

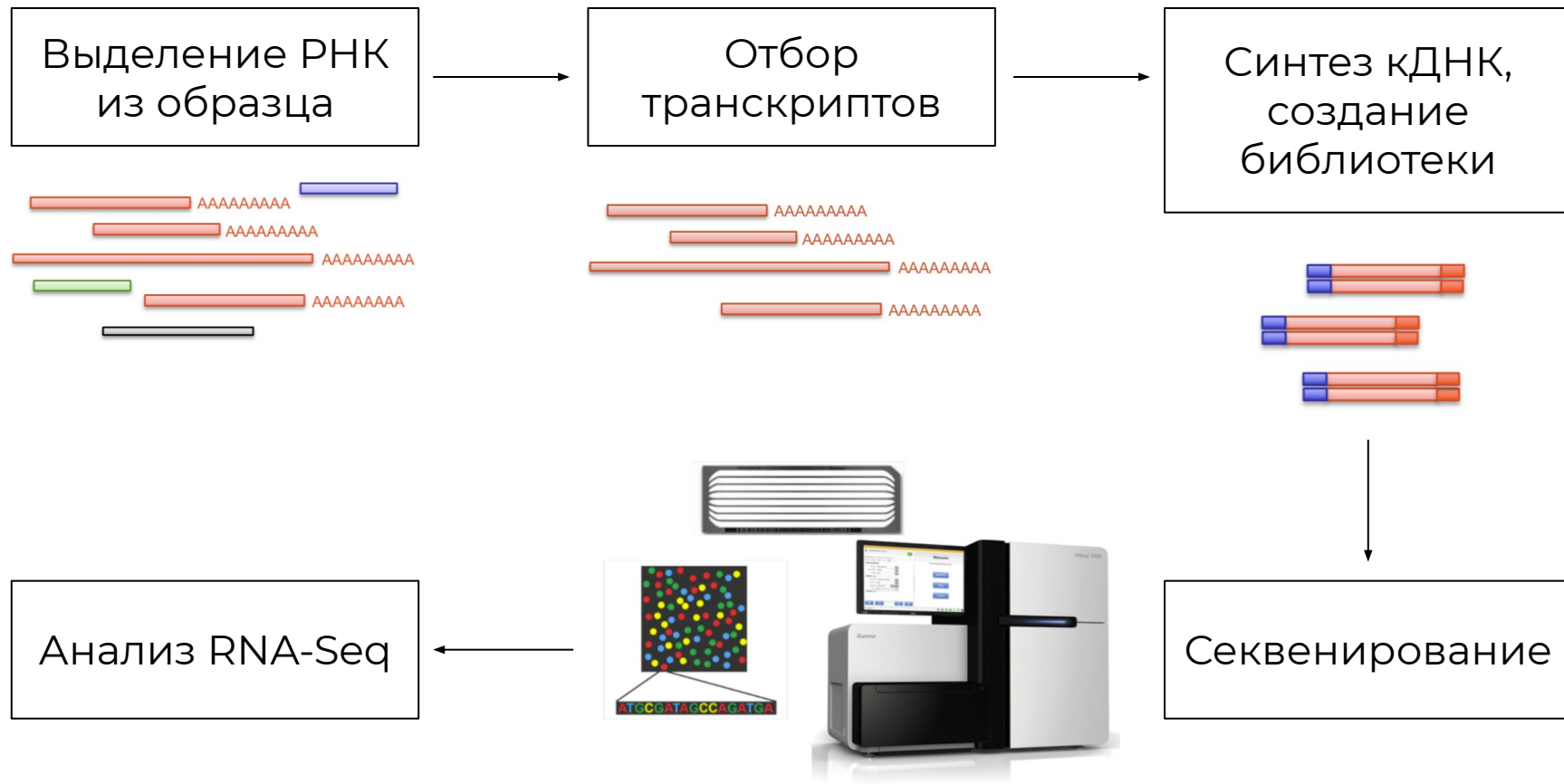
Базовая биоинформатика

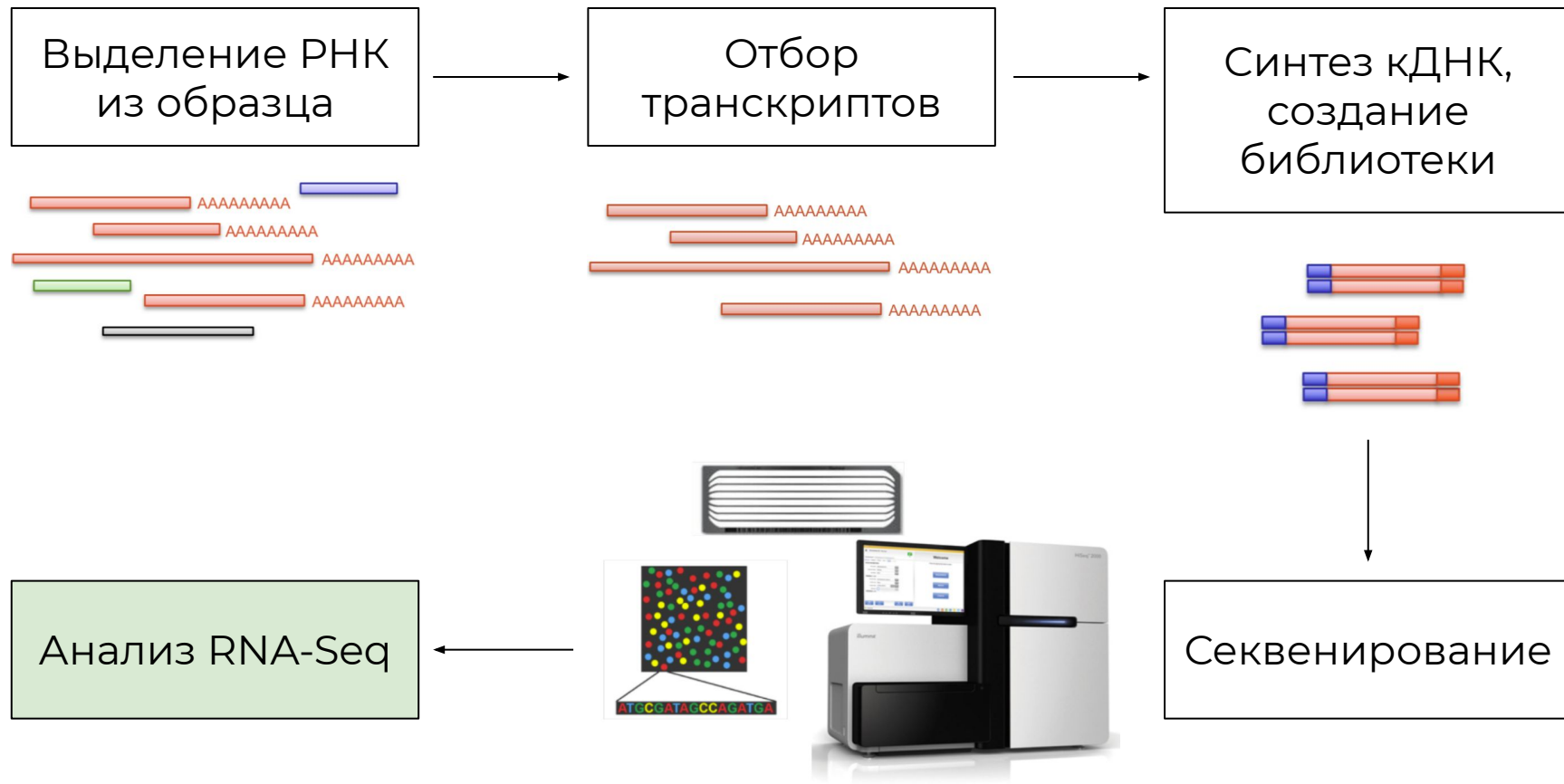
Лекции 7 и 8
Обработка RNA-Seq

12 / 11 / 2020

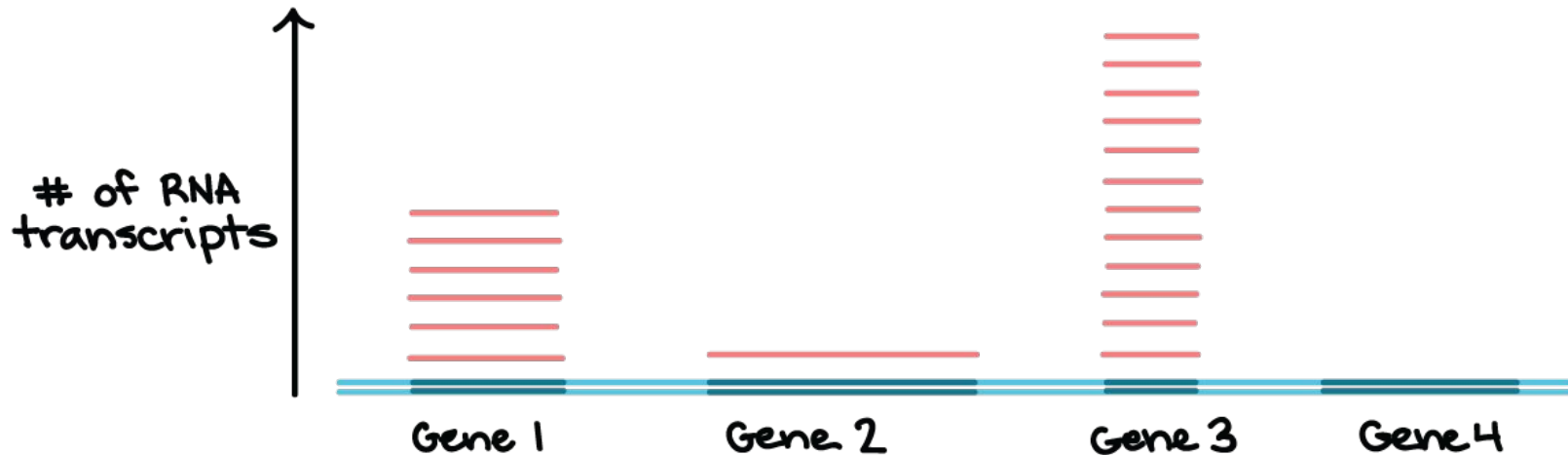


В предыдущей серии...



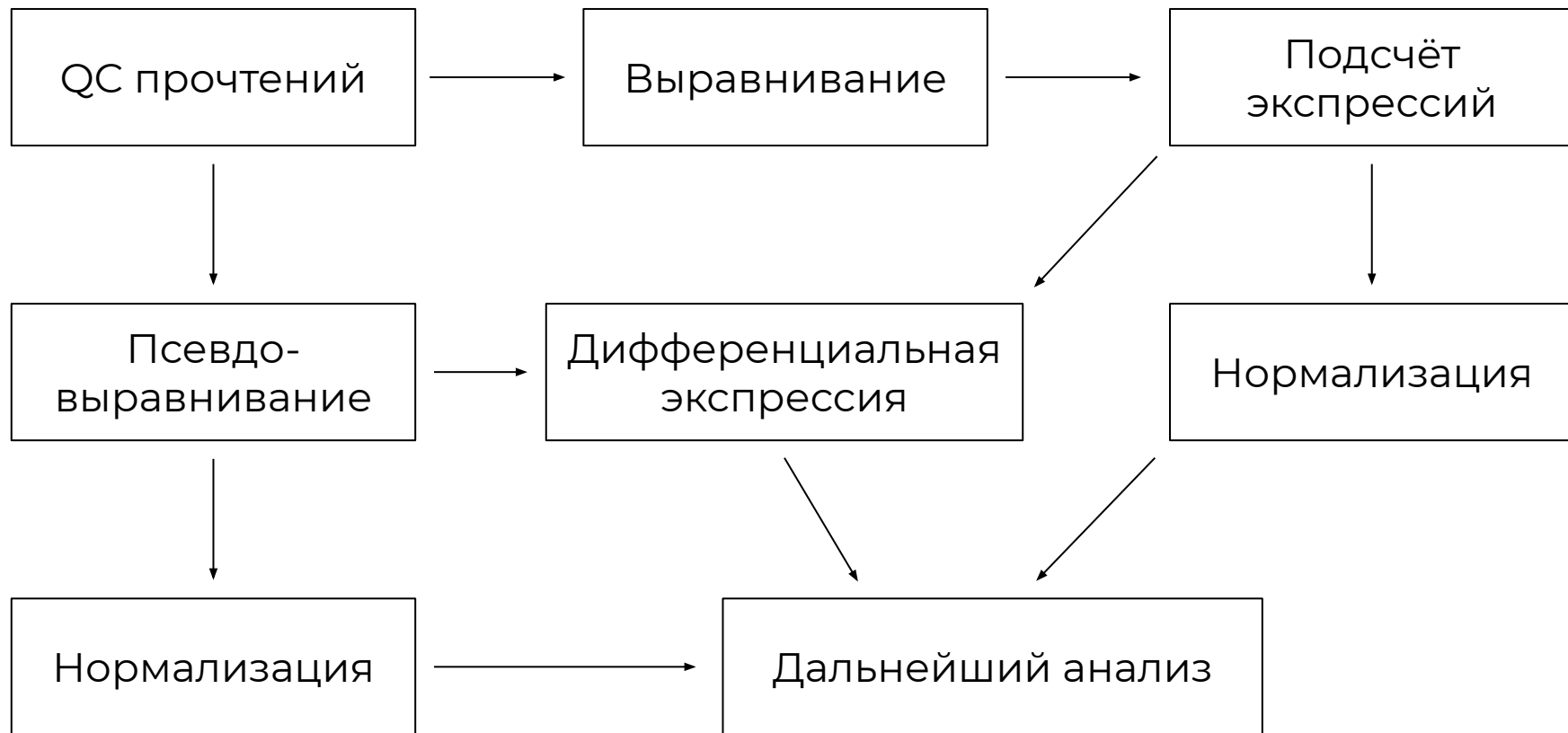


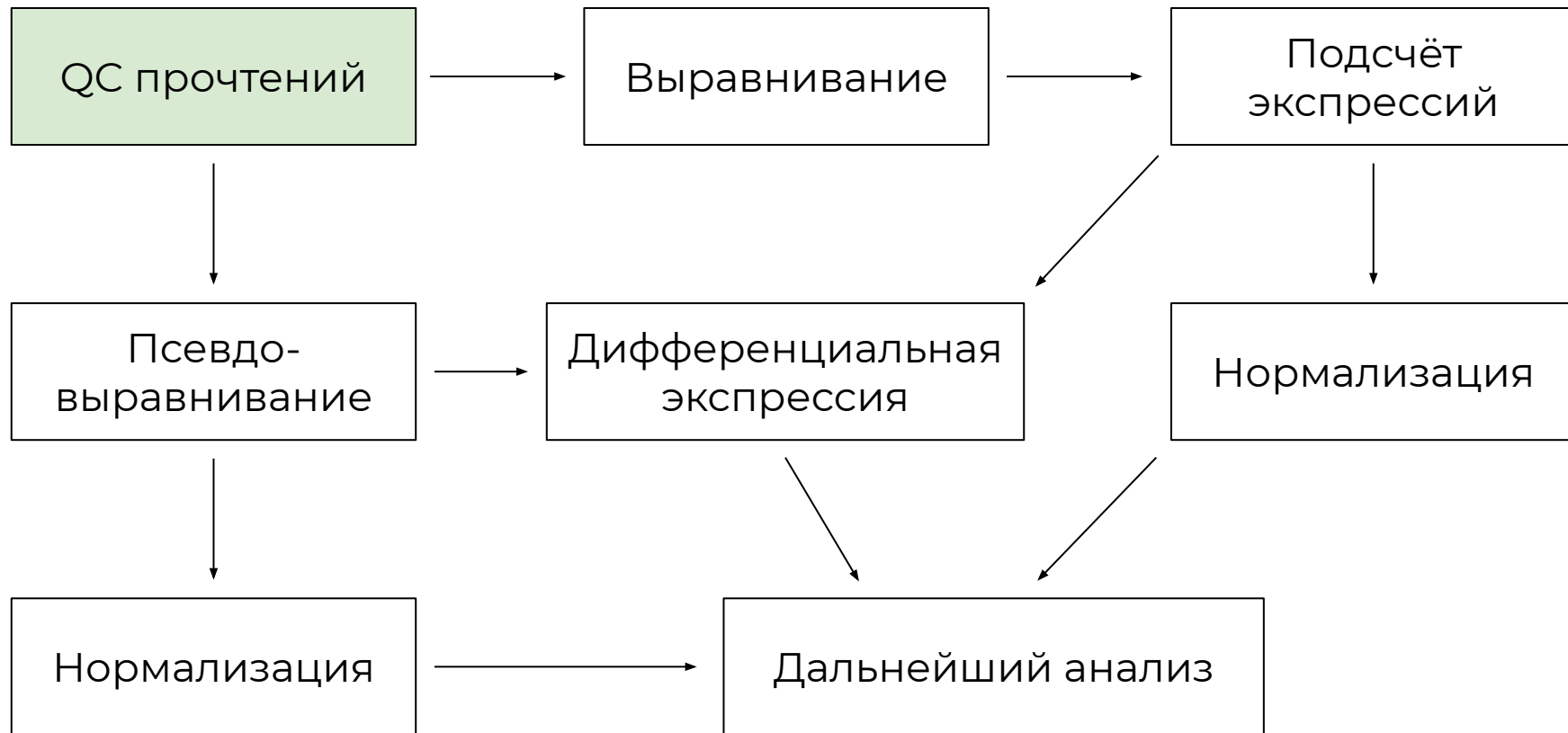
Процессинг RNA-Seq



Оценить экспрессию гена можно по его **покрытию**

“Дорожная карта” анализа RNA-Seq

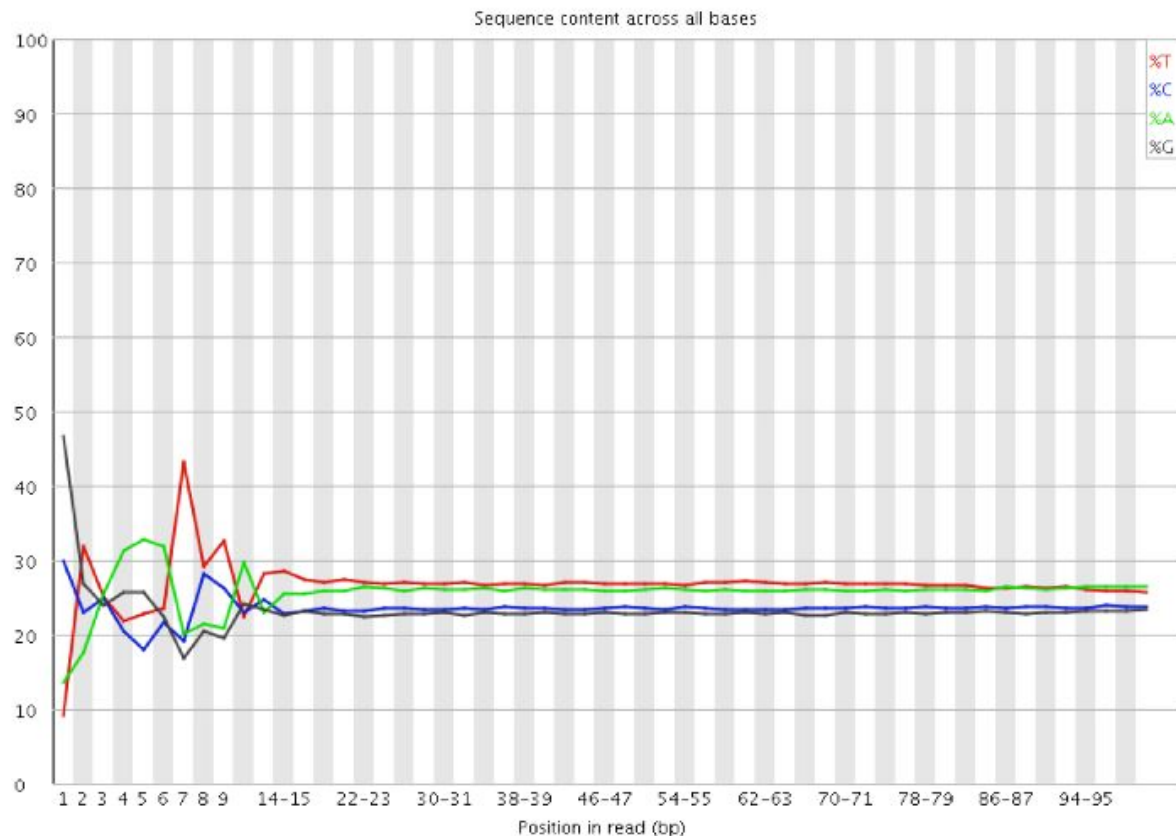




- Прочтения могут иметь “плохие концы” по краям и в целом быть засорены чем-то.
- Какие метрики качества можно использовать для того, чтобы понять, что перед нами адекватные прочтения?

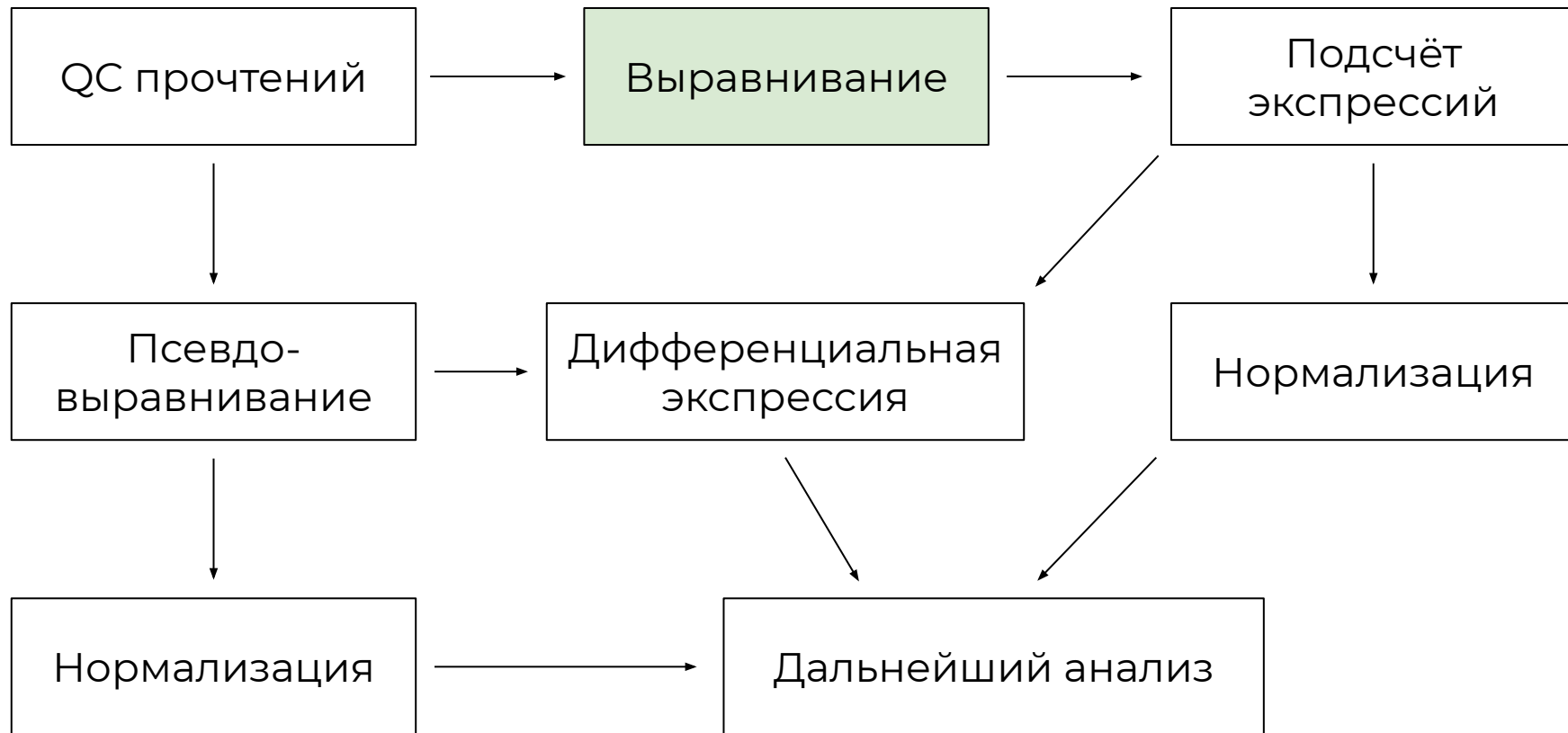
- FastQC — программа, которая оценивает базовые метрики качества прочтений, а также делает легко интерпретируемый графический отчёт.
- Пример отчёта FastQC до тримминга:
https://kodomo.fbb.msu.ru/~ann_karpukhina/files/chr11_fastqc.html
- Пример отчёта FastQC после тримминга:
https://kodomo.fbb.msu.ru/~ann_karpukhina/files/chr11_trim_fastqc.html

Per base sequence content in RNA-Seq

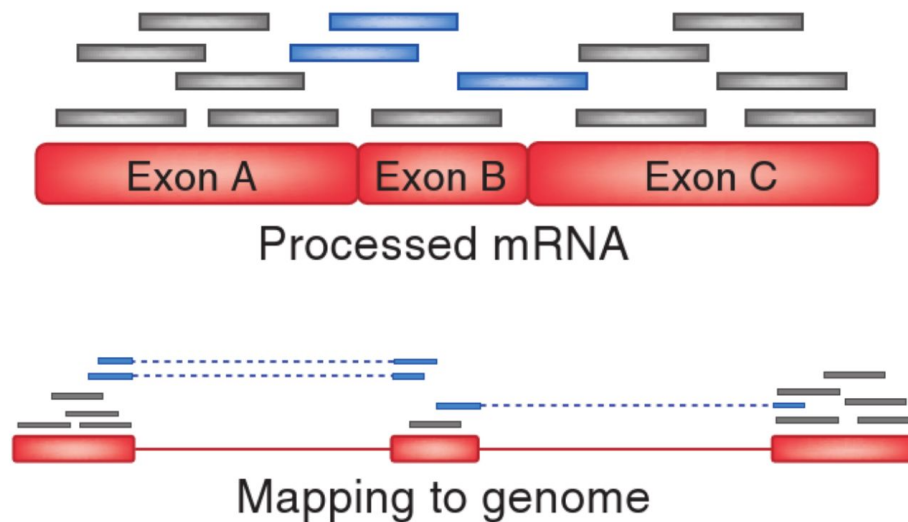


Per base sequence content в RNA-Seq

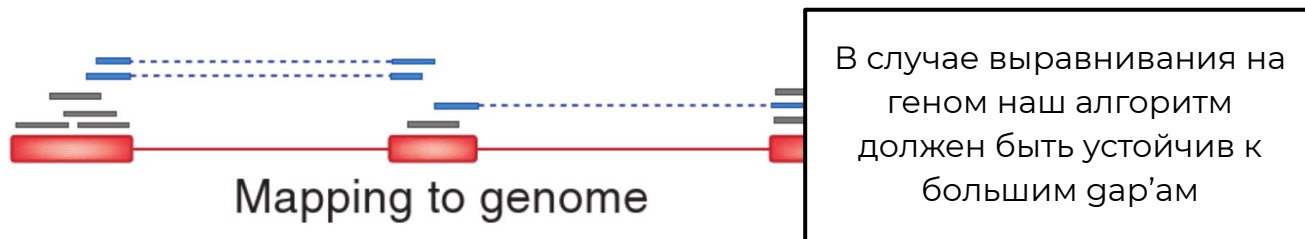
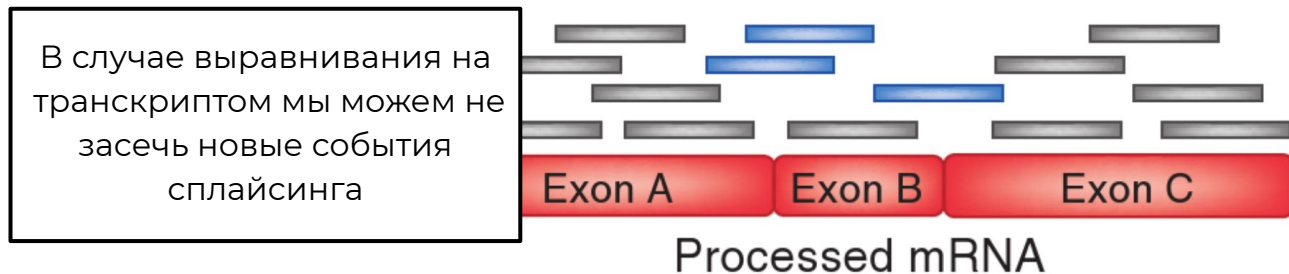




- Выравнивать можно как на референсный геном, так и на референсный транскриптом



- Выравнивать можно как на референсный геном, так и на референсный транскриптом



Sequence analysis

Advance Access publication October 25, 2012

STAR: ultrafast universal RNA-seq aligner

Alexander Dobin^{1,*}, Carrie A. Davis¹, Felix Schlesinger¹, Jorg Drenkow¹, Chris Zaleski¹, Sonali Jha¹, Philippe Batut¹, Mark Chaisson² and Thomas R. Gingeras¹

¹Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA and ²Pacific Biosciences, Menlo Park, CA, USA

Associate Editor: Inanc Birol

- Около 15 000 цитирований за 8 лет. Рекомендован ENCODE.
- Хорошо работает даже с большими отличиями от референса. Прост в использовании.
- Требуется очень большое количество RAM.

- В работах вы можете встретить и другие программы, которые используются для выравнивания прочтений RNA-Seq.

Name	Version	Mapping	Reference
Bowtie	2.2.6	Unspliced read aligner	[31]
BWA	0.7.12-r1039	Unspliced read aligner	[33]
TopHat	2.10	Spliced read aligner	[18]
STAR	2.5.3	Spliced read aligner	[34]
kallisto	0.43.1	pseudo-alignment	[35]
Salmon	0.8.2	pseudo-alignment	[36]

<https://doi.org/10.1371/journal.pone.0190152.t001>

Файл выравнивания: SAM/BAM

```
Coord      12345678901234  5678901234567890123456789012345
ref        AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1      TTAGATAAAGGATA*CTG
+r002        aaaAGATAA*GGATA
+r003        gcctaAGCTAA
+r004          ATAGCT.....TCAGC
-r003          ttagctTAGGC
-r001/2          CAGCGGCAT
```

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001  99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M * 0  0 AAAAGATAAGGATA *
r003   0 ref  9 30 5S6M * 0  0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004   0 ref 16 30 6M14N5M * 0  0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0  0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

Файл выравнивания: SAM/BAM

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45

r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

Header
section

Alignment
section

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

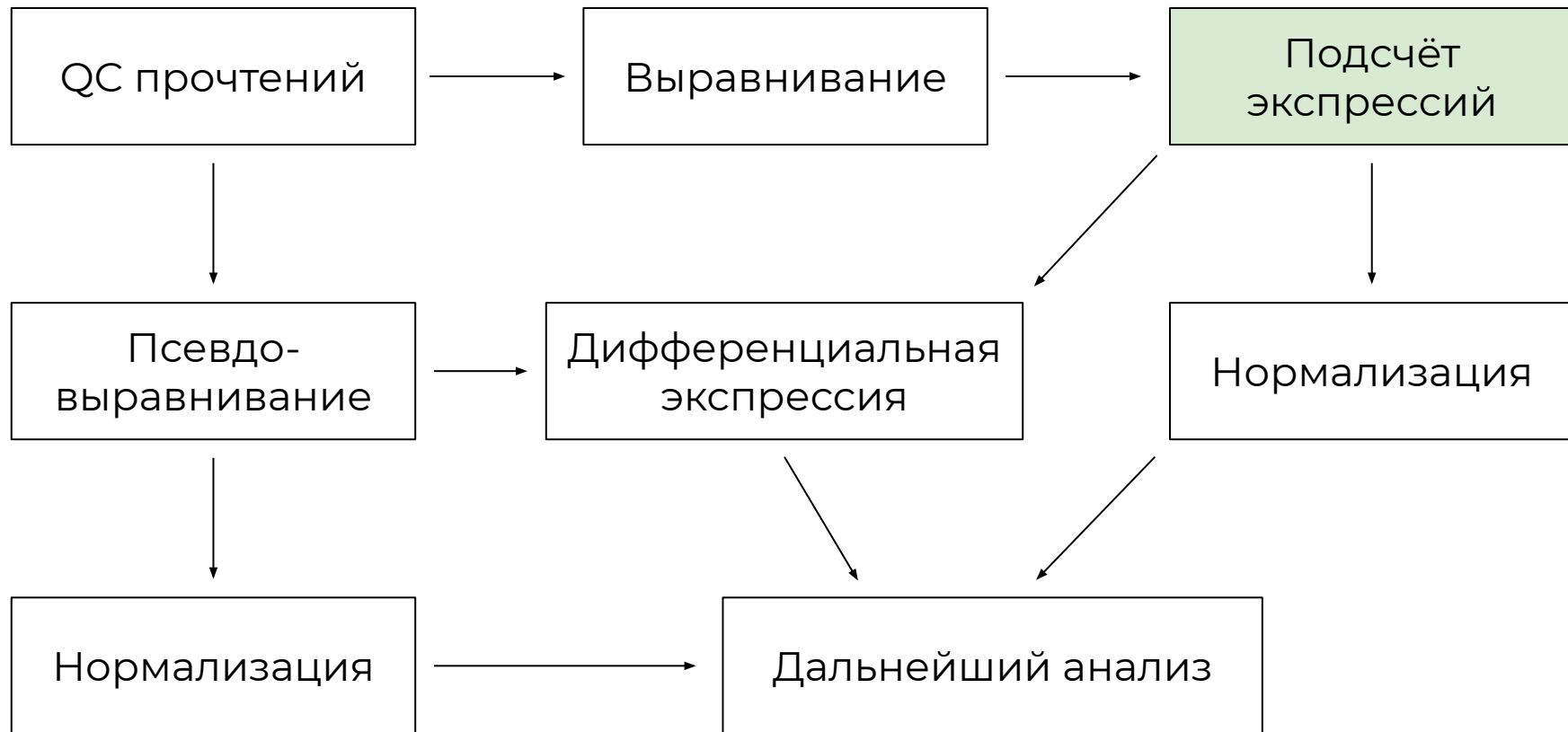
POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

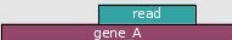

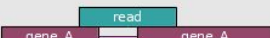

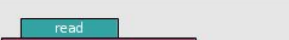
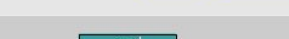
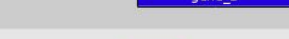

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

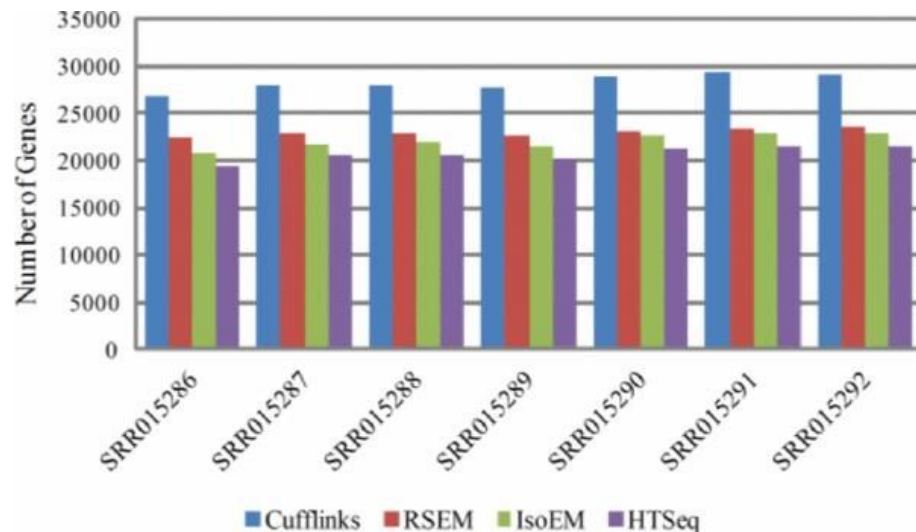
Подсчёт числа ридов на ген



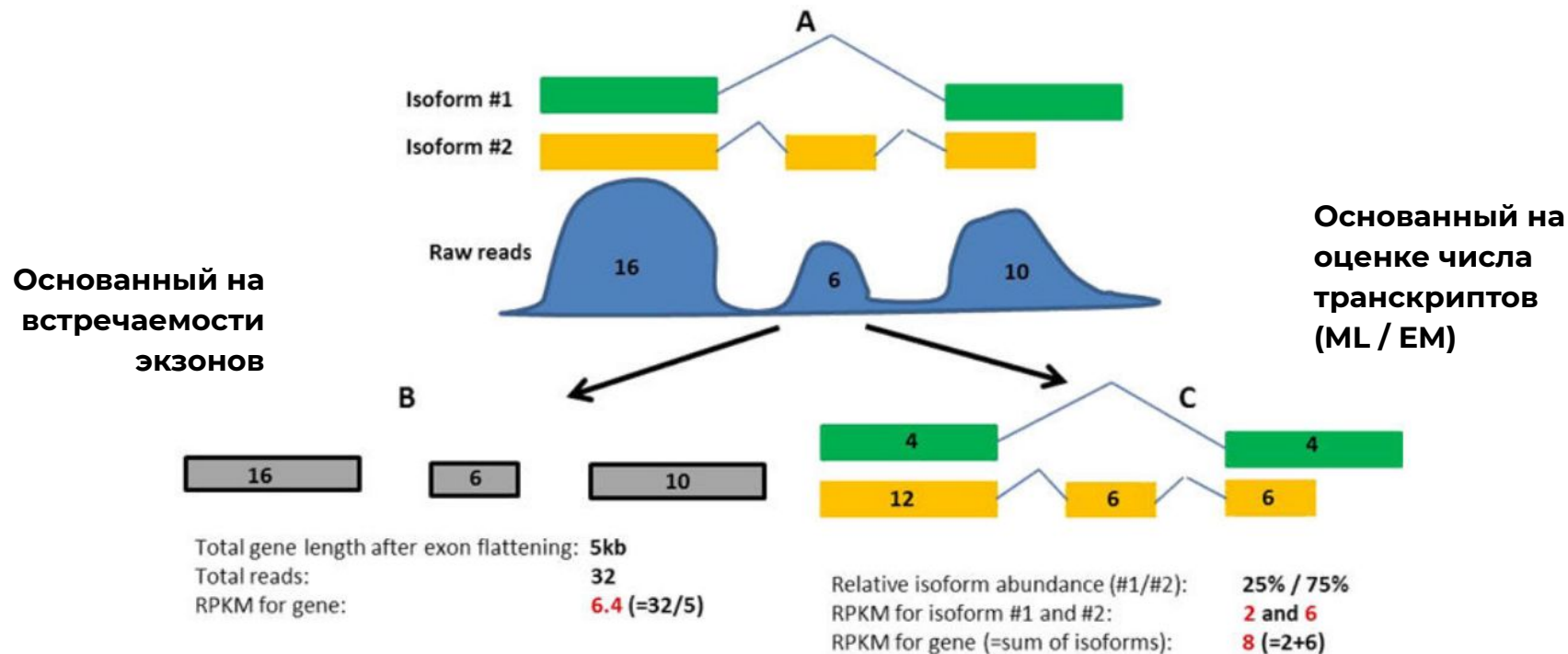
Подсчёт числа ридов на ген

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous (both genes with --nonunique all)	gene_A	gene_A
	ambiguous (both genes with --nonunique all)		
	alignment_not_unique (both genes with --nonunique all)		

- В STAR по умолчанию “вшит” **HTSeq**, который подсчитывает число прочтений на ген.

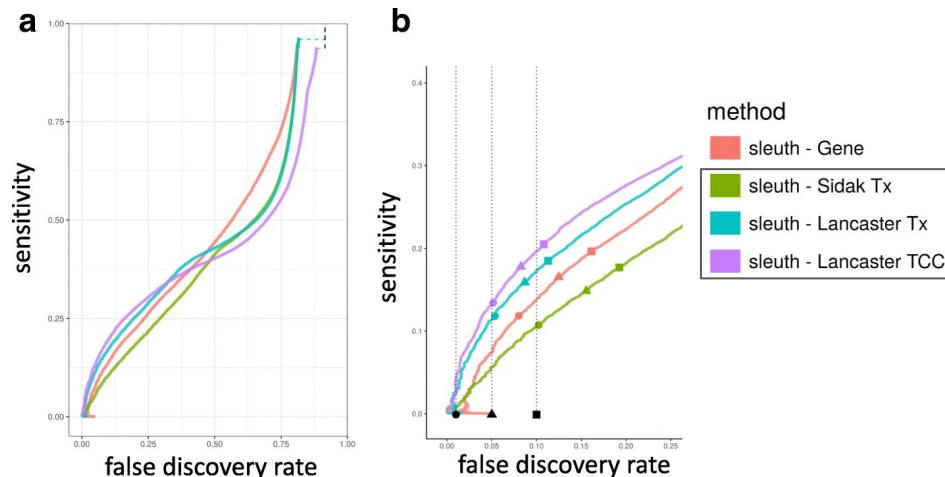
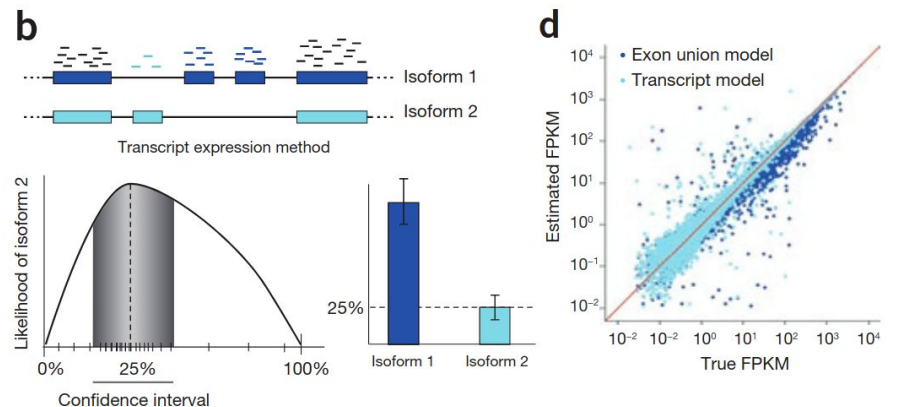


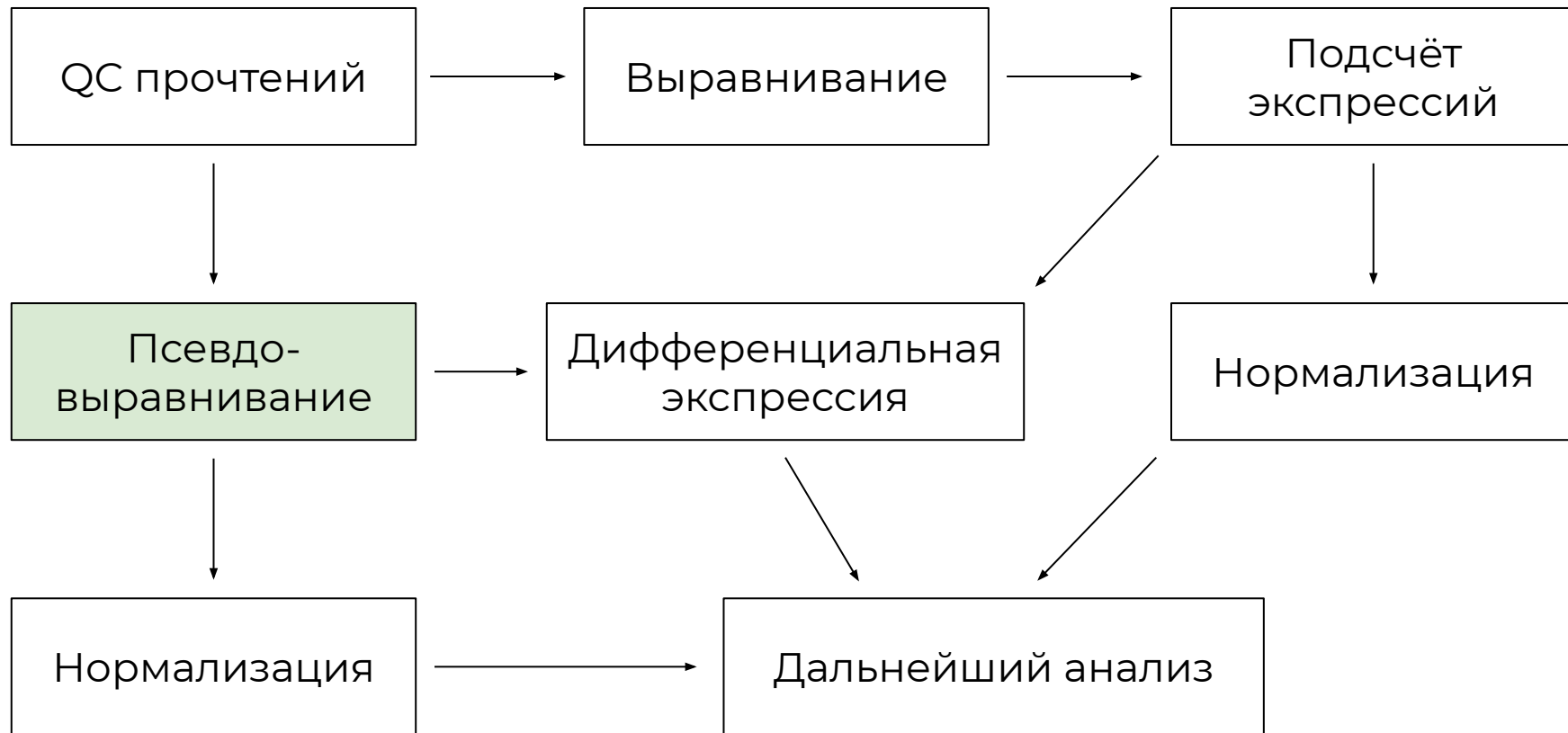
Подсчёт числа ридов на транскрипт

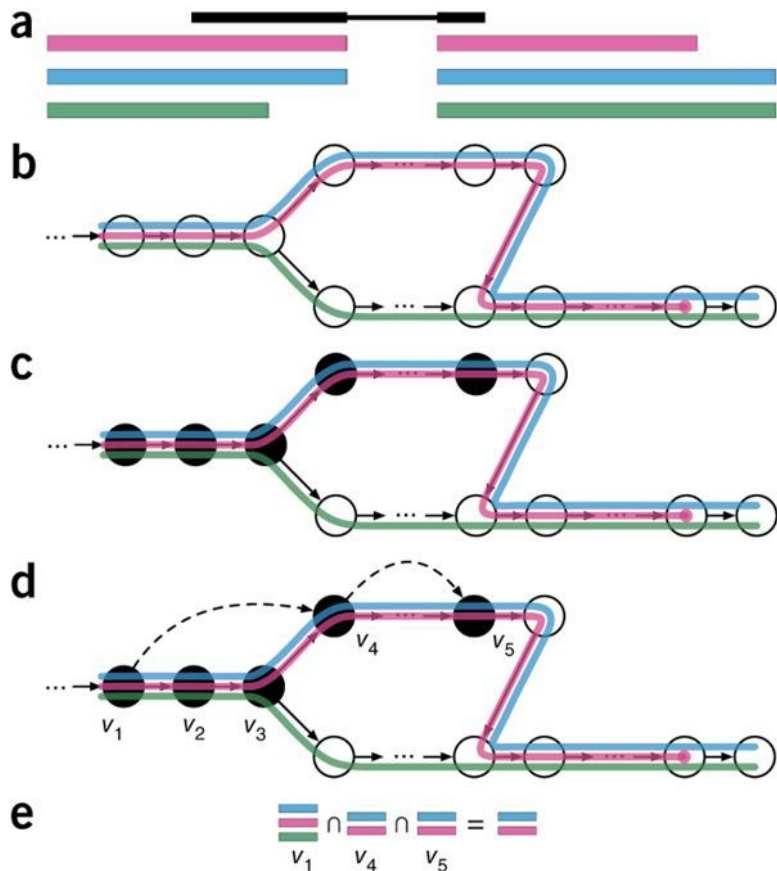


Подсчёт числа ридов на транскрипт

- Чаще всего подсчёт представленности различных изоформ реализуется при помощи ML/EM (рисунок сверху).
- Дифференциальная экспрессия, посчитанная на уровне транскриптов, более точна, чем посчитанная на уровне генов (рисунок снизу).







Published: 04 April 2016

Near-optimal probabilistic RNA-seq quantification

Nicolas L Bray, Harold Pimentel, Páll Melsted & Lior Pachter

Nature Biotechnology **34**, 525–527(2016) | [Cite this article](#)

23k Accesses | **1793** Citations | **167** Altmetric | [Metrics](#)

- **Kallisto** имеет около двух тысяч цитирований.
- Работает **очень** быстро и в последнее время широко используется.
- Не выдаёт выравнивание.

Выравнивания или псевдовыравнивания?

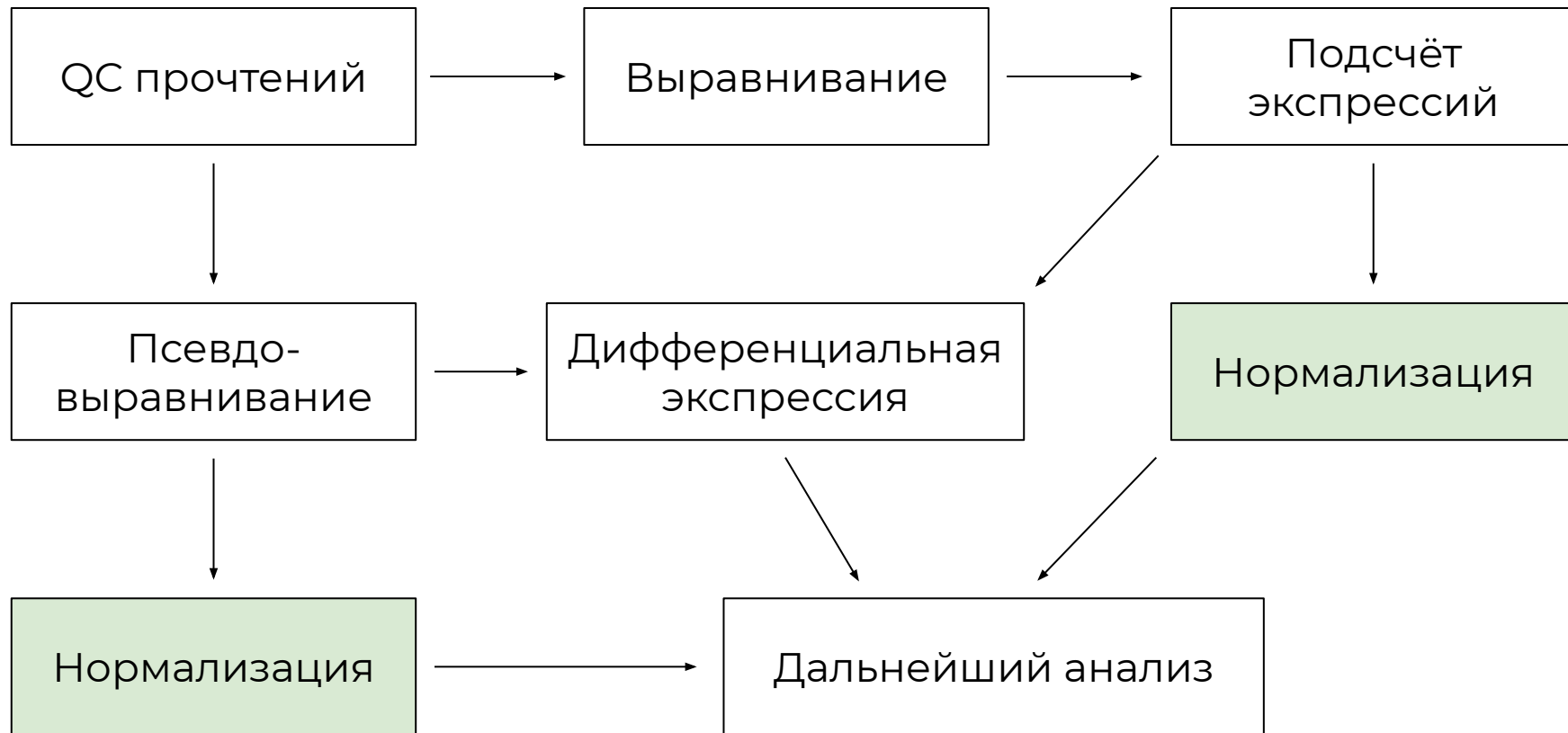
- В целом методы достаточно хорошо скоррелированы.
- В зависимости от задач используют разные подходы, в BostonGene в основном используют kallisto (в TCGA тоже).

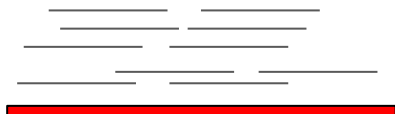
Table 1: Pearson Correlations on Transcript Counts ($\log(\text{counts}+1)$), Sim

	kallisto	Salmon	HISAT2	STAR	ground truth
kallisto	1	0.998	0.986	0.977	0.951
Salmon	0.998	1	0.987	0.977	0.951
HISAT2	0.986	0.987	1	0.977	0.949
STAR	0.977	0.977	0.977	1	0.941
ground truth	0.951	0.951	0.949	0.941	1

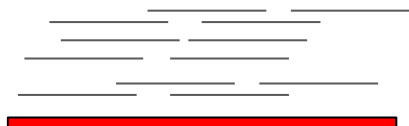
Table 5: Pearson correlation on Transcript Counts ($\log(\text{counts}+1)$), Zika

	kallisto	Salmon_0.8.2	Salmon_0.11.2	HISAT2	STAR
kallisto	1	0.998	0.966	0.936	0.934
Salmon_0.8.2	0.998	1	0.966	0.936	0.934
Salmon_0.11.2	0.966	0.966	1	0.970	0.966
HISAT2	0.936	0.936	0.970	1	0.976
STAR	0.934	0.934	0.966	0.976	1





Как сравнить уровни экспрессии этих генов?



Каунты — это число ридов, откартированных на ген или транскрипт. Число каунтов зависит от:

1. экспрессии гена,
2. размера библиотеки,
3. длины и
4. GC-состава гена.

	Wild-type		Mutant	
	Sample 1	Sample 2	Sample 3	Sample 4
Gene 1	24	31	76	59
Gene 2	0	3	7	2
Gene 3	1988	1125	3052	2450
Gene 4	5	0	0	1
...
Total	22341961	20739175	15669423	23711320

$$\theta_i = P(\text{read from transcript } i) = Z^{-1} \tau_i \ell'_i$$

$$Z = \sum_i \tau_i \ell'_i$$

expression level

length

- Для того, чтобы учесть длину гена и глубину секвенирования, придумали метрику **RPKM**.

$$RPKM = 10^9 \times \frac{C}{N * L}$$

- **C** is the number of mappable reads mapped onto the gene's exons.
- **N** is the total number of mappable reads in the experiment.
- **L** is the total length of the exons in base pairs.
- Fragments Per Kilobase of exon per Million fragments mapped (FPKM),

RPKM

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

Total reads: 35 45 106

Tens of reads: 3.5 4.5 10.6



Gene Name	Rep1 RPM	Rep2 RPM	Rep3 RPM
A (2kb)	2.86	2.67	2.83
B (4kb)	5.71	5.56	5.66
C (1kb)	1.43	1.78	1.43
D (10kb)	0	0	0.09



Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009

- Существует и другая метрика — **TPM**.

$$\text{TPM}_i = \frac{X_i}{\tilde{l}_i} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\tilde{l}_j}} \right) \cdot 10^6$$

- В чём её смысл?

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1



Gene Name	Rep1 RPK	Rep2 RPK	Rep3 RPK
A (2kb)	5	6	15
B (4kb)	5	6.25	15
C (1kb)	5	8	15
D (10kb)	0	0	0.1
Total RPK:	15	20.25	45.1



Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02

TPM vs. RPKM

- Для того, чтобы сравнить экспрессию какого-то гена между образцами, лучше использовать TPM.
- Домашнее задание на дополнительный балл: написать программу, которая переводит каунты из RPKM в TPM (для данного транскриптома).

RPKM

Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009

Total: 4.29 4.5 4.25

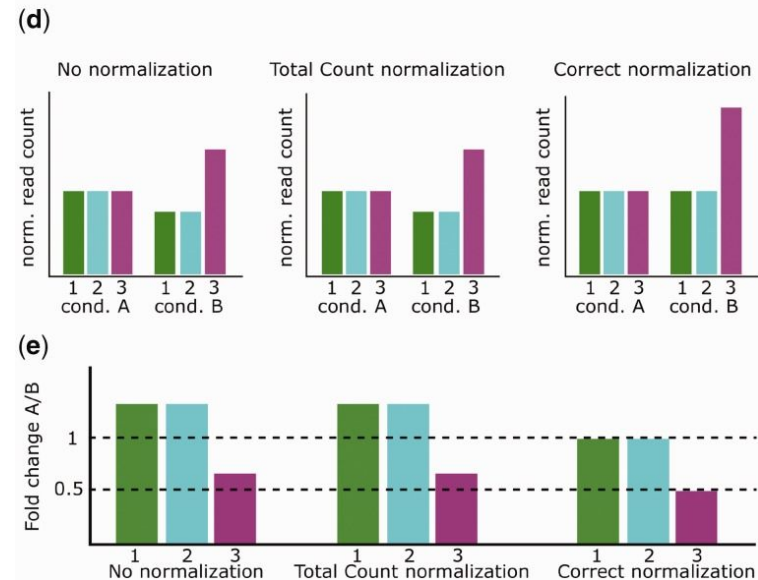
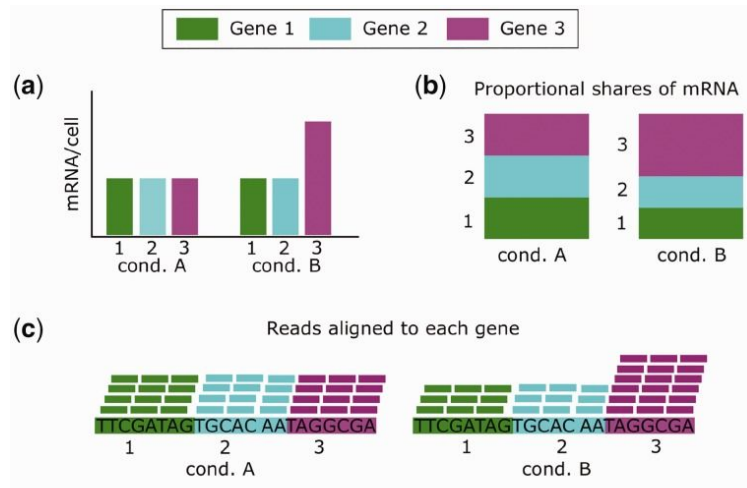
TPM

Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02

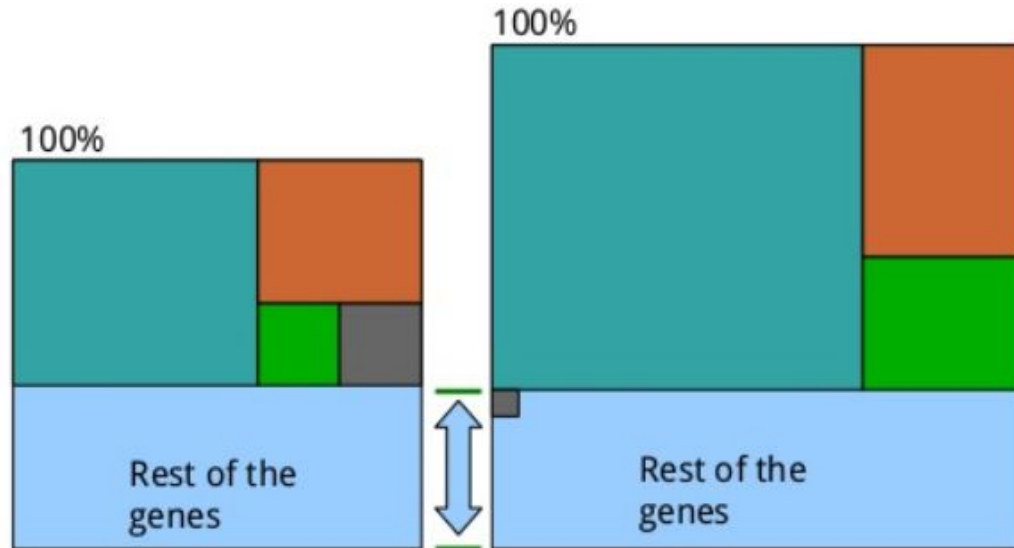
Total: 10 10 10

Проблемы RPKM и TPM

- Есть две среды с разными условиями (**A** и **B**). В средах группы **A** клетки начинают экспрессировать ген **3**, экспрессия остальных не меняется. При нормализации при помощи TPM/RPKM мы увидим “уменьшение” экспрессии других генов.

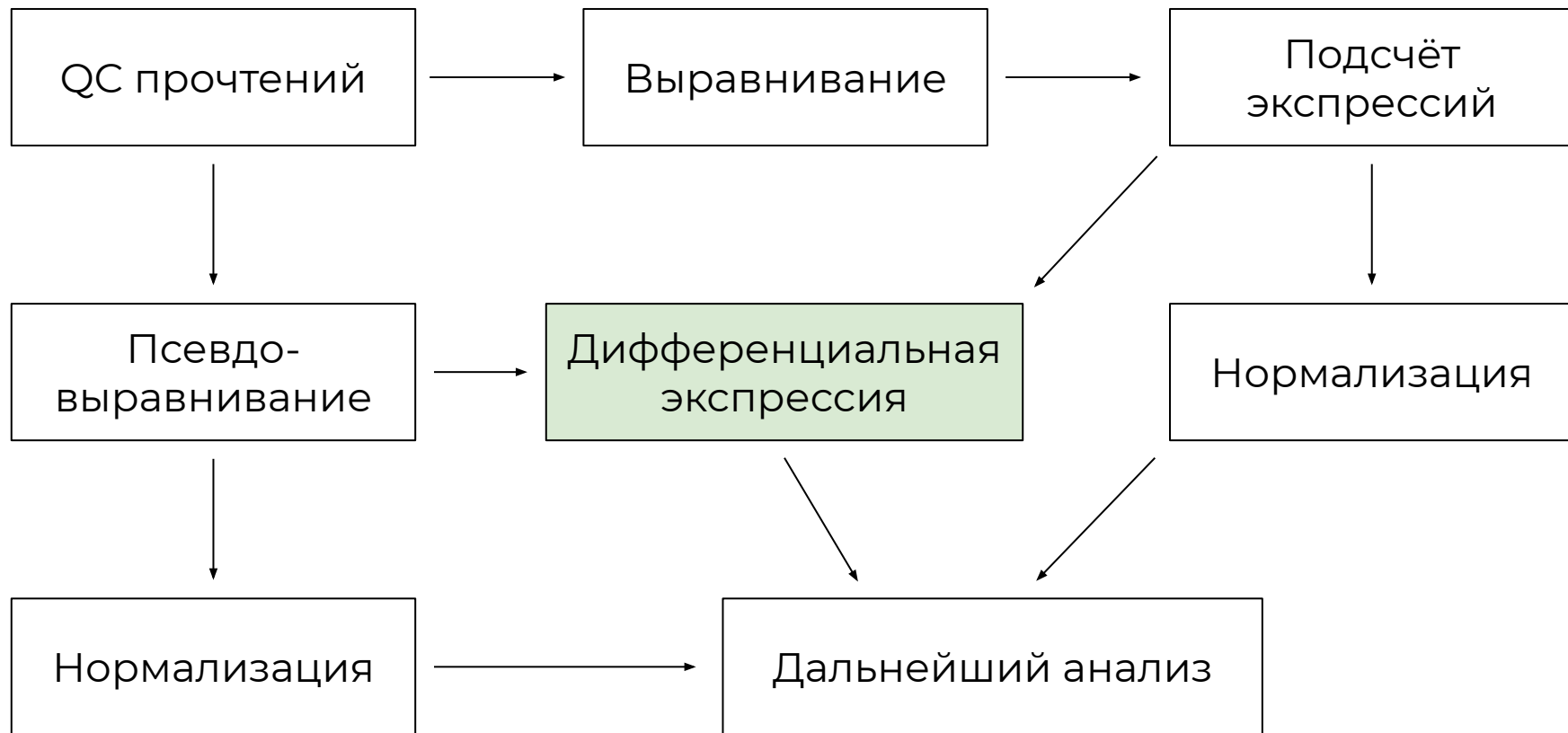


- Предположение: основное количество генов не дифференциально экспрессированы. Можно оценить их разницу, и так мы узнаем **эффективный размер** изучаемой библиотеки.



- Предположение: основное количество генов не дифференциально экспрессированы. Можно оценить их разницу, и так мы узнаем **эффективный размер** изучаемой библиотеки.
- В явном виде мы не считаем скорректированные значения экспрессий, однако это предположение “вшито” в многие библиотеки для определения дифференциальной экспрессии.

Дифф. экспрессия



Статистические методы поиска
дифференциально экспрессированных генов



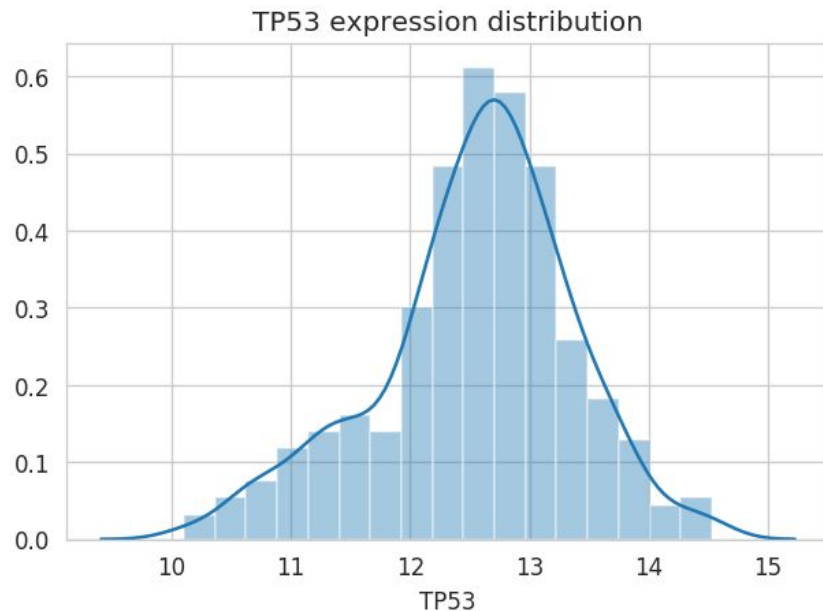
```
graph TD; A[Статистические методы поиска дифференциально экспрессированных генов] --> B[Параметрические (GLM, LM): DESeq2, edgeR, limma+voom]; A --> C[Непараметрические: NOIseq, SAMseq];
```

Параметрические (GLM, LM):
DESeq2, edgeR, limma+voom

Представляют экспрессию
как линейную комбинацию
предикторов (с некоторыми
добавками)

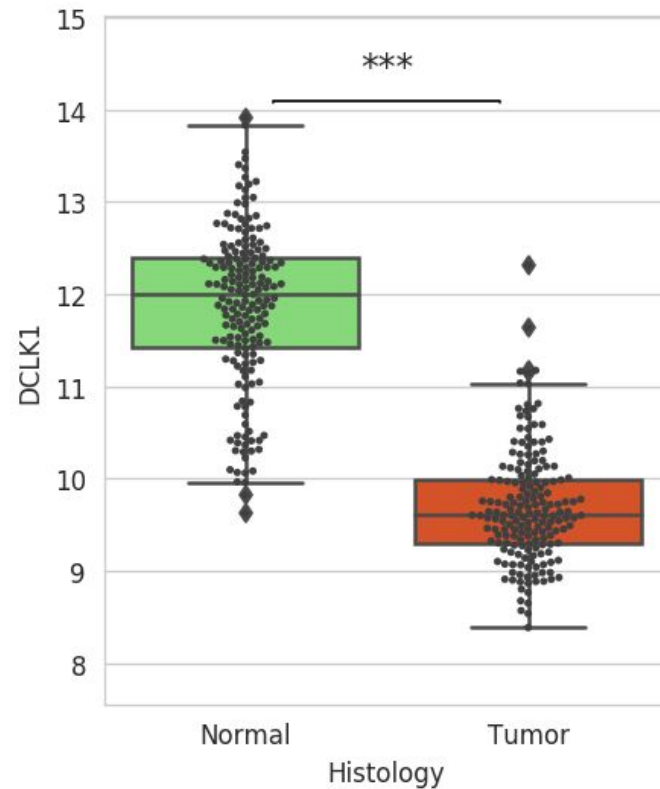
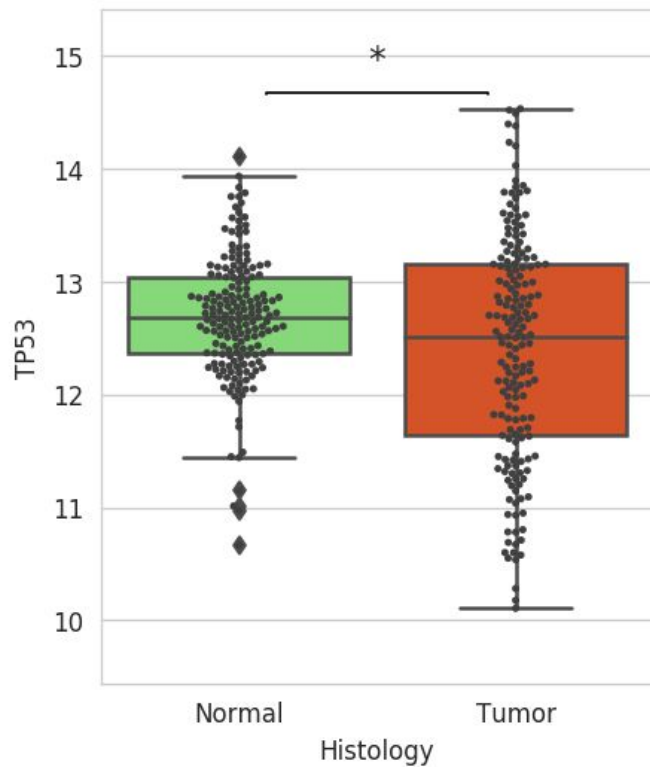
Непараметрические:
NOIseq, SAMseq

Используют
непараметрические тесты
для сравнения двух или
более групп



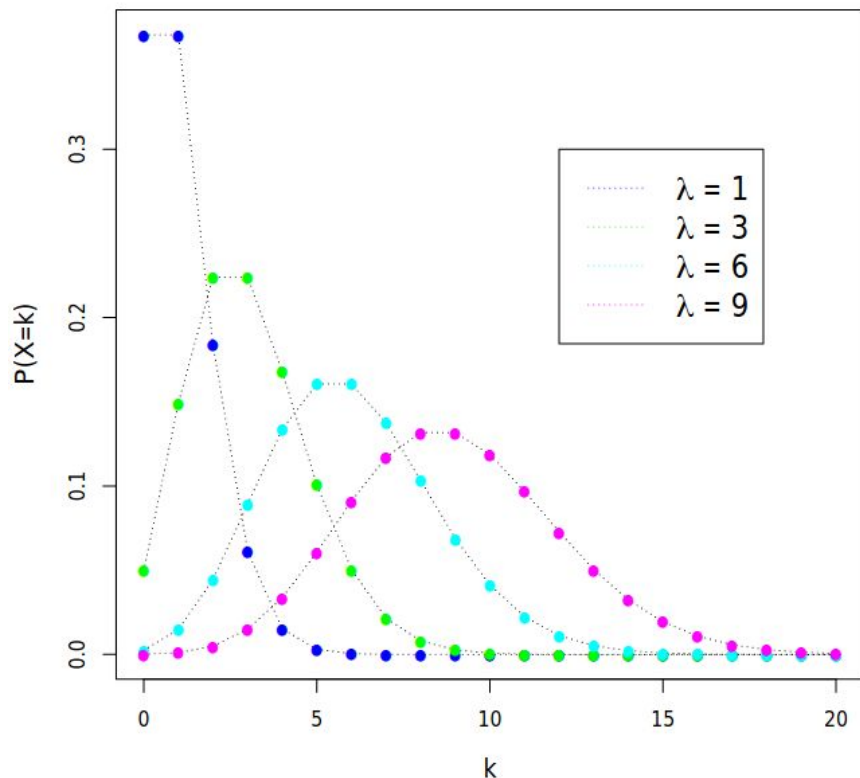
Это нормальное распределение?

Непараметрические тесты



Аппроксимирующие распределения: Распределение Пуассона

Poisson distribution



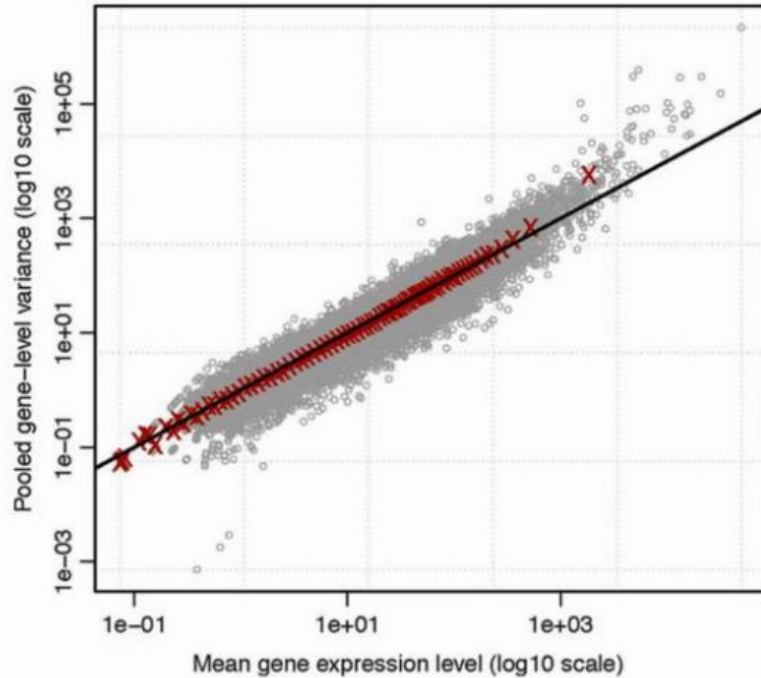
$$p(k) \equiv \mathbb{P}(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

где

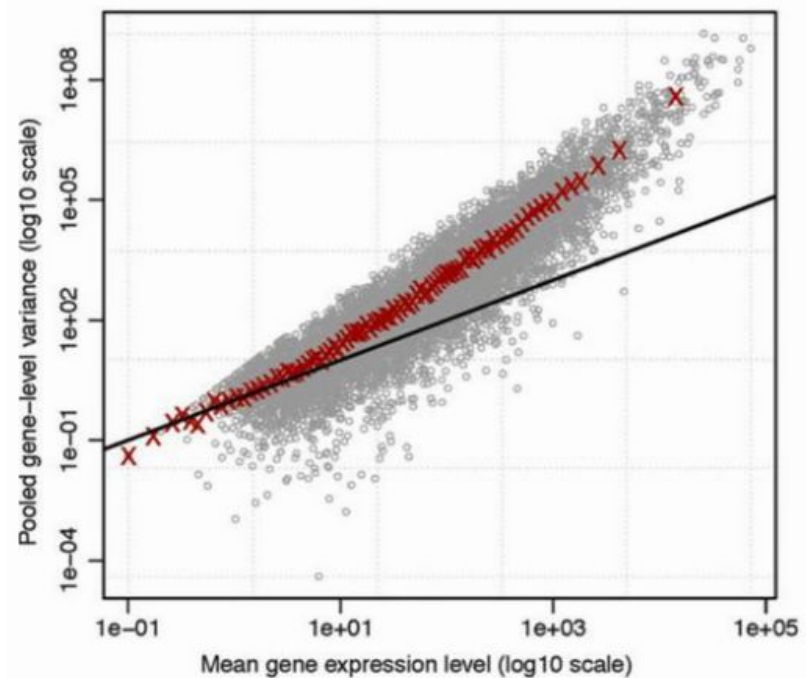
- λ — математическое ожидание случайной величины (среднее количество событий за фиксированный промежуток времени),
- $k!$ обозначает **факториал** числа k ,
- $e = 2,718281828 \dots$ — **основание натурального логарифма**.

Аппроксимирующие распределения: Распределение Пуассона

Технические вариации



Биологические вариации



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

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

Математическое

ожидание

Дисперсия

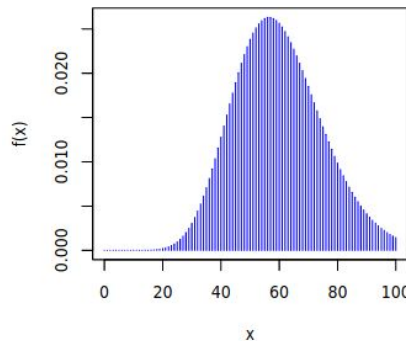
$$\frac{rq}{p}$$

$$p$$

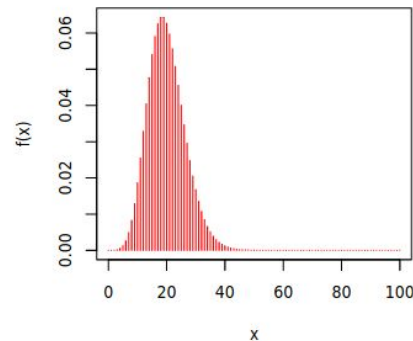
$$\frac{rq}{p^2}$$

$$p^2$$

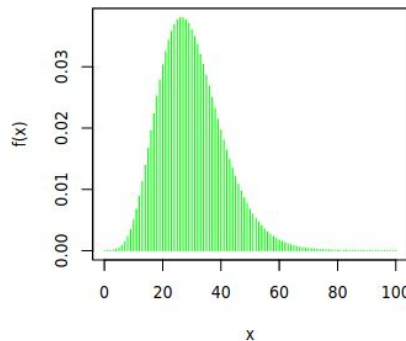
NB(20 , 0.25)



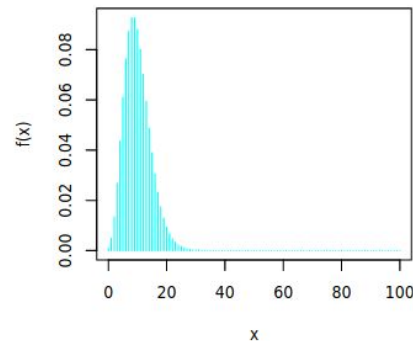
NB(20 , 0.5)



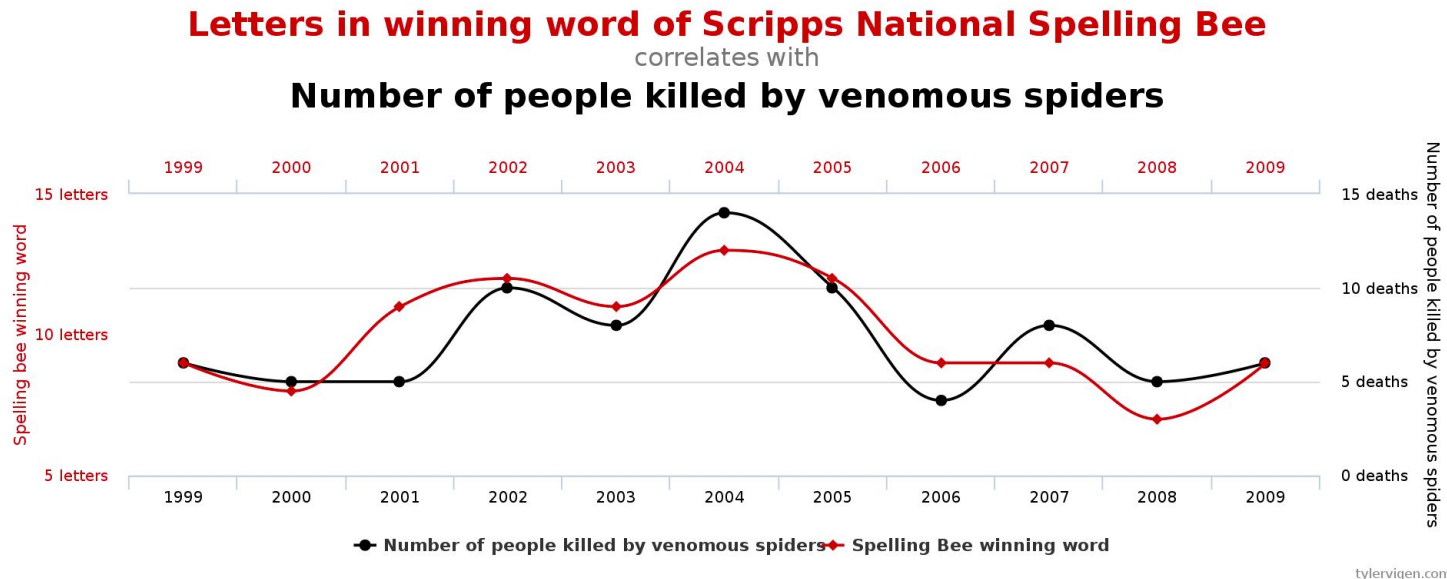
NB(10 , 0.25)



NB(10 , 0.5)



- Параметрические модели используются в различных программах для определения дифференциальной экспрессии (DESeq2, edgeR) — в основном это обратное биномиальное распределение.
- Они используют простейшие методы машинного обучения (генерализованные линейные модели) для определения факторов, влияющих на длину гена.
- В них “вшита” оценка эффективного размера библиотеки, а также искажений, связанных с длиной гена.



Решение

Пермутации

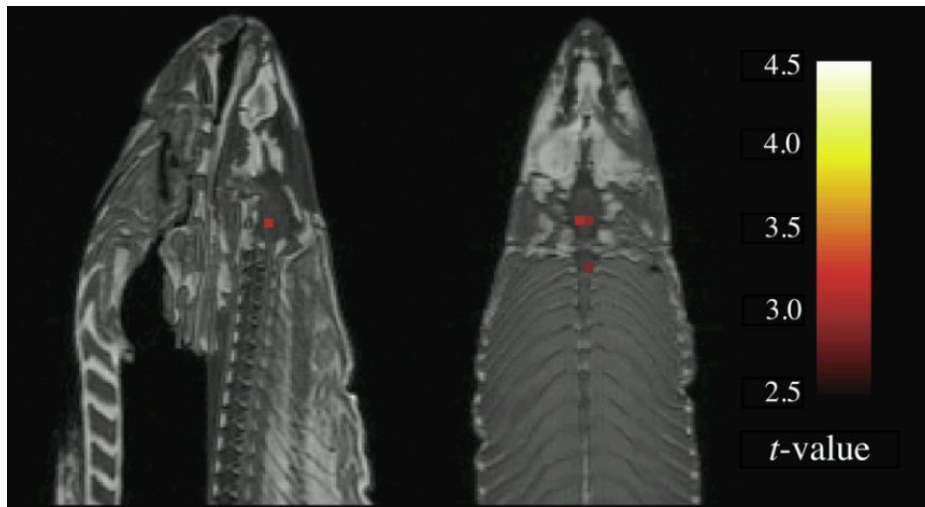
Поправка p-value

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

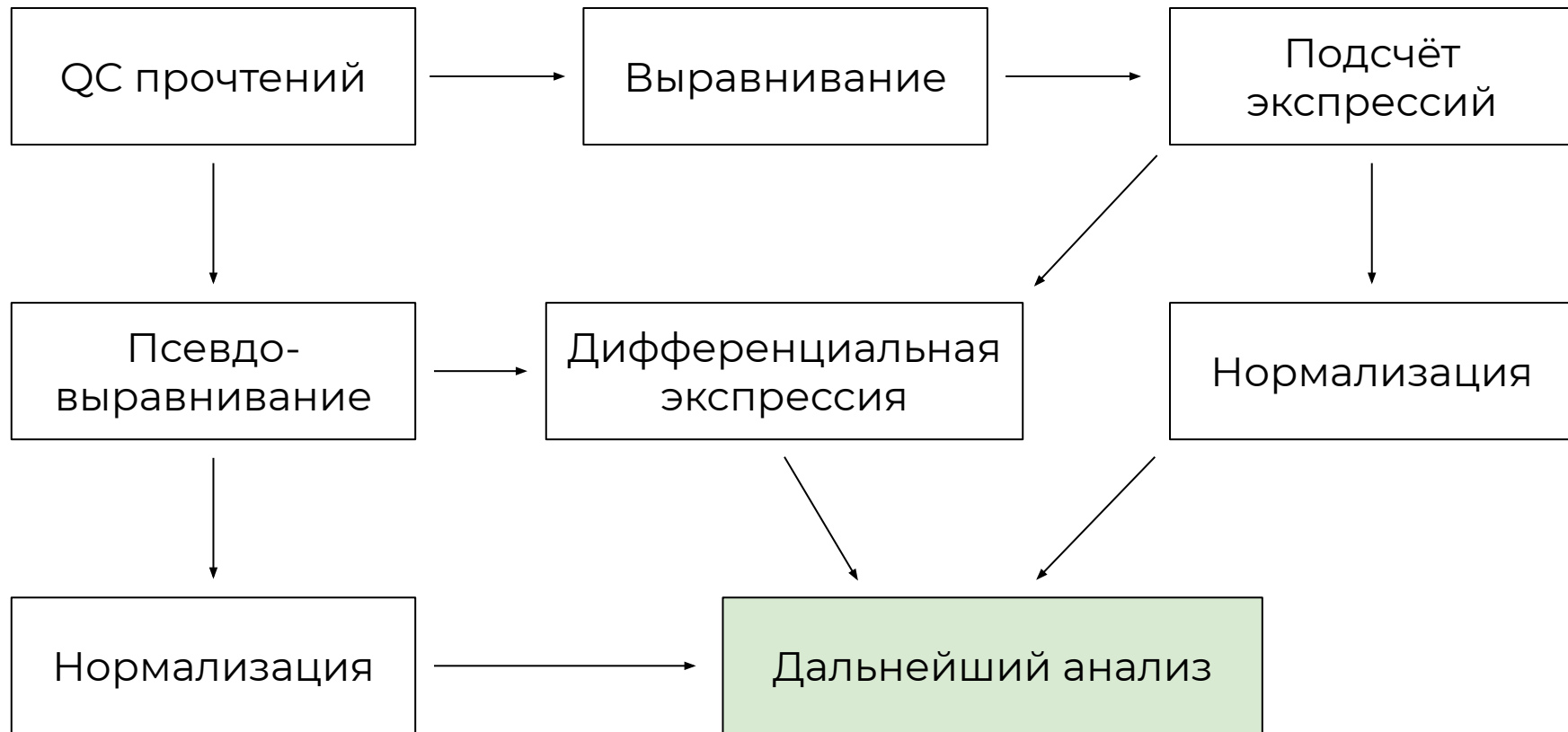
Craig M. Bennett¹, Abigail A. Baird², Michael B. Miller¹, and George L. Wolford³

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

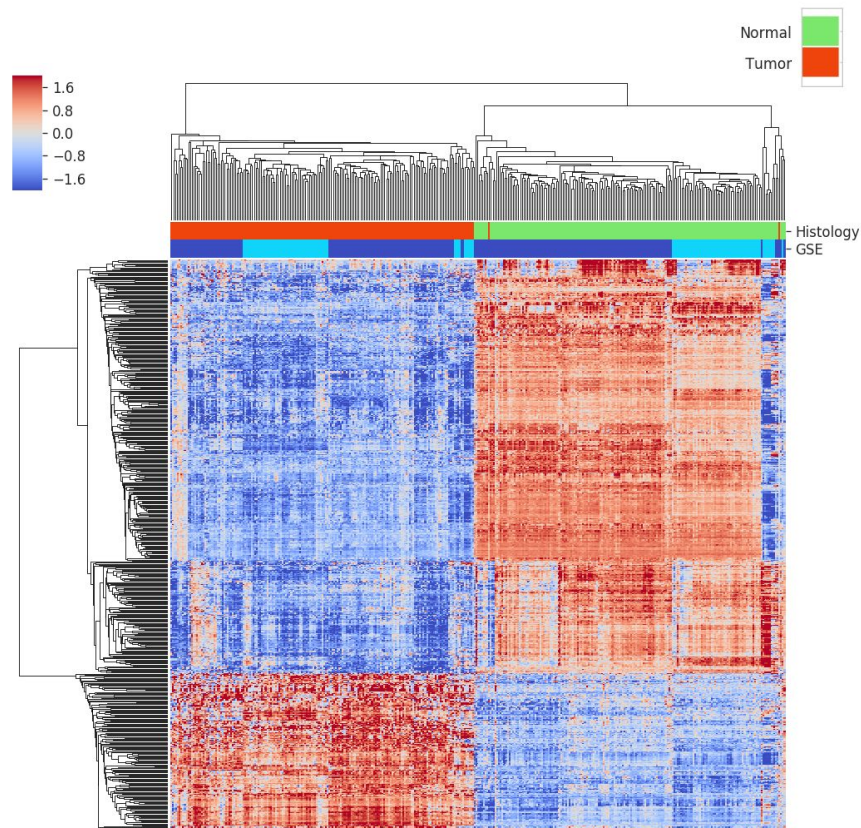
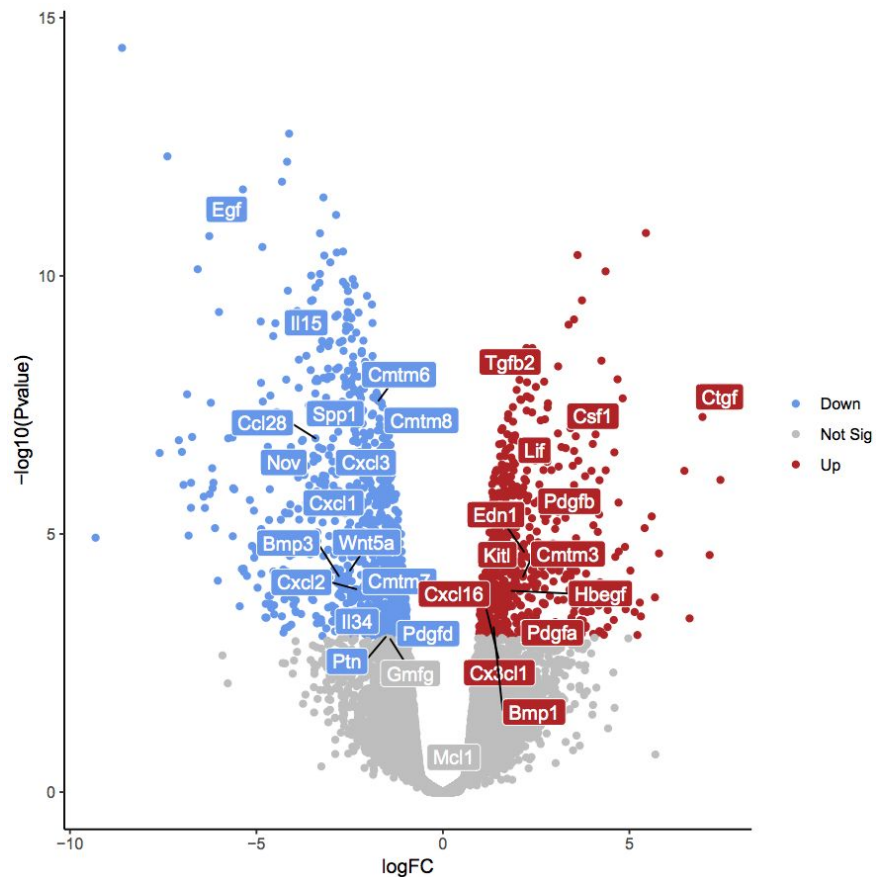
³ Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH



Дальнейший анализ



Анализ профилей экспрессии



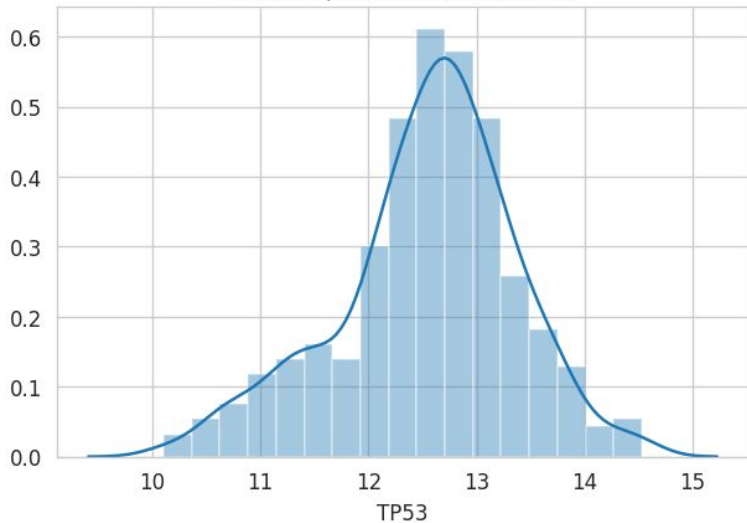
Z-скорирование

$$Z = \frac{x - \mu_x}{\sigma_x}$$

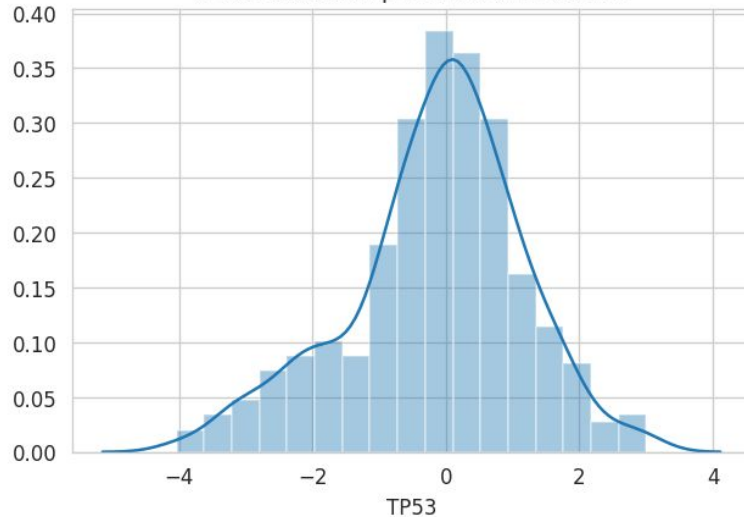
Diagram illustrating the Z-score formula with labels:

- score (points to x)
- Population mean (points to μ_x)
- Population standard deviation (points to σ_x)

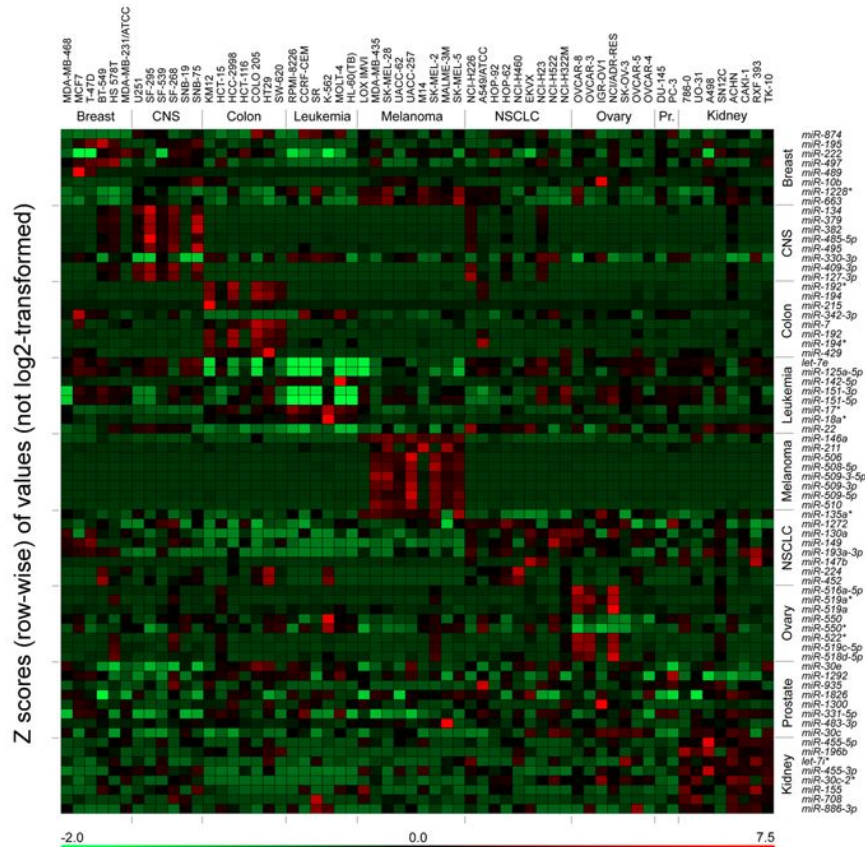
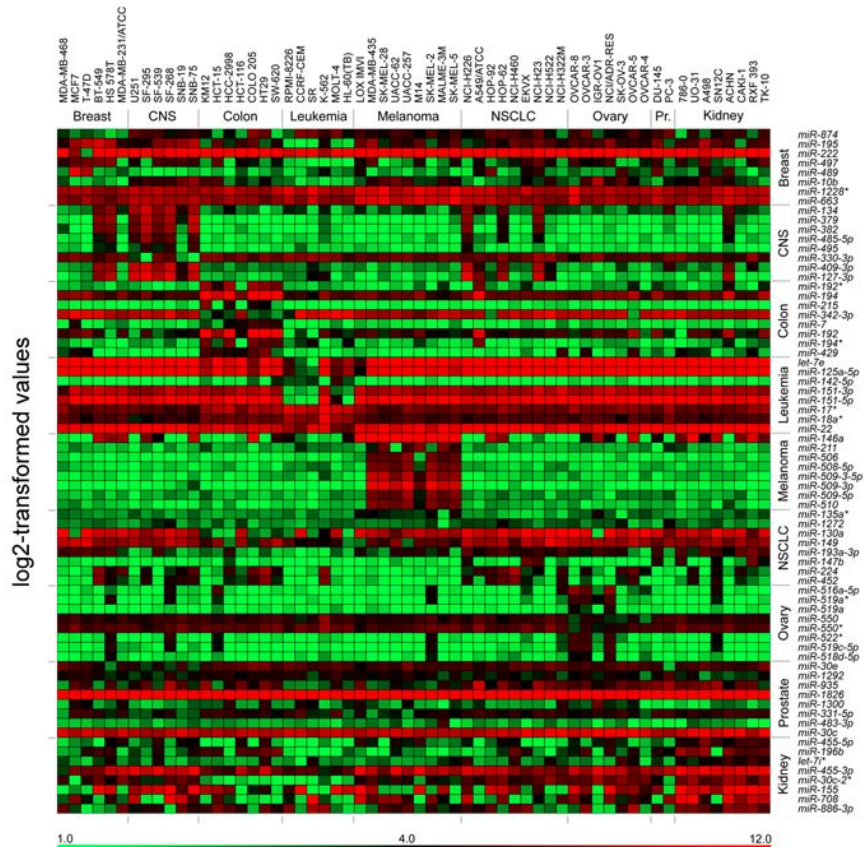
TP53 expression distribution



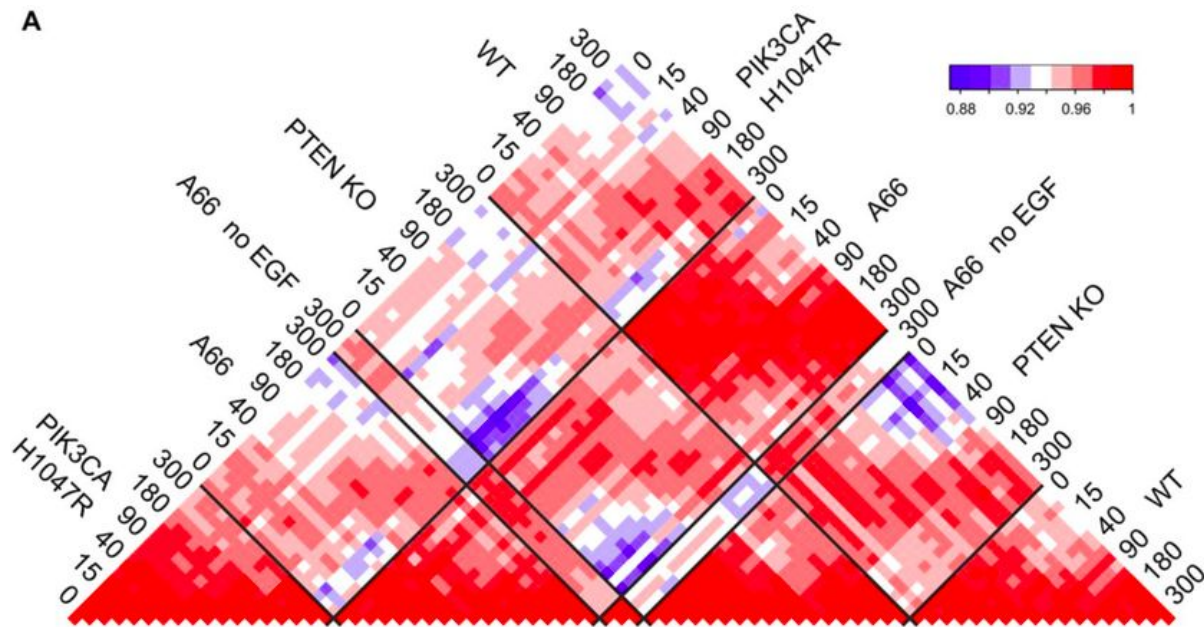
TP53 scaled expression distribution



Z-скорирование

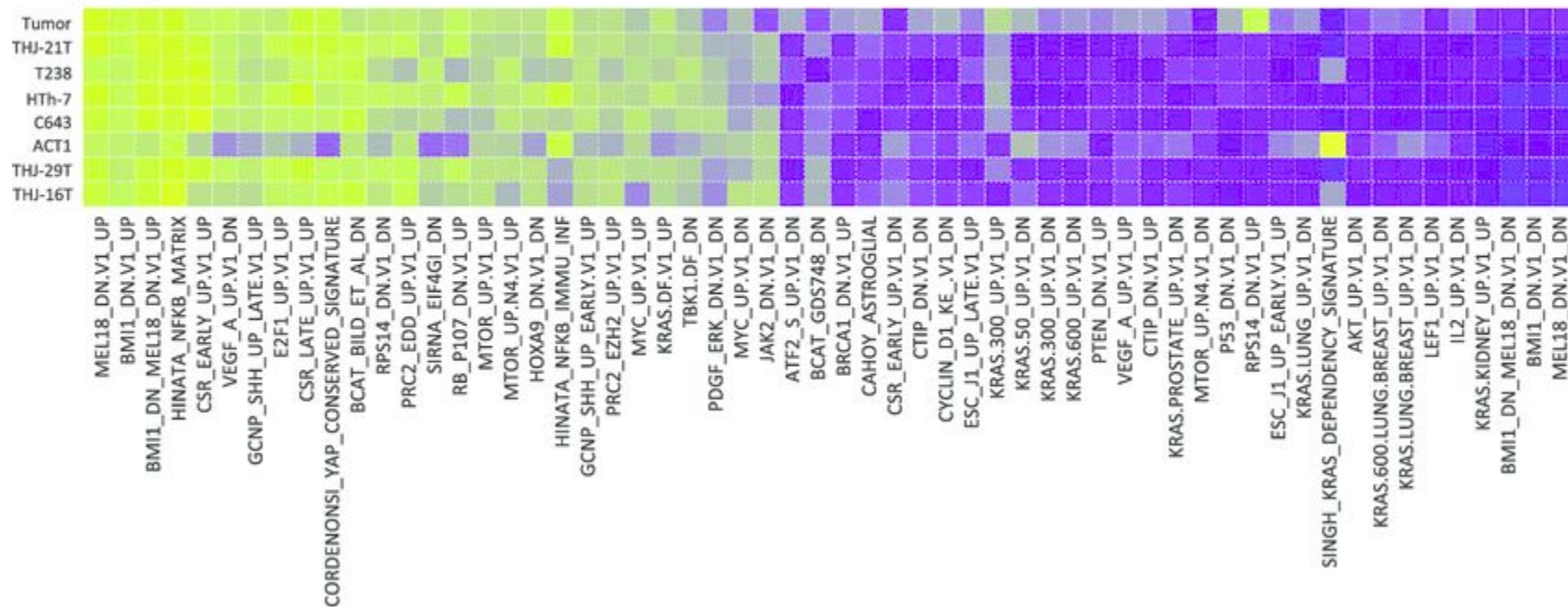


- Для улучшения кластеризации в качестве метрики можно использовать корреляцию между образцами



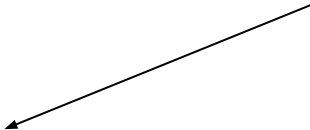
Genes up-regulated in DAOY cells (medulloblastoma) upon knockdown of BMI1 or PCGF2 or both

Genes down-regulated in DAOY cells (medulloblastoma) upon knockdown of BMI1 or PCGF2 or both



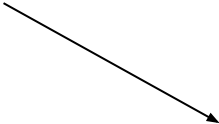
Enriched in ATC overexpressed genes Enriched in ATC underexpressed genes

Деконволюция bulk RNA-Seq проб — это процесс определения того, в каких долях какие клеточные типы содержатся в пробе.

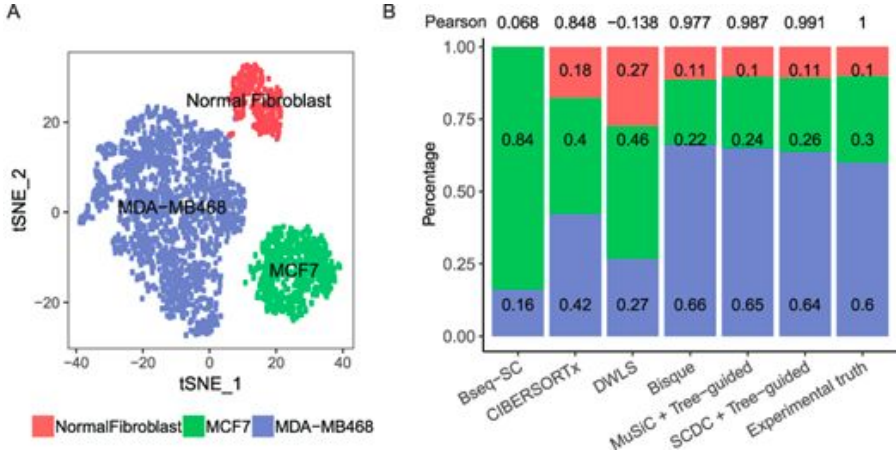


Основанная на маркерных генах
(*BisqueRNA*)

gene	cluster	avg_logFC
Gene 1	Neurons	0.82
Gene 2	Neurons	0.59
Gene 3	Astrocytes	0.68
Gene 4	Oligodendrocytes	0.66
Gene 5	Microglia	0.71
Gene 6	Endothelial Cells	0.62



Основанная на референсе
(*SCDC*, *BisqueRNA*).



Контакты

Оля — вопросы по биоинформатике & Feedback по курсу:



@olya_kudryashova



olga.kudryashova@bostongene.com

Серёжа — вопросы насчёт блока по транскриптомике:



@sergisa



sergei.isaev@bostongene.com

Телеграмм-группа курса иммунологии и биоинформатики от BostonGene:



<https://t.me/joinchat/B1VA6B1Qe1zGiBeZuTC2vQ>

Катя Титова (HR) — вопросы о стажировке в BostonGene летом 2021:



vk.com/titovakate



ekaterina.titova@bostongene.com