

План курса

- Лекция 1
 - Введение,
 - качество данных
 - Картирование, samtools
- Лекция 2
 - сборка транскриптома, и подсчёт транскриптомных ридов
 - Проверка самосогласованности: корреляционная тепловая карта, PCA/MDS
 - Нормализация
- Лекция 3
 - Дифф. экспрессия (edgeR)
 - Функциональный анализ (goseq) Дифф. сплайсинг (cuffdiff, DEXseq, MISO, SAJR)
 - Визуализация

Formats: gff (gtf, gff3)

GFF: general feature format

Tab-separated, 8 mandatory fields plus one optional attribute field

seqname	source	feature	start	end	score	strand	frame	group [attribute]
X	Ensembl	Repeat	2419108	2419128	42	.	.	group_id

Feature: gene, transcript, CDS, exon, site, promoter, etc

Score: coverage, prediction weight

Strand: +, -

Frame: 0, 1, 2

Group: id to group several features

Dot is used for undefined (empty) values.

GTF: general transfer format (or GFF2)

Differ only in the last field, it should contains attribute pairs: **name "value"**; In some cases formats like **name=value**; can be used as well.

Some sources (like UCSC) says that there are two mandatory attributes: **gene_id** and **transcript_id**

```
1 StringTie transcript 14399 15902 1000 - .
  gene_id "na.1"; transcript_id "na.1.1"; cov "16.654898"; FPKM "4.974452";
1 StringTie exon14399 14829 1000 - .
  gene_id "na.1"; transcript_id "na.1.1"; exon_number "1"; cov "18.297413";
```

GFF3

Attributes must be in form of **name=value**; Predefined attributes:

ID: should be unique

Parent: points to parent id (for example transcript for exons)

Etc

ctg123	.	mRNA	1300	9000	.	+	.	ID=mrna0001;Name=sonichedgehog
ctg123	.	exon	1300	1500	.	+	.	ID=exon00001;Parent=mrna0001

Создание аннотации по данным РНК-Сек: stringtie

- Сортируем bam файл по геномным координатам `samtools sort -m 500M -o out.bam in.bam`.
- Создаем аннотацию для каждого образца `stringtie bam -o sample.id.gtf -G ref.gtf`

Соединение аннотаций

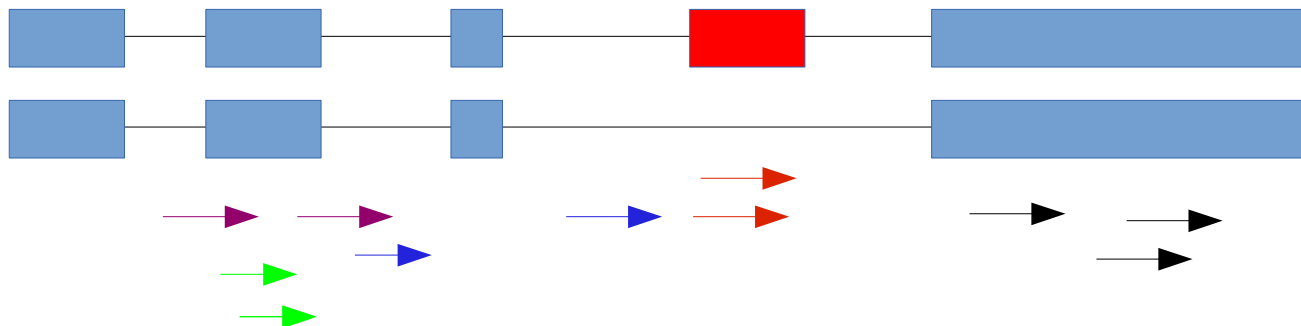
- Соединяем вместе
ls -1 ann/*gtf > ann/gtf.list #list of gtf files
stringtie --merge gtf.list -G ref.gtf -o merged.gtf
- Альтернативы:
 - cufflinks (старый, медленный)
 - scripture
 - etc

Подсчёт ридов

Задача: найти риды пересекающие ген

- сплайсинг:
 - **риды внутри интрона**
 - **риды пересекающие интрон**
 - **риды из альтернативных экзонов**
 - новые экзон-экзонные границы
- первые/последние экзоны
- перекрывание генов
- картирование в несколько позиций
- цепь, парность

HTSeq, stringtie, etc



Подсчет ридов: htseq-count

htseq-count

--stranded=no — данные не цепь-специфичны

-f bam — входной формат - bam

-m intersection-strict использовать наиболее строгий подход к приписыванию ридов к гену

in.bam merged.gtf > counts.out

входные и выходной файл

htseq-count кроме информации о риде, попавшем на ген, печатает еще несколько строчек. На них стоит посмотреть для оценки качества, но для дальнейшего анализа их надо удалить:

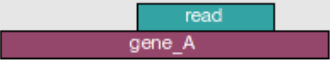
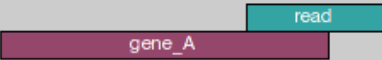


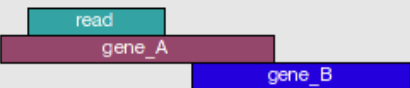

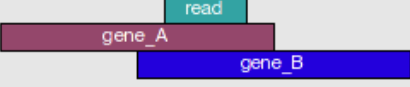
__no_feature 73305

__ambiguous 6658

__too_low_aQual 0

__not_aligned 0

__alignment_not_unique 26782

	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	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Подсчет ридов: stringtie

Stringtie позволяет посчитать количество ридов попавших в данный ген

- B — estimate coverage for genes from -G
- e — do not assemble new genes
- G test/merged.gtf

To transform coverage to read counts and merge it into single files prepDE.py (download it from stringtie website) script can be used.

prepDE.py -i **sample.list.txt**

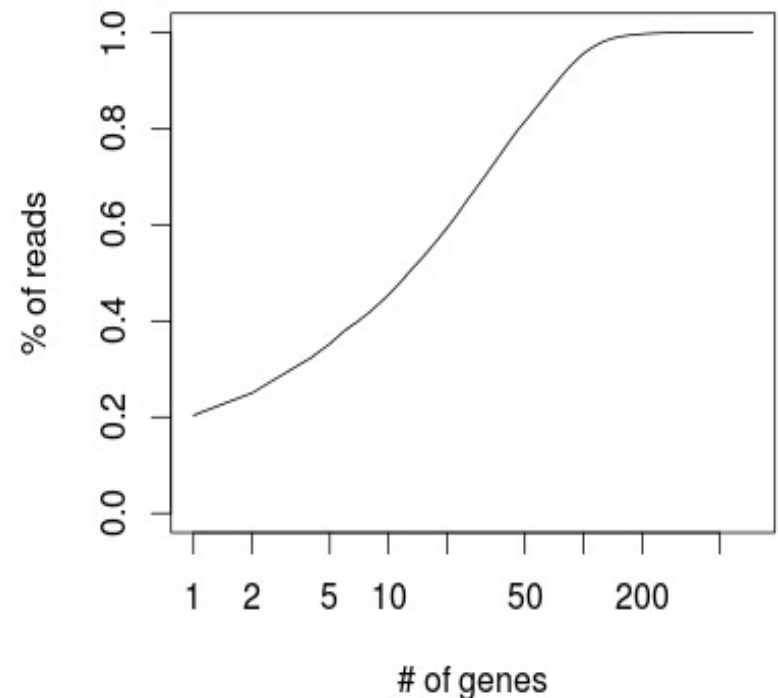
- g ./gene.counts.csv — output file for gene counts
- t ./transc.counts.csv — output file transcript counts
- l 100 — read length

Формат файла **sample.list.txt**:

```
sample1_id path_to_gtf1  
sample2_id path_to_gtf2
```

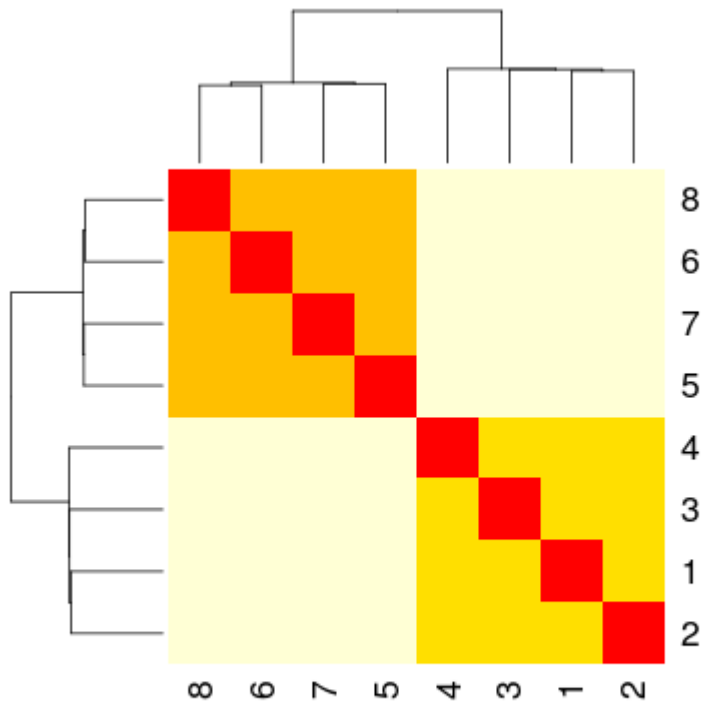
Оценка качества данных РНК-сек

- % картирующихся ридов
- из них на гены
- риды с рРНК, тРНК, митохондриальные гены
- гены с очень высоким покрытием



Самосогласованность: корреляция

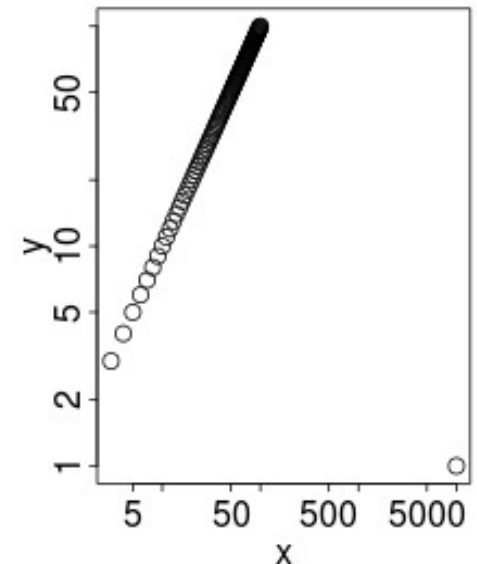
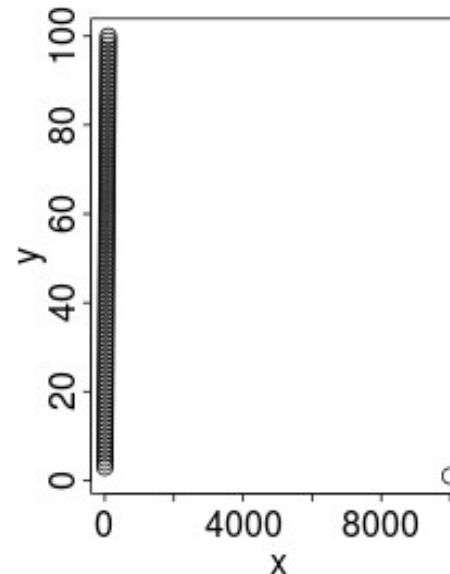
- Образцы для одного состояния должны коррелировать лучше, чем для разных
- Матрицы корреляции принято изображать при помощи тепловых карт (heatmap)



R:

```
correlation = cor(data, method = 'spearman')  
heatmap(correlation,  
        symm = TRUE,  
        distfun = function(x){as.dist(1-x)},  
        ColSideColors=tissue.col)
```

Пирсон: -0.15
Пирсон, логарифм: 0.51
Спирман: 0.94



Более практический пример

сгенерируем данные для 5000 генов в 4 контрольных и 4 больных образцах

```
x = rnorm(5000,sd=0.6)
```

```
y = rnorm(5000,sd=0.6)
```

```
d = cbind(rnorm(5000,x),rnorm(5000,y),rnorm(5000,y),rnorm(5000,x),  
          rnorm(5000,y),rnorm(5000,y),rnorm(5000,x),rnorm(5000,x))
```

```
pca=prcomp(t(d))
```

первые два гена (как и любые случайные) не разделяют образцы

```
col=c(1,2,2,1,2,2,1,1)
```

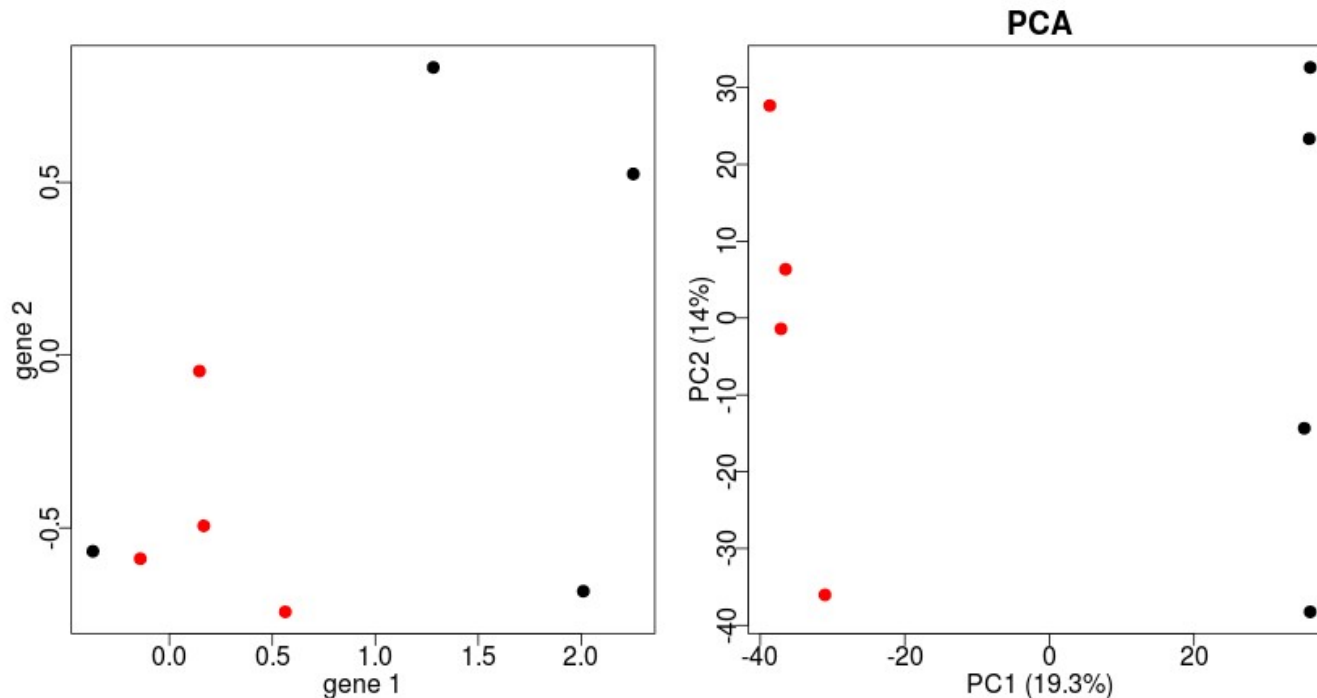
```
plot(d[1,],d[2,],pch=19,col=col,xlab='gene 1',ylab='gene 2')
```

```
dev = round(pca$sdev/sum(pca$sdev)*100,1)
```

а главные компоненты - разделяют

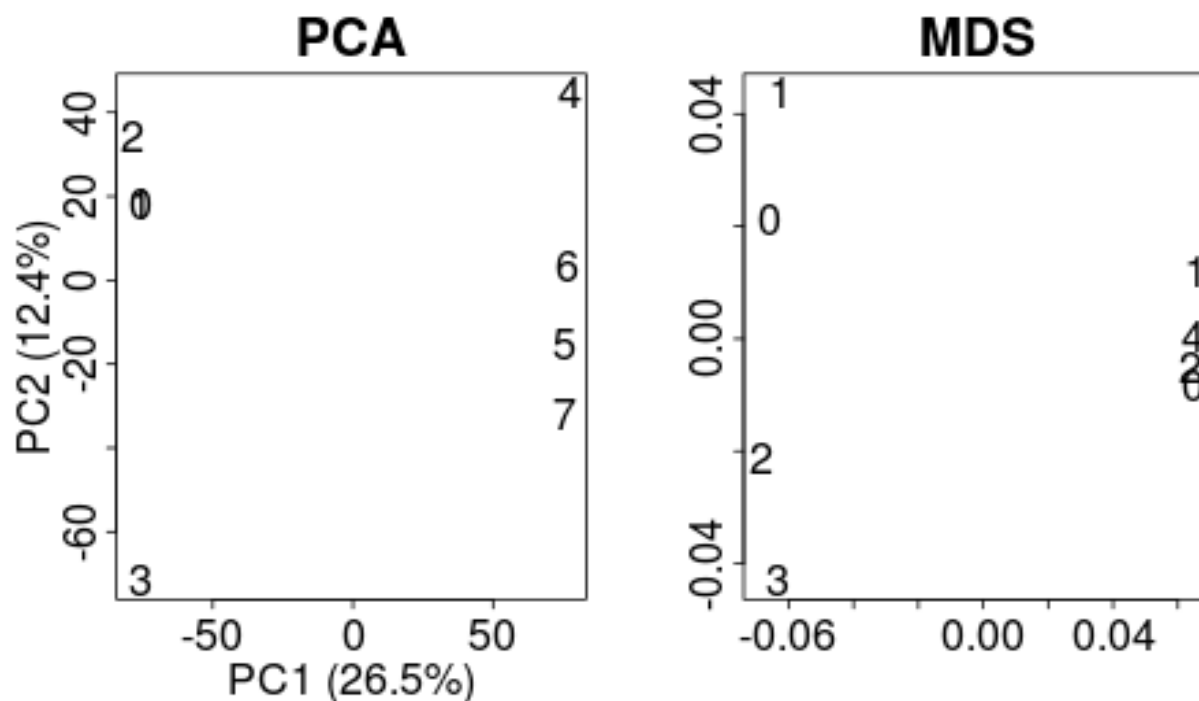
```
plot(-pca$x[,1],-pca$x[,2],pch=19,col=col,main='PCA',
```

```
      xlab=paste('PC1 (',dev[1],'%)',sep=''),ylab=paste('PC2 (',dev[2],'%)',sep=''))
```



Самосогласованность: MDS

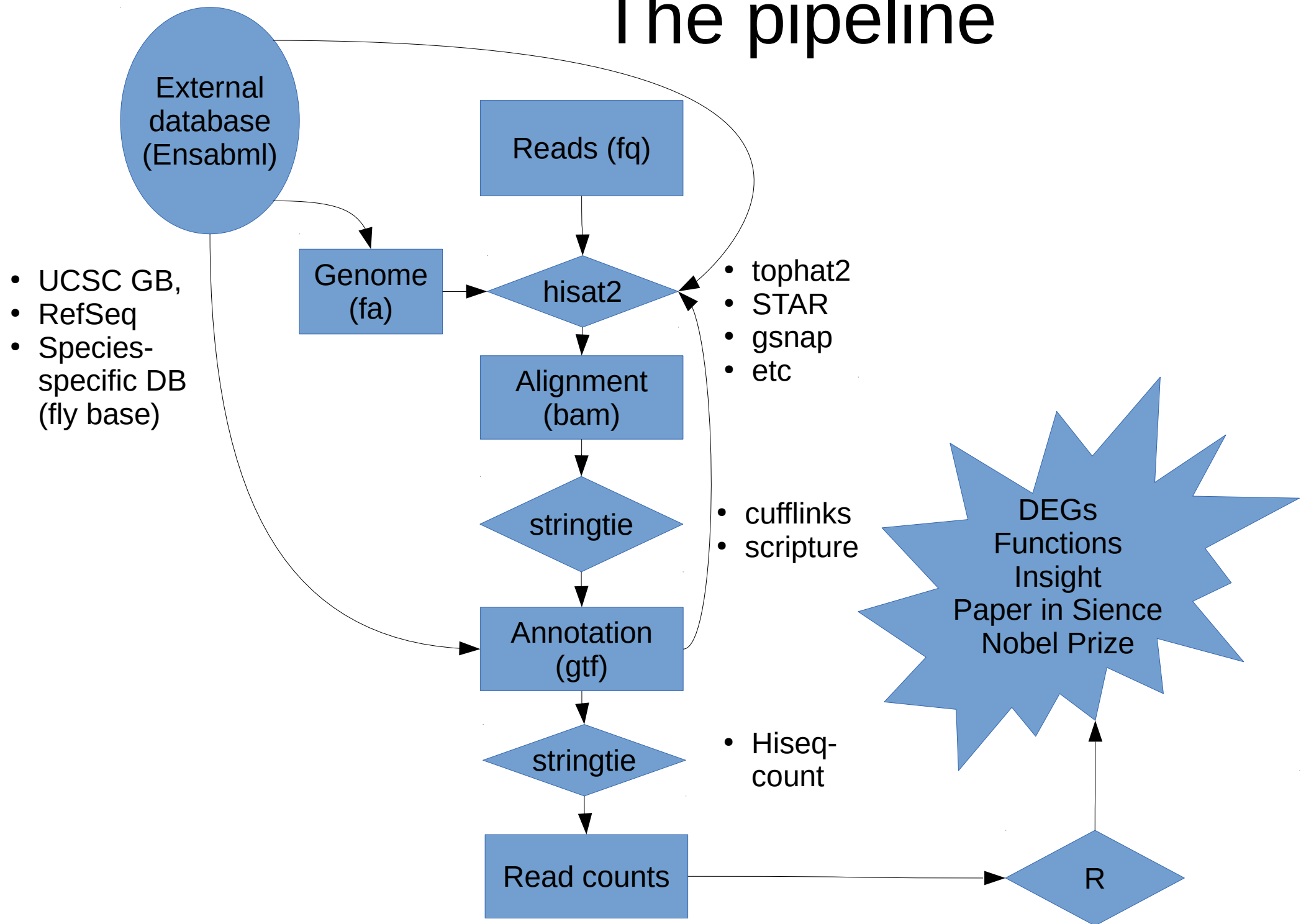
Иногда евклидово расстояние не подходит или его вовсе нет, а есть только матрица попарных расстояний. Например 1 — корреляция. Тогда используем MDS:



R:

```
mds=cmdscale(1-correlation,k=2)
plot(mds[,1],mds[,2],
col=tissue.col,
pch=19,
cex=age.size)
```

The pipeline



RPKM или FPKM

Reads (Fragments) per Kilobase per Million reads

$$[RF]PKM = \frac{r_g}{l_g R} \times 10^9$$

$$R = \sum r_g$$

r_g — число ридов в гене

l_g — длина гена

R — общее число ридов
(приписанных к генам)

Проблемы:

- Что делать с альтернативным сплайсингом?
 - считать ли риды с альтернативных экзонов?
 - что такое длина гена?
- изменения экспрессии нескольких мажорных генов приводят к видимым изменениям экспрессий других генов. Что делать если все изменения в одну сторону?



Mapping and quantifying mammalian transcriptomes
by RNA-Seq

RPKM vs TMP — массовая vs молярная концентрация

TMP = Transcripts **P**er **M**illion

$$E_g \sim \frac{r_g}{l_g}$$

$$TPM = nE_g = \frac{E_g}{\sum_g E_g} \times 10^6$$

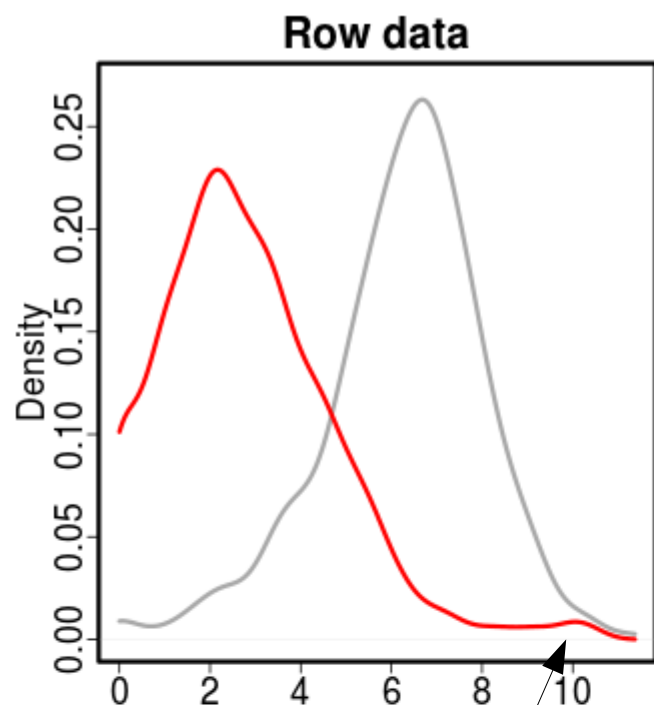
Measurement of mRNA abundance using RNA-seq data:
RPKM measure is inconsistent among samples

Но проще всего использовать
CPM — count per million

$$CPM = \frac{r_g}{R} \times 10^6$$

Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments

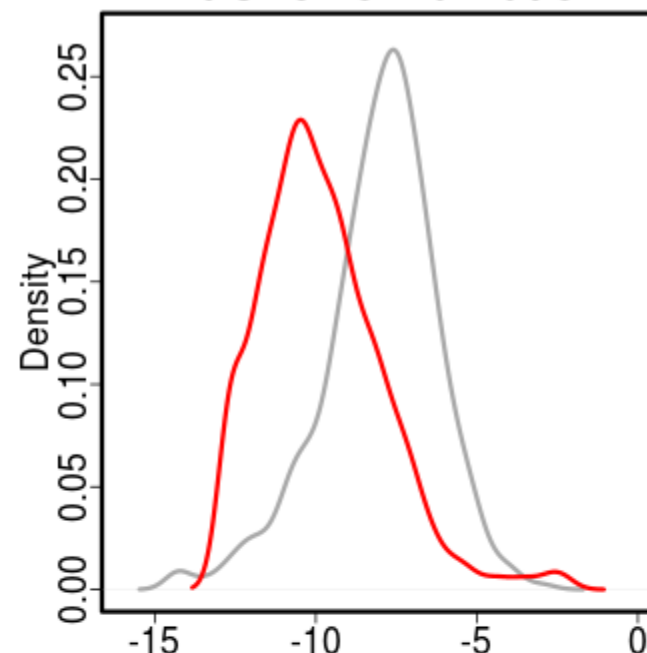
James H Bullard^{1*}, Elizabeth Purdom^{2†}, Kasper D Hansen¹, Sandrine Dudoit^{1,2}



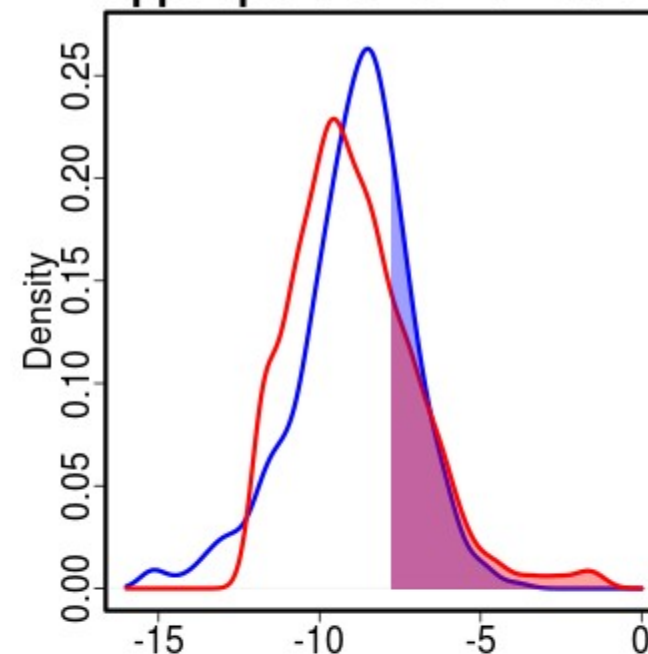
Гены более всего влияющие на
размер библиотеки

Нормировка

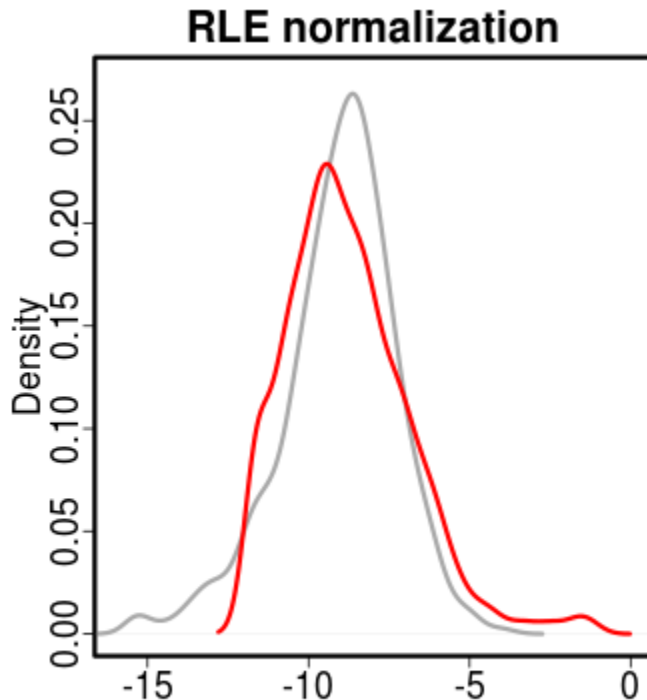
Lib size normalization



Upperquartile normalization



Нормировка: RLE



$$nf_s = \underset{g \in \text{genes}}{\text{median}} \frac{r_{g,s}}{\left(\prod_{\text{smpl}=1}^m r_{g,\text{smpl}} \right)^{1/m}}$$

Anders and Huber *Genome Biology* 2010, **11**:R106
<http://genomebiology.com/2010/11/10/R106>



METHOD

Open Access

Differential expression analysis for sequence count data

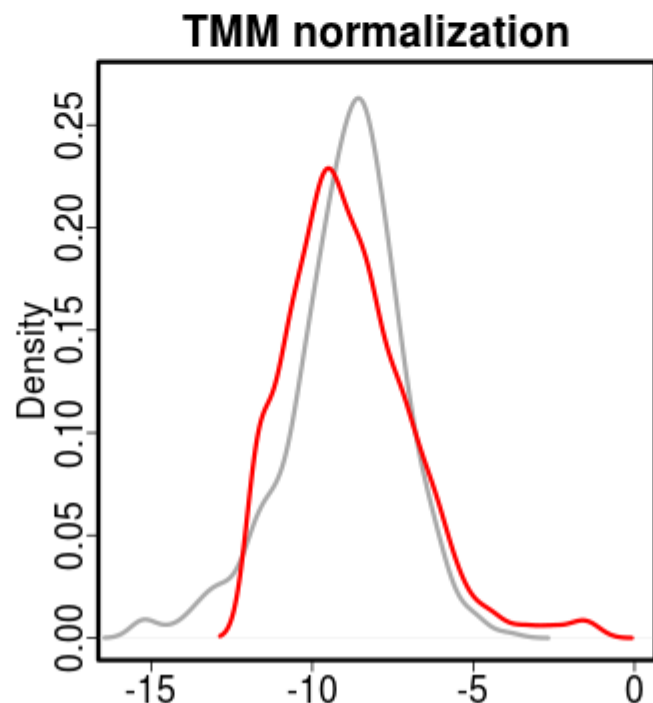
Simon Anders*, Wolfgang Huber

Нормировка: TMM

$$M_g = \log_2 \frac{Y_{gk} / N_k}{Y_{gk'} / N_{k'}}$$

$$\log_2(TMM_k^{(r)}) = \frac{\sum_{g \in G^*} w_{gk}^r M_{gk}^r}{\sum_{g \in G^*} w_{gk}^r} \text{ where } M_{gk}^r = \frac{\log_2 \left(\frac{Y_{gk}}{N_k} \right)}{\log_2 \left(\frac{Y_{gr}}{N_r} \right)} \text{ and } w_{gk}^r = \frac{N_k - Y_{gk}}{N_k Y_{gk}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}};$$

$$Y_{gk}, Y_{gr} > 0.$$



Тримированный по экспрессии и по изменению взвешенный (по экспрессии) средний логарифм изменений

Robinson and Oshlack *Genome Biology* 2010, 11:R25
<http://genomebiology.com/2010/11/3/R25>



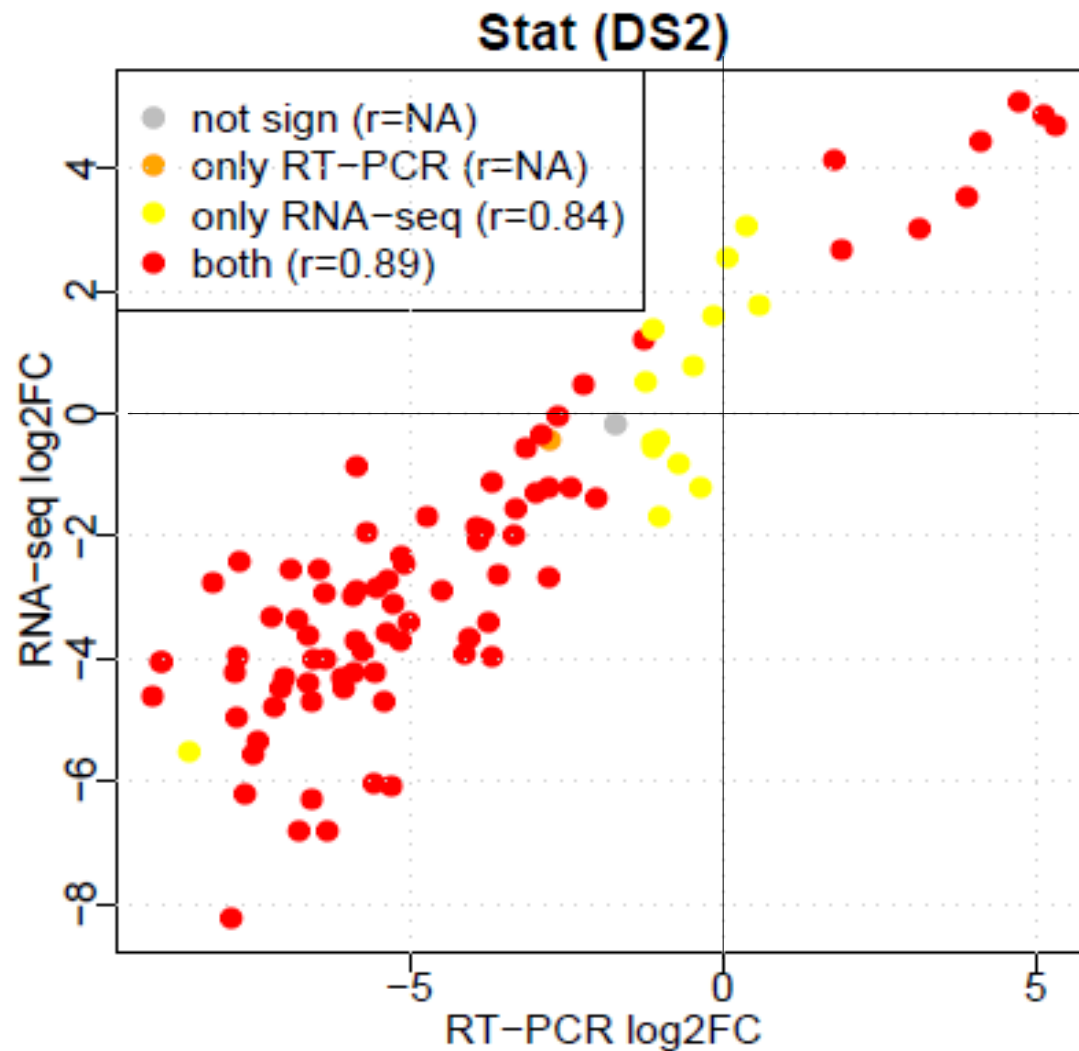
METHOD

Open Access

A scaling normalization method for differential expression analysis of RNA-seq data

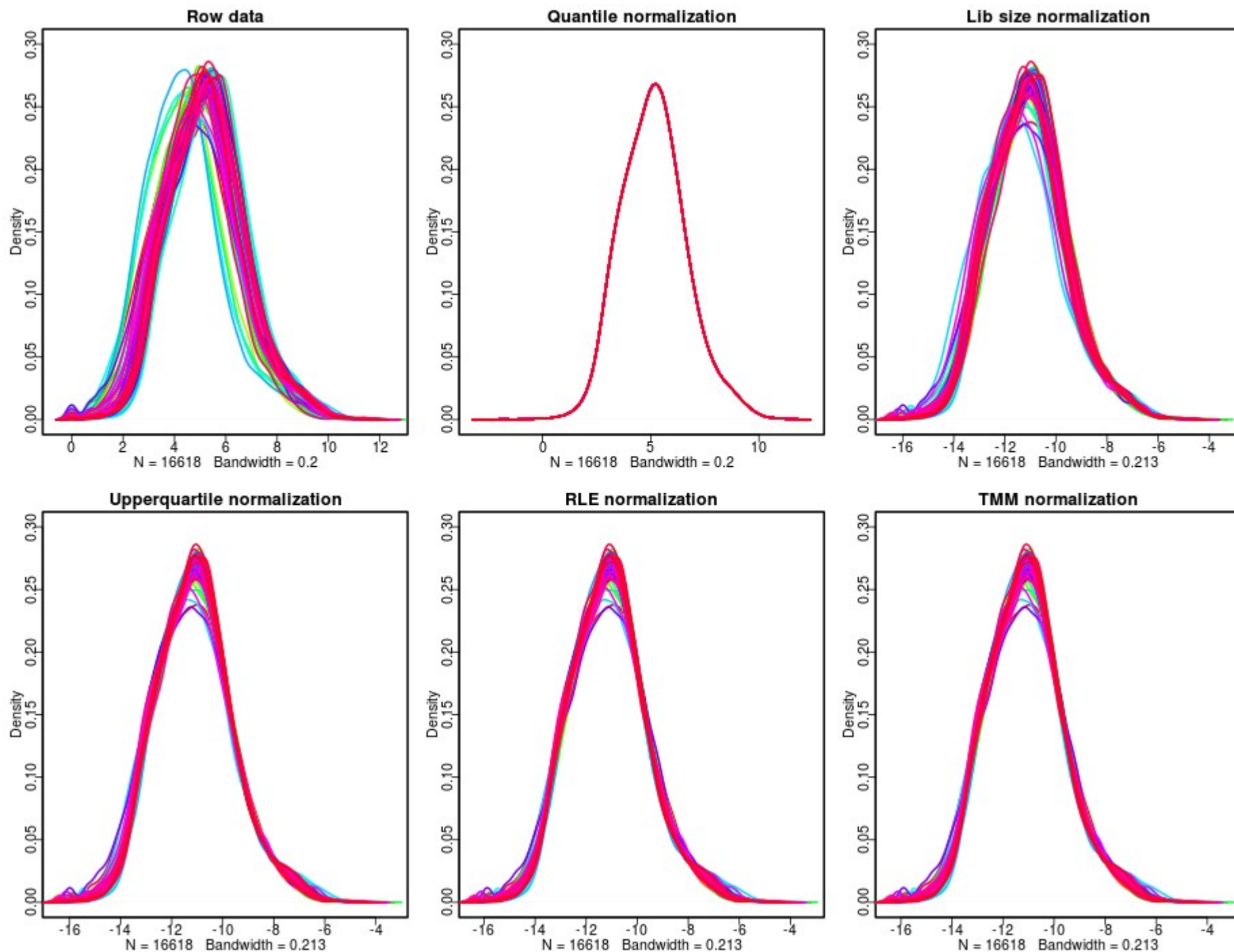
Mark D Robinson^{1,2*}, Alicia Oshlack^{1*}

А на самом деле почти все гены подавляются



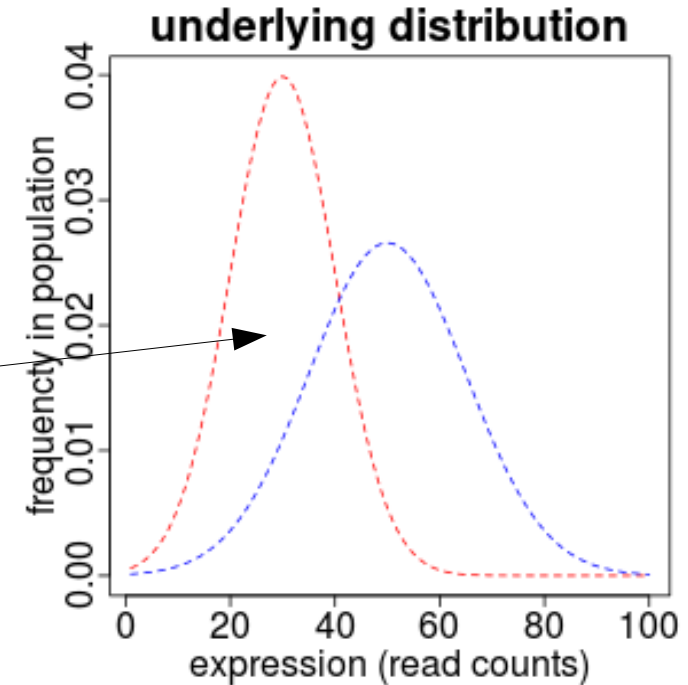
однако такое
случается редко

Нормировка "нормальных" данных



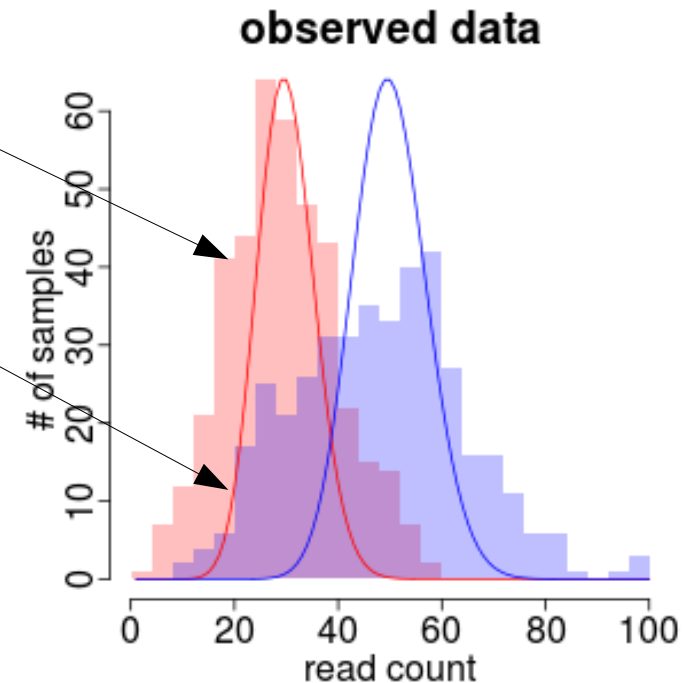
Биологическая вариабельность

распределение доноров по
уровню экспрессии гена
(~число ридов)



наблюдаемое
распределение образцов по
уровню экспрессии гена

пуассоновское (ожидаемое)
распределение числа ридов



Негативно-биномиальное распределение

- распределение количества удач в бернулевских испытаниях с вероятностью успеха p до получения r неудач

$$f(k; r, p) \equiv \Pr(X = k) = \binom{k + r - 1}{k} p^k (1 - p)^r \quad \text{for } k = 0, 1, 2, \dots$$

Стандартная
параметризация
 p, r

$$mean = \frac{pr}{1 - p}$$

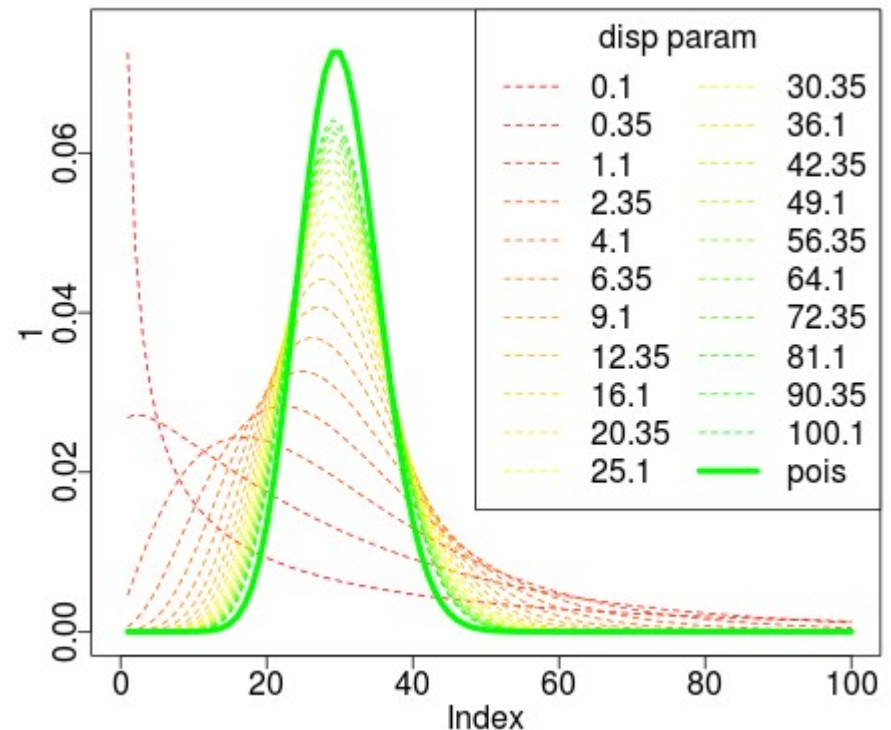
$$var = \frac{pr}{(1 - p)^2}$$

Альтернативная
параметризация
 m, r

$$mean = m$$

$$var = m + \frac{m^2}{r}$$

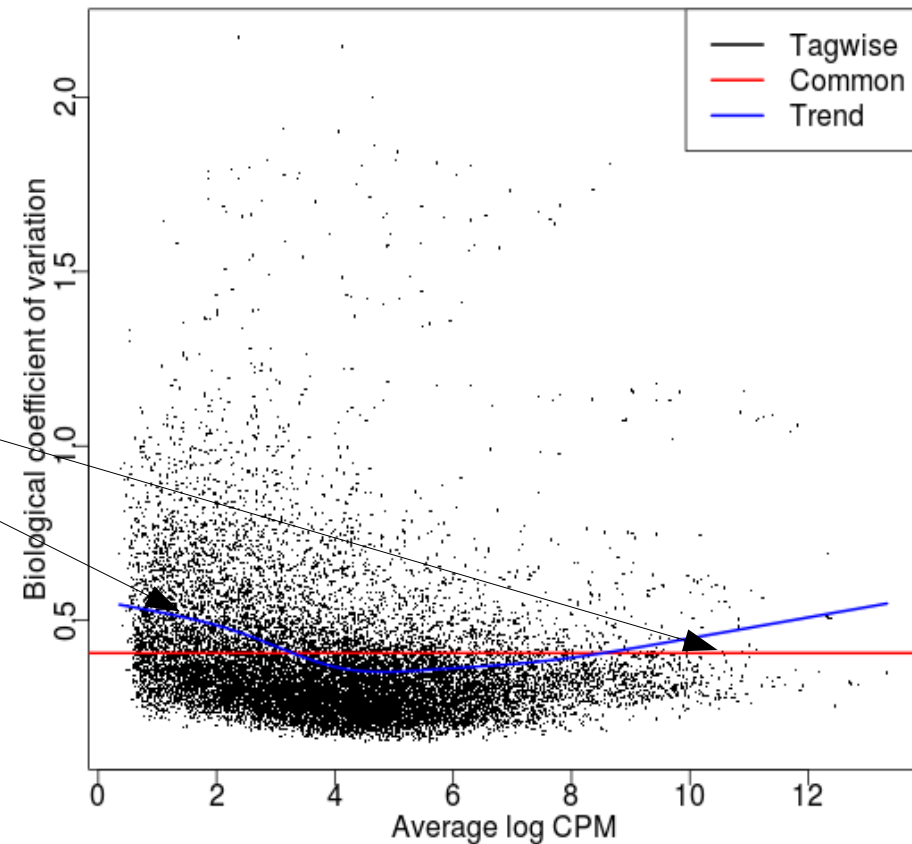
$$p = \frac{m}{m + r}$$



edgeR: оценка дисперсионного параметра

counts — таблица: гены — образцы
gender — предиктор. Например пол донора

```
edgeR = DGEList(counts)
edgeR = calcNormFactors(edgeR, method='RLE')
design = model.matrix(~ gender)
edgeR = estimateGLMCommonDisp(edgeR, design)
edgeR = estimateGLMTrendedDisp(edgeR, design)
edgeR = estimateGLMTagwiseDisp(edgeR, design)
strict.disp =
pmax(edgeR$tagwise.dispersion, edgeR$trended.dispersion, edgeR$common.dispersion)
plotBCV(edgeR)
```



edgeR: многофакторный анализ

```
formula = ~ a + s + a:s  
design = model.matrix(formula)  
glm = glmFit(edgeR, design, dispersion=strict.disp)
```

Указываем функции glmLRT номер тестируемого фактора:

```
pv.age      = glmLRT(glm, 2)$table$PValue  
pv.sex      = glmLRT(glm, 3)$table$PValue  
pv.agesex = glmLRT(glm, 4)$table$PValue
```


Поправка на множественное тестирование

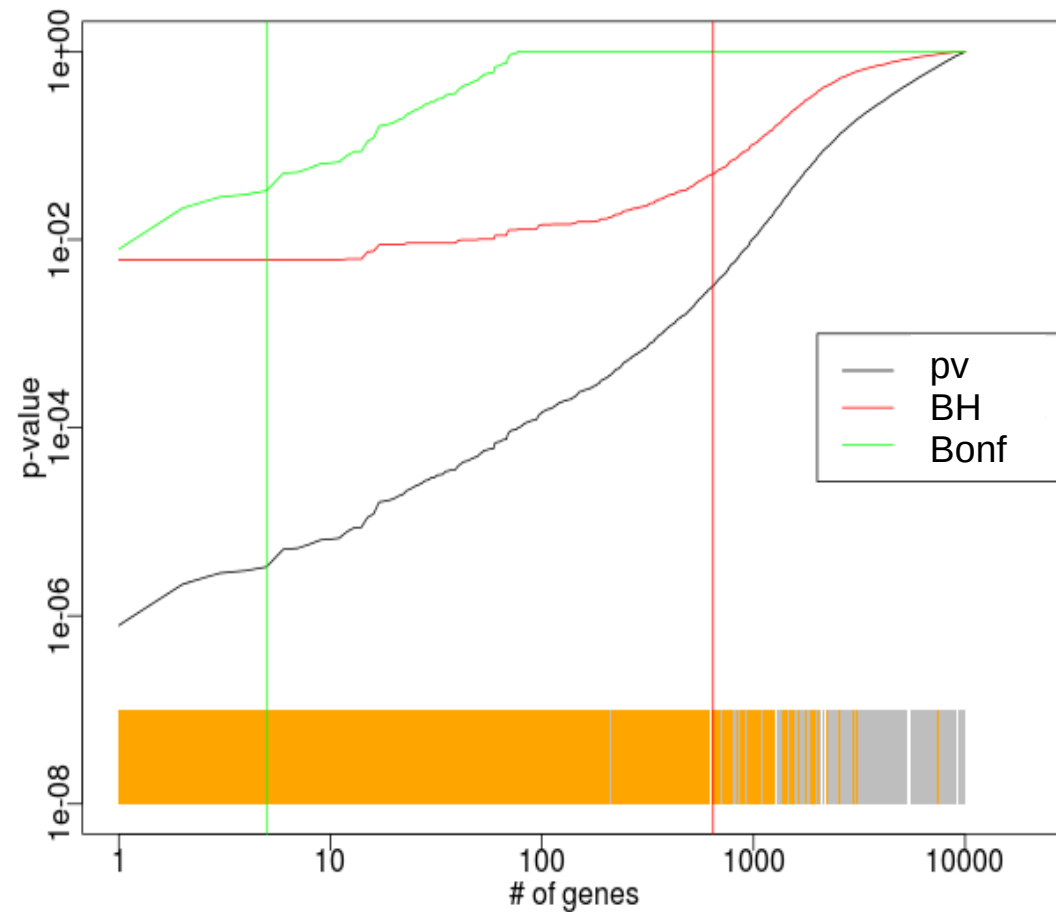
- поправка Бонферони:
контролируем вероятность хоть одного ложного:

$$pv.corr_i = pv_i * N$$

- поправка Бенджамини-Хочберга:
контролируем долю ложных

$$pv.corr_i = \frac{pv_i * N}{i}$$

R: p.adjust



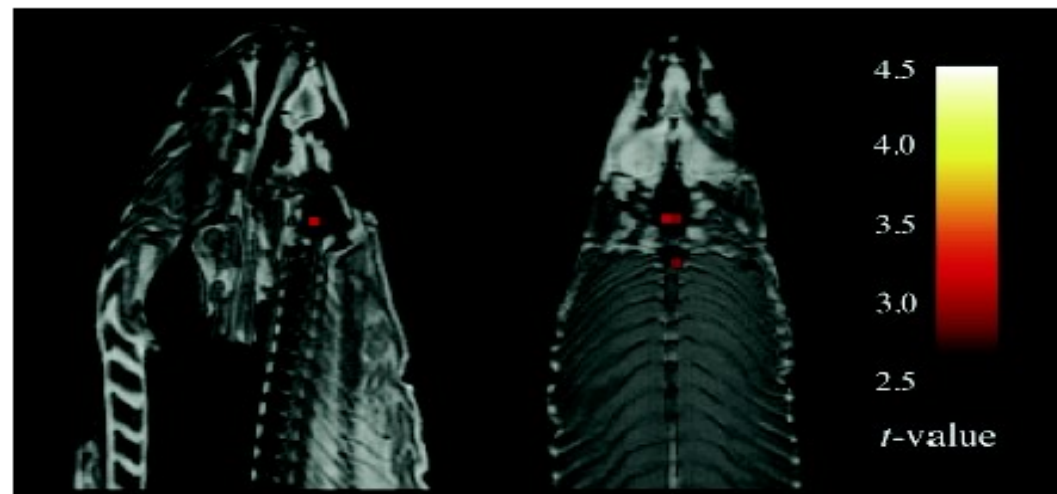
Помни о мёртвом лососе

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



Домашнее задание

- Прокартируйте все образцы при помощи hisat2
- Соберите транскрипты при помощи stringtie для каждого образца используя аннотацию из ensembl (-G)
- Перекартируйте риды используя новую аннотацию
- Оцените экспрессию генов в каждом образце при помощи stringtie, получите таблицу read counts при помощи prepDE.py (можно использовать htseq-count — он установлен на расчетном узле)
- Постройте PCA и heatmap (коэффициент корреляции Спирмана) для образцов