



«Анализ транскриптомных данных»

# Лекция #14. **Определение типов клеток**

**Серёжа Исаев**

аспирант **MedUni Vienna**

# Содержание курса

## 1. Bulk RNA-Seq:

- a. экспериментальные подходы,
- b. выравнивания и псевдовыравнивания,
- c. анализ дифференциальной экспрессии,
- d. функциональный анализ;

## 2. Single-cell RNA-Seq:

- a. экспериментальные подходы,
- b. отличия от процессинга bulk RNA-Seq,
- c. методы снижения размерности,
- d. кластера и траектории,**
- e. мультимодальные омики одиночных клеток.

# Что такое тип клетки?

В целом сейчас нет однозначного ответа на этот вопрос

Мы знаем, что есть какие-то состояния клетки — некоторые состояния уже не могут выйти в другое, кроме как в апоптоз (эритроцит не может стать незритроцитом)

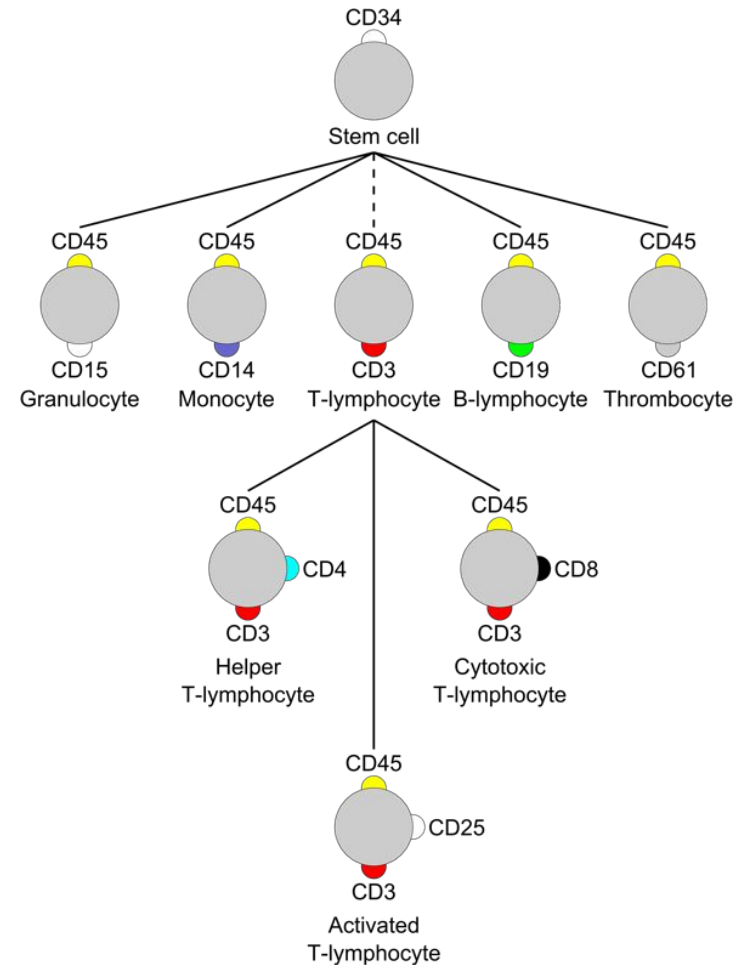
Бывают случаи, когда у нас есть более-менее стабильное состояние клетки (M1 и M2 макрофаги), однако при должном желании мы можем запустить превращение одного типа клетки в другой

В целом тип клетки — это некоторая транскрипционно (почти) гомогенная популяция (гетерогенность обуславливается разными клеточными состояниями) с определённой функцией в организме

# Тимы иммунных клеток

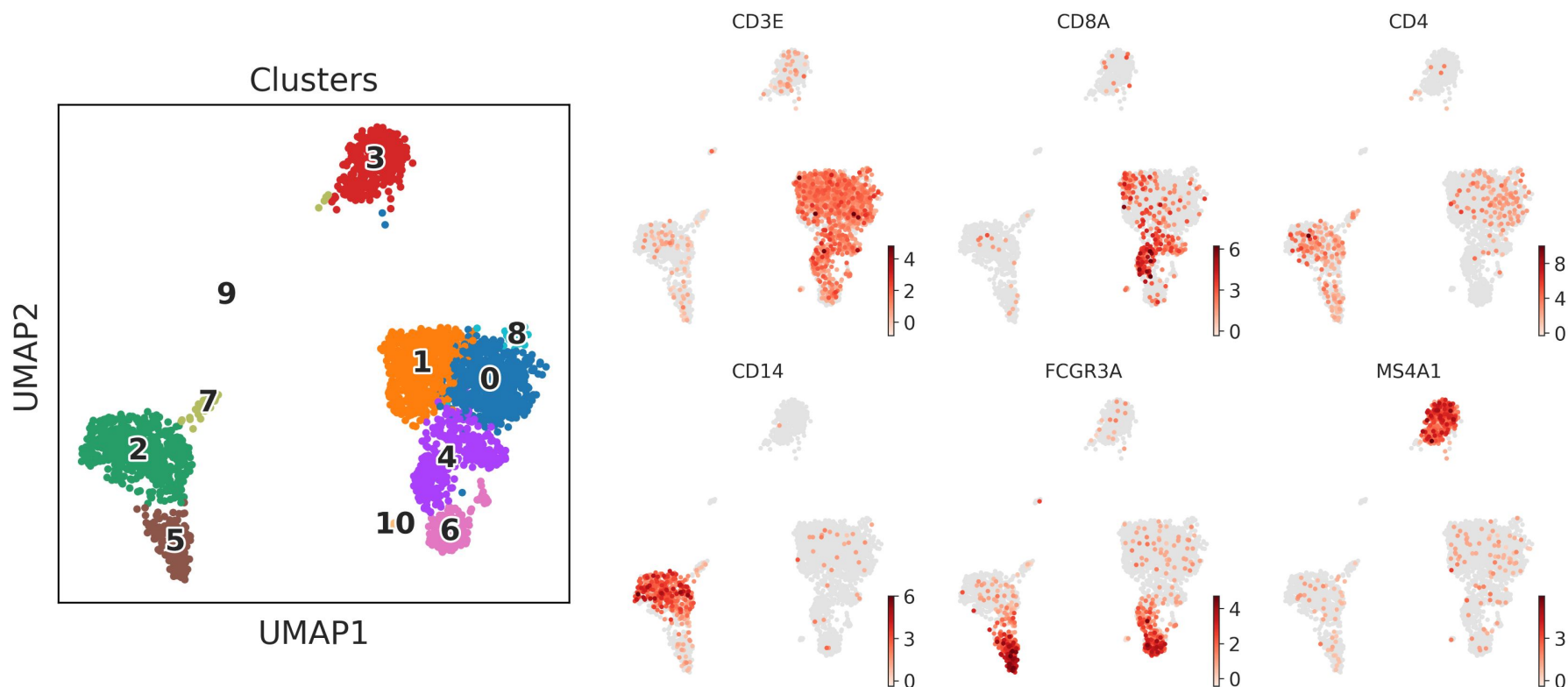
Для описания гетерогенности иммунных клеток уже не одно десятилетие используется набор поверхностных белков, представленных на клетках

Однако поверхностные белки != РНК в клетке, поэтому необходимо быть очень аккуратным в выводах



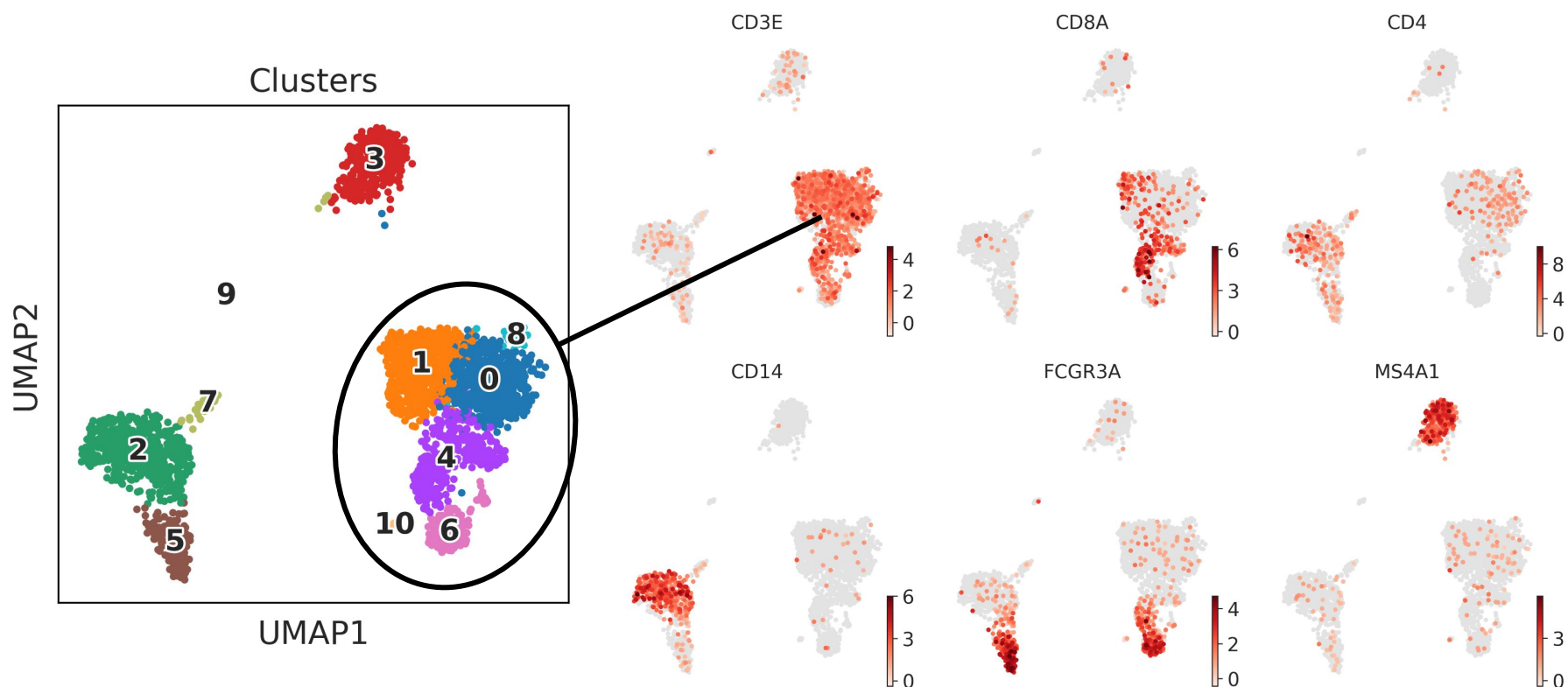
# Мануальные подходы к аннотации

Мы можем оценить, в каком кластере экспрессируются какие из известных нам маркерных генов



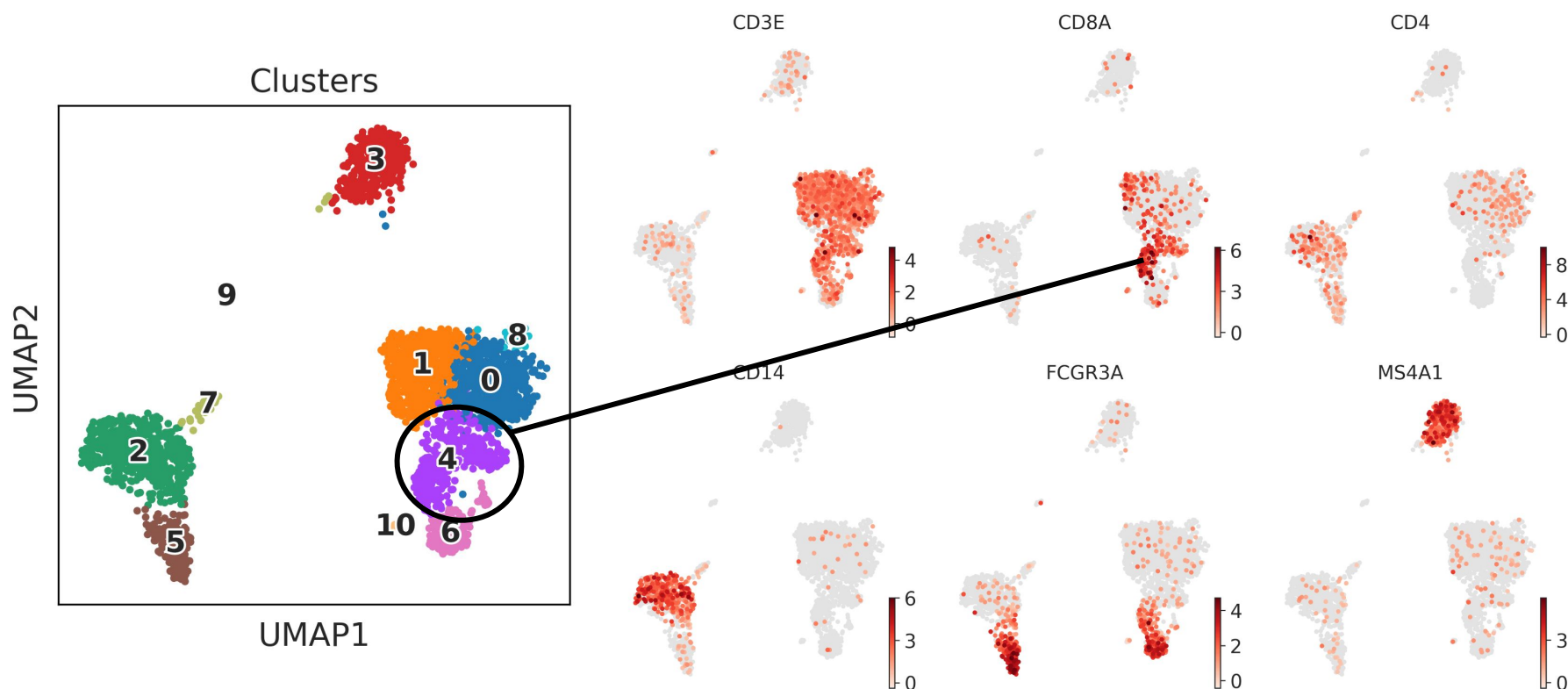
# Мануальные подходы к аннотации

Мы можем оценить, в каком кластере экспрессируются какие из известных нам маркерных генов



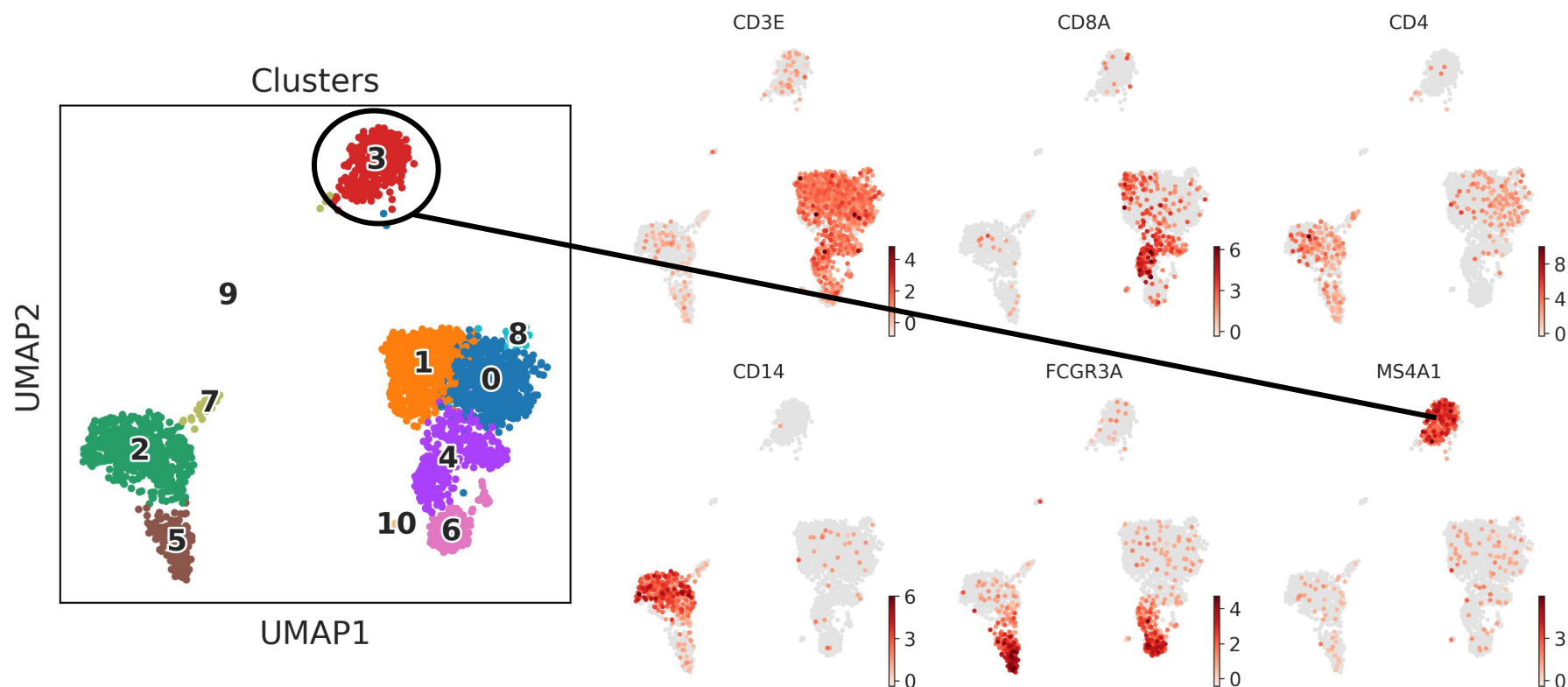
# Мануальные подходы к аннотации

Мы можем оценить, в каком кластере экспрессируются какие из известных нам маркерных генов



# Мануальные подходы к аннотации

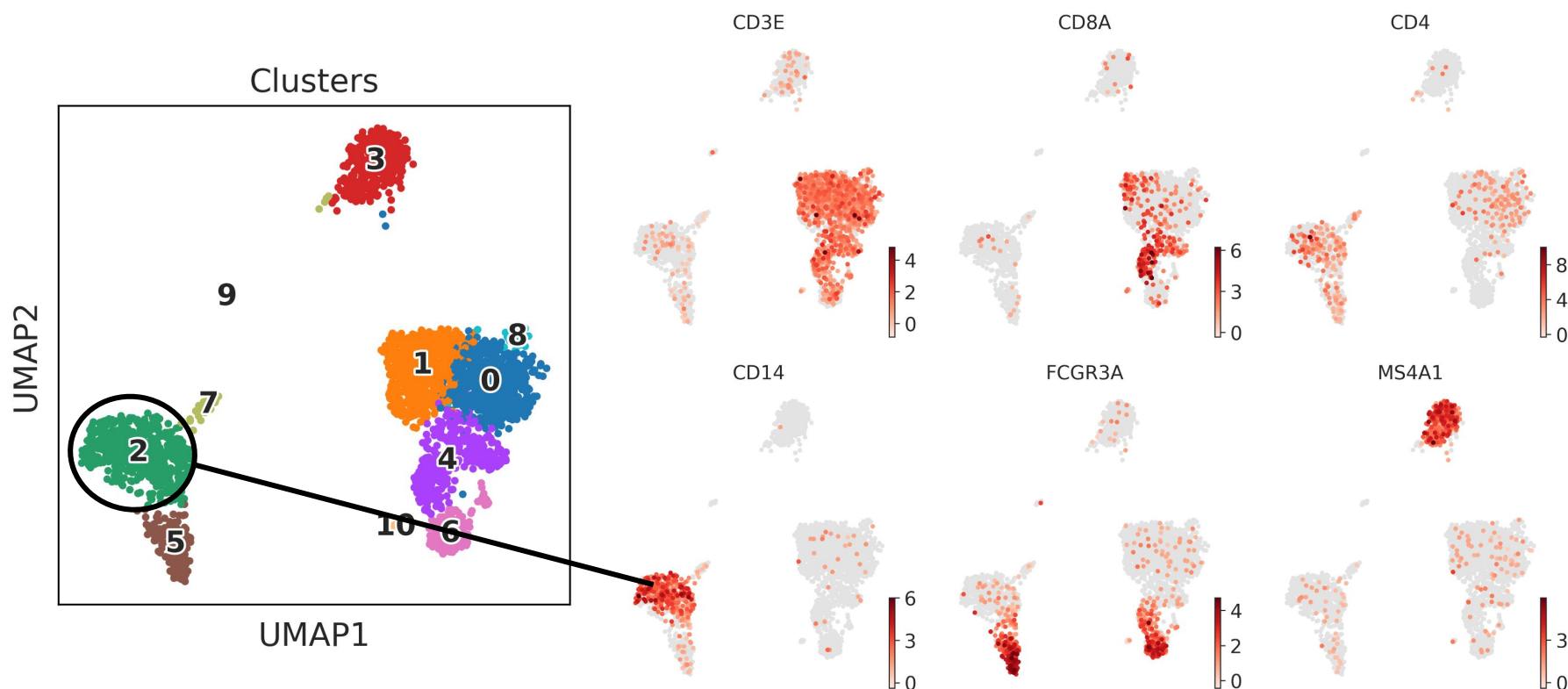
Мы можем оценить, в каком кластере экспрессируются какие из известных нам маркерных генов





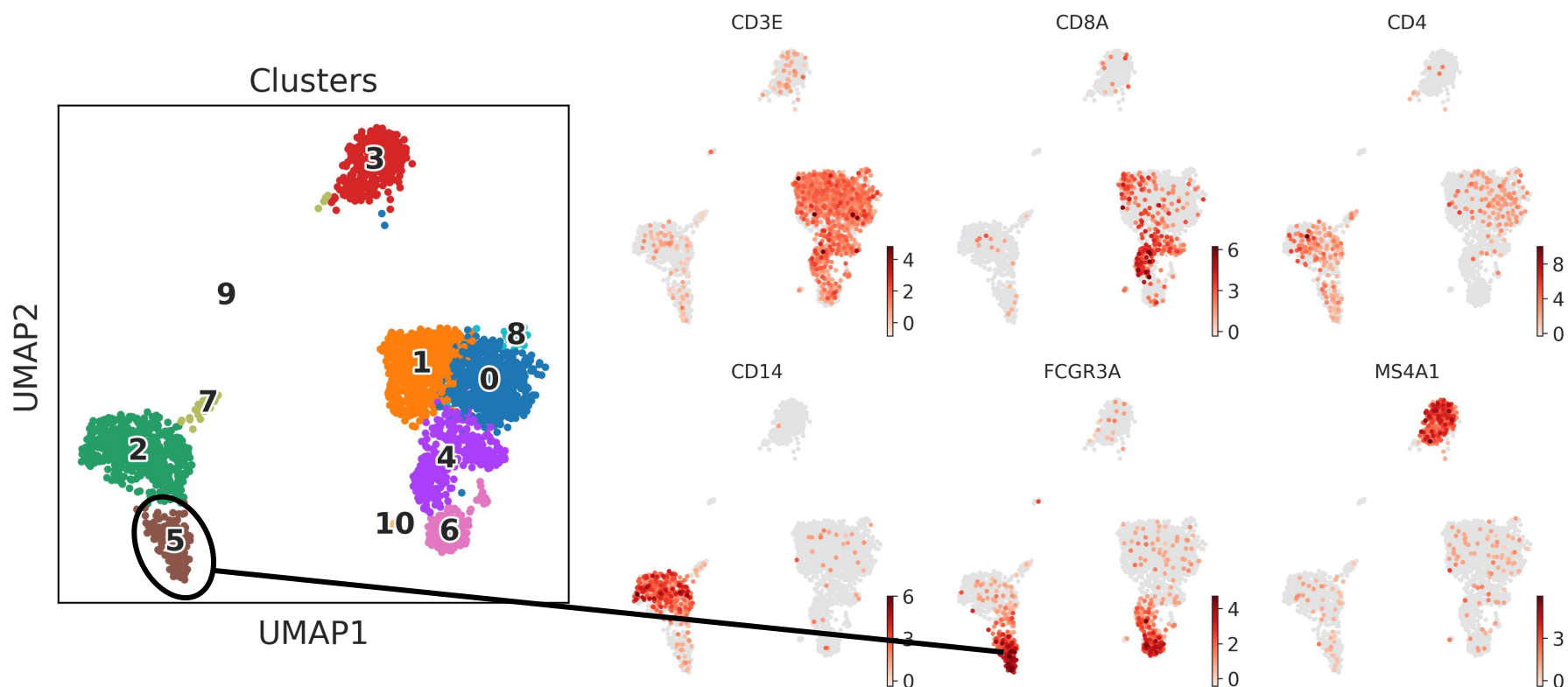
# Мануальные подходы к аннотации

Мы можем оценить, в каком кластере экспрессируются какие из известных нам маркерных генов



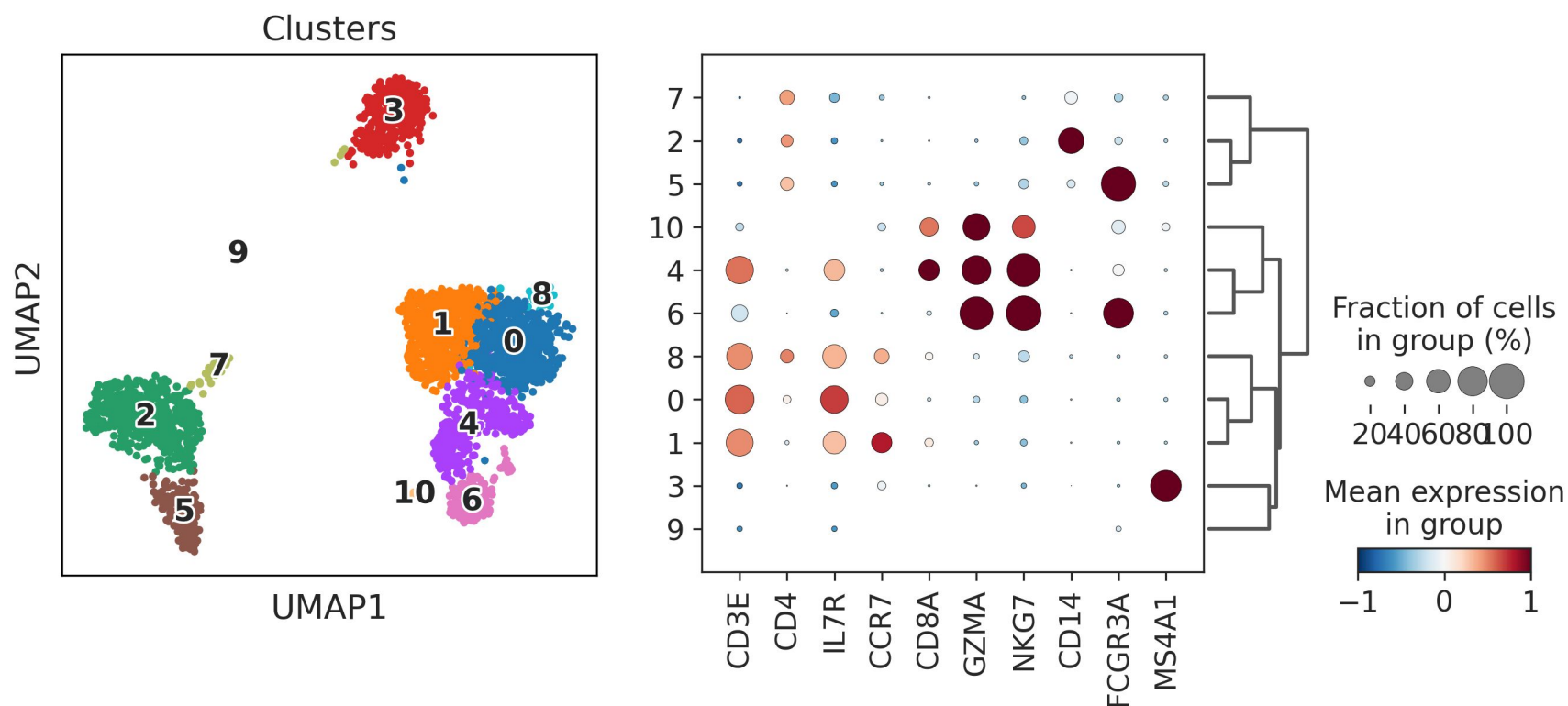
# Мануальные подходы к аннотации

Мы можем оценить, в каком кластере экспрессируются какие из известных нам маркерных генов



# Мануальные подходы к аннотации

Удобным и информативным в свете этого всего будет график, который называют отражает среднюю экспрессию в группе и % клеток с этой экспрессией



# Поиск маркеров

Когда закончились идеи насчёт маркерных генов, которые можно было бы оценить, можно зайти с другого конца — определить маркерные гены для каждого из кластеров *de novo*, а потом предугадать, что это был за кластер, опираясь на список генов

Есть два концептуальных подхода к поиску маркеров:

1. поиск дифференциально экспрессированных генов,
2. поиск генов, экспрессия которых является лучшим предиктором для отличия этого типа клетки от другого (AUC ROC)

# AUC ROC

1 = целевой кластер  
0 = другой кластер

	group	expr
0	0	0.000000
1	0	0.000000
2	0	0.000000
3	1	0.000000
4	0	0.000000
...	...	...
2695	1	0.819593
2696	0	0.000000
2697	0	0.000000
2698	0	0.000000
2699	0	0.000000



Сортированные по  
возрастанию экспрессии

	group	expr
0	0	0.000000
1723	0	0.000000
1724	0	0.000000
1725	0	0.000000
1726	0	0.000000
...	...	...
443	1	2.078248
2598	1	2.121581
98	1	2.218295
2506	1	2.251932
1326	1	2.505317

# ROC

Затём вы идёте по возрастающей экспрессии и ставите порог классификатора по этому значению экспрессии

Выше? Это клетка из вашего кластера

Ниже? Это клетка не из вашего кластера



Какое количество True Positive (TP)? False negative (FN)?  $TPR = TP / (TP + FN)$

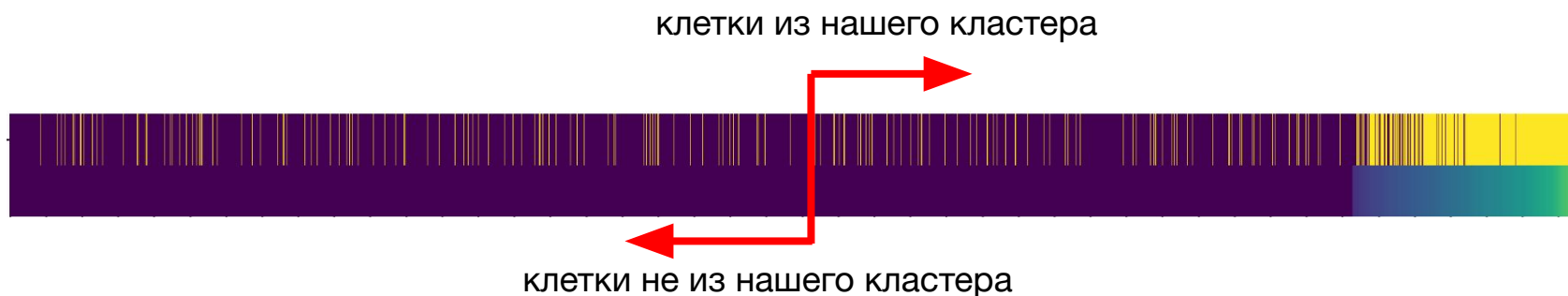
Какое количество False Positive (FP)? True Negative (TN)?  $FPR = FP / (FP + TN)$

# ROC

Затём вы идёте по возрастающей экспрессии и ставите порог классификатора по этому значению экспрессии

Выше? Это клетка из вашего кластера

Ниже? Это клетка не из вашего кластера



Какое количество True Positive (TP)? False negative (FN)?  $TPR = TP / (TP + FN)$

Какое количество False Positive (FP)? True Negative (TN)?  $FPR = FP / (FP + TN)$

# ROC

Затём вы идёте по возрастающей экспрессии и ставите порог классификатора по этому значению экспрессии

Выше? Это клетка из вашего кластера

Ниже? Это клетка не из вашего кластера



Какое количество True Positive (TP)? False negative (FN)?  $TPR = TP / (TP + FN)$

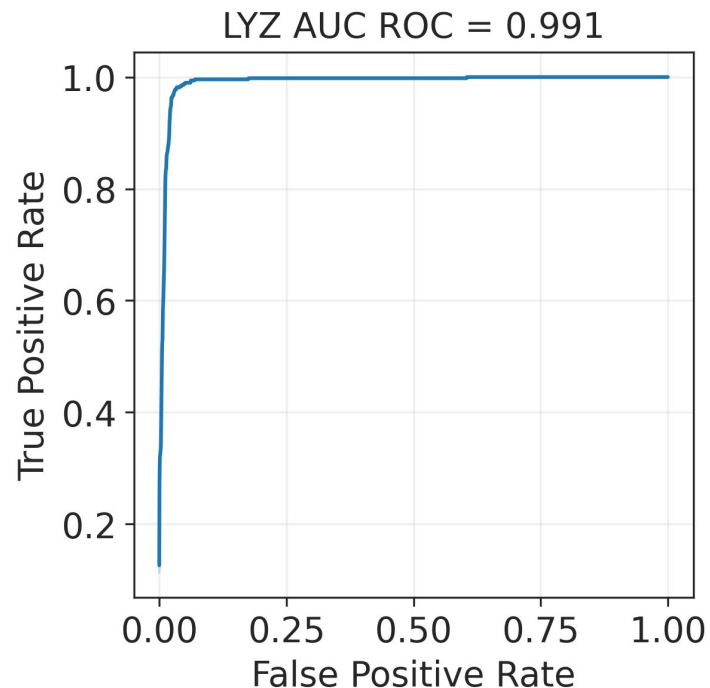
Какое количество False Positive (FP)? True Negative (TN)?  $FPR = FP / (FP + TN)$



# AUC ROC

А теперь отрисуем все значения TPR и FPR на одном графике в виде кривой и найдём площадь под этой кривой

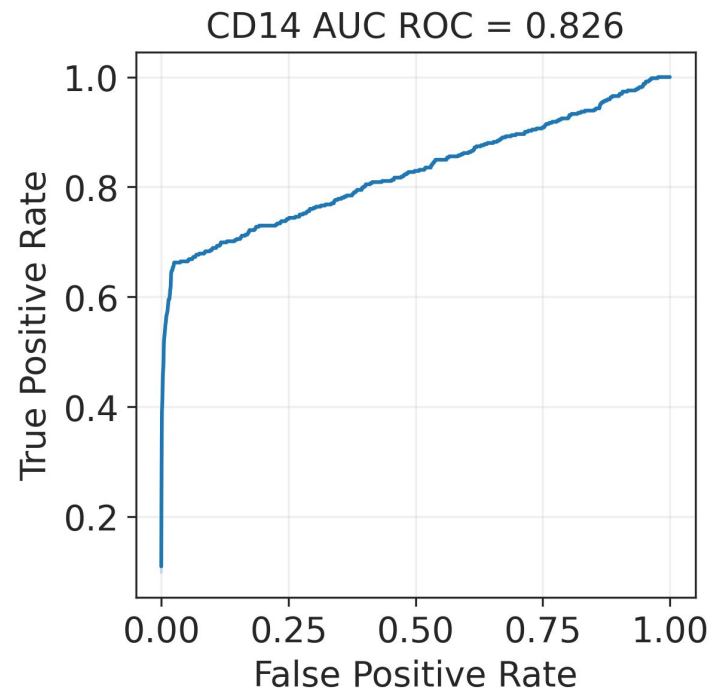
Мы можем повторить это для каждого гена и выбрать в итоге гены с наибольшим значением AUC — это и будут маркерные гены



# AUC ROC

А теперь отрисуем все значения TPR и FPR на одном графике в виде кривой и найдём площадь под этой кривой

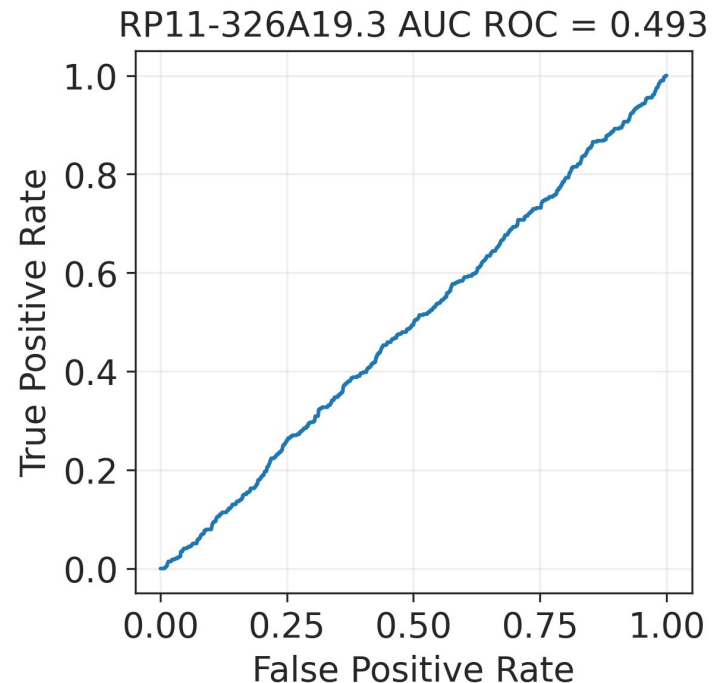
Мы можем повторить это для каждого гена и выбрать в итоге гены с наибольшим значением AUC — это и будут маркерные гены



# AUC ROC

А теперь отрисуем все значения TPR и FPR на одном графике в виде кривой и найдём площадь под этой кривой

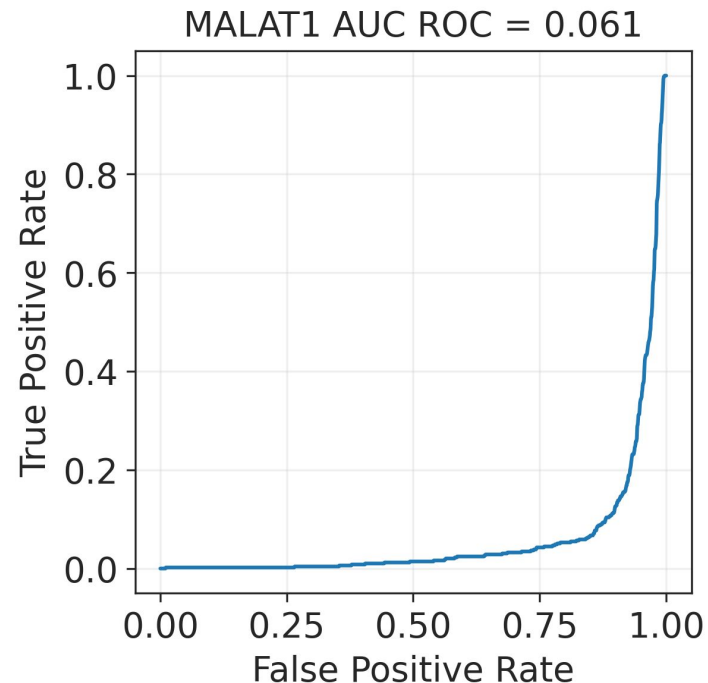
Мы можем повторить это для каждого гена и выбрать в итоге гены с наибольшим значением AUC — это и будут маркерные гены



# AUC ROC

А теперь отрисуем все значения TPR и FPR на одном графике в виде кривой и найдём площадь под этой кривой

Мы можем повторить это для каждого гена и выбрать в итоге гены с наибольшим значением AUC — это и будут маркерные гены



# Дифференциальная экспрессия

Также маркерные гены можно искать при помощи дифференциальной экспрессии, однако подходы здесь обычно значительно проще, чем в bulk RNA-Seq из-за очень больших размеров выборок

Можно использовать:

1. T-test с неравными дисперсиями (с натяжкой, но даст представление о маркерах) на  $\log_{10}CP10k$  — очень быстрый
2. **Тест Манна-Уитни на  $\log_{10}CP10k$**  — медленнее, но подходит лучше. В основном используют его
3. NB-модели на каунтах — вычислительно сложные, требуют много ресурсов, не очень ясно, зачем

# Проблема множества батчей

Часто мы имеем дело с датасетами, которые состоят из нескольких батчей

Как в таком случае считать дифференциальную экспрессию?

1. На каждом батче индивидуально (например, при помощи Манна-Уитни), а потом усредняем  $\log FC$  и аккумулируем  $p$ -value
2. Делаем псевдо-балки из каждого батча (если их больше, чем три-четыре), после чего используем edgeR / DESeq2 для поиска дифференциальной экспрессии между батчами

# Тест Манна-Уитни и AUC

Тест Манна-Уитни

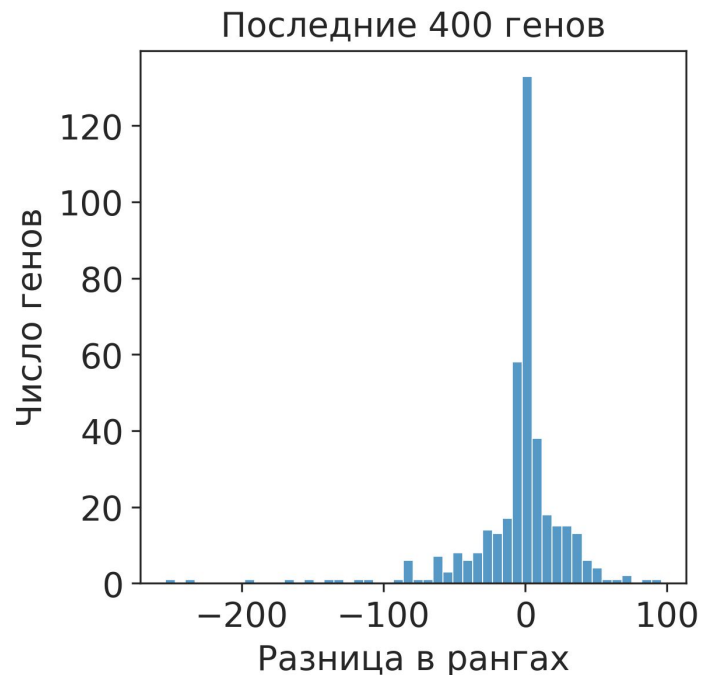
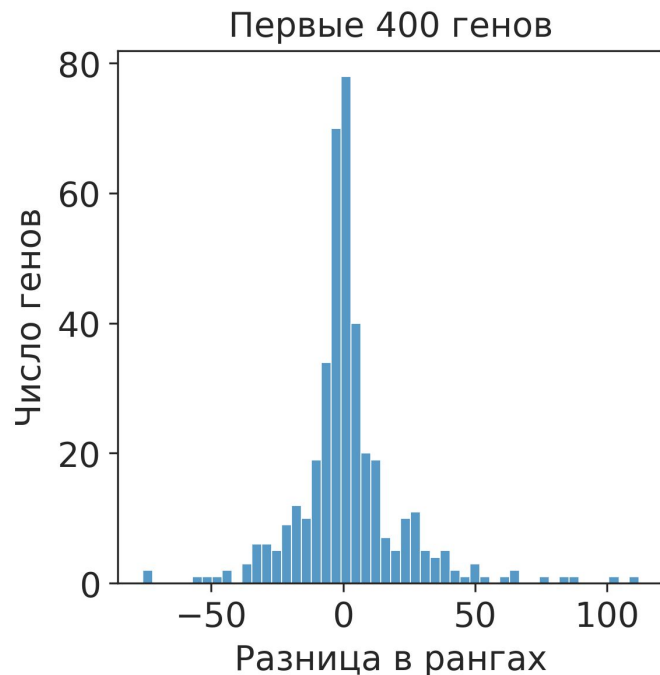
	names	scores	logfoldchanges	pvals	pvals_adj
0	LYZ	34.076332	5.554640	1.653918e-254	5.414596e-250
1	S100A9	34.022038	6.539301	1.052143e-253	1.722253e-249
2	S100A8	33.251110	6.832742	1.967234e-242	2.146777e-238
3	TYROBP	32.434139	4.540583	9.065962e-231	7.420036e-227
4	FTL	31.846174	3.446953	1.487022e-222	9.736426e-219
...	...	...	...	...	...
32733	RPSA	-25.772381	-1.679404	1.809493e-146	1.974639e-143
32734	RPL23A	-25.978748	-1.195210	8.610019e-149	1.006696e-145
32735	RPS27	-26.104221	-1.198215	3.264942e-150	4.453654e-147
32736	RPS27A	-27.741587	-1.906760	2.201332e-169	3.431771e-166
32737	MALAT1	-30.530386	-1.625752	1.029974e-204	3.065389e-201

AUC

	AUC
LYZ	0.990816
S100A9	0.990045
S100A8	0.978425
TYROBP	0.966149
FTL	0.958402
...	...
RPSA	0.128895
RPL23A	0.126083
RPS27	0.124240
RPS27A	0.100790
MALAT1	0.060538

В пределе статистика Манна-Уитни сходится к AUC, поэтому в принципе можно пользоваться только одним из этих тестов (Манн-Уитни даёт к тому же значение p-value, однако AUC удобнее интерпретировать)

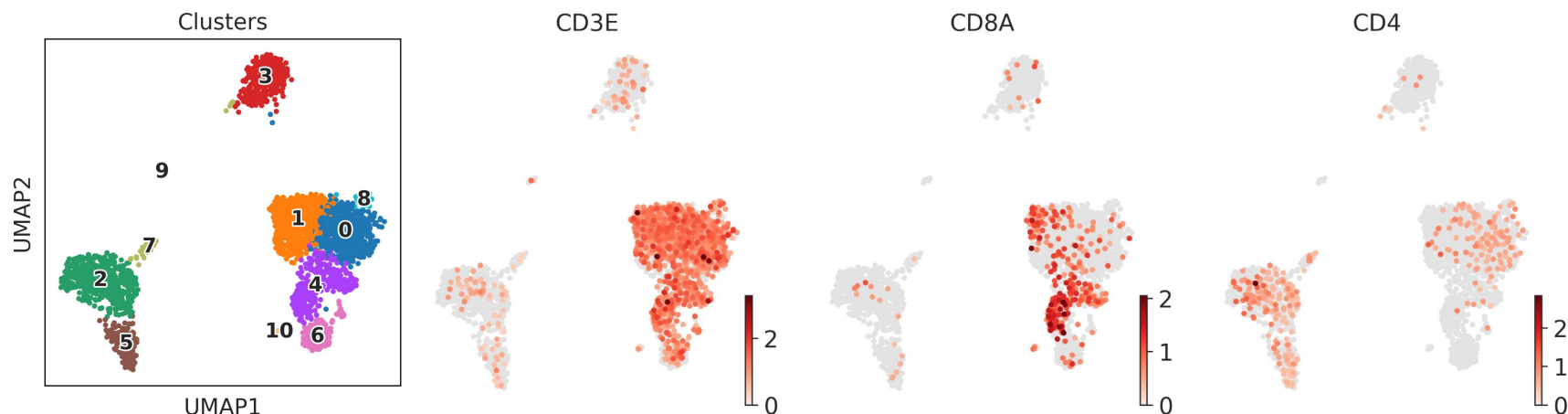
# Тест Манна-Уитни и AUC



В пределе статистика Манна-Уитни сходится к AUC, поэтому в принципе можно пользоваться только одним из этих тестов (Манн-Уитни даёт к тому же значение  $p$ -value, однако AUC удобнее интерпретировать)



# Проблемы с маркерными генами



Кластер 4 — это определённо CD8<sup>+</sup> Т-клетки (скорее всего, памяти). Соответственно, дифференциальная экспрессия должна выдать нам маркеры этого типа клеток

# Проблемы с маркерными генами

	names	scores	logfoldchanges	pvals	pvals_adj		names	scores	logfoldchanges	pvals	pvals_adj
0	CCL5	26.200100	4.760021	2.650470e-151	8.677110e-147	10	CD3D	15.700964	1.724816	1.489616e-55	4.433367e-52
1	NKG7	23.580784	4.148077	6.069330e-123	9.934887e-119	11	PTPRCAP	13.961631	1.369620	2.672486e-44	6.730142e-41
2	B2M	20.350811	0.793257	4.566622e-92	4.983403e-88	12	LYAR	13.860597	2.838753	1.097574e-43	2.566599e-40
3	GZMA	19.658840	3.405880	4.857022e-86	3.975230e-82	13	GZMM	13.172859	2.242082	1.257444e-39	2.744414e-36
4	CST7	19.377089	3.492904	1.204750e-83	7.888220e-80	14	CD8A	12.877013	3.518669	6.063743e-38	1.240718e-34
5	IL32	19.057325	2.118002	5.712215e-81	3.116775e-77	15	HCST	12.408269	1.327444	2.356944e-35	4.538920e-32
6	CTSW	18.355267	2.710095	2.996152e-75	1.401258e-71	16	MALAT1	12.320384	0.677299	7.035862e-35	1.279667e-31
7	HLA-C	17.823990	0.956840	4.603284e-71	1.883779e-67	17	KLRG1	12.160748	3.459849	5.029613e-34	8.666288e-31
8	HLA-A	16.113411	0.911328	2.053779e-58	7.470736e-55	18	HLA-B	11.554396	0.636137	7.013920e-31	1.148108e-27
9	GZMK	16.089947	4.630021	3.000959e-58	9.824540e-55	19	PRF1	10.618041	2.075951	2.456628e-26	3.351046e-23

Однако ген CD3D на 10-ом месте, CD8A — на 14-ом, а гена CD3E вообще нет. Почему? Потому что у нас в датасете есть ещё Т-клетки, в том числе и CD8 Т-клетки

=> **ВСЕГДА** надо представлять, какие типы клеток ожидаются

# Автоматическое определение типов клеток

Методы можно разделить на несколько групп:

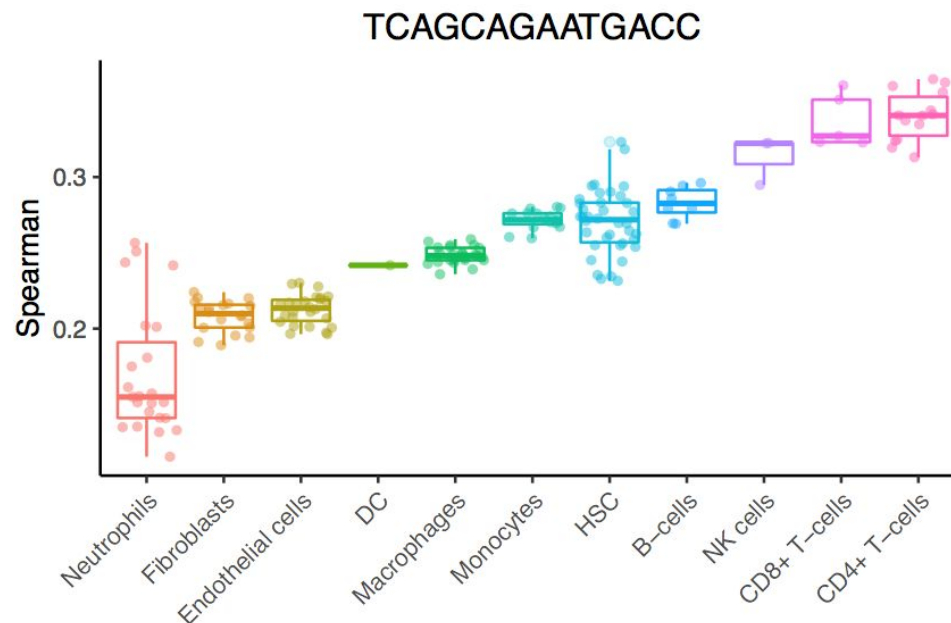
1. методы, определяющие максимально похожие bulk-датасеты с клеточными линиями (SingleR),
2. методы, использующие маркерные гены,
3. label transferring

# SingleR

Помогает автоматически предсказать, что за типы клеток перед нами, если совсем нет идей

Много референсных датасетов, но недостаточно для глубокого типирования

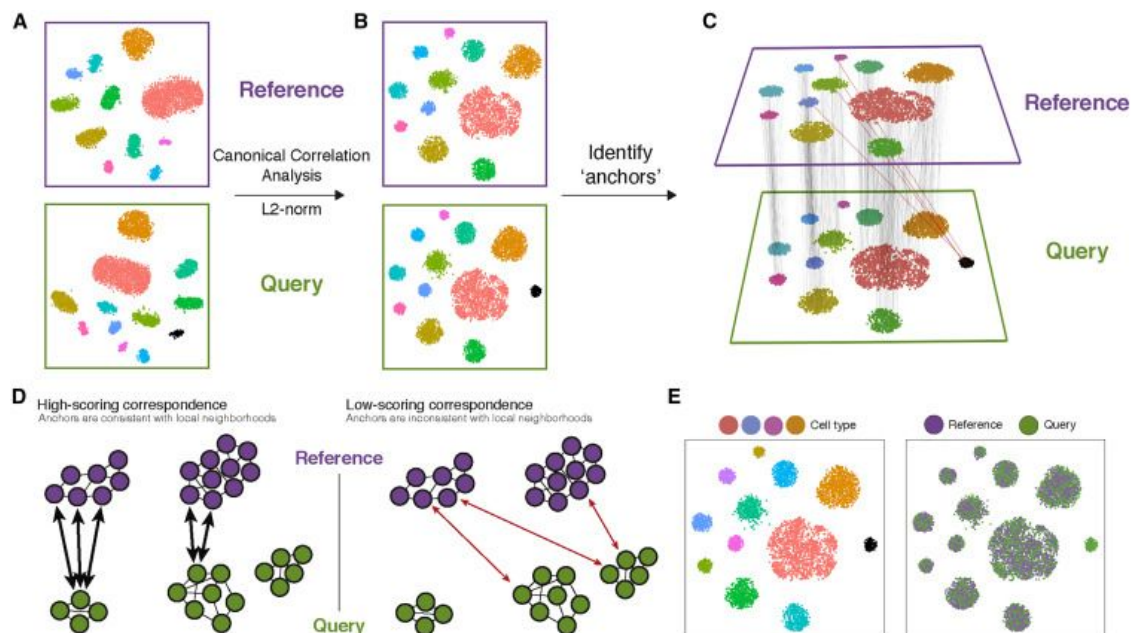
Предположения о типе клеток основывается на корреляции с известными датасетами (bulk!)



# Label transfer

Принцип основывается на том, что есть уже достаточно большое количество хорошо проаннотированных датасетов

Принцип достаточно прост: “проинтегрируем” (в первом приближении) датасет с референсным и по окружению неизвестных клеток поймём, какого они типа

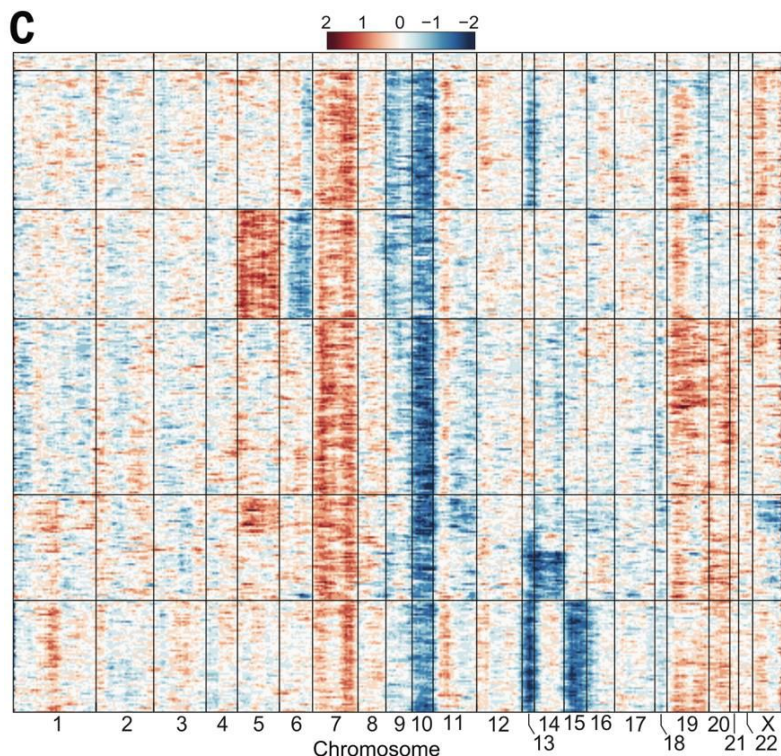


# Label transfer

Существует достаточно много инструментов для того, чтобы автоматически определять типы клеток по референсу:

1. conos (~ conos / pagoda2 ecosystem),
2. scANVI (~ scVI ecosystem),
3. Azimuth / Seurat (~ Seurat ecosystem),
4. Symphony (~ Seurat / Harmony ecosystem, не адаптирован для Python)

# Определение раковых клеток: InferCNV



Часто бывает сложно отделить нормальные клетки от опухолевых

Мы знаем, что в нормальных клетках практически всегда происходят серьёзные хромосомные перестройки — можно воспользоваться этим и идентифицировать раковые кластеры благодаря им