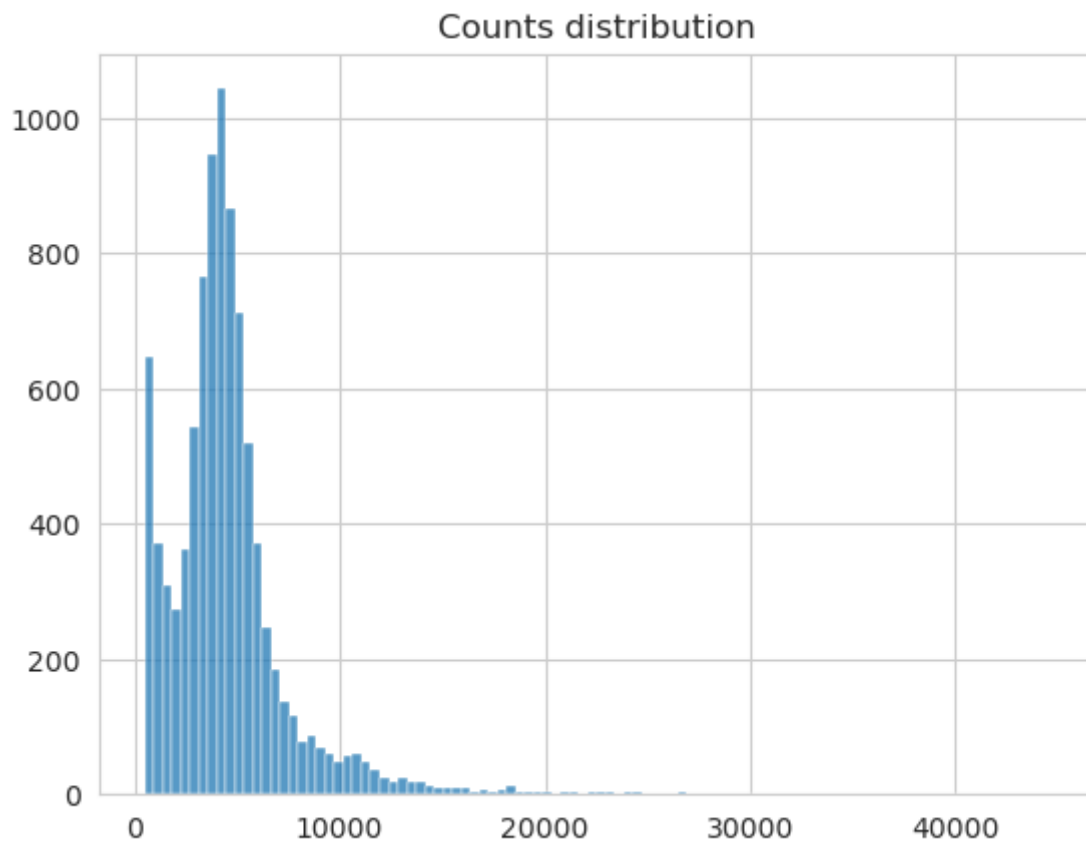


```
AnnData object with n_obs x n_vars = 9234 x 36601
var: 'gene_symbols', 'feature_types'
```

	gene_symbols	feature_types	gene_id
<b>MIR1302-2HG</b>	MIR1302-2HG	Gene Expression	ENSG00000243485
<b>FAM138A</b>	FAM138A	Gene Expression	ENSG00000237613
<b>OR4F5</b>	OR4F5	Gene Expression	ENSG00000186092
<b>AL627309.1</b>	AL627309.1	Gene Expression	ENSG00000238009
<b>AL627309.3</b>	AL627309.3	Gene Expression	ENSG00000239945

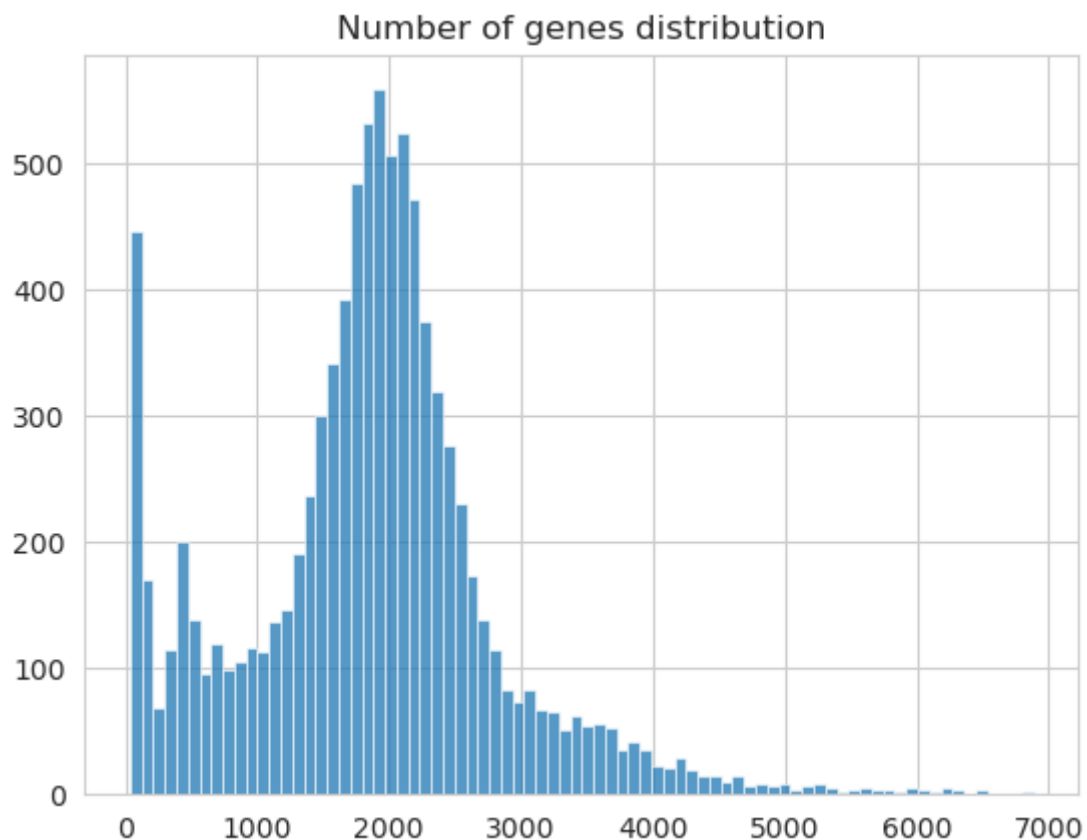


По оси X - количество молекулы мРНК в одной клетке.

По оси Y -много клеток имеют такое количество мРНК

```
: np.median(cell_counts)|
```

```
: 4183.5
```



Медиана числа обнаруженных генов на клетку составила 4183.5, что свидетельствует о достаточной глубине секвенирования. Большинство клеток экспрессируют 2000–3000 генов

Top 30 genes:

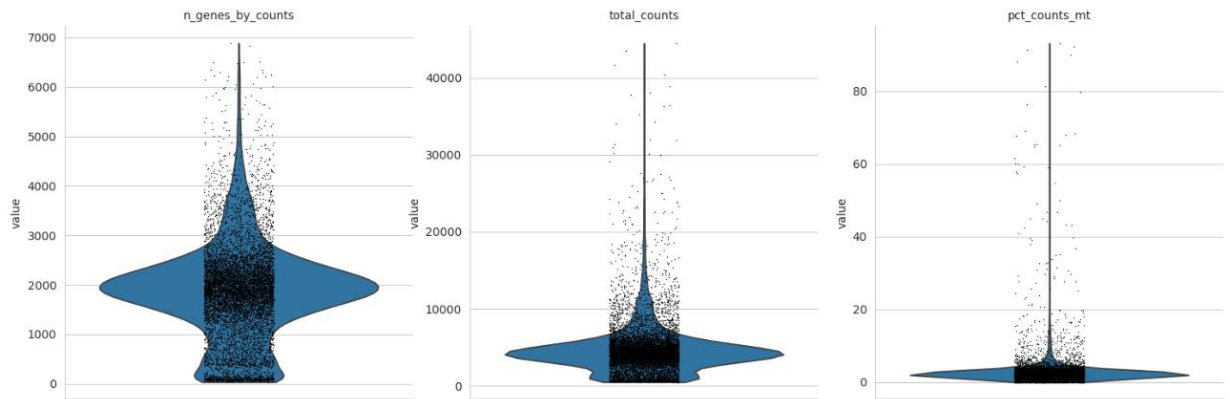
MIR1302-2HG FAM138A OR4F5 AL627309.1 AL627309.3 AL627309.2 AL627309.5 AL627309.4 AP006222.2 AL732372.1 OR4F29 AC114498.1 OR4F16 AL669831.2 LINC01409 FAM87B LINC01128 LINC00115 FAM41C AL645608.6 AL645608.2 AL645608.4 LINC02593 SAMD11 NOC2L KLHL17 PLEKHN1 PERM1 AL645608.7 HES4

Index(['MT-ND1', 'MT-ND2', 'MT-CO1', 'MT-CO2', 'MT-ATP8', 'MT-ATP6', 'MT-CO3', 'MT-ND3', 'MT-ND4L', 'MT-ND4', 'MT-ND5', 'MT-ND6', 'MT-CYB'], dtype='object')

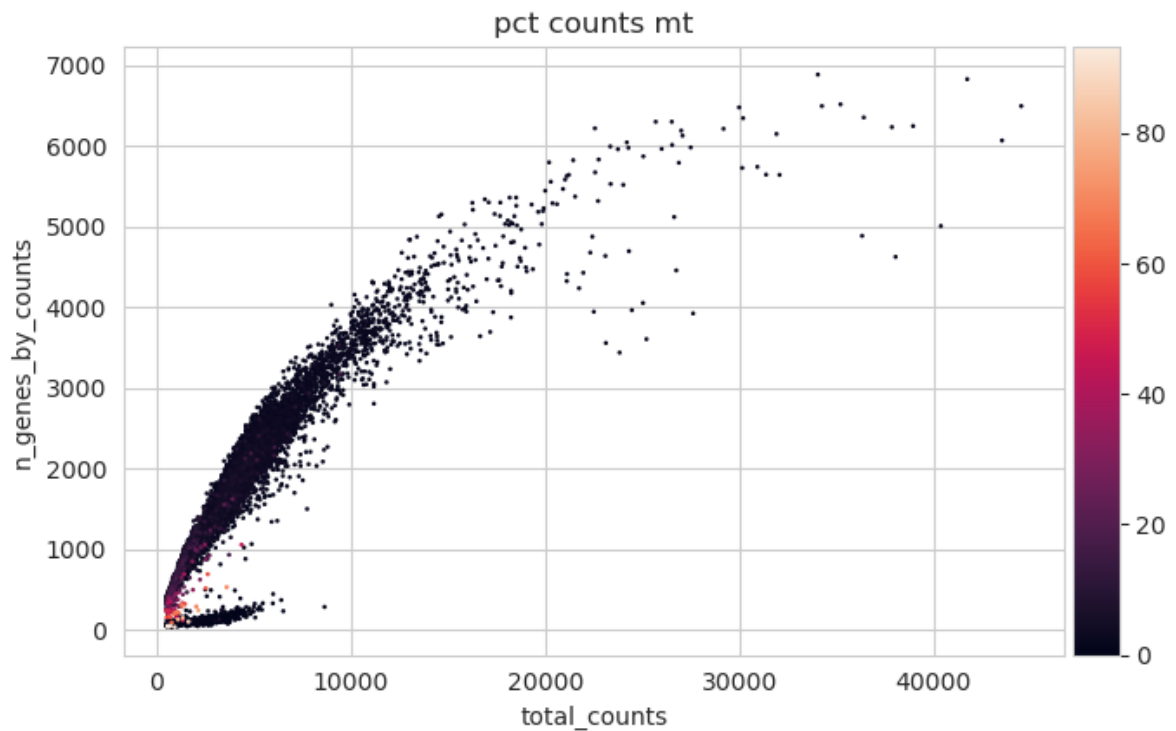
выделяем где много митохондриальных генов, так клетки их содержащие в основном мёртвые) так как мембрана клеток разрушается

	gene_symbols	feature_types	gene_id	mt	ribo
<b>MIR1302-2HG</b>	MIR1302-2HG	Gene Expression	ENSG00000243485	False	False
<b>FAM138A</b>	FAM138A	Gene Expression	ENSG00000237613	False	False
<b>OR4F5</b>	OR4F5	Gene Expression	ENSG00000186092	False	False
<b>AL627309.1</b>	AL627309.1	Gene Expression	ENSG00000238009	False	False
<b>AL627309.3</b>	AL627309.3	Gene Expression	ENSG00000239945	False	False

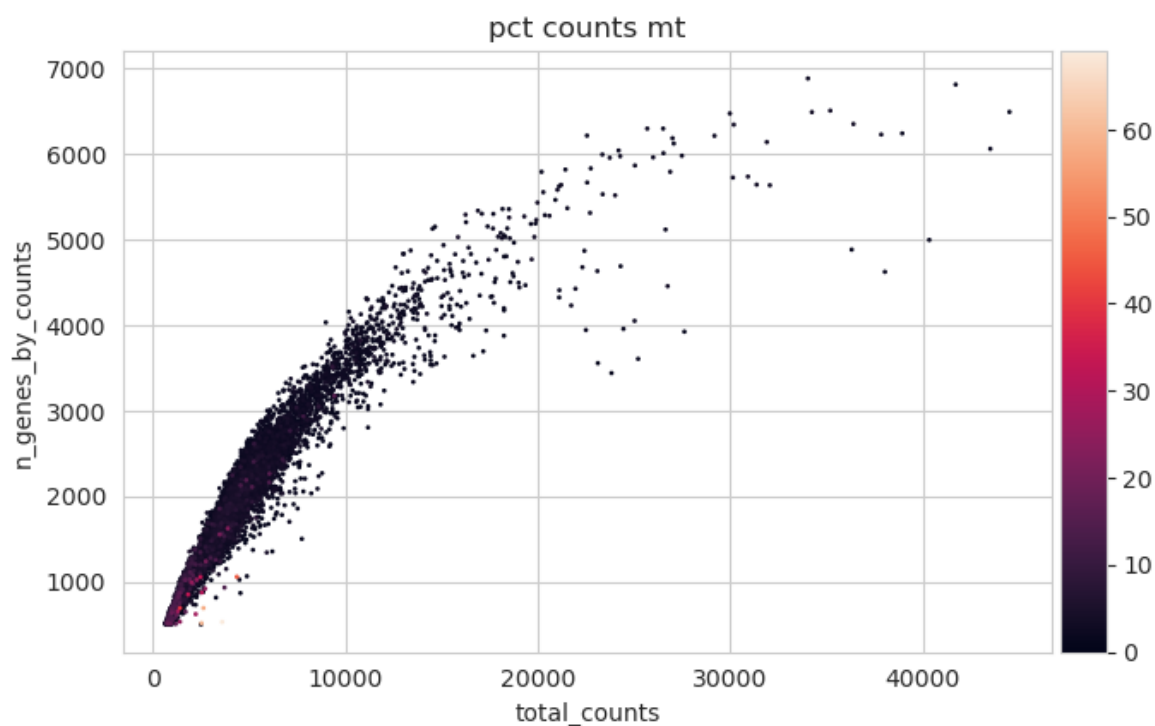
Пометили гены)



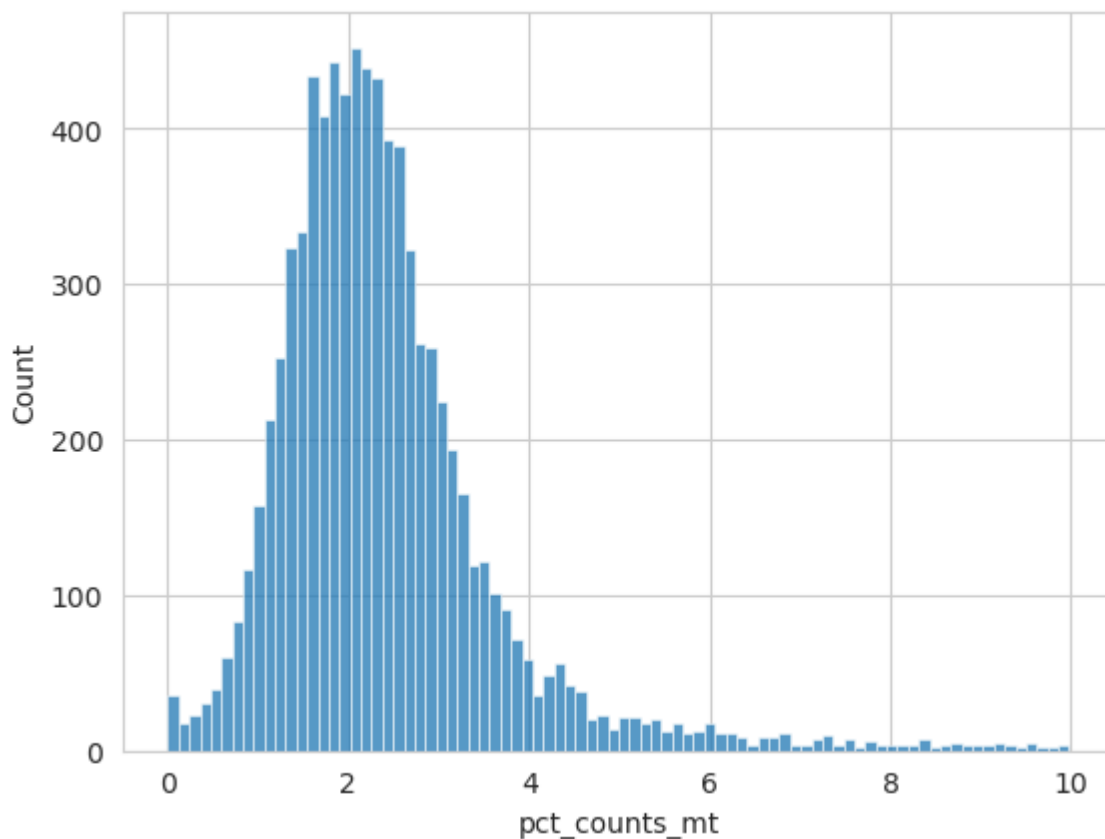
большинство клеток имеют достаточное количество генов около 2000–3000 и мрнк 5000–8000. Процент митохондриальной РНК сосредоточен на уровне  $< 10\%$ , что указывает на хорошее качество образца))))))



Хорошая корреляция между `total_counts` и `n_genes_by_counts`, что ожидаемо, чем больше генов, тем больше мрнк. Большинство клеток (чёрные точки) имеют низкий процент мит.-генов (0–10%), что свидетельствует о хорошем качестве. Выявлены потенциальные дубликаты (точки с очень высокими значениями обеих метрик) и единичные мёртвые клетки (красные точки в нижней части)



Сделали фильтрацию, убрали клетки, где меньше 500 генов, выкидываем гены, которые экспр. Менее чем 3 клетках, убираем мертвяк, смотрим график красных точек (мёртвые клетки) стало гораздо меньше фильтрация более менее сработала



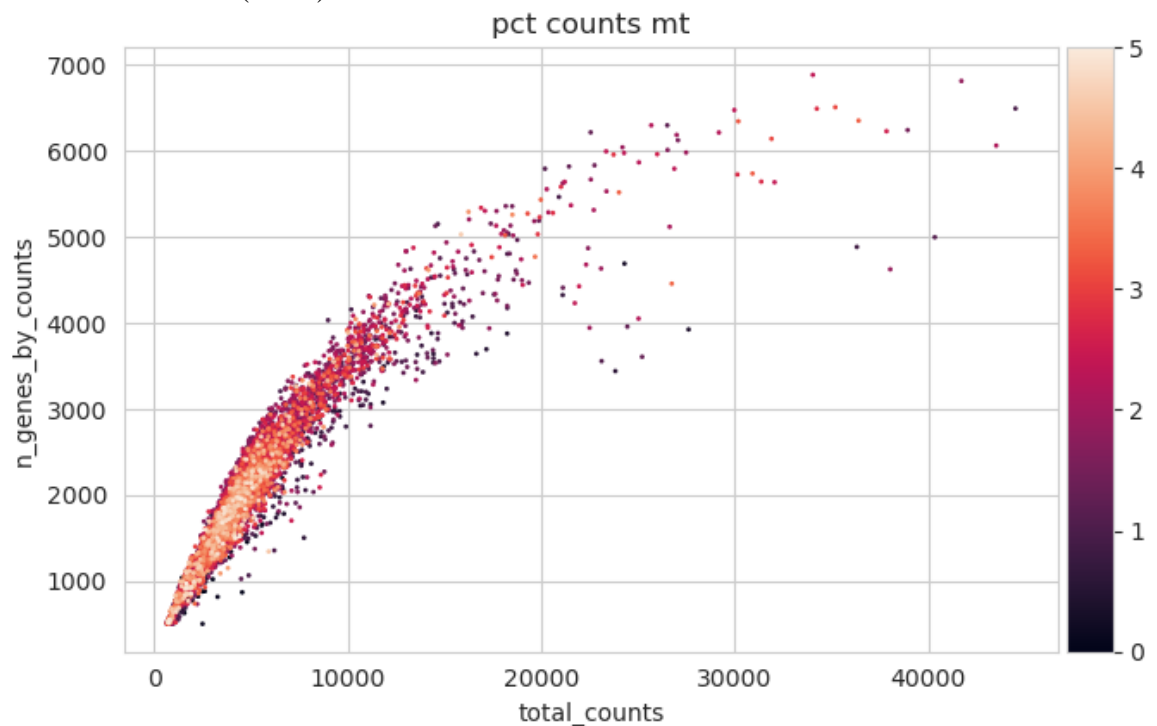
Распределение доли митохондриальной РНК по клеткам пик около 2% и плавный спад к большим значениям, большинство клеток имеют менее 5% митохондриальных чтений, что соответствует хорошему качеству образца, значит мёртвых клеток не так и много

Сделаем порог фильтрации меньше 5 процентов) раньше стоял 10

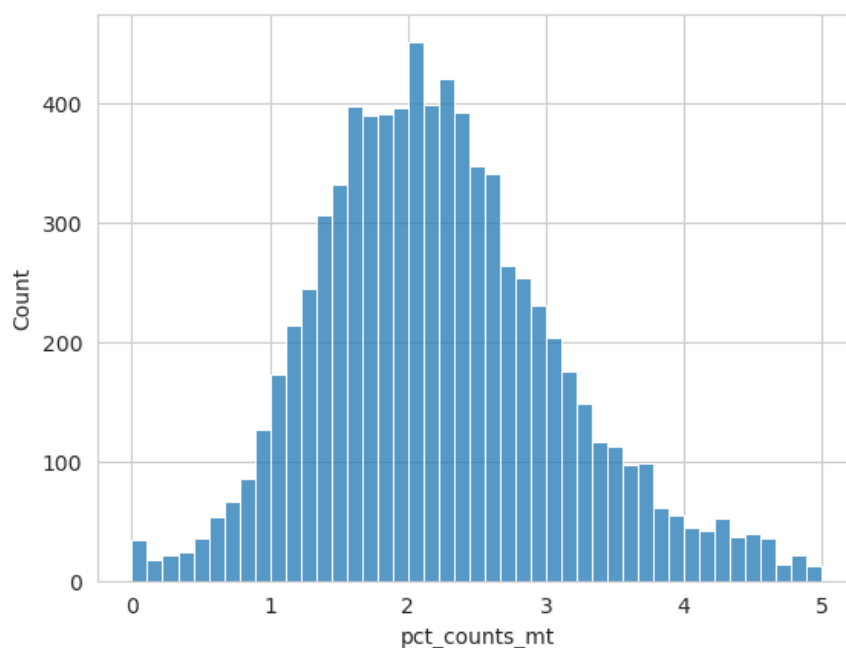
Cell before filterring: 8082

Cell after filterring: 7764

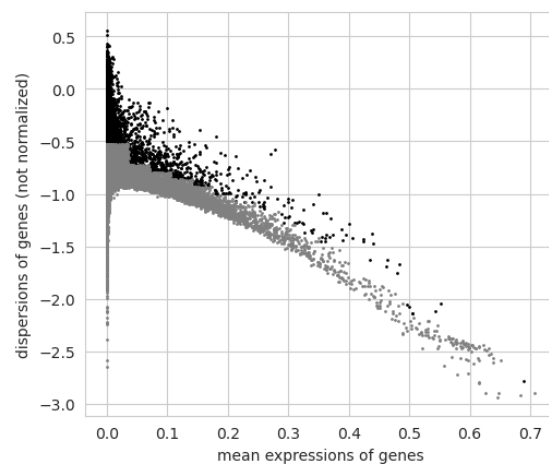
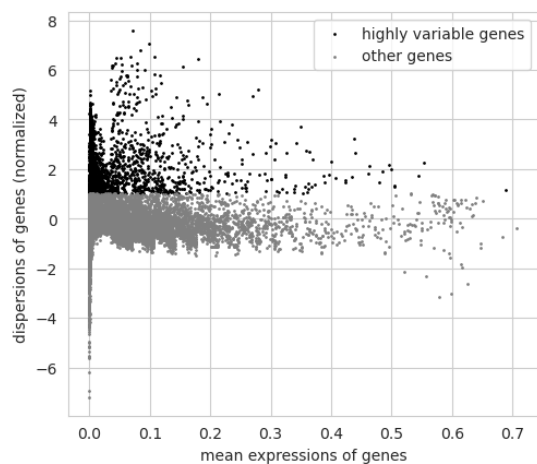
Filtered out: 318 cells (3.9%)



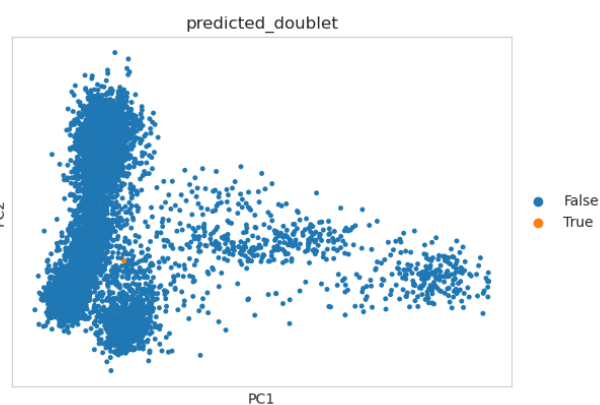
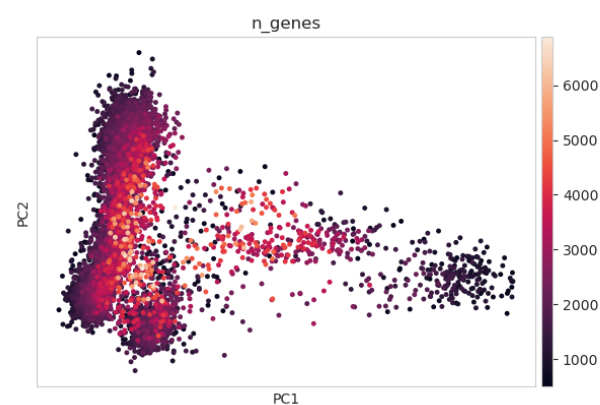
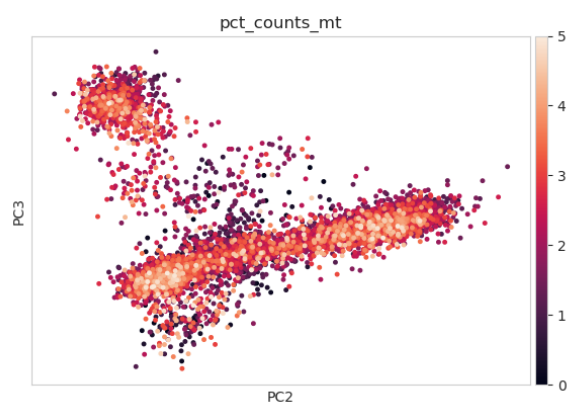
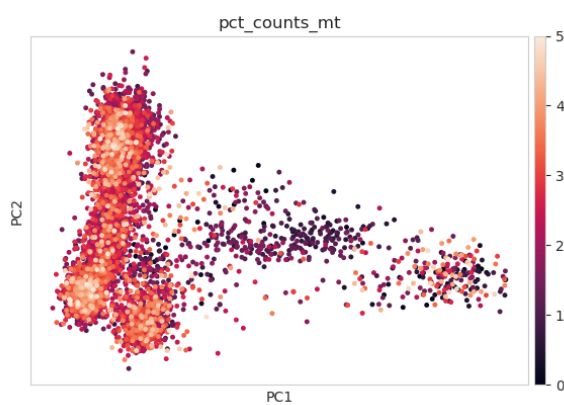
Практически нет ярко выделяющихся точек с высоким процентом митохондриальной РНК значит, подозрительные повреждённые клетки были удалены, и для дальнейшего анализа используется "очищенный" набор клеток хорошего качества ну в конце есть небольшой разброс)) скорее всего клетки, которые посчитались два раза или очень высоко экспр. (далее их уберём)

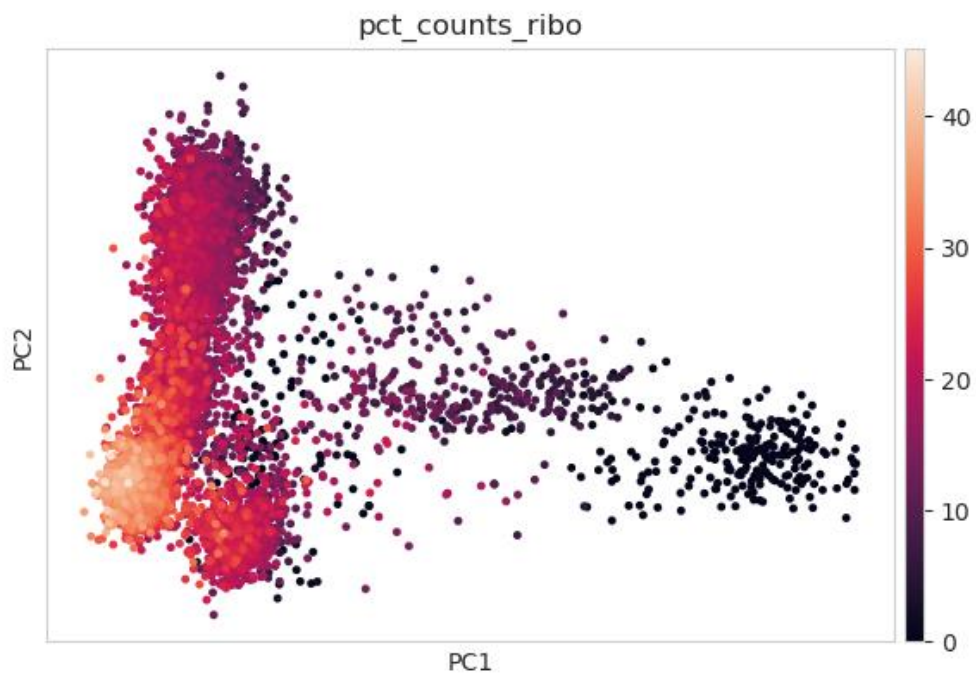


Клеток с экстремально высоким содержанием митохондриальной РНК не осталось

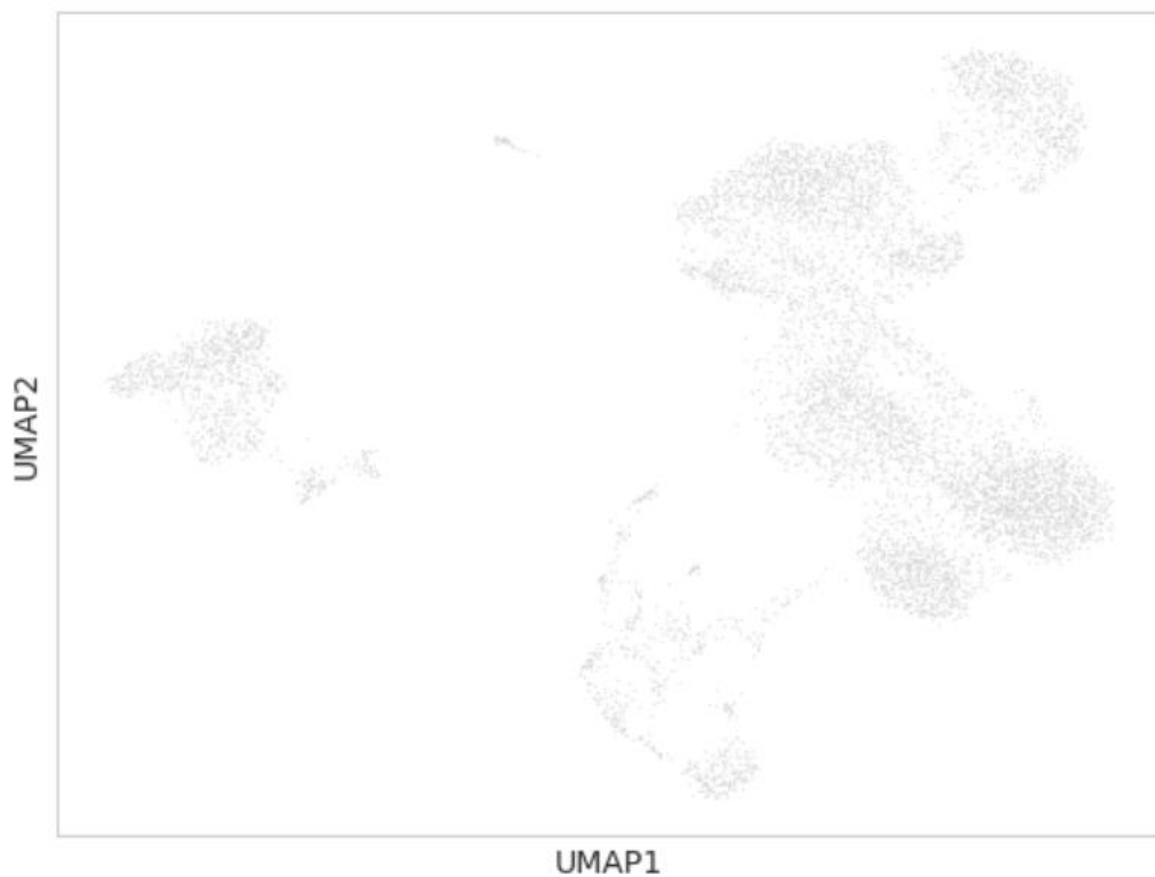


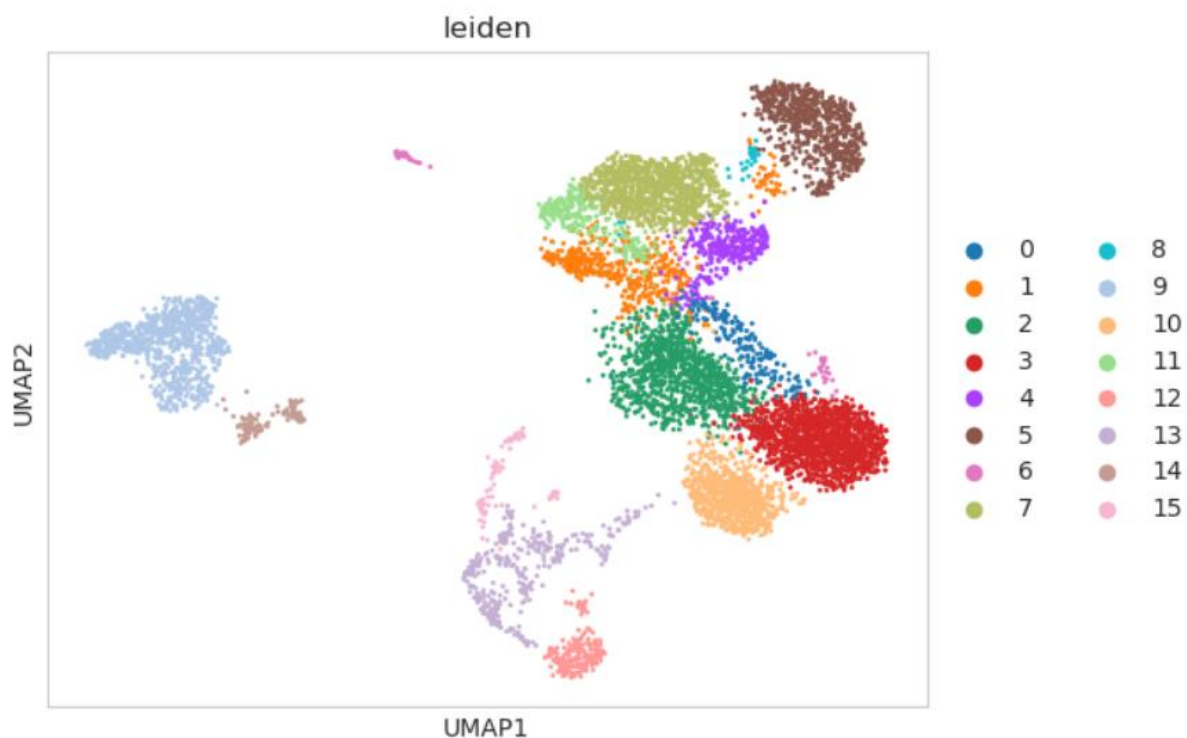
выбираем гены, которые сильно варьируют между клетками (высокая нормализованная дисперсия), а не просто сильно или слабо которые экспрессируются



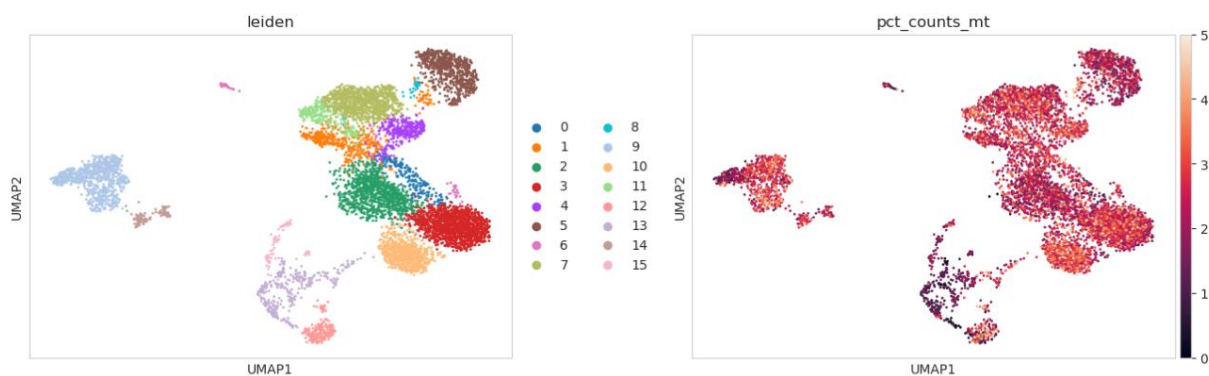


В PCA-пространстве не наблюдается отдельного скопления клеток с высокой митохондриальной или рибосомной долей, а доля предсказанных дублетов минимальна, что подтверждает хорошее качество данных)

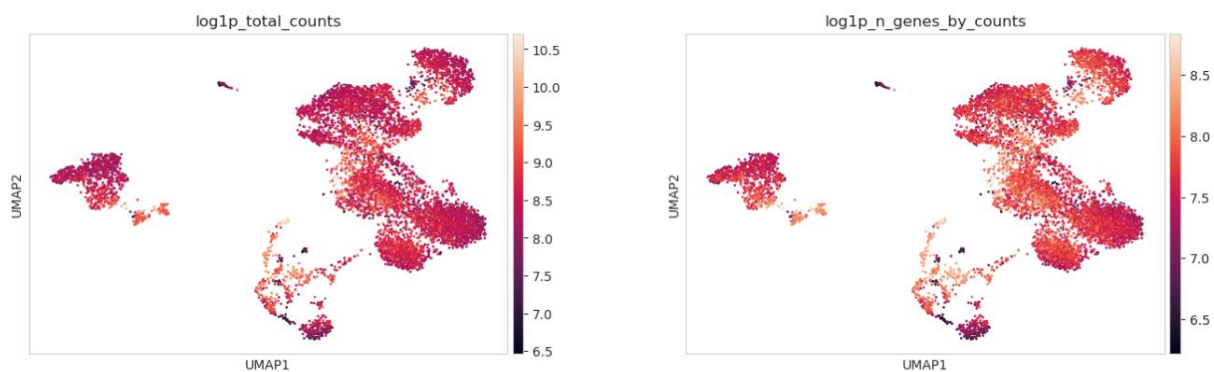




На UMAP-проекции алгоритм Leiden разделил клетки на 16 областей (кластеров) более менее компактных

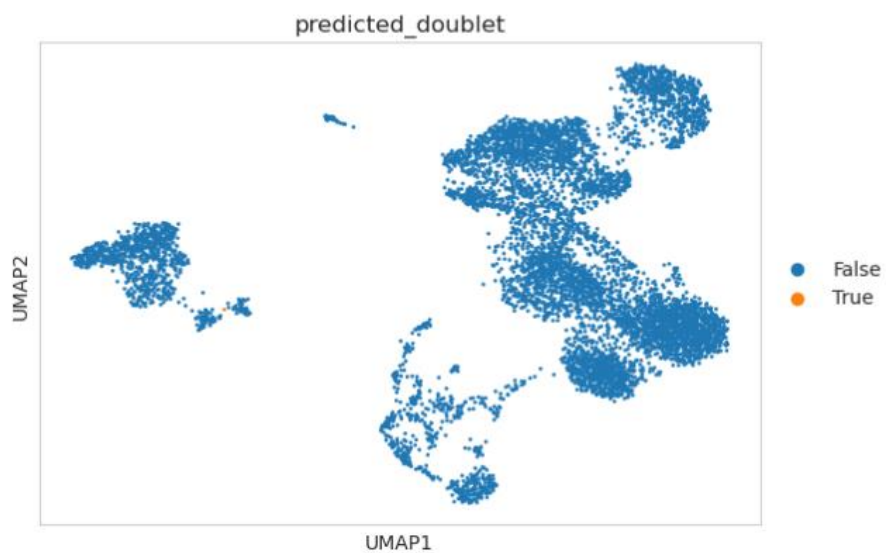


митохондриальная доля в целом низкая (0–5%) во всех кластерах

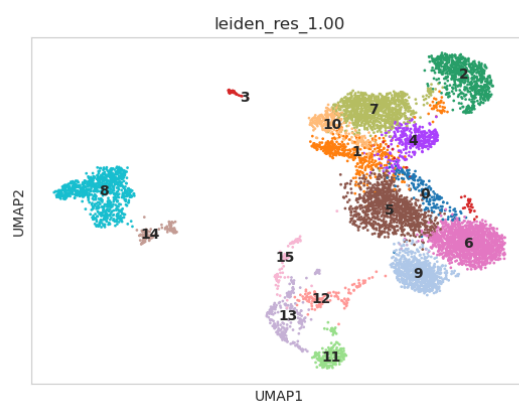
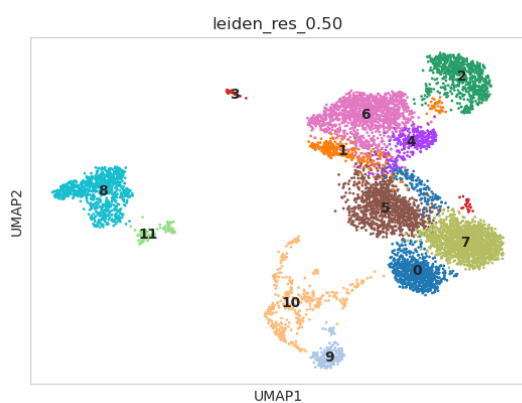
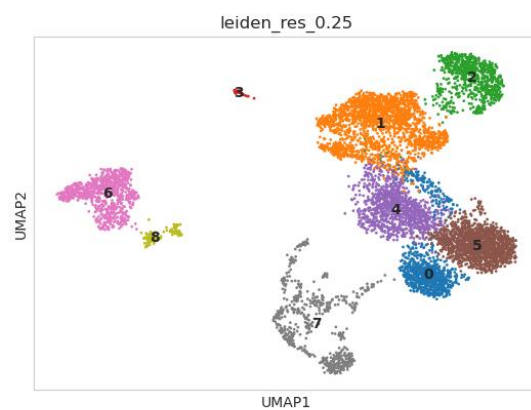
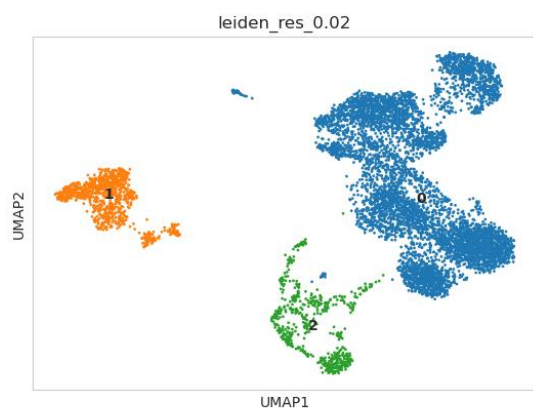


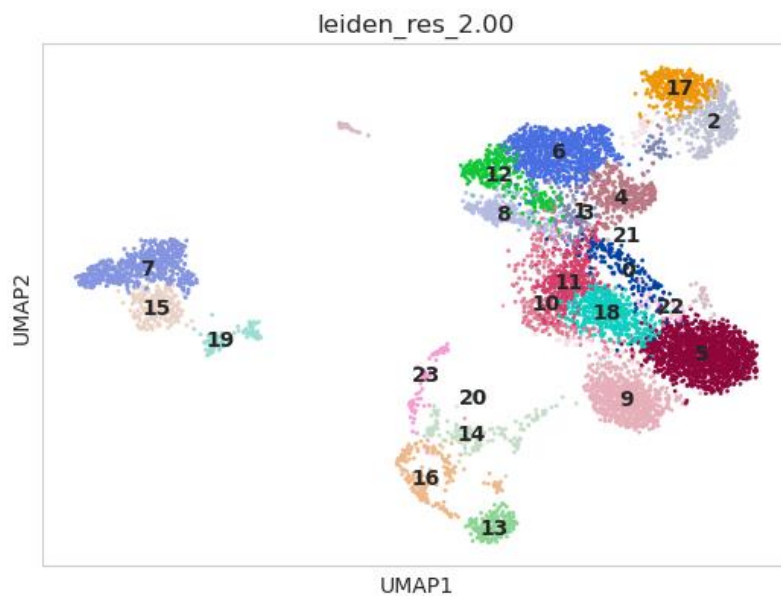
градиенты глубины (суммарных мрнк и числа генов) распределены плавно по UMAP, без локальных кусков с экстремально низкими или высокими значениями, кроме кластеров как мне кажется 13,15



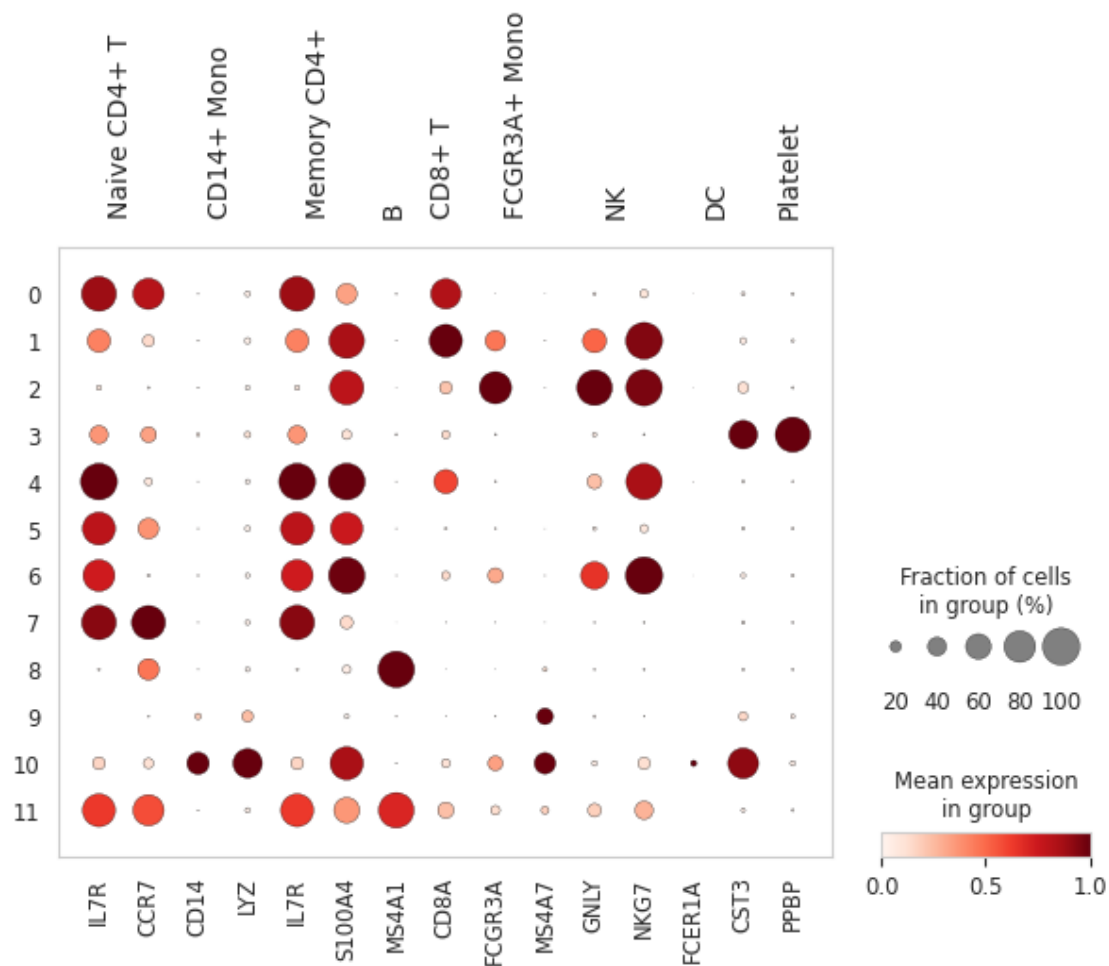


Дублетов не видно)



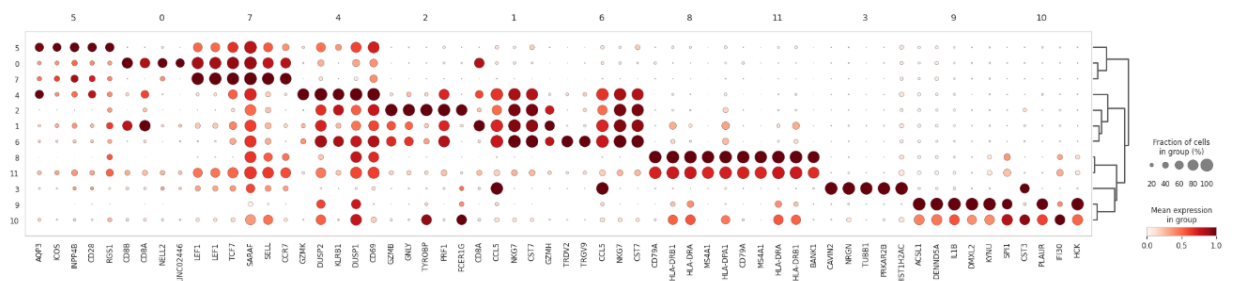
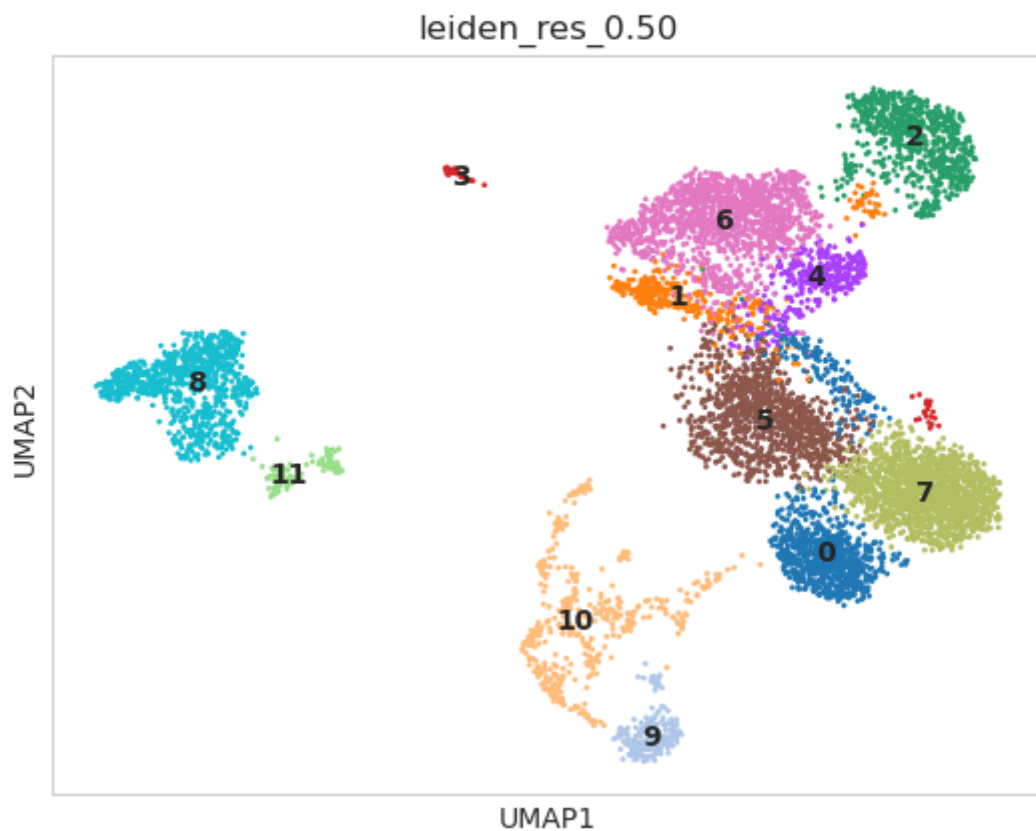


leiden\_res\_0.50, возьмём так как при нём как раз кластеры 13 и 15 один



Каждый маркер концентрируется в ожидаемых кластерах Naive/Memory CD4+ T в кластерах 0, 4–6, CD8+ T в кластерах и тд

```
sc.pl.umap(
    adata,
    color=["leiden_res_0.50"],
    legend_loc="on data",
)
```



показывает иерархию кластеров то есть какие кластеры похожи друг на друга к примеру можно понять какой вид клеток, может быть это и один, а может и разные подтипы

Далее берём кластер любой

Ищешь в нём top маркерные гены те, что экспрессируются сильнее всего в этом кластере и слабо в остальных (по результатам rank genes groups).

Рисуем эти гены на UMAP каждый на отдельной панели, окрашенный по уровню экспрессии.

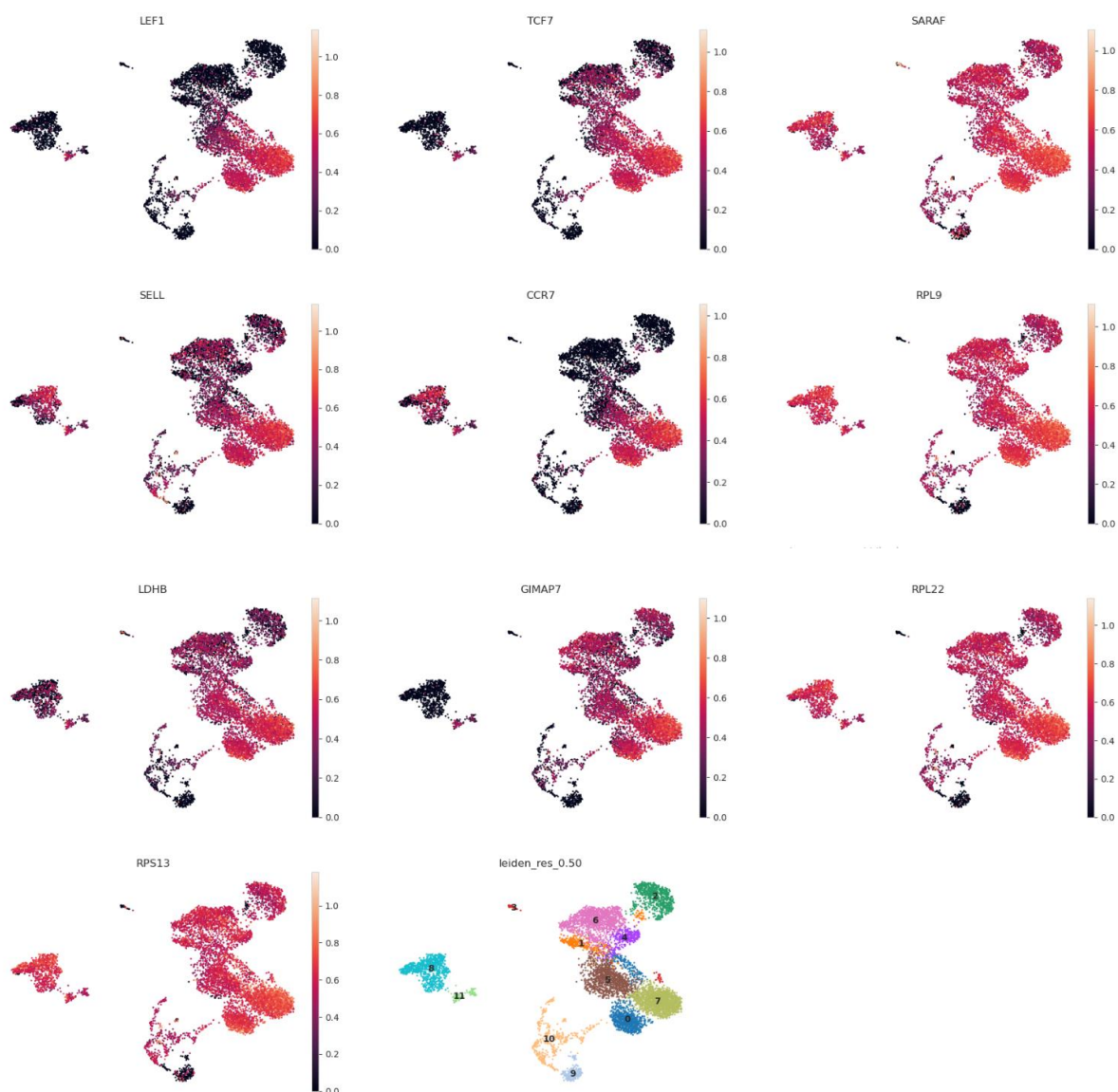
Смотришь, совпадают ли красные облака (высокая экспрессия этих генов) с регионом кластера который выбрали на UMAP

```
[98]: sc.get.rank_genes_groups_df(adata, group="7").head(7)
```

[98]:	names	scores	logfoldchanges	pvals	pvals_adj
0	LEF1	49.217827	2.263160	0.0	0.0
1	TCF7	45.231152	1.605779	0.0	0.0
2	SARAF	43.004902	0.584164	0.0	0.0
3	SELL	41.837601	1.352558	0.0	0.0
4	CCR7	41.464260	2.050642	0.0	0.0
5	RPL9	40.956863	0.524194	0.0	0.0
6	LDHB	40.685944	1.124914	0.0	0.0

К примеру

Взяли кластер 7



Можно кластеры 7,0,5 это один вид клеток, но разный подтип скорее всего