«Анализ транскриптомных данных»

Лекция #4.

Дифференциальная экспрессия

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;

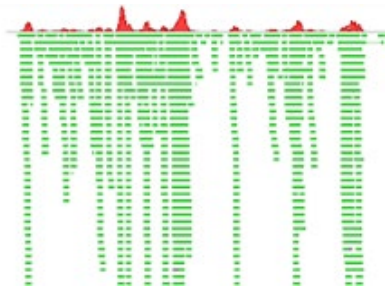
1. Single-cell RNA-Seq:
    a. экспериментальные подходы,
    b. отличия от процессинга bulk RNA-Seq,
    c. методы снижения размерности,
    d. кластера и траектории,
    e. мультимодальные омики одиночных клеток.

Дорожная карта анализа RNA -Seq

QC прочтений

FastQC Report

Выравнивание

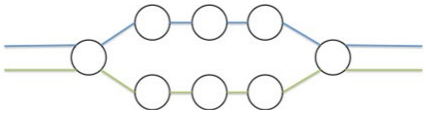


Подсчёт экспрессий

	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

Псевдовыравнивание



Дифференциальная экспрессия

GO, GSEA, ssGSEA и прочее

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

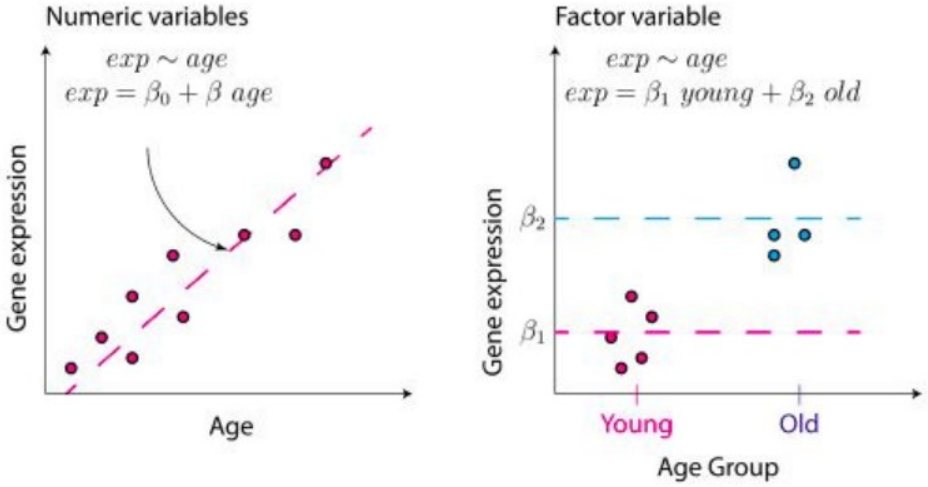
Проблема в том, что тест Манна-Уитни будет слишком слабый, так как чаще всего у нас мало точек в каждой из выборок, а t-test просто не подойдёт потому, что наши данные распределены не нормально

Что делать?

Причём тут регрессия?

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

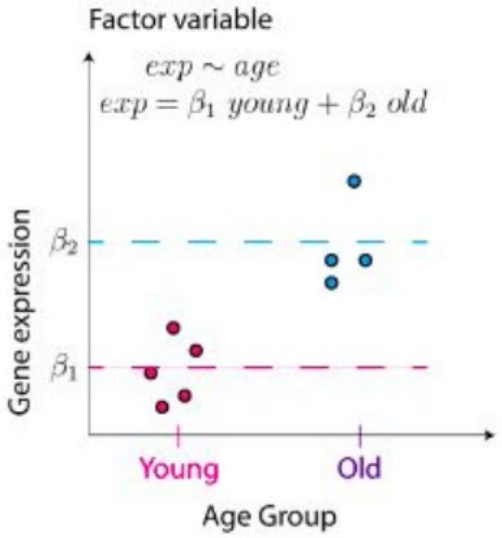
С другой стороны, GLM позволяют обобщить регрессию на ненормальные распределения



Причём тут регрессия?

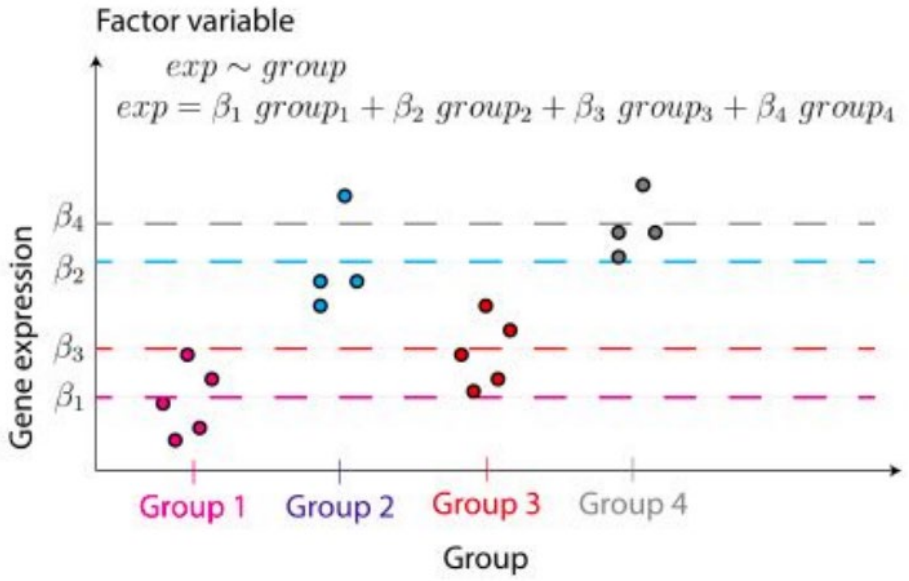
Статистический вопрос, который мы будем извлекать из регрессии, — значимо ли различаются параметры β1 и β2?

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



Причём тут регрессия?

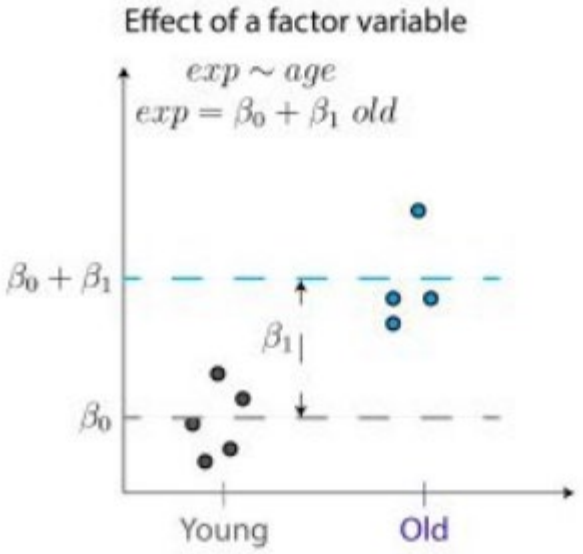
Линейную модель можно обобщить и добавить более двух уровней фактора, чтобы сравнивать сразу несколько категорий



Intercept

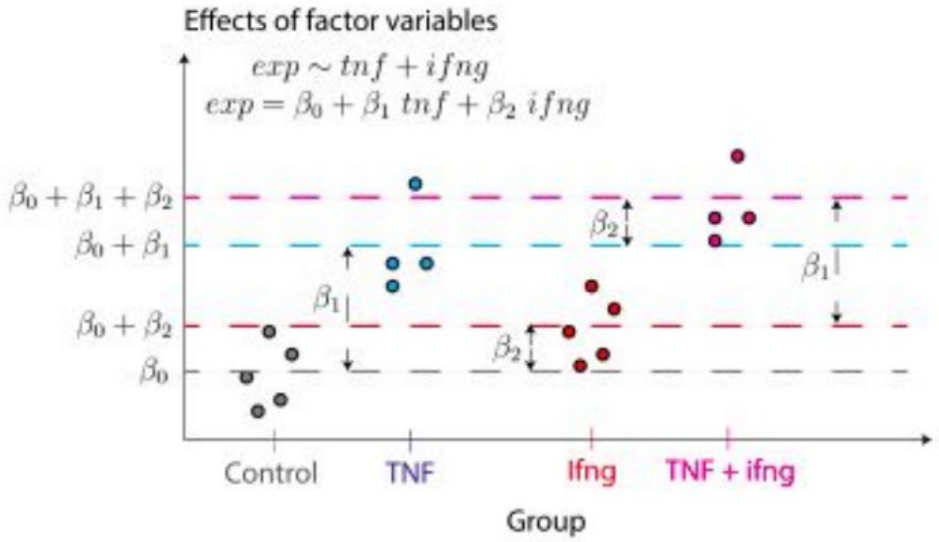
Вместо того, чтобы сравнивать значимость разницы между β1 и β2, обычно используют модель со свободным членом β0 и после этого вычисляют значимость β1

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



Intercept

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



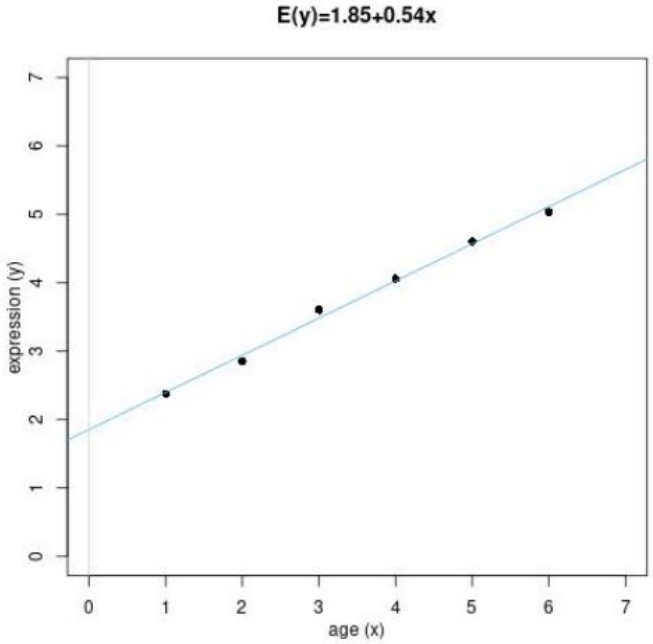
Линейные модели

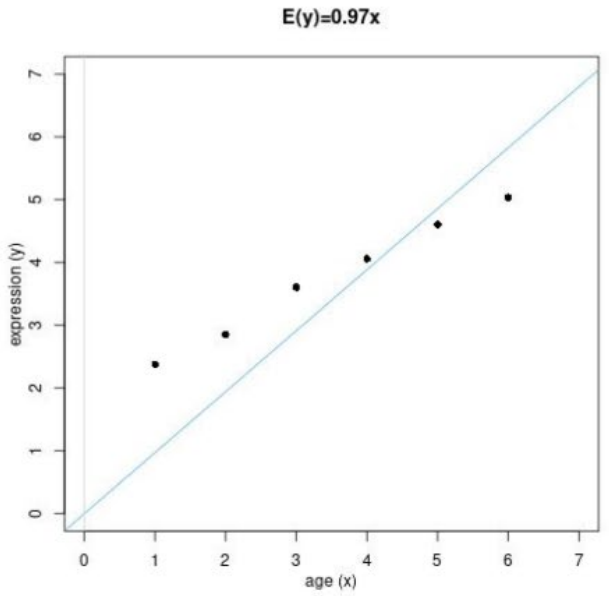
$$y \sim 0 + feature1 + feature2 + \ldots$$

без intercept

$$y \sim 1 + feature1 + feature\ 2 + \ldots$$

с intercept





Какие переменные включают в модель?

Таргет:

- экспериментальные условия,

сопутствующие факторы:

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

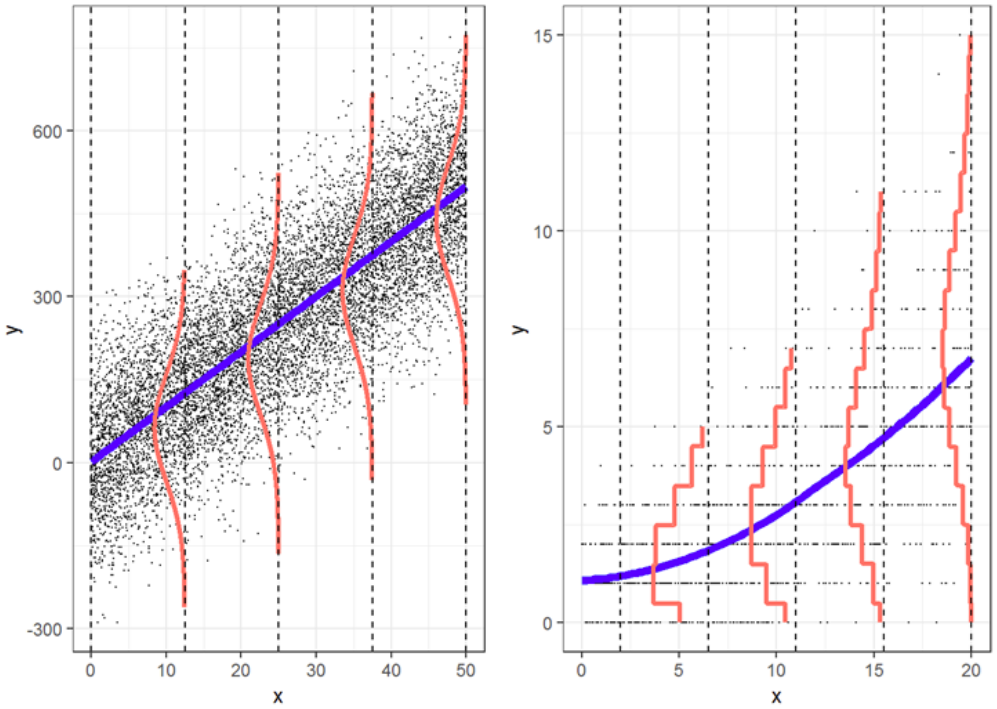
Что не включают:

- техническую повторность

Обобщённые линейные модели (GLM)

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

Коэффициенты определяются при помощи MLE



Модель DESeq2

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

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

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

$$log2(p_{i,j}) = x_{j,A}\beta_{\,i,A} + x_{j,B}\,\beta_{i,B}$$

- 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 (разбиралось на прошлом занятии),
2. потом происходит оценка дисперсии и затем
3. происходит оценка параметров β модели при помощи GLM

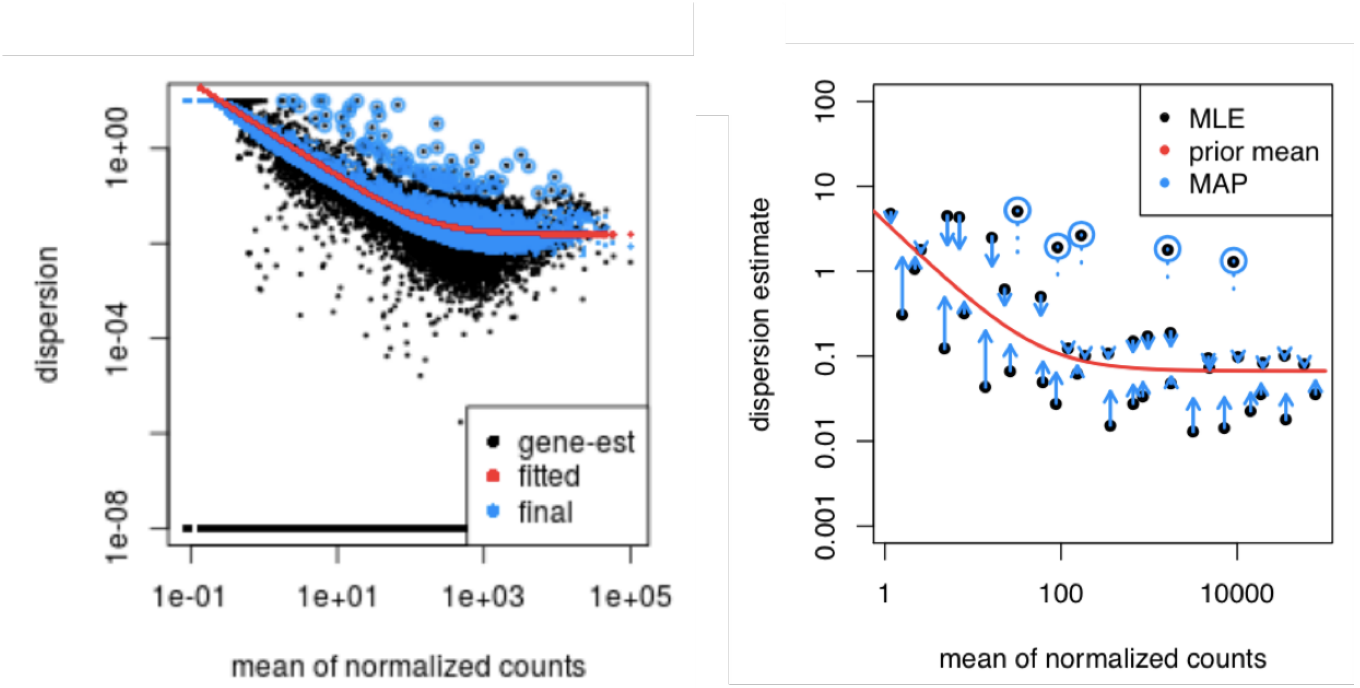
$$K_{i,j} \sim NB(\mu_{i,j}, \alpha_i)$$

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

$$log2(p_{i,j}) = x_{j,A}\beta_{\,i,A} + x_{j,B}\,\beta_{i,B}$$

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`

Expr = $\beta_0 + \beta_1$ CondSun

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: `~ 0 + condition`

Expr = β_0 + β_1 CondSun

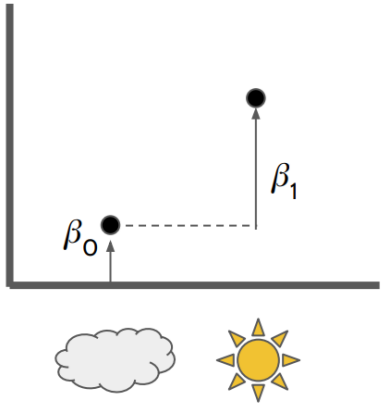
Коэффициенты из DESeq:

β_0 = `Intercept`

β_1 = `condition_sun_vs_shade`

Null hypothesis:

β_1 = 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, два уровня

Condition:

colData

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

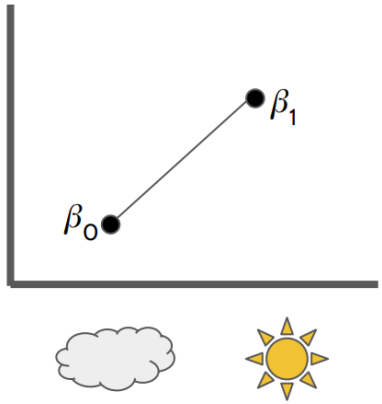
Иногда можно немного переписать модель для упрощенной интерпретации

Design: `~ 0 + condition`

Expr = β_0Shade + β_1Sun

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:

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

Design: `~ 1 + colour`

Коэффициенты из DESeq:

β_0 = Intercept
β_1 = colour_pink_vs_white
β_2 = colour_yellow_vs_white

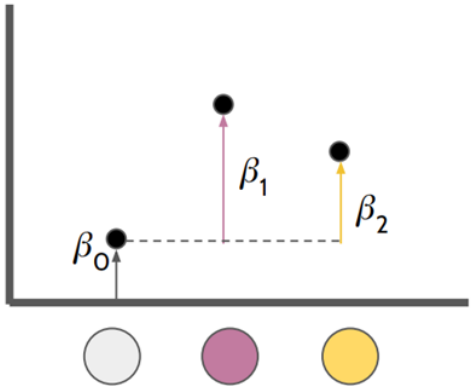
Expr = β_0 + β_1ColPink + β_2ColYellow

Нулевая гипотеза:

Pink vs White
$\beta_1 = 0$

Yellow vs White
$\beta_2 = 0$

Pink vs Yellow
$\beta_1 - \beta_2 = 0$



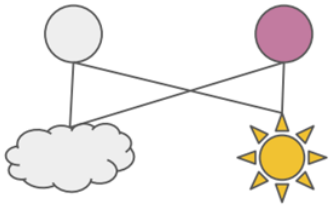
Model matrix

	(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:

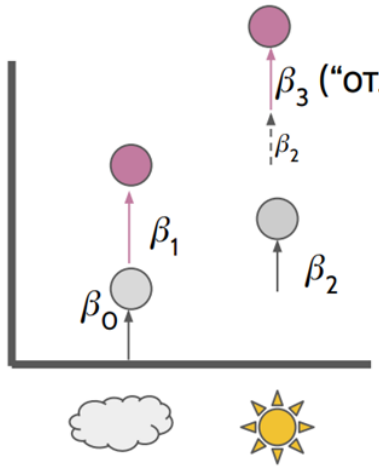
Pink vs White (Shade)
$\beta_1 = 0$

Pink vs White (Sun)
$\beta_1 + \beta_3 = 0$

Sun vs Shade (White):
$\beta_2 = 0$

Sun vs Shade (Pink):
$\beta_2 + \beta_3 = 0$

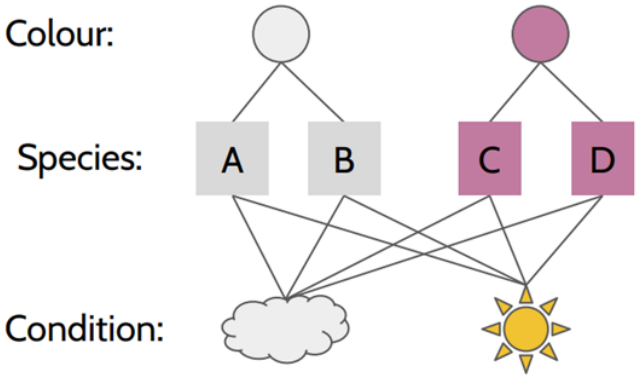
Interaction:
$\beta_3 = 0$



Коэффициенты из DESeq:

β_0 = Intercept
β_1 = colour_pink_vs_white
β_2 = condition_sun_vs_shade
β_3 = colourpink.conditionsun

Три фактора с вложенностью



Species вложен в colour.

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

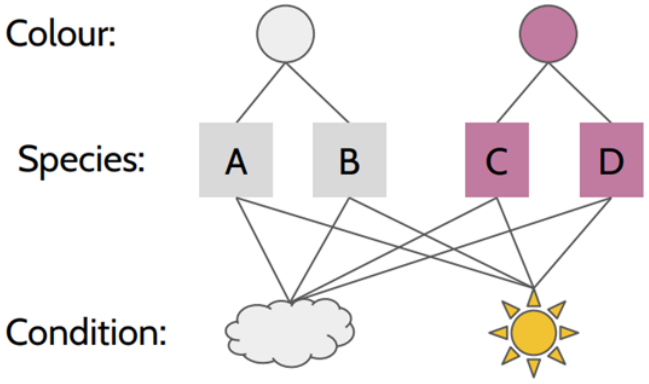
Design:

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

Contrasts (example):

Weights: (average) 0.5 0.5 0.5 0.5

Три фактора с вложенностью

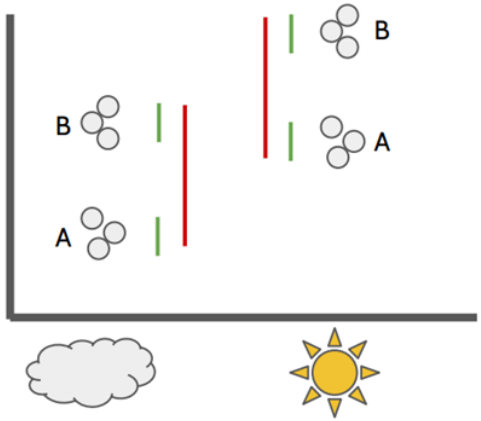


Species вложен в colour.

Species полностью входит в colour, поэтому в дизайн colour не включаем (но это есть смысл учесть про 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

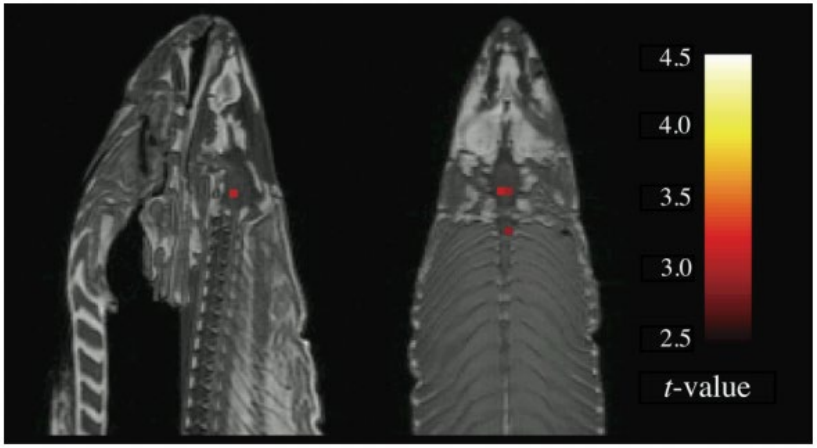
Проблема множественного сравнения

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

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[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(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.

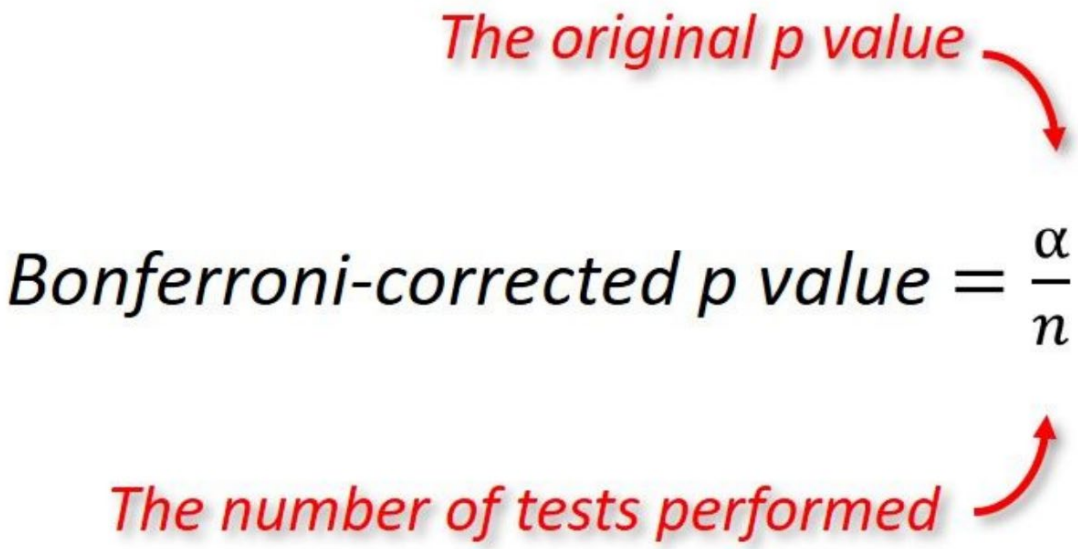
Identical *t*-contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds ($p = 0.25$).

Принципы принятия решений

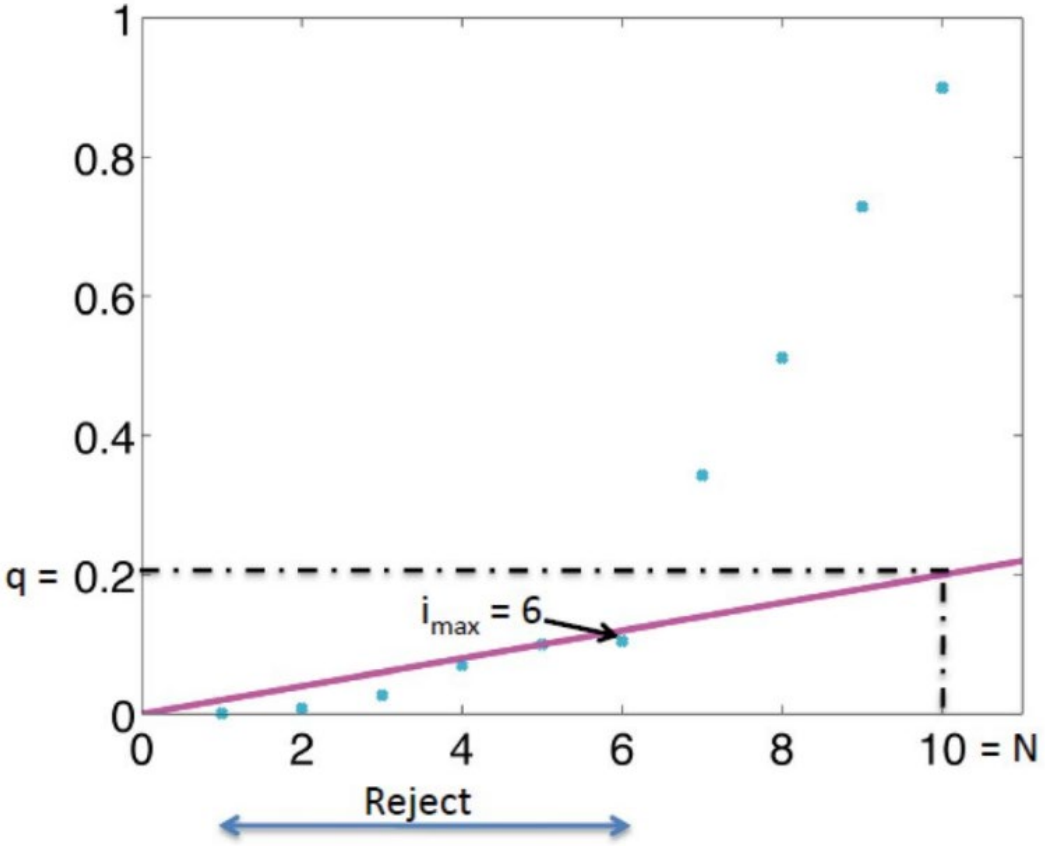
Некоторые обобщения ошибки первого рода:

- 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

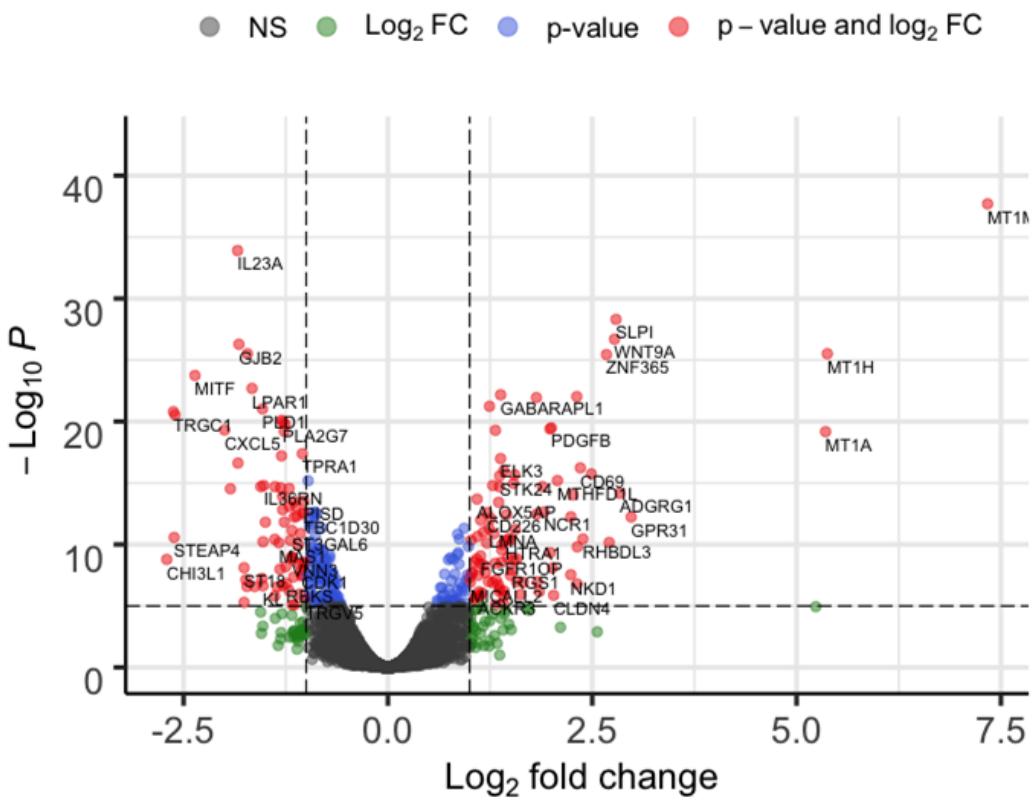
Поправка Бонферрони



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