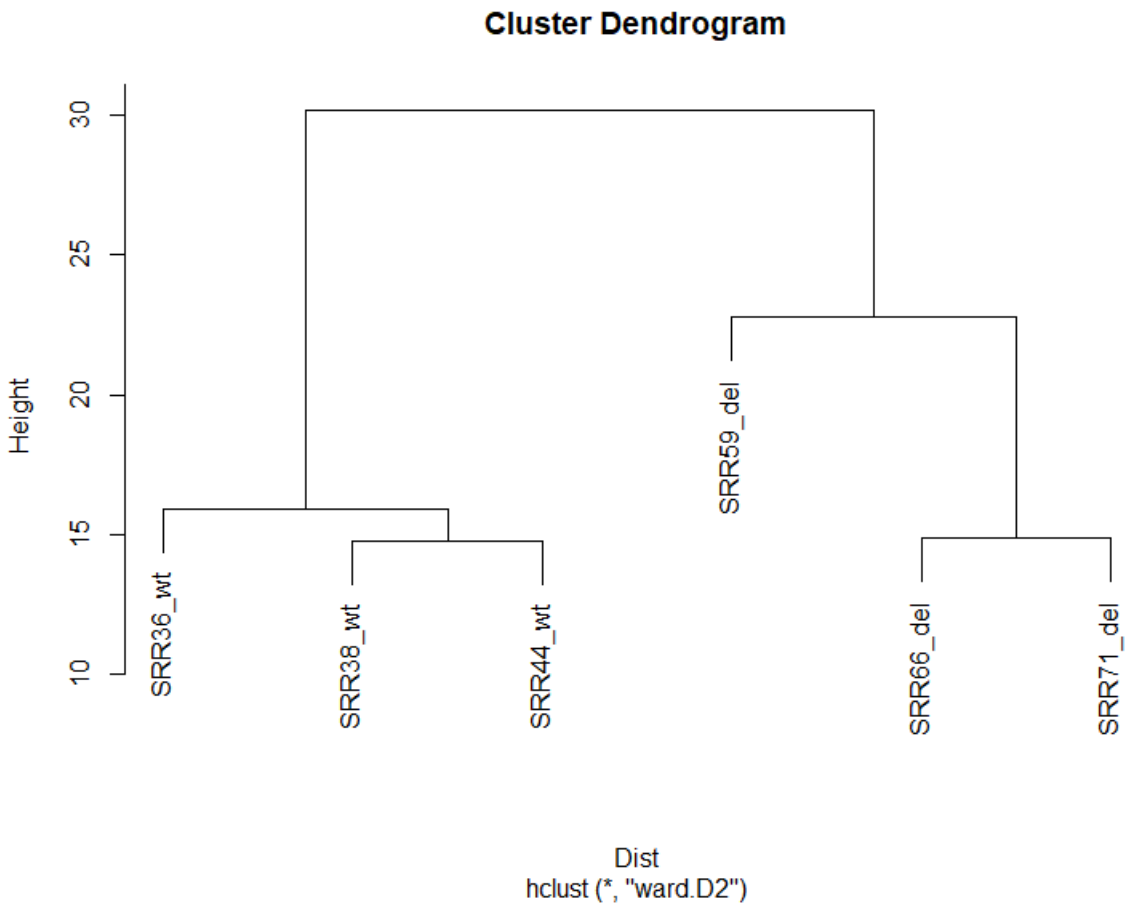
1. Объединила информацию о прочтениях в единую таблицу формата .csv (counts_Reshetnikova.csv)

1.2. Для того, чтобы оценивать дифф. экспрессию данные необходимо нормализовать. Для этого подготовила матрицу, сделала проверку на отсутствие прочтений, оценила нормализацию путём построения графика норм. данных с использованием DESeq2. Дальше работаем с нормализованными данными.

2. Проверка сходимости повторностей с помощью:

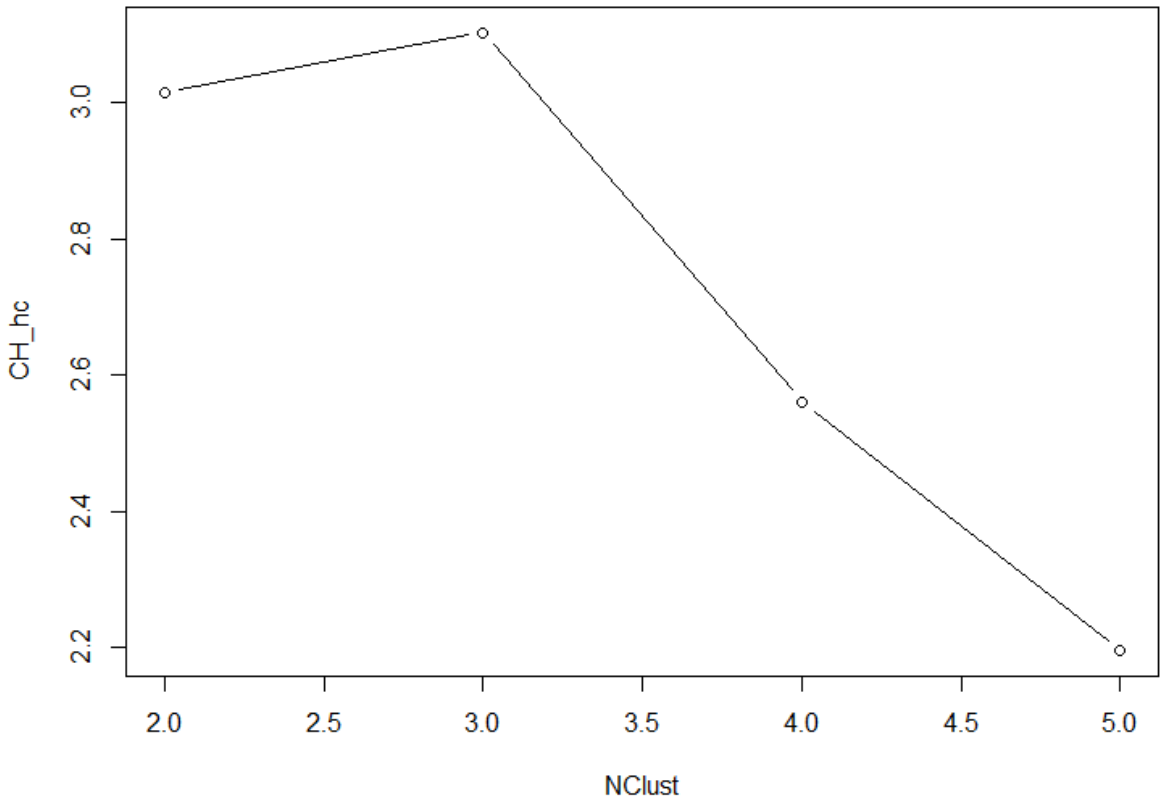
а) Кластерного анализа

Сперва строим кластерную дендрограмму:

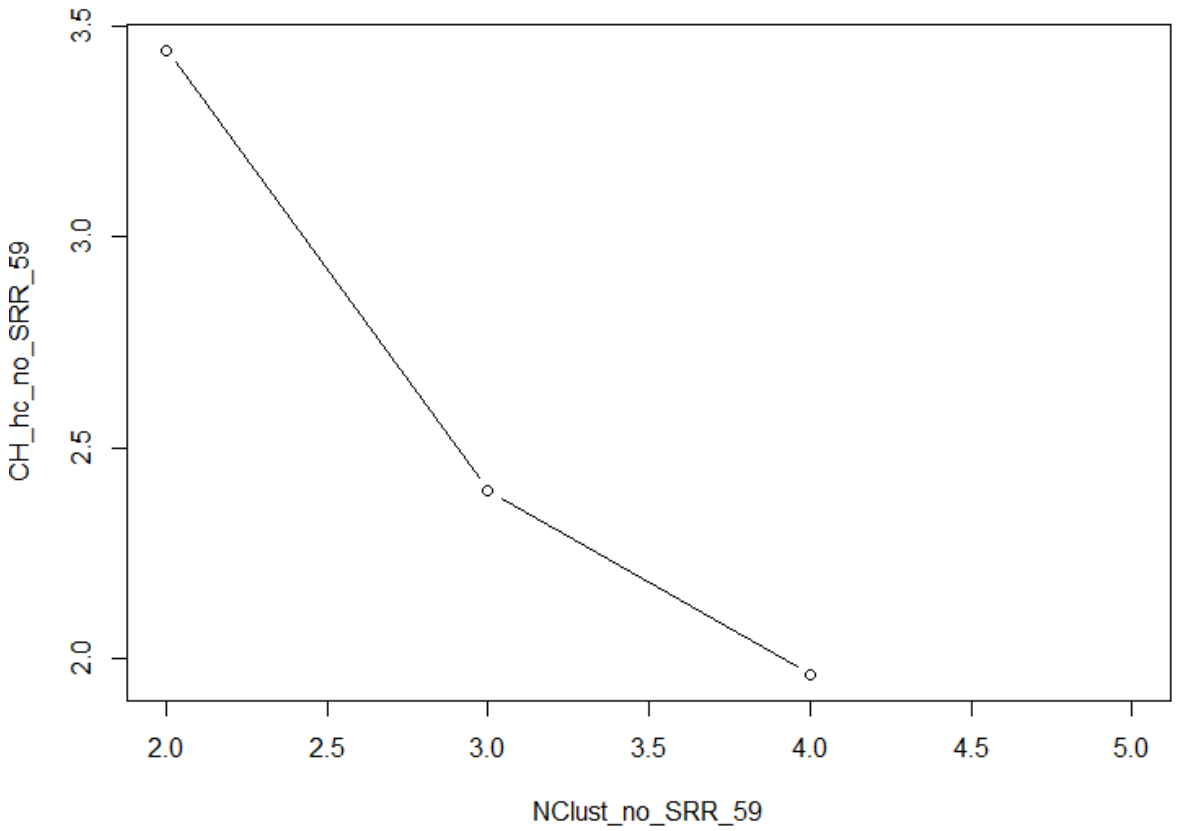


Видим, что значение для SRR_59 выбивается.

Подбор оптимального количества кластеров так же указывает, что данные “делятся” на 3 части:

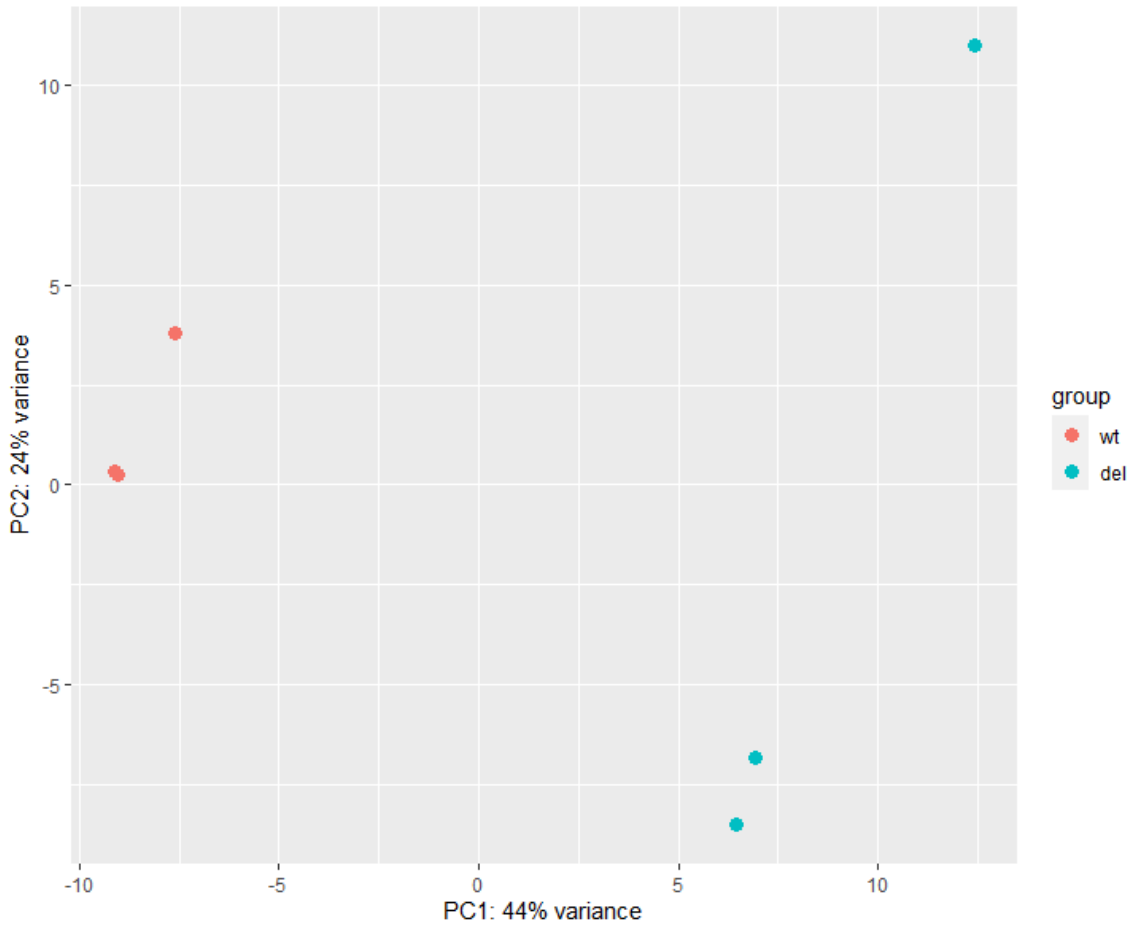


Мы знаем, что у нас есть 2 группы данных – штамм WT и штамм с делецией. Поэтому, принимаем решение работать с данными без выбивающейся пробы. Определяем оптимальное количество кластеров (2), что видно по графику.



б) Анализ PCA

Помимо кластерного анализа для проведения работы можно использовать анализ PCA. Создаём график в соответствии с тем, как наши данные записаны в матрице. График анализа PCA явно показывает выбивающуюся пробу (синяя точка в правом верхнем углу).



Результаты кластерного анализа и анализа PCA соответствуют друг другу.

3. Приступаем к поиску генов в соответствие с условием задачи (*В качестве уровня значимости выбирайте 5%, в качестве порогового значения для изменения экспрессии возьмите 2 раза.*)

Если анализировать данные без поправки, команда summary выдаёт нулевые значения для повышенной и пониженной в 2 раза экспрессии.

```
> summary(DEresults)

out of 4019 with nonzero total read count
adjusted p-value < 0.05
LFC > 1.00 (up)    : 0, 0%
LFC < -1.00 (down) : 0, 0%
outliers [1]       : 1, 0.025%
low counts [2]     : 0, 0%
(mean count < 1)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

Зачастую у генов с низкой экспрессией LFC в среднем выше. Этот артефакт обычно всегда воспроизводится и ожидаем. Если в среднем прочтений на ген мало, то даже небольшое изменение в сравниваемых группах будет давать сильное изменение LFC. Посему, имеет смысл использовать поправки. Так, в работе использована поправка apelglm – она позволит нам уменьшить число ложноотрицательных результатов.

Действительно, после поправки имеем 5 генов, для которых экспрессия увеличилась в 2 раза и 4 гена, для которых экспрессия уменьшилась в 2 раза.

```
> summary(DEresults_LFS_apeglm)

out of 4019 with nonzero total read count
s-value < 0.005
LFC > 1.00 (up)    : 5, 0.12%
LFC < -1.00 (down) : 4, 0.1%
```

Теперь стало возможным узнать названия этих генов через вывод их кодификаторов.

```
> down
[1] "cds-NP_010050.1" "cds-NP_010563.3" "cds-NP_010893.3" "cds-NP_012097.1"
```

Пониженная экспрессия при делеции ULP2 - таблица “кодификатор - белок”:

cds-NP_010050.1	Predicted membrane protein
cds-NP_010563.3	Glucose transmembrane transport
cds-NP_010893.3	Pyrimidine metabolism
cds-NP_012097.1	Protein of unknown function, secreted when constitutively expressed

зеленым подчеркнула то, что затем будет найдено по анализу GO

```
> up
[1] "cds-NP_011438.1" "cds-NP_012817.1" "cds-NP_013208.1" "cds-NP_013441.1" "cds-NP_013937.1"
```

Повышенная экспрессия при делеции ULP2 - таблица “кодификатор - белок”:

cds-NP_011438.1	Amino acid transporters
cds-NP_012817.1	Eisosome protein SEG2
cds-NP_013208.1	RNA exonuclease 3
cds-NP_013441.1	actin cortical patch, actin binding, regulation of cytokinesis
cds-NP_013937.1	putative carboxylic ester hydrolase

4. Для отображения данных построили диаграмму Венна:

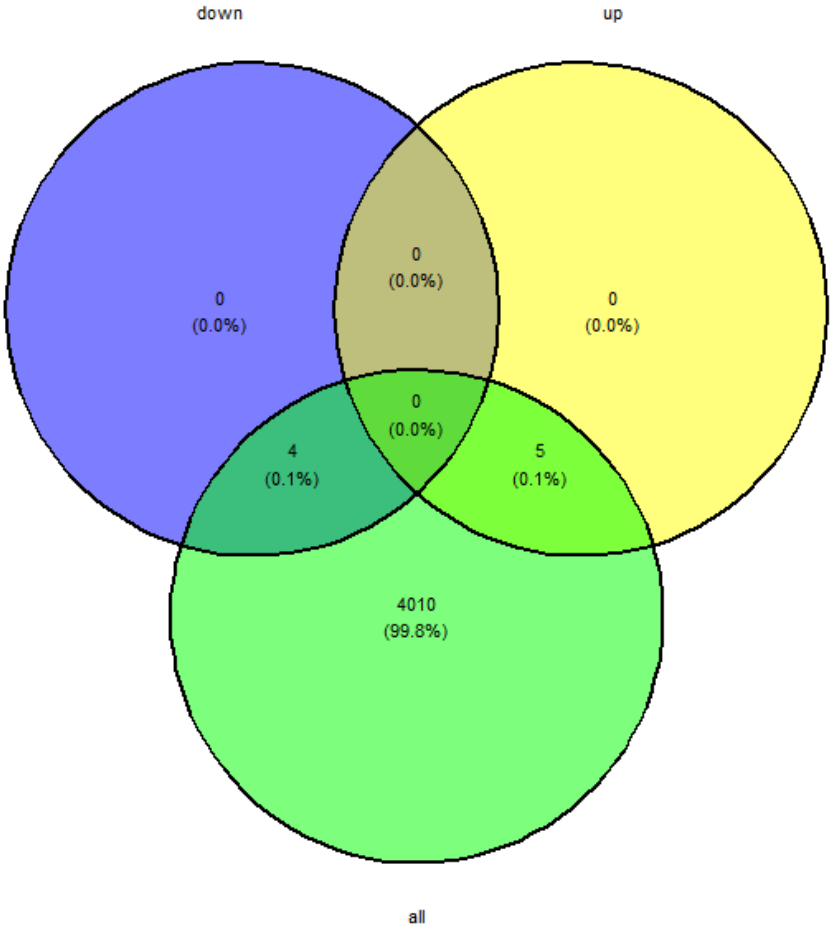
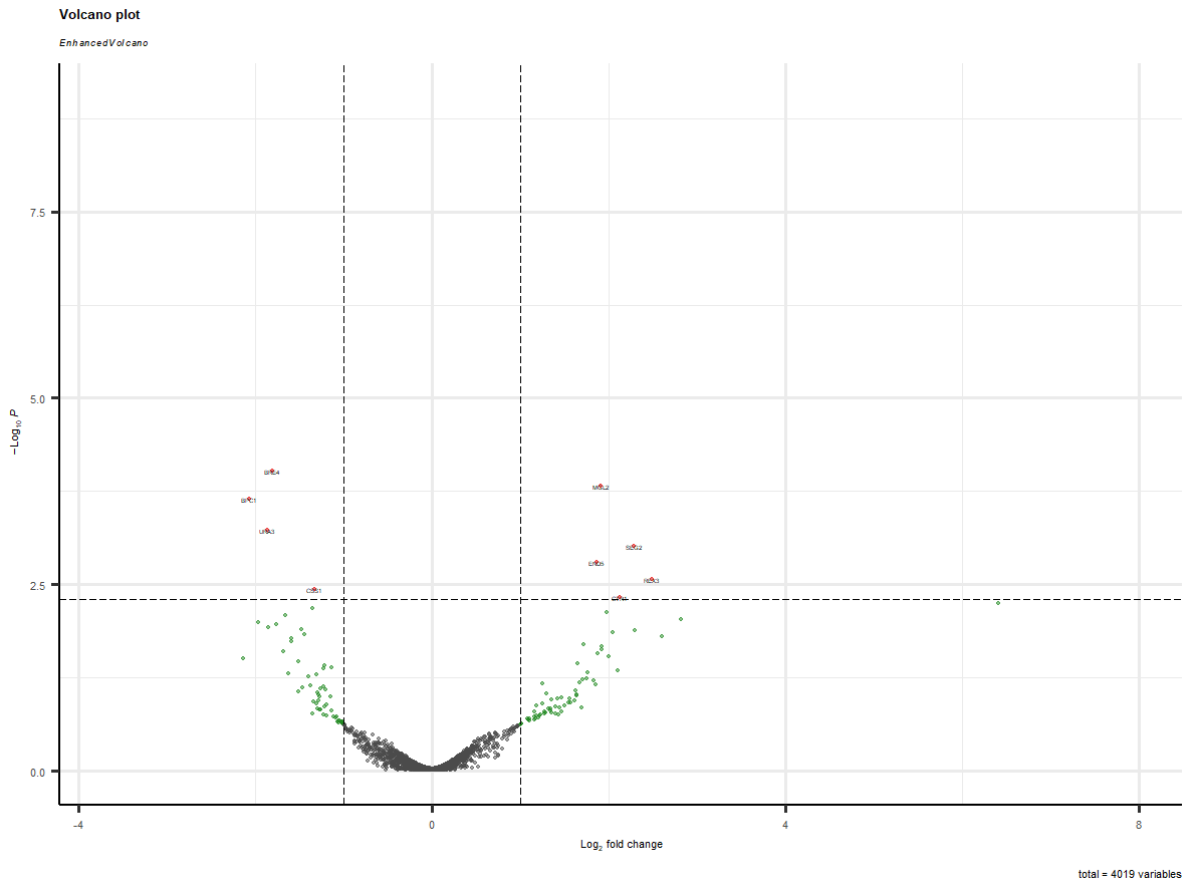


Диаграмма Венна



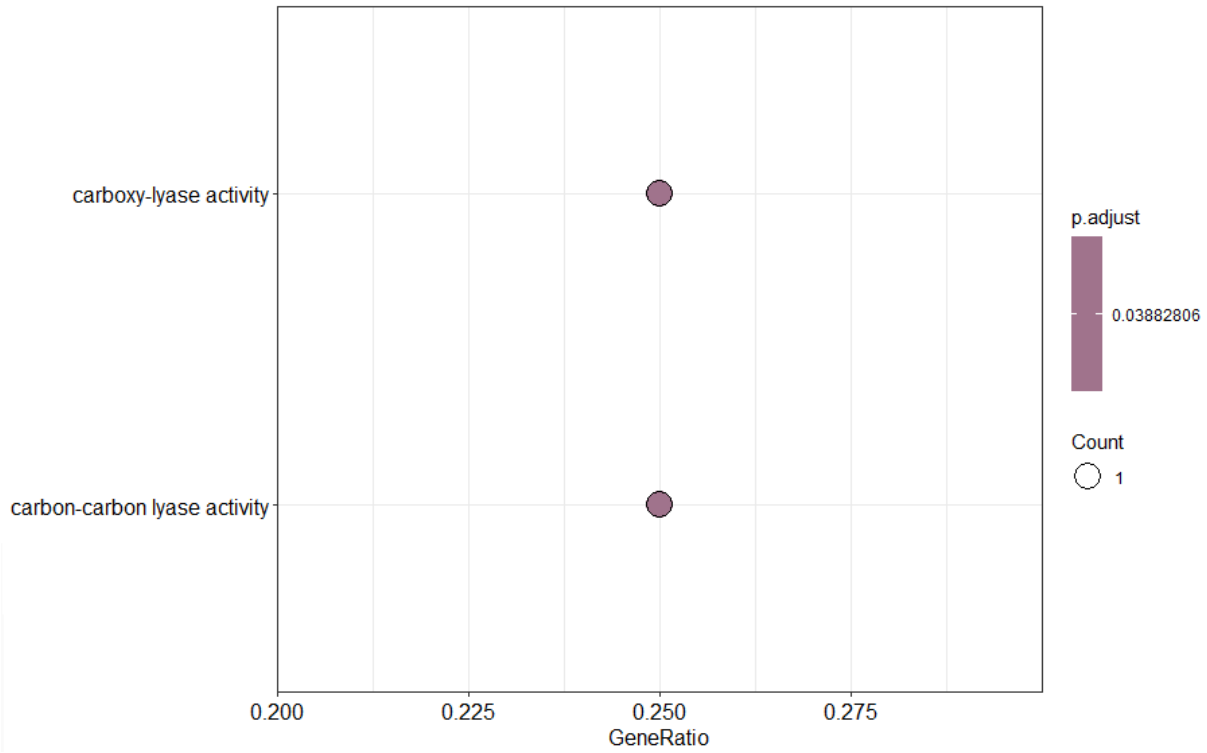
EnchancedVolcano plot

Здесь на рисунке плохо различимо, но всё же. 4 выбивающиеся красные точки слева – это как раз те гены, экспрессия которых понижена. 5 справа – гены, экспрессия которых повышена. Что соответствует предыдущим анализам.
(ремарка для EnchancedVolcano – отсечки pCutoff = 0.005, так как это именно то значение s-value по которому происходила сортировка. P-value после apeglm поправки – нет. Значение 0,005 взяла из summary)

Диаграмма Венна и EnchancedVolcano соответствуют друг другу.

5. Анализ обогащения терминами Gene Ontology.

Для уменьшенной экспрессии: есть результаты только для 2/4 генов по домену MF.



По анализу GO – Делеция гена ULP2 влияет на уменьшение карбон-лиазной и карбон-карбон-лиазной активности – присоединение-отщепление карбоксильных групп.

Для повышенной экспрессии: нет соответствий ни для одного гена ни по одному из доменов.

Эти результаты объяснимы возможным несоответствием БД между собой. Поэтому, по всей видимости, анализ стоит продолжить вручную. Это сделать легко, так как уже ранее мы получили ID для генов с повышенной и пониженной экспрессией (таблица выше).

По таблице выше, которую составляла в соответствие с идентификаторами. При делеции ULP2 наблюдаем повышение экспрессии генов:

- Возможного мембранного белка
- Глюкозосодержащего (?) трансмембранного белка
- Белка-участника пиримидинового метаболизма
- Белка непонятной функции

Видимо, клетка “копит” в себе урацил для, возможно, синтеза ДНК.
Понижение экспрессии генов:

- Транспортера аминокислот
- Эйзосомного белка
- РНК-экзонуклеазы
- Цитоскелетного белка, который участвует во многих процессах
- Предполагаемой гидролазы сложного эфира карбоновой кислоты

Наблюдаем тенденцию понижения экспрессии генов, связанных с эндо- (может и экзо-) цитоза. Дополнительно затронут важный (как я поняла..) цитоскелетный белок, что,

возможно, обуславливает сложности внутриклеточного транспорта и построения цитоскелетных структур.