

«Анализ данных NGS»

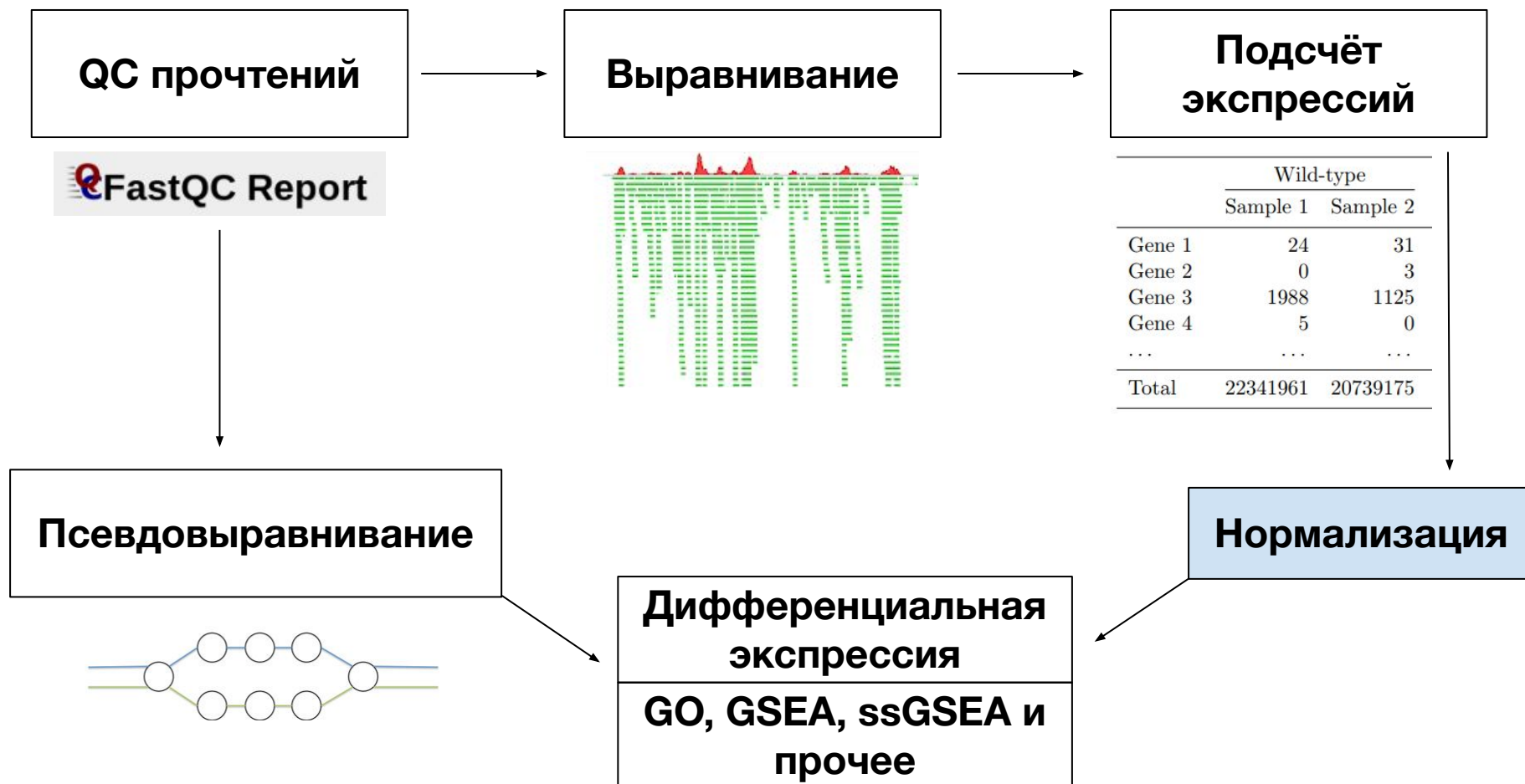
Лекция #3

Анализ bulk RNA-Seq

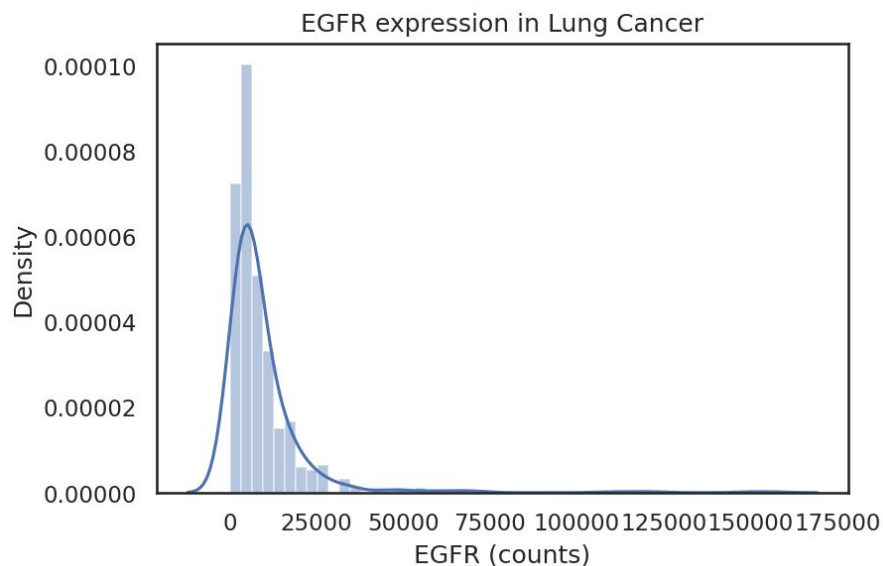
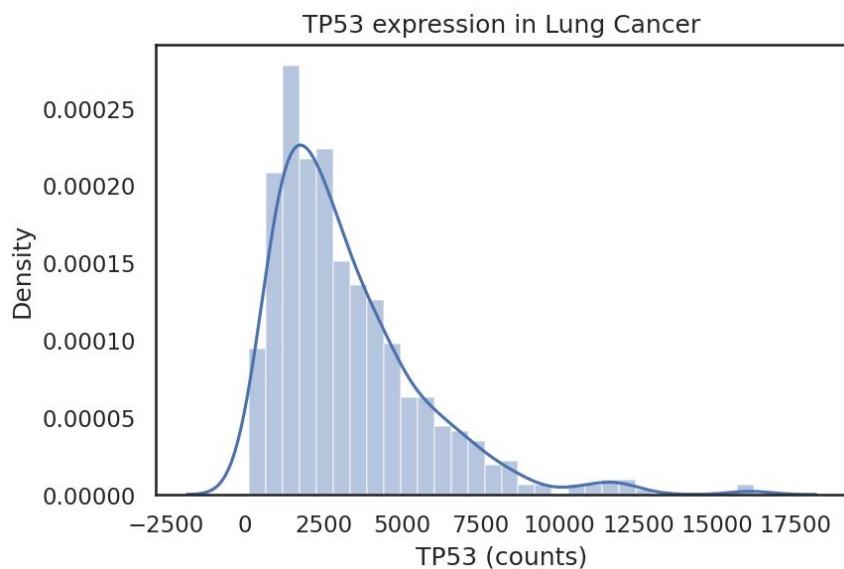
Серёжа Исаев

аспирант **MedUni Vienna**

Дорожная карта анализа RNA-Seq



Распределение каунтов генов

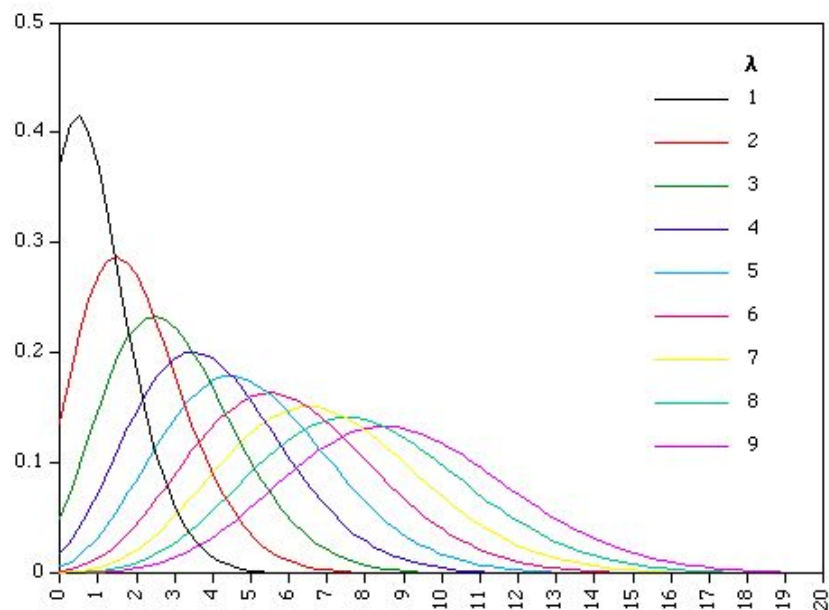


Экспрессии генов TP53 и EGFR в образцах рака лёгкого

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).

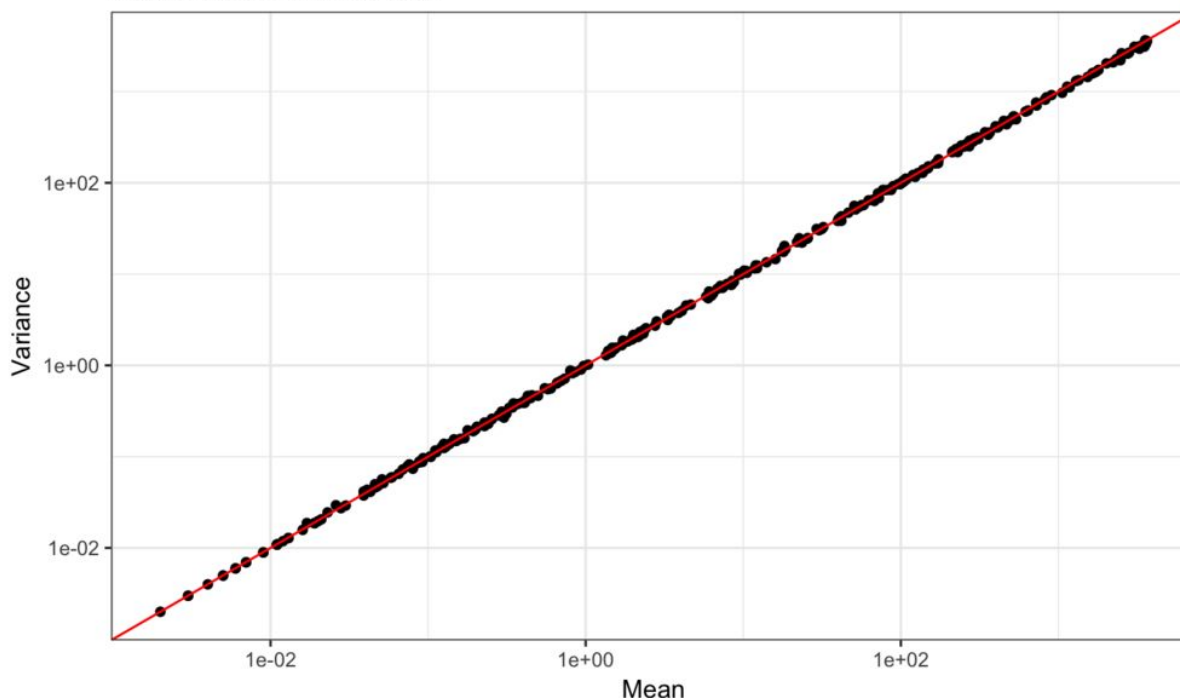
Мы забираем из шляпы 300 шариков, то есть в среднем мы увидим красный шарик 3 раза

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

- Шарик = прочтение
- Цвет шарика = ген

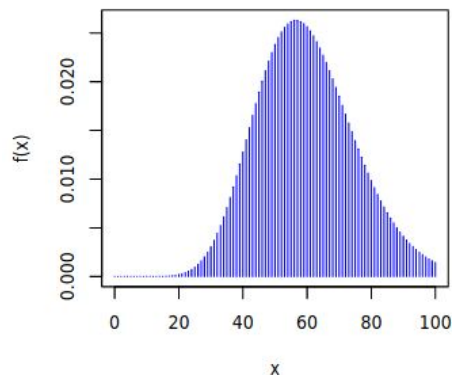
Среднее и дисперсия распределения Пуассона

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

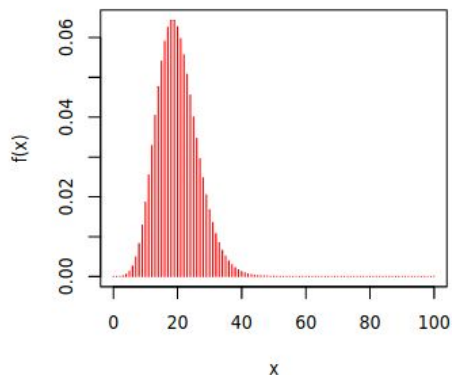


Отрицательное биномиальное распределение

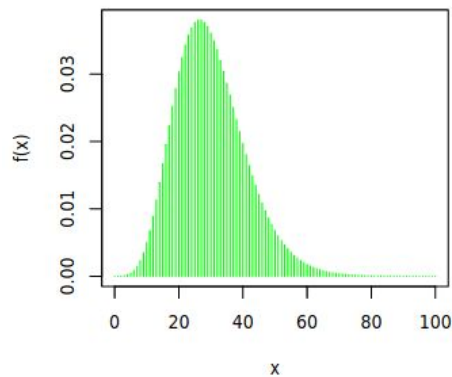
NB(20 , 0.25)



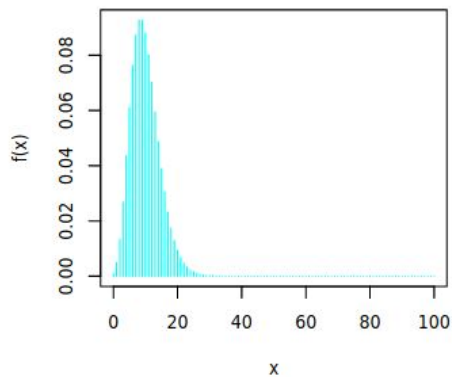
NB(20 , 0.5)



NB(10 , 0.25)



NB(10 , 0.5)



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

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 $1 - p$

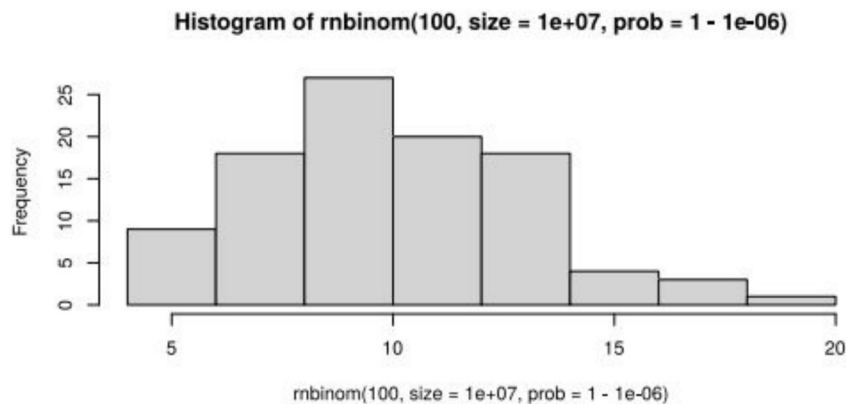
- Допустим, я беру по одному прочтению из образца **X**
- Если прочтение будет из гена **g**, то это успех (число удач = число каунтов гена)
- Если нет, то неудача (число неудач = глубина секвенирования)
- p — вероятность успеха (= экспрессия гена)

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 отсеку гена? $r = 1e7$ 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))
```

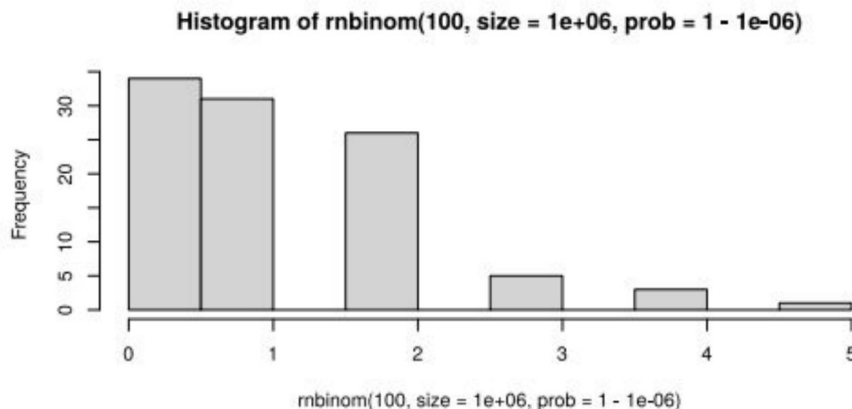


Отрицательное биномиальное распределение

- 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 отсеку гена? $r = 1e6$ 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))
```

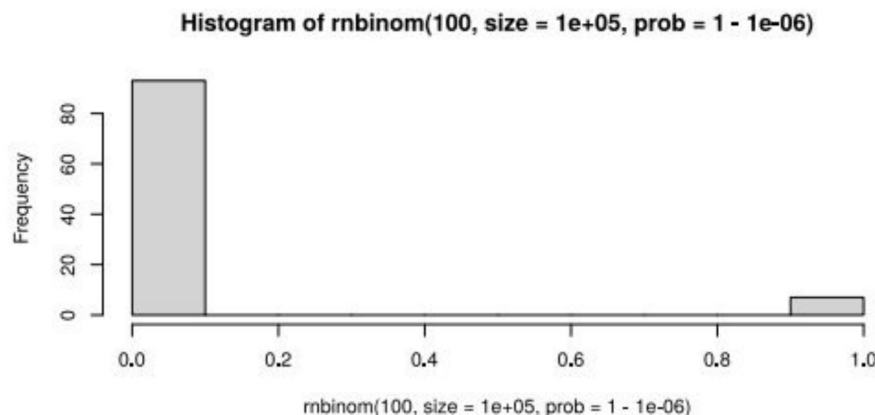


Отрицательное биномиальное распределение

- 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 отсеку гена? $r = 1e5$ 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

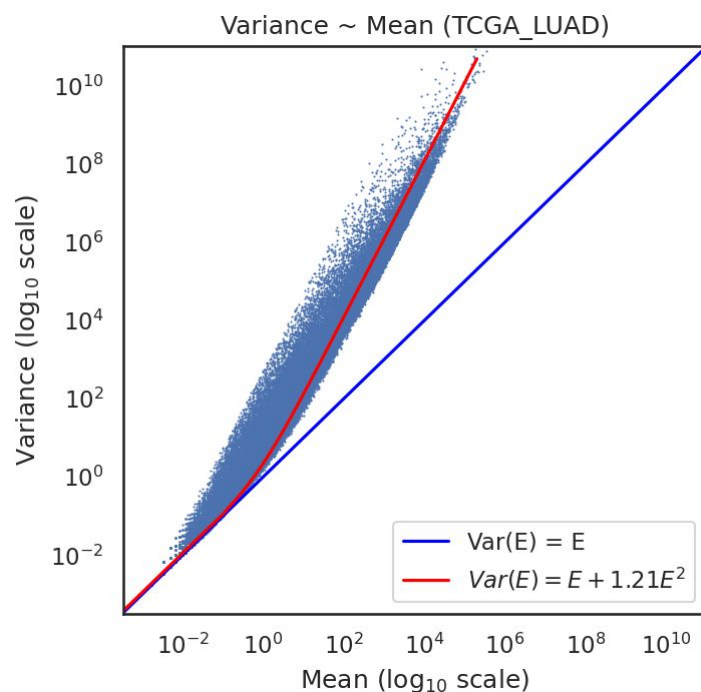
Это свойство называют **овердисперсией**

$$\begin{aligned}\mathbb{E}[X] &= \frac{r(1-p)}{p}, \\ \text{Var}[X] &= \frac{r(1-p)}{p^2} = \frac{r(1-p)(p + (1-p))}{p^2} = \frac{r(1-p)p + r(1-p)^2}{p^2} = \\ &= \frac{r(1-p)}{p} + \frac{r(1-p)^2}{p^2} = \mathbb{E}[X] + \frac{1}{r} \frac{r^2(1-p)^2}{p^2} = \mathbb{E}[X] + \frac{1}{r} \mathbb{E}[X]^2,\end{aligned}$$

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

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

RPKM_{*i*} is gene's *i* RPKM metrics,
r_{*i*} is a number of reads mapped on the gene *i*,
l_{*i*} is an effective length of the gene *i*

TPM_{*i*} is gene's *i* TPM metrics,
r_{*i*} is a number of reads mapped on the gene *i*,
l_{*i*} is an effective length of the gene *i*

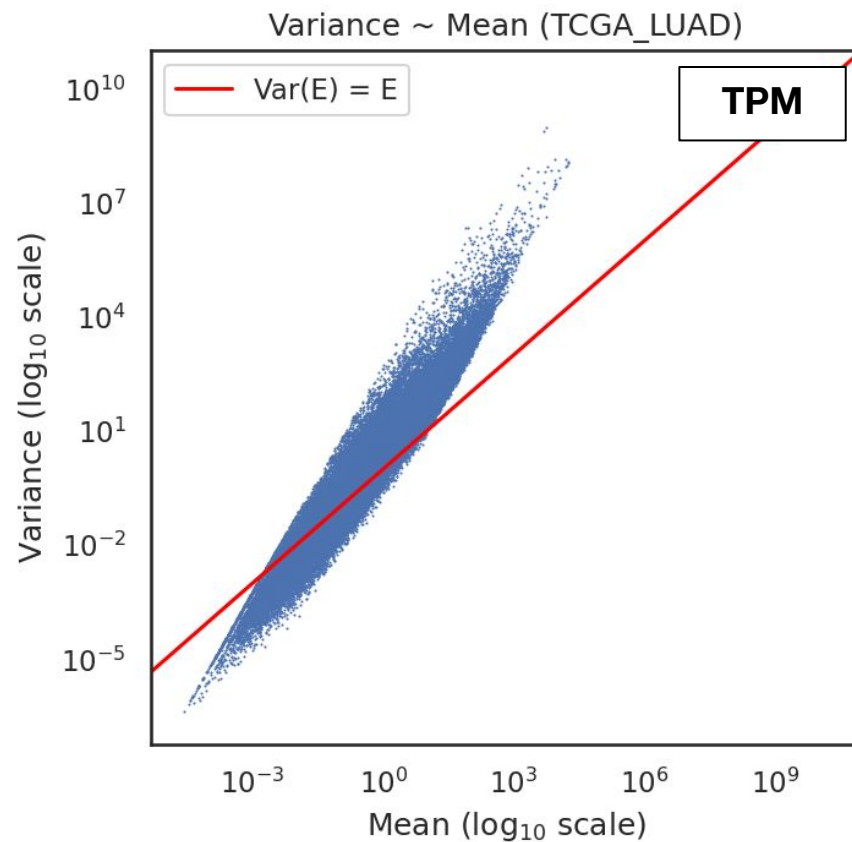
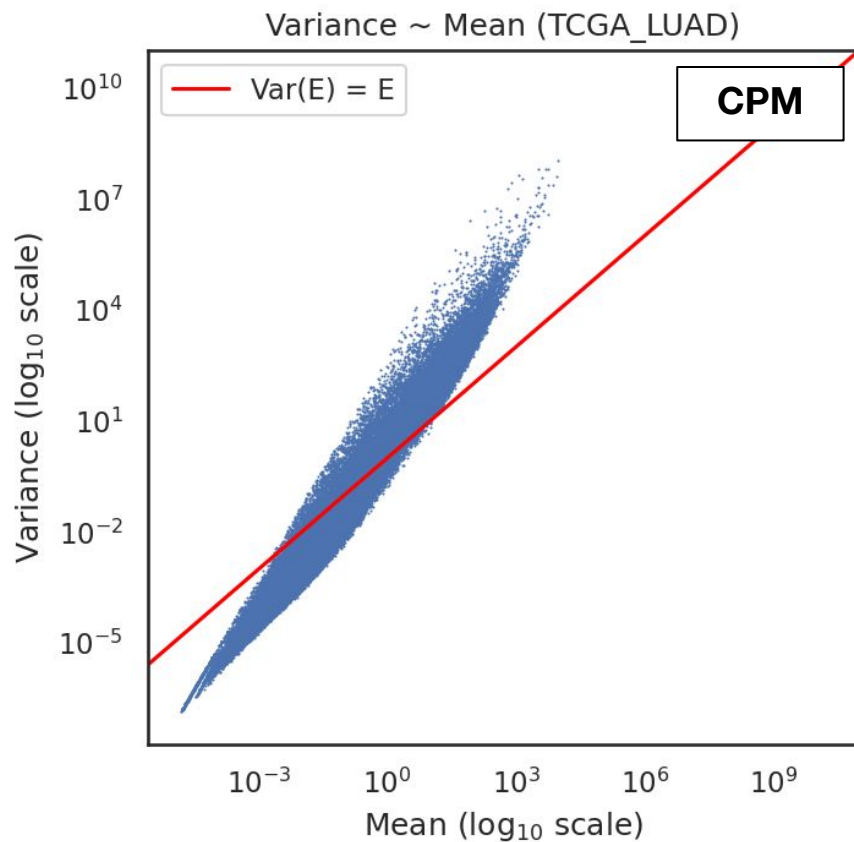
В чём разница?

Связь TPM и RPKM

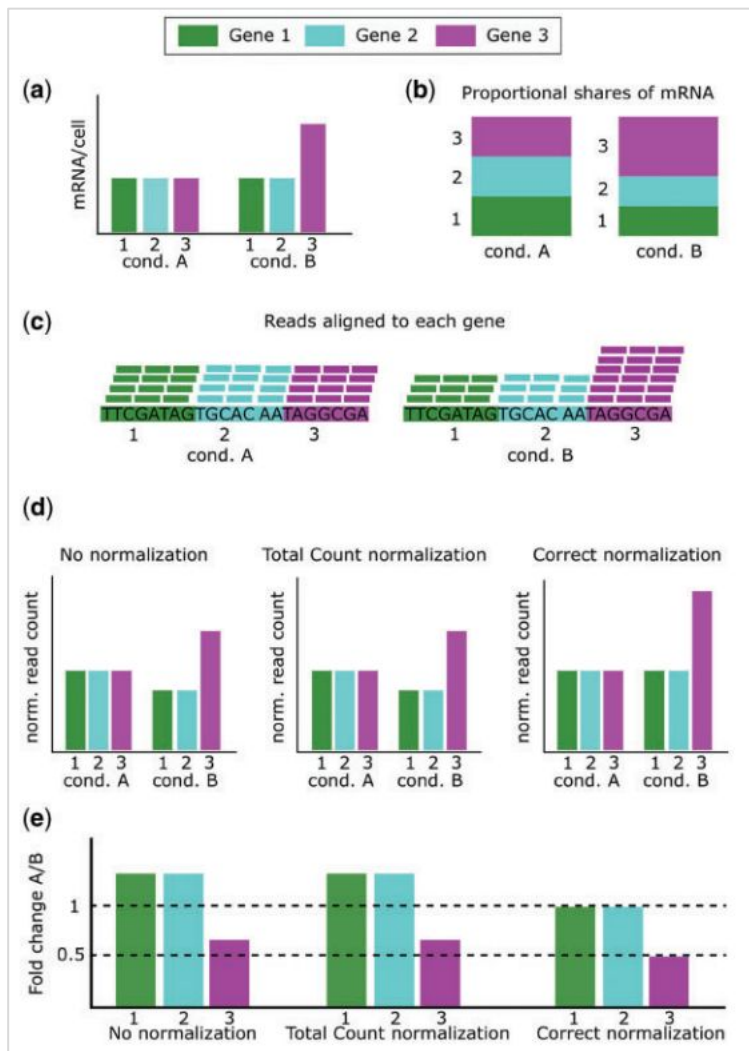
$$\sum_i \text{RPKM}_i = \sum_i \frac{r_i}{l_i \sum_j r_j} \cdot 10^9 = \frac{10^9}{\sum_j r_j} \sum_i \frac{r_i}{l_i},$$

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

Распределение CPM / TPM



Проблемы TPM и RPKM

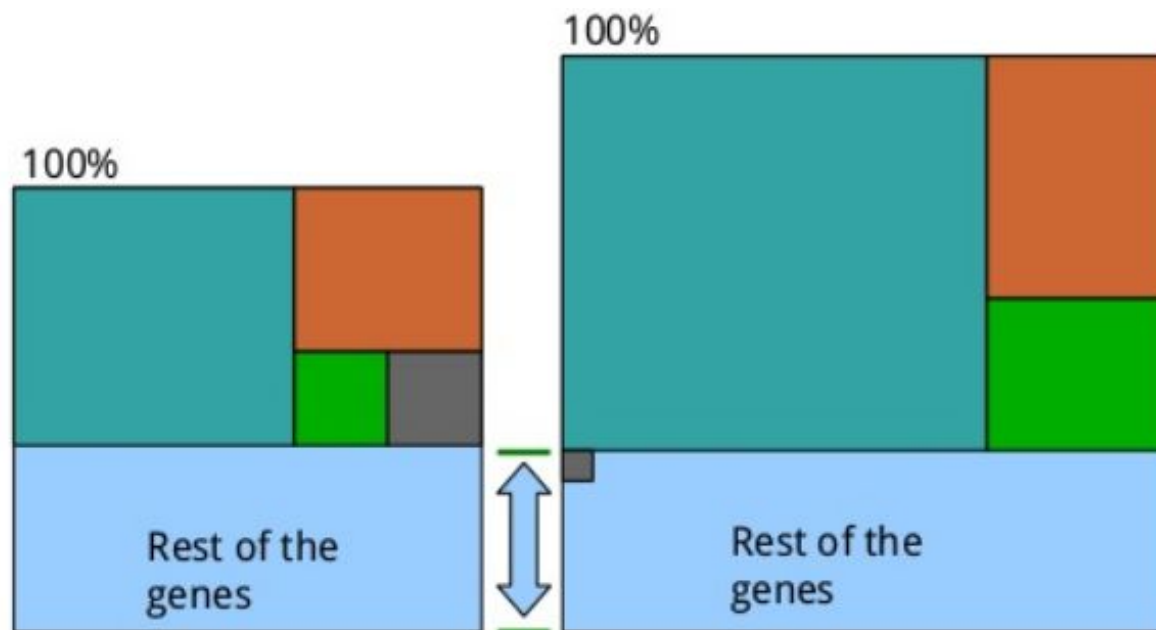


Нормализация на глубину библиотеки предполагает, что суммарное “истинное” количество РНК в клетке константно

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	$\text{sqrt}(1489 * 906) = \mathbf{1161.5}$
ABCD1	22	13	$\text{sqrt}(22 * 13) = \mathbf{17.7}$
...

Нормализация в DESeq2 (RLE)

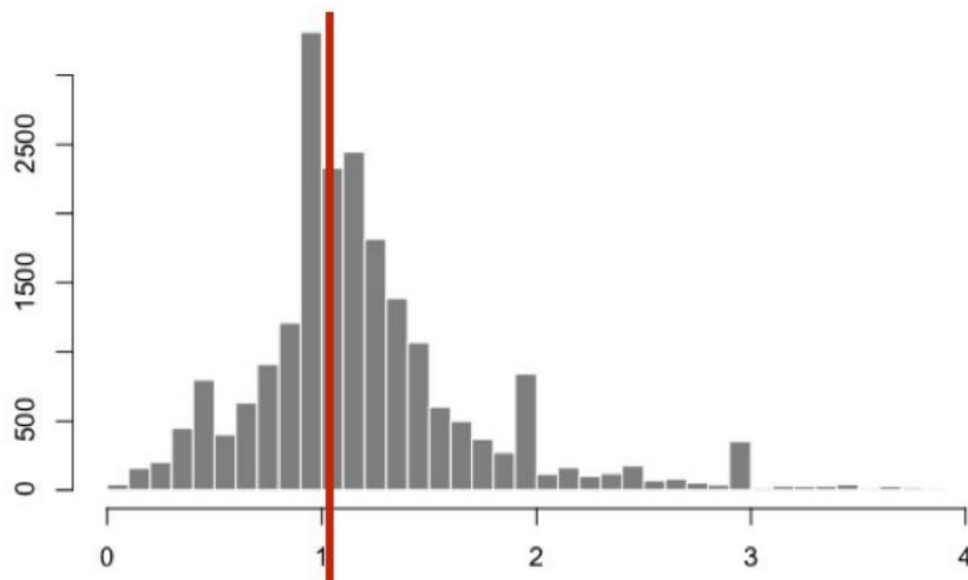
gene	sampleA	sampleB	pseudo-reference sample	ratio of sampleA/ref	ratio of sampleB/ref
EF2A	1489	906	1161.5	$1489/1161.5 = \mathbf{1.28}$	$906/1161.5 = \mathbf{0.78}$
ABCD1	22	13	16.9	$22/16.9 = \mathbf{1.30}$	$13/16.9 = \mathbf{0.77}$
MEFV	793	410	570.2	$793/570.2 = \mathbf{1.39}$	$410/570.2 = \mathbf{0.72}$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

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))
```

sample 1 / pseudo-reference sample



Нормализация в DESeq2 (RLE)

SampleA median ratio = 1.3

SampleB median ratio = 0.77

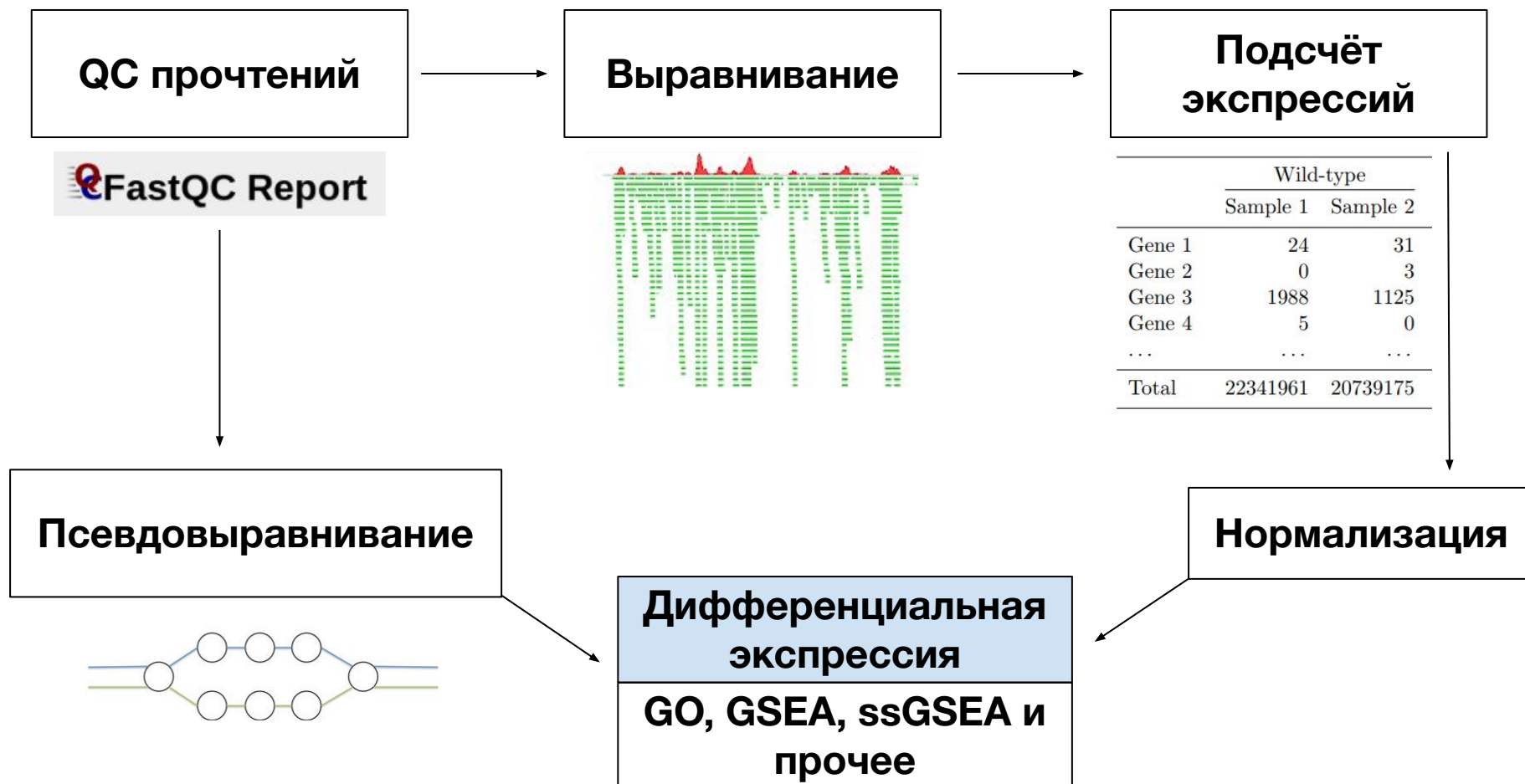
gene	sampleA	sampleB
EF2A	1489	906
ABCD1	22	13
...

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)

Дорожная карта анализа RNA-Seq



Суть задачи

Нам необходимо статистически сравнить среднее экспрессий между двумя выборками образцов

Что бы мы сделали в классическом случае?

1. Тест Манна-Уитни,
2. t-test

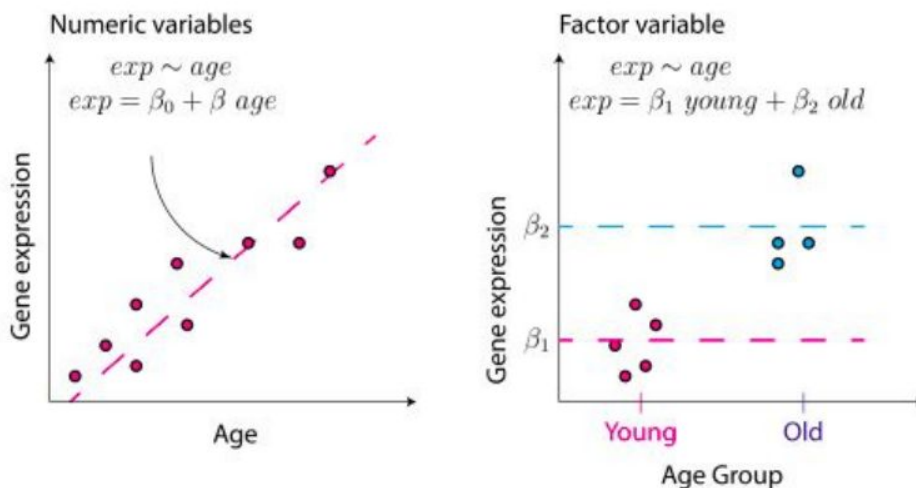
Проблема в том, что тест Манна-Уитни будет слишком слабый, так как чаще всего у нас мало точек в каждой из выборок, а t-test просто не подойдёт потому, что наши данные распределены не нормально

Что делать?

Причём тут регрессия?

С одной стороны, регрессионные модели могут позволить нам оценить статистическую достоверность разниц в средних

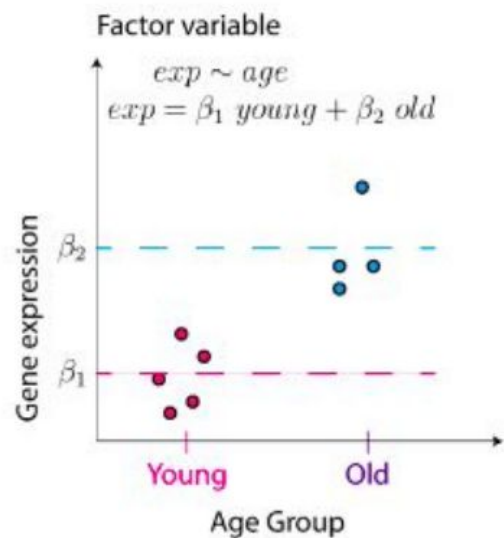
С другой стороны, GLM позволяют обобщить регрессию на ненормальные распределения



Причём тут регрессия?

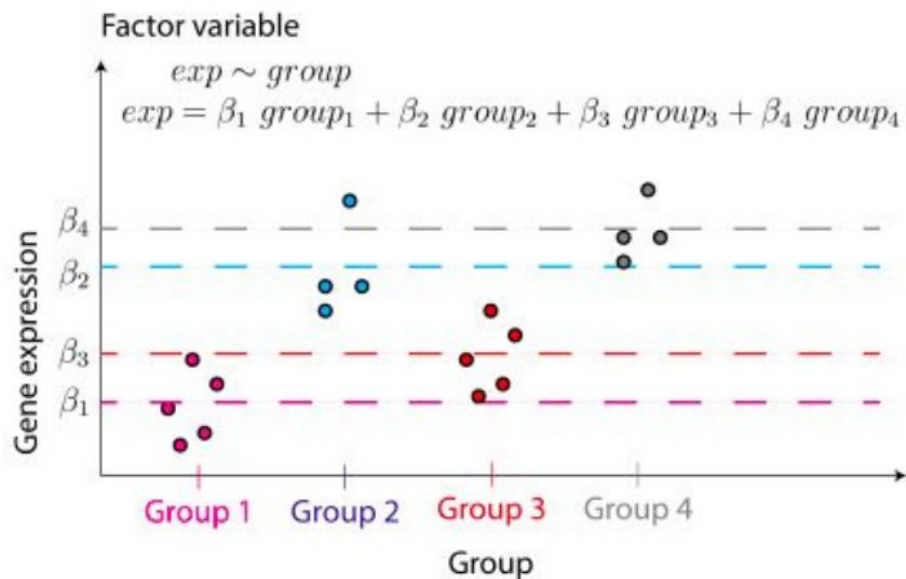
Статистический вопрос, который мы будем извлекать из регрессии, — значимо ли различаются параметры β_1 и β_2 ?

Это можно сказать, сравнив правдоподобия моделей или при помощи других подходов (будет оговорено дальше)



Причём тут регрессия?

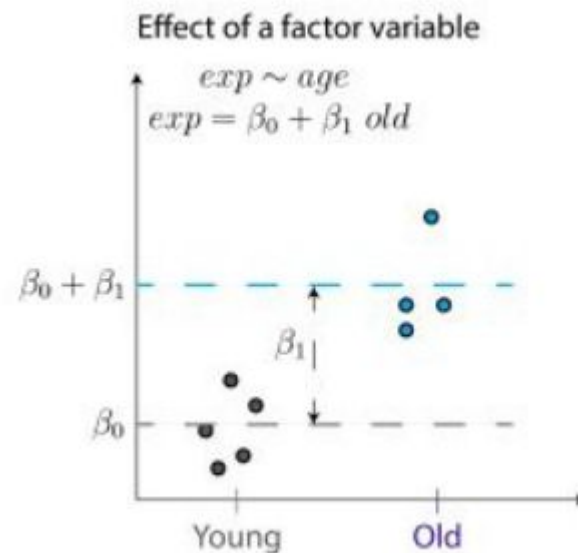
Линейную модель можно обобщить и добавить более двух уровней фактора, чтобы сравнивать сразу несколько категорий



Intercept

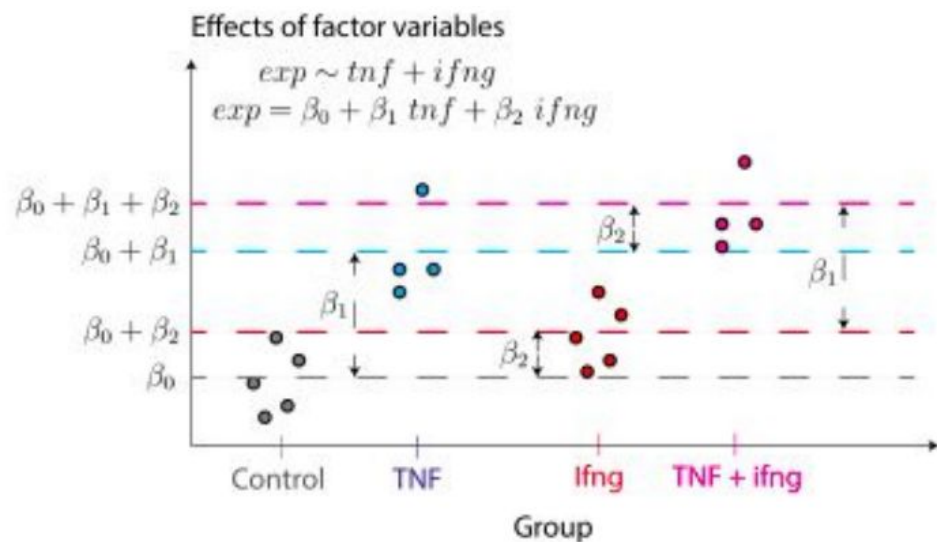
Вместо того, чтобы сравнивать значимость разницы между β_1 и β_2 , обычно используют модель со свободным членом β_0 и после этого вычисляют значимость β_1

Свободный член в данном случае называют словом **intercept**



Intercept

Эту же логику можно обобщить и на модели с несколькими категориями в целевой переменной



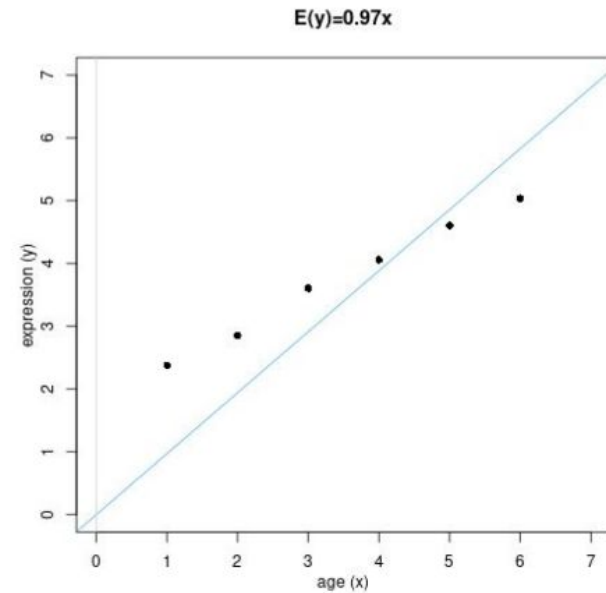
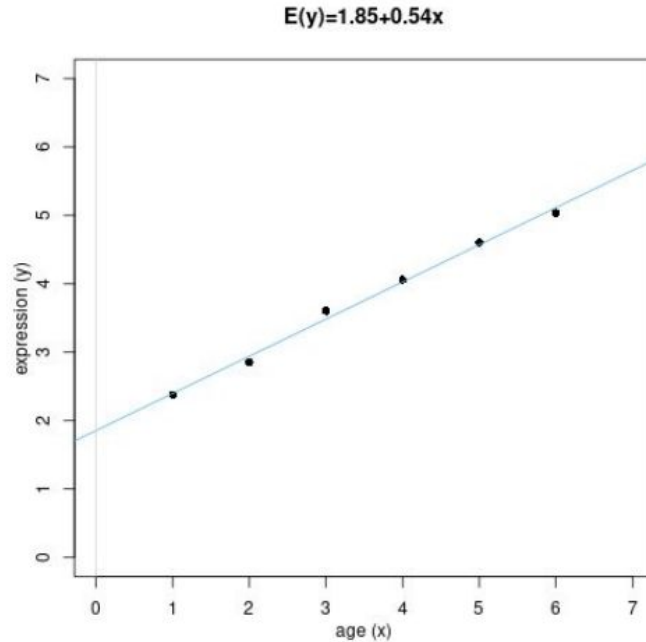
Линейные модели

$$y \sim 0 + \text{feature1} + \text{feature2} + \dots$$

без intercept

$$y \sim 1 + \text{feature1} + \text{feature2} + \dots$$

c intercept



Какие переменные включают в модель?

Таргет:

- экспериментальные условия,

сопутствующие факторы:

- пациент,
- пол,
- возраст,
- ... (всё, что может иметь влияние на экспрессию)

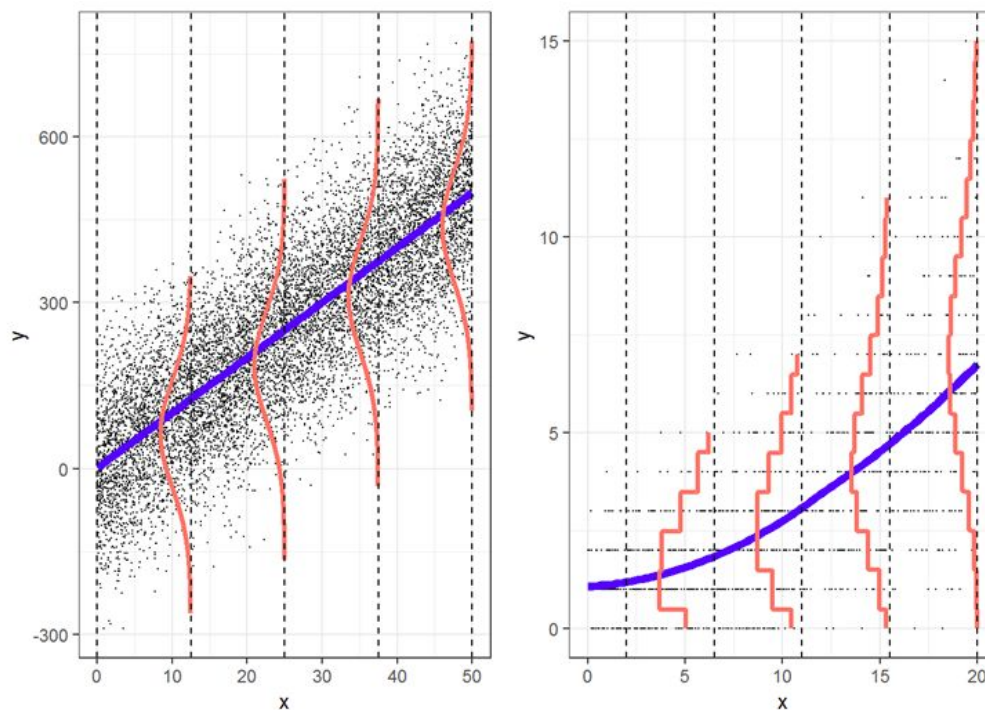
Что не включают:

- техническую повторяемость

Обобщённые линейные модели (GLM)

В обобщённой линейной модели нет требования к **нормальности и гомоскедастичности остатков**

Коэффициенты определяются при помощи MLE



Модель DESeq2

Модель, которая вшита в DESeq2, может описываться следующим образом:

$$K_{i,j} \sim NB(\mu_{i,j}, \alpha_i)$$

$$\mu_{i,j} = s_j p_{i,j}$$

$$\log_2(p_{i,j}) = x_{j,A} \beta_{i,A} + x_{j,B} \beta_{i,B}$$

- Where, $K_{i,j}$ is matrix of observed counts (known),
- $\mu_{i,j}$ is a mean for NB distribuion,
- $p_{i,j}$ is a probability to get read i from sample j
- s_j is a scaling factor (will be calculated), α_i are gene dispersions (will be calculated),
- matrix x is model coefficients (zero or one depending on conditions) and most importantly
- $\beta_{i,j}$ (log-)probability to get read from gene i if a sample is from condition

Последовательность действий DESeq2

1. Сначала происходит оценка size factor'a,
2. потом происходит оценка дисперсии и затем
3. происходит оценка параметров β модели при помощи GLM

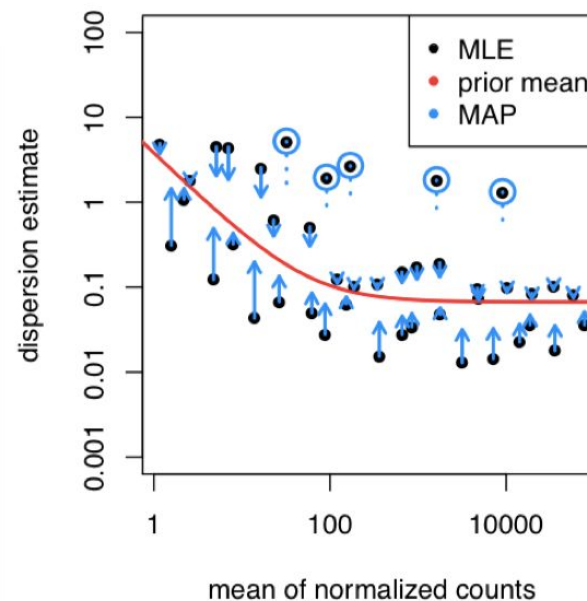
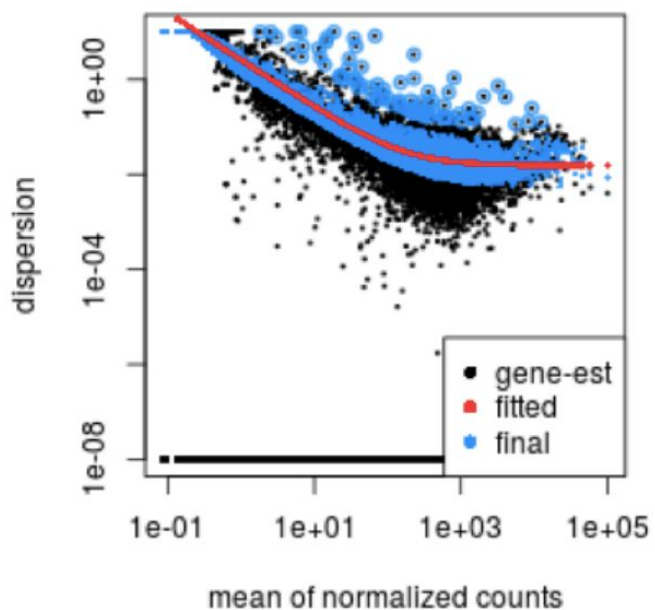
$$K_{i,j} \sim NB(\mu_{i,j}, \alpha_i)$$

$$\mu_{i,j} = s_j p_{i,j}$$

$$\log_2(p_{i,j}) = x_{j,A} \beta_{i,A} + x_{j,B} \beta_{i,B}$$

Подрезание дисперсии

При малых размерах выборки оценка дисперсии становится достаточно неточной, поэтому используют процедуру *подрезание дисперсии*



Взаимодействие переменных

Удобным способом понимания и отображения того, что с чем сравнивается в дизайне экспериментов по секвенированию РНК могут служить модельные матрицы

Модельные матрицы содержат 0 или 1 для каждого из элементов линейной модели

```
model.matrix(~1+condition+time+condition:time, samples)
```

Рассмотрим примеры модельных матриц для разных дизайнов (по материалам Hugo Tavares)

Один фактор, два уровня

Condition:



colData

	condition <factor>
sample1	shade
sample2	shade
sample3	shade
sample4	sun
sample5	sun
sample6	sun

Один фактор, два уровня

Condition:



colData

	condition <factor>
sample1	shade
sample2	shade
sample3	shade
sample4	sun
sample5	sun
sample6	sun

Design: ~ 1 + condition

$$\text{Expr} = \beta_0 + \beta_1 \text{CondSun}$$

Один фактор, два уровня

Condition:



colData

	condition <factor>
sample1	shade
sample2	shade
sample3	shade
sample4	sun
sample5	sun
sample6	sun

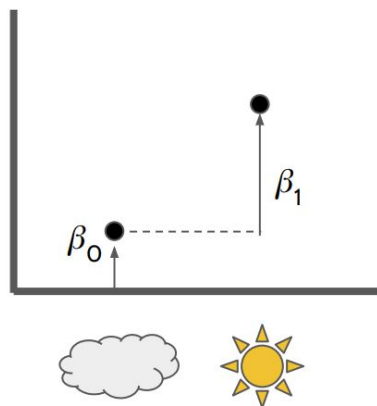
Коэффициенты из DESeq:

β_0 = Intercept

β_1 = condition_sun_vs_shade

Null hypothesis:

$\beta_1 = 0$



Иногда можно немного переписать модель для упрощенной интерпретации

Design: $\sim 0 + \text{condition}$

Expr = $\beta_0 + \beta_1 \text{CondSun}$

Кодируется переменной со значениями 0/1

Model matrix

	(Intercept)	conditionsun
sample1	1	0
sample2	1	0
sample3	1	0
sample4	1	1
sample5	1	1
sample6	1	1

Один фактор, два уровня

Condition:



colData

	condition <factor>
sample1	shade
sample2	shade
sample3	shade
sample4	sun
sample5	sun
sample6	sun

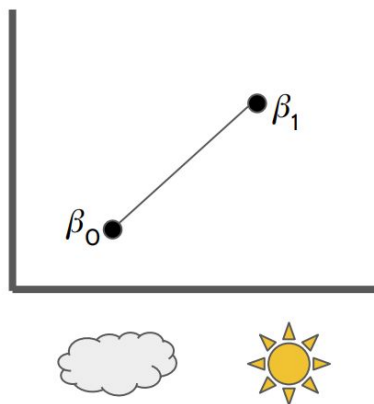
Иногда можно немного переписать модель для упрощенной интерпретации

Design: $\sim 0 + \text{condition}$

$$\text{Expr} = \beta_0 \text{Shade} + \beta_1 \text{Sun}$$

Null hypothesis:

$$\beta_1 - \beta_0 = 0$$



Кодируется переменной со значениями 0/1

Model matrix

	(Intercept)	conditionsun
sample1	1	0
sample2	1	0
sample3	1	0
sample4	1	1
sample5	1	1
sample6	1	1

Один фактор, три уровня

Colour:



Коэффициенты из DESeq:

β_0 = Intercept

β_1 = colour_pink_vs_white

β_2 = colour_yellow_vs_white

```
colour
<factor>
sample1 pink
sample2 pink
sample3 pink
sample4 yellow
sample5 yellow
sample6 yellow
sample7 white
sample8 white
sample9 white
```

Design: $\sim 1 + \text{colour}$

Expr = $\beta_0 + \beta_1 \text{ColPink} + \beta_2 \text{ColYellow}$

Нулевая гипотеза:

Pink vs White

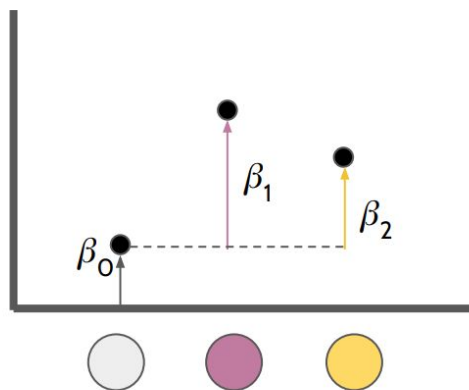
$\beta_1 = 0$

Yellow vs White

$\beta_2 = 0$

Pink vs Yellow

$\beta_1 - \beta_2 = 0$

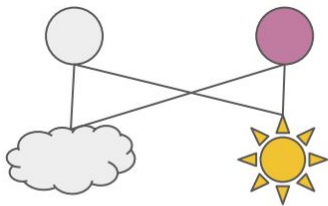


Model matrix

	(Intercept)	colourpink	colouryellow
sample1	1	1	0
sample2	1	1	0
sample3	1	1	0
sample4	1	0	1
sample5	1	0	1
sample6	1	0	1
sample7	1	0	0
sample8	1	0	0
sample9	1	0	0

Два фактора и взаимодействие

Colour:



Condition:

Design:

```
~ 1 + colour + condition + colour:condition
```

Нулевая гипотеза:

Pink vs White (Shade)

$$\beta_1 = 0$$

Pink vs White (Sun)

$$\beta_1 + \beta_3 = 0$$

Sun vs Shade (White):

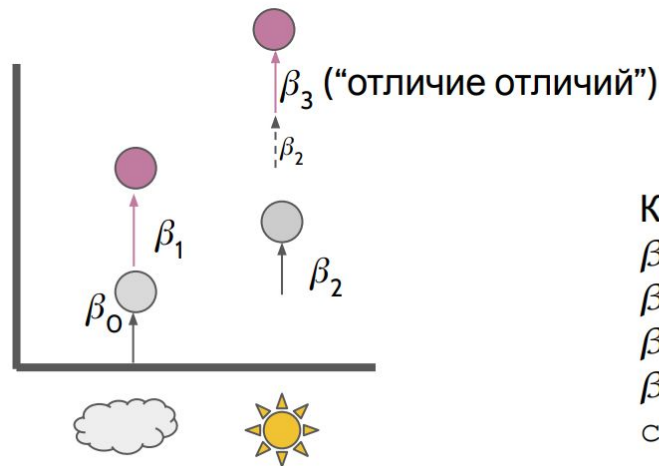
$$\beta_2 = 0$$

Sun vs Shade (Pink):

$$\beta_2 + \beta_3 = 0$$

Interaction:

$$\beta_3 = 0$$



Коэффициенты из DESeq:

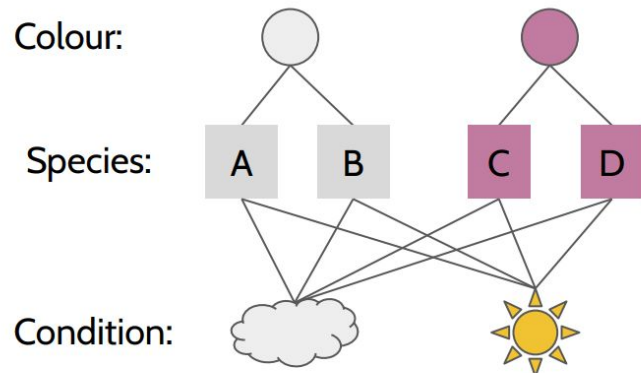
β_0 = Intercept

β_1 = colour_pink_vs_white

β_2 = condition_sun_vs_shade

β_3 =
colourpink.condition_sun

Три фактора с вложенностью



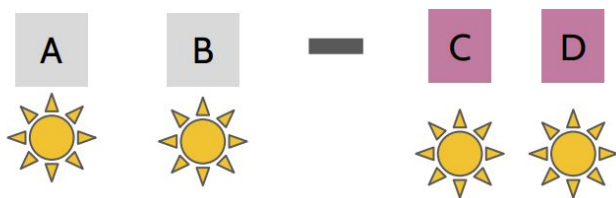
Species вложен в colour.

Species полностью входит в colour, поэтому в дизайн colour не включаем (но это есть смысл учесть при создании контрастов).

Design:

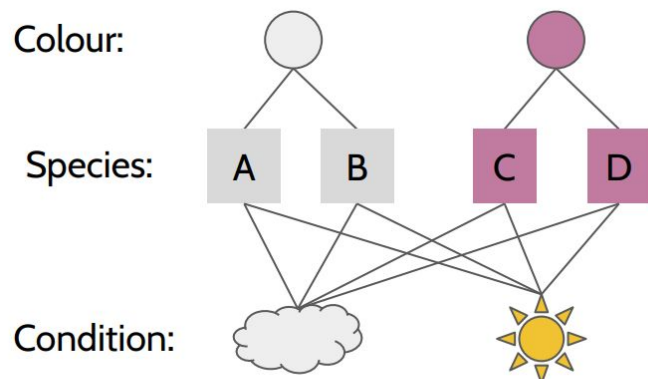
```
~ 1 + species + condition +  
species:condition
```

Contrasts (example):



Weights: 0.5 0.5 0.5 0.5
(average)

Три фактора с вложенностью



Species вложен в colour.

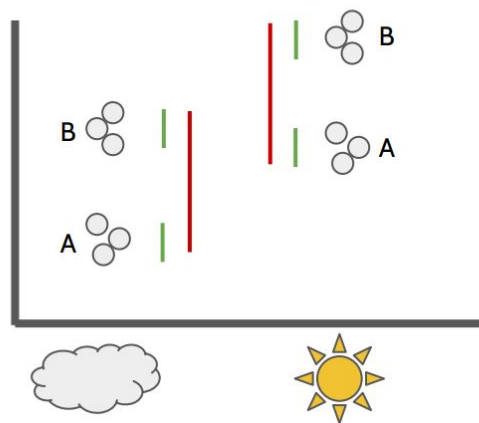
Species полностью входит в colour, поэтому в дизайн colour не включаем (но это есть смысл учесть при создании контрастов).

Design:

```
~ 1 + species + condition +  
species:condition
```

Почему не?


```
~ 1 + colour + condition + colour:condition
```



Можно переоценить или недооценить ошибку (либо тест теряет мощность, либо больше ошибок I рода (по сравнению с использованием вложенного фактора))

P-value

Способы определения достоверности
коэффициентов линейной модели



Likelihood-Ratio Test (LRT)

Рассматривает отношение
правдоподобий H_0 и H_a , логарифм
их отношения распределён как χ^2

Тест Вальда

Похож на LRT, но в явном виде
сравнивает не правдоподобия
моделей, а коэффициенты

p-value = NA?

Если в строке все значения = 0, что изменение экспрессии и дисперсию не посчитать

Если в строке есть очень большой выброс, то p-value назначается NA

Строка не прошла фильтрацию по средней экспрессии

Проблема множественного сравнения



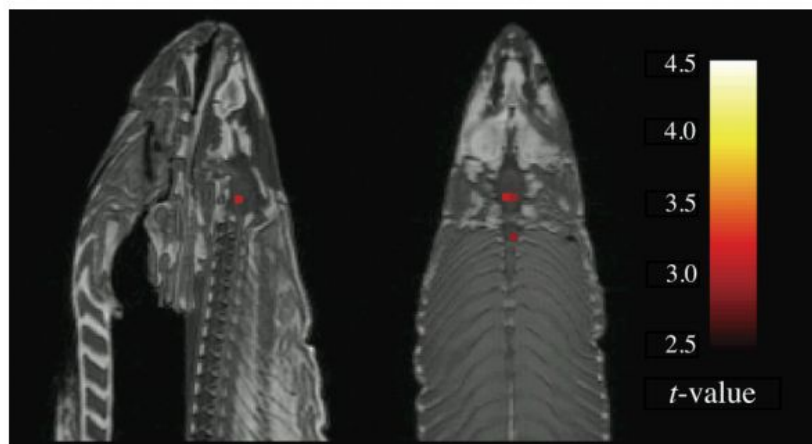
Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett¹, Abigail A. Baird², Michael B. Miller¹, and George L. Wolford³

¹ Psychology Department, University of California Santa Barbara, Santa Barbara, CA; ² Department of Psychology, Vassar College, Poughkeepsie, NY;

³ Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

GLM RESULTS



A t -contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, $p(\text{uncorrected}) < 0.001$, 3 voxel extent threshold.

Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

Identical t -contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds ($p = 0.25$).

Принципы принятия решений

Некоторые обобщения ошибки первого рода:

- **FWER** — **family-wise error rate**, групповая вероятность ошибки первого рода. Используется при поправке методом Бонферрони
- **FDR** — **false discovery rate**, средняя доля ложных отклонений гипотез (среди всех отклонений). Используется при поправке методом Бенджамини — Хохберга

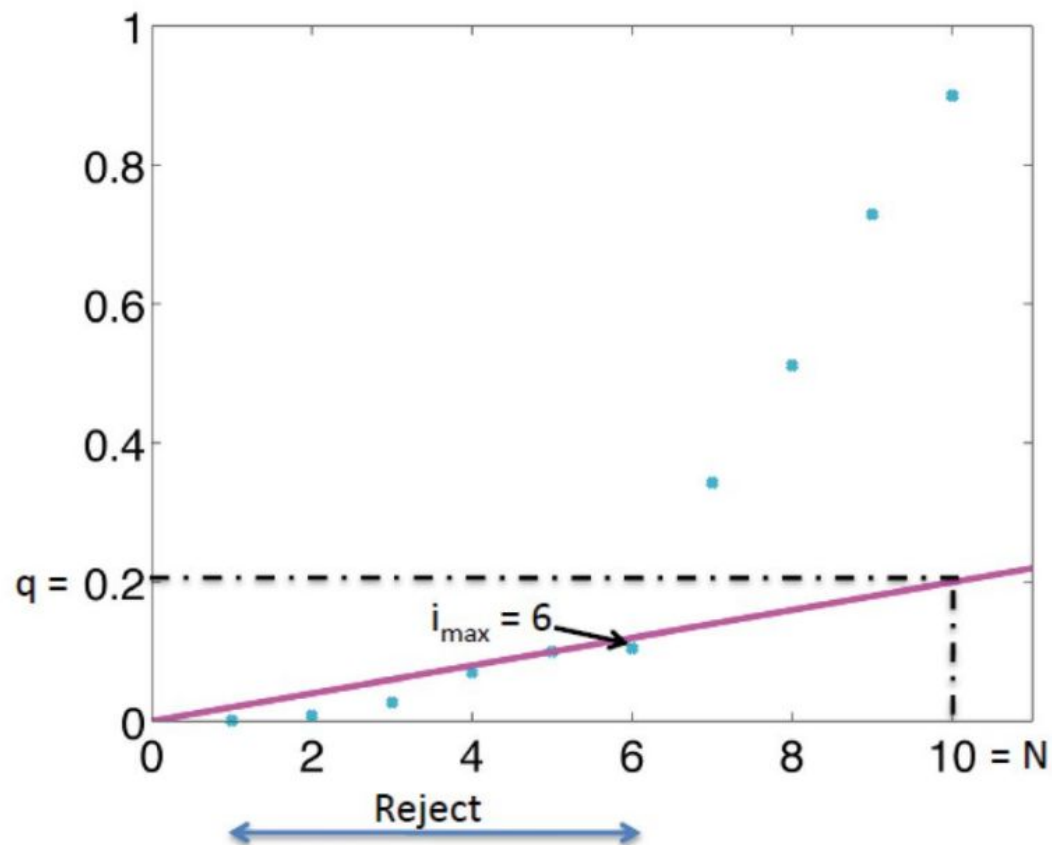
Поправка Бонферрони

The original p value 

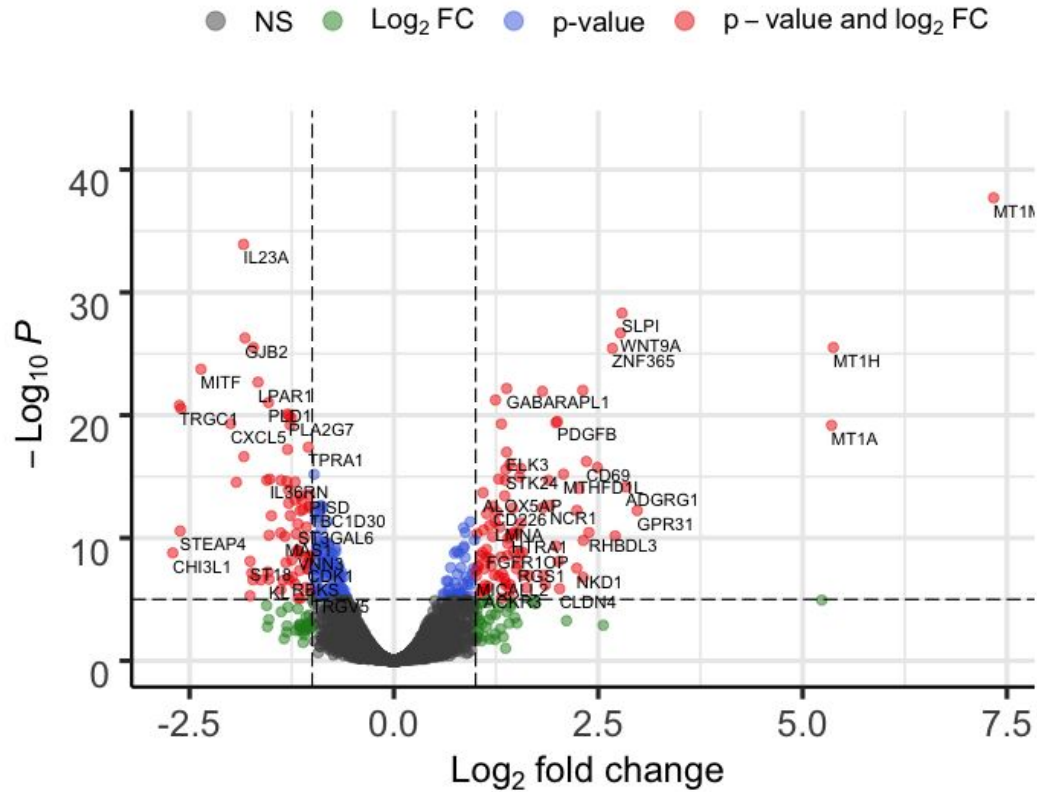
$$\text{Bonferroni-corrected } p \text{ value} = \frac{\alpha}{n}$$

The number of tests performed 

Поправка Бенджамини-Хохберга



Volcano plot



От генов к транскриптам: tximport

Как мы уже говорили ранее, самой правильной стратегией будет проводить анализ дифференциальной экспрессии на уровне транскриптов, а потом уже агрегировать информацию до уровня генов

