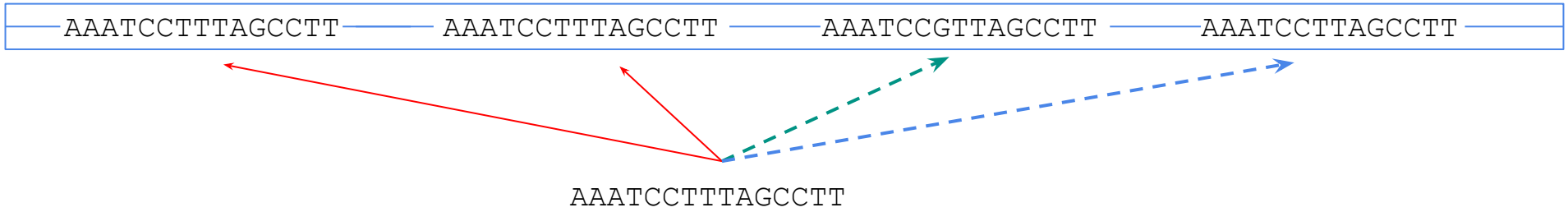
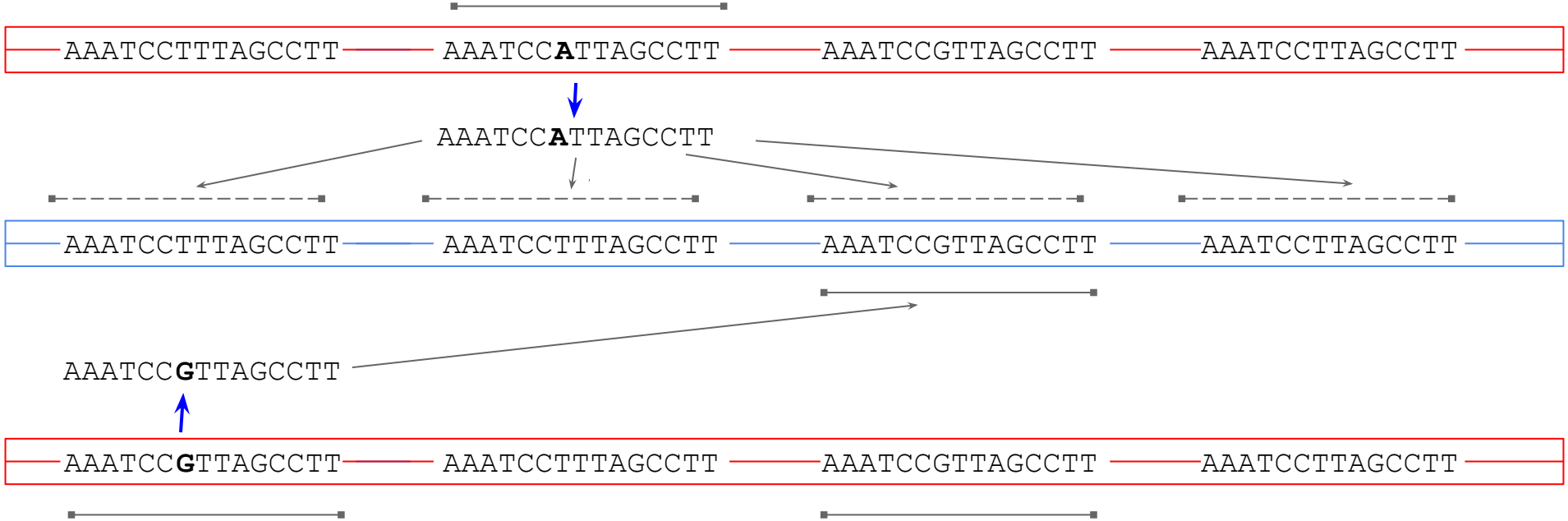
Ресеквенирование

Продолжение разговора

Множественное картирование



Притяжение к референсу

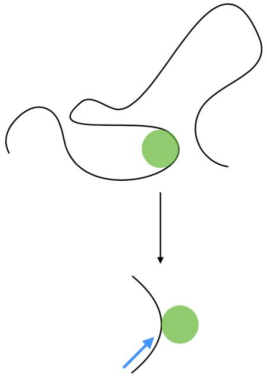


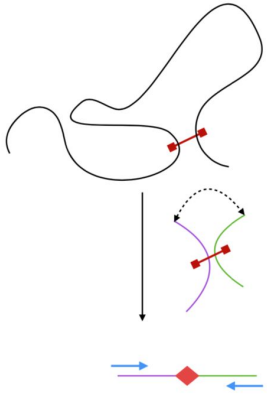
Ресеквенирование

Объект: собранный и аннотированный геном







ChIP-seq


HiC

Транскриптомика

Анализ экспрессии генов

РНК-секвенирование

- Разрушение клеток и тканей
- Очистка РНК (в том числе от ДНК!)
- Таргетное обогащение или, наоборот, обеднение образца
- Синтез кДНК
- Секвенирование

Проблемы:

- РНК и ДНК во многом химически схожи
- но РНК гораздо менее стабильная, чем ДНК
- ДНКаза работает не всегда на 100%
- большинство РНК в препарате представлено рРНК, рРНК и тРНК стабильнее, чем мРНК

Критерий оценки качества препарата **RIN** (RNA integrity number)

РНК-секвенирование

Смелое предположение:

количество ридов пропорционально количеству мРНК гена

Один пример (2 образца):

Размер библиотеки		Библиотека 1	Библиотека 2
		20M	10M
Ген 1	Read count	1000	500
	mRNA	10	10

Второй пример (тот же образец):

	Ген 1	Ген 2
Длина	1000	5000
mRNA	10	10
Read count	100	500

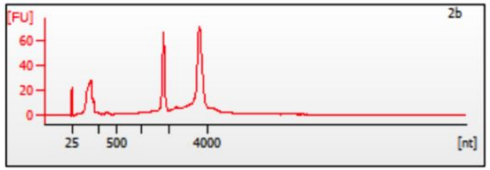
Данные:
- повторности технические
- повторности биологические

Обычно данные состоят из двух и более образцов (samples), отражающих разные условия

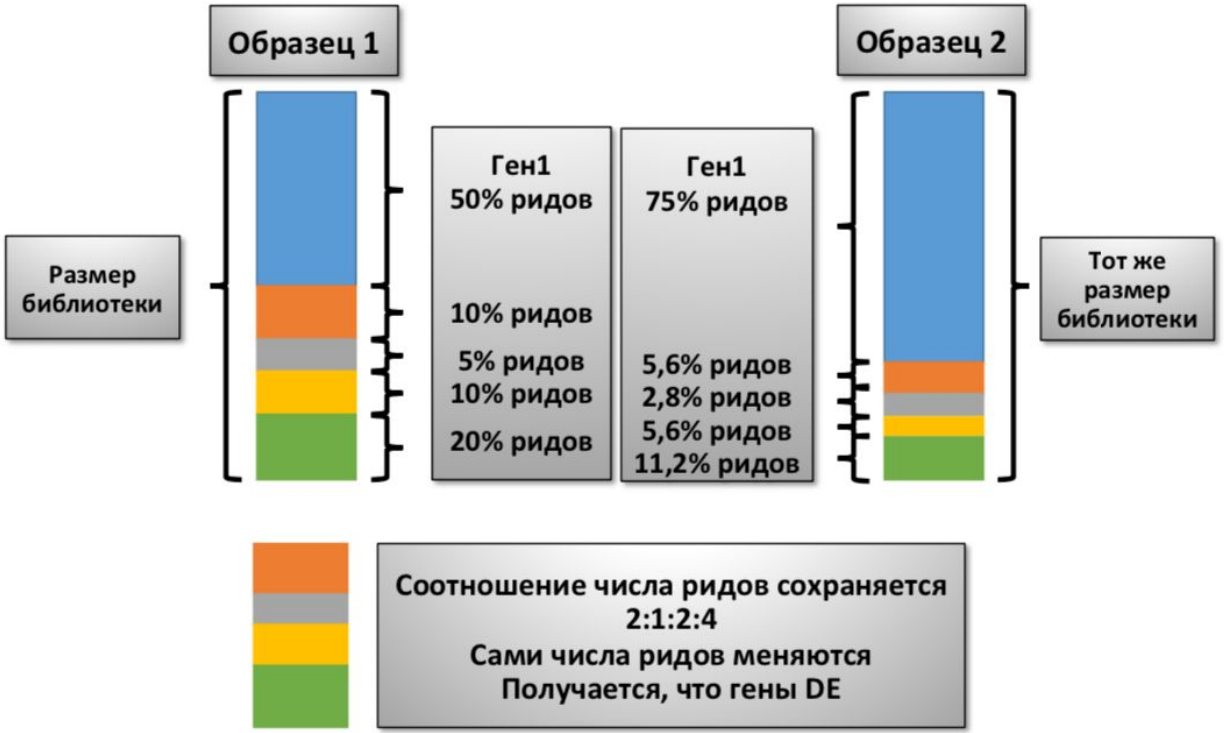
Пробоподготовка

Обычно реализуется таргетный подход:
- полиА-обогащение;
- деплеция рРНК;
- IP

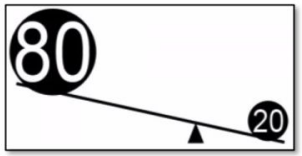
* повторности обеспечивают основной принцип всех научных исследований - воспроизводимость



Уровень экспрессии гена

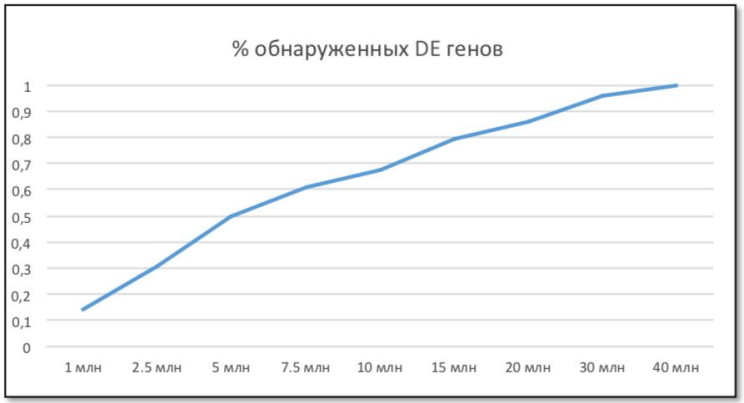


Особенности покрытия



20% ридов приходится на 80% генов

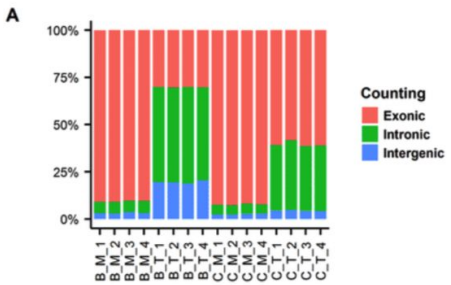
Длина рида?
Глубина покрытия?

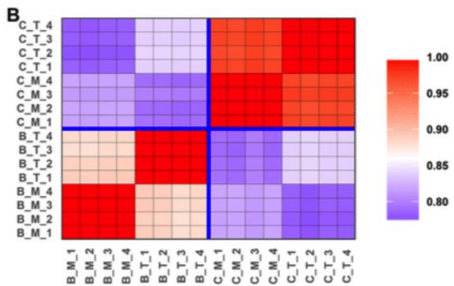


Обычно всем программам для анализа ДЭ на вход идут числа ридов, пришедшиеся на единицу аннотации

Что делать дальше?

gene count / exon count ← здесь добавятся ошибки аннотации





Нормализация

Наиболее популярные способы

- RPKM (Reads Per Kilobase Million)
- FPKM (Fragments Per Kilobase Million)
- TPM (Transcripts Per Kilobase Million)

RPKM:

1) число всех ридов в образце делим на 1 млн (per-million)
2) ридкаунты гена делим на per-M, получаем RPM (нормализация на глу
3) делим RPM на длину гена в т.п.н. (нормализация на длину)

TPM:

1) ридкаунты гена делятся на длину гена в т.п.н. (RPK)
2) суммируем все RPK в образце и делим это число на 1 млн (per-million)
3) делим PRK каждого гена на per-million фактор

When you use TPM, the sum of all TPMs in each sample are the same. This makes it easier to compare the proportion of reads that mapped to a gene in each sample. In contrast, with RPKM and FPKM, the sum of the normalized reads in each sample may be different, and this makes it harder to compare samples directly.

Here's an example. If the TPM for gene A in Sample 1 is 3.33 and the TPM in sample B is 3.33, then I know that the exact same proportion of total reads mapped to gene A in both samples. This is because the sum of the TPMs in both samples always add up to the same number (so the denominator required to calculate the proportions is the same, regardless of what sample you are looking at.)

(from RNAseq blog)

Некоторые определения

Вариация - различие значений какого-либо признака в некоторой выборке

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

выборочное среднее (математическое ожидание)

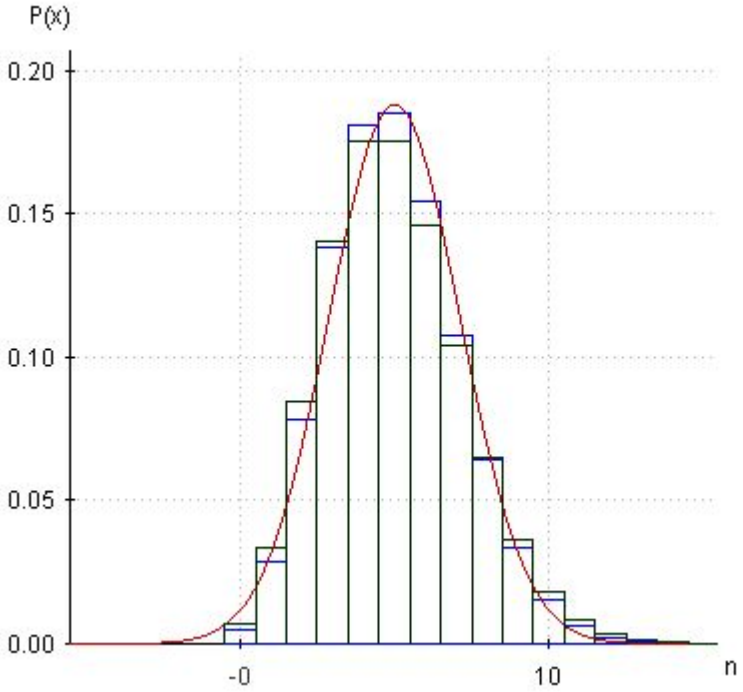
$$\widetilde{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

несмещенная оценка дисперсии (математического ожидания квадрата отклонения случайной величины от математического ожидания). Англ. *variance* (!!!)

$\sqrt{S^2}$ - среднеквадратичное отклонение (стандартное отклонение). Измеряется в тех же единицах, что сама X

prior (distribution) - априорное распределение вероятностей, выражает предположение об X до учета экспериментальных данных

Поиск **prior distribution** (“фитирование”)



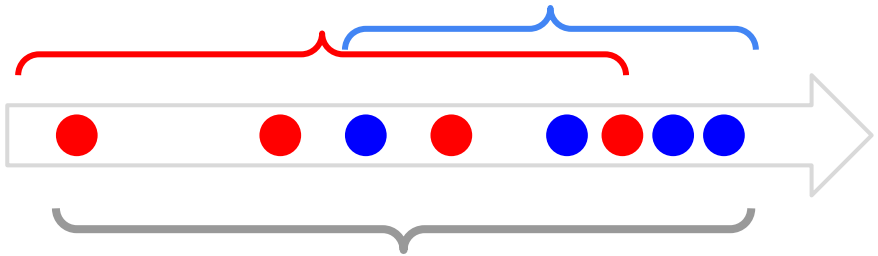
DESeq2: мотивация

We would like to use statistical testing to decide whether, for a given gene, an observed difference in read counts is significant, that is, whether it is greater than what would be expected just due to natural random variation.

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma_{ij}^2),$$

i - ген
j - образец

μ и σ на практике неизвестны,
их нужно оценивать из данных



Слайды о DESeq2 основаны на статьях:
10.1186/s13059-014-0550-8 (M. Love)
10.1186/gb-2010-11-10-r106 (S. Anders)

DESeq2: рид-каунты как случайные величины

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma^2_{ij}),$$

$$\mu_{ij} = q_{i,\rho(j)} s_j.$$

$$\sigma^2_{ij} = \underbrace{\mu_{ij}}_{\text{shot noise}} + \underbrace{s^2_j v_{i,\rho(j)}}_{\text{raw variance}}. \quad \longrightarrow \quad v_{i,\rho(j)} = v_\rho(q_{i,\rho(j)}).$$

это необходимо допустить, потому что малое число повторностей обычно не позволяет надежно оценить дисперсию гена i

DESeq2: нормализация на размер

1. Сделаем референс, где ридкаунты равны среднему геометрическому

gene	sampleA	sampleB	pseudo-reference sample
EF2A	1489	906	sqrt(1489 * 906) = 1161.5
ABCD	22	13	sqrt(24 * 13) = 17.7
...	...	...	...

2. Посчитаем отношение каждого образца к референсу

gene	sampleA	sampleB	pseudo-reference sample	ratio sampleA/ref	ratio sampleB/ref
EF2A	1489	906	1161.5	1489/1161.5 = 1.28	906/1161.5 = 0.78
ABCD	22	13	16.9	22/16.9 = 1.30	13/16.9 = 0.77
MEF3	793	410	570.2	793/570.2 = 1.39	410/570.2 = 0.72
BBC1	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
...	...	...	...		

gene	sampleA	sampleB
EF2A	1489 / 1.3 = 1145.39	906 / 0.77 = 1176.62
ABCD	22 / 1.3 = 16.92	13 / 0.77 = 16.88
...	...	...

3. Выберем медиану этого значения для каждого образца

DESeq2: рид-каунты как случайные величины

$$\hat{q}_{i\rho} = \frac{1}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{k_{ij}}{\hat{s}_j},$$

$$z_{i\rho} = \frac{\hat{q}_{i\rho}}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{1}{\hat{s}_j}.$$

$$w_{i\rho} = \frac{1}{m_\rho - 1} \sum_{j:\rho(j)=\rho} \left(\frac{k_{ij}}{\hat{s}_j} - \hat{q}_{i\rho} \right)^2$$

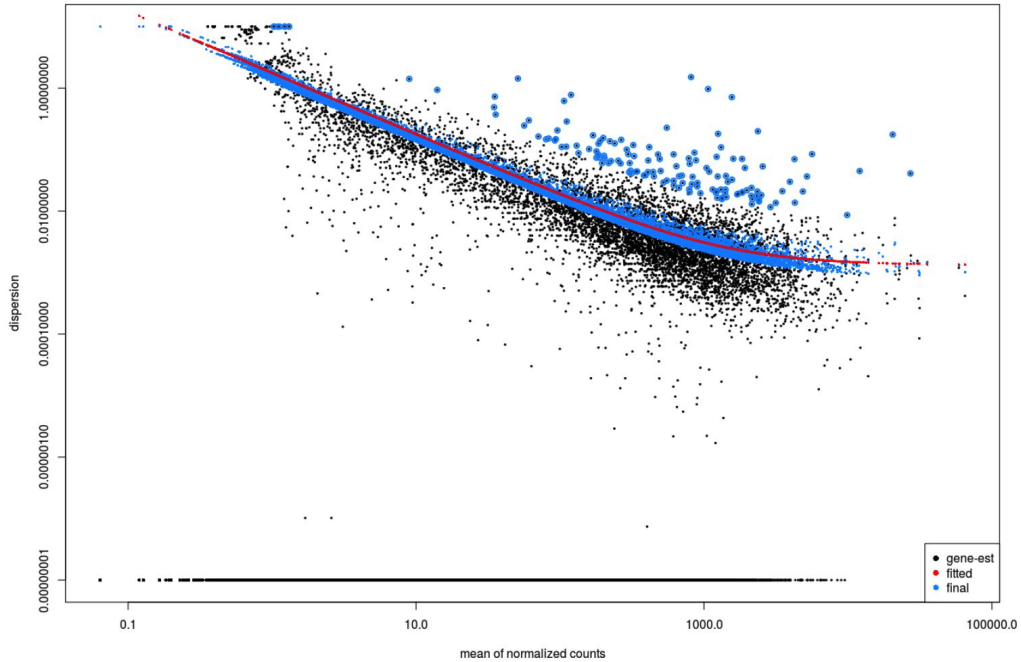
$$\hat{v}_\rho(\hat{q}_{i\rho}) = w_\rho(\hat{q}_{i\rho}) - z_{i\rho}$$

однако, при небольшом числе j,
w_{ip} будет плохой оценкой для дисперсии,
поэтому дисперсия оценивается путем
локальной регрессии на графике (q_{ip}, w_{ip})

DESeq2: mean-dispersion model

Ген 1: 1000, 990, 1000, 1010
(среднее: 1000, дисперсия: 50, dm: 0.7%)

Ген 2: 0, 10, 10, 20
(среднее: 10, дисперсия: 50, dm: 70%)



D - мера вариабельности между образцами (BCV)

$$\mathbf{D} = (Var - \mu) / \mu^2$$

пример взят из http://varianceexplained.org/r/empirical_bayes_baseball/

Empirical Bayes Shrinkage (EBS)

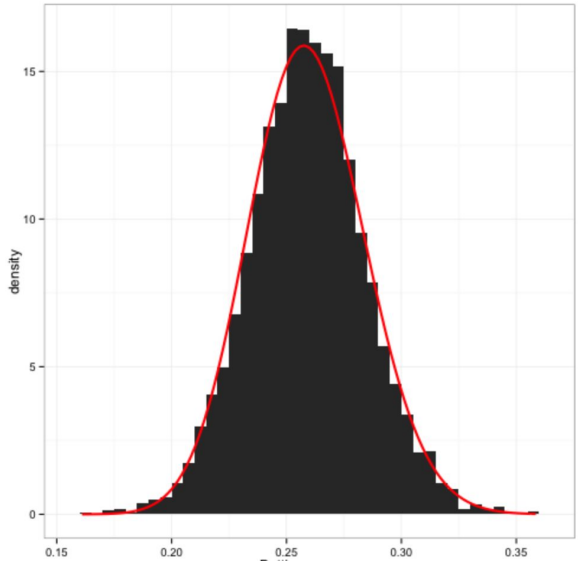
4 / 10 или 300 / 1000 ?

Среднее α (число побед) и β (число поражений) из распределения?

Тогда для среднего игрока можно ожидать:
результат сезона = α / (α + β)

Представим, что сезон 1 игрока был бы длиннее на 14 матчей, тогда:
результат сезона 1 игрока = (4 + α) / (10 + α + β),

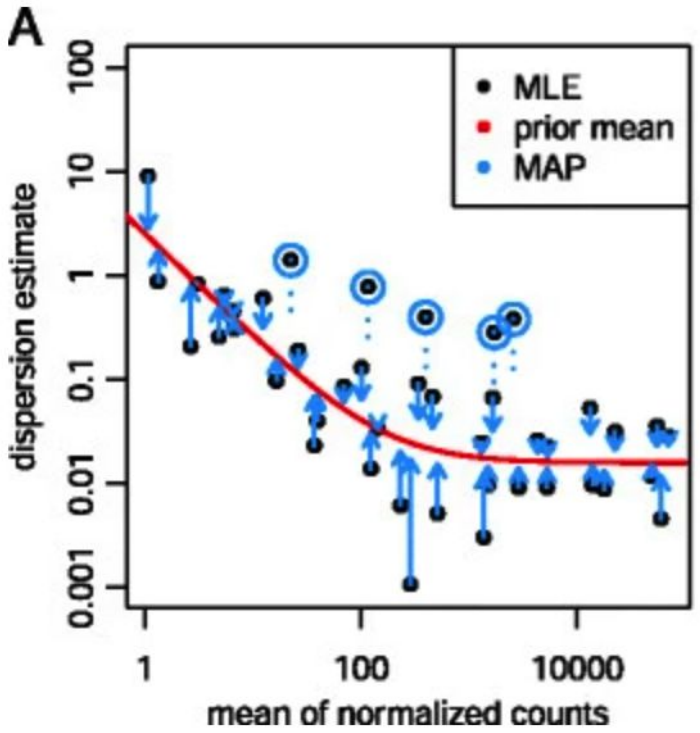
пусть α = 300, β = 1000. Тогда для игроков имеем:
1 игрок = 304 / 1310 = 0.23
2 игрок = 600 / 2300 = 0.26



DESeq2: mean-dispersion model

Ген 1: 1000, 990, 1000, 1010
(среднее: 1000, дисперсия: 50, dm: 0.7%)

Ген 2: 0, 10, 10, 20
(среднее: 10, дисперсия: 50, dm: 70%)



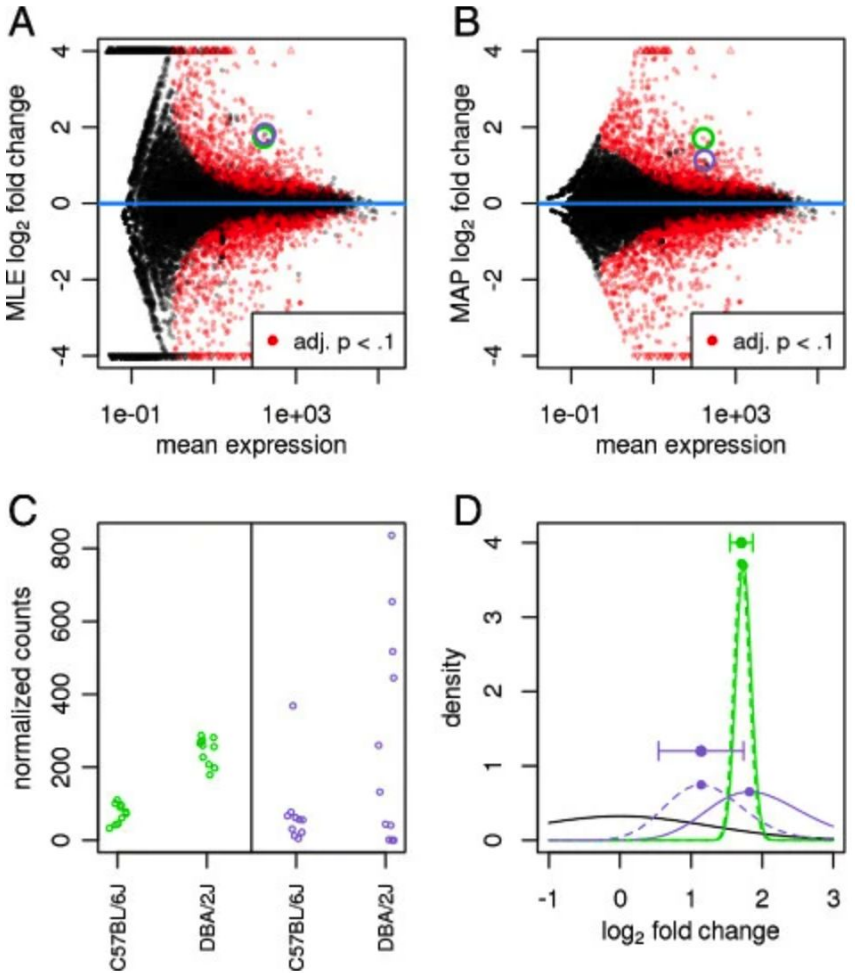
D - мера вариабельности между образцами (BCV)

$\mathbf{D} = (Var - \mu) / \mu^2$

DESeq2: EBS of log (Fold Change)

“The strength of shrinkage does not depend simply on the mean count, but rather on the amount of information available for the fold change estimation” (D)

Figure 2



Проверка статистической гипотезы **(Wald test)**

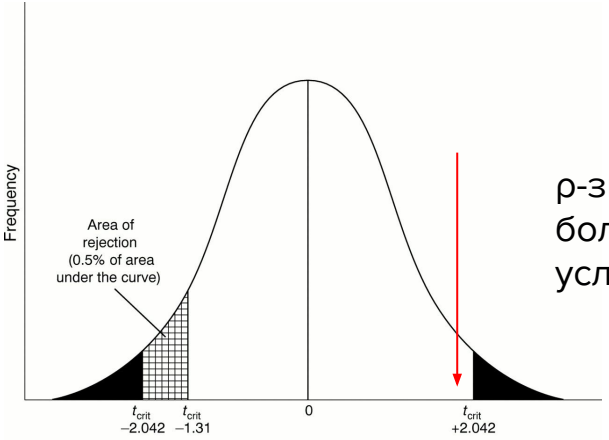
Дано: MAP оценки для LFC и стандартных отклонений

$$z_i = \frac{x_i - \bar{x}}{s}$$

(Z-статистика)

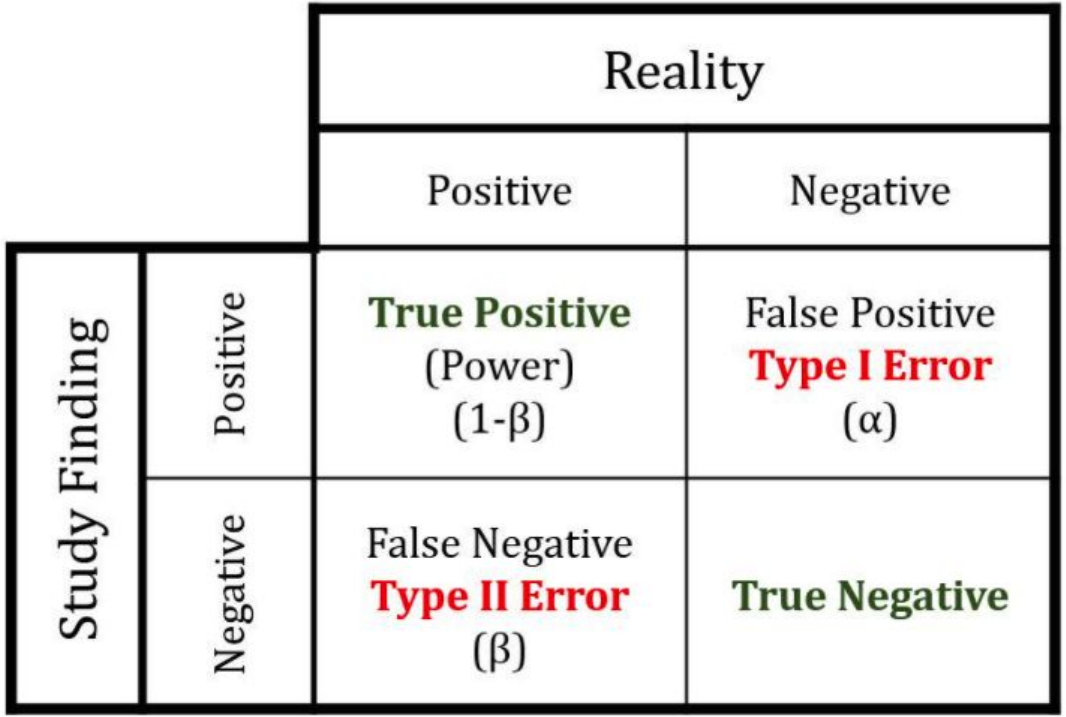
$$W = \frac{(\hat{\theta} - \theta_0)^2}{\mathrm{var}(\hat{\theta})}$$

Статистика теста Вальда для Θ - параметра, оцененного с MLE



ρ-значение - вероятность получить такое же или более экстремальное значение статистики, при условии, что верна H_0

Ошибки первого и второго рода



		Reality	
		Positive	Negative
Study Finding	Positive	**True Positive** (Power) (1-β)	False Positive **Type I Error** (α)
	Negative	False Negative **Type II Error** (β)	**True Negative**

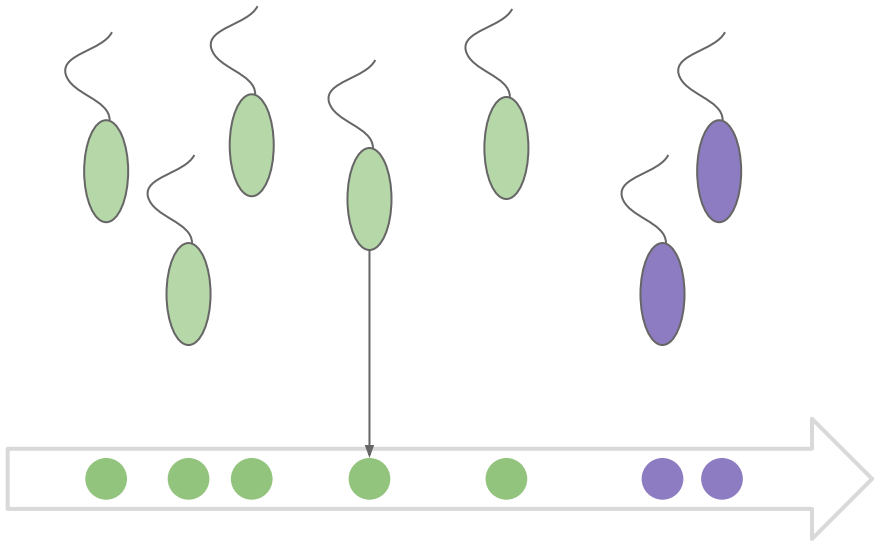
ошибка 1 рода (ложноположительный): найти эффект там, где его нет.

уровень значимости (α) - вероятность ошибки 1 рода,
(1-α) - специфичность метода

ошибка 2 рода (ложноотрицательный): не найти эффект там, где он есть.
(1-β) - чувствительность метода (или мощность критерия)

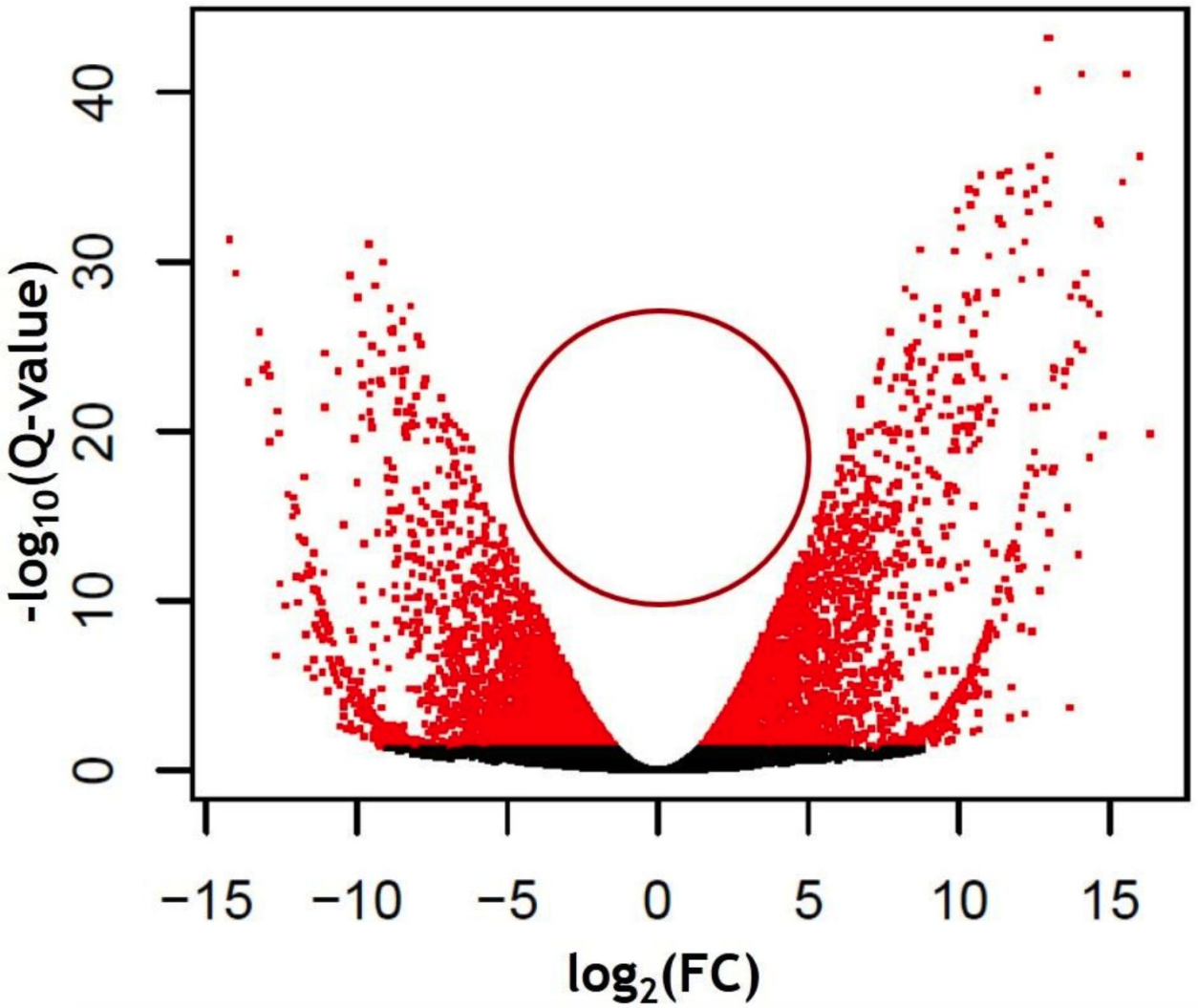
Некоторые усложнения

или почему методов оценки ДЭ генов больше одного



$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma^2_{ij}), \quad ?$$

DESeq2:
результаты
на графике
**Vulcano
plot**



Отношения между **scatter, vulcano** и **MA**