MICHIGAN STATE UNIVERSITY

ChenLab
Michigan State University
http://binchenlab.org

Big Data in Pharmacology

Bin Chen

Associate 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

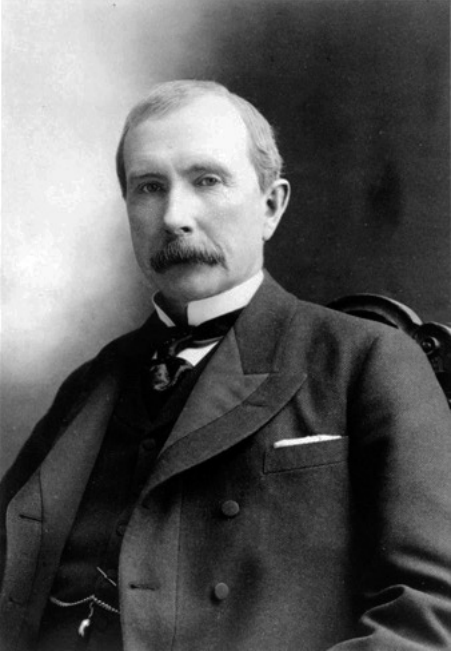
IPSTP boot camp, Summer, 2024

Learning objectives

- Understand pharmacology-related big data technologies and resources

Beyond Excel, GraphPad …..

Data = Oil



John D. Rockefeller in 1885

Company	Founding year	Investment (by 2021)
Insitro	2018	$734M
XtalPi	2014	$386M
Exscientia	2012	$148M

https://www.nanalyze.com/2021/04/companies-ai-drug-discovery/

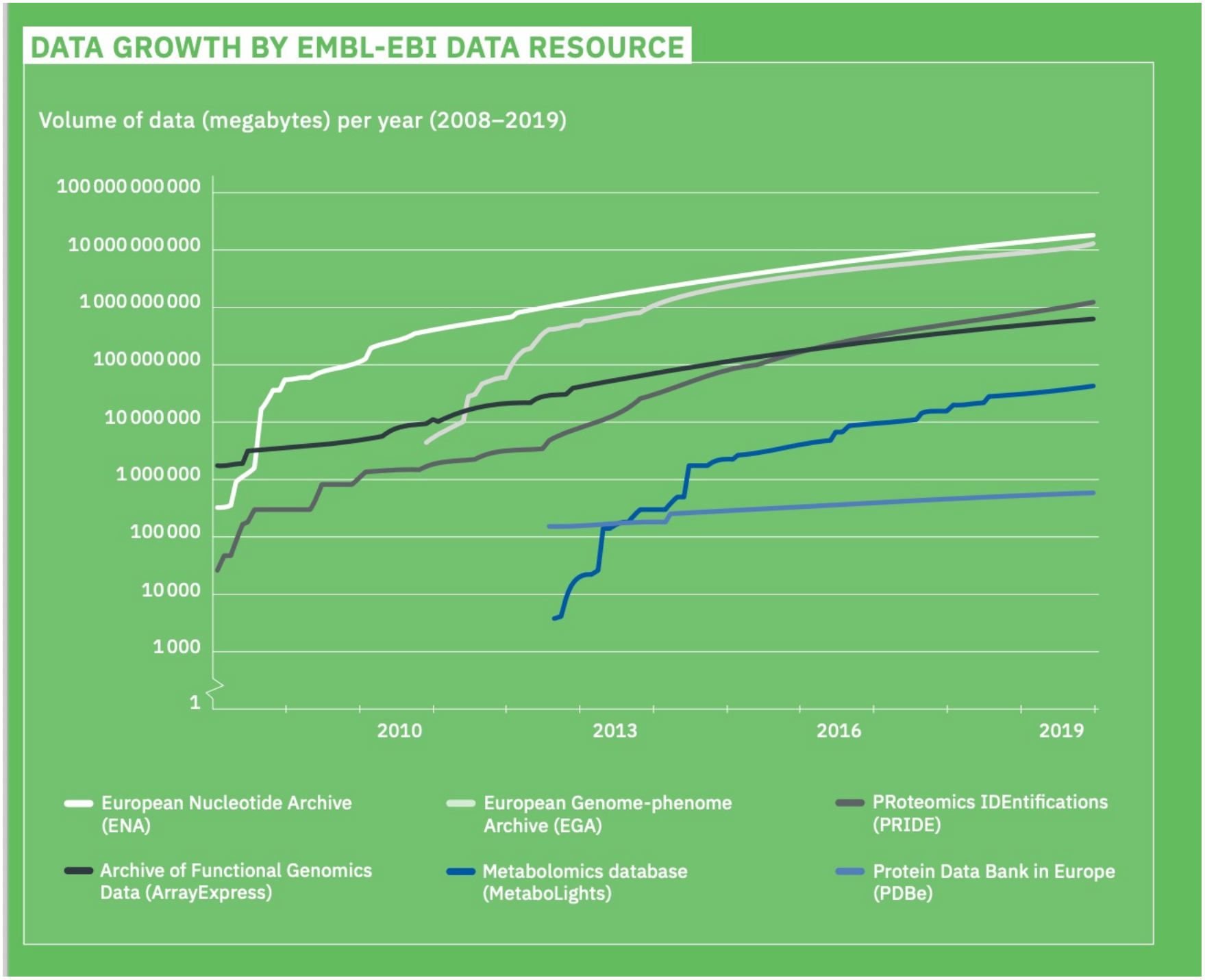
EMBL-EBI

25 petabytes 2014

307 petabytes 2019

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

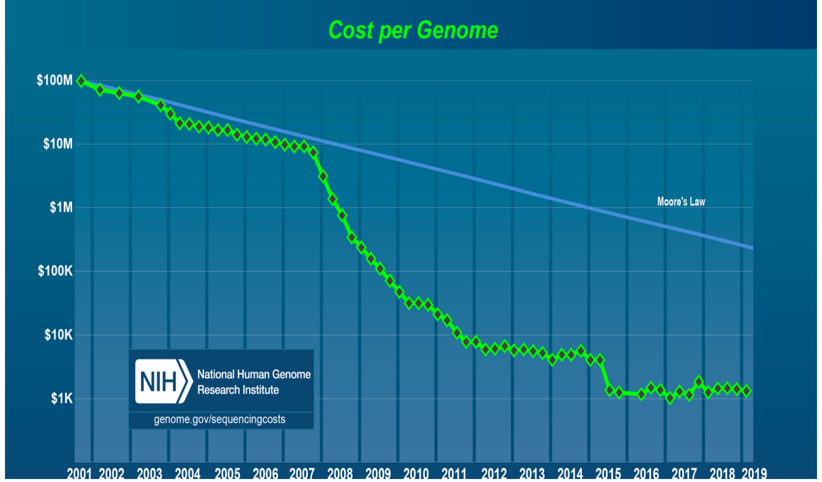
307 petabytes=  × 150,000

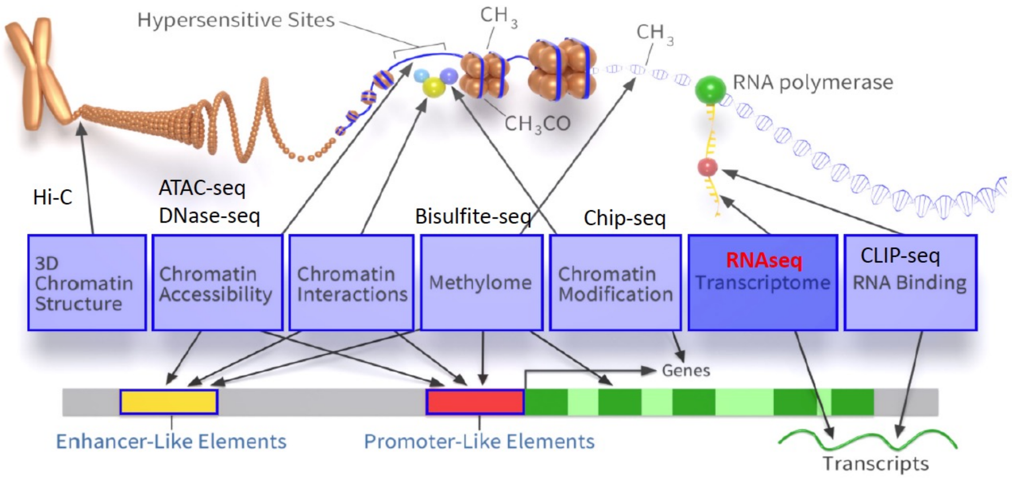


shorturl.at/tyFGV

Why so much data available today?

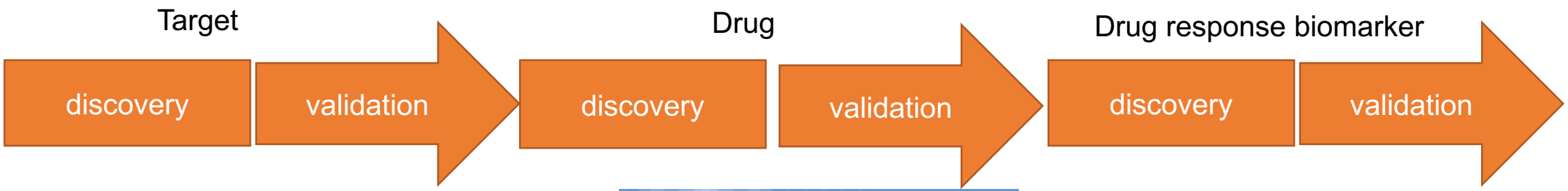
- Low sequencing costs
- High throughput technologies
- High performance computing





hael Pazin (NHGRI)

Big Data in drug discovery



1000 Genomes
A Deep Catalog of Human Genetic Variation

GEO
Gene Expression Omnibus

NIH LINCS PROGRAM

Cancer Therapeutics Response Portal

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

CONNECTIVITY MAP

IMMPORT
BIOINFORMATICS FOR THE FUTURE OF IMMUNOLOGY

GTEx Portal

ENCODE

ChEMBL

ClinicalTrials.gov

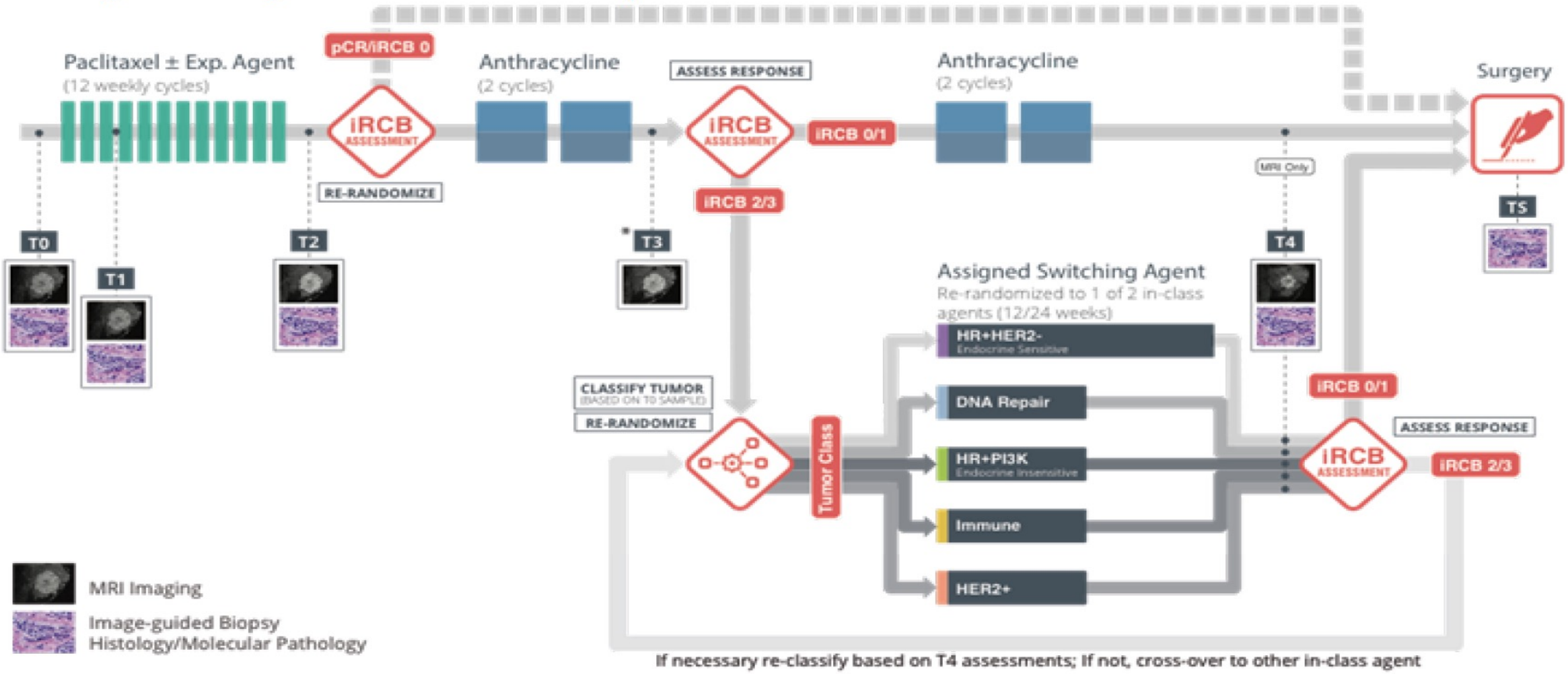
THE HUMAN PROTEIN ATLAS

CCLE Cancer Cell Line Encyclopedia

PubChem

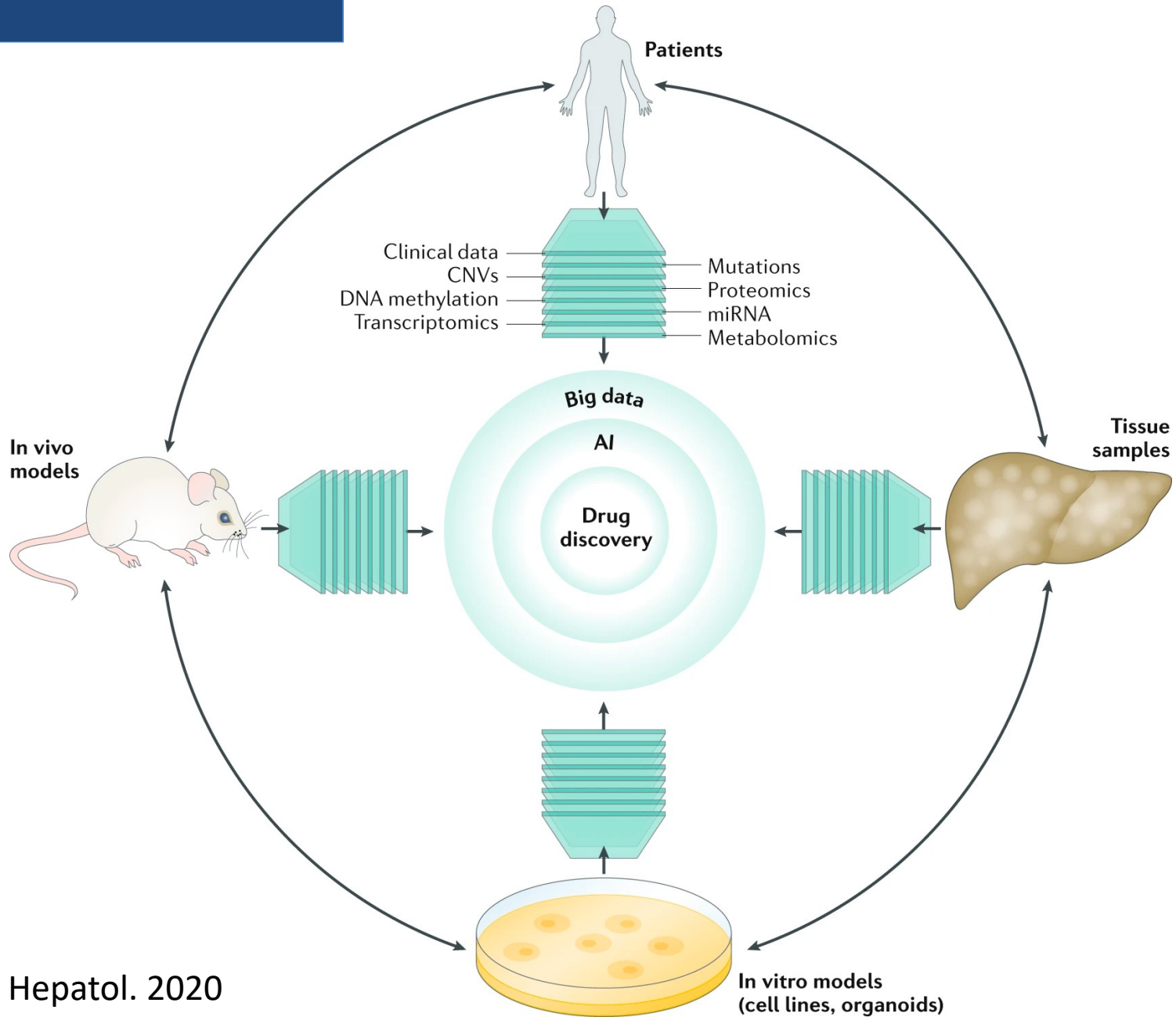
PharmGKB

Big Data in the clinic



I-SPY 2 TRIAL: Neoadjuvant and Personalized Adaptive Novel Agents to Treat Breast Cancer (I-SPY 2)

Big Data in translational research

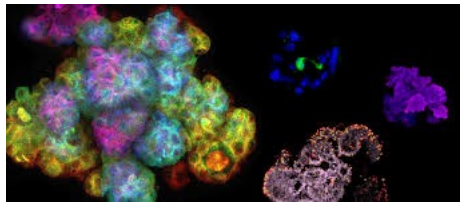
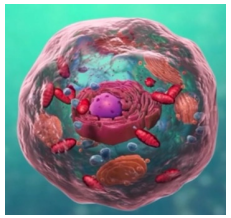


Chen B, Nat Rev Gastroenterol Hepatol. 2020

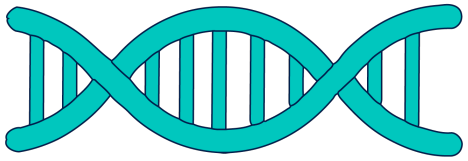
Big data you often encounter

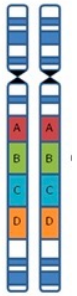
Molecular Profiling of biological systems with/without perturbation



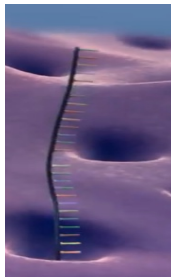



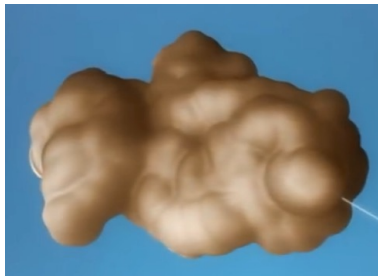
With/without perturbation
(e.g., drug treatment, gene knock-down)

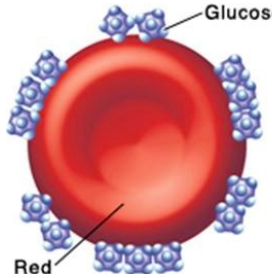

DNA



Chromosome
(copy number)

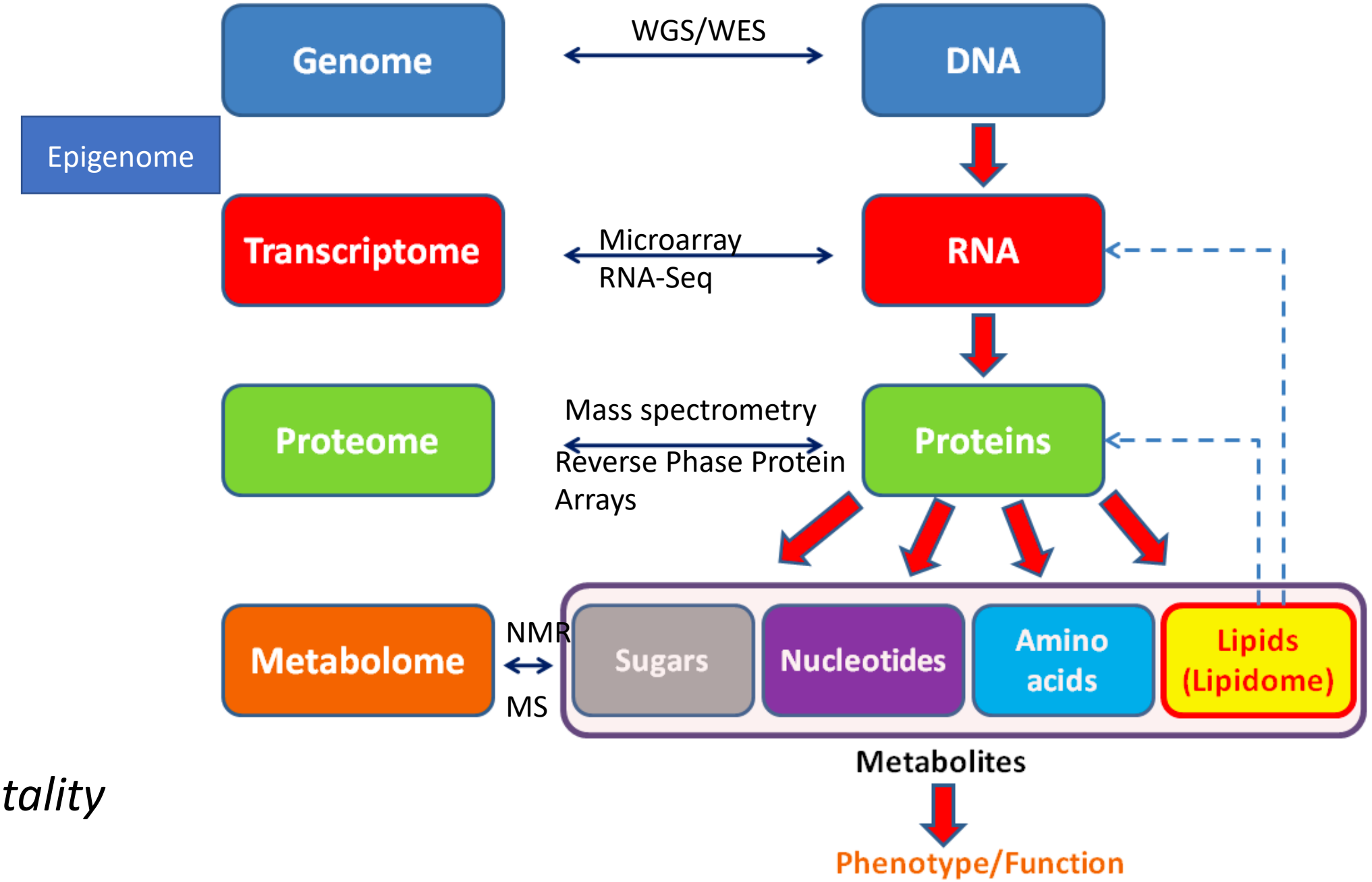

mRNA (gene)


Protein



metabolite

.....



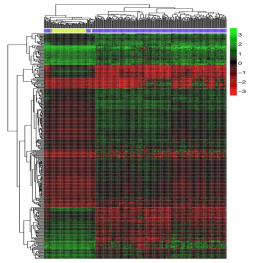
-Ome: *totality*

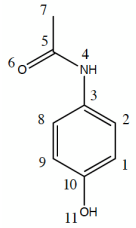
https://en.wikipedia.org/wiki/Genomics

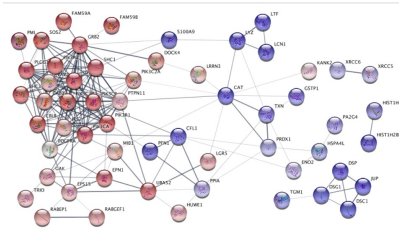
Commonly used big data types

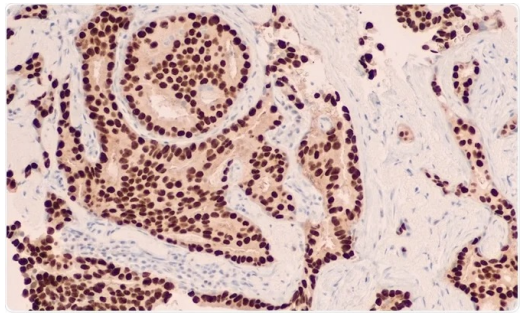
- OMICS
  - Transcriptomics
  - Proteomics
- Graph
  - Chemical structures
  - Protein-Protein Interactions
- Image
  - MRI
  - IHC

What do you see?

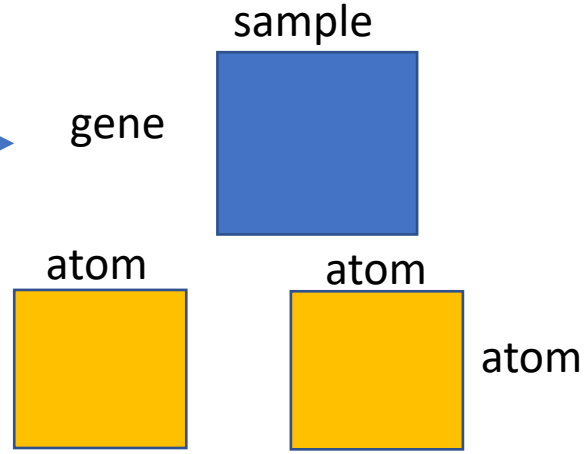




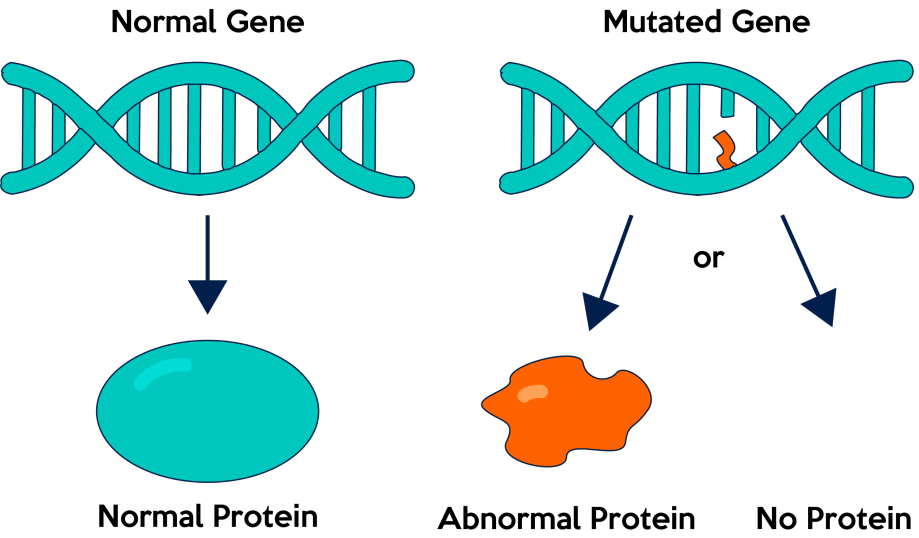


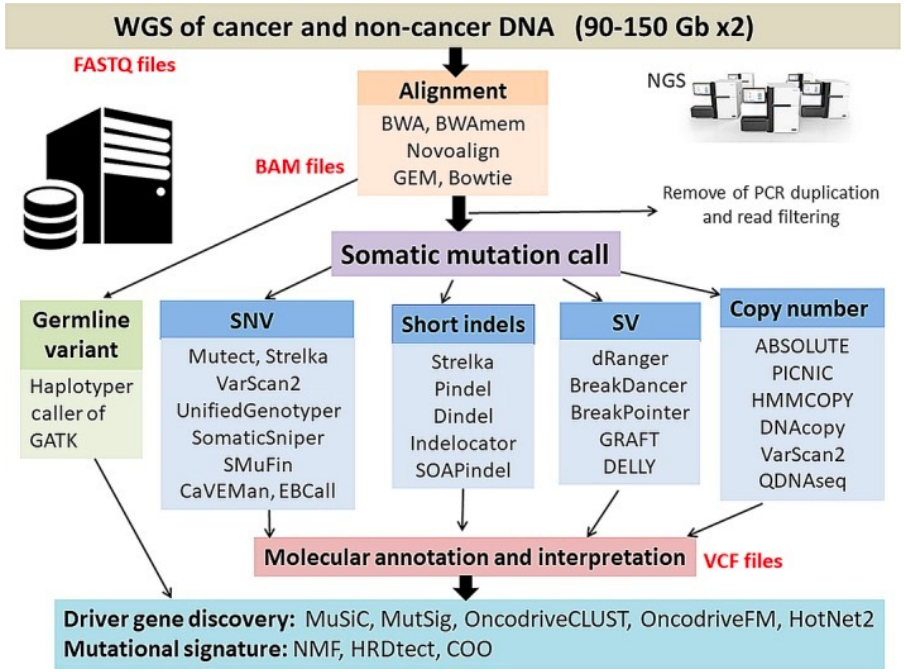


What does computer see?



Essentially a group of matrices





Genome/mutation

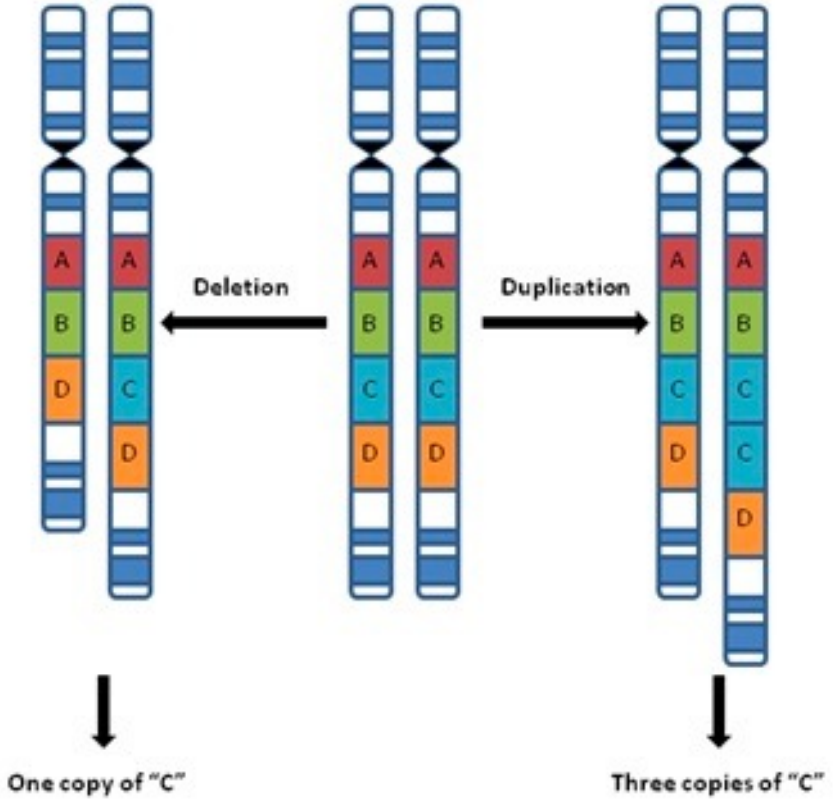
Hugo_Symb	Entrez_Gene	NCBI_Build	Chromosom	Start_positio	End_positio	Strand	Variant_Classifi	Variant_Typ	Reference_A	Tumor_Seq_	dbSNP_RS	dbSNP_Val_	Genome_Change	cDNA_Change	Codon_Change	Protein_Change
VPS13D	55187	37	1	12359347	12359347	+	Nonsense_Mut	SNP	C	A	NA	NA	g.chr1:12359347C>A	c.6122C>A	c.(6121-6123)tCa>tAa	p.S2041*
AADACL4	343066	37	1	12726308	12726322	+	In_Frame_Del	DEL	CTGGCGTG/	-	rs5821842	byFrequenc	g.chr1:12726308_12726	c.786_800delCTGGCGTG	c.(784-801)tcctggcgtgac	p.WRDAI263del
IFNLR1	163702	37	1	24484172	24484172	+	Silent	SNP	G	A			g.chr1:24484172G>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	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	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	c.277G>C	c.(277-279)Ggc>Cgc	p.G93R
MAST2	23139	37	1	46498028	46498028	+	Silent	SNP	C	T			g.chr1:46498028C>T	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	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	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	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	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	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	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	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	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	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	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	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	c.406A>G	c.(406-408)Agc>Ggc	p.S136G
PCBD1	5092	37	10	72645567	72645567	+	Silent	SNP	T	C			g.chr10:72645567T>C	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	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	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	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	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	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	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	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	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	c.13164G>C	c.(13162-13164)acG>ac	p.T4388T

1656 cell lines

19K genes

Mutation status
(1 or 0)

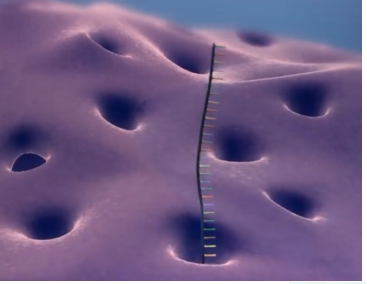
Genome/copy number variation



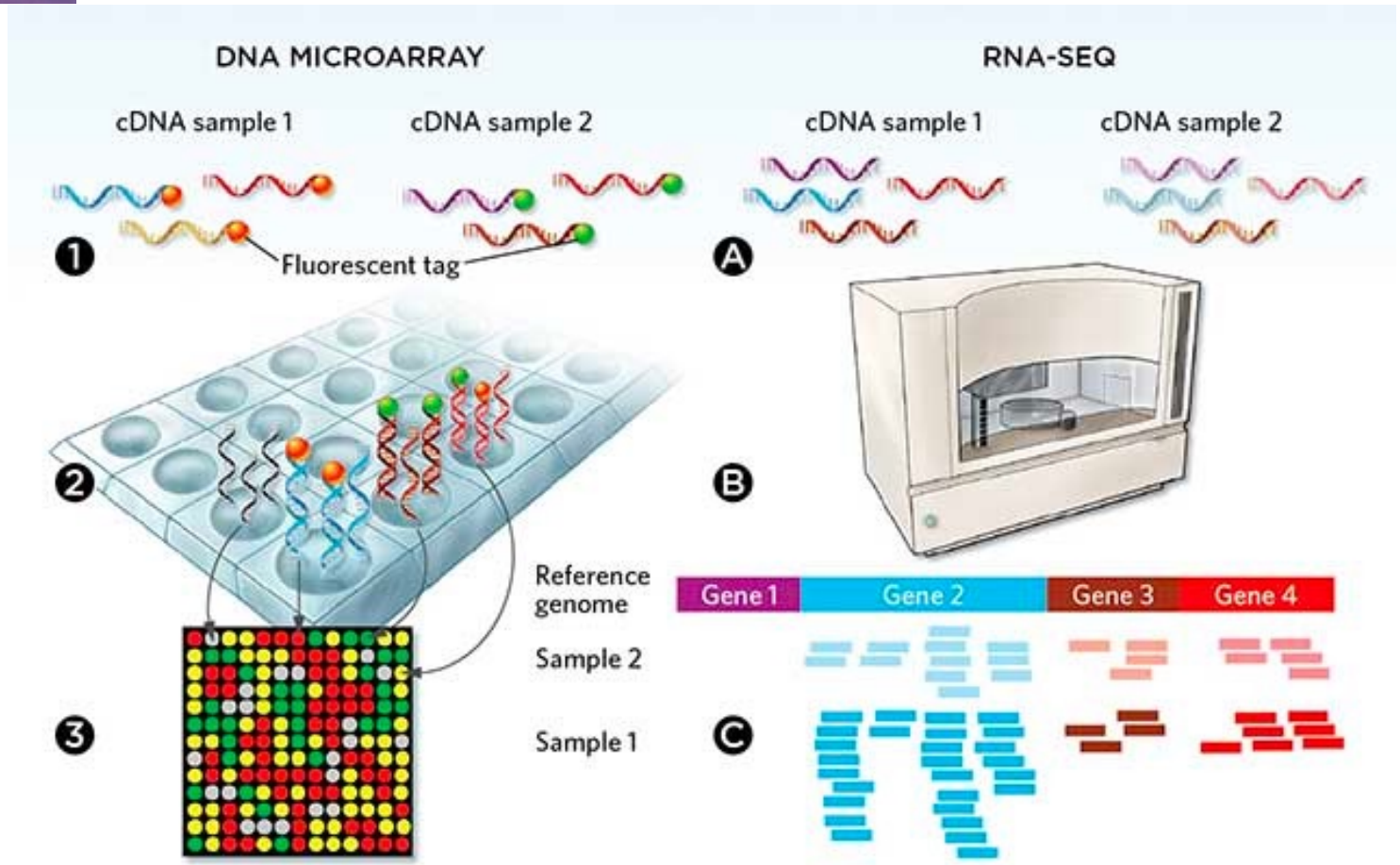
1657 cell lines

28K genes

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



Transcriptome



1210 cell lines

60K genes

Gene Expression (numeric)

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

cBioPortal FOR CANCER GENOMICS

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

Modify Query

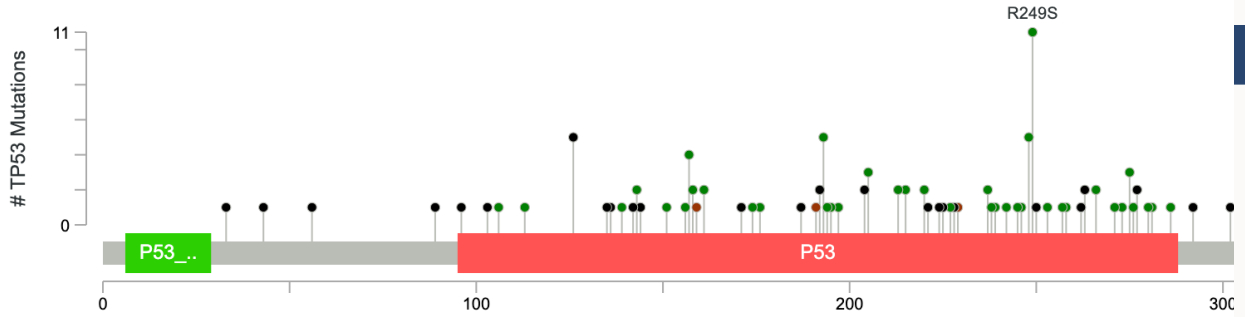
Liver Hepatocellular Carcinoma (TCGA, Firehose Legacy)
Samples with mutation and CNA data (366 patients/samples) - TP53

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.

cinoma, V600E, BRCA-UK, Campbell SEARCH

ClinVar Genomic variation as it relates to human health

Search ClinVar

Advanced search

About | Access | Submit | Stats | FTP | Help

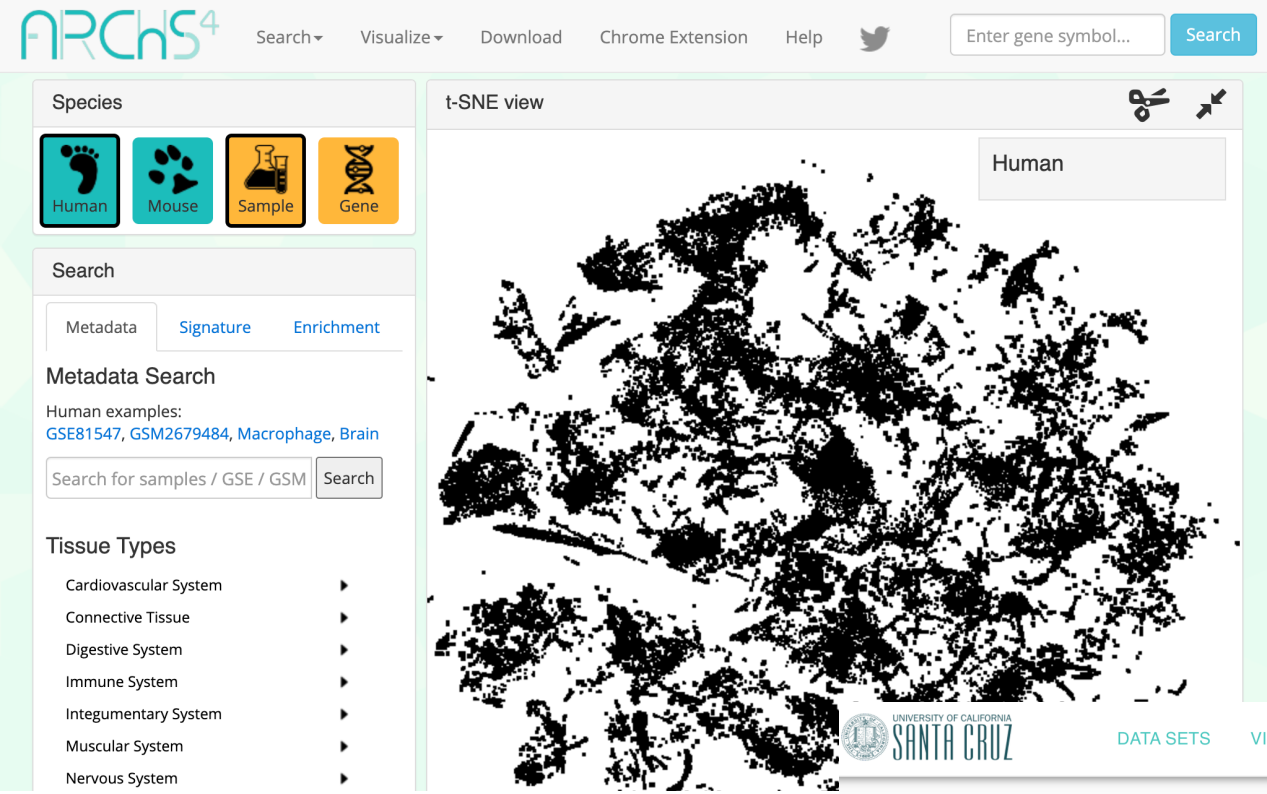
Was this helpful?

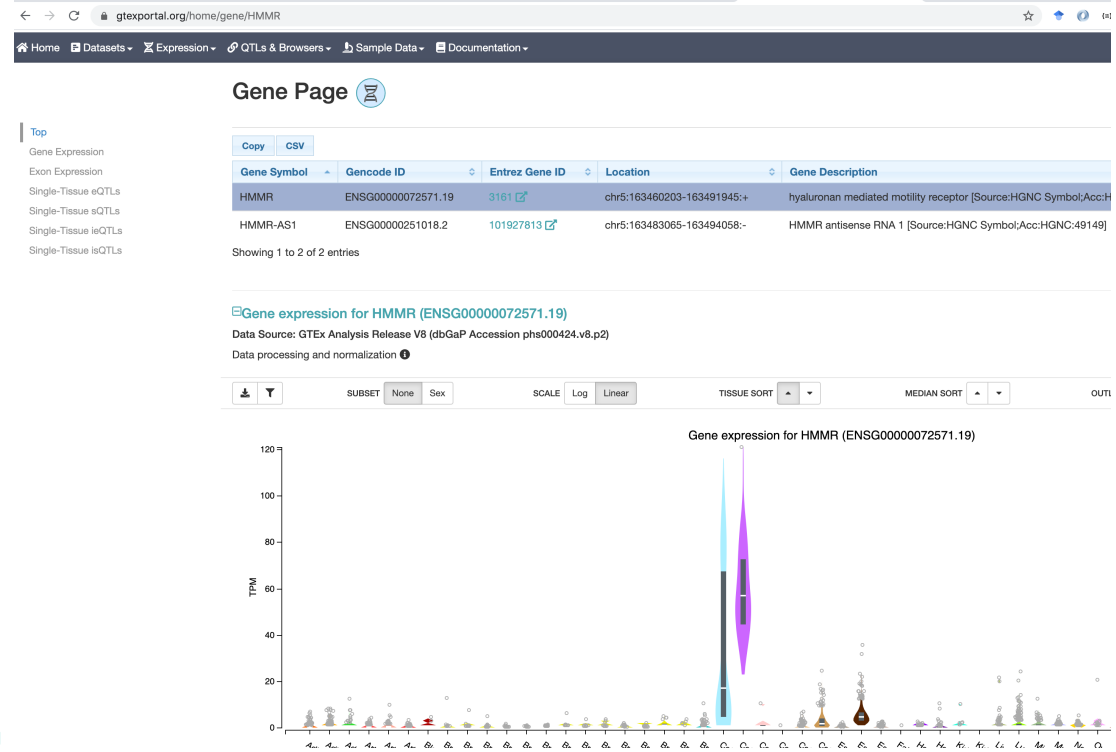
Follow | Print | Downloa

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


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
2
4
5
7c
1c
2
3
1d
2
4
5
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

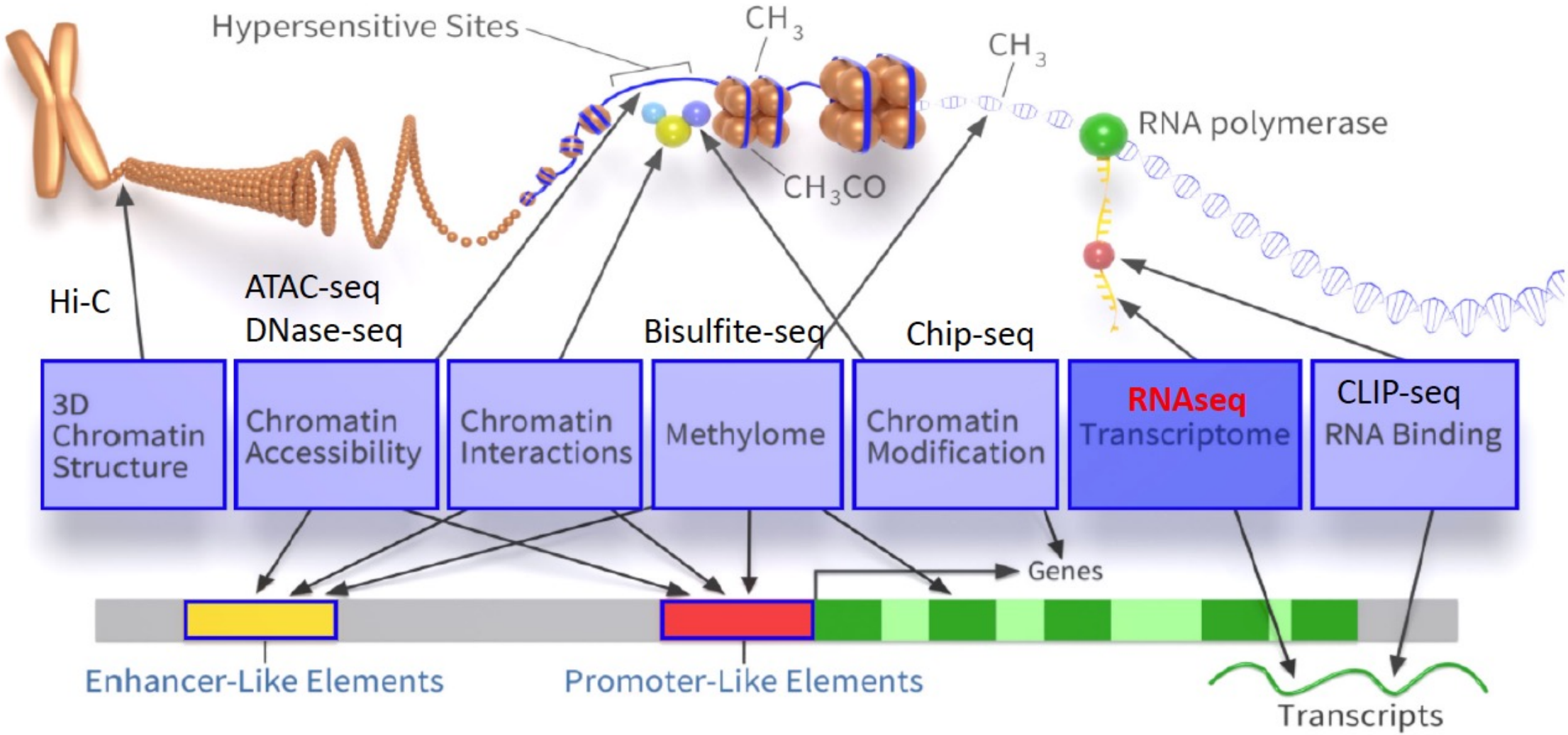
- Users FAQ
- Submitters FAQ
- Using your EGA account
- Contact Us
- EGA mailing list

ArrayExpress – function

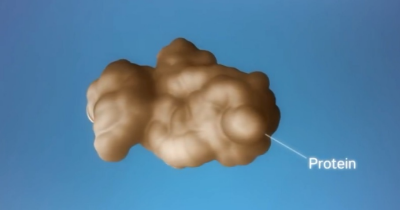
ArrayExpress Archive of Functional Genomics Data st… genomics experiments, and provides these data for reuse to the research community.

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

Browse ArrayExpress

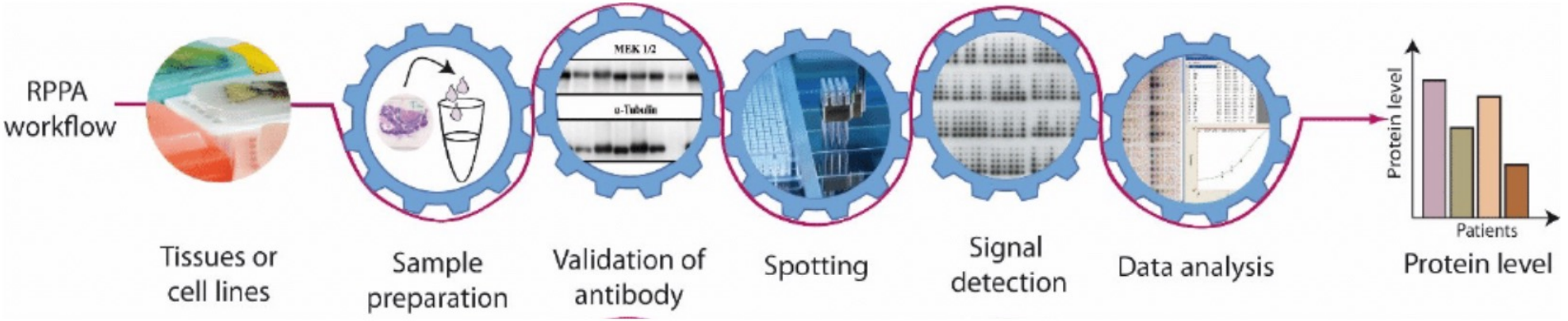


Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

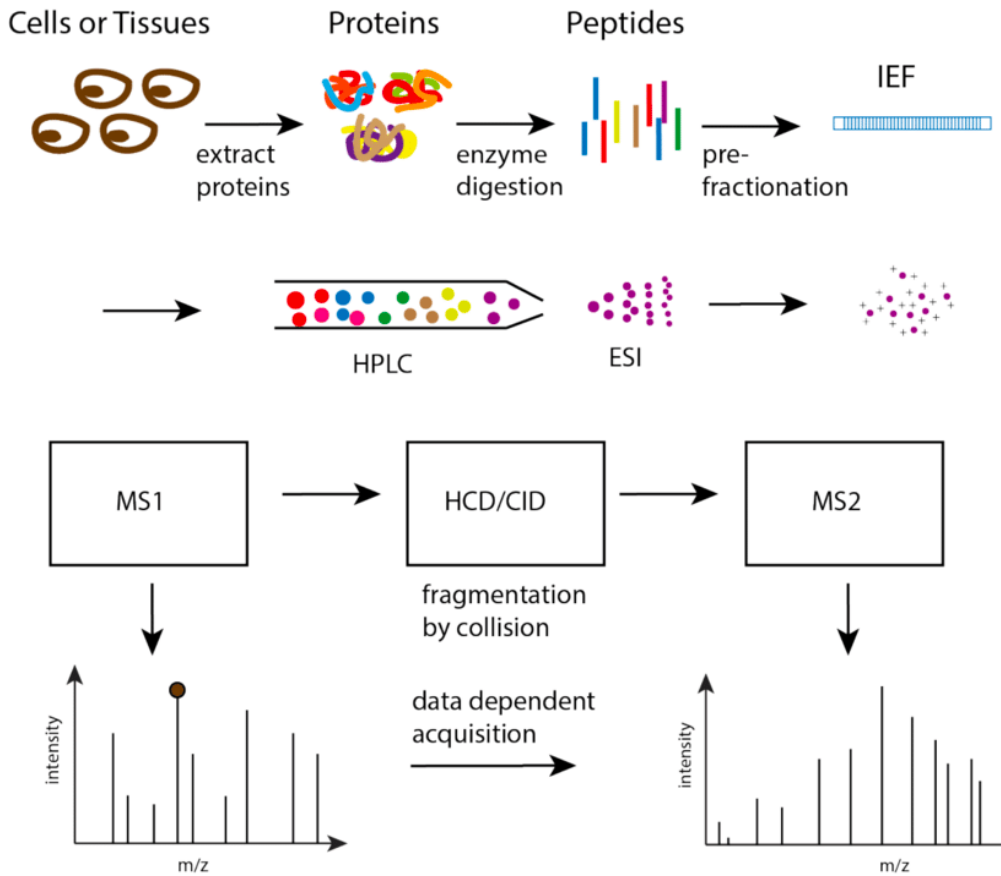


Proteome
Phosphoproteome

RPPA



MS



899 cell lines

214 proteins

Protein Expression
(numeric)

Stefanie Boellner, 2015

THE HUMAN PROTEIN ATLAS

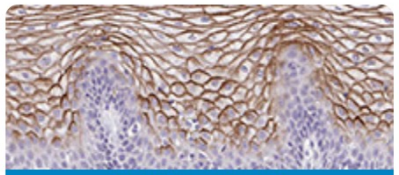
MENU HELP NEWS

SEARCH[i]

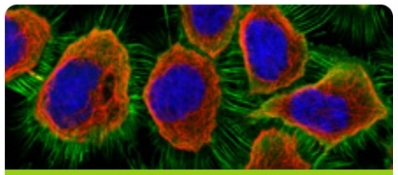
Search Fields »

e.g. RBM3, insulin, CD36

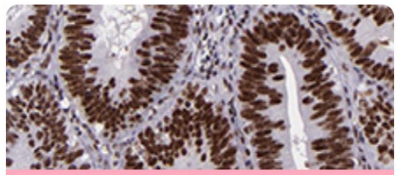
SARS-CoV-2 relevant proteins



TISSUE ATLAS



CELL ATLAS



PATHOLOGY ATLAS

PRIDE Archive

PRoteomics IDEntifications Database

Home Resources Tools Docs About

Search

NIH 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

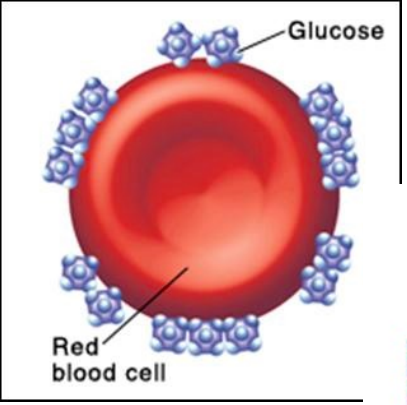
Proteogenomic Characterization Reveals Therap

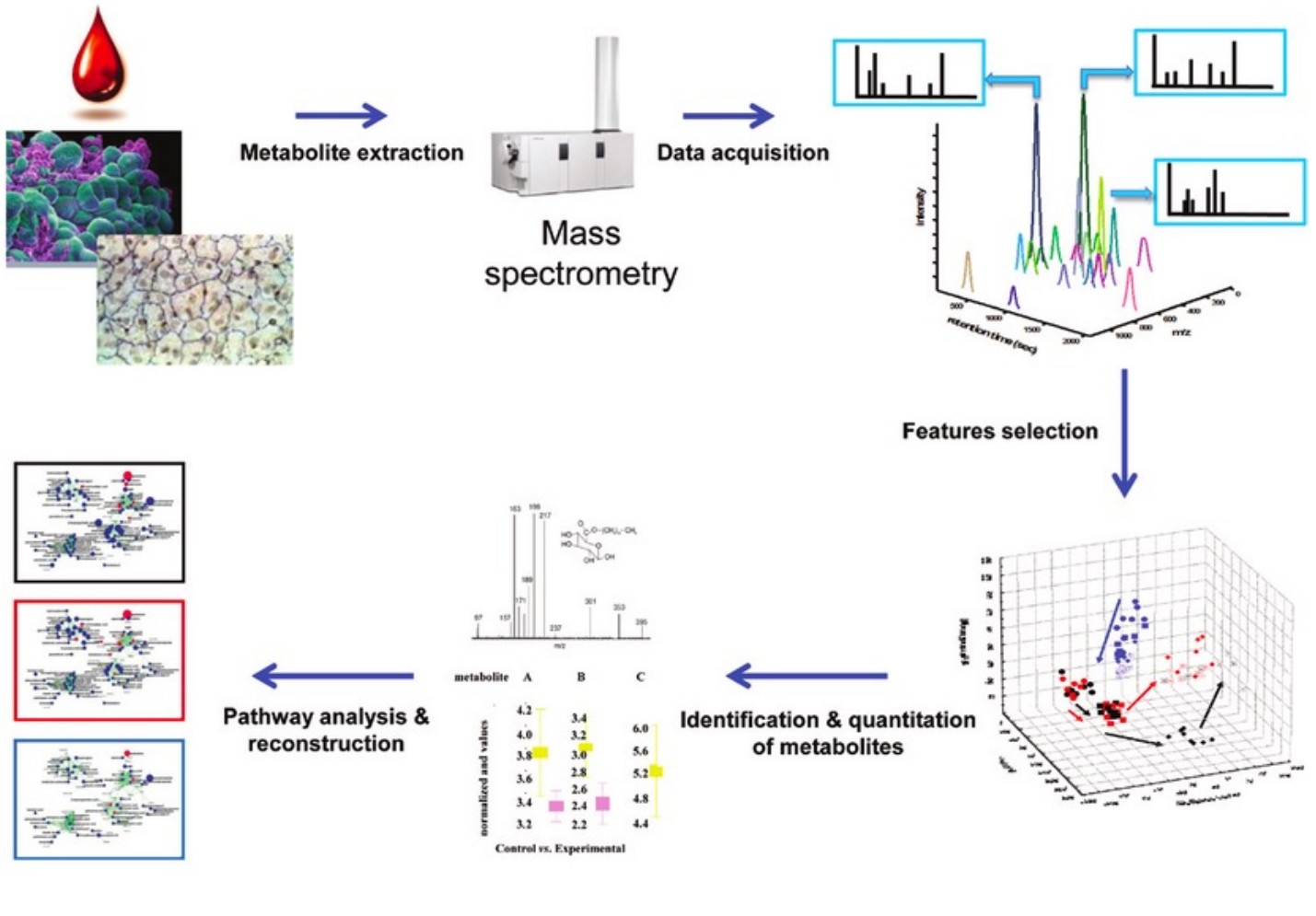
Michael A. Gillette, Shankha Satpathy, Song Cao,

May 2020

CPTAC Head and Neck Squamous Cell Carcinom

metabolome





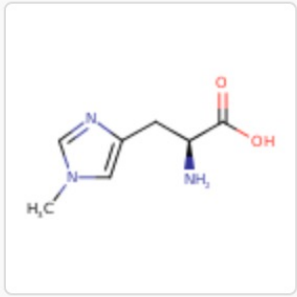
928 cell lines

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

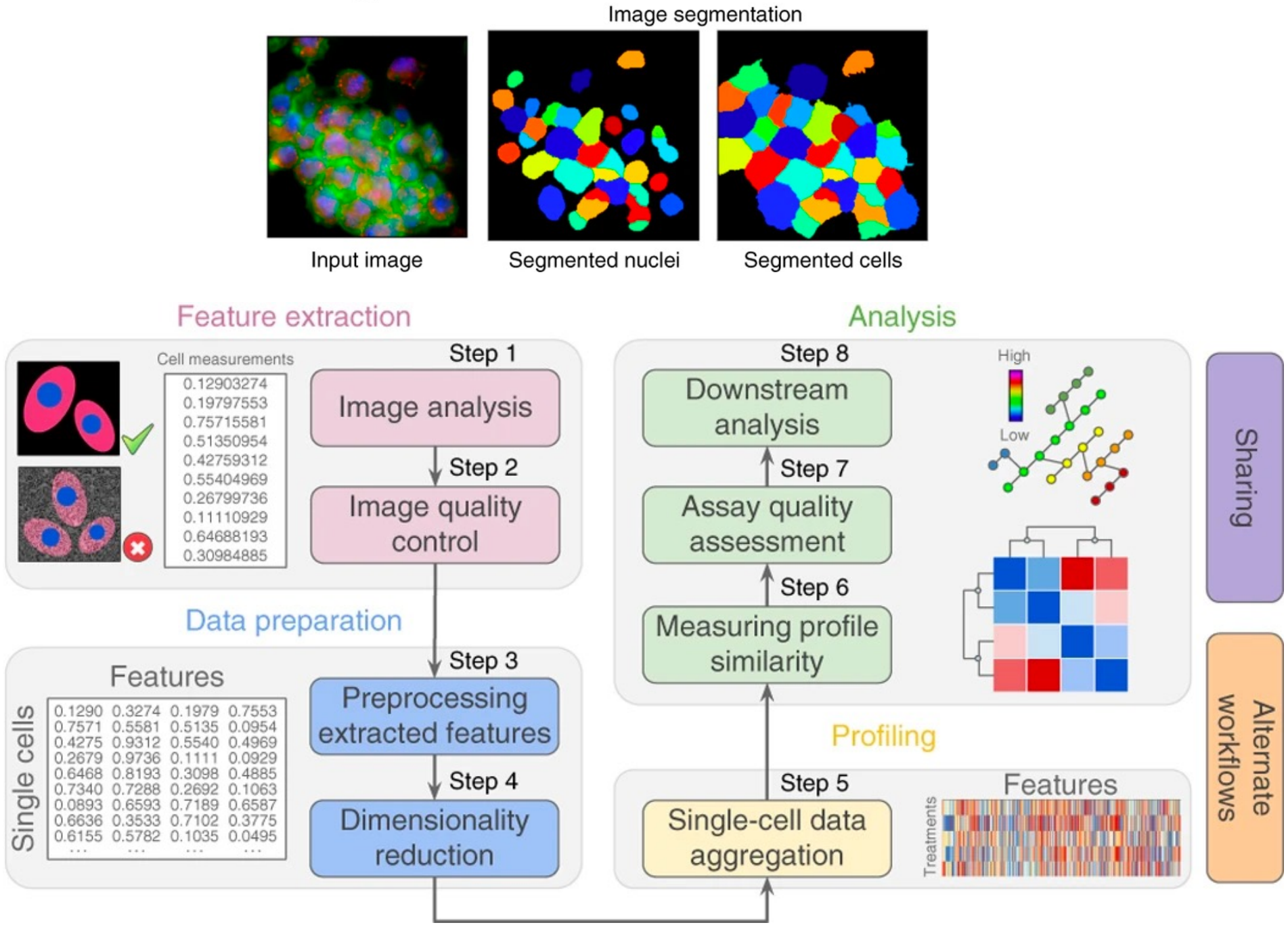
Organism

Organism Part

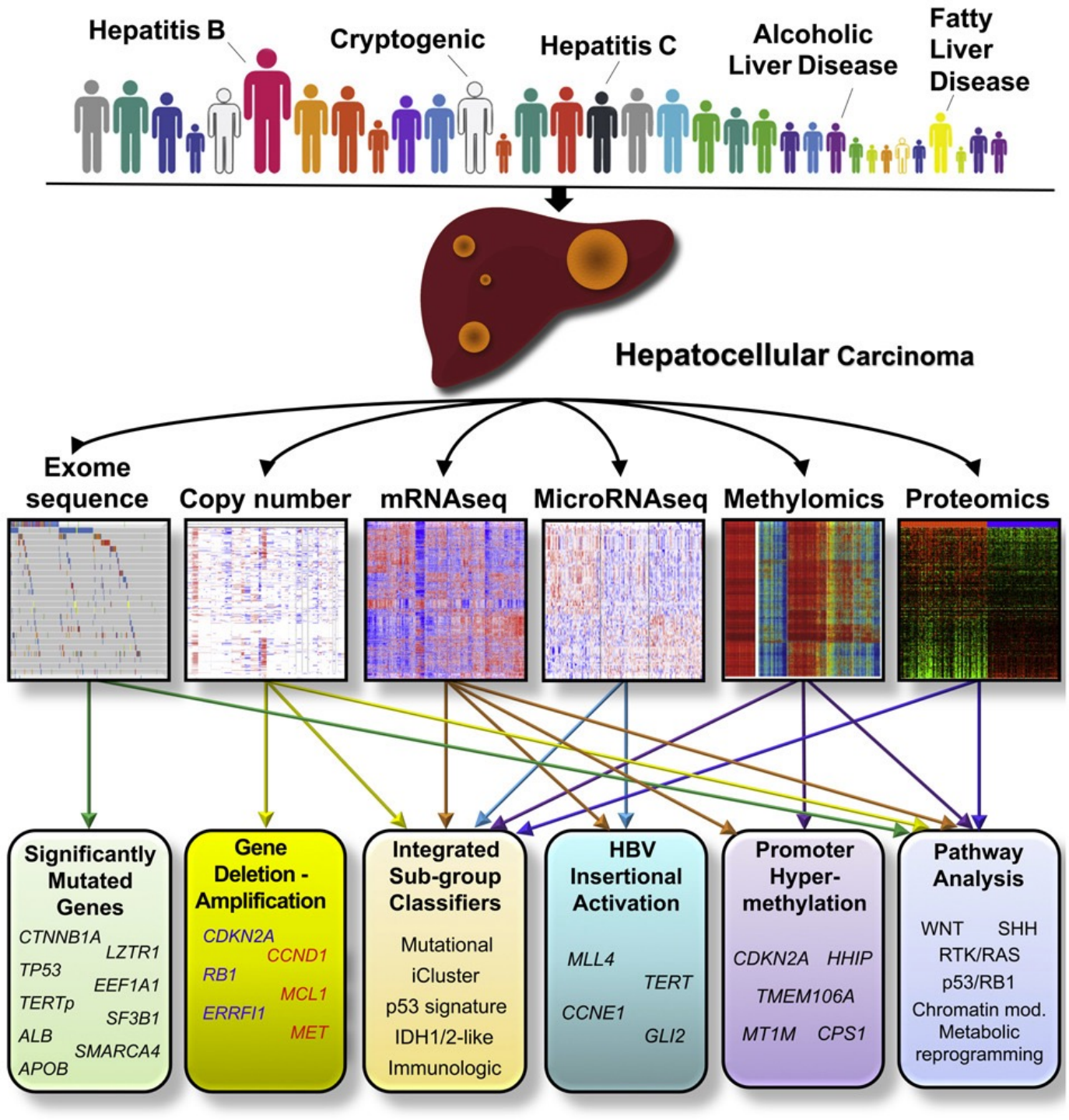
676 results , showing 1 to 10

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

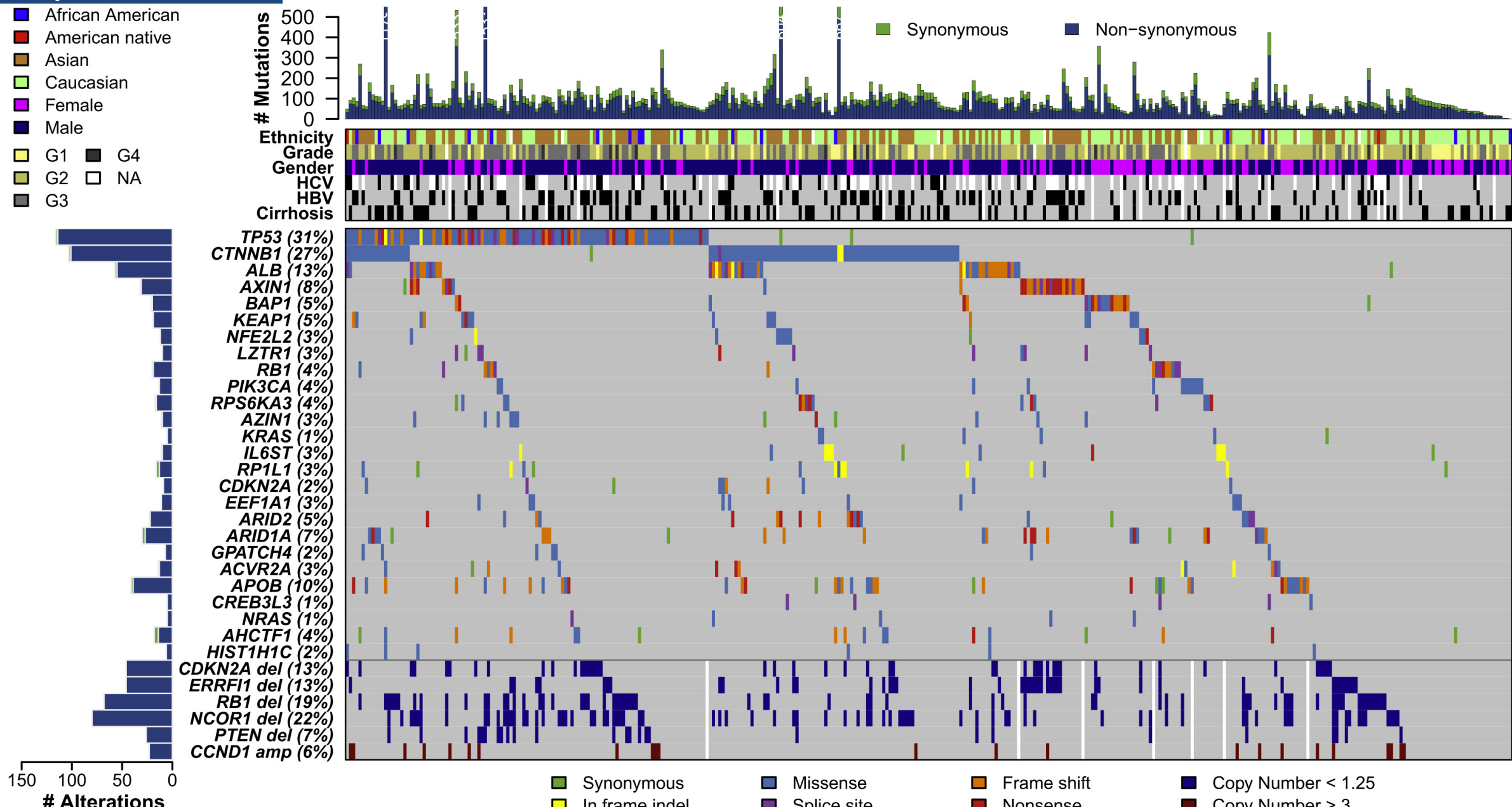
Study Identifier	MTBLS1217	Organi
Study Size	939.11MB	Study
Submitted by	Christoph Bueschl	

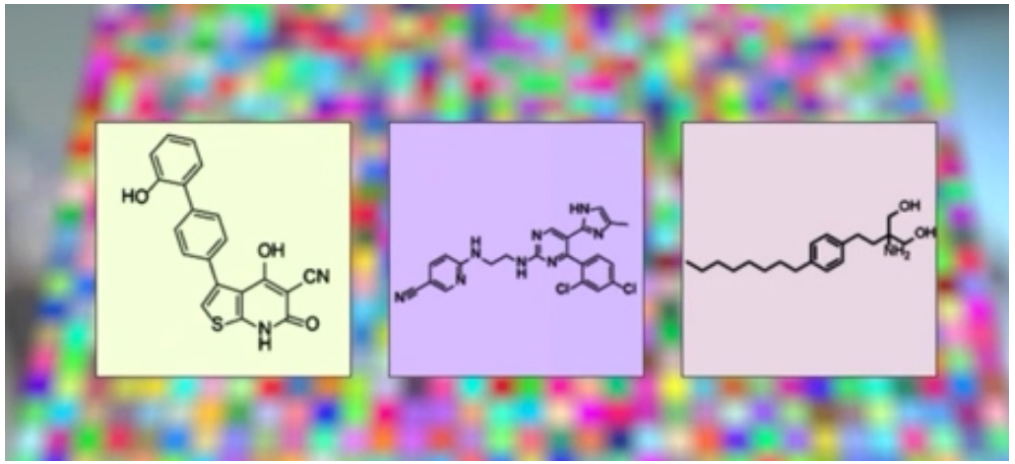


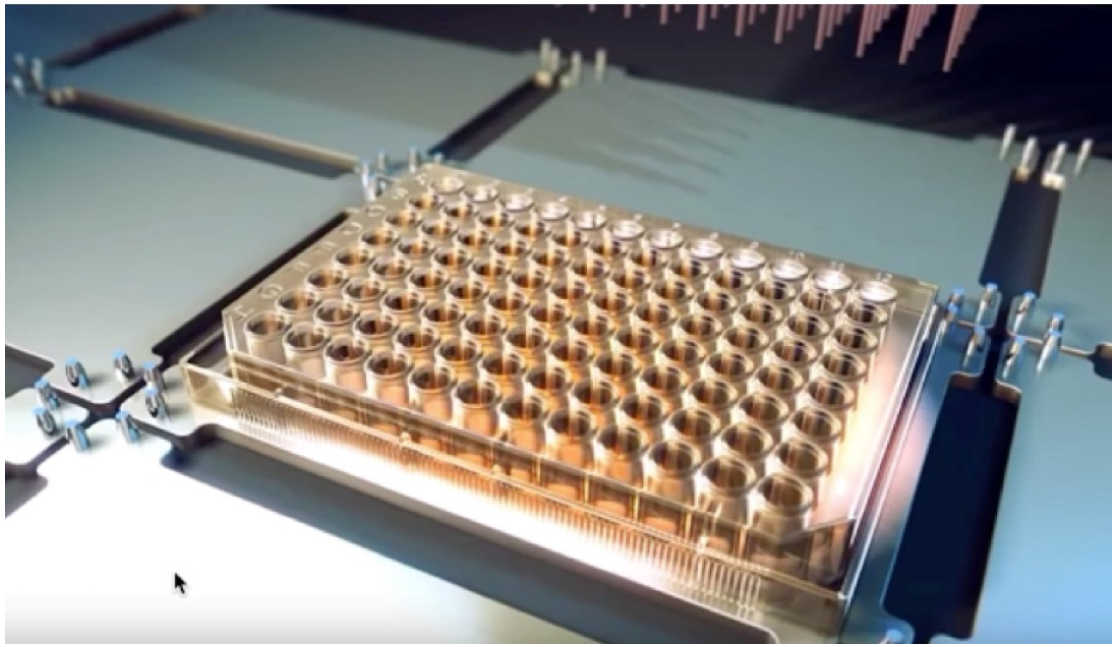
Juan C Caicedo, 2017


Hepatitis B
Cryptogenic
Hepatitis C
Alcoholic Liver Disease
Fatty Liver Disease
Hepatocellular Carcinoma
Exome sequence
Copy number
mRNAseq
MicroRNAseq
Methylomics
Proteomics
Significantly Mutated Genes
CTNNB1A
LZTR1
TP53
EEF1A1
TERTp
SF3B1
ALB
SMARCA4
APOB
Gene Deletion - Amplification
CDKN2A
CCND1
RB1
MCL1
ERRFI1
MET
Integrated Sub-group Classifiers
Mutational
iCluster
p53 signature
IDH1/2-like
Immunologic
HBV Insertional Activation
MLL4
TERT
CCNE1
GLI2
Promoter Hyper-methylation
CDKN2A
HHIP
TMEM106A
MT1M
CPS1
Pathway Analysis
WNT
SHH
RTK/RAS
p53/RB1
Chromatin mod.
Metabolic reprogramming

HCC Genetic Landscape





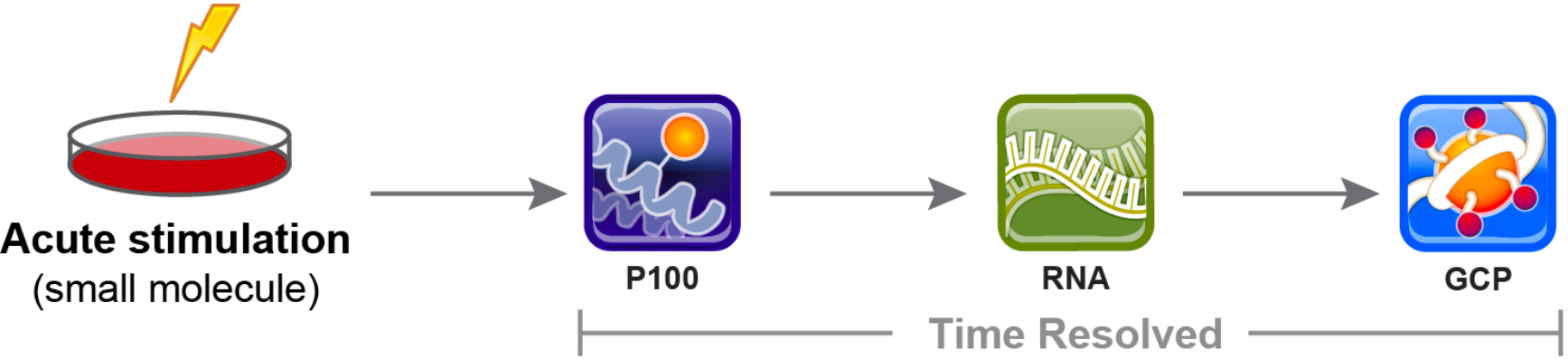


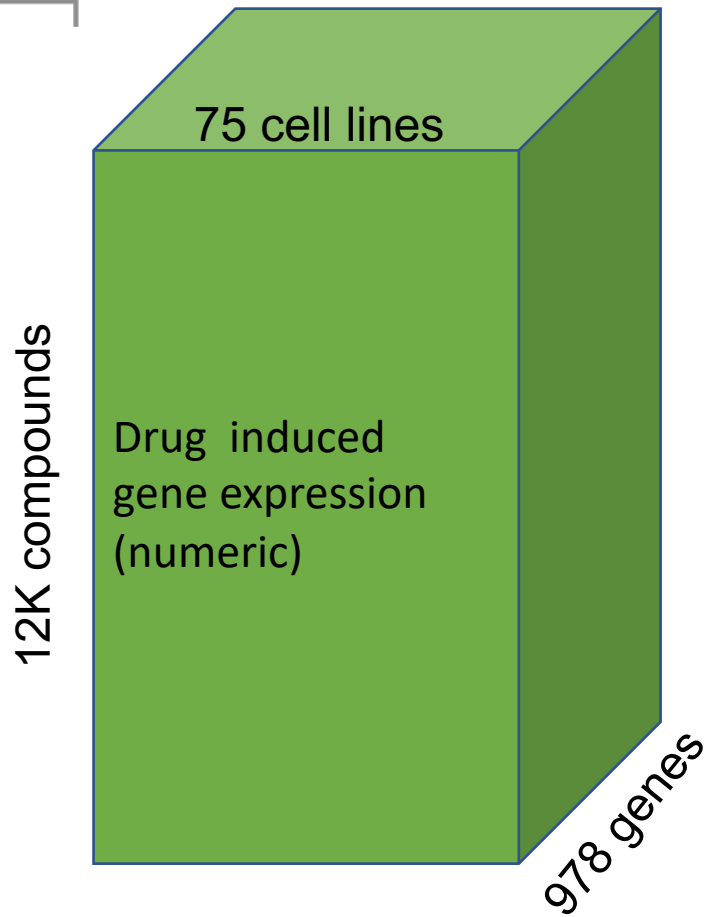
578 cell lines

5K compounds

Drug sensitivity score
(numeric)

Perturbation data (Drug sensitivity)





Perturbation data (Drug induced gene expression)

625 cell lines

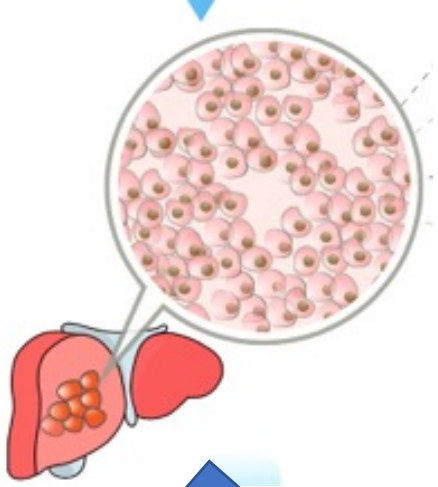
18K genes

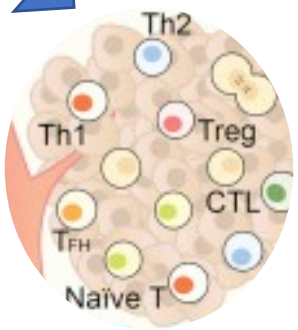
Gene essentiality score
(numeric)

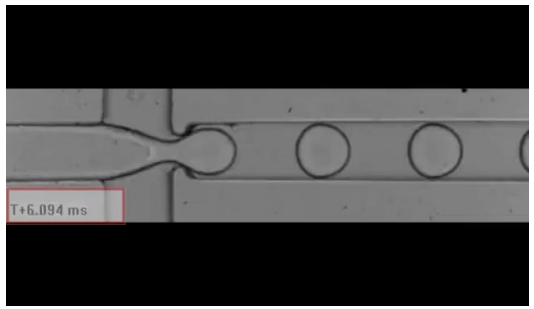
Perturbation data (Genome wide knock-out)

Demo

- https://depmap.org/portal/







20K patient tissues

60K genes

Gene Expression
(numeric)

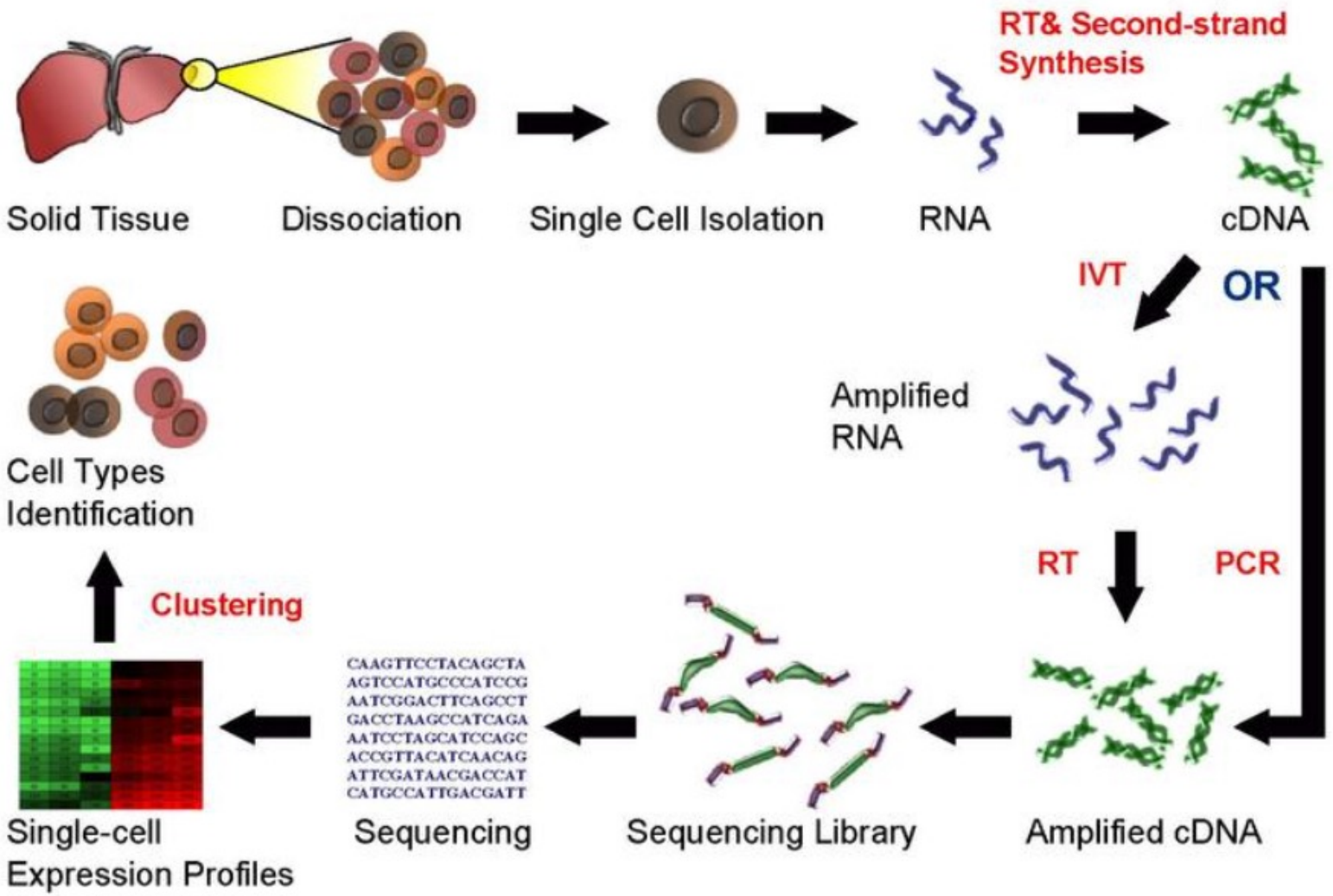
Millions of single cells

60K genes

Gene Expression
(numeric)

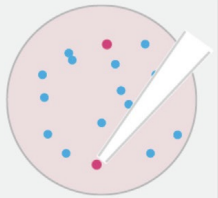
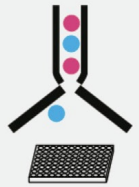
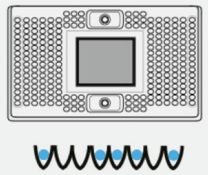
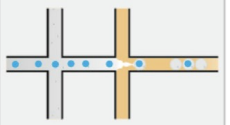
Advance topic: Single cell profiling

Work flow of single cell RNA-seq



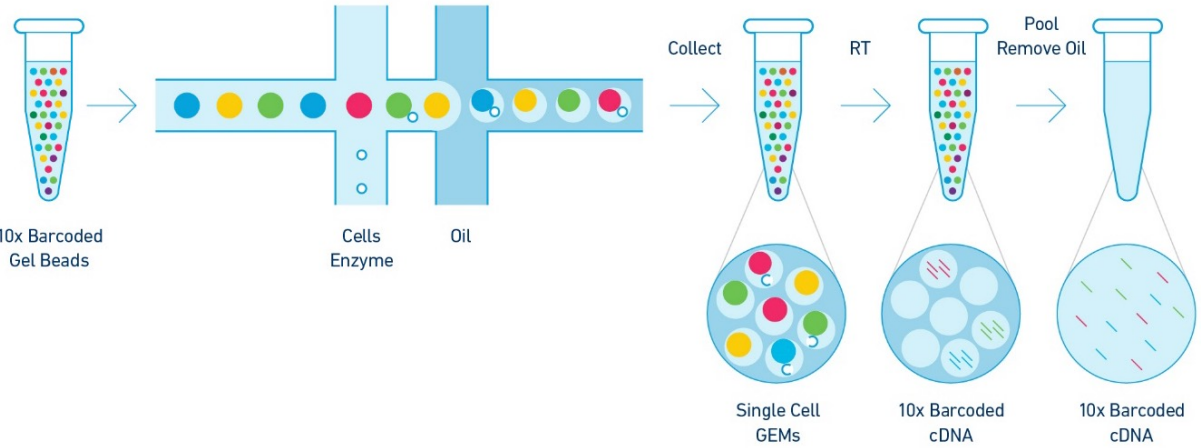
Advance topic: Single cell profiling

Overview about the scRNA-seq methods

	Micro-manipulation / Automated Pipetting	FACS	Microwell encapsulation	Droplet encapsulation
				
Cell Stress	Low	Moderate	Moderate	Moderate
Selection	Yes	Yes	No* / Yes++	No*
Doublet	Low	Low	Low-High	Moderate
Throughput	Low	Moderate	Moderate	High
Capture efficency	Low	Moderate	Moderate	Low-Moderate
Academic / Commerical scRNA workflow	- CellenONE (Cellenion)‡ - Smart-Seq2 (42)	- MARS-Seq (39) - Smart-Seq2 (42)	- C1 (Fluidigm) - ddSeq (Biorad / Illumina) - ICell8 (Clontech)++ - Rhapsody (BD)	- InDrop (1CellBio) - DropSeq (Dolomite-bio) - 10X (Chromium)
Example of use	Fragile rare cells	Rare cells based on phenotype or marking	Large cell numbers	Large cell numbers

	FACS		Microwell encapsulation				Droplet encapsulation		
	Smart-Seq2	MARS-Seq	C1	ddSeq	ICell8	Rhapsody	InDrop	DropSeq	10X
Singlet Capture efficency	82%	92%	39%	2.6%	37%++	Not reported	7%	Not reported	50%
Doublet rate	Not reported	2.27%	3-30%	5.8%	1.3-4%	0.6%	4%	0.36-11.3	1.6-3%
Reference	42	39	37 FWP	PB	PB	PB	36	37	26

‡Automated pipetting system
*Preselection or enrichment can be performed prior
++Only reagents added to wells containing singlets, determined by system
FWP: Fluidigm white paper
PB: Product brochure / manual

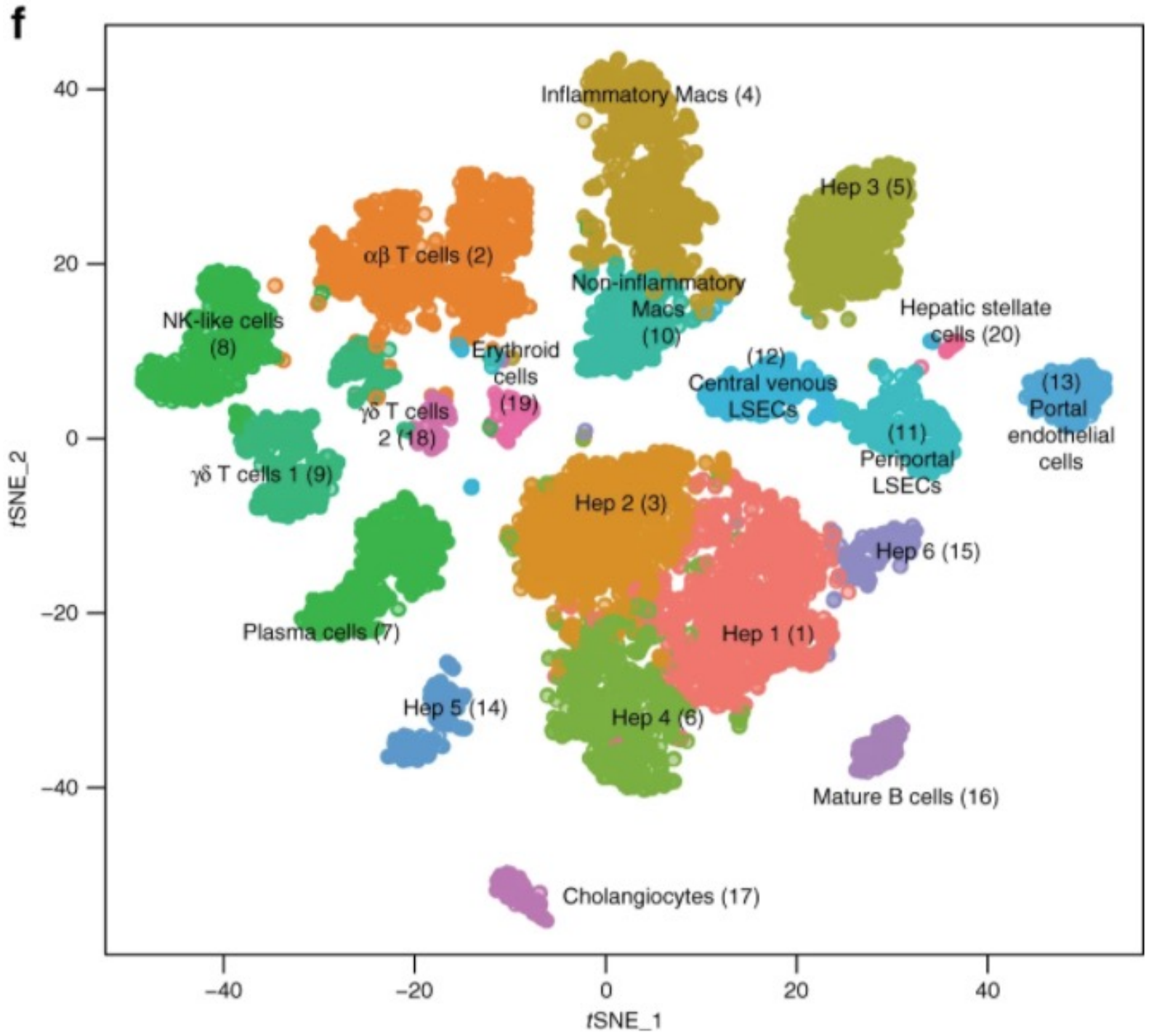


Advance topic: Single cell profiling

Nguyen et al., 2018, Front. Immunol.

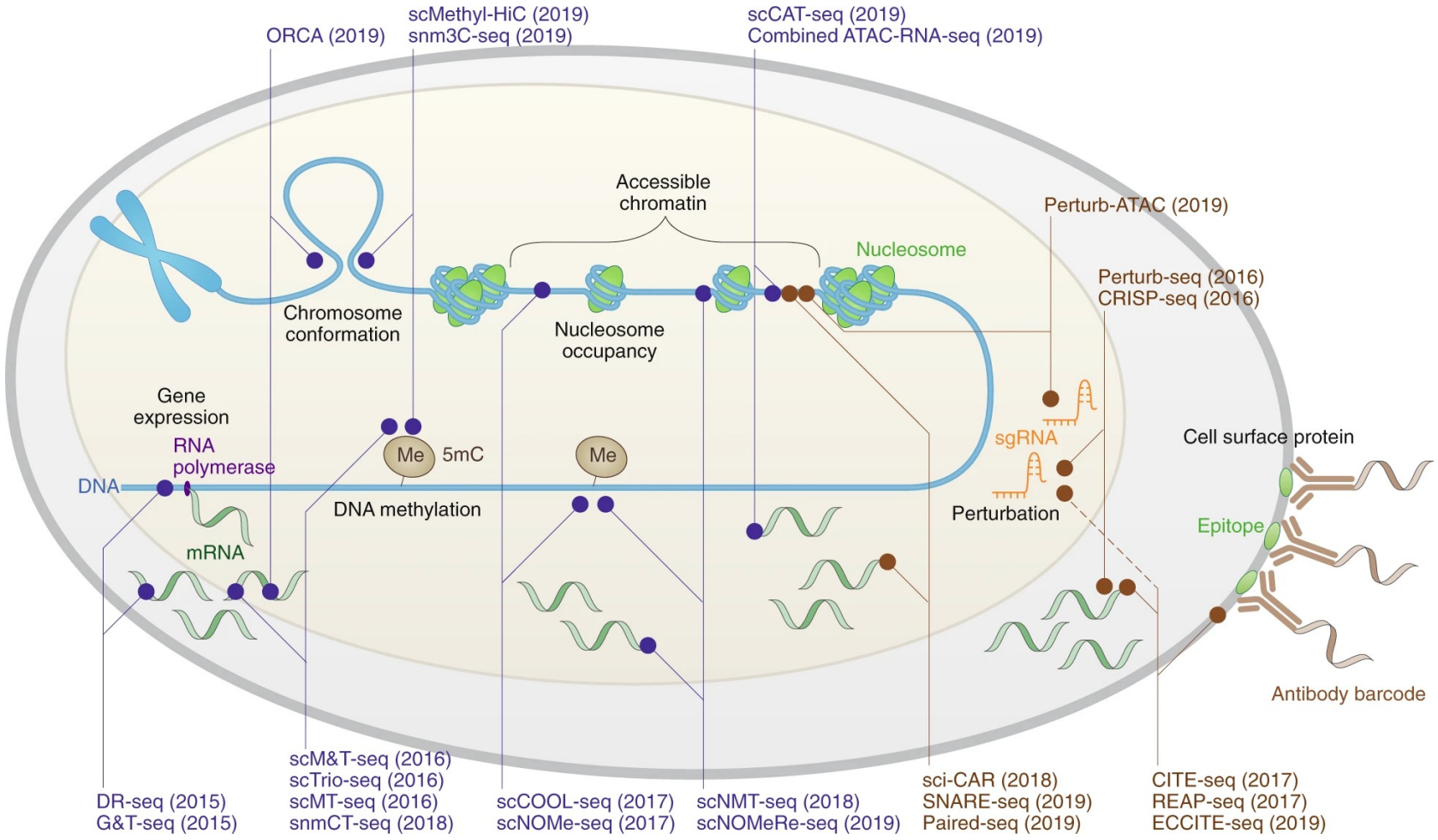
Main topics in single cell RNASeq

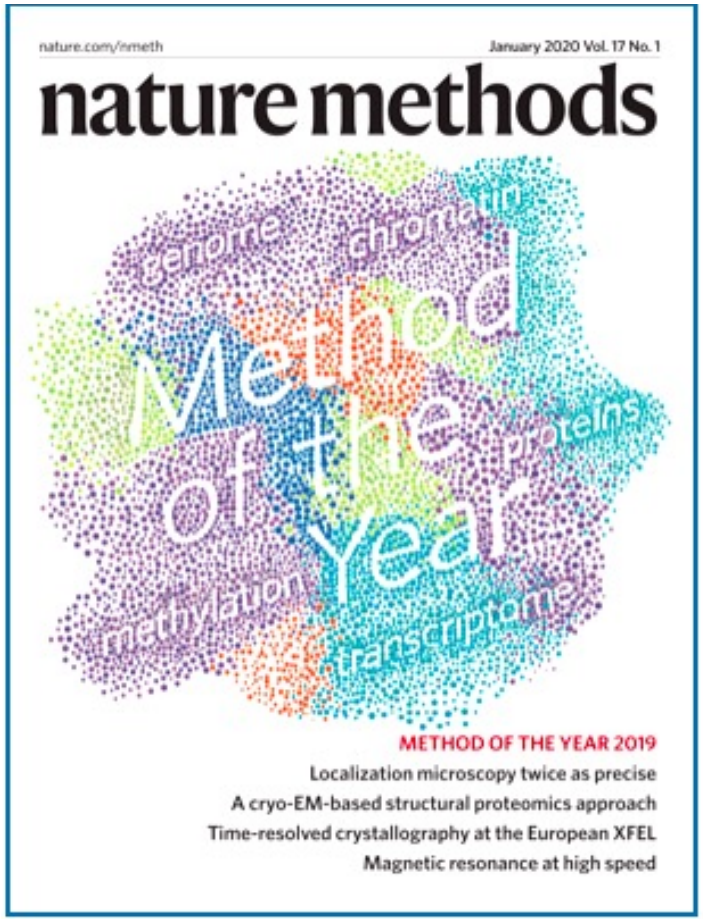
- Cell type identification
- Cell development
- Biomarker discovery



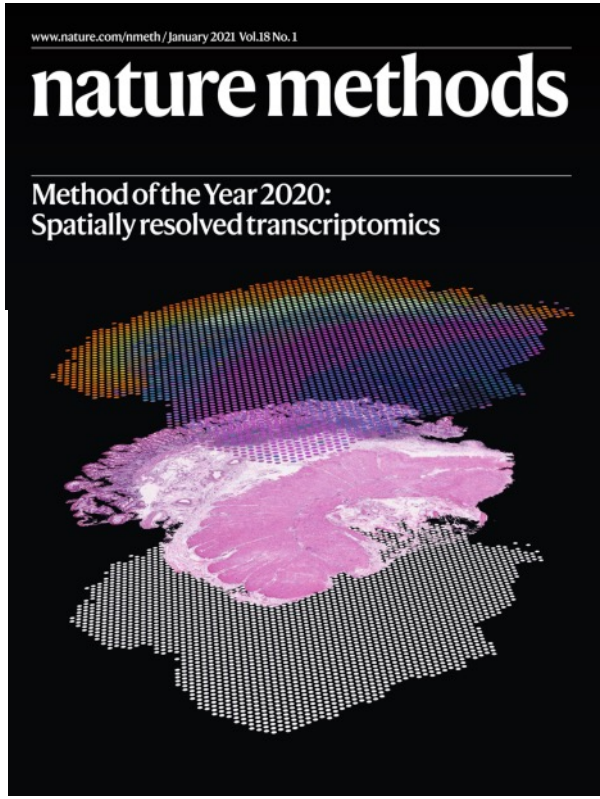
Advance topic: Single cell profiling

Multi-modal single cell

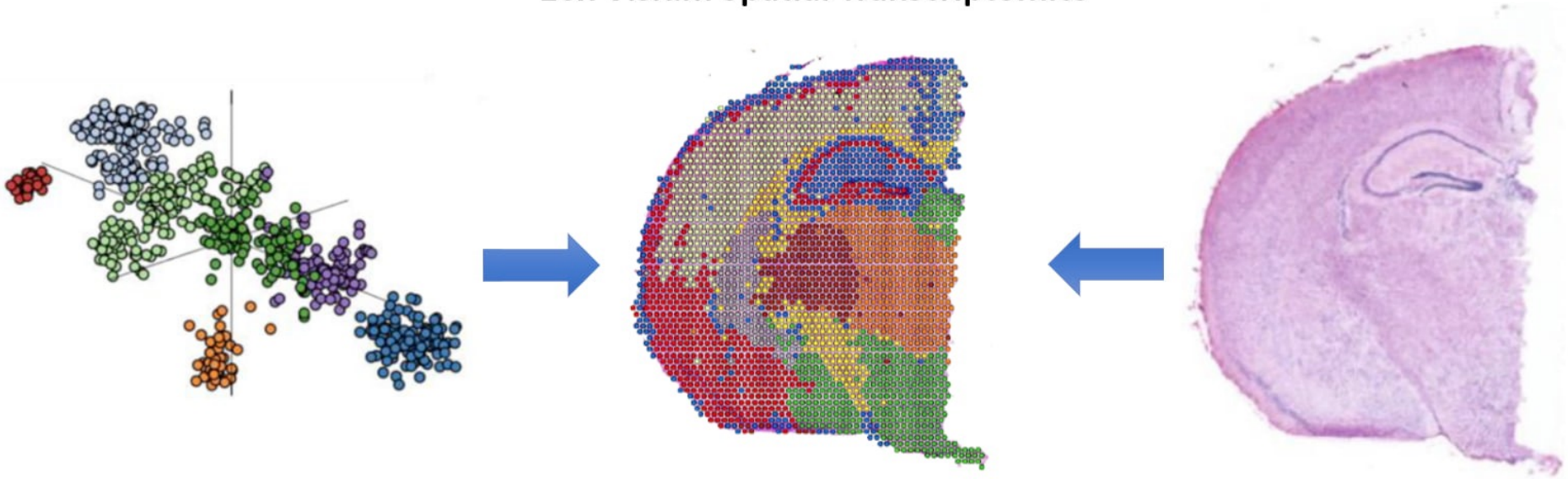




Spatial biology



10x Visium Spatial Transcriptomics



Single Cell Gene Expression | **Spatially Resolved Gene Expression** | **Tissue Section**

Adapted from 10x Genomics

Demo

- https://singlecell.broadinstitute.org/single_cell

Demo (Chemical Biology)

- https://pubchem.ncbi.nlm.nih.gov/
- https://www.ebi.ac.uk/chembl/
- https://www.alphafold.ebi.ac.uk/

AlphaFold Protein Structure Database

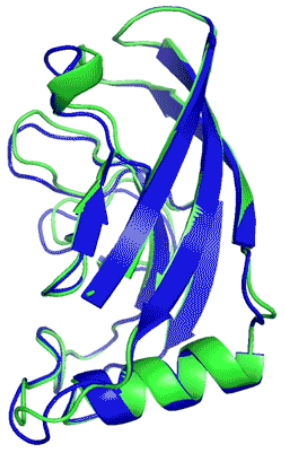
Developed by DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism BETA

Search

Examples: Free fatty acid receptor 2 | At1g58602 | Q5VSL9 | E. coli Help: AlphaFold DB search help

T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result

Computational prediction

Data properties

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

Bioinformatics

Big Data + AI ---> discovery

- Search
  - Search expression of MYC in lung cancer
- Compare
  - Is MYC expression higher in lung cancer than liver cancer?
- Predict
  - Predict drug sensitivity using pharmacogenomics data

Resources good for your publication

- GitHub for code sharing
- Figshare/synapse for data sharing
- Zenodo for code and data sharing

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
  - https://liulab-dfci.github.io/bioinfo-combio
- Troubleshooting
  - https://stackoverflow.com/
  - https://www.biostars.org/
- GitHub code repository
  - https://github.com/Bin-Chen-Lab/Awesome_BigData_AI_DrugDiscovery

ChatGPT would be the best resource for bioinformatics learning and coding.