«Анализ транскриптомных данных»

Лекция #3.

Распределения в омиках. Методы нормализации

Серёжа Исаев

аспирант **ФБМФ МФТИ**
аспирант **MedUni Vienna**

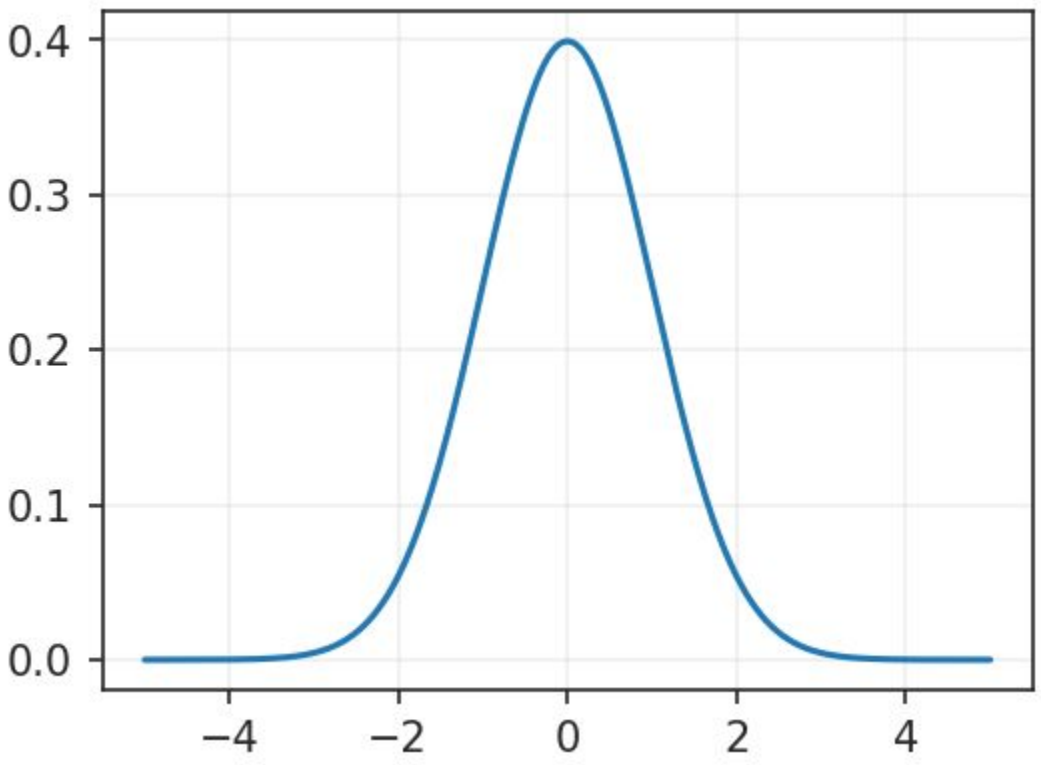
Содержание курса

1. Bulk RNA-Seq:
    a. экспериментальные подходы,
    b. выравнивания и псевдовыравнивания,
    c. анализ дифференциальной экспрессии,
    d. функциональный анализ;

2. Single-cell RNA-Seq:
    a. экспериментальные подходы,
    b. отличия от процессинга bulk RNA-Seq,
    c. методы снижения размерности,
    d. кластера и траектории,
    e. мультимодальные омики одиночных клеток.

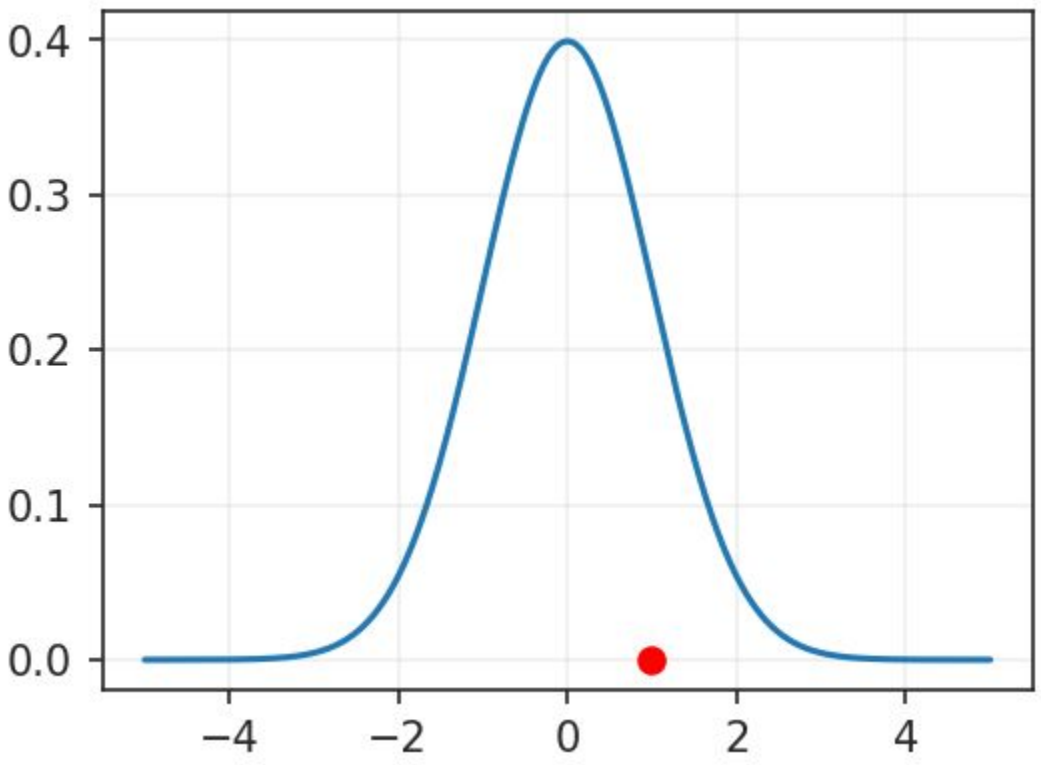
Функция правдоподобия

Дано некоторое распределение. Для простоты будем считать, что у нас есть нормальное распределение с центром в 0 и дисперсией 1



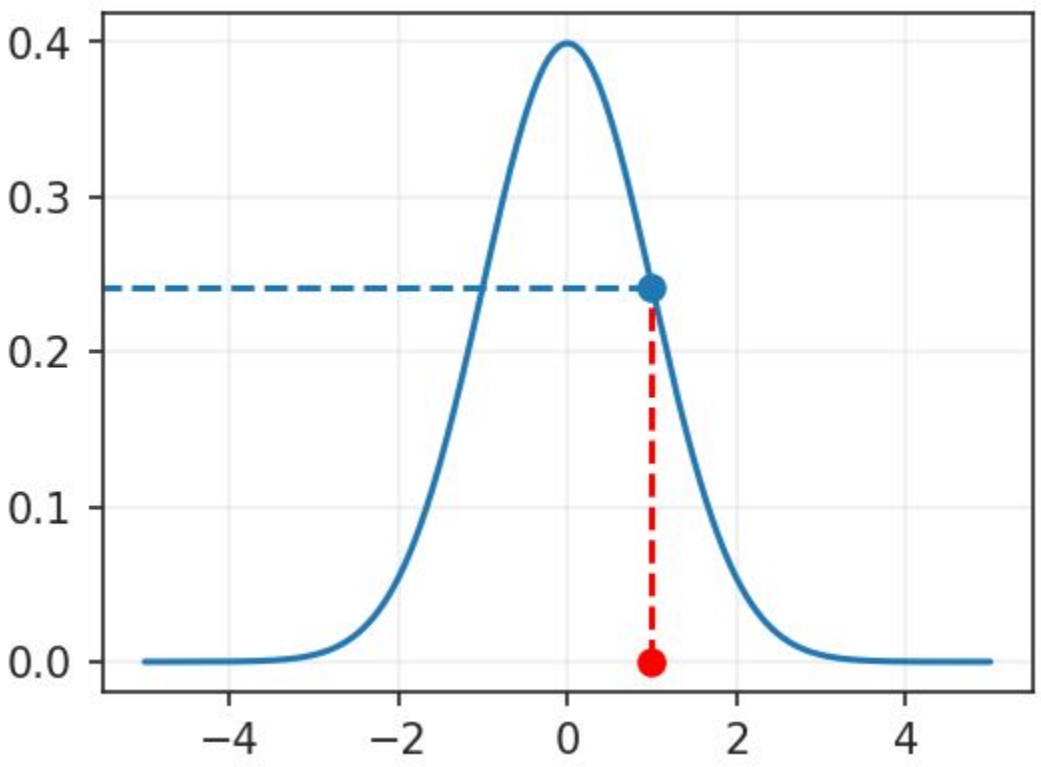
Функция правдоподобия

Представим, что мы видим некоторое событие — точку со значением 1.
Какова плотность вероятности в этой точке?



Функция правдоподобия

Значение плотности распределения в точке 1 равно 0.24, то есть $f(1 \mid N(0, 1)) = 0.24$.



Функция правдоподобия

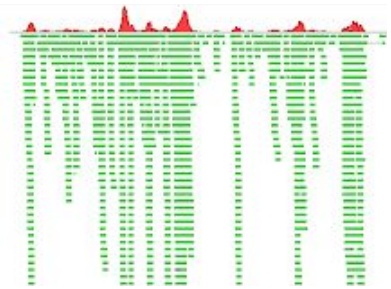
А теперь представим, что нам дана только эта точка, зато не даны параметры распределения, то есть нам необходимо их оценить. Оценка плотности вероятности распределения при известных данных — это и есть правдоподобие $L(N(0, 1) \mid 1) = f(1 \mid N(0, 1)) = 0.24$. То есть мы оцениваем плотность вероятности параметров системы при уже имеющихся наблюдениях.

Дорожная карта анализа RNA-Seq

QC прочтений

FastQC Report

Выравнивание



Подсчёт экспрессий

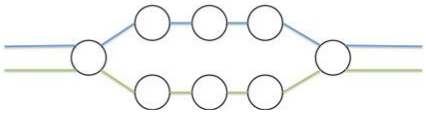
	Wild-type	
	Sample 1	Sample 2
Gene 1	24	31
Gene 2	0	3
Gene 3	1988	1125
Gene 4	5	0
...	...	...
Total	22341961	20739175

Нормализация

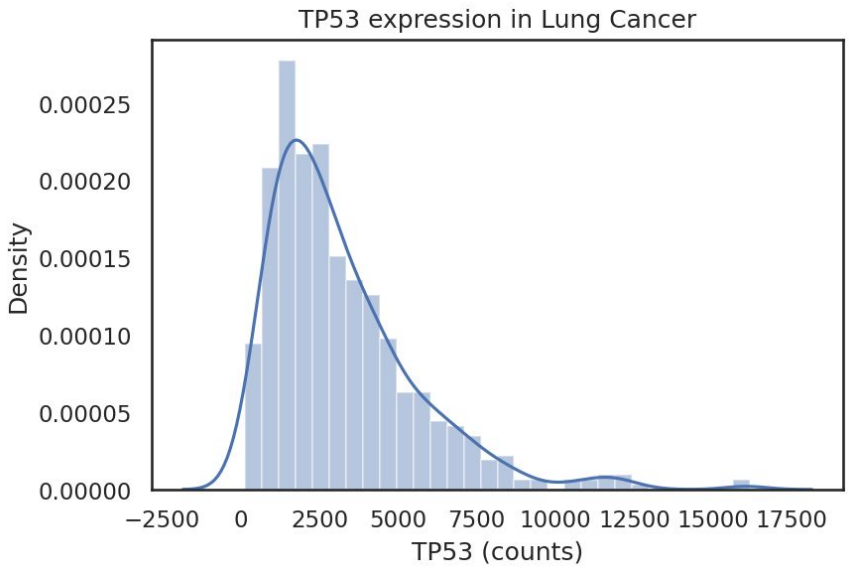
Дифференциальная экспрессия

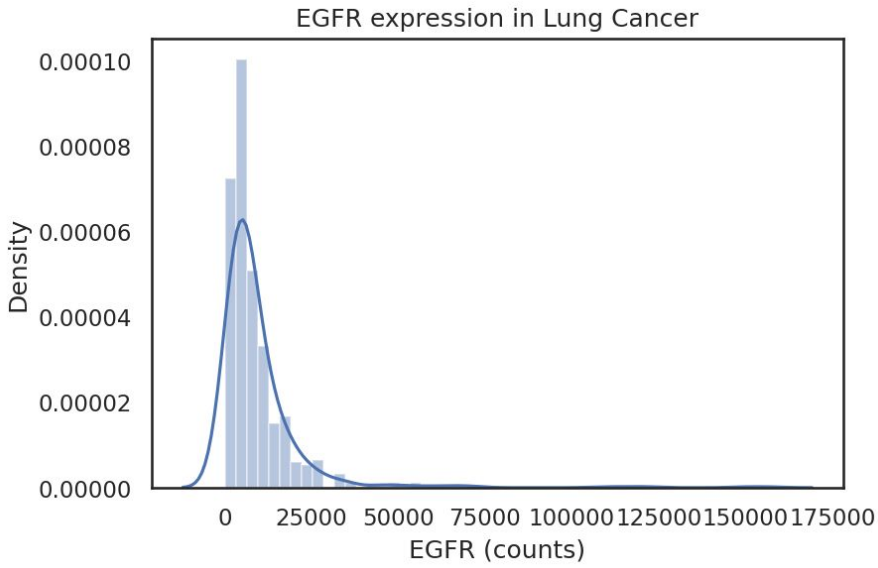
GO, GSEA, ssGSEA и прочее

Псевдовыравнивание



Распределение каунтов генов



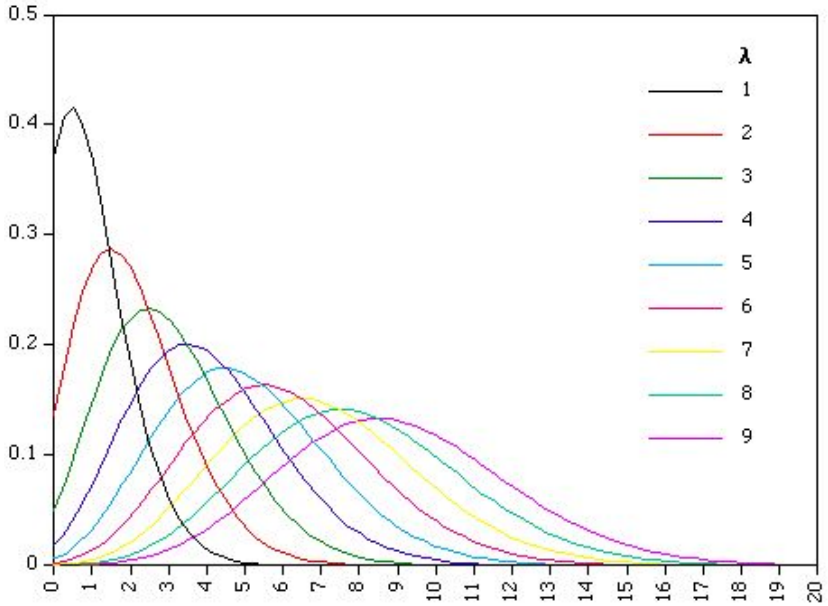


Экспрессии генов TP53 и EGFR в образцах рака лёгкого

Какое это распределение?

Распределение Пуассона

$$p(k) \equiv \mathbb{P}(Y = k) = \frac{\lambda^k}{k!}\, e^{-\lambda},$$



Распределение Пуассона отражает число событий, произошедших за фиксированное время, при условии, что данные события происходят с некоторой фиксированной **средней интенсивностью** и **независимо друг от друга**

Распределение Пуассона

Представим, что у нас есть бесконечно большая шляпа, в которой есть несколько типов шариков — красные, синие, зелёные, … Сфокусируемся на красном шарике, доля красных шариков 0.01 (то есть вероятность вытащить красный шарик — 1 из 100).

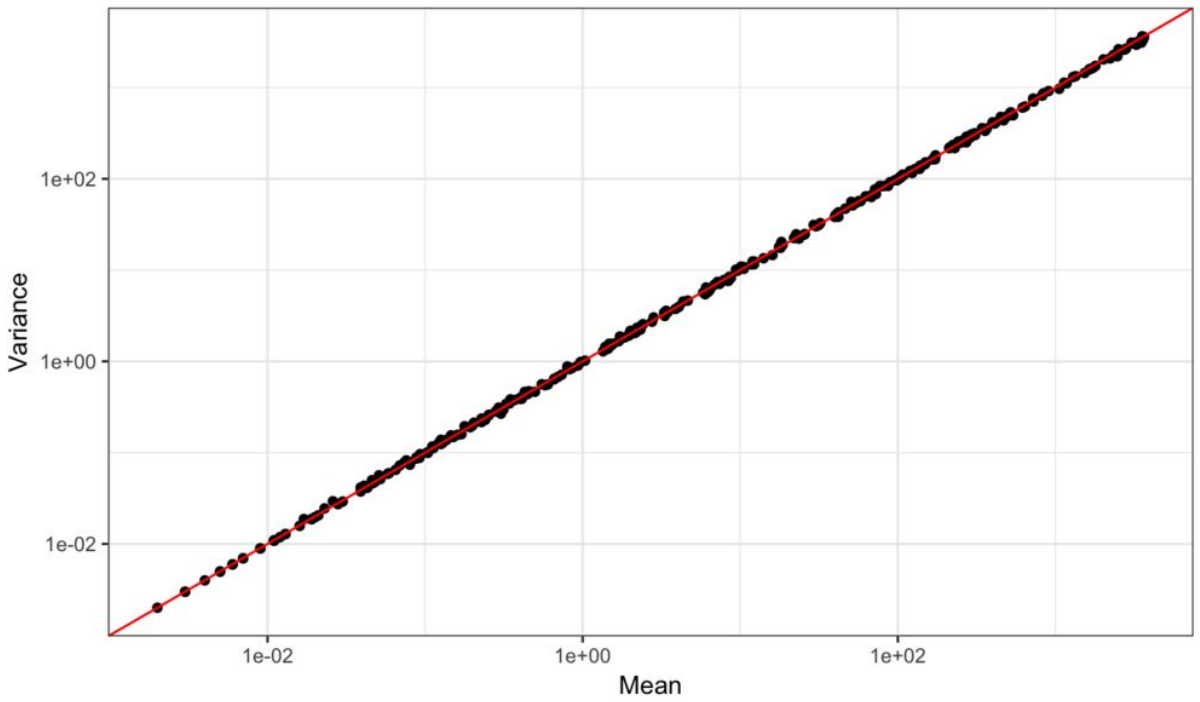
Мы забираем из шляпы 300 шариков, то есть в среднем мы увидим красный шарик 3 раза

Какое будет распределение вероятности различного количества красных шариков, которые мы увидим? Это как раз Пуассон

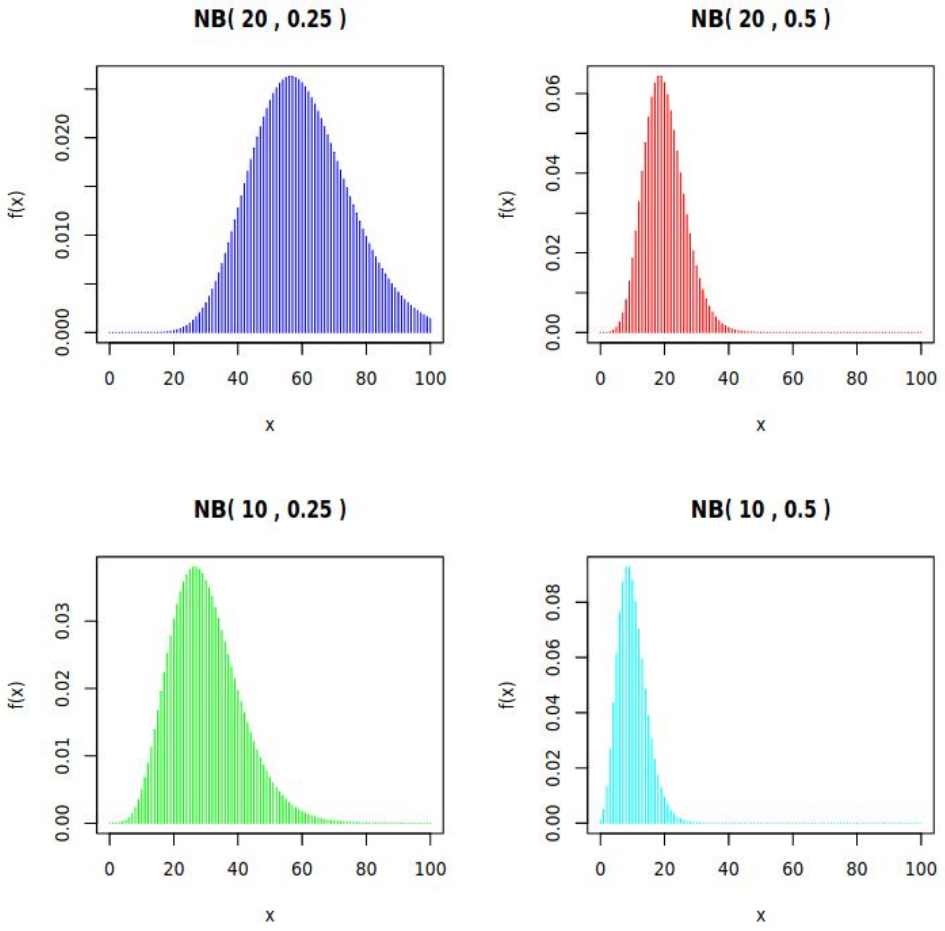
- Шарики = прочтения
- Цвет шарика = ген

Среднее и дисперсия распределения Пуассона

В распределении Пуассона среднее равно дисперсии, а потому достаточно легко понять, если несколько случайных величин распределены по Пуассону



Отрицательное биномиальное распределение



$$NB(K = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k$$

Отрицательное биномиальное распределение определяется как **количество произошедших неудач** в последовательности испытаний Бернулли с вероятностью успеха p, проводимой **до r-го успеха**.

Отрицательное биномиальное распределение

Несложно заметить, что можно таким же образом подсчитать число удач до n-ой неудачи, только теперь в вероятность мы подставим не p, а 1 — p

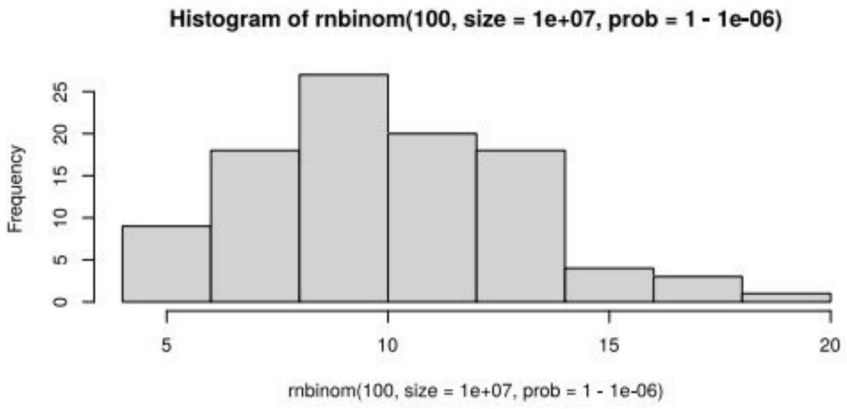
- Допустим, я беру по одному прочтению из образца **X**
- Если прочтение будет из гена ***g***, то это успех (число удач = число каунтов гена)
- Если нет, то неудача (число неудач = глубина секвенирования)
- p — вероятность успеха (= экспрессия гена)

Отрицательное биномиальное распределение

- Допустим ген имеет не очень высокую экспрессию, например, p = 10e-6, а мы секвенируем прочтения по штучке за раз
- Сколько прочтений из этого гена я получу пока не отсеквенирую r = 1e7 прочтений не из этого гена?

Для этого воспользуемся формулой NB(r, 1 — p), которое будет показывать число удач до r-ой неудачи

```
hist(rnbinom(100, size=1e+7, prob=1 - 1e-6))
```

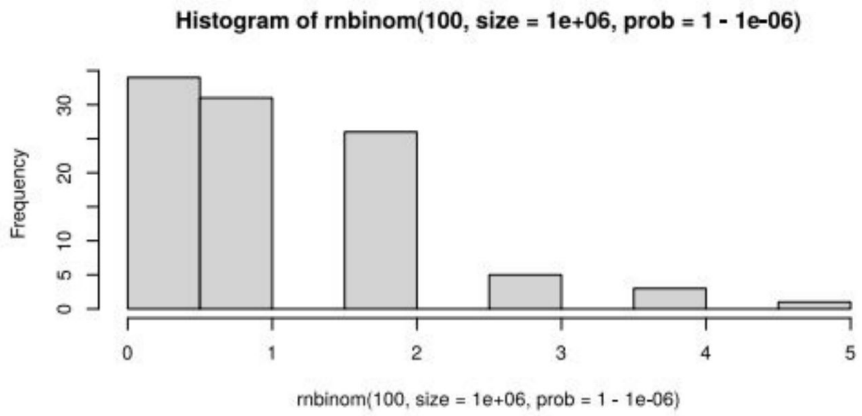


Отрицательное биномиальное распределение

- Допустим ген имеет не очень высокую экспрессию, например, p = 10e-6, а мы секвенируем прочтения по штучке за раз
- Сколько прочтений из этого гена я получу пока не отсеквенирую r = 1e6 прочтений не из этого гена?

Для этого воспользуемся формулой NB(r, 1 — p), которое будет показывать число удач до r-ой неудачи

```
hist(rnbinom(100, size=1e+6, prob=1 - 1e-6))
```

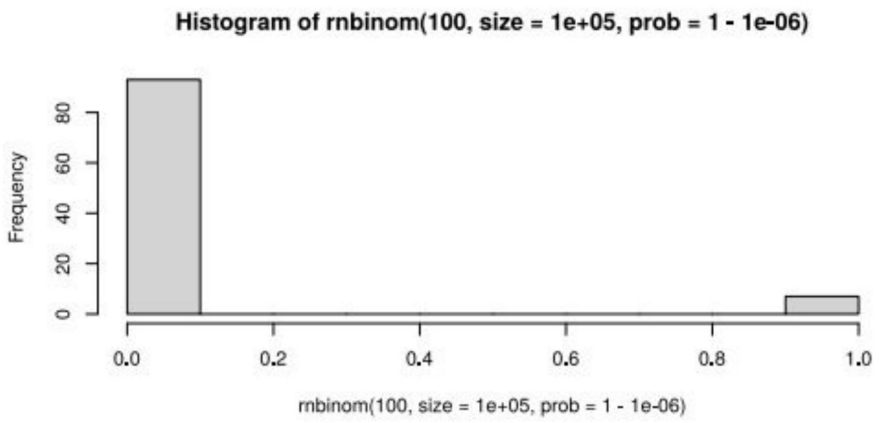


Отрицательное биномиальное распределение

- Допустим ген имеет не очень высокую экспрессию, например, p = 10e-6, а мы секвенируем прочтения по штучке за раз
- Сколько прочтений из этого гена я получу пока не отсеквенирую r = 1e5 прочтений не из этого гена?

Для этого воспользуемся формулой NB(r, 1 — p), которое будет показывать число удач до r-ой неудачи

```
hist(rnbinom(100, size=1e+5, prob=1 - 1e-6))
```



Среднее и дисперсия NB-распределения

Среднее и дисперсия отрицательного биномиального распределения связаны, благодаря чему мы можем инспектировать наши распределения даже без каких-либо тестов на Goodness of Fit

Это свойство называют **овердисперсией**

$$\mathbb{E}[X] = \frac{r(1-p)}{p},$$

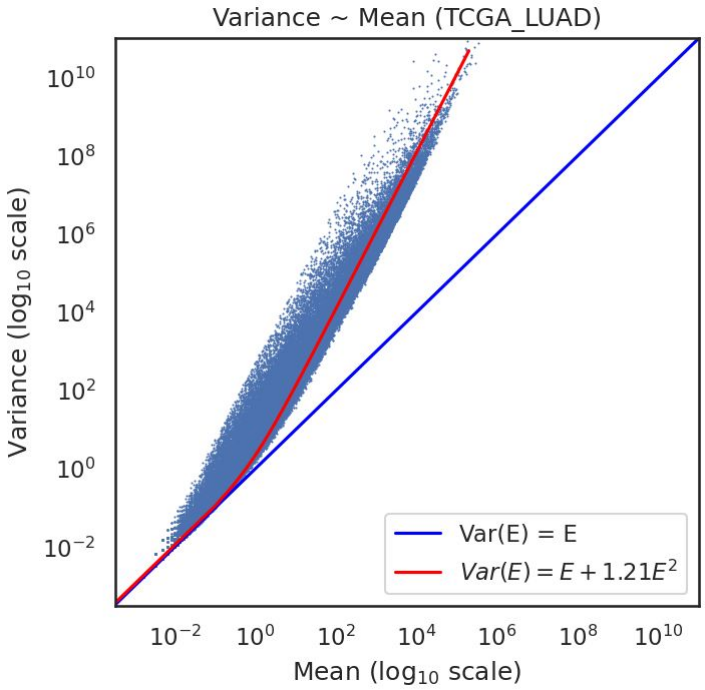
$$\mathrm{Var}[X] = \frac{r(1-p)}{p^2} = \frac{r(1-p)(p+(1-p))}{p^2} = \frac{r(1-p)p + r(1-p)^2}{p^2} =$$

$$= \frac{r(1-p)}{p} + \frac{r(1-p)^2}{p^2} = \mathbb{E}[X] + \frac{1}{r}\frac{r^2(1-p)^2}{p^2} = \mathbb{E}[X] + \frac{1}{r}\mathbb{E}[X]^2,$$

Среднее и дисперсия NB-распределения

Среднее и дисперсия отрицательного биномиального распределения связаны, благодаря чему мы можем инспектировать наши распределения даже без каких-либо тестов на Goodness of Fit

Это свойство называют **овердисперсией**



Как понять распределение наших данных?

1. Допустим, мы считаем, что наши значения описываются некоторым распределением X(a, b)
2. При помощи MLE мы можем оценить наиболее правдоподобные значения параметров этого распределения a и b
3. После этого мы можем посчитать правдоподобие того, что наши данные порождены данной моделью
4. В итоге, используя информацию о правдоподобии данных в контексте данного распределения и числе параметров распределения, мы можем сравнить Goodness of Fitness наших данных различными распределениями

Нормализации

Количество каунтов гена, которые мы видим, зависит от нескольких параметров:

- от длины гена,
- от глубины библиотеки,
- от экспрессии гена,
- от дополнительных факторов, которые сложно оценить.

Для того, чтобы убрать влияние глубины секвенирования и длины (а в особенности чтобы суммировать информацию по экспрессии транскриптов в экспрессию гена, отнормировав на длину каждого из транскриптов), придумали ряд метрик

RPKM и TPM

$$\mathrm{RPKM}_i = \frac{r_i}{l_i \sum_j r_j} \cdot 10^9 \qquad \mathrm{TPM}_i = \frac{r_i}{l_i \sum_j \frac{r_j}{l_j}} \cdot 10^6$$

$\mathbf{RPKM}_i$ is gene's i RPKM metrics,
$\mathbf{r}_i$ is a number of reads mapped on the gene i,
$\mathbf{l}_i$ is an effective length of the gene i

$\mathbf{TPM}_i$ is gene's i TPM metrics,
$\mathbf{r}_i$ is a number of reads mapped on the gene i,
$\mathbf{l}_i$ is an effective length of the gene i

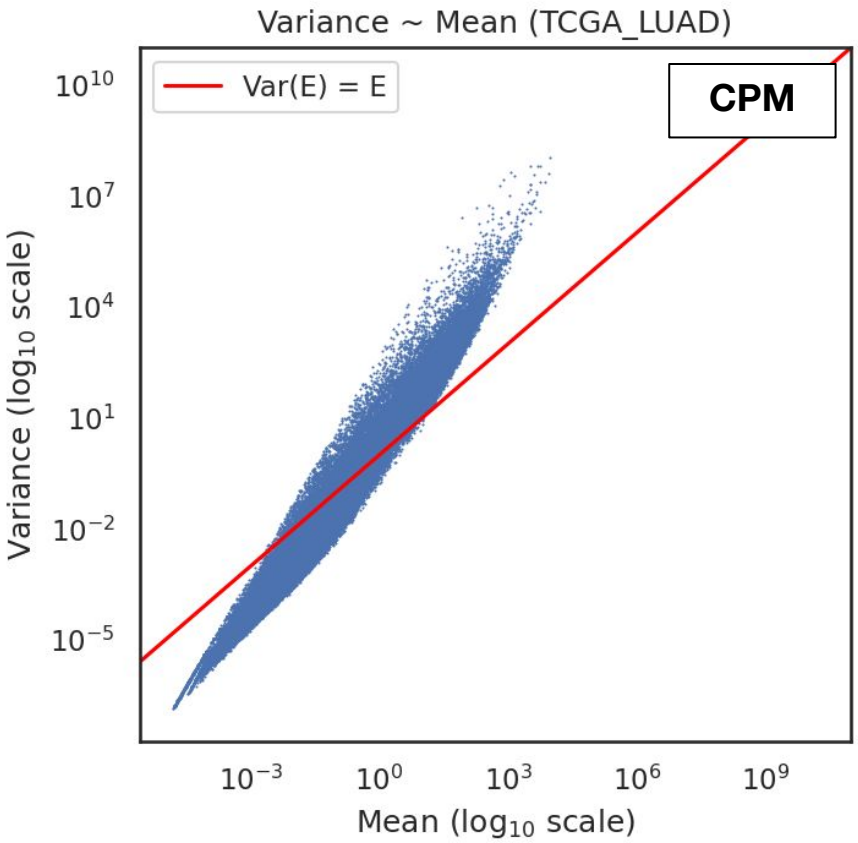
В чём разница?

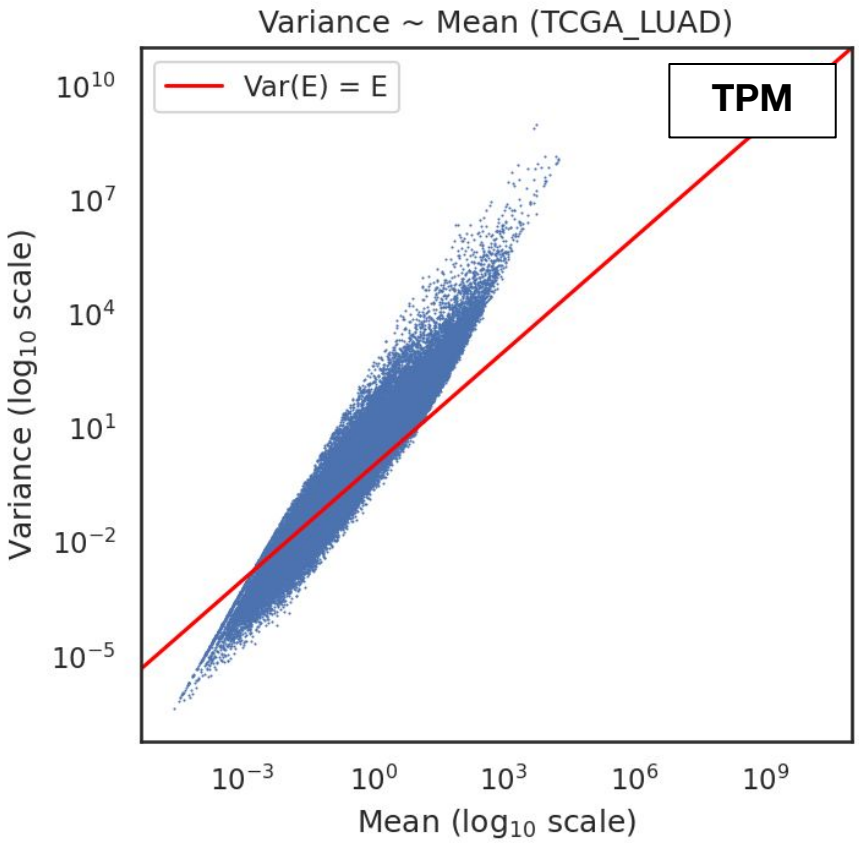
Связь TPM и RPKM

$$\sum_i \mathrm{RPKM}_i = \sum_i \frac{r_i}{l_i \sum_j r_j} \cdot 10^9 = \frac{10^9}{\sum_j r_j} \sum_i \frac{r_i}{l_i},$$

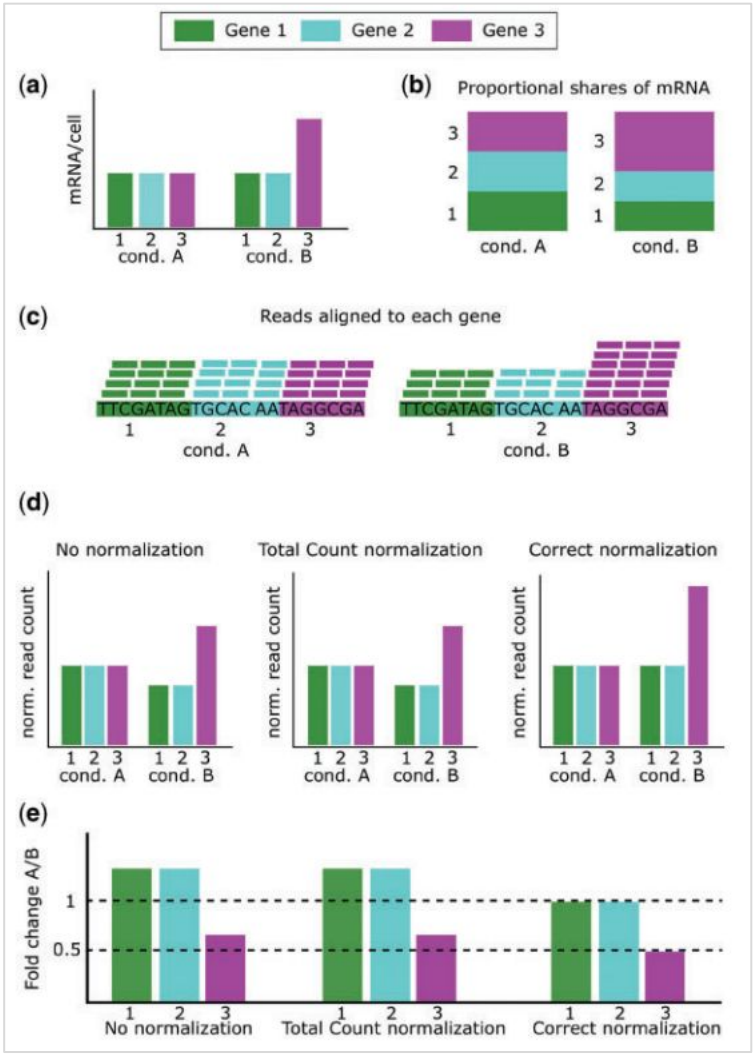
$$\begin{aligned}
\mathrm{TPM}_i &= \frac{r_i}{l_i \sum_j \frac{r_j}{l_j}} \cdot 10^6 = \\
&= \frac{r_i}{l_i \sum_j r_j} \cdot 10^9 \cdot \frac{1}{\sum_j \mathrm{RPKM}_j} \cdot 10^6 = \\
&= \frac{\mathrm{RPKM}_i}{\sum_j \mathrm{RPKM}_j} \cdot 10^6
\end{aligned}$$

Распределение CPM / TPM





Проблемы TPM и RPKM



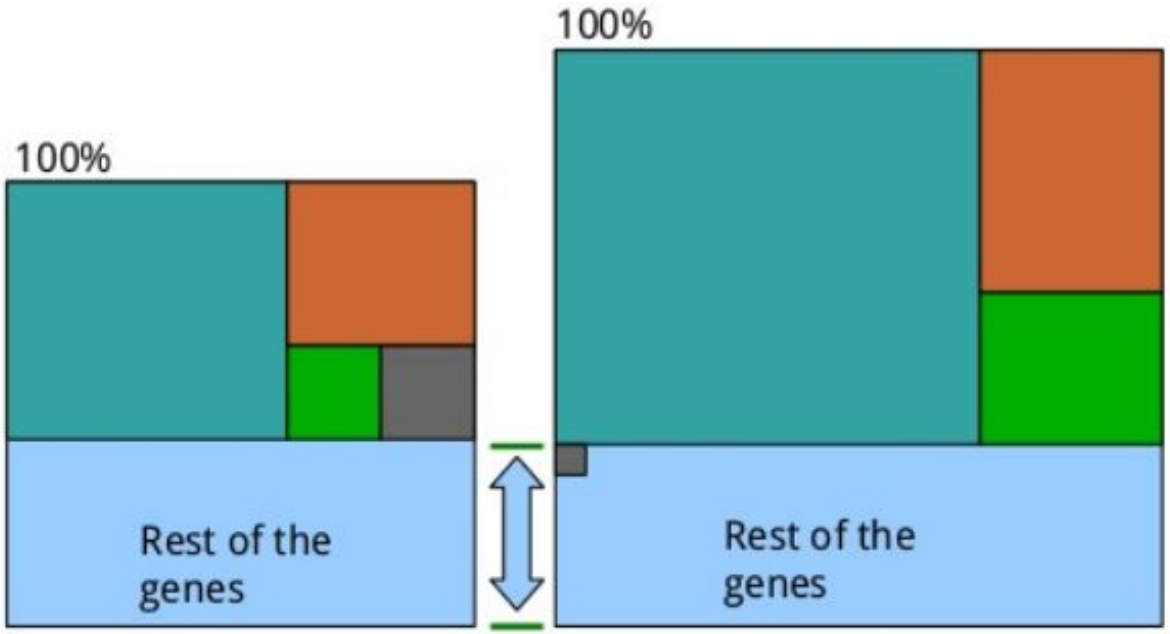
Evans et al., **Brief Bioinform**, 2017

Нормализация на глубину библиотеки предполагает, что суммарное “истинное” количество РНК в клетке константно

Это не работает в случае, когда, например, экспрессия одного набора генов увеличилась, а других — не поменялась

Корректная нормализация

При корректной нормализации (которую, например, выполняет DESeq2 или edgeR) мы принимаем во внимание, что большая часть генов не меняет свою экспрессию между образцами



Нормализация в DESeq2 (RLE)

gene	sampleA	sampleB	pseudo-reference sample
EF2A	1489	906	sqrt(1489 * 906) = **1161.5**
ABCD1	22	13	sqrt(22 * 13) = **17.7**
…	…	…	…

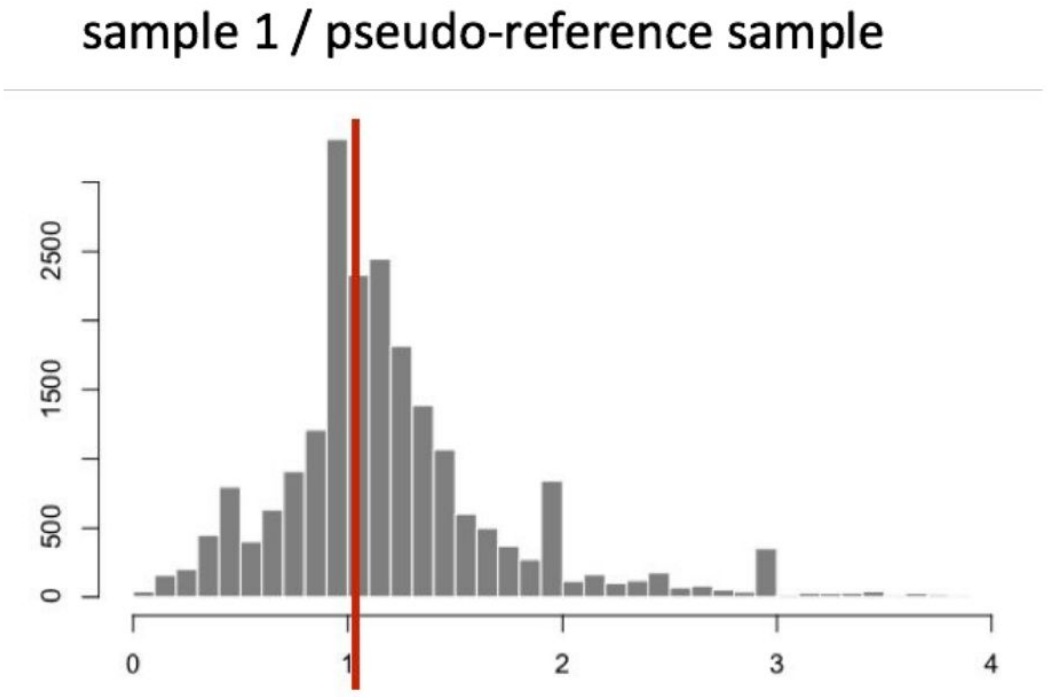
Нормализация в DESeq2 (RLE)

gene	sampleA	sampleB	pseudo-reference sample	ratio of sampleA/ref	ratio of sampleB/ref
EF2A	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ABCD1	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**
MEFV	793	410	570.2	793/570.2 = **1.39**	410/570.2 = **0.72**
BAG1	76	42	56.5	76/56.5 = **1.35**	42/56.5 = **0.74**
MOV10	521	1196	883.7	521/883.7 = **0.590**	1196/883.7 = **1.35**

Нормализация в DESeq2 (RLE)

```
normalization_factor_sampleA <- median(c(1.28, 1.3, 1.39, 1.35, 0.59))
```

```
normalization_factor_sampleB <- median(c(0.78, 0.77, 0.72, 0.74, 1.35))
```



Нормализация в DESeq2 (RLE)

SampleA median ratio = 1.3

SampleB median ratio = 0.77

gene	sampleA	sampleB
EF2A	1489	906
ABCD1	22	13
…	…	…

gene	sampleA	sampleB
EF2A	1489 / 1.3 = **1145.39**	906 / 0.77 = **1176.62**
ABCD1	22 / 1.3 = **16.92**	13 / 0.77 = **16.88**
…	…	…

Итого по нормализациям

- **CPM** — простое сравнение одинаковых генов каунтов между разными образцами, грубая нормировка только на глубину библиотеки. Не для DE
- **RPKM** — сравнение генов внутри одного образца (например, для ранговых методов, о которых поговорим дальше). Не для DE
- **TMP** — сравнение генов как внутри одного образца (для ранговых методов), так и грубого — между образцами (но не для DE!)
- **RLE** и **TMM** — сравнение генов между разными образцами (в том числе и для DE), но не внутри одного образца (отсутствует нормировка на длину)