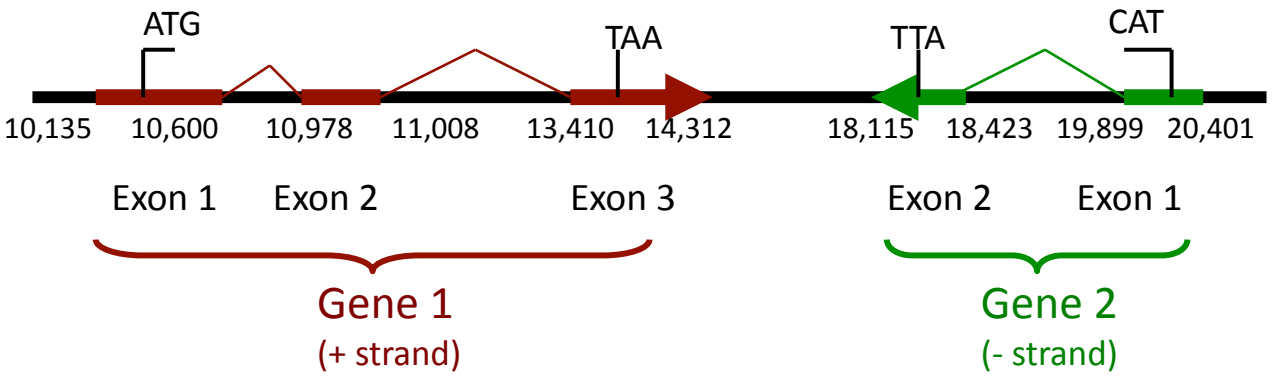
Genomic features

- Genome annotation = determine the precise location and structure (intervals, or lists of intervals, and associated biological information) of genomic features along the genome
- Genomic features: genes, promoters, protein binding sites, translation start/stop site, DNaseI sites, etc.
- Example – gene annotations:
  - Exon/intron structure (exon and intron start-end coordinates)
  - Strand (+ or -)
  - Start and end sites for translation (ORF)



Representation: BED format

Basic format (columns 1-3 required):

```
#chr    start    end          -> 0-based
chr7    10134    10600
chr7    10977    11008
chr7    13409    14312
chr7    18114    18423
chr7    19898    20401
```

Single intervals, e.g. exons

Extended format:

```
#chr  start   end    name   score strand thick_start thick_end rgb
chr7  10134   10600  Exon1  100   +      10180       10600     255,0,0
chr7  10977   11008  Exon2  100   +      10977       11000     255,0,0
chr7  13409   14312  Exon3  100   +      13409       14300     255,0,0
chr7  18114   18423  Exon4  100   -      18150       18423     0,0,255
chr7  19898   20401  Exon5  100   -      19898       20350     0,0,255
```

Representation: BED format

Extended format – groups:

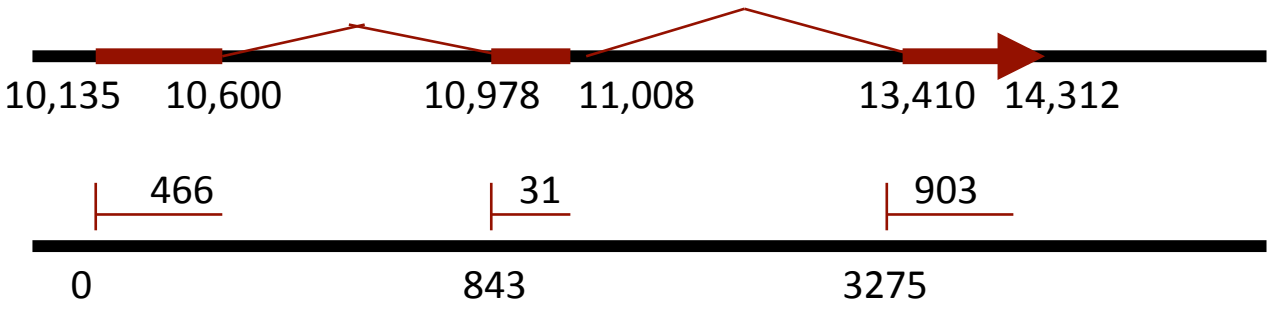
Block count (col 10)

Block sizes (col 11)

```
chr7 10134 14312 GeneA  10  +  10180  11000 0 3 466,31,903
0,843,3275
chr7 18114 20401 GeneB  10  -  18150  20350 0 2 309,603 0,1284
```

Block starts (col 12)

Multiple intervals, e.g. transcripts



Representation: GTF format

```
#chr program feature start end strand frame gene_id; txpt_id

chr7 GF exon 10135 10600 100 + . gene_id “genA”; transcript_id “genA.1”;
chr7 GF exon 10978 11008 100 + . gene_id “genA”; transcript_id “genA.1”;
chr7 GF exon 13410 14312 100 + . gene_id “genA”; transcript_id “genA.1”;
chr7 GF exon 18115 18423 100 - . gene_id “genB”; transcript_id “genB.1”;
chr7 GF exon 19899 20401 100 - . gene_id “genB”; transcript_id “genB.1”;
```

- Each interval feature takes one line
- Columns 1-9 separated by tab ‘\t’; fields within column 9 separated by space ‘ ‘
- Column 9 can have additional attributes
- Coordinates are 1-based

Representation: GFF3 format

```
#chr source feature start end strand frame ID;Name;Parent

##gff-version 3
chr7 GF mRNA 10135 14312 100 + . ID=mrna001;Name=genA
chr7 GF exon 10135 10600 100 + . ID=exon00001;Parent=mrna001
chr7 GF exon 10978 11008 100 + . ID=exon00002;Parent=mrna001
chr7 GF exon 13410 14312 100 + . ID=exon00003;Parent=mrna001
Chr7 GF mRNA 18115 20401 100 - . ID=mrna002;Name=genB
chr7 GF exon 18115 18423 100 - . ID=exon00004;Parent=mrna002
chr7 GF exon 19899 20401 100 - . ID=exon00005;Parent=mrna002
```