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）

2023 年度 研究開発実施報告

概要

研究開発課題名	（和文）空間オミックスデータ解析用データベースの開発 （英文）Development of a database for spatial genomics data analysis
開発対象データベースの名称（URL）	Spatial Genomics Atlas of Cells and Tissues（仮） (https://genomics.virus.kyoto-u.ac.jp/sgact/)
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（2024 年 3 月時点）

目次

§1. 研究実施体制

グループ名	研究代表者または主たる共同研究者氏名	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	Spatial Genomics Atlas of Cells and Tissues	SGACT	https://genomics.virus.kyoto-u.ac.jp/sgact/ （未公開）

【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	singleCellHaystack	https://github.com/alexisvdb/singleCellHaystack

§3. 実施内容

(1) 本年度の研究開発計画と達成目標

During fiscal year 2023, the two large goals are preparing the necessary data for the beta release of the Spatial Genomics Atlas of Cells and Tissues (SGACT), and starting the implementation of the beta version of the database.

1) Preparation of the data for the database beta version

a. Visium sample data analysis

The main goal for fiscal year 2023 is to set up a data processing pipeline for collection, annotation, quality control, and normalization of spatial transcriptomics samples. As far as possible, this pipeline will be automated. We will first focus on human and mouse data generated using the 10x Genomics Visium platform. In brief, Visium samples will be collected from public databases, including their transcriptome data, image data, and meta data. Metadata will be processed to extract the source of the tissue section (organism, tissue, sex, age, condition, etc), date of publication of the data, data related to the used platform (although we will limit our database to the Visium platform initially), and scientific literature related to the sample, and organize this metadata in a systematic manner. Quality control steps will be implemented to filter out low-quality samples or spots. Data will be normalized to enable comparisons of the transcriptome data of between spots and between samples, and their integration into one large dataset per organism.

Public metadata typically does not include detailed annotation of the different locations within a tissue slice. We will attempt a number of approaches for adding location-specific annotations to each sample. One approach is to predict the cell type or cell type composition at each location within the tissue slice. A second approach is to predict the active biological pathways at each location. Finally, although the main focus will be on the analysis of transcriptomics data of each tissue section, spatial transcriptomics data samples also typically include an image of the tissue. A third possible approach therefore is to analyze the image data of each tissue and assign additional annotations based on features detected in the images.

b. Tissue microenvironment analysis

Even within a single tissue sample, there is a variety of substructures or microenvironments. Depending on the surrounding microenvironment, even cells of the same cell type can be in drastically different states. The exploration of such microenvironments in spatial genomics data is still very limited. In fiscal year 2023, we will initiate the exploration of our large collection of spatial transcriptomics data to get an overview of the variety of tissue microenvironments across all tissues.

2) Start the implementation of the database beta version

The field of spatial transcriptomics is at present receiving a lot of attention, and it is possible

that other groups are constructing similar databases now. Therefore, it would be preferable to release our database as soon as possible. We decided to start the implementation of database prototypes earlier than originally planned, from January 2024. This way, we can transition more easily to the development of the beta version of the full database in fiscal year 2024. Depending on the progress, we might be able to release the database earlier than originally planned.

(2) 進捗状況

1) Preparation of the data for the database beta version

a. Visium sample data analysis

During fiscal year 2023 we collected publicly available Visium samples from the Gene Expression Omnibus (GEO) database of the National Center for Biotechnology Information (NCBI) and the 10x Genomics website. Regarding the collection of samples from GEO, we prepared a pipeline to download data that has been submitted to GEO, judge whether the data is a suitable Visium sample, and collects the necessary data (such as barcode data and image data). Samples with lacking data are excluded. Next, data of all collected samples is processed to consistent file names, and processed into a data object of the R Seurat package. A number of quality measures are calculated, such as the number of reads and the number of detected genes per spot, as well as the number of isolated spots within the tissue (i.e., spots that have no nearby neighboring spots). Based on trends we observed in the collected samples, we currently define low quality spots as spots that have no nearby neighboring spots and/or have fewer than 100 detected genes. The fraction of low-quality spots per sample is recorded, and 11 samples with >25% low quality spots were filtered out. The quality control also includes the detection and removal of duplicate samples. As of March 2024, a total of 700 Visium samples (366 human and 334 mouse samples), including 642 from GEO and 58 from the 10x Genomics website were successfully processed. This corresponds to about 1.5 million spots. Data of each sample was further normalized using the R Seurat package, and data was stored into suitable formats. To support the computation and to host the database, a server was purchased and set up in our institute.

Annotation of samples was more difficult to automate, and was therefore done manually. As far as possible, for every sample, we collected the species, tissue of origin, condition, sex, age or developmental stage, ethnicity or mouse strain, date of publication, as well as a URL of the source of the data and related scientific literature. A short description of the sample was also prepared. For annotation of the tissue of origin, we used the Uberon anatomy ontology, and for conditions the Human disease (DOID) and Mondo disease ontologies. For both tissue and condition, we assigned a broad and (where possible) a more detailed annotation. Currently, the samples include 35 unique tissues (the most frequent ones being brain, breast and liver) and 49 unique more detailed tissue annotations (including cardiac ventricle, skin epidermis, etc). Conditions include 25 unique terms (the most frequent ones being cancer and cardiovascular system disease) and 64 unique more detailed conditions (including glioblastoma, myocardial infarction, etc).

To predict the cell type composition of each spot inside each tissue sample, we used 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. We therefore collected a large number of scRNA-seq samples from the CellxGene database, and prepared 68 human and 84 mouse scRNA-seq reference datasets, covering 27 human and 17 mouse tissues. In addition, we also prepared a very broad reference for each species, covering a wide range of cell types. We used these reference datasets to predict the cell type compositions of all Visium samples, using – as far as possible – references originating from the same or similar tissues and conditions as each Visium sample.

To further improve the interpretation of tissue substructures and cell types, we estimated the activity of biological pathways in each spot of each Visium sample. For this, we used sets of genes known to be involved in biological processes, based on Gene Ontology annotations. In addition, inspection of genes or pathways with a clear, non-random pattern of activity inside the tissue section is another valuable approach to interpret spatial transcriptomics data. We therefore also predicted such spatially variable features using our singleCellHaystack approach.

Finally, after discussion and consideration, we decided that the image-based annotation of 700 images covering a wide range of tissues and conditions would require an excessive amount of time and effort from a pathologist or histologist. We are considering alternative approaches for utilizing the image data.

b. Tissue microenvironment analysis

In fiscal year 2023, we started exploratory analysis of microenvironments inside tissues. We defined microenvironments as each spot and its 6 neighboring spots in the hexagonal grid of spots in Visium samples. We averaged the gene expression of the 7 spots in each microenvironment, and grouped resulting microenvironments by their similarity in gene expression patterns. We will continue this analysis in fiscal year 2024.

2) Start the implementation of the database beta version

In the latter half of fiscal year 2023, we started implementing a first prototype of the database using the Shiny package in R, and later a second prototype using the Flask framework in Python (図 1). A technical assistant (programmer) was hired to implement the second prototype. The prototypes include the following features:

- An overview of the database is shown, including the number of samples, tissues, and conditions per species.
- The user is able to search the database for the species, tissue, condition of interest. When the user selects a sample, a preview and more detailed information is shown. A link to the source of the data and related scientific literature is shown.
- After selecting a sample of interest, several types of information can be inspected, including the image, location of the sample in the database with regard to other samples, a clustering

result of the spots in the sample, visualizations of the quality measures of the sample and a comparison of the quality with that of other samples in the database, spatially variable genes and biological pathways, and predicted cell type compositions.

- The R Shiny prototype allows users to interactively select two subsets of spots within a sample, and compare gene expression between them. It also allows users to search the entire database for similar spots.

Both prototypes are not publicly accessible, but they will be used as a foundation for the public version of the database.

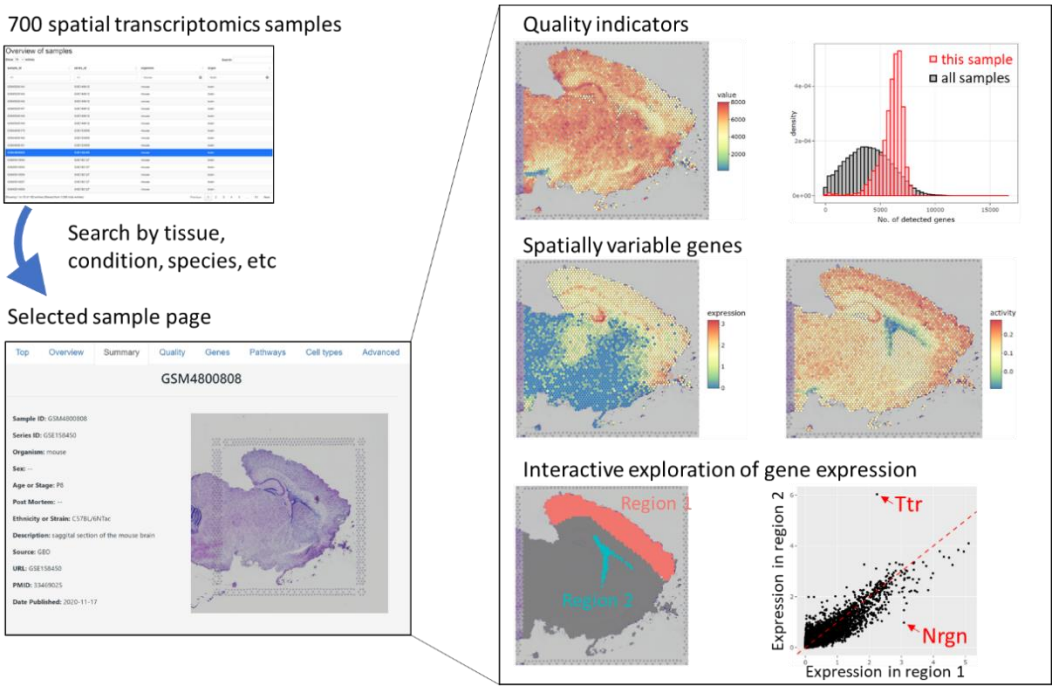

図 1. Example view of some of the prototype features as of March 2024.

§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)	0 件
	国際(欧文)	0 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)

該当なし

(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	3 件
	国際	2 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, “SingleCellHaystack: A universal differential expression prediction tool for single-cell and spatial genomics data”, Informatics in Biology, Medicine and Pharmacology 2023 (IIBMP2023), Kashiwa (online), 2023年9月7日.

Alexis Vandenbon, “Spatial transcriptomics analysis reveals how murine breast cancers disorganize the liver transcriptome in a zonated manner”, NGS EXPO 2023, Osaka, 2023年11月15日.

Alexis Vandenbon, “SingleCellHaystack: A universal tool for predicting differentially active features in single-cell and spatial genomics data”, 第46回 日本分子生物学会年会, Kobe (online), 2023年12月1日.

〈国際〉

1. Alexis Vandenbon, “SingleCellHaystack: A universal differential expression prediction tool for single-cell and spatial genomics data”, GIW ISCB-Asia, Singapore, 2023年11月19日.

④ ポスター発表

〈国内〉

1. Alexis Vandenbon and Diego Diez, “SingleCellHaystack: A universal differential expression prediction tool for single-cell and spatial genomics data”, Informatics in Biology, Medicine and Pharmacology 2023 (IIBMP2023), Kashiwa (online), 2023年9月7-8日.
2. Alexis Vandenbon, Rin Mizuno, Riyo Konishi, Masaya Onishi, Kyoko Masuda, Yuka Kobayashi, Hiroshi Kawamoto, Ayako Suzuki, Chenfeng He, Yuki Nakamura, Kosuke Kawaguchi, Masakazu Toi, Masahito Shimizu, Yasuhito Tanaka, Yutaka Suzuki and Shinpei Kawaoka, “Spatial transcriptomics analysis reveals how murine breast cancers disorganize the liver transcriptome in a zonated manner”, NGS EXPO 2023, Osaka, 2023年11月16 日.
3. Alexis Vandenbon and Diego Diez, “SingleCellHaystack: A universal tool for predicting differentially active features in single-cell and spatial genomics data”, 第46回日本分子生物学会年会, Kobe (online), 2023年12月6日.

〈国際〉

1. Alexis Vandenbon, Rin Mizuno, Riyo Konishi, Masaya Onishi, Kyoko Masuda, Yuka Kobayashi, Hiroshi Kawamoto, Ayako Suzuki, Chenfeng He, Yuki Nakamura, Kosuke Kawaguchi, Masakazu Toi, Masahito Shimizu, Yasuhito Tanaka, Yutaka Suzuki and Shinpei Kawaoka, “Spatial transcriptomics analysis reveals how murine breast cancers disorganize the liver transcriptome in a zonated manner”, 14th International Workshop on Advanced Genomics (14AGW), Tokyo, 2023年10月4日.
2. Alexis Vandenbon, “SingleCellHaystack: A universal differential expression prediction tool for single-cell and spatial genomics data”, GIW ISCB-Asia, Singapore, 2023年11月19日.

(4) 知的財産権の出願 （国内の出願件数のみ公開）

出願件数

種別		件数
特許出願	国内	0 件

(5) 受賞・報道等

① 受賞

該当なし

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

公開前

② 分析

公開前

(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

公開前

(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）

公開前

(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）

公開前

§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
2023 年 10 月 1 日～ 2024 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
2023 年 4 月 7 日	Meeting with advisor(非公開)	オンライン	2人	研究に関する意見交換
2023 年 4 月 19 日	Meeting with advisor(非公開)	オンライン	2人	同上
2023 年 5 月 12 日	Meeting with advisor(非公開)	オンライン	2人	同上
2023 年 6 月 28 日	Meeting with advisor(非公開)	オンライン	2人	同上
2023 年 7 月 4 日	Meeting with advisor(非公開)	オンライン	2人	同上
2023 年 8 月 28 日	Meeting with advisor(非公開)	オンライン	2人	同上
2023 年 9 月 14 日	Meeting with advisors(非公開)	オンライン	3人	同上
2023 年 9 月 27 日	Meeting with advisor(非公開)	オンライン	2人	同上
2023 年 10 月 4 日	Meeting with advisor and expert (非公開)	14AGW (東京)	3人	今後の発展についての議論
2023 年 10 月 17 日	Meeting with advisors(非公開)	オンライン	4人	研究に関する意見交換
2023 年 12 月 1 日	Meeting with advisor(非公開)	オンライン	2人	同上
2023 年 12 月 28 日	Meeting with advisors(非公開)	オンライン	4人	同上
2024 年 1 月 22 日	Meeting with advisor(非公開)	オンライン	2人	同上
2024 年 2 月 5 日	Meeting with advisor(非公開)	オンライン	2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年2月22日	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