План курса

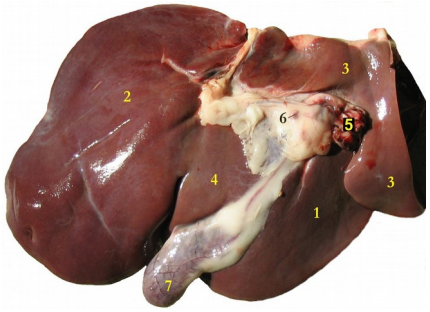
- Лекция 1
  - Введение, картирование, сборка транскриптома, и подсчёт транскриптомных ридов
  - Проверка самосогласованности: корреляционная тепловая карта, PCA/MDS
- Лекция 2
  - Нормализация
  - lm/ANOVA; glm/ANODEV;
  - Дифф. экспрессия (edgeR)
  - Кластеризация (hclust, k-means, PAM)
  - Функциональный анализ (goseq)
- Лекция 3
  - Дифф. сплайсинг (cuffdiff, DEXseq, MISO, SAJR)
  - Визуализация

Дедлайн по всем ДЗ блока РНК-Сек наступает через 4 недели после занятия, на котором было дано ДЗ

Транскриптомика

- Анализ клеточной РНК (полученной в результате ДНК-зависимого синтеза)

Образец

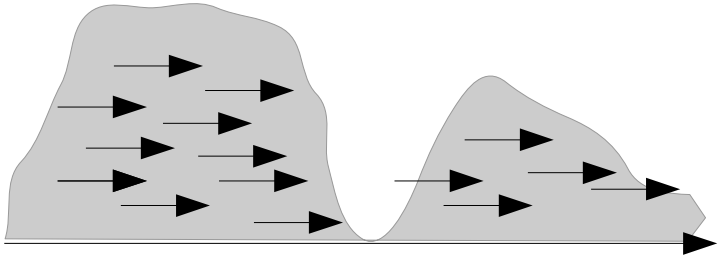


Выделение РНК

Секвенирование

```
read1 ATTAGGCTAAGTTAGGT
read2 GTAATCCCGCCGGGAGG
read3 TCCCTGGAGGACGATGC
read4 GCTTAGCTTGAAATTTC
```

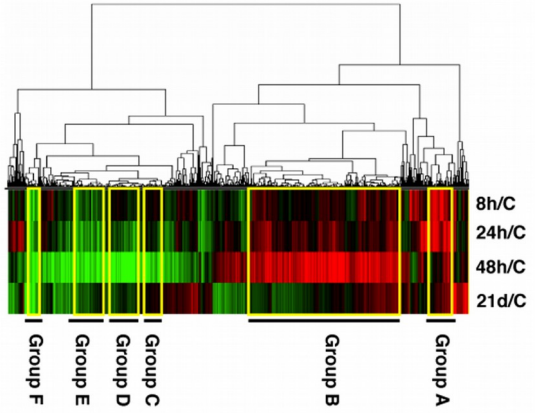
Картирование на геном



	sample1	sample1
gene1	124	78
gene2	171	63
gene3	126	87

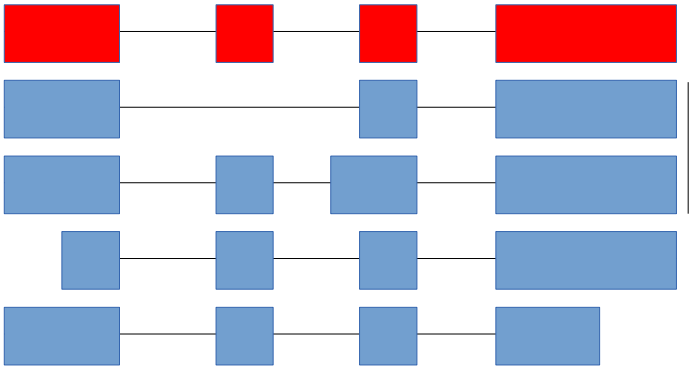
Подсчёт ридов

Статистический анализ



- Задачи транскриптомики:
  - определение структуры (нт последовательности молекул РНК)
  - определение концентраций молекул РНК
  - **сравнение концентраций в нескольких образцах**

Структура [м]РНК



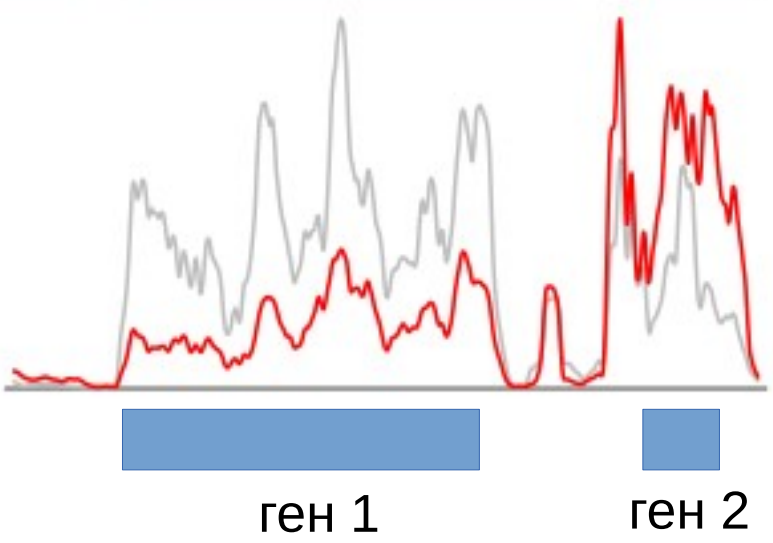
Альтернативный сплайсинг

+ RNA-editing

Альтернативный промотор (CAGE)

Альтернативный терминатор (полиА риды)

Изменения экспрессии

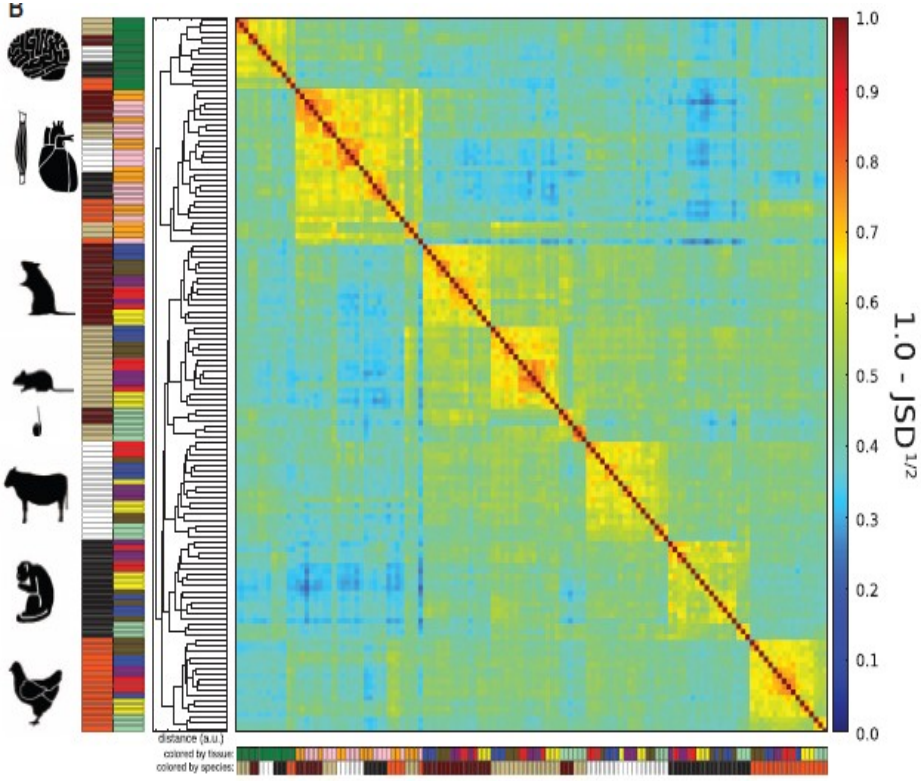


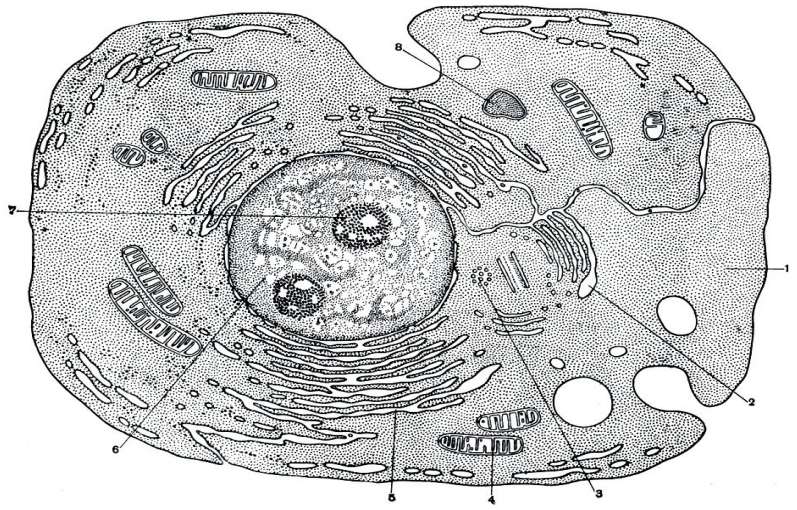
контроль

тепловой шок

Объект исследования

- Биологические отличия
  - биологический вид
  - орган (ткани)
  - состояние (болезнь, воздействие)
- Фракция РНК
  - общая (все подряд)
  - по наличию полиА
  - CAP/5'-3p
  - удаление тРНК и рРНК
  - по размеру
  - по внутриклеточной локализации
    - ядро
      - нуклеоплазма
      - ядрышки
      - хромосомы
    - цитоплазма
      - свободная
      - связанная с рибосомой

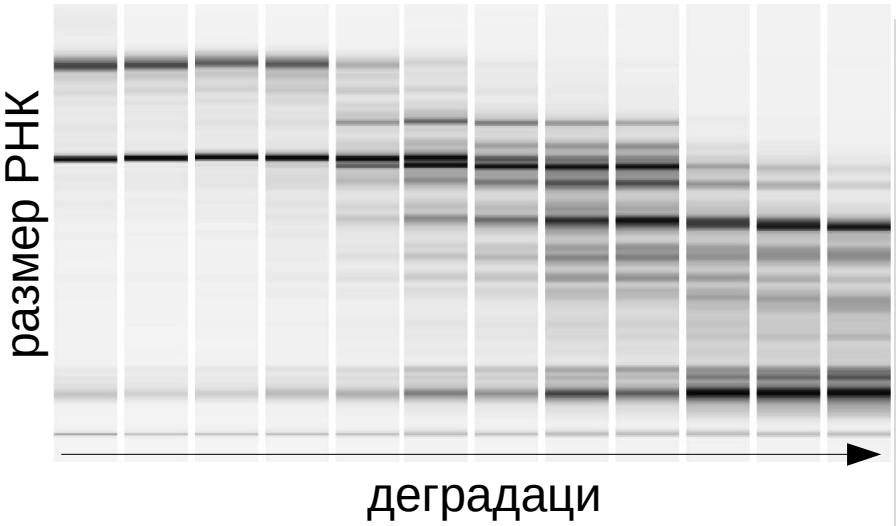


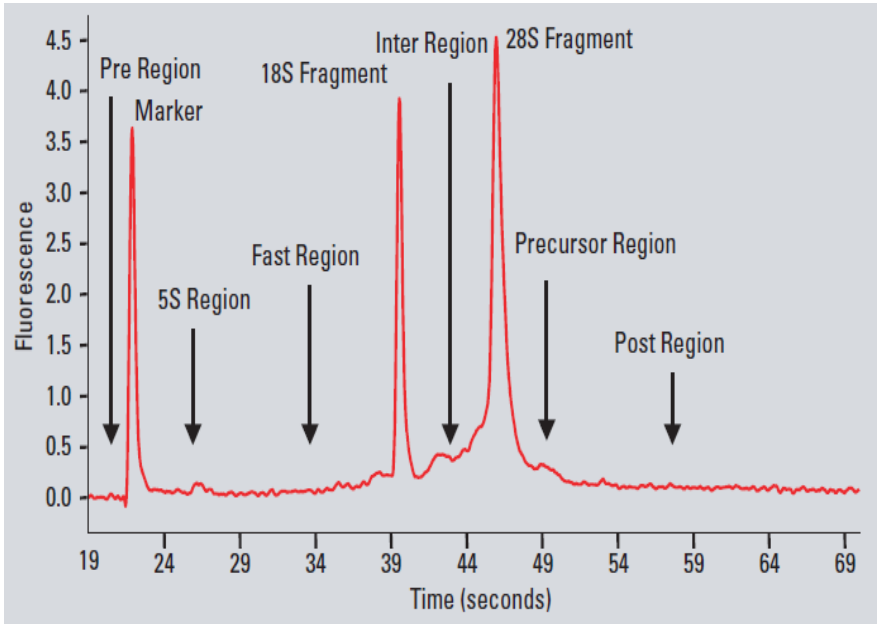


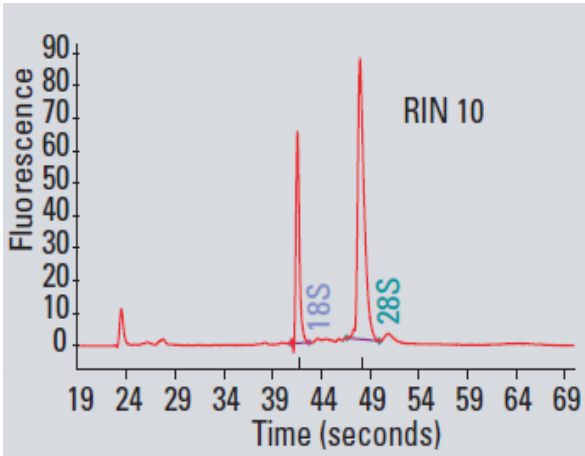
Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues
Jason Merkin *et al.*
Science **338**, 1593 (2012);
DOI: 10.1126/science.1228186

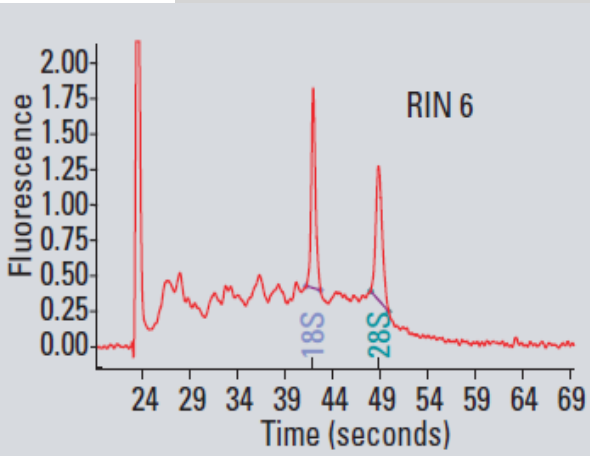
Качество РНК: RNA Integrity Number (RIN)

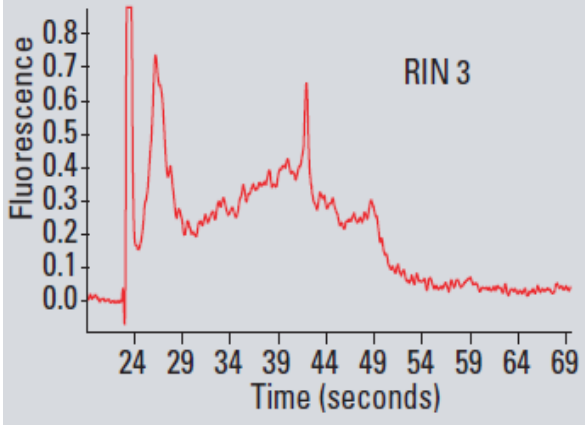
Agilent Technologies

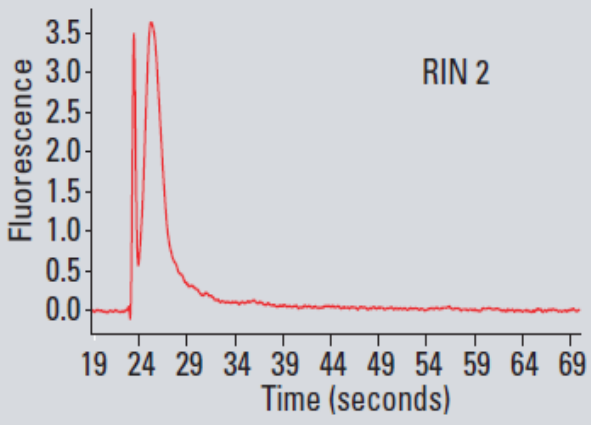






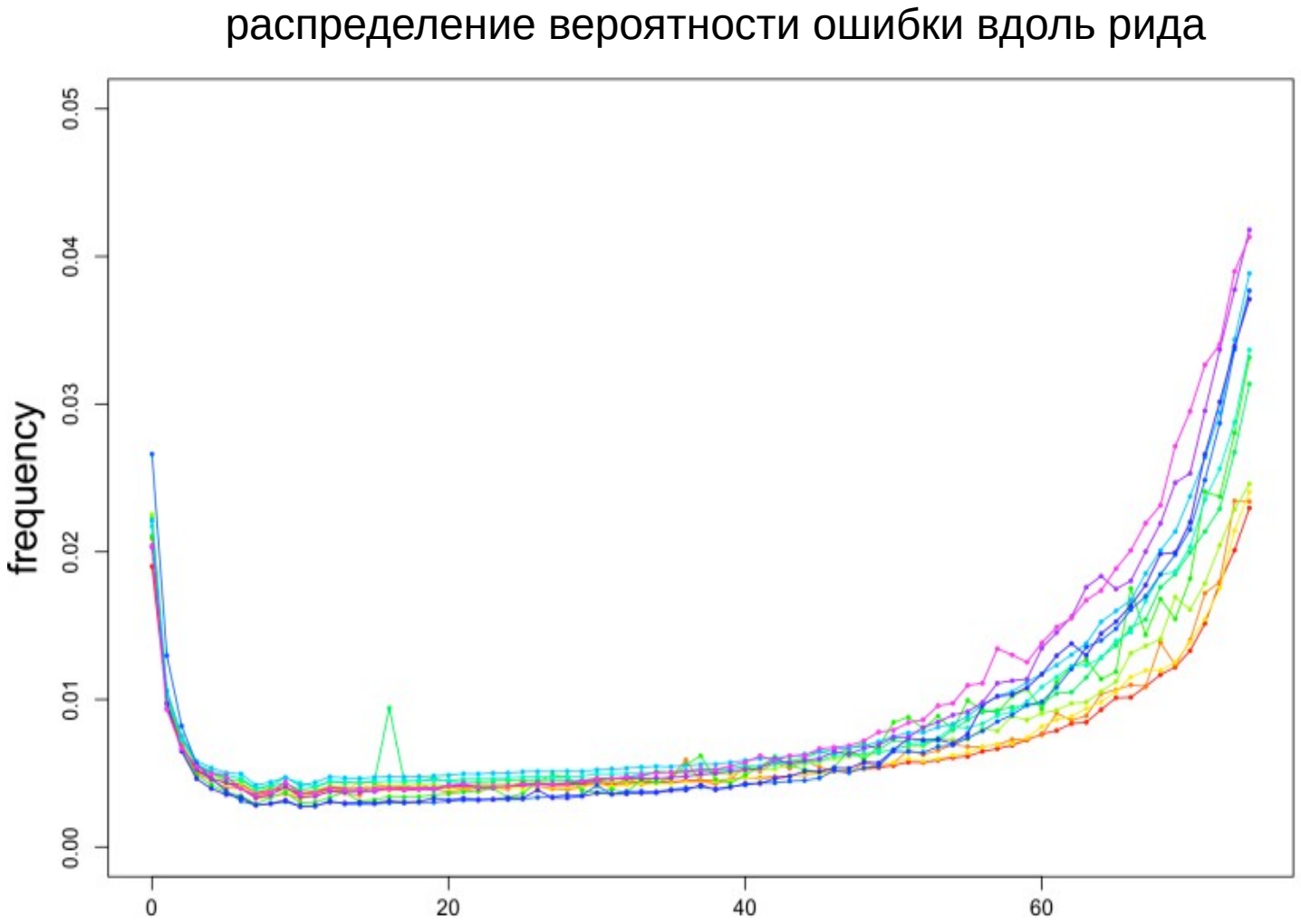






Результаты секвенирования

- Качество ридов
- цепь-специфичность
- Парность (SRA!)
- To trim or not to trim?
- SOLiD!



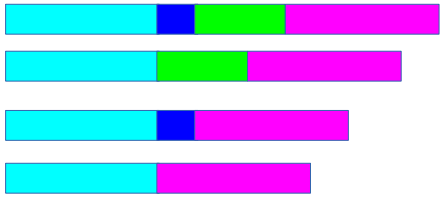
Practice: FastQC

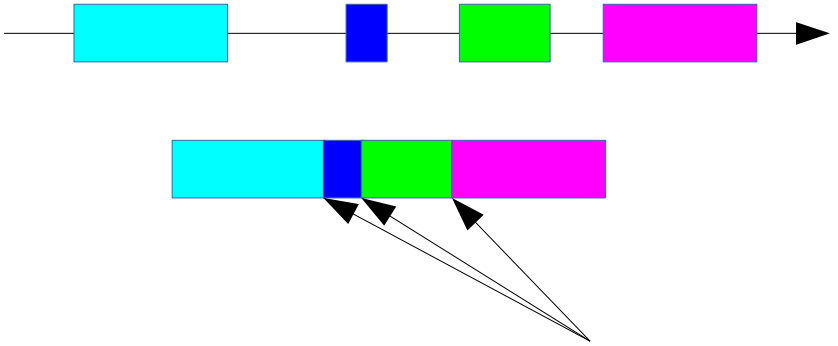
- Google for FastQC
- Download (wget) it to cluster and unzip
- Chmod +x FastQC/fastqc
- Run fastqc on any of /mnt/local/bioinf_labs/home/mazin2/fq/*gz
- Copy results to your computer (WinSCP, scp, etc)

Картирование ридов

Как картировать риды на экзон-экзонные границы?

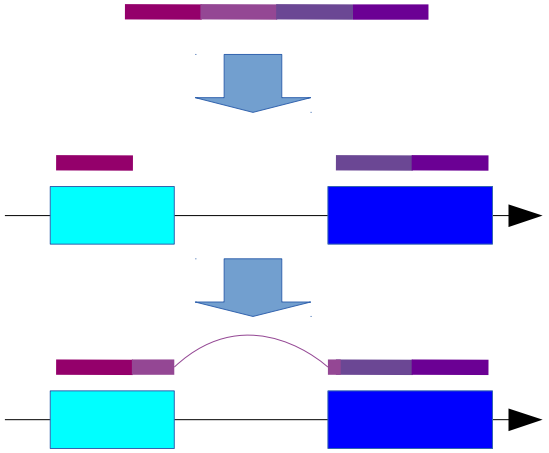
- использование аннотации
  - только аннотированные экзон-экзонные границы
  - все возможные экзон-экзонные границы





экзон-экзонная граница (junction)

- Предсказание аннотации из данных



разрезаем рид на несколько кусков и картируем на геном по отдельности

продолжаем куски и ищем положение разрыва — потенциальный интрон

находим интрон

Hisat2

- Замена tophat2
- Очень быстрый (~15*16 мин*ядер на 80 млн ридов)
- Может включать полиморфизмы в референс
- Как использовать:
  - построить индекс (hisat2-build), [используя доступную информацию о генах]
  - картировать fq.gz на индекс

Hisat2, подробности

- Скрипты **hisat2_extract_splice_sites.py** и **hisat2_extract_exons.py** позволяют извлечь информацию о экзонах и сайтах сплайсинга из аннотации
- Которую надо передать **hisat2-build** при помощи параметров **--ss** и **--exon.** Это правильно, но тогда **hisat2-build** будет требовать до 200G при построении индекса. Альтернатива - передать координаты сайту самому hisat2 при помощи параметра **--known-splicesite-infile**
- Параметры **hisat2**
  --dta-cufflinks — искать интроны более строго и добавить атрибут XS необходимый для работы cufflinks
  --no-unal не печатать невыровненные риды
  -x путь к индексу
  -U путь к ридам
- hisat2 по умолчанию печатает риды в стандартный выход. Чтобы записать их в файл надо указать параметр -S (будет печатать sam-файл) или использовать пайпы:
  hisat2 …. | samtools view -Sb - > out.bam

Hisat2: scoring options

--mp MX,MN — границы штрафа за ошибку. Пропорционален качеству нуклеотида. **Default: MX = 6, MN = 2**

--sp MX,MN — штраф за обрезание рида **Default: MX = 2, MN = 1.**

--no-softclip — отключить обрезание

--np — штраф за N. **Default: 1**

--rdg/--rfg <int1>,<int2> - штраф за открытие и продолжение гэпа. **Default: 5, 3**.

--pen-… - Штрафы связанные с интронами

--score-min <func> - Минимальный допустимый вес выравнивания в зависимости от длины рида. **Default is L,0,-0.2.**

Функции бывают константные (C), линейные (L), корень (S), логарифм (G).

Еще есть :

- STAR
- gsnap
- tophat2 — медленный, заменен на hisat2

samtools

samtools view in.bam [chr:start-end] | less -S

-h, -H — print header

-S, -b — input sam, output bam

-f, -F — filter by flag

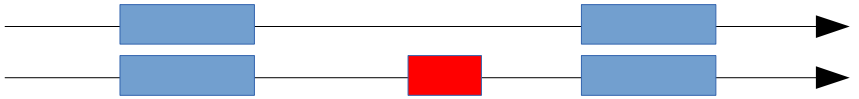
samtools sort

samtools index

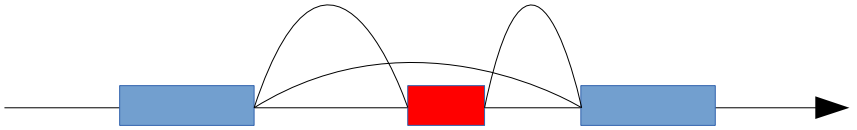
samtools faidx

Картирование: аннотация это важно!

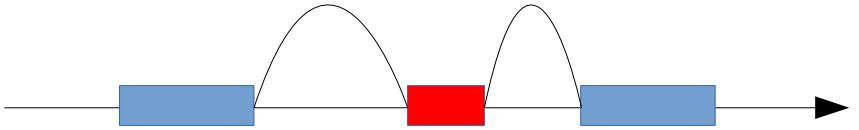
- эффективность предсказания экзон-экзонных границ зависит от глубины покрытия:



Хорошо покрытый образец: все экзон-экзонные границы нашлись



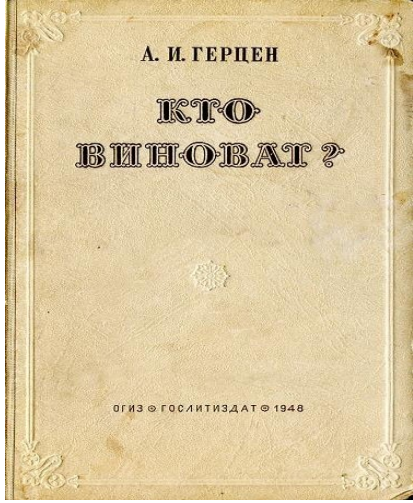
Плохо покрытый образец: один вариант потерялся



Ложные выводы:

Уникальный альтернативный сплайсинг?

Часть ридов не удаётся картировать → пониженная экспрессии?



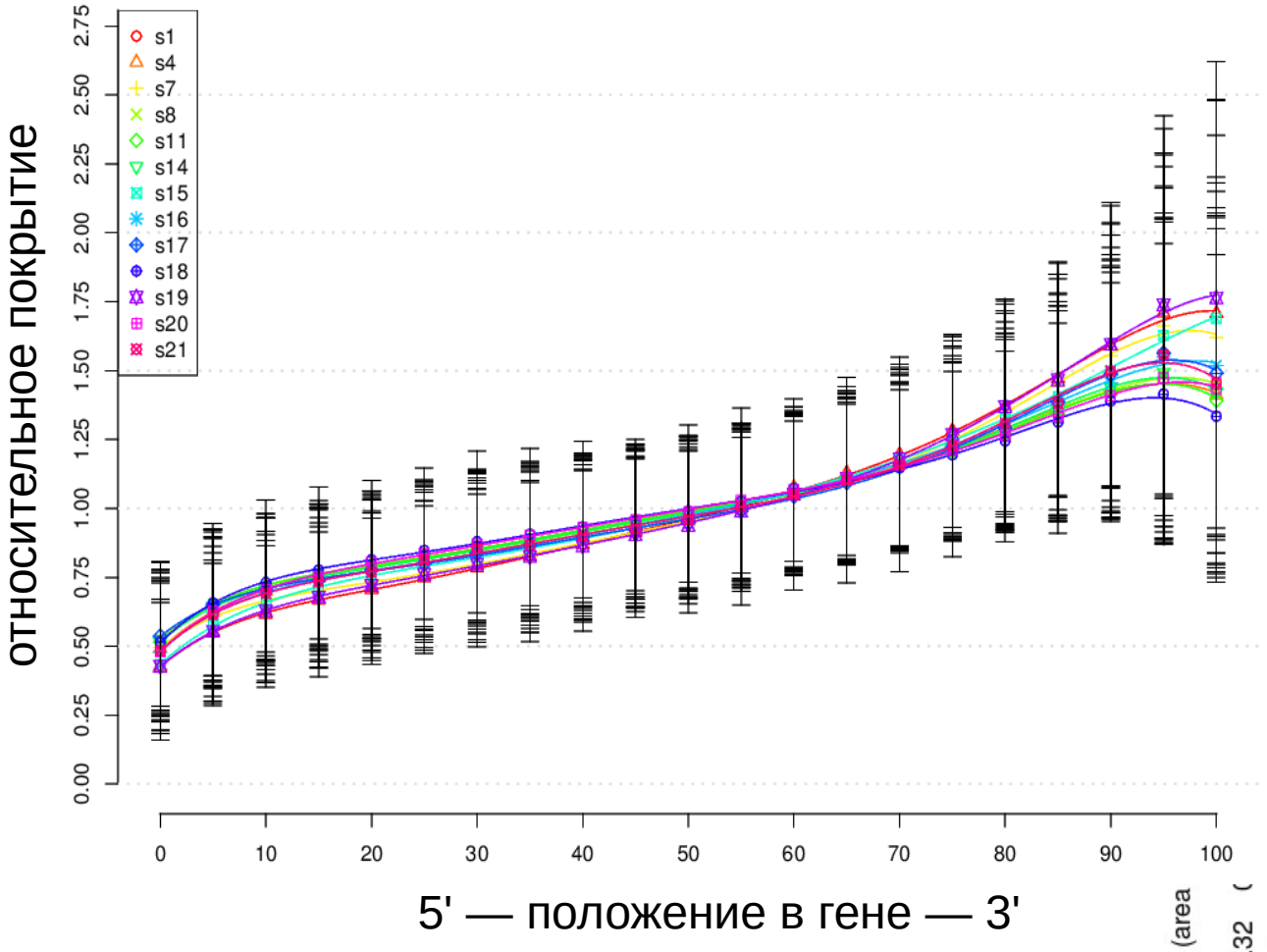
Что же делать?

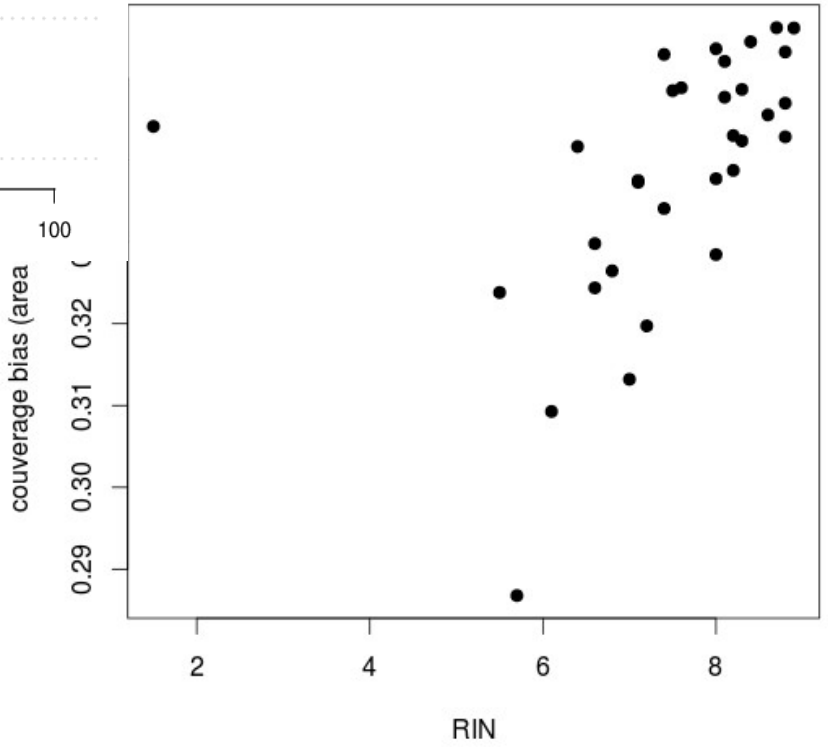
- По возможности картирование проводить с использованием аннотации. Возможно следует сделать два картирования.
- Использовать существующую аннотацию если возможно (человек — GENCODE)
- Использовать всю имеющуюся информацию (существующая аннотация, все образцы текущего исследования) для создания аннотации *de novo* (или улучшения):
  - картируем каждый образец предсказывая новые экзон-экзонные границы (hisat2, STAR, etc)
  - делаем аннотацию (stringtie, cufflinks, cuffmerge)
  - перекартируем образцы по одному.
- При сравнении нескольких видов аннотации должны быть унифицированы

Ошибки и проблемы картирования

- Неправильное картирование на экзон-экзоное соединение:
  - Маленькое перекрывание (overhang)
  - Только один вариант перекрывания (все риды картировались с одним overhang)
- Картирование на псевдогены вместо генов
  - Надо перекартировать
- Множественное картирование
  - Удалять по NH тагу в бам файле (большинство программ делают это сами)

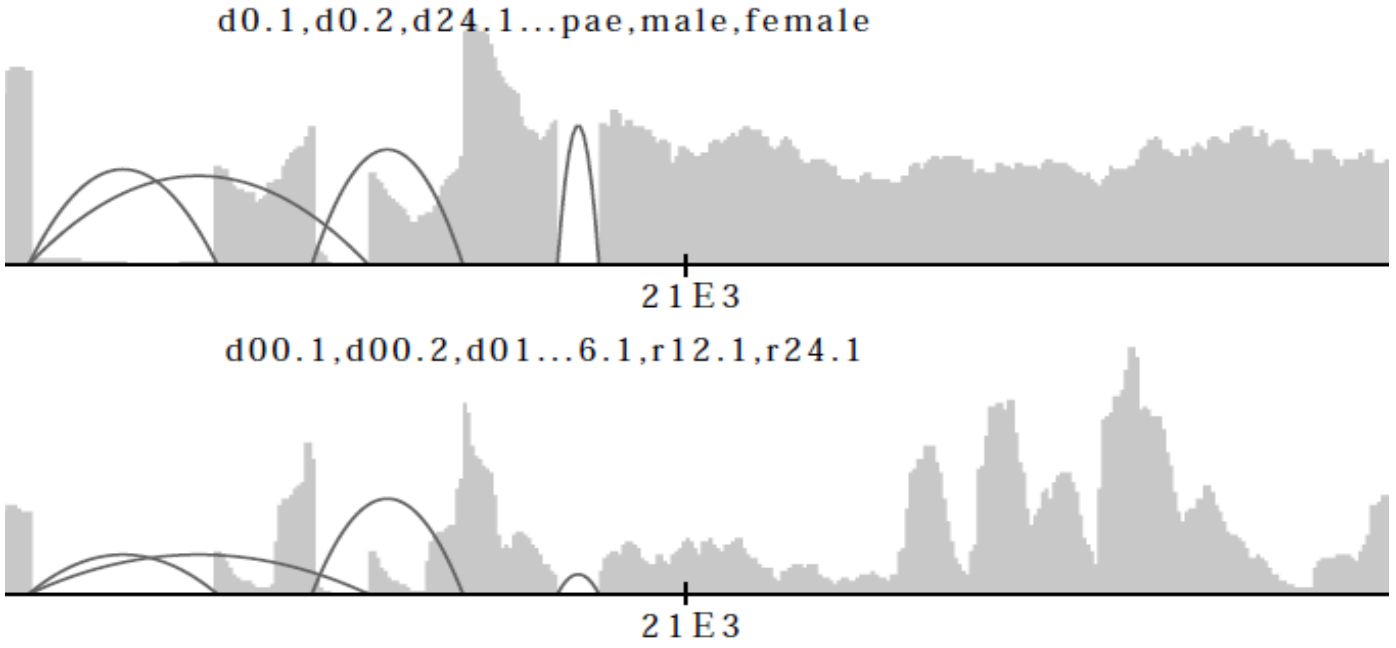
5'-3' перекос покрытия





Ровность покрытия

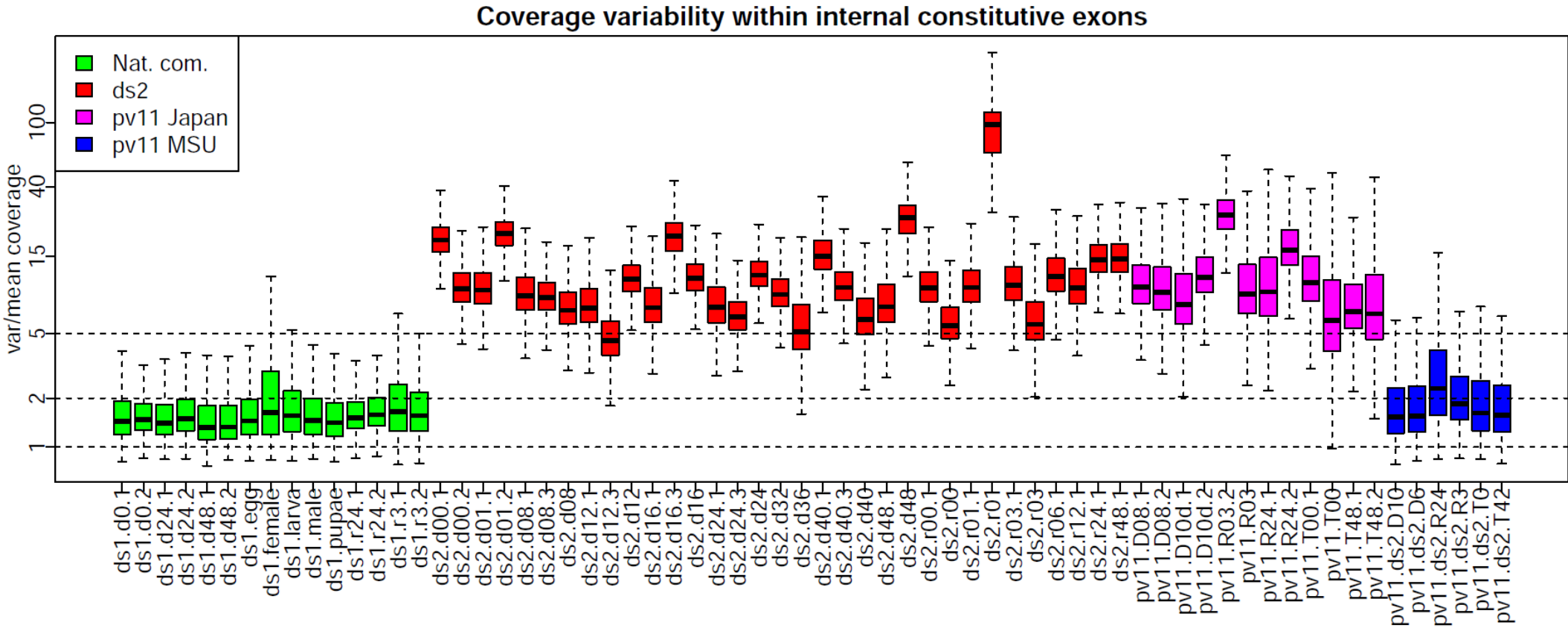
Низкое качество исходной РНК или оверамплификация в ходе подготовки библиотеки, может приводить к тому, что покрытие генов ридами будет очень не ровным, вплоть до появления «стопок» идентичных ридов



Что делать:

- Ничего
- Убрать идентично-картируемые риды (не работает для непарных ридов)
- Менять экспериментальную процедуру (меньше циклов ПЦР, использовать баркодирование)

Оценка неровности покрытия: отклонение от распределения Пуассона



Formats: sam (bam)

Sequence Alignment/Map format, tab-delimited, consists of two sections:

- Header (optional)
  - All lines starts with '@'
  - @HD — first header line
  - Lines are tab-delimited, consist of pairs: TAG:VALUE

    ```
    @HD VN:1.5 SO:coordinate
    @SQ SN:ref LN:45
    ```
- Alignment 11 fields:
  - QNAME — read name
  - FLAG — bit flag that says whether read was mapped, whether it was mapped in multiple places, strand, is it first or second mate (for paired reads), etc
  - RNAME — reference (chromosome) name
  - POS — 1-based leftmost mapping POSition
  - MAPQ — mapping (usually just read) quality
  - CIGAR — junction, indels, etc: 8M200N10M (I, D — indels, S, H - clipping)
  - RNEXT, PNEXT — RNAME and POS for the mate
  - TLEN - observed Template LENgth
  - SEQ — read sequence
  - Optional attrobutes in form of TAG:TYPE:VALUE
    - NH:i:1 — number of alignments
    - NM:i:0 — number of mismatches
    - XS:A:- - strand (by junctions)

For more details: http://samtools.github.io/hts-specs/SAMv1.pdf

Доступ на кластер

- Под windows используйте putty
- Host: mg.uncb.iitp.ru
- Port: 9022
- Получить логин и пароль у преподавателя

Bash

ls -lh — lists files in current directory
cd new.path — change directory
- ~ home directory
- ./ current directory
- ../ parent directory
- / root directory

less,more,cat — see text file
cp — copy file
rm — remove file
mv from to — move file
mkdir — create directory
grep — look for pattern
find — search for files
wget — download from Internet
gzip, gunzip, tar -xzf — compress/decompress files
man cp — get manual for cp command

ls > ls.out — redirects output of ls command into file
Ls -1 | grep txt — redirectd output of ls command as input of grep

echo PATH

Loops:

```
for i in `ls -1`
do
    echo $i
done
```

Vim

- i — insert
- Esc — escape insert mode
- :wq — write and exit

Практикум он же ДЗ1

Сегодня мы будем работать с данными РНК-сек полученными из коры (B) и мозжечка (C), мышей разного возраста (от 15.5 до 34 дней от зачатия). Данные лежат на кластере в папке /mnt/local/vse2019/shared/rnaseq. Для сокращения объема расчетов вы получите данные только для 19ой хромосомы.
Ответом на данный практикум будет текстовый файл содержащий необходимый код (на bash и/или других языках при необходимости, включая скачивание и распаковывание программ, картирование и подсчет количества ридов — пункты 7-11)

Задание:

1) Найдите в интернете и скачайте бинарники для последних версий hisat2
2) Зайти на ensembl.org → downloads → Download data via FTP → скачать последовательность 19 хромосомы мыши и её аннотацию в формате gtf (для всего генома).
3) отфильтруйте из аннотации только 19ую хромосому при помощи команды grep -P '^19\t'
4) Постройте индекс по последовательности 19ой хромосомы при помощи команды hisat2-build (без координат сайтов)
5) Прокартируйте все fq файлы (начните с одного) на 19ую хромосому при помощи hisat2 не допуская обрезания ридов и сообщив hisat2 координаты сайтов сплайсинга
6) Выберите случайно один образец
7) Сколько ридов картируется в регион 19:12485000-12490000 в этом образце?
8) Сколько из них картируются только в одно место генома?
9) Сколько ридов картровалось без замен? Сколько с 1, 2 и т. д. заменами?
10) Сколько ридов картировалось на экзон-экзонные границы? Перечислите координаты всех интронов в данном интервале подтверждённых хотя бы одним ридом в формате:
11) chr:from-to coverage