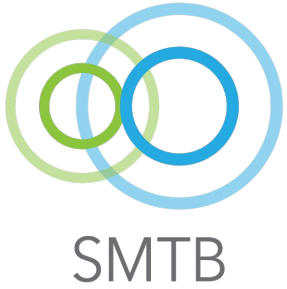


Предсказание представленности поверхностных белков по данным scRNA-Seq

Васильев Д. А.*, Мурашов И. Г. *, Панова Ю. С.*, Исаев С. В.

* — авторы внесли одинаковый вклад

scRNA-Seq — схема эксперимента

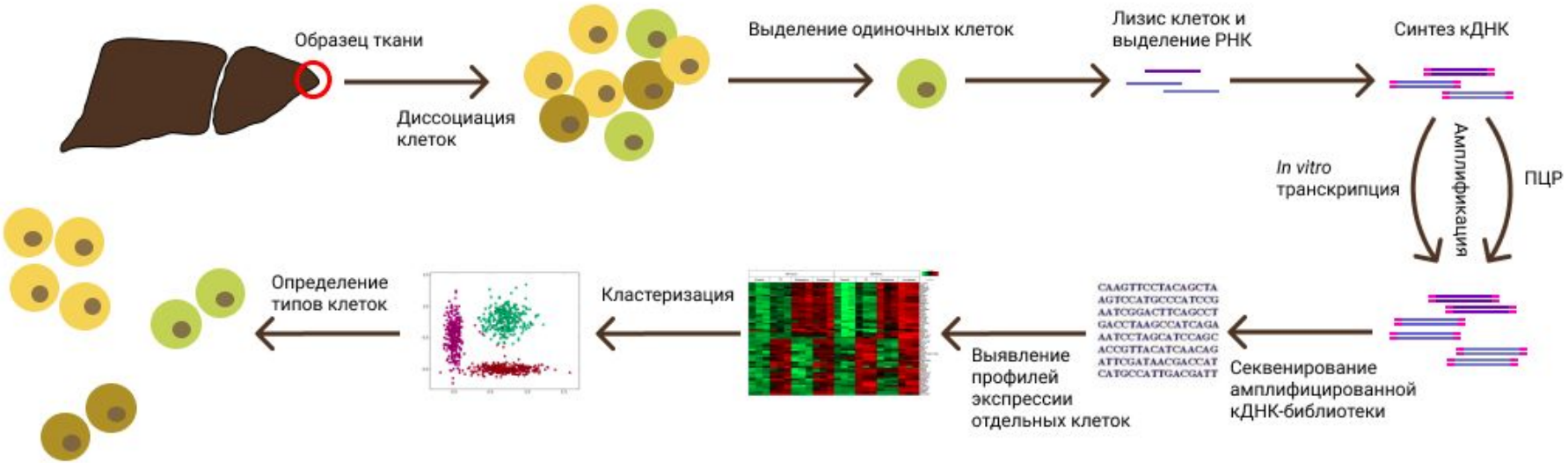
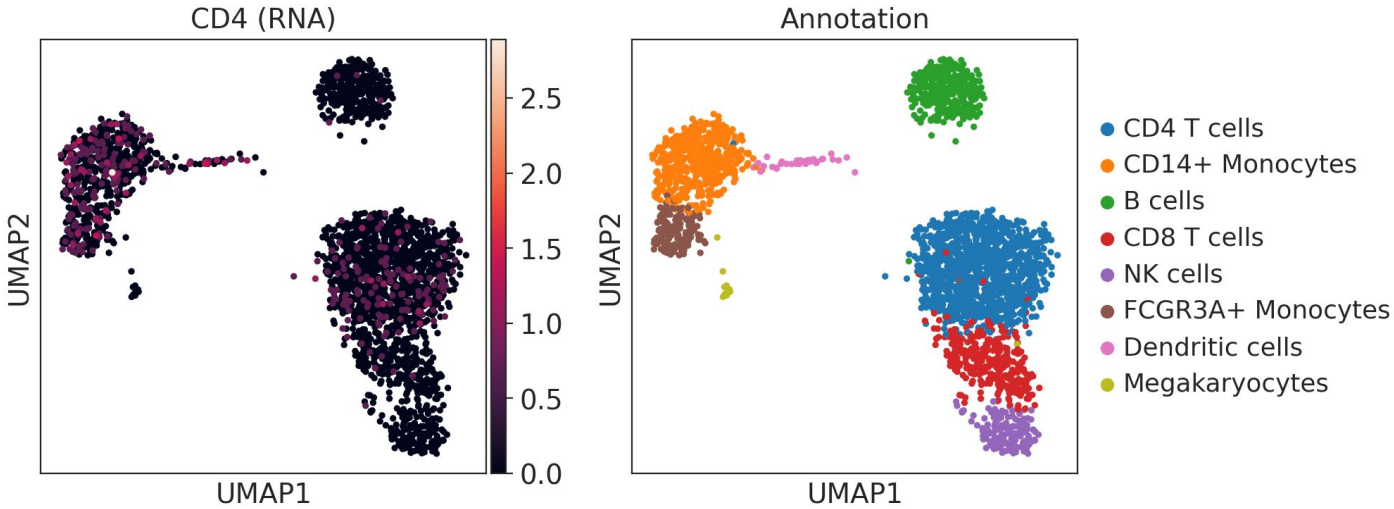


Иллюстрация из Википедии

Проблема

Не всегда поверхностные маркеры клеток можно детектировать на уровне экспрессии РНК при помощи scRNA-Seq.



Известный пример клеточного типирования открытого датасета PBMC 3k — несмотря на низкую экспрессию гена CD4, большой кластер клеток обозначен как CD4+ Т-лимфоциты..

CITE-Seq — схема эксперимента

CITE-Seq позволяет оценить помимо **уровней экспрессии РНК** каждой конкретной клетки также и **наличие или отсутствие поверхностных белков** на тех же клетках.

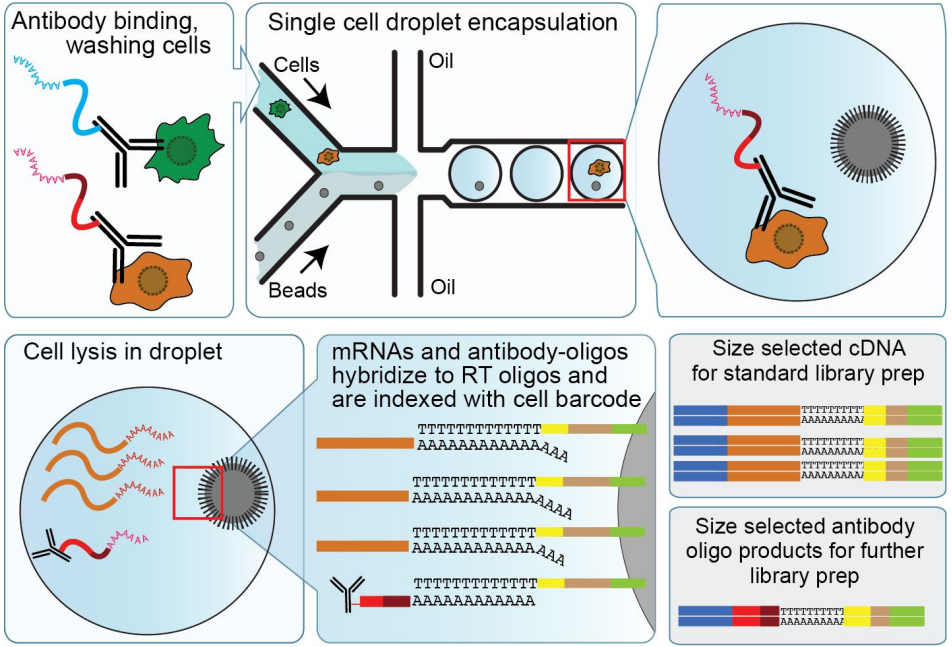


Иллюстрация с сайта cite-seq.com

SMTB

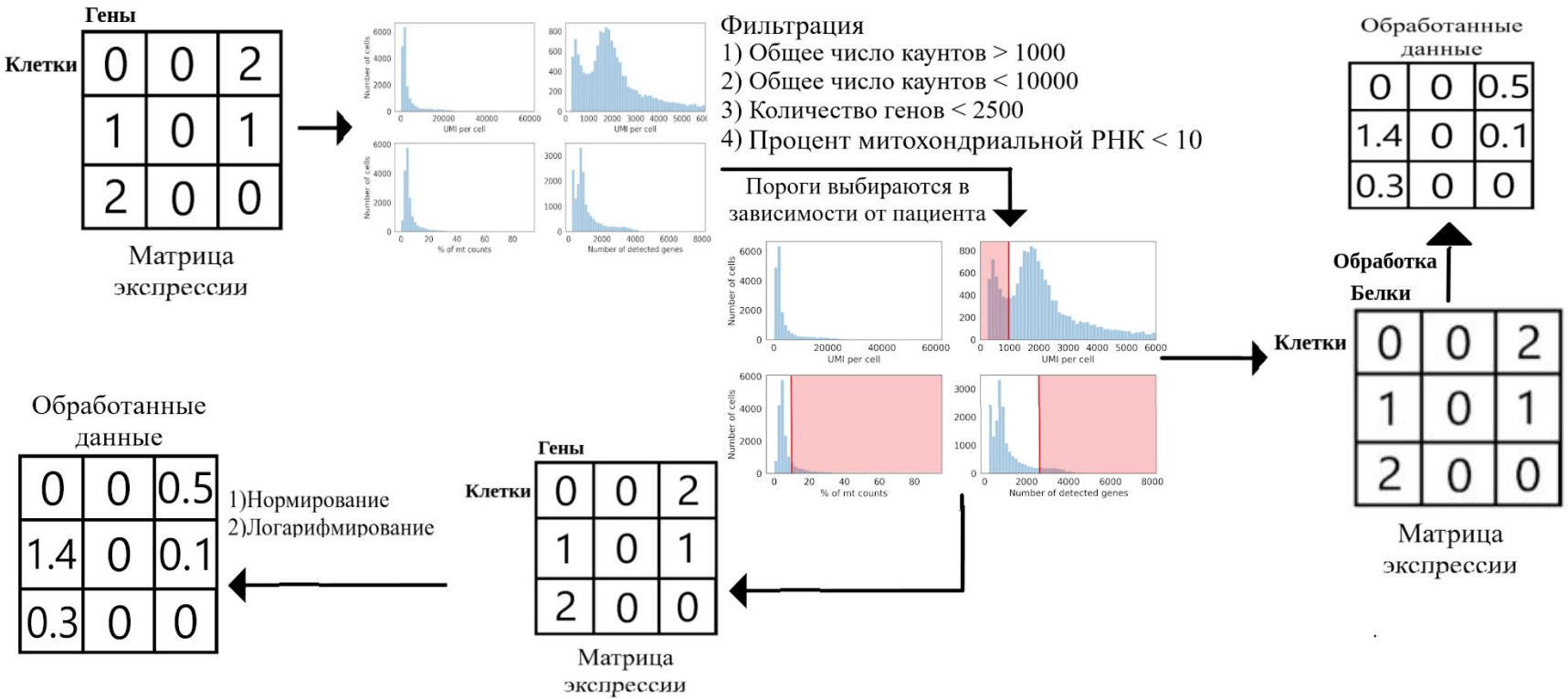
Задачи

Определить связь между уровнями экспрессии РНК и белка, а также попробовать создать предсказательную модель экспрессии белка по данным CITE-Seq.

Материалы и методы

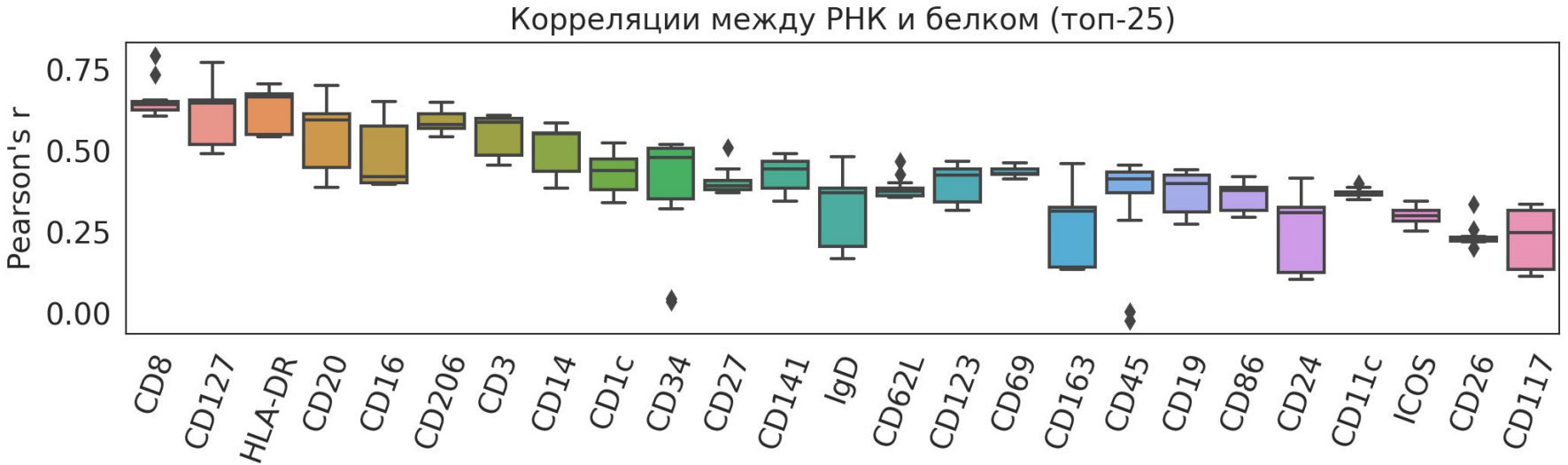
1. **14 матриц экспрессии** CD45+-клеток, полученных в результате CITE-Seq-эксперимента по сравнению иммунного микроокружения аденокарциномы лёгкого (1 матрица = 1 пациент). — *загружено из GEO*
2. Библиотеки **Python 3**: numpy, pandas, seaborn, scanpy, scipy, sklearn и другие.

Препроцессинг данных

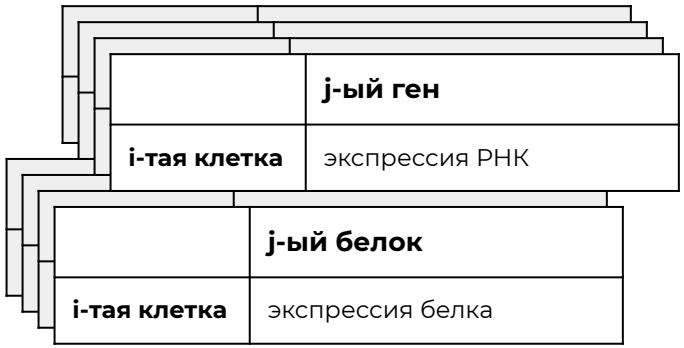


РНК vs. белок

Экспрессия различных белков по-разному скоррелирована с экспрессией соответствующих РНК.



Предсказательная модель



Train:
13 пар экспрессия РНК / экспрессия белка

Test:
1 пара экспрессия РНК / экспрессия белка

линейная регрессия

представленность белка

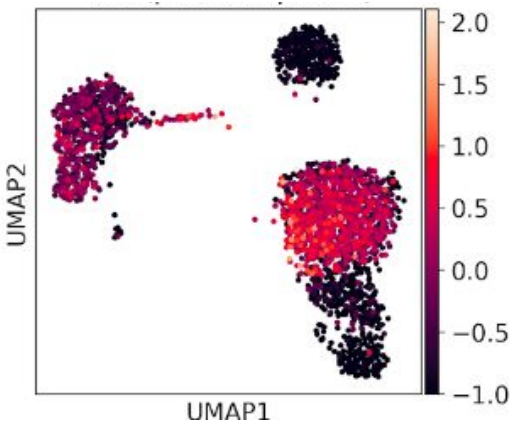
$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

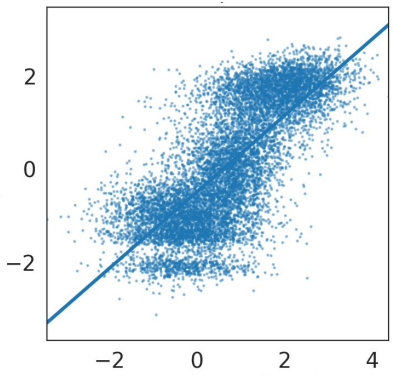
коэффициенты, определяемые с помощью обучения

экспрессия *i*-го транскрипта

обучение и оценка качества модели на тестовом датасете

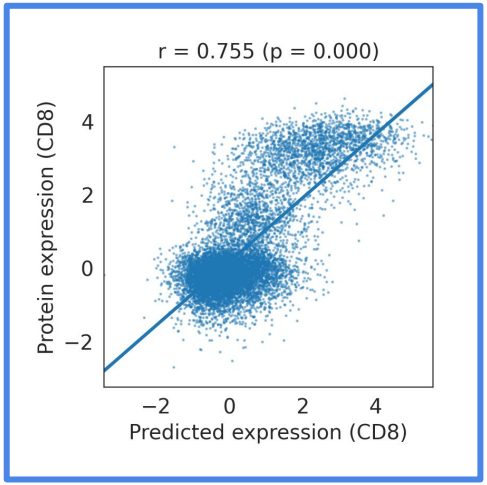
использование модели для клеточного типирования

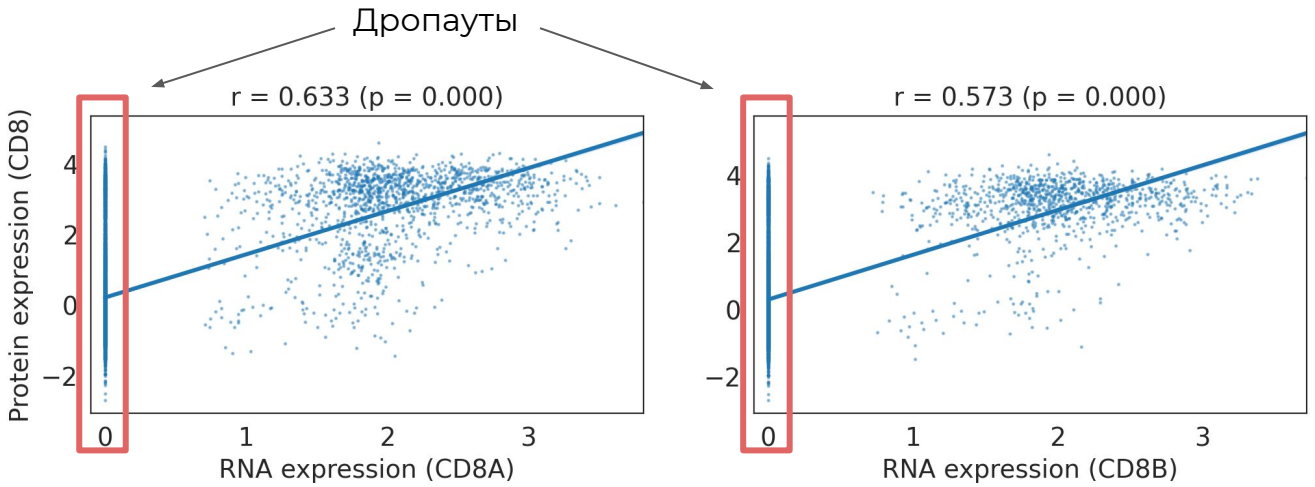


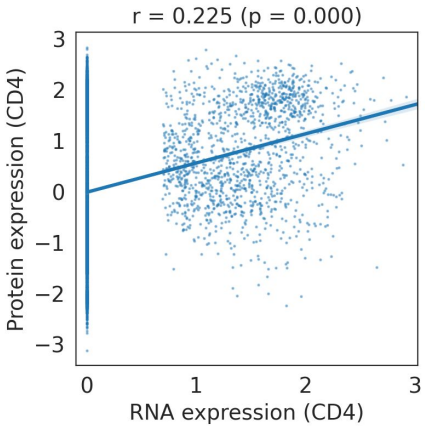


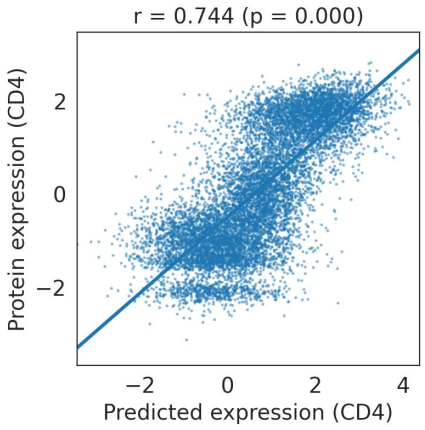
Оценка качества предсказания

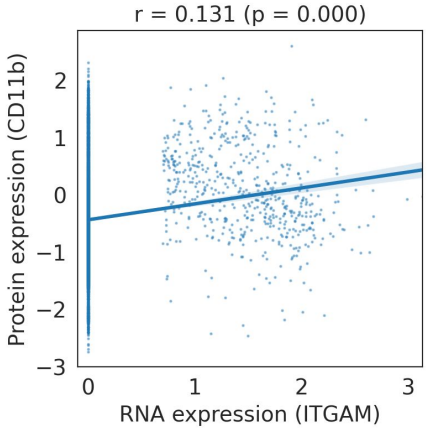
Оценка представленности белка на поверхности клетки при помощи линейной регрессии неплохо скоррелирована с истинным значением.

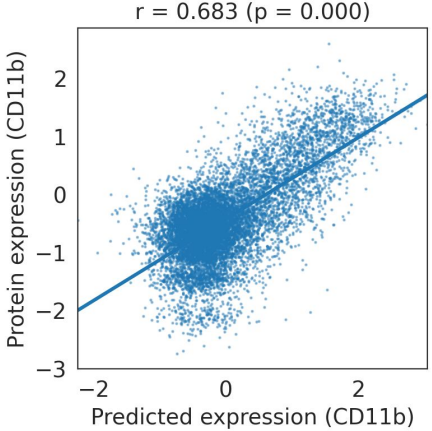


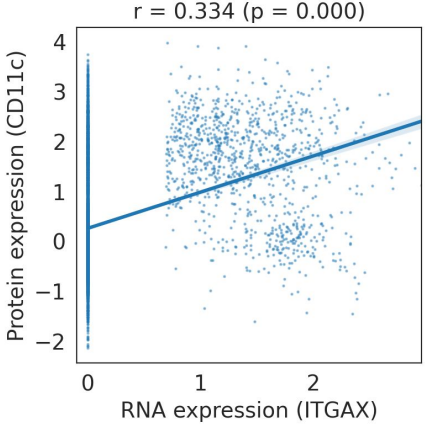


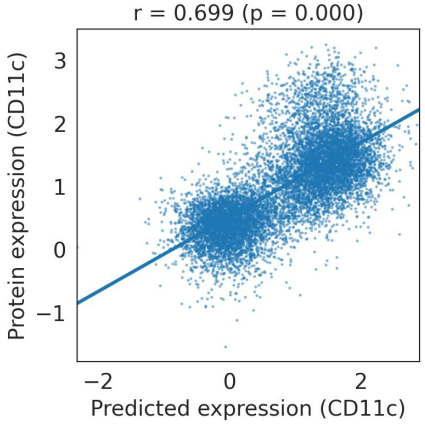

r = 0.225 (p = 0.000)
Protein expression (CD4)
RNA expression (CD4)

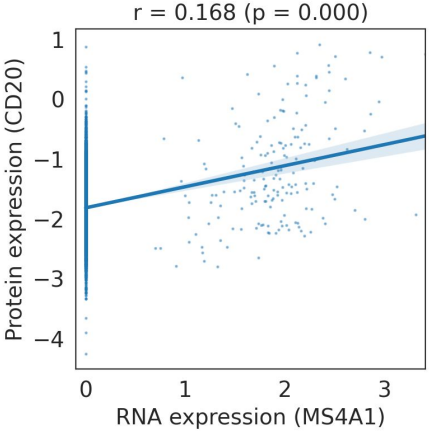

r = 0.744 (p = 0.000)
Protein expression (CD4)
Predicted expression (CD4)

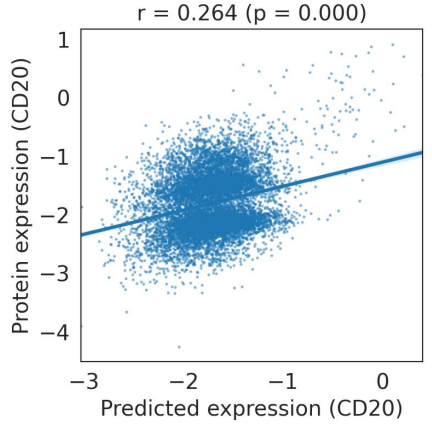

r = 0.131 (p = 0.000)
Protein expression (CD11b)
RNA expression (ITGAM)


r = 0.683 (p = 0.000)
Protein expression (CD11b)
Predicted expression (CD11b)

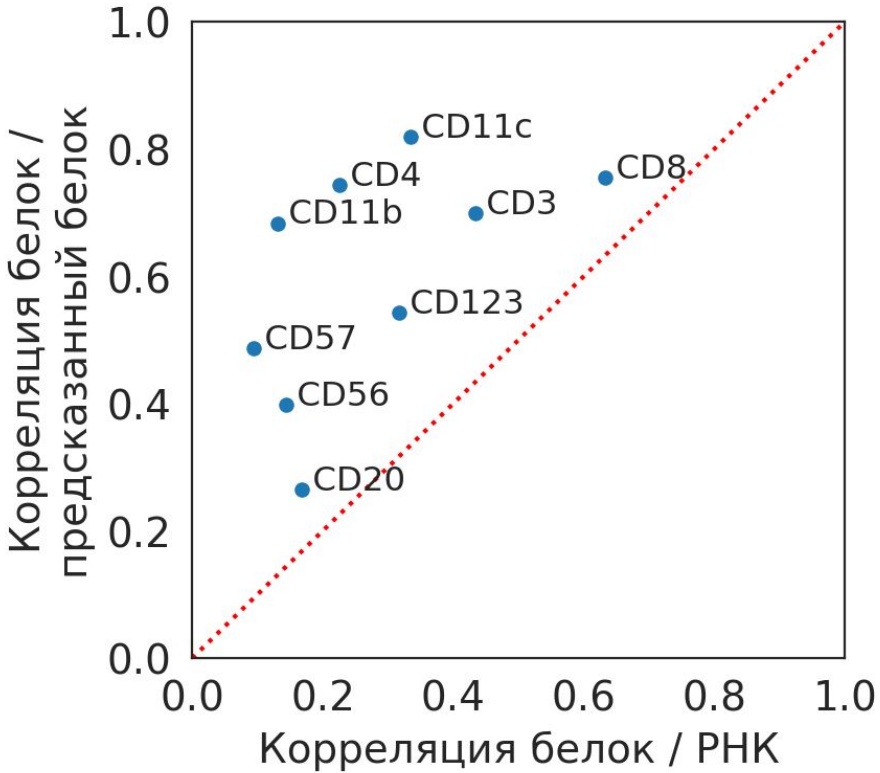

r = 0.334 (p = 0.000)
Protein expression (CD11c)
RNA expression (ITGAX)


r = 0.699 (p = 0.000)
Protein expression (CD11c)
Predicted expression (CD11c)


r = 0.168 (p = 0.000)
Protein expression (CD20)
RNA expression (MS4A1)


r = 0.264 (p = 0.000)
Protein expression (CD20)
Predicted expression (CD20)

Оценка качества предсказания



Оценка представленности белка на поверхности клетки при помощи линейной регрессии **информативнее**, чем при помощи уровня экспрессии соответствующего гена.

Визуализация данных

Test sample	~ 20k генов
~ 10k клеток	экспрессия РНК

Отбор 3 000 самых высоко вариабельных генов

Test sample	3k генов
~ 10k клеток	экспрессия РНК

Шкалирование

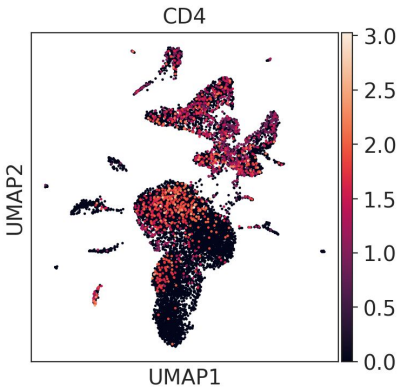
Test sample	3k генов
~ 10k клеток	исправленная экспрессия РНК

PCA и отбор первых 30 компонент

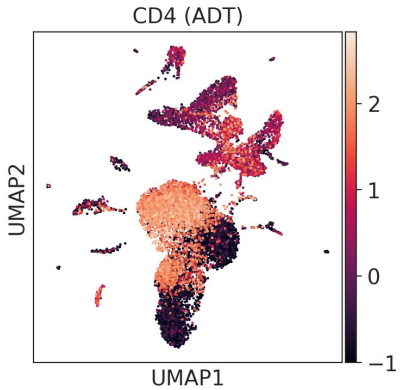
Test sample	30 компонент
~ 10k клеток	Координаты PC

UMAP

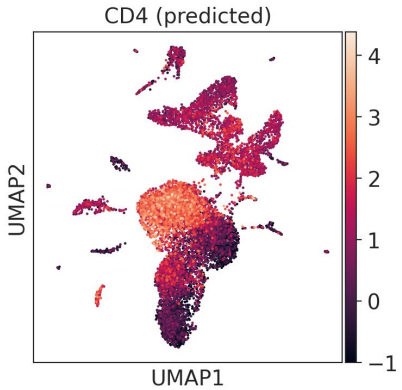
Визуализация экспрессии РНК



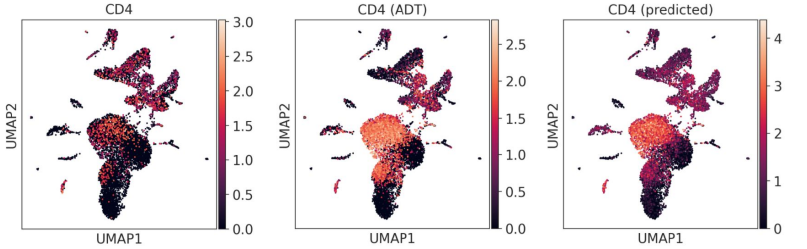
Визуализация экспрессии белка

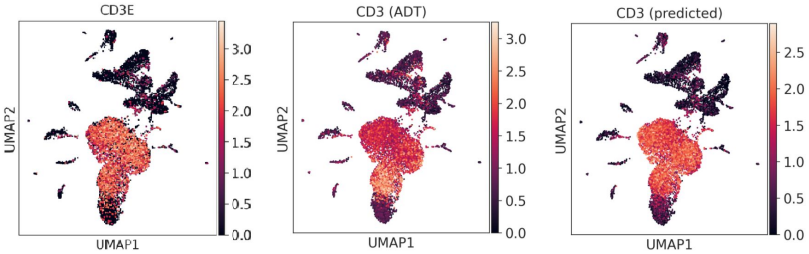


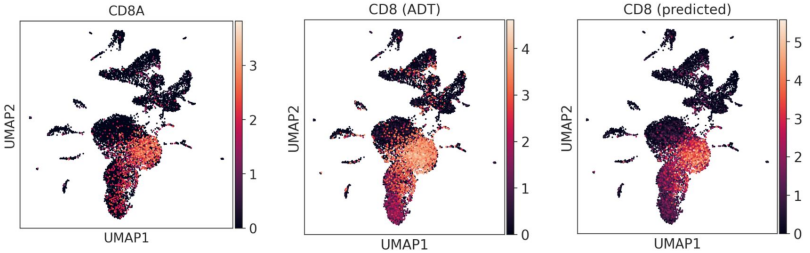
Визуализация предсказанной экспрессии белка

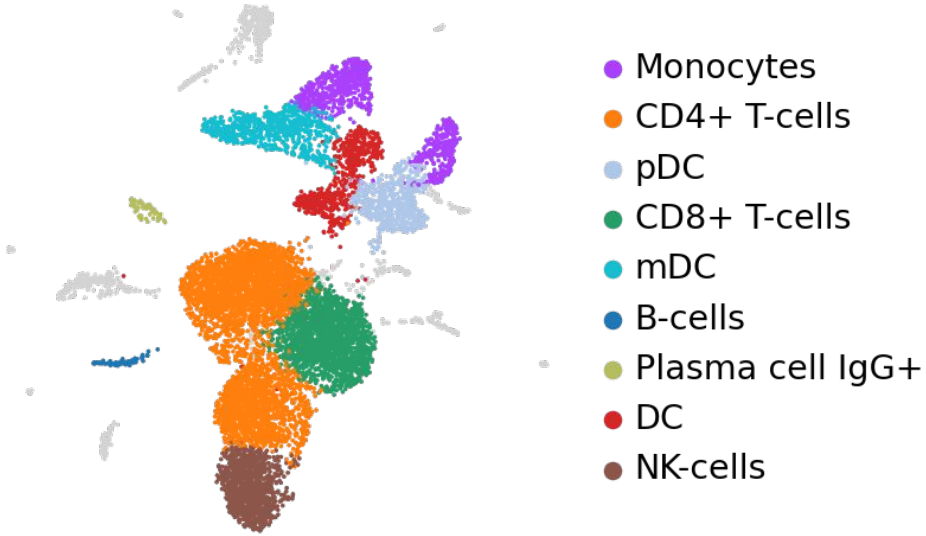


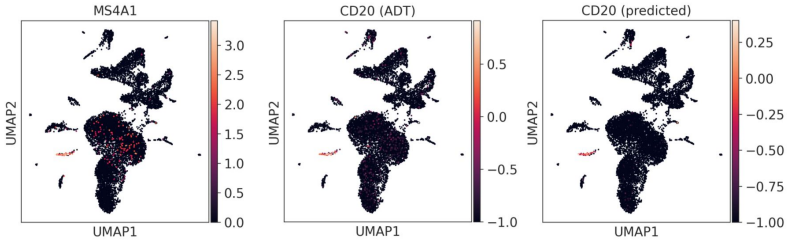
Определение типов клеток





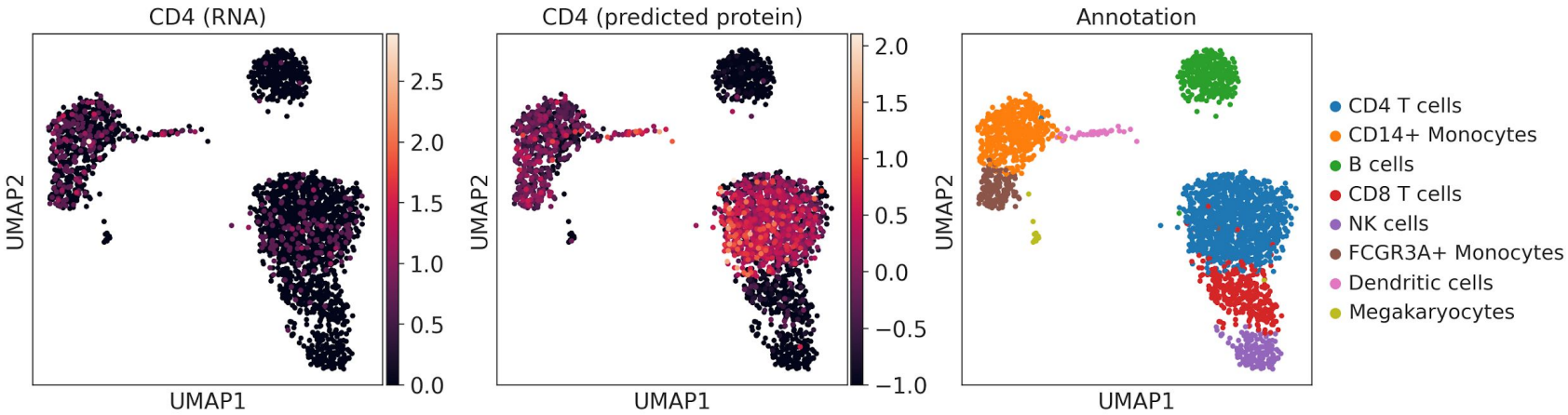






Использование модели

Созданная модель может использоваться для определения клеточных типов в датасетах с scRNA-Seq (**не** CITE-Seq).



Известный пример клеточного типирования открытого датасета PBMC 3k — несмотря наша модель позволяет оценить, какой кластер является CD4+ T-клетками.

Использование модели

Коэффициенты линейного регрессора могут помочь объяснить биологические зависимости, существующие в клетке, а также могут позволить определить генные сети, которые вовлечены в клеточный сигналинг.

index	CDC25A	RASAL1	RNF175	RP11-452C13.1	AC012074.2	RP11-343H5.6	KRTAP5-AS1	LRRC26	CA6	RP11-1299A16.3	C1QTNF4	CD4
Coefficient	0.204197	0.203634	0.198346	0.191418	0.188253	0.180848	0.176435	0.172346	0.168617	0.163068	0.162569	0.162495

index	TMEM262	CD8B	CD8A	SLC4A10	PTGDR2	GNG3	RP11-291B21.2	SCN1B	RP11-480D4.6	SCGB1B2P	KLRC1	AC083949.1
Coefficient	-0.37442	-0.370146	-0.310913	-0.256833	-0.228408	-0.217252	-0.187065	-0.186127	-0.180028	-0.173513	-0.159369	-0.154999

Топ первых и последних по значению коэффициента регрессии генов (на примере модели для определения представленности белка CD4).

SMTB

Выводы

1. В ходе работы была произведена оценка корреляции экспрессии РНК и представленности соответствующего белка на поверхности клетки.
2. Также нами была построена масштабируемая математическая модель для определения представленности поверхностных белков клетки на основе данных scRNA-Seq.
3. Мы провели оценку качества работы этой модели на данных, не участвовавших в обучении.
4. Было проиллюстрировано, что предсказания модели можно использовать для более эффективного фенотипирования клеток.

Вклад авторов

Три первых автора внесли одинаковый вклад в развитие проекта.

Васильев Д. А. осуществлял препроцессинг необработанных экспрессионных матриц и проводил аннотацию типов клеток у 56 пациента.

Мурашов И. Г. разрабатывал дизайн ML-эксперимента и занимался предсказанием представленности белков на поверхности клеток.

Панова Ю. С. занималась сбором данных, визуализацией зависимостей и осуществляла литературную поддержку биомаркеров, используемых для аннотации типов клеток 56 пациента.

Исаев С. В. курировал проект.

Благодарности

Спасибо всем **организаторам ШМТБ** за силы и инициативу провести Школу даже в условиях мировой пандемии Covid-19 и в необычном для Школы формате.

Особую благодарность мы выражаем **Лёше Дорошенко**, **Феде Гагарину** и **Ане Пузырёвой** за организацию облачного вычислительного сервера и его постоянную поддержку.

Идея работы была развита в ходе плодотворных обсуждений с выпускниками ШМТБ, сотрудниками компании BostonGene **Яриком Лозинским** и **Катей Нуждиной**.