Translational Bioinformatics Workshop in R @MSU

Aug 10-14, 2020

Objectives:

1) Be comfortable to use R and open big data in research
2) Get to know essential resources you can refer to
3) Get to know someone who may help your future research

- Date: Aug 10-14, 2020
- Cost: $100 per person
- 1 session: 45min tutorial and then 15min break
- five-day workshop, primarily in R, occasionally in Python
- Morning (9-12pm): tutorial (required), Thursday starts at 8:00am
- Afternoon (1pm - 3pm): lab & QA (optional)

Workshop structure

- Introduction to big data and R (Dr. Bin Chen, PI in Chen lab)
- Data manipulation and visualization (Dr. Eugene Chekalin, postdoc in Chen lab)
- Basic statistical analysis (Dr. Yuehua Cui, Professor in statistics)
- Machine learning (Dr. Jiayu Zhou, Assistant Professor in computer science)
- RNA-Seq (Dr. Ke Liu, Postdoc in Chen lab)
- Single cell RNA-Seq (Dr. Eric Kort, Scientist at VARI and Assistant Professor at MSU)
- From big data to drug discovery (Dr. Bin Chen, PI in Chen lab)
- Structure-based drug discovery (Dr. Jing Xing, Postdoc in Chen lab)
- R markdown/R package/Shiny (Paul Egeler, Senior Data Engineer at Spectrum Health)

MICHIGAN STATE UNIVERSITY

Introduction to Big Data and R

Bin Chen

Assistant Professor
Dept. of Pediatrics and Human Development
Dept. of Pharmacology and Toxicology
College of Human Medicine
Michigan State University
Bin.Chen@hc.msu.edu @DrBinChen
http://binchenlab.org

Aug 2020

Myself

Trained as a chemist in college

Worked as a software engineer for three years

PhD in informatics (cheminformatics/bioinformatics)

Postdoc training in translational bioinformatics

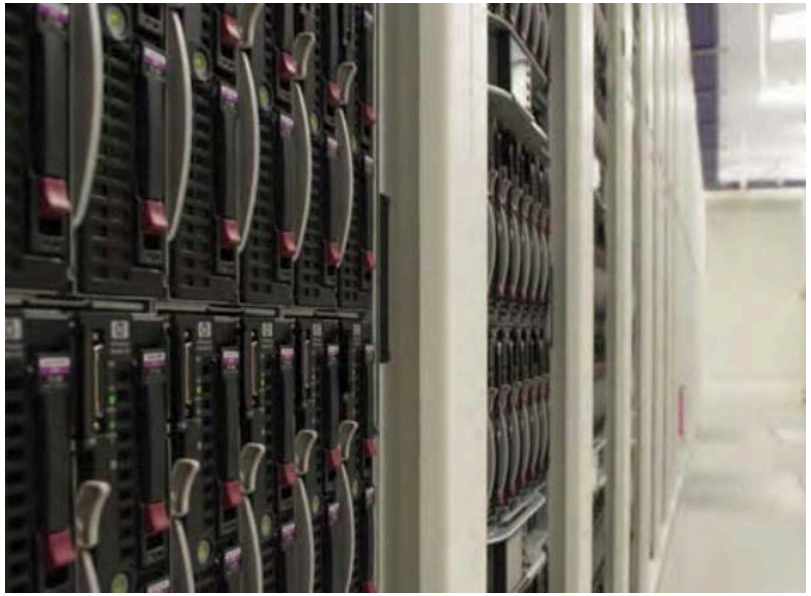
Work experience at Novartis, Pfizer and Merck

PI at UCSF/MSU

Chen lab mission: Leverage Big Data and AI to Discover New Therapeutics

Session 1: Big Data in translational bioinformatics
Session 2: Big Data in R

EMBL-EBI

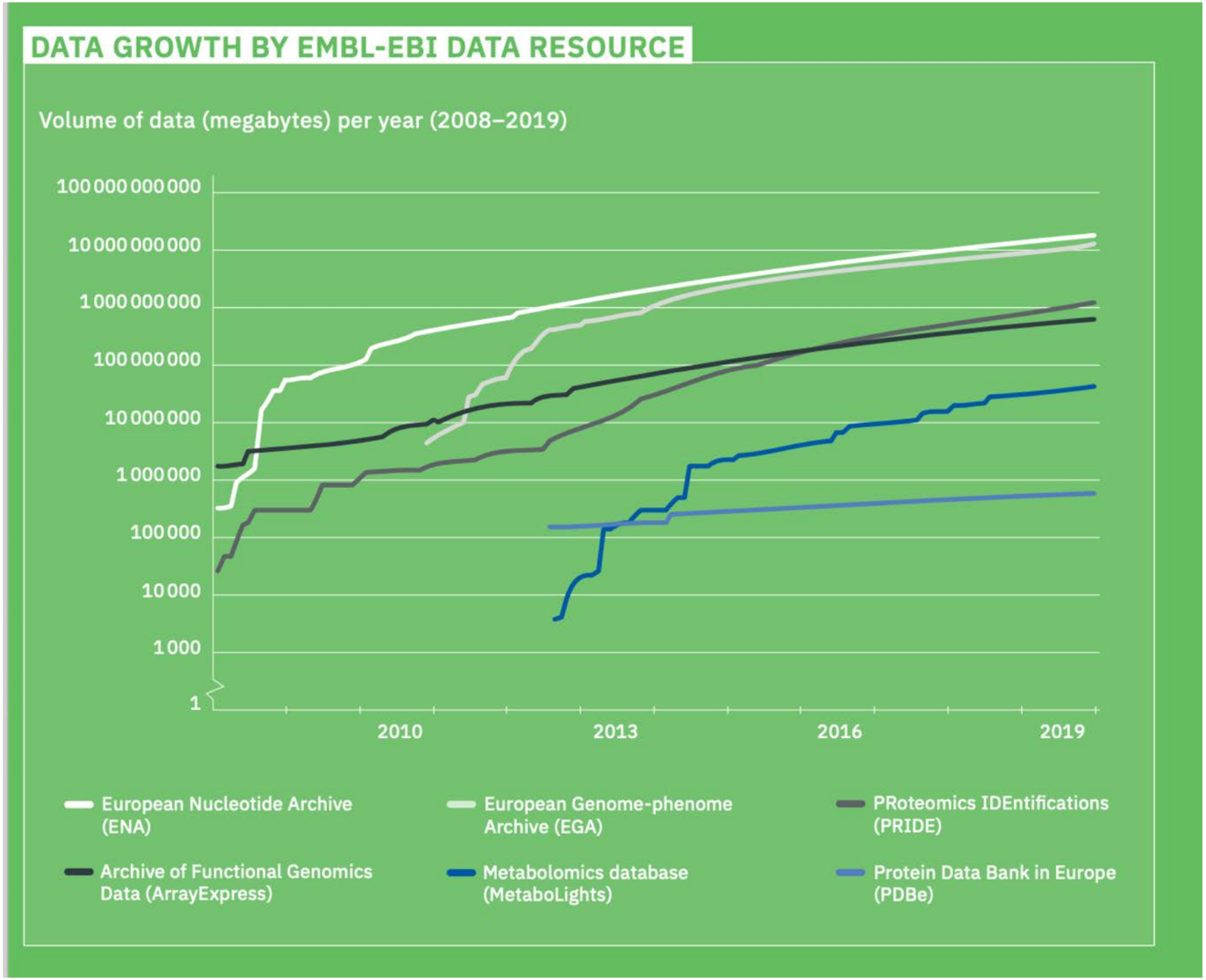


25 petabytes 2014

307 petabytes 2019

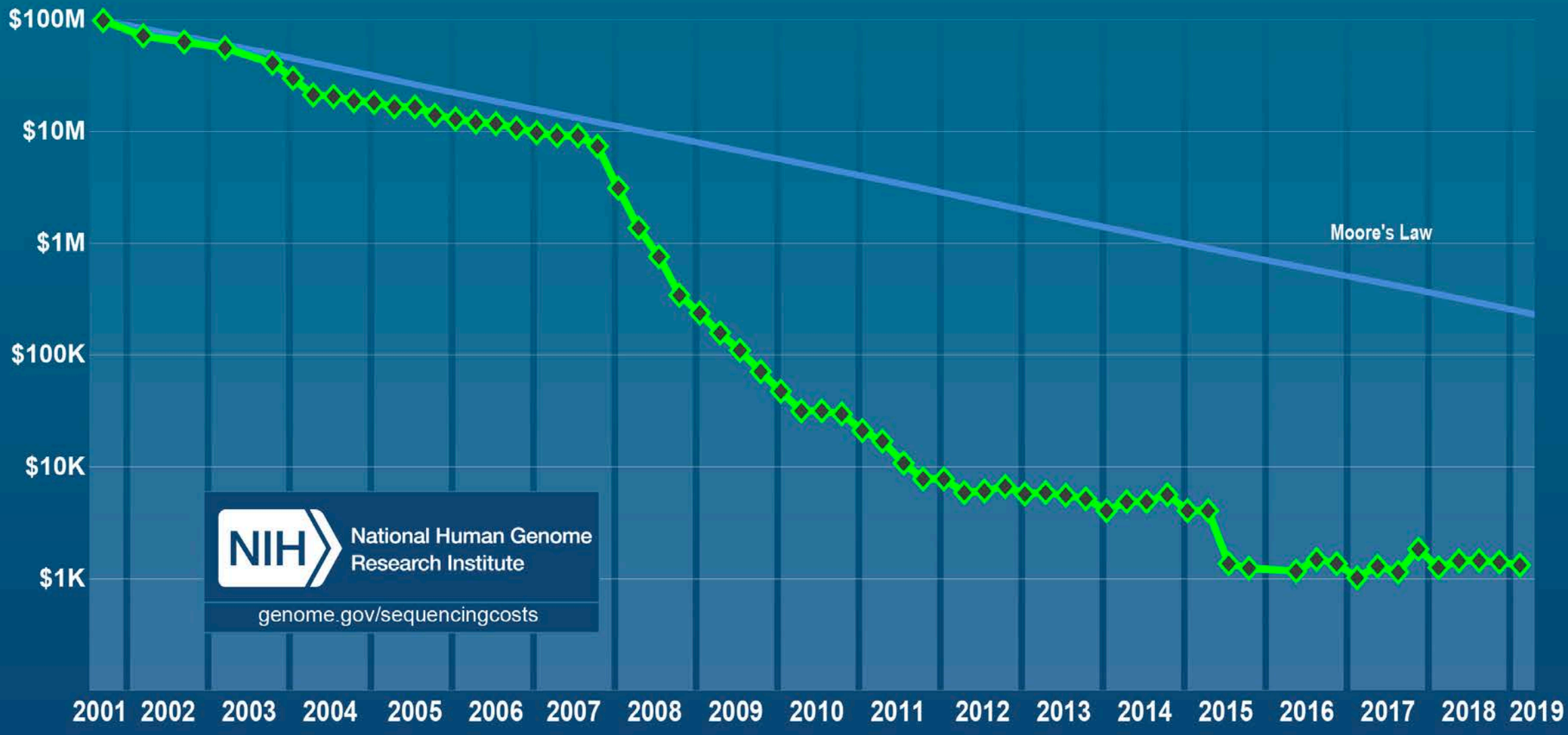
http://bit.ly/1OyTuqZ
http://bit.ly/2o3QJdy
shorturl.at/tyFGV

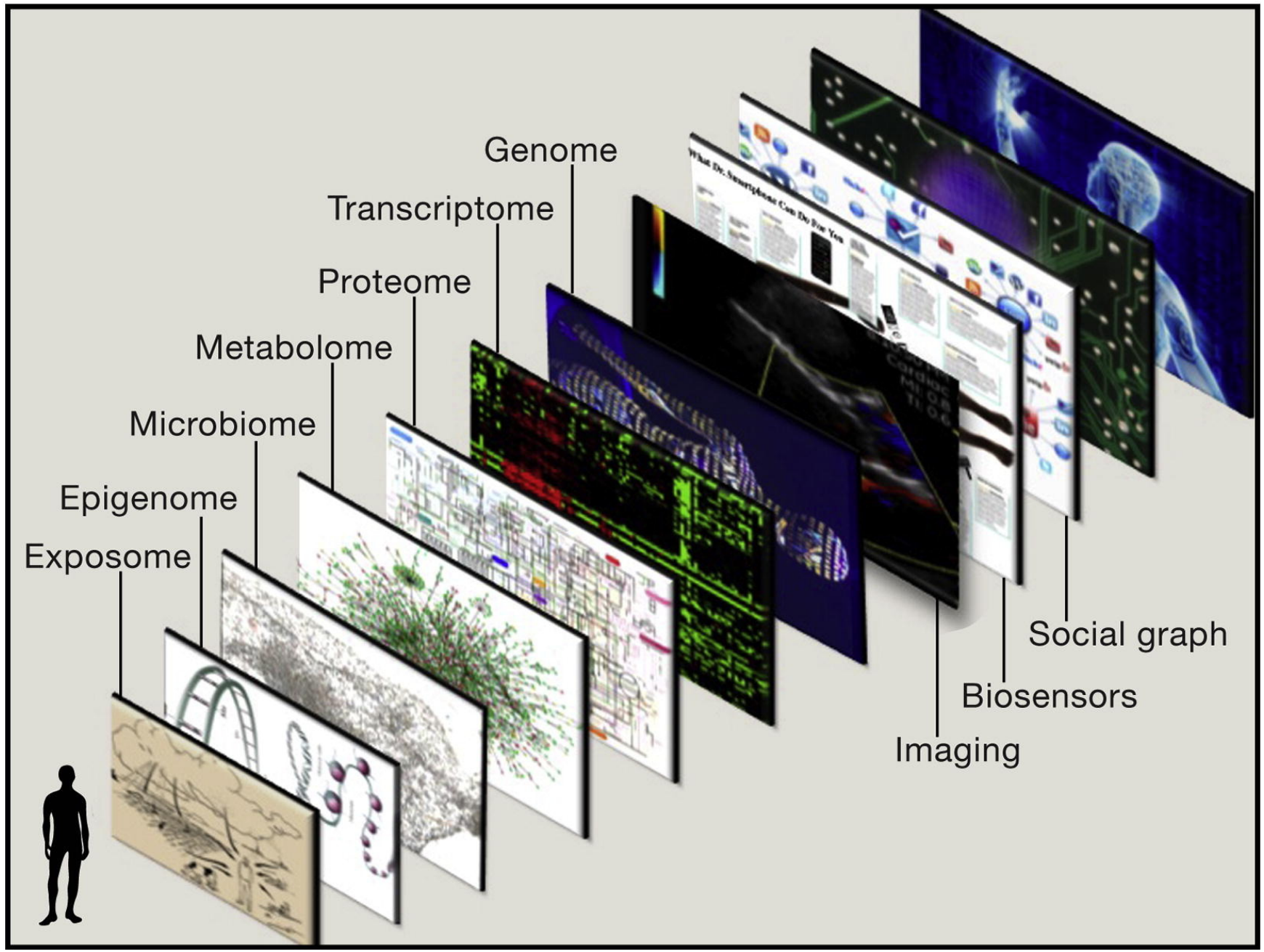
307 petabytes=  × 150,000



shorturl.at/tyFGV

Cost per Genome





Eric Topol, Cell, 2014

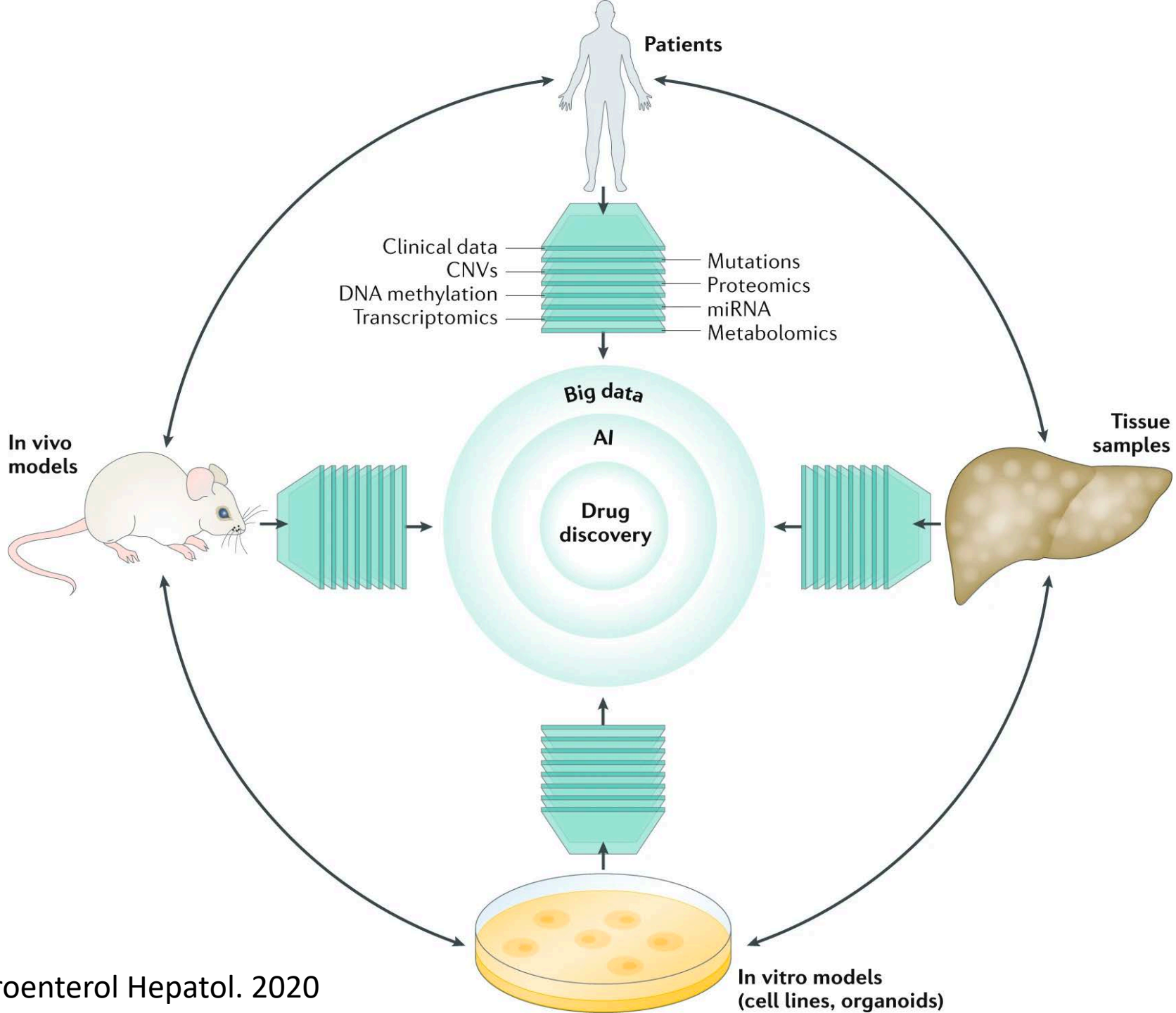
Scale (unit)	Imaging technology	Use
Molecular (angstrom)	Single-particle cryo-electron microscopy (EM) and electron tomography averaging	Structural analysis, molecular function
Molecular machines (nanometer)	Cryo-EM, super-resolution light microscopy (SRM)	Biochemistry, molecular mechanisms
Cells (micrometer)	Transmission EM, volume EM, light microscopy (wide-field, confocal, SRM), electron tomography, 3D scanning EM, soft X-ray tomography	Cellular morphology, activity within cells, mechanism
Tissues (millimeter)	Volume EM, scanning EM, light microscopy (multiphoton, light sheet, OPT, etc.), X-rays (micro-CT), fluorescence imaging, mass spectrometry imaging	Protein localization, tissue morphology and anatomy, interactions between cells
Organism/organ (centimeter)	Photography, X-rays, magnetic resonance imaging, optical tomography technologies, computerized tomography, luminescence imaging	Mechanistic understanding of development and disease

Jan Ellenberg, Nature Methods, 2018

What is Translational Bioinformatics

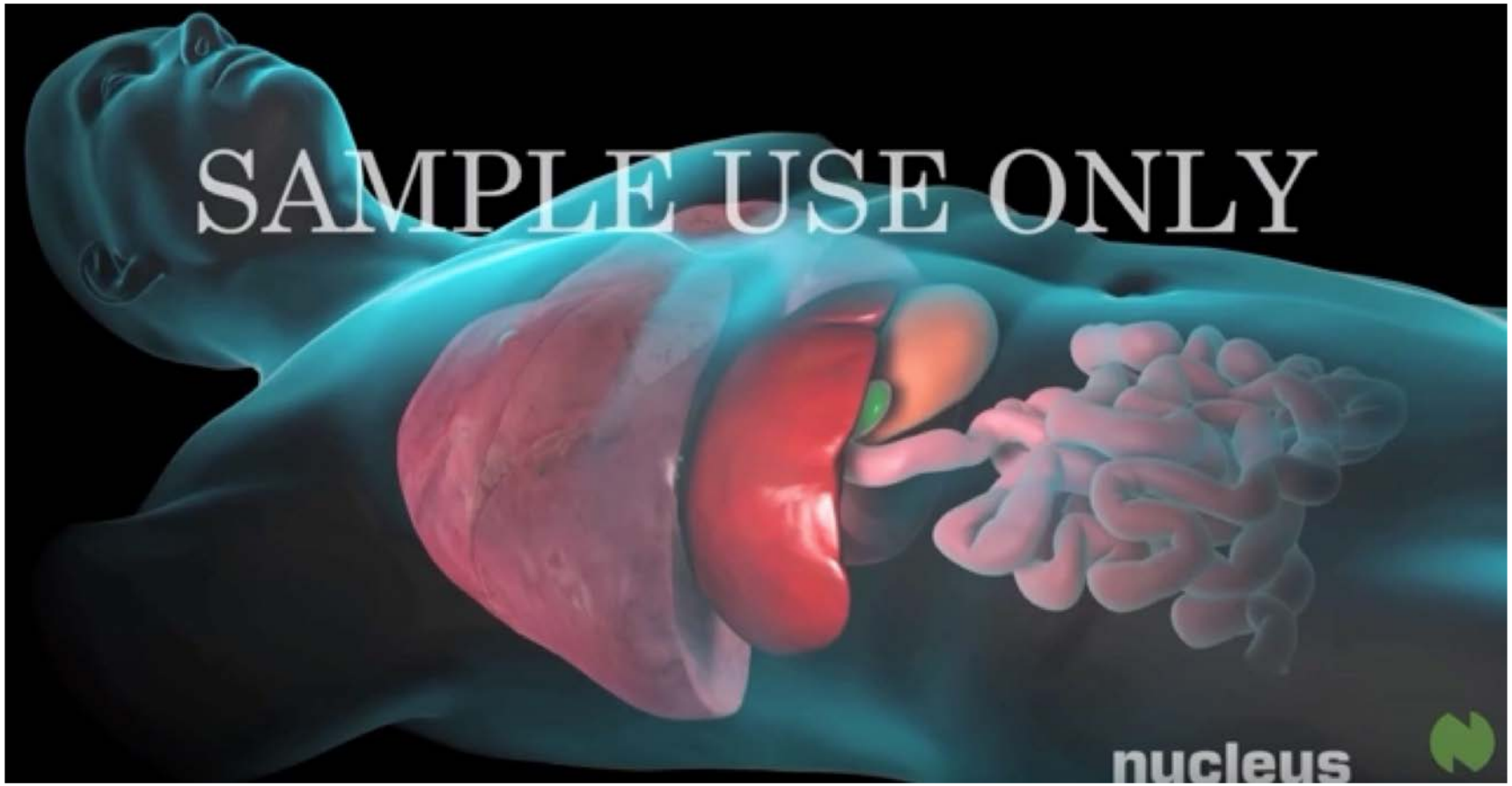
Translational Bioinformatics is the development of storage, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, and genomic data, into proactive, predictive, preventive, and participatory health.

https://www.amia.org/



Chen B, Nat Rev Gastroenterol Hepatol. 2020

Tissue



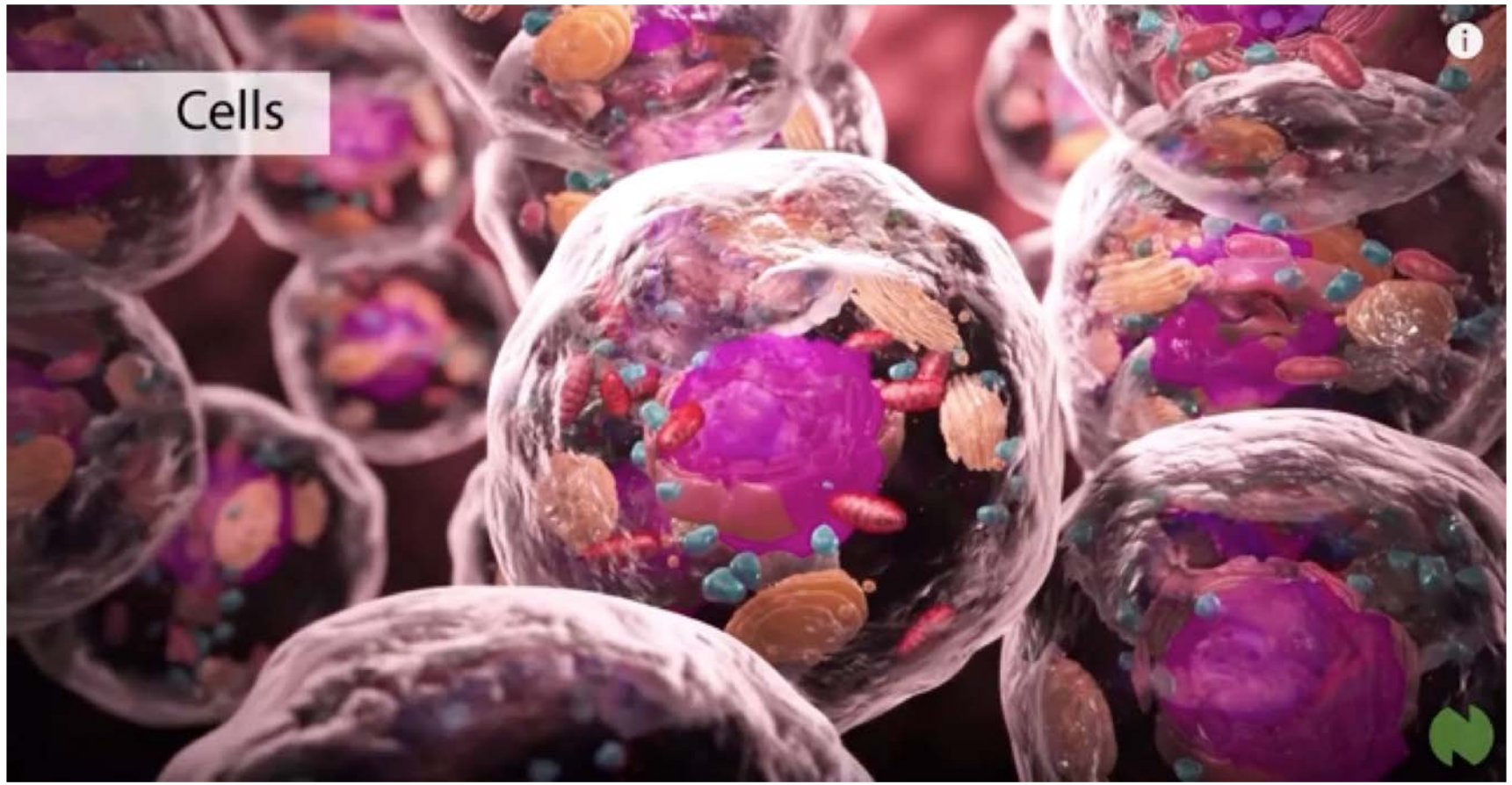
https://www.youtube.com/watch?v=mnHPx5XEvfQ

Cell line

https://www.youtube.com/watch?v=CMRKKl9XSDU

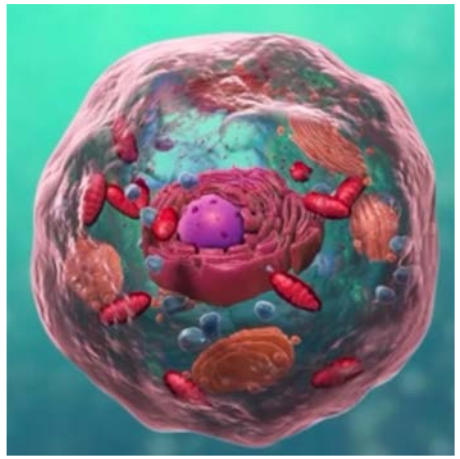
Animal Model

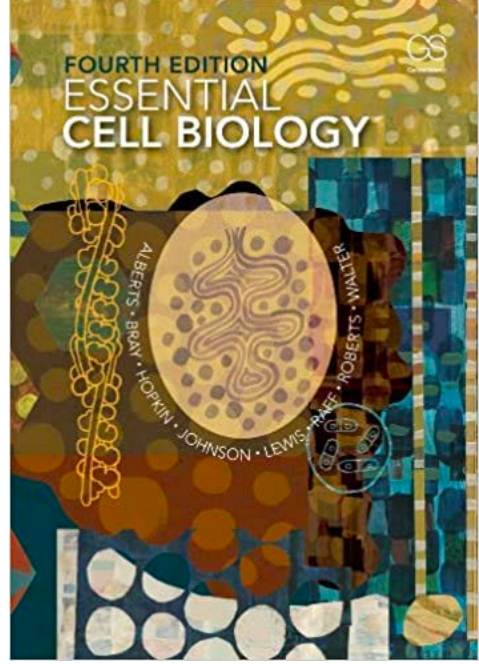
https://www.youtube.com/watch?v=CK78IXTRH0s

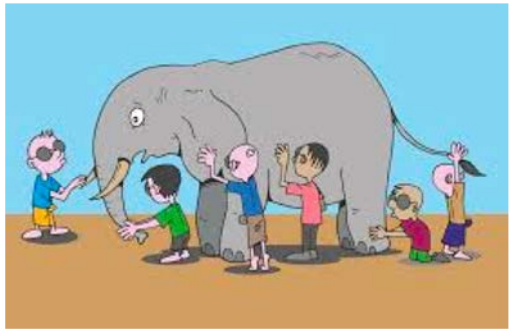


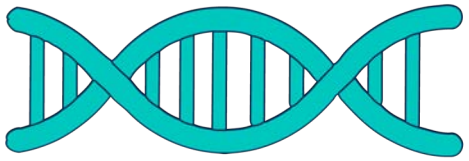
https://www.youtube.com/watch?v=URUJD5NEXC8

- Ignore spatial info
- Ignore dynamic
- Ignore cell-cell variation

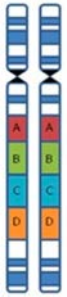








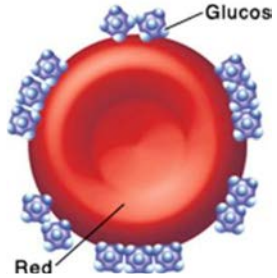
DNA



Chromosome
(copy number)

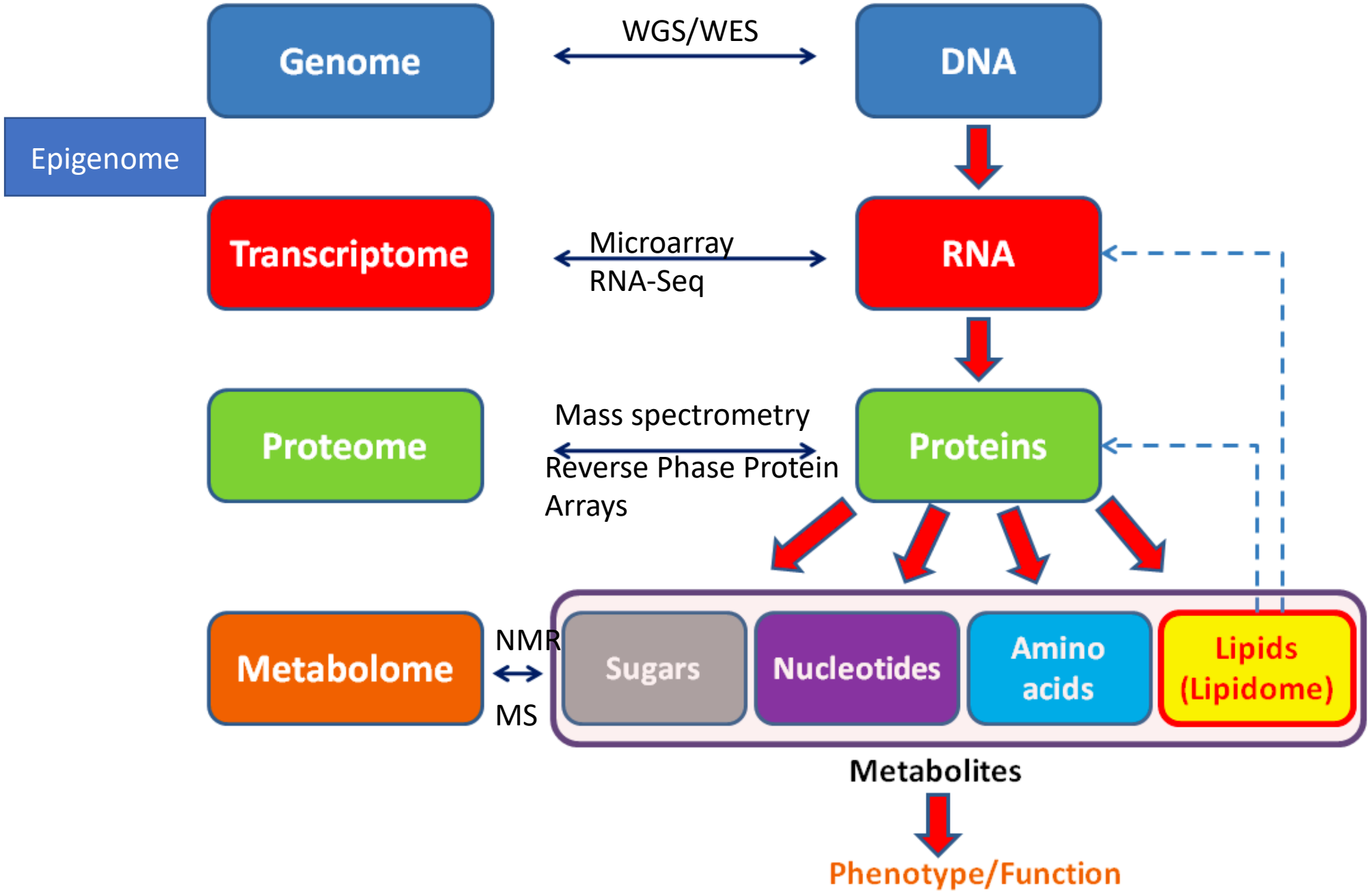
mRNA (gene)

Protein

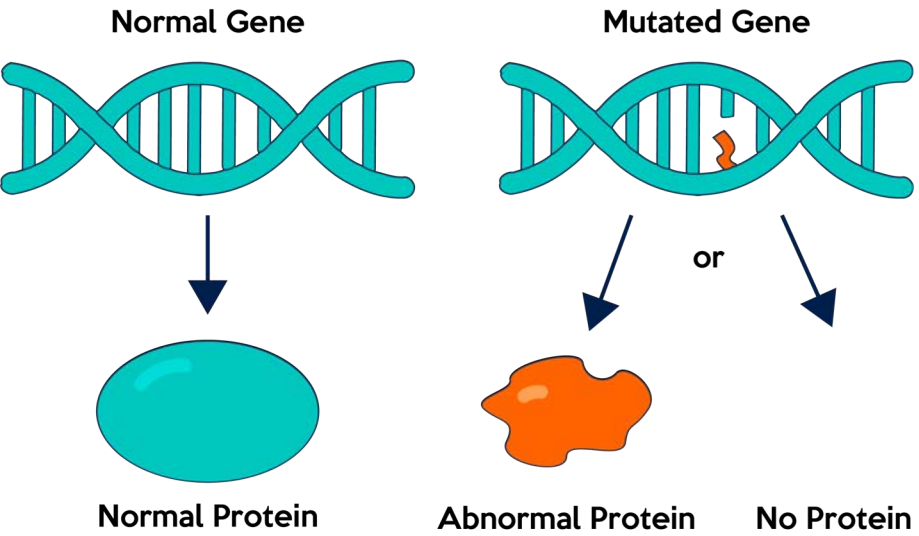


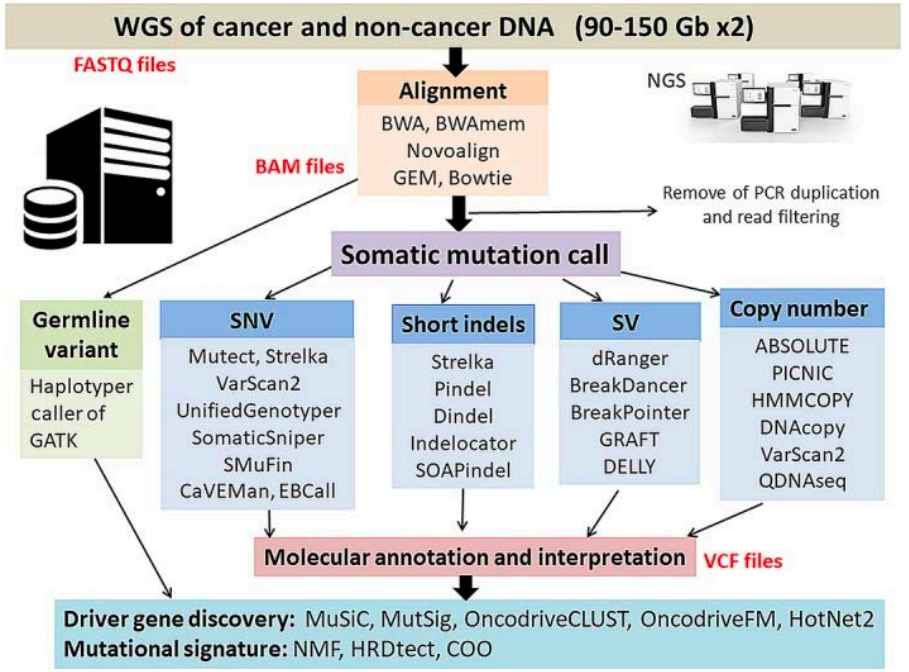
metabolite

. . . . . .



https://en.wikipedia.org/wiki/Genomics





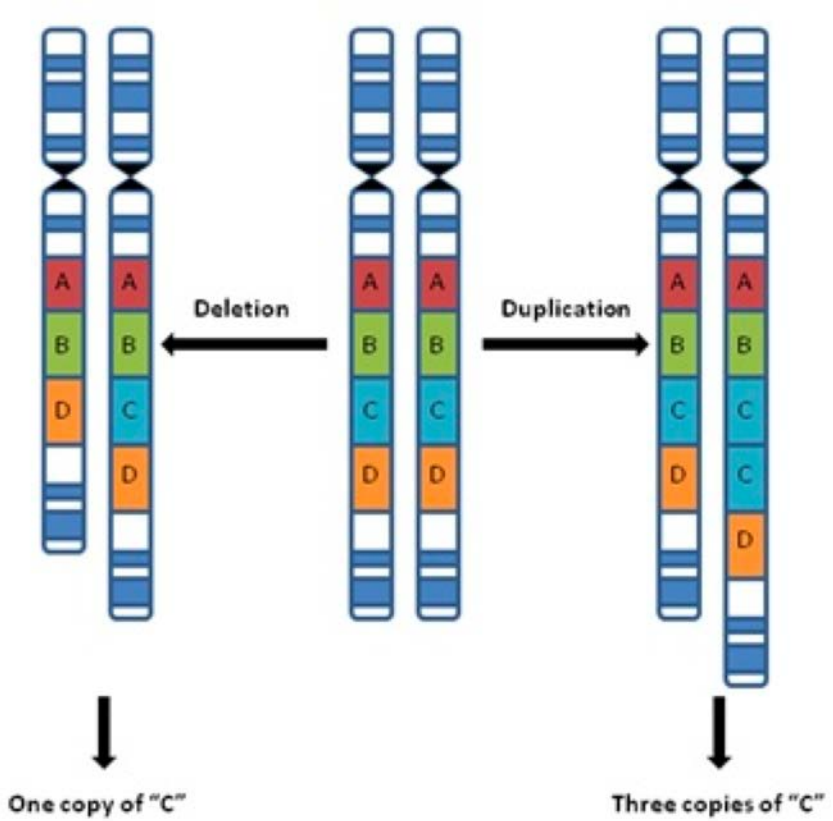
Genome/mutation

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Hugo_Symb	Entrez_Gene	NCBI_Build	Chromosom	Start_positi	End_positio	Strand	Variant_Classifi	Variant_Typ	Reference_A	Tumor_Seq_	dbSNP_RS	dbSNP_Val_	Genome_Change			cDNA_Change	Codon_Change	Protein_Change
VPS13D	55187	37	1	12359347	12359347	+	Nonsense_Mut	SNP	C	A	NA	NA	g.chr1:12359347C>A		A	c.6122C>A	c.(6121-6123)tCa>tAa	p.S2041*
AADACL4	343066	37	1	12726308	12726322	+	In_Frame_Del	DEL	CTGGCGTG	-	rs5821842	byFrequenc	g.chr1:12726308_12726		A	c.786_800delCTGGCGTG	c.(784-801)tcctggcgtgac	p.WRDAI263del
IFNLR1	163702	37	1	24484172	24484172	+	Silent	SNP	G	A			g.chr1:24484172G>A		A	c.1011C>T	c.(1009-1011)ggC>ggT	p.G337G
TMEM57	55219	37	1	25785018	25785019	+	Frame_Shift_Ins	INS	-	A			g.chr1:25785018_25785		A	c.789_790insA	c.(790-792)aaafs	p.K264fs
ZSCAN20	7579	37	1	33954141	33954141	+	Missense_Mut	SNP	T	G	NA	NA	g.chr1:33954141T>G		A	c.494T>G	c.(493-495)gTg>gGg	p.V165G
POU3F1	5453	37	1	38512139	38512139	+	Missense_Mut	SNP	C	G	NA	NA	g.chr1:38512139C>G		A	c.277G>C	c.(277-279)Ggc>Cgc	p.G93R
MAST2	23139	37	1	46498028	46498028	+	Silent	SNP	C	T			g.chr1:46498028C>T		A	c.3366C>T	c.(3364-3366)cgC>cgT	p.R1122R
GBP4	115361	37	1	89657103	89657103	+	Silent	SNP	G	T	rs146643676		g.chr1:89657103G>T		A	c.757C>A	c.(757-759)Cgg>Agg	p.R253R
VAV3	10451	37	1	1.08E+08	1.08E+08	+	Splice_Site	SNP	A	G			g.chr1:108247170A>G			ACH-000001	c.e17+1	
NBPF20	1E+08	37	1	1.48E+08	1.48E+08	+	Missense_Mut	SNP	T	G	rs3692234	NA	g.chr1:148346689T>G		A	c.68A>C	c.(67-69)aAa>aCa	p.K23T
FLG2	388698	37	1	1.52E+08	1.52E+08	+	Missense_Mut	SNP	C	T			g.chr1:152331291C>T		A	c.70G>A	c.(70-72)Gag>Aag	p.E24K
RHBG	57127	37	1	1.56E+08	1.56E+08	+	Frame_Shift_De	DEL	C	-	NA	NA	g.chr1:156339132delC		A	c.92delC	c.(91-93)gctfs	p.A31fs
FCRL5	83416	37	1	1.57E+08	1.57E+08	+	Missense_Mut	SNP	A	T			g.chr1:157497658A>T		A	c.1709T>A	c.(1708-1710)cTc>cAc	p.L570H
SPTA1	6708	37	1	1.59E+08	1.59E+08	+	Missense_Mut	SNP	C	A	NA	NA	g.chr1:158639228C>A		A	c.1803G>T	c.(1801-1803)aaG>aaT	p.K601N
C1orf192	257177	37	1	1.61E+08	1.61E+08	+	Missense_Mut	SNP	A	G			g.chr1:161334861A>G		A	c.428T>C	c.(427-429)aTa>aCa	p.I143T
DUSP12	11266	37	1	1.62E+08	1.62E+08	+	Missense_Mut	SNP	T	C			g.chr1:161719830T>C		A	c.239T>C	c.(238-240)gTg>gCg	p.V80A
QSOX1	5768	37	1	1.8E+08	1.8E+08	+	Silent	SNP	C	T			g.chr1:180153105C>T		A	c.807C>T	c.(805-807)ctC>ctT	p.L269L
EPRS	2058	37	1	2.2E+08	2.2E+08	+	Silent	SNP	G	A			g.chr1:220156537G>A		A	c.3294C>T	c.(3292-3294)gcC>gcT	p.A1098A
RGS7	6000	37	1	2.41E+08	2.41E+08	+	Missense_Mut	SNP	G	C	NA	NA	g.chr1:241031950G>C		A	c.546C>G	c.(544-546)gaC>gaG	p.D182E
ZBTB18	10472	37	1	2.44E+08	2.44E+08	+	Missense_Mut	SNP	A	G			g.chr1:244217482A>G		A	c.406A>G	c.(406-408)Agc>Ggc	p.S136G
PCBD1	5092	37	10	72645567	72645567	+	Silent	SNP	T	C			g.chr10:72645567T>C		A	c.123A>G	c.(121-123)aaA>aaG	p.K41K
ECD	11319	37	10	74899185	74899186	+	Frame_Shift_Ins	INS	-	T	NA	NA	g.chr10:74899185_7489		A	c.1302_1303insA	c.(1300-1305)aaagaafs	p.E435fs
ZNF503	84858	37	10	77160064	77160064	+	Silent	SNP	G	A	NA	NA	g.chr10:77160064G>A		A	c.384C>T	c.(382-384)ccC>ccT	p.P128P
PDZD7	79955	37	10	1.03E+08	1.03E+08	+	Missense_Mut	SNP	C	T			g.chr10:102781635C>T		A	c.787G>A	c.(787-789)Ggt>Agt	p.G263S
SORCS1	114815	37	10	1.08E+08	1.08E+08	+	Missense_Mut	SNP	A	C	NA	NA	g.chr10:108434902A>C		A	c.1845T>G	c.(1843-1845)gaT>gaG	p.D615E
GFRA1	2674	37	10	1.18E+08	1.18E+08	+	Missense_Mut	SNP	C	G	rs200978034		g.chr10:118030598C>G		A	c.70G>C	c.(70-72)Gga>Cga	p.G24R
CPXM2	119587	37	10	1.26E+08	1.26E+08	+	Missense_Mut	SNP	T	C	rs561902091		g.chr10:125516817T>C		A	c.1829A>G	c.(1828-1830)tAc>tGc	p.Y610C
AL355531.2	0	37	10	1.31E+08	1.31E+08	+	Silent	SNP	C	T	rs5777868	byFrequenc	g.chr10:131309064C>T		A	c.462G>A	c.(460-462)agG>agA	p.R154R
TCERG1L	256536	37	10	1.33E+08	1.33E+08	+	Missense_Mut	SNP	G	A	NA	NA	g.chr10:132915133G>A		A	c.1324C>T	c.(1324-1326)Ccg>Tcg	p.P442S
MUC5B	727897	37	11	1271274	1271274	+	Silent	SNP	G	C	rs3709080	NA	g.chr11:1271274G>C		A	c.13164G>C	c.(13162-13164)acG>aa	p.T4388T

1656 cell lines

19K genes

Mutation status
(1 or 0)



1657 cell lines

28K genes

Copy number
(-2, -1, 0, 1, 2)

cBioPortal FOR CANCER GENOMICS

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

Modify Query

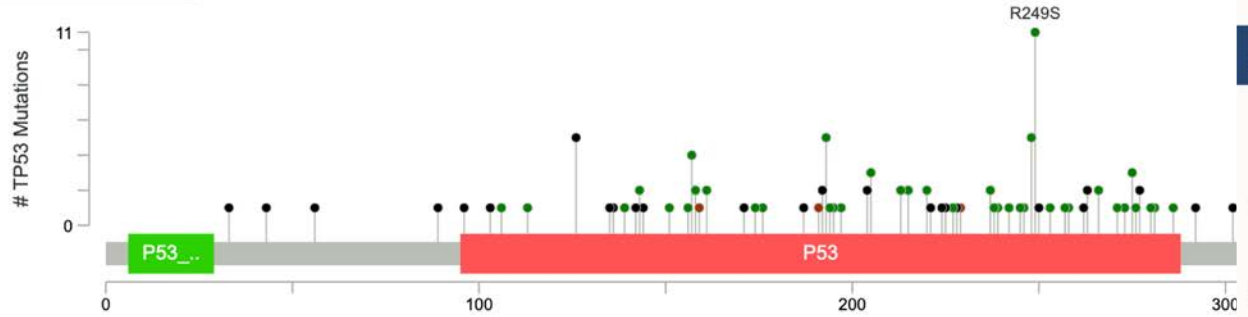
Liver Hepatocellular Carcinoma (TCGA, Firehose Legacy)
Samples with mutation and CNA data (366 patients/samples) - TP53

Queried gene is

OncoPrint | Cancer Types Summary | Plots | Mutations | Co-expression | Comparison/Survival | CN Segments | Pathways | Download

TP53

Add annotation tracks | Y-Axis Max: 11



COSMIC
Catalogue Of Somatic Mutations In Cancer

Projects | Data | Tools | News | Help | About | Genome Version

COSMIC v91, released 07-APR-20

COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer

earching for a gene, cancer type, mutation, etc. below.

inoma, V600E, BRCA-UK, Campbell SEARCH

ClinVar Genomic variation as it relates to human health

Search ClinVar

Advanced search

About | Access | Submit | Stats | FTP | Help

Was this helpful?

Follow | Print | Downloa

NM_004958.4(MTOR):c.7500T>G (p.Ile2500Met)

Cite this record

Interpretation:	**Likely pathogenic**
Review status:	★☆☆☆ criteria provided, single submitter
Submissions:	5 (Most recent: Jan 21, 2020)
Last evaluated:	Apr 11, 2019
Accession:	VCV000376455.2
Variation ID:	376455
Description:	single nucleotide variant

SRA

SRA

Advanced

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/

Full

Send to:

Design: HiSeq X Ten 150bp paired end sequencing of sample 90888_N5. Library was created with index CGCTCATT.

Submitted by: Cancer Research UK Cambridge Institute

Study: Liver Cancer Evolution - Lesion segregation
PRJEB37808 • ERP121138 • All experiments • All runs
show Abstract

Library:

European Genome-phenome Archive

All

Examples: EGAS00000000001, Cancer

Search

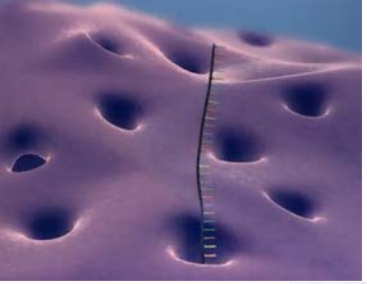
EGA home | About | Studies | Datasets | Data access committees | Data providers | Submit to EGA | Contact Us | Login

The https://ega-archive.org/ site adds additional functionality not found on this site.

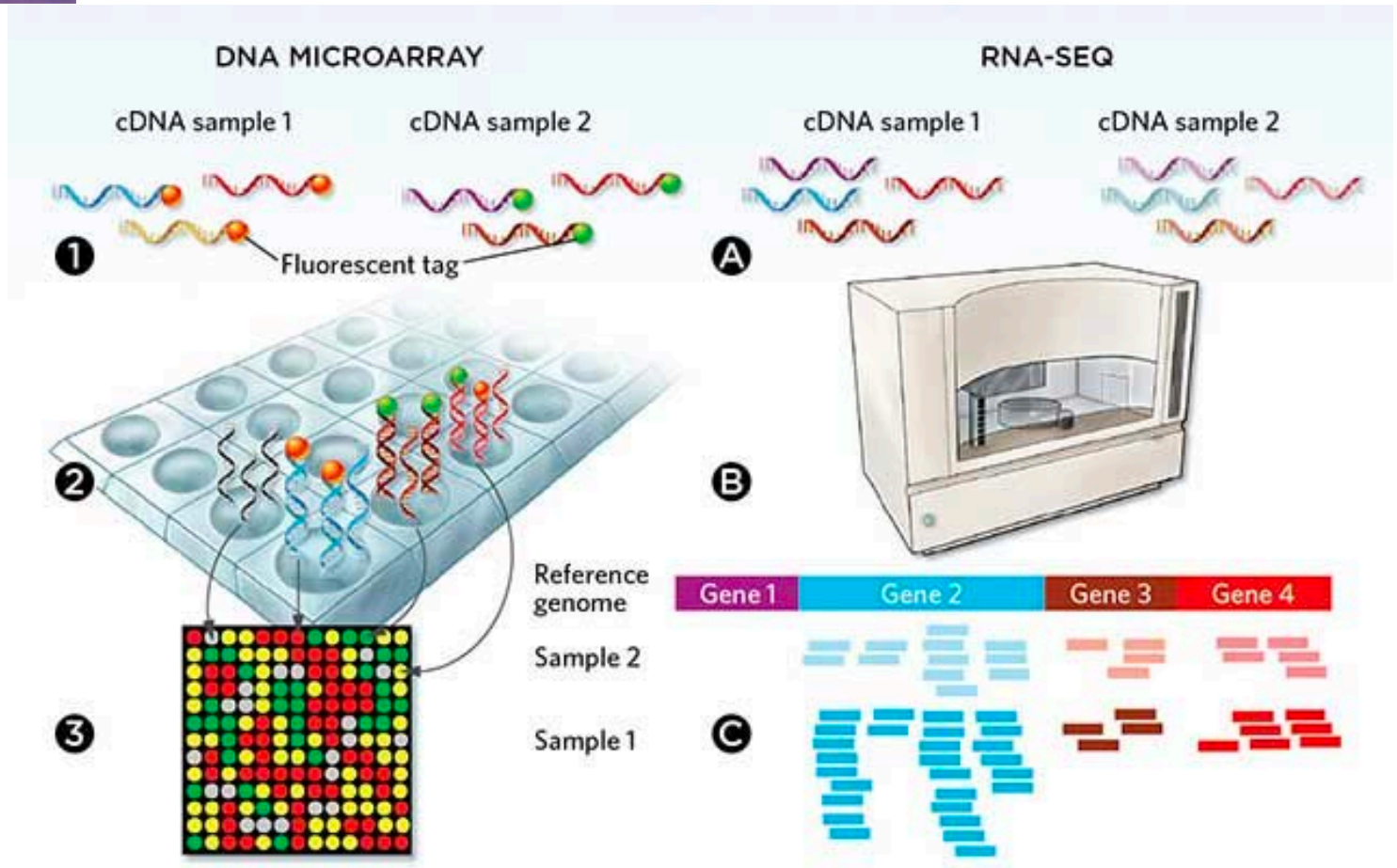
For queries or feedback related to the new site please contact ega-helpdesk@ega-archive.org

Help

- Users FAQ
- Submitters FAQ
- Using your EGA account
- Contact Us
- EGA mailing list



Transcriptome

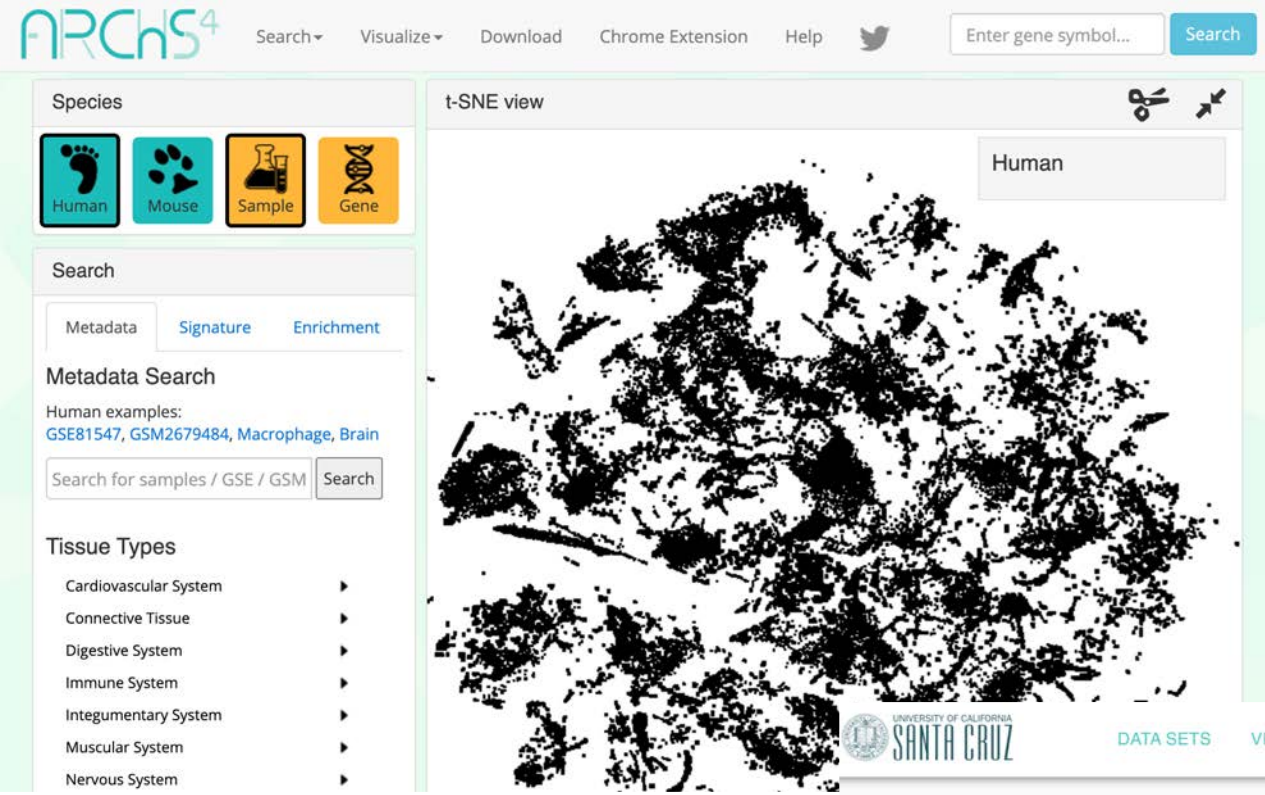


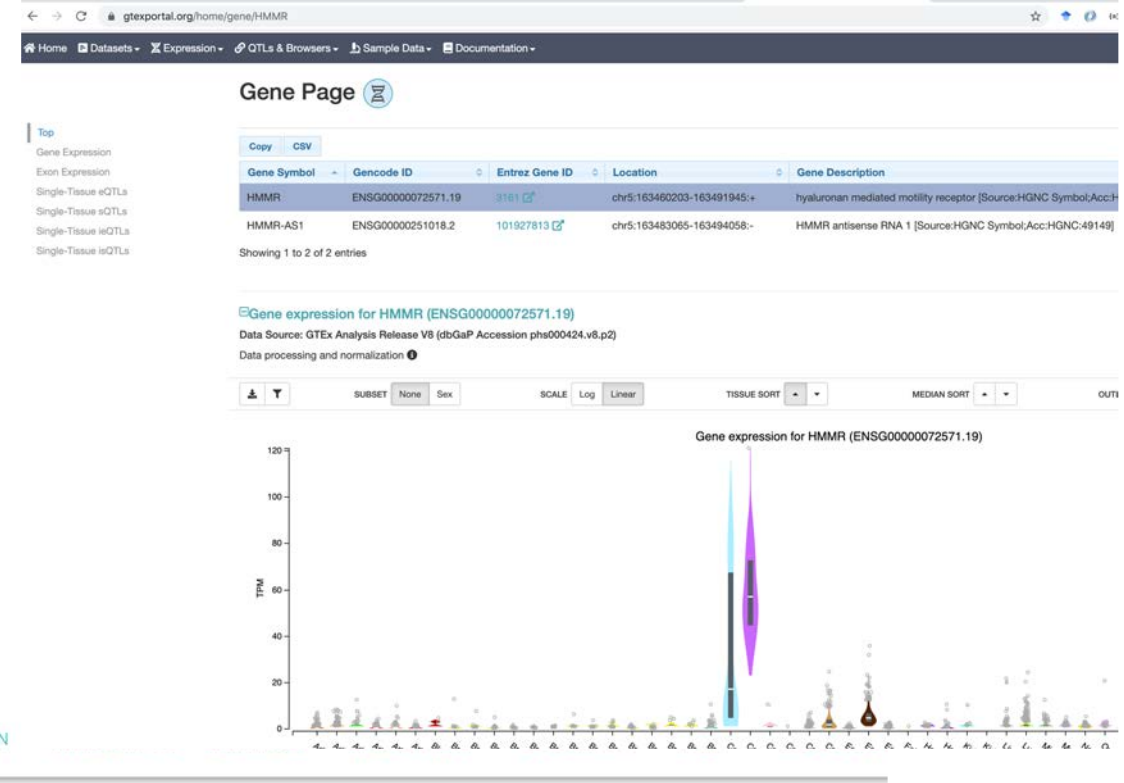
1210 cell lines

60K genes

Gene Expression (numeric)

https://www.the-scientist.com/lab-tools/an-array-of-options-35381


ARCHS4
Search
Visualize
Download
Chrome Extension
Help
Enter gene symbol...
Search
Species
Human
Mouse
Sample
Gene
Search
Metadata
Signature
Enrichment
Metadata Search
Human examples:
GSE81547, GSM2679484, Macrophage, Brain
Search for samples / GSE / GSM
Search
Tissue Types
Cardiovascular System
Connective Tissue
Digestive System
Immune System
Integumentary System
Muscular System
Nervous System
t-SNE view
Human


gtexportal.org/home/gene/HMMR
Home
Datasets
Expression
QTLs & Browsers
Sample Data
Documentation
Gene Page
Top
Gene Expression
Exon Expression
Single-Tissue eQTLs
Single-Tissue sQTLs
Single-Tissue ieQTLs
Single-Tissue isQTLs
Copy
CSV
Gene Symbol
Gencode ID
Entrez Gene ID
Location
Gene Description
HMMR
ENSG00000072571.19
3161
chr5:163460203-163491945:+
hyaluronan mediated motility receptor [Source:HGNC Symbol;Acc:H
HMMR-AS1
ENSG00000251018.2
101927813
chr5:163483065-163494058:-
HMMR antisense RNA 1 [Source:HGNC Symbol;Acc:HGNC:49149]
Showing 1 to 2 of 2 entries
Gene expression for HMMR (ENSG00000072571.19)
Data Source: GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2)
Data processing and normalization
SUBSET
None
Sex
SCALE
Log
Linear
TISSUE SORT
MEDIAN SORT
Gene expression for HMMR (ENSG00000072571.19)
TPM
120
100
80
60
40
20
0


UNIVERSITY OF CALIFORNIA
SANTA CRUZ
DATA SETS
VISUALIZATION
Help with transcripts
Add Gene (e.g. KRAS)
KRAS
Study A
GTEX Adipose Tissue
Study B
TCGA Uveal Melanoma
Expression Unit
TPM
log2 (TPM)
-7.5
-5
-2.5
0
2.5
density
1.14
0.38
0.18
KRAS ENST00000557334.5
KRAS ENST00000311936.7
KRAS ENST00000556131.1
KRAS ENST00000256078.8
1a
2
7b
1b
4
5
7c
1c
3
1d
6
7a

Gene Expression Omnibus

Keyword or GEO Accession **Search**

Browse Content

Repository Browser

DataSets:	4348
Series:	134024
Platforms:	21247
Samples:	3740858

SRA

SRA

Advanced

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2

Full

Send to:

Design: HiSeq X Ten 150bp paired end sequencing of sample 90888_N5. Library was created with index CGCTCATT.

Submitted by: Cancer Research UK Cambridge Institute

Study: Liver Cancer Evolution - Lesion segregation
PRJEB37808 • ERP121138 • All experiments • All runs
show Abstract

European Genome-phenome Archive

All

Examples: EGAS00000000001, Cancer

Search

EGA home | About | Studies | Datasets | Data access committees | Data providers | Submit to EGA | Contact Us | Login

The https://ega-archive.org/ site adds additional functionality not found on this site.

For queries or feedback related to the new site please contact ega-helpdesk@ega-archive.org

Help

- Users FAQ
- Submitters FAQ
- Using your EGA account
- Contact Us
- EGA mailing list

ArrayExpress – function

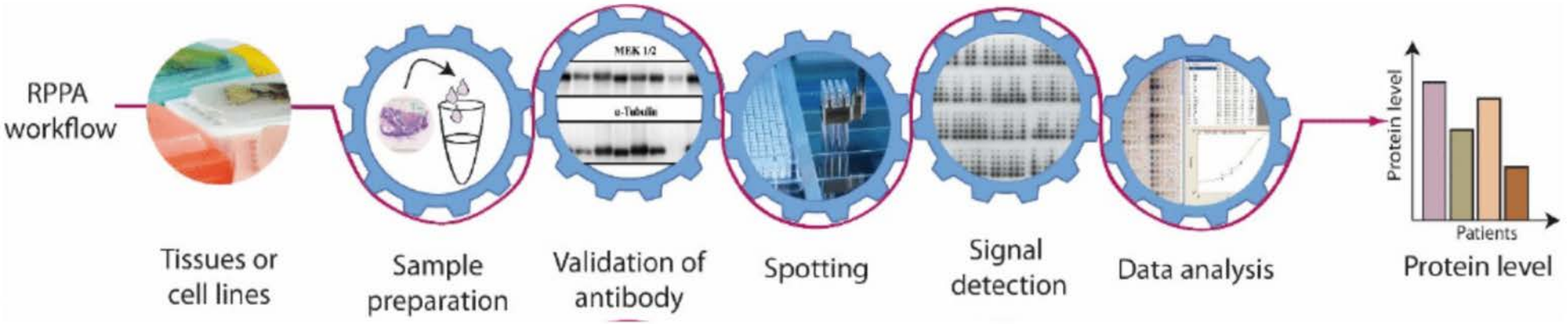
ArrayExpress Archive of Functional Genomics Data st… genomics experiments, and provides these data for reuse to the research community.

- 73320 experiments
- 2461459 assays
- 57.81 TB of archived data

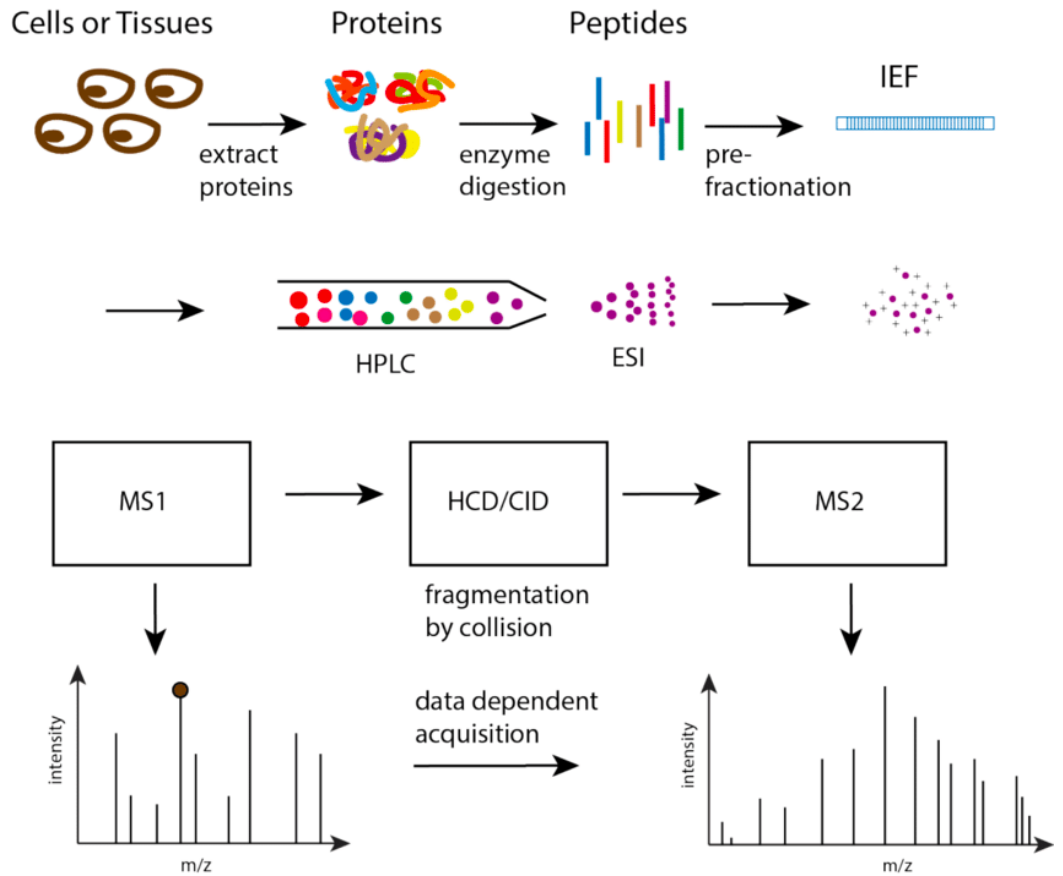
Browse ArrayExpress

Proteome
Phosphoproteome

RPPA



MS



899 cell lines

214 proteins

Protein Expression (numeric)

Stefanie Boellner, 2015

THE HUMAN PROTEIN ATLAS

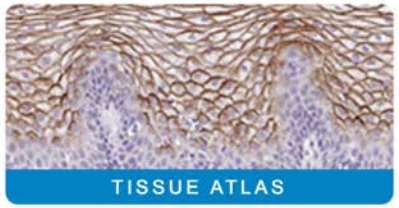
MENU HELP NEWS

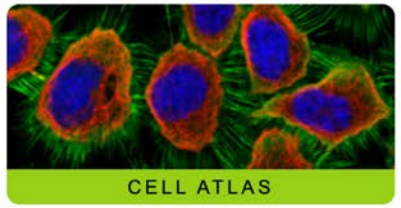
SEARCH

Search Fields »

e.g. RBM3, insulin, CD36

SARS-CoV-2 relevant proteins


TISSUE ATLAS


CELL ATLAS


PATHOLOGY ATLAS

PRIDE Archive

PRoteomics IDEntifications Database

Home Resources Tools Docs About

Search

NATIONAL CANCER INSTITUTE Office of Cancer Clinical Proteomics Research

Center for Strategic Scientific Initiatives

DATA PORTAL HOME ASSAY PORTAL AN

Data Portal

CPTAC 3 (2016-present)

CPTAC 2 (2011-2016)

CPTC (2006-2011)

Latest Data Release and Publications:

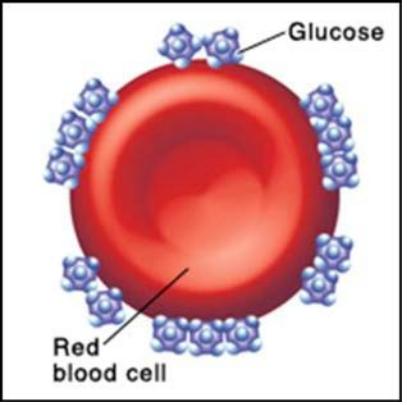
July 2020

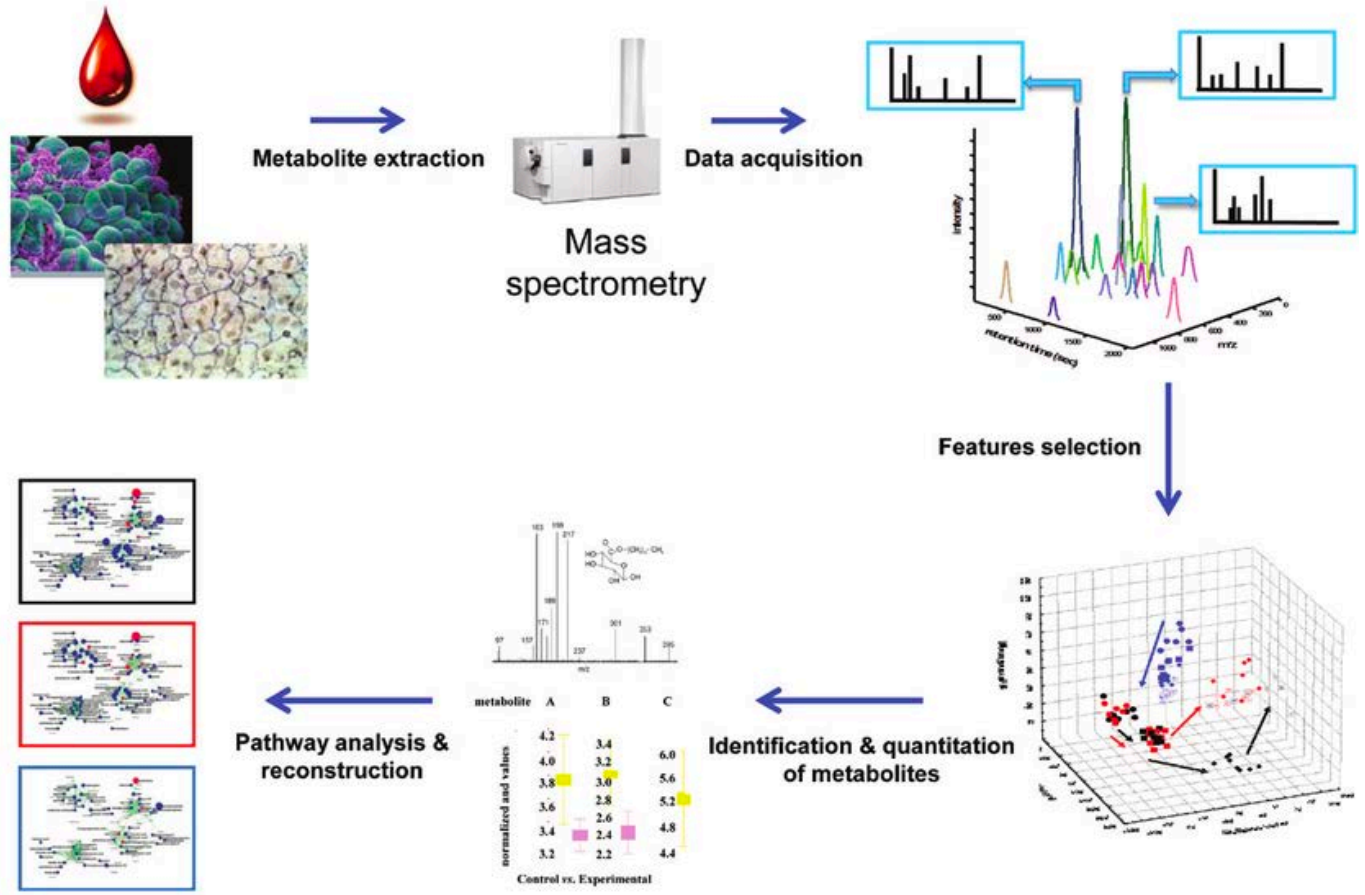
Proteogenomic Characterization Reveals Therap

Michael A. Gillette, Shankha Satpathy, Song Cao,

May 2020

CPTAC Head and Neck Squamous Cell Carcinom





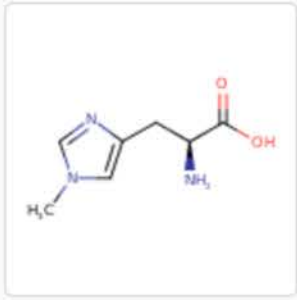
928 cell lines

225 metabolite

Metabolite abundance
(numeric)

Do Yup Lee, 2010

HMDB Browse Search Downloads About Contact Us Search

HMDB ID / CAS Number	Name	Structure	Formula / Average Mass / Monoisotopic Mass	Biospecimen
HMDB0000001 332-80-9	1-Methylhistidine		$C_7H_{11}N_3O_2$ 169.1811 169.085126611	Blood Cerebrospinal Feces Saliva Urine
HMDB0000002 109-76-2	1,3-Diaminopropane			

MetaboLights

Examples: Ala

Home | Browse Studies | Browse Compounds | Browse Species | Download | Help | Give us feedback | About

MetaboLights / Search

Filter your results

Type

- [x] study
- [ ] compound

Technology

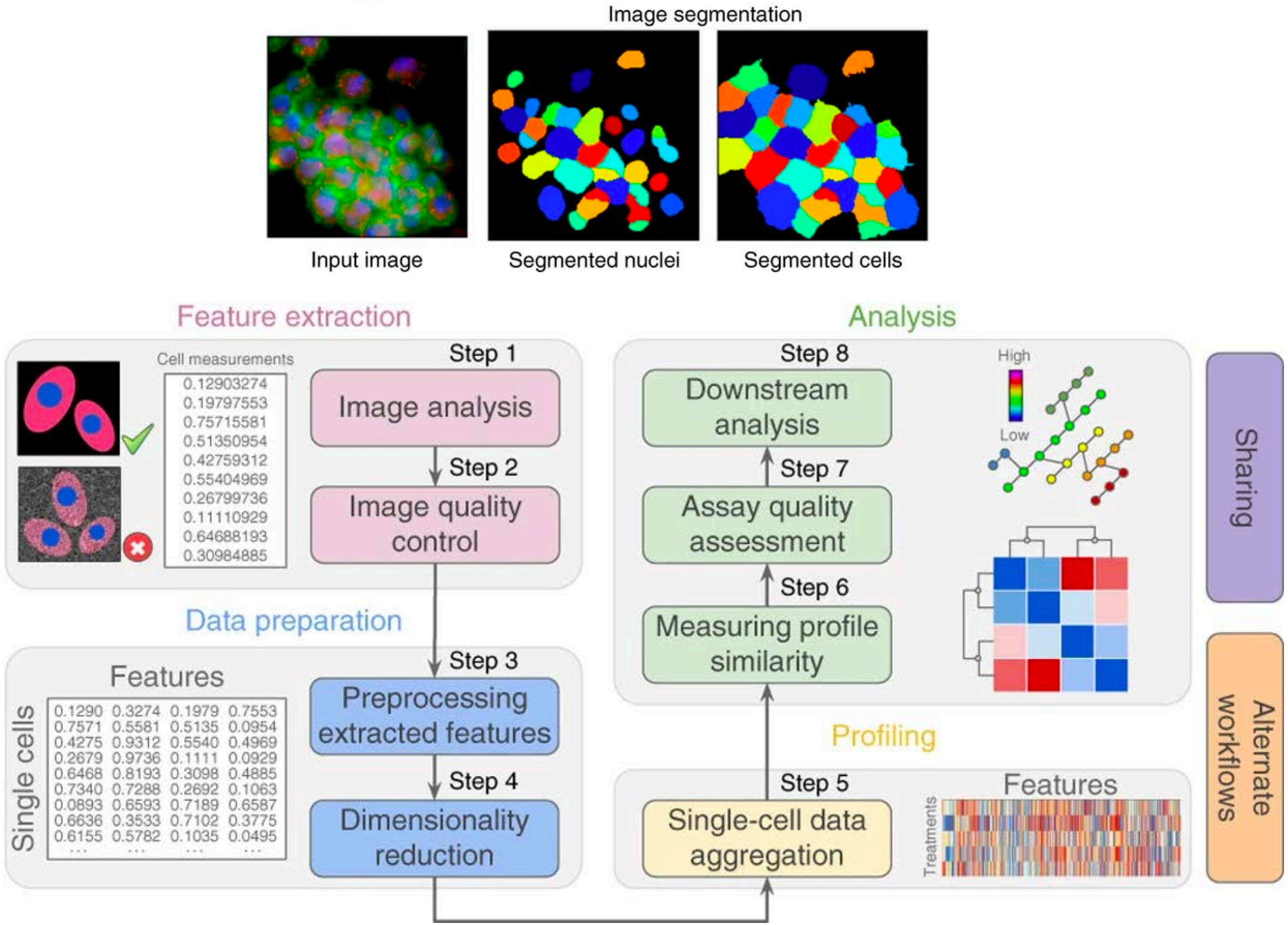
Organism

Organism Part

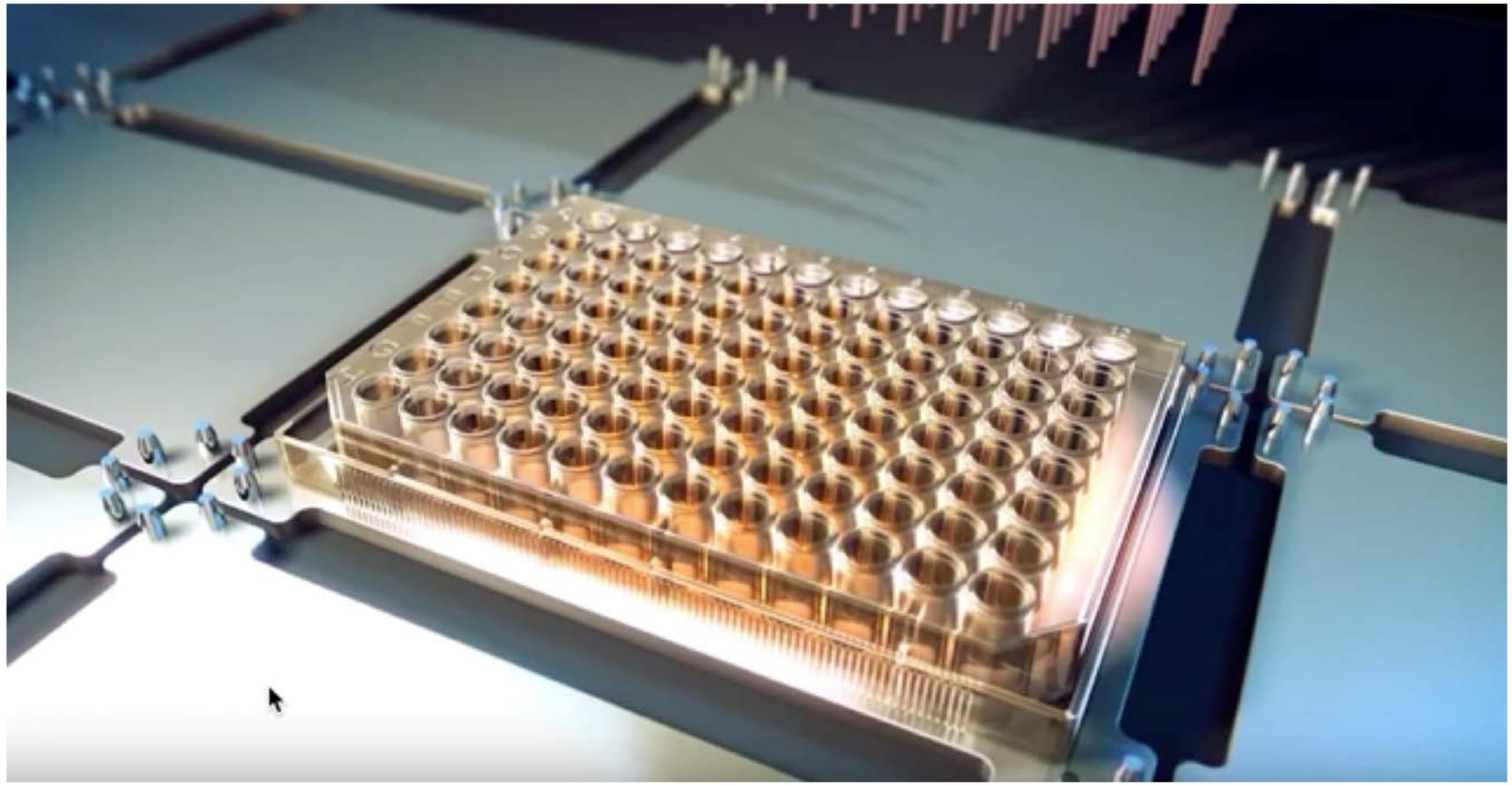
676 results , showing 1 to 10

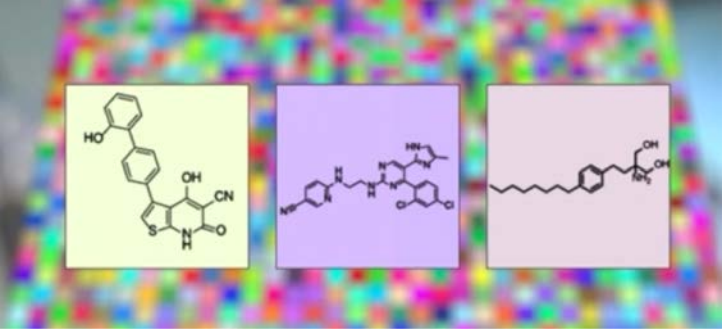
Stable Isotope–Assisted Plant Metabolomics: Combi Tracer-Based Labeling for Enhanced Untargeted Pro Annotation

Study Identifier	MTBLS1217	Organi
Study Size	939.11MB	Study
Submitted by	Christoph Bueschl	

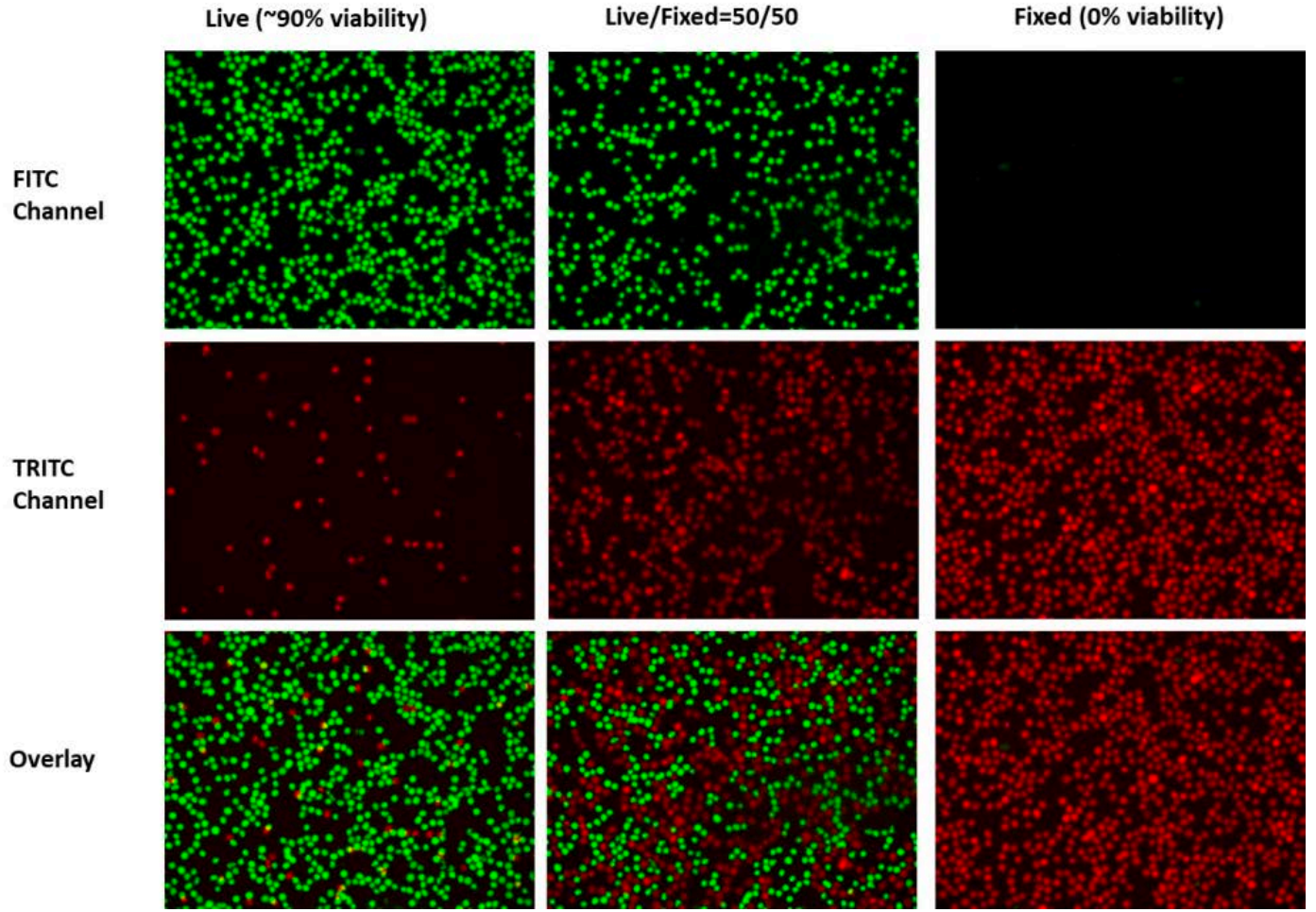


Juan C Caicedo, 2017





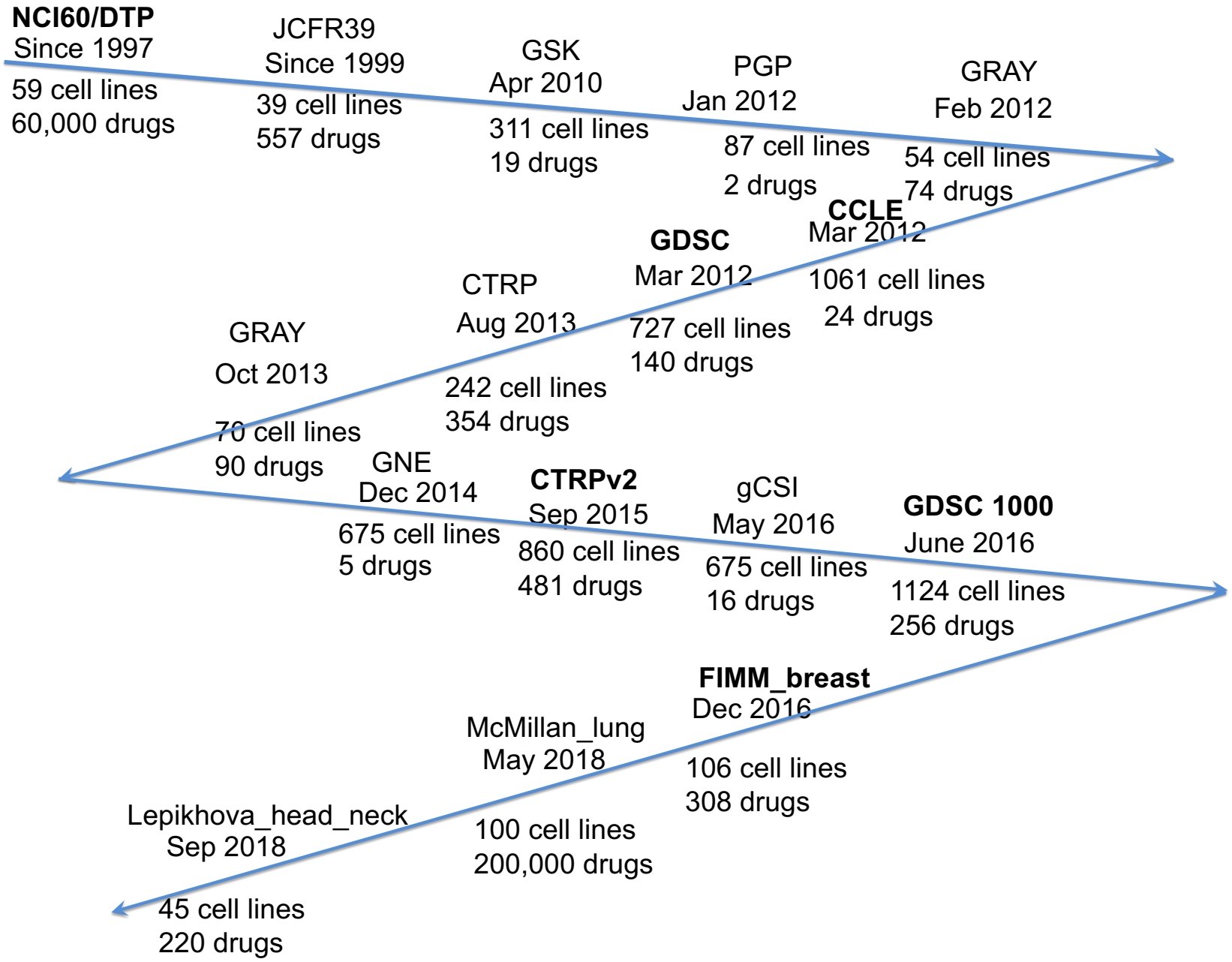
Cellular response/drug sensitivity



578 cell lines

5K compounds

Drug sensitivity score (numeric)


NCI60/DTP
Since 1997
59 cell lines
60,000 drugs
JCFR39
Since 1999
39 cell lines
557 drugs
GSK
Apr 2010
311 cell lines
19 drugs
PGP
Jan 2012
87 cell lines
2 drugs
GRAY
Feb 2012
54 cell lines
74 drugs
CCLE
Mar 2012
1061 cell lines
24 drugs
GDSC
Mar 2012
727 cell lines
140 drugs
CTRP
Aug 2013
242 cell lines
354 drugs
GRAY
Oct 2013
70 cell lines
90 drugs
GNE
Dec 2014
675 cell lines
5 drugs
CTRPv2
Sep 2015
860 cell lines
481 drugs
gCSI
May 2016
675 cell lines
16 drugs
GDSC 1000
June 2016
1124 cell lines
256 drugs
FIMM_breast
Dec 2016
106 cell lines
308 drugs
McMillan_lung
May 2018
100 cell lines
200,000 drugs
Lepikhova_head_neck
Sep 2018
45 cell lines
220 drugs

Cellular response/dependency

depmap

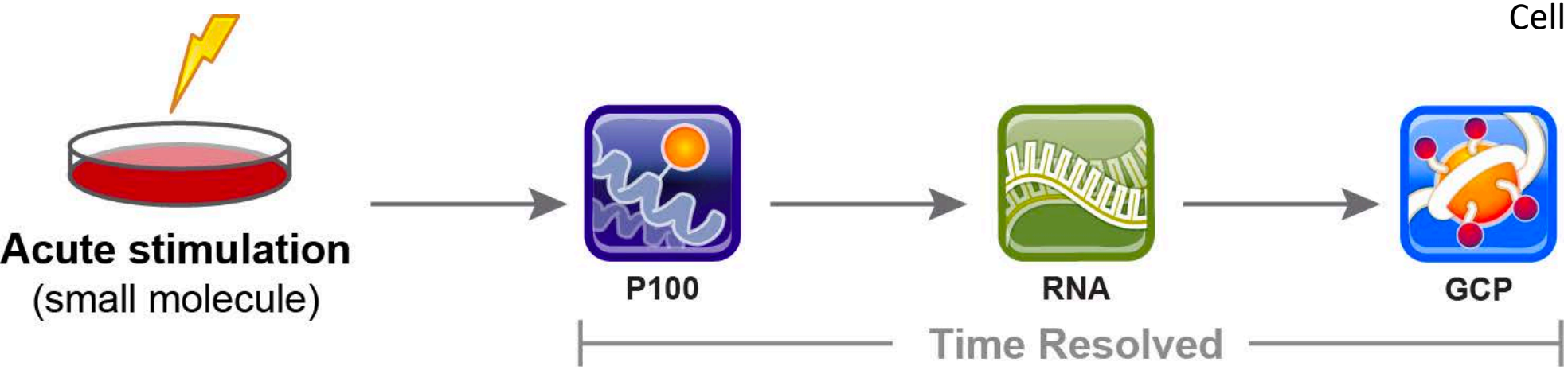
A Cancer Dependency Map to systematically identify genetic and pharmacologic dependencies and the biomarkers that predict them.

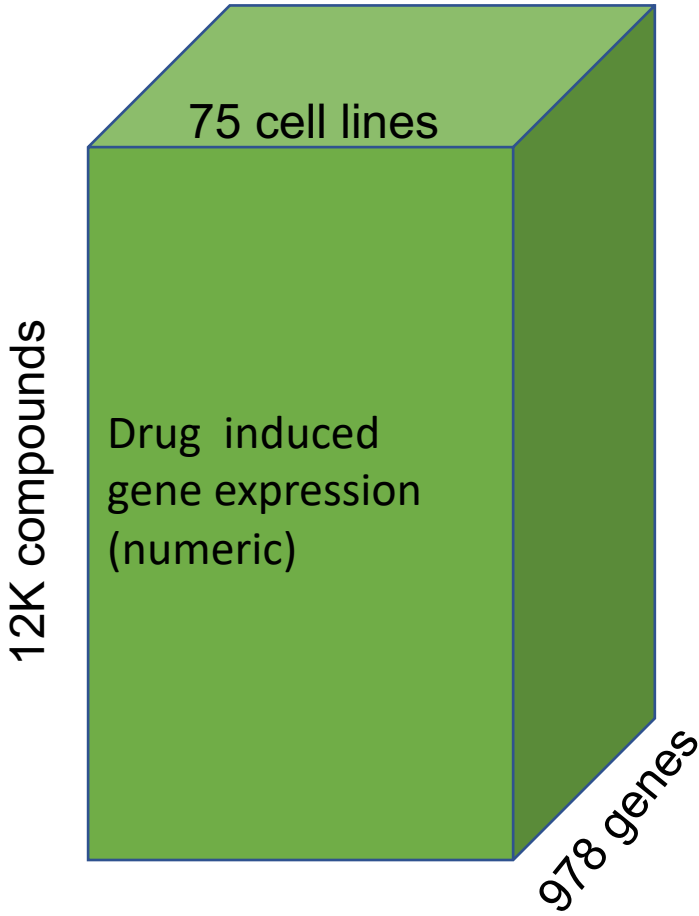
625 cell lines

18K genes

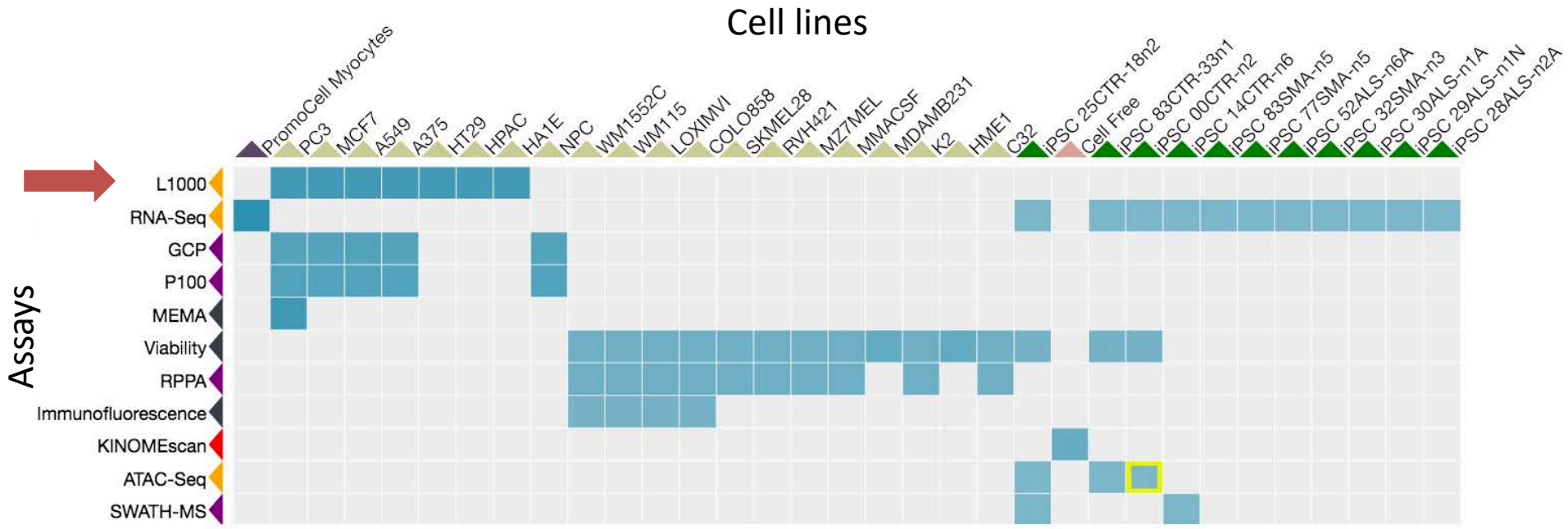
Gene essentiality score (numeric)

Cellular response/gene expression



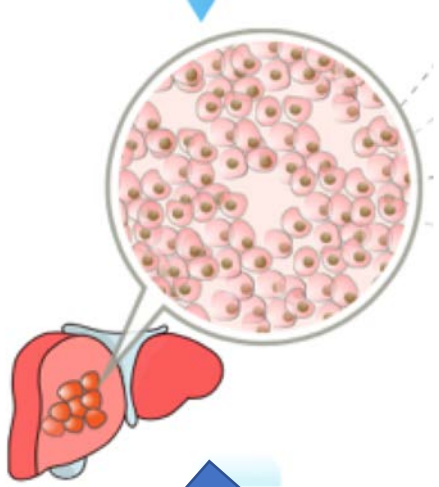


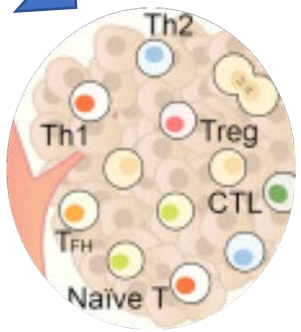
Library of Network-Based Cellular Signatures (LINCS)

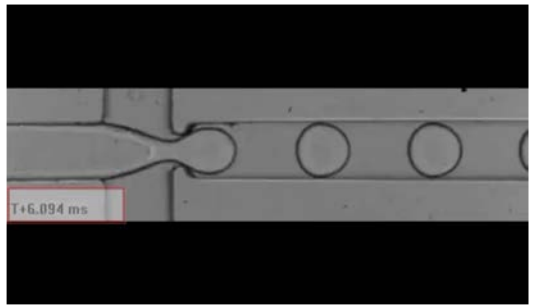


- >20,000 Small-molecule compounds
  - ~1,300 FDA-approved drugs
- >20,000 genomic perturbagens
- > 70 cell types
  - Cancer cell lines
  - Primary cells

http://www.lincsproject.org/data/data-releases/







20K patient tissues

Single cells

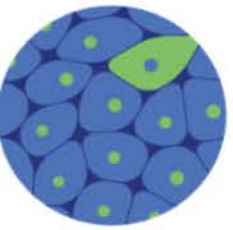
60K genes

Gene Expression
(numeric)

Millions of single cells

60K genes

Gene Expression
(numeric)



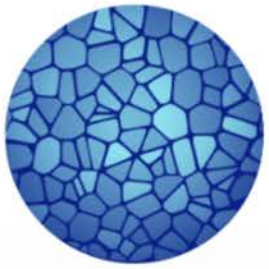
Single Cell PORTAL BETA

Reducing barriers and accelerating single-cell research

Featuring
272 studies
9,584,113 cells

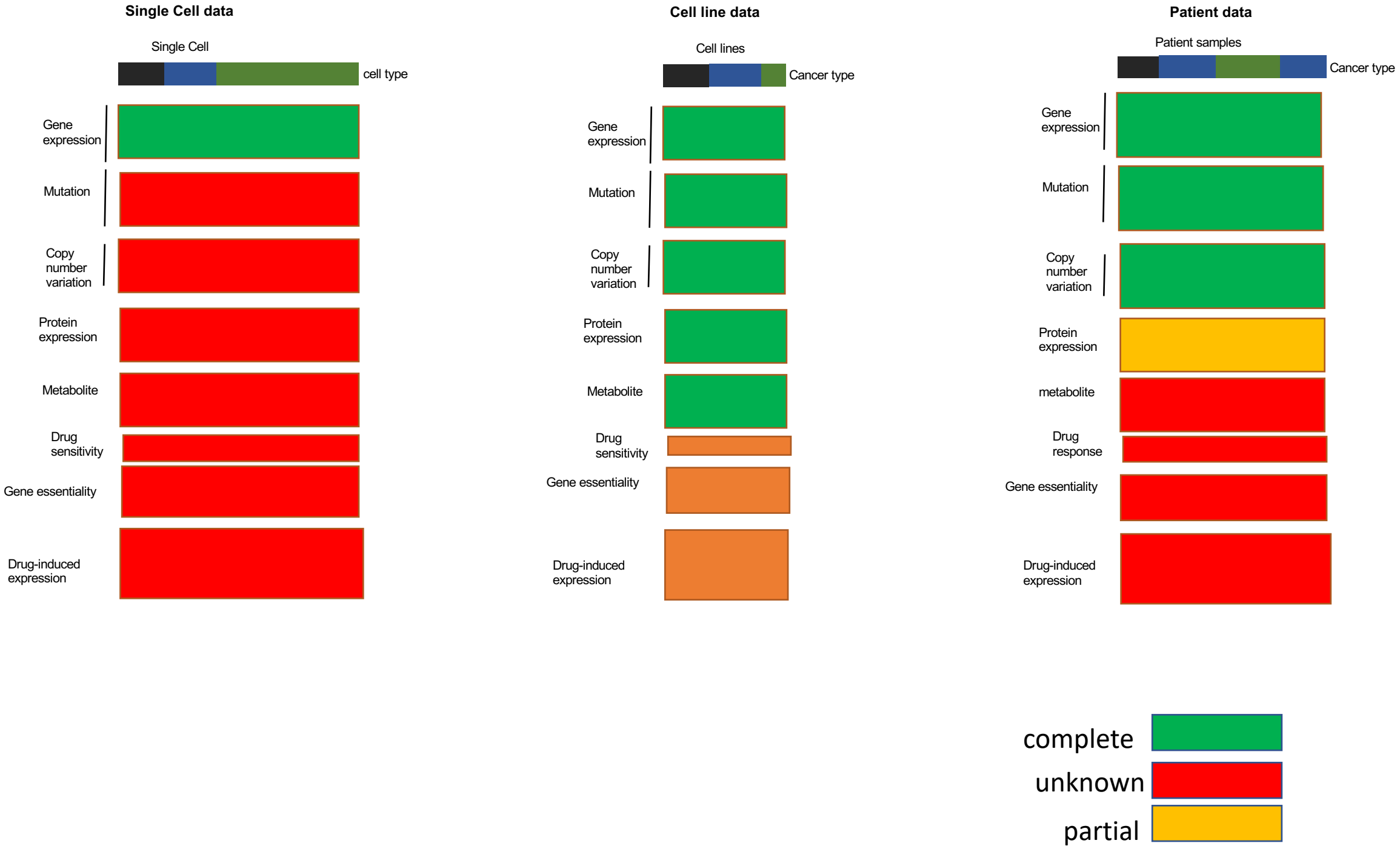
COVID-19 Cell Atlas

wellcome sanger institute



HUMAN CELL ATLAS

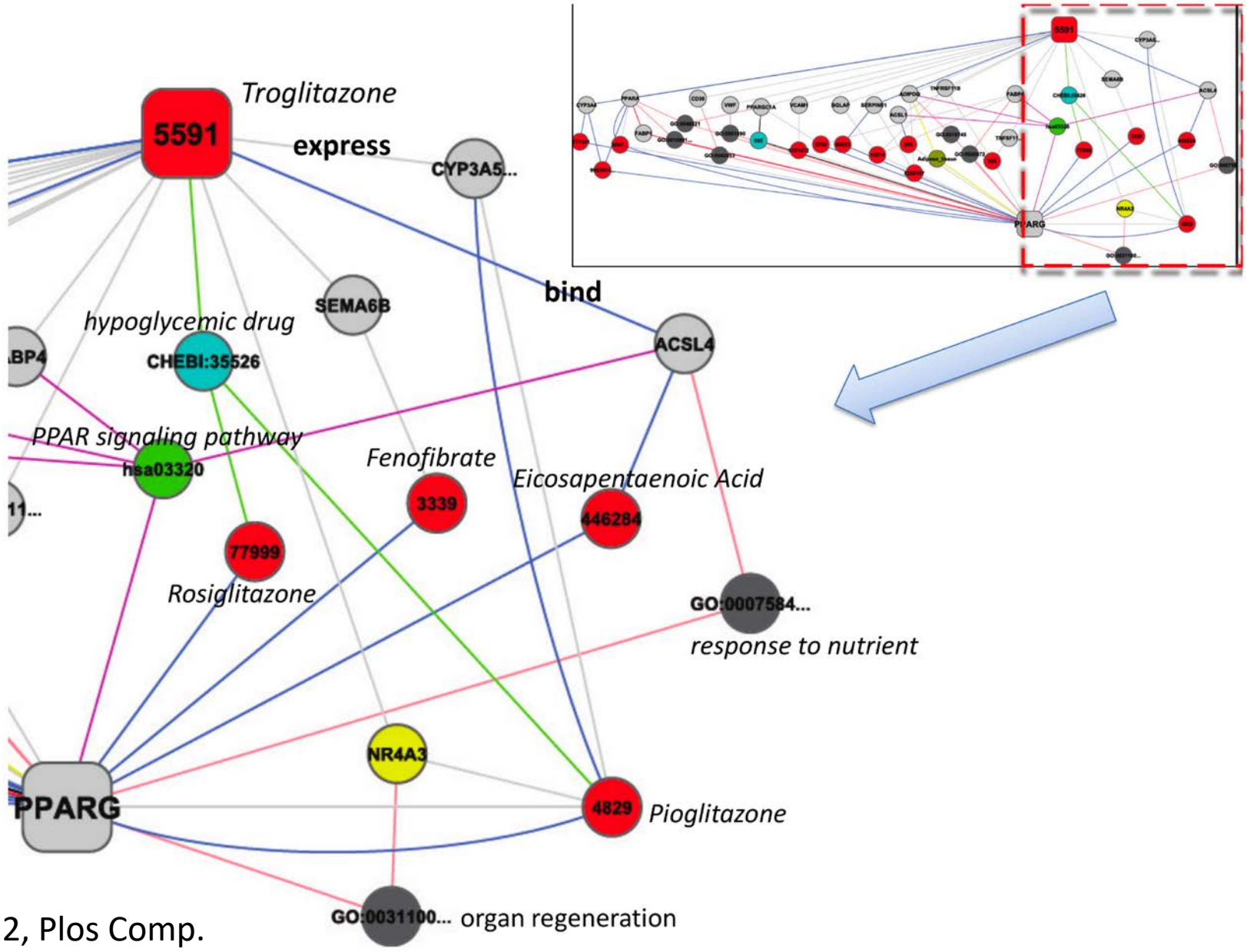
CHAN ZUCKERBERG INITIATIVE


Single Cell data
Single Cell
cell type
Gene expression
Mutation
Copy number variation
Protein expression
Metabolite
Drug sensitivity
Gene essentiality
Drug-induced expression
Cell line data
Cell lines
Cancer type
Gene expression
Mutation
Copy number variation
Protein expression
Metabolite
Drug sensitivity
Gene essentiality
Drug-induced expression
Patient data
Patient samples
Cancer type
Gene expression
Mutation
Copy number variation
Protein expression
metabolite
Drug response
Gene essentiality
Drug-induced expression
complete
unknown
partial

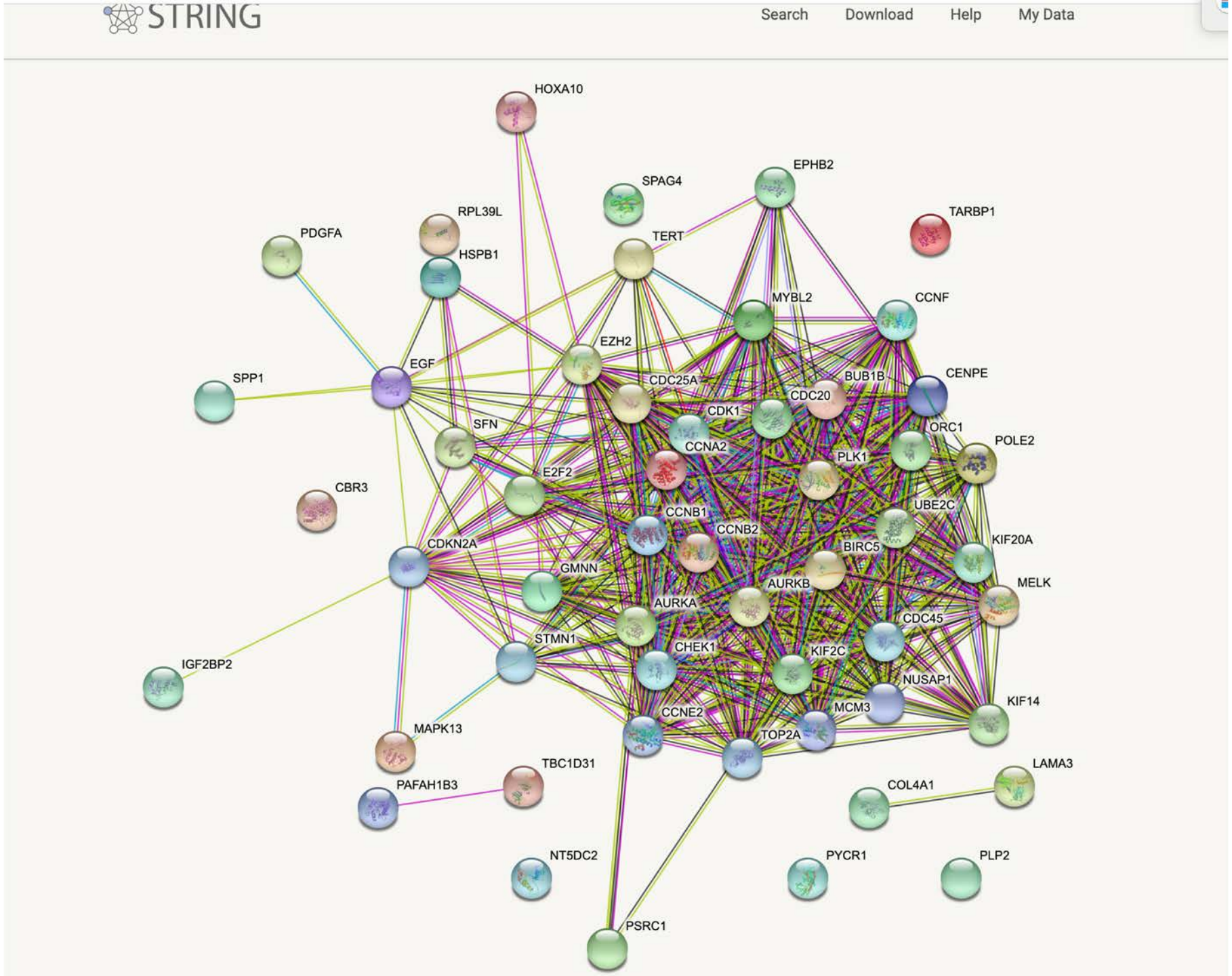
Data properties

- Matrix (small-samples, big features)
- Sparse
- Noisy (High-throughput)
- Batch effect
- Processed data

Network (knowledge graph)



Bin Chen, 2012, Plos Comp.



Stringdb.org

Unstructured data

1 INDICATIONS AND USAGE
1.1 EGFR Mutation-Positive, Metastatic Non-Small Cell Lung Cancer
GILOTRIF is indicated for the first-line treatment of patients with metastatic non-small cell lung cancer (NSCLC) whose tumors have non-resistant epidermal growth factor receptor (EGFR) mutations as detected by an FDA-approved test [see Clinical Pharmacology (12.1) and Clinical Studies (14.1)]. Limitation of Use: The safety and efficacy of GILOTRIF have not been established in patients whose tumors have resistant EGFR mutations [see Clinical Studies (14.1)].

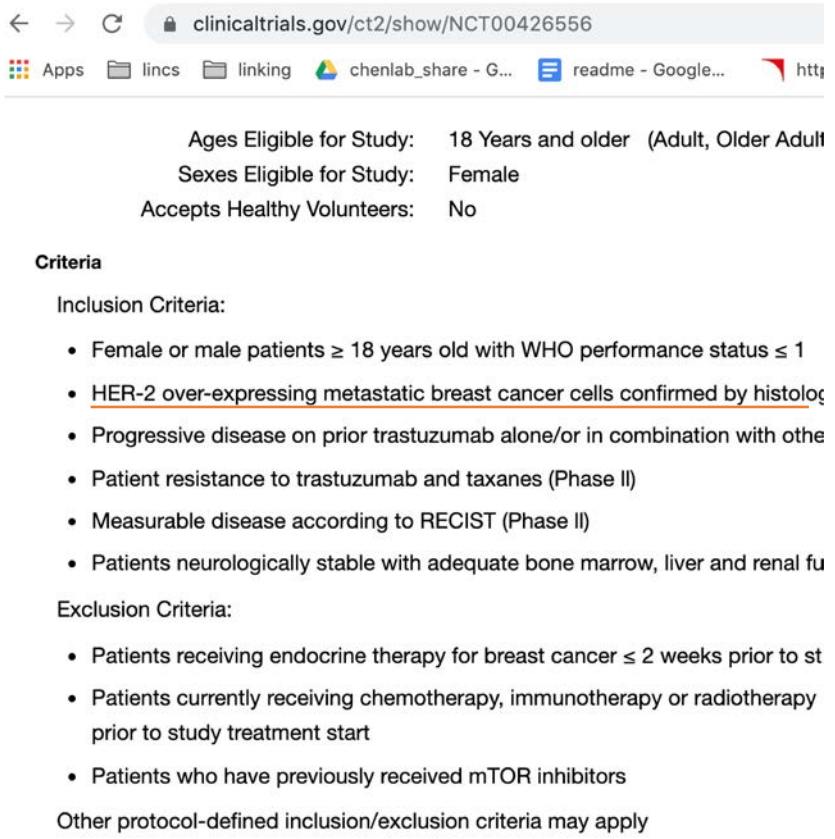
2 DOSAGE AND ADMINISTRATION
2.1 Patient Selection for EGFR Mutation-Positive Metastatic NSCLC
2.1 Patient Selection for Non-Resistant EGFR Mutation-Positive Metastatic NSCLC Select patients for first-line treatment of metastatic NSCLC with GILOTRIF based on the presence of nonresistant EGFR mutations in tumor specimens [see Indications and Usage (1.1) and Clinical Studies (14.1)]. Information on FDA-approved tests for the detection of EGFR mutations in NSCLC is available at: http://www.fda.gov/CompanionDiagnostics.

6 ADVERSE REACTIONS
6.1 Clinical Trials Experience
The data described below reflect exposure to GILOTRIF as a single agent in LUX-Lung 3, a randomized, active-controlled trial conducted in patients with EGFR mutation-positive, metastatic NSCLC, and in LUX-Lung 8, a randomized, active controlled trial in patients with metastatic squamous NSCLC progressing after platinum-based chemotherapy.(…)
EGFR Mutation-Positive, Metastatic NSCLC

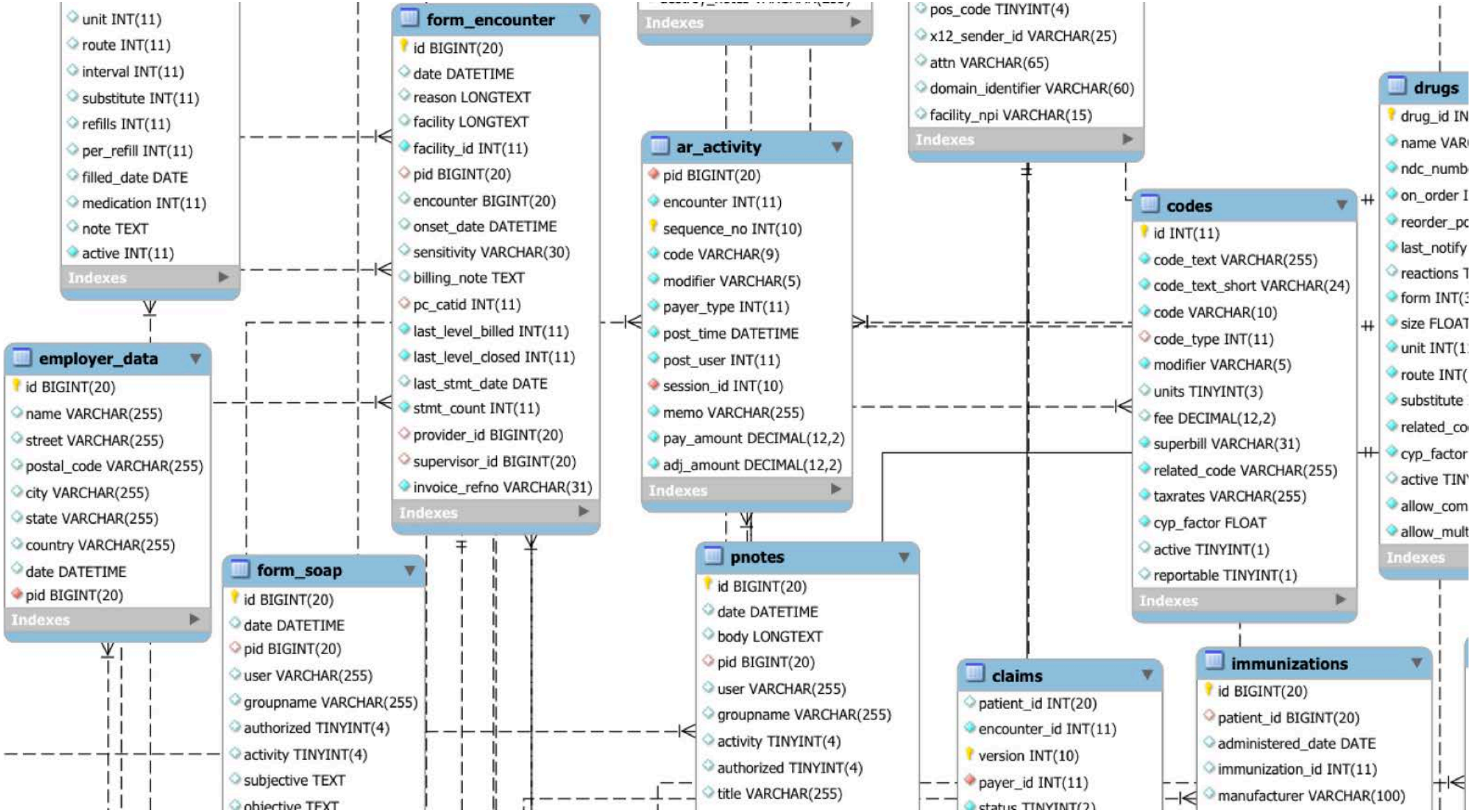
Summary

Go to:

The high rate of clinical response to protein kinase-targeting drugs matched to cancer patients with specific genomic alterations has prompted efforts to use cancer cell-line (CCL) profiling to identify additional biomarkers of small-molecule sensitivities. We have quantitatively measured the sensitivity of 242 genomically characterized CCLs to an Informer Set of 354 small molecules that target many nodes in cell circuitry, uncovering protein dependencies that: 1) associate with specific cancer-genomic alterations and 2) can be targeted by small molecules. We have created the Cancer Therapeutics Response Portal (www.broadinstitute.org/ctrp) to enable users to correlate genetic features to sensitivity in individual lineages and control for confounding factors of CCL profiling. We report a candidate dependency, associating activating mutations in the oncogene β-catenin with sensitivity to the Bcl2-family antagonist, navitoclax. The resource can be used to develop novel therapeutic hypotheses and accelerate discovery of drugs matched to patients by their cancer genotype and lineage.



Electronic Medical Records



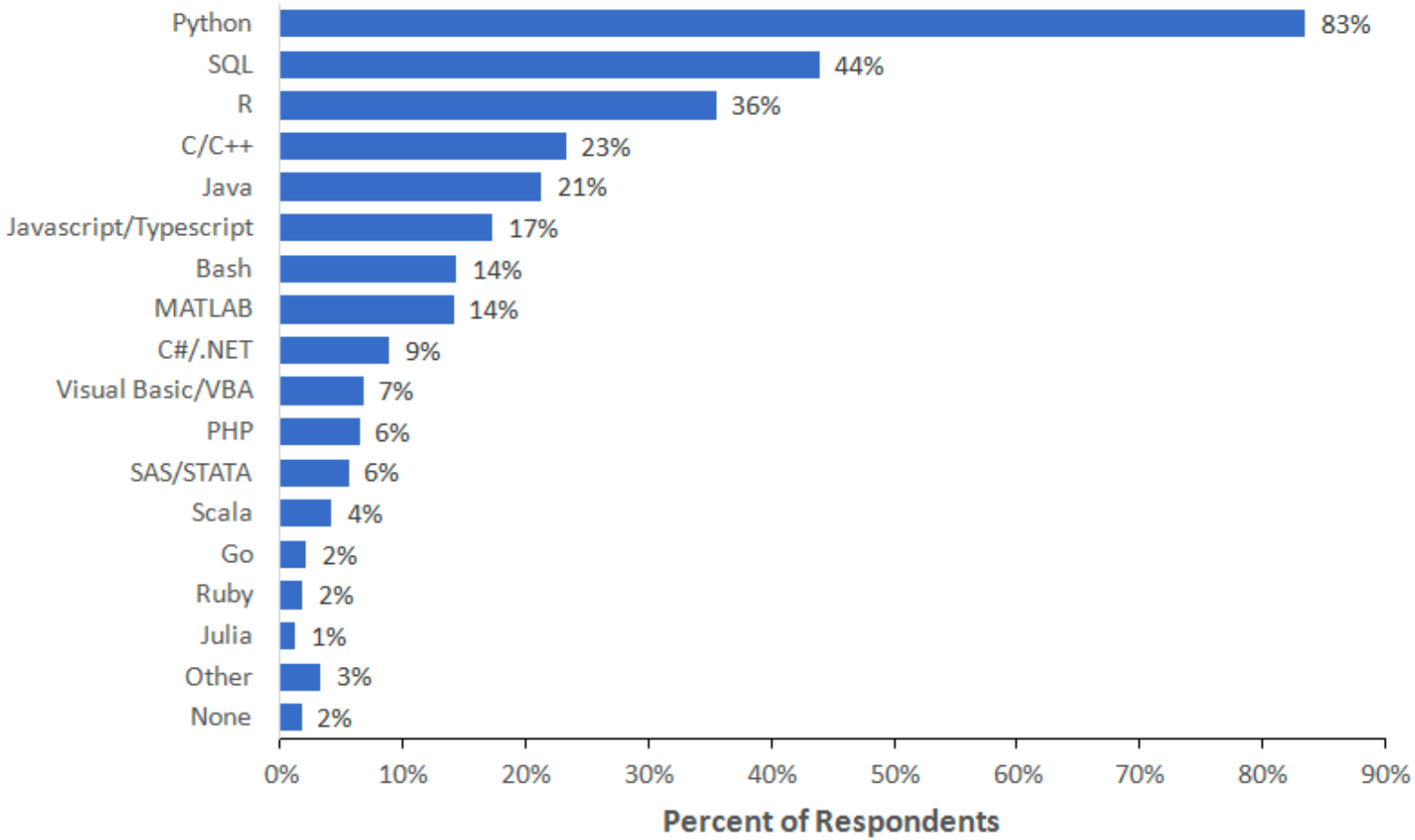
Summary of big data

- Unstructured
  - Free text (literature, clinical notes)
- Structured
  - Network
    - Protein-protein interaction
    - Knowledge-graph (e.g., drug-target interaction, drug-disease relation, target-pathway)
  - Matrix
    - Imaging
    - Omics (disease-based, cellular-responses)
      - Genomics
      - Transcriptomics
      - Proteomics
      - Metabolomics
      - …

Session 1: Big Data in translational bioinformatics

Session 2: Big Data in R

What programming language do you use on a regular basis?



Note: Data are from the 2018 Kaggle Machine Learning and Data Science Survey. You can learn more about the study here: http://www.kaggle.com/kaggle/kaggle-survey-2018. A total of 18827 respondents answered the question.

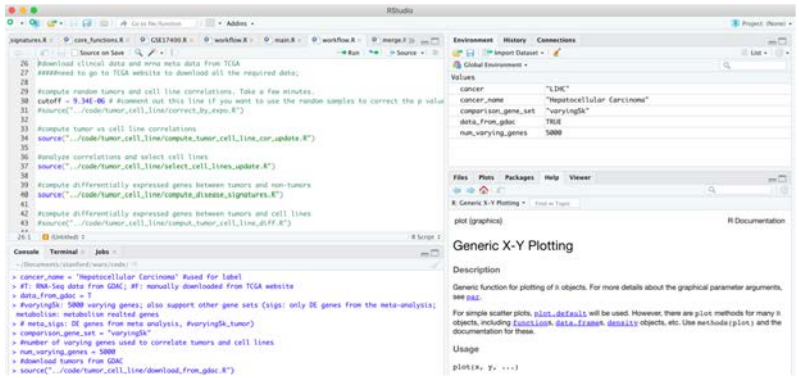
BUSINESS BROADWAY
DATA SCIENCE | CUSTOMER ANALYTICS | MACHINE LEARNING

R resources

- Tutorial
  - http://manuals.bioinformatics.ucr.edu/home/R_BioCondManual/
  - https://rstudio.com/resources/cheatsheets/
  - https://www.r-bloggers.com/
  - http://rafalab.github.io/pages/harvardx.html
- Troubleshooting
  - https://stackoverflow.com/
  - https://www.biostars.org/
- GitHub code repository

R

- Install R
- Install Rstudio



RStudio

Basic grammar:

- Variable
- Data type
- Data input/output
- Basic operator
- Conditionals and loops
- Function

R Packages
>10K

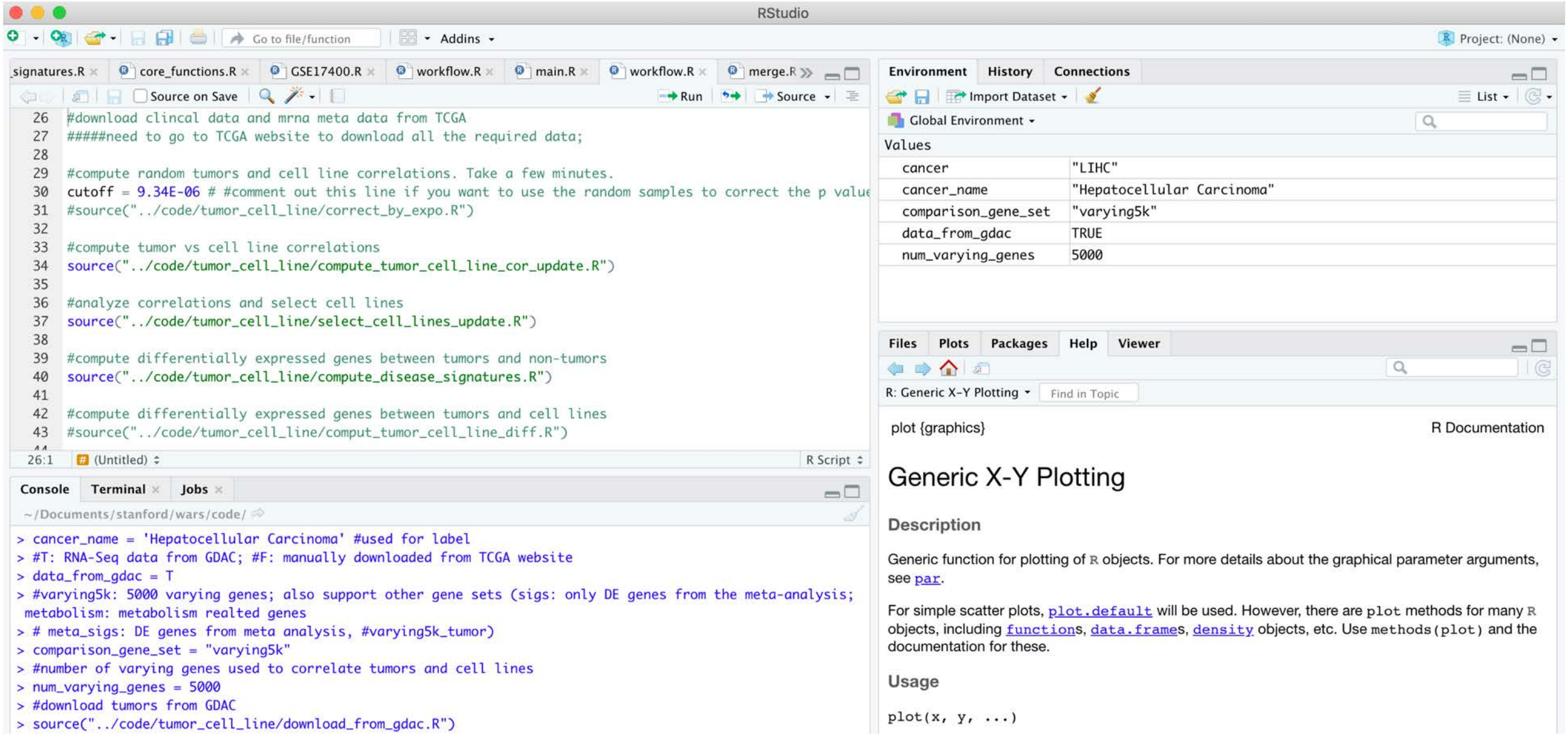
Base

customized

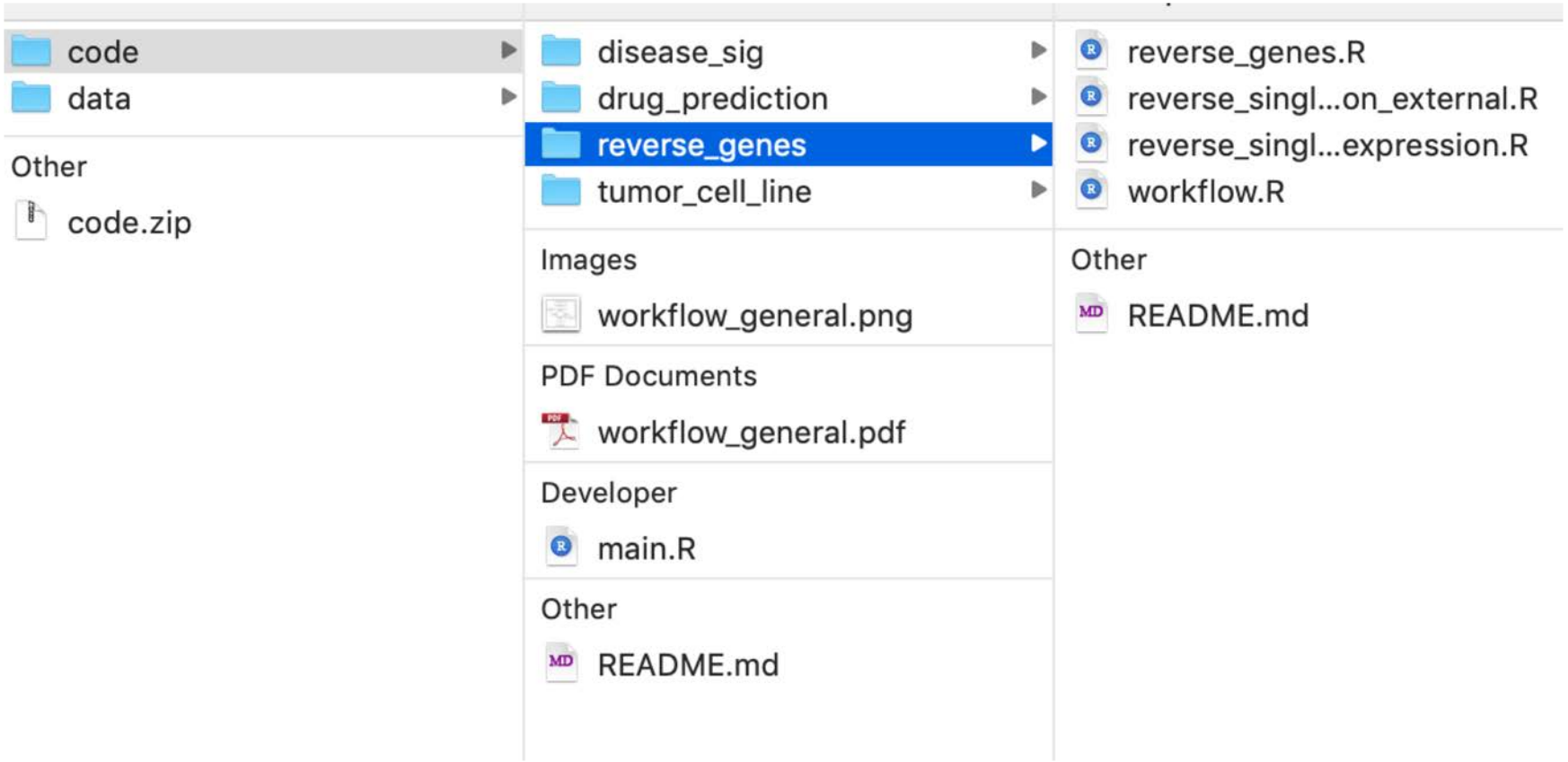
- Data manipulation: dplyr
- Data visualization: ggplot
- Web developer: shiny
- Documentation: Rmarkdown
- Text processing: stringr
- Machine learning: e1071

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

~2000


RStudio
Go to file/function
Addins
Project: (None)
signatures.R
core_functions.R
GSE17400.R
workflow.R
main.R
workflow.R
merge.R
Source on Save
Run
Source
#download clincal data and mrna meta data from TCGA
#####need to go to TCGA website to download all the required data;
#compute random tumors and cell line correlations. Take a few minutes.
cutoff = 9.34E-06 # #comment out this line if you want to use the random samples to correct the p value
#source("../code/tumor_cell_line/correct_by_expo.R")
#compute tumor vs cell line correlations
source("../code/tumor_cell_line/compute_tumor_cell_line_cor_update.R")
#analyze correlations and select cell lines
source("../code/tumor_cell_line/select_cell_lines_update.R")
#compute differentially expressed genes between tumors and non-tumors
source("../code/tumor_cell_line/compute_disease_signatures.R")
#compute differentially expressed genes between tumors and cell lines
#source("../code/tumor_cell_line/comput_tumor_cell_line_diff.R")
26:1
(Untitled)
R Script
Console
Terminal
Jobs
~/Documents/stanford/wars/code/
> cancer_name = 'Hepatocellular Carcinoma' #used for label
> #T: RNA-Seq data from GDAC; #F: manually downloaded from TCGA website
> data_from_gdac = T
> #varying5k: 5000 varying genes; also support other gene sets (sigs: only DE genes from the meta-analysis;
metabolism: metabolism realted genes
> # meta_sigs: DE genes from meta analysis, #varying5k_tumor)
> comparison_gene_set = "varying5k"
> #number of varying genes used to correlate tumors and cell lines
> num_varying_genes = 5000
> #download tumors from GDAC
> source("../code/tumor_cell_line/download_from_gdac.R")
Environment
History
Connections
Import Dataset
List
Global Environment
Values
cancer "LIHC"
cancer_name "Hepatocellular Carcinoma"
comparison_gene_set "varying5k"
data_from_gdac TRUE
num_varying_genes 5000
Files
Plots
Packages
Help
Viewer
R: Generic X-Y Plotting
Find in Topic
plot {graphics}
R Documentation
Generic X-Y Plotting
Description
Generic function for plotting of R objects. For more details about the graphical parameter arguments, see par.
For simple scatter plots, plot.default will be used. However, there are plot methods for many R objects, including functions, data.frames, density objects, etc. Use methods(plot) and the documentation for these.
Usage
plot(x, y, ...)

Demo R project



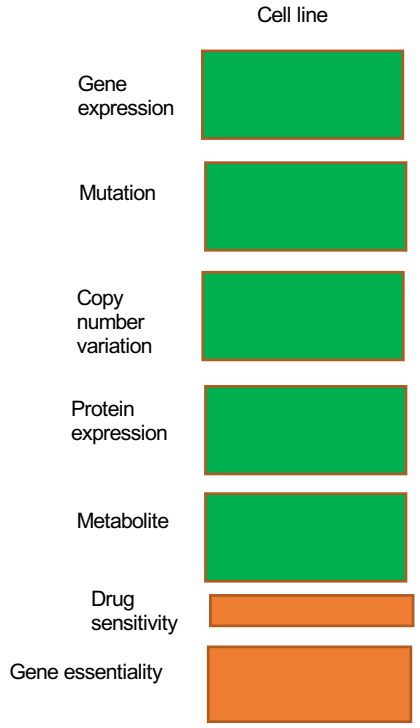
https://github.com/Bin-Chen-Lab/HCC_NEN

OCTAD_cell_line.RData

octad_cell_line_features

	id	name	type
VPS13D	mutation_VPS13D	VPS13D	mutation
AADACL4	mutation_AADACL4	AADACL4	mutation
TMEM57	mutation_TMEM57	TMEM57	mutation
ZSCAN20	mutation_ZSCAN20	ZSCAN20	mutation
POU3F1	mutation_POU3F1	POU3F1	mutation
VAV3	mutation_VAV3	VAV3	mutation

octad_cell_line_matrix



octad_cell_line_meta

DepMap_ID	stripped_cell_line_name	CCLE.Name	alias	COSMIC_ID	lineage	lineage_subtype
ACH-000001	NIHOVCAR3	NIHOVCAR3_OVARY	OVCAR3	905933	ovary	ovary_adenocarcinoma
ACH-000002	HL60	HL60_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE		905938	leukemia	AML
ACH-000003	CACO2	CACO2_LARGE_INTESTINE	CACO2, CaCo-2	NA	colorectal	
ACH-000004	HEL	HEL_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE		907053	leukemia	AML
ACH-000005	HEL9217	HEL9217_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE		NA	leukemia	AML
ACH-000006	MONOMAC6	MONOMAC6_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE		908148	leukemia	AML

Data input and output

Data Type

- Numeric
- Character
- Logical
- Factor

Data Type

- Vector
- Data.frame
- Matrics
- Arrays
- List
- RData

Advanced data Type

- Image
- Unstructured text
- Class

Big files

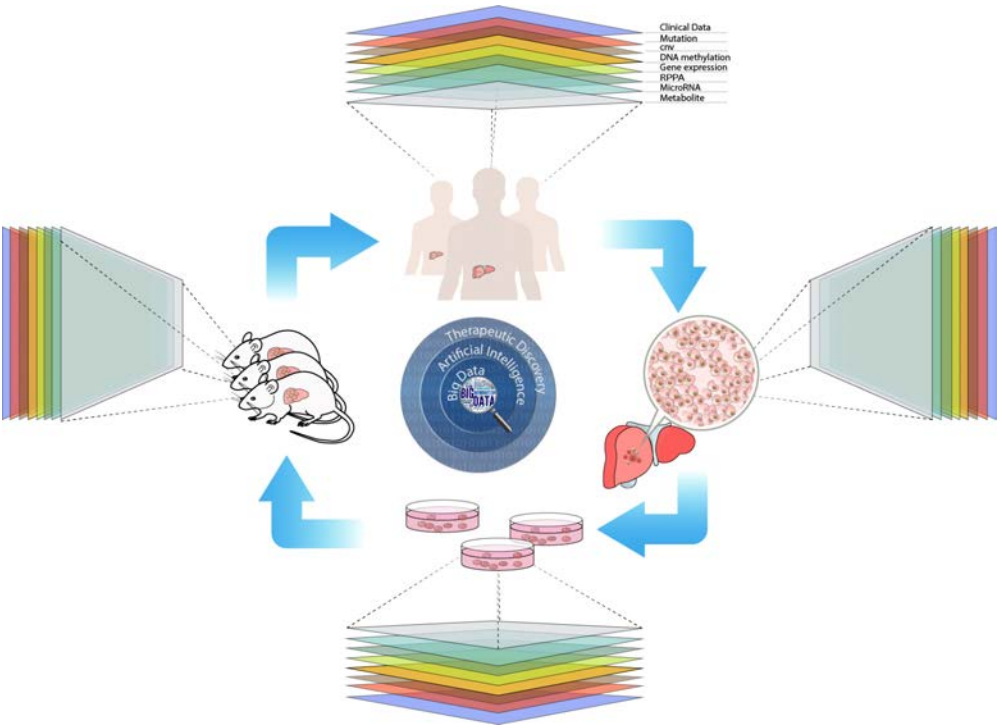
- Data.table
- H5

Subsetting

Basic Operators and Calculations

Data summary

Connecting open data points to facilitate translational research



Data (grey: to be processed)

Cell lines

- Gene expression (1210 cell lines X 19k genes)
- Mutation (1656 cell lines X 19K genes)
- Copy number (1657 cell lines X 28K genes)
- Protein expression (899 cell lines X 214 proteins)
- Metabolite abundance (928 cell lines X 225 metabolites)

- CRISPR (625 cell lines X 18K genes)
- RNAi (712 cell lines X 17K genes)
- Drug sensitivity (578 cell lines X 5K cmpds)
- Drug expression profile (75 cell lines X 12K cmpds X 978 genes)

Animals

- Ad-hoc

Patient tissues

- Bulk disease (18K samples X 60K transcripts)
- Bulk normal (7K samples X 60K transcripts)
- Single cell

Patient EMR (Spectrum Health)

- Medication, lab test, bill, outcome, disease condition

Tool/Model

- Query
- Visualization
- Statistical analysis
- Predictive models