



华大生命科学研究院
BGI·Research

Visualization and Computation platforms

Weiwen (Raymond) WANG

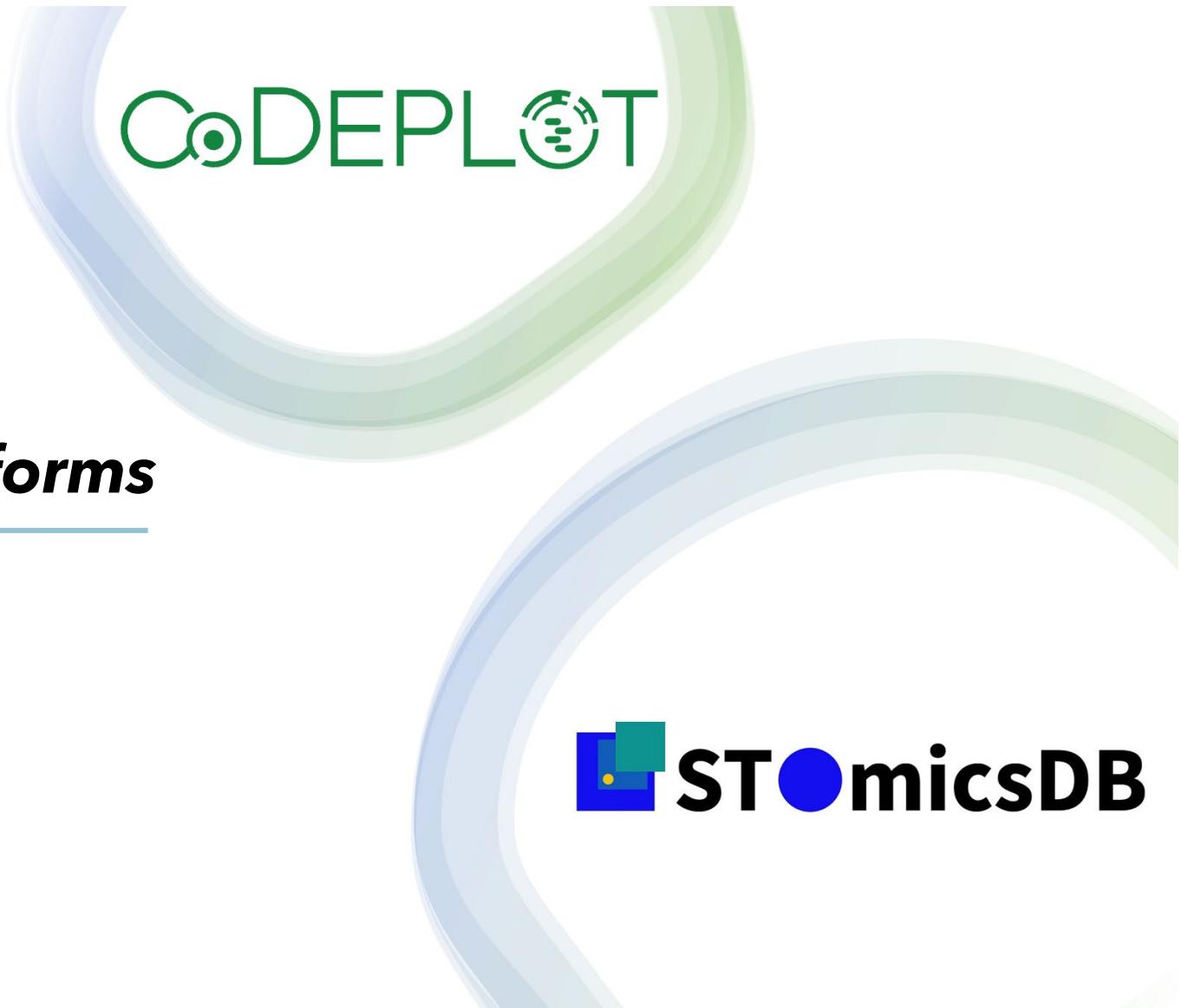
王伟文

2023.9

Overview



Why we build these platforms



CNSA: data archiving, preservation, and sharing

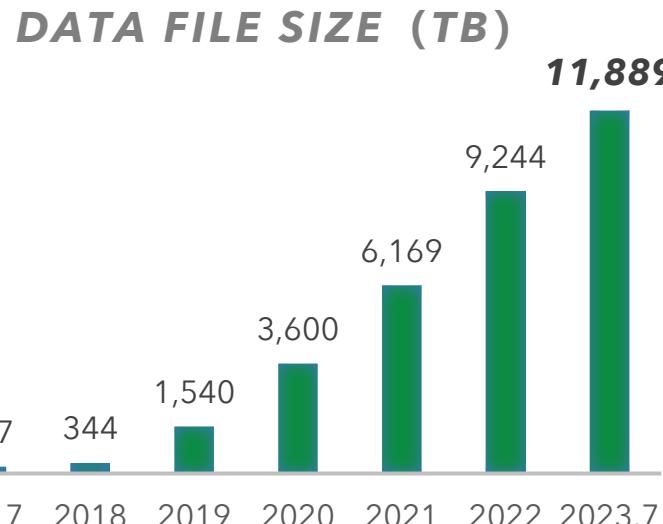
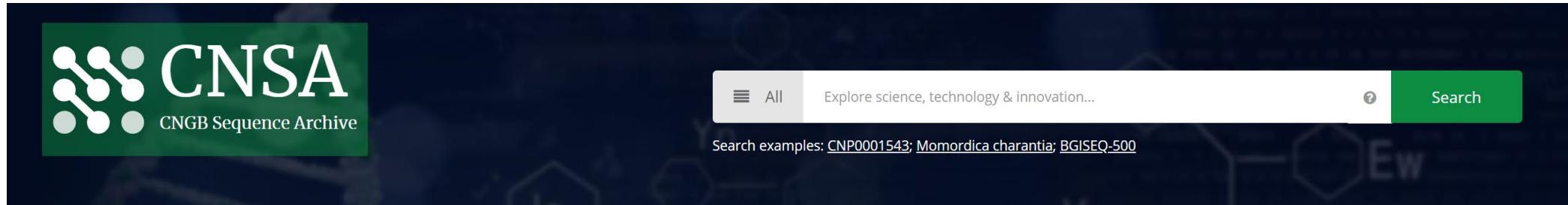
 DATA
11,889TB

 PROJECTS
4,573

 INSTITUTIONS
481

 PUBLICATIONS
1,258

 VISITS
> 10 Millions



- ***Challenge for data reuse and sharing***

1. Difficulty in finding data
2. Difficulty in downloading large datasets
3. Difficulty in analysing large-scale data
4. Difficulty in identifying high-quality data

- ***What are we doing?***

- Specialized Databases (Manual curation, high-quality data)
- Computation platform (Online analysis, high performance)

Overview



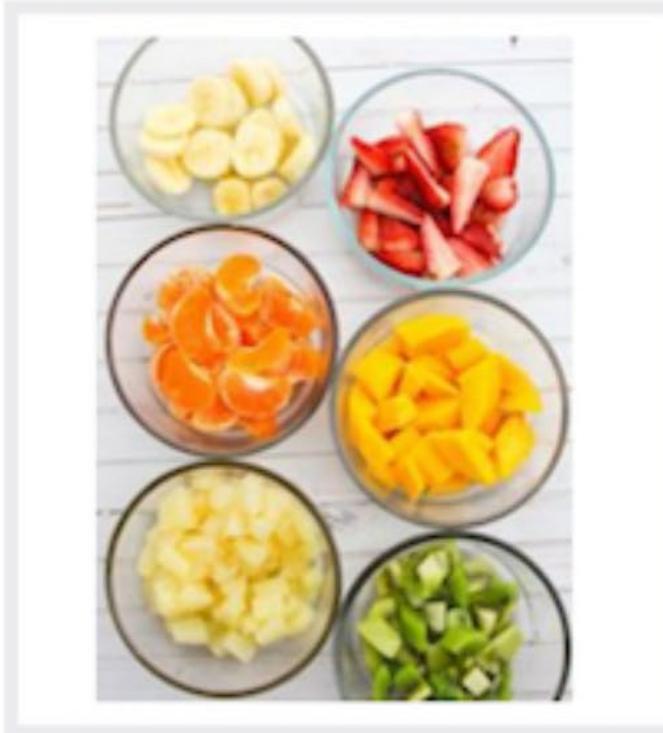
1. Background

► **What is spatial transcriptomics**

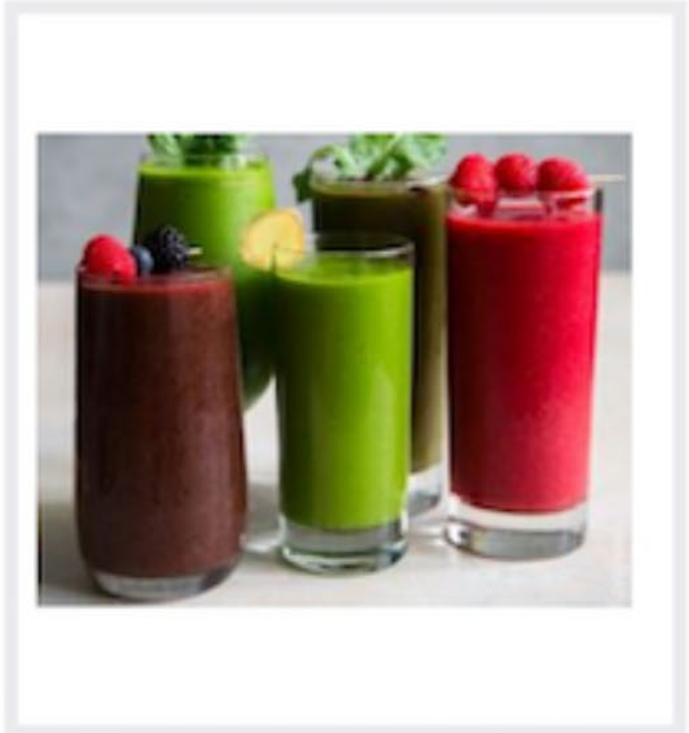
Spatial



Single cell



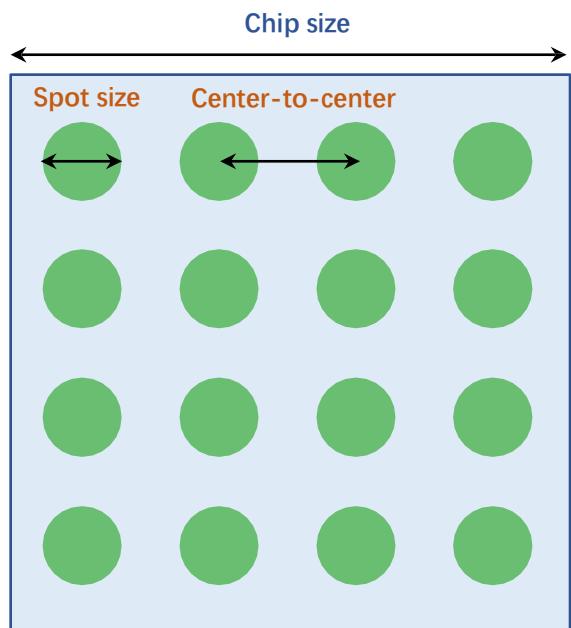
Bulk



1. Background

Comparison

Stereo-seq: smallest spot size, largest field of view

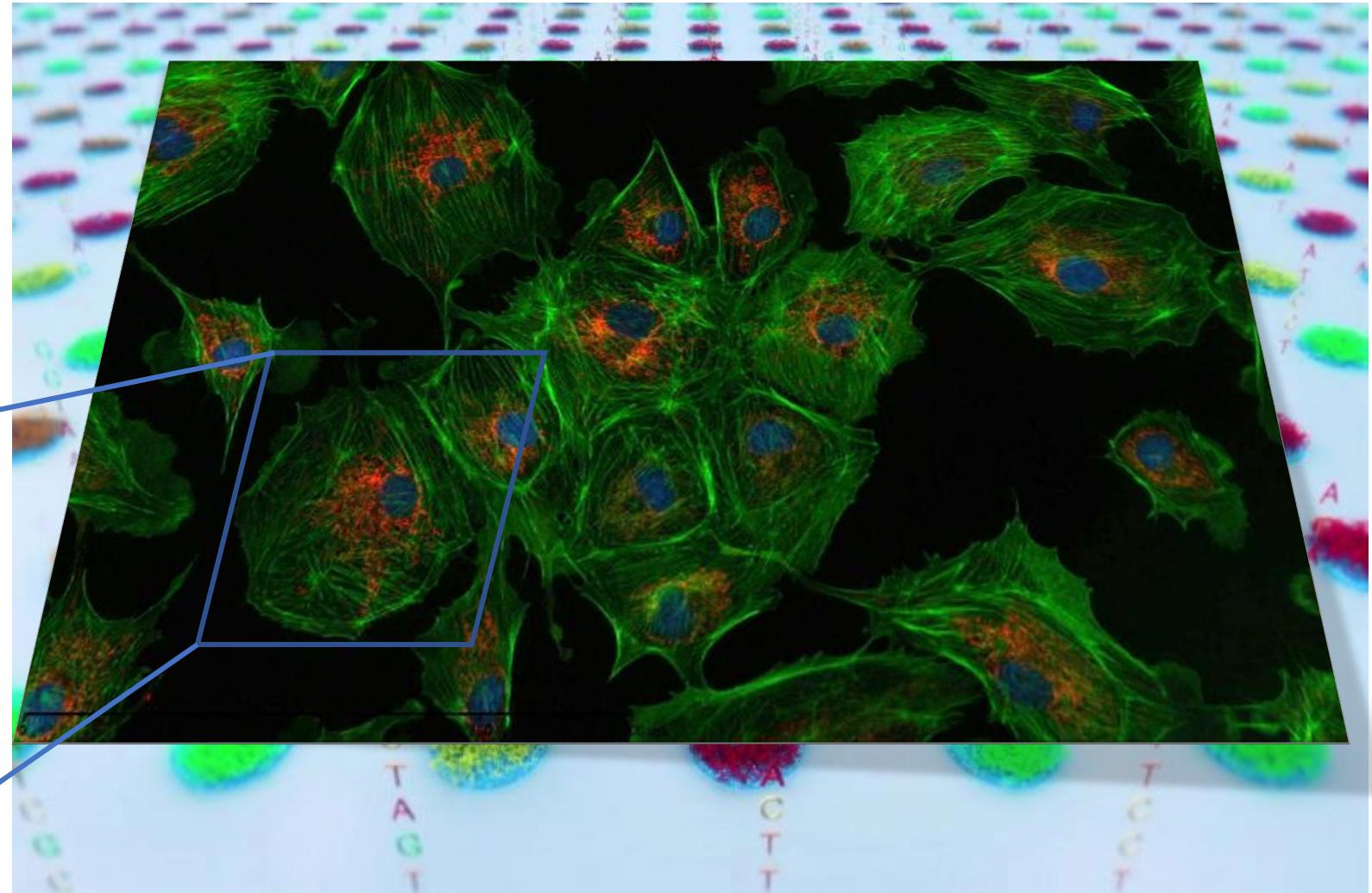
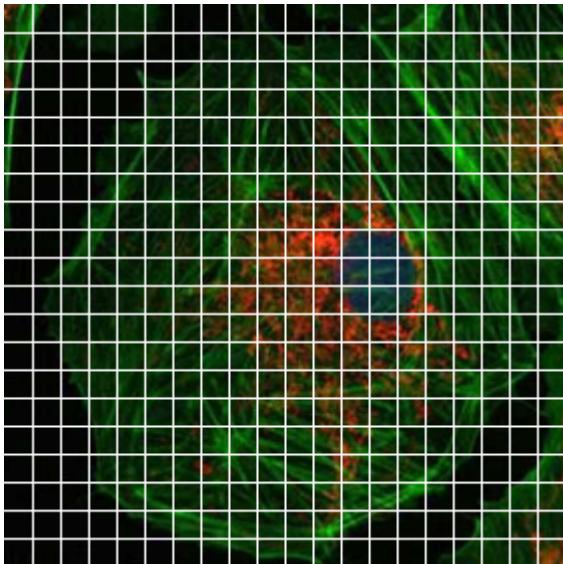


	DBiT-seq	Slide-seq	HDST	10x	Stereo-seq
Spot size (μm)	10	10	2	55	0.5
Center-to-center (μm)	20	10	2	100	0.7
Field of view (mm)	1.0 x 1.0	Φ 3.0	5.7 x 2.4	6.5 x 6.5	120 x 120

1. Background

► Resolution

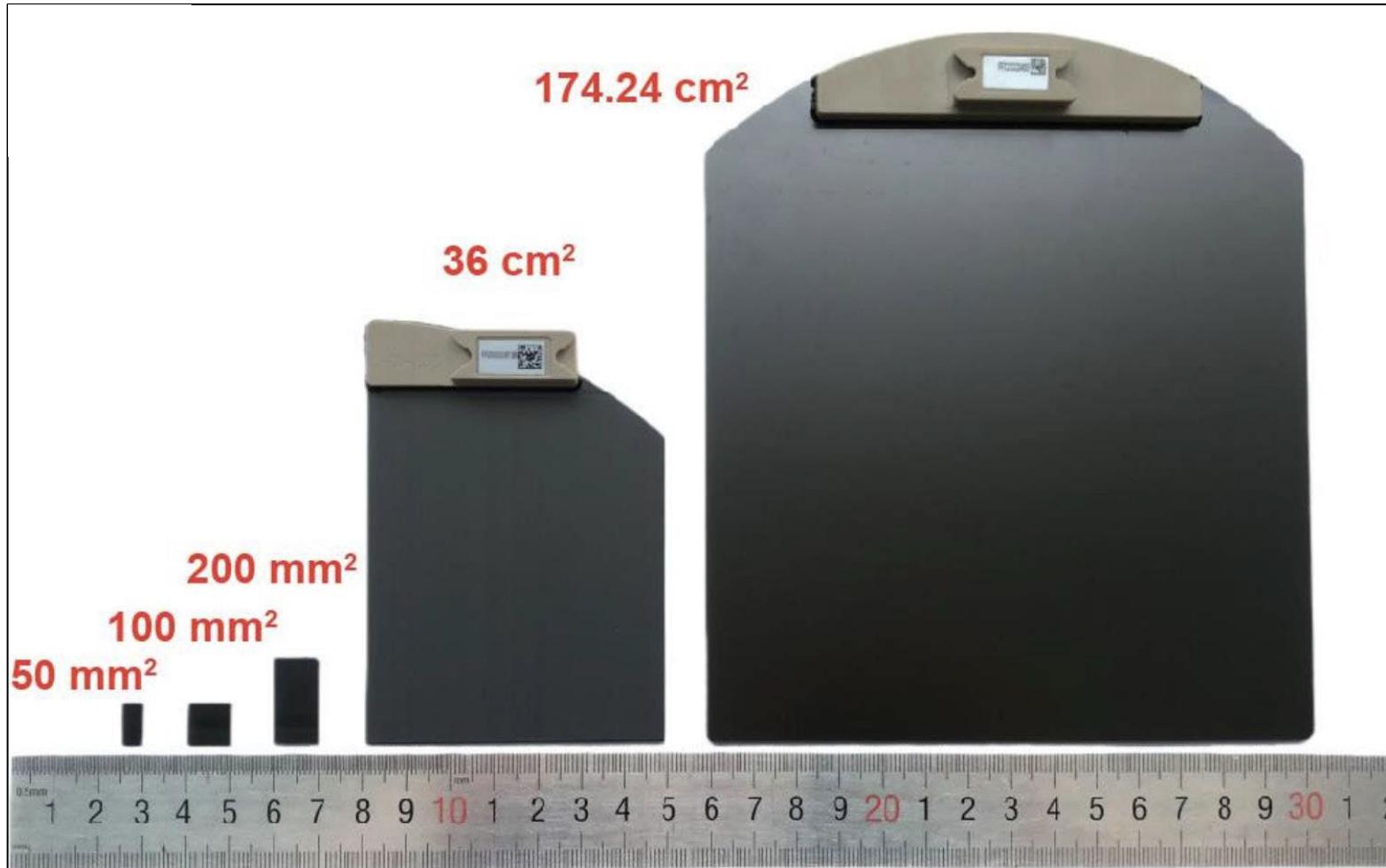
- **Cell size:** 3-30 μm
- **Stereo-seq :**
0.5 μm (Subcellular resolution)



1. Background

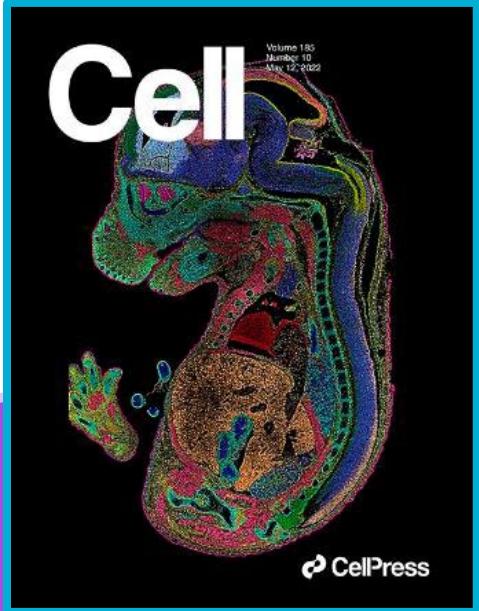
Field of view

- **Stereo-seq chips:** ranging from 50 mm^2 to 174.24 cm^2



1. Background

Publication examples



2022.5.4

Mouse Organogenesis
Spatiotemporal
Transcriptomic Atlas



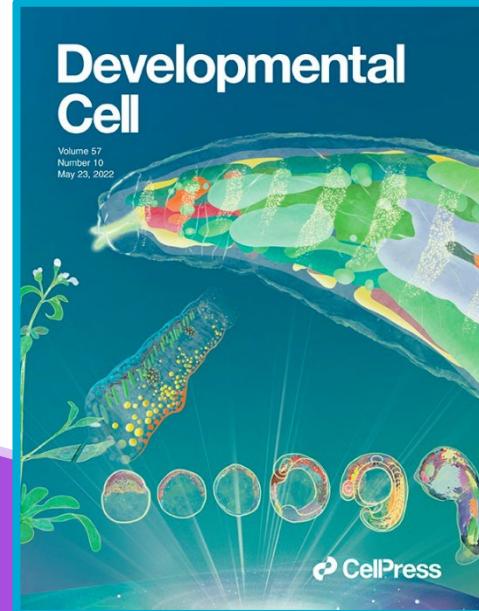
2022.9.2

The cellular and molecular features of the **axolotl** telencephalon during development and injury-induced regeneration.



2022.9.21

A cellular hierarchy in **melanoma** uncouples growth and metastasis



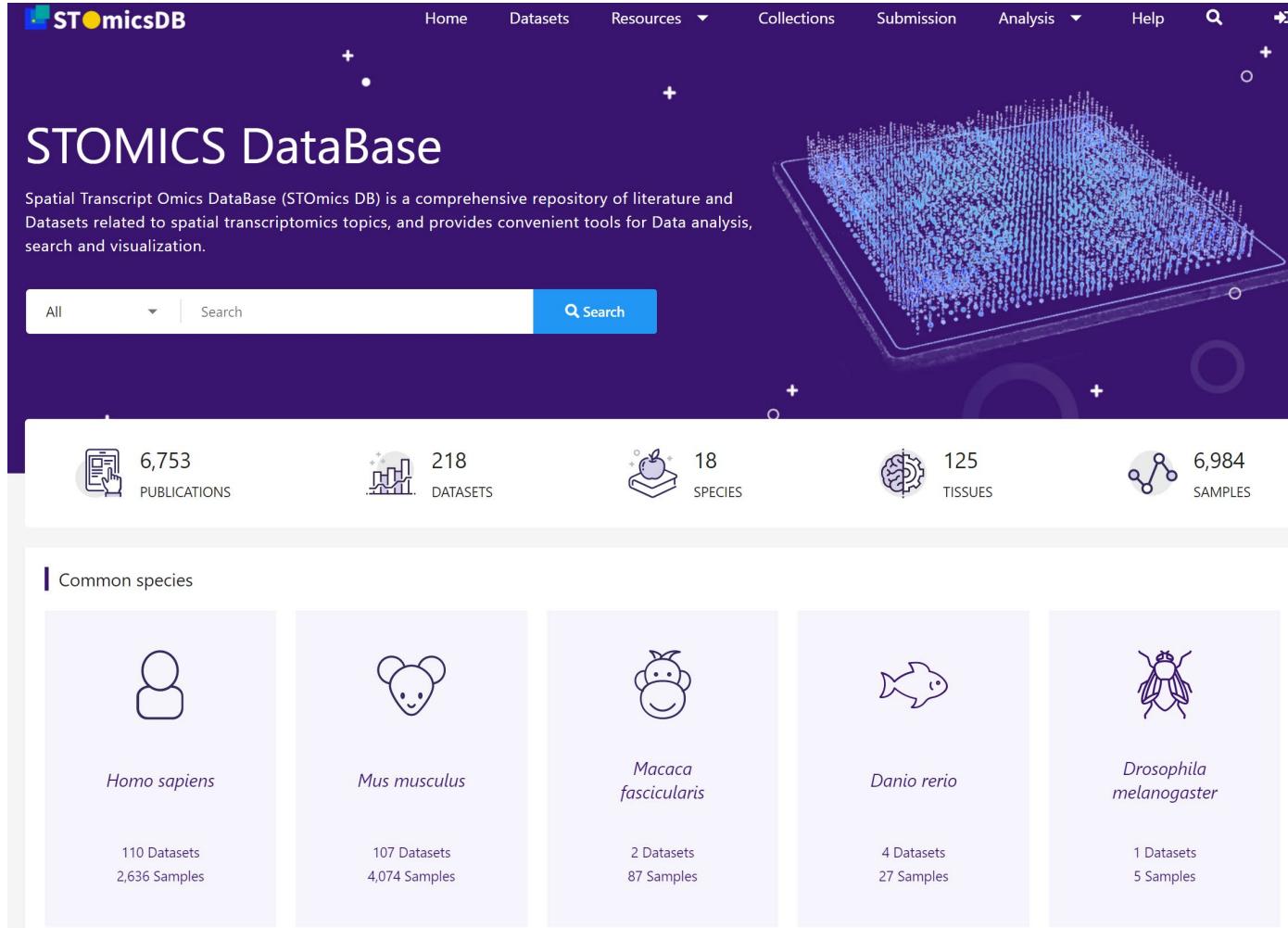
2022.5.5

Zebrafish/Drosophila
Embryogenesis
Spatiotemporal
Transcriptomic Atlas

2. STomicsDB



Spatial Transcriptomics DataBase

A screenshot of the STomicsDB website. The header includes a logo, navigation links (Home, Datasets, Resources, Collections, Submission, Analysis, Help), a search bar, and a 3D visualization of a tissue sample. Below the header, a main heading "STOMICS DataBase" is displayed, followed by a brief description of the database as a comprehensive repository of literature and datasets related to spatial transcriptomics. A search bar and a "Search" button are present. The main content area shows statistics for publications (6,753), datasets (218), species (18), tissues (125), and samples (6,984). A section titled "Common species" lists five species with their respective icons and dataset counts:

- Homo sapiens: 110 Datasets, 2,636 Samples
- Mus musculus: 107 Datasets, 4,074 Samples
- Macaca fascicularis: 2 Datasets, 87 Samples
- Danio rerio: 4 Datasets, 27 Samples
- Drosophila melanogaster: 1 Datasets, 5 Samples

- Curated 200+ datasets
- Spatial transcriptomic data exploration and visualization
- Customized collections/databases

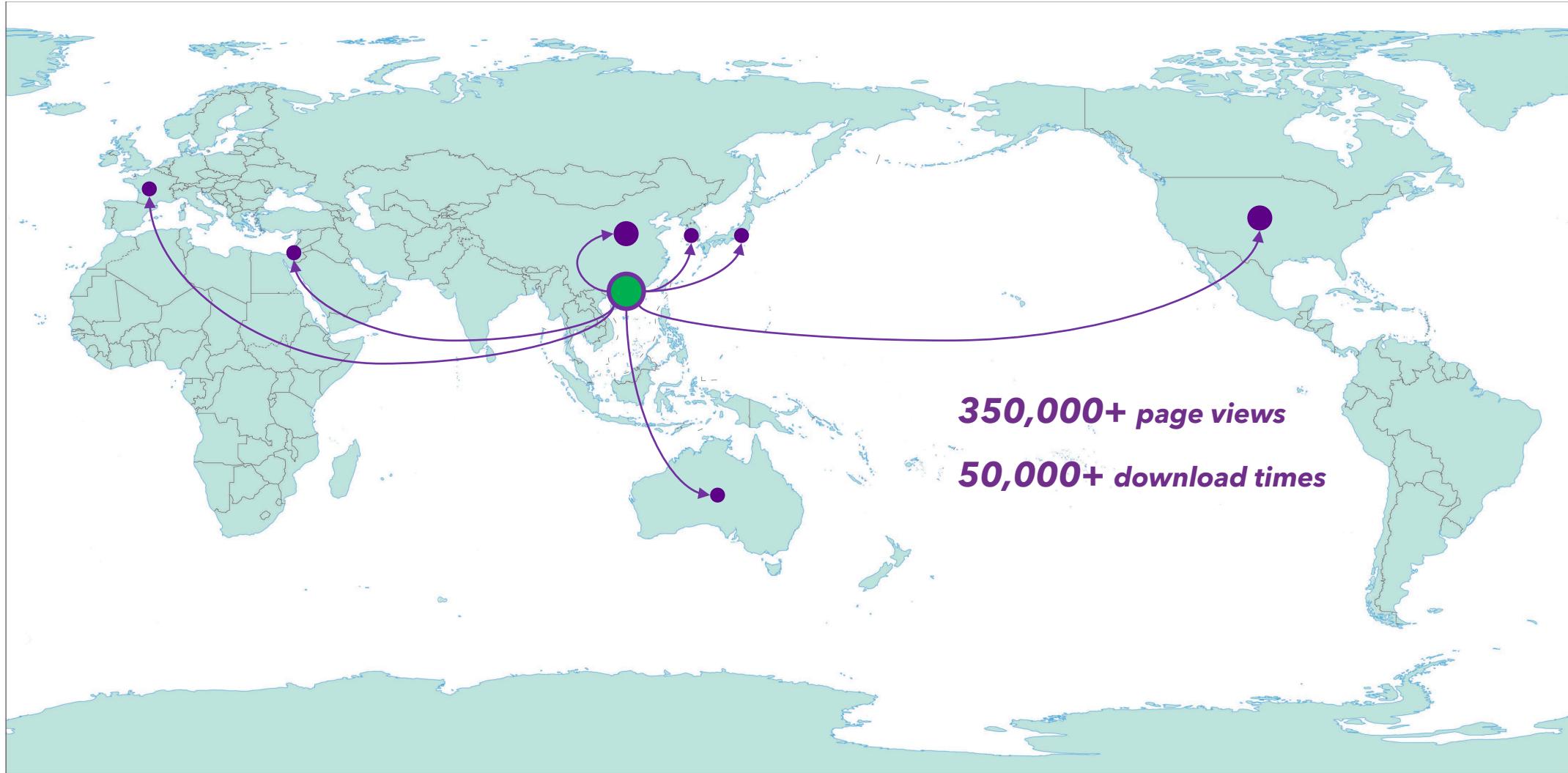
db.cngb.org/stomics



2. STOmicsDB



Overview



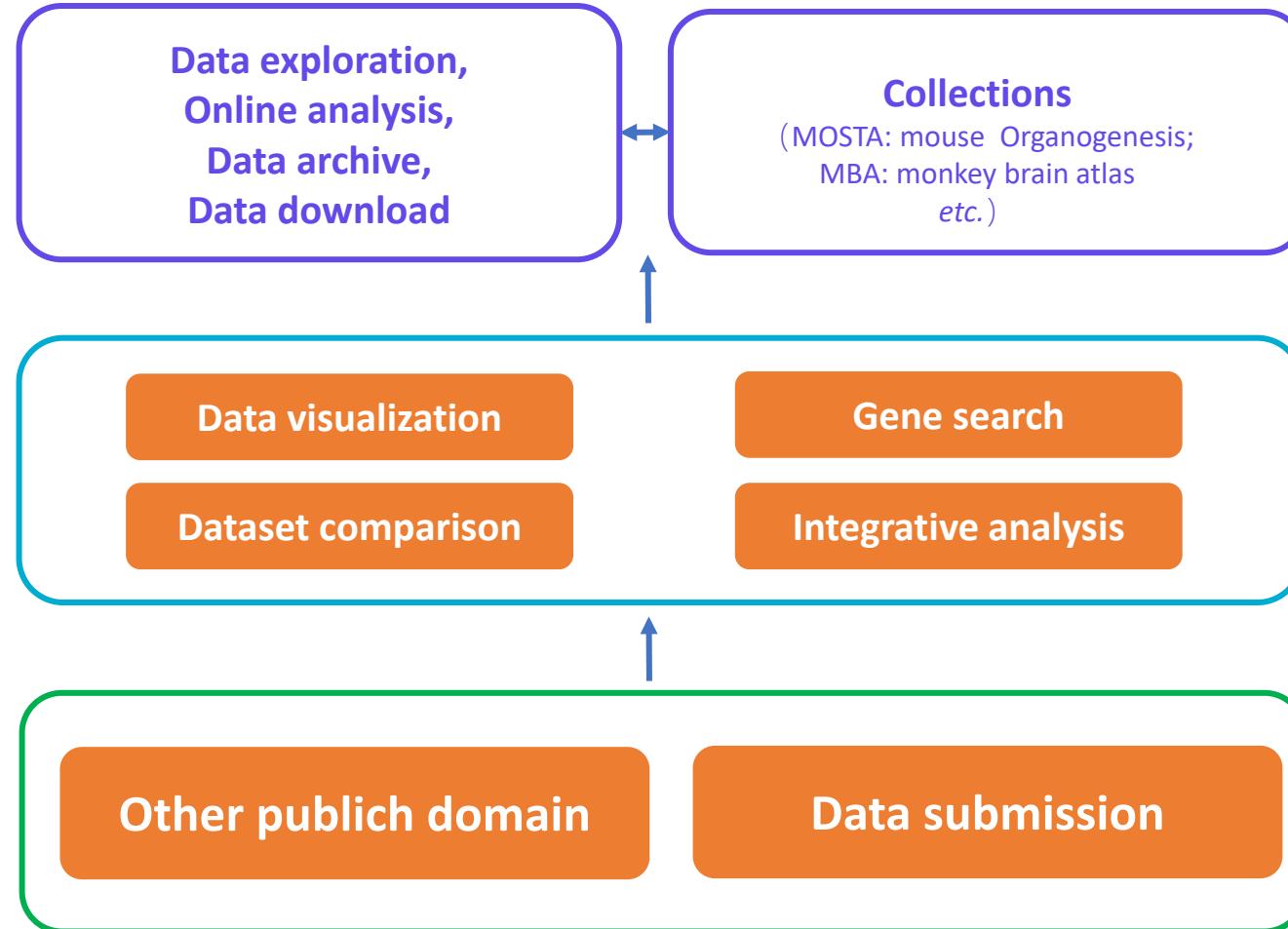
2. STOmicsDB

Structure

Application

Tool

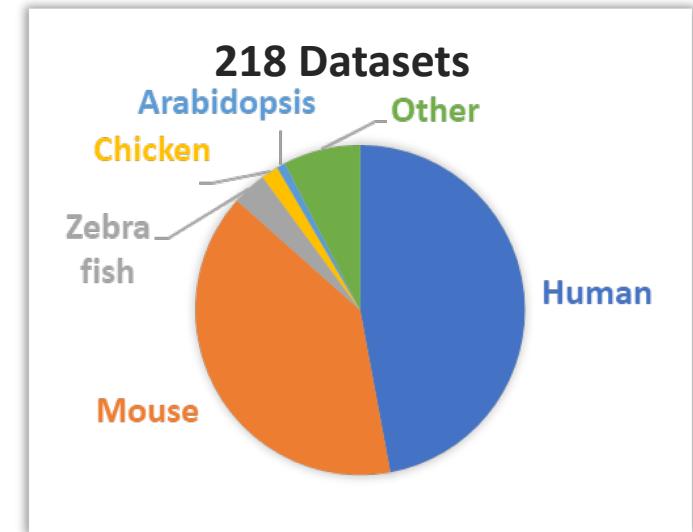
Data



2. STOmicsDB

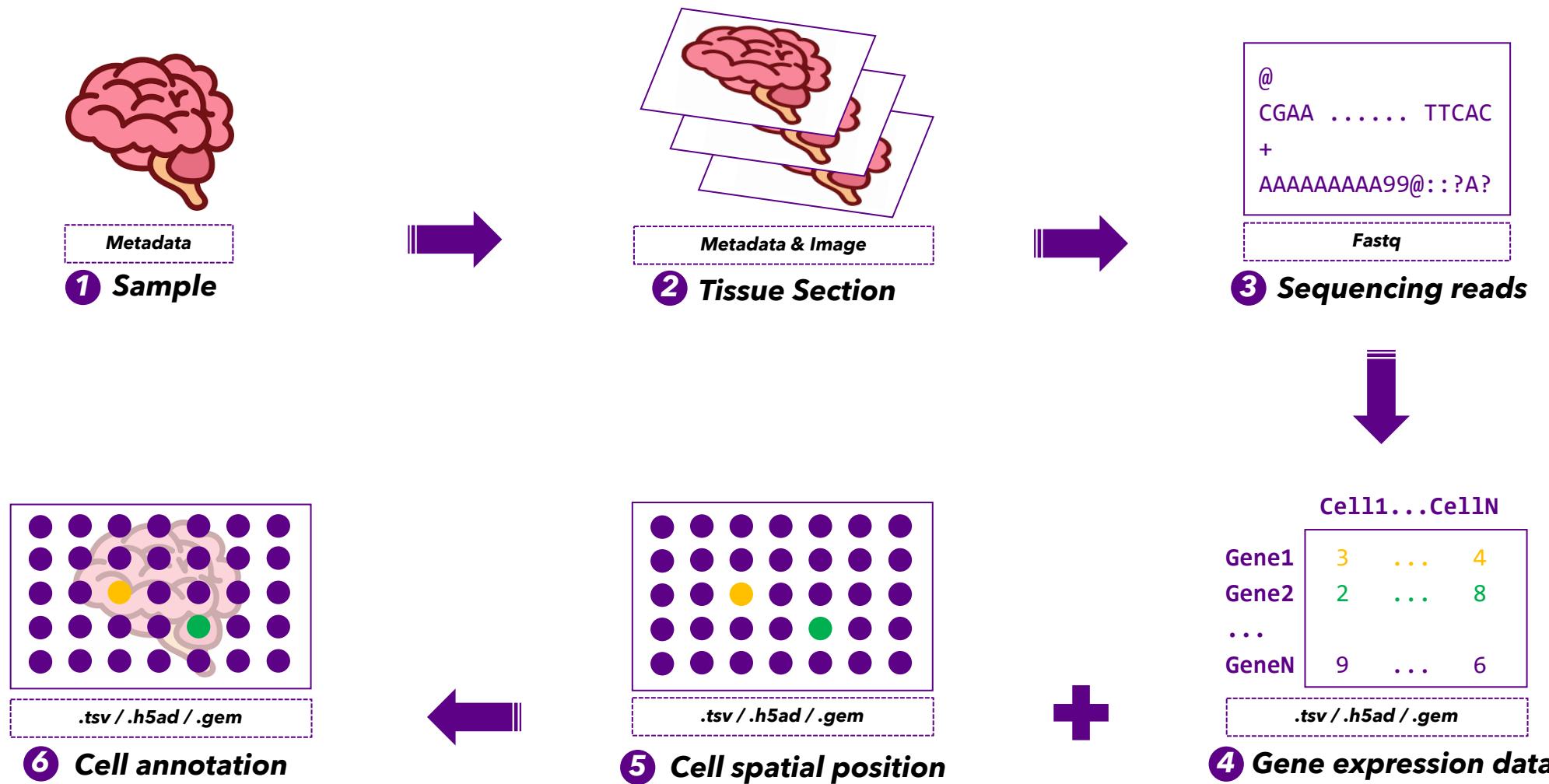
Data

- **Collected (218 datasets so far)**
 - from NCBI, EBI, DDBJ, etc.
 - from papers
 - from submission
- **Curated**
 - display general information of each dataset
(summary, overall design, species, cell types,
development stage, sample number, section number, etc)



2. STomicsDB

Data archiving system



2. STomicsDB



Data archiving system

The screenshot shows the 'STomics Sub' section of the STomics DB website. It displays a table of four submissions:

Submission ID	Project	Data access manner	Status	Release date	Update date
sts0000025	STT0000013: Large field of view-spatially resolved transcript...	Public	Processing	2022-01-31	2022-01-06
sts0000024	STT0000012: Large field of view-spatially resolved transcript...	Public	Processing	2022-01-31	2022-01-06
sts0000023	STT0000011: Spatiotemporal transcriptomic atlas of mouse ...	Public	Processing	2022-01-31	2022-01-06
sts0000020	STT0000010: test	Public	Processing	2022-12-10	2021-12-30

Showing 1 to 4 of 4 result(s).

The screenshot shows a detailed view of a project entry. The project summary is "Large field of view-spatially resolved transcriptomics at nanoscale resolution". The project ID is STT0000001. The project details include:

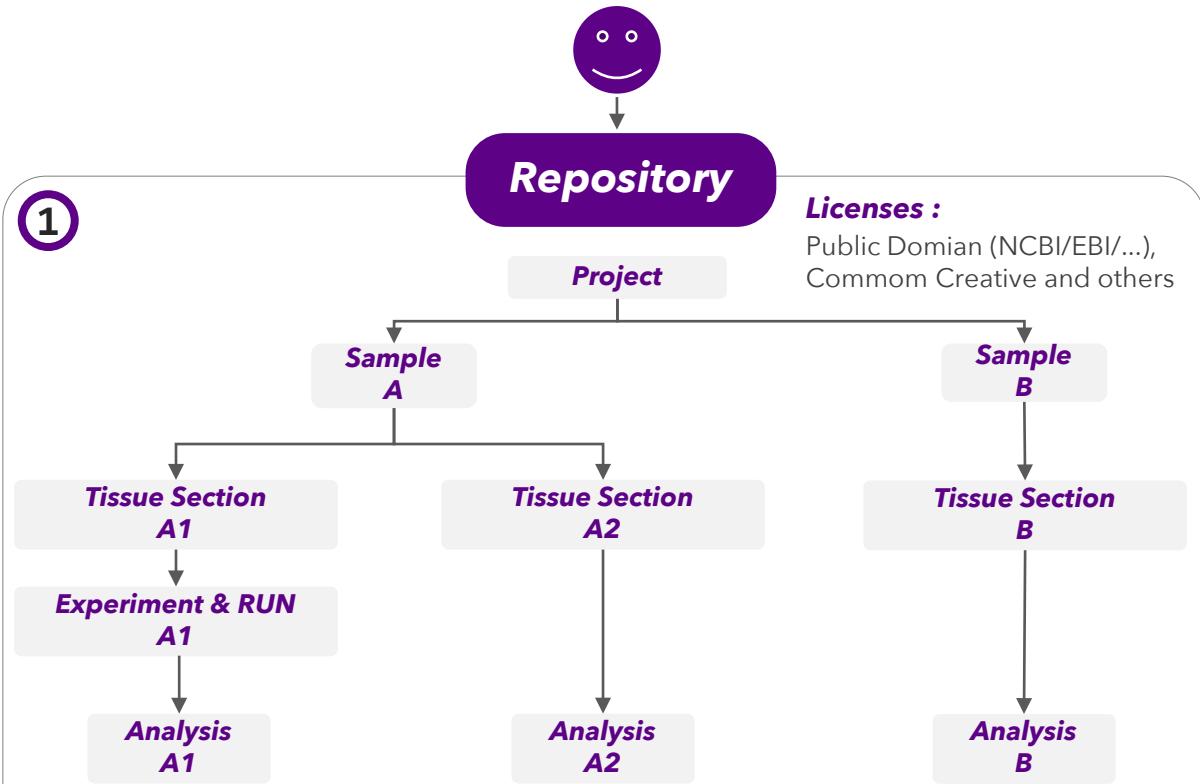
- Organism: Homo sapiens
- Data type: STomics, Raw sequence reads
- Sample scope: Multisample
- Summary: High-throughput profiling (DNB) patterned array chips and *in situ* applied STomics-seq to the adult mouse of tissues and organisms.
- Contributors: Zheng C, Hu Y
- Publication: Niu Y, Sun N, Li C, Lei Y et al.
- Submitter: 高圆圆 (Liang Wu), BGI-Shenzhen
- DOI: 10.26036/6FTT00000001
- Release date: 2018-03-29
- Updated: 2018-03-30
- Reference project: CNP0001543
- Statistics:
 - Sample: 8
 - Tissue Section: 20
 - Experiment: 20
 - Run: 20
 - Dataset: 20
- Data size: 100GB

A modal window titled "Stereo-seq" is open, showing a circular heatmap visualization of gene expression across tissue sections. Below the heatmap is a table of spatial gene expression data:

geneID	x	y	MIDCount
Cr1l	105228	101186	1
Cr1l	113258	105281	1
Cr1l	113258	105281	1
Cr1l	113586	106877	1
Cd46	109348	106877	1
Cd46	109348	106877	1
Gb22250	112910	100030	1
Gb22250	112910	100030	1
Gb34	105952	100030	1
Gb34	112123	108884	1
Gb34	112568	102785	1
Gb34	113048	109062	2

Buttons for "Download" and "Close" are visible at the bottom of the modal.

Submitter



Datasets & Collections

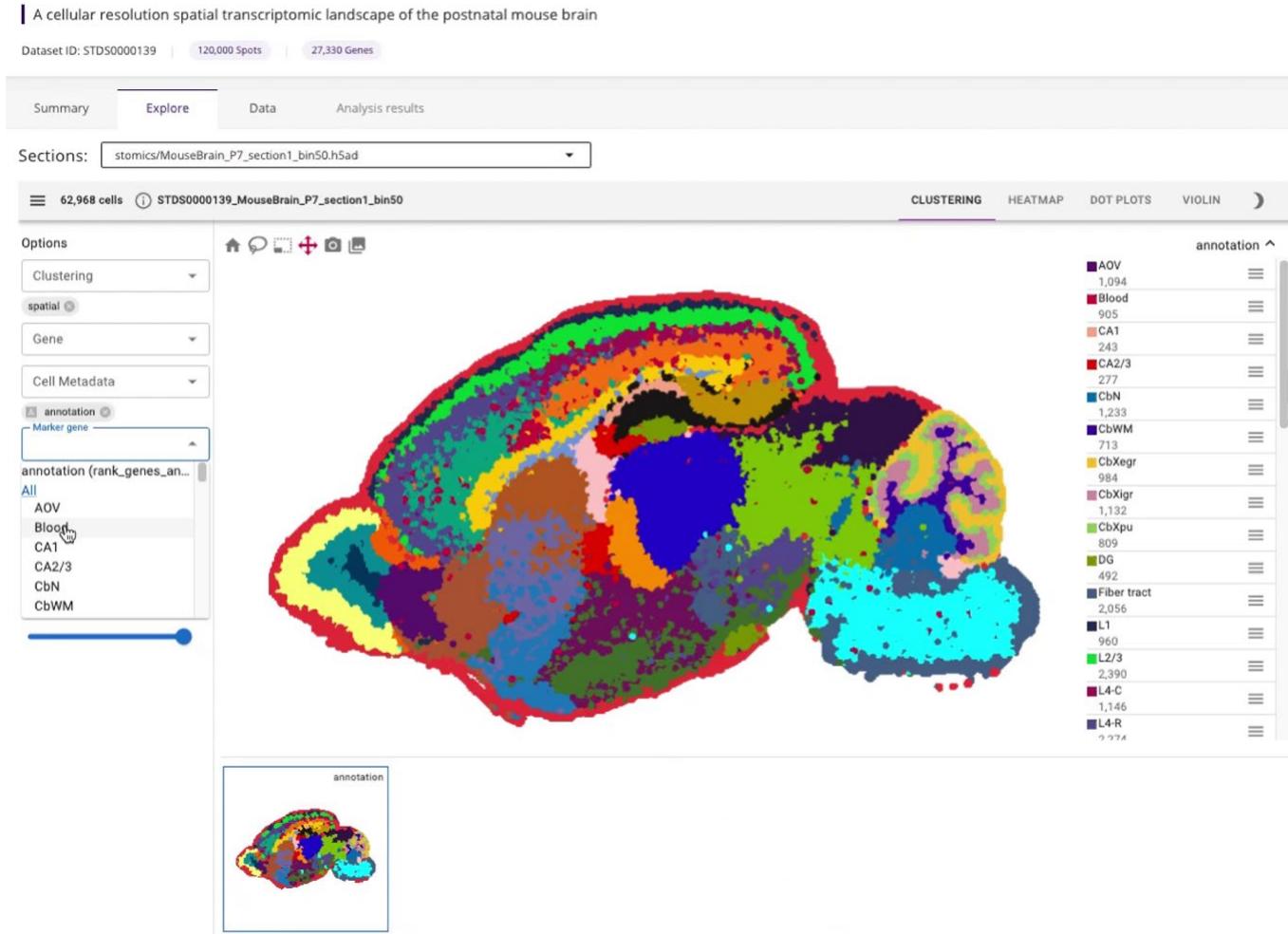
Visualization

Discovery

Knowledge

2. STomicsDB

Visualization



1

Independent public datasets

- #1 Researchers have different filter criteria and nomenclature for data quality.
- #2 Different articles have different levels of cell annotation.

2

Standardized Analysis

- #1 Normalize and logarithmize the gene expression data.
- #2 Conduct principal component analysis (PCA).
- #3 Calculate the neighborhood map with PCA results

3

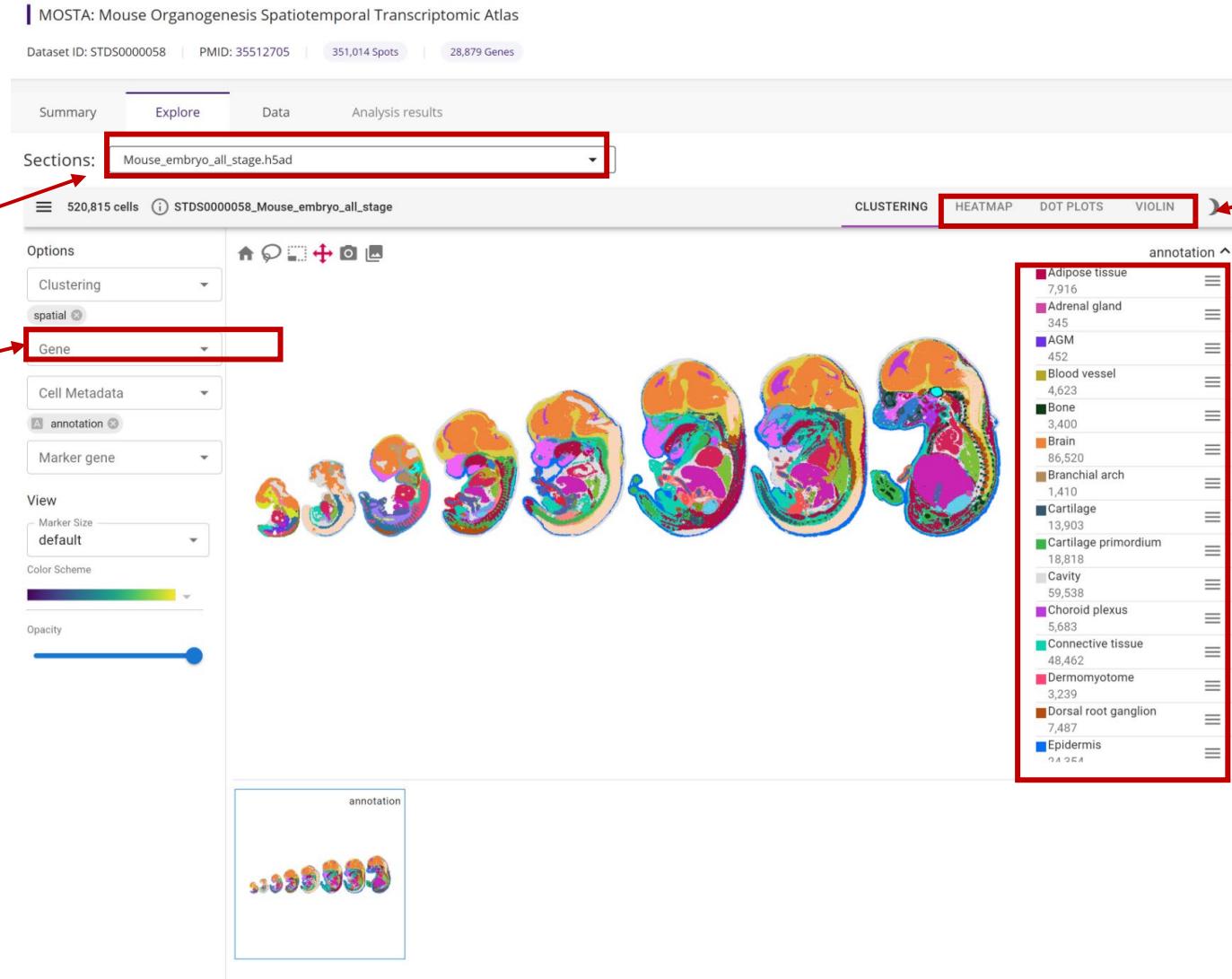
Consistently comparable datasets

- #1 Standardized data quality control and cell grouping.
- #2 Standardized named subcell type annotations.

2. Introduction

Visualization

<https://db.cngb.org/stomics/>, then search ‘MOSTA’ in the search bar



Select different sample

Select the gene of interest

Switch to the comparison of the gene expression module

Cell annotation

2. STomicsDB

Gene search

6,000+
curated sample slides

*Species Tissue *Gene

All species

Input a gene

Search

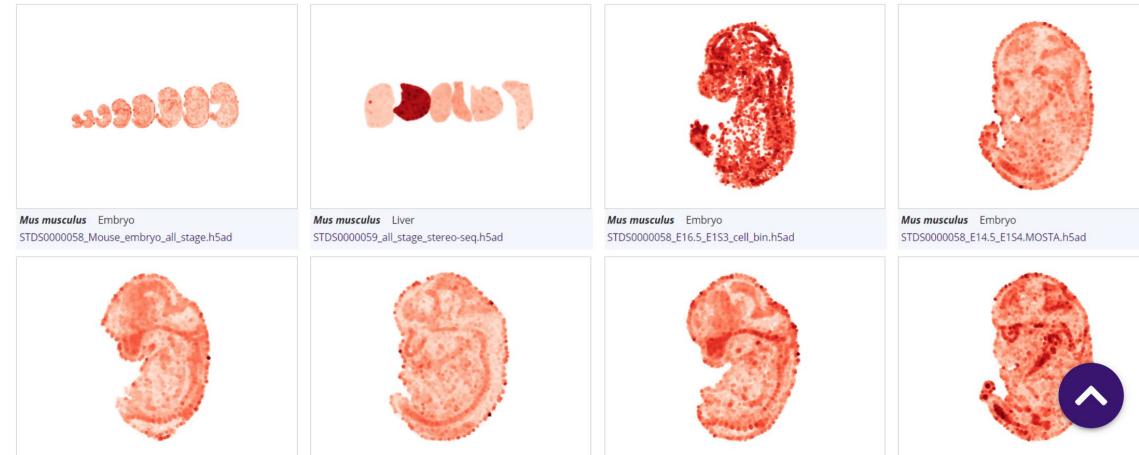
Q Select species & gene

Show the spatial map of corresponding gene among different sections of different dataset.

ⓘ You can observe that

- multiple perspectives
- different development stages
- different cancer types
- different tissues, normal and diseased

RUN DEMO >>



This function shows the spatial expression of a specific gene.

Allows users to search gene of interest among all datasets, and they can select the species or tissue to narrow down the results.

Help users to efficiently find genes with the spatial feature. We sort the gene by the spatial pattern. Users could easily find the dataset which contains gene of interest with specific spatial expression.

2. Introduction

Gene search

https://db.cngb.org/stomics/analysis/gene_search

Select species

Select tissue

Type the gene

Red arrows point from the text labels to the corresponding input fields in the search interface.

*Species

All species

Tissue

*Gene

Input a gene

Search

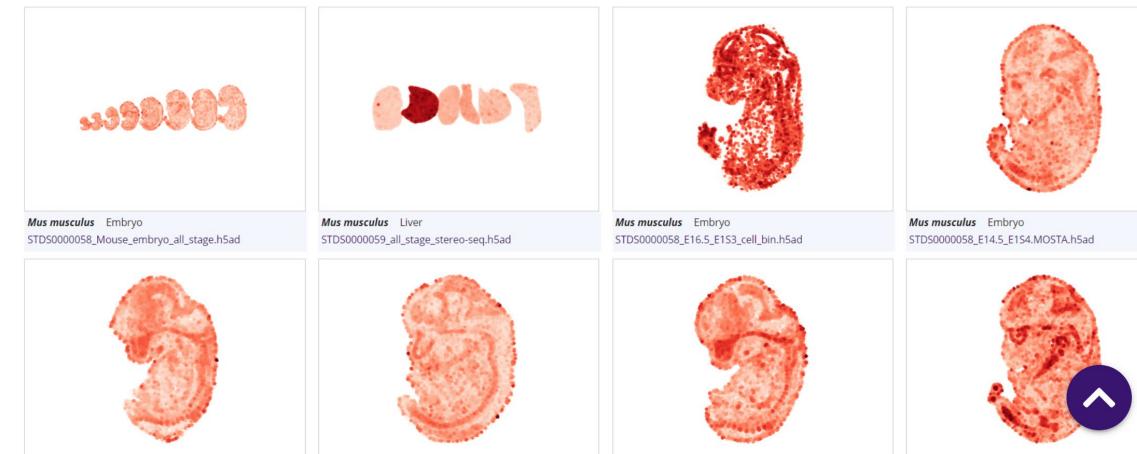
Q Select species & gene

Show the spatial map of corresponding gene among different sections of different dataset.

ⓘ You can observe that

- multiple perspectives
- different development stages
- different cancer types
- different tissues, normal and diseased

RUN DEMO >>



2. Introduction

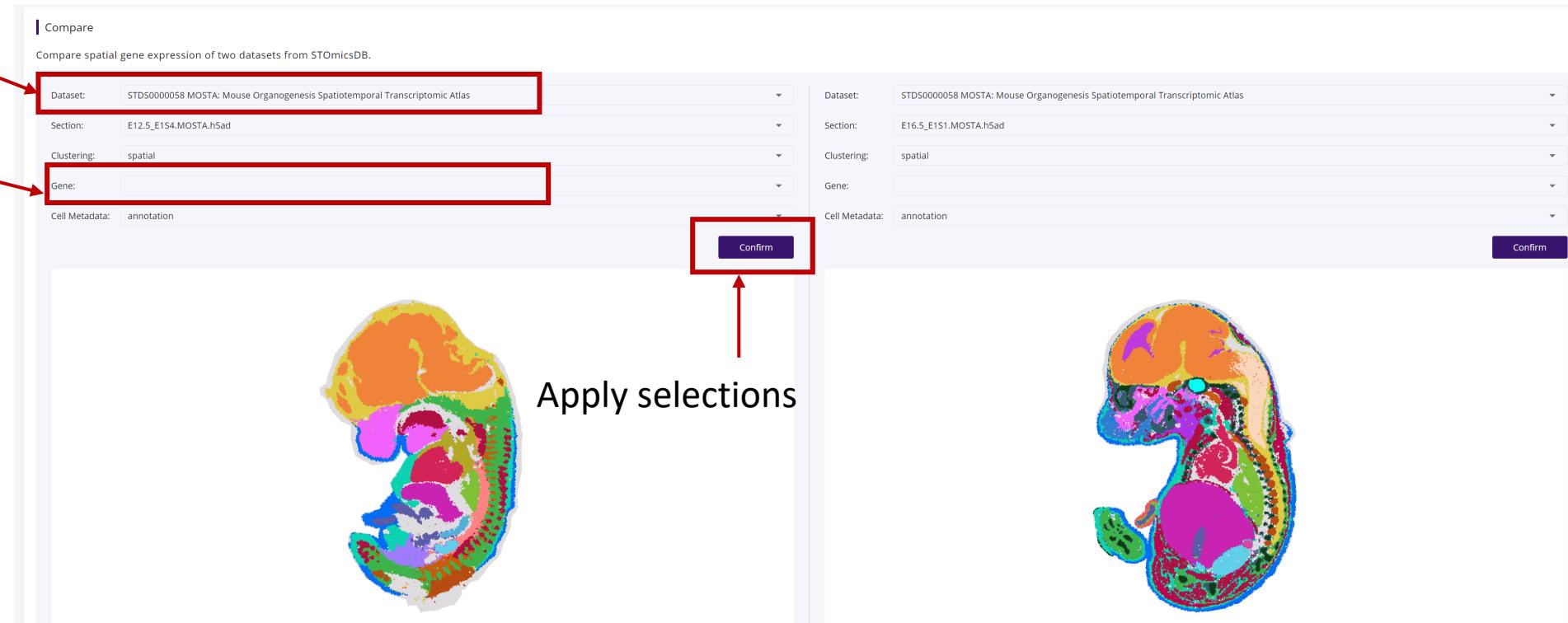
Dataset comparison

<https://db.cngb.org/stomics/analysis/compare>

Select dataset

Select gene

Apply selections



Compare
Compare spatial gene expression of two datasets from STomicsDB.

Dataset: STDS0000058 MOSTA: Mouse Organogenesis Spatiotemporal Transcriptomic Atlas

Section: E12.5_E154.MOSTA.h5ad

Clustering: spatial

Gene:

Cell Metadata: annotation

Confirm

Dataset: STDS0000058 MOSTA: Mouse Organogenesis Spatiotemporal Transcriptomic Atlas

Section: E16.5_E151.MOSTA.h5ad

Clustering: spatial

Gene:

Cell Metadata: annotation

Confirm

2. Introduction

Collections

<https://db.cngb.org/stomics/collections>

- Collaborated with 6 research groups so far
- Welcome collaboration



Collection entry

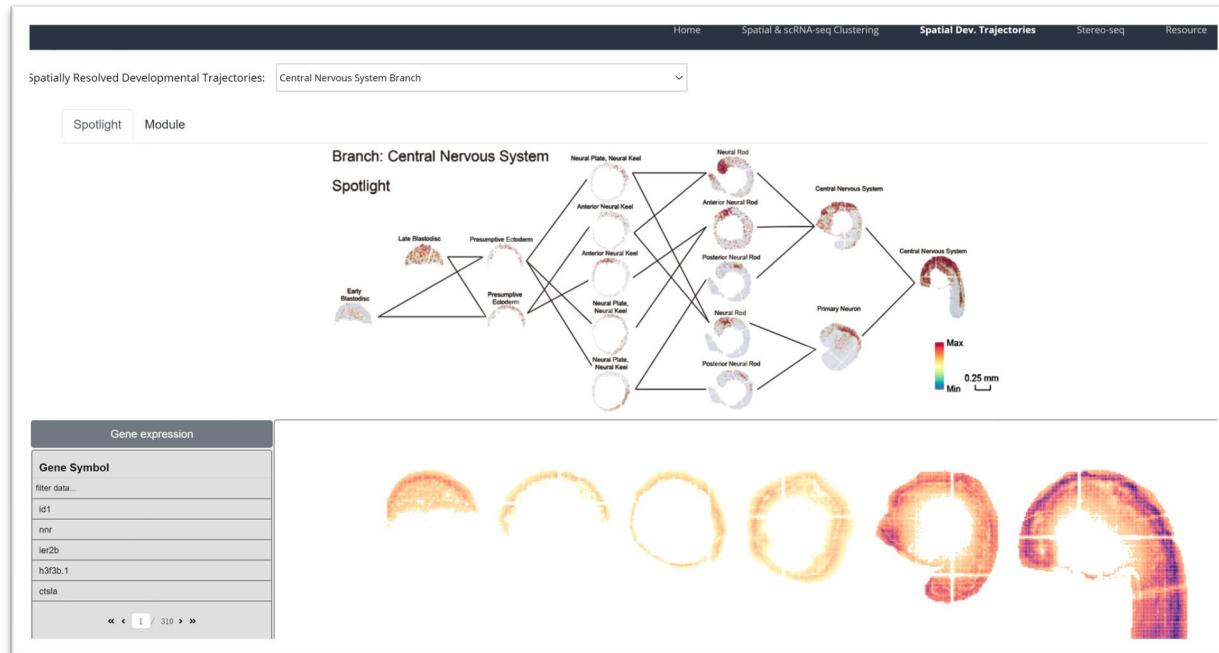
The screenshot shows the STomicsDB website interface. At the top, there is a navigation bar with links: Home, Resources, Datasets, Collections (which is highlighted with a red box and an arrow pointing to it from the text 'Collection entry'), Submission, Analysis, Help, and a search icon. Below the navigation bar, there are six collection entries arranged in a grid:

- MOSTA**: Mouse Organogenesis Spatiotemporal Transcriptomic Atlas. 300,000+ bins.
- ZESTA**: Zebrafish Embryogenesis Spatiotemporal Transcriptomic Atlas. 152,977 bin15, 91 Section, 86,307 Cells.
- Flysta3D**: High-resolution 3D spatiotemporal transcriptomic maps of developing Drosophila embryos and larvae. 90 Section, 5 Samples.
- ACSTA**: Arabidopsis Cell-type-specific Spatiotemporal Transcriptomic Atlas. 26 Samples, 13,950 Cells.
- MBA**: Macaque Brain Atlas. 358,237 Cells.
- ARTISTA**: Axolotl Regenerative Telencephalon Interpretation via Spatiotemporal Transcriptomic Atlas. 36 Samples.

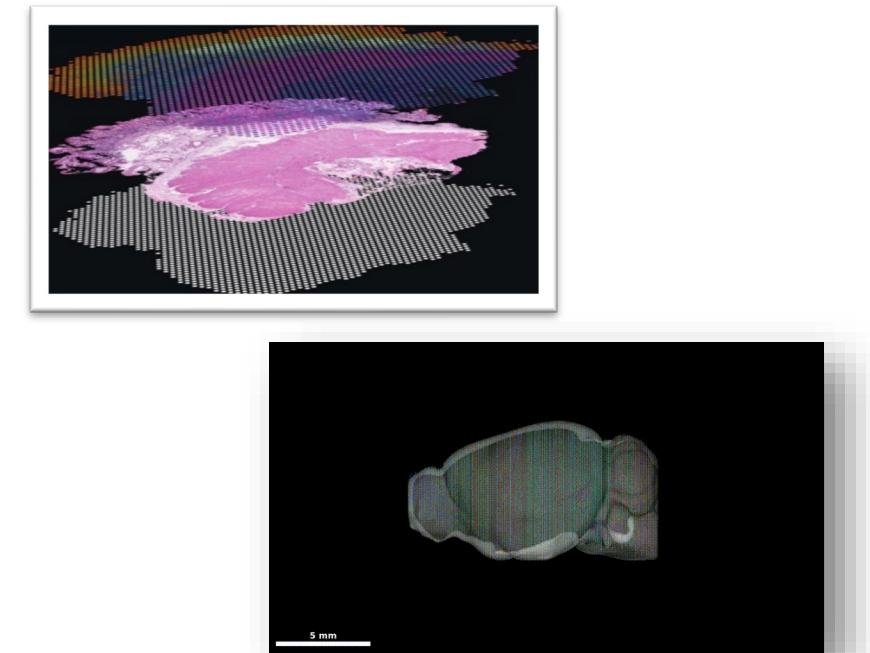
2. STOmicsDB

Collections

- Customizable visualization



Development Trajectory of Zebrafish



3D model of *Drosophila*



Codeplot: a platform for code-free analyses

The screenshot shows the CoDEPLOT platform interface. At the top, there is a navigation bar with links for Home, My workspace, Datasets, Tools, Blockchain, Help, and user authentication (Login, Signup). A green banner on the left side of the main content area reads "Encrypted for security" and "Code-free analysis". Below this banner, a paragraph of text describes the platform's reliability and security features, mentioning automatic bioinformatics analysis, block chain, multi-party secure computing, and other cutting-edge technologies. Two buttons are present: "START >>" in a green box and "TOUR >>" in a white box. On the right side, a section titled "What services we could provide" is displayed, featuring four cards with icons and descriptions: "Secure computing environment", "Personalized workspace", "Freely accessible datasets", and "Open community supported workflows". Each card has a "More" button at the bottom.

Encrypted for security
Code-free analysis

This is a reliable and flexible computing platform. Users can do automatic bioinformatics analysis without programming background. At the same time, block chain, multi-party secure computing and other cutting-edge technologies are employed to ensure the security of users' data.

START >> TOUR >>

What services we could provide

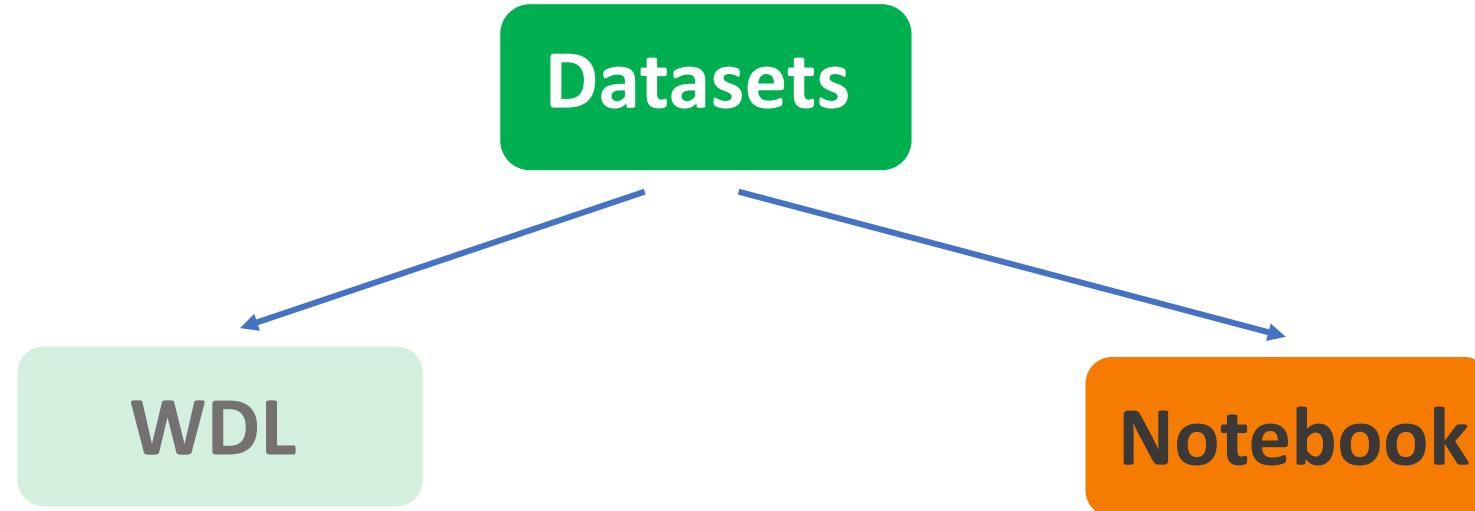
Secure computing environment
Multi-level authority control, data storage & transmission encryption, blockchain and secure multi-party computing are used to ensure the security of platform datasets and analysis.

Personalized workspace
Building computing sandbox in workspace, users can collaborate with researchers or team members in different regions by sharing space.

Freely accessible datasets
High-quality and well curated datasets collects datasets covering fields of plants, animals, microorganisms and diseases.

Open community supported workflows
This platform provides a highly efficient and human readable WDL based developing environment to build bioinformatic workflows.

Function



● Zero-code | streamline batch analysis

Based on standardized **WDL** language

Customize tuning parameters

● Low-code | costume analysis with notebook

The **Jupyter notebook** is deployed to provide
Python, R and other packages

Datasets

- Curated datasets (21 datasets so far)
 - Ensemble plant datasets (96 plant genomes)
 - COVID-19 datasets (~10 million seqs)
 - Single-cell datasets (21 species)
- User-owned datasets
- Publication support

首页 / 数据集 / The Cycas genome and the early evolution of seed plants

The Cycas genome and the early evolution of seed plants

The cycad genome project is an integration of genomic data of cycads and other related seed plants, including the raw sequencing data, assembly and annotation.

数据量: 444 更新时间: 2022-04-19 克隆

[概述](#) [数据](#) [工作流程](#)

1. Backgroud

Introduction to cycads.
Cycads are long-lived, woody and dioecious gymnosperms that develop cones and reproduce by seeds characterized by their frond like leaves. Today, they compose one of the largest lineages of gymnosperms comprising ca. 360 living species (<http://www.cycadlist.org>) that widely distributed across tropical and subtropical regions. As cycads are among the most ancient lineages of living seed plants, the cycad genome project provides great resources for a better understanding of the origin and early evolution of seed plants.

Cycad genome database
The cycad genome database is an integration of genomic data of cycads and other related seed plants, including the raw sequencing data, assembly and annotation. Assemblies are from cycad genomes, female and male specific regions of cycad genomes, and transcriptomes of cycads and other gymnosperm species. The annotations included repeat, gene, and functional annotation of the cycad genome, as well as open reading frame predictions of transcriptomes.

2. Data description

2.1 Genome

A Cycas panzhihuaensis genome was assembled and polished by modified softwares NextDenovo and NextPolish. After conjunction with Hi-C chromosome conformation, the C. panzhihuaensis genome comprises 10.5 Gb in 5,123 contigs (N50 = 12 Mb), with 95.3% of the assembled contigs anchored to the largest 11 pseudomolecules, corresponding to the 11 chromosomes (n = 11) of the C. panzhihuaensis karyotype.



WDL: enter and get results

<https://db.cngb.org/codeplot/>

- **30+ WDLs**
 - **Spatial transcriptomics analysis**
 - **Single-cell analysis**
 - **GWAS**
 - **Others**
- **Self-defined WDL**

Select the WDL



Select inputFile

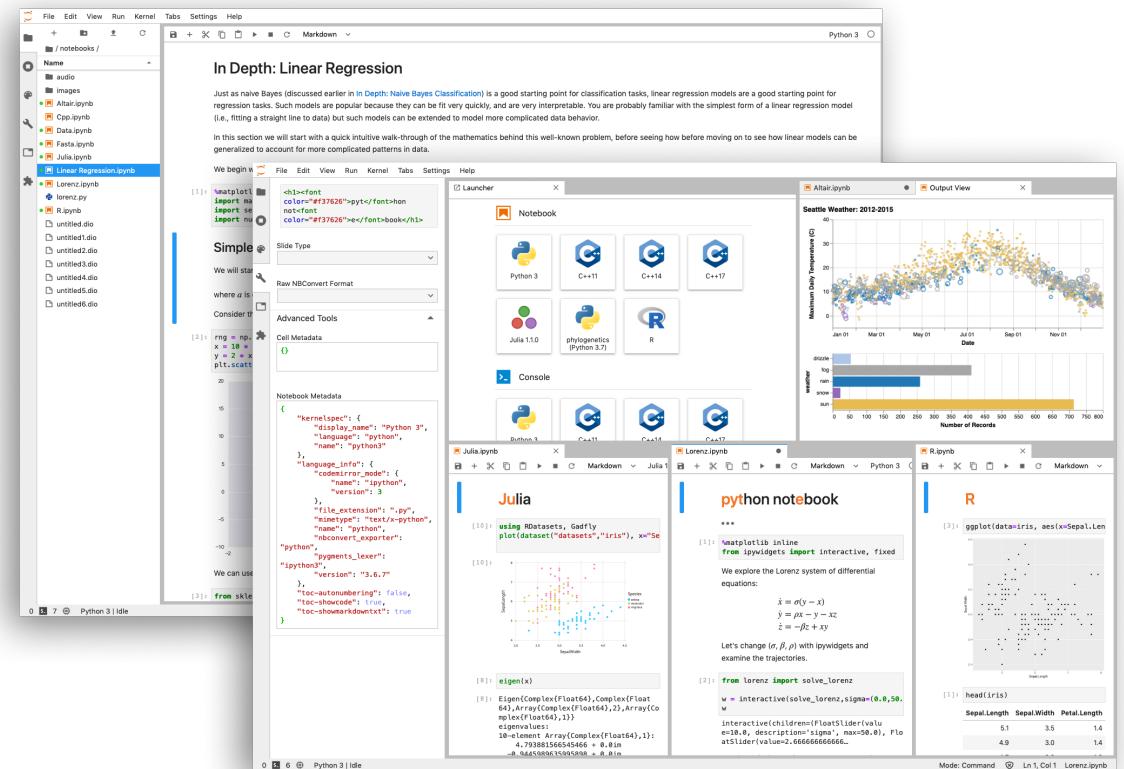


Get results

Notebook

- Cell block
- Data visualization
- Interactive analysis

<https://db.cngb.org/codeplot/>



Summary



- **Visualization**
- **Data archive**
- **Collection**



- **Codeplot is a reliable, flexible computing platform for bioinformatic analyses, which could facilitate biological data sharing and analysis.**