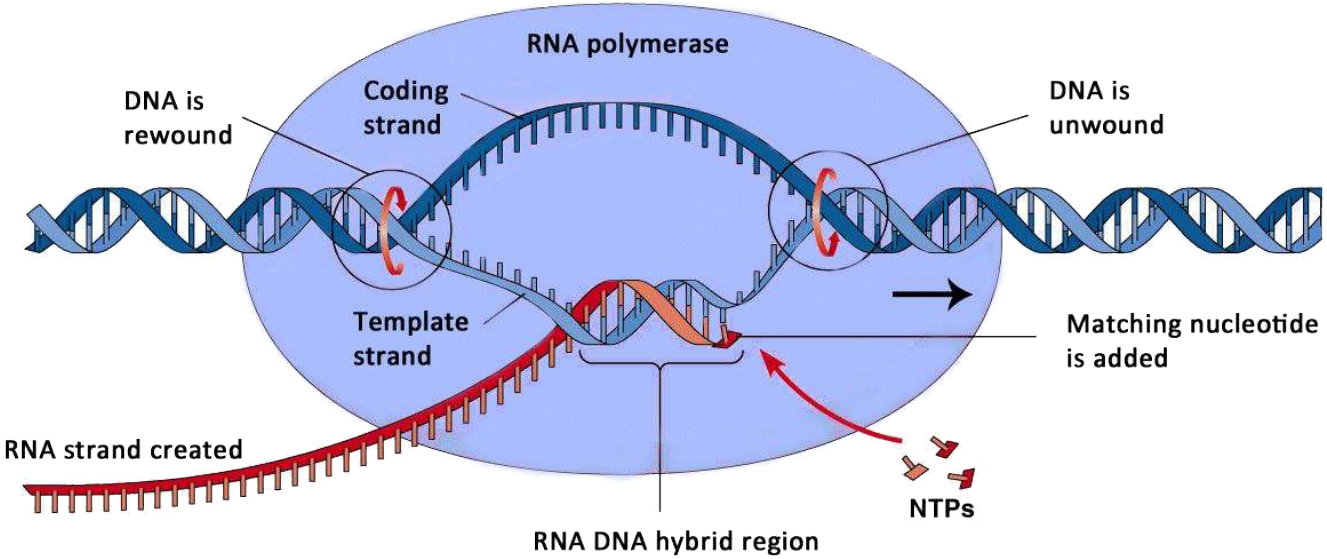


Лекция 2
Транскриптомика:

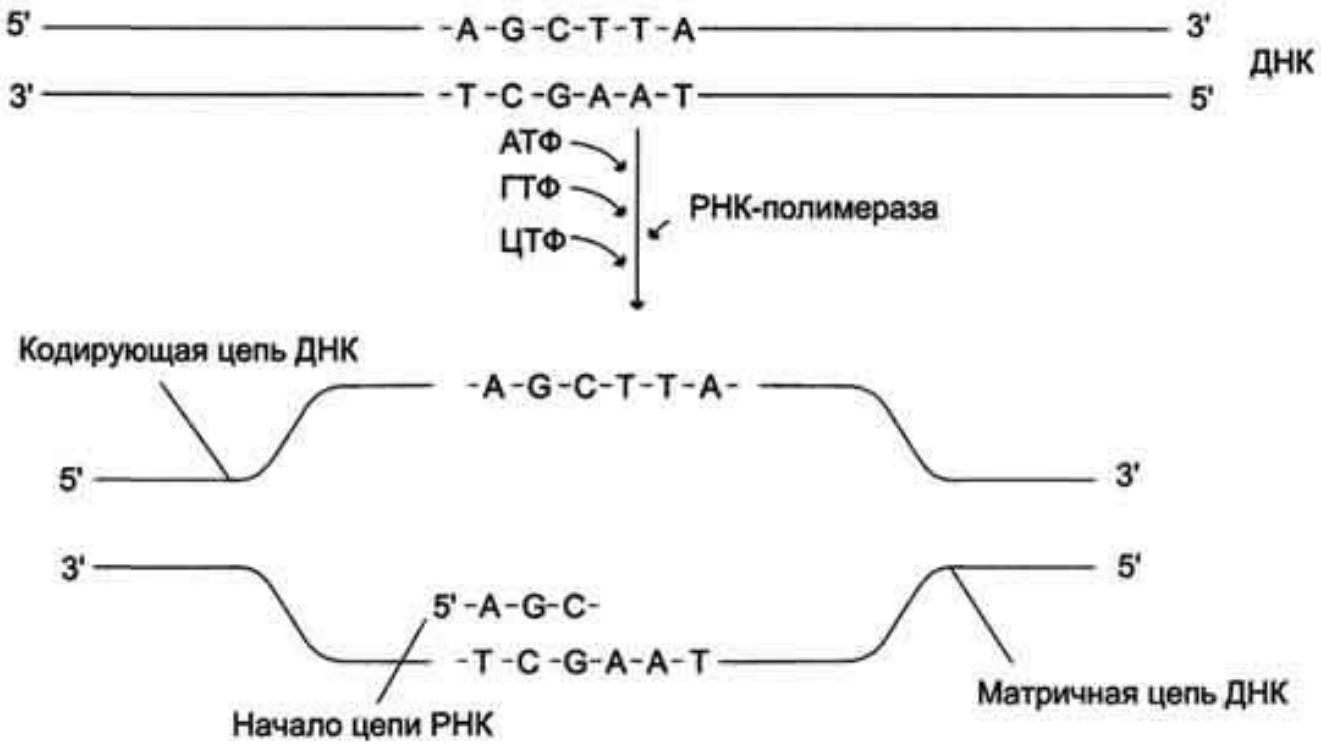
практические методы и применяемые алгоритмы

Александр Предеус
Институт Биоинформатики

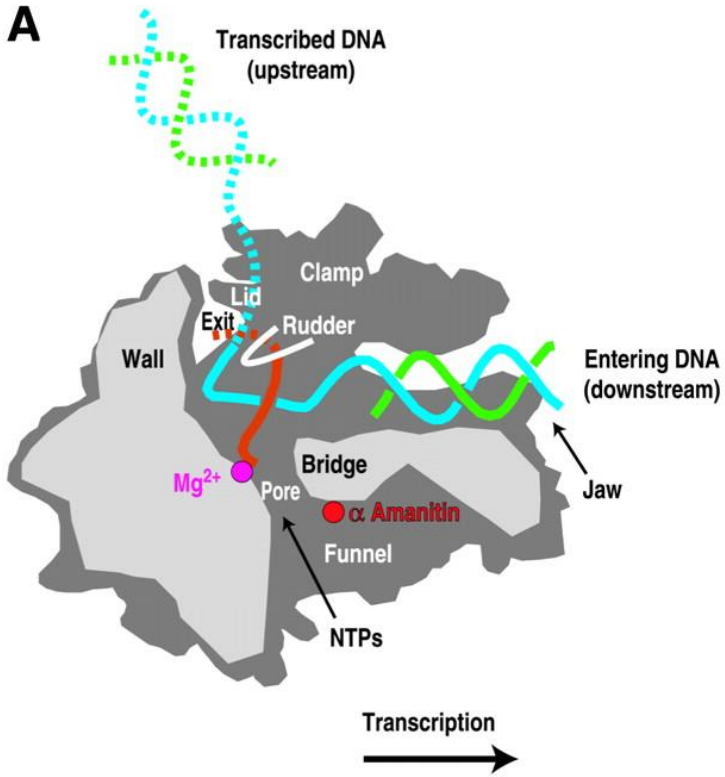
Краткое содержание

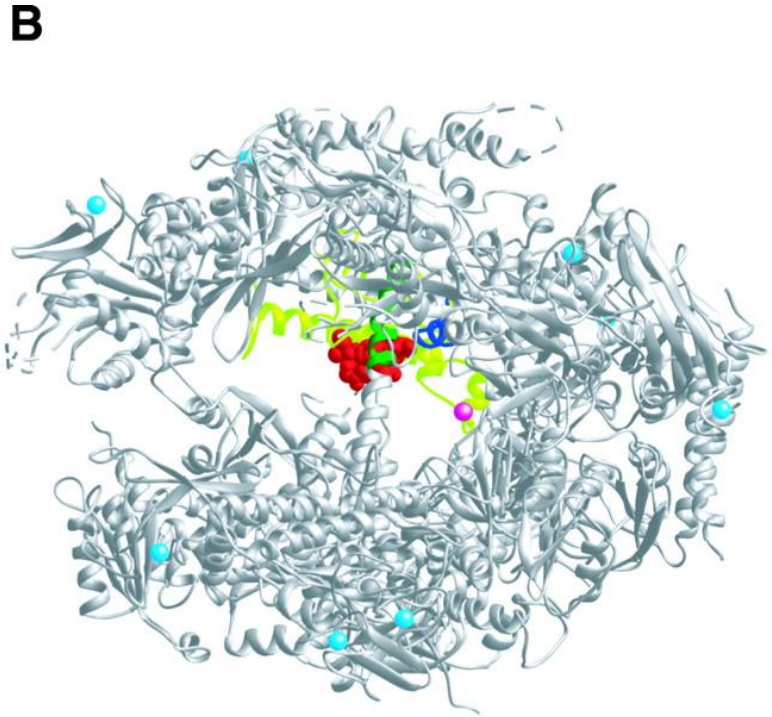
- Микрочипы
  - Типы
  - Обработка
  - Ограничения
- РНК-сек
  - Особенности
  - Выравнивание

Экспрессия гена = транскрипция



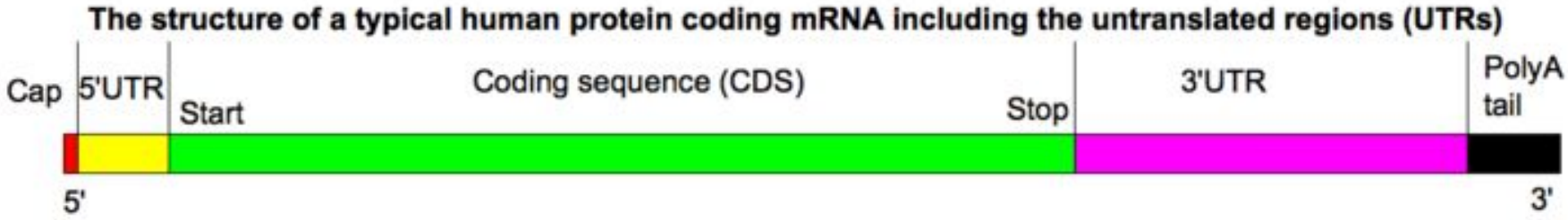
Экспрессия гена = транскрипция





мРНК

- Кодирует протеины
- 0 .. ~300000 копий каждого типа на клетку *H.s.*



Сколько генов у человека?

Сколько генов у человека?

- Протеин-кодирующих – 19797 (Gencode v23)
- Вообще – 60484
  - Псевдогены
  - Микро-РНК
  - Длинные некодирующие РНК
  - Иммунные гены

Аннотации

- Для млекопитающих: RefSeq, ENSEMBL, GENCODE

Frankish *et al*. *BMC Genomics* 2015, **16**(Suppl 8):S2
http://www.biomedcentral.com/1471-2164/16/S8/S2

BMC Genomics

RESEARCH Open Access

Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction

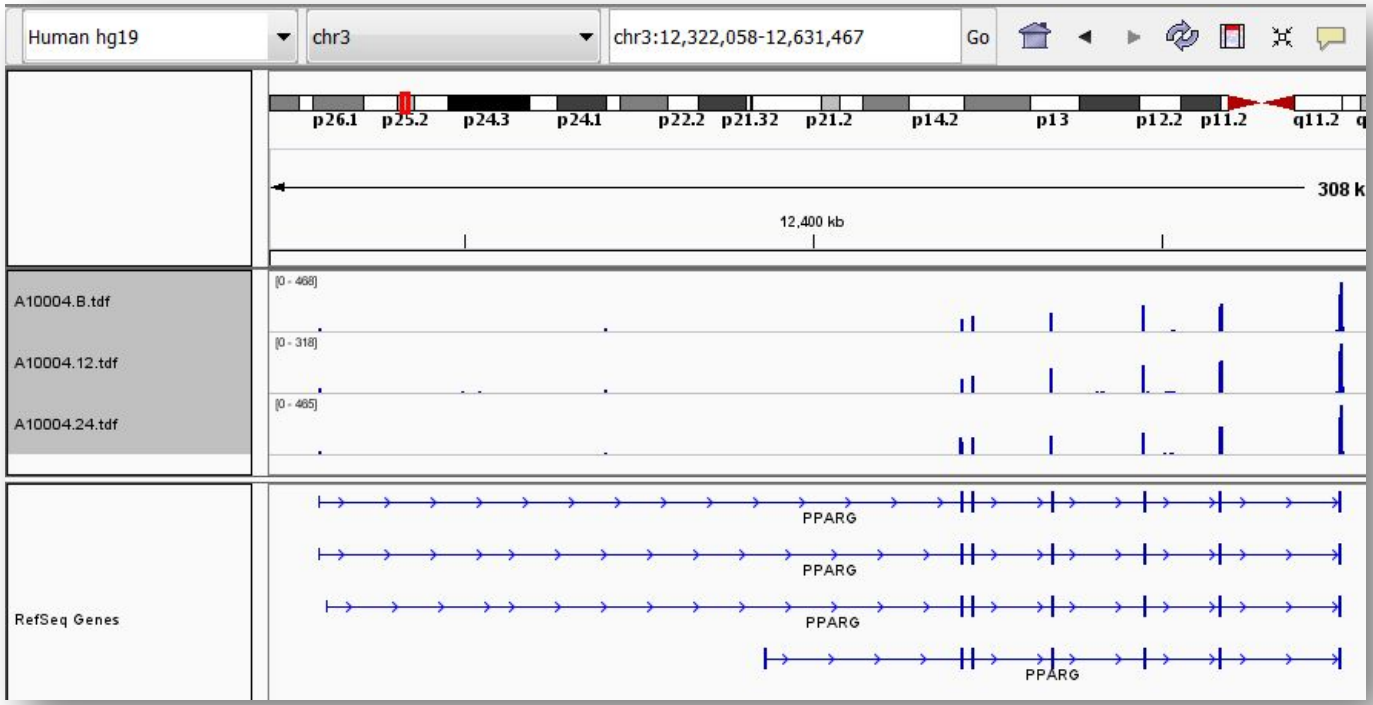
Adam Frankish[1*], Barbara Uszczynska[2], Graham RS Ritchie[1,3], Jose M Gonzalez[1], Dmitri Pervouchine[2,4], Robert Petryszak[3], Jonathan M Mudge[1], Nuno Fonseca[3], Alvis Brazma[3], Roderic Guigo[2], Jennifer Harrow[1*]

From VarI-SIG 2014: Identification and annotation of genetic variants in the context of structure, function and disease
Boston, MA, USA. 12 July 2014

GENCODE vs. RefSeq:
http://www.ncbi.nlm.nih.gov/pubmed/26110515

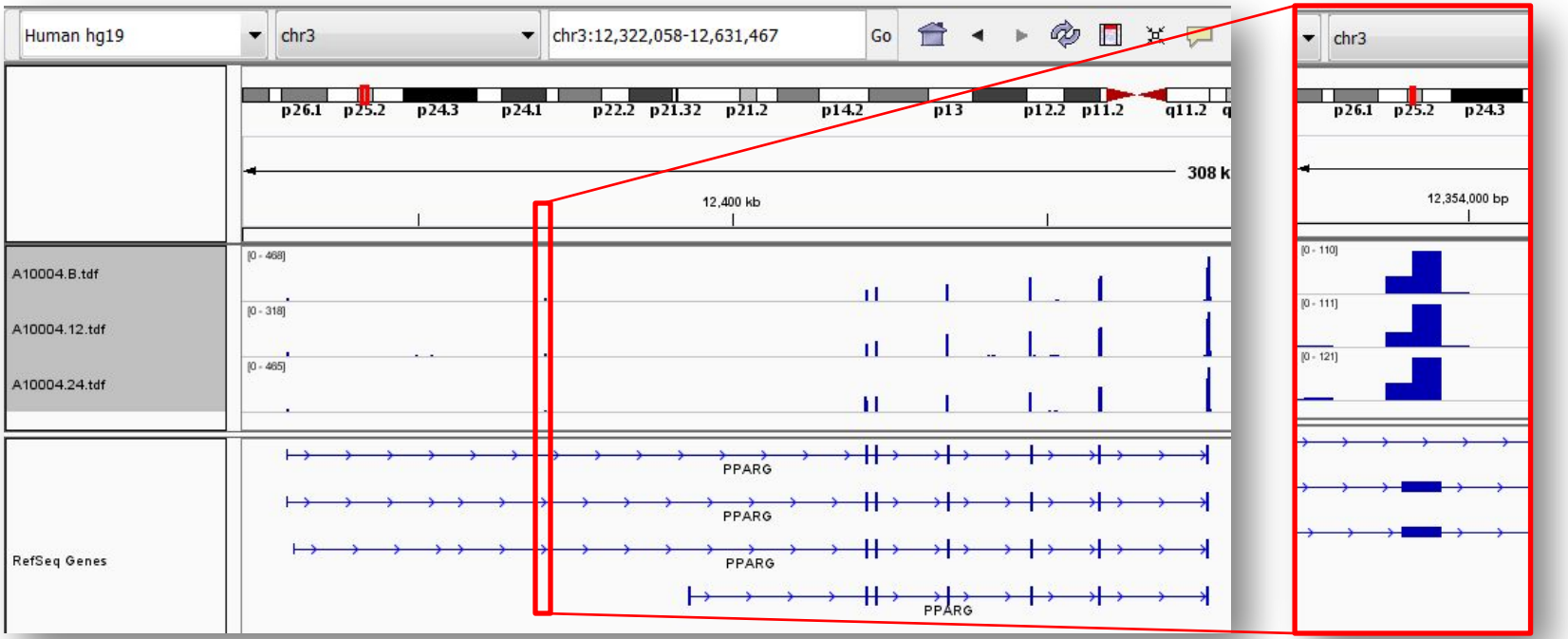
Ген = несколько транскриптов

- Причина - сплайсинг в генах эукариот



Ген = несколько транскриптов

- PPARg изоформы – разной адипогенности



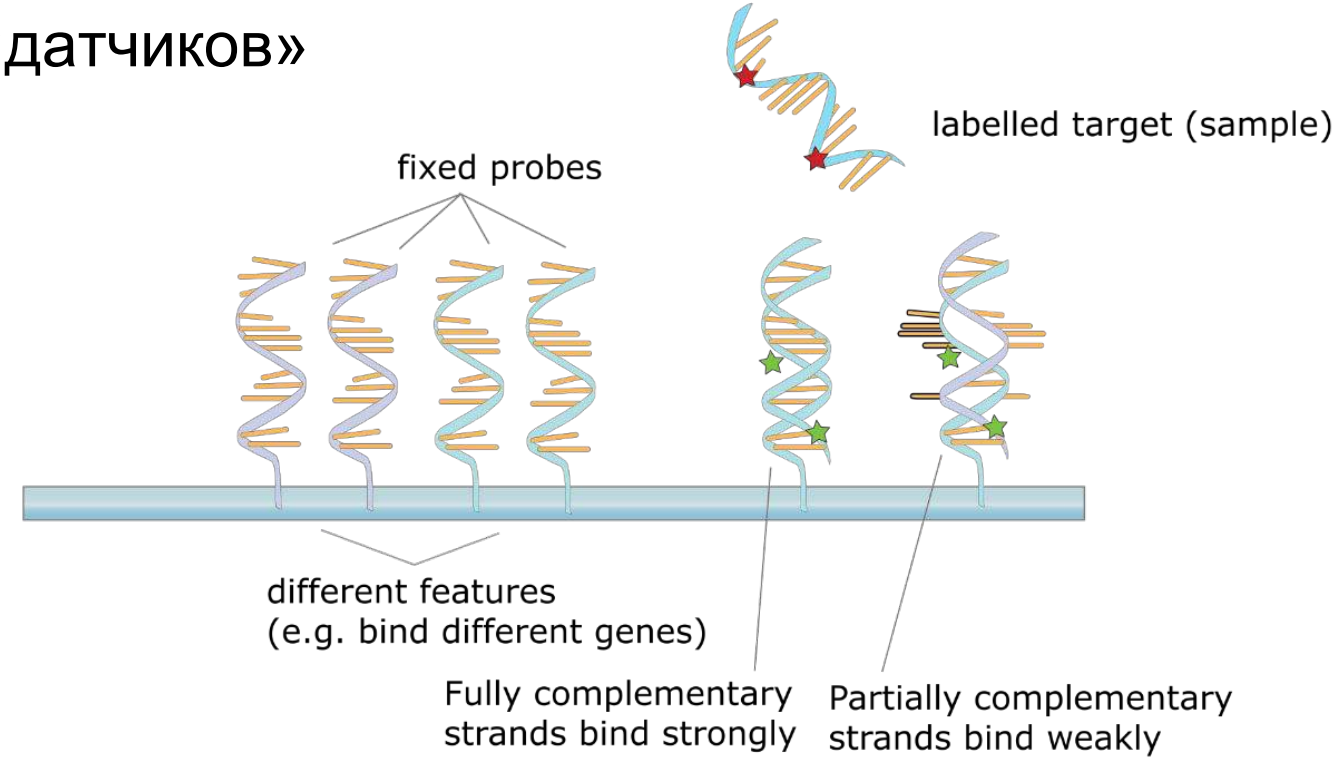
Уровни мРНК и протеинов

- Значительная, но не идеальная корреляция

Organism	r_p	r_s	Data-set size	Reference
Saccharomyces cerevisiae	0.36	n.d.	73	[40]
Saccharomyces cerevisiae	0.76	0.74	148	[39]
Mus musculus	0.59	n.d.	425	[46]
Saccharomyces cerevisiae	n.d.	0.45	678	[43]
Desulfovibrio vulgaris	0.50	n.d.	703	[45]
Escherichia coli	0.57	0.50	1103	[32]
Schizosaccharomyces pombe	0.58	0.61	1367	[44]
Saccharomyces cerevisiae	0.66	n.d.	2044	[41]
Saccharomyces cerevisiae	n.d.	0.57	4251	[38]

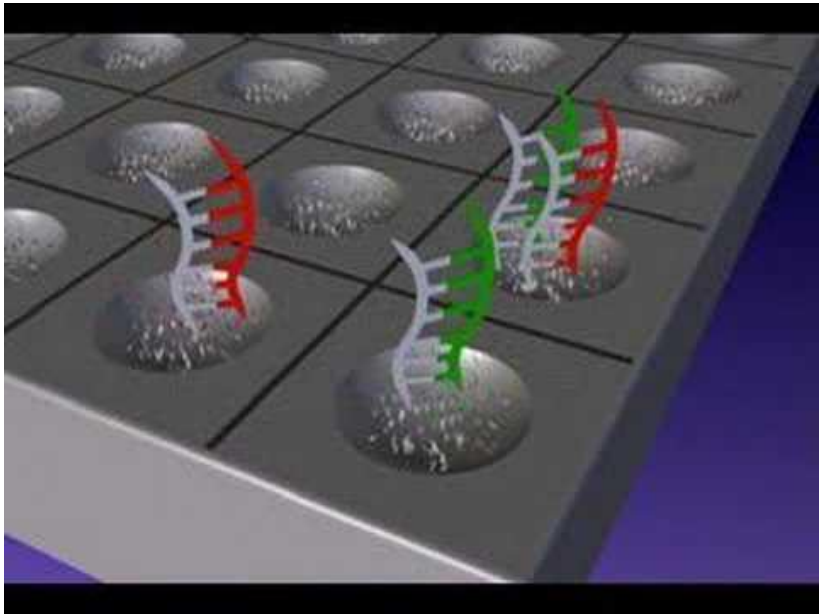
Микрочип экспрессии

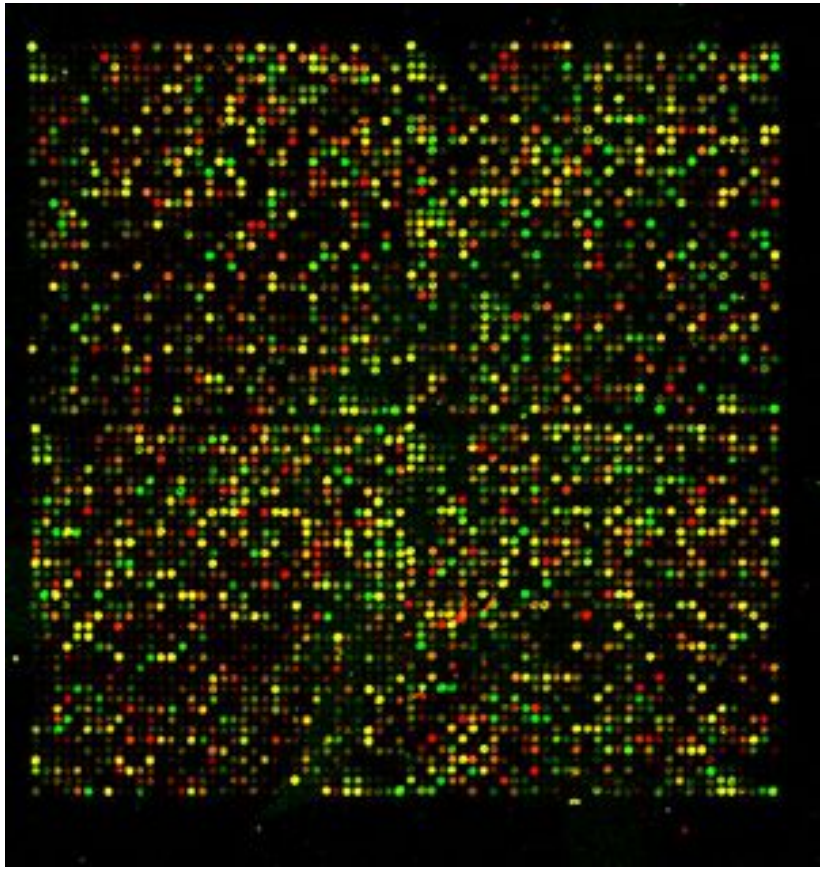
- 10-50000 «датчиков» (probes)



Микрочип

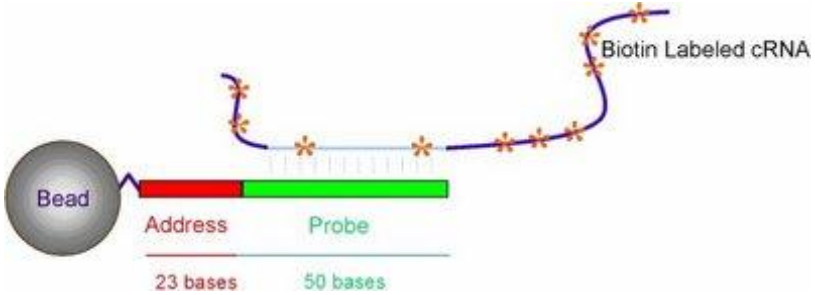
- Измеряется флуоресценция

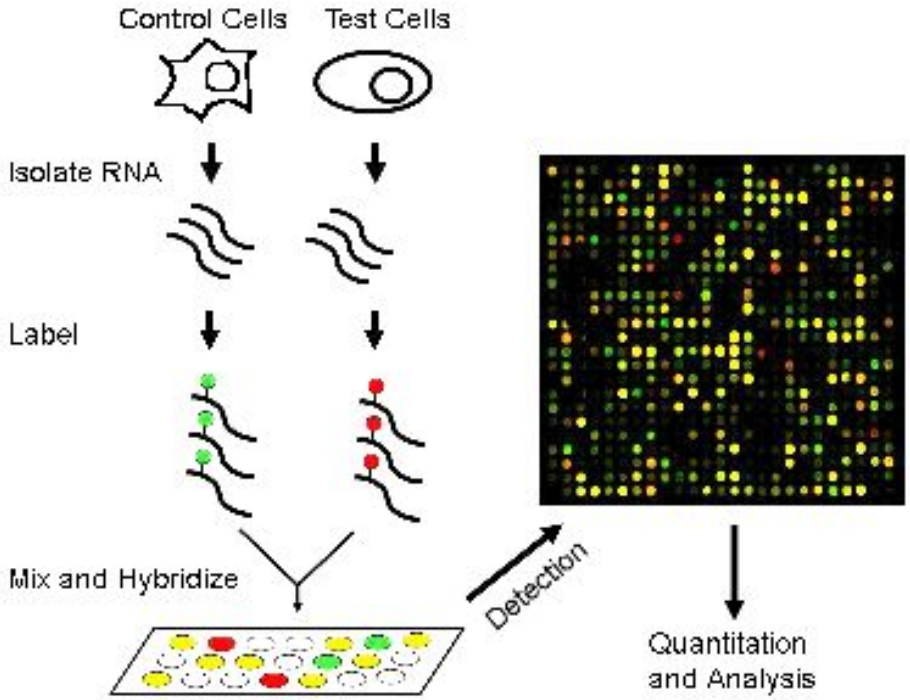




Производители

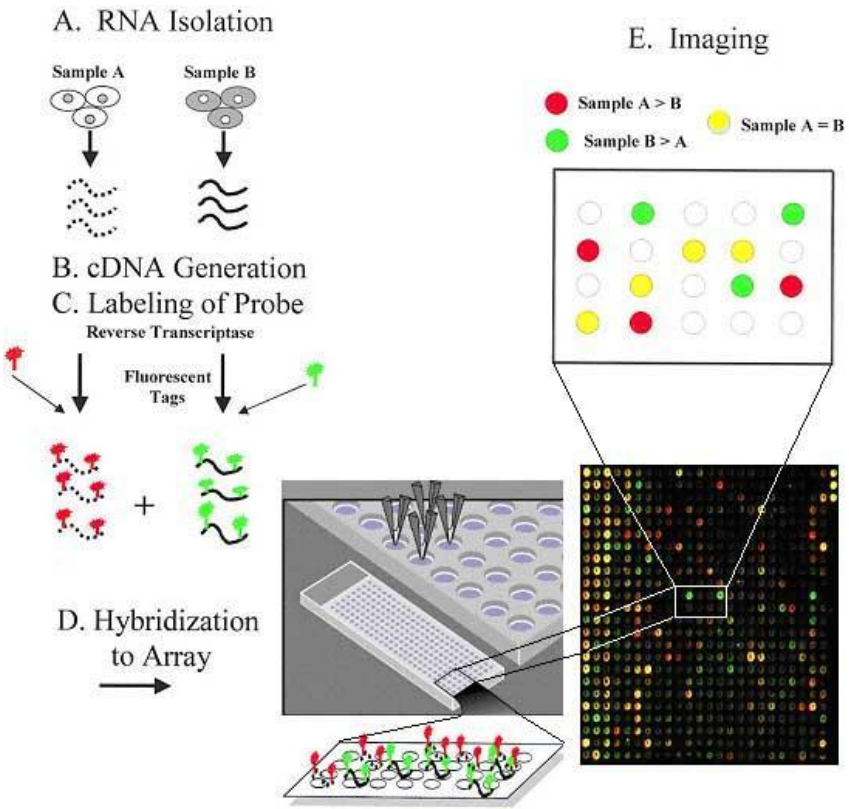
- Affymetrix (75% для мыши и человека)
- Agilent (1-color & 2-color)
- Illumina - beads





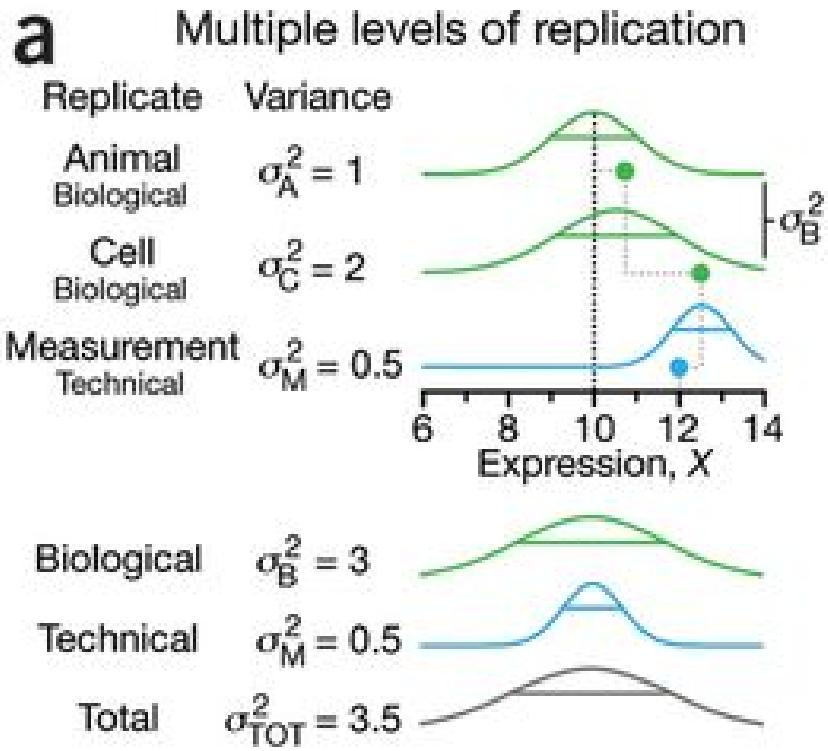
Дифференциальная экспрессия

- Несколько образцов + несколько контролей



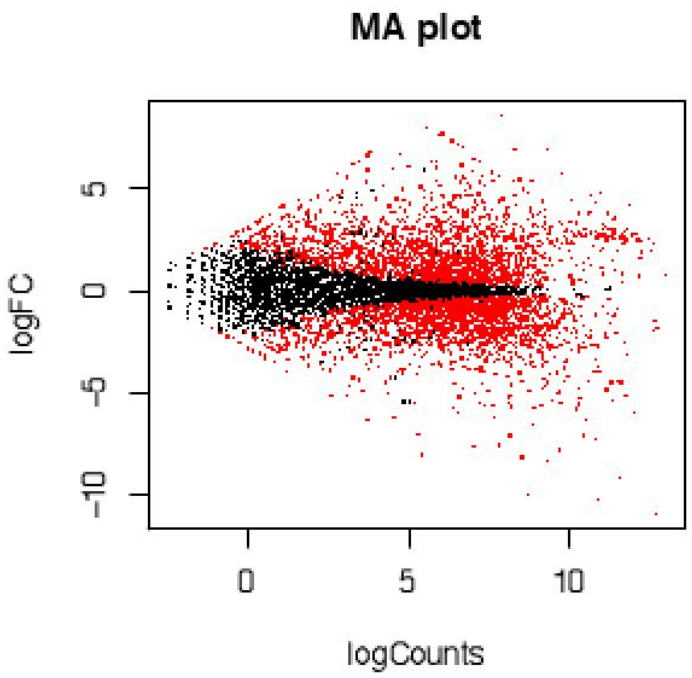
Вариация экспрессии

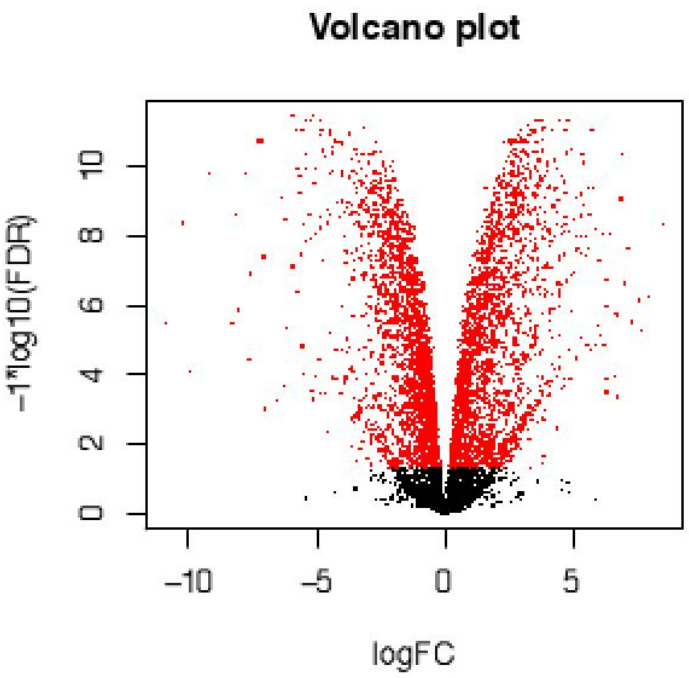
- Разделить техническую и био-вариацию!
- Считается что соотношение био/техн вариации ~ 2:1



Volcano & MA plots

- logFC – всегда log2



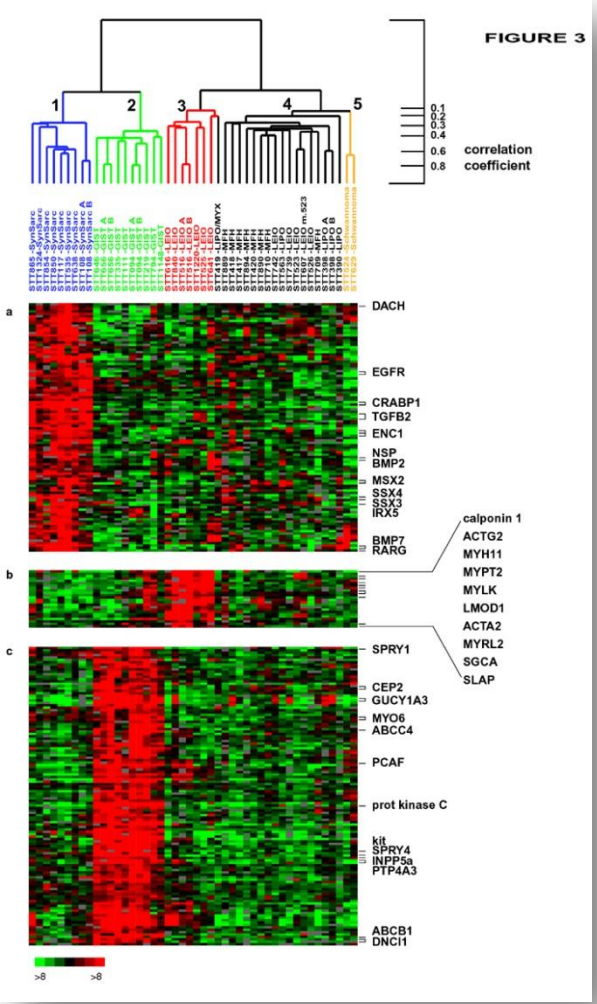


Множественные сравнения

- Family-wise Error Rate (FWER)
  - Для ошибки в любом из измерений
  - Пример - Бонферрони
    - Делим p-value на число датчиков
    - На 12,000 генов, p-value 0.000004 дает 5% FP
- False Discovery Rate (FDR)
  - Отношение FP/(FP+TP)
  - Более мягкий критерий – 5 ошибок на 100 измерений

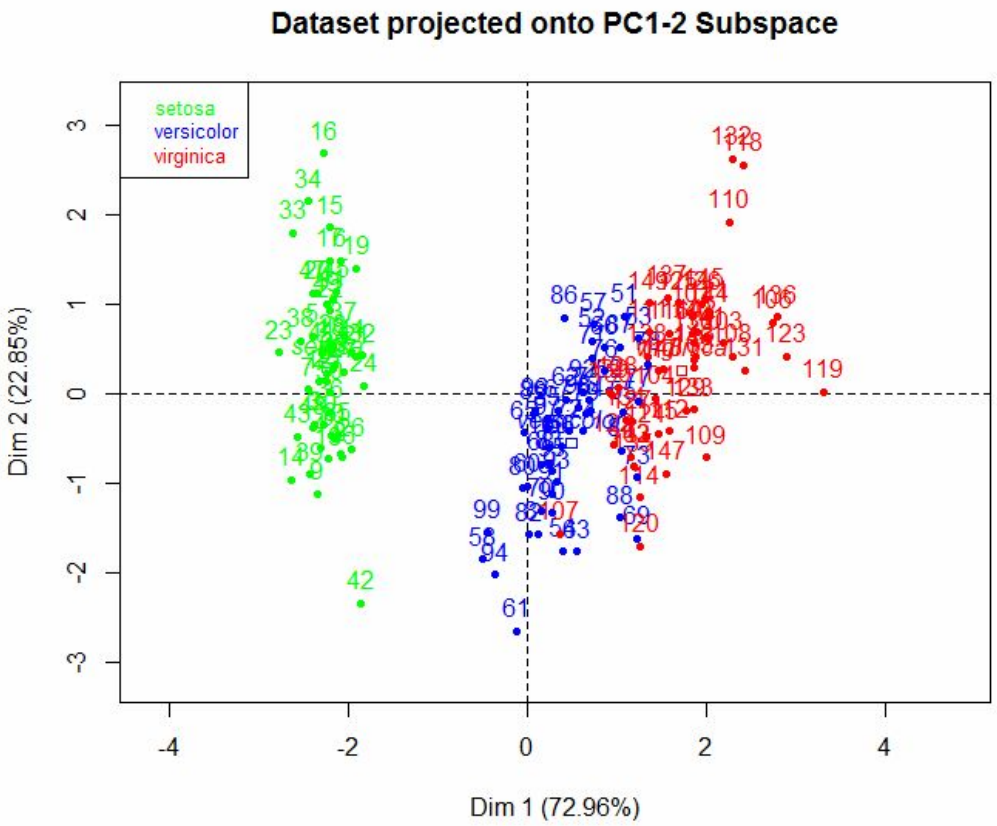
Кластеризация

- По образцам - чтобы увидеть структуру данных
- Метрика - Евклидово расстояние



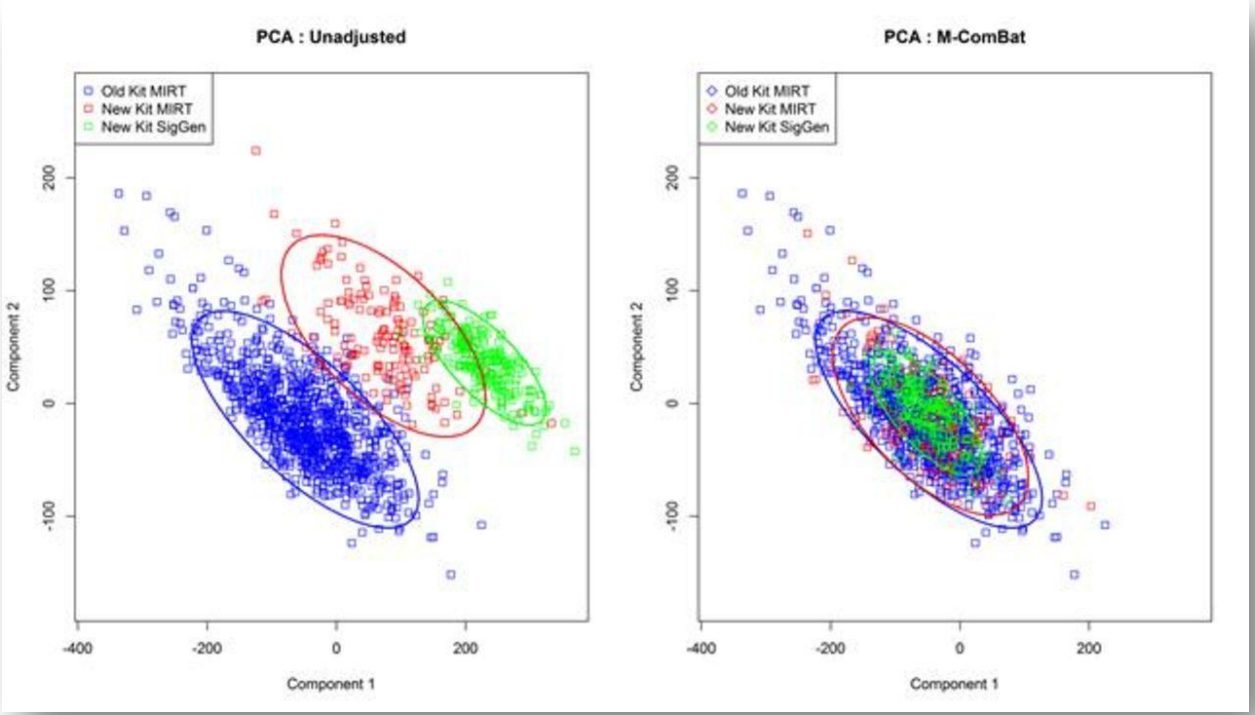
Что такое PCA?

- PCA, как и кластеризация - метод уменьшения размерности
- PCA показывает направления максимальной вариации данных



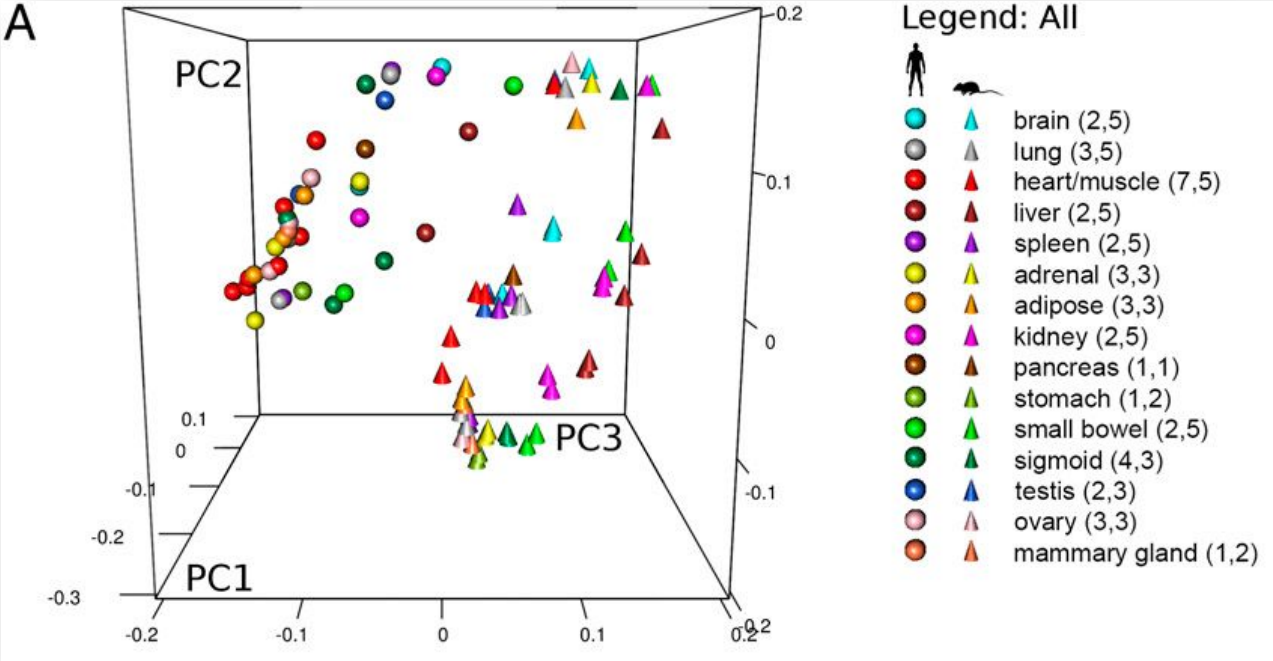
Зачем это нужно?

- Эффект донора (или batch-effect) легко может сменить интерпретацию эксперимента на прямо противоположную

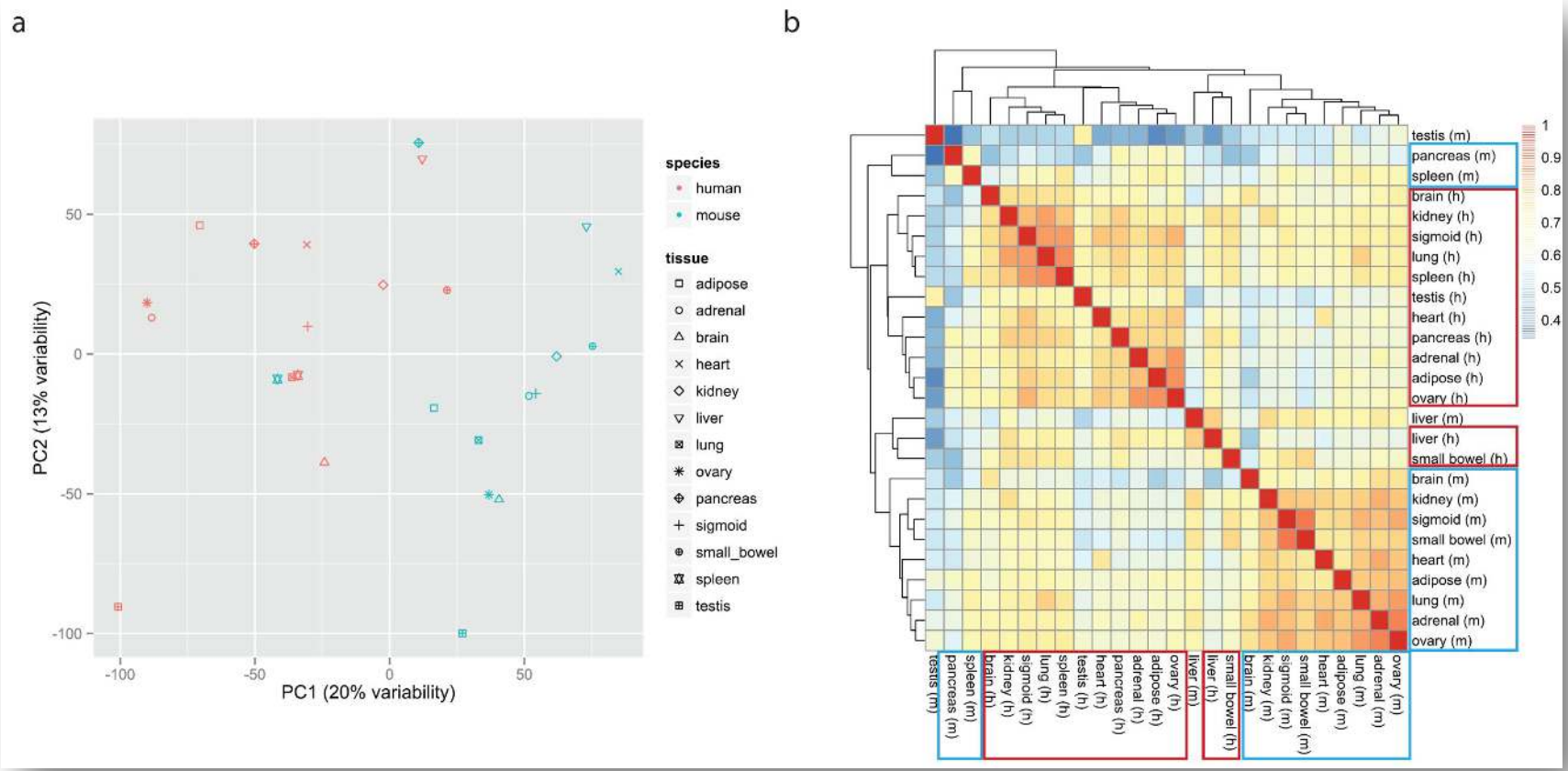


mouseENCODE и batch-effect

- мышь-человек
- РНК-сек на 15 тканях
- результат - сильные отличия!
- мышь - плохая модель?!



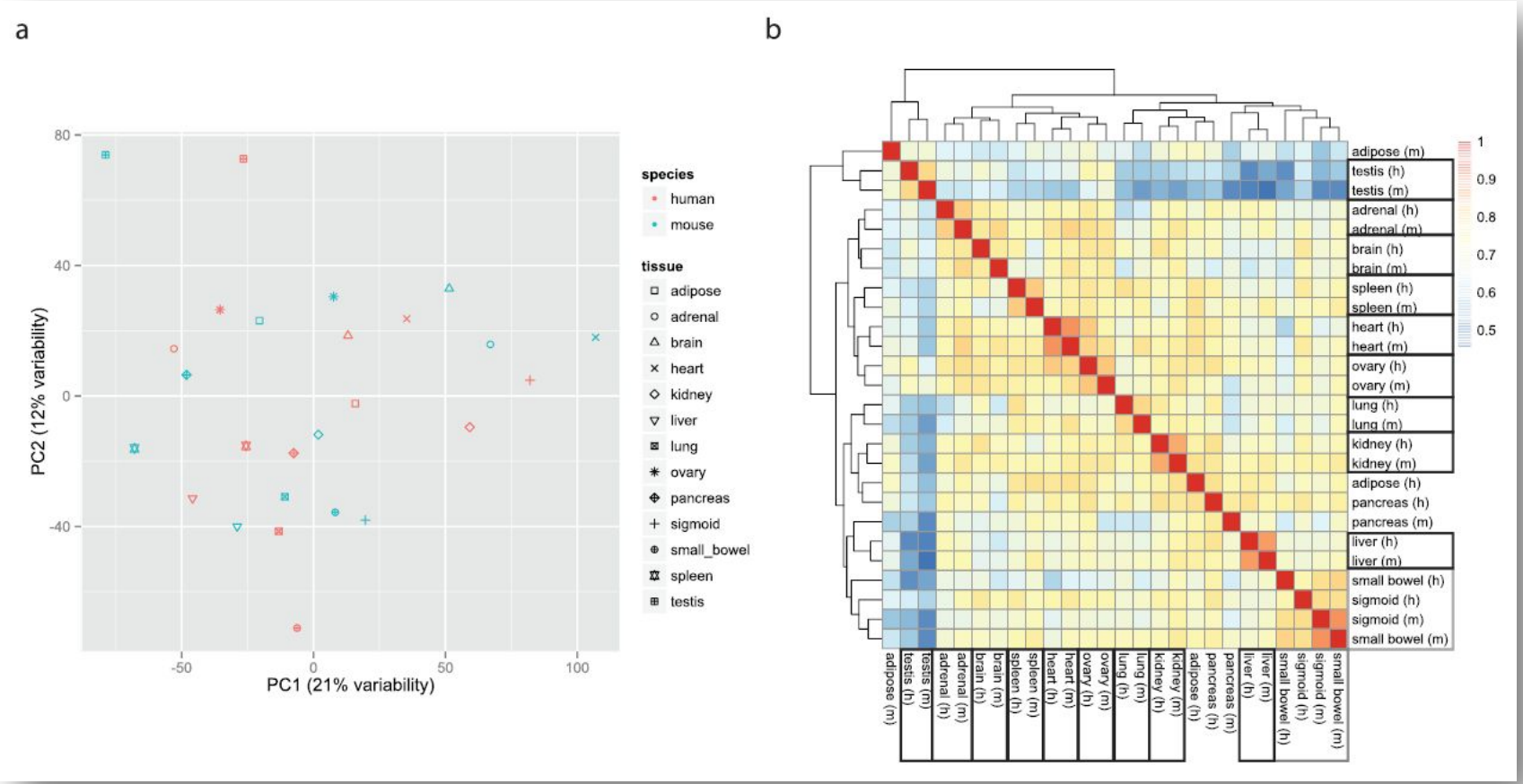
Исходная картина



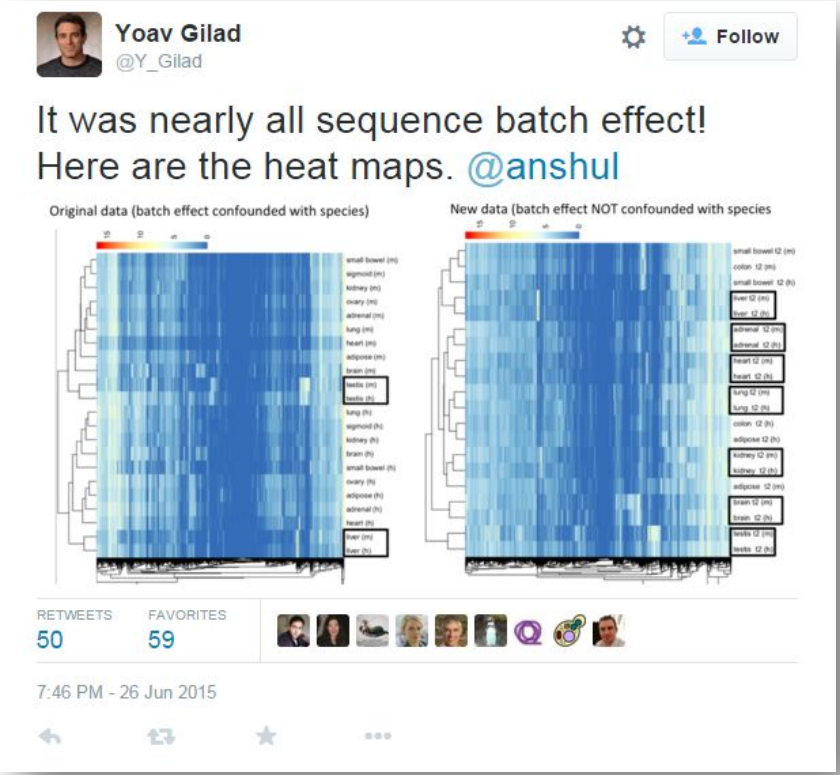
В чем проблема?

- Мышиные данные делались одной группой (или группами), а человеческие - другой
- Весь набор ковариат отличается в мышиных/человеческих данных
- Gilad group решила посмотреть, что будет, если убрать этот эффект (sva R package, ComBat)

Картина после коррекции



Fail of the year award?



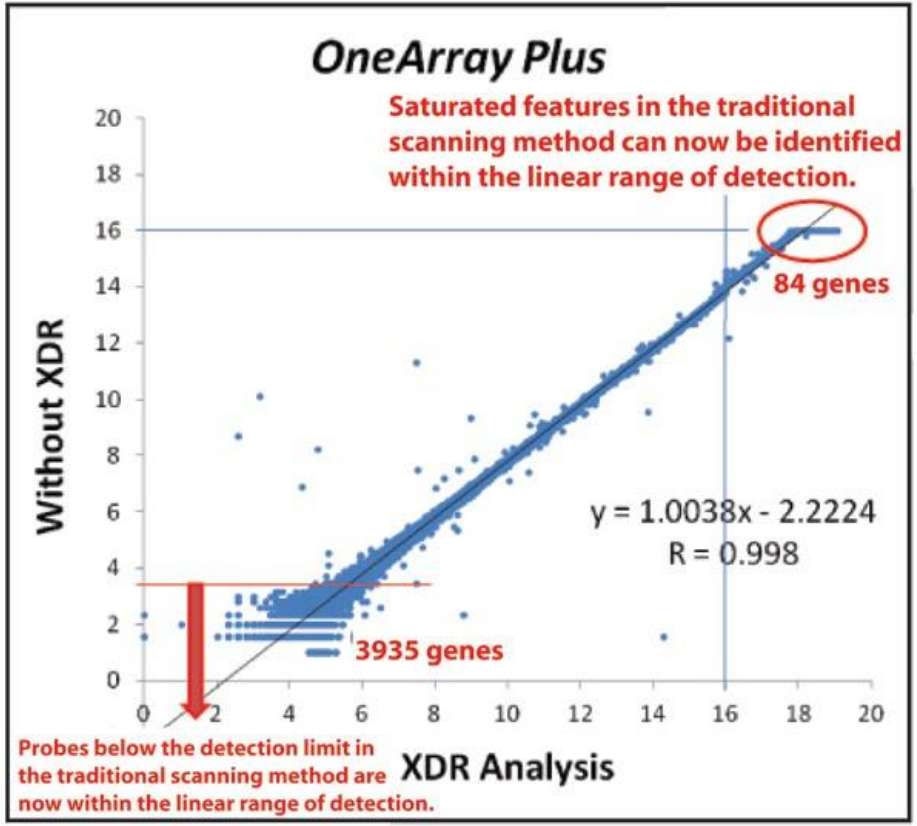
https://twitter.com/Y_Gilad

sva

- Прямое моделирование для известных ковариатов (пол, лаборатория, генотип, итд)
- Непрямое моделирование – «the truth is out there»
- comBat – удалить эффект одного или нескольких ковариатов

Ограничения микрочипов

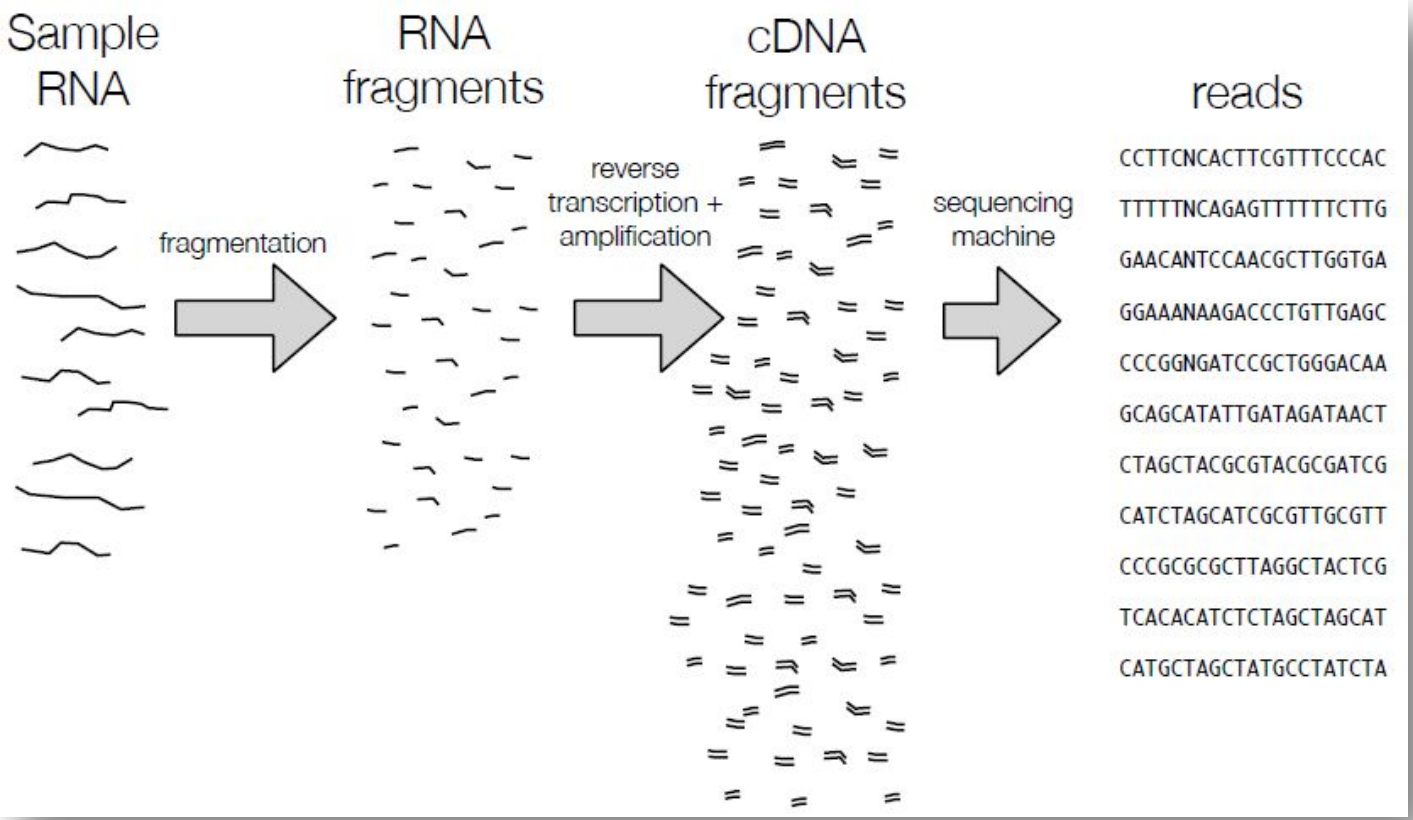
- Шум – только ~топ 6-8к генов удается померить надежно
- Динамический диапазон – неверно оценивает самые сильные гены



Краткое содержание

- Микрочипы
  - Типы
  - Обработка
  - Ограничения
- РНК-сек
  - Особенности
  - Выравнивание

Что такое РНК-сек



Квантификация РНК-сек

- Два способа:
  - На транскриптом (десятки-сотни тысяч «хромосом»)
  - На аннотированный геном (десятки хромосом)
- И там и там надо выравнивание
- Выровнять на геном - труднее

Выравнивание

- Концепция
- Модель
- Алгоритм
- Имплементация

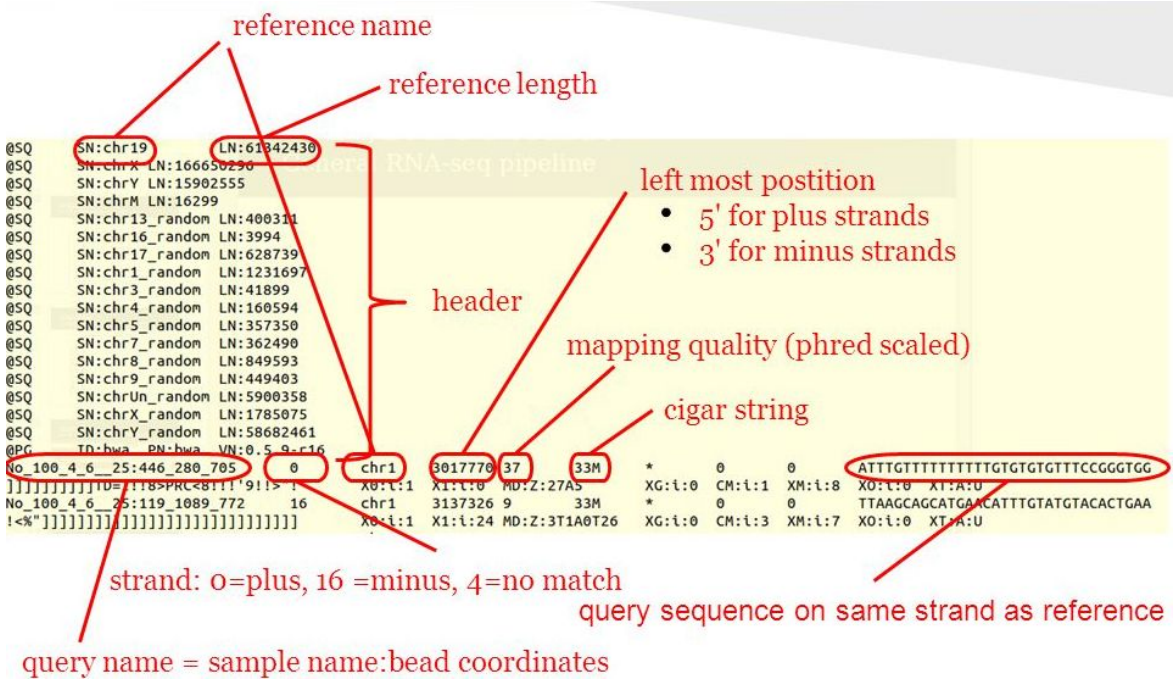
```
CTTTATAGAGCATAAGCAGCAGCGCAACACCCTTAT mutation

CTTTATAGAGCATA---AGCAGCGCAACACCCTTAT reference

CTTTATAGAGCATA---A                   read 1 →  no end trimming
CTTTATAGAGCATA---AGCAG               read 2 →
      AGAGCATAAGCAGCAGCGCAA          read 3 →
          CATAAGCAGCAGCGCAACACCCTTAT read 4 ←
                 AGCAGCGCAACACCCTTAT read 5 ←
                    AGCGCAACACCCTTAT read 6 ←

CTTTATAGAGCATaa                      read 1 → with end trimming
CTTTATAGAGCATAagcag                  read 2 →
      agagCATAAGCAGCAGCGCaa          read 3 →
          cATAAGCAGCAGCGCAACACCCTTAT read 4 ←
                 agcagcGCAACACCCTTAT read 5 ←
                    agcGCAACACCCTTAT read 6 ←
```

Форматы SAM/BAM

Col	Field	Description
1	QNAME	Query (pair) NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIAGR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe ('=' if same as RNAME)
8	MPOS	1-based Mate POSistion
9	ISIZE	Inferred insert SIZE
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

SAM format: FLAG field

numeric	binary	description
1	00000001	template has multiple fragments in sequencing
2	00000010	each fragment properly mapped according to aligner
4	00000100	fragment is unmapped
8	00001000	mate is unmapped
16	00010000	sequence is reverse complemented
32	00100000	sequence of mate is reversed
64	01000000	is first fragment in template
128	10000000	is second fragment in template



Строка CIGAR

- Строка CIGAR позволяет воспроизвести выравнивание

```
Coor      12345678901234   5678901234567890123456789012345
Ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001/1          TTAGATAAAGGATA*CTG
r002          aaaAGATAA*GGATA
r003       gcctaAGCTAA
r004                       ATAGCT..............TCAGC
r003                              ttagctTAGGC
r001/2                                            CAGCGGCAT
```

```
@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       *
r004  0     ref 16  30  6M14N5M     *  0   0   ATAGCTTCAGC       *
r003  2064  ref 29  17  6H5M        *  0   0   TAGGC             *
r001  147   ref 37  30  9M          = 7  -39   CAGCGGCAT         *  NM:i:1;
```

Качество выравнивания

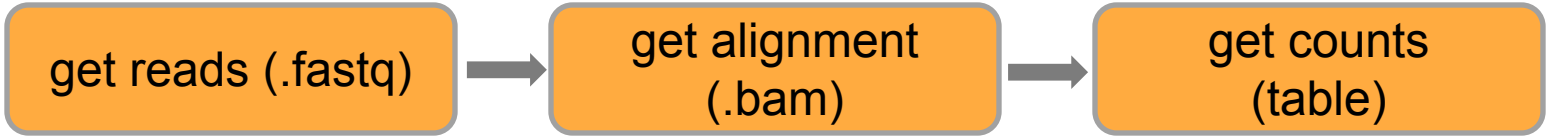
- Оценивает шанс некорректного выравнивания, $-10\log_{10}(p)$
- Уникальное выравнивание - максимум MapQ
- Повторы: MapQ ~ 0

Качество против скорости

- Точная оценка mapq требует вычислительных затрат
- Хорошие алгоритмы – решают
- Возможно полностью точное выравнивание (e.g. RazerS3), хотя оно и очень медленно

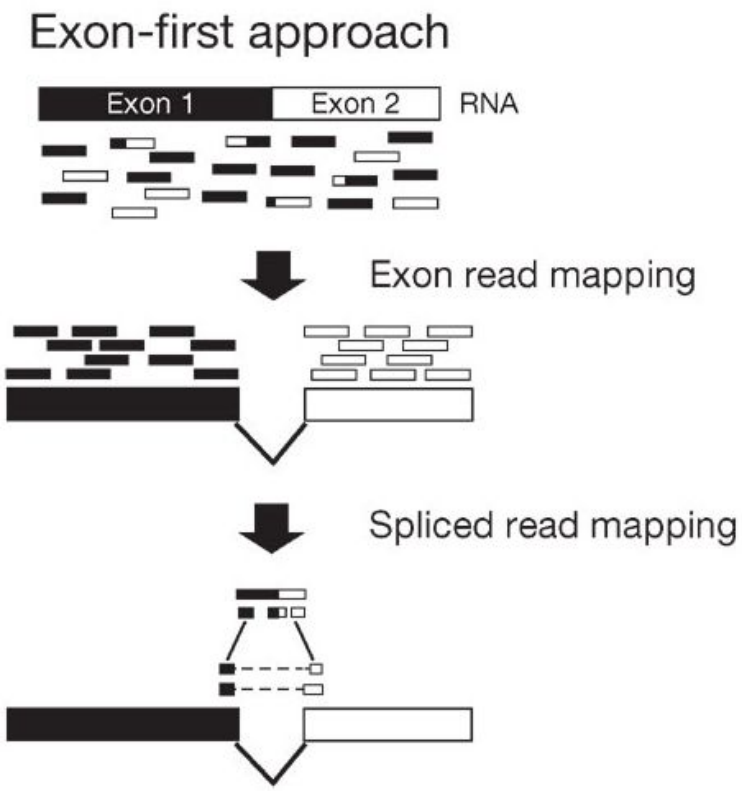
Квантификация РНК-сек

- Два способа:
  - На транскриптом
  - На аннотированный геном
- И там и там надо выравнивание
- Выровнять на геном - труднее



Выравнивание транскриптома

- На геном – трудно: большие «инделы» = сплайсинг



Выравнивание транскриптома

Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 9061–9066, August 1996
Genetics

Gene recognition via spliced sequence alignment

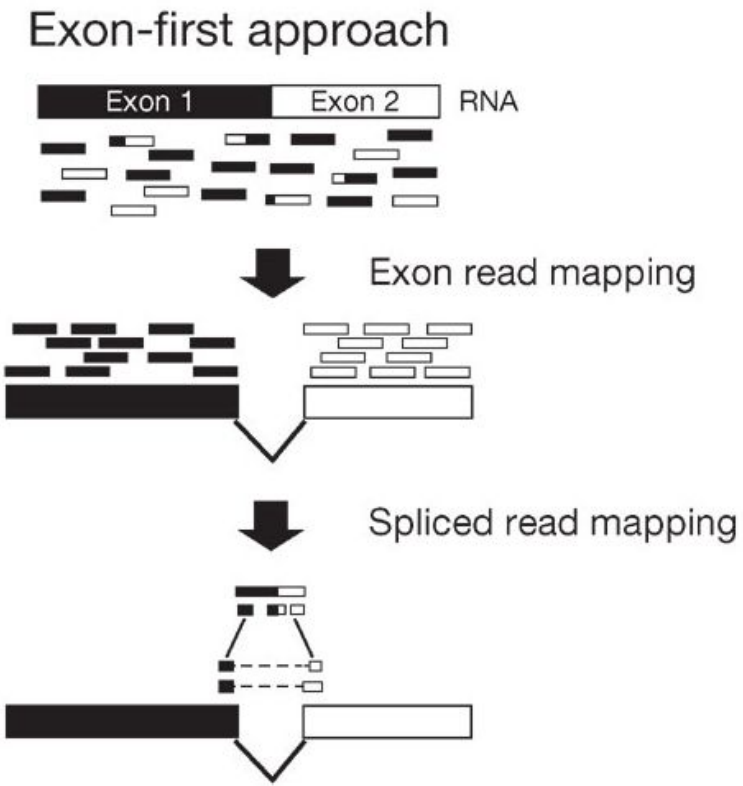
MIKHAIL S. GELFAND[†], ANDREY A. MIRONOV[‡], AND PAVEL A. PEVZNER[§¶]

[†]Institute of Protein Research, Russian Academy of Sciences, Puschino, Moscow, 142292, Russia; [‡]Laboratory of Mathematical Methods, National Center for Biotechnology NIIGENETIKA, Moscow, 113545, Russia; and [§]Departments of Mathematics and Computer Science, University of Southern California, Los Angeles, CA 90089-1113

Communicated by Charles R. Cantor, Boston University, Boston, MA, April 19, 1996 (received for review January 15, 1996)

Геномное выравнивание транскриптов

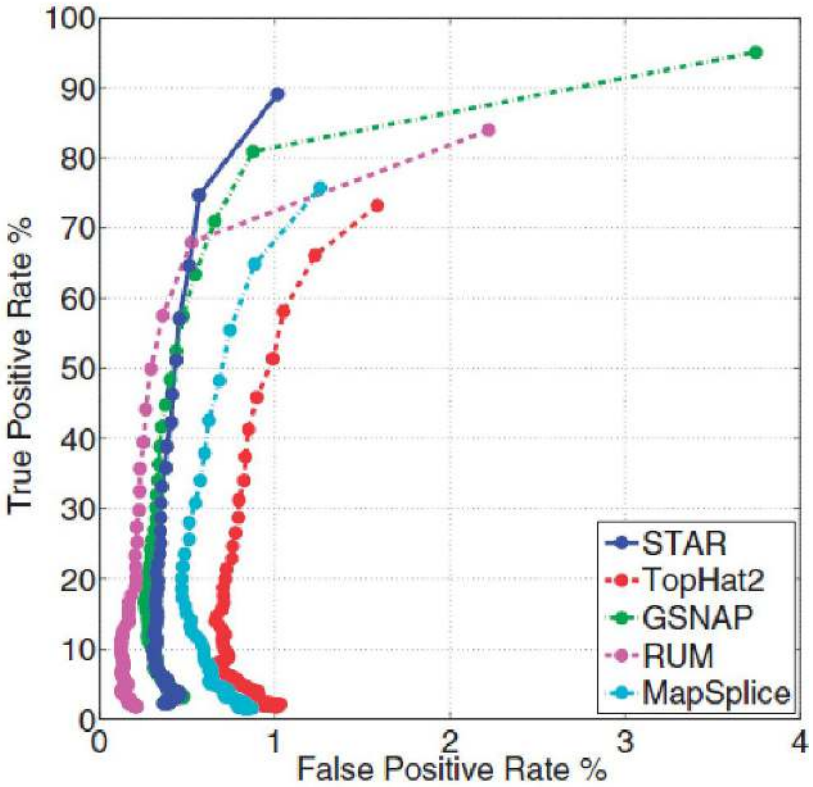
- 10-20% прочтений - сплайс
- Tophat - плохо
- STAR - хорошо
- hisat2 - тоже хорошо
- hera - последние новости



STAR (rna-star)

- Создатель - Александр Добин (CSHL)
- Очень быстрый (сотни М ридов в час)
- Высокая чувствительность
- Требования к памяти – 32 Гб
- Весь ENCODE за день на 32 CPU кластере (!)

Чувствительность и скорость



Aligner	Mapping speed: million read pairs/hour		Peak physical RAM, GB	
	6 threads	12 threads	6 threads	12 threads
STAR	309.2	549.9	27.0	28.4
STAR sparse	227.6	423.1	15.6	16.0
TopHat2	8.0	10.1	4.1	11.3
RUM	5.1	7.6	26.9	53.8
MapSplice	3.0	3.1	3.3	3.3
GSNAP	1.8	2.8	25.9	27.0

Новые (неаннотированные) сплайсы

H1ES			HUVEC		
Read count per junction from two replicates	Number of tested junctions	Proportion of junctions validated by at least two 454 reads (%)	Read count per junction from two replicates	Number of tested junctions	Proportion of junctions validated by at least two 454 reads (%)
2	192	72.4	2	192	74.0
3	192	77.6	3	192	75.0
4	96	74.0	4	96	76.0
5	96	82.3	5–6	96	84.4
6–7	96	79.2	7–8	96	84.4
8–11	96	81.3	9–12	96	86.5
12–24	96	87.5	13–23	96	94.8
≥25	96	88.5	≥24	96	90.6

Простая квантификация

- htseq-count
  - Мультимапперы - выбросить
  - Неоднозначные - выбросить
  - результат в прочтениях
- Не различает транскрипты (изоформы)

	union	intersection _strict	intersection _nonempty
read, gene_A	gene_A	gene_A	gene_A
read, gene_A	gene_A	no_feature	gene_A
read, gene_A, gene_A	gene_A	no_feature	gene_A
read, read, gene_A, gene_A	gene_A	gene_A	gene_A
read, gene_A, gene_B	gene_A	gene_A	gene_A
read, gene_A, gene_B	ambiguous	gene_A	gene_A
read, gene_A, gene_B	ambiguous	ambiguous	ambiguous

Htseq: http://www.ncbi.nlm.nih.gov/pubmed/25260700

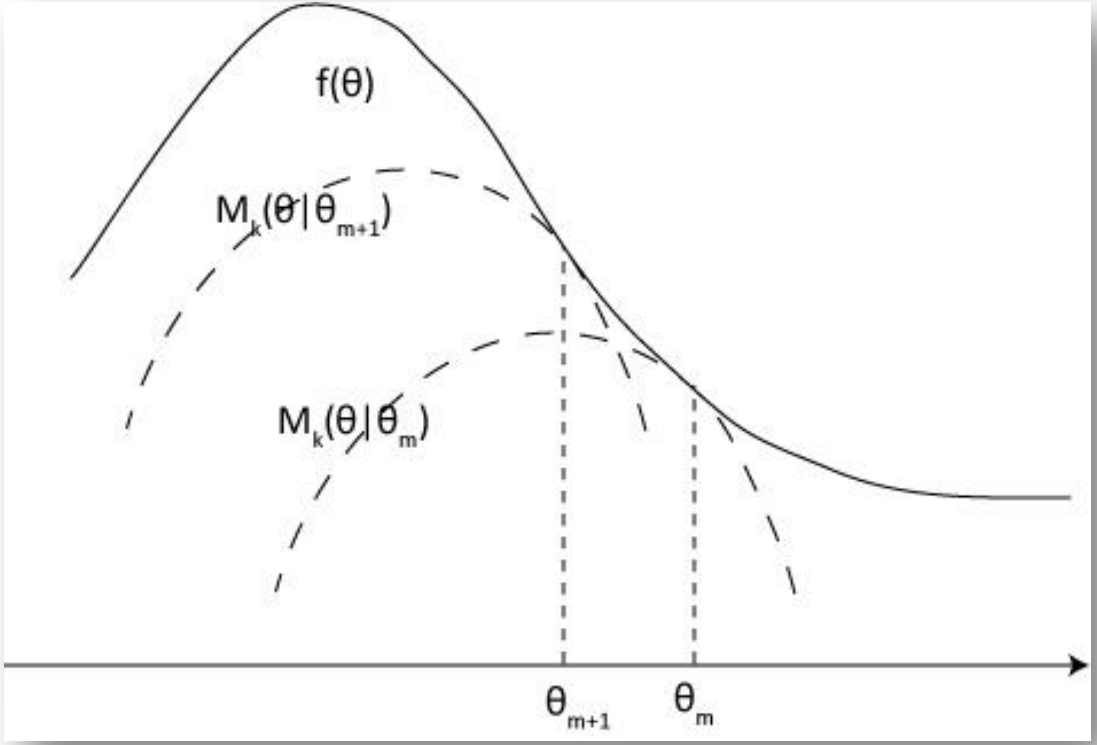
Квантификация на транскрипт

- Выровнять bowtie/bwa/STAR
- Считать RSEM (использует EM алгоритм)

RSEM:
http://www.ncbi.nlm.nih.gov/pubmed/21816040

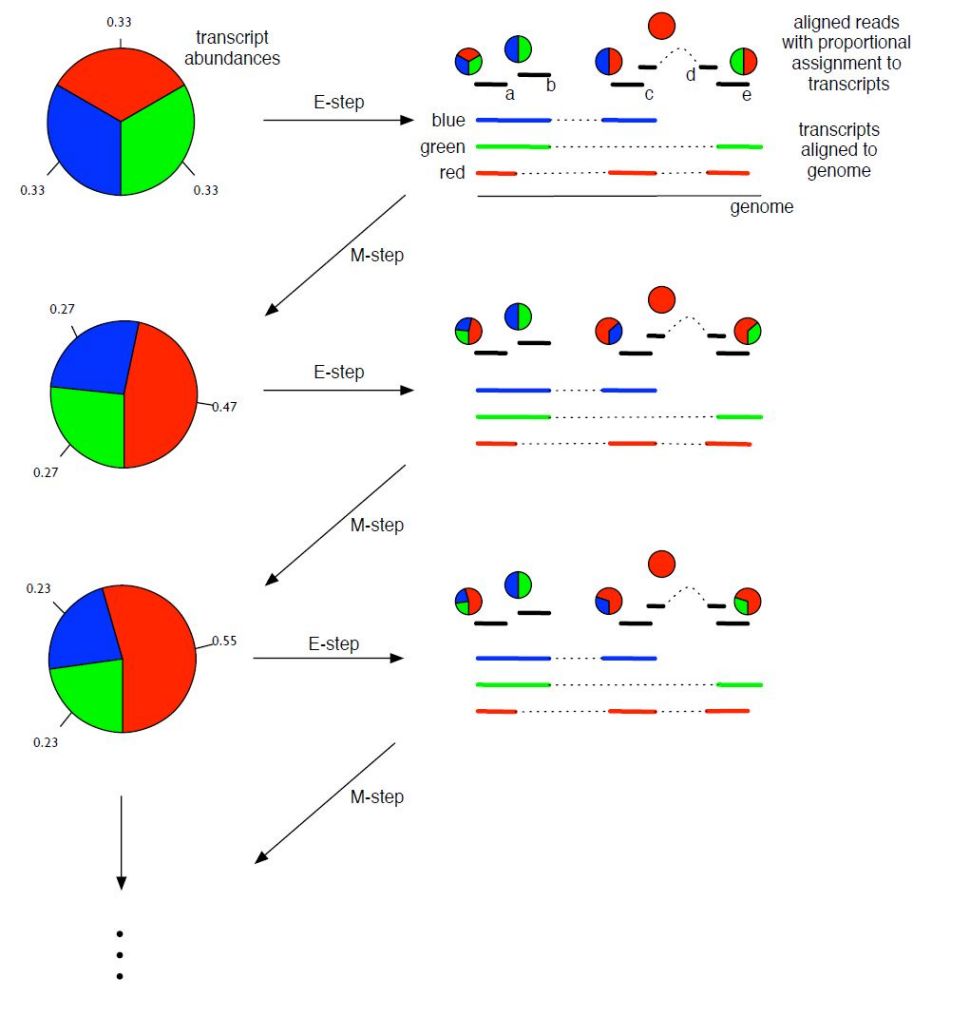
Expectation Maximization (EM)

- Ищем *maximum a posteriori* (MAP)
- Хорошая сходимость



EM для РНК-сек

- Используя распределение прочтений, итеративно улучшает нашу исходную догадку



Как оценить точность?

- MAQC - qPCR для ~ 1000 genes
- Коррекция на GC% улучшает совпадение с qPCR

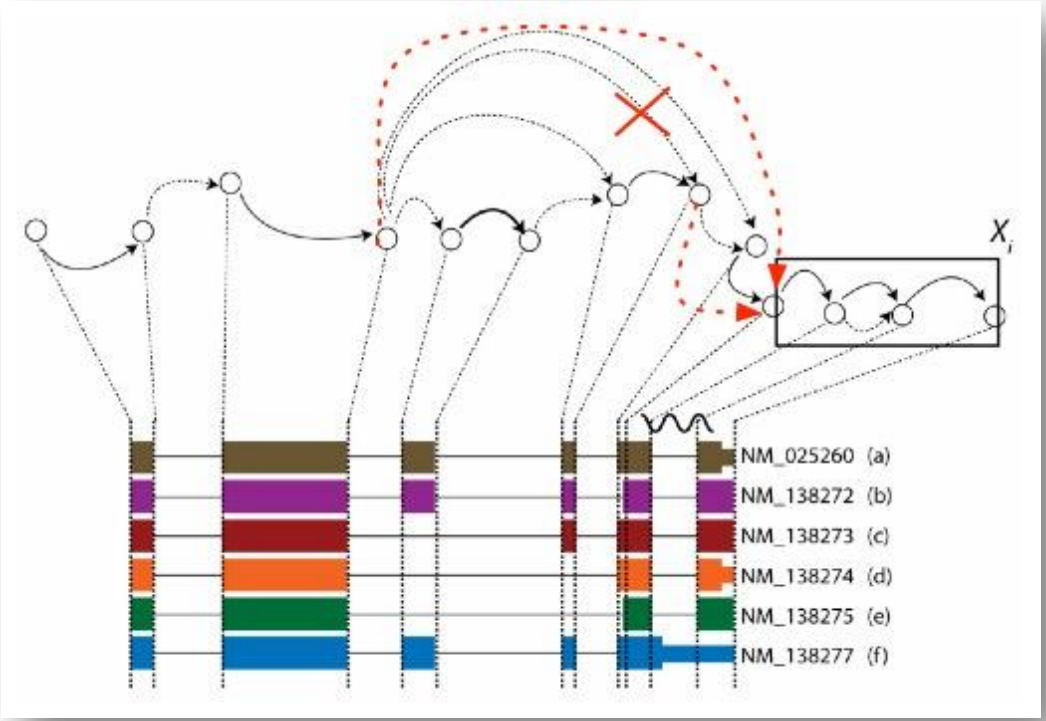
Table 2 Correlation of quantification method predictions with MAQC qRT-PCR values

SRA ID	Read type	Sample	RSEM	IsoEM	IsoEM (C)	Cufflinks	Cufflinks (C)	rQuant
SRX016366	SE	HBR	0.69	0.68	0.68	0.71	**0.79**	0.72
SRX003926	SE	HBR	0.68	0.67	0.67	0.7	**0.73**	0.71
SRX018974	PE	HBR	0.69	0.69	0.69	0.69	**0.78**	NA
SRX016368	SE	UHR	0.71	0.71	0.72	0.72	**0.77**	0.72
SRX016369	SE	UHR	0.73	0.74	0.74	0.73	**0.76**	0.74
SRX016370	SE	UHR	0.74	0.75	0.75	0.74	**0.77**	0.75
SRX016371	SE	UHR	0.74	0.75	0.75	0.74	**0.77**	0.75
SRX016372	SE	UHR	0.75	0.75	0.75	0.74	**0.77**	0.76
SRX003927	SE	UHR	0.72	0.71	0.72	0.71	**0.74**	0.72

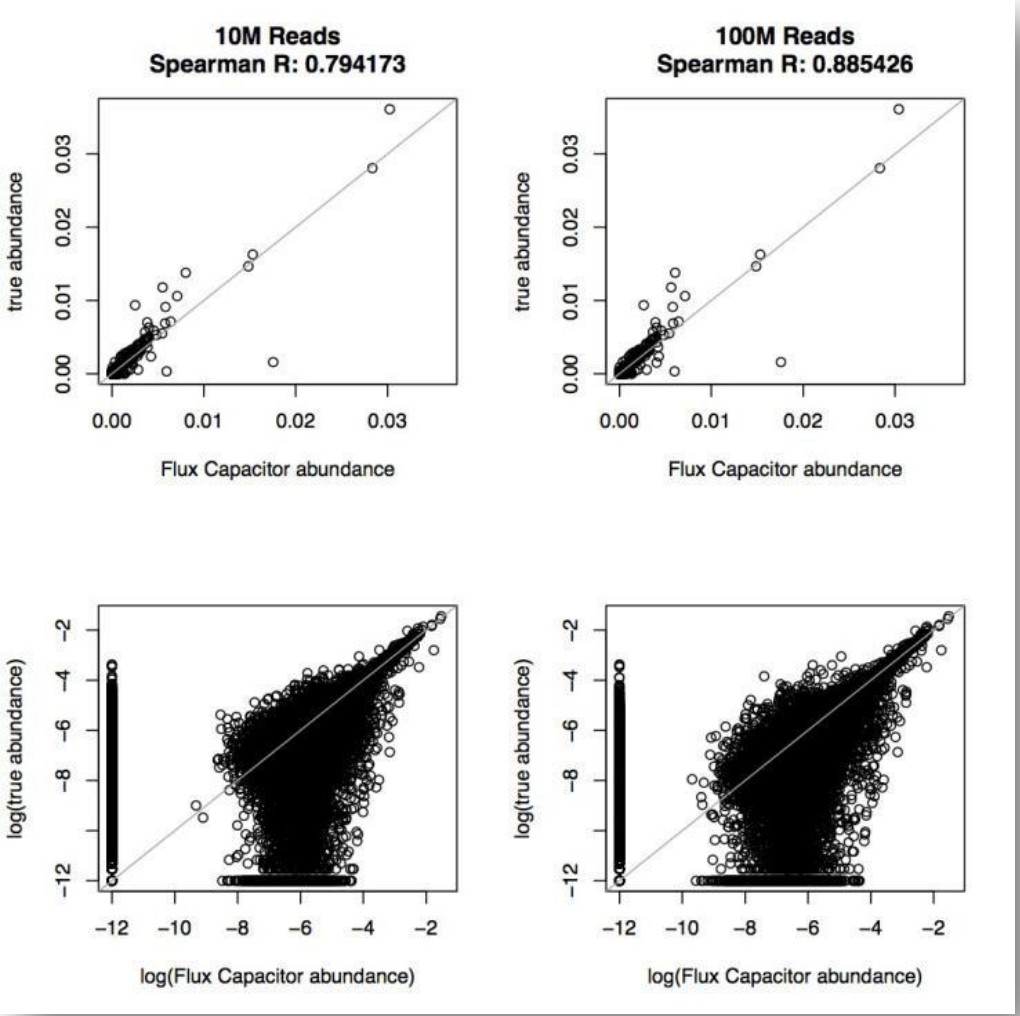
Correlation values (Pearson r^2 of log-transformed abundance values) were computed between the predictions of four methods and "gold-standard" values from qRT-PCR for nine different RNA-Seq data sets. IsoEM and Cufflinks were run with (C) and without their bias correction modes.

История с FluxCapacitor

- Использовался консорциумом GTEx
- Нигде не был толком опубликован

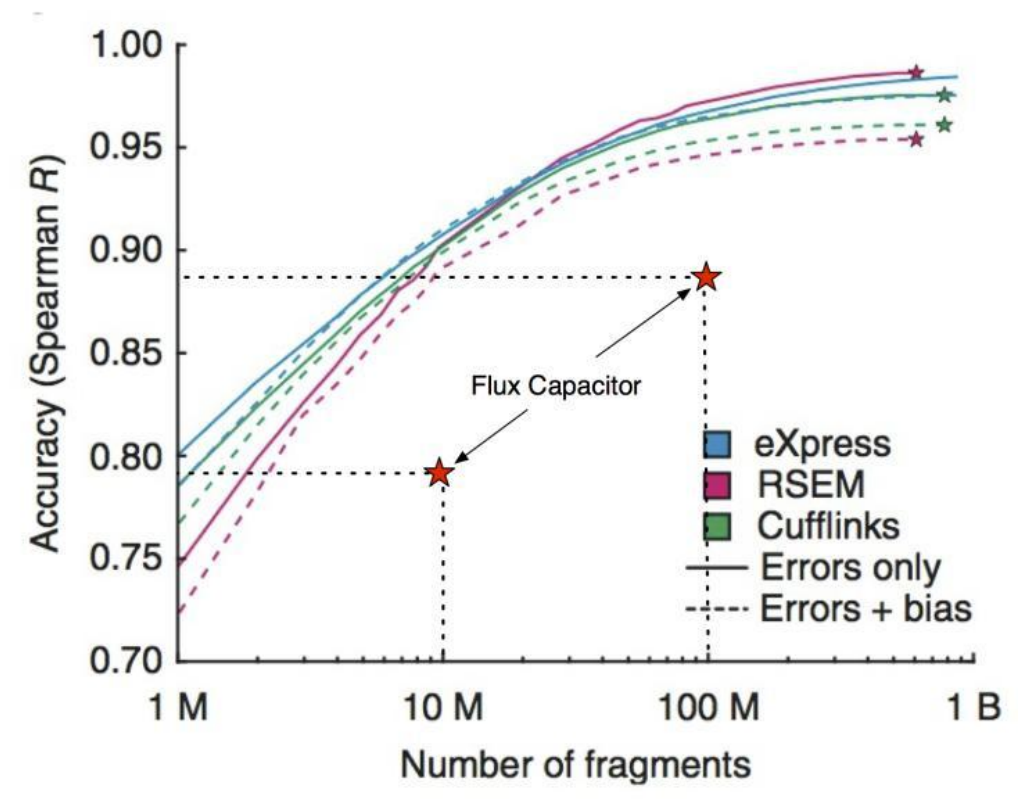


Большие проблемы с точностью



Выводы: все плохо

- FluxCapasitor fails
- Почему - толком не ясно



“The most embarrassing citation ever”

Using Flux Capacitor is equivalent to throwing out 90% of the data!

As a result, BAM alignment files were obtained and used to generate genome-wide normalized profiles using RSeQC software. Exon quantifications (summarized per-genes) were used for expression level determination, either as raw read counts or as reads per kilo-base per million mapped reads (RPKM) using Flux Capacitor (http://liorpachter.wordpress.com/tag/flux-capacitor/).

What is most embarrassing?

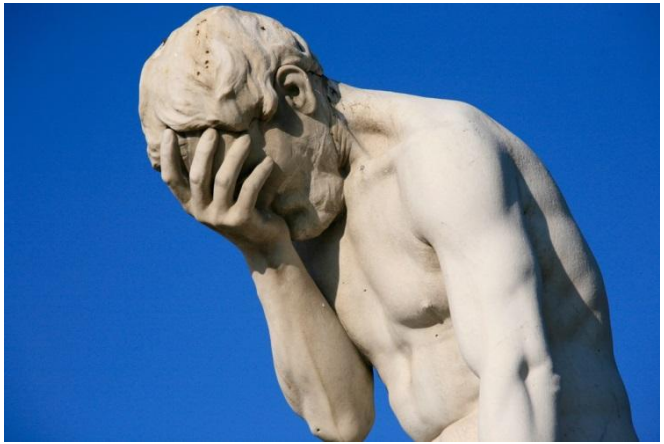
- The use of my blog as a citation for the methods of Flux Capacitor
- Neither reviewers nor readers noticed the problematic citation
- The use of Flux Capacitor by Iannone et al.
- The use of Flux Capacitor by the GTEx consortium in their main paper
- All of the above are equally embarrassing!

View Results

Polldaddy.com

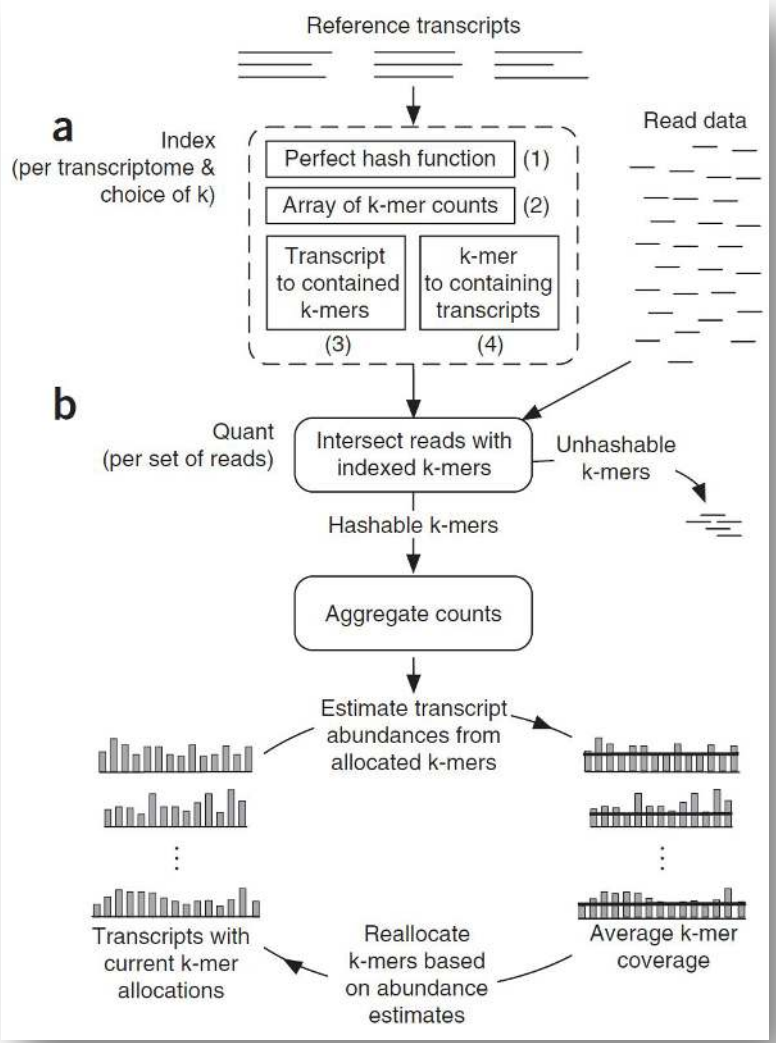
vote


SHAME!



Sailfish

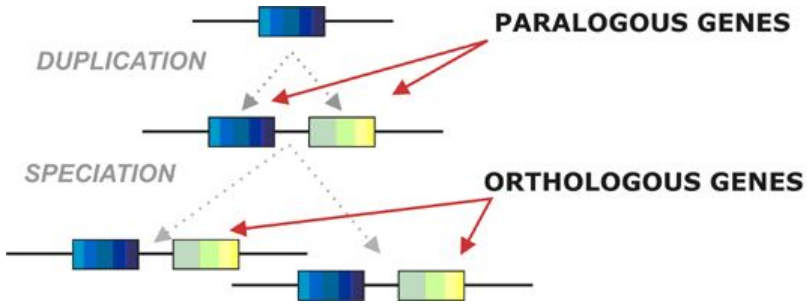
- Квантификация без выравнивания:
  - Сделать к-меры
  - Выровянть к-меры идеально
  - EM на транскриптах



Sailfish: http://www.ncbi.nlm.nih.gov/pubmed/24752080

Гены-паралоги

- Хорошая мера эффективности работы алгоритма квантификации



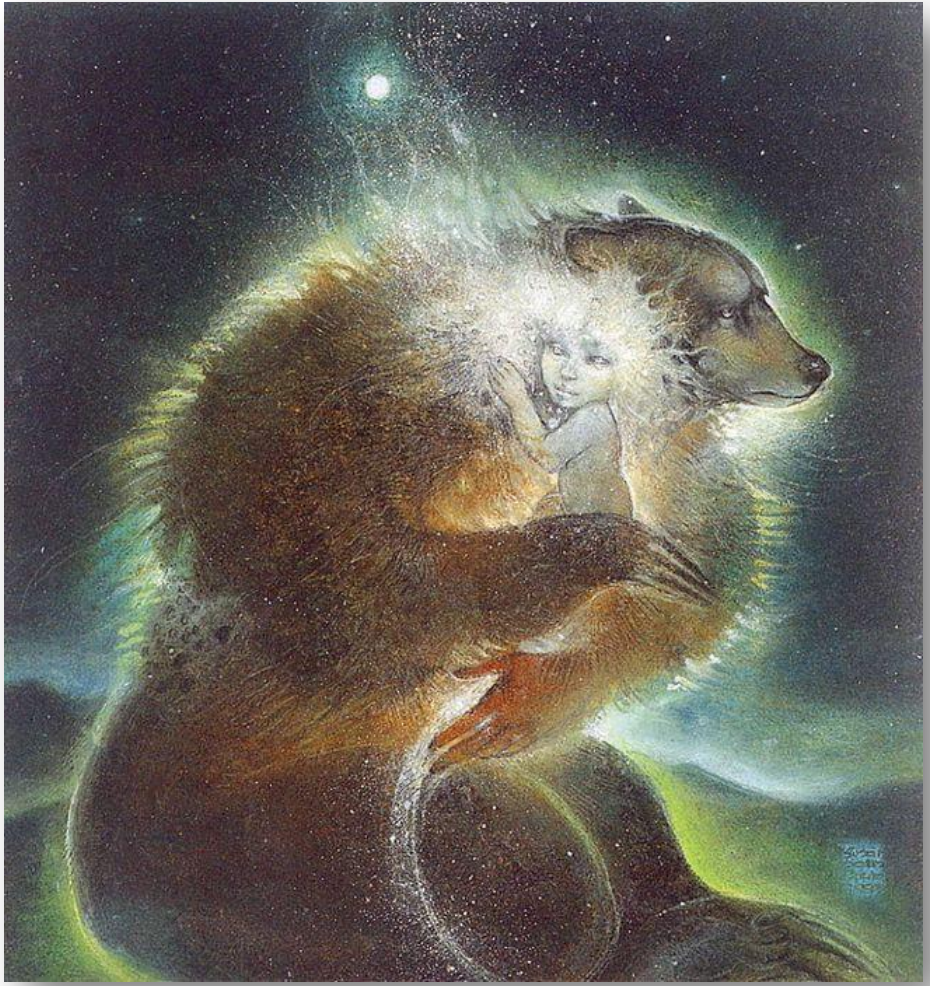
Supplementary Table 1: Abundance estimation in sequence-redundant human genes

	Sailfish	RSEM	eXpress	Cufflinks
Pearson	0.93	0.95	0.92	0.76
Spearman	0.88	0.89	0.88	0.80
RMSE	21.33	21.71	22.76	41.31
medPE	6.27	9.28	12.05	81.76

Kallisto

- Lior Pachter group
- *Artemis Kalliste*

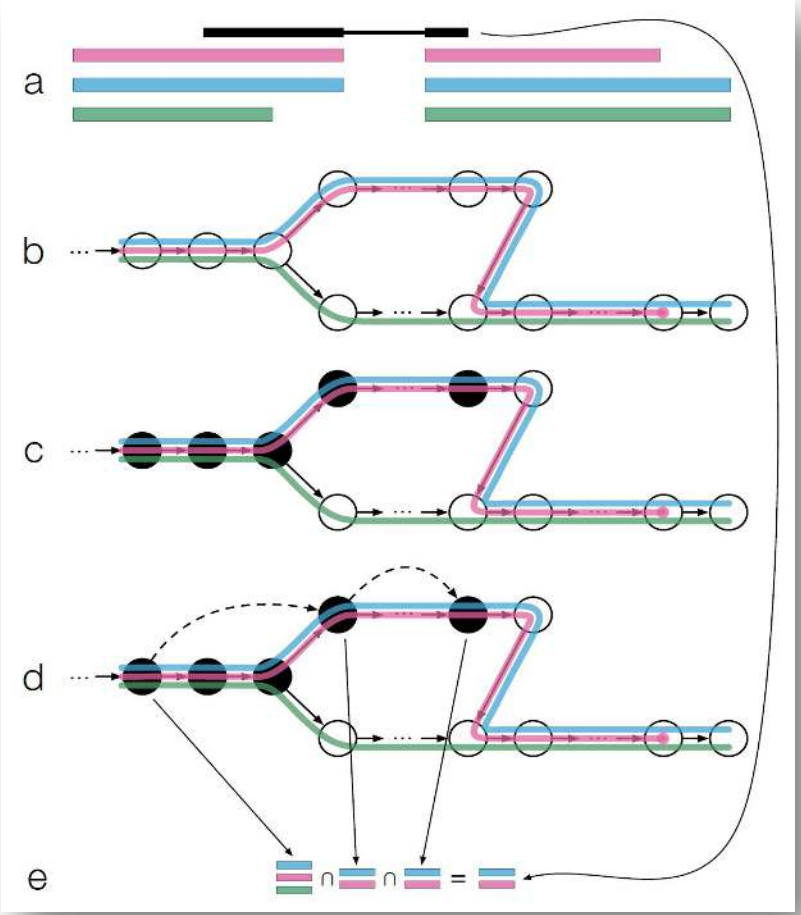
(Άρτεμις Καλλίστη)



kallisto preprint:
http://arxiv.org/abs/1505.02710

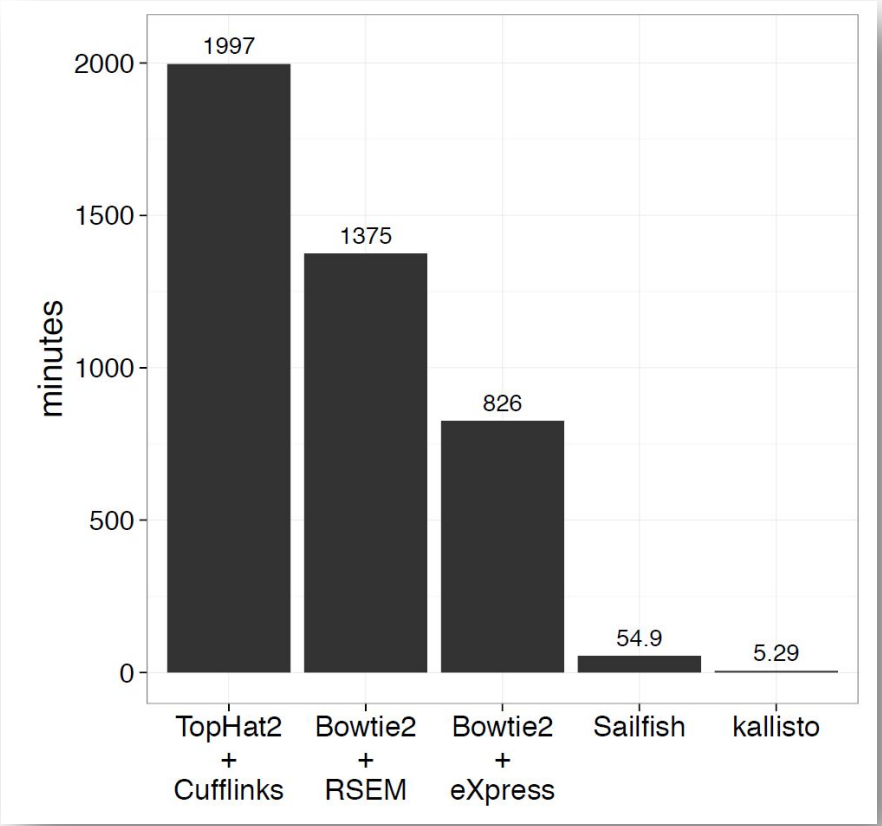
Транскриптомный граф де Брейна

- T-DBG:
  - Ноды = к-меры
  - Транскрипт = путь
- Вместо выравнивания – ищем классы эквивалентности



Скорость!

- Около-оптимальная скорость
- *«it’s only 5 times slower than counting words in fastq file with wc»*



Спасибо за внимание.