공개된 scRNA-seq 데이터 구하기

AUTHOR
Taeyoon kim

PUBLISHED
October 16, 2023

1 배경 지식

오늘날 대부분의 저널과 연구비 지원 기관은 과학자가 생성한 모든 게놈/염기서열 데이터를 출판과 함께 공개 리포지토리에 보관하도록 요구합니다. 또한 `GitHub`과 같은 코드 저장소를 통해 분석을 위해 개발된 코드를 공유해 재현성과 오픈 액세스를 촉진할 수 있습니다. 여기에서는 `RNA-seq` 데이터에 대한 데이터베이스를 간략히 소개한 다음 scRNA-seq 데이터를 불러오는 방법을 설명합니다.

2 데이터베이스 종류

scRNA-seq 연구가 증가함에 따라 기존 데이터베이스에 단일 세포 데이터를 수용하도록 조정되고 있기도 하지만 새로운 데이터베이스도 만들어지고 있습니다. 아래 목록은 현재 가장 많이 사용되는 데이터베이스입니다.

1. GEO/SRA
2. Single Cell Expression Atlas
3. Single Cell Portal
4. CZ Cell x Gene Discover

`GEO/SRA`는 마이크로어레이, 차세대 시퀀싱 및 기타 형태의 고처리량 기능 게놈 데이터의 포괄적인 세트를 보관하고 자유롭게 배포하는 공개 저장소입니다. 데이터 저장소 외에도, 사용자가 GEO에 저장된 연구와 유전자 발현 패턴을 쿼리하고 다운로드할 수 있는 웹 기반 인터페이스와 애플리케이션 모음을 사용할 수 있습니다. `Single Cell Expression Atlas`는 EMBL에서 호스팅하는 데이터베이스로 다양한 데이터셋을 탐색하고 다운로드 할 수 있습니다.

scRNA-seq 전용으로 만들어진 새로운 데이터베이스에는 MIT와 하버드의 Broad Institute에서 호스팅하는 `Single Cell Portal`, 챈-주커버그 재단에서 호스팅하는 `CZ Cell x Gene Discover` 데이터베이스가 있습니다.

3 보관 데이터들의 유형

연구 결과를 발표하는 저널에서 시퀀스 데이터를 `GEO`와 같은 MIAME 또는 MINSEQE 호환 공개 저장소에 예치하도록 요구합니다. 따라서 데이터베이스에는 제출자가 제공한 원본 데이터(시리즈, 샘플 및 플랫폼)와 처리된 데이터셋이 저장됩니다. 다만 저널 출판은 데이터를 제출하기 위한 필수 요건이 아닙니다. 데이터베이스에서 제공되는 데이터의 유형은 크게 다음 2가지가 있습니다.

1. 3개 파일로 분리된 데이터 매트릭스 파일(barcodes.tsv.gz, features.tsv.gz, and matrix.mtx.gz)
2. 단일 `HDF5` 포멧 파일

> **Note**
> HDF5(Hierarchical Data Format Version 5)는 과학 데이터를 파일에 저장하기 위한 차세대 범용 표준

4 데이터 불러오기

scRNA-seq 데이터 분석 도구 또한 크게 2가지 `Seurat`과 `Scanpy`로 구분할 수 있으며, 각각 `R`언어와 `Python`을 사용한다는 중요한 차이점이 존재합니다. 아래는 각각의 도구와 데이터 유형에 따른 예시 코드입니다.

4.1 `Seurat`을 사용하는 방법

4.1.1 데이터 매트릭스 파일

```r
library(Seurat)

file_path <- "path/to/data/directory"
data <- Read10X(data.dir=file_path)
seurat_object <- CreateSeuratObject(counts = data$`Gene Expression`)
```

4.1.2 HDF5 포멧 파일

```
library(Seurat)

data <- Read10X_h5("../{filtered_feature_bc_matrix}.h5")
seurat_obj <- CreateSeuratObject(counts=data)
```

4.2 `Scanpy` 을 사용하는 방법

4.2.1 데이터 매트릭스 파일

```
import scanpy as sc

file_path = "path/to/data/directory"
adata = sc.read_10x_mtx(file_path, var_names="gene_symbols", cache=True)
```

4.2.2 HDF5 포멧 파일

```
import scanpy as sc

adata = sc.read_10x_h5("../{filtered_feature_bc_matrix}.h5")
adata.var_names_make_unique()
```

5 결론

데이터를 불러온 다음 가장 먼저 할 일은 데이터의 전처리와 품질 관리를 신중하게 수행하는 것입니다. 이후 데이터를 적절하게 시각화하고 분석을 수행하여 세포 유형의 특성과 생물학적 의미를 파악할 수 있을 것입니다.