



EMBO Workshop 2024

Hackathon

The Spatial Biology of Cancer

Reminders

- The Hackathon is based on **R programming language** (v. \geq 4.4.0.)
- You must have already download all the needed packages/dependencies needed.
- All information is available @
https://github.com/matteocereda/EMBO_SBC_hackathon

Organization of the files

```
se) mgrieco@dhcp-162-24 repo % cd EMBO_SBC_hackathon  
se) mgrieco@dhcp-162-24 EMBO_SBC_hackathon %  
se) mgrieco@dhcp-162-24 EMBO_SBC_hackathon % ll  
al 2232  
-r--r--@ 1 mgrieco staff 1131123 Aug 20 10:31 00_dependencies.html  
-r--r--@ 1 mgrieco staff 1535 Jul 23 11:21 00_environment.R  
-r--r--@ 1 mgrieco staff 6350 Aug 6 10:49 00_helper_functions.R  
xr-xr-x@ 35 mgrieco staff 1120 Aug 23 10:50 input  
xr-xr-x@ 53 mgrieco staff 1696 Aug 26 11:44 level1  
xr-xr-x@ 46 mgrieco staff 1472 Aug 24 09:00 level2
```

Instructions and questions for level 1

Instructions and questions for level 2

Input folder containing ST datasets

Working directory of the repository available [here](#):

Required packages for the hackathon

Where to set your path to the current working directory

Where to set your path to the current working directory

To download this repository locally on your laptop, you can paste this line to the bash console:

```
(base) mgrieco@dhcp-162-24 repo %  
(base) mgrieco@dhcp-162-24 repo % git clone https://github.com/matteocereda/EMBO_SBC_hackathon.git
```

Reminders

- You must have chosen your **team**mates
- **Team** should be as much as **heterogeneous** as possible
- Team must have a **leader**
- ONLY team **leaders** will send the final solution to the organisers.



Final solution

- The final **solution** should contain:
 1. a **PDF** file with answers to our questions
 2. a **R** file with answers to our questions
- The team **leader** will send these files to:
mariachiara.grieco@unimi.it
- The quickest and most correct solution will win the competition.

Organization of the hackathon

1 Level 1

Visium 10X Genomics

Map the whole transcriptome within the tissue context

2 Level 2

CosMx Spatial Molecular Imaging Nanostring

Resolve the precise position of RNA molecules in tissue section at highest spatial resolution

Timetable

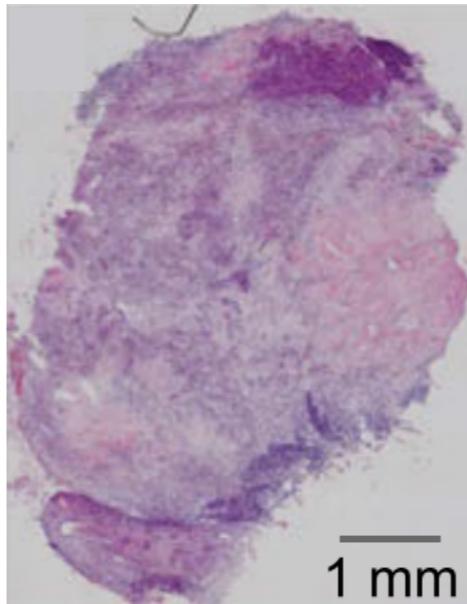
13:30-14:00	Introduction
14:00-15:30	LEVEL 1
15:30-16:00	Disclosure solution level 1
16:00-17:30	LEVEL 2
17:30 -18:00	Conclusion

Helpers

Helpers are here to give
support and direction but
not solutions.

THE CASE STUDY

High-grade
serous ovarian
carcinoma
(HGSOC)



THE TASK

Identify a receptor tyrosine kinases
involved in OC progression

Level steps

- 1 Loading raw input data into R**
- 2 Quality check (QC)**
- 3 Data pre-processing and clustering**
- 4 Functional interpretation of cluster marker genes**
- 5 Cell type annotation for the identified clusters using cell type gene signatures**
- 6 Ligand-receptor signalling analysis**

Tools

Seurat

Satija R. et al. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol (2015)

Clustermole

<https://cran.r-project.org/web/packages/clustermole/clustermole.pdf>

ClusterProfiler

Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012

LIANA (Ligand-Receptor Interactions Analysis)

Dimitrov, D., Türei, D., Garrido-Rodríguez, M. et al. Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. Nat Commun (2022)

Databases

Molecular Signatures Database (MSigDB)

Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell systems (2015)

Network of Cancer Genes (NCG)

Dressler, L., Bortolomeazzi, M., Keddar, M.R. et al. Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the Network of Cancer Genes (NCG) resource. Genome Biol (2022)

Clustermole markers database

<https://cran.r-project.org/web/packages/clustermole/vignettes/clustermole-intro.html>

OmniPath database

Türei D, Valdeolivas A, Gul L, et al. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. Mol Syst Biol (2021)

Supporting Information

High-grade serous ovarian carcinoma is the most common type of ovarian cancer



Incidence

High-grade serous ovarian carcinoma (HGSOC) accounts for approximately 75% of epithelial ovarian cancers.



Nomenclature

High-grade refers to ovarian carcinomas that are classified as Grade 3.

A tumor's grade refers to the type of cells composing tumor, based on how the cells and tissue look under a microscope.

HGSOC cells and tissue appear highly abnormal compared to healthy cells and tissue under microscope.
HGSOC cells are poorly differentiated: do not have a clear structure or pattern

Serous means that the tumor arose from the serous membrane



Molecular characteristics

Genetically is characterized by almost universal TP53 mutations and copy number alterations (CNAs)

Genomic heterogeneity

Commonly affected genes: FAT3, CSMD3, BRCA1, BRCA2, NF1, CDK12, GABRA6, RB1, PTEN, RAD51B

Alterations in these genes are found only in a small fraction of tumors

Chromosomal Instability

- Most HGSOCs are polyclonal
- As it HGSOC progresses, clonal diversity increases → worse prognosis and development of chemoresistance

Our case study

The analysis aims to characterise **molecular mechanisms** driving tumor onset resolving **spatial tissue architecture**.

We will characterise a tissue section from a primary tumour of a **high-grade serous ovarian carcinoma (HGSO)**.

The tumour sample was collected during interval de-bulking surgery from a patient with a good response to taxane- and platinum-based neo-adjuvant chemotherapy (NACT) treatment.

The same tumour sample has been subsequently profiled using **different spatially-resolved technologies**.

The available tissue section includes a large stromal/immune compartment, together with the tumour epithelium. Therefore, it is suitable for studying the crosstalk between tumour cells and their associated microenvironment.

nature communications

Article

<https://doi.org/10.1038/s41467-024-47271-y>

Spatial transcriptomics reveals discrete tumour microenvironments and autocrine loops within ovarian cancer subclones

Received: 5 October 2022

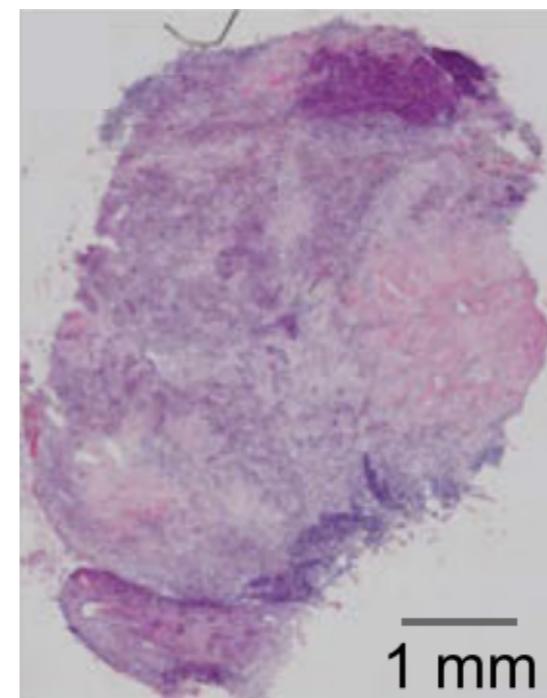
Elena Denisenko  , Leanne de Kock  , Adeline Tan², Aaron B. Beasley  , Maria Beilin⁴, Matthew E. Jones¹, Rui Hou¹, Dáithí Ó Muirí¹, Sanelia Bilic⁴, G. Raj K. A. Mohan^{4,5}, Stuart Salfinger⁶, Simon Fox  , Khaing P. W. Hmon¹, Yen Yeow  , Youngmi Kim⁷, Rhea John⁷, Tami S. Gilderman⁷, Emily Killingbeck  , Elin S. Gray  , Paul A. Cohen  & Alistair R. R. Forrest  

Accepted: 26 March 2024

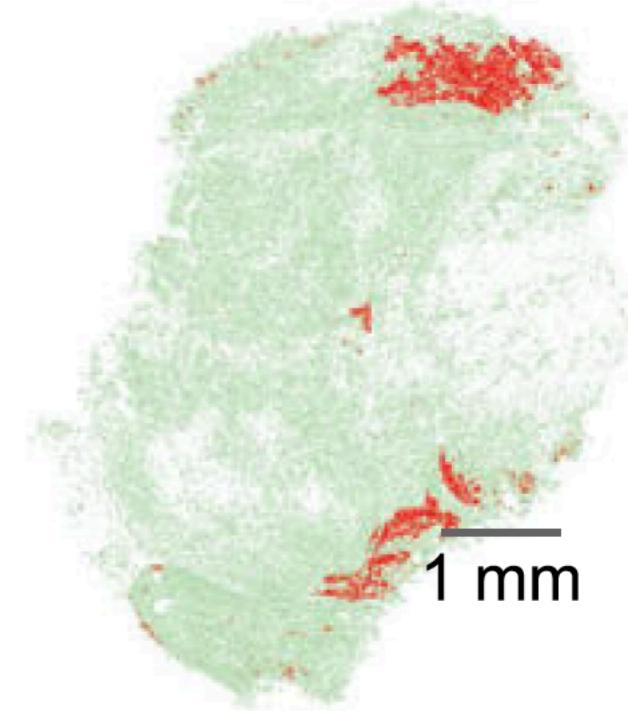
Published online: 03 April 2024

 Check for updates

H&E staining



QuPath cell classification



Spatial Transcriptomics (ST) technologies exploited

Visium (10x Genomics)

- Technology Type: **Sequencing-based spatial transcriptomics**
- Spatial mapping of gene expression within tissue sections
- Pseudo-bulk resolution
- Useful for understanding the spatial organization of different cell types within a tumor and how they interact within their microenvironment

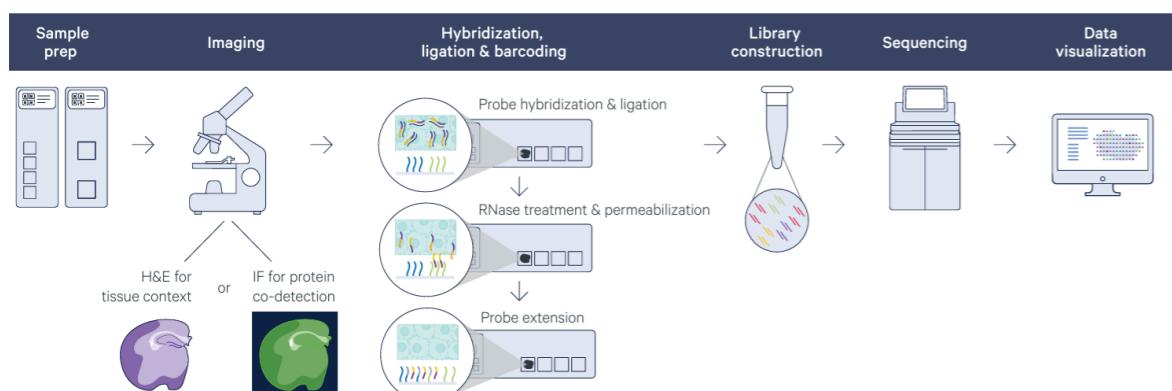
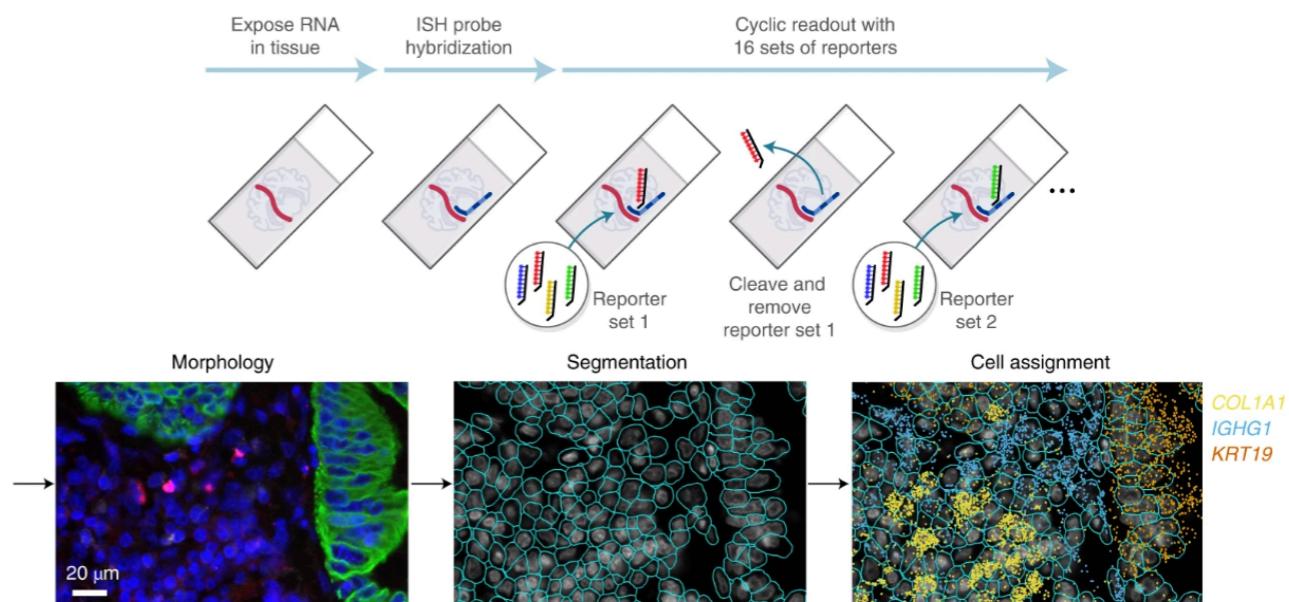


Image taken from <https://www.10xgenomics.com>

CosMx™ Spatial Molecular Imaging (SMI) (NanoString®)

- Technology Type: **Imaging-based spatial single-cell transcriptomics**
- In situ visualization and quantification of RNA transcripts and/or proteins
- Single cell boundaries are segmented based on immunofluorescent staining markers
- Enables detailed mapping of molecular interactions within the tissue, providing insights into the cellular architecture and microenvironment



He, S., Bhatt, R., Brown, C., Brown, E.A., Buhr, D.L., Chantranuvatana, K., et al. High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nat Biotechnol.* (2022)

Level 1 - Analysis of 10X Genomics Visium data

1 Loading raw input data into R

10X Genomics Visium raw data must be loaded and arranged to build a Seurat object that will be used for downstream analysis.

2 Quality check (QC)

Visually explore 10X Genomics Visium data to get an idea of its sequencing quality

3 Data pre-processing and clustering

Normalize 10X Genomics Visium gene expression data and perform unsupervised clustering following the standard Seurat processing pipeline.

4 Functional interpretation of cluster marker genes

Identify cancer cell-driven clusters by interrogating pathways (MSigDB) and cancer driver genes (NCG) databases.

5 Cell type annotation for the identified clusters using cell type gene signatures

Use clustermole gene signature database to assign each 10X Genomics Visium spot cluster its most similar cell type.

6 Ligand-receptor signalling analysis

Use LIANA toolkit to explore the most relevant ligand-receptor interactions occurring between cancer and immune cells.

Level 2 - Analysis of CosMx data

1 Loading raw input data into R

The CosMx raw data available for our HGSOC sample shows a slightly different format compared to the one typically recognised by Seurat. Therefore, you will need to adjust the input data format before being able to build the Seurat object that you will use for downstream analysis.

2 Quality check (QC)

Visually explore CosMx data to get an idea of its sequencing quality, and remove single cells showing insufficient or no transcriptional features.

3 Data pre-processing and clustering

Normalize single-cell gene expression data and perform unsupervised clustering following the standard Seurat processing pipeline.

4 Functional interpretation of cluster marker genes

Identify cancer cell-driven clusters by interrogating pathways (MSigDB) and cancer driver genes (NCG) databases.

5 Cell type annotation for the identified clusters using cell type gene signatures

Use clustermole gene signature database to assign each CosMx single-cell cluster to its most similar cell type.

6 Ligand-receptor signalling analysis

Use LIANA toolkit to explore the most relevant ligand-receptor interactions occurring between cancer and immune cells.

Databases

Molecular Signatures Database (MSigDB)

Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell systems (2015)

- Is a comprehensive collection of annotated gene sets used for gene set enrichment analysis
- In the analysis, the hallmarks category (H) is used for identifying pathway enrichments within the spatially-resolved clusters

Network of Cancer Genes (NCG)

Dressler, L., Bortolomeazzi, M., Keddar, M.R. et al. Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the Network of Cancer Genes (NCG) resource. Genome Biol (2022)

- Is a literature-based repertoire of 3,355 well-known or predicted drivers of cancer and non-cancer somatic evolution in 122 cancer types and 12 non-cancer tissues
- Includes genes affected by somatic modifications have known or predicted cancer driver roles
- In the analysis, this database is used for identifying cancer driver genes within the dataset

Clustermole markers database

<https://cran.r-project.org/web/packages/clustermole/vignettes/clustermole-intro.html>

- Is a comprehensive resource designed to facilitate the identification and annotation of cell types
- It includes a vast collection of cell-type markers for both human and mouse tissues, covering thousands of different cell types
- In the analysis, this database is used for assigning cell types to spatially-resolved clusters

OmniPath database

Türei D, Valdeolivas A, Gul L, et al. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. Mol Syst Biol (2021)

- Is a comprehensive and curated database and resource of intra- and inter-cellular signalling, specifically focusing on protein-protein interactions, post-translational modifications, and transcriptional regulation

Tools

Seurat

Satija R. et al. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol (2015)

- Is a powerful and versatile tool that provides a structured framework for storing and accessing single-cell data along with methods for different types of analyses
- In the context of ST analysis, Seurat allows to study gene expression within the context of tissue architecture

Clustermole

<https://cran.r-project.org/web/packages/clustermole/clustermole.pdf>

- Is a tool designed to enhance the analysis of spatial transcriptomics
- It helps in the process of identifying and labelling cell types within tissue samples from various databases

ClusterProfiler

Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012

- Is a tool designed for functional enrichment analysis and visualization of gene clusters
- It facilitates efficient data interpretation by providing tools to access, manipulate, and visualize enrichment results

Tools

LIANA (Ligand-Receptor Interactions Analysis)

Dimitrov, D., Türei, D., Garrido-Rodriguez, M. et al. Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. Nat Commun (2022)

- Is a tool designed for analyzing and interpreting cell-cell communication networks.
- It focuses on identifying and evaluating interactions between ligands and receptors within transcriptomics data.

LIANA integrates several methods to infer cell-cell communication by analyzing ligand-receptor interactions:

► **NATMI** - Network Analysis Toolkit for Multicellular Interactions

Hou, R., Denisenko, E., Ong, H.T. et al. Predicting cell-to-cell communication networks using NATMI. Nat Commun (2020)

It predicts and visualizes cell-to-cell communication networks by leveraging connectomeDB2020, a curated list of ligand-receptor pairs, or user-defined pairs

► **iTALK** - identifying and illustrating alterations in intercellular signaling network

Wang Y, Wang R, Zhang S, et al. iTALK: an R Package to Characterize and Illustrate Intercellular Communication. bioRxiv (2019)

Designed for profiling ligand-receptor-mediated intercellular communication from single-cell RNA-seq (scRNA-seq) data.

It identifies ligand-receptor pairs, tracks changes in intercellular communication, and provides functional annotations using a curated database.

► **CellChat**

Jin, S., Guerrero-Juarez, C.F., Zhang, L. et al. Inference and analysis of cell-cell communication using CellChat. Nat Commun (2021)

Designed to infer and analyze intercellular communication networks using scRNA-seq data.

It quantitatively predicts key signaling inputs and outputs for cells, analyzing how these signals coordinate to perform functions.

It employs network analysis, pattern recognition, and manifold learning to classify signaling pathways and identify both conserved and context-specific pathways across different datasets

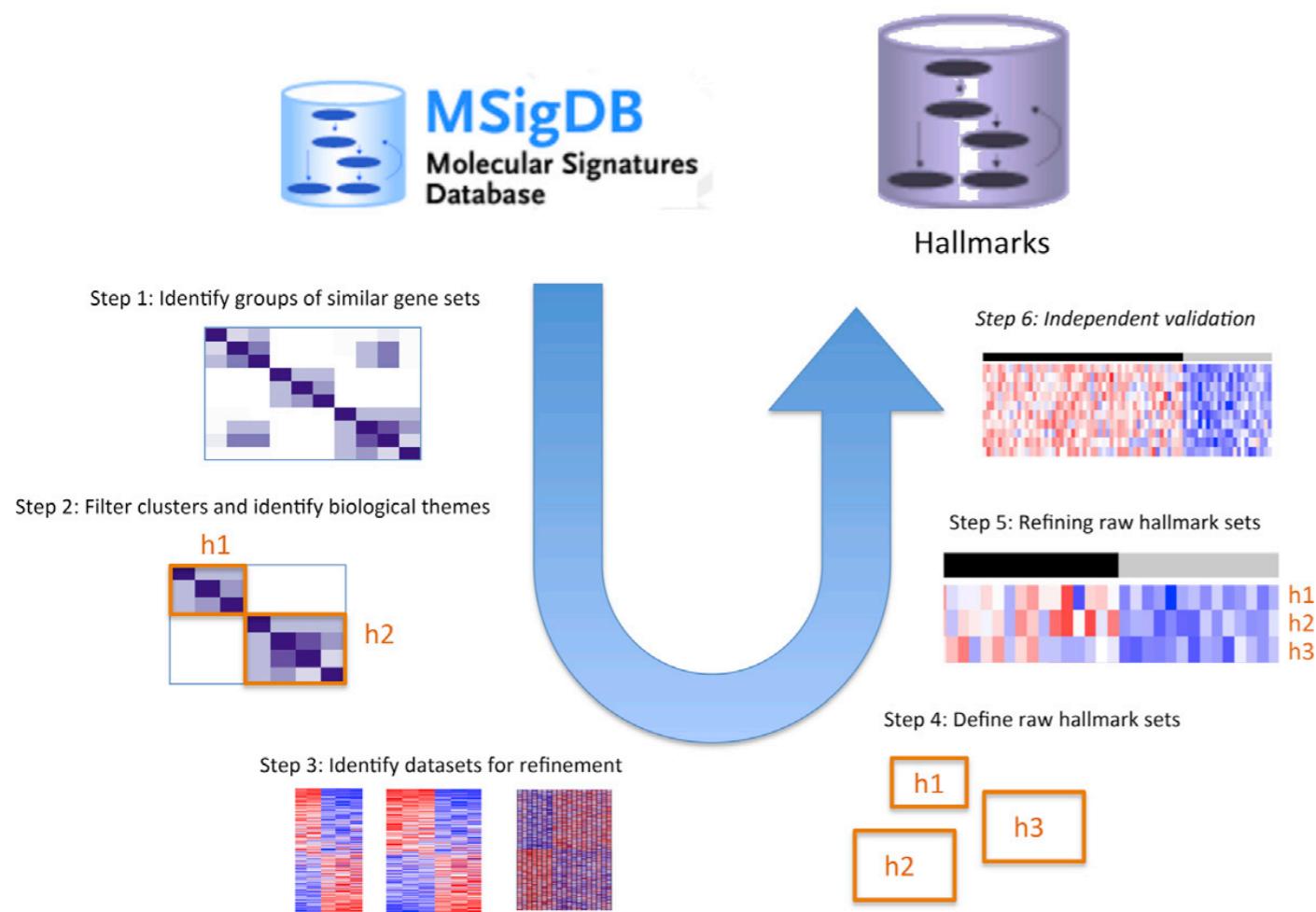
► **SCA** - SingleCellSignalR

Cabello-Aguilar S. et al. SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. NAR (2020)

It processes raw reads count matrices performing data normalization, clustering, and cell-type identification, or it can focus solely on LR interaction inference.

MSigDB hallmark gene set collection

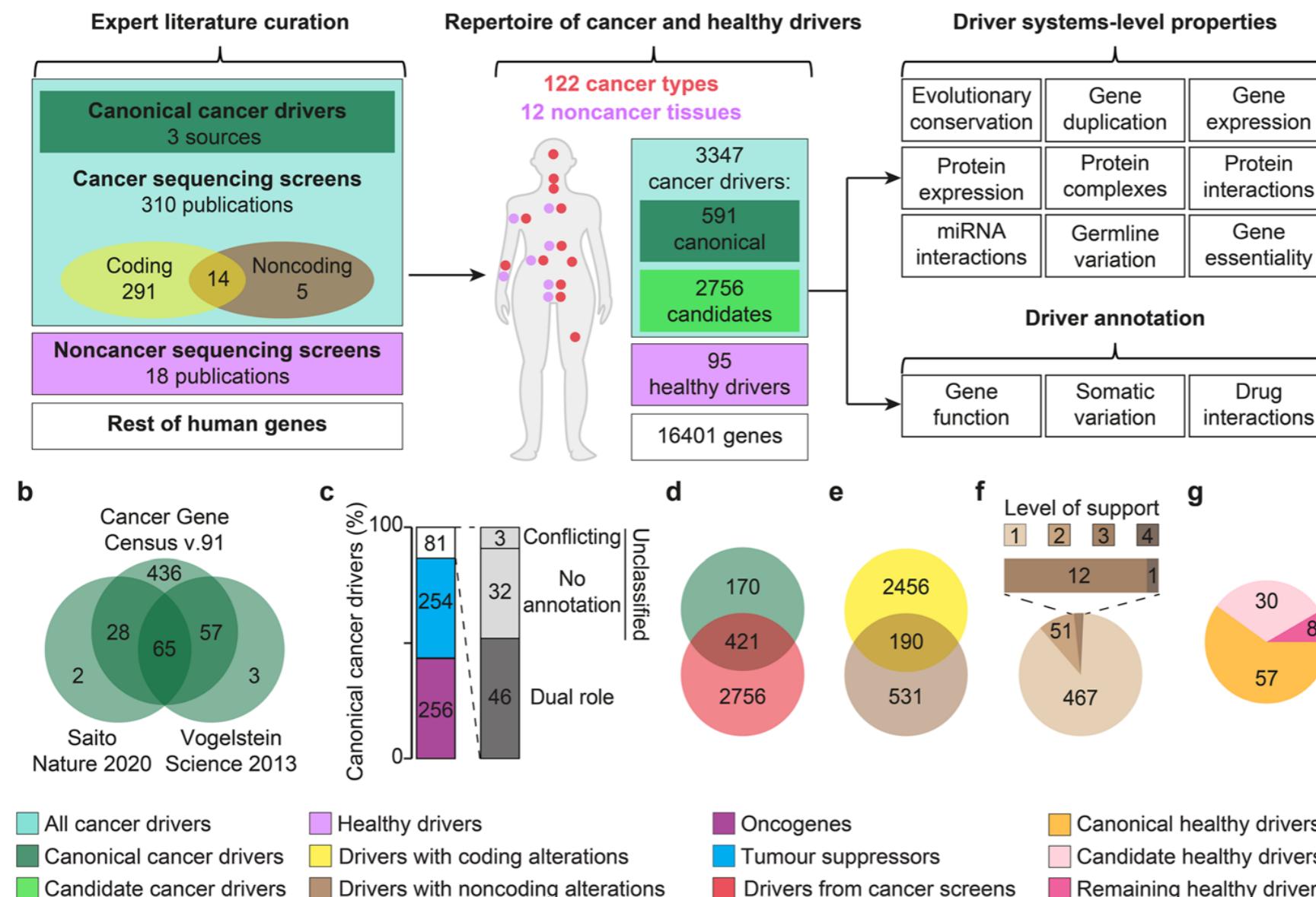
- Hallmark gene sets summarize and represent specific well-defined biological states or processes
- Comprises a set of 50 hallmarks including 4,022 of the original 8,380 MSigDB gene sets
- These gene sets were generated by a computational methodology based on identifying overlaps between gene sets in other MSigDB collections (C1-C6) and retaining genes that display coordinate expression



Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell systems* (2015)

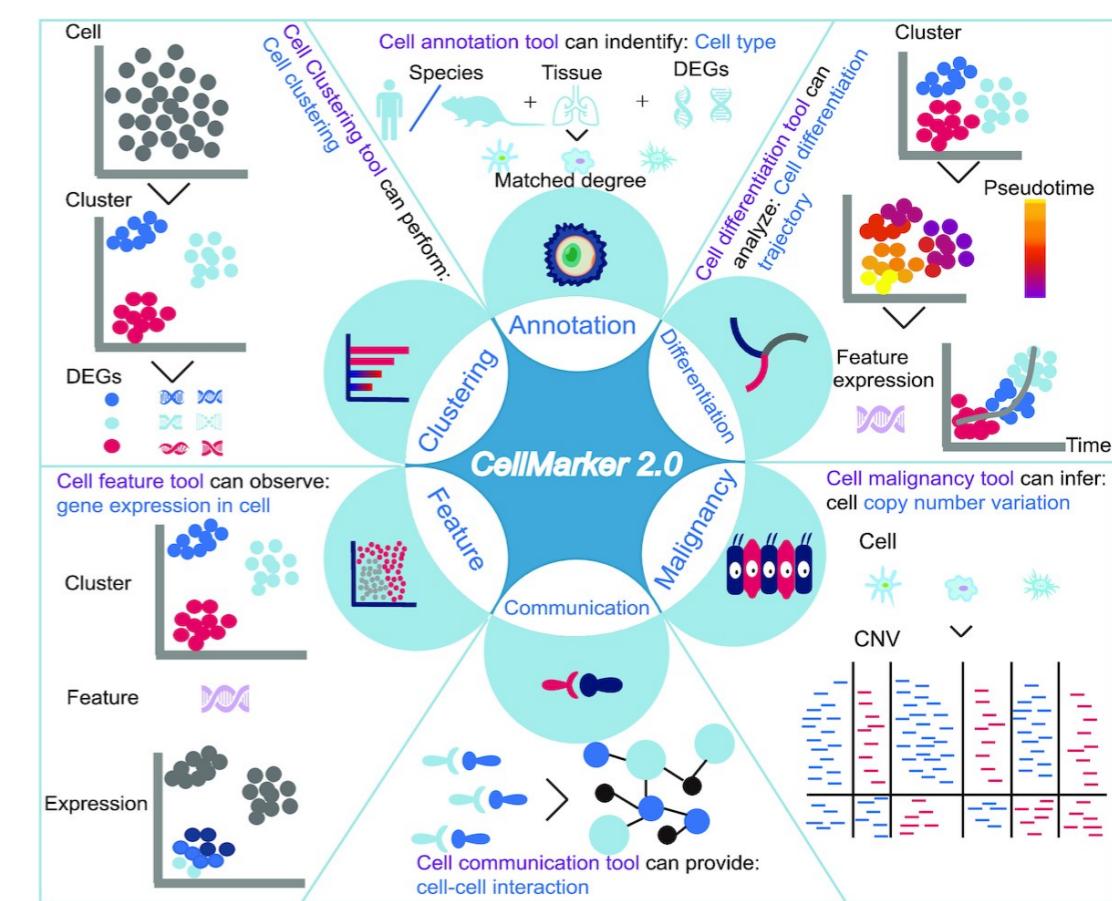
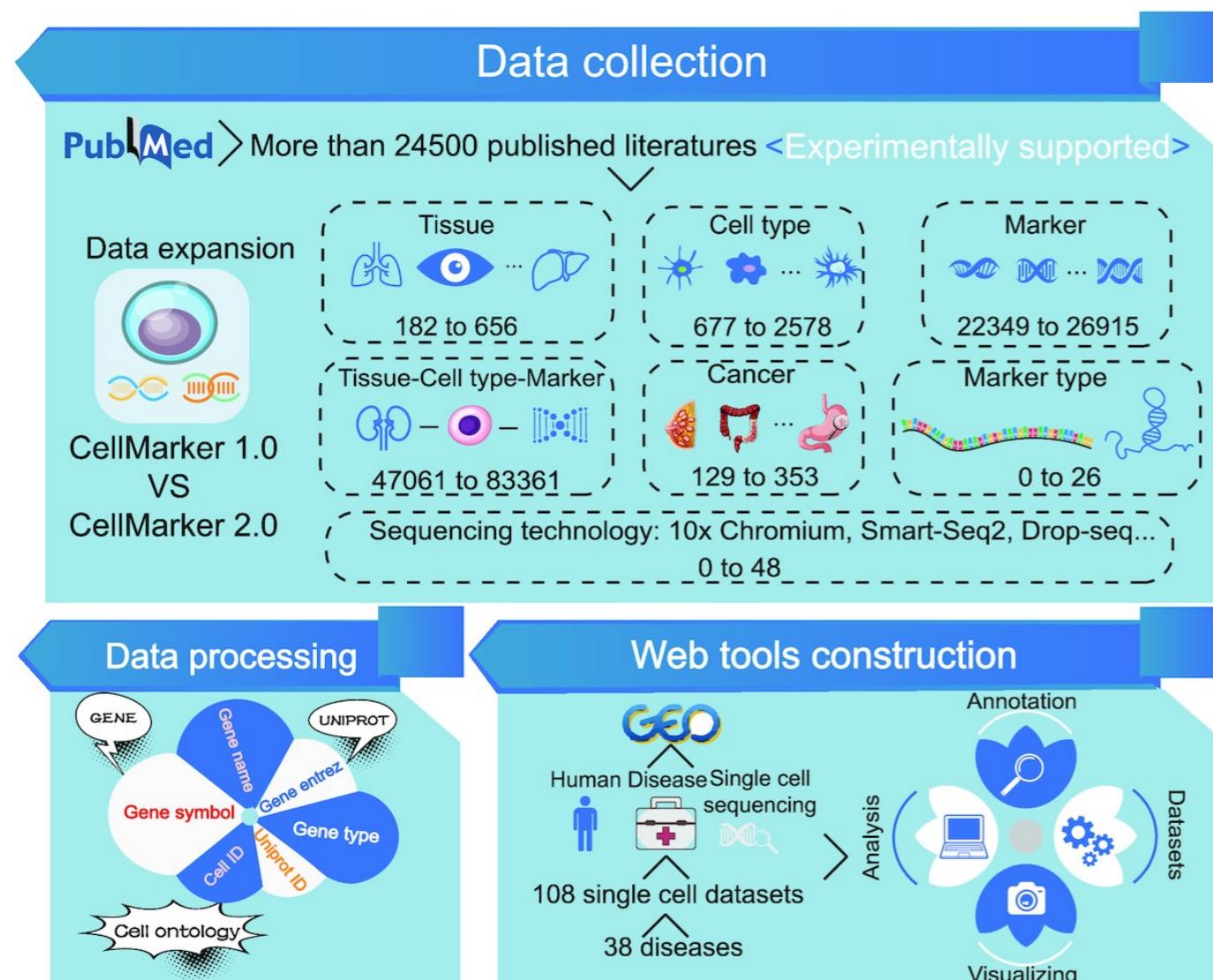
Network of Cancer Genes and Healthy Drivers

- Collection of a comprehensive repertoire of 3355 drivers, 3347 in 122 cancer types and 95 in 12 non-cancer tissues, respectively
- Combining multiple data sources, a set of properties and annotations was computed for all these drivers
- Candidates are defined those relying only on statistical support, whereas canonical drivers were confirmed in sequencing screens as well



Cell Marker

- Clustermole performs cell type inference based on a list of markers retrieved from multiple sources: ARCHS4, Cell Marker, MSigDB, PanglaoDB, Savan, TISSUES and xCell. Among these, Cell Marker is the largest resource.
- Cell Marker is a manually curated collection of experimentally supported markers of various cell types in different tissues of human and mouse

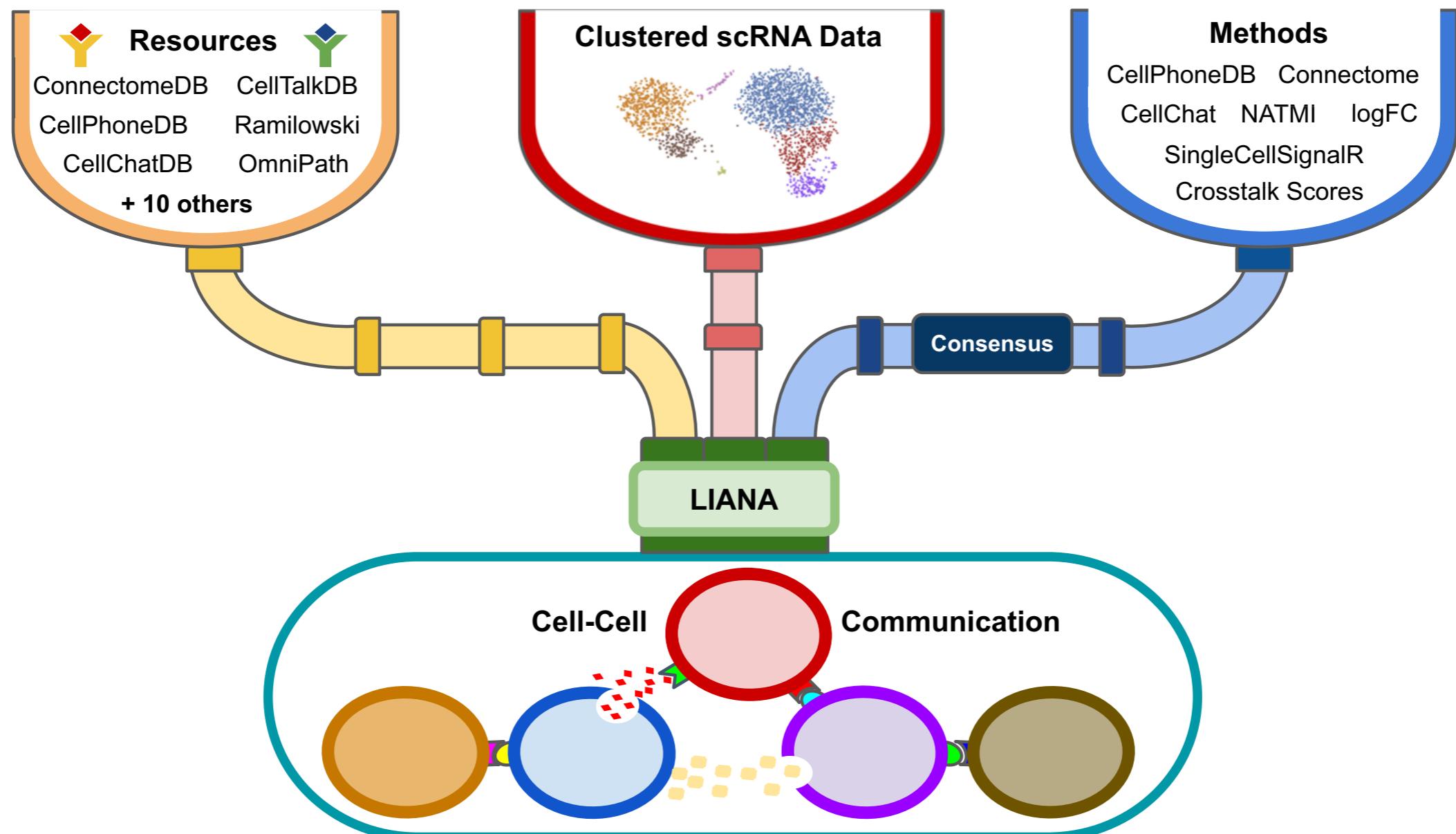


Congxue Hu, Tengyue Li, Yingqi Xu, Xinxin Zhang, Feng Li, Jing Bai, Jing Chen, Wenqi Jiang, Kaiyue Yang, Qi Ou, Xia Li, Peng Wang, Yunpeng Zhang, CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data, *Nucleic Acids Research*, Volume 51, Issue D1, 6 January 2023, Pages D870–D876, <https://doi.org/10.1093/nar/gkac947>

Congxue Hu, et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data, NAR (2023)

LIANA— a Ligand-receptor ANalysis frAmeWork

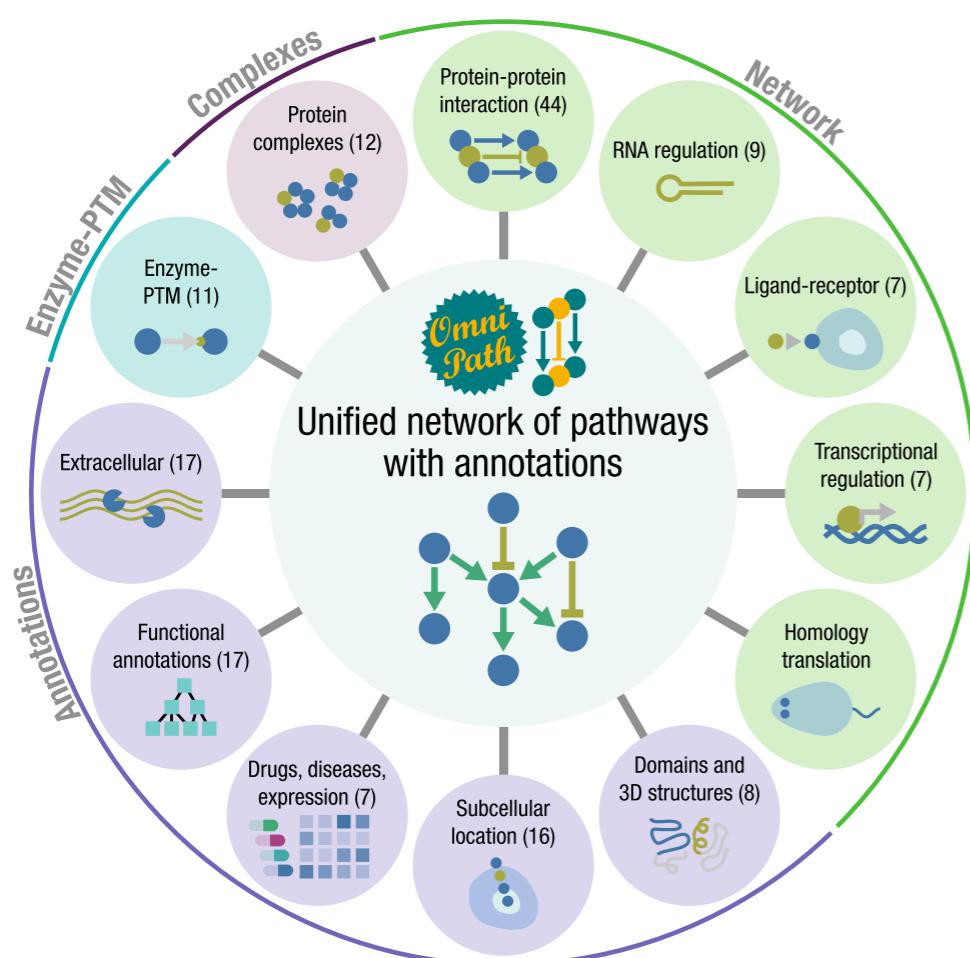
- Is a tool designed for analyzing and interpreting cell-cell communication networks.
- It combines 16 resources and 7 cell-cell communication inference methods, plus their consensus computed by Robust Rank Aggregation approach.



Dimitrov, D., Türei, D., Garrido-Rodríguez, M. et al. Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. Nat Commun (2022)

OmniPath

- OmniPath is an integrated resource of biological knowledge valuable for network analysis and modeling of bulk and single-cell omics data
- Combines data from more than 100 resources and contains protein-protein and gene regulatory interactions, enzyme-PTM relationships, protein complexes, protein annotations and intercellular communication.



Workflow	Components			
Original resources	Network 61 101,962 PPI 29,967 transcriptional 13,486 RNA related			
Database build	Enzyme-PTM 11 39,201 relationships 4,408 proteins			
Database	Complexes 12 22,005 complexes			
Annotations	Annotations 49 4,165,413 records 20,365 proteins			
	python™			
Query interfaces	Web service: http://omnipathdb.org/			
Downstream methods	HTTP (curl, browser, etc) R/Bioconductor (OmnipathR) Cytoscape			
Applications	Python client			
	Third party software (modeling, network analysis, etc)			

Türei D, Valdeolivas A, Gul L, et al. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol Syst Biol* (2021)

LET'S GO

13:30-14:00	Introduction
14:00-15:30	LEVEL 1
15:30-16:00	Disclosure solution level 1
16:00-17:30	LEVEL 2
17:30 -18:00	Conclusion

mariachiara.grieco@unimi.it