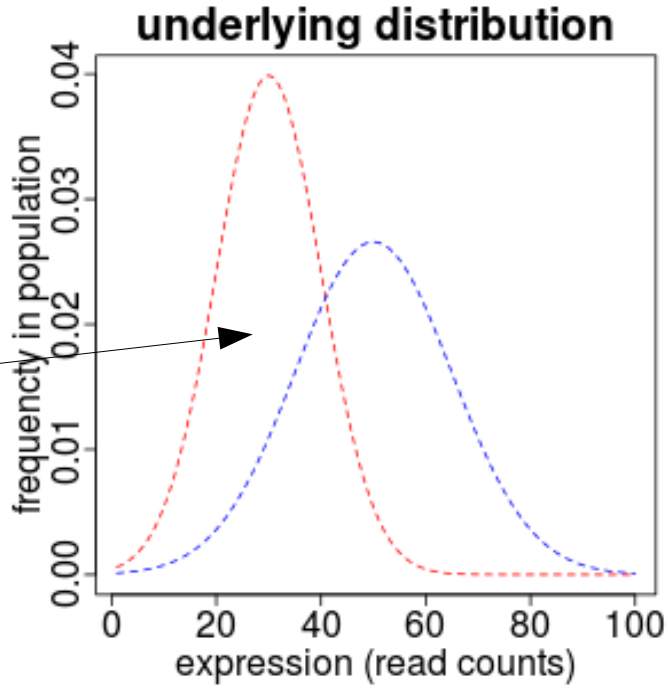
План

- нормализация
- Линейные и обобщенные линейные модели, ANOVA
- Дифф. экспрессия (edgeR)
- поправка на множественное тестирование
- Кластеризация (hclust, k-means, PAM, SOM)
- Функциональный анализ (goseq)
- Дифф. сплайсинг (DEXseq, SAJR)
- Визуализация

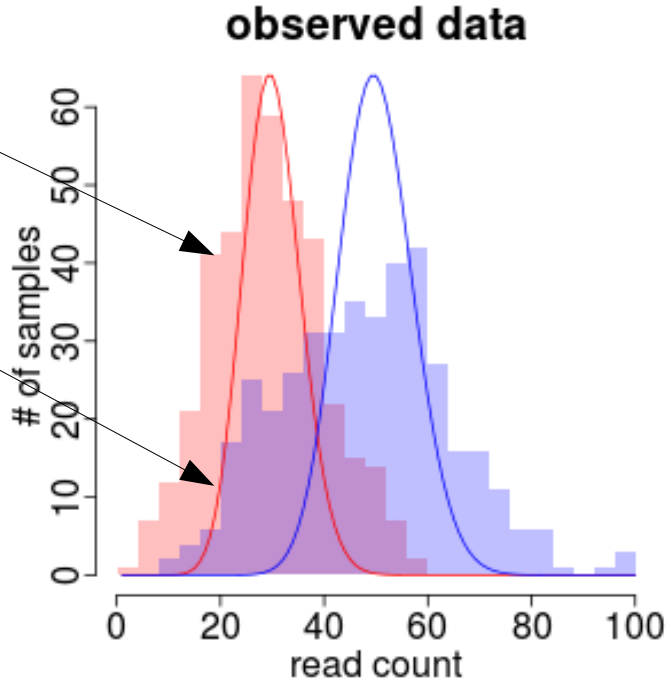
Биологическая вариабельность

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

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

Стандартная параметризация

p, r

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

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

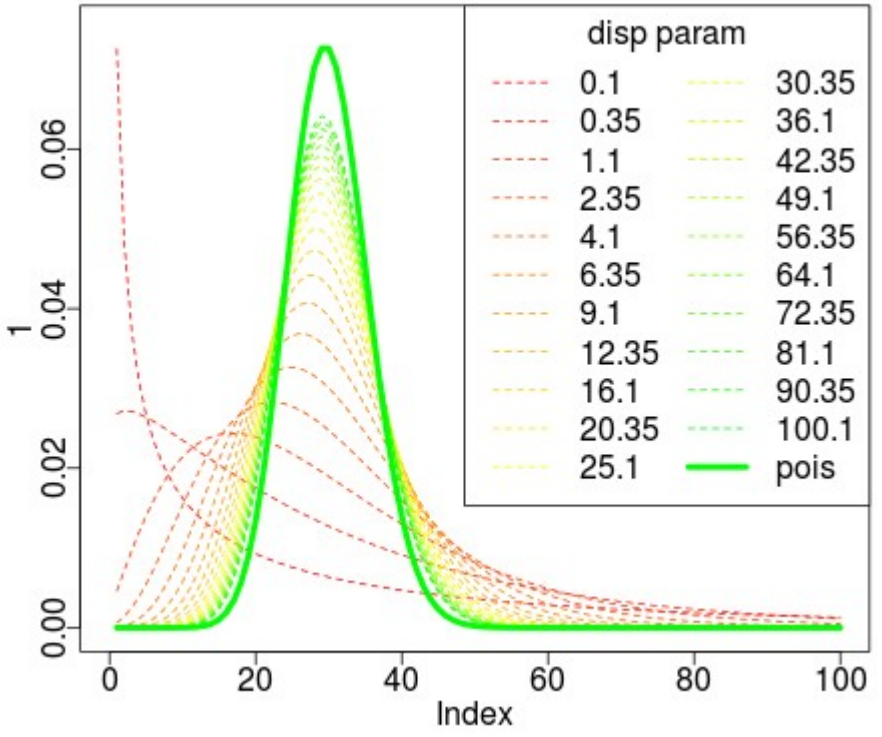
Альтернативная параметризация

m, r

$$mean = m$$

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

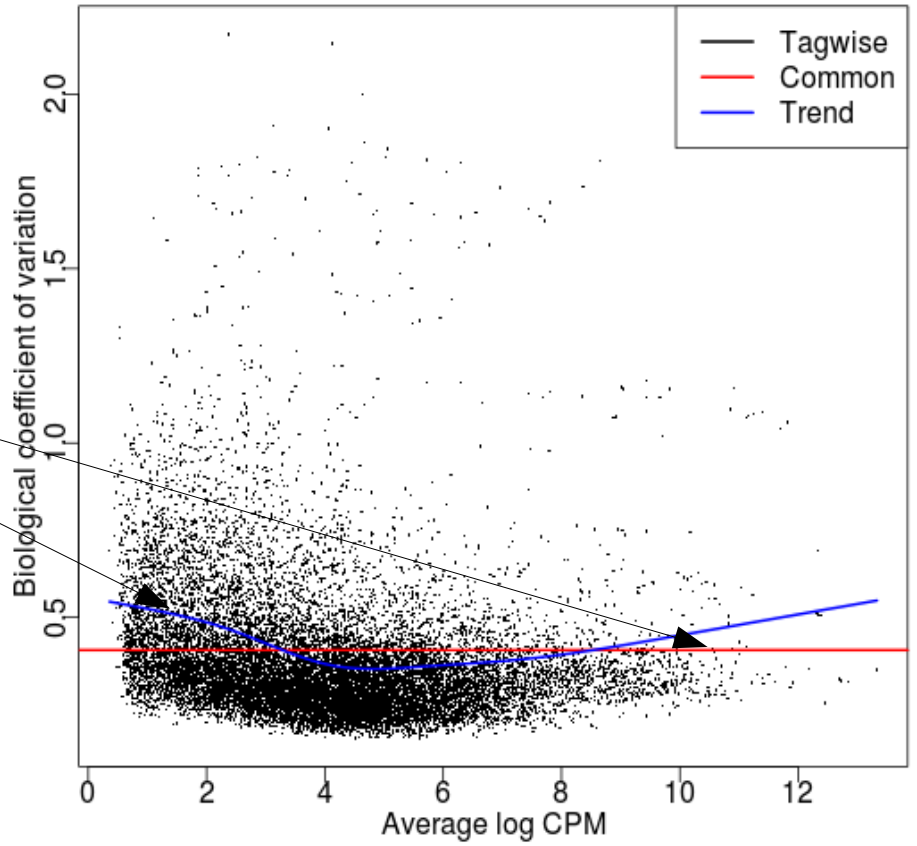
$$p = \frac{m}{m+r}$$



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.disp
ersion,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

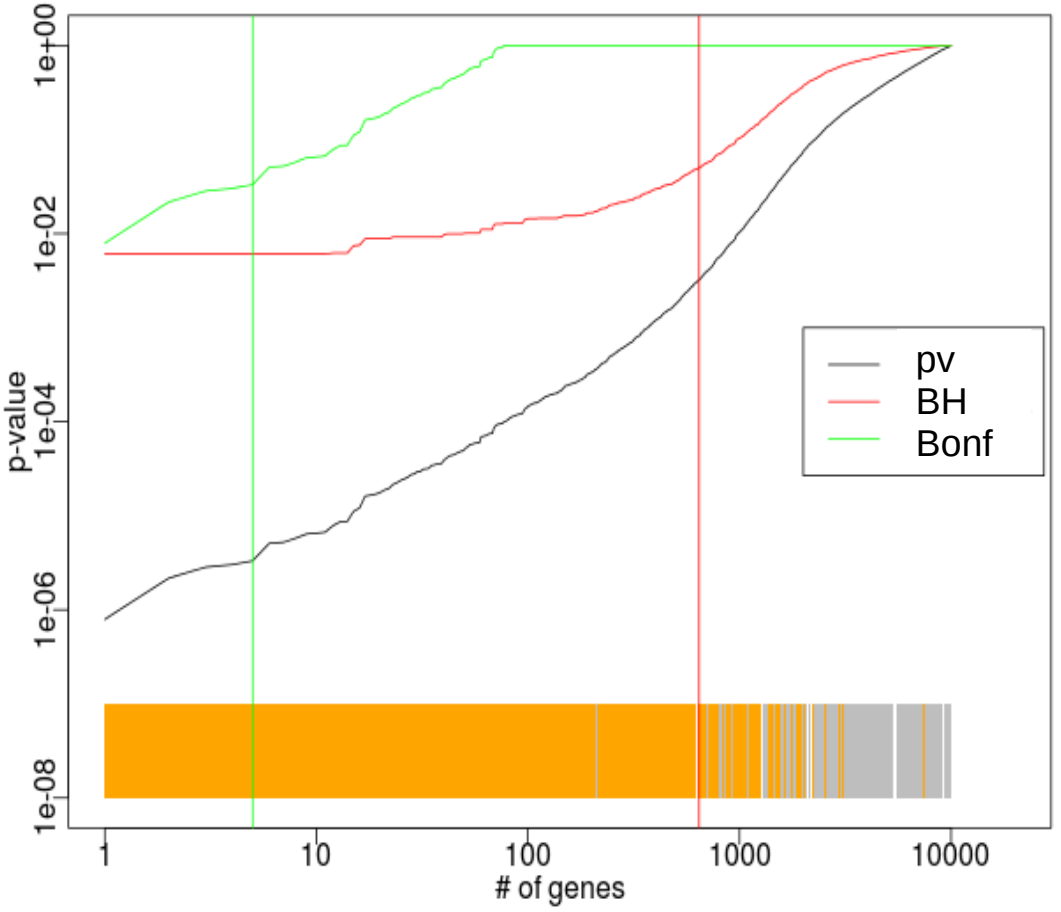
- поправка Бонферони: контролируем вероятность хоть одного ложного:

$$pv.corr_i = pv_i * N$$

- поправка Бенджамини-Хочберга: контролируем долю ложных

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

R: p.adjust

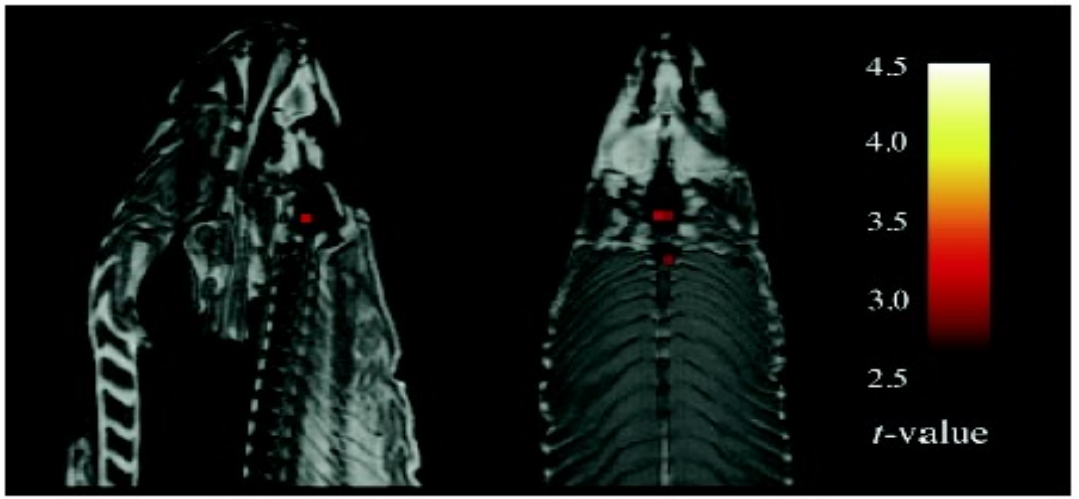


Помни о мёртвом лососе

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

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[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



Функциональная аннотация

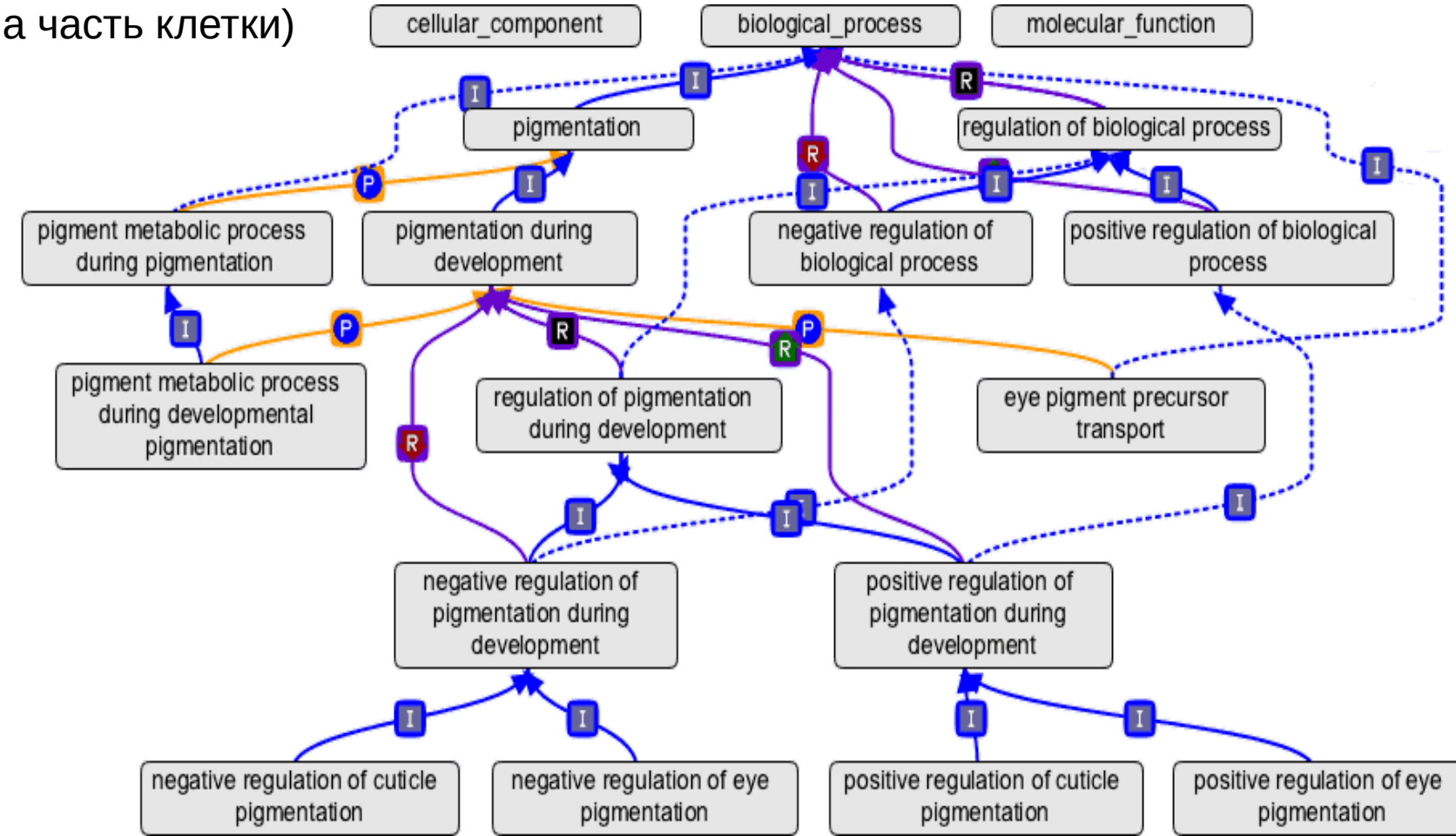
- **Augustus** - http://bioinf.uni-greifswald.de/augustus/ - предсказывает гены (CDS) только по геному. Может использовать RNA-seq
- **transdecoder** - https://transdecoder.github.io/ - находит CDS в транскриптах
- **interproscan** - https://code.google.com/p/interproscan/ - приписывает аминокислотным последовательностям (или нуклеотидным — находя в них ORF) семейства interpro (и, через них GO, и описание) а так же некоторые другие свойства: PANTHER (классификация по функциям путям, etc), gene3D (структурная классификация), Pfam и т. д.
- **blast2go** - www.blast2go.com - коммерческая (есть свободная версия) для функционального анализа аминокислотных последовательностей.

Gene ontology

Три классификации (онтологии) генов по:
- биологическому процессу (**B**iological **P**rocess)
- молекулярной функции (**M**olecular **F**unction)
- клеточная локализация (**C**ellular **C**ompartment)

Каждая онтология направленный, ациклический граф

Связи между категориями:
- is a (подтип, митохондрия это органелла)
- part of (мембрана часть клетки)
- regulate

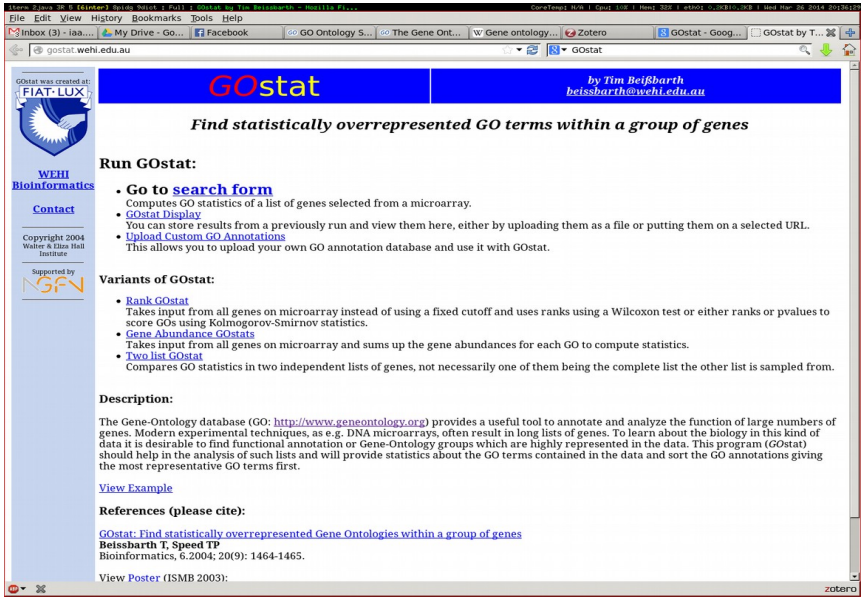


GO-enrichment

	Относятся к данной GO-категории	Не относятся к данной GO-категории
значимые гены	30	781
незначимые гены	42	4542

тест Фишера, поправка на множественное тестирования (для каждой категории)

http://gostat.wehi.edu.au/



Пакеты в R:

- topGO
- GOstat
- goseq

topGO

```
library(topGO)
options(stringsAsFactors = FALSE)
go = read.csv('~/skoltech/projects/evo.devo/input/GO/Homo_sapiens.GRCh37.74.GO.csv.gz')
go[1:2,]
# Ensembl.Gene.ID GO.Term.Accession
# 1 ENSG00000261657        GO:0006810
# 2 ENSG00000261657        GO:0016021
go = split(go$GO.Term.Accession,go$Ensembl.Gene.ID)
u = setNamesnames(go)
s = setNames(factor(as.integer(u %in% sample(u,1000))),u)
str(s)

# use topGO annotation
tgo1 <- new("topGOdata", ontology = "BP",
                    allGenes = s,
                    nodeSize = 10,
                    annotationFun = annFUN.org,mapping='org.Hs.eg.db',ID='Ensembl')

r1<- runTest(tgo1, algorithm = "classic", statistic = "fisher")
hist(score(r1))
table(p.adjust(score(r1),m='BH')<0.5)
GenTable(tgo1,r1,topNodes=10)
showSigOfNodes(tgo1, score(r1), firstSigNodes = 5, useInfo ='all')

# use external annotation
tgo2 <- new("topGOdata", ontology = "BP",
            allGenes = s,
            nodeSize = 10,
            annotationFun = annFUN.gene2GO,gene2GO=go)
r2<- runTest(tgo2, algorithm = "classic", statistic = "fisher")
showSigOfNodes(tgo2, score(r2), firstSigNodes = 5, useInfo ='all')
```

	fisher	ks	t	globaltest	sum
classic	✓	✓	✓	✓	✓
elim	✓	✓	✓	✓	✓
weight	✓	–	–	–	–
weight01	✓	✓	✓	✓	✓
lea	✓	✓	✓	✓	✓
parentchild	✓	–	–	–	–

- classic — каждая категория тестируется независимо
- elim — категории тестируются снизу, если категория значима, входящие в нее гены выкидываются из анализа родителей
- weight — если родительская и дочерняя категория значимы, то берется более значимая
- weight01 — что-то среднее между elim и weight
- parentchild — тест для данной категори выполняется внутри множется генов относящихся к родительской категории

- Fisher — обычный тест Фишера
- Ks — тест Колмогорова-Смирнова на scores (p-value)
- T — T-test на scores
-

Выбор фонового набора генов, эффект покрытия

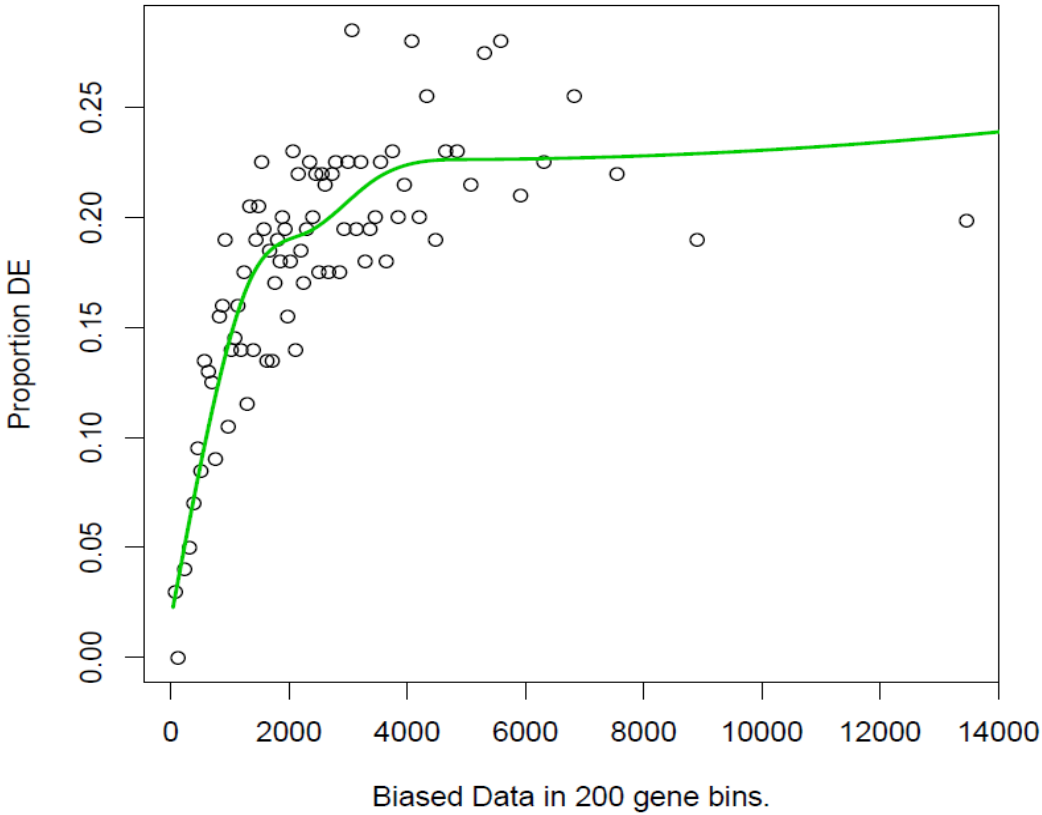
Проблема:

- Многие гены экспрессирующиеся в мозгу связаны с нервной системой. Любые гены меняющие экспрессию в мозгу в ответ на стресс будут обогащены категориями связанными с нервной системой: **надо выбирать правильный фоновый набор генов!**
- Гены с большим числом ридов (длинные и/или высоко экспрессирующиеся) имеют больше шансов получить низкое p-value. Это можно контролировать при помощи пакета goseq.

R

```
#deg — named vector. 1 and 0 for
#significant and not significant genes
t = nullp(deg,
        genome = 'mm10',
        id='ensGene',
        bias.data = tot.exp[names(clusters)])

go.clust[[i]]=goseq(t,
        genome = 'mm10',
        id='ensGene',
        method = 'Hypergeometric')
```



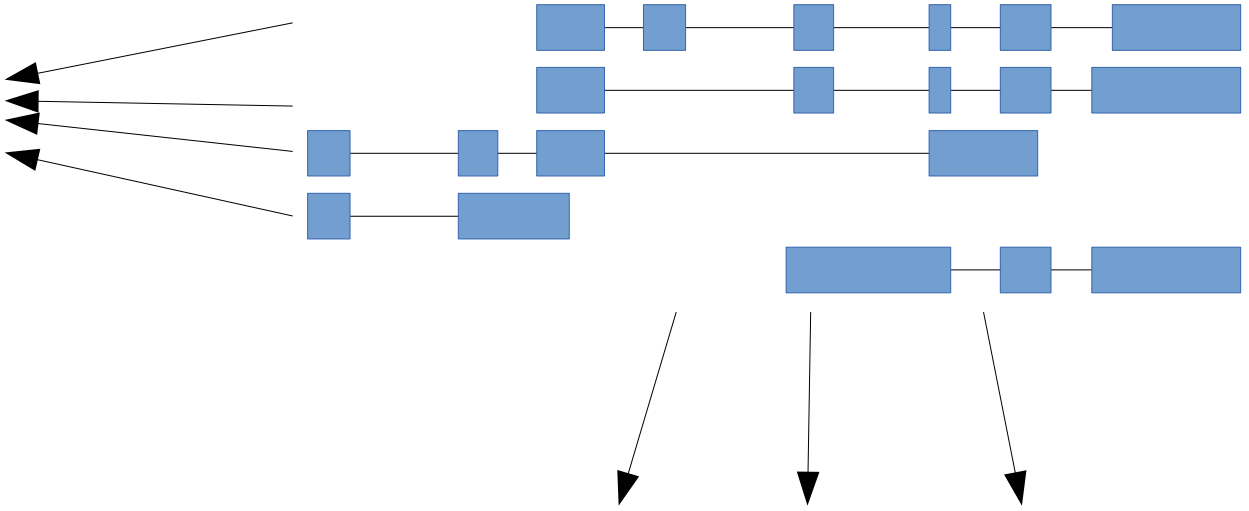
Альтернативный сплайсинг

РНК-сек и альтернативный сплайсинг

Идеальный ген в вакууме: непрерывный, один ген — один транскрипт

Настоящий эукариотический ген: набор [плохо] перекрывающихся транскриптов

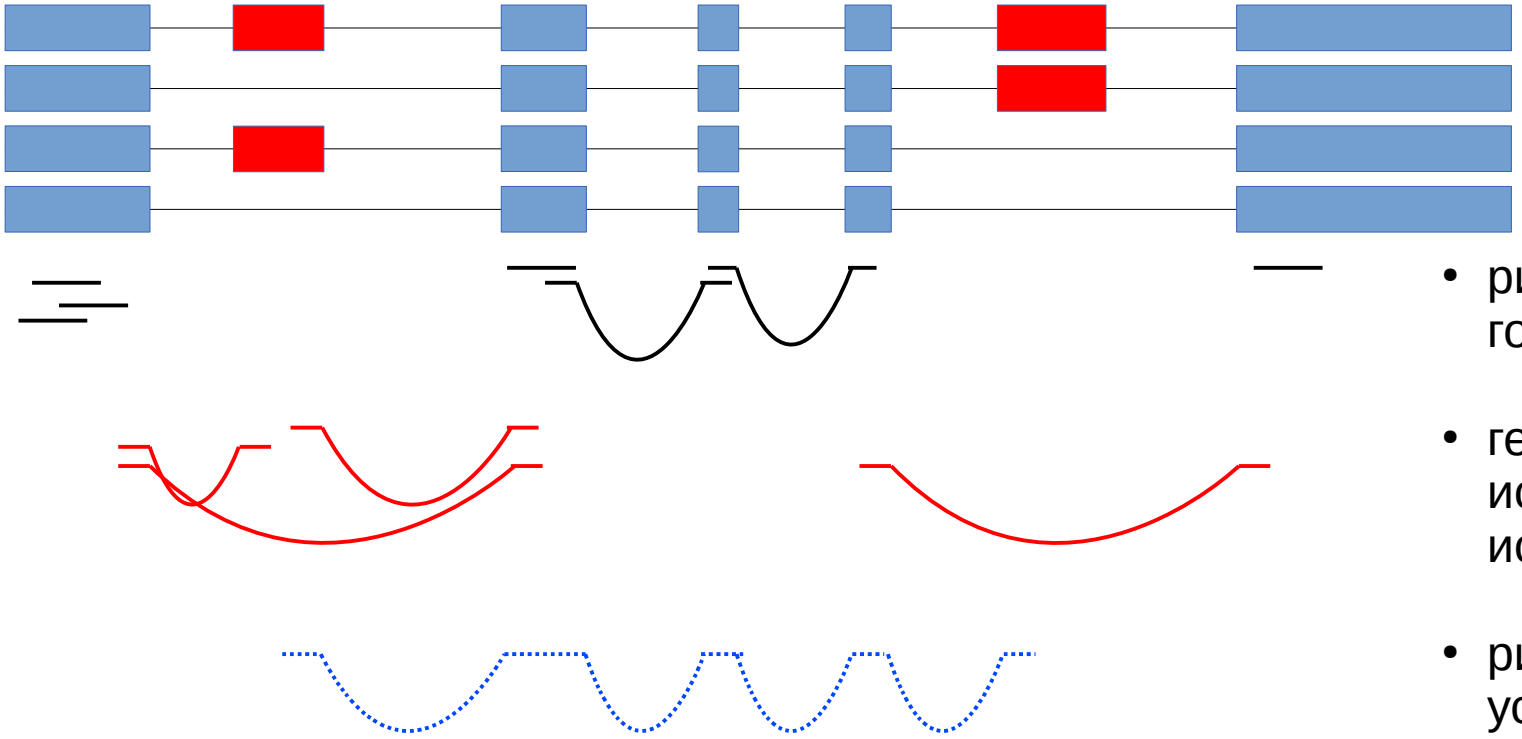
транскрипт-центричный подход



экзон/сайт-центричный подход

Транскрипт или экзон?

- Транскрипт является биологическим объектом с некоторой (может быть известной) функцией. Именно транскрипт кодирует белок.
- Однако данные РНК-сек в общем случае не позволяют установить транскрипт:



- риды ничего не говорящие о АС
- гены говорящие о использовании/не использовании экзона
- рид, который мог бы установить транскрипт

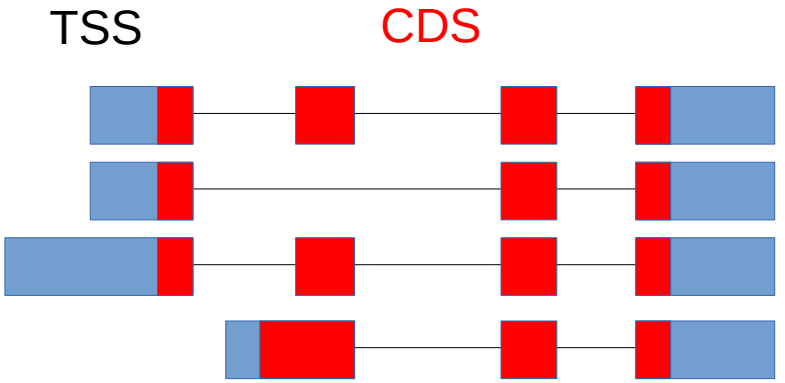
Cuffdiff: транскрипт-центричный подход

Подготавливаем аннотацию:

```
cuffcompare -o cuff -CG -r input.gtf output
```

Проводим сравнение:

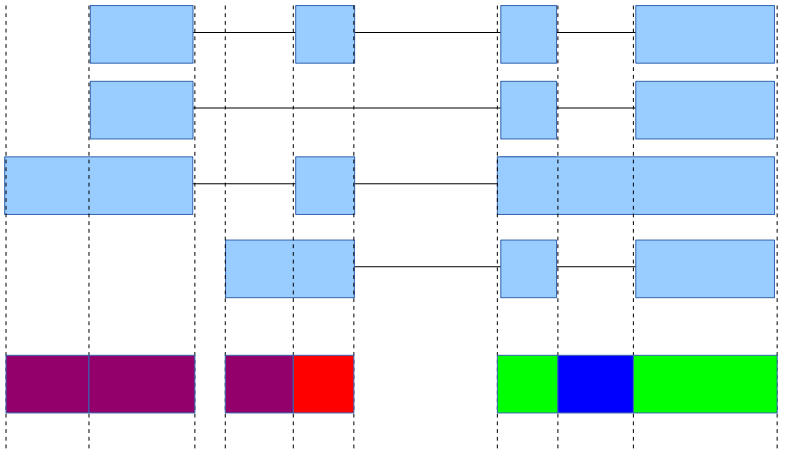
```
cuffdiff
    --min-reps-for-js-test 1
    output.combined.gtf
    sam1.rep1.bam,sam1.rep2.bam,...
    sam2.rep1.bam,sam2.rep2.bam,...
```



Получаем на выходе:

- FPKM и числа ридов для
  - генов
  - изоформ
  - CDS
- результаты дифф экспрессии для:
  - генов
  - изоформ
  - CDS
  - TSS
- Изменение относительных представленностей
  - AC-изоформ (для одного TSS)
  - Альтернативных TSS
  - CDS одного гена

DEXseq



Экзоны

Сегменты
(counting bins)

Альт. TSS

Альт.
сплайсинг

Удержанный
интрон

Константные
экзоны

GLM, НБ распределение

$$\mu_{s,g,b} \sim \beta^{G}_{g} + \beta^{S}_{b} + \beta^{DE}_{s,g} + \beta^{AS}_{s,b}$$

- **G:** средняя экспрессия гена
- **S:** среднее включения сегмента
- **DE:** изменение экспрессии гена в данном образце
- **AS:** изменение включения сегмента в данном образце

DEXseq: как пользоваться

Готовим аннотацию

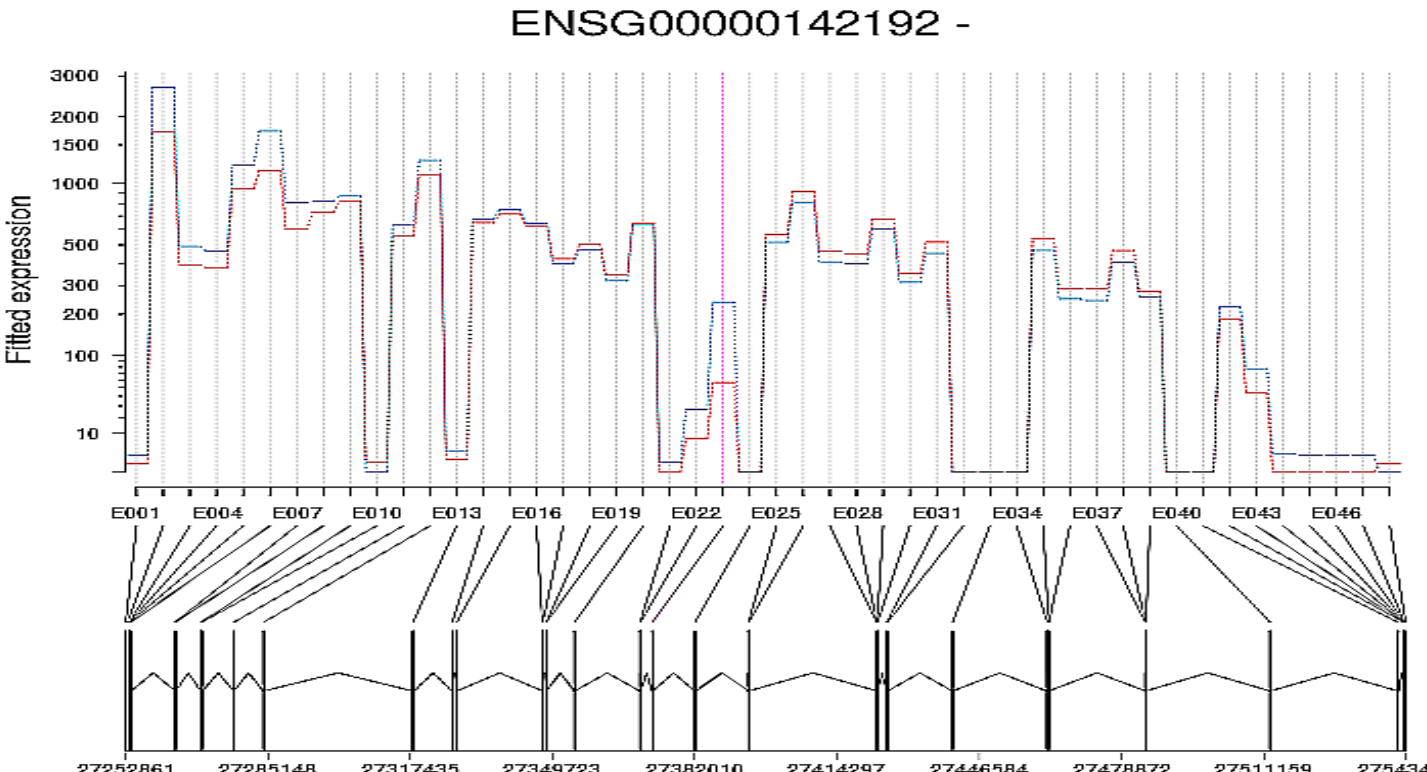
```
python dexseq_prepare_annotation.py input.gtf output.gff
```

Считаем риды в каждом сегменте

```
samtools view in.bam | python dexseq_count.py
    -p no
    -s no
    output.gff
    -
    out.txt
```

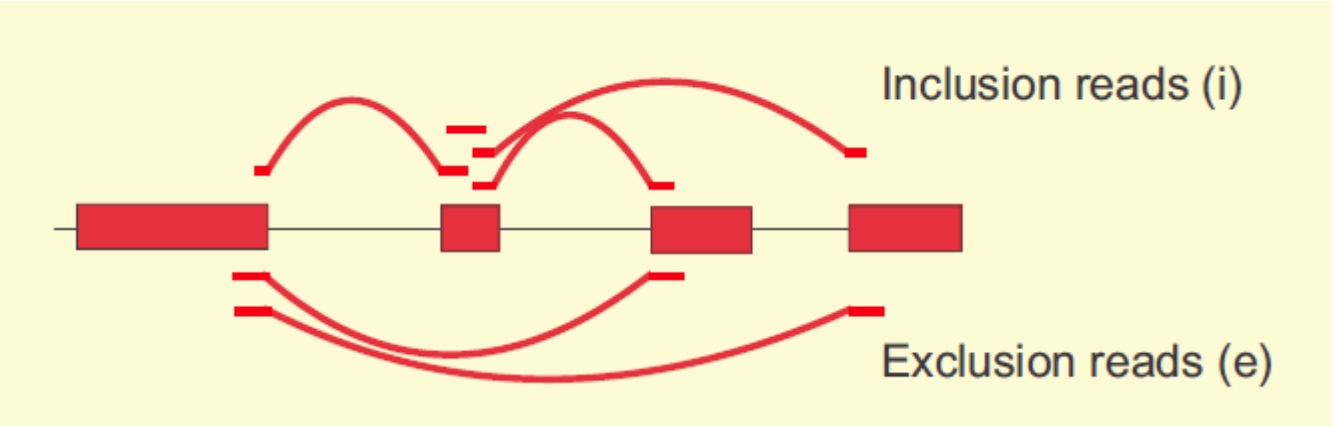
Загружаем риды в R, и проводим статистический анализ

estimateSizeFactors
estimateDispersions
fitDispersionFunction
testForDEU
DEUresultTable



SAJR

http://storage.bioinf.fbb.msu.ru/~mazin/index.html



$$\Psi = \frac{\frac{i}{(ls + lr - 1)}}{\frac{i}{(ls + lr - 1)} + \frac{e}{(lr - 1)}}$$

(i,e) ~ биномиальное распределение
GLM, quasibinomial
Готовим аннотацию

```
java -jar sajr.jar
    gff2sajr
    -ann_foreign=input.gtf
    -ann_out=output.gtf
```

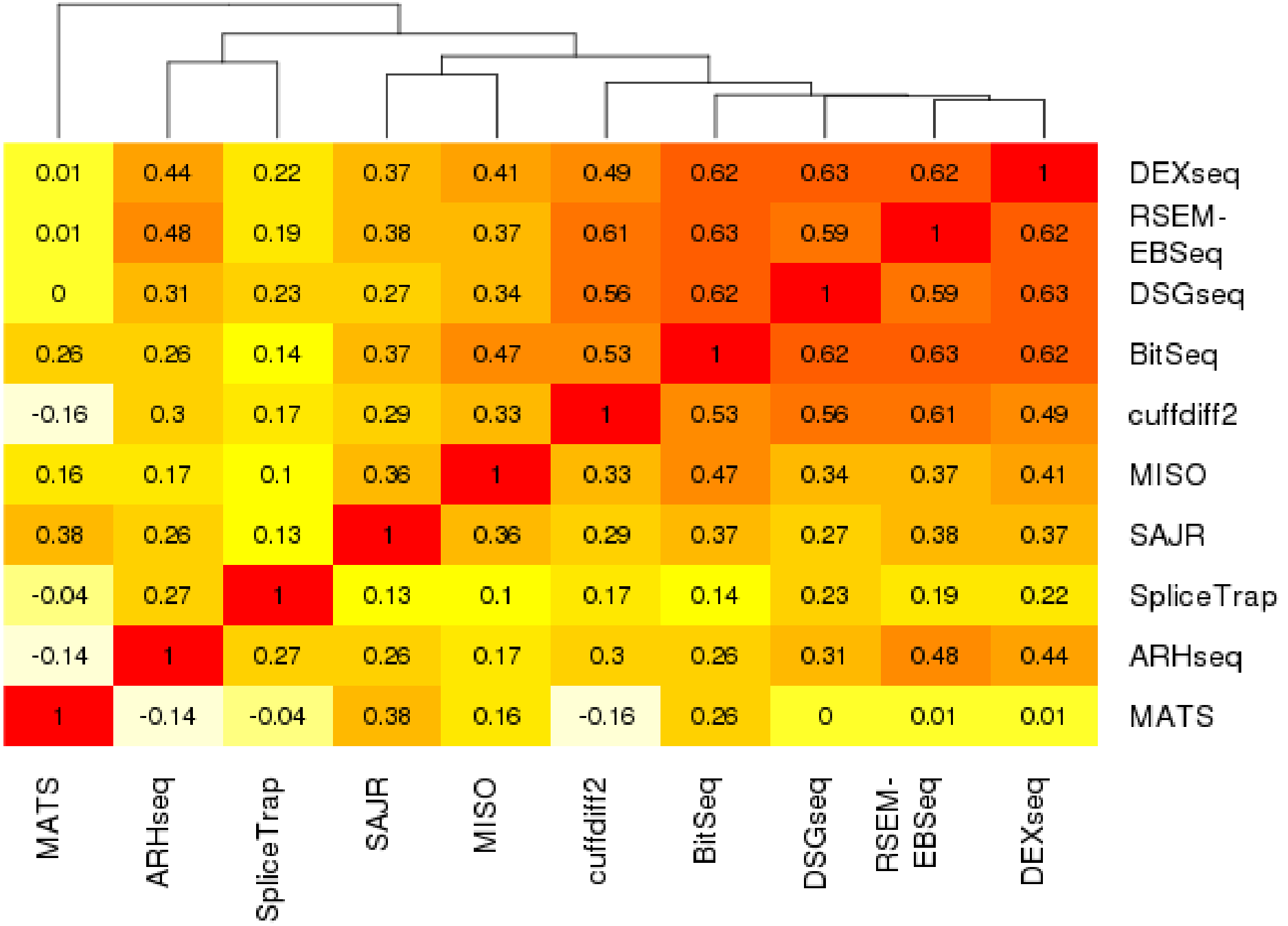
Считаем риды

```
java -jar sajr.jar
    count_reads
    -batch_in=1.bam,2.bam,...
    -batch_out: 0,1,2,3
    -paired=0
    -ann_in=output.gtf
    -effective_read_length=100
```

Работа в R

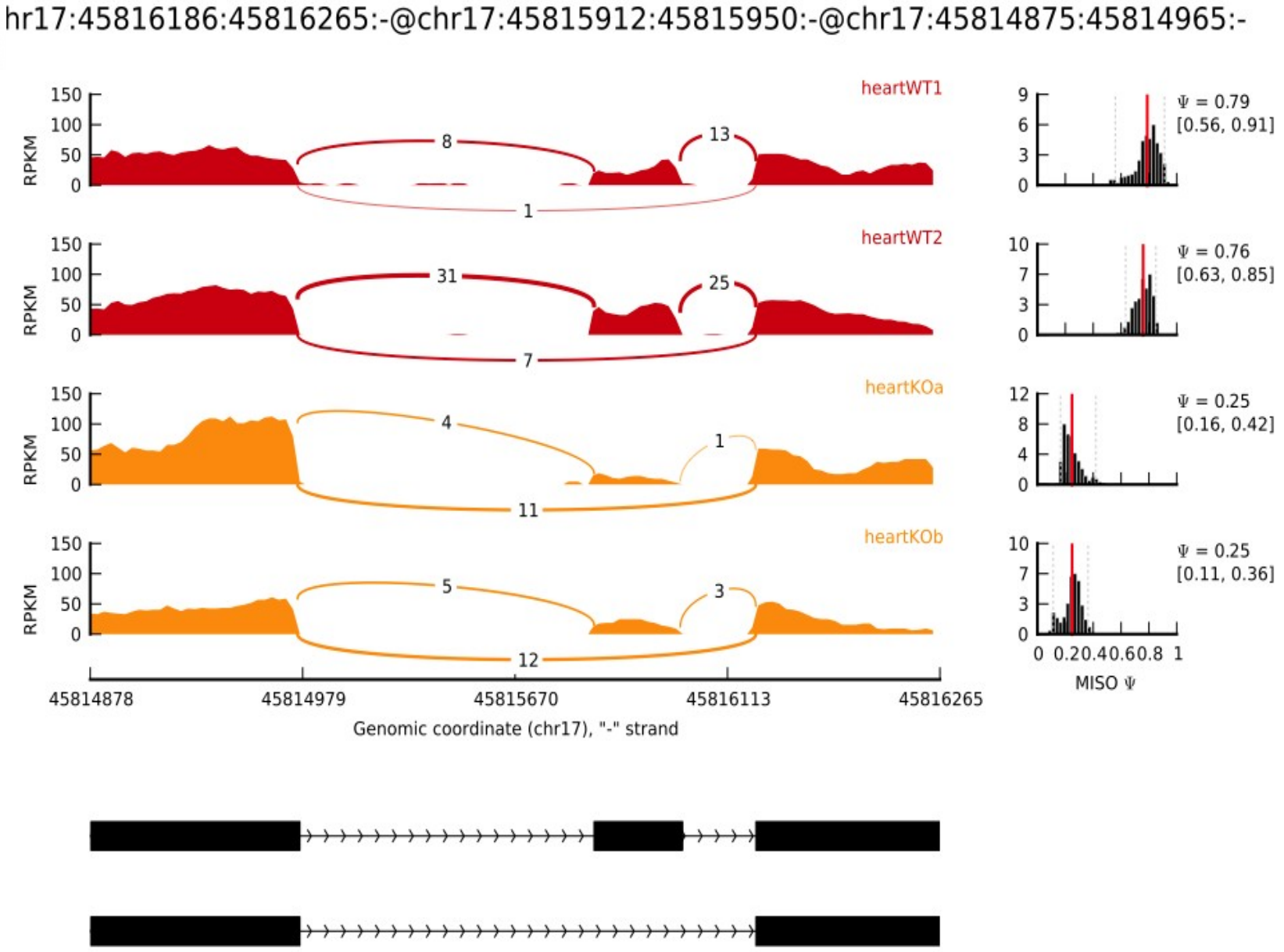
- loadSAData
- fitSAGLM
- calcSAPvalue

Результаты этих методов имеют мало общего

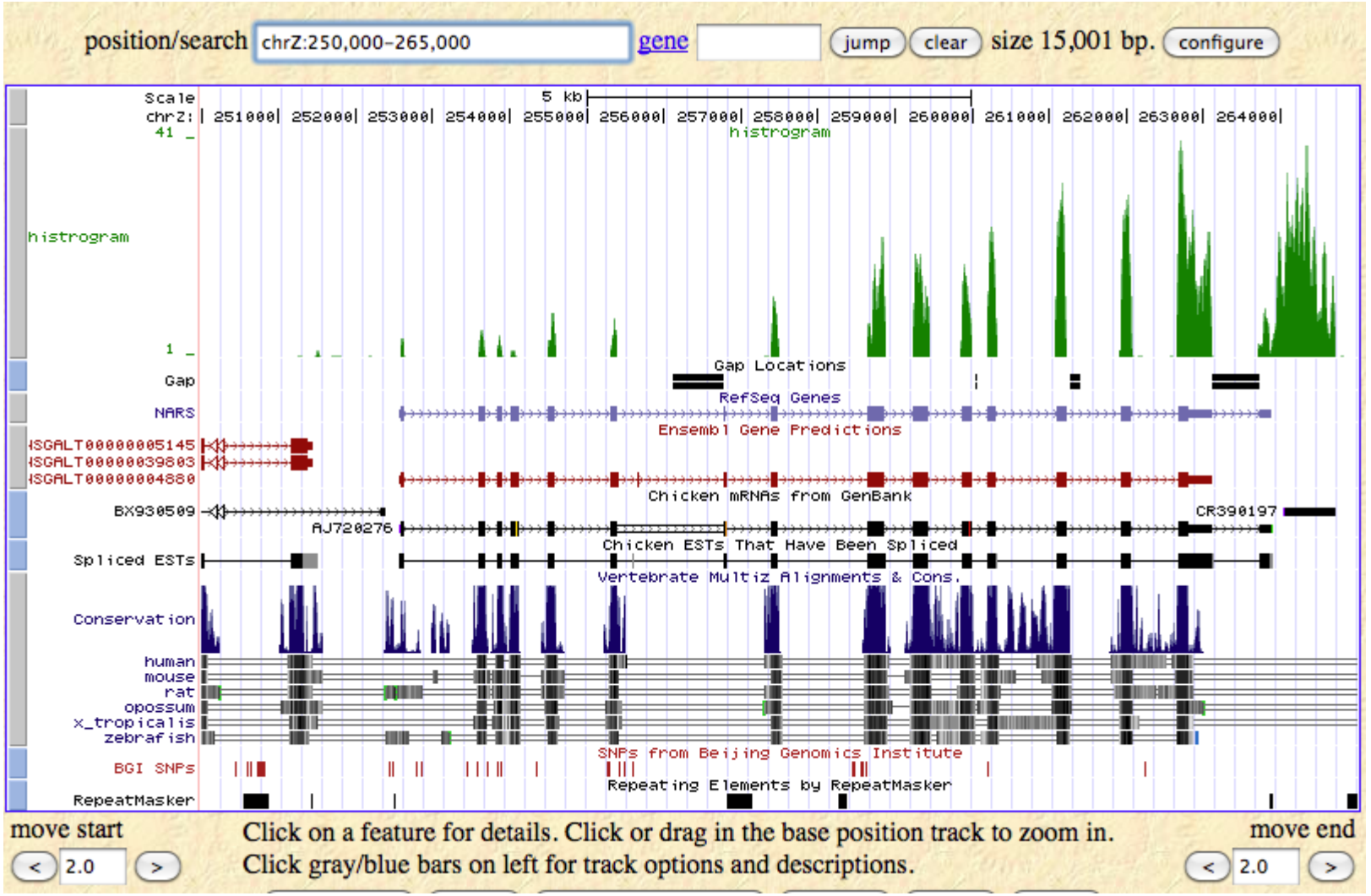


MISO: sashimi plot

http://genes.mit.edu/burgelab/miso/

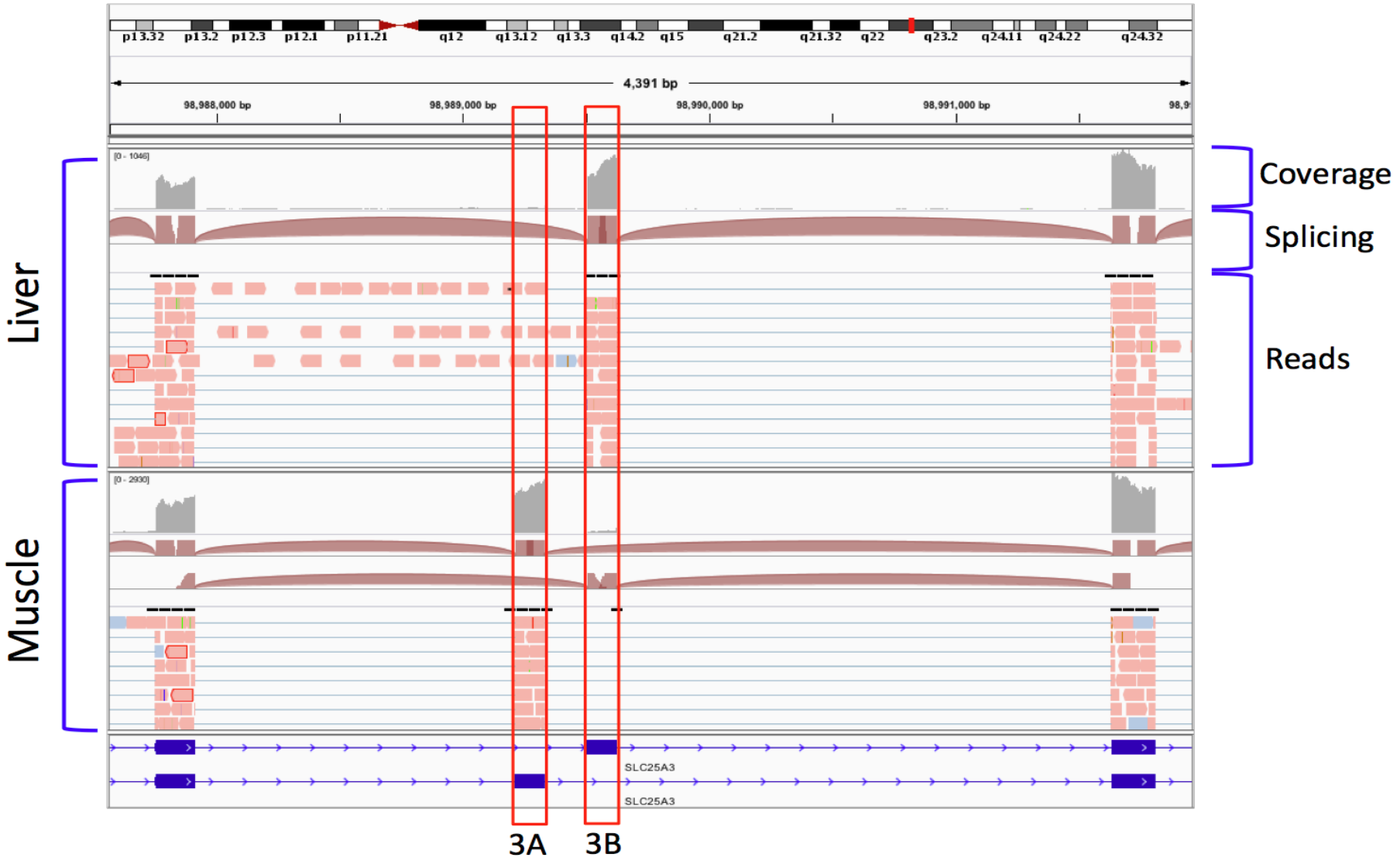


GenomeBrowser: просто сделай Bed[Graph]



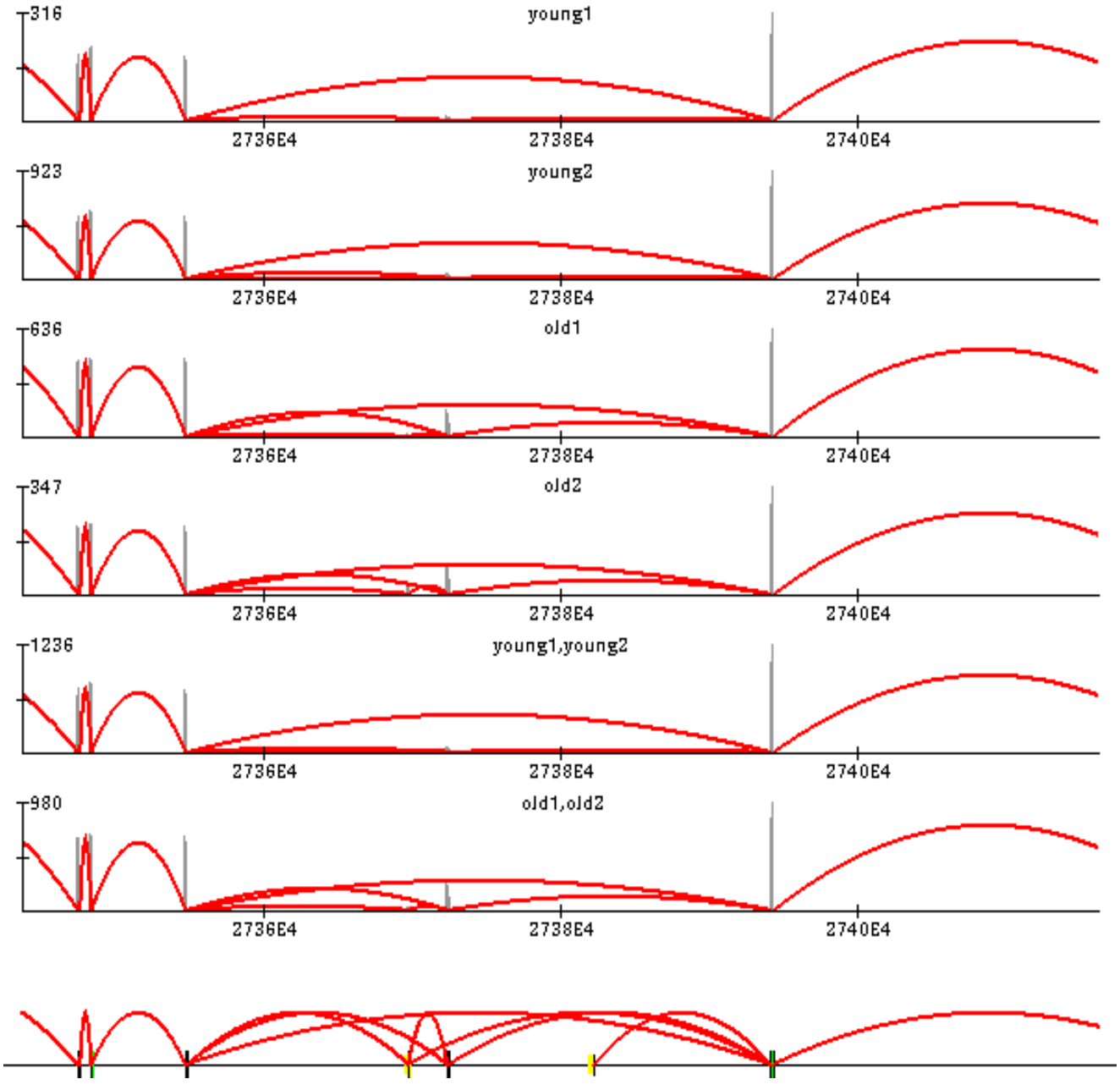
IGV

http://www.broadinstitute.org/igv/home



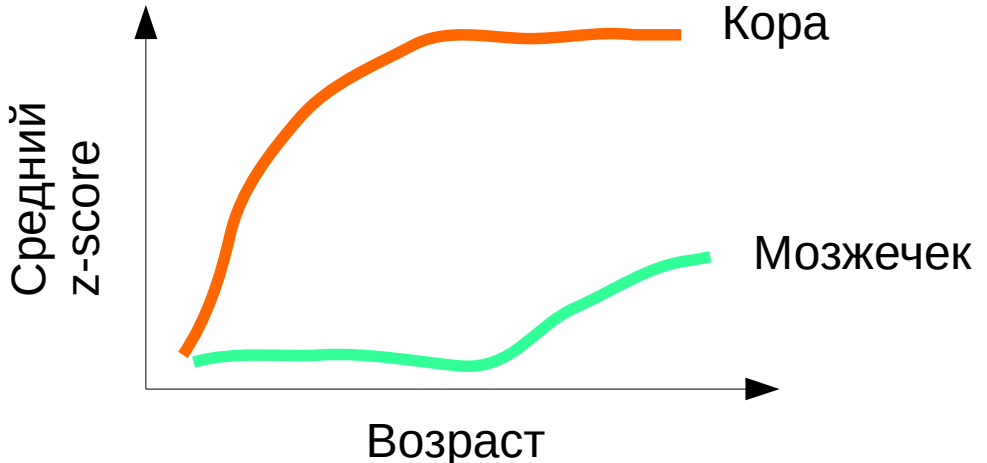
SAJR: SJV

http://storage.bioinf.fbb.msu.ru/~mazin/downloads.html
java -jar sjv.jar show



Домашнее задание

- Используя данные из предыдущего ДЗ при помощи edgeR найти гены с межтканевыми и/или возрастными изменениями экспресии (корректированное p-value < 0.05, межтканевые отличия должны быть не менее чем в два раза), возрастные изменения можно считать линейными по возрасту. Используйте модель ~ tissue + age
- Скалстеризуйте гены значимые хотя бы по одному фактору при помощи иерархической кластеризации (расстояние 1 — коэффициент корреляции Спирмана) в 6 кластеров.
- Отшкалируйте экспрессию каждого гена к среднему ноль и дисперсии один (z-score). Нарисуйте для каждого кластера зависимость среднего z-score от возраста для обоих тканей



Материалы к лекции (R - код)

- Код в R иллюстрирующий PCA, MDS, heatmap, lm/glm и anova http://rpubs.com/iaaka/14464
- Код в R иллюстрирующий edgeR, DEXseq и limma, GO-анализ, диаграммы Венна http://rpubs.com/iaaka/14560
- http://rpubs.com/iaaka/260288 — кластеризация
- http://rpubs.com/iaaka/14859 - дифф. сплайсинг с DEXseq