China National GeneBank 国家基因库

华大生命科学研究院 BGI·Research

Visualization and Computation platforms

Weiwen (Raymond) WANG

王伟文

2023.9

Overview

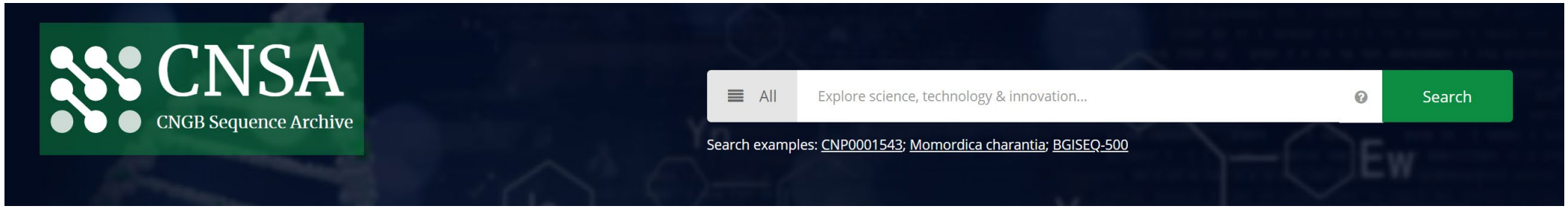
CoDEPLOT

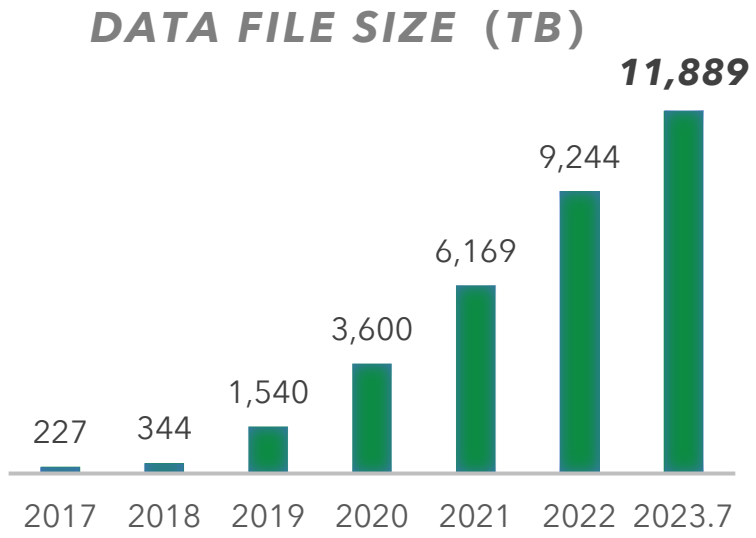
Why we build these platforms

STomicsDB

CNSA: data archiving, preservation, and sharing

DATA	PROJECTS	INSTITUTIONS	PUBLICATIONS	VISITS
11,889TB	4,573	481	1,258	> 10 Millions





Data management certifications

FAIRsharing.org standards, databases, policies

re3data.org

OpenDOAR

CORE TRUST SEAL

- ***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

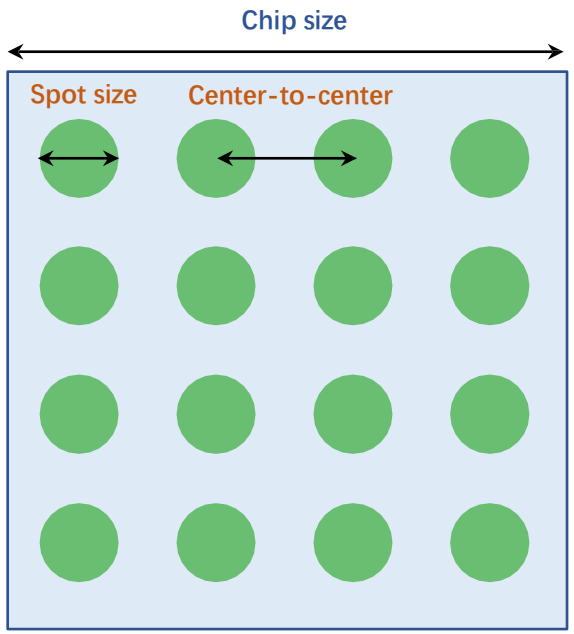
Bulk

Figures from 10X Visium company

1. Background

Comparison

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



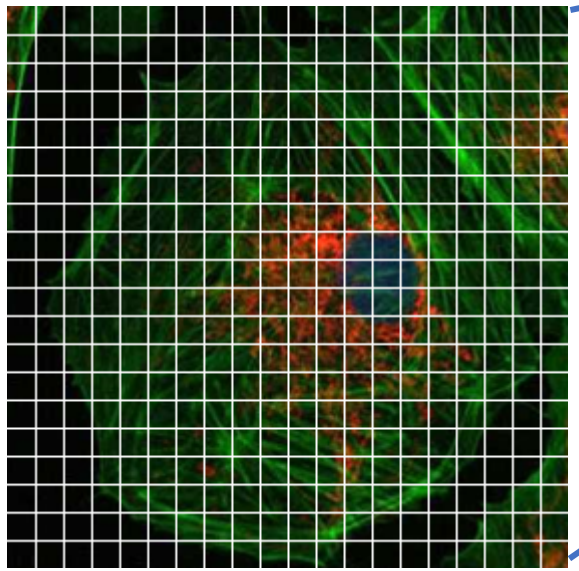
	DBiT-seq	Slide-seq	HDST	10x	Stereo-seq
Spot size (μm)	10	10	2	55	0.5
Center-to-center (μm)	20	10	2	100	0.7
Field of view (mm)	1.0 x 1.0	Φ 3.0	5.7 x 2.4	6.5 x 6.5	120 x 120

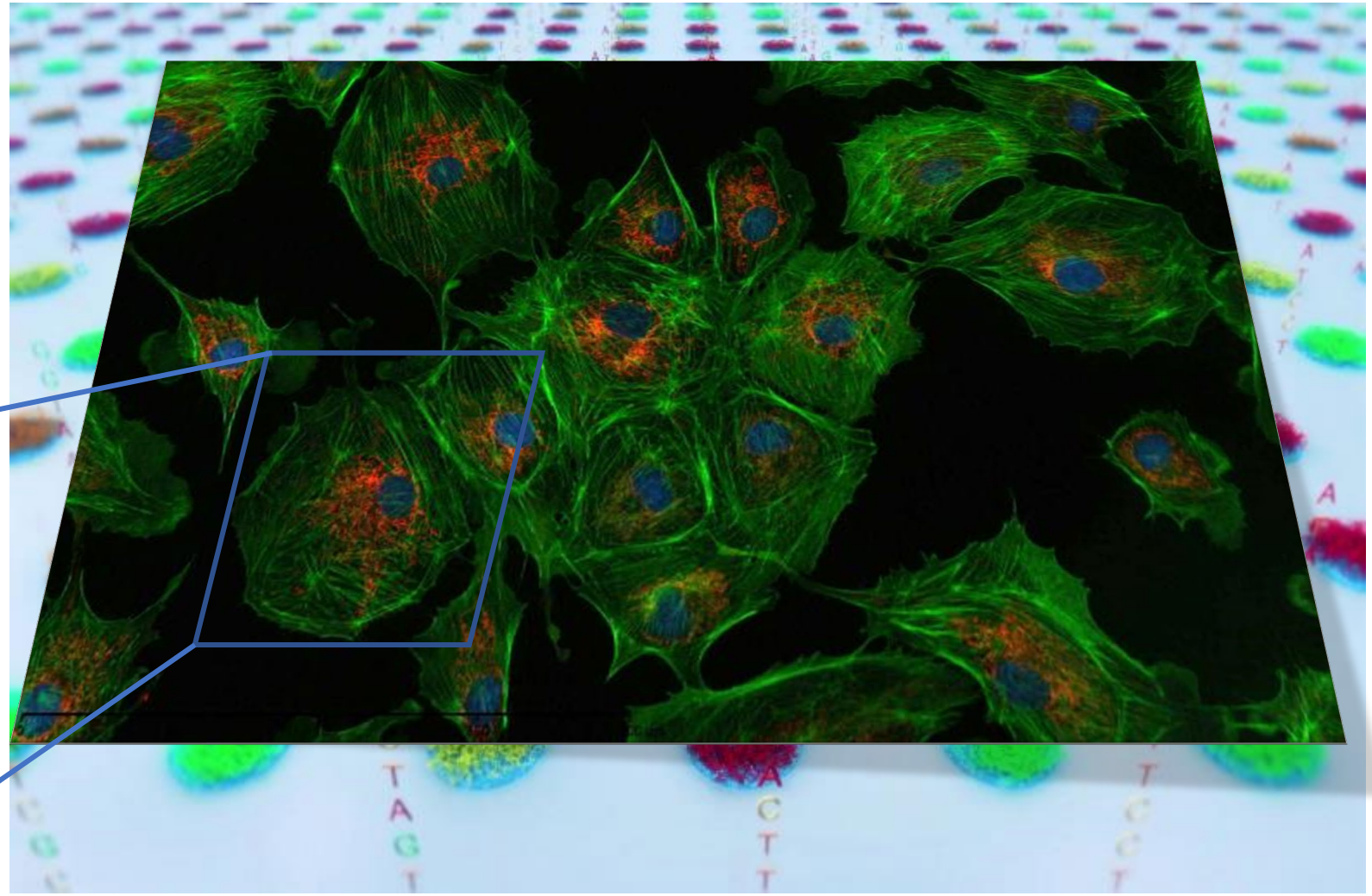
1. Background

Resolution

- **Cell size:** 3-30 μm
- **Stereo-seq :**

  0.5 μm (Subcellular resolution)

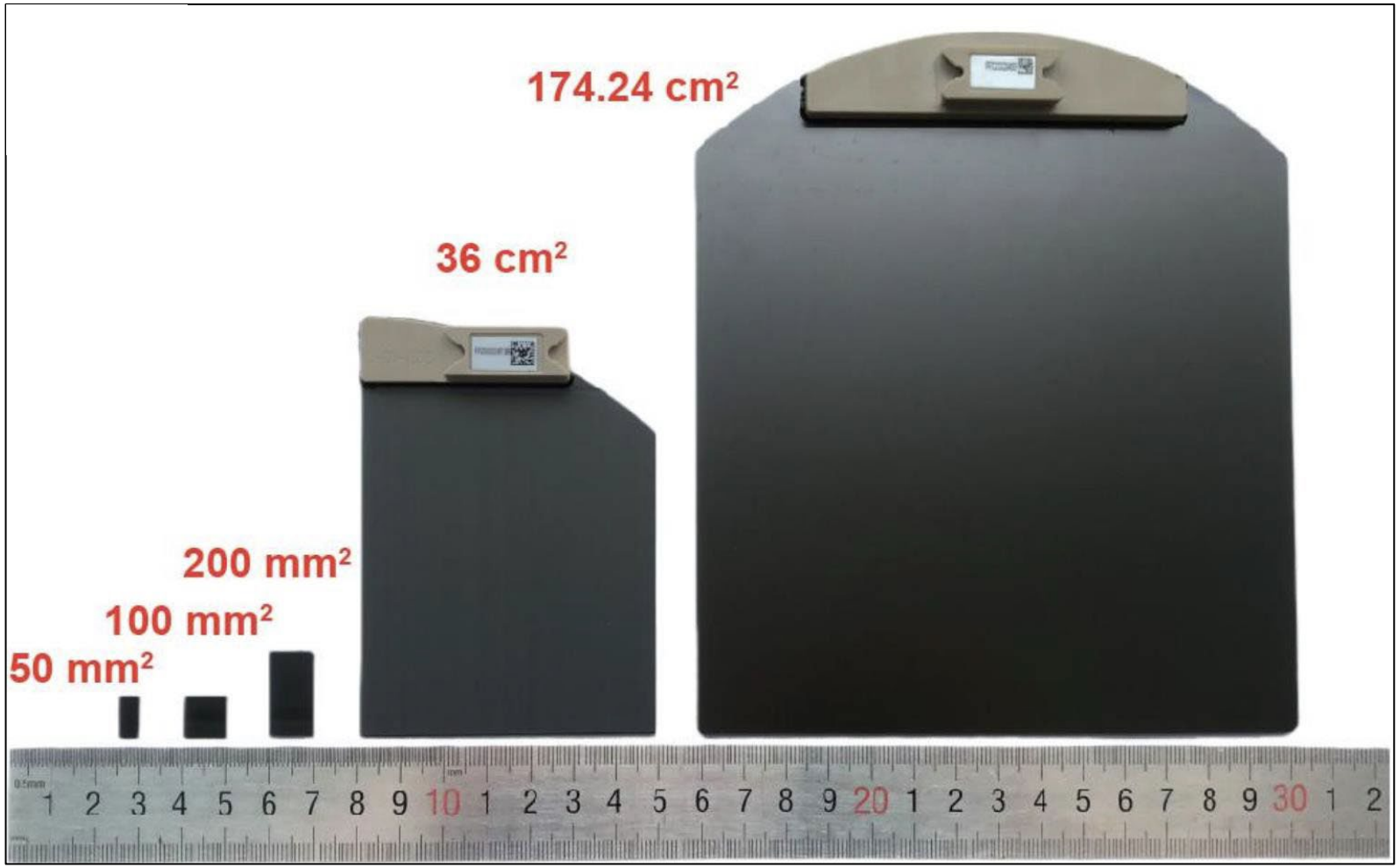


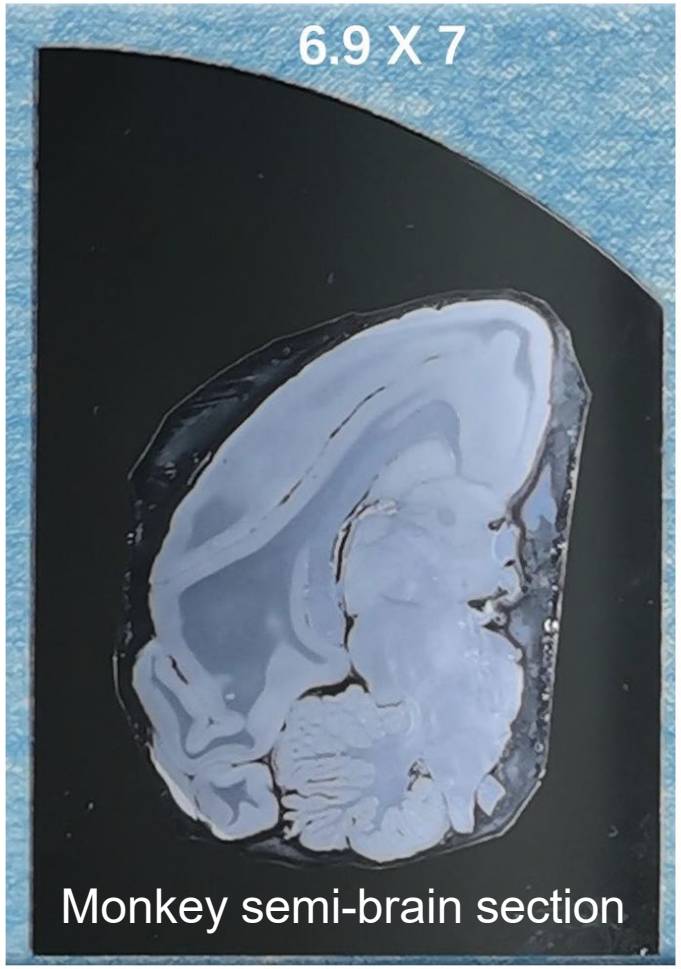


1. Background

Field of view

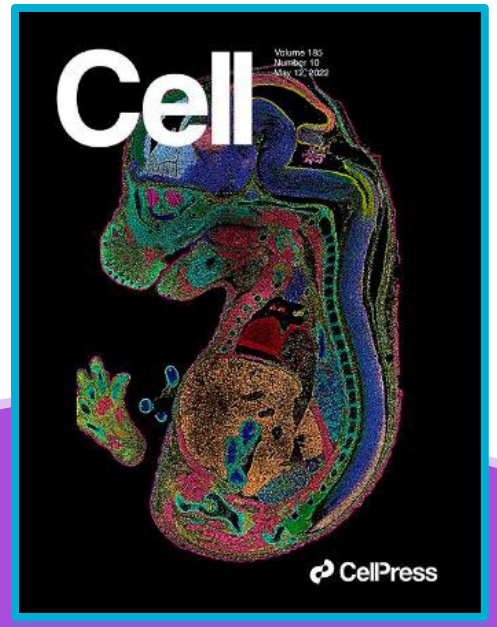
- **Stereo-seq chips:** ranging from 50 mm^2 to 174.24 cm^2





1. Background

Publication examples



2022.5.4

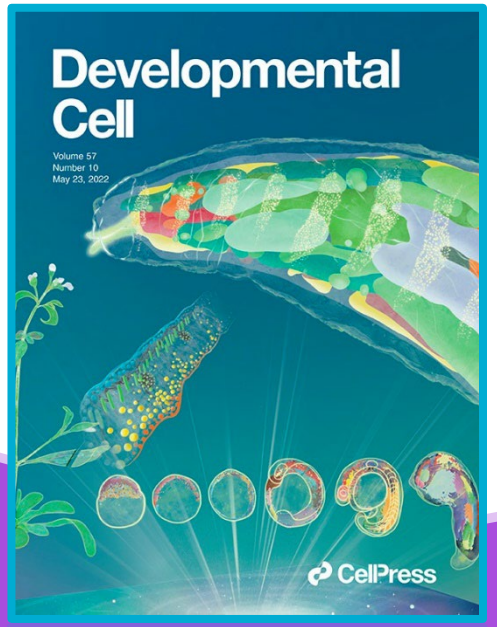
Mouse Organogenesis Spatiotemporal Transcriptomic Atlas

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

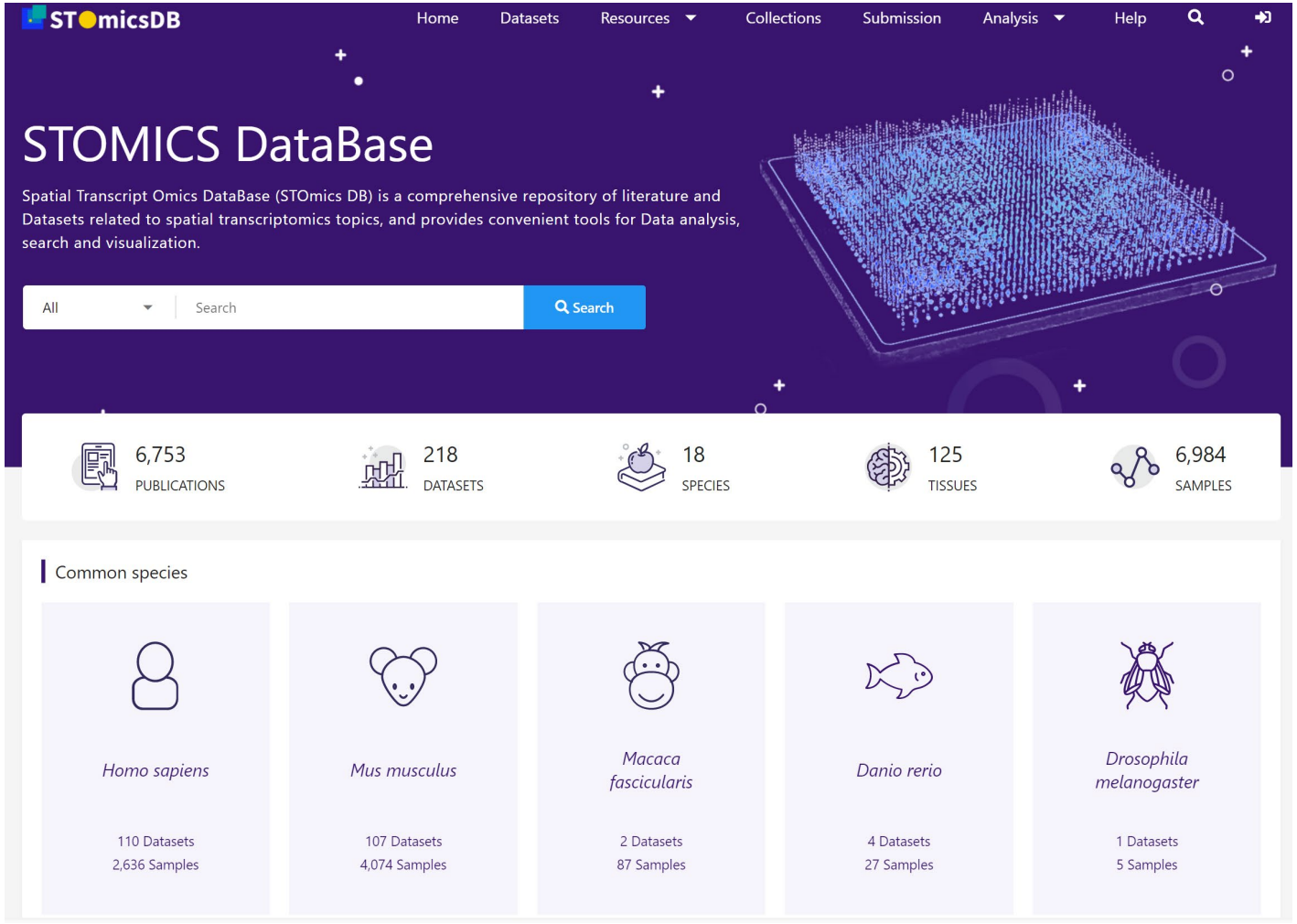


2022.5.5

Zebrafish/Drosophila Embryogenesis Spatiotemporal Transcriptomic Atlas

2. STOmicsDB

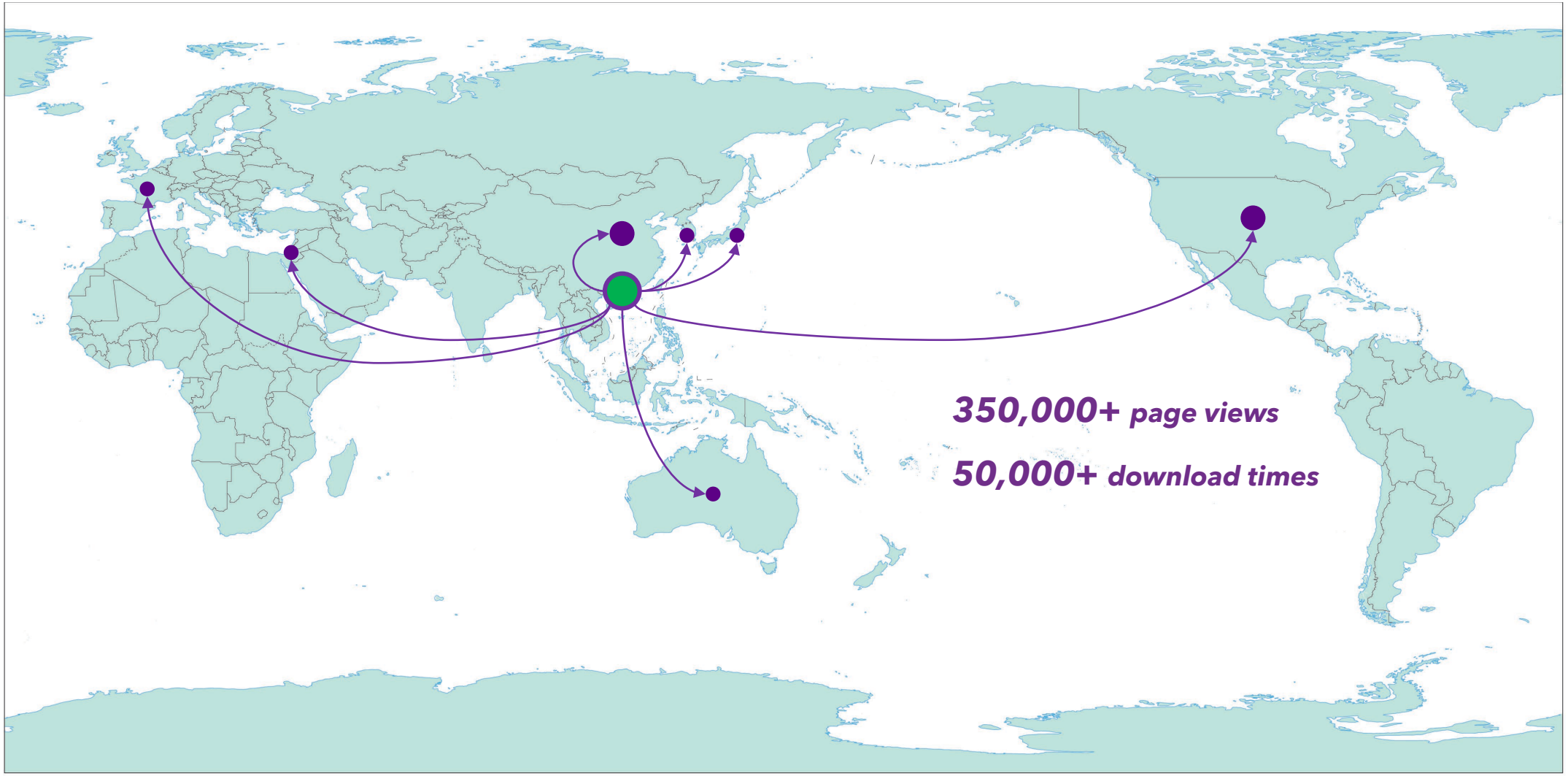
Spatial TranscriptOmics DataBase



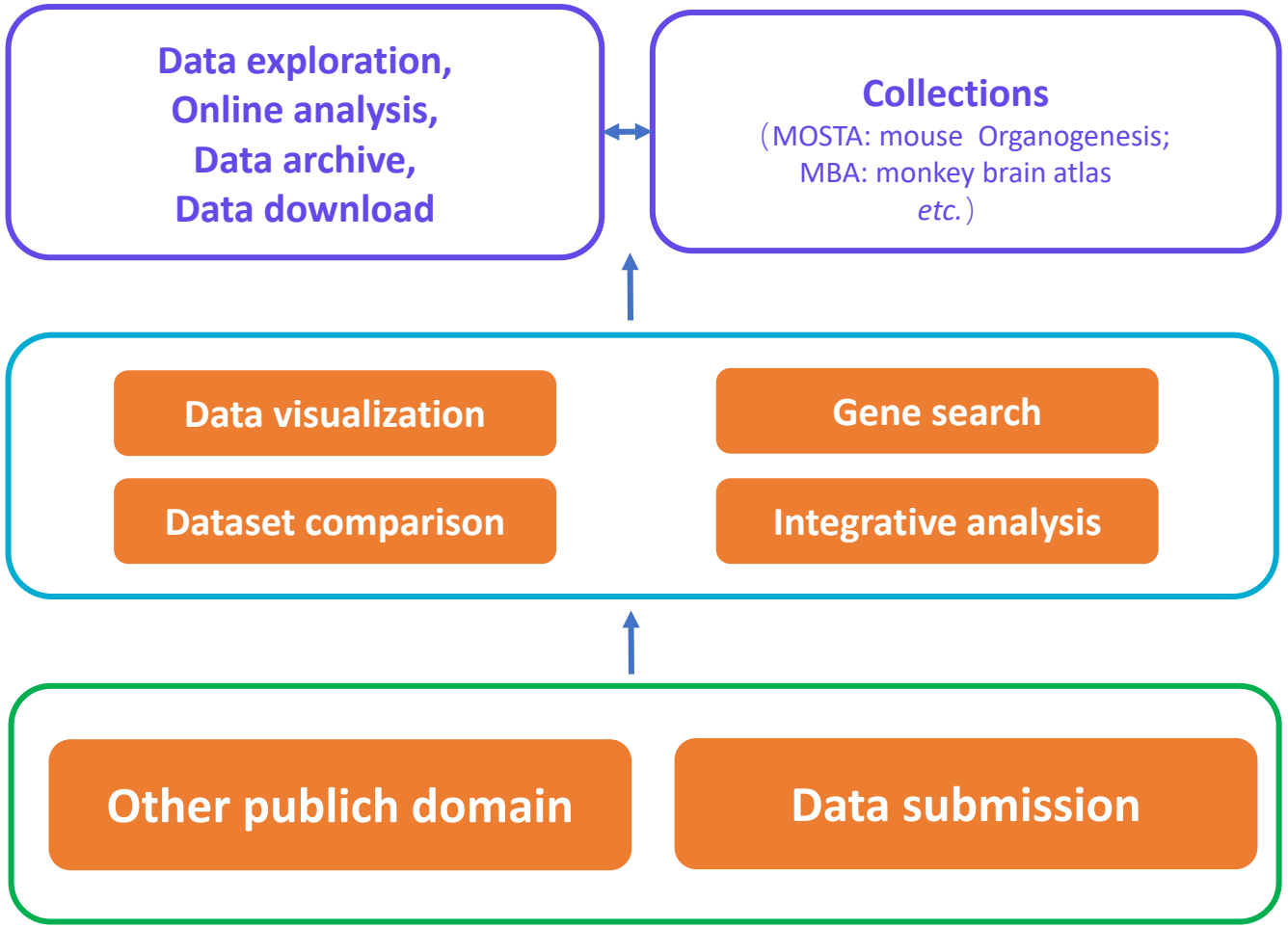
- ***Curated 200+ datasets***
- ***Spatial transcriptomic data exploration and visualization***
- ***Customized collections/databases***

db.cngb.org/stomics

Overview

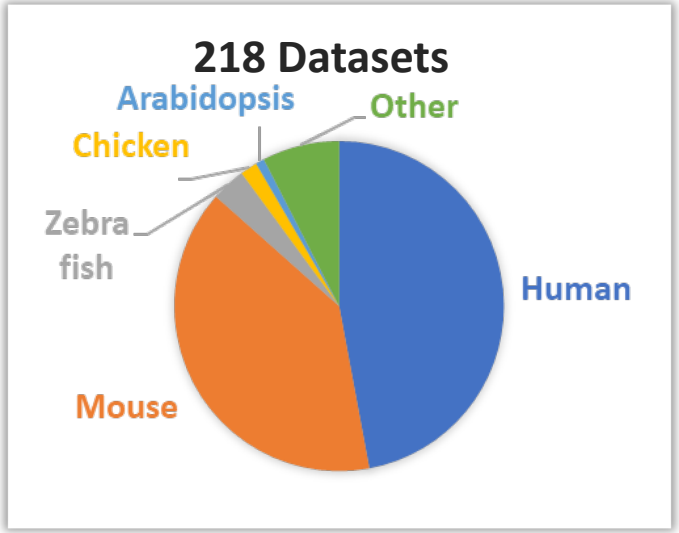


Structure


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 publich domain
Data submission

Data

- **Collected (218 datasets so far)**
  - from NCBI, EBI, DDBJ, etc.
  - from papers
  - from submission

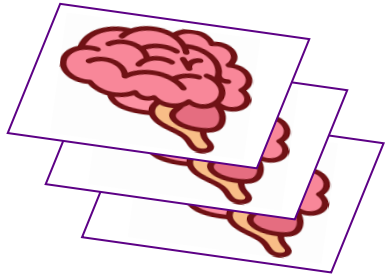


- **Curated**
  - display general information of each dataset (summary, overall design, species, cell types, development stage, sample number, section number, etc)

Data archiving system

Metadata

❶ Sample



Metadata & Image

❷ Tissue Section

```
@
CGAA ..... TTCAC
+
AAAAAAAAA99@::?A?
```

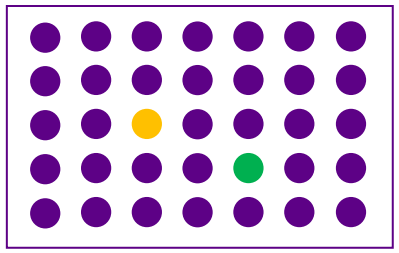
Fastq

❸ Sequencing reads

	Cell1	...	CellN
Gene1	3	...	4
Gene2	2	...	8
...			
GeneN	9	...	6

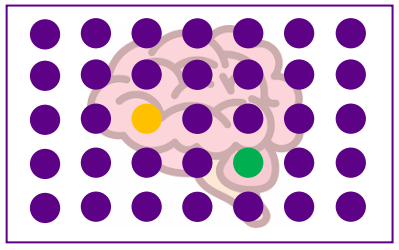
.tsv / .h5ad / .gem

❹ Gene expression data



.tsv / .h5ad / .gem

❺ Cell spatial position

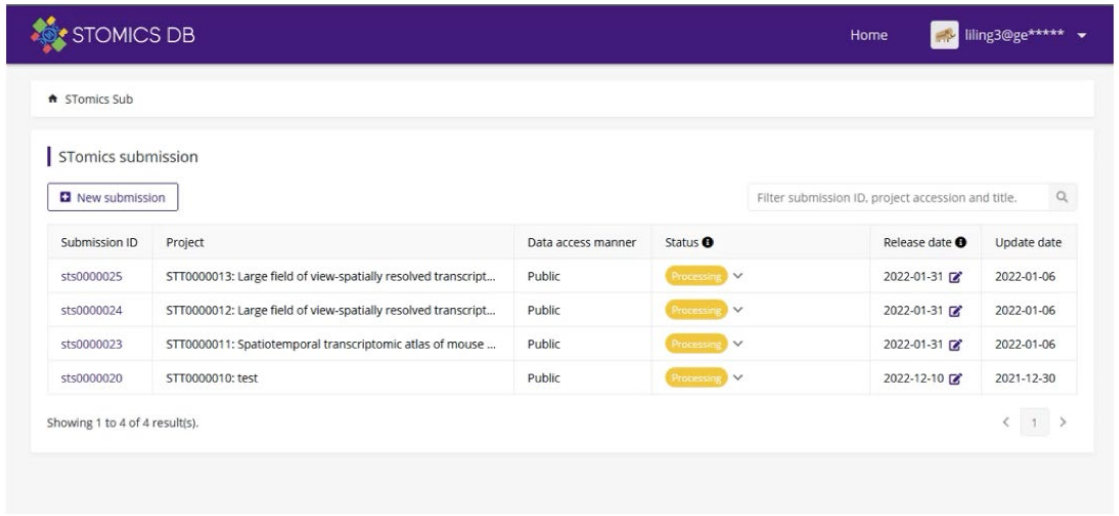


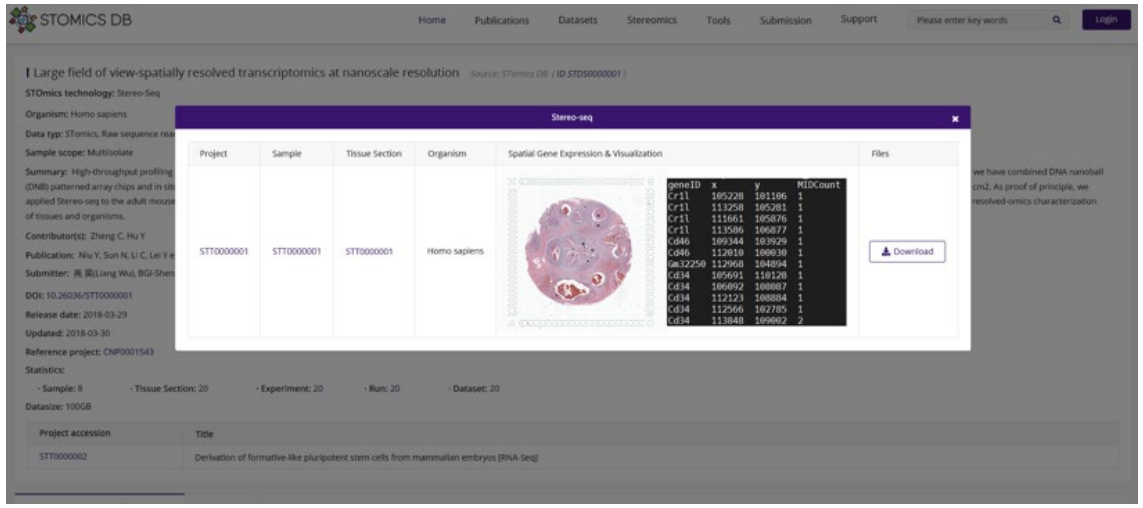
.tsv / .h5ad / .gem

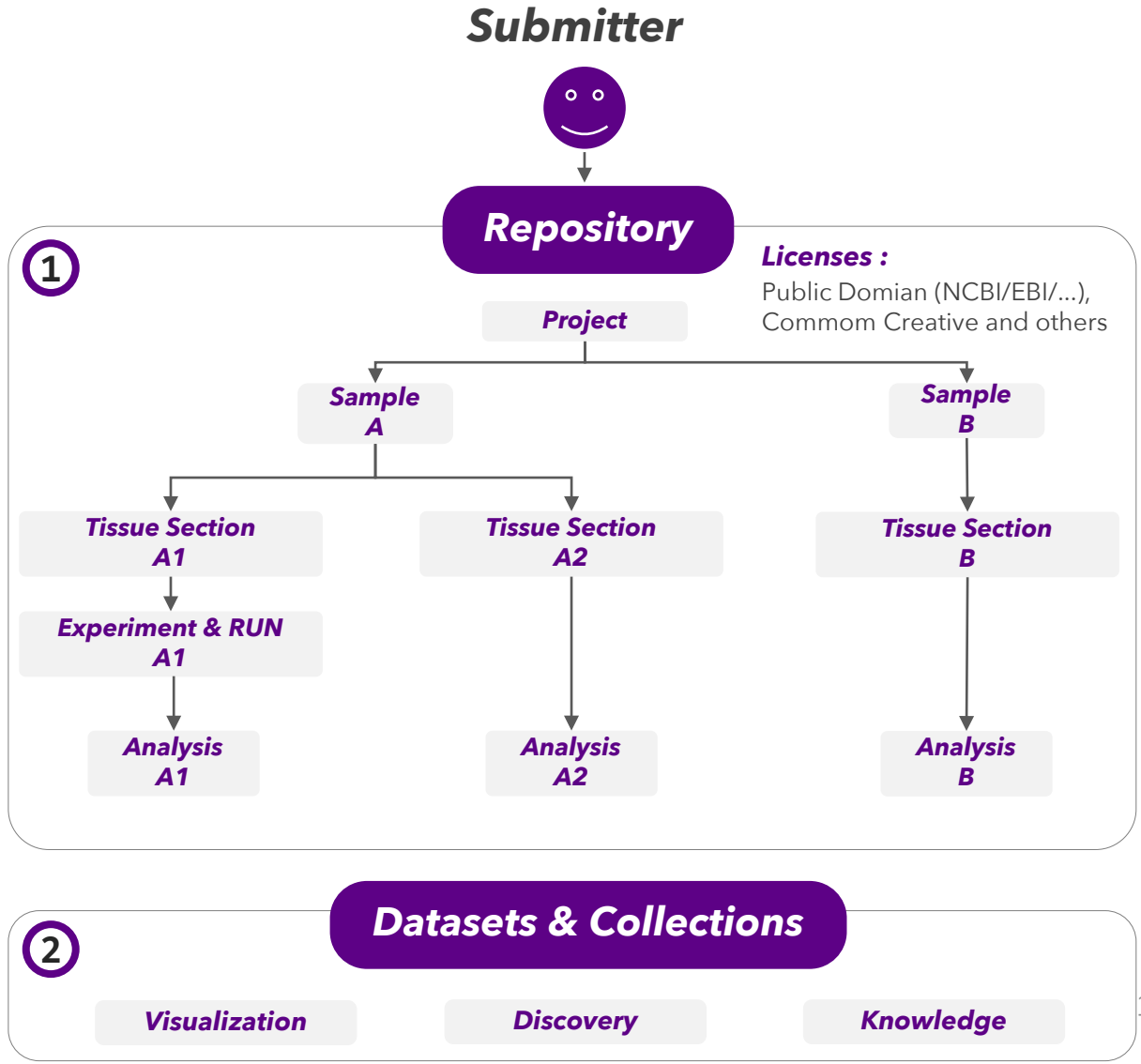
❻ Cell annotation

STOmicsDB

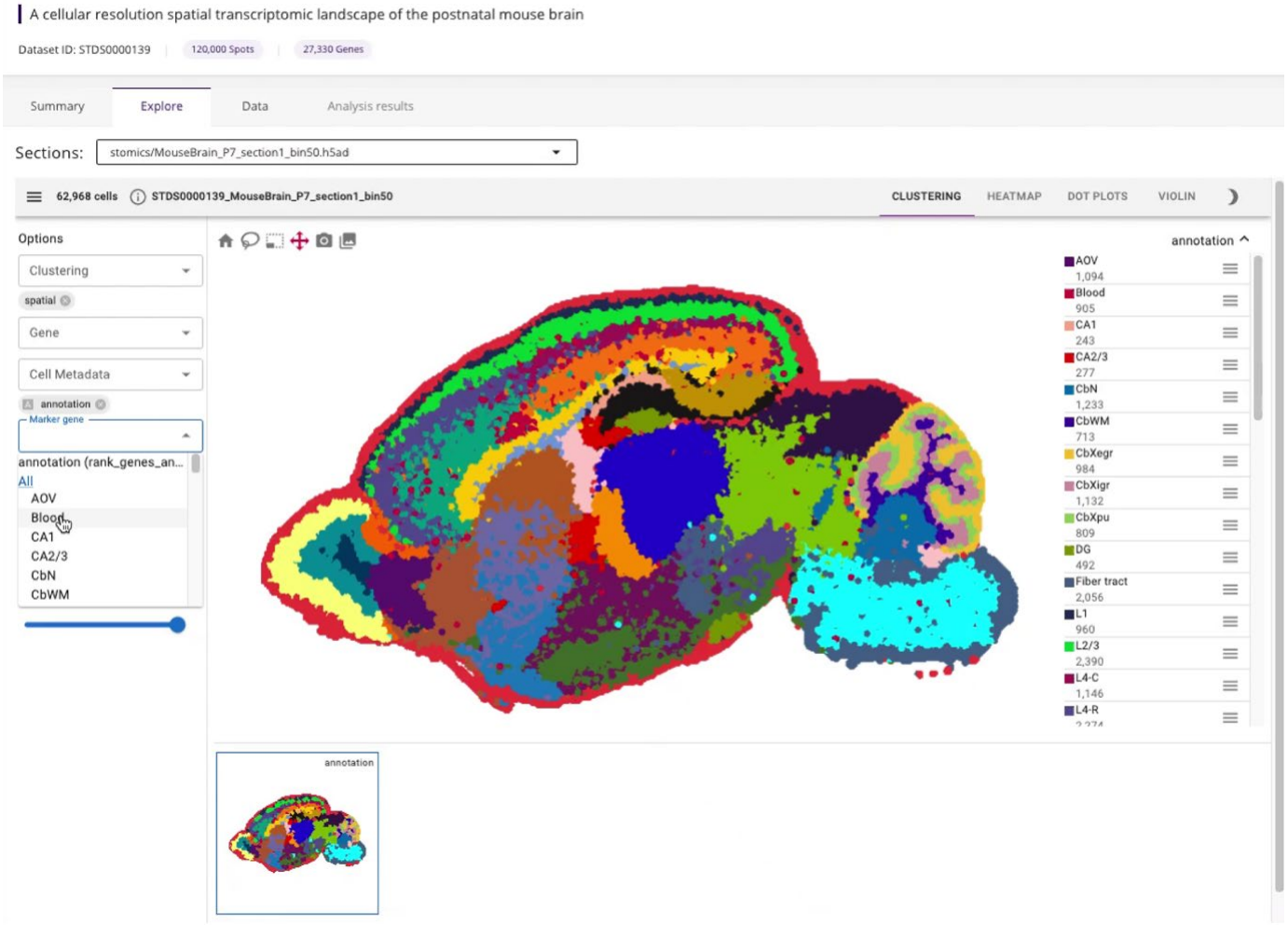
Data archiving system







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.

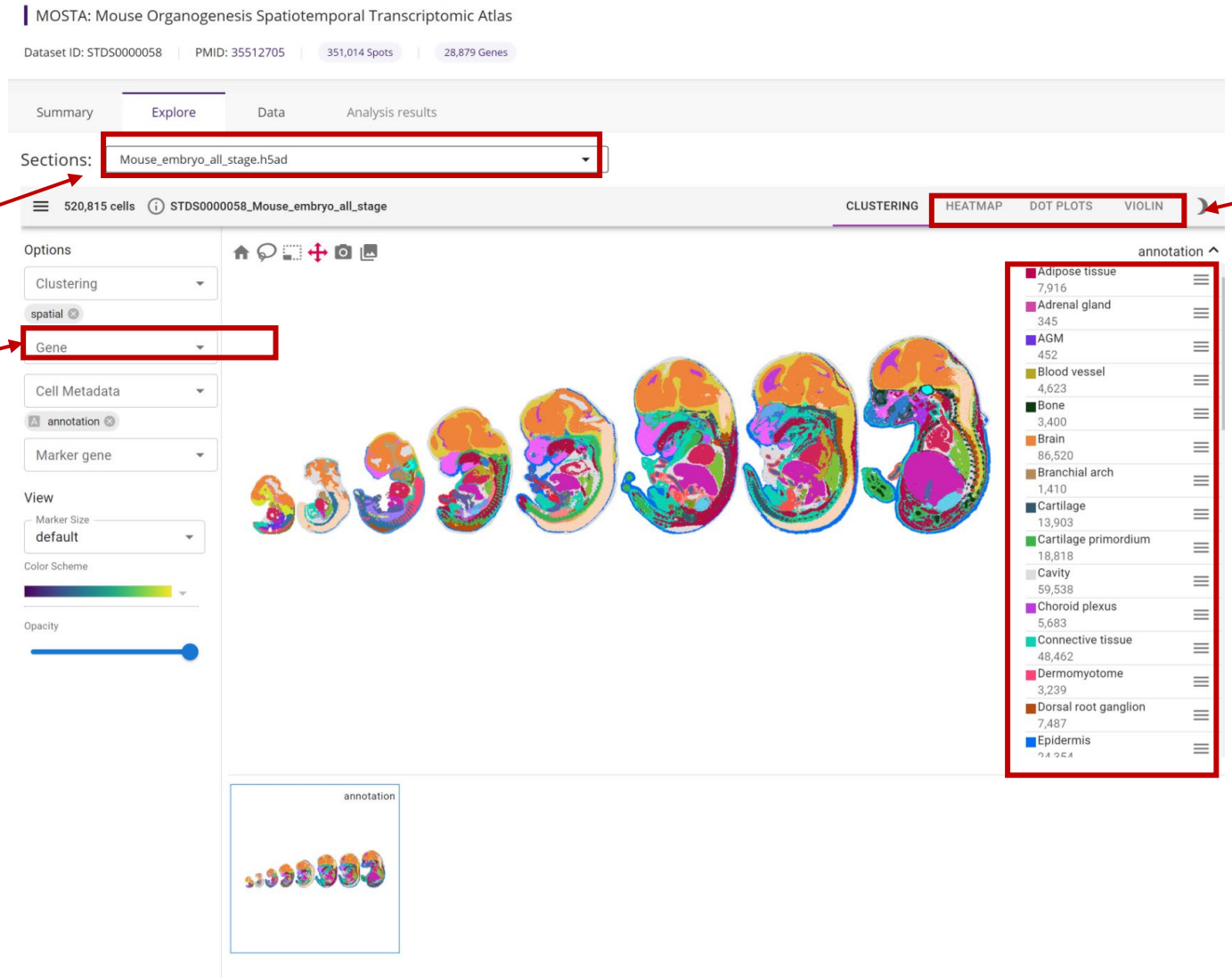
#2 Standardized named subcell type annotations.

2. Introduction

STOmicsDB

Visualization

https://db.cngb.org/stomics/, then search 'MOSTA' in the search bar



Select different sample

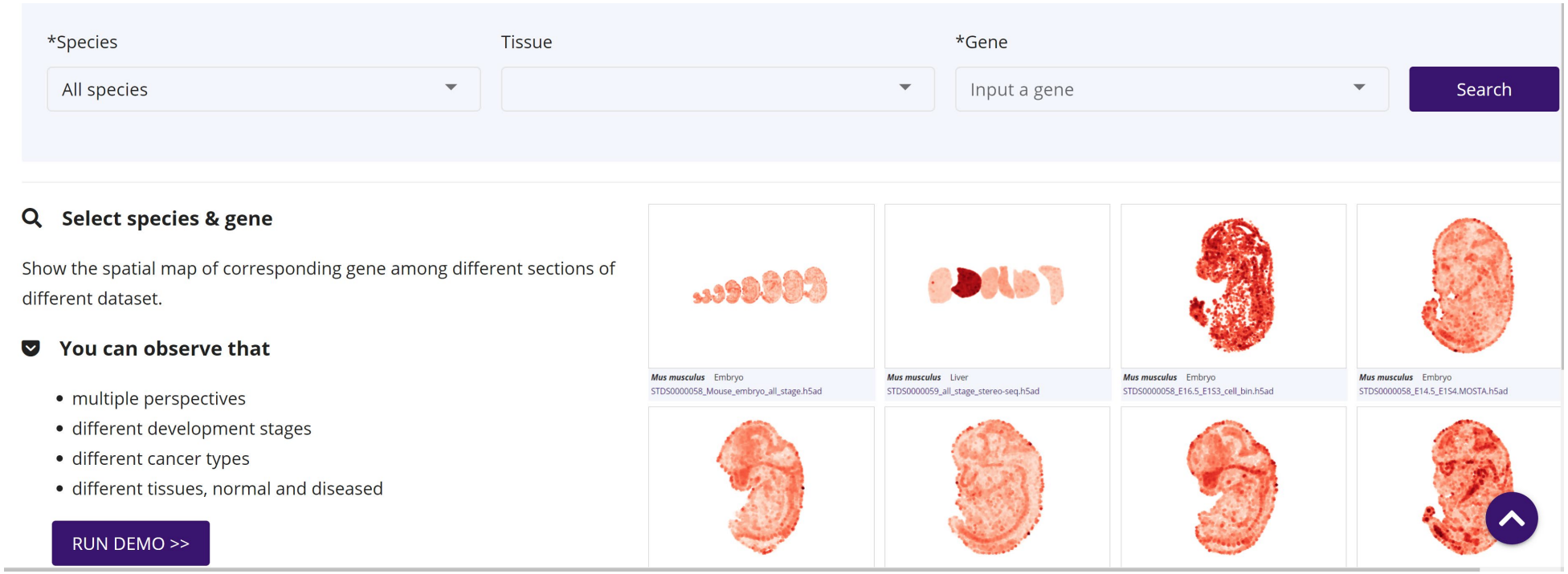
Select the gene of interest

Switch to the comparison of the gene expression module

Cell annotation

2. STOmicsDB

Gene search



6,000+

curated sample slides

This function shows the spatial expression of a specific gene.

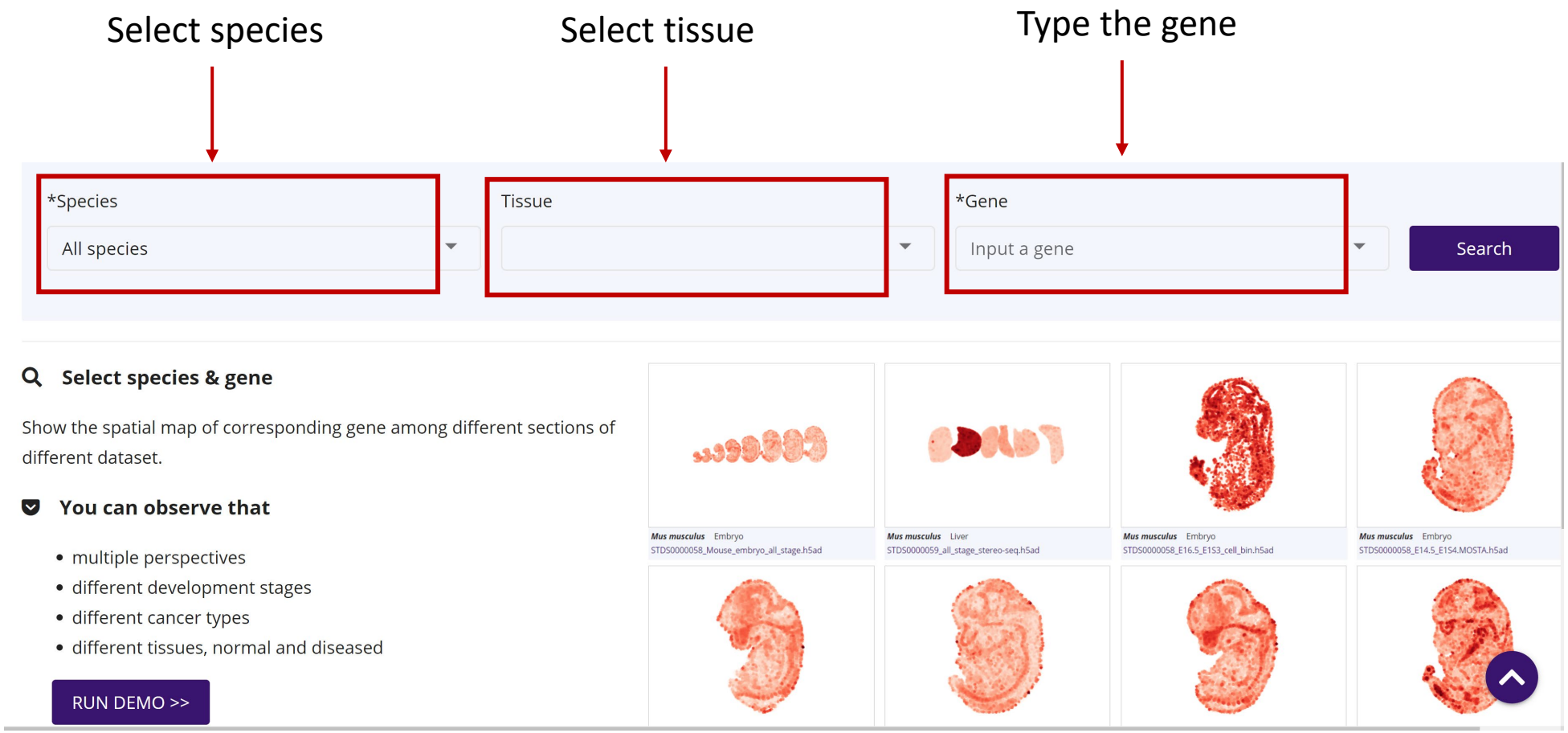
Allows users to search gene of interest among all datasets, and they can select the species or tissue to narrow down the results.

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



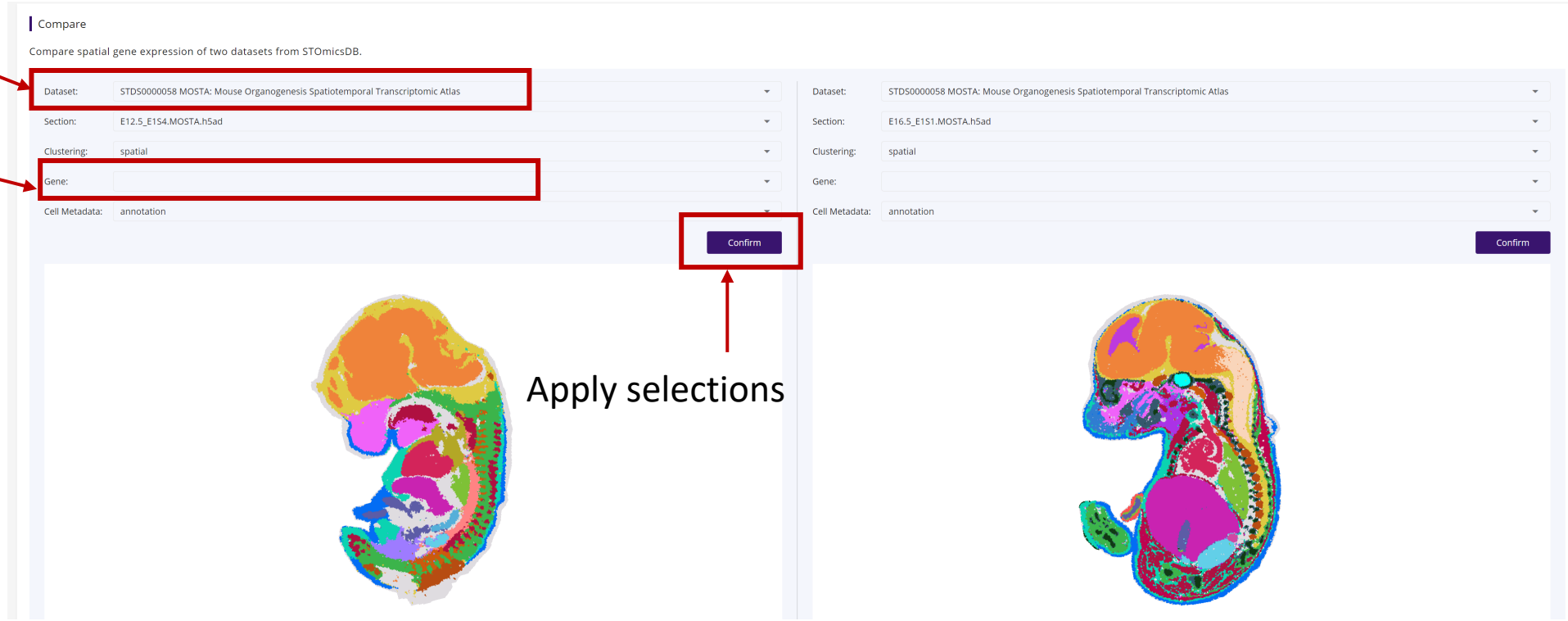
2. Introduction

Dataset comparison

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

Select dataset

Select gene



Apply selections

2. Introduction

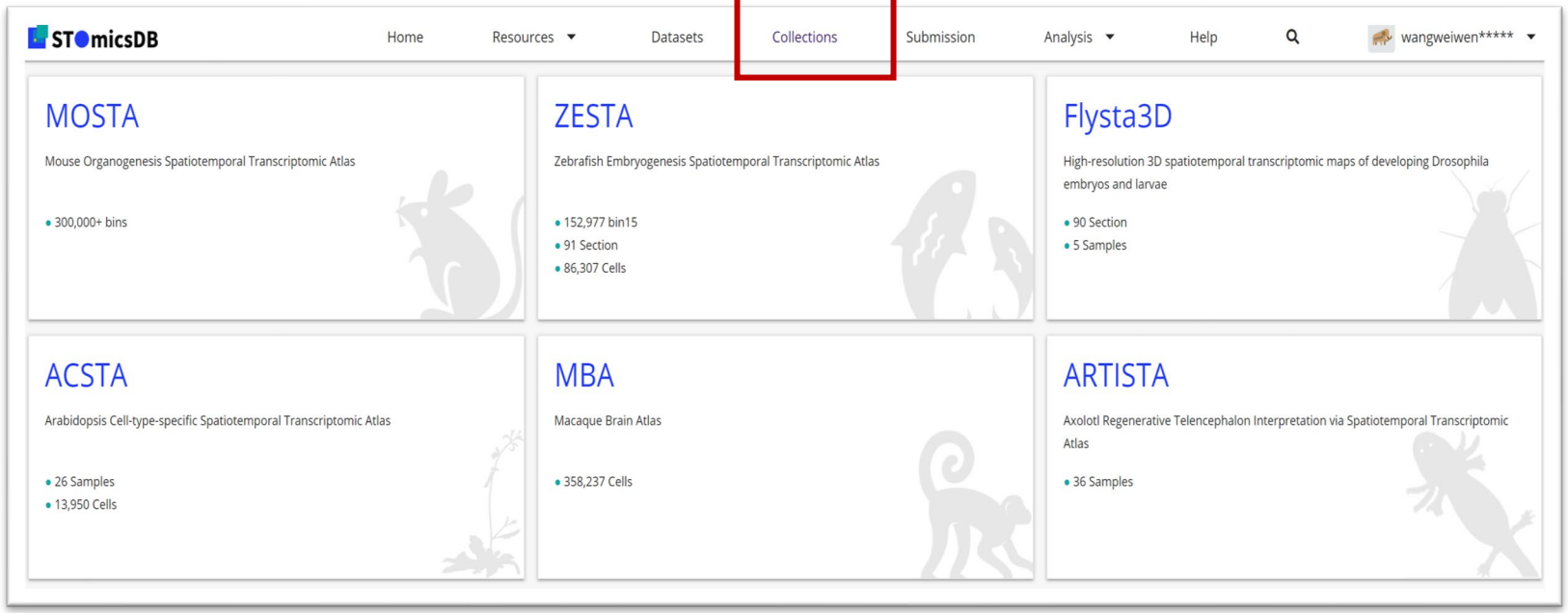
STomicsDB

Collections

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

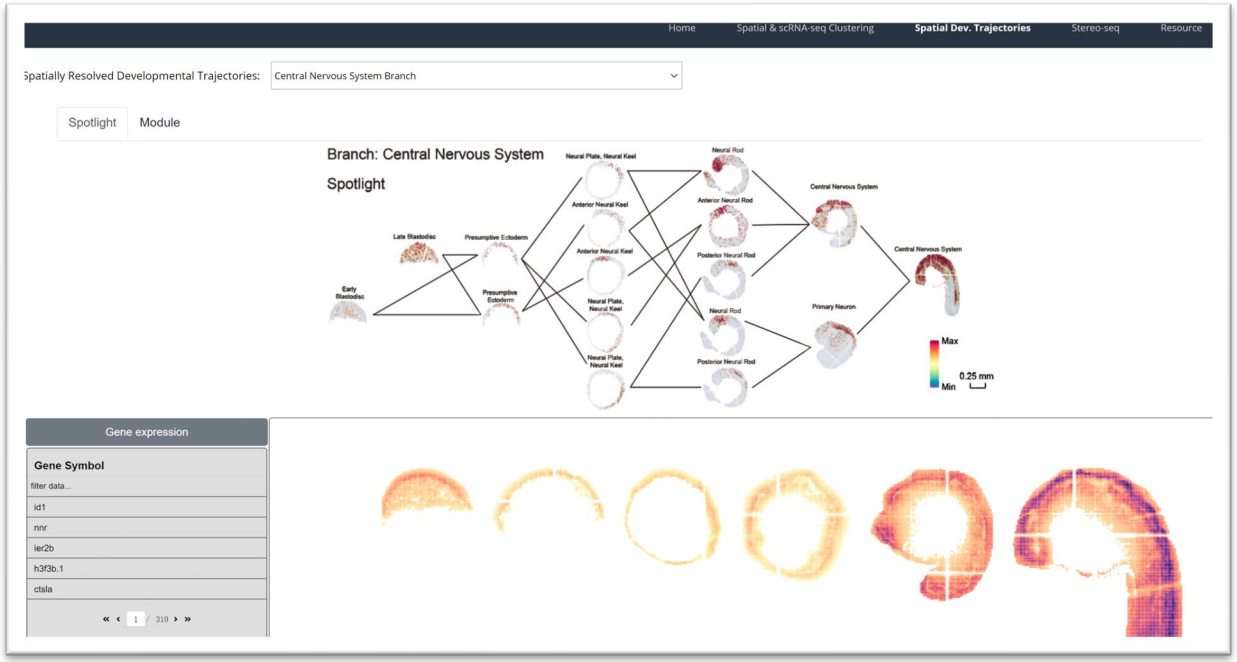
- Collaborated with 6 research groups so far
- Welcome collaboration

Collection entry

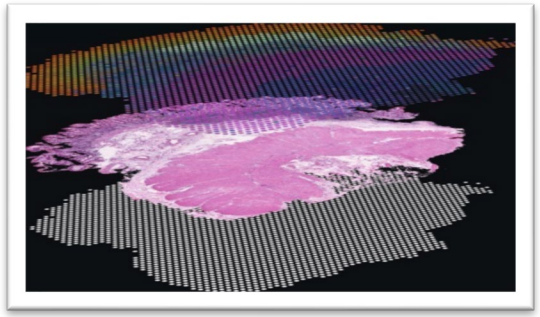


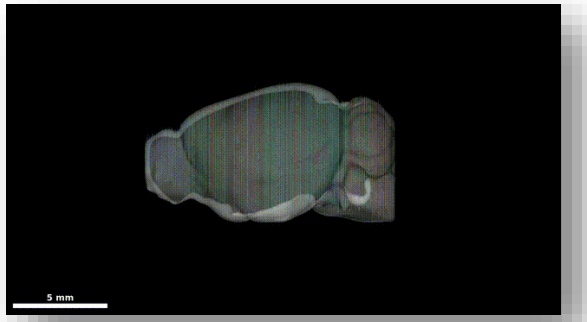
Collections

- Customizable visualization



Development Trajectory of Zebrafish



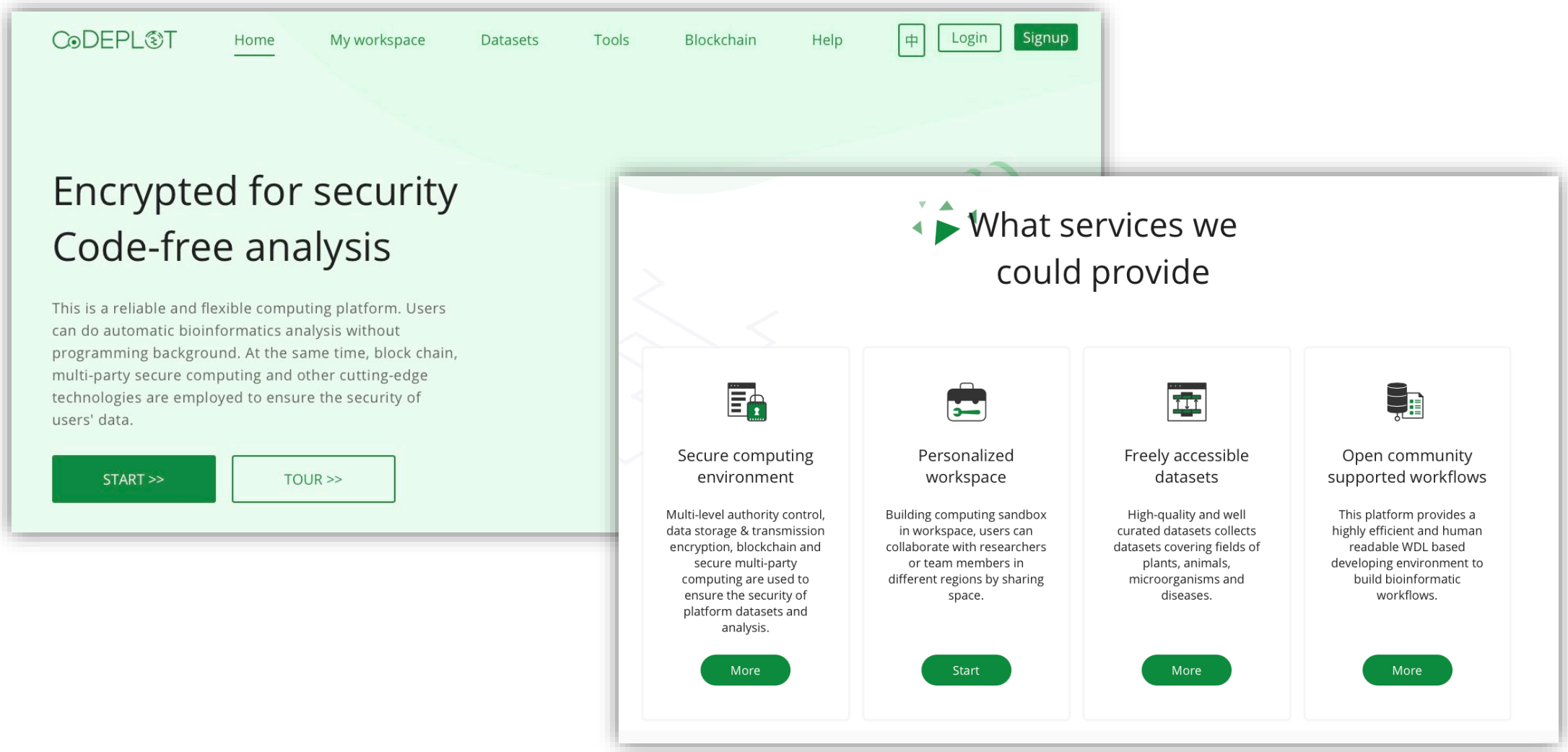


3D model of *Drosophila*

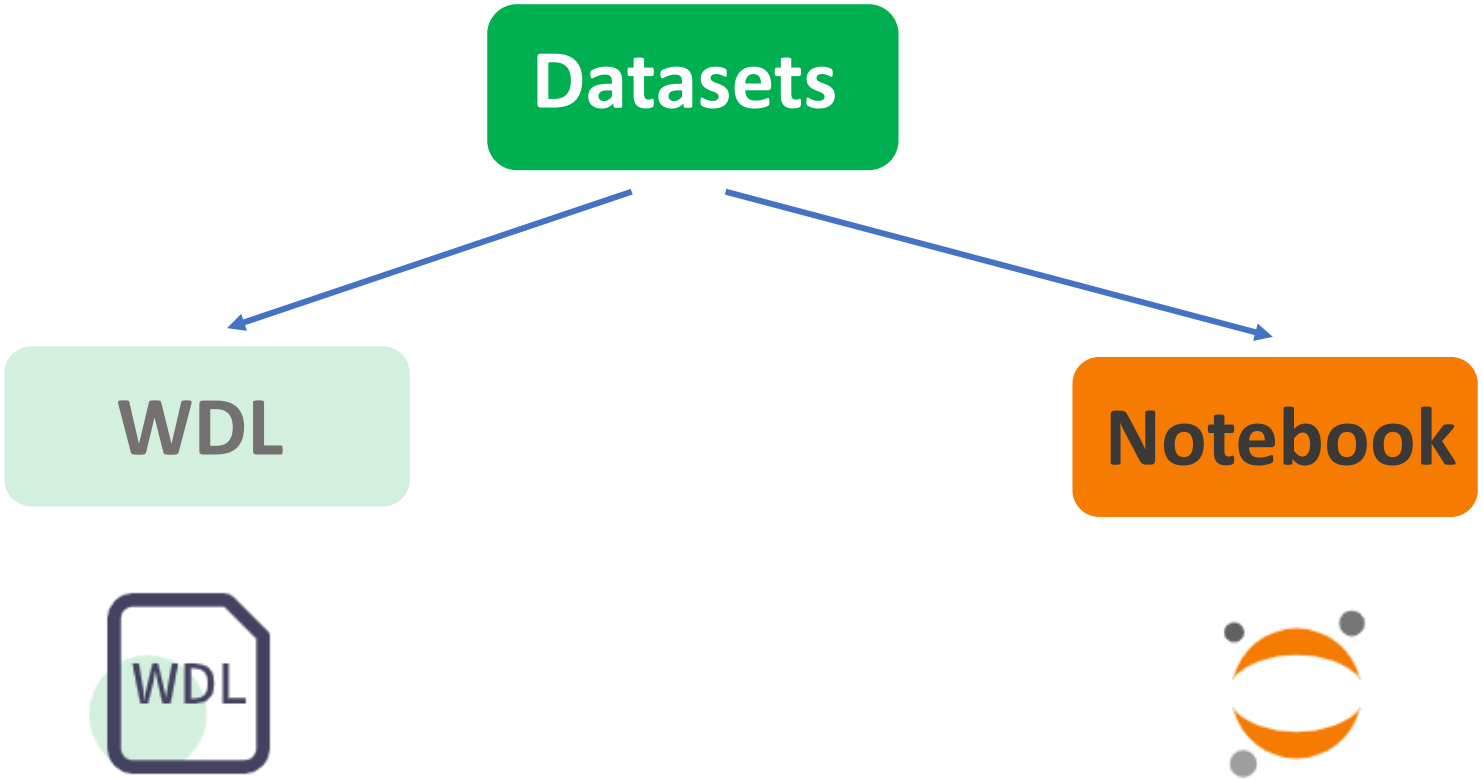
1
STOmicsDB


CoDEPLOT
2

Codeplot: a platform for code-free analyses



Function



- **Zero-code** | *streamline batch analysis*

  Based on standardized WDL language

  Customize tuning parameters

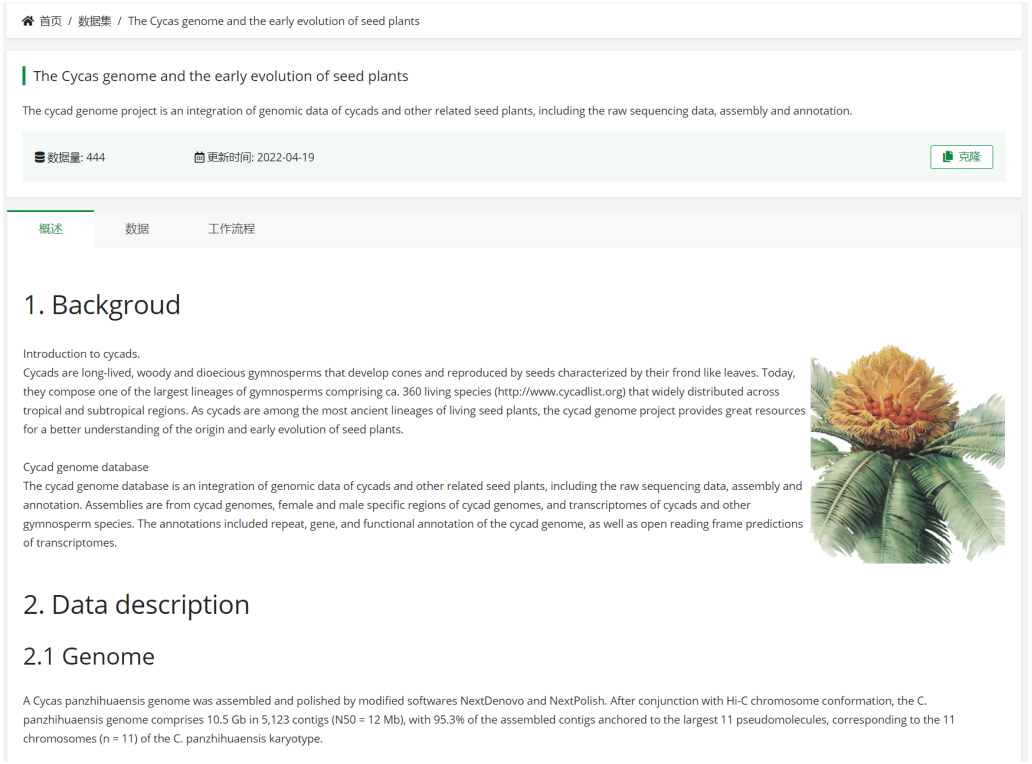
- **Low-code** | *costume analysis with notebook*

  The Jupyter notebook is deployed to provide

  Python, R and other packages

Datasets

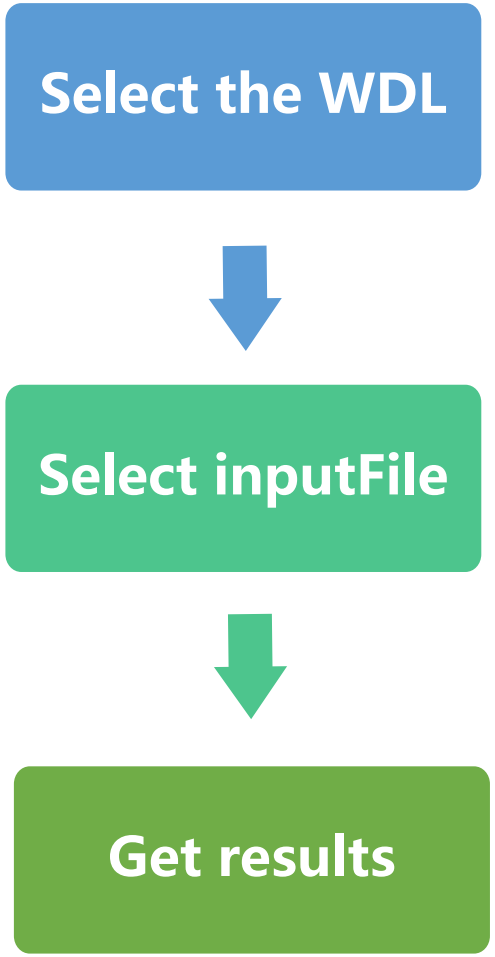
- **Curated datasets (21 datasets so far)**
  - Ensemble plant datasets (96 plant genomes)
  - COVID-19 datasets (~10 million seqs)
  - Single-cell datasets (21 species)
- **User-own datasets**
- **Publication support**



WDL: enter and get results

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

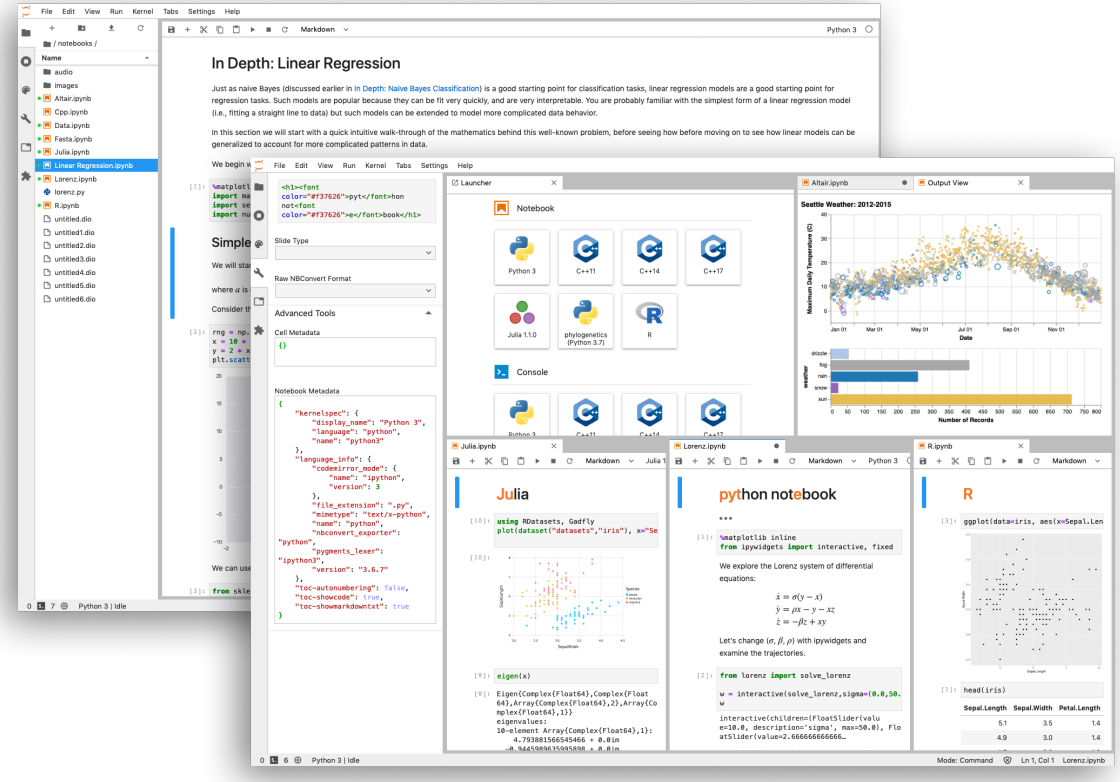
- **30+ WDLs**
  - **Spatial transcriptomics analysis**
  - **Single-cell analysis**
  - **GWAS**
  - **Others**
- **Self-defined WDL**



Notebook

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

- 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.**