



# Аннотация геномов. Формат GFF и GTF. Другие форматы аннотации геномов

Что такое аннотация, gff, gtf, другие форматы аннотации геномов

Page • Tag • 1 backlink

Аннотация генома

RepeatMasker

Структура гена эукариот

Предсказание и аннотация генов

Псевдогены

Длины генов и геномов

Аннотация модельных организмов

UCSC Tables

ENSEMBL

Аннотация генома человека

Gencode

Аннотация BED

Аннотация gtf, gtp

Формат GFF/GTF

Колонки:

Пример строки в GFF:

Иерархические построения

## Аннотация генома

### 1. Структурная аннотация

- поиск повторов
- поиск протеин-кодирующих генов
- идентификация кодирующих последовательностей
- поиск псевдогенов и некодирующих РНК

### 2. Функциональная аннотация:

- какова роль генов и регуляторных элементов.

## RepeatMasker

- создается библиотека повторов (de novo или нет)
- RepeatMasker находит и "маскирует" их
- *softmask* - замена AGCAGCAGC на agcagcagc
- *hardmask* - замена AGCAGCAGC на NNNNNNNNNN

**FASTA** ACAGACTGGTATGAAGGTGGCCACAATTCAGAAAGAAAAAGAAGAGC

**BED**



**FASTA'** ACANNNNGGTANNNNNNNGGCCACANNNNNNAAGAANNNNNAGAGC

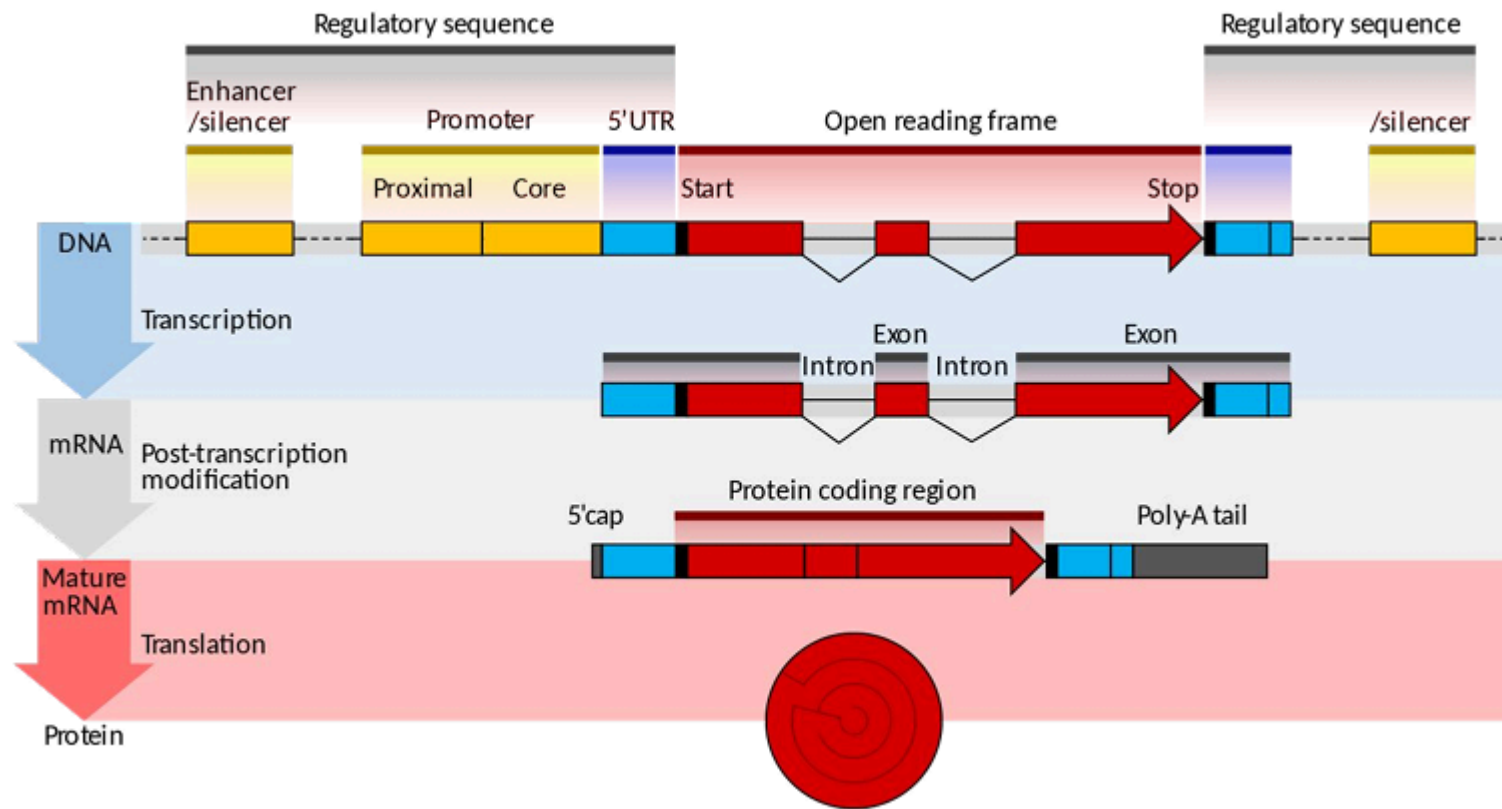
Повторы бывают: **тандемные** (повторяющиеся) и **диспергированные** (на расстоянии, часто в результате дупликации)

**Маскировка:** последовательность собранная заменяется нижнем регистре или N

fasta → bedtools → fasta с **маркированными повторами** (теперь можно определить за счет регистра или N)

### Структура гена эукариот

1. промотор
2. нетранслируемые последовательности (5'UTR, 3'UTR)
3. экзоны и интроны

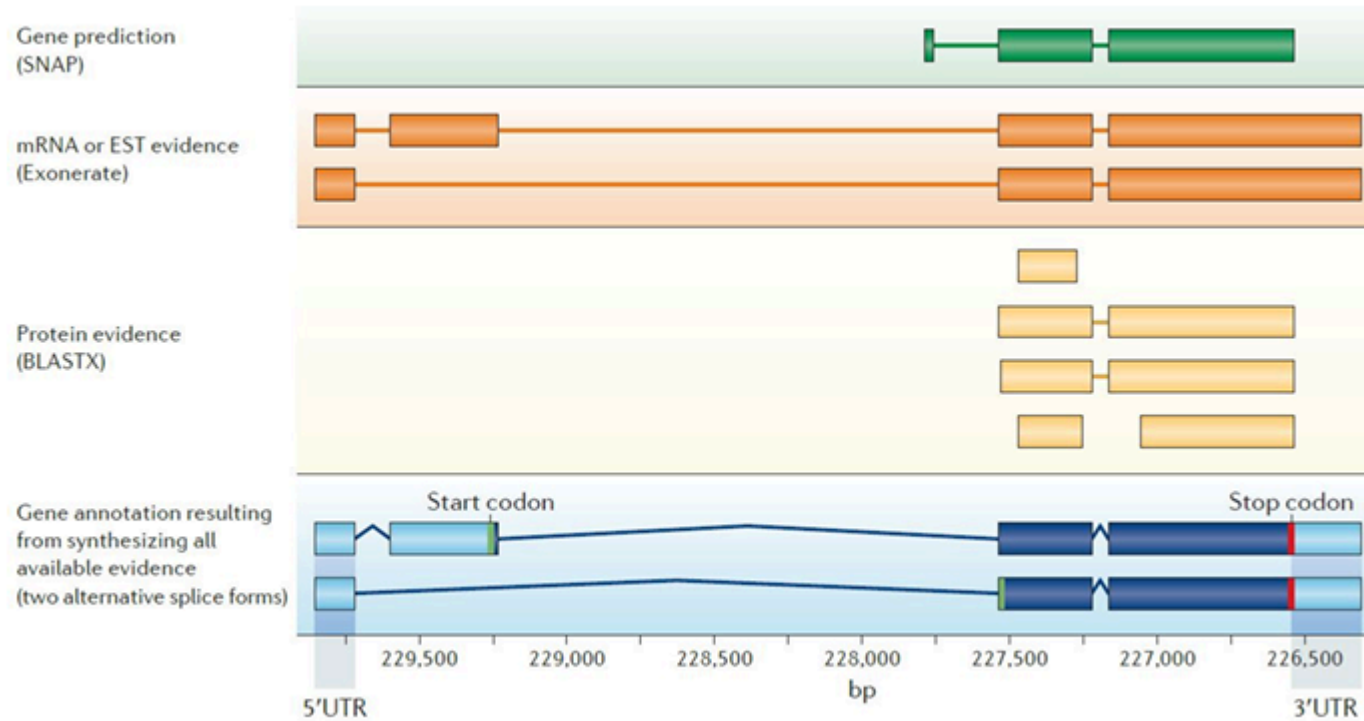


- сплайсинг — вырезание интронов и склеивание протеинов
- нетранслируемые последовательности — 5'UTR со стороны промотора, регуляторные функции и интенсивности трансляции
- 5-кэп — транспорт мРНК из ядра
- 3'-полиА — 100-200 аденозинов, защита от деградации мРНК
- энхансеры, сайленсеры — регуляция транскрипции

## Предсказание и аннотация генов

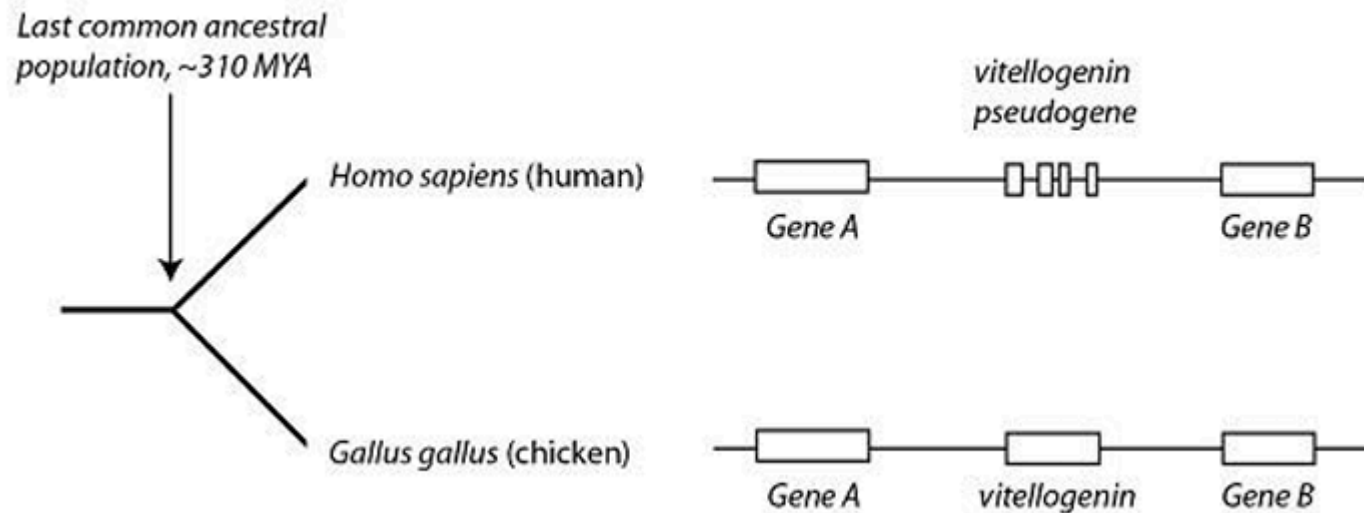
- предсказание использует только последовательность генома (нахождение всех открытых рамок считывания)

- аннотация требует данных экспрессии и протеинов



## Псевдогены

- ген, у которого в процессе эволюции произошла поломка, что привело к его полной неактивности, нетранскрибируемости
- около 15к в геноме человека (20к протеин-кодирующих)
- поломки в промоторах (нет транскрипции)
- ранние стоп-кодны (нет трансляции) → NMD
- типы: процессированные, неprocessированные, унитарные



**Процессированные** — получены в ходе обратной транскрипции (встраивания) ретротранспозонами уже сплайсированных генов (процессированная мРНК)

**Непроцессированные** — в ходе дупликации ломается вторая копия гена (может сформироваться в паралог, к примеру алкогольдегидрогеназа)

**Унитарные** — один ген в ходе эволюции сломался и утерю функциональность

#### Процессированные псевдогены

- **Механизм образования:** Возникают в результате ретротранспозиции, когда мРНК функционального гена обратно транскрибируется в ДНК и встраивается в геном.
- **Характеристики:**
  - Не имеют интронов, поскольку образуются из зрелой мРНК.
  - Часто лишены промоторных последовательностей, что делает их неспособными к транскрипции.
  - Могут иметь полиа(А) хвост, который остаётся после ретротранспозиции.
- **Пример:** Многие ретрогены (ретропсевдогены) возникают таким образом.

### Непроцессированные псевдогены

- **Механизм образования:** Возникают в результате дупликации и последующей мутации функционального гена.
- **Характеристики:**
  - Сохраняют структуру, схожую с оригинальным геном, включая интроны и экзоны.
  - Со временем накапливают мутации, которые нарушают их способность кодировать белки.
  - Могут сохранять элементы, такие как промоторы, но, из-за мутаций, остаются нефункциональными.
- **Пример:** Гены-дубликаты, утратившие свою функциональность.

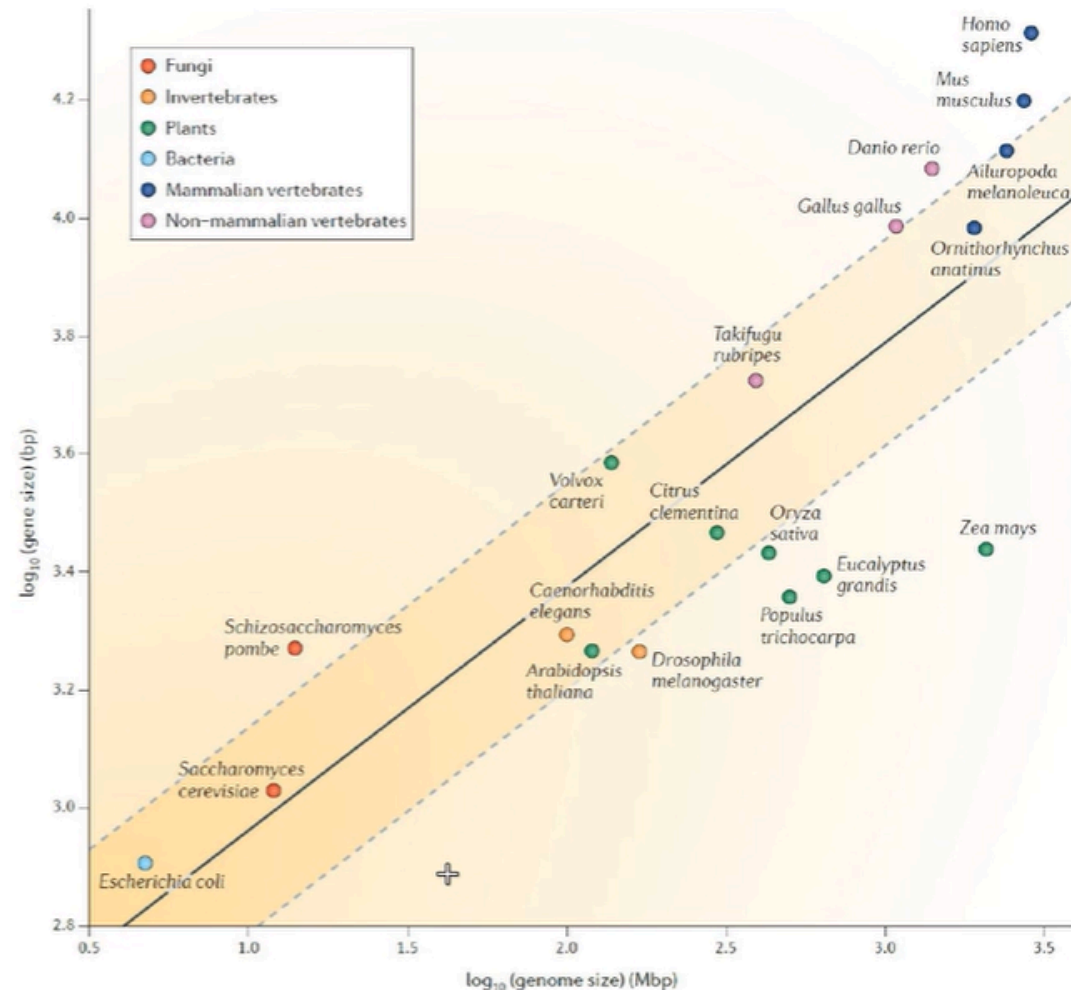
### Унитарные псевдогены

- **Механизм образования:** Возникают вследствие накопления мутаций в функциональном гене, что приводит к потере его функции, без дупликации или ретротранспозиции.
- **Характеристики:**
  - Возникают в результате мутаций (например, нонсенс-мутации, фреймшифты) в единственном экземпляре гена, который ранее был функциональным.
  - В отличие от процессированных и непроцессированных псевдогенов, они не имеют функционального аналога в геноме.
- **Пример:** Унитарные псевдогены могут возникать при утрате важного для вида гена, например, у человека псевдоген гена GULO, ответственного за синтез витамина С, который функционален у многих других млекопитающих.

## Длины генов и геномов

# Длины генов и геномов

- длина гена коррелирует с длиной генома
- для успешной аннотации больших геномов нужны сборки с БОльшим N50



## Аннотация модельных организмов

Сборкой, аннотацией и функциональной аннотацией, как правило, занимаются одни и те же консорциумы.

1. [FlyBase](#)
2. [WormBase](#)
3. [TAIR](#)



4. [SGD](#)

5. [EcoGene](#)

## UCSC Tables

[genome.ucsc.edu/cgi-bin/hgTables](http://genome.ucsc.edu/cgi-bin/hgTables)

The screenshot shows the UCSC Table Browser interface. At the top is a navigation bar with links: Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Help, and About Us. Below this is the 'Table Browser' title and a descriptive paragraph. The main form includes several dropdown menus and buttons. The 'clade' is set to 'Mammal', 'genome' to 'Human', and 'assembly' to 'Dec. 2013 (GRCh38/hg38)'. The 'group' dropdown is open, showing 'Genes and Gene Predictions' selected. The 'track' is set to 'RefSeq Genes'. There are buttons for 'add custom tracks' and 'track hubs'. The 'table' dropdown is also open, showing 'Genes and Gene Predictions' selected. There is a 'describe table schema' button. The 'region' is set to 'chr9:133252000-133280861', with 'lookup' and 'define regions' buttons. There are 'paste list' and 'upload list' buttons. The 'filter' dropdown is open, showing 'Regulation' selected. The 'intersect' dropdown is open, showing 'Comparative Genomics' selected. The 'correlate' dropdown is open, showing 'Repeats' selected. The 'output' dropdown is open, showing 'All Tracks' selected. There is a 'Send output to' section with checkboxes for 'Galaxy', 'GREAT', and 'GenomeSpace'. The 'output file' is set to 'gencode\_v19\_tong\_bed.gz', with a note '(leave blank to keep output in browser)'. The 'file type returned' section has radio buttons for 'plain text' and 'gzip compressed'. At the bottom are 'get output' and 'summary/statistics' buttons.

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To export data, use [GREAT](#). Send data to [GenomeSpace](#) for use with diverse computational tools. Refer to the [Credits](#) page for the list of contributors. Downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.

clade: Mammal genome: Human assembly: Dec. 2013 (GRCh38/hg38)

group: Genes and Gene Predictions track: RefSeq Genes add custom tracks track hubs

table: Genes and Gene Predictions describe table schema

region: chr9:133252000-133280861 lookup define regions

identify: mRNA and EST

filter: Regulation

intersect: Comparative Genomics

correlate: Repeats

output: All Tracks

output file: gencode\_v19\_tong\_bed.gz (leave blank to keep output in browser)

Send output to ☐ Galaxy ☐ GREAT ☐ GenomeSpace

file type returned: ☐ plain text ☒ gzip compressed

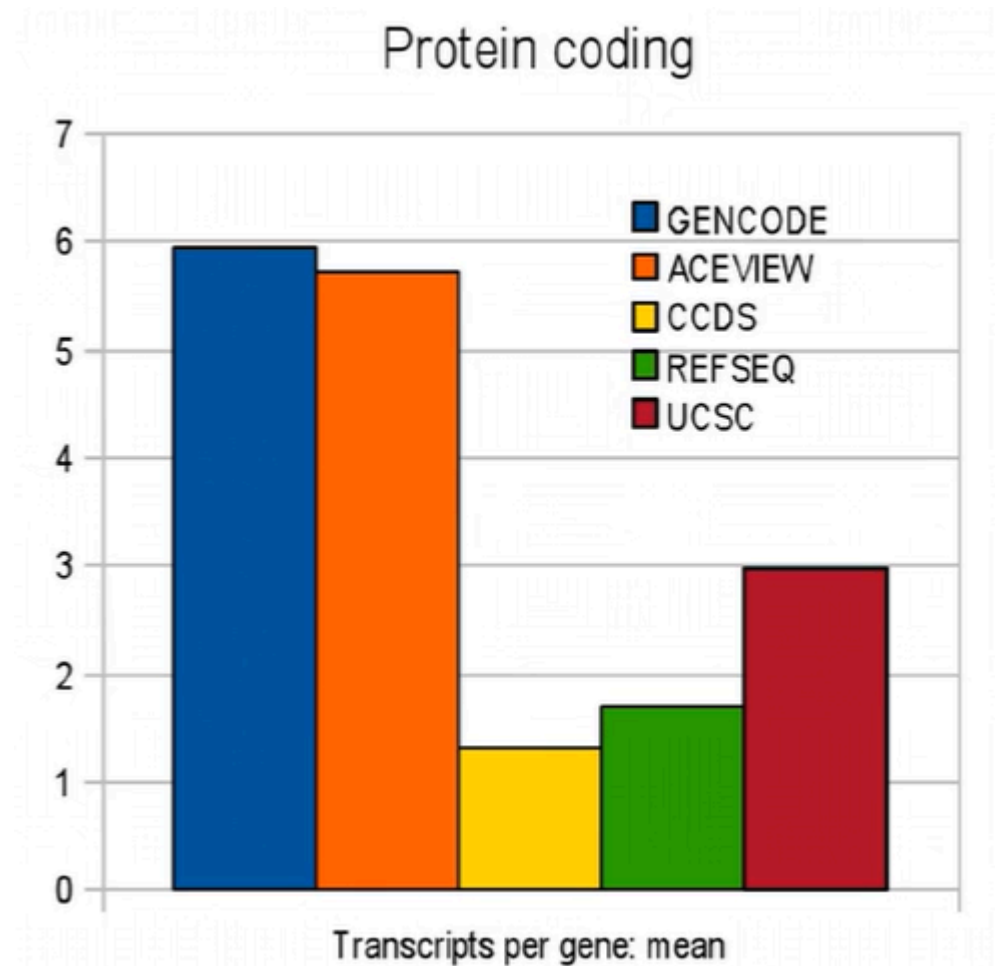
get output summary/statistics

## ENSEMBL

- группа под руководством Европейского института молекулярной биологии

- выпуски — release
- включает сборки геномов и их аннотации для различных биологических видов
- поддерживает единые форматы данных, API для программистов и быстрый доступ через FTP
- предоставляет инструменты для удобной работы с геномной информацией

## Аннотация генома человека

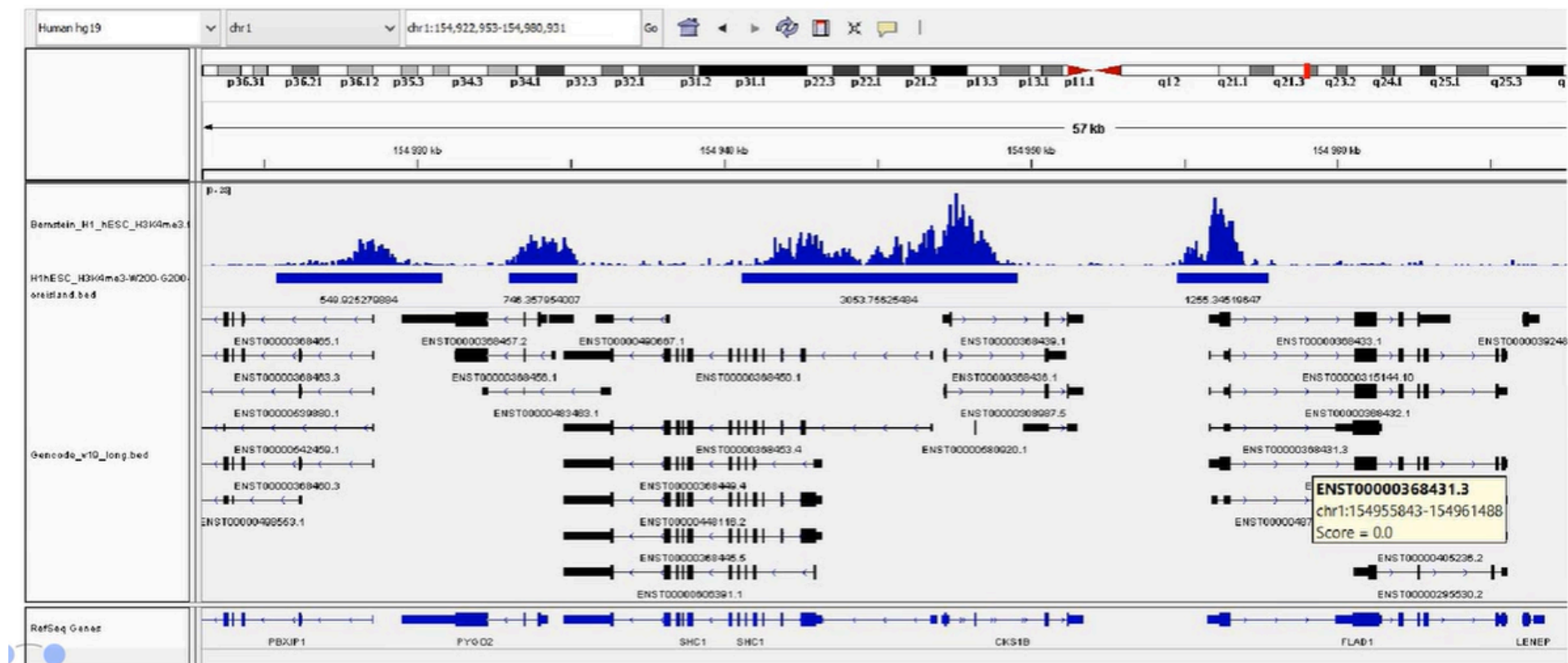


- CCDS
- Gencode
- UCSC
- REFSEQ
- Ensembl

## **Gencode**

- аннотация для человека и мыши
- использует нотацию генов Ensembl
- ENSG0000 — для гена человека, ENST0000 — для транскриптов человека
- GRCh вариации

## **Аннотация BED**



Аннотация 12-колоночного BED-файла позволяет увидеть различное направление транскриптов, название, кодирующие и некодирующие участки

Но мы не можем в таком формате показать принадлежность транскриптов (черное) к определенному гену (например синим *SNC1*)

## Аннотация gtf, gtr

GTF = GFF2, GFF3

### Формат GFF/GTF

- Структура: 9 колонок, разделённых табуляцией ( `tab` ).

- **Название:**
  - **GFF** (*General Feature Format*)
  - **GTF** (*General Transfer Format*)

### Колонки:

1. **seqname** — хромосома или скаффолд (например, `chr1`).
2. **source** — источник аннотации (например, `Ensembl`, `NCBI`).
3. **feature** — тип элемента (`gene`, `transcript`, `exon`, `CDS` и т. д.).
4. **start** — начальная позиция (целое число).
5. **end** — конечная позиция (целое число).
6. **score** — оценка (часто `.` или `0`, если не применяется).
7. **strand** — нить ДНК (`+`, `-` или `.` для неопределённой).
8. **frame** — рамка считывания (`0`, `1`, `2` или `.`):
  - `0` — первый нуклеотид совпадает с кодоном.
  - `1`, `2` — сдвиг на 1 или 2 нуклеотида соответственно.
9. **attribute** — дополнительные атрибуты (например, `ID=gene1;Name=TP53`).

---

### Пример строки в GFF:

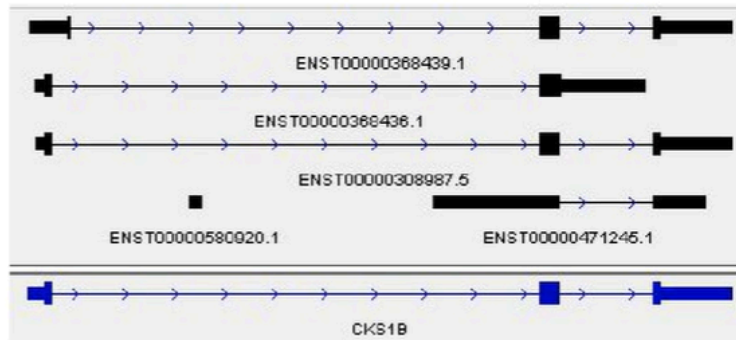
Plain Text ▾

```
chr1 Ensembl gene 1000 5000 . + . ID=gene1;Name=BRCA2
```

### Примечание:

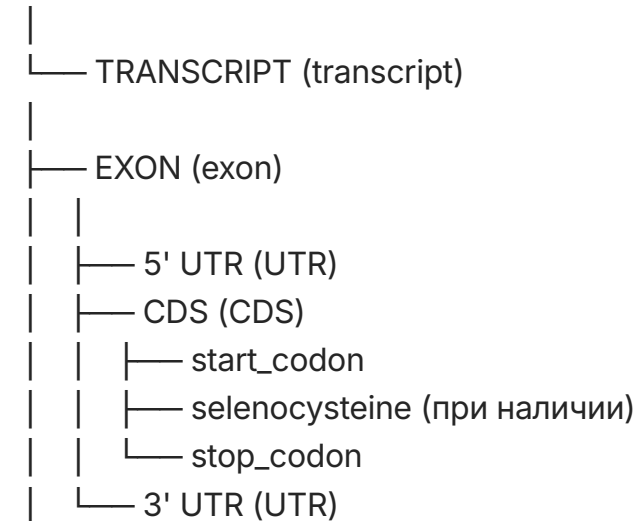
- В GTF (разновидность GFF) атрибуты имеют строгий формат (например, `gene_id "gene1"; transcript_id "t1";`).
- Используется для хранения геномных аннотаций в биоинформатике.

## Иерархические построения



- ENSG##### Ensembl **Gene** ID
  - ENST##### Ensembl **Transcript** ID
  - ENSP##### Ensembl **Peptide** ID
  - ENSE##### Ensembl **Exon** ID
- For non-human species a suffix is added:  
 ENSMUSG      MUS (*Mus musculus*)      for mouse

GENE (gene)



# 0-based или 1-based

- формат BED - начинается с 0, не включает правое значение: [1,5)
- форматы GTF/GFF - начинается с 1, включают правое значение: [1,5]

**GTF**

AGCAGC

1 2 3 4 5 6

**BED**

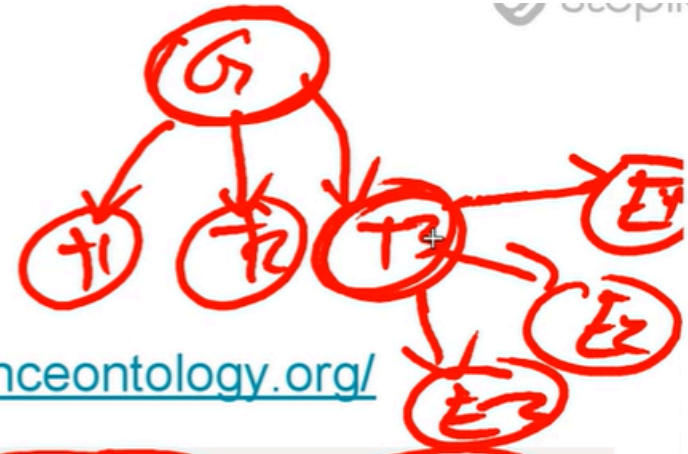
AGCAGC

0 1 2 3 4 5



# Формат GFF3

- больше уровней иерархии
- фиксированное направление иерархии
- поля в третьей колонке только из [sequenceontology.org/](http://sequenceontology.org/)



##gff-version 2

KJ660346	demo	exon	1000	2000	.	+	.	gene_id "Happy"; transcript_id "Tic";
KJ660346	demo	exon	3000	4000	.	+	.	gene_id "Happy"; transcript_id "Tic";
KJ660346	demo	exon	5000	6000	.	+	.	gene_id "Happy"; transcript_id "Tic";
KJ660346	demo	exon	7000	8000	.	+	.	gene_id "Happy"; transcript_id "Tic";

##gff-version 3

KJ660346	demo	exon	1000	2000	.	+	.	Parent=Tic,Tac,Toe;
KJ660346	demo	exon	3000	4000	.	+	.	Parent=Tic;
KJ660346	demo	exon	5000	6000	.	+	.	Parent=Tic,Tac,Toe;

Формат **GFF3** задает **направление иерархии** между элементами аннотации. В форматах **GTF**, **GFF2** и **GFF3** 9 колонок, разделенных знаком **табуляции**. Превращение форматов можно проводить через различные genome tools (например, из gff в gff3)

- Важно всегда соблюдать сборку аннотации и сборки генома!
  - ensembl: 1-22, x,y, MT
  - ucsc: chr1-22, chrX