ДЗ14, Сорокина Тамара, Б06-204

1. Подготовка данных и фильтрация

В рамках анализа были использованы сырые данные по экспрессии генов (файл raw_counts_ici_samples.tsv) и информация об ответе на терапию (файл meta_responses.tsv). Перед началом анализа из датасета были удалены гены с низким уровнем экспрессии. Для этого была рассчитана медианная экспрессия по каждому гену, и все гены, медиана которых оказалась ниже 1-го квартиля распределения медианных значений, были исключены из дальнейшего анализа. Это позволило сосредоточиться на более информативных, высокоэкспрессированных генах.

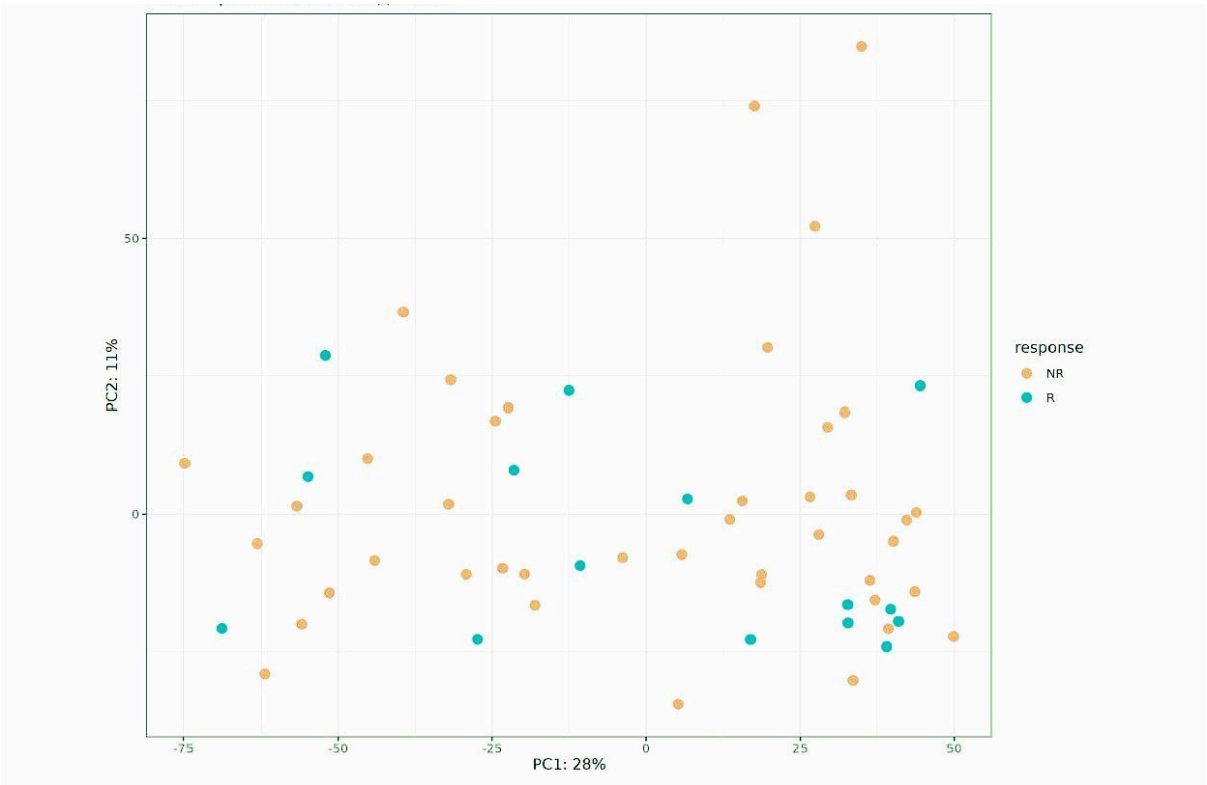
2. Нормализация данных

Для нормализации сырых каунтов была использована функция DESeq2::vst(), реализующая вариационно-стабилизирующее преобразование. Это позволяет устранить зависимость дисперсии от среднего значения, что критично для корректной визуализации и последующего анализа.

3. PCA-анализ

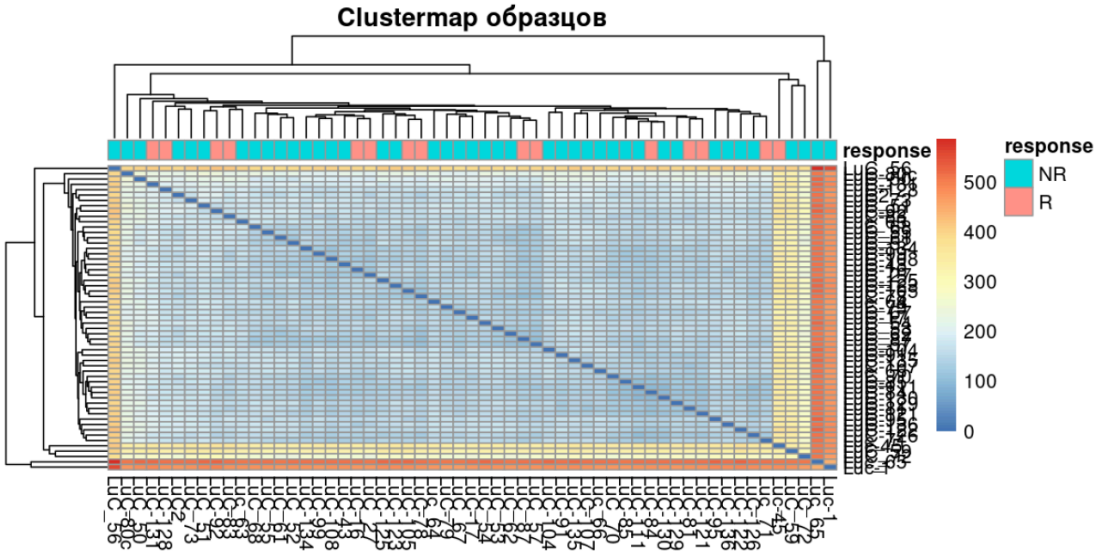
На основе нормализованных данных была построена диаграмма главных компонент (PCA). Главные компоненты (PC1 и PC2) объясняют 28% и 11% дисперсии соответственно. Образцы на графике PCA не группируются чётко по ответу на терапию (response), что указывает на отсутствие выраженного различия между группами R (responders) и NR (non-responders) на этом уровне визуализации.

Однако, четыре образца, расположенные в верхней части PCA-графика, существенно отклоняются от основной массы точек. Эти образцы могут быть потенциальными выбросами либо представлять собой биологически обоснованные особенности, например, специфическую подгруппу пациентов, не ответивших на терапию. Для окончательного вывода потребуется дополнительная биологическая или клиническая информация.



4. Clustermap (кластеризация образцов)

Также был построен кластерный тепловой график (clustermap) на основе попарных расстояний между образцами. По результатам кластеризации не наблюдается чёткой группировки образцов в соответствии с ответом на терапию. Это дополнительно подтверждает выводы, сделанные по PCA: различия между группами R и NR не проявляются явно на уровне глобальной экспрессии всех отобранных генов.



5. Анализ дифференциальной экспрессии

После фильтрации и нормализации данных был проведён анализ дифференциальной экспрессии между группами R и NR с использованием DESeq2. Были рассчитаны значения log2 Fold Change и p-value для каждого гена.

Среди всех анализируемых генов, 675 генов показали статистически значимую дифференциальную экспрессию (с поправкой на множественное сравнение, FDR < 0.05 и |log2FoldChange| > 1). Эти гены могут быть потенциально биологически значимыми и требовать дальнейшего функционального анализа.

Вывод

Анализ показал, что по результатам PCA и кластеризации образцы не разделяются чётко по ответу на терапию. Однако были выявлены 675 генов с достоверной дифференциальной экспрессией между группами, что может указывать на молекулярные различия и требует дальнейшего исследования.