Intro to R for Data Analysis and Visualization
Summer 2022

Connecting Patient Datasets to Preclinical Hypotheses in Cancer

Riyue Sunny Bao PhD
Associate Professor of Medicine
Co-Director of Cancer Bioinformatics (CBS)
Co-Director of Translational Immuno-Informatics Lab (TIIL)

July 8th, 2022

UPMC | HILLMAN CANCER CENTER

University of Pittsburgh

Outline

Schedule

- 8:30 AM – 9:30 AM : Lecture
- 9:30 AM – 10:00 AM : Hands-on practice

Key take home messages

- How to associate gene expression data with clinical outcome
- Hands on: Use (public) gene expression data to discover tumor subtypes and survival analysis

Class materials

Download class materials from GitHub

- MSTP_summer2022.lecture.pdf
- MSTP_summer2022.handson.Rmd

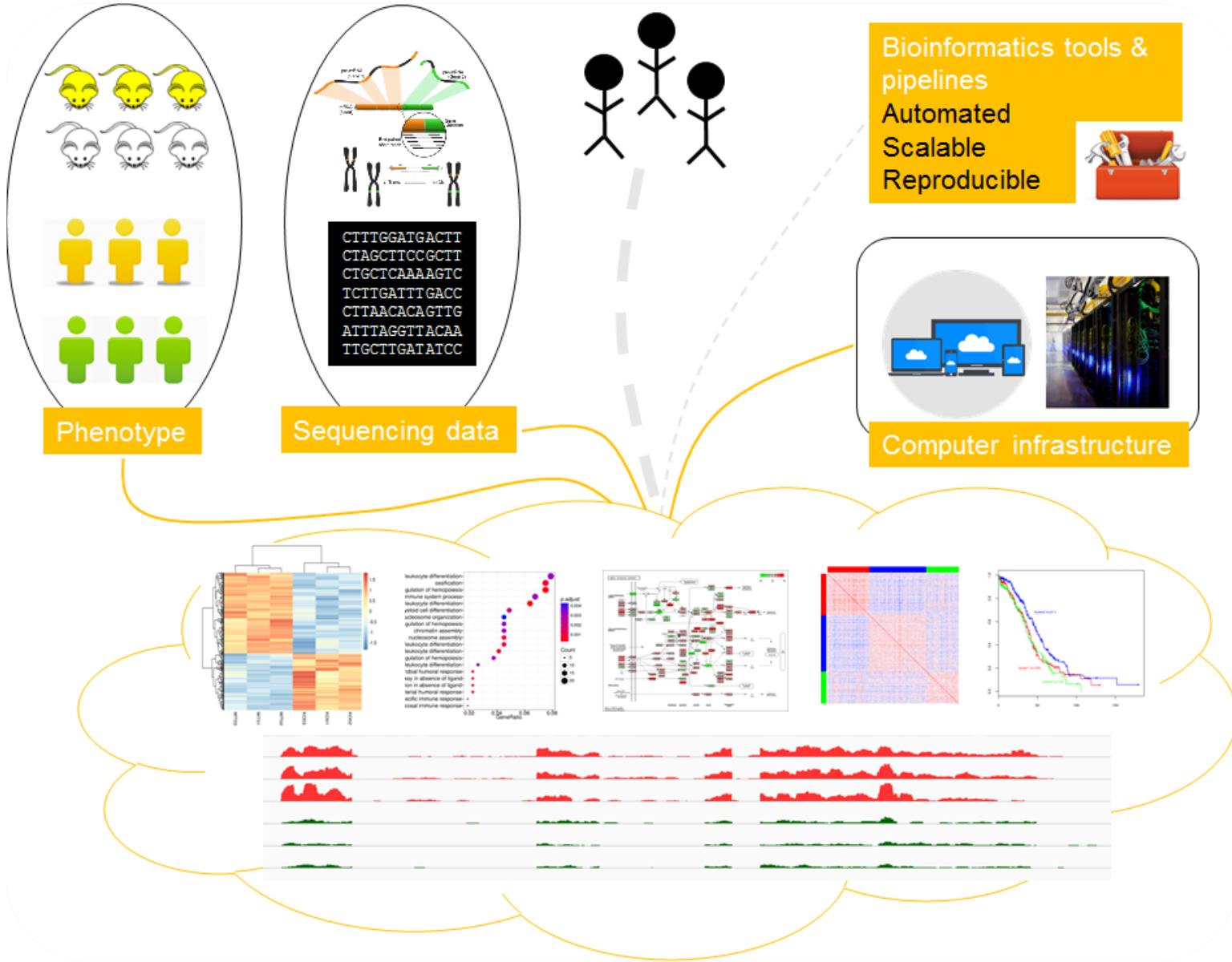
https://github.com/riyuebao/MSTP_summer2022_class07

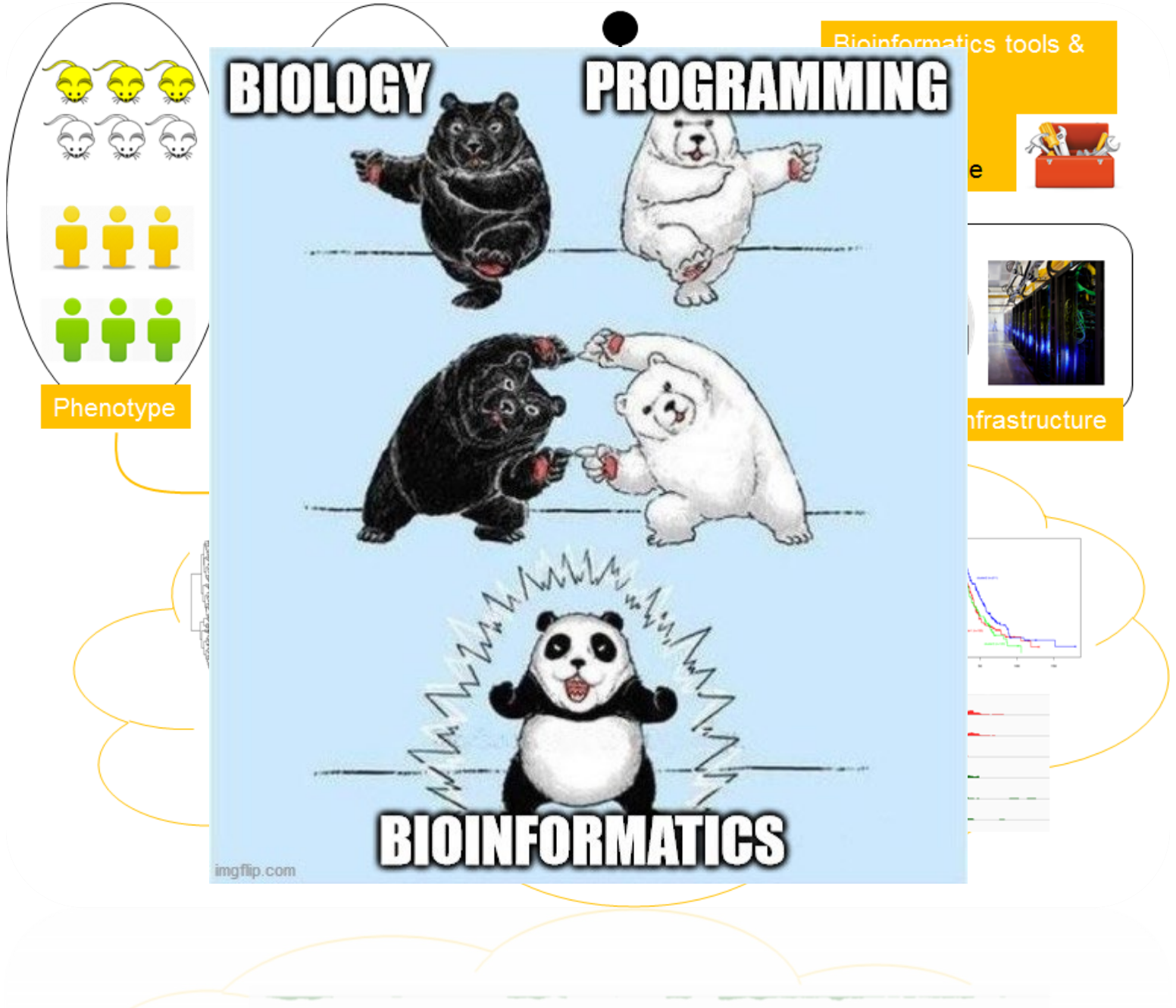
What you need

- Rstudio (or R console) on personal computers (hands on practice)

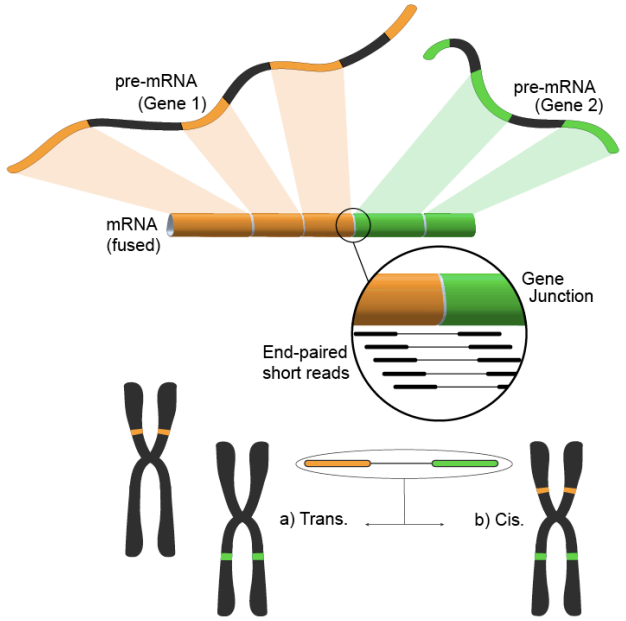
Objective

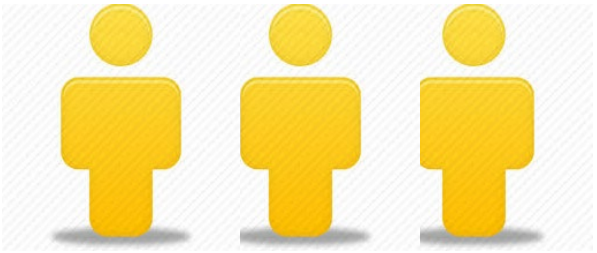
- Learn the background and application of The Cancer Genome Atlas (TCGA)
- Learn the structure and access of Genomics Data Commons (GDC)
- Explore datasets hosted on GDC
- Practice how to connect gene expression with clinical data
  - Use gene expression to identify tumor subtype
  - Detect survival difference between subtypes
  - Produce high-quality plots for publication


Phenotype
CTTTGGATGACTT
CTAGCTTCCGCTT
CTGCTCAAAAGTC
TCTTGATTTGACC
CTTAACACAGTTG
ATTTAGGTTACAA
TTGCTTGATATCC
Sequencing data
Bioinformatics tools & pipelines
Automated
Scalable
Reproducible
Computer infrastructure


Bioinformatics tools &
e
Phenotype
nfrastructure
BIOLOGY
PROGRAMMING
BIOINFORMATICS
imgflip.com

Background

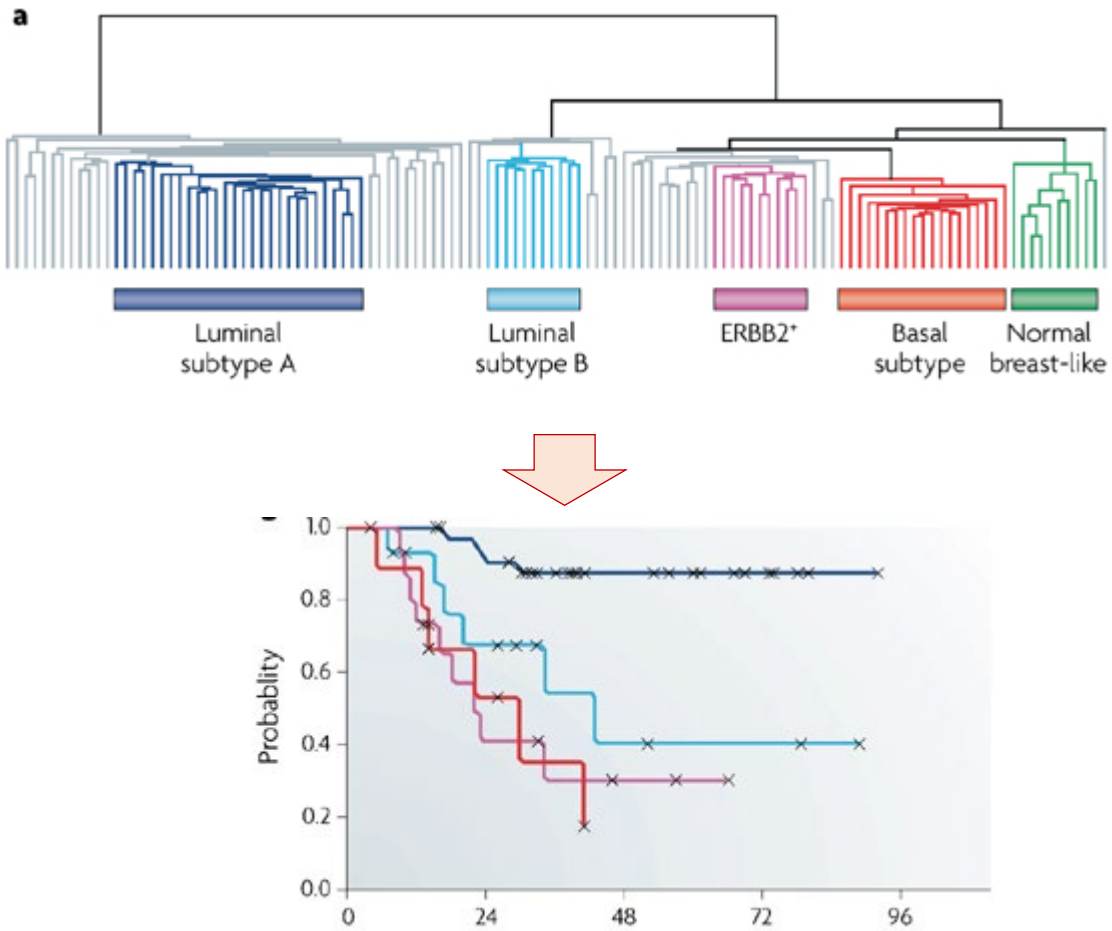




Gene Expression

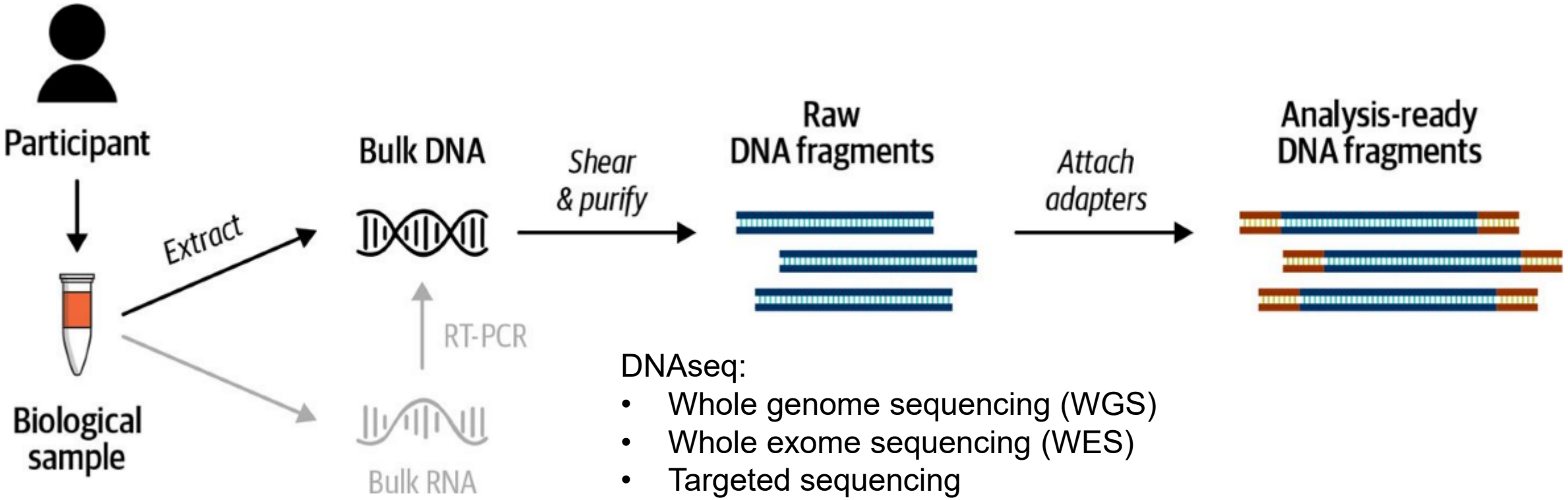
Patient’s clinical data

Background



- Correlate gene expression with clinical data (tumor stage, tumor grade, time to death, time to relapse, etc.)
- Identify tumor subtypes through sample clustering
- Detect survival difference between tumor subtypes
- Discover gene signatures to predict patient classes

NGS experiments



DNAseq:

- Whole genome sequencing (WGS)
- Whole exome sequencing (WES)
- Targeted sequencing

RNAseq:

- Whole transcriptome sequencing (WTS)
- mRNA sequencing
- Small RNA sequencing (miRNA, snoRNA, etc.)

Main types of public cancer omics databases

Sequencing based dbs

Protein based dbs

Imaging based dbs

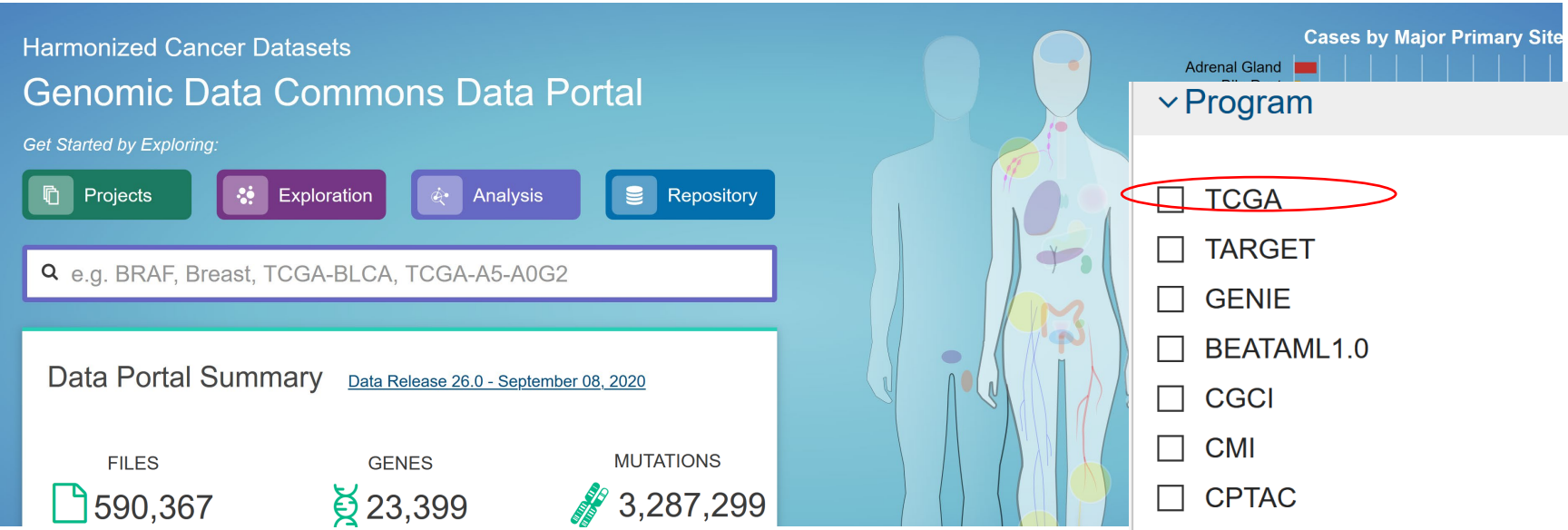
Bulk tissue databases

Single cell databases

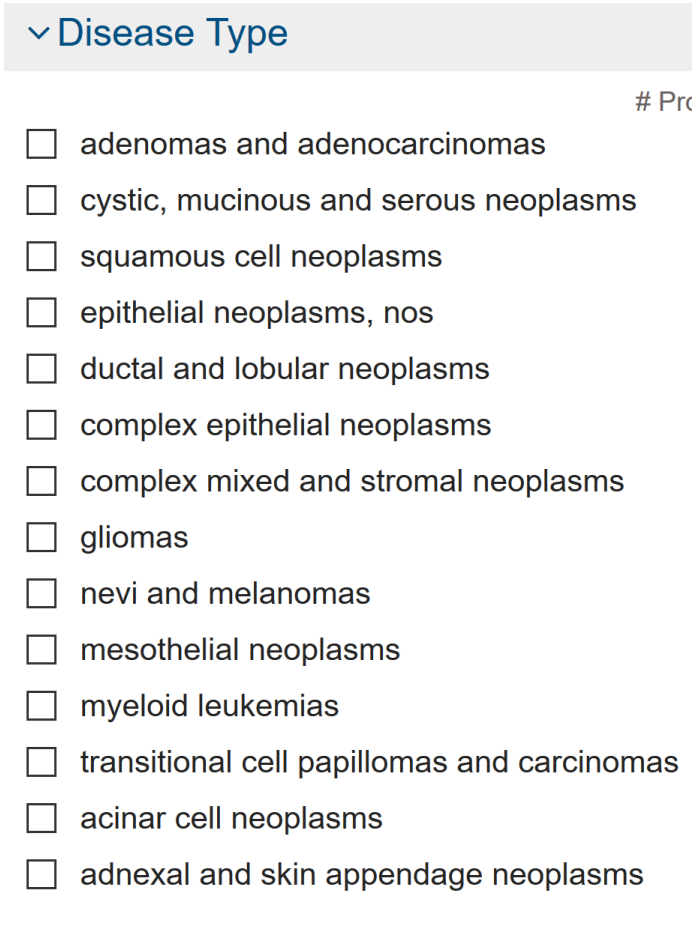
- Mine data and form new hypothesis
- Validate your existing discoveries using a much larger cohort
- Identify novel targets for cancer therapy and bring it to clinic
- Identify new patient subtypes and develop new biomarkers
- Develop innovative computational tools
- Meta-analysis across many cohorts
- Perform machine learning and deep learning using cohorts of sufficient sample sizes

Genomic Data Commons (GDC)

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



54 disease types



3.53 PB cancer genomics data!!
Open to the scientific community!!
Data download is very stable and fast!!

With the latest Gen3 technology

https://stats.gen3.org/

Genomic Data Commons (GDC)

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home | Projects | Exploration | Analysis | Repository | Manage Sets | 18

- CMI 3
- BEATAML1.0 2

13 More...

Disease Type

Projects

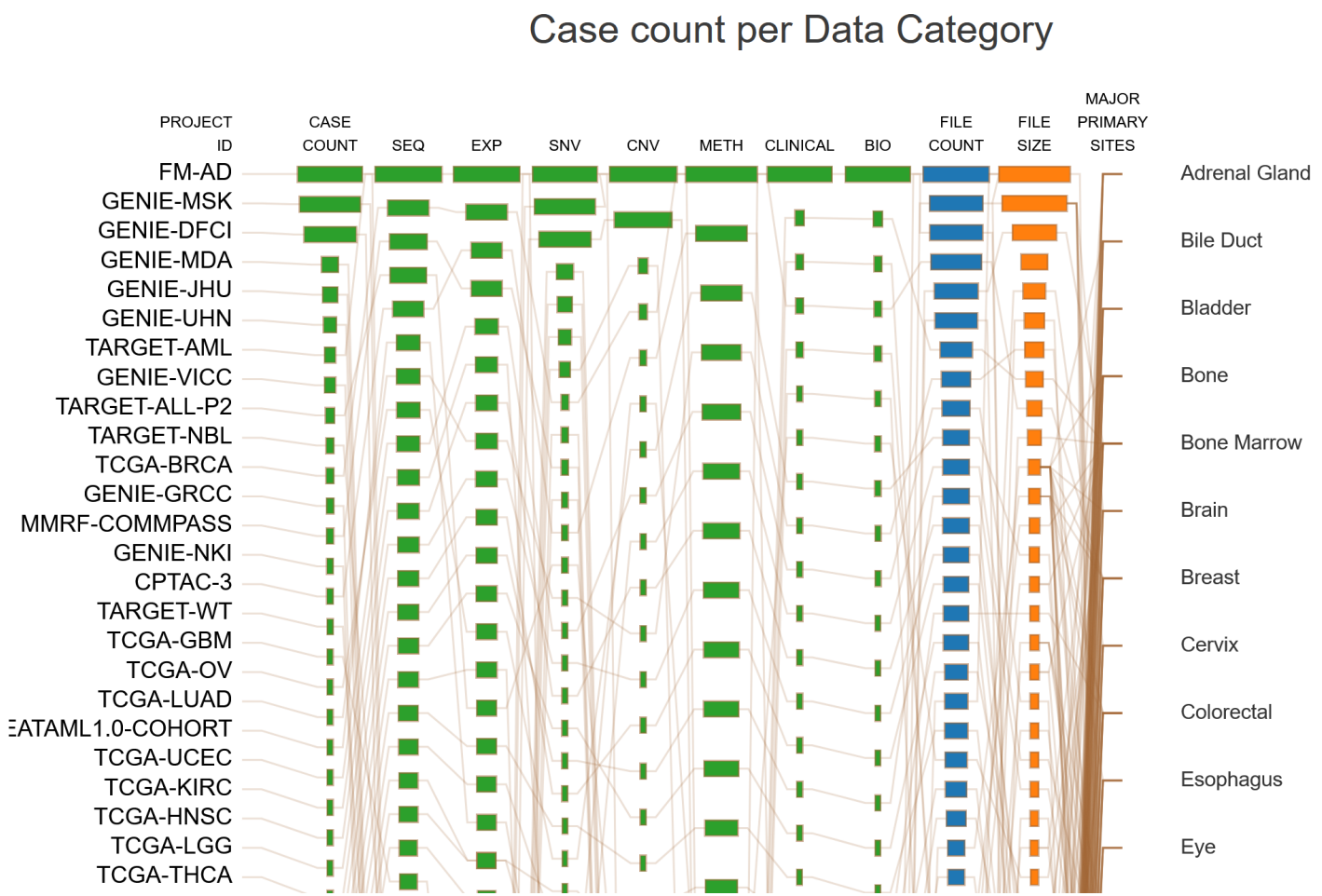
- adenomas and adenocarcinomas 35
- cystic, mucinous and serous neoplasms 22
- squamous cell neoplasms 19
- epithelial neoplasms, nos 18
- ductal and lobular neoplasms 16

54 More...

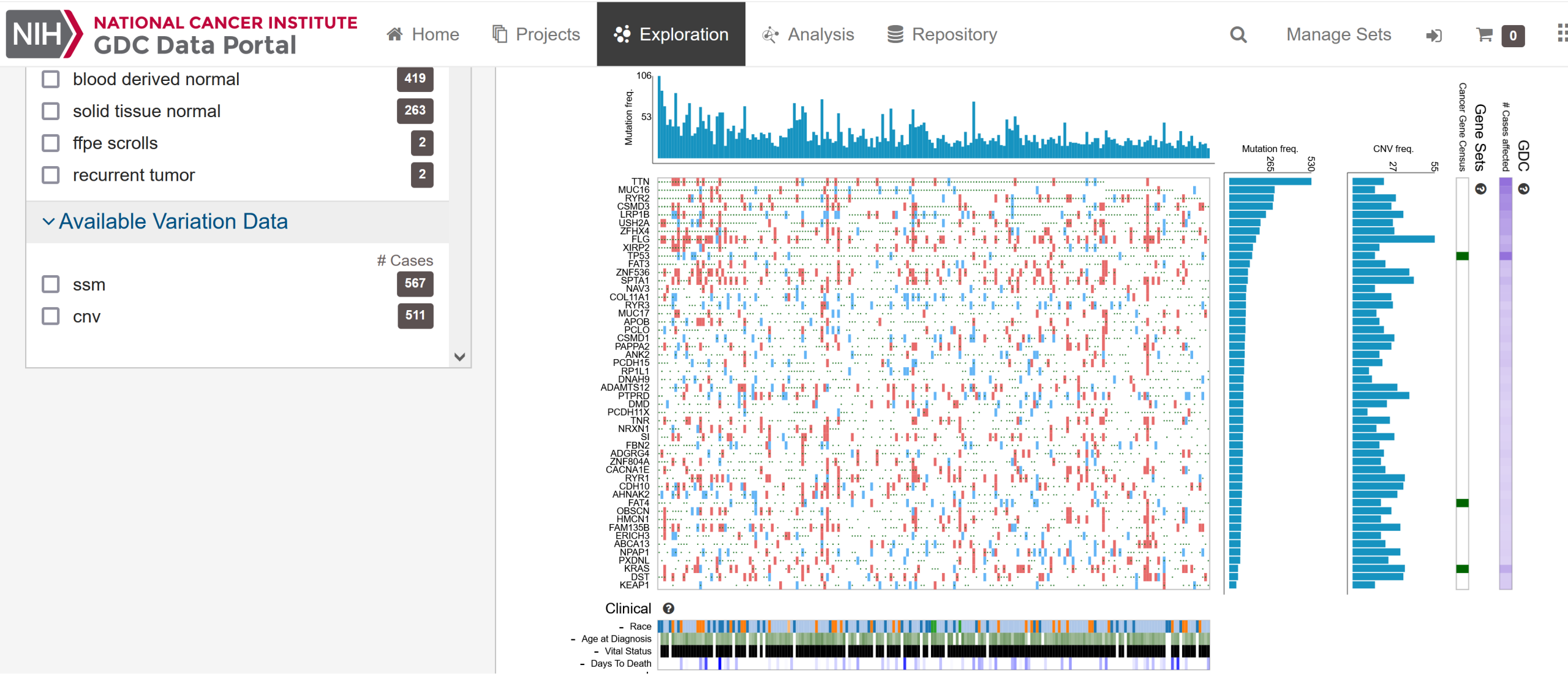
Data Category

Projects

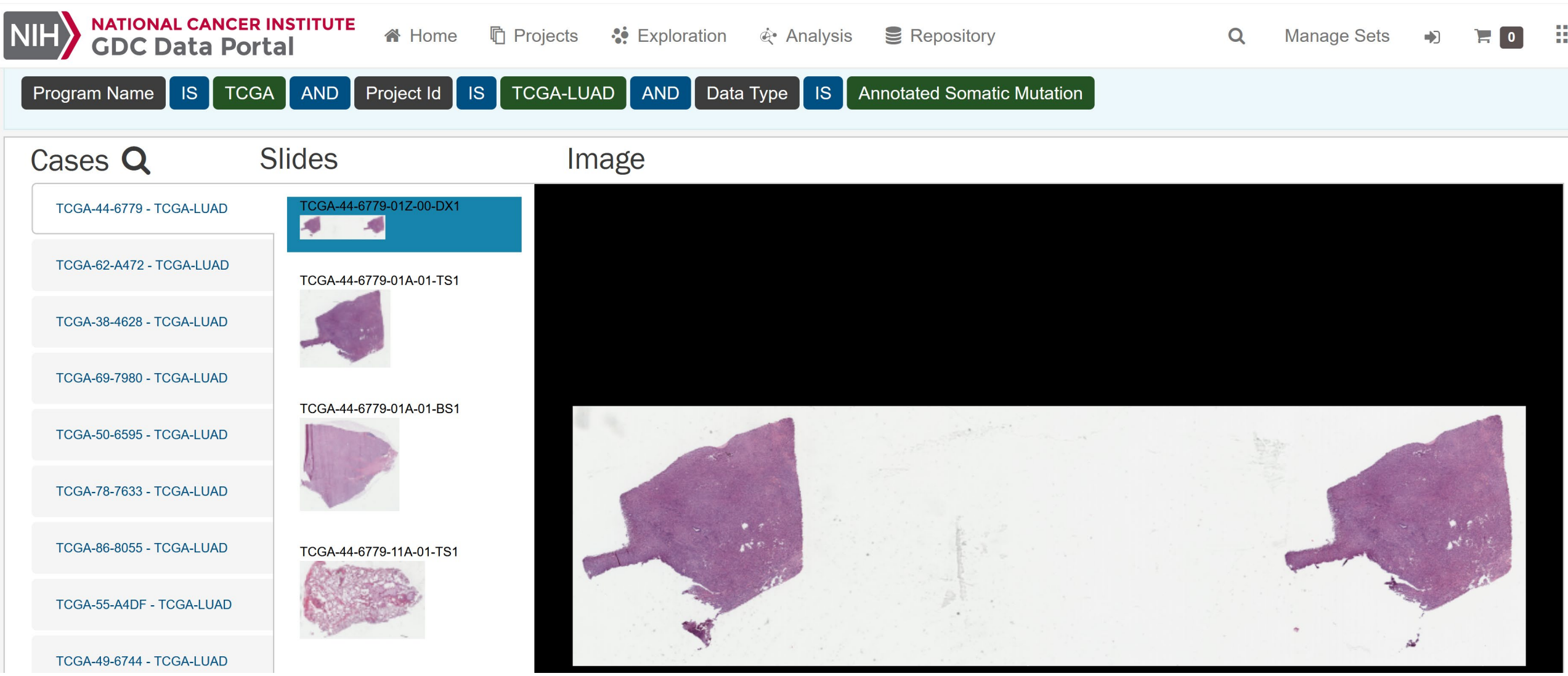
- simple nucleotide variation 62
- sequencing reads 61
- transcriptome profiling 57
- clinical 47
- biospecimen 46



Genomic Data Commons (GDC): TCGA-LUAD



Genomic Data Commons (GDC) : TCGA-LUAD



Genomic Data Commons (GDC): TCGA-LUAD

ARTICLE

OPEN

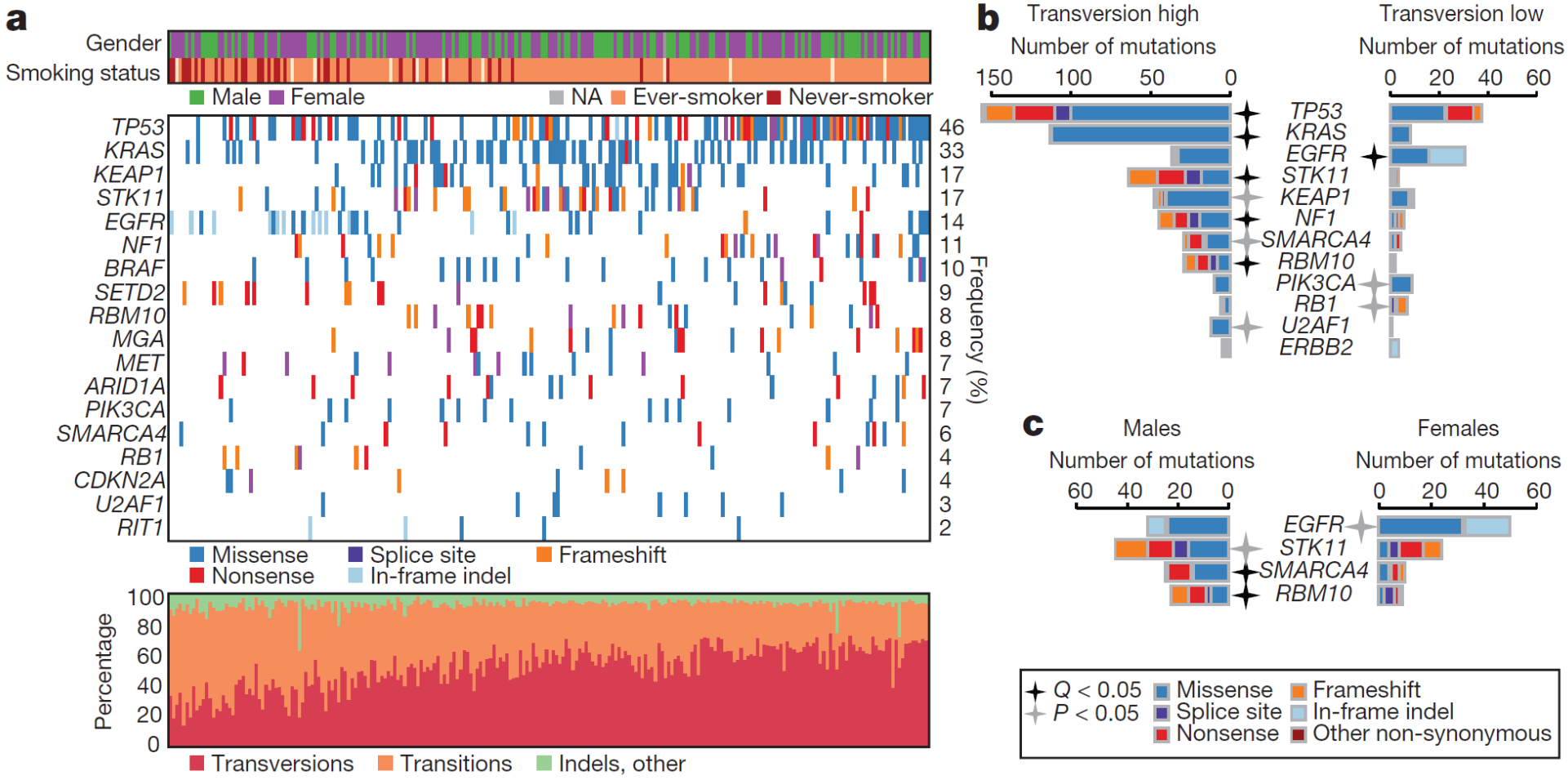
doi:10.1038/nature13385

Comprehensive molecular profiling of lung adenocarcinoma

The Cancer Genome Atlas Research Network*

Adenocarcinoma of the lung is the leading cause of cancer death worldwide. Here we report molecular profiling of 230 resected lung adenocarcinomas using messenger RNA, microRNA and DNA sequencing integrated with copy number, methylation and proteomic analyses. High rates of somatic mutation were seen (mean 8.9 mutations per megabase). Eighteen genes were statistically significantly mutated, including *RIT1* activating mutations and newly described loss-of-function *MGA* mutations which are mutually exclusive with focal *MYC* amplification. *EGFR* mutations were more frequent in female patients, whereas mutations in *RBM10* were more common in males. Aberrations in *NF1*, *MET*, *ERBB2* and *RIT1* occurred in 13% of cases and were enriched in samples otherwise lacking an activated oncogene, suggesting a driver role for these events in certain tumours. DNA and mRNA sequence from the same tumour highlighted splicing alterations driven by somatic genomic changes, including exon 14 skipping in *MET* mRNA in 4% of cases. MAPK and PI(3)K pathway activity, when measured at the protein level, was explained by known mutations in only a fraction of cases, suggesting additional, unexplained mechanisms of pathway activation. These data establish a foundation for classification and further investigations of lung adenocarcinoma molecular pathogenesis.

Genomic Data Commons (GDC): TCGA-LUAD



Genomic Data Commons (GDC): TCGA-OV

ARTICLE

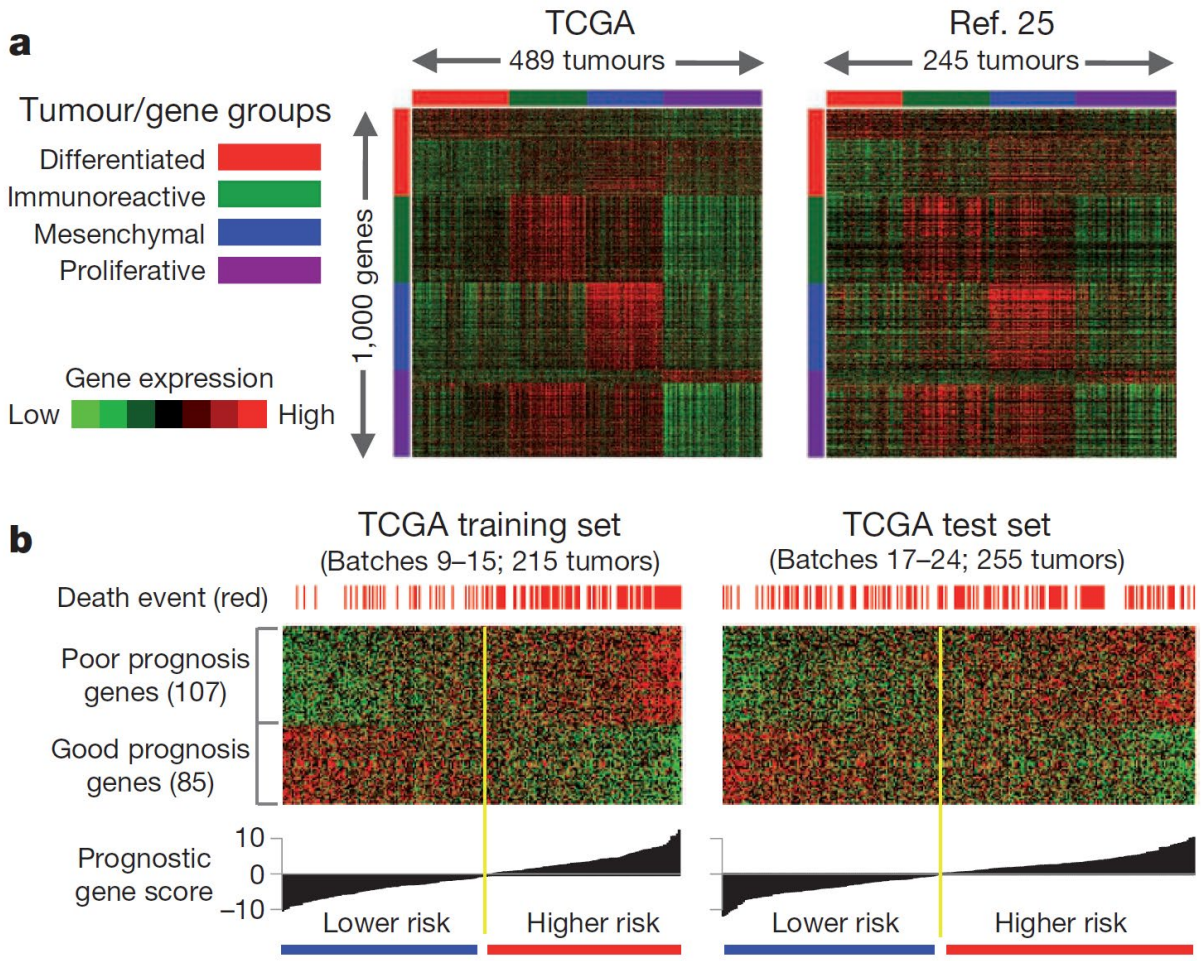
doi:10.1038/nature10166

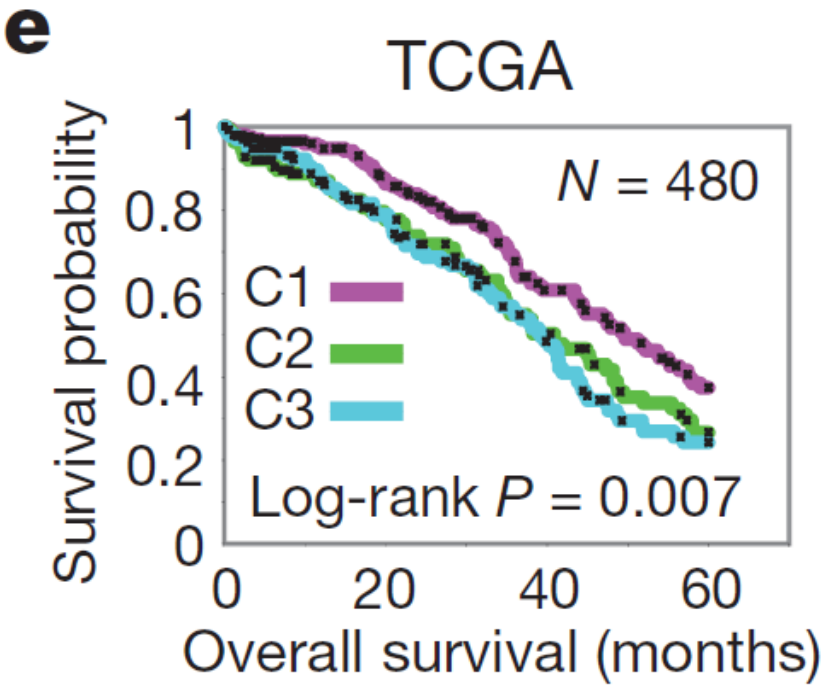
Integrated genomic analyses of ovarian carcinoma

The Cancer Genome Atlas Research Network*

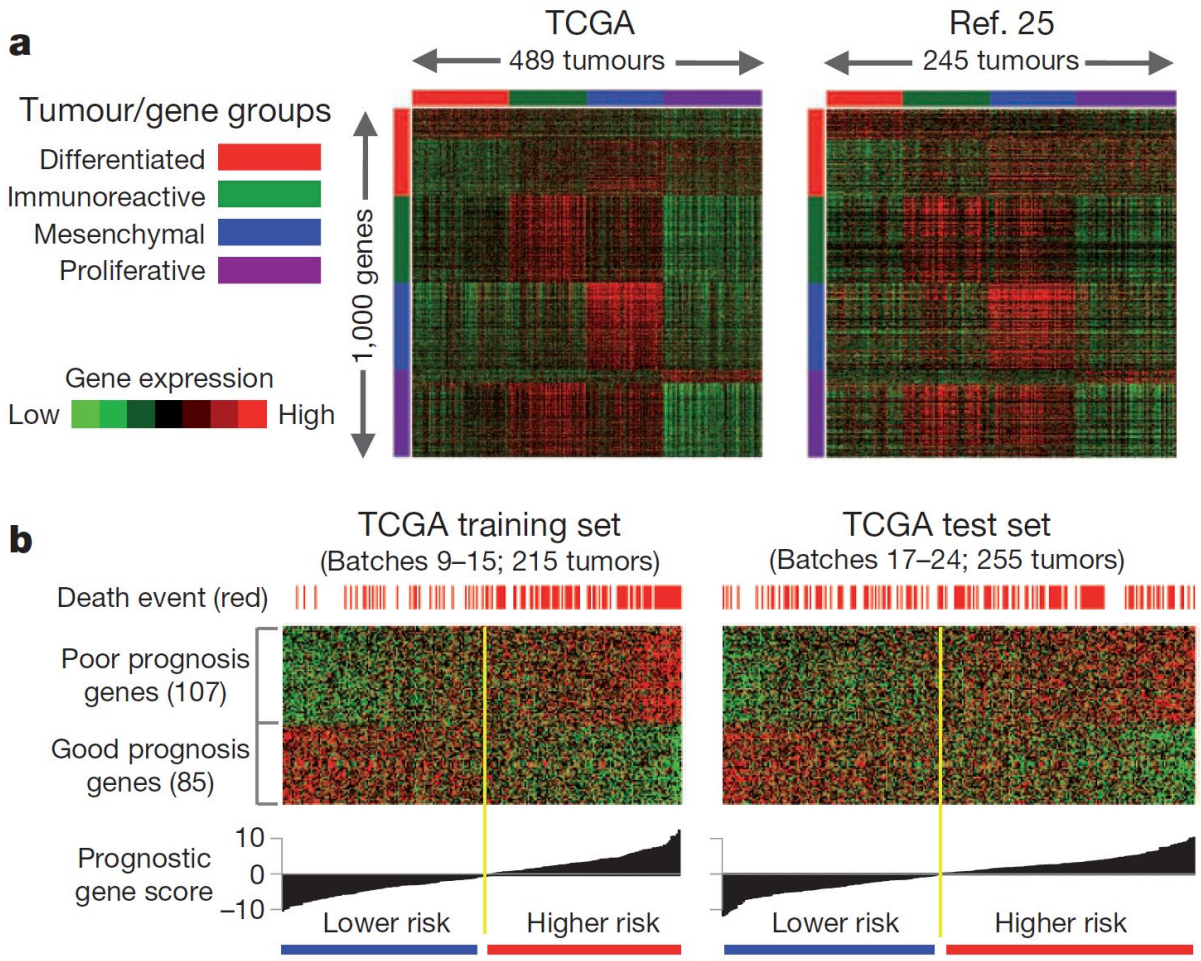
A catalogue of molecular aberrations that cause ovarian cancer is critical for developing and deploying therapies that will improve patients' lives. The Cancer Genome Atlas project has analysed messenger RNA expression, microRNA expression, promoter methylation and DNA copy number in 489 high-grade serous ovarian adenocarcinomas and the DNA sequences of exons from coding genes in 316 of these tumours. Here we report that high-grade serous ovarian cancer is characterized by *TP53* mutations in almost all tumours (96%); low prevalence but statistically recurrent somatic mutations in nine further genes including *NF1*, *BRCA1*, *BRCA2*, *RB1* and *CDK12*; 113 significant focal DNA copy number aberrations; and promoter methylation events involving 168 genes. Analyses delineated four ovarian cancer transcriptional subtypes, three microRNA subtypes, four promoter methylation subtypes and a transcriptional signature associated with survival duration, and shed new light on the impact that tumours with *BRCA1/2* (*BRCA1* or *BRCA2*) and *CCNE1* aberrations have on survival. Pathway analyses suggested that homologous recombination is defective in about half of the tumours analysed, and that NOTCH and FOXM1 signalling are involved in serous ovarian cancer pathophysiology.

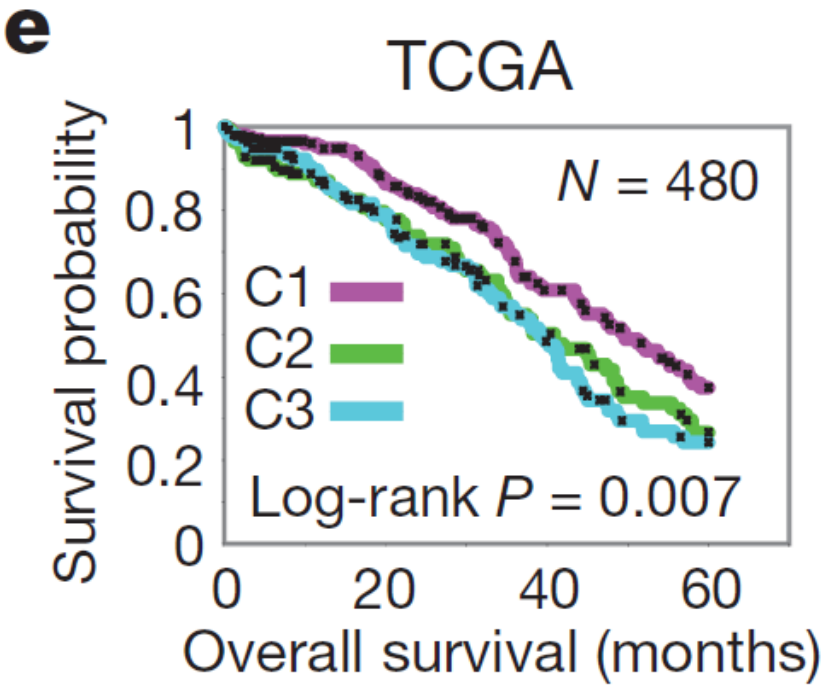
Genomic Data Commons (GDC): TCGA-OV





Genomic Data Commons (GDC): TCGA-OV





Today's hands on practice!

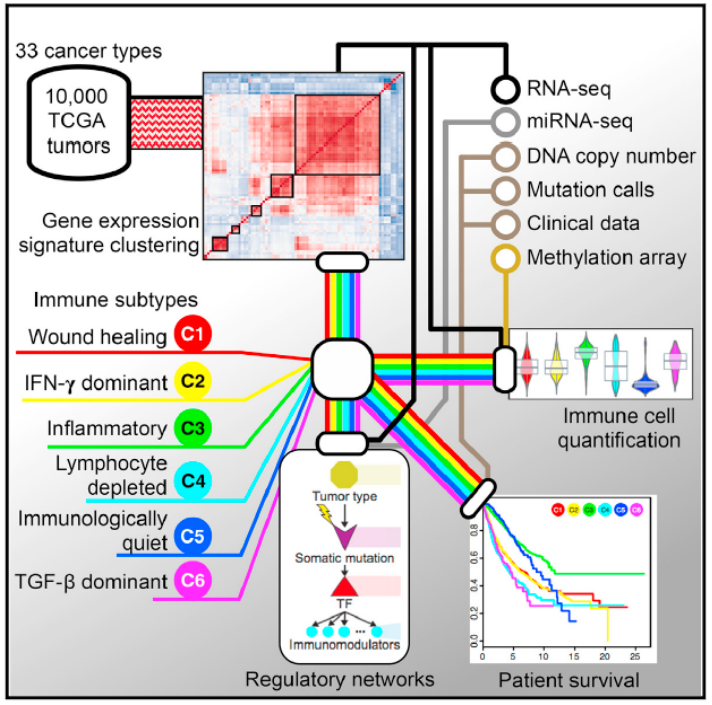
Genomic Data Commons (GDC): TCGA-pan-cancer

Resource

Immunity

The Immune Landscape of Cancer

Graphical Abstract



Authors

Vésteinn Thorsson, David L. Gibbs, Scott D. Brown, ..., Mary L. Disis, Benjamin G. Vincent, Ilya Shmulevich

Correspondence

vesteinn.thorsson@systemsbiology.org (V.T.), benjamin.vincent@unchealth.unc.edu (B.G.V.), ilya.shmulevich@systemsbiology.org (I.S.)

In Brief

Thorsson et al. present immunogenomics analyses of more than 10,000 tumors, identifying six immune subtypes that encompass multiple cancer types and are hypothesized to define immune response patterns impacting prognosis. This work provides a resource for understanding tumor-immune interactions, with implications for identifying ways to advance research on immunotherapy.

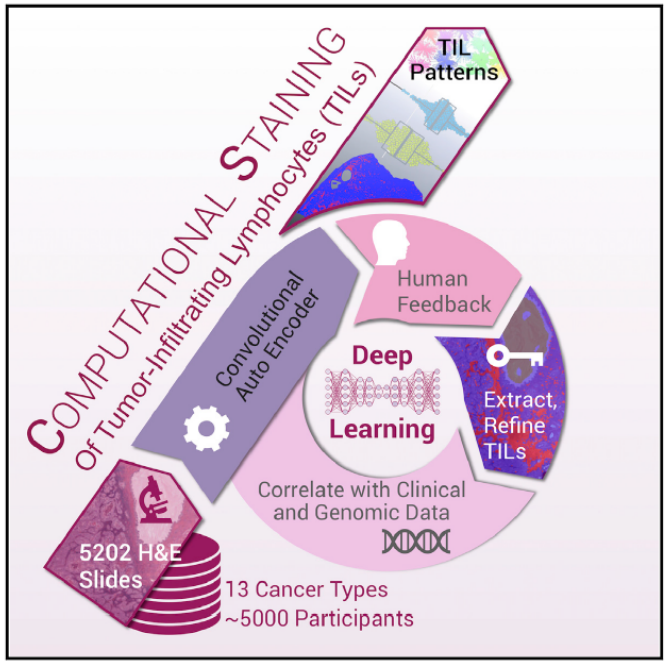
Highlights

Resource

Cell Reports

Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images

Graphical Abstract



Authors

Joel Saltz, Rajarsi Gupta, Le Hou, ..., Alexander J. Lazar, Ashish Sharma, Vésteinn Thorsson

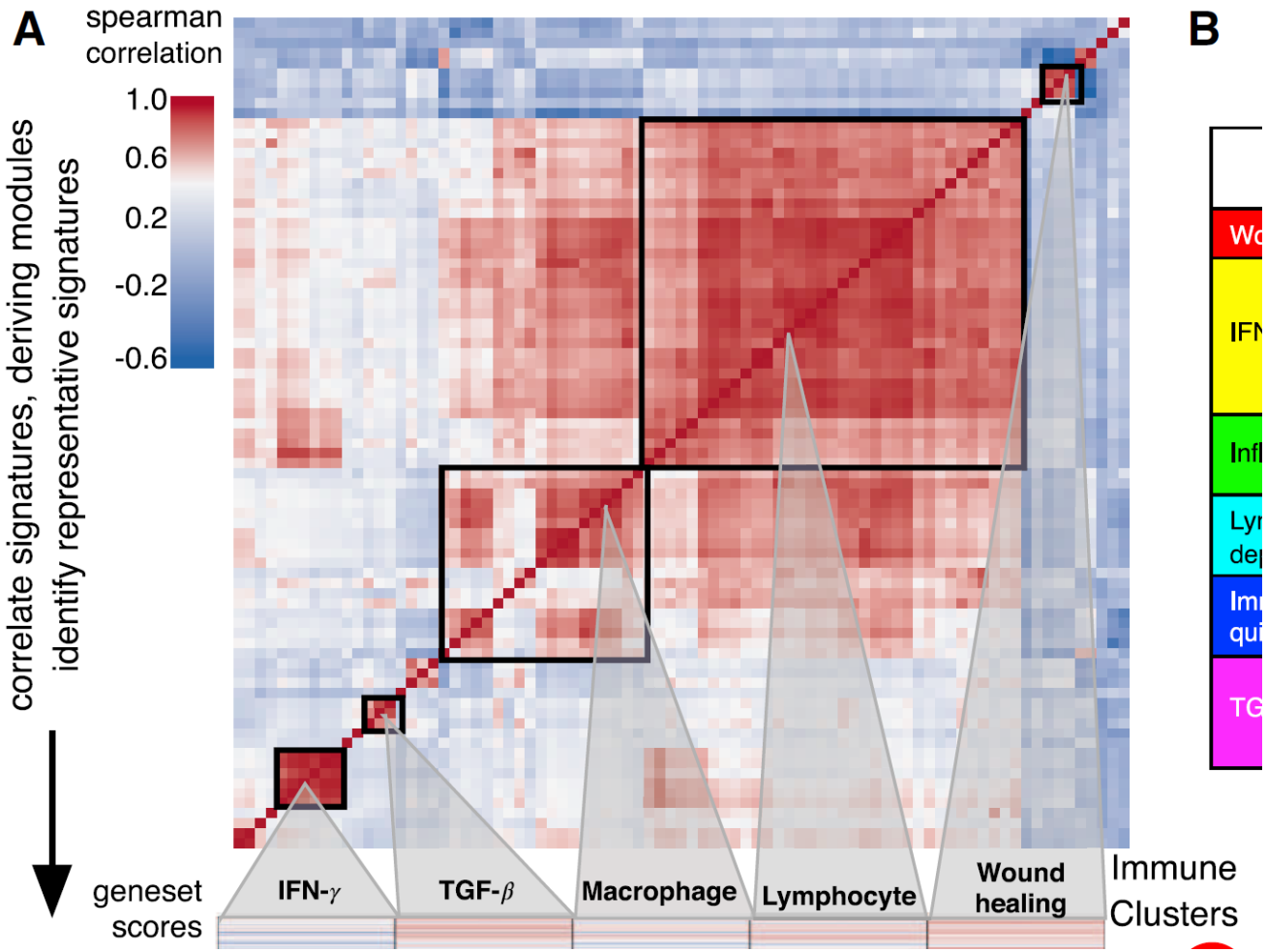
Correspondence

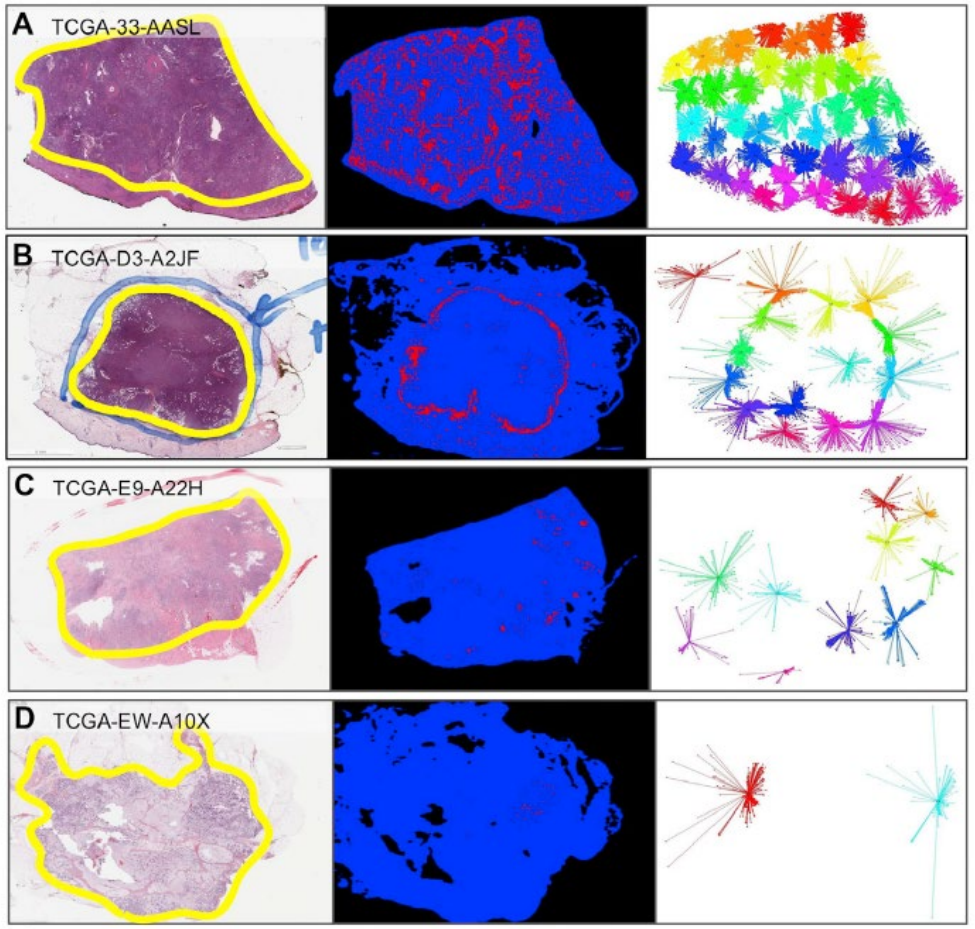
joel.saltz@stonybrookmedicine.edu (J.S.), vesteinn.thorsson@systemsbiology.org (V.T.)

In Brief

Tumor-infiltrating lymphocytes (TILs) were identified from standard pathology cancer images by a deep-learning-derived "computational stain" developed by Saltz et al. They processed 5,202 digital images from 13 cancer types. Resulting TIL maps were correlated with TCGA molecular data, relating TIL content to survival, tumor subtypes, and immune profiles.

Genomic Data Commons (GDC): TCGA-pan-cancer





E

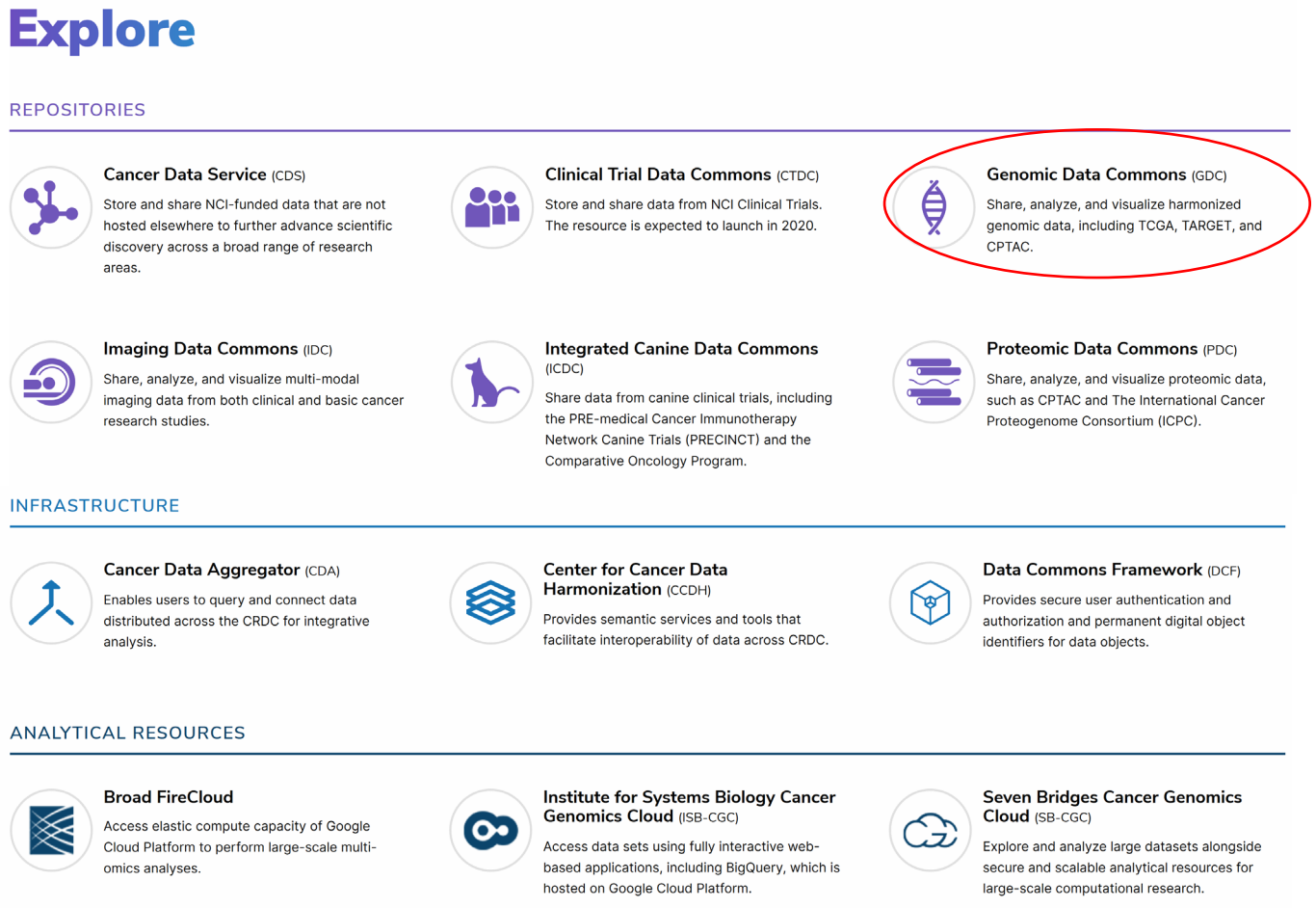
Participant Barcode	Study	Number of TIL Patches	TIL fraction	Number of TIL Clusters	Cluster Size Mean	Within-Cluster Dispersion Mean	Cluster Extent Mean	Ball Hall Index	Banfield Raftery Index	C Index	Determinant Ratio Index	Global Pattern
TCGA-33-AASL	LUSC	26245	20.6	40	656.1	293456	41.0	447	159518	0.015	2065.4	Brisk Diffuse
TCGA-D3-A2JF	SKCM	6832	4.9	18	379.6	238600	82.1	771	43456	0.022	790.0	Brisk Band-like
TCGA-E9-A22H	BRCA	1000	1.5	10	100.0	54876	51.9	560	6174	0.025	343.0	Non-Brisk Multifocal
TCGA-EW-A10X	BRCA	285	0.1	2	142.5	430332	223.0	3093	2283	0.000	29.6	Non-Brisk Focal

Genomic Data Commons (GDC)

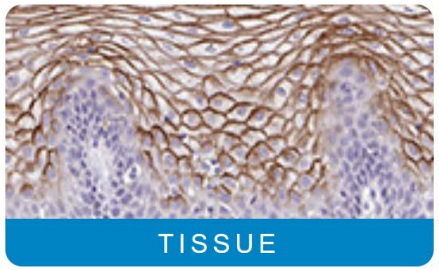
https://datacommons.cancer.gov/

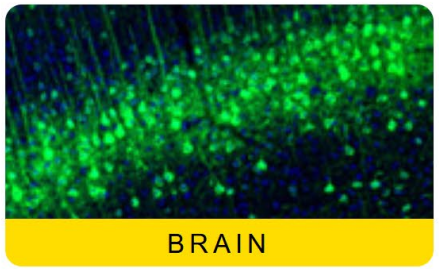
Connecting Data to Accelerate Cancer Research

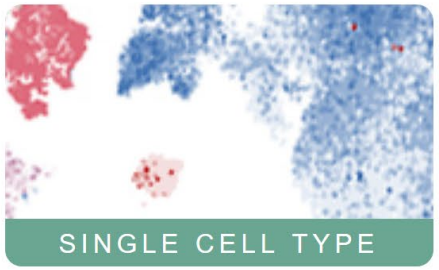
The NCI Cancer Research Data Commons (CRDC) is a cloud-based data science infrastructure that provides secure access to a large, comprehensive, and expanding collection of cancer research data. Users can explore and use analytical and visualization tools for data analysis in the cloud.

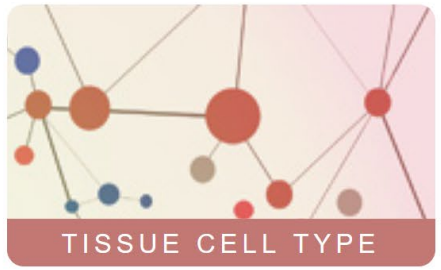


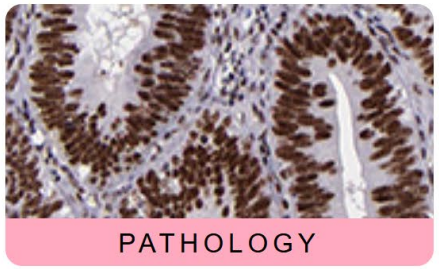
The Human Protein Atlas





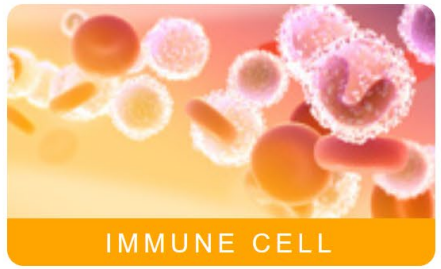


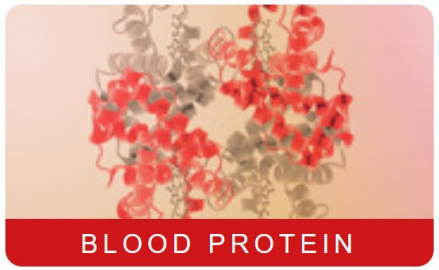


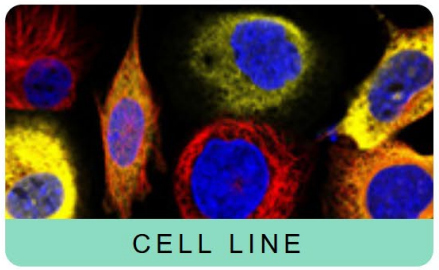


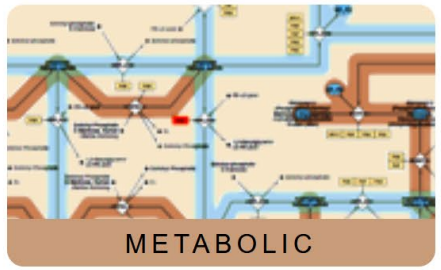
The open access resource for human proteins

Search for specific genes/proteins or explore the 10 different sections









The Human Protein Atlas

ACE2

SUMMARY | TISSUE | RNA BRAIN | RNA SINGLE CELL | RNA TISSUE CELL | PATHOLOGY | RNA IMMUNE | MS BLOOD | RNA SUBCELL | RNA CELL LINE | METABOLIC

PROTEIN SUMMARY

RNA DATA

GENE/PROTEIN

ANTIBODIES AND VALIDATION

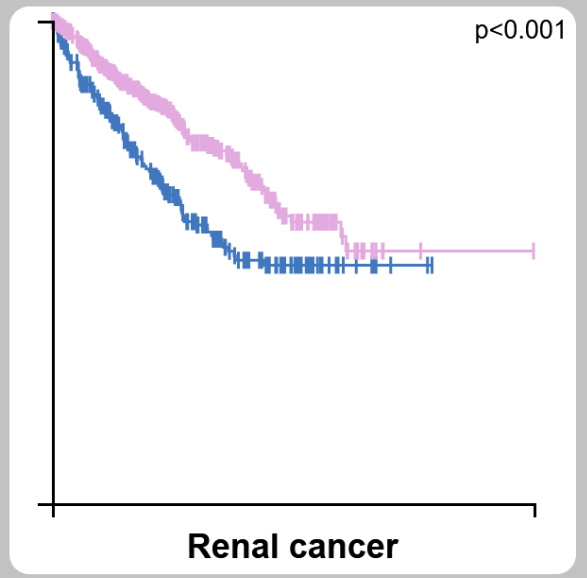
HUMAN PROTEIN ATLAS SUMMARY[i]

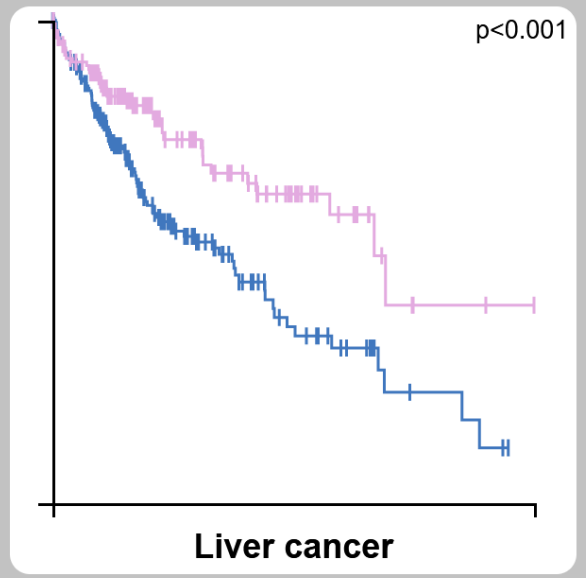
Protein[i]	Angiotensin I converting enzyme 2
Gene name[i]	ACE2
Tissue specificity[i]	Tissue enhanced (gallbladder, intestine, kidney)
Tissue expression cluster[i]	Intestine & Liver - Lipid metabolism (mainly)
Single cell type specificity[i]	Cell type enriched (Proximal enterocytes)
Single cell type expression cluster[i]	Enterocytes - Digestion (mainly)
Immune cell specificity[i]	Not detected in immune cells
Brain specificity[i]	Not detected in human brain
Cancer prognostic summary	Prognostic marker in renal cancer (favorable) and liver cancer (favorable)
Predicted location[i]	Membrane, Secreted (different isoforms)

The Human Protein Atlas

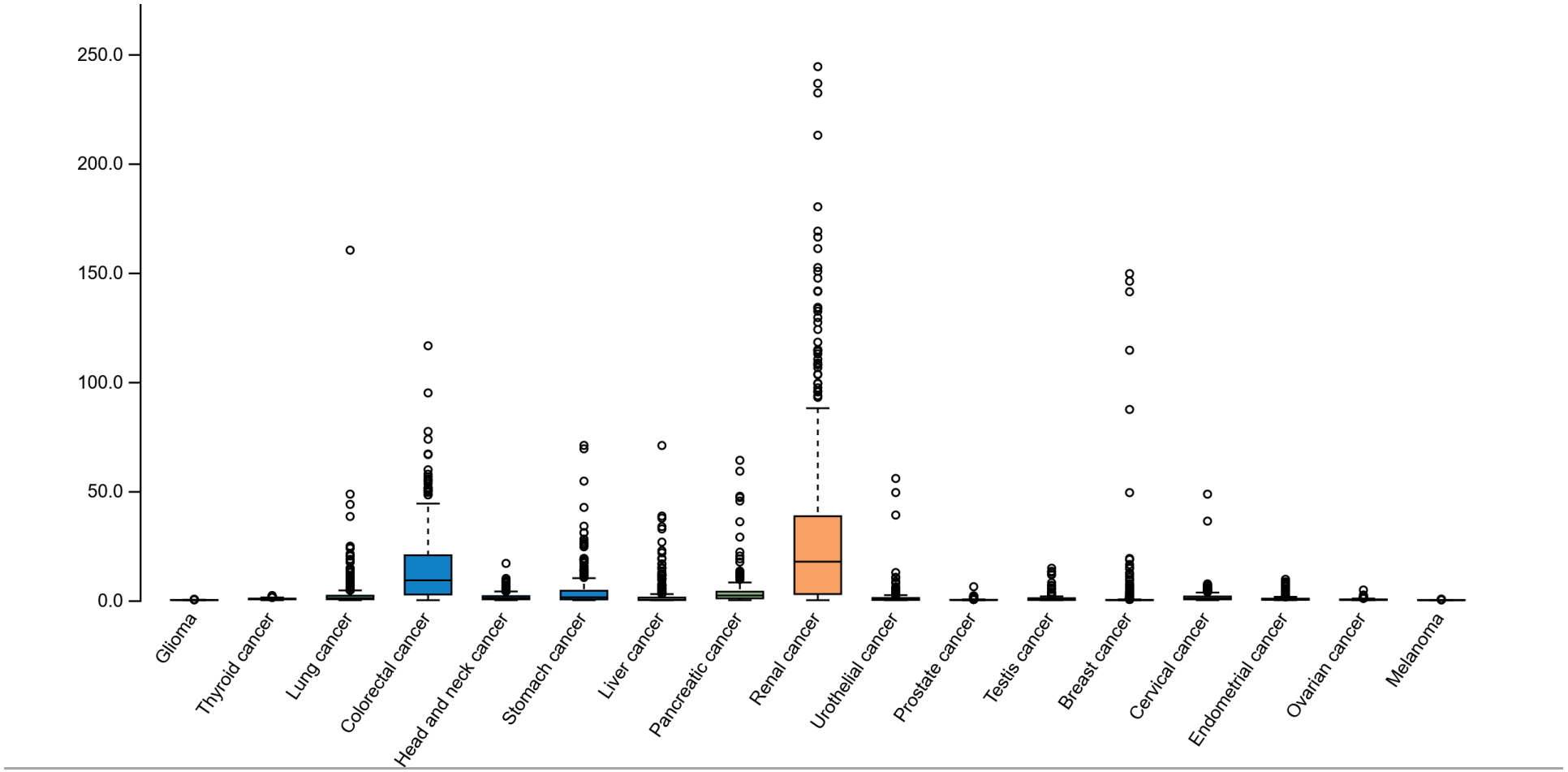
PROGNOSTIC SUMMARY[i]

Prognostic marker in renal cancer (favorable) and liver cancer (favorable)



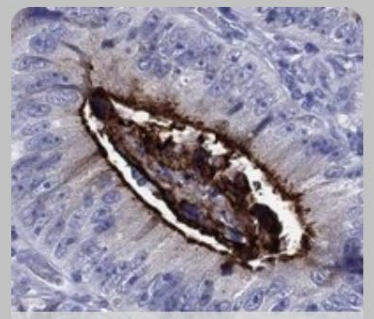


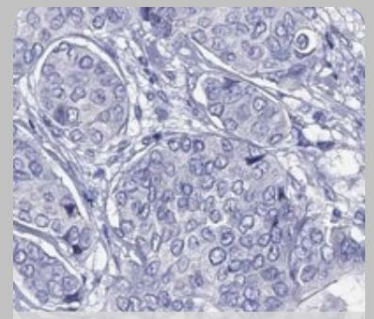
The Human Protein Atlas

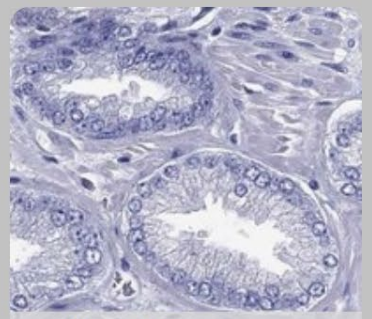


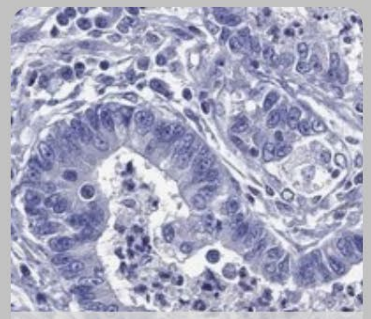
The Human Protein Atlas

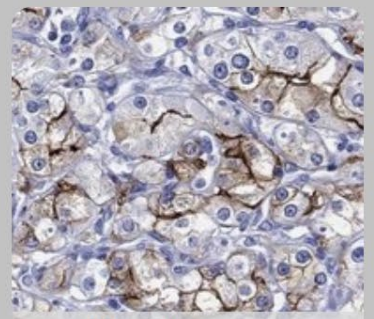
PROTEIN EXPRESSION[i]


Colorectal cancer


Breast cancer


Prostate cancer


Lung cancer

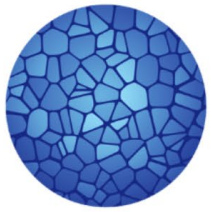

Renal cancer

PROTEIN EXPRESSION SUMMARY[i]

Organ | Expression | Alphabetical

HPA000288 CAB026174 CAB026213 CAB080024 CAB080025 CAB080027 CAB080028

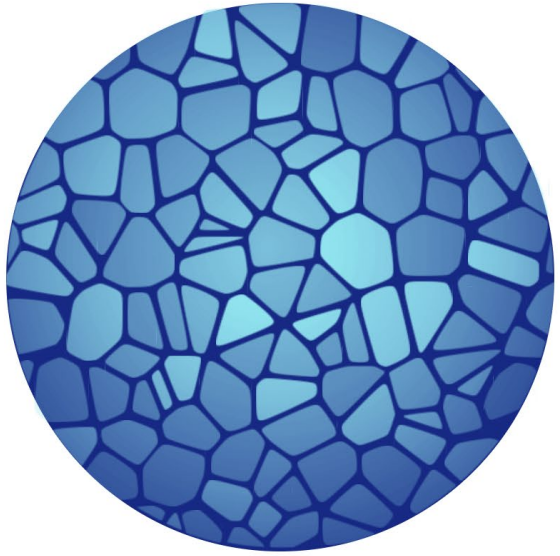
The Human Cell Atlas



HUMAN
CELL
ATLAS

Home | About HCA | COVID-19 | Research | News | Publications | Data/DCP | Resources | Join/Contact

HCA RESEARCH ON COVID-19 | HCA LIFETIME ALLIANCE ON COVID-19



HUMAN
CELL
ATLAS

The Human Cell Atlas

Immune Cell Atlas | Study overview | Help & resources | Create a study | Sign in

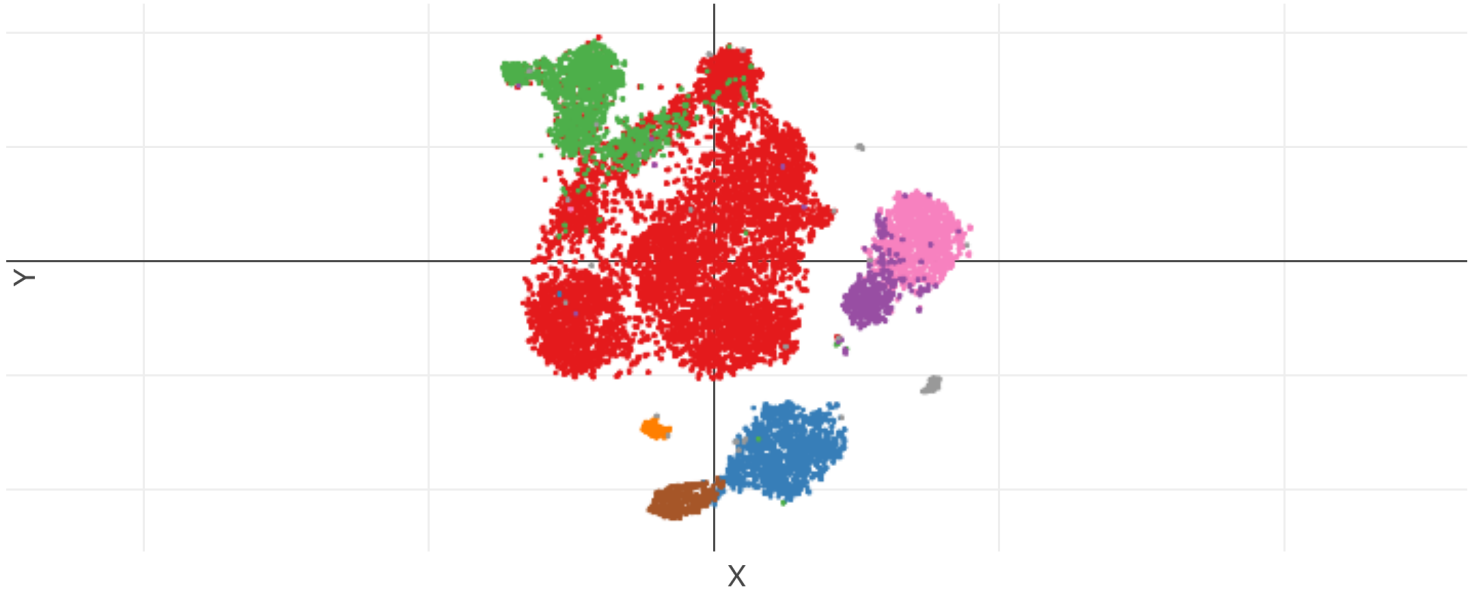
Study: ICA: Blood Mononuclear Cells (2 donors, 2 sites) 13316 cells

Summary | Explore | Download

Genes (e.g. "PTEN NF2")

Scatter

lineage_kmeans



AllCells - Primary Lineages

Show all | Hide all

1. T	8326
2. CD14+ Monocyte	1264
3. NK	1394
4. Memory B cell	491
5. DC	142
6. CD16+ Monocyte	398
7. Naive B cell	1169

OPTIONS

Clustering

lineage_kmeans

Annotation

AllCells - Primary L...

+ Create annotation

Subsampling

All Cells

Reset view

The Human Cell Atlas

nature COMMUNICATIONS

ARTICLE

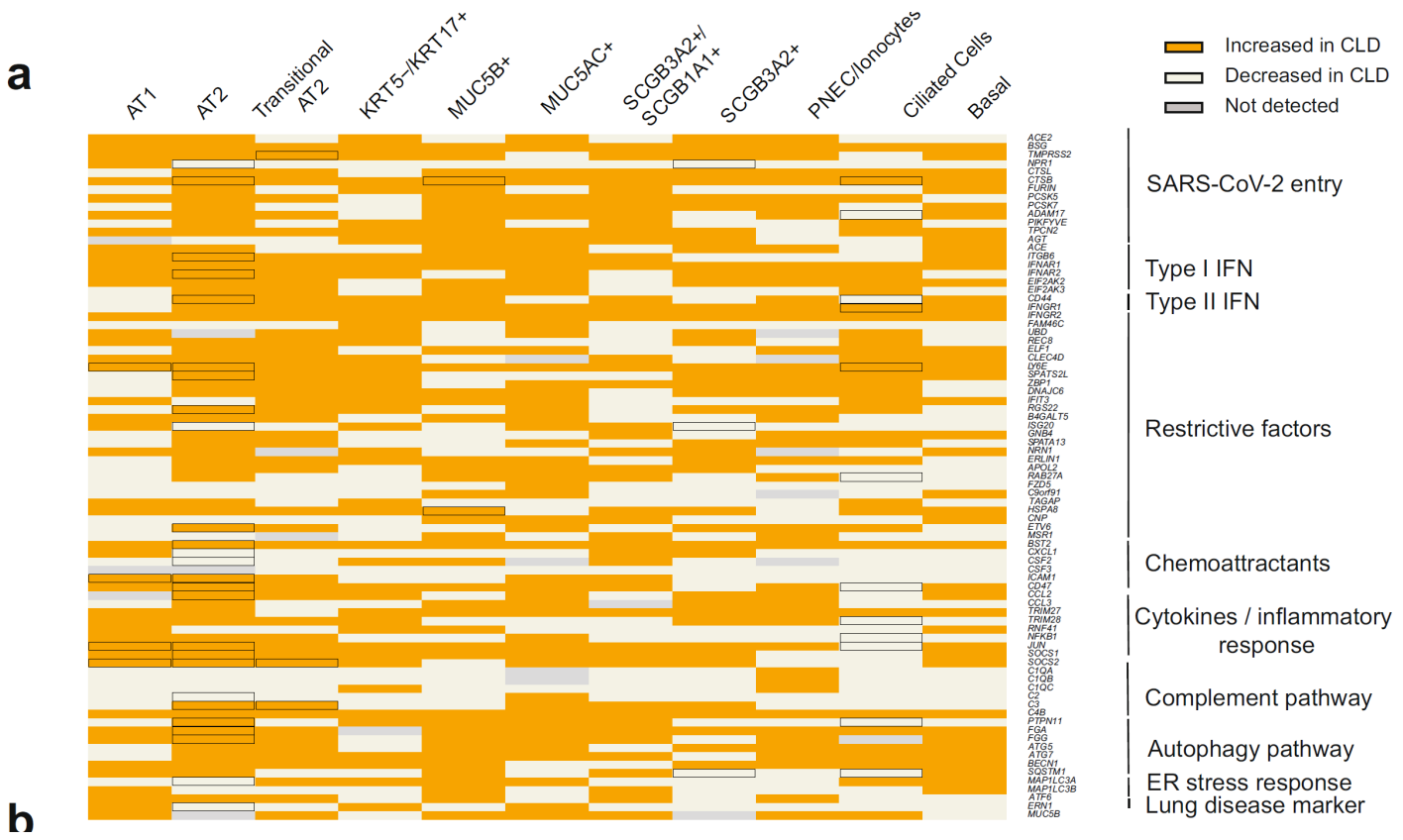
https://doi.org/10.1038/s41467-021-24467-0 OPEN

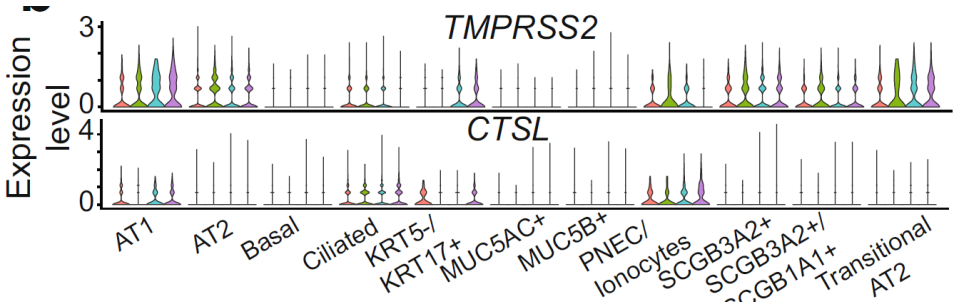
Chronic lung diseases are associated with gene expression programs favoring SARS-CoV-2 entry and severity

Linh T. Bui[1,87], Nichelle I. Winters[2,87], Mei-I Chung[1], Chitra Joseph[3], Austin J. Gutierrez[1], Arun C. Habermann[2], Taylor S. Adams[4], Jonas C. Schupp[4], Sergio Poli[5], Lance M. Peter[1], Chase J. Taylor[2], Jessica B. Blackburn[2], Bradley W. Richmond[2,6], Andrew G. Nicholson[7,8], Doris Rassl[9], William A. Wallace[10,11], Ivan O. Rosas[12], R. Gisli Jenkins[3], Naftali Kaminski[4], Jonathan A. Kropski[2,6,13,88], Nicholas E. Banovich[1,88]✉, and the Human Cell Atlas Lung Biological Network*

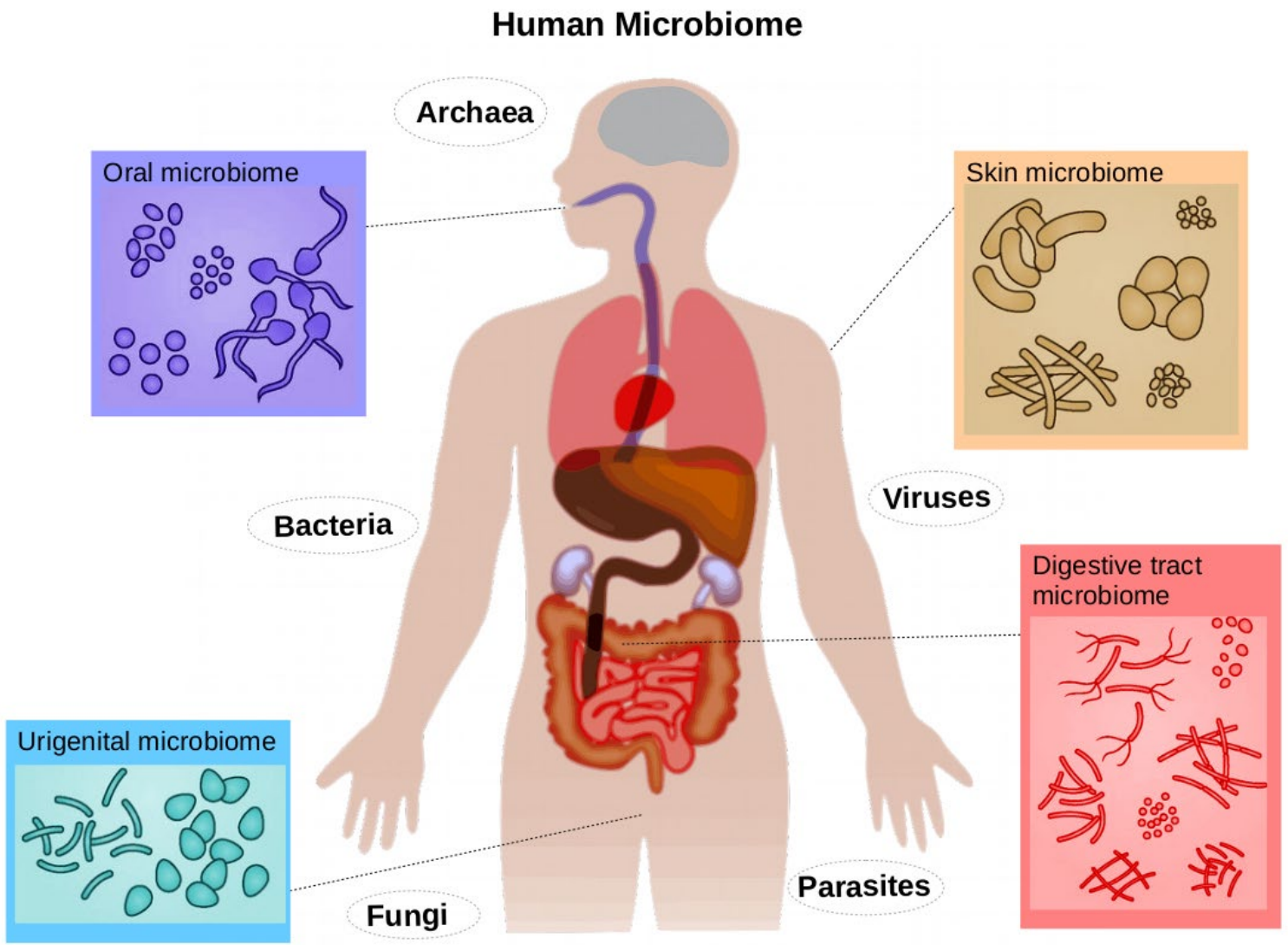
Patients with chronic lung disease (CLD) have an increased risk for severe coronavirus disease-19 (COVID-19) and poor outcomes. Here, we analyze the transcriptomes of 611,398 single cells isolated from healthy and CLD lungs to identify molecular characteristics of lung cells that may account for worse COVID-19 outcomes in patients with chronic lung diseases. We observe a similar cellular distribution and relative expression of SARS-CoV-2 entry factors in control and CLD lungs. CLD AT2 cells express higher levels of genes linked directly to the efficiency of viral replication and the innate immune response. Additionally, we identify basal differences in inflammatory gene expression programs that highlight how CLD alters the inflammatory microenvironment encountered upon viral exposure to the peripheral lung. Our study indicates that CLD is accompanied by changes in cell-type-specific gene expression programs that prime the lung epithelium for and influence the innate and adaptive immune responses to SARS-CoV-2 infection.

The Human Cell Atlas





The Human Microbiome



The Human Microbiome – Cancer Microbiome

RESEARCH

CANCER IMMUNOTHERAPY

The commensal micr… associated with anti-… metastatic melanom…

Vyara Matson,[1*] Jessica Fessler,[1*] Riyue Bao,[2,3*] Maria-Luisa Alegre,[4] Jason J. Luke,[4] Thomas F.

Anti–PD-1–based immunotherapy has had a majo… only benefited a subset of patients. Among the v… interpatient heterogeneity is differential composi… has been shown to affect antitumor immunity and… mouse models. We analyzed baseline stool sampl… before immunotherapy treatment, through an inte… sequencing, metagenomic shotgun sequencing, a… reaction for selected bacteria. A significant assoc… microbial composition and clinical response. Bact… responders included *Bifidobacterium longum*, *Col… faecium*. Reconstitution of germ-free mice with fe… could lead to improved tumor control, augmented… anti–PD-L1 therapy. Our results suggest that the… mechanistic impact on antitumor immunity in hu…

RESEARCH

CANCER IMMUNOTHERAPY

Gut microbiome modulates response to anti–PD-1 imm… melanoma patie…

V. Gopalakrishnan,[1,2*] C. N. Spencer, T. V. Karpinets,[3] P. A. Prieto,[1†] D. Vi… L. Zhao,[3] C. W. Hudgens,[6] D. S. Hutc… T. Cotechini,[8] T. Kumar,[3] W. S. Chen… J. Galloway-Pena,[11] H. Jiang,[1] P. L. C… R. F. Chemaly,[11] S. Shelburne,[3,11] L. M… A. G. Swennes,[7] F. McAllister,[14] E. M… E. Le Chatelier,[15] L. Zitvogel,[16] N. Po… E. M. Burton,[1] J. M. Gardner,[1] E. Sir… A. Diab,[17] H. Tawbi,[17] I. C. Glitza,[17] W… R. N. Amaria,[17] M. A. Davies,[17] J. E. … L. M. Coussens,[8] Z. A. Cooper,[1,3¶] P. … J. F. Petrosino,[7] M. T. Tetzlaff,[6,9] P. S… R. R. Jenq,[3#] J. A. Wargo[1,3#**]

RESEARCH

CANCER IMMUNOTHERAPY

Gut microbiome influences efficacy of PD-1–based immunotherapy against epithelial tumors

Bertrand Routy,[1,2,3] Emmanuelle Le Chatelier,[4] Lisa Derosa,[1,2,3] Connie P. M. Duong,[1,2,5] Maryam Tidjani Alou,[1,2,3] Romain Daillère,[1,2,3] Aurélie Fluckiger,[1,2,5] Meriem Messaoudene,[1,2] Conrad Rauber,[1,2,3] Maria P. Roberti,[1,2,5] Marine Fidelle,[1,3,5] Caroline Flament,[1,2,5] Vichnou Poirier-Colame,[1,2,5] Paule Opolon,[6] Christophe Klein,[7] Kristina Iribarren,[8,9,10,11,12] Laura Mondragón,[8,9,10,11,12] Nicolas Jacquelot,[1,2,3] Bo Qu,[1,2,3] Gladys Ferrere,[1,2,3] Céline Clémenson,[1,13] Laura Mezquita,[1,14] Jordi Remon Masip,[1,14] Charles Naltet,[15] Solenn Brosseau,[15] Coureche Kaderbhai,[16] Corentin Richard,[16] Hira Rizvi,[17] Florence Levenez,[4] Nathalie Galleron,[4] Benoit Quinquis,[4] Nicolas Pons,[4] Bernhard Ryffel,[18] Véronique Minard-Colin,[1,19] Patrick Gonin,[1,20] Jean-Charles Soria,[1,14] Eric Deutsch,[1,13] Yohann Loriot,[1,3,14] François Ghiringhelli,[16] Gérard Zalcman,[15] François Goldwasser,[9,21,22] Bernard Escudier,[1,14,23] Matthew D. Hellmann,[24,25] Alexander Eggermont,[1,2,14] Didier Raoult,[26] Laurence Albiges,[1,3,14] Guido Kroemer,[8,9,10,11,12,27,28*] Laurence Zitvogel[1,2,3,5*]

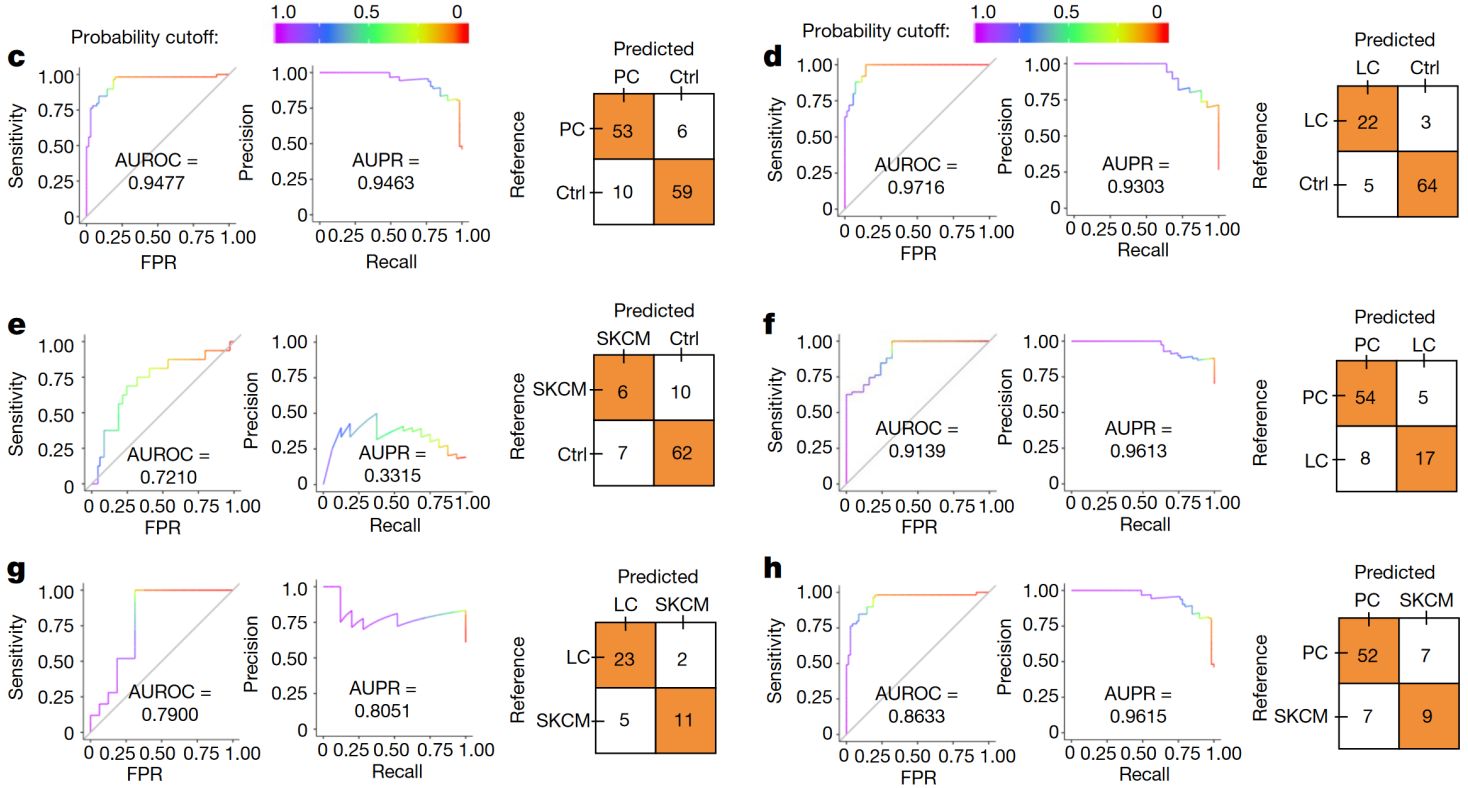
The Human Microbiome – Cancer Microbiome

Microbiome analyses of blood and tissues suggest cancer diagnostic approach

https://doi.org/10.1038/s41586-020-2095-1
Received: 7 June 2019
Accepted: 6 February 2020
Published online: 11 March 2020
Check for updates

Gregory D. Poore[1,12], Evguenia Kopylova[2,9,12], Qiyun Zhu[2], Carolina Carpenter[3], Serena Fraraccio[3], Stephen Wandro[3], Tomasz Kosciolek[2,10], Stefan Janssen[2,11], Jessica Metcalf[4], Se Jin Song[3], Jad Kanbar[5], Sandrine Miller-Montgomery[1,3], Robert Heaton[6], Rana Mckay[7], Sandip Pravin Patel[3,7], Austin D. Swafford[3] & Rob Knight[1,2,3,8]

Systematic characterization of the cancer microbiome provides the opportunity to develop techniques that exploit non-human, microorganism-derived molecules in the diagnosis of a major human disease. Following recent demonstrations that some types of cancer show substantial microbial contributions[1–10], we re-examined whole-genome and whole-transcriptome sequencing studies in The Cancer Genome Atlas[11] (TCGA) of 33 types of cancer from treatment-naive patients (a total of 18,116 samples) for microbial reads, and found unique microbial signatures in tissue and blood within and between most major types of cancer. These TCGA blood signatures remained predictive when applied to patients with stage Ia–IIc cancer and cancers lacking any genomic alterations currently measured on two commercial-grade cell-free tumour DNA platforms, despite the use of very stringent decontamination analyses that discarded up to 92.3% of total sequence data. In addition, we could discriminate among samples from healthy, cancer-free individuals ($n = 69$) and those from patients with multiple types of cancer (prostate, lung, and melanoma; 100 samples in total) solely using plasma-derived, cell-free microbial nucleic acids. This potential microbiome-based oncology diagnostic tool warrants further exploration.



Sample data

ARTICLE

doi:10.1038/nature10166

Integrated genomic analyses of ovarian carcinoma

The Cancer Genome Atlas Research Network*

A catalogue of molecular aberrations that cause ovarian cancer is critical for developing and deploying therapies that will improve patients' lives. The Cancer Genome Atlas project has analysed messenger RNA expression, microRNA expression, promoter methylation and DNA copy number in 489 high-grade serous ovarian adenocarcinomas and the DNA sequences of exons from coding genes in 316 of these tumours. Here we report that high-grade serous ovarian cancer is characterized by *TP53* mutations in almost all tumours (96%); low prevalence but statistically recurrent somatic mutations in nine further genes including *NF1*, *BRCA1*, *BRCA2*, *RB1* and *CDK12*; 113 significant focal DNA copy number aberrations; and promoter methylation events involving 168 genes. Analyses delineated four ovarian cancer transcriptional subtypes, three microRNA subtypes, four promoter methylation subtypes and a transcriptional signature associated with survival duration, and shed new light on the impact that tumours with *BRCA1/2* (*BRCA1* or *BRCA2*) and *CCNE1* aberrations have on survival. Pathway analyses suggested that homologous recombination is defective in about half of the tumours analysed, and that NOTCH and FOXM1 signalling are involved in serous ovarian cancer pathophysiology.

Sample data

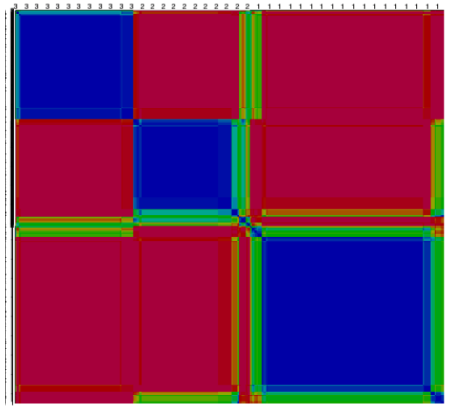
ARTICLE

doi:10.1038/nature10166

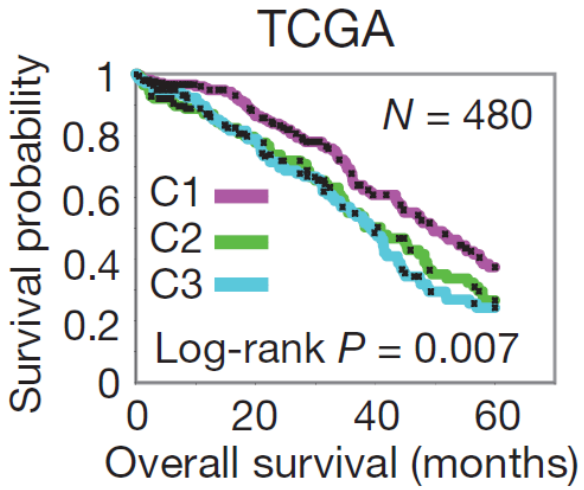
Integrated genomic analyses of ovarian carcinoma

The Cancer Genome Atlas Research Network*

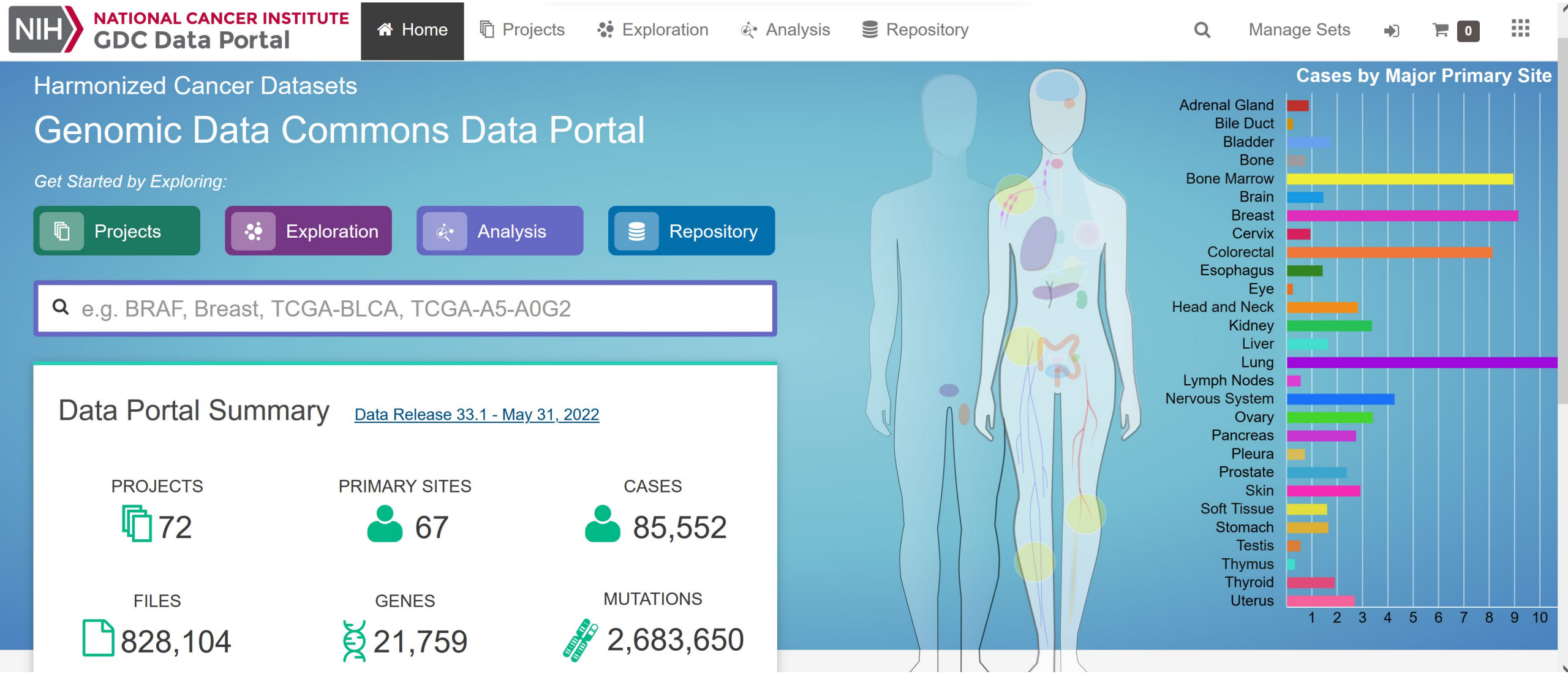
150 most variable miRNAs for sample clustering



Cluster 1 has significantly better survival



https://portal.gdc.cancer.gov



https://portal.gdc.cancer.gov

The Cancer Genome Atlas (TCGA)
Ovarian cancer

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home | Projects | Exploration | Analysis | Repository | Manage Sets | 0

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

Files | Cases

Add a File Filter

Clear | Program Name | IS | TCGA | AND | Project Id | IS | TCGA-OV

Search Files

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

Data Category

	# Files
simple nucleotide variation	8,408
copy number variation	3,578
sequencing reads	2,684
biospecimen	2,601
dna methylation	1,869

4 More...

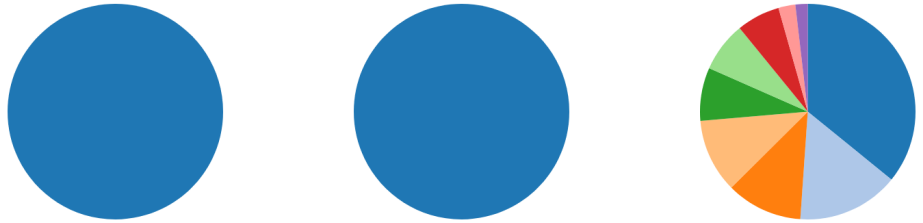
Data Type

	# Files
Annotated Somatic Mutation	4,900
Aligned Reads	2,684

Files (23,453) | Cases (608)

Add All Files to Cart | Manifest | View Images

Primary Site | Project | Data Category



Show More

Showing 1 - 20 of 23,453 files 46.67 TB

JSON | TSV

Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
	ec5ab2eb-37b0-4f3c-a35c-						

ID	Disease Type	Primary Site	Program	Cases
TCGA-BRCA	Breast Invasive Carcinoma	Breast	TCGA	1,098
TCGA-GBM	Glioblastoma Multiforme	Brain	TCGA	617
TCGA-OV	Ovarian Serous Cystadenocarcin…	Ovary	TCGA	608
TCGA-LUAD	Lung Adenocarcinoma	Lung	TCGA	585
TCGA-UCEC	Uterine Corpus Endometrial Carci…	Uterus	TCGA	560
TCGA-KIRC	Kidney Renal Clear Cell Carcinoma	Kidney	TCGA	537
TCGA-HNSC	Head and Neck Squamous Cell C…	Head and Neck	TCGA	528
TCGA-LGG	Brain Lower Grade Glioma	Brain	TCGA	516
TCGA-THCA	Thyroid Carcinoma	Thyroid	TCGA	507
TCGA-LUSC	Lung Squamous Cell Carcinoma	Lung	TCGA	504
TCGA-PRAD	Prostate Adenocarcinoma	Prostate	TCGA	500
TCGA-SKCM	Skin Cutaneous Melanoma	Skin	TCGA	470
TCGA-COAD	Colon Adenocarcinoma	Colorectal	TCGA	461
TCGA-STAD	Stomach Adenocarcinoma	Stomach	TCGA	443
TCGA-BLCA	Bladder Urothelial Carcinoma	Bladder	TCGA	412
TCGA-LIHC	Liver Hepatocellular Carcinoma	Liver	TCGA	377
TCGA-CESC	Cervical Squamous Cell Carcino…	Cervix	TCGA	307
TCGA-KIRP	Kidney Renal Papillary Cell Carci…	Kidney	TCGA	291
TCGA-SARC	Sarcoma	Soft Tissue	TCGA	261
TCGA-LAML	Acute Myeloid Leukemia	Bone Marrow	TCGA	200
TCGA-PAAD	Pancreatic Adenocarcinoma	Pancreas	TCGA	185
TCGA-ESCA	Esophageal Carcinoma	Esophagus	TCGA	185
TCGA-PCPG	Pheochromocytoma and Paragan…	Adrenal Gland	TCGA	179
TCGA-READ	Rectum Adenocarcinoma	Colorectal	TCGA	172
TCGA-TGCT	Testicular Germ Cell Tumors	Testis	TCGA	150
TCGA-THYM	Thymoma	Thymus	TCGA	124
TCGA-KICH	Kidney Chromophobe	Kidney	TCGA	113
TCGA-ACC	Adrenocortical Carcinoma	Adrenal Gland	TCGA	92
TCGA-MESO	Mesothelioma	Pleura	TCGA	87
TCGA-UVM	Uveal Melanoma	Eye	TCGA	80
TCGA-DLBC	Lymphoid Neoplasm Diffuse Larg…	Lymph Nodes	TCGA	58
TCGA-UCS	Uterine Carcinosarcoma	Uterus	TCGA	57
TCGA-CHOL	Cholangiocarcinoma	Bile Duct	TCGA	51
Total				11,315

TCGA cancer types (n=33)

- TCGA raw data were harmonized by NCI's GDC team
  - Release 6 June 2017
  - Release 33.1 (May 2022; current)
- Data types: from raw files to compiled results
- Result access: public or protected
- Apply for access: dbGap
- Download: GDC

Access GDC data

```
[rbao@cri16in002 ~]$ gdc-client -h
usage: gdc-client [-h] [--version] {download,upload,interactive} ...

The Genomic Data Commons Command Line Client

optional arguments:
  -h, --help            show this help message and exit
  --version             show program's version number and exit

commands:
  {download,upload,interactive}
                        for more information, specify -h after a command
    download            download data from the GDC
    upload              upload data to the GDC
    interactive         run in interactive mode
```

Bioconductor packages

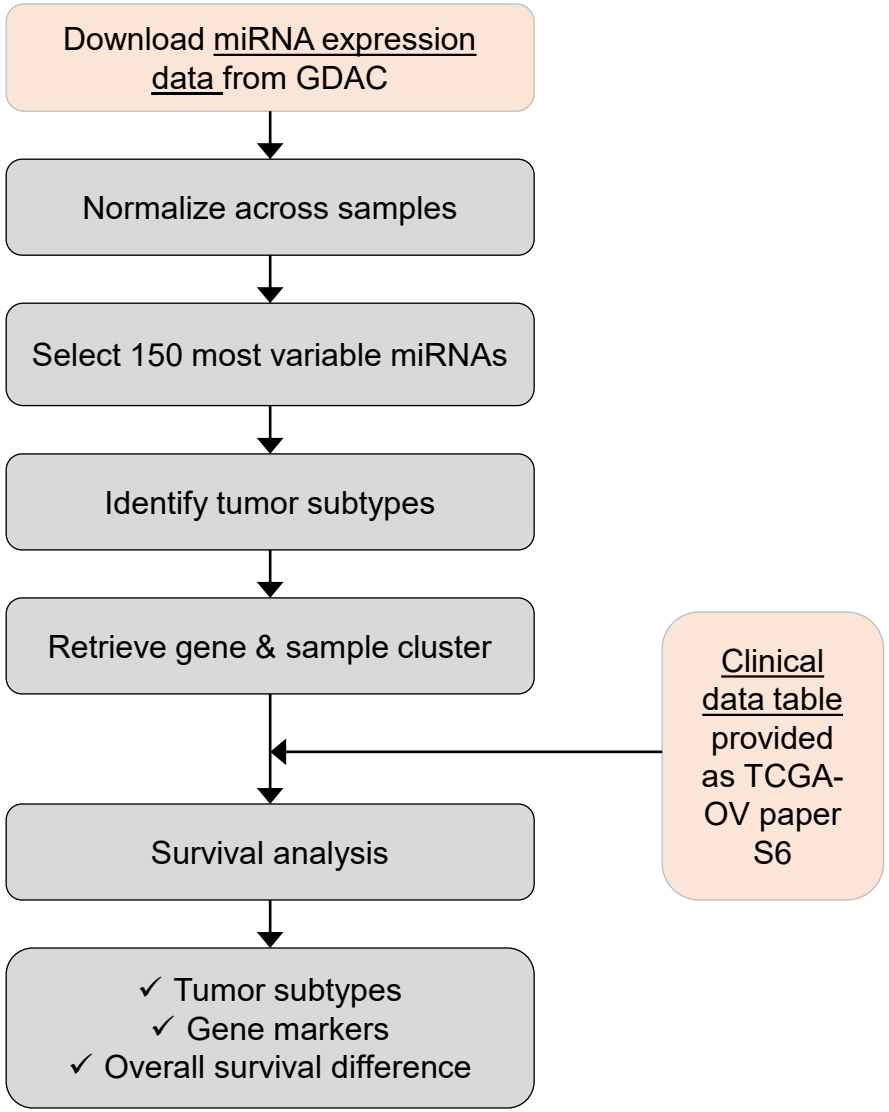
- library(TCGAbiolinks)
  - https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html
- library(GenomicDataCommons)
  - https://bioconductor.org/packages/release/bioc/html/GenomicDataCommons.html

The GDC Application Programming Interface (API): An Overview

The GDC API drives the GDC Data and Submission Portals and provides programmatic access to GDC functionality. This includes searching for, downloading, and submitting data and metadata. The GDC API uses JSON as its communication format, and standard HTTP methods like `GET`, `PUT`, `POST` and `DELETE`.

```
curl https://api.gdc.cancer.gov/files/e6fe614f-7fb4-4d6c-85c1-311ec25af938?pretty=true
```

Integrate Expression with Clinical Data



Integrate Expression with Clinical Data

Download miRNA expression data from GDAC

Normalize across samples

Select 150 most variable miRNAs

Identify tumor subtypes

Retrieve gene & sample cluster

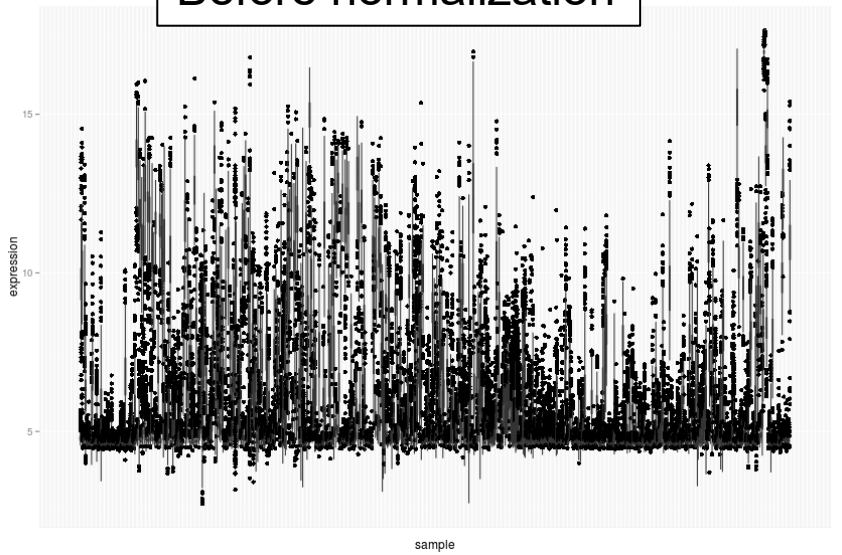
Clinical data table provided as TCGA-OV paper S6

Survival analysis

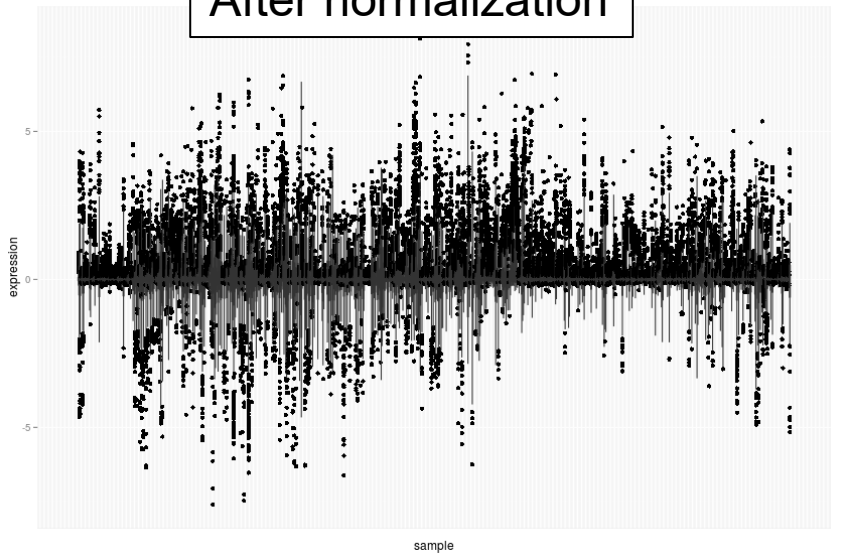
✓ Tumor subtypes
✓ Gene markers
✓ Overall survival difference

miRNA	Sample Total	Sample in TCGA-OV paper
821	594	486

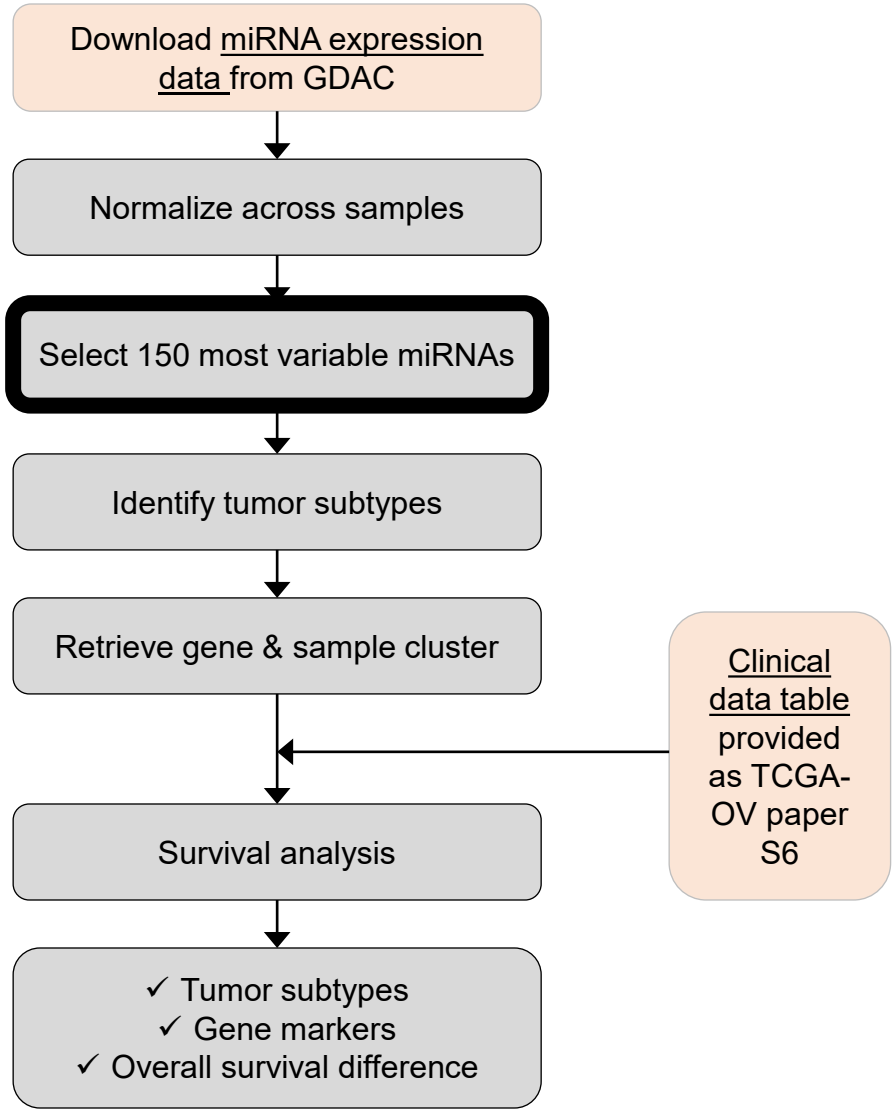
Before normalization



After normalization

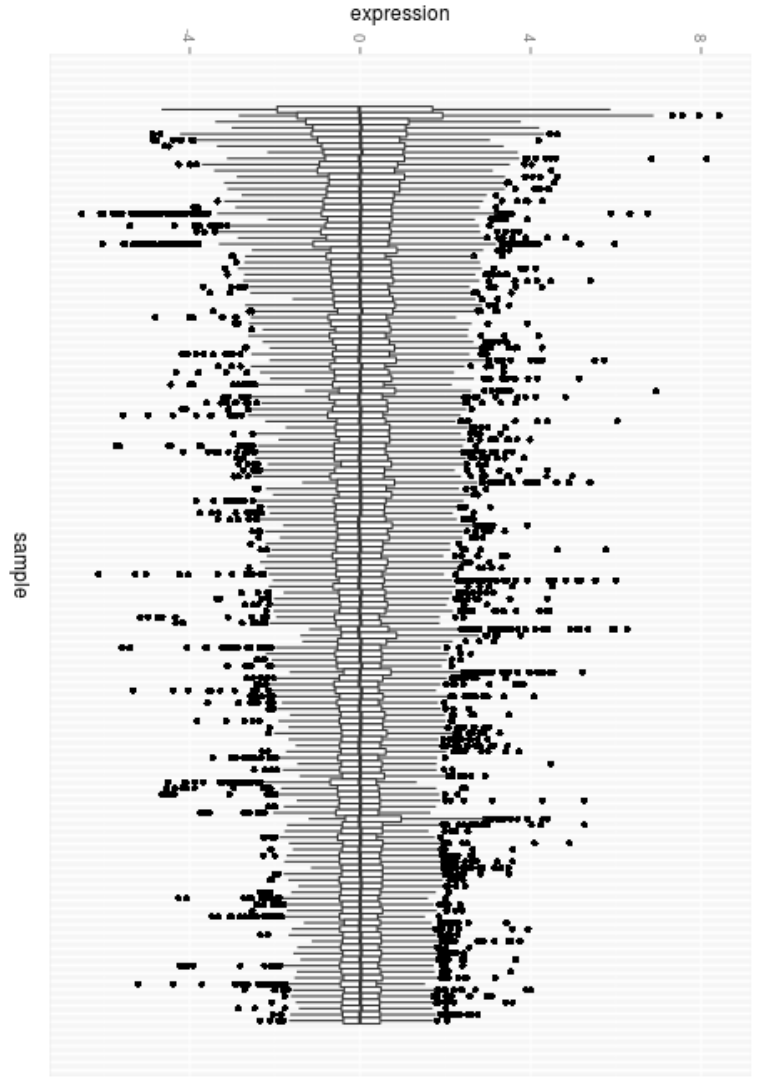


Integrate Expression with Clinical Data



	mad
HSA-MIR-205	2.6824804
HSA-MIR-449A	2.3890068
HSA-MIR-31	1.8097683
HSA-MIR-224	1.6650244
HSA-MIR-451	1.6289215
HSA-MIR-10A	1.4811346
HSA-MIR-10B	1.4523870
HSA-MIR-31*	1.4370410
HSA-MIR-363	1.3751314
HSA-MIR-96	1.3709655
HSA-MIR-203	1.3489193
HSA-MIR-494	1.2932659

Showing 1 to 13 of 150 entries



Median absolute deviation (MAD)

$$\mathrm{MAD} = \mathrm{median}_i\,(\,|X_i - \mathrm{median}_j(X_j)|\,)$$

Integrate Expression with Clinical Data

Download miRNA expression data from GDAC

Normalize across samples

Select 150 most variable miRNAs

Identify tumor subtypes

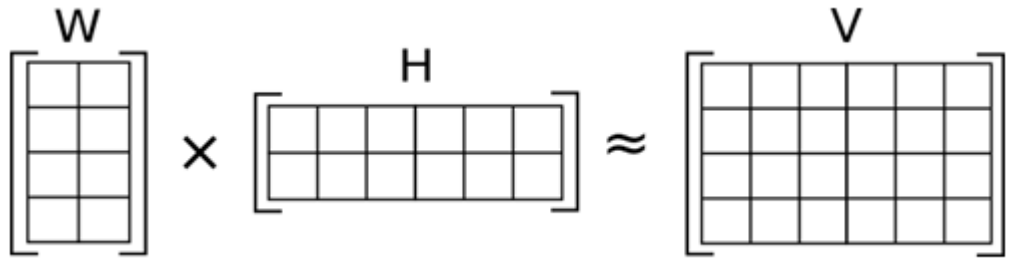
Retrieve gene & sample cluster

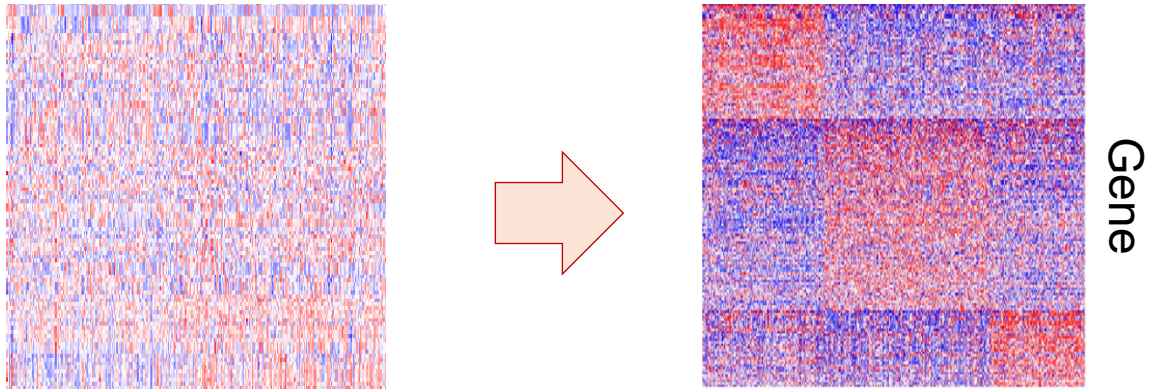
Clinical data table provided as TCGA-OV paper S6

Survival analysis

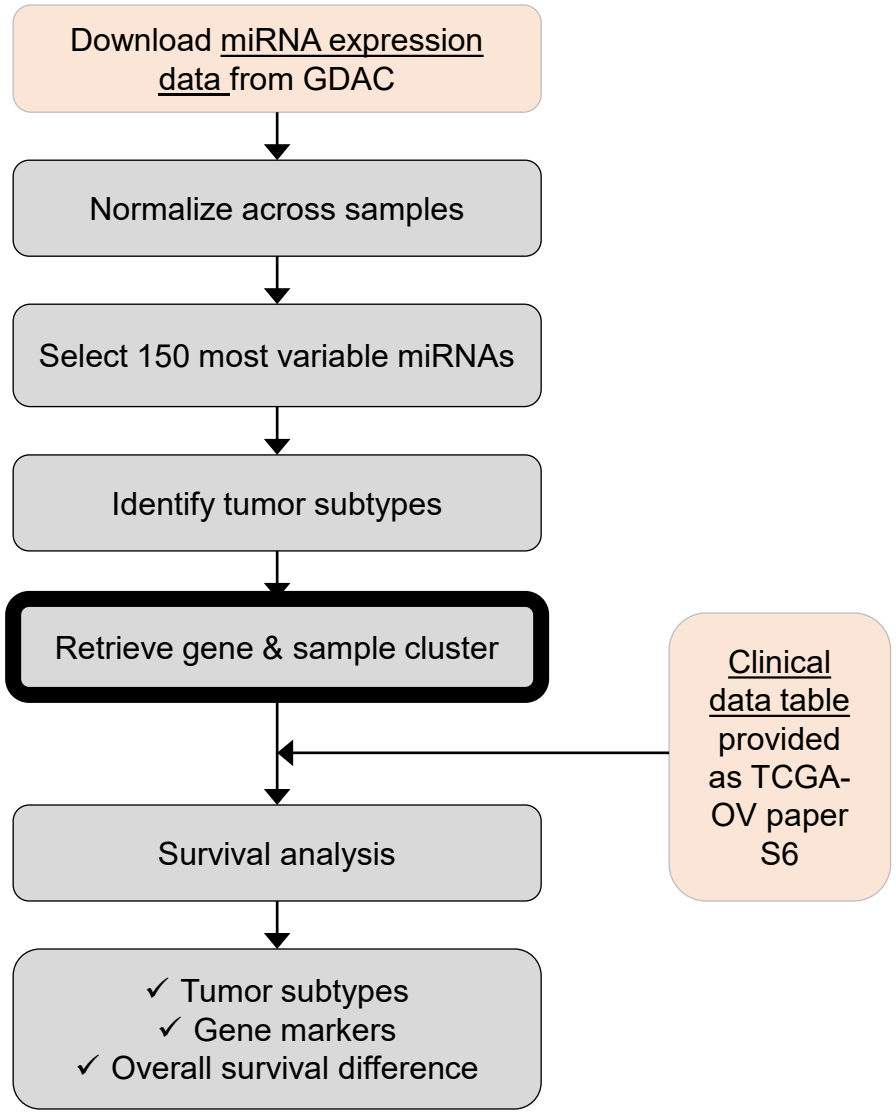
✓ Tumor subtypes
✓ Gene markers
✓ Overall survival difference

Non-negative matrix factorization (NMF), also non-negative matrix approximation is a group of algorithms in multivariate analysis and linear algebra where a matrix V is factorized into (usually) two matrices W and H, with the property that all three matrices have no negative elements.





Integrate Expression with Clinical Data



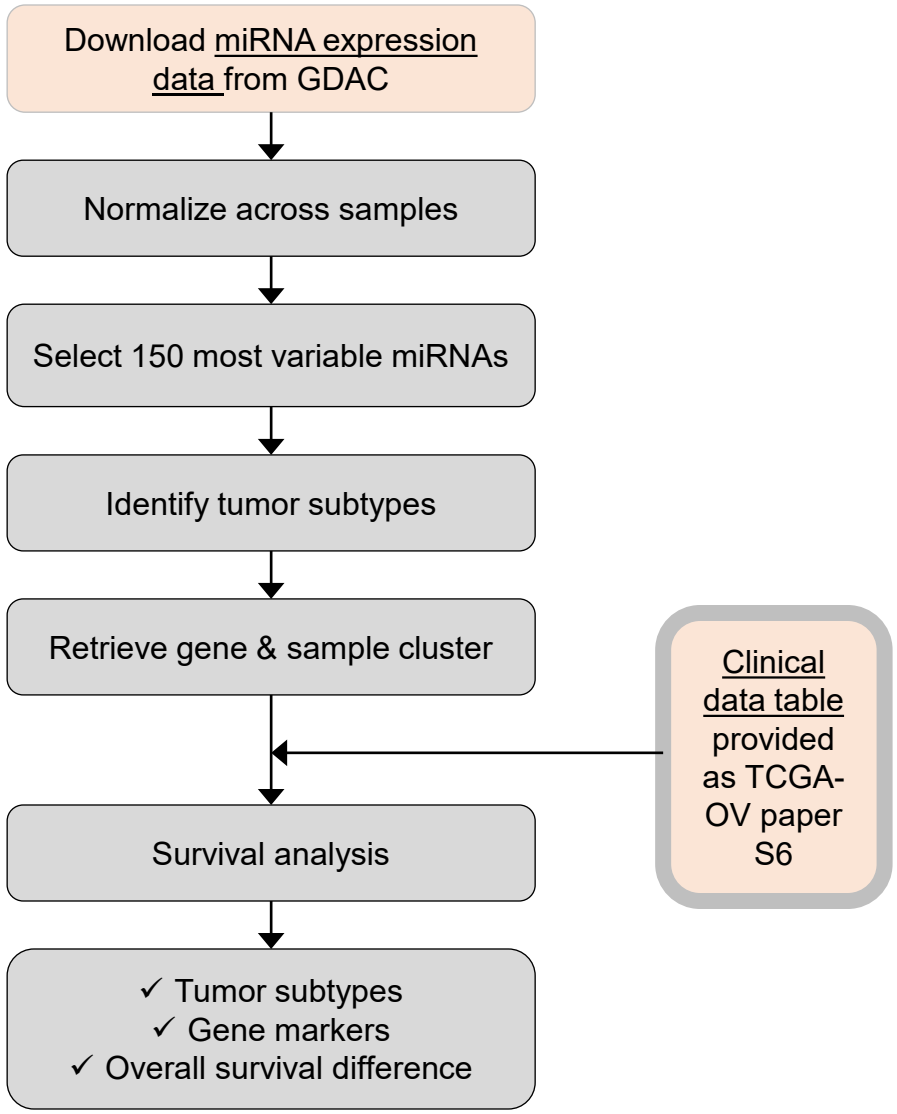
sample	cluster
TCGA.04.1331	1
TCGA.04.1338	1
TCGA.04.1341	1
TCGA.04.1343	1
TCGA.04.1348	1
TCGA.04.1362	1
TCGA.04.1365	1
TCGA.04.1530	1
TCGA.04.1542	1
TCGA.04.1648	1
TCGA.04.1649	1
TCGA.04.1651	1

Sample clusters

gene	cluster
HSA-MIR-205	1
HSA-MIR-494	1
HSA-MIR-144	1
HSA-MIR-142-5P	1
HSA-MIR-181A	1
HSA-MIR-151-3P	1
HSA-MIR-1225-5P	1
HSA-MIR-222	1
HSA-MIR-638	1
EBV-MIR-BART19-3P	1
HSA-MIR-21*	1
HSA-MIR-630	1

Gene clusters

Integrate Expression with Clinical Data

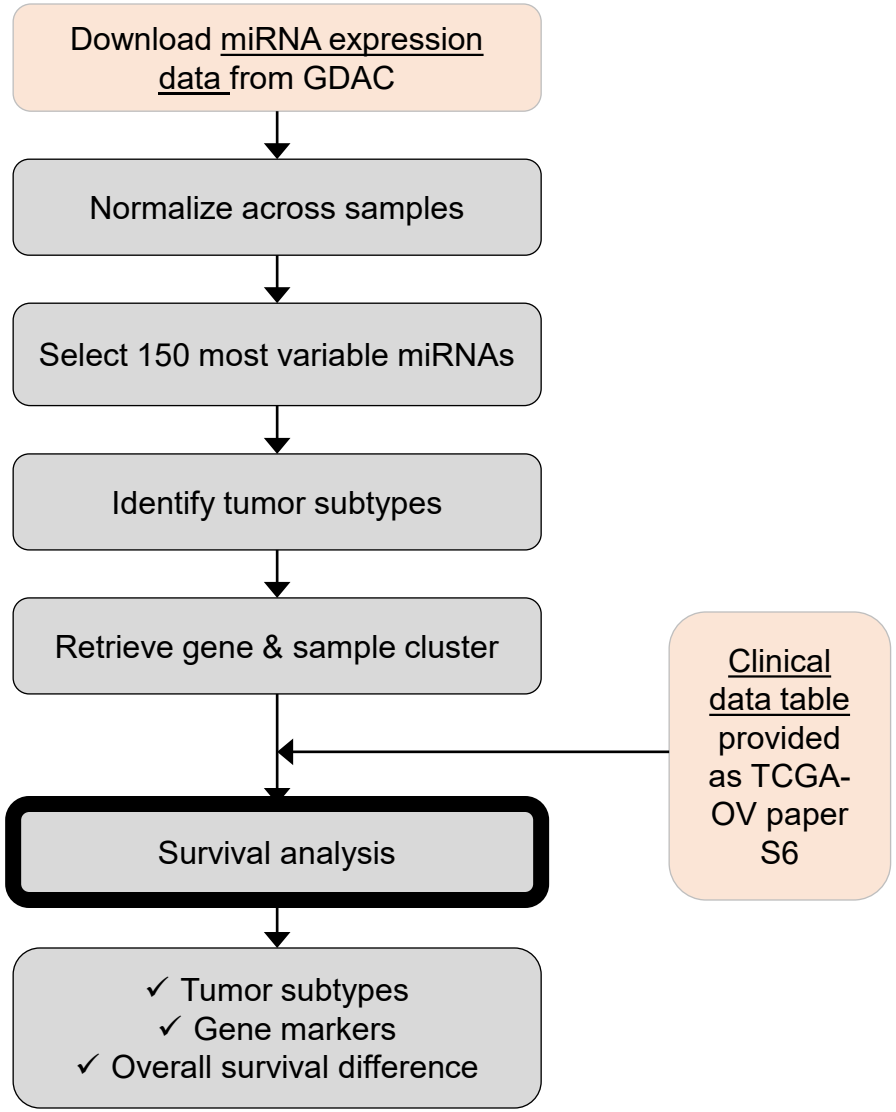


	sample	vital.status	overall.survival.month	age.at.diagnosis.year	tumor.stage	tumor.grade
1	TCGA.04.1331	DECEASED	43.80	79.04	IIIC	G3
5	TCGA.04.1338	LIVING	46.49	78.87	IIIC	G3
6	TCGA.04.1341	LIVING	NA	85.52	NA	G3
8	TCGA.04.1343	DECEASED	11.84	72.41	IV	G3
11	TCGA.04.1348	DECEASED	48.62	44.48	IIIB	G3
17	TCGA.04.1362	DECEASED	44.20	59.58	IIC	G3
19	TCGA.04.1365	LIVING	76.33	87.47	IIIB	G3
25	TCGA.04.1530	DECEASED	118.75	68.53	IIIC	G3
26	TCGA.04.1542	DECEASED	83.97	52.78	IIIB	G2
29	TCGA.04.1648	DECEASED	28.56	57.84	IIIC	G2
30	TCGA.04.1649	LIVING	64.46	74.42	IIIC	G3
31	TCGA.04.1651	DECEASED	36.07	53.78	IIIC	G3

Showing 1 to 13 of 486 entries

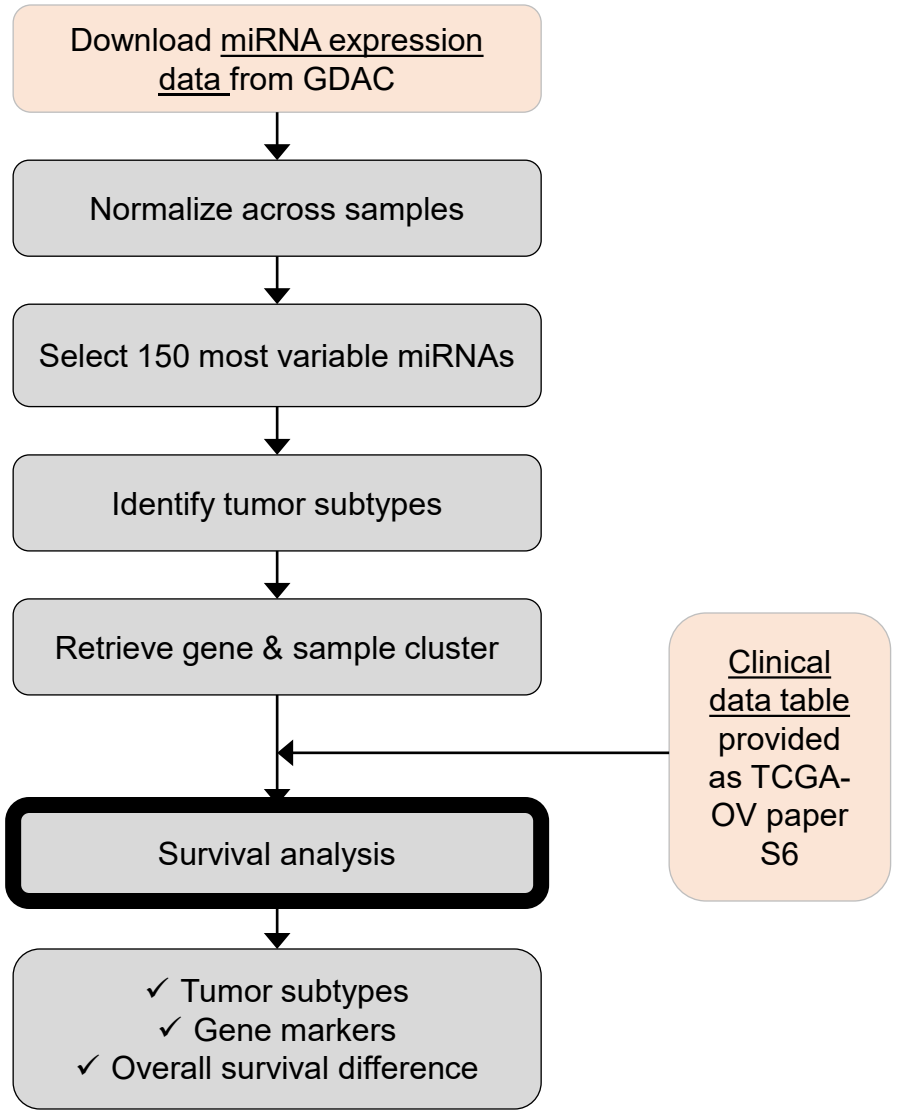
- Sample ID
- Vital status
- Overall survival
- Age at diagnosis
- Tumor stage
- Tumor grade
- **Sample cluster** (from the previous step)

Integrate Expression with Clinical Data



- Survival analysis
  - Study the time between entry to a study or an event (such as death)
  - Calculate survival/risk difference and detect significance between groups
  - Build models to predict prognosis
- Survival data
  - Time to event (in year, month, etc.)
  - Status (whether the event has happened?)
    - Censoring: only some individuals have experienced the event by the last follow up, while for others, the time is unknown
  - "cumulative" survival time
- Survival methods
  - Kaplan-Meier estimator
  - Log-rank test (Mantel-Haenzel test)
  - Cox regression model (proportional hazard model)

Integrate Expression with Clinical Data



Right Censoring

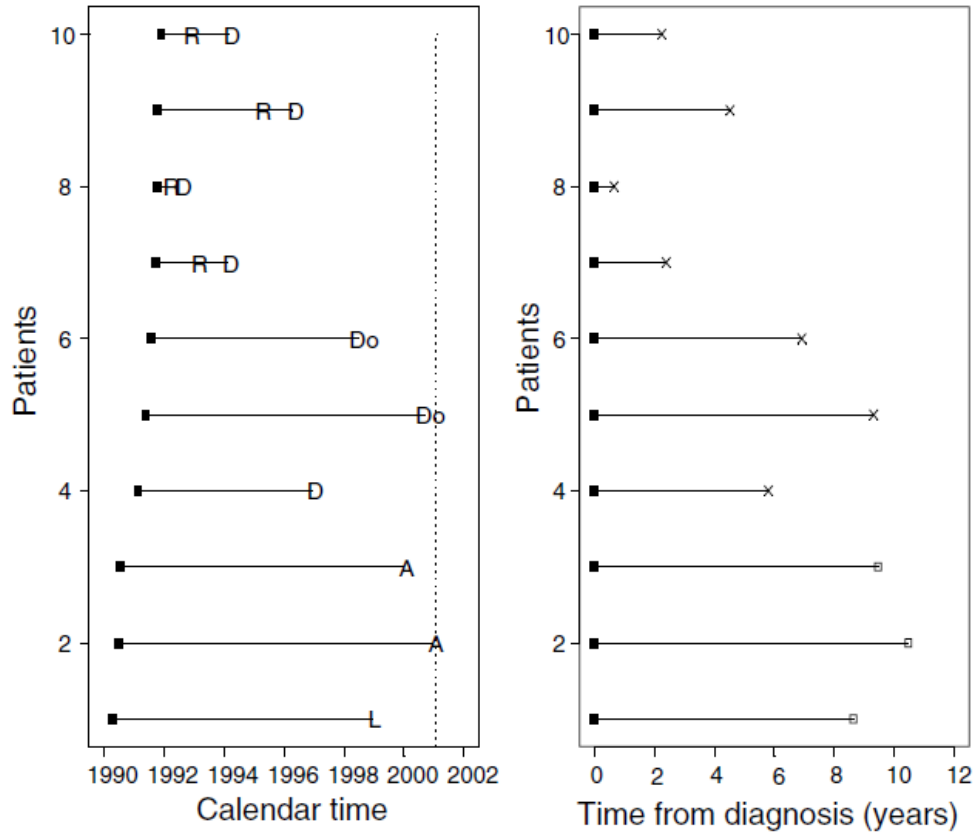
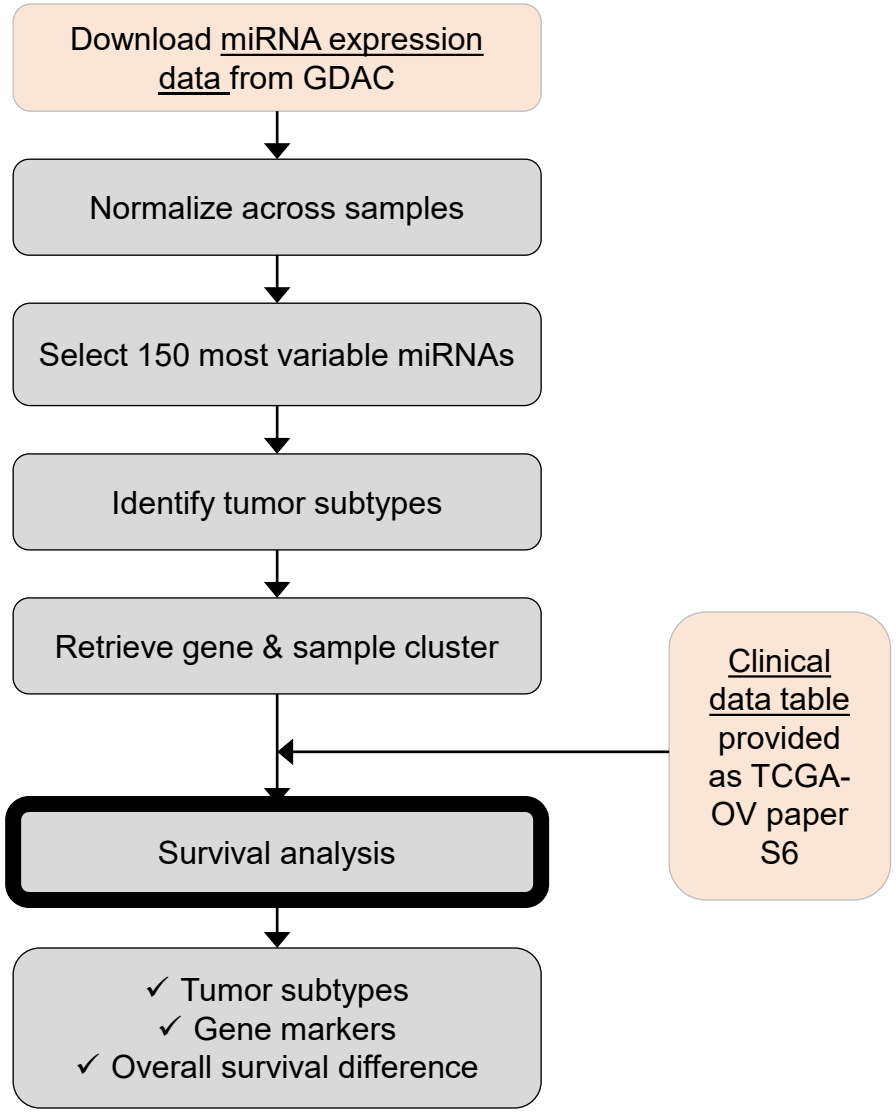


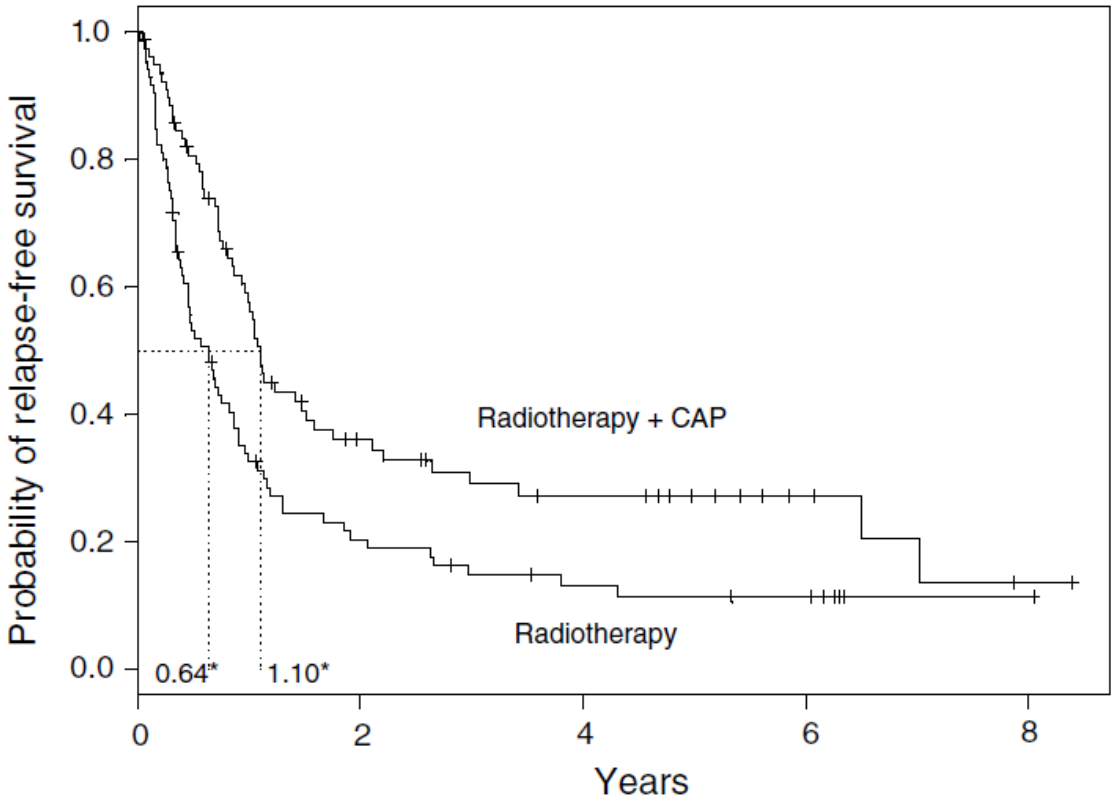
Figure 1 Converting calendar time in the ovarian cancer study to a survival analysis format. Dashed vertical line is the date of the last follow-up, R = relapse, D = death from ovarian cancer, Do = death from other cause, A = attended last clinic visit (alive), L = loss to follow-up, X = death, □ = censored.

Clark et al. Survival Analysis

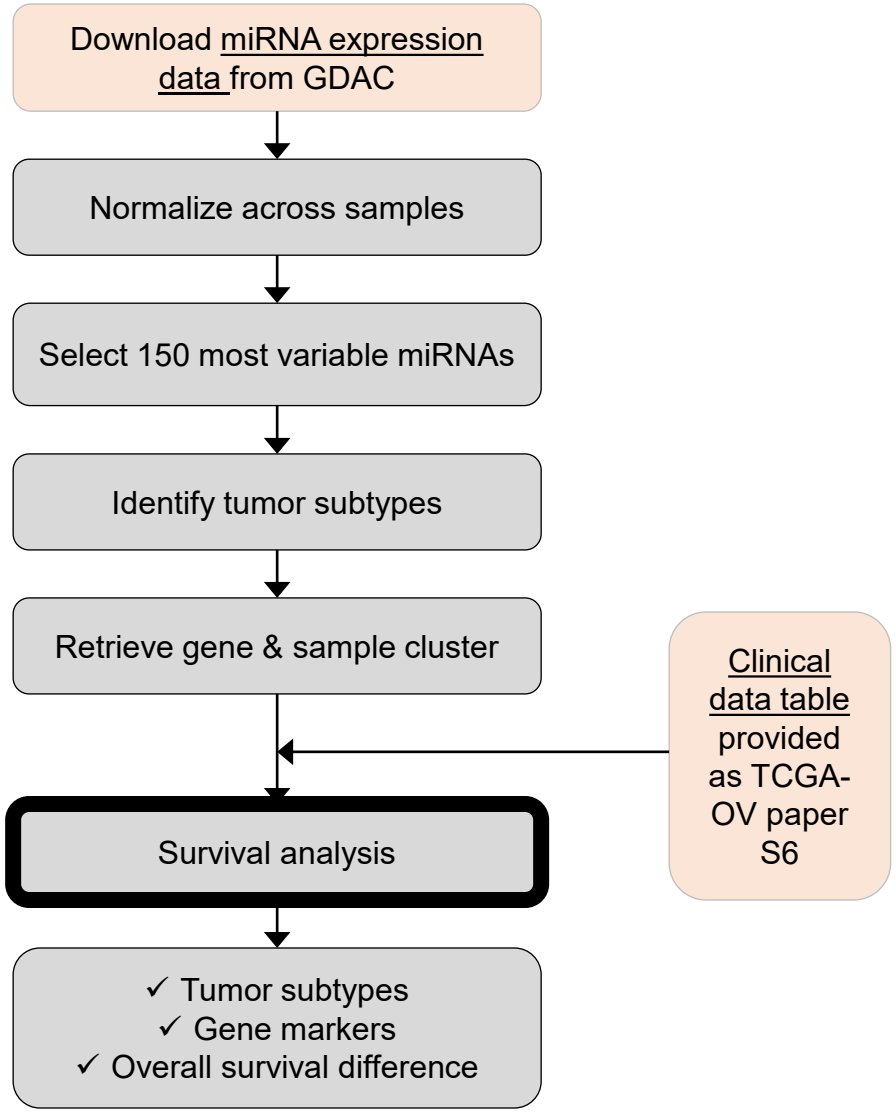
Integrate Expression with Clinical Data



- **Kaplan-Meier estimator**
  - **Stepwise function**
  - **Does not account for effect of other covariates (univariable test)**



Integrate Expression with Clinical Data



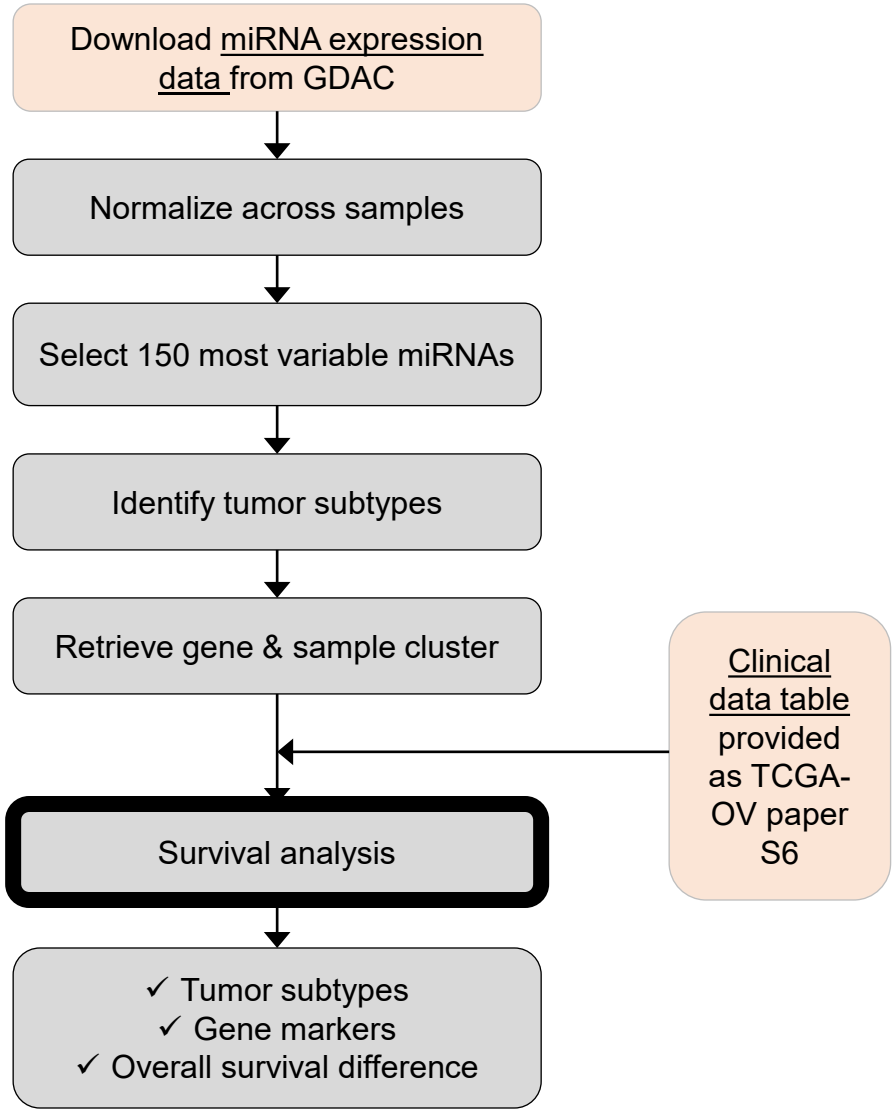
- Log-rank test
  - Chi-square test
  - Efficient in comparing groups differed by categorical variables, but not continuous ones
  - Univariable test

$$\mathrm{HR} = \frac{O_1/E_1}{O_2/E_2}$$

The hazard ratio (HR) is a measure of the relative survival experience in the two groups

O: Observed
E: Expected (if no difference between group 1 and 2)

Integrate Expression with Clinical Data



- Cox Proportional hazard model
  - Conveniently access the effect of continuous and categorical variables
  - Test the significance of factor of interest adjusting for other factors
  - Multivariable test!

```
S ~ sample.cluster +
    patient.age +
    tumor.grade
```

```
> summary(m3)
Call:
coxph(formula = surv ~ (clinical.sub$cluster + clinical.sub$age.at.diagnosis.year))

  n= 558, number of events= 292

                                      coef exp(coef) se(coef)     z Pr(>|z|)
clinical.sub$cluster2              0.451874  1.571254 0.145955 3.096 0.001962 **
clinical.sub$cluster3              0.309421  1.362636 0.138750 2.230 0.025744 *
clinical.sub$age.at.diagnosis.year 0.019574  1.019766 0.005354 3.656 0.000257 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Cox Proportional hazard model

- β: coefficient of explanatory variables or predictors
- exp(β): the ratio of the hazards between two individuals whose values of x differ by one unit when all other covariates are held constant (**hazard ratio,** analogous to an **odds ratio** in the setting of multiple logistic regression analysis)
- Z: Wald statistics calculated by dividing β by its standard error
- P: P-value that corresponds to Z statistics. If P<0.05, then the null hypothesis of β equal to zero can be rejected at 95% confidence level

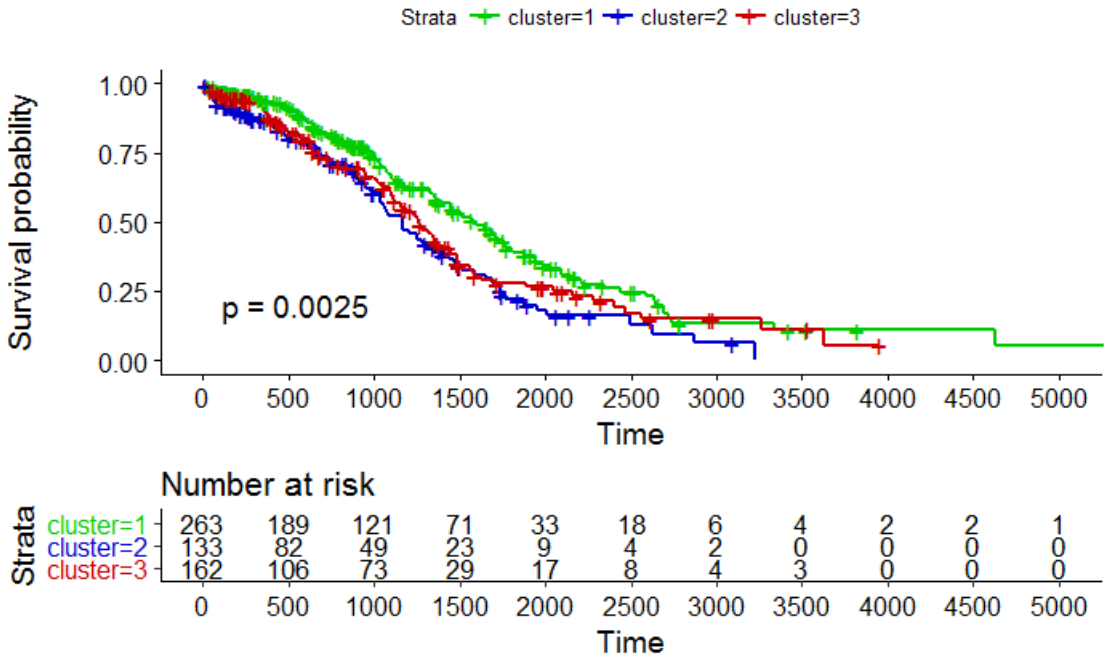
```
> summary(m3)
Call:
coxph(formula = surv ~ (clinical.sub$cluster + clinical.sub$age.at.diagnosis.year))

  n= 558, number of events= 292

                                    coef exp(coef) se(coef)     z Pr(>|z|)
clinical.sub$cluster2              0.451874  1.571254 0.145955 3.096 0.001962 **
clinical.sub$cluster3              0.309421  1.362636 0.138750 2.230 0.025744 *
clinical.sub$age.at.diagnosis.year 0.019574  1.019766 0.005354 3.656 0.000257 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Cox Proportional hazard model

- β: coefficient of explanatory variables or predictors
- exp(β): the ratio of the hazards between two indivi
  unit when all other covariates are held constant (**h**
  ratio in the setting of multiple logistic regression a
- Z: Wald statistics calculated by dividing β by its sta
- P: P-value that corresponds to Z statistics. If P<0.0
  to zero can be rejected at 95% confidence level



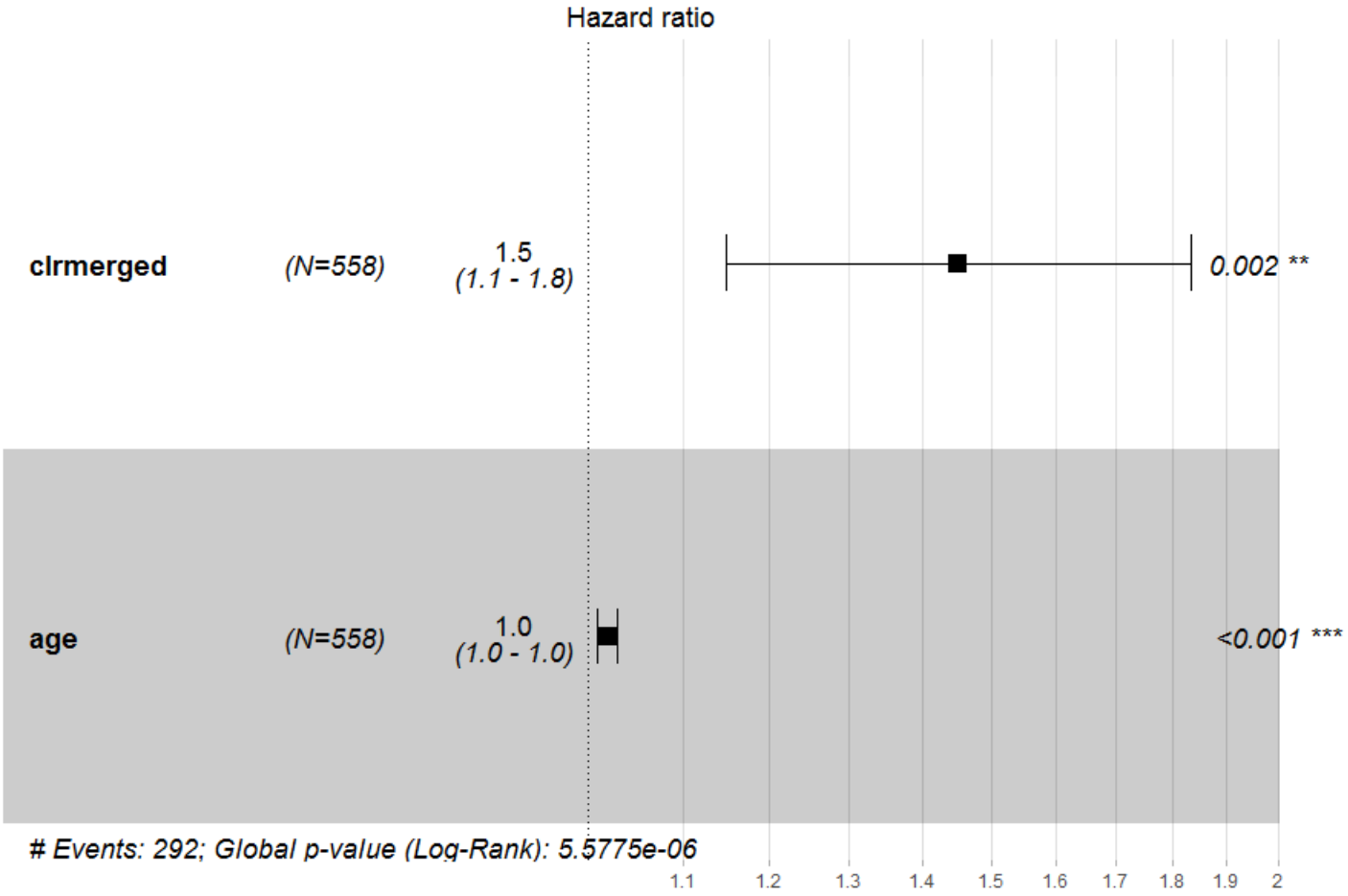

```
> summary(m3)
Call:
coxph(formula = surv ~ (clinical.sub$cluster + c

  n= 558, number of events= 292

                                    coef exp(coef) se(coef)     z Pr(>|z|)
clinical.sub$cluster2              0.451874  1.571254 0.145955 3.096 0.001962 **
clinical.sub$cluster3              0.309421  1.362636 0.138750 2.230 0.025744 *
clinical.sub$age.at.diagnosis.year 0.019574  1.019766 0.005354 3.656 0.000257 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Cox Proportional hazard model: visualize using forest plot

- β: coefficient of explanatory variables
- exp(β): the ratio of the hazards betwe unit when all other covariates are hel **ratio** in the setting of multiple logistic
- Z: Wald statistics calculated by dividi
- P: P-value that corresponds to Z stat to zero can be rejected at 95% confic

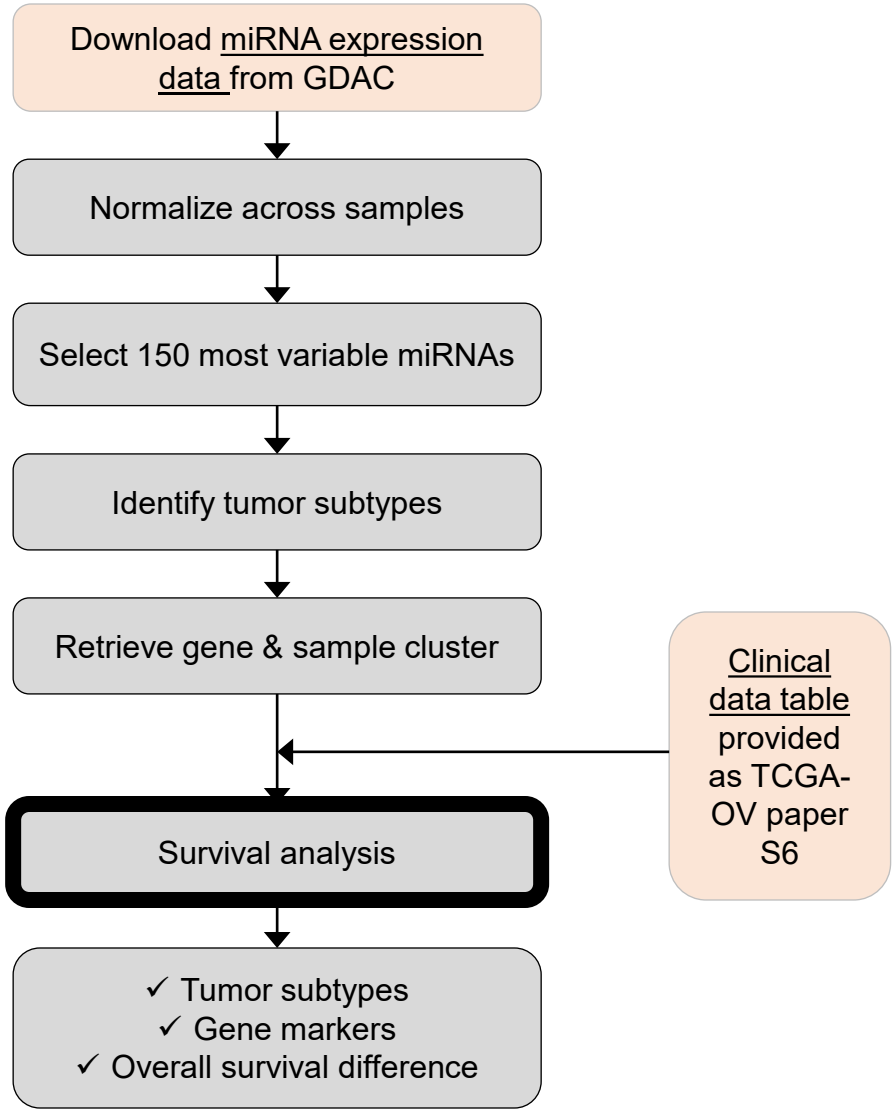



```
> summary(m3)
Call:
coxph(formula = surv ~ (clinica

  n= 558, number of events= 292

clinical.sub$cluster2
clinical.sub$cluster3
clinical.sub$age.at.diagnosis.year 0.019574  1.019766 0.005354 3.656 0.000257 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Integrate Expression with Clinical Data



- Survival methods
  - Kaplan-Meier estimator
  - Log-rank test (Mantel-Haenzel test)
  - Cox regression model (proportional hazard model)

Library(survival)

survfit

survdiff

coxph

`MSTP_summer2022.handson.Rmd`

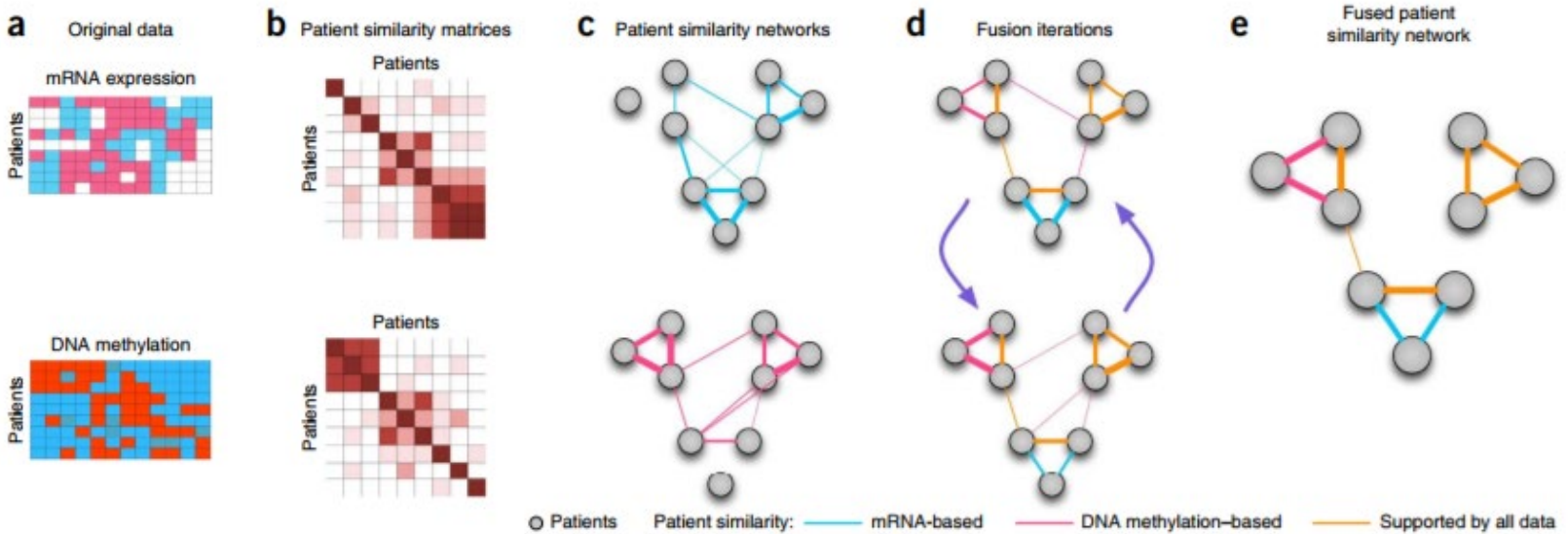
Integration of multi-omics data

- Multiple types of genomic data, e.g. mRNA, miRNA, methylation
- SNF: similarity network fusion
  http://compbio.cs.toronto.edu/SNF/SNF/Software.html
- iClusterPlus: a joint latent variable model for integrative clustering
  https://bioconductor.org/packages/release/bioc/html/iClusterPlus.html

Similarity network fusion for aggregating data types on a genomic scale

Bo Wang[1,5], Aziz M Mezlini[1,2], Feyyaz Demir[1,2], Marc Fiume[2], Zhuowen Tu[3], Michael Brudno[1,2], Benjamin Haibe-Kains[4,5] & Anna Goldenberg[1,2]

Recent technologies have made it cost-effective to collect diverse types of genome-wide data. Computational methods are needed to combine these data to create a comprehensive view of a given disease or a biological process. Similarity network fusion (SNF) solves this problem by constructing networks of samples (e.g., patients) for each available data type and then efficiently fusing these into one network that represents the full spectrum of underlying data. For example, to create a comprehensive view of a disease given a cohort of patients, SNF computes and fuses patient similarity networks obtained from each of their data types separately, taking advantage of the complementarity in the data. We used SNF to combine mRNA expression, DNA methylation and microRNA (miRNA) expression data for five cancer data sets. SNF substantially outperforms single data type analysis and established integrative approaches when identifying cancer subtypes and is effective for predicting survival.



Thank you!

Questions

Hands on practice!