China National GeneBank 国家基因库

华大生命科学研究院 BGI·Research

Biodata & AI

"The Role of CNGB in Advancing Life Sciences"

Xiaofeng(Vincent) WEI

魏晓锋

2023.9

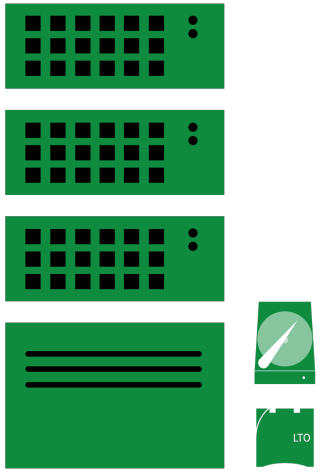
China National GeneBank

国家基因库

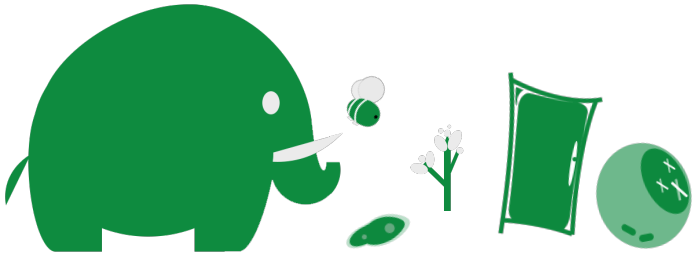
- **Started from 2011**
  Approved and funded by the Chinese government
- **Launched in SEP 2016**
  "Owned by All, Completed by All and Shared by All"
- **Effective bioresource conservation, digitalization and utilization.**

CNGB

Digitalization Platform



Bio-informatics Data Center



Biorepository

A Leading Comprehensive Biorepository

- Capable of storing tens of millions of samples
- Automated, high-throughput and cost-efficient
- Covering samples of plants, animals, microbes and human beings

An Automated, Cost-Efficient and High-Throughput Biorepository

Informatized Management

- Whole-process and informatized sample management system
- Covering the whole process from sample collection, transportation, preprocessing, inventory management, to sharing and application

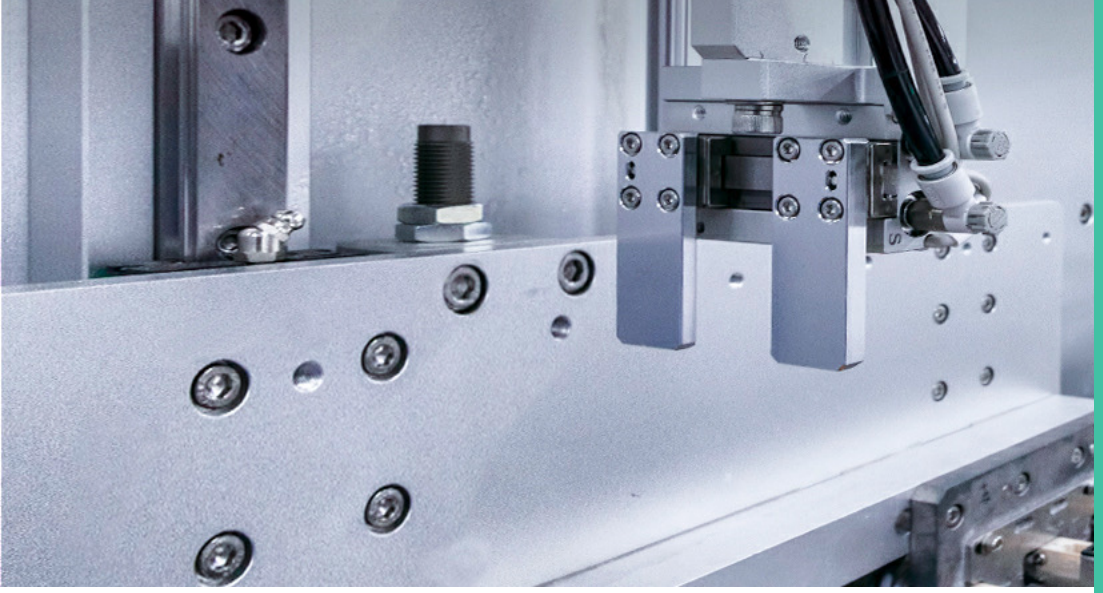
Biological Resource Center

- Integrating genetic resources and bio-information
- An open platform to conserve and digitalize living bio-resources
- In strict compliance with Chinese laws and regulations, and international conventions

A Digitalized Biodiversity Conservation Base and Biological Resource Center

Digitalized Biodiversity Base

- The world's first digitalized botanical garden – Ruili Botanical Garden



A Leading "Reading" Platform Producing Petabases of Data Annually

High Data Output Capacity

- Petabase-level annual output capacity
- Annual processing capacity: 500,000 samples
- 24/7/365 operations

Highly Informatized and Automated

- Throughout the whole process, from library preparation to data output
- Efficient, speedy, and traceable

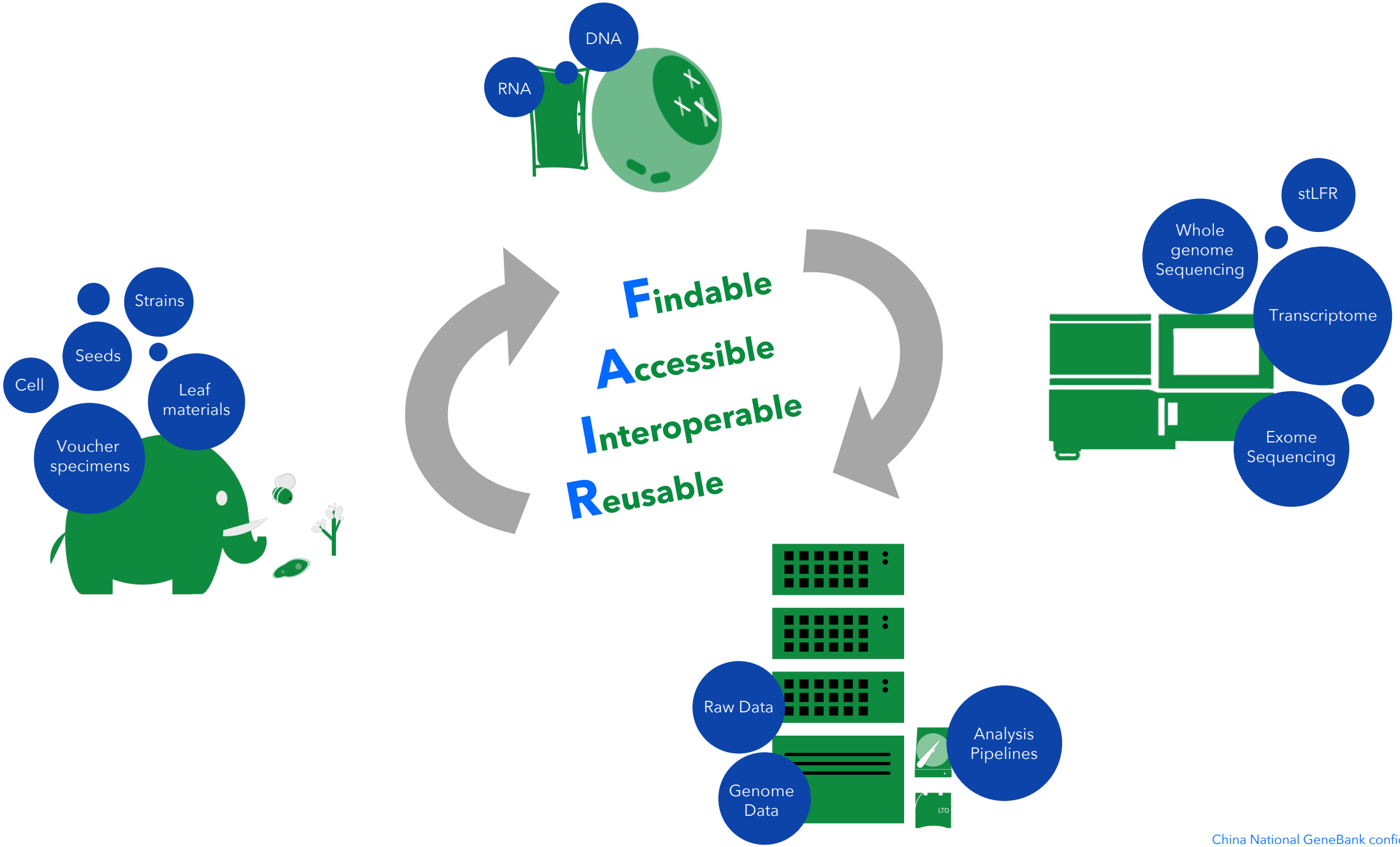
A Secured Database and Highly Efficient Bio-Informatics Analysis Platform

High Performance Computing

- Storage throughput: >150 GB/sec
- Computational capacity: 691 teraflops

Secured and Stabilized

- ISO/IEC 27001 certification for information security
- 24/7/365 operations with high stability and zero major security failure
- Energy-efficient data center



A platform is needed to facilitate resource sharing

Application Services

Analysis Pipelines

Data Management System

Laboratory information management system

Sample Tracking & Management System

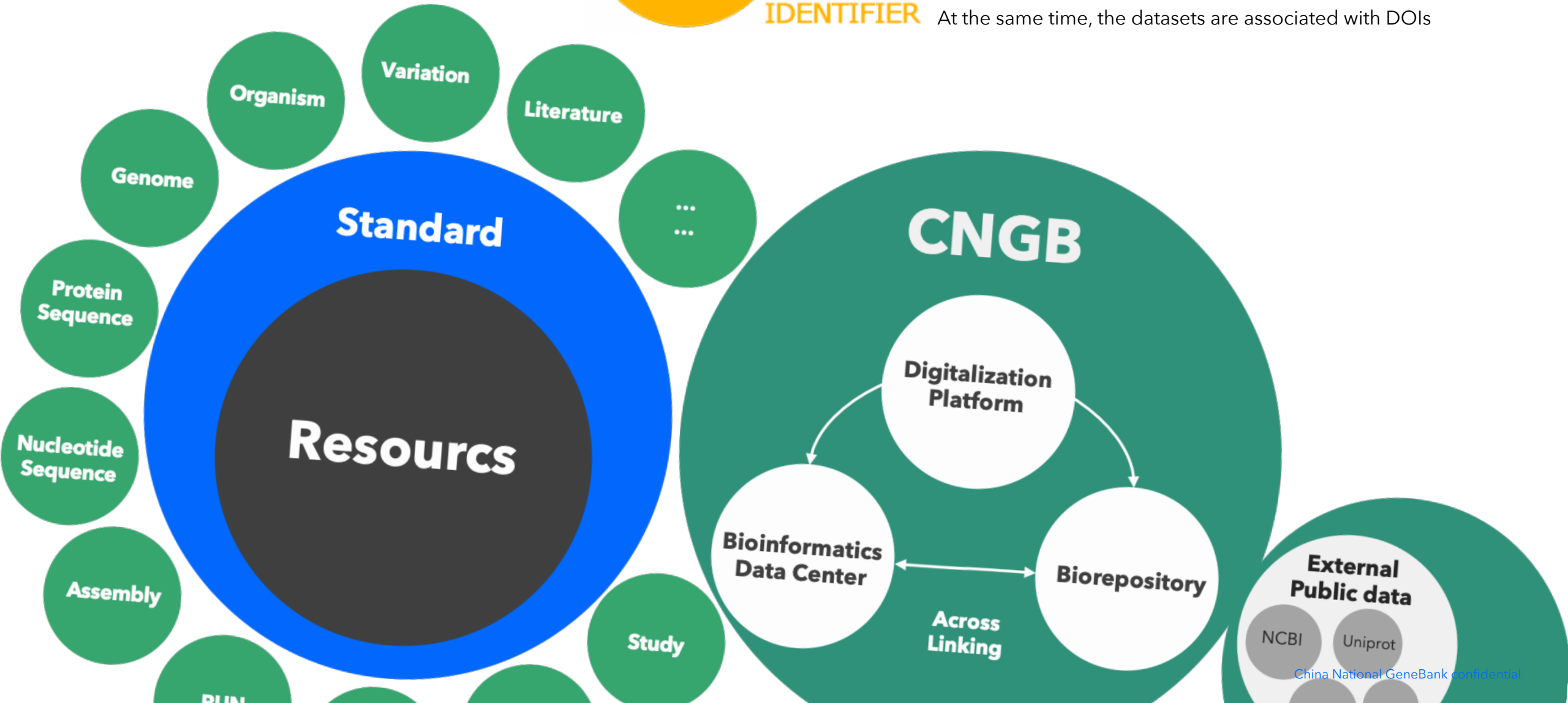
Project Management System

DIGITAL
OBJECT
IDENTIFIER

Core structure is based on INSDC

Combined with GGBN's and GA4GH's

At the same time, the datasets are associated with DOIs



CNSA
CNGB Sequence Archive

All | Explore science, technology & innovation... | Search

Search examples: CNP0001543; Momordica charantia; BGISEQ-500

CNGBdb has been certificated by FAIRsharing, and is included in re3data and OpenDOAR. CNGBdb has been recognized by many journals/publishers, such as Elsevier, Cell Press, Science, Wiley, Taylor & Francis, Oxford, etc. CNGBdb has been designated as a supported data repository for the Earth BioGenome Project (EBP). To promote data sharing, CNGBdb uniformly assigns a DOI (Digital Object Identifier) for each submitted project.

11902TB	1288TB	4601	843690	706416	828107
Total data	Public data	Project	Sample	Experiment	Run

One of the largest global archives for raw gene sequencing data.

4,600 projects
11.9 PB

480 institutions, covering **1,260** articles and **230** journals

Recommended by 21 international presses and journals, including the world's top five academic publishing groups such as **Elsevier** and **Wiley**

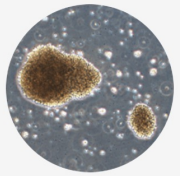
619,000 samples

- **Animals:** **Nearly 1000 cell lines from 38 animal species**, including immortalized B lymphocytes, fibroblasts, mesenchymal stem cells and renal epithelial cells
- **Plant:** Seeds, voucher specimens, and molecular material from more than **100 plant species**
- **Microorganisms:** **More than 2000 strains** were isolated from human symbiosis, pig intestine and environment

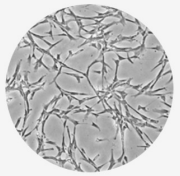
EBB (E-Biobank) + BRC-PAM Biological Resource Center of Plants, Animals and Microorganisms

Animal Cell Bank

Animal Cell Bank has been committed to the establishment and cryopreservation of different types of cell cultures from animal species. Multiple techniques have been employed to ensure the high-quality and availability of the cell cultures.



Mammalian cells



Avian cells

Reptilian cells

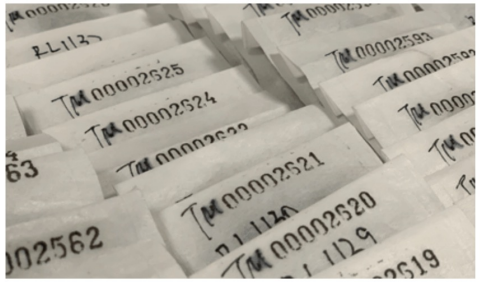
Amphibian cells

Bank of Plant Resource

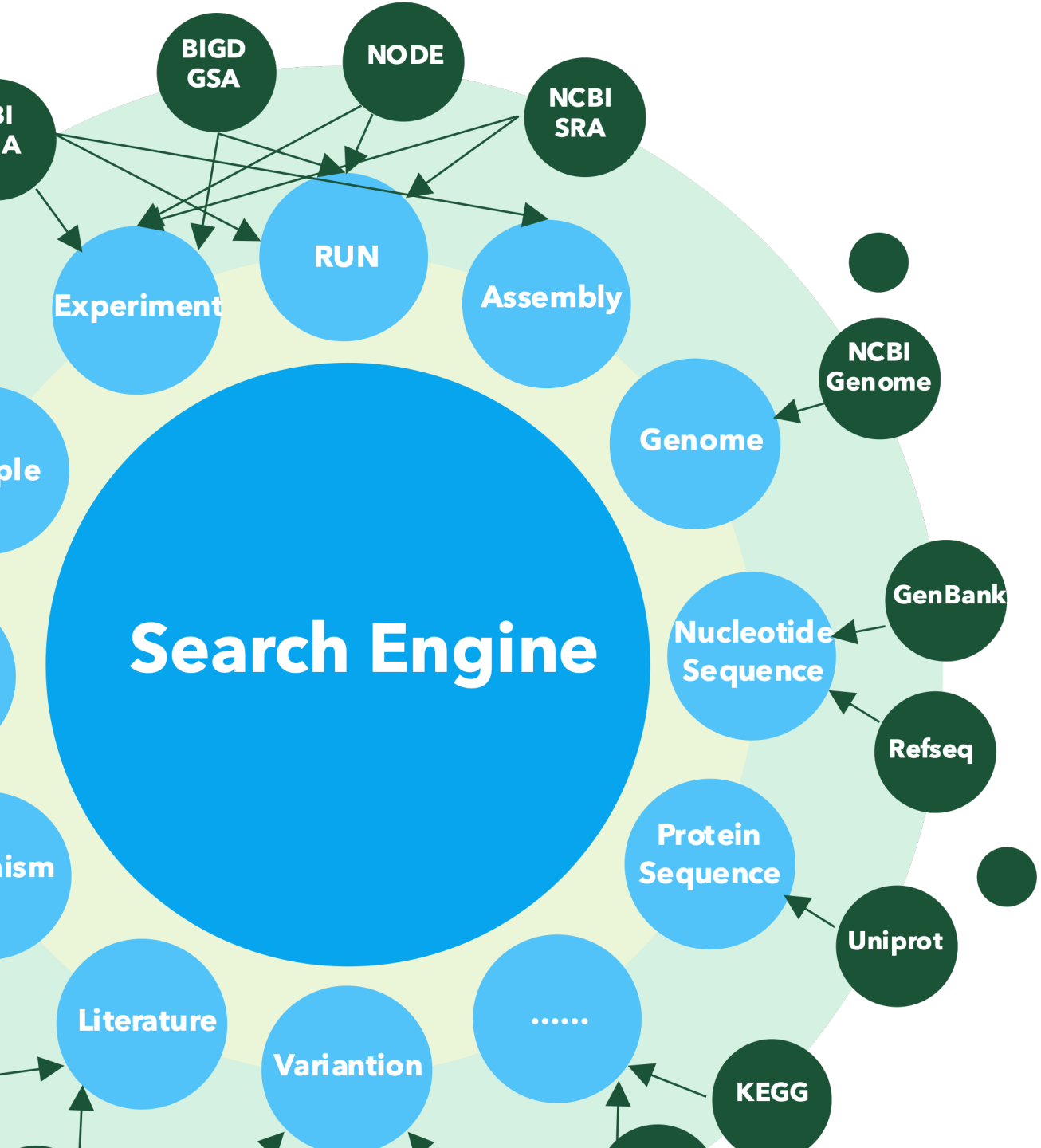
Bank of Plant Resource aims to build an open-sharing platform to conserve and distribute the plant resources that have been integrated with bio-information. Various plant samples such as seeds and voucher specimens and associated genomic data are now accessible to researchers.

Seed

Voucher specimen



Tissue materials



6 Billion + index (Elasticsearch)

10TB+ metadata

7 billion + triples (Knowledge graph | Neo4j)

- Aggregate large amounts of molecular data and associated information from CNGB, NCBI, EBI, DDBJ and other platforms.
- Data are interconnected and indexed by the search engine for efficient retrieval.
- The data are also correlated with bio-samples and even living organisms to ensure full lifecycle traceability.

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 >>

Secure computing environment, flexible computing resources, multi-field research data and tools.

If you lack programming skills, also be able to do bioinformatics analysis and collaborative collaborative computing.

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

Secure computing environment, flexible computing resources, multi-field research data and tools.

If you lack programming skills, also be able to do bioinformatics analysis and collaborative collaborative computing.

Intelligence

11.9 PB

Data Files

6 Billion

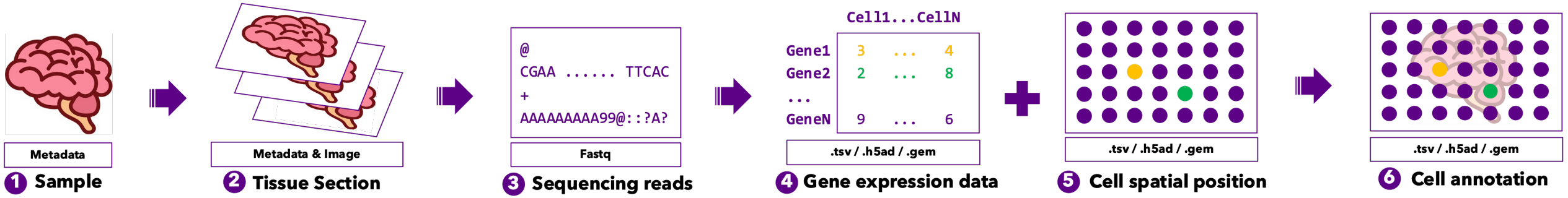
Information

? Knowledge

One of the largest database for Spatial Transcriptomics

1,377,720 Request

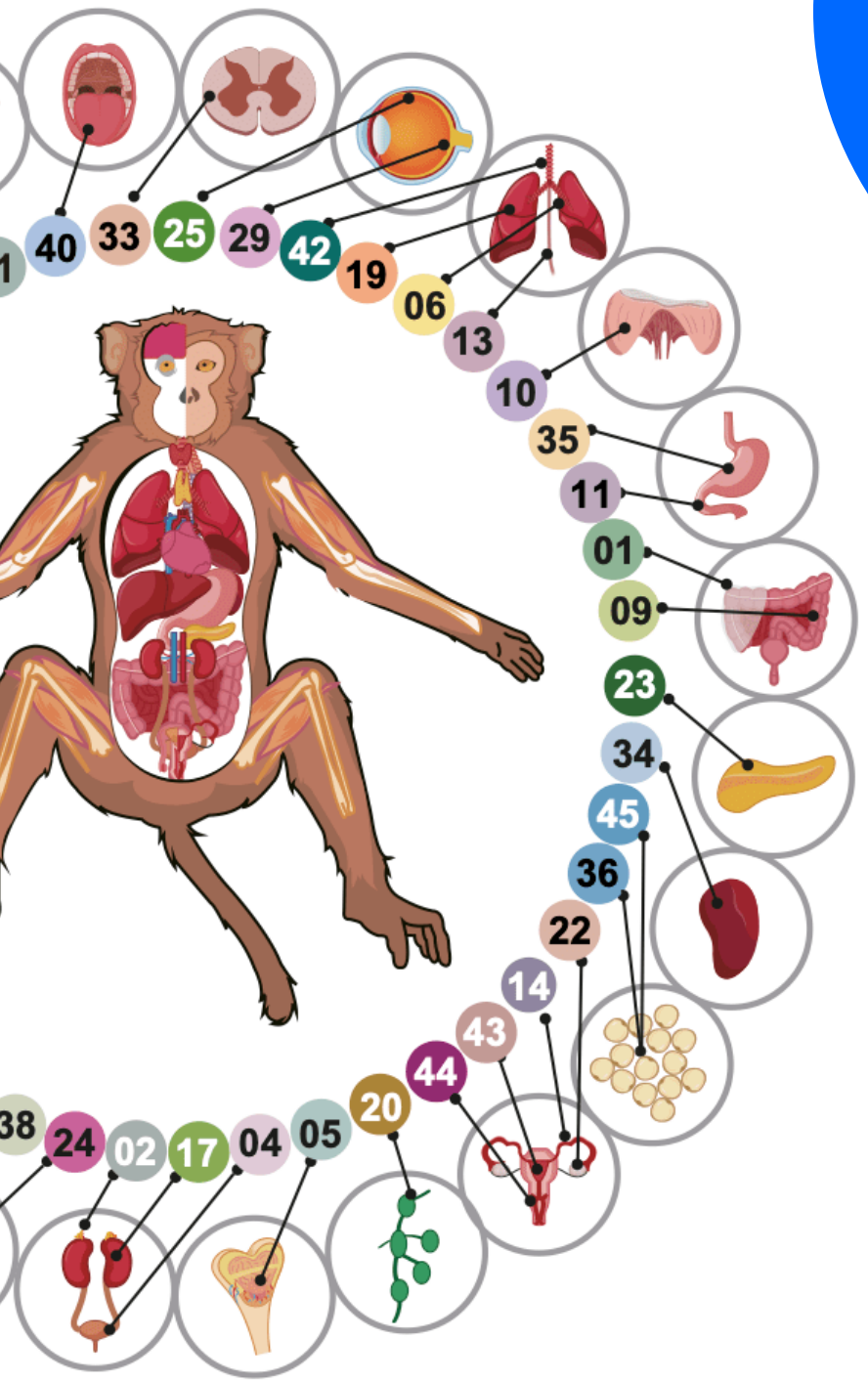
6,753	218	18	125	6,984
PUBLICATIONS	DATASETS	SPECIES	TISSUES	SAMPLES



Stereomap

- A ultra-high-definition visualization software designed for viewing the spatial omics data
- Capable of displaying over a million cells in a large field of view with ultra-high resolution
- Integrates a variety of tools for further exploration and mining of the data.

This lays the foundation for researchers to understand spatial cell types, annotations, locations, and development.



Non-Human Primate Cell Atlas

1.14 million cells

from 45 organs of adult macaques

LettuceDB

Home Omics Search JBrowse Tools Download About

Lettuce Database

An integrated multi-omics database for cultivated lettuce

1. Germplasm

Collection of representative ***Lactuca*** accessions worldwide

2. Genome

Genome assemblies and annotations of lettuce

3. Genome

Sequence variations including SNPs, indels, and SVs

4. Phenome

Agronomic traits of cultivated and wild lettuce

5. Microbiome

Microbial taxa from rhizosphere soil samples

6. Spatial Omics

Single-cell and spatial transcriptome of lettuce tissue

A new knowledge system to digitize a species from "6 Dimensions"

Intelligence

11.9 PB

Data Files

6 Billion

Information

? Knowledge

** created by AI

** created by photographer

** created by AI