## «Ribo-seq»

Загрузите файл (Google Drive ID: 1iqYLFVKySREVm5Kn5NagMeHaDi-dVLcj, файл 01. RiboSeq_RNASeq_HCC_counts.tsv), в котором содержится матрица каунтов результатов Ribo-Seq и RNA-Seq экспериментов больных гепатоцеллюлярной карциномой. Матрица каунтов была получена стандартным воркфлоу STAR. На каждого пациента приходится 4 столбца в таблице (RNA-Seq нормы и опухоли, Ribo-Seq нормы и опухоли), записанные в виде sample_number-tissue_type-experiment.

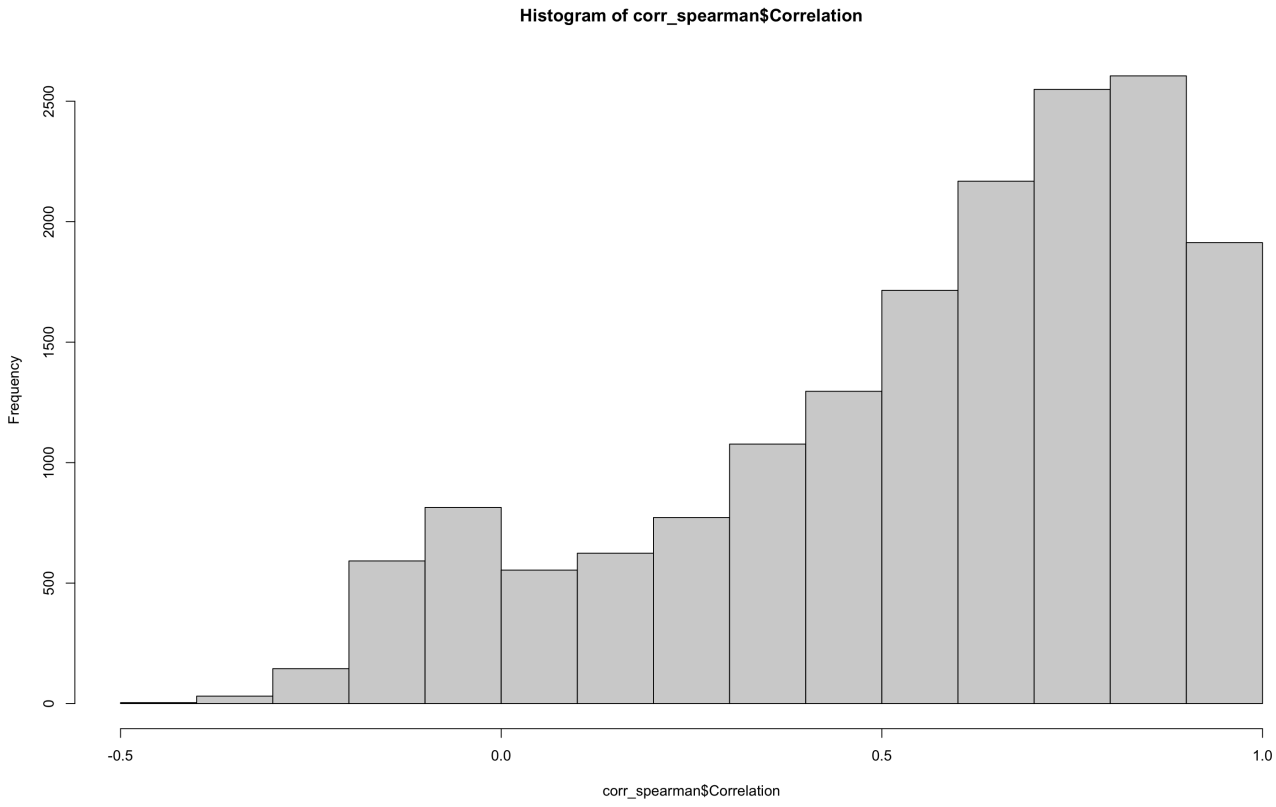
Файл в папке HSE_RiboSeq_HT : 01. *RiboSeq_RNASeq_HCC_counts.tsv*
Код в проекте R – *Ribo-seq*

1. (2) Для каждого гена найдите коэффициент корреляции числа каунтов между экспериментами RNA-Seq и Ribo-Seq. Какую корреляцию вы выберете — Пирсона или Спирмена? Обратите внимание, что перед вами не нормированные на глубину библиотеки каунты, учтите это при анализе.

Код для выполнения задания в файле **Ribo.R**

Для нахождения коэффициента корреляции числа каунтов между экспериментами RNA-Seq и Ribo-Seq была применена функция cor для каждого гена. В папке result в файлах corr_spearman и corr_pearson.csv для каждого гена посчитаны коэффициенты корреляции.
Я бы выбрала корреляцию Спирмена, непараметрический тест, так как число каунтов не распределены нормально.

2. (2) Постройте гистограмму распределения этих коэффициентов корреляции. Для каких генов корреляция самая высокая, а для каких — самая низкая? Как вы можете это объяснить?



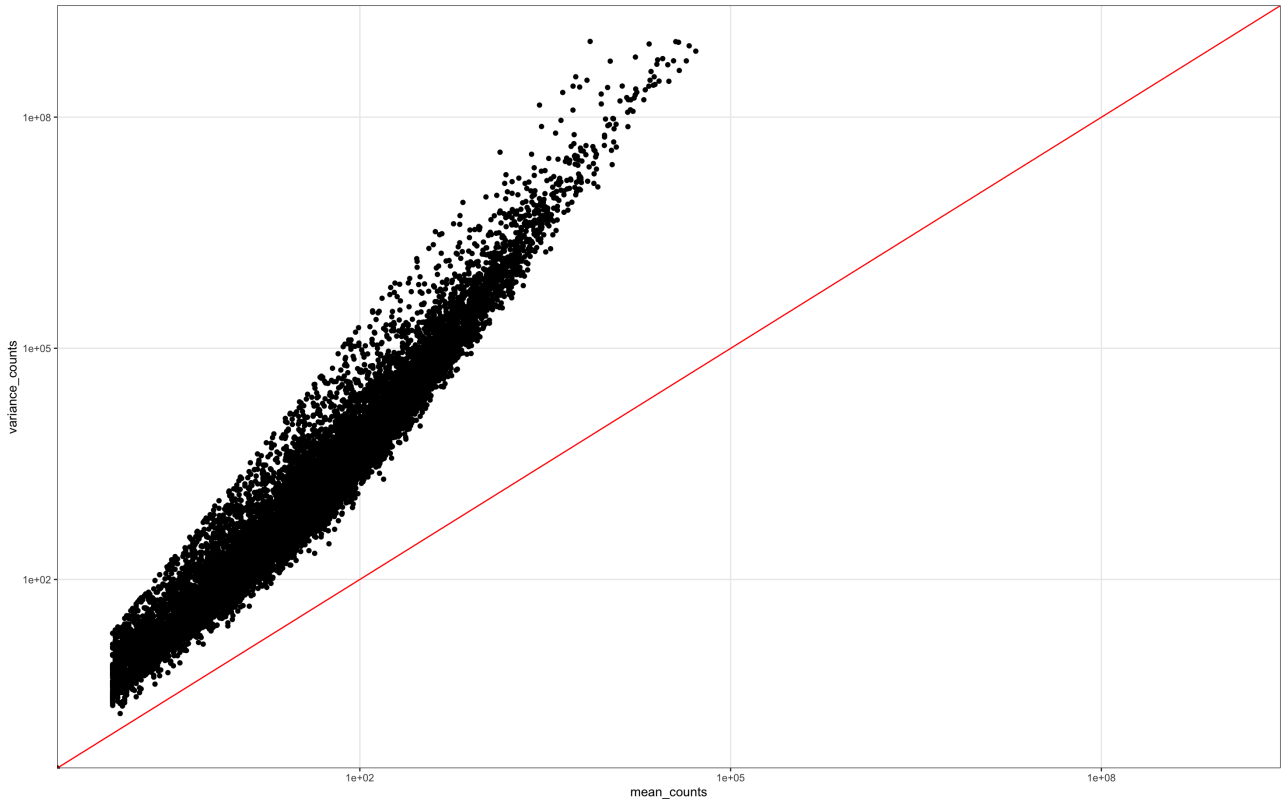
| Гены с самой высокой корреляцией | | Гены с самой низкой корреляцией | |
|---|---|---|---|
| Gene | Correlation | Gene | Correlation |
| IQCF3 | 1.0000000 | FAM58A | -0.3684335 |
| GAGE10 | 1.0000000 | TMEM88 | -0.3694606 |
| DYNAP | 1.0000000 | ABCG4 | -0.3837030 |
| TRIM60 | 1.0000000 | MBD3 | -0.3864696 |
| FAM46D | 1.0000000 | LAT | -0.3887943 |
| KCNA10 | 1.0000000 | GRM2 | -0.3931655 |
| TMPRSS7 | 1.0000000 | CEACAM4 | -0.4034470 |
| PRG3 | 1.0000000 | REM2 | -0.4303548 |
| AGR3 | 0.9999795 | ZNF358 | -0.4505198 |
| SI | 0.9999689 | C14orf28 | -0.4997039 |

Высокий коэффициент корреляции обусловлен тем, что количество каунтов у этих генов равно 0 в экспериментах RNA и Ribo seq.

3. (3) Проанализируйте распределение каунтов Ribo-seq: постройте зависимость дисперсии от среднего. Похоже ли это на NB-распределение? Если нет, то на какое похоже? Если да, то какие статистические тесты можно использовать для подтверждения того, что перед вами NB-распределение? Проведите такую оценку.

Код для выполнения задания в файле **Ribo_task3.R**

Для анализа распределения каунтов Ribo-seq были рассчитаны дисперсия и среднее для каждого гена. А далее был построен график, который показывает зависимость дисперсии от среднего. На графике каждая точка данных представляет ген, а красная линия представляет собой x = y.
На графике видно, что среднее значение не равно дисперсии (разброс точек данных не попадает на диагональ). Для генов с высокой средней экспрессией дисперсия по репликам имеет тенденцию превышать среднее значение (разброс выше красной линии).
Распределение, которое лучше всего соответствует нашим данным представляет собой отрицательное биномиальное распределение (так как среднее значение < дисперсии).
Среднее и дисперсия отрицательного биномиального распределения связаны, благодаря чему мы можем инспектировать наши распределения даже без каких-либо тестов на Goodness of Fit.

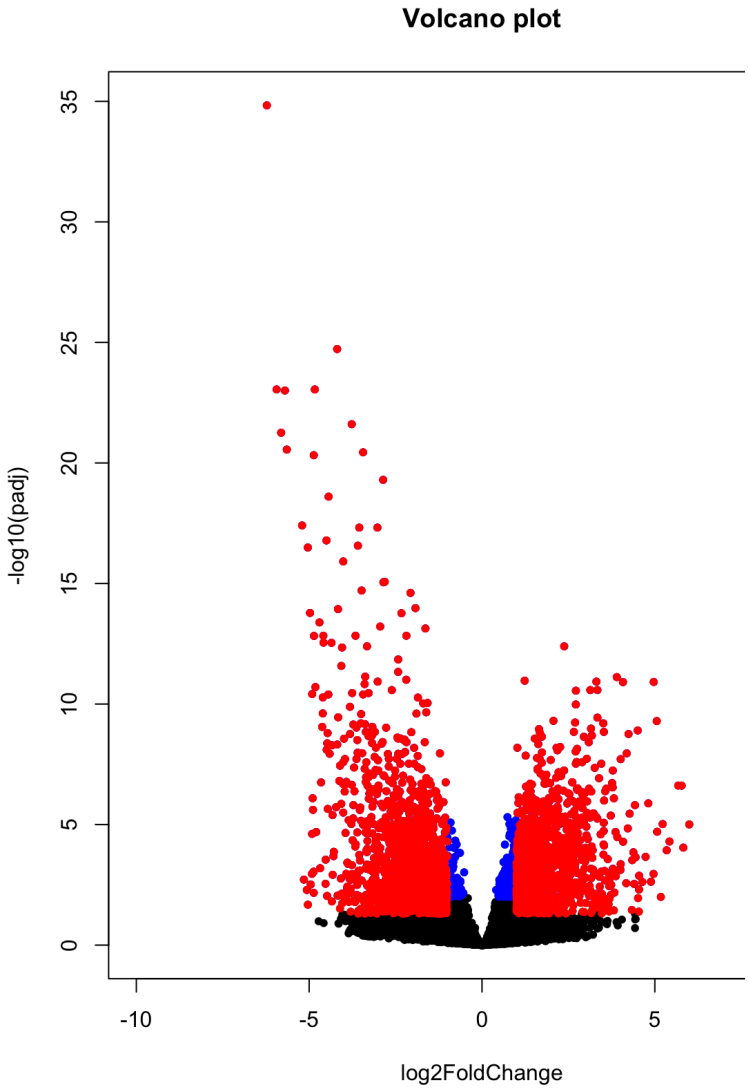


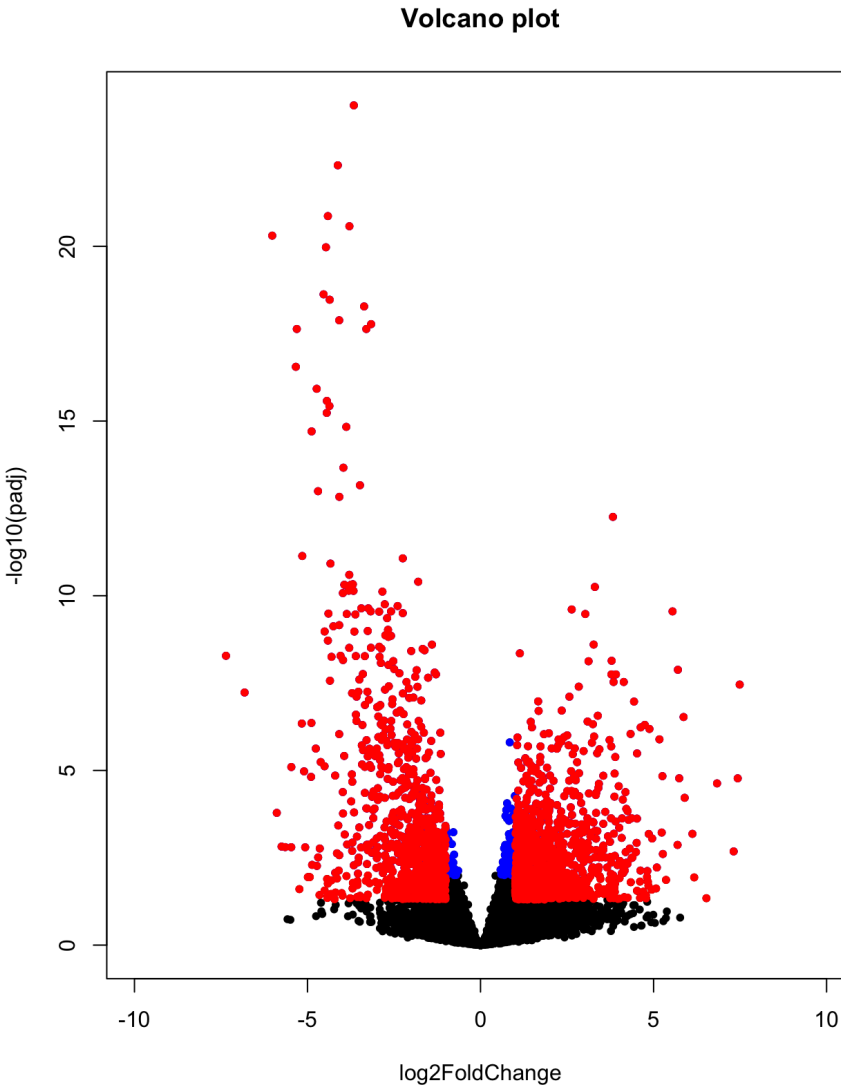

Зависимость дисперсии от среднего

4. (3) Найдите дифференциально экспрессированные гены по RNA-Seq между нормой и опухолью. Для этого можете использовать DESeq2 или edgeR. Постройте volcano plot. То же самое сделайте и для Ribo-Seq эксперимента. Совпадают ли результаты дифференциальной экспрессии? Какие гены оказались значимо различны по экспрессии в одном случае, а какие — в другом? Проведите похожий анализ, только с поиском генов с разницей в эффективности трансляции (`~ method + condition + method:condition`). Что вы скажете о наборе генов, который вы получили этим методом? Выводы подкрепите графиками.

Код для выполнения задания в файле **Ribo_task4.R**

Для нахождения дифференциально экспрессированных генов между нормой и опухолью использовался DESeq2. Результаты для RNA-Seq и Ribo-Seq не совпадают. В RNA-Seq дифференциально экспрессированных генов немного больше.
Гены, которые значимо различны по экспрессии, находятся в папке result в файлах result_RNA.csv и result_Ribo.csv.




Volcano plot RNA-Seq

Volcano plot Ribo-Seq

(синие padj<0.01, красные log2FC>1 и padj<0.05)

Топ 5 значимо различных по экспрессии генов (RNA-Seq)

| | baseMean | log2FoldChange | lfcSE | stat | pvalue | padj |
|---|---|---|---|---|---|---|
| ITGA6 | 473.99053 | 2.378037 | 0.2958812 | 8.037135 | 9.196356E-16 | 4.03115E-13 |
| PBK | 18.27054 | 3.899668 | 0.5097642 | 7.649944 | 2.010672E-14 | 7.611766E-12 |
| PPP1CC | 344.58846 | 1.235945 | 0.1626314 | 7.599666 | 2.968958E-14 | 1.075086E-11 |
| CENPF | 163.02374 | 3.309567 | 0.4364688 | 7.582597 | 3.387057E-14 | 1.175379E-11 |
| GPC3 | 2950.58499 | 4.967709 | 0.6557331 | 7.575810 | 3.568945E-14 | 1.213223E-11 |

Топ 5 значимо различных по экспрессии генов (Ribo-Seq)

| | baseMean | log2FoldChange | lfcSE | stat | pvalue | padj |
|---|---|---|---|---|---|---|
| CGREF1 | 52.171200 | 3.825921 | 0.4765743 | 8.027965 | 9.910297E-16 | 5.534703E-13 |
| HIST1H3B | 292.788867 | 3.303179 | 0.4463973 | 7.399640 | 1.365545E-13 | 5.607569E-11 |
| SLC26A6 | 81.113164 | 2.634797 | 0.3675748 | 7.168057 | 7.606936E-13 | 2.469954E-10 |
| GPC3 | 7131.337141 | 5.548405 | 0.7766696 | 7.143843 | 9.075701E-13 | 2.805167E-10 |
| HIST1H3C | 141.578703 | 3.027347 | 0.4261856 | 7.103354 | 1.217646E-12 | 3.333485E-10 |