoma
ERBB2	breast and gastric cancer
KIT	gastrointestinal stromal tumor
PDGFRA	leukemia, MDS
PDGFRB	dermatofibrosarcoma protuberans
BRCA1 and BRCA2 (germline)	ovarian cancer
Mutations Used to Select against Targeted Therapy	
KRAS	colorectal cancer
NRAS	colorectal cancer
BRAF	colorectal cancer

Hyman et al., Cell 2017

Table 2. Genes with Clinical Evidence Supporting Them as Targets for Drug Development

Gene	Alteration(s)
RET	M918, fusions
MET	exon 14 splice, amplifications
AKT1	E17K
ERBB2	activating missense mutations
FGFR1/2/3	fusions, amplifications, activating missense mutations
FLT3	ITD, D835
IDH1	R132
IDH2	R140
MAP2K1 (MEK1)	activating missense mutations
MTOR	activating missense mutations
BRAF	non-V600-activating mutations, fusions
NTRK1/2/3	fusions
NRG1	fusions
PIK3CA	activating missense mutations
Homologous recombination deficiencies (BRCA1, BRCA2, PALB2, RAD50, ATM, RAD51, RAD51B/C/D, FANCA, CHEK1/2)	inactivating alterations
ARAF	S214
EGFR	rare activating missense mutations, insertions
TSC1/2	inactivating alterations
SMARCA4	inactivating alterations
SMARRB1	inactivating alterations

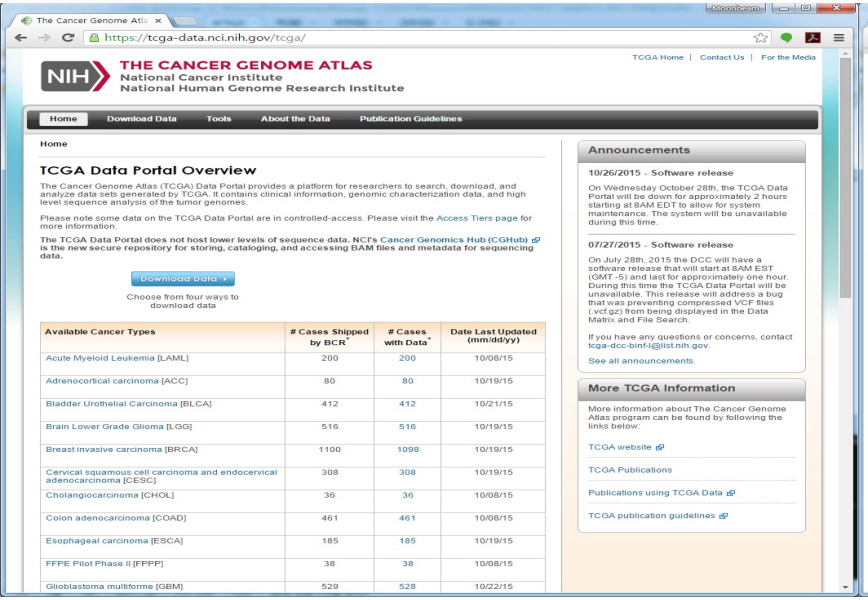
Cancer drugs

- Vemurafenib and trametinib
  BRAF V600E mutations in melanoma
- Erlotibin and osimertinib
  EGFR mutations in NSCLC; L858R, Del (19), T790M
- Pembrolizumab (immune checkpoint inhibitor): FDA approved
  SOLID tumors from any tissue type
  Mismatch repair deficiency (dMMR)
- Nivolumab, ipilimumab and atezolizumab
  High tumor mutation burden

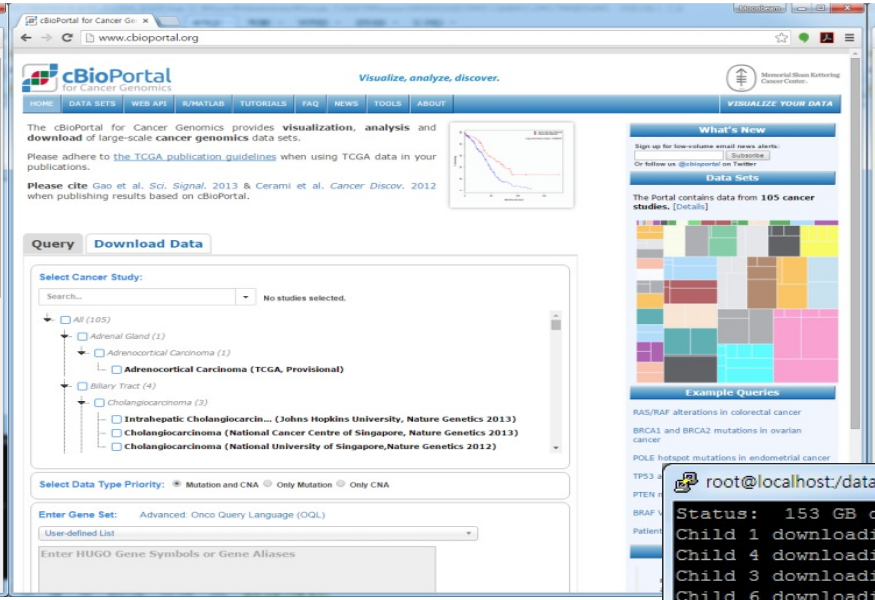
The first version of TCGA(The Cancer Genome Atlas)

Next Generation Sequencing data

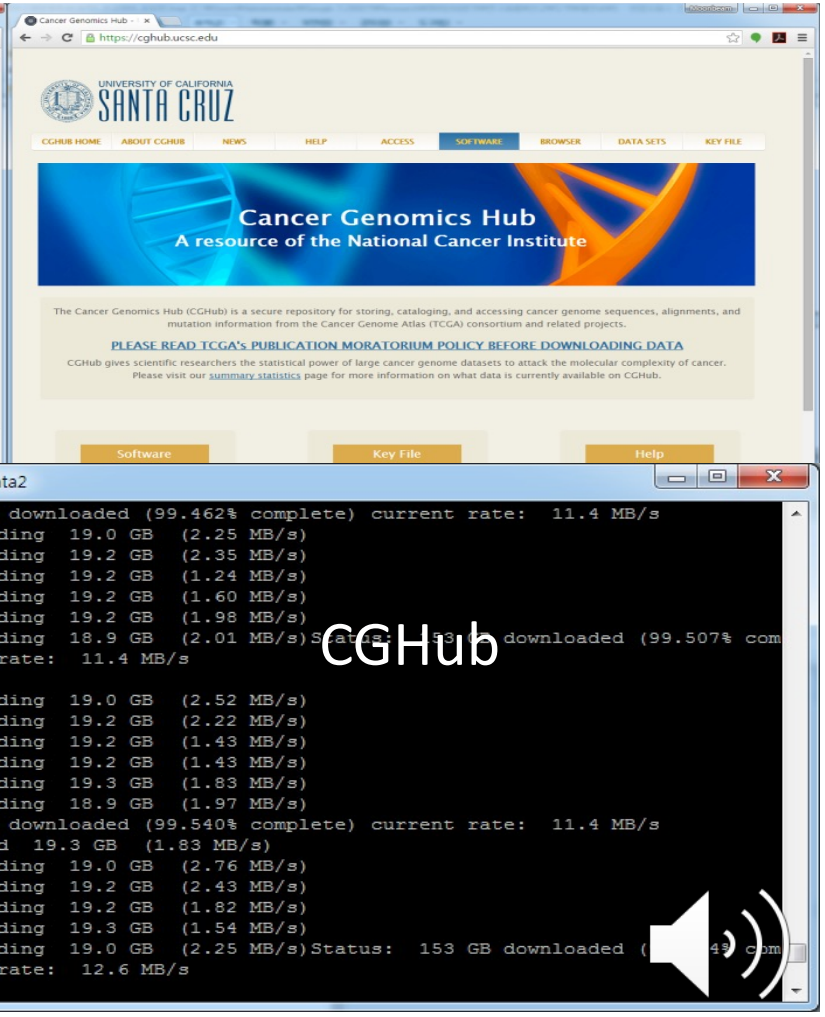
> 20 peta bytes



TCGA data portal



cbioportal

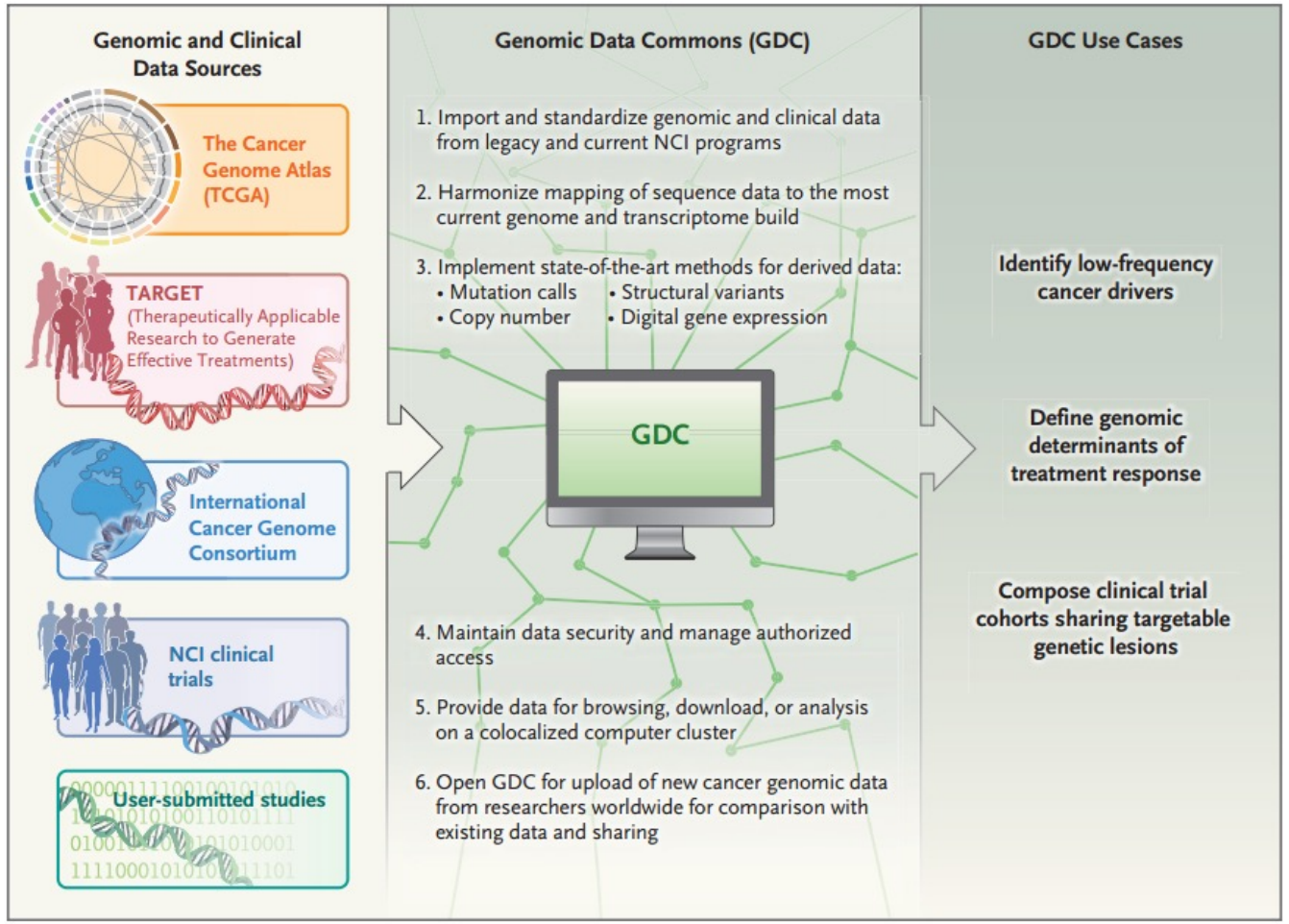


CGHub

Clinical Genomic Data (2020)

~ 25,000 peta

Toward a Shared Vision for Cancer Genomic Data



NEJM, Sep, 2016

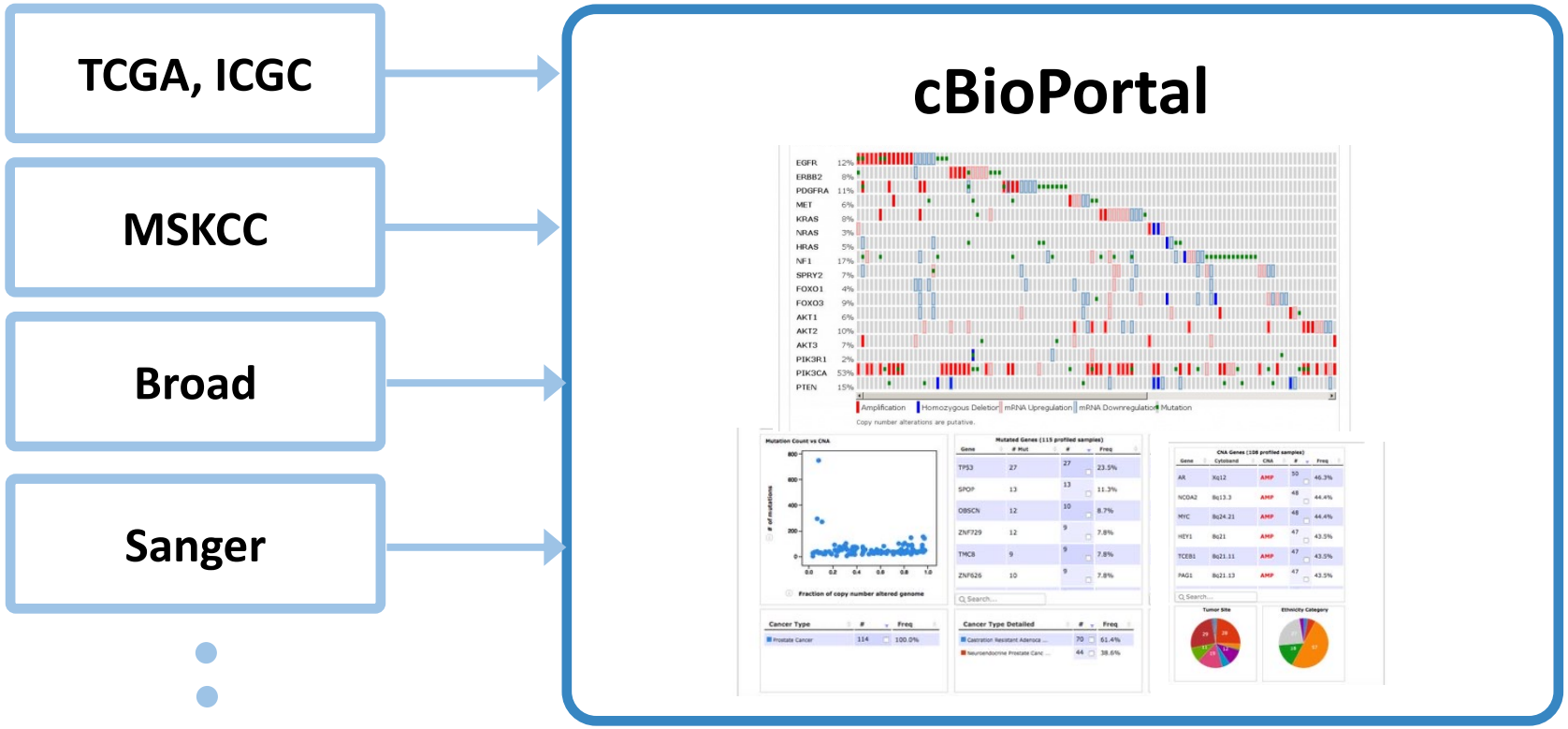
Table 1. Acronyms for Data Sharing Projects and Groups

Acronym	Full Name	What is it?
CEDAR	Center for Expanded Data Annotation and Retrieval	A Stanford University center sponsored by the National Institutes of Health's Big Data to Knowledge Initiative. It aims to improve the metadata used to label cancer data for easier retrieval.
ClinGen	Clinical Genome Resource	A National Institutes of Health project to define the clinical relevance of genes and variants in various diseases, including cancer.
ClinVar	Clinical Variant Resource	A partner project of ClinGen that aggregates information about the relationships between genetic variation and human health.
dbGaP	The database of Genotypes and Phenotypes	An archive run by the National Center for Biotechnology Information that collects data on the interaction between human genotypes and phenotypes.
GA4GH	Global Alliance for Genomics and Health	A coalition of more than 300 health-related institutions promoting the sharing of clinical and genomic data.
GENIE	Genomics, Evidence, Neoplasia, Information, Exchange	An American Association for Cancer Research project that collects clinical and genomic information.
ORIEN	Oncology Research Information Exchange Network	A research center collaborative that shares a common protocol and provides clinical trial matching for patients.
TARGET	Therapeutically Applicable Research to Generate Effective Treatments	A National Cancer Institute project tracking the molecular changes driving childhood cancers.
TCGA	The Cancer Genome Atlas	A joint project of the National Cancer Institute and National Human Genome Research Institute that makes maps o genetic changes in 33 types of cancer.

Cell, 2017

cBioPortal FOR CANCER GENOMICS

Genomic/
Clinical Data



(1) **Visualization** (2) Analysis (3) Download

Access cBioPortal http://www.cbioportal.org/

cBioPortal FOR CANCER GENOMICS

Data Sets | Web API | R/MATLAB | Tutorials | FAQ | News | Visualize Your Data | About

The cBioPortal for Cancer Genomics provides **visualization**, **analysis** and **download** of large-scale **cancer genomics** data sets.
Please cite Gao et al. *Sci. Signal.* 2013 & Cerami et al. *Cancer Discov.* 2012 when publishing results based on cBioPortal.

QUERY

Prostate Adenocarcinoma (TCGA, Provisional) 선택

Select Studies: 0 studies selected (0 samples) Select all

Search...

PanCancer Studies 2
Cell lines 2
Adrenal Gland 1
Ampulla of Vater 1
Biliary Tract 5
Bladder/Urinary Tract 7
Blood 8
Bone 2
Bowel 5
Breast 10

Prostate

Prostate Adenocarcinoma

- Genomic Hallmarks of Prostate Adenocarcinoma (CPC-GENE, Nature … 477 samples
- MSK-IMPACT Clinical Sequencing Cohort (MSKCC): Prostate Cancer 504 samples
- Metastatic Prostate Cancer, SU2C/PCF Dream Team (Robinson et al., … 150 samples
- Neuroendocrine Prostate Cancer (Trento/Cornell/Broad 2016) 114 samples
- Prostate Adenocarcinoma (Broad/Cornell, Cell 2013) 57 samples
- Prostate Adenocarcinoma (Broad/Cornell, Nat Genet 2012) 112 samples
- Prostate Adenocarcinoma (Fred Hutchinson CRC, Nat Med 2016) 176 samples
- Prostate Adenocarcinoma (MSKCC, Cancer Cell 2010) 216 samples
- 333 samples
- Prostate Adenocarcinoma (TCGA, Provisional) 499 samples
- Prostate Adenocarcinoma CNA study (MSKCC, PNAS 2014) 104 samples
- Prostate Adenocarcinoma Organoids (MSKCC, Cell 2014) 12 samples
- Prostate Adenocarcinoma, Metastatic (Michigan, Nature 2012) 61 samples

Summary

Skin

Cutaneous Squamous Cell Carcinoma

Select Data Type Priority: Mutation and CNA | Only Mutation | Only CNA

Enter Gene Set:
Advanced: Onco Query Language (OQL)

User-defined List

Enter HUGO Gene Symbols or Gene Aliases

*** 20가지의 암종에 대하여 총 292 스터디 107,299 샘플 (116,057 시퀀싱) 포함**

What's New @cbioportal

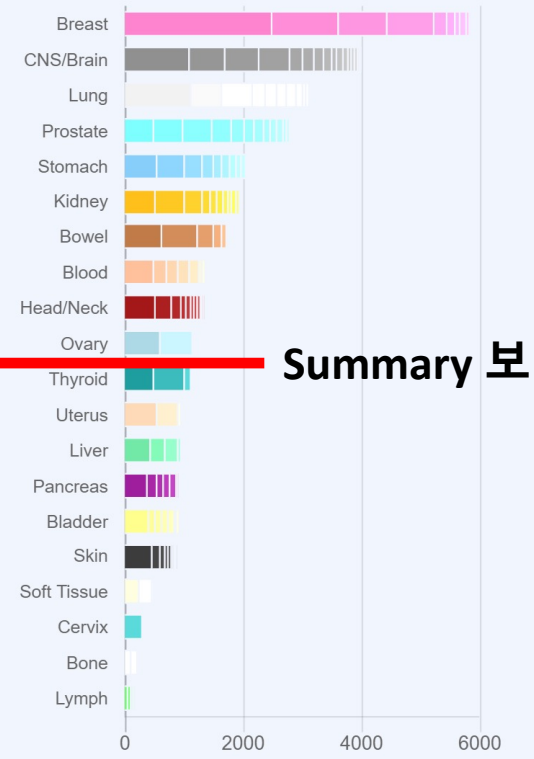
Sign up for low-volume email news alerts:

Subscribe

Cancer Studies

The portal contains 164 cancer studies (details)

Cases by Top 20 Primary Sites



Summary 보기

Example Queries

Q

- Googling
- Cohort abbreviation
  - For example
  - LUAD: Lung Adenocarcinoma
  - Breast cancer or prostate cancer ??

https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations

TCGA Study Abbreviations

Study Abbreviation	Study Name
LAML	Acute Myeloid Leukemia
ACC	Adrenocortical carcinoma
BLCA	Bladder Urothelial Carcinoma
LGG	Brain Lower Grade Glioma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
LCML	Chronic Myelogenous Leukemia
COAD	Colon adenocarcinoma
CNTL	Controls
ESCA	Esophageal carcinoma
FPPP	FFPE Pilot Phase II
GBM	Glioblastoma multiforme
HNSC	Head and Neck squamous cell carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma

Q

- Sample type

- For example
  - Primary tissue (cancer)
  - Normal

  - https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes

Sample Type Codes

Code	Definition	Short Letter Code
01	Primary Solid Tumor	TP
02	Recurrent Solid Tumor	TR
03	Primary Blood Derived Cancer - Peripheral Blood	TB
04	Recurrent Blood Derived Cancer - Bone Marrow	TRBM
05	Additional - New Primary	TAP
06	Metastatic	TM
07	Additional Metastatic	TAM
08	Human Tumor Original Cells	THOC
09	Primary Blood Derived Cancer - Bone Marrow	TBM
10	Blood Derived Normal	NB
11	Solid Tissue Normal	NT
12	Buccal Cell Normal	NBC
13	EBV Immortalized Normal	NEBV
14	Bone Marrow Normal	NBM
15	sample type 15	15SH
16	sample type 16	16SH
20	Control Analyte	CELLC
40	Recurrent Blood Derived Cancer - Peripheral Blood	TRB
50	Cell Lines	CELL
60	Primary Xenograft Tissue	XP
61	Cell Line Derived Xenograft Tissue	XCL
99	sample type 99	99SH

Tissue Source Site Codes

Prostate Adenocarcinoma (TCGA, Provisional) Study Summary 보기

cBioPortal FOR CANCER GENOMICS

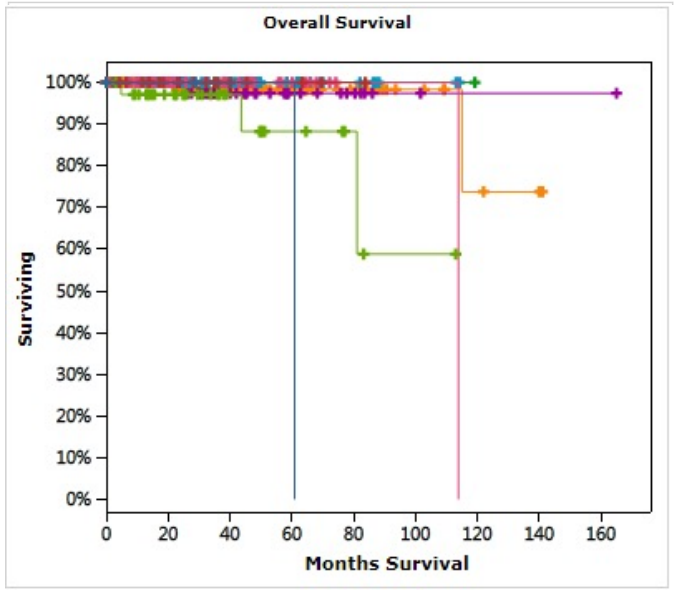
Data Sets | Web API | R/MATLAB | Tutorials | FAQ | News | Visualize Your Data | About

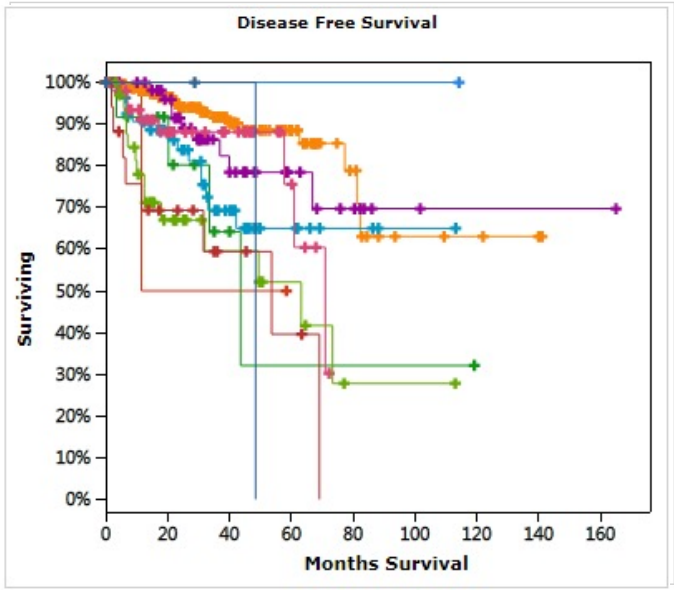
Prostate Adenocarcinoma (TCGA, Provisional) Query this study | Download data

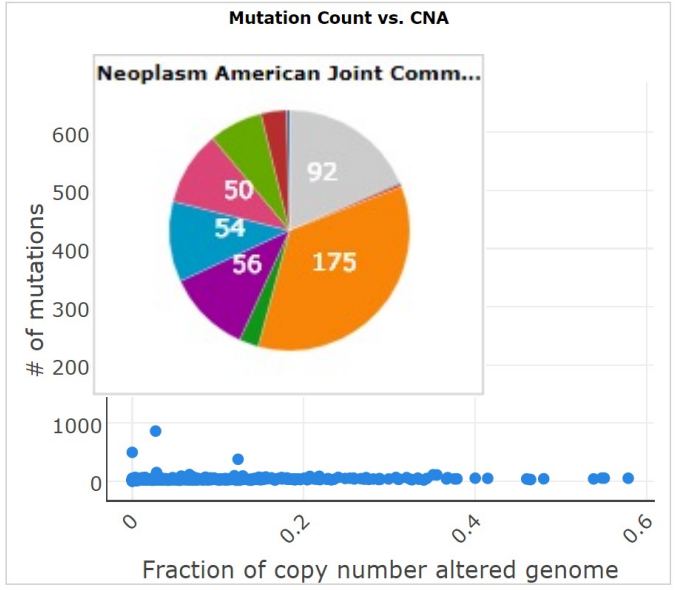
TCGA Prostate Adenocarcinoma; raw data at the NCI.

Study Summary | Clinical Data | Mutated Genes | Copy Number Alterations

Selected: 499 samples / 498 patients | query genes - click to expand | Query | Select cases by IDs | Add Chart





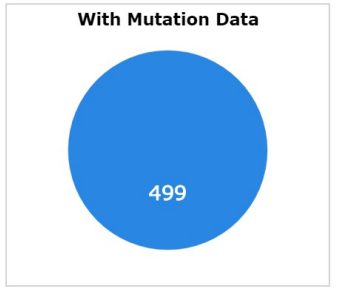


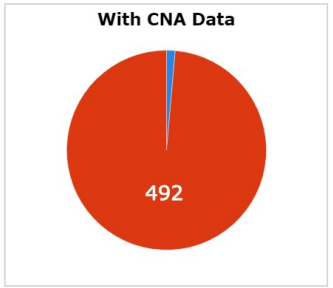
Mutated Genes (499 profiled samples)

Gene	# Mut	#	Freq
FRG1BP	134	97	19.44%
TP53Ⓜ	64	61	12.22%
SPOPⓂ	58	57	11.42%
MUC17Ⓜ	36	33	6.61%
KMT2C	35	30	6.01%

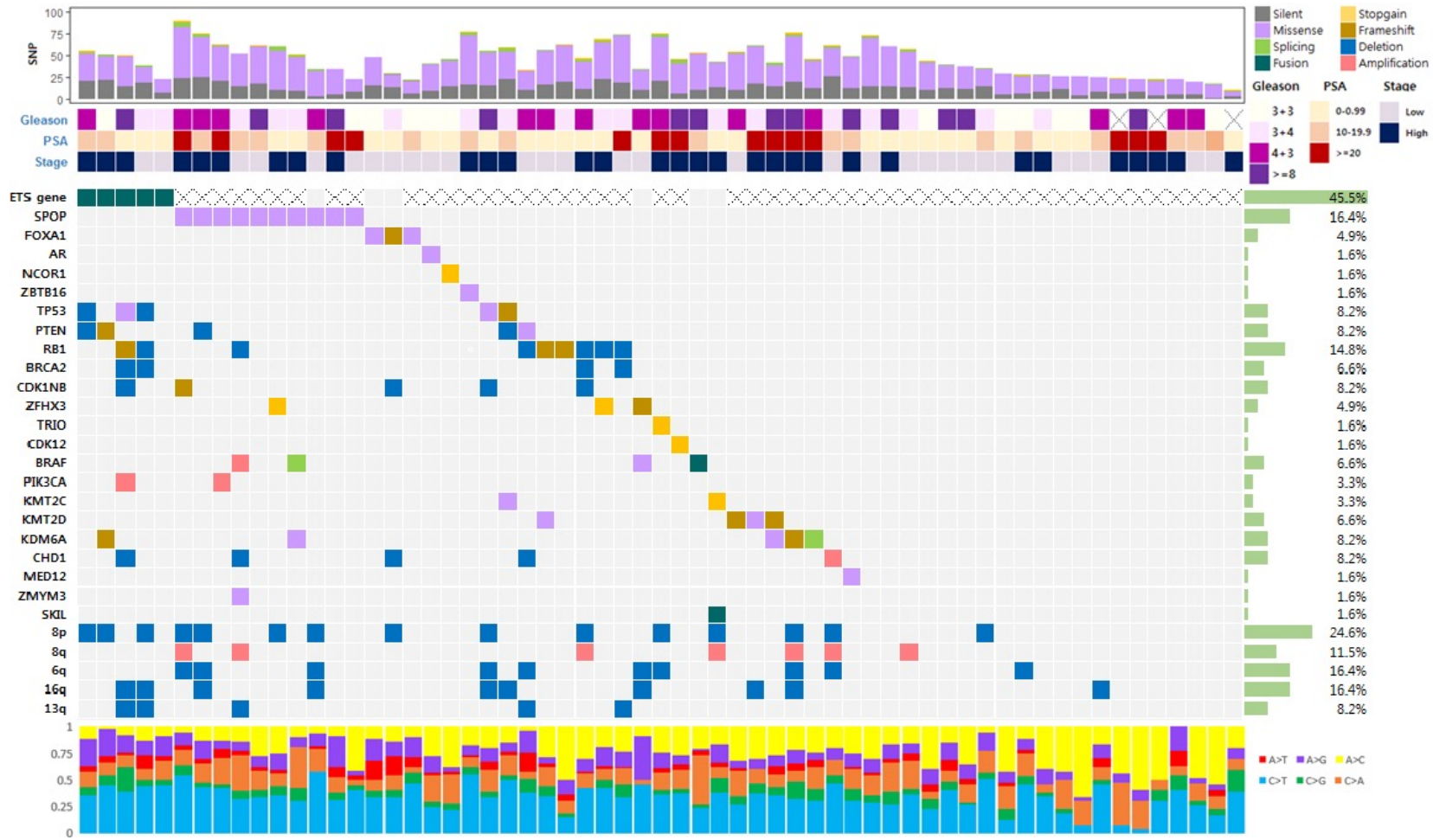
CNA Genes (492 profiled samples)

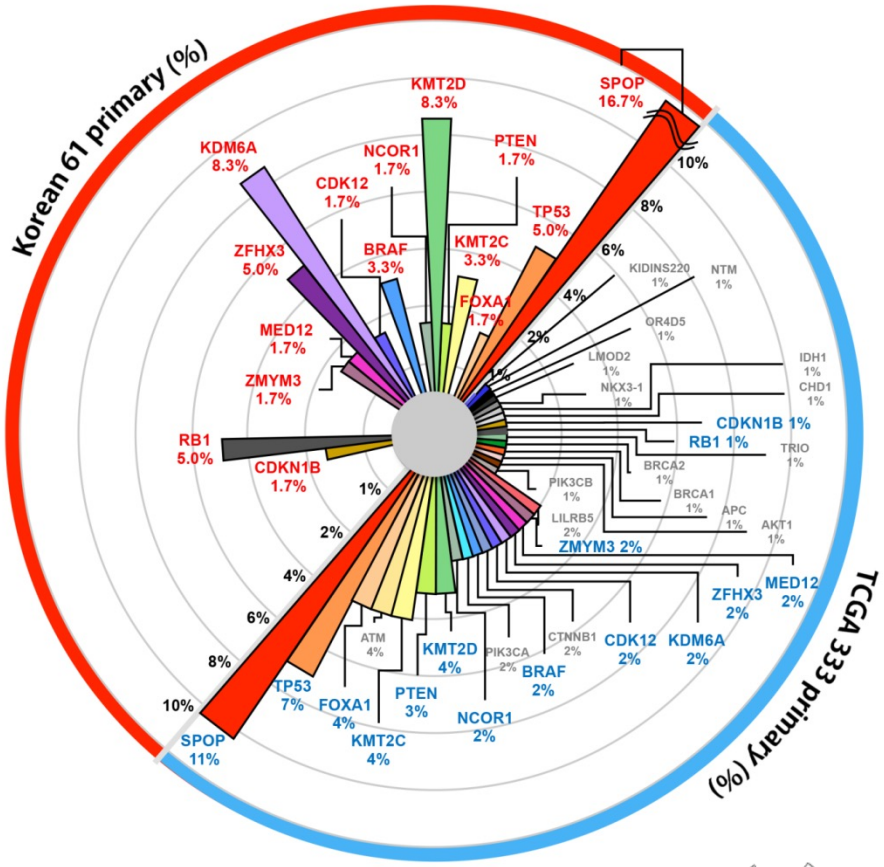
Gene	Cytoband	CNA	#	Freq
PTEN	10q23.3	DEL	95	19.31%
LCP1	13q14.3	DEL	85	17.28%
EGR3	8p23-p21	DEL	83	16.87%
RB1	13q14.2	DEL	81	16.46%
CYSLTR2	13q14.2	DEL	81	16.46%



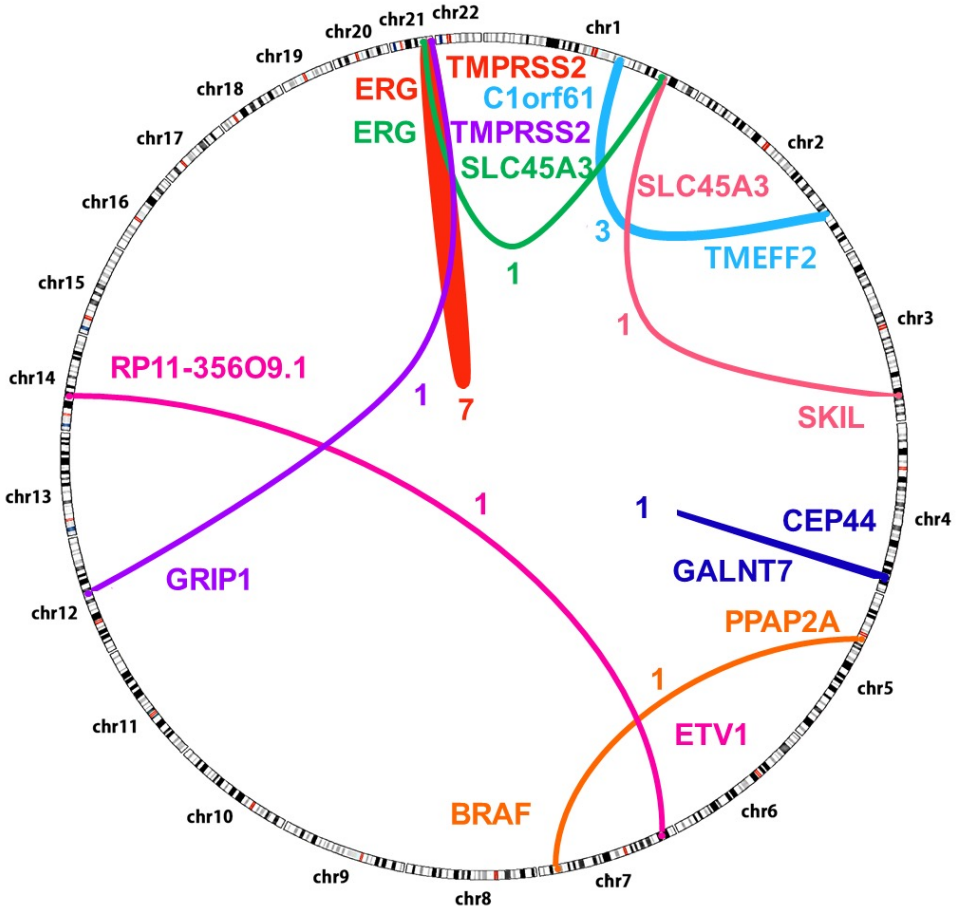


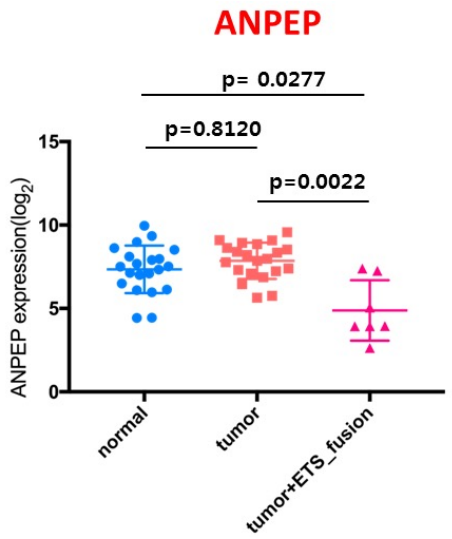
Distribution of genomic aberrations in Korean primary prostate cancer

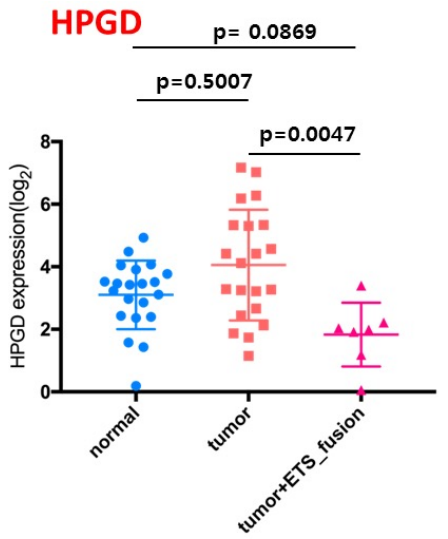


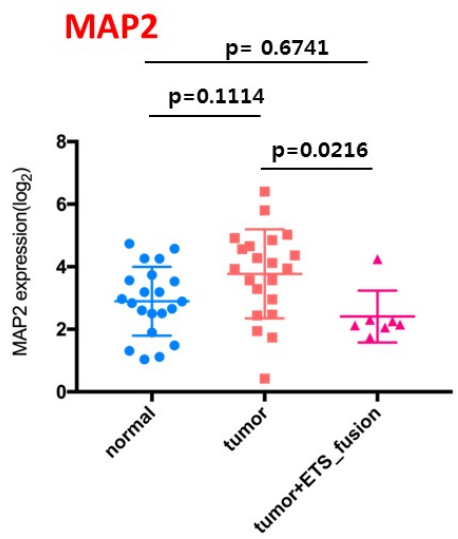


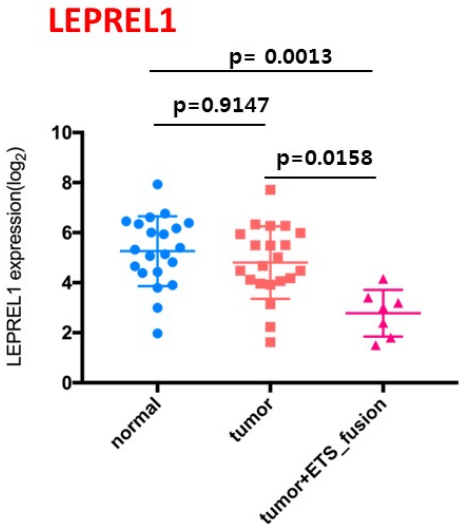
Gene fusions in Korean prostate cancer

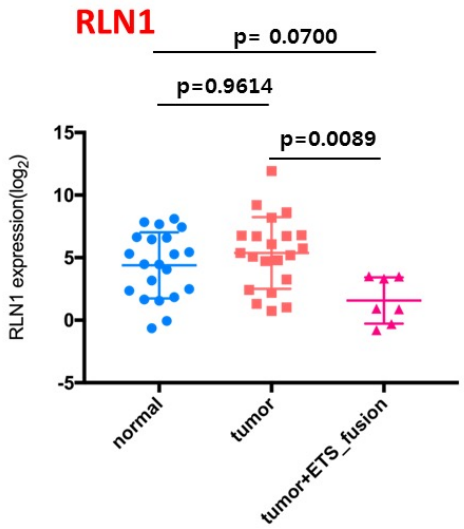


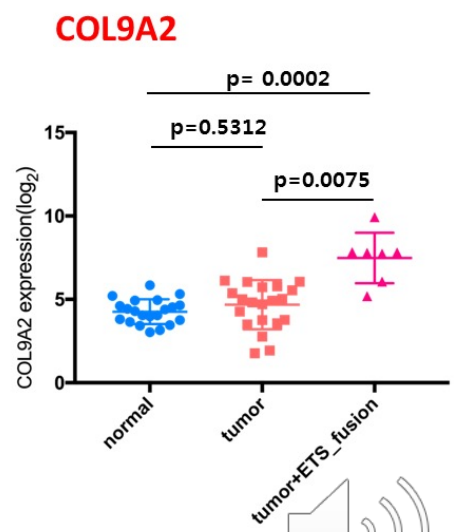




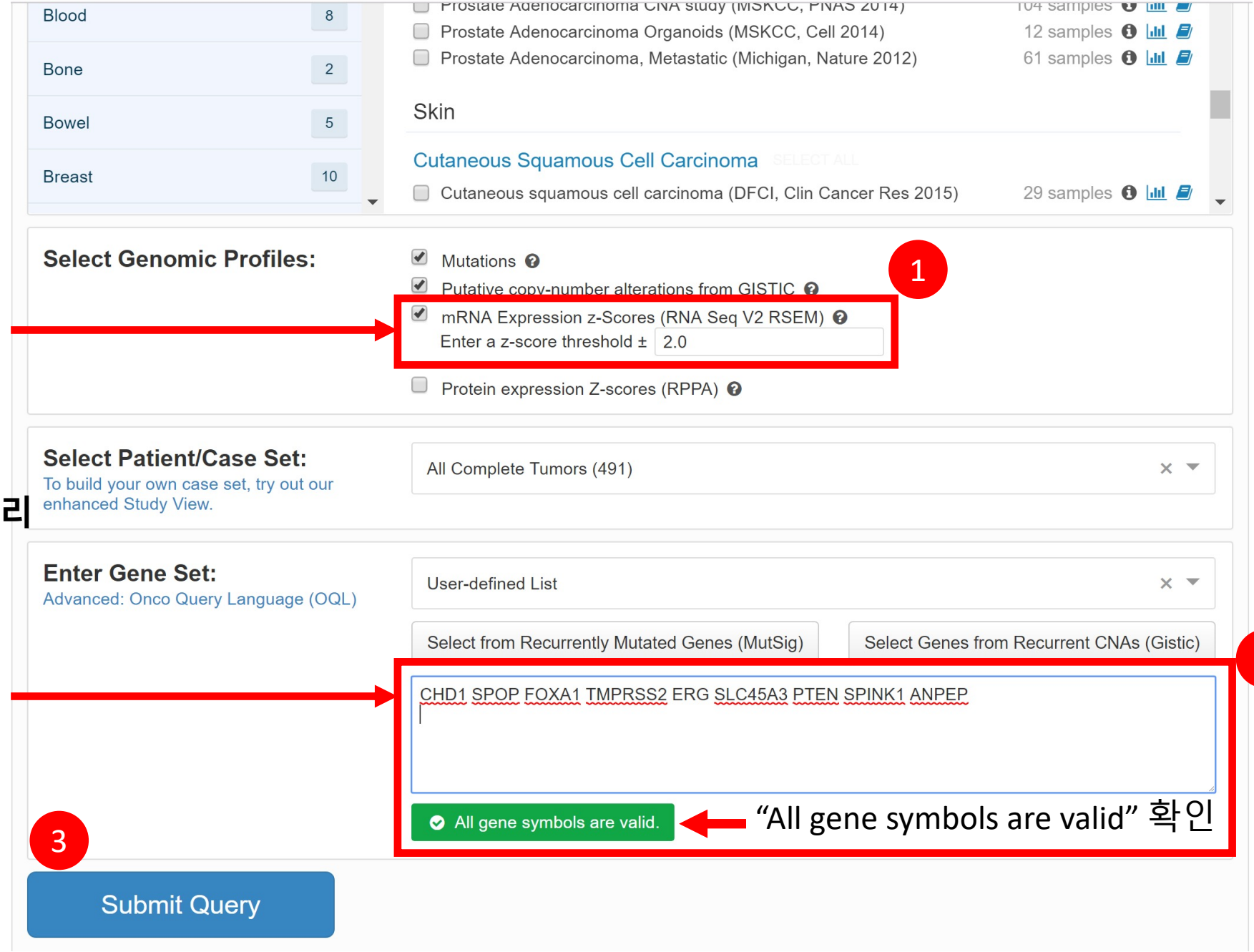








Prostate cancer related genes



mRNA Expression z-score 선택/설정

공간 (space)로 분리

- CHD1
- SPOP
- FOXA1
- TMPRSS2
- ERG
- SLC45A3
- PTEN
- SPINK1
- ANPEP

"All gene symbols are valid" 확인

Summary of cancer genes

cBioPortal FOR CANCER GENOMICS

Data Sets | Web API | R/MATLAB | Tutorials | FAQ | News | Visualize Your Data | About

Modify Query | Prostate Adenocarcinoma (TCGA, Provisional)

Gene Set / Pathway is altered in 354 (72.1%) of queried samples

변이가 일어난 샘플 비율 (유전자 클릭/움직여서 위치 수정 가능)

OncoPrint | Cancer Types Summary | Mutual Exclusivity | Plots | Mutations | Co-Expression | Enrichments | Survival | Network | CN Segments | Download | Bookmark

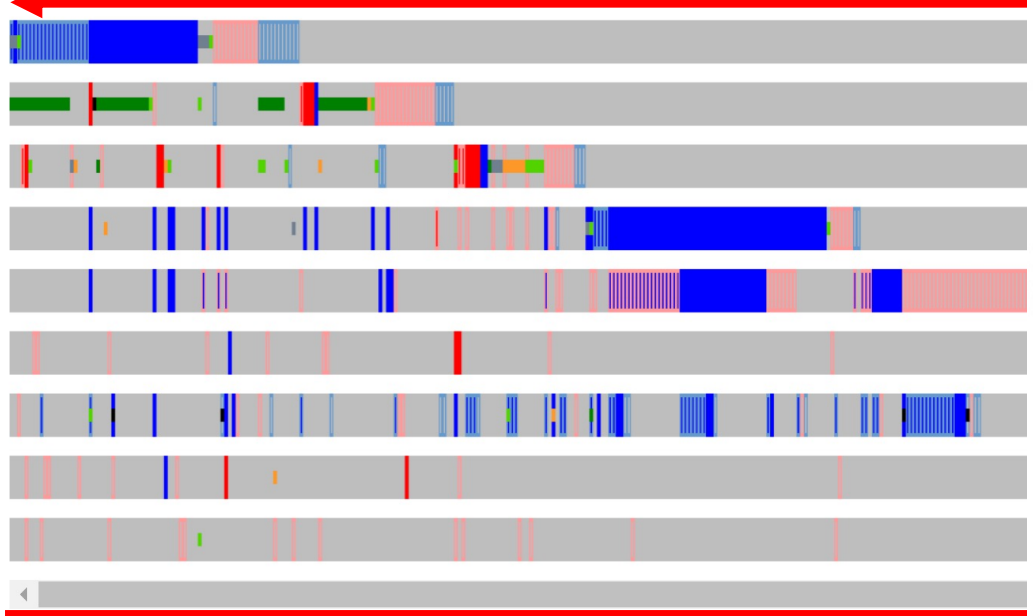
9개의 유전자 변이들이 있는 총 샘플 수

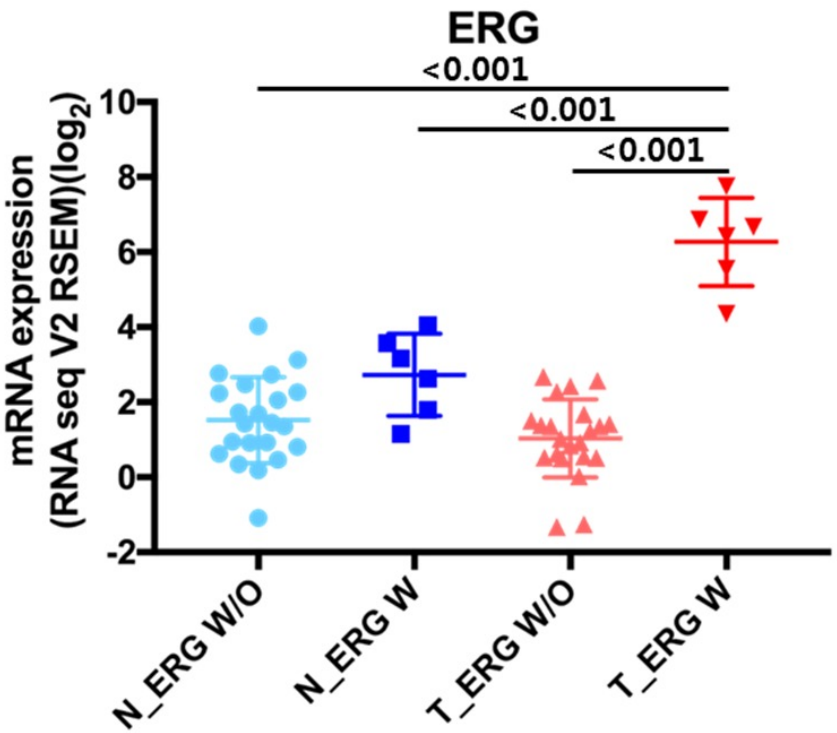
Case Set: All Complete Tumors (491 patients / 491 samples)

Altered in 354 (72%) of 491 sequenced cases/patients (491 total)

각 sample들에 대한 ger

CHD1	16%
SPOP	17%
FOXA1	11%
TMPRSS2	20%
ERG	24%
SLC45A3	4%
PTEN	28%
SPINK1	3%
ANPEP	6%





Genetic Alteration

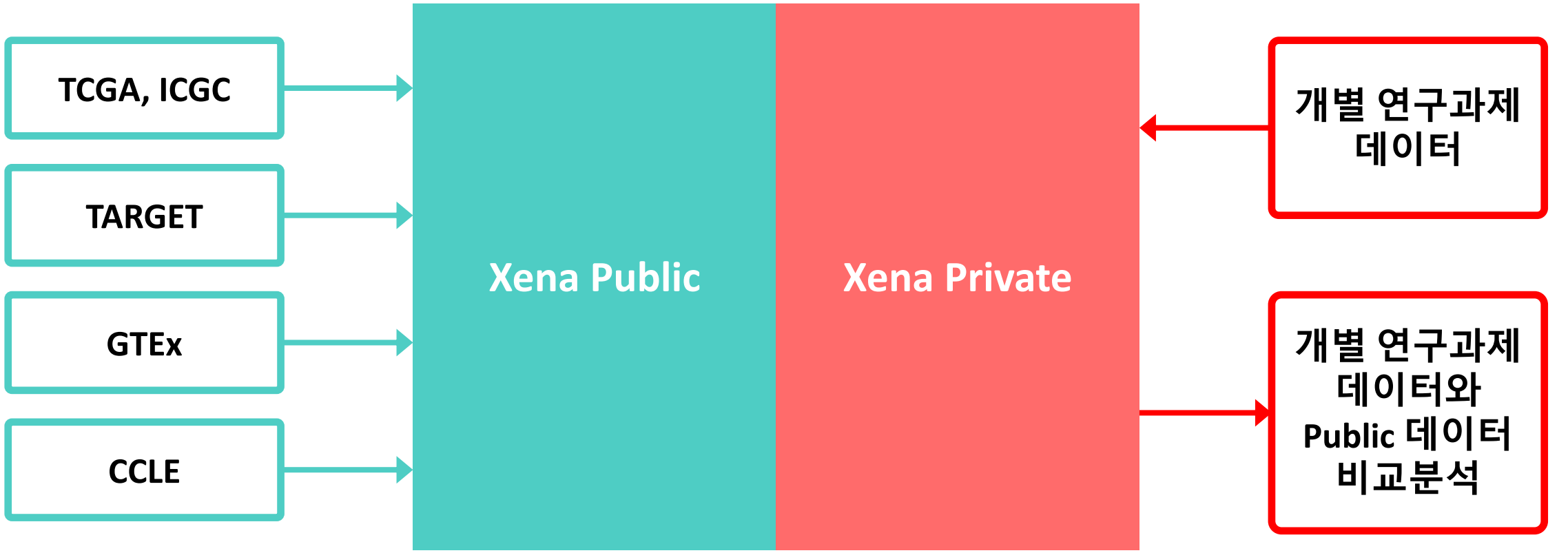
Amplification | Deep Deletion | mRNA Upregulation | mRNA Downregulation | Truncating Mutation (putative driver)

Truncating Mutation (putative passenger) | Inframe Mutation (putative passenger) | Missense Mutation (putative driver)

Missense Mutation (putative passenger)

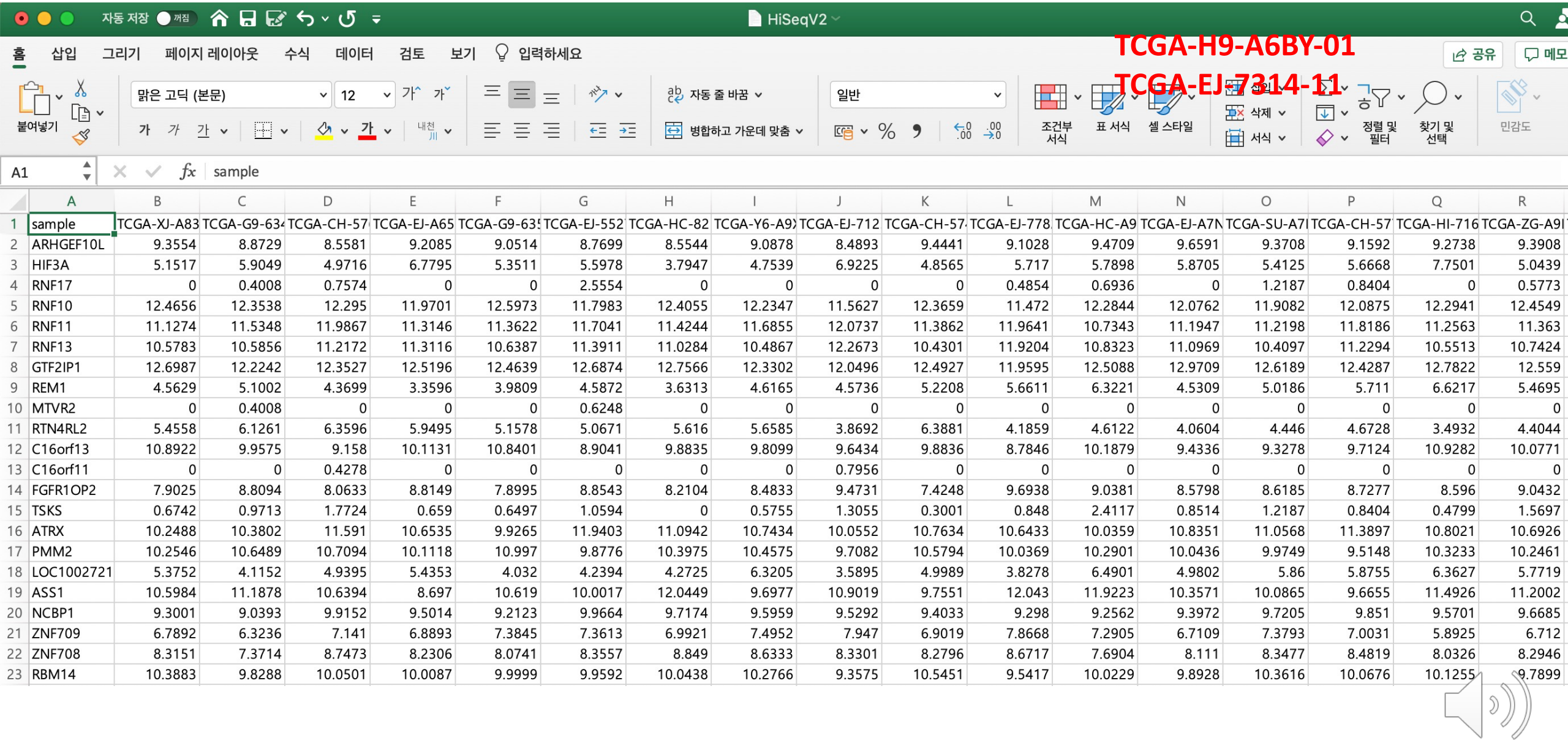
Xena browser (https://xenabrowser.net)

UCSC Xena Browser



(1) Visualization (2) Analysis (3) Download

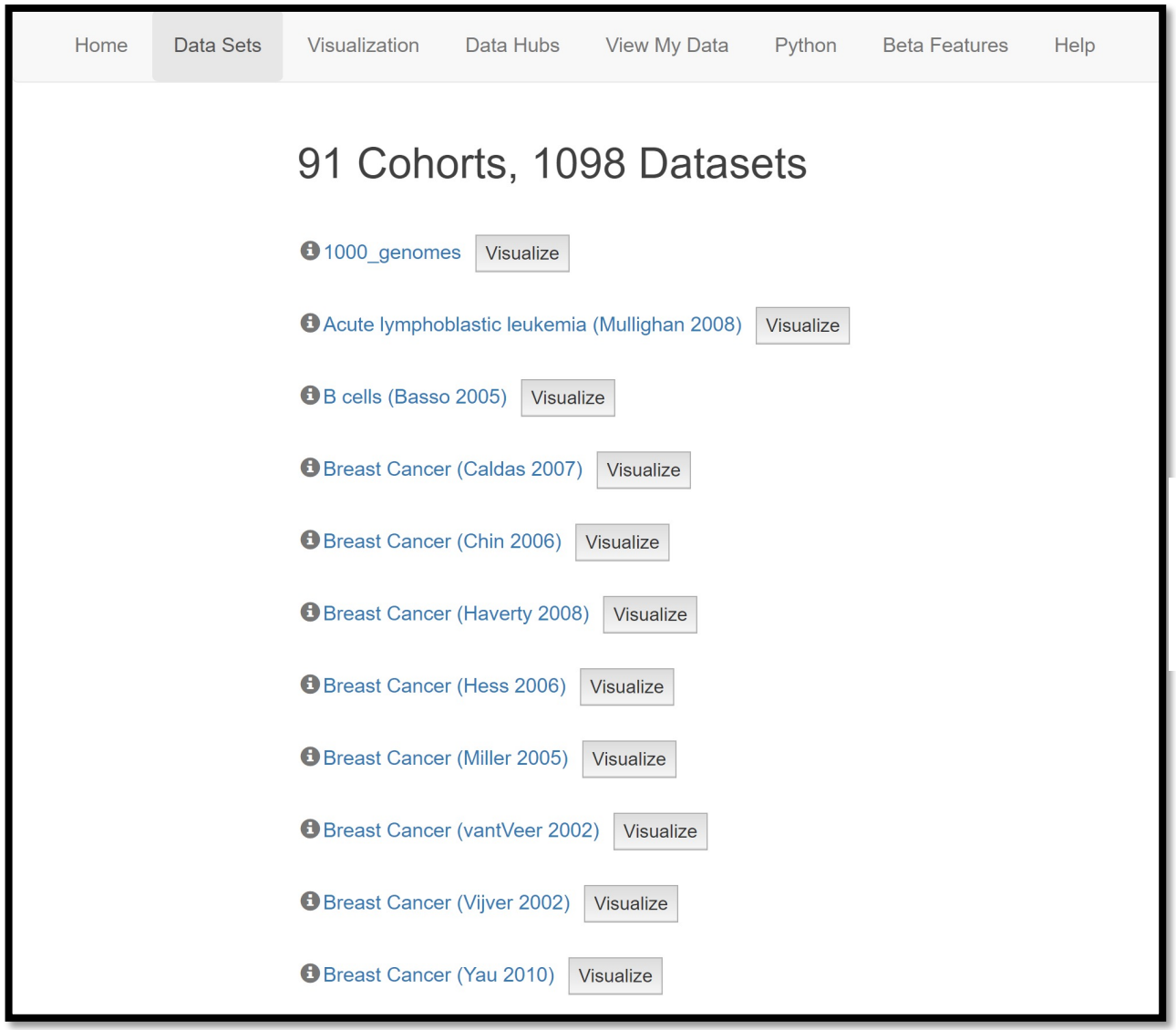
20,530 rows, 550 columns



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	sample	TCGA-XJ-A83	TCGA-G9-634	TCGA-CH-57	TCGA-EJ-A65	TCGA-G9-635	TCGA-EJ-552	TCGA-HC-82	TCGA-Y6-A9X	TCGA-EJ-712	TCGA-CH-57	TCGA-EJ-778	TCGA-HC-A9	TCGA-EJ-A7N	TCGA-SU-A7I	TCGA-CH-57	TCGA-HI-716	TCGA-ZG-A9I
2	ARHGEF10L	9.3554	8.8729	8.5581	9.2085	9.0514	8.7699	8.5544	9.0878	8.4893	9.4441	9.1028	9.4709	9.6591	9.3708	9.1592	9.2738	9.3908
3	HIF3A	5.1517	5.9049	4.9716	6.7795	5.3511	5.5978	3.7947	4.7539	6.9225	4.8565	5.717	5.7898	5.8705	5.4125	5.6668	7.7501	5.0439
4	RNF17	0	0.4008	0.7574	0	0	2.5554	0	0	0	0	0.4854	0.6936	0	1.2187	0.8404	0	0.5773
5	RNF10	12.4656	12.3538	12.295	11.9701	12.5973	11.7983	12.4055	12.2347	11.5627	12.3659	11.472	12.2844	12.0762	11.9082	12.0875	12.2941	12.4549
6	RNF11	11.1274	11.5348	11.9867	11.3146	11.3622	11.7041	11.4244	11.6855	12.0737	11.3862	11.9641	10.7343	11.1947	11.2198	11.8186	11.2563	11.363
7	RNF13	10.5783	10.5856	11.2172	11.3116	10.6387	11.3911	11.0284	10.4867	12.2673	10.4301	11.9204	10.8323	11.0969	10.4097	11.2294	10.5513	10.7424
8	GTF2IP1	12.6987	12.2242	12.3527	12.5196	12.4639	12.6874	12.7566	12.3302	12.0496	12.4927	11.9595	12.5088	12.9709	12.6189	12.4287	12.7822	12.559
9	REM1	4.5629	5.1002	4.3699	3.3596	3.9809	4.5872	3.6313	4.6165	4.5736	5.2208	5.6611	6.3221	4.5309	5.0186	5.711	6.6217	5.4695
10	MTVR2	0	0.4008	0	0	0	0.6248	0	0	0	0	0	0	0	0	0	0	0
11	RTN4RL2	5.4558	6.1261	6.3596	5.9495	5.1578	5.0671	5.616	5.6585	3.8692	6.3881	4.1859	4.6122	4.0604	4.446	4.6728	3.4932	4.4044
12	C16orf13	10.8922	9.9575	9.158	10.1131	10.8401	8.9041	9.8835	9.8099	9.6434	9.8836	8.7846	10.1879	9.4336	9.3278	9.7124	10.9282	10.0771
13	C16orf11	0	0	0.4278	0	0	0	0	0	0.7956	0	0	0	0	0	0	0	0
14	FGFR1OP2	7.9025	8.8094	8.0633	8.8149	7.8995	8.8543	8.2104	8.4833	9.4731	7.4248	9.6938	9.0381	8.5798	8.6185	8.7277	8.596	9.0432
15	TSKS	0.6742	0.9713	1.7724	0.659	0.6497	1.0594	0	0.5755	1.3055	0.3001	0.848	2.4117	0.8514	1.2187	0.8404	0.4799	1.5697
16	ATRX	10.2488	10.3802	11.591	10.6535	9.9265	11.9403	11.0942	10.7434	10.0552	10.7634	10.6433	10.0359	10.8351	11.0568	11.3897	10.8021	10.6926
17	PMM2	10.2546	10.6489	10.7094	10.1118	10.997	9.8776	10.3975	10.4575	9.7082	10.5794	10.0369	10.2901	10.0436	9.9749	9.5148	10.3233	10.2461
18	LOC1002721	5.3752	4.1152	4.9395	5.4353	4.032	4.2394	4.2725	6.3205	3.5895	4.9989	3.8278	6.4901	4.9802	5.86	5.8755	6.3627	5.7719
19	ASS1	10.5984	11.1878	10.6394	8.697	10.619	10.0017	12.0449	9.6977	10.9019	9.7551	12.043	11.9223	10.3571	10.0865	9.6655	11.4926	11.2002
20	NCBP1	9.3001	9.0393	9.9152	9.5014	9.2123	9.9664	9.7174	9.5959	9.5292	9.4033	9.298	9.2562	9.3972	9.7205	9.851	9.5701	9.6685
21	ZNF709	6.7892	6.3236	7.141	6.8893	7.3845	7.3613	6.9921	7.4952	7.947	6.9019	7.8668	7.2905	6.7109	7.3793	7.0031	5.8925	6.712
22	ZNF708	8.3151	7.3714	8.7473	8.2306	8.0741	8.3557	8.849	8.6333	8.3301	8.2796	8.6717	7.6904	8.111	8.3477	8.4819	8.0326	8.2946
23	RBM14	10.3883	9.8288	10.0501	10.0087	9.9999	9.9592	10.0438	10.2766	9.3575	10.5451	9.5417	10.0229	9.8928	10.3616	10.0676	10.1255	9.7899

Access Xena Browser, Select TCGA PRAD

https://xenabrowser.net/datapages/



1 리스트에서 TCGA Prostate Cancer (PRAD) 찾기

2 TCGA Prostate Cancer (PRAD) 클릭

Gene Expression 데이터 다운로드

dataset: gene expression RNAseq (polyA+ IlluminaHiSeq)

Visualize

1 Download

TCGA prostate adenocarcinoma (PRAD) gene expression by RNAseq (polyA+ IlluminaHiSeq)

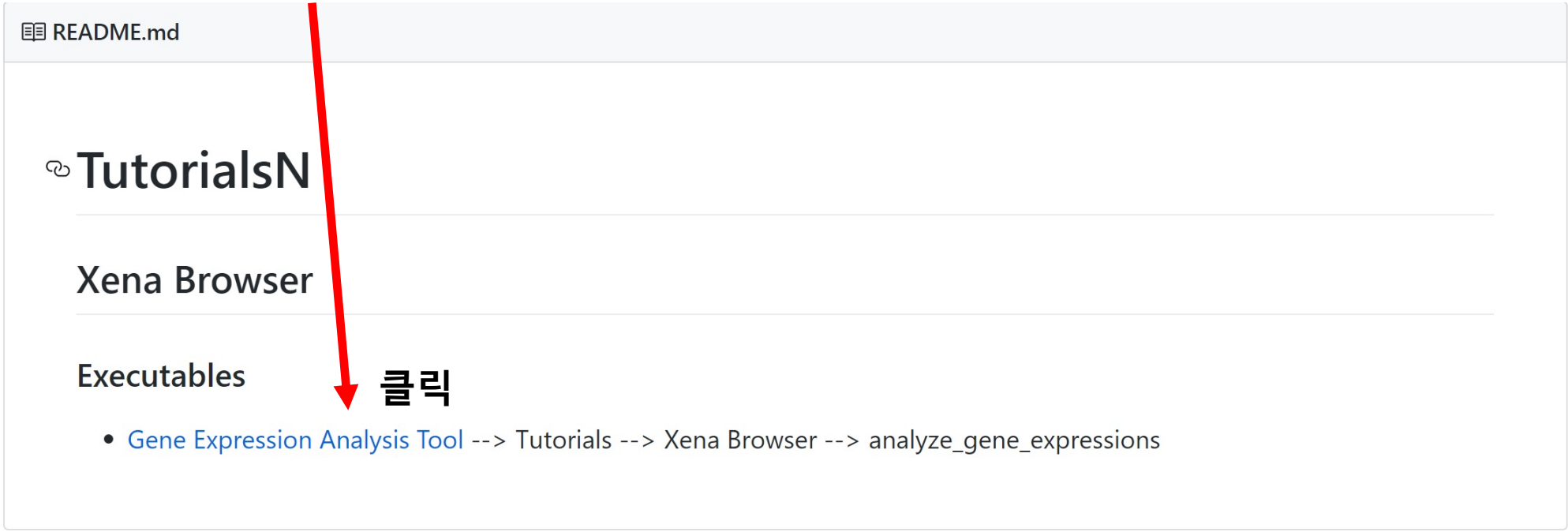
The gene expression profile was measured experimentally using the Illumina HiSeq 2000 RNA Sequencing platform by the University of North Carolina TCGA genome characterization center. Level 3 data was downloaded from TCGA data coordination center. This dataset shows the gene-level transcription estimates, as in log2(x+1) transformed RSEM normalized count. Genes are mapped onto the human genome coordinates using UCSC Xena HUGO probeMap (see ID/Gene mapping link below for details). Reference to method description from University of North Carolina TCGA genome characterization center: DCC description

2 **HiSeqV2.gz 다운로드됨**

Gene Expression 데이터로 간단한 분석하기

1. https://github.com/cgab-ncc/Tutorials 로 이동
   (Tutorials 소스코드 공유 페이지)

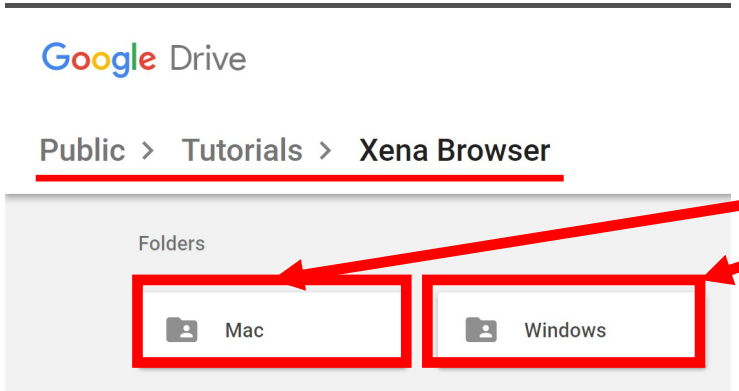
2. Google Drive로 이동



Gene Expression 데이터로 간단한 분석하기

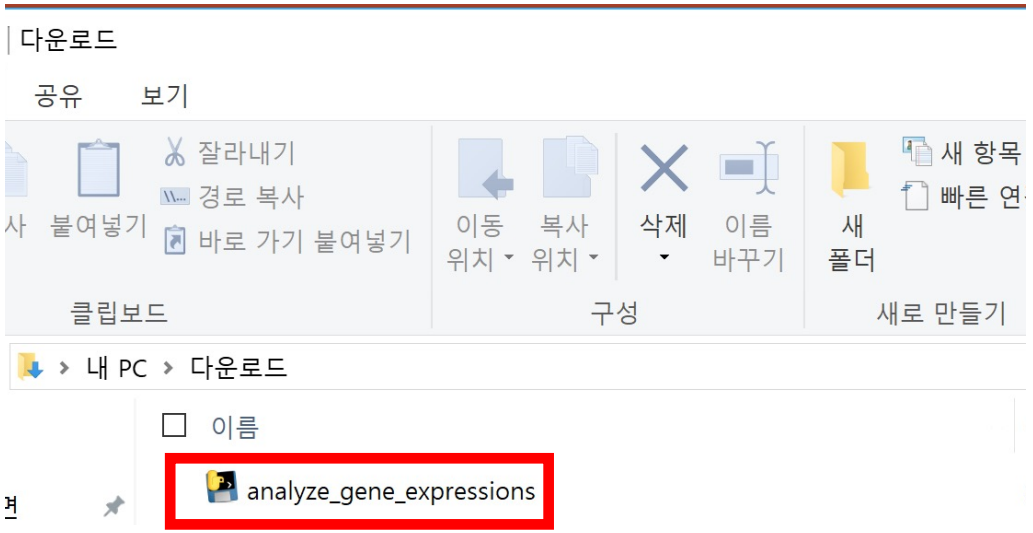
3. Tutorials → Xena Browser 로 이동

4. OS (운영체제)에 맞게 Mac 혹은 Windows 버전 (analyze_gene_expressions) 다운로드



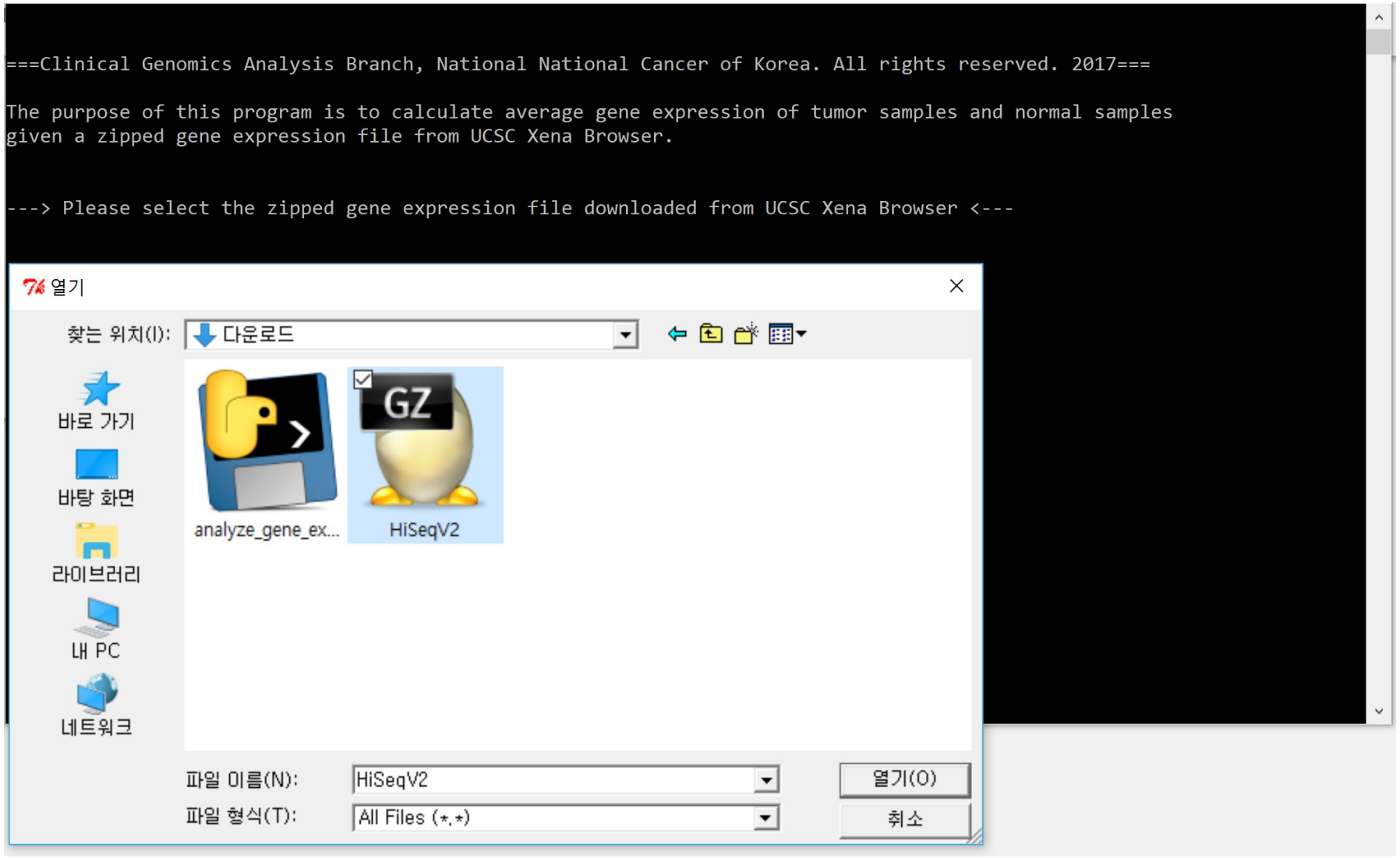
폴더 안에 있는
analyze_gene_expressions
응용 프로그램 다운로드

5. analyze_gene_expressions 더블 클릭하여 프로그램 실행



Gene Expression 데이터로 간단한 분석하기

6. "열기" 창이 나오면 Step 1 에서 다운로드 했던 HiSeqV2 (gene expression file) 선택/열기 (**프로그램 시작 시 초기화 시간: 약 15초**)

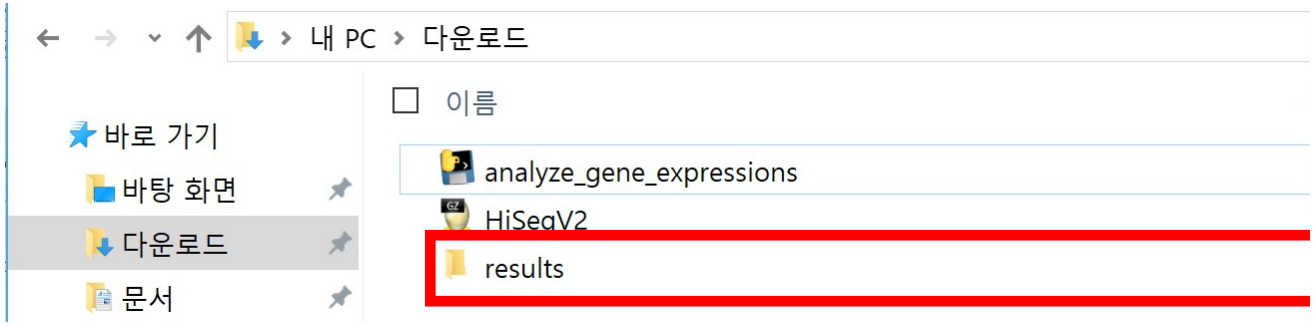


Gene Expression 데이터로 간단한 분석하기

7. 프로그램 수행 중인지의 확인

```
===Clinical Genomics Analysis Branch, National National Cancer of Korea. All rights reserved. 2017===

The purpose of this program is to calculate average gene expression of tumor samples and normal samples
given a zipped gene expression file from UCSC Xena Browser.


---> Please select the zipped gene expression file downloaded from UCSC Xena Browser <---


Calculating average gene expression ratios between tumor and normal
```

8. 프로그램이 다 끝나면 analyze_gene_expressions 생성(컴퓨터 사양에 따라 최대 10분 소요 가능)




```
dwhong — analyze_gene_expressions — 80×24
Last login: Fri Aug 25 09:56:36 on ttys001
dwhong-ui-MacBook-Pro:~ dwhong$ /Users/dwhong/Documents/2017/Analysis/analyze_ge
ne_expressions ; exit;


===Clinical Genomics Analysis Branch, National National Cancer of Korea. All rig
hts reserved. 2017===

The purpose of this program is to calculate average gene expression of tumor sam
ples and normal samples
given a zipped gene expression file from UCSC Xena Browser.


---> Please select the zipped gene expression file downloaded from UCSC Xena Bro
wser <---


Calculating average gene expression ratios between tumor and normal samples 99%
Finished calculating average gene expression value ratio
Started writing data to .csv files
Writing data to analysis.csv file
Finished writing data to analysis.csv file
Writing data to gene_expressions.csv file
Finished writing data to .csv files
```

Gene Expression 데이터로 간단한 분석하기

9. results 폴더에 생성된 있는 파일 필드들에 대한 설명

analysis.csv

	필드	설명
1	gene_name	유전자 이름
2	tumor_samples_count	tumor 샘플 개수

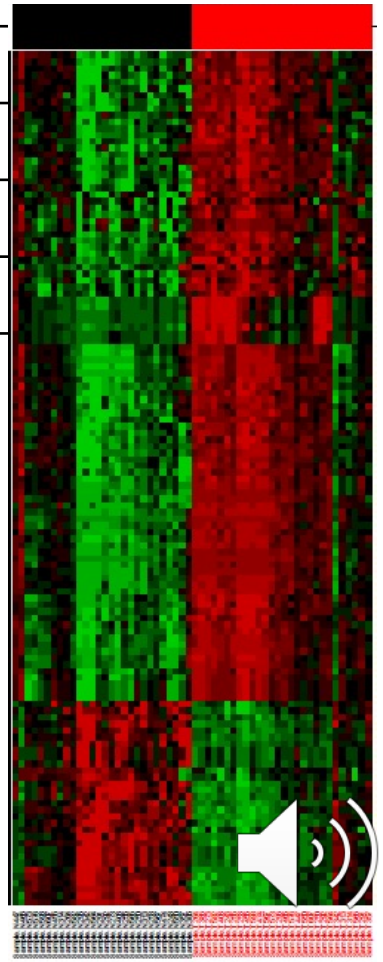
gene_name	tumor_samples_count	normal_samples_count	avg(tumor)	avg(normal)	avg(tumor)/avg(normal)	avg(tumor)=0_count	avg(normal)=0_count
NKX2-3	497	52	4.364812072	0.857811538	5.088311216	13	21
ZIC2	497	52	5.140035815	1.246317308	4.124179118	14	12
SLC45A2	497	52	5.236834406	1.359369231	3.852400281	3	5
DLX2	497	52	5.094922334	1.327380769	3.838327669	11	11
LOC100287718	497	52	2.125735815	0.572592308	3.712477074	46	21
MMP26	497	52	3.777932394	1.140948077	3.311222019	15	16
LOC100128675	497	52	3.508176459	1.075101923	3.263110579	4	10
DLX1	497	52	7.984439034	2.697932692	2.959465615	1	2
ANGPTL3	497	52	2.56153501	0.902688462	2.837673371	38	19
EPHA8	497	52	2.863710463	1.025198077	2.793324068	15	14
C15orf50	497	52	2.157590141	0.801048077	2.693458986	28	16
GHRHR	497	52	2.839915694	1.079544231	2.63066173	25	16
FOXN4	497	52	2.469121932	0.942075	2.620939874	32	6
RPRML	497	52	2.270581891	0.875782692	2.59263161	46	17
PHGR1	497	52	4.465725956	1.731853846	2.578581308	8	7
MIOX	497	52	2.012503622	0.812188462	2.477877632	24	13
LOC339674	497	52	2.04588672	0.834109615	2.452779206	12	1
TDRD1	497	52	6.499556338	2.667855769	2.436247271	2	
AMH	497	52	3.219250704	1.329751923	2.420940815	18	6
HOXC5	497	52	4.580359155	1.903648077	2.406095544	3	7

Gene Expression 데이터로 간단한 분석하기

9. results 폴더에 생성된 있는 파일 필드들에 대한 설명

gene_expressions.csv

	필드	설명
1	gene_name	유전자 이름
2	T	Tumor group
3	N	Normal group



Matrix

Experiment or Sample

Gene

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{array}$$

NORMAL (CONTROL)

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{array}$$

TUMOR (TREATMENT)

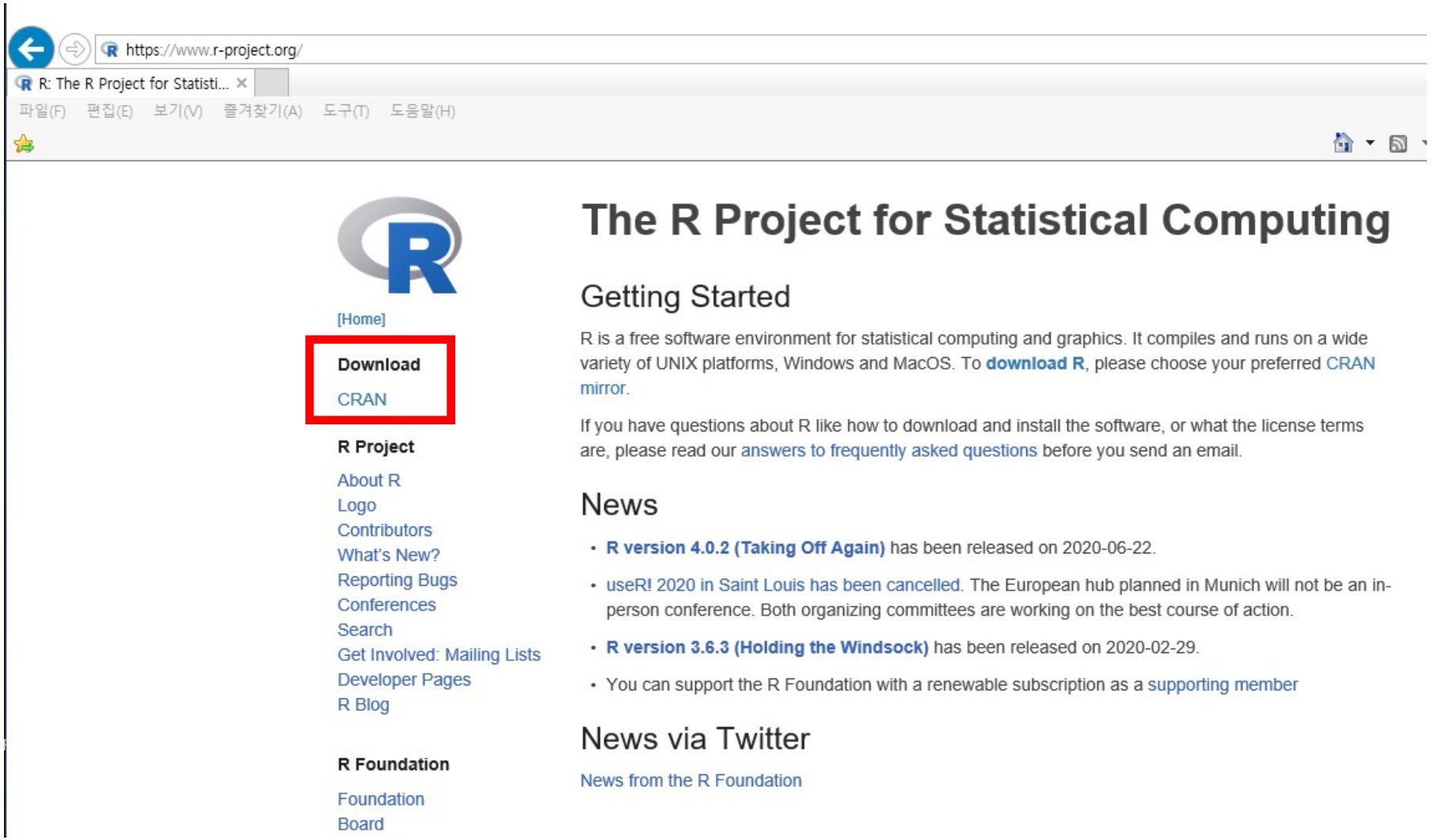
An $m \times n$ matrix: the m rows are horizontal and the n columns are vertical. Each element of a matrix is often denoted by a variable with two subscripts. For example, $a_{2,1}$ represents the element at the second row and first column of the matrix.

https://en.wikipedia.org/wiki/Matrix_(mathematics)

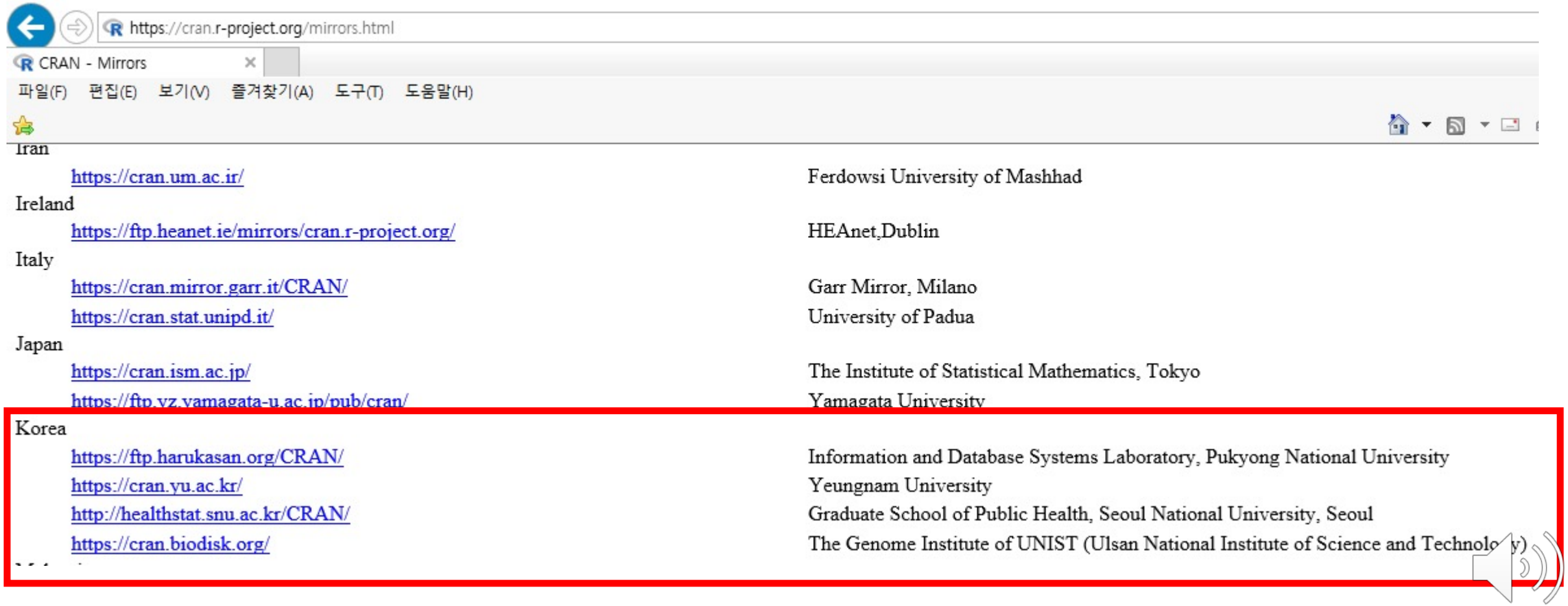
To download and install R software

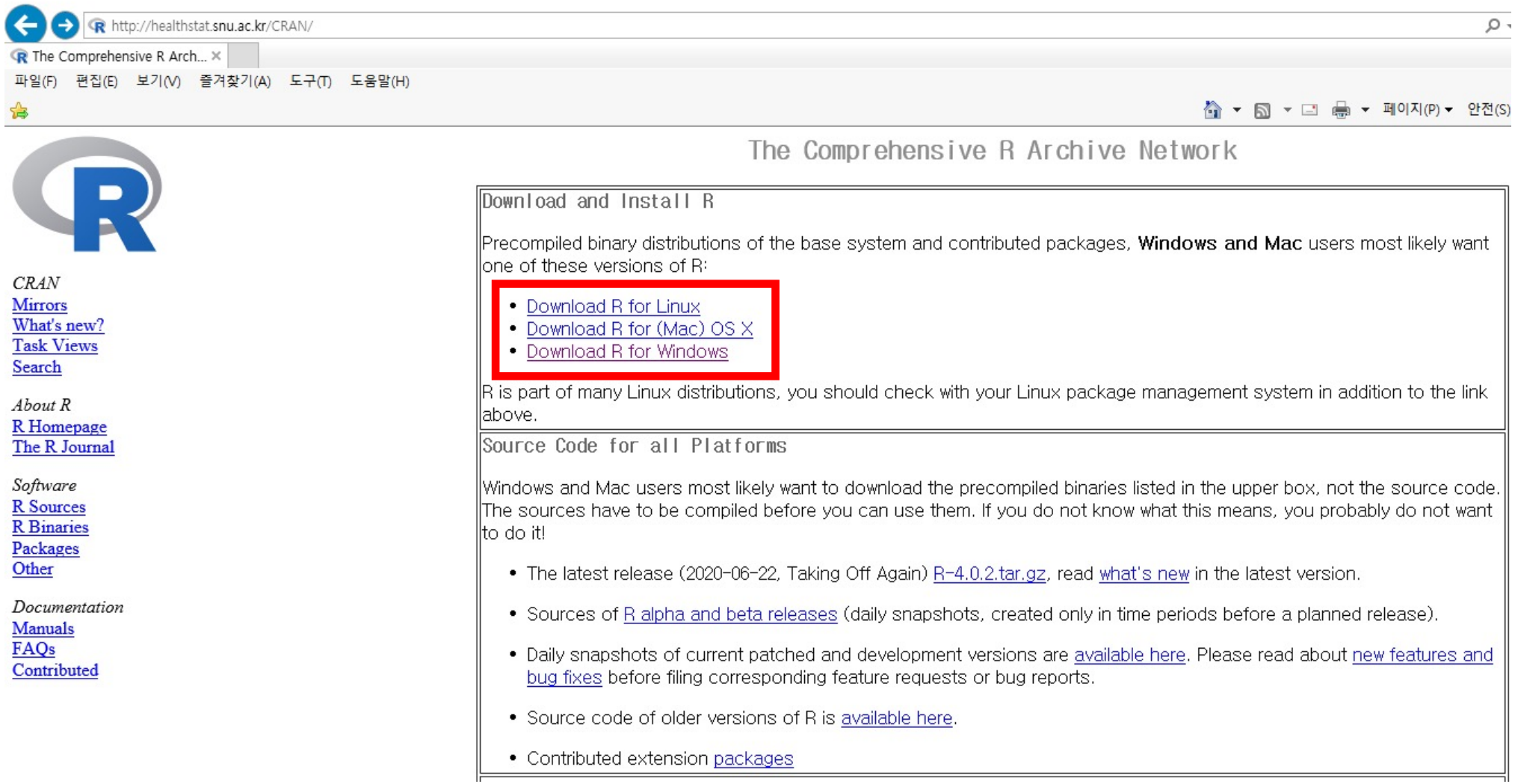
- R packages
  - https://cran.seoul.go.kr
- R studio
  - https://www.rstudio.com/products/rstudio/download/#download

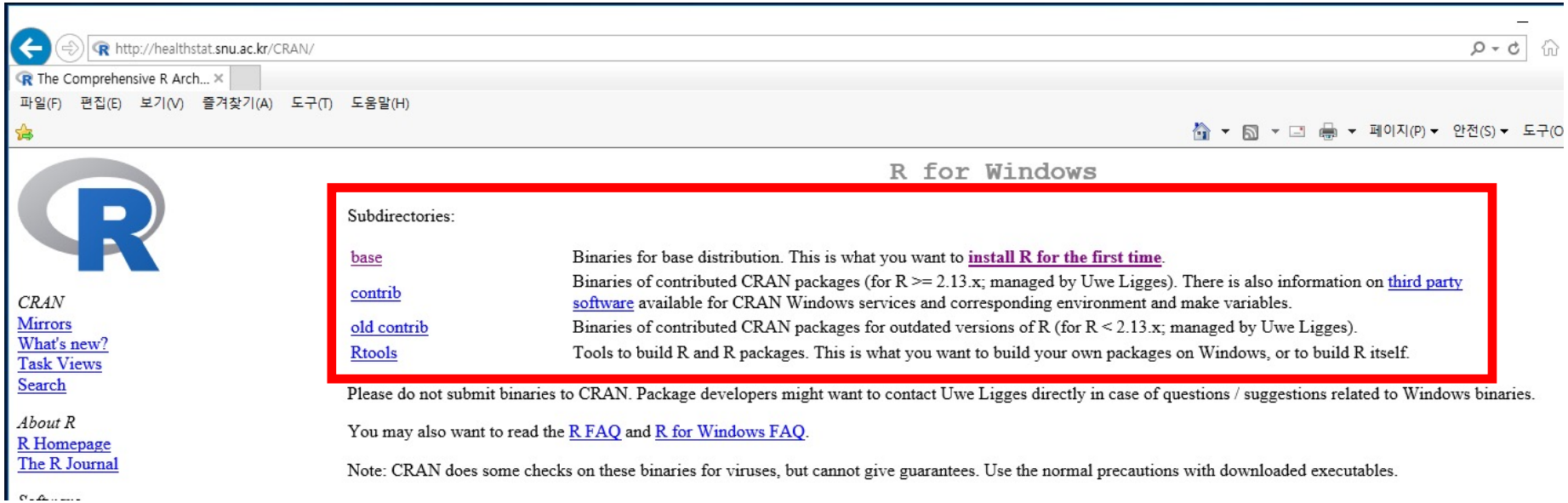
To download and install R software

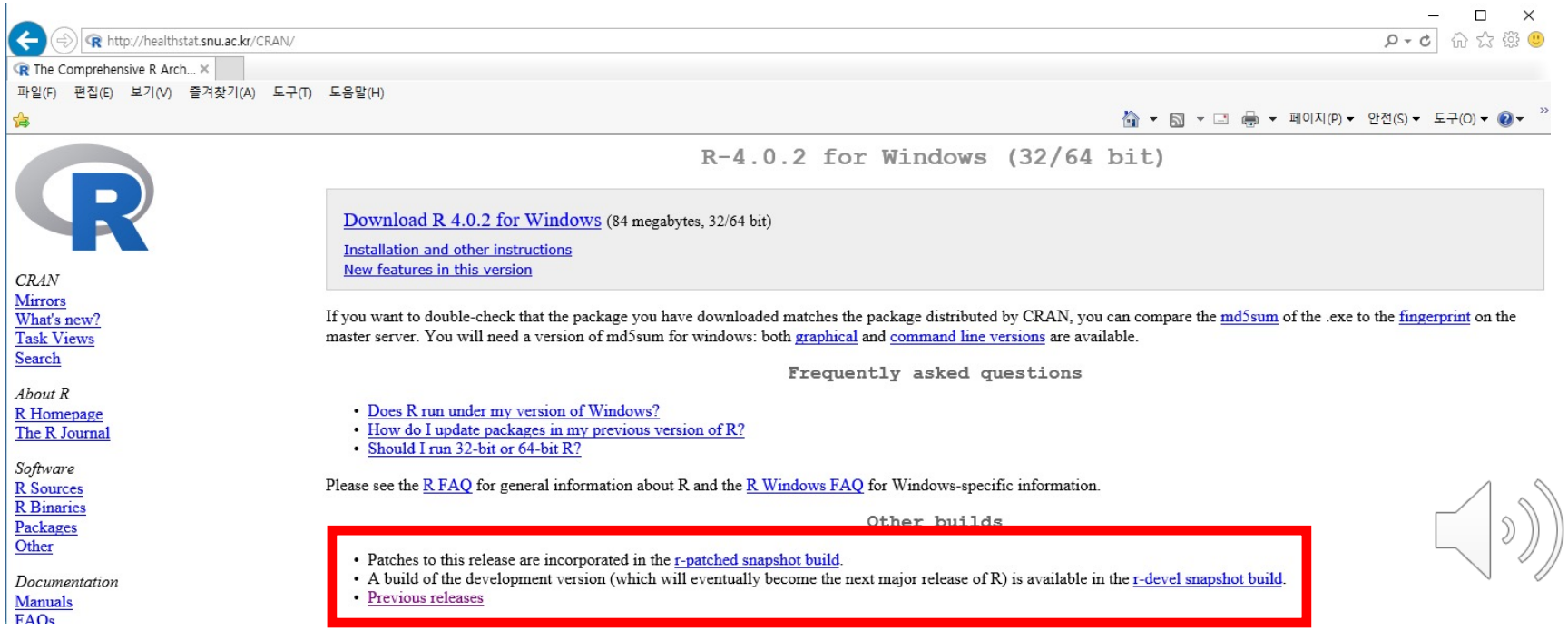


- R packages
  - https://cran.seoul.go.kr

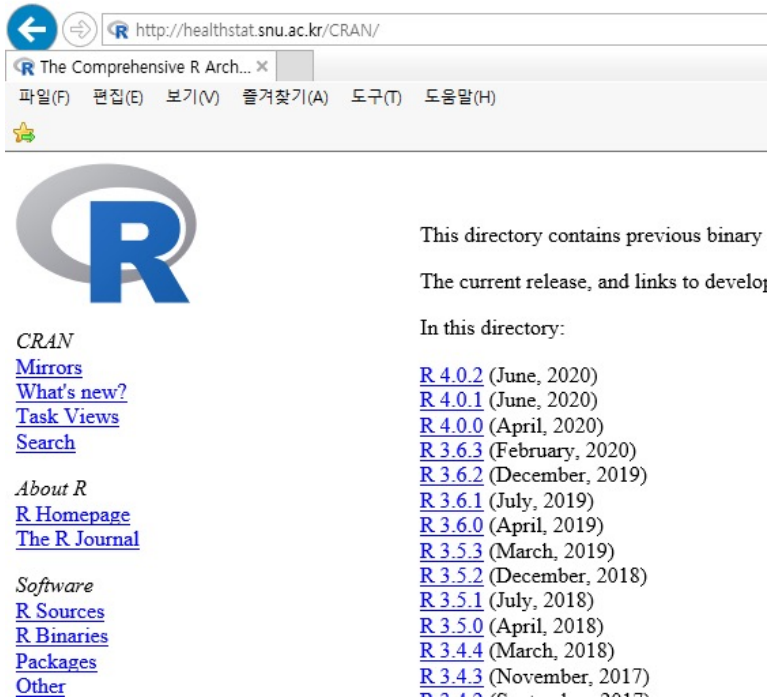


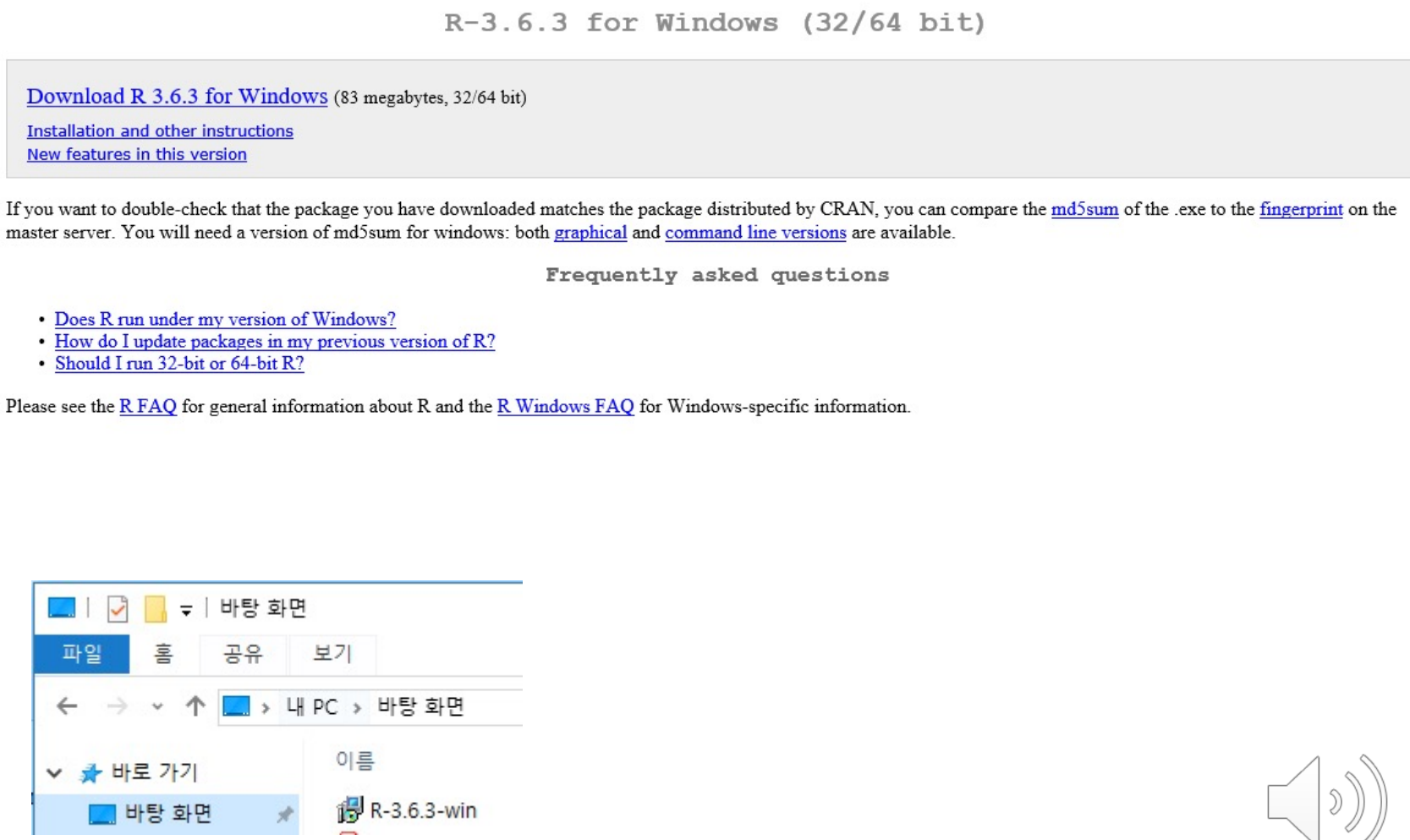

http://healthstat.snu.ac.kr/CRAN/
The Comprehensive R Arch...
파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)
페이지(P) 안전(S)
The Comprehensive R Archive Network
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
Download and Install R
Precompiled binary distributions of the base system and contributed packages, Windows and Mac users most likely want one of these versions of R:
Download R for Linux
Download R for (Mac) OS X
Download R for Windows
R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.
Source Code for all Platforms
Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!
The latest release (2020-06-22, Taking Off Again) R-4.0.2.tar.gz, read what's new in the latest version.
Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
Source code of older versions of R is available here.
Contributed extension packages


http://healthstat.snu.ac.kr/CRAN/
The Comprehensive R Arch...
파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)
페이지(P) 안전(S) 도구(O
R for Windows
Subdirectories:
base
Binaries for base distribution. This is what you want to install R for the first time.
contrib
Binaries of contributed CRAN packages (for R >= 2.13.x; managed by Uwe Ligges). There is also information on third party software available for CRAN Windows services and corresponding environment and make variables.
old contrib
Binaries of contributed CRAN packages for outdated versions of R (for R < 2.13.x; managed by Uwe Ligges).
Rtools
Tools to build R and R packages. This is what you want to build your own packages on Windows, or to build R itself.
Please do not submit binaries to CRAN. Package developers might want to contact Uwe Ligges directly in case of questions / suggestions related to Windows binaries.
You may also want to read the R FAQ and R for Windows FAQ.
Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded executables.
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal

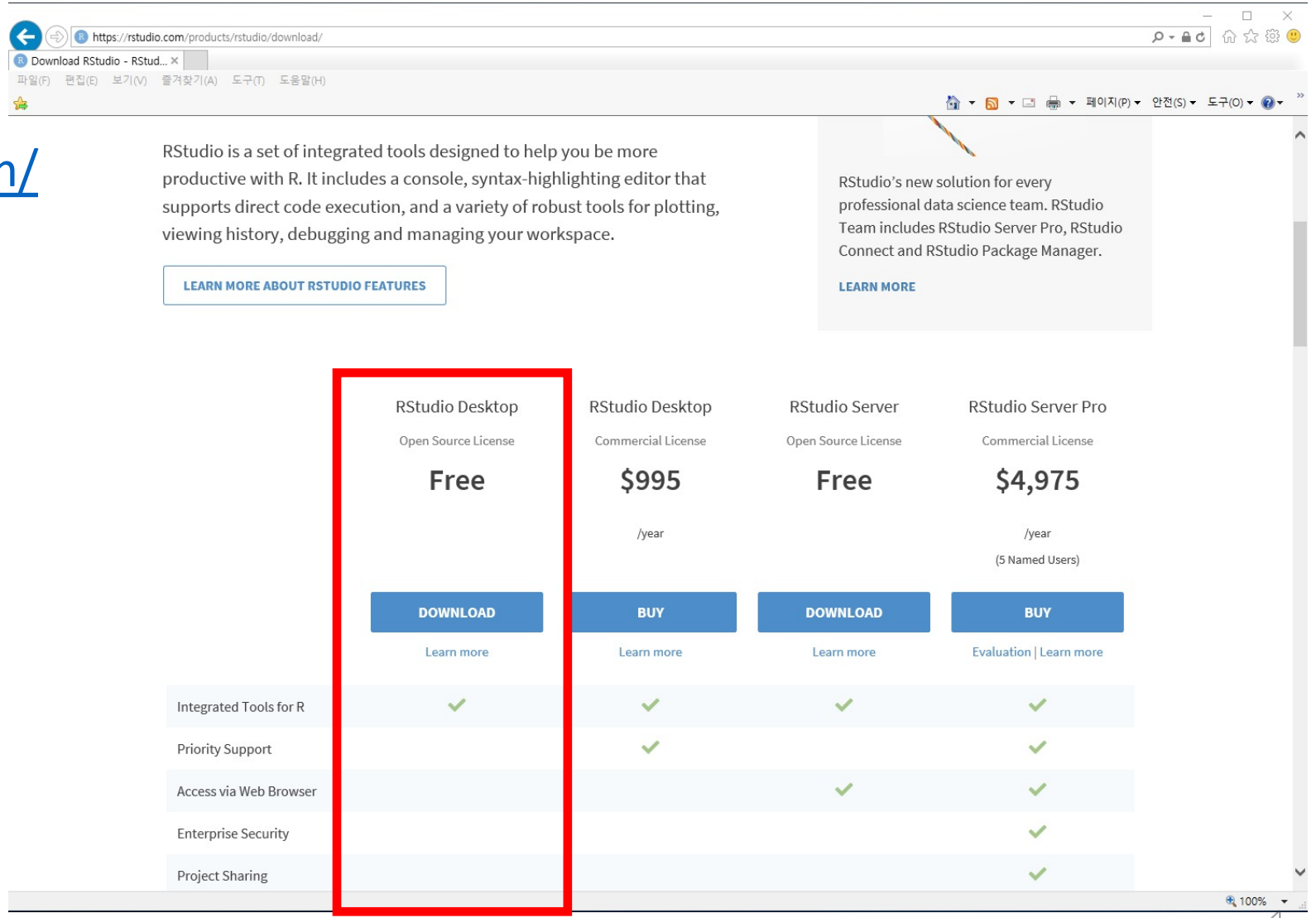

http://healthstat.snu.ac.kr/CRAN/
The Comprehensive R Arch...
파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)
페이지(P) 안전(S) 도구(O)
R-4.0.2 for Windows (32/64 bit)
Download R 4.0.2 for Windows (84 megabytes, 32/64 bit)
Installation and other instructions
New features in this version
If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the md5sum of the .exe to the fingerprint on the master server. You will need a version of md5sum for windows: both graphical and command line versions are available.
Frequently asked questions
Does R run under my version of Windows?
How do I update packages in my previous version of R?
Should I run 32-bit or 64-bit R?
Please see the R FAQ for general information about R and the R Windows FAQ for Windows-specific information.
Other builds
Patches to this release are incorporated in the r-patched snapshot build.
A build of the development version (which will eventually become the next major release of R) is available in the r-devel snapshot build.
Previous releases
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs

Stable version (3.6.3)





- R studio
  - https://www.rstudio.com/products/rstudio/download/#download



https://rstudio.com/products/rstudio/download/#download

Download RStudio - RStud...

파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)

페이지(P) 안전(S) 도구(O)

RStudio Desktop 1.3.1073 - Release Notes

1. Install R. RStudio requires R 3.0.1+.

2. Download RStudio Desktop. Recommended for your system:

DOWNLOAD RSTUDIO FOR WINDOWS
1.3.1073 | 171.62MB

Requires Windows 10/8/7 (64-bit)

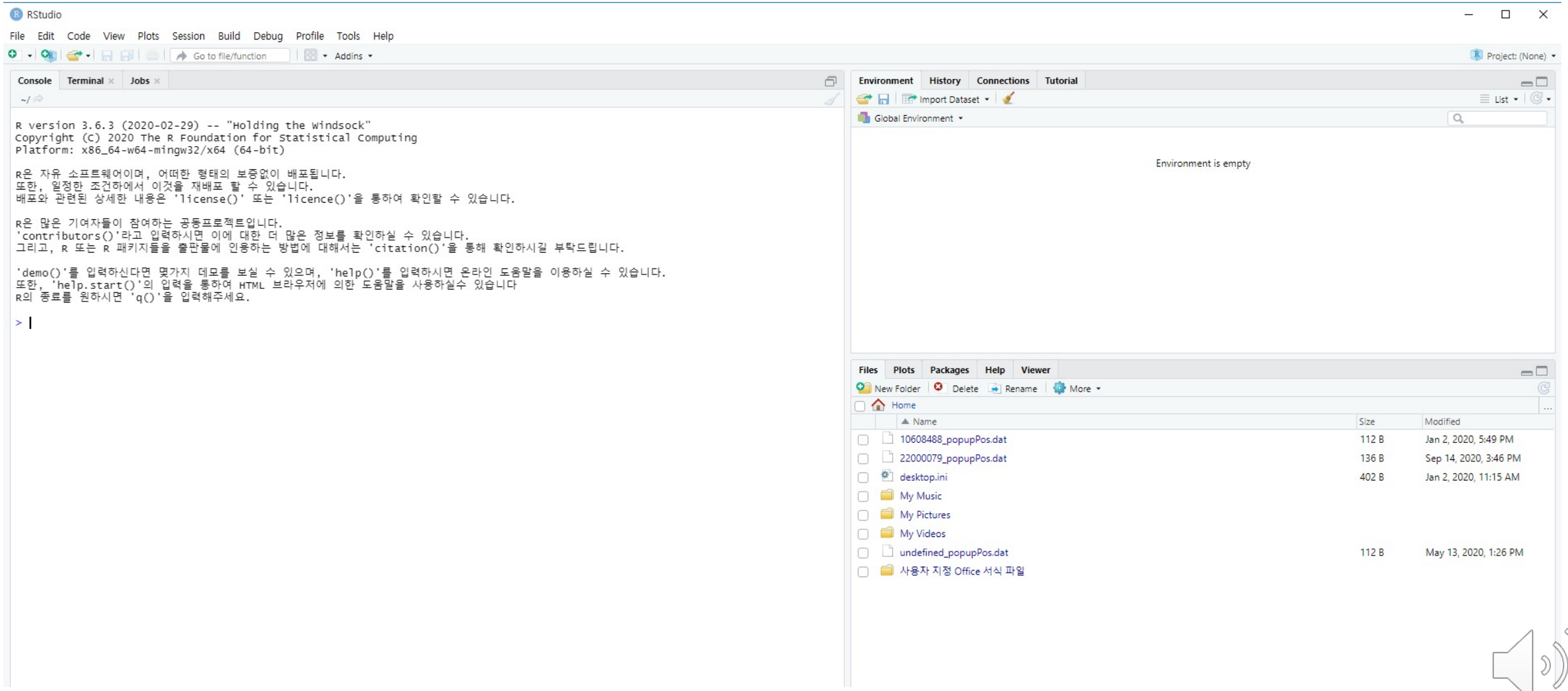
All Installers

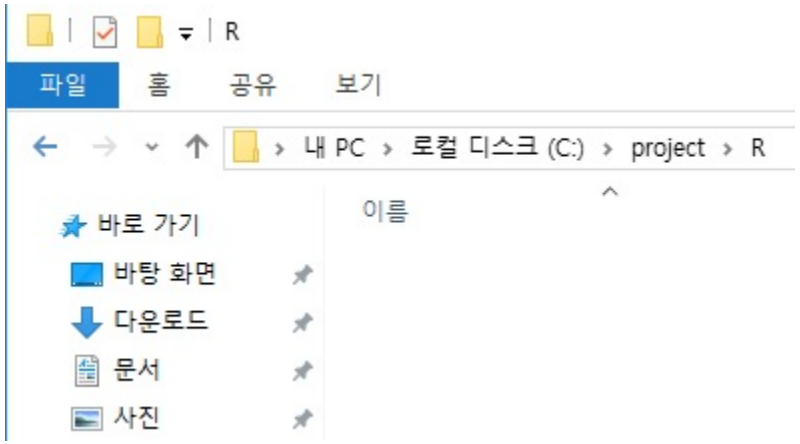
Linux users may need to import RStudio's public code-signing key prior to installation, depending on the operating system's security policy.

RStudio requires a 64-bit operating system. If you are on a 32 bit system, you can use an older version of RStudio.

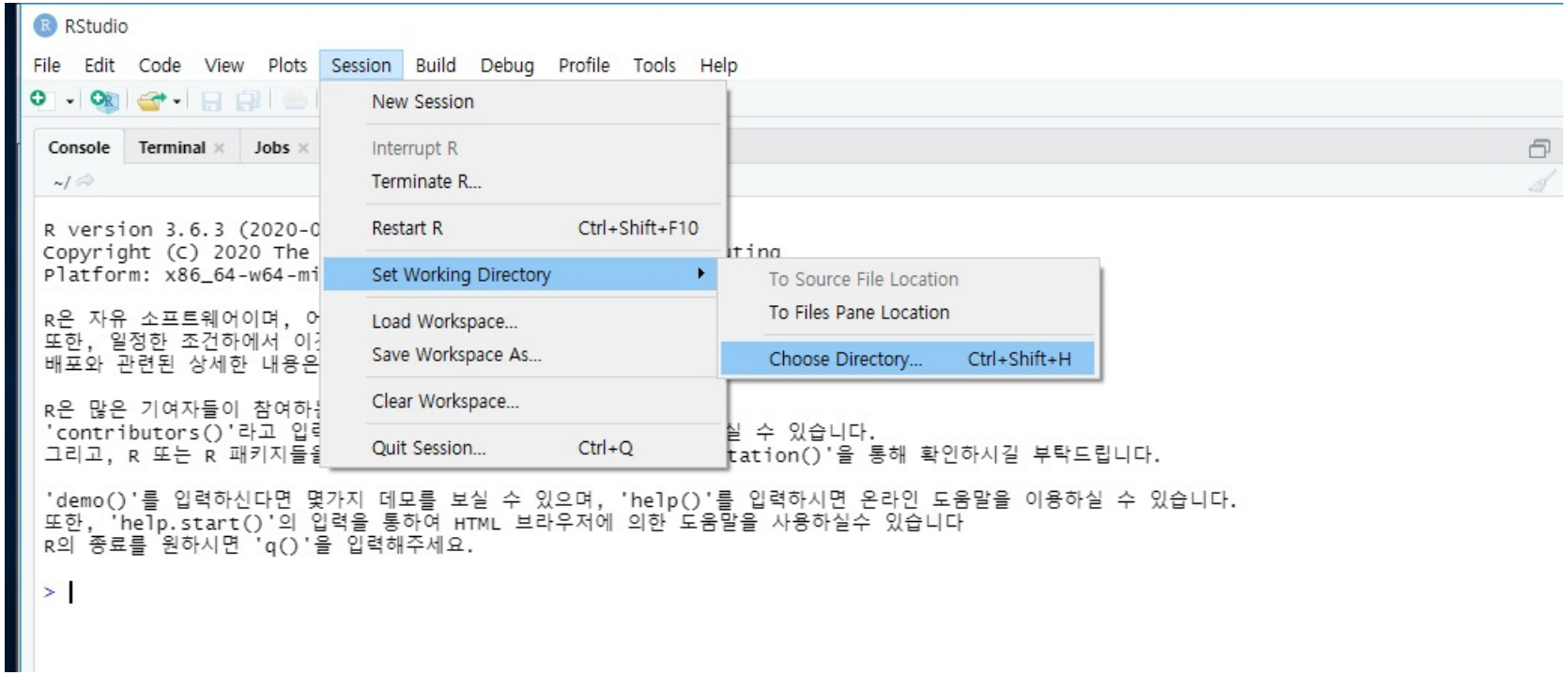
OS	Download	Size	SHA-256
Windows 10/8/7	RStudio-1.3.1073.exe	171.62 MB	2fea472a
macOS 10.13+	RStudio-1.3.1073.dmg	148.66 MB	0878b305
Ubuntu 16	rstudio-1.3.1073-amd64.deb	124.07 MB	6d71c5ff
Ubuntu 18/Debian 10	rstudio-1.3.1073-amd64.deb	126.78 MB	86be9352

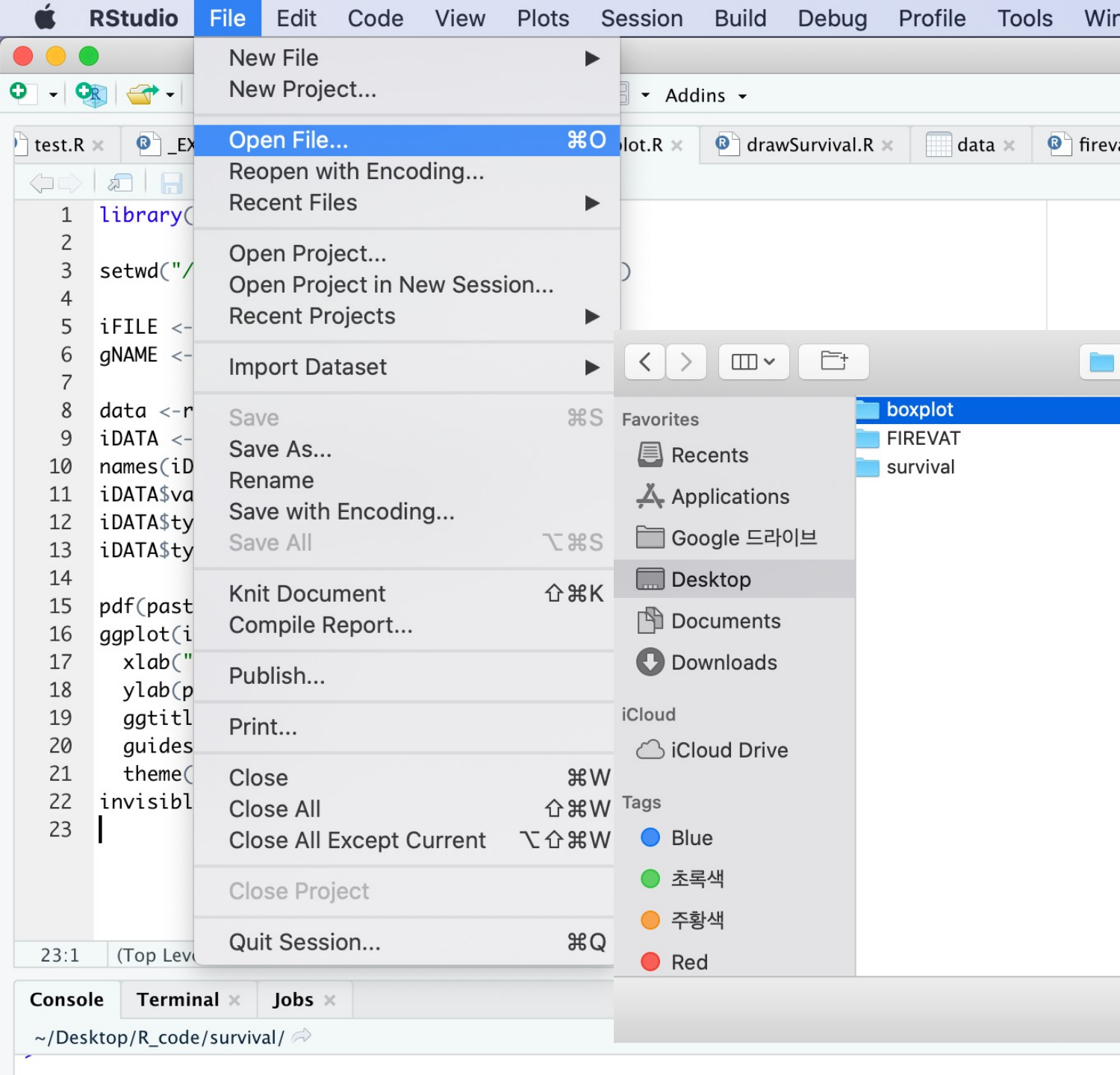
IDE (Integrated Development Environment)

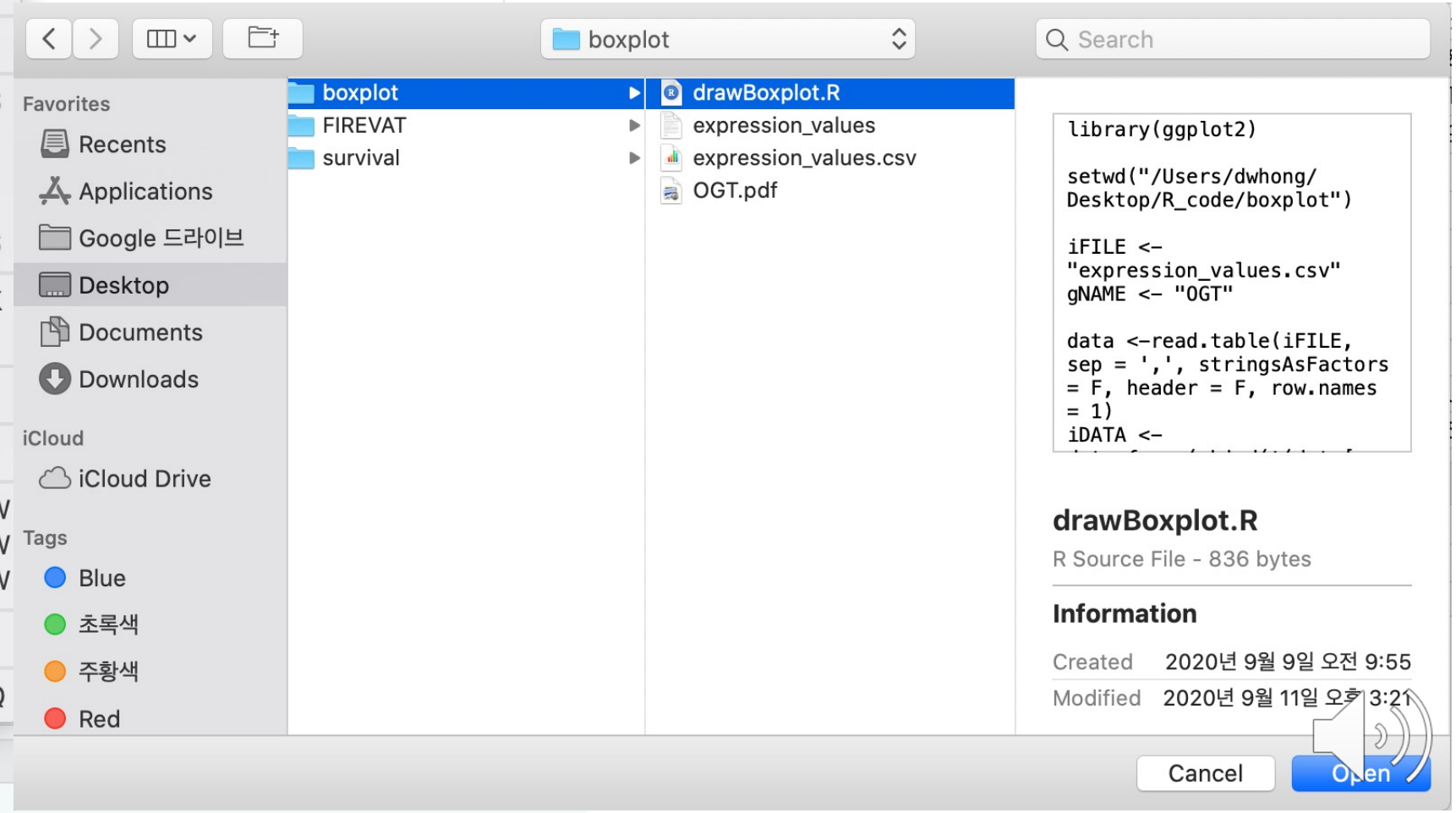




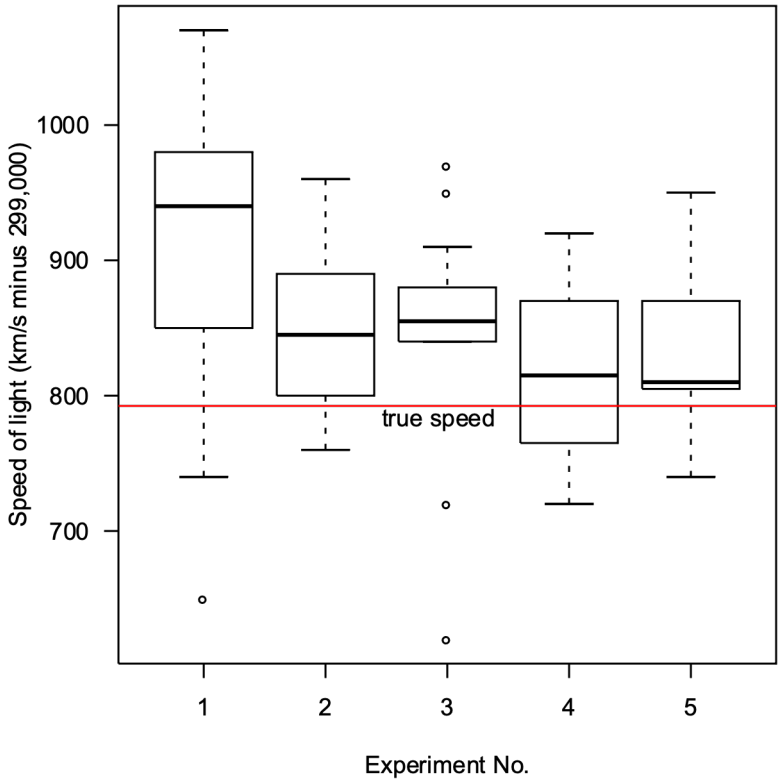
setwd("C://project//R")




RStudio
File
Edit
Code
View
Plots
Session
Build
Debug
Profile
Tools
New File
New Project...
Open File... ⌘O
Reopen with Encoding...
Recent Files
Open Project...
Open Project in New Session...
Recent Projects
Import Dataset
Save ⌘S
Save As...
Rename
Save with Encoding...
Save All ⌥⌘S
Knit Document ⇧⌘K
Compile Report...
Publish...
Print...
Close ⌘W
Close All ⇧⌘W
Close All Except Current ⌥⇧⌘W
Close Project
Quit Session... ⌘Q
Addins
test.R
drawSurvival.R
data
library(
setwd("/
iFILE <-
gNAME <-
data <-r
iDATA <-
names(iD
iDATA$va
iDATA$ty
iDATA$ty
pdf(past
ggplot(i
xlab("
ylab(p
ggtitl
guides
theme(
invisibl
23:1
(Top Lev
Console
Terminal
Jobs
~/Desktop/R_code/survival/


boxplot
Search
Favorites
Recents
Applications
Google 드라이브
Desktop
Documents
Downloads
iCloud
iCloud Drive
Tags
Blue
초록색
주황색
Red
boxplot
FIREVAT
survival
drawBoxplot.R
expression_values
expression_values.csv
OGT.pdf
library(ggplot2)
setwd("/Users/dwhong/
Desktop/R_code/boxplot")
iFILE <-
"expression_values.csv"
gNAME <- "OGT"
data <-read.table(iFILE,
sep = ',', stringsAsFactors
= F, header = F, row.names
= 1)
iDATA <-
drawBoxplot.R
R Source File - 836 bytes
Information
Created 2020년 9월 9일 오전 9:55
Modified 2020년 9월 11일 오후 3:21
Cancel
Open

BOXPLOT


Speed of light (km/s minus 299,000)
1000
900
800
700
true speed
1
2
3
4
5
Experiment No.

```
> install.packages("ggplot2")
trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.6/ggplot2_3.3.2.tgz'
Content type 'application/x-gzip' length 4068619 bytes (3.9 MB)
==================================================
downloaded 3.9 MB
```

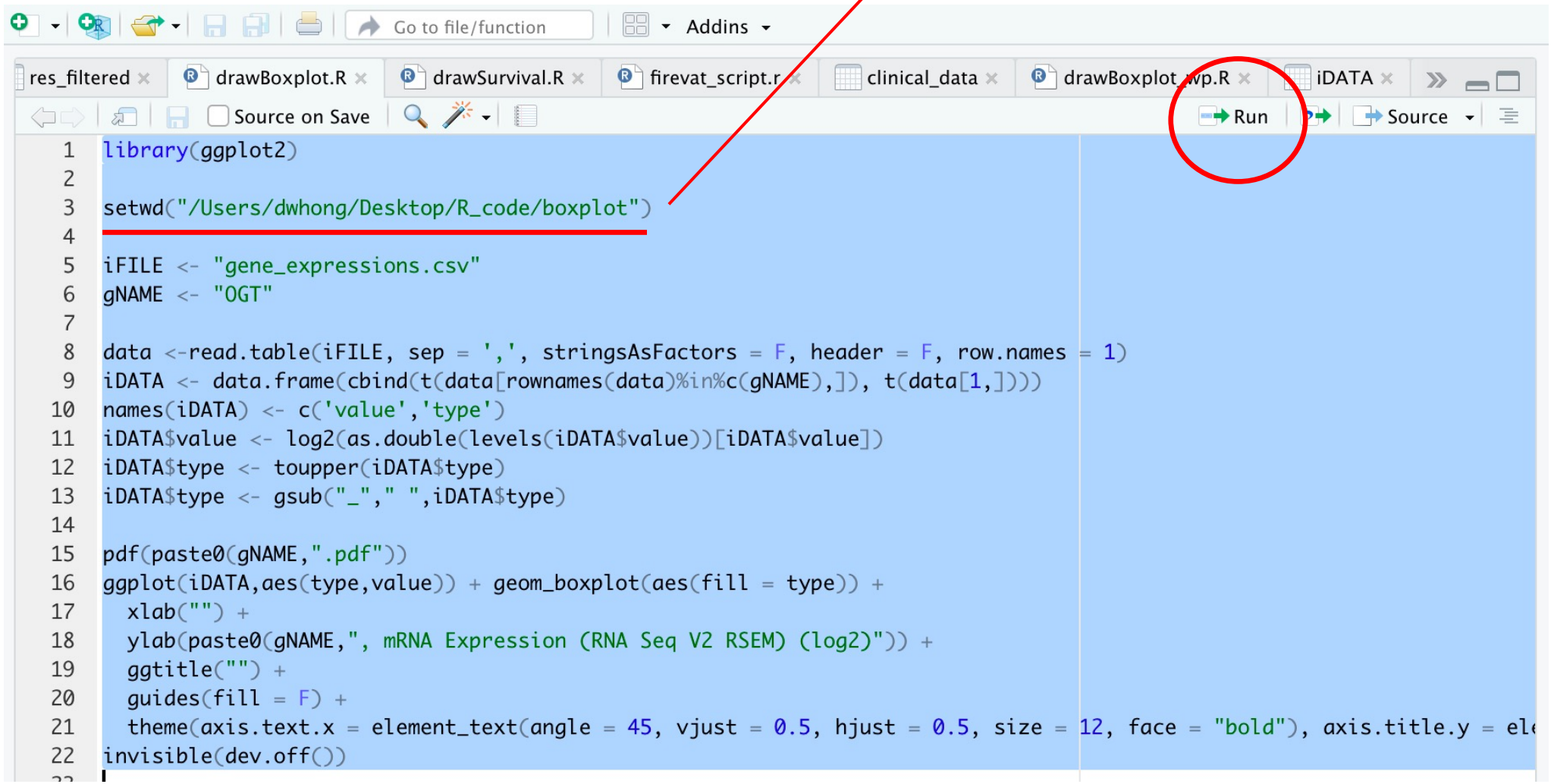
The downloaded binary packages are in
/var/folders/45/v4b_fjns00jg9ytjprdsxkd00000gn/T//Rtmp0BReEZ/downloaded_packages

```
library(ggplot2)

setwd("/Users/dwhong/Desktop/R_code/boxplot")

iFILE <- "gene_expressions.csv"
gNAME <- "OGT"

data <-read.table(iFILE, sep = ',', stringsAsFactors = F, header = F, row.names = 1)
iDATA <- data.frame(cbind(t(data[rownames(data)%in%c(gNAME),]), t(data[1,])))
names(iDATA) <- c('value','type')
iDATA$value <- log2(as.double(levels(iDATA$value))[iDATA$value])
iDATA$type <- toupper(iDATA$type)
iDATA$type <- gsub("_"," ",iDATA$type)

pdf(paste0(gNAME,".pdf"))
ggplot(iDATA,aes(type,value)) + geom_boxplot(aes(fill = type)) +
  xlab("") +
  ylab(paste0(gNAME," mRNA Expression (RNA Seq V2 RSEM) (log2)")) +
  ggtitle("") +
  guides(fill = F) +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust = 0.5, size = 12, face = "bold"), axis.title.y = ele
invisible(dev.off())
```

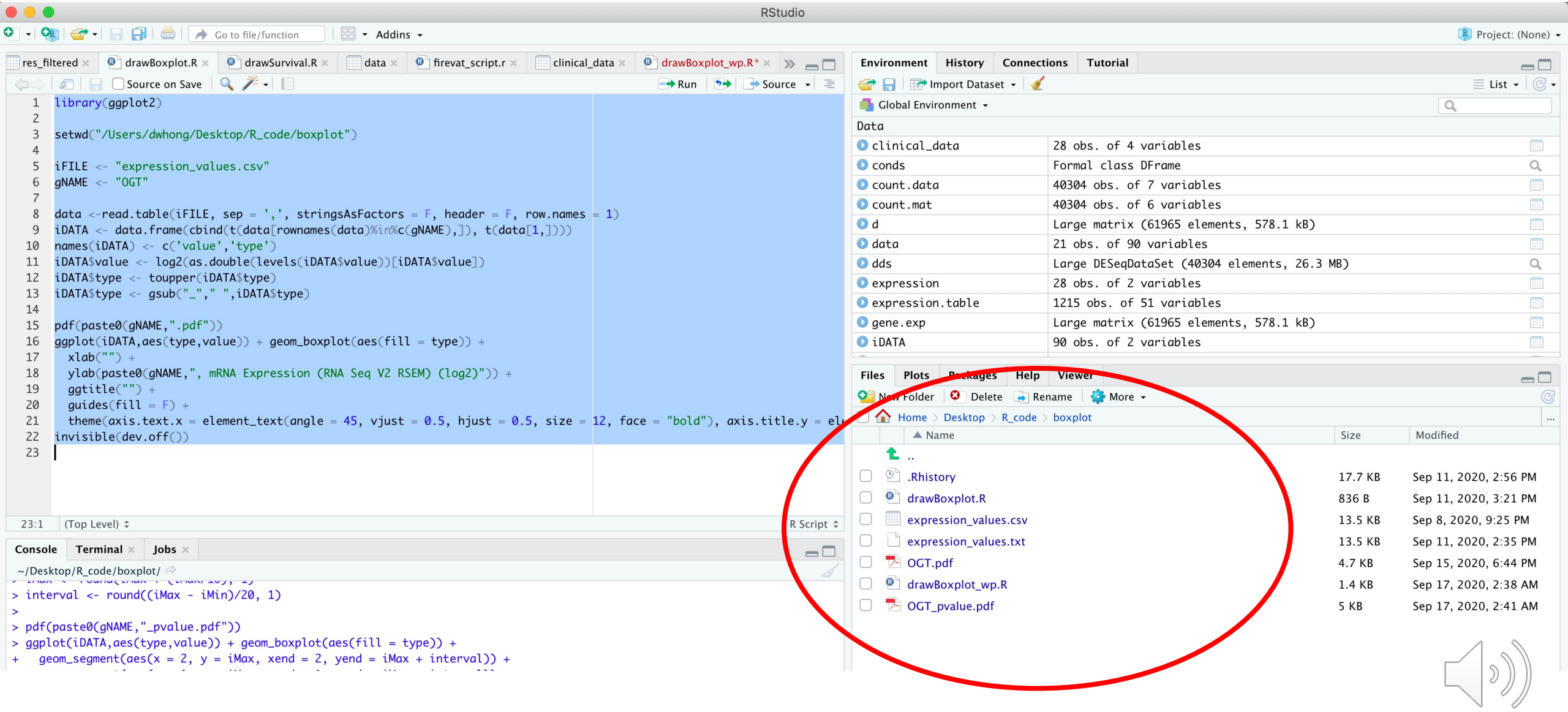
```
> data <-read.table(iFILE, sep = ',', stringsAsFactors = F, header = F, row.names = 1)
> iDATA <- data.frame(cbind(t(data[rownames(data)%in%c(gNAME),]), t(data[1,])))
> names(iDATA) <- c('value','type')
> iDATA$value <- log2(as.double(levels(iDATA$value))[iDATA$value])
> iDATA$type <- toupper(iDATA$type)
> iDATA$type <- gsub("_"," ",iDATA$type)
>
> pdf(paste0(gNAME,".pdf"))
> ggplot(iDATA,aes(type,value)) + geom_boxplot(aes(fill = type)) +
+   xlab("") +
+   ylab(paste0(gNAME," mRNA Expression (RNA Seq V2 RSEM) (log2)")) +
+   ggtitle("") +
+   guides(fill = F) +
+   theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust = 0.5, size = 12, face = "bold"), axis.title.y = elemen
t_text(size = 16, face = "bold"))
> invisible(dev.off())
>
```

Run selected R scripts

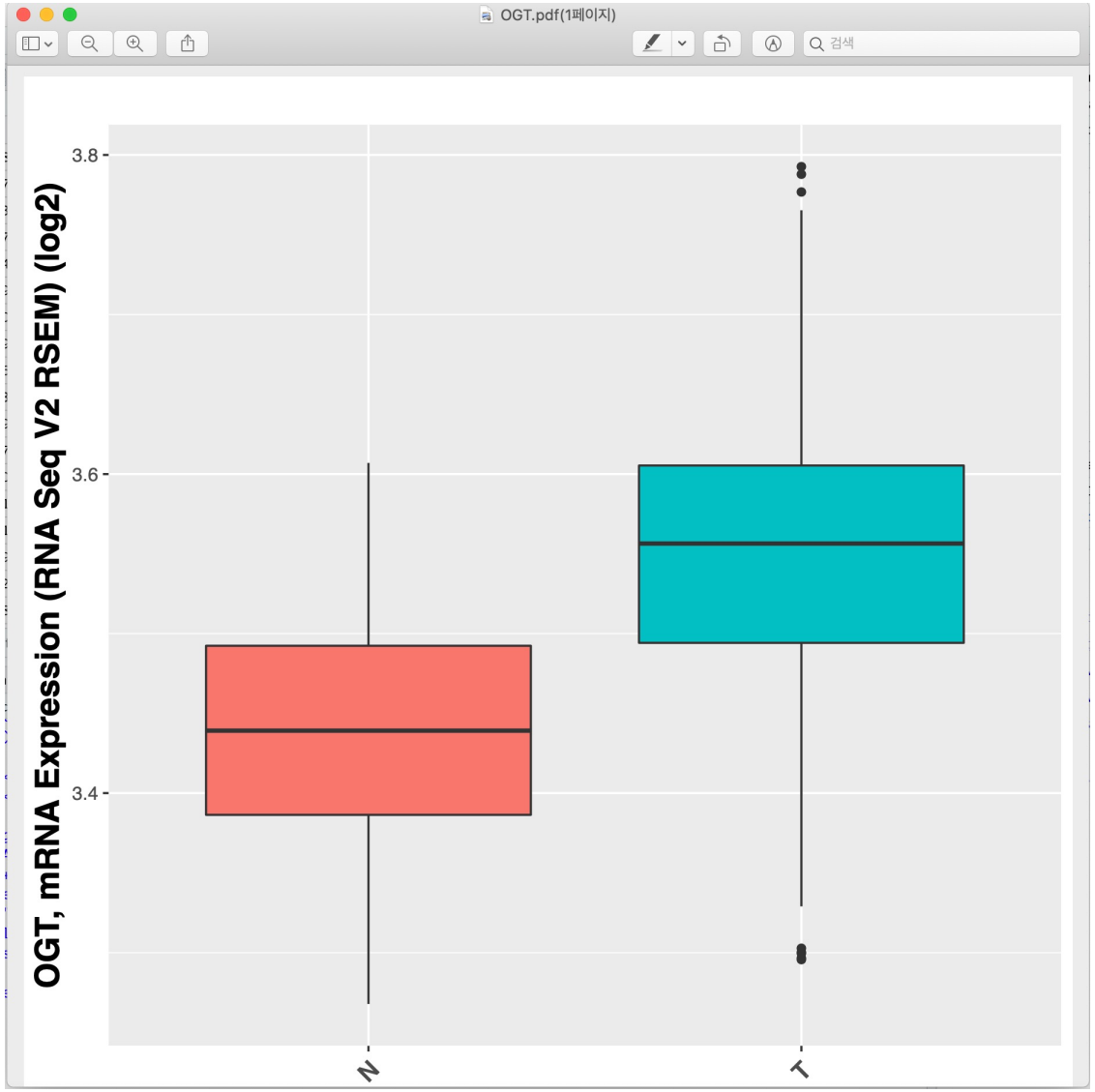
setwd("C://project//R")



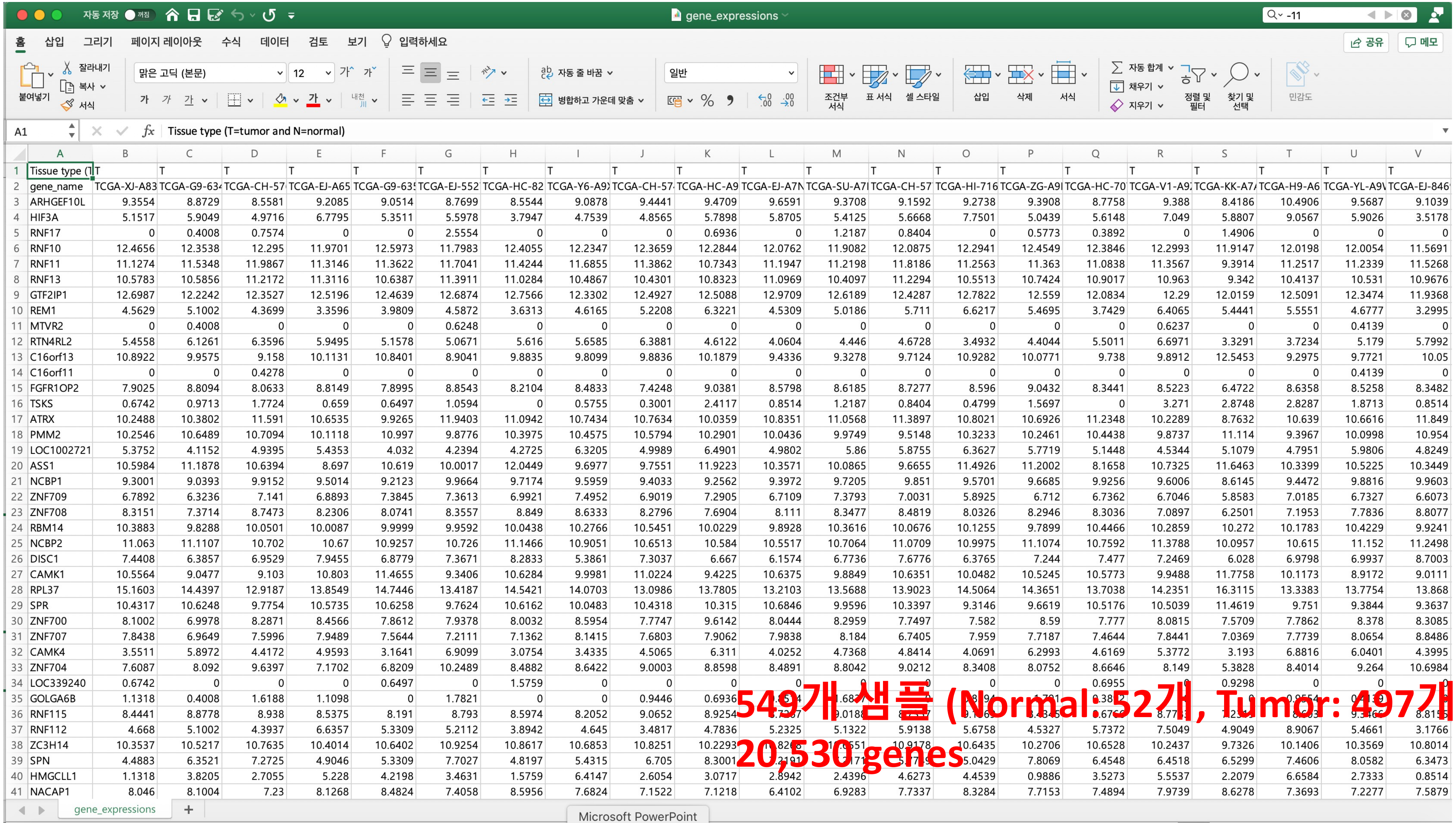
Browsing the output



Browsing the output



Load “expression_values.csv” in EXCEL



A1: Tissue type (T=tumor and N=normal)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Tissue type (T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
2	gene_name	TCGA-XJ-A83	TCGA-G9-634	TCGA-CH-57	TCGA-EJ-A65	TCGA-G9-635	TCGA-EJ-552	TCGA-HC-82	TCGA-Y6-A9	TCGA-CH-57	TCGA-HC-A9	TCGA-EJ-A7N	TCGA-SU-A7I	TCGA-CH-57	TCGA-HI-716	TCGA-ZG-A9I	TCGA-HC-70	TCGA-V1-A9	TCGA-KK-A7	TCGA-H9-A6	TCGA-YL-A9V	TCGA-EJ-846
3	ARHGEF10L	9.3554	8.8729	8.5581	9.2085	9.0514	8.7699	8.5544	9.0878	9.4441	9.4709	9.6591	9.3708	9.1592	9.2738	9.3908	8.7758	9.388	8.4186	10.4906	9.5687	9.1039
4	HIF3A	5.1517	5.9049	4.9716	6.7795	5.3511	5.5978	3.7947	4.7539	4.8565	5.7898	5.8705	5.4125	5.6668	7.7501	5.0439	5.6148	7.049	5.8807	9.0567	5.9026	3.5178
5	RNF17	0	0.4008	0.7574	0	0	2.5554	0	0	0	0.6936	0	1.2187	0.8404	0	0.5773	0.3892	0	1.4906	0	0	0
6	RNF10	12.4656	12.3538	12.295	11.9701	12.5973	11.7983	12.4055	12.2347	12.3659	12.2844	12.0762	11.9082	12.0875	12.2941	12.4549	12.3846	12.2993	11.9147	12.0198	12.0054	11.5691
7	RNF11	11.1274	11.5348	11.9867	11.3146	11.3622	11.7041	11.4244	11.6855	11.3862	10.7343	11.1947	11.2198	11.8186	11.2563	11.363	11.0838	11.3567	9.3914	11.2517	11.2339	11.5268
8	RNF13	10.5783	10.5856	11.2172	11.3116	10.6387	11.3911	11.0284	10.4867	10.4301	10.8323	11.0969	10.4097	11.2294	10.5513	10.7424	10.9017	10.963	9.342	10.4137	10.531	10.9676
9	GTF2IP1	12.6987	12.2242	12.3527	12.5196	12.4639	12.6874	12.7566	12.3302	12.4927	12.5088	12.9709	12.6189	12.4287	12.7822	12.559	12.0834	12.29	12.0159	12.5091	12.3474	11.9368
10	REM1	4.5629	5.1002	4.3699	3.3596	3.9809	4.5872	3.6313	4.6165	5.2208	6.3221	4.5309	5.0186	5.711	6.6217	5.4695	3.7429	6.4065	5.4441	5.5551	4.6777	3.2995
11	MTVR2	0	0.4008	0	0	0	0.6248	0	0	0	0	0	0	0	0	0	0	0.6237	0	0	0.4139	0
12	RTN4RL2	5.4558	6.1261	6.3596	5.9495	5.1578	5.0671	5.616	5.6585	6.3881	4.6122	4.0604	4.446	4.6728	3.4932	4.4044	5.5011	6.6971	3.3291	3.7234	5.179	5.7992
13	C16orf13	10.8922	9.9575	9.158	10.1131	10.8401	8.9041	9.8835	9.8099	9.8836	10.1879	9.4336	9.3278	9.7124	10.9282	10.0771	9.738	9.8912	12.5453	9.2975	9.7721	10.05
14	C16orf11	0	0	0.4278	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4139	0
15	FGFR1OP2	7.9025	8.8094	8.0633	8.8149	7.8995	8.8543	8.2104	8.4833	7.4248	9.0381	8.5798	8.6185	8.7277	8.596	9.0432	8.3441	8.5223	6.4722	8.6358	8.5258	8.3482
16	TSKS	0.6742	0.9713	1.7724	0.659	0.6497	1.0594	0	0.5755	0.3001	2.4117	0.8514	1.2187	0.8404	0.4799	1.5697	0	3.271	2.8748	2.8287	1.8713	0.8514
17	ATRX	10.2488	10.3802	11.591	10.6535	9.9265	11.9403	11.0942	10.7434	10.7634	10.0359	10.8351	11.0568	11.3897	10.8021	10.6926	11.2348	10.2289	8.7632	10.639	10.6616	11.849
18	PMM2	10.2546	10.6489	10.7094	10.1118	10.997	9.8776	10.3975	10.4575	10.5794	10.2901	10.0436	9.9749	9.5148	10.3233	10.2461	10.4438	9.8737	11.114	9.3967	10.0998	10.954
19	LOC1002721	5.3752	4.1152	4.9395	5.4353	4.032	4.2394	4.2725	6.3205	4.9989	6.4901	4.9802	5.86	5.8755	6.3627	5.7719	5.1448	4.5344	5.1079	4.7951	5.9806	4.8249
20	ASS1	10.5984	11.1878	10.6394	8.697	10.619	10.0017	12.0449	9.6977	9.7551	11.9223	10.3571	10.0865	9.6655	11.4926	11.2002	8.1658	10.7325	11.6463	10.3399	10.5225	10.3449
21	NCBP1	9.3001	9.0393	9.9152	9.5014	9.2123	9.9664	9.7174	9.5959	9.4033	9.2562	9.3972	9.7205	9.851	9.5701	9.6685	9.9256	9.6006	8.6145	9.4472	9.8816	9.9603
22	ZNF709	6.7892	6.3236	7.141	6.8893	7.3845	7.3613	6.9921	7.4952	6.9019	7.2905	6.7109	7.3793	7.0031	5.8925	6.712	6.7362	6.7046	5.8583	7.0185	6.7327	6.6073
23	ZNF708	8.3151	7.3714	8.7473	8.2306	8.0741	8.3557	8.849	8.6333	8.2796	7.6904	8.111	8.3477	8.4819	8.0326	8.2946	8.3036	7.0897	6.2501	7.1953	7.7836	8.8077
24	RBM14	10.3883	9.8288	10.0501	10.0087	9.9999	9.9592	10.0438	10.2766	10.5451	10.0229	9.8928	10.3616	10.0676	10.1255	9.7899	10.4466	10.2859	10.272	10.1783	10.4229	9.9241
25	NCBP2	11.063	11.1107	10.702	10.67	10.9257	10.726	11.1466	10.9051	10.6513	10.584	10.5517	10.7064	11.0709	10.9975	11.1074	10.7592	11.3788	10.0957	10.615	11.152	11.2498
26	DISC1	7.4408	6.3857	6.9529	7.9455	6.8779	7.3671	8.2833	5.3861	7.3037	6.667	6.1574	6.7736	7.6776	6.3765	7.244	7.477	7.2469	6.028	6.9798	6.9937	8.7003
27	CAMK1	10.5564	9.0477	9.103	10.803	11.4655	9.3406	10.6284	9.9981	11.0224	9.4225	10.6375	9.8849	10.6351	10.0482	10.5245	10.5773	9.9488	11.7758	10.1173	8.9172	9.0111
28	RPL37	15.1603	14.4397	12.9187	13.8549	14.7446	13.4187	14.5421	14.0703	13.0986	13.7805	13.2103	13.5688	13.9023	14.5064	14.3651	13.7038	14.2351	16.3115	13.3383	13.7754	13.868
29	SPR	10.4317	10.6248	9.7754	10.5735	10.6258	9.7624	10.6162	10.0483	10.4318	10.315	10.6846	9.9596	10.3397	9.3146	9.6619	10.5176	10.5039	11.4619	9.751	9.3844	9.3637
30	ZNF700	8.1002	6.9978	8.2871	8.4566	7.8612	7.9378	8.0032	8.5954	7.7747	9.6142	8.0444	8.2959	7.7497	7.582	8.59	7.777	8.0815	7.5709	7.7862	8.378	8.3085
31	ZNF707	7.8438	6.9649	7.5996	7.9489	7.5644	7.2111	7.1362	8.1415	7.6803	7.9062	7.9838	8.184	6.7405	7.959	7.7187	7.4644	7.8441	7.0369	7.7739	8.0654	8.8486
32	CAMK4	3.5511	5.8972	4.4172	4.9593	3.1641	6.9099	3.0754	3.4335	4.5065	6.311	4.0252	4.7368	4.8414	4.0691	6.2993	4.6169	5.3772	3.193	6.8816	6.0401	4.3995
33	ZNF704	7.6087	8.092	9.6397	7.1702	6.8209	10.2489	8.4882	8.6422	9.0003	8.8598	8.4891	8.8042	9.0212	8.3408	8.0752	8.6646	8.149	5.3828	8.4014	9.264	10.6984
34	LOC339240	0.6742	0	0	0	0.6497	0	1.5759	0	0	0	0	0	0	0	0	0.6955	0	0.9298	0	0	0
35	GOLGA6B	1.1318	0.4008	1.6188	1.1098	0	1.7821	0	0	0.9446	0.6936	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	[illegible]	[illegible]	[illegible]
36	RNF115	8.4441	8.8778	8.938	8.5375	8.191	8.793	8.5974	8.2052	9.0652	8.9254	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
37	RNF112	4.668	5.1002	4.3937	6.6357	5.3309	5.2112	3.8942	4.645	3.4817	4.7836	5.2325	5.1322	5.9138	5.6758	4.5327	5.7372	7.5049	4.9049	8.9067	5.4661	3.1766
38	ZC3H14	10.3537	10.5217	10.7635	10.4014	10.6402	10.9254	10.8617	10.6853	10.8251	10.2293	[illegible]	[illegible]	10.9178	10.6435	10.2706	10.6528	10.2437	9.7326	10.1406	10.3569	10.8014
39	SPN	4.4883	6.3521	7.2725	4.9046	5.3309	7.7027	4.8197	5.4315	6.705	8.3001	[illegible]	[illegible]	[illegible]	5.0429	7.8069	6.4548	6.4518	6.5299	7.4606	8.0582	6.3473
40	HMGCLL1	1.1318	3.8205	2.7055	5.228	4.2198	3.4631	1.5759	6.4147	2.6054	3.0717	2.8942	2.4396	4.6273	4.4539	0.9886	3.5273	5.5537	2.2079	6.6584	2.7333	0.8514
41	NACAP1	8.046	8.1004	7.23	8.1268	8.4824	7.4058	8.5956	7.6824	7.1522	7.1218	6.4102	6.9283	7.7337	8.3284	7.7153	7.4894	7.9739	8.6278	7.3693	7.2277	7.5879

gene_expressions

549개 샘플 (Normal: 52개, Tumor: 497개)

20,530 genes

> View (data)

res_filtered | drawBoxplot.R | data | drawSurvival.R | firevat_script.r | clinical_data | drawBoxplot_wp.R

Filter | Cols: « < 1 - 50 > »

	V2	V3	V4	V5	V6	V7
Tissue type (T=tumor and N=normal)	T	T	T	T	T	T
gene_name	TCGA-XJ-A83F-01	TCGA-G9-6348-01	TCGA-CH-5766-01	TCGA-EJ-A65G-01	TCGA-G9-6354-01	TCGA-EJ-5527-(
ARHGEF10L	9.3554	8.8729	8.5581	9.2085	9.0514	8.7699
HIF3A	5.1517	5.9049	4.9716	6.7795	5.3511	5.5978
RNF17	0.0	0.4008	0.7574	0.0	0.0	2.5554
RNF10	12.4656	12.3538	12.295	11.9701	12.5973	11.7983
RNF11	11.1274	11.5348	11.9867	11.3146	11.3622	11.7041
RNF13	10.5783	10.5856	11.2172	11.3116	10.6387	11.3911
GTF2IP1	12.6987	12.2242	12.3527	12.5196	12.4639	12.6874
REM1	4.5629	5.1002	4.3699	3.3596	3.9809	4.5872
MTVR2	0.0	0.4008	0.0	0.0	0.0	0.6248
RTN4RL2	5.4558	6.1261	6.3596	5.9495	5.1578	5.0671
C16orf13	10.8922	9.9575	9.158	10.1131	10.8401	8.9041
C16orf11	0.0	0.0	0.4278	0.0	0.0	0.0
FGFR1OP2	7.9025	8.8094	8.0633	8.8149	7.8995	8.8543
TSKS	0.6742	0.9713	1.7724	0.659	0.6497	1.0594
ATRX	10.2488	10.3802	11.591	10.6535	9.9265	11.9403
PMM2	10.2546	10.6489	10.7094	10.1118	10.997	9.8776

Showing 1 to 19 of 20,532 entries, 549 total columns

Environment | History | Connections | Tutorial

Import Dataset

Global Environment

Data	
data	20532 obs. of 549 variables
iDATA	549 obs. of 2 variables
Values	
gNAME	"OGT"
iFILE	"gene_expressions.csv"

Files | Plots | Packages | Help | Viewer

New Folder | Delete | Rename | More

Home > Desktop > R_code > boxplot

Name

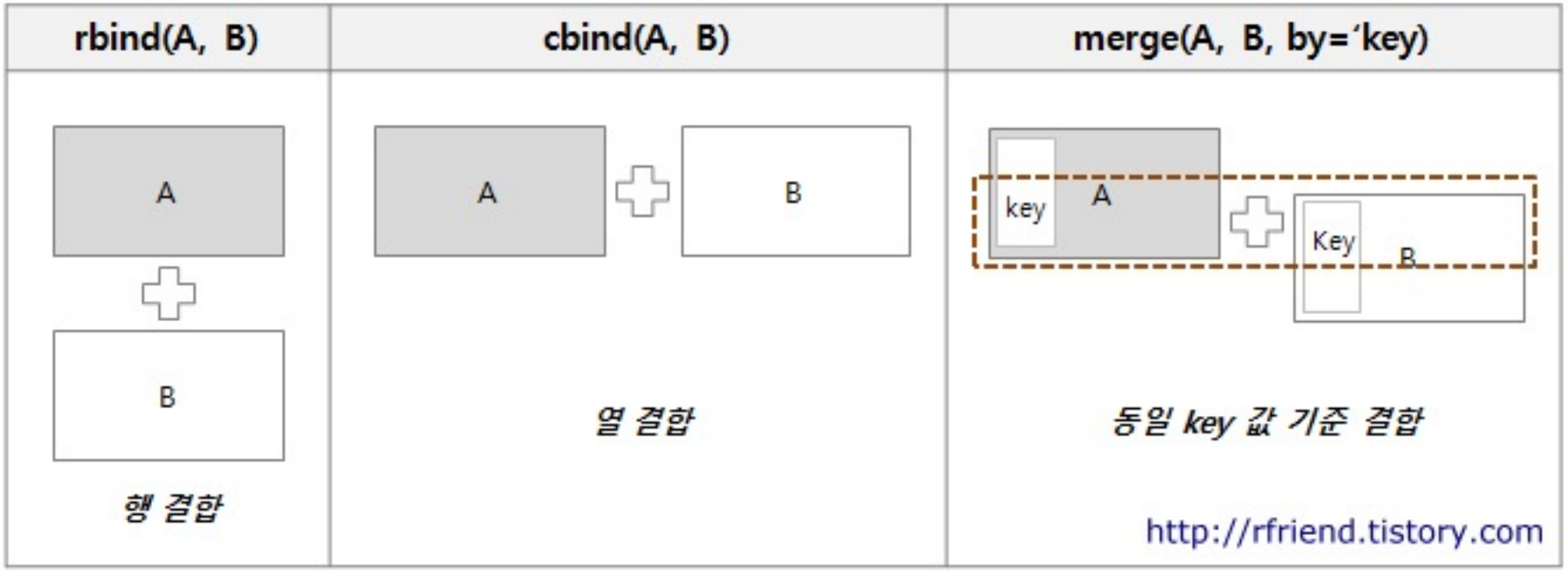
..

.Rhistory

drawBoxplot.R

drawBoxplot_wp.R

Variable (변수) <- value (값)

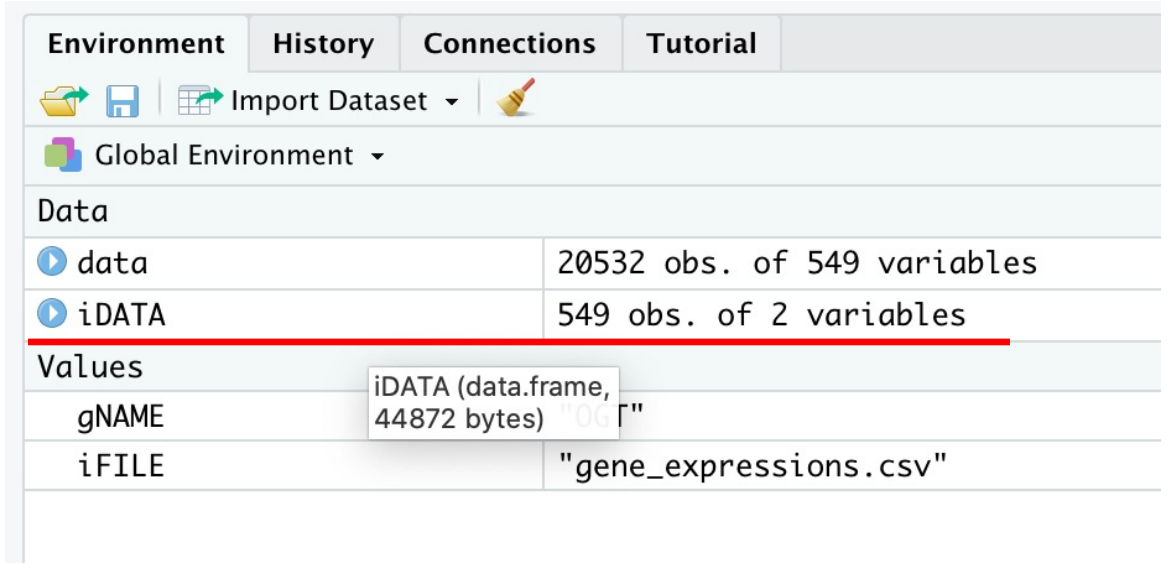


> View (iDATA)

Filter

	value	type
V2	3.594584	T
V3	3.364670	T
V4	3.582339	T
V5	3.619378	T
V6	3.530046	T
V7	3.504099	T
V8	3.548905	T
V9	3.592696	T
V10	3.518661	T
V11	3.701837	T
V12	3.567898	T
V13	3.664779	T
V14	3.500700	T
V15	3.556417	T
V16	3.574114	T
V17	3.498991	T
V18	3.443620	T
V19	3.443580	T

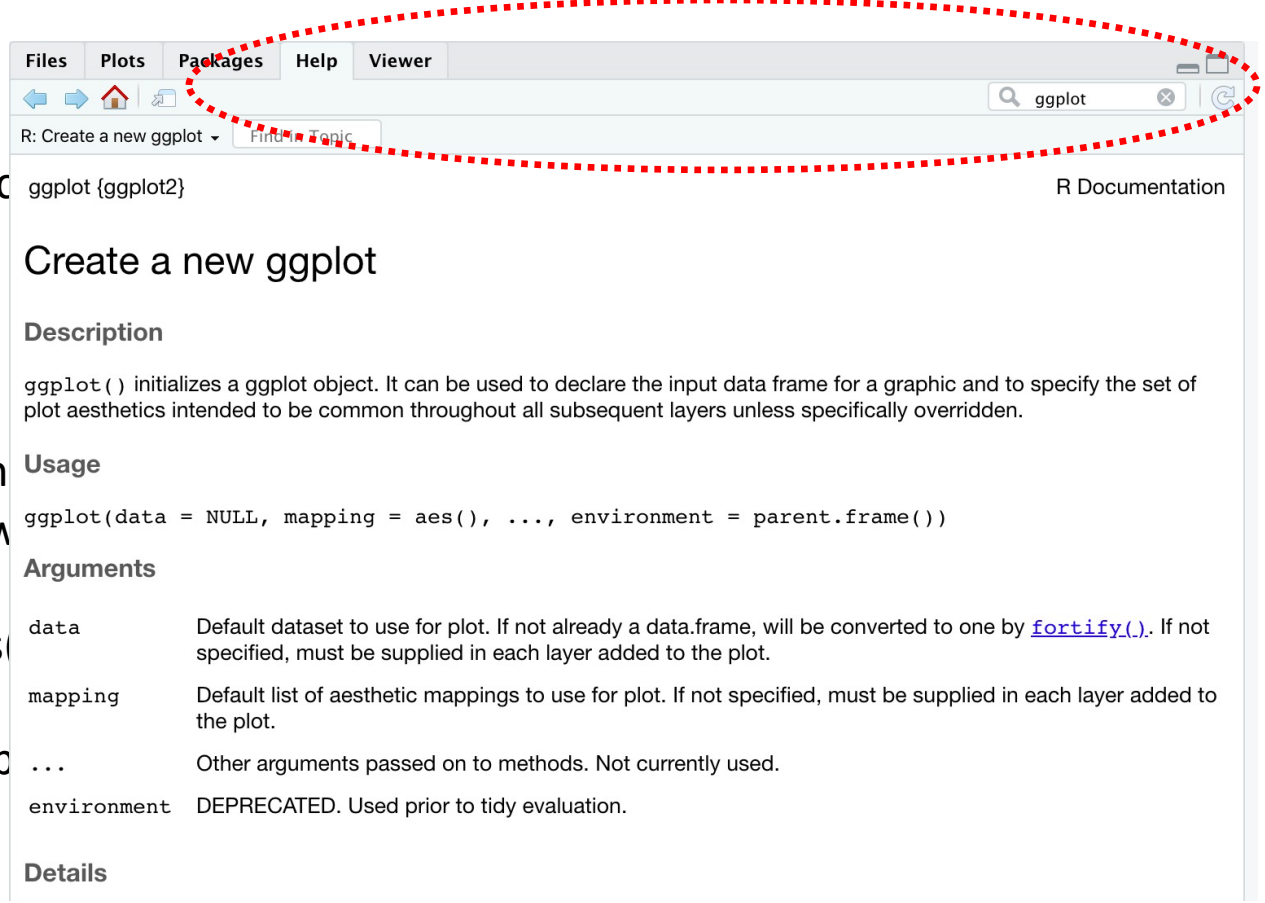
Showing 1 to 19 of 549 entries, 2 total columns



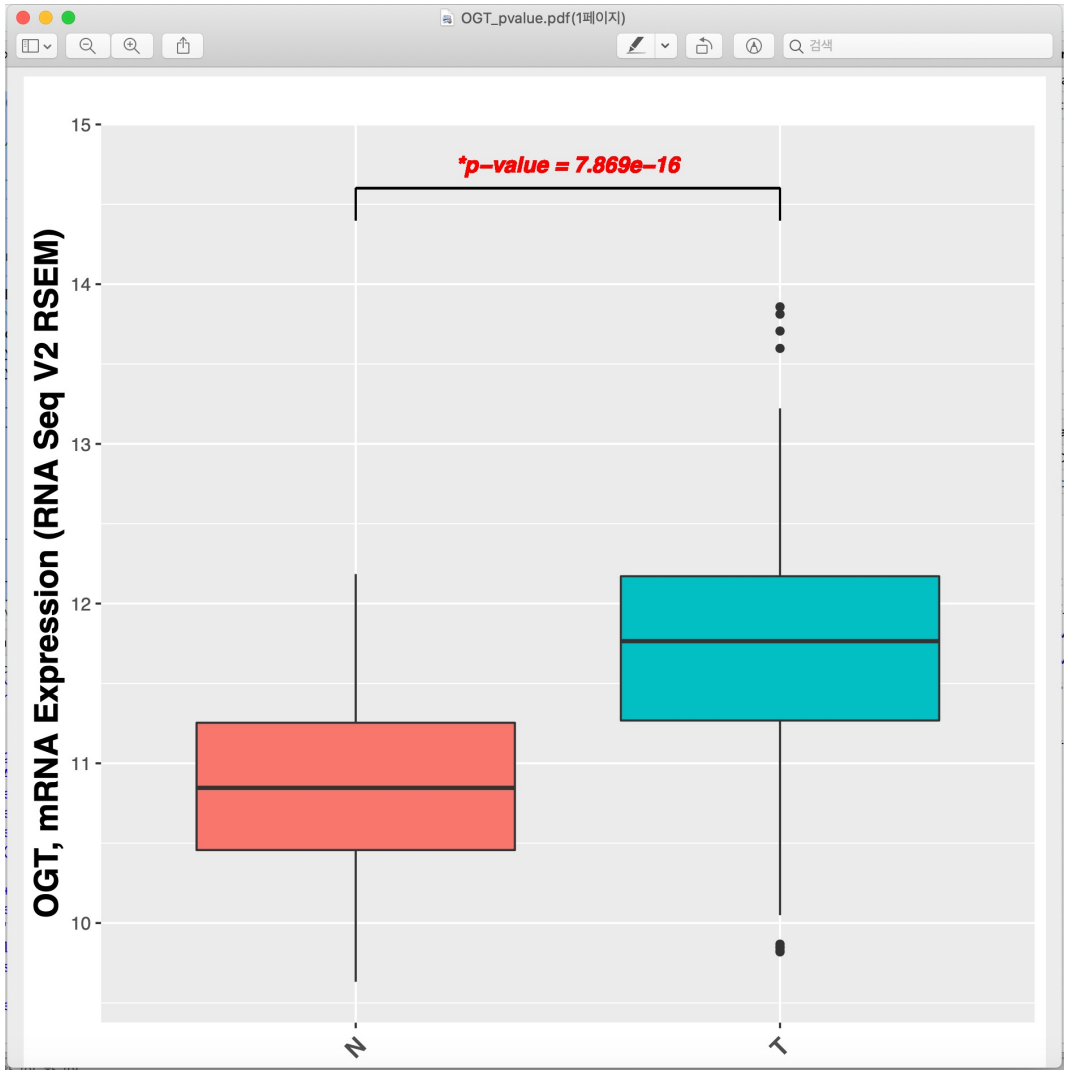
drawBoxplot.R

```
library(ggplot2)

setwd("/Users/dwhong/Desktop/R_code/boxplot")

iFILE <- "gene_expressions.csv"
gNAME <- "OGT"

data <-read.table(iFILE, sep = ',', stringsAsFactors = F, header = F, row.names = 1)
iDATA <- data.frame(cbind(t(data[rownames(data)%in%c(gNAME),]), t(data[1,])))
names(iDATA) <- c('value','type')
iDATA$value <- log2(as.double(levels(iDATA$value))[iDATA$value])
iDATA$type <- toupper(iDATA$type)
iDATA$type <- gsub("_"," ",iDATA$type)

pdf(paste0(gNAME,".pdf"))
ggplot(iDATA,aes(type,value)) + geom_boxplot(aes(fill = type)) +
  xlab("") +
  ylab(paste0(gNAME,", mRNA Expression (RNA Seq V2 RSEM) (log2)")) +
  ggtitle("") +
  guides(fill = F) +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust = 0.5, size = 12, face = "bold"), axis.title.y = element_text(size =
16, face = "bold"))
invisible(dev.off())
```

```
library(ggplot2)

setwd("/Users/dwhong/Desktop/R_c

iFILE <- ”gene_expressions.csv"
gNAME <- "OGT"

data <-read.table(iFILE, sep = ',', strin
iDATA <- data.frame(cbind(t(data[row
names(iDATA) <- c('value','type')
iDATA$value <- log2(as.double(levels(
iDATA$type <- toupper(iDATA$type)
iDATA$type <- gsub("_"," ",iDATA$typ

pdf(paste0(gNAME,".pdf"))
ggplot(iDATA,aes(type,value)) + geom_boxplot(aes(fill = type)) +
  xlab("") +
  ylab(paste0(gNAME,", mRNA Expression (RNA Seq V2 RSEM) (log2)")) +
  ggtitle("") +
  guides(fill = F) +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust = 0.5, size = 12, face = "bold"), axis.title.y = element_text(size =
16, face = "bold"))
invisible(dev.off())
```



Parametric test vs Non-parametric test



Parametric and Non-parametric tests for comparing two or more groups

Parametric test	Non-Parametric equivalent
Paired t-test	Wilcoxon Rank sum Test
Unpaired t-test	Mann-Whitney U test
Pearson correlation	Spearman correlation
One way Analysis of variance	Kruskal Wallis Test

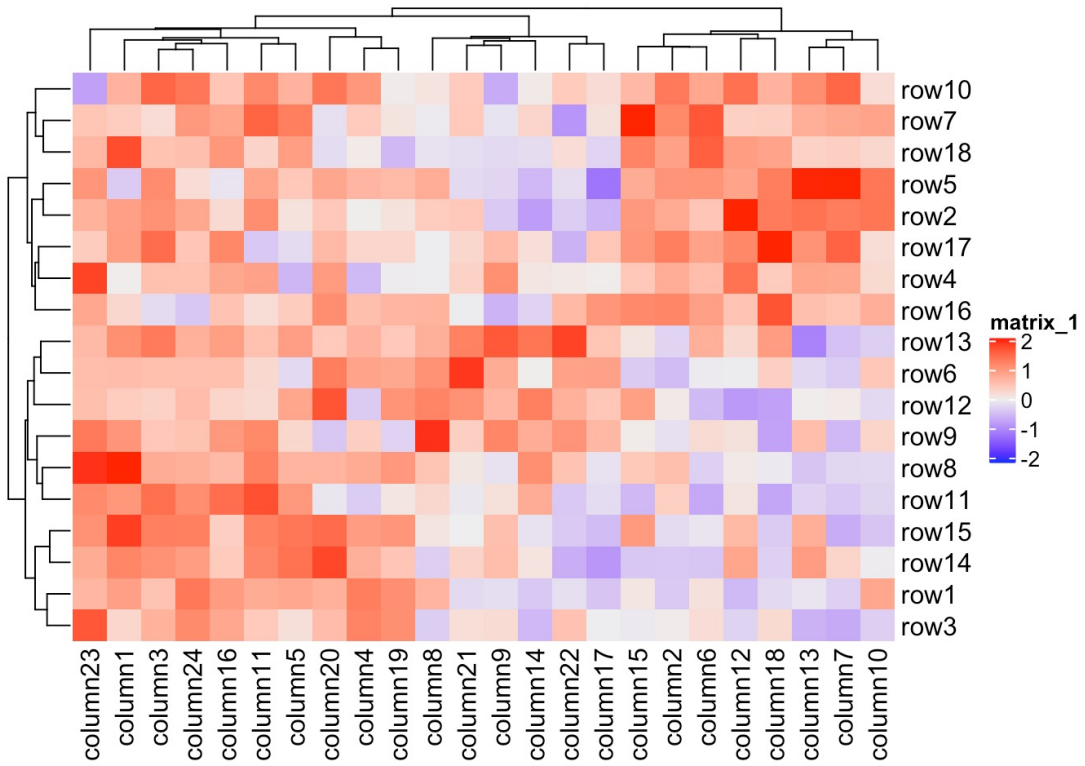
source: https://www.healthknowledge.org.uk

drawBoxplot_wp.R

```
library(ggplot2)

setwd("/Users/dwhong/Desktop/R_code/boxplot")

iFILE <- "expression_values.csv"
gNAME <- "OGT"

data <-read.table(iFILE, sep = ',', stringsAsFactors = F, header = F, row.names = 1)
iDATA <- data.frame(cbind(t(data[rownames(data)%in%c(gNAME),]), t(data[1,])))
names(iDATA) <- c('value','type')
iDATA$value <- as.double(levels(iDATA$value))[iDATA$value]
iDATA$type <- toupper(iDATA$type)
iDATA$type <- gsub("_"," ",iDATA$type)

nor <- iDATA[iDATA$type == "N",1]
tur <- iDATA[iDATA$type != "N",1]
pVal <- format(t.test(nor,tur)$p.value, digits = 4)

iMin <- min(nor,tur)
iMax <- max(nor,tur)
iMax <- round(iMax + (iMax/10), 1)
interval <- round((iMax - iMin)/20, 1)
```

```
pdf(paste0(gNAME,"_pvalue.pdf"))
ggplot(iDATA,aes(type,value)) + geom_boxplot(aes(fill = type)) +
  geom_segment(aes(x = 2, y = iMax-(interval*0.2), xend = 2, yend = iMax)) +
  geom_segment(aes(x = 1, y = iMax-(interval*0.2), xend = 1, yend = iMax)) +
  geom_segment(aes(x = 1, y = iMax, xend = 2, yend = iMax)) +
  geom_text(aes(x = 1.5, y = iMax + (interval*0.15), label = paste0("*p-value = ",pVal)), fontface = "bold.italic",
size = 4, color = "red") +
  xlab("") +
  ylab(paste0(gNAME,", mRNA Expression (RNA Seq V2 RSEM)")) +
  ggtitle("") +
  guides(fill = F) +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust = 0.5, size = 12, face = "bold"), axis.title.y =
element_text(size = 16, face = "bold"))
invisible(dev.off())
```

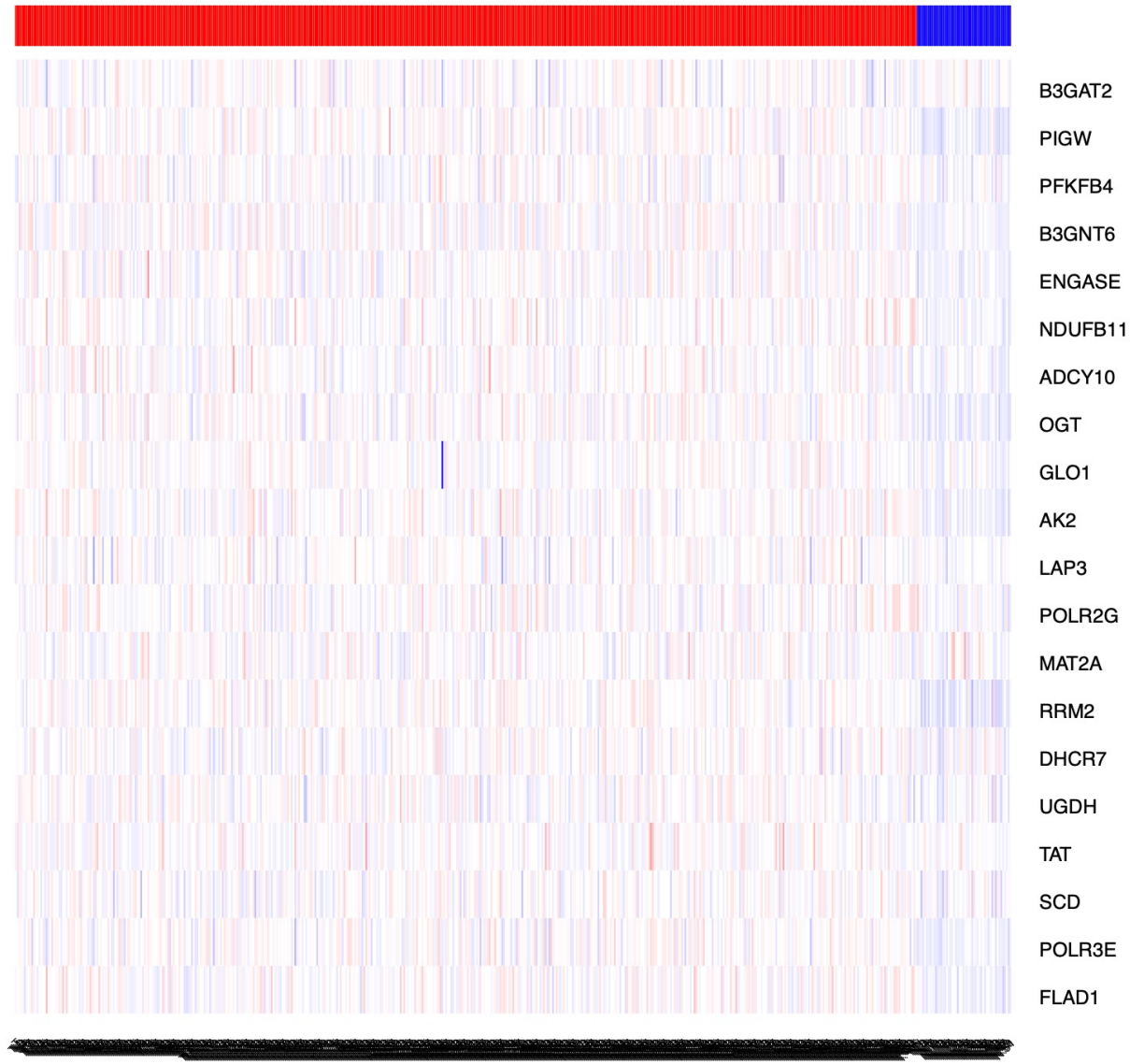
HEATMAP


row10
row7
row18
row5
row2
row17
row4
row16
row13
row6
row12
row9
row8
row11
row15
row14
row1
row3
column23
column1
column3
column24
column16
column11
column5
column20
column4
column19
column8
column21
column9
column14
column22
column17
column15
column2
column6
column12
column18
column13
column7
column10
matrix_1
2
1
0
-1
-2

20 metabolic genes

heatmap.R

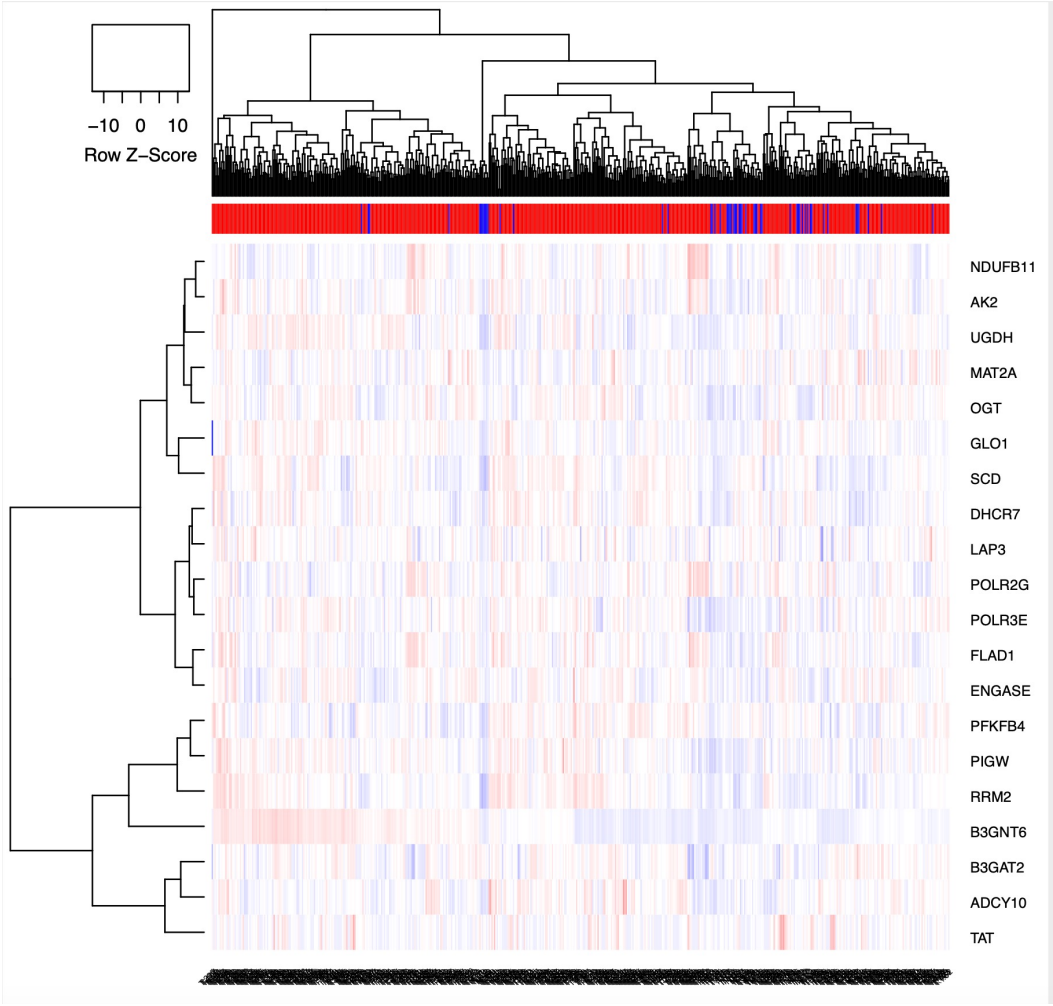
```
library(gplots)

setwd("/Users/dwhong/Desktop/R_code/heatmap")

a <- read.table('gene_expressions.csv',sep=',',stringsAsFactors = F,header=T, row.names=1)

sample_color <- c(rep("#FF0000",497),rep("#0000FF",52))
a <- a[rownames(a)%in%c('TAT', 'B3GNT6', 'OGT', 'MAT2A', 'UGDH', 'SCD', 'LAP3', 'PIGW', 'PFKFB4', 'NDUFB11',
'DHCR7', 'POLR2G', 'POLR3E', 'B3GAT2', 'RRM2', 'ADCY10', 'AK2', 'GLO1', 'FLAD1', 'ENGASE'),]

pval <- c(0.)
pdf('heatmap.pdf', width = 7,height = 7)
heatmap.2(data.matrix(a), col=bluered(100), scale = 'row', ColSideColors=sample_color, Colv=NA, Rowv =
NA, dendrogram = "none", cexRow = 0.90,
        srtCol=45,  adjCol = c(1,1), srtRow=0, adjRow=c(0, 1),
        key=TRUE, keysize = 1, key.title = NA,symkey=FALSE, density.info="none",
trace="none",,na.color="black")
dev.off()
```

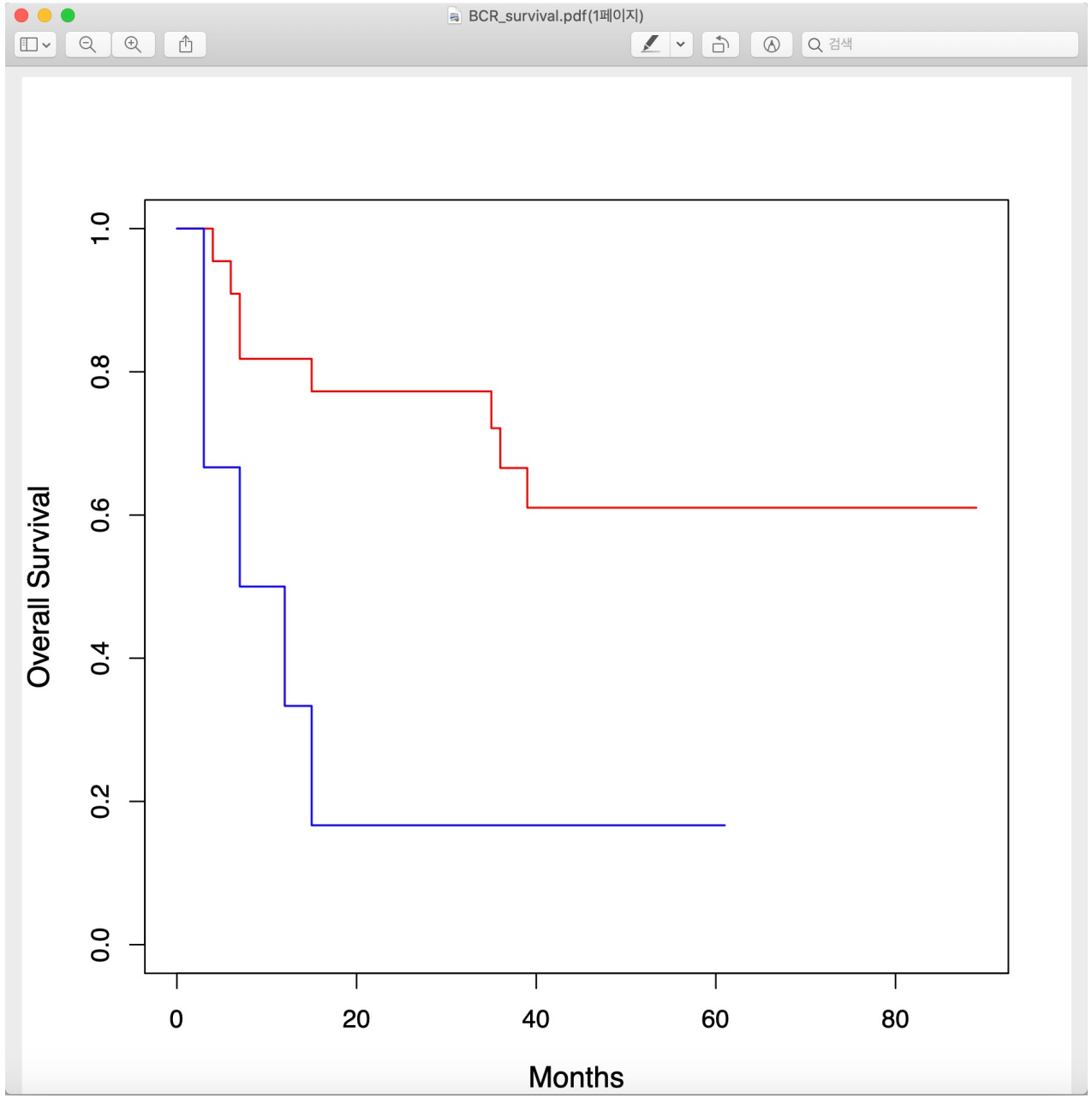

B3GAT2
PIGW
PFKFB4
B3GNT6
ENGASE
NDUFB11
ADCY10
OGT
GLO1
AK2
LAP3
POLR2G
MAT2A
RRM2
DHCR7
UGDH
TAT
SCD
POLR3E
FLAD1

heatmap.2(data.matrix(a), col=bluered(100), scale = 'row', ColSideColors=sample_color, **Colv=TRUE, Rowv = TRUE, dendrogram = ”both"**, cexRow = 0.90,

srtCol=45, adjCol = c(1,1), srtRow=0, adjRow=c(0, 1),

key=TRUE, keysize = 1, key.title = NA,symkey=FALSE, density.info="none", trace="none",,na.color="black")




BCR_survival.pdf(1페이지)
검색
Overall Survival
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Months

자동 저장 꺼짐

홈 삽입 그리기 페이지 레이아웃

붙여넣기 맑은 고딕 (본문)

가 가 가

J26

	A	B	C
1	sample_id	expression	
2	P-403	DOWN	
3	P-447	UP	
4	P-550	DOWN	
5	P-564	DOWN	
6	P-565	UP	
7	P-567	DOWN	
8	P-573	DOWN	
9	P-576	DOWN	
10	P-582	DOWN	
11	P-598	DOWN	
12	P-607	DOWN	
13	P-243	DOWN	
14	P-250	DOWN	
15	P-253	UP	
16	P-255	DOWN	
17	P-264	DOWN	
18	P-273	DOWN	
19	P-291	DOWN	
20	P-292	DOWN	
21	P-295	UP	
22	P-313	UP	
23	P-314	DOWN	
24	P-342	DOWN	
25	P-357	DOWN	
26	P-385	UP	
27	P-455	DOWN	
28	P-465	DOWN	
29	P-474	DOWN	
30			
31			
32			

expression_groups

준비

자동 저장 꺼짐

홈 삽입 그리기 페이지 레이아웃 수식

붙여넣기 맑은 고딕 (본문)

가 가 가

A1 sample_id

	A	B	C
1	sample_id	os_status	os_months
2	P-403	1	36
3	P-447	1	3
4	P-550	0	31
5	P-564	1	7
6	P-565	1	7
7	P-567	0	53
8	P-573	0	46
9	P-576	0	51
10	P-582	0	48
11	P-598	0	25
12	P-607	0	35
13	P-243	0	89
14	P-250	0	81
15	P-253	0	61
16	P-255	0	83
17	P-264	0	68
18	P-273	0	79
19	P-291	1	6
20	P-292	0	51
21	P-295	1	15
22	P-313	1	12
23	P-314	1	15
24	P-342	0	78
25	P-357	1	4
26	P-385	1	3
27	P-455	1	35
28	P-465	1	39
29	P-474	1	7
30			
31			

OS_DATA

준비

Go to file/function | Addins

d | drawBoxplot.R | data | drawSurvival.R* | firevat_script.r | clinical_data | drawBoxplot_wp.R* | iDATA

Filter

	sample_ID	OS_STATUS	OS_MONTHS	expression
1	P-243	0	89	DOWN
2	P-250	0	81	DOWN
3	P-253	0	61	UP
4	P-255	0	83	DOWN
5	P-264	0	68	DOWN
6	P-273	0	79	DOWN
7	P-291	1	6	DOWN
8	P-292	0	51	DOWN
9	P-295	1	15	UP
10	P-313	1	12	UP
11	P-314	1	15	DOWN
12	P-342	0	78	DOWN
13	P-357	1	4	DOWN
14	P-385	1	3	UP
15	P-403	1	36	DOWN
16	P-447	1	3	UP
17	P-455	1	35	DOWN
18	P-465	1	39	DOWN

Showing 1 to 19 of 28 entries, 4 total columns

Environment | History | Connections | Tutorial

Import Dataset

Global Environment

Data	
clinical_data	28 obs. of 4 variables
data	20532 obs. of 549 variables
expression	28 obs. of 2 variables
iDATA	549 obs. of 2 variables
surv_test	List of 17
survdiff_result	List of 6
Values	
gNAME	"OGT"
iFILE	"gene_expressions.csv"
iMax	4
iMin	3.26786534807899

Files | Plots | Packages | Help | Viewer

New Folder | Delete | Rename | More

Home > Desktop > R_code > survival

Name

..

BCR_survival.pdf

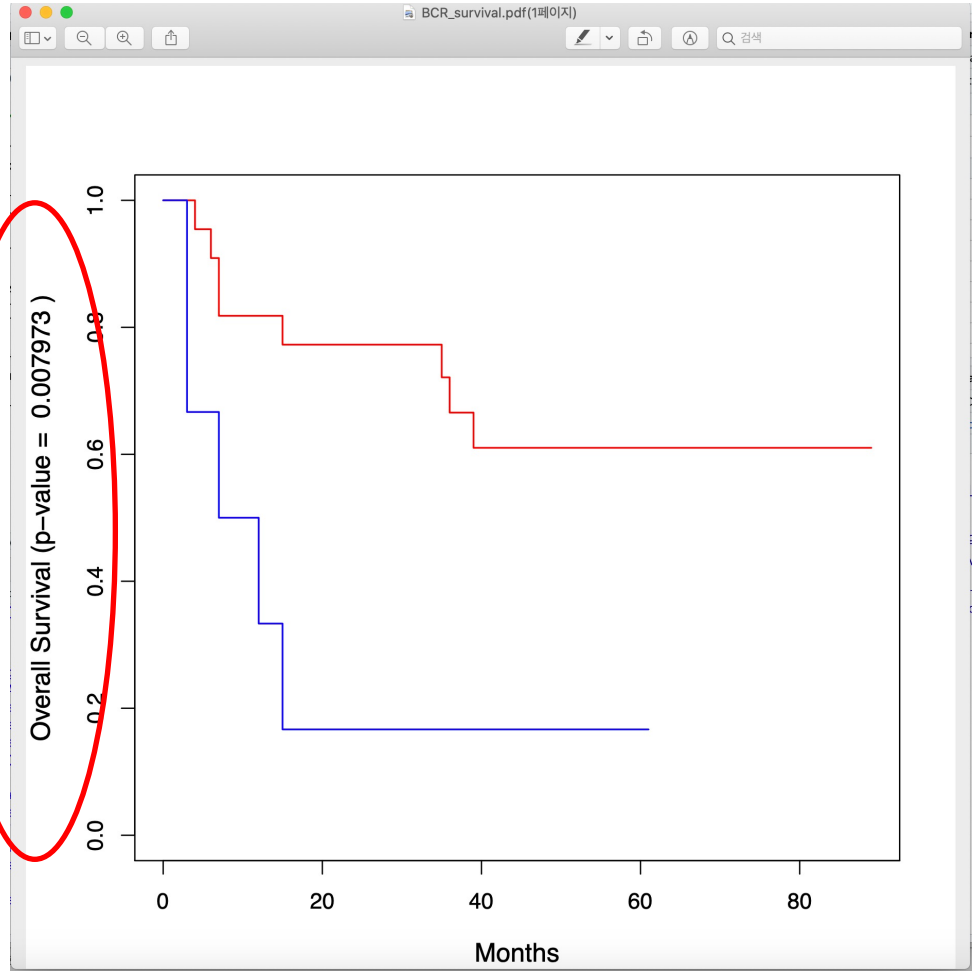
drawSurvival.R

expression_groups.csv

drawSurvival.R

```
library(survival)

setwd("/Users/dwhong/Desktop/R_code/survival")

expression <- read.csv("expression_groups.csv", header = TRUE, stringsAsFactors = FALSE, sep = ",")
names(expression) <- c("sample_ID", "expression")

clinical_data <- read.csv("OS_DATA.csv", header = TRUE, stringsAsFactors = FALSE, sep = ",")
names(clinical_data) <- c("sample_ID", "OS_STATUS", "OS_MONTHS")

clinical_data <- merge(clinical_data, expression, by = "sample_ID")

surv_test <- survfit(Surv(OS_MONTHS, OS_STATUS) ~ expression, data = clinical_data, conf.int = T)
survdiff_result <- survdiff(Surv(OS_MONTHS, OS_STATUS) ~ expression, data = clinical_data)
os_p_value <- 1 - pchisq(survdiff_result$chisq, df = (sum(1 * (survdiff_result$exp > 0))) - 1)

pdf(paste0(”OS_survival.pdf"))
plot(surv_test, lty = 1, lwd = 1.3, col = c("red", "blue"), xlab = "Months", ylab = "Overall Survival" , cex = 1.3, cex.lab=1.2,
xaxt="n")
axis(1, at = seq(0, 1200, 20))
invisible(dev.off())
```

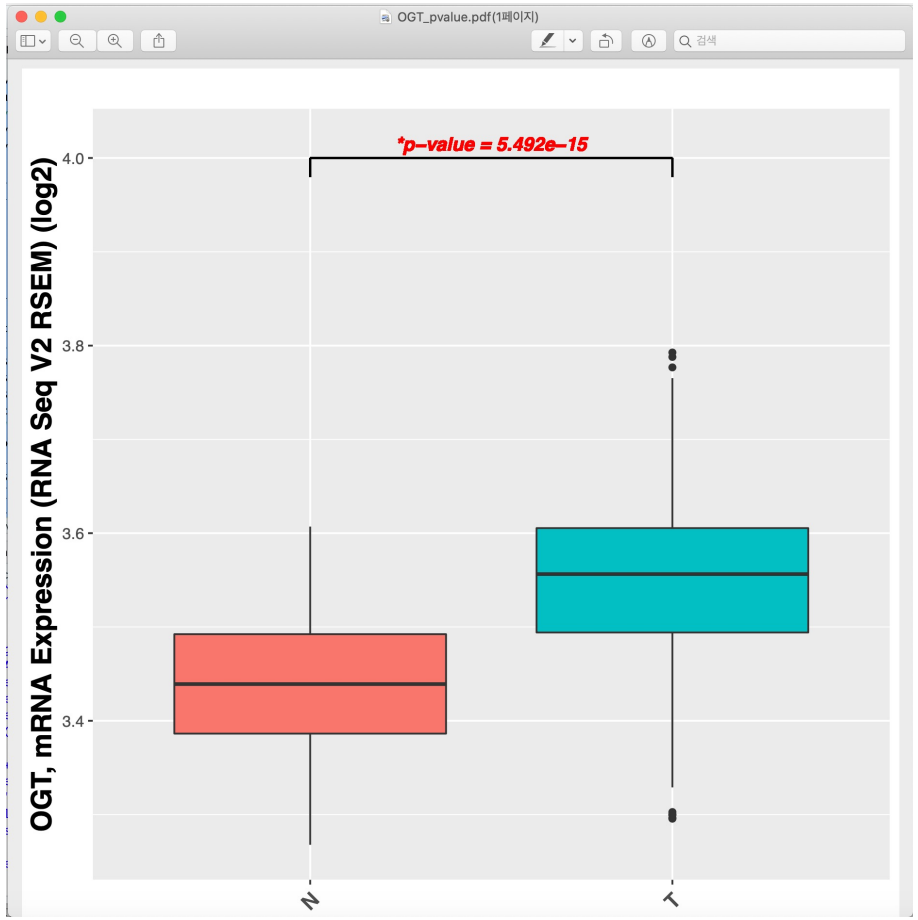
Adding a p-value




```
pdf(paste0("BCR_survival.pdf"))
plot(surv_test, lty = 1, lwd = 1.3, col = c("red", "blue"),
xlab = "Months", ylab = paste("Overall Survival (p-value
= ", format(os_p_value, digits = 4) , ")") , cex = 1.3,
cex.lab=1.2, xaxt="n")
axis(1, at = seq(0, 1200, 20))
invisible(dev.off())
```

Homework #1

- Y 축 legend에 “(log2)”를 추가하시오.
- 전체 스케일을 log2 로 변환하시오.

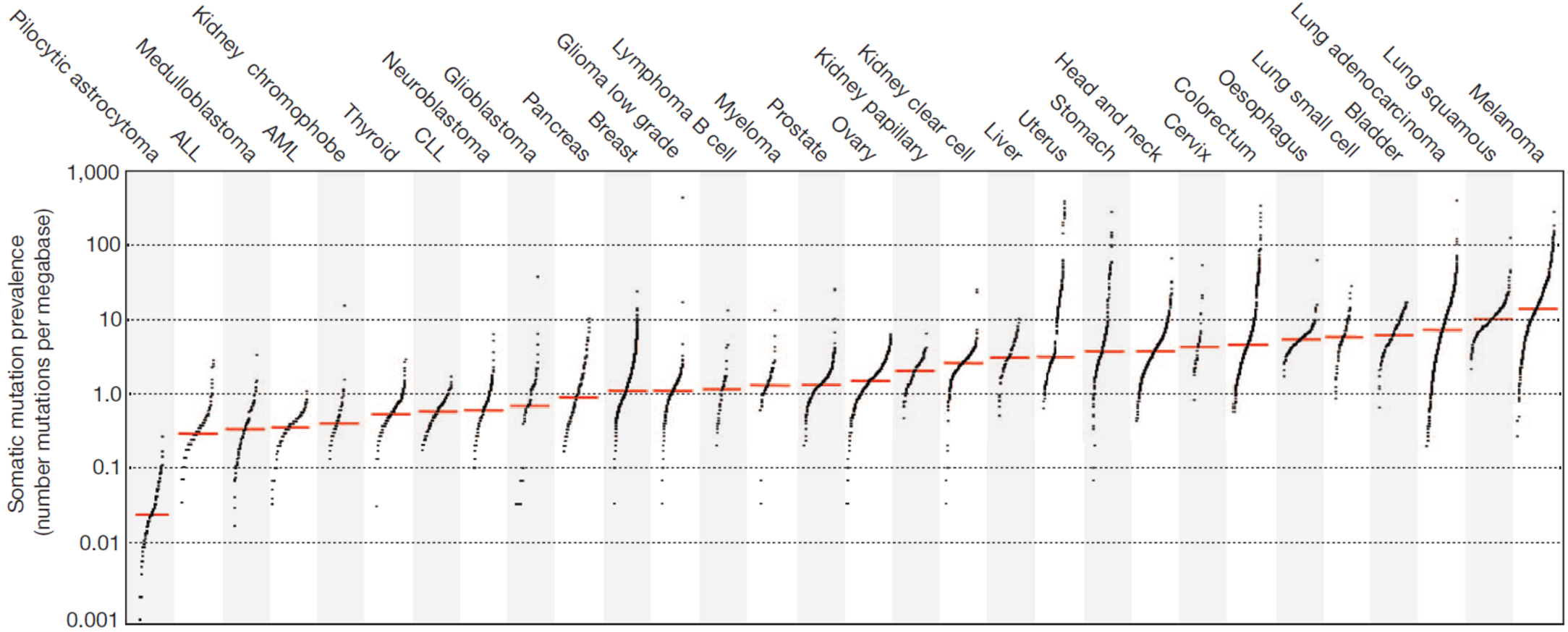


Homework #2

- Xena browser에서 LUAD 유전자 발현 데이터를 받으시오
- ALDH 관련 모든 유전자의 heatmap을 그리시오
  - 연관 샘플과 유전자 각각에 대해서 클러스터링 진행한 것과 row, column 클러스터링을 진행하지 않은 heatmap 데이터 2개를 각각 그리시오.

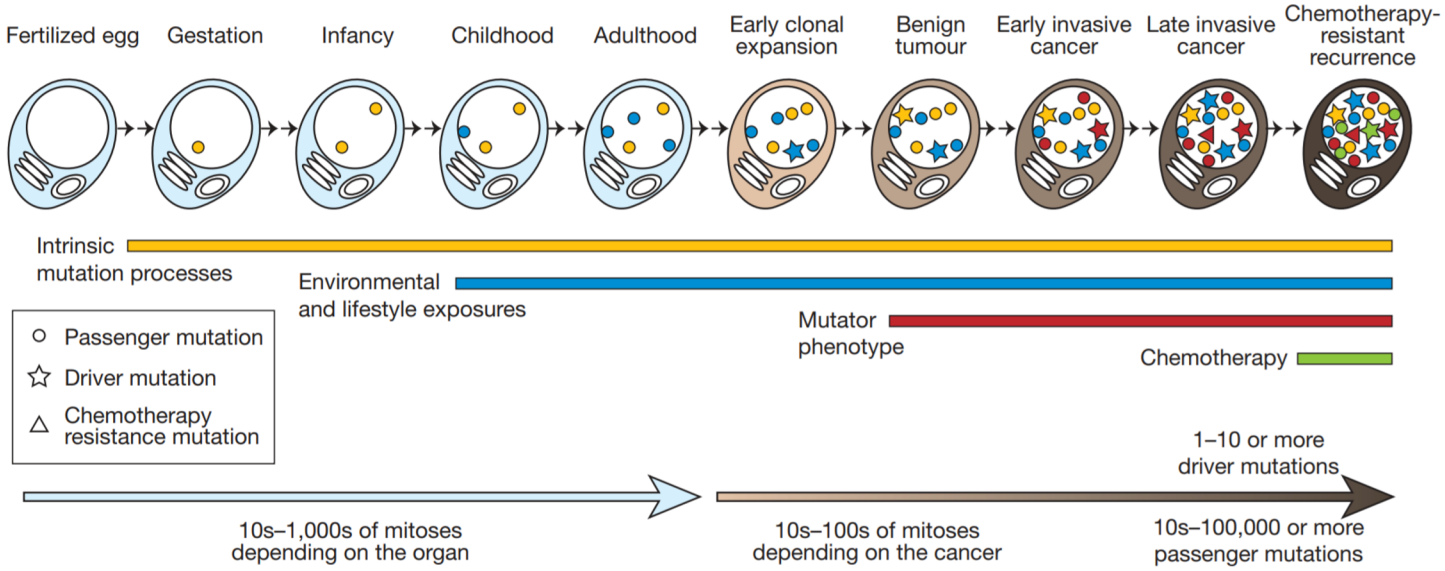
GDC Data Portal (https://portal.gdc.cancer.gov)

Somatic mutation rates vary by tumor tissue type

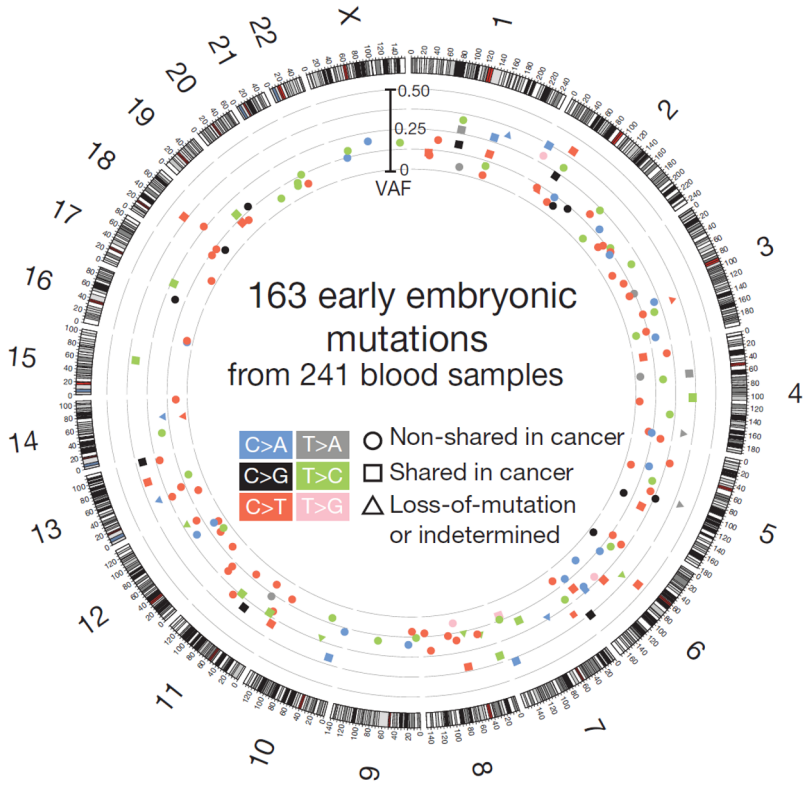


Alexandrov et al., Nature 2013

Cancer is a genetic disease caused by somatic mutations



Stratton et al., Nature 2009



Ju et al., Nature 2017

Q.

- Which gene has the most mutations in lung adenocarcinoma ?

http://firebrowse.org

FIREBROWSE

Search analysis results

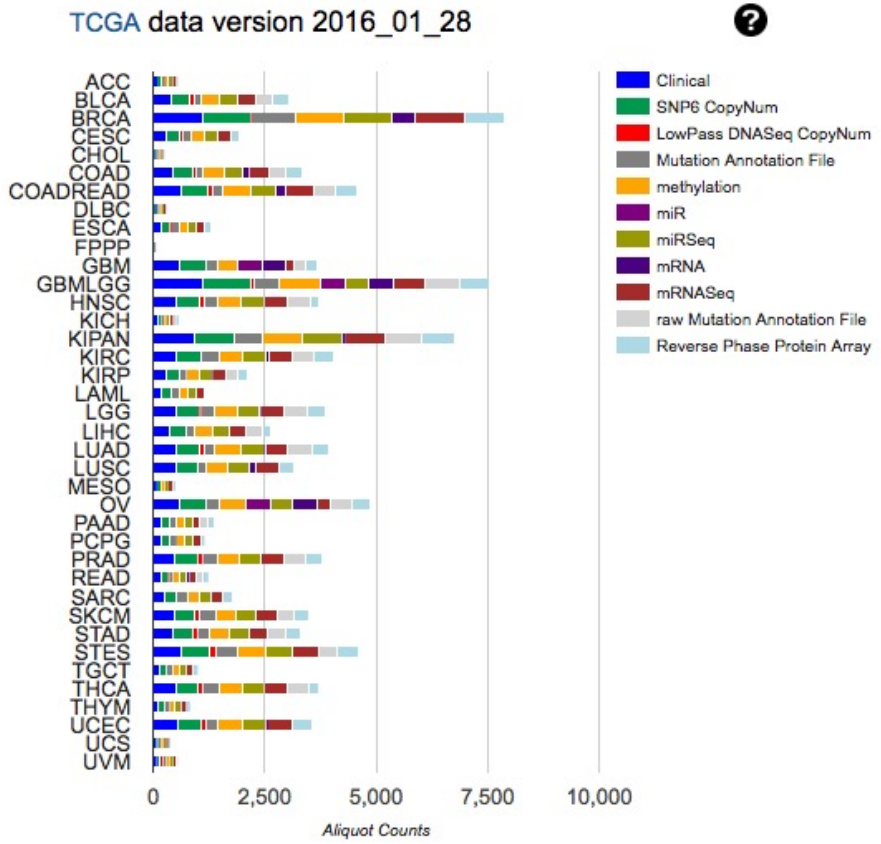
HOME | BROAD GDAC | WEB API | FAQ | SAMPLES REPORT | AWG RESULTS | OLD RUNS | TUTORIAL | RELEASE NOTES | CONTACT

View Expression Profile | Enter gene name | TP53

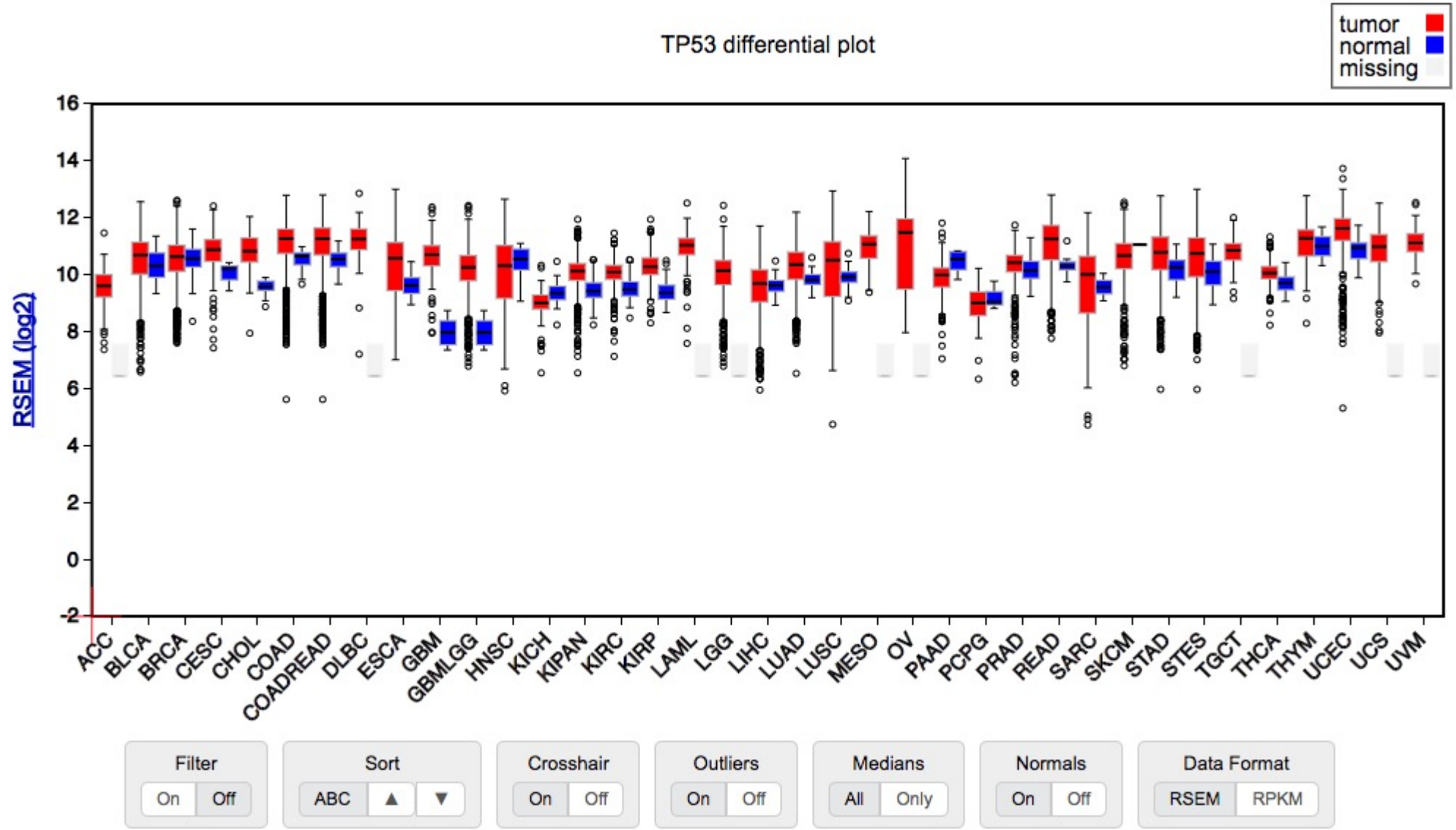
Enter cohort abbrev | LUAD | View Analysis Profile

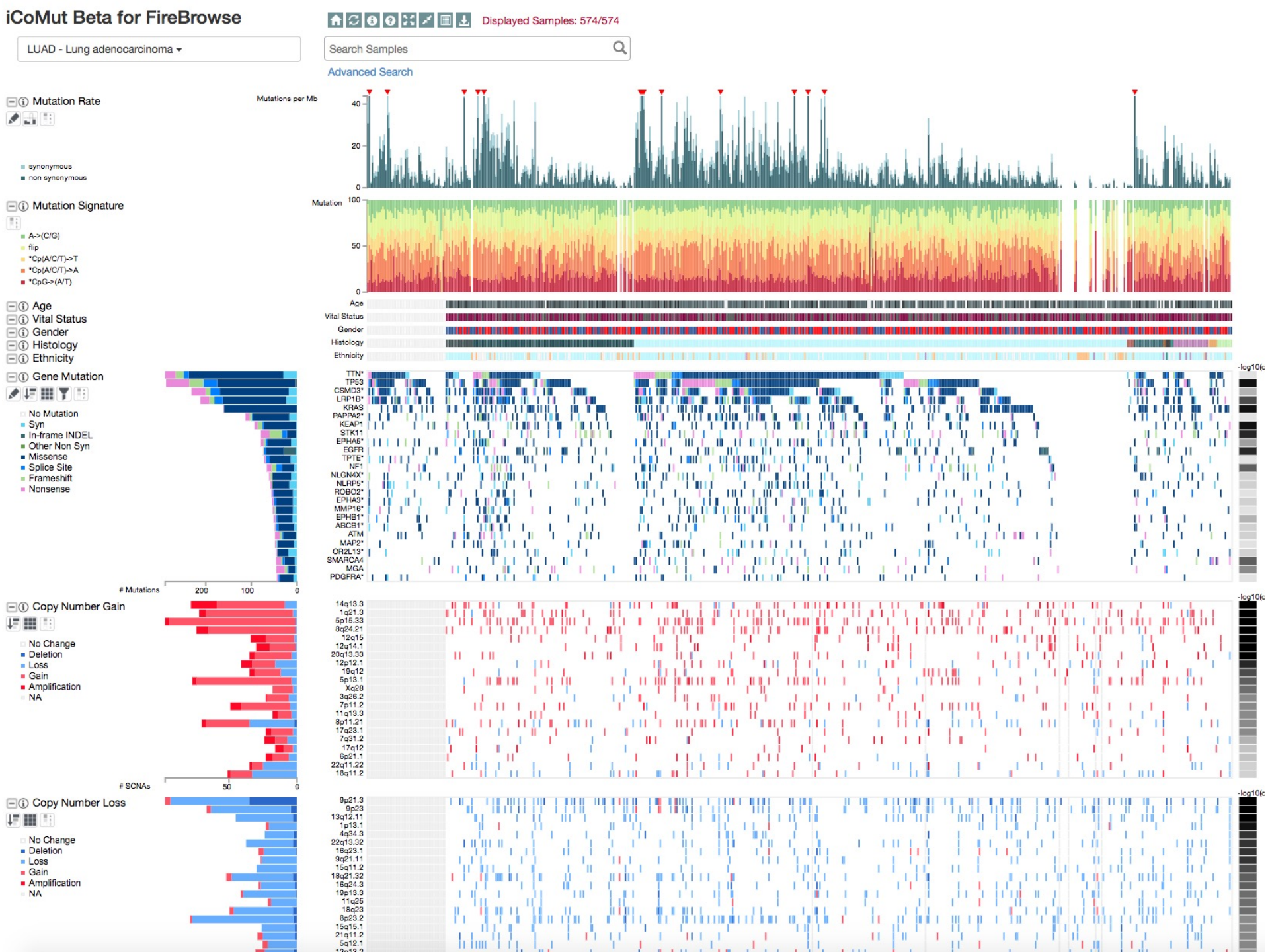
SELECT COHORT

- Clinical Analyses
- CopyNumber Analyses
- Correlations Analyses
- miR Analyses
- miRseq Analyses
- mRNA Analyses
- mRNAseq Analyses
- Mutation Analyses
- Pathway Analyses
- RPPA Analyses



Version: 1.1.38 (2018-02-26 12:20:33 484103261f6ef681a05cf163)

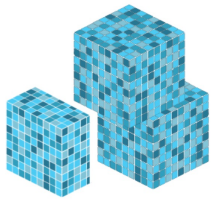

TP53 differential plot
tumor
normal
missing
RSEM (log2)
16
14
12
10
8
6
4
2
0
-2
ACC
BLCA
BRCA
CESC
CHOL
COAD
COADREAD
DLBC
ESCA
GBM
GBMLGG
HNSC
KICH
KIPAN
KIRC
KIRP
LAML
LGG
LIHC
LUAD
LUSC
MESO
OV
PAAD
PCPG
PRAD
READ
SARC
SKCM
STAD
STES
TGCT
THCA
THYM
UCEC
UCS
UVM
Filter
On
Off
Sort
ABC
▲
▼
Crosshair
On
Off
Outliers
On
Off
Medians
All
Only
Normals
On
Off
Data Format
RSEM
RPKM


iCoMut Beta for FireBrowse
LUAD - Lung adenocarcinoma
Displayed Samples: 574/574
Search Samples
Advanced Search
Mutation Rate
synonymous
non synonymous
Mutations per Mb
Mutation Signature
A->(C/G)
flip
*Cp(A/C/T)->T
*Cp(A/C/T)->A
*CpG->(A/T)
Mutation
Age
Vital Status
Gender
Histology
Ethnicity
Gene Mutation
No Mutation
Syn
In-frame INDEL
Other Non Syn
Missense
Splice Site
Frameshift
Nonsense
Mutations
TTN*
TP53
CSMD3*
LRP1B*
KRAS
PAPPA2*
KEAP1
STK11
EPHA5*
EGFR
TPTE*
NF1
NLGN4X*
NLRP5*
ROBO2*
EPHA3*
MMP16*
EPHB1*
ABCB1*
ATM
MAP2*
OR2L13*
SMARCA4
MGA
PDGFRA*
Copy Number Gain
No Change
Deletion
Loss
Gain
Amplification
NA
SCNAs
14q13.3
1q21.3
5p15.33
8q24.21
12q15
12q14.1
20q13.33
12p12.1
19q12
5p13.1
Xq28
3q26.2
7p11.2
11q13.3
8p11.21
17q23.1
7q31.2
17q12
6p21.1
22q11.22
18q11.2
Copy Number Loss
No Change
Deletion
Loss
Gain
Amplification
NA
9p21.3
9p23
13q12.11
1p13.1
4q34.3
22q13.32
16q23.1
9q21.11
15q11.2
18q21.32
16q24.3
19p13.3
11q25
18q23
8p23.2
15q15.1
21q11.2
5q12.1

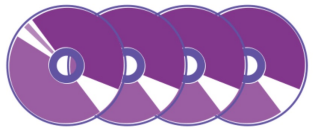
International Cancer Research Projects : TCGA, ICGC, PCAWG

NIH **THE CANCER GENOME ATLAS**
National Cancer Institute
National Human Genome Research Institute

TCGA produced over **2.5** PETABYTES of data



To put this into perspective, **1 petabyte** of data is equal to **212,000** DVDs



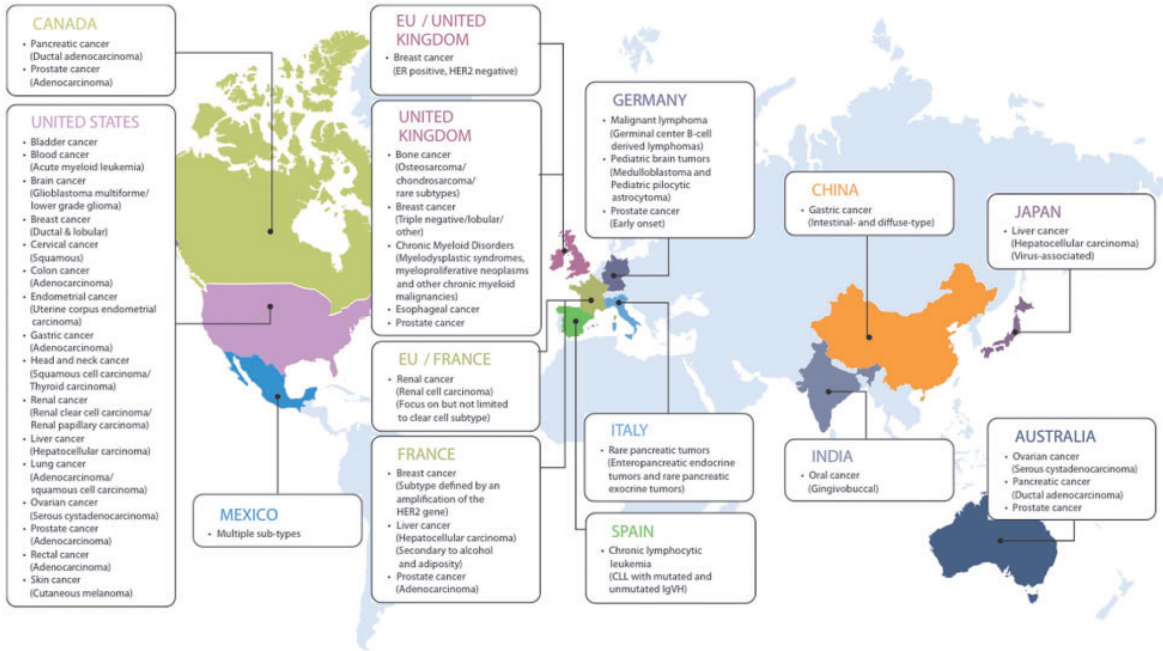
TCGA data describes **33** DIFFERENT TUMOR TYPES ...including **10** RARE CANCERS

...based on paired tumor and normal tissue sets collected from **11,000** PATIENTS

...using **7** DIFFERENT DATA TYPES

International Cancer Genome Consortium

Zhang et al (Database 2011)



PCAWG PanCancer Analysis OF WHOLE GENOMES

2800 pairs of whole genomes amount to **~800TB raw data**			
Workflow	Compute (Cores /RAM)	Average runtimes	storage per donor
BWA-MEM alignment	8 / 16GB	5 days / specimen X 2	240GB
Sanger	8 / 32GB	4 days / donor	2GB
DKFZ/EMBL	16 /64GB	2 days / donor	5GB
Broad	32 / 256GB	3 days / donor	35GB
Total per donor		19 days	282GB
Total for 2800 donors		>53,000 days (145 years)	**~800TB (30 years HD movie)**

AACR 2018

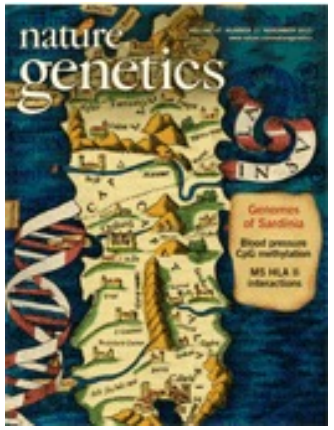
Intron retention is a widespread mechanism of tumor-suppressor inactivation

Hyunchul Jung, Donghoon Lee, Jongkeun Lee, Donghyun Park, Yeon Jeong Kim, Woong-Yang Park, Dongwan Hong, Peter J Park & Eunjung Lee

Affiliations | Contributions | Corresponding authors

Nature Genetics **47**, 1242–1248 (2015) | doi:10.1038/ng.3414

Received 30 November 2014 | Accepted 08 September 2015 | Published online 05 October 2015



Higher impact papers through reusing public data

Dr. Myles Axton
Chief Editor,
Nature Genetics.

EDITORIAL

nature genetics

NATURE GENETICS | VOLUME 46 | NUMBER 3 | MARCH 2014

Call for data analysis papers

Community standards for data access, interoperability and metadata only make sense if data are creatively reused to further research. We are therefore inviting the submission of Analysis papers that reformat and integrate existing data sets to generate substantial novel insights into gene expression in cell differentiation transitions and different cell fates.

Examples of abnormal splicing related intron retention

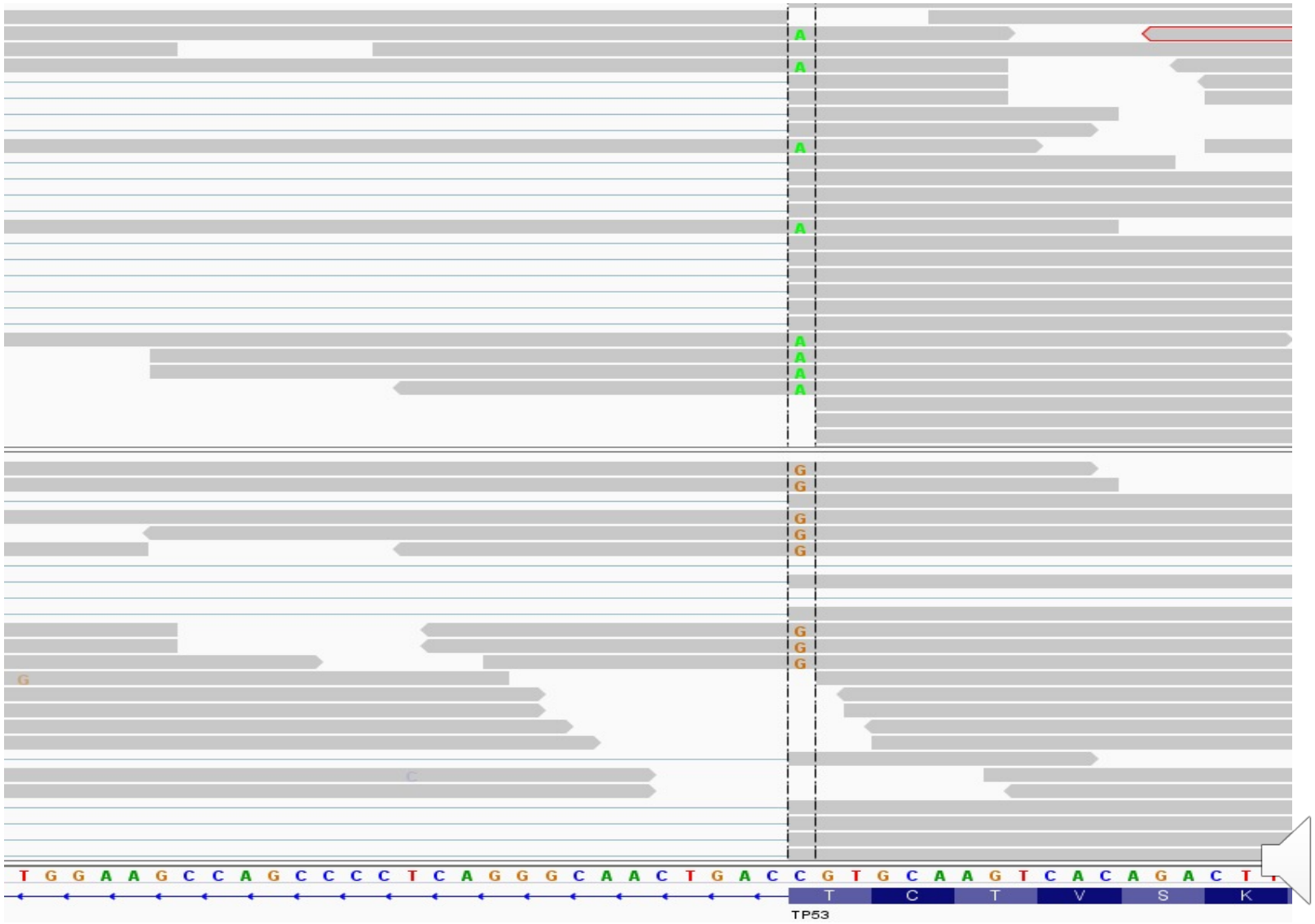
LUNG : TCGA-60-2712

(C > A, silent mutation)

BREAST : TCGA-AO-A12F

(C > G, silent mutation)

TP53 (chr17 : 7579312)



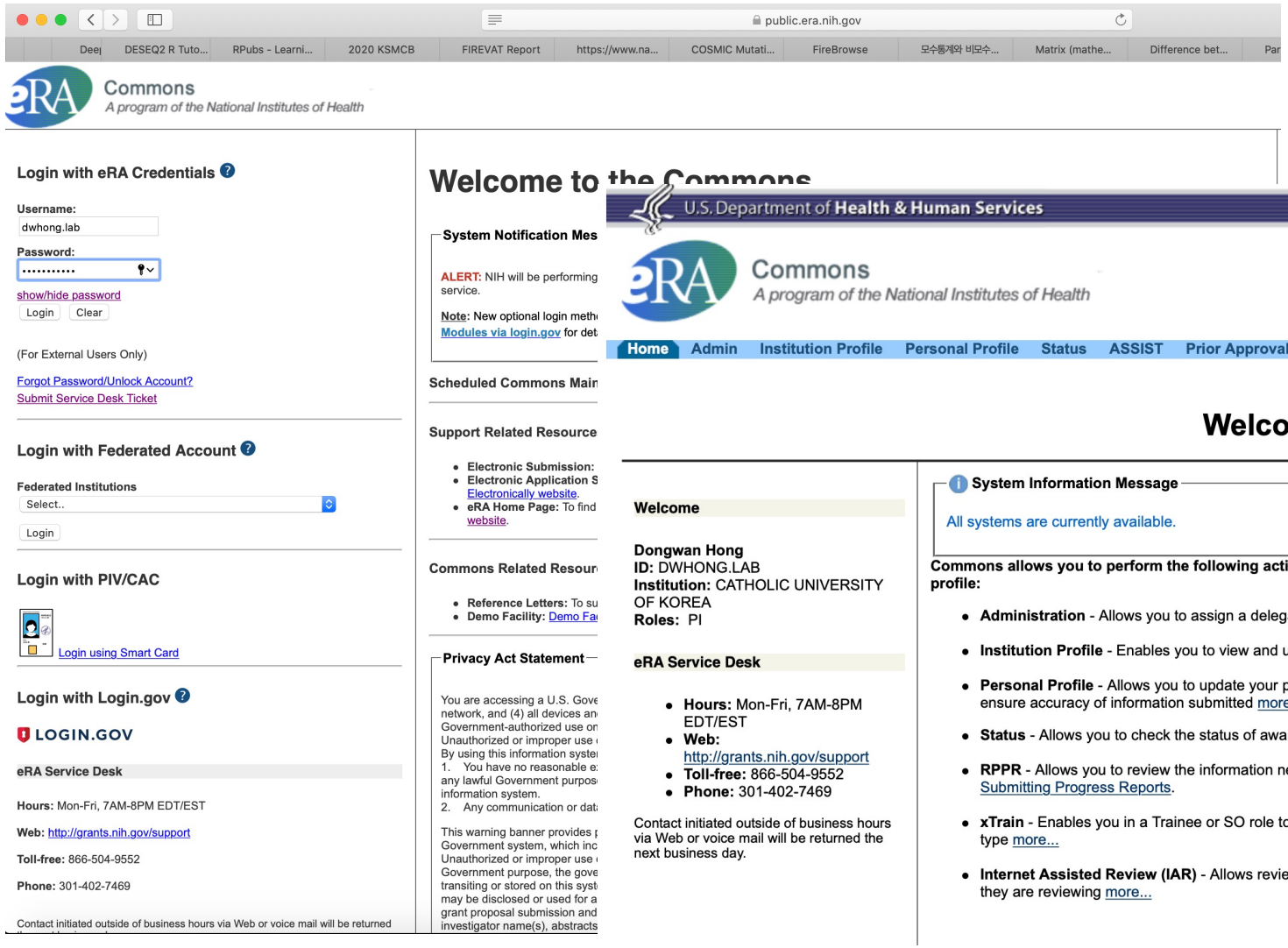
Cancer sequencing statistics

Cancer Type	Number of samples	RNA-Seq read length
BRCA	503	2 x 50 bp
COAD	217	76 bp
KIRC	400	2 x 50 bp
LUSC	178	2 x 50 bp
OV	273	2 x 75 bp
UCEC	241	76 bp
Total	1,812	

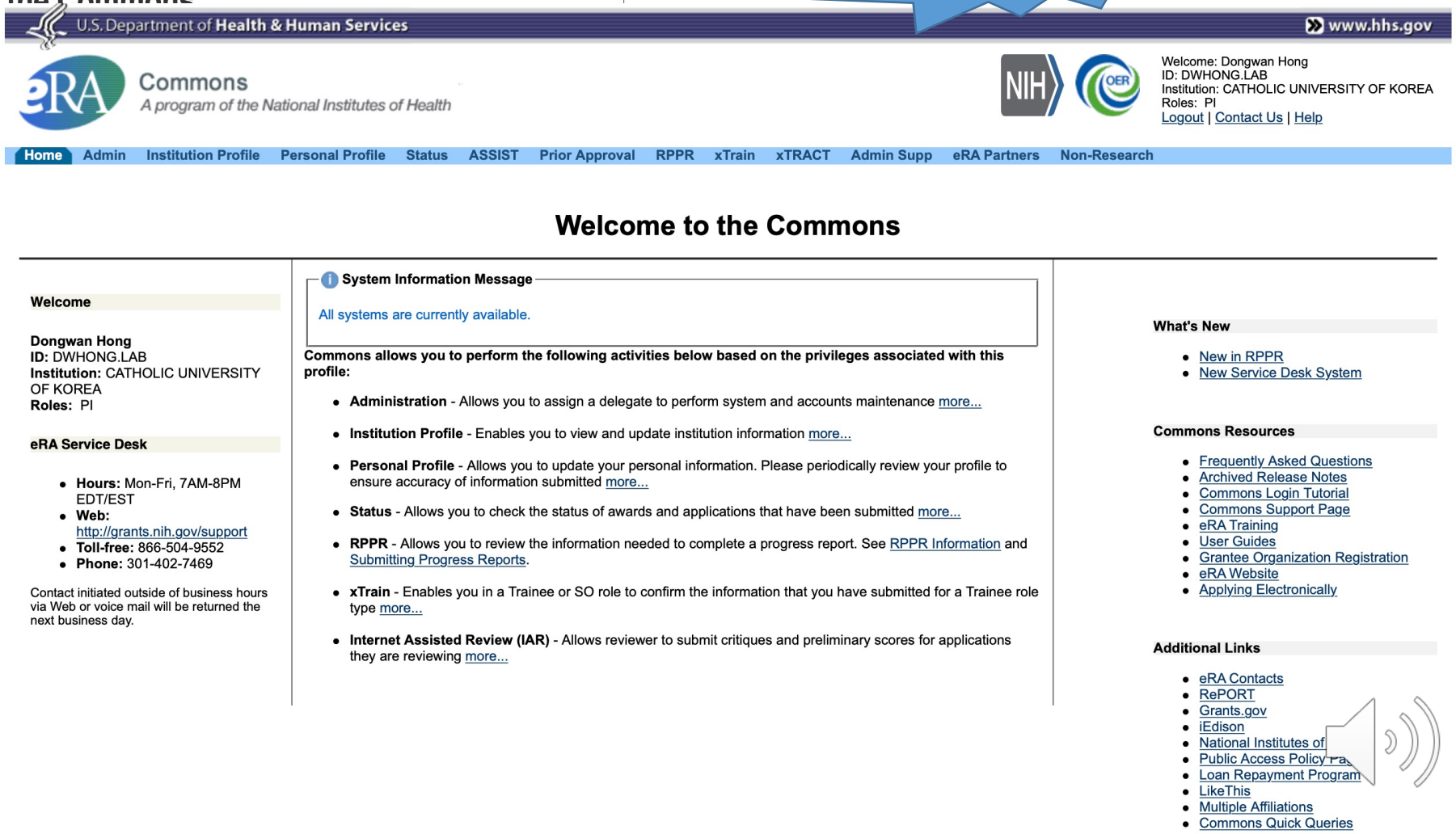
Case: coverage 100x RNA-Seq
File size: ~8.8GByte
8,800,000,000 Byte
70,400,000,000 bit (~70Gbit)

~7,040 sec = ~2 hours / 1 file
1,812 patients -> 3,624 hours
151 days

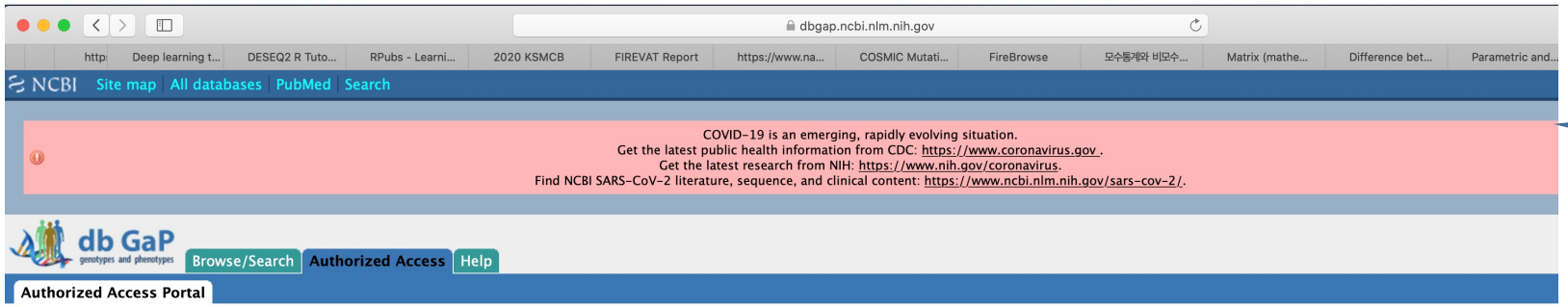
Getting an account of eRA commons



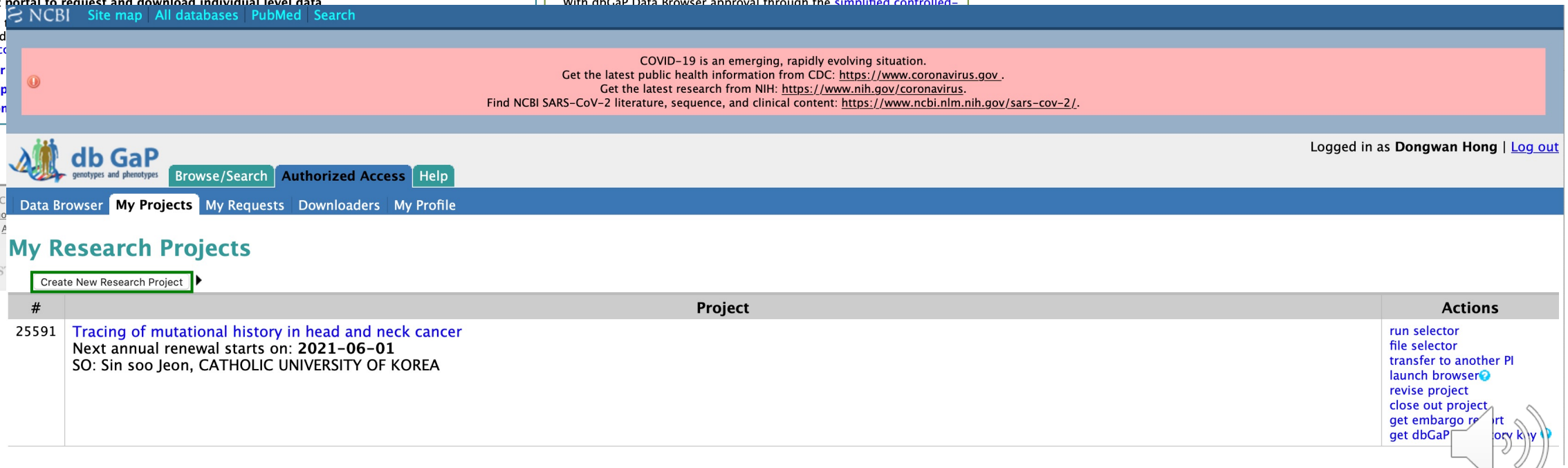
Controlled-access



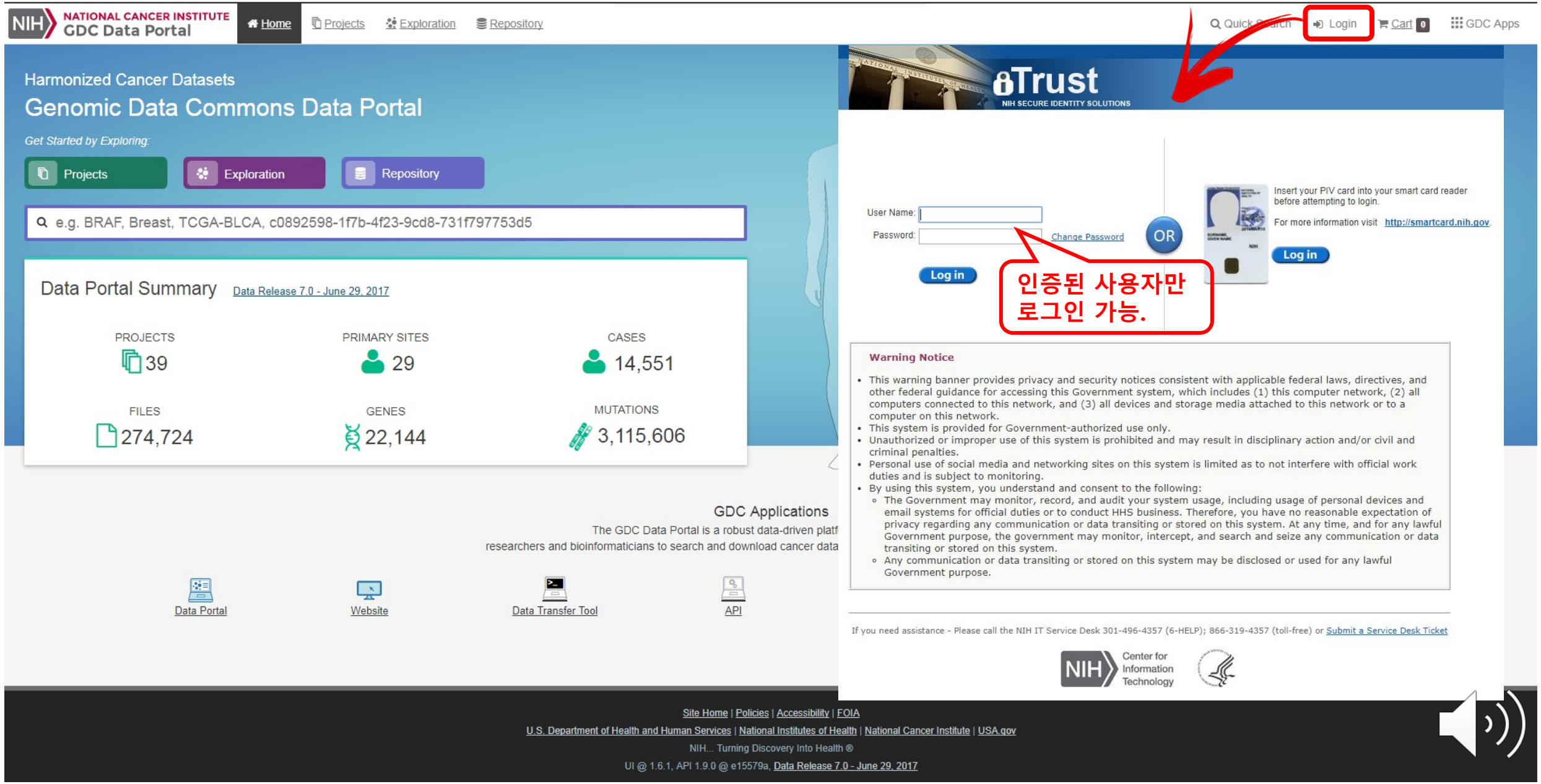
Getting an account of dbGaP



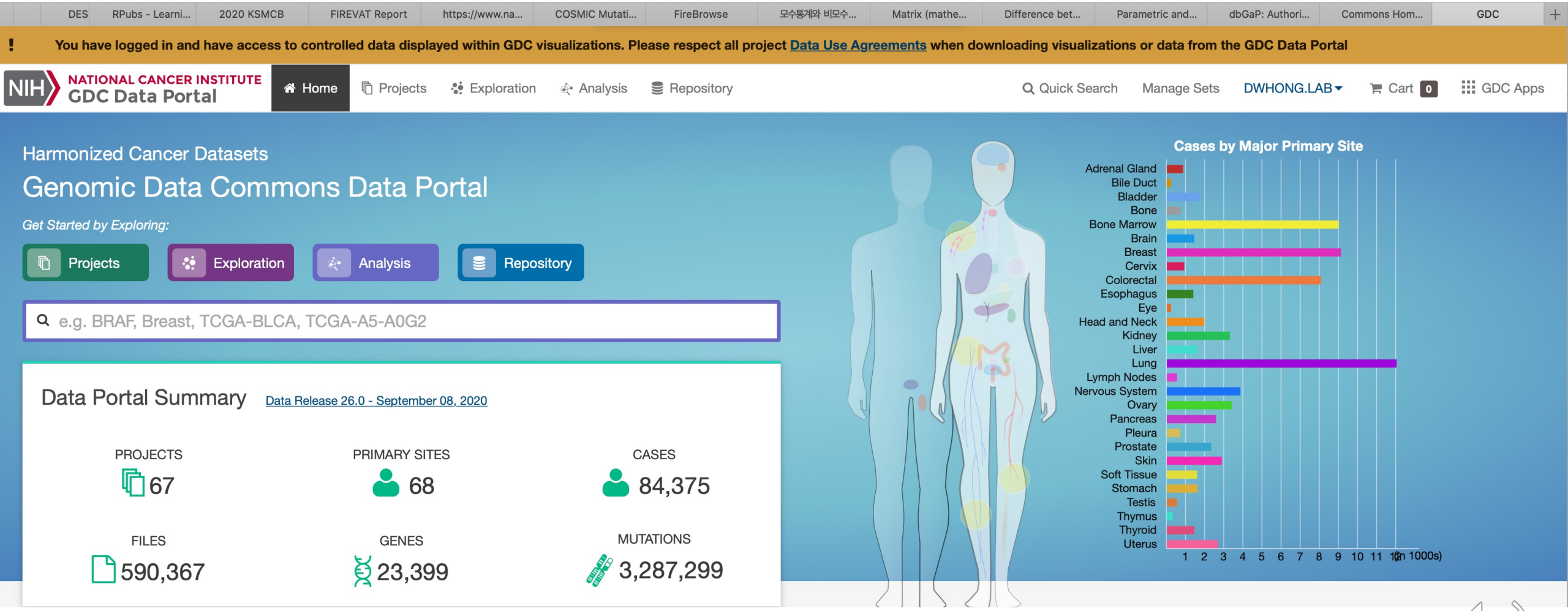
Controlled-access

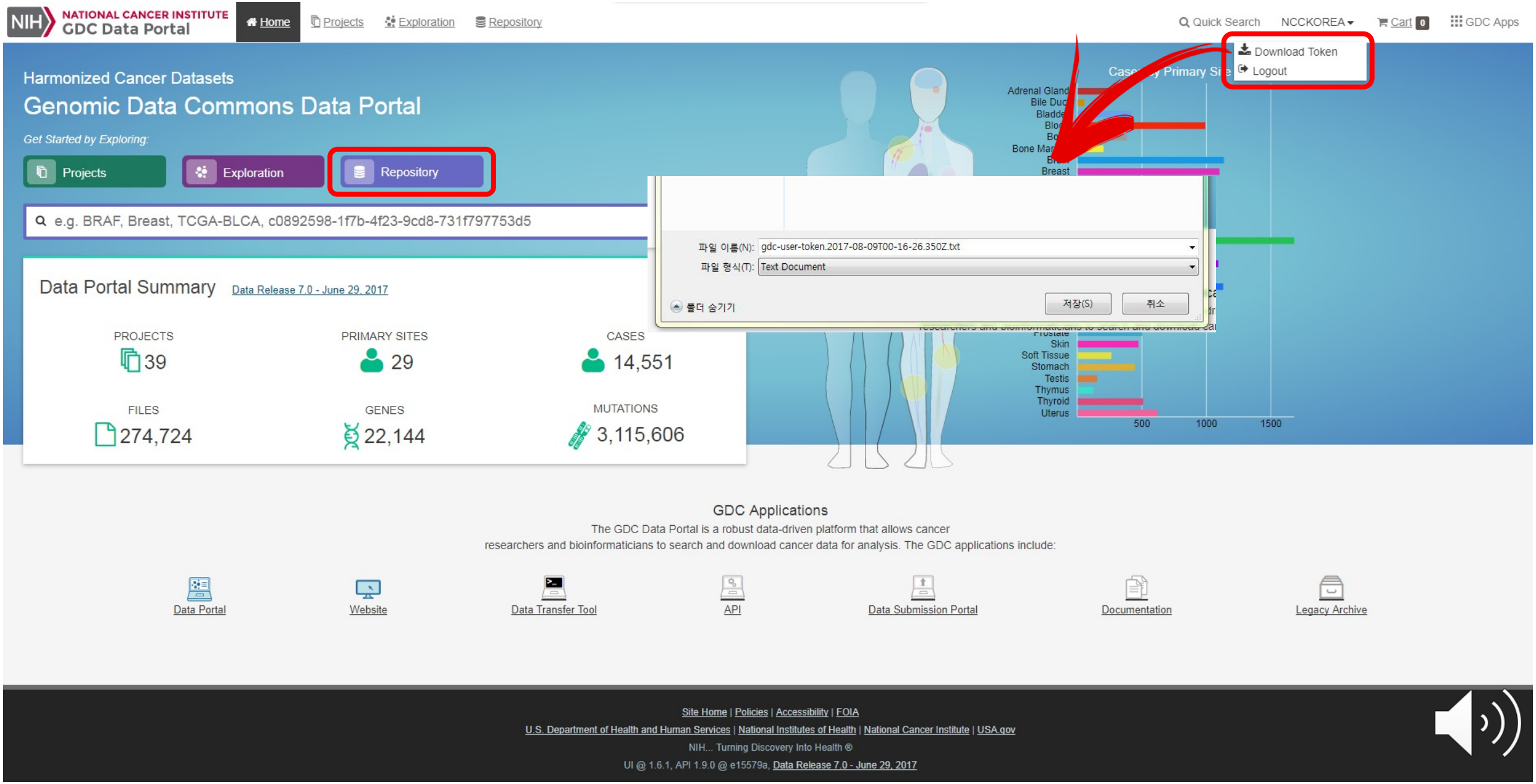


GDC Data Portal



Data Portal Summary




NIH NATIONAL CANCER INSTITUTE GDC Data Portal
Home
Projects
Exploration
Repository
Quick Search
NCCKOREA
Cart 0
GDC Apps
Download Token
Logout
Harmonized Cancer Datasets
Genomic Data Commons Data Portal
Get Started by Exploring:
Projects
Exploration
Repository
e.g. BRAF, Breast, TCGA-BLCA, c0892598-1f7b-4f23-9cd8-731f797753d5
Adrenal Gland
Bile Duc
Bladde
Blo
Bo
Bone Mar
Br
Breast
Prostate
Skin
Soft Tissue
Stomach
Testis
Thymus
Thyroid
Uterus
500
1000
1500
파일 이름(N): gdc-user-token.2017-08-09T00-16-26.350Z.txt
파일 형식(T): Text Document
폴더 숨기기
저장(S)
취소
Data Portal Summary
Data Release 7.0 - June 29, 2017
PROJECTS
39
PRIMARY SITES
29
CASES
14,551
FILES
274,724
GENES
22,144
MUTATIONS
3,115,606
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
Data Transfer Tool
API
Data Submission Portal
Documentation
Legacy Archive
Site Home | Policies | Accessibility | FOIA
U.S. Department of Health and Human Services | National Institutes of Health | National Cancer Institute | USA.gov
NIH... Turning Discovery Into Health ®
UI @ 1.6.1, API 1.9.0 @ e15579a, Data Release 7.0 - June 29, 2017

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home | Projects | Exploration | Repository | Quick Search | NCCKOREA | Cart 0 | GDC Apps

Cases | Files

Add a Case/Biospecimen Filter

Case

Search for Case ID

Case Submitter ID

eg. TCGA-DD*, *DD*, TCGA-DD-AAVP | Go!

Primary Site

Program

Project

Disease Type

Gender

Age at Diagnosis

Vital Status

Days to Death

Race

Ethnicity

← Start searching by selecting a facet

Add All Files to Cart | Download Manifest

Cases (14,551) | Files (274,724)

Primary Sites

Showing 1 - 20 of 14,551 cases

Cart	Case UUID	Submitter ID
	1a20f675	TCGA-HT-A74J
	4d6b5b30	TCGA-43-A56U
	65cac997	TCGA-GM-A3XL
	08de63a2	TCGA-A1-A0SQ
	78c0d30c	TCGA-K1-A6RV
	096bd95f	TCGA-J2-A4AD
	ec4b4d34	TCGA-XR-A8TE
	9d8c4693	TCGA-CM-6170
	69d0a566	TCGA-27-2526
	a36a4440	TCGA-24-1844
	1a02b991	TCGA-63-A5MI

Cases | Files

Add a Case/Biospecimen Filter

Case

Search for Case ID

Case Submitter ID

eg. TCGA-DD*, *DD*, TCGA-DD-AAVP | Go!

Primary Site

Kidney	1,681
Brain	1,133
Nervous System	1,127
Breast	1,098
Lung	1,089

24 More...

Program

TCGA	11,315
TARGET	3,236

Project

TARGET-NBL	1,127
TCGA-BRCA	1,098
TARGET-AML	988
TARGET-WT	652
TCGA-GBM	617

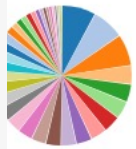
34 More...

Advanced Search

Browse Annotations

470.59 TB

Disease Type



Gender

Vital Status

JSON | TSV

Gender	Files	Available Files per Data Category							Annotations
		Seq	Exp	SNV	CNV	Meth	Clinical	Bio	
Male	29	3	3	16	4	1	1	1	0
Female	32	4	5	16	4	1	1	1	0
Female	32	4	5	16	4	1	1	1	0
Female	30	4	5	16	2	1	1	1	0
Male	32	4	5	16	4	1	1	1	0
Female	32	4	5	16	4	1	1	1	0
Male	32	4	5	16	4	1	1	1	0
Female	32	4	5	16	4	1	1	1	0
Female	39	5	3	24	4	1	1	1	1
Female	32	4	5	16	4	1	1	1	0
Male	30	4	5	16	2	1	1	1	0

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home | Projects | Exploration | Repository

Quick Search | NCCKOREA | Cart 0 | GDC Apps

Cases | Files

Add a File Filter

File

Search for File ID or File Submitter ID

> Data Category
> Data Type
> Experimental Strategy
> Workflow Type
> Data Format
> Platform
> Access

← Start searching by selecting a facet

Add All Files to Cart | Download Manifest

Primary Sites

Showing 1 - 20 of 14,551 cases

Cart	Case UUID	Submitter ID
	1a20f675	TCGA-HT-A74J
	4d6b5b30	TCGA-43-A56U
	65cac997	TCGA-GM-A3XL
	08de63a2	TCGA-A1-A0SQ
	78c0d30c	TCGA-K1-A6RV
	096bd95f	TCGA-J2-A4AD
	ec4b4d34	TCGA-XR-A8TE
	9d8c4693	TCGA-CM-6170
	69d0a566	TCGA-27-2526
	a36a4440	TCGA-24-1844
	1a02b991	TCGA-63-A5MI

Cases | Files

Add a File Filter

File

Search for File ID or File Submitter ID

Data Category

Simple Nucleotide Variation	5,368
Transcriptome Profiling	2,916
Raw Sequencing Data	2,462
Copy Number Variation	2,294
DNA Methylation	657

2 More...

Data Type

Annotated Somatic Mutation	2,680
Raw Simple Somatic Mutation	2,680
Aligned Reads	2,462
Gene Expression Quantification	1,782
Copy Number Segment	1,147

8 More...

Experimental Strategy

WXS	6,668
RNA-Seq	2,377
Genotyping Array	2,294
miRNA-Seq	1,701
Methylation Array	657

Advanced Search

Browse Annotations

470.59 TB

se Type | Gender | Vital Status

JSON | TSV

Gender	Files	Available Files per Data Category							Annotations
		Seq	Exp	SNV	CNV	Meth	Clinical	Bio	
Male	29	3	3	16	4	1	1	1	0
Female	32	4	5	16	4	1	1	1	0
Female	32	4	5	16	4	1	1	1	0
Female	30	4	5	16	2	1	1	1	0
Male	32	4	5	16	4	1	1	1	0
Female	32	4	5	16	4	1	1	1	0
Male	32	4	5	16	4	1	1	1	0
Female	32	4	5	16	4	1	1	1	0
Female	39	5	3	24	4	1	1	1	1
Female	32	4	5	16	4	1	1	1	0
Male	30	4	5	16	2	1	1	1	0

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home | Projects | Exploration | Repository | Quick Search | NCCKOREA | Cart 0 | GDC Apps

Cases | Files

Add a File Filter

File

Search for File ID or File Submitter ID

Data Category
- [x] Simple Nucleotide Variation 670

Data Type
- [x] Annotated Somatic Mutation 670

Experimental Strategy
- [ ] WXS 670

Workflow Type
- [ ] MuSE Annotation 670
- [x] MuTect2 Annotation 670
- [ ] SomaticSniper Annotation 670
- [ ] VarScan2 Annotation 670

Data Format
- [ ] VCF 670

Platform

No data for this field

Access
- [ ] controlled 670

Clear | Primary Site | IS | Lung | AND | Project Id | IS | TCGA-LUAD | AND | Workflow Type | IS | MuTect2 Annotation | AND | Data Category | IS | Simple Nucleotide Variation | AND | Data Type | IS | Annotated Somatic Mutation

Advanced Search

Add All Files to Cart | Download Manifest

Browse Annotations

전체 샘플 다운로드

Cases (569) | Files (670)

1.36 GB

Primary Sites | Projects | Disease Type | Gender | Vital Status

Showing 1 - 20 of 569 cases

개별 샘플 다운로드

JSON | TSV

Cart	Case UUID	Submitter ID	Project	Primary Site	Gender	Files	Available Files per Data Category							Annotations
							Seq	Exp	SNV	CNV	Meth	Clinical	Bio	
	096bd63f	TCGA-J2-A4AD	TCGA-LUAD	Lung	Female	32	4	5	16	4	1	1	1	0
	66763a0c	TCGA-05-4424	TCGA-LUAD	Lung	Male	32	4	5	16	4	1	1	1	0
	7b89166e	TCGA-86-8669	TCGA-LUAD	Lung	Male	32	4	5	16	4	1	1	1	0
	e43a2b72	TCGA-55-A493	TCGA-LUAD	Lung	Female	32	4	5	16	4	1	1	1	0
	a905d275	TCGA-78-8655	TCGA-LUAD	Lung	Female	32	4	5	16	4	1	1	1	0
	ddeacccf	TCGA-44-5645	TCGA-LUAD	Lung	Female	135	15	14	88	12	4	1	1	0
	a56d2b48	TCGA-17-Z021	TCGA-LUAD	Lung	--	20	2	0	16	0	1	0	1	0
	243c6fd6	TCGA-69-7764	TCGA-LUAD	Lung	Male	32	4	5	16	4	1	1	1	0
	bf34664d	TCGA-44-A4SU	TCGA-LUAD	Lung	Female	32	4	5	16	4	1	1	1	0
	47062ad2	TCGA-MP-A4T7	TCGA-LUAD	Lung	Female	32	4	5	16	4	1	1	1	0
	8b3a111a	TCGA-17-Z026	TCGA-LUAD	Lung	--	20	2	0	16	0	1	0	1	0
	0232d299	TCGA-91-6848	TCGA-LUAD	Lung	Male	32	4	5	16	4	1	1	1	0
	e60a8f8a	TCGA-62-A470	TCGA-LUAD	Lung	Male	30	4	5	16	2	1	1	1	0

FILES
670

CASES
569

FILE SIZE
1.36 GB

File Counts by Project

Project	Cases	Files	File Size
TCGA-LUAD	569	670	1.36 GB

File Counts by Authorization Level

Level	Files	File Size
Authorized	670	1.36 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.

Cart Items

Metadata | Download | Remove From Cart

Manifest
Cart

Showing **1** - **20** of **670** files

TSV

Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
controlled	3978ca8b-8802-4f4a-a83d-15b0d95c1ba0.vep.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF	1.18 MB	1
controlled	8dd7ac3c-3a90-4738-b70f-56fa4f48d725.vep.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF	1.49 MB	0
controlled	08751465-d6e7-43ef-8eaf-f9ec33957f5e.vep.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF	1.42 MB	0
controlled	81d3e78a-5cc2-4db4-adc4-8c9652a37b1e.vep.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF	2.49 MB	0
controlled	4493aeec-ce91-4b1f-b68a-528cb80de7ed.vep.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF	1.67 MB	0
controlled	6d1da113-0ee4-4705-b538-58a3299234a0.vep.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF	2.6 MB	0
controlled	42ee3cb8-03e5-4908-af31-a4a424672eb3.vep.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF	1.45 MB	0
controlled	46f09f9b-6e3b-45f3-bb17-afbfdf5c5ecf.vep.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF	1.06 MB	2
controlled	c2e38573-7112-4da8-98a3-e2518685daec.vep.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF	1.74 MB	0
controlled	465451f7-ade8-45fd-992a-14bb3d159d69.vep.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF	5.06 MB	0
controlled	6a4bc808-0c77-462a-bee1-e5b24a35efdf.vep.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF	1.8 MB	1
controlled	0ca4e2e8-0ff6-44ce-991f-5903721d5d5c.vep.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF	1.14 MB	0
controlled	6c14562b-f1c9-4a0b-8416-e3b3611afe8f.vep.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF	1.52 MB	0
controlled	1bbf5142-6bd5-4d57-a099-f51bfa8f4379.vep.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF	1.73 MB	0
controlled	1764cec0-6cb0-468f-924f-728a1ba5e9cd.vep.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF	1.38 MB	0
controlled	8767a4dc-9e81-4b4f-b9ce-22474a94d081.vep.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF	1.33 MB	0
controlled	8c3f0ea1-4181-4045-bf06-3d557dfb0e78.vep.vcf.gz	1	TCGA-LUAD	Simple Nucleotide Variation	VCF	1.83 MB	1

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home | Projects | Exploration | Repository | Quick Search | Login | Cart 670 | GDC Apps

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Projects | Exploration | Repository

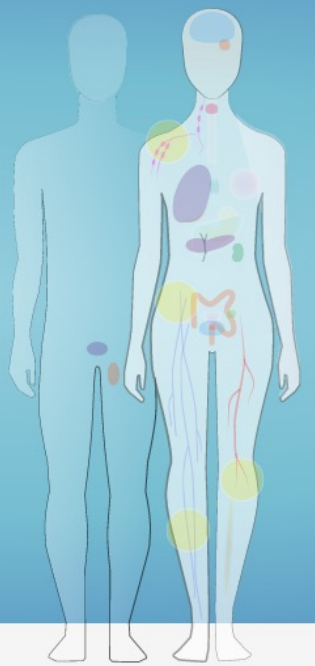
e.g. BRAF, Breast, TCGA-BLCA, c0892598-1f7b-4f23-9cd8-731f797753d5

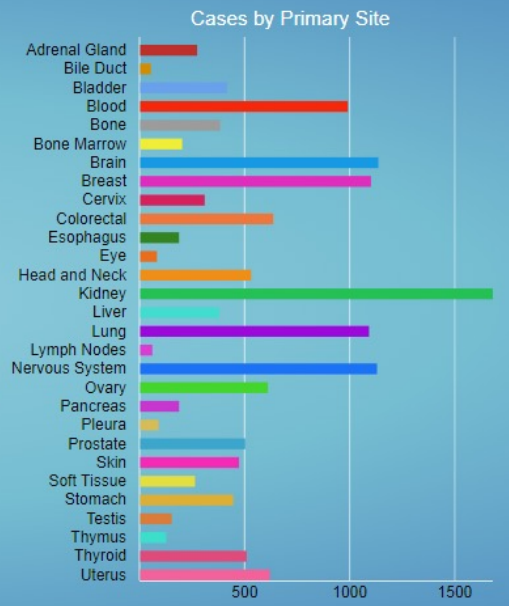
Data Portal Summary

Data Release 7.0 - June 29, 2017

PROJECTS	PRIMARY SITES	CASES
39	29	14,551

FILES	GENES	MUTATIONS
274,724	22,144	3,115,606





Data Portal | Website | API | Data Transfer Tool | Documentation | Data Submission Portal | Legacy Archive

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 | Data Transfer Tool | API | Data Submission Portal | Documentation | Legacy Archive

GDC Data Transfer Tool

The GDC provides a standard client-based mechanism in support of high performance data downloads and submission.

The raw sequence files, typically stored as BAM or FASTQ, make up the bulk of data. The size for a single file can vary greatly depending on the specific analysis; However, some of the whole genome BAM files in The Cancer Genome Atlas (TCGA) reach sizes of 200-300 GB. In such cases, a high performance data download and submission client is essential.

Below are basic instructions and links for downloading the GDC Data Transfer Tool. For additional instructions, please visit the GDC Data Transfer Tool User's Guide.

Downloading the GDC Data Transfer Tool

System Recommendations

The system recommendations for using the GDC Data Transfer Tool are as follows:

- OS: Linux (Ubuntu 14.x or later), OS X (10.9 Mavericks or later), or Windows (7 or later)
- CPU: Eight 64-bit cores, Intel or AMD
- RAM: 8 GiB or more
- Storage: Enterprise-class storage system capable of ≥1 Gb/s (Gigabit per second) write throughput and sufficient free space for BAM files, most of which are in the 50 MB - 40 GB size range, with some reaching sizes of 200-300 GB.

Binary Distributions

Links to the binary distributions for supported platforms are provided below.

Access Data

Data Access Processes and To
Data Access Policies
GDC Community Tools
Documentation for Data Portal
Documentation for Data Transfe
GDC Data Portal
GDC Data Transfer Tool
Launch Data Portal
Obtaining Access to Controlled

Get Started

GDC Data Transfer Tool Guide »

Get the GDC Data Transfe

Downloading the GDC Data Transfer Tool

System Recommendations

The system recommendations for using the GDC Data Transfer Tool are as follows:

- OS: Linux (Ubuntu 14.x or later), OS X (10.9 Mavericks or later), or Windows (7 or later)
- CPU: Eight 64-bit cores, Intel or AMD
- RAM: 8 GiB or more
- Storage: Enterprise-class storage system capable of ≥1 Gb/s (Gigabit per second) write throughput and sufficient free space for BAM files, most of which are in the 50 MB - 40 GB size range, with some reaching sizes of 200-300 GB.

Binary Distributions

Links to the binary distributions for supported platforms are provided below.

- gdc-client_v1.2.0_Windows_x64.zip
- gdc-client_v1.2.0_Ubuntu14.04_x64.zip
- gdc-client_v1.2.0_OSX_x64.zip

If you are a user of CentOS 6 or RedHat Enterprise Release 6 and wish to use the Data Transfer Tool, contact the GDC Help Desk for assistance.

Source Code

Access GitHub Repository

Release Notes

Release Notes are available on the GDC Data Transfer Tool Release Notes page.

Support

Please visit the GDC Help Desk

Token file to download data

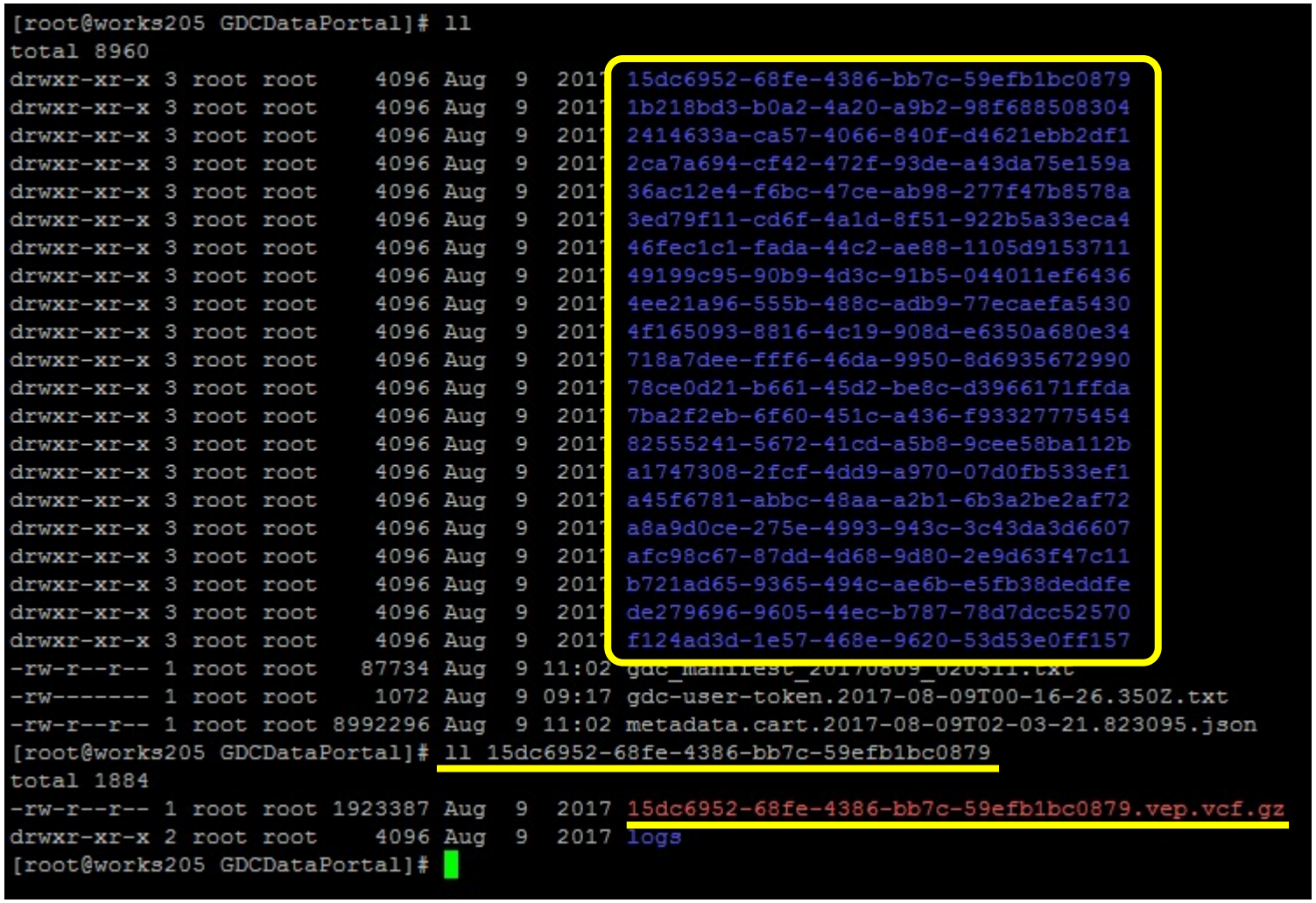
```
E10KXAiOiJKV1QiLCJhbGciOiJFUzI1NiJ9.eyJzdWIiOiJjYjBjMmZmZjFmM2I0OWVlYjY2YWMxMTk4M2VkNzBlYyIsImlhdCI6
TYyNzk1Njk5MywiZXhwIjoxNjMwNTQ4OTkz8jJv4k8Cc3RhY2tfbWV0aG9kcyI6WyJzYW1sMiJdLCJvcGVuc3RhY2tfYXVkaXRfa
RzIjpbImc2DLL1MGM4VDNPYUtUZGE0dXBndGciXSwib3BlbnN0YWNrX2dyb3VwX2lkcyI6W3siaWQiOiJmNzQ5ZjM4ZmM3YjI0OT
jYTVhZjI3MzNjOGFkODJjOSJ9XSwib3BlbnN0YWNrX2lkcF9pZCI6ImVyYV9jb21tb24iLCJvcGVuc3RhY2tfcHJvdG9jb2xfaWQ
OiJzYW1sMiJ9.b2xE1spUjnfIeAaRpi41yGYpcOzXogFo0q0O8D6jHQbkAcvmQhhFGGP0SOYf9FypWu6k_jn0ASG0FoWvNbk-sQ
```

Manifest file to download data

다운로드 — vi gdc_manifest.2021-08-03_WXS.txt — 80×24

```
id,filename,md5,size,state
f69aca82-d6ff-472a-9387-004d70bc69f1,C440.TCGA-MX-A5UG-10A-01D-A31J-08.3.bam,f24b3b69f89fded445c924bfbd0014d6,10539055916,live
42665c2f-125b-4726-9e7f-2f2a197bc777,C440.TCGA-CG-5730-01A-11D-1600-08.6.bam,3d19f40d0a52e789e7a5cd8b9a098bcd,19844923712,live
6fe66f7e-40fd-4cfe-af91-ae9303abd794,C440.TCGA-BR-6801-01A-11D-1882-08.6.bam,543bd85ed815cbfe11e3de71cd812aa8,13280872429,live
4f88da11-35c4-43a4-be61-1da5d56b2ae3,C440.TCGA-BR-4357-01A-01D-1158-08.9.bam,5ca1cb2d9e97792fafc221280633de6d,27216630031,live
3c010c58-b4b6-4492-b325-c65f6a5118db,C440.TCGA-D7-A4YY-01A-11D-A25D-08.6.bam,faf623382d1175b7daada1b11b215e08,13391796191,live
4899aa59-e251-41b2-a5c3-670e48b6e5dd,C440.TCGA-BR-8059-10A-01D-2340-08.6.bam,4d01a6b20a792bb623ea15eef4a76f51,15303841085,live
cfaa25d5-8b06-47e6-8b5c-d06a0ed90bf0,C440.TCGA-FP-8099-10A-01D-2341-08.8.bam,1d58b70d741f155f74c15f974dc83d6d,15633790296,live
440da841-7ebe-401c-856c-c83143e6c5fe,C440.TCGA-VQ-AA6F-10A-01D-A413-08.1.bam,0c92823b477f72b154c2b399fd8ea4a9,7755493851,live
8d7f0133-b203-49b0-bc4a-f619be4b59a1,C440.TCGA-BR-6453-01A-11D-1800-08.3.bam,9864c3d29ca5a447d3a431802cf30755,14758011447,live
297f53aa-3cce-4c9f-ace0-f852b41fa841,C440.TCGA-R5-A7ZF-10A-01D-A34X-08.1.bam,b60ac6e27db98551a5cc2ab06c03705f,8041228291,live
fefe5d1b-5659-45a0-aa78-f6da8256b13d,C440.TCGA-HU-A4GY-11A-11D-A24F-08.3.bam,4b3929492c0963e8da9cc803780ffde3,14649518189,live
```

```
[root@works205 GDCDataPortal]# gdc-client download -t gdc-user-token.2017-08-09T00-16-26.350Z.txt -m gdc_manifest_20170809_020311.txt
Downloading a1747308-2fcf-4dd9-a970-07d0fb533ef1.vep.vcf.gz (UUID a1747308-2fcf-4dd9-a970-07d0fb533ef1):
100% [##################################################################################################################] Time: 0:00:03 464.50 kB/s
Validating checksum...
Downloading a8a9d0ce-275e-4993-943c-3c43da3d6607.vep.vcf.gz (UUID a8a9d0ce-275e-4993-943c-3c43da3d6607):
100% [##################################################################################################################] Time: 0:00:04 533.30 kB/s
Validating checksum...
Downloading 49199c95-90b9-4d3c-91b5-044011ef6436.vep.vcf.gz (UUID 49199c95-90b9-4d3c-91b5-044011ef6436):
100% [##################################################################################################################] Time: 0:00:04 572.24 kB/s
Validating checksum...
Downloading 718a7dee-fff6-46da-9950-8d6935672990.vep.vcf.gz (UUID 718a7dee-fff6-46da-9950-8d6935672990):
100% [##################################################################################################################] Time: 0:00:04 503.79 kB/s
Validating checksum...
Downloading 46fec1c1-fada-44c2-ae88-1105d9153711.vep.vcf.gz (UUID 46fec1c1-fada-44c2-ae88-1105d9153711):
100% [##################################################################################################################] Time: 0:00:03 400.10 kB/s
Validating checksum...
Downloading 2ca7a694-cf42-472f-93de-a43da75e159a.vep.vcf.gz (UUID 2ca7a694-cf42-472f-93de-a43da75e159a):
100% [##################################################################################################################] Time: 0:00:04 611.81 kB/s
Validating checksum...
Downloading 7ba2f2eb-6f60-451c-a436-f93327775454.vep.vcf.gz (UUID 7ba2f2eb-6f60-451c-a436-f93327775454):
100% [##################################################################################################################] Time: 0:00:03 237.20 kB/s
Validating checksum...
Downloading de279696-9605-44ec-b787-78d7dcc52570.vep.vcf.gz (UUID de279696-9605-44ec-b787-78d7dcc52570):
100% [##################################################################################################################] Time: 0:00:03 363.61 kB/s
Validating checksum...
Downloading 1b218bd3-b0a2-4a20-a9b2-98f688508304.vep.vcf.gz (UUID 1b218bd3-b0a2-4a20-a9b2-98f688508304):
100% [##################################################################################################################] Time: 0:00:03 423.61 kB/s
Validating checksum...
Downloading b721ad65-9365-494c-ae6b-e5fb38deddfe.vep.vcf.gz (UUID b721ad65-9365-494c-ae6b-e5fb38deddfe):
100% [##################################################################################################################] Time: 0:00:03 301.38 kB/s
Validating checksum...
```


[root@works205 GDCDataPortal]# ll
total 8960
drwxr-xr-x 3 root root 4096 Aug 9 2017 15dc6952-68fe-4386-bb7c-59efb1bc0879
drwxr-xr-x 3 root root 4096 Aug 9 2017 1b218bd3-b0a2-4a20-a9b2-98f688508304
drwxr-xr-x 3 root root 4096 Aug 9 2017 2414633a-ca57-4066-840f-d4621ebb2df1
drwxr-xr-x 3 root root 4096 Aug 9 2017 2ca7a694-cf42-472f-93de-a43da75e159a
drwxr-xr-x 3 root root 4096 Aug 9 2017 36ac12e4-f6bc-47ce-ab98-277f47b8578a
drwxr-xr-x 3 root root 4096 Aug 9 2017 3ed79f11-cd6f-4a1d-8f51-922b5a33eca4
drwxr-xr-x 3 root root 4096 Aug 9 2017 46fec1c1-fada-44c2-ae88-1105d9153711
drwxr-xr-x 3 root root 4096 Aug 9 2017 49199c95-90b9-4d3c-91b5-044011ef6436
drwxr-xr-x 3 root root 4096 Aug 9 2017 4ee21a96-555b-488c-adb9-77ecaefa5430
drwxr-xr-x 3 root root 4096 Aug 9 2017 4f165093-8816-4c19-908d-e6350a680e34
drwxr-xr-x 3 root root 4096 Aug 9 2017 718a7dee-fff6-46da-9950-8d6935672990
drwxr-xr-x 3 root root 4096 Aug 9 2017 78ce0d21-b661-45d2-be8c-d3966171ffda
drwxr-xr-x 3 root root 4096 Aug 9 2017 7ba2f2eb-6f60-451c-a436-f93327775454
drwxr-xr-x 3 root root 4096 Aug 9 2017 82555241-5672-41cd-a5b8-9cee58ba112b
drwxr-xr-x 3 root root 4096 Aug 9 2017 a1747308-2fcf-4dd9-a970-07d0fb533ef1
drwxr-xr-x 3 root root 4096 Aug 9 2017 a45f6781-abbc-48aa-a2b1-6b3a2be2af72
drwxr-xr-x 3 root root 4096 Aug 9 2017 a8a9d0ce-275e-4993-943c-3c43da3d6607
drwxr-xr-x 3 root root 4096 Aug 9 2017 afc98c67-87dd-4d68-9d80-2e9d63f47c11
drwxr-xr-x 3 root root 4096 Aug 9 2017 b721ad65-9365-494c-ae6b-e5fb38deddfe
drwxr-xr-x 3 root root 4096 Aug 9 2017 de279696-9605-44ec-b787-78d7dcc52570
drwxr-xr-x 3 root root 4096 Aug 9 2017 f124ad3d-1e57-468e-9620-53d53e0ff157
-rw-r--r-- 1 root root 87734 Aug 9 11:02 gdc_manifest_20170809_020311.txt
-rw------- 1 root root 1072 Aug 9 09:17 gdc-user-token.2017-08-09T00-16-26.350Z.txt
-rw-r--r-- 1 root root 8992296 Aug 9 11:02 metadata.cart.2017-08-09T02-03-21.823095.json
[root@works205 GDCDataPortal]# ll 15dc6952-68fe-4386-bb7c-59efb1bc0879
total 1884
-rw-r--r-- 1 root root 1923387 Aug 9 2017 15dc6952-68fe-4386-bb7c-59efb1bc0879.vep.vcf.gz
drwxr-xr-x 2 root root 4096 Aug 9 2017 logs
[root@works205 GDCDataPortal]#

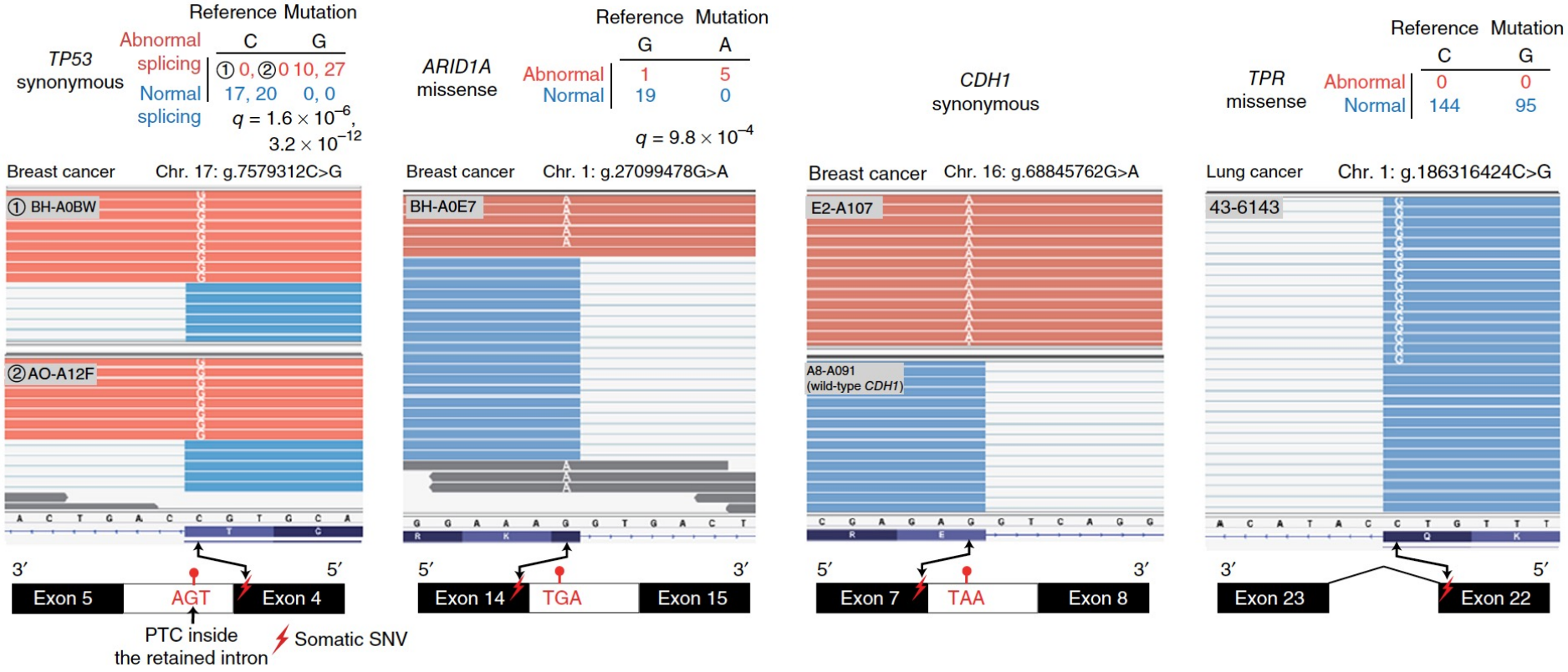
Tumor suppressor genes with LBMEs causing intron retention

Tumor type	Confirmed intron retention		Predicted intron retention		Samples with mutations (% truncating mutations, % LBEMs)
	Silent	Missense	Silent	Missense	
Breast invasive carcinoma (n = 503)	*MLL3*[a] *TP53* (2)	*ARID1A*	*CDH1*		*ARID1A* (2%, 12%) *CDH1* (6%, 3%) *MLL3* (4%, 5%) *TP53* (14%, 3%)
Colorectal carcinoma (n = 217)	*TP53*			*TP53* *VPS13A*	*TP53* (18%, 5%)
Kidney renal clear cell carcinoma (n = 400)	*SETD2*	*CDKN2A* *SPSB3* (2)[b] *VHL* (2)		*FAT1*	*VHL* (30%, 2%) *SETD2* (8%, 3%)
Lung squamous cell carcinoma (n = 178)	*ARID1A* *TP53* (3)	*CDKN2A* (exon 2) *CIC* *MLL2*, *MLL2*[a] *PSMC3* *TP53* (2) *WRN*	*TP53* (2)	*CDKN2A* (exon 1) *LRP1B* *PTCH1*	*ARID1A* (4%, 14%) *CDKN2A* (8%, 14%) *LRP1B* (5%, 12%) *MLL2* (12%, 10%) *TP53* (28%, 14%)
Ovarian serous cystadenocarcinoma (n = 273)	*TP53*				*TP53* (34%, 1%)
Uterine corpus endometrial carcinoma (n = 241)		*BAP1* *PTCH1* *KDM5C*		*PTEN*	*PTCH1* (2%, 25%) *PTEN* (47%, 1%)

Identification of SNVs disrupting splicing

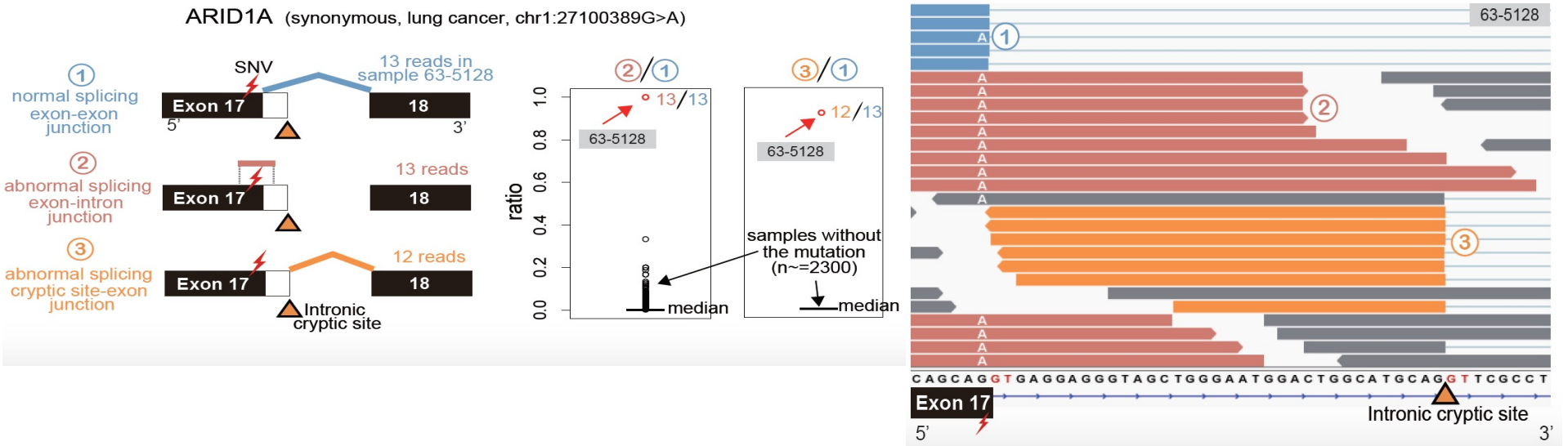
- **Allele-specific splicing analysis** of LBEMs causing intron retention
  - Statistical significance test using Fisher's exact test
    - Normal Splicing: Wild Type
    - Abnormal Splicing: Mutant Allele

Ref Ab.	Alt Ab.
Ref Nor.	Alt Nor.



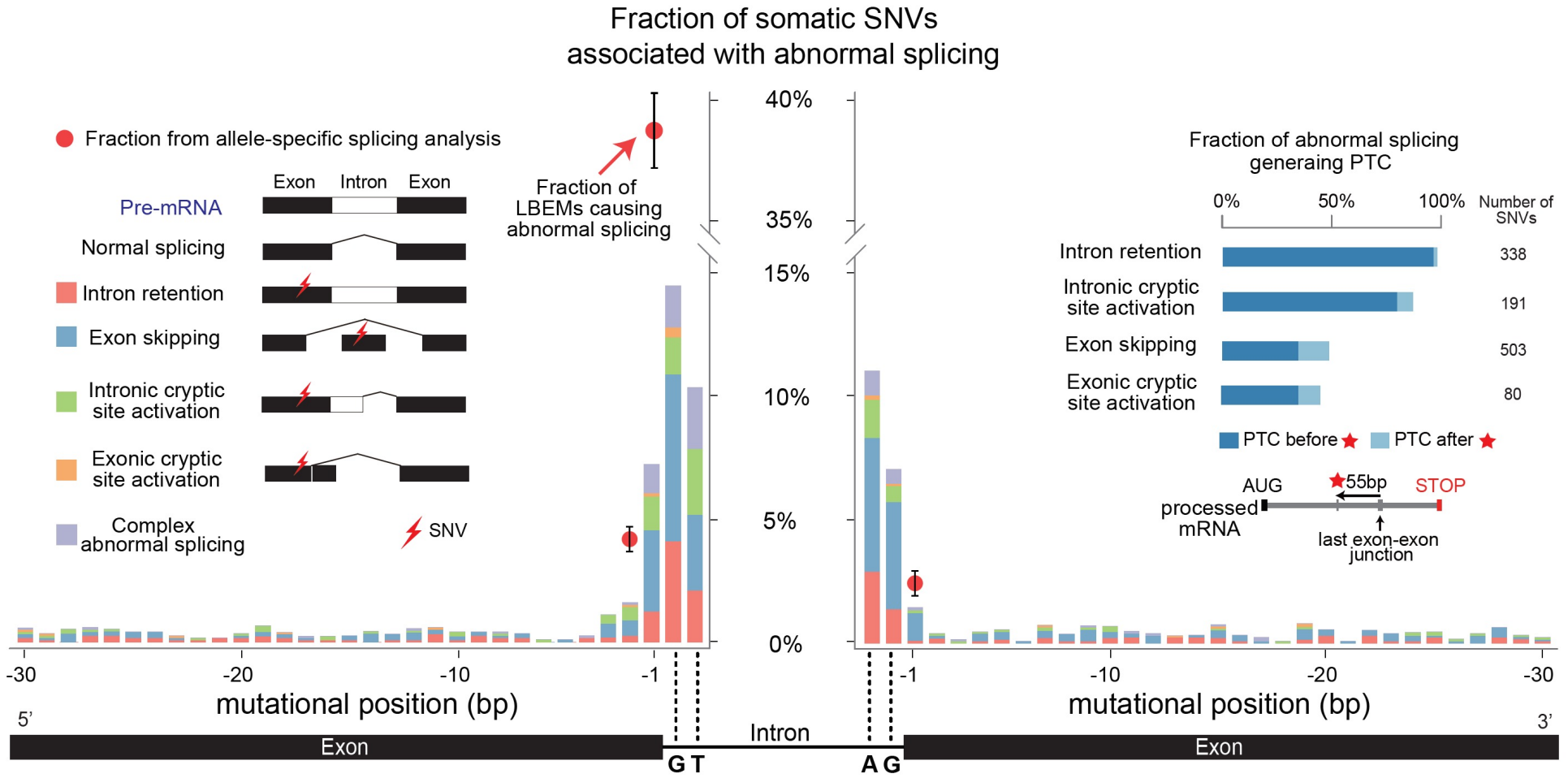
Identification of SNVs disrupting splicing

- Ratio-based splicing analysis
  - Complex abnormal splicing
    - Somatic SNVs are associated with different types of abnormal splicing
      - Intron retention, exon skipping and intronic and exonic cryptic site activation
  - The ratio of abnormally spliced reads / normally spliced reads



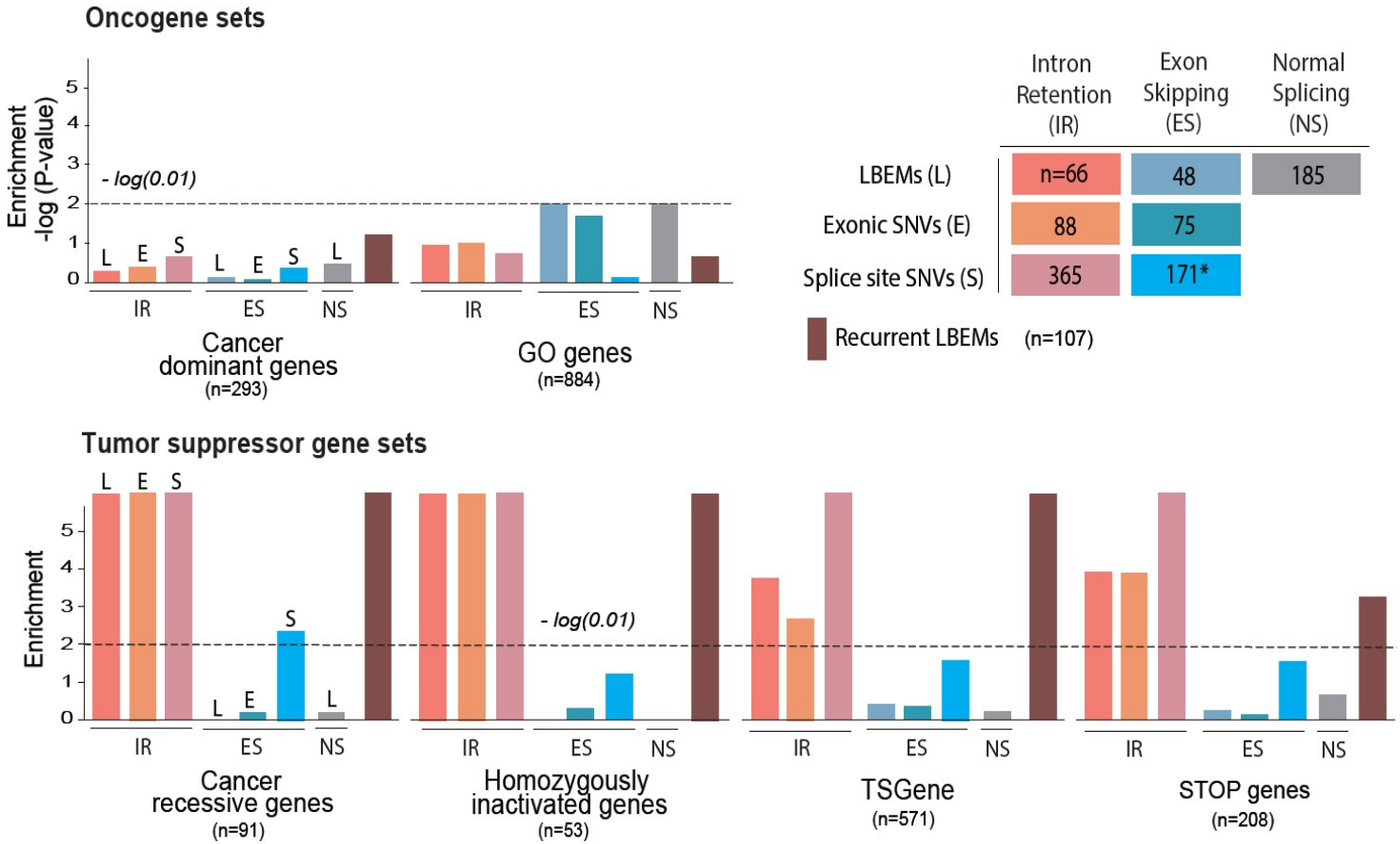
Frequently altered splicing by last-base exonic mutations (LBEM)s

- Positional association of somatic SNVs with abnormal splicing



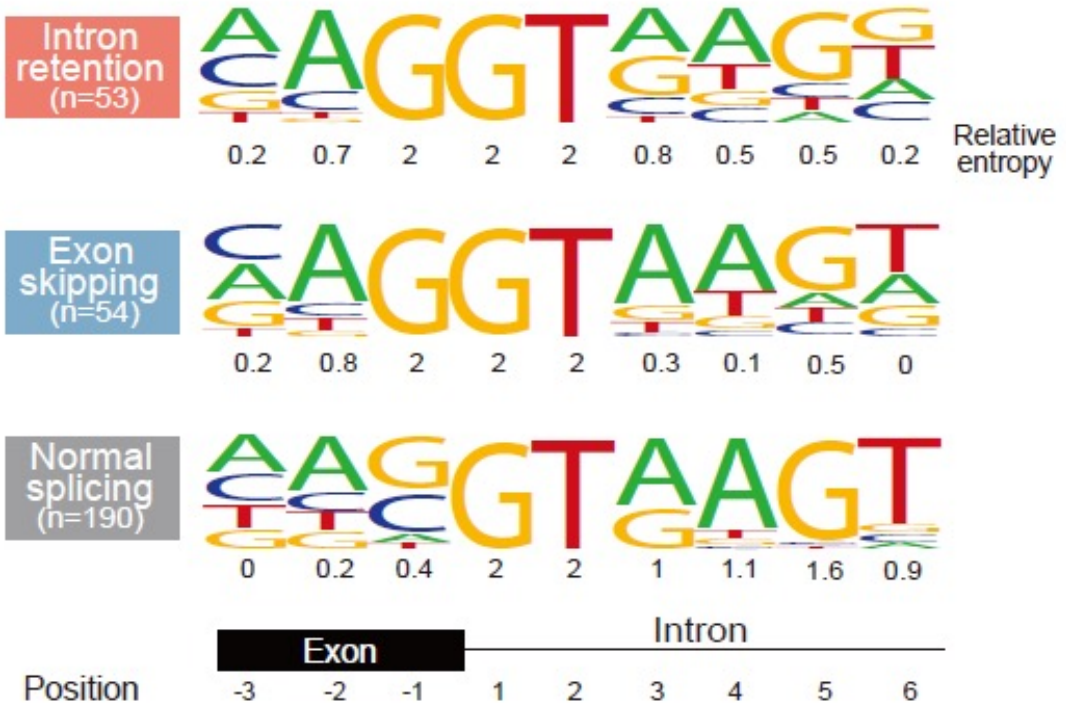
Enrichment of intron retention-causing SNVs in TSGs

- 107 LBEMs
  - They were occurred in two or more patients, with 23 in known TSGs
  - They also showed significant enrichment in the TSG sets but not in the oncogene sets



Characterization of discriminative features for splicing aberration and construction of prediction models

- Distinct LBEM sequence motifs
  - LBEMs in different splicing groups show distinct sequence motifs near-intron junctions



Summary

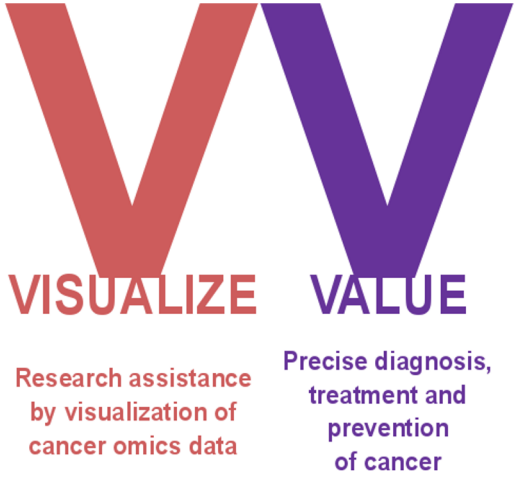
- The first comprehensive characterization of somatic mutations that disrupt pre-mRNA splicing in cancer

- **Intron retention is a frequent mechanism of tumor suppressor inactivation**, with loss of function resulting from NMD or truncated proteins

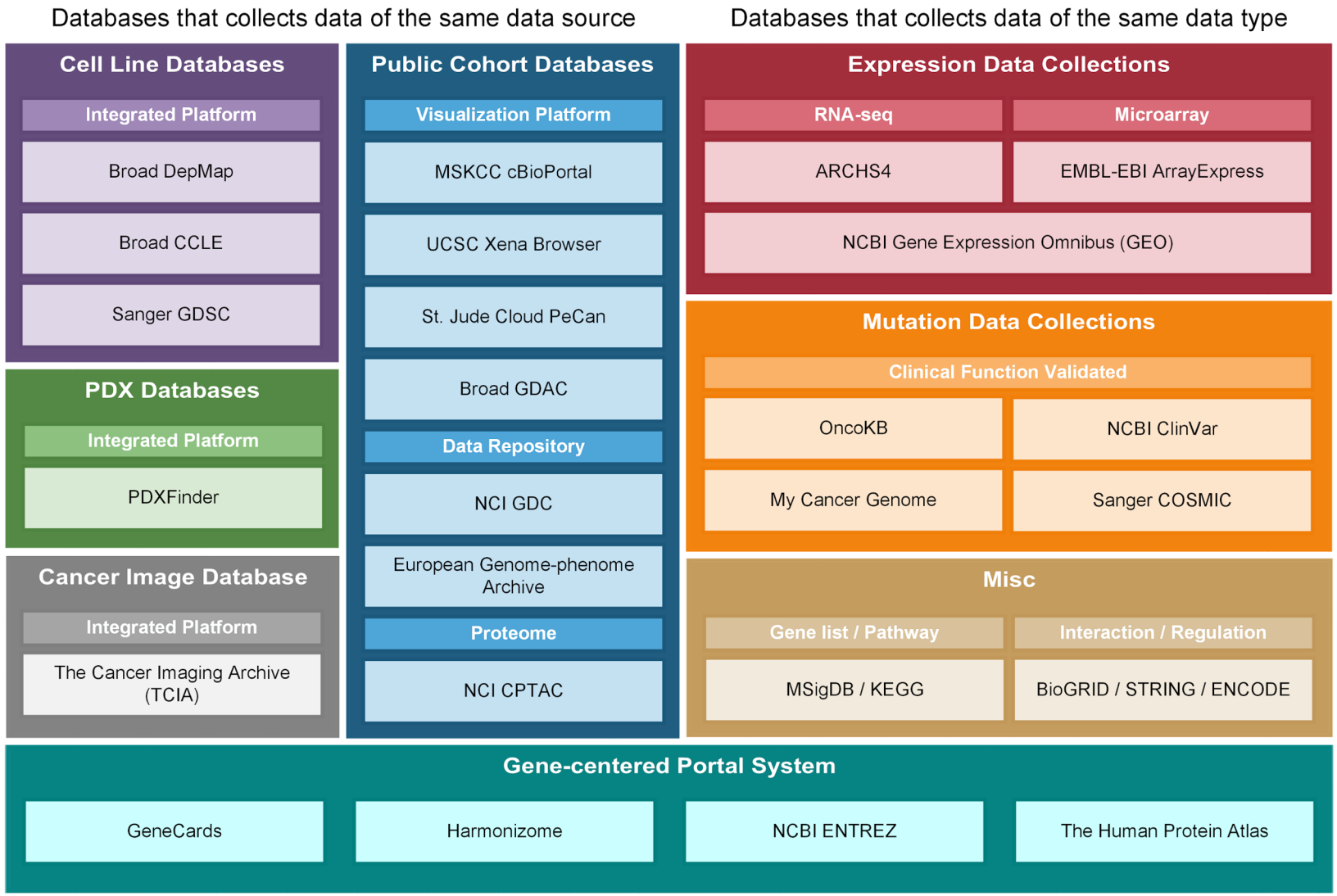
  - The importance of **mutations in the last base of exon** in splicing regulation
  - Enrichment of intron retention caused by these mutations in tumor suppressors such as TP53, ARID1A and VHL
  - Distinct genomic context for these mutations

Big data is defined by these specific attributes;
four **V**s.

Two more **V**s are necessary for **cancer genomic data**



Genomic databases in cancer research



= 100Tb

= 1 million $

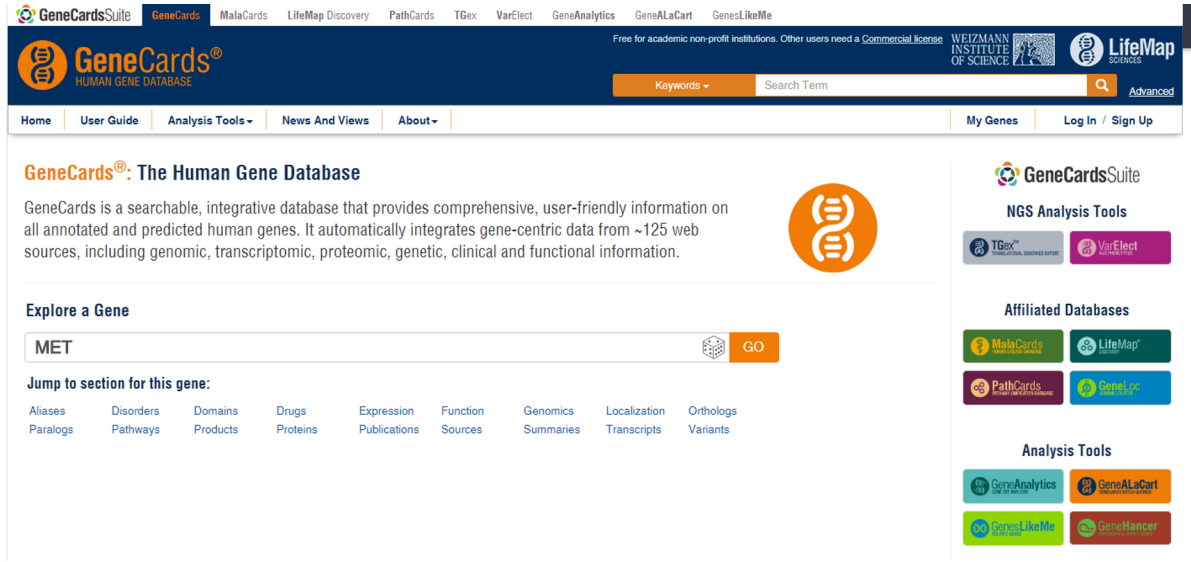
	개발인원 (PI수)	데이터 보유 현황	개발금액
GDC	총 20명 이상 NCI: 4명 Chicago 6명 OICR: 4명 NCI / University of Chicago / OICR	4.1페타 바이트	약 220억
cBioPortal	총 53명 이상 MSKCC 15명 Dana-Farber 8명 MSKCC / Dana-Farber cancer institute	약 70,000개 시퀀싱 데이터 …	약 88억
CPTAC	총 23명 이상 각 7, 10, 6명 Proteome Characterization Centers / Proteogenomic Data Analysis Centers / Proteogenomic Translational Research Centers	단백질: 약 11테라 약 2만명의 사용자가 300테라 다운로드 10.9Tb	약 55억

Trending analytical approaches

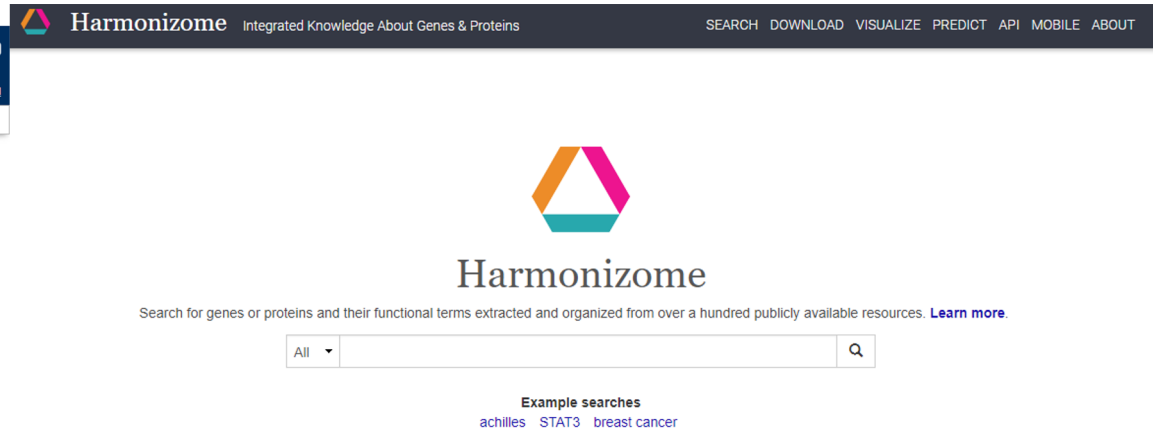
Keywords:

- **Connectivity** between heterogeneous data types
  - Harmonizome, Genecards, NCBI Entrez, cBioPortal
  - Links between databases became necessary
- **Systematic** experiments
  - DepMap; CRISPR / RNAi / drug sensitivity
- **Standardization** for machine learning & **deep-learning**
  - ARCHS4: RNA-seq data collection in one pipeline
  - ENCODE: DNA binding site identification. Used in deep-learning based prediction (DeepBind)

Connectivity between heterogeneous data types



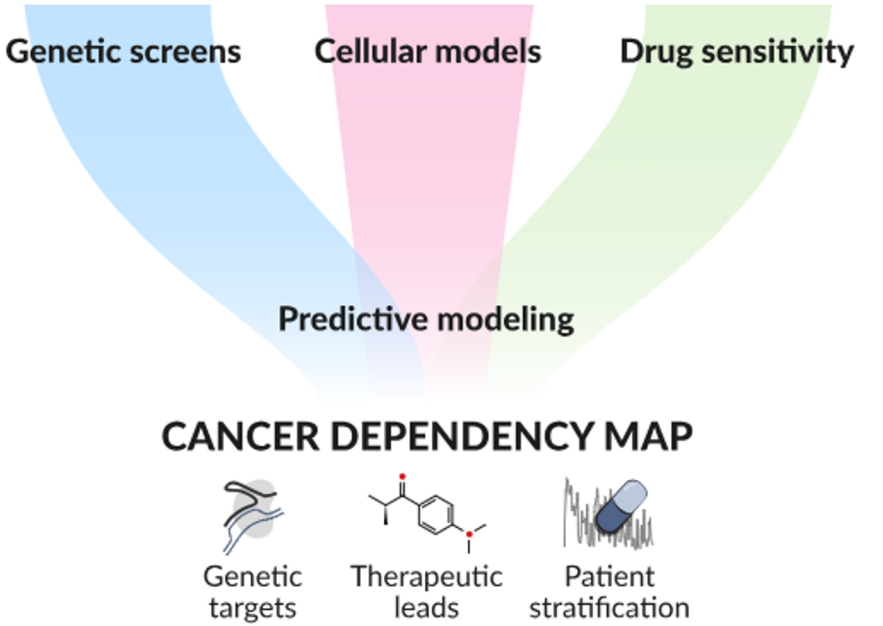
Stelzer, Gil, et al., *Current protocols in bioinformatics* 2016.



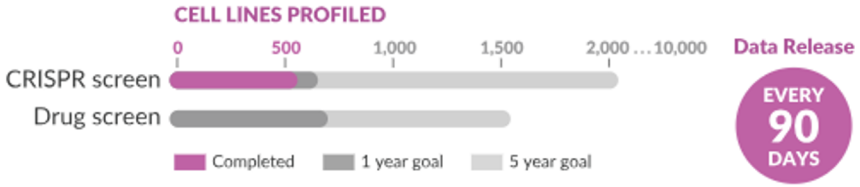
http://amp.pharm.mssm.edu/Harmonizome/

Rouillard, Andrew D., et al., *Database* 2016.

Systematic experiments

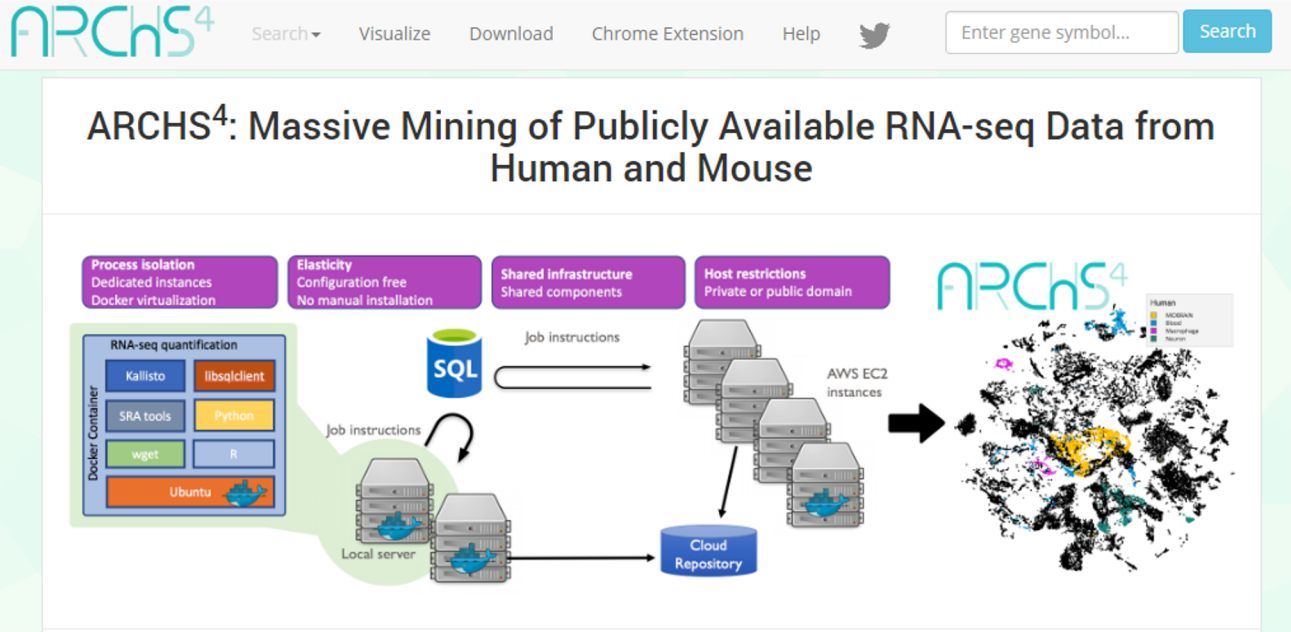


To date DepMap has profiled more than 500 cell lines. Over the next several years we will greatly expand the diversity of cell lines profiled for genetic vulnerabilities with quarterly data release. Additionally, limited drug sensitivity data are available.



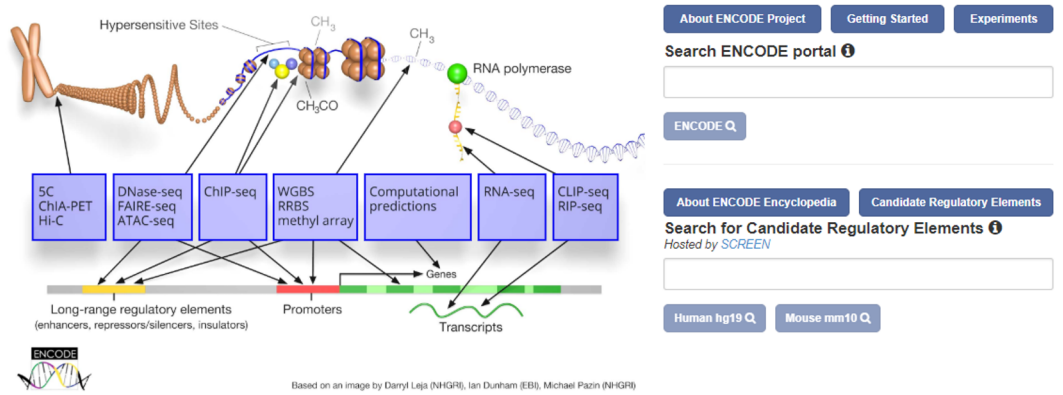
DepMap; Tsherniak, Aviad, et al., *Cell* 2017.

Standardization for machine learning & deep-learning



Lachmann, Alexander, et al., *Nature communications* 2018.

ENCODE: Encyclopedia of DNA Elements



ENCODE Project Consortium. *Science* 2004.

_computational BIOLOGY

ANALYSIS

Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

Babak Alipanahi[1,2,6], Andrew Delong[1,6], Matthew T Weirauch[3–5] & Brendan J Frey[1–3]

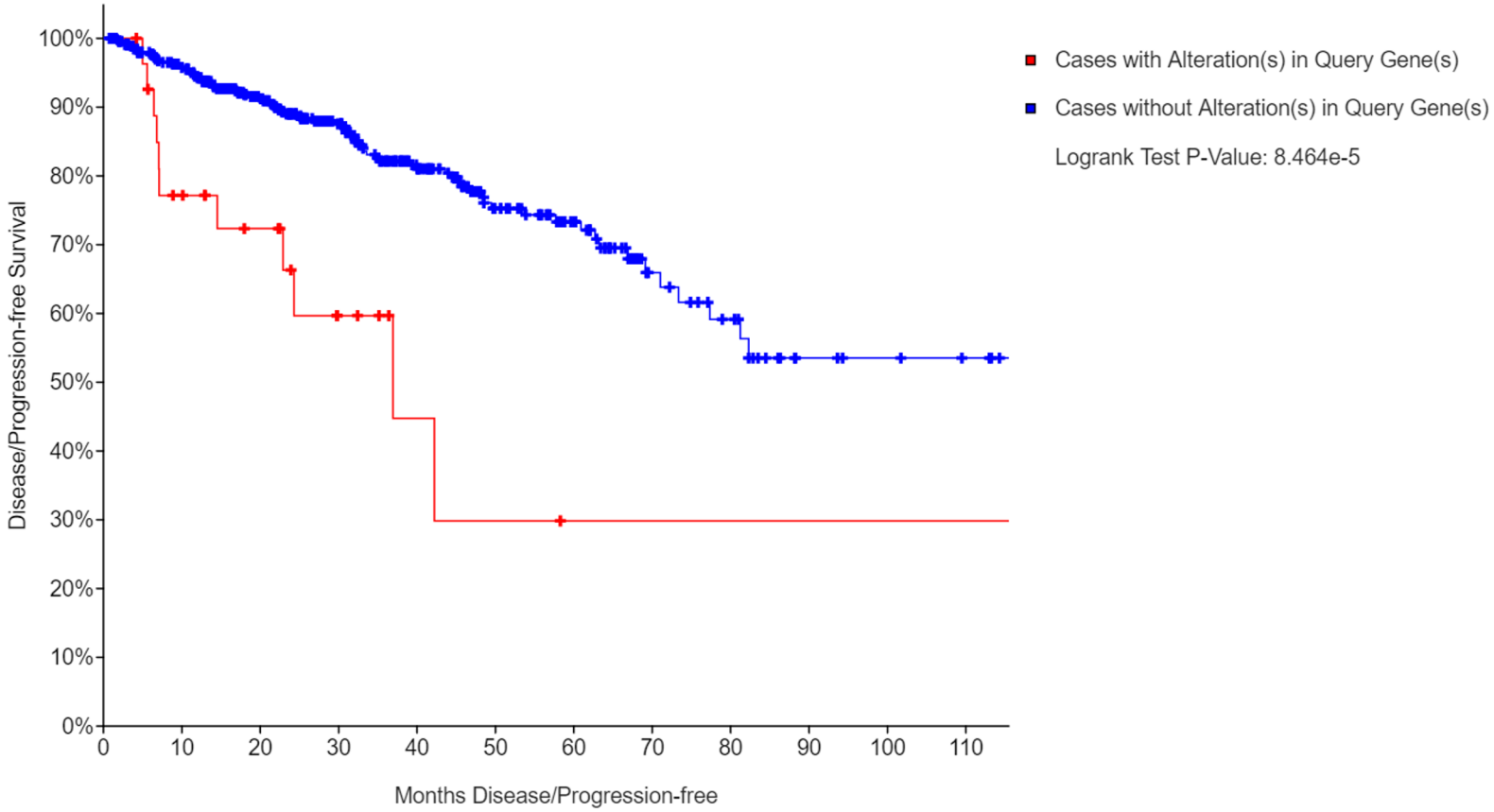
Alipanahi, Babak, et al., *Nature biotechnology* 2015

Usage of genomic databases in cancer research - Scenario #1

Scenario #1.

With cBioPortal,

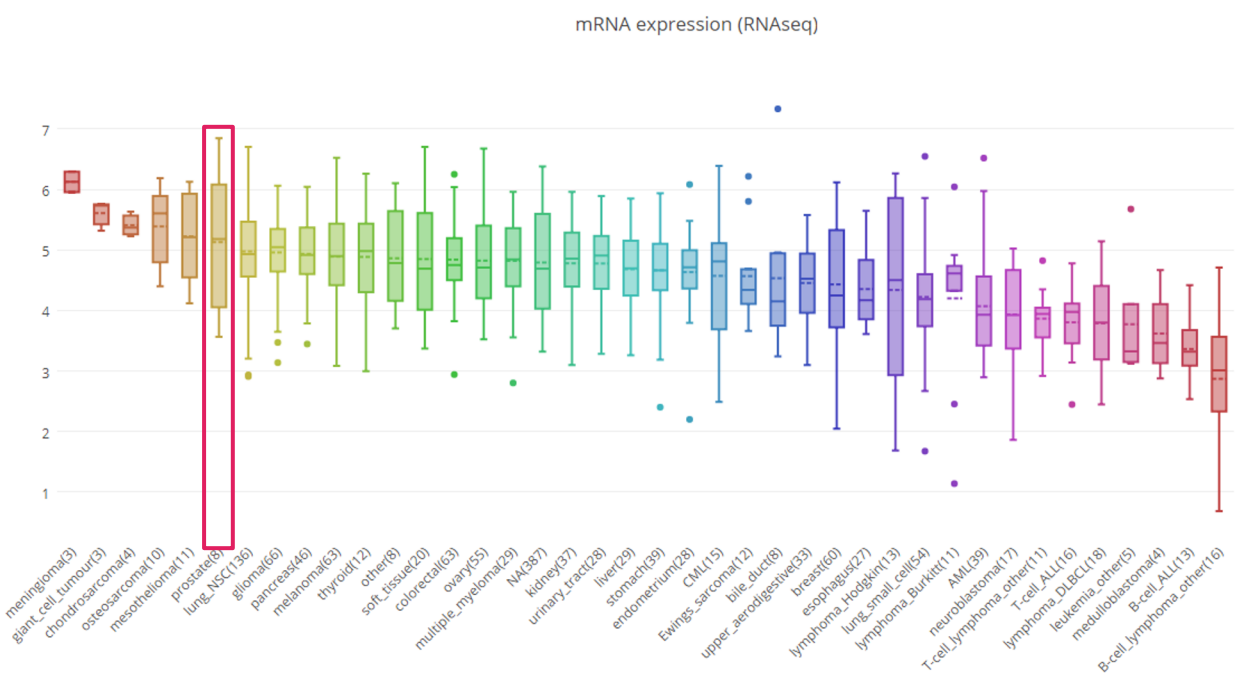
High expression of Gene X identified as a significant disease progression marker in TCGA PRAD cohort



Gao, Jianjiong, et al., *Sci. Signal.* 2013.

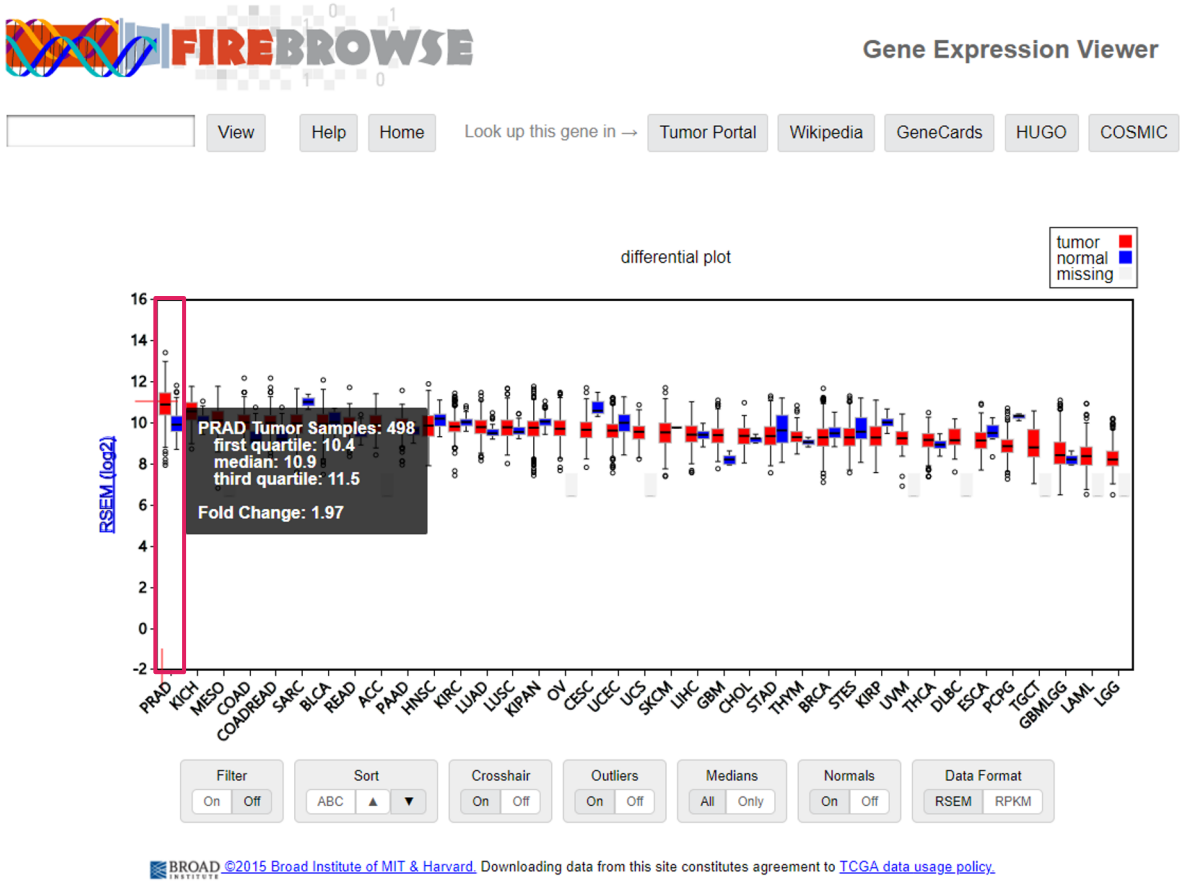
Usage of genomic databases in cancer research - Scenario #1

CCLE; Barretina, Jordi, et al., *Nature* 2012.

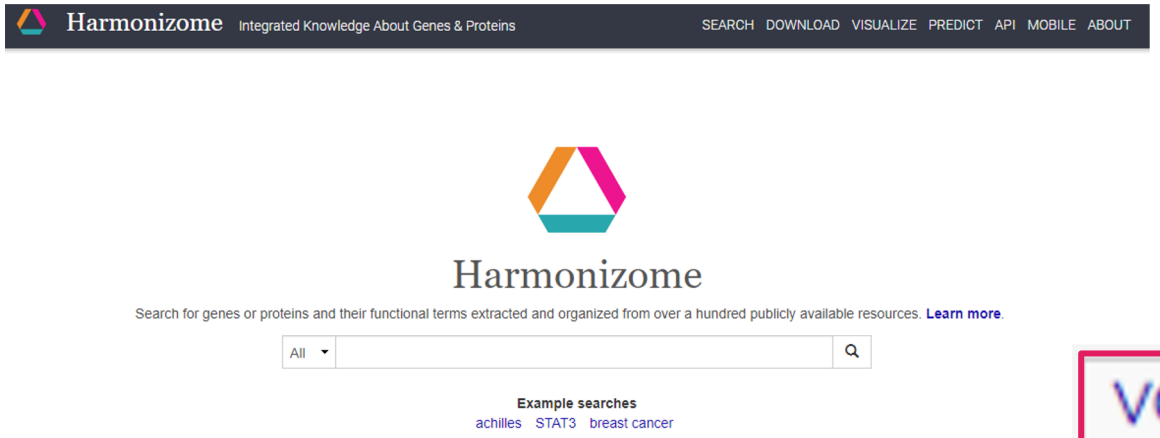


We confirmed expression of gene X at both cell line (**CCLE**) and patient cohort level (**Broad GDAC**). Expression level of gene X in prostate cancer was higher than in other cancers.

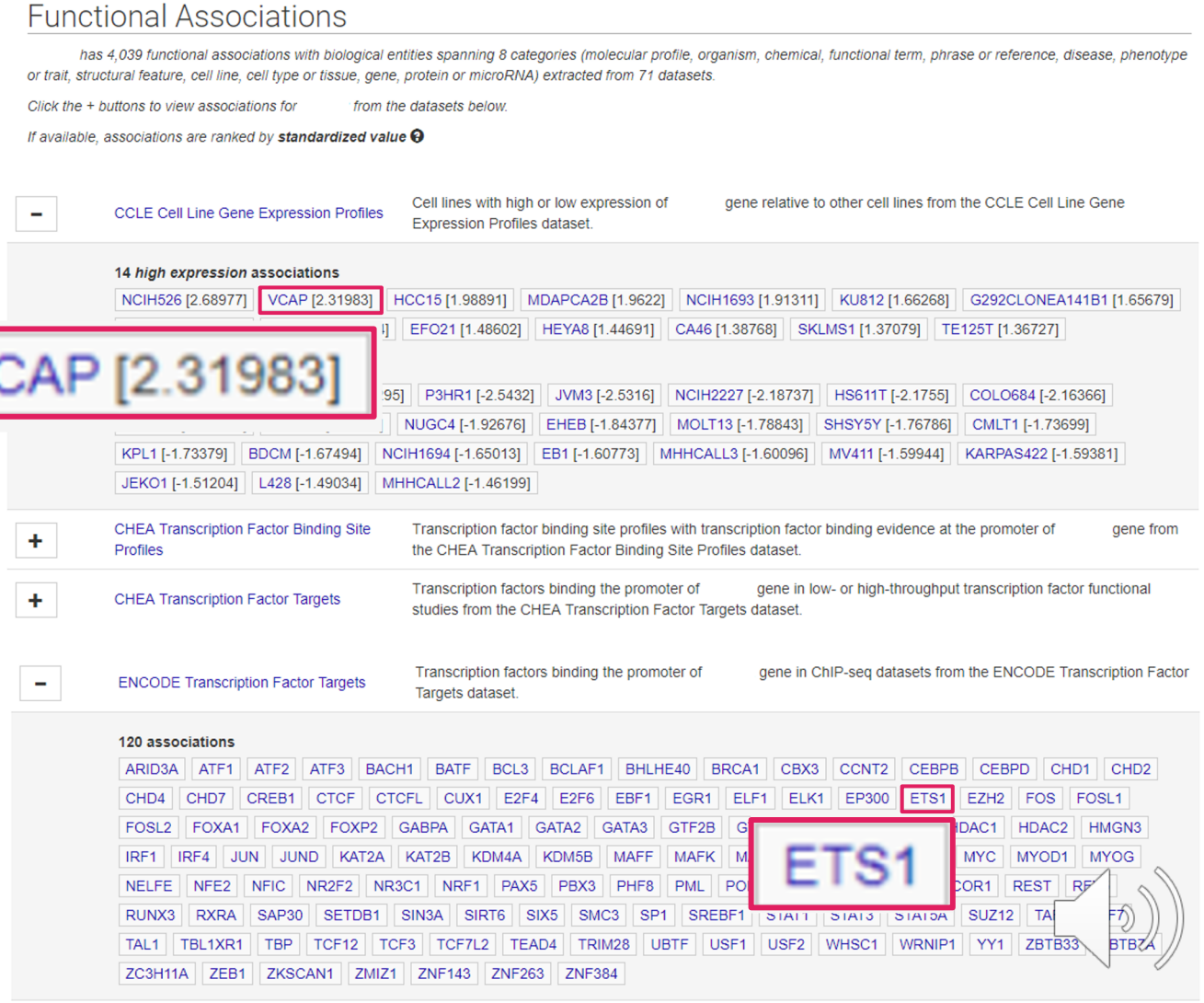
http://firebrowse.org



Usage of genomic databases in cancer research - Scenario #1

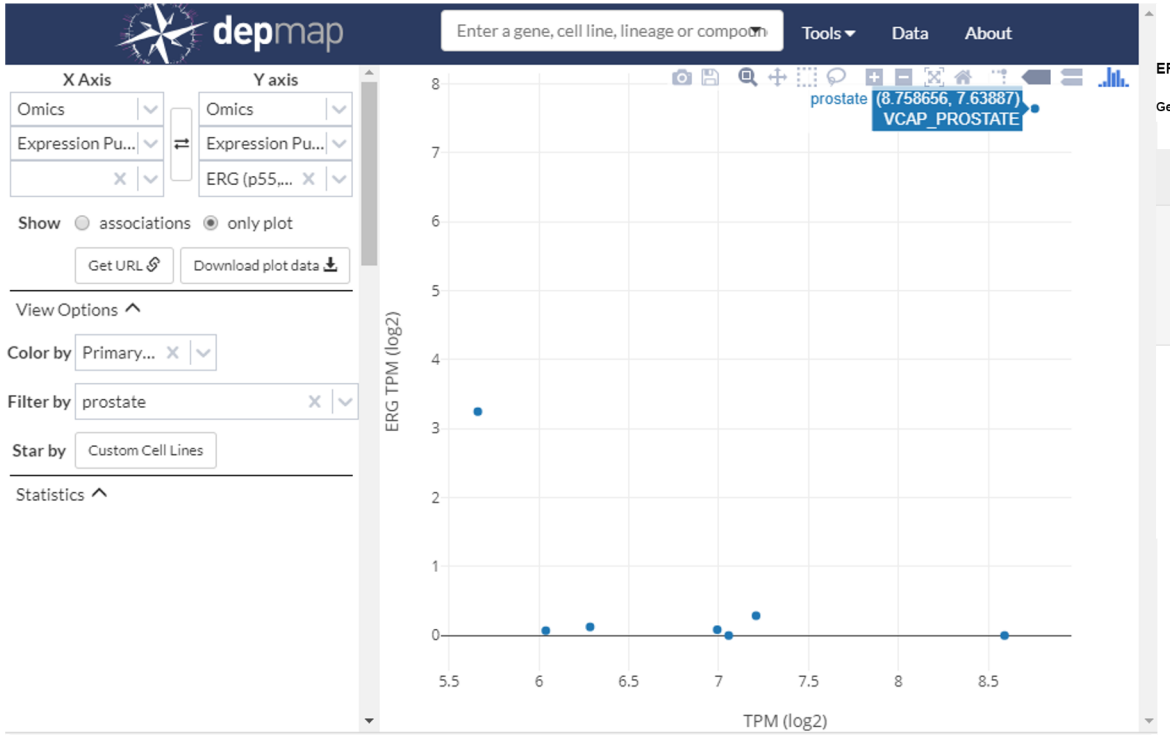


http://amp.pharm.mssm.edu/Harmonizome/



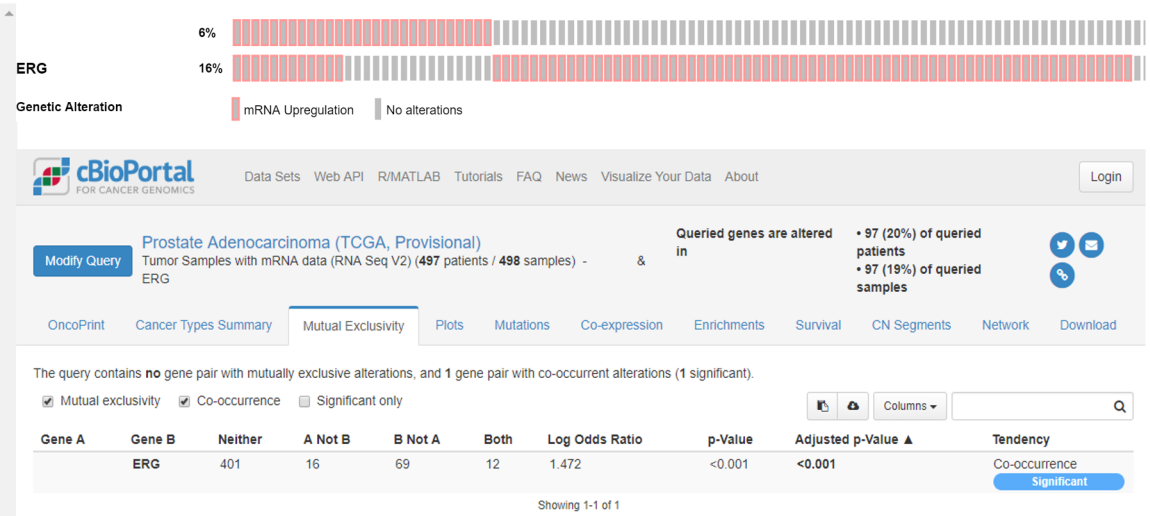
In Harmonizome, we identified high expression of gene X in **VCAP cell line**, which is well known as a TMPRSS2-ERG fusion positive model of prostate cancers.
Also, this gene was predicted to be a **target of ETS1**, which is an ETS-family transcription factor like ERG.

Usage of genomic databases in cancer research - Scenario #1



TMPRSS2--ERG	TMPRSS2	chr21:42880008:-	1	ERG	chr21:39817544:-	0.97	522	3	3.8331	0.022	117
TMPRSS2--ERG	TMPRSS2	chr21:42879877:-	0.6	ERG	chr21:39817544:-	0.92	124	60	0.9106	0.4406	117
TMPRSS2--ERG	TMPRSS2	chr21:42879877:-	0.37	ERG	chr21:39846044:-	0.91	7	65	0.0514	0.4773	61
TMPRSS2--ERG	TMPRSS2	chr21:42878372:-	0.33	ERG	chr21:39817544:-	0.81	2	63	0.0147	0.4626	56
TMPRSS2--ERG	TMPRSS2	chr21:42880008:-	1	ERG	chr21:39846044:-	0.85	36	3	0.2644	0.022	61
TMPRSS2--ERG	TMPRSS2	chr21:42879877:-	0.15	ERG	chr21:39795483:-	0.02	1	21	0.0073	0.1542	59

https://depmap.org/portal/



Among 8 prostate cancer cell lines,
VCaP cell line was the one with highest gene X and ERG expression level.

Significant co-occurrence of high expression level of ERG and gene X was confirmed in cBioPortal.

Usage of genomic databases in cancer research - Scenario #1

NEOPLASIA
www.neoplasia.com

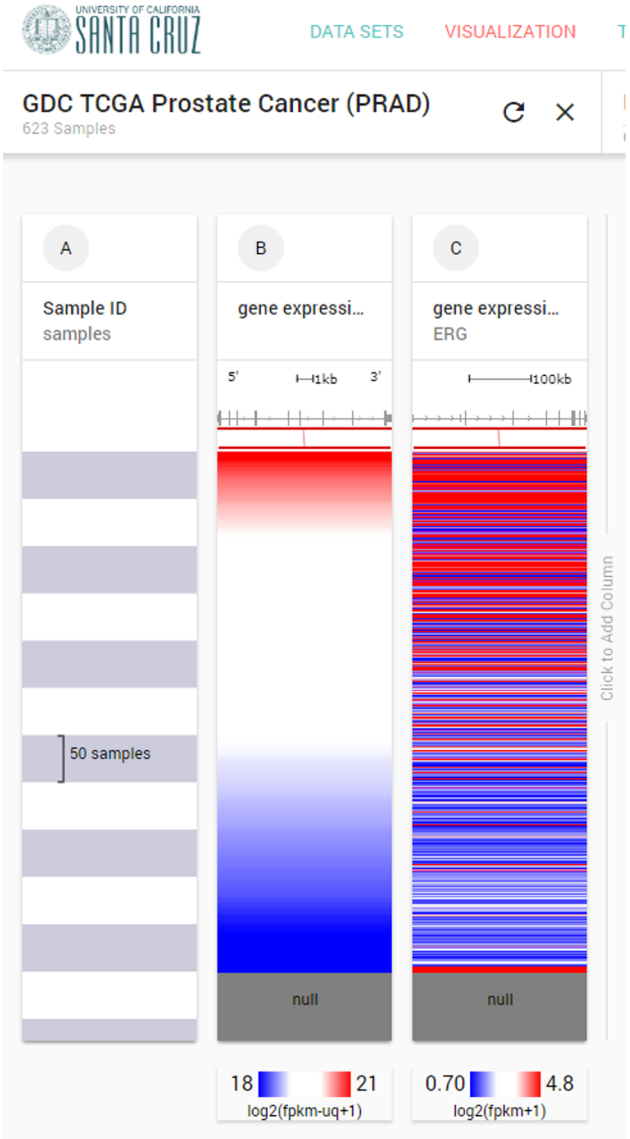
Volume 14 Number 7 July 2012 pp. 600–611 600

Molecular Subtyping of Primary Prostate Cancer Reveals Specific and Shared Target Genes of Different ETS Rearrangements[1,2]

Paula Paulo*,†,‡,§, Franclim R. Ribeiro*,†,‡,§, Joana Santos*,†, Diana Mesquita*,†, Mafalda Almeida*,¶, João D. Barros-Silva*,†, Harri Itkonen#, Rui Henrique¶,**,††, Carmen Jerónimo*,¶,††, Anita Sveen‡,§, Ian G. Mills‡,#,‡‡, Rolf I. Skotheim‡,§, Ragnhild A. Lothe‡,§ and Manuel R. Teixeira*,†,§,††

*Department of Genetics, Portuguese Oncology Institute, Porto, Portugal; †Group of Cancer Genetics, Research Centre of the Portuguese Oncology Institute, Porto, Portugal; ‡Department of Cancer Prevention, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway; §Centre for Cancer Biomedicine, University of Oslo, Oslo, Norway; ¶Group of Cancer Epigenetics, Research Centre of the Portuguese Oncology Institute, Porto, Portugal; #Centre for Molecular Medicine Norway, Nordic European Molecular Biology Laboratory Partnership, Biotechnology Centre, University of Oslo, Oslo, Norway; **Department of Pathology, Portuguese Oncology Institute, Porto, Portugal; ††Institute of Biomedical Sciences Abel Salazar, University of Porto, Porto, Portugal; ‡‡Department of Urology, Oslo University Hospital, Oslo, Norway

Paulo, Paula, et al., *Neoplasia* 2012.



Co-expression of Gene X and ERG was re-confirmed with **UCSC Xena Browser**. This result is consistent with the results from the above study which showed that gene X is a target of ETS-family transcription factors.

경청해 주셔서 고맙습니다.