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）
研究開発実施報告書　様式

2024 年度 研究開発実施報告

概要

研究開発課題名	空間オミックスデータ解析用データベースの開発
開発対象データベースの名称（URL）	DeepSpaceDB (https://deepspacedb.com/)
研究代表者氏名	VANDENBON Alexis（60570140）
所属・役職	京都大学医生物学研究所・准教授（2025 年 3 月時点）

□目次

§1. 研究実施体制

グループ名	研究代表者・研究分担者氏名	所属機関・役職名	研究題目
Vandenbon Group	Vandenbon Alexis	京都大学・准教授	Data analysis and database implementation

§2. 研究開発対象とするデータベース・ツール等

(1) データベース一覧

【主なデータベース】

No.	名称	別称・略称	URL
1	DeepSpaceDB		https://deepspacedb.com/

【その他のデータベース】

No.	名称	別称・略称	URL
1			

(2) ツール等一覧

No.	名称	別称・略称	URL
1	singleCellHaystack		https://github.com/alexisvdb/singleCellHaystack
2	oCELLoc		https://github.com/afeefa-zainab/oCELLoc (未公開)

§3. 実施内容

(1) 本年度に計画されていた研究開発項目・タスク

Visium sample annotation supported by RNA-seq data analysis

- Further prepare reference bulk RNA-seq and single-cell RNA-seq datasets
- Apply methods for cell type composition prediction to all spatial transcriptomics samples
- Cluster Visium spots by similarity of gene expression patterns
- Assign annotations to each cluster

Tissue microenvironment data exploration

- Explore the transcriptomes of microenvironments, and define groups of similar microenvironments in the database
- Prepare the annotation and downstream analysis of the microenvironments

Tissue section image data analysis

- Use spot-level annotations to assign annotations to corresponding parts of images
- Initiate collaboration with histologist focusing on subset of images for one tissue
- Attempt to implement function that allows users to add comments to images

Implementation beta version of database

- Complete a few prototypes using different platforms and decide on the better platform to use
- Implement the user interface
- Implement analysis functions
- Beta testing of the database
- Release the beta version of the database at the latest in March 2025

Preparation of the database release version

- Regularly update our collection of Visium samples to include newly published samples
- Track the availability of samples of other spatial transcriptomics technologies, and consider other popular platforms to add in the future

(2) 進捗状況

During fiscal year 2024, we continued the collection and processing of 10x Genomics Visium samples, we worked on tissue section image annotations, we implemented our database (DeepSpaceDB, https://deepspacedb.com/), and made it public in September 2024. We are also working on further updates for the Visium data, addition of samples of the Xenium platform, and the development of a bioinformatics tool (oCELLoc) for aiding cell type predictions in spatial transcriptomics data.

Below, the progress will be introduced following the structure described in (1) 本年度に計画されていた研究開発項目・タスク above.

<u>Visium sample annotation supported by RNA-seq data analysis</u>

In March 2024, our collection of data included 700 Visium samples (366 human and 334 mouse samples). During FY2024, we increased the number of samples to 1,674 Visium samples

(1,011 human and 663 mouse samples; end of March 2025), in addition to 79 samples of the 10x Genomics Xenium platform (further discussed below).

Because of this increase in samples, the number of tissues and conditions (e.g., disease states) included in our collection has also increased. One task in Visium data analysis is the prediction of the cell type composition of each spot in a sample. To predict the cell type composition of each spot inside each tissue sample, we have been using spacexr's RCTD (Robust Cell Type Decomposition) method. This method uses single-cell RNA-seq (scRNA-seq) reference datasets with known cell type annotations to predict the cell type composition of each spatial location in a spatial transcriptomics dataset. The new tissues and conditions in our data collection resulted in the need for additional scRNA-seq reference datasets. However, finding suitable reference datasets is difficult because of the many possible combinations of tissues and conditions. During FY2024, we realized that our current approach of manually finding suitable reference datasets is not sustainable. Therefore, we are developing a bioinformatics approach, oCELLoc, that predicts which cell types or cell states are present in a spatial transcriptomics dataset in an automated manner. This method is still under development and has not yet been made public.

Another approach to facilitate the interpretation of Visium samples is to cluster spots by their similarity of gene expression patterns. We did this in two ways: One is clustering the spots of each sample separately, using a default workflow (using the R Seurat package). A second way is to first merge the spots of all samples together into one large atlas of spots, and subsequently cluster all spots in the atlas. We refer to the resulting clusters of the latter approach as the "general clusters". By inspecting the tendencies of these general clusters (e.g., is a cluster enriched in spots of a certain tissue?) we manually assigned annotations to each of them, as far as possible. Several clusters were found to frequently overlap with tumor tissue. An example of a human breast cancer sample with annotations by a pathologist is shown in **Figure 1A**, including invasive and non-invasive tumor tissue. **Figure 1B** shows spot clusters that are often found in cancer-related samples. As can be seen from the figure, spots in the cancer-associated clusters roughly overlap with the tumor tissues annotated by the pathologist.

<u>Tissue microenvironment data exploration</u>

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 averaged the gene expression data of the spots in each microenvironment, and explored patterns by clustering microenvironments by similarity. Similar to the general clustering of spots, we observed that some microenvironment clusters were associated with particular tissues or conditions. An example of hepatocyte-associated and cancer-associated microenvironments is shown in **Figure 1C-D**. 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. In addition, exploring microenvironments using methods for detecting spatial domains could result in more biologically meaningful results. For these reasons, we concluded that the priority for including our current microenvironments in the database is low.

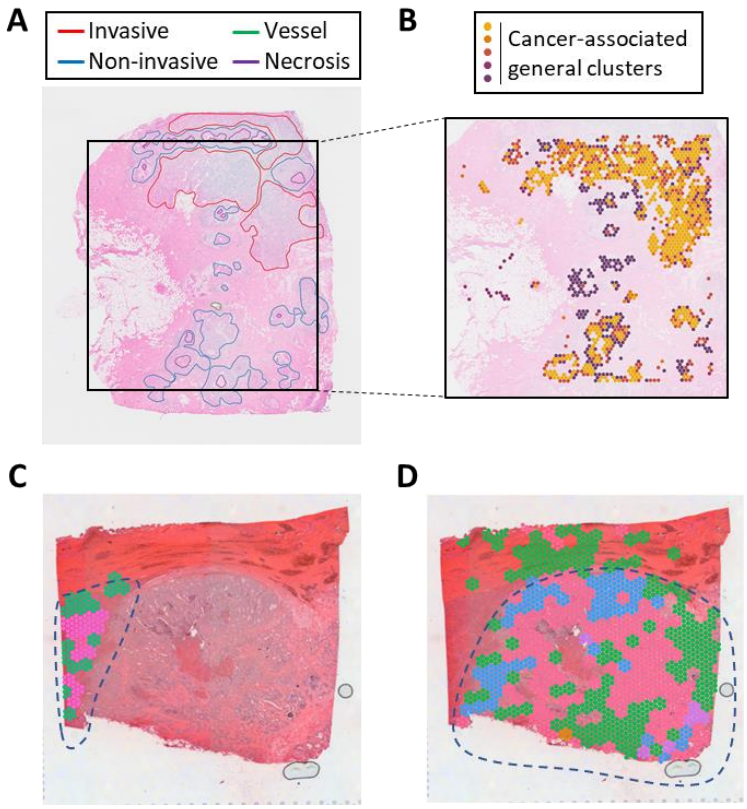

Figure 1: (**A-B**) Example of a human breast cancer sample with annotation by a pathologist (**A**) and spots assigned to cancer-associated general clusters (**B**). (**C-D**) Example of "microenvironments" in a human liver sample with colorectal cancer metastasis. Microenvironment clusters associated with hepatocytes (**C**) and with cancer tissue (**D**) are shown respectively.

Tissue section image data analysis

We initiated a collaboration with Dr. HORIMOTO Yoshiya (Tokyo Medical University Hospital and the Juntendo University Faculty of Medicine) who is a pathologist specializing in breast pathology. He conducted pathological annotations of the H&E image data for all human breast cancer samples in DeepSpaceDB, indicating invasive and non-invasive tumor tissue as well as necrosis tissue (see for example **Fig. 1A**). However, it should be noted that the quality of Visium H&E images is low compared to images used for routine diagnosis. Therefore, a high-quality annotation was not always possible. The 69 image annotations by Dr. HORIMOTO have been included in DeepSpaceDB.

Regarding the image data of other organs and tissues, we have not been successful in finding an expert who is willing to manually annotate them. Manual annotation requires expertise, time, and effort. Recently, a number of AI driven approaches for clinical image annotation have been published (such as PathChat and LLaVa-Med). We therefore explored the use of generative AI approaches for annotating tissue slice images. To the best of our knowledge, PathChat is not yet accessible, and LLaVa-Med can only be run locally and requires a relatively powerful GPU. We therefore focused on using OpenAI's GPT-4o. In brief, we used an automated pipeline to cut each tissue slice image into a 5-by-5 grid of rectangular parts and used the Python OpenAI package to

obtain prediction by GPT-4o about the structural and pathological features in each part. 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.

Implementation beta version of database

We continued the implementation of two prototypes: one using R Shiny 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. We implemented a user interface and various analysis functions, and after beta testing the database, we made it public in early September 2024, just before the NGS Expo 2024 conference. The release was earlier than originally planned (planned by March 2025). In brief, as of March 2025, DeepSpaceDB includes the following features:

- 1,674 Visium samples (1,011 human and 663 mouse samples).
- A table that allows users to filter for species, tissues, and conditions of interest, amongst others (**Fig. 2A**).
- For each tissue slice, a page summarizes (**Fig. 2B**) the background annotation information (species, tissue, condition, data source, link to publication, etc); the tissue image; quality indicators, image annotations obtained from GPT-4o and in a subset of samples provided by Dr. HORIMOTO Yoshiya; clustering results; predicted spatially variable genes and biological pathways; predicted cell type compositions; a "tissue explorer" tool where users can freely select and compare gene expression patterns between regions of interest; and downloadable data files.
- Users can select two samples and compare various features between them, including quality measures and gene expression patterns.
- Users can upload their own Visium data files and explore their own data using similar tools as for the data included in DeepSpaceDB. The uploaded data is automatically removed from the server after some time.
- A tool to search the entire database using a gene (or pathway) of interest. The results returned by this tool show all samples in the database ordered by the significance of a gene's (or pathway's) activity. This tool can be useful for exploring in tissues in which a gene (or pathway) plays a critical role. When searching the entire database using human gene *EPCAM*, top scoring samples are enriched for cancer-related samples (**Fig. 2C**). The top scoring sample is a colorectal cancer sample, which shows a very distinct pattern of expression for EPCAM (**Fig. 2D**).
- We added documentation and short videos introducing the functions in DeepSpaceDB.

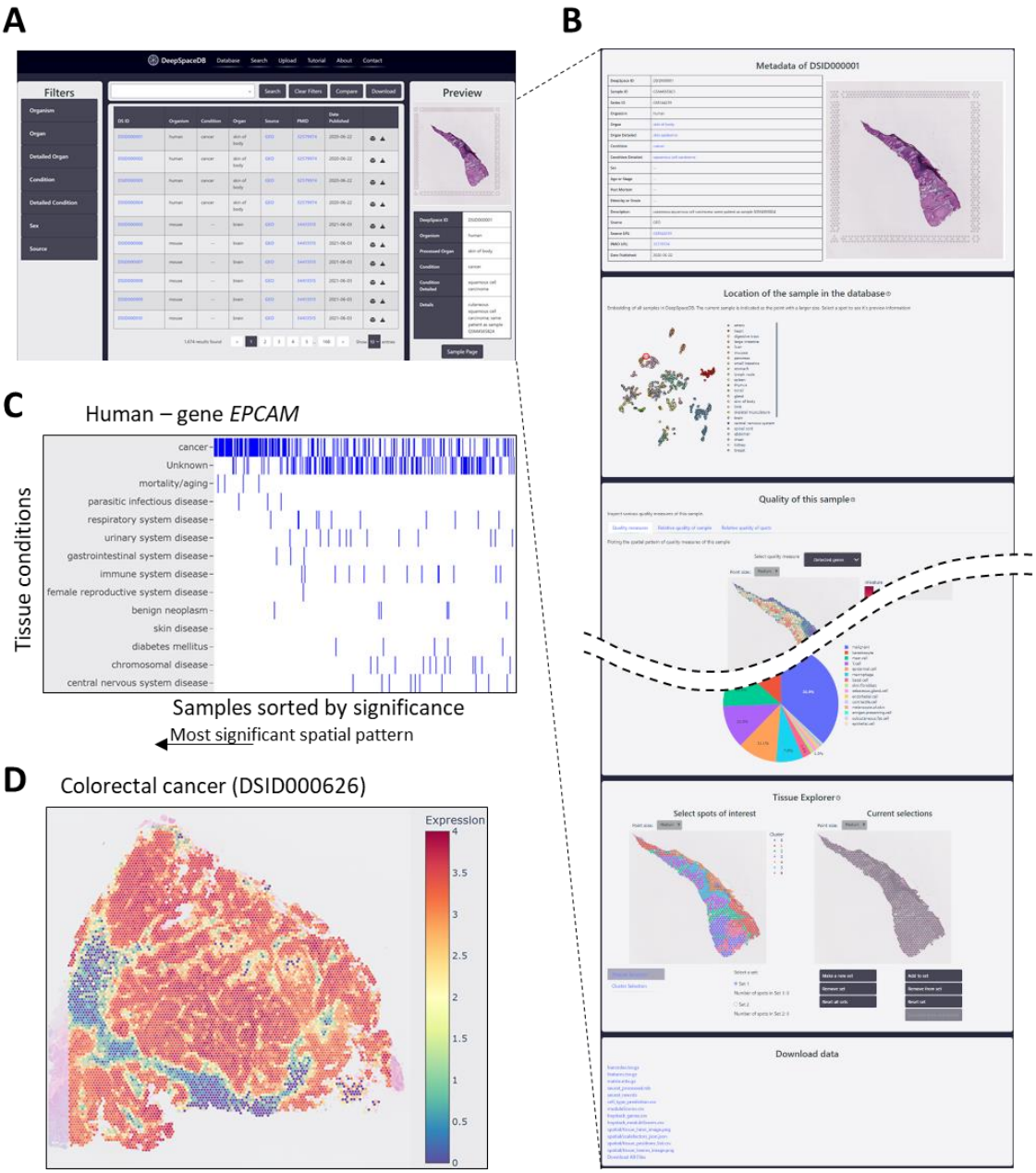

Figure 2: A selection of features in DeepSpaceDB. (**A**) A table of all samples in the database. Users can search for keyworks or select organism, organ, condition, etc of interest. A preview of the selected sample is shown at the right side of the screen. (**B**) Selecting a sample opens the sample page where a wide variety of features and functions is included. (**C**) Users can search the entire database using a gene of interest, and sort all samples in the order of significance of the spatial expression pattern of the gene. Here, we show the human samples sorted by the expression pattern of *EPCAM* with their tissue conditions. We can see that many of the samples with a strong expression pattern of *EPCAM* are cancer-associated samples. (**D**) A plot showing the expression pattern of *EPCAM* in a colorectal cancer sample, the sample with the most significant pattern of expression of *EPCAM*.

Preparation of the database release version

As described above, we made DeepSpaceDB public in early September 2024. During FY2024, the number of publicly available Visium samples has been steadily increasing, and therefore we

conducted 2 updates to our collection, as shown in the table below. We plan to continue regular updates in the future.

Date	Total	Human	Mouse
March 2024	700	366	334
Sept 2024	1,039	627	412
March 2025	1,674	1,011	663

Table 1: Numbers of Visium samples in DeepSpaceDB.

We are also tracking developments of other platforms and their available samples. Especially, we are focusing on the 10x Genomics Xenium and Xenium Prime 5K platforms. Below, we will refer to both platforms as "Xenium". We have prepared an initial test collection of 79 Xenium samples, collected from NCBI GEO and other sources. We made a data processing pipeline which normalizes this data and processes it into a format that allows it to be easily and quickly accessed (using so-called zarr stores). We also implemented a first version of the database interface for exploring Xenium samples (not yet public). At present, it allows users to zoom in and out, plot gene expression patterns at several resolutions, and select regions of interest within a tissue to inspect and compare gene expression within them.

§4. 成果発表等

(1) 原著論文発表

① 論文数概要

種別	国内外	件数
発行済論文	国内(和文)	1 件
	国際(欧文)	1 件
未発行論文 (accepted, in press 等)	国内(和文)	0 件
	国際(欧文)	0 件

② 論文詳細情報

該当なし

(2) その他の著作物(総説、書籍など)

1. Vandenbon A. and Takemoto K., "DeepSpaceDB：空間トランスクリプトミクスデータの探索的解析", *実験医学*, 43(6), 2025.
2. Honcharuk V., Zainab A., Horimoto Y., Takemoto K., Diez D., Kawaoka S., Vandenbon A., DeepSpaceDB: a spatial transcriptomics atlas for interactive in-depth analysis of tissues and tissue microenvironments, *bioRxiv*, 2025.

(3) 国際学会および国内学会発表

① 概要

種別	国内外	件数
招待講演	国内	0 件
	国際	0 件
口頭発表	国内	3 件
	国際	1 件
ポスター発表	国内	8 件
	国際	1 件

② 招待講演

〈国内〉

該当なし

〈国際〉

該当なし

③ 口頭講演

〈国内〉

1. Alexis Vandenbon, “DeepSpaceDB: a spatial transcriptomics atlas that allows interactive in-depth analysis of tissues and tissue microenvironments”, NGS EXPO 2024, Osaka, 2024年9月4日.
2. Alexis Vandenbon, “SingleCellHaystack: A universal differential expression prediction tool for single-cell and spatial genomics data”, Bio"Pack"athon 2024, Osaka, 2024年9月20日.
3. Alexis Vandenbon, “DeepSpaceDB: a spatial transcriptomics atlas for interactive in-depth analysis of tissues and tissue microenvironments”, Asia & Pacific Bioinformatics Joint Conference 2024 (APJBC2024), Okinawa, 2024年10月23日.

〈国際〉

1. Alexis Vandenbon, “DeepSpaceDB: an Interactive Database for Spatial Transcriptomics Data”, 30th East Asia Joint Symposium, Taiwan, 2024年10月29日.

④ ポスター発表

〈国内〉

1. Vladyslav Honcharuk, “A guided tour of DeepSpaceDB, a spatial transcriptomics database”, NGS Expo 2024, Osaka, 2024年9月4日.
2. Afeefa Zainab, “The Importance of Suitable Reference Data for Improved Cell Type Annotation Prediction in Spatial Transcriptomics Samples”, NGS Expo 2024, Osaka, 2024年9月4日.
3. Alexis Vandenbon, “DeepSpaceDB: a spatial transcriptomics atlas that allows interactive in-depth analysis of tissues and tissue microenvironments”, NGS EXPO 2024, Osaka, 2024年9月4日.
4. Alexis Vandenbon, “DeepSpaceDB: an interactive database for spatial transcriptomics data”, トーゴ―の日シンポジウム 2024, Tokyo, 2024年10月5日.
5. Alexis Vandenbon, “DeepSpaceDB: a spatial transcriptomics atlas for interactive in-depth analysis of tissues and tissue microenvironments”, Asia & Pacific Bioinformatics Joint Conference 2024 (APJBC2024), Okinawa, 2024年10月23日.

6. Afeefa Zainab, “The Importance of Suitable Reference Data for Improved Cell Type Annotation Prediction in Spatial Transcriptomics Samples”, Asia & Pacific Bioinformatics Joint Conference 2024 (APJBC2024), Okinawa, 2024年10月23日.

7. Alexis Vandenbon, “DeepSpaceDB: an interactive spatial transcriptomics database that allows in-depth analysis of tissues and their substructures”, The 47th Annual Meeting of the Molecular Biology Society of Japan (MBSJ), Hakata, 2024年11月28日.

8. Afeefa Zainab, “The Importance of Suitable Reference Data for Improved Cell Type Annotation Prediction in Spatial Transcriptomics Samples”, The 47th Annual Meeting of the Molecular Biology Society of Japan (MBSJ), Hakata, 2024年11月29日.

〈国際〉

1. Afeefa Zainab, “The Importance of Suitable Reference Data for Improved Cell Type An notation Prediction in Spatial Transcriptomics Samples”, Intelligent Systems for Molec ular Biology (ISMB), Canada, 2024年7月15日.

(4) 知的財産権の出願 （国内の出願件数のみ公開）

① 出願件数

種別		件数
特許出願	国内	0 件

② 一覧

1) 国内出願

該当なし

(5) 受賞・報道等

① 受賞

1. Best Poster Award at the Asia & Pacific Bioinformatics Joint Conference 2024 (APJB C2024), Alexis Vandenbon, 2024年10月25日.

② メディア報道

該当なし

③ その他の成果発表

該当なし

§5. 主要なデータベースの利活用状況

1. アクセス数

① 実績

表 1 研究開発対象の主要なデータベースの利用状況

名称	種別	2024 年度(月間平均値)
DeepSpaceDB ※1	訪問者数	8,921 (1,274 per month)
	訪問数	26,037 (3,720 per month)
	ページ数	153,756 (21,965 per month)

※1: public since September 4, 2024. Data is from September 2024 to March 2025.

② 分析

The plots in **Figure 3** show the unique visitors, the number of visits, and the pages per month over the period from June 2024 to March 2025. Overall, the trends of the statistics look promising. We made the database public on September 2024. Since then, we have made oral and poster presentations in several conferences and meetings. This, together with other activity such as the news articles by NBDC, have caused the number of visitors and visits to increase gradually over time. We hope that the publication of our manuscript in the future will further improve these statistics.

However, it is not clear how many of the visitors are bots from search engines such as Google (see 2. データベースの利用状況を示すアクセス数以外の指標 below).

Figure 3: The unique visitors, the number of visits, and the pages per month over the period from June 2024 to March 2025. DeepSpaceDB was made public in early September 2024.

2. データベースの利用状況を示すアクセス数以外の指標

In the statistics reported by AWStats, we can also see the duration of visits. For example, this is the data for March 2025:

Visit durations	Number of visits	Percent
0s-30s	4,856	77.7 %
30s-2mn	284	4.5 %
2mn-5mn	298	4.7 %
5mn-15mn	350	5.6 %
15mn-30mn	179	2.8 %
30mn-1h	192	3 %
1h+	88	1.4 %

Table 2: Visit durations for March 2025.

We can see that 77,7% of the visits in this month are short (<30 seconds). It is possible that a large portion of these visits are bots. On the other hand, about 12.8% of the visits are longer than 5 minutes. We suspect that these are researchers who are exploring several samples. In the future, after the publication of papers about DeepSpaceDB, we hope the percentage of long visits will increase.

In recent months, the monthly bandwidth of the database is roughly 200 GB per month. However, in March 2025 there was a peak in the bandwidth, increasing to about 507 GB. This was mainly caused by (we suspect) the automated downloading of a large number of sample data by one user or one institute. We are thinking about whether it is needed to restrict such accesses. In April 2025, the bandwidth returned to a typical level (about 200 GB).

3. データベースの利活用により得られた研究成果（生命科学研究への波及効果）

Our database was made public only recently. There are no such research results yet.

4. データベースの利活用によりもたらされた産業への波及効果や科学技術のイノベーション（産業や科学技術への波及効果）

Our database was made public only recently. There are no such research results yet.

§6. 研究開発期間中に主催した活動(ワークショップ等)

(1) 進捗ミーティング

年月日	名称	場所	参加人数	目的・概要
2024年4月1～2025年3月31日(毎週開催)	チーム内ミーティング(非公開)	医生物学研究所	2人～4人	研究進捗報告のためのミーティング
2024年5月8日	Meeting with advisors (非公開)	オンライン	3人	研究に関する意見交換
2024年6月24日	Meeting with advisors (非公開)	オンライン	3人	同上
2024年7月11日	Meeting with advisors (非公開)	オンライン	3人	同上
2024年7月26日	Meeting with advisors (非公開)	オンライン	3人	Discussing image annotation by pathologist
2024年8月7日	Meeting with advisors (非公開)	オンライン	8人	研究に関する意見交換
2024年8月19日	NBDC DICP site visit (非公開)	医生物学研究所	約20人	Site visit and progress report
2024年9月11日	Meeting with advisors (非公開)	オンライン	3人	Discussing image annotation by pathologist
2024年11月13日	Meeting with advisors (非公開)	オンライン	6人	研究に関する意見交換
2024年11月20日	Meeting with advisors (非公開)	オンライン	3人	Discussing image annotation by pathologist
2024年12月12日	Meeting with advisor (非公開)	オンライン	2人	研究に関する意見交換
2024年12月18日	Meeting with advisors (非公開)	オンライン	3人	Discussing image annotation by pathologist
2024年12月27日	国立がん研究センター・AI委員会(非公開)	オンライン	約10人	研究に関する意見交換
2025年1月7日	Meeting with advisors (非公開)	オンライン	5人	Discussing protocol paper
2025年2月5日	Meeting with advisors (非公開)	オンライン	5人	同上
2025年2月13日	Meeting with advisors (非公開)	オンライン	6人	同上

(2) 主催したワークショップ、シンポジウム、アウトリーチ活動等

該当なし

以上

別紙１　既公開のデータベース･ウェブツール等

No.	正式名称	別称･略称	概要	URL	公開日	状態	分類	関連論文
1	singleCellHaystack	singleCellHaystack	A tool for predicting differentially expressed genes or pathways in single cell and spatial transcriptomics data.	https://github.com/alexisvdb/singleCellHaystack	2019/2/22	維持･発展	ツール等	Vandenbon A. and Diez D., "A universal differential expression prediction tool for single-cell and spatial genomics data", Scientific Reports, 2023, 13 (1) Vandenbon A.and Diez D., "A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data", Nature Communications, 2020, 11 (1), 4318