«Анализ транскриптомных данных»

Лекция #15.

Анализ мультимодальных омик одиночных клеток

Серёжа Исаев

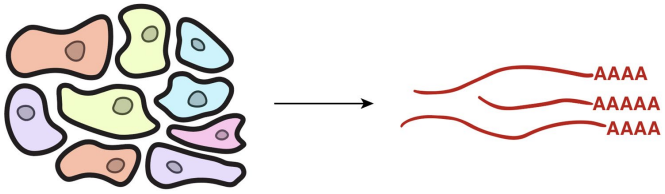
аспирант **ФБМФ МФТИ**
аспирант **MedUni Vienna**

Унимодальные омики одиночных клеток

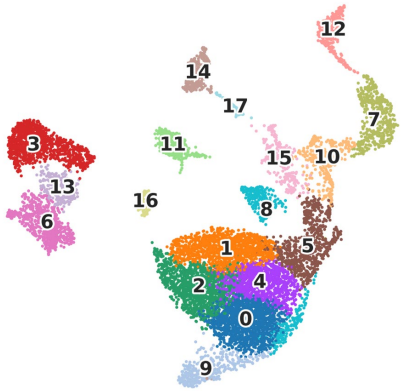
Назовём данные, которые имеют схожую природу (тип биологической молекулы), данными **одной модальности**

scRNA-Seq — классический результат унимодальной омики одиночных клеток

Также в принципе под это определение попадают и данные CyTOF, проточной цитометрии и проч.



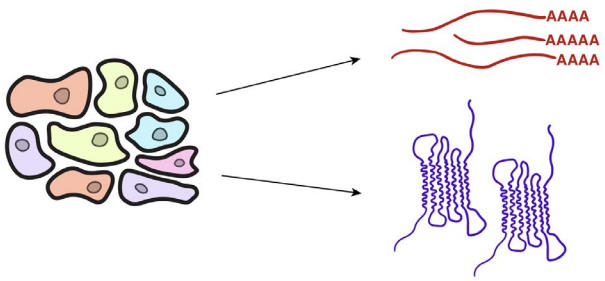
	Cell 1	Cell 2	Cell 3	...	Cell N-1	Cell N
Gene 1	0	0	0	...	0	0
Gene 2	1	0	2	...	0	0
...	...	...	...	...	...	...
Gene M-1	0	3	5	...	2	0
Gene M	1	2	0	...	0	4



Мультимодальные омики одиночных клеток

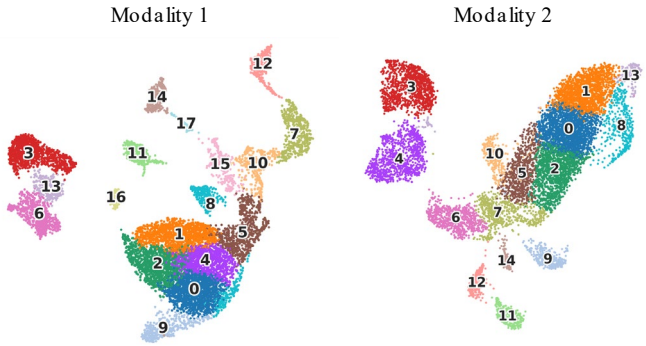
Существует ряд экспериментов, в ходе которых можно получить информацию о данных сразу **нескольких модальностей** для каждого из образцов (клеток)

Каждая из модальностей может быть представлена в виде собственного эмбеддинга с собственными кластерами и закономерностями в данных



	Cell 1	Cell 2	Cell 3	...	Cell N-1	Cell N
Gene 1	0	0	0	...	0	0
Gene 2	1	0	2	...	0	0
...	...	...	...	...	...	...
Gene M-1	0	3	5	...	2	0
Gene M	1	2	0	...	0	4

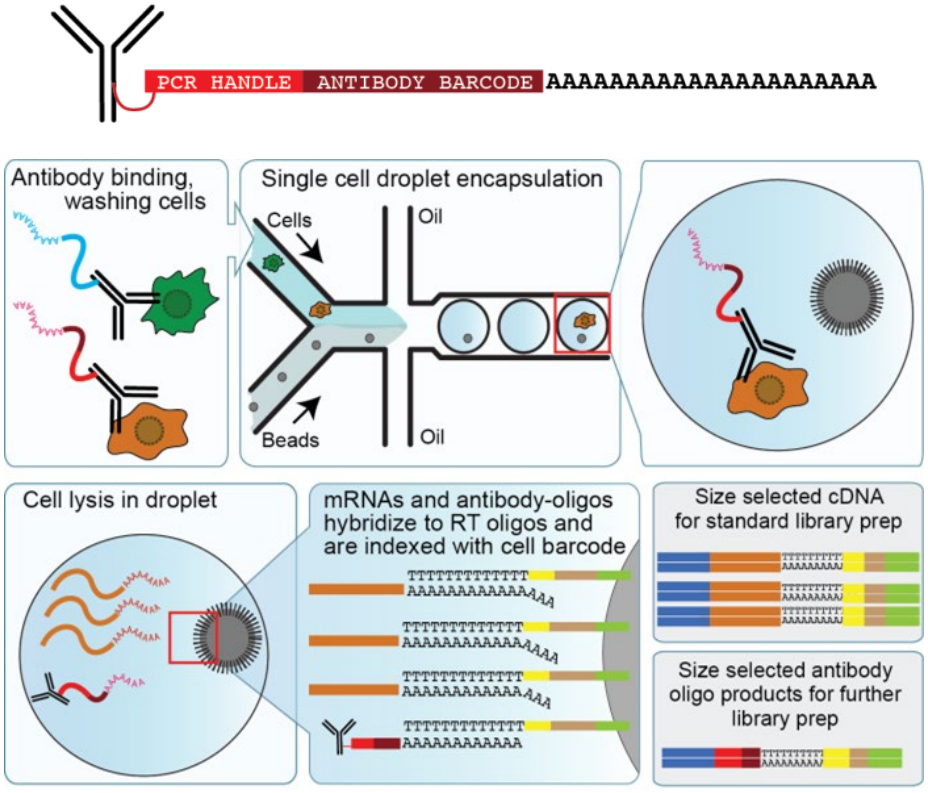
Protein 1	120	98	32	...	63	11
Protein 2	43	3	40	...	19	34
...	...	...	...	...	...	...
Protein K-1	12	140	23	...	10	2
Protein K	82	270	1	...	12	4



CITE-Seq: белки (ADT) + РНК

В ходе CITE-Seq-эксперимента для каждой из клеток мы можем определить представленность поверхностных белков

Количество замеряемых белков лимитировано только панелью и может достигать 250

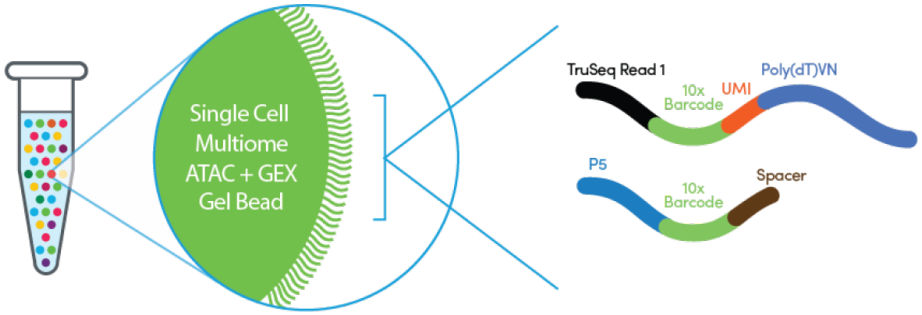


cite-seq.com

10x Multiome

Проводится на ядрах, которые предварительно были обработаны транспозазой (как при подготовке scATAC-Seq)

Теперь шарики содержат праймеры не только на спэйсеры вырезанных регионов хроматина, но и на поли-А РНК



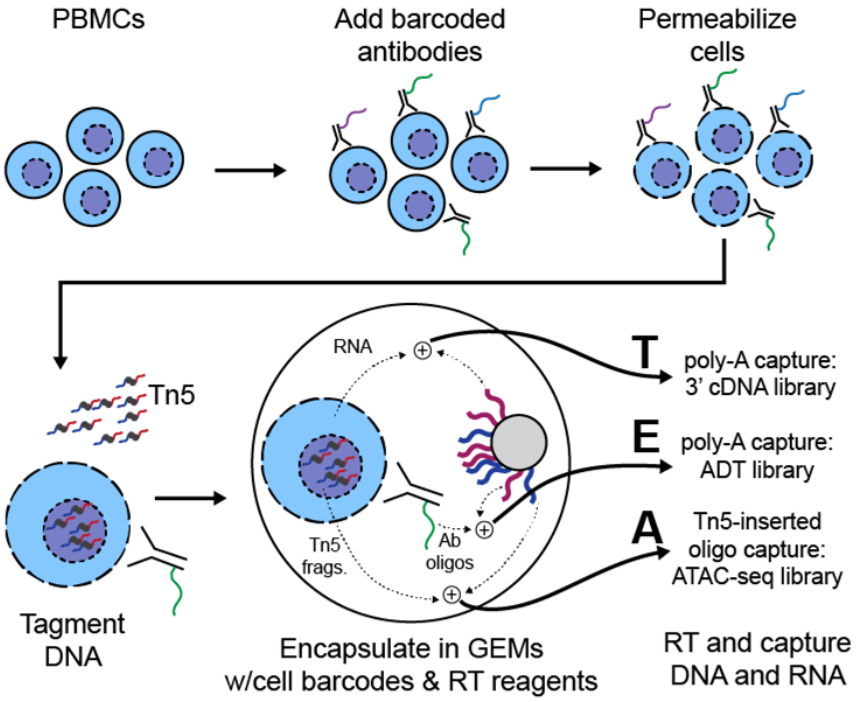
10xgenomics.com

Эксперименты с тремя и более модальностями

Существует ряд экспериментов, которые позволяют измерить больше, чем две модальности

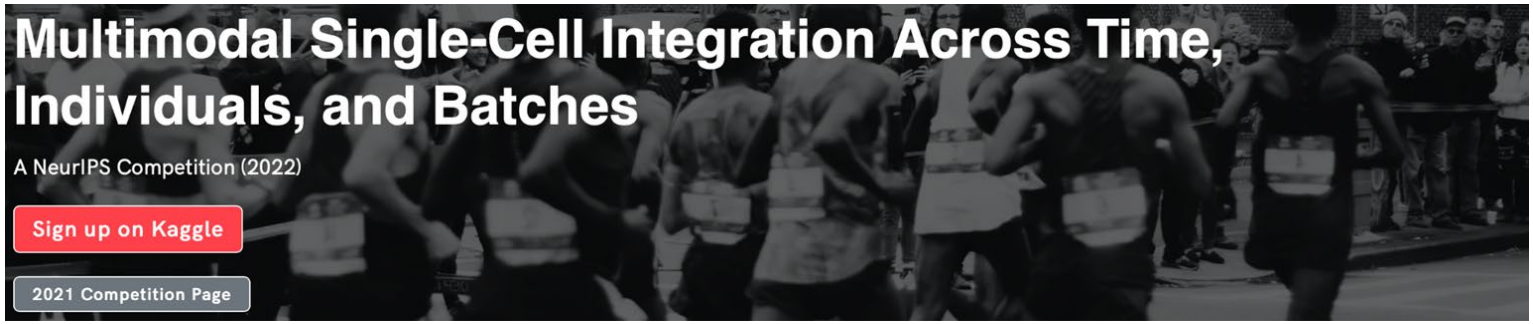
TEA-Seq: РНК + ADT + ATAC
DOGMA-Seq: РНК + ADT + ATAC + мтДНК

Мультимодальные методы продолжают появляться каждый год



protocols.io

Проблема мультимодальной интеграции



Совершенно неясно, какие вычислительные методы позволяют лучше всего работать с мультимодальными омиксными данными одиночных клеток

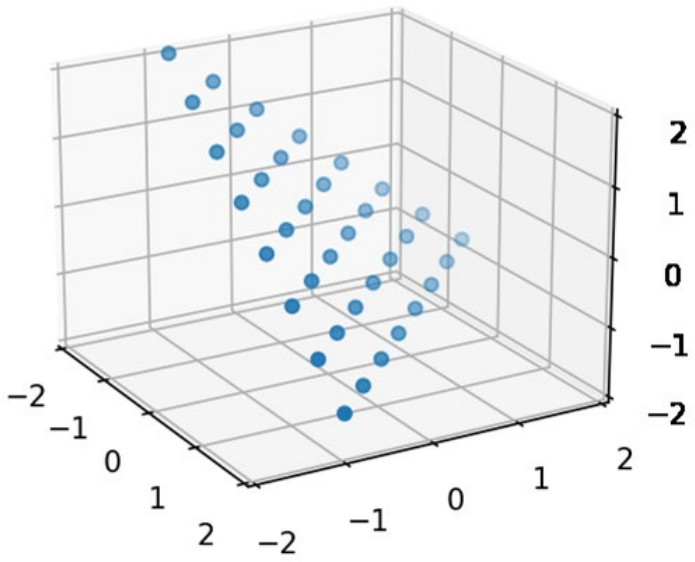
В последние два года проводится открытое соревнование, в ходе которого участники должны предложить самые эффективные методы интеграции таких омик

Процессинг модальности ADT : основные шаги

В целом принцип схож с процессингом РНК и включает в себя следующие шаги:

1. контроль за дисперсией (log1p, **clr**, …),
2. **отсутствует выделение HVG** (у нас и так мало измерений),
3. PCA для уменьшения шума,
4. построение графа k ближайших соседей (kNN) и кластеризация (Louvain, **Leiden**, SNN, …),
5. tSNE или UMAP.

Centered Log -ratio (CLR) transformation



$$\mathrm{clr}(\mathbf{x}) = \left[\ln \frac{x_1}{\mathrm{g}(\mathbf{x})}, \ln \frac{x_2}{\mathrm{g}(\mathbf{x})}, \ldots, \ln \frac{x_D}{\mathrm{g}(\mathbf{x})}\right] = \boldsymbol{\xi}$$

CLR-трансформация данных пытается справиться с той проблемой, что на самом деле наблюдаемые нами значения — это не абсолютные, а относительные экспрессии (ограниченные глубиной покрытия)

Процессинг модальности ATAC : основные шаги

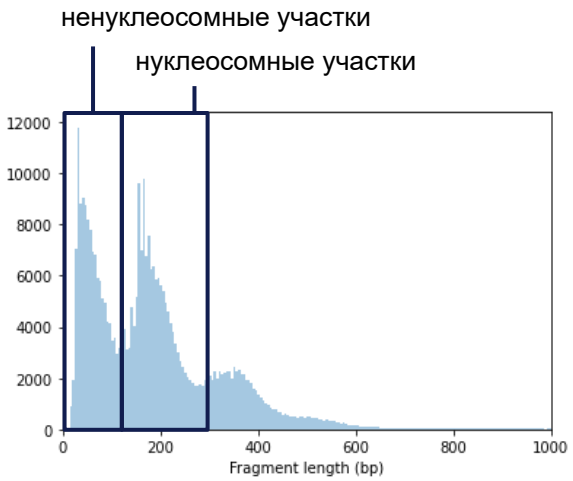
В целом элементы анализа также похожи на процессинг scRNA -Seq, но следует учитывать большую разреженность матриц, с которыми идёт работа:

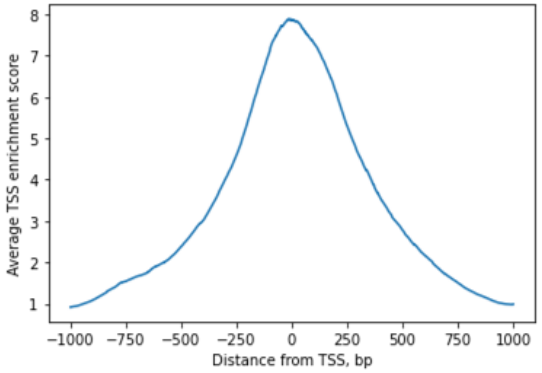
1. QC (**поиск нуклеосомного сигнала** и **обогащённость в регионах TSS**),
2. в некоторых случаях **бинаризация матрицы** ,
3. контроль за дисперсией (log1p, **TF-IDF**),
4. определение HVG,
5. PCA / **SVD** для уменьшения шума,
6. построение графа k ближайших соседей (kNN) и кластеризация (Louvain, **Leiden** , SNN, …),
7. tSNE или UMAP.

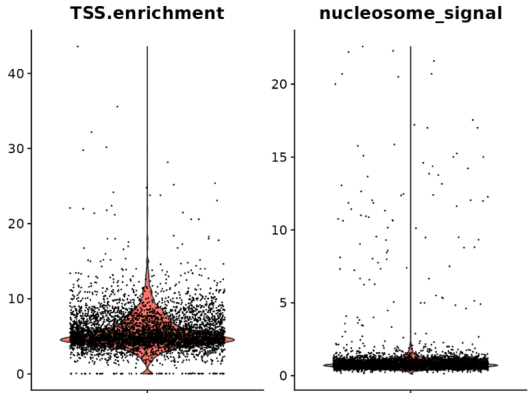
scATAC-Seq-специфический контроль качества

Участки открытого хроматина длиной от 147 до 294 нуклеотидов считаются сигналом от нуклеосомы. Отношение между числом таких отрезков и числом отрезков меньшего размера называют **нуклеосомным сигналом** клетки (чем ниже, тем лучше)

Также покрытие должно быть повышено в регионе TSS, исходя из этого считается **TSS enrichment score**







TF-IDF и LSI

TF-IDF — это метод, пришедший из NLP и позволяющий определить важность слов в документе для того, чтобы определить характеристики документа

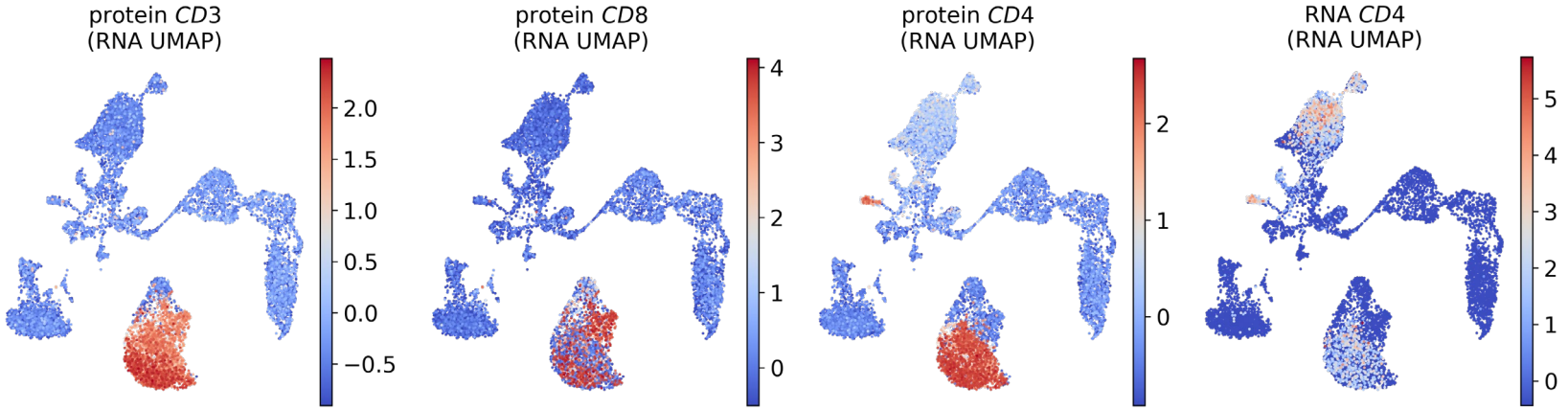
Логика проста — мы умножаем частоту встречи слова в документе (= долю UMI данного пика от всех UMI в клетке) на логарифм отношения всех документов (числа всех клеток) к числу документов, содержащих данное слово (число клеток, содержащих этот пик)

LSI — это следующий за TF-IDF SVD

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

Анализ с опорой на scRNA -Seq

Иногда нам достаточно сделать представление данных, основываясь на модальности РНК, и дальше использовать другую модальность для более точного определения типов клеток или выявления иных закономерностей

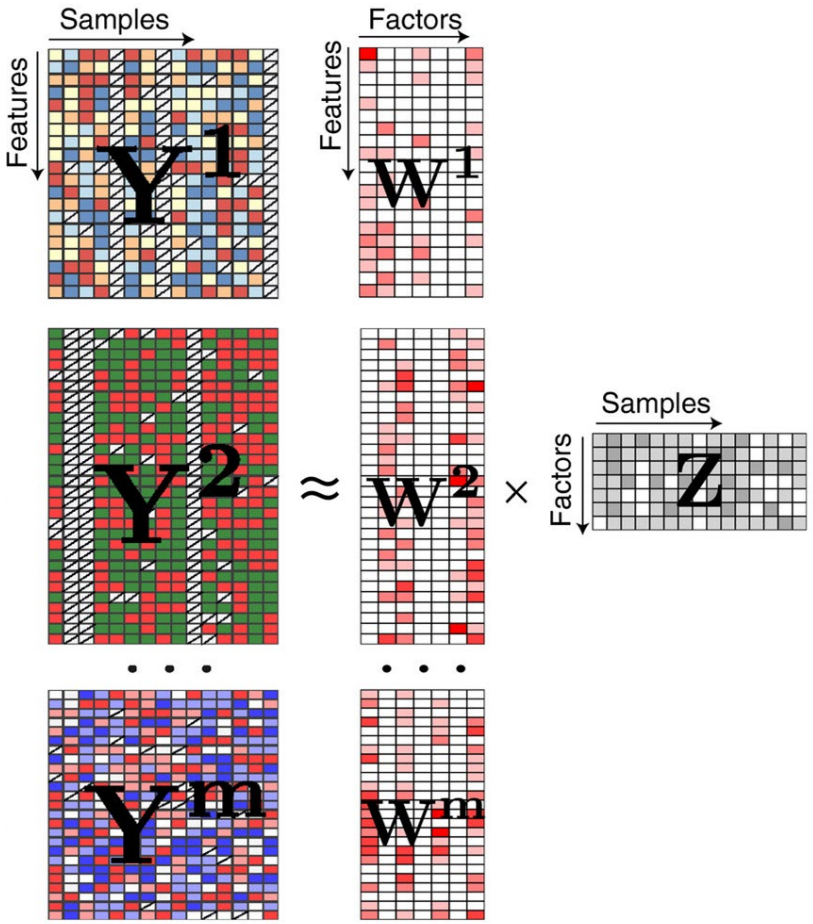


MOFA (MultiOmics Factor Analysis)

Задача мультиомиксного факторного анализа (MOFA) — это представить каждую из модальностей в виде произведения матрицы факторов на матрицу вкладов факторов в каждый из признаков модальностей

Таким образом, каждый фактор в представлении клетки при помощи MOFA — это линейная комбинация признаков из всех модальностей

Некоторый аналог PCA для нескольких матриц

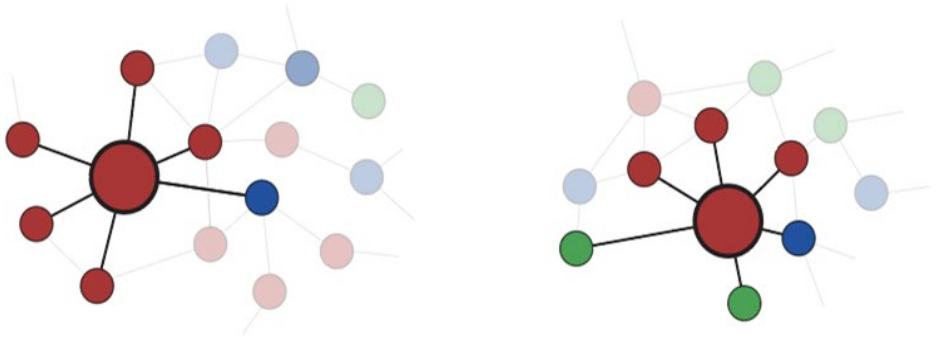


WNN (Weighted Nearest Neighbors)

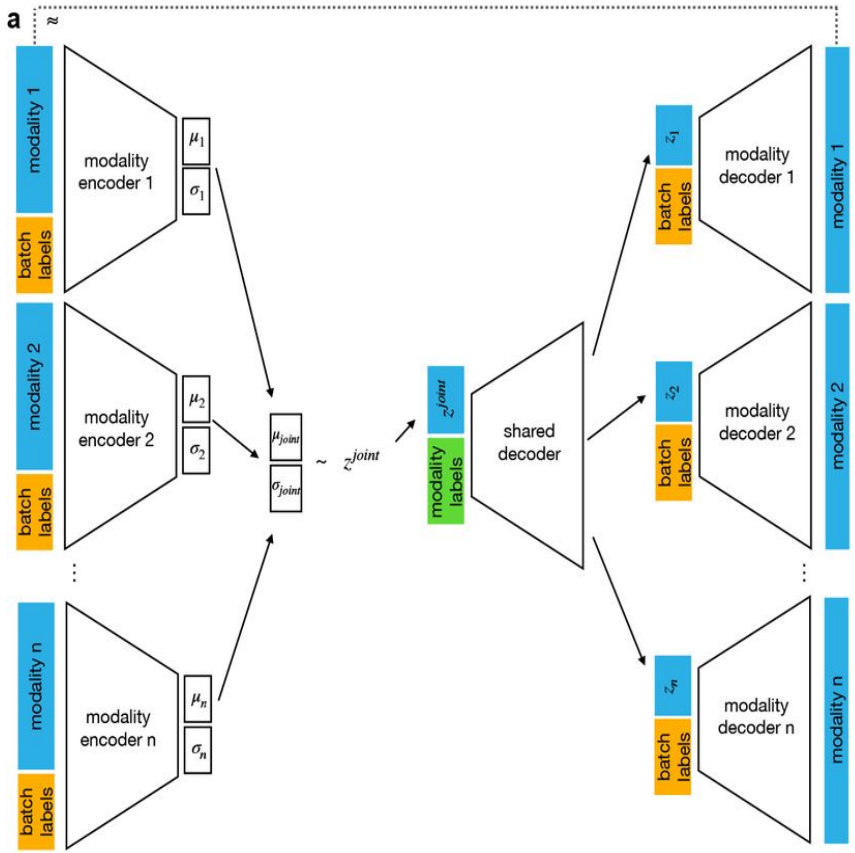
При построении взвешенного графа ближайших соседей (WNN) используются два kNN-графа обеих модальностей

Рёбра, соединяющие точку с соседями, взвешиваются для каждой модальности прямо пропорционально тому, насколько хорошо эта модальность предсказывает другую, и наоборот

То есть, например, если граф по РНК хорошо предсказывает экспрессию белка, то тогда оставляем только граф РНК



totalVI (CITE-Seq) / multiVI (10x Multiome)

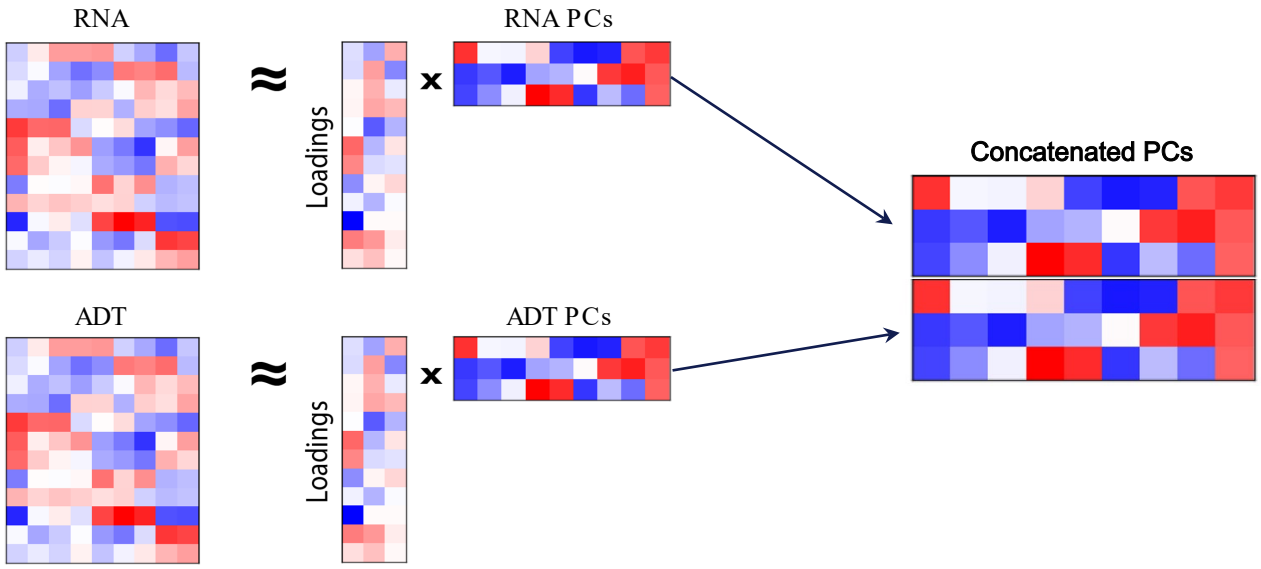


Ряд подходов, основанных на scVI, основаны на аутоэнкодерах, у которых есть свой энкодер и декодер для каждой модальности, однако латентный слой у них один

Иначе говоря, в латентном слое выучивается некоторое низкоразмерное представление данных, а в частях энкодера и декодера — связь этого представления с экспрессиями

Простое объединение PCA

Оказывается (один из результатов **NeurIPS Competition**), что объединение главных компонент из обеих модальностей в некоторой пропорции (по всей видимо, специфичной для набора белков и датасета) даёт в целом неплохие результаты



Коротко о батч -коррекции

Батч-коррекция может проводиться на нескольких уровнях:

1. нормализованная экспрессия (Seurat CCA, Seurat RPCA, scanorama),
2. низкоразмерное представление данных (Harmony, scanorama),
3. граф ближайших соседей (conos, bbkNN).

В зависимости от того, в какой момент проводится батч -коррекция, разные методы по-разному могут комбинироваться с этой процедурой

Сценарии использования батч -коррекции с MOFA

Вообще, в MOFA есть встроенная батч-коррекция, но она работает плохо.
Поэтому можно использовать следующие комбинации:

1. **Seurat CCA** (на каждой модальности) → **MOFA** → kNN
2. **MOFA** → **Harmony** на представлении MOFA → kNN

В целом эти комбинации работают достаточно неплохо

Сценарии использования батч -коррекции с WNN

Так как WNN интегрирует модальности на уровне создания графа, мы не ограничены в использовании методов батч-коррекции до того, как произойдёт интеграция. Сценарии следующие:

1. **Seurat CCA** (на каждой модальности) → PCA → **WNN**
2. **Harmony** на PCA каждой модальности → **WNN**

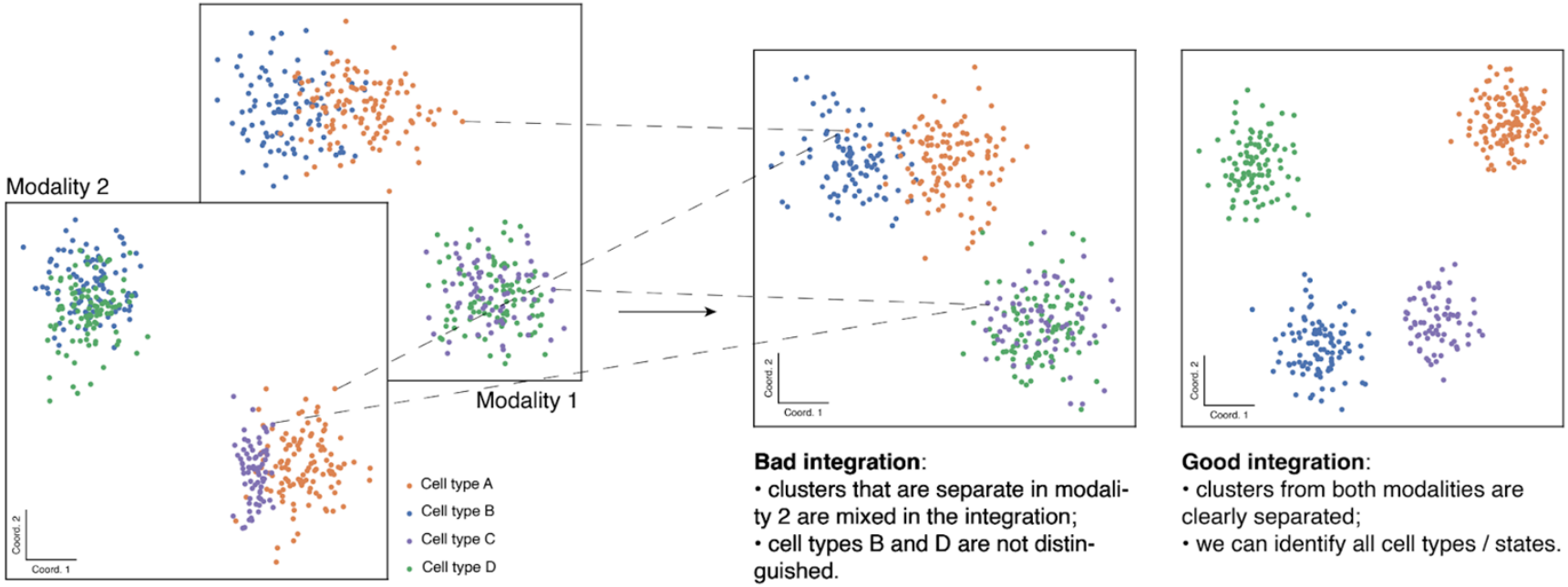
В оригинальной работе с WNN авторы предлагают корректировать батч независимо на каждой модальности при помощи Seurat CCA

Сценарии использования батч -коррекции с totalVI

Так как totalVI предполагает работу на непроцессированных каунтах, то единственный способ батч-коррекции, который применим тут, — это непосредственно коррекция при помощи инструментов scVI (при обучении аутоэнкодера можно указать флаги батчей)

Что мы хотим от интеграции?

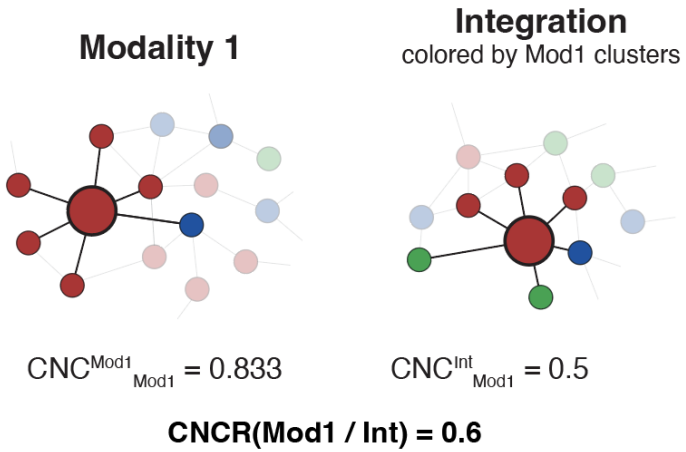
Главное требование к интеграции — это сохранение информации из обеих модальностей

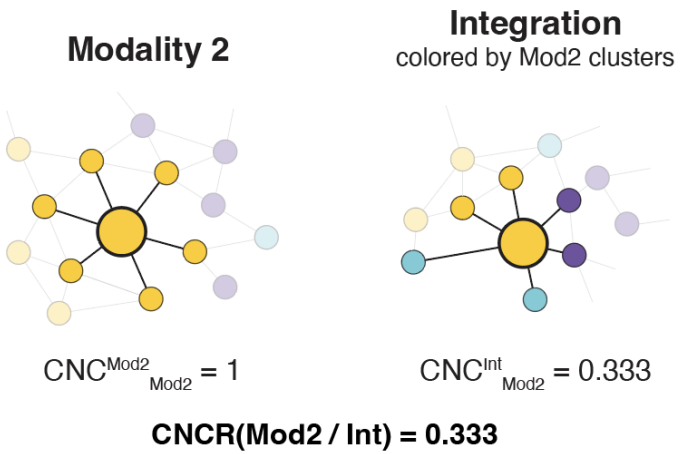


Bad integration:
• clusters that are separate in modality 2 are mixed in the integration;
• cell types B and D are not distinguished.

Good integration:
• clusters from both modalities are clearly separated;
• we can identify all cell types / states.

Метрика CNCR (Cell Neighborhood Consistency Ratio)





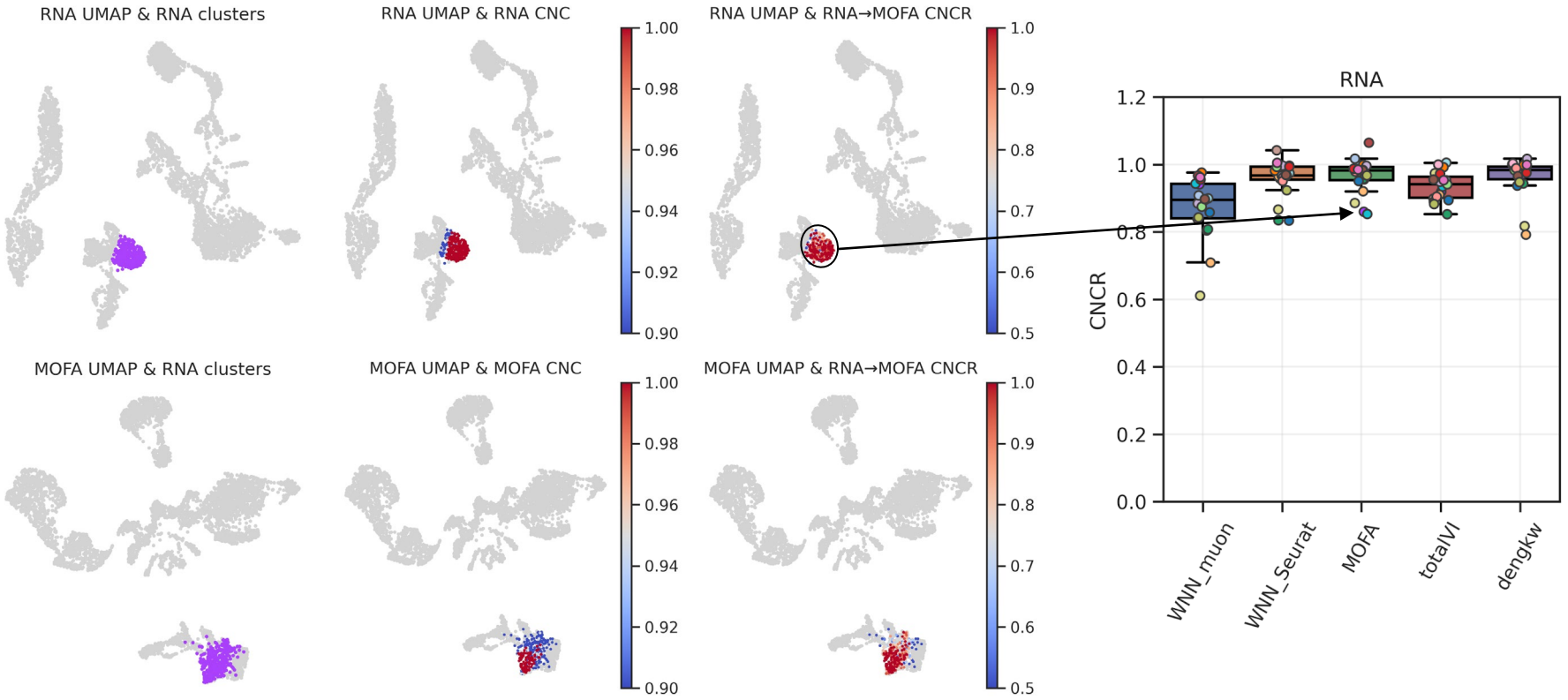
Для оценки локального сохранения информации о кластеризации мы ввели метрику CNCR

Чем выше CNCR, тем лучше в интеграции сохраняется принадлежность к изначальному кластеру соседей клеток

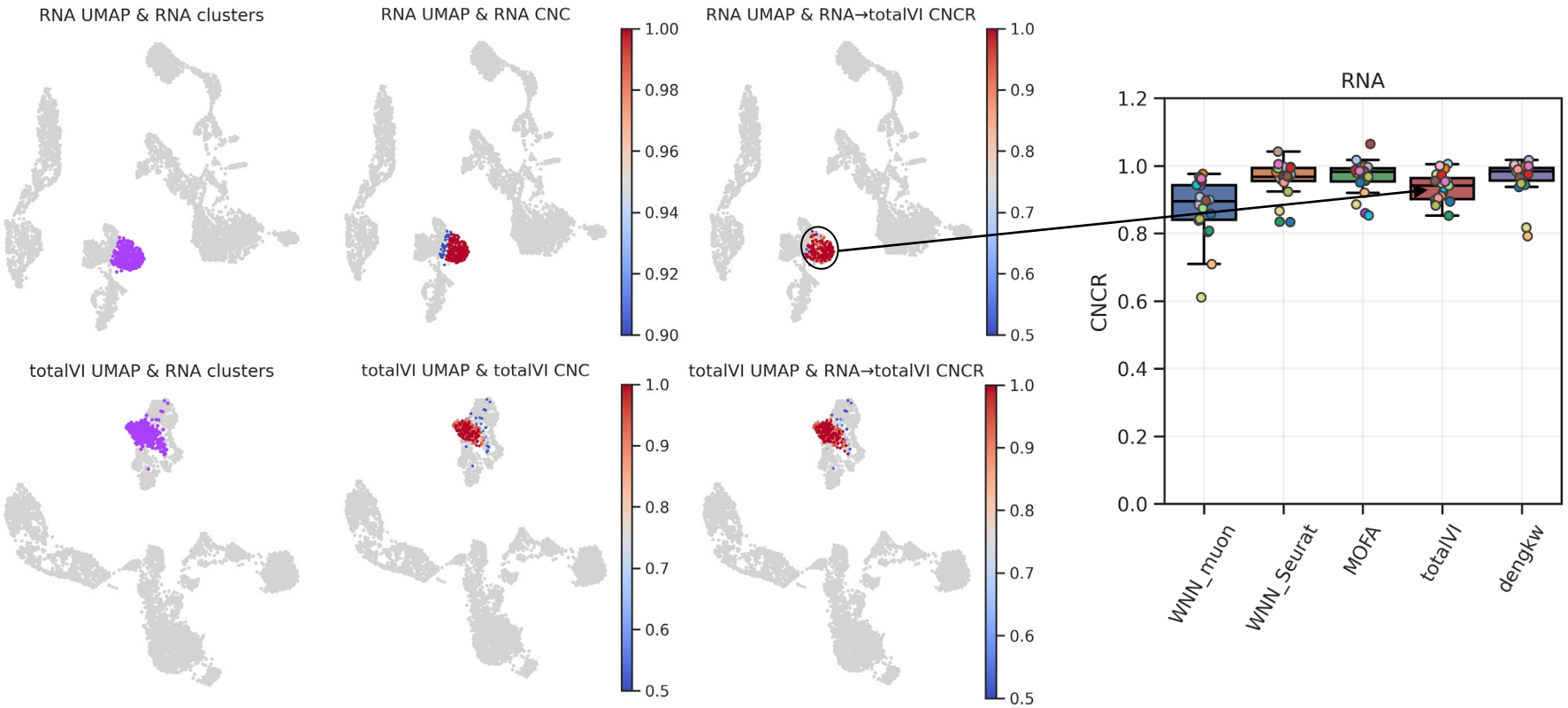
$$\mathrm{CNC} = \frac{M_{\mathrm{SameClusterNeighbors}}}{N_{\mathrm{AllNeighbors}}}$$

$$\mathrm{CNCR} = \mathrm{CNC}_{\mathrm{Ratio}} = \frac{\mathrm{CNC}_{\mathrm{Integration}}}{\mathrm{CNC}_{\mathrm{Modality}}}$$

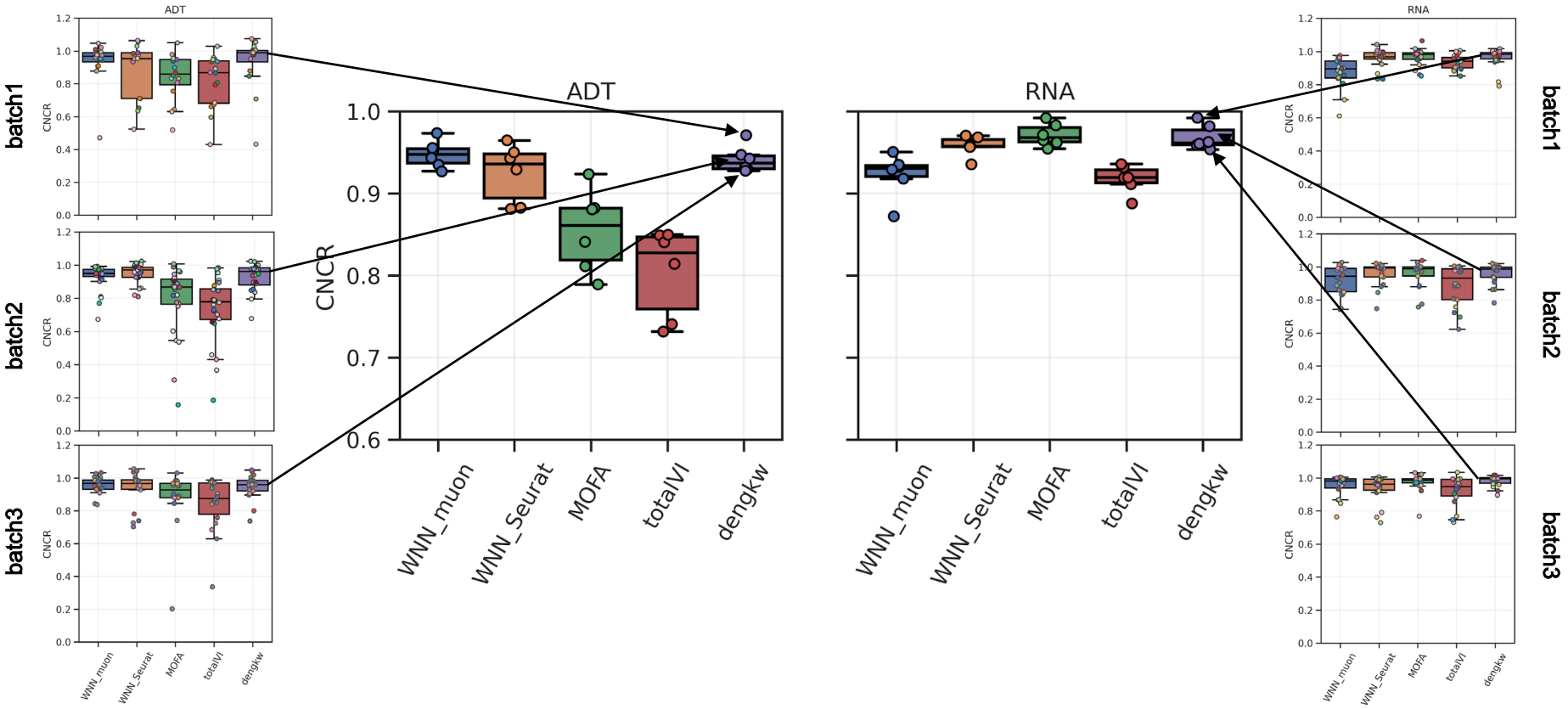
Пример работы метрики



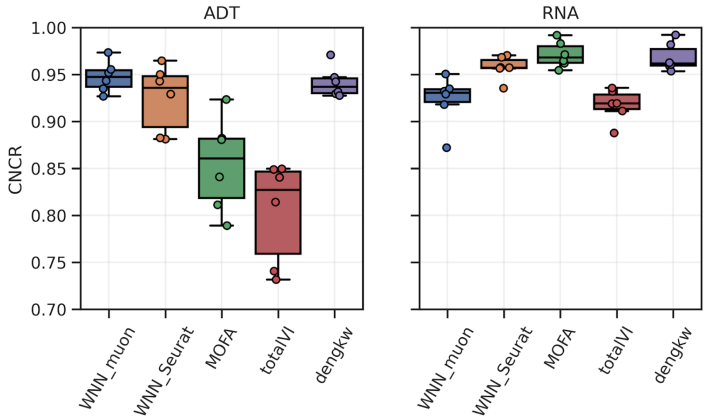
Пример работы метрики

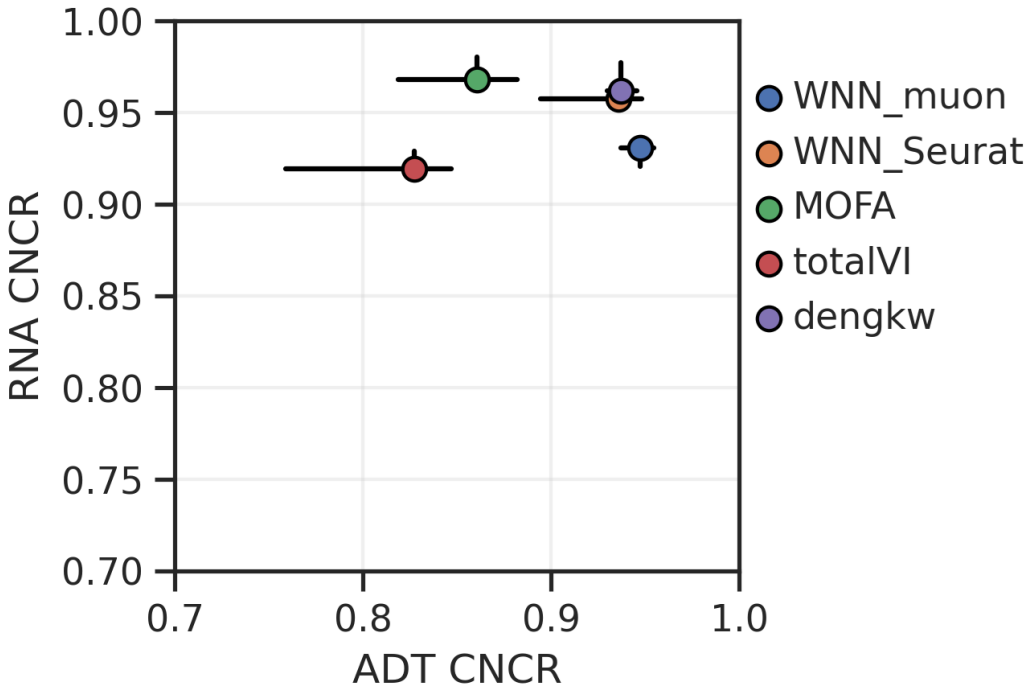


Пример работы метрики



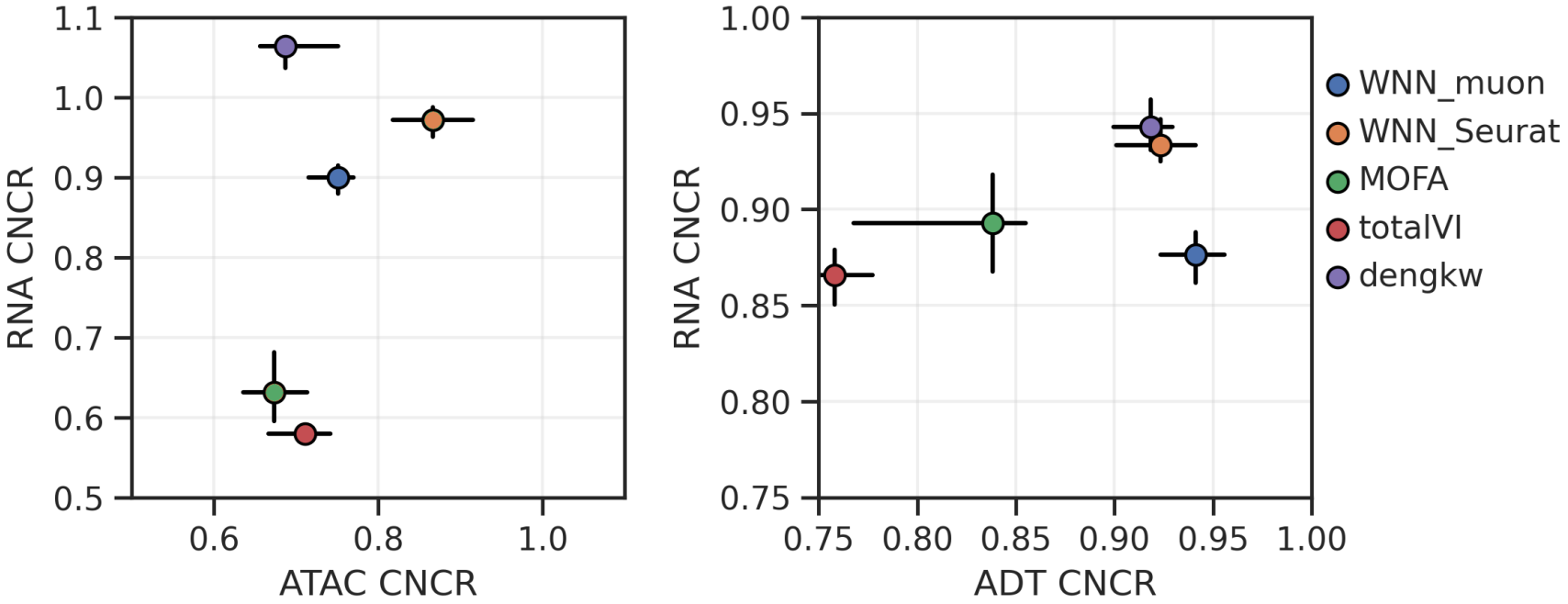
Пример работы метрики



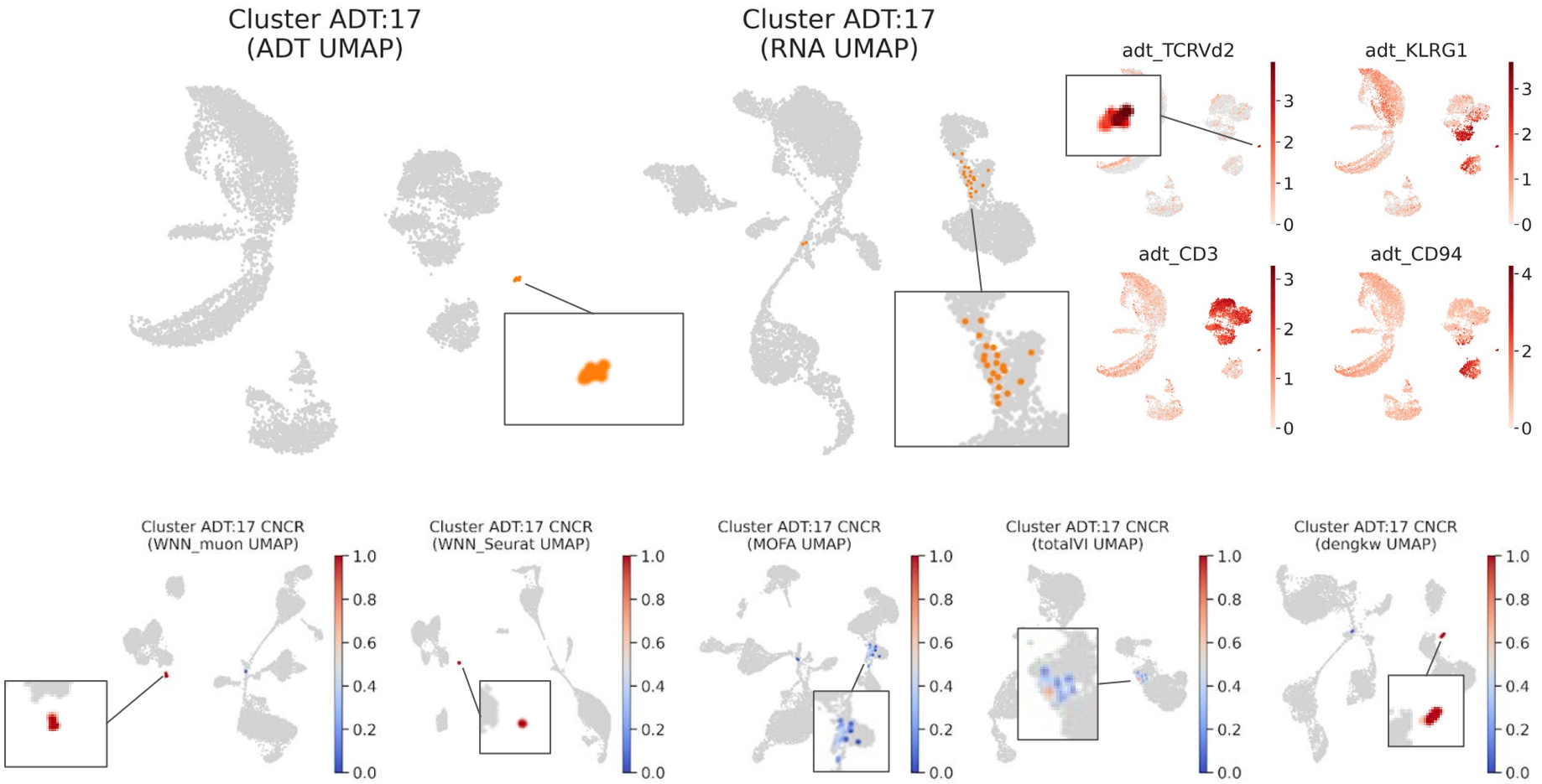


Результаты бенчмаркинга

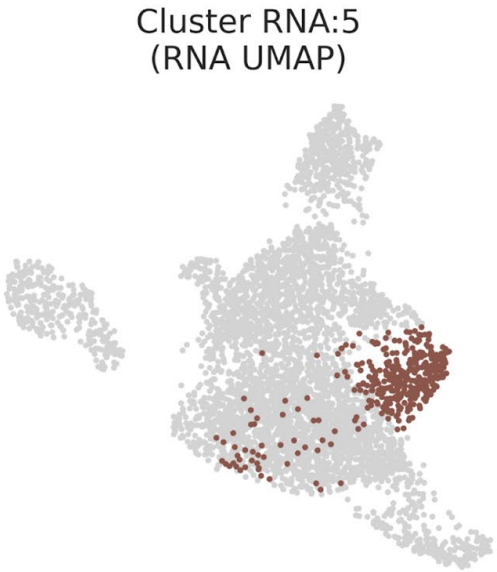
По всей видимости, лучше всего на нашей метрике показывают себя методы **WNN** (из Seurat) и простое склеивание PCA (здесь назван **dengkw**)

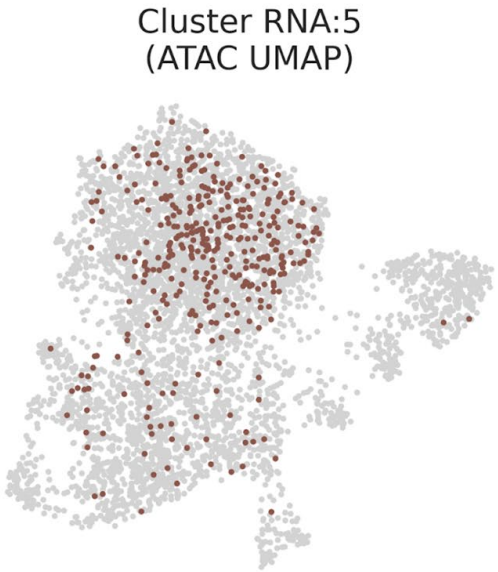


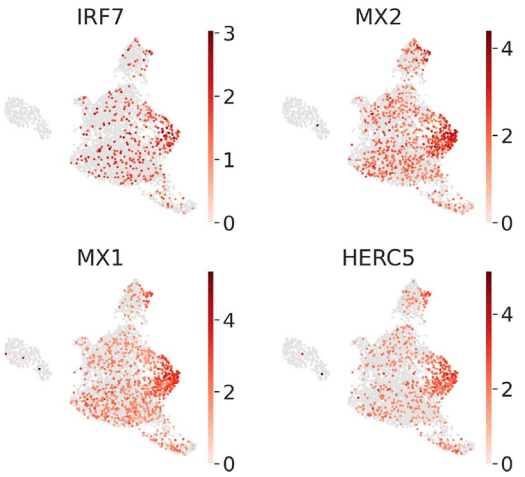
Почему это важно? (Пример из CITE-Seq)

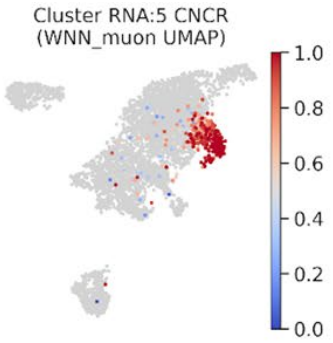


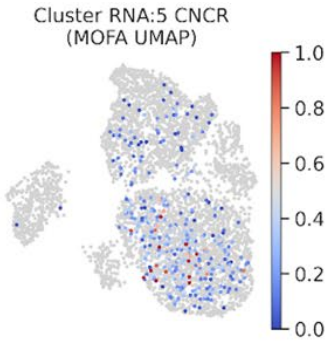
Почему это важно? (Пример из 10x Multiome)

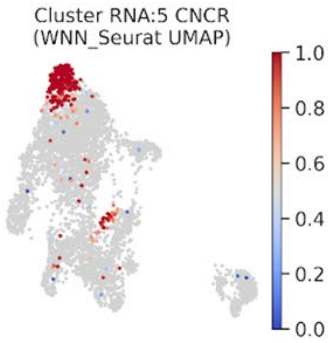


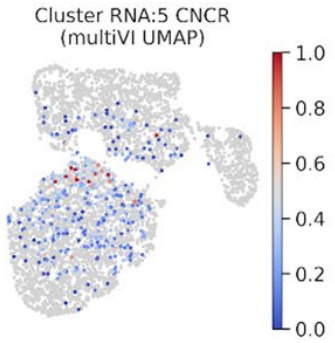


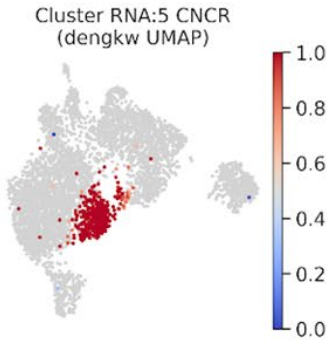




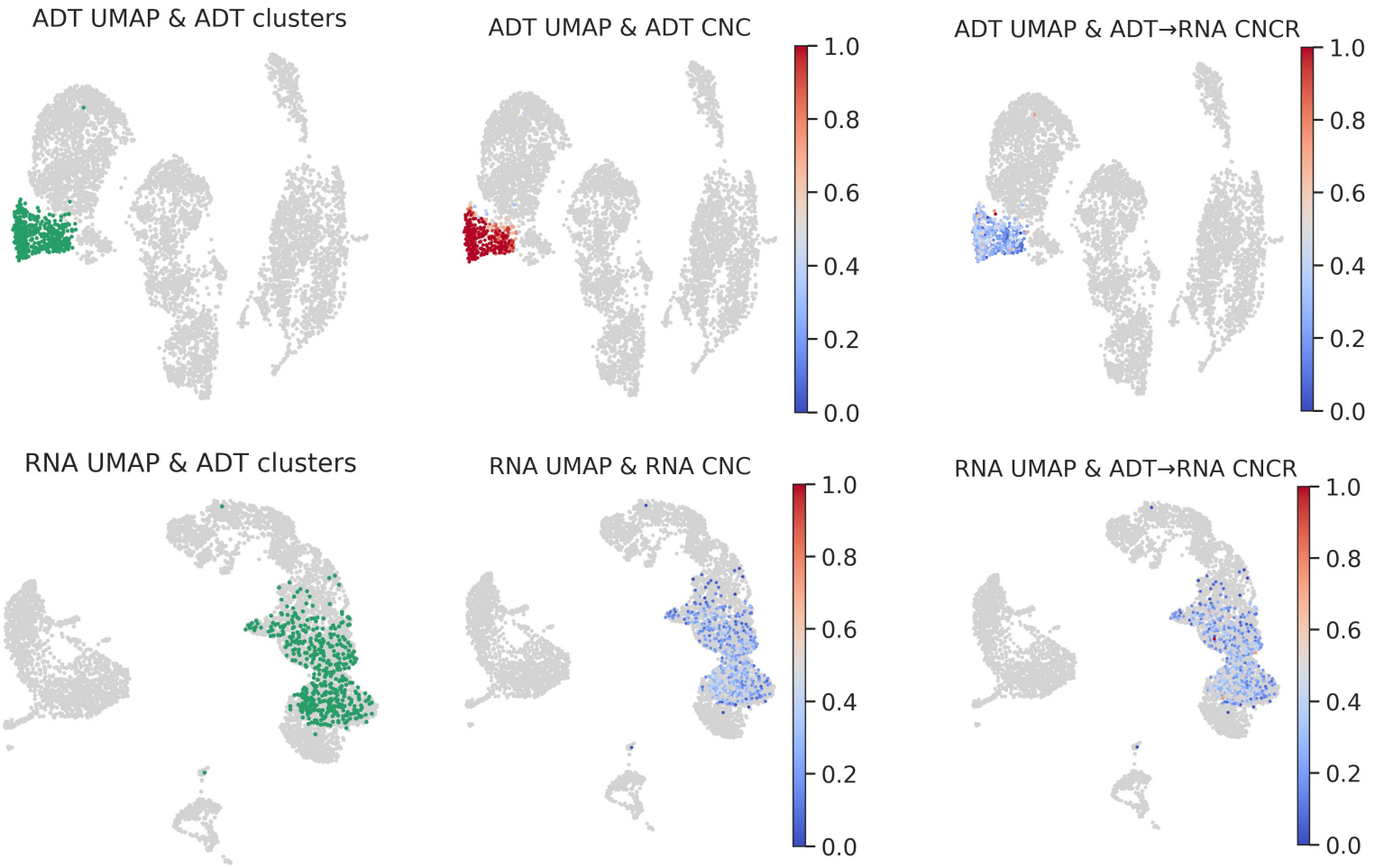




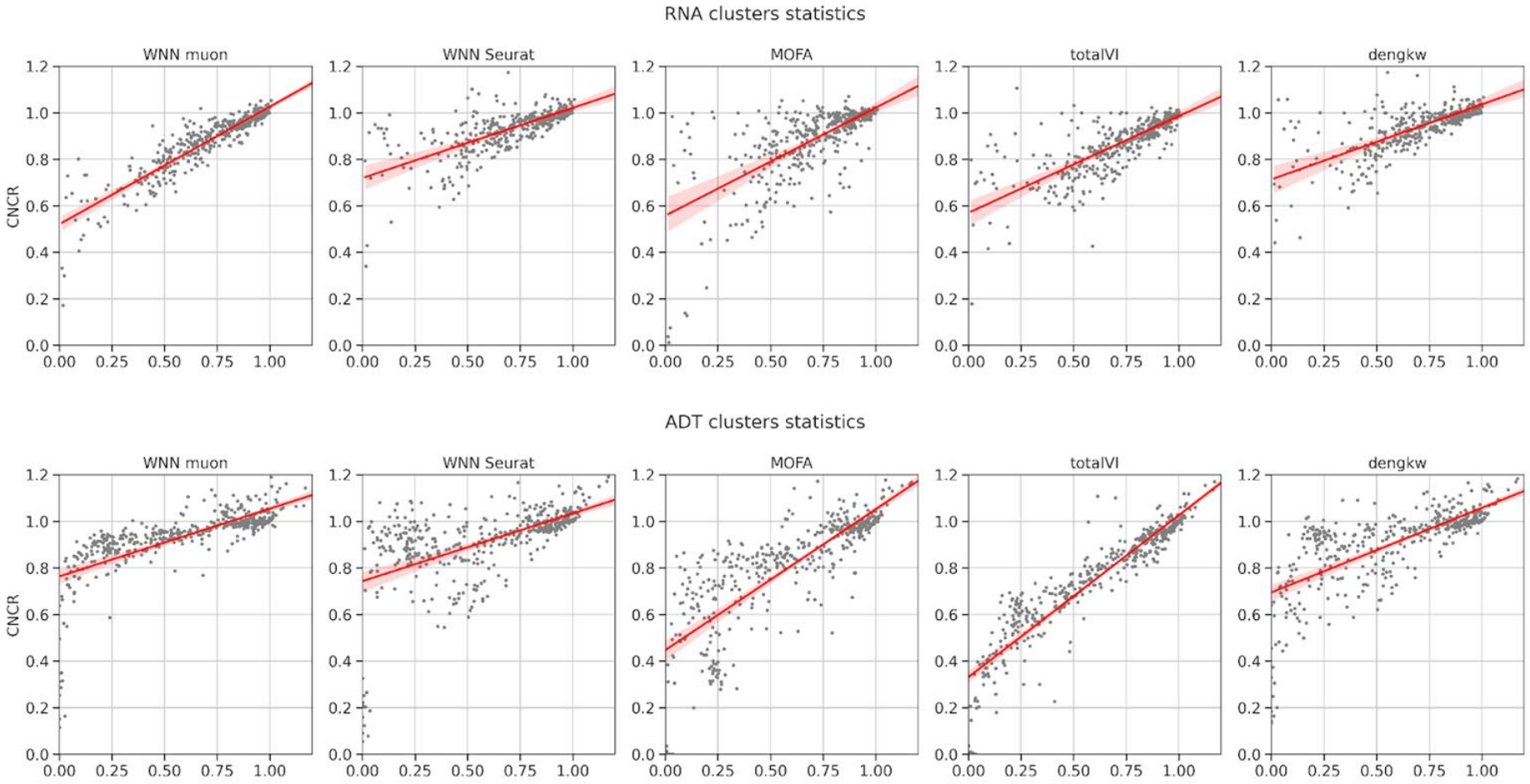




Cross-modality CNCR — причина неудач методов



Cross-modality CNCR — причина неудач методов



Как работать с этими данными?

Seurat v. 4 (пакет для R) поддерживает работу с мультимодальными омиками одиночных клеток

- имплементирован WNN,
- постоянно развивается, открытое активное коммьюнити.

Muon (пакет для Python) основан на **scanpy** и написан специально для работы с мультимодальными омиками одиночных клеток

- имплементированы WNN и MOFA,
- легко адаптировать для работы с totalVI и multiVI,
- поддерживается чуть хуже, чем Seurat.