

AJACS「シングルセルRNA-seqを知って・学んで・使う」
(2024年12月23日・オンライン)

公共データベースから シングルセルRNA-seqデータを取得する

理化学研究所 生命医科学研究センター(IMS)
生命医科学大容量データ技術研究チーム
チームリーダー

粕川 雄也

講義内容の概要

• 目的

- シングルセルRNA-seqのデータをまだ持っていない人や、取得前に解析の練習をしたい人が、公開リソースからデータを入手できるようになること
- その他、シングルセルRNA-seqに関する情報を入手できる場所を知ること

• 内容

- シングルセルRNA-seqの主なデータ形式
- シングルセルRNA-seqが入手できるリポジトリ
- ダウンロードしたファイルを解析ソフトウェアで使う
- シングルセルRNA-seq解析に有用なウェブサイト

シングルセルRNA-seqの主なデータ形式

シングルセルRNA-seqの主な公開データ形式

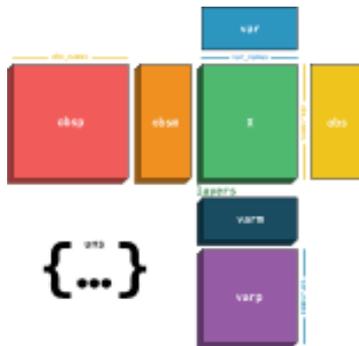
- FASTQ形式 – 配列ならびにそのクオリティスコアが含まれるファイル

```
@SRR9291388.1 K00125:97:HLHLYBBXX:3:1101:2980:998 length=26
NGCACCTAGTCTCAACGTTCTACCAA
+SRR9291388.1 K00125:97:HLHLYBBXX:3:1101:2980:998 length=26
#AAFFJJJJFFJJJJJJJJJJJJFJJ
```

- BAM形式 – ゲノムへのマッピング結果のファイル

```
A00228:279:HFVFDVDMXX:1:1201:20754:18396 272 1 11274 0 91M * 0 0
TGGTGGCCAGCGCCCCCTGCTGGCGCCGGGGCACTGCAGGGCCCTCTTGCTTACTGTATAGTGGTGGCAGCCGCCTGCTGGCAGCTAGGG
F:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
NH:i:5 HI:i:4 AS:i:89 nM:i:0 RE:A:I li:i:0 BC:Z:CAGCCCAC QT:Z:FFFFFFF,
CR:Z:ATTCGTTGTGTAGCAG CY:Z:FFFFF:FFFFF:FFF CB:Z:ATTCGTTGTGTAGCAG-1 UR:Z:CCGTTCCGTTGG
UY:Z:F:F:FF:FFF:F UB:Z:CCGTTCCGTTGG RG:Z:pbmc_1k_v3:0:1:HFVFDVDMXX:1
```

- H5AD形式 – 遺伝子ごとの発現量テーブル、細胞アノテーション、次元削減(クラスタリング)の結果が含まれるファイル



<https://anndata.readthedocs.io/>

シングルセルRNA-seqが入手できるリポジトリ

INSDC (International Nucleotide Sequence Database Collaboration)
(DDBJ / NCBI / EMBL-EBI)

INSDC (DDBJ / NCBI / EMBL-EBI)

- 新規シーケンスを用いた論文を発表するときには、そのシーケンスデータを公開することが原則的に義務化されている(個人情報関連については例外等もあり)
- その登録先のリポジトリを DDBJ (日本)、NCBI(米国)、EBI(欧州)が運営しており、誰でもデータを入手可能である

	DDBJ	NCBI	EMBL-EBI
機能ゲノムクスデータ	GEA	GEO	ArrayExpress
シーケンスデータ	DDBJ SRA (DRA)	SRA	ENA
サンプル情報	BioSample	BioSample	BioSamples
プロジェクト情報	BioProject	BioProject	BioProject

INSDC - アクセション番号

公開データの参照に使われる「アクセション番号」

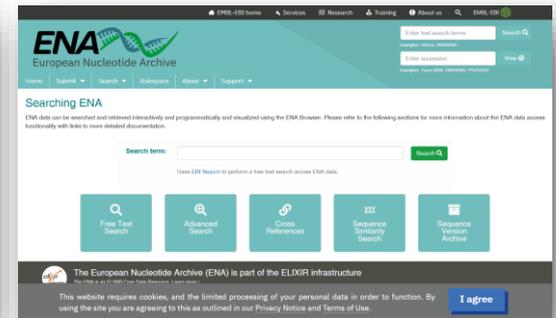
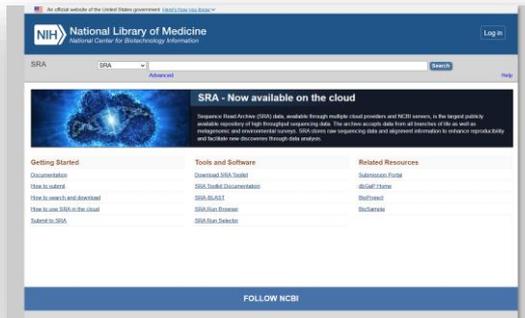
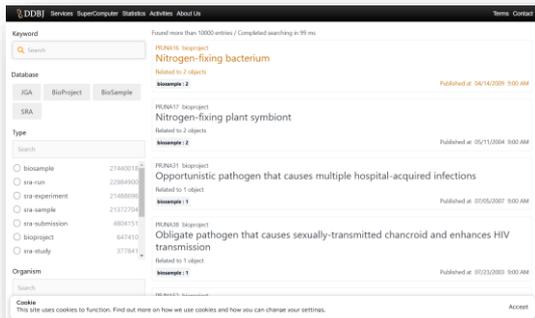
<https://ena-docs.readthedocs.io/en/latest/submit/general-guide/accessions.html>

- BioProject: “PRJ” + “E/N/D” + (アルファベット1文字) + (数字)
- SRA/DRA/ENAの “submission”: “E/D/S” + “RA” + (数字)
- SRA/DRA/ENAの “Run”: “E/D/S” + “RR” + (数字)
- NCBI GEO: “GSE” + (数字)
 - ArrayExpress上では “E-GEOD” + (数字)
- ArrayExpress: “E-MTAB-” + (数字)
- DDBJ GEA: “E-GEAD” + (数字)

INSDC - ウェブサイトからダウンロード

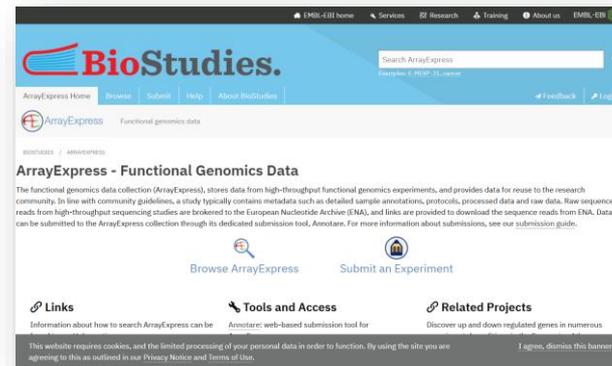
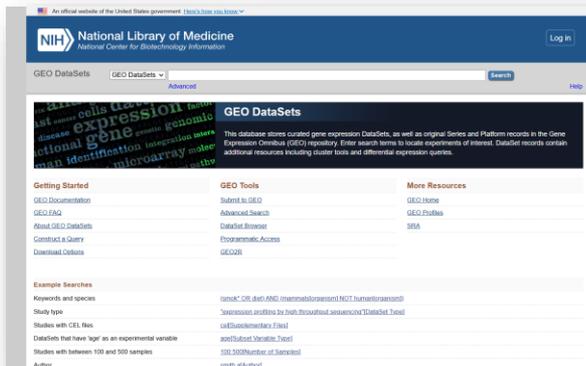
- BioProject や SRA/DRA/ENA のアクセッション番号

- DDBJ DRA: <https://ddbj.nig.ac.jp/search/>
- NCBI SRA: <https://www.ncbi.nlm.nih.gov/sra/>
- EMBL ENA: <https://www.ebi.ac.uk/ena/browser/search>



- NCBI GEOやArrayExpressionのアクセッション番号

- NCBI GEO: <https://www.ncbi.nlm.nih.gov/gds/>
- ArrayExpress: <https://www.ebi.ac.uk/biostudies/arrayexpress>



INSDC - ダウンロードツールの利用

- SRAToolkit の `fasterq-dump` コマンドを使って、直接FASTQファイルをダウンロードすることも可能
 - 欲しいデータの SRA/DRA/ENAの“Run”アクセッション番号のリスト
 - BioProjectやSRAの他のアクセッション番号しか分からない場合は、ウェブサイト検索で取得するか、Entrez Direct がインストールされていれば、以下で取得可能

```
$ esearch -db sra -query (アクセッション番号) | efetch -format runinfo |  
cut -f 1 -d ',' ↵
```

- `fasterq-dump` コマンドの基本的な書き方は以下のとおり

```
$ fasterq-dump -S --include-technical (アクセッション番号) ↵
```

インストール方法について

[SRA Toolkit] <https://github.com/ncbi/sra-tools/wiki/02.-Installing-SRA-Toolkit>

[Entrez Direct] <https://www.ncbi.nlm.nih.gov/books/NBK179288/>

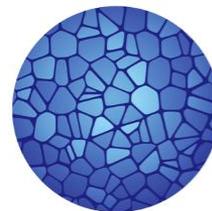
Debian/Ubuntu ではapt から、pip/biocondaでもインストール可能なものもあり

シングルセルRNA-seqが入手できるリポジトリ

Human Cell Atlas

Human Cell Atlas

- ヒトのすべての細胞のリファレンスマップの構築を目指す、国際共同プロジェクト
- 現在は肝臓や脂肪組織など18種類の「biological network」を対象にトランスクリプトームベースのリファレンスマップ (V1 Atlas) を構築中
 - その後、マルチオミックスのデータや、空間トランスクリプトームのデータへ拡張予定
- 2024年11月に first draft atlas に関する一連の論文が発表
 - <https://www.nature.com/collections/jccbbdahji>
- HCAのウェブサイト: <https://www.humancellatlas.org/>



**HUMAN
CELL
ATLAS**

Human Cell Atlas – データプラットフォーム

- 3種類のデータプラットフォームがある(最近再構成された模様)
 - CellxGene Discover: <https://cellxgene.cziscience.com/collections>
 - 遺伝子発現プロファイル+データセットに関する最小限のメタデータ
 - HCA Data Repository:
<https://explore.data.humancellatlas.org/projects>
 - FASTQファイル+サンプルに関する詳細なメタデータ(reference atlas構築に必要な詳細情報)
 - 今後は制限アクセス(ヒト由来サンプル等)用に使われる
 - Cell Annotation Platform: <https://celltype.info/>
 - 遺伝子発現プロファイル+細胞に関する詳細情報(細胞種等)

Platform	CellxGene Discover	HCA Data Repository	Cell Annotation Platform (CAP)
File format	Matrix - AnnData (h5ad)	FASTQ	Matrix - AnnData (h5ad)
Metadata	Tier 1 metadata Technical information required for integration	Tier 2 metadata Biological information required for downstream analysis & protected by managed access	Cell annotation metadata Metadata related to cell nomenclature, captured mostly on a 'per cell' basis

Human Cell Atlas – CellxGene discover (1)

<https://cellxgene.cziscience.com/collections>

3. データの選択

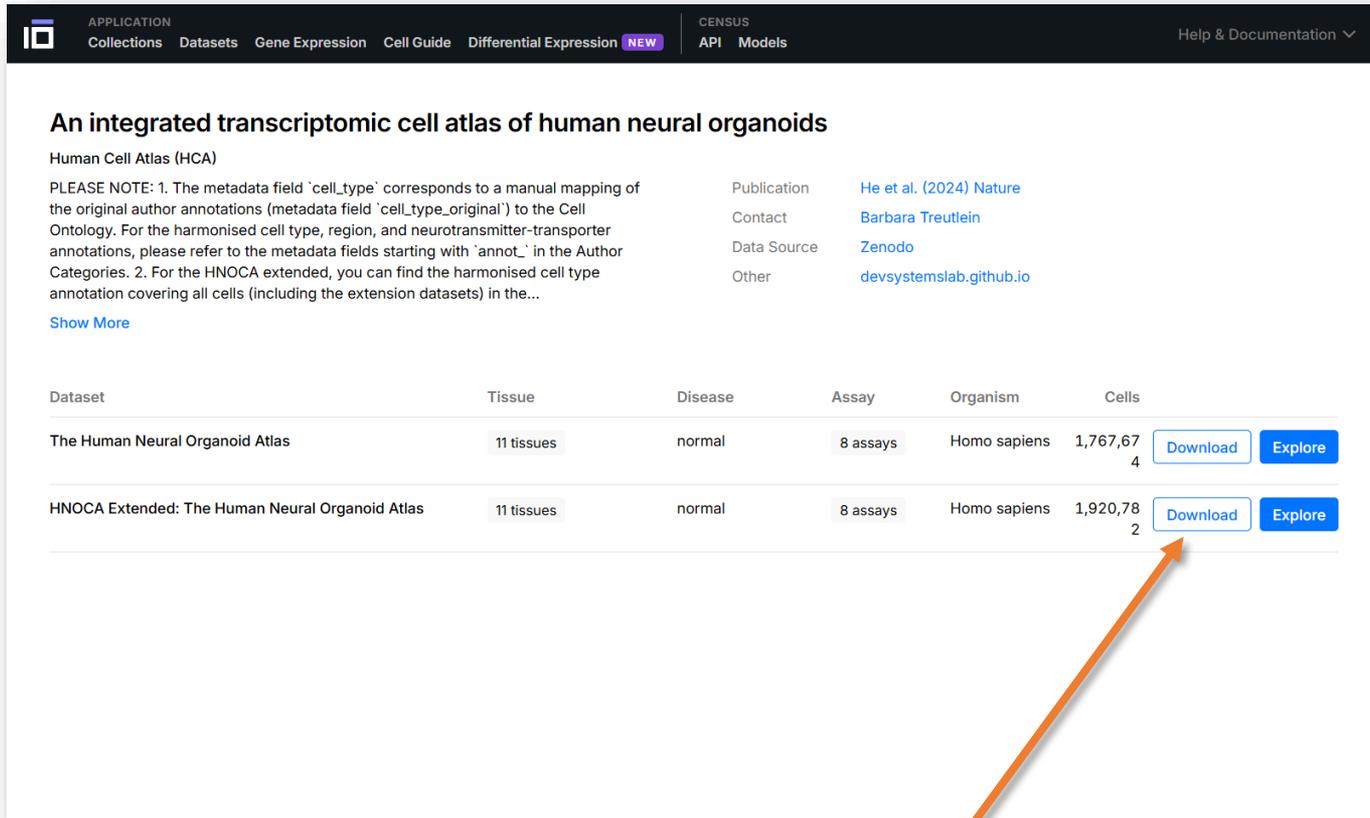
The screenshot shows the CellxGene Collections page. On the left, there is a 'Filters' sidebar with various dropdown menus. The main area displays a table of collections. An orange box highlights the 'Filters' sidebar, and another orange box highlights the table. An arrow points from the '3. データの選択' label to the table.

Collections	Publication	Tissue	Disease	Organism
Single cell transcriptomic analyses of the dynamic local and systemic response to cardiac injury in mice and zebrafish.	Cortada et al. (2024) Commun Biol	5 tissues	myocardial infarction normal	Mus musculus
Multimodal single-cell profiling reveals neuronal vulnerability and pathological cell states in focal cortical dysplasia	Galvão et al. (2024) iScience	4 tissues	isolated focal cortical dysplasia type II	Homo sapiens
A multi-region single nucleus transcriptomic atlas of Parkinson's disease	N. M. et al. (2024) Sci Data	5 tissues	normal Parkinson disease	Homo sapiens
A spatial human thymus cell atlas mapped to a continuous tissue axis	Yan et al. (2024) Nature	thymus	normal	Homo sapiens
An integrated transcriptomic cell atlas of human neural organoids	He et al. (2024) Nature	11 tissues	normal	Homo sapiens
Exploring the Utility of snRNA-seq in Profiling Human Bladder Tissue: A Comprehensive Comparison with scRNA-seq	No publication	urinary bladder	normal	Homo sapiens
Human and mouse dermal fibroblast atlas	No publication	15 tissues	normal 3 diseases	Homo sapiens Mus musculus
Molecular Signatures of Resilience to Alzheimer's Disease in Neocortical Layer 4 Neurons	Dharshini et al. (2024) bioRxiv	3 tissues	Alzheimer disease normal	Homo sapiens

1. データの選択条件の指定

2. 条件に合致する候補データの一覧

Human Cell Atlas – CellxGene discover (2)



The screenshot shows the CellxGene discover interface. At the top, there is a navigation bar with 'APPLICATION' (Collections, Datasets, Gene Expression, Cell Guide, Differential Expression **NEW**) and 'CENSUS' (API, Models) sections. The main heading is 'An integrated transcriptomic cell atlas of human neural organoids'. Below this, there is a 'Human Cell Atlas (HCA)' section with a 'PLEASE NOTE' and a 'Show More' link. To the right, there are links for 'Publication' (He et al. (2024) Nature), 'Contact' (Barbara Treutlein), 'Data Source' (Zenodo), and 'Other' (devsystems-lab.github.io).

Dataset	Tissue	Disease	Assay	Organism	Cells		
The Human Neural Organoid Atlas	11 tissues	normal	8 assays	Homo sapiens	1,767,674	Download	Explore
HNOCA Extended: The Human Neural Organoid Atlas	11 tissues	normal	8 assays	Homo sapiens	1,920,782	Download	Explore

An orange arrow points from a text box at the bottom to the 'Download' button of the 'HNOCA Extended' dataset.

4. ダウンロードの選択

Human Cell Atlas – CellxGene discover (3)

The screenshot shows the CellxGene Discover web application. A modal window titled "Download Dataset" is open, displaying the following information:

- NAME:** HNOCA Extended: The Human Neural Organoid Atlas
- DATA FORMAT:** .h5ad (AnnData v0.10) .rds (Seurat v5)
- DOWNLOAD DETAILS:** 19082MB
- URL:** `https://datasets.cellxgene.cziscience.com/029f6418-f2d6-4c09-8d5c-92859f95cd59.h5ad` (with a "Copy" button)
- Buttons:** "Cancel" and "Download"

Background text from the interface includes: "An integrated Human Cell Atlas (HCA)", "PLEASE NOTE: 1. The original author annotation ontology. For the harmonized annotations, please refer to the HCA Ontology. For the harmonized categories, please refer to the HCA Categories. 2. For the harmonized annotation covering all cells, please refer to the HCA annotation covering all cells.", "Dataset", "The Human Neural Organoid Atlas", "HNOCA Extended: The Human Neural Organoid Atlas", "download", "Explore", "download", "Explore".

5. ダウンロード形式の指定

6. 直接ダウンロード、もしくは指定されたURLでcurl等を用いてダウンロード

Human Cell Atlas – HCA Data Repository (1)

<https://explore.data.humancellatlas.org/projects>

4. データの選択

The screenshot displays the 'HUMAN CELL ATLAS DATA EXPLORER' interface. On the left, a 'Filters' sidebar includes categories like PROJECT, DONOR, and SAMPLE with various dropdown menus. The main area is titled 'Explore Data' and shows a table of search results. A warning banner at the top states: 'Access to data explorer is limited for unauthorized users. To view all available data please contact the data manager.' The table has columns for Project Title, Access, Biological Network, and other details. The 'Access' column shows 'Required' (orange) and 'Granted' (green) statuses. A callout box points to the 'Required' status, explaining that it indicates restricted access and that a procedure is needed to obtain the data, while 'Granted' indicates that the data is freely available.

Project Title	Access	Biological Network	Other Info
A temporal cortex cell atlas highlights gene expression dynamics during human brain maturation	Required	Inspecifie...	Homo sapiens brain 2 library
Single-cell connectomic analysis of adult mammalian lungs	Granted	Lung	Homo sapiens lung 10x 3' v2
Myofibroblast transcriptome indicates SFRP2hi fibroblast progenitors in systemic sclerosis skin	Granted	Skin	Homo sapiens skin of body 2 library
Evolution of neuronal cell classes and types in the vertebrate retina	Granted	Eye	Homo sapiens eye 10x 3' v2

3. アクセス方法の情報
“Required”のものは制限アクセスデータのため入手に手続き必要
“Granted”は自由に入手可能

1. データの選択条件の指定

2. 条件に合致する候補データの一覧

Human Cell Atlas – HCA Data Repository (2)

The screenshot displays the 'Single-cell connectomic analysis of adult mammalian lungs' dataset page. The 'Download' tab is selected, showing options for file type and shell. The 'fastq.gz' file type is selected, and the 'Request curl Command' button is visible. The 'Selected Data Summary' panel on the right provides details about the data, including estimated cells, file size, and number of files.

5. Downloadタブを選択

6. 必要なファイルを選択

7. ダウンロード用のコマンドを発行

File Type	Count	File Size
<input checked="" type="checkbox"/> fastq.gz	6	33.98 GB
<input type="checkbox"/> mtz.gz	14	113.80 MB
<input type="checkbox"/> tsv.gz	28	4.10 MB

Category	Value
Estimated Cells	17.9k
File Size	33.98 GB
Files	6
Projects	1
Species	Homo sapiens
Donors	2

Human Cell Atlas – HCA Data Repository (3)

The screenshot displays the Human Cell Atlas Data Explorer interface. The main heading is "Single-cell connectomic analysis of adult mammalian lungs". Below the heading, there are tabs for "Overview", "Metadata", "Matrices", "Download", and "Export". The "Download" tab is active. A section titled "Your curl Command is Ready" provides instructions on how to execute the curl command in a terminal. A code block contains the following command:

```
curl --location --fail
https://service.azul.data.humancellatlas.org/manifest/files/ksQw1KVkY3A0M6RjdX.JsxBAyo8LHXyVdMr0X6jT-
rnFxxBA-zvMrA5JWILycxVob0Zx1xCAkxQ9Y5sitC6a0waJf-MuVKncGs-nZUuDwKkHFjnZ9nQ | curl --fail-early --
continue-at - --retry 15 --retry-delay 10 --config -
```

To the right of the code block, there is a "Current Query" section with the following details:

- Project: MammalLungConnectomeRareDon
- Genus Species: Homo sapiens
- File Format: fastq.gz

Below that is a "Selected Data Summary" section:

- Estimated Cells: 17.9k
- File Size: 33.98 GB
- Files: 6
- Projects: (partially visible)

An orange arrow points from the code block to a callout box at the bottom of the page.

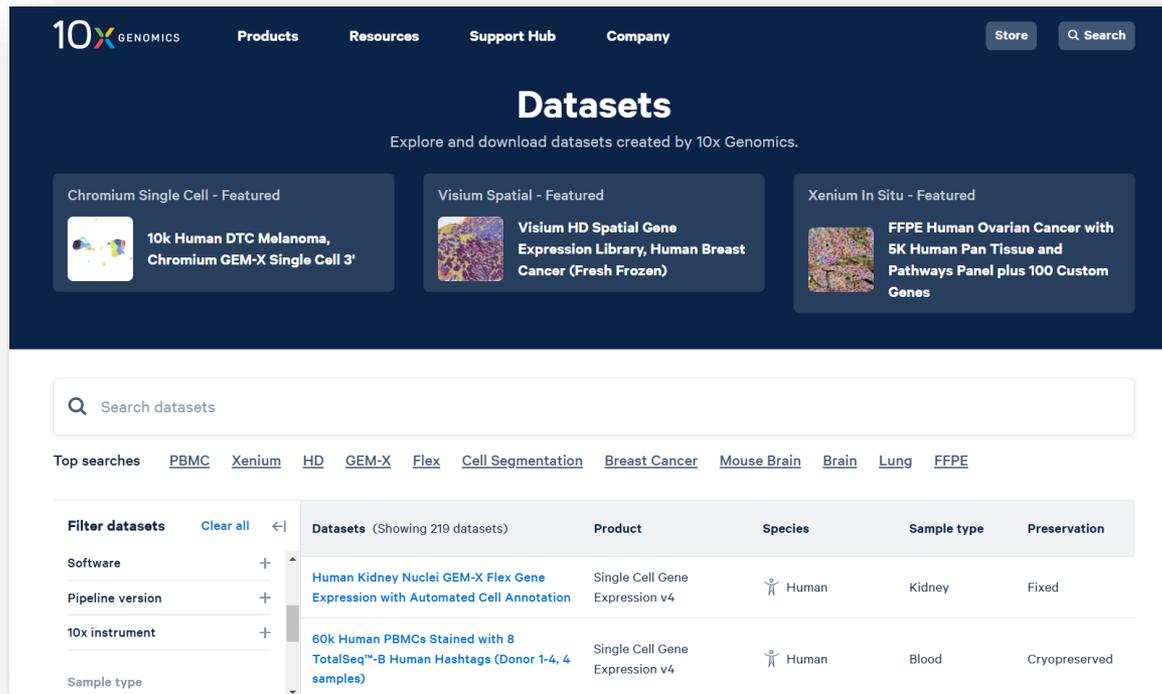
8. ダウンロード用の(curlの)コマンド

シングルセルRNA-seqが入手できるリポジトリ

10x Genomics社の Dataset サイト

10x Genomics社の Dataset サイト

- 10x Genomics 社のウェブサイトで、10x Genomics社で取得したデータが公開されている
 - <https://www.10xgenomics.com/datasets>
 - Creative Commons Attribution licenseで公開
- 何種類かのサンプルに対して、10x Genomics 社の様々なプロトコルで生産したデータが入手



The screenshot shows the 10x Genomics Datasets website. The header includes the 10x Genomics logo, navigation links for Products, Resources, Support Hub, and Company, and buttons for Store and Search. The main heading is "Datasets" with the subtitle "Explore and download datasets created by 10x Genomics." Below this, three featured datasets are displayed: Chromium Single Cell (10k Human DTC Melanoma, Chromium GEM-X Single Cell 3'), Visium Spatial (Visium HD Spatial Gene Expression Library, Human Breast Cancer (Fresh Frozen)), and Xenium In Situ (FFPE Human Ovarian Cancer with 5K Human Pan Tissue and Pathways Panel plus 100 Custom Genes). A search bar is present, and below it are top search categories like PBMC, Xenium, HD, GEM-X, Flex, Cell Segmentation, Breast Cancer, Mouse Brain, Brain, Lung, and FFPE. A filter sidebar on the left allows filtering by software, pipeline version, 10x instrument, and sample type. The main content area shows a table of datasets with columns for Datasets, Product, Species, Sample type, and Preservation.

Filter datasets	Clear all	←	Datasets (Showing 219 datasets)	Product	Species	Sample type	Preservation
Software	+	▲	Human Kidney Nuclei GEM-X Flex Gene Expression with Automated Cell Annotation	Single Cell Gene Expression v4	Human	Kidney	Fixed
Pipeline version	+	▲	60k Human PBMCs Stained with 8 TotalSeq™-B Human Hashtags (Donor 1-4, 4 samples)	Single Cell Gene Expression v4	Human	Blood	Cryopreserved
10x instrument	+	▲					
Sample type		▼					

10x Genomics社の Dataset サイト (1)

<https://www.10xgenomics.com/datasets>

3. データの選択

The screenshot shows the 10x Genomics datasets website interface. At the top, there is a search bar labeled "Search datasets". Below it, there are "Top searches" for various categories like PBMC, Xenium, HD, GEM-X, Flex, Cell Segmentation, Breast Cancer, Mouse Brain, Brain, Lung, and FFPE. On the left, a "Filter datasets" sidebar is visible with various filter options such as "10x Genomics product", "Platform", "Product", "Chemistry version", "Additional application", "Software", "Pipeline version", "10x instrument", "Sample type", "Species", "Sample/tissue type", "Preservation method", "Disease state", and "Cells or nuclei". The main content area displays a table of datasets with columns for "Datasets (Showing 219 datasets)", "Product", "Species", "Sample type", and "Preservation". An orange arrow points from the "3. データの選択" label to the "5k Human PBMCs (Donor 4)" entry in the table.

Datasets (Showing 219 datasets)	Product	Species	Sample type	Preservation
Human Kidney Nuclei GEM-X Flex Gene Expression with Automated Cell Annotation	Single Cell Gene Expression v4	Human	Kidney	Fixed
60k Human PBMCs Stained with 8 TotalSeq™-B Human Hashtags (Donors 1-4, 4 samples)	Single Cell Gene Expression v4	Human	Blood	Cryopreserved
5k Human PBMCs (Donor 4)	Single Cell Gene Expression v4	Human	Blood	Cryopreserved
5k Human PBMCs (Donor 3)	Single Cell Gene Expression v4	Human	Blood	Cryopreserved
5k Human PBMCs (Donor 2)	Single Cell Gene Expression v4	Human	Blood	Cryopreserved
5k Human PBMCs (Donor 1) with Automated Cell Annotation	Single Cell Gene Expression v4	Human	Blood	Cryopreserved
20k Human PBMCs Multiplex Sample (Donors 1-4)	Single Cell Gene Expression v4	Human	Blood	Cryopreserved
10k Human Diseased PBMCs (Myelofibrosis) Freshly Processed	Single Cell Gene Expression v4	Human	Blood	Fresh

1. データの選択条件・キーワードの指定

2. 条件に合致する候補データの一覧

10x Genomics社の Dataset サイト (2)

4. 該当のタブを選択

“Input files”: FASTQファイルやサンプル情報のファイル

“Output and Supplemental files”: CellRanger等のプログラムの出力ファイル

10x GENOMICS Products Res

< All datasets

5k Human PBMCs (Donor 4)

Single Cell Gene Expression dataset analyzed using Cell Ranger 9.0.0

Mapped **95.7%**

Assess data quality

View summary metrics to assess data quality and more.

Visualize and explore data

Discover differentially expressed genes, visualize your favorite genes, and explore your data with our visualization software.

View summary

Explore data

Dataset overview

Output and supplemental files

Input files

Learn about Chromium analysis

Batch download

File type	Size	md5sum
TAR	17.3 GB	97f56ba33471572d4d7dc98830180a24
CSV	243 B	c56cfc73b177221de7d70df6e34f5a6b

FASTQs

Multi Config CSV

5. 該当ファイルの選択

ダウンロードしたファイルを解析ソフトウェアで使う

ダウンロードしたFASTQデータを CellRangerで使う

ダウンロードしたFASTQデータを CellRangerで使う

- 必要となるFASTQファイル

- シングルセルRNA-seqの場合は、基本的に R1 (read1) と R2 (read2)のファイルがあればよい(ただし、3' v1 chemistry を使っている場合は I1 (index1) ファイルが必要)

10x Genomics Library: Chemistry version	Recommended	Alternative
Gene Expression*: 3' v3.1, 3' LT v3.1, 3' v3, 3' v2, 5' v2, 5' v1.1, 5' v1	R1 and R2 FASTQs	10x BAM
Visium FF & FFPE	1. R1 and R2 FASTQs; 2. Image (submitted as processed data file)	1. 10x BAM; 2. Image (submitted as processed data file)
Feature Barcode: Cell Surface Protein, CRISPR Screening	R1 and R2 FASTQs	10x BAM
3' CellPlex**	10x per-sample BAM (see here)	N/A
5' TCR/BCR***: v2, v1.1, v1	R1 and R2 FASTQs	N/A
Multiome Gene Expression	R1 and R2 FASTQs	10x BAM
Multiome ATAC****	10x BAM	N/A
ATAC****: v1.1, v1	10x BAM	N/A
Chromium Genome, Single Cell DNA	R1 and R2 FASTQs	10x BAM

*Note that for 3' v1 chemistry, uploading only R1 and R2 FASTQ files is NOT sufficient for others to reproduce the analysis. In this version, 10x Genomics BAM file is the recommended file.

<https://kb.10xgenomics.com/hc/en-us/articles/360024716391-What-format-of-10x-Genomics-data-should-I-submit-to-NCBI-GEO-SRA>

ダウンロードしたFASTQデータを CellRangerで使う

- 実験chemistryごとのread長の構成

	Read1 (R1)	i7 index (I1)	i5 index (I2)	Read2 (R2)
3' v4	<u>28</u>	10	10	<u>Insert >=90</u>
3' v3 DL	<u>28</u>	10	10	<u>Insert >=90</u>
3' v3/v3.1	<u>28</u>	8	-	<u>Insert >=91</u>
3' v2	<u>26</u>	8	-	<u>Insert >=98</u>
3' v1	<u>Insert >=98</u>	<u>14 (R3)</u>	8 (I1)	<u>10</u>
5' v3	<u>28(+TSO+Insert)</u>	10	10	<u>Insert >=90</u>
5' v2 DL	<u>28(+TSO+Insert)</u>	10	10	<u>Insert >=90</u>
5' v1/v1.1	<u>26(+TSO+Insert)</u>	8	-	<u>Insert >=91</u>

ダウンロードして得られた複数のFASTQファイルのうち、
どれが Read1 (R1)、Read2 (R2)、i7 index (I1) をread長で推定する

ダウンロードしたFASTQデータを CellRangerで使う

• 処理の流れ(例: SRR9291388)

1. FASTQファイルをダウンロード

```
$ fasterq-dump -S --include-technical SRR9291388
```

以下の3つのファイルが生成される

- SRR9291388_1.fastq
- SRR9291388_2.fastq
- SRR9291388_3.fastq

2. read長の確認をする

```
$ head -1 SRR9291388_*.fastq
```

```
==> SRR9291388_1.fastq <==  
@SRR9291388.1 K00125:97:HLHLYBBXX:3:1101:2980:998 length=26  
  
==> SRR9291388_2.fastq <==  
@SRR9291388.1 K00125:97:HLHLYBBXX:3:1101:2980:998 length=98  
  
==> SRR9291388_3.fastq <==  
@SRR9291388.1 K00125:97:HLHLYBBXX:3:1101:2980:998 length=8
```

ダウンロードしたFASTQデータを CellRangerで使う

- 処理の流れ(例: SRR9291388)

3. ファイルとReadの種類に対応づけと、ファイル名の変更

ダウンロードファイル名	Readの種類	CellRanger用ファイル名
SRR9291388_1.fastq	R1	SRR9291388_S1_L001_R1_001.fastq
SRR9291388_2.fastq	R2	SRR9291388_S1_L001_R2_001.fastq
SRR9291388_3.fastq	I1	SRR9291388_S1_L001_I1_001.fastq (なくてもよい)

CellRanger用のFASTQファイル名に変更する必要がある

“[Sample Name]_S1_L00[Lane Number]_[Read Type]_001.fastq.gz”

(CellRanger 4.0以降は L00[Lane Number] は省略可能)

ダウンロードしたFASTQデータを CellRangerで使う

- 処理の流れ(例: SRR9291388)

4. CellRanger の実行

```
$ cellranger count --id=(任意のID) --create-bam=true  
--transcriptome=(reference transcriptomeファイルのあるディレクトリ)  
--fastqs=(3.用意したFASTQファイルのあるディレクトリ) ↵
```

実行例:

```
$ cellranger count --id=SRR9291388 --create-bam=true  
--transcriptome=/data/refdata-gex-GRCh38-2024-A/  
--fastqs=$HOME/work/fastqs/ ↵
```

Chemistry は自動判定されるが、判定に問題がある場合は “--chemistry” オプションで明示的に指定する

CellRangerプログラムと、reference transcriptome の設定は完了しているものとする

ダウンロードしたファイルを解析ソフトウェアで使う

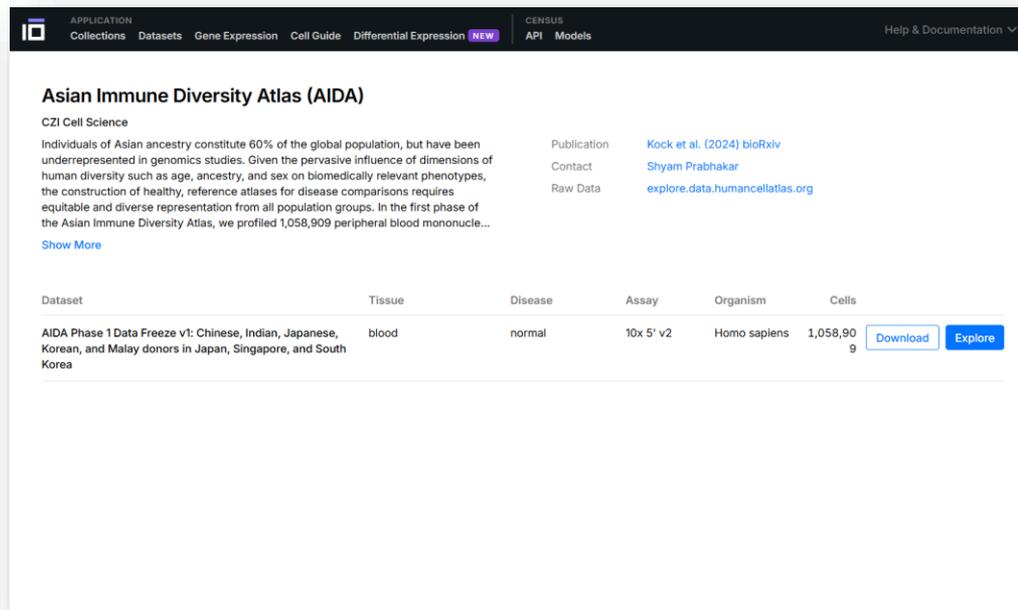
ダウンロードしたH5ADファイルを Scanpy で使う

ダウンロードしたH5ADファイルを Scanpy で使う

• 処理の流れ (Asian Immune Diversity Atlasのデータ)

1. H5AD形式のファイルのダウンロード

<https://cellxgene.cziscience.com/collections/ced320a1-29f3-47c1-a735-513c7084d508>



The screenshot shows the Cellxgene website interface for the Asian Immune Diversity Atlas (AIDA) dataset. The page includes a navigation bar with options like 'APPLICATION', 'Collections', 'Datasets', 'Gene Expression', 'Cell Guide', 'Differential Expression', 'CENBUS', 'API', 'Models', and 'Help & Documentation'. The main content area features the title 'Asian Immune Diversity Atlas (AIDA)' and a description: 'CZI Cell Science. Individuals of Asian ancestry constitute 60% of the global population, but have been underrepresented in genomics studies. Given the pervasive influence of dimensions of human diversity such as age, ancestry, and sex on biomedically relevant phenotypes, the construction of healthy, reference atlases for disease comparisons requires equitable and diverse representation from all population groups. In the first phase of the Asian Immune Diversity Atlas, we profiled 1,058,909 peripheral blood mononucle...'. Below the description is a 'Show More' link. To the right, there are links for 'Publication' (Kock et al. (2024) bioRxiv), 'Contact' (Shyam Prabhakar), and 'Raw Data' (explore.data.humancellatlas.org). At the bottom, there is a table with columns for Dataset, Tissue, Disease, Assay, Organism, and Cells. The table contains one row: 'AIDA Phase 1 Data Freeze v1: Chinese, Indian, Japanese, Korean, and Malay donors in Japan, Singapore, and South Korea', 'blood', 'normal', '10x 5' v2', 'Homo sapiens', and '1,058,909'. There are 'Download' and 'Explore' buttons next to the cell count.

Dataset	Tissue	Disease	Assay	Organism	Cells
AIDA Phase 1 Data Freeze v1: Chinese, Indian, Japanese, Korean, and Malay donors in Japan, Singapore, and South Korea	blood	normal	10x 5' v2	Homo sapiens	1,058,909

最終的に “2a99fd19-9a29-48c3-9d65-47467fd7cefe.h5ad” というファイルがダウンロードされる

ファイル名が長いので、以下“AIDA.h5ad” に変更して説明する

ダウンロードしたH5ADファイルを Scanpy で使う

- 処理の流れ (Asian Immune Diversity Atlasのデータ)

2. Python の scanpy パッケージをロードしデータの読み込み

```
import scanpy↵  
aida = scanpy.read_h5ad('AIDA.h5ad',backed='r')↵
```

ダウンロードしたH5ADファイルを Scanpy で使う

- 処理の流れ (Asian Immune Diversity Atlasのデータ)

3. 読み込んだデータの確認

```
aida
```

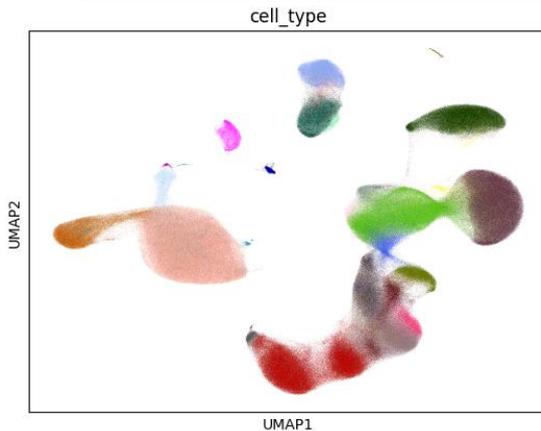
```
AnnData object with n_obs × n_vars = 1058909 × 36161 backed at 'AIDA.h5ad'  
  obs: 'mapped_reference_assembly', 'alignment_software', 'library_uuid',  
'assay_ontology_term_id', 'library_starting_quantity', 'is_primary_data',  
'cell_type_ontology_term_id', 'author_cell_type', 'sample_uuid',  
'tissue_ontology_term_id', 'development_stage_ontology_term_id',  
'sample_derivation_process', 'donor_BMI_at_collection',  
'suspension_derivation_process', 'suspension_enriched_cell_types',  
'suspension_percent_cell_viability', 'suspension_uuid', 'suspension_type',  
'donor_id', 'self_reported_ethnicity_ontology_term_id',  
'donor_living_at_sample_collection', 'organism_ontology_term_id',  
'disease_ontology_term_id', 'sex_ontology_term_id', 'Country', 'nCount_RNA',  
'nFeature_RNA', 'Ethnicity_Selfreported', 'TCR_VDJdb', 'TCRa_V_gene',  
'TCRa_D_gene', 'TCRa_J_gene', 'TCRa_C_gene', 'TCRb_V_gene', 'TCRb_D_gene',  
'TCRb_J_gene', 'TCRb_C_gene', 'TCR_Clonality', 'TCR_Clone_ID', 'BCR_VDJ_V_call',  
'BCR_VDJ_D_call', 'BCR_VDJ_J_call', 'BCR_VDJ_C_call', 'BCR_VJ_V_call',  
'BCR_VJ_J_call', 'BCR_VJ_C_call', 'BCR_Clonality', 'BCR_Clone_size',  
'BCR_mu_freq', 'tissue_type', 'cell_type', 'assay', 'disease', 'organism', 'sex',  
'tissue', 'self_reported_ethnicity', 'development_stage', 'observation_joinid'  
  var: 'feature_is_filtered', 'feature_name', 'feature_reference',  
'feature_biotype', 'feature_length', 'feature_type'  
  uns: 'citation', 'default_embedding', 'schema_reference', 'schema_version',  
'title'  
  obsm: 'X_umap'
```

ダウンロードしたH5ADファイルを Scanpy で使う

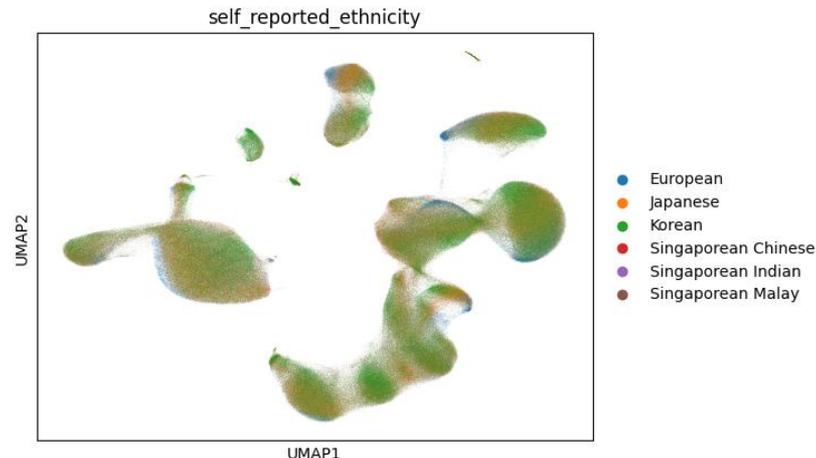
• 処理の流れ (Asian Immune Diversity Atlasのデータ)

4. scanpy上での解析(以下はデータに含まれている UMAPプロットの作成)

```
scanpy.pl.umap(aida, color='cell_type') ↵  
scanpy.pl.umap(aida, color='self_reported_ethnicity') ↵
```



- T cell
- erythrocyte
- platelet
- B cell
- dendritic cell
- monocyte
- natural killer cell
- CD4-positive, alpha-beta T cell
- CD8-positive, alpha-beta T cell
- plasmacytoid dendritic cell
- mature B cell
- plasma cell
- memory B cell
- naive B cell
- CD8-positive, alpha-beta cytotoxic T cell
- gamma-delta T cell
- regulatory T cell
- naive thymus-derived CD4-positive, alpha-beta T cell
- naive thymus-derived CD8-positive, alpha-beta T cell
- central memory CD4-positive, alpha-beta T cell
- effector memory CD4-positive, alpha-beta T cell
- CD8-positive, alpha-beta memory T cell
- CD4-positive, alpha-beta cytotoxic T cell
- CD16-negative, CD56-bright natural killer cell, human
- CD16-positive, CD56-dim natural killer cell, human
- mucosal invariant T cell
- conventional dendritic cell
- CD14-positive monocyte
- innate lymphoid cell
- CD141-positive myeloid dendritic cell
- CD14-low, CD16-positive monocyte
- CD1c-positive myeloid dendritic cell
- double negative T regulatory cell



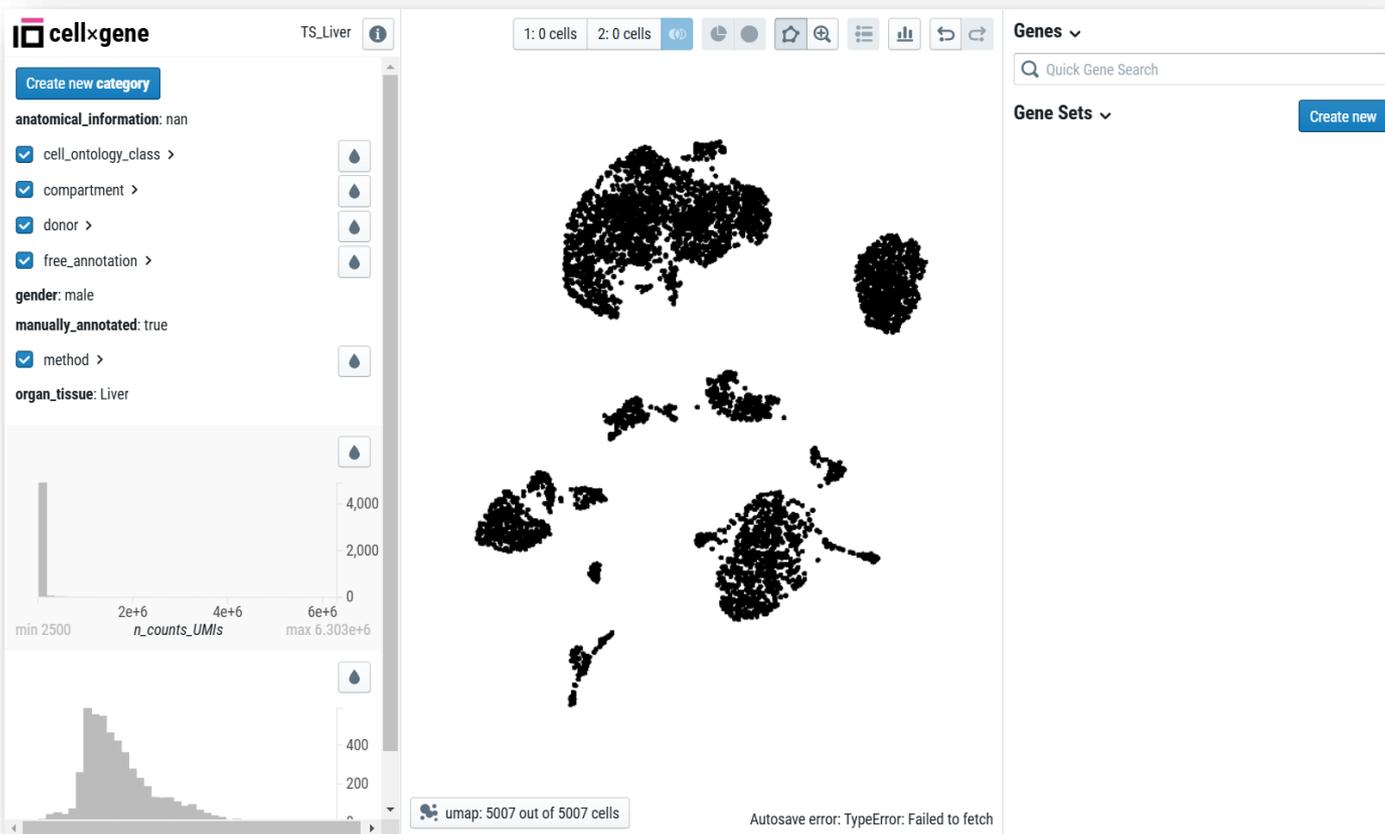
- European
- Japanese
- Korean
- Singaporean Chinese
- Singaporean Indian
- Singaporean Malay

ダウンロードしたファイルを解析ソフトウェアで使う

ダウンロードしたH5ADファイルを cellxgene で使う

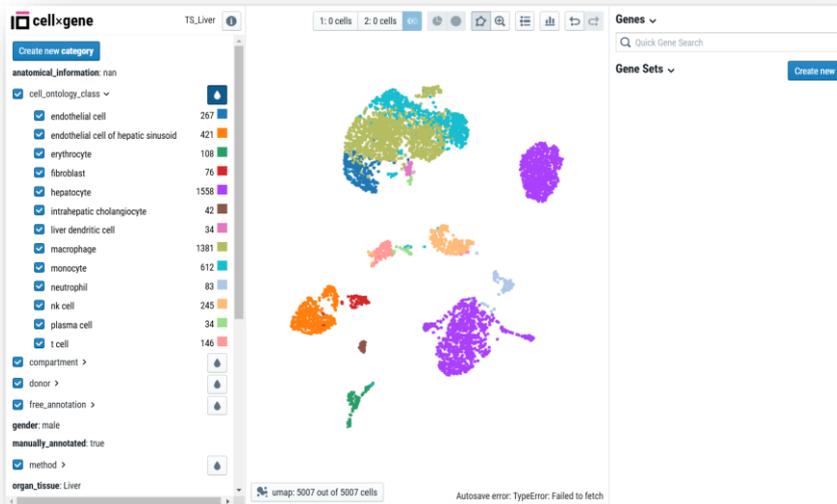
ダウンロードしたH5ADファイルを cellxgene で使う

- cellxgene (<https://github.com/chanzuckerberg/cellxgene>)
 - シングルセルRNA-seqの結果(クラスタリングや細胞アノテーション情報)を参照したり、自分で細胞へのアノテーションを付与したりできるウェブベースのツール



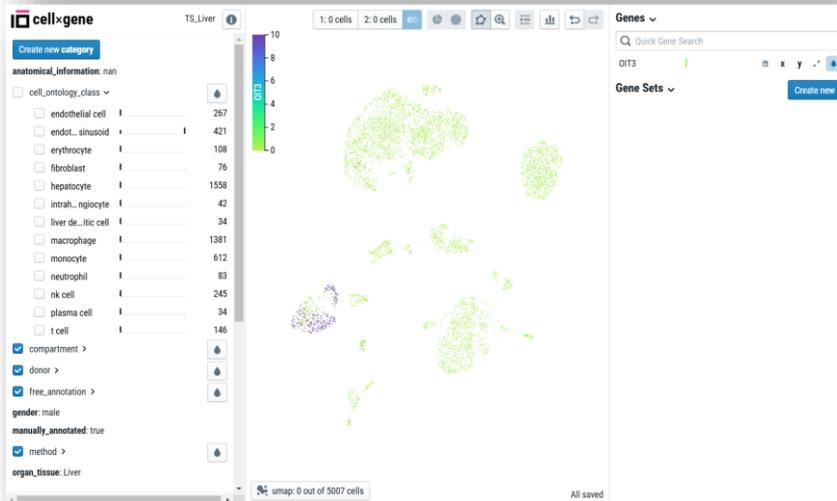
ダウンロードしたH5ADファイルを cellxgene で使う

- cellxgene (<https://github.com/chanzuckerberg/cellxgene>)



提供されている細胞アノテーションによる色付け表示

(自分で細胞アノテーション/カテゴリ分けを行うこともできる)



指定した遺伝子の発現量による色付け表示

ダウンロードしたH5ADファイルを cellxgene で使う

- cellxgeneで必要となるH5ADファイルの要件

- Matrix data (usually raw or normalized expression values) in `anndata.X`
- At least one embedding (e.g., tSNE, UMAP) in `anndata.obsm`, specified with the prefix `X_` (e.g., by default scanpy stores UMAP coordinates in `anndata.obsm['X_umap']`)
- A unique identifier is required for each cell, which by default will be pulled from the `obs` DataFrame index. If the index is not unique or does not contain the cell ID, an alternative column can be specified with `--obs-names`
- A unique identifier is required for each gene, which by default will be pulled from the `var` DataFrame index. If the index is not unique or does not contain the gene ID, an alternative column can be specified with `--var-names`

https://cellxgene.cziscience.com/docs/05__Annotate%20and%20Analyze%20Your%20Data/5_3__Preparing%20Data

ダウンロードしたH5ADファイルを cellxgene で使う

- cellxgene のインストール (pipが利用可能とする)

```
$ pip install cellxgene ↵
```

他の方法でのインストール方法については(Docker利用等)、以下を参照

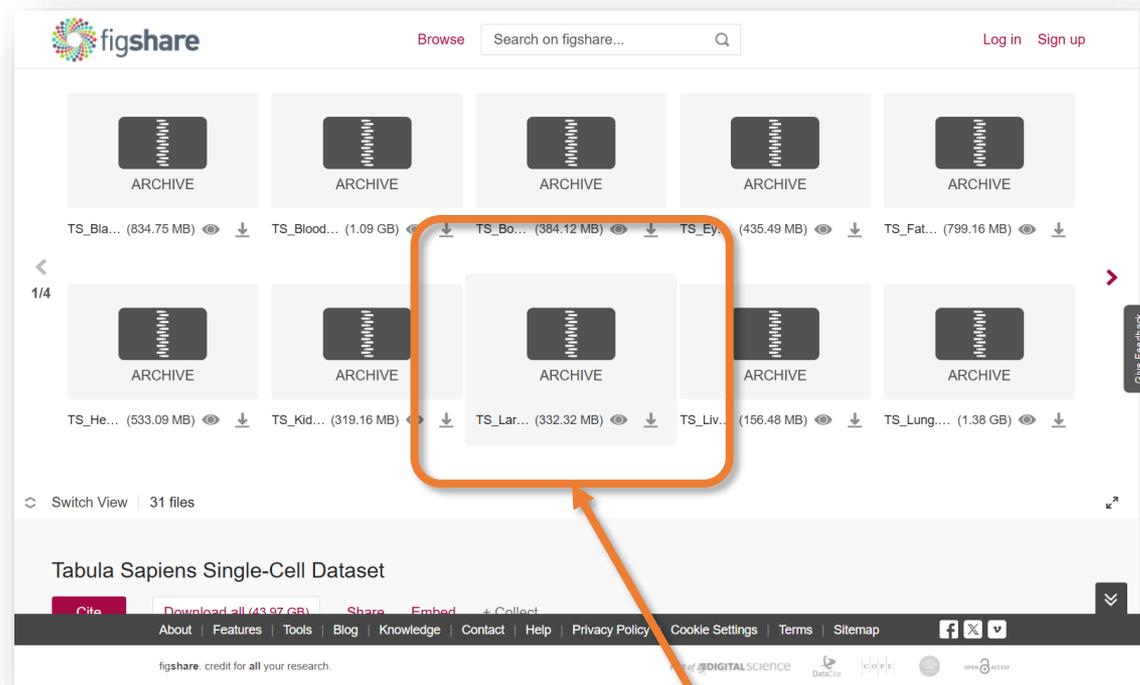
https://cellxgene.cziscience.com/docs/05__Annotate%20and%20Analyze%20Your%20Data/5_1__Getting%20Started:%20Install,%20Launch,%20Quick%20Start

ダウンロードしたH5ADファイルを cellxgene で使う

- 処理の流れ (Tabula sapiensのLiverデータ)

1. H5AD形式のファイルのダウンロード (figshareより)

https://figshare.com/articles/dataset/Tabula_Sapiens_release_1_0/14267219



下向き矢印のアイコンをクリックするとZIPファイルがダウンロードされるので unzipコマンドや、7-zip等のツールで展開すると“TS_Liver.h5ad”というファイルが作成される

ダウンロードしたH5ADファイルを cellxgene で使う

• 処理の流れ (Tabula sapiensのLiverデータ)

2. cellxgene の起動

```
$ cellxgene launch TS_Liver.h5ad
```

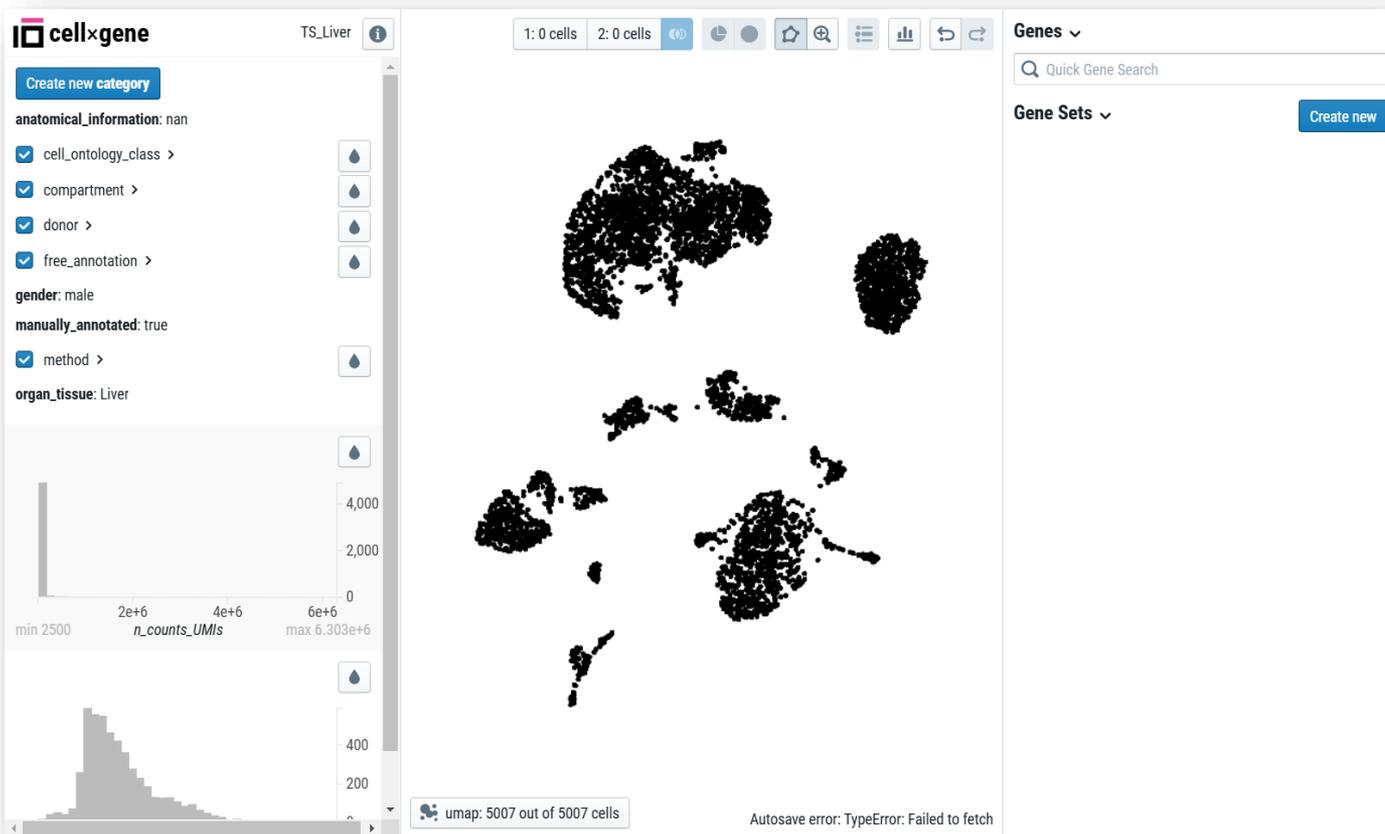
```
[cellxgene] Starting the CLI...  
[cellxgene] Loading data from TS_Liver.h5ad, this may take a while...  
[cellxgene] Warning: Anndata data matrix is sparse, but not a CSC (columnar) matrix.  
Performance may be improved by using CSC.  
[cellxgene] Warning: Var annotation 'gene_symbol' has 57316 categories, this may be  
cumbersome or slow to display. We recommend setting the --max-category-items option to  
500, this will hide categorical annotations with more than 500 categories in the UI  
WARNING:root:Type float64 will be converted to 32 bit float and may lose precision.  
WARNING:root:Type float64 will be converted to 32 bit float and may lose precision.  
WARNING:root:Type float64 will be converted to 32 bit float and may lose precision.  
WARNING:root:Type float64 will be converted to 32 bit float and may lose precision.  
[cellxgene] Launching! Please go to http://localhost:5005 in your browser.  
[cellxgene] Type CTRL-C at any time to exit.
```

このアドレスをGoogle Chrome等のウェブブラウザで開く

ダウンロードしたH5ADファイルを cellxgene で使う

• 処理の流れ (Tabula sapiensのLiverデータ)

3. ウェブブラウザでの表示



シングルセルRNA-seq解析に有用なウェブサイト

EBI – Single-cell Expression Atlas

- <https://www.ebi.ac.uk/gxa/sc/>
- 様々な生物種の公開シングルセルRNA-seqデータを収集して再解析し、例えば、遺伝子名等でデータセット横断的に検索して、マーカー遺伝子や細胞クラスターの情報などが見られるようにしたウェブサイト

The screenshot displays the EBI Single Cell Expression Atlas interface. At the top, there is a navigation bar with links for Home, Browse experiments, Download, Release notes, Help, and Support. The main header features the site logo, the title "Single Cell Expression Atlas", and the subtitle "Single cell gene expression across species". A search bar is present with the gene ID "657" entered. The results section shows that gene 657 (ENSG00000107779) is expressed in Homo sapiens. The results are presented in a table with columns for Species, Marker genes, Title, Experimental variables, and Number of assays. The table shows that the gene is expressed in Homo sapiens, with a title of "GTEx: snRNAseq atlas" and 209,126 assays. The experimental variables include organism part, sampling site, inferred cell type - authors labels, and inferred cell type - ontology labels.

EMBL-EBI Services Research Training About us

Single Cell Expression Atlas

Single cell gene expression across species

Query bulk expression
Back to Expression Atlas

Home Browse experiments Download Release notes Help Support

Gene ID or gene symbol: 657 Species: Any

Search

Organism part: Select...

657 (ENSG00000107779) is expressed in:

Species	Marker genes	Title	Experimental variables	Number of assays
Homo sapiens	<ul style="list-style-type: none">See cluster 1, 10, 11, 12, 13, 16, 2, 3, 4, 5, 7, 9 for k = 19See cluster 11 for k = 45	GTEx: snRNAseq atlas	<ul style="list-style-type: none">organism partsampling siteinferred cell type - authors labelsinferred cell type - ontology labels	209,126

This website requires cookies, and the limited processing of your personal data in order to function. By using the site you are agreeing to this as outlined in our [Privacy Notice](#) and [Terms of Use](#). [I agree, dismiss this banner](#)

Broad Institute – single-cell portal

- <https://singlecell.broadinstitute.org/>
- 公開シングルセルRNA-seqデータの再解析を行い、各データセットごとのクラスタリング結果の図を参照したり、興味のある遺伝子について細胞ごとの発現プロファイルを参照したりできるウェブサイト

The screenshot displays the Single Cell Portal interface. At the top, there is a navigation bar with the logo, 'Help', 'Create study', and 'Sign in'. Below this is a large banner with the 'Single Cell PORTAL' logo and the tagline 'Reducing barriers and accelerating single-cell research'. A circular graphic on the right of the banner states 'Featuring 789 studies' and '55,477,292 cells', with a 'New feature' badge. The main content area includes search filters for 'organ', 'species', 'disease', and 'cell type', along with a 'Search by text' input field. Below the search filters, it shows '789 total studies found' and 'Page 1 of 79'. The first study listed is 'Transcriptional profile of the rat cardiovascular system at single cell resolution' with 504,278 cells. The abstract describes an snRNA-seq dataset from a healthy Wistar rat. A table below the abstract lists study details:

Disease	Organ	Species	Sex	Library preparation protocol
normal	aorta, atrioventricular node, 8 more...	Rattus norvegicus	male	10x 3' v2

The second study listed is 'HRCA: snRNA-seq of the human retina - retinal ganglion cells' with 0 cells.

RIKEN – scPortalen 2

- <https://single-cell.riken.jp/>
- 公開シングルセルRNA-seqのメタデータの修正や再計算を行い、再利用を促進することを目的としたデータベース
- 解析結果のダウンロードやCellxGeneでの参照などが可能となっている

SCPortalen2

Toggle Sidebar

SCPortalen2
— A truly cell-centric database, from RIKEN IMS.

NEW - We have added a new Database Discovery Environment, accessible from the side-bar. Here, using common metadata metrics, you can explore publicly available scRNAseq datasets, indexed from NCBI Sequence Read Archive (SRA).

The SCPortalen Database can be cited with the main publication:

Abugessaisa, I., Noguchi, S., Böttcher, M., Hasegawa, A., Kouno, T., Kato, S., Tada, Y., Ura, H., Abe, K., Shin, J. W., Plessey, C., Corinici, P., & Kasukawa, T. (2018). **SCPortalen: Human and mouse single-cell centric database**. *Nucleic Acids Research*, 46(D1), D781–D787. <https://doi.org/10.1093/nar/gkx949>

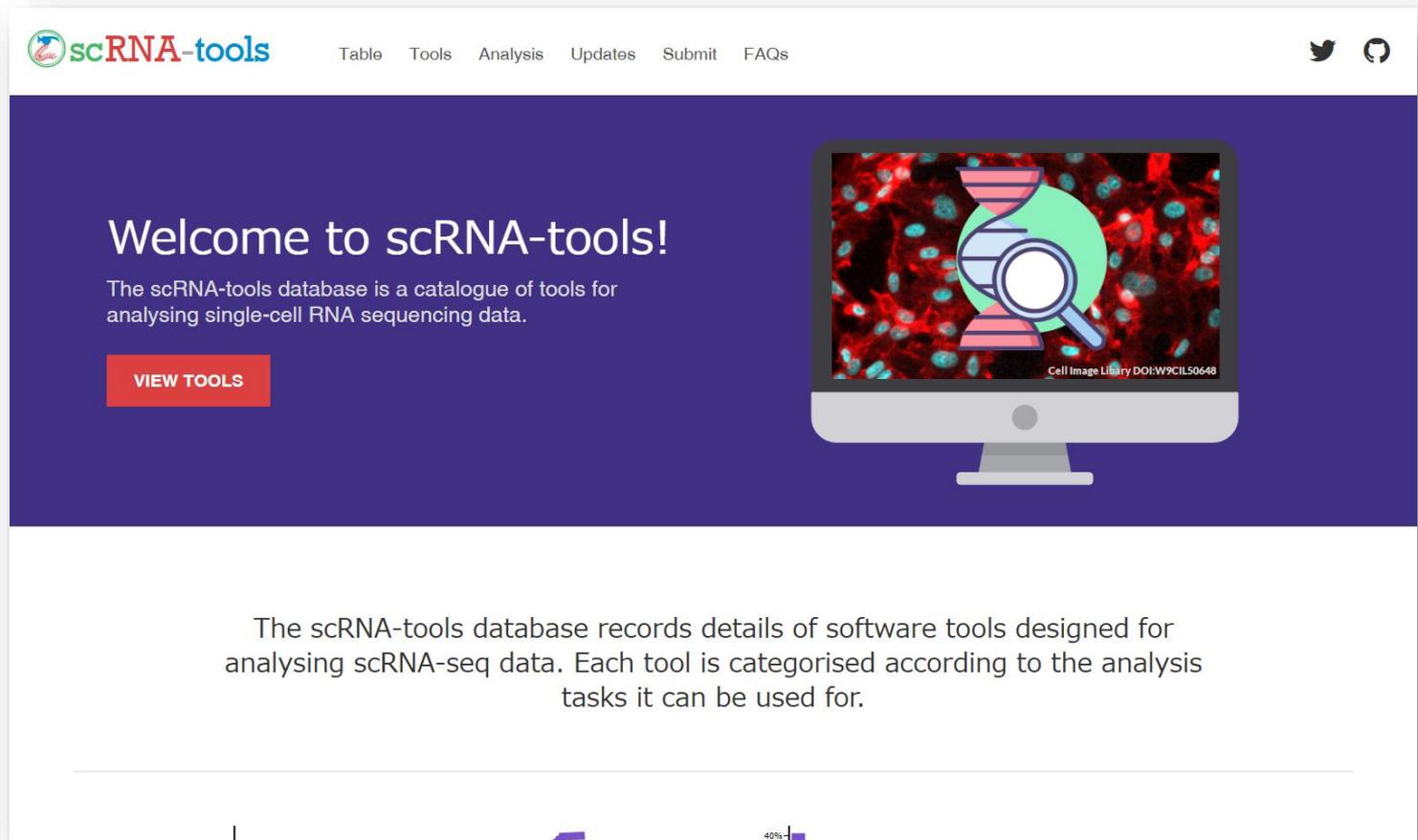
RIKEN Center for Integrative Medical Sciences
理化学研究所 生命医学研究センター

SCPortalen2: 5' end data, new data visualization and limitless exploration.

We now present SCPortalen2! But what's changed? In this new version we've made some changes to improve the quantity and quality of data as well as a new service to explore thousands of scRNA-seq

scRNA-tools

- <https://www.scrna-tools.org/>
- シングルセルRNA-seqの解析に用いられる様々なソフトウェアに関するデータベース



The screenshot shows the homepage of the scRNA-tools website. The header features the logo "scRNA-tools" on the left, navigation links "Table", "Tools", "Analysis", "Updates", "Submit", and "FAQs" in the center, and social media icons for Twitter and GitHub on the right. The main content area has a dark blue background with the text "Welcome to scRNA-tools!" and a sub-header "The scRNA-tools database is a catalogue of tools for analysing single-cell RNA sequencing data." Below this is a red button labeled "VIEW TOOLS". To the right is an illustration of a computer monitor displaying a DNA double helix and a magnifying glass over a cell image. The footer contains the text: "The scRNA-tools database records details of software tools designed for analysing scRNA-seq data. Each tool is categorised according to the analysis tasks it can be used for."

まとめ – Take-home messages

- すでに多くのシングルセルRNA-seqのデータが公共リポジトリ等から入手可能である
 - INSDC (DDBJ / NCBI / EMBL-EBI)
 - Human Cell Atlas 関連サイト
 - 10x Genomics 社の提供するデータセット など
- これらを利用することで今からでもシングルセルRNA-seqのデータ解析がスタートできる
 - データ解析手法の勉強
 - 自分の研究対象に近いデータセットを解析することで、研究の見通しを立てられる
- シングルセルRNA-seqは関連情報も多く存在する
 - シングルセルRNA-seqデータや解析方法も用意に入手可能
 - ただし、シングルセルRNA-seqの解析手法の取得も重要であるが、常にアップデートし続けていくことも重要