

# 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 at Chen lab)
- Data manipulation and visualization (Dr. Eugene Chekalin, postdoc om 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  
<http://binchenlab.org>

# 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](mailto:Bin.Chen@hc.msu.edu) @DrBinChen

<http://binchenlab.org>

# 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

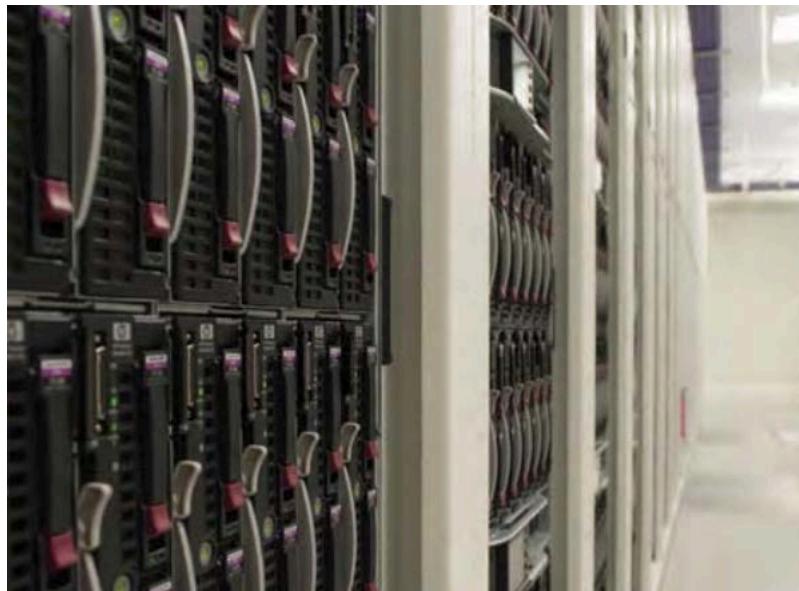
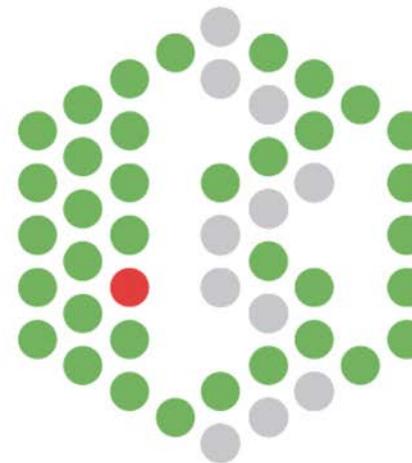
Work experience at Novartis, Pfizer and Merck

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

307 petabytes =



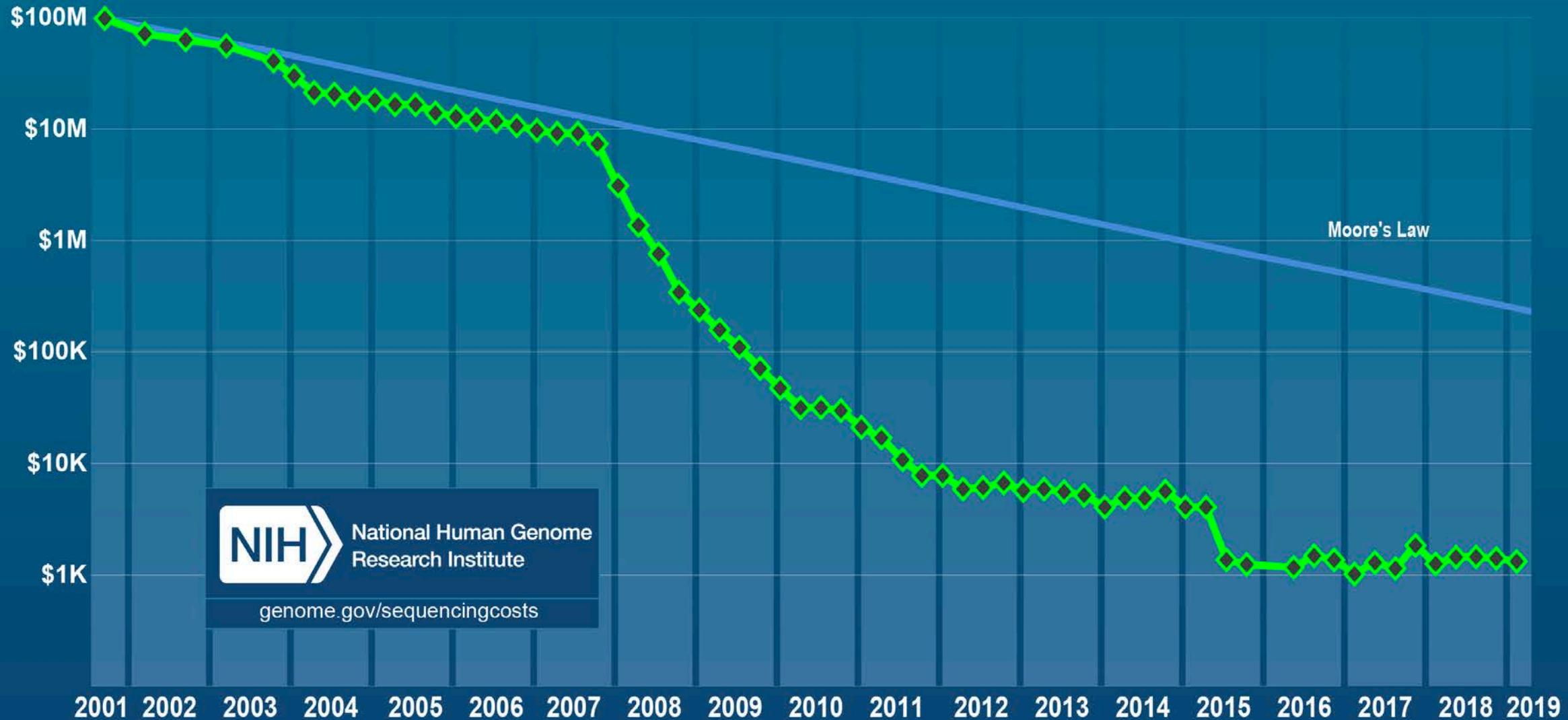
x 150,000

## DATA GROWTH BY EMBL-EBI DATA RESOURCE

Volume of data (megabytes) per year (2008–2019)

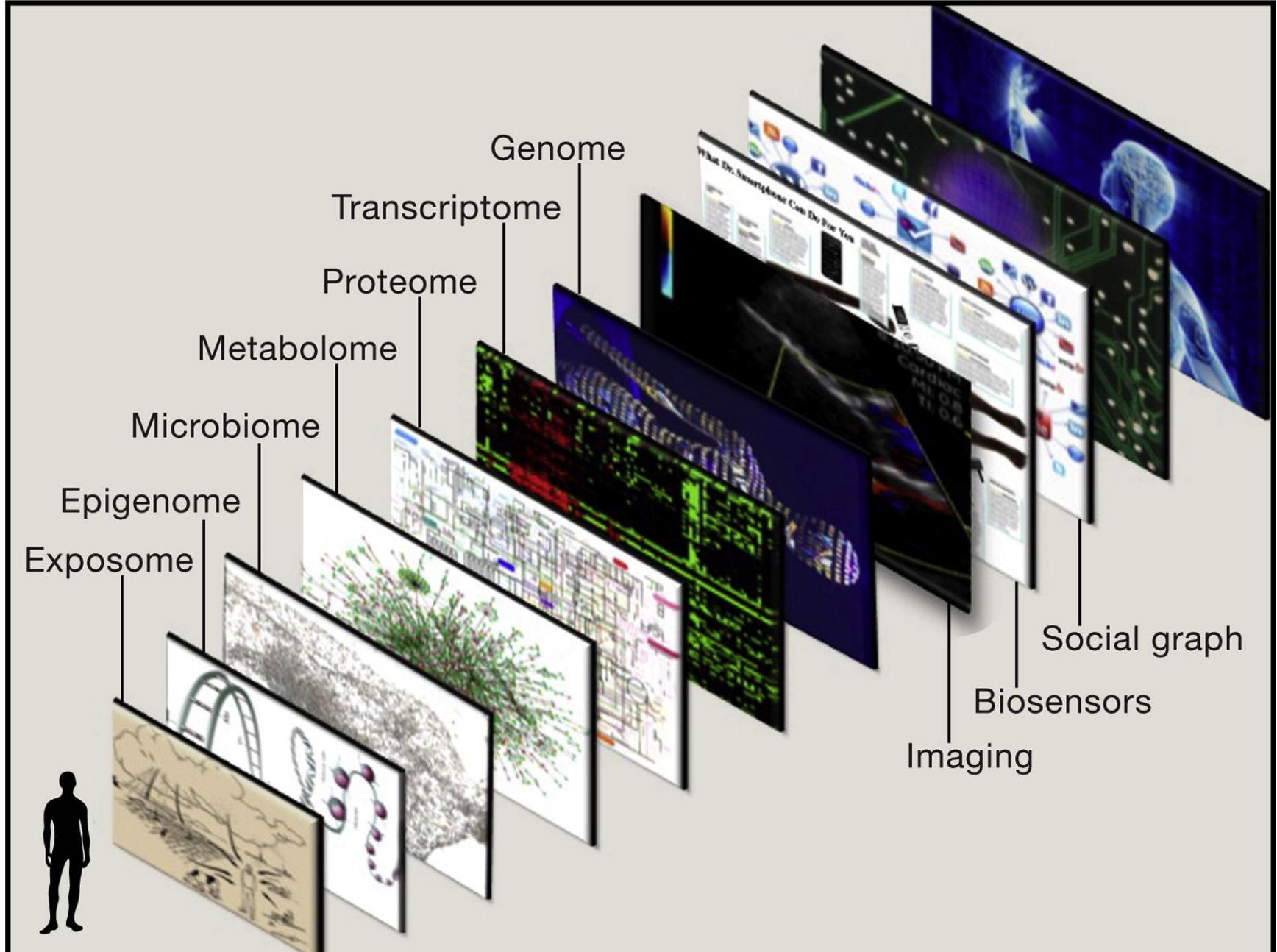


## *Cost per Genome*



National Human Genome  
Research Institute

[genome.gov/sequencingcosts](http://genome.gov/sequencingcosts)

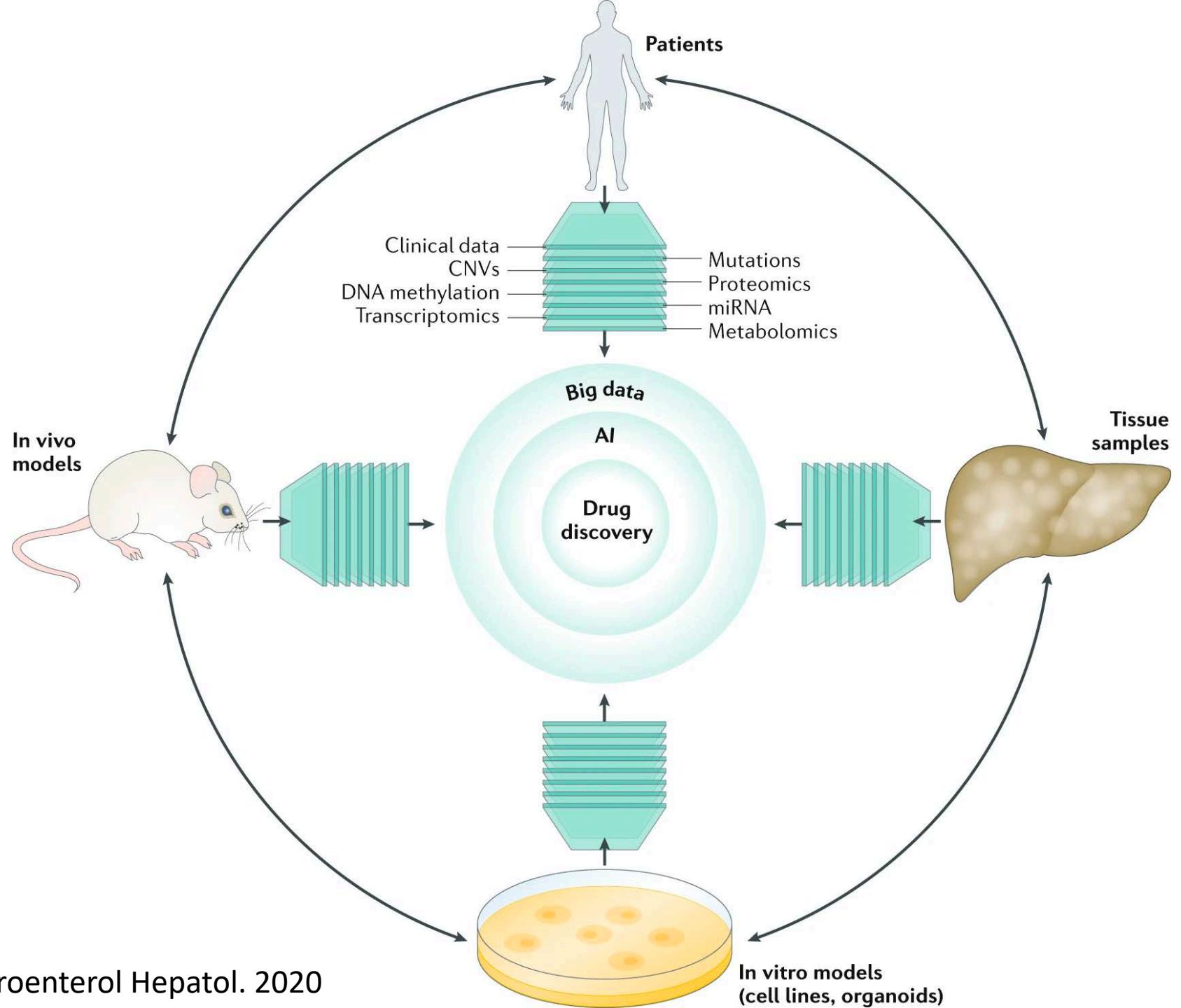


| Scale (unit)                   | Imaging technology                                                                                                                                 | Use                                                                             |
|--------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------|
| 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                            |

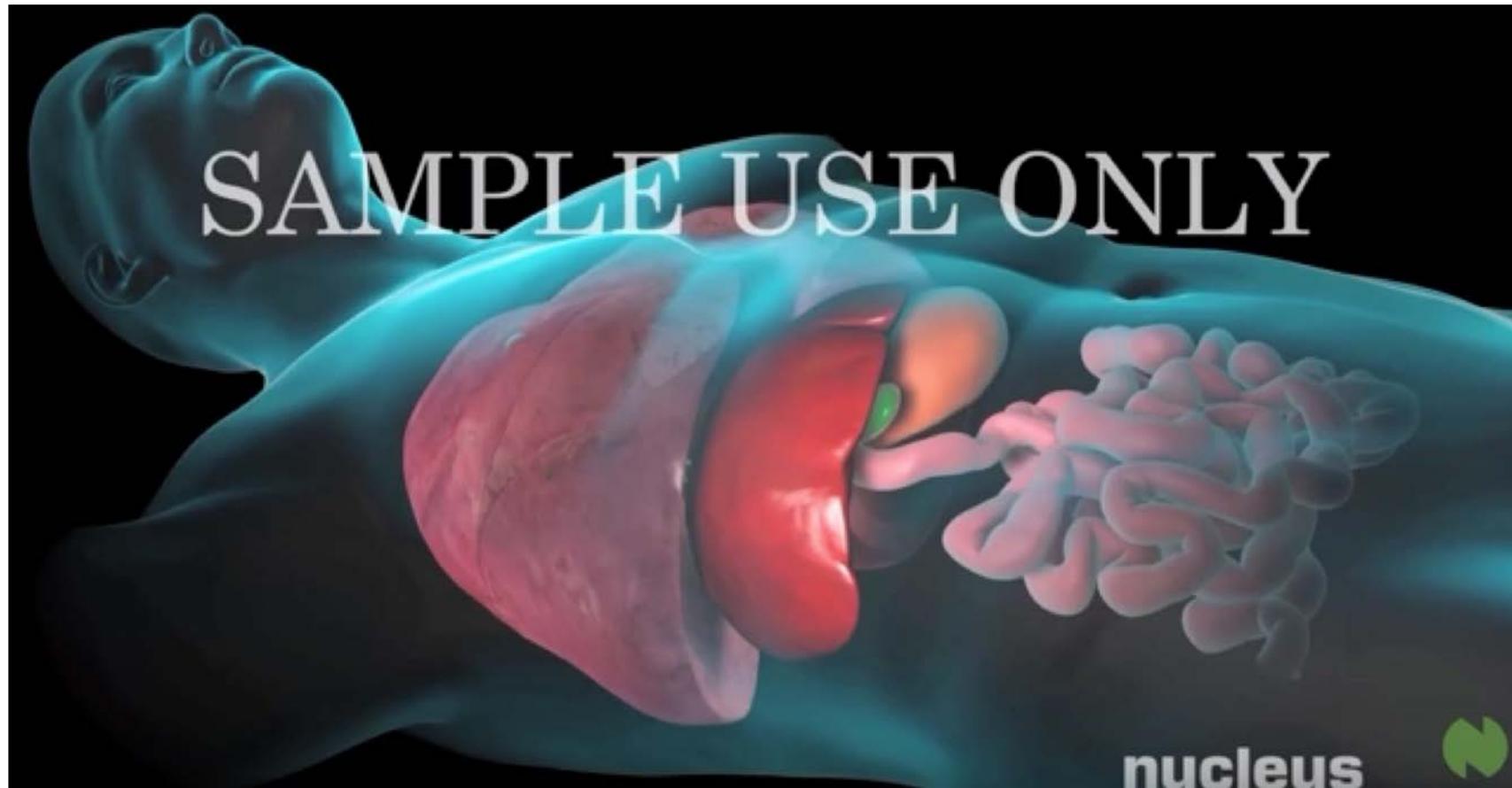
# What is Translational Bioinformatics

**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/>



# Tissue



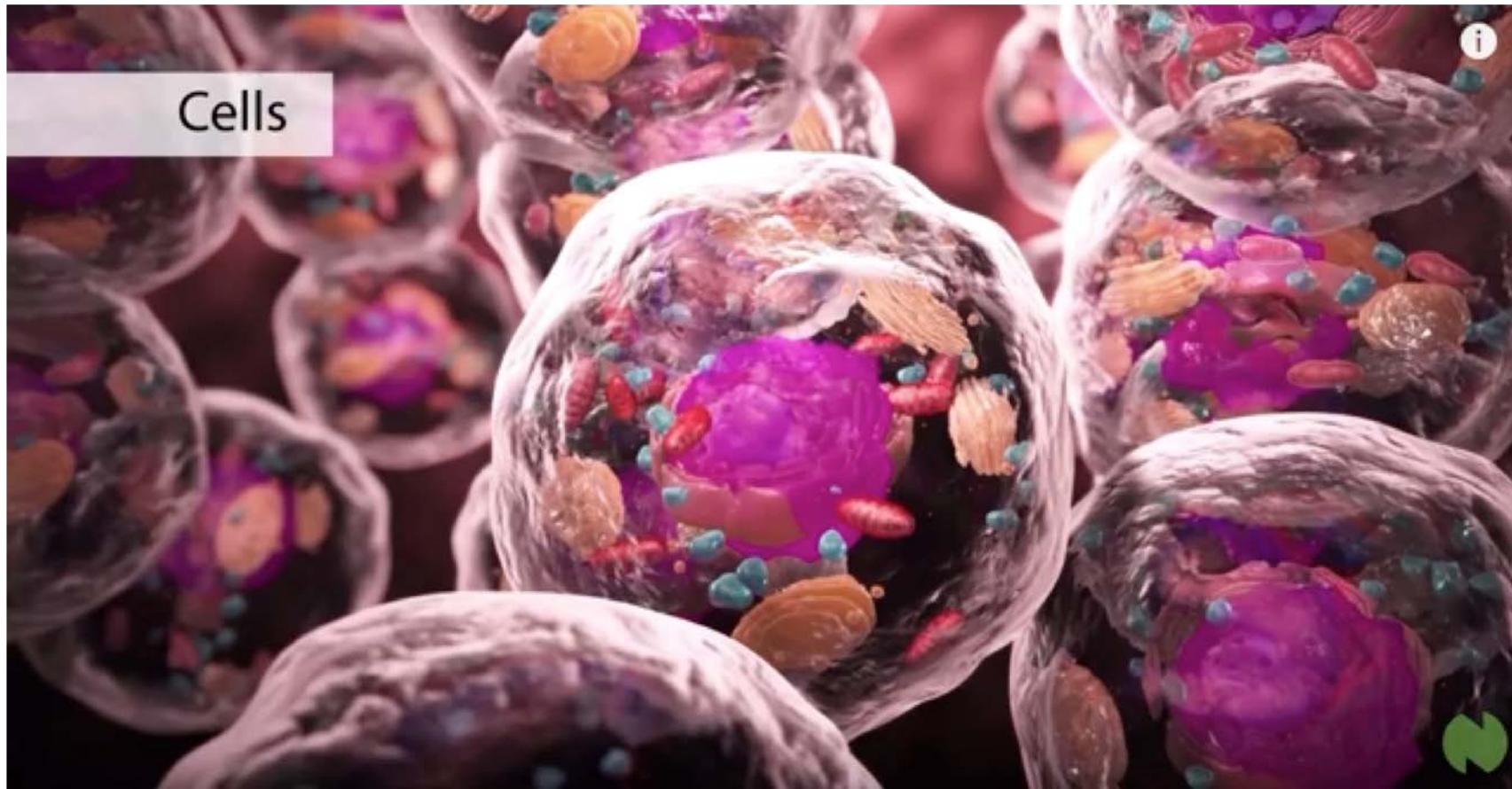
## Cell line



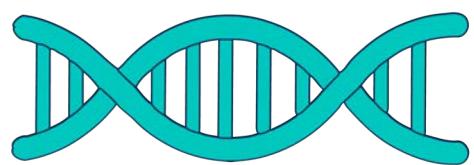
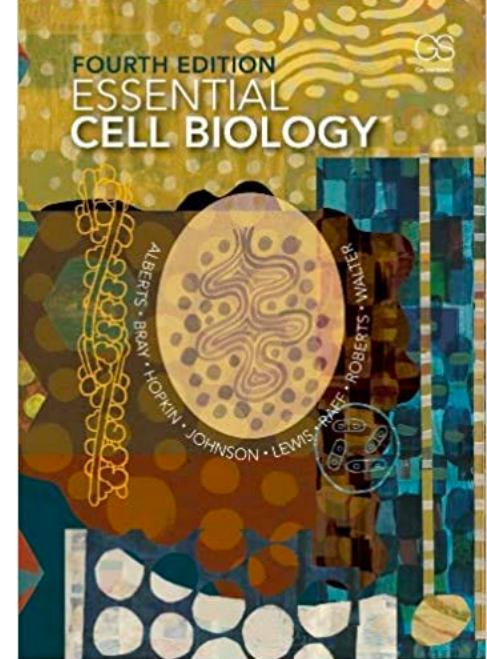
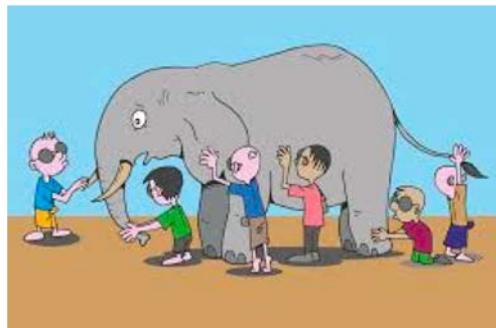
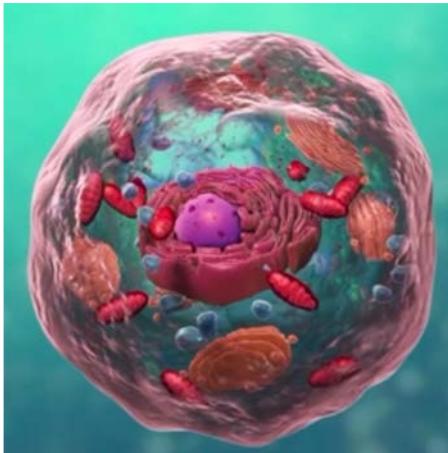
## Animal Model



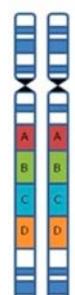
<https://www.youtube.com/watch?v=CK78IXTRH0s>



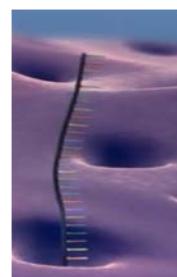
- Ignore spatial info
- Ignore dynamic
- Ignore cell-cell variation



DNA



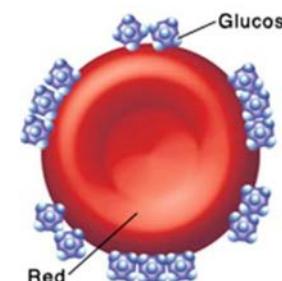
Chromosome  
(copy number)



mRNA (gene)

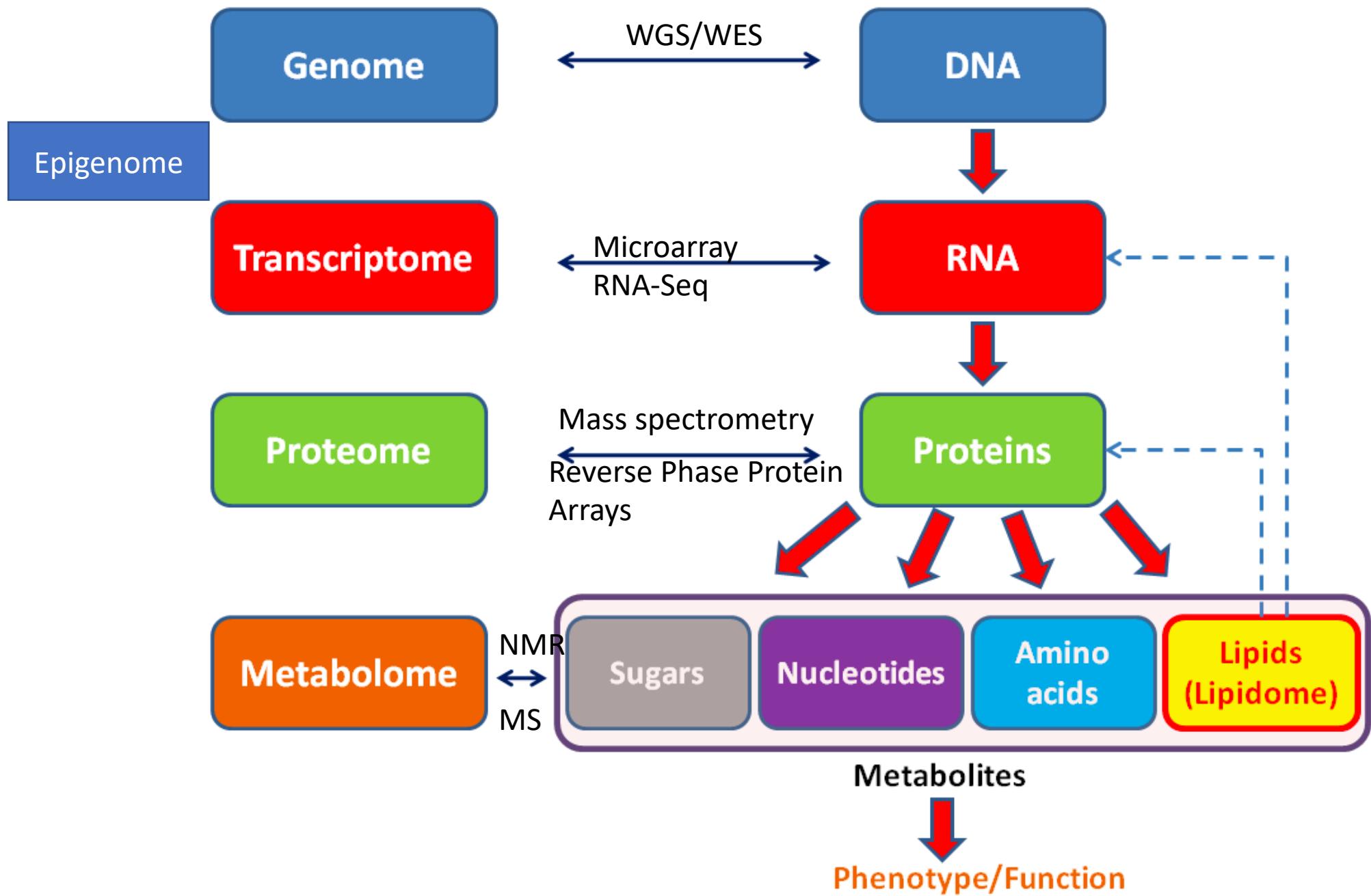


Protein

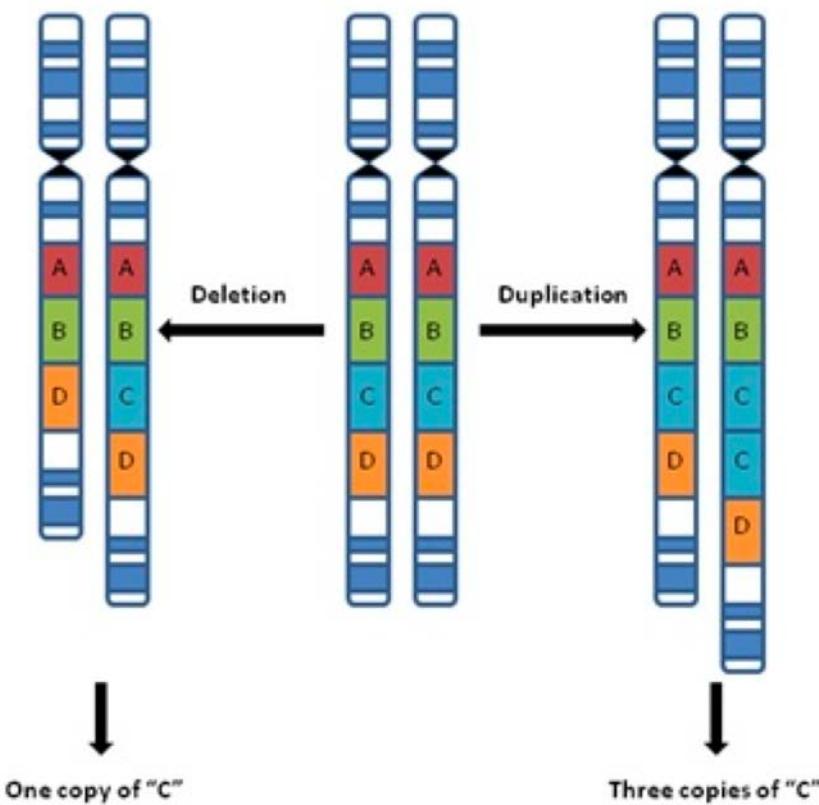


metabolite

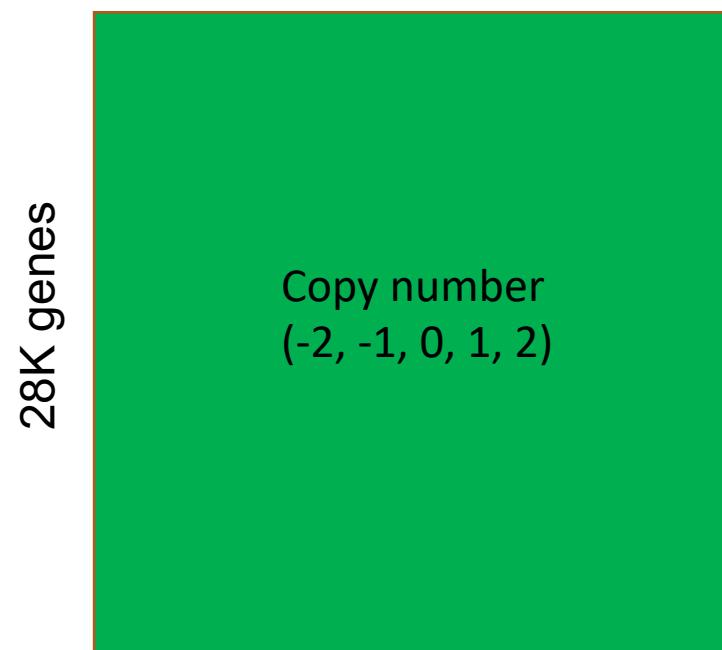
• • • •







1657 cell lines



[Modify Query](#)


## Liver Hepatocellular Carcinoma (TCGA, Firehose Legacy)

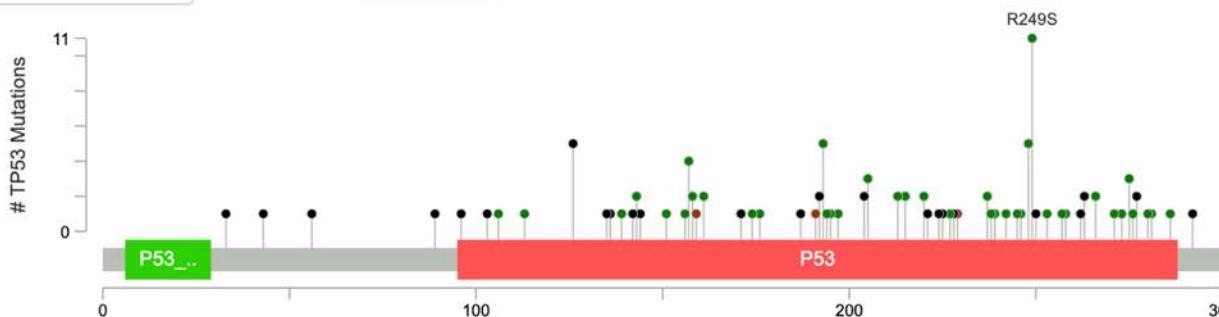
Samples with mutation and CNA data (366 patients/samples) - TP53

Queried gene is

[Oncoprint](#)
[Cancer Types Summary](#)
[Plots](#)
[Mutations](#)
[Co-expression](#)
[Comparison/Survival](#)
[CN Segments](#)
[Pathways](#)
[Download](#)

TP53

[Add annotation tracks](#)

Y-Axis Max: 

[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

## ClinVar

Genomic variation as it relates to human health

 [Search ClinVar](#)
[Advanced search](#)
[About](#)
[Access](#)
[Submit](#)
[Stats](#)
[FTP](#)
[Help](#)

Was this helpful?

[Follow](#)

[Print](#)
[Download](#)
**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



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: EGAS000000000001, Cancer

Search

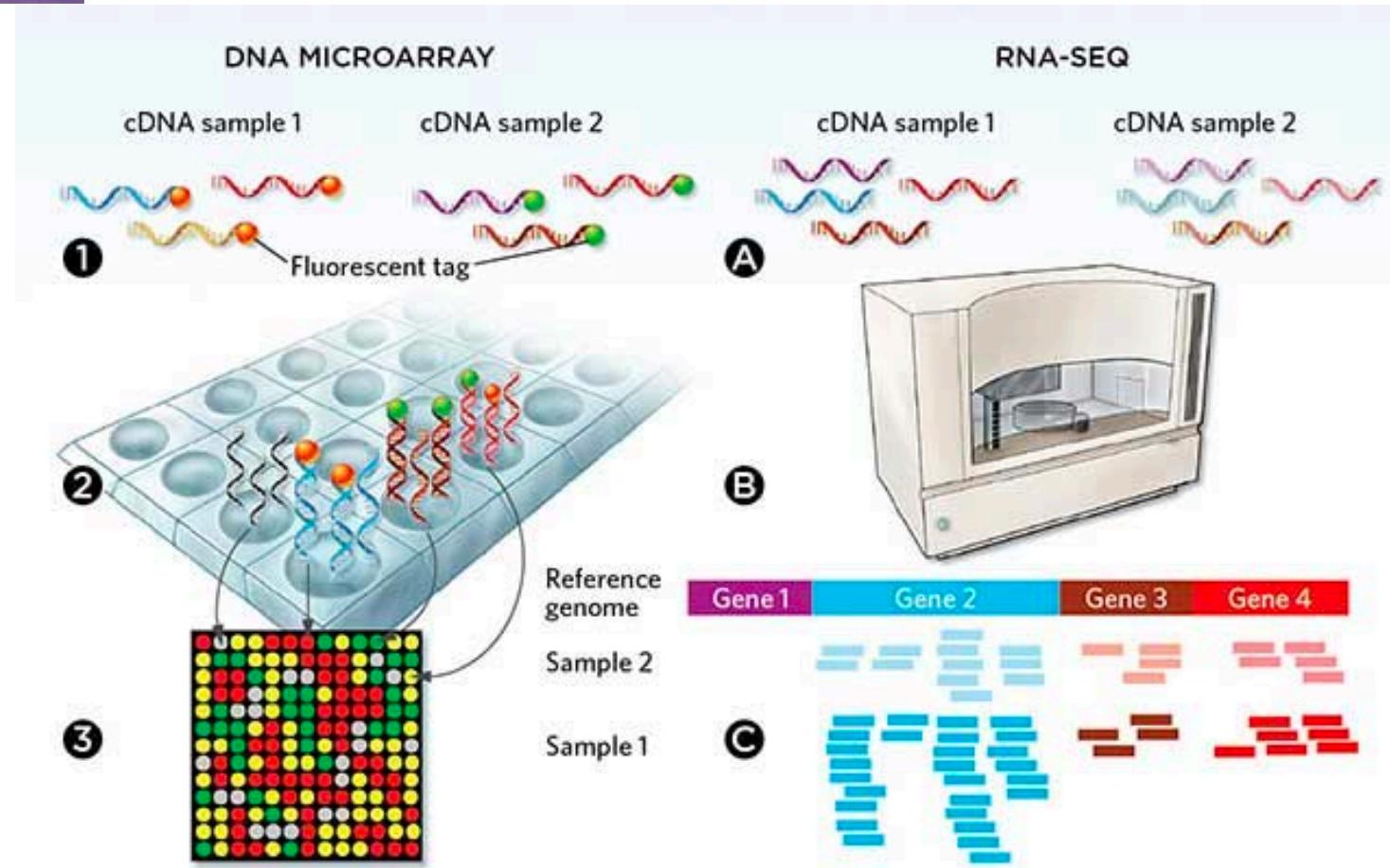
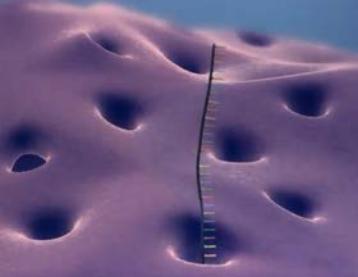
EGA home About Studies Datasets Data access committees Data providers Submit to EGA Contact Us Login

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](mailto:ega-helpdesk@ega-archive.org)

Help

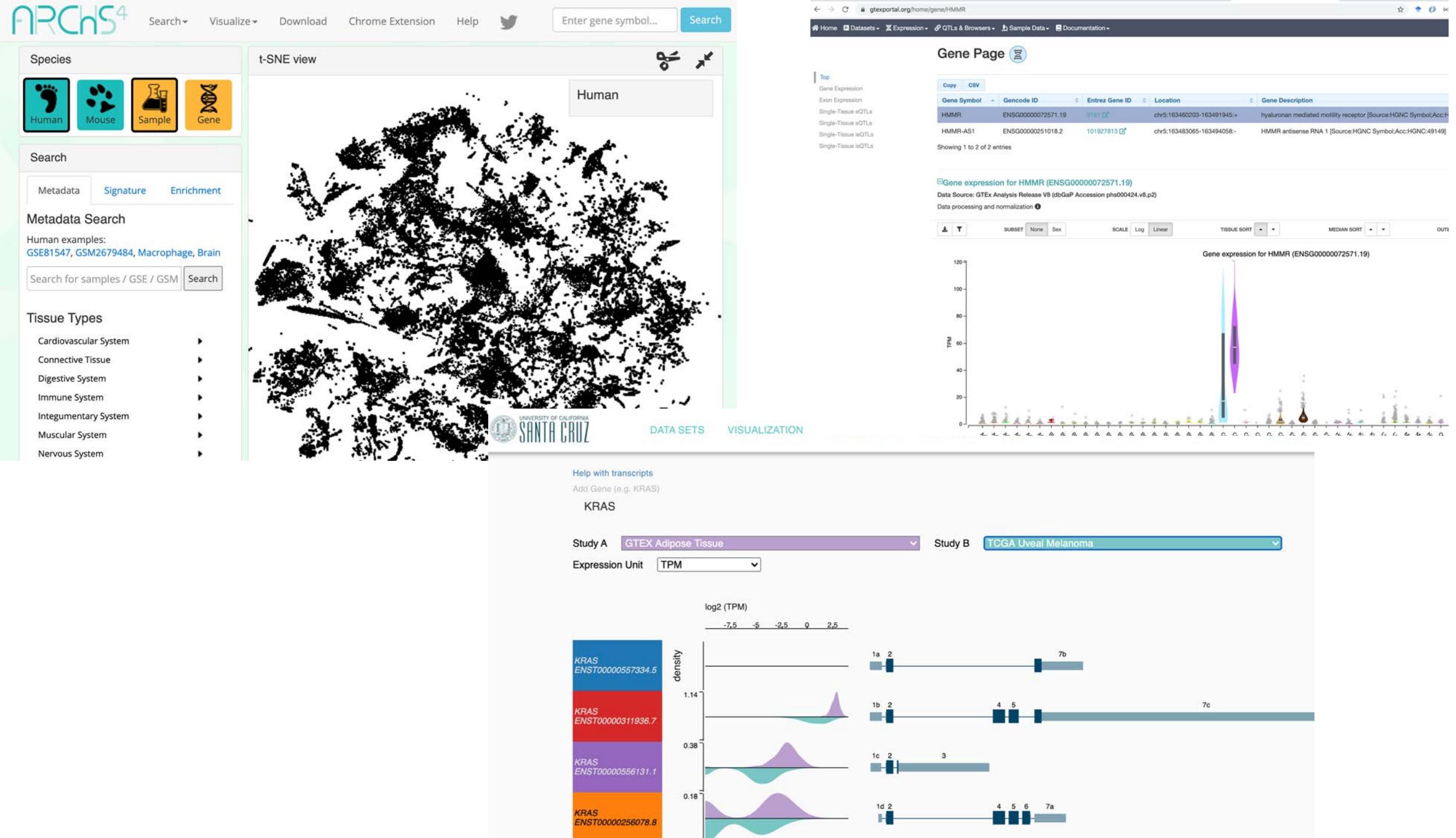
- [Users FAQ](#)
- [Submitters FAQ](#)
- [Using your EGA account](#)
- [Contact Us](#)
- [EGA mailing list](#)



1210 cell lines

60K genes

Gene Expression  
(numeric)





Keyword or GEO Accession

Search

## Browse Content

### Repository Browser

DataSets: 4348

Series: 134024

Platforms: 21247

Samples: 3740858

# ArrayExpress – function

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



Browse ArrayExpress

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

### Help

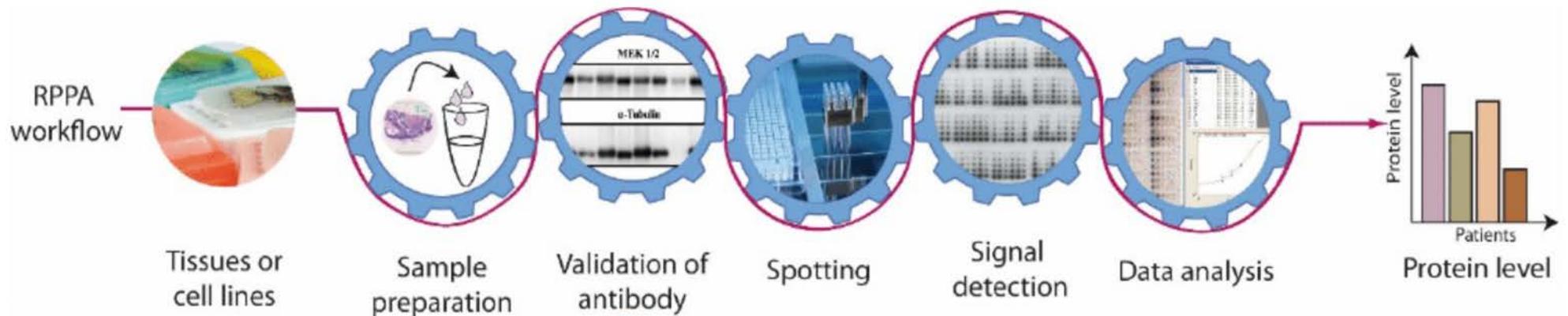
- [Users FAQ](#)
- [Submitters FAQ](#)
- [Using your EGA account](#)
- [Contact Us](#)
- [EGA mailing list](#)

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

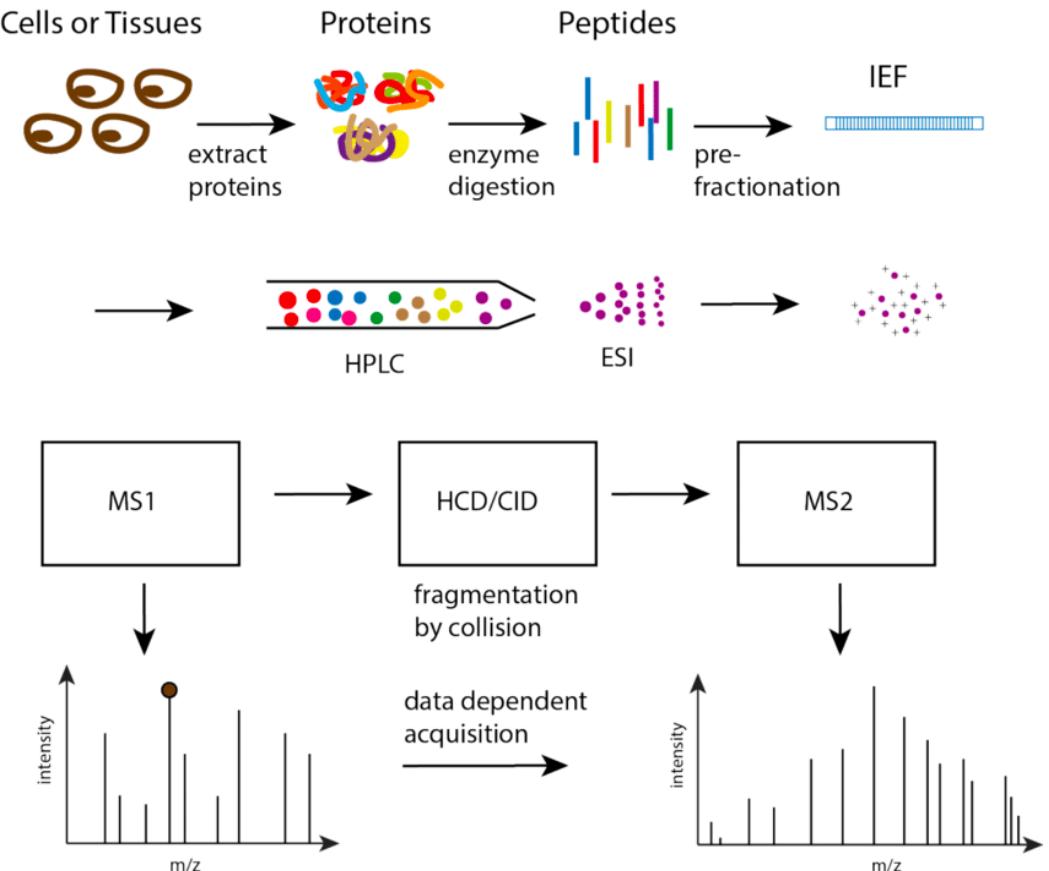


Proteome  
Phosphoproteome

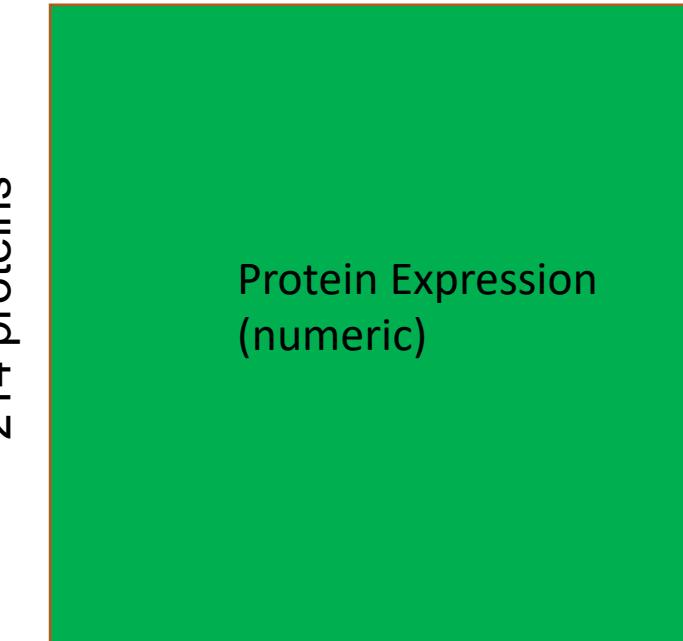
RPPA



MS



899 cell lines



# THE HUMAN PROTEIN ATLAS

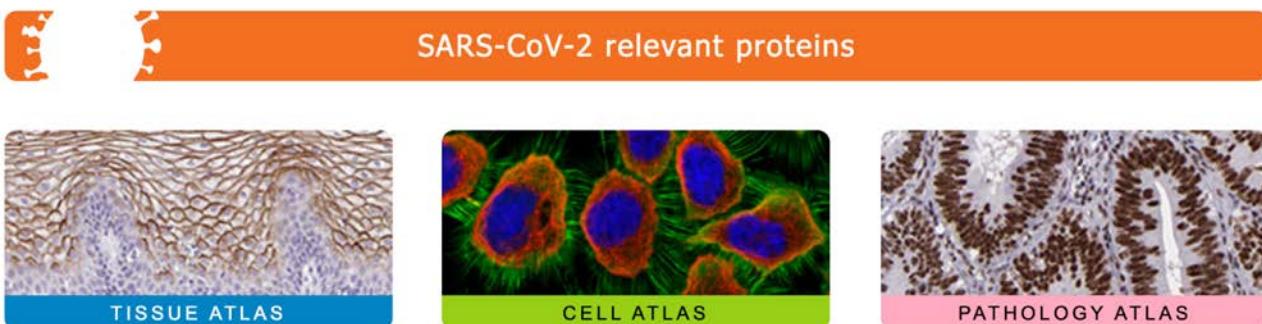


≡ MENU HELP NEWS

SEARCH<sup>i</sup>

e.g. RBM3, insulin, CD36

Search Fields »



## PRIDE Archive

PRoteomics IDEntifications Database

Home Resources Tools Docs About



NATIONAL CANCER INSTITUTE  
Office of Cancer Clinical  
Proteomics Research

Center for Strategic Scientific Initiatives

DATA PORTAL HOME



Data Portal

### Latest Data Release and Publications:

July 2020

Proteogenomic Characterization Reveals Therap

Michael A. Gillette, Shankha Satpathy, Song Cao,

CPTAC 3

(2016-present)

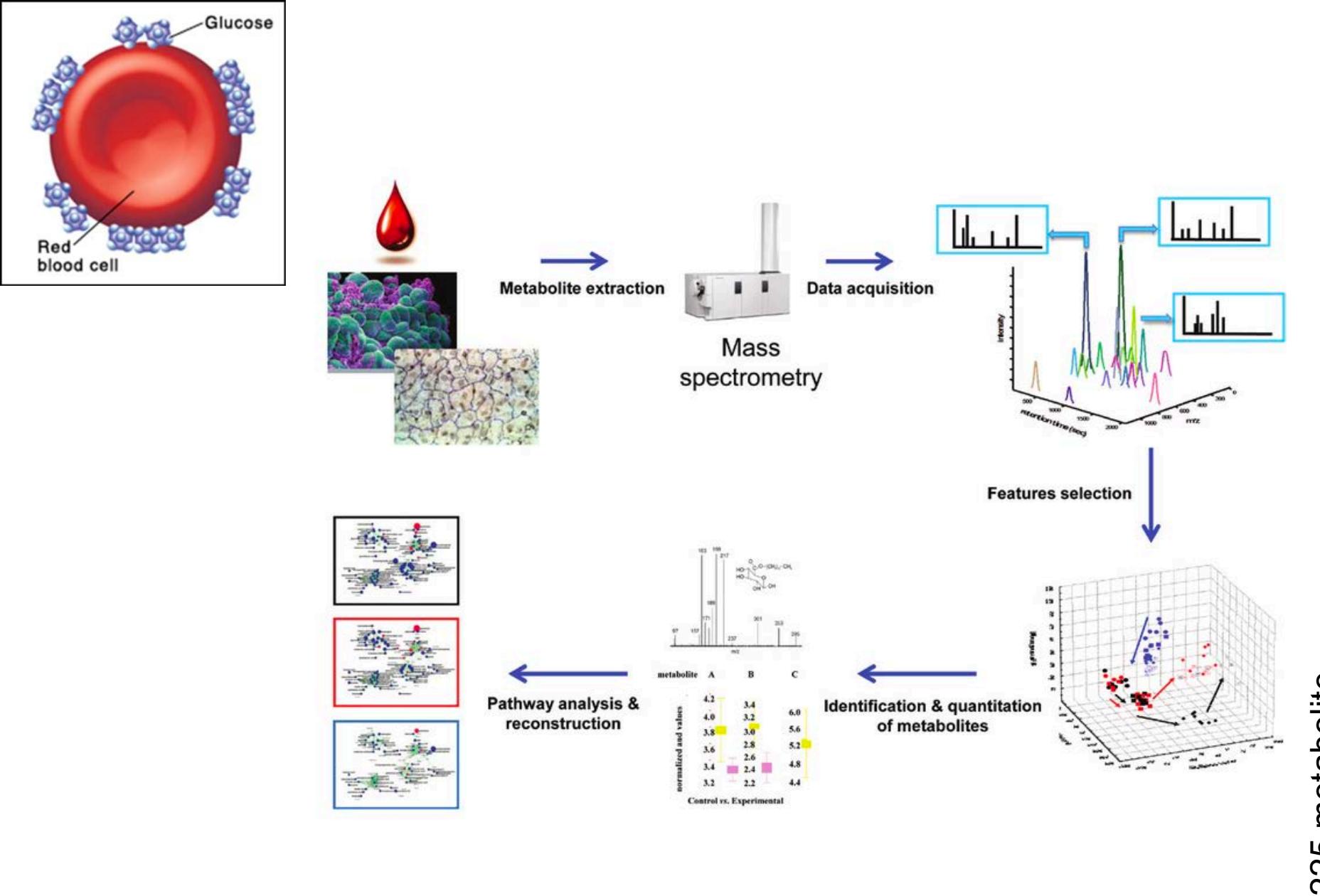
CPTAC 2 (2011-2016)

CPTC (2006-2011)

May 2020

CPTAC Head and Neck Squamous Cell Carcinom

Search



928 cell lines

Metabolite abundance (numeric)

HMDB ID ↗

CAS Number

Name ↗

Structure

Formula

Average Mass ↗

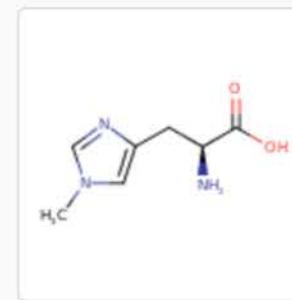
Monoisotopic Mass ↗

Biospecimen

HMDB0000001

332-80-9

1-Methylhistidine

C<sub>7</sub>H<sub>11</sub>N<sub>3</sub>O<sub>2</sub>

169.1811

169.085126611

Blood  
Cerebrospinal fluid  
Feces  
Saliva  
Urine

HMDB0000002

109-76-2

1,3-Diaminopropane

MetaboLights

Examples: Alanine, Glucose

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

MetaboLights / Search

Filter your results

Type

- study
- compound

Technology

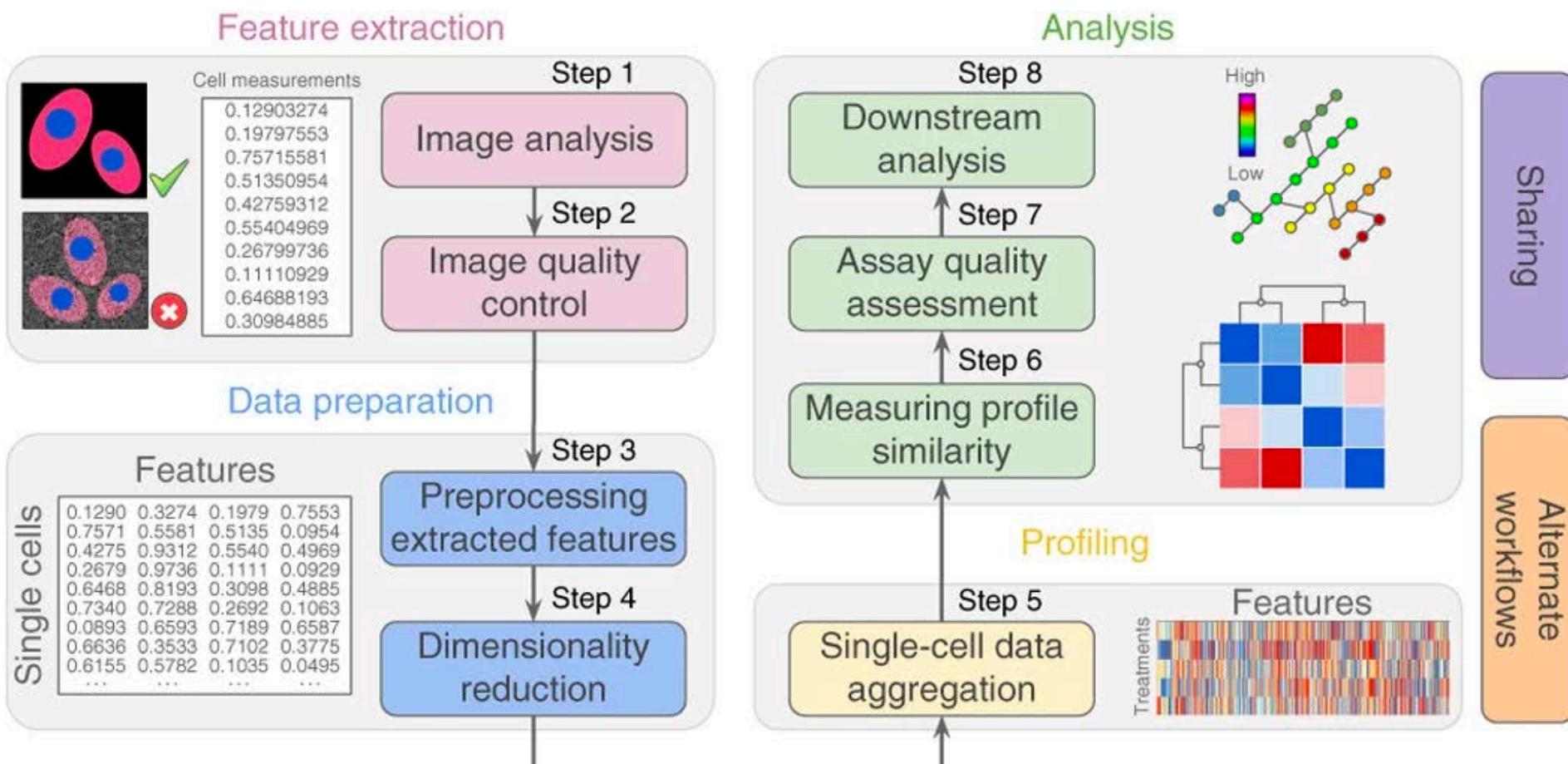
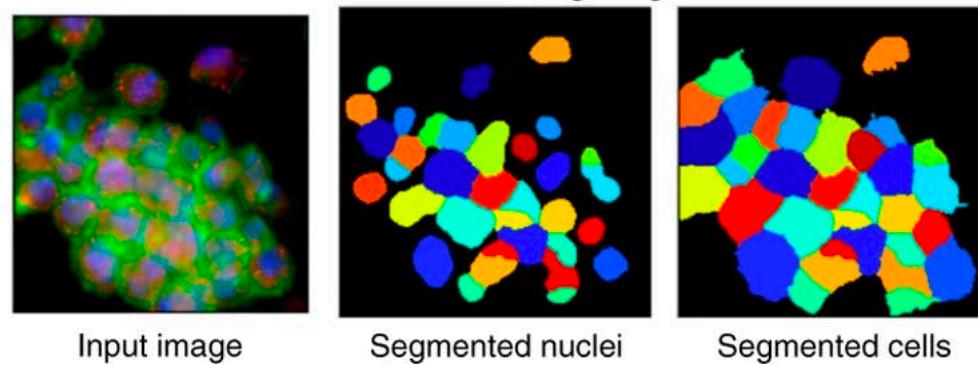
Organism

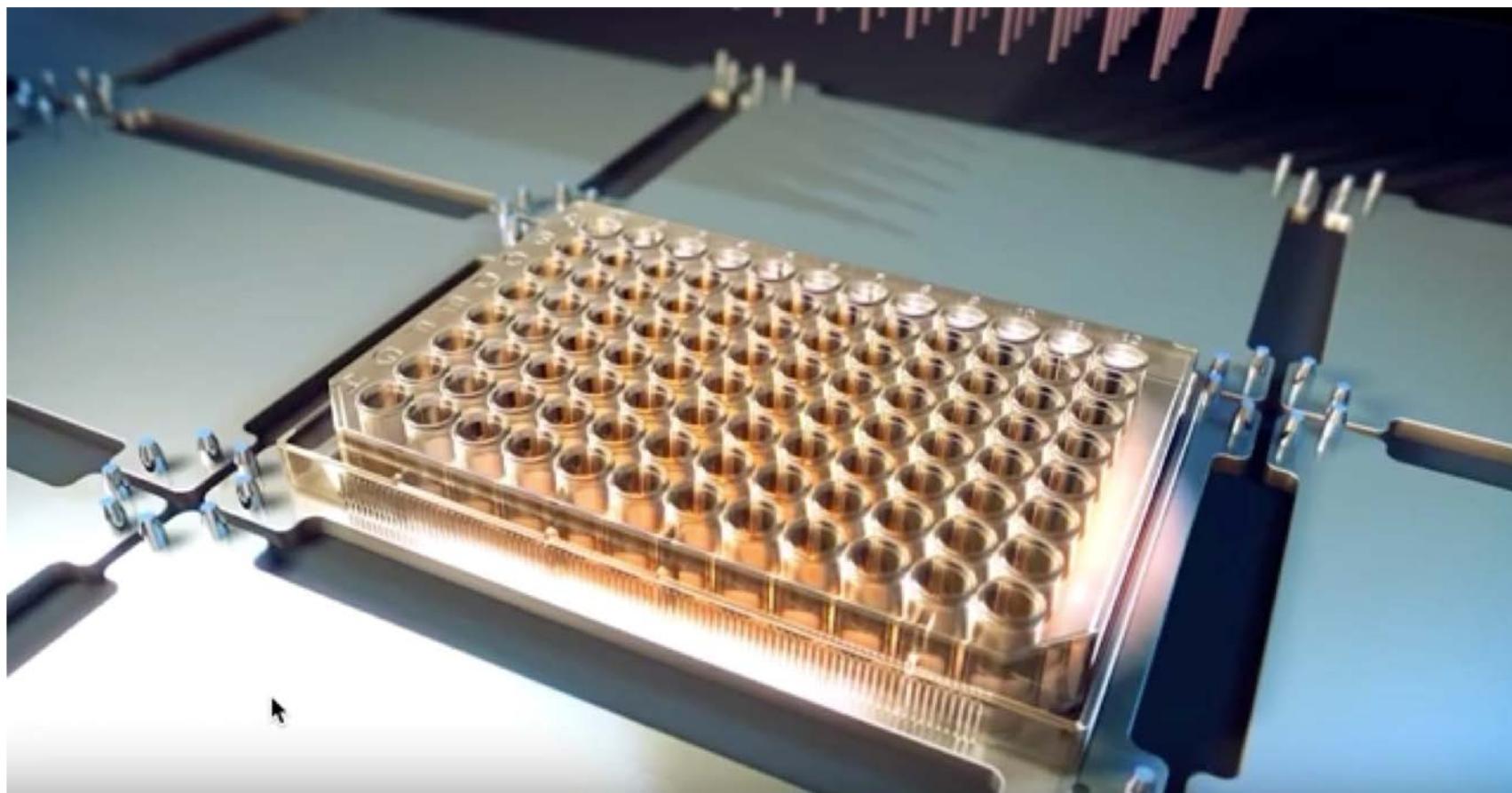
Organism Part

676 results , showing 1 to 10

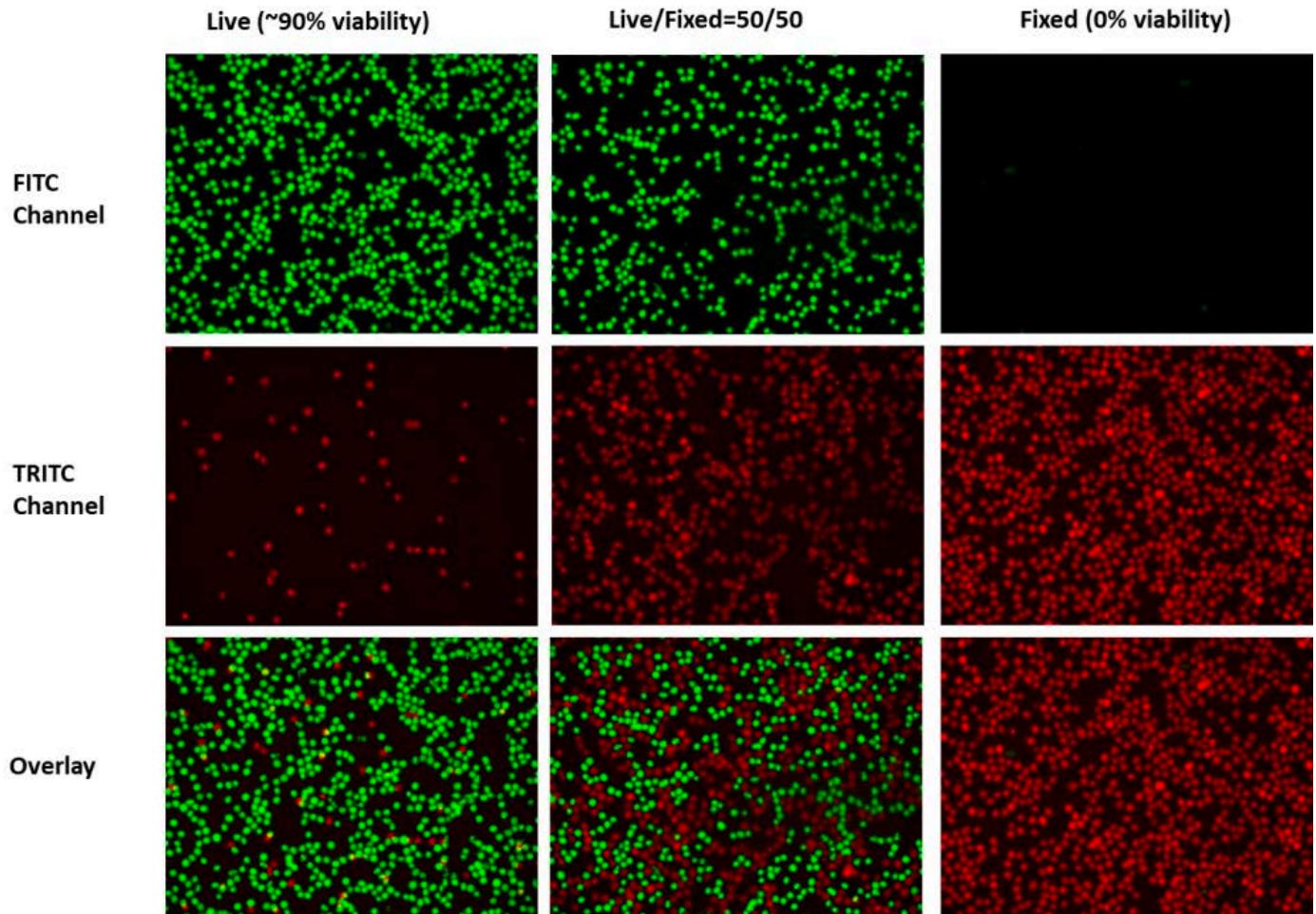
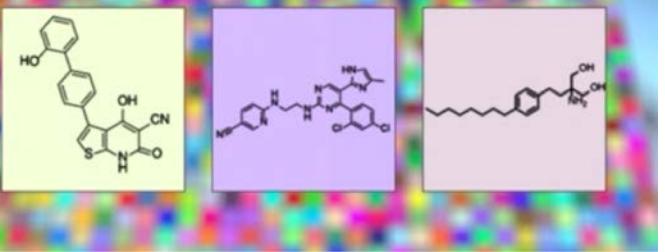
**Stable Isotope-Assisted Plant Metabolomics: Combining Tracer-Based Labeling for Enhanced Untargeted Protein Annotation**

|                  |                   |                |
|------------------|-------------------|----------------|
| Study Identifier | MTBLS1217         | Organism       |
| Study Size       | 939.11MB          | Study Location |
| Submitted by     | Christoph Bueschl | Email          |





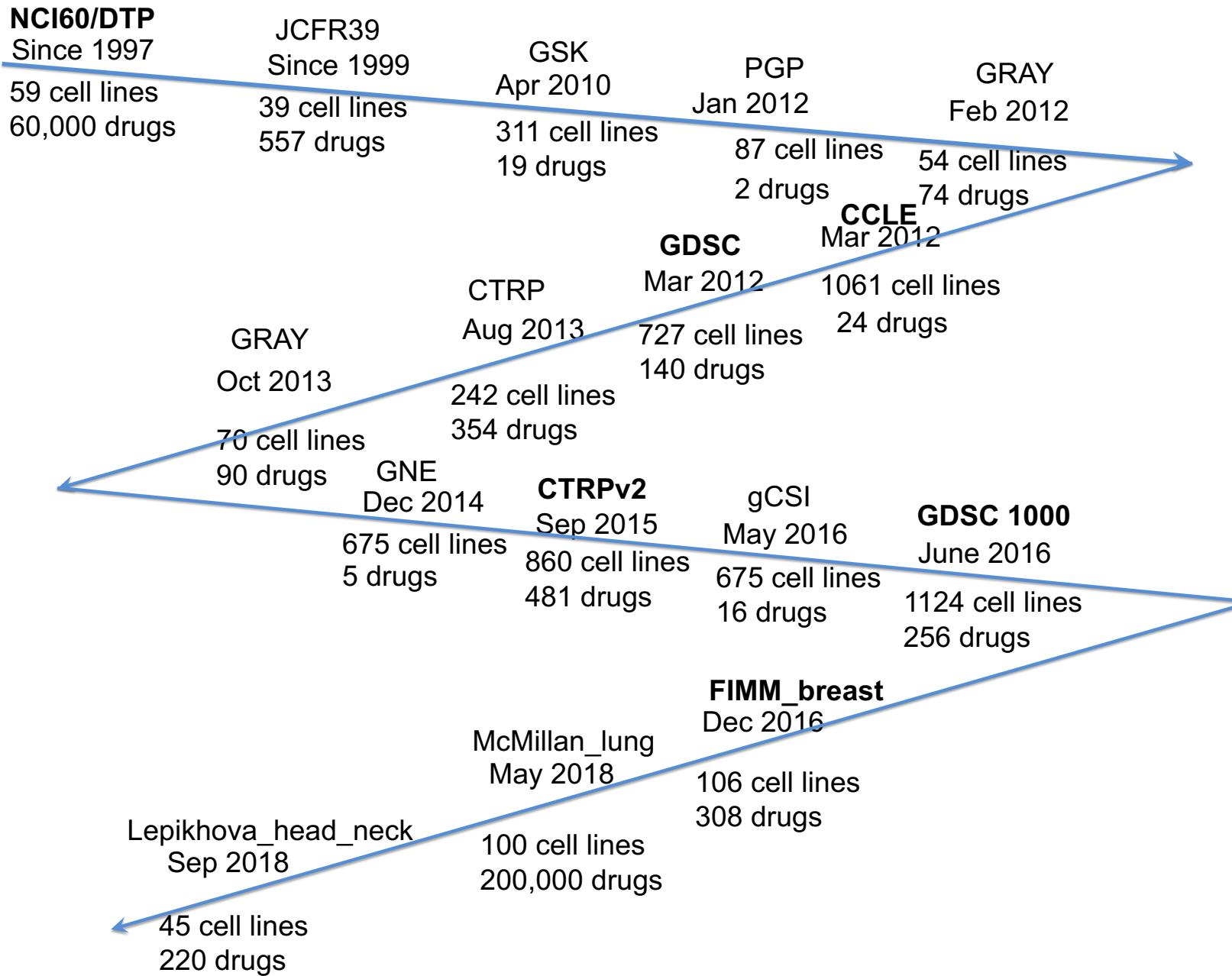
Cellular response/drug sensitivity



5K compounds

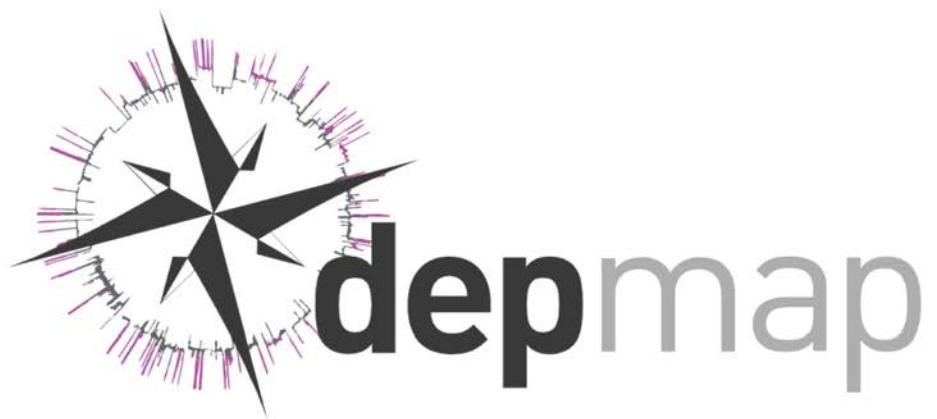
578 cell lines

Drug sensitivity score  
(numeric)





Cellular response/dependency



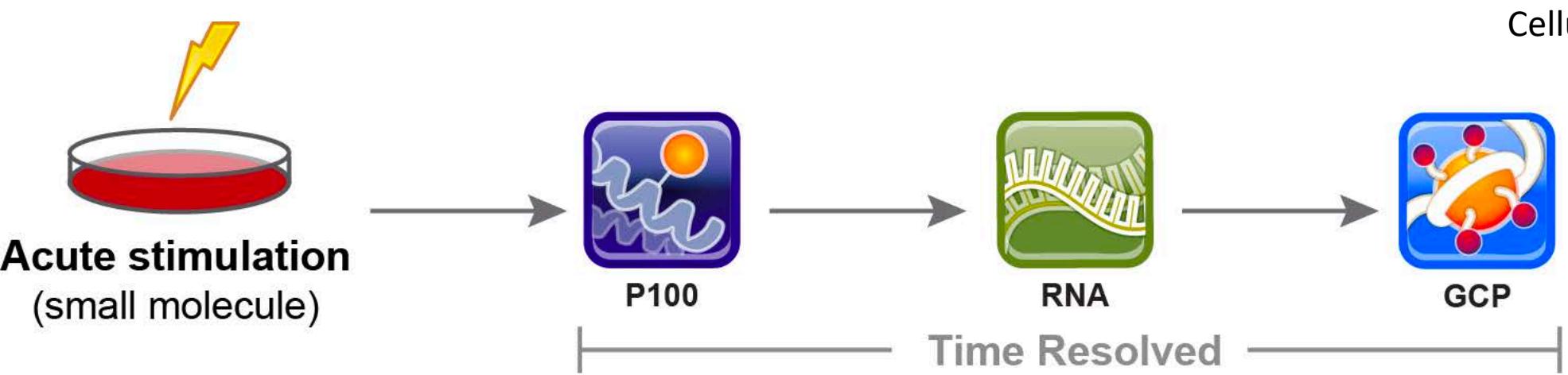
**A Cancer Dependency Map to systematically identify  
genetic and pharmacologic dependencies and the  
biomarkers that predict them.**

18K genes

625 cell lines

Gene essentiality score  
(numeric)

Cellular response/gene expression

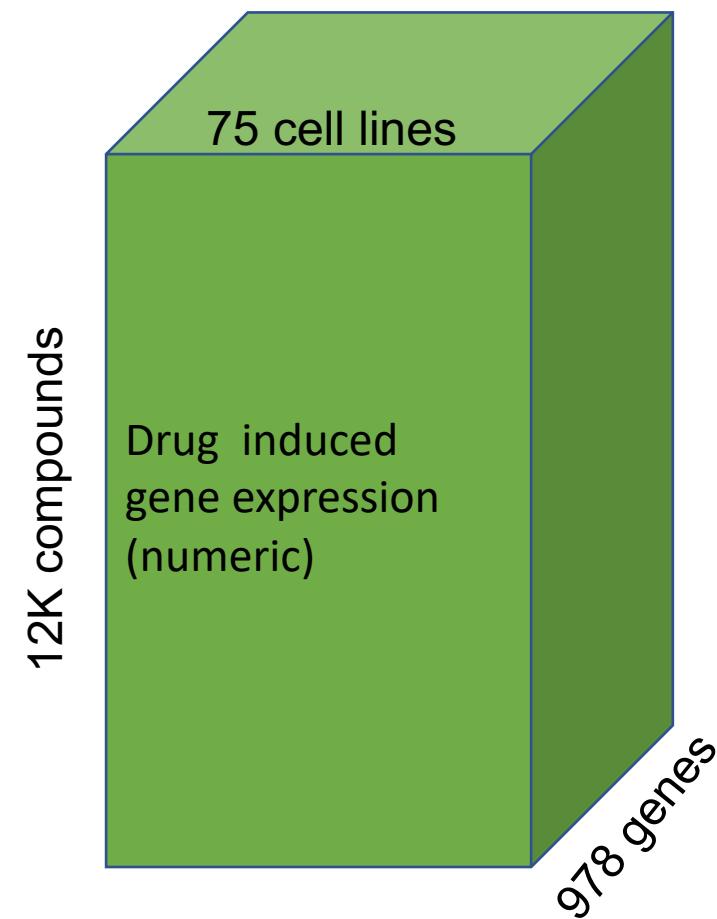


NIH LINCS  
PROGRAM

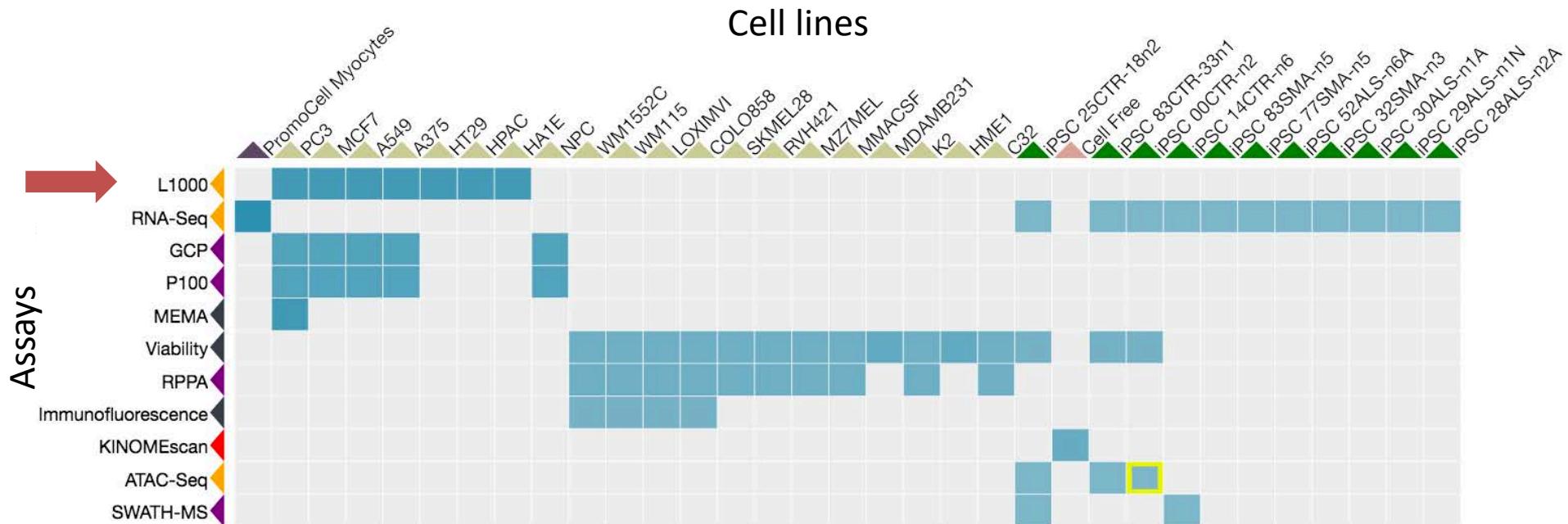
LI

Home About Centers Data Tools Community Publications News

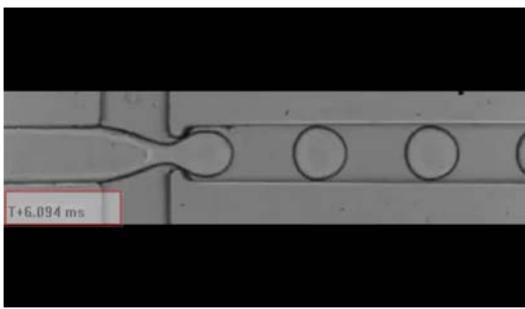
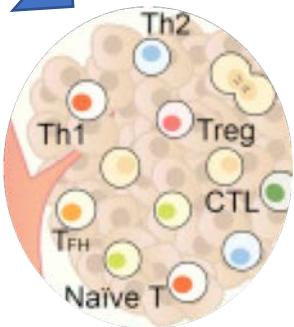
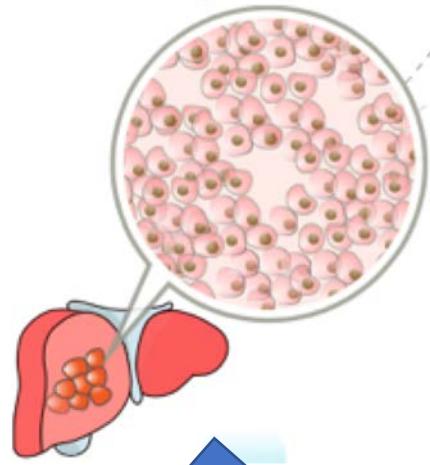
A screenshot of the NIH LINCS Program website. The header features the text 'NIH LINCS PROGRAM' with a 3D cube icon to the left. Below the header is a navigation bar with links: Home, About, Centers, Data, Tools, Community, Publications, and News. The 'About' link is highlighted with a blue background. At the bottom of the page is a decorative footer section featuring a circular graphic and a molecular diagram showing the relationship between MAP3K1 and MAPK14.



# 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



20K patient tissues

Single cells

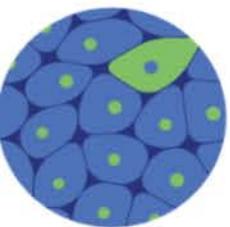
60K genes

Gene Expression  
(numeric)

Millions of single cells

60K genes

Gene Expression  
(numeric)

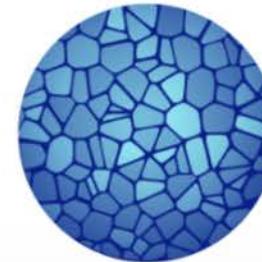


**Single Cell** BETA  
**PORTAL**

Reducing barriers and accelerating single-cell research

Featuring  
272 studies  
9,584,113 cells

# COVID-19 Cell Atlas



HUMAN  
CELL  
ATLAS



CHAN  
ZUCKERBERG  
INITIATIVE

### Single Cell data



Gene expression

Mutation

Copy number variation

Protein expression

Metabolite

Drug sensitivity

Gene essentiality

Drug-induced expression

### Cell line data



Gene expression

Mutation

Copy number variation

Protein expression

Metabolite

Drug sensitivity

Gene essentiality

Drug-induced expression

### Patient data



Gene expression

Mutation

Copy number variation

Protein expression

metabolite

Drug response

Gene essentiality

Drug-induced expression

complete

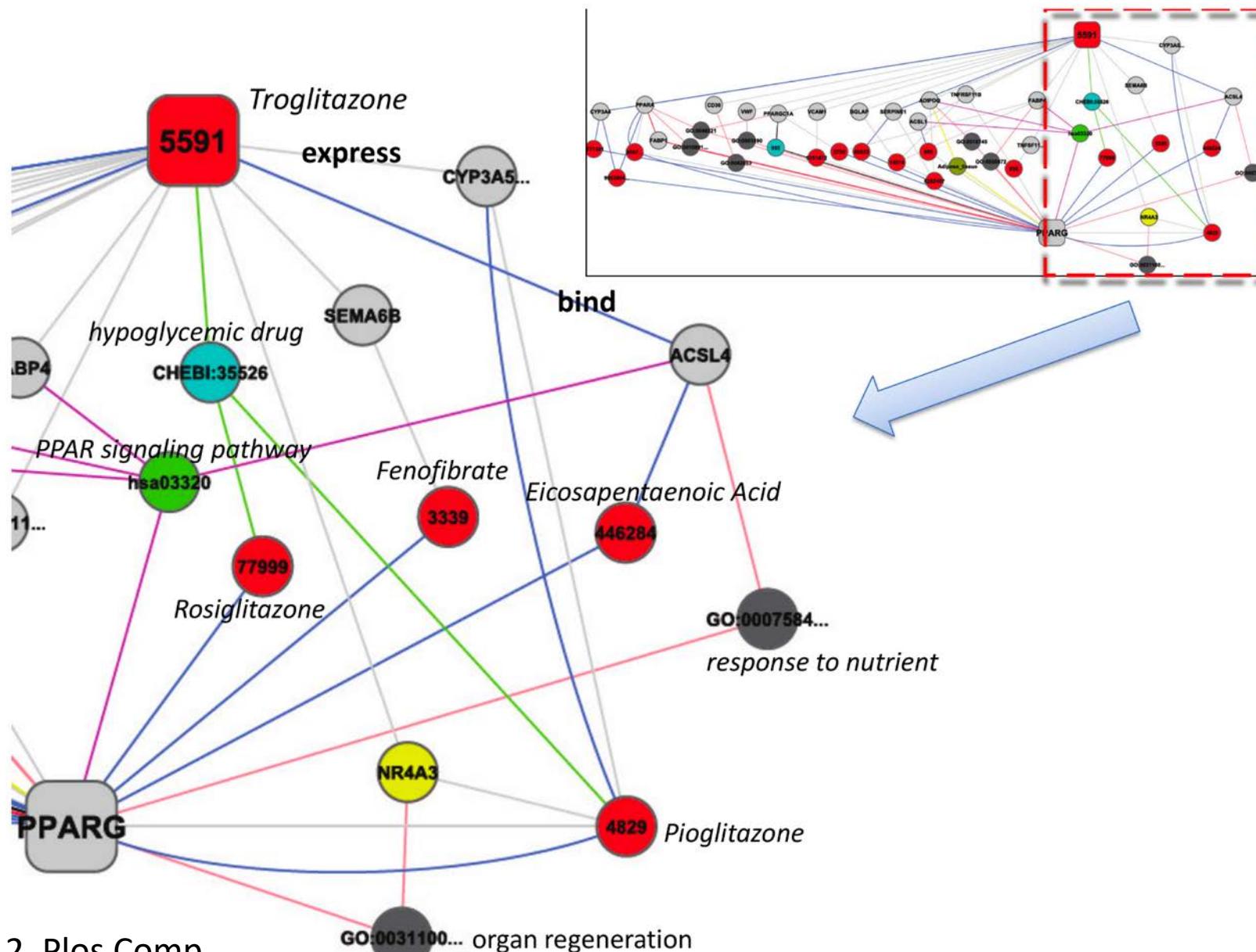
unknown

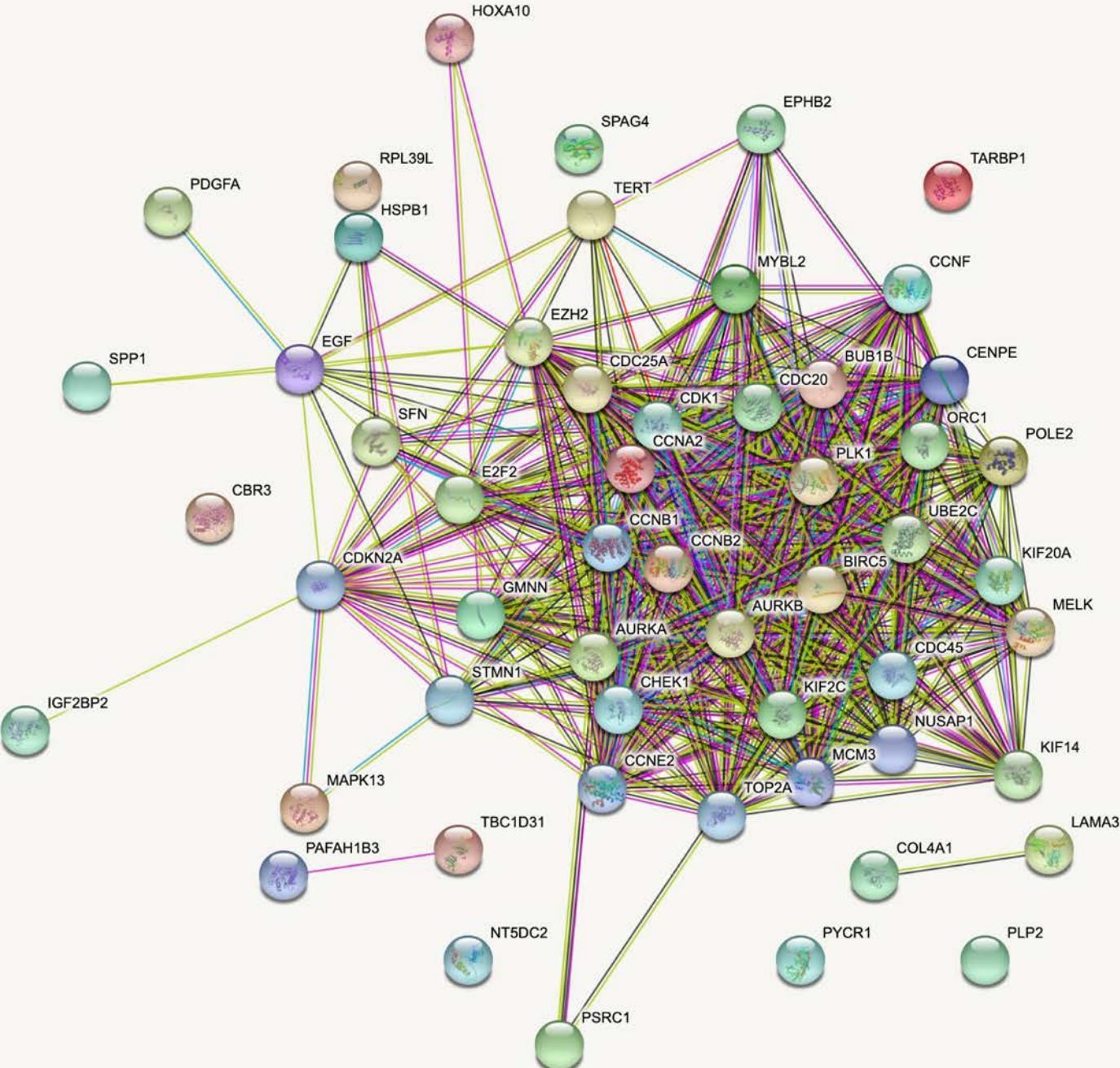
partial

# Data properties

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

# Network (knowledge graph)





# Unstructured data

## 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 genetically 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](http://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  $\beta$ -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.

### Criteria

#### Inclusion Criteria:

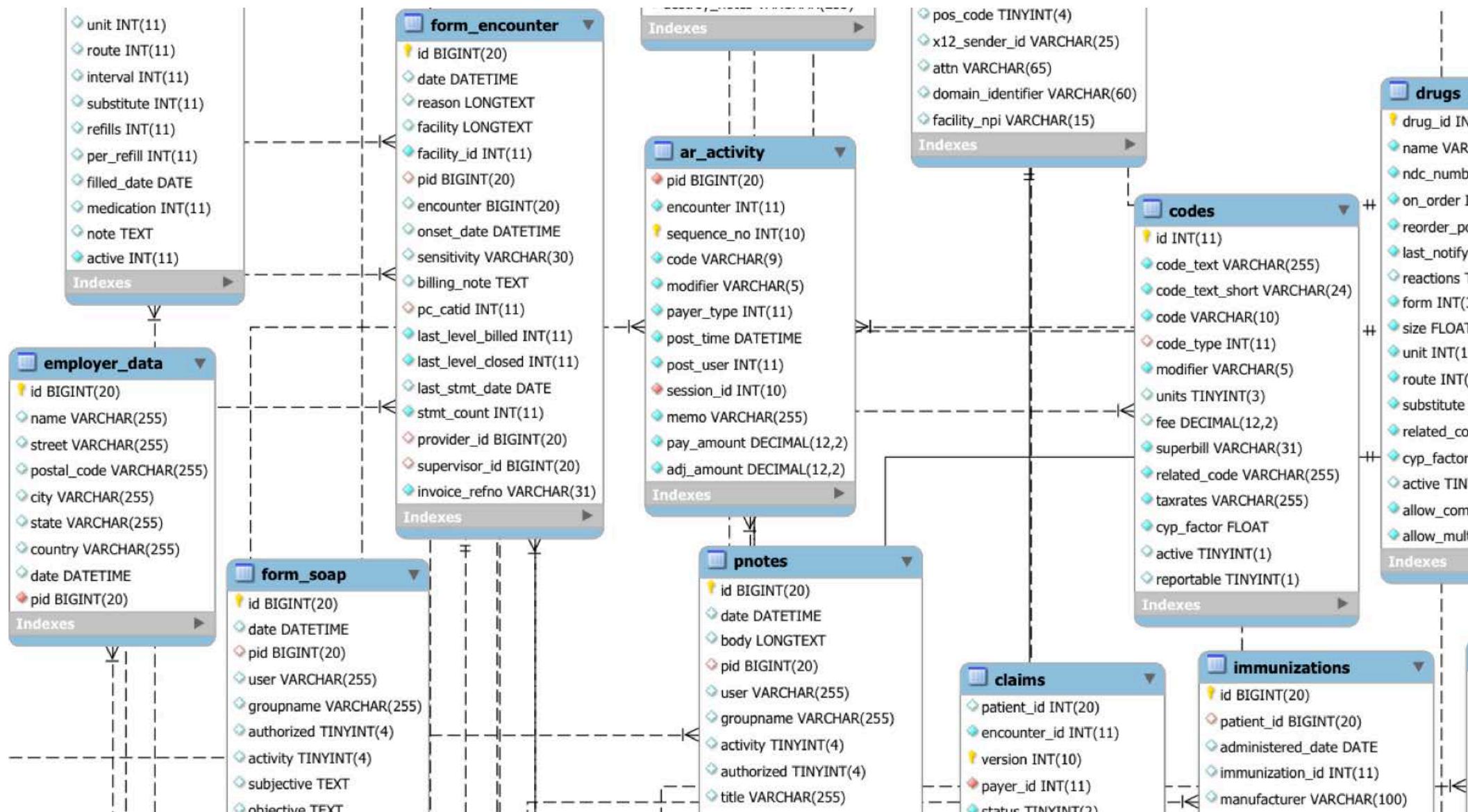
- Female or male patients  $\geq 18$  years old with WHO performance status  $\leq 1$
- HER-2 over-expressing metastatic breast cancer cells confirmed by histologic analysis
- Progressive disease on prior trastuzumab alone/or in combination with other agents
- Patient resistance to trastuzumab and taxanes (Phase II)
- Measurable disease according to RECIST (Phase II)
- Patients neurologically stable with adequate bone marrow, liver and renal function

#### Exclusion Criteria:

- Patients receiving endocrine therapy for breast cancer  $\leq 2$  weeks prior to study treatment start
- Patients currently receiving chemotherapy, immunotherapy or radiotherapy prior to study treatment start
- Patients who have previously received mTOR inhibitors

Other protocol-defined inclusion/exclusion criteria may apply

# Electronic Medical Records

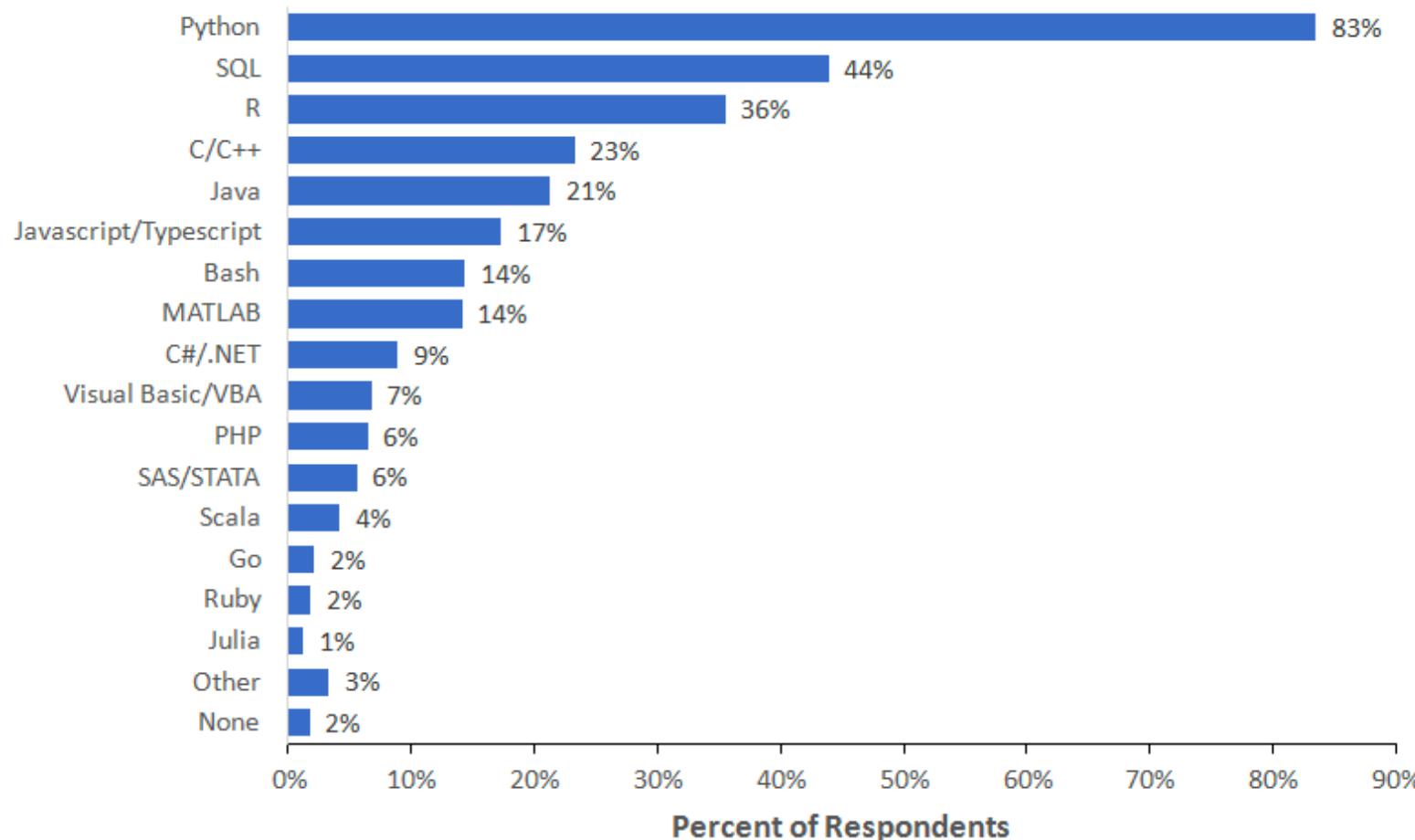


# Summary of big data

- Unstructured
  - Free text (literature, clinical notes)
- Structured
  - Network
    - Protein-protein interaction
    - Knowledge-graph (e.g., drug-target interaction, drug-disease relation, target-pathway)
  - Matrix
    - Image
    - 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.

# R resources

- Tutorial
  - [http://manuals.bioinformatics.ucr.edu/home/R\\_BioCondManual/](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/>

# R

- Install R
- Install Rstudio

RStudio

Project: (None)

File Edit View Insert Plots Packages Help

signatures.R core\_functions.R GSE17400.R workflow.R main.R workflow.R merge.R

Source on Save Run Source

```
26 #download clincal data and mrna meta data from TCGA
27 #####need to go to TCGA website to download all the required data;
28
29 #compute random tumors and cell line correlations. Take a few minutes.
30 cutoff = 9.34E-06 # comment out this line if you want to use the random samples to correct the p value
31 #source("../code/tumor_cell_line/correct_by_expo.R")
32
33 #compute tumor vs cell line correlations
34 source("../code/tumor_cell_line/compute_tumor_cell_line_cor_update.R")
35
36 #analyze correlations and select cell lines
37 source("../code/tumor_cell_line/select_cell_lines_update.R")
38
39 #compute differentially expressed genes between tumors and non-tumors
40 source("../code/tumor_cell_line/compute_disease_signatures.R")
41
42 #compute differentially expressed genes between tumors and cell lines
43 #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 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

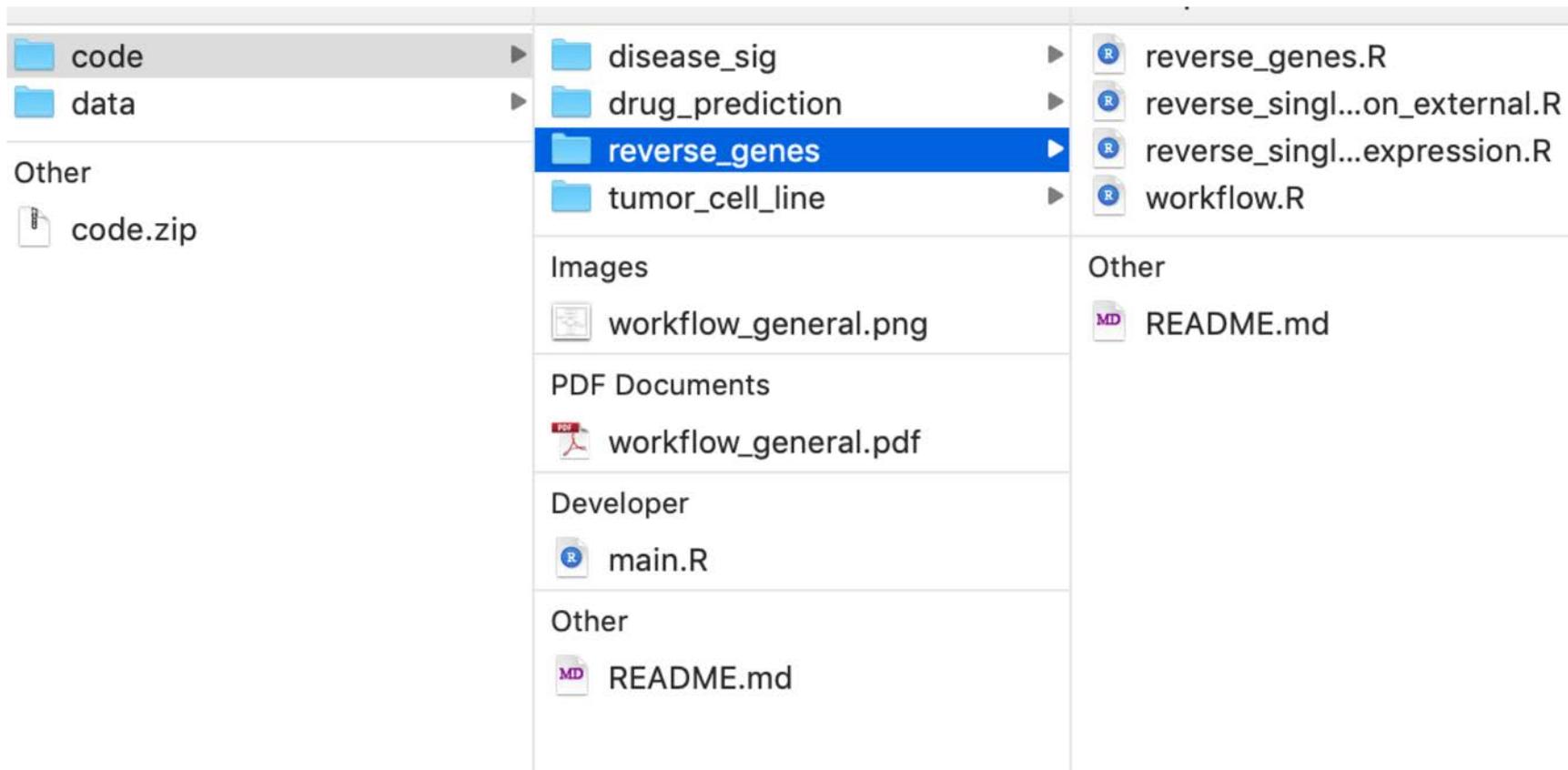
Generic function for plotting of R objects. For more details about the graphical parameter arguments, see [par](#).

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



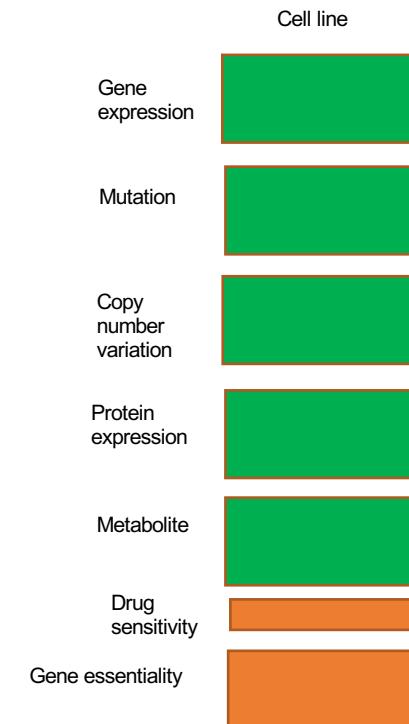
[https://github.com/Bin-Chen-Lab/HCC\\_NEN](https://github.com/Bin-Chen-Lab/HCC_NEN)

# OCTAD\_cell\_line.RData

## octad\_cell\_line\_features

|                | <b>id</b>        | <b>name</b> | <b>type</b> |
|----------------|------------------|-------------|-------------|
| <b>VPS13D</b>  | mutation_VPS13D  | VPS13D      | mutation    |
| <b>AADACL4</b> | mutation_AADACL4 | AADACL4     | mutation    |
| <b>TMEM57</b>  | mutation_TMEM57  | TMEM57      | mutation    |
| <b>ZSCAN20</b> | mutation_ZSCAN20 | ZSCAN20     | mutation    |
| <b>POU3F1</b>  | mutation_POU3F1  | POU3F1      | mutation    |
| <b>VAV3</b>    | 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

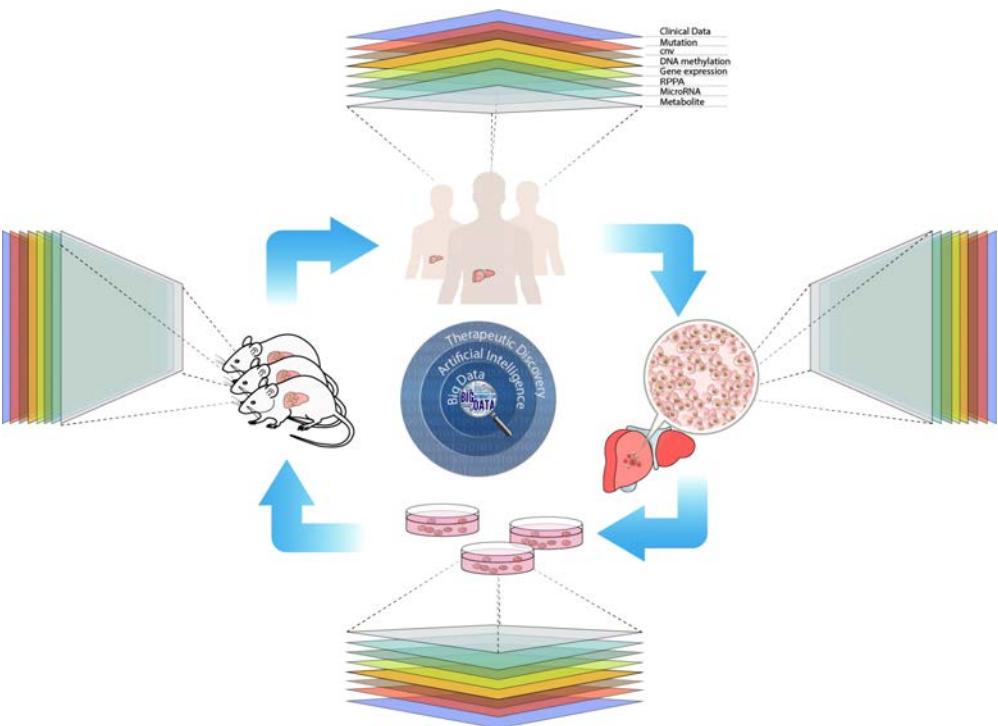
- Data.table
- H5

# 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
- Correlation analysis
- Clustering
- Predictive models