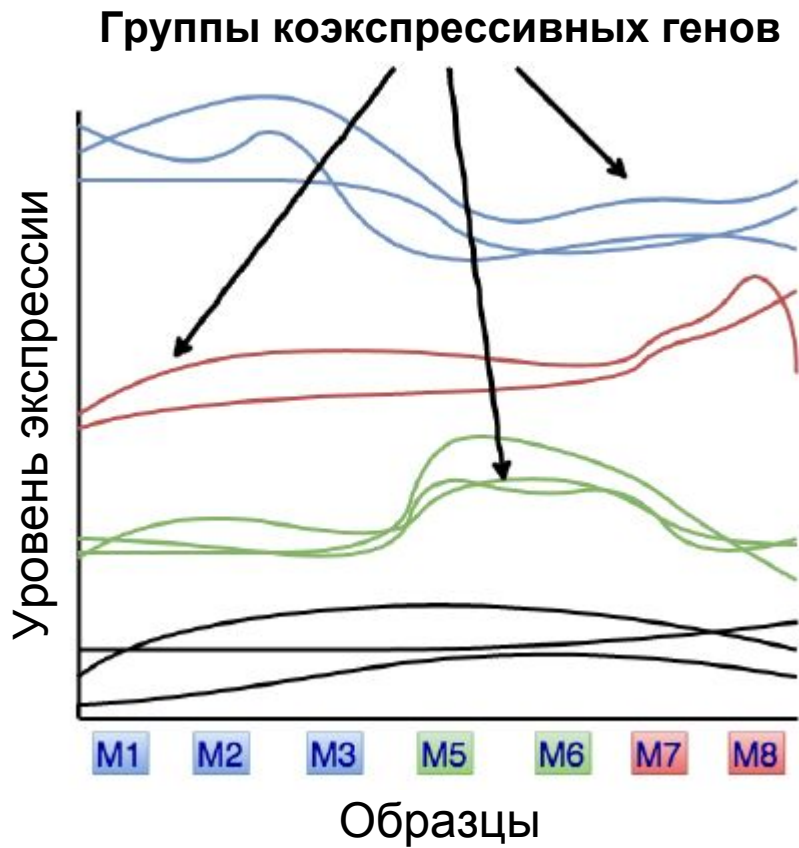
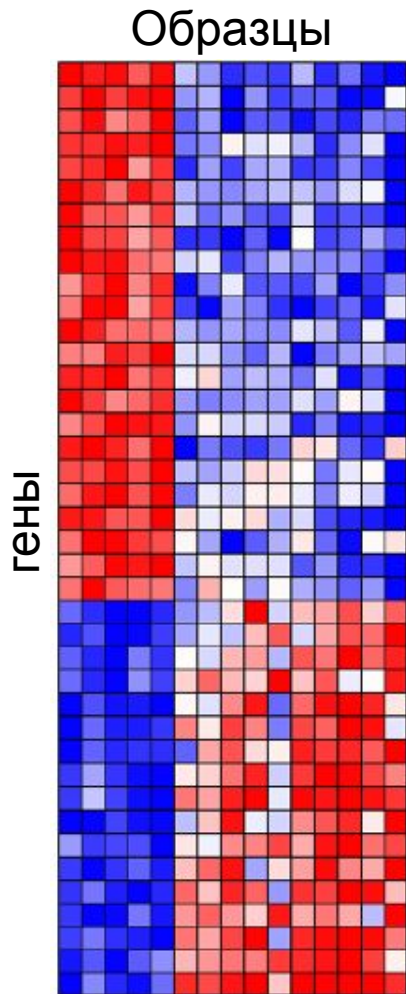


# Разработка и улучшение поисковых стратегий в системе GeneQuery

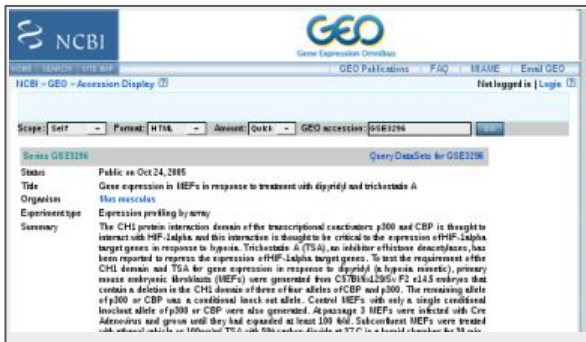
Научный руководитель: Предеус Александр (Институт биоинформатики)  
Студенты: Богданова Валерия, Плеханова Елена

# Дифференциальная экспрессия



# Анализ представленности

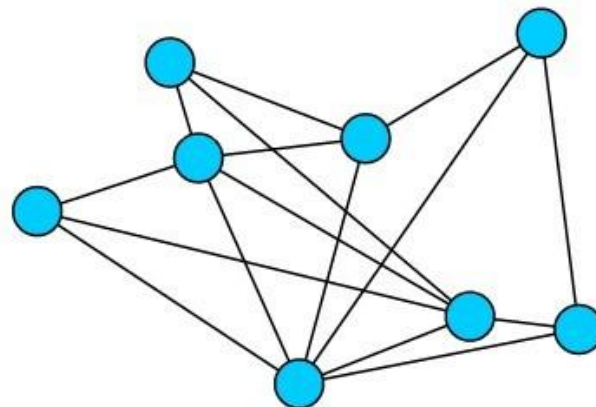
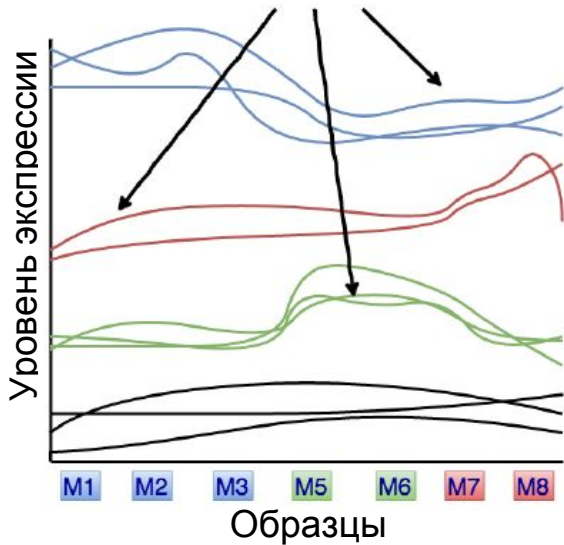




# Тест Фишера

	в запросе	вне запроса	всего
в модуле	$k$	$n - k$	$n$
вне модуля	$D - k$	$N - D - n + k$	$N - n$
всего	$D$	$N - D$	$N$

## Группы коэкспрессивных генов



# Транскриптомная сеть



## Глобальные цели:

- Автоматическая классификация по категориям
- Улучшение кластеризации генов на подгруппы коэкспрессивных
- Удобная и информативная визуализация кластеризации

# Задачи

- Найти способ автоматической аннотации образцов (Лена) и реализовать его (Лера)
- Визуализировать кластеризацию по образцам с помощью PCA (Лера и Лена)
- Определить закономерности кластеризации (Лера)
- Придумать метрику, хорошо отражающую расстояния между образцами и учесть ее при кластеризации GeneQuery (наши потомки)

# Аннотация образцов

Чтобы выявить ошибку кластеризации, надо посмотреть на кластеризацию...

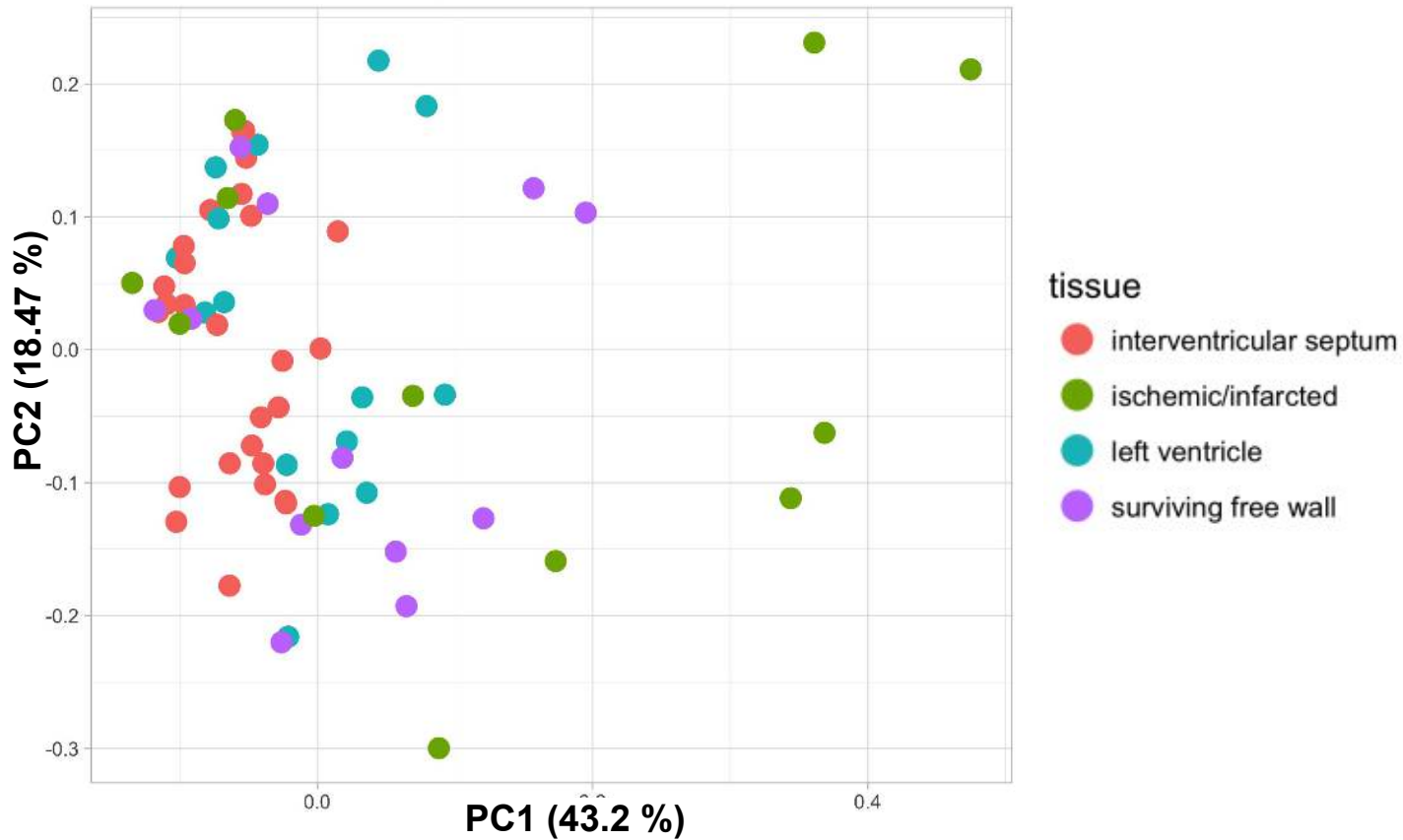
## База данных GEO

Series	77607
DataSets (with annotations)	4348
Samples	2022591



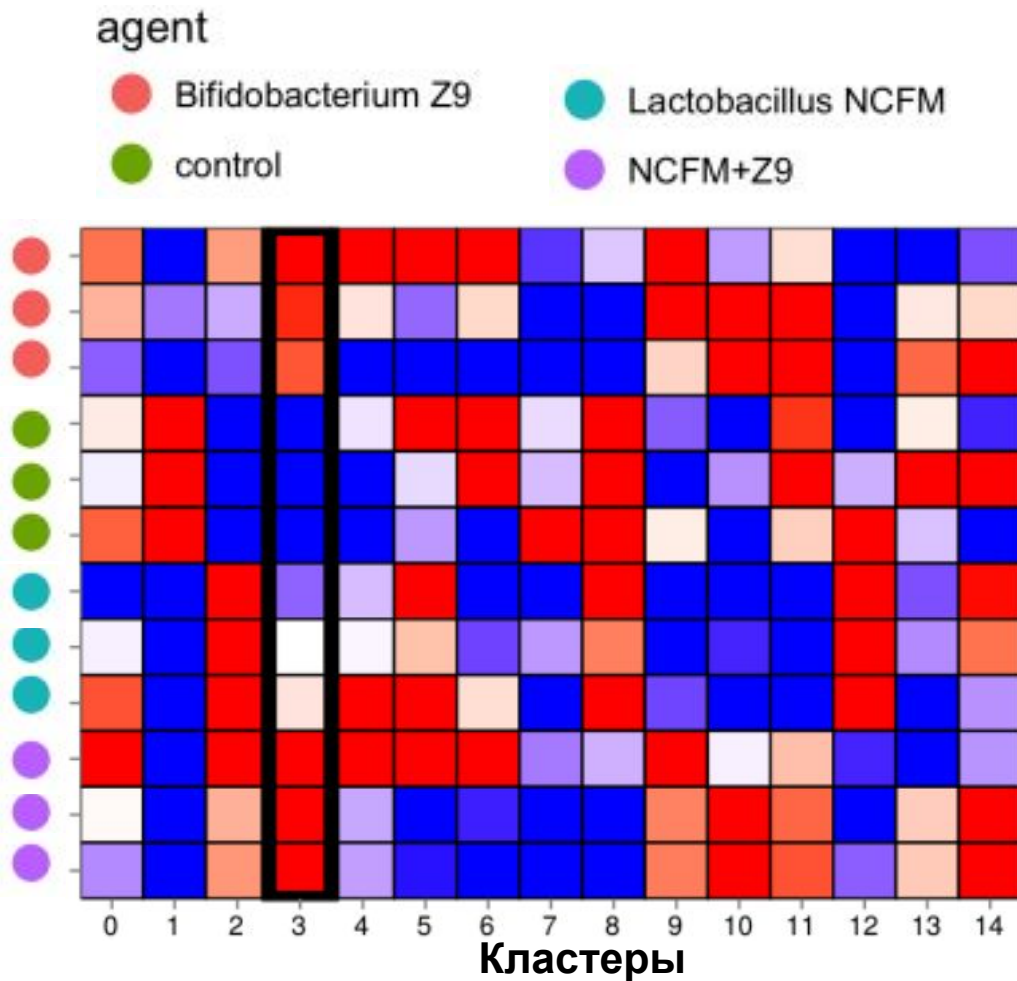
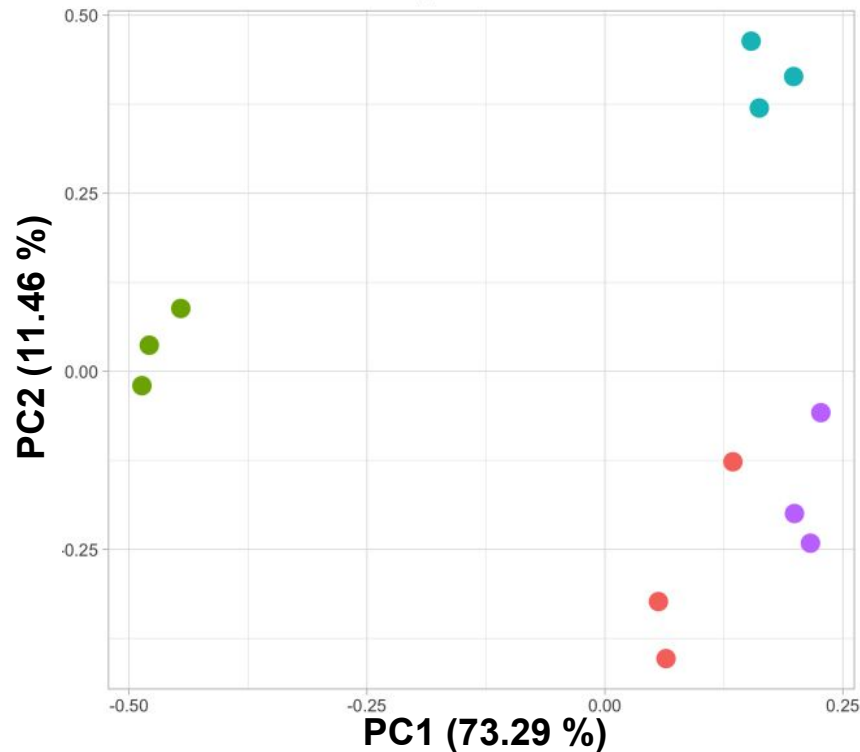
# Плохой случай

GSE4648 GDS2329 PCA plot



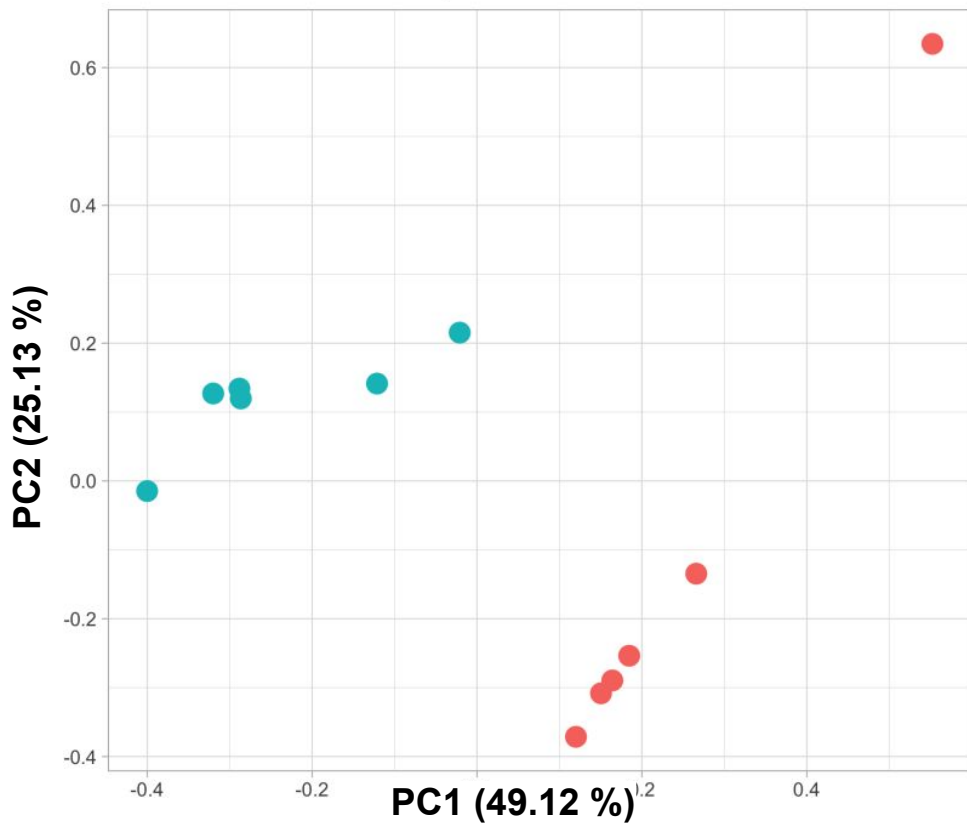
# Хороший случай

GSE20302 GDS3813 PCA plot



# Случай с выбросом

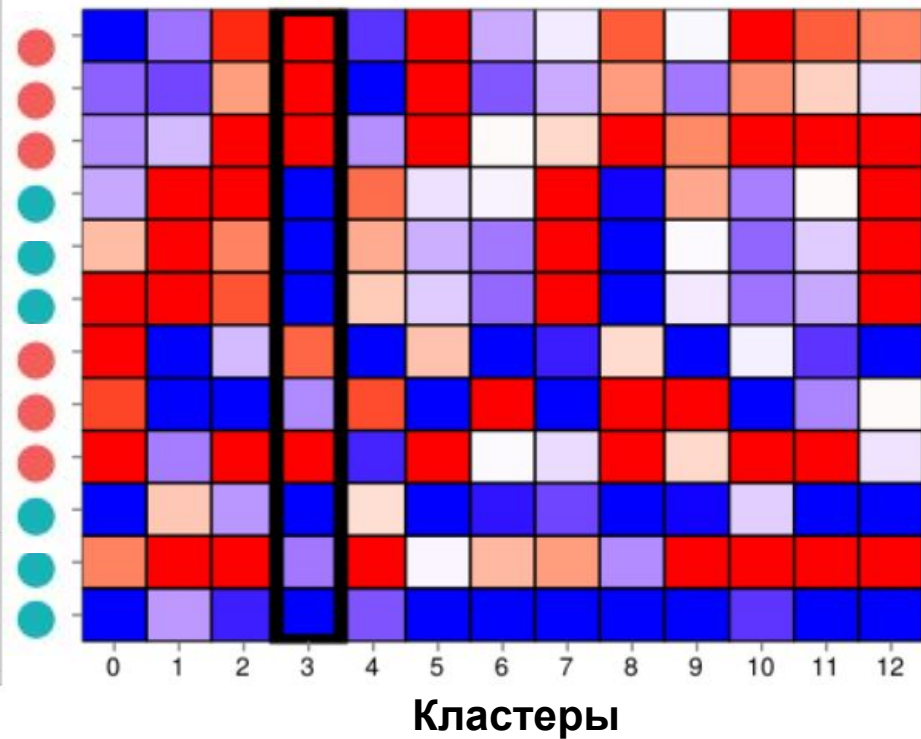
GSE6623 GDS2720 PCA plot



cell.type

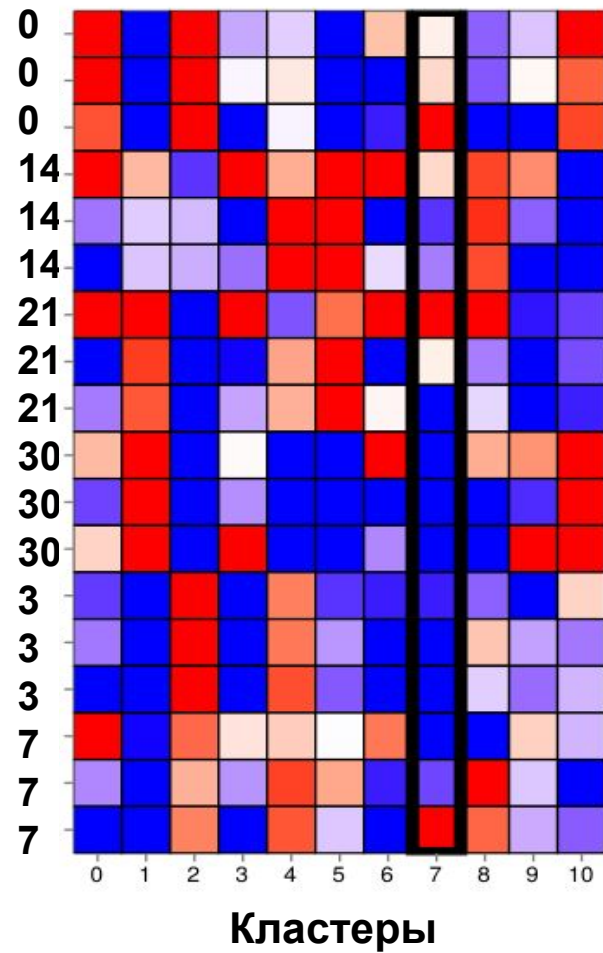
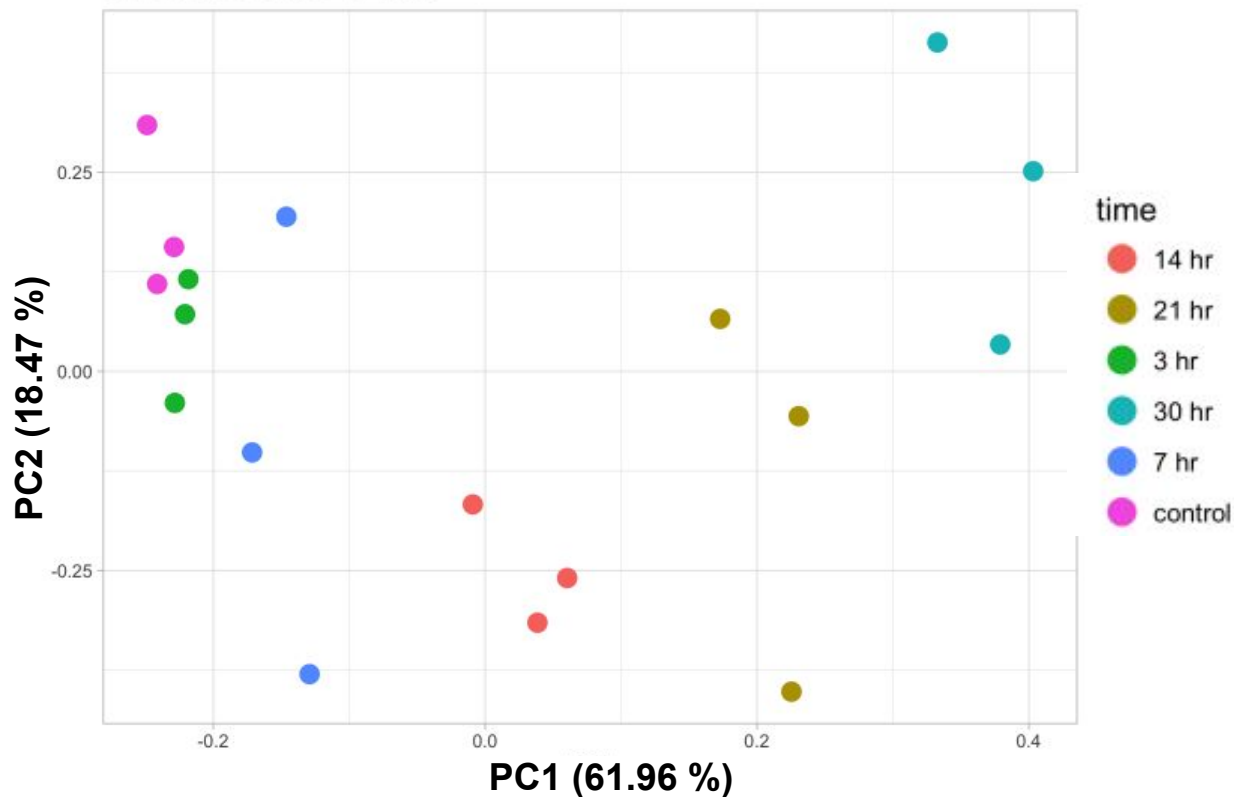
● LSK

● myeloid progenitor



# Случай с time series

GSE628 GDS568 PCA plot



# Выводы

Были выявлены следующие типы кластеризации:

- все хорошо
- есть выбросы
- все плохо

Также замечено хорошее разделение time series.

Спасибо за терпение!