

Visualization and Computation platforms

Weiwen (Raymond) WANG

王伟文

2023.9

Overview

The logo for CoDEPLOT, featuring the text 'CoDEPLOT' in a green, sans-serif font. The 'o' is replaced by a green circle with a dot inside, and the 'T' is replaced by a green circle with a vertical line and a dot inside. The logo is set against a background of two overlapping, curved, light green and blue shapes.

CoDEPLOT



Why we build these platforms

The logo for STomicsDB, featuring the text 'STomicsDB' in a bold, black, sans-serif font. The 'S' is replaced by a blue square with a green square inside it, and the 'o' is replaced by a blue circle. The logo is set against a background of two overlapping, curved, light blue and green shapes.

STomicsDB

CNSA: data archiving, preservation, and sharing

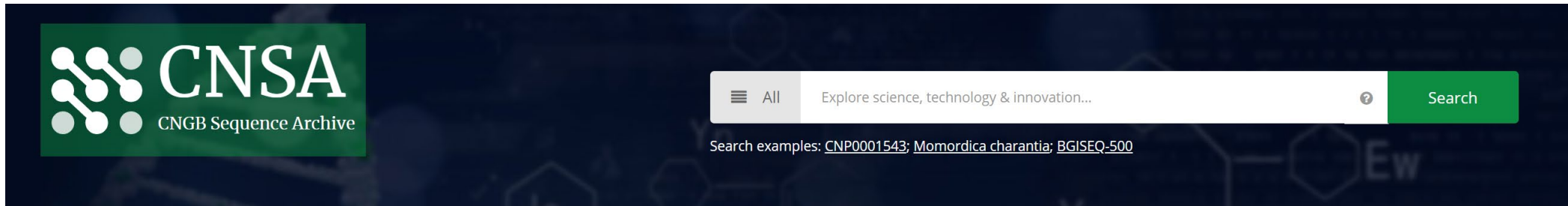

DATA
11,889TB


PROJECTS
4,573

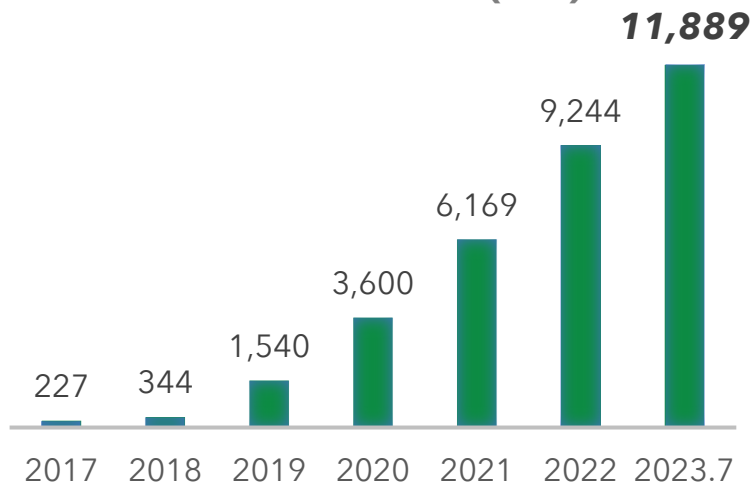

INSTITUTIONS
481


PUBLICATIONS
1,258


VISITS
> 10 Millions



DATA FILE SIZE (TB)



Data management certifications


FAIRsharing.org
standards, databases, policies


Open
DOAR


re3data.org



- ***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  **STomicsDB**

CoDEPLOT 

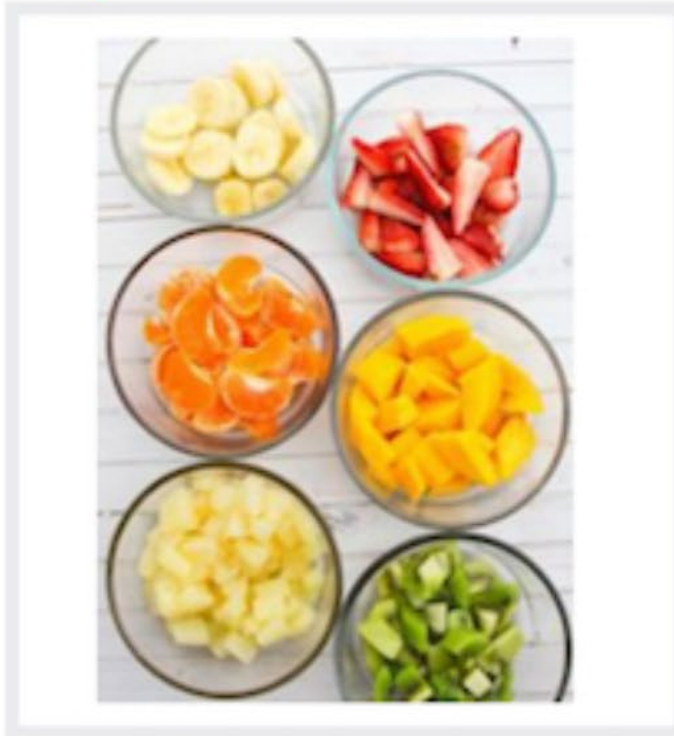
1. Background

What is spatial transcriptomics

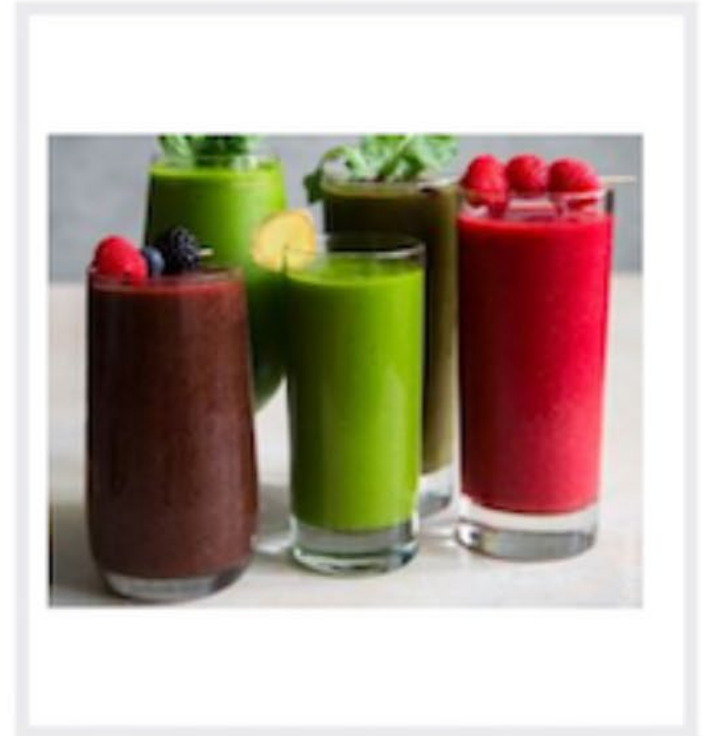
Spatial



Single cell



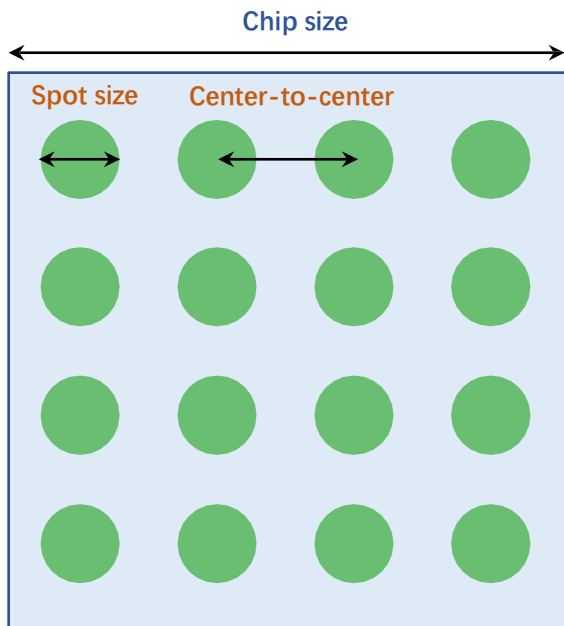
Bulk



1. Background

Comparison

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

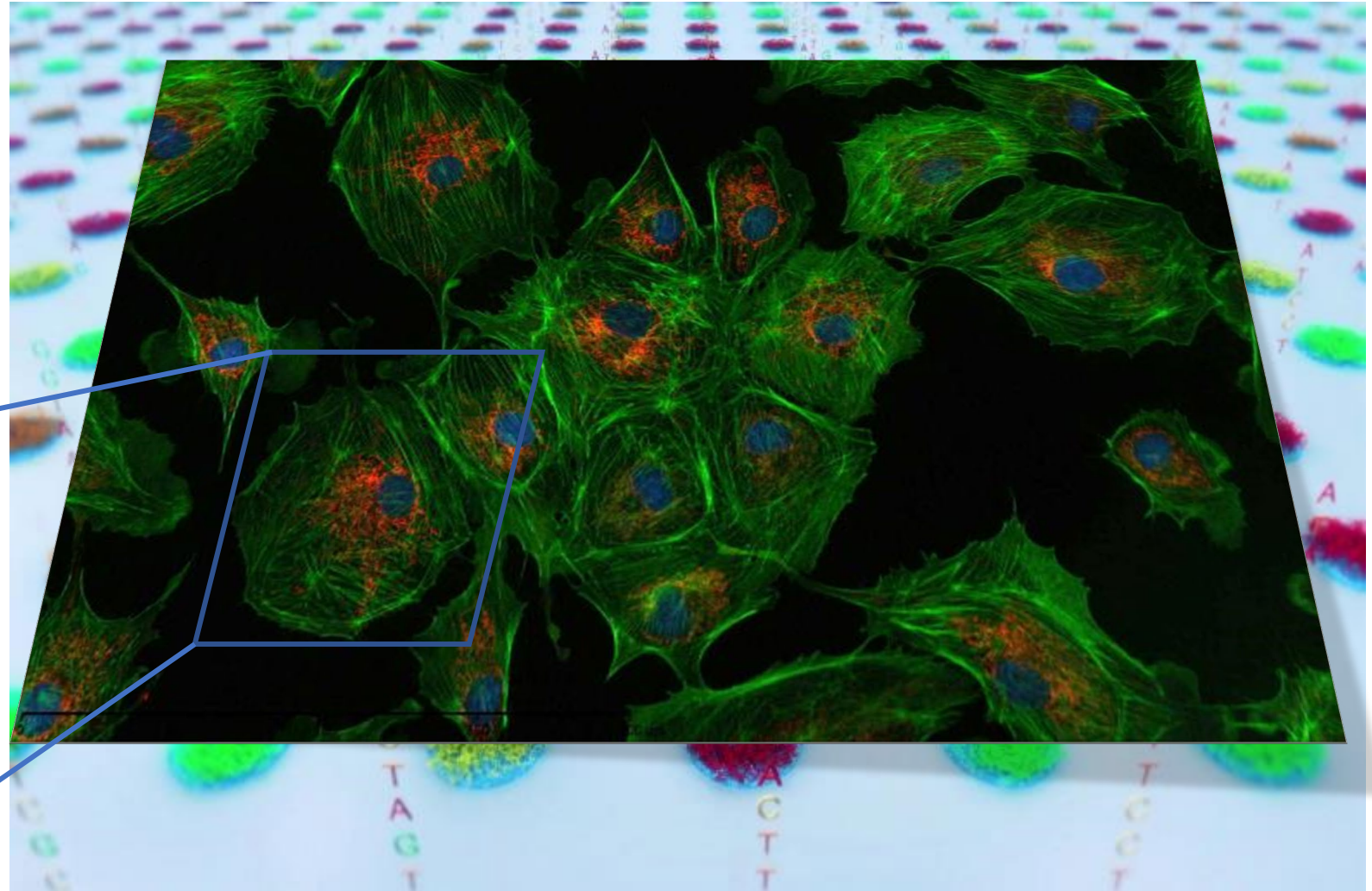
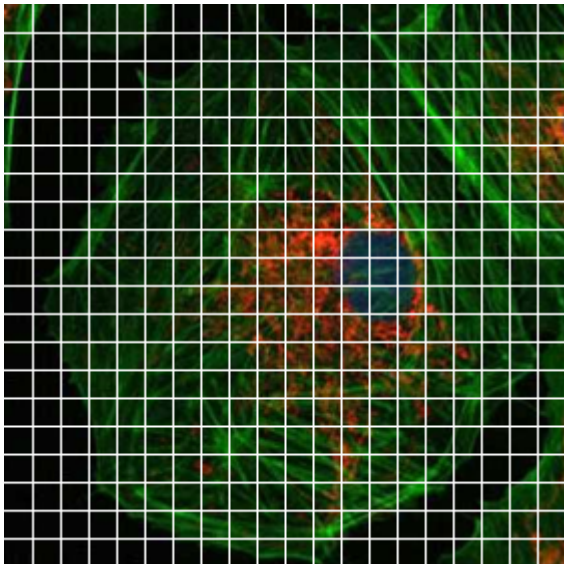


| | DBiT-seq | Slide-seq | HDST | 10x | Stereo-seq |
|------------------------------------|-----------|------------|-----------|-----------|------------|
| Spot size (μm) | 10 | 10 | 2 | 55 | 0.5 |
| Center-to-center (μm) | 20 | 10 | 2 | 100 | 0.7 |
| Field of view (mm) | 1.0 x 1.0 | Φ 3.0 | 5.7 x 2.4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

1. Background

Resolution

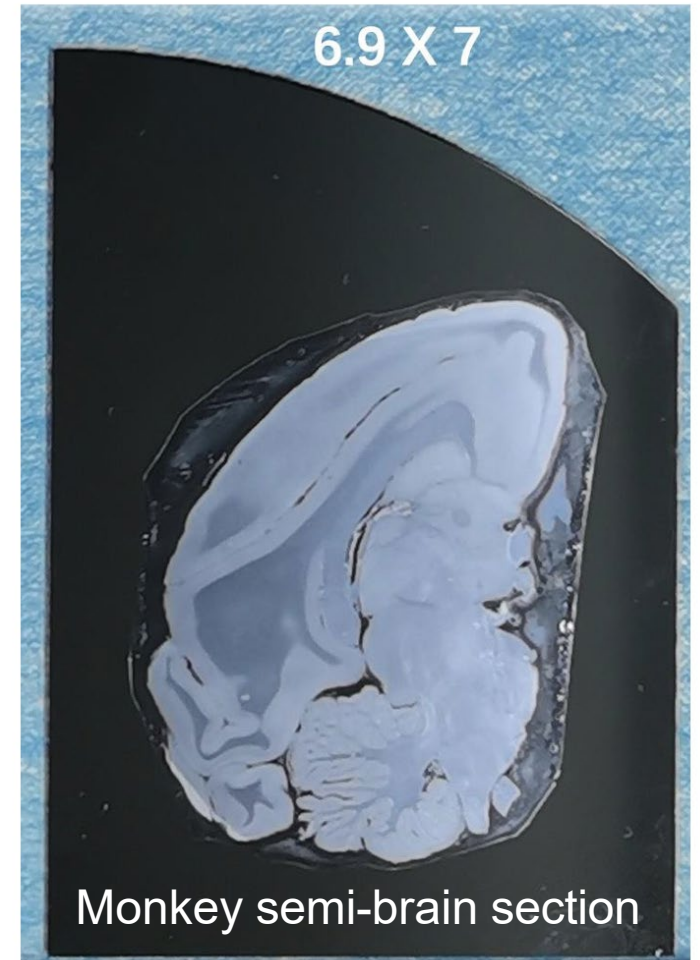
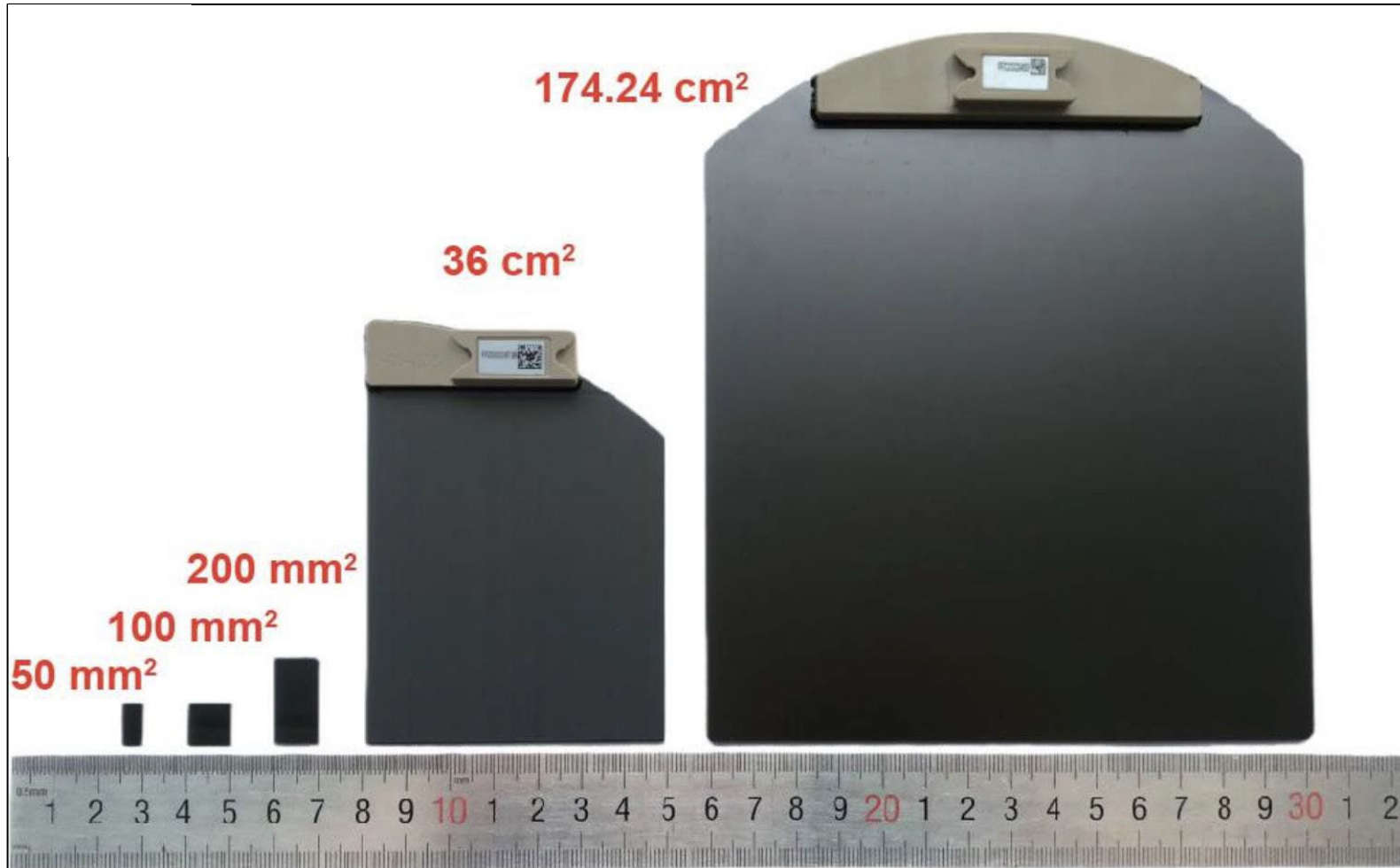
- Cell size: 3-30 μm
- Stereo-seq :
0.5 μm (Subcellular resolution)



1. Background

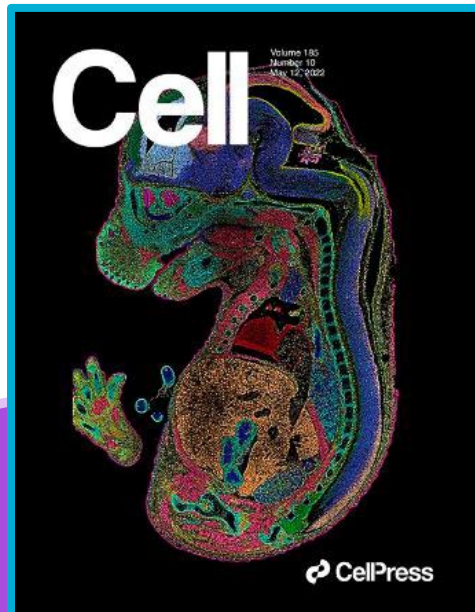
Field of view

- **Stereo-seq chips:** ranging from 50 mm² to 174.24 cm²



1. Background

Publication examples



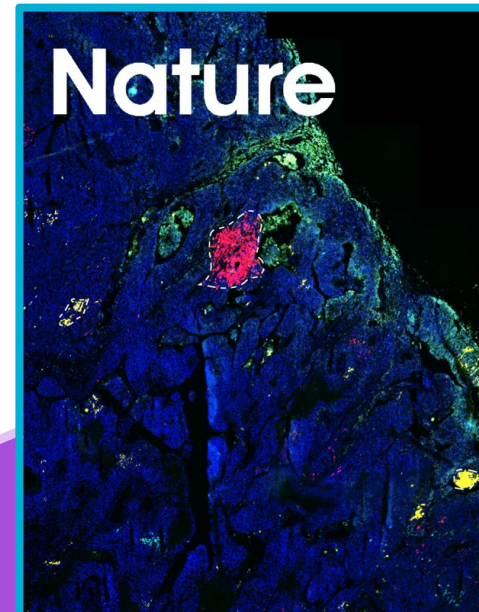
2022.5.4

Mouse Organogenesis
Spatiotemporal
Transcriptomic Atlas



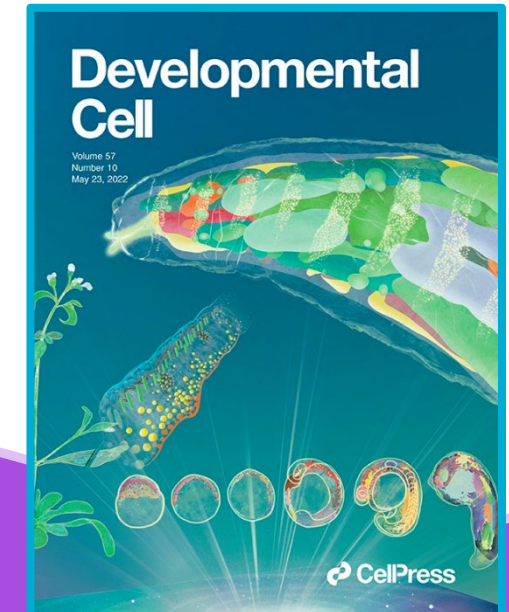
2022.9.2

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



2022.9.21

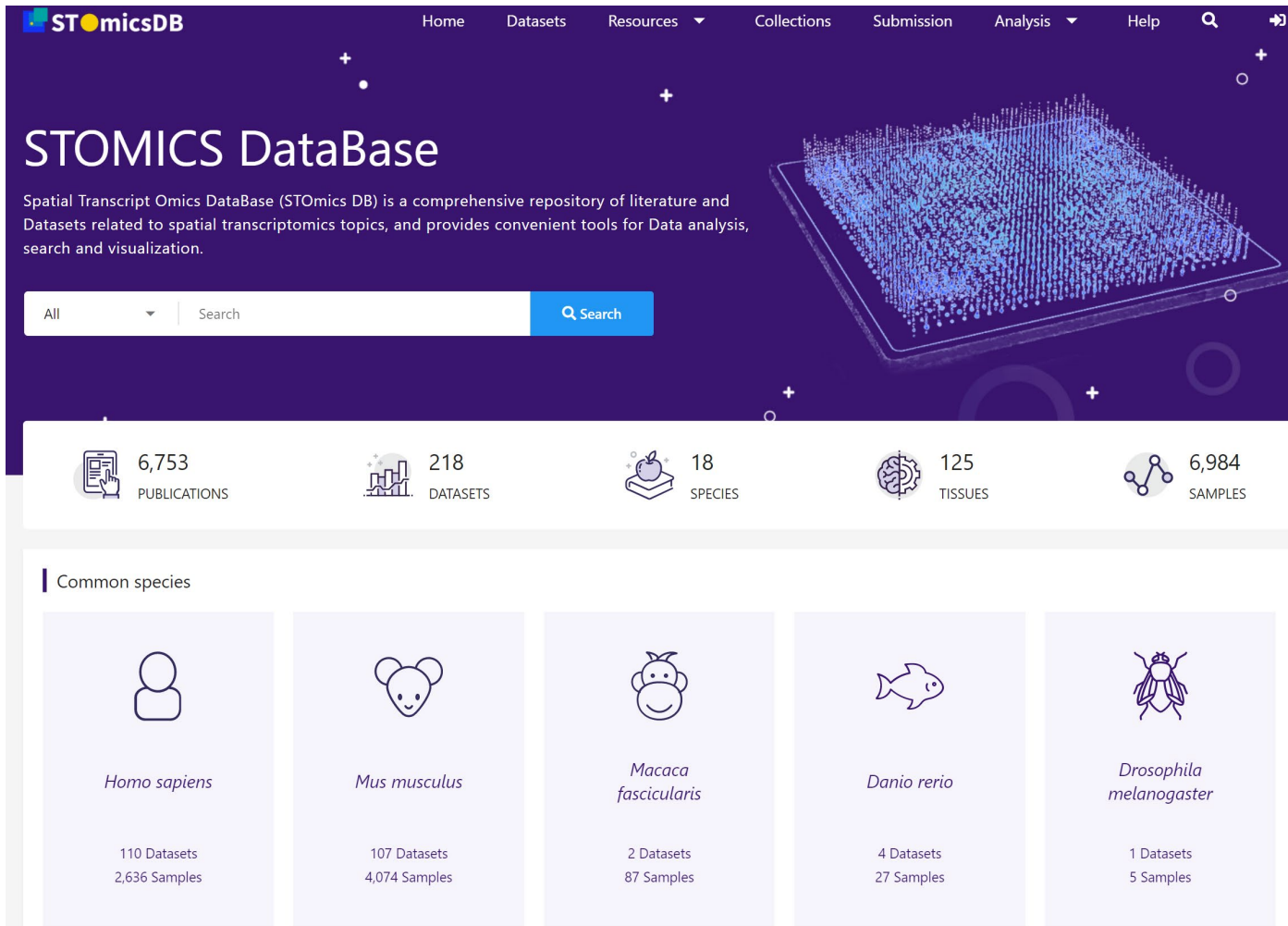
A cellular hierarchy in
melanoma uncouples
growth and metastasis



2022.5.5

Zebrafish/Drosophila
Embryogenesis
Spatiotemporal
Transcriptomic Atlas

Spatial TranscriptOmics DataBase



The screenshot shows the STOmicsDB website interface. At the top, there is a navigation bar with links for Home, Datasets, Resources, Collections, Submission, Analysis, and Help. Below the navigation bar, the main header reads "STOMICS DataBase" and includes a brief description: "Spatial Transcript Omics DataBase (STOmics DB) is a comprehensive repository of literature and Datasets related to spatial transcriptomics topics, and provides convenient tools for Data analysis, search and visualization." A search bar is located below the header. The main content area features a grid of statistics: 6,753 PUBLICATIONS, 218 DATASETS, 18 SPECIES, 125 TISSUES, and 6,984 SAMPLES. Below this, a section titled "Common species" displays five cards for Homo sapiens, Mus musculus, Macaca fascicularis, Danio rerio, and Drosophila melanogaster, each with its respective icon and dataset/sample counts.

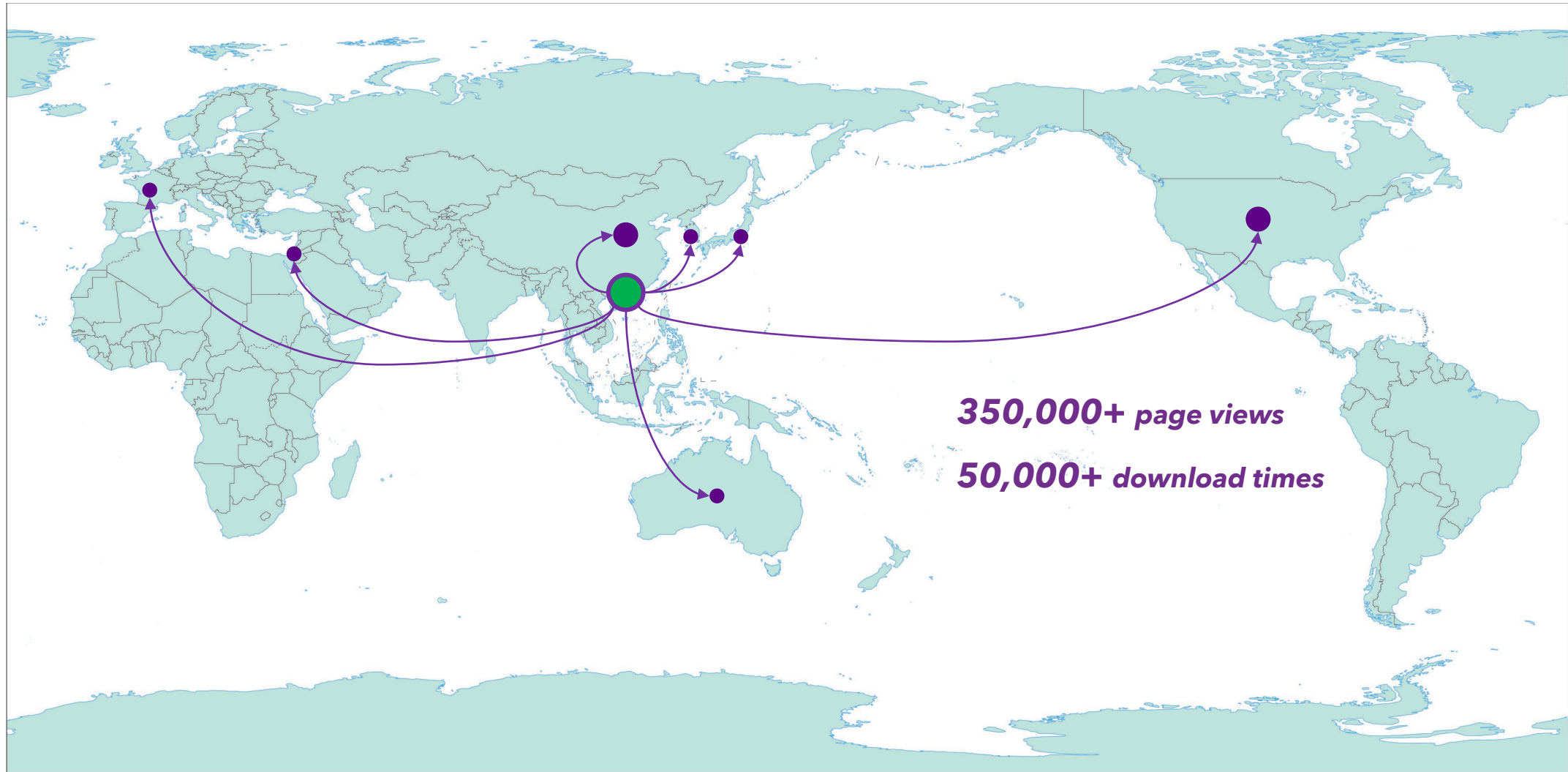
| Species | Datasets | Samples |
|--------------------------------|----------|---------|
| <i>Homo sapiens</i> | 110 | 2,636 |
| <i>Mus musculus</i> | 107 | 4,074 |
| <i>Macaca fascicularis</i> | 2 | 87 |
| <i>Danio rerio</i> | 4 | 27 |
| <i>Drosophila melanogaster</i> | 1 | 5 |

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

db.cngb.org/stomics



Overview



Structure

Application

Data exploration,
Online analysis,
Data archive,
Data download



Collections

(MOSTA: mouse Organogenesis;
MBA: monkey brain atlas
etc.)

Tool

Data visualization

Gene search

Dataset comparison

Integrative analysis



Data

Other public domain

Data submission

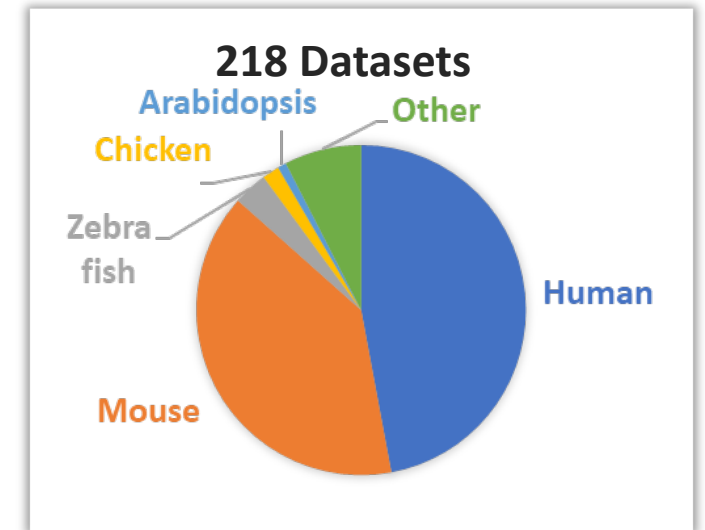
Data

- **Collected (218 datasets so far)**

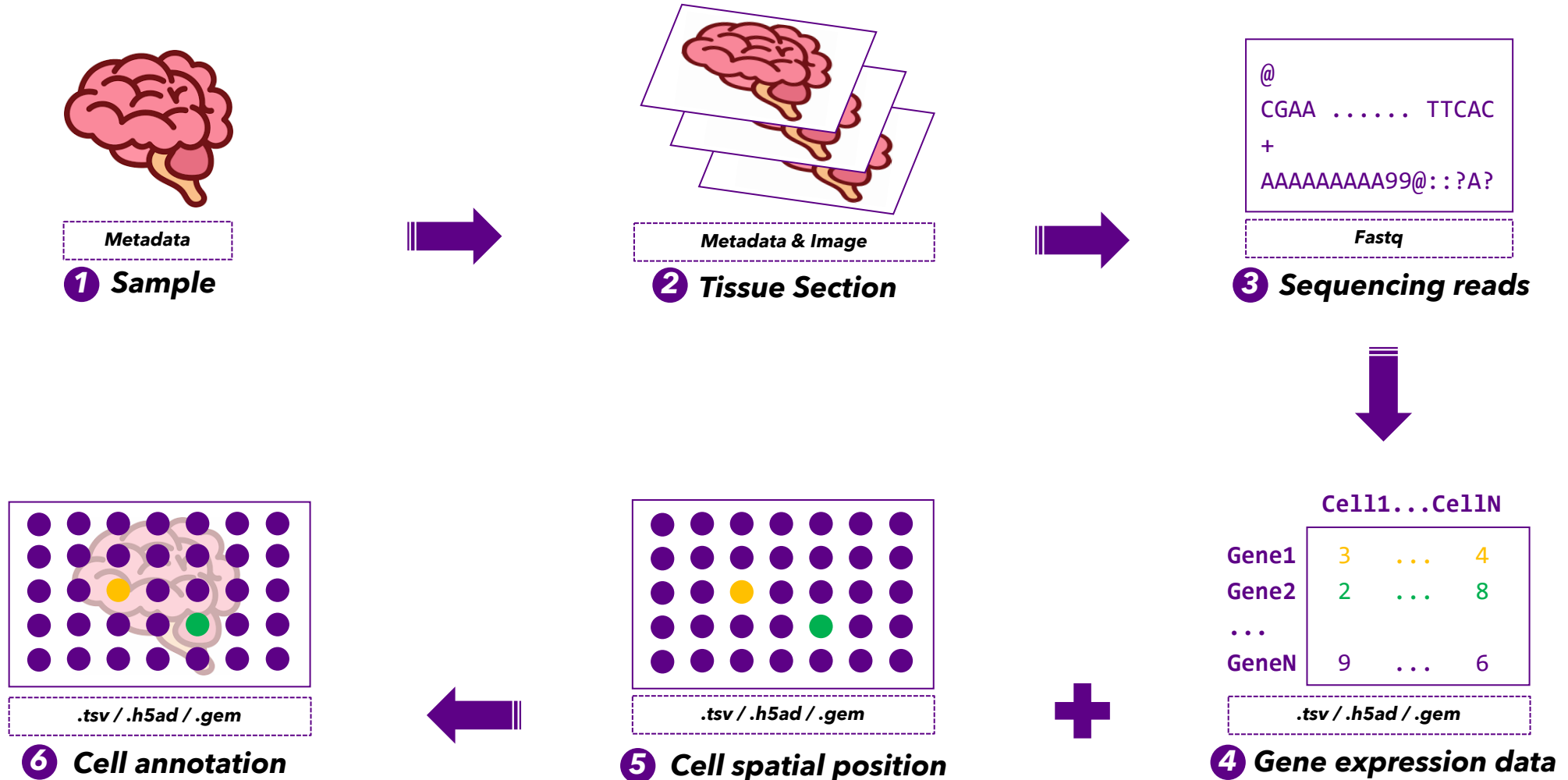
- from NCBI, EBI, DDBJ, etc.
- from papers
- from submission

- **Curated**

- display general information of each dataset (summary, overall design, species, cell types, development stage, sample number, section number, etc)



Data archiving system



Data archiving system

STOMICS DB Home liling3@ge*****

STomics Sub

STomics submission

New submission

Filter submission ID, project accession and title.

| 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).

STOMICS DB Home Publications Datasets Stereomics Tools Submission Support Please enter key words Login

Large field of view-spatially resolved transcriptomics at nanoscale resolution

STOmics technology: Stereo-Seq

Organism: Homo sapiens

Data type: STomics, Raw sequence read

Sample scope: Multisite

Summary: High-throughput profiling (DNA) patterned array chips and in situ applied Stereo-seq to the adult mouse of tissues and organisms.

Contributors: Zheng C, Hu Y

Publication: Niu Y, Sun N, Li C, Lin Y et al. *Nature* 2018; 561: 52-57

Submitter: 梁良 (Liang Wu), BGI-Shenzhen

DOI: 10.26035/STT0000001

Release date: 2018-03-29

Updated: 2018-03-30

Reference project: CNP001543

Statistics: Sample: 8 Tissue Section: 20 Experiment: 20 Run: 20 Dataset: 20

Project accession: STT0000001 Title: Derivation of formative-like pluripotent stem cells from mammalian embryos [RNA-Seq]

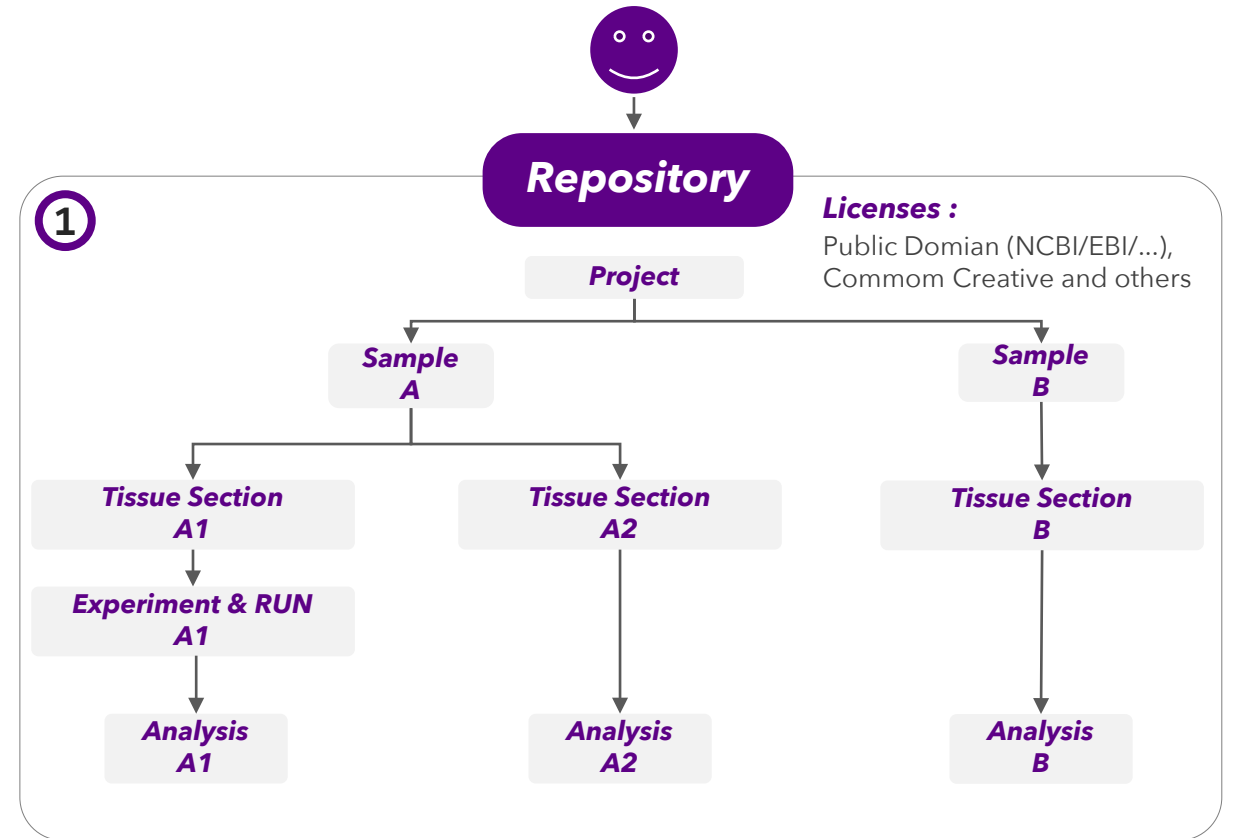
Stereo-seq

| Project | Sample | Tissue Section | Organism | Spatial Gene Expression & Visualization | Files |
|------------|------------|----------------|--------------|---|-------|
| STT0000001 | STT0000001 | STT0000001 | Homo sapiens | | |

geneID x y MIDCount

| | | | |
|---------|--------|--------|---|
| Cr11 | 105228 | 101106 | 1 |
| Cr11 | 112259 | 105201 | 1 |
| Cr11 | 111661 | 105876 | 1 |
| Cr11 | 113586 | 106877 | 1 |
| C846 | 109344 | 103929 | 1 |
| C446 | 112019 | 106939 | 1 |
| Ga32250 | 112968 | 104894 | 1 |
| C634 | 106601 | 110120 | 1 |
| C634 | 106992 | 108887 | 1 |
| C434 | 112123 | 108884 | 1 |
| C634 | 112566 | 102785 | 1 |
| C434 | 113048 | 109092 | 2 |

Submitter



Datasets & Collections

2

Visualization

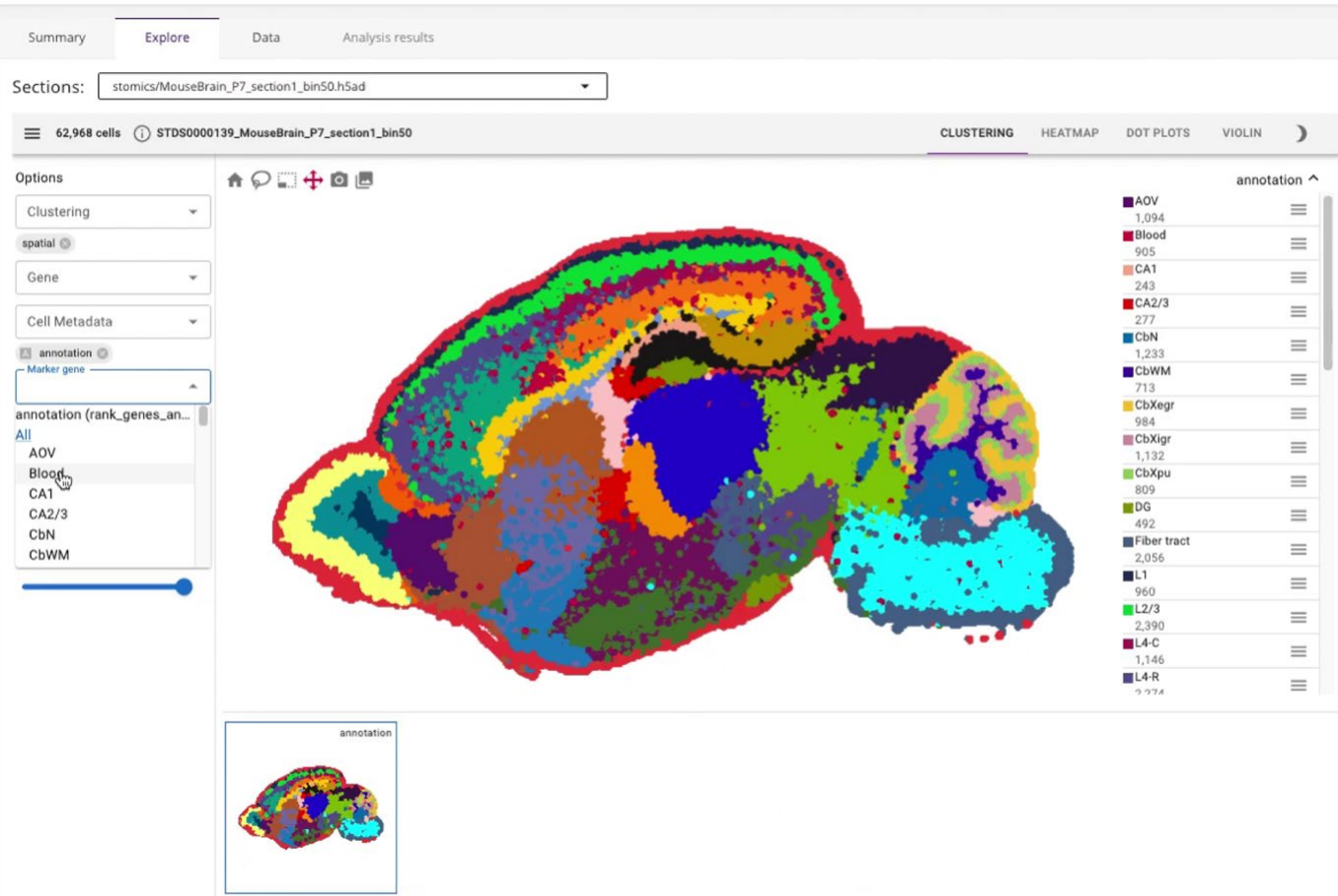
Discovery

Knowledge

Visualization

A cellular resolution spatial transcriptomic landscape of the postnatal mouse brain

Dataset ID: STDS0000139 | 120,000 Spots | 27,330 Genes



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

MOSTA: Mouse Organogenesis Spatiotemporal Transcriptomic Atlas
Dataset ID: STDS0000058 | PMID: 35512705 | 351,014 Spots | 28,879 Genes

Summary | Explore | Data | Analysis results

Sections: Mouse_embryo_all_stage.h5ad

520,815 cells | STDS0000058_Mouse_embryo_all_stage

CLUSTERING | HEATMAP | DOT PLOTS | VIOLIN

Options

- Clustering
- spatial
- Gene
- Cell Metadata
- annotation
- Marker gene

View

- Marker Size: default
- Color Scheme
- Opacity

annotation

- Adipose tissue: 7,916
- Adrenal gland: 345
- AGM: 452
- Blood vessel: 4,623
- Bone: 3,400
- Brain: 86,520
- Branchial arch: 1,410
- Cartilage: 13,903
- Cartilage primordium: 18,818
- Cavity
- Choroid plexus: 5,683
- Connective tissue: 48,462
- Dermomyotome: 3,239
- Dorsal root ganglion: 7,487
- Epidermis

Select different sample

Select the gene of interest

Switch to the comparison of the gene expression module

Cell annotation

Gene search

*Species: All species | Tissue: | *Gene: Input a gene | Search

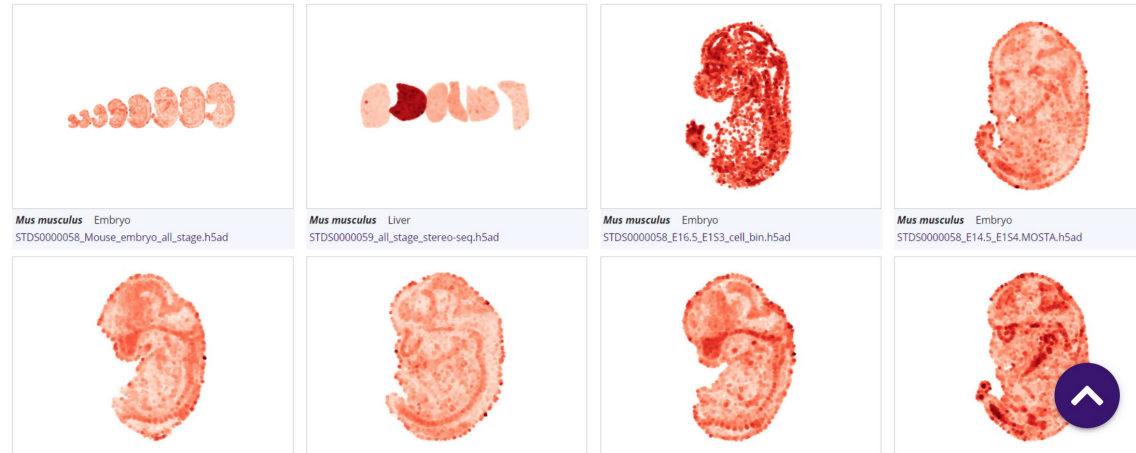
Select species & gene

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 >>



6,000+
curated sample slides

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.

Gene search

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

Select species Select tissue Type the gene

*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 >>

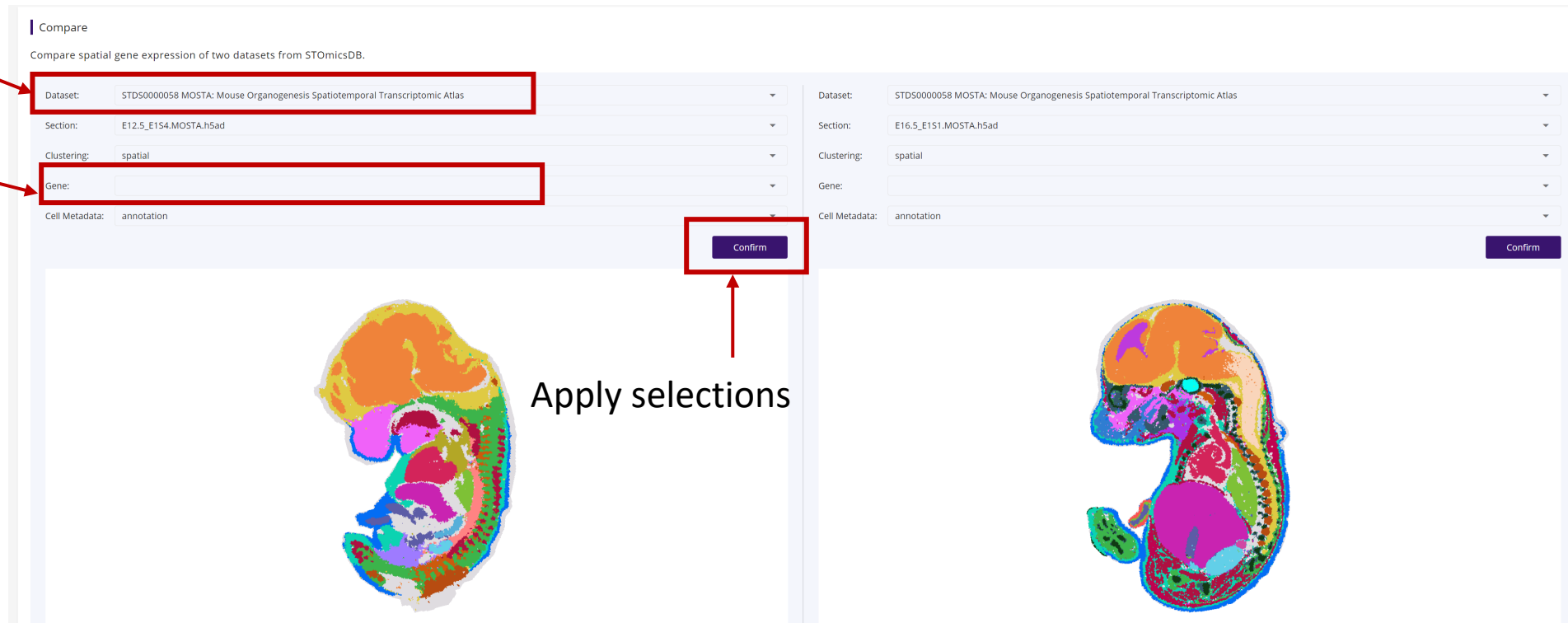
| | | | |
|---|--|--|---|
| <i>Mus musculus</i> Embryo STD50000058_Mouse_embryo_all_stage.h5ad | <i>Mus musculus</i> Liver STD50000059_all_stage_stereo-seq.h5ad | <i>Mus musculus</i> Embryo STD50000058_E16.5_E1S3_cell_bin.h5ad | <i>Mus musculus</i> Embryo STD50000058_E14.5_E1S4.MOSTA.h5ad |
| | | | |

Dataset comparison

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

Select dataset

Select gene



Compare

Compare spatial gene expression of two datasets from STOmicsDB.

Dataset: STDS0000058 MOSTA: Mouse Organogenesis Spatiotemporal Transcriptomic Atlas

Section: E12.5_E1S4.MOSTA.h5ad

Clustering: spatial

Gene:

Cell Metadata: annotation

Confirm

Apply selections

Dataset: STDS0000058 MOSTA: Mouse Organogenesis Spatiotemporal Transcriptomic Atlas

Section: E16.S_E1S1.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

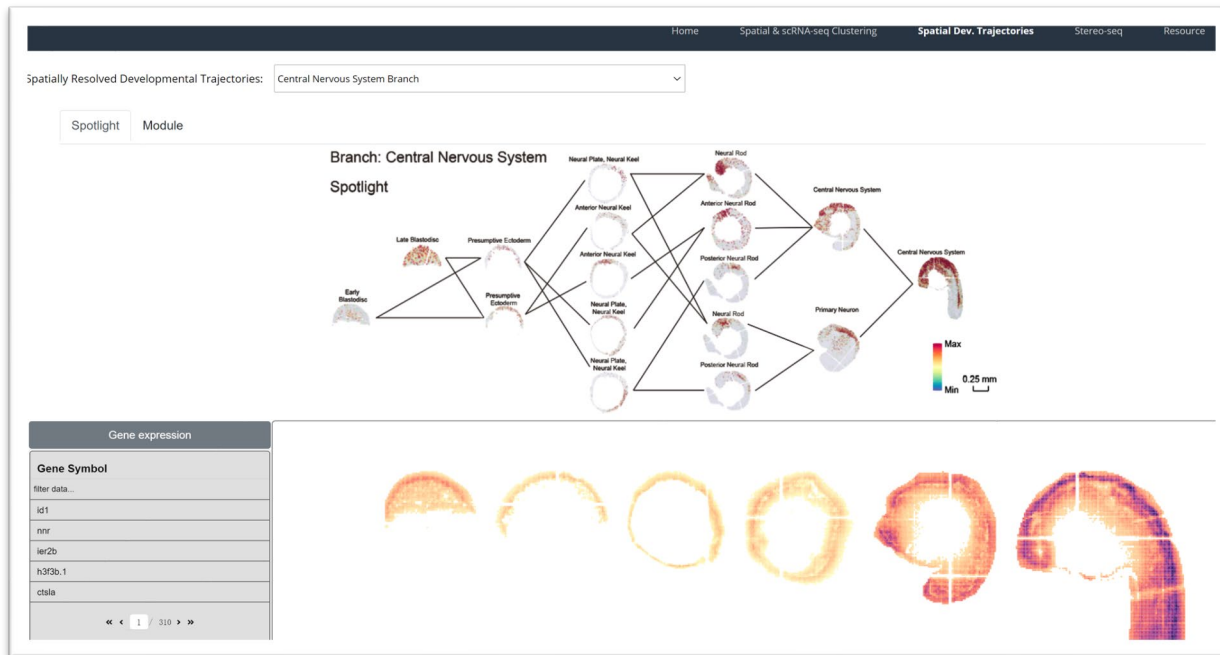
Collection entry

A screenshot of the STomicsDB website's 'Collections' page. The navigation bar includes 'Home', 'Resources', 'Datasets', 'Collections' (highlighted with a red box and an arrow), 'Submission', 'Analysis', and 'Help'. The user 'wangweiwen*****' is logged in. The main content area displays six collection entries in a 2x3 grid:

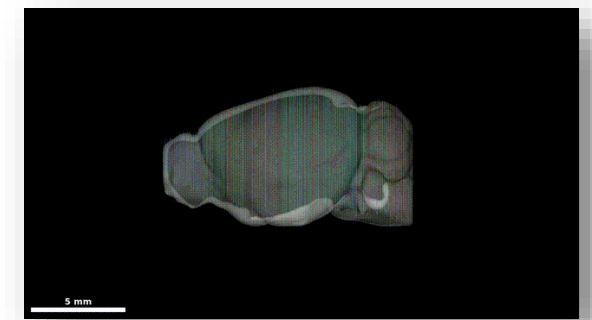
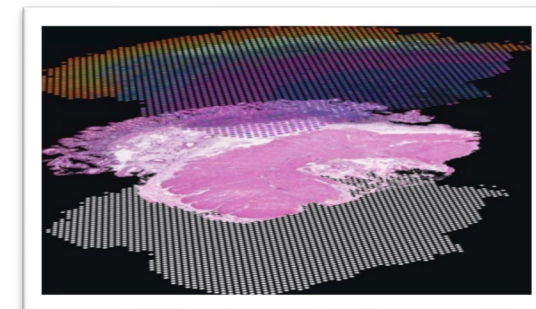
- MOSTA**: Mouse Organogenesis Spatiotemporal Transcriptomic Atlas. 300,000+ bins.
- ZESTA**: Zebrafish Embryogenesis Spatiotemporal Transcriptomic Atlas. 152,977 bin15, 91 Section, 86,307 Cells.
- Flysta3D**: High-resolution 3D spatiotemporal transcriptomic maps of developing Drosophila embryos and 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.

Collections

- Customizable visualization



Development Trajectory of Zebrafish

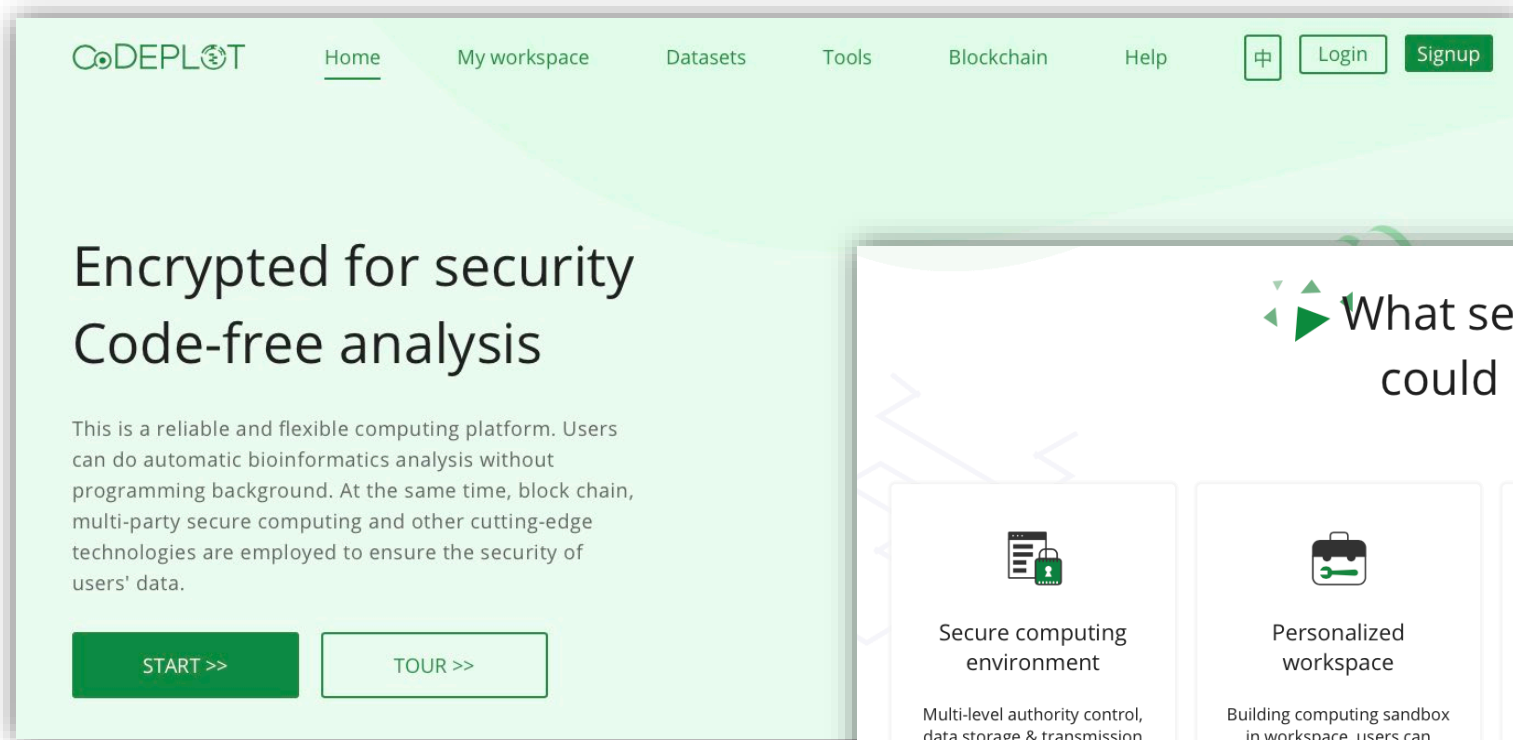


3D model of *Drosophila*

1  STomicsDB

  2

Codeplot: a platform for code-free analyses



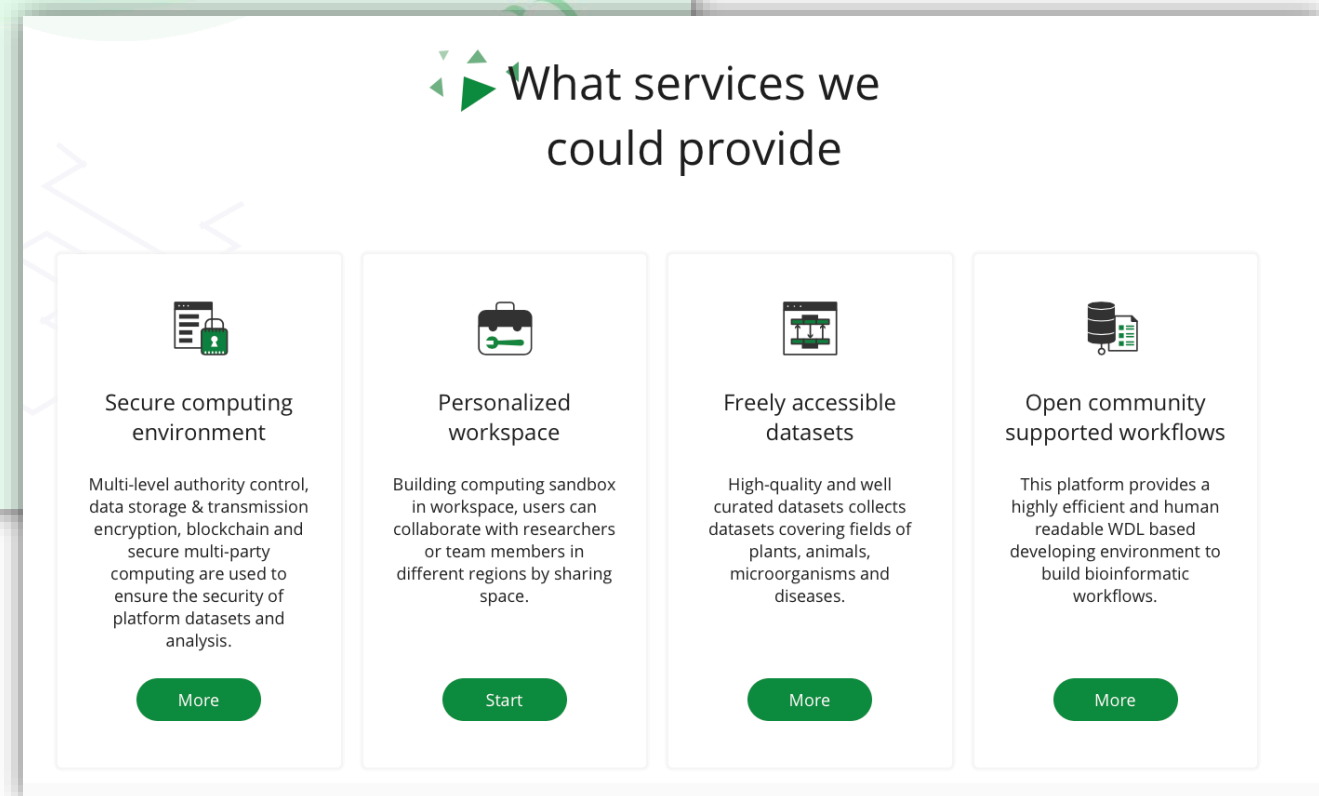
The screenshot shows the top navigation bar of the CoDEPLOT website. It includes the logo, a home link, and menu items for 'My workspace', 'Datasets', 'Tools', 'Blockchain', and 'Help'. There are also 'Login' and 'Signup' buttons. The main content area features the heading 'Encrypted for security Code-free analysis' and a paragraph describing the platform's security features. At the bottom of this section are two buttons: 'START >>' and 'TOUR >>'.

CoDEPLOT [Home](#) [My workspace](#) [Datasets](#) [Tools](#) [Blockchain](#) [Help](#) [Login](#) [Signup](#)

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 >>](#)




This section is titled 'What services we could provide' and features four service cards. Each card has an icon, a title, a description, and a button.


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.

[More](#)
- **Personalized workspace**

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

[Start](#)
- **Freely accessible datasets**

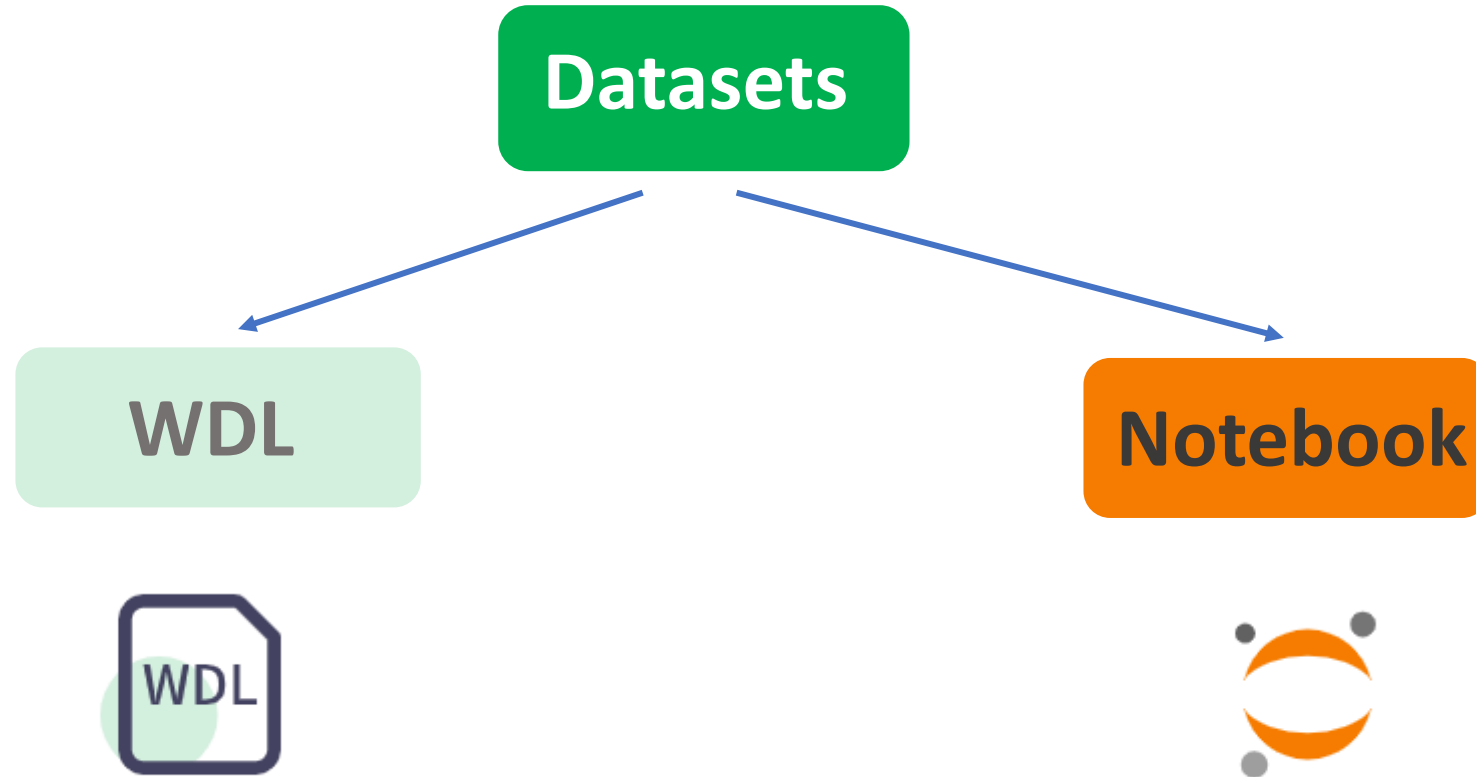
High-quality and well curated datasets collects datasets covering fields of plants, animals, microorganisms and diseases.

[More](#)
- **Open community supported workflows**

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

[More](#)

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-own 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. Background

Introduction to cycads.

Cycads are long-lived, woody and dioecious gymnosperms that develop cones and reproduced 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

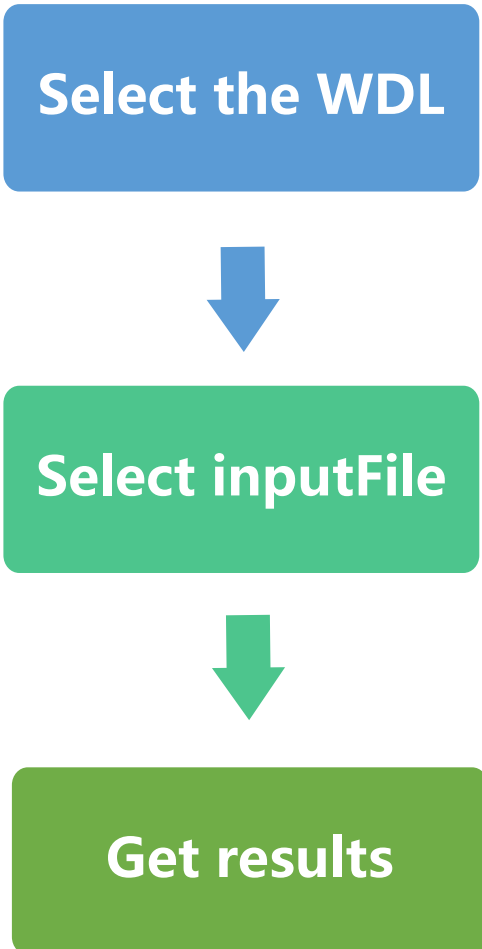
2.1 Genome

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/>

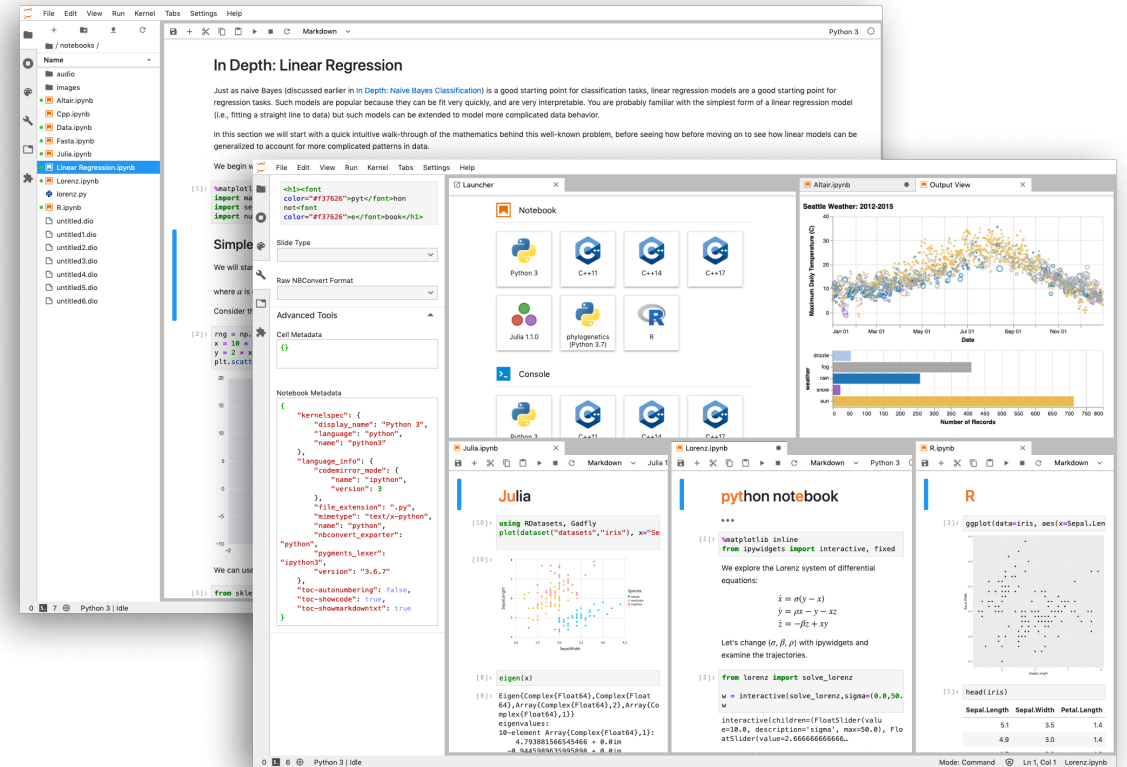
- **30+ WDLs**
 - Spatial transcriptomics analysis
 - Single-cell analysis
 - GWAS
 - Others
- **Self-defined WDL**



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

Notebook

- Cell block
- Data visualization
- Interactive analysis



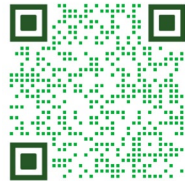
Summary

 **STomicsDB**



- Visualization
- Data archive
- Collection

 **CODEPLOT**



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