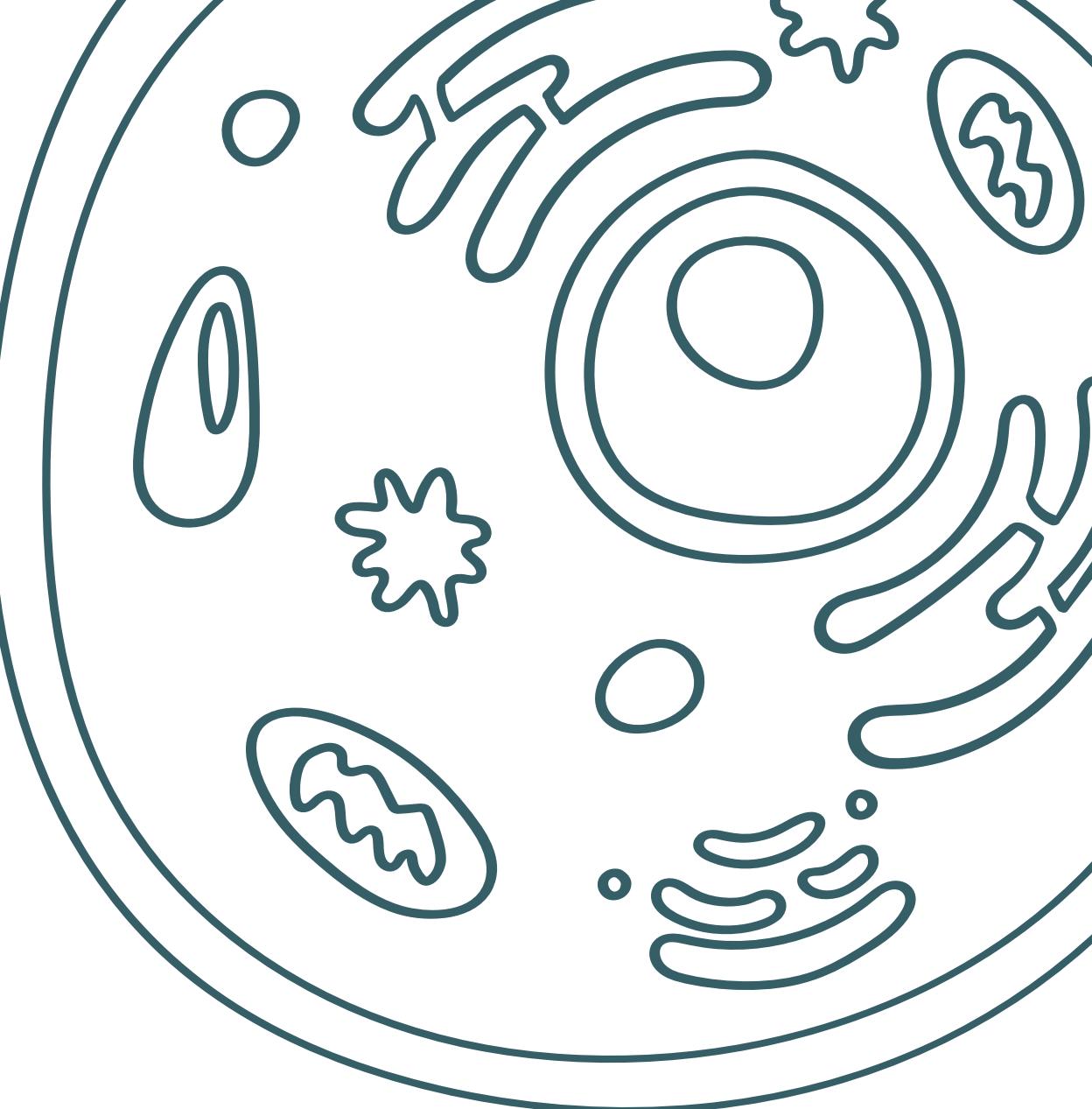


# Single cell RNA-seq analysis

Сергеева Юлия  
Закирова Марфа  
Богдан Елизавета  
Б06-907



# План

- Постановка задачи
- Single cell RNA-seq
- Предобработка данных
- Методы машинного обучения
- Результаты



# Постановка задачи

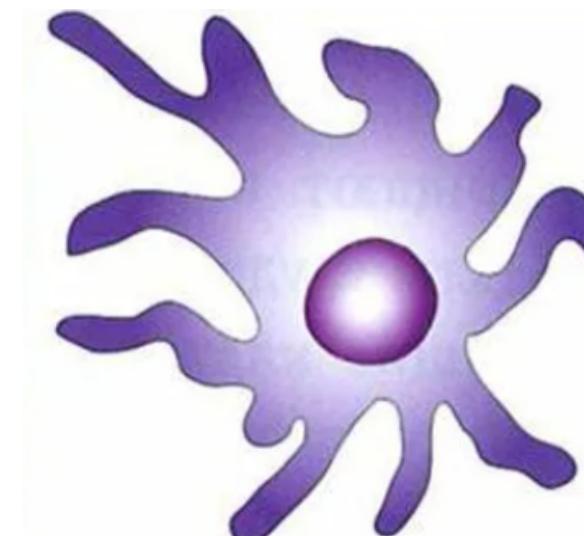
## Цель:

- Разделение мононуклеарных периферических клеток крови на типы после scRNA-seq

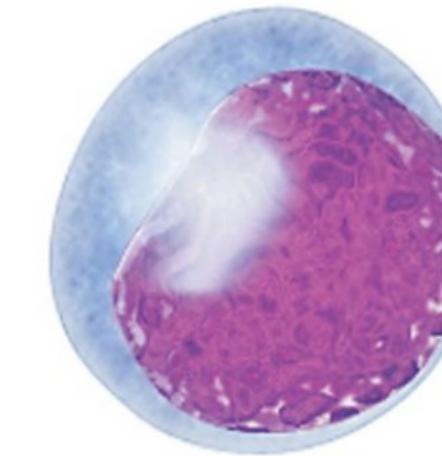
## Данные:

- **matrix**: уровни экспрессии всех генов каждого образца - подсчет уникальных UMI для данного barcode
- **barcodes**: баркоды каждой клетки
- **gene**: названия генов

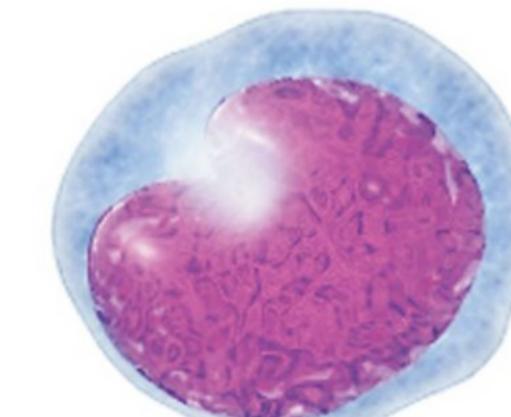
# Мононуклеарные периферические клетки крови



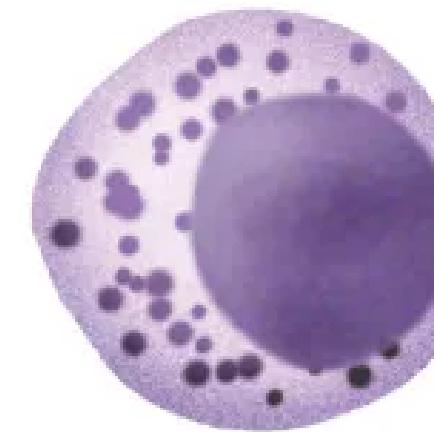
Dendritic cell



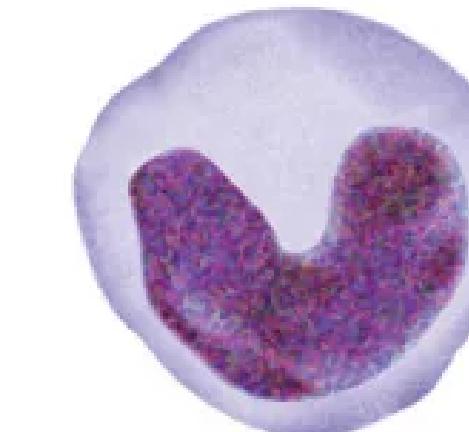
T lymphocyte  
(T cell)



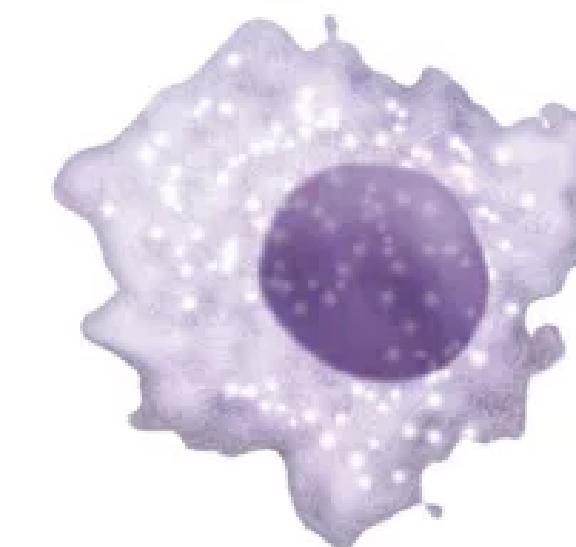
B lymphocyte  
(B cell)



Natural killer cell



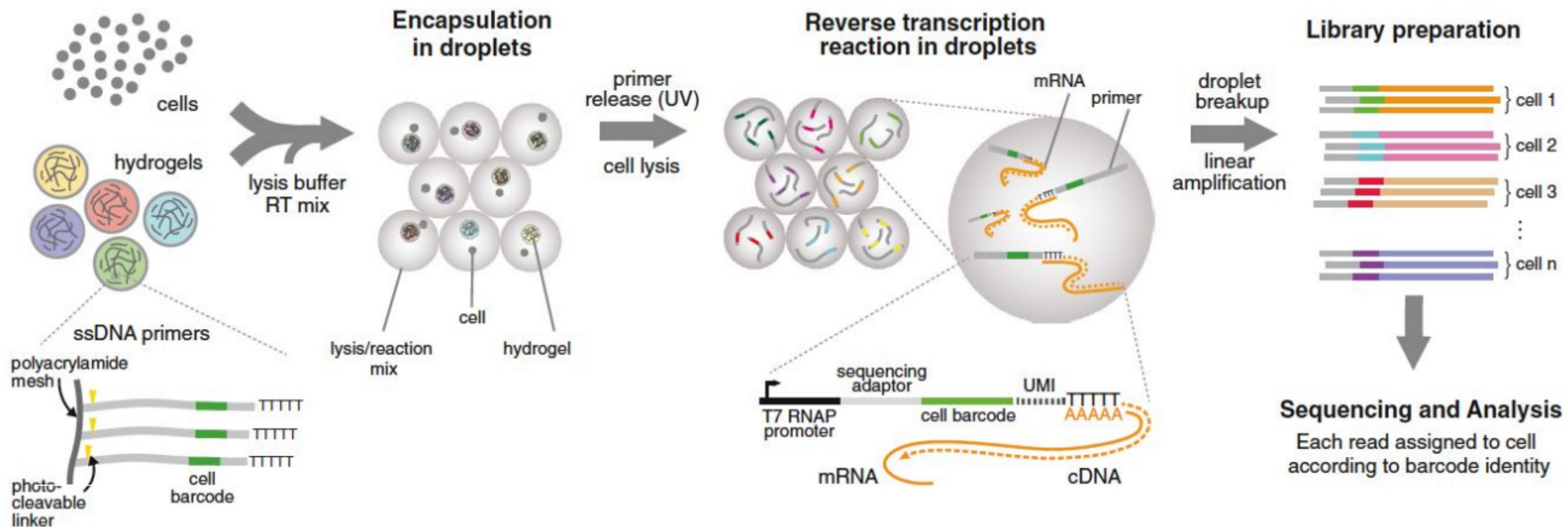
Monocyte



Macrophage

Fig. 10.1, p.177

# SINGLE CELL RNA-SEQ: KLEIN ET AL., 2015



**Figure 1. A Platform for DNA Barcoding Thousands of Cells**

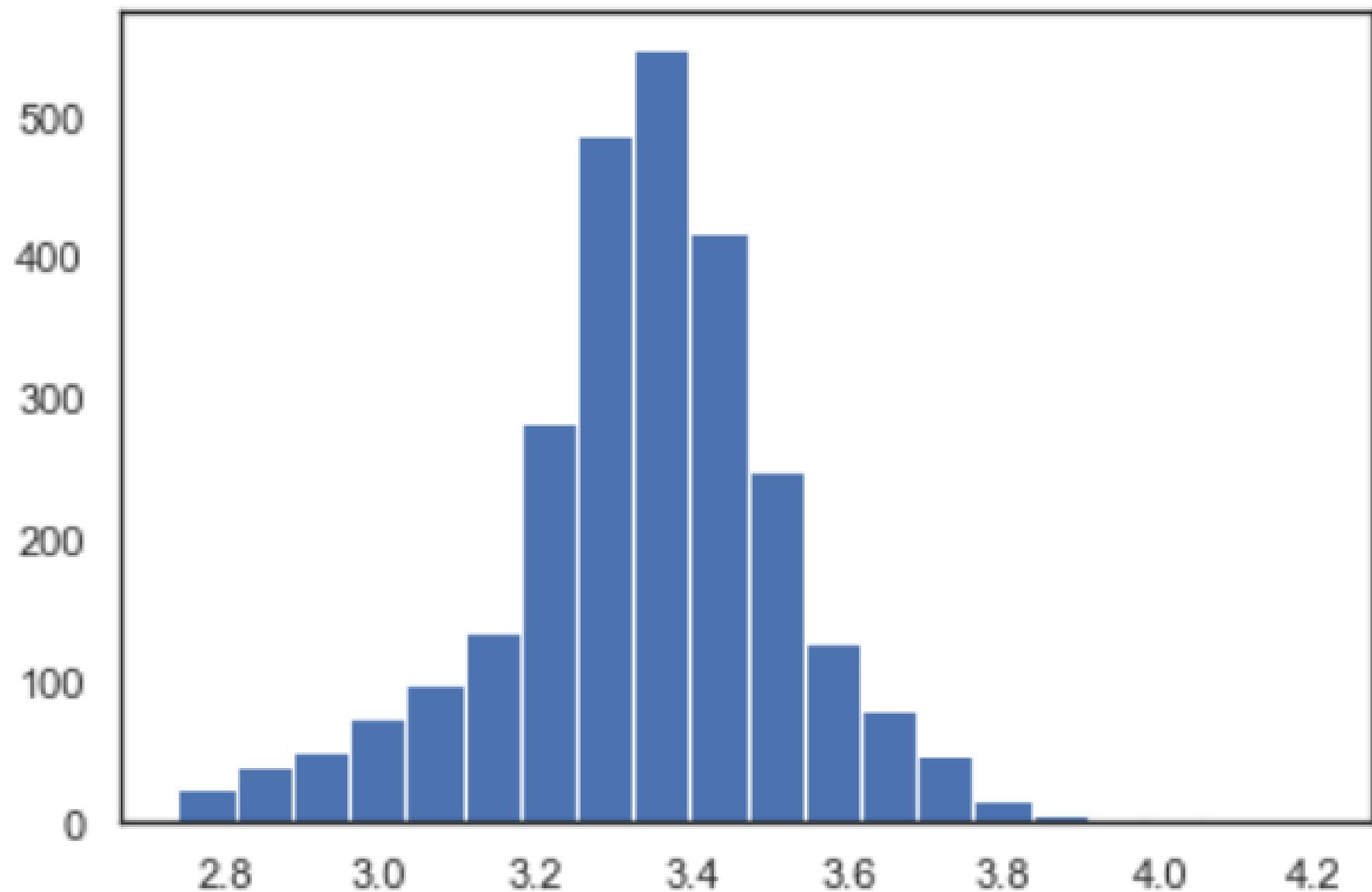
Cells are encapsulated into droplets with lysis buffer, reverse-transcription mix, and hydrogel microspheres carrying barcoded primers. After encapsulation primers are released. cDNA in each droplet is tagged with a barcode during reverse transcription. Droplets are then broken and material from all cells is linearly amplified before sequencing. UMI = unique molecular identifier.

# Ход работы

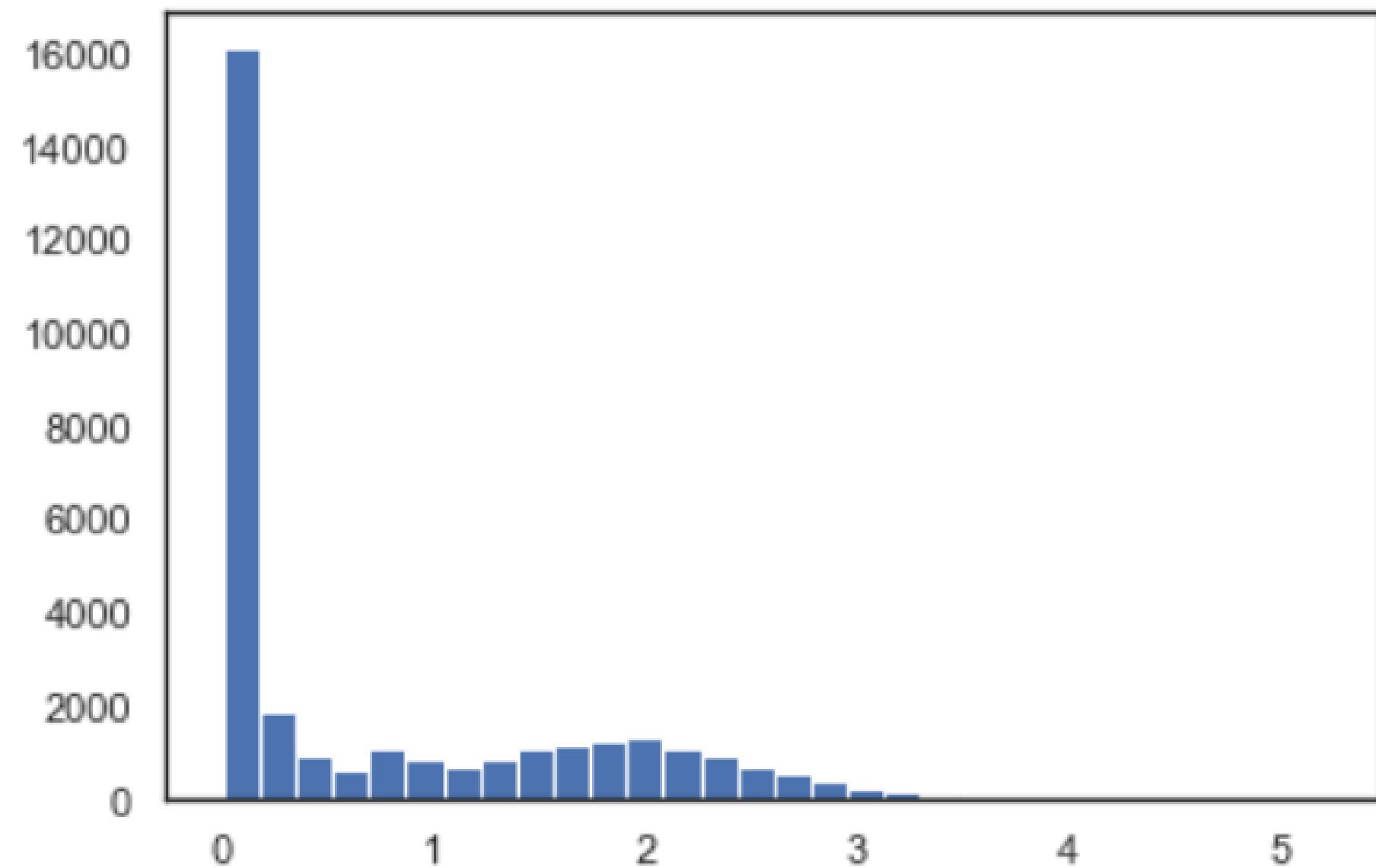
- подготовка
- нормализация данных
- выбор наиболее вариабельных генов
- стандартизация данных
- понижение размерности(PCA)
- кластеризация
- визуализация в двумерном пространство(UMAP)

# Визуализация данных

Hist cells

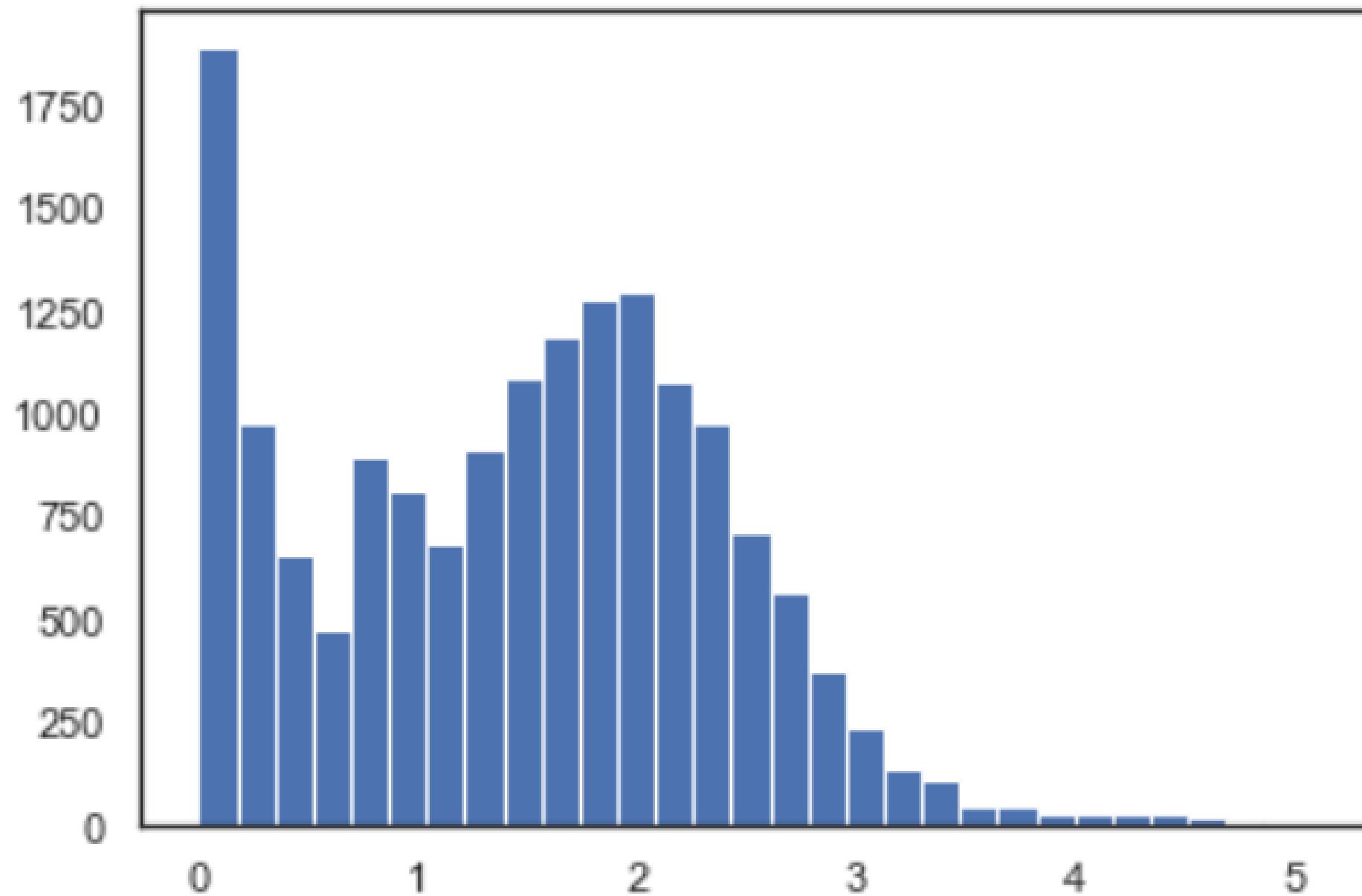


Hist genes



# Визуализация данных

Hist genes without zeros



Mean cells: 2366.9

Median cells: 2197.0

Std cells: 1094.0

Mean genes: 195.2

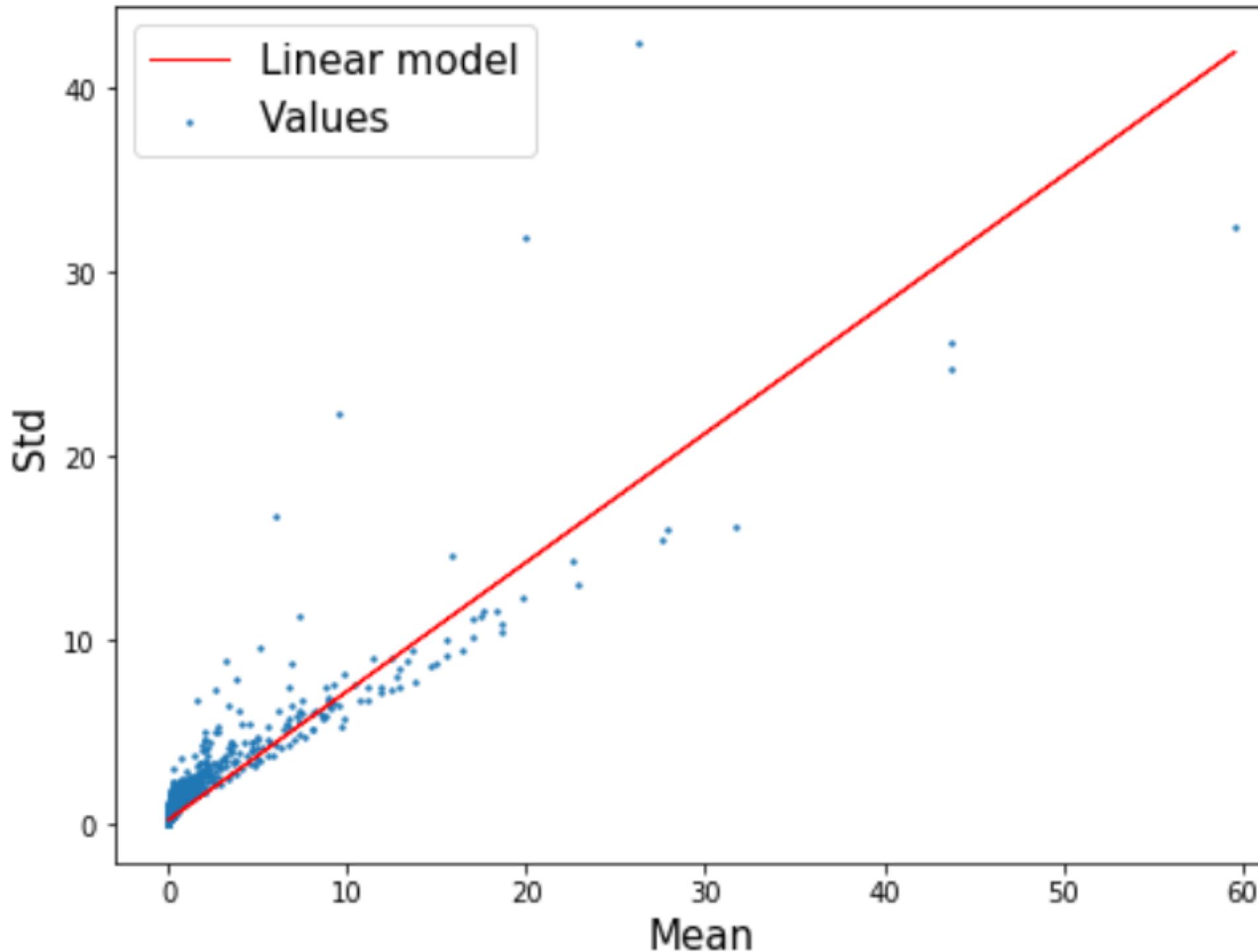
Median genes: 1.0

Std genes: 2301.4

# Подготовка

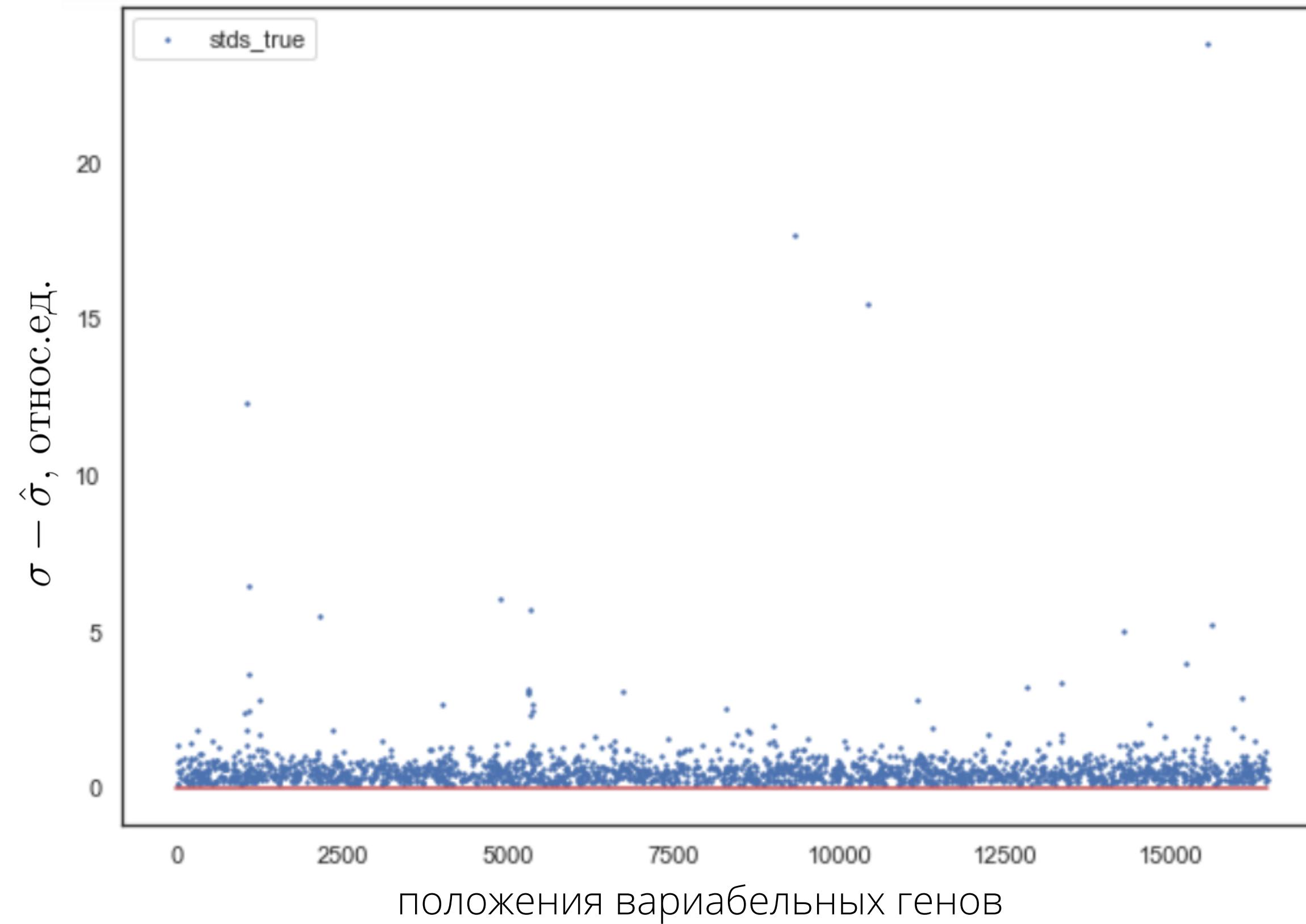
- удаление выбросов  
(клеток с экспрессией меньше 250 и больше 5000)
- удаление генов с суммарной экспрессией равной 0

# Предсказание ожидаемой дисперсии по среднему



- Все значения нормализуются по предсказанной дисперсии
- Высчитываются дисперсии нормализованных значений
- Отбираются 2000 генов с наибольшей дисперсией (наиболее вариабельные)

# Отбор наиболее вариабельных признаков(генов)



# Масштабирование(StandardScaler)

$$\tilde{x} = \frac{x - \text{loc}}{\text{scale}}$$

- Все признаки приводятся к одному масштабу:  
среднее =0  
дисперсия =1

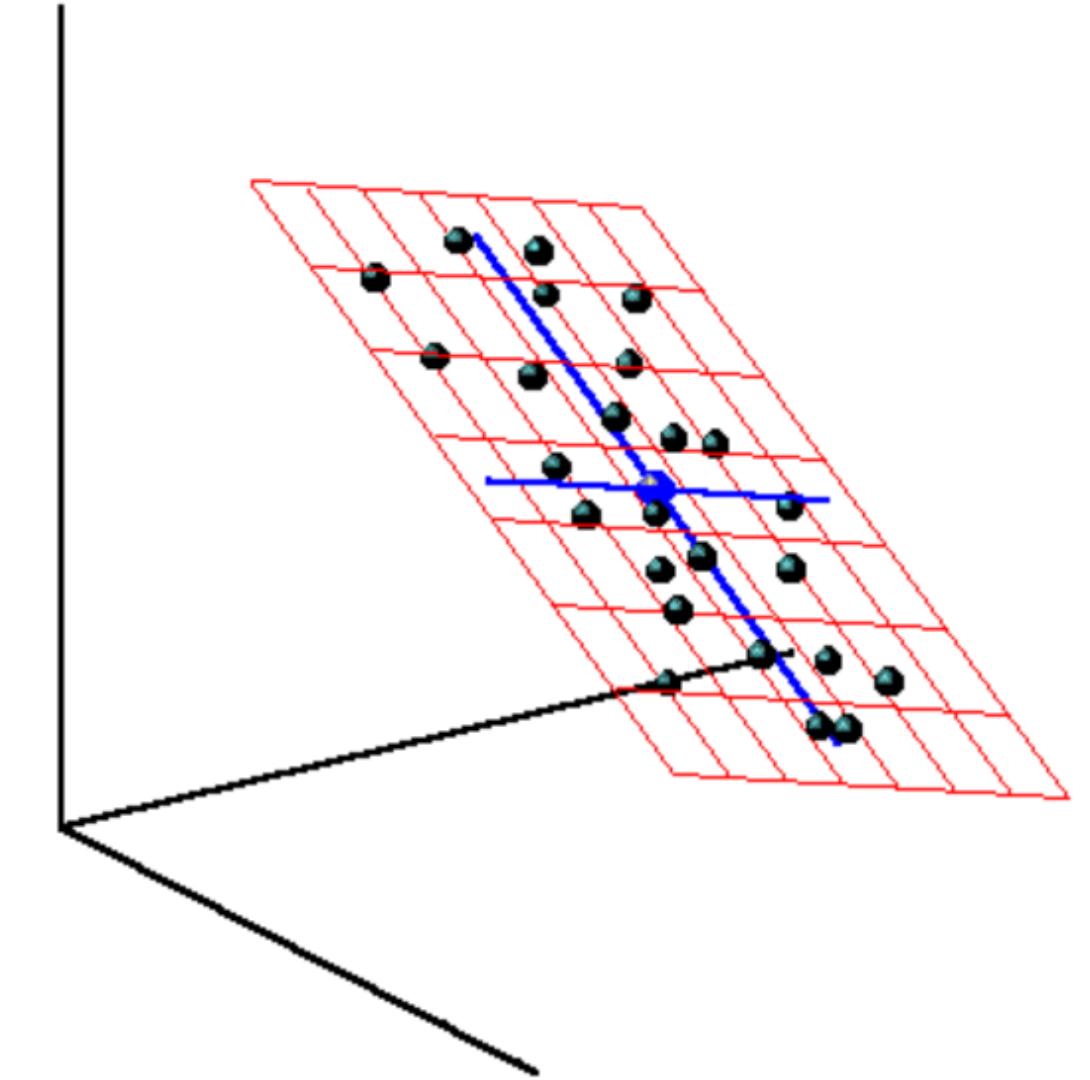
# Понижение размерности

## PCA - метод выделения главных компонент

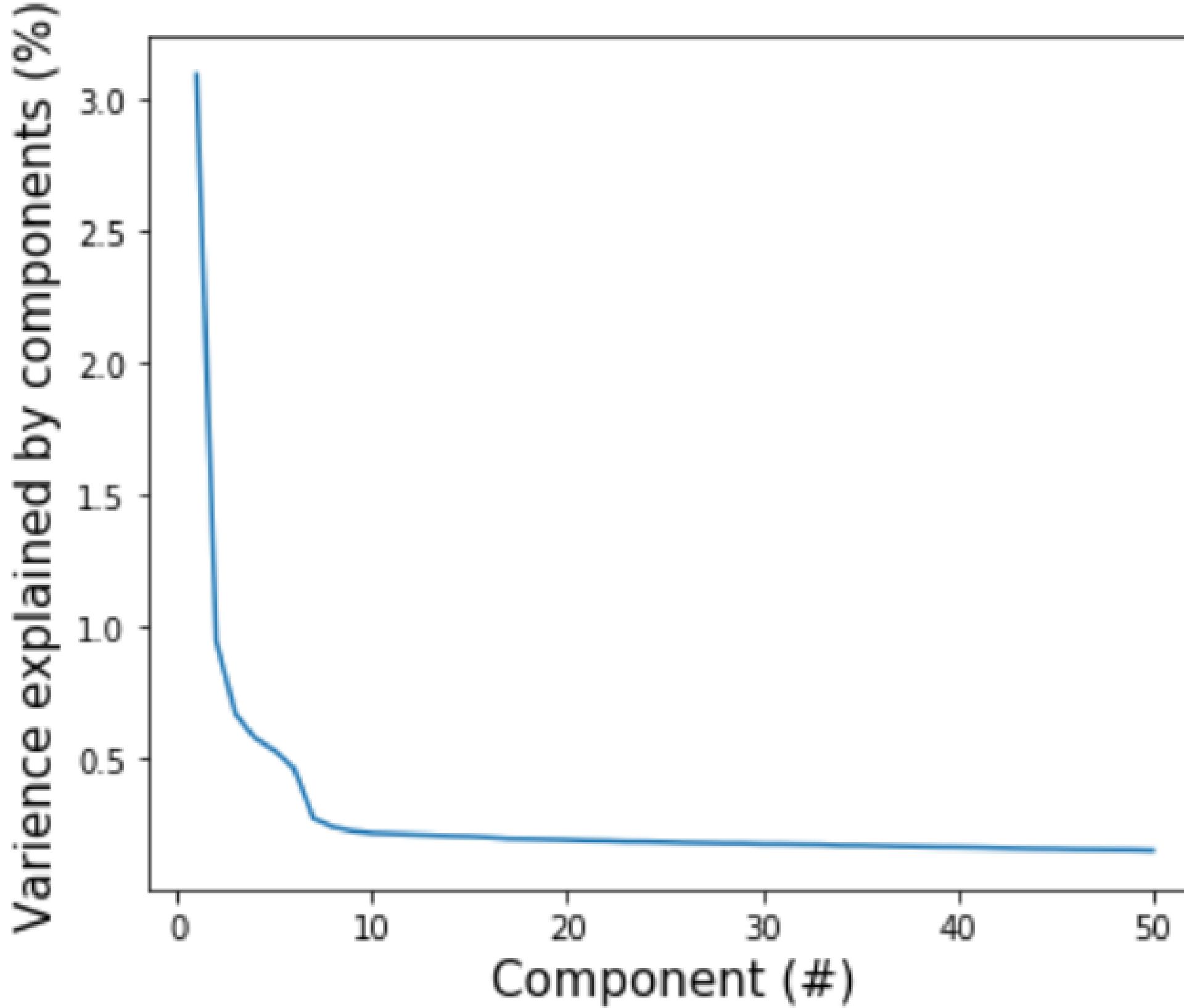
**Цель:** понизить размерность, но сохранить глобальную структуру данных

- PCA находит линейно независимые комбинации признаков, которые могут практически без потерь информации представлять данные.
- Математически: переход в пространство из  $k$  собственных векторов матрицы ковариаций собственных признаков.

Дисперсия значений по этим направлениям максимальна



# Выбор главных компонент



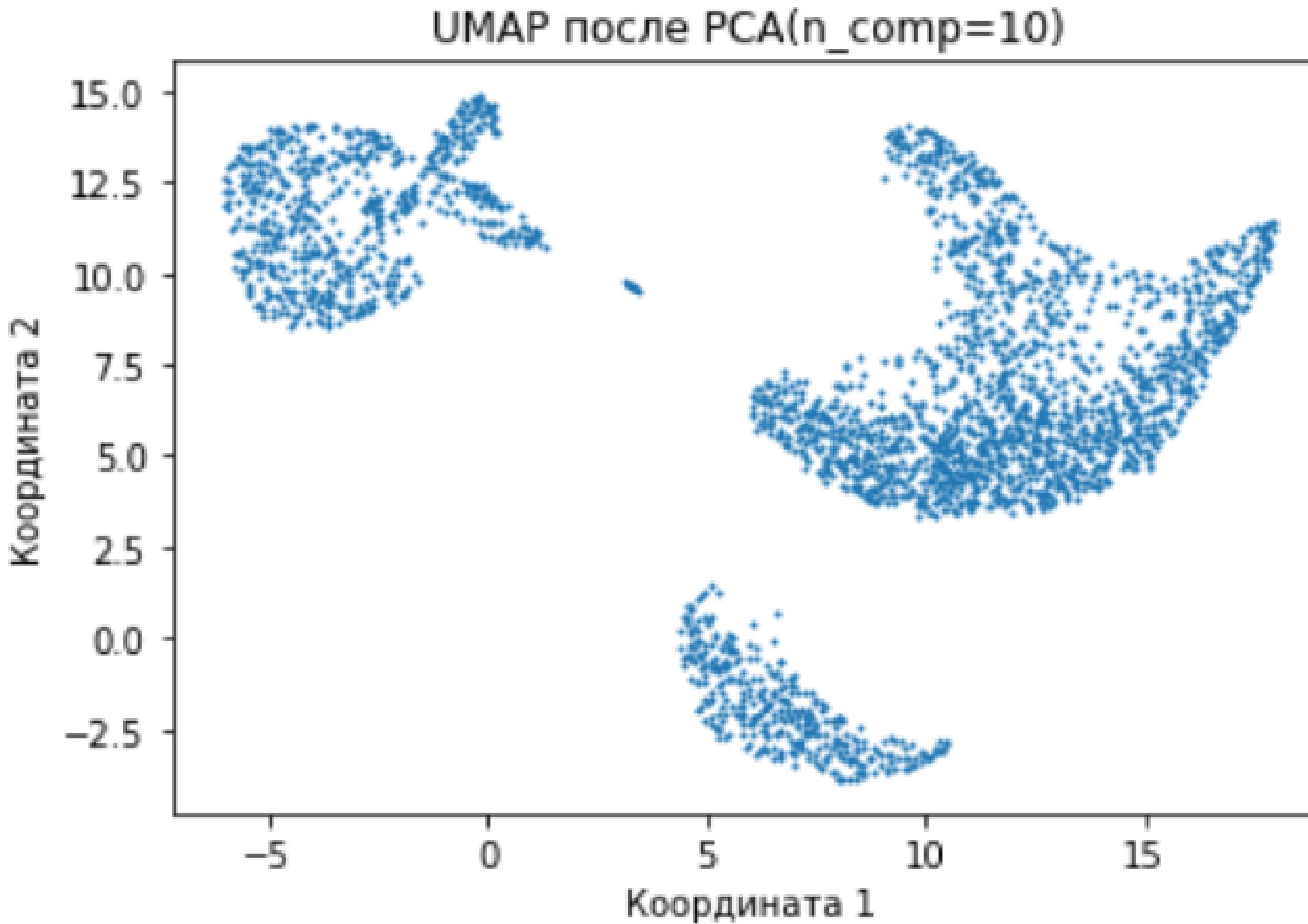
Отсекаем главные компоненты с малой дисперсией, выходящей на плато на графике

# УМАР

УМАР - нелинейное преобразование для уменьшения размерности

- Стремится сохранить расстояние между объектами: близкие объекты после проецирования остаются близкими, далекие - далекими. Подходит для визуализации
- У УМАР нет ограничений на размерность исходного пространства признаков, он достаточно быстрый и вычислительно эффективный

# Визуализация с UMAP



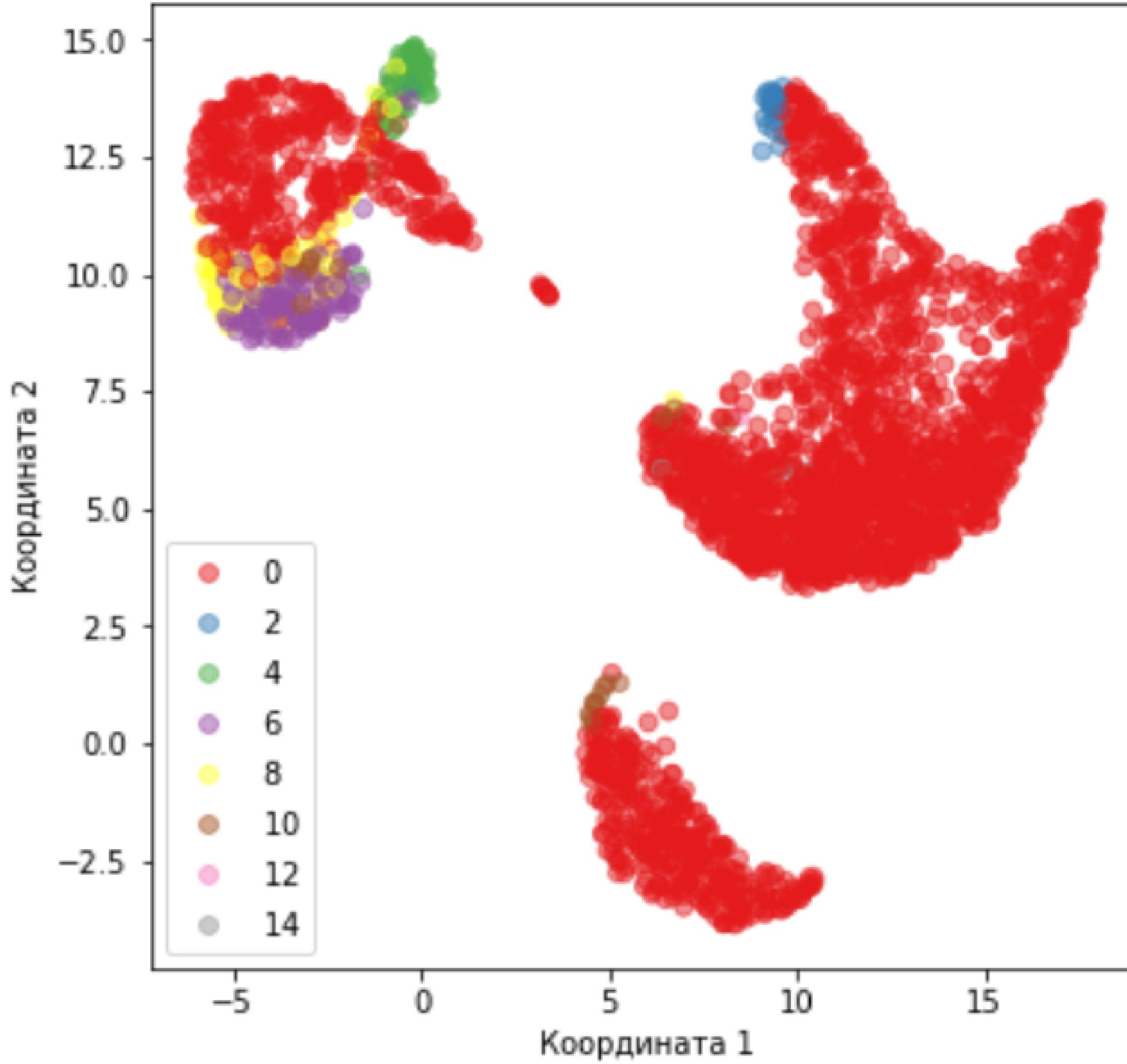
# Задача кластеризации

- задача обучения без учителя
- задача поставлена некорректно
- результат заранее не известен

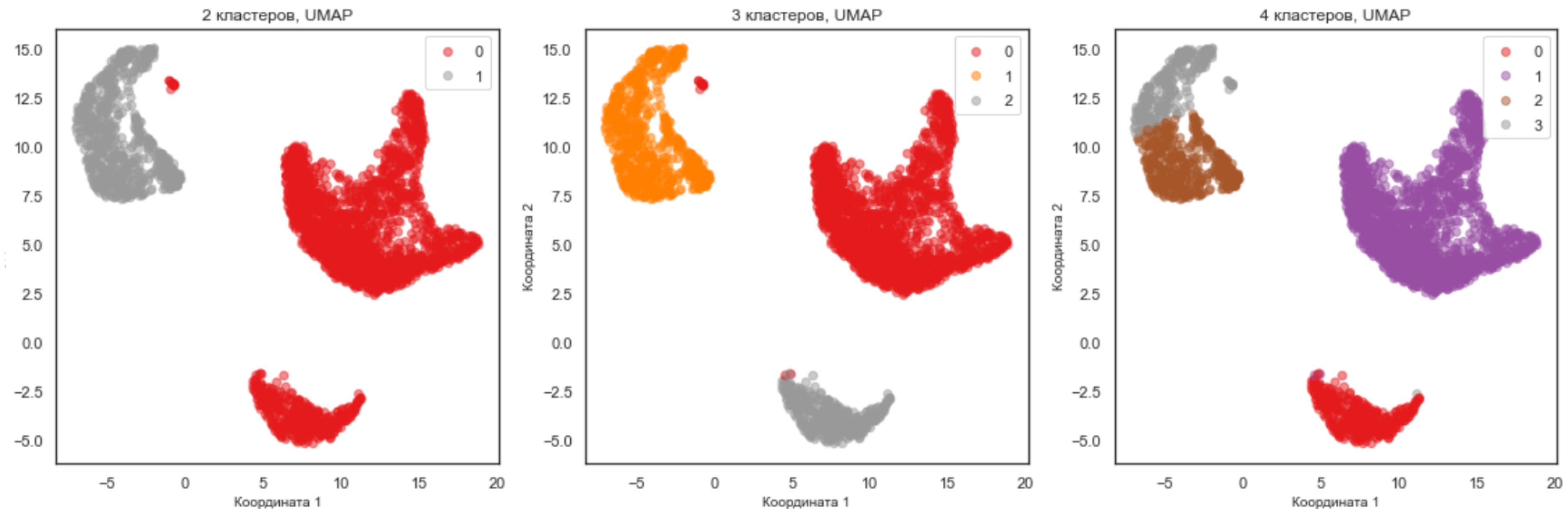
Метод	Параметры	Масштабируемость	Use-case	Метрика
k-means	Число кластеров	Очень много объектов; среднее число кластеров	Выпуклые; примерно одинаковые кластеры	Евклидово расстояние
Spectral Clustering	Число кластеров	Среднее число объектов, малое число кластеров	Несколько кластеров, равномерный размер кластера	Граф расстояний
DBSCAN	Радиус окрестности, $\varepsilon$ ; число соседей, $m$	Много объектов; среднее число кластеров	Неравные; невыпуклые кластеры; выбросы	Евклидово расстояние
Agglomerative Clustering	Число кластеров; тип расстояния между кластерами; расстояние	Много объектов и много кластеров	Лучше, когда много кластеров, нужно задать метрику	Любая метрика
Ward	Количество кластеров или пороговое значение расстояния	Большое число объектов и кластеров	Множество кластеров	Евклидово расстояние

# Результаты кластеризации

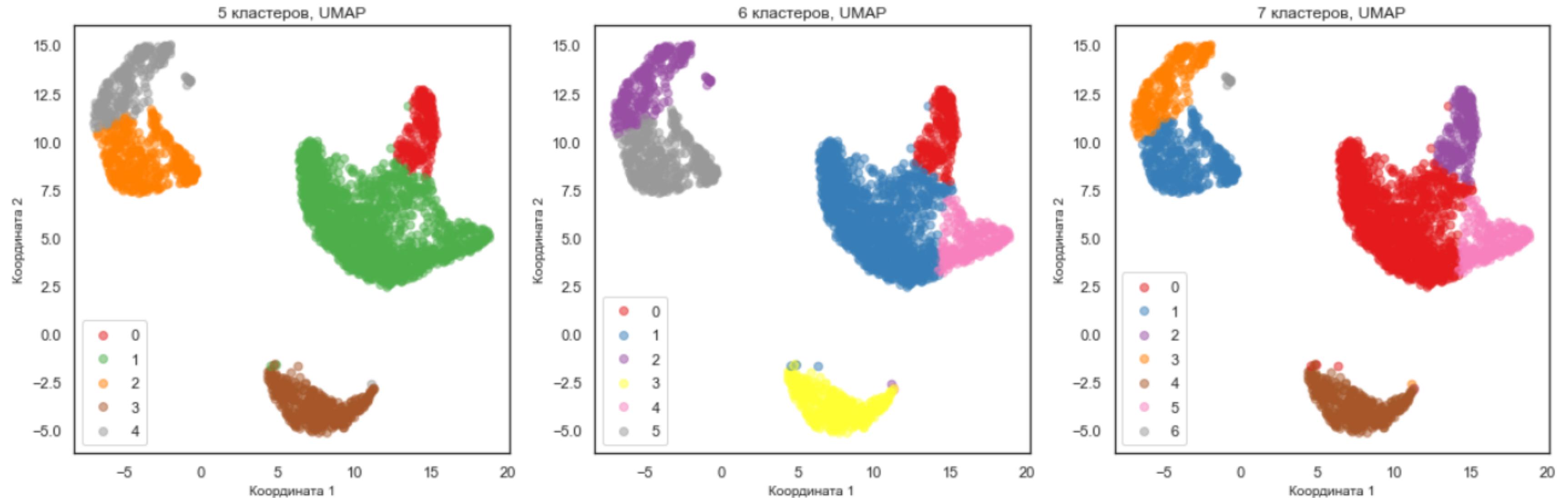
## Визуализация работы MeanShift с помощью UMAP



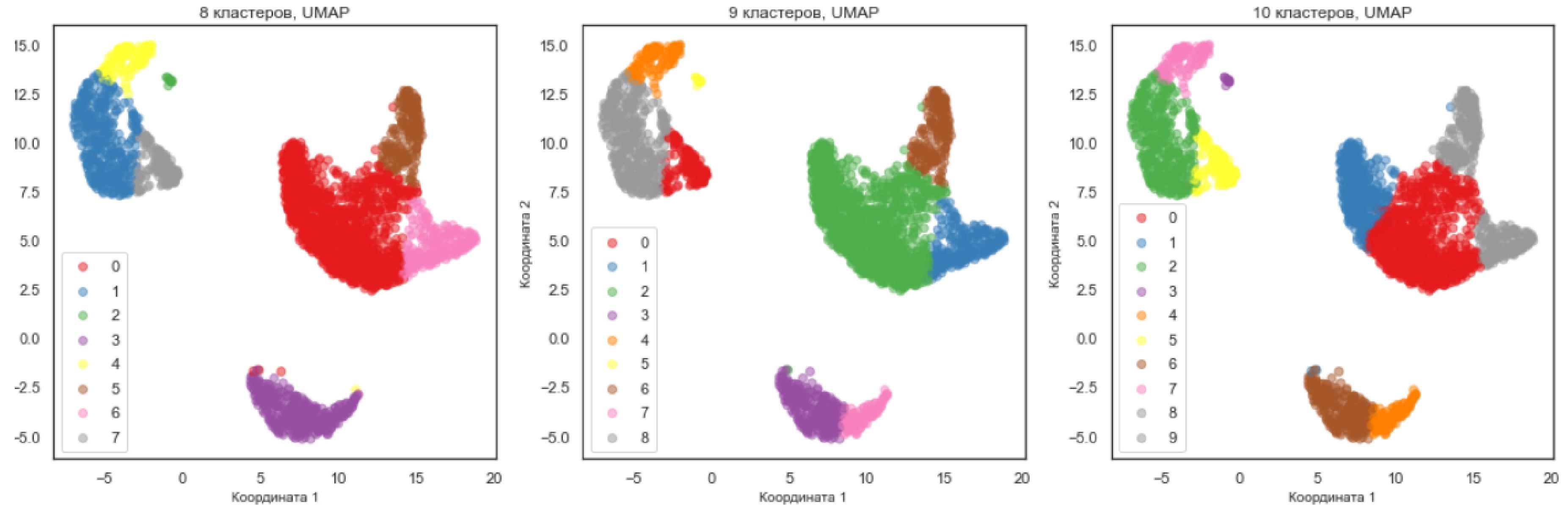
# SpectralClustering



# SpectralClustering

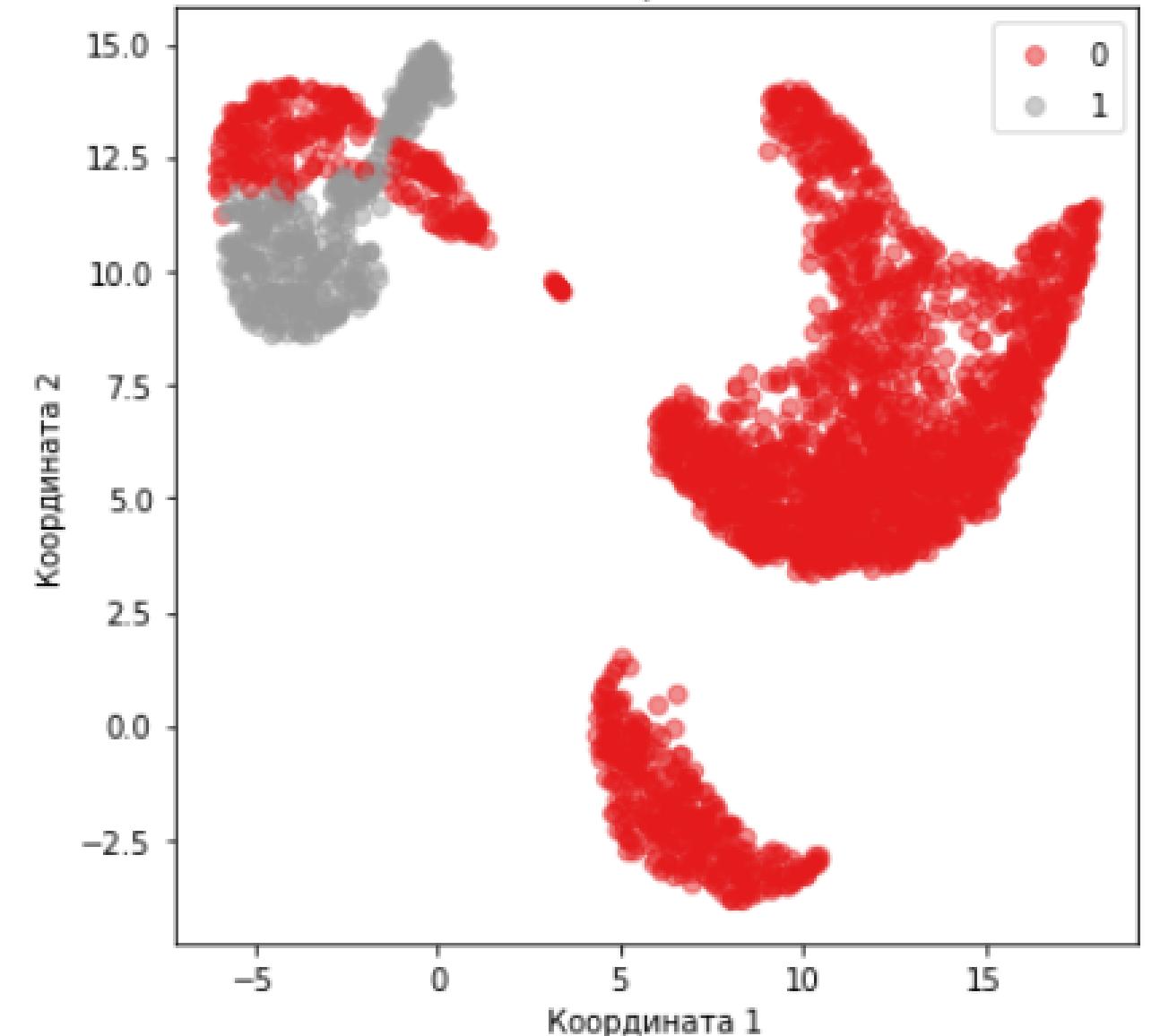


# SpectralClustering

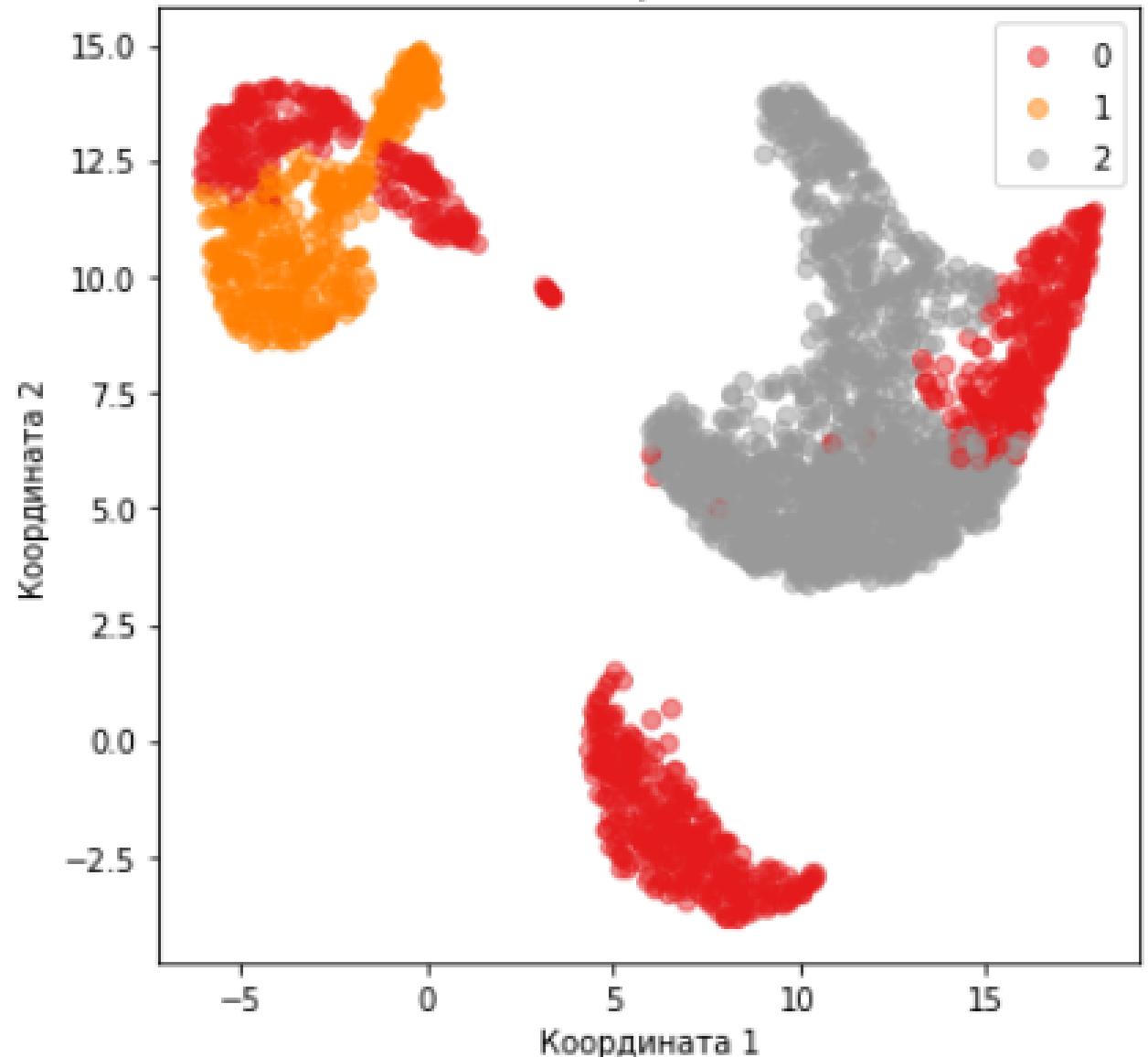


# Ward

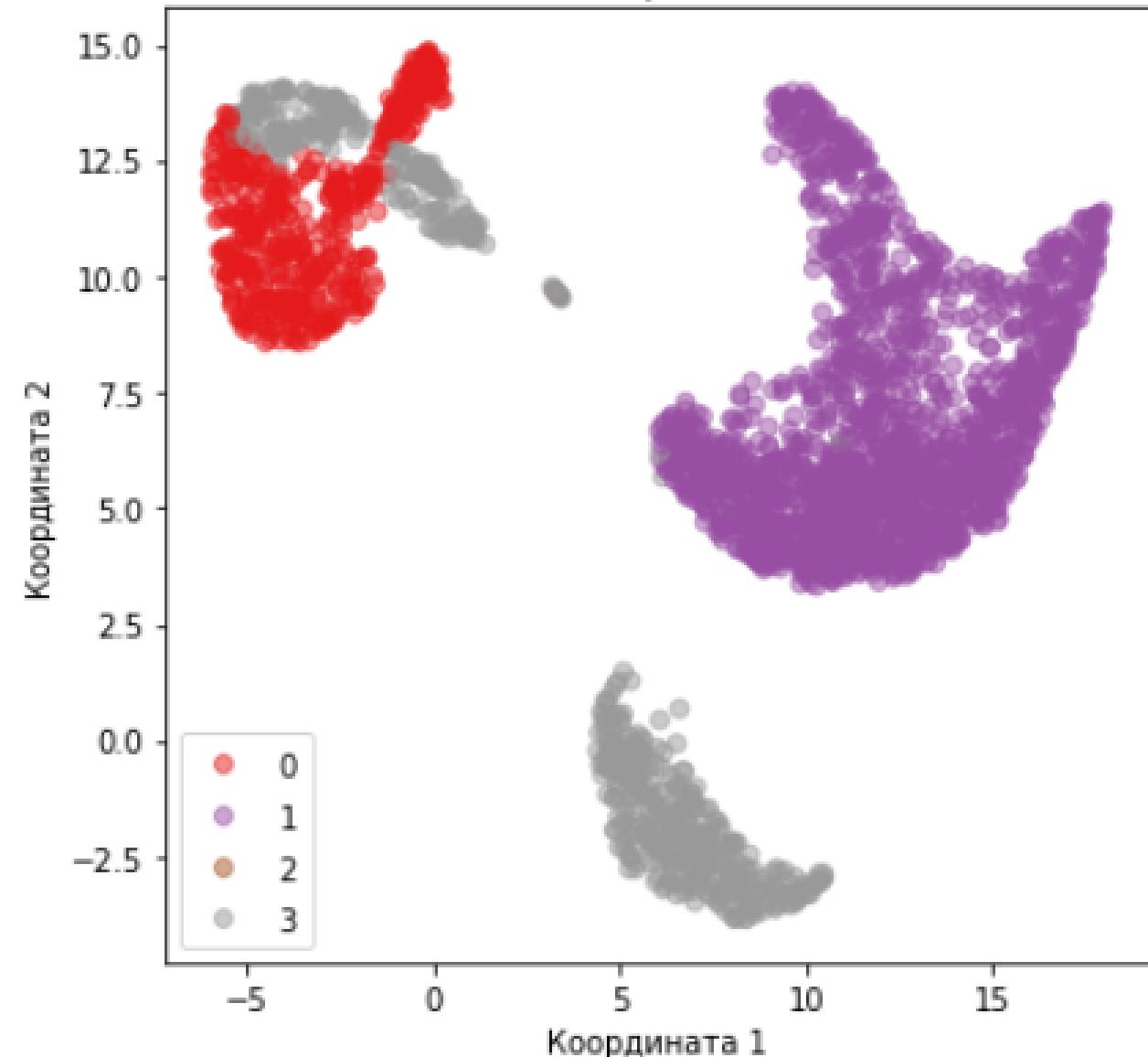
2 кластеров, UMAP



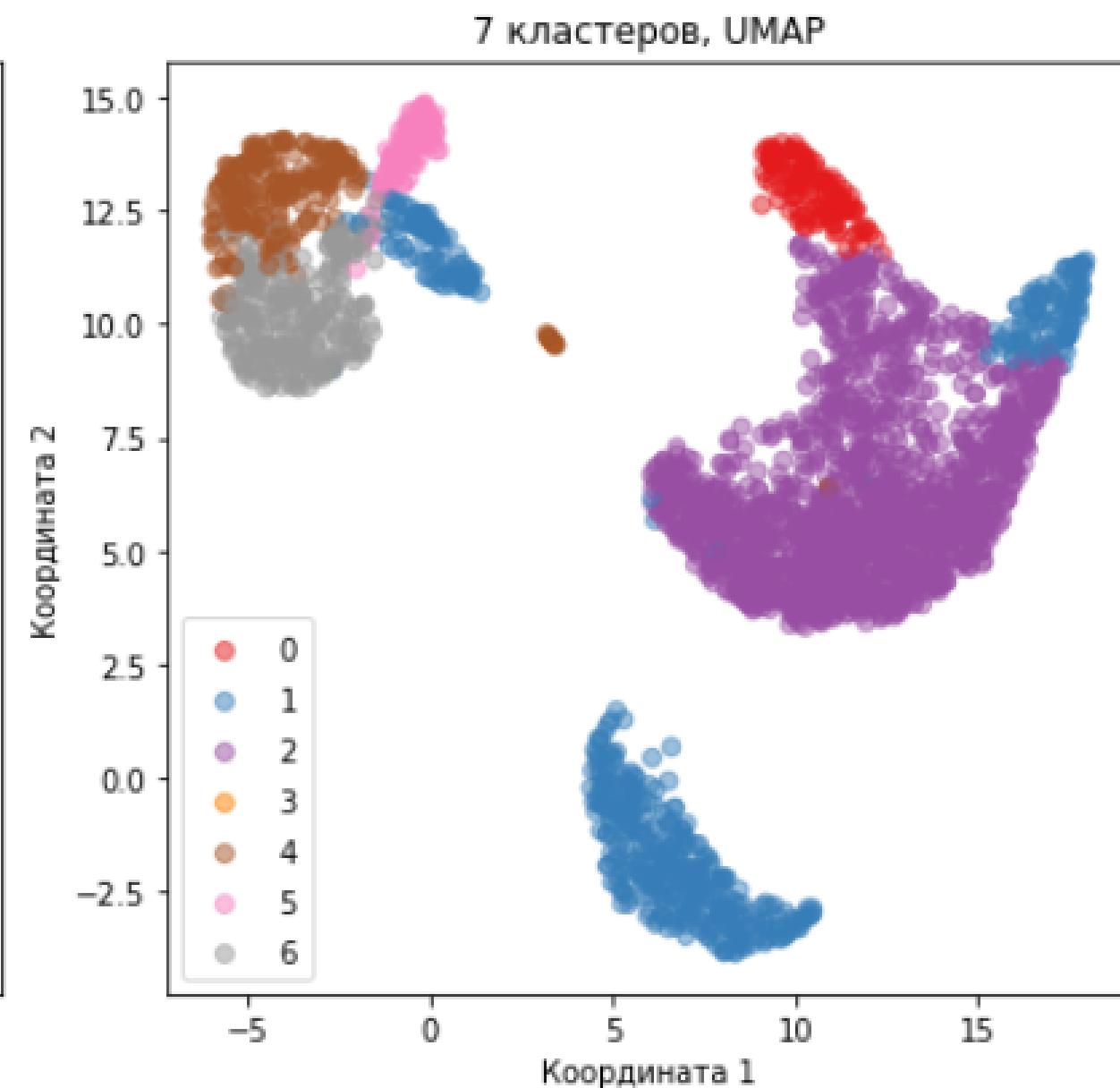
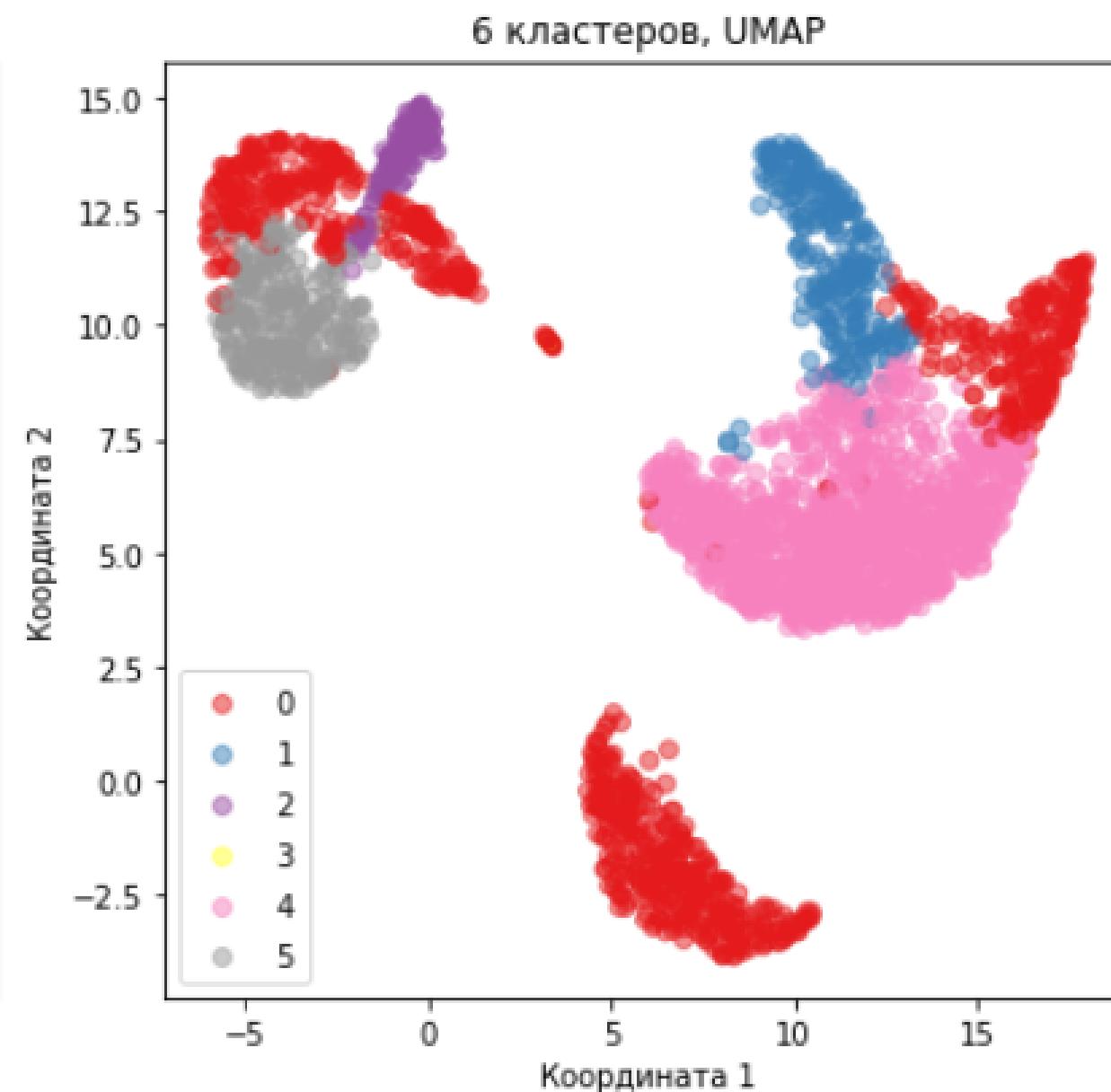
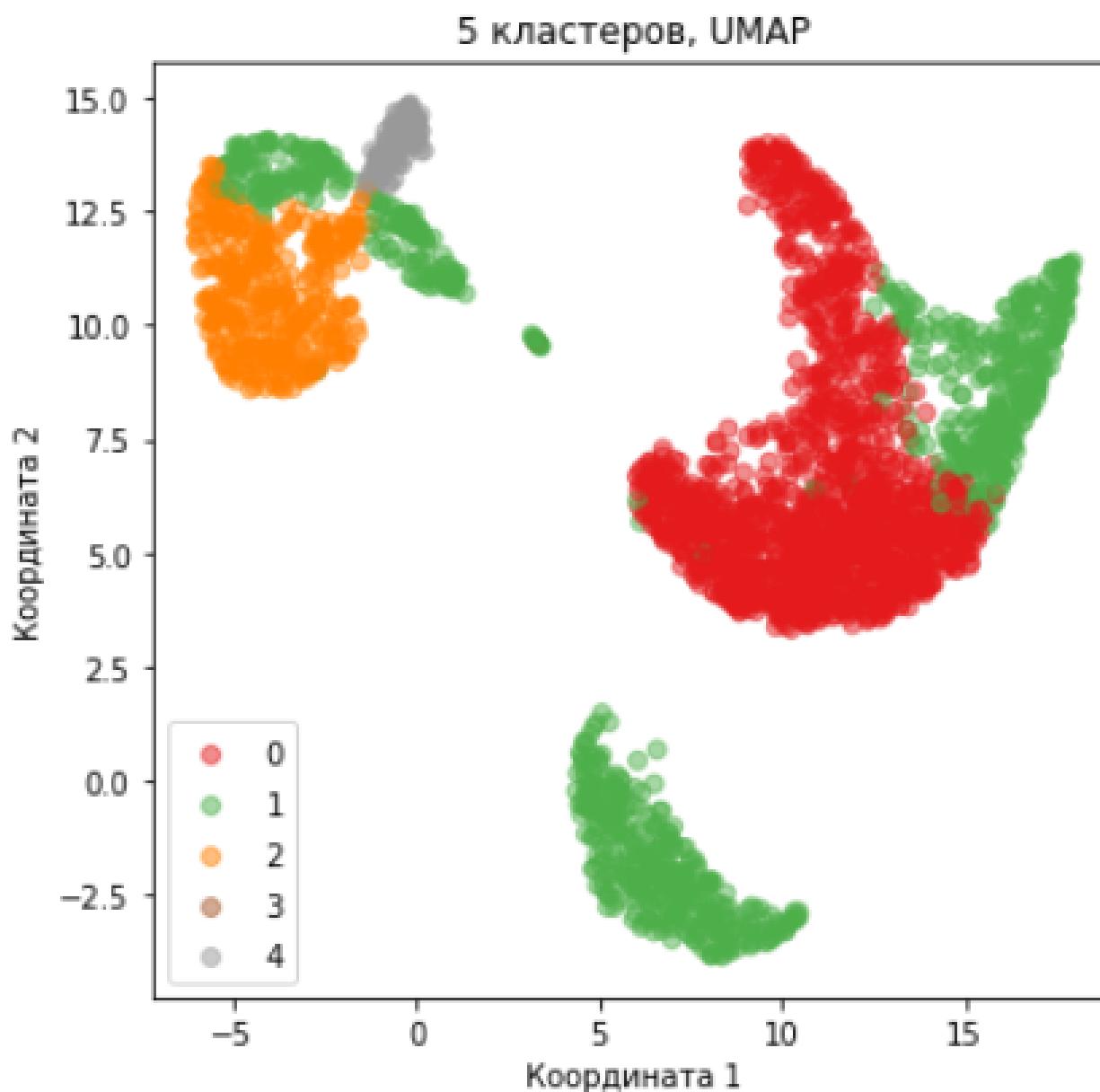
3 кластеров, UMAP



4 кластеров, UMAP

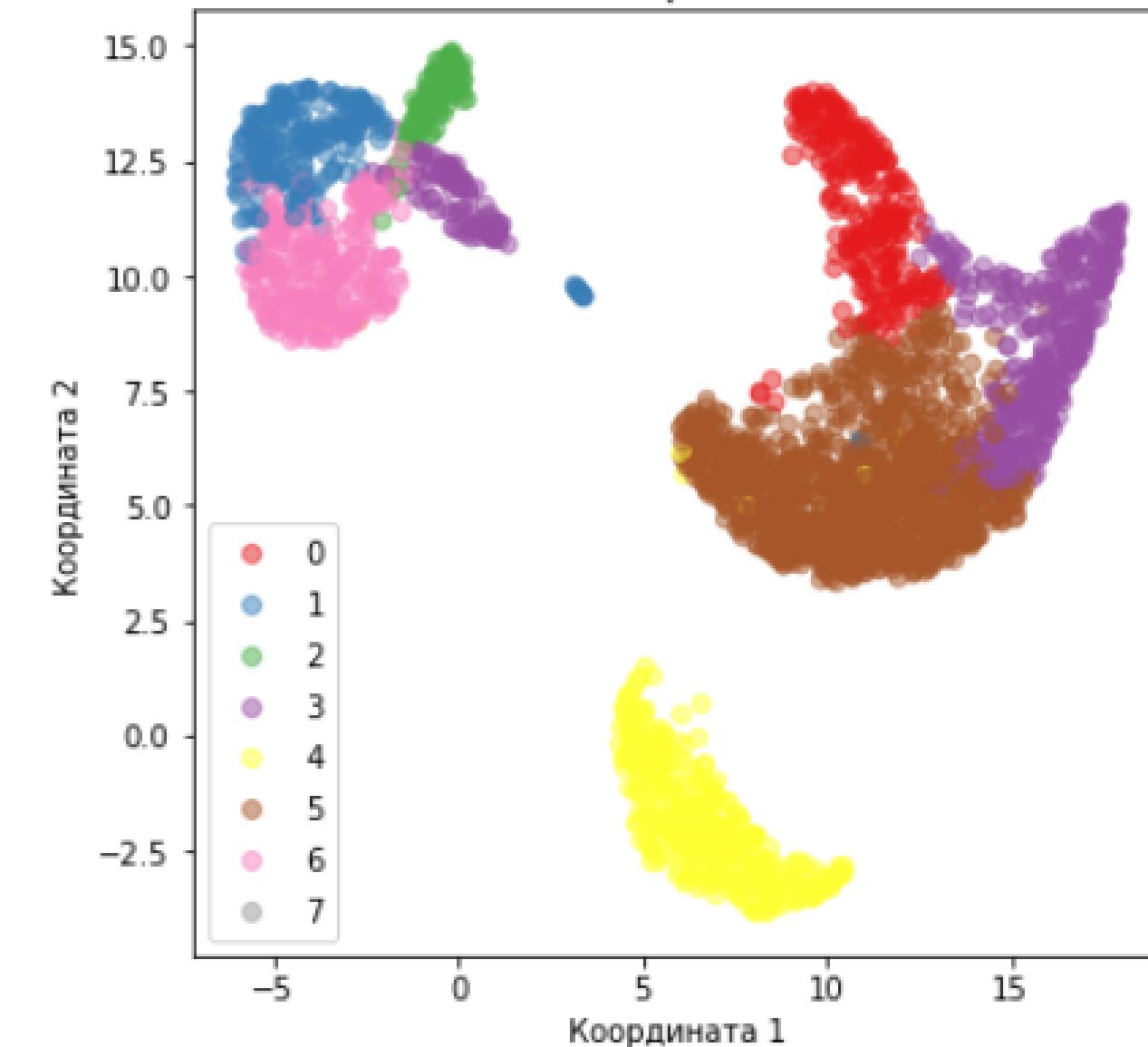


# Ward

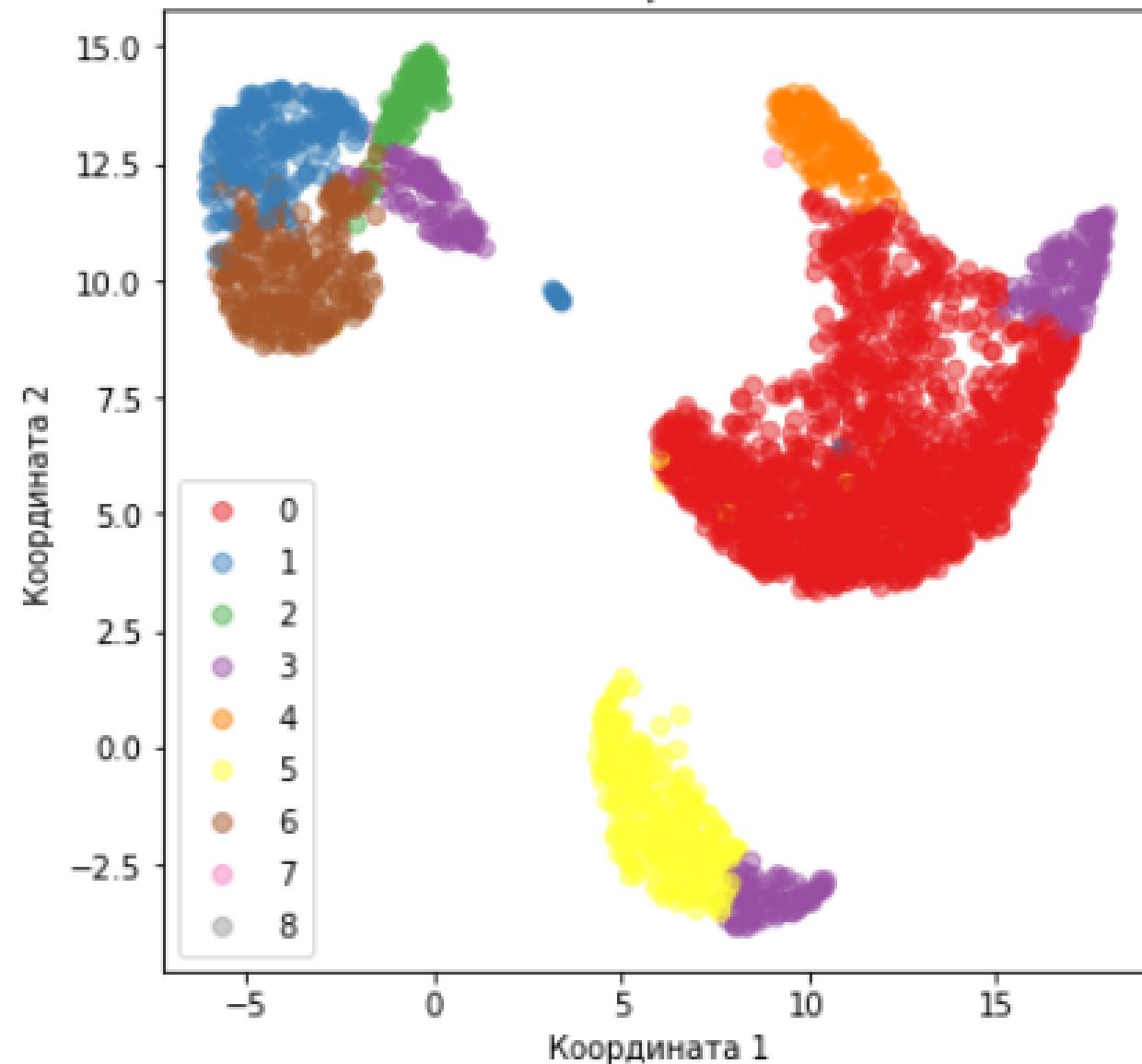


# Ward

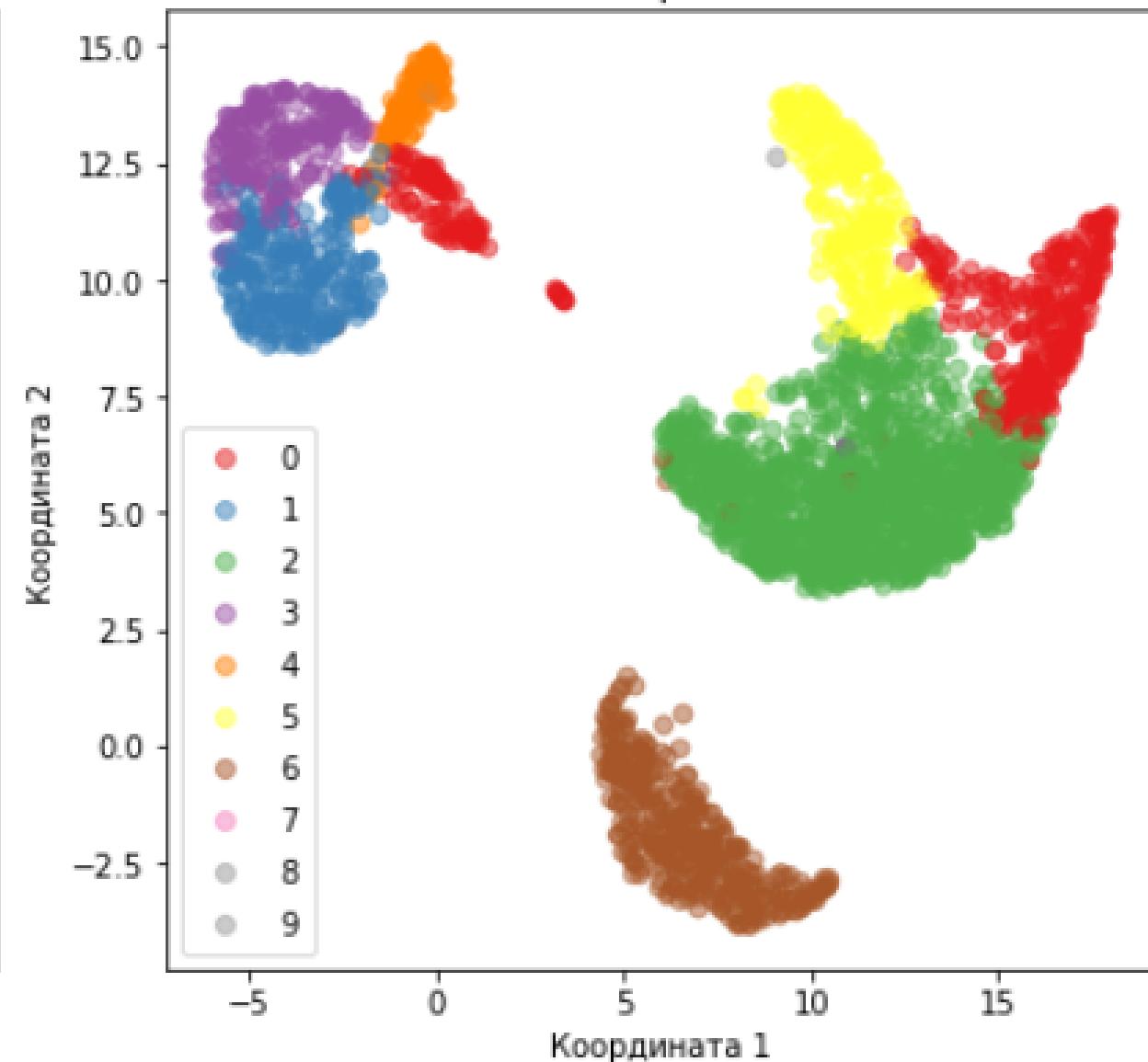
8 кластеров, UMAP



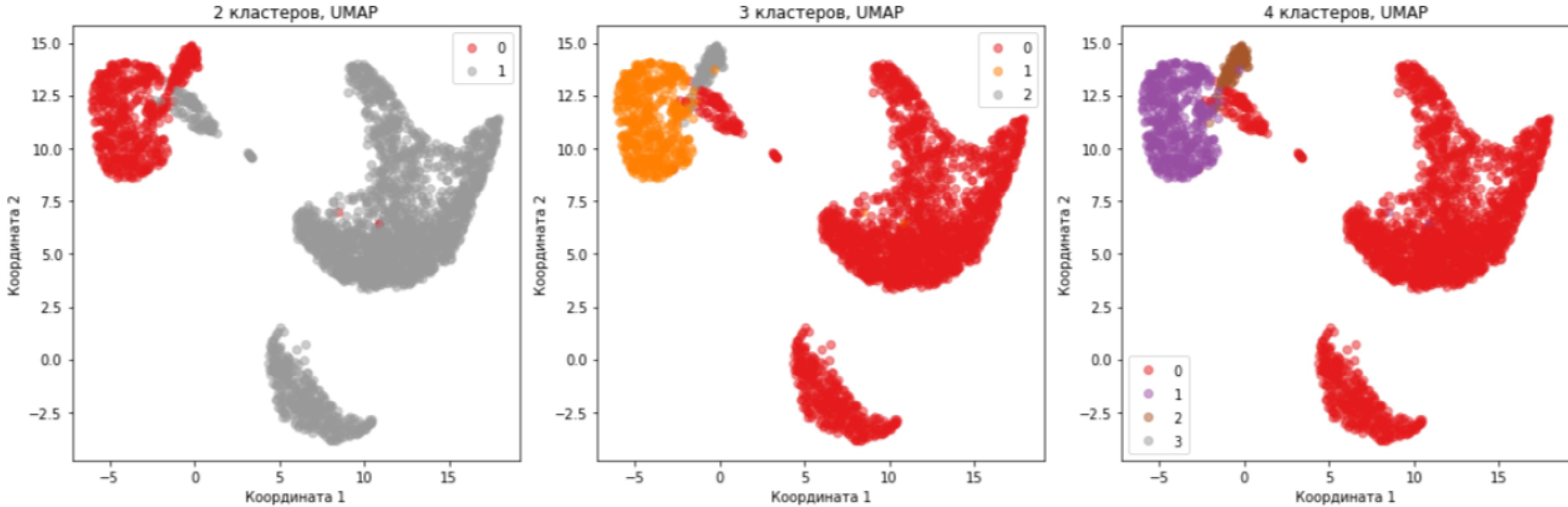
9 кластеров, UMAP



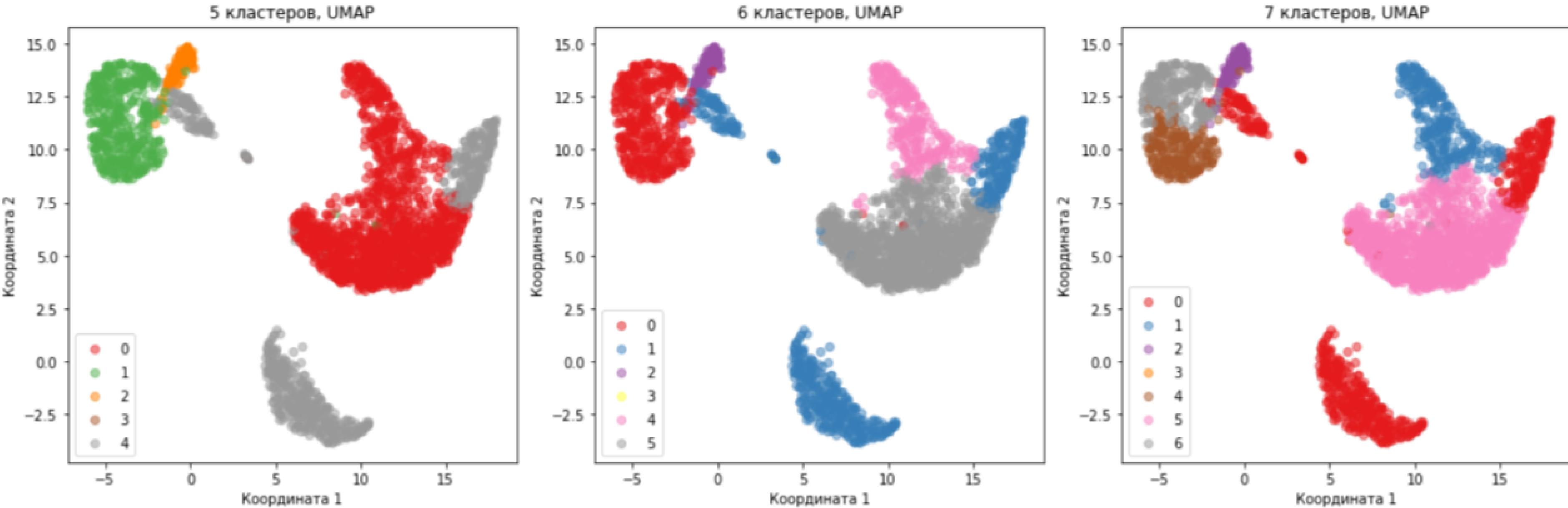
10 кластеров, UMAP



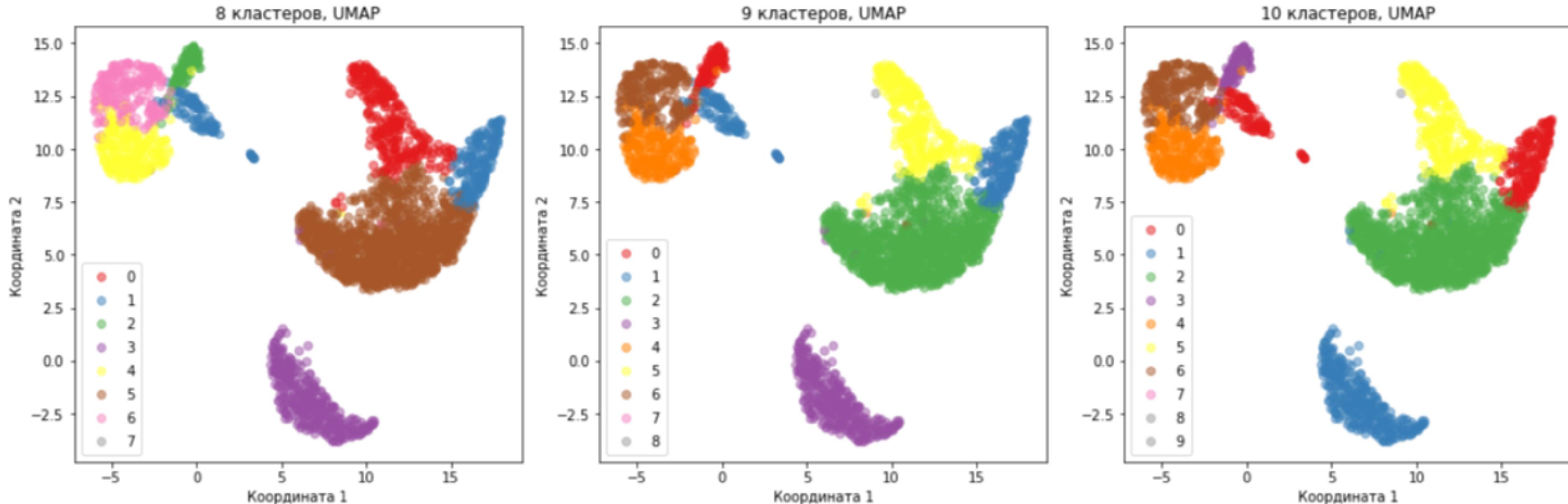
# AgglomerativeClustering



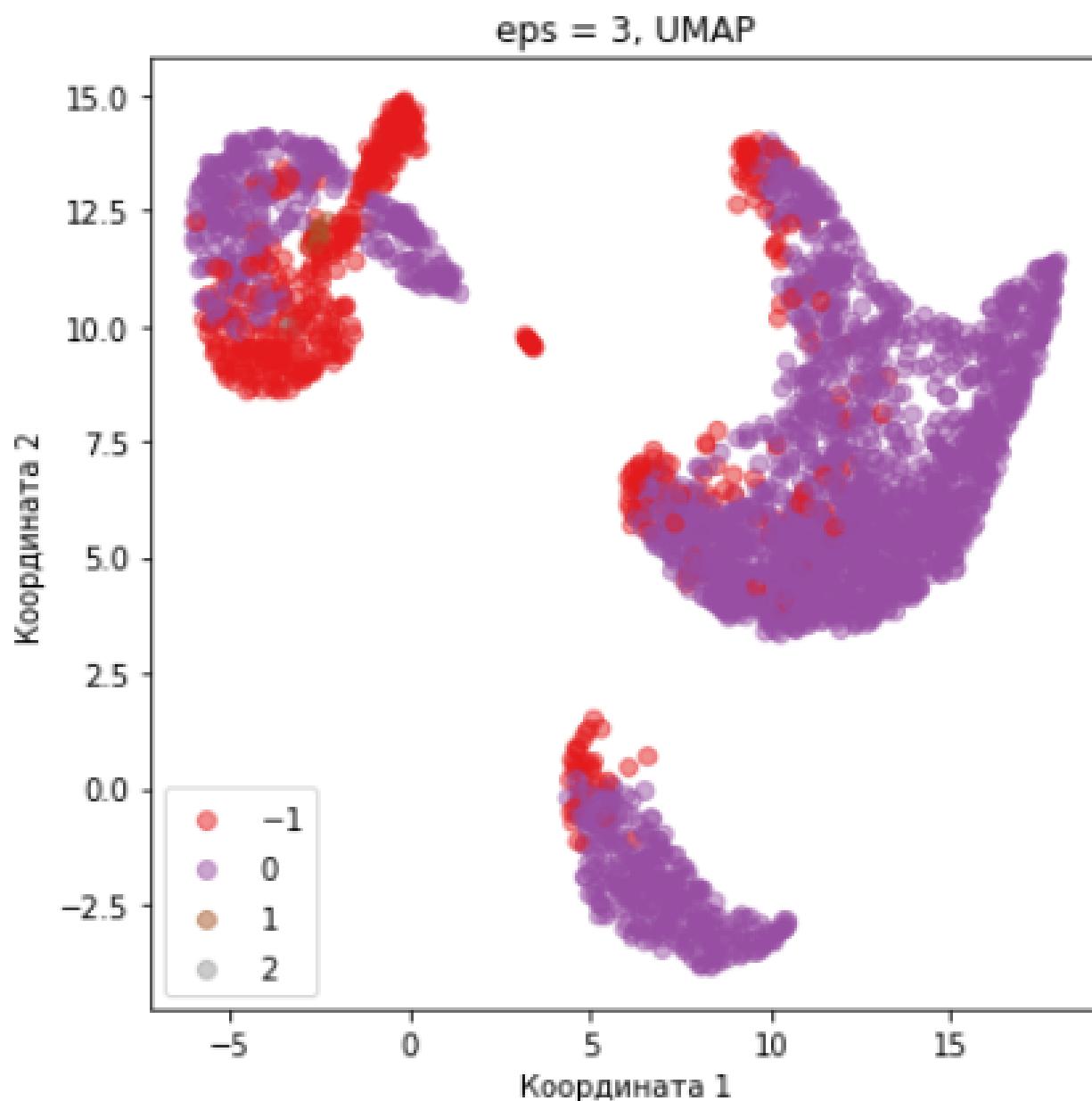
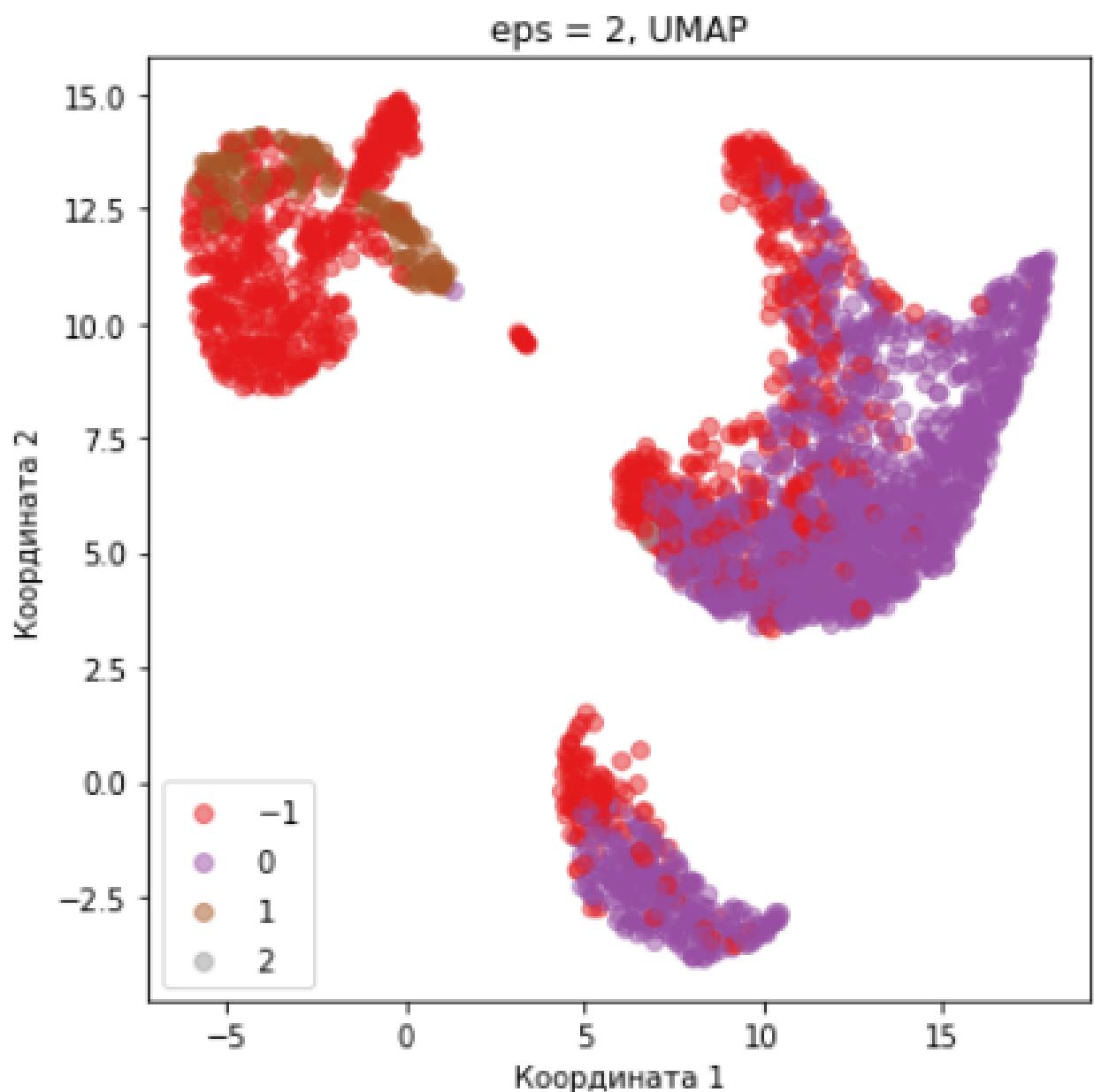
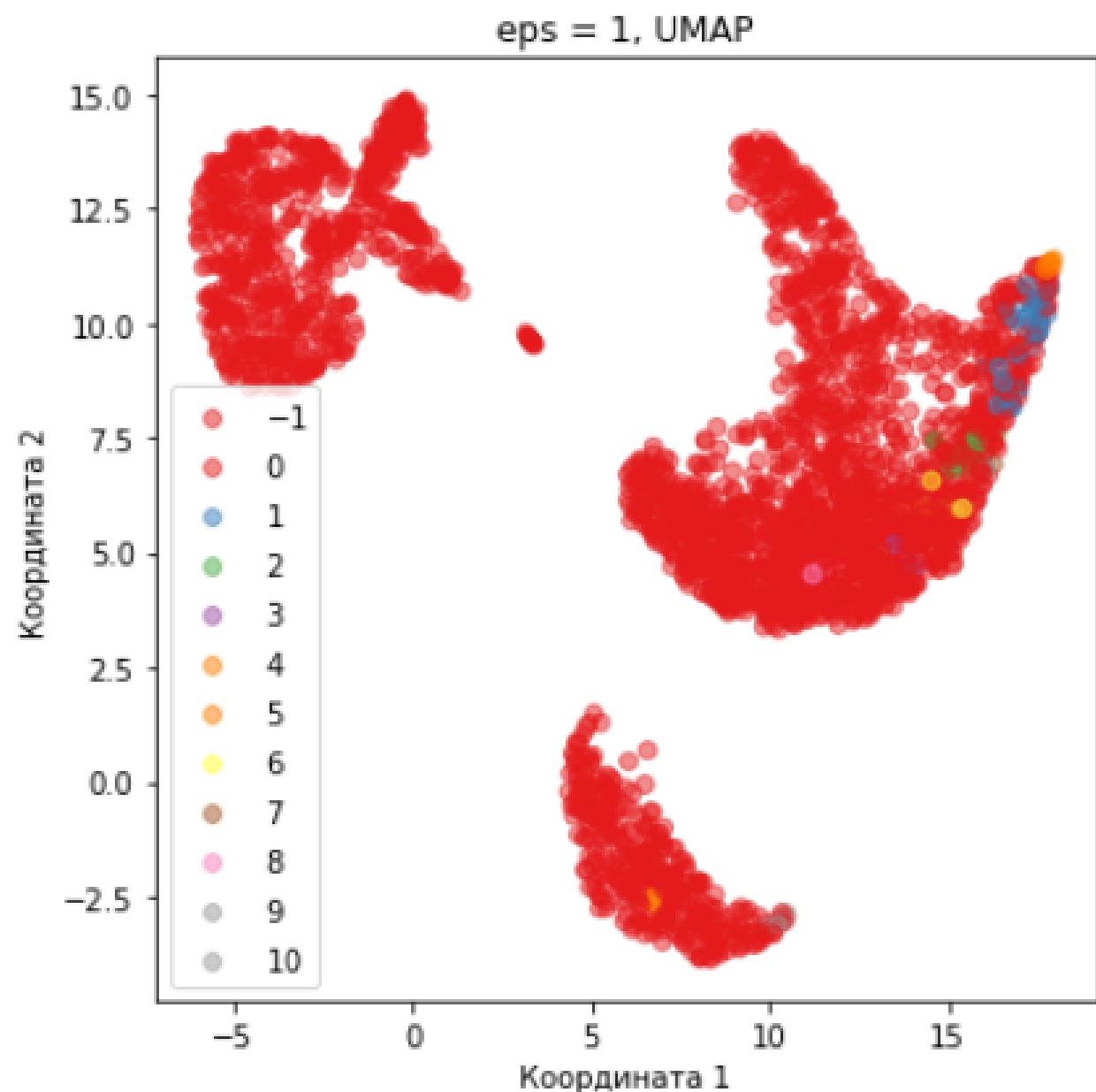
# AgglomerativeClustering



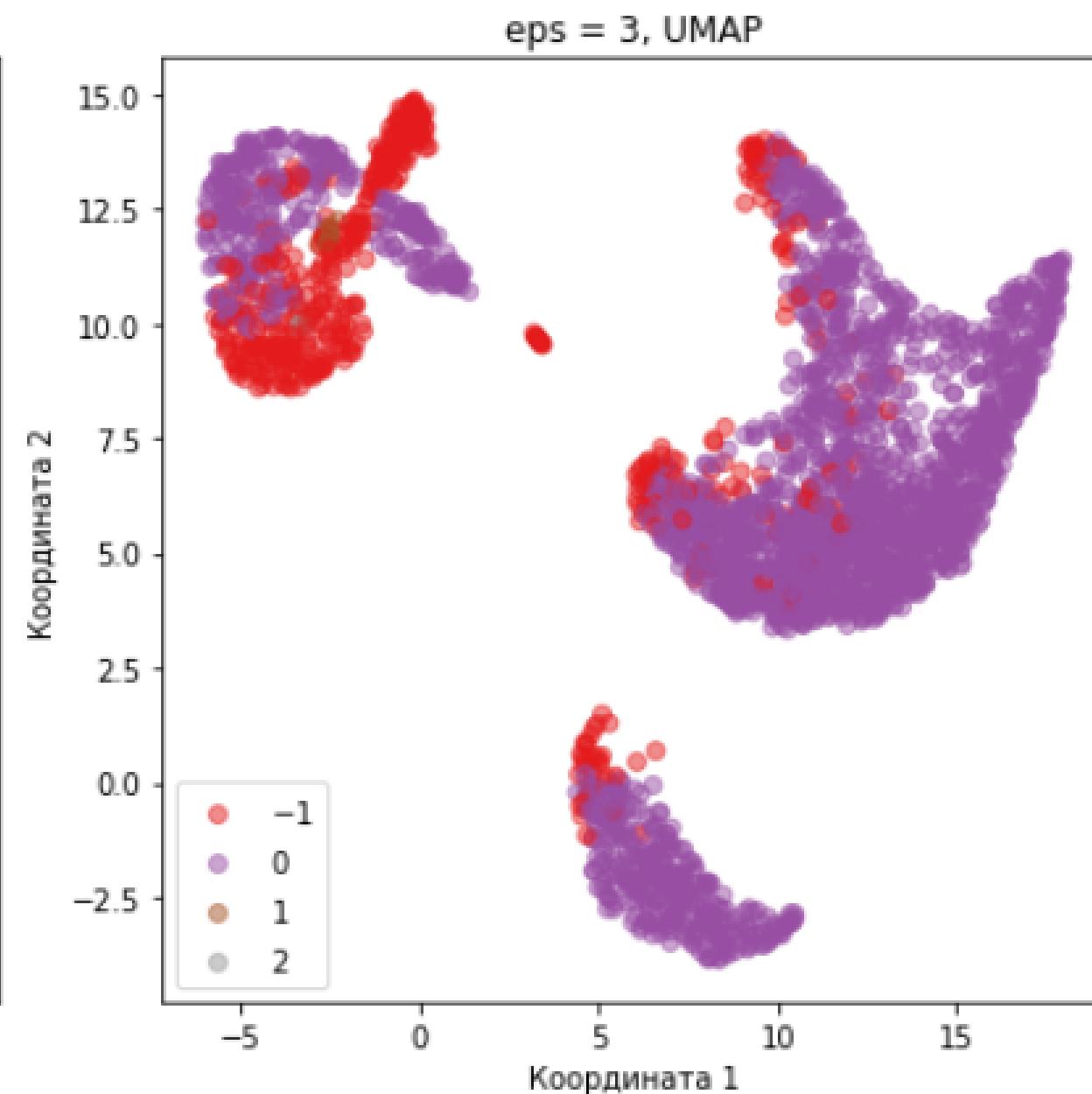
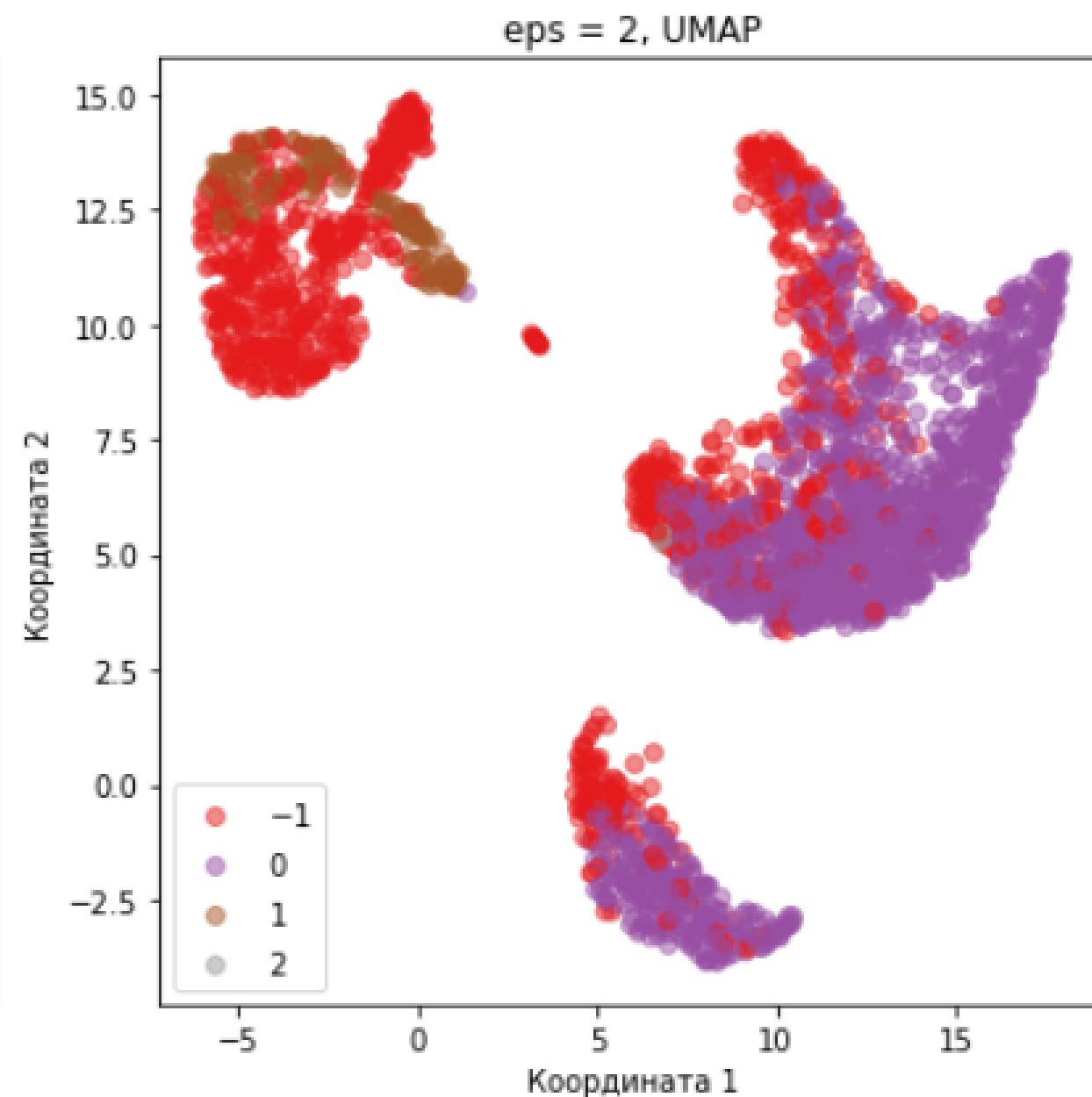
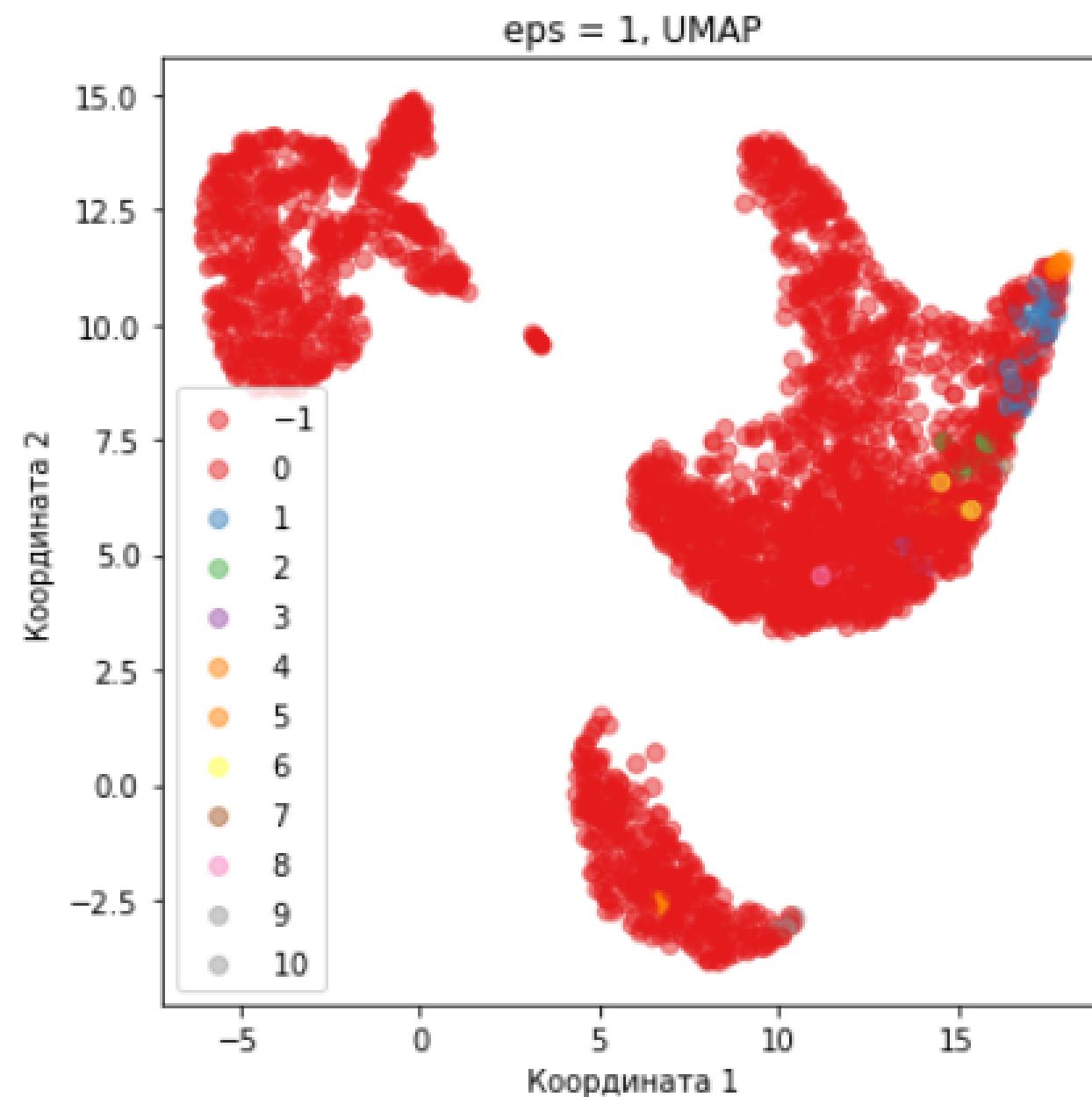
# AgglomerativeClustering



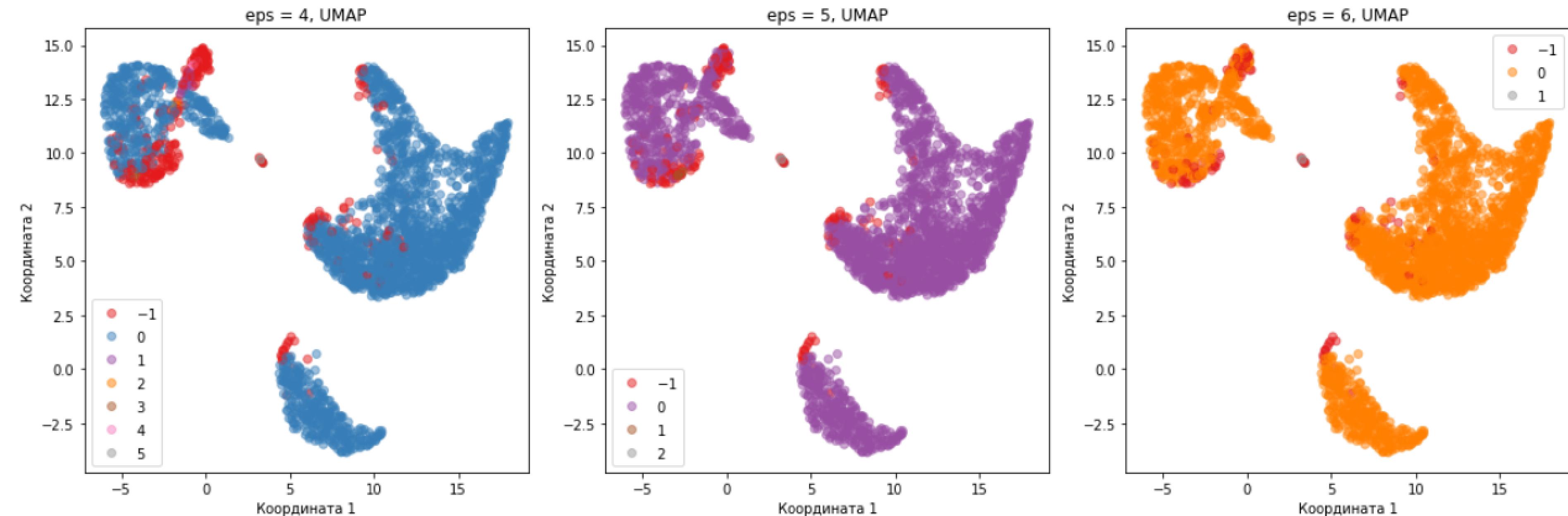
# DBSCAN



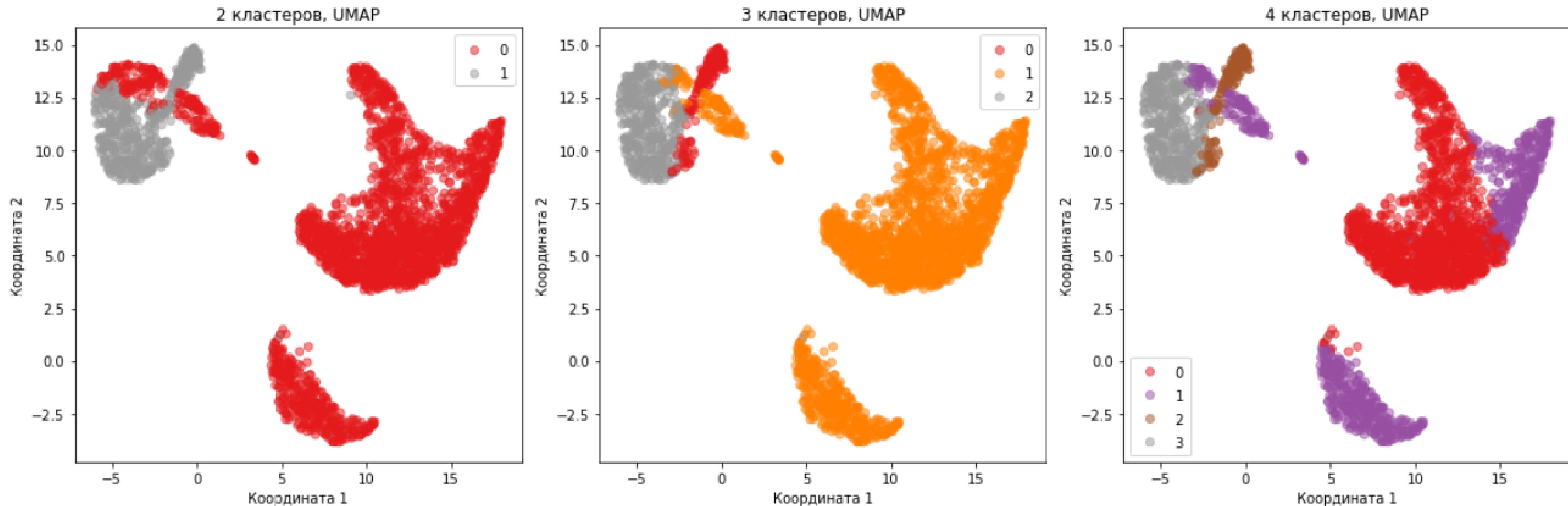
# DBSCAN



# DBSCAN

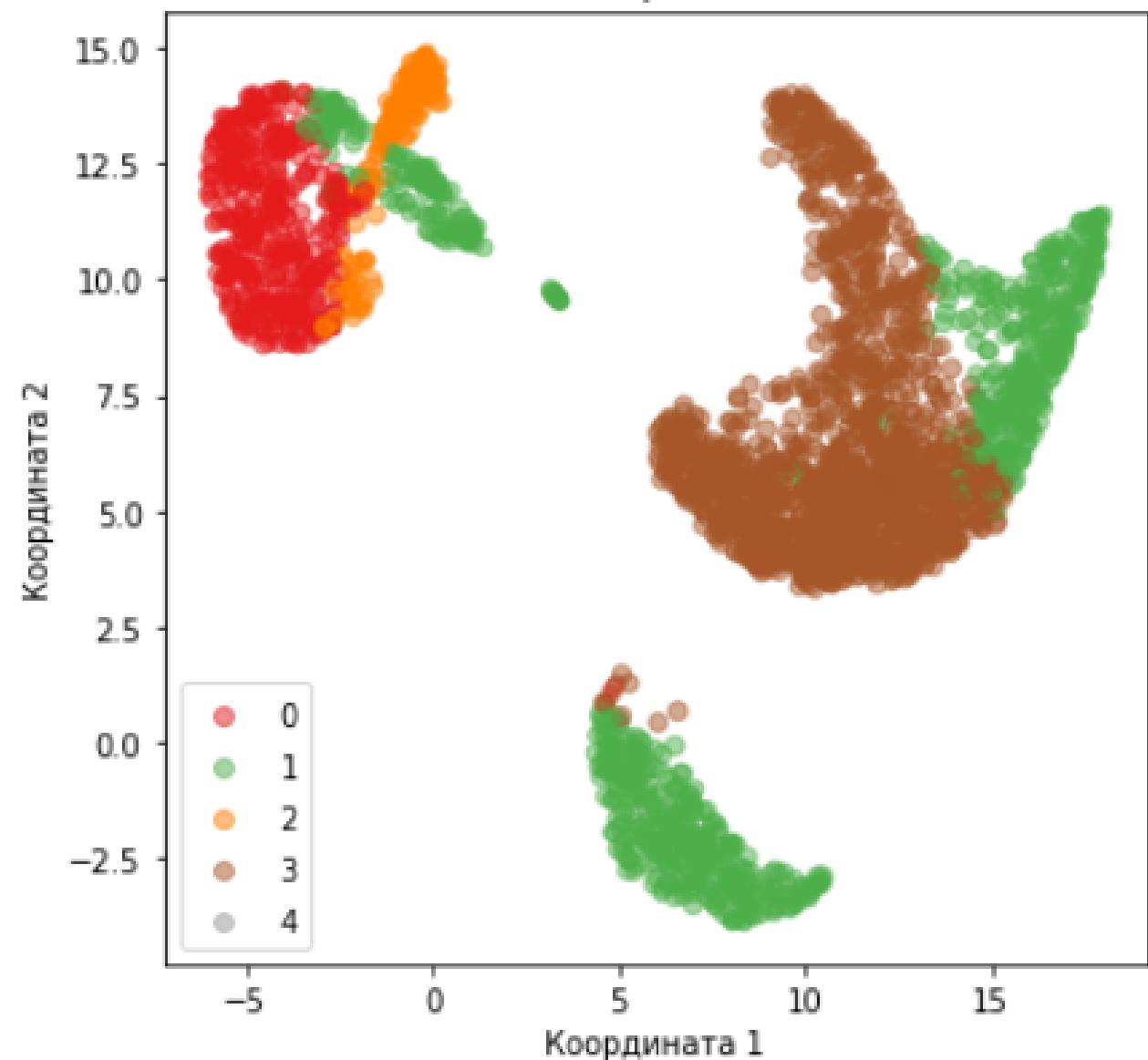


# KMeans

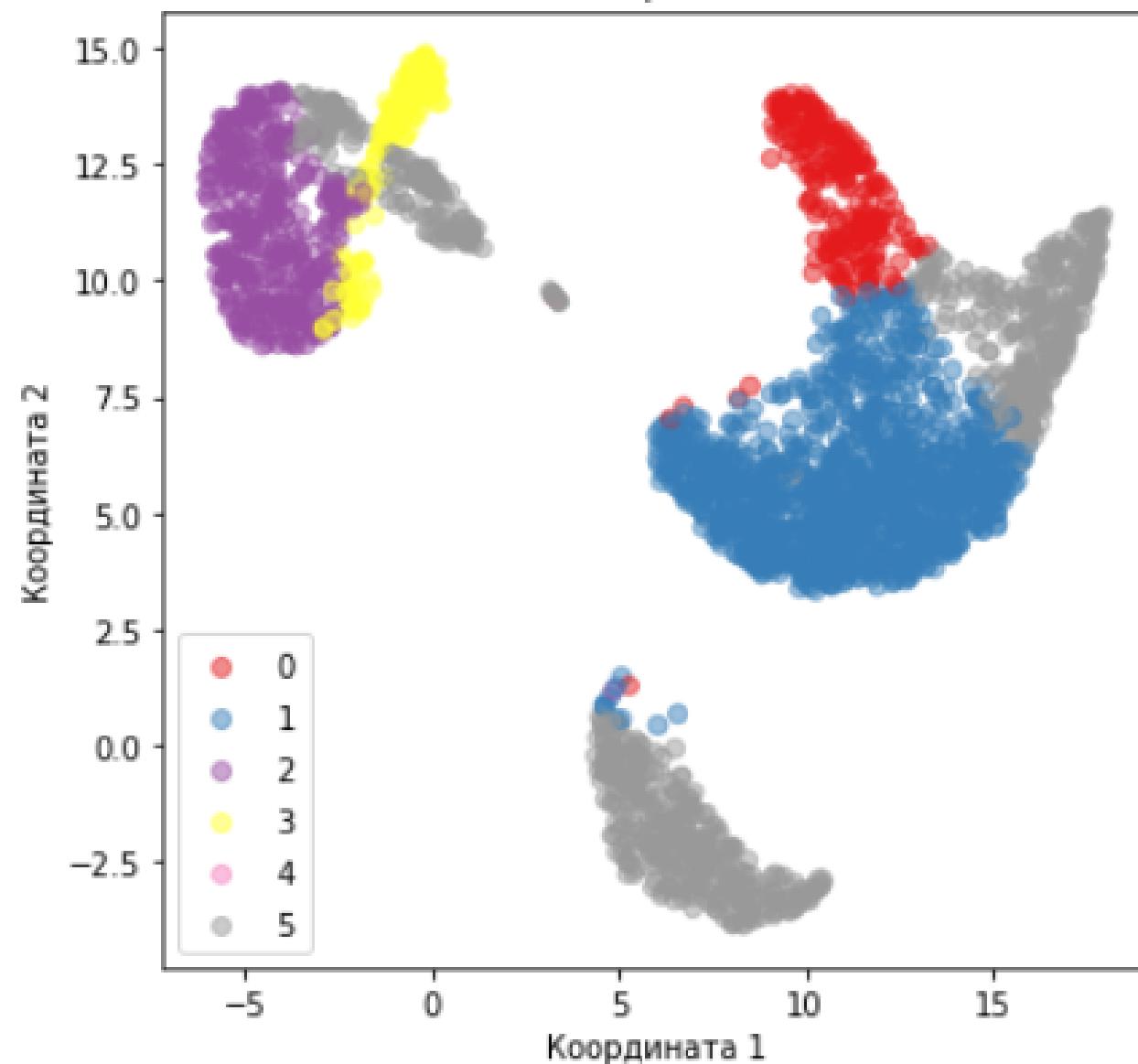


# KMeans

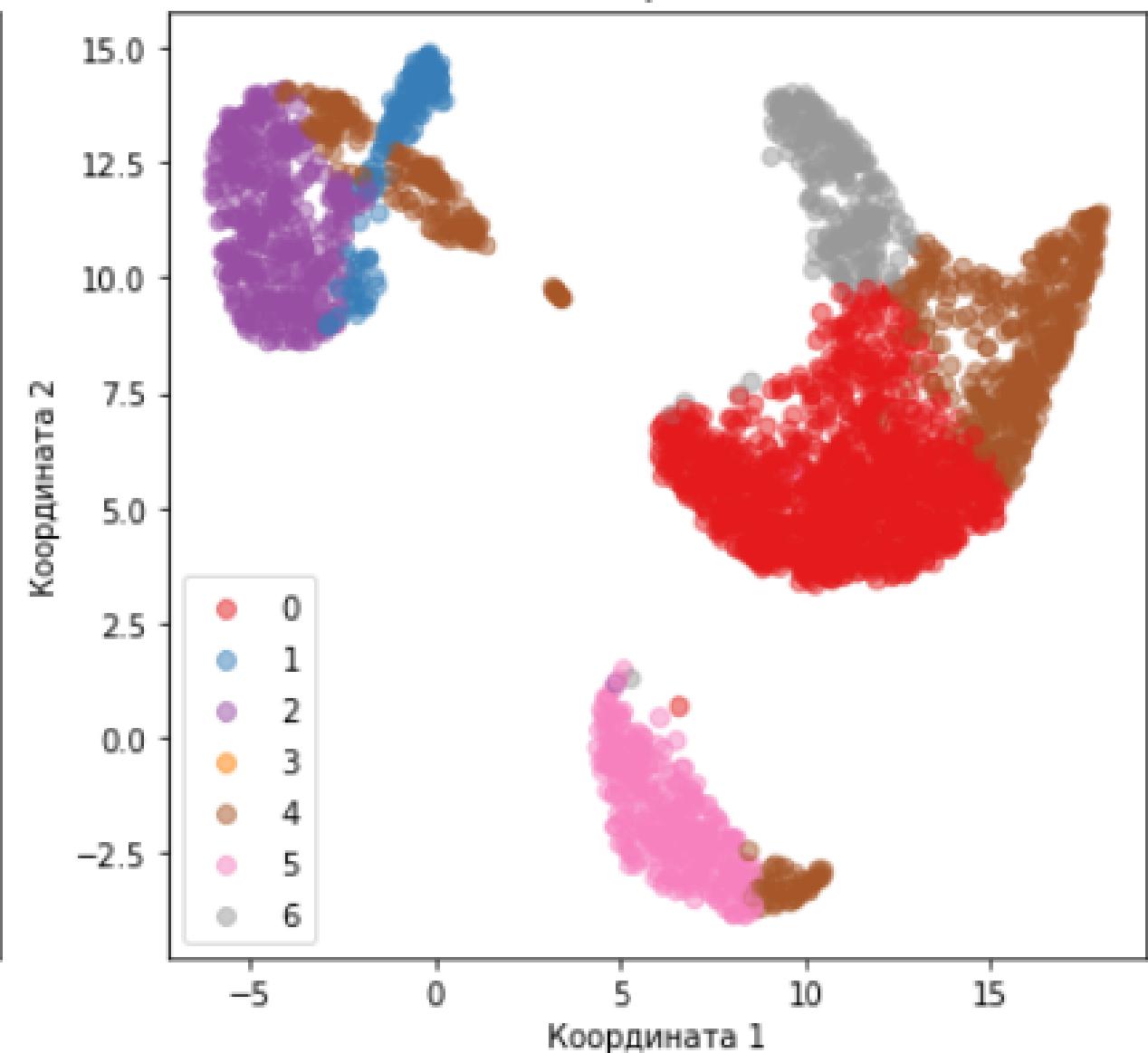
5 кластеров, UMAP



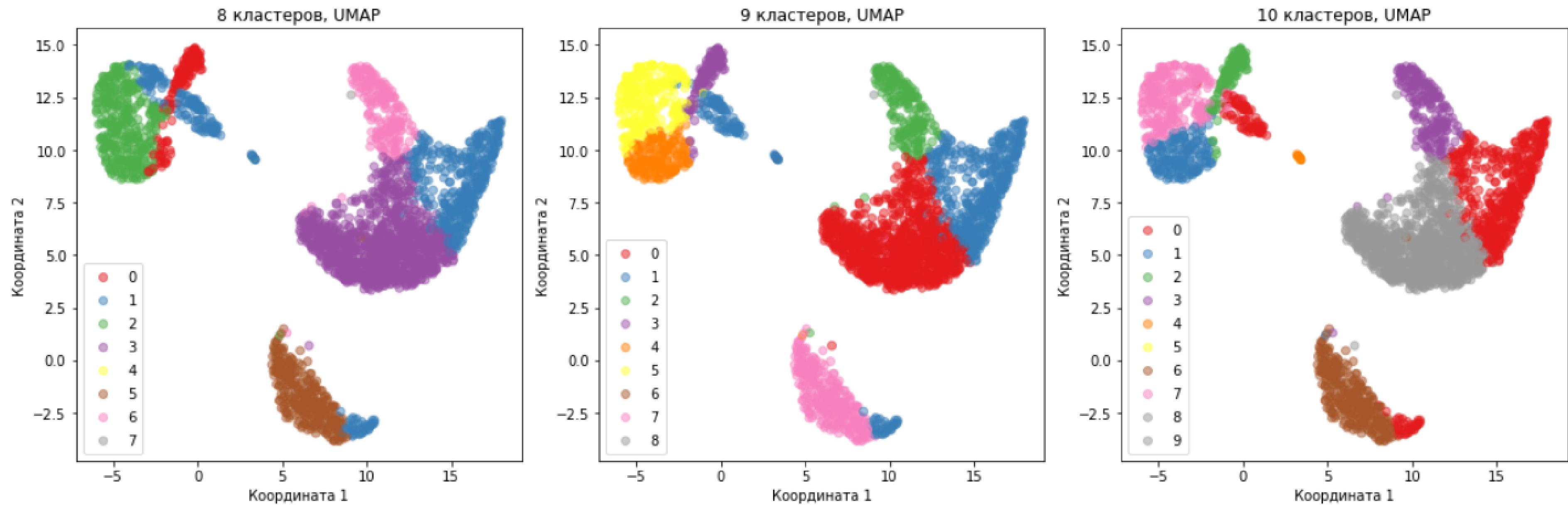
6 кластеров, UMAP



7 кластеров, UMAP



# KMeans



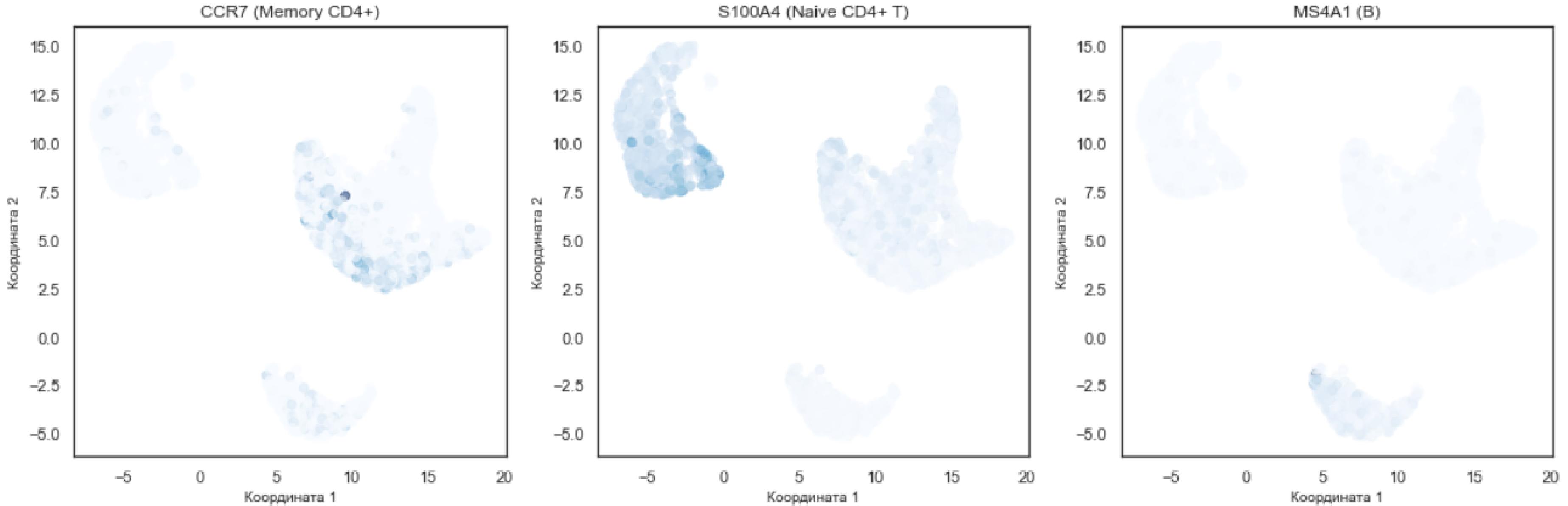
# Выделение кластеров по экспрессии маркерных генов

Тип клеток	Доля от общего числа МКПК, %	Маркеры
Наивные CD4 <sup>+</sup> Т клетки	25-40	IL7R, S100A4
CD4 <sup>+</sup> Т клетки памяти	25-40	IL7R, CCR7
CD14+ Моноциты	5-7	LYZ, CD14
FCGR3A+ Моноциты	5-7	FCGR3A, MS4A7
МК, мезодермальные киллеры	Нет данных	PPBP
CD8 <sup>+</sup> Т клетки	5-30	CD8A
NK-киллеры	10-30	GNLY, NKG7
DC, дендритные клетки	1-2	FCER1A', 'CST3
В клетки	5-10	MS4A1

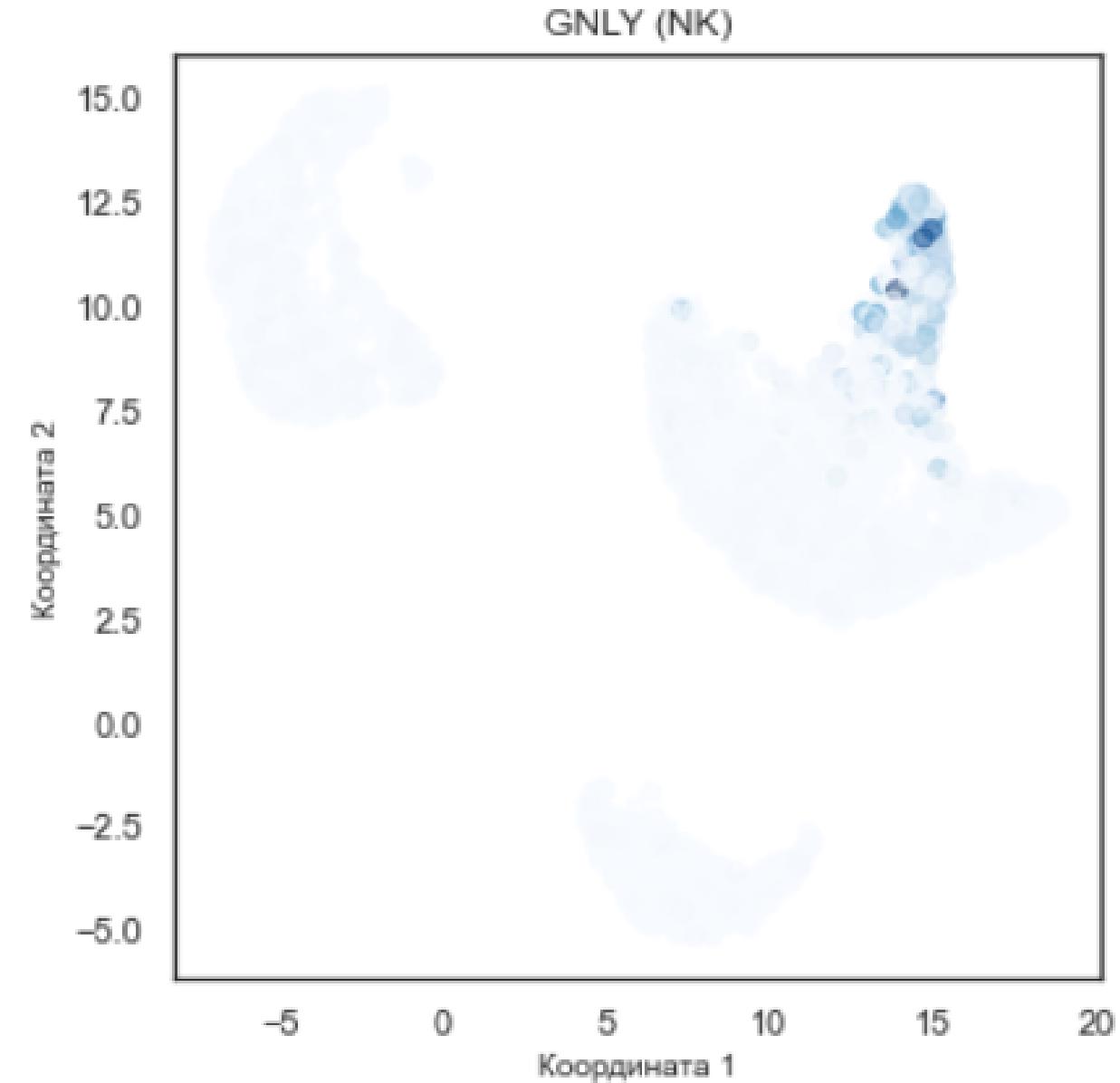
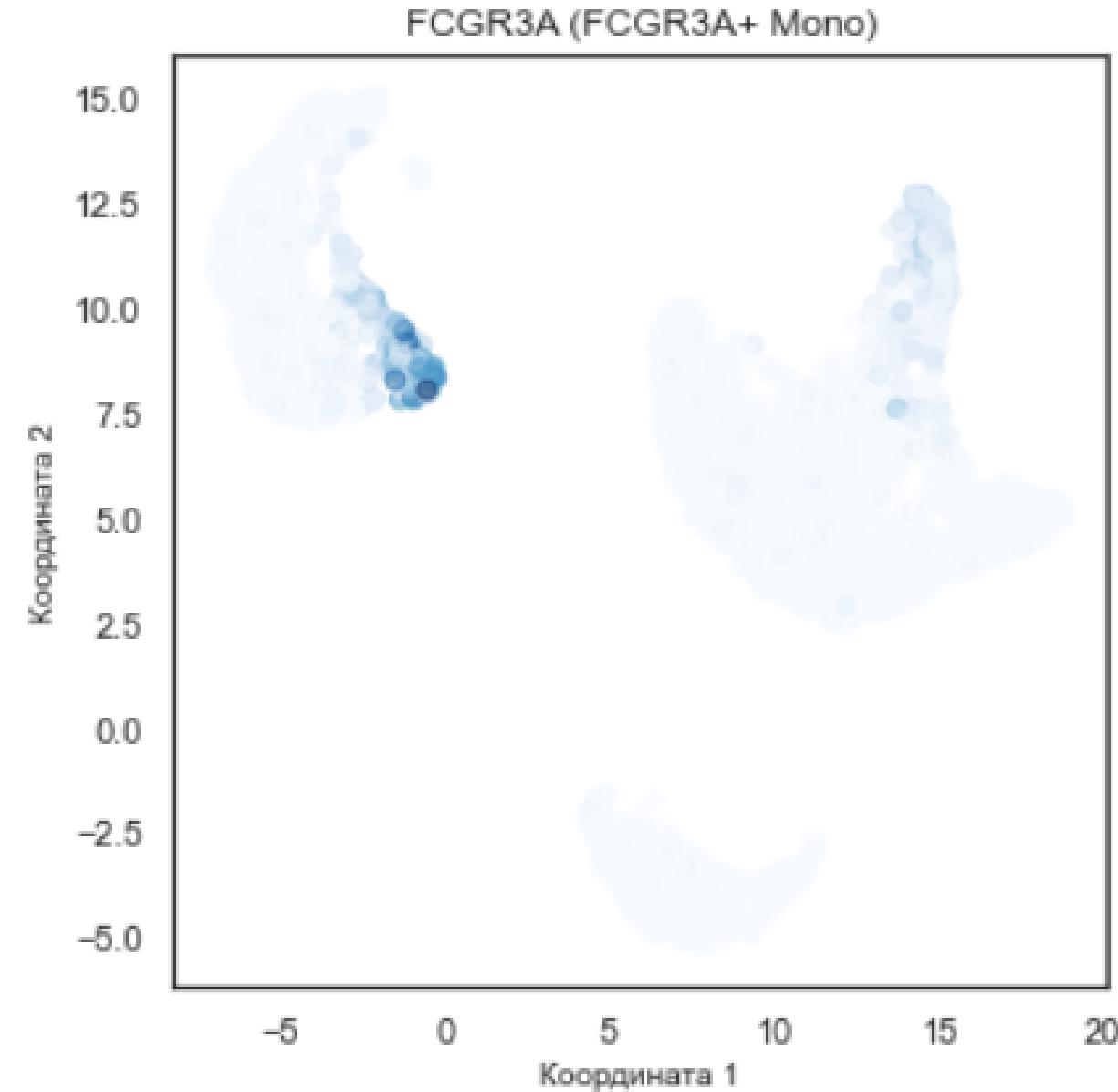
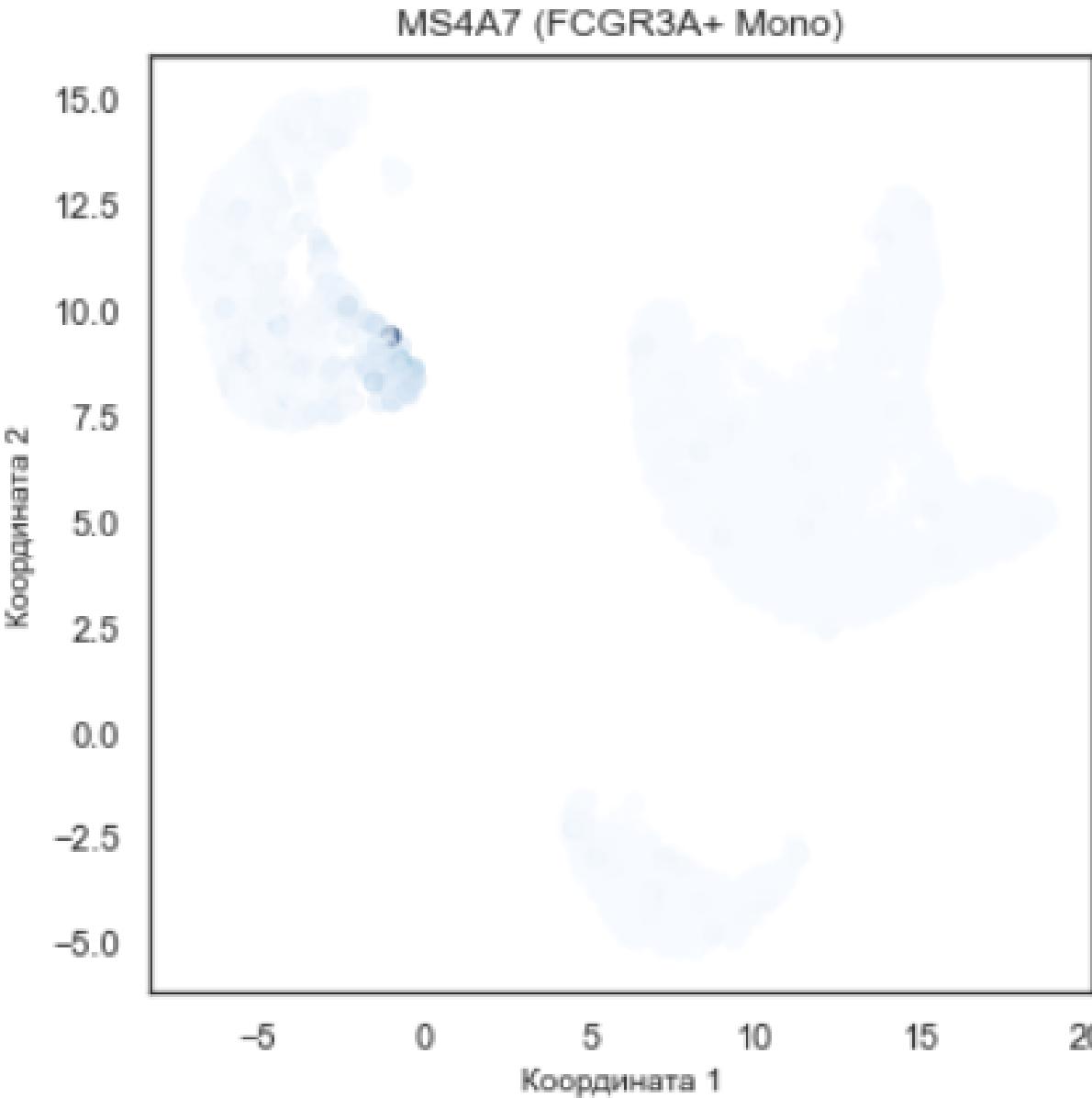
# Визуализация экспрессии маркерных генов



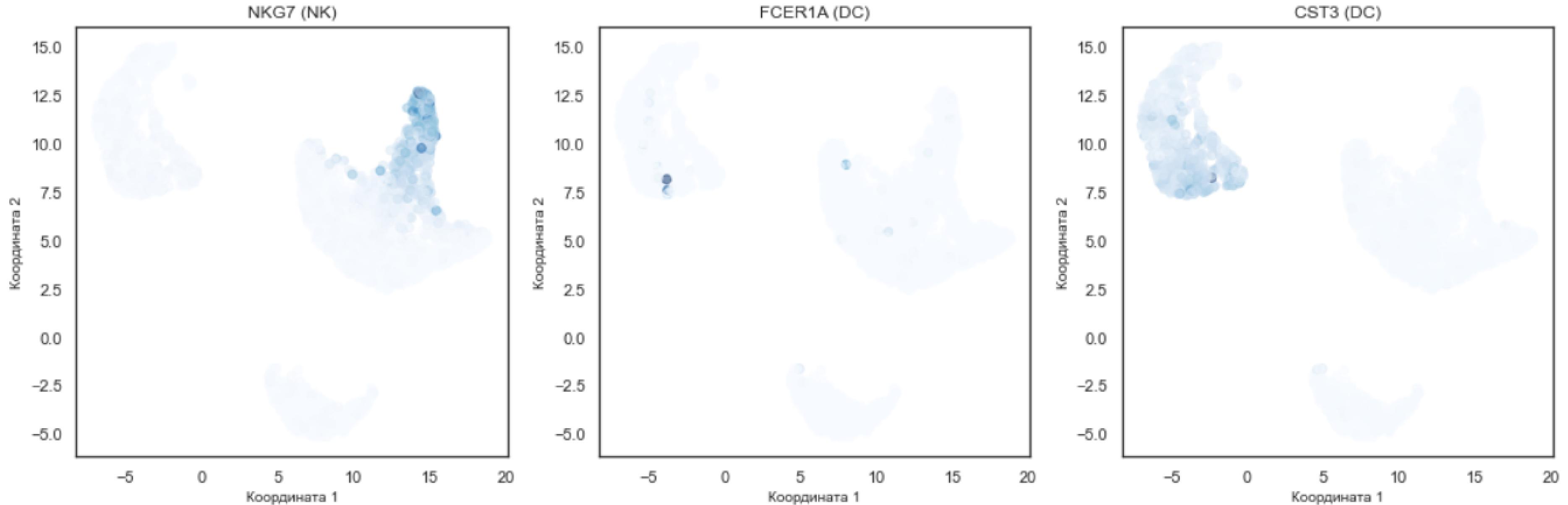
# Визуализация экспрессии маркерных генов



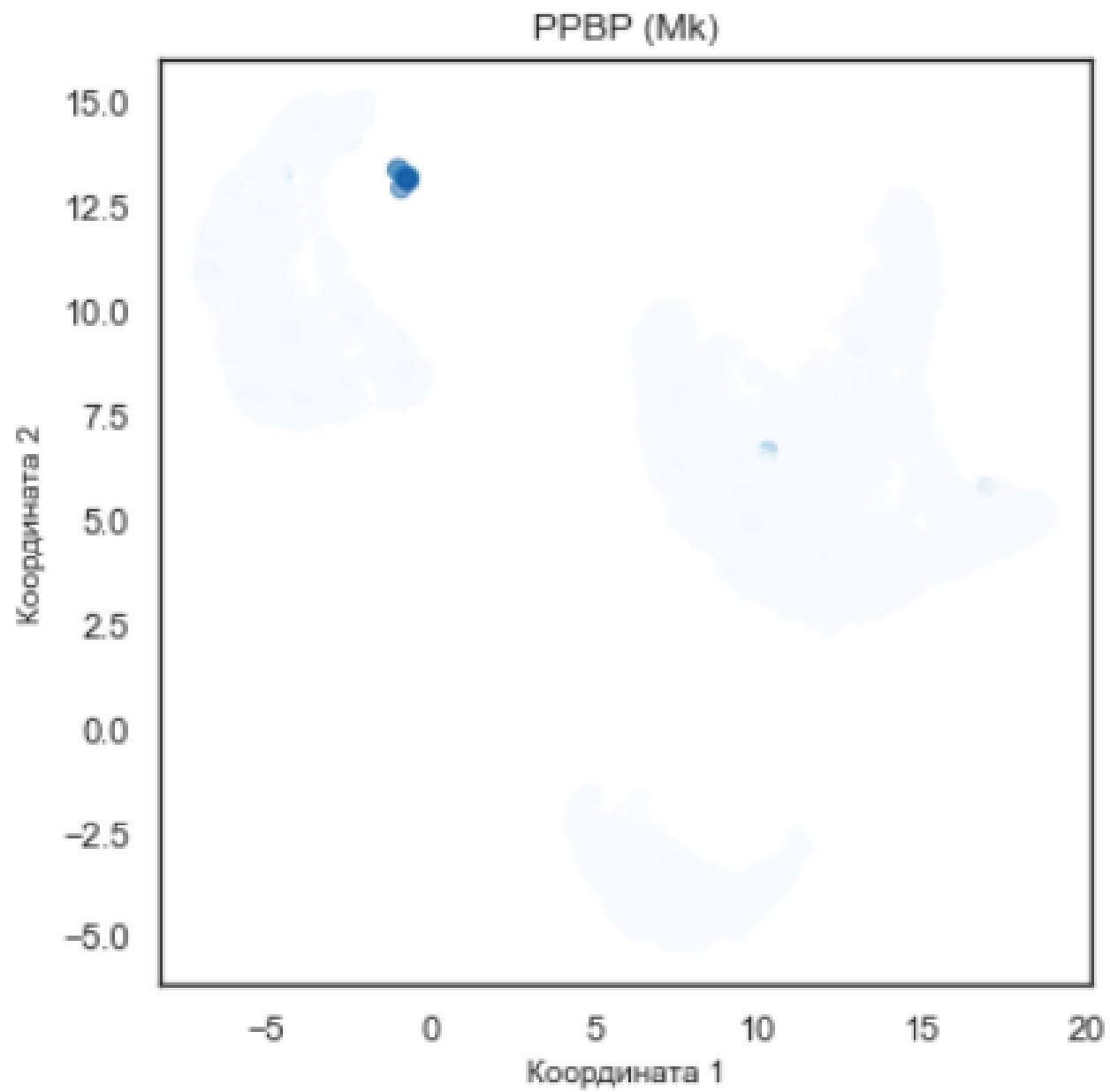
# Визуализация экспрессии маркерных генов



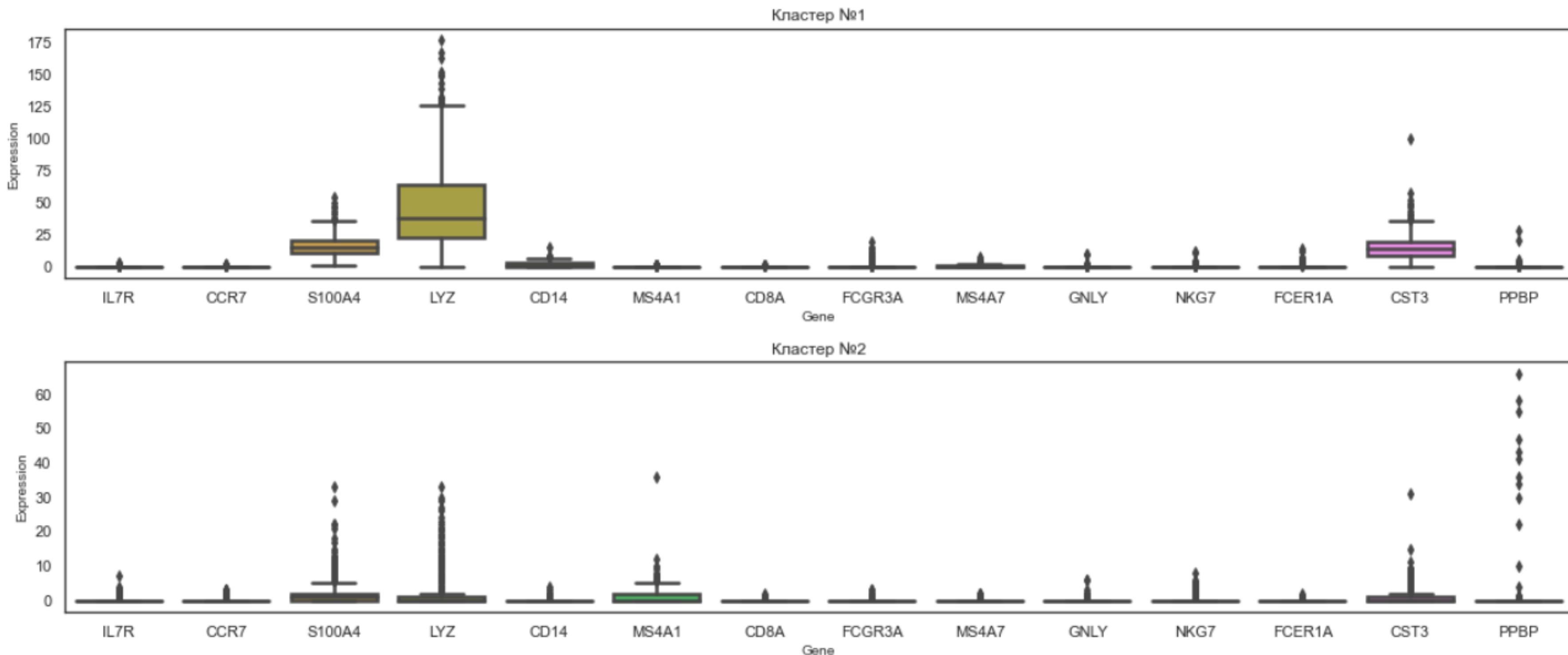
# Визуализация экспрессии маркерных генов



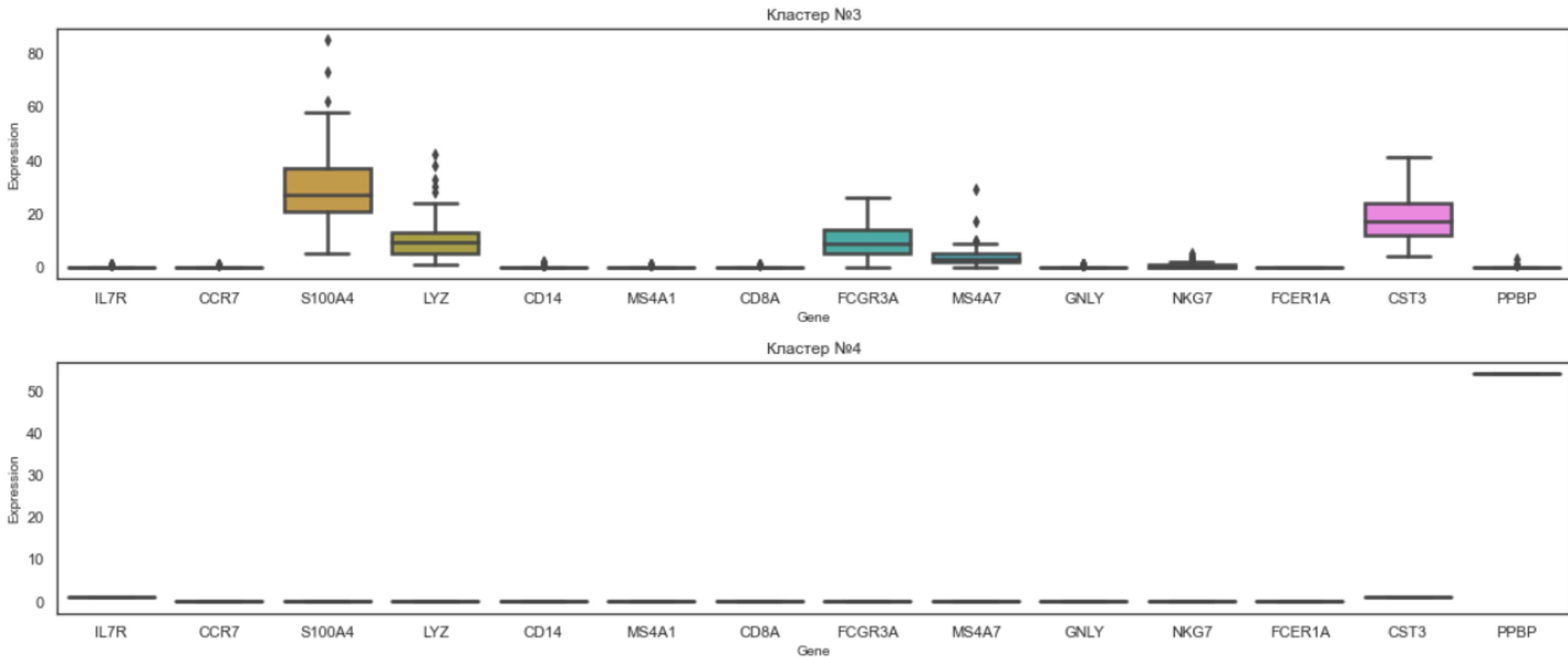
# Визуализация экспрессии маркерных генов



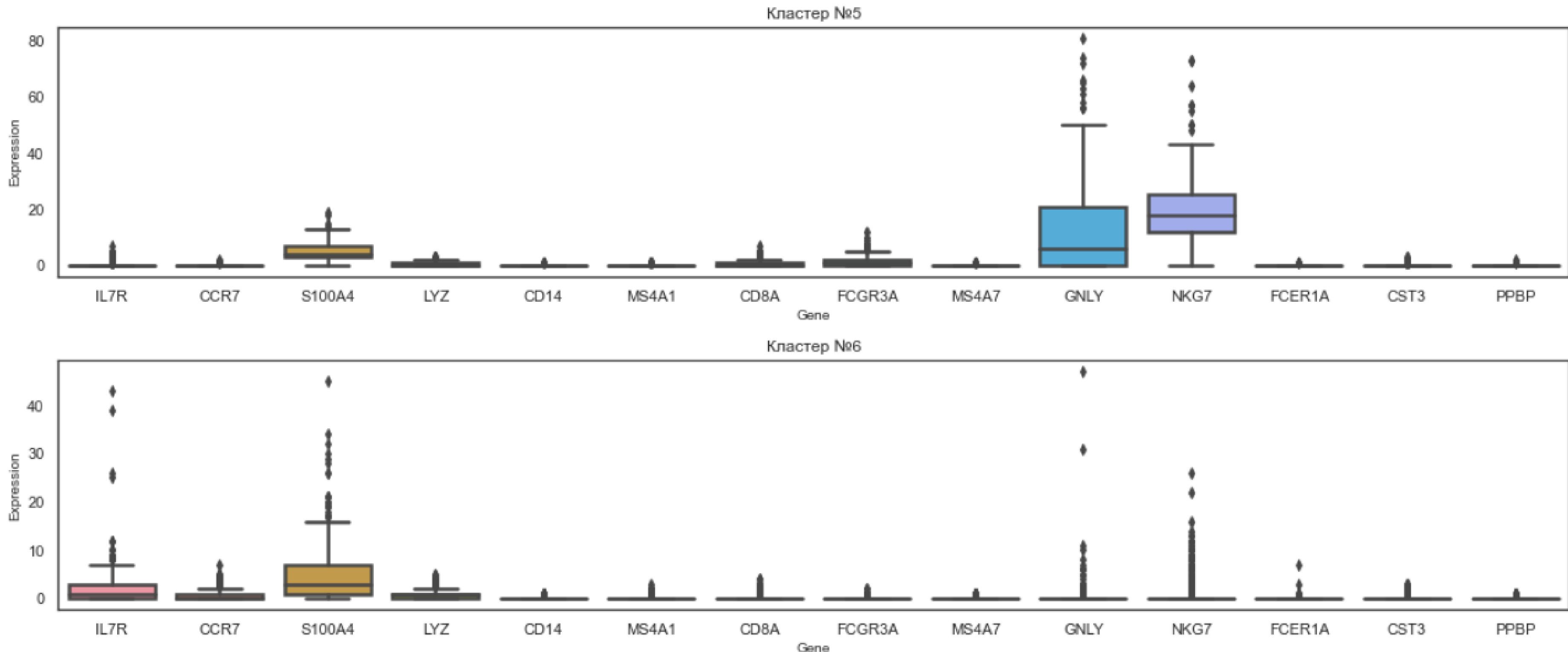
# AgglomerativeClustering, 6 кластеров



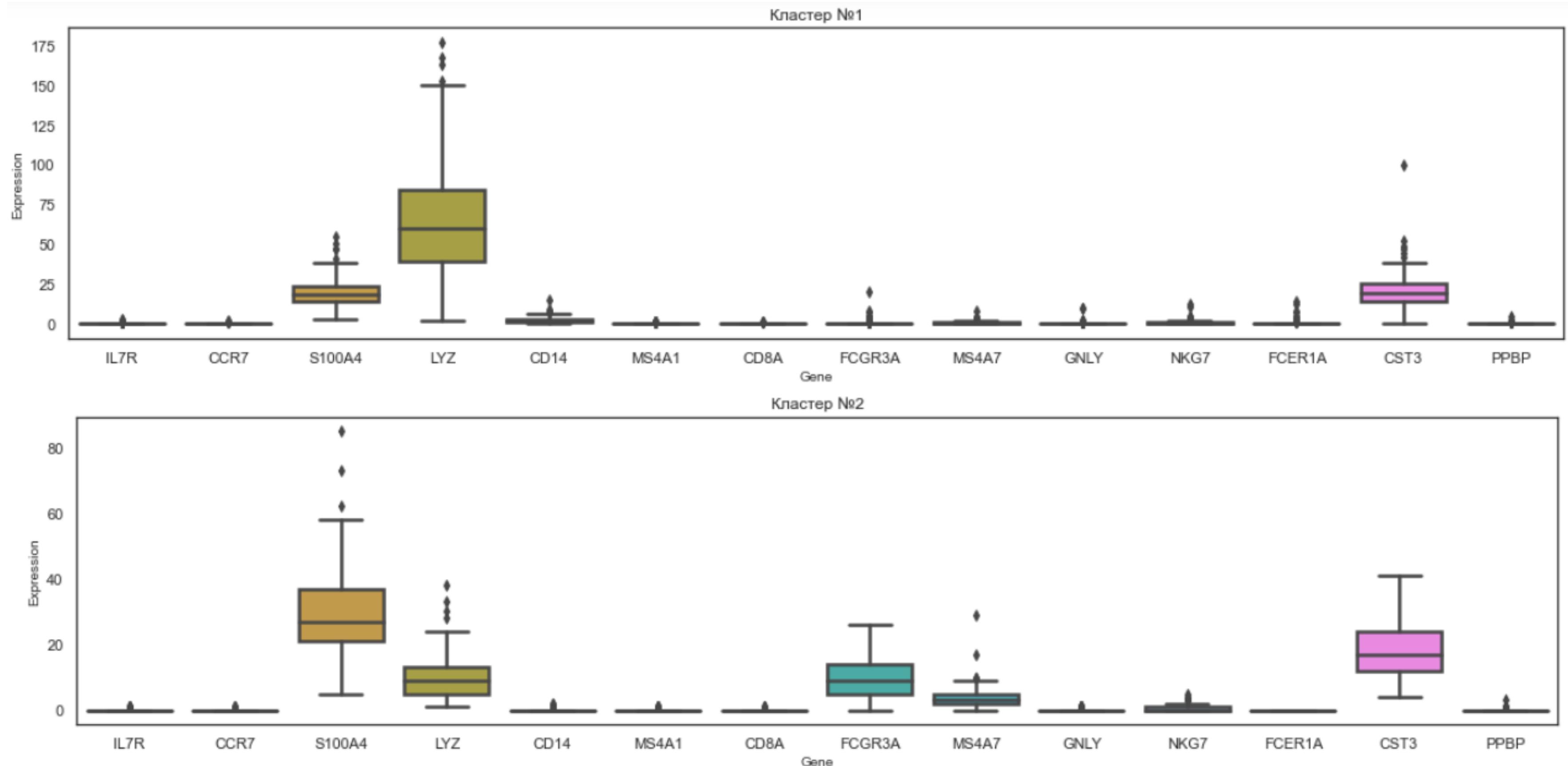
# AgglomerativeClustering, 6 кластеров



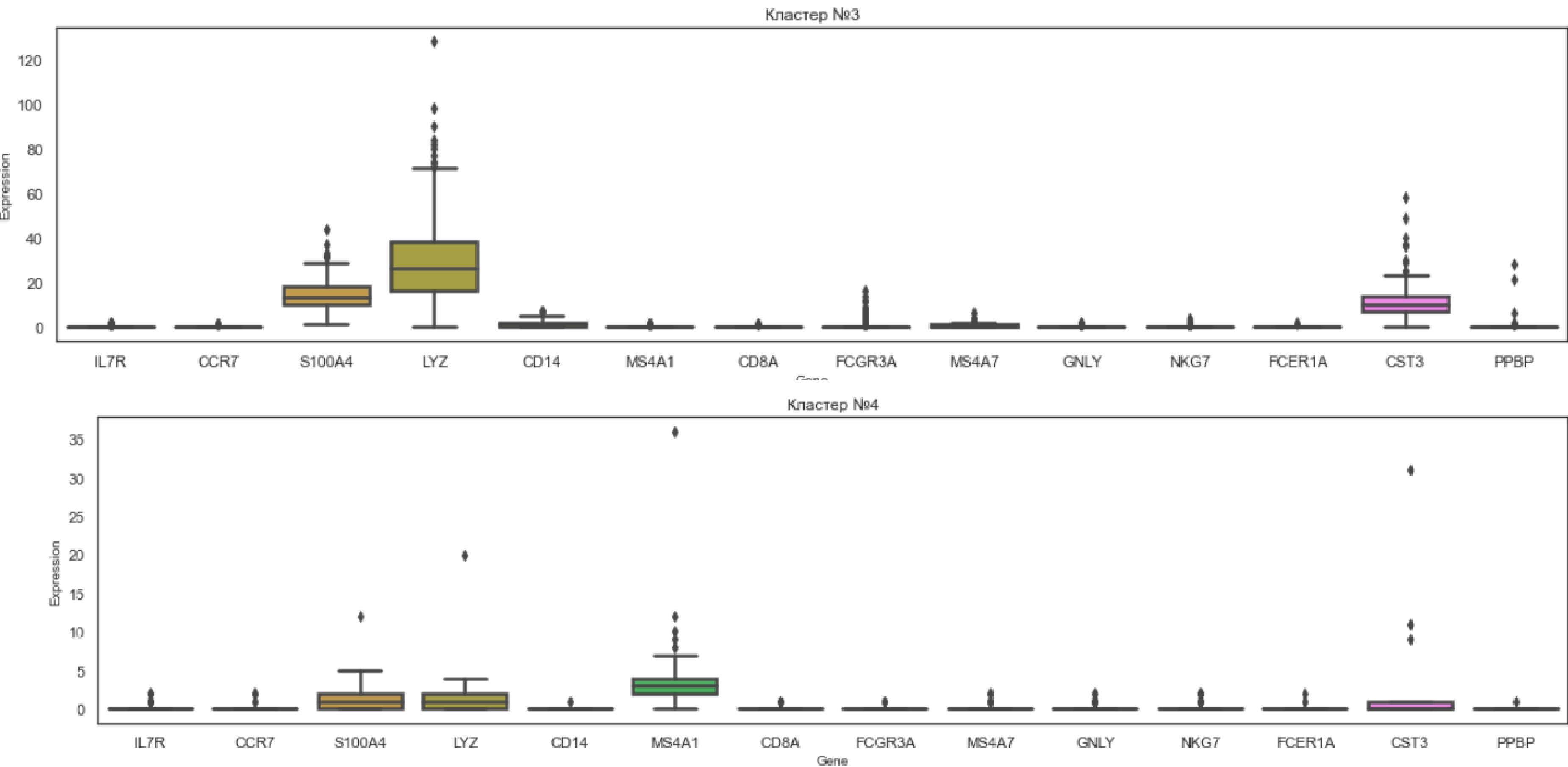
# AgglomerativeClustering, 6 кластеров



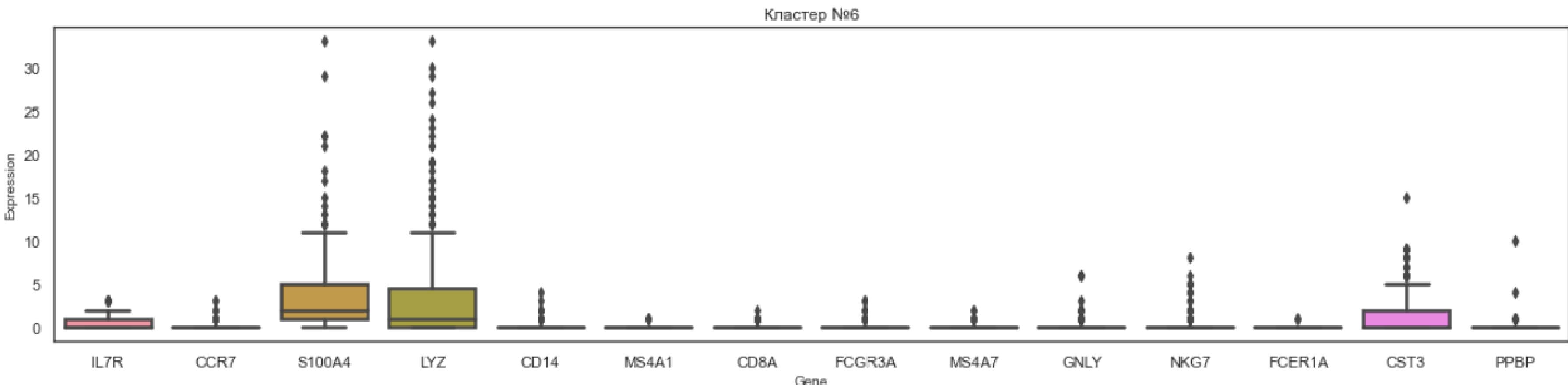
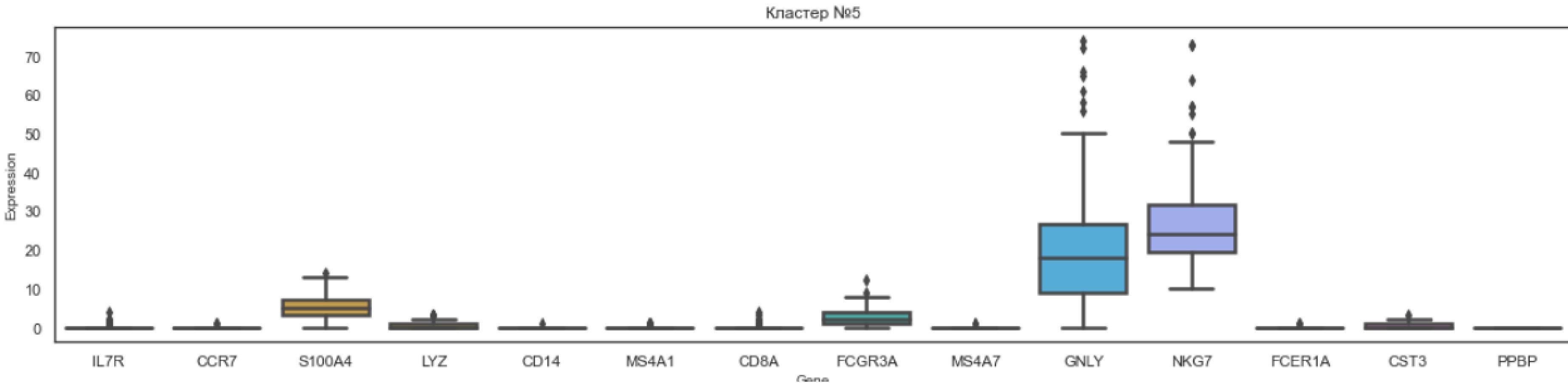
# AgglomerativeClustering, 15 кластеров



# AgglomerativeClustering, 15 кластеров

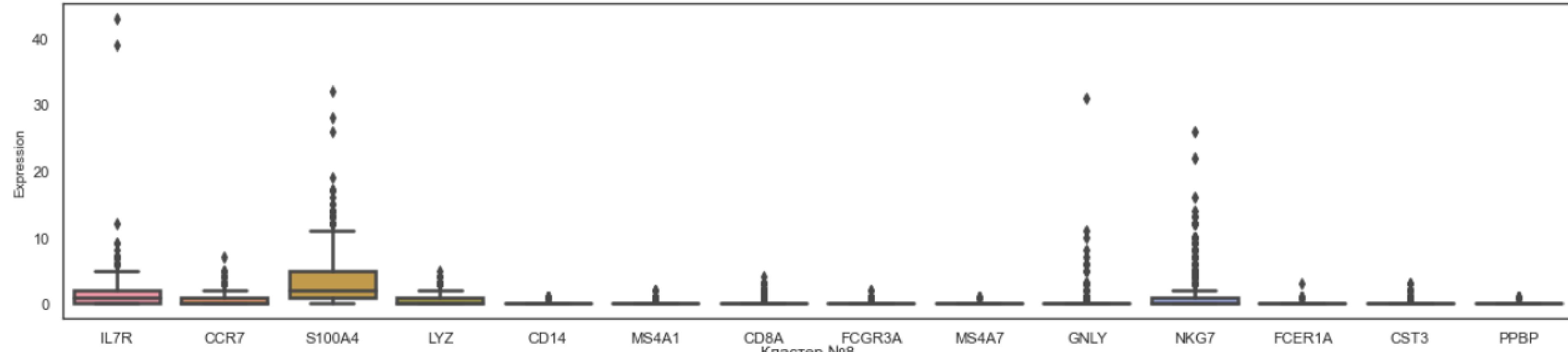


# AgglomerativeClustering, 15 кластеров

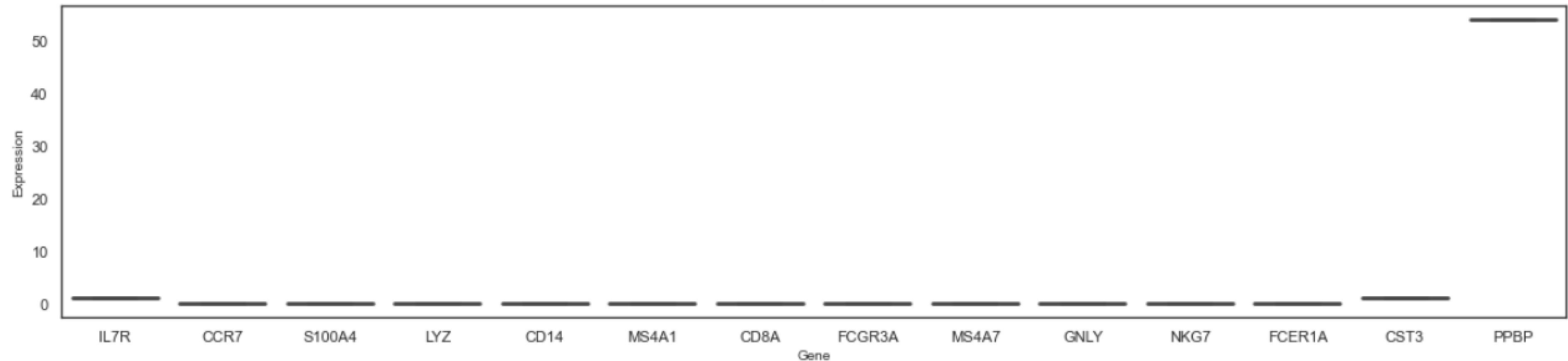


# AgglomerativeClustering, 15 кластеров

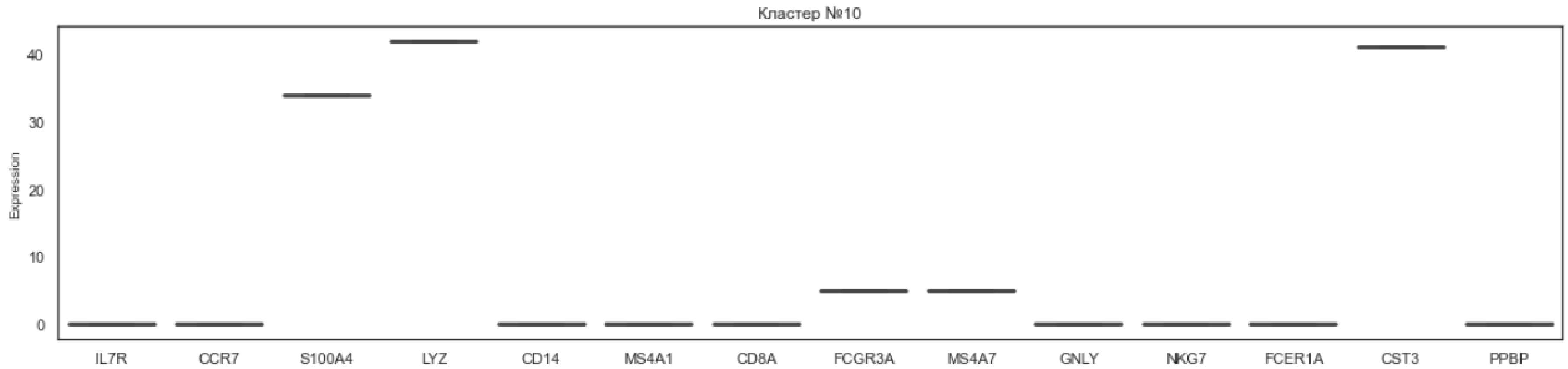
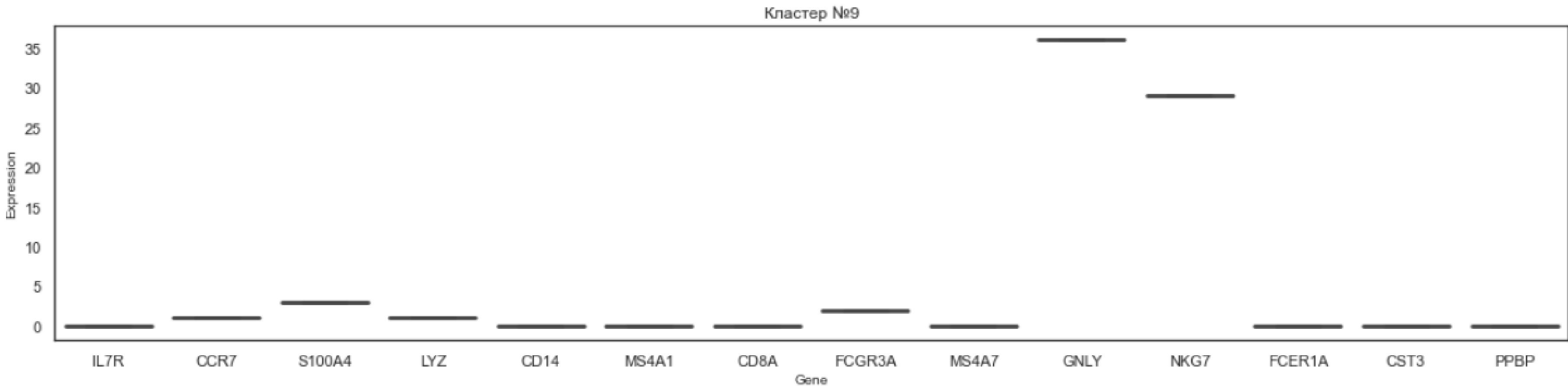
Кластер №7



Кластер №8

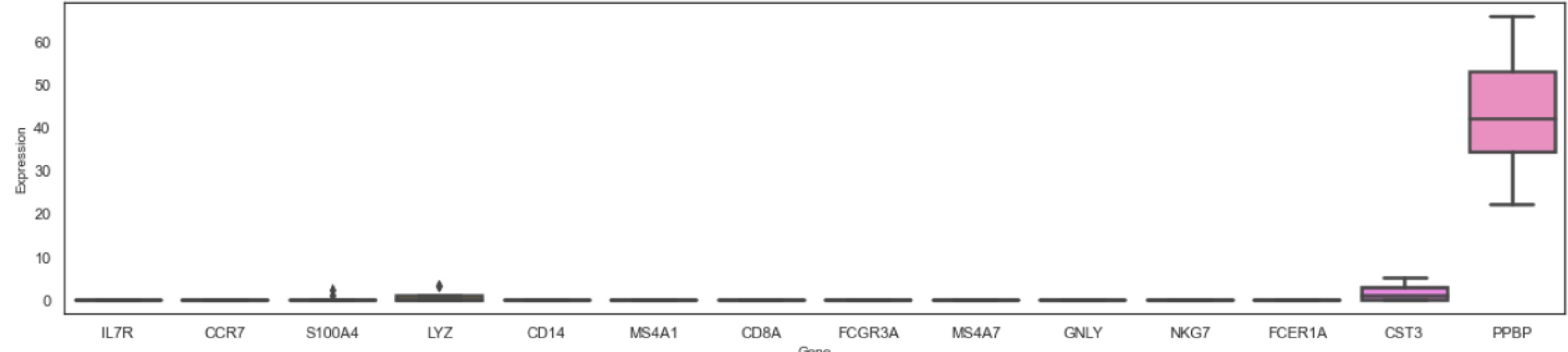


# AgglomerativeClustering, 15 кластеров

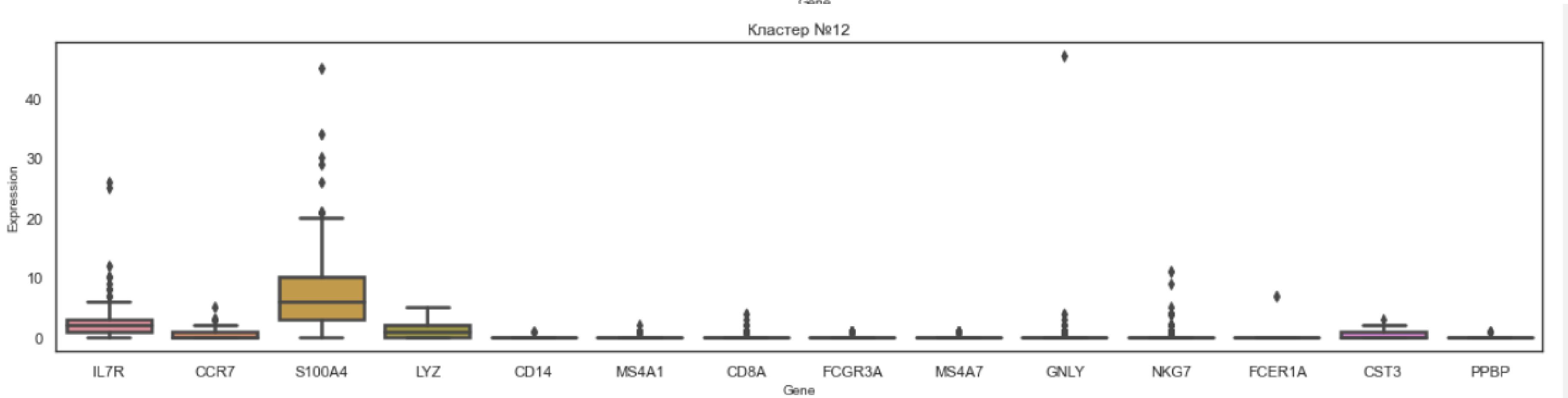


# AgglomerativeClustering, 15 кластеров

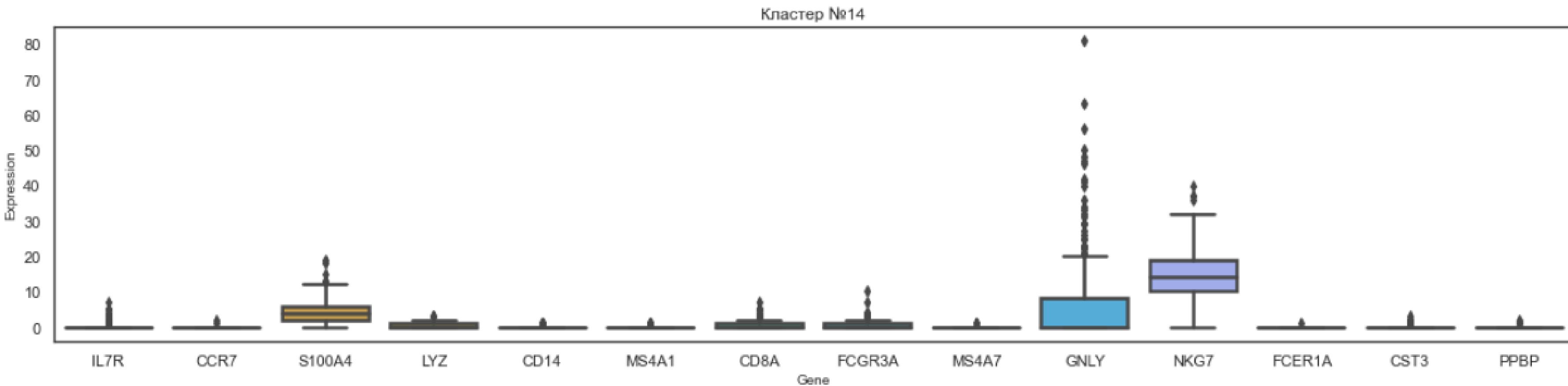
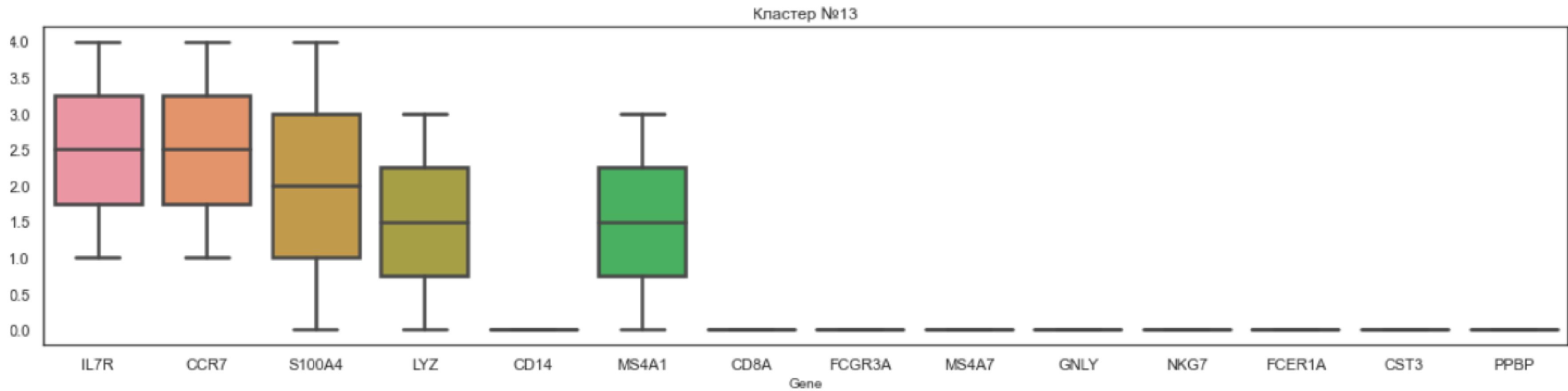
Кластер №11



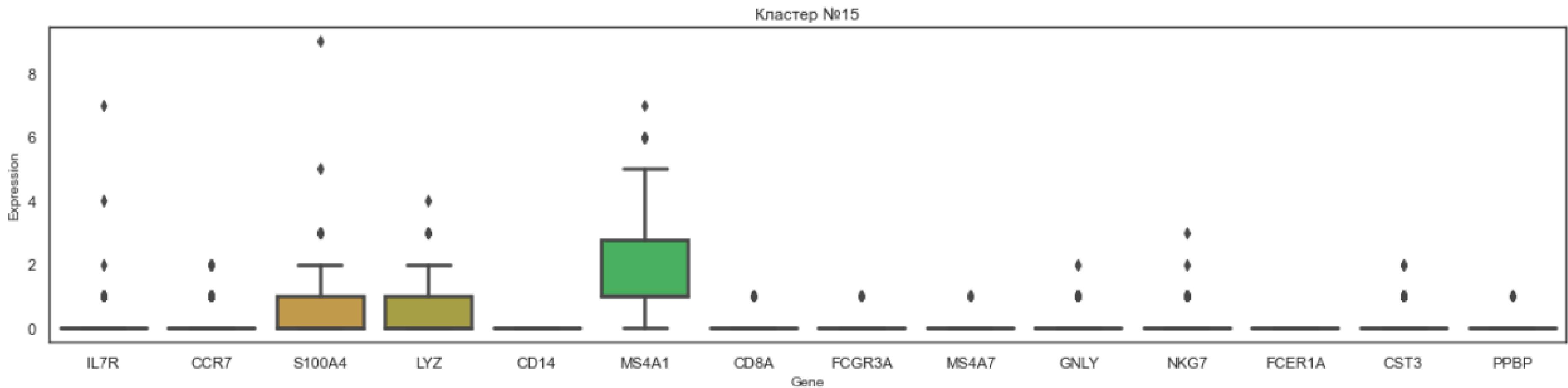
Кластер №12



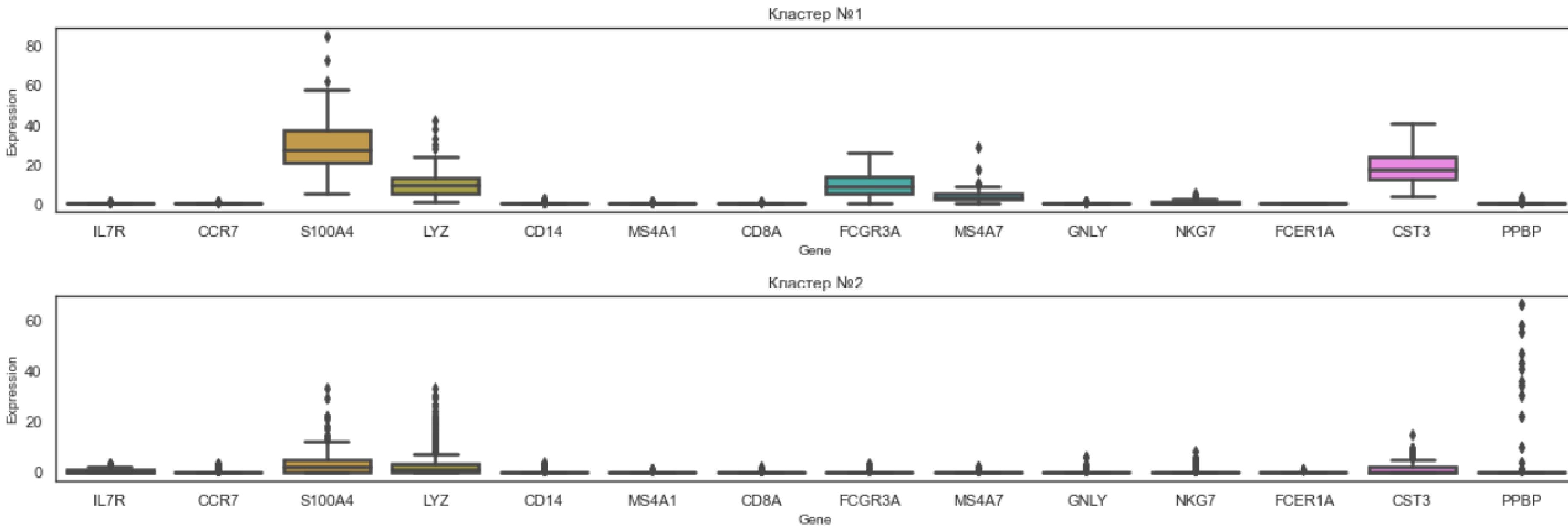
# AgglomerativeClustering, 15 кластеров



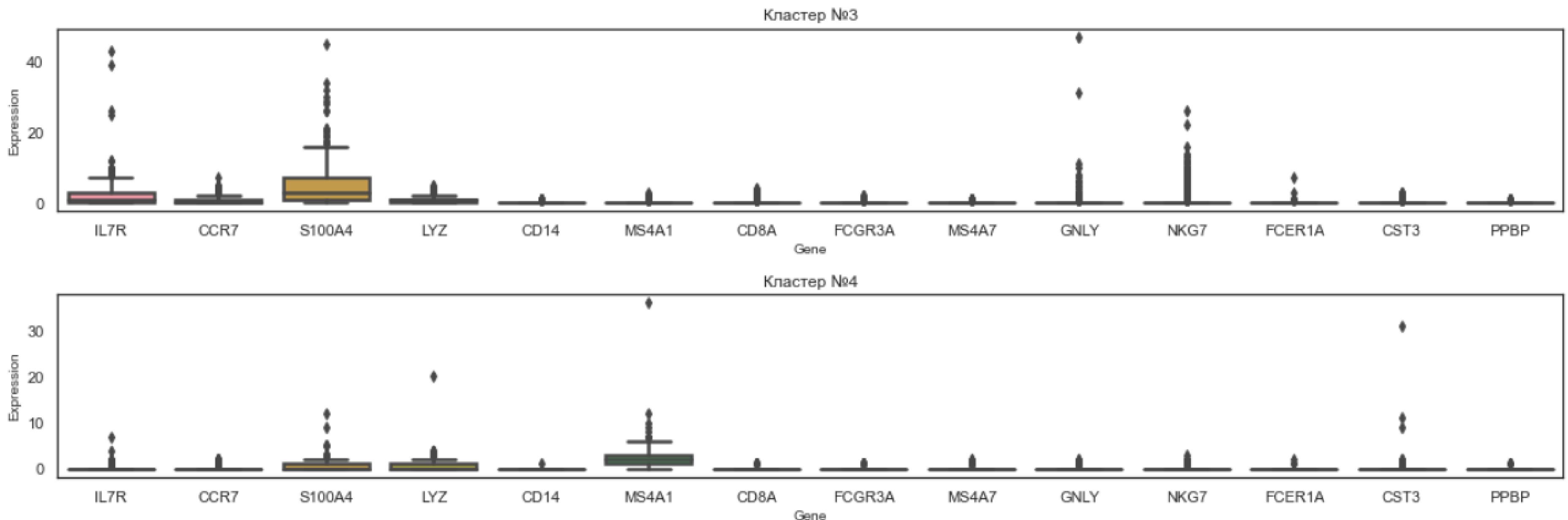
# AgglomerativeClustering, 15 кластеров



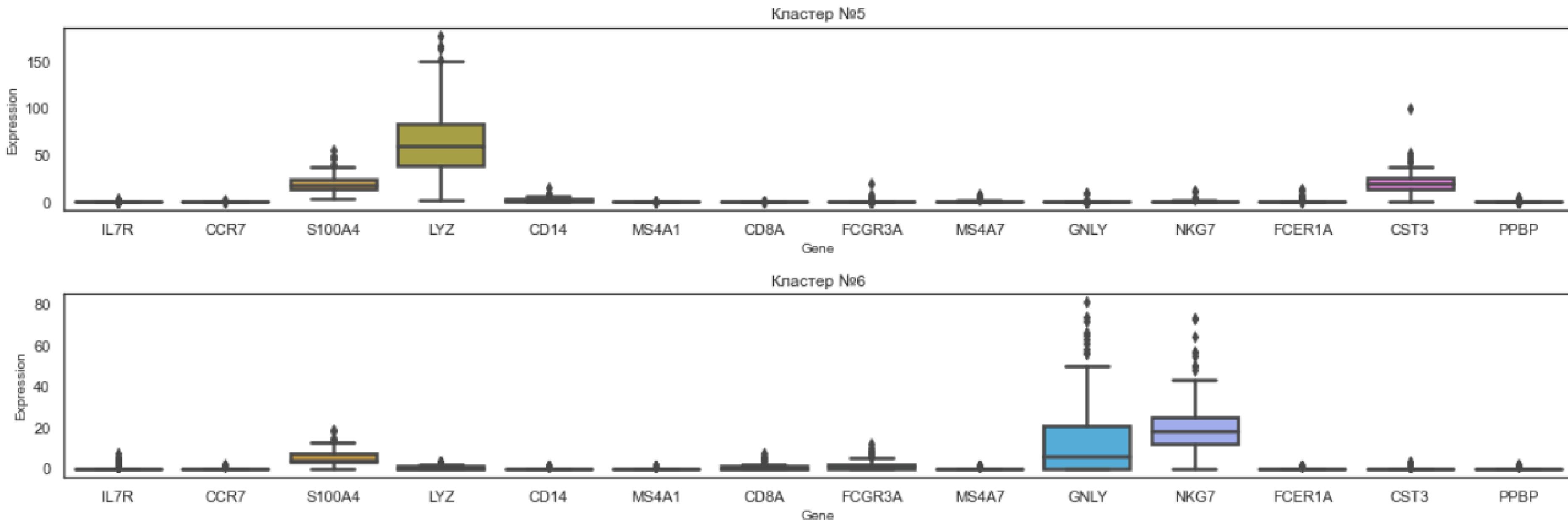
# AgglomerativeClustering, 9 кластеров



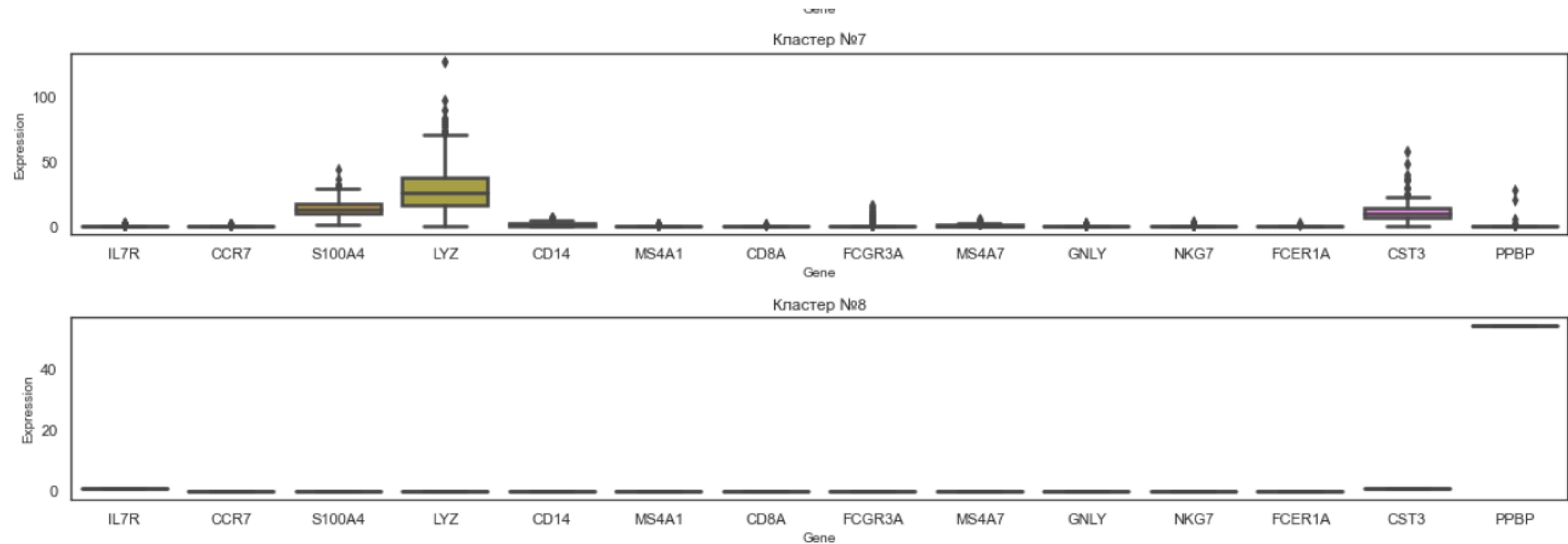
# AgglomerativeClustering, 9 кластеров



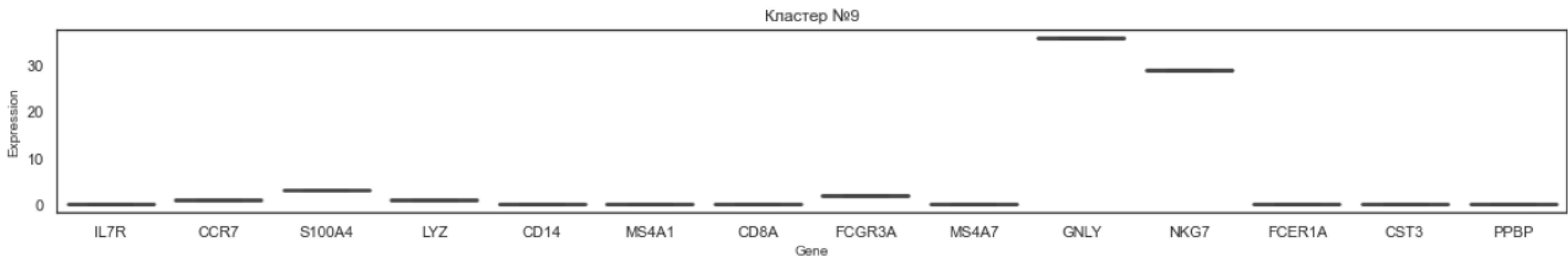
# AgglomerativeClustering, 9 кластеров



# AgglomerativeClustering, 9 кластеров



# AgglomerativeClustering, 9 кластеров



# Выводы из полученных кластеризаций

- При разбиении на 6 кластеров по экспрессии генов-маркеров хорошо выделялись типы клеток, такие как NK-клетки, M<sub>k</sub>, В-клетки и моноциты. Остальные типы клеток выделялись хуже
- При разбиении на 15 кластеров некоторые группы клеток разделялись на большее число кластеров (разделились CD14 и FCGR3A моноциты и NK)
- При разбиении на 9 кластеров выделялись NK клетки, M<sub>k</sub>, CD4+ клетки памяти и наивные клетки, моноциты CD14, моноциты FCGR3A+, В клетки. Однако, явно выделить все 9 известных типов клеток не удалось.

**Таким образом**, для кластеризации мононуклеарных клеток крови лучше пользоваться 9-ти кластерной системой.

Спасибо за внимание!

