Введение в транскриптомный анализ

Анастасия Жарикова

azharikova89@gmail.com

30 ноября 2021

Задача

Секвенировать транскриптом

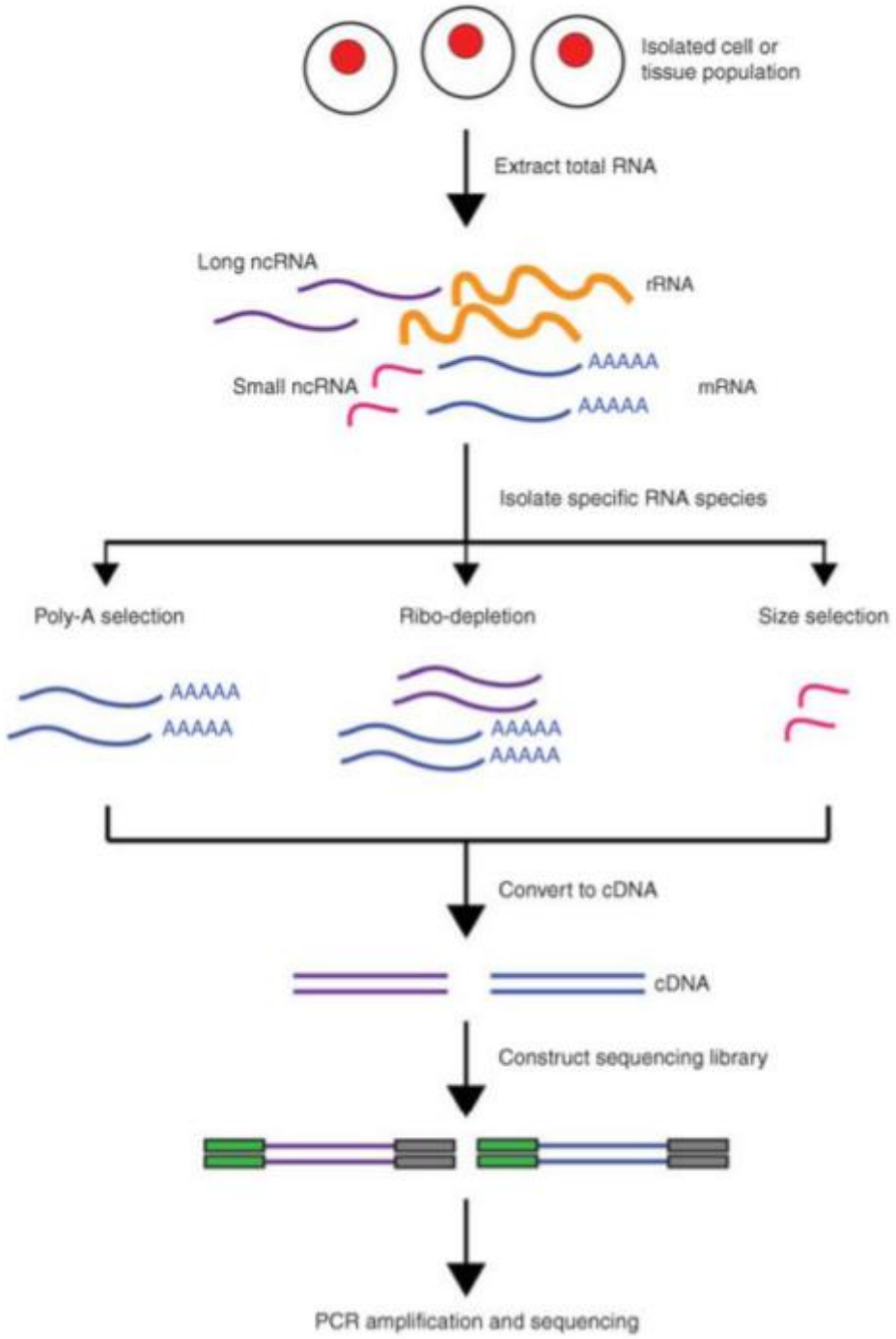
Типы РНК в клетке

- **Тотальная РНК**
- **полиА**
- **Без фракции рРНК**
- **По размеру:**
  - Малые РНК:
    - микроРНК
    - малые ядерные РНК
    - малые ядрышковые РНК
    - малые интерферирующие РНК
    - пиРНК
  - «Длинные» РНК:
    - мРНК
    - длинные некодирующие РНК
- **По внутриклеточной локализации:**
  - Ядерные
  - Цитоплазмаические
- ...

Процесс

1. Подготовка нужной фракции РНК
2. Проверка качества РНК
3. Обратная транскрипция => кДНК
4. Фрагментация (~ 200-300 нк)
5. Секвенирование (чем глубже, тем лучше)

Технические реплики – повторный анализ одного и того же образца
Биологические реплики – повторное взятие образца и анализ


Isolated cell or
tissue population
Extract total RNA
Long ncRNA
rRNA
Small ncRNA
AAAAA
AAAAA
mRNA
Isolate specific RNA species
Poly-A selection
Ribo-depletion
Size selection
AAAAA
AAAAA
AAAAA
AAAAA
Convert to cDNA
cDNA
Construct sequencing library
PCR amplification and sequencing

ref

Цепь-специфичные библиотеки

При секвенировании сохраняем информацию о том, с какой цепи ДНК шла транскрипция

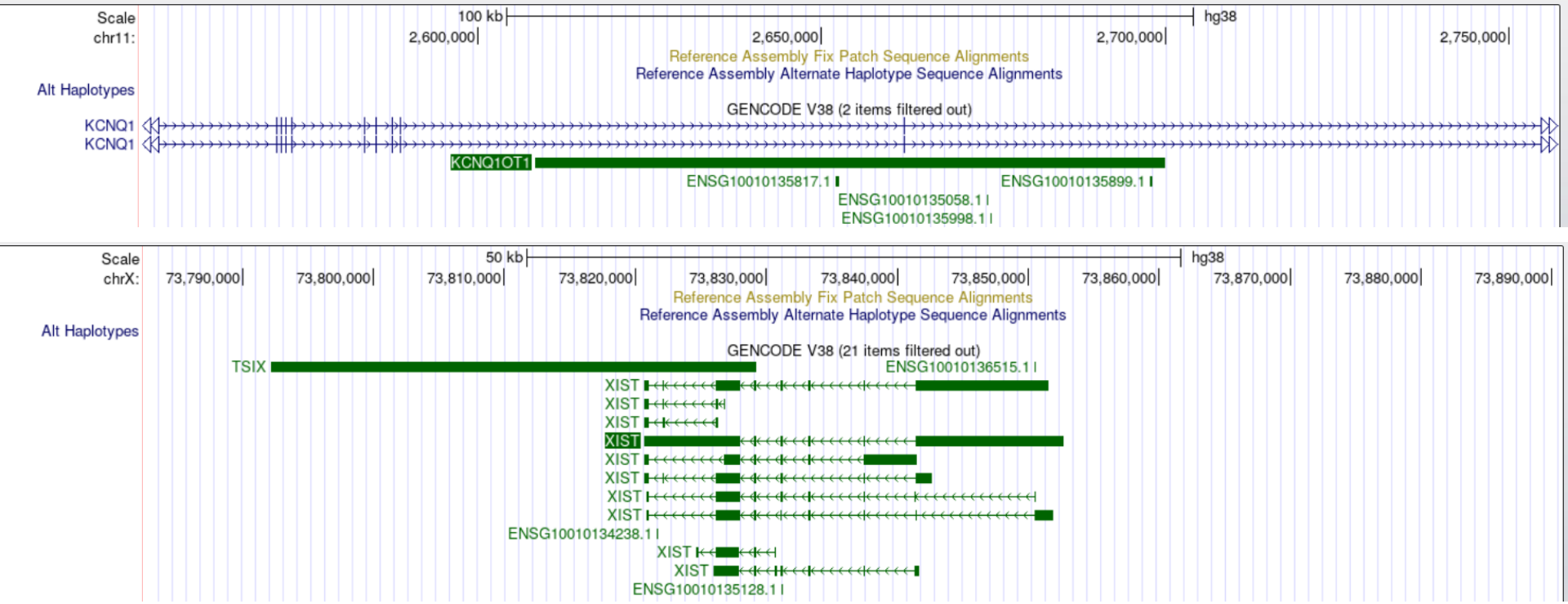
ЗАЧЕМ?

Цепь-специфичные библиотеки

При секвенировании сохраняем информацию о том, с какой цепи ДНК шла транскрипция

ЗАЧЕМ?

мультик

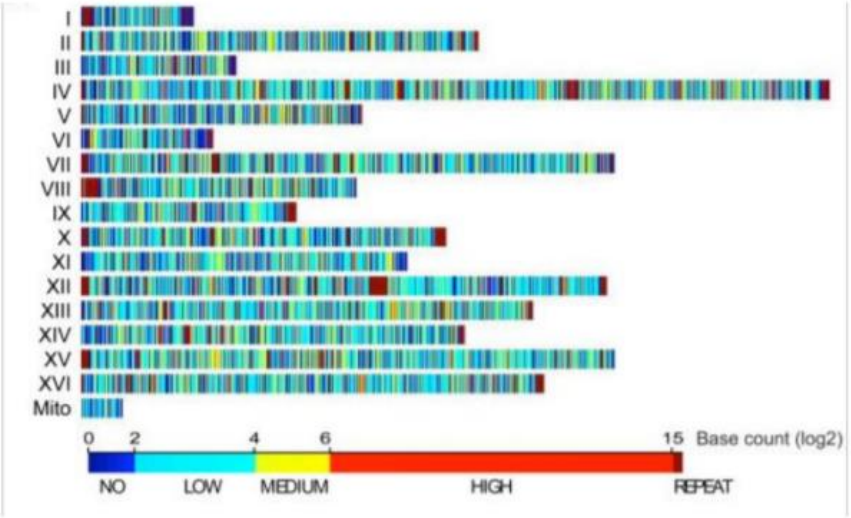


Задачи

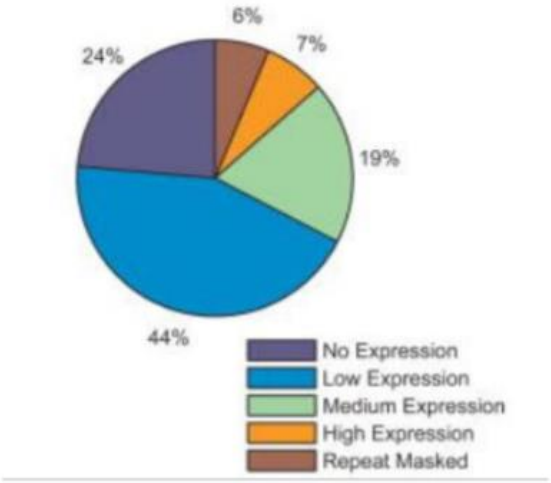
1. Определение концентрации РНК
2. Сравнение уровня экспрессии одного гена в разных образцах
3. Выявление однонуклеотидных полиморфизмов
4. Детекция мест сплайсинга, а также исследование альтернативного сплайсинга
5. Поиск некодирующих РНК
6. Редактирование РНК
7. Анализ транскриптомов единичной клетки

Анализ транскриптомов

Первые работы по секвенированию транскриптомов появились в 2008 году

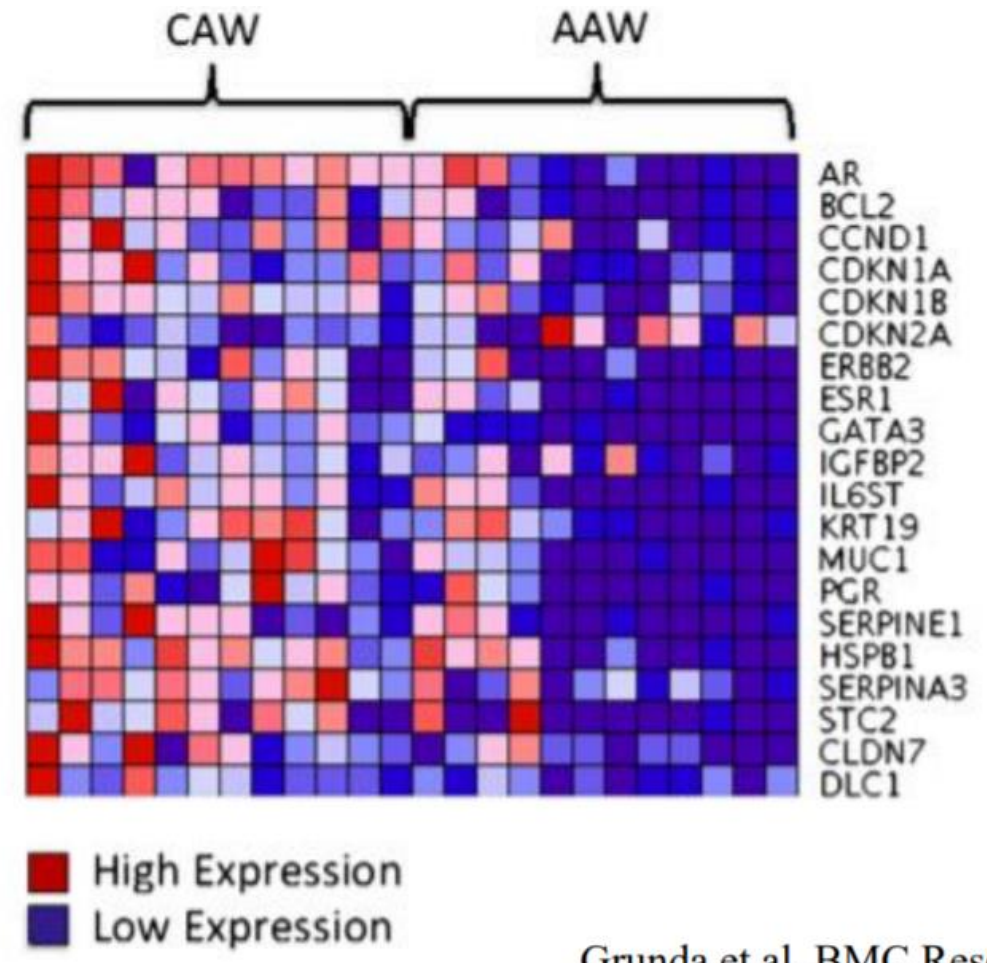


дрожжи



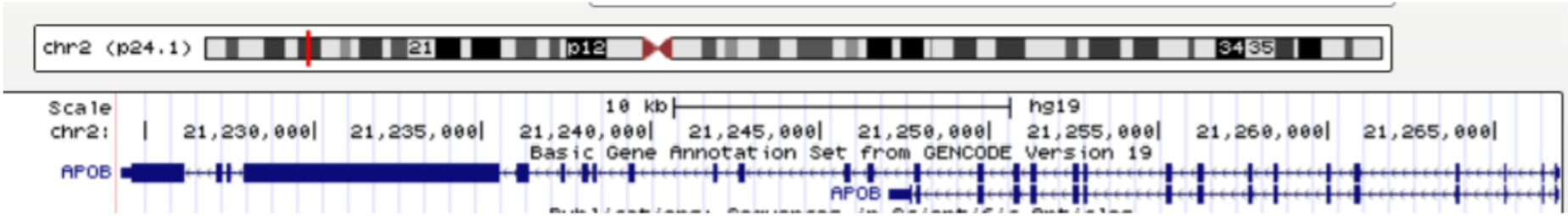
Science. 2008 Jun 6; 320(5881): 1344–1349.

Дифференциальная экспрессия



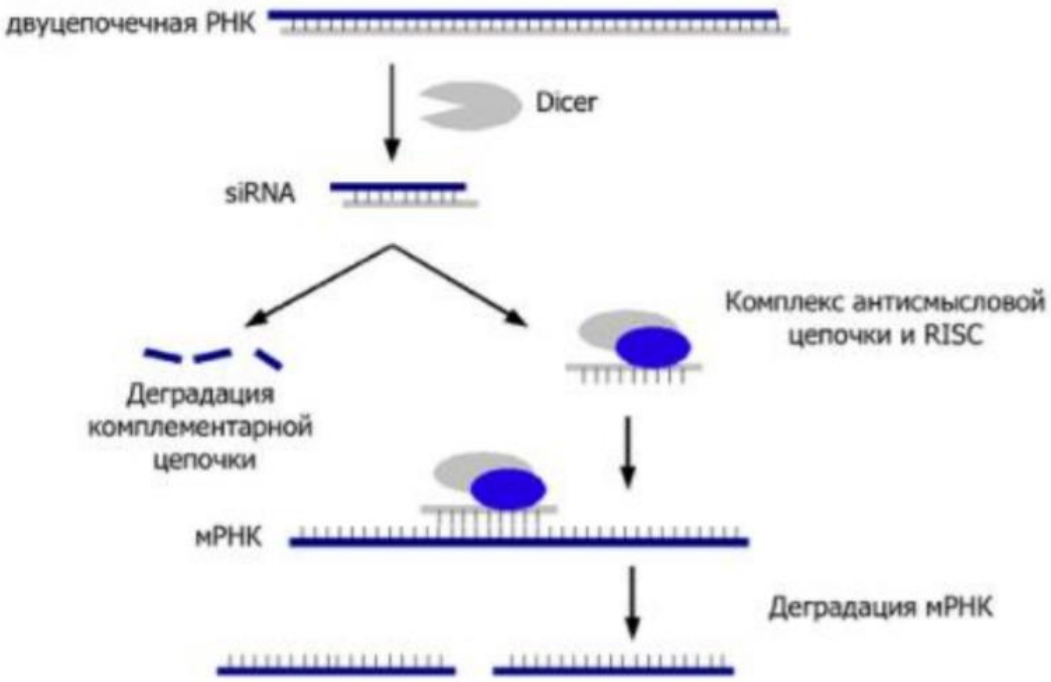
Grunda et al. BMC Research Notes 2012, 5:248

Альтернативный сплайсинг



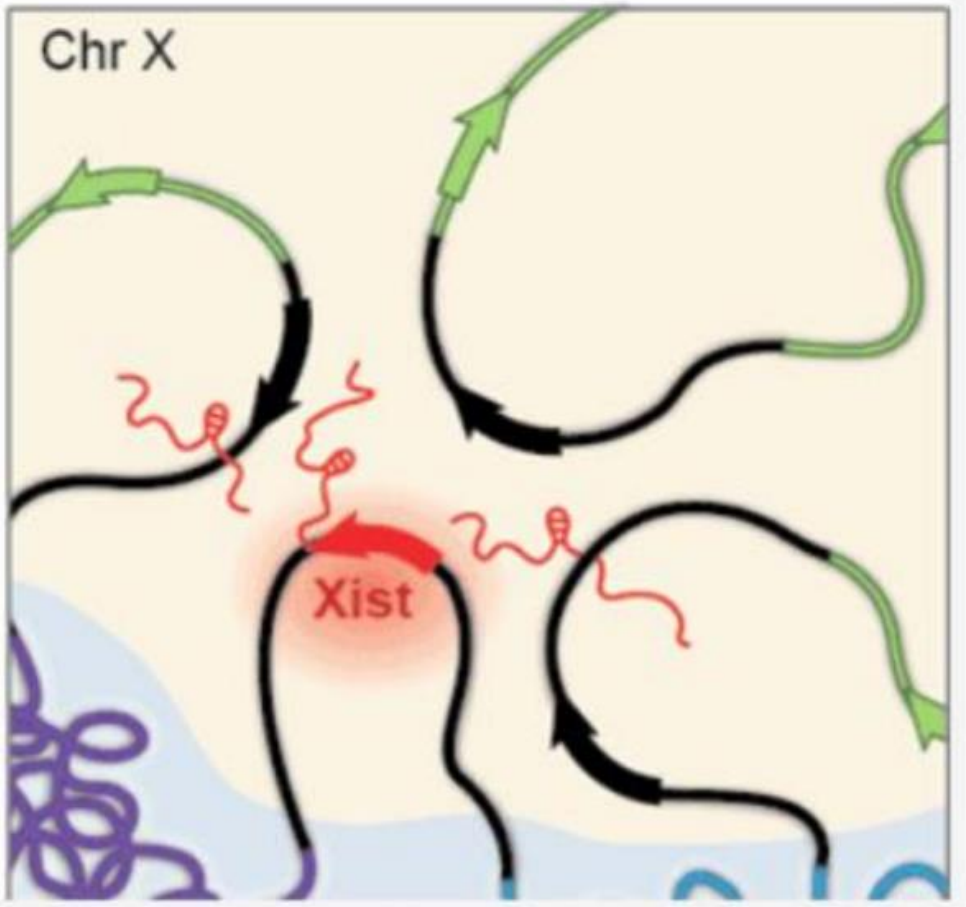
АпоВ-100 – длинный транскрипт – синтезируется в печени.
АпоВ-48 – коротки транскрипт – синтезируется в кишечнике.
Синтезирующиеся белки входят в состав разных групп липопротеинов, которые в последующем идут каждый своим путем метаболизма

Некодирующие РНК

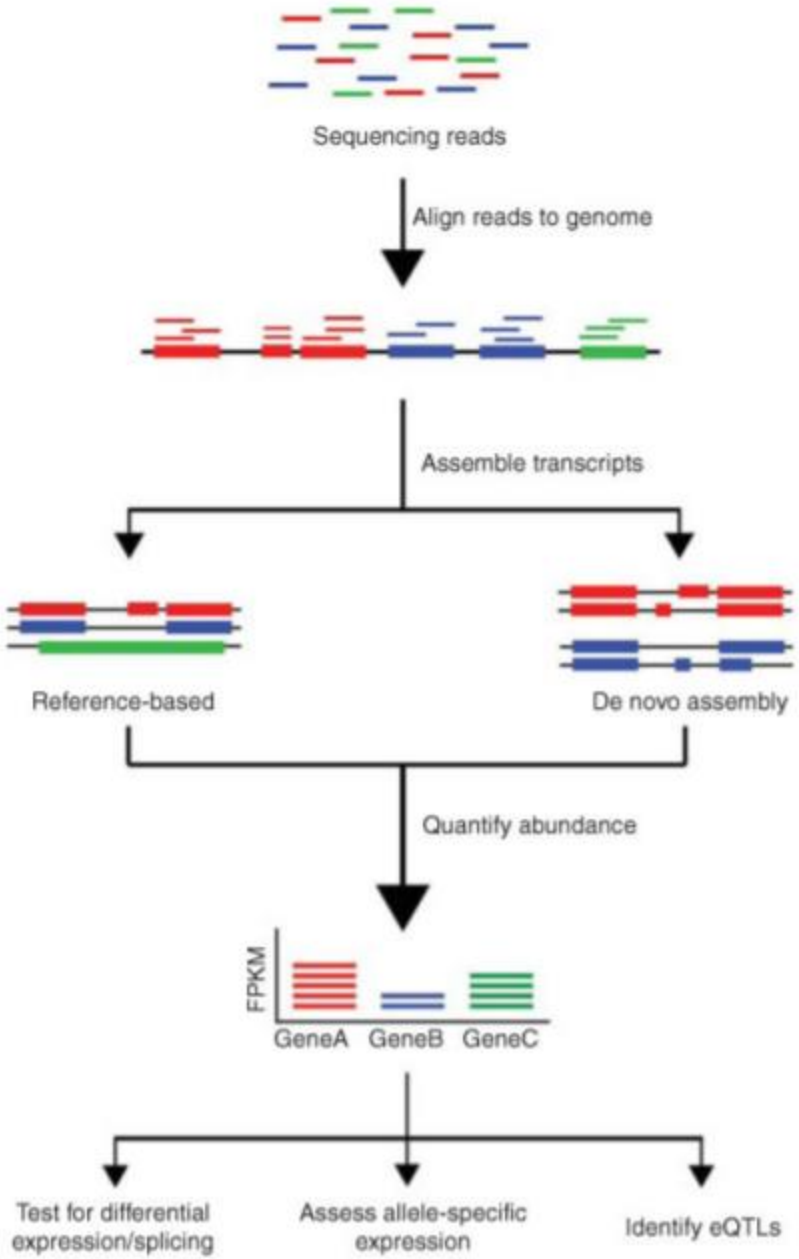


За открытие механизма РНК-интерференции в 2006 году присуждена Нобелевская премия по медицине

Некодирующие РНК

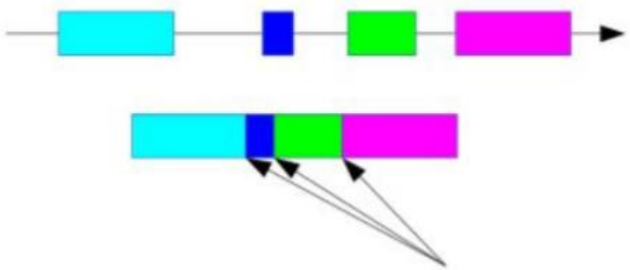


Science, 2013


Sequencing reads
Align reads to genome
Assemble transcripts
Reference-based
De novo assembly
Quantify abundance
FPKM
GeneA
GeneB
GeneC
Test for differential expression/splicing
Assess allele-specific expression
Identify eQTLs

ref

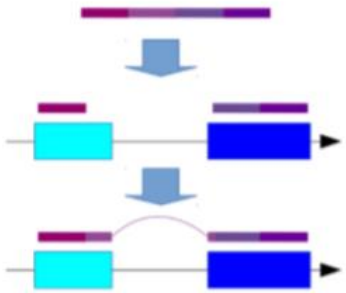
Картирование



Использование аннотации:

- Только аннотированные экзон-экзонные границы
- Все возможные экзон-экзонные границы

Предсказание аннотации из данных:

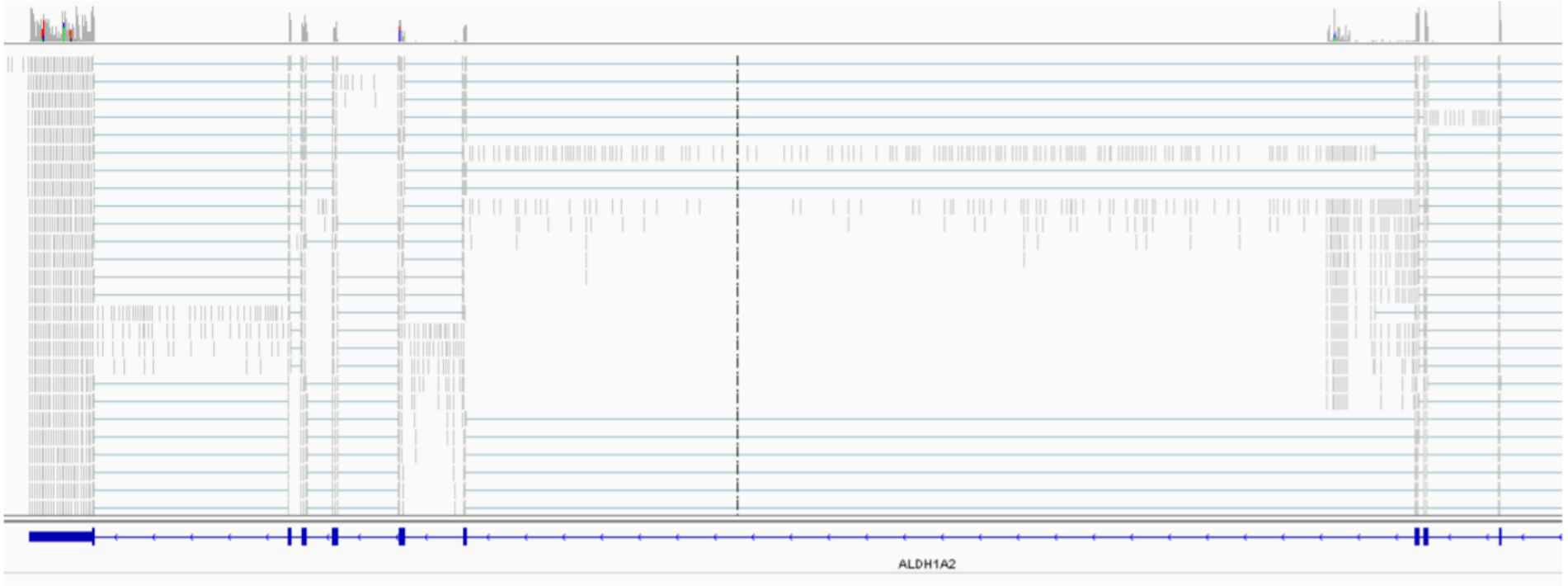


Индексирование генома

1. Аналогично задаче картирования экзомного секвенирования
2. Индексирование с учетом разметки (.gtf)
   - Экстракция из аннотации экзонов
   - Экстракция из аннотации сайтов сплайсинга
   - Индексирование с использованием списка экзонов и сайтов сплайсинга

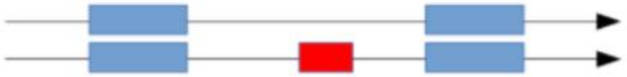
hisat2

IGV



Глубина покрытия важна

В клетке присутствуют 2 варианта транскриптов одного гена:



Секвенировали два образца: один получился с хорошим покрытием, другой – с плохим
Видим:

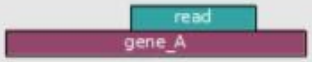
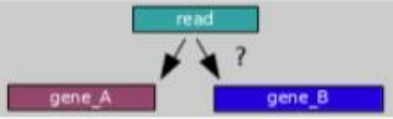
Хорошо покрытый образец: все экзон-экзонные границы нашлись

Плохо покрытый образец: один вариант потерялся

Дифференциальный альтернативный сплайсинг?
Дифференциальная экспрессия?

Нужно нормировать на размер библиотеки и оценивать нормировочные коэффициенты!

Подсчет чтений – Htseq-count

	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)		

```
ENSG00000000005.5        0
ENSG00000000419.8        23
ENSG00000000457.9        397
ENSG00000000460.12       239
ENSG00000000938.8        0
ENSG00000000971.11       13
ENSG00000001036.9        19
ENSG00000001084.6        1
ENSG00000001167.10       12
ENSG00000001460.13       23
```

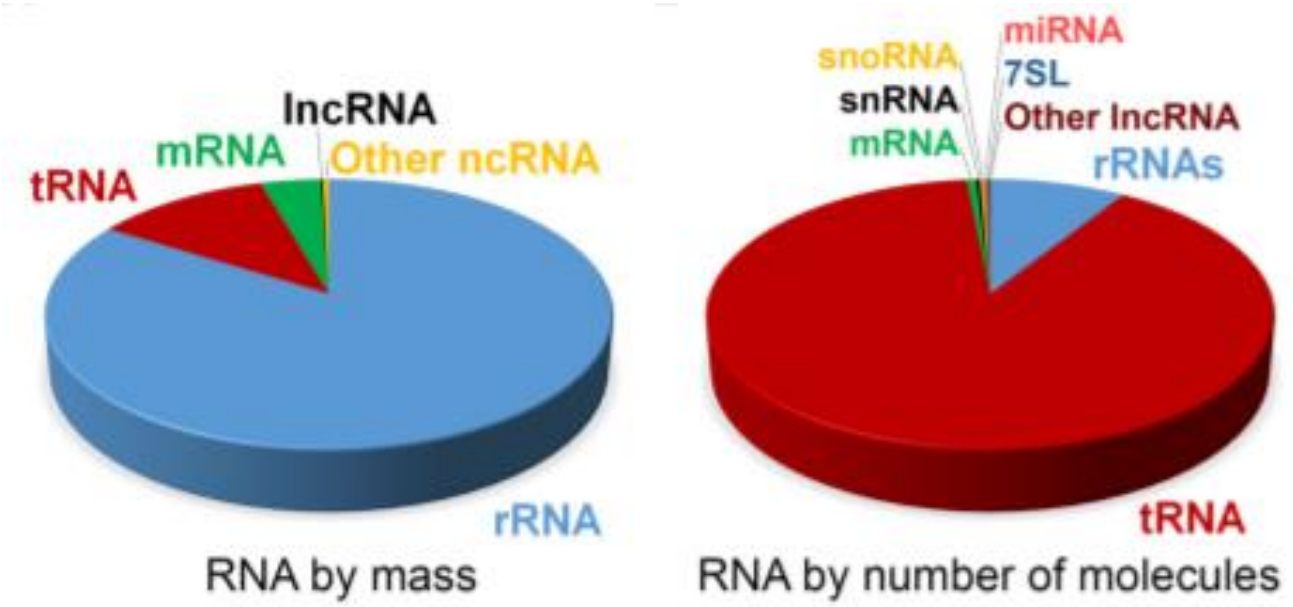
```
__no_feature    13696354
__ambiguous     66168
__too_low_aQual 0
__not_aligned   0
__alignment_not_unique  0
```

Разметка генов

https://www.gencodegenes.org/

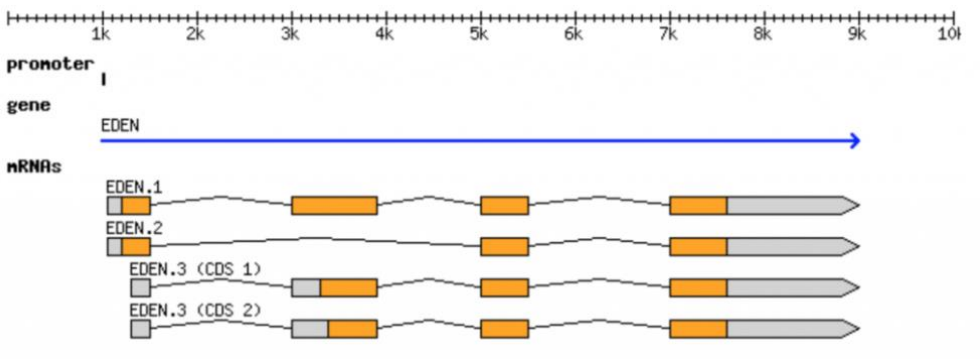
Statistics about the current GENCODE Release (version 38)

Total No of Genes	60649	Total No of Transcripts	237012
Protein-coding genes	19955	Protein-coding transcripts	86757
Long non-coding RNA genes	17944	- full length protein-coding	61015
Small non-coding RNA genes	7567	- partial length protein-coding	25742
Pseudogenes	14773	Nonsense mediated decay transcripts	18881
- processed pseudogenes	10667	Long non-coding RNA loci transcripts	48752
- unprocessed pseudogenes	3565		
- unitary pseudogenes	241		
- polymorphic pseudogenes	49		
- pseudogenes	15	Total No of distinct translations	63968
Immunoglobulin/T-cell receptor gene segments		Genes that have more than one distinct translations	13689
- protein coding segments	409		
- pseudogenes	236		



https://doi.org/10.3389/fgene.2015.00002

GFF3




```
##gff-version 3.2.1
##sequence-region ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             5000  5500  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             7000  7600  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
ctg123 . CDS             5000  5500  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
ctg123 . CDS             7000  7600  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```

https://learn.gencore.bio.nyu.edu/ngs-file-formats/gff3-format/

Что можно сделать дальше?

- Оценить самосогласованность образцов
- Сколько чтений легло в границы разметки?
- Подсчет дифференциальной экспрессии
- GO аннотация
- Анализ альтернативного сплайсинга
- Сборка аннотации
- …