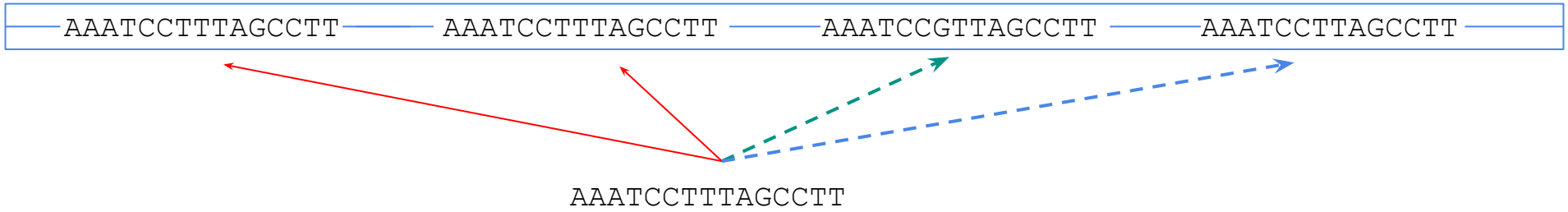
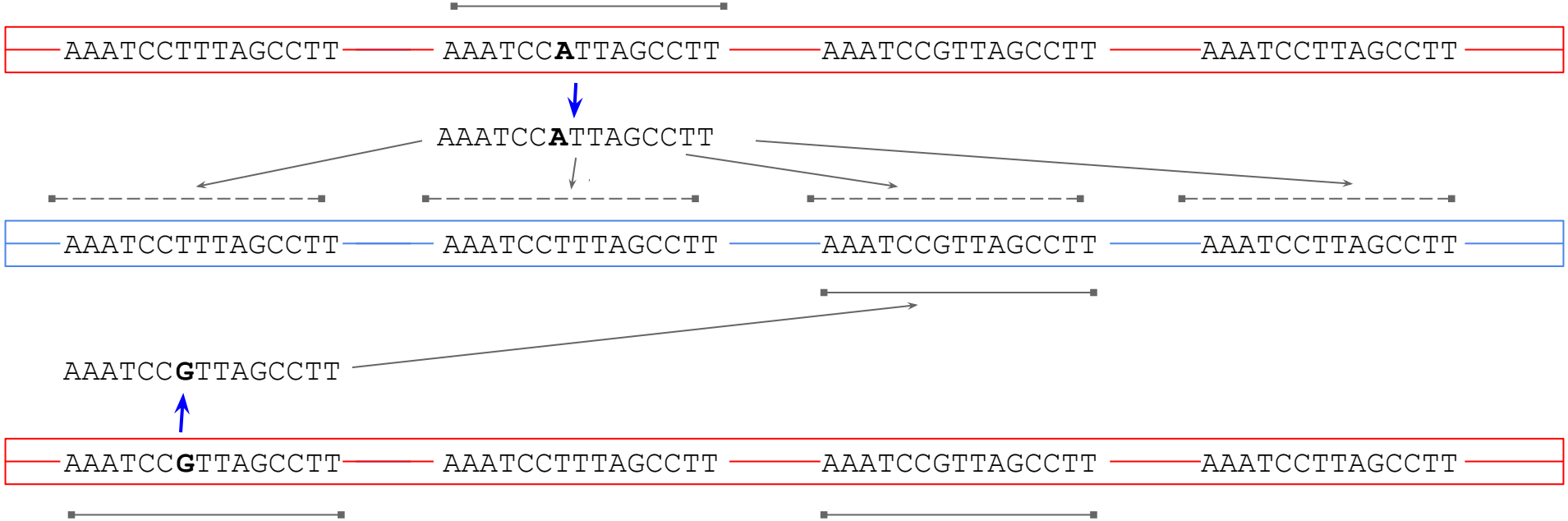
Ресеквенирование

Продолжение разговора

Множественное картирование



Притяжение к референсу

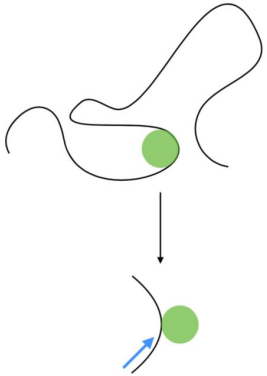


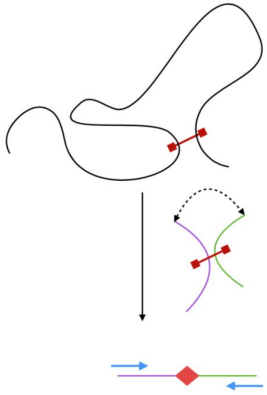
Ресеквенирование

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

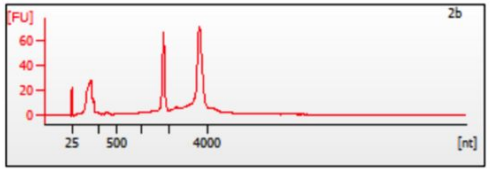
Данные:
- повторности технические
- повторности биологические

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

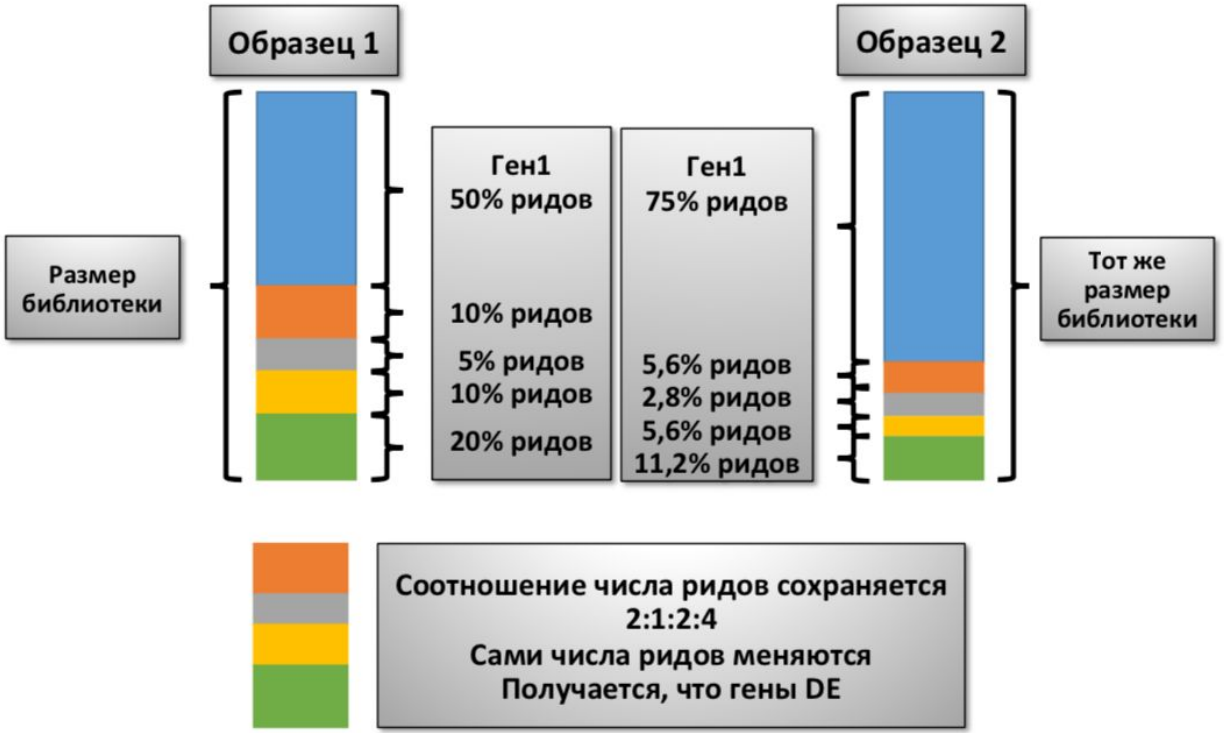
Пробоподготовка

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

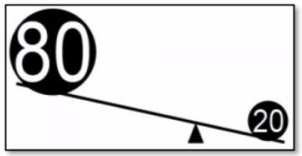
* повторности обеспечивают основной принцип всех научных исследований - воспроизводимость



Уровень экспрессии гена

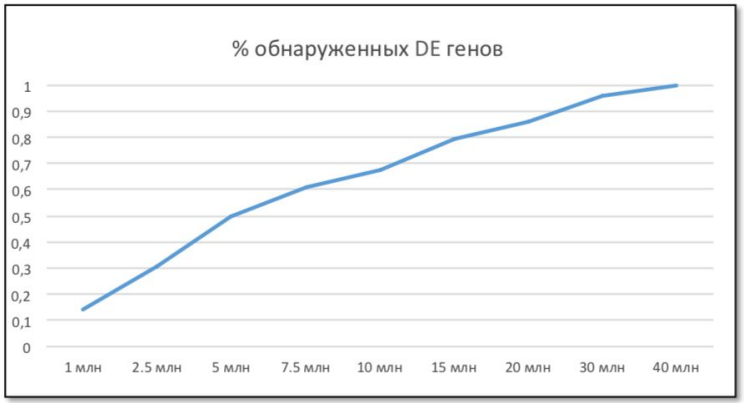


Особенности покрытия



20% ридов приходится на 80% генов

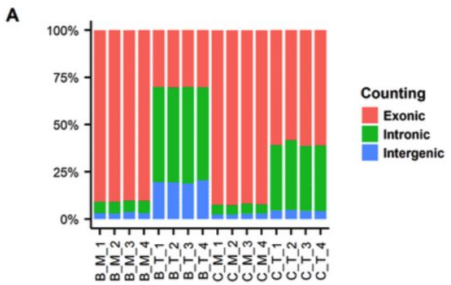
Длина рида?
Глубина покрытия?

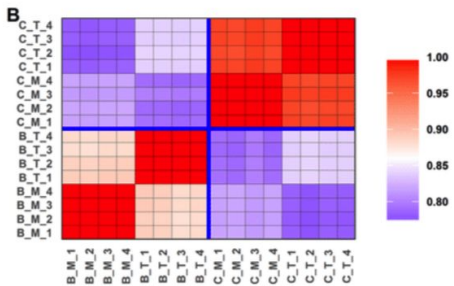


Обычно всем программам для анализа ДЭ на вход идут числа ридов, пришедшиеся на единицу аннотации

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)

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
...	...	...	...		

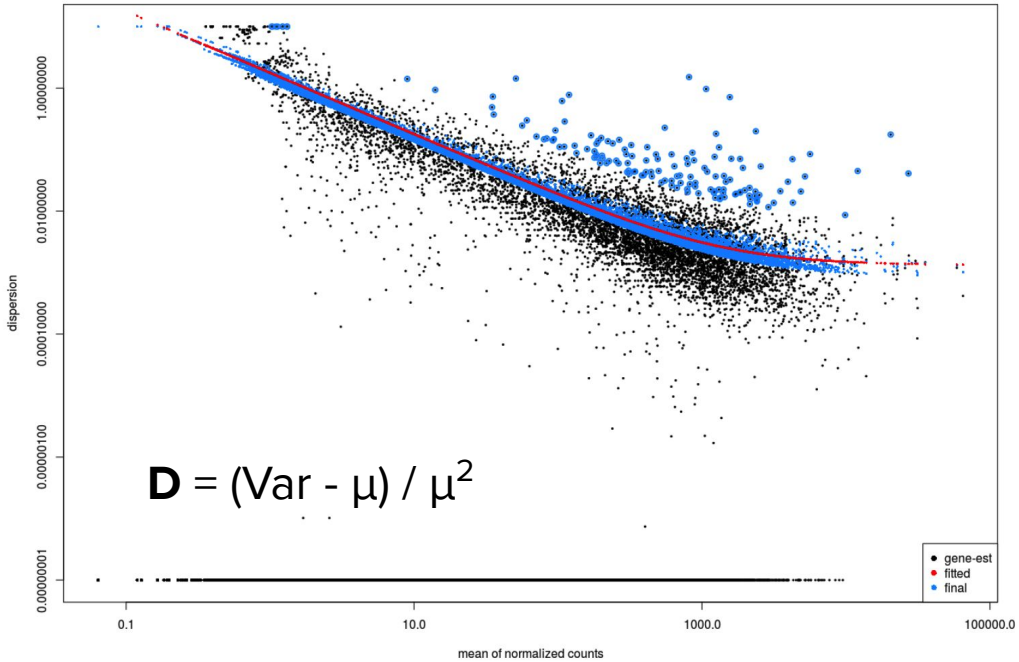
gene	sampleA	sampleB
EF2A	1489 / 1.3 = 1145.39	906 / 0.77 = 1176.62
ABCD	22 / 1.3 = 16.92	13 / 0.77 = 16.88
...	...	...

3. Выберем медиану этого значения для каждого образца

DESeq2: mean-dispersion model

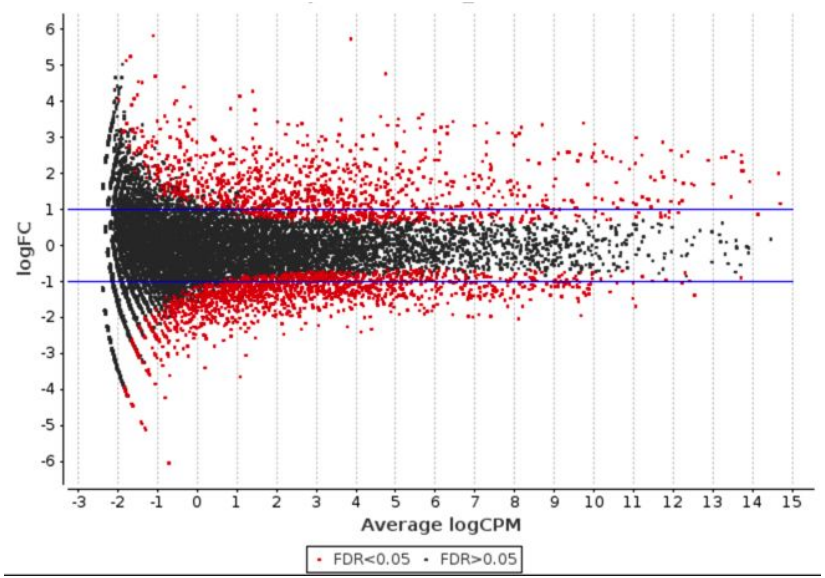
Ген 1: 1000, 990, 1000, 1010
(среднее: 1000, дисперсия: 50, dm: 0.7%)

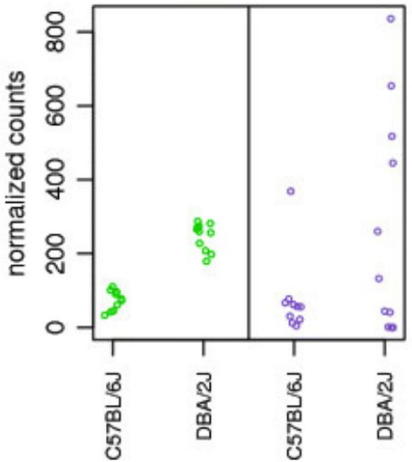
Ген 2: 0, 10, 10, 20
(среднее: 10, дисперсия: 50, dm: 70%)

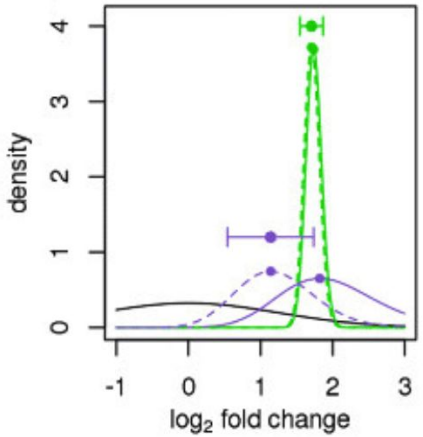


D - мера вариабельности между образцами (BCV)

DESeq2: EBS of LFC, MA-plot

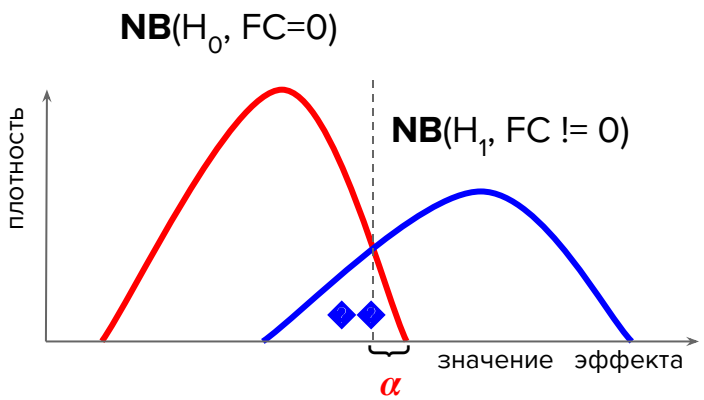






DESeq2: статистический тест

		Верная гипотеза	
		H_0	H_1
Результат применения критерия	H_0	H_0 верно принята	H_0 неверно принята (Ошибка *второго* рода)
	H_1	H_0 неверно отвергнута (Ошибка *первого* рода)	H_0 верно отвергнута



H_0 - нет разницы между экспериментами, logFC = 0.

α ошибка первого рода

1-α специфичность метода

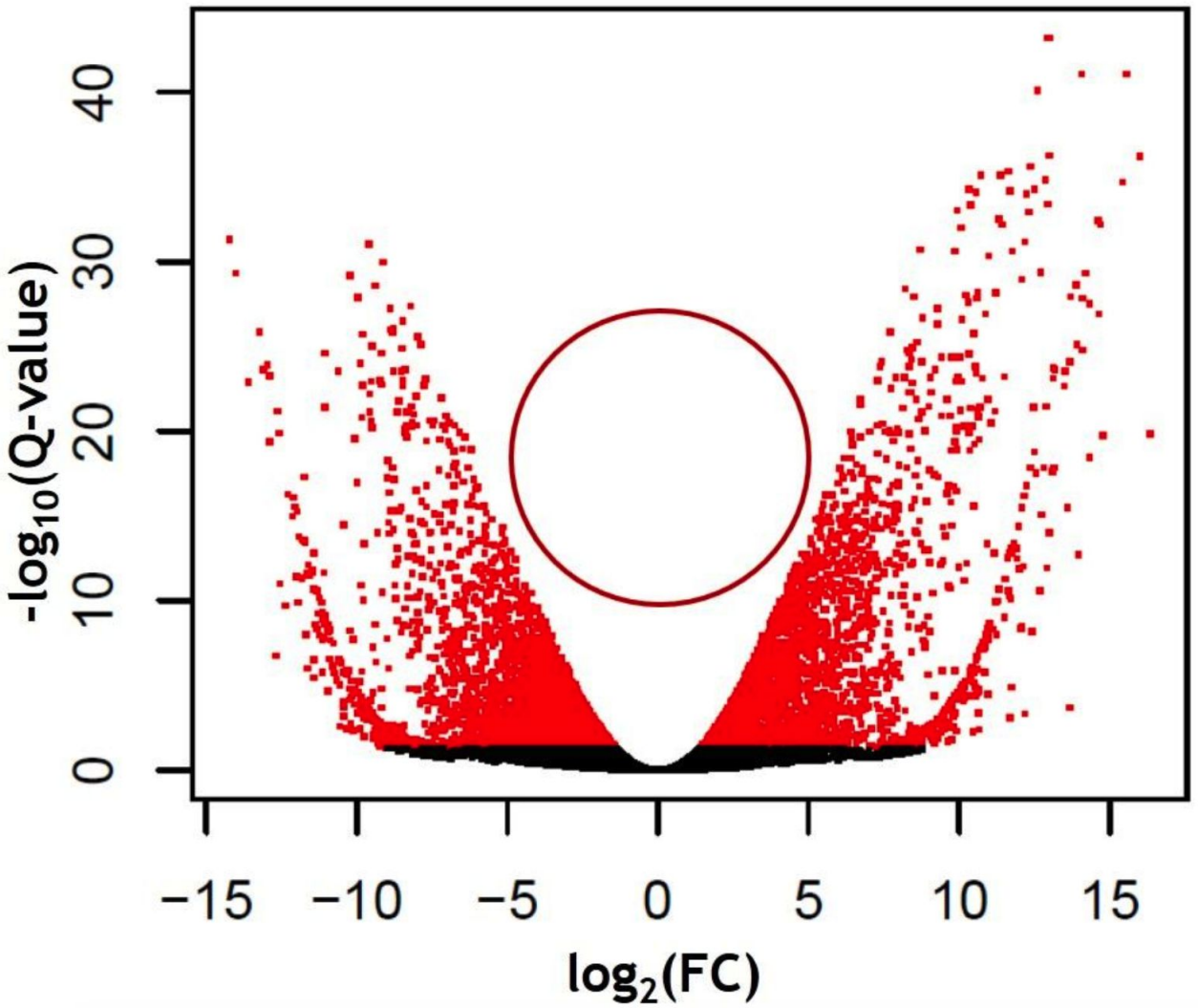
(H_0 при верной H_0)

β ошибка второго рода

1-β чувствительность метода

(H_1 при верной H_1: найти эффект там, где он был)

DESeq2:
результаты
на графике
**Vulcano
plot**



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