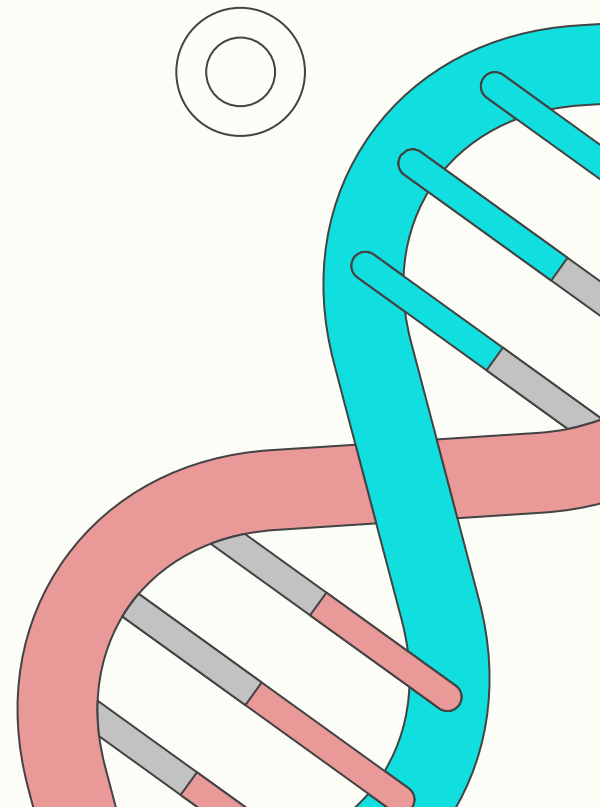


Dec 17, 2025

Анализ архетипов vs. обычная кластеризация

Ваничкин Алексей





01

Введение

В начале было слово



Проблема анализа гетерогенности в single-cell данных

Современные single-cell технологии (в частности scRNA-seq) выявляют **высокую степень клеточной гетерогенности**, однако:

- стандартные методы анализа ориентированы на **дискретную классификацию** клеток. Такие методы предполагают, что клетки можно разделить на **конечное число** устойчивых типов

Но это **допущение** часто не соответствует биологической реальности, поскольку:

- клетки могут находиться в **континууме функциональных состояний**
- различия между клетками могут отражать градиенты функций, а не типы

Концепция компромиссов (trade-offs) в биологических системах

Ключевая биологическая гипотеза статьи:

- **Биологические системы** подвержены **ограничениям ресурсов** и поэтому не могут **одновременно** оптимизировать все функции.

Из этого следует, что:

- клетки вынуждены **распределять** ресурсы **между** несколькими **конкурирующими задачами**
- **улучшение** одной функции приводит к **ухудшению** другой
- такие системы эволюционно приходят к состоянию **Парето-оптимальности**

Pareto Task Inference как альтернатива кластеризации

Авторы прямо **противопоставляют ParTI** стандартной **кластеризации**:

- кластеризация ищет **центры плотности**;
- ParTI ищет **экстремальные состояния**.

ParTI не предполагает:

- существования **дискретных типов клеток**
- **жёсткого присваивания** клетки к одному состоянию

Вместо этого метод моделирует:

- **континуум состояний**, возникающий как компромисс между задачами
- клетки как **смеси** специализированных **фенотипов**

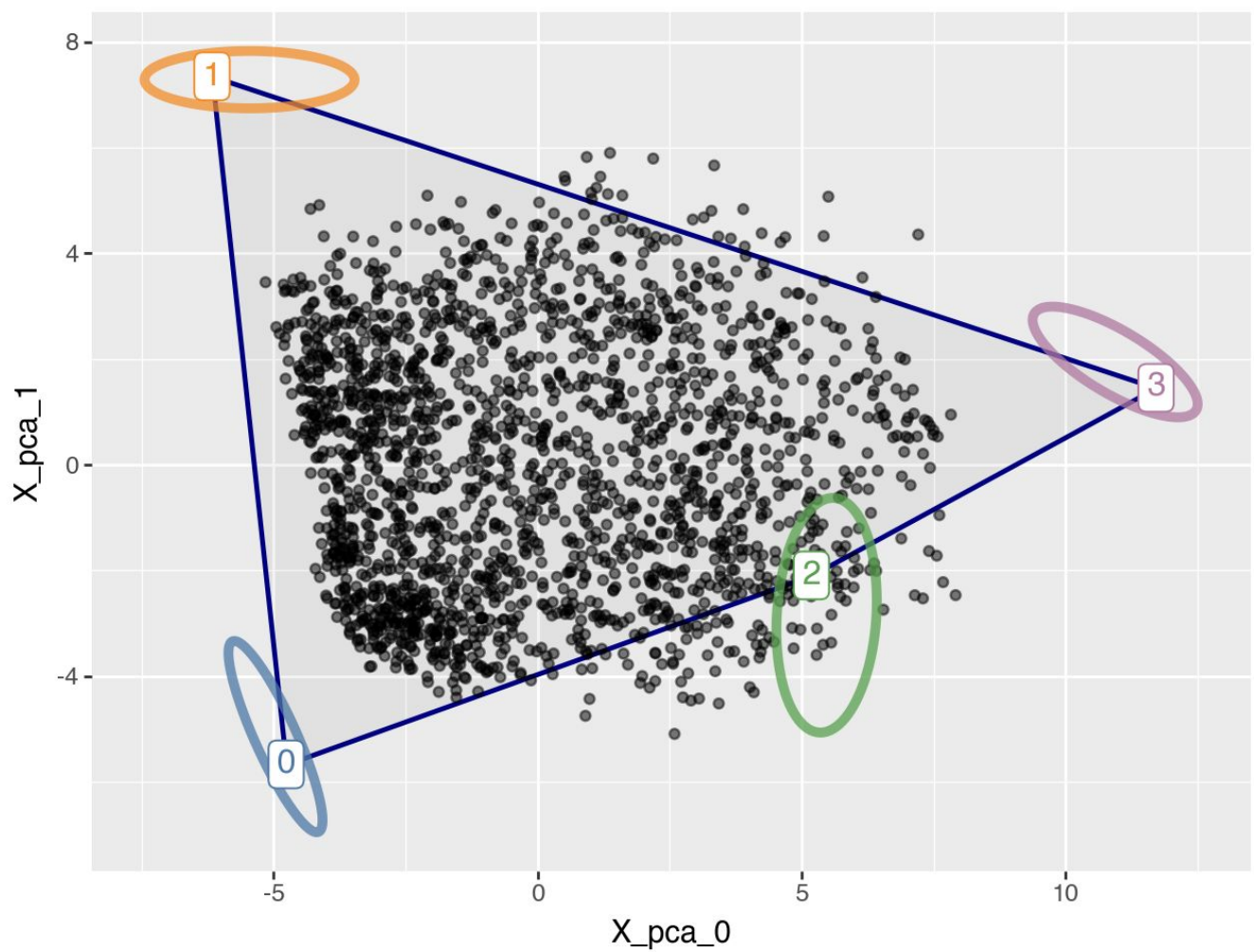
Геометрическая формулировка гипотезы ParTI

Ключевое теоретическое **утверждение** статьи:

- если система оптимизирует **K конкурирующих задач**, допустимые фенотипы лежат **внутри выпуклого многогранника** с **K** вершинами.

В контексте scRNA-seq:

- **каждая клетка** — точка в пространстве признаков
- **многогранник** — геометрическое отражение ограничений
- **вершины** — архетипы, соответствующие оптимизации отдельных задач



Резюме

Кластеризация	ParTI
Дискретные типы	Континуум состояний
Центры плотности	Экстремальные фенотипы
Классы	Компромиссы
Гипотеза устойчивых типов	Гипотеза trade-offs



02

Выбор датасета и предобработка

Не бывает мрачных датасетов, бывают только мрачные люди

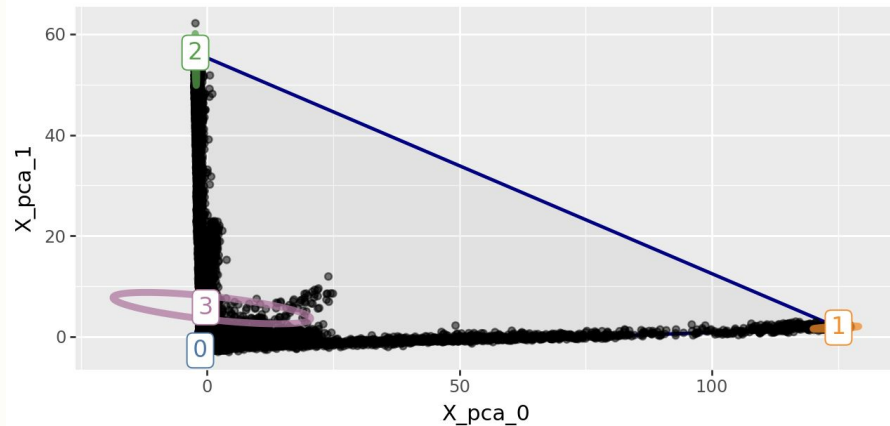
Первые попытки

Выбранный датасет: *Tabula Muris Consortium*

- Использовался ранее для сравнения различных пайплайнов кластеризации.

Основные сложности:

- Содержит данные из **20 различных тканей**
- Имеет **предобработанные UMI**, что делало PCA и последующий анализ менее информативным
- Из-за большого числа клеток и тканей методы кластеризации выделяли **около 150 кластеров**, что сильно усложняло визуализацию и интерпретацию



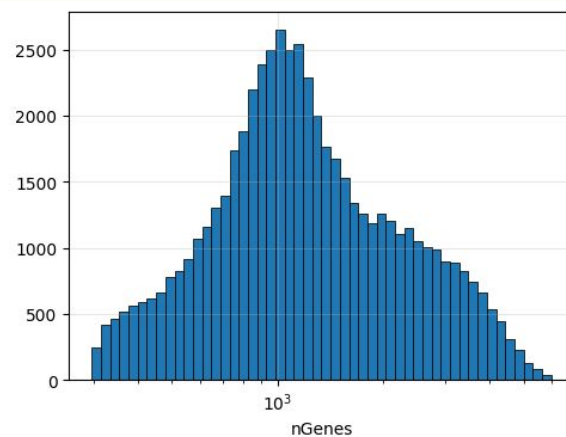
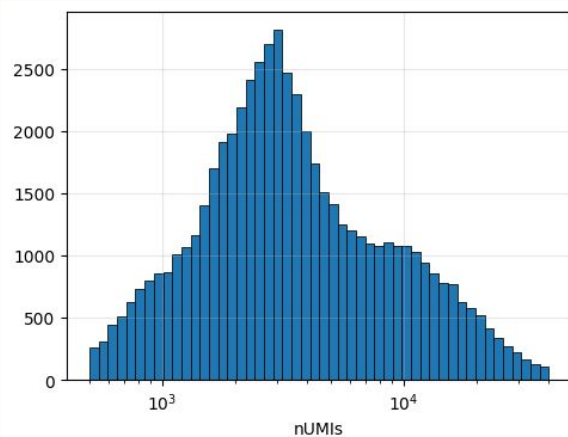
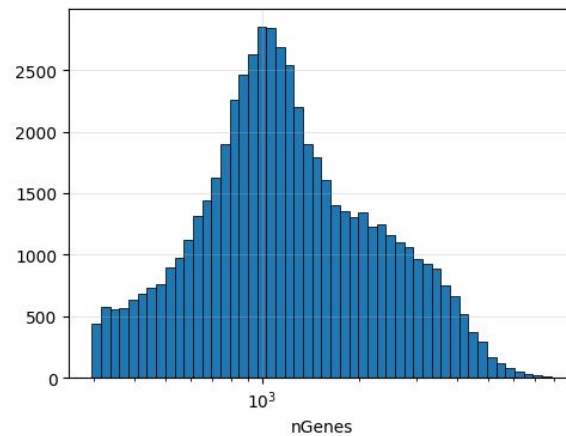
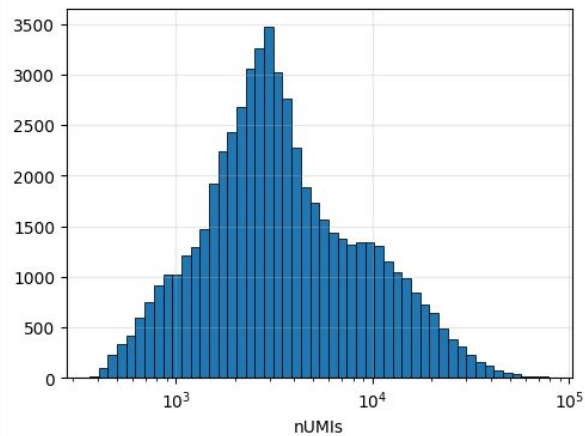
Выбранный датасет

Характеристики датасета: *Lung Parenchyma*

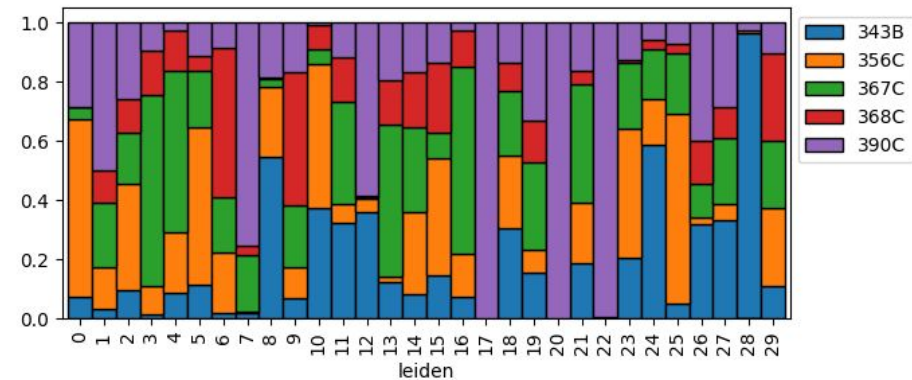
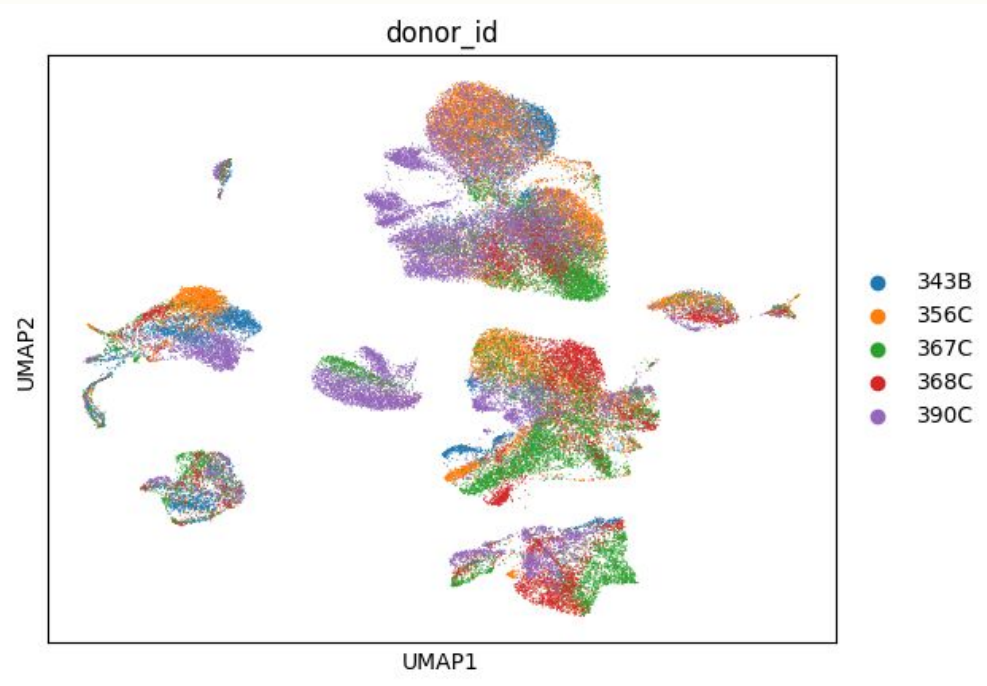
- **Ткань:** лёгочное паренхима (lung parenchyma)
- **Состояние:** нормальная ткань
- **Технология секвенирования:** 10x Genomics 3' v2
- **Организм:** Homo sapiens
- **Объём:** 57 019 клеток

<https://cellxgene.cziscience.com/collections/4d74781b-8186-4c9a-b659-ff4dc4601d91>

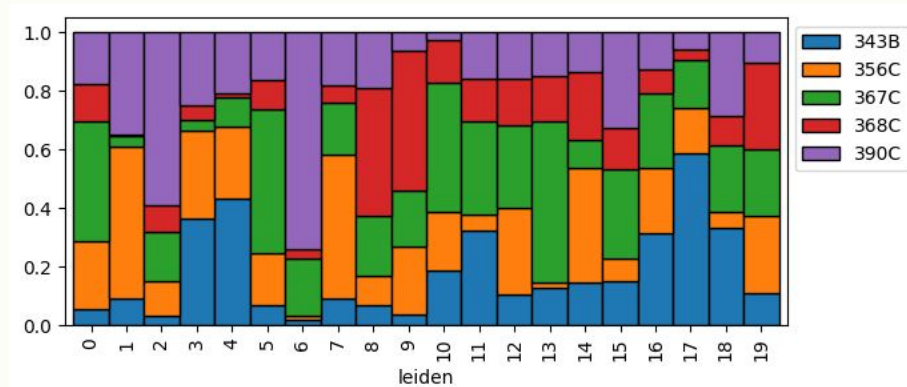
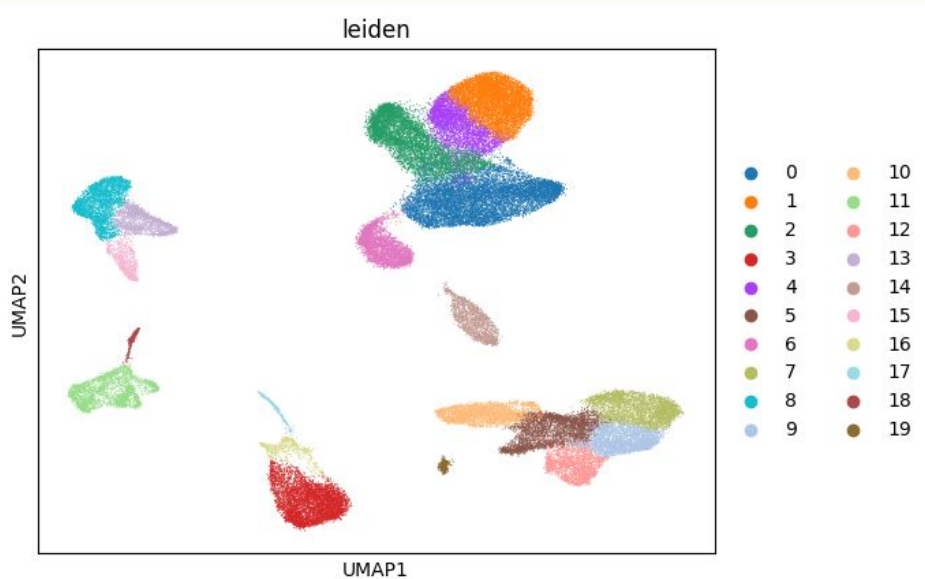
Предобработка



До батч-коррекции



После батч-коррекции



Был использован метод **Batch balanced kNN (bbknn)**

ARI: 0.62

NMI: 0.80

Accuracy: 0.71

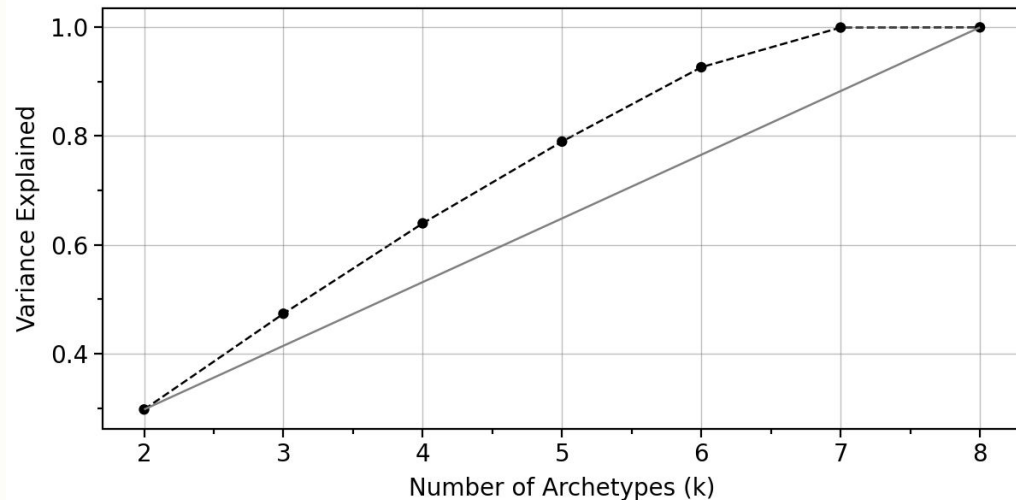


03

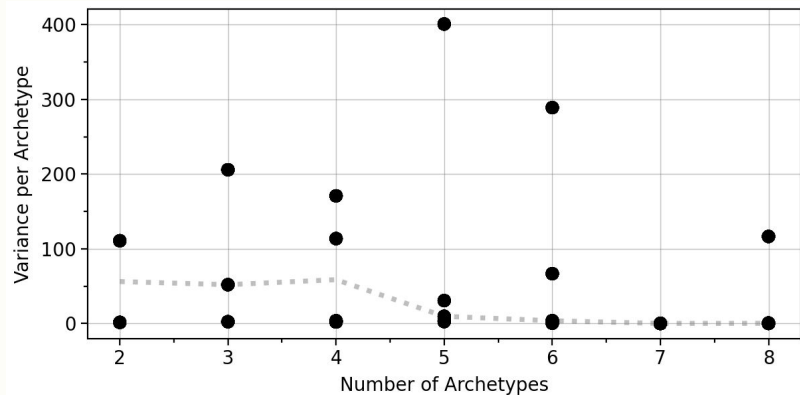
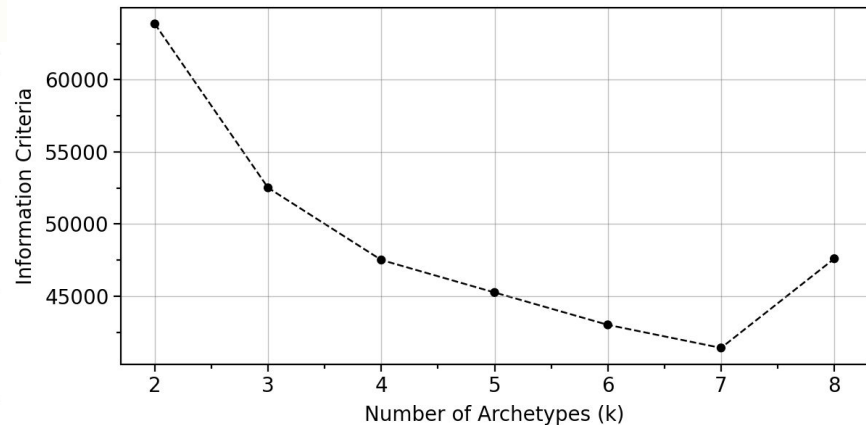
ParTipy

Архетип - всему голова

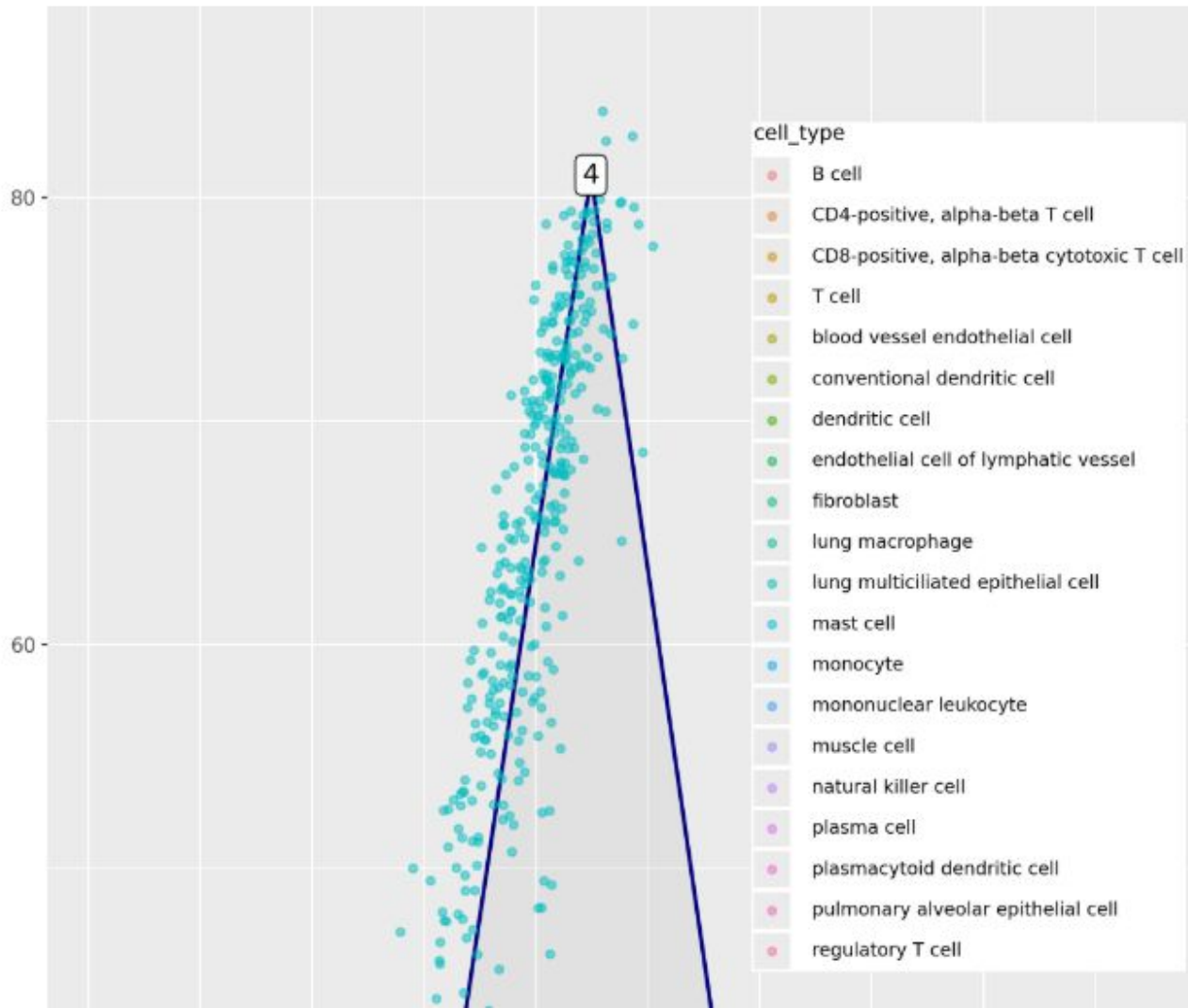
Выбор количества архетипов

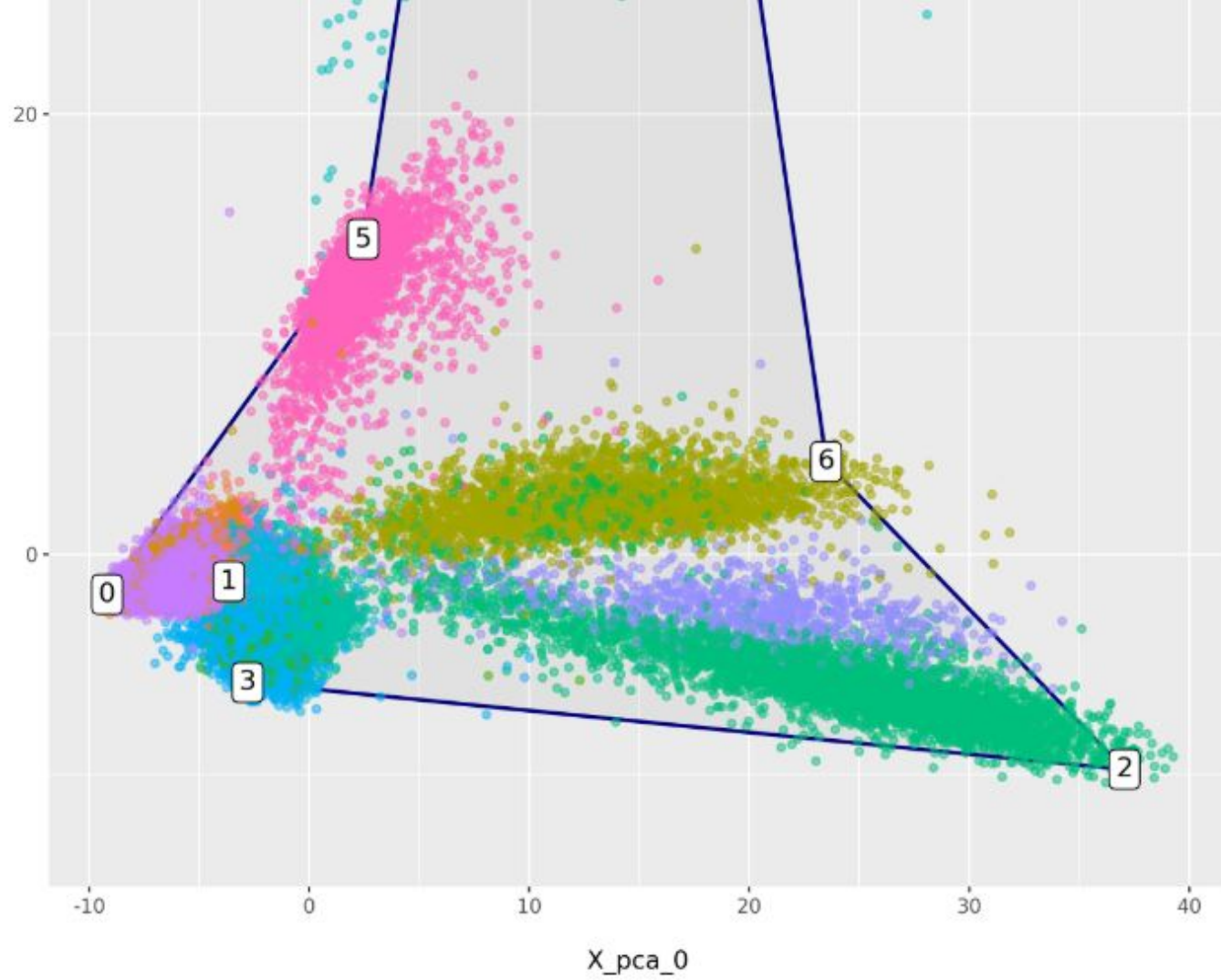


Было выбрано в дальнейшем
использовать **7 архетипов**

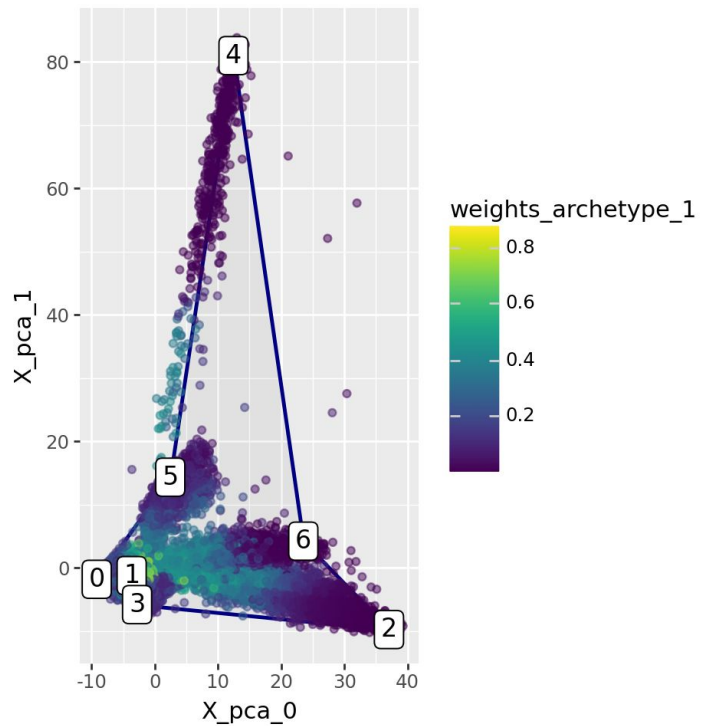
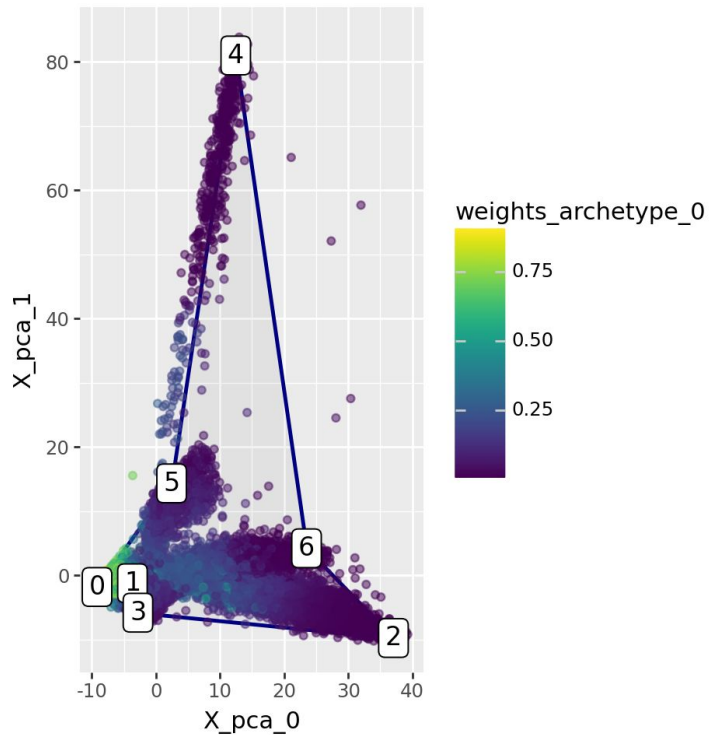




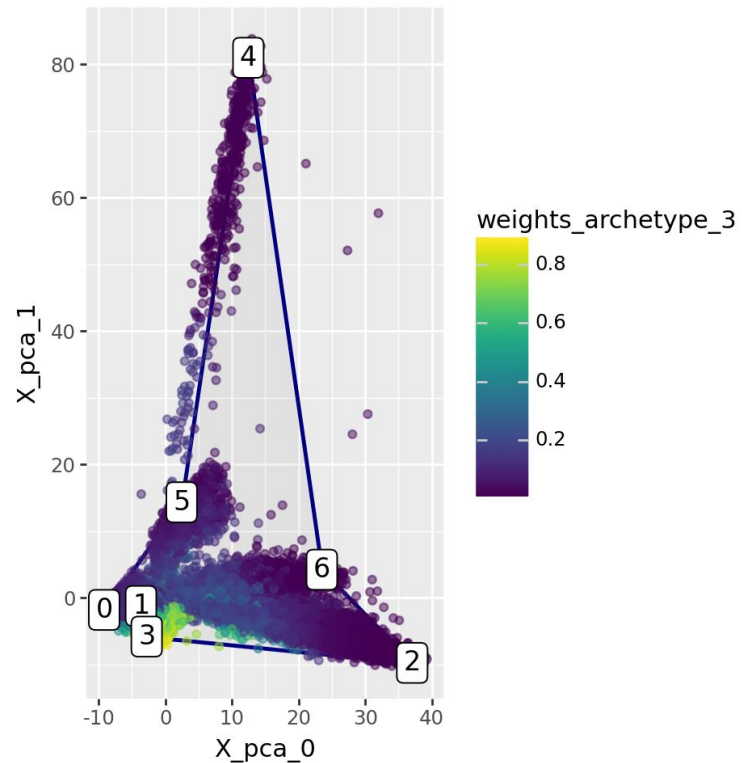
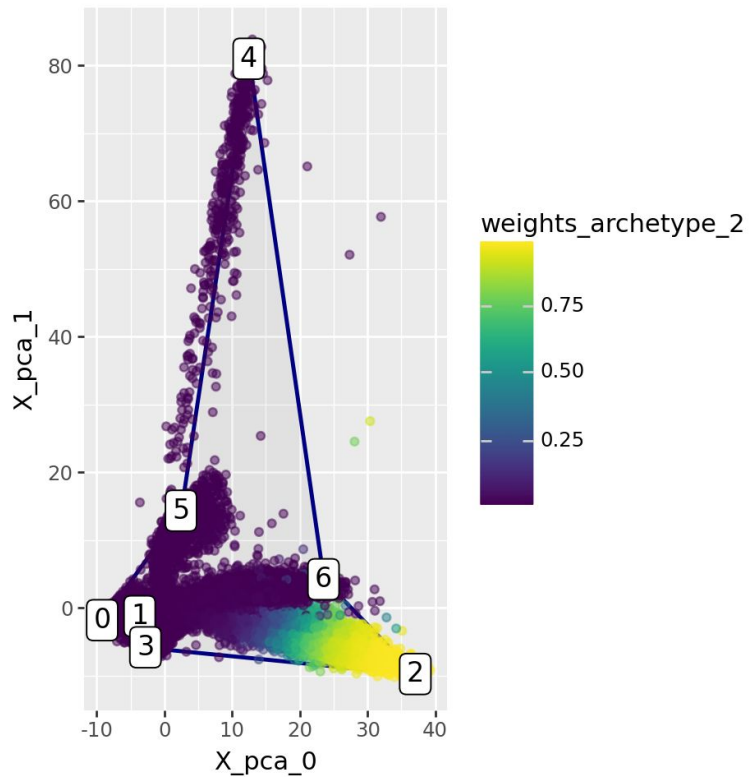




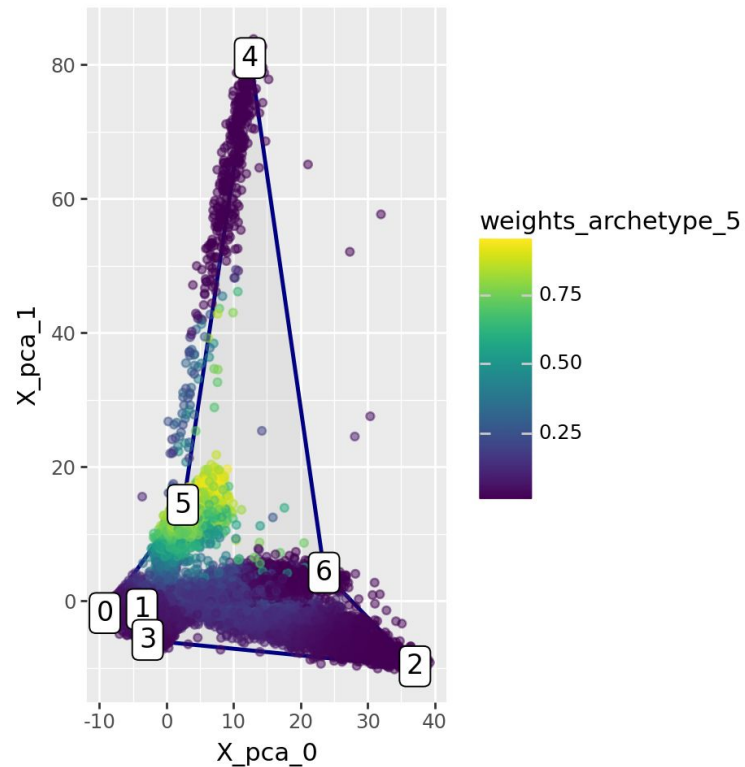
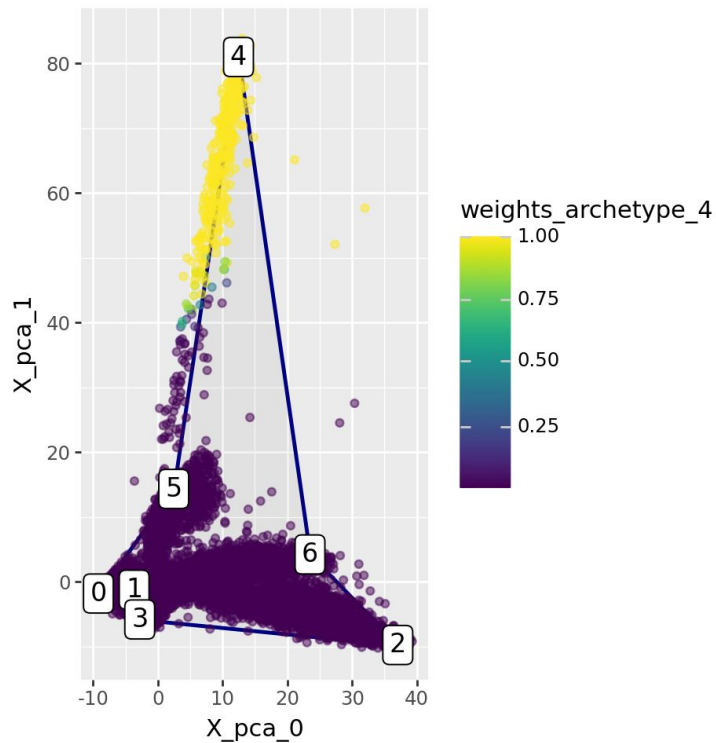
Характеристика архетипов



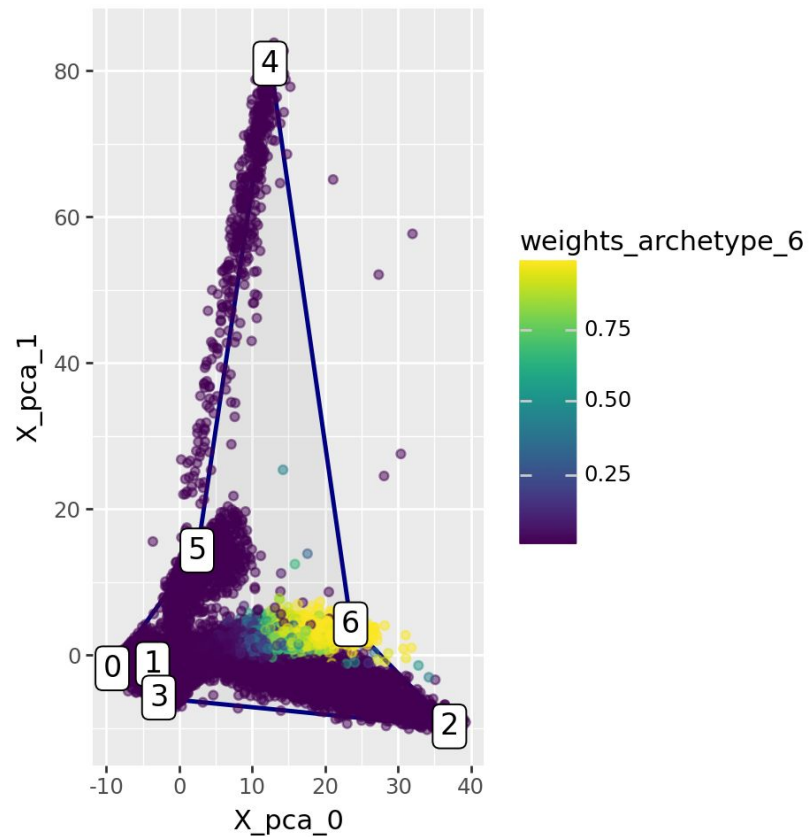
Характеристика архетипов



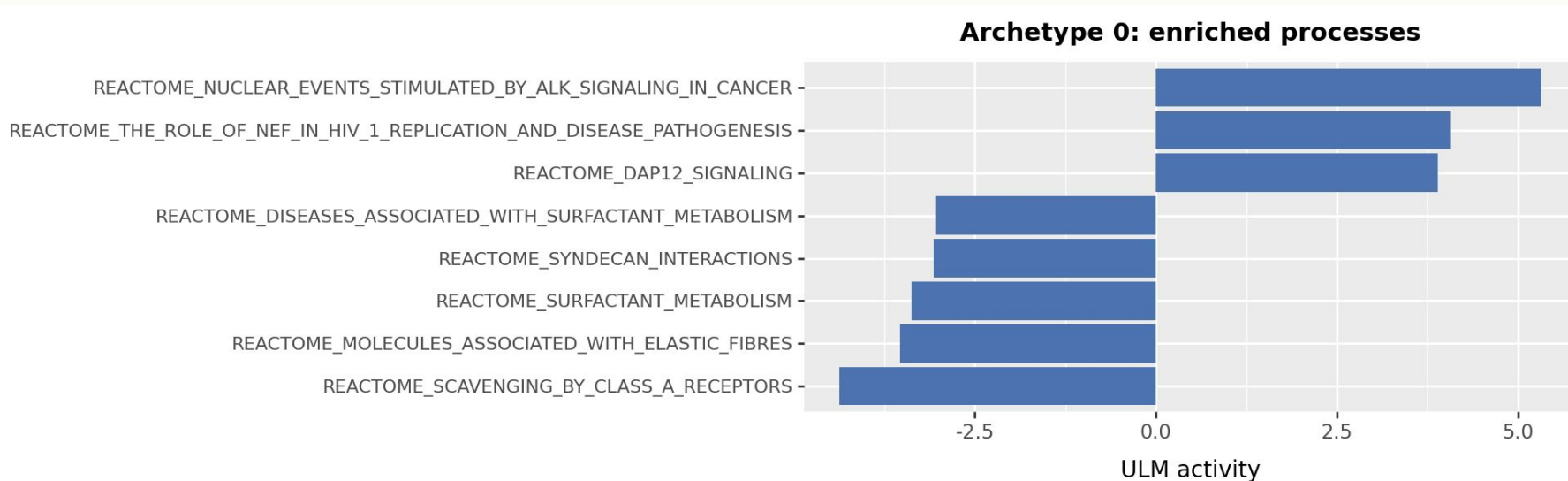
Характеристика архетипов



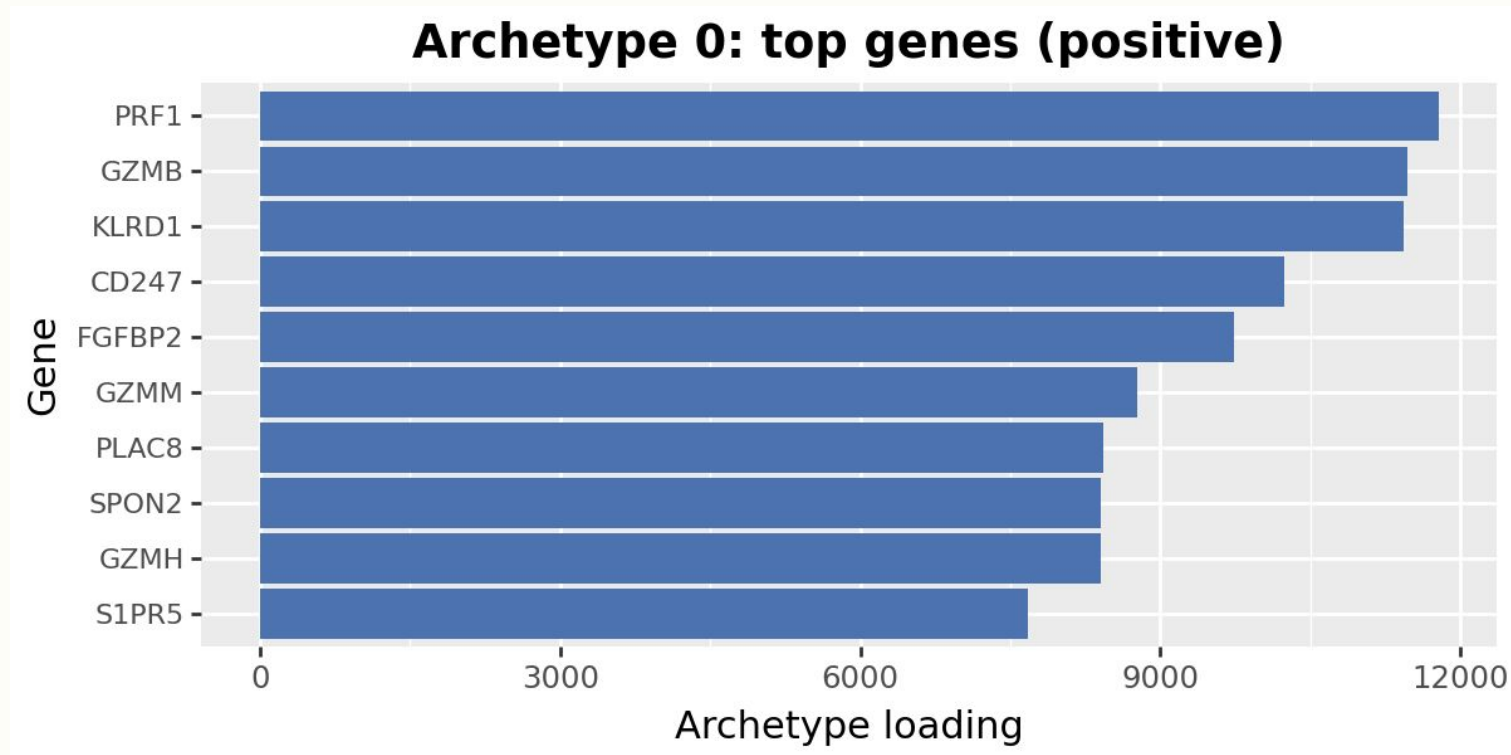
Характеристика архетипов



Характеристика 0 архетипа

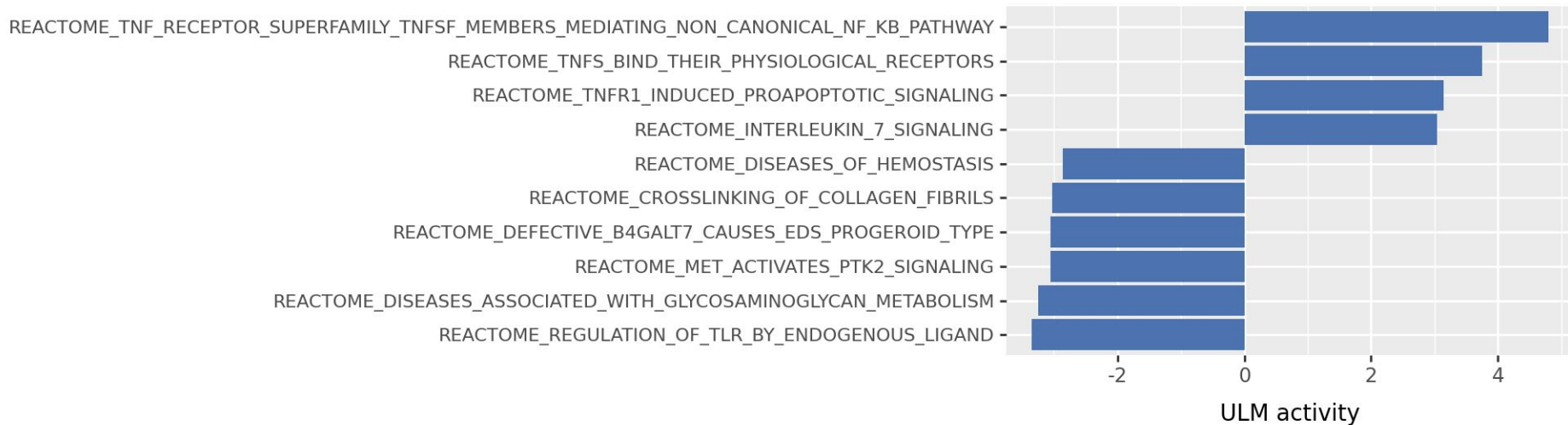


Характеристика 0 архетипа



Характеристика 1 архетипа

Archetype 1: enriched processes



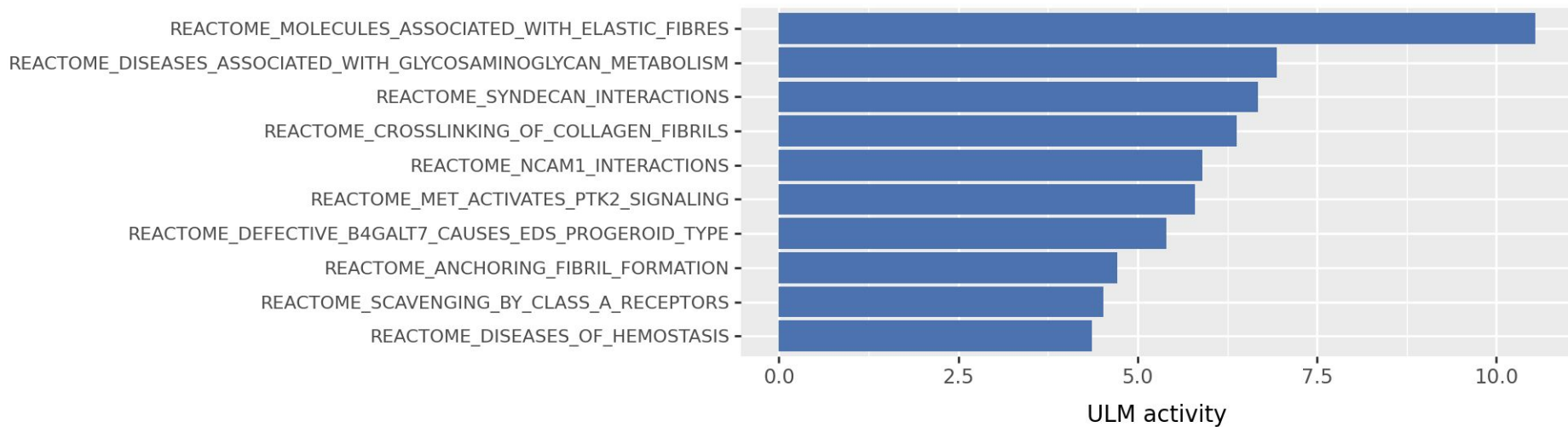
Характеристика 1 архетипа

Archetype 1: top genes (positive)



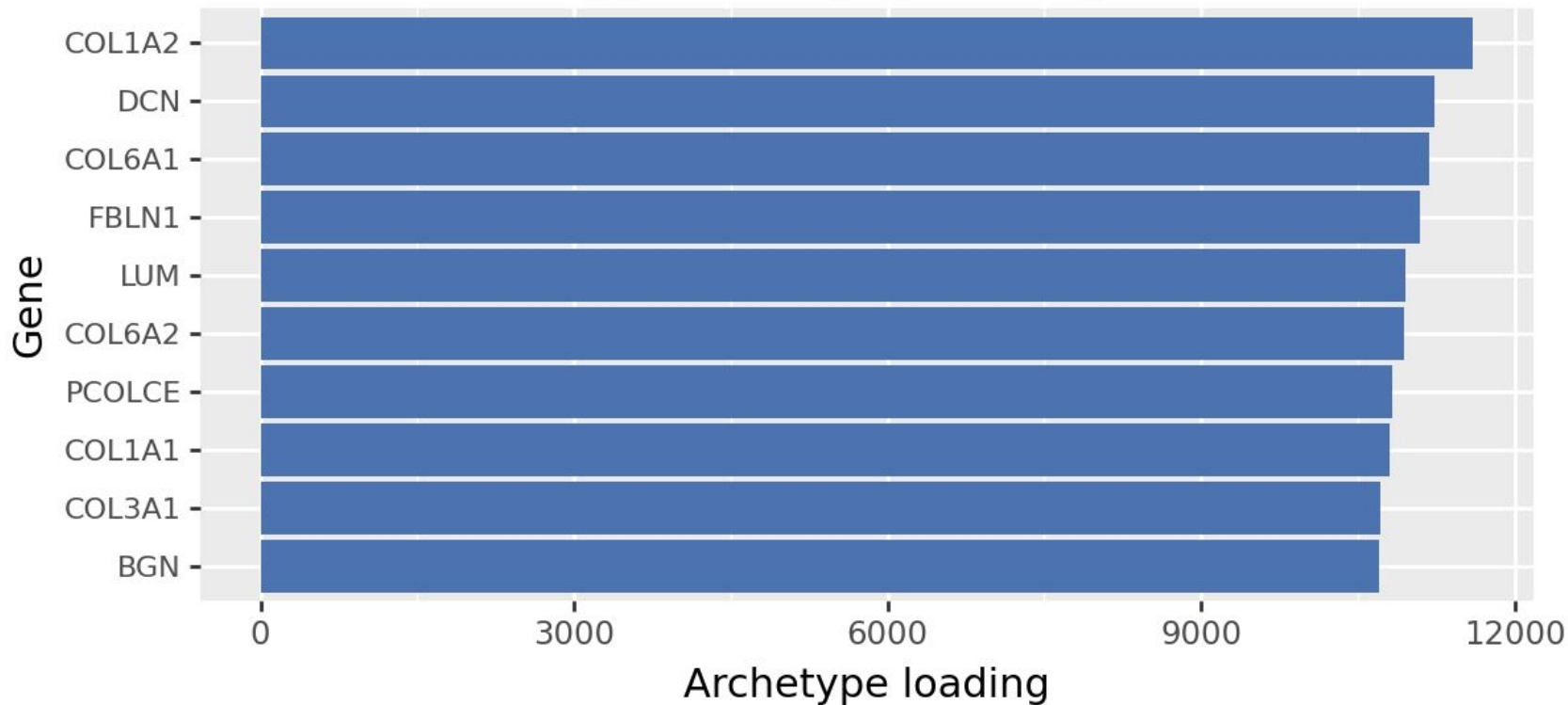
Характеристика 2 архетипа

Archetype 2: enriched processes



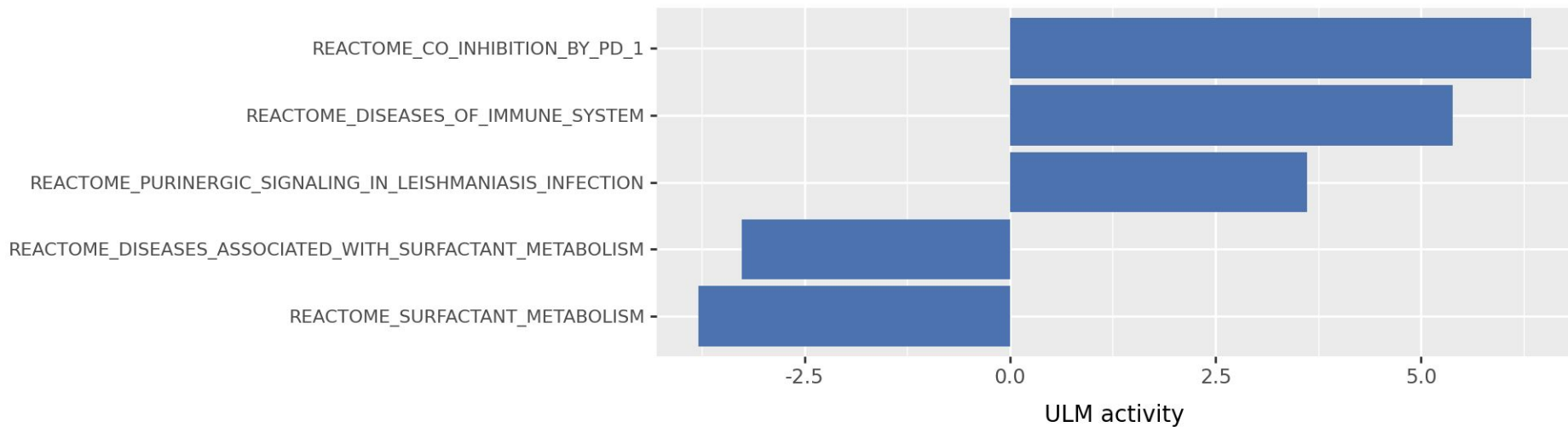
Характеристика 2 архетипа

Archetype 2: top genes (positive)

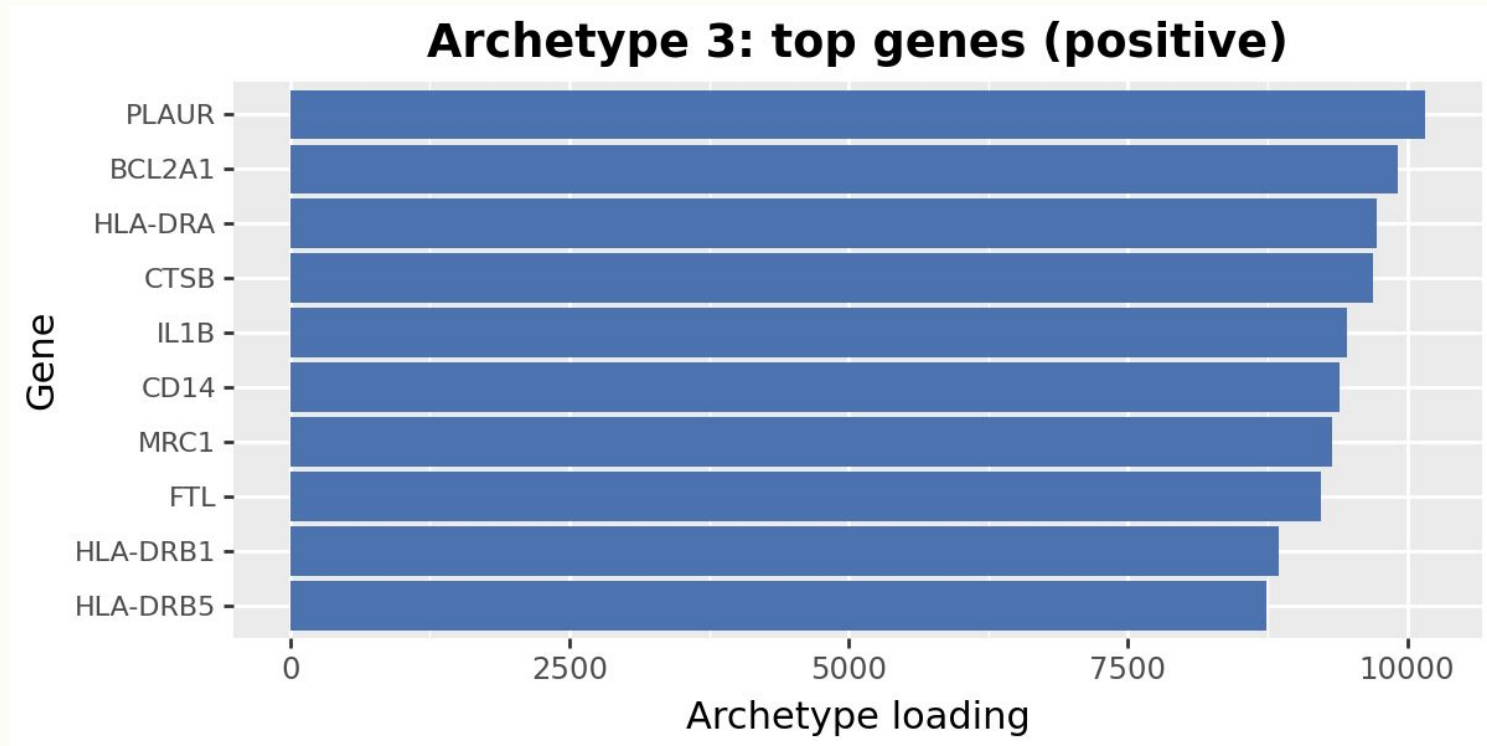


Характеристика 3 архетипа

Archetype 3: enriched processes

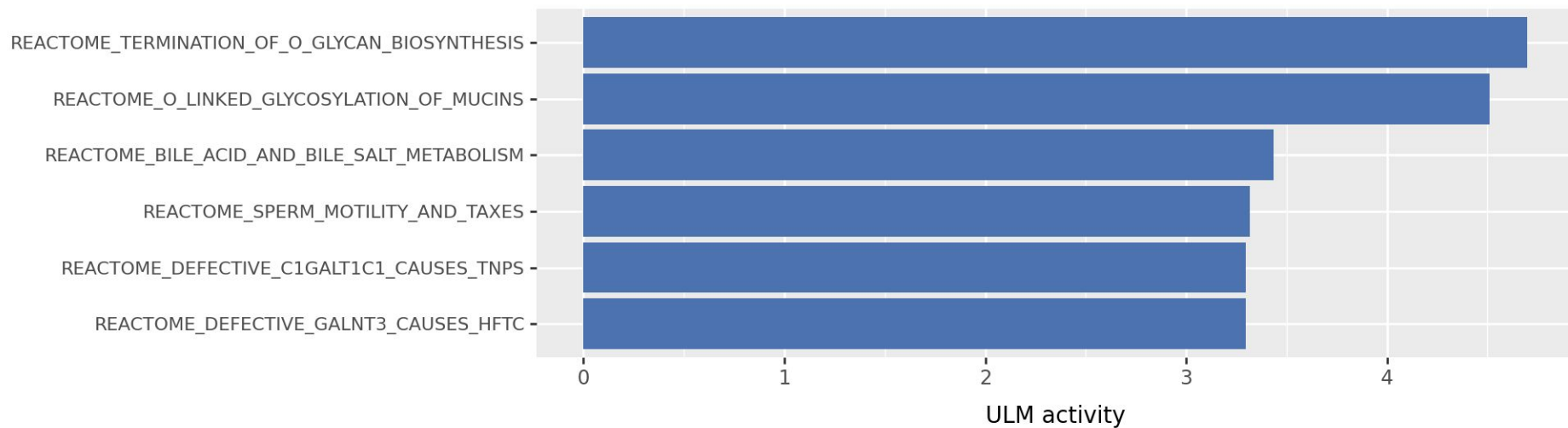


Характеристика 3 архетипа

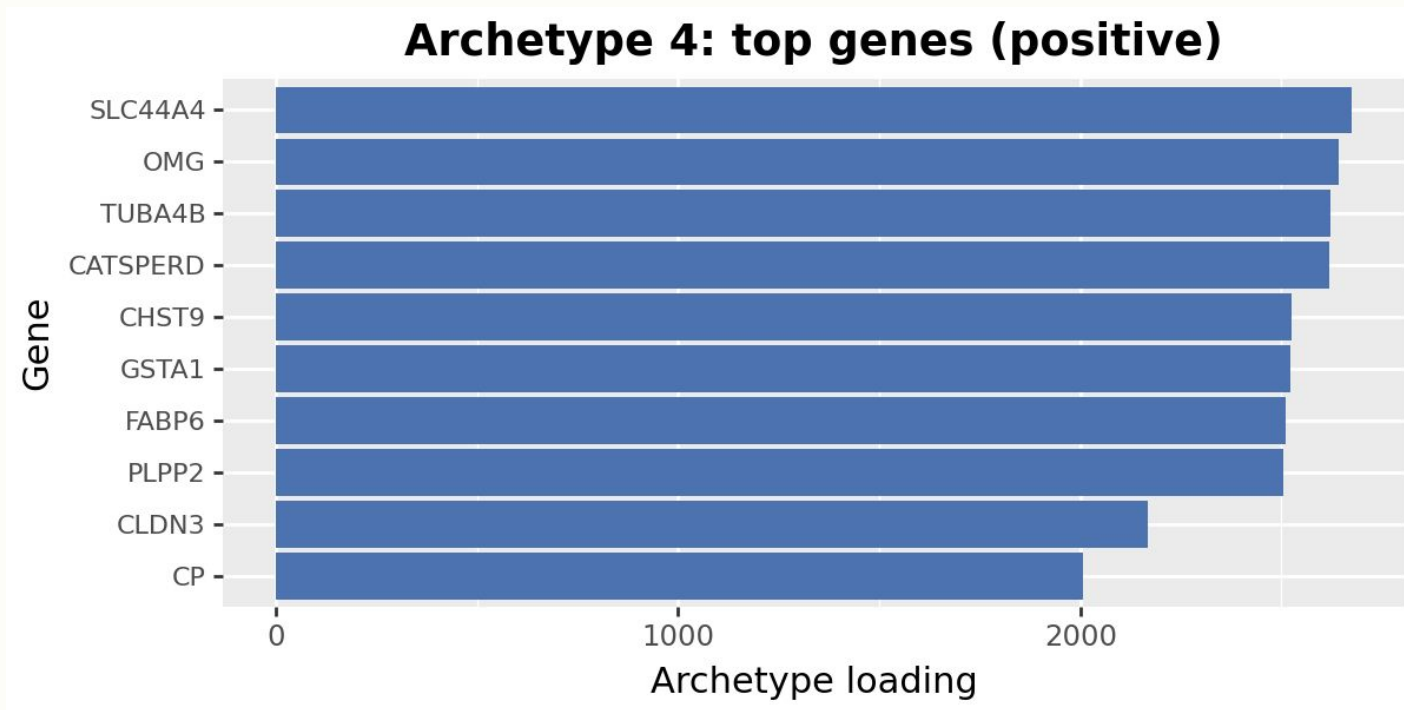


Характеристика 4 архетипа

Archetype 4: enriched processes

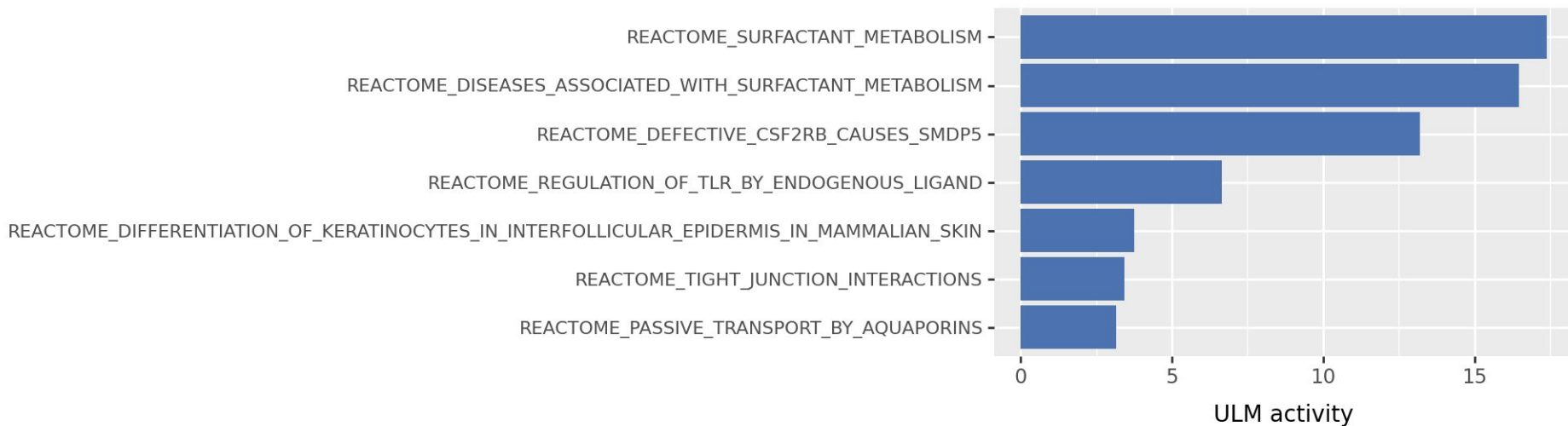


Характеристика 4 архетипа

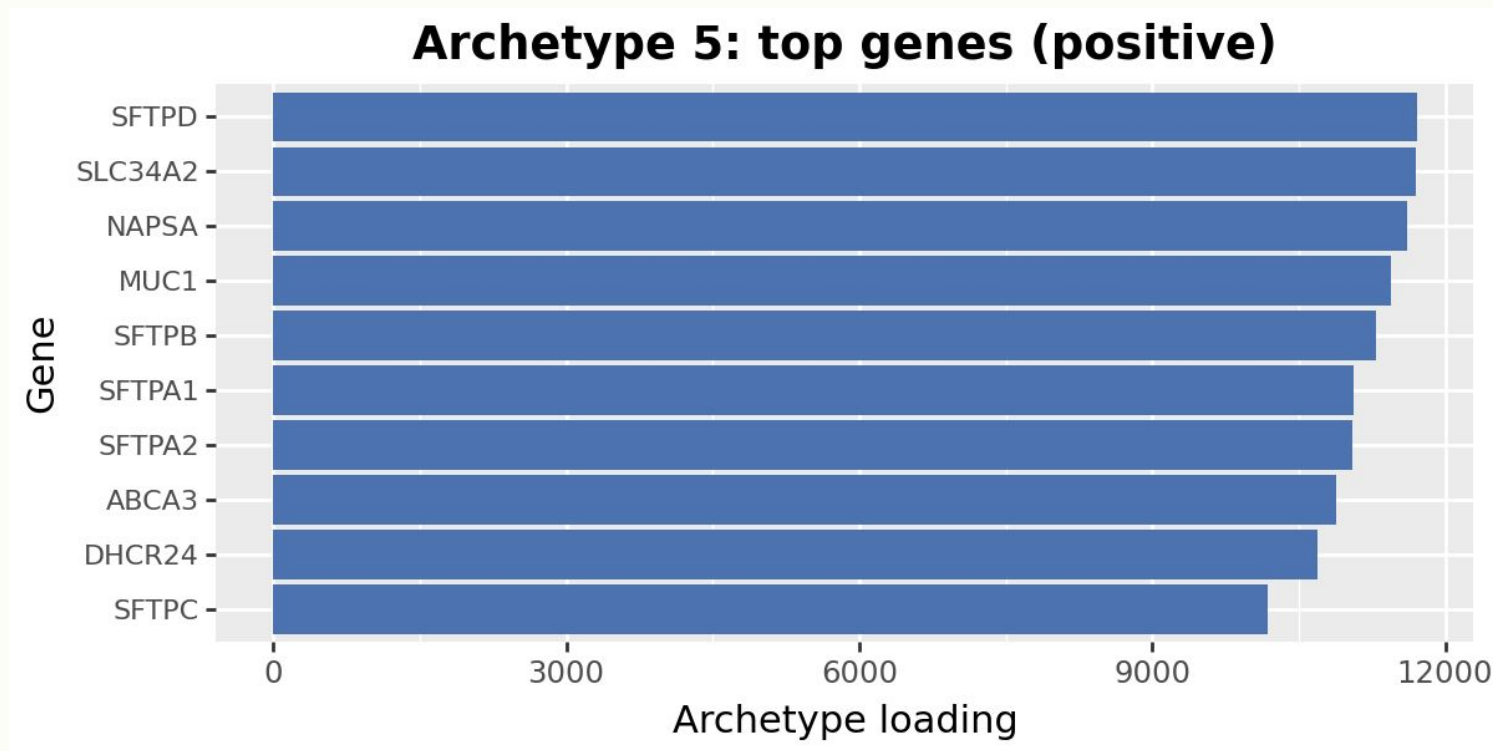


Характеристика 5 архетипа

Archetype 5: enriched processes

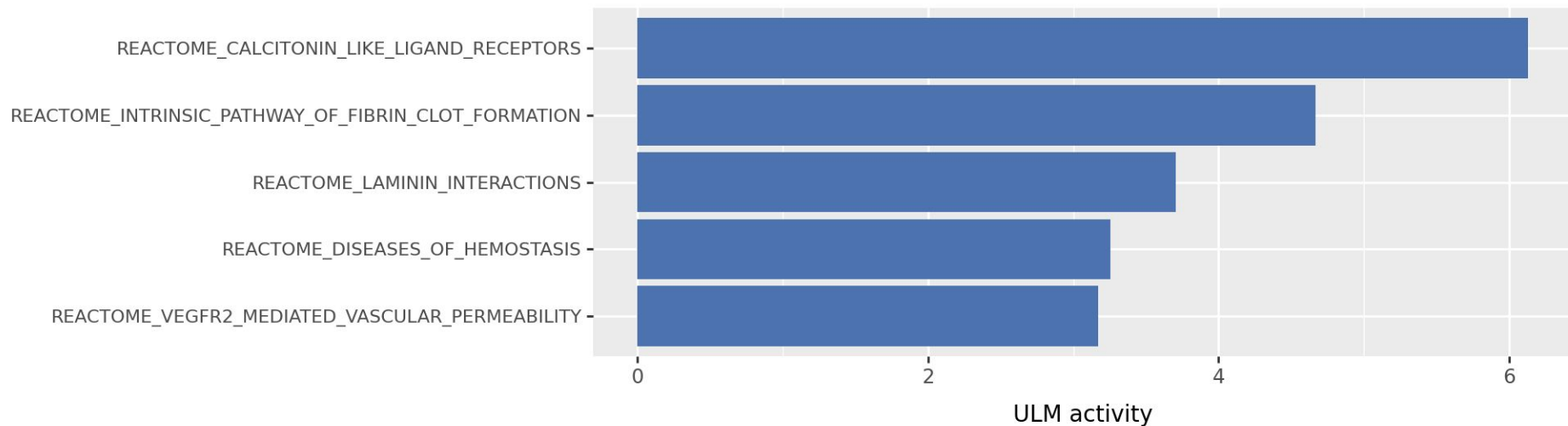


Характеристика 5 архетипа

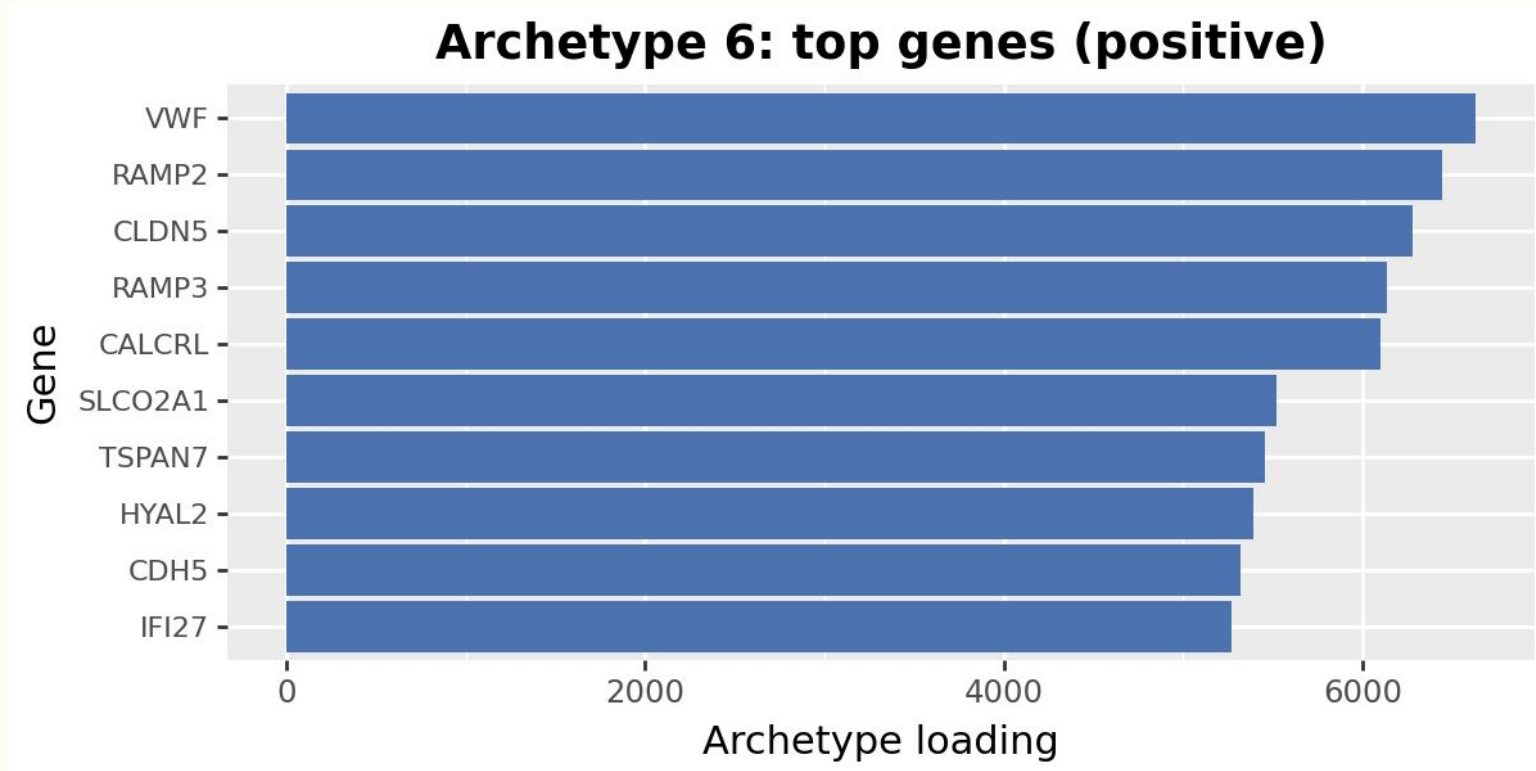


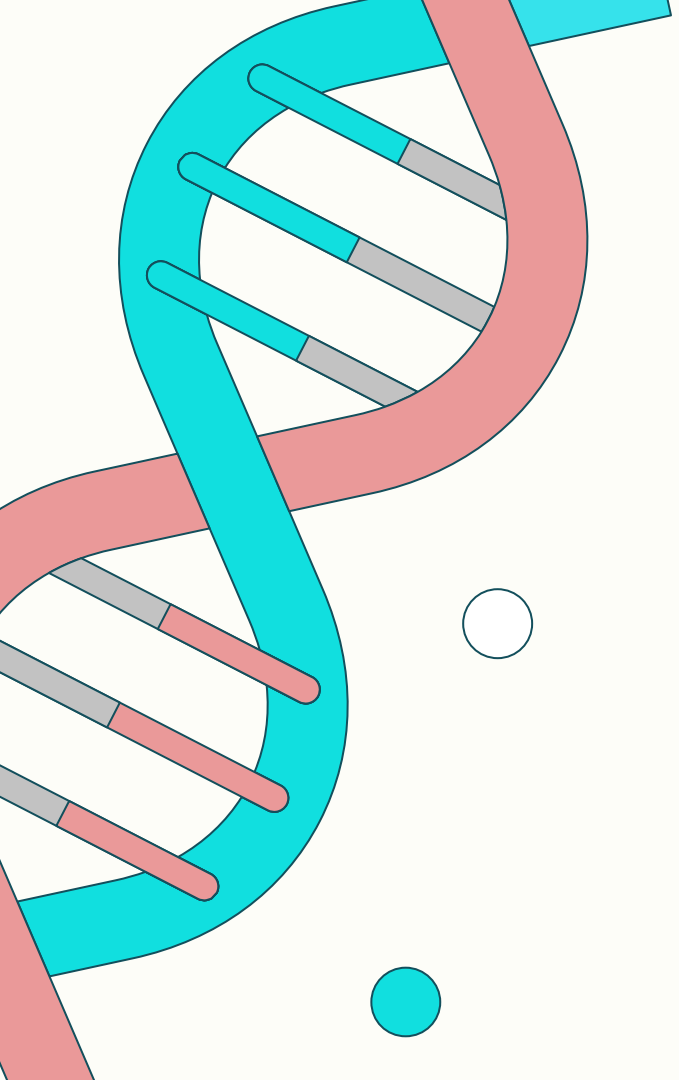
Характеристика 6 архетипа

Archetype 6: enriched processes



Характеристика 6 архетипа





**Thank
you for your
interest!**

