

KSBi-BIML 2021

Bioinformatics & Machine Learning (BIML)
Workshop for Life Scientists

생물정보학 & 머신러닝 워크샵(온라인)

Cancer Genome Study in Big Data Era

홍동완

Bioinformatics & Machine Learning for Life Scientists

BIML-2021

안녕하십니까?

한국생명정보학회의 동계 워크샵인 BIML-2021을 2월 15부터 2월 19일까지 개최합니다. 생명정보학 분야의 융합이론 보급과 실무역량 강화를 위해 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하였으며 올해로 7차를 맞이하게 되었습니다. 유례가 없는 코로나 대유행으로 인해 올해의 BIML 워크숍은 온라인으로 준비했습니다. 생생한 현장 강의에서만 느낄 수 있는 강의자와 수강생 사이의 상호교감을 가질 수 없다는 단점이 있지만, 온라인 강의의 여러 장점을 살려서 최근 생명정보학에서 주목받고 있는 거의 모든 분야를 망라한 강의를 준비했습니다. 또한 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다.

BIML 워크샵은 전통적으로 크게 생명정보학과 AI, 두 개의 분야로 구성되어오고 있으며 올해 역시 유사한 방식을 채택했습니다. AI 분야는 Probabilistic Modeling, Dimensionality Reduction, SVM 등과 같은 전통적인 Machine Learning부터 Deep Learning을 이용한 신약개발 및 유전체 연구까지 다양한 내용을 다루고 있습니다. 생명정보학 분야로는, Proteomics, Chemoinformatics, Single Cell Genomics, Cancer Genomics, Network Biology, 3D Epigenomics, RNA Biology, Microbiome 등 거의 모든 분야가 포함되어 있습니다. 연사들은 각 분야 최고의 전문가들이라 자부합니다.

이번 BIML-2021을 준비하기까지 너무나 많은 수고를 해주신 BIML-2021 운영위원회의 김태민 교수님, 류성호 교수님, 남진우 교수님, 백대현 교수님께 커다란 감사를 드립니다. 또한 재정적 도움을 주신, 김선 교수님 (AI-based Drug Discovery), 류성호 교수님, 남진우 교수님께 감사를 표시하고 싶습니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 강의자료를 만드는데 노력하셨을 뿐만 아니라 실시간 온라인 Q&A 세션까지 참여해 수고해 주시는 모든 연사분들께 깊이 감사드립니다.

2021년 2월

한국생명정보학회장 김동섭

The Cancer genome study in big data era

Prof. Dongwan Hong (dwhong@catholic.ac.kr)

Catholic University of Korea, College of Medicine

Cancer genome study plays an important role in precision medicine as well as big data. To make sense of cancer diseases, various types of mutations such as somatic point mutations, copy number alterations and structural variants and so on, should be investigated together in multi-layered level. The study results from international big cancer projects (TCGC, ICGC PCAWG and CPTAC etc) have been published and opened continuously to cancer clinicians and researchers to popular databases including GDC, PDC and ICGC data portal. Currently, we have reached easily to public cancer databases and can examine somatic mutations from pan-cancers. Through the cancer studies, we have found that somatic genome mutations occur due to combinations of various intrinsic/extrinsic mutational processes and DNA repair mechanisms. In this lecture, 1) we will survey the history of cancer and understand origin, progression, recurrence and metastasis of cancer. 2) In addition, we will find out the various ngs technologies used in cancer study. 3) Finally, cancer databases and analysis tools will be introduced and the tutorials of them provided with some instances on lecture notes.

KSBi-BIML 2021

Cancer Genome Study in Big Data Era

Dongwan Hong, Ph.D.
(dwhong@catholic.ac.kr)

Catholic University of Korea, College of Medicine

본 강의 자료는 한국생명정보학회가 주관하는 KSBi-BIML 2021 워크숍 온라인 수업을 목적으로 제작된것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다. 수업 목적으로 배포 및 전송 받은 경우에도 이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없습니다.

만약 이러한 사항을 위반할 경우 발생하는 모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고합니다.

Overview

- **Part I: Cancer Big Database**
 - History of cancer genome study
 - Cbioportal, Xena browser and GDC data portal
 - Study example of cancer big database
- **Part II: Cancer Genome Analysis**
 - Mutational Signature Analysis: Mutalisk and FIREVAT
 - Data driven subtyping of cancer
- **Part III: Artificial Intelligence Study using Cancer Big Data**



PART I

**What is the title of the paper
you are preparing for publication ?**

sciendo

Data and Information Management, 2018; 2(2): 83–90



Research Article

Open Access

Tolga Yuret*

**Citation performance of publications grouped by
keywords, titles, and abstracts**

Table B1: Word performance metrics

Performance Metric	Description
YK	Average annual performance
YT	Year-adjusted performance
YA	Year-adjusted abstract performance

Table B2: Top 30 Best Performing Words (YK)

Rank	Keyword	Rank	Keyword
1.	Embryonic stem-cells	16.	Acute myeloid-leukemia
2.	Carbon nanotubes	17.	Long-term potentiation
3.	Field-effect transistors	18.	Activated protein-kinase
4.	Graphite	19.	NF-κB
5.	Genome-wide association	20.	Genome
6.	Caenorhabditis-elegans	21.	Placebo-controlled trial
7.	DNA methylation	22.	Arabidopsis-thaliana
8.	Living cells	23.	Growth-factor receptor
9.	Regulatory T-cells	24.	Synaptic Plasticity
10.	Gold nanoparticles	25.	Mouse model
11.	TGF-β	26.	Mesenchymal stem-cells
12.	One-pot synthesis	27.	Growth-factor-β
13.	Quantum dots	28.	Drug-delivery
14.	Functionalization	29.	Human brain
15.	Electrodes	30.	Innate immunity

Table B3: Top 30 Best Performing Words (YT)

Rank	Keyword	Rank	Keyword
1.	Graphene	16.	Histone
2.	Batteries	17.	Genome
3.	Sequencing	18.	Inflammation
4.	Randomized	19.	Alzheimer's
5.	Meta-analysis	20.	Recent
6.	Genome-wide	21.	Mammalian
7.	Guidelines	22.	High-performance
8.	Society	23.	Arabidopsis
9.	Advances	24.	Recommendations
10.	Systematic	25.	Reveals
11.	2010	26.	Solar
12.	Update	27.	Regulates
13.	Immunity	28.	Photovoltaic
14.	Methylation	29.	Targeting
15.	Photocatalytic	30.	Topological

Best performing words

Table B4: Top 30 Best Performing Words (YA)

Rank	Keyword	Rank	Keyword
1.	Graphene	16.	EMT
2.	BRAF	17.	Progression-free
3.	Batteries	18.	Nanomaterials
4.	Reads	19.	RNAs
5.	Microbiota	20.	Plasmonic
6.	Autophagy	21.	Genome-wide
7.	Nanosheets	22.	Pluripotent
8.	microRNAs	23.	Biofuels
9.	Reprogramming	24.	AACR
10.	miRNAs	25.	NRF2
11.	SIRT1	26.	Non-coding
12.	Person-years	27.	Transcriptome
13.	Trastuzumab	28.	Self-renewal
14.	KRAS	29.	Epigenetic
15.	Next-generation	30.	Functionalization

Table B5: Top 30 Worst Performing Words (YK)

Rank	Keyword	Rank	Keyword
1.	Spaces	16.	Trauma
2.	Sheep	17.	Geometry
3.	Dogs	18.	Convergence
4.	Brazil	19.	Ceramics
5.	Existence	20.	Waves
6.	Cattle	21.	Students
7.	Law	22.	Politics
8.	Injuries	23.	Alloys
9.	Cultivars	24.	Fish
10.	Steel	25.	Leaves
11.	Equations	26.	Flows
12.	Yield	27.	Spain
13.	Education	28.	Gender
14.	Turkey	29.	Competition
15.	Constituents	30.	Patient

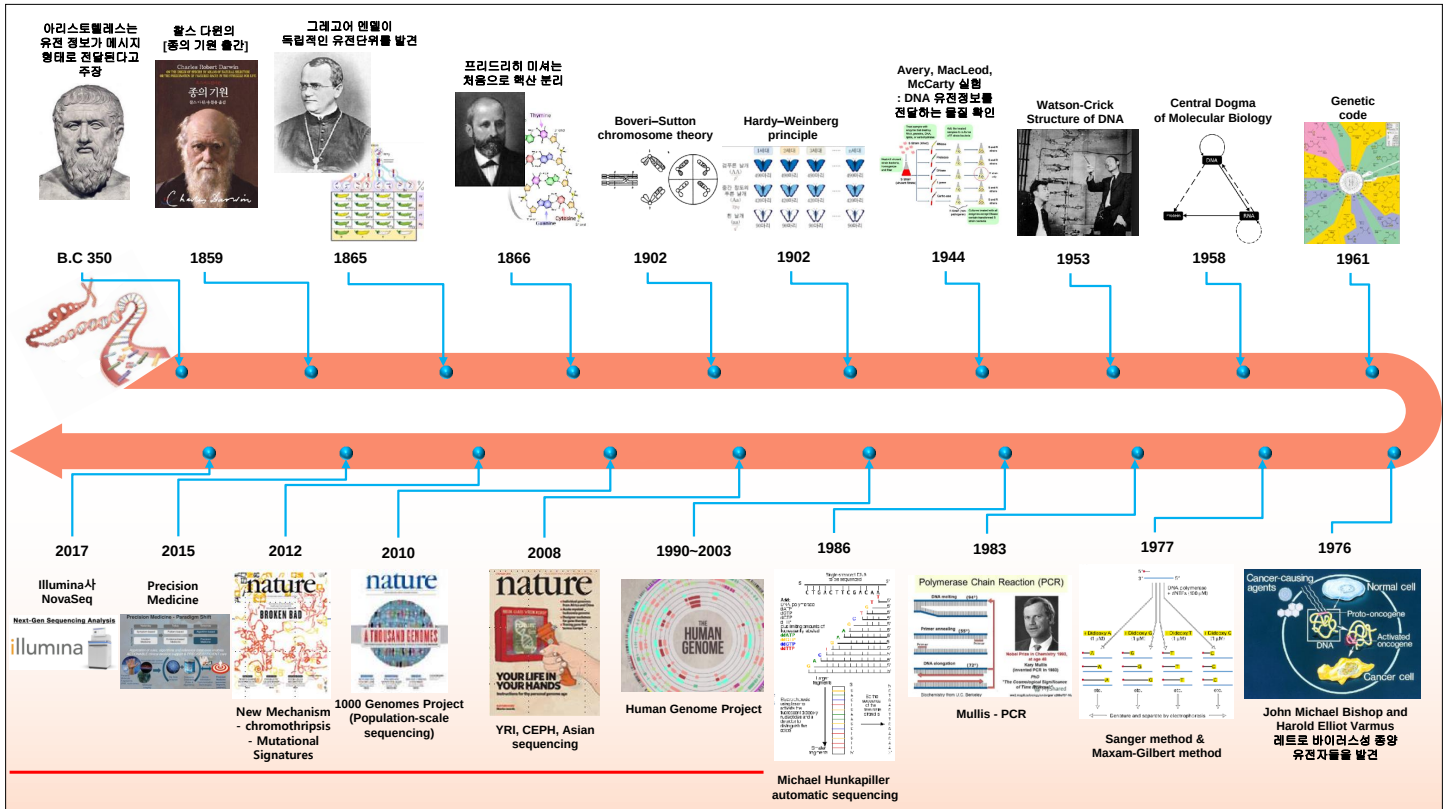
Table B6: Top 30 Worst Performing Words (YT)

Rank	Keyword	Rank	Keyword
1.	Note	16.	Some
2.	Theorem	17.	Mexico
3.	Graphs	18.	Existence
4.	Turkey	19.	Politics
5.	Operators	20.	Integral
6.	Genus	21.	Presenting
7.	University	22.	Report
8.	Spaces	23.	Symmetric
9.	Case	24.	Unusual
10.	Asymptotic	25.	Nursing
11.	Teaching	26.	Equations
12.	Law	27.	Education
13.	Brazil	28.	Brazilian
14.	Iran	29.	Korean
15.	Bilateral	30.	Korea

Worst performing words

Table B7: Top 30 Worst Performing Words (YA)

Rank	Keyword	Rank	Keyword
1.	Espana	16.	Abelian
2.	Boy	17.	Replications
3.	Girl	18.	Eyelid
4.	Let	19.	(p > 0.05)
5.	Anstrom	20.	Turkish
6.	Banach	21.	Projective
7.	Algebras	22.	Irreducible
8.	His	23.	Polish
9.	Tritium	24.	Paulo
10.	She	25.	Russian
11.	Opt	26.	Buffalo
12.	Woman	27.	Essay
13.	Algebra	28.	Courts
14.	Court	29.	Man
15.	Colonial	30.	Rio



The History of Cancer Genomics

REVIEWS

The cancer genome
2011
Michael R. Stratton^{1,2}, Peter J. Campbell^{1,2}
Nature. 2011;474(7353):609-615. doi:10.1038/nature10661

2012
Comprehensive molecular characterization of human breast cancer
Nature. 2012;487(7407):330-337. doi:10.1038/nature11246

2013
Integrated genomic characterization of endometrial cancer
Nature. 2013;497(7447):67-73. doi:10.1038/nature12462

2014
Comprehensive molecular characterization of urothelial bladder carcinoma
Nature. 2014;507(7492):315-322. doi:10.1038/nature12965

UK Sanger Nature 2009

Cell

Perspective on Oncogenic Processes at the Beginning of Cancer Genomics
Graphical Abstract
Ding et al., *Cell* 2018

THE CANCER GENOME ATLAS
11,000 genes, 33 cancer types

PAN-CANCER ATLAS
Substrates: DNA, RNA & Protein, Cellular
Associations: Germ-line and somatic, Driver & molecular subtypes, Immune cell & tumor

Article

Pan-cancer analysis of whole genomes
The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium
Nature 578, 82-93(2020) | Cite this article
262k Accesses | 169 Citations | 1189 Altmetric | Metrics

~2,700 PanCan WGS

Article | Open Access | Published: 05 February 2020

Authors
Li Ding, Matthew H. Bailey, Eduard Porta-Pardo, ..., David J. Gad Getz, The Cancer Genome Research Network

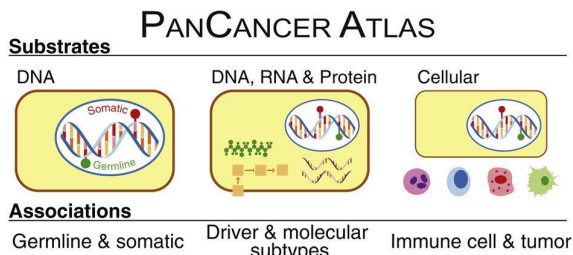
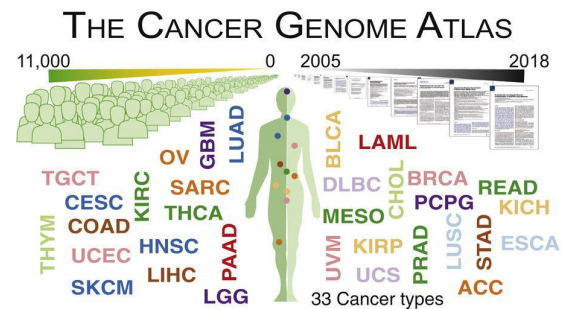
Correspondence
lding@wustl.edu (L.D.), wheeler@bcm.edu (D.A.W.), gadgett@broadinstitute.org (G)

In Brief
A synthesized view on oncogenic processes based on PanCancer analyses highlights the complex of genome alterations on the s and multi-omic profiles of human cells as well as their influence on the microenvironment.

TCGA Research Network Publication

Nature, 2020

Cancer Big Project



Li Ding et. al., Cell 2018

F-22 Raptor stealth fighters

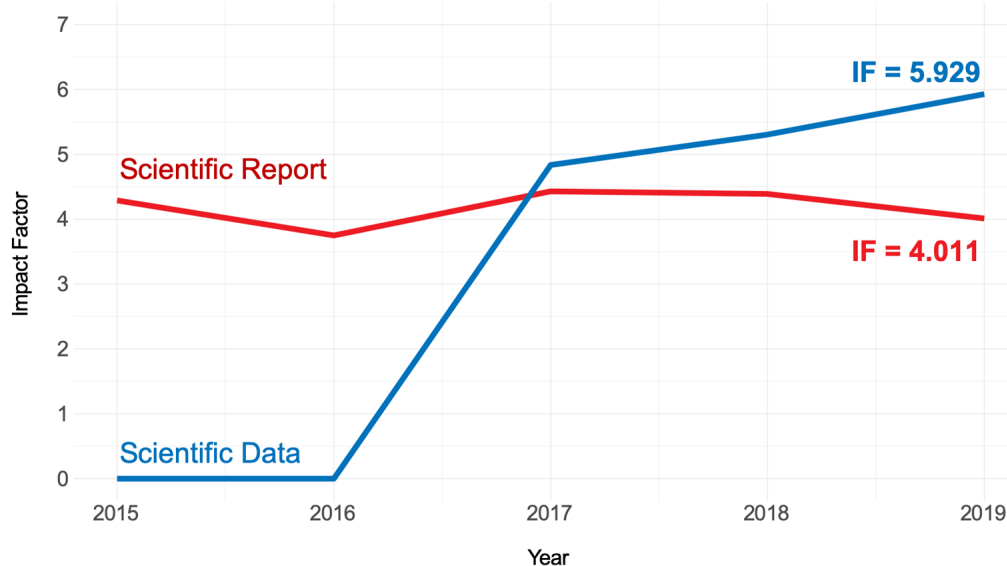


- TCGA project was finished with advancements in cancer genomics
 - Systemized large-scale genomics-based cancer research (33 types / 11,000 tumors)

HISTORY OF CANCER TREATMENT MODALITIES

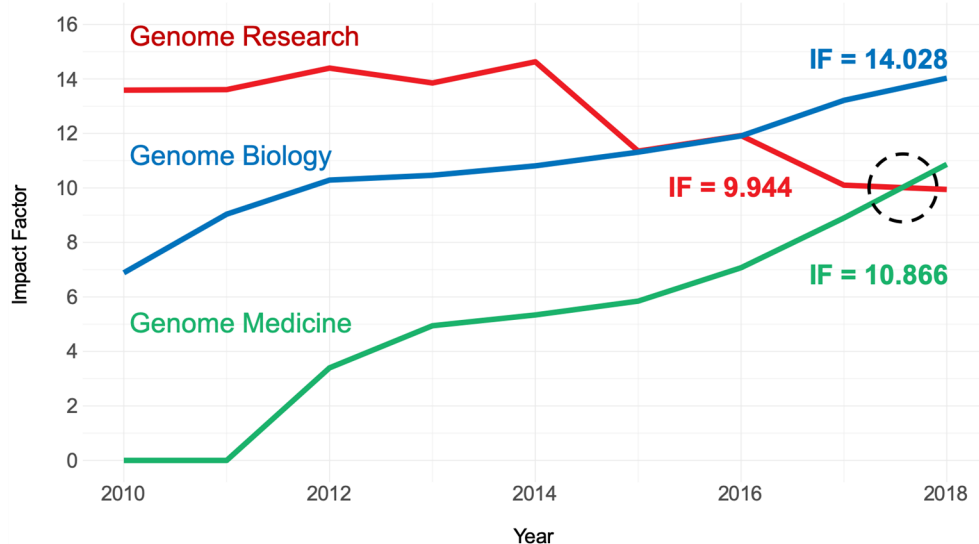
	SURGERY	RADIATION	CHEMO-THERAPY	TARGETED DRUGS	IMMUNO-THERAPY
APPROACH	Cut out accessible tumor cells to stop growth and prevent their spread	Use highly concentrated X-rays or radioactive isotopes to kill cancerous cells	Use cytotoxic drugs to kill or inhibit cancer cells	Interfere with a mechanism required for, or that supports tumor growth	Support the immune system's innate ability to recognize and eliminate tumor cells
SINCE	1800s	early 1900s	late 1940s	2000s	2010s
LIMITATIONS	Many inaccessible tumors ineligible; limited effectiveness if tumor has already begun to spread	Limited effectiveness if tumor has already begun to spread; potentially dangerous for tumors near vital organs	High toxicity and often does not destroy the whole tumor, leading to high rates of recurrence	Limited tumor types eligible; high efficiency but short durability driving high rates of recurrence	Applicable to all tumors at all stages of disease including metastatic tumors; responses are highly durable; potential for lower toxicity profiles; synergistic with other treatments

Trend 1. Data are More Impactful Than Individual Reports



Report < **Data**

Trend 2. Medicine Outweighs Research



Research < Medicine

Clinical Genomic Big Data

500 PB (2012) -> 25,000 PB (2020)

Exa -> Zeta -> Yotta

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

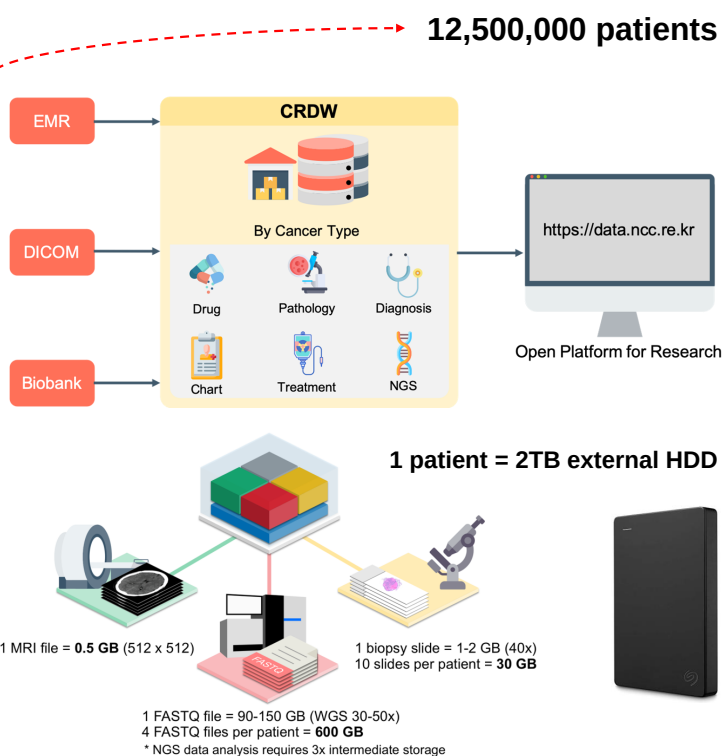
TCGA data describes
33
DIFFERENT
TUMOR TYPES

...including
10
RARE
CANCERS

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

...using
7
DIFFERENT
DATA TYPES

TCGA Research Network



• Part I: Cancer Big Database

- History of cancer genome study
- Cbioportal, Xena browser and GDC data portal
- Study example of cancer big database

The first version of TCGA(The Cancer Genome Atlas)

Next Generation Sequencing data
> 20 peta bytes

The image displays three web portals used for cancer genomics data analysis:

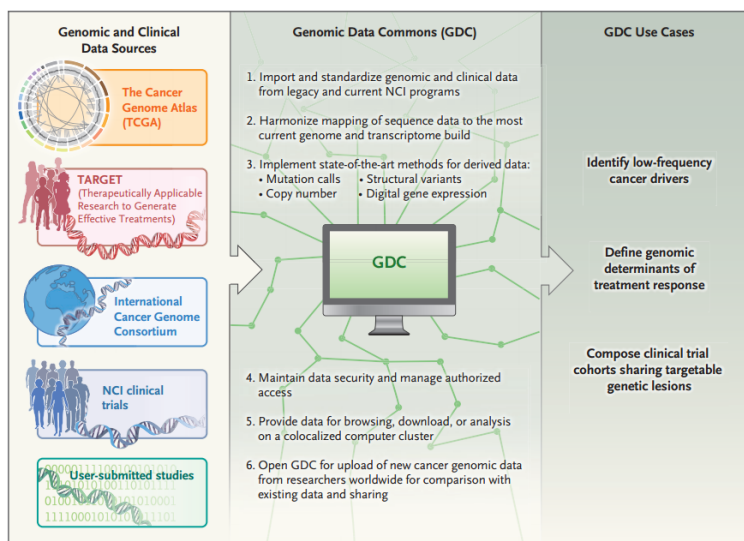
- TCGA data portal:** The Cancer Genome Atlas (TCGA) Data Portal Overview. It provides a platform for researchers to search, download, and analyze data generated by TCGA. The portal includes a table of available cancer types and a list of data types.
- cBioportal:** The cBioPortal for Cancer Genomics. It provides a platform for researchers to search, download, and analyze data generated by TCGA. The portal includes a table of available cancer types and a list of data types.
- CGHub:** The Cancer Genomics Hub. It provides a platform for researchers to search, download, and analyze data generated by TCGA. The portal includes a table of available cancer types and a list of data types.

Below the screenshots, a terminal window shows the download progress of data from CGHub:

```

root@localhost:~# curl -O https://cghub.cancer.gov/data/TCGA/
Status: 153 GB downloaded (99.462% complete) current rate: 11.4 MB/s
Child 1 downloading 19.0 GB (2.25 MB/s)
Child 4 downloading 19.2 GB (2.35 MB/s)
Child 3 downloading 19.2 GB (1.24 MB/s)
Child 6 downloading 19.2 GB (1.60 MB/s)
Child 5 downloading 19.2 GB (1.98 MB/s)
Child 8 downloading 19.9 GB (2.01 MB/s)
Status: 153 GB downloaded (99.540% complete) current rate: 11.4 MB/s
Child 5 finished 19.3 GB (1.83 MB/s)
Child 1 downloading 19.0 GB (2.76 MB/s)
Child 4 downloading 19.2 GB (2.43 MB/s)
Child 3 downloading 19.2 GB (1.82 MB/s)
Child 6 downloading 19.3 GB (1.54 MB/s)
Child 8 downloading 19.0 GB (2.25 MB/s)
Status: 153 GB downloaded (99.584% complete) current rate: 12.6 MB/s
  
```

Toward a Shared Vision for Cancer Genomic Data

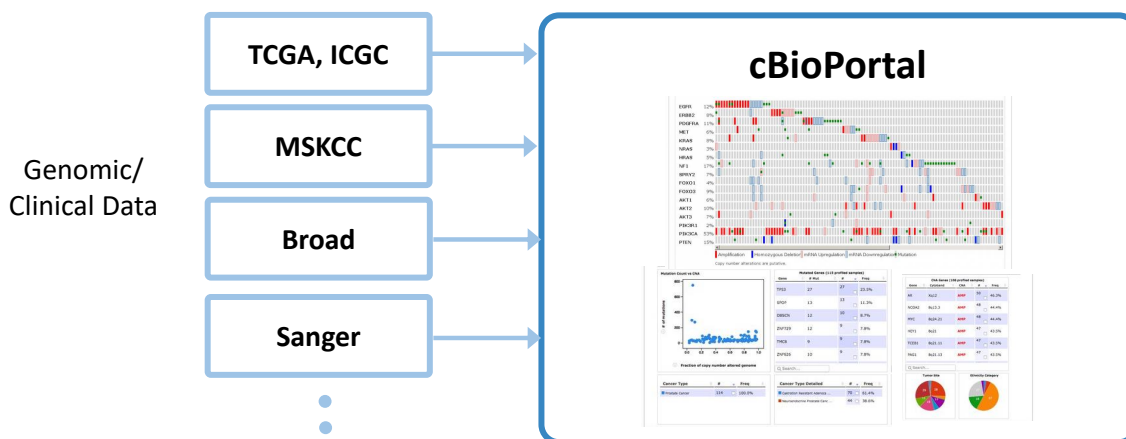


NEJM, Sep, 2016

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 of key genetic changes in 33 types of cancer.

Cell, 2017



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

Access cBioPortal

<http://www.cbioportal.org/>

Prostate Adenocarcinoma (TCGA, Provisional) 선택

What's New @cbioportal

• 20가지의 암종에 대하여 총 292 스타터. 107,299 샘플 (116,057 시퀀싱) 포함

Summary 보기

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

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

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

Select Data Type Priority: ☒ Mutation and CNA ☐ Only Mutation ☐ Only CNA

Enter Gene Set: User-defined List

Advanced: Onco Query Language (OQL)

Enter HUGO Gene Symbols or Gene Aliases

Cancer Studies

The portal contains 164 cancer studies (details)

Cases by Top 20 Primary Sites

Breast

CNS/Brain

Lung

Prostate

Stomach

Kidney

Bowel

Blood

Head/Neck

Ovary

Thyroid

Uterus

Liver

Pancreas

Bladder

Skin

Soft Tissue

Cervix

Bone

Lymph

0 2000 4000 6000

Example Queries

Prostate Adenocarcinoma (TCGA, Provisional) Study Summary 보기

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

Overall Survival

Disease Free Survival

Mutation Count vs. CNA

Neoplasm American Joint Comm...

With Mutation Data

With CNA Data

Surviving

Months Survival

Surviving

Months Survival

of mutations

Fraction of copy number altered genome

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%

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%

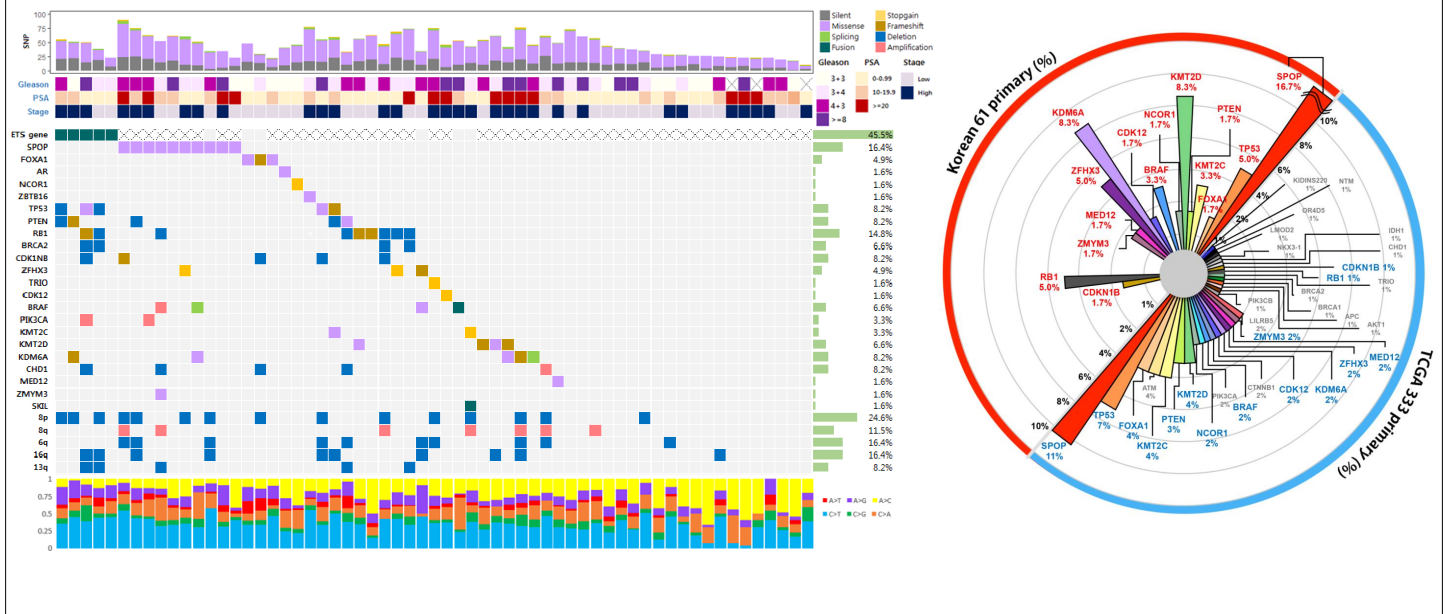
RB1 13q14.2 DEL 81 16.46%

CYSLTR2 13q14.2 DEL 81 16.46%

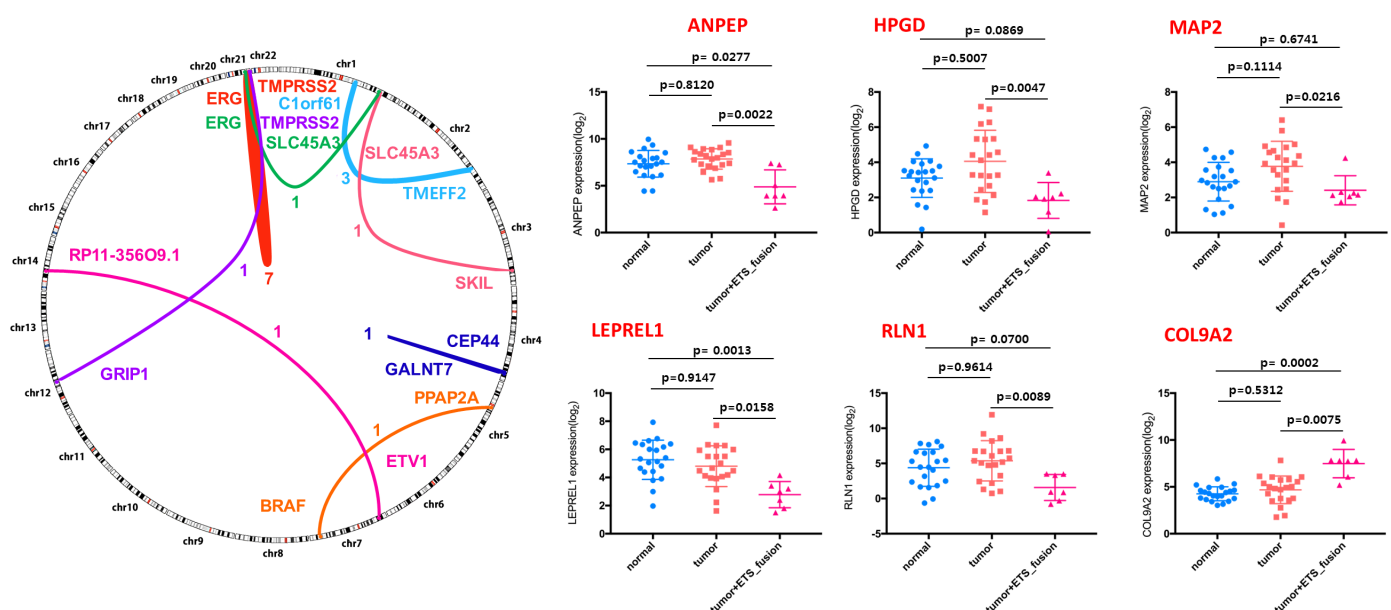
499

492

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

Select Genomic Profiles:

- ☒ Mutations
- ☒ Putative copy-number alterations from GISTIC
- ☒ mRNA Expression z-scores (RNA Seq V2 RSEM)

Enter a z-score threshold $\pm$ 2.0
- ☐ Protein expression z-scores (RPPA)

Select Patient/Case Set: All Complete Tumors (491)

Enter Gene Set: User-defined List

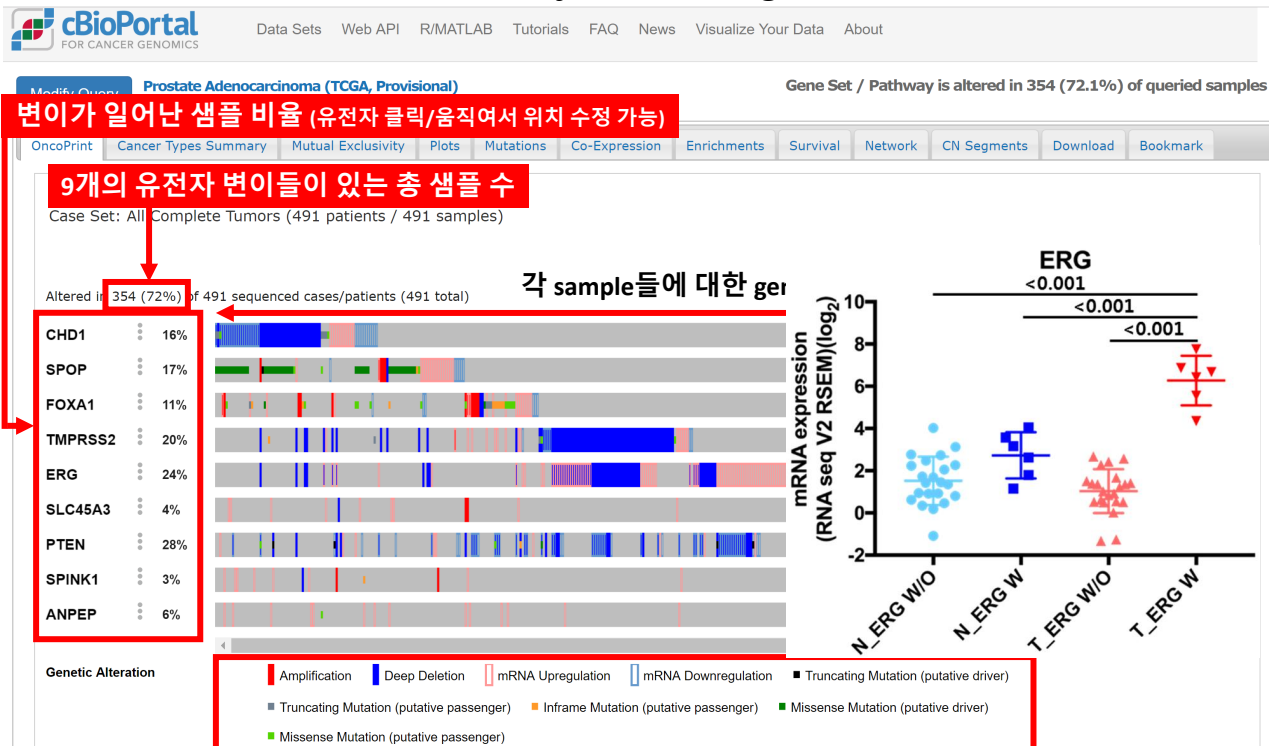
Select from Recurrently Mutated Genes (MutSig) | Select Genes from Recurrent CNAs (Gistic)

CHD1 SPOP FOXA1 TMPRSS2 ERG SLC45A3 PTEN SPINK1 ANPEP

All gene symbols are valid.

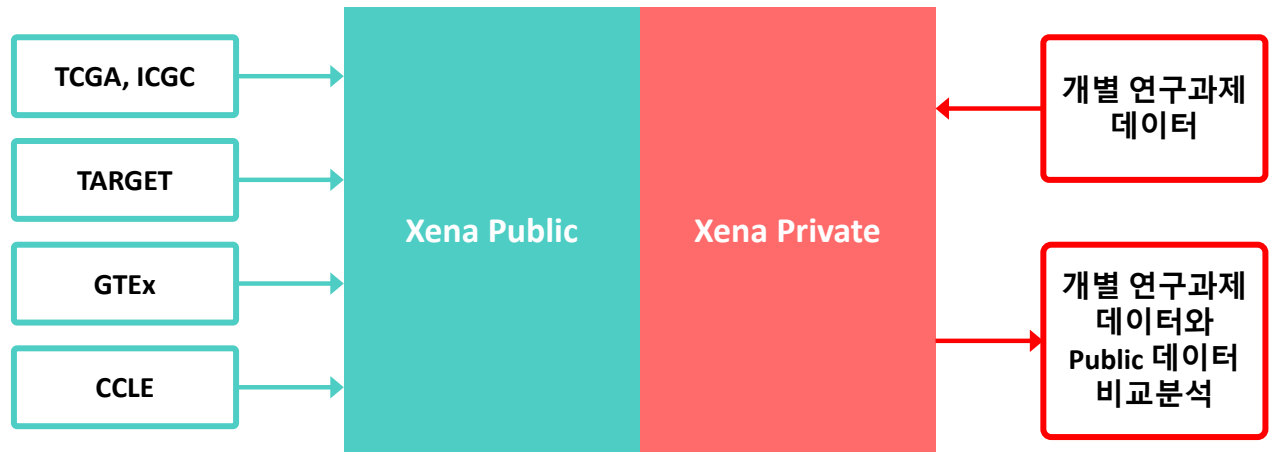
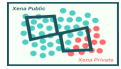
Submit Query

Summary of cancer genes



UCSC Xena Browser

UCSC Xena
See the bigger picture



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

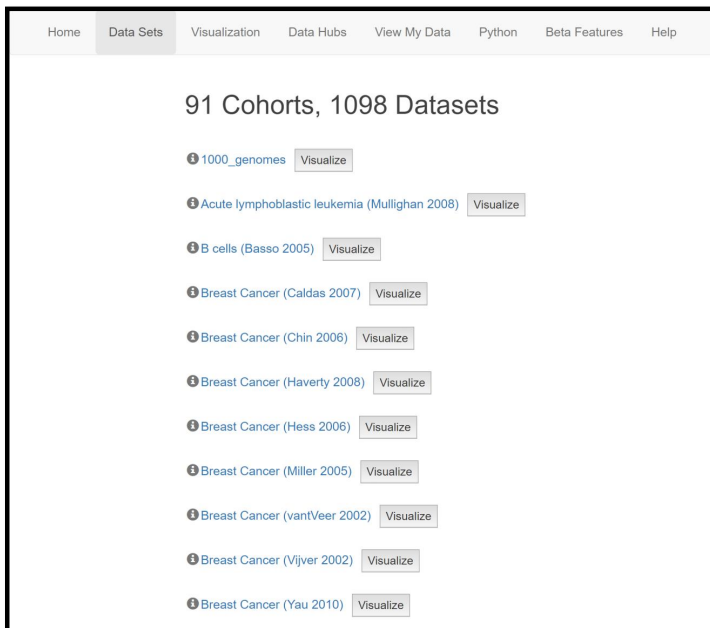
20,530 rows, 550 columns

TCGA-H9-A6BY-01
TCGA-EJ-7314-11

sample	TCGA-XJ-A83	TCGA-G9-63	TCGA-CH-57	TCGA-EJ-A65	TCGA-G9-63	TCGA-EJ-552	TCGA-HC-82	TCGA-Y6-A9	TCGA-EJ-712	TCGA-CH-57	TCGA-EJ-778	TCGA-HC-A9	TCGA-EJ-A7	TCGA-SU-A7	TCGA-CH-57	TCGA-HI-716	TCGA-ZG-A9
1	sample																
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
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
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
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
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
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
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
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
10	MTVR2	0	0.4008	0	0	0	0.6248	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
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
13	C16orf11	0	0	0.4278	0	0	0	0	0	0.7956	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
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
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
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
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
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
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
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
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
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

Access Xena Browser, Select TCGA PRAD

<https://xenabrowser.net/datapages/>



1 리스트에서 TCGA Prostate Cancer (PRAD) 찾기

2 TCGA Prostate Cancer (PRAD) 클릭



i TCGA Prostate Cancer (PRAD) Visualize

Gene Expression 데이터 다운로드

dataset: gene expression RNAseq (polyA+ IlluminaHiSeq)

Visualize

1

Download

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

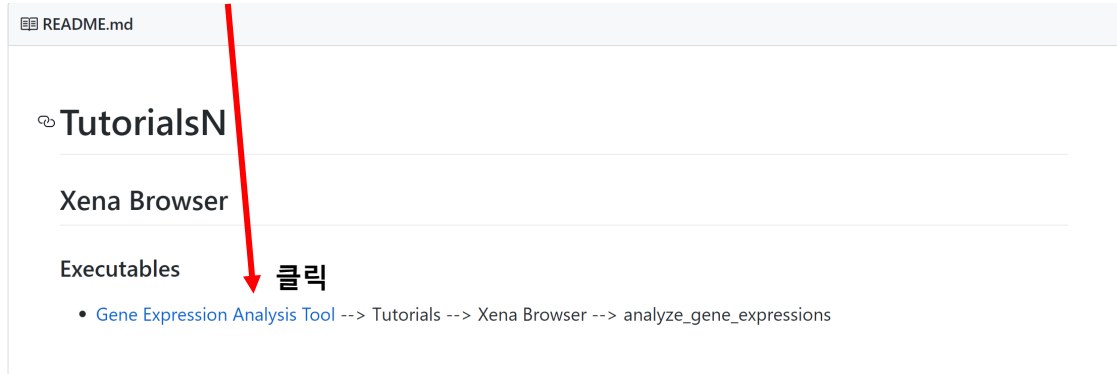
The gene expression profile was measured experimentally using the Illumina HiSeq 2000 RNA Sequencing platform by the University of North Carolina TCGA genome characterization center. Level 3 data was downloaded from TCGA data coordination center. This dataset shows the gene-level transcription estimates, as in $\log_2(x+1)$ 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)

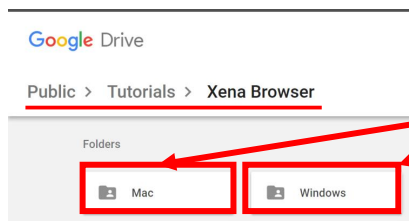
2. Google Drive로 이동



Gene Expression 데이터로 간단한 분석하기

3. Tutorials → Xena Browser 로 이동

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



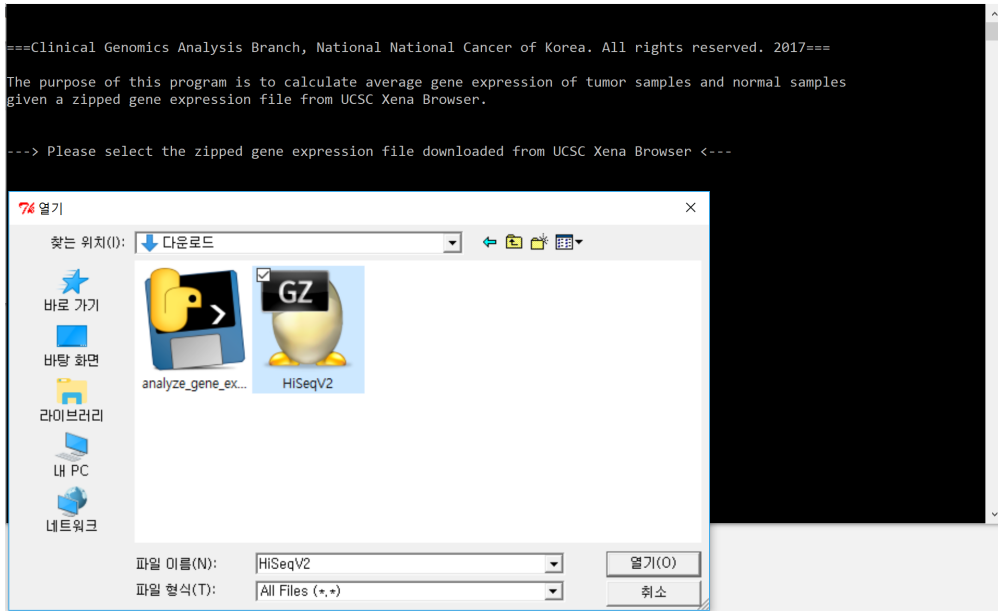
폴더 안에 있는
analyze_gene_expressions
응용 프로그램 다운로드

5. analyze_gene_expressions 더블 클릭하여 프로그램 실행



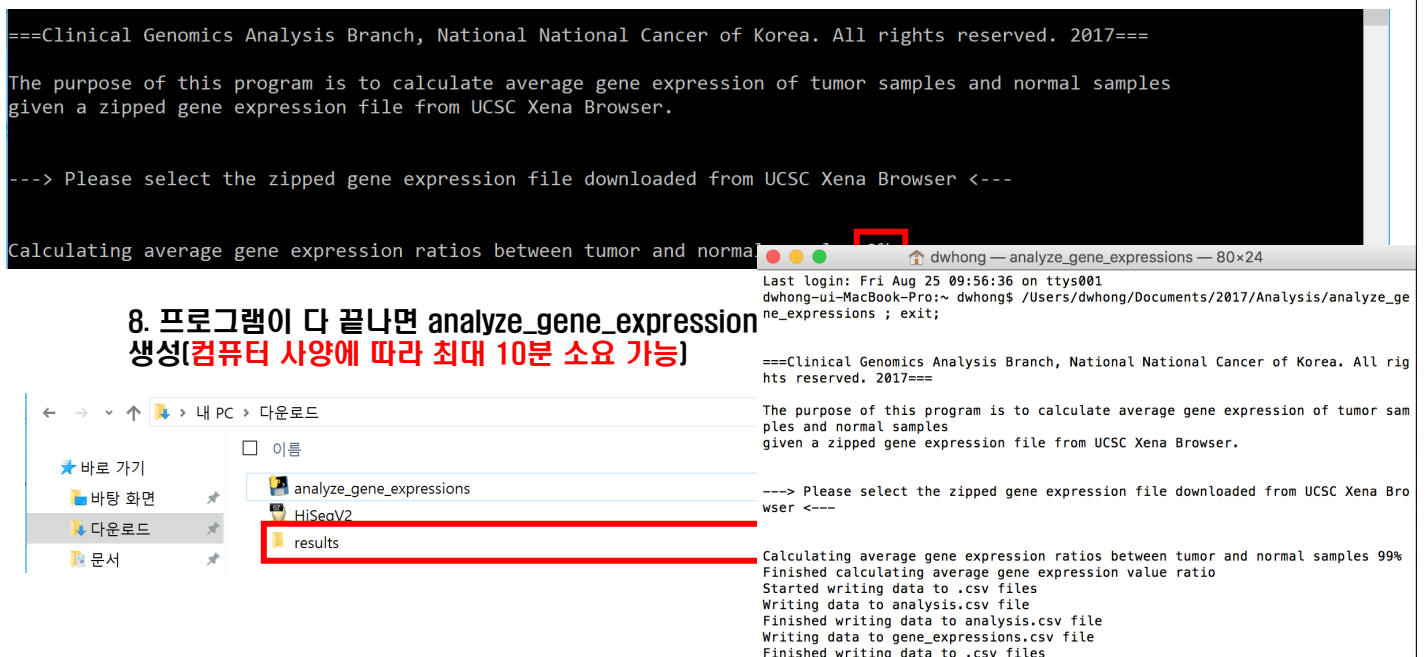
Gene Expression 데이터로 간단한 분석하기

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



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

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

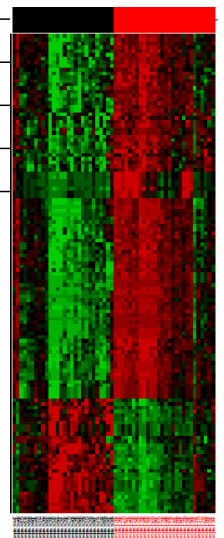
gene_name	tumor_samples_count	normal_samples_count	avg(tumor)	avg(normal)	avg(tumor)/avg(normal)	avg(tumor)=0_count	avg(normal)=0_count
NKX2-3	497	52	4.364812072	0.857811538	5.088311216	13	21
ZIC2	497	52	5.140035815	1.246317308	4.124179118	14	12
SLC45A2	497	52	5.236834406	1.359369231	3.852400281	3	5
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
FOXP4	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	10
TDOD1	497	52	6.499556338	2.667855769	2.436247271	2	1
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

$$\begin{array}{c}
 \text{Experiment or Sample} \\
 \begin{array}{c} 1 \quad 2 \quad \dots \quad n \\
 \text{Gene} \begin{array}{c} 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array} \left[\begin{array}{cccc} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{array} \right]
 \end{array}
 \end{array}
 \quad
 \begin{array}{c}
 \begin{array}{c} 1 \quad 2 \quad \dots \quad n \\
 \begin{array}{c} 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array} \left[\begin{array}{cccc} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{array} \right]
 \end{array}
 \end{array}
 \begin{array}{c}
 \text{NORMAL (CONTROL)} \qquad \qquad \text{TUMOR (TREATMENT)}
 \end{array}$$

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

To download and install R software

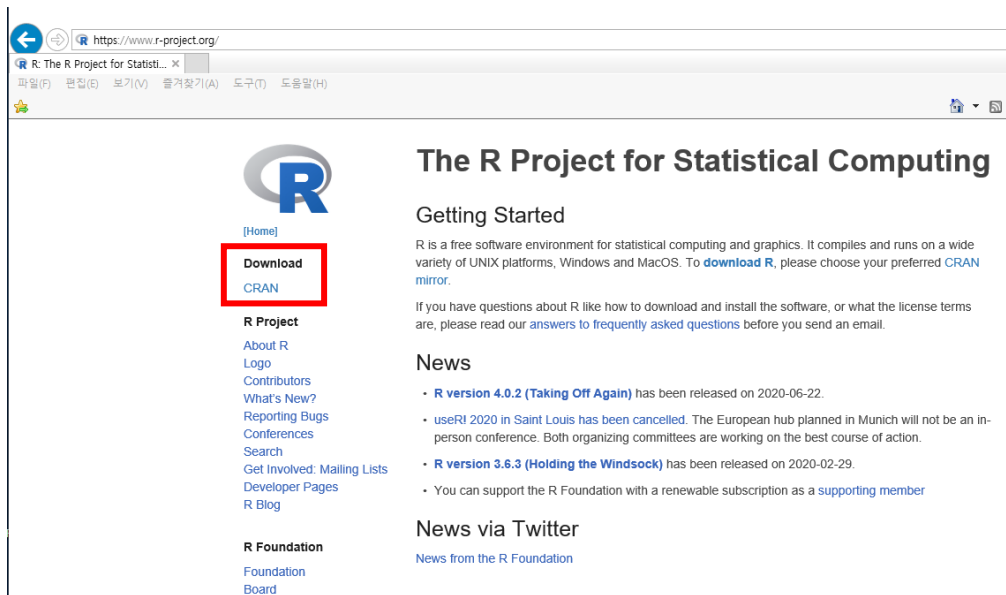
- R packages

- <https://cran.seoul.go.kr>

- R studio

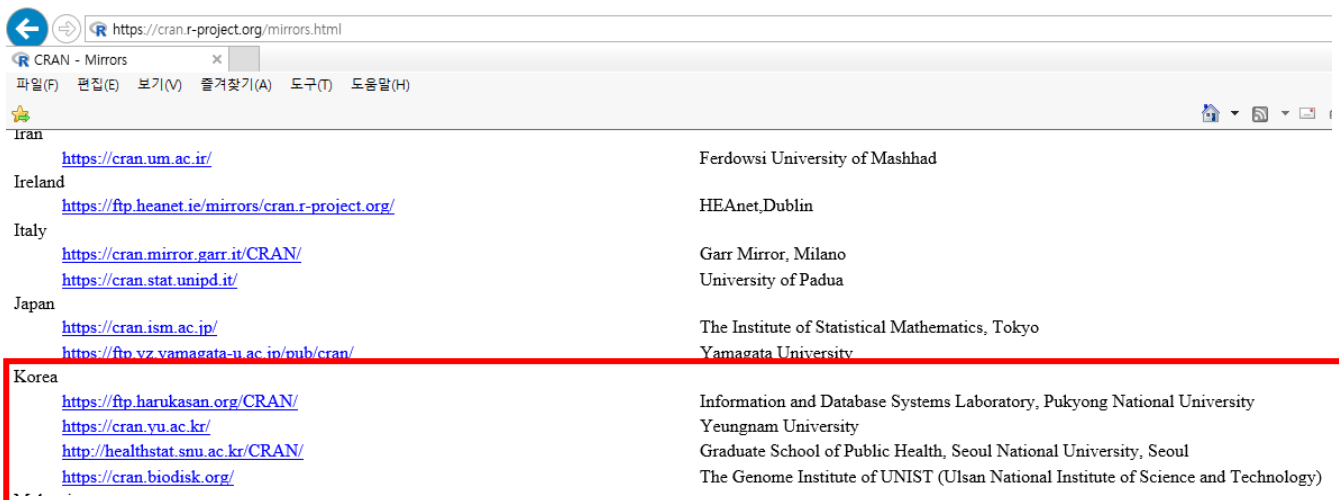
- <https://www.rstudio.com/products/rstudio/download/#download>

To download and install R software



• R packages

- <https://cran.seoul.go.kr>



← → http://healthstat.snu.ac.kr/CRAN/

The Comprehensive R Archive Network

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](#)

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](#)

← → http://healthstat.snu.ac.kr/CRAN/

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.

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](#)

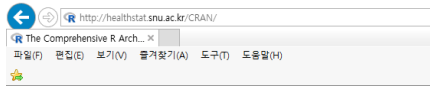
Documentation

[Manuals](#)

[FAQs](#)

Stable version (3.6.3)

R-3.6.3 for Windows (32/64 bit)



CRAN
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

[About R](#)
[R Homepage](#)
[The R Journal](#)

[Software](#)
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

This directory contains previous binary
The current release, and links to develop

In this directory:

[R 4.0.2](#) (June, 2020)
[R 4.0.1](#) (June, 2020)
[R 4.0.0](#) (April, 2020)
[R 3.6.3](#) (February, 2020)
[R 3.6.2](#) (December, 2019)
[R 3.6.1](#) (July, 2019)
[R 3.6.0](#) (April, 2019)
[R 3.5.3](#) (March, 2019)
[R 3.5.2](#) (December, 2018)
[R 3.5.1](#) (July, 2018)
[R 3.5.0](#) (April, 2018)
[R 3.4.4](#) (March, 2018)
[R 3.4.3](#) (November, 2017)

[Download R 3.6.3 for Windows](#) (83 megabytes, 32/64 bit)

[Installation and other instructions](#)

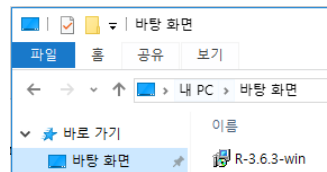
[New features in this version](#)

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.



• R studio

- <https://www.rstudio.com/products/rstudio/download/#download>

RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging and managing your workspace.

[LEARN MORE ABOUT RSTUDIO FEATURES](#)

RStudio's new solution for every professional data science team. RStudio Team includes RStudio Server Pro, RStudio Connect and RStudio Package Manager.

[LEARN MORE](#)

	RStudio Desktop Open Source License	RStudio Desktop Commercial License	RStudio Server Open Source License	RStudio Server Pro Commercial License
	Free	\$995 /year	Free	\$4,975 /year (5 Named Users)
	DOWNLOAD	BUY	DOWNLOAD	BUY
	Learn more	Learn more	Learn more	Evaluation Learn more
Integrated Tools for R	✓	✓	✓	✓
Priority Support		✓		✓
Access via Web Browser			✓	✓
Enterprise Security				✓
Project Sharing				✓

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	2fca472a
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/Bebian 10	rstudio-1.3.1073-amd64.deb	126.78 MB	86be9352

IDE (Integrated Development Environment)

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Console Terminal Jobs

```
R version 3.6.2 (2020-02-29) -- "Holding the windsock"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R은 자유 소프트웨어이며, 어떠한 형태의 배포권도 없습니다.
또한, 일정한 조건하에서 이것을 재배포 할 수 있습니다.
배포와 관련된 상세한 내용은 'license()' 또는 'licence()'를 통하여 확인할 수 있습니다.

R은 많은 기여자들이 참여하는 공동프로젝트입니다.
'contributors()'라고 입력하시면 이에 대한 더 많은 정보를 확인할 수 있습니다.
그리고, R 또는 R 패키지를 출판물에 인용하는 방법에 대해서는 'citation()'를 통해 확인하실 수 부탁드립니다.

'demo()'를 입력하신다면 몇가지 데모를 보실 수 있으며, 'help()'를 입력하시면 온라인 도움말을 이용하실 수 있습니다.
또한, 'help.start()'의 입력을 통하여 HTML 브라우저에 의한 도움말을 이용하실 수 있습니다.
R의 종류를 원하시면 'q()'를 입력해주세요.

> |
```

Environment History Connections Tutorial

Import Dataset

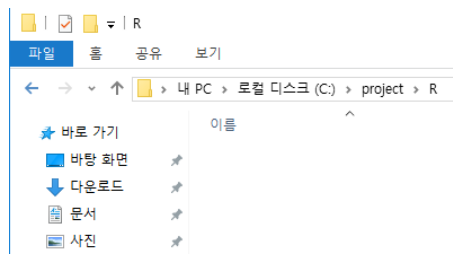
Global Environment

Environment is empty

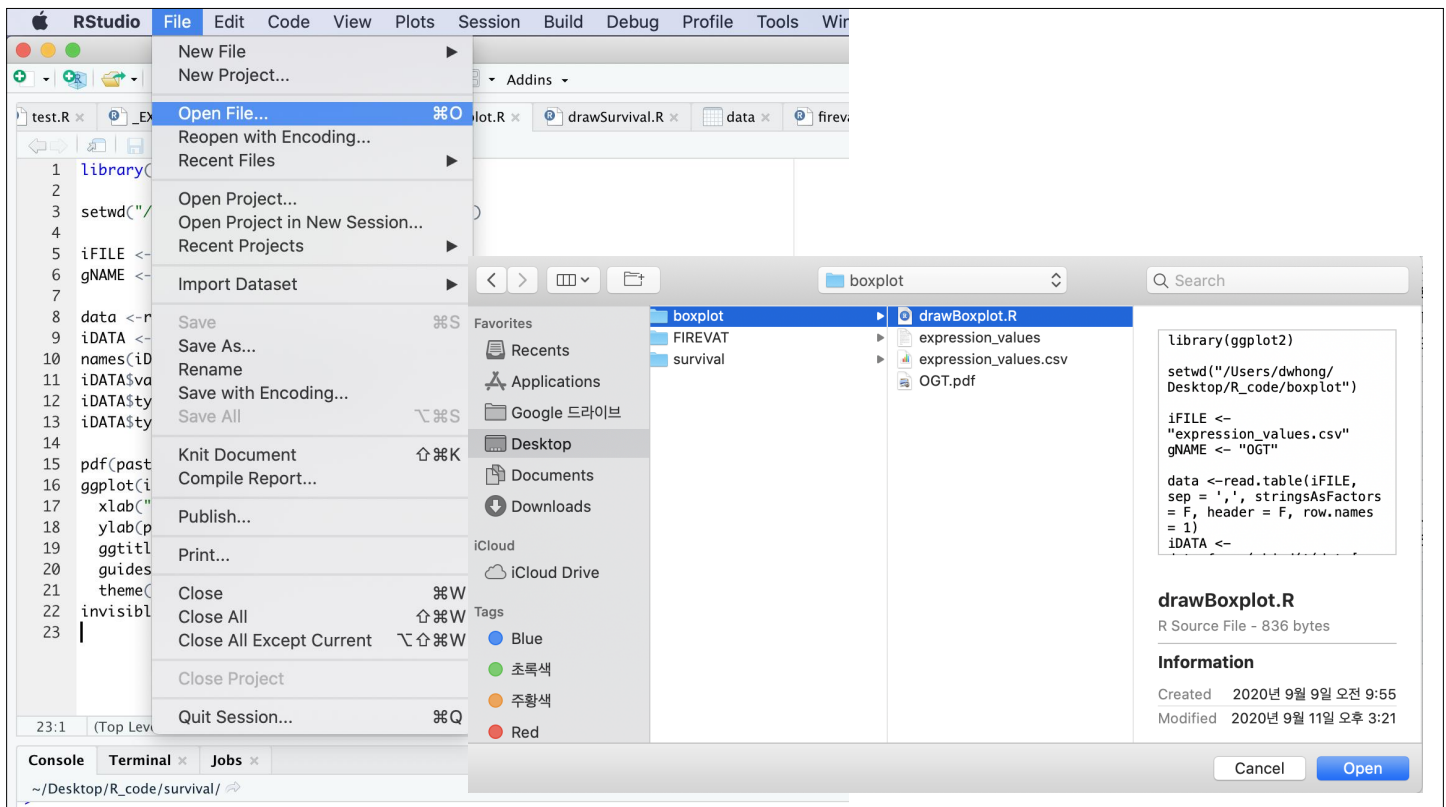
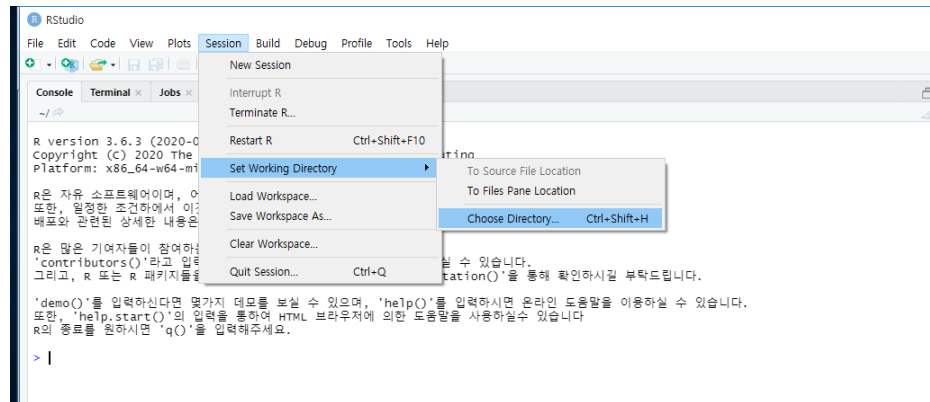
Files Plots Packages Help Viewer

New Folder Delete Rename More

Name	Size	Modified
10608488_popupPos.dat	112 B	Jan 2, 2020, 5:49 PM
230000079_popupPos.dat	136 B	Sep 14, 2020, 3:46 PM
desktop.ini	402 B	Jan 2, 2020, 11:15 AM
My Music		
My Pictures		
My Videos		
undefined_popupPos.dat	112 B	May 13, 2020, 1:26 PM
사용자 지정 Office 서식 파일		



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



> 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

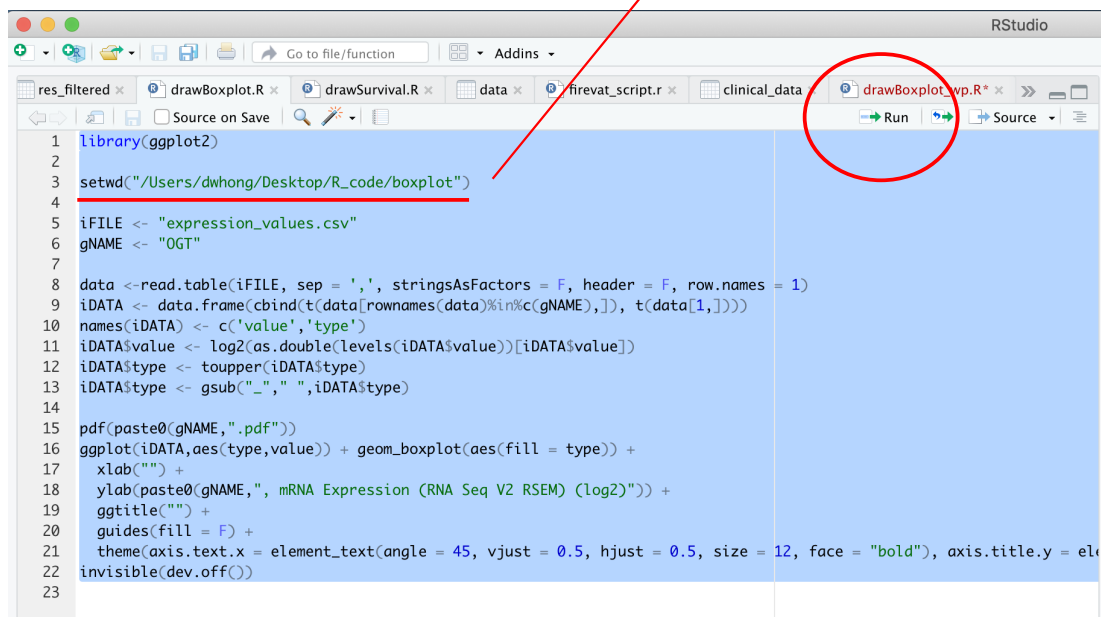
/var/folders/45/v4b_fjns00jg9ytjprdsxkd00000gn/T/Rtmp0BReEZ/downloaded_packages

```
> pdf(paste0("BCR_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())
> install.packages("ggplot2")
Error in install.packages : Updating loaded packages
> 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
>
```

Run selected R scripts

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



```
1 library(ggplot2)
2
3 setwd("/Users/dwhong/Desktop/R_code/boxplot")
4
5 iFILE <- "expression_values.csv"
6 gNAME <- "OGT"
7
8 data <- read.table(iFILE, sep = ',', stringsAsFactors = F, header = F, row.names = 1)
9 iDATA <- data.frame(cbind(t(data[rownames(data)%in%c(gNAME),]), t(data[1,])))
10 names(iDATA) <- c('value', 'type')
11 iDATA$value <- log2(as.double(levels(iDATA$value))[iDATA$value])
12 iDATA$type <- toupper(iDATA$type)
13 iDATA$type <- gsub("-", "", iDATA$type)
14
15 pdf(paste0(gNAME, ".pdf"))
16 ggplot(iDATA, aes(type, value)) + geom_boxplot(aes(fill = type)) +
17   xlab("") +
18   ylab(paste0(gNAME, ", mRNA Expression (RNA Seq V2 RSEM) (log2)")) +
19   ggtitle("") +
20   guides(fill = F) +
21   theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust = 0.5, size = 12, face = "bold"), axis.title.y = element_text(angle = 0, size = 12, face = "bold"))
22 invisible(dev.off())
23
```

Browsing the output

The screenshot shows the RStudio interface. The main editor displays R code for generating a boxplot. The Environment pane on the right lists data objects. The Files pane at the bottom right, circled in red, shows the file structure of the project.

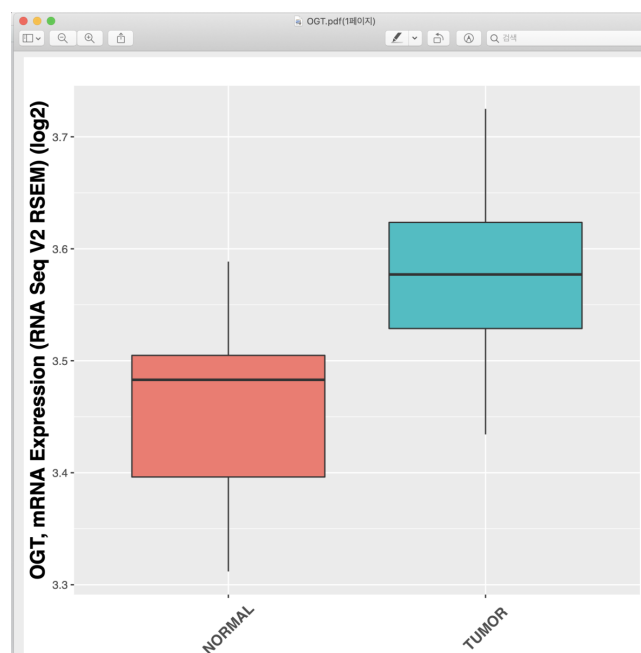
Environment Pane:

Object	Details
clinical_data	28 obs. of 4 variables
conds	Formal class DFrame
count_data	40304 obs. of 7 variables
count.mat	40304 obs. of 6 variables
d	Large matrix (61965 elements, 578.1 kB)
data	21 obs. of 90 variables
dds	Large DESeqDataSet (40304 elements, 26.3 MB)
expression	28 obs. of 2 variables
expression.table	1215 obs. of 51 variables
gene.exp	Large matrix (61965 elements, 578.1 kB)
iDATA	90 obs. of 2 variables

Files Pane:

Name	Size	Modified
..		
.Rhistory	17.7 KB	Sep 11, 2020, 2:56 PM
drawBoxplot.R	836 B	Sep 11, 2020, 3:21 PM
expression_values.csv	13.5 KB	Sep 8, 2020, 9:25 PM
expression_values.txt	13.5 KB	Sep 11, 2020, 2:35 PM
OGT.pdf	4.7 KB	Sep 15, 2020, 6:44 PM
drawBoxplot_wp.R	1.4 KB	Sep 17, 2020, 2:38 AM
OGT_pvalue.pdf	5 KB	Sep 17, 2020, 2:41 AM

Browsing the output



Load “expression_values.csv” in EXCEL

gene_name	normal	normal	normal	normal	normal	normal	normal	normal	normal	normal	normal	normal	normal	normal	normal	normal	normal	normal	normal	normal	normal
TAT	0.8958	0.8672	1.2714	0.5408	2.512	0	1.2445	0.938	1.1568	1.7502	0.4668	1.4784	1.0107	0.7716	0.983	0.8842	1.7301				
B3GNT6	0	1.9577	0	1.4952	4.0601	0.7176	0	2.2214	2.8377	1.7502	6.6538	2.8935	7.1735	6.7802	0	6.1959	4.829				
OGT	11.1784	11.3547	10.0148	10.6018	9.9306	11.2622	11.4902	11.181	10.5841	11.3482	11.251	11.4939	12.0301	11.2389	10.8557	10.3244	10.1287				
MAT2A	11.5439	13.5627	11.6553	11.4857	12.4825	11.5974	12.2364	12.0035	12.2314	12.0627	11.3862	13.3059	13.9321	11.9116	11.3162	12.4472	12.1896				
UGDH	8.9438	11.4139	10.4629	10.315	10.9648	9.3947	10.608	11.312	10.9958	10.9448	13.123	10.685	10.8975	10.7436	9.1704	11.4489	10.9039				
SCD	7.9392	11.2766	11.9138	10.8692	12.6961	9.4274	12.4205	12.9763	14.0059	14.3601	14.5127	12.5572	12.767	13.0293	8.5166	13.8213	13.2844				
LAP3	10.156	10.3446	10.4812	10.2748	10.3643	9.8901	10.5145	10.3429	10.4444	10.1088	10.6386	10.0503	10.415	10.2513	10.1426	10.4152	10.2124				
PIGW	6.9248	7.5552	6.2265	6.0149	7.2984	6.7284	6.9807	7.1077	7.01	6.6761	8.4875	7.266	7.5251	7.5523	6.4975	7.5076	7.0521				
PFKFB4	6.6743	6.6424	6.7563	7.3152	8.2443	5.8181	8.1881	7.8961	8.098	8.4418	8.1826	7.602	7.6706	7.7783	5.9775	8.1852	7.6872				
NDUF811	10.8777	10.1862	10.2686	11.0024	11.5313	10.283	10.9653	10.7783	10.3895	10.993	10.6711	11.0263	10.3235	10.3984	10.4835	10.8734	11.0743				
DHCR7	10.1141	9.5762	9.9227	9.7881	11.2564	9.3049	10.2361	10.9987	11.1945	11.5415	10.8288	10.9455	10.3004	10.4401	9.3837	11.56	11.3175				
POLR2G	9.9999	9.5051	9.4266	9.9692	10.5754	9.5316	10.6829	9.881	9.9832	10.2068	9.8711	10.1701	9.5912	9.604	9.5833	10.1329	10.4324				
POLR3E	9.6127	9.9049	9.4406	9.158	9.6472	9.4355	9.6623	9.6767	9.7803	9.9581	9.8903	9.6352	9.8618	9.4608	9.6133	9.8198	9.635				
B3GAT2	3.3651	4.2859	4.099	1.9385	2.3395	3.3589	3.5141	2.7424	3.5652	3.4687	3.2861	3.2419	4.2558	3.4806	3.5264	2.538	2.7175				
RRM2	2.7212	3.6482	4.9062	5.7724	5.7632	1.1946	3.9746	6.2585	5.1217	3.0165	10.1432	4.4646	6.4682	3.7029	2.4315	5.1568	4.3981				
ADCY10	0.5163	1.7959	1.0429	0	2.1789	0	2.9721	1.5017	2.0266	2.0538	3.3459	1.9479	1.0107	1.6423	0.5737	1.4284	2.392				
AK2	10.7419	10.7659	10.3982	10.6466	10.9903	10.397	10.9091	10.551	10.5719	10.7067	10.9951	10.8341	10.803	10.3686	10.1685	10.5777	10.6418				
GLO1	11.519	12.0438	11.5369	11.5738	12.9207	10.8486	12.1822	12.4179	12.7297	12.6109	13.2661	13.5214	11.7888	14.7447	11.0139	12.8664	13.0468				
FLAD1	9.233	9.3388	8.8658	8.9953	9.3014	8.8985	9.2713	9.0583	8.9096	9.1091	9.5427	9.4435	9.0948	8.8095	9.0116	9.0567	9.2122				
ENGASE	8.1129	9.2586	8.2307	8.5253	7.8549	8.716	9.2887	9.032	8.6126	9.4241	8.9587	9.6423	9.4067	8.8049	8.7796	8.6058	7.993				

90개 샘플 (normal: 18개, Tumor: 72개)
20 metabolic genes

> View (data)

```

1 library(ggplot2)
2
3 setwd("~/Users/dwhong/Desktop/R_code/boxplot")
4
5 iFILE <- "expression_values.csv"
6 gNAME <- "OGT"
7
8 data <- read.table(iFILE, sep = ',', stringsAsFactors = F, header = F, row.names = 1)
9 iDATA <- data.frame(cbind(t(data[rownames(data)%in%c(gNAME),]), t(data[1,])))
10 names(iDATA) <- c('value', 'type')
11 iDATA$value <- log2(as.double(levels(iDATA$value))[iDATA$value])
12 iDATA$type <- toupper(iDATA$type)
13 iDATA$type <- gsub("-", "", iDATA$type)
14
15 pdf(paste0(gNAME, ".pdf"))
16 ggplot(iDATA, aes(type, value)) + geom_boxplot(aes(fill = type)) +
17   xlab("") +
18   ylab(paste0(gNAME, ", mRNA Expression (RNA Seq V2 RSEM) (log2)")) +
19   ggtitle("") +
20   guides(fill = F) +
21   theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust = 0.5, size = 12, face = "bold"), axis.title.y = element_text(face = "bold")) +
22   invisible(dev.off())
23

```

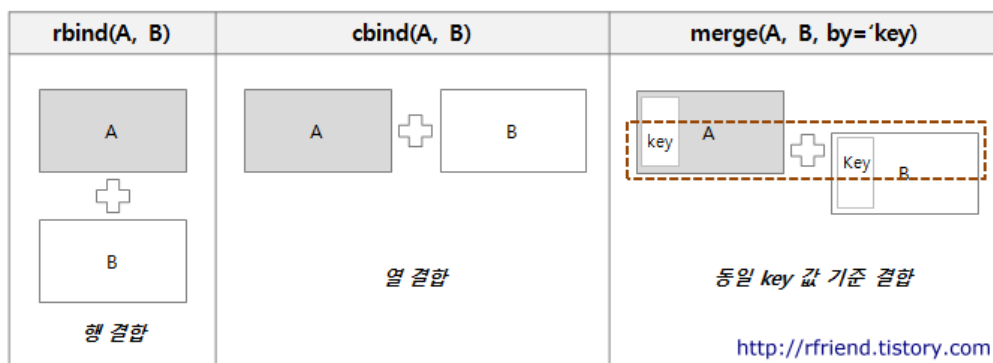
Environment

Object	Class	Size
clinical_data	data.frame	28 obs. of 4 variables
conds	Formal class DFrame	
count.data	matrix	40304 obs. of 7 variables
count.mat	matrix	40304 obs. of 6 variables
d	Large matrix	(61965 elements, 578.1 kB)
data	data.frame	21 obs. of 90 variables
dds	Large DESeqDataSet	(40304 elements, 26.3 MB)
expression	matrix	28 obs. of 2 variables
expression.table	matrix	1215 obs. of 51 variables
gene.exp	matrix	Large matrix (61965 elements, 578.1 kB)
iDATA	data.frame	90 obs. of 2 variables

> View (data)

gene_name	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16
TAT	0.8958	0.8672	1.2714	0.5408	2.512	0	1.2445	0.938	1.1568	1.7502	0.4668	1.4784	1.0107	0.7716	0.9
B3GNT6	0	1.9577	0	1.4952	4.0601	0.7176	0	2.2214	2.8377	1.7502	6.6538	2.8935	7.1735	6.7802	0
OGT	11.1784	11.3547	10.0148	10.6018	9.9306	11.2622	11.4902	11.181	10.5841	11.3482	11.251	11.4939	12.0301	11.2389	10.
MAT2A	11.5439	13.5627	11.6553	11.4857	12.4825	11.5974	12.2364	12.0035	12.2314	12.0627	11.3862	13.3059	13.9321	11.9116	11.
UGDH	8.9438	11.4139	10.4629	10.315	10.9648	9.3947	10.608	11.312	10.9958	10.9448	13.123	10.685	10.8975	10.7436	9.1
SCD	7.9392	11.2766	11.9138	10.8692	12.6961	9.4274	12.4205	12.9763	14.0059	14.3601	14.5127	12.5572	12.767	13.0293	8.5
LAP3	10.156	10.3446	10.4812	10.2748	10.3643	9.8901	10.5145	10.3429	10.4444	10.1088	10.6386	10.0503	10.415	10.2513	10.
PIGW	6.9248	7.5552	6.2265	6.0149	7.2984	6.7284	6.9807	7.1077	7.01	6.6761	8.4875	7.266	7.5251	7.5523	6.4
PFKFB4	6.6743	6.6424	6.7563	7.3152	8.2443	5.8181	8.1881	7.8961	8.098	8.4418	8.1826	7.602	7.6706	7.7783	5.9
NDUF811	10.8777	10.1862	10.2686	11.0024	11.5313	10.283	10.9653	10.7783	10.3895	10.993	10.6711	11.0263	10.3235	10.3984	10.
DHCR7	10.1141	9.5762	9.9227	9.7881	11.2564	9.3049	10.2361	10.9987	11.1945	11.5415	10.8288	10.9455	10.3004	10.4401	9.3
POLR2G	9.9999	9.5051	9.4266	9.9692	10.5754	9.5316	10.6829	9.881	9.9832	10.2068	9.8711	10.1701	9.5912	9.604	9.5
POLR3E	9.6127	9.9049	9.4406	9.158	9.6472	9.4355	9.6623	9.6767	9.7803	9.9581	9.8903	9.6352	9.8618	9.4608	9.6
B3GAT2	3.3651	4.2859	4.099	1.9385	2.3395	3.3589	3.5141	2.7424	3.5652	3.4687	3.2861	3.2419	4.2558	3.4806	3.5
RRM2	2.7212	3.6482	4.9062	5.7724	5.7632	1.1946	3.9746	6.2585	5.1217	3.0165	10.1432	4.4646	6.4682	3.7029	2.4

Variable (변수) <- value (값)



> View (iDATA)

	value	type
V2	11.1784	NORMAL
V3	11.3547	NORMAL
V4	10.0148	NORMAL
V5	10.6018	NORMAL
V6	9.9306	NORMAL
V7	11.2622	NORMAL
V8	11.4902	NORMAL
V9	11.1810	NORMAL
V10	10.5841	NORMAL
V11	11.3482	NORMAL
V12	11.2510	NORMAL
V13	11.4939	NORMAL
V14	12.0301	NORMAL
V15	11.2389	NORMAL
V16	10.8557	NORMAL
V17	10.3244	NORMAL
V18	10.1287	NORMAL
V19	11.4049	NORMAL

Global Environment	Format Class # of items
count.data	40304 obs. of 7 variables
count.mat	40304 obs. of 6 variables
d	Large matrix (61965 elements, 578.1 kB)
data	21 obs. of 90 variables
dds	Large DESeqDataSet (40304 elements, 26.3 MB)
expression	28 obs. of 2 variables
expression.table	1215 obs. of 51 variables
gene.exp	Large matrix (61965 elements, 578.1 kB)
iDATA	90 obs. of 2 variables
l2	51 obs. of 1 variable
label	51 obs. of 2 variables
m	40304 obs. of 19 variables

```
library(ggplot2)
```

drawBoxplot.R

```
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$value <- log2(as.double(levels(iDATA$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())
```

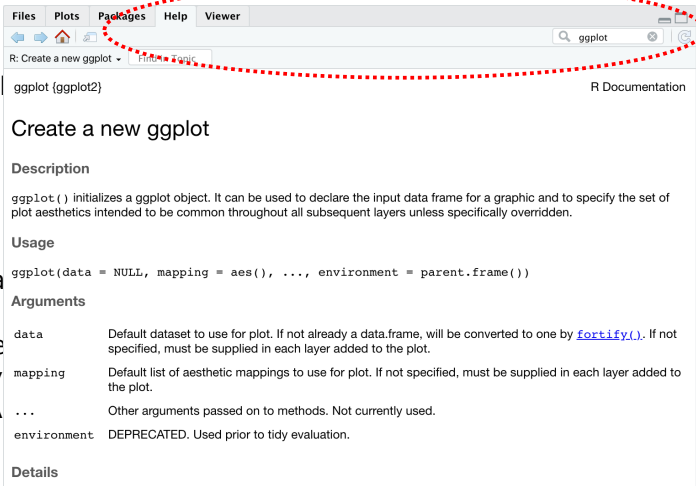
library(ggplot2)

setwd("/Users/dwhong/Desktop/

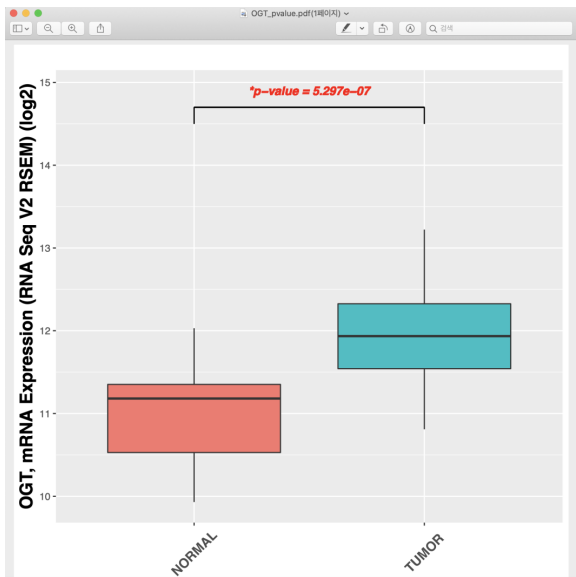
iFILE <- "expression_values.csv"
gNAME <- "OGT"

data <- read.table(iFILE, sep = ',',
iDATA <- data.frame(cbind(t(data
names(iDATA) <- c('value','type')
iDATA\$value <- log2(as.double(le
iDATA\$type <- toupper(iDATA\$ty
iDATA\$type <- gsub("_"," ",iDATA

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>

```
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 == "NORMAL",1]
```

```
tur <- iDATA[iDATA$type != "NORMAL",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, xend = 2, yend = iMax + interval)) +
```

```
  geom_segment(aes(x = 1, y = iMax, xend = 1, yend = iMax + interval)) +
```

```
  geom_segment(aes(x = 1, y = iMax + (interval*1), xend = 2, yend = iMax + (interval*1))) +
```

```
  geom_text(aes(x = 1.5, y = iMax + (interval*2), label = paste0("*p-value = ",pVal)), fontface = "bold.italic",  
size = 4, color = "red") +
```

```
  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())
```

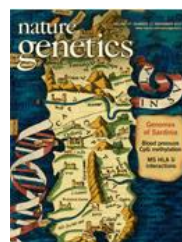

Intron retention is a widespread mechanism of tumor-suppressor inactivation

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

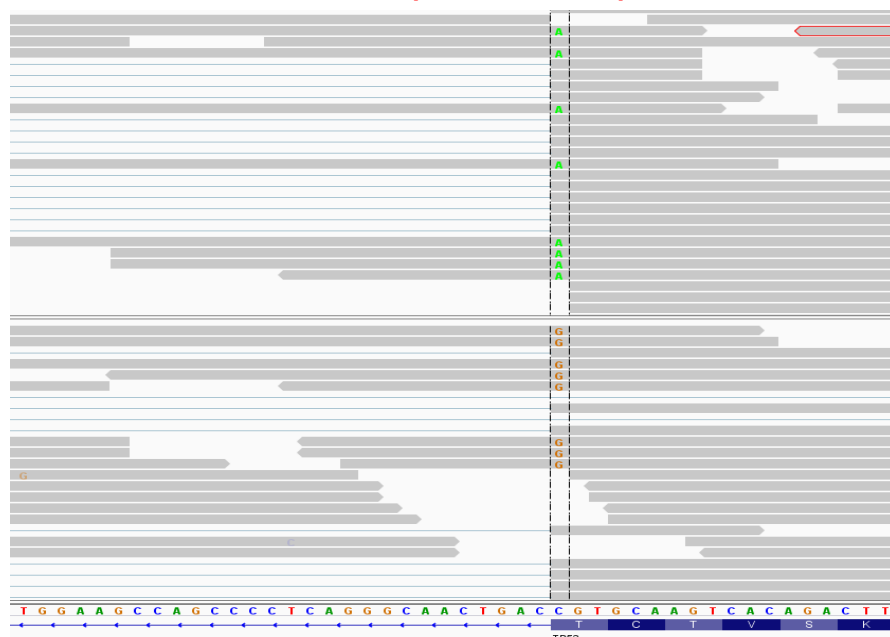
LUNG : TCGA-60-2712

(**C > A**, silent mutation)

BREAST : TCGA-AO-A12F

(**C > G**, silent mutation)

TP53 (chr17 : 7579312)



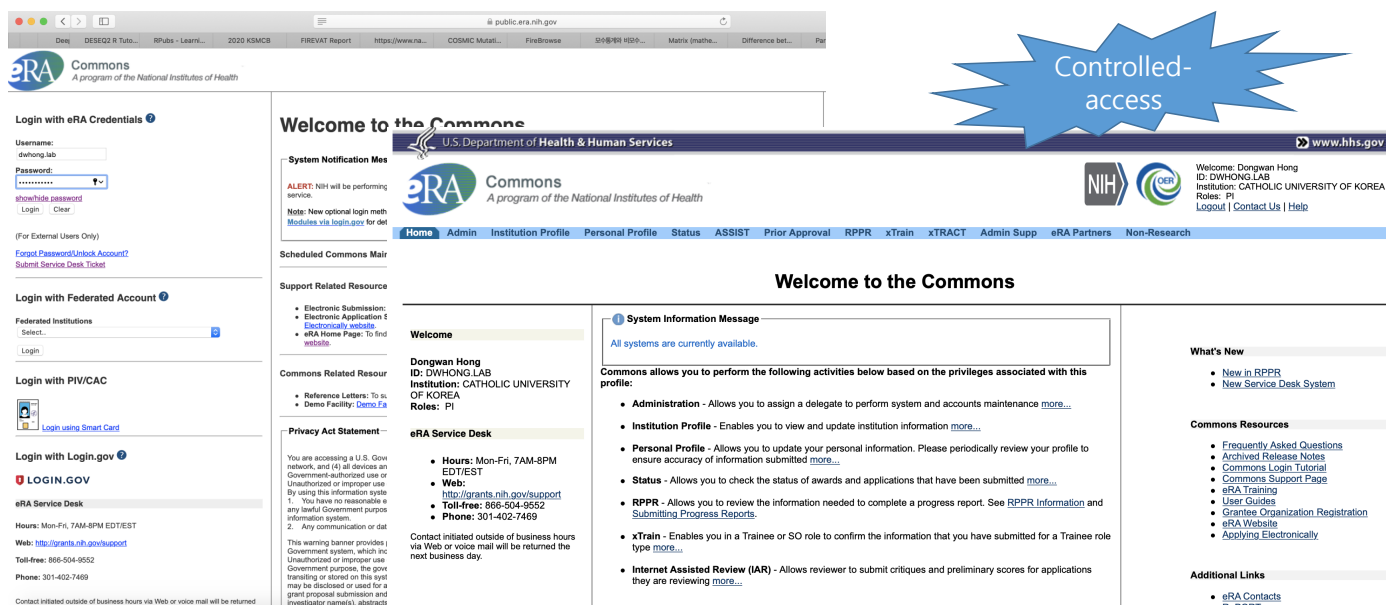
Cancer sequencing statistics

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)

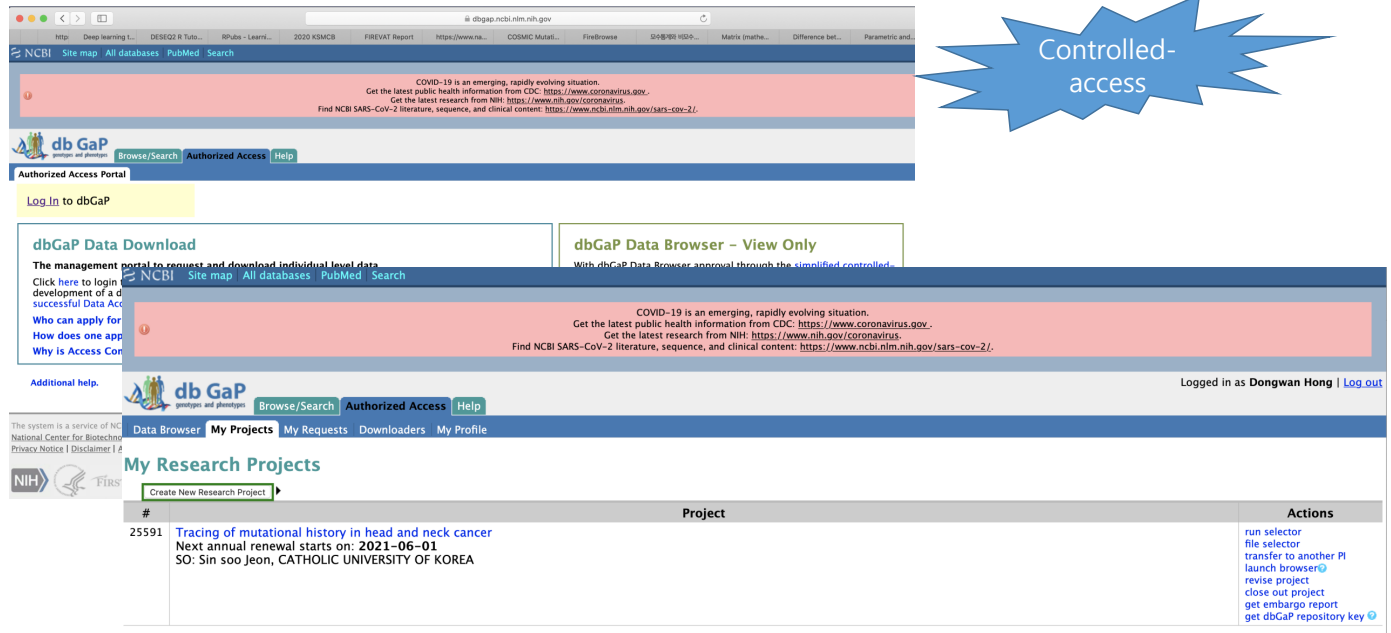
~7,040 sec = ~2 hours / 1 file
 1,812 patients -> 3,624 hours
 151 days

Getting an account of eRA commons



The screenshot displays the eRA Commons website interface. On the left, there are login options for eRA Credentials, Federated Institutions, PIV/CAC, and Login.gov. The main content area shows a 'Welcome to the Commons' message with a 'System Information Message' stating 'All systems are currently available.' Below this, a list of activities is provided, including Administration, Institution Profile, Personal Profile, Status, RPPR, xTrain, and Internet Assisted Review (IAR). On the right, there are sections for 'What's New' and 'Commons Resources'.

Getting an account of dbGaP



Controlled-access

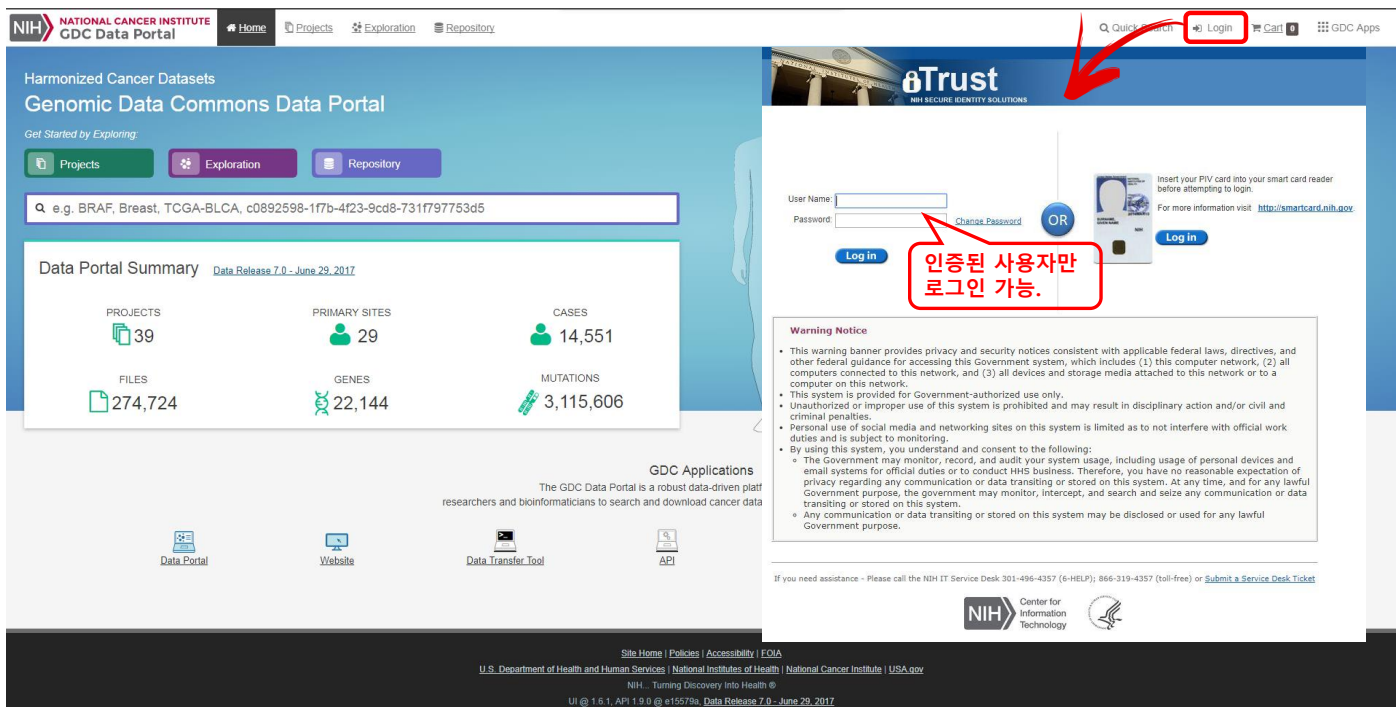
dbGaP Data Download

dbGaP Data Browser - View Only

My Research Projects

#	Project	Actions
25591	Tracing of mutational history in head and neck cancer Next annual renewal starts on: 2021-06-01 SO: Sin soo Jeon, CATHOLIC UNIVERSITY OF KOREA	run selector file selector transfer to another PI launch browser revise project close out project get embargo report get dbGaP repository key

GDC Data Portal



Genomic Data Commons Data Portal

Data Portal Summary

PROJECTS: 39

PRIMARY SITES: 29

CASES: 14,551

FILES: 274,724

GENES: 22,144

MUTATIONS: 3,115,606

GDC Applications

Log in

인증된 사용자만 로그인 가능.

Warning Notice

Data Portal Summary

DES RPubS - Learn... 2020 KSMCB FIREVAT Report https://www.na... COSMIC Mutati... FireBrowse 모수통계와 비모수... Matrix (mathe... Difference bet... Parametric and... dbGaP: Authori... Commons Hom... GDC +

You have logged in and have access to controlled data displayed within GDC visualizations. Please respect all project [Data Use Agreements](#) when downloading visualizations or data from the GDC Data Portal

NIH NATIONAL CANCER INSTITUTE GDC Data Portal Home Projects Exploration Analysis Repository Q Quick Search Manage Sets DWHONG.LAB Cart 0 GDC Apps

Harmonized Cancer Datasets
Genomic Data Commons Data Portal

Get Started by Exploring:

Projects Exploration Analysis Repository

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

Data Portal Summary Data Release 26.0 - September 08, 2020

PROJECTS 67	PRIMARY SITES 68	CASES 84,375
FILES 590,367	GENES 23,399	MUTATIONS 3,287,299

Cases by Major Primary Site

Primary Site	Cases (approx.)
Lung	10,000
Breast	9,000
Prostate	8,000
Colon Rectal	7,000
Stomach	6,000
Bladder	5,000
Esophagus	4,000
Head and Neck	3,000
Soft Tissue	2,000
Testis	1,000
Thyroid	1,000
Uterus	1,000
Adrenal Gland	1,000
Bile Duct	1,000
Bladder	1,000
Bone	1,000
Bone Marrow	1,000
Brain	1,000
Cervix	1,000
Colon Rectal	1,000
Esophagus	1,000
Eye	1,000
Head and Neck	1,000
Kidney	1,000
Liver	1,000
Lymph Nodes	1,000
Nervous System	1,000
Ovary	1,000
Pancreas	1,000
Pleura	1,000
Prostate	1,000
Skin	1,000
Soft Tissue	1,000
Stomach	1,000
Testis	1,000
Thymus	1,000
Thyroid	1,000
Uterus	1,000

NIH NATIONAL CANCER INSTITUTE GDC Data Portal Home Projects Exploration Repository Q Quick Search NCKKOREA Cart 0 GDC Apps

Harmonized Cancer Datasets
Genomic Data Commons Data Portal

Get Started by Exploring:

Projects Exploration **Repository**

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

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

Download Token Logout

파일 이름(N): gdc-user-token.2017-08-09T00-16-26.350Z.txt
파일 형식(T): Text Document

다운로드 하기 저장(S) 취소

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

NATIONAL CANCER INSTITUTE
GDC Data Portal

[Home](#)
[Projects](#)
[Exploration](#)
[Repository](#)

[Login](#)
[Cart](#)
0

[GDC Apps](#)

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

[Projects](#)
[Exploration](#)
[Repository](#)

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

Cases by Primary Site

Primary Site	Cases
Adrenal Gland	10
Bile Duct	10
Bladder	10
Blood	10
Bone	10
Bone Marrow	10
Brain	10
Breast	10
Cervix	10
Colorectal	10
Esophagus	10
Eye	10
Head and Neck	10
Kidney	10
Liver	10
Lung	10
Lymph Nodes	10
Nervous System	10
Ovary	10
Pancreas	10
Pleura	10
Prostate	10
Skin	10
Soft Tissue	10
Stomach	10
Testis	10
Thyroid	10
Uterus	10

[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](#)

[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](#)

NATIONAL CANCER INSTITUTE
Genomic Data Commons

[CCG Web Site](#)
[Contact Us](#)
[Launch Data Portal](#)

[About the GDC](#)
[About the Data](#)
[Analyze Data](#)
[Access Data](#)
[Submit Data](#)
[For Developers](#)
[Support](#)

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.

- [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](#)

Access Data

- [Data Access Processes and Tools](#)
- [Data Access Policies](#)
- [GDC Community Tools](#)
- [Documentation for Data Portal](#)
- [Documentation for Data Transfer Tool](#)
- [GDC Data Portal](#)
- [GDC Data Transfer Tool](#)
- [Launch Data Portal](#)
- [Obtaining Access to Controlled Data](#)

Get Started

- [GDC Data Transfer Tool Guide](#)
- [Get the GDC Data Transfer Tool](#)


```
[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.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.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.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.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.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.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.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.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.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.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-abb0-48aa-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.vcf.gz
drwxr-xr-x 2 root root 4096 Aug 9 2017 logs
[root@works205 GDCDataPortal]# █
```

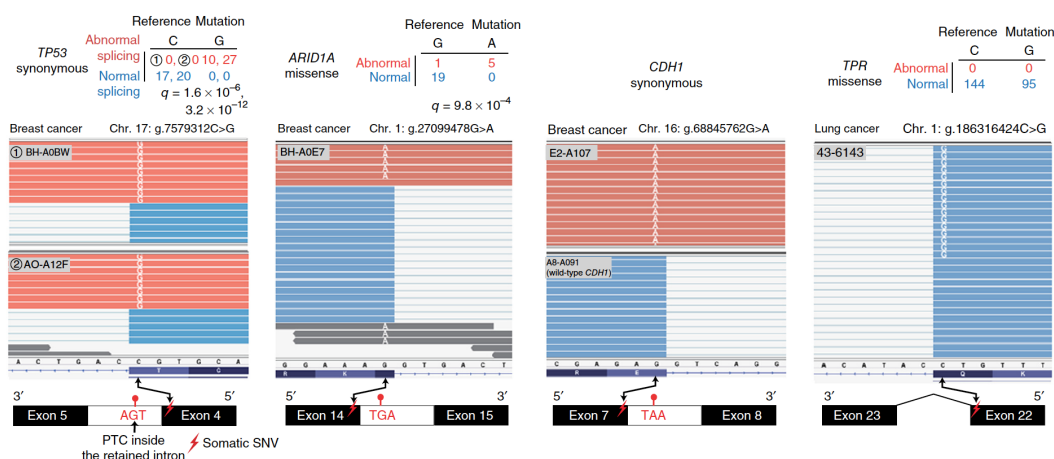
Tumor suppressor genes with LBMEs causing intron retention

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

Identification of SNVs disrupting splicing

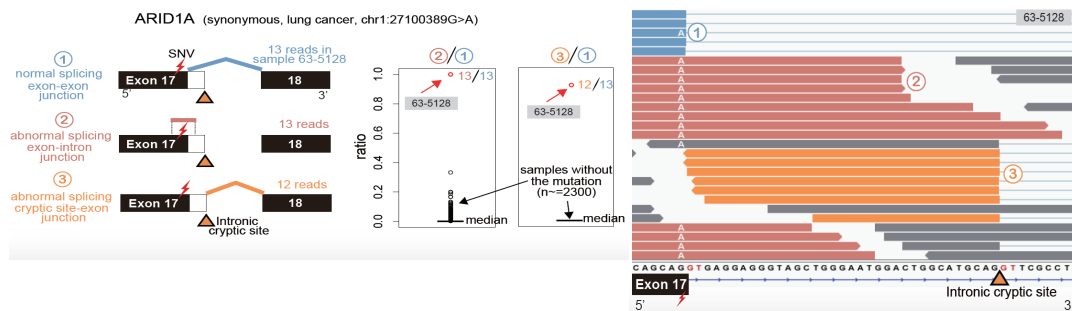
- **Allele-specific splicing analysis** of LBMEs causing intron retention
 - Statistical significance test using Fisher's exact test
 - Normal Splicing: Wild Type
 - Abnormal Splicing: Mutant Allele

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



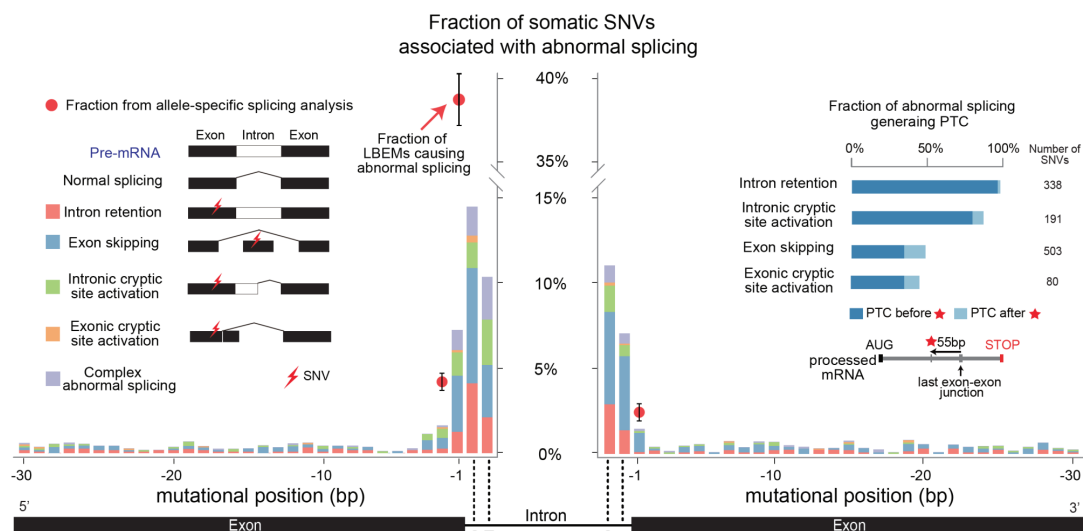
Identification of SNVs disrupting splicing

- 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



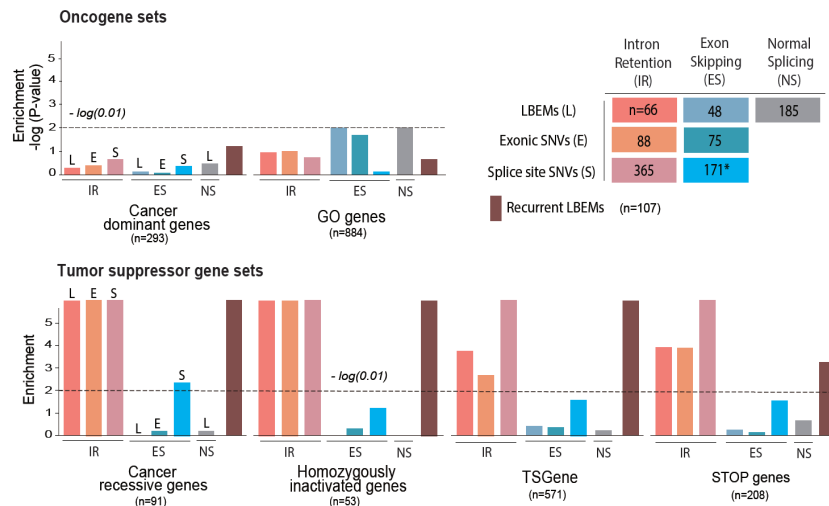
Frequently altered splicing by last-base exonic mutations (LBEM)s

- Positional association of somatic SNVs with abnormal splicing



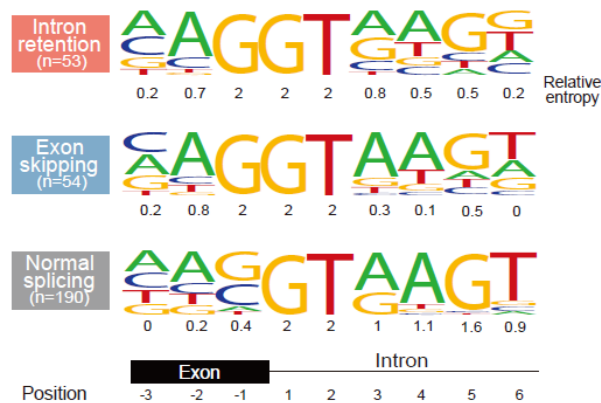
Enrichment of intron retention-causing SNVs in TSGs

- 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

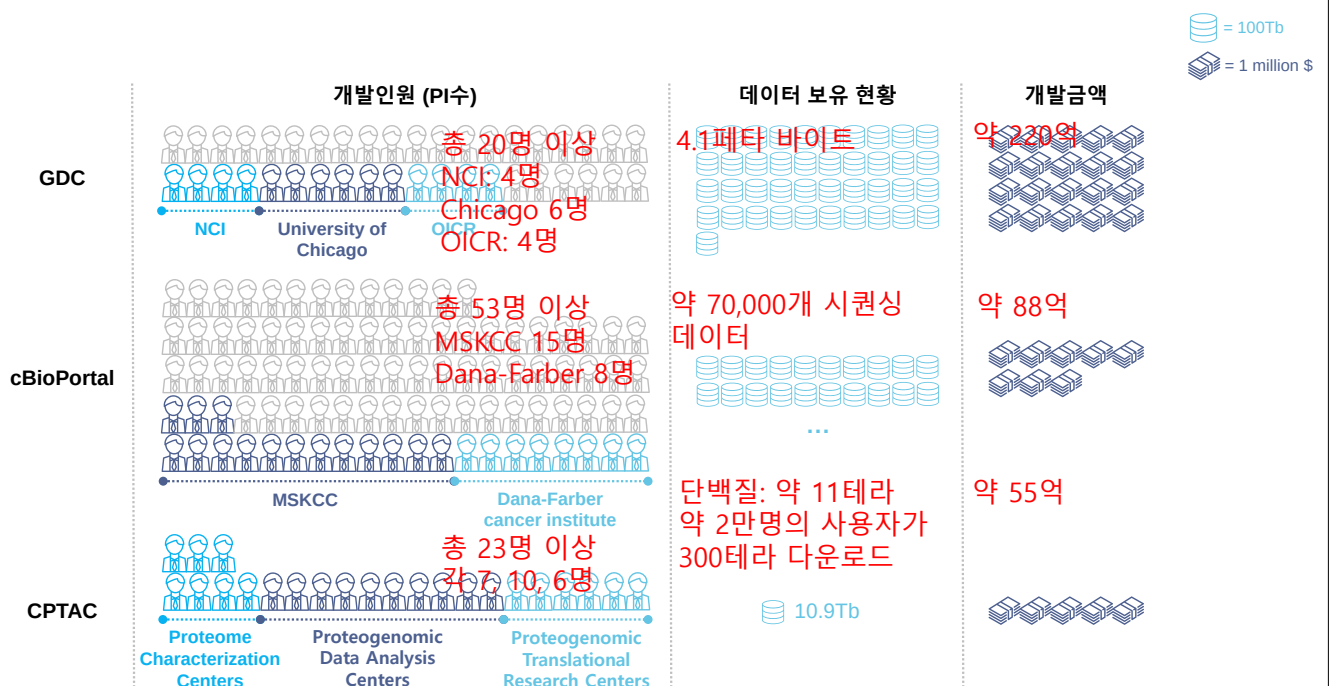
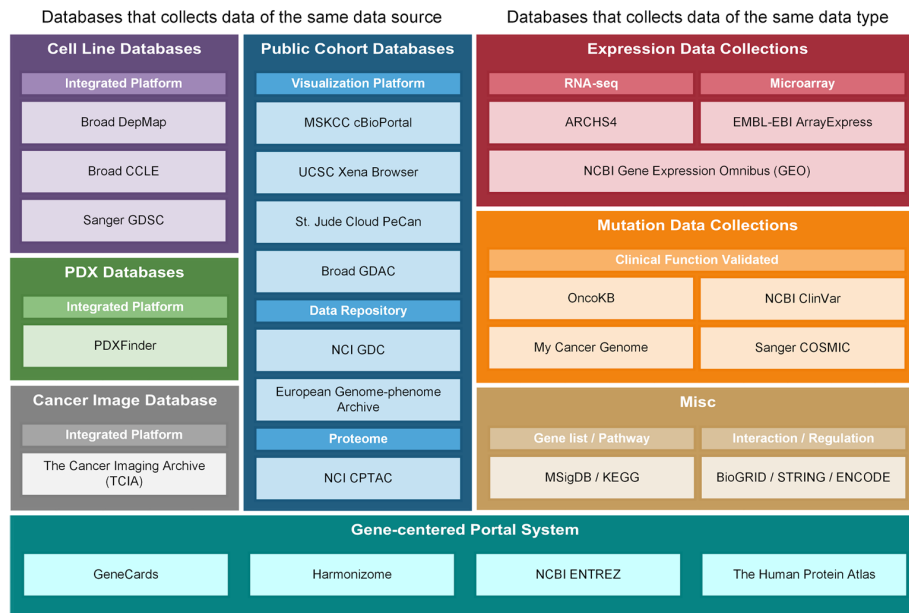
Big data is defined by these specific attributes;
four Vs.



Two more Vs are necessary for **cancer genomic data**



Genomic databases in cancer research

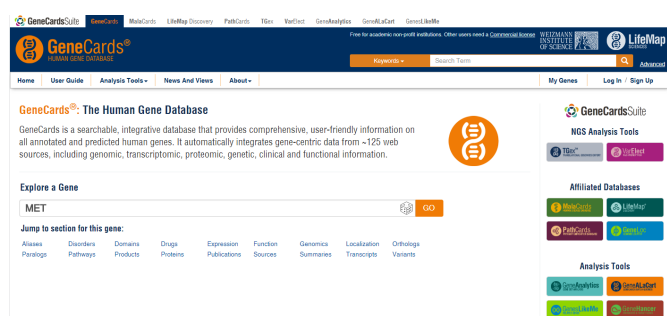


Trending analytical approaches

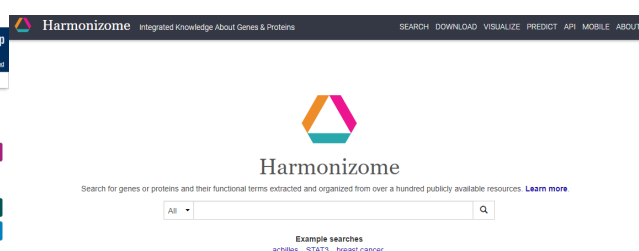
Keywords:

- **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



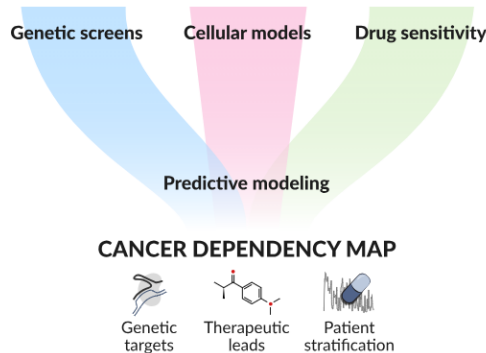
Stelzer, Gil, et al., *Current protocols in bioinformatics* 2016.



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

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

Systematic experiments

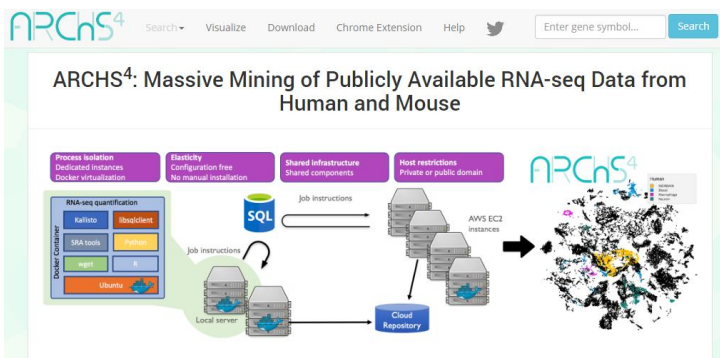


To date DepMap has profiled more than 500 cell lines. Over the next several years we will greatly expand the diversity of cell lines profiled for genetic vulnerabilities with quarterly data release. Additionally, limited drug sensitivity data are available.

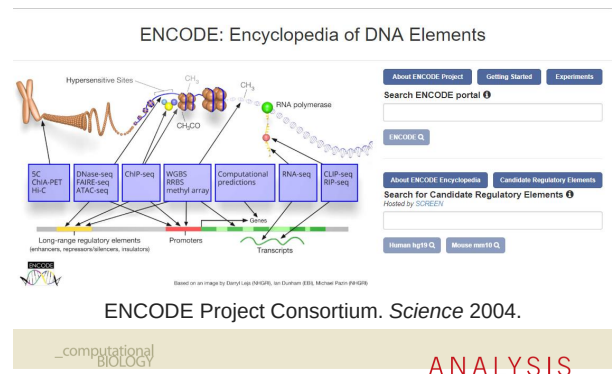


DepMap; Tsherniak, Aviad, et al., *Cell* 2017.

Standardization for machine learning & deep-learning



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



ENCODE Project Consortium. *Science* 2004.

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

Babak Alipanahi^{1,2,4}, Andrew Delong^{1,6}, Matthew T Weirauch³⁻⁵ & Brendan J Frey¹⁻³

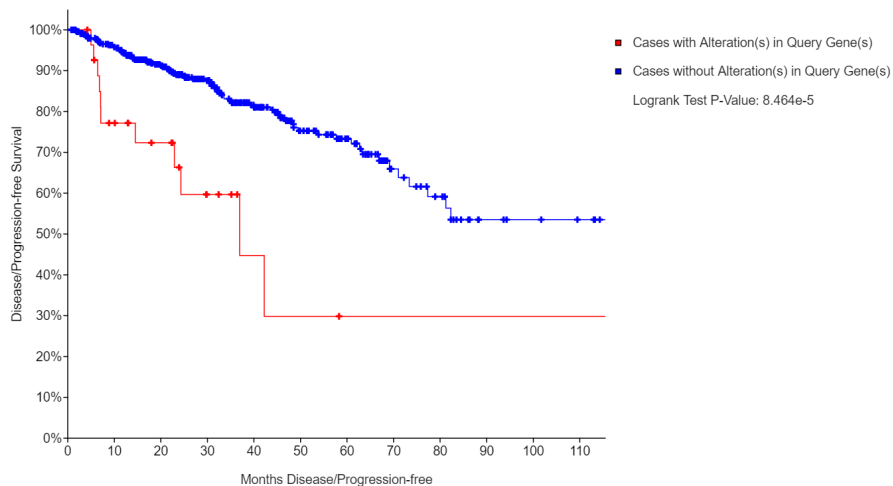
Alipanahi, Babak, et al., *Nature biotechnology* 2015

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.

89

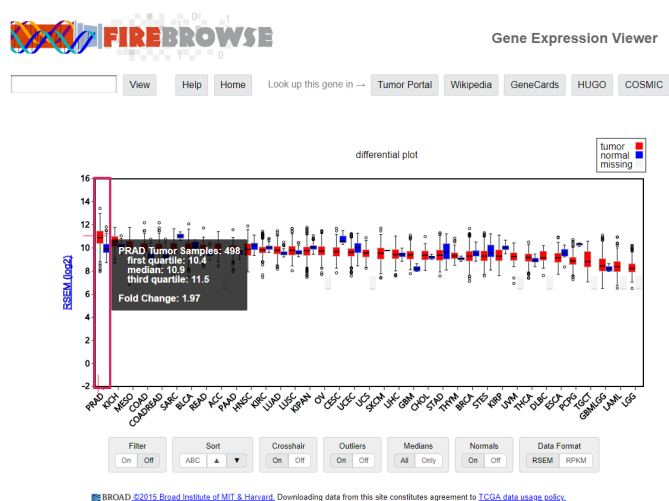
Usage of genomic databases in cancer research - Scenario #1

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

<http://firebrowse.org>

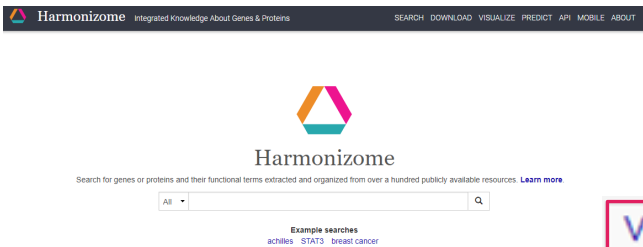


We confirmed expression of gene X at both cell line (**CCLE**) and patient cohort level (**Broad GDAC**). Expression level of gene X in prostate cancer was higher than in other cancers.



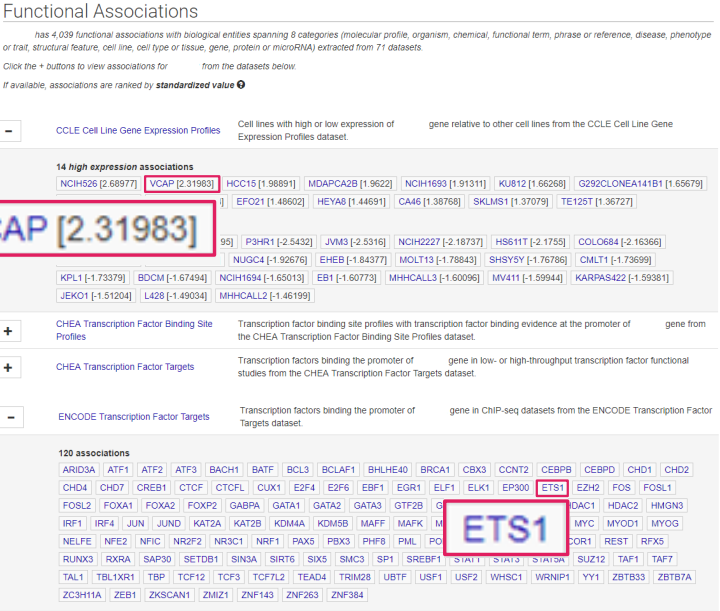
90

Usage of genomic databases in cancer research - Scenario #1

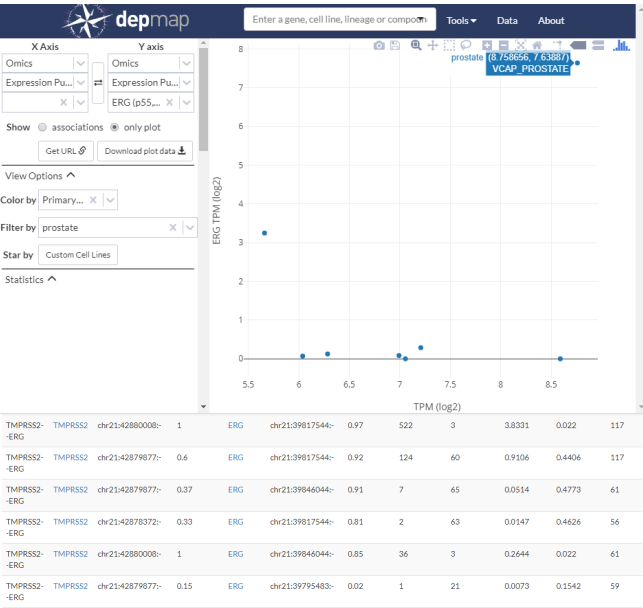


<http://amp.pharm.mssm.edu/Harmonizome/>

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



<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

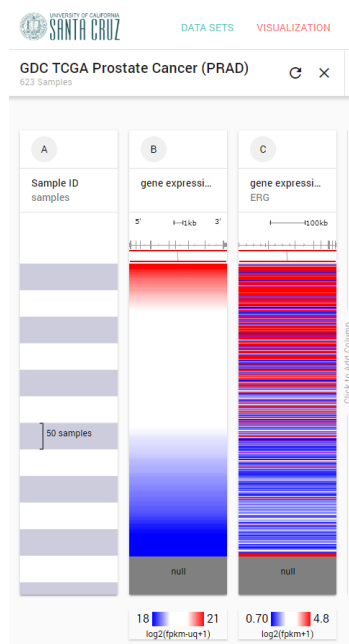
Molecular Subtyping of Primary Prostate Cancer Reveals Specific and Shared Target Genes of Different ETS Rearrangements^{1,2}

Paulo Paulo^{*,1,2,3,4}, Francis R. Ribeiro^{*,1,2,3,4}, Joana Santos^{*,1}, Diana Mesquita^{*,1}, Mafalda Almeida^{*,1}, João D. Barros-Silva^{*,1}, Harri Ikonen^{*,1}, Rui Henrique^{*,1,2}, Carmen Jardim^{*,1,2}, Anita Sveen^{*,1}, Ian G. Mills^{*,1,2}, Rolf I. Skotheim^{*,1}, Ragnhild A. Lothe^{*,1} and Manuel R. Teixeira^{*,1,2,3,4}

¹Department of Genetics, Portuguese Oncology Institute, Porto, Portugal; ²Group of Cancer Genetics, Research Centre of the Portuguese Oncology Institute, Porto, Portugal; ³Department of Cancer Prevention, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway; ⁴Centre for Cancer Biomedicine, University of Oslo, Oslo, Norway; ⁵Group of Cancer Epigenetics, Research Centre of the Portuguese Oncology Institute, Porto, Portugal; ⁶Centre for Molecular Medicine Norway, Nordic European Molecular Biology Laboratory Partnership, Biotechnology Centre, University of Oslo, Oslo, Norway; ⁷Department of Pathology, Portuguese Oncology Institute, Porto, Portugal; ⁸Institute of Biomedical Sciences Abel Salazar, University of Porto, Porto, Portugal; ⁹Department 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.



An example of cancer genome study

Survival graph using omics data in TCGA and clinical data in UCSC

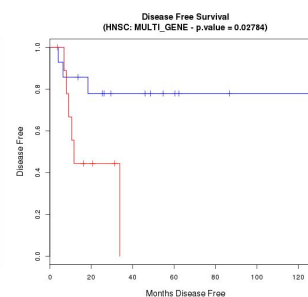
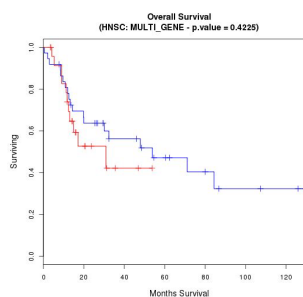
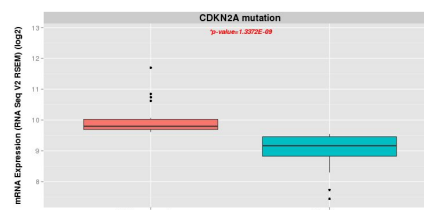
Showing 1 to 4 of 4 entries

TCGA Data

Show 100 entries

Cancer Type	Hugo Symbol	Chrom	Start position	Variant Classification	Reference Allele	Tumor Seq Allele2	cDNA Change	Codon Change
HNSC	CDKN2A	9	21968242	Splice_Site	C	T		
HNSC	CDKN2A	9	21971000	Nonsense_Mutation	CCCA	CDKN2A	5	41271120
HNSC	CDKN2A	9	21971028	Nonsense_Mutation		CDKN2A	9	21971000
HNSC	CDKN2A	9	21971029	Nonsense_Mutation		CDKN2A	9	21971028
HNSC	CDKN2A	9	21971036	Missense_Mutation		CDKN2A	9	21971029
HNSC	CDKN2A	9	21971096	Nonsense_Mutation		CDKN2A	9	21971096
HNSC	CDKN2A	9	21971120	Nonsense_Mutation		CDKN2A	9	21971120
HNSC	CDKN2A	9	21971153	Nonsense_Mutation		CDKN2A	9	21971153
HNSC	CDKN2A	9	21971186	Nonsense_Mutation		CDKN2A	9	21971186
HNSC	CDKN2A	9	21971208	Splice_Site				

Showing 1 to 10 of 10 entries



An example of cancer genome study

Showing 1 to 4 of 4 entries

TCGA Data

Show 100 entries

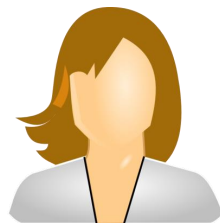
Cancer Type	Hugo Symbol	Chrom	Start position	Variant Classification	Reference Allele	Tumor Seq Allele2	cDNA Change	Codon Change	Protein Change	Recurrent Count
HNSC	CDKN2A	9	21968242	Splice_Site	C	T				5
HNSC	CDKN2A	9	21971000	Nonsense_Mutation	C	A	c.358G>T	c.(358-360)GAG>TAG	p.E120*	5
HNSC	CDKN2A	9	21971028	Nonsense_Mutation	C	T	c.339G>A	c.(329-330)TGG>TGA	p.W110*	6
HNSC	CDKN2A	9	21971029	Nonsense_Mutation	C	T	c.329G>A	c.(328-330)TGG>TAG	p.W110*	4
HNSC	CDKN2A	9	21971036	Missense_Mutation	C	A	c.322G>T	c.(322-324)GAT>TAT	p.D108Y	2
HNSC	CDKN2A	9	21971096	Nonsense_Mutation	C	A	c.262G>T	c.(262-264)GAC>TAA	p.Q87*	4
HNSC	CDKN2A	9	21971120	Nonsense_Mutation	G	A	c.238C>T			4
HNSC	CDKN2A	9	21971153	Nonsense_Mutation	C	A	c.205G>T			4
HNSC	CDKN2A	9	21971186	Nonsense_Mutation	G	A	c.172C>T			4
HNSC	CDKN2A	9	21971208	Splice_Site	C	T	c.151G>T			4

Showing 1 to 10 of 10 entries

NCC Data

Show 100 entries

Sample Name	Hugo Symbol	Chrom	Start position	Variant Classification	Reference Allele	Tumor Seq Allele2
HNT-070	CDKN2A	chr9	21969532	Intron	A	G
HNT-070	CDKN2A	chr9	21971082	In_Frame_Ins	-	TCC
HNT-081	CDKN2A	chr9	21971074	In_Frame_Del	ACCAGCGTGTCCAGGAAGCCCTCCCGGGCAGCGTCGTGCACGGGT	-
HNT-095	CDKN2A	chr9	21971096	Missense_Mutation	C	T



A 연구원

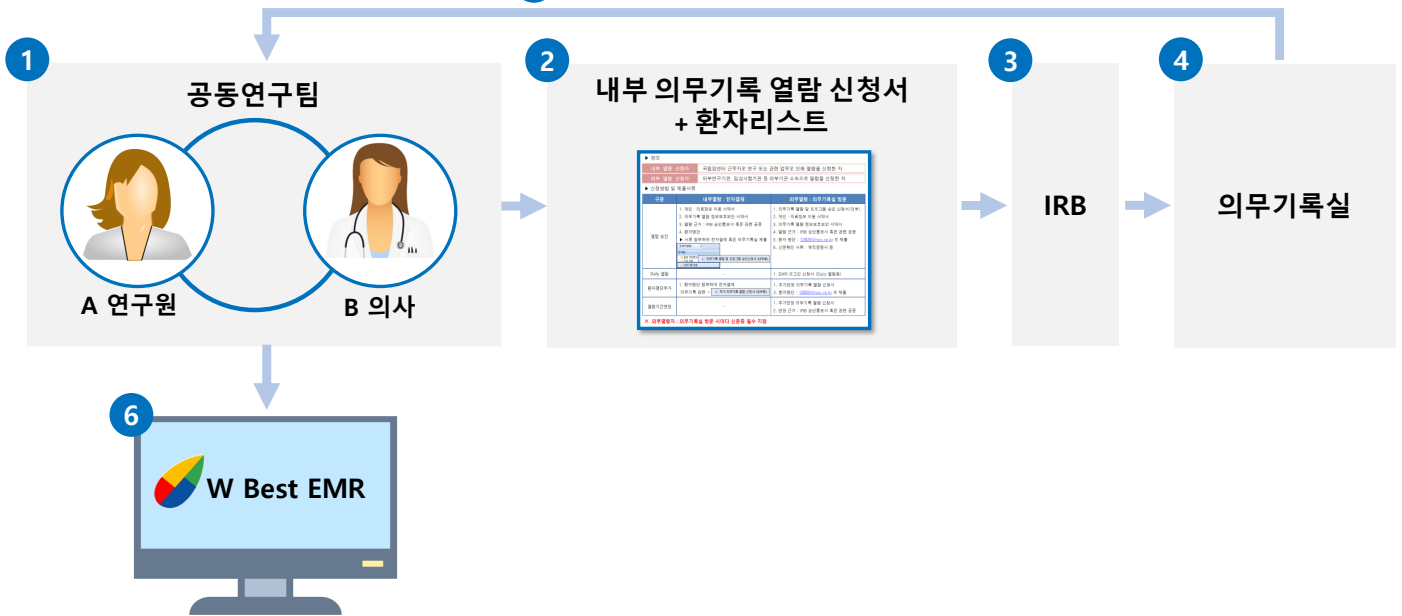
연구과제질문

- (1) 2010년부터 2016년 사이에 내원한
- (2) HNSCC 환자 중
- (3) p16 마커의 변이 리스트와 PIK3CA의 발현 검색

EMR을 접속하여 (1), (2)에 해당하는 환자 리스트를 얻어야 함.
 임상유전체분석실의 오믹스 데이터베이스 시스템에서 (3)의 정보를 제공 받아야 함

현재 연구용 의무기록 열람 신청 과정

5 연구용 의무기록 열람 권한 부여



보완해야할 현재 연구용 의무기록 열람 과정



분석 가능한 형식으로
의무기록 다운로드

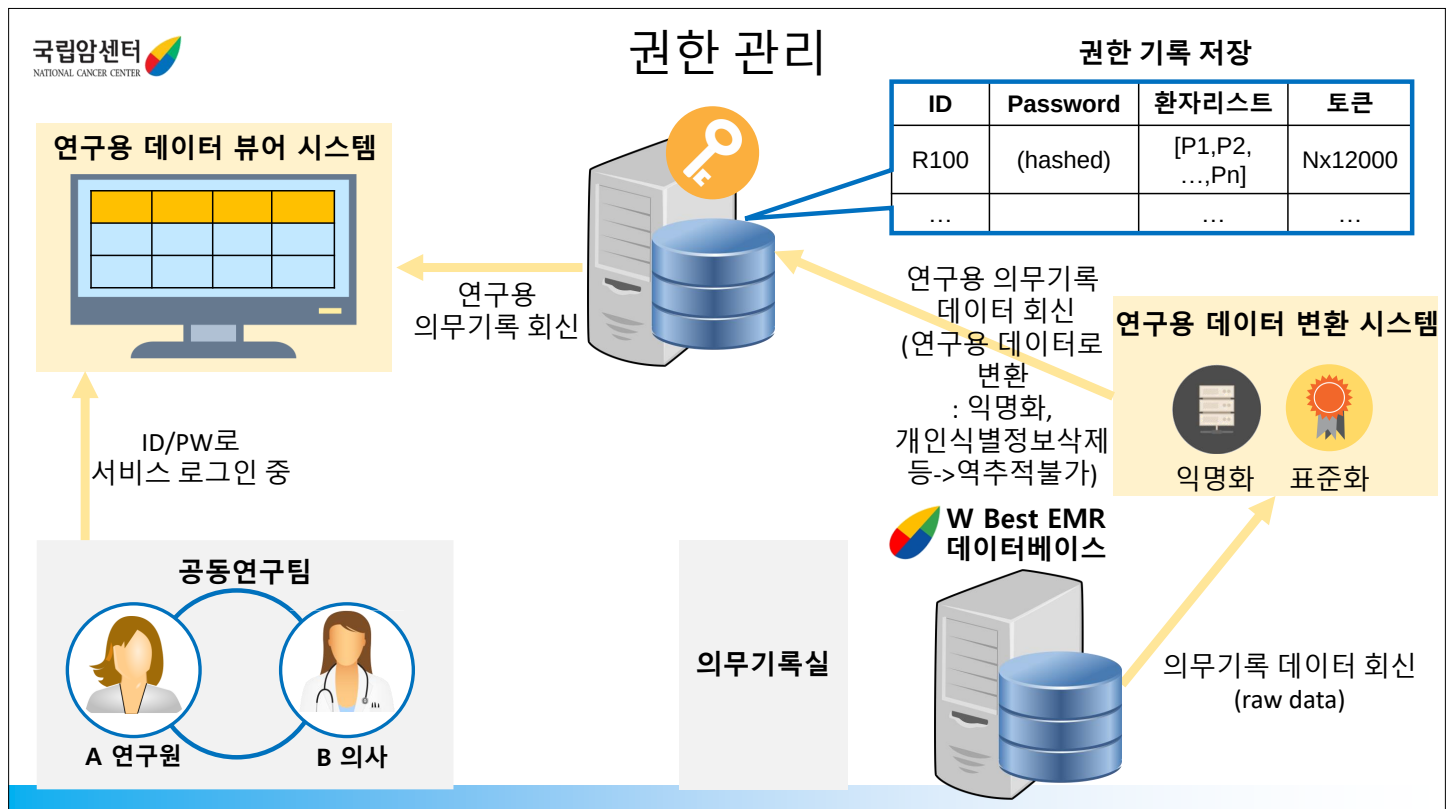
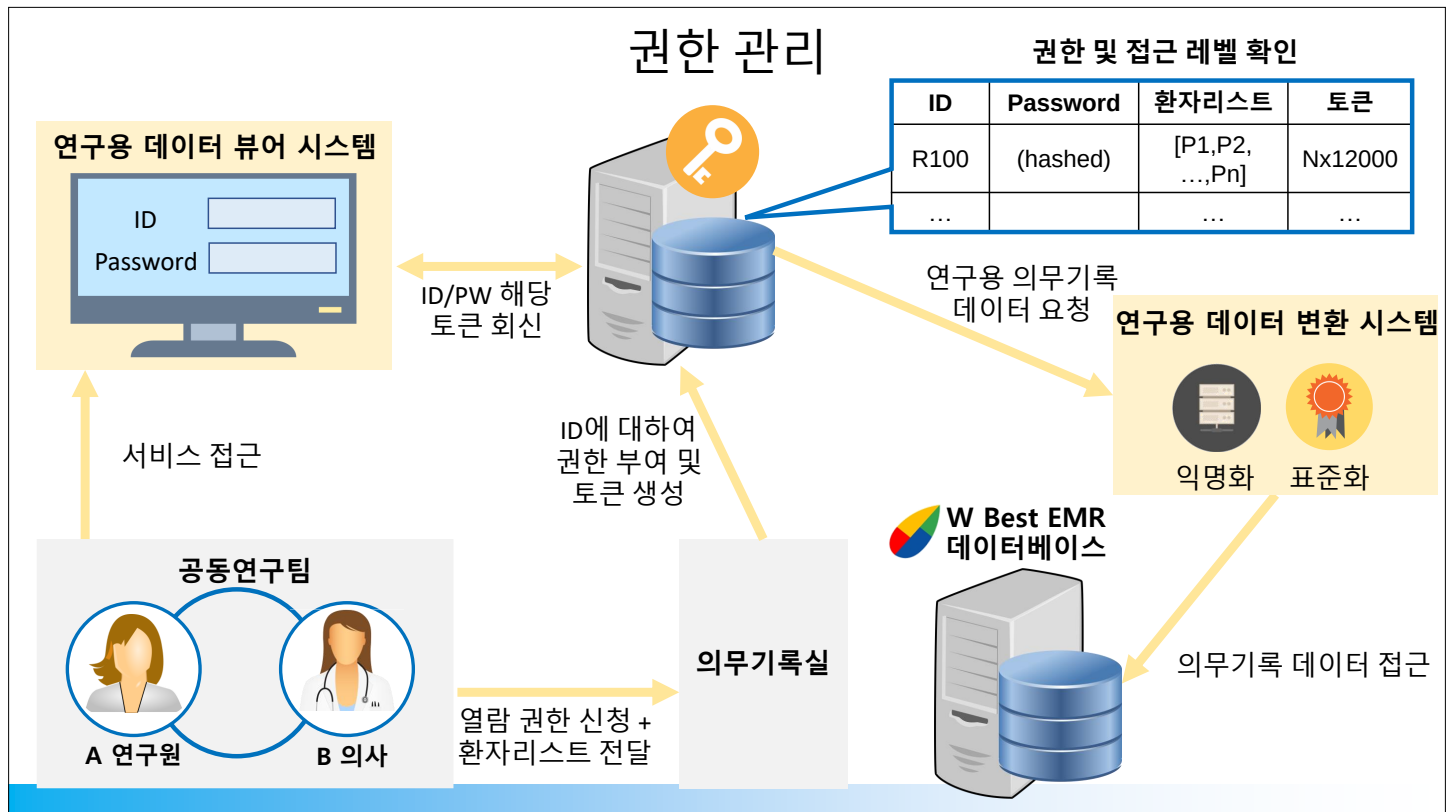
불가능

연구에 불필요한
개인식별정보

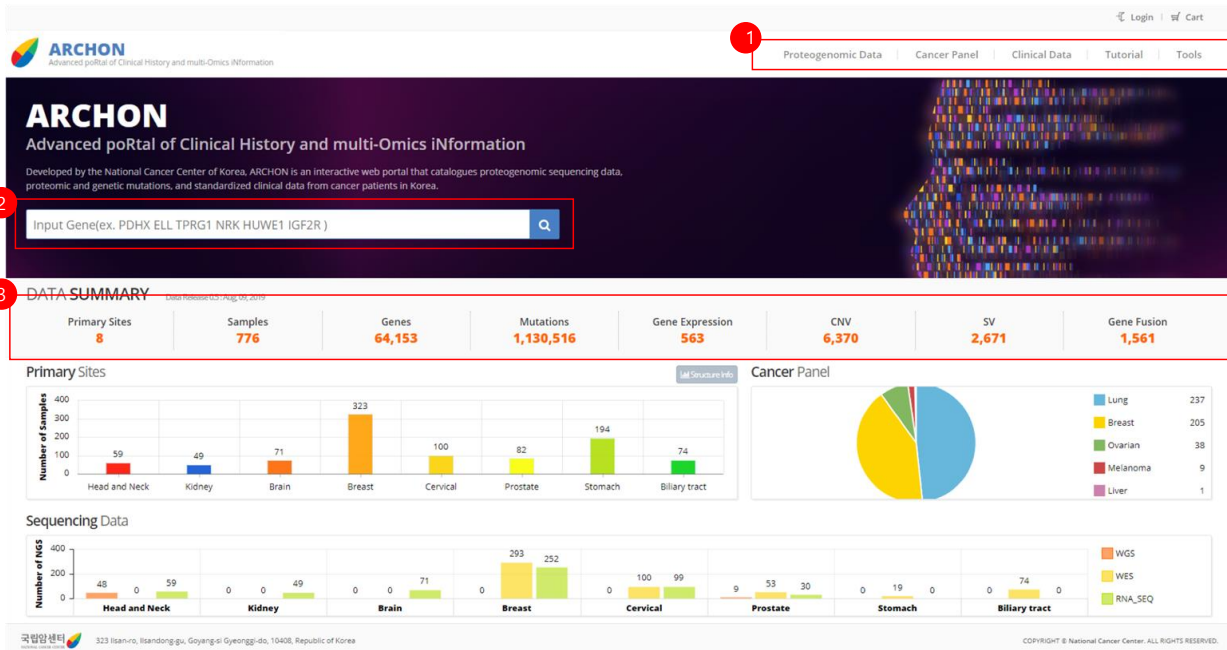
노출

역 추적 가능
의무기록

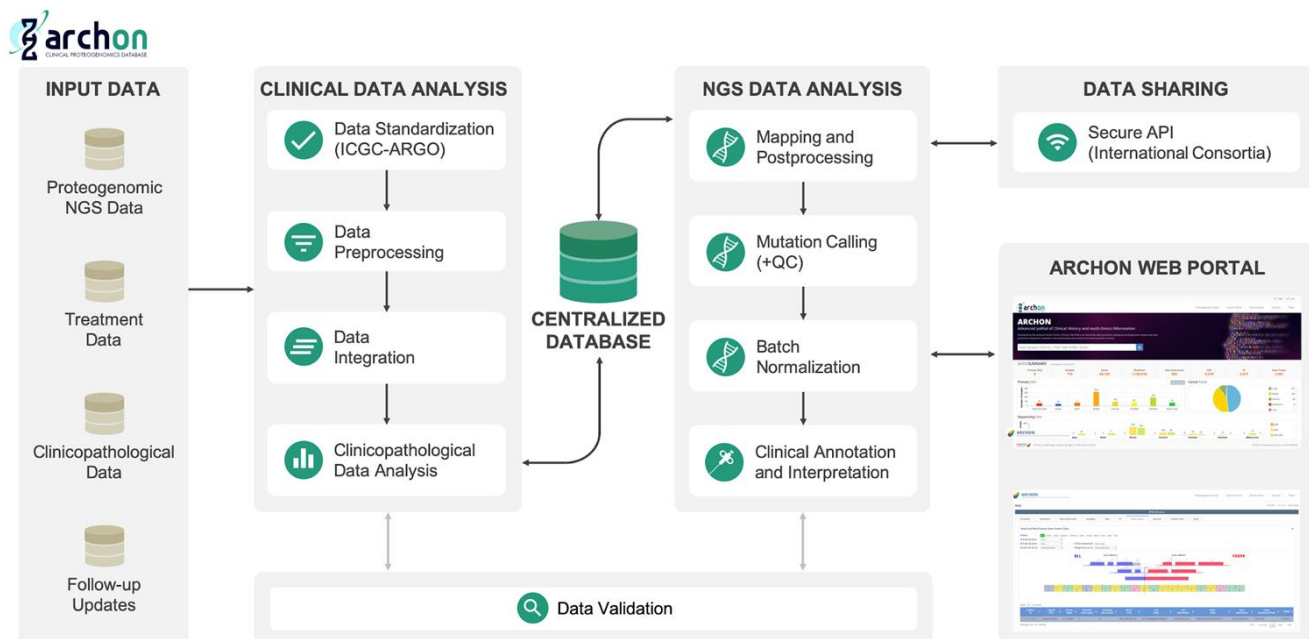
노출



Introducing ARCHON



ARCHON Under the Hood



QBE by metadata

ARCHON
Advanced portal of Clinical History and multi-Omics Information

Sample List

Show 10 entries

Patient Identifier

- CGAB-01-00001
- CGAB-01-00002
- CGAB-01-00003
- CGAB-01-00004
- CGAB-01-00005
- CGAB-01-00006
- CGAB-01-00007
- CGAB-01-00008
- CGAB-01-00009
- CGAB-01-00010

Showing 1 to 10 of 59 entries

Sample Summary

CGAB-01-00005 - Clinical

Age at Diagnosis: 54
Nationality: KOREA, REPUBLIC OF
Cancer Type: TUBULOUS CARCINOMA
Current Status: Alive
First Smoking: 0
Second Smoking Year: 0
Pack Per Day Smoking: 0.0
Any Cancer Stage Edition: 7th
Lymph Status: Negative

Person Gender Test Type: Male
Race Category Test: ACAN
Death Date: -
Current Status: Alive
Start Year Alcohol: 2003
Stop Year Alcohol: 4073
Frequency Alcohol: 5.0
Primary Site of Disease Anatomic Site Name: Stomach
Tumor Hist Status: Negative
International Classification of Diseases, Tenth Revision: C23.9
Tumor Hist Status: Negative
International Classification of Diseases, Tenth Revision: C23.9
Tumor Hist Status: Negative

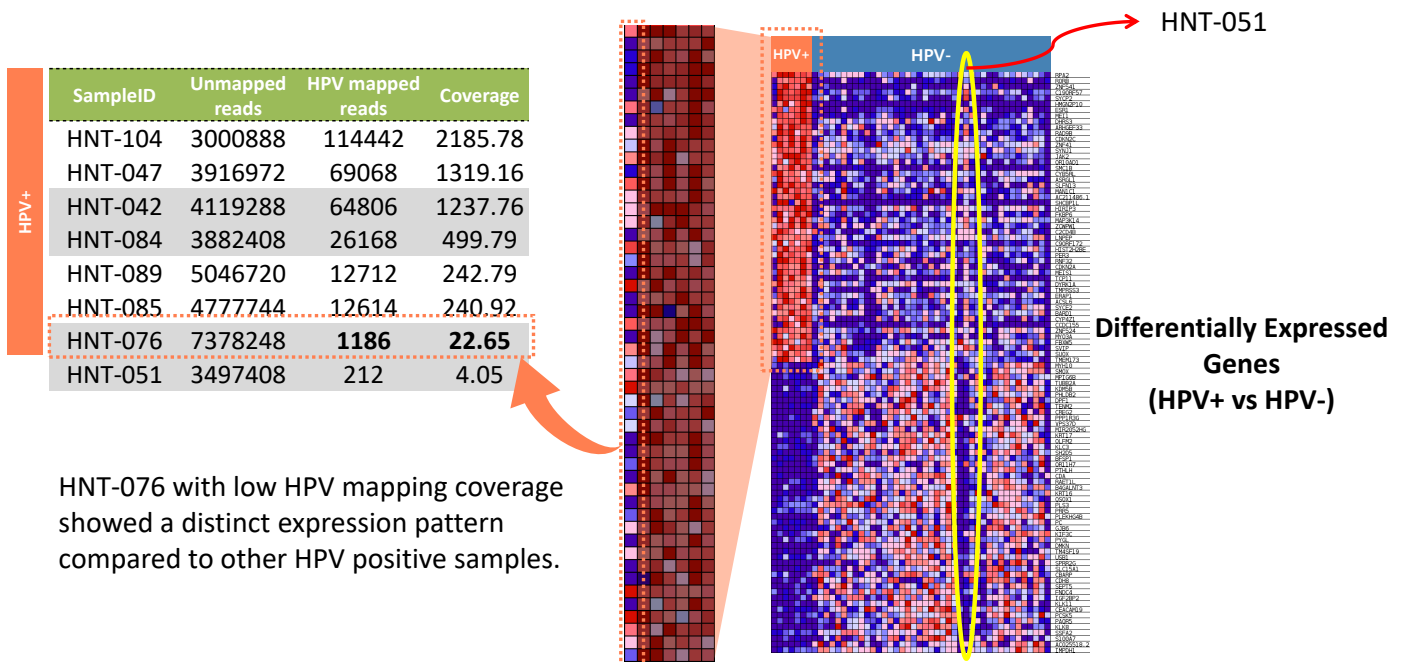
International Classification of Diseases for Oncology: 80713
Third Edition ICD-O-3.1 Site Code
Neoplasms American Joint Committee on Cancer Clinical TNM: Regional lymph node N1 Stage
Head and Neck Tumor Pathology T Stage: pT4a
Head and Neck Tumor Pathology M Stage: M0
Head and Neck Tumor Pathology TNM Stage: IIIA

CGAB-01-00005 - Metabolomics Stage 20

Metabolomics Stage 20

CGAB-01-00005 - Mutations List

SAMPLE ID	GENE NAME	CHROMOSOME	START POSITION	END POSITION	MUTATION TYPE	REFERENCE ALLELE	ALTERNATIVE ALLELE	PROTEIN CHANGE	BIOMAP	TESSARS	CLINVAR	CONCISE
CGAB-01-00005	BRCA1	17	41178101	41178101	Point SNV G to A	G	A	p.R1797G	0.00	-	0	-
CGAB-01-00005	BRCA1	17	41178101	41178101	Point SNV G to A	G	A	p.R1797G	0.00	-	0	-
CGAB-01-00005	BRCA1	17	41178101	41178101	Point SNV G to A	G	A	p.R1797G	0.00	-	0	-
CGAB-01-00005	BRCA1	17	41178101	41178101	Point SNV G to A	G	A	p.R1797G	0.00	-	0	-
CGAB-01-00005	BRCA1	17	41178101	41178101	Point SNV G to A	G	A	p.R1797G	0.00	-	0	-
CGAB-01-00005	BRCA1	17	41178101	41178101	Point SNV G to A	G	A	p.R1797G	0.00	-	0	-
CGAB-01-00005	BRCA1	17	41178101	41178101	Point SNV G to A	G	A	p.R1797G	0.00	-	0	-
CGAB-01-00005	BRCA1	17	41178101	41178101	Point SNV G to A	G	A	p.R1797G	0.00	-	0	-
CGAB-01-00005	BRCA1	17	41178101	41178101	Point SNV G to A	G	A	p.R1797G	0.00	-	0	-
CGAB-01-00005	BRCA1	17	41178101	41178101	Point SNV G to A	G	A	p.R1797G	0.00	-	0	-
CGAB-01-00005	BRCA1	17	41178101	41178101	Point SNV G to A	G	A	p.R1797G	0.00	-	0	-



CRDW (Clinical Research Data Warehouse)

환자
임상진단: R/O Pancreatic Cancer and Duodenal Cancer
중환자실 진단: 1
From 1) Lymph node, #13, biopsy: Negative (E0H)
From 2) Lymph node, #16, biopsy: Negative (E0H)
발견 조직은 신장 상부의 duodenum, common bile duct, pancreas, gallbladder를 포함하고 있는 pylorus s
충담관, 십이지장, 위장을 포함하여 결핵에 보았을 때 위장 내부는 경계가 불명확한 점막의 종괴가 관찰됨.
합해 온 담낭의 크기는 8x3x2cm이며 내강을 따라 열어보았을 때 내부에는 velvety한 양상을 잘 유지하고 있는

CELLULAR_DIFFERENTIATION
대표어 수종 간
moderately differentiated

SIZE_OF_TUMOR
대표어 수종 간
3.2

EXTENT_OF_INVASION
대표어 수종 간
Invasion of adjacent large vessels: absent

BLOOD_VESSEL_INVASION
대표어 수종 간
absent

SURGICAL_MARGINS
대표어 수종 간
free (safety margin: 2cm)

1) Size of tumor : **3.2x2x1cm**
2) Extent of invasion:
Invasion of : peripancreatic tissue and duodenum

Variable	Value
SIZE_OF_TUMOR	3.2
INVASION	Invasion of adjacent large vessels: absent

Example of **NLP-based data extraction** on clinicopathology report.

Our CRDW system automatically updates all clinicopathology reports for all inpatients undergoing surgery

CRDW Use Case

Figure 1. Cumulative Incidence of Diabetes by Cancer Status

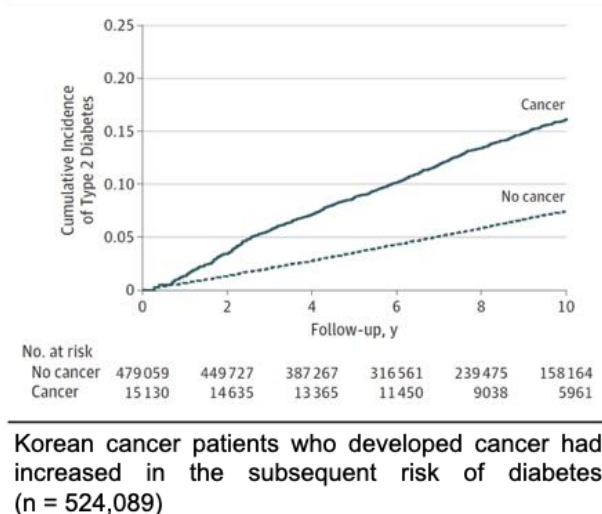
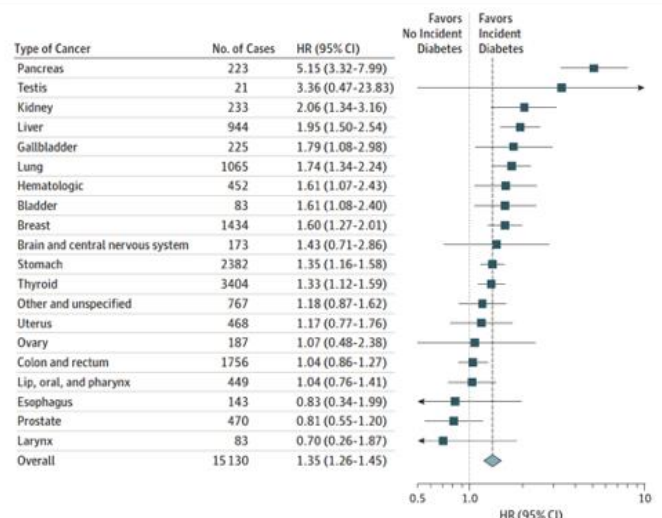


Figure 3. Hazard Ratios (HRs) for Incident Diabetes Associated With Cancer Development by Type of Cancer



Hwangbo et al., JAMA Oncology 2018

PART II

- Part I: Cancer Big Database
 - History of cancer genome study
 - Cbioportal, Xena browser and GDC data portal
 - Study example of cancer big database
- **Part II: Cancer Genome Analysis**
 - Mutational Signature Analysis: Mutalisk and FIREVAT
 - Data driven subtyping of cancer
- Part III: Artificial Intelligence Study using Cancer Big Data

ICGC Mutational Signatures Working Group

With which methods?



- For SBS, DBS and indel signatures, should we:
 - stick to the original ICGC PCAWG pipelines **SigProfiler** (Sanger) and **SigAnalyzer** (Broad)
 - test other methods (>25 **published tools** including EMu, SomaticSignatures, Palimpsest, **Mutalisk**, BayesNMF...). For review: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221235>
 - Choose one approach or combine several?
- For SV signatures:
 - not sure the PCAWG method can be easily reproduced (**ClusterSV** available to generate footprints but extracting signatures from that is not obvious)
 - alternatives include **Palimpsest** or **Signal**
- For signatures of repair deficiencies:
 - Mismatch repair deficiency -> **MSIsensor**, **MSIseq**, **MOSAIC**...
 - Homologous recombination defects -> **Signature 3**, **HRDetect**, **SigMA**, combined **HRD scores**...

Participants: Alvin Ng (**Sanger**), Dongwan Hong, Paz Polak (**Broad;NY**), Eric Letouzé (Paris genome center)



a web-based somatic MUTation AnaLyls toolKit for genomic, transcriptional and epigenomic signatures

Nucleic Acids Research (2018)

COSMIC
Catalogue Of Somatic Mutations In Cancer

Projects ▾ Data ▾ Tools ▾ News ▾ Help ▾ About ▾ Genome Version ▾ Search COSMIC...

Mutational Signatures (v3.1 - June 2020)

Introduction

Somatic mutations are present in all cells of the human body and occur throughout life. They are the consequence of multiple mutational processes, including the intrinsic slight infidelity of the DNA replication machinery, exogenous or endogenous mutagen exposures, enzymatic modification of DNA and defective DNA repair. Different mutational processes generate unique combinations of mutation types, termed "Mutational Signatures".

In the past few years, large-scale analyses have revealed many mutational signatures across the spectrum of human cancer types, including the latest effort by the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Network (Alexandrov, L.B. et al., 2020) using data from more than 23,000 cancer patients.



<https://cancer.sanger.ac.uk/cosmic/signatures>

Signature-based websites

As the number of mutational signatures and variant classes considered has increased, the need for a curated census of signatures has become apparent. Here, we deliver such a resource by providing a comprehensive overview of the key information known, suspected or widely discussed in the scientific literature for each of the identified mutational signatures on a dedicated website.

This summary includes the mutational profile, proposed aetiology and tissue distribution of each signature, as well as potential associations with other mutational signatures and how the signature has changed during iterations of analysis.

Currently, three different variant classes are considered, resulting in the following sets of mutational signatures.

Single Base Substitution (SBS)
Signatures

Doublet Base Substitution (DBS)
Signatures

Small Insertion and Deletion (ID)
Signatures

Versions

Mutational signatures version 3 was released as part of COSMIC release v89 (May 2019) and updated to version 3.1 in COSMIC release v91 (June 2020). The version 3.1 update expands and improves upon the version 2 signatures (March 2015) that were part of earlier COSMIC releases and can still be consulted.

Mutational Signatures
Version 2

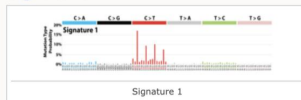
Bioinformatic tools

The current set of mutational signatures has been extracted using SigProfiler, a compilation of publicly available bioinformatic tools addressing all the steps needed for signature identification. SigProfiler functionalities include mutation matrix generation from raw data and signature extraction, among others.

SigProfiler
Bioinformatic Tools

Mutational Signature v2

Signature 1



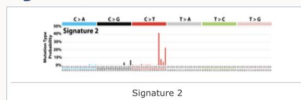
Cancer types: Signature 1 has been found in all cancer types and in most cancer samples.

Proposed aetiology: Signature 1 is the result of an endogenous mutational process initiated by spontaneous deamination of 5-methylcytosine.

Additional mutational features: Signature 1 is associated with small numbers of small insertions and deletions in most tissue types.

Comments: The number of Signature 1 mutations correlates with age of cancer diagnosis.

Signature 2



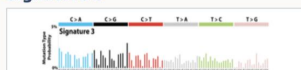
Cancer types: Signature 2 has been found in 22 cancer types, but most commonly in cervical and bladder cancers. In most of these 22 cancer types, Signature 2 is present in at least 10% of samples.

Proposed aetiology: Signature 2 has been attributed to activity of the AID/APOBEC family of cytidine deaminases. On the basis of similarities in the sequence context of cytosine mutations caused by APOBEC enzymes in experimental systems, a role for APOBEC1, APOBEC3A and/or APOBEC3B in human cancer appears more likely than for other members of the family.

Additional mutational features: Transcriptional strand bias of mutations has been observed in exons, but is not present or is weaker in introns.

Comments: Signature 2 is usually found in the same samples as Signature 13. It has been proposed that activation of AID/APOBEC cytidine deaminases is due to viral infection, retrotransposon jumping or to tissue inflammation. Currently, there is limited evidence to support these hypotheses. A germline deletion polymorphism involving the phenomenon of local hypermutation present in some cancers, known as kataegis, potentially implicating AID/APOBEC enzymes in this process as well.

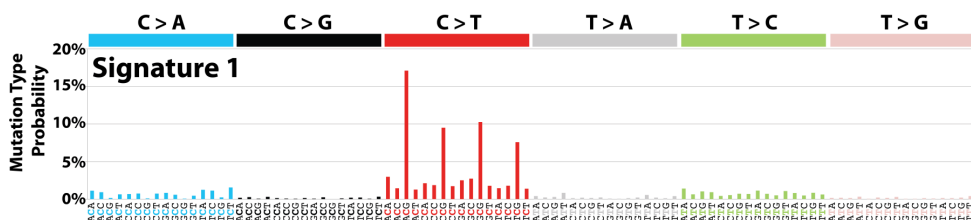
Signature 3



Cancer types: Signature 3 has been found in breast, ovarian, and pancreatic cancers.

Proposed aetiology: Signature 3 is associated with failure of DNA double-strand break-repair by homologous recombination.

Additional mutational features: Signature 3 associates strongly with elevated numbers of large (longer than 3bp) insertions and deletions with overlapping microhomology at breakpoint junctions.



Mutational Signature (v3 – May 2019)

- Single Base Substitution (SBS): 65 signatures (18 possible sequencing artifacts)
- Double Base Substitution (DBS): 11 signatures
- Small Insertion and Deletion Substitution (ID): 17 signatures



Mutational Signature v3.1




[Projects](#)
[Data](#)
[Tools](#)
[News](#)
[Help](#)
[About](#)
[Genome Version](#)

[SEARCH](#)
[Login](#)

Mutational Signatures (v3.1 - June 2020)

[Mutational Signatures Home](#)
[Single Base Substitution \(SBS\) Signatures](#)
[Doublet Base Substitution \(DBS\) Signatures](#)
[Small Insertion and Deletion \(ID\) Signatures](#)
[Mutational Signatures Version 2](#)
[SigProfiler Bioinformatic Tools](#)

Single Base Substitution (SBS) Signatures

Single base substitutions (SBS), also known as single nucleotide variants, are defined as a replacement of a certain nucleotide base. Considering the pyrimidines of the Watson-Crick base pairs, there are only six different possible substitutions: C>A, C>G, C>T, T>A, T>C, and T>G. These SBS classes can be further expanded considering the nucleotide context.

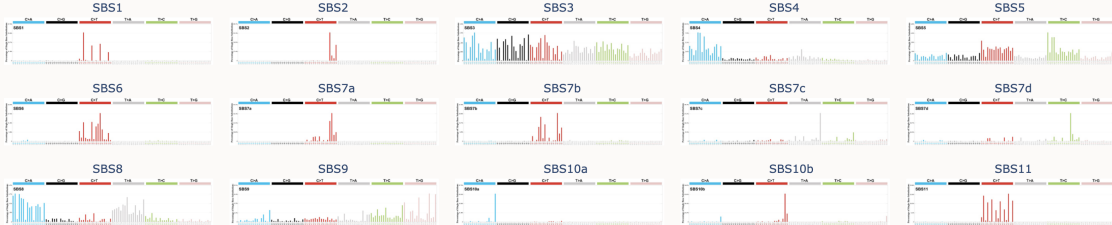
Current SBS signatures have been identified using 96 different contexts, considering not only the mutated base, but also the bases immediately 5' and 3'.

Click on any signature below to learn more about its details.

Signature extraction methods

With a few exceptions, the signatures were extracted using [SigProfiler](#) (as described in [Alexandrov, L.B. et al., 2020](#)) from the 2,780 whole-genome variant calls produced by the ICGC/TCGA [Pan Cancer Analysis of Whole Genomes \(PCAWG\)](#) Network. The stability and reproducibility of the signatures were assessed on somatic mutations from an additional 1,865 whole genomes and 19,184 exomes. All input data and references for original sources are available from synapse.org ID [syn11801889](#).

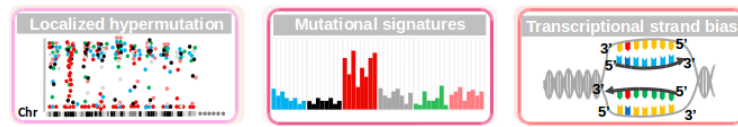
The COSMIC v3 signatures are available in numerical form in [syn12009743](#), and attributions of the signatures to mutations in tumors are available in [syn11804040](#) and [syn11804058](#). The COSMIC v3.1 signatures can be downloaded [here](#).



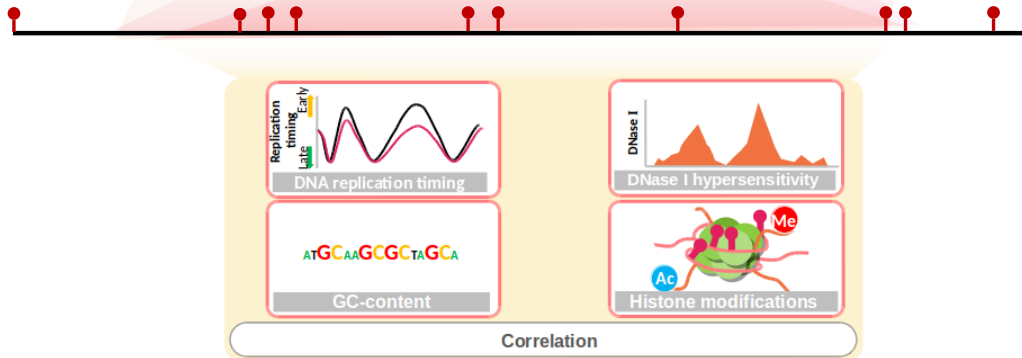
If there are 3,000 mutations
we'd expect a random dispersion of mutations (approx. 1 mutation/Mb)



Human genome $\approx$ 3,000-megabase (Mb)

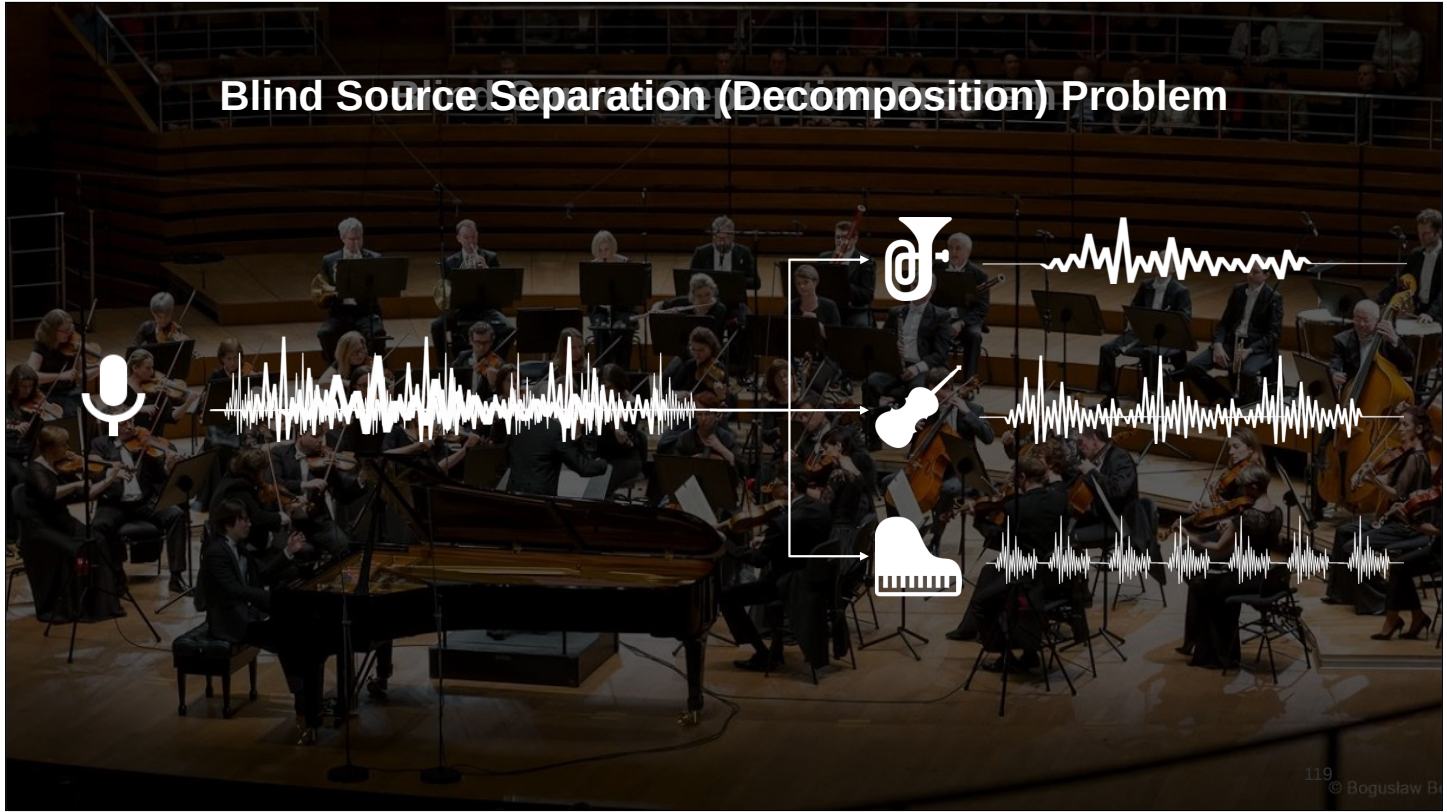


However, in reality, we observe uneven distribution of mutations in genome due to favored contexts



117







WIKIPEDIA
The Free Encyclopedia

Article [Talk](#)

Non-negative matrix factorization

From Wikipedia, the free encyclopedia

$$\begin{bmatrix} & & & & \\ & & & & \\ & & & & \\ & & & & \end{bmatrix}^{\mathbf{W}} \times \begin{bmatrix} & & & & & & \\ & & & & & & \\ & & & & & & \end{bmatrix}^{\mathbf{H}} \approx \begin{bmatrix} & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \end{bmatrix}^{\mathbf{V}}$$

History [\[edit \]](#)

In [chemometrics](#) non-negative matrix factorization has a long history under the name "self modeling curve resolution".^[10] In this framework the vectors in the right matrix are continuous curves rather than discrete vectors. Also early work on non-negative matrix factorizations was performed by a Finnish group of researchers in the 1990s under the name *positive matrix factorization*.^{[11][12][13]} It became more widely known as *non-negative matrix factorization* after Lee and [Seung](#) investigated the properties of the algorithm and published some simple and useful algorithms for two types of factorizations.^{[14][15]}

Background [\[edit \]](#)

Let matrix $\mathbf{V}$ be the product of the matrices $\mathbf{W}$ and $\mathbf{H}$,

$$\mathbf{V} = \mathbf{WH}.$$

Matrix multiplication can be implemented as computing the column vectors of $\mathbf{V}$ as linear combinations of the column vectors in $\mathbf{W}$ using coefficients supplied by columns of $\mathbf{H}$. That is, each column of $\mathbf{V}$ can be computed as follows:

$$\mathbf{v}_i = \mathbf{W}\mathbf{h}_i,$$

where $\mathbf{v}_i$ is the i -th column vector of the product matrix $\mathbf{V}$ and $\mathbf{h}_i$ is the i -th column vector of the matrix $\mathbf{H}$.

When multiplying matrices, the dimensions of the factor matrices may be significantly lower than those of the product matrix and it is this property that forms the basis of NMF. NMF generates factors with significantly reduced dimensions compared to the original matrix. For example, if $\mathbf{V}$ is an $m \times n$ matrix, $\mathbf{W}$ is an $m \times p$ matrix, and $\mathbf{H}$ is a $p \times n$ matrix then p can be significantly less than both m and n .

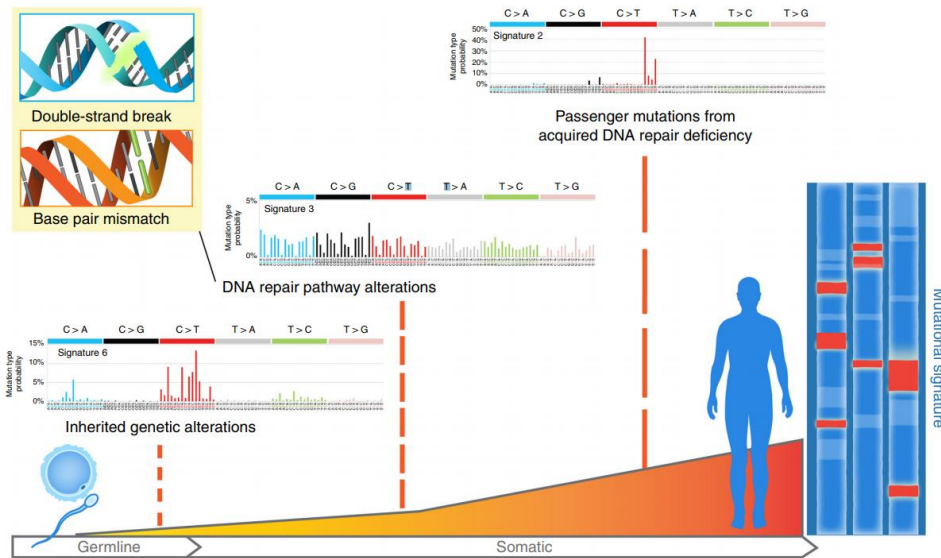
Here is an example based on a text-mining application:

- Let the input matrix (the matrix to be factored) be $\mathbf{V}$ with 10000 rows and 500 columns where words are in rows and documents are in columns. That is, we have 500 documents indexed by 10000 words. It follows that a column vector $\mathbf{v}$ in $\mathbf{V}$ represents a document.
- Assume we ask the algorithm to find 10 features in order to generate a *features matrix* $\mathbf{W}$ with 10000 rows and 10 columns and a *coefficients matrix* $\mathbf{H}$ with 10 rows and 500 columns.
- The product of $\mathbf{W}$ and $\mathbf{H}$ is a matrix with 10000 rows and 500 columns, the same shape as the input matrix $\mathbf{V}$ and, if the factorization worked, it is a reasonable approximation to the input matrix $\mathbf{V}$.
- From the treatment of matrix multiplication above it follows that each column in the product matrix $\mathbf{WH}$ is a linear combination of the 10 column vectors in the features matrix $\mathbf{W}$ with coefficients supplied by the coefficients matrix $\mathbf{H}$.

This last point is the basis of NMF because we can consider each original document in our example as being built from a small set of hidden features. NMF generates these features.

It is useful to think of each feature (column vector) in the features matrix $\mathbf{W}$ as a document archetype comprising a set of words where each word's cell value defines the word's rank in the feature: The higher a word's cell value the higher the word's rank in the feature. A column in the coefficients matrix $\mathbf{H}$ represents an original document with a cell value defining the document's rank for a feature. We can now reconstruct a document (column vector) from our input matrix by a linear combination of our features (column vectors in $\mathbf{W}$) where each feature is weighted by the feature's cell value from the document's column in $\mathbf{H}$.

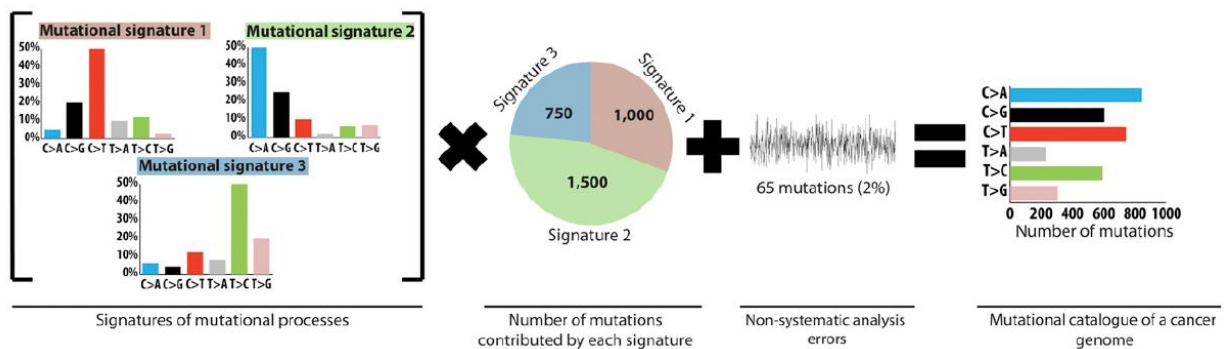
Mutations accumulated over lifetime can be decomposed



Jennifer Ma et al., *Nat Comm* 2017

121

Mutational signature decomposition



Alexandrov et al., *Cell* 2013

122

Mutational signatures

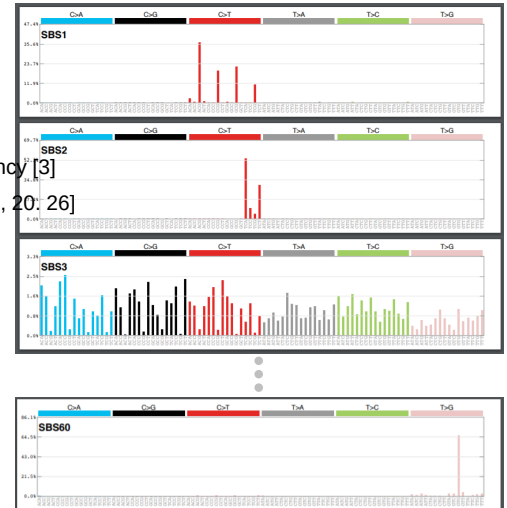
30 Catalogue of Somatic Mutations in Cancer (COSMIC) signatures



Alexandrov et al., *Nature* 2013

- Age [1, 5]
- APOBEC activity [2, 13]
- Homologous recombination deficiency [3]
- Defective DNA mismatch repair [15, 20, 26]
- Tobacco smoking/chewing [4, 29]
- Ultraviolet light exposure [7]
- POLE* mutations [10]
- Exposure to aristolochic acid [22]
- Alkylating agent temozolomide [11]
- Aflatoxin [24]

65 PanCancer Analysis of Whole Genomes (PCAWG) signatures

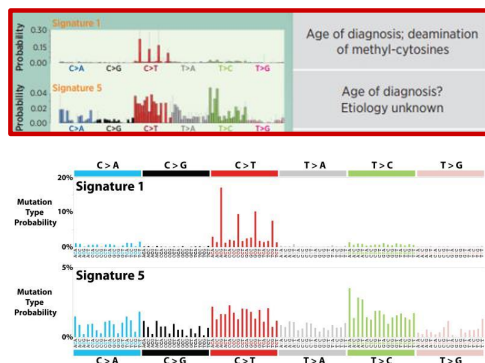


Campbell et al., *bioRxiv* 2017

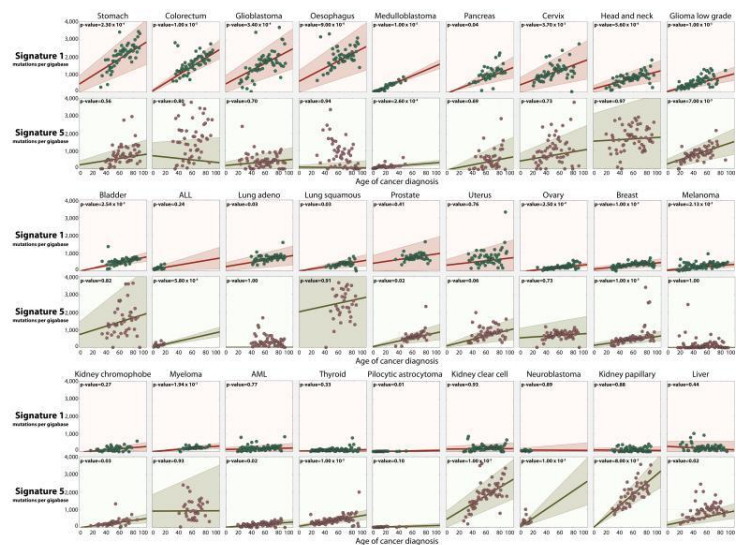
123

Clock-like mutational signatures

Nik-Zainal et al., *Clinical Cancer Research* 2017



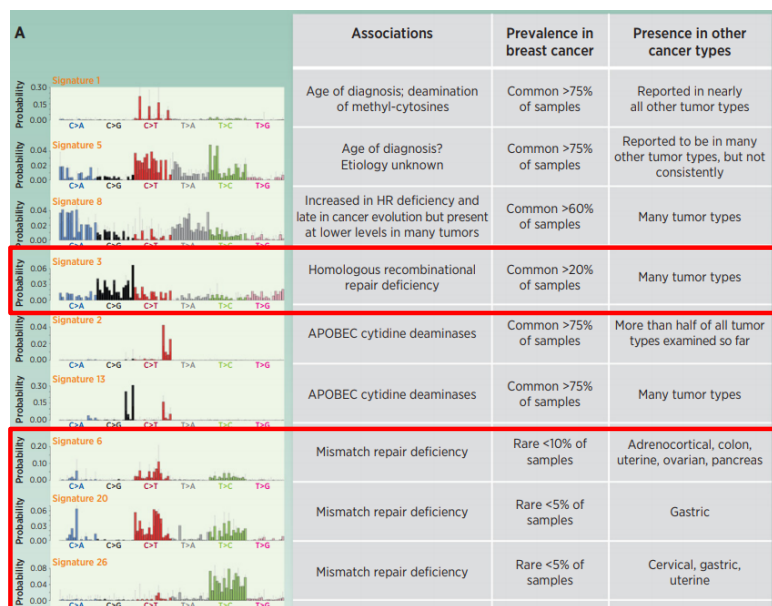
Signature 1 and 5 is positively correlated to the age of patients



Alexandrov, Ludmil B., et al. *Nat Gen* 2015

124

Mutational signatures are clinically relevant



Nik-Zainal et al., *Clinical Cancer Research* 2017

“... **mutational signatures** [3, 6, 20, 26] are a direct pathophysiologic readout of the abrogation of a DNA repair gene/pathway and could be used as a **biomarker** to report **DNA repair/deficiency** in a tumor.”

125

nature
medicine

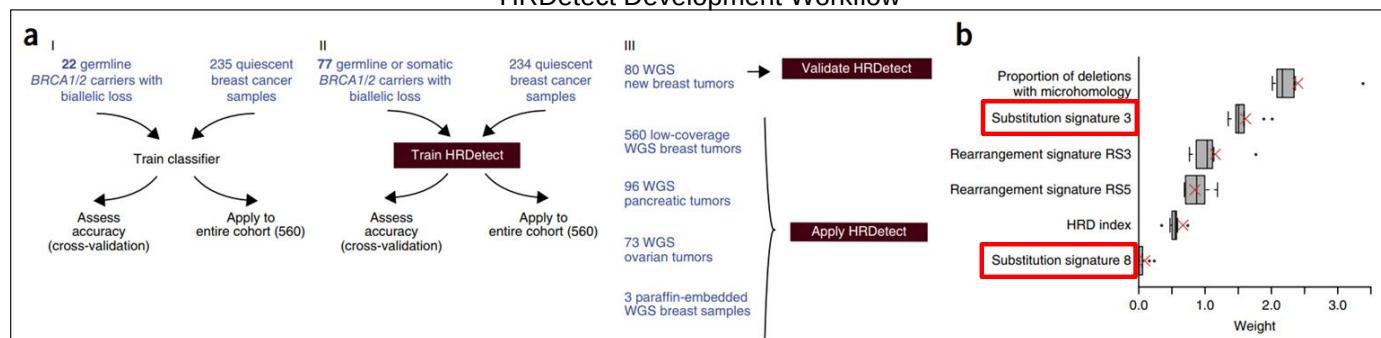
ANALYSIS

Mutational signature #3 (homologous recombination defect)

HRDetect is a predictor of *BRCA1* and *BRCA2* deficiency based on mutational signatures

Helen Davies^{1,2,22}, Dominik Glodzik^{1,32}, Sandro Morganello¹, Lucy R Yates^{1,2}, Johan Staaf³, Xueqing Zou¹, Manasa Ramakrishna^{1,4}, Sancha Martin¹, Sandrine Boyault⁵, Anieta M Sieuwerts⁶, Peter T Simpson⁷, Tari A King⁸, Keiran Raine¹, Jorunn E Eyfjord⁹, Gu Kong¹⁰, Åke Borg³, Ewan Birney¹¹, Hendrik G Stunnenberg¹², Marc J van de Vijver¹³, Anne-Lise Borresen-Dale^{14,15}, John W M Martens⁶, Paul N Span^{16,17}, Sunil R Lakhani^{7,18}, Anne Vincent-Salomon^{19,20}, Christos Sotiropoulos²¹, Andrew Tutt^{22,23}, Alastair M Thompson²⁴, Steven Van Laere^{25,26}, Andrea L Richardson^{27,28}, Alain Viari^{29,30}, Peter J Campbell¹, Michael R Stratton¹ & Serena Nik-Zainal^{1,31}

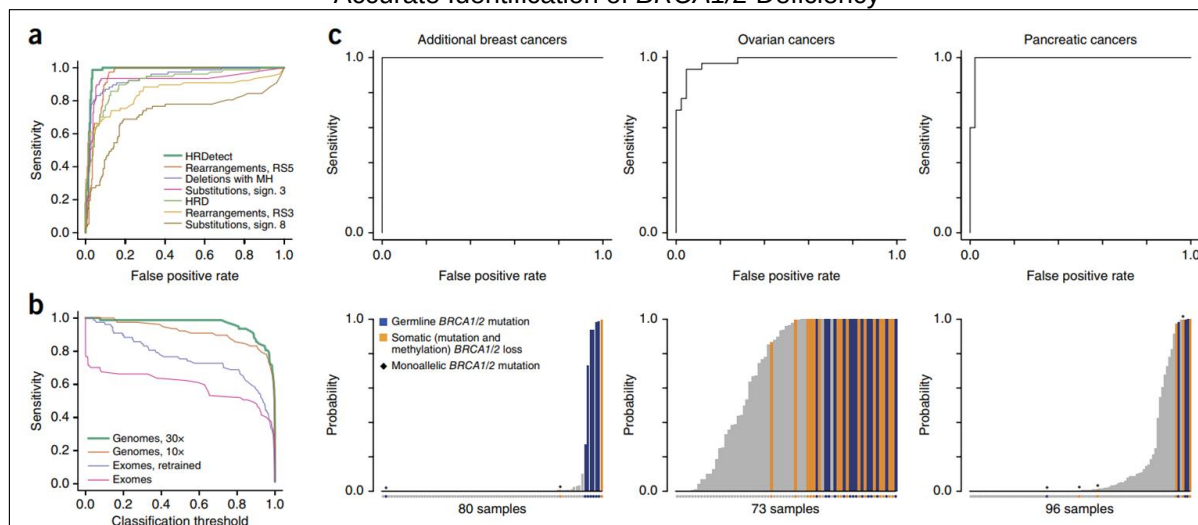
HRDetect Development Workflow



Helen Davies et al., *Nature Medicine* 2017

Mutational signature #3 (homologous recombination defect)

Accurate Identification of *BRCA1/2* Deficiency

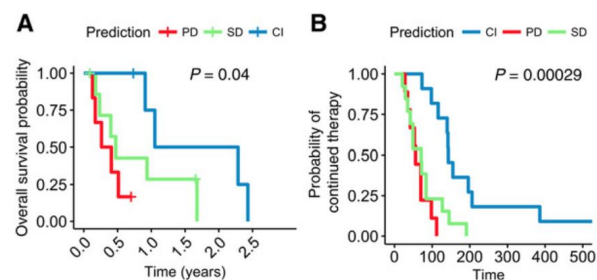
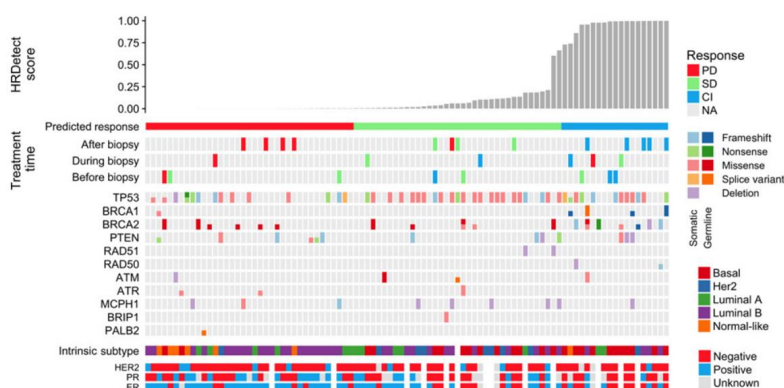


Helen Davies et al., *Nature Medicine* 2017

127

Mutational signature #3 (homologous recombination defect)

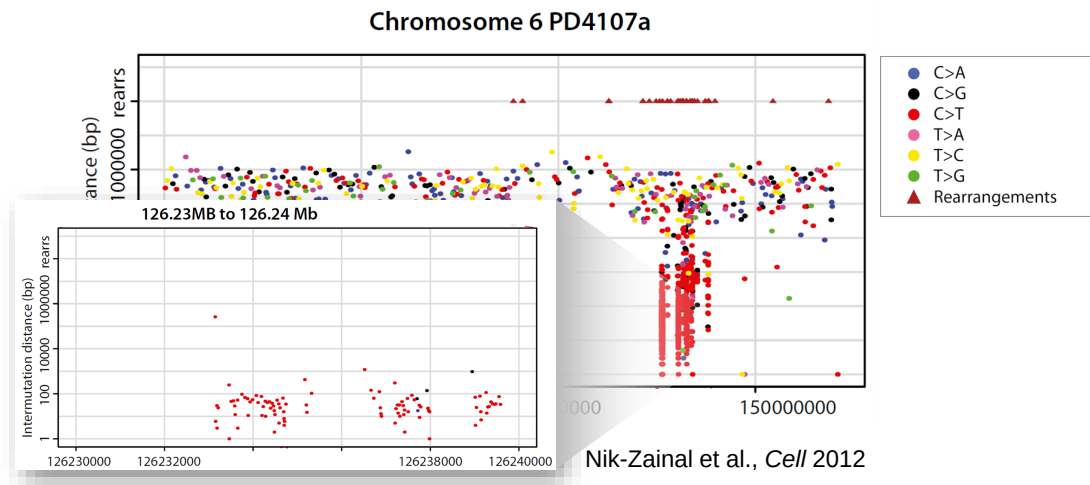
Homologous Recombination Deficiency and Platinum-Based Therapy Outcomes in Advanced Breast Cancer



Eric Y. Zhao et al., *Clinical Cancer Research* 2017

128

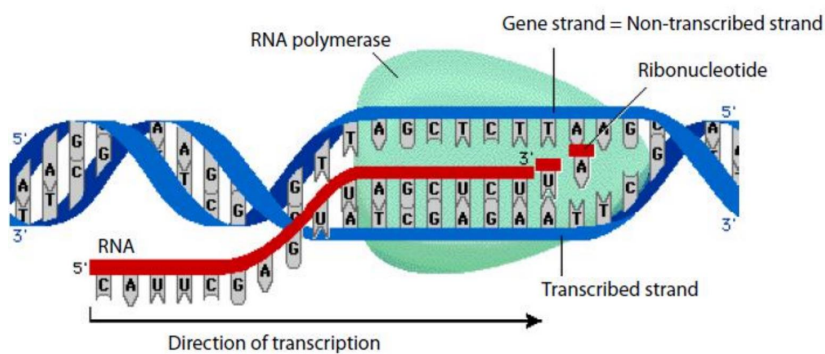
Localized hypermutations (kataegis)



- There are clusters of mutations with short distances between mutations (kataegis)
- Kataegis is co-localized with rearrangements that have features of chromothripsis

129

Transcriptional strand bias

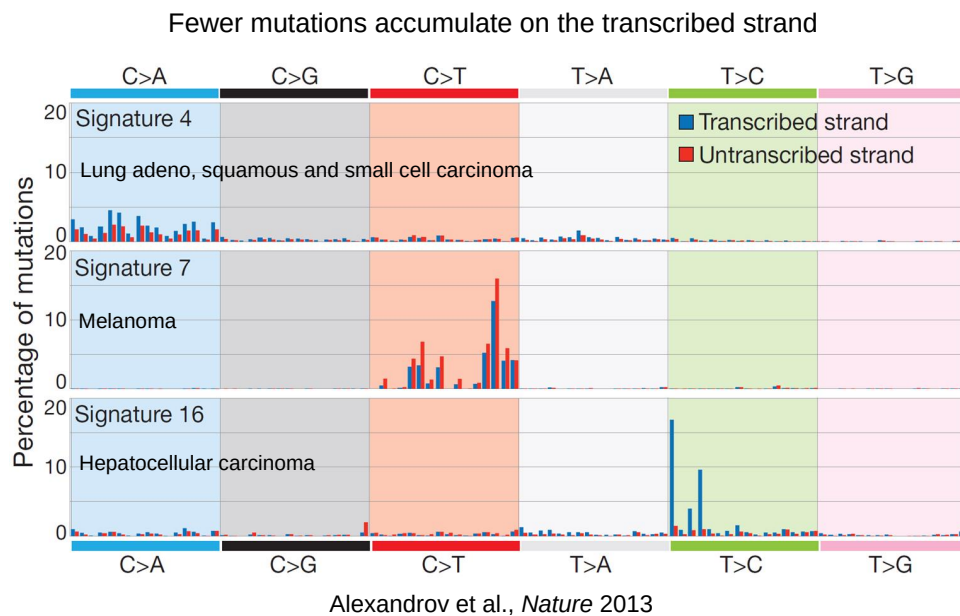


<ftp://ftp.sanger.ac.uk/pub/resources/theses/snz/chapter6.pdf>

Cause of transcriptional strand bias is transcription-coupled repair (TCR) of nucleotide excision repair (NER); DNA damage is repaired more efficiently on the transcribed strand

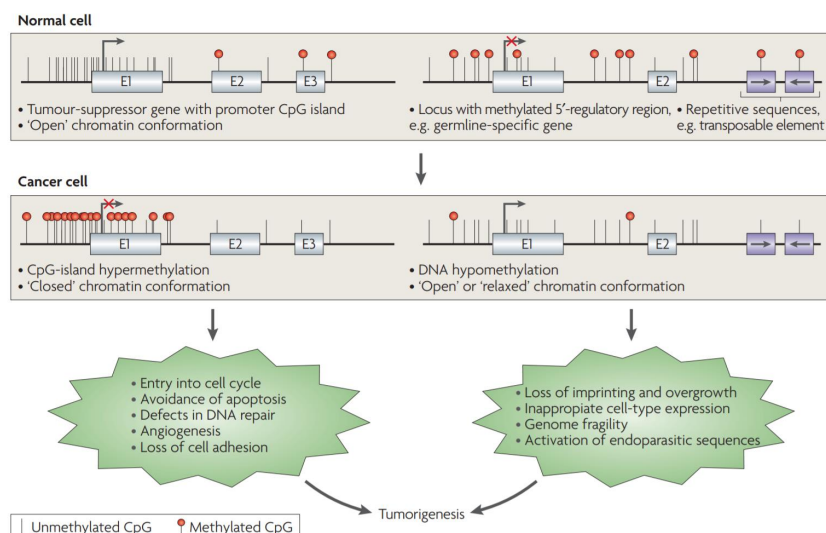
130

Somatic mutation rates are biased by transcriptional strand



131

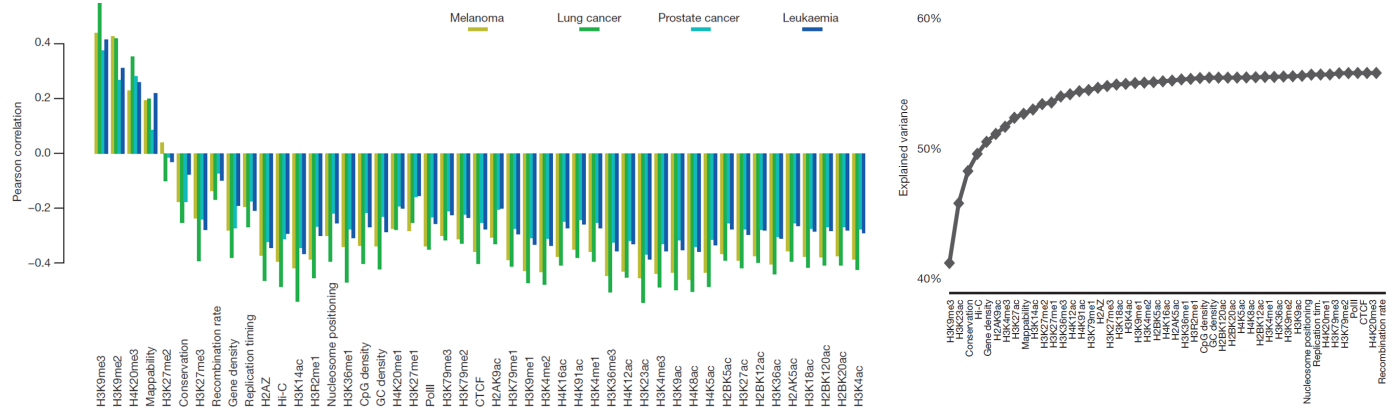
Cancer and epigenomics



Esteller, *Nature Reviews Genetics* 2007

132

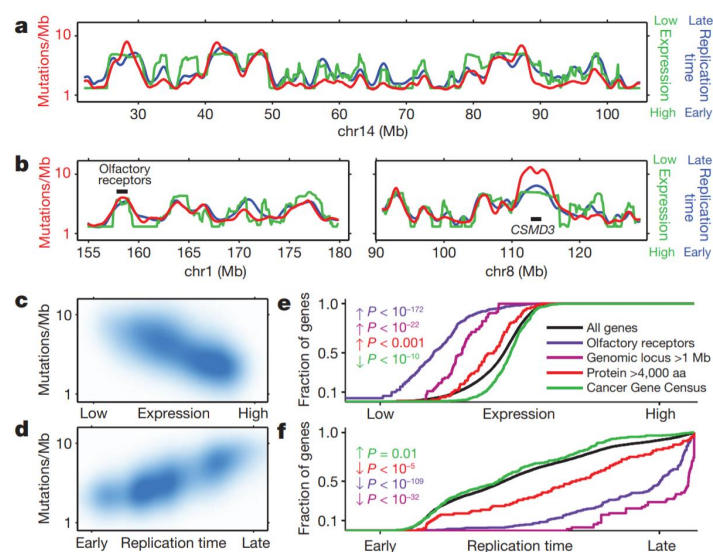
Mutation rates are correlated with genomic/epigenomic modifications



Schuster-Bockler and Lehner, *Nature* 2012

133

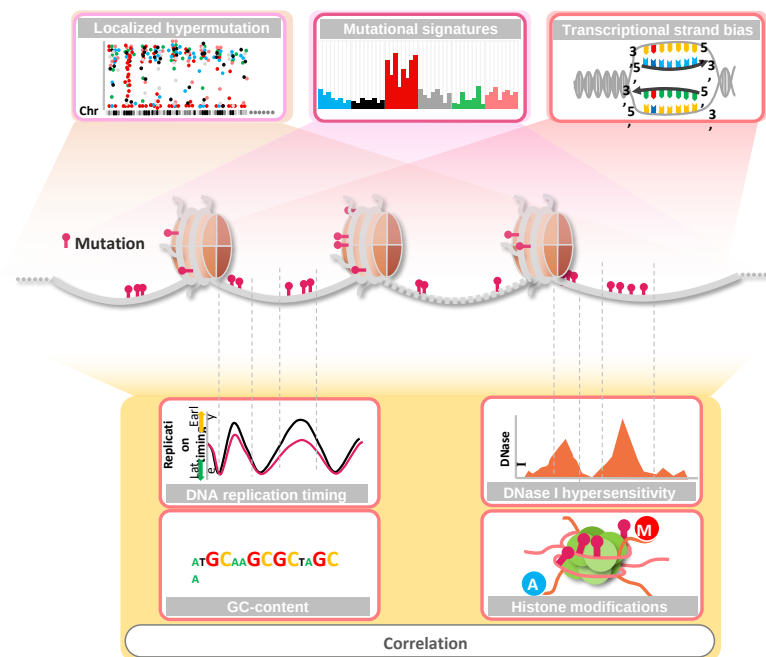
Mutation rates also vary with DNA replication timing



Lawrence et al., *Nature* 2013

134

Putting them altogether to better understand mutations operative in cancer



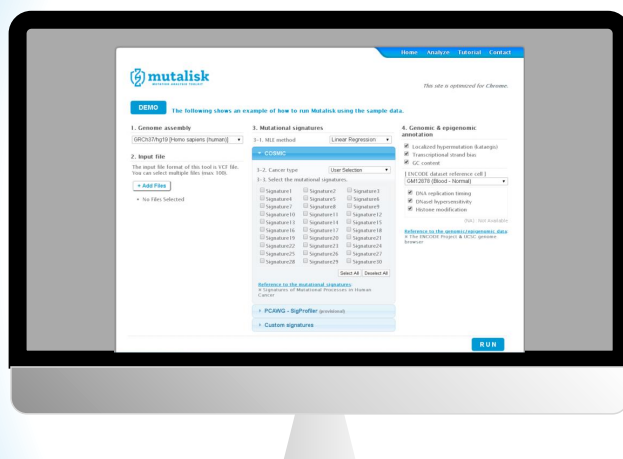
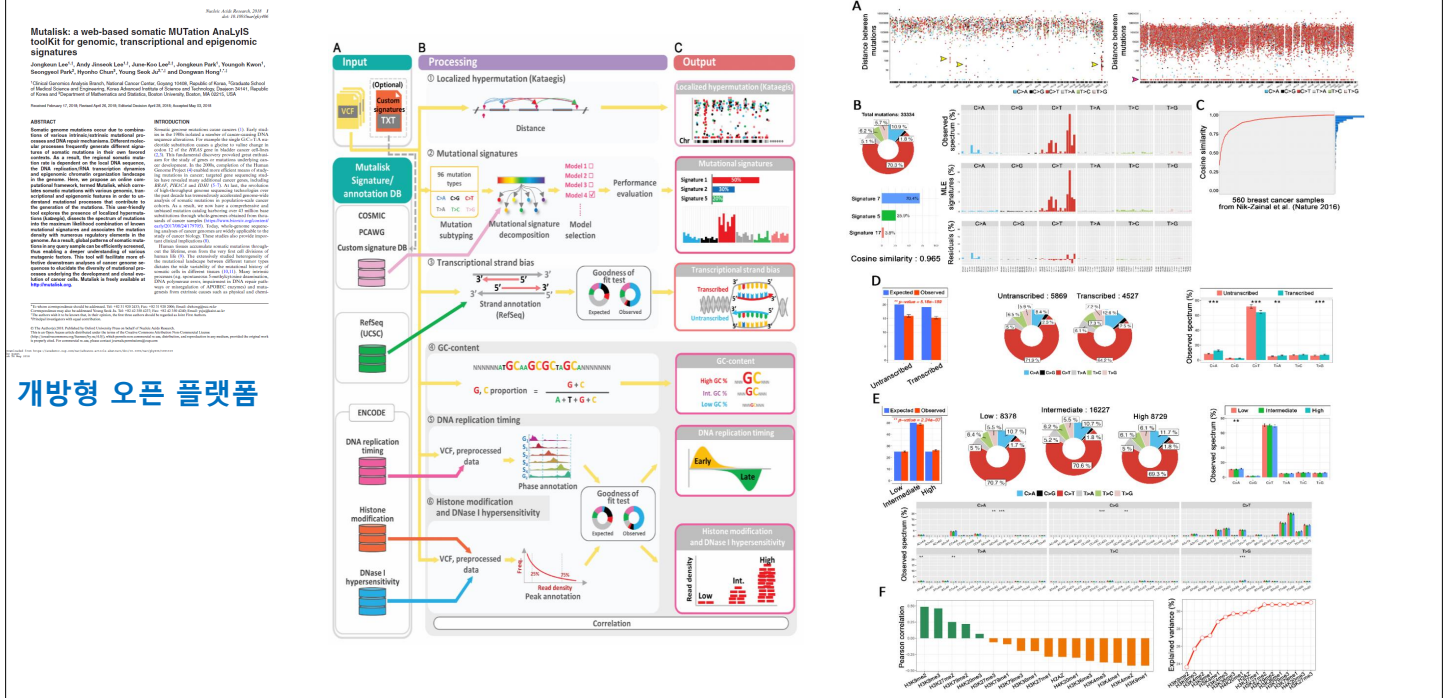
135

Existing mutational signature decomposition tools

Tool	Graphical user interface available?	Batch analysis possible?	Results downloadable?
WTSI (Alexandrov et al., Cell 2013)	No	Yes	Yes
deconstructSigs (Rosenthal et al., Genome Biology 2016)	No	Yes	Yes
MutaGene (Goncarenco et al., Nucleic Acids Research 2017)	Yes	No	No

136

Mutalisk: a web-based somatic MUTation AnaLysis toolKit for genomic, transcriptional and epigenomic signatures (Nucleic Acids Research, 2018)



Available at <http://mutalisk.org/>

mutalisk
 MUTATION ANALYSIS TOOLKIT

1. Identification of mutational signatures
1. Analysis of associations between various genome/epigenome regulatory elements and somatic mutation rates

Mutalisk (1) – analysis options

The screenshot shows the Mutalisk web application interface with the following components and annotations:

- Header:** Navigation links (Home, Analyze, Tutorial, Contact) and a note "This site is optimized for Chrome."
- DEMO Section:** A heading "The following shows an example of how to run Mutalisk using the sample data."
- Step 1: Genome assembly**
 - Dropdown menu: GRCh37/hg19 [Homo sapiens (human)]
 - Annotation: "GRCh37/hg19, GRCh38/hg38, GRCh38/mm10, WBcel236/ce11, WS220/ce10" (points to the dropdown)
- Step 2: Input file**
 - Text: "The input file format of this tool is VCF file. You can select multiple files (max 100)." (points to the step heading)
 - Button: "+ Add Files"
 - Text: "• No Files Selected"
 - Annotation: "Multiple vcf files (max 100)" (points to the "+ Add Files" button)
- Step 3: Mutational signatures**
 - Section: "3-1. MLE method" with a dropdown set to "Linear Regression"
 - Section: "3-2. Cancer type" with a dropdown set to "User Selection"
 - Section: "3-3. Select the mutational signatures."
 - Grid of checkboxes for 30 COSMIC signatures (Signature 1 to Signature 30)
 - Annotation: "COSMIC, PCAWG (SigProfiler) signatures" (points to the grid)
 - Annotation: "User/custom signatures" (points to the "PCAWG - SigProfiler (provisional)" and "Custom signatures" buttons)
- Step 4: Genomic & epigenomic annotation**
 - Checkboxes: "Localized hypermutation (kataegis)", "Transcriptional strand bias", "GC content" (all checked)
 - Text: "[ENCODE dataset reference cell]"
 - Dropdown menu: "GM12878 (Blood - Normal)"
 - Annotation: "31 cell lines" (points to the dropdown)
 - Checkboxes: "DNA replication timing", "DNaseI hypersensitivity", "Histone modification" (all checked)
 - Text: "(NA) : Not Available"
 - Text: "Reference to the genomic/epigenomic data: * The ENCODE Project & UCSC genome browser"
 - Annotation: "42 cancer types" (points to the "DNA replication timing" checkbox)
- Footer:** "National Cancer Center. 323 Ilhan-ro, Ilsandong-gu, Goyang-si Gyeonggi-do, 10408, Republic of Korea"
- Buttons:** "RUN" (bottom right)

Mutalisk (1) – analysis options

Custom signatures

3-2. Select the mutational signatures.

Please upload your own signature file for the decomposition of mutational signatures. A tab-delimited sample signature file is available below:

[[Sample signature file](#)]

For additional information on the formatting of the signature file, please refer to the [Tutorials](#) page.

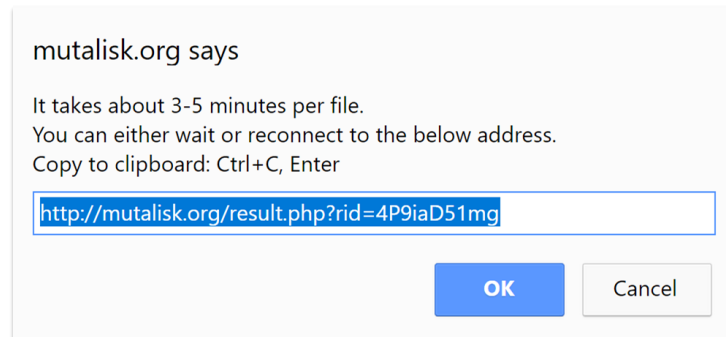
+ Add file

• No File Selected

Provide a tab-delimited .txt signature file

Substitution Type	Trinucleotide	Somatic Mutation Type	Signature 1	...	Signature 7
C>A	ACA	A[C>A]A	0.011098326		0.0004
C>A	ACC	A[C>A]C	0.009149341		0.0005
C>A	ACG	A[C>A]G	0.00149007	...	0
C>A	ACT	A[C>A]T	0.006233885		0.0004
C>G	ACA	A[C>G]A	0.001801068		0
		⋮			⋮
T>G	TTT	T[T>G]T	0.004030128	...	0.0014

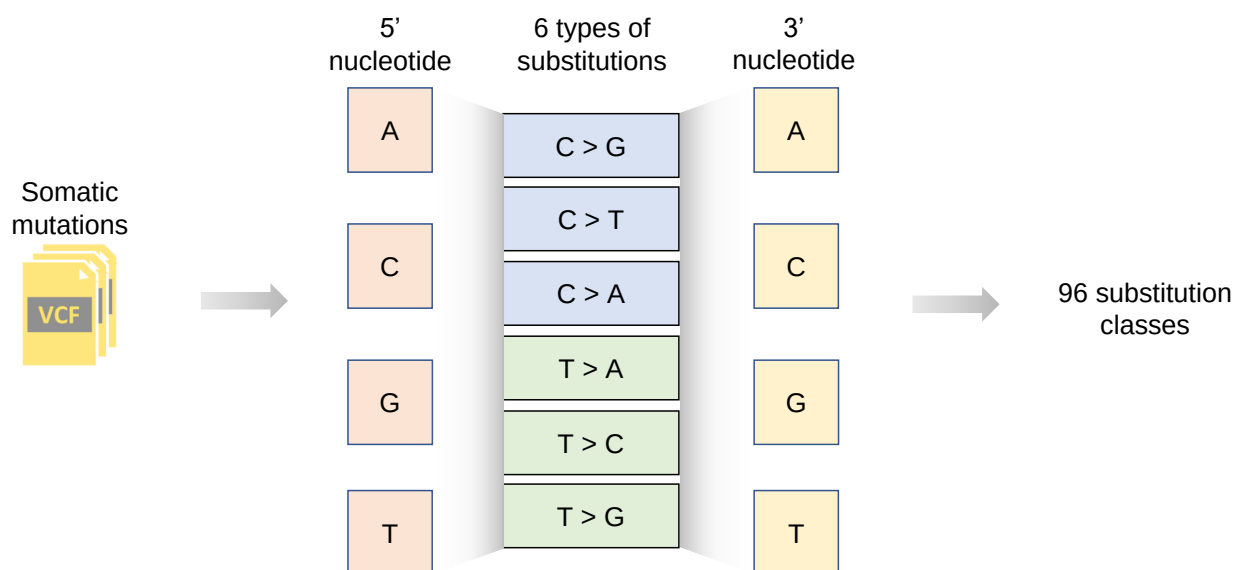
Mutalisk (1) – analysis options



User-uploaded data are permanently deleted after 48 hours;
you may access the analysis results for 48 hours

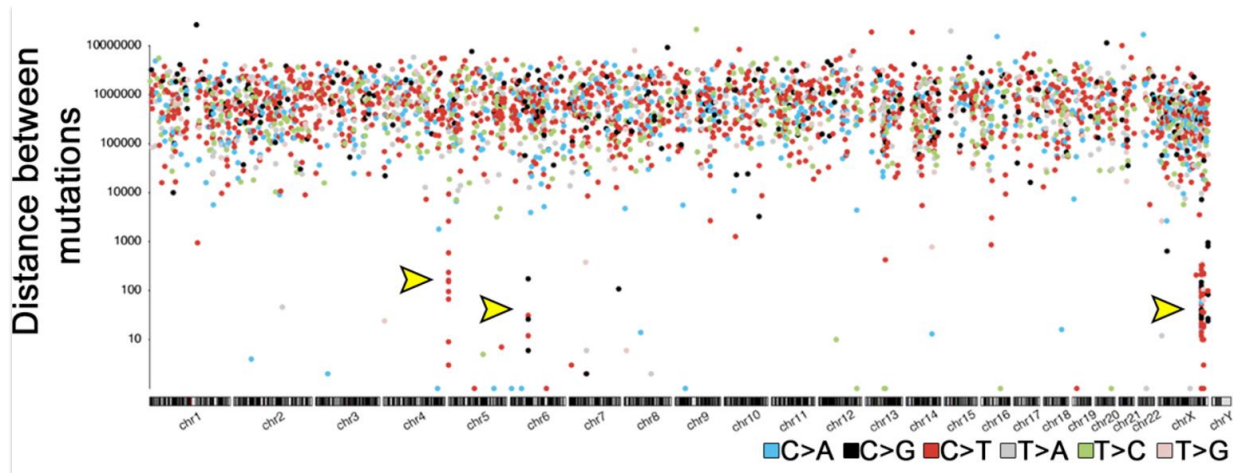
141

Mutalisk (2) – preprocessing



142

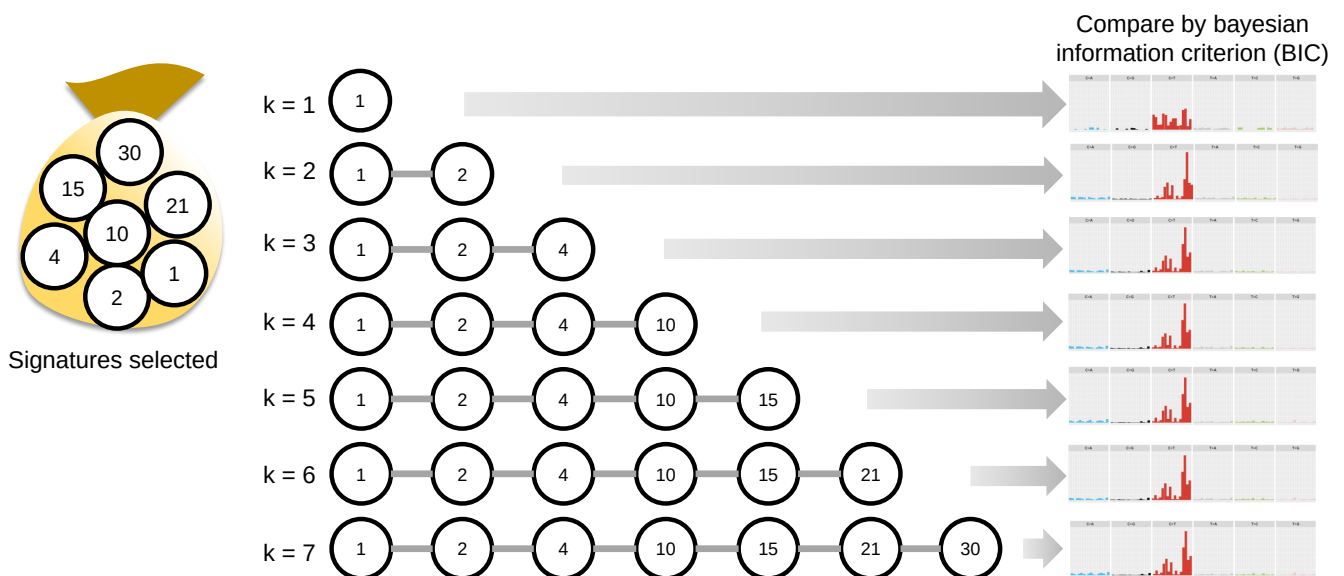
Mutalisk (3) – rainfall plot (*kataegis*)



Lung cancer sample (Lee et al., *J Clin Oncol* 2017)

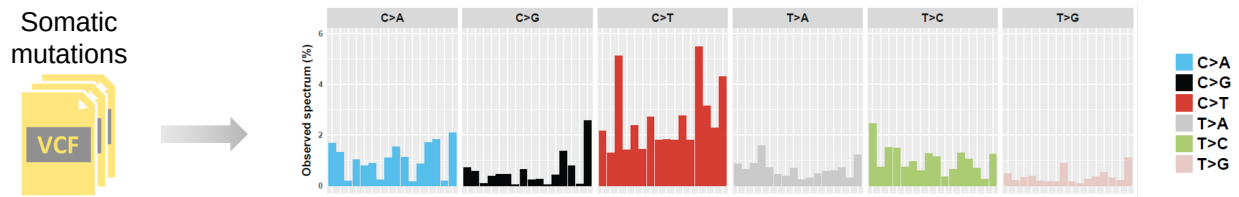
143

Mutalisk (4) – mutational signature identification



144

Mutalisk (4) – mutational signature identification



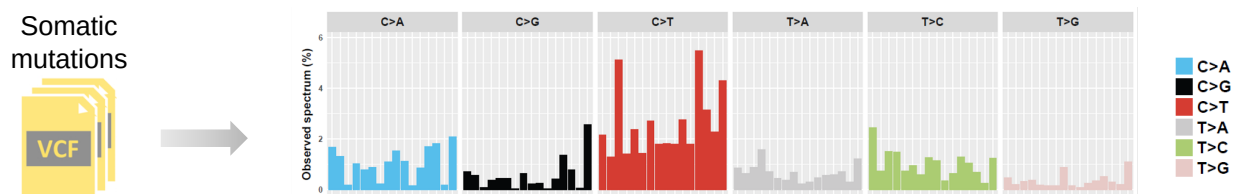
If we conjecture that there are 7 COSMIC signatures underlying the mutational profile, then

$$C(30,7) = \frac{30!}{(7! (30 - 7)!)}$$

2,035,800 different combinations of signatures

145

Mutalisk (4) – mutational signature identification



If we conjecture that there are 7 PCAWG signatures underlying the mutational profile, then

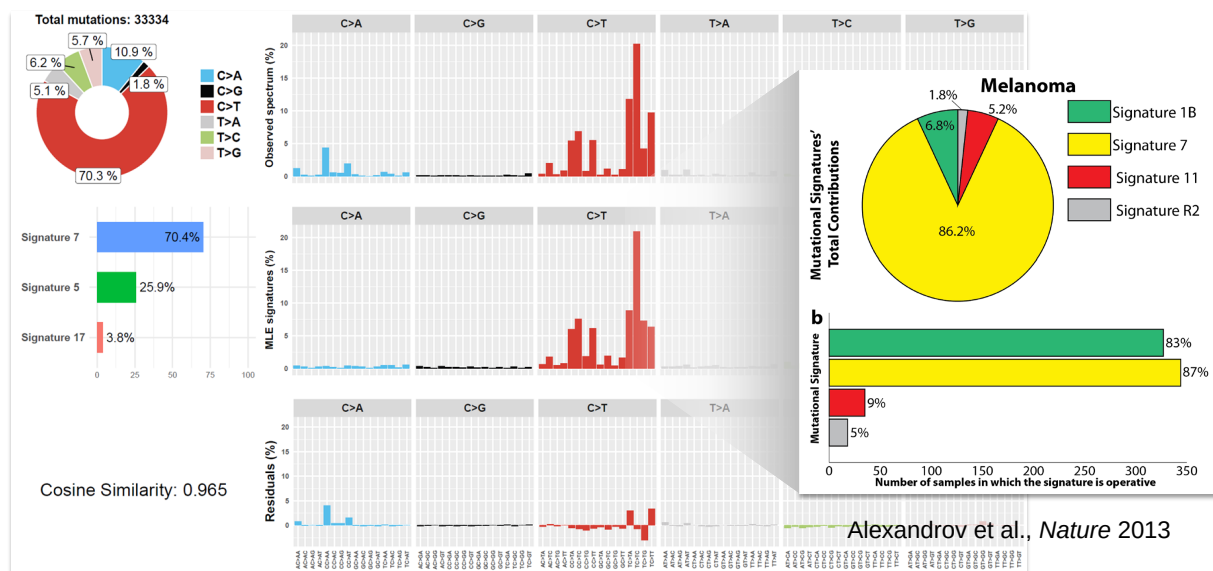
$$C(65,7) = \frac{65!}{(7! (65 - 7)!)}$$

696,190,560 different combinations of signatures

This is a computationally expensive task

146

Mutalisk (4) – mutational signature identification



Melanoma sample (Plesance et al., Nature 2010)

147

Mutalisk (4) – mutational signature identification

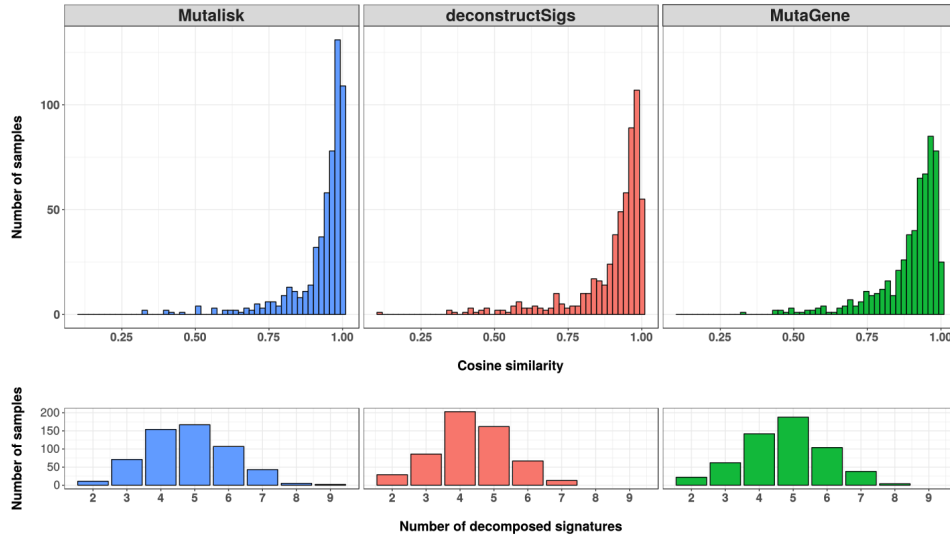
Mutalisk identifies up to 7 decomposition models

Mutational signatures						
No.	Signatures	Probabilities	Cosine similarity	BIC	Confidence Interval	
7	7 8 2 11 18 17 12 Our Best	0.59097 0.11749 0.10825 0.06858 0.05707 0.03374 0.02390	0.98200	222713.700		View
6	7 8 2 11 18 17	0.59163 0.14049 0.10844 0.06984 0.05328 0.03632	0.98200	222820.200		View
5	7 8 2 11 17	0.59492 0.19932 0.10851 0.06440 0.03285	0.98200	223568.700		View
3	7 8 2	0.64330 0.25263 0.10407	0.98000	224619.100		View
4	7 8 2 11	0.59826 0.22793 0.10916 0.06466	0.98100	224828.400		View
2	7 8	0.71863 0.28137	0.96800	225142.600		View
1	7	1.00000	0.78700	273566.100		View

148

Mutalisk (4) – mutational signature identification

Comparison of decomposition results using 560 breast cancer samples (Nik-Zainal et al., Nature 2016)



149

Mutalisk (5) – transcriptional strand bias analysis

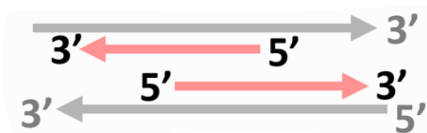
Somatic mutations



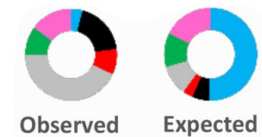
UCSC RefSeq



Strand annotation



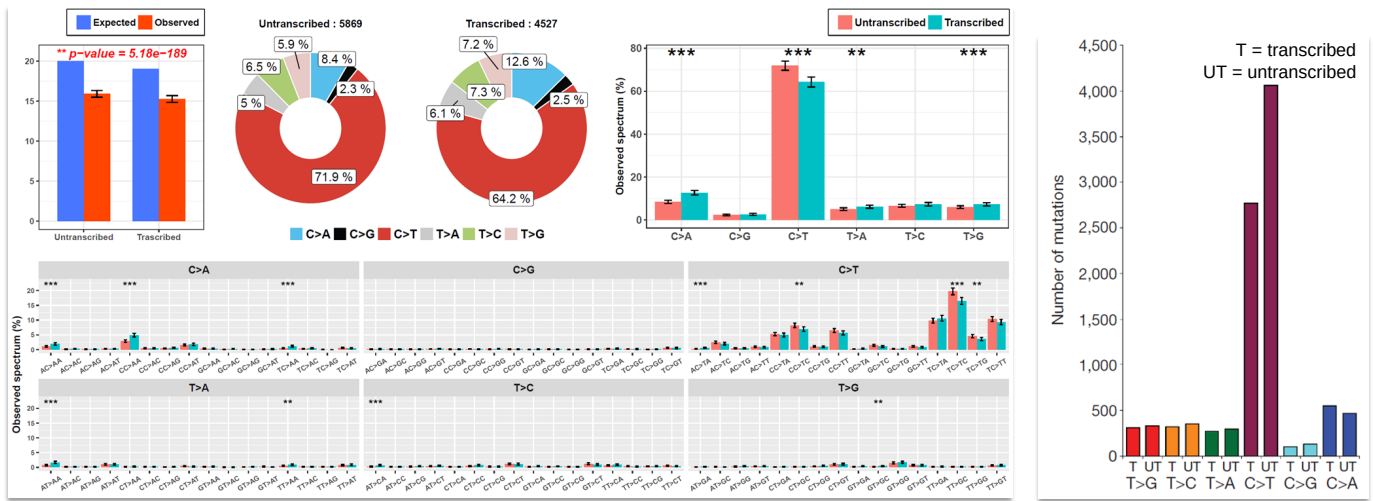
Goodness of fit test



Reference genome	Transcribed strand region proportion	Untranscribed strand region proportion
CRCh37/hg19	0.19044	0.20010
GRCh38/hg38	0.18885	0.28435

150

Mutalisk (5) – transcriptional strand bias analysis

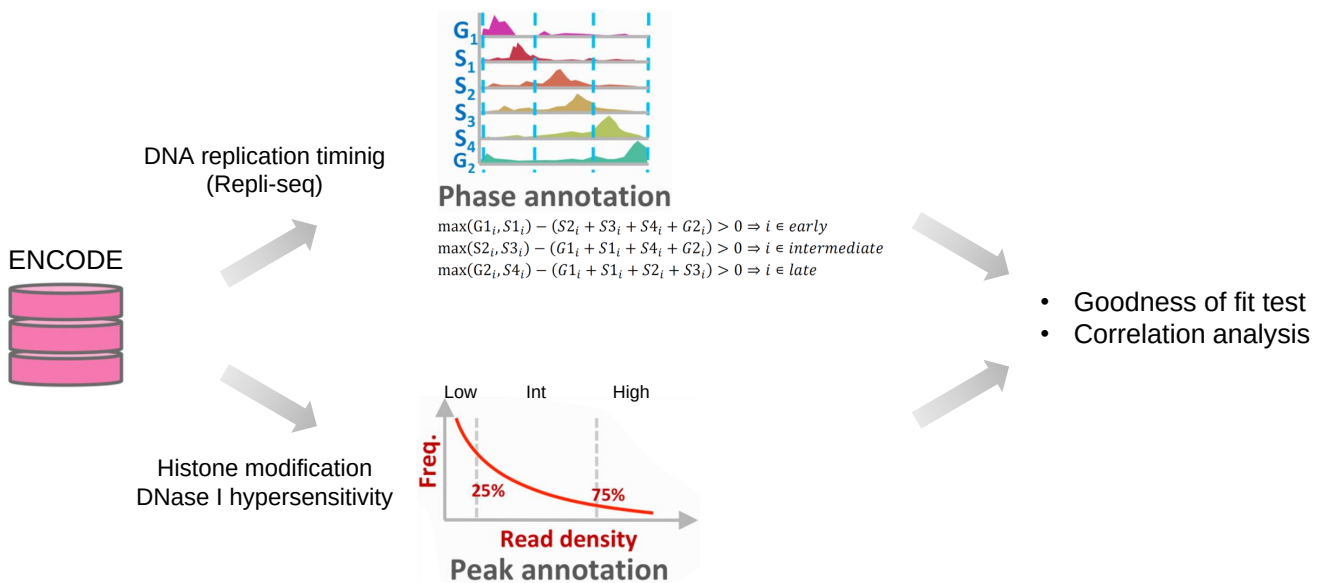


Melanoma sample (Pleasant et al., *Nature* 2010)

Pleasant et al., *Nature* 2010

151

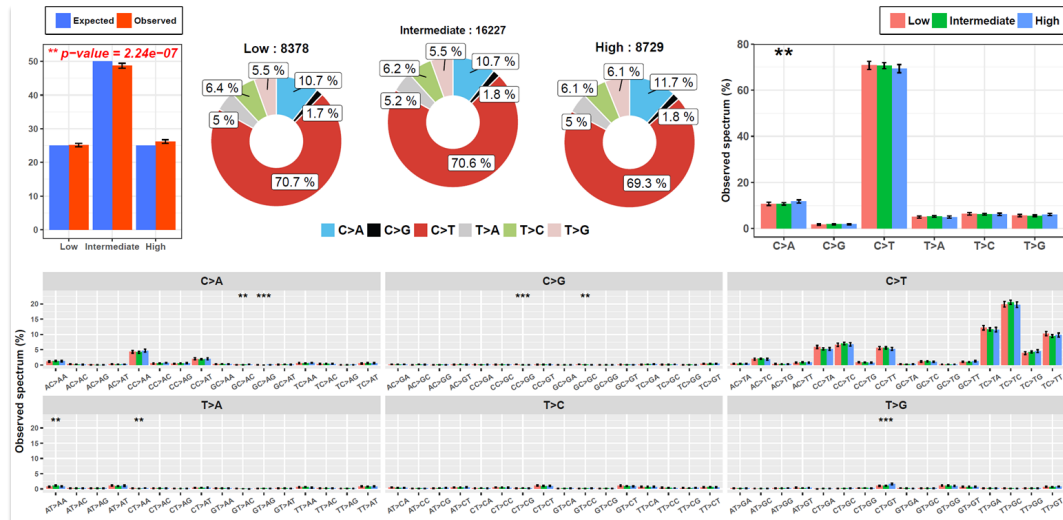
Mutalisk (6) – genomic/epigenomic modification analysis



152

Mutalisk (6) – genomic/epigenomic modification analysis

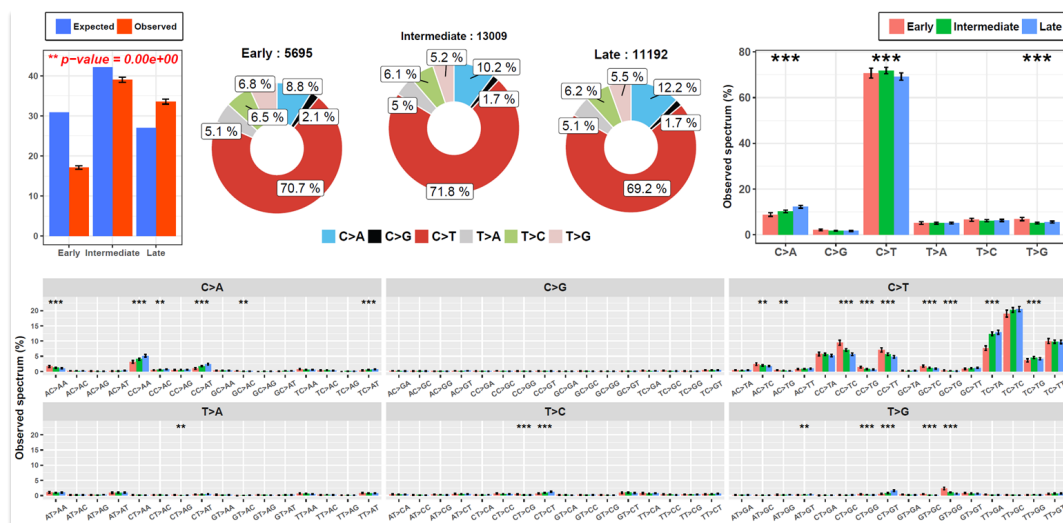
H3k9me3 , melanoma sample (Pleasant et al., Nature 2010)



153

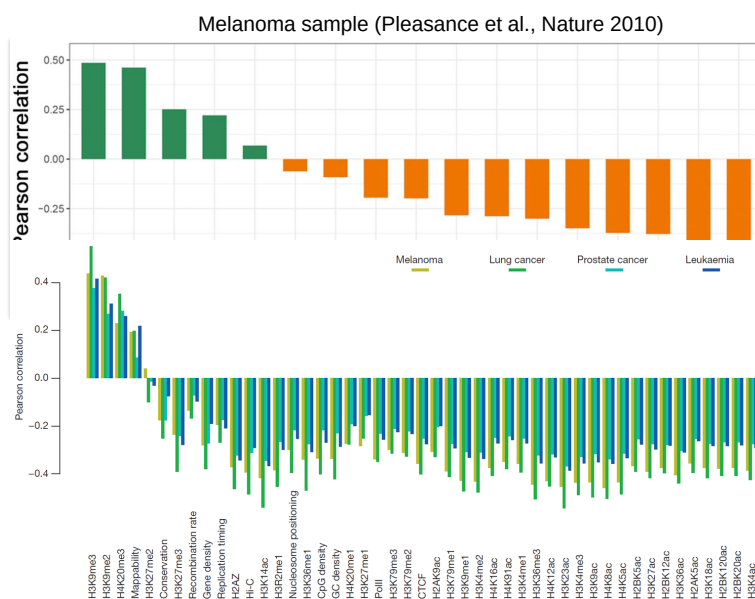
Mutalisk (6) – genomic/epigenomic modification analysis

DNA replication timing, melanoma sample (Pleasant et al., Nature 2010)



154

Mutalisk (6) – genomic/epigenomic modification analysis

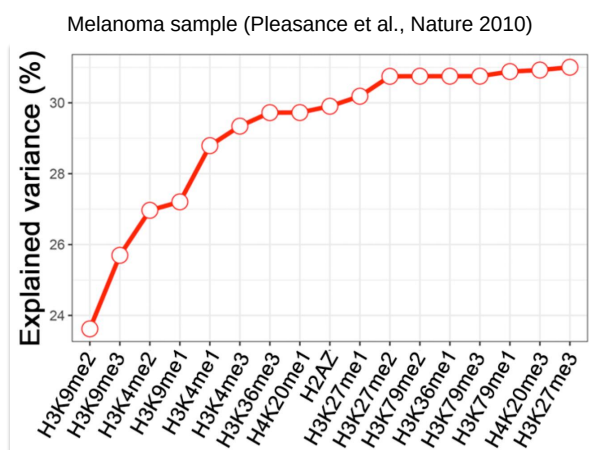


Mitalisk

Schuster-Bockler and
Lehner, *Nature* 2012

155

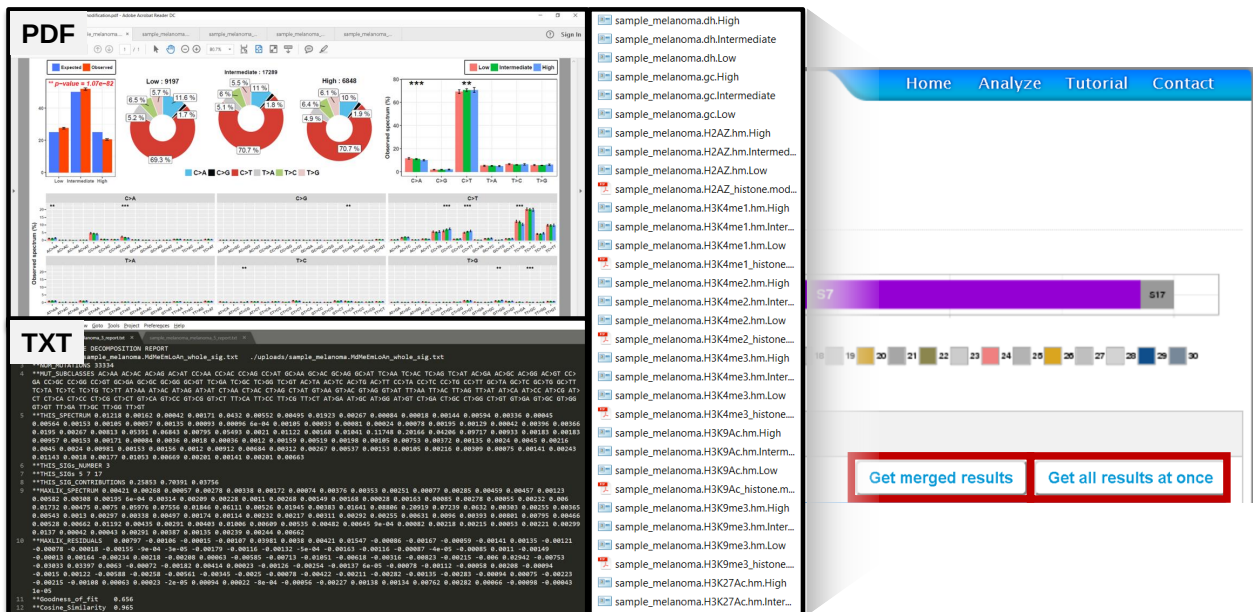
Mutalisk (6) – genomic/epigenomic modification analysis



- Forward feature selection (generalized least-squares model and Akaike information criterion)
- Explained variance computed by linear regression

156

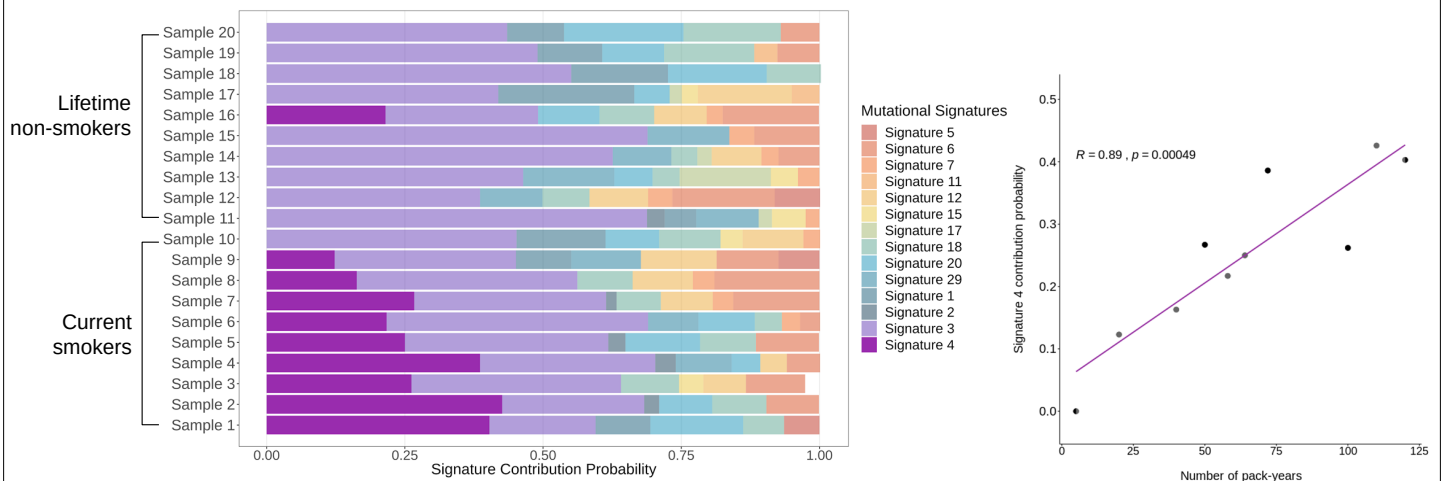
Mutalisk (7) – downloading results



157

Tobacco-associated mutational signature #4

TCGA Lung Adenocarcinoma Samples



158

Run Mutalisk

Home Analyze Tutorial Contact



This site is optimized for Chrome.

DEMO

The following shows an example of how to run Mutalisk using the sample data.

Melanoma data example

Lung data example

§ Download the sample data: [Melanoma / Lung]

[Source of the sample data](#) : [Melanoma cancer] *L. D. Pleasance et. al. Nature 2010.* & [Lung cancer] *June-Koo Lee et. al. JCO. 2017.*

sample_lung.vcf

Variant Call Format (VCF): sample_lung.vcf

```
1 ##fileformat=VCFv4.2
2 #CHROM POS ID REF ALT QUAL FILTER INFO
3 1 1910980 . A G . . .
4 1 5148788 . G C . . .
5 1 5867505 . C T . . .
6 1 7146801 . G A . . .
7 1 7665765 . C T . . .
8 1 7749485 . T G . . .
9 1 8710829 . G A . . .
10 1 9804956 . C T . . .
11 1 12381586 . G A . . .
12 1 17274833 . G T . . .
13 1 17365503 . T A . . .
14 1 18099894 . A T . . .
15 1 19665267 . G A . . .
16 1 20753216 . C A . . .
17 1 22905721 . T C . . .
18 1 28496930 . T C . . .
19 1 29464355 . A G . . .
20 1 30253004 . G T . . .
21 1 30346067 . G A . . .
22 1 30627515 . T C . . .
23 1 32251636 . T A . . .
24 1 34081408 . G A . . .
25 1 34435930 . A T . . .
26 1 35909337 . T C . . .
27 1 36062312 . T C . . .
28 1 36414737 . T A . . .
29 1 36729500 . A T . . .
30 1 36936453 . A T . . .
31 1 37261912 . G T . . .
32 1 37277934 . G A . . .
33 1 37708424 . C T . . .
34 1 39672350 . G C . . .
35 1 42038123 . C T . . .
36 1 45810549 . C T . . .
37 1 46950051 . C T . . .
38 1 46984444 . C T . . .
39 1 47898023 . G T . . .
40 1 47446596 . G T . . .
41 1 47619512 . C A . . .
42 1 48781219 . A G . . .
43 1 50032004 . G A . . .
44 1 50642579 . G T . . .
```

행 1, 열 1 — 3399 행

DEMO

The following shows an example of how to run Mutalisk using the sample data.

1. Genome assembly

GRCh37/hg19 [Homo sapiens (human)]

2. Input file

The input file format of this tool is VCF file. You can select multiple files (max 300). The total size of multiple files should be less than 1GB.

+ Add Files

• sample_lung.vcf

3. Mutational signatures

3-1. MLE method

Linear Regression

COSMIC

3-2. Cancer type

Lung Adeno

3-3. Select the mutational signatures

- | | | |
|---|---|--|
| ☒ Signature1 | ☒ Signature2 | ☐ Signature3 |
| ☒ Signature4 | ☒ Signature5 | ☒ Signature6 |
| ☐ Signature7 | ☐ Signature8 | ☐ Signature9 |
| ☐ Signature10 | ☐ Signature11 | ☐ Signature12 |
| ☒ Signature13 | ☐ Signature14 | ☐ Signature15 |
| ☐ Signature16 | ☒ Signature17 | ☐ Signature18 |
| ☐ Signature19 | ☐ Signature20 | ☐ Signature21 |
| ☐ Signature22 | ☐ Signature23 | ☐ Signature24 |
| ☐ Signature25 | ☐ Signature26 | ☐ Signature27 |
| ☐ Signature28 | ☐ Signature29 | ☐ Signature30 |

Select All Deselect All

[Reference to the mutational signatures:](#)

* Signatures of Mutational Processes in Human Cancer

▶ PCAWG - SigProfiler (provisional)

▶ Custom signatures

4. Genomic & epigenomic annotation

- ☒ Localized hypermutation (kataegis)
- ☒ Transcriptional strand bias
- ☒ GC content

[ENCODE dataset reference cell]

GM12878 (Blood - Normal)

- ☒ DNA replication timing
- ☒ DNaseI hypersensitivity
- ☒ Histone modification

(NA) : Not Available

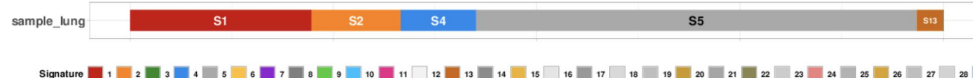
[Reference to the genomic/epigenomic data:](#)

* The ENCODE Project & UCSC genome browser

RUN

National Cancer Center, 323 Ilsan-ro, Ilsandong-gu, Goyang-si Gyeonggi-do, 10408, Republic of Korea

Summary of the best results from our tool



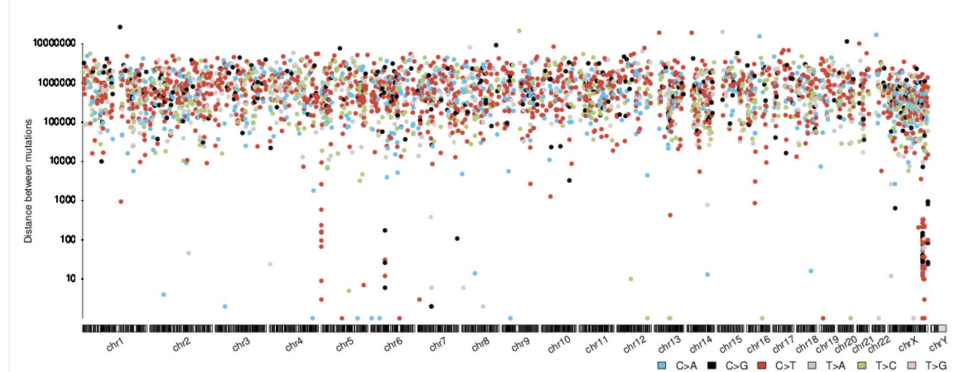
sample_lung

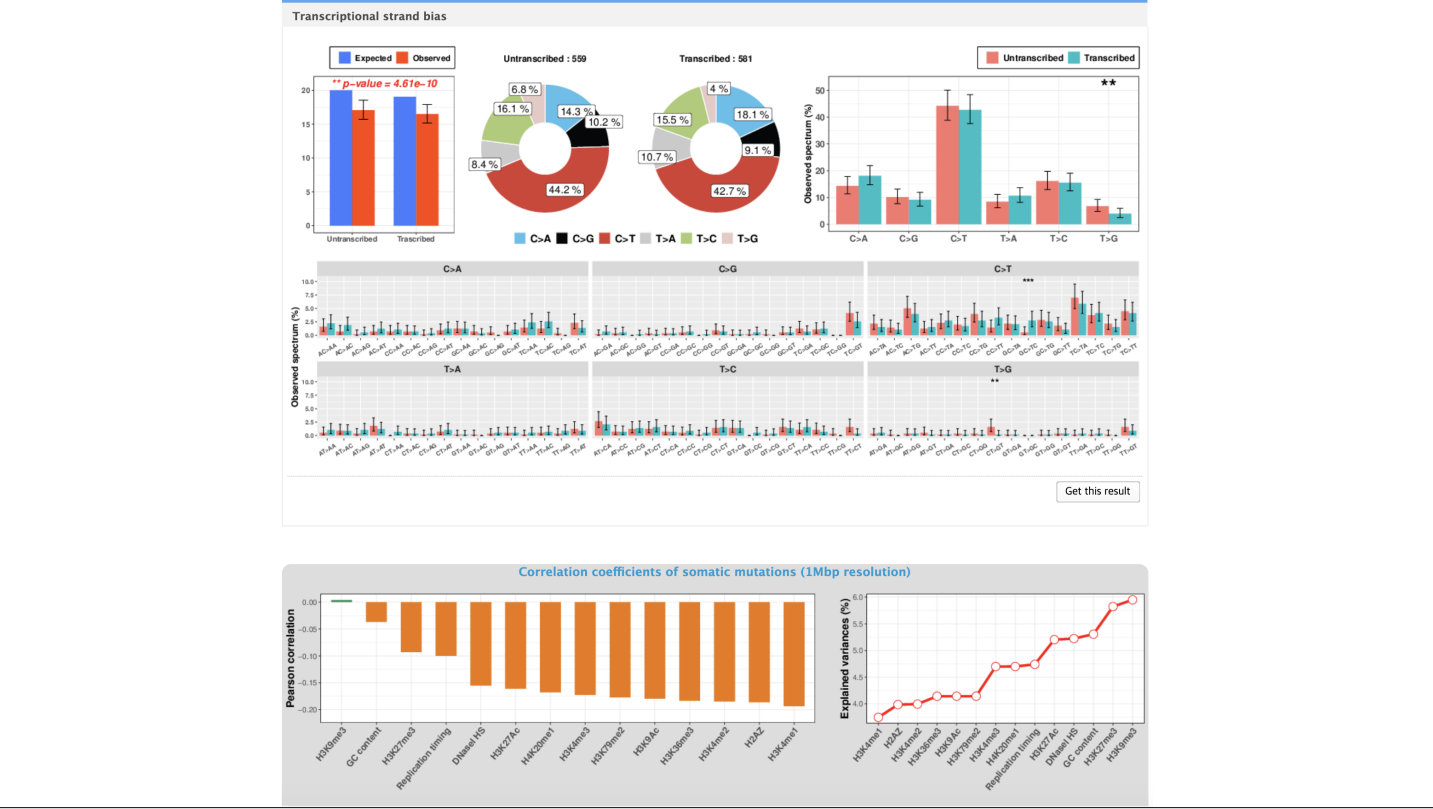
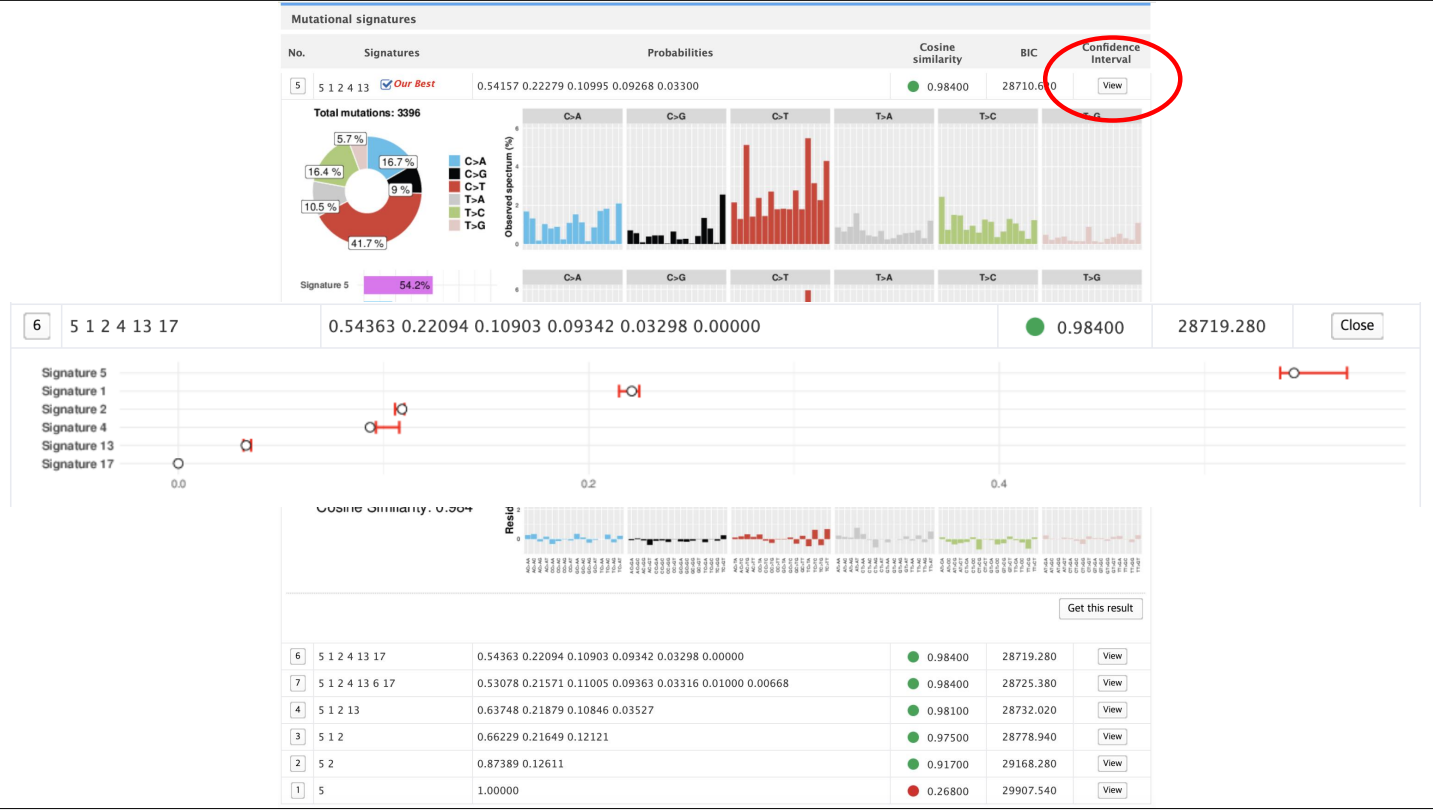
File : sample_lung

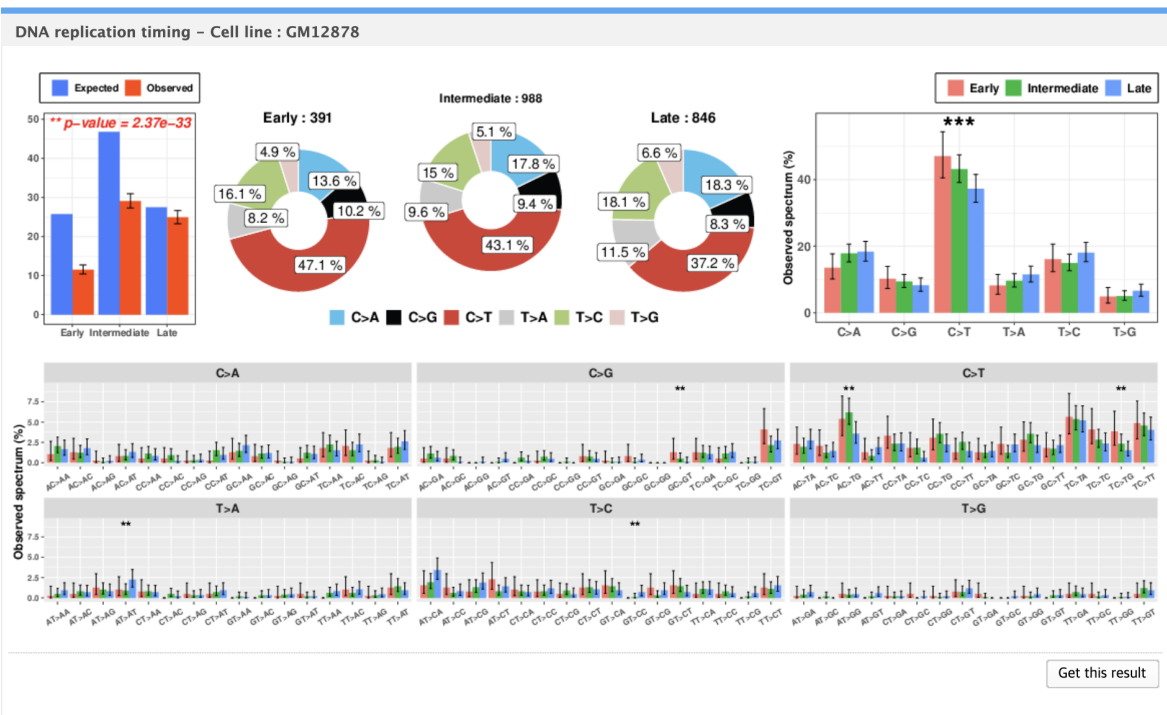
Get merged results

Get all results at once

Localized hypermutation (kataegis)







Home

Analyze

Tutorial

Contact



Run MutaLisk (v3)

This site is optimized for Chrome.

DEMO

The following shows an example of how to run MutaLisk using the sample data.

1. Genome assembly

GRCh37/hg19 [Homo sapiens (human)]

2. Input file

The input file format of this tool is VCF file. You can select multiple files (max 300). The total size of multiple files should be less than 1GB.

+ Add Files

sample_lung.vcf

3. Mutational signatures

3-1. MLE method

Linear Regression

COSMIC

PCAWG - SigProfiler (provisional)

3-2. Cancer type

User Selection

3-3. Select the mutational signatures.

Full screening

Random sampling

SBS1 SBS2 SBS3 SBS4

SBS5 SBS6 SBS7a SBS7b

SBS9 SBS12 SBS16 SBS19

SBS23 SBS27 SBS31 SBS35

SBS39 SBS43 SBS47 SBS51

SBS55 SBS56 SBS57 SBS58

SBS59 SBS60

Select All

Deselect All

Custom signatures

4. Genomic & epigenomic annotation

Localized hypermutation (kataegis)

Transcriptional strand bias

GC content

[ENCODE dataset reference cell]

GM12878 (Blood - Normal)

DNA replication timing

DNAseI hypersensitivity

Histone modification

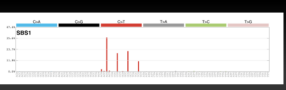
(NA) : Not Available

Reference to the genomic/epigenomic data:

The ENCODE Project & UCSC genome browser

RUN

SBS1



An endogenous mutational process initiated by spontaneous or enzymatic deamination of 5-methylcytosine to thymine which generates C:T mismatches in double stranded DNA. Failure to detect and remove these mismatches prior to DNA replication results in fixation of the T substitution for C.

Summary of the best results from our tool

sample_lung

S1

S2

S5

S8

S13

S40

S42

1 2 3 4 5 6 7a 7b 7c 7d 8 9 10a 10b 11 12 13 14 15 16 17a 17b 18 19 20 21 22

SBS 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50

52 53 54 55 56 57 58 59 60

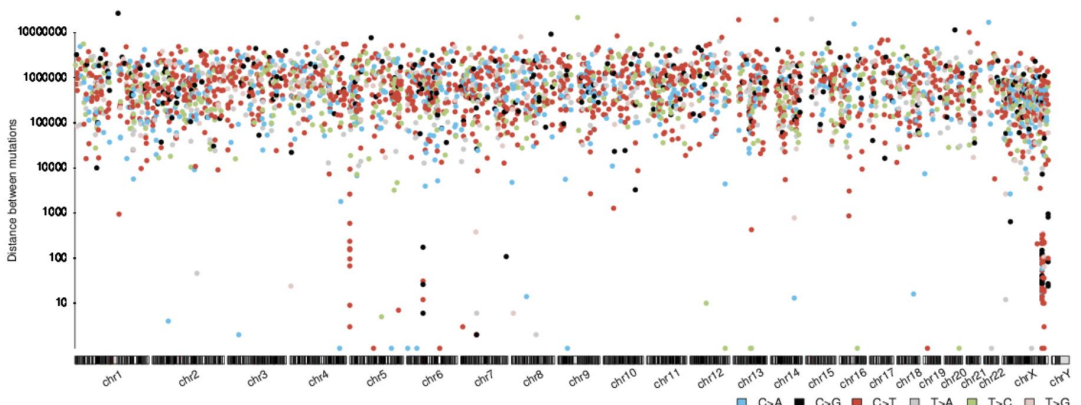
sample_lung

File : sample_lung

Get merged results

Get all results at once

Localized hypermutation (kataegis)



-84-



Interpretation of signature results

Proposed aetiology

Unknown.

Comments

Numbers of mutations attributed to [SBS40](#) are correlated with patients' ages for some types of human cancer.

Proposed aetiology

Unknown. [SBS5](#) mutational burden is increased in bladder cancer samples with ERCC2 mutations and in many cancer types due to tobacco smoking.

Comments

[SBS5](#) is clock-like in that the number of mutations in most cancers and normal cells correlates with the age of the individual. Rates of acquisition of [SBS5](#) mutations over time differ between different cancer types and different normal cell types. These differences do not clearly correlate with estimated rates of stem cell division in different tissues nor with differences in [SBS1](#) mutation rates. [SBS5](#) may be contaminated by [SBS16](#).

Figure 1: Genomic profiles of transcribed and untranscribed regions.

Top Left: Expected vs. Observed GC Content

Region	Expected (%)	Observed (%)
Untranscribed	~20.0	~17.0
Transcribed	~19.0	~16.5

**** p-value = 4.61e-10**

Top Middle: GC Content Distribution (Donut Charts)

Region	C>A (%)	C>G (%)	C>T (%)	T>A (%)	T>C (%)	T>G (%)
Untranscribed (n=559)	14.3	10.2	44.2	8.4	16.1	6.8
Transcribed (n=581)	18.1	9.1	42.7	10.7	15.5	4.0

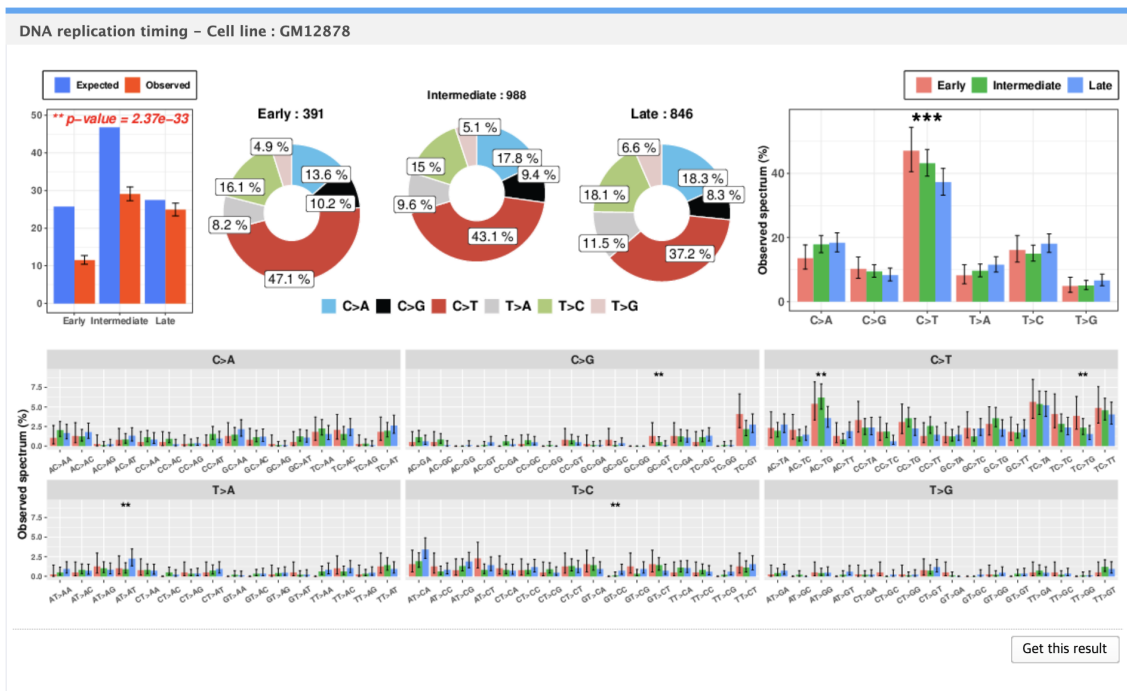
Top Right: Observed GC Spectrum

Dinucleotide Context	Untranscribed (%)	Transcribed (%)
C>A	~13	~19
C>G	~10	~9
C>T	~44	~43
T>A	~9	~11
T>C	~17	~16
T>G	~7	~4

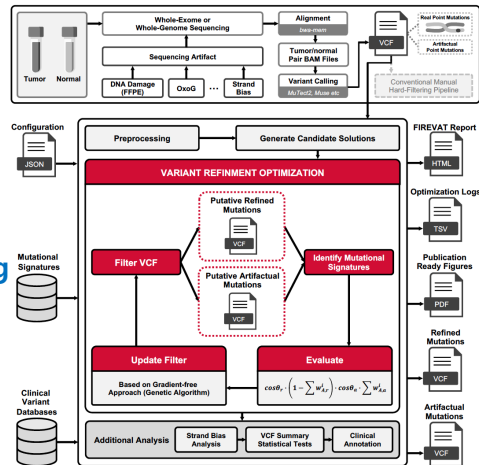
Bottom: Observed GC Spectrum by Dinucleotide Context

The bottom section displays 12 bar charts showing the observed GC spectrum (%) for each dinucleotide context (AC>AA, AC>AG, AC>AT, CC>AA, CC>AG, CC>AT, GC>AA, GC>AG, GC>AT, TC>AA, TC>AG, TC>AT). The contexts are grouped by the first base (A, C, G) and then by the second base (A, G, T). The y-axis represents the observed spectrum (%) from 0.0 to 10.0. The legend indicates the expected (blue) and observed (red) values. Significant differences are marked with asterisks: *** for C>T, ** for T>G, and * for T>A.





FIREVAT: finding reliable variants without artifacts in human cancer samples using etiologically relevant mutational signatures



Sample ID TCGA-V7-A7H0-01A-11D-A33E-09_TCGA-V7-A7H0-10A-01D-A33H-09_mutect_annotated_BRCA

Sample VCF File TCGA-V7-A7H0-01A-11D-A33E-09_mutect_annotated_BRCA.vcf

Sample VCF Genome hg38

Sample VCF Mutation Calling Method Mutect2

Sample VCF Total Point Mutations 2,354

FIREVAT Execution Start Date/Time 2019-01-29 14:49:26

FIREVAT Execution End Date/Time 2019-01-29 17:23:13

FIREVAT Genetic Algorithm (GA) Parameters

GA Population Size 1,000

GA Maximum Iteration 200

GA Run 50

GA Mutation Probability 0.250

Filter Variable	Filter Direction	Optimized Cutoff
Avg Tumor Ref Q88	>	1
Avg Tumor Alt Q88	>	19
Avg Normal Ref Q88	>	14
Tumor REF AD	>	0
Tumor ALT AD	>	0
Normal REF AD	>	0
Normal ALT AD	>	15
Tumor VAF (%)	>	35
Normal VAF (%)	>	35

Objective Value	C.refined	V.refined	C.artifact	W.artifact
G.434	0.993	0.193	0.915	0.592

2.1. Identified Signatures

Original VCF	Refined VCF	Artifact VCF
SB560 0%	SB560 0%	SB560 16.8%
SB506 19.7%	SB506 18.6%	SB506 1.2%
SB504 0%	SB504 0%	SB504 12.4%
SB505 18%	SB505 16.9%	SB505 0%
SB549 3.5%	SB549 4.1%	SB549 0%
SB546 9.9%	SB546 11.7%	SB546 0%
SB543 5.4%	SB543 4%	SB543 12.9%
SB540 0%	SB540 0%	SB540 45.6%
SB54 27.5%	SB54 20.8%	SB54 0%
SB520 0%	SB520 7.4%	SB520 0%
SB518 18%	SB518 21%	SB518 0%
SB51 0%	SB51 0%	SB51 10.5%

Anticlines

- Age
- Damage by Reactive Oxygen Species
- Possible Sequencing Artifact
- Defective DNA Mismatch Repair
- Potential Sequencing Artifact
- Known Sequencing Artifact
- Tobacco Smoking/Chewing
- Likely Sequencing Artifact
- Unknown

3. Optimized VCF Statistics

https://commons.wikimedia.org/wiki/File:2012-06-21_FFPE_block_white_background.jpg

Metastatic colorectal cancer *KRAS* genotyping in routine practice: results and pitfalls

Aude Lamy^{1,2}, France Blanchard^{*1}, Florence Le Pessot^{*1,3}, Richard Sesboué^{6,2}, Frédéric Di Fiore^{2,4}, Jessie Bossut¹, Elodie Fiant¹, Thierry Frébourg^{2,5,6} and

¹Laboratory of Tumor Genetics, University Hospital, Rouen, France; ²Insertm U614, Faculty of Medicine, Institute for Medical Research, Rouen University, France; ³Department of Pathology, University Hospital, Rouen, France; ⁴Digestive Oncology Unit, Department of Gastroenterology, University Hospital, Rouen, France; ⁵Laboratory of Molecular Genetics, University Hospital, Rouen, France and ⁶Department of Genetics, University Hospital, Rouen, France

Sources of erroneous sequences and artifact chimeric reads in next generation sequencing of genomic DNA from formalin-fixed paraffin-embedded samples

Simon Haile¹, Richard D. Corbett¹, Steve Bilobram¹, Morgan H. Bye¹, Heather Kirk¹, Pawan Pandoh¹, Eva Trinh¹, Tina MacLeod¹, Helen McDonald¹, Miruna Baia¹, Diane Miller¹, Karen Novik¹, Robin J. Coope¹, Richard A. Moore¹, Yongjun Zhao¹, Andrew J. Mungall¹, Yussanne Ma¹, Rob A. Holt¹, Steven J. Jones¹ and Marco A. Marra^{1,2,*}

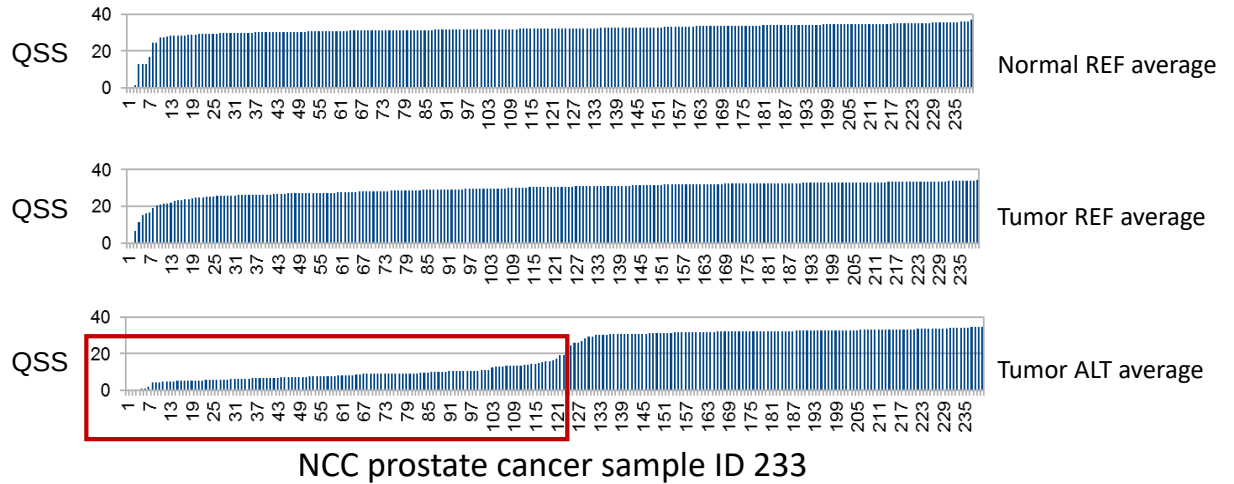
High T790M Detection Rate in TKI-Naive NSCLC with *EGFR* Sensitive Mutation
Truth or Artifact?

Xin Ye, MD, PhD,* Zhong-Zheng Zhu, MD, PhD,† Lei Zhong, M.D.‡, Yachao Lu, MS,*
Yun Sun, BS,* Xiaolu Yin, MD, PhD,* Zhenfan Yang, MD, PhD,*
Guanshan Zhu, MD, PhD,* and Qunsheng Ji, MD, PhD*

Haile *et al.*, NAR 2018

Common problem in variant filtering

How to decide the optimal QSS value ?



177

Installation FIREVAT software in R

- Minimum requirement: R version $\geq 3.4.2$

Getting started

Installation

You can install the released version of `FIREVAT` from CRAN:

```
install.packages("FIREVAT")
```

You can also install the developmental version of `FIREVAT` from GitHub:

```
install.packages("devtools")  
library(devtools)  
install_github("cgab-ncc/FIREVAT")
```

178

Running FIREVAT in R

```
library(FIREVAT)

# Output directory
output.dir <- "" # assign this path

# VCF file
sample.vcf.file <- system.file("extdata", "DCC_PCAWG_Cell_Lines_HCC1954.vcf", package = "FIREVAT")

# Configuration file
config.file <- system.file("config", "PCAWG_DKFZ_Cell_Line_Filtering_Params.json", package = "FIREVAT")

# Run FIREVAT
results <- RunFIREVAT(vcf.file = sample.vcf.file,
                      vcf.file.genome = 'hg19',
                      config.file = config.file,
                      df.ref.mut.sigs = GetPCAWGMutSigs(),
                      target.mut.sigs = GetPCAWGMutSigsNames(),
                      sequencing.artifact.mut.sigs = PCAWG.All.Sequencing.Artifact.Signatures,
                      output.dir = output.dir,
                      objective.fn = Default.Obj.Fn,
                      num.cores = 2,
                      ga.pop.size = 100,
                      ga.max.iter = 5,
                      ga.run = 5,
                      perform.strand.bias.analysis = TRUE,
                      ref.forward.strand.var = "TumorDPRefForward",
                      ref.reverse.strand.var = "TumorDPRefReverse",
                      alt.forward.strand.var = "TumorDPAltForward",
                      alt.reverse.strand.var = "TumorDPAltReverse",
                      annotate = FALSE)
```

179

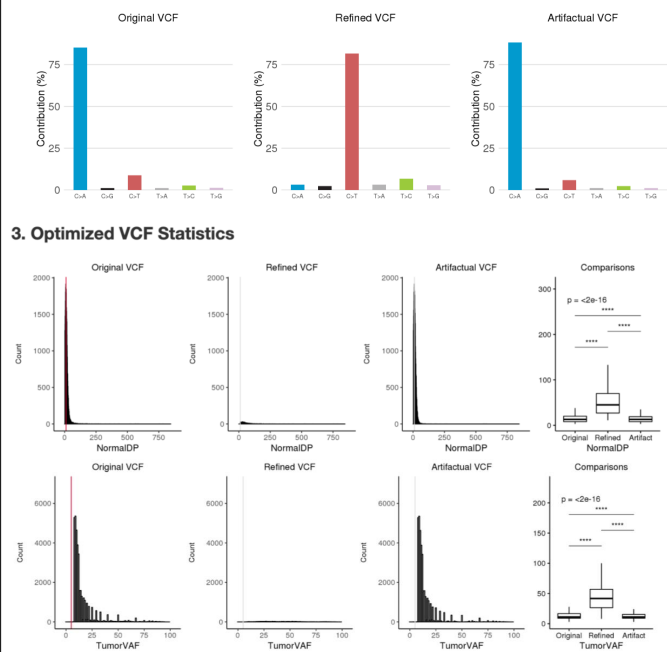
```
library(FIREVAT)
setwd("/Users/dwhong/Desktop/R_code/firevat")

# SKCM example
SKCM.example.vcf <- "/C828.TCGA-EB-A24D-01A-11D-A197-08.2.varscan.snp.annotated.tcga_filtered.reheadered_Original.vcf"
SKCM.config.file <- "/MC3_Varscan_filtering_params.json"
SKCM.output.dir <- "/SKCM_example/"

# Run FIREVAT
SKCM.example.results <- RunFIREVAT(
  vcf.file = SKCM.example.vcf,
  vcf.file.genome = 'hg19',
  config.file = SKCM.config.file,
  df.ref.mut.sigs = GetPCAWGMutSigs(),
  target.mut.sigs = GetPCAWGMutSigsNames(),
  sequencing.artifact.mut.sigs = PCAWG.All.Sequencing.Artifact.Signatures,
  output.dir = SKCM.output.dir,
  num.cores = 3,
  ga.pop.size = 100,
  ga.max.iter = 10,
  ga.run = 10,
  perform.strand.bias.analysis = FALSE,
  annotate = FALSE
)
```

firevat_run.R

FIREVAT refinement on TCGA-EB-A24D (SKCM)



2.2. Trinucleotide Spectrums

	Original VCF	Refined VCF	Artfactual VCF
Mutations Count (%)	36,453 (100%)	1,367 (3.75%)	35,086 (96.25%)
Cosine Similarity Score	0.989	0.996	0.989
Residual Sum of Squares (RSS)	0.00267	0.000795	0.00291

4. Variants with Strand Bias

4.1. Refined VCF

None to display.

4.2. Artfactual VCF

CHROM	POS	REF	ALT	PrimaryDPRefForward	PrimaryDPRefReverse	PrimaryDPAITForward	PrimaryDPAITReverse	StrandBiasQValue
chr1	16,918,473	G	A	434	275	2	80	0.0000000
chr2	95,513,889	T	A	136	209	4	69	0.0000199
chr2	95,513,893	A	G	126	210	4	71	0.0000316
chr2	95,513,877	A	G	172	173	5	46	0.0000829
chr2	95,513,880	C	T	160	180	5	51	0.0000829

5. VCF Annotation (ClinVar)

5.1. Refined VCF

CHROM	POS	REF	ALT	GENEINFO	CLNSIG
chr7	140,453,134	T	C	BRAF:673	Pathogenic
chr7	140,501,337	T	G	BRAF:673	Pathogenic/Likely_pathogenic

5.2. Artfactual VCF

None to display.

183

Conclusions

1. FIREVAT is a quick and efficient variant refinement tool using VCF files
2. Variant refinement can be performed using mutational signatures
3. <https://github.com/cgab-ncc/FIREVAT>

184

Self Test #1

- Mutalisk 웹 사이트 (<http://mutalisk.org>)에서 제공하는 melanoma 환자의 돌연변이 데이터 (sample_melanoma.vcf)를 다운로드 받아서 분석을 진행하고 다음 물음에 답하시오.
- Version 2와 Version 3을 이용한 시그니처 분석 후 1st, 2nd signature 번호 (타입)는 각각 무엇인가?
- Replication timing에서 돌연변이가 가장 많은 구간과 유의미한 돌연변이는 무엇인가? (Version 2, 3 각각에 대해 답하시오.)

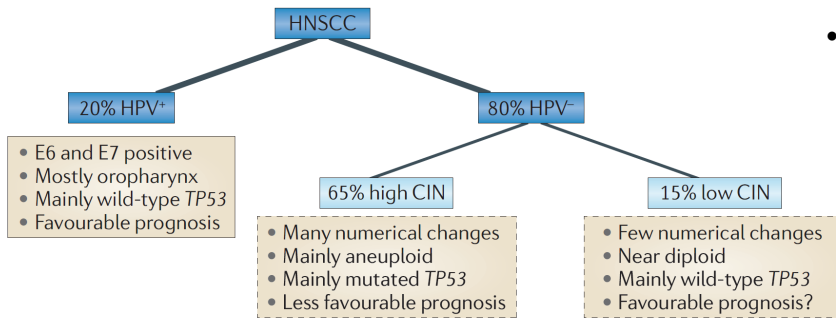
185

Self Test #2

- Data 폴더에서 제공하는 두경부암 돌연변이 파일(TCGA-CR-7399-01A-11D-2012-08_TCGA-CR-7399-10A-01D-2013-08_mutect_annotated.vcf)에서 artifacts 시그니처와 남은 정제된 시그니처 타입 (번호)은 각각 무엇인가?
 - 분석 시 parameter option은 Data 폴더에서 제공하는 파일 (MuTect2_Filtering_Params.json)을 활용하시오.
- 전체 돌연변이 개수, artifact 그룹의 돌연변이 개수, refined 그룹의 돌연변이 개수를 각각 구하시오.

186

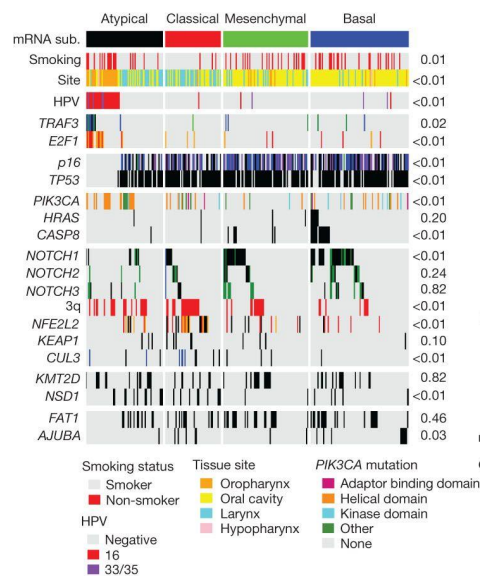
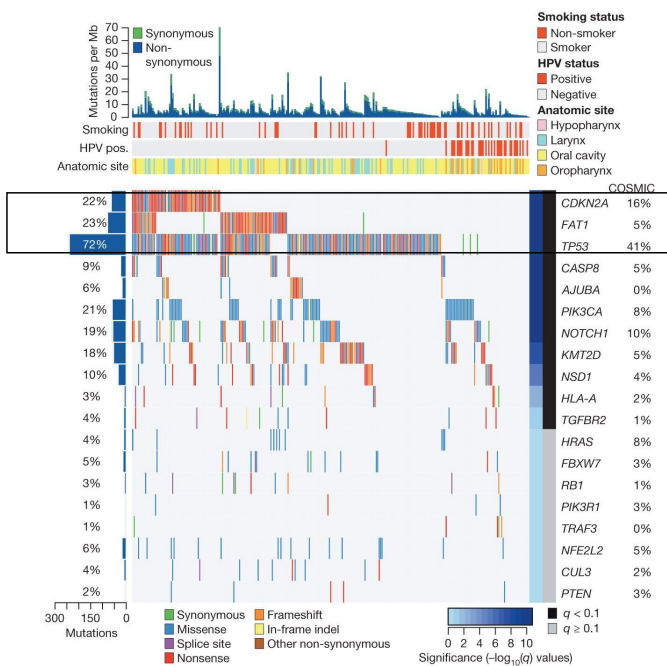
Head and Neck Cancer Study



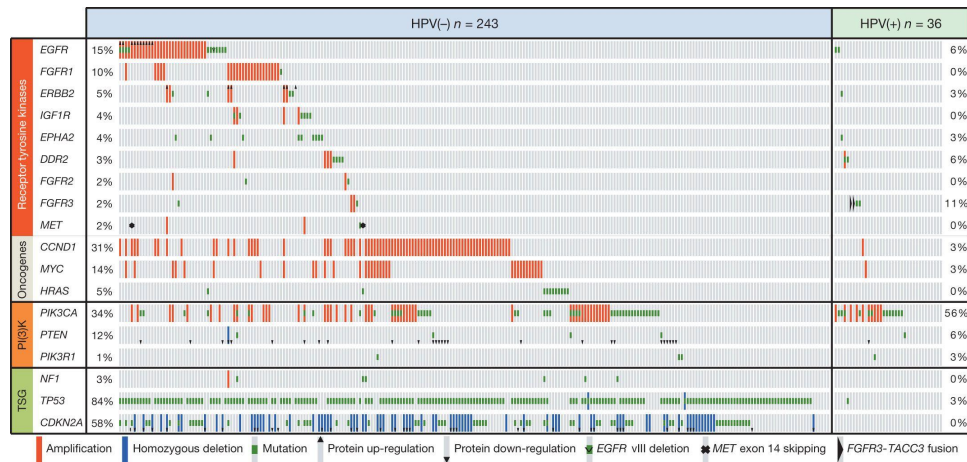
- Three genetic subclasses of HNSCCs
 - **HPV(+)** (human papillomavirus)
 - **HPV(-)**
 - **High CIN** (chromosome instability)
 - **Low CIN**

Leemans et al., Nature reviews cancer (2011)

TCGA HNSCC study (2015, Nature)

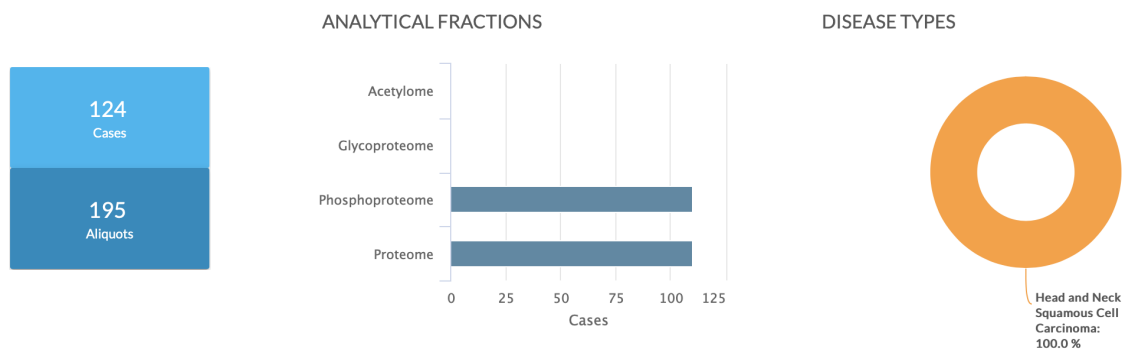


Targetable genes in HNSCC?



개별 환자의 therapeutic target을 확인했을 뿐,
암 발병의 기전이나
임상, 치료적 측면에 활용가능한 환자군 구별 방법은 없음

CPTAC



PDC Study ID	Study	Project	Program	Disease Type	Primary Site	Analytical Fraction	Experiment Type	Cases #	Raw
PDC000222	CPTAC HNSCC Discovery Study - Phosphoproteome	CPTAC3-Discovery	Clinical Proteomic Tumor Analysis Consortium	Head and Neck Squamous Cell Carcinoma	Head and Neck	Phosphoproteome	TMT11	110	260
PDC000221	CPTAC HNSCC Discovery Study - Proteome	CPTAC3-Discovery	Clinical Proteomic Tumor Analysis Consortium	Head and Neck Squamous Cell Carcinoma	Head and Neck	Proteome	TMT11	110	500

Article

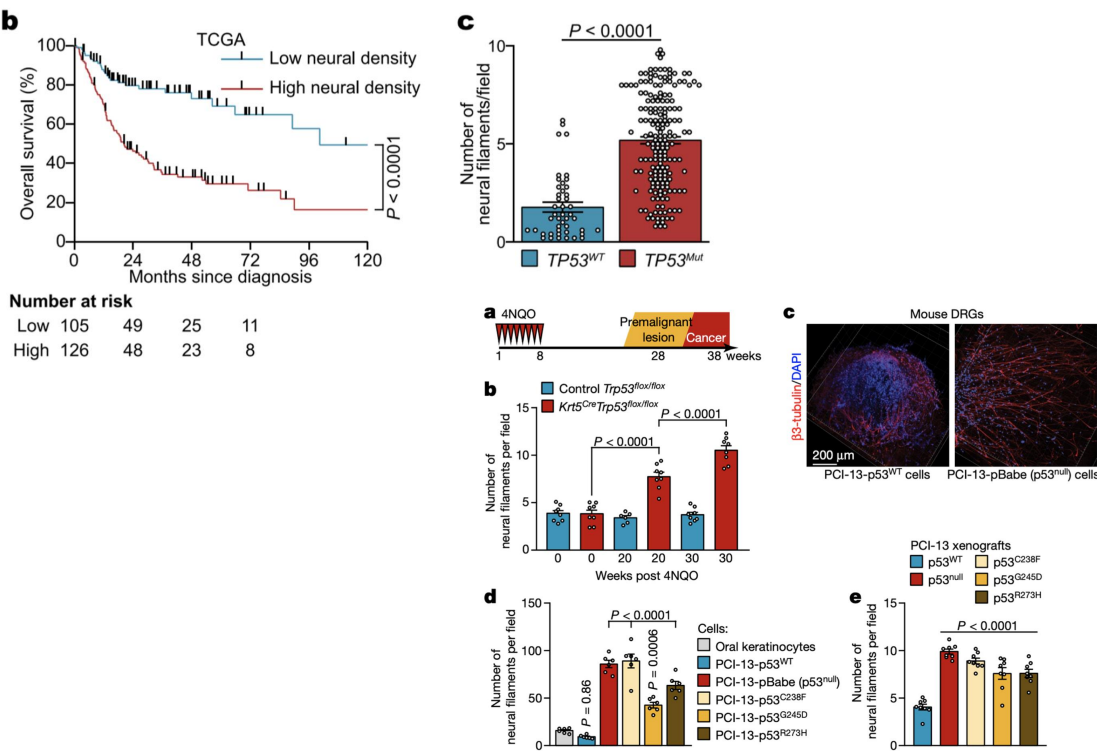
Loss of p53 drives neuron reprogramming in head and neck cancer

https://doi.org/10.1038/s41586-020-1996-3
Received: 3 December 2018
Accepted: 15 January 2020
Published online: 12 February 2020
Check for updates

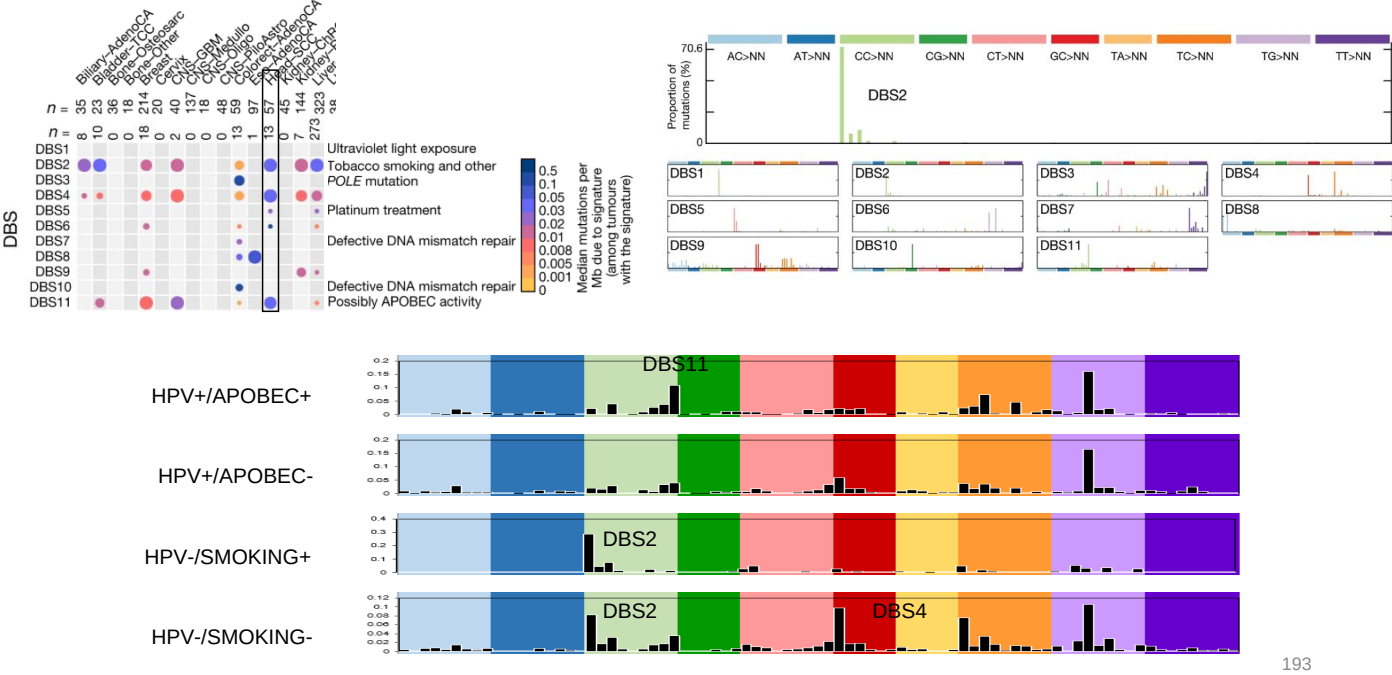
Moran Amit^{1,2,3,4}, Hideaki Takahashi^{1,2,3,5}, Mihnea Paul Dragomir³, Antje Lindemann¹, Frederico O. Gleber-Netto¹, Curtis R. Pickering¹, Simone Anfossi¹, Abdullah A. Osman¹, Yu Cai¹, Rong Wang¹, Erik Knutsen^{1,4}, Masayoshi Shimizu^{1,5}, Cristina Ivan^{1,5}, Xiayu Rao⁶, Jing Wang⁶, Deborah A. Silverman⁷, Samantha Tam¹, Mei Zhao¹, Carlos Caulin^{8,9}, Assaf Zinger^{10,11}, Ennio Tasciotti^{10,11}, Patrick M. Dougherty¹², Adel El-Naggar¹³, George A. Calin^{1,5,12} & Jeffrey N. Myers^{14,15}

Loss of p53 alters neural milieu

To evaluate the impact of tumour innervation in head and neck cancer, we analysed survival data from The Cancer Genome Atlas. High neural density in oral cavity squamous cell carcinoma (OCSCC) was associated with poorer overall survival ($P < 0.0001$, log-rank test) and with the presence of *TP53* mutations ($P < 0.0001$, Extended Data Fig. 1a–c). To study the interplay between nerves and epithelial cell p53 function throughout tumorigenesis, we evaluated neuritogenesis over the course of the progression from precursor lesions to high-grade lesions in *Trp53*^{flox/flox} wild-type control and *Krt5*^{cre}*Trp53*^{flox/flox} (*Trp53*^{-/-} in the recombined epithelial tissue) mouse models of oral cancer⁸ (Fig. 1a).



MutaliskR DBS signature

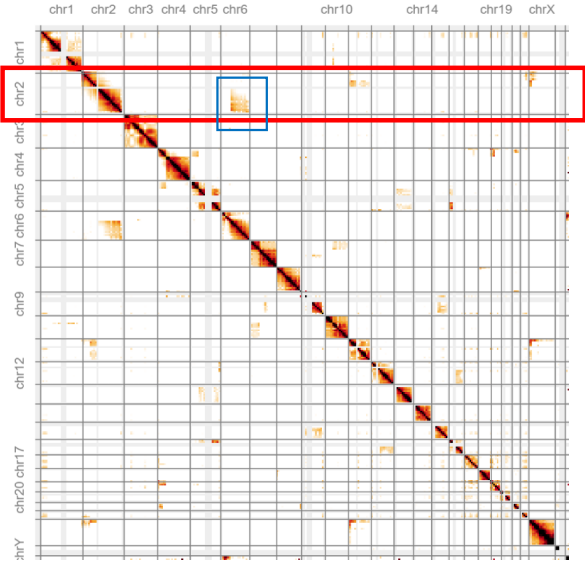
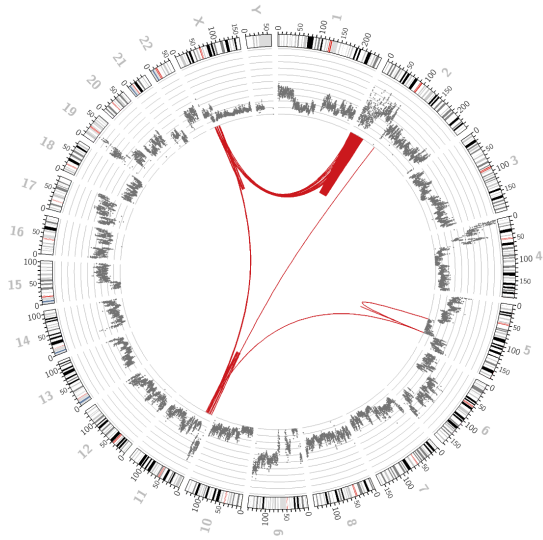


193

Integrative SV analysis with Hi-C (Chromothripsis)

HNT-073

sample	cluster	n.sv	svaf1	svaf2	circos.color	classification
HNT-073	24	244	0.124	0.134	red	multichromosomal_chromothripsis



TO Part III

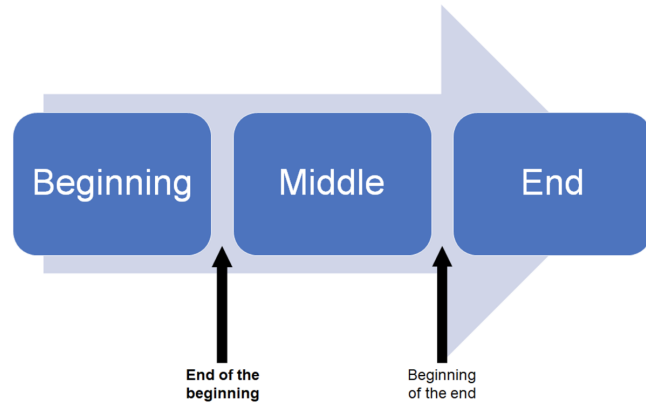


PART III

- **Part I: Cancer Big Database**
 - History of cancer genome study
 - Cbioportal, Xena browser and GDC data portal
 - Study example of cancer big database
- **Part II: Cancer Genome Analysis**
 - Mutational Signature Analysis: Mutalisk and FIREVAT
 - Data driven subtyping of cancer
- **Part III: Artificial Intelligence Study using Cancer Big Data**

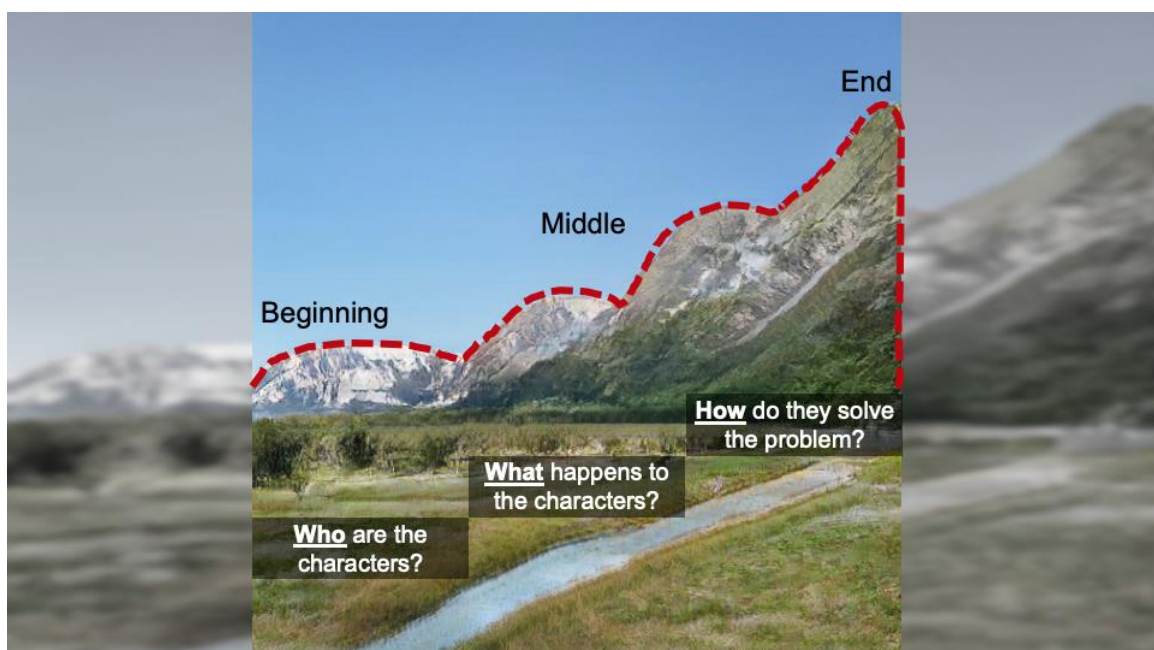
PART III

At the End of the Beginning of Cancer Genomics



	Beginning	Middle
Interest	Genome research	Genome medicine
Value	Individual report	Well curated data
Research size	Case reports	Systemic analysis of large cohort
Therapy	Chemotherapy	Immunotherapy

Archetypal Story Arc



Actually, it's All About AI



<http://nvidia-research-mingyuliu.com/gaugan>

AI

Deep-learning is the hottest
topic in cancer genomic
research area

Latest Cancer Research Trend Analysis - Top 15 Bigrams

~250,000 articles / 1 year (pubmed)

As of August 1, 2019

Rank	Keyword	2015	2016	2017	2018	2019	Cumulative growth ratio (2018/2015)
1	deep learning	127	284	790	2,107	2,285	16.69
2	pd1 blockade	0	6	8	9	9	10.00
3	circular rna	53	96	249	512	458	9.50
4	health-related quality	0	3	4	8	5	9.00
5	checkpoint inhibitor	99	239	488	784	679	7.85
6	circular rnas	80	140	308	612	567	7.57
7	mutational burden	30	61	107	225	230	7.29
8	blood-brain barrier	5	8	9	40	23	6.83
9	epstein-barr virus	2	1	5	18	8	6.33
10	next-generation sequencing	2	9	15	17	18	6.00
11	progression-free survival	4	5	10	28	20	5.80
12	checkpoint inhibitors	318	767	1,292	1,844	1,594	5.78
13	endogenous rna	82	119	239	467	463	5.64
14	receptor t	57	110	194	320	262	5.53
15	liquid biopsies	47	124	193	250	179	5.23

Latest Cancer Research Trend Analysis - Cancer Genomics

Keyword	2015	2016	2017	2018	2019	Cumulative growth ratio (2018/2015)
deep learning	127	284	790	2,107	2,285	16.47
nextgeneration sequencing	2	9	15	17	18	6.00
cancer database	213	274	515	685	598	3.21
machine learning	1,731	2,280	3,287	5,405	4,834	3.12
cancer genome	789	1,017	1,359	1,904	1,614	2.41
germline variants	71	85	117	159	143	2.22

AI ?

Strong AI = General AI

공상과학 영화



터미네이터



A.I



아이로봇



업그레이드

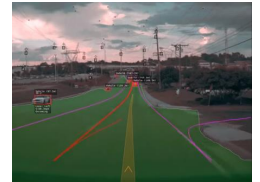
Weak AI = Narrow AI



Google 광고 개인 최적화



Apple Watch 부정맥 감지



TESLA
Tesla autopilot recognition

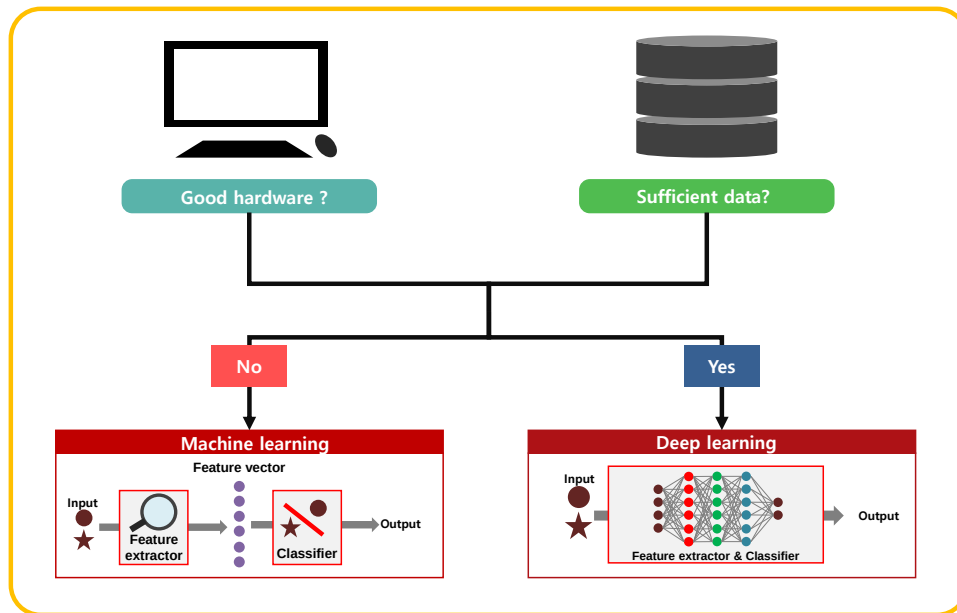
모든 부분에서 AR 가능한 로봇 또는 인간 수준 AI

특정 임무를 수행하는 AI

Cancer drugs

- Vemurafenib and trametinib
BRAF V600E mutations in melanoma
- Erlotinib 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

Which goal ?



컴퓨팅 파워의 발전

PS3
PlayStation 3



최소 출시: 2006년 11월 11일
최초 가격: \$499
CPU: 3.2 GHz 셀 브로드밴드 엔진 1
PPE, 8 SPEs
RAM: 시스템 RAM: 256MB (XDR DRAM)
비디오 RAM: 256MB (GDDR3)
GPU: 550 MHz 엔비디아/SCEI RSX 리얼리 신시사이저

PS3 1대 FLOPS = 230.4 GFLOPS

미공군 336대의 ps3 를 이용한 도시감시용 슈퍼컴퓨터

500 TFLOPS



XBOX
SERIES X



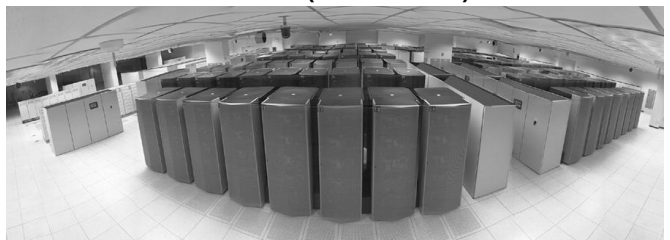
최소 출시: 2020년 11월 10일
최초 가격: \$499
CPU: AMD ZEN 2 기반 커스텀 마이크로 아키텍처
제원: 8코어 16스레드, 3.6 GHz
RAM: 16 GB, GDDR6 ECC SGRAM
GPU: AMD RDNA 2 기반 커스텀 마이크로 아키텍처

XBOX 1대 FLOPS = 12.1472 TFLOPS



2001 슈퍼 컴퓨터 ACSI white
핵무기 관련 시뮬레이션용 사용 (농구장 2배 크기)

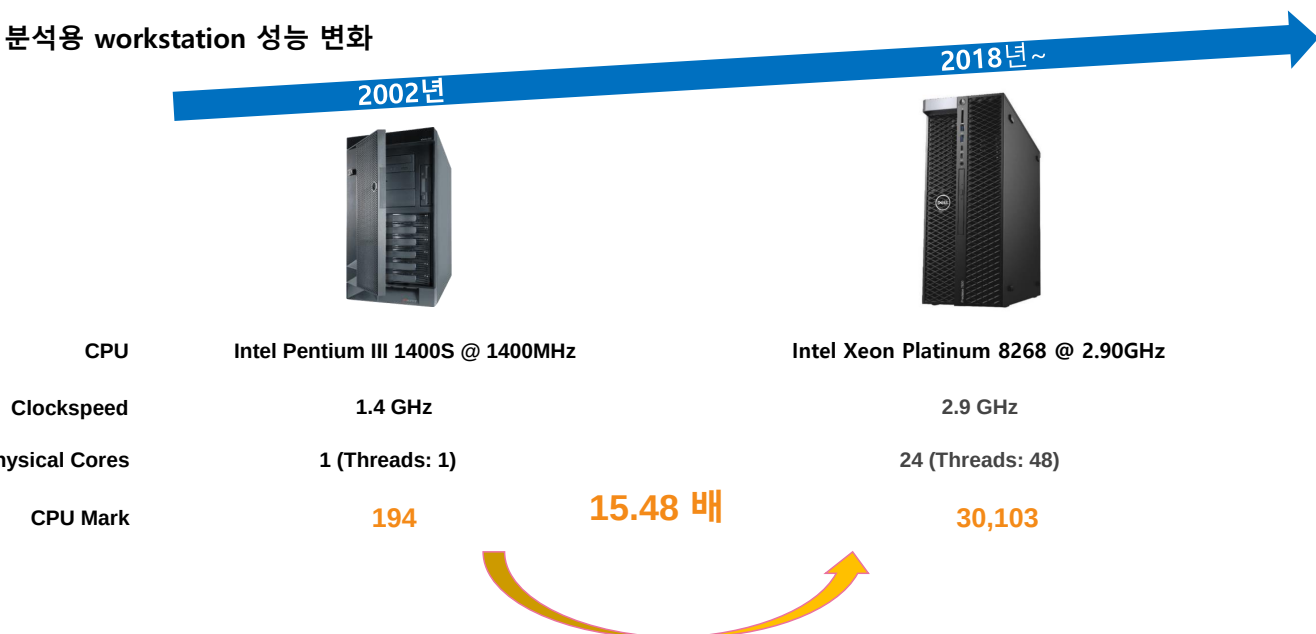
12.3 TFLOPS



*: Teraflops: 초당 10¹²회의 연산을 처리 할 수 있는 계산 능력

컴퓨팅 능력의 저변 확대

Data 분석용 workstation 성능 변화



인공 지능을 이용한 연구 환경 조성 ??

ex)

AlphaGo
AlphaGo Zero



3hours

19hours

70hours

$$361! = 2.6 \times 10^{845}$$

가장 큰 수의 단위: 10^{68} 무량대수
읽을 수 있는 수: 10^{71} 천무량대수

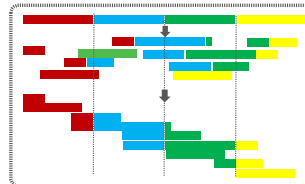
인간의 한계

오믹스 데이터

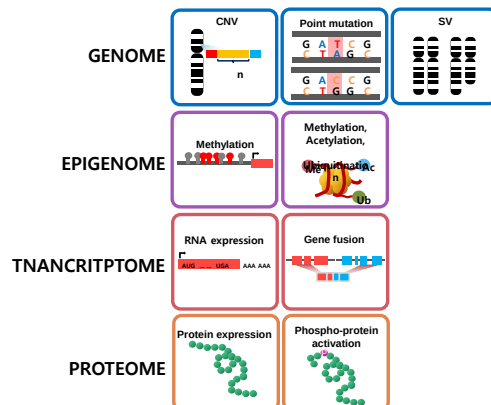
1) 일종의 디지털 데이터

A T G C
(30쌍)

2) 대량의 정보

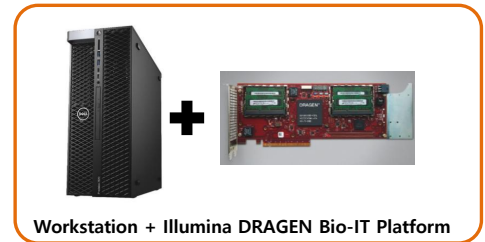
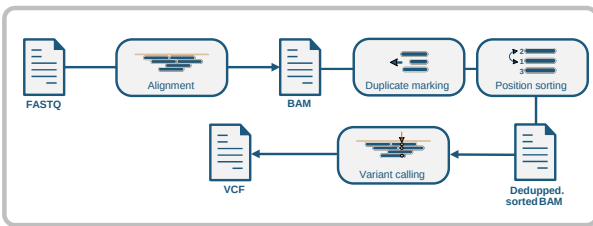


3) 복잡성

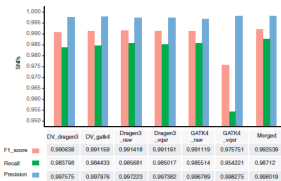


Pipeline 계산 능력

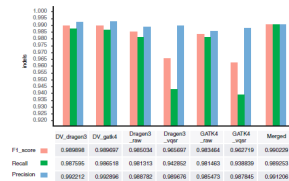
Germline variant calling pipelines



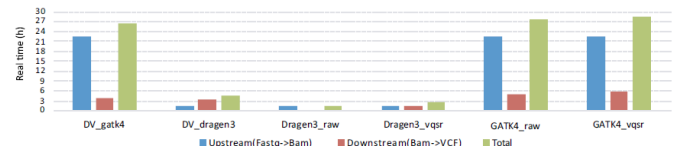
A. NA12878_PrecisionFDA - SNPs



B. NA12878_PrecisionFDA - indels



A. NA12878_PrecisionFDA



Sci Rep 2020 Nov 19;10(1):20222.

DRAGEN3 vs GATK4 : 오차 범위 내 동일 수준 정확성

동일 데이터 처리 (fastq to vcf) 기준;
9배(GATK4; 27시간, dragen; 3시간) 시간 소요

2.7. Machine learning 과 Deep learning

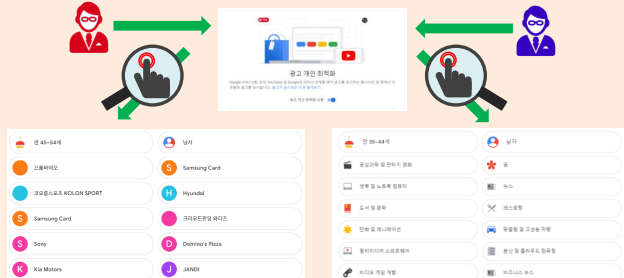
Machine learning

: 컴퓨터가 데이터를 통해 스스로 학습하여 예측이나 판단을 제공하는 기술

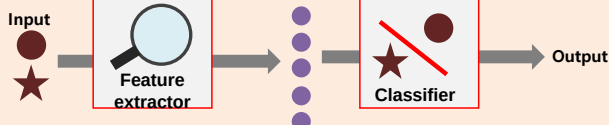
ex)

구글 광고 개인 최적화

(<https://adssettings.google.com/authenticated>)



Human Feature vector

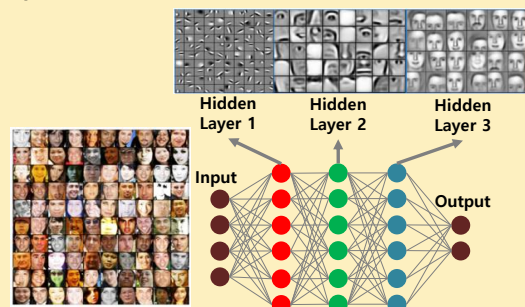


Deep learning

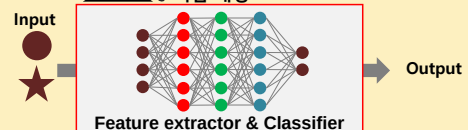
: 깊은 인공신경망을 알고리즘을 활용하는 머신러닝 기술

ex)

딥러닝 알고리즘을 활용한 얼굴 인식 프로세스



학습 대상



Three types of **research questions** driving application of **AI** in **genomics**
(Ching *et al.*, J R Soc Interface 2018)

1. Disease and patient categorization
(**Gene expression, DNA methylation**)
2. Functional and biological study (DNA sequences)
3. Treatment of patients (Gene expression)

Three types of **research questions** driving application of **AI** in **genomics**
(Ching *et al.*, J R Soc Interface 2018)

1. **Disease and patient categorization**
2. Functional and biological study
3. Treatment of patients

Here, NBA players are clustered in 8 newly suggested positions by a machine-learning approach.

[illegible]

17 histological types

Three types of **research questions** driving application of **AI** in **genomics**
(Ching *et al.*, J R Soc Interface 2018)

1. Disease and patient categorization
2. **Functional and biological study**
3. Treatment of patients

Different Tissues & Physiology



Caterpillar



Butterfly

Different Tissues & Physiology, Same Genome

Nucleotide sequences

```
...CGCCGCTGACCTATCATCAGTTC
CAAGCGCTGATAGCGAGCATGCC
CCCGCCTCCGTCCGCCGAACCCA
CCATCAGTTTGGAGACAOTCAACC
GCGCCGTTACACCTATCTCGGATA
ATCACGACGAACGATTTGGAGTGC
CGACACTCGAAGAACTTGGCTTCG
ATACGGAAGGTCTTAAACCTCCAA
TATGGATCGGCGGAGAAAACGAA
GCTCTGTTGAGACT...
```

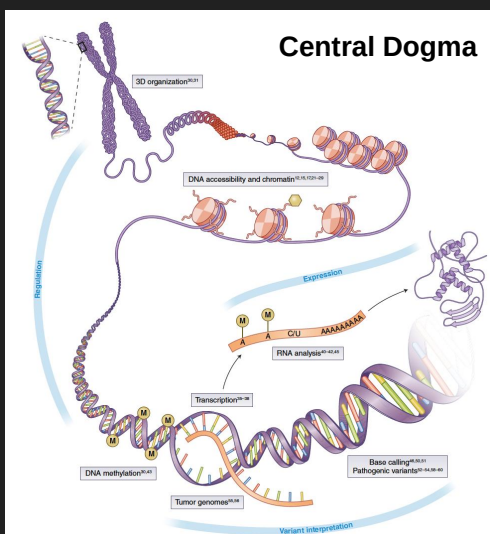
=

```
...CGCCGCTGACCTATCATCAGTTC
CAAGCGCTGATAGCGAGCATGCC
CCCGCCTCCGTCCGCCGAACCCA
CCATCAGTTTGGAGACACTCAACC
GCGCCGTTACACCTATCTCGGATA
ATCACGACGAACGATTTGGAGTGC
CGACACTCGAAGAACTTGGCTTCG
ATACGGAAGGTCTTAAACCTCCAA
TATGGATCGGCGGAGAAAACGAA
GCTCTGTTGAGACT...
```

Caterpillar

Butterfly

Many, So Many Components in Central Dogma

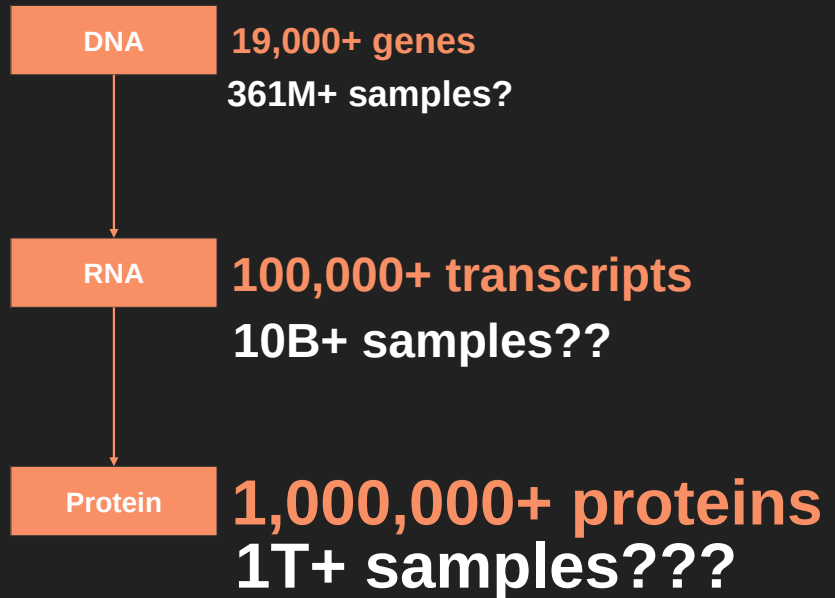


Zou *et al.*, Nat Genetics 2019

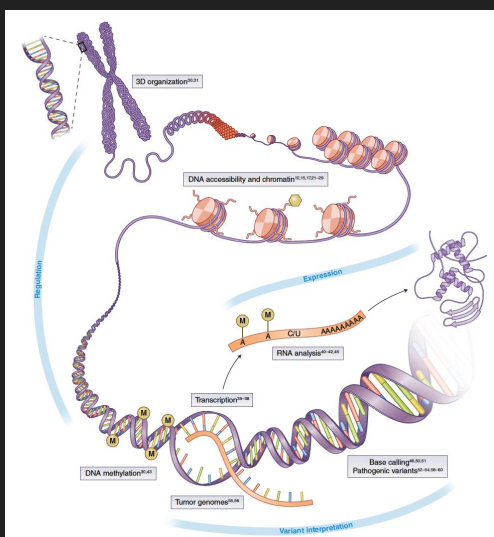


Current Problems in Applying Deep Learning to Genomics

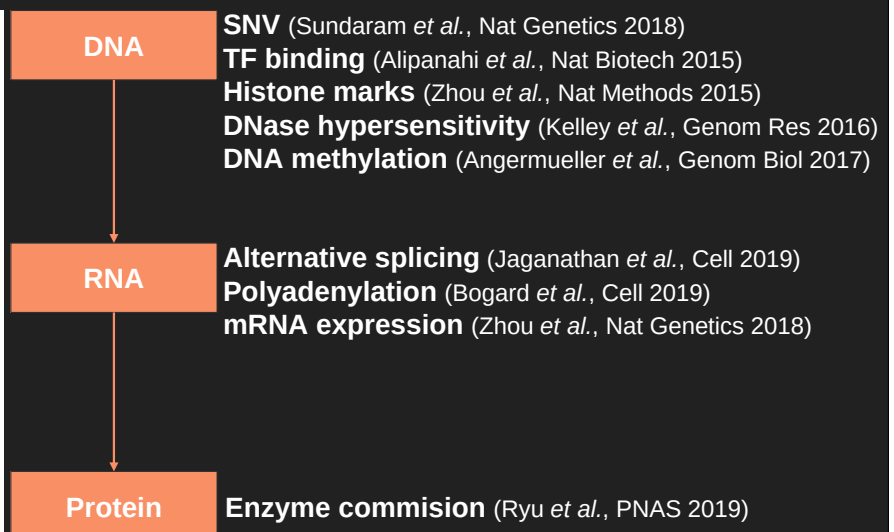
To conduct optimal learning, we need n^2 samples for n features for correlated features
(Hua *et al.*, Bioinformatics 2005)



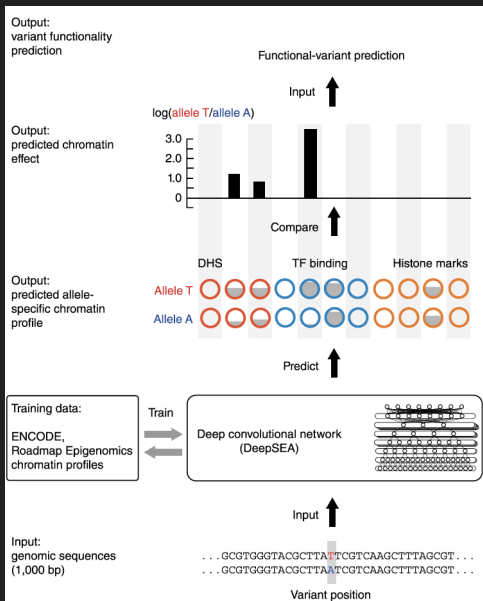
Deep Learning Research in Genomics



Zou *et al.*, Nat Genetics 2019

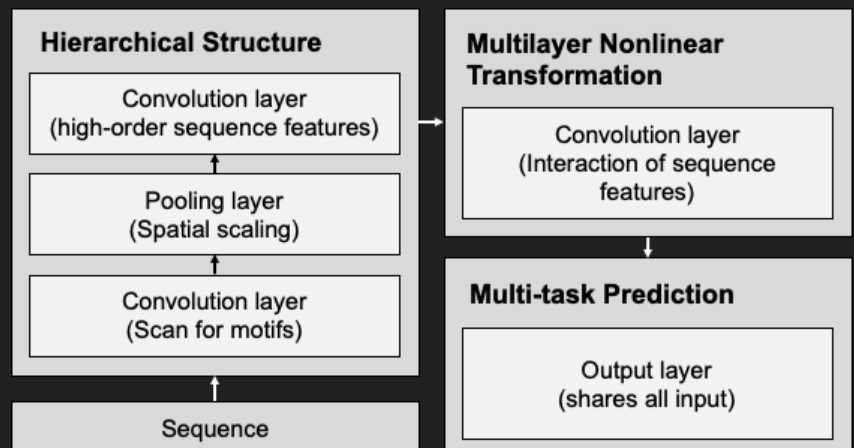


Deep Learning Research in Genomics - DeepSEA



DeepSEA (Zhou *et al.*, Nat Methods 2015)

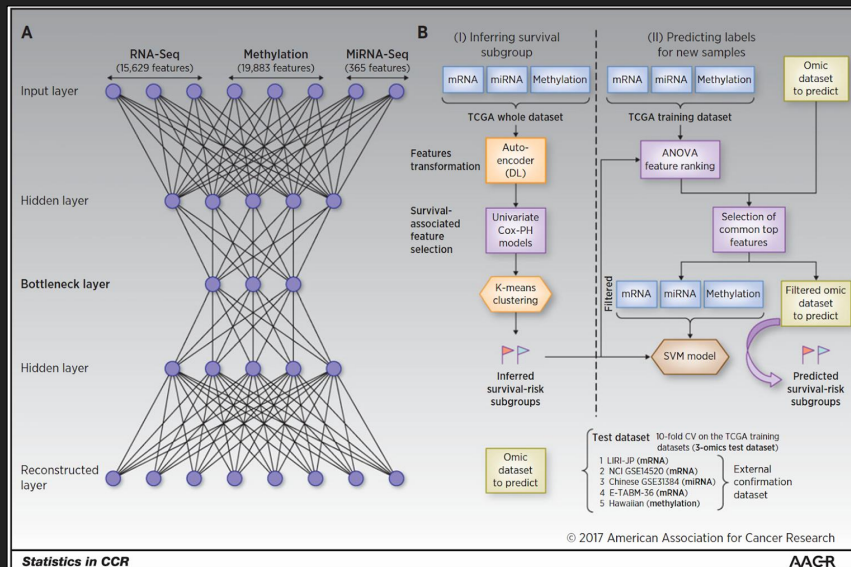
Using genomic sequences (1,000 bp) predict chromatin organization (transcription factor binding, histone marks, DNase sensitivity).



Three types of **research questions** driving application of **AI** in **genomics** (Ching *et al.*, J R Soc Interface 2018)

1. Disease and patient categorization
2. Functional and biological study
3. **Treatment of patients**

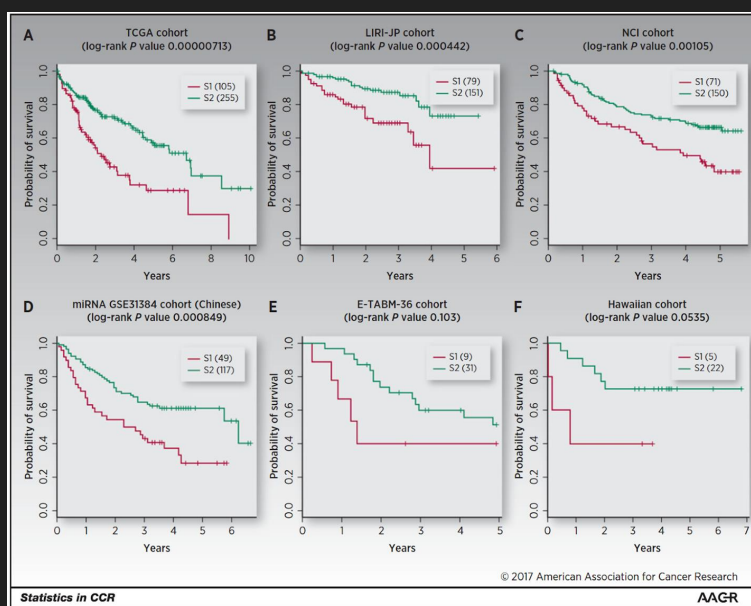
Deep Learning Research in Genomics - Predicting Prognosis



Chaudhary *et al.*, Clin Can Res 2018

- Predicting survival of liver cancer patients by training an **autoencoder**.
- Trained on 360 TCGA-HCC samples' **mRNA** and **miRNA sequencing** as well as **methylation data**

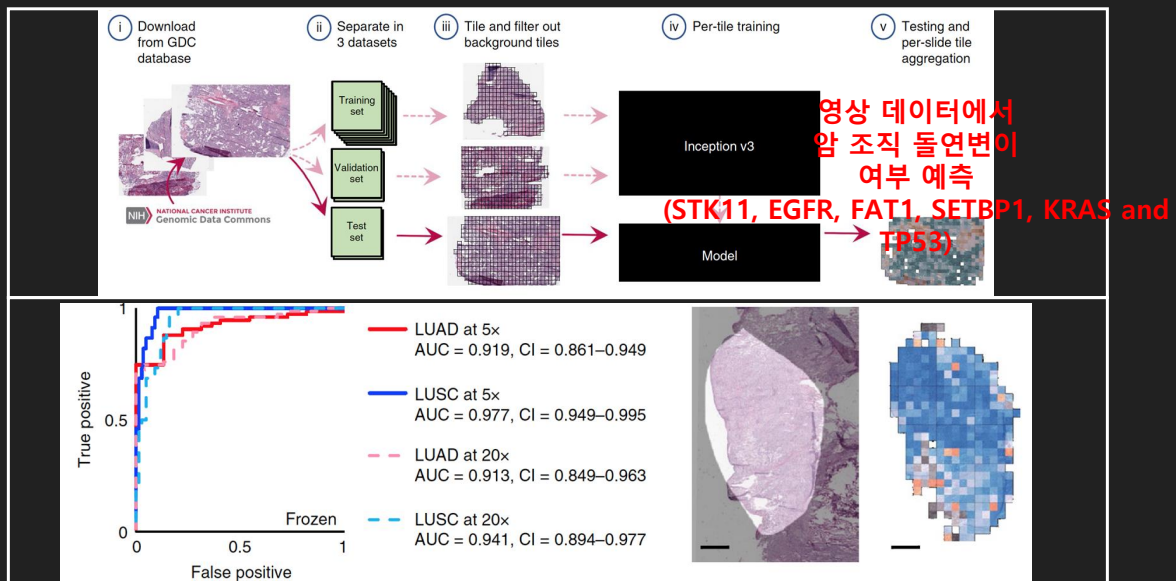
Deep Learning Research in Genomics - Predicting Prognosis



Chaudhary *et al.*, Clin Can Res 2018

- Validated the deep learning model on 684 patients across 5 independent cohorts

Deep Learning Research in Biomedicine



Coudray et al., Nature Medicine 2018

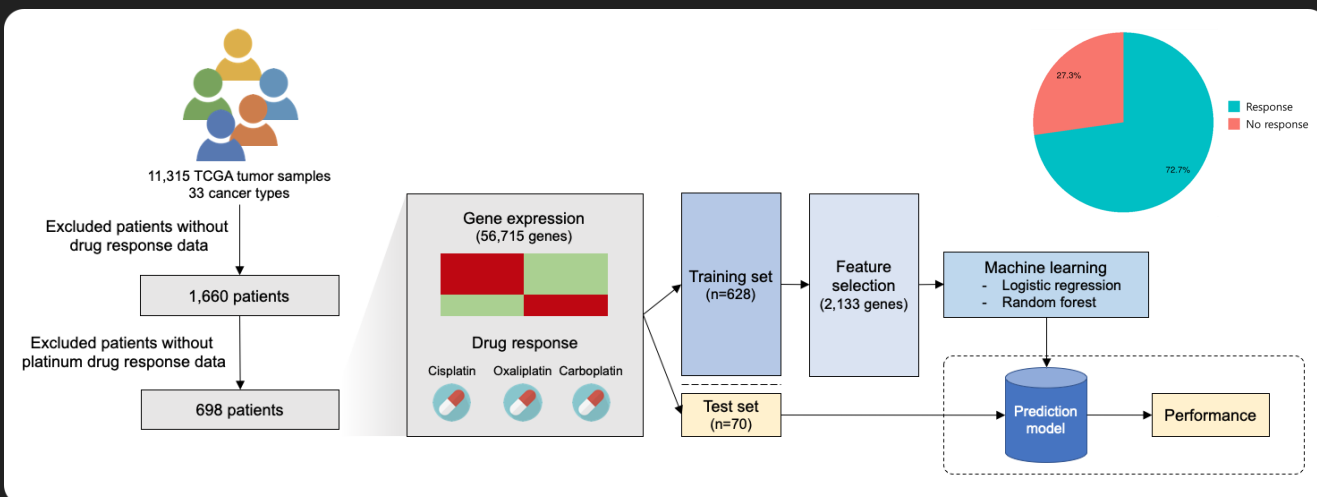
Use machine learning algorithm



ARBTR: An aRtificially intelligent Bayesian approach to predicting predisposition and evoluTionary deteRminants in human cancer

Challenges in Cancer Genomic Data for Deep Learning

Our pilot study predicting platinum therapy response using omics data



Challenges in Cancer Genomic Data for Deep Learning

Our pilot study predicting platinum therapy response using omics data

Method/Model	AUC
Random forest	0.619*
SVM	0.612
Ada boost	0.609
Logistic regression	0.593
MLP	0.593

Genomic/transcriptomic features (selection) matter a great deal.

Due to currently limited sample size in publicly available datasets, we must devise new ways of tackling the problem of predicting therapeutic response in cancer patients

So why are genomic/transcriptomic features problematic (besides the dimensionality problem)?

Because of tumor heterogeneity

Dagogo-Jack *et al.*, Nat Rev Clin Oncol 2018

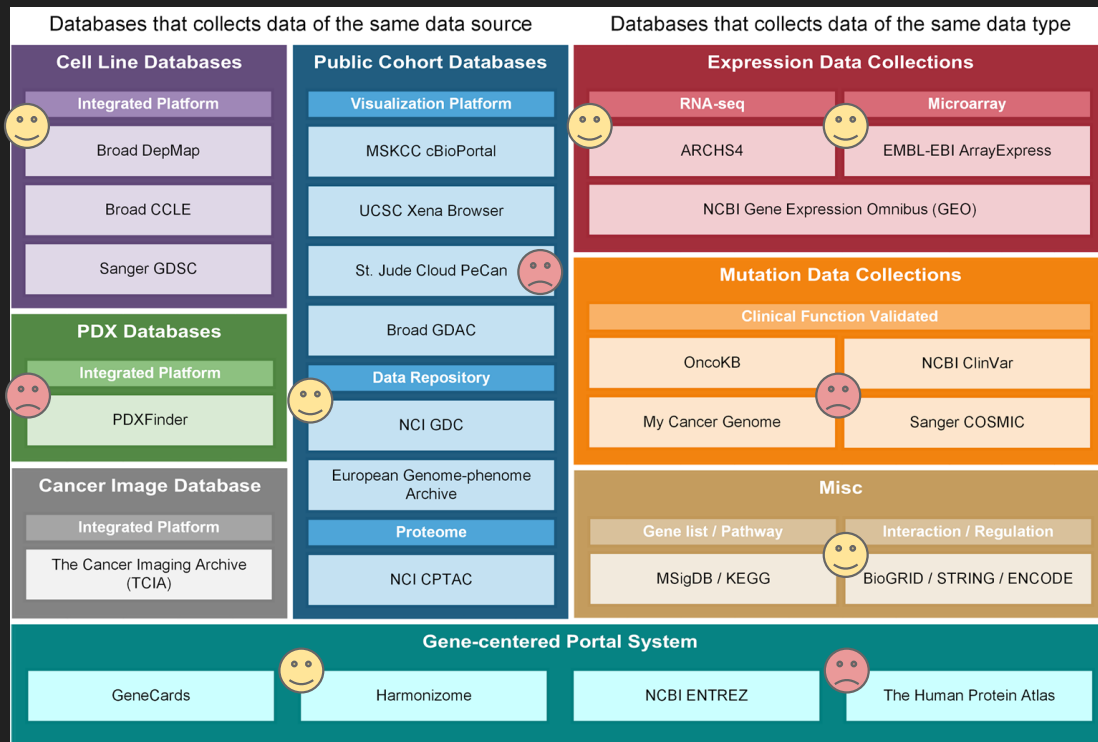
Current Problems in Applying Deep Learning to Genomics



Suitable for deep-learning



Not suitable for deep-learning



Cancer Omics Data

Cell line data



Genomics of Drug Sensitivity in Cancer

<https://www.cancerrxgene.org/>



Cancer Cell Line Encyclopedia

<https://portals.broadinstitute.org/ccle>

Summary

	GDSC1	GDSC2
Cell line	987	809
Compounds	367	198
IC50	310,904	135242

Methods & dataset

Exome sequencing
Array exome sequencing
Mutation
Copy Number
Methylation
Expression
Drug Screening – IC50s

Summary

	CCLE
Cell line	1,457
Compounds	24 (504 cell line test)

Dataset

WES
Mutation
Fusion
structural variant
miRNA
Global Chromatin
RPPA
Antibody
RPKM

RSEM
Methylation
copy number
metabolomics

Cancer Omics Data

Human genomic data



<https://portal.gdc.cancer.gov/>

Summary

GDC

Projects	67
Primary sites	68
Cases	84,392
Genes	23,399
Files	596,758
Mutations	3,287,299

Method

Clinical data	Diagnostic image
Biospecimen data	Tissue image
Pathology Reports	Radiological image
SNP microarray	Bisulfite sequencing
Copy number microarray	Bead array
Low-Pass DNA Sequencing	miRNA Sequencing
Whole exome	Total RNA Sequencing
Whole genome	Microarray
SNP microarray	Reverse-Phase Protein Array
Sequence trace	



<https://dcc.icgc.org/>

Summary

ICGC

Data release 28	86
cancer project	22
Cancer primer sites	22,230
Donor with molecular data in DCC	24,289
Total donors	
Simple somatic mutations	81,782,588

Method

Clinical data
Biospecimen Data
WXS
WGS
RNA-Seq
miRNA-seq
Bisulfite-seq



<https://cptac-data-portal.georgetown.edu/>

Summary

ICGC

Studies	55
Tumor sites	12
Cases	2,549
Samples	3,639
Files	107,493
Data	25,576GB

Method

Proteome
Phosphoproteome
Acetylome
Glycoproteome
Ubiquitylome

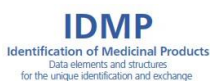


Issues to be considered

데이터 표기의 문제

▶ 각각의 데이터 표기 기준 안에 따른 데이터 구성의 문제 발생

◇ 약품 표기법



◇ 의학용어 표기법



ex)1. IL2 = IL-2 = interleukin - 2 = interleukin-2
2. typing error

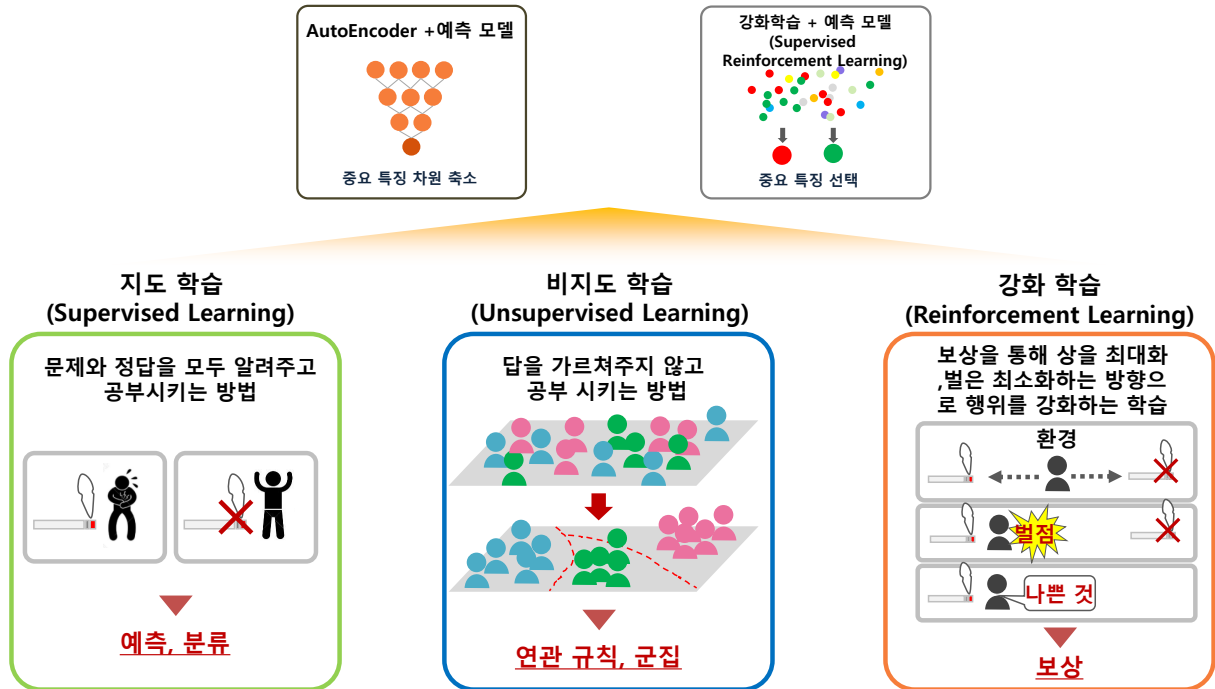
Table S1: Correcting drug names from TCGA to standard names

Recorded Name from TCGA	DrugBank ID	Standard drug name
IL2	DB00041	Aldesleukin
IL-2	DB00041	Aldesleukin
Interleukin - 2	DB00041	Aldesleukin
Interleukin-2	DB00041	Aldesleukin
Alvesin	NA	NA
Anastrozole	DB01217	Anastrozole
ANASTROZOLE	DB01217	Anastrozole
Arimidex	DB01217	Anastrozole
ARIMIDEX	DB01217	Anastrozole
PF-04605412	NA	anti-ASB1 integrin monoclonal antibody PF-04605412
MORAb-004	NA	anti-endothelialin/TEM1 monoclonal antibody MORAb-004?
autologous vaccine	NA	autologous vaccine
Axitinib	DB06626	Axitinib
Axitinib	DB06626	Axitinib
Cediranib	DB04849	AZD2171
Bacillus Calmette-Guerin (BCG)	NA	BCG
BCG	NA	BCG
avastin	DB00112	Bevacizumab
Avastin	DB00112	Bevacizumab
bevacizumab	DB00112	Bevacizumab
Bevacizumab	DB00112	Bevacizumab
BEVACIZUMAB	DB00112	Bevacizumab
Bicalutamide	DB01128	Bicalutamide
casodex	DB01128	Bicalutamide
Casodex	DB01128	Bicalutamide
bleomycin	DB00290	Bleomycin
Bleomycin	DB00290	Bleomycin
BRAF inhibitor	NA	BRAF inhibitor
cabazitaxel	DB06772	Cabazitaxel
jevtana	DB06772	Cabazitaxel
Cabozantinib	DB08875	Cabozantinib
Cancer Vax	NA	Cancer Vax
capecitabine	DB01101	Capecitabine
Capecitabine	DB01101	Capecitabine

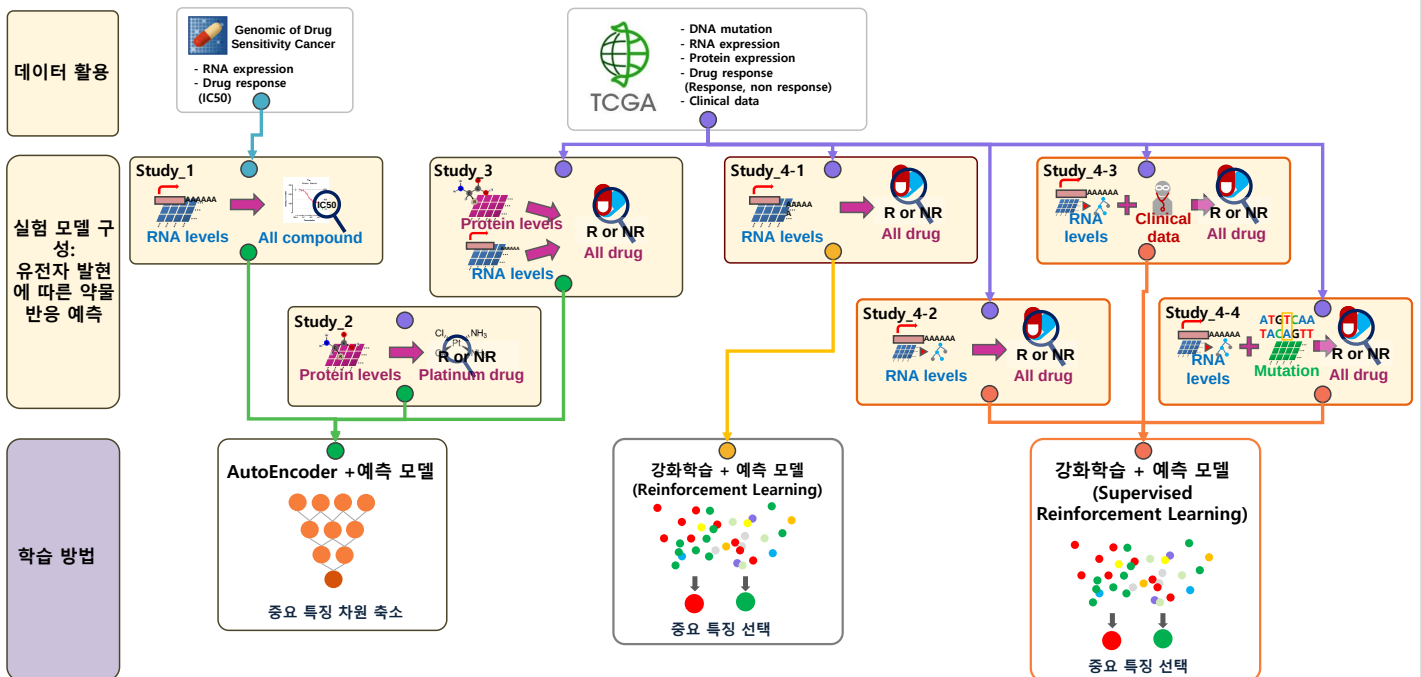
Bioinformatics, 32(19), 2891-2895.



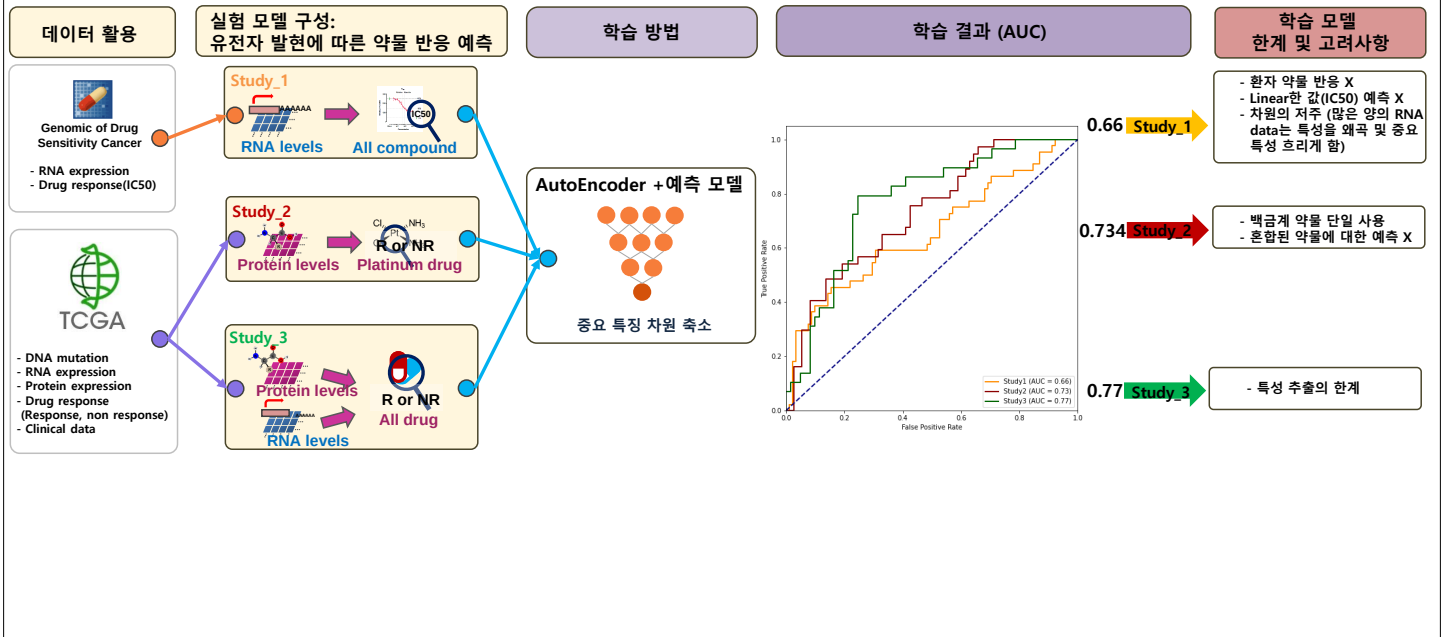
Preliminary study - methodology



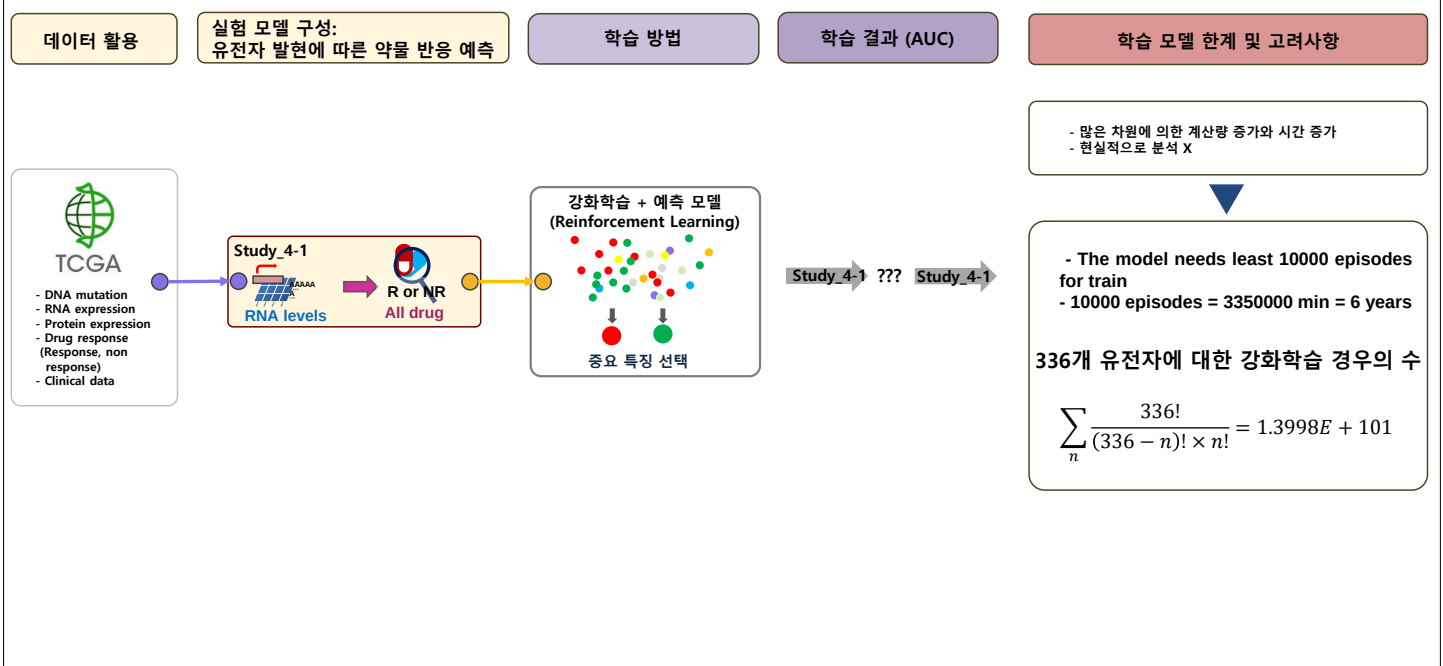
Preliminary study



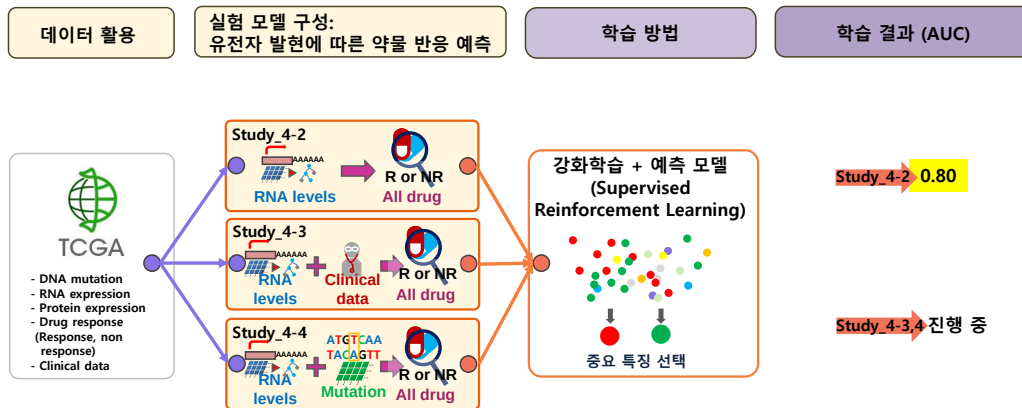
Results



Results



Results



Results

학습 결과

학습 모델 한계 및 고려사항

Study 4-2 결과

Cancer type	Sample size (R:NR)	Total gene (20531 genes)		KEGG cancer pathway (525 genes)		COSMIC cancer gene census (683 genes)	
		AUROC	Loss	AUROC	Loss	AUROC	Loss
PANCAN	1794 (42:58)	0.791	0.795	0.782	0.875	0.720	0.715
BRCA	278 (15:85)	0.902	0.579	0.913	0.722	0.910	0.624
LGG	176 (80:20)	0.939	0.480	0.908	0.591	0.840	0.930
STAD	150 (35:65)	0.586	2.651	0.620	1.598	0.584	0.941
HNSC	59 (25:75)	0.989	0.10	0.958	0.225	0.989	0.139
KIRC	18 (89:11)	0.750	21965.61	0.750	8.888	0.750	12.087
KIRP	11 (73:27)	0.333	141.36	1.000	0.141	1.000	0.041
KICH	8 (87.5:12.5)	0.500	32591.15	0.500	13.325	0.500	17.119
LUAD	112 (47:53)	0.752	1.143	0.811	0.921	0.750	1.016
PRAD	43 (33:67)	0.593	1.230	0.704	1.371	0.580	1.827

- 특성 추출을 위한
여러 조건 연구

Results

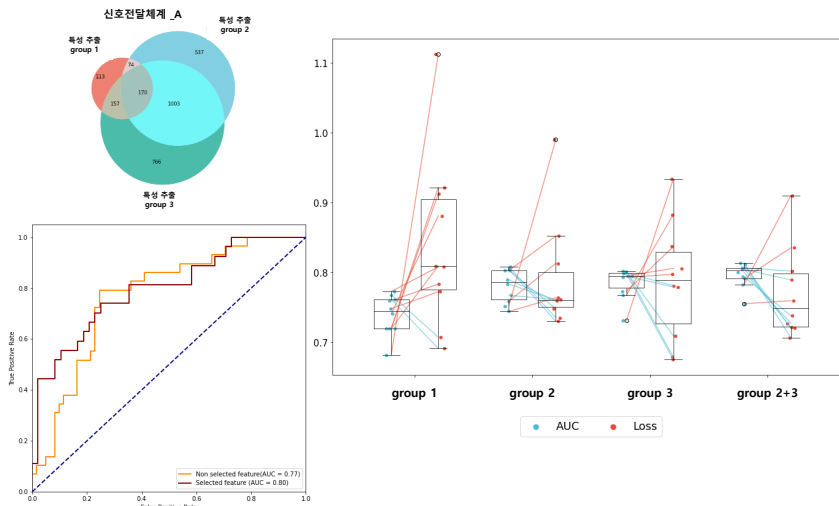
학습 결과

개선점

학습 모델 한계 및 고려사항

Study 4-2 결과

- 암세포의 신호전달체계 중 항암제 관련 신호전달체계 선택 후, 특성 추출
- 추출된 특성으로 예측모델 수행



- 항암제 관련 신호전달체계에서 특성 추출 후 예측모델 수행 시 학습 능력 향상되었고 (AUC=0.80) 예측의 불확실 정도가 개선됨.

- 특성 추출 고려
- 예를 들어, cell signaling, cellular processes 등

Results

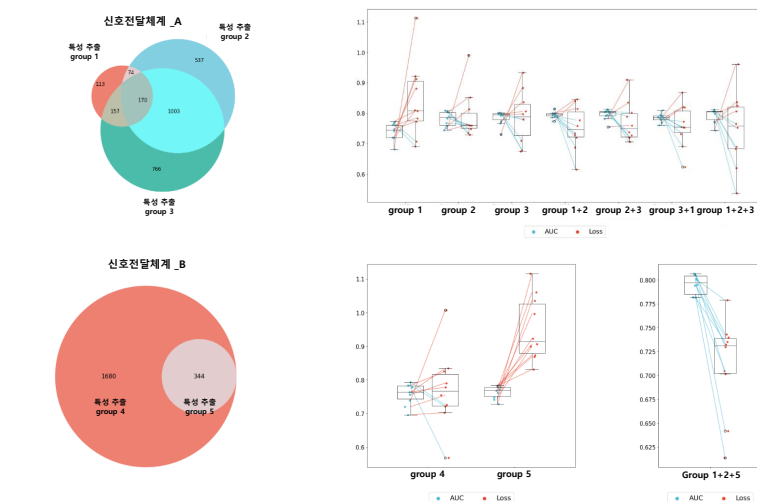
학습 결과

개선점

학습 모델 한계 및 고려사항

Study 4-2 결과

- 암세포의 신호전달체계 중 항암제 관련 신호전달체계 선택 후, 특성 추출
- 추출된 특성으로 예측모델 수행



- 항암제 관련 신호전달체계에서 특성 추출 후 예측모델 수행 시 학습 능력 향상되었고 (AUC=0.80) 예측의 불확실 정도가 개선됨.

- 특성 추출 고려
- 예를 들어, cell signaling, cellular processes 등

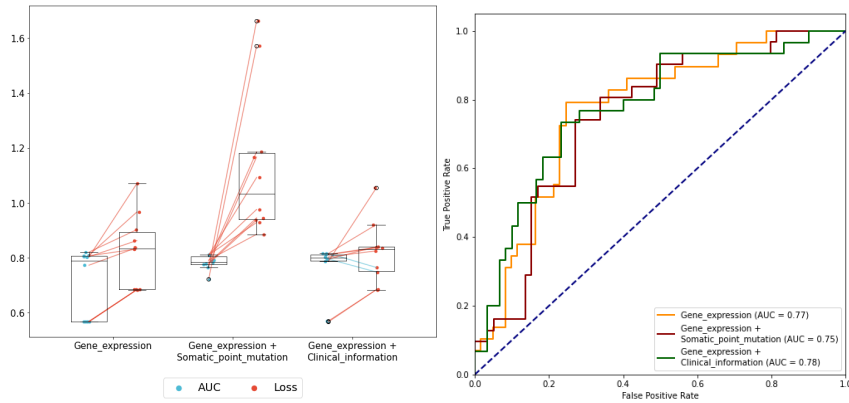
Results

학습 결과

개선점

학습 모델 한계 및 고려사항

Study 4-3,4



- 특성 중 항암제 관련성이 반영되지 않는 임상정보 및 돌연변이를 사용했을 경우 학습 능력 (+돌연변이: 0.78, +임상정보: AUC=0.75)이 떨어지고 예측의 불확실 정도가 나쁨.
- Layer 마다 특성을 고려해야 함.

- Layer 병합을 이용한 실험 모델에서는 각각 특성 추출 고려
- Layer와 layer 사이 관계 이해

Acknowledgements

Catholic Univ. of Korea

Dr. Jongkeun Park Dr. Jeessoo Chae
Mr. Dawoon Jung

National Cancer Center

Dr. Yu Seog Jung Dr. Sung Min Park
Dr. Jongkeun Lee Dr. Yul Hwangbo
Mr. Jinseok Lee Dr. Gwisun Choi
Mr. Hyunbin Kim Dr. Sun Young Gong
Mr. Soohwan Eho Dr. Jong Bae Park
Ms. Jee Hyun Kim Dr. Sangjae Park
Dr. Kyunghee Kim Dr. Eunsook Lee

Korea Basic Science Institute

Dr. Jin Young Kim Dr. Hee Yeon Hwang

Funding Sources:

- National Cancer Center, Korea
- Korea Health Industry Development Institute (KHIDI)
- Ministry of Health & Welfare
- National Research Foundation of Korea (NRF)

KISTI

Dr. Min Seon Yeom
Dr. Jun hak Lee

SIG

Dr. Woong Yang Park
Dr. Jin Ho Kim

Yonsei Univ.

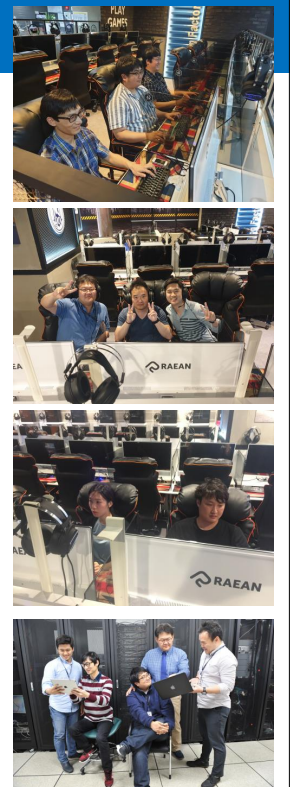
Dr. Jae Ho Jeong
Dr. Sun Young Rha
Dr. Nam Hoon Cho

KAIST

Dr. Young Seok Ju

Univ. of Seoul

Dr. Dong Sung Lee



경청해 주셔서 고맙습니다.

Big Data Analysis Lab.

교수: 홍동완 (dwhong@catholic.ac.kr, 010-5363-5404)

연구실: 별관 교204호, 실험실: 별관 10층 암연구소

연구 주제

- Cancer Multi-Omics (WGS, WES, Panel, RNA-Seq, Proteome, Hi-C 등) 분석
- 분석 소프트웨어 및 통합 데이터베이스 개발
- 인공지능 기반 약물반응성, 적정 치료제, 치료법 개발
- 빅데이터 분석 알고리즘 개발

