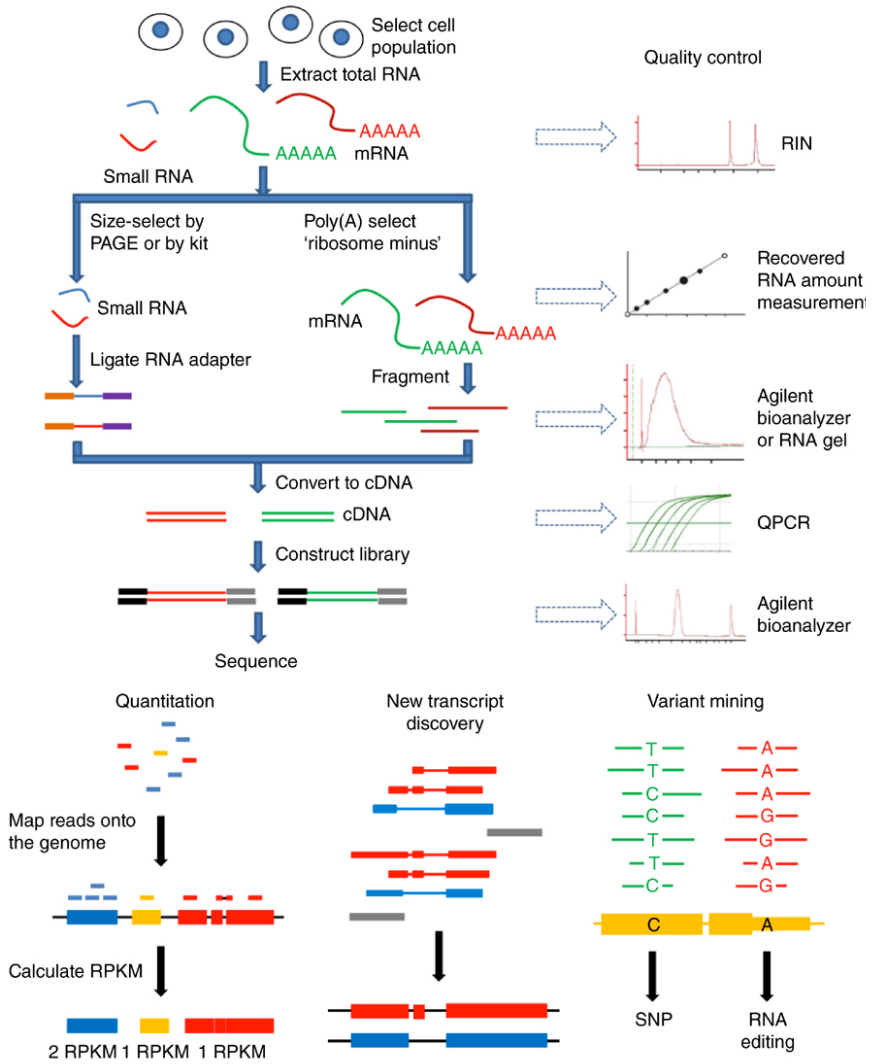
Intro to Gene Annotations and RNA-seq

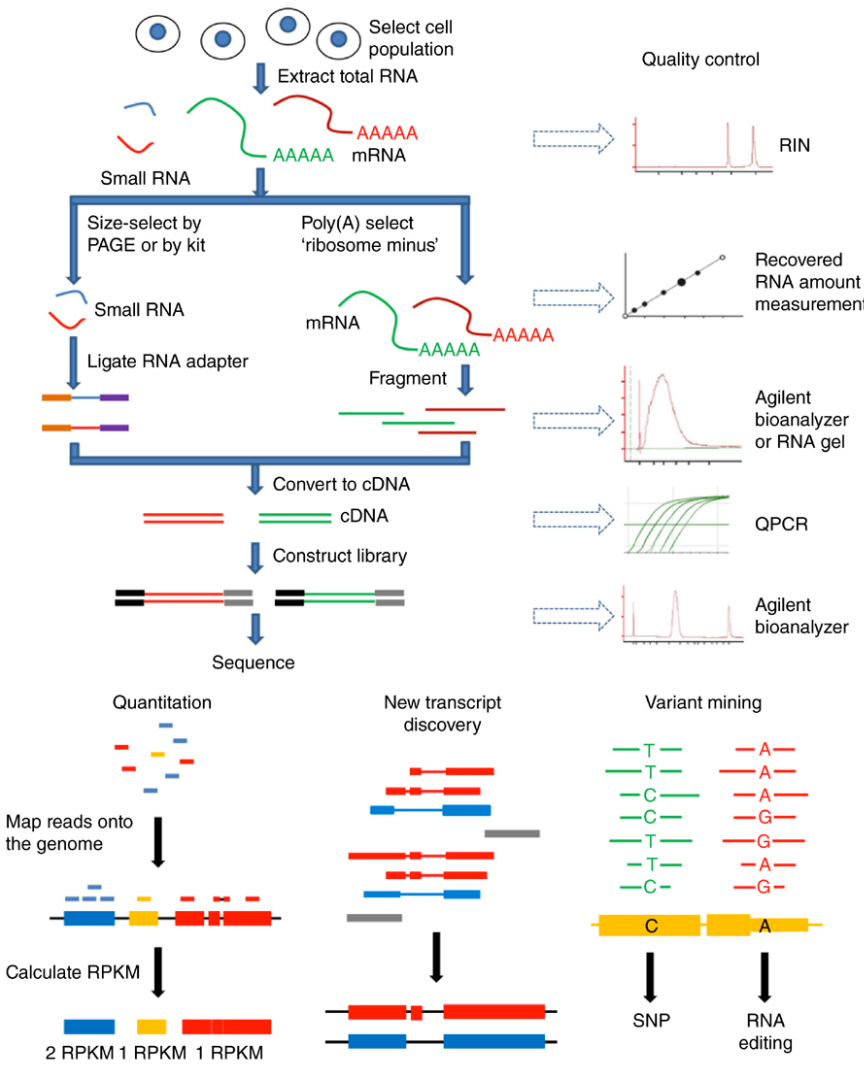
Overview

- Overview of RNA-seq
  - Alignment
  - Visualization
- Genome Annotations
  - Coordinate Systems
  - Introduction to File Formats
  - Explore GTF File
- RNA-seq analysis
  - Quantification
  - Differential Expression
- Astrocyte Workflows

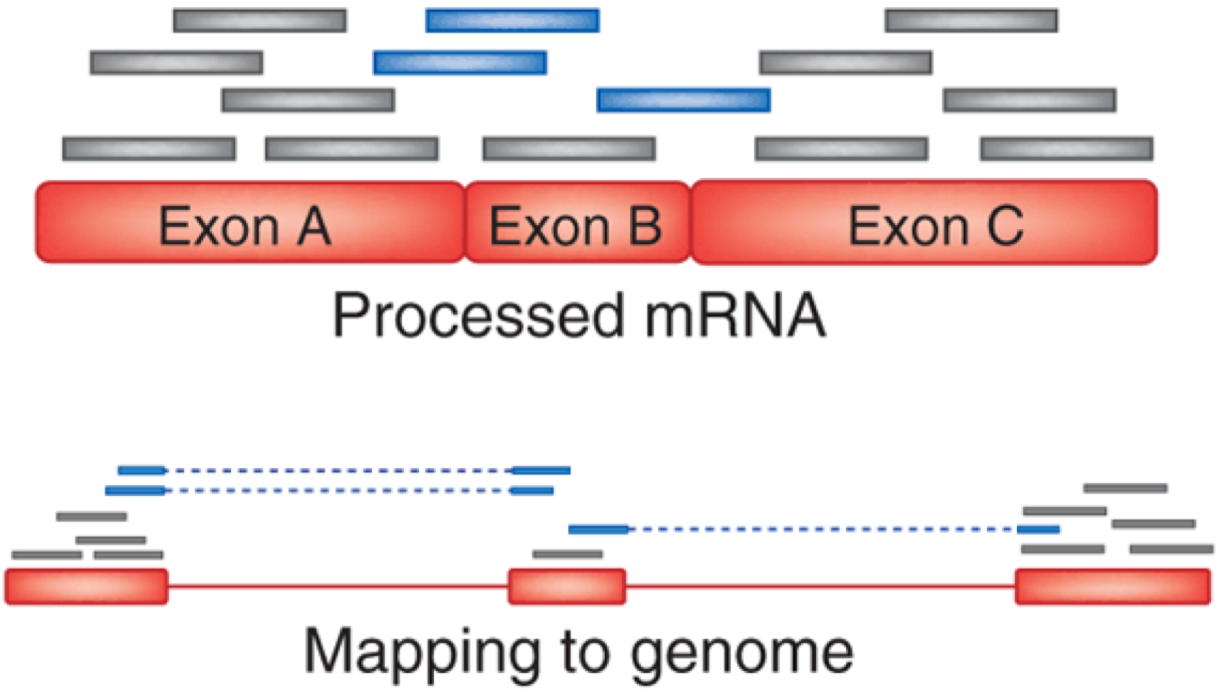


Intro to RNA-seq and review

Overview of RNA-seq

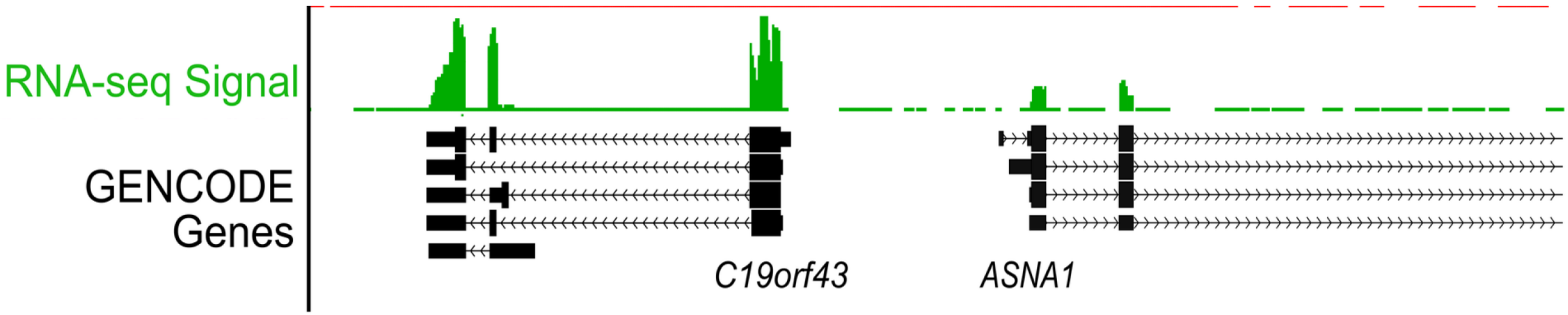


Alignment to Genome and Reference Annotations



https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

Visualization of Alignment



Review of Alignment and Intro to RNA-seq

What is a genome annotation?

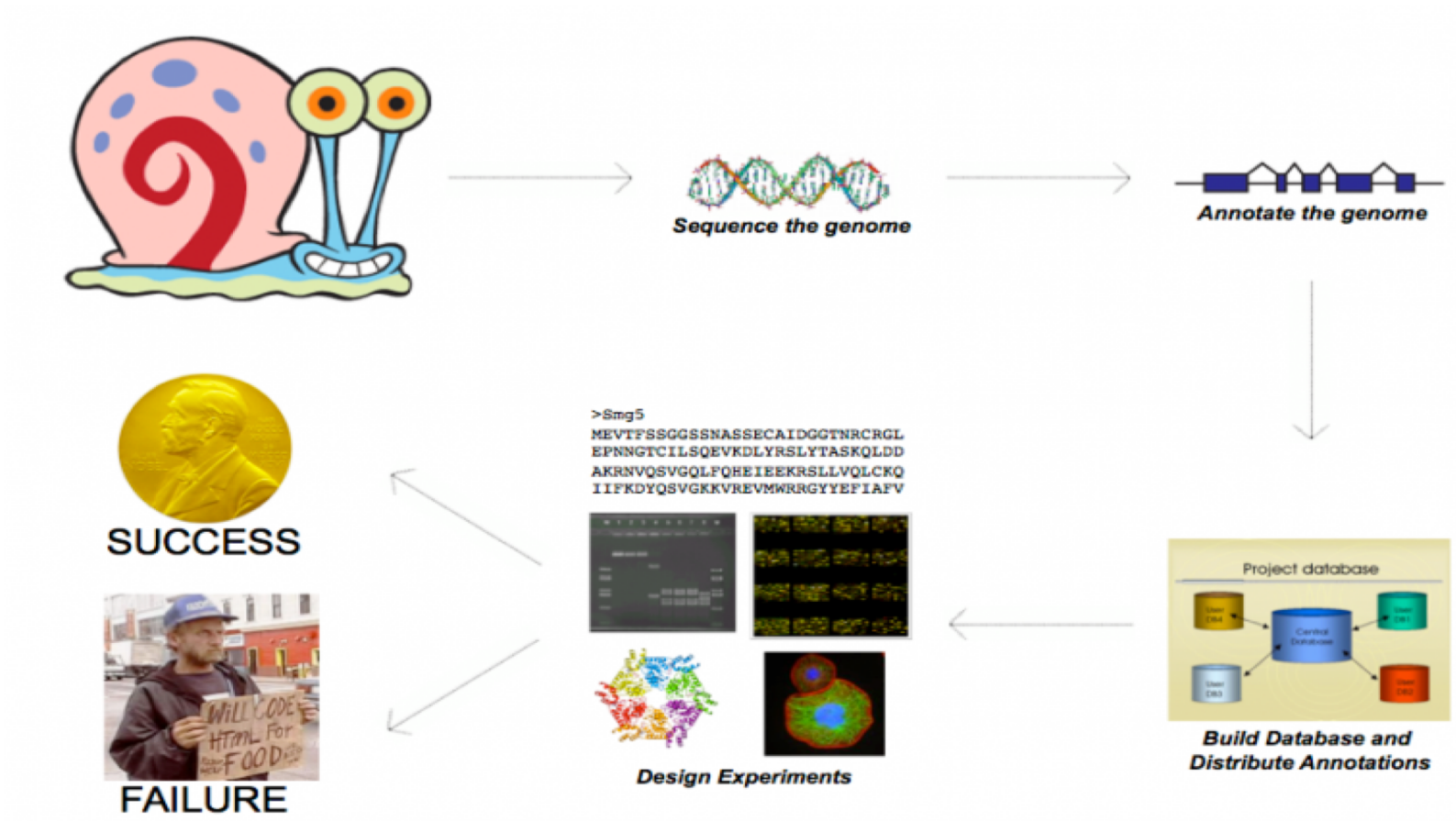
CTGCCGTCTGCTGCCATCGGAGCCCAAAGCCGGGCTGTGACTGCTCAGAC
CAGCCGGCTGGAGGGAGGGGCTCAGCAGGTCTGGCTTTGGCCCTGGGAGA
GCAGGTGGAAGATCAGGCAGGCCATCGCTGCCACAGAACCCAGTGGATTG
GCCTAGGTGGGATCTCTGAGCTCAACAAGCCCTCTCTGGGTGGTAGGTGC
AGAGACGGGAGGGGCAGAGCCGCAGGCACAGCCAAGAGGGCTGAAGAAAT
GGTAGAACGGAGCAGCTGGTGATGTGTGGGCCCACCGGCCCCAGGCTCCT
GTCTCCCCCCAGGTGTGTGGTGATGCCAGGCATGCCCTTCCCCAGCATCA
GGTCTCCAGAGCTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTG
AGTGTCCCCAGTGTTGCAGAGGTGAGAGGAGAGTAGACAGTGAGTGGGAG
TGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCGTCTCCTGTCTCCTGGAG
AGGCTTCGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTCATCT
GGAGCCCTGCTGCTTGCGGTGGCCTATAAAGCCTCCTAGTCTGGCTCCAA
GGCCTGGCAGAGTCTTTCCCAGGGAAAGCTACAAGCAGCAAACAGTCTGC
ATGGGTCATCCCCTTCACTCCCAGCTCAGAGCCCAGGCCAGGGGCCCCCA
AGAAAGGCTCTGGTGGAGAACCTGTGCATGAAGGCTGTCAACCAGTCCAT
AGGCAAGCCTGGCTGCCTCCAGCTGGGTCGACAGACAGGGGCTGGAGAAG
GGGAGAAGAGGAAAGTGAGGTTGCCTGCCCTGTCTCCTACCTGAGGCTGA
GGAAGGAGAAGGGGATGCACTGTTGGGGAGGCAGCTGTAACTCAAAGCCT
TAGCCTCTGTTCCCACGAAGGCAGGGCCATCAGGCACCAAAGGGATTCTG
CCAGCATAGTGCTCCTGGACCAGTGATACACCCGGCACCCTGTCCTGGAC
ACGCTGTTGGCCTGGATCTGAGCCCTGGTGGAGGTCAAAGCCACCTTTGG
TTCTGCCATTGCTGCTGTGTGGAAGTTCACTCCTGCCTTTTCCTTTCCCT
AGAGCCTCCACCACCCCGAGATCACATTTCTCACTGCCTTTTGTCTGCCC
AGTTTCACCAGAAGTAGGCCTCTTCCTGACAGGCAGCTGCACCACTGCCT
GGCGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACGGTGTTTGTCATG

What is an annotation?

- Gene
- Repeat
- Disease Locus
- Something interesting

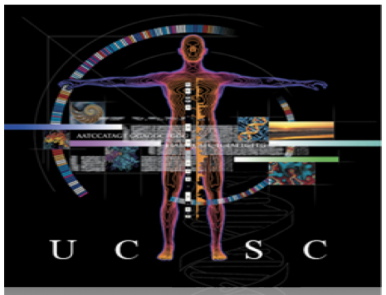
```
CTGCCGTCTGCTGCCATCGGAGCCCAAAGCCGGGCTGTGACTGCTCAGAC
CAGCCGGCTGGAGGGAGGGGCTCAGCAGGTCTGGCTTTGGCCCTGGGAGA
GCAGGTGGAAGATCAGGCAGGCCATCGCTGCCACAGAACCCAGTGGATTG
GCCTAGGTGGGATCTCTGAGCTCAACAAGCCCTCTCTGGGTGGTAGGTGC
AGAGACGGGAGGGGCAGAGCCGCAGGCACAGCCAAGAGGGCTGAAGAAAT
GGTAGAACGGAGCAGCTGGTGATGTGTGGGCCCACCGGCCCCAGGCTCCT
GTCTCCCCCCAGGTGTGTGGTGATGCCAGGCATGCCCTTCCCCAGCATCA
GGTCTCCAGAGCTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTG
AGTGTCCCCAGTGTTGCAGAGGTGAGAGGAGAGTAGACAGTGAGTGGGAG
TGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCGTCTCCTGTCTCCTGGAG
AGGCTTCGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTCATCT
GGAGCCCTGCTGCTTGCGGTGGCCTATAAAGCCTCCTAGTCTGGCTCCAA
GGCCTGGCAGAGTCTTTCCCAGGGAAAGCTACAAGCAGCAAACAGTCTGC
ATGGGTCATCCCCTTCACTCCCAGCTCAGAGCCCAGGCCAGGGGCCCCCA
AGAAAGGCTCTGGTGGAGAACCTGTGCATGAAGGCTGTCAACCAGTCCAT
AGGCAAGCCTGGCTGCCTCCAGCTGGGTCGACAGACAGGGGCTGGAGAAG
GGGAGAAGAGGAAAGTGAGGTTGCCTGCCCTGTCTCCTACCTGAGGCTGA
GGAAGGAGAAGGGGATGCACTGTTGGGGAGGCAGCTGTAACTCAAAGCCT
TAGCCTCTGTTCCCACGAAGGCAGGGCCATCAGGCACCAAAGGGATTCTG
CCAGCATAGTGCTCCTGGACCAGTGATACACCCGGCACCCTGTCCTGGAC
ACGCTGTTGGCCTGGATCTGAGCCCTGGTGGAGGTCAAAGCCACCTTTGG
TTCTGCCATTGCTGCTGTGTGGAAGTTCACTCCTGCCTTTTCCTTTCCCT
AGAGCCTCCACCACCCCGAGATCACATTTCTCACTGCCTTTTGTCTGCCC
AGTTTCACCAGAAGTAGGCCTCTTCCTGACAGGCAGCTGCACCACTGCCT
GGCGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACGGTGTTTGTCATG
```

Why are genome annotations important?



Slide from Mark Yandell

Annotations Provide Context



Conservation
Repeat elements
Genome Gaps
Cytobands
Gene annotations
"Mappability"
DeCIPHER
ISGA

dbSNP
Short Genetic Variations

gnomAD browser

ClinVar

OMIM
Online Mendelian Inheritance in Man

1000 Genomes
A Deep Catalog of Human Genetic Variation

Genetic variation

...CCTCATGCATGGAAA...
...CCTCATGTATGGAAA...
...CCTCATGCATGGAAA...
...CCTCATGCATGGAAA...
...CCTCATGTATGGAAA...
...CCTCATGCATGGAAA...
...CCTCATGTATGGAAA...

Pfam

KEGG
Kyoto Encyclopedia of Genes and Genomes

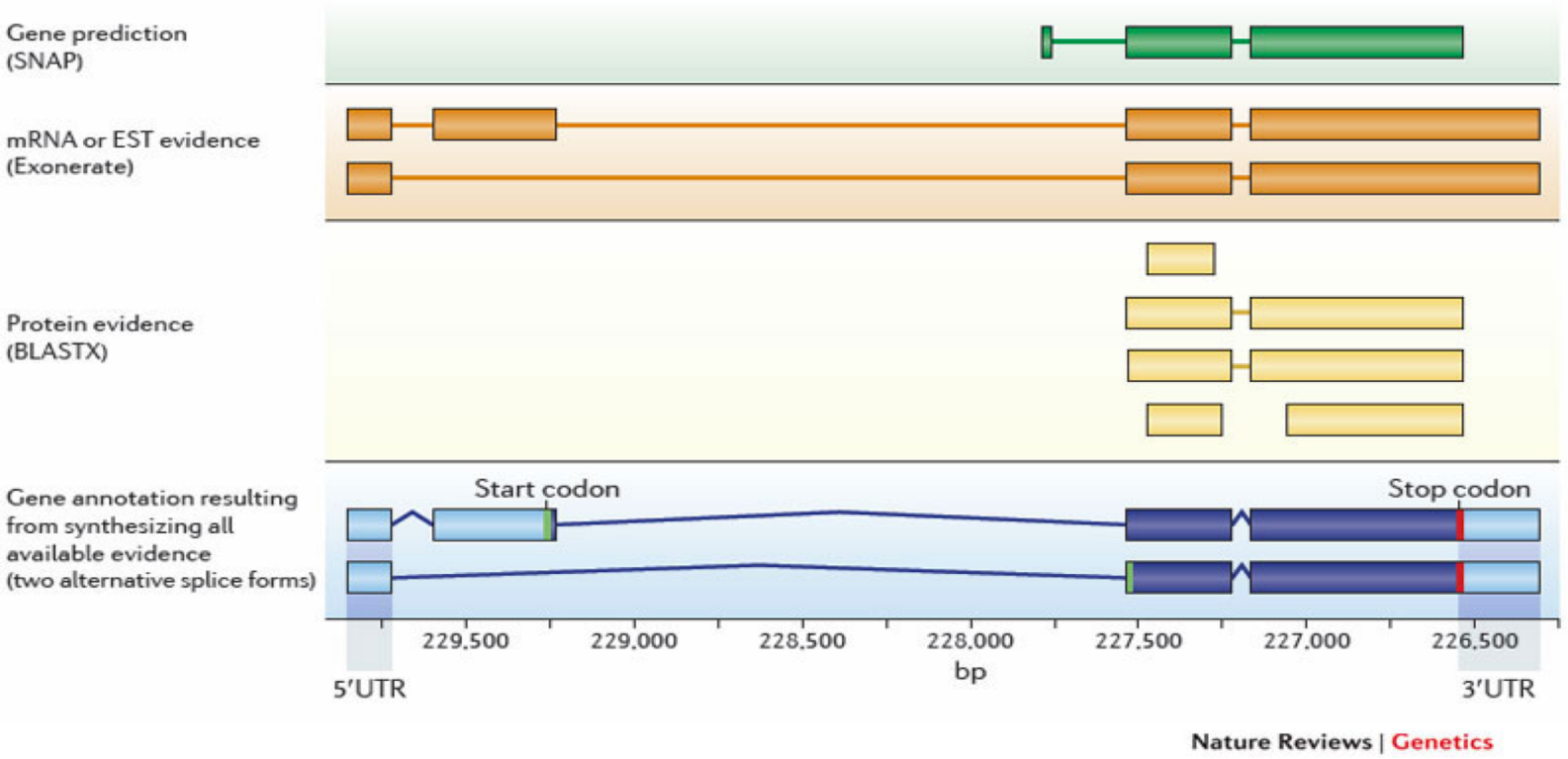
HGMD®

Human Protein Reference Database

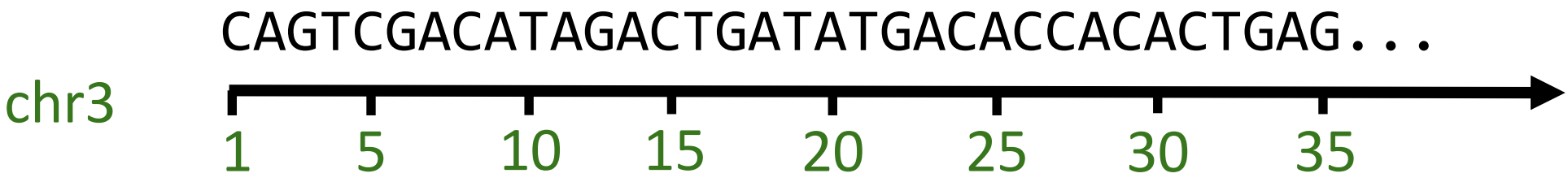
ENCODE

Chromatin marks
DNA methylation
RNA expression
TF binding

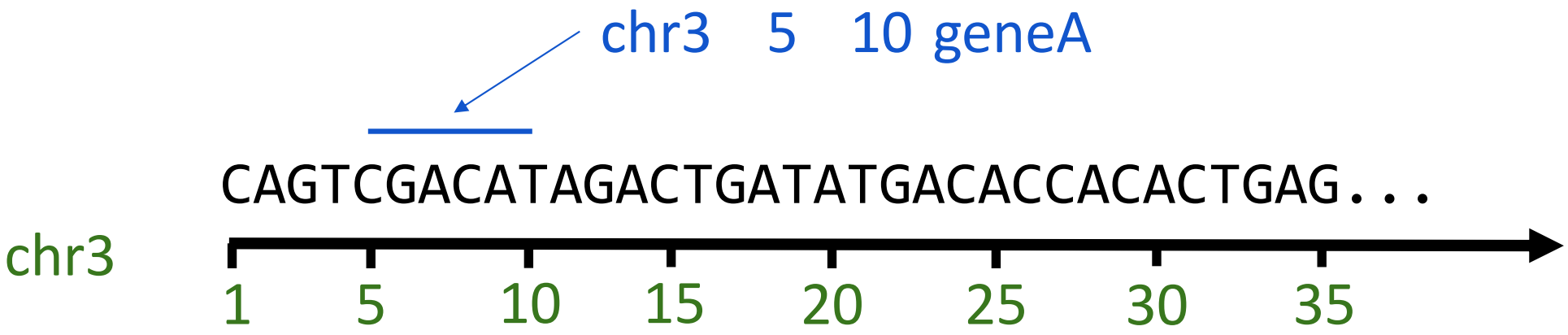
How do we annotate the genome?



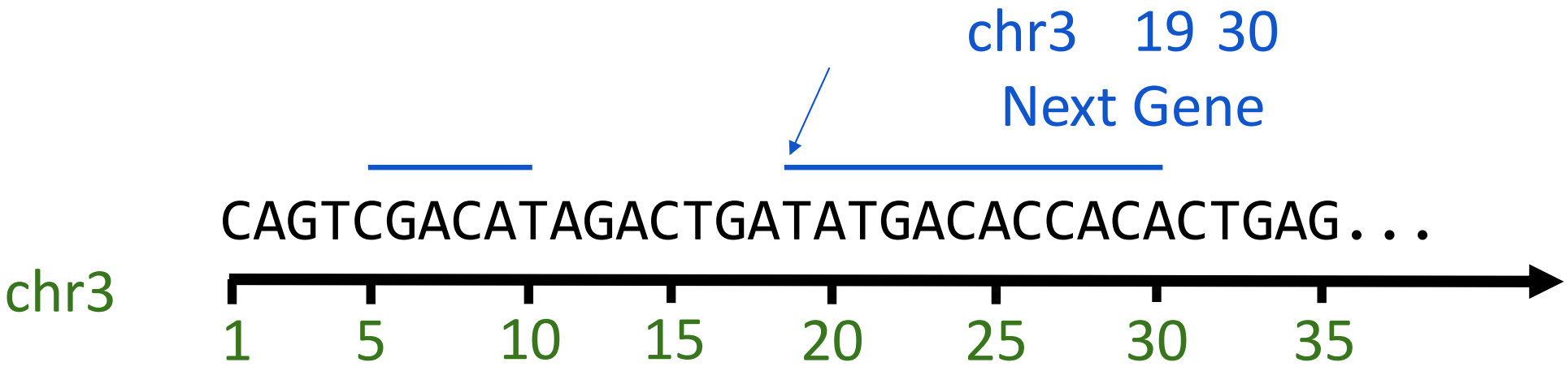
The genome as coordiantes



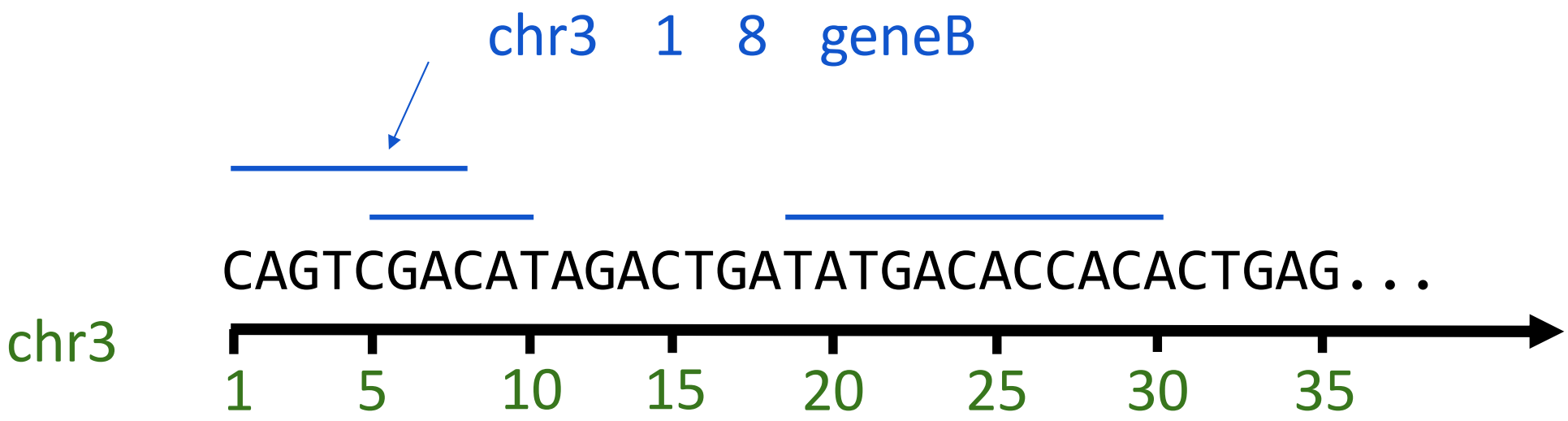
How do you describe a gene?



What about 2 consecutive genes?



What about overlapping genes?



File Formats

- BED (Browser Extensible Format)
- GFF (General Feature Format)
- GTF (Gene Transfer Format, GTF2.2)

BED File Format

BED format Index ▷

BED format provides a flexible way to define the data lines that are displayed in an annotation track. BED lines have three required fields and nine additional optional fields. The number of fields per line must be consistent throughout any single set of data in an annotation track. The order of the optional fields is binding: lower-numbered fields must always be populated if higher-numbered fields are used.

If your data set is BED-like, but it is very large (over 50MB) and you would like to keep it on your own server, you should use the bigBed data format.

The first three required BED fields are:

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as *chromStart=0, chromEnd=100*, and span the bases numbered 0-99.

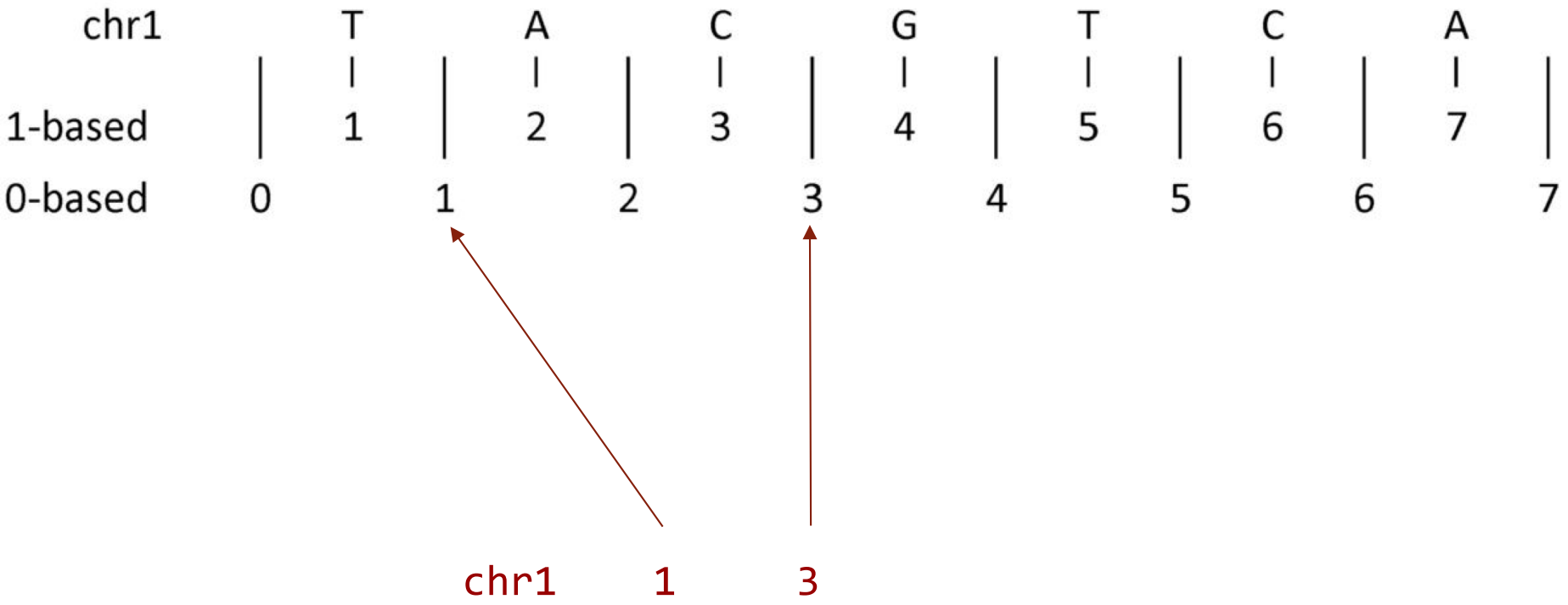
The 9 additional optional BED fields are:

4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser window when the track is open to full display mode or directly to the left of the item in pack mode.
5. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the *score* value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). This table shows the Genome Browser's translation of BED score values into shades of gray:

shade								
score in range	≤ 166	167-277	278-388	389-499	500-611	612-722	723-833	834-944	≥ 945

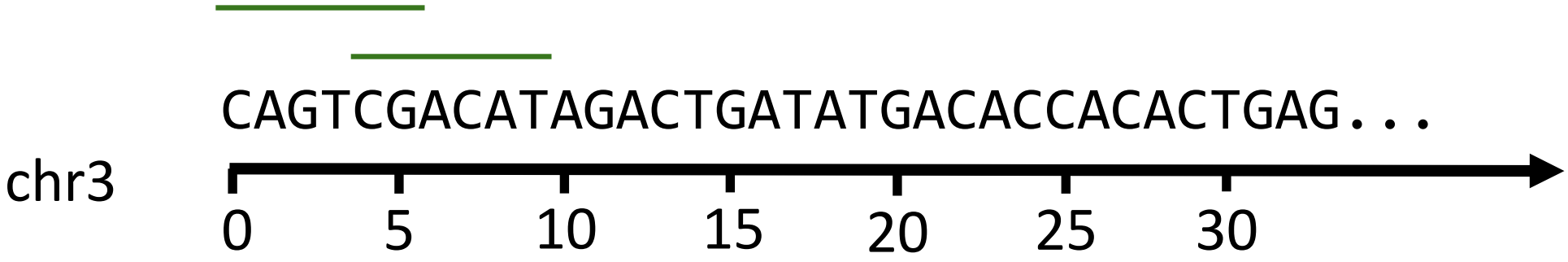
6. **strand** - Defines the strand - either '+' or '-'.
7. **thickStart** - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays). When there is no thick part, thickStart and thickEnd are usually set to the chromStart position.
8. **thickEnd** - The ending position at which the feature is drawn thickly (for example, the stop codon in gene displays).
9. **itemRgb** - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line *itemRgb* attribute is set to "On", this RBG value will determine the display color of the data contained in this BED line. NOTE: It is recommended that a simple color scheme (eight colors or less) be used with this attribute to avoid overwhelming the color resources of the Genome Browser and your Internet browser.
10. **blockCount** - The number of blocks (exons) in the BED line.
11. **blockSizes** - A comma-separated list of the block sizes. The number of items in this list should correspond to *blockCount*.
12. **blockStarts** - A comma-separated list of block starts. All of the *blockStart* positions should be calculated relative to *chromStart*. The number of items in this list should correspond to *blockCount*.

BED File use 0-based, half open intervals



https://www.biostars.org/p/84686/

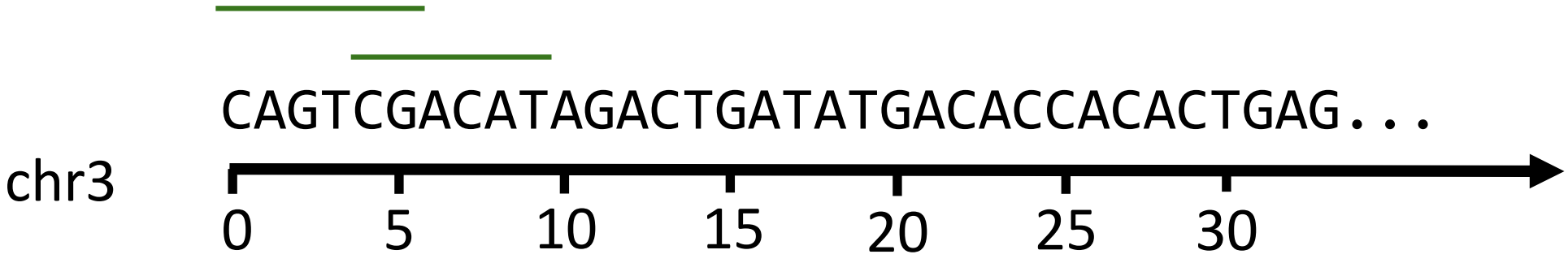
Example BED3




```
chr3    0    7
chr3    4    10
```


https://www.biostars.org/p/84686/

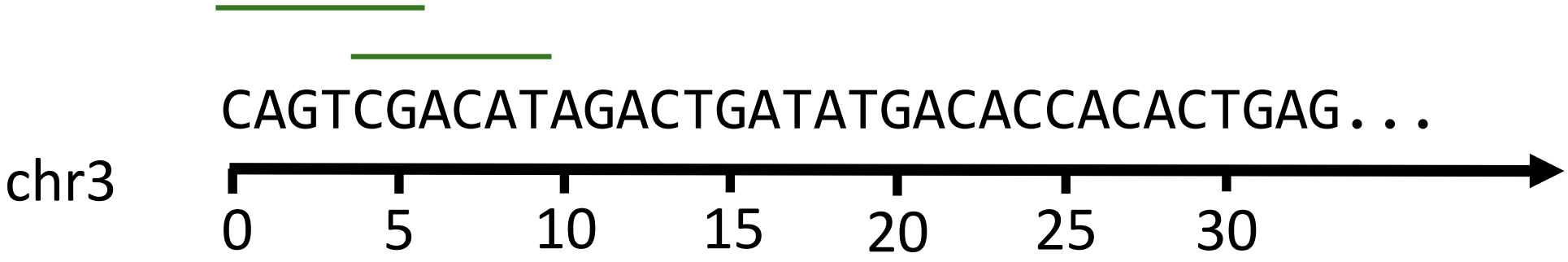
Add Name in 4th column to BED format




```
chr3    0     7    first
chr3    4     10   second
```


https://www.biostars.org/p/84686/

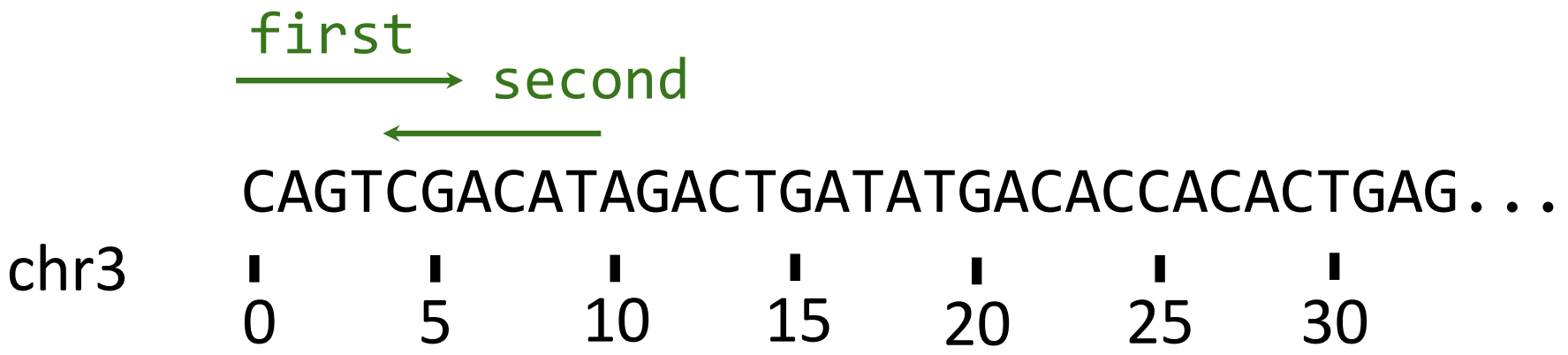
Add Intensity in 5th column to BED format




```
chr3    0     7    first    7.39
chr3    4     10   second   1e-8
```


https://www.biostars.org/p/84686/

Add orientation/strand to the 6th field




```
chr3    0     7    first    7.39  +
chr3    4     10   second   1e-8  -
```


https://www.biostars.org/p/84686/

BED can also track Genome Annotations

BED format

Index ▷

BED format provides a flexible way to define the data lines that are displayed in an annotation track. BED lines have three required fields and nine additional optional fields. The number of fields per line must be consistent throughout any single set of data in an annotation track. The order of the optional fields is binding: lower-numbered fields must always be populated if higher-numbered fields are used.

If your data set is BED-like, but it is very large (over 50MB) and you would like to keep it on your own server, you should use the bigBed data format.

The first three required BED fields are:

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as *chromStart=0, chromEnd=100*, and span the bases numbered 0-99.

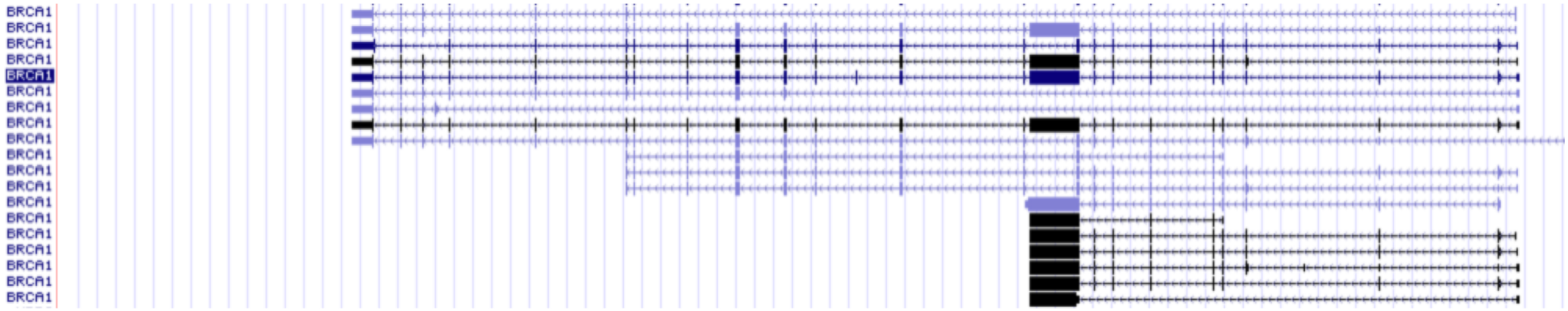
The 9 additional optional BED fields are:

4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser window when the track is open to full display mode or directly to the left of the item in pack mode.
5. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the *score* value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). This table shows the Genome Browser's translation of BED score values into shades of gray:

shade									
score in range	≤ 166	167-277	278-388	389-499	500-611	612-722	723-833	834-944	≥ 945

6. **strand** - Defines the strand - either '+' or '-'.
7. **thickStart** - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays). When there is no thick part, thickStart and thickEnd are usually set to the chromStart position.
8. **thickEnd** - The ending position at which the feature is drawn thickly (for example, the stop codon in gene displays).
9. **itemRgb** - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line *itemRgb* attribute is set to "On", this RBG value will determine the display color of the data contained in this BED line. NOTE: It is recommended that a simple color scheme (eight colors or less) be used with this attribute to avoid overwhelming the color resources of the Genome Browser and your Internet browser.
10. **blockCount** - The number of blocks (exons) in the BED line.
11. **blockSizes** - A comma-separated list of the block sizes. The number of items in this list should correspond to *blockCount*.
12. **blockStarts** - A comma-separated list of block starts. All of the *blockStart* positions should be calculated relative to *chromStart*. The number of items in this list should correspond to *blockCount*.

BED12 Example




```
chr17   41196311        41277340        uc010whm.2      0       -       41197694        41277202        0       8       1508,61,74,55,84,41,78,142,     0,3348,4826,6768,12757,19038,19579,80887,
chr17   41196311        41277340        uc002icp.4      0       -       41197694        41258496        0       23      1508,61,74,55,84,41,78,88,311,191,127,172,89,3426,77,46,103,140,89,56,54,99,142,
0,3348,4826,6768,12757,19038,19579,23313,26633,30036,32193,38109,46649,47140,51551,52949,55480,59827,60573,62183,71431,79722,80887,
chr17   41196311        41277468        uc002icu.3      0       -       41197800        41276113        0       22      1508,61,55,84,41,78,88,311,191,124,172,89,117,77,46,106,140,89,78,54,99,175,
0,3348,6768,12757,19038,19579,23313,26633,30036,32193,38109,46649,50449,51551,52949,55480,59827,60573,62161,71431,79722,80982,
chr17   41196311        41277468        uc010cyx.3      0       -       41197694        41258543        0       22      1508,61,74,55,84,41,78,88,311,191,127,172,89,3426,77,46,106,140,89,78,99,175,
0,3348,4826,6768,12757,19038,19579,23313,26633,30036,32193,38109,46649,47140,51551,52949,55480,59827,60573,62161,79722,80982,
chr17   41196311        41277500        uc002ict.3      0       -       41197694        41276113        0       24      1508,61,74,55,84,41,78,88,311,191,124,66,172,89,3426,77,46,106,140,89,78,54,99,213,
0,3348,4826,6768,12757,19038,19579,23313,26633,30036,32193,35039,38109,46649,47140,51551,52949,55480,59827,60573,62161,71431,79722,80976,
chr17   41196311        41277500        uc010whn.2      0       -       41197694        41226495        0       11      1508,61,74,55,84,41,78,88,311,191,213,  0,3348,4826,6768,12757,19038,19579,23313,26633,30036,80976,
chr17   41196311        41277500        uc010who.3      0       -       41197694        41202109        0       5       1508,61,74,129,213,     0,3348,4826,5767,80976,
chr17   41196311        41277500        uc002icq.3      0       -       41197694        41276113        0       23      1508,61,74,55,84,41,78,88,311,191,127,172,89,3426,77,46,106,140,89,78,54,99,213,
0,3348,4826,6768,12757,19038,19579,23313,26633,30036,32193,38109,46649,47140,51551,52949,55480,59827,60573,62161,71431,79722,80976,
chr17   41196311        41322420        uc010whp.2      0       -       41197694        41258543        0       22      1508,61,74,55,84,41,78,88,311,191,124,172,89,117,77,46,106,140,89,78,54,278,
0,3348,4826,6768,12757,19038,19579,23313,26633,30036,32193,38109,46649,50449,51551,52949,55480,59827,60573,62161,71431,125831,
chr17   41215349        41256973        uc010whq.1      0       -       41215349        41256198        0       12      41,78,88,311,191,127,172,89,117,106,140,89,
0,541,4275,7595,10998,13155,19071,27611,31411,36442,40789,41535,
chr17   41215349        41277468        uc002idc.1      0       -       41215349        41276113        0       18      41,78,88,311,191,127,172,89,117,77,46,103,140,89,78,54,99,175,
0,541,4275,7595,10998,13155,19071,27611,31411,32513,33911,36442,40789,41535,43123,52393,60684,61944,
chr17   41215349        41277468        uc010whr.1      0       -       41215349        41258543        0       17      41,78,88,311,191,127,172,89,117,77,46,106,140,89,78,99,175,
0,541,4275,7595,10998,13155,19071,27611,31411,32513,33911,36442,40789,41535,43123,60684,61944,
chr17   41243116        41276132        uc002idd.3      0       -       41243347        41276113        0       9       3761,77,46,106,140,89,78,54,99, 0,4746,6144,8675,13022,13768,15356,24626,32917,
chr17   41243451        41256973        uc002ide.1      0       -       41243452        41256198        0       4       3426,103,140,89,        0,8340,12687,13433,
chr17   41243451        41277340        uc010cyy.1      0       -       41243452        41276113        0       10      3426,77,46,106,140,89,78,54,99,142,     0,4411,5809,8340,12687,13433,15021,24291,32582,33747,
chr17   41243451        41277468        uc010whs.1      0       -       41243452        41276113        0       10      3426,77,46,106,140,89,78,54,99,175,     0,4411,5809,8340,12687,13433,15021,24291,32582,33842,
chr17   41243451        41277500        uc010cyz.2      0       -       41243452        41258543        0       11      3426,77,46,106,140,89,78,116,54,99,213, 0,4411,5809,8340,12687,13433,15021,19030,24291,32582,33836,
chr17   41243451        41277500        uc010cza.2      0       -       41243452        41276113        0       9       3426,77,46,106,140,89,54,99,213,        0,4411,5809,8340,12687,13433,24291,32582,33836,
chr17   41243451        41277500        uc010wht.1      0       -       41243452        41246659        0       2       3426,213,       0,33836,
chr17   41277599        41292342        uc002idf.3      0       +       41277599        41277599        0       4       188,63,182,1669,        0,5625,7373,13074,
chr17   41277599        41292342        uc010czb.2      0       +       41277599        41277599        0       2       188,1669,       0,13074,
chr17   41277599        41297125        uc002idg.3      0       +       41277599        41277599        0       5       188,63,266,468,381,     0,5625,13074,14233,19145,
chr17   41277599        41305688        uc002idh.3      0       +       41277599        41277599        0       8       188,63,125,182,205,266,120,70,  0,5625,7016,7373,12573,13074,16874,28019,
```

GTF/GFF

GFF (General Feature Format)

GTF (General Transfer Format)

Note: GFFv2 == GTF

Fields:

1. Seqname – name of chrom
2. Source – name of data source
3. Feature - type
4. Start – Start position
5. End – End position
6. Score
7. Strand
8. Frame – Indicates that base of codon
9. Attribute – Semicolon list

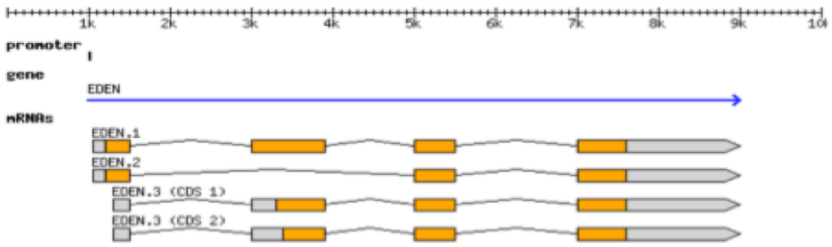
```
chr22  TeleGene enhancer  10000000  10001000  500 +  .  touch1
chr22  TeleGene promoter  10010000  10010100  900 +  .  touch1
chr22  TeleGene promoter  10020000  10025000  800 -  .  touch2
```

Note that the start and end coordinates are 1-based versus 0-based BED format

GTF/GFF Example

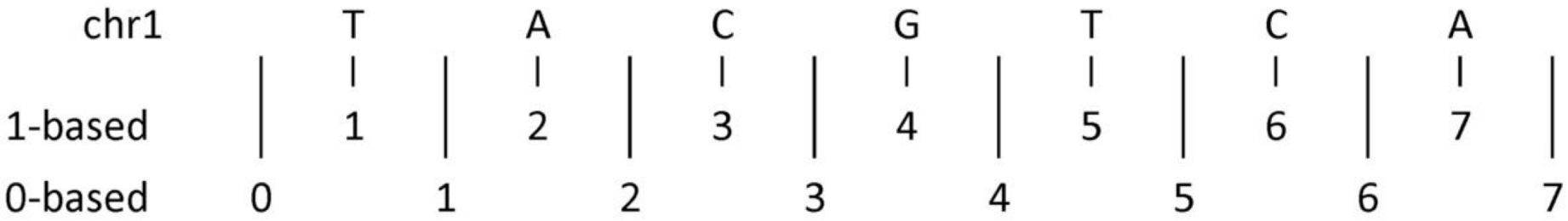
GFF Example

Gene "EDEN" with 3 alternatively spliced transcripts, isoform 3 has two alternative translation start sites




```
##gff-version 3
##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
```

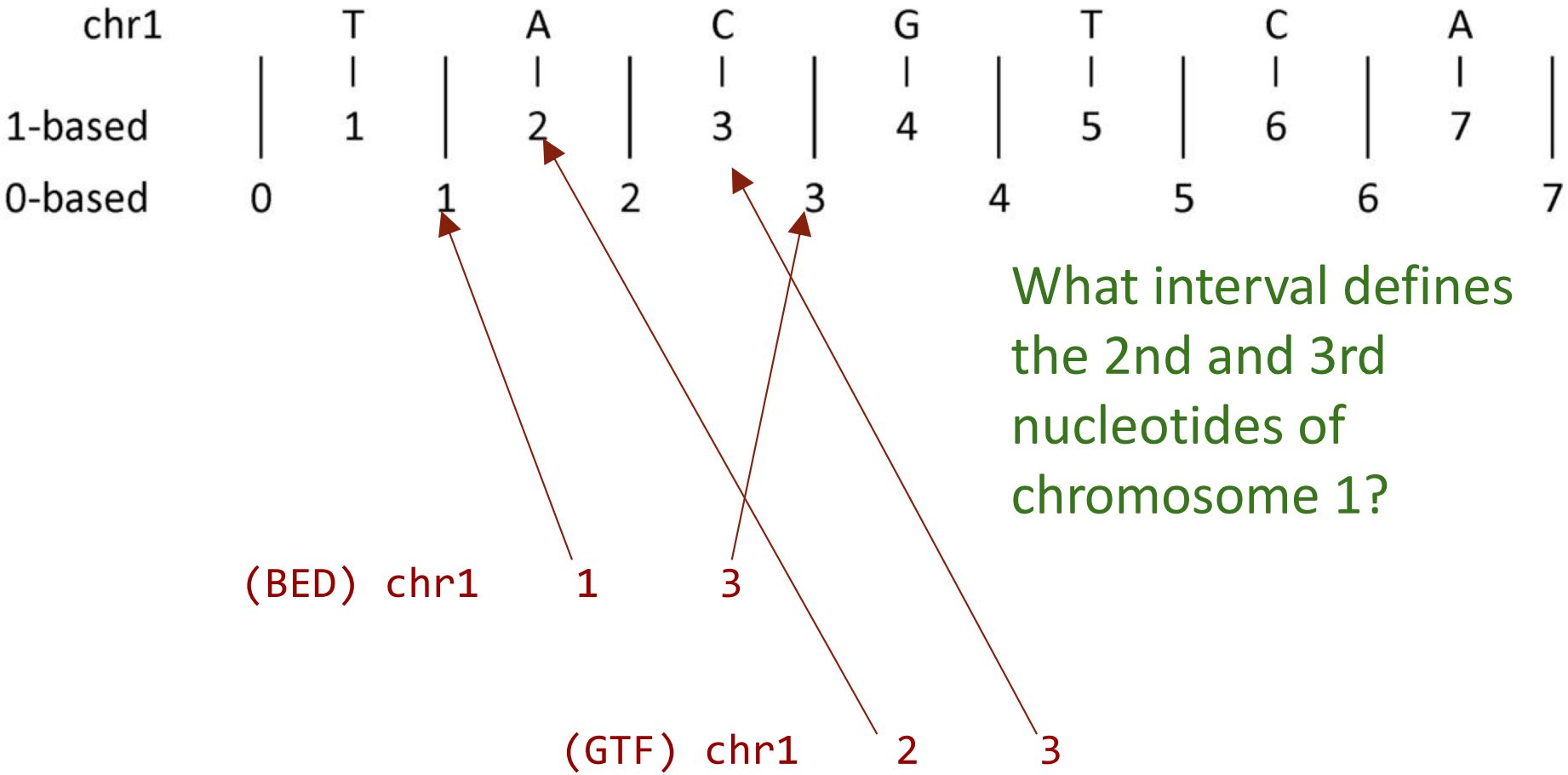
How does BED and GTF define a genomic range?



What interval defines the 2nd and 3rd nucleotides of chromosome 1?

https://www.biostars.org/p/84686/

How does BED and GTF define a genomic range?



What interval defines the 2nd and 3rd nucleotides of chromosome 1?

https://www.biostars.org/p/84686/

Not all file formats are created Equal

BED: 0-based, half-open

GFF: 1-based, closed

SAM: 1-based, closed

BAM: 0-based, half-open.

VCF: 1-based, closed

…