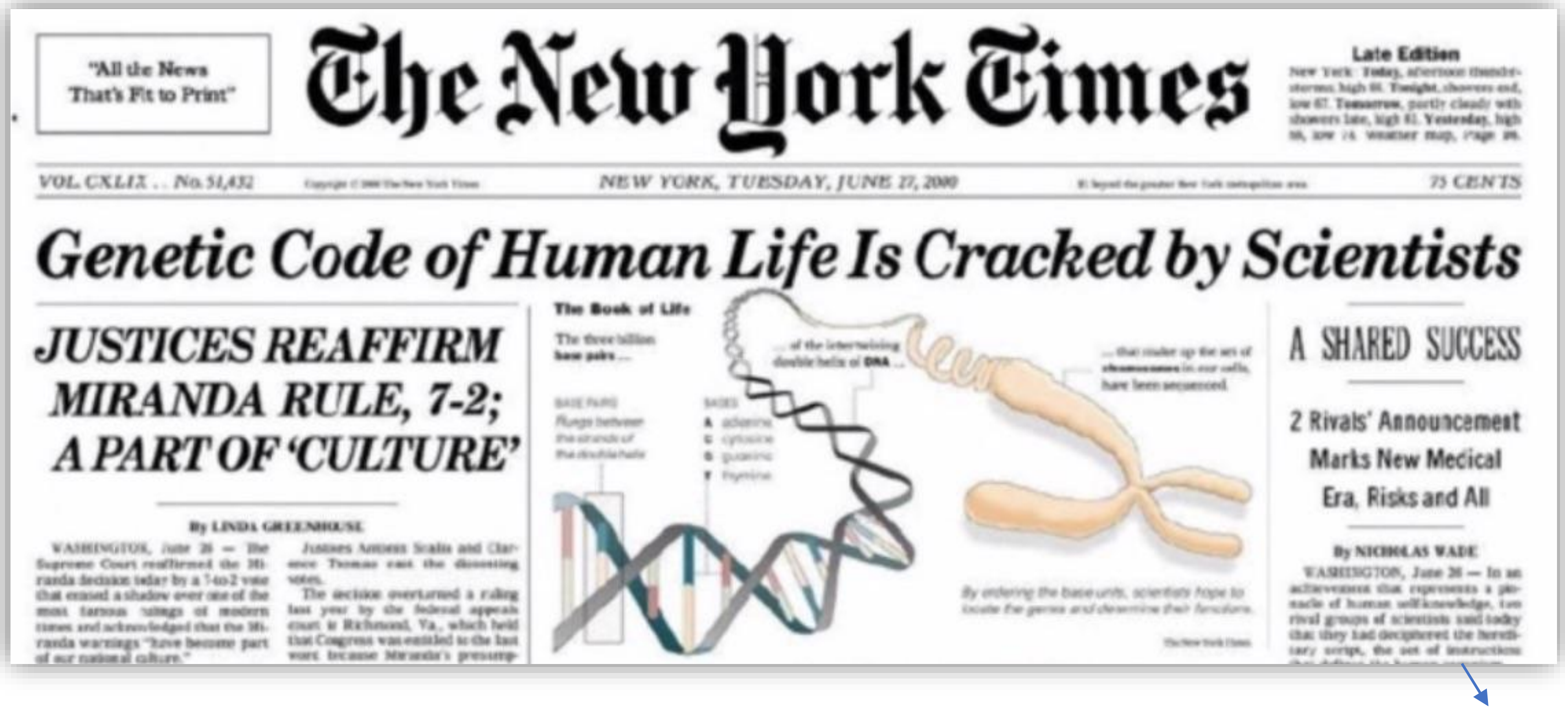
A framework for building a single-cell transcriptome treasure chest

Yiwen Wang

2024/05/17

CHENG LAB

More than 20 years ago, the code was "cracked"



"All the News That's Fit to Print"

The New York Times

Late Edition

New York: Today, afternoon thunderstorms, high 88. Tonight, showers end, low 67. Tomorrow, partly cloudy with showers late, high 81. Yesterday, high 88, low 74. Weather map, Page 26.

VOL. CXLIX . . No. 51,432 Copyright © 2000 The New York Times NEW YORK, TUESDAY, JUNE 27, 2000 $1 beyond the greater New York metropolitan area 75 CENTS

Genetic Code of Human Life Is Cracked by Scientists

JUSTICES REAFFIRM MIRANDA RULE, 7-2; A PART OF 'CULTURE'

By LINDA GREENHOUSE

WASHINGTON, June 26 — The Supreme Court reaffirmed the Miranda decision today by a 7-to-2 vote that erased a shadow over one of the most famous rulings of modern times and acknowledged that the Miranda warnings "have become part of our national culture."

Justices Antonin Scalia and Clarence Thomas cast the dissenting votes.

The decision overturned a ruling last year by the federal appeals court in Richmond, Va., which held that Congress was entitled to the last word because Miranda's presump-

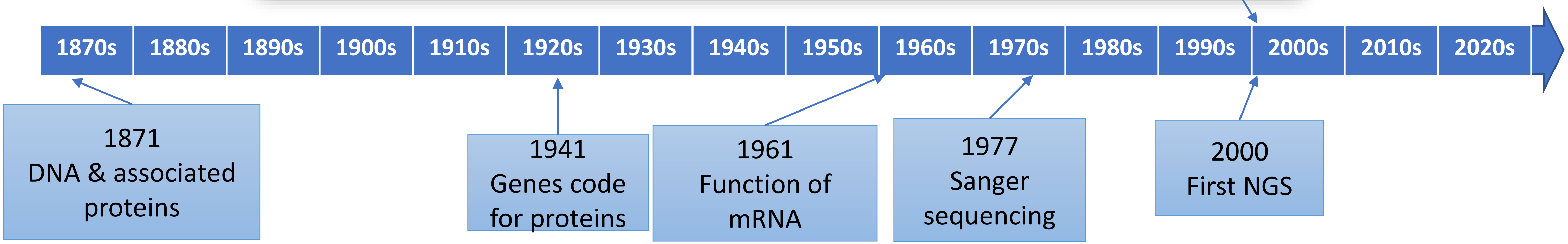
The Book of Life
The three billion base pairs ...
... of the intertwining double helix of DNA ...
... that make up the set of chromosomes in our cells, have been sequenced.
BASE PAIRS Rungs between the strands of the double helix
BASES
A adenine
C cytosine
G guanine
T thymine
By ordering the base units, scientists hope to locate the genes and determine their functions.
The New York Times

A SHARED SUCCESS

2 Rivals' Announcement Marks New Mecical Era, Risks and All

By NICHOLAS WADE

WASHINGTON, June 26 — In an achievement that represents a pinnacle of human self-knowledge, two rival groups of scientists said today that they had deciphered the hereditary script, the set of instructions



1871
DNA & associated proteins

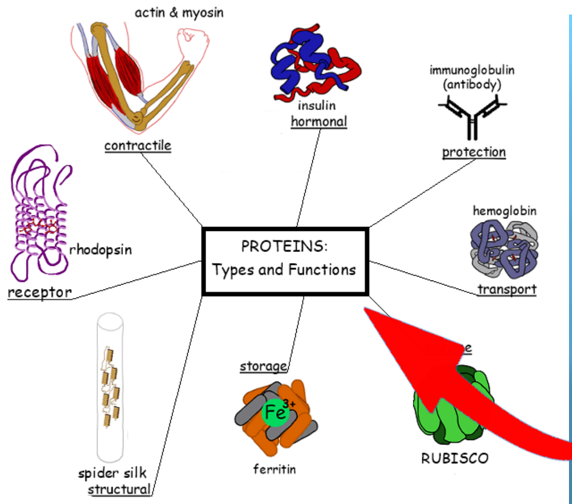
1941
Genes code for proteins

1961
Function of mRNA

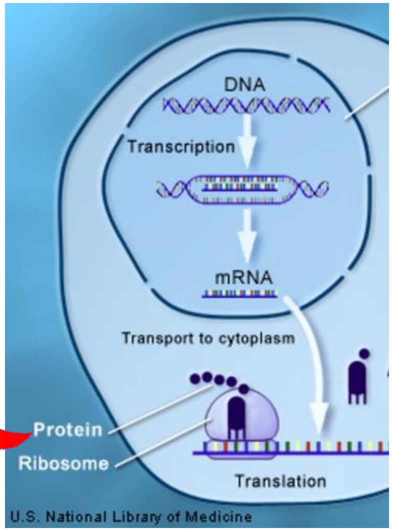
1977
Sanger sequencing

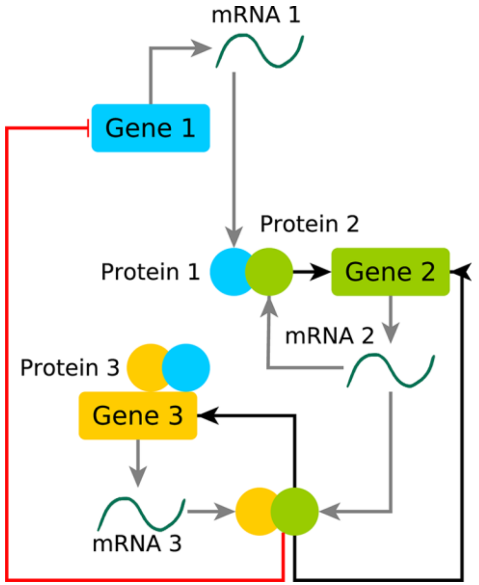
2000
First NGS

A key for deciphering how the code is executed: Single-cell transcriptome (SCT)

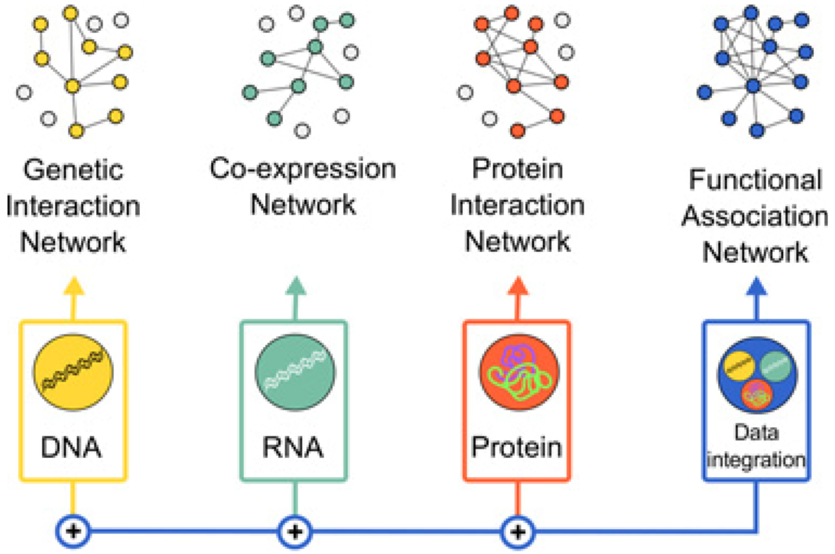


Ciechanover A, Front. Young Minds (2021)

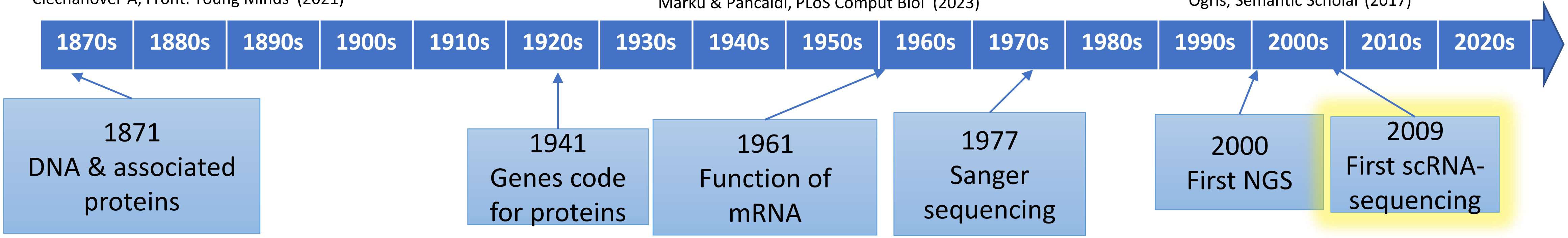




Marku & Pancaldi, PLoS Comput Biol (2023)


Genetic Interaction Network
Co-expression Network
Protein Interaction Network
Functional Association Network
DNA
RNA
Protein
Data integration

Ogris, Semantic Scholar (2017)



1871
DNA & associated proteins

1941
Genes code for proteins

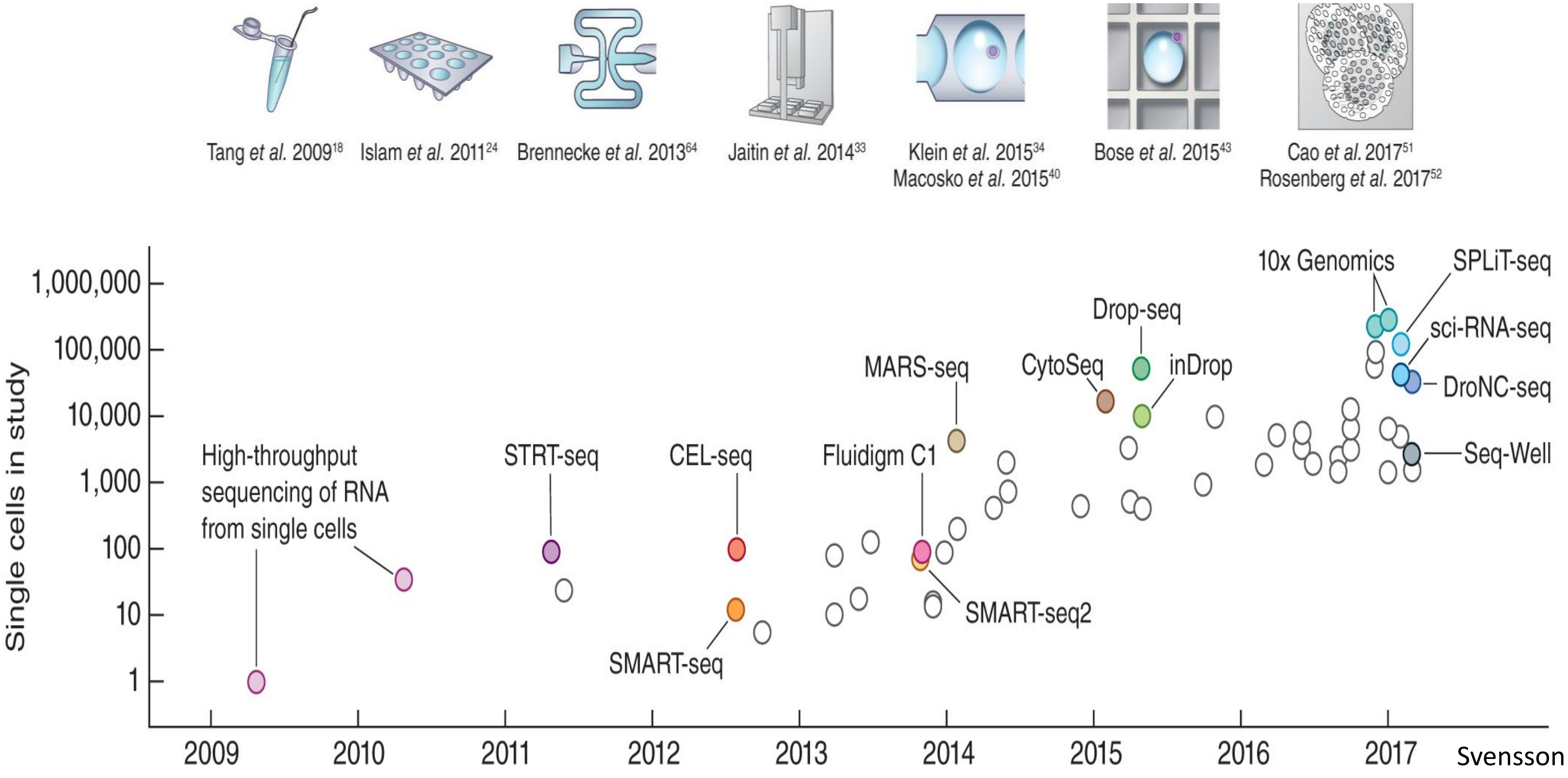
1961
Function of mRNA

1977
Sanger sequencing

2000
First NGS

2009
First scRNA-sequencing

Rapid advances of single-cell RNA sequencing technologies since 2009

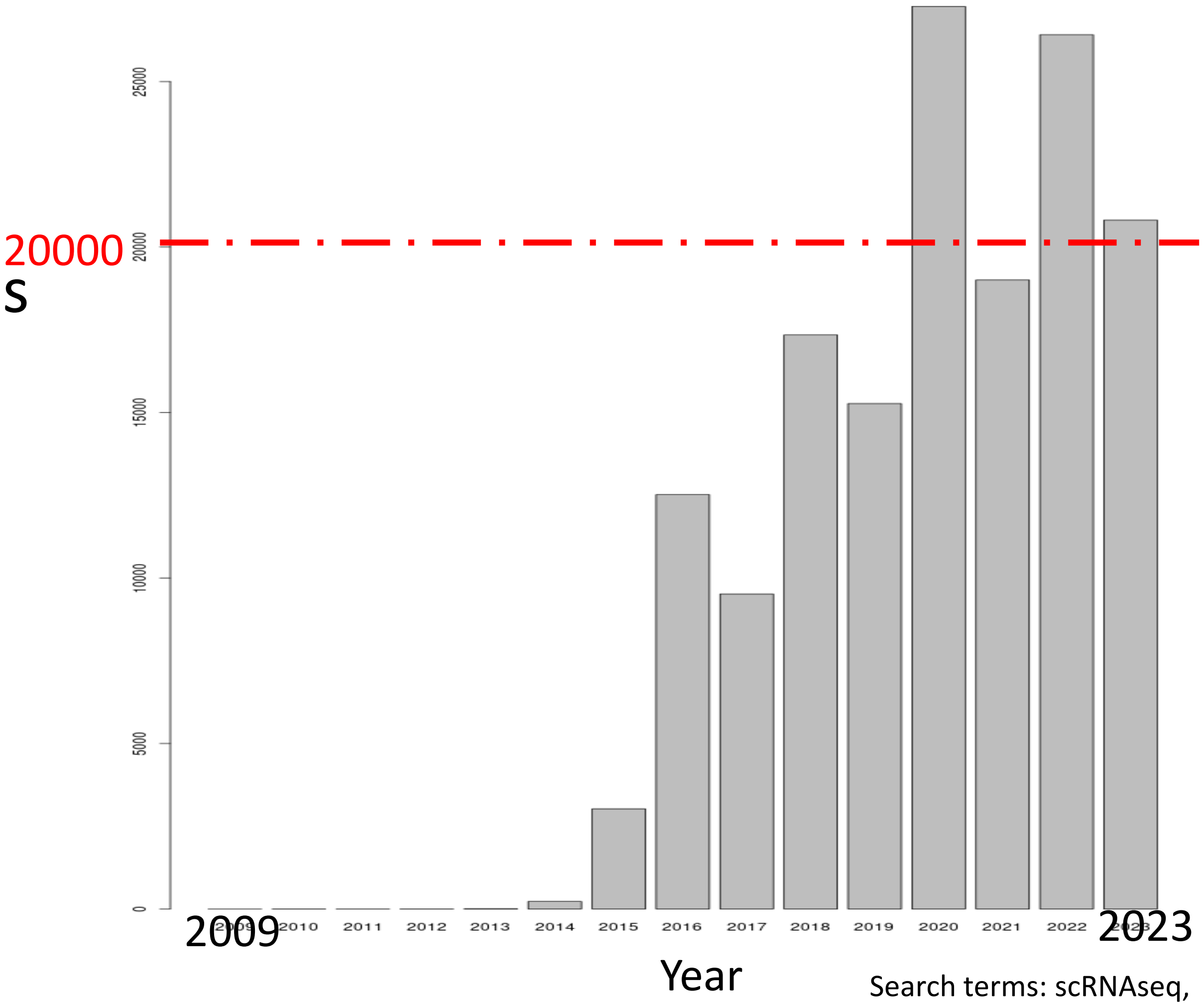


Svensson et al., Nat Protoc (2018)

1870s 1880s 1890s 1900s 1910s 1920s 1930s 1940s 1950s 1960s 1970s 1980s 1990s 2000s **2010s** **2020s**

SCT data in GDS: where the treasure's buried

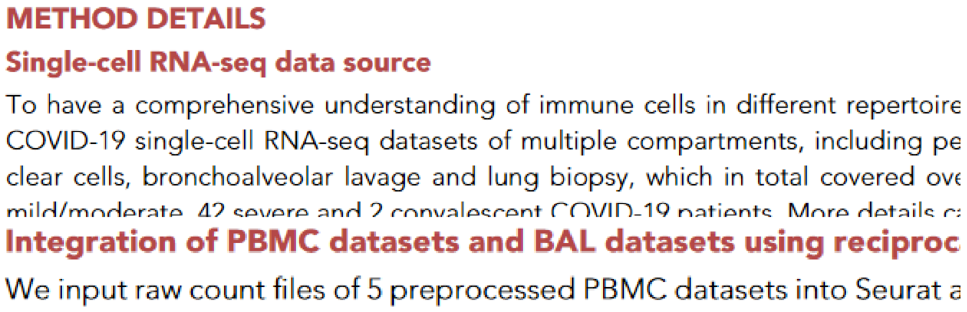
- More than 20,000 scRNA-seq studies per year since 2022 uploaded to GEO DataSet (GDS)
- SCT treasure hunt: dig out values buried in GDS
    - ✓ Build curated database (treasure chest)
    - ✓ Build bioinfo tools
    - ✓ Build cell atlas
    - ✓ Benchmark
    - ✓ Generate hypothesis
    - ✓ Find supporting evidence



Search terms: scRNAseq, scRNA-seq, “single-cell RNA”

SCT treasure hunt representative examples

Curated database (treasure chest)	Build bioinfo tools	Cell atlas
CZ CELL×GENE DISCOVER GepLiver GREIN SIB IMMUCAN scLiverDB TISCH2	SEURAT CellTypist SCENIC CellChat	HuBMAP Azimuth THE UNIVERSITY OF TEXAS MD Anderson ~~Cancer~~ Center T Cell Map HUSCH Cincinnati Children's changing the outcome together ToppCell HCCDB



METHOD DETAILS

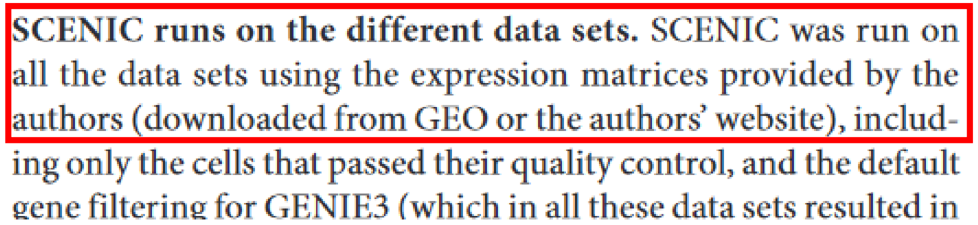
Single-cell RNA-seq data source

To have a comprehensive understanding of immune cells in different repertoire COVID-19 single-cell RNA-seq datasets of multiple compartments, including pe clear cells, bronchoalveolar lavage and lung biopsy, which in total covered ove mild/moderate, 42 severe and 2 convalescent COVID-19 patients. More details c

Integration of PBMC datasets and BAL datasets using reciproc

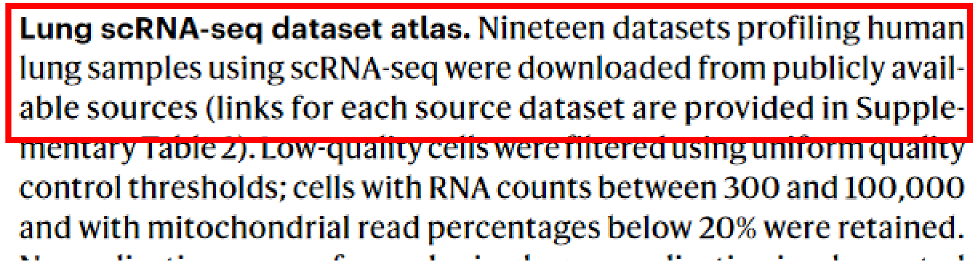
We input raw count files of 5 preprocessed PBMC datasets into Seurat a

Jin et al., iScience, 2021



SCENIC runs on the different data sets. SCENIC was run on all the data sets using the expression matrices provided by the authors (downloaded from GEO or the authors' website), including only the cells that passed their quality control, and the default gene filtering for GENIE3 (which in all these data sets resulted in

Aibar et al., Nature Methods, 2017

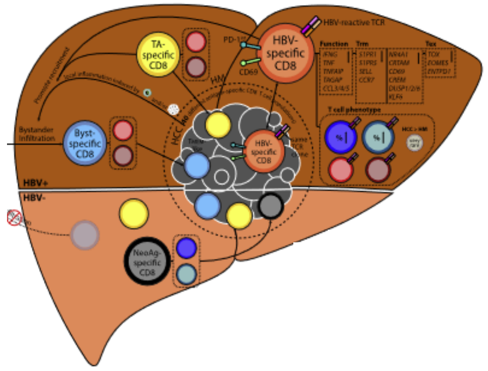


Lung scRNA-seq dataset atlas. Nineteen datasets profiling human lung samples using scRNA-seq were downloaded from publicly available sources (links for each source dataset are provided in Supplementary Table 2). Low-quality cells were filtered using uniform quality control thresholds; cells with RNA counts between 300 and 100,000 and with mitochondrial read percentages below 20% were retained.

Hao et al., Nature Biotechnology, 2023

Start SCT treasure hunt for our research

Define	
Research topic	Goal
scRNA-seq Hepatocellular carcinoma human liver tissue	T-cell responses in HBV vs. non-HBV



Data from GDS

- Query on website
- Query by scripts (R, Python, bash…)
- Download

Pipeline for

✓Re-analysis

✓Benchmark

✓Meta-analysis

✓Integrate and investigate

CellTypist

SCTK

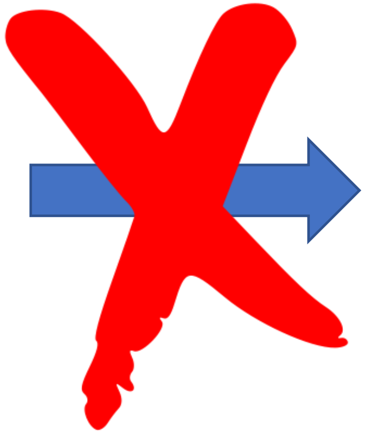
SCALA
Single Cell Analysis for All

Obstacles to SCT treasure hunt

Define	
Research topic	Goal
scRNA-seq Hepatocellular carcinoma human liver tissue	T-cell responses in HBV vs. non-HBV

Data from GDS

- Diverse file formats
- Messy annotations



Pipeline for

✓Re-analysis

✓Benchmark

✓Meta-analysis

✓Integrate and investigate

R

CellTypist

SCTK

SCALA
Single Cell Analysis for All

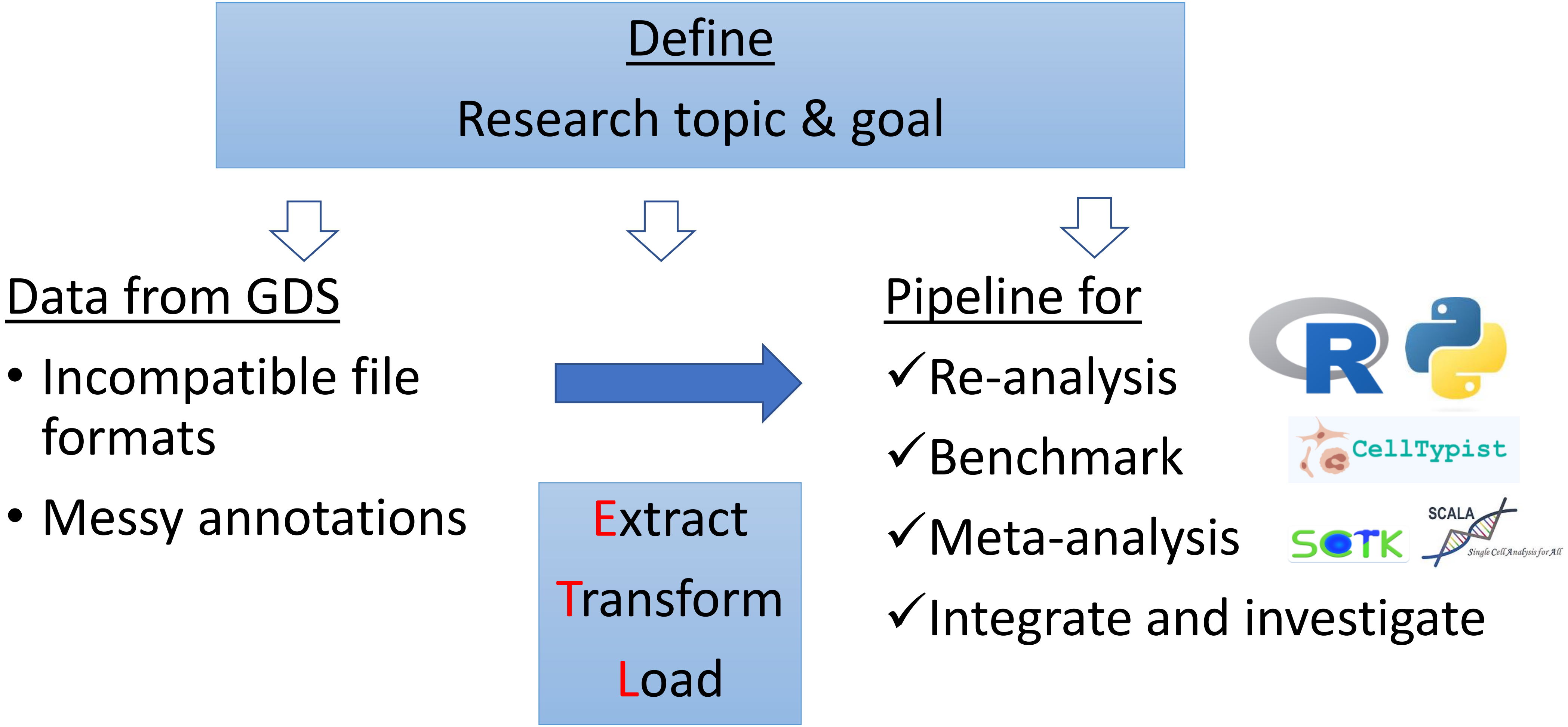
Peek into SCT treasure chests built by others

Define	
Research topic	**Goal**
scRNA-seq Hepatocellular carcinoma human liver tissue	T-cell responses in HBV vs. non-HBV

✗ ✗

Curated database	HCC & human liver tissue	HBV vs. non-HBV
CZ CELL×GENE DISCOVER	0	N/A
SIB IMMUCAN	4	N/A
GREIN	3	N/A
GepLiver	15	N/A
scLiver DB	6	N/A

ETL is critical for SCT treasure hunt



Define

Research topic & goal

Data from GDS

- Incompatible file formats
- Messy annotations

Extract

Transform

Load

Pipeline for

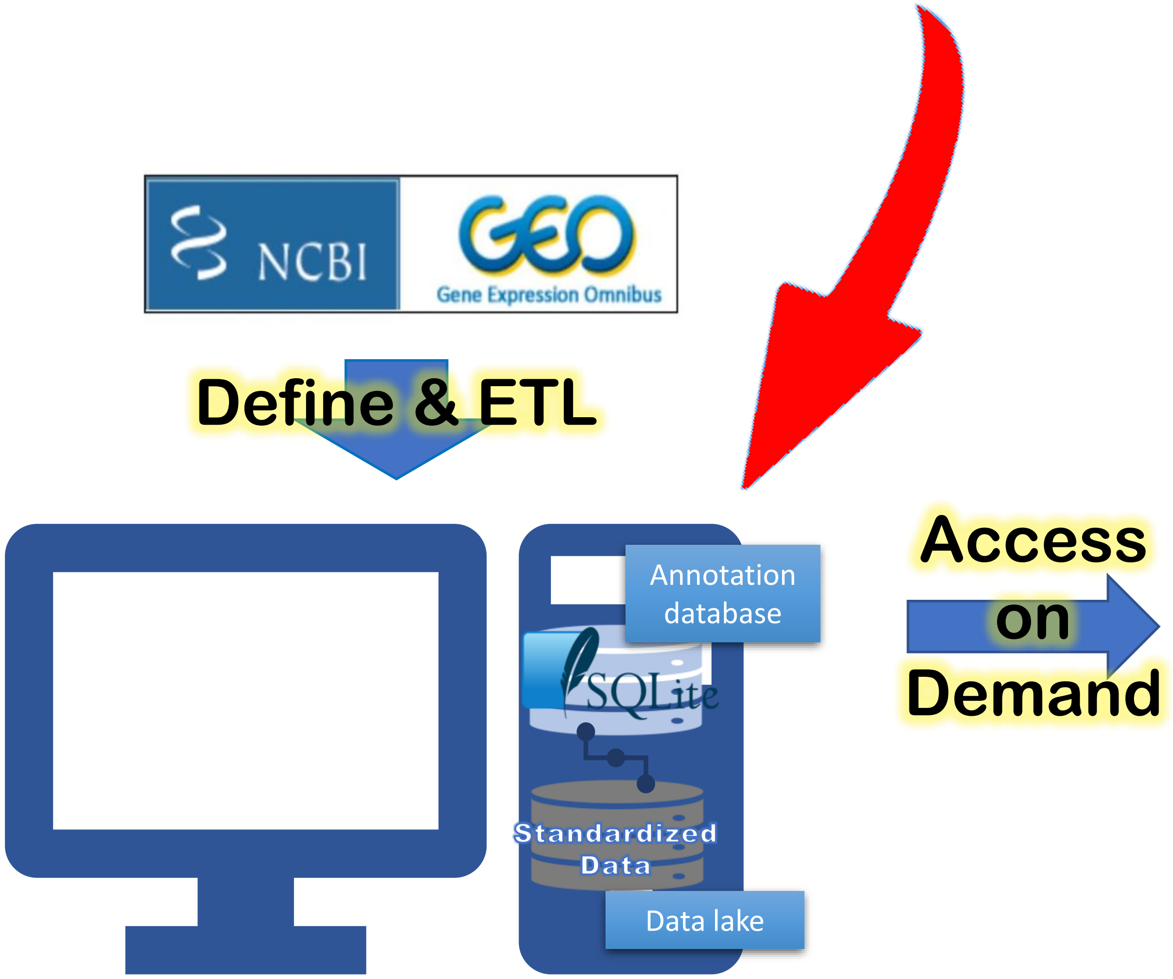
✓Re-analysis

✓Benchmark

✓Meta-analysis

✓Integrate and investigate

Build a SCT treasure chest of our own



Define & ETL

Annotation database

Standardized Data

Data lake

Access on Demand

<u>Pipeline for</u>

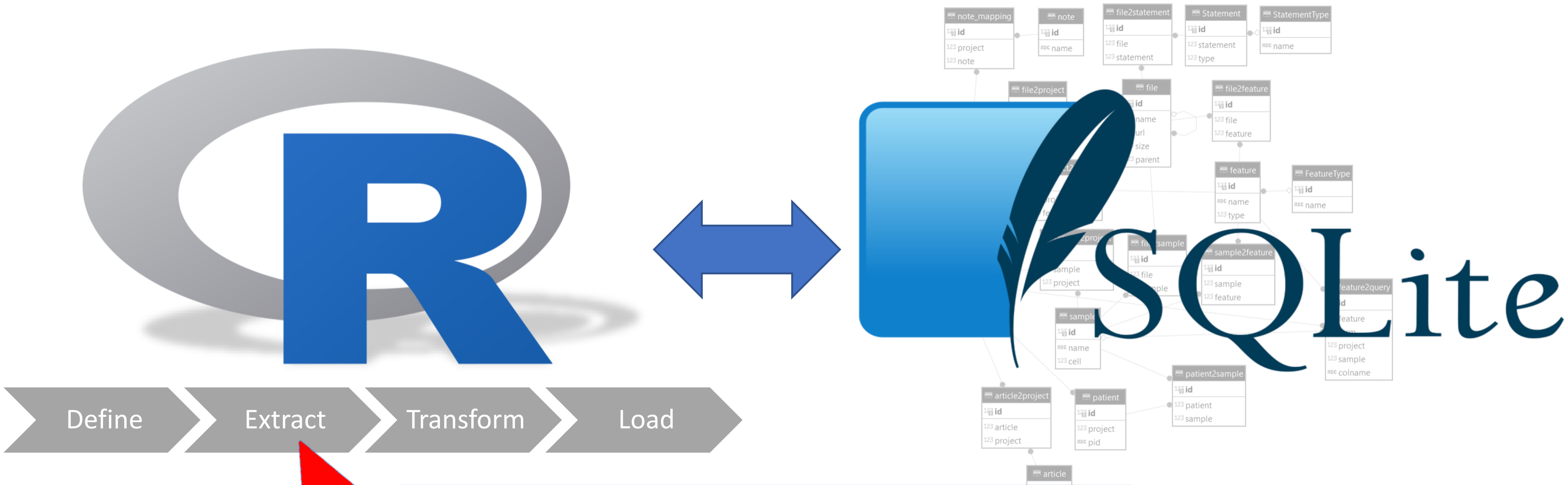
✓Re-analysis

✓Benchmark

✓Meta-analysis

✓Integrate and investigate

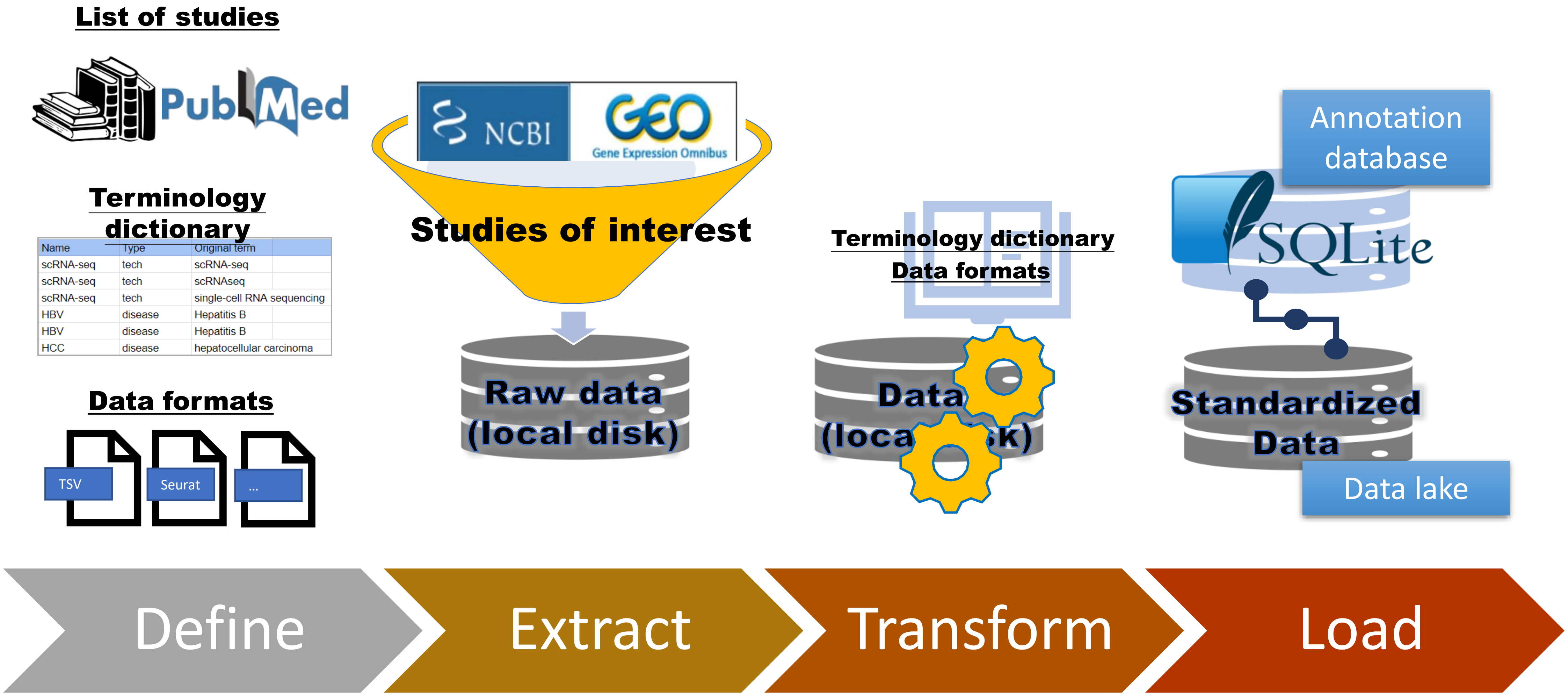
A framework for building SCT treasure chest



Define → Extract → Transform → Load

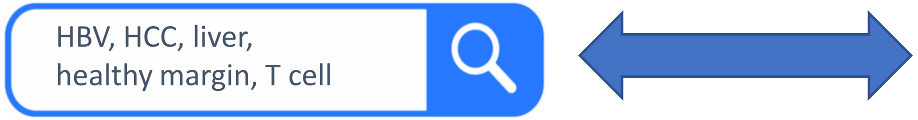
- ✓No server needed
- ✓Only R programming required
- ✓Set up in a few days

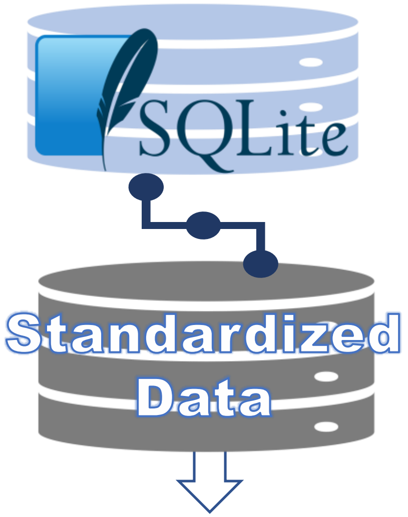
Pipeline for building SCT treasure chest



Name	Type	Original term
scRNA-seq	tech	scRNA-seq
scRNA-seq	tech	scRNAseq
scRNA-seq	tech	single-cell RNA sequencing
HBV	disease	Hepatitis B
HBV	disease	Hepatitis B
HCC	disease	hepatocellular carcinoma

On-demand access for downstream pipelines





Annotation database

Data lake

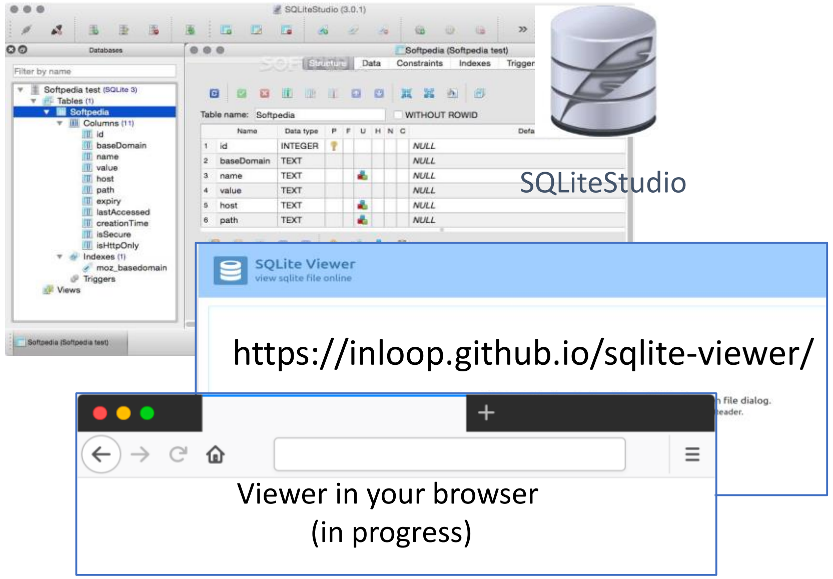


Table name: Softpedia ☐ WITHOUT ROWID

	Name	Data type	P	F	U	H	N	C	Defa
1	id	INTEGER	🔑						*NULL*
2	baseDomain	TEXT							*NULL*
3	name	TEXT			●				*NULL*
4	value	TEXT							*NULL*
5	host	TEXT			●				*NULL*
6	path	TEXT			●				*NULL*

SQLiteStudio

https://inloop.github.io/sqlite-viewer/

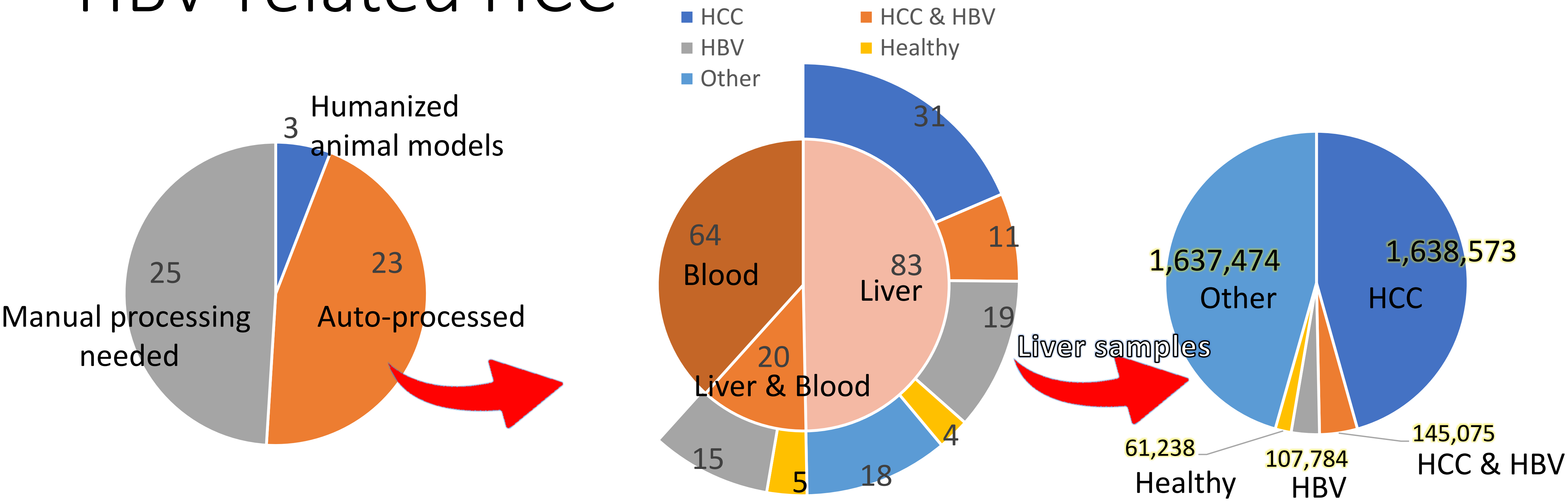
Viewer in your browser
(in progress)

Pipeline for

- ✓Re-analysis
- ✓Benchmark
- ✓Meta-analysis
- ✓Integrate and investigate

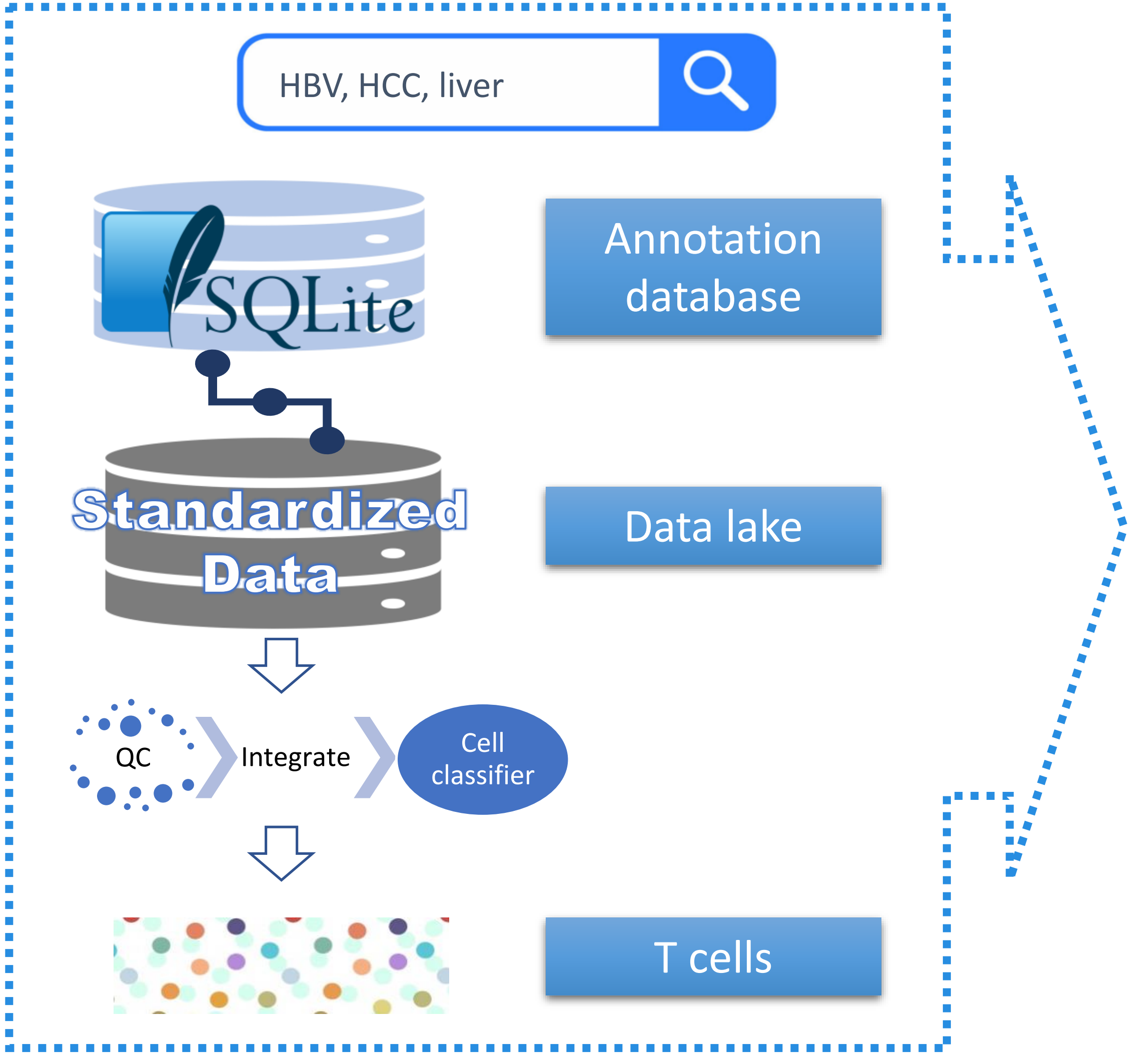
R CellTypist SCTK SCALA *Single Cell Analysis for All*

Use case: a SCT treasure chest for studying HBV-related HCC



61 datasets found on GDS queried using keywords: scRNA-seq & (HCC | HBV)	**167 human subjects** **in 23 auto-processed datasets** grouped by sources and diseases (4,088,738 cells)	**3,590,144 cells** **from liver samples** grouped by available disease labels

Tap in SCT treasure chest for T-cell responses in HBV-related HCC



HBV, HCC, liver

SQLite

Annotation database

Standardized Data

Data lake

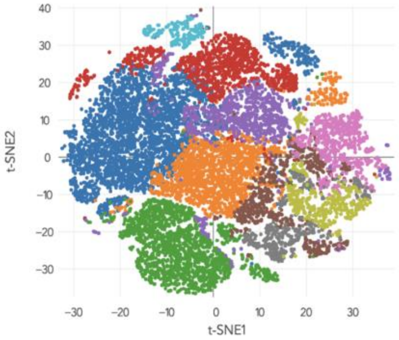
QC

Integrate

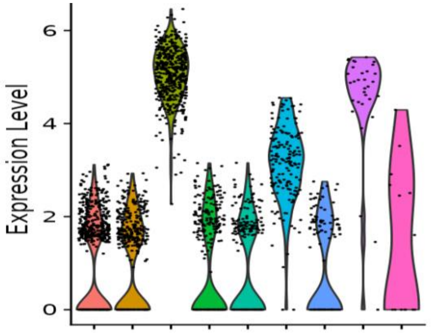
Cell classifier

T cells

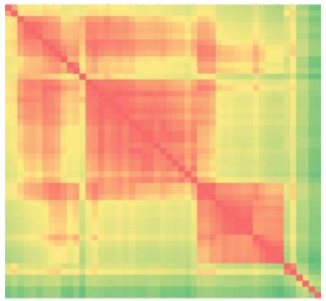
Identify subtypes of T cells



Compare gene expression across conditions



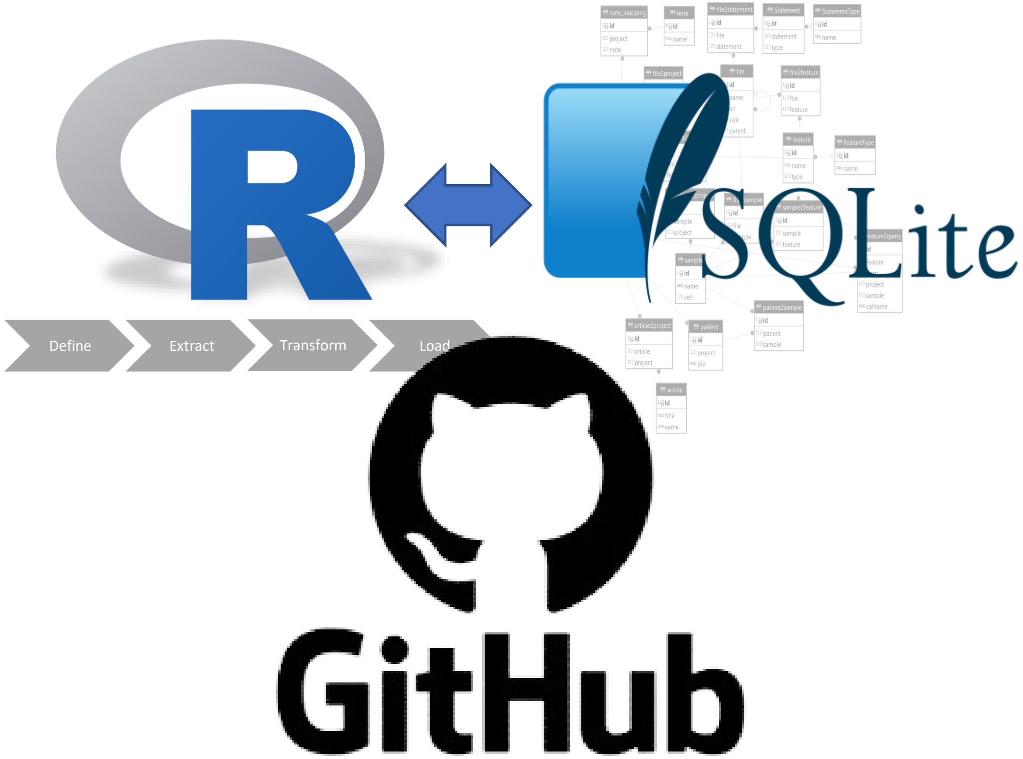
Identify marker genes

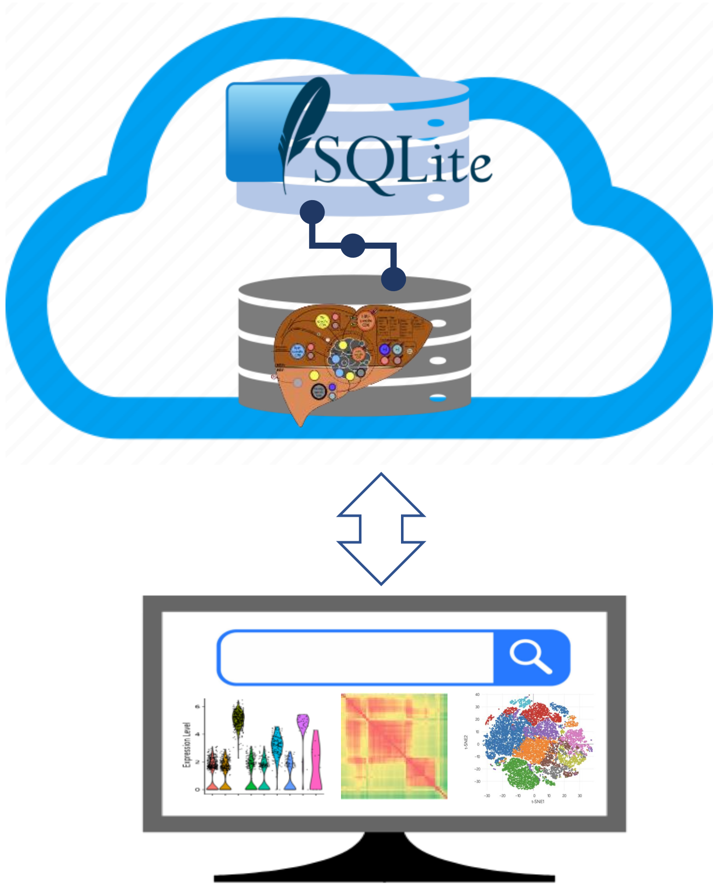


Deconvolute cells in spatial transcriptome

Note: Figures are for illustration purposes only.

The framework and our curated database will be available online soon





Contact us if you are interested

- Have a taste with our HBV-related HCC treasure chest
- Build a customized SCT treasure chest for your own research
- Suggest features that might be useful for you
- Join the project

Thanks for your attention!