Introduction to
Korea BioData Station (K-BDS)

kobic — Korea Bioinformation Center

KRIBB — Korea Research Institute of Bioscience & Biotechnology

Seungwoo Hwang

2023. 10. 11.
The 20th ABC Symposium

2023. 10. 12.
The 32nd Int'l KOGO Annual Conference: Session 2

Bio-data Management Policy in Korea

☐ **Data management policy in Korea is getting stricter**

- Funding agencies are reinforcing their data management policies.

- More and more grants are requiring data submission upon the completion of grants.

- KOBIC was designated as the data center for such submissions.

☐ **Enforcement Decree of the National Research And Development Innovation Act** (2021)

- Researchers who receive government R&D funding should submit their R&D outcomes to designated centers.

  Could be many different things: Data, Softwares, Biomaterials, Chemicals, Articles, Patents, etc.

☐ **Act on the Promotion of Management and Utilization of National Research Data** (**NEW** September 2023, Pre-Announcement of Legislation)

- Instead of individual researchers, their institutions should be responsible for management of research data of all their researchers.

- **Responsibility of data management:** Individual researchers → Institutional level

- Because, in practice, it is harder to hold individual researchers responsible.

A Test of 'Data are available on request'

Availability of data and materials

From a 2023 paper

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.
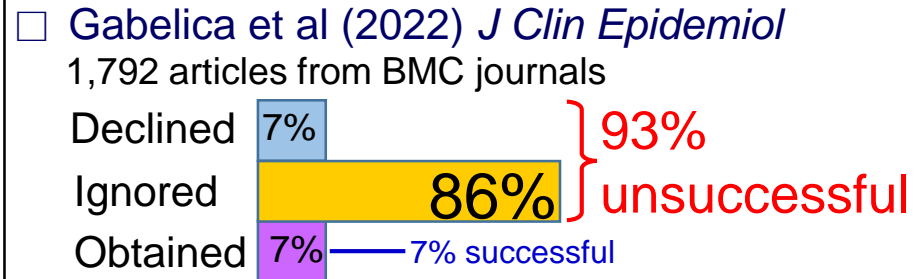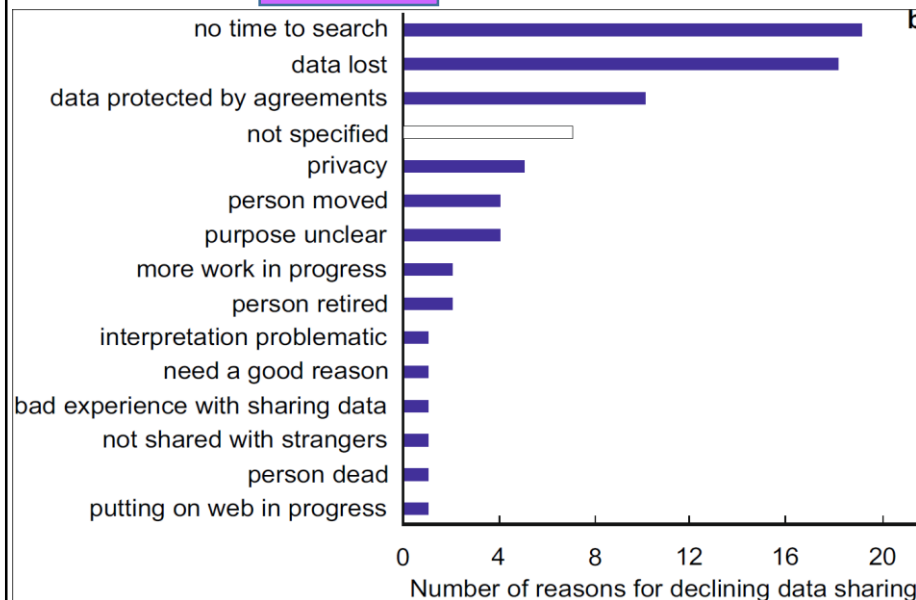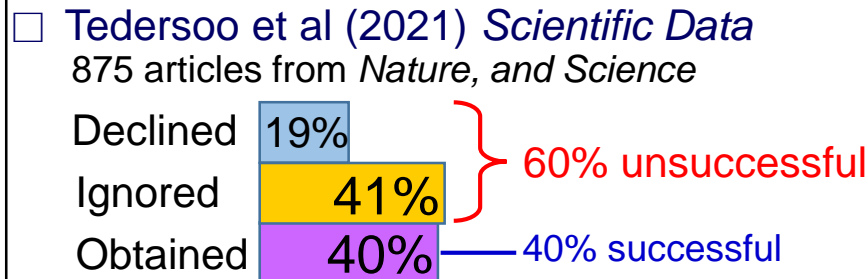

□ Tedersoo et al (2021) *Scientific Data*
875 articles from *Nature, and Science*

Declined 19%
Ignored 41%
Obtained 40%

} 60% unsuccessful
— 40% successful

Number of reasons for declining data sharing

(categories shown, top to bottom: no time to search, data lost, data protected by agreements, not specified, privacy, person moved, purpose unclear, more work in progress, person retired, interpretation problematic, need a good reason, bad experience with sharing data, not shared with strangers, person dead, putting on web in progress)

□ Gabelica et al (2022) *J Clin Epidemiol*
1,792 articles from BMC journals

Declined 7%
Ignored 86%
Obtained 7%

} 93% unsuccessful
— 7% successful

Table 3. Reasons for not sharing requested data (N = 132)

| Reason	N (%)
The authors asked for more information about our study, but after our detailed response and clarification, we did not receive further response from them	23 (17.42)
Their informed patient consent did not include sharing data with other researchers, or the ethical committee prohibited external data sharing and use	14 (10.6)
They cannot access the data, either because they are no longer in the institution that conducted the research or they are no longer active on the project	12 (9.09)
They do not want to share the data or in any way participate in our study without a specific explanation	11 (8.33)
No reply after we signed and sent an NDA or DTA that the authors requested	10 (7.57)
The corresponding author instructed us to use a web service to request the data	8 (6.06)
The data do not belong to them	6 (4.54)
The study was still ongoing	5 (3.78)
Privacy concerns; specifically, they did not want to share deidentified patient data, and other concern was third party data storage and safekeeping from external access to patient data	4 (3.03)
This was a summary article, and there are no data to share	3 (2.27)
We were unable to meet their specific requests regarding the NDA	2 (1.51)
The author requested reimbursement for data sharing	2 (1.51)
The author wanted ethics approval translated into English	2 (1.51)
I will send you the data (however, the authors did not send the data subsequently)	2 (1.51)
After both parties signed the NDA, the author wrote back to say that data sharing is against their ethical committee suggestions	1 (0.75)
A conceptual article, no data to share	1 (0.75)
The data set has 8 Gb and is unable to share such a large file	1 (0.75)
The most important data and the data set is in the supplementary file in the manuscript (however, this was not the case; the supplementary file did not contain raw data from the study)	1 (0.75)
The author misplaced the data	1 (0.75)
The author was not sure what data we needed (the response remained the same even after we sent three emails to explain	1 (0.75)

□ Individual researchers do not well comply to data management policy.

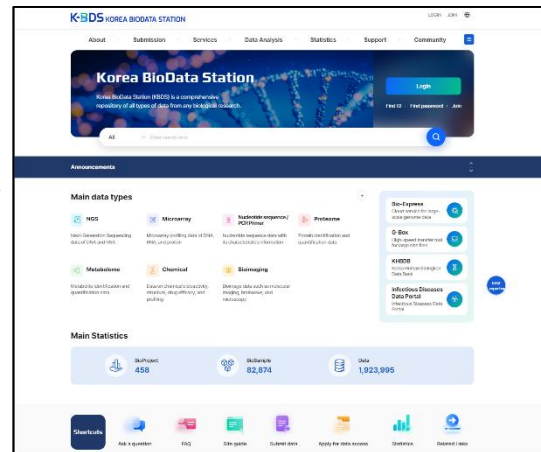▪ Hold institutions responsible for data management instead of individual researchers

Korea BioData Station (K-BDS)

☐ Our response to Korea's data management policy

☐ **A repository of all types of data from biological research**

☐ **URL:** https://kbds.re.kr/

☐ In operation since December 2021
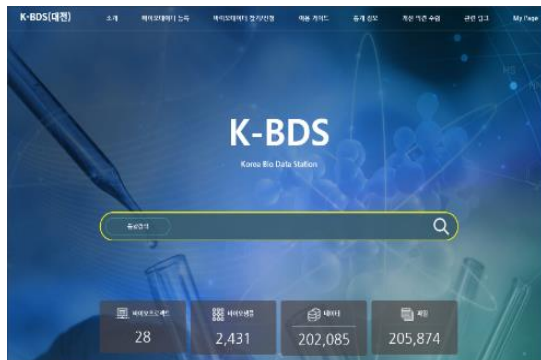
Pilot version (2021)	The first official version (2022)	An upgraded version (2023, upcoming)
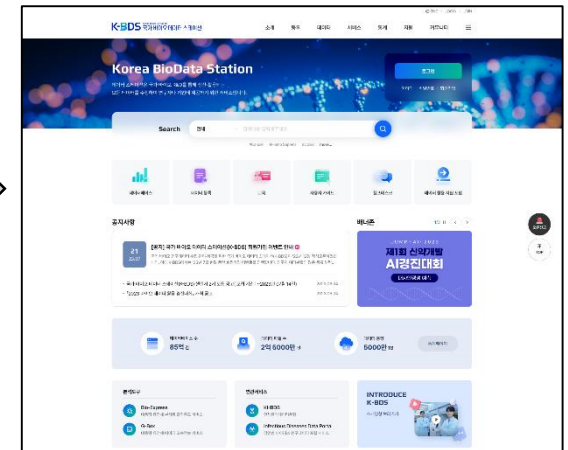
Types of Data in K-BDS

Data Curation Center

Genomics DCC (KOBIC)

Proteomics DCC (Korea U.)

Metabolomics DCC (Seoul Nat'l U.)

Chemical DCC (Kor Res Inst Chem Tech)

BioImage DCC (Ewha U.)

Major data types

Domain-specific repositories exist worldwide

NGS Microarray Nucleotide sequence

Quantitative proteomics

Metabolomics

Chemical's bioactivity, structure, drug efficacy, profiling

Microscopy, molecular imaging, brain signal

Others

Topic-specific miscellaneous data
- Medical device, Brain science, Drug
- Pesticide, Food, Fishery
- Environment

All other data (Unstructured data)
- Generalist data

Correspond to those types of data submitted to NIH GREI* repositories
*Generalist Repository Ecosystem Initiative

figshare DRYAD
zenodo Vivli OSF
The Dataverse Project MENDELEY DATA

EBI BioStudies
EMBL-EBI
BioStudies.

NGDC OMIX
OMIX
Miscellaneous data

Subdomains in K-BDS

Major data types

https://kbds.re.kr/KRA — NGS

https://kbds.re.kr/KArray — Microarray

https://kbds.re.kr/KNA — Nucleotide sequence

https://kbds.re.kr/KPOP — Quantitative proteomics

https://kbds.re.kr/KMAP — Metabolomics

https://kbds.re.kr/KCODE — Chemical's bioactivity, structure, drug efficacy, profiling

https://kbds.re.kr/KBI — Microscopy, molecular imaging, brain signal

KRIBB Korea Research Institute of Bioscience & Biotechnology

kobic 국가생명연구자원정보센터 Korea Bioinformation Center

Existing KoNA will be integrated into and replaced by K-BDS

KoNA: **Ko**rean **N**ucleotide **A**rchive
https://kobic.re.kr/kona/

By December 2023

Major data types

https://kbds.re.kr/**KRA** — NGS

https://kbds.re.kr/**KArray** — Microarray

https://kbds.re.kr/**KNA** — Nucleotide sequence

https://kbds.re.kr/**KPOP** — Quantitative proteomics

https://kbds.re.kr/**KMAP** — Metabolomics

https://kbds.re.kr/**KCODE** — Chemical's bioactivity, structure, drug efficacy, profiling

https://kbds.re.kr/**KBI** — Microscopy, molecular imaging, brain signal

Korea Research Institute of Bioscience & Biotechnology

kobic 국가생명연구자원정보센터 Korea Bioinformation Center

K-BDS is offered in both Korean and English languages

☐ Correct language version will automatically show up based on the preferred language setting in your web browser.

☐ In addition, you can directly access a language specific-URL.

- https://kbds.re.kr/kr for Korean
- https://kbds.re.kr/en for English

☐ In addition, you can toggle between the two versions.
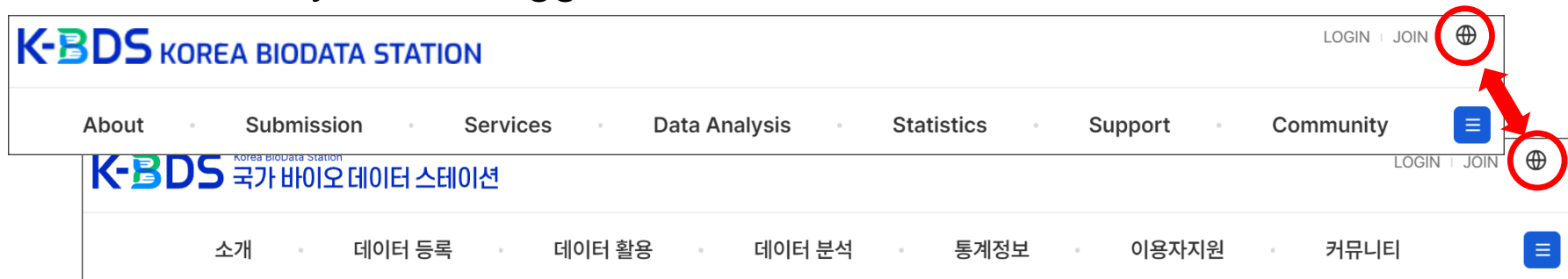
K-BDS is offered in both Korean and English languages

☐ In either version, submitter's text input will be strictly in English, with just a very few exceptions. For example, for "BioProject title" field:

Version	When you are **submitting**	When you are **browsing**
In **Korean** version	Two fields will appear: (1) English title (2) Korean title	Two fields will be displayed: (1) English title (2) Korean title
In **English** version	Only one field will appear: English title	Only one field will be displayed: English title

☐ So there are some fields where Korean text will be inputted, but we never allow Korean text input to the fields that can be shared internationally.

☐ There should be no language problem when

- sharing our data internationally

- getting data submissions internationally

Administration webpages (Not accessible to users)

Visible only to administrators

Move to administration web page

Many functionalities for various aspects of administration

Check meta data and raw data
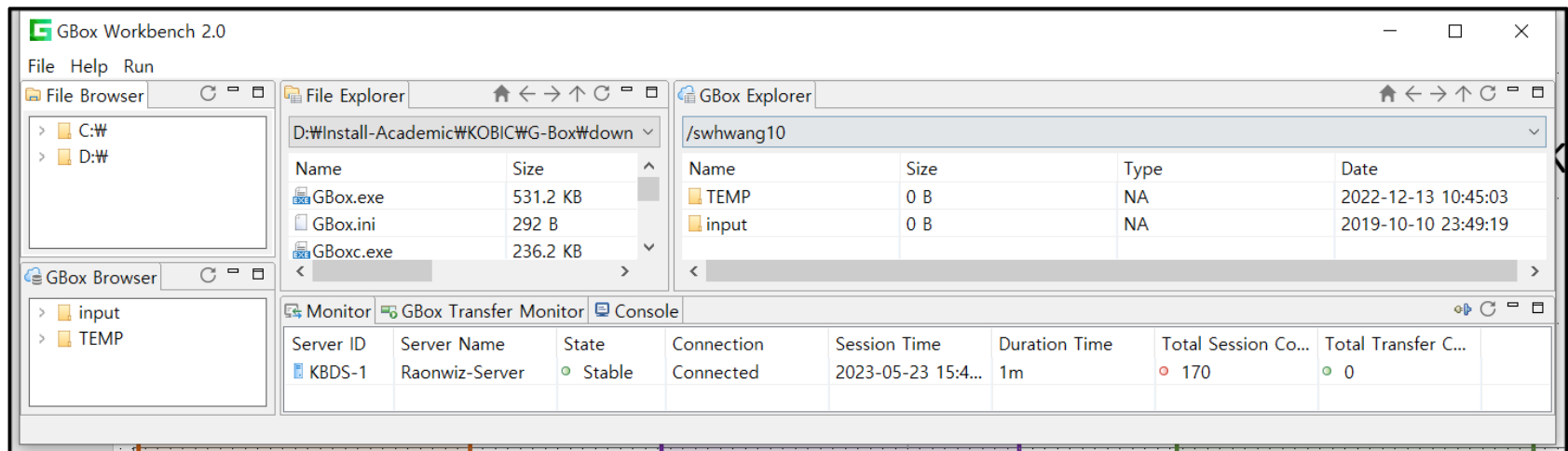
Approve the submission

Send a decline message to the submitter

G-Box

G-Box: A high-speed file transfer application

- Used for both for uploading to and downloading from K-BDS
- ~ 10 times faster than HTTP or FTP
- Based on Samsung's RAPIDANT solution (~comparable speed to IBM's Aspera)
- Available for Windows, Linux, and Mac
  - Also available as both a GUI version and a command-line version

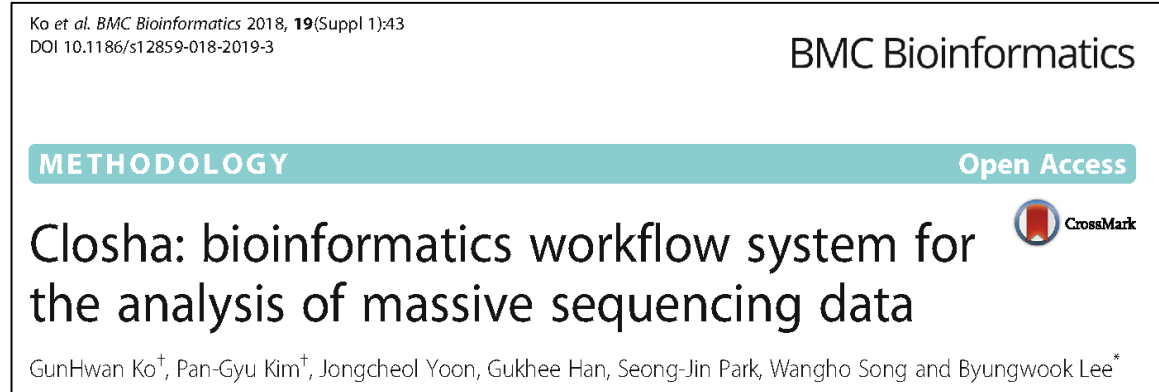| Windows 8, 9, 10		Linux Ubuntu 20.x +, CentOS 7/8		Mac macOS 11.0.1 +		
Bio-Express CLOSHA Workbench Windows	⬇	Bio-Express CLOSHA Workbench Linux	⬇	Bio-Express Workbench Mac	Intel	Bio-Express
Bio-Express GBox Windows	⬇	Bio-Express GBox Linux	⬇	Bio-Express GBox Mac	Intel	Bio-Express
Bio-Express GBox-CLI Windows	⬇	Bio-Express GBox-CLI Linux	⬇	Bio-Express GBox-CLI Mac	Intel	Bio-Express

Korea Research Institute of Bioscience & Biotechnology

kobic 국가생명연구자원정보센터 Korea Bioinformation Center

Bio-Express

Bio-Express: A cloud-based large-scale genome data analysis application

- A workflow (pipeline)-based system (like Galaxy)

- Began its service in 2016, published in 2018

BMC Bioinformatics

METHODOLOGY   **Open Access**

Closha: bioinformatics workflow system for the analysis of massive sequencing data

GunHwan Ko[†], Pan-Gyu Kim[†], Jongcheol Yoon, Gukhee Han, Seong-Jin Park, Wangho Song and Byungwook Lee[*]

- **URL:** https://www.kobic.re.kr/bioexpress/

- Standalone version is available for Windows, Linux, and Mac

| Windows 8, 9, 10		Linux Ubuntu 20.x +, CentOS 7/8		Mac macOS 11.0.1 +		
Bio-Express CLOSHA Workbench Windows	⬇	Bio-Express CLOSHA Workbench Linux	⬇	Bio-Express Workbench Mac	Intel	Bio-Express
Bio-Express GBox Windows	⬇	Bio-Express GBox Linux	⬇	Bio-Express GBox Mac	Intel	Bio-Express
Bio-Express GBox-CLI Windows	⬇	Bio-Express GBox-CLI Linux	⬇	Bio-Express GBox-CLI Mac	Intel	Bio-Express

- A web application version is available as well, which runs directly on web brower

Korea Research Institute of Bioscience & Biotechnology

kobic 국가생명연구자원정보센터
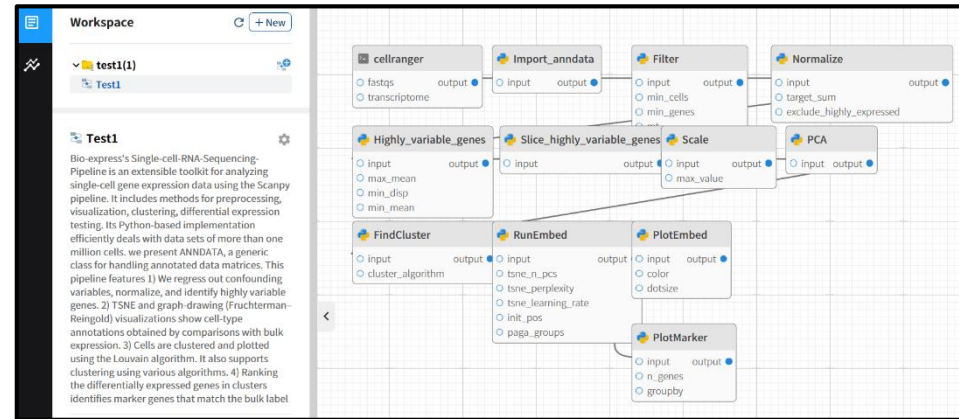Korea Bioinformation Center

Bio-Express

Both versions use KOBIC's cloud computing resource
(not the end user's local PC resource)
Current quota per-user: 100 CPU core, 400 GB memory, and 1TB storage

Stand-alone version
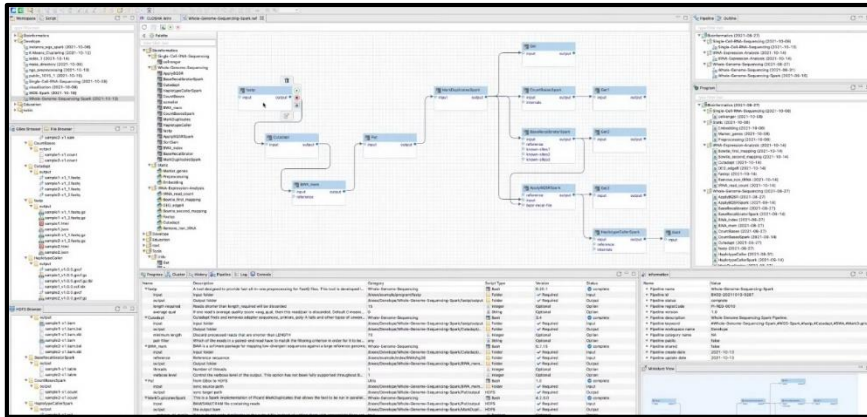
Web application version

☐ The data files downloaded from or uploaded to K-BDS through G-Box can be readily imported into Bio-Express

Upload

Download

This PC

G GBox

K-BDS KOREA BIODATA STATION

BX BIO EXPRESS

kobic 국가생명연구자원정보센터
Korea Bioinformation Center

Conclusion

K-BDS KOREA BIODATA STATION

Multiple data archives are housed in a single unified repository

NGS   Microarray   Nucleotide sequence

Quantitative proteomics

Metabolomics

Chemical's bioactivity, structure, drug efficacy, profiling

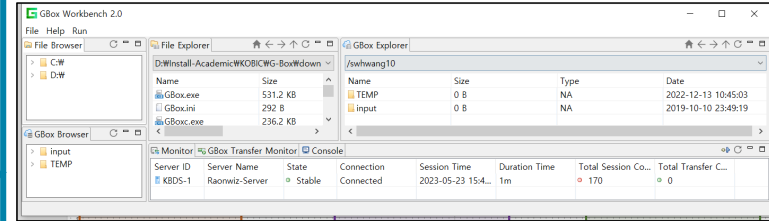Microscopy, molecular imaging, brain signal

Seamless coupling of these three systems

G GBox

High speed file transfer is available in multiple OSs and both GUI / CLI

BX BIO EXPRESS

A cloud- and workflow-based data analysis application is available in both standalone and web versions

- A solution for data management policy in Korea

- A useful addition to global ecosystem of data sharing resources