Отчет по теме введение в секвенирование единичных клеток

Горохов Никита

Ноябрь 2019

Исследуем субпопуляции клеток

Были установлены и загружены следующие пакеты:

```
library(Seurat)
library(velocyto.R)
library(SeuratWrappers)
library(ggplot2)
library(cowplot)
library(gridExtra)
library(dplyr)
library(monocle3)
```

Теперь загрузим наши данные

```
load(file = "hse.object.Rdata")
data <- sc.object
```

В этом объекте хранятся два объекта: RNA (все гены) и integrated (остались только наиболее вариабельные гены)

```
dim(data@assays$RNA@data)
[1] 19602  2233
dim(data@assays$integrated@data)
[1] 2684 2233
```

Мы будем работать с integrated. Последовательно применим линейные и нелинейные преобразования (PCA, tSNE, uMAP) и разобьем данные на кластеры.

```
data <- RunPCA(object = data, verbose = FALSE, assay = "integrated", npcs = 1
data <- RunTSNE(object = data , dims = 1:50, verbose = FALSE)
data <- RunUMAP(object = data , dims = 1:50, verbose = FALSE)
data <- FindNeighbors(object = data , dims = 1:50, verbose = FALSE)
data <- FindClusters(object = data , resolution=0.2, verbose = FALSE)
DimPlot(object = data, reduction = "pca")
DimPlot(object = data, reduction = "tsne")
DimPlot(object = data, reduction = "umap")
```

Результаты можно увидеть на Рис. 1 – 3. Как и ожидалось, нелинейные методы (uMAP и tSNE) преобразуют данные лучше, чем линейные (PCA). Также посмотрим на типы

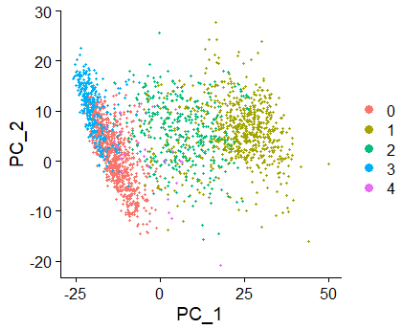


Рис. 1: PCA

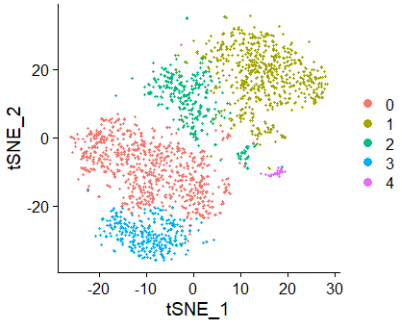


Рис. 2: tSNE

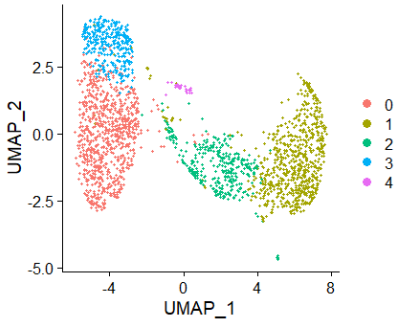


Рис. 3: uMAP

клеток:

```
table(data$cells)
  1   2   3   4
124 944 403 762
```

Больше всего у нас представлено клеток 2 и 4 класса. Однако алгоритм разбил на 5 классов:

```
table(data$seurat_clusters)
  0   1   2   3   4
816 708 354 328  27
```

Дифференциально-экспрессионный анализ и анализ значимых генов

Теперь найдем значимые гены (гены которые экспрессируют только в определенных кластерах)

```
pbmc.markers.cell <- FindAllMarkers(object = data, only.pos = TRUE,
                                    min.pct = 0.5, logfc.threshold = 0.2,
                                    test.use = "roc")
dim(pbmc.markers.cell)
[1] 901   7
```

По заданым критерями нашелся 901 ген. Посмотрим каким кластерам принадлежат эти гены.

```
table(pbmc.markers.cell$cluster)
  0   1   2   3   4
215 144  45 373 124
```

Выберем топ 2 генов из каждого кластера.

```
top2 <- pbmc.markers.cell %>% group_by(cluster) %>% top_n(n = 2, wt = avg_dif
```

Построим heatmap и featureplot.

```
DoHeatmap(object = data, features = top2$gene)
FeaturePlot(object = data, features = top2$gene)
```

Теперь попробуем сделать все тоже самое, только дополнительно применим информацию о классах (cells)

```
Idents(data) = data$cells
all.markers.type = FindAllMarkers(object = data, min.pct = 0.5, logfc.thresho
new_top2 = all.markers.type %>% group_by(cluster) %>% top_n(n = 2, wt = avg_l
DoHeatmap(object = data, features = new_top2$gene)
FeaturePlot(object = data, features = new_top2$gene)
```

И теперь сравним полученные результаты. На рис 4 и 5 изображены два heatmap. На левом heatmap граница между кластерами 1 и 2, 3 и 4 особо не прослеживается. Также гены PTTG1 и HMGB2 одинаково экспрессируют в 0 и 1 кластерах. После применения функции Ident() картина поменялась в лучшую сторону. Рис 6 7 подтверждают также подтверждают это.

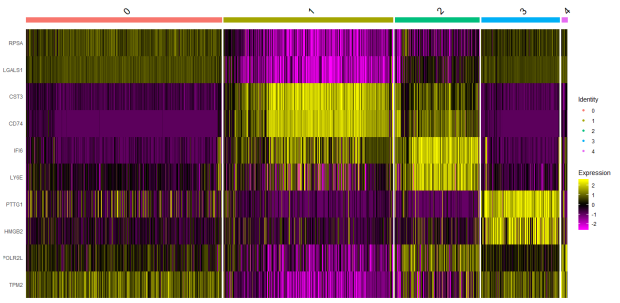


Рис. 4: Heatmap before

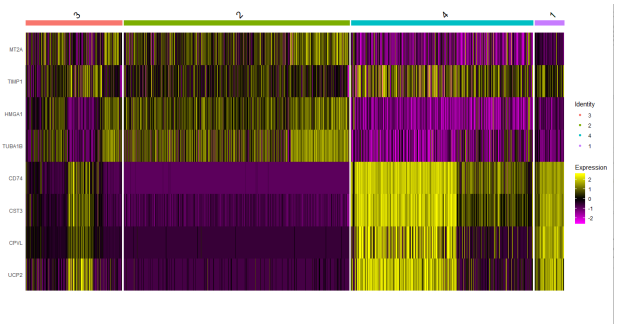


Рис. 5: Heatmap after

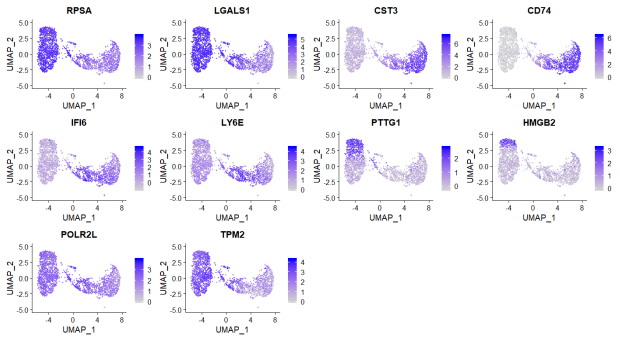


Рис. 6: Feature plot before

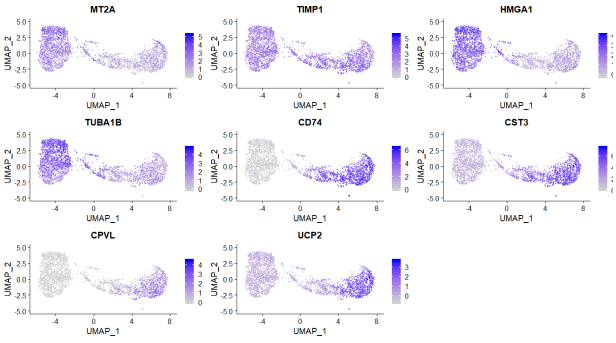


Рис. 7: Feature plot after

Найдем биологический смысл полученных результатов. Для этого подсчитаем топ 5 генов для каждого кластера (см. табл.1). Попробуем найти связь между этими 20 генами. Для этого используем сервис **Enrichr**. Переходим по первому совпадению (ChEA

2016 - содержит результаты Chip-Seq). И выбираем вкладку Clustergram. На рис. 8 показана кластерграмма (по оси x наборы генов, а по y гены). Попробуем проанализировать полученные результаты. Можно отметить два набора, для которых найдено больше всех совпадений:

- CIITA: 3 гена из 4 кластера и 2 гена из 1 кластера.
- XRN2: 4 гена из 2ого кластера и по 1 гену из 1ого и 4ого.

Итого, с биологической точки зрения можно сделать вывод, что наиболее экспрессионные гены из 2 и 4 кластеров связаны между собой. Также это подтвержает очень низконе значение p-value.

p-val-adj	cluster	gene
3.16e-3	3	PNRC1
1.83e-2	3	MT2A
1.62e-1	3	VMP1
3.33e-1	3	TIMP1
1.00e+0	3	CITED2
1.69e-258	2	HNRNPA1
7.92e-227	2	LGALS1
1.91e-208	2	HMGA1
9.02e-161	2	TAGLN
1.07e-154	2	TUBA1B
3.38e-268	4	CD74
6.37e-247	4	IFI6
1.64e-237	4	HLA-DRA
6.98e-237	4	HLA-DRB1
4.77e-236	4	CST3
4.23e- 52	1	HLA-DQB1
1.33e- 49	1	CPVL
1.10e- 37	1	HLA-DPA1
1.12e- 30	1	UCP2
1.46e- 27	1	C1orf54

Таблица 1: 5 наиболее значимых генов в каждом кластере

Разбор статьи

Я выбрал статью [1]. В ней использовался протокол Drop-Seq. Пайплайн изображен на Рис. Этапы:

- Выделение единичных клеток
- Изолирование каждой клетки с последующим навешиванием баркода в капле (droplet)

- лизирование клетки
- получение мРНК, формирую STAMP (single-cell transcriptomes attached to microparticles)
- Обратная транскрипция, амплификация
- секвенирование тысячи STAMPs
- декодирование (используя баркоды)
- каждый олигонуклеотид состоит из 4 частей: PCR id, barcode, UMI и TTTT...
-

Обработка:

- В качестве иструмента они использовали библиотеку Seurat
- Всего было 49.300 клеток и из них отобрали клетки, в которых обнаружилось более 900 генов. Итого получилось 13,155 клеток.
- Затем на отобранных клетках применили PCA и отобрали 32 статистически значимых компоненты (используя перестановочный тест)
- После применили t-SNE и получили двумерную матрицу (2 компоненты остались).
- Основываясь на полученных компонентах они спроецировали оставшиеся клетки.
- Разбили на 39 кластеров (39 потому что столько популяций было). В каждом кластере было от 50 до 29,400 клеток.

Список литературы

[1] Evan Z. Macosko et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 2015.

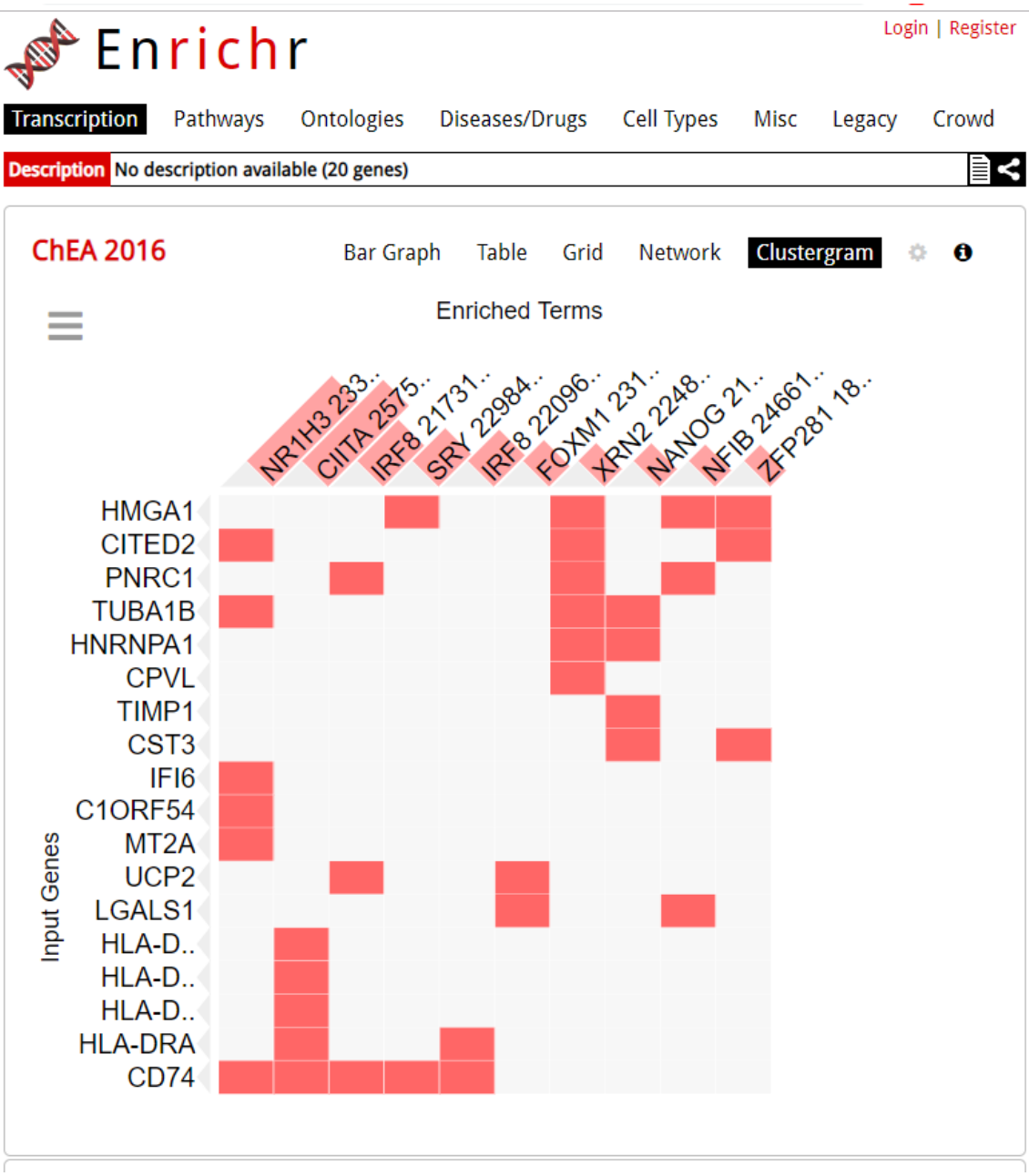
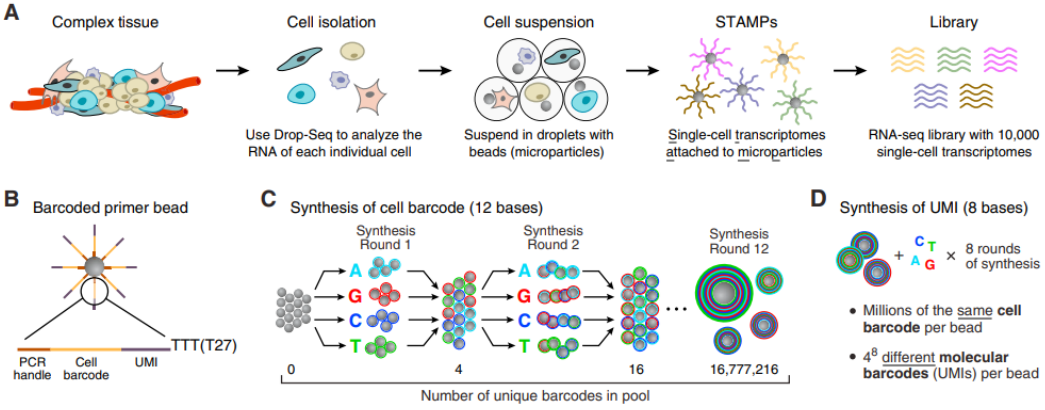


Рис. 8: Кластерграмма 20 генов


A
Complex tissue
Cell isolation
Cell suspension
STAMPs
Library
Use Drop-Seq to analyze the RNA of each individual cell
Suspend in droplets with beads (microparticles)
Single-cell transcriptomes attached to microparticles
RNA-seq library with 10,000 single-cell transcriptomes
B
Barcoded primer bead
TTT(T27)
PCR handle
Cell barcode
UMI
C
Synthesis of cell barcode (12 bases)
Synthesis Round 1
Synthesis Round 2
Synthesis Round 12
A
G
C
T
0
4
16
16,777,216
Number of unique barcodes in pool
D
Synthesis of UMI (8 bases)
8 rounds of synthesis
Millions of the same cell barcode per bead
4^8 different molecular barcodes (UMIs) per bead

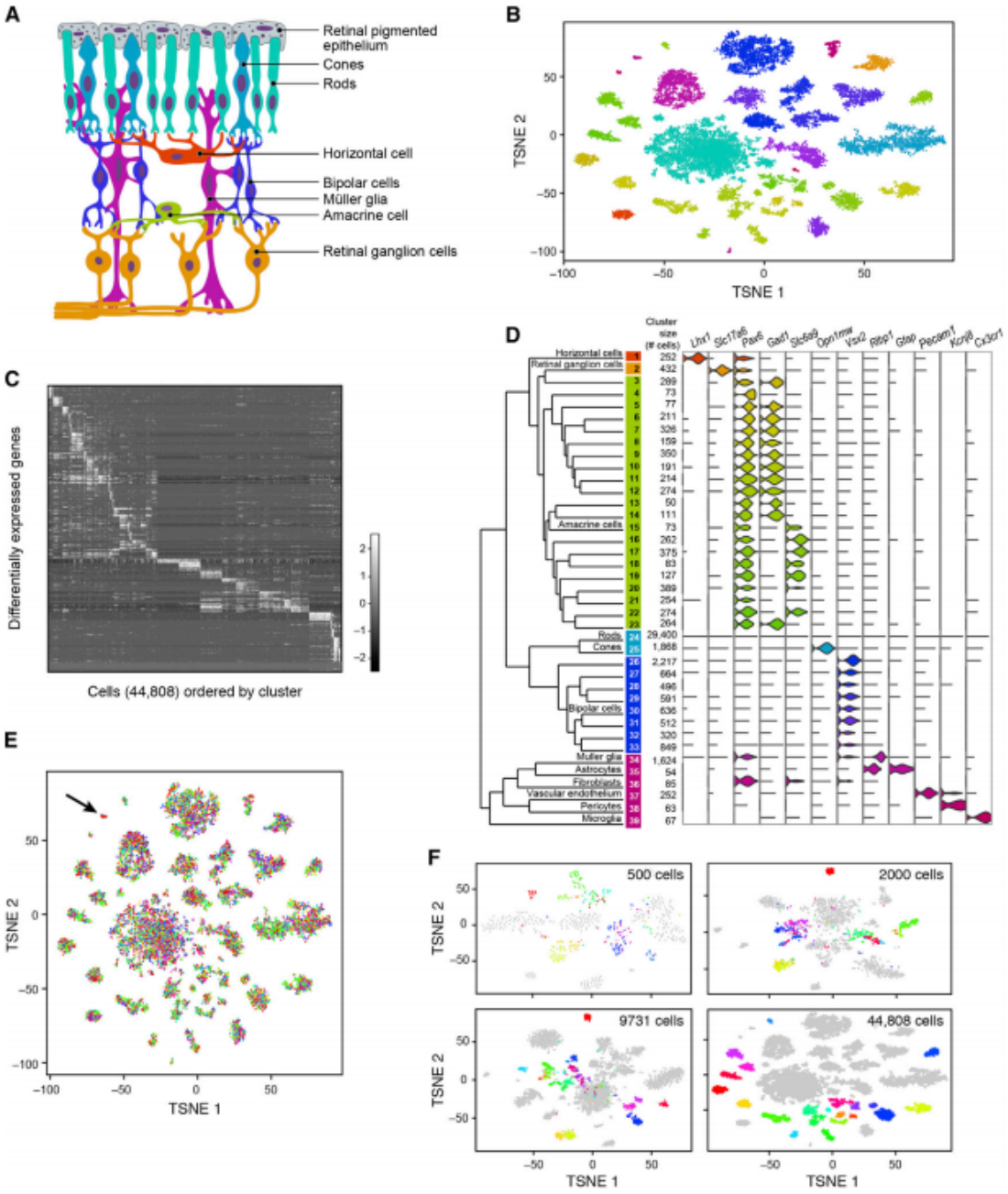


Рис. 9: (A): классы клеток, (B): покрасили согласно классам после кластеризации (D): heatmap + кластеризация (E): покрасили согласно кластерам