2025 年度終了報告書

ライフサイエンスデータベース統合推進事業
統合化推進プログラム

【研究課題名】「空間オミックスデータ解析用データベースの開発」

研究代表者：
VANDENBON Alexis（京都大学 医生物学研究所 准教授）

研究分担者：
なし。

1. 概要（Overview）

（1）研究および計画の概要（Research and Project Overview）

Spatial transcriptomics is a new technology that allows researchers to study gene expression patterns within tissue slices. We are constructing a database that allows users to easily access publicly available spatial transcriptomics data of various tissues, without the need for bioinformatics experience. This database allows anyone to explore spatial transcriptomics data and generate new hypotheses.

（2）成果の概要（Summary of Results）

At the beginning of the project, we focused on collecting and processing spatial transcriptomics samples generated using the 10X Genomics Visium platform. After implementing prototypes, we developed a full version containing 1,038 Visium samples and several tools for data exploration. We released DeepSpaceDB (version 1.0) in September 2024, earlier than planned.

As of October 2025 (version 1.1), DeepSpaceDB includes 2,144 Visium samples. Users can search for samples by tissue or condition, view sample quality, predicted spatial domains, spatially variable genes and pathways, image annotations (by a pathologist or AI), predicted cell-type compositions, and cell-cell communication. The “Tissue Explorer” tool allows users to freely select multiple parts of a tissue section (or of two different sections) and compare gene expression patterns between them. Various data files are also available for download.The database also offers a “Search” function to find samples where a query gene shows distinct spatial patterns. Users can also upload and analyze their own Visium data. Such interactive tools are not available in other databases, making DeepSpaceDB a unique resource for accessible spatial transcriptomics data.

Since 2024, we have also been processing Xenium platform samples. Over 1,000 have been processed, and interface development is ongoing. The Xenium interface differs from the Visium one but supports similar interactive analyses, including region-based comparisons. A manuscript describing this tool is planned for submission by the end of 2025.

名称（Name）	概要（Summary）
DeepSpaceDB	A database that allows both wet and dry biologists to easily explore the growing number of publicly available spatial genomics datasets—covering gene expression, pathways, and cell type distributions across tissues and microenvironments. It also offers consistently processed and well-annotated samples for download, facilitating re-analysis, hypothesis generation, and comparison with new spatial data.

2. 目的・目標の達成状況（Achievement status of objectives and goals）

（1）達成目標と達成状況（Goals and Status of Achievement）

達成目標（Goals）	達成状況(Status of Achievement)
【Item 1】Preparation data for database beta version	
Visium samples data analysis	In total we successfully collected, analyzed, and processed 2,144 Visium samples.
Tissue microenvironment analysis	We conducted analyses and predicted spatial domains in all Visium samples in FY2025.
【Item 2】Implementation beta version of database	
Implementation simple prototypes of the database	Two prototypes, using R and Python, had been successfully implemented by March 2024.
Implementation data browsing layers	We successfully implemented several ways to browse the data from various points of view.
Implementation analysis functions	A variety of analysis functions was successfully implemented.
Beta testing and Beta release	We tested beta version of the database and published the database in September 2024.
【Item 3】Preparation database release version	
Visium data update	We regularly updated the Visium data, finally increasing the number of samples to 2,144.
Expansion of covered platforms	We have collected and processed >1,000 Xenium samples and are implementing a new interface.
Database implementation	Implementation was mostly ready by September 2024. Additional tools were added later.
RDF implementation	We have not started this item yet.
Release the release version	The database was successfully made public in September 2024, earlier than planned.

（2）実施状況の詳細（Details of implementation status）

All items were conducted by the Vandenbon Group.

【Item 1】Preparation data for database beta version

（Item 1.1）Visium samples data analysis

This item is the main foundation of the database. In brief, we collected a large number of Visium samples, covering many tissues and conditions. We analyzed them from various points of view, and results were made available in the database (see below). We can say that we completed this task successfully. Below follows a more detailed description.

The data of 10X Genomics Visium samples were collected from various sources (NCBI's GEO, etc), including transcriptome and image data. We manually annotated each sample, as far as possible, by assigning a tissue of origin (using the UBERON Ontology), a disease status (using the Disease Ontology), and other meta data including species, accompanying scientific papers, date of publication, age or developmental stage, and others.

Each Visium sample contains transcriptome and image data (typically an H&E staining of the tissue slice). The transcriptome data of each sample was processed using a standard pipeline, including calculation of quality metrics, normalization, dimensionality reduction, and clustering. We excluded 30 samples of low quality. In the remaining 2,144 samples, we predicted items to add to the database: spatially variable genes, biological pathway activities, spatial domains (i.e., anatomical compartment with distinct gene expression patterns), cell type compositions, and cell-cell interactions, using published methods. To improve the accuracy of cell type predictions, we have been developing a method, called oCELLoc, which uses regularization techniques to select a small subset of relevant cell types from a larger reference. Work on this method is ongoing.

The collected samples were produced by different laboratories, and it was therefore unclear if comparisons between samples are biologically meaningful. To check this, each Visium sample was converted into a pseudo-bulk profile and analyzed, confirming that samples from the same organ generally showed similar expression patterns. We also aggregated all 5.4 million Visium spots and, after batch correction, clustered them by similarity. Each cluster was then manually assigned a broad label (e.g., tumor, epithelial), based on dominant tissue type or condition, providing useful context for spatial data interpretation.

Image annotations were conducted using 2 approaches. One is through a collaboration with Dr. HORIMOTO Yoshiya (Tokyo Medical University Hospital and the Juntendo University Faculty of Medicine) who is a pathologist specialized in breast pathology. He annotated the H&E image data for 69 human breast cancer samples in DeepSpaceDB, indicating invasive and non-invasive tumor tissue as well as necrosis tissue. However, it should be noted that the low quality of Visium H&E images often makes high-quality annotations impossible. As a second approach, we used OpenAI's GPT-4o for image annotation. In brief, each image was cut into a 5-by-5 grid which we passed on to GPT-4o to obtain descriptions of the structural and pathological features. The resulting predictions were included in DeepSpaceDB. However, it is difficult to systematically evaluate the accuracy of these predictions. Therefore, we have added a message, advising users to treat these predictions with care.

(Item 1.2) Tissue microenvironment analysis

To explore patterns larger than individual spots, we also explored microenvironments within tissues. Here, making use of the hexagonal pattern of Visium spots, we defined "microenvironments" as a central spot and its 6 neighbors. We calculated expression patterns in each microenvironment and explored patterns through clustering analysis. Here too, we observed that some microenvironment clusters were associated with particular tissues or conditions (ex: cancer-associated microenvironments). However, after conducting these initial analyses, we concluded that the concept of microenvironments is more suitable to explore with technologies with a higher resolution (such as Xenium) compared to Visium. Therefore, we concluded that the priority for including our current microenvironments in the database is low. Instead, we predicted spatial domains (mentioned above) in each sample using two existing methods and added the results to the database.

【Item 2】 Implementation beta version of database

(Item 2.1) Implementation simple prototypes of the database

In the latter half of fiscal year 2023, we prepared two prototypes: one using the Shiny package in R, and one using the Flask framework in Python. After comparison, we decided to continue development using the Flask version. The current version of the database is implemented using Python's Flask framework for the server-side and JavaScript for the client-side functionality.

(Item 2.2) Implementation data browsing layers

Originally, we planned to implement 3 layers: 1) tissues (samples), 2) spots or cells, and 3) microenvironments. However, the final implementation was based on the following pages:

- **Database**: a table of all samples in the database, which can be searched using keywords and various filters (**Fig. 1A**).
- **Sample**: a page which lists various information and data for the selected sample, organized into multiple sections.
- **Search**: users can search the database using their gene (or pathway) of interest, and easily see in which samples their gene of interest has clear spatial patterns of activity.
- **Upload**: users can upload the data files of their own Visium samples, and explore it using a subset of the tools of DeepSpaceDB.

In addition, the database has Tutorial (including text and videos), About (describing the aim and funding of the project) and Contact pages.

The Sample page is the core of the database, and includes the following sections:

- **Metadata**: (**Fig.1B**) the metadata of the sample, such as the tissue of origin and condition.

- **Location**: a 2D embedding of all samples which allows users to easily find similar samples.
- **Quality**: quality indicators (such as the number of detected genes per spot; **Fig.1C**) can be visualized within the tissue slice. The quality indicators of the selected sample can also be compared to those of other samples in the database.
- **Image annotation**: shows the image of the tissue slice with AI-based image annotations. For 69 human breast cancer samples, annotations by Dr. HORIMOTO Yoshiya are provided.
- **Clusters**: visualization of the clustering result of the spots of the sample (**Fig. 1D**). This section also includes the spatial domains predicted by two methods.
- **Genes**: three methods were used to predict spatially variable genes in each sample (**Fig. 1E**). Users can easily visualize the spatial expression pattern of genes.
- **Pathways**: We predicted the activity of biological pathways from the expression of the genes that are involved in them. In this section, users can visualize the spatial activity patterns of these pathways.
- **Cell types**: We predicted the cell type compositions of all spots in each sample using

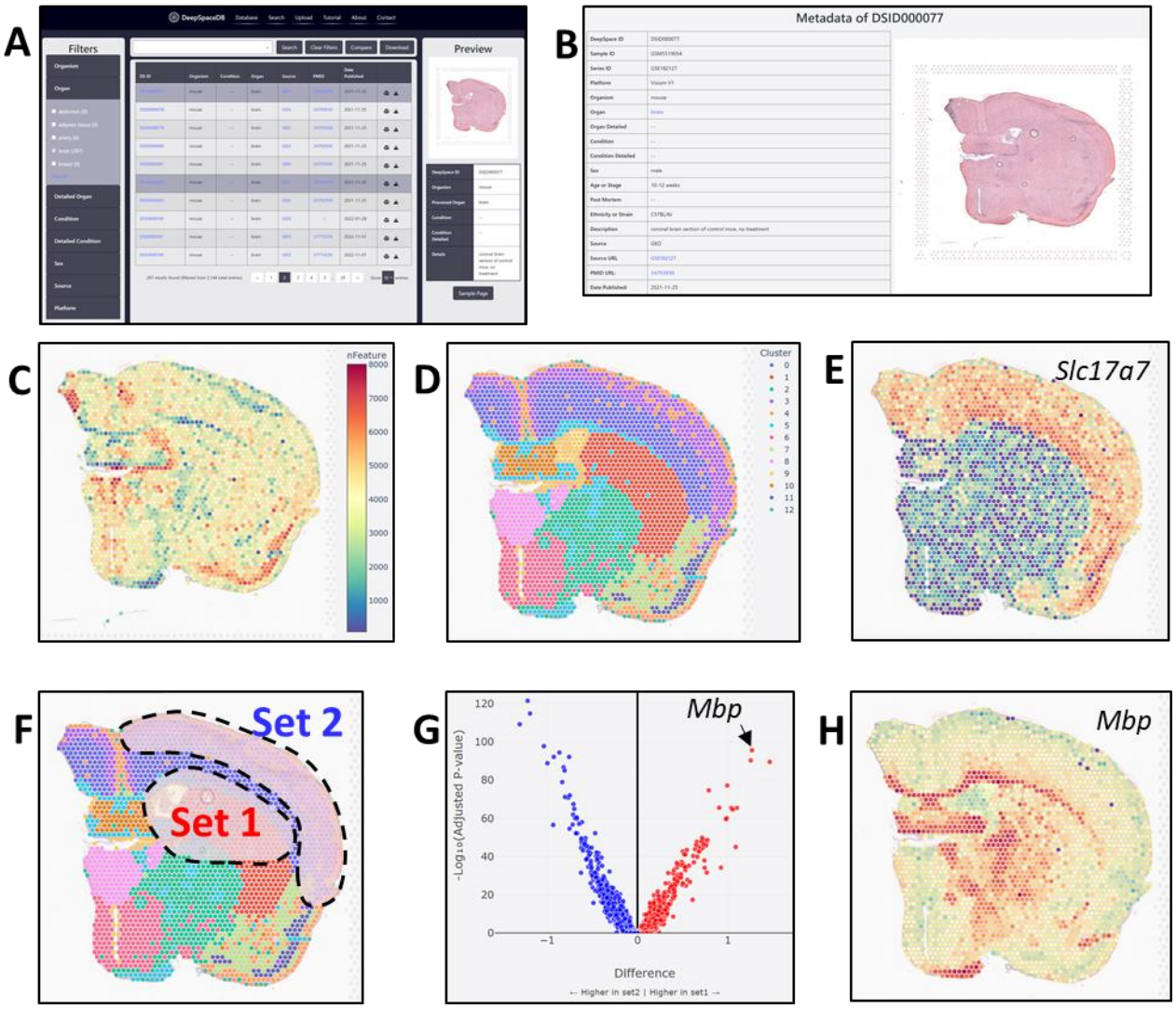

Figure 1: A selection of features in DeepSpaceDB.

2 existing methods. Here, users can visualize the spatial distribution of cell types.

- **Tissue Explorer**: (also explained in Item 2.3; **Fig. 1F-H**) Users can manually select several regions of interest in the tissue slice using their mouse cursor, and compare gene expression levels in the selected regions.
- **CCI**: We predicted cell-cell communication in each sample. The prediction result can be downloaded here.
- **Download**: The raw data as well as several prediction results can be downloaded here. Users can also download the processed data as a Seurat object, which they can use for further analysis on their own computer.

(Item 2.3) Implementation analysis functions

Most of the items listed above involve the visualization of pre-calculated results. However, a number of tools require interactive processing and calculation on our server. In the Tissue Explorer, users can interactively select regions of interest within a tissue slice (Fig. 1F). On our server, the average gene expression is subsequently calculated within the selected regions, and a t-test is performed to evaluate differences in expression between the selected regions. The result is returned to the user in the form of a table, but also as a scatterplot and a volcano plot (Fig. 1G). The same analysis can also be performed for pathways. This tool allows users to interactively explore spatial samples and find genes or pathways of interest (Fig. 1H). Such interactive analysis tools are absent in existing databases. Later, we added a tool which allows comparisons between two different tissue slices (ex: a control and a treatment sample).

The Search tool allows users to search the database using their gene (or pathway) of interest. In response to a query, on our server, all samples are sorted by the strength of the spatial pattern of the query gene (or pathway). Subsequently, a statistical test is applied to find tissues and conditions that are enriched among the samples with the strongest spatial patterns.

Using the Upload tool, users can upload their own Visium data files. The uploaded data is processed on our server, and after that, users can interactively explore their uploaded sample using a subset of the tools present in the database.

(Item 2.4) Beta testing and Beta release

By April 2024, we had implemented many of the functionalities of the database. We shared a private version with collaborators and advisors, and made corrections and changes according to their feedback. We also presented the database in an internal seminar in our institute in July 2024, and in the DICP site visit in August 2024. We made DeepSpaceDB public in early September 2024, just before the NGS Expo 2024 conference. The release was earlier than

originally planned (planned to be March 2025).

【Item 3】Preparation database release version

(Item 3.1) Visium data update

We are continuously collecting and processing new samples, and have updated the database regularly. Table 1 shows the number of samples over time until the present status.

Date	Total	Human	Mouse
March 2024	700	366	334
Sept 2024 (first release)	1,039	627	412
March 2025	1,674	1,011	663
Oct 2025 (present)	2,144	1,361	783

Table 1: Numbers of Visium samples in DeepSpaceDB.

(Item 3.2) Expansion of covered platforms

This item is still ongoing. So far, we have collected and processed roughly 1,000 Xenium samples, including both samples using a small gene panel and Xenium Prime samples (covering about 5,000 genes). Samples are collected from various sources, and processed using a standard pipeline. Since Xenium data is considerably larger than Visium data, we are processing it into Zarr stores for easier access, and we are implementing a new interface. We aim to make this interface public by the end of 2025.

(Item 3.3) Database implementation

Although we originally planned to implement the database release version during fiscal year 2025, we started and completed the implementation earlier. The implementation was mostly finished by August 2024.

(Item 3.4) RDF implementation

In our original application, we planned to implement RDF during fiscal year 2025. However, we have not yet started this.

(Item 3.5) Release the release version

We released DeepSpaceDB to the public in September 2024, ahead of plan. At present, the public version only includes samples of the Visium platform. We plan to release the Xenium interface by end of 2025.

(3) 主な成果論文等 (Major research papers, etc.)

1. Vandenbon Alexis, *et al.*, “DeepSpaceDB: a spatial transcriptomics atlas for interactive

in-depth analysis of tissues and tissue microenvironments", Nucleic Acids Research, 2026, Database Issue, *in press* (DOI: 10.1093/nar/gkaf1117).
This paper introduces DeepSpaceDB in detail, including collection and processing of the Visium data, the methodology underlying analysis tools, and several example applications.
2. Vandenbon Alexis, *et al.*, "Mining Spatial Transcriptomics Datasets using DeepSpaceDB", Journal of Visualized Experiments, 2025, 223 (DOI: 10.3791/68892).
This protocol article briefly introduces the database's structure and demonstrates its use through examples, including mouse brain and liver analyses. This journal focuses on visualized protocols; the staff of the journal is currently preparing a video version of the protocol (about 10 minutes), showing example usage of the database.

(4) 主要なデータベースの利活用状況 (Usage of major databases)

The database was made public in September 2024. Since then, the monthly number of visits has steadily increased, suggesting expanding recognition and growing interest in the database, likely reflecting both continued outreach efforts and word-of-mouth among researchers. Overall, the access numbers are considered appropriate and show a positive trend. At present, a large proportion of accesses is from within Japan. With the publication of our paper in the *Nucleic Acids Research* Database Issue, we expect the number of accesses from abroad to further increase.

(5) データベースを利用して得られた研究成果・産業応用の例 (Examples of research results and industrial applications obtained using the database)

Since the database was made public only relatively recently, we have not heard of research results obtained by other researchers based on our database, except for collaborators.

3. 今後の計画および展望 (Future plans and outlook)

We plan to continue the processing of Xenium data and the implementation of the Xenium interface. We aim to make this interface public by the end of 2025, and prepare a manuscript. We plan to continue making updates to the samples in the database. We will make one update for the Visium data by the March 2026. After the publication of the current Xenium data, we will also perform regular updates of this data.

One remaining bottleneck is the manual annotation of new samples. We will explore the use of large language models (LLMs) for partly automating this annotation. We also plan to make improvements to the LLM-based image annotations. Using the GPU we acquired as part of this project, we will use more specialized AI models for adding more meaningful and detailed annotations to the H&E images. We plan to use the transcriptomics data to verify the validity of the annotations (ex: in regions annotated as tumors, can we confirm high expression of tumor marker genes?).

Finally, by March 2026, we plan to start preparations for adding a third platform to DeepSpaceDB. One candidate is the Visium HD platform. We will collect samples and prepare a processing pipeline.

Through the steps listed here, we hope to ensure the long-term growth, sustainability and scientific relevance of DeepSpaceDB.

4. 計画・実施体制等の妥当性 (Appropriateness of the plan and research groups)

(1) 各グループの担当項目 (Items of responsibility of each group)

(1)-1. VANDENBON Group (Kyoto University)

The Vandenbon Group conducted all practical steps of the research plan, including the collection of samples, processing, quality check, database implementation, testing, and maintenance. However, the image pathological annotation of the human breast cancer sample images was conducted by Dr. HORIMOTO Yoshiya.