CITE-seq analysis proof-of-concept

Patrick Cherry

2024-06-20

```
[1] "GSE100866_CBMC_8K_13AB_10X-ADT_clr-transformed.csv.gz"
[2] "GSE100866_CBMC_8K_13AB_10X-ADT_umi.csv.gz"
[3] "GSE100866_CBMC_8K_13AB_10X-RNA_umi.csv.gz"
[4] "GSE100866_CD8_merged-ADT_clr-transformed.csv.gz"
[5] "GSE100866_CD8_merged-ADT_umi.csv.gz"
[6] "GSE100866_CD8_merged-RNA_umi.csv.gz"
[7] "GSE100866_PBMC_vs_flow_10X-ADT_clr-transformed.csv.gz"
[8] "GSE100866_PBMC_vs_flow_10X-ADT_umi.csv.gz"
[9] "GSE100866_PBMC_vs_flow_10X-RNA_umi.csv.gz"
```

```
/Users/patrick/bfx git projects/CITE-seq/2024-06-20/CITE-seq
```

Intro

CITE-seq using data from Stoeckius, et al (Nat Methods 14, 865–868 (2017) 10.1038/nmeth.4380) in GSE100866

Seurat multi-modal clustering

Methods Intro

From Supplementary Figure 1 CITE-seq library preparation:

> Illustration of the DNA-barcoded antibodies used in CITE-seq. (b) Antibody-oligonucleotide complexes appear as a high-molecular-weight smear when run on an agarose gel (1). Cleavage of the oligo from the antibody by reduction of the disulfide bond collapses the smear to oligo length (2). (c) Drop-seq beads are microparticles with conjugated oligonucleotides comprising a common PCR handle, a cell barcode, followed by a unique molecular identifier (UMI) and a polyT tail. (d) Schematic illustration of CITE-seq library prep in Drop-seq (downstream of Fig. 1b). Reverse transcription and template switch is performed in bulk after emulsion breakage. After amplification, full length cDNA and antibody-oligo products can be separated by size and amplified independently (also shown in d) (e) Reverse transcription and amplification produces two product populations with distinct sizes (left panel). These can be size separated and amplified independently to obtain full length cDNAs (top panel, capillary electrophoresis trace) and ADTs (bottom panel, capillary electrophoresis trace).

Read in data

count data

```
[1] 36280  8617
```

`dim()` outputs (columns, rows); the CBMC matrix (cord blood mononuclear cells) contains 36280 features and 8617 samples (single cell droplets).

```
[1] 20501  8617
```

`CollapseSpeciesExpressionMatrix()` is a convenience function for slimming down a multi-species expression matrix, when only one species is primarily of interest. Given the default parameter of `ncontrols = 100`, this command keeps only the top 100 features detected from each species in each sample. This matrix went from 36280 to 20501 features, which is a 43% reduction.

ADT UMI matrix

```
[1]   13 8617
```

Quick matrix QC

The number of rows (samples / UMIs) matches the RNA counts matrix; we have corresponding sample data.

```
[1] 1
```

And the names of the samples all match.

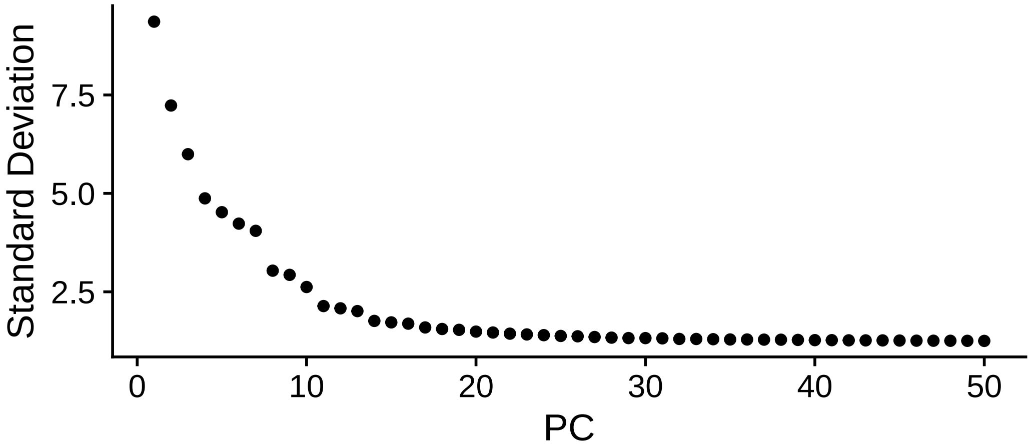
Seurat object & cluster

```
Warning: Feature names cannot have underscores ('_'), replacing with dashes
('-')
```

```
Normalizing layer: counts
```

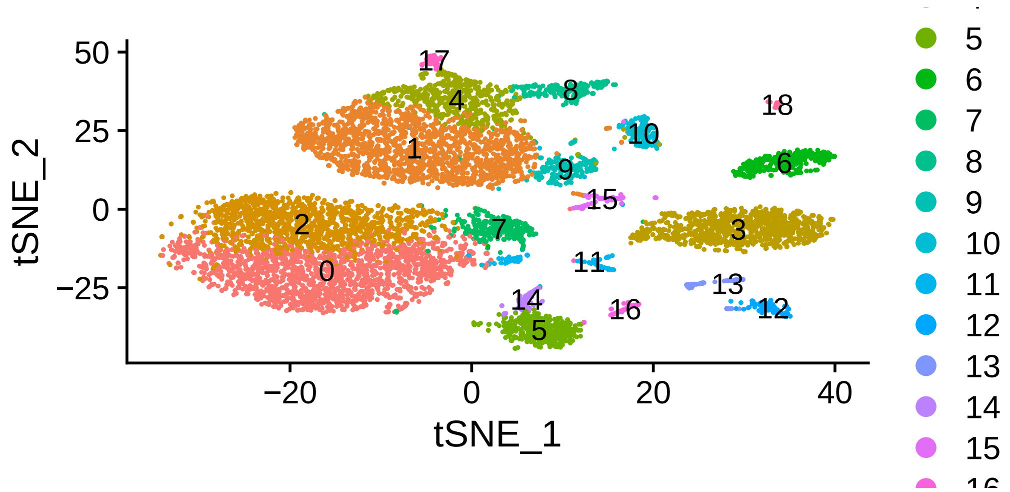
```
Finding variable features for layer counts
```

```
Centering and scaling data matrix
```



```
Computing nearest neighbor graph


Computing SNN


Modularity Optimizer version 1.3.0 by Ludo Waltman and Nees Jan van Eck

Number of nodes: 8617
Number of edges: 343912

Running Louvain algorithm...
Maximum modularity in 10 random starts: 0.8613
Number of communities: 19
Elapsed time: 0 seconds
```

Clustering t-SNE plot



Incorporate protein expression (antibody barcodes) to the Seurat object

The above code adds a new assay called "ADT" to the Seurat object `cmbc`. We can confirm it's added with the following `GetAssayData()` command.

```
3 x 3 sparse Matrix of class "dgCMatrix"
    CTGTTTACACCGCTAG CTCTACGGTGTGGCTC AGCAGCCAGGCTCATT
CD3                60               52               89
CD4                72               49              112
CD8                76               59               61
```

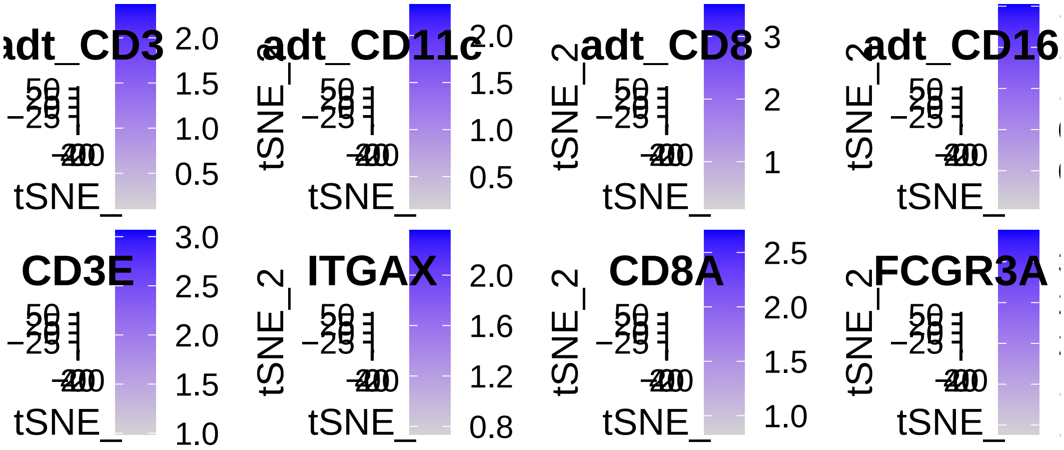
```
 [1] "CD3"    "CD4"    "CD8"    "CD45RA" "CD56"   "CD16"   "CD10"   "CD11c"
 [9] "CD14"   "CD19"   "CD34"   "CCR5"   "CCR7"
```

Now we can repeat the pre-processing (normalization and scaling) steps that we typically run with RNA, but modifying the 'assay' argument.

(For CITE-seq data, the Broad does not recommend typical LogNormalization. Instead, they use a centered log-ratio (CLR) normalization, computed independently for each feature. This is a slightly improved procedure from the original publication.)

```
Normalizing across features
```

```
Centering and scaling data matrix
```


adt_CD3
tSNE_
adt_CD11c
tSNE_2
tSNE_
adt_CD8
tSNE_2
tSNE_
adt_CD16
tSNE_2
tSNE_
CD3E
tSNE_
ITGAX
tSNE_2
tSNE_
CD8A
tSNE_2
tSNE_
FCGR3A
tSNE_2
tSNE_