Bioinfo seminar: The first step towards scRNA-seq data analysis in python

Kenji Kamimoto, Ph.D.
微生物病研究所生物情報解析分野
システム生物学グループ
准教授　神元健児

What we will cover today

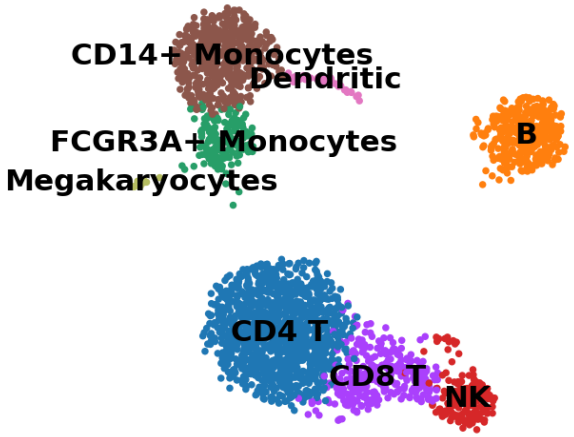
Part 1. Introduction to single-cell biology. ~15mins
The history and principle of single-cell transcriptome technologies.

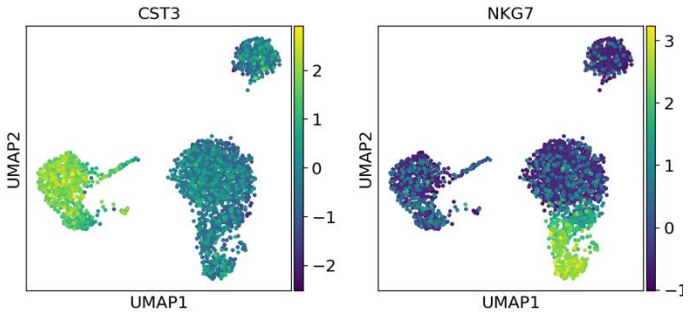
Part 2. Fundamental single-cell analysis with Scanpy ~1-2h
How we analyze such high-dimensional datasets.

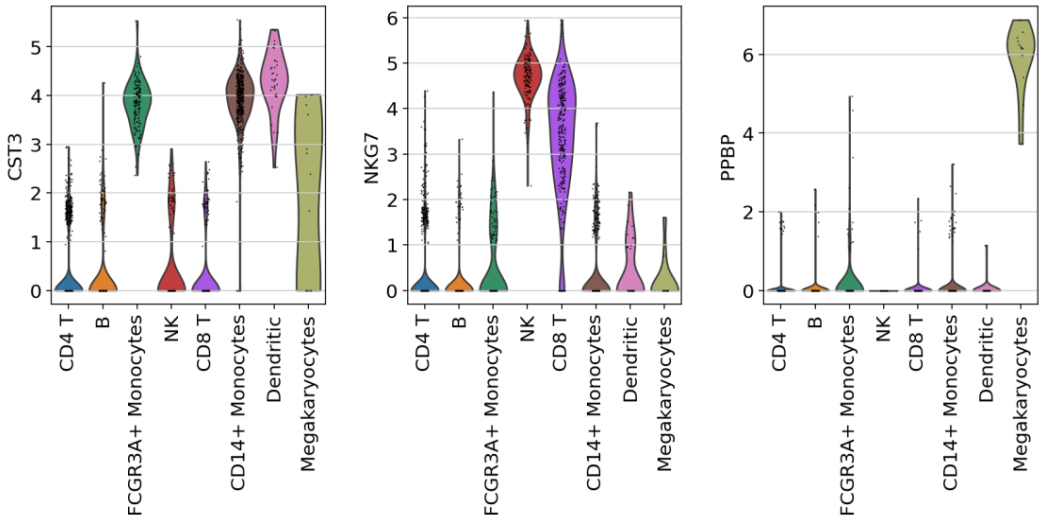
Part 3. How to use single-cell data in a public database. ~15mins
- CellxGene dataset

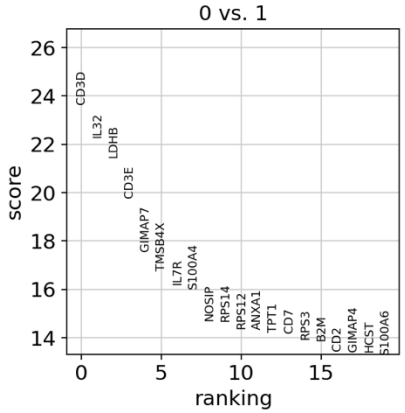
Part 4. Getting started with explorative single-cell data analysis ~15mins
- CellxGene interactive analysis tool.

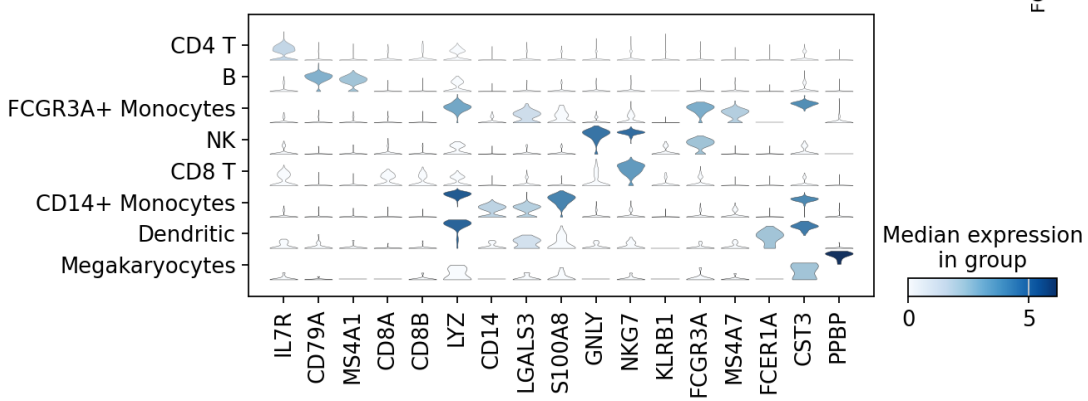
今日のハンズオン目標







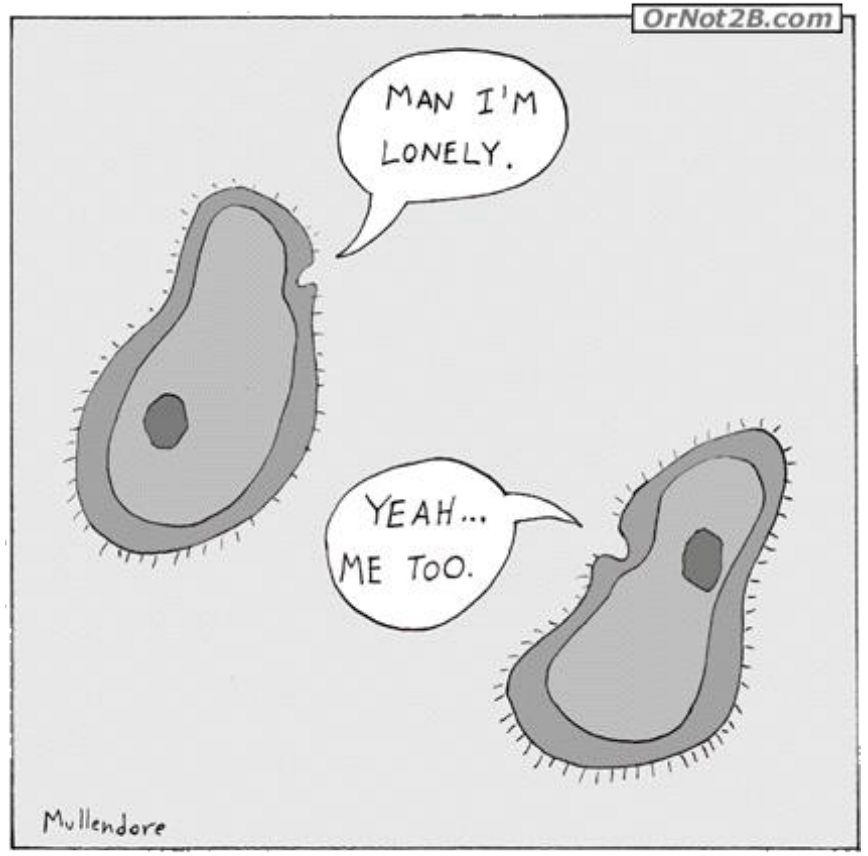




このような作図・解析をできるようにします

Part1: Introduction to single-cell biology

- The recent history of scRNA-seq
- Molecular biology underlying the technology
- How we analyze such high-dimensional datasets
- Different single-cell modalities
- Limitations of current single-cell approaches



Single cells.

Why single-cell?

Bulk

Single-cell

Single-cell modalities

- RNA-seq
- Nucleus-seq
- ATAC-seq
- Spatial
- Lineage tracing
- Transcription factor binding
- Proteomics
- … https://www.nature.com/collections/sxnwgntqsk

History of single-cell sequencing

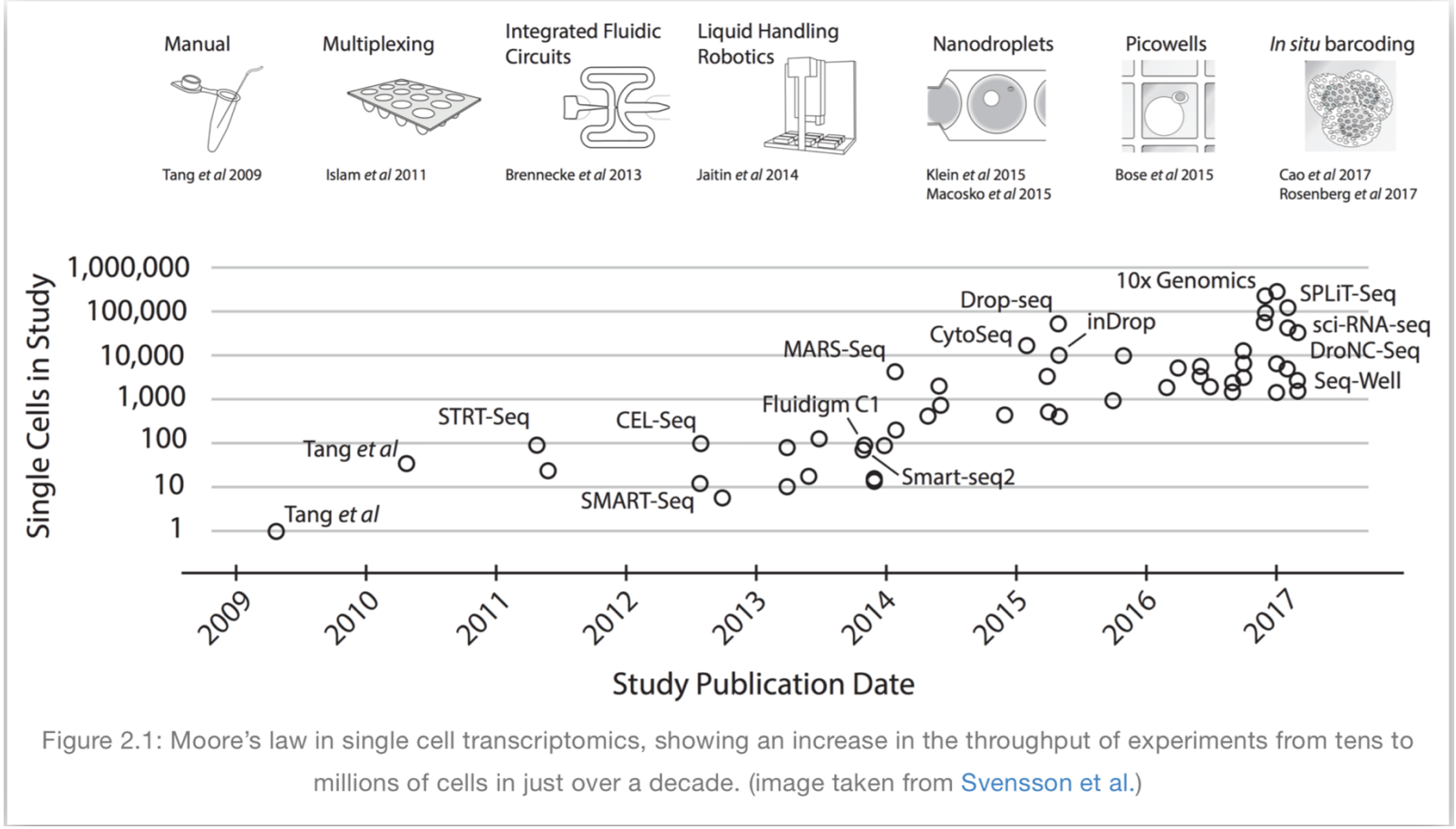


Figure 2.1: Moore's law in single cell transcriptomics, showing an increase in the throughput of experiments from tens to millions of cells in just over a decade. (image taken from Svensson et al.)

1990: Single-cell PCR

- In 1990 (Methods in Molecular and Cellular Biology):

Representative in Vitro cDNA Amplification From Individual Hemopoietic Cells and Colonies

GERARD BRADY, MARY BARBARA, and NORMAN N. ISCOVE*
Ontario Cancer Institute, Toronto, Canada M4X 1K9

- Exponential amplification of cDNA from single hemopoietic cells
- Followed by PCR analysis on specific targets

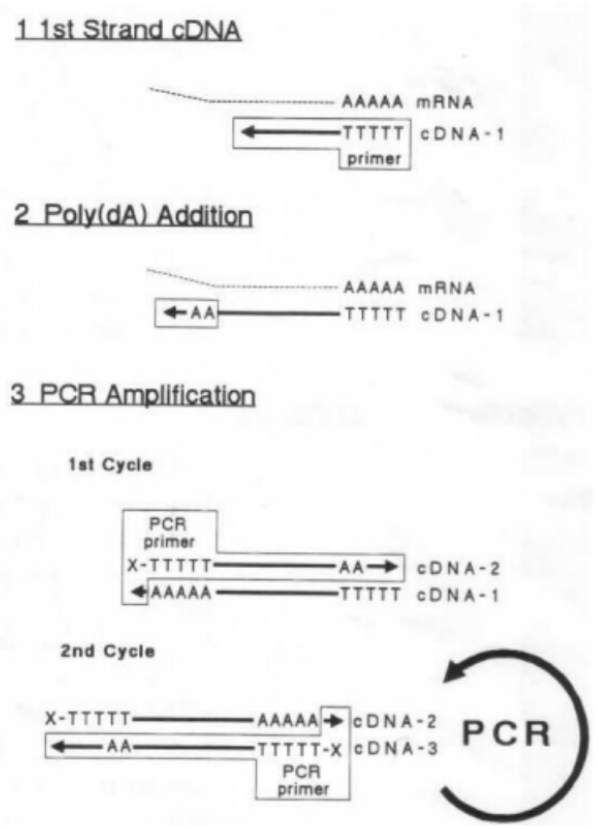


Figure 1. Sequence-independent amplification of total cDNA. Boxes enclose the reactions occurring at each step. The PCR primer contains oligo(dT) as well as a unique 36 base sequence labelled X (see Methodology.)

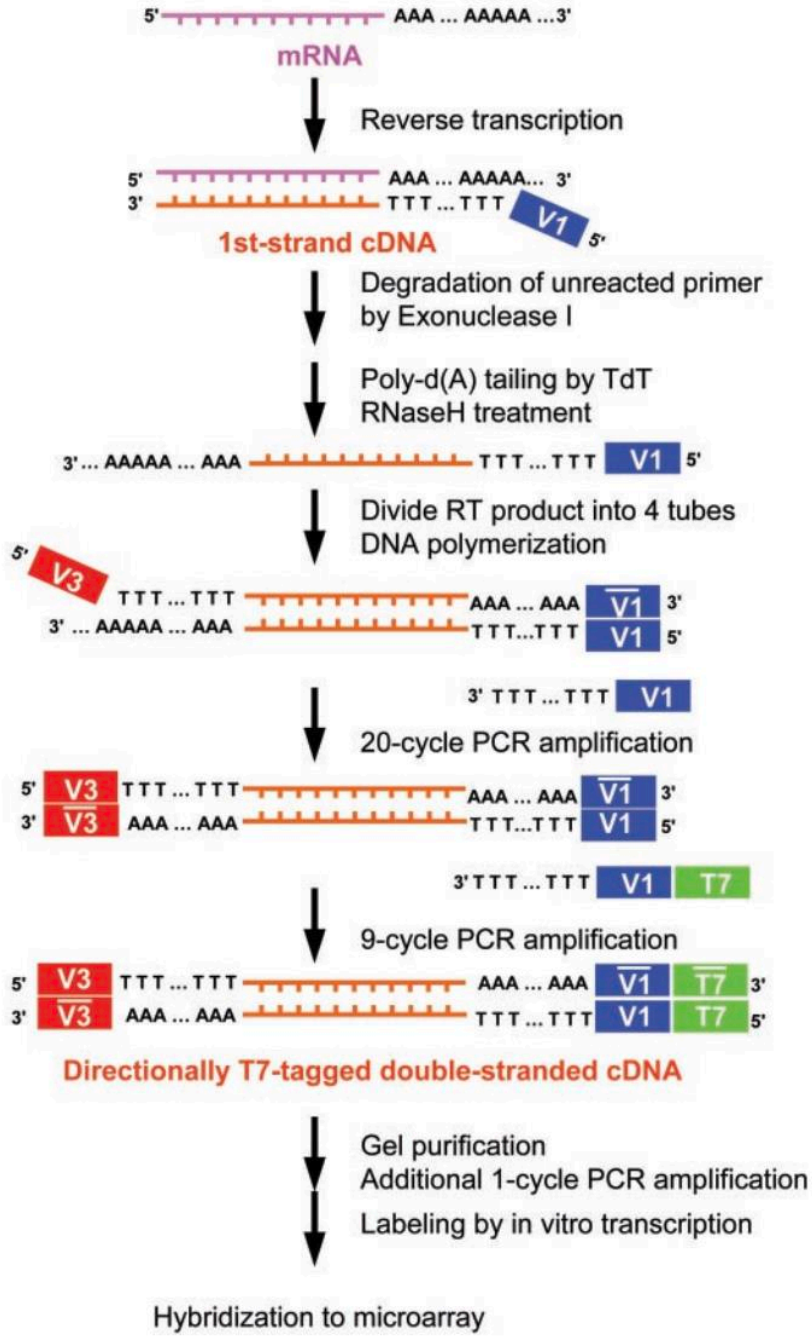
2006: Single-cell Microarray

Nucleic Acids Research, 2006, Vol. 34, No. 5 ***e42***
doi:10.1093/nar/gkl050

An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis

Kazuki Kurimoto[1], Yukihiro Yabuta[1], Yasuhide Ohinata[1], Yukiko Ono[1,4], Kenichiro D. Uno[2], Rikuhiro G. Yamada[3], Hiroki R. Ueda[2,3] and Mitinori Saitou[1,5,6,*]

- Applied to single-cell analysis of mouse preimplantation embryos



2009: single-cell RNA-seq is born

mRNA-Seq whole-transcriptome analysis of a single cell

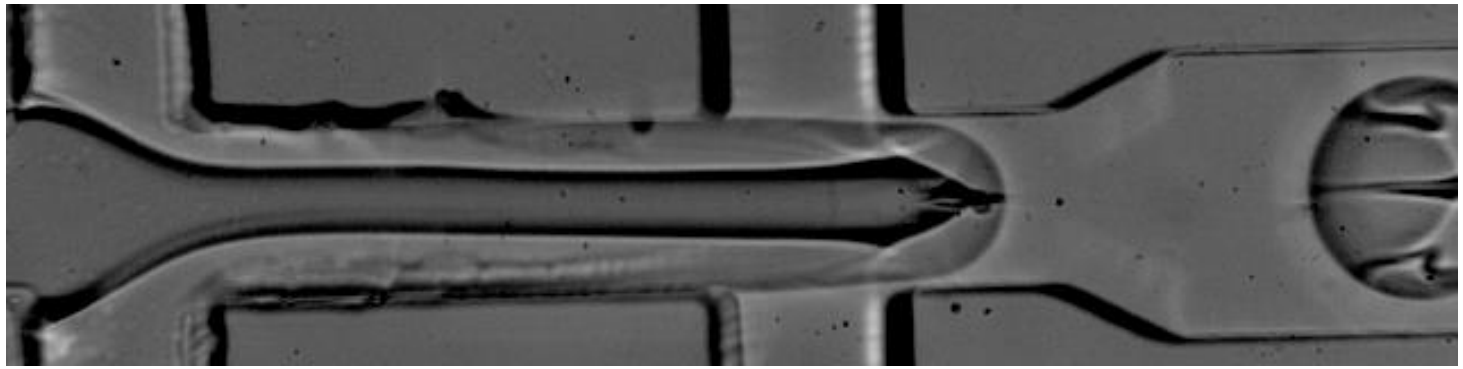
Fuchou Tang[1,3], Catalin Barbacioru[2,3], Yangzhou Wang[2], Ellen Nordman[2], Clarence Lee[2], Nanlan Xu[2], Xiaohui Wang[2], John Bodeau[2], Brian B Tuch[2], Asim Siddiqui[2], Kaiqin Lao[2] & M Azim Surani[1]

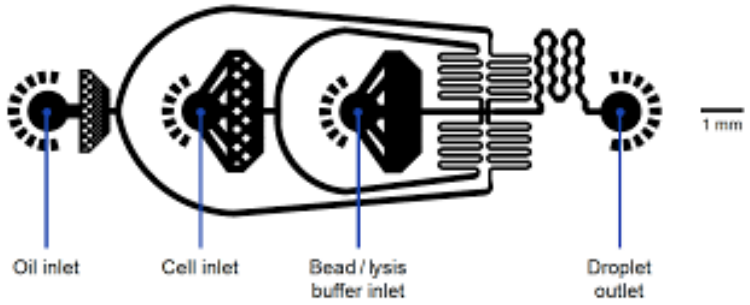
- A modified version of the previous protocol developed for scMicroarray

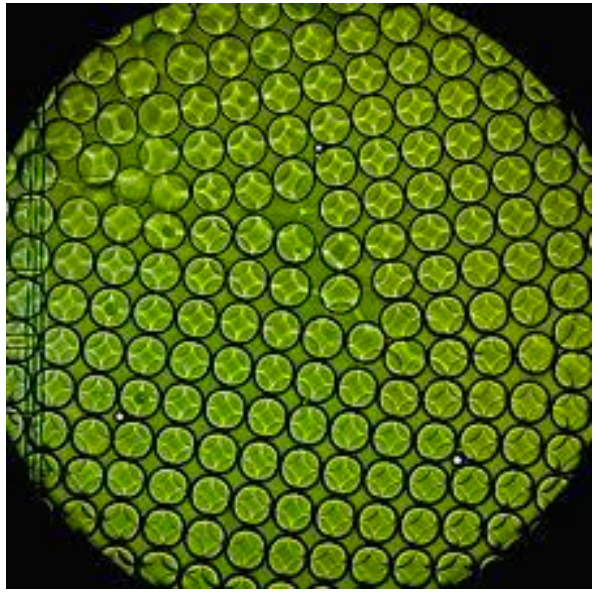
1. Time for RT incubation increased to get full-length first-strand cDNAs (PCR extension time was increased too).
2. Eliminated end bias during sequencing
3. SOLiD seq library prepped from amplified single-cell cDNAs (fragmentation, ligation etc).

2015: Highly-parallel scRNA-seq

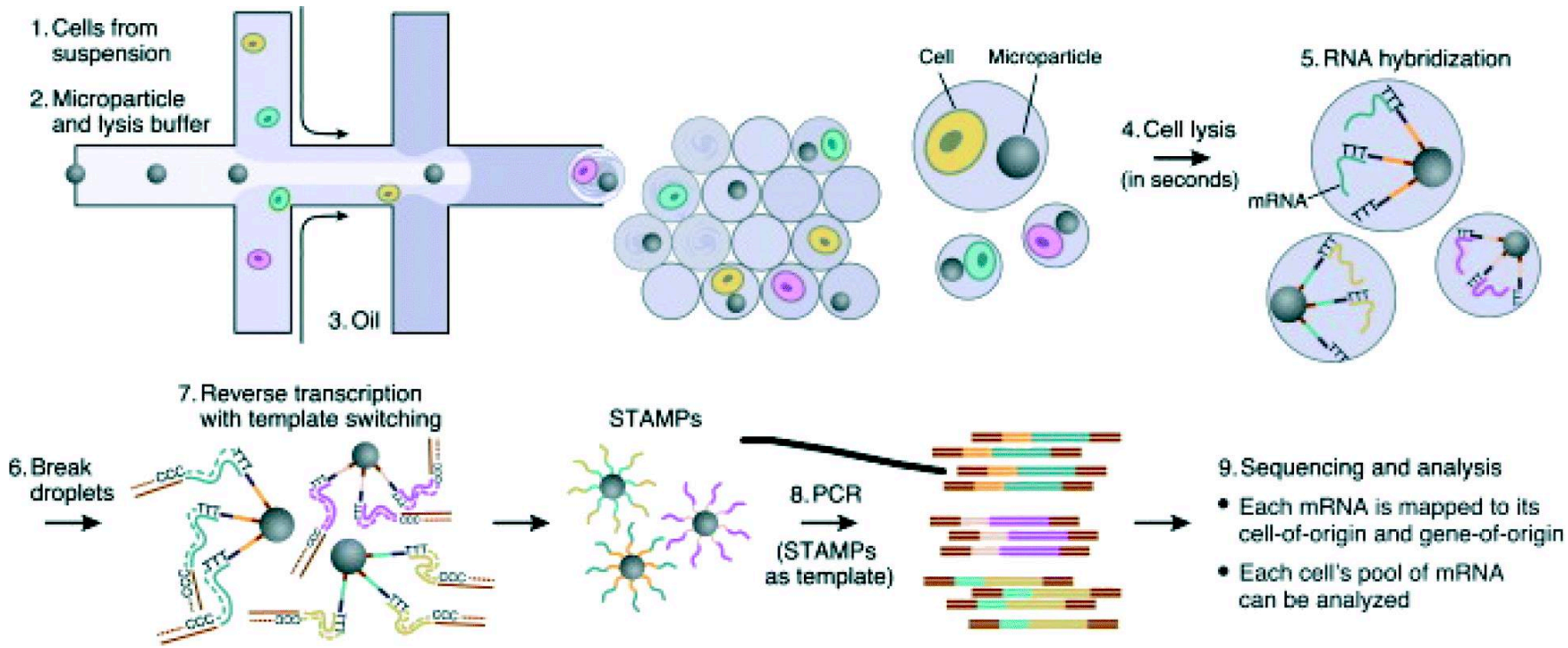
- Drop-seq
- Co-encapsulation of cells with beads massively parallelizes single-cell isolation and reduces cost





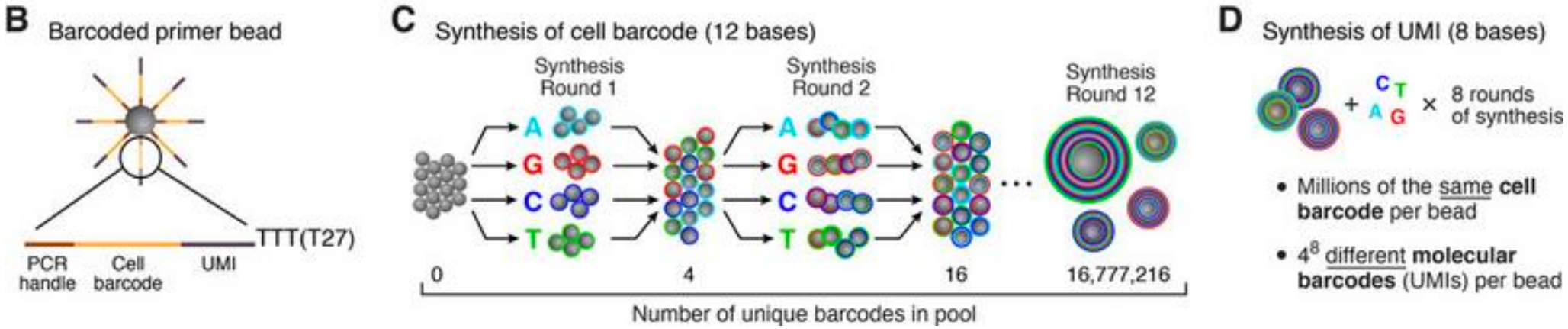


Drop-seq essentials



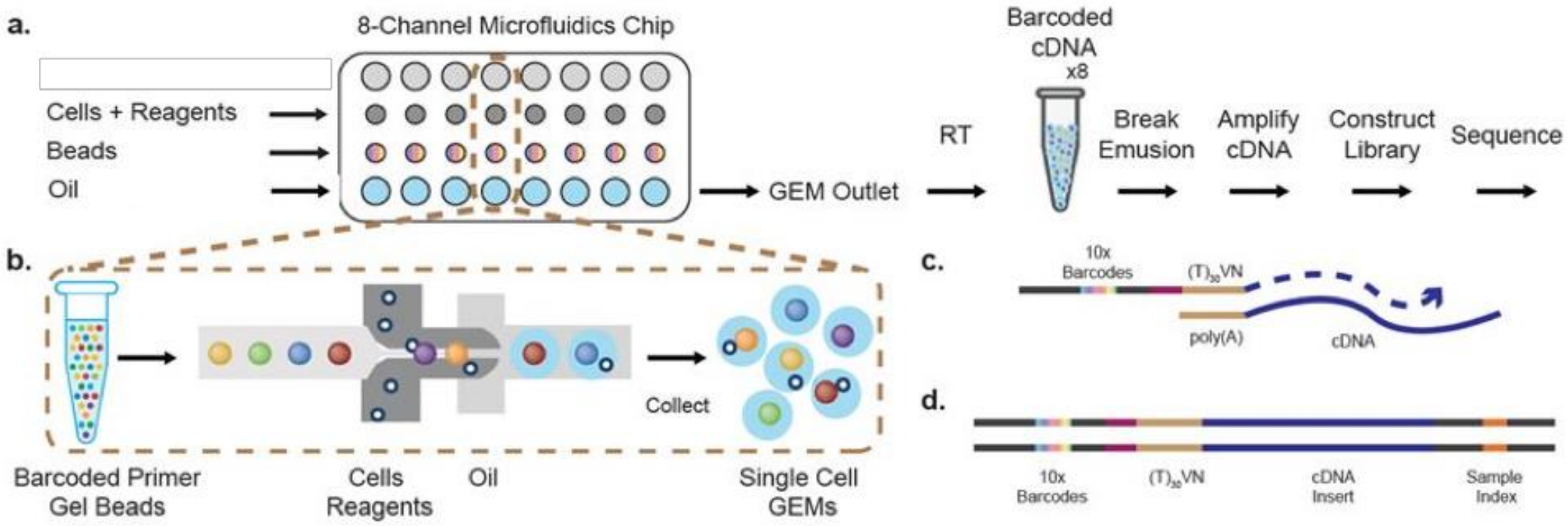
Dropseq.org

Key to Drop-seq: Cell barcode synthesis on beads



Dropseq.org

Successful commercialization: 10x Genomics



Some limitations to consider:

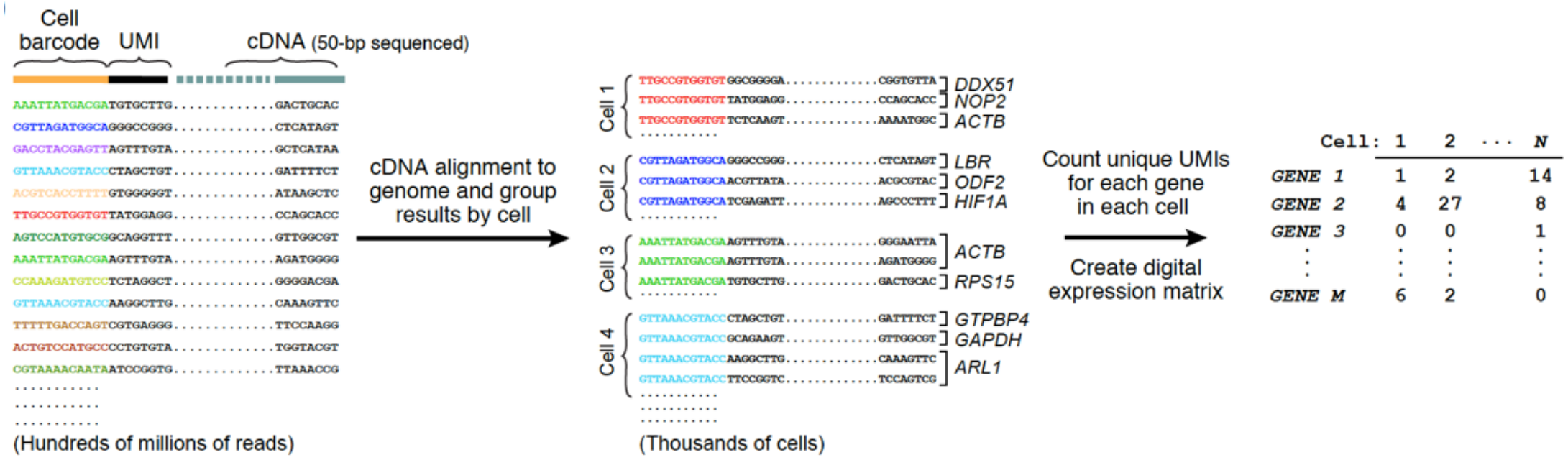
- Single-cell datasets are sparse and suffer dropout – meaning that lowly expressed transcripts (e.g. those of transcription factors) are more likely to escape detection
- It is very VERY easy to produce poor-quality datasets, resulting from poor cell handling
- Likewise, it’s easy to create artifacts – e.g. cell multiplets
- Loss of spatial, temporal and lineage

Key points:

- scRNA-seq is ideally suited to study heterogeneous populations of cells.
- A typical sample preparation workflow consists of isolating single cells (or nuclei), converting the RNA into cDNA, preparing a sequencing library (Illumina) and sequencing.
- Many single-cell protocols have been developed, some openly available, others provided commercially. These mainly differ in their throughput (how many cells are captured per experiment), the type of quantification (full-length or tag-based) and also cost.
- SMART-seq2 is a popular low-throughput method, providing full-length transcript quantification.
- 10x Chromium is a popular high-throughput method, using UMIs for transcript quantification (from either 3’ or 5’ ends).
- When planning an experiment, care should be taken to avoid confounding due to batch effects as well as ensuring an adequate level of replication to address questions of interest.

Analysis

The DGE: Digital Gene Expression Matrix

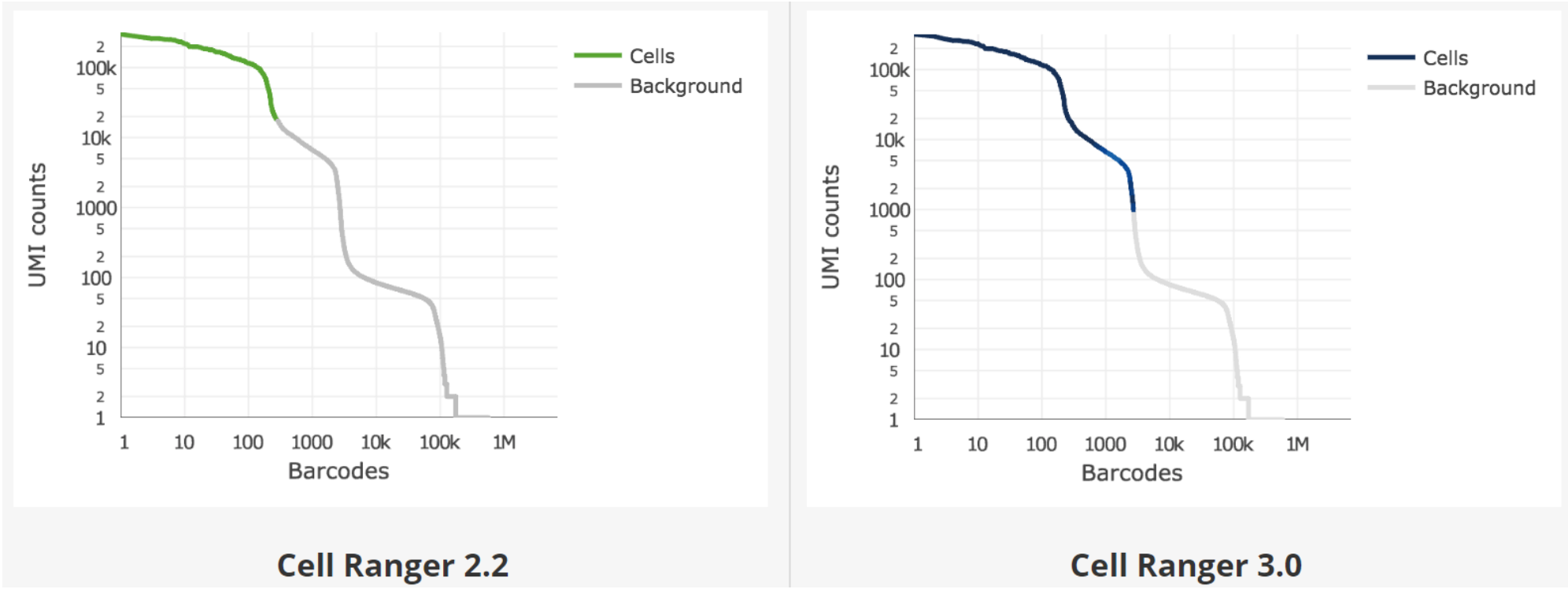


Dropseq.org

Alignment needs to be performed to get DGE, but today, we will skip the alignment process in this class.

Quality control

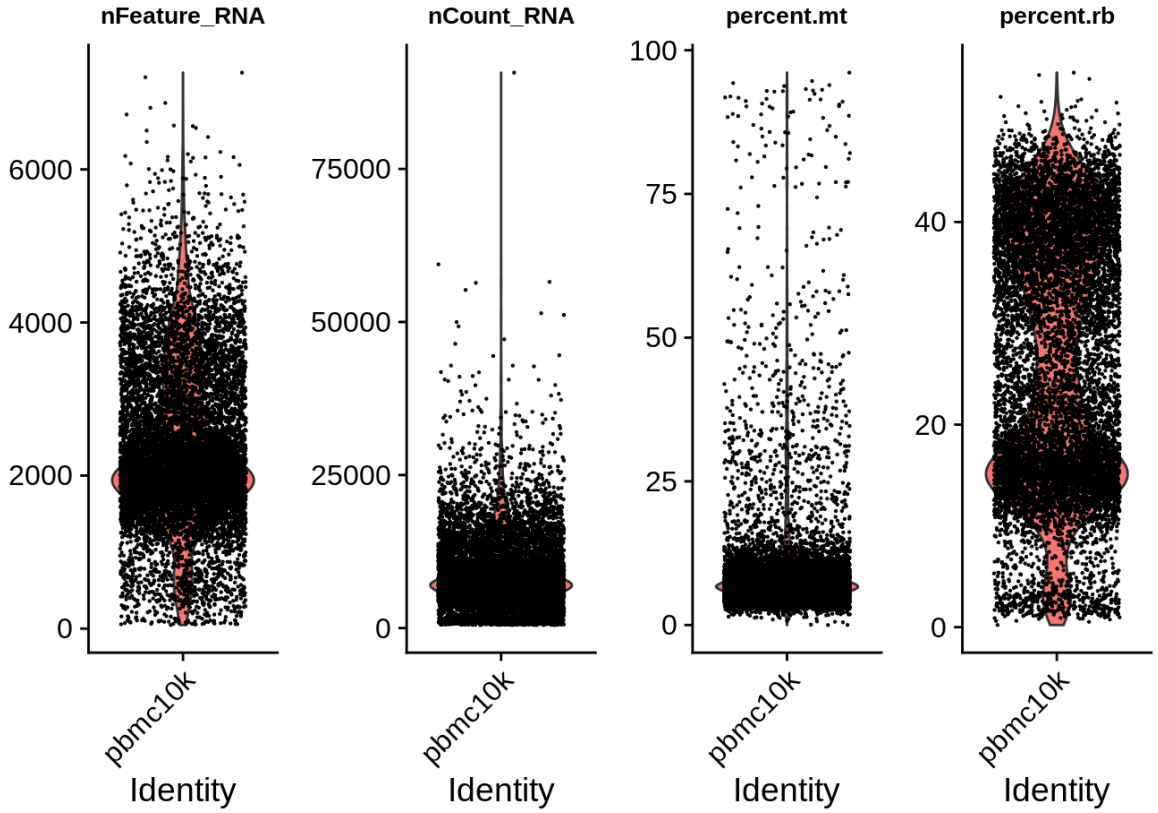
Cell filtering



Dropseq.org

Quality control

Cell filtering



Percent of mitochondrial genes, RNA counts per cell, Gene counts per cell are widely used features for quality control.

Comprehending high-dimensional datasets

- Dimensionality in statistics refers to **how many attributes a dataset has**

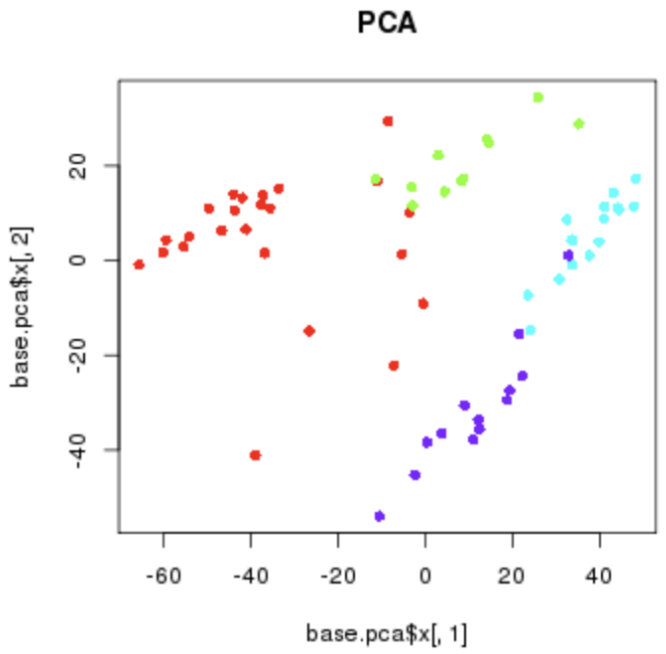
- In single-cell analysis, for each cell, we have collected measurements on thousands of genes

- To be able to comprehend this information, we visualize high-dimensional data by projecting it into a low-dimensional space

Methods for reducing dimensionality: PCA

- Principal Component Analysis (PCA) is the classic, fast and easy approach
- PCA creates low-dimensional embeddings that best preserves the **overall variance** of the dataset
- PCA is linear dimensionality reduction

and typically does not cleanly segregate different cell types into distinct clusters

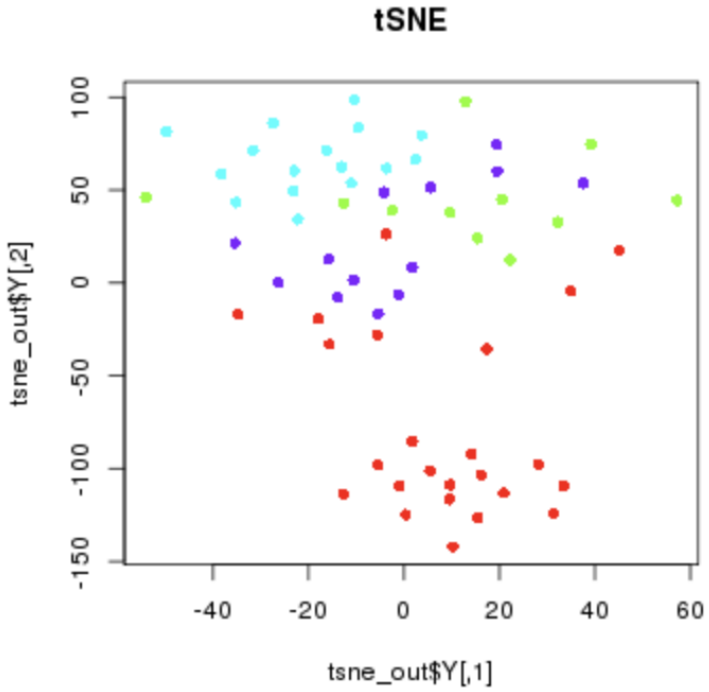


Methods for reducing dimensionality: t-SNE

- t-SNE: t-distributed Stochastic Neighbor Embedding
- tSNE is a non-linear dimensionality reduction method
- 2 central components of t-SNE:

1: Create a probability distribution in the high-dimensional space that dictates the relationships between various neighboring points

2. Recreate a low dimensional space that follows that probability distribution as best as possible (in This step – the t-distribution is used)



Methods for reducing dimensionality: UMAP

- Limitations of t-SNE: it is very good at preserving local structure but not global structure
- UMAP: Uniform Manifold Approximation and Projection
- UMAP advantages over t-SNE: faster runtime and consistency, meaningful organization of cell clusters and preservation of continuums

Dimensionality reduction for visualizing single-cell data using UMAP

Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W H Kwok, Lai Guan Ng, Florent Ginhoux & Evan W Newell ✉

Nature Biotechnology **37**, 38–44 (2019) | Download Citation ↓

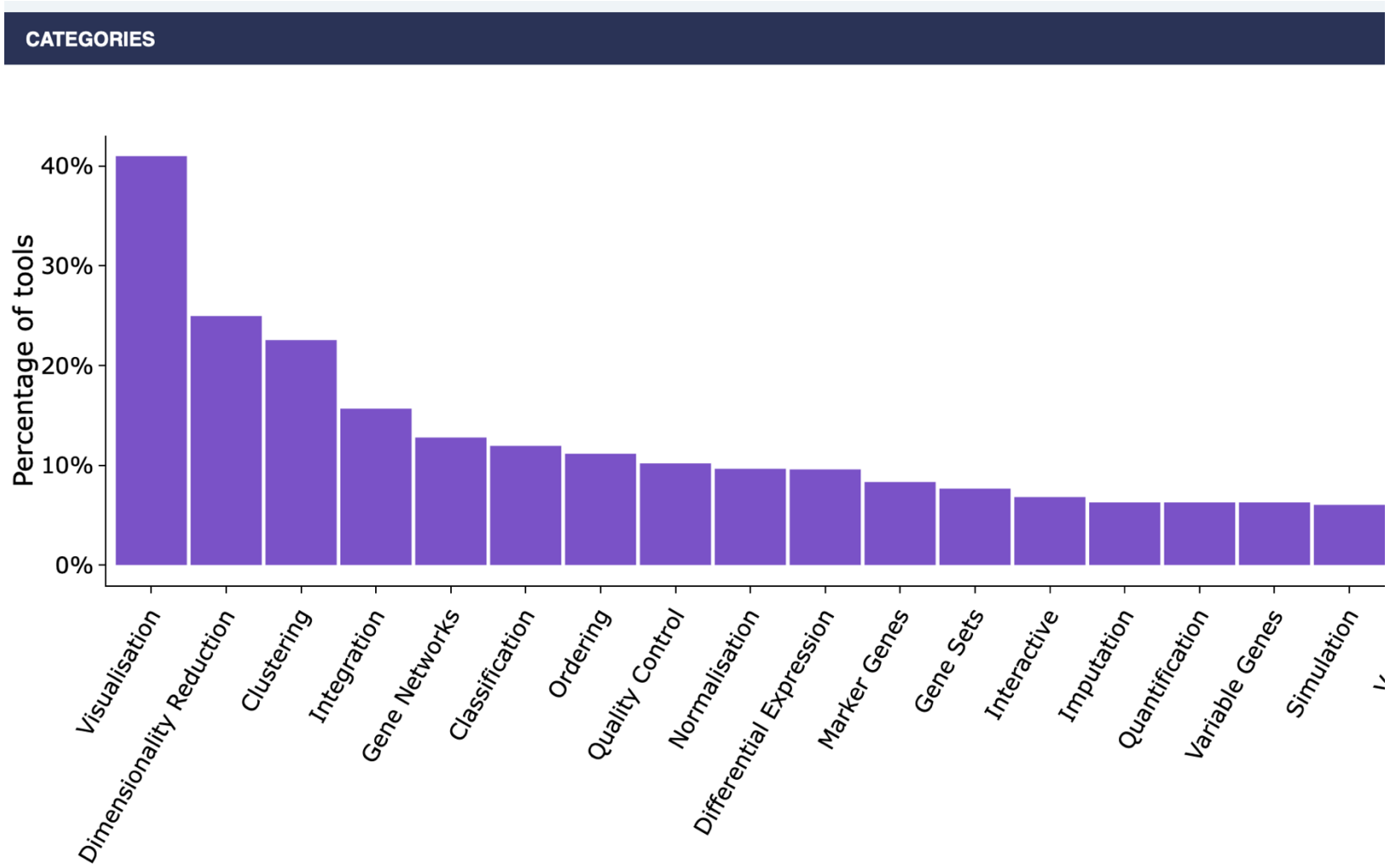
Getting started on analysis

R toolkit: https://satijalab.org/seurat/

Python: https://scanpy.readthedocs.io/en/stable/

Application of single-cell omics

- Many software programs have been published for single-cell analysis.
- Most of these tools are based on general statistical or data science methods: Visualization, Dimensionality reduction, clustering, etc. (https://www.scrna-tools.org/analysis)



Prepare a folder in Google Colab

1. Google driveへアクセス　https://drive.google.com/drive/home
2. 左上の　+新規（ + New) のボタンをクリック
   => 新しいフォルダ
   => scRNAseq_hands-on　という名前でフォルダ作成

3. https://github.com/KenjiKamimotoLab/scRNA_hands_on/ にアクセス
4. [Single_cell_RNA_seq_basics_with_Colab]　というところをクリック（左側に青いアイコンがあるところ）
5. [1_getting_sstarted_with_scanpy.ipynb] をクリック
6. 右側のダウンロードボタンをクリックしてファイルをダウンロード

7. ブラウザの戻るボタンで1つ前の画面に戻る

8. [2_CellxGene.ipynb]をクリック
6. 右側のダウンロードボタンをクリックしてファイルをダウンロード

Download files

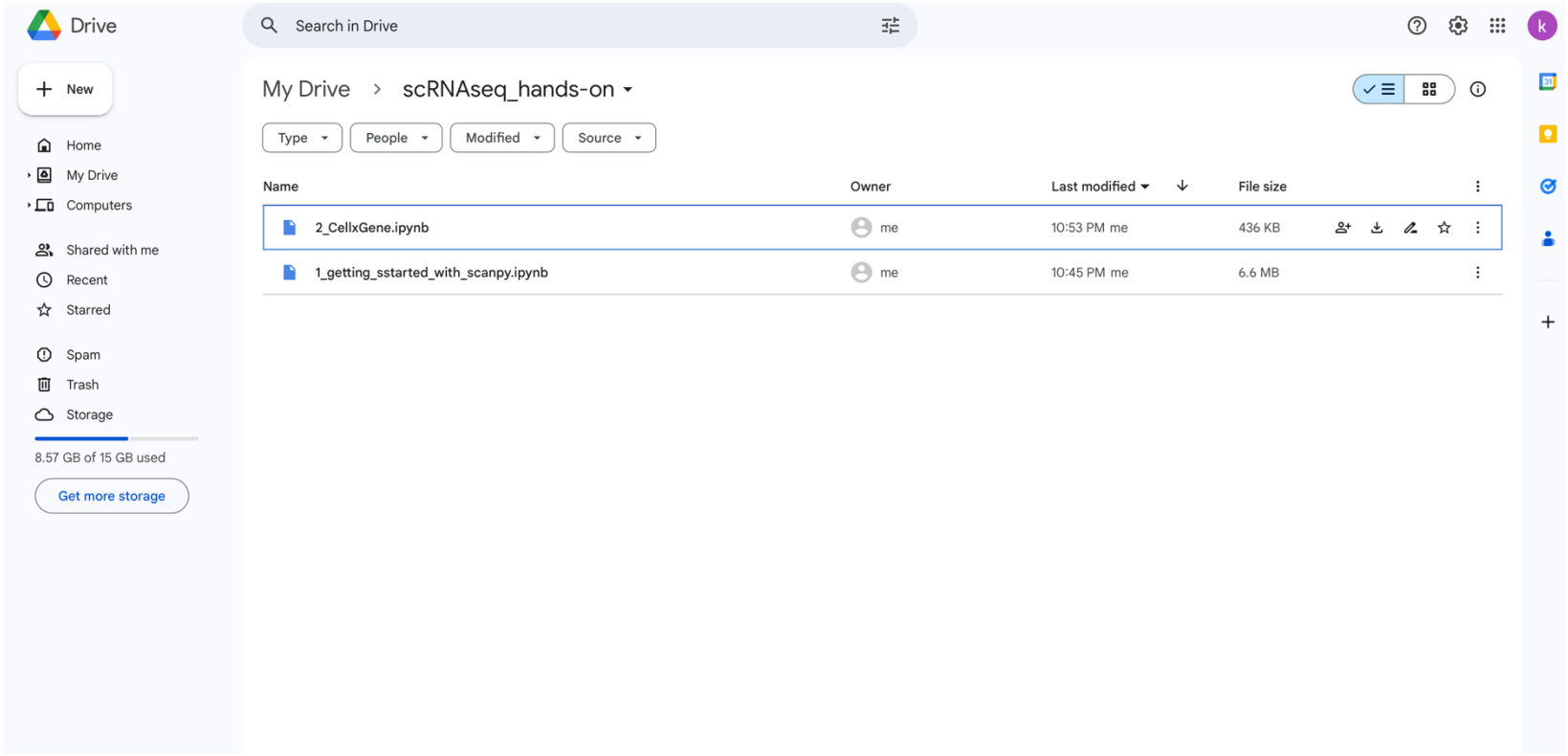
1. https://github.com/KenjiKamimotoLab/scRNA_hands_on/ にアクセス
2. [Single_cell_RNA_seq_basics_with_Colab]　というところをクリック（左側に青いアイコンがあるところ）
3. [1_getting_sstarted_with_scanpy.ipynb] をクリック
4. 右側のダウンロードボタンをクリックしてファイルをダウンロード

5. ブラウザの戻るボタンで1つ前の画面に戻る

6. [2_CellxGene.ipynb]をクリック
7. 右側のダウンロードボタンをクリックしてファイルをダウンロード

Place files in Google Colab

ダウンロードした2つのファイルを、Google drive の　scRNAseq_hands-on のフォルダ（一番最初に作成したもの）に配置する。



scRNAseq_hands-on　のフォルダに2つのファイルが上のように配置されていればOK.

Cell Ranger

Explore the output of cellranger count

The `cellranger count` pipeline outputs are in the pipestance directory in the outs folder. List the contents of this directory with `ls -1`.

Copy

```
ls -1 run_count_1kpbmcs/outs
```

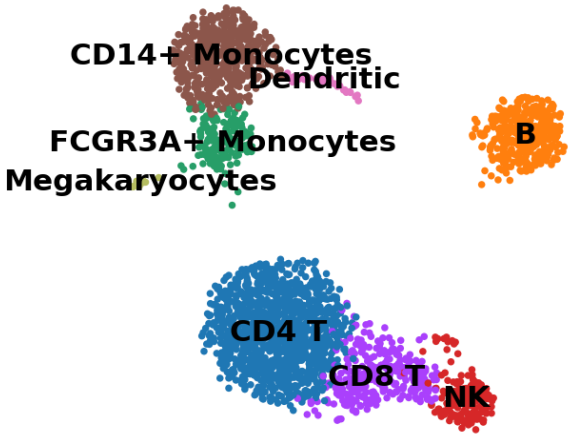
The output is similar to the following:

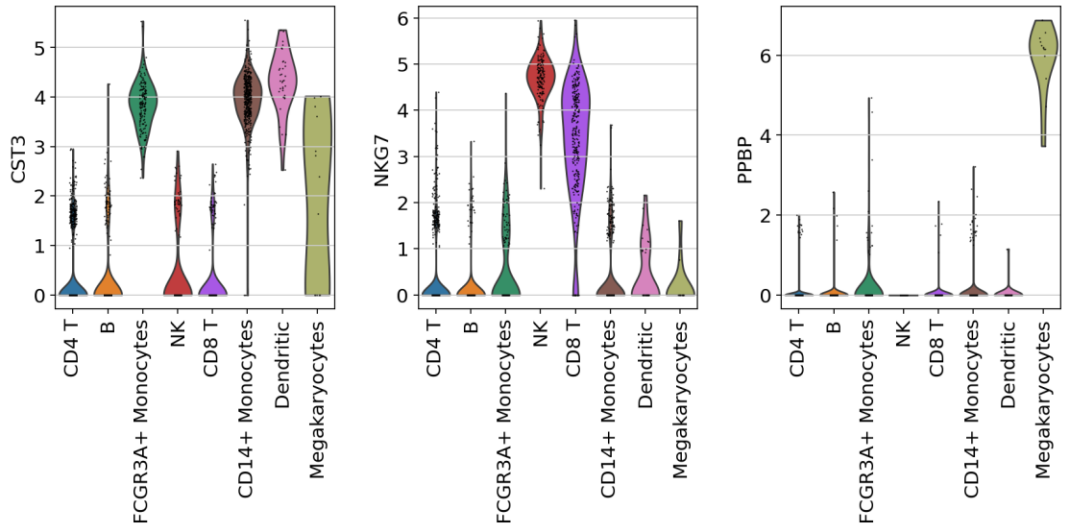
Copy

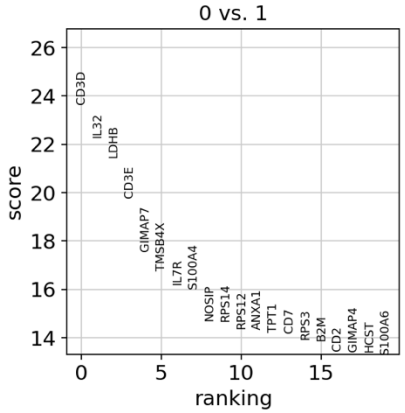
```
├── analysis
├── cloupe.cloupe
├── filtered_feature_bc_matrix
├── filtered_feature_bc_matrix.h5
├── metrics_summary.csv
├── molecule_info.h5
├── possorted_genome_bam.bam
├── possorted_genome_bam.bam.bai
├── raw_feature_bc_matrix
├── raw_feature_bc_matrix.h5
└── web_summary.html
```

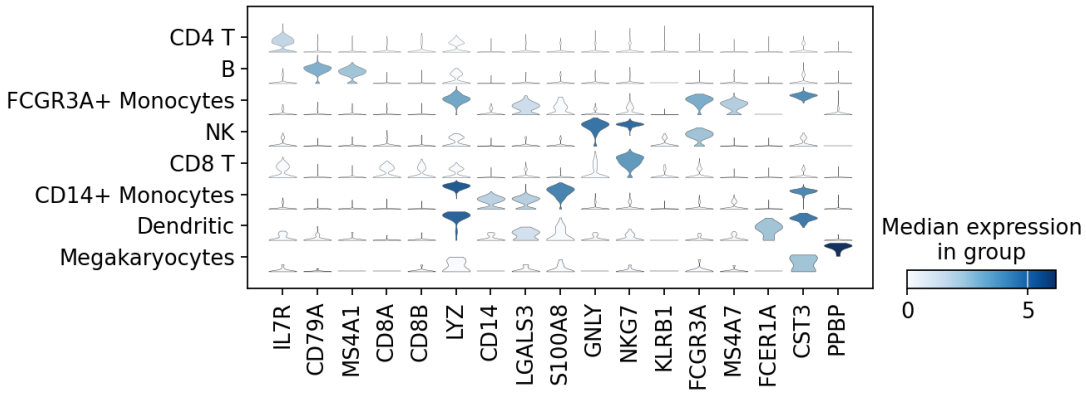
Check the web_summary.html to see results of the experiment. You can also load the cloupe.cloupe file into the Loupe Browser and start an analysis. This `outs/` directory also contains a number of outputs that can be used as input for software tools developed outside of 10x Genomics, such as the Seurat R package.

今日のハンズオン目標









このような作図・解析をできるようにします

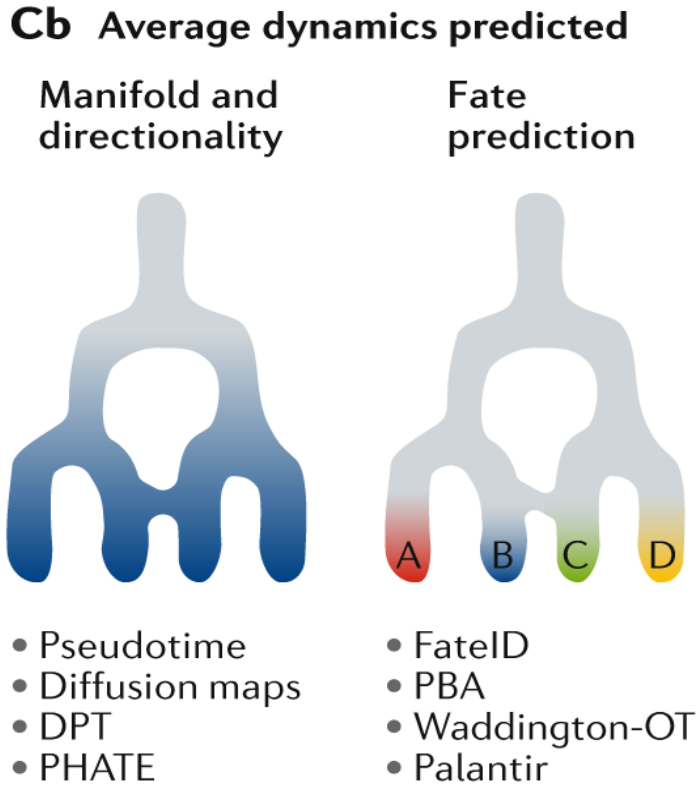
scRNA-seq 解析の流れ

1. 入力データの準備
2. データの品質チェック、前処理のための可視化
3. データのフィルタリング（解析に使用できない細胞、遺伝子を除く）
4. データの変換
5. Highly variable geneの検出
6. 次元削減（PCA, tSNE, UMAPなど）
7. クラスタリング
8. クラスター毎に特徴的な遺伝子を検出
9. DEG解析
10. 可視化
11. データの保存

次元削減について直感的に理解する

https://pair-code.github.io/understanding-umap/

Trajectory analysis / Pseudotime analysis



Review article Daniel E. Wagner and Allon M. Klein. (2020).

発生過程・分化誘導過程のタイムコースデータ等の変化の途中のscRNA-seqデータを並べ替えて、分化の道筋や時間情報を推論する手法

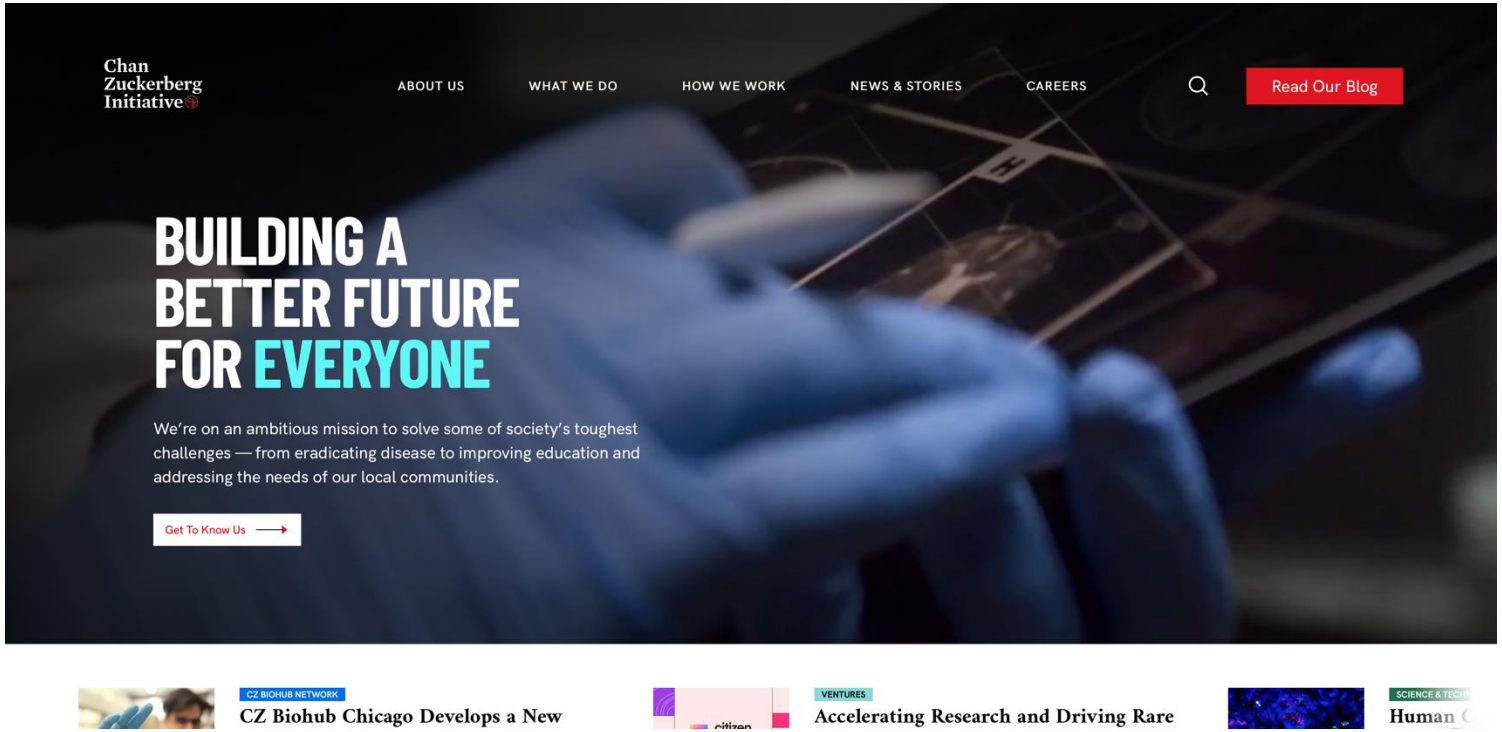
Pros

- 時間情報や発生のダイナミクスに関する多くの情報を提供する

Cons

- 次元削減と混同されることも多いが同一ではない
- 系譜に関する証拠として使えるわけではない
- どんなデータに使ってもいいわけではない

CellxGene： single-cell analysis program created by CZI

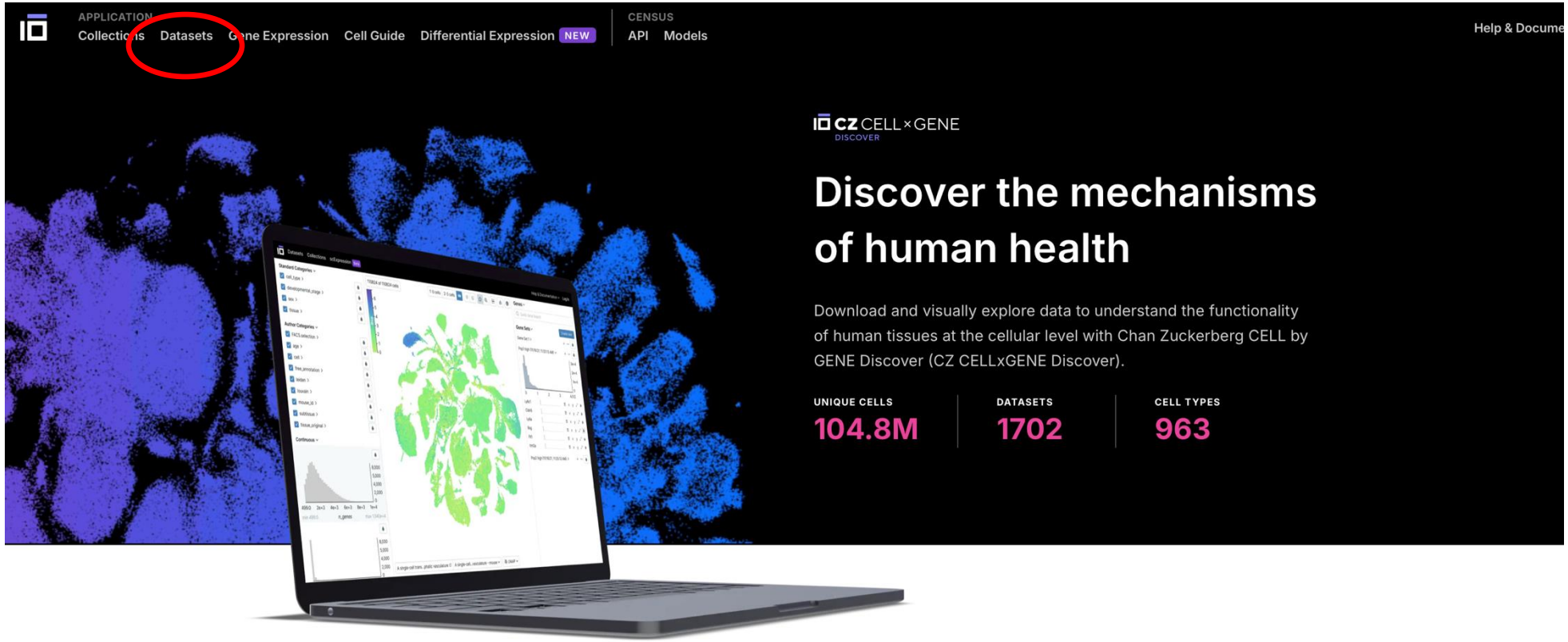


CZI: Chan Zuckerberg Initiative
Facebookの創設者Mark Zuckerbergとその妻Chan Zuckerbergが作った教育・研究団体

CZIはAIやsingle-cell関連の技術開発に力を入れており、CellxGeneはその成果の一つ

CellxGene Database

1. https://cellxgene.cziscience.com に移動。googleなどの検索エンジンでcellxgeneと検索してもよい
2. 左上のDatatesをクリック



CellxGene Database

組織・臓器名

病気名など

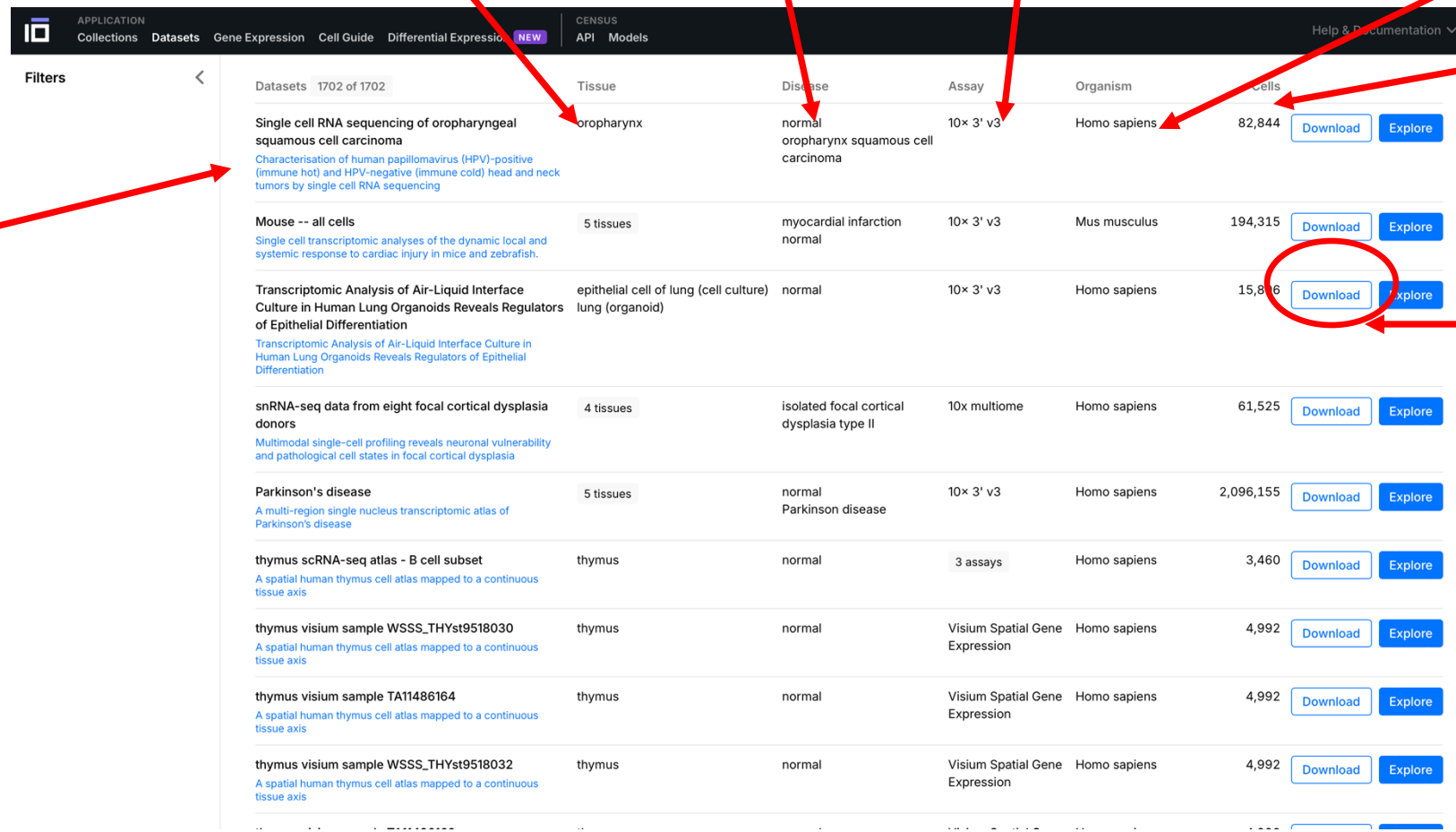
計測のプラットフォーム、
試薬の情報

生物種

データ内の
細胞の数

データ名、
説明

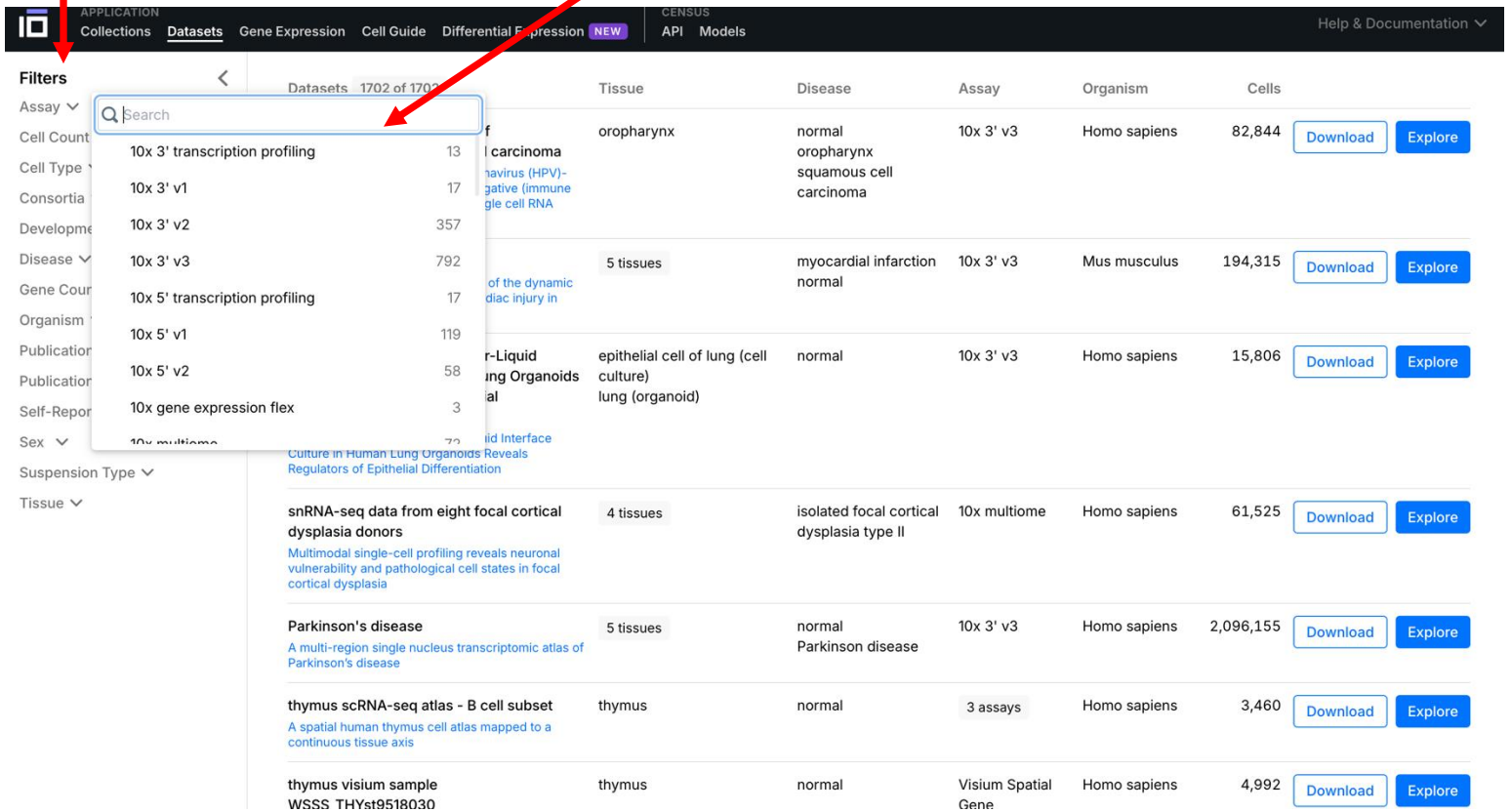
ダウンロード
ボタン



CellxGene Database

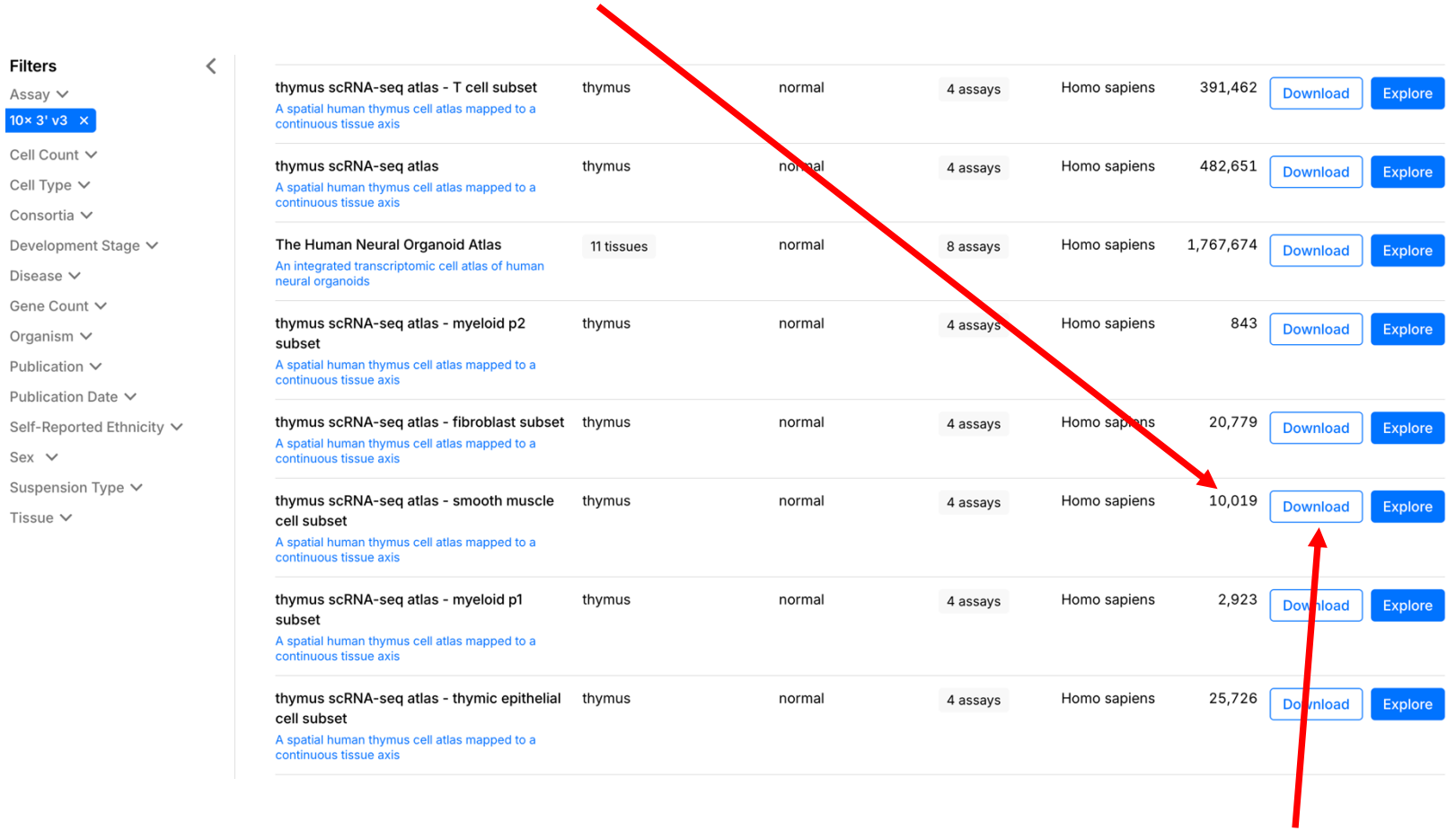
1. Filters　の下のAssayをクリック

２．「10x 3’ ~ 」もしくは「10x 5’~」，「10x gene expression flex」のどれかを選択



CellxGene Database

1. 好きなデータを選択。練習用であれば細胞数が少なめのものがいい（10万以下など）

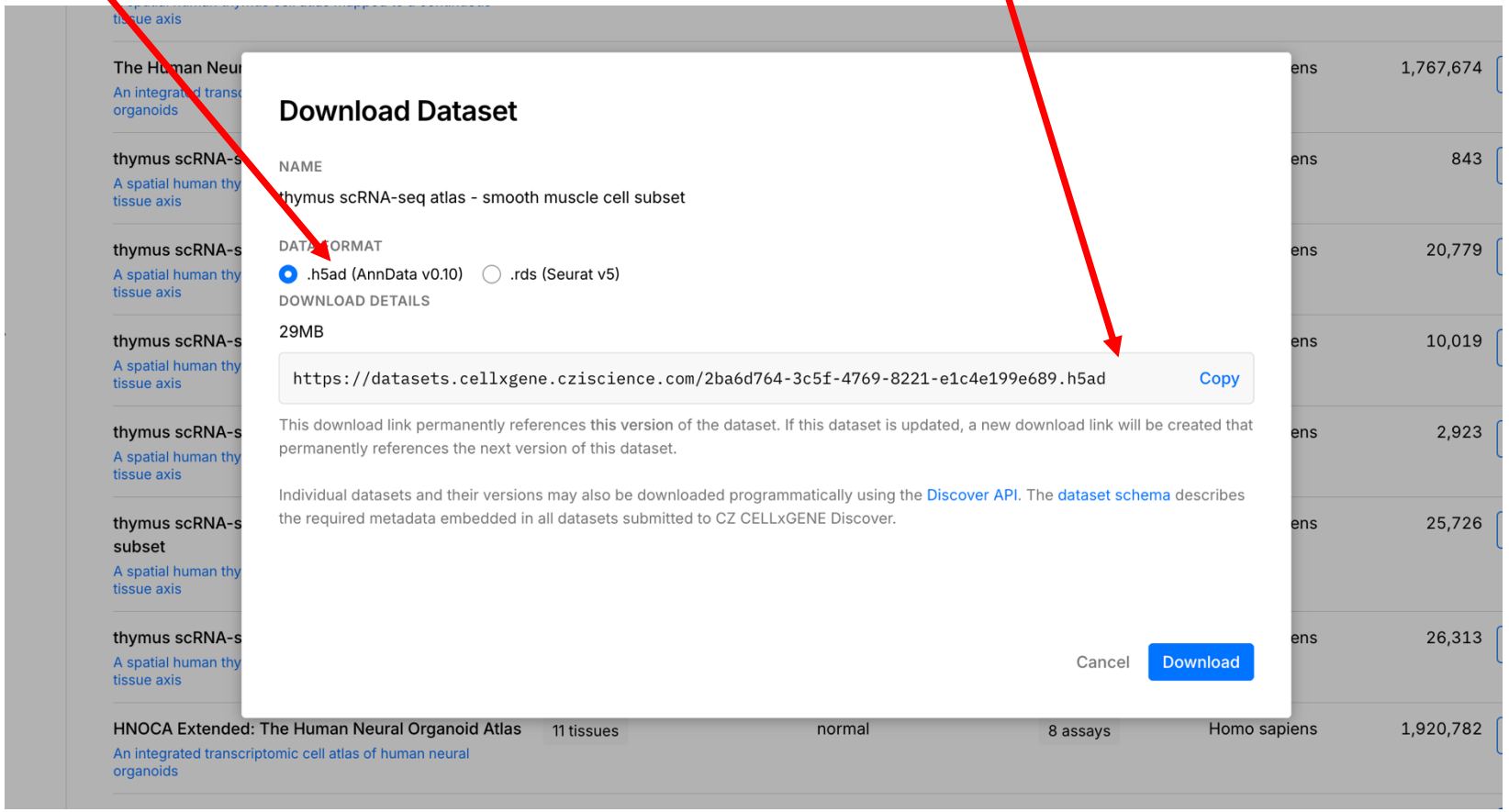


2．ダウンロードボタンをクリック

CellxGene Database

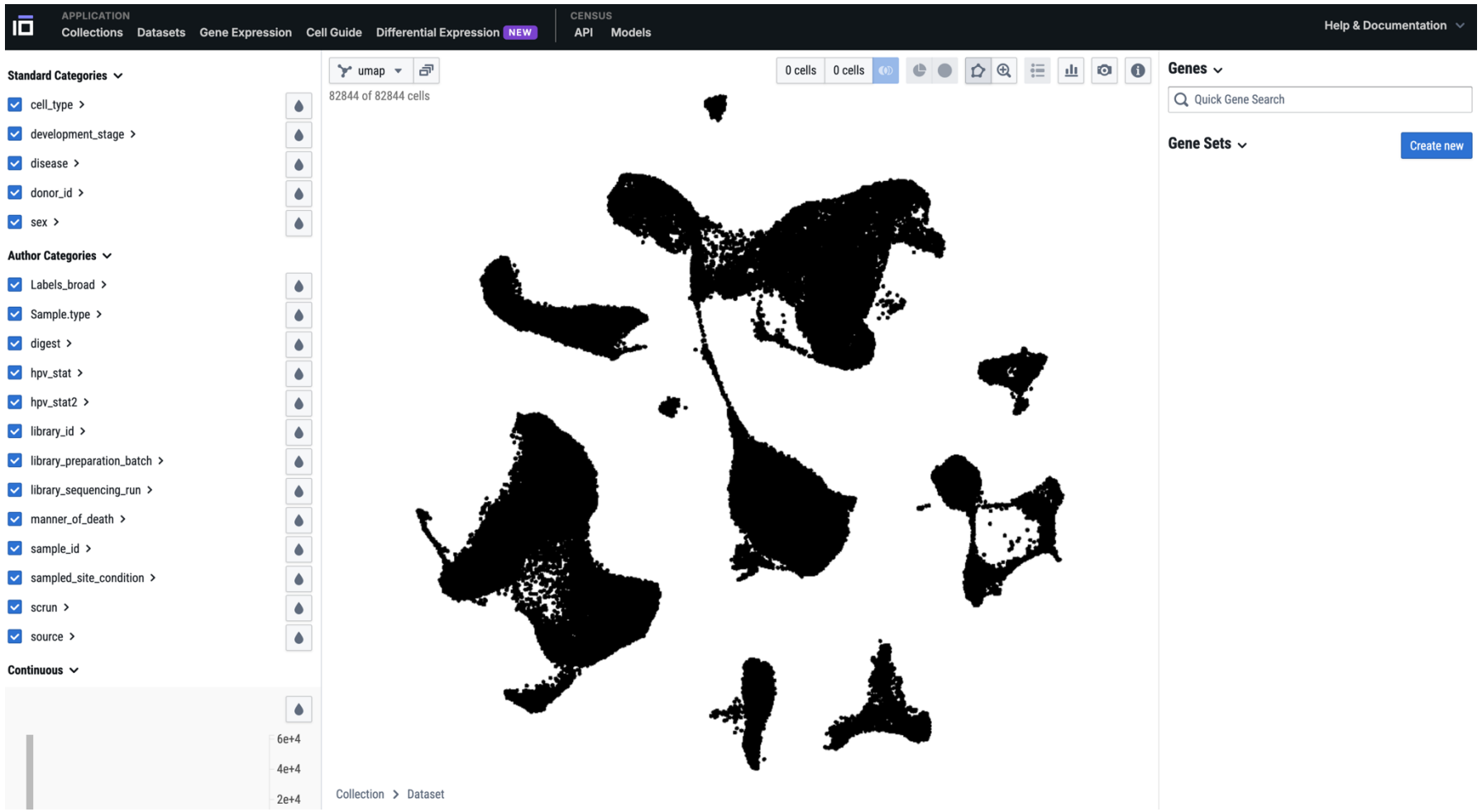
1．h5adフォーマットを選択

2．URLをコピー



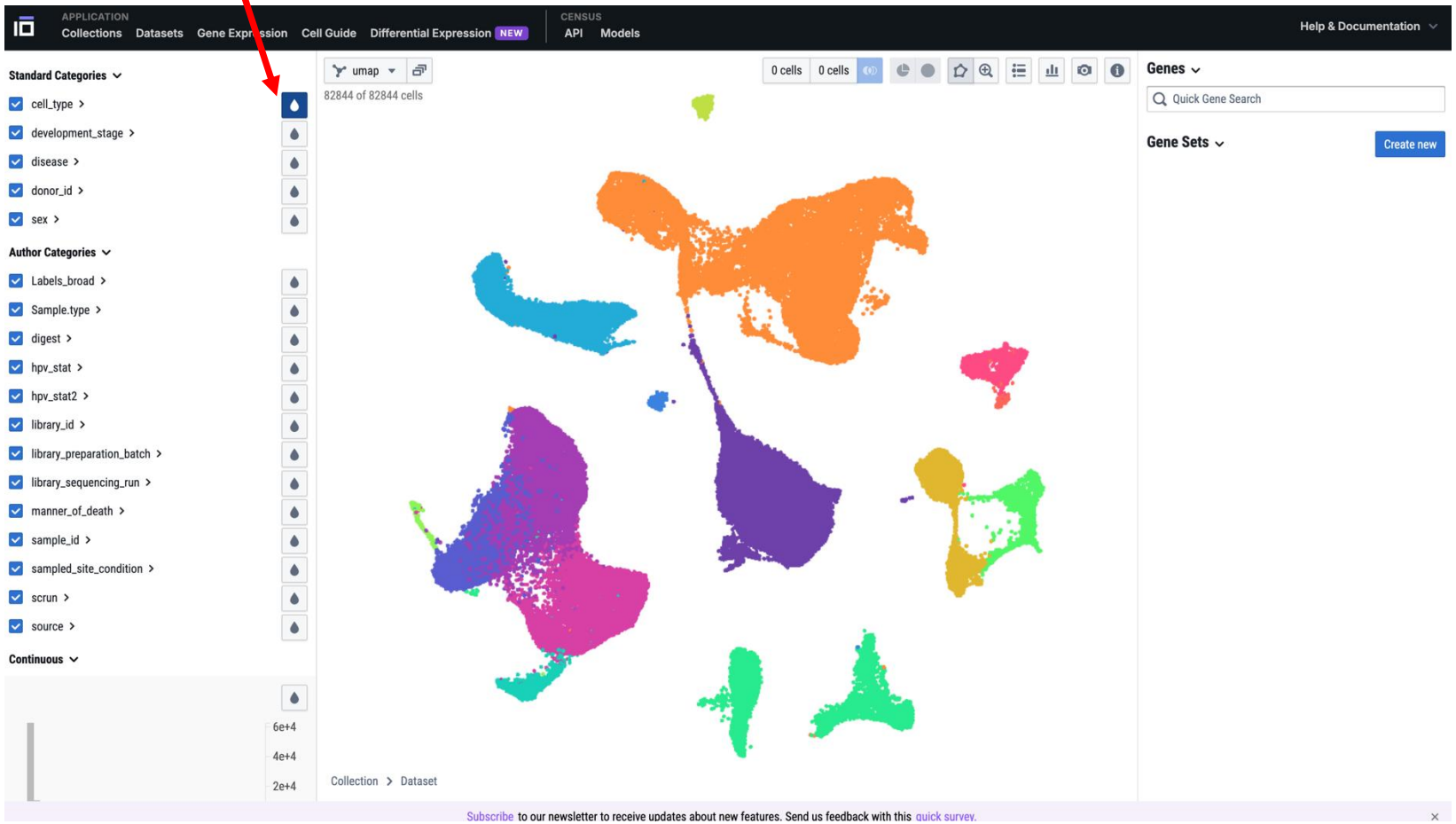
CellxGene data analysis

メタデータ　（クラスタ、病気、ドナーなど）



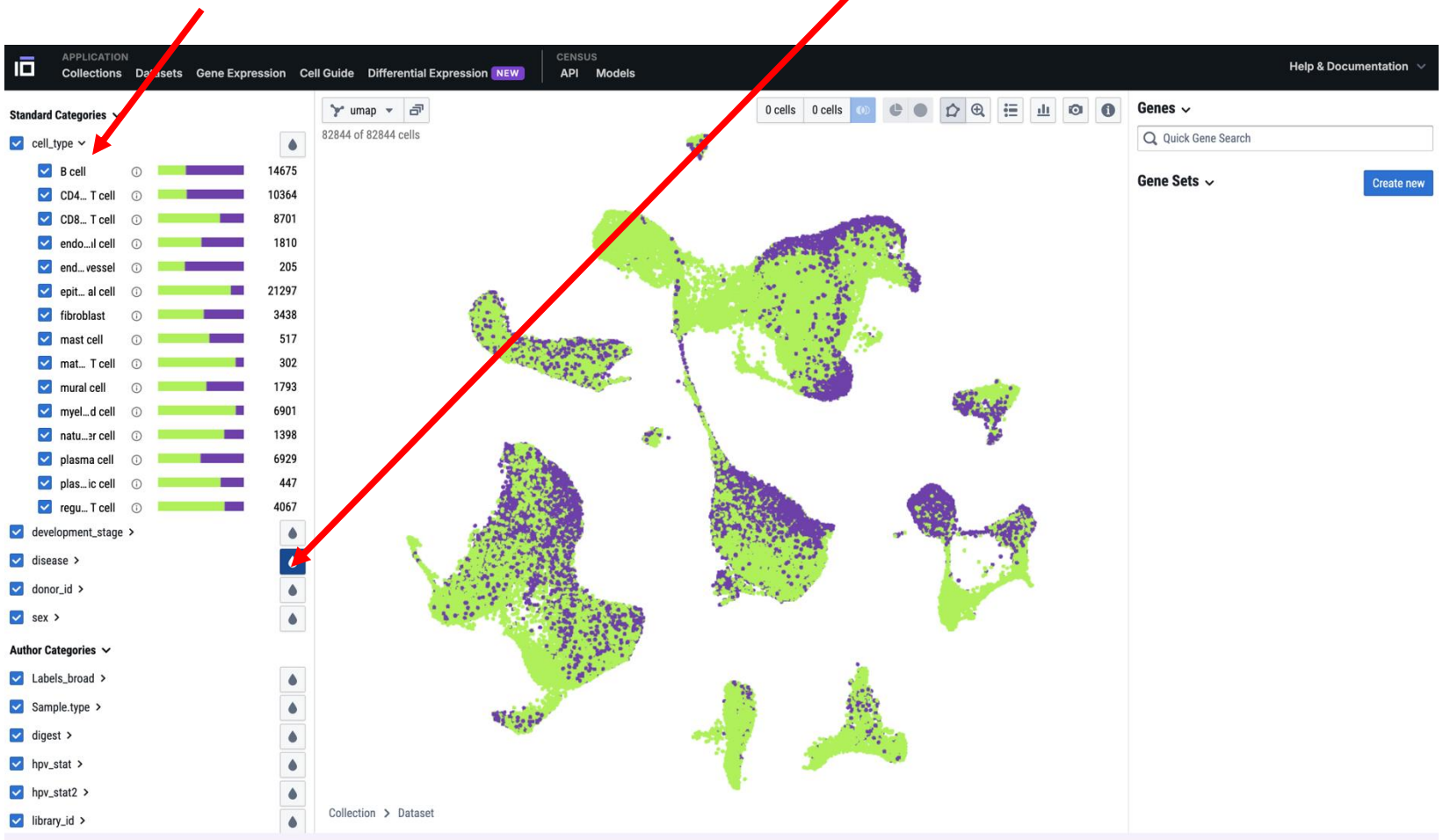
CellxGene data analysis

色の選択：注目する項目の💧マークをクリック＝＞そのカテゴリごとに色が付く



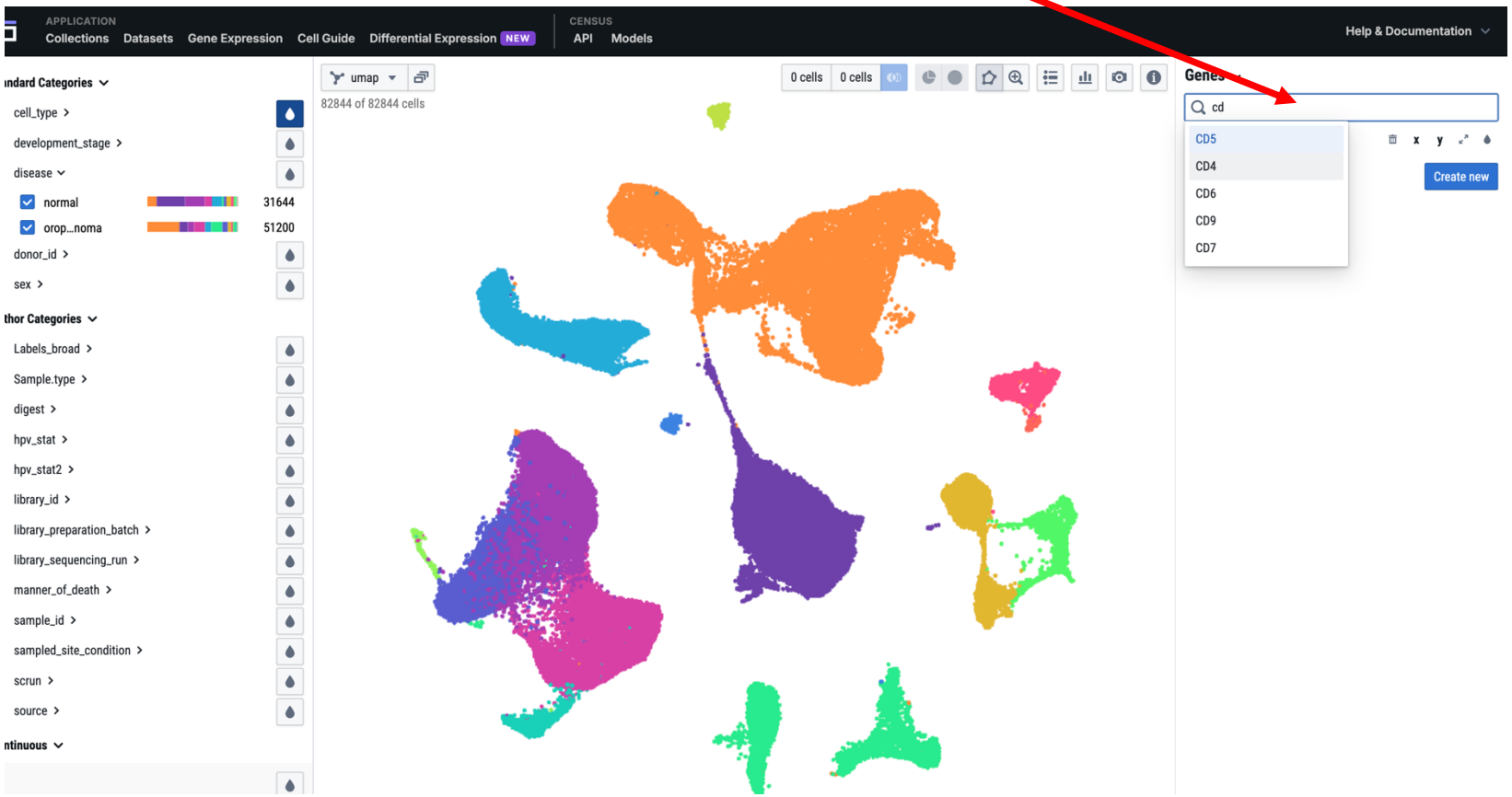
CellxGene data analysis

ある項目の文字列をクリックた後に別の項目で色付けするとその関連がわかる



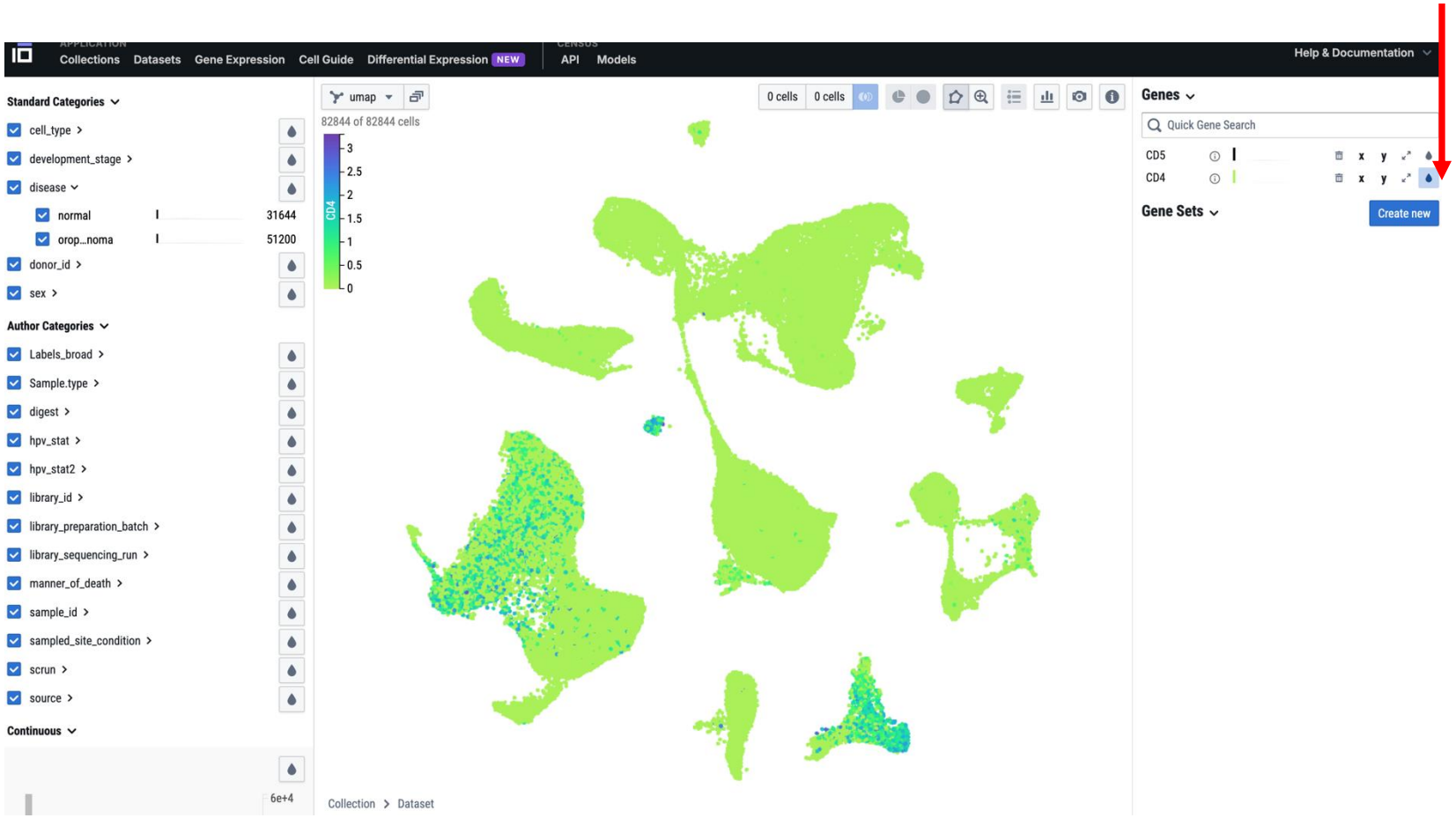
CellxGene data analysis

遺伝子の可視化：1. 遺伝子名を入れる（途中までいれると候補が表示される）



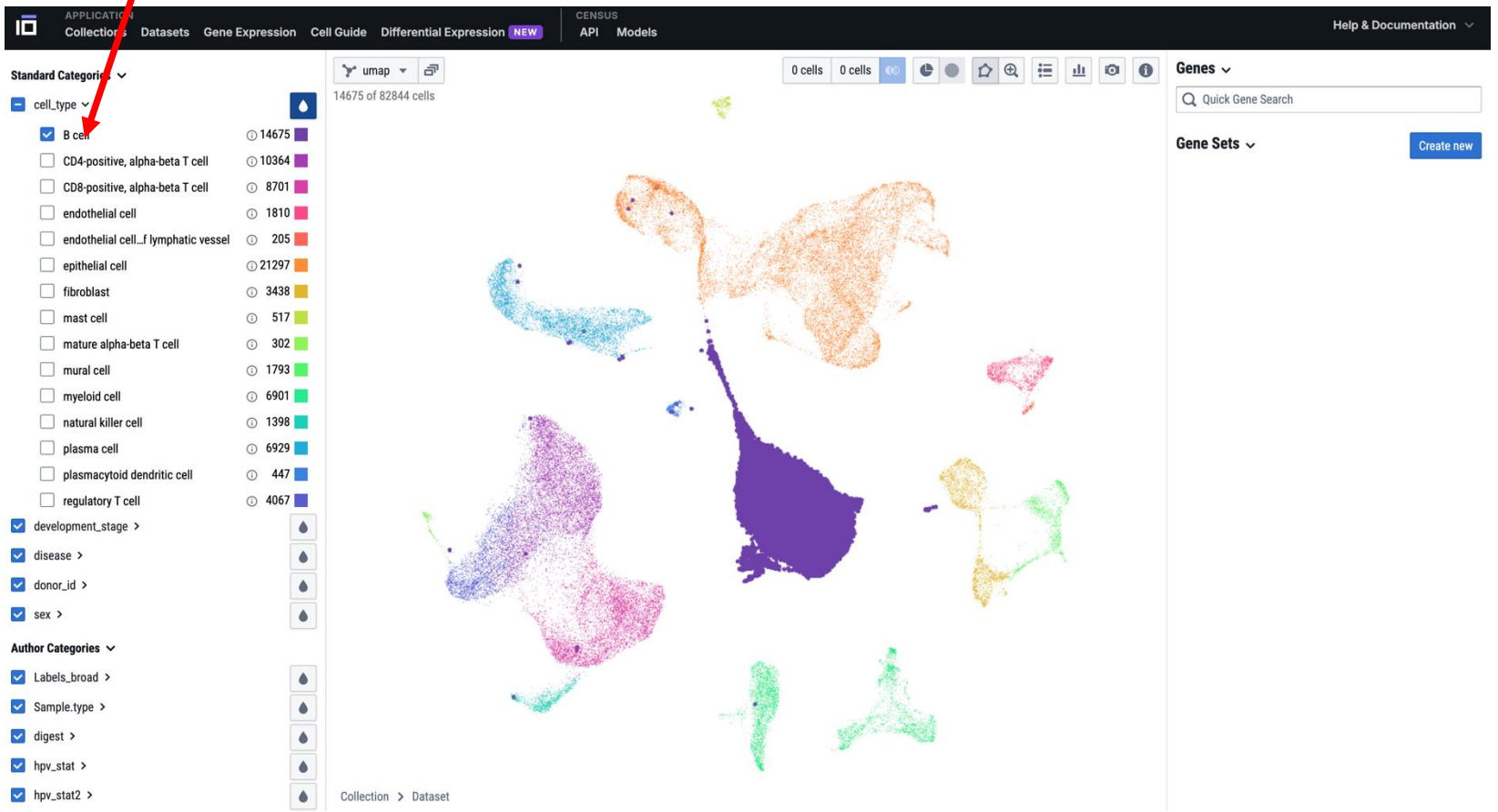
CellxGene data analysis

遺伝子の可視化：2.遺伝子の💧をクリックすると可視化される



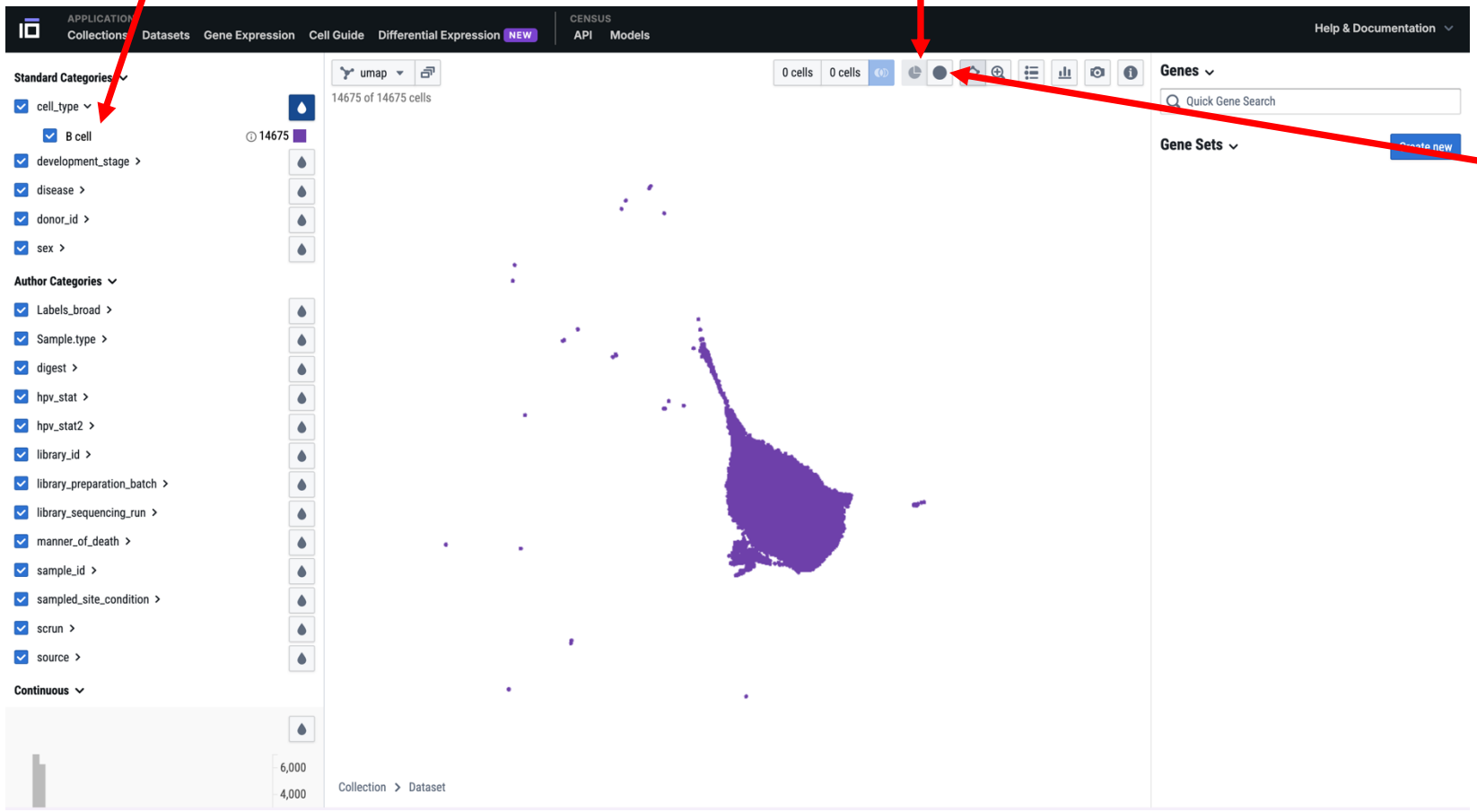
CellxGene data analysis

データの選択：注目するグループをのチェックボックスで選択（注目するもの以外のチェックを外す）



CellxGene data analysis

データの選択２：
1. 注目するグループをのチェックボックスで選択　2. サブセットボタンを押すと選択された集団だけ残る

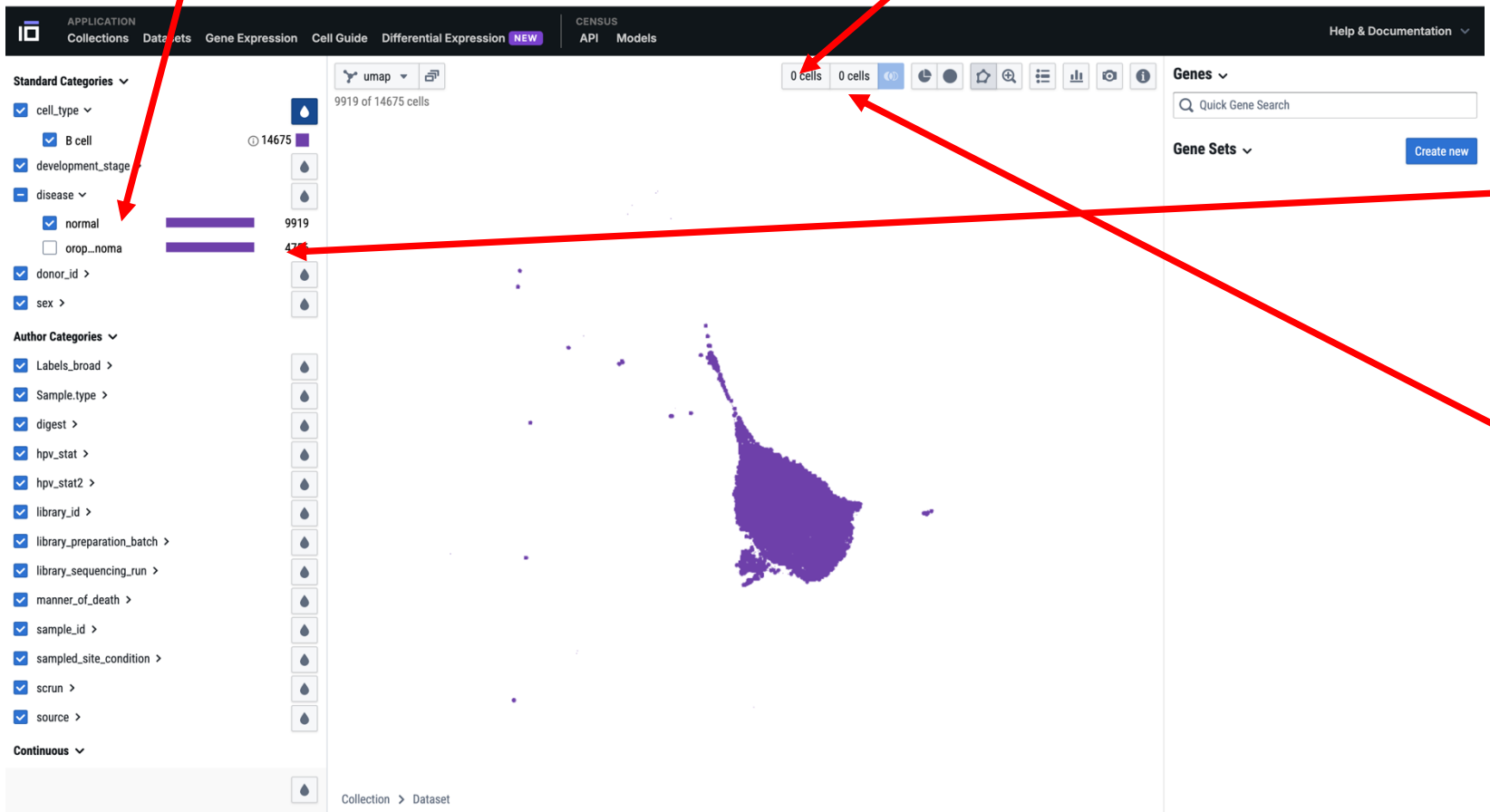


3. 隣の●ボタンでリセットできる

CellxGene data analysis

変動遺伝子解析（Bcell内でnormalとdiseaseの比較）：

1. 注目するグループ（例：normal）を選択　2. [0 cells]というボタンの左を押すと、選択されたグループが登録される



3. 比較対象となる別のグループ（disease）を選択

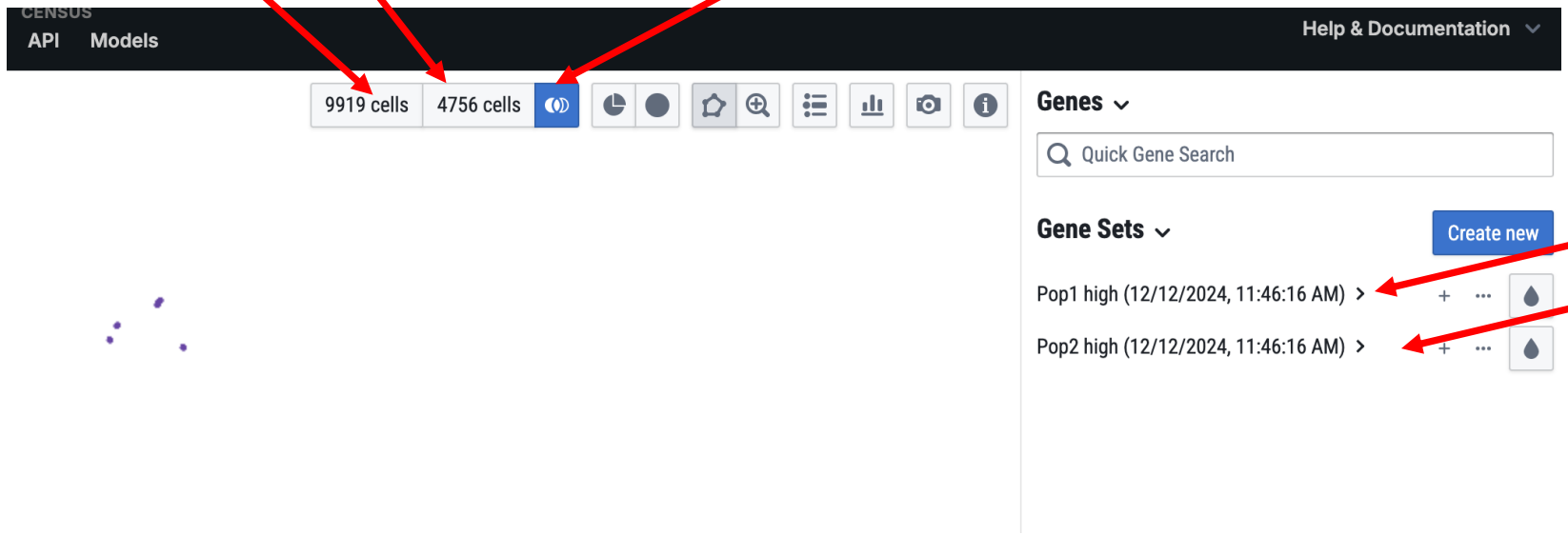
4. 右側の[0cells]ボタンを押して登録

CellxGene data analysis

変動遺伝子解析（Bcell内でnormalとdiseaseの比較）：

1. 2つのグループが登録された状態

2. となりのボタンを押すと２グループの比較で変動遺伝子を検出



3. 発現に差がある遺伝子のリストが作成される