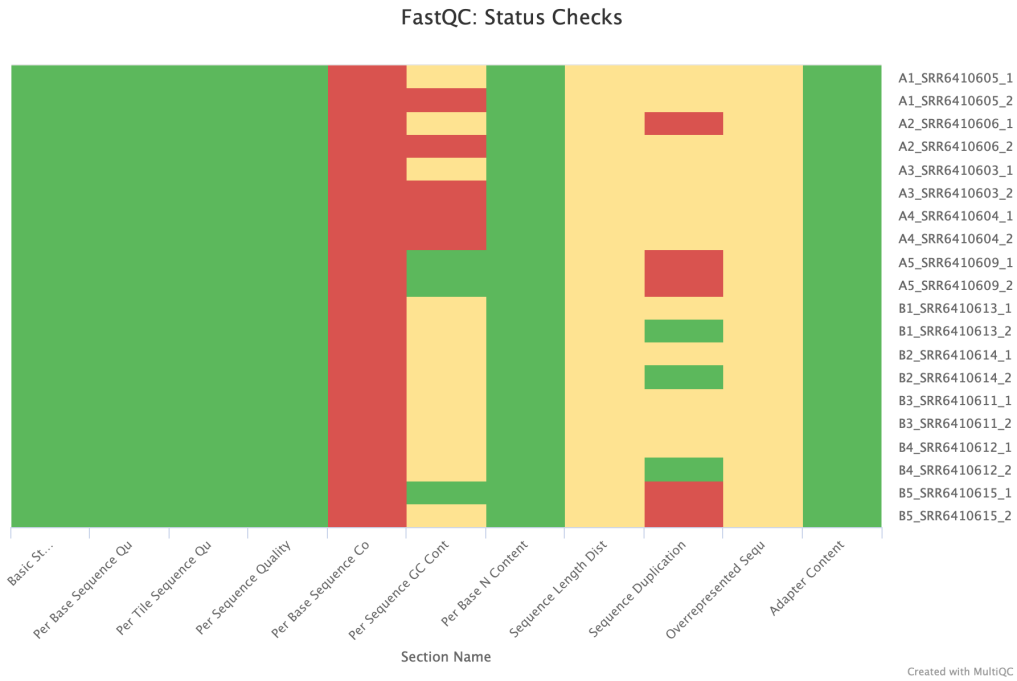
«Bulk RNA-Seq»

1. (1) Загрузите прочтения по ссылкам, указанным в конце задания. Запустите контроль качества прочтений на каждом образце индивидуально. При помощи программы MultiQC аггрегируйте результаты по всем образцам и словесно опишите качество данных, с которыми вам предстоит работать.

Были загружены 5 образцов из группы B и 5 образцов из группы A (код в блокноте *Punko_HW1_task1_2.ipynb)*.
Запустили контроль качества прочтений на каждом образце индивидуально с помощью FastQC (папка *fastqc_results* на сервере).
С помощью MultiQC аггрегировали результаты по всем образцам (файл *multiqc_report.html*). Средняя длина прочтений – 75 п.н., по 50 млн последовательностей в каждом образце. Распределение качества по позициям у нас очень хорошее во всех образцах. На графике (Per Sequence Quality Scores), который показывает, какое количество ридов относится к тому или иному качеству, видно, что во всех образцах большинство ридов с хорошим качеством. На графике Per Base Sequence Content мы видим, что есть сильная неслучайность в распределении нуклеотидов, однако для RNA-seq-экспериментов это неудивительно, так как случайные праймеры не на сто процентов отжигаются случайно. Распределение GC состава на образец показывает, что в некоторых образцах есть второй пик, который может говорить о контаминации. Также в некоторых образцах высокое количество дублирующихся прочтений, что говорит об артефактах ПЦР (это справедливо для, например, геномных библиотек, в которых встретить два одинаковых рида очень маловероятно), однако для RNA-Seq-экспериментов это нормально, потому что какие-то гены могут встречаться очень часто из-за высокой копийности их РНК. Образцов с загрязнением адаптера > 0,1 % не обнаружено.



2. (1) Подготовьте референсный транскриптом для `kallisto` (с геномом человека) и проведите подсчёт экспрессии на уровень транскриптов. Откуда вы взяли референсный транскриптом? Не забудьте, что для следующих шагов вам потребуется таблица с соответствием генов транскриптам, также вам пригодится понимание того, какие конкретно гены являются белок-кодирующими (дифференциальную экспрессию будем выполнять именно на них).

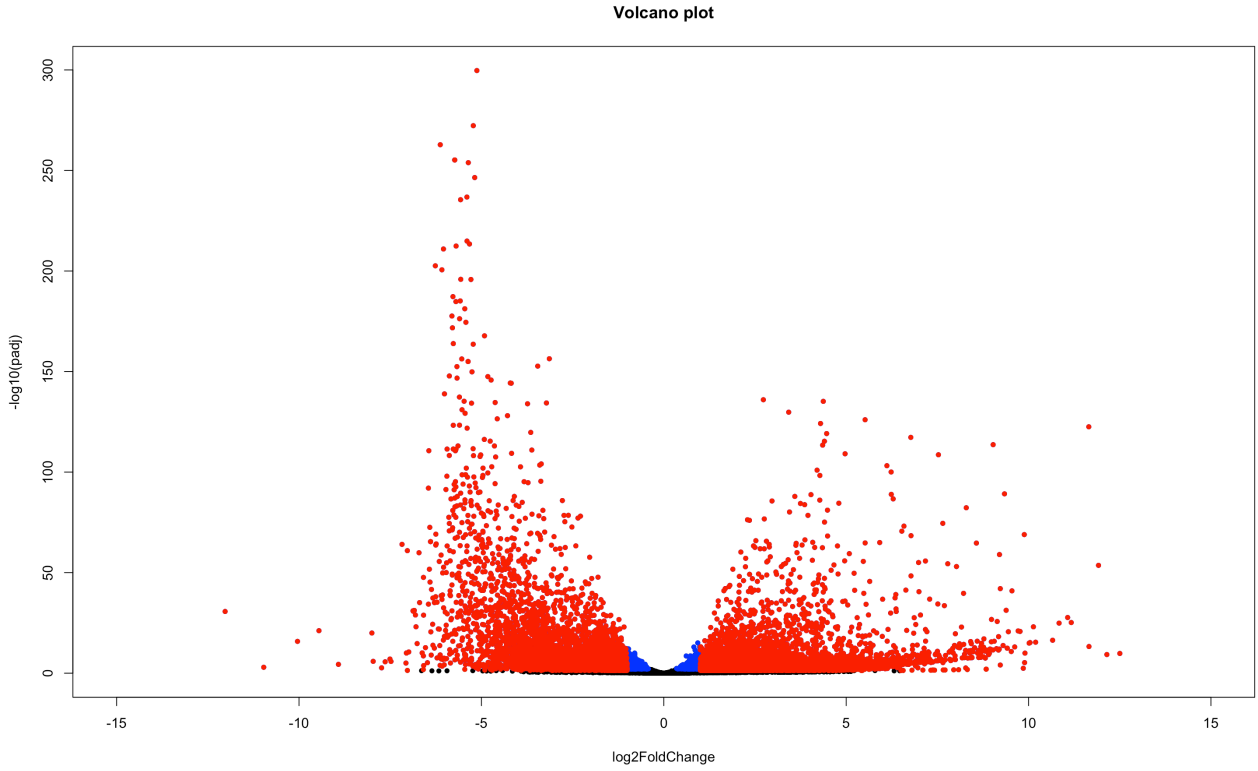
Подготовили референсный транскриптом для `kallisto` (с геномом человека) и провели подсчёт экспрессии на уровень транскриптов (код в блокноте *Punko_HW1_task1_2.ipynb*; папка *Kallisto*).
Референсный транскриптом был загружен из базы Ensembl.org. Файл *Homo_sapiens.GRCh38.cdna.all.fa.gz* содержит последовательности кДНК, соответствующие генам Ensembl, за исключением генов нкРНК. кДНК состоит из последовательностей транскриптов реальных и возможных генов, включая псевдогены, NMD и т.п.

3. -
4. (0.5) С использованием `tximport` загрузите в `DESeq2` результат обсчёта экспрессий из пункта (2) и проведите дифференциальную экспрессию между образцами из групп A и B на уровне генов.

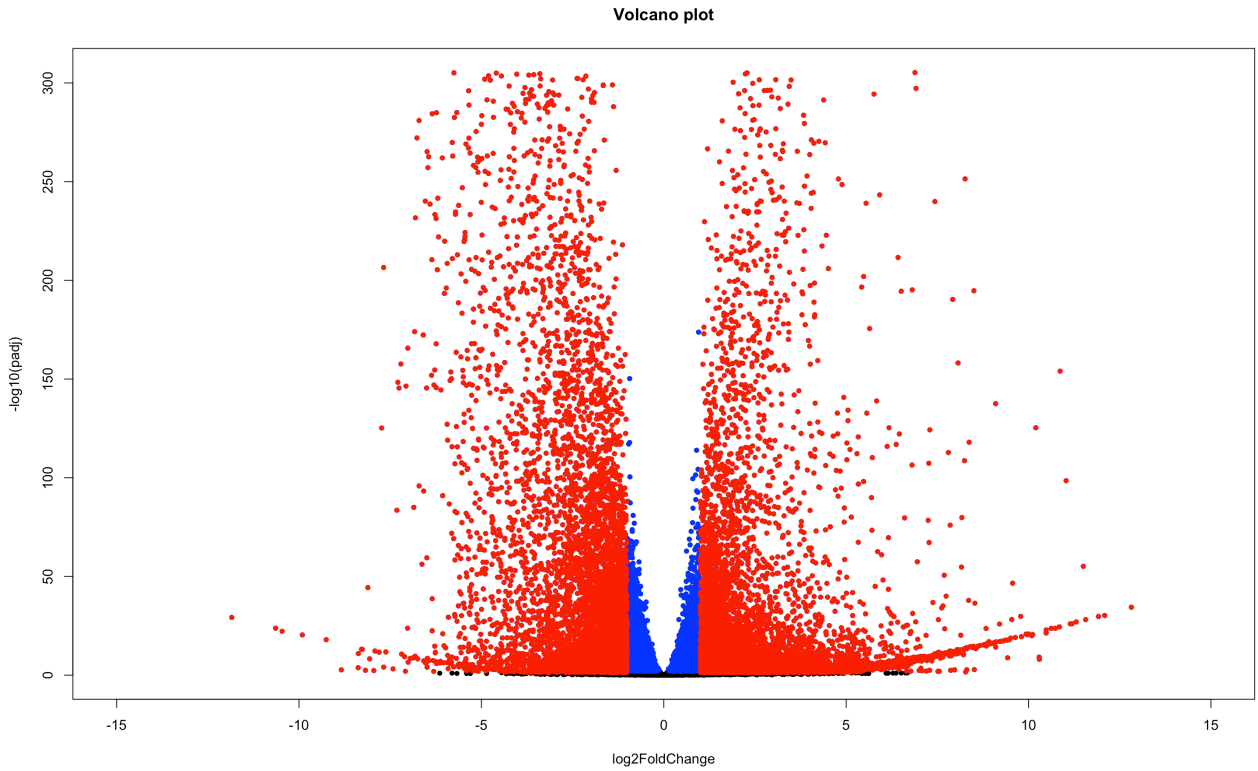
С использованием `tximport` загрузили в `DESeq2` результат обсчёта экспрессий и провели дифференциальную экспрессию между образцами из групп A и B на уровне генов. Код в проекте R *HW1_NGS_Punko;* скрипт *de_script.R;* для *tximport* файл *tx2gene_grch38_ens94.txt* и папка *data*.

5. -
6. (1) Проведите анализ самосогласованности данных при помощи PCA. Выбросите ненужные образцы и проведите анализ дифференциальной ещё раз. Влияет ли наличие аутлаеров на результат дифференциальной экспрессии? Выводы подкрепите графически.

График PCA в проекте R *HW1_NGS_Punko* в папке *results* (*plotPCA*).
После исключения образцов B5 и A5 был проведен анализ дифференциальной экспрессии еще раз. Результаты в папке *results without A5B5*.
Наличие аутлаеров влияет на результат дифференциальной экспрессии, количество значимых дифф экспрессированных генов увеличилось. Наглядно разницу мы можем увидеть ниже на графиках volcano plot.



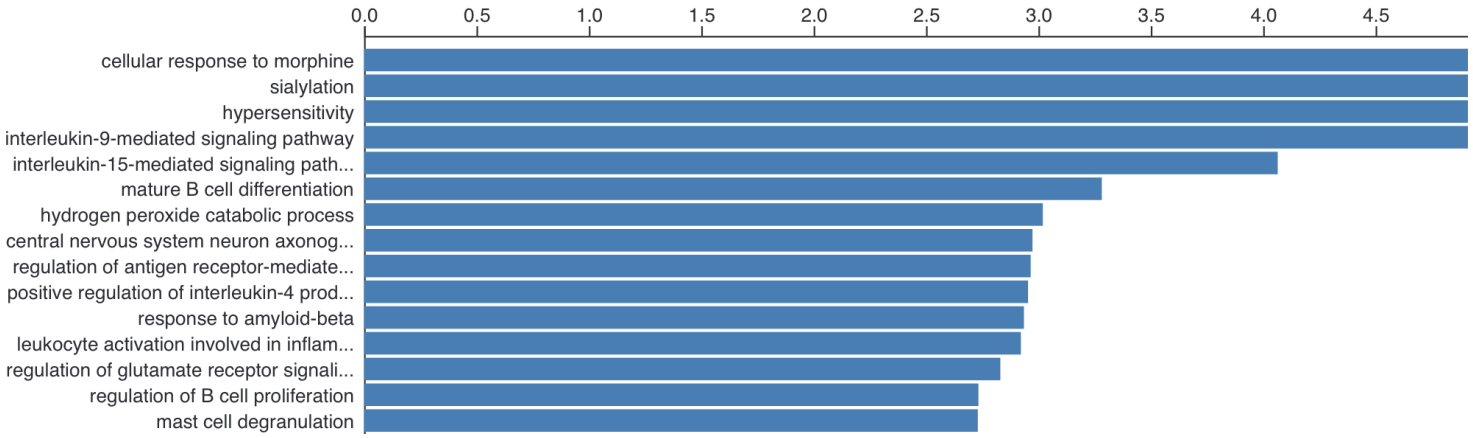
Volcano plot с аутлаерами (синие padj<0.01, красные log2FC>1 и padj<0.05)



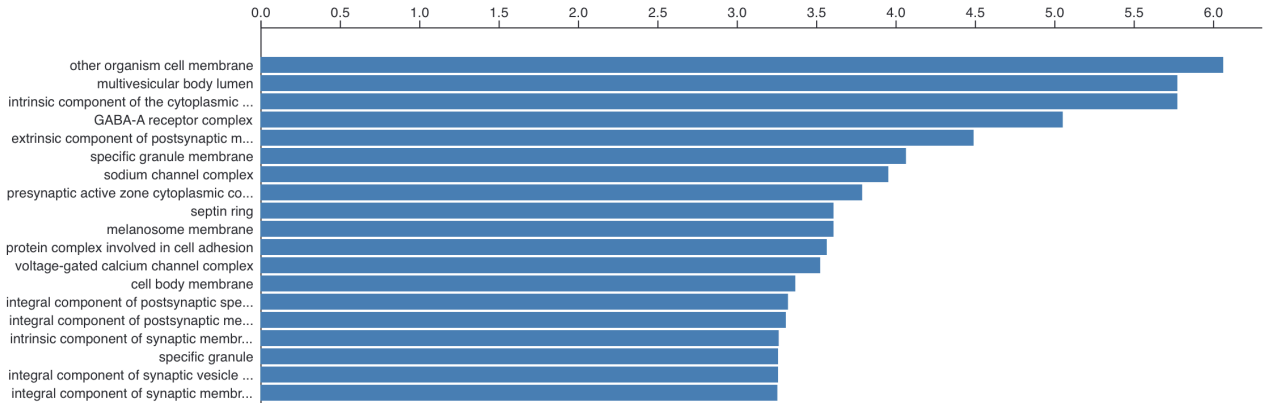
Volcano plot без аутлаеров (синие padj<0.01, красные log2FC>1 и padj<0.05)

7. (1.5) Проведите функциональный анализ полученных данных (feel free в выборе инструментов). Какую основную разницу между образцами вы видите?

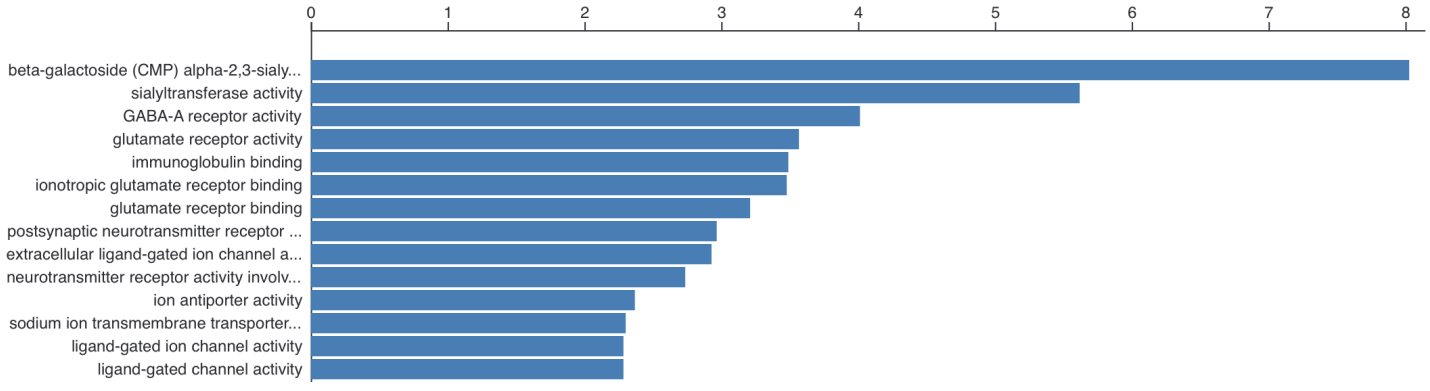
Функциональный анализ данных проводился с помощью *WebGestalt (WEB-based Gene SeT AnaLysis Toolkit)*.



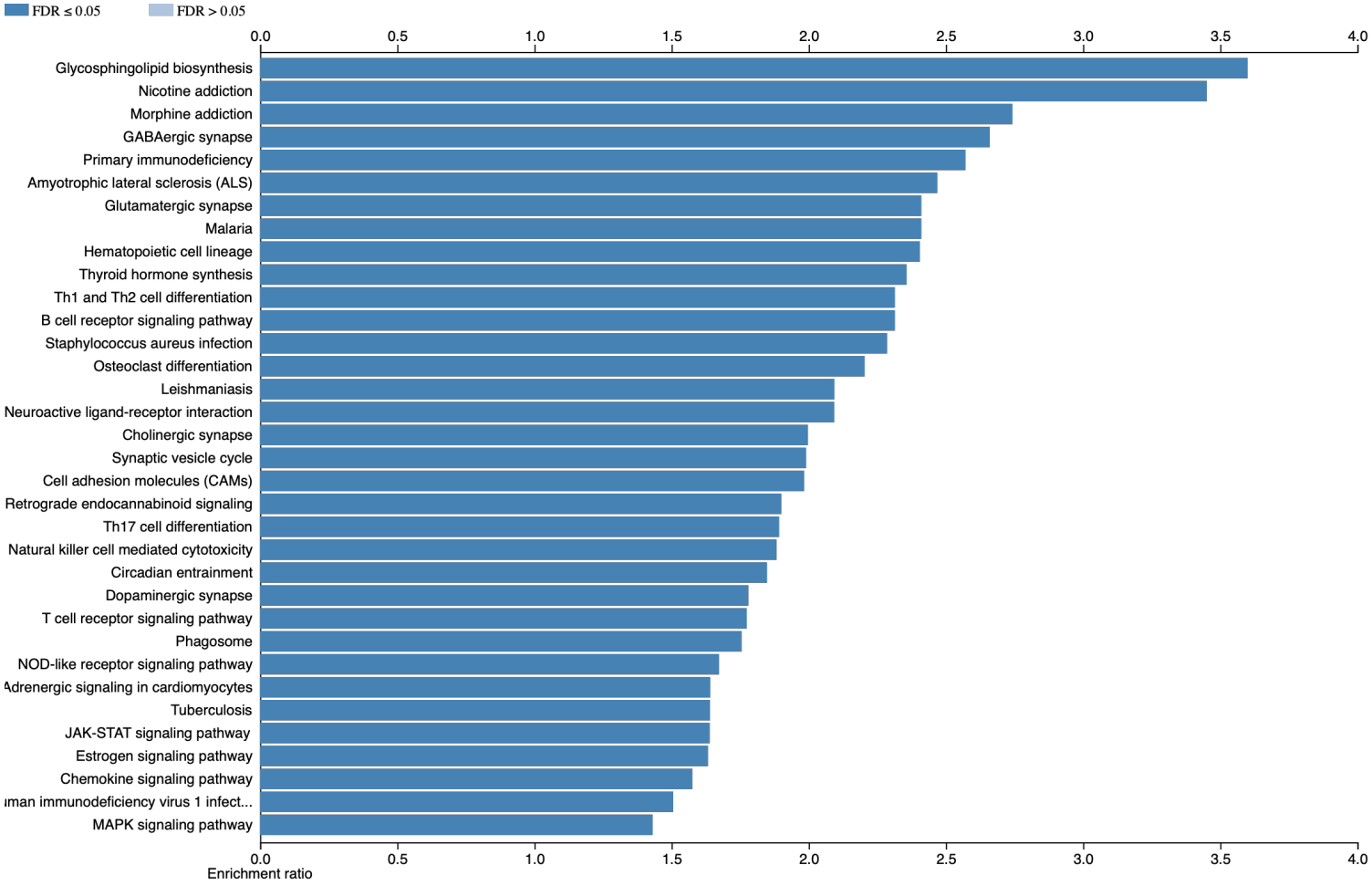
Результат для базы Geneontology (Biological Process)



Результат для базы Geneontology (Cellular Component)



Результат для базы Geneontology (Molecular Function)



Результат для базы KEGG

8. -