FRAMESHIFT

USTAR Center for Genetic Discovery

HEALTH
UNIVERSITY OF UTAH

Genome annotation formats

Alistair Ward

BED 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.

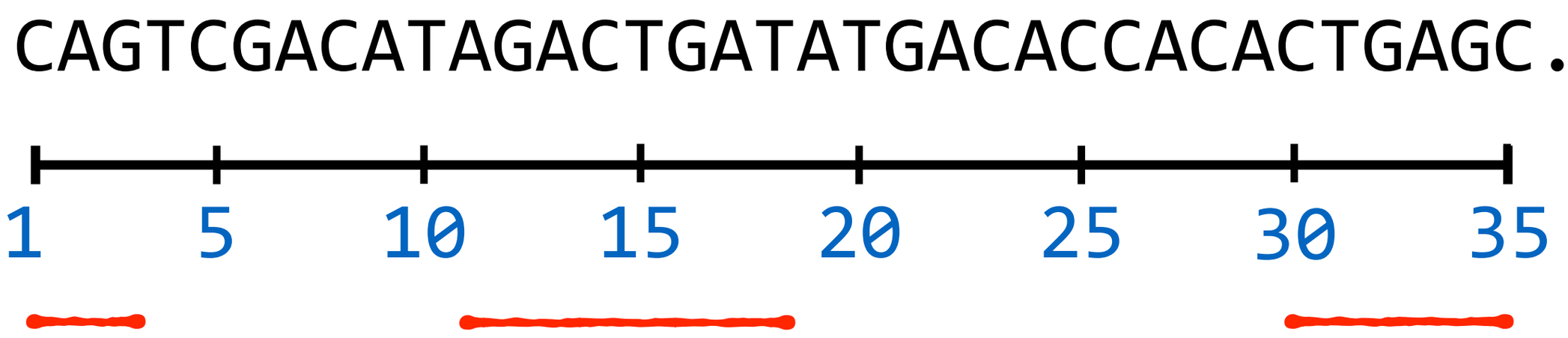
The 9 additional optional BED fields are:

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

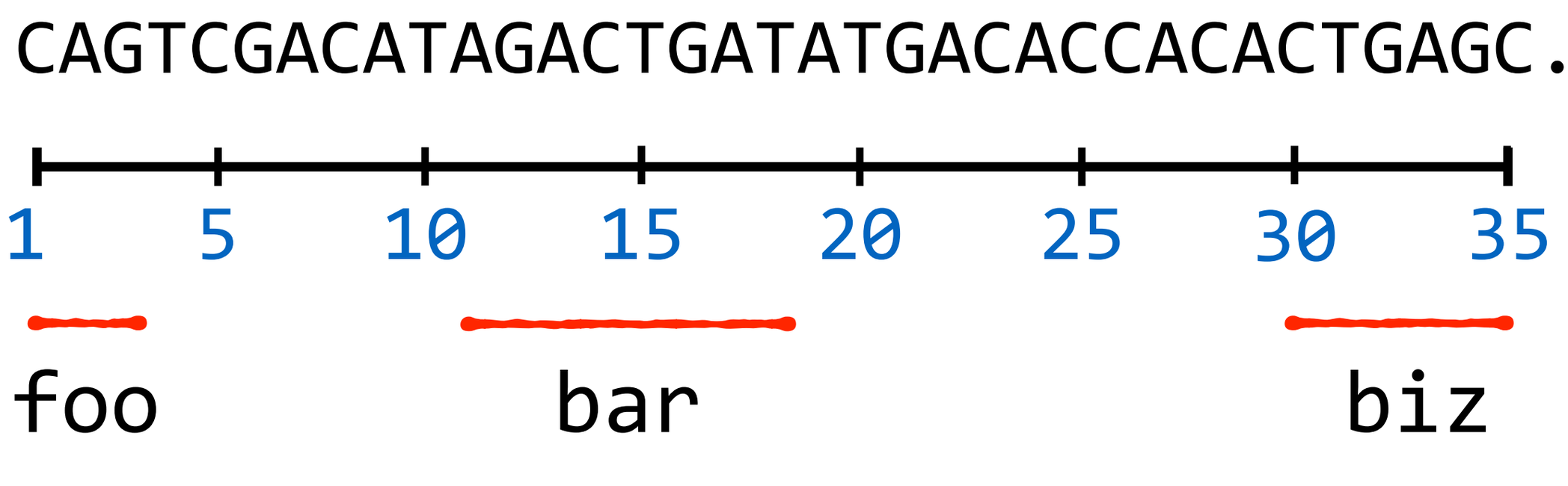
6. **strand** - Defines the strand - either '+' or '-'.
7. **thickStart** - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays). When there is no thick part, thickStart and thickEnd are usually set to the chromStart position.
8. **thickEnd** - The 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*.

Minimal BED format. So-called BED3 format.



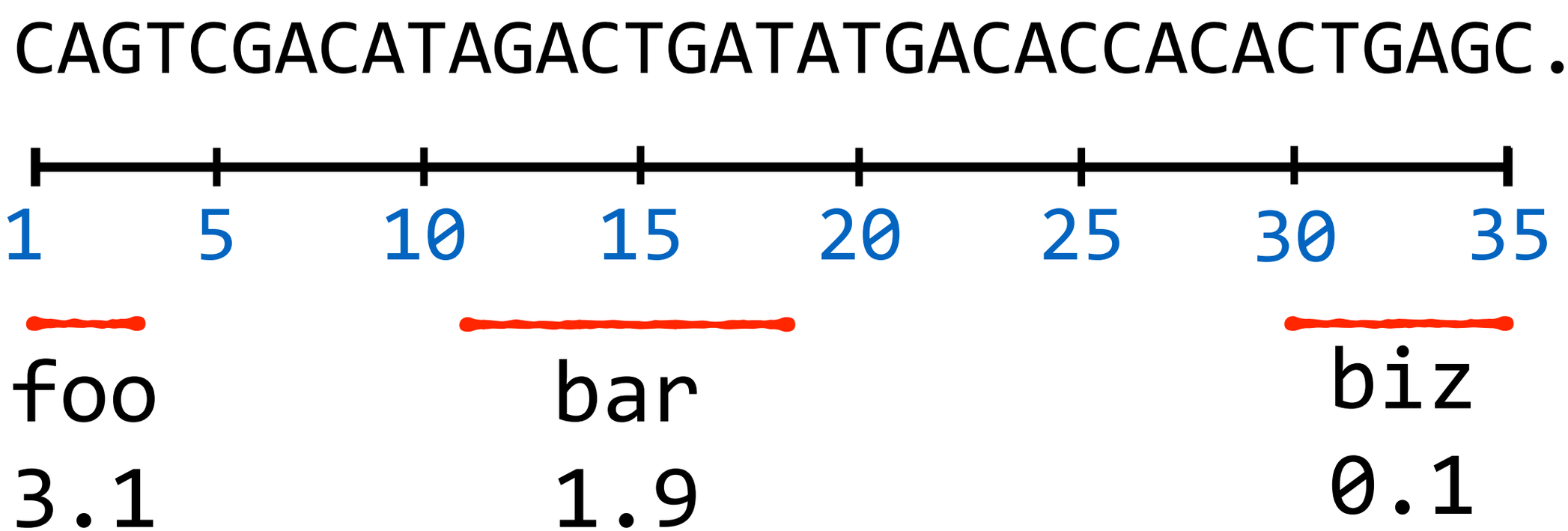

```
chr1   0    3
chr1   11   18
chr1   29   35
```

BED format supports “labels”



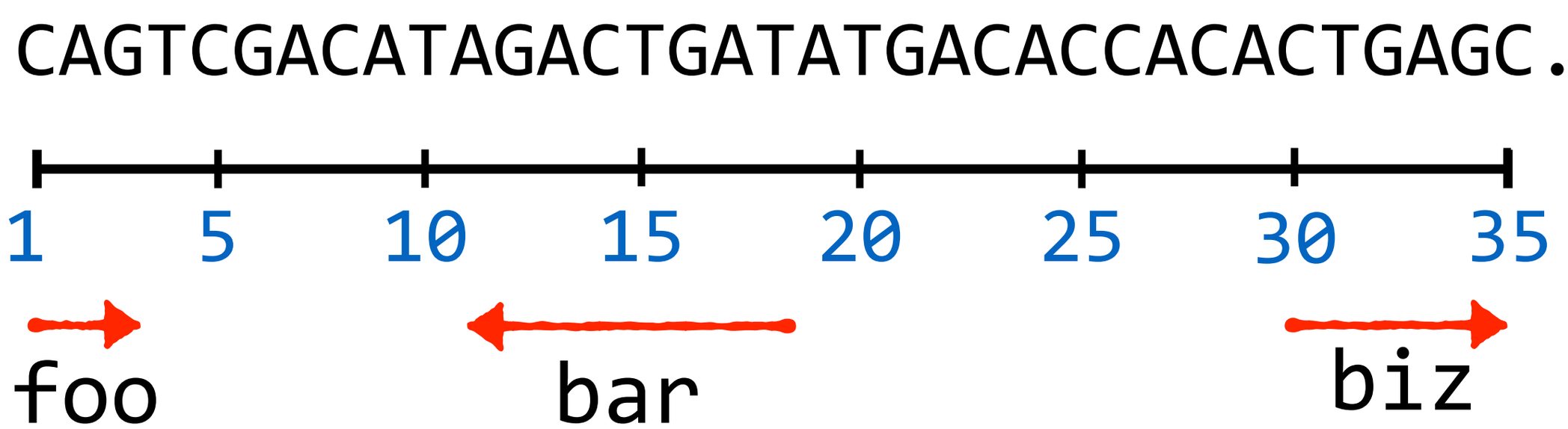

```
chr1   0    3    foo
chr1   11   18   bar
chr1   29   35   biz
```

And scores




```
chr1  0   3   foo  3.1
chr1  11  18  bar  1.9
chr1  29  35  biz  0.1
```

And strands. This is so-called BED6 format.




```
chr1   0    3    foo   3.1    +
chr1   11   18   bar   1.9    -
chr1   29   35   biz   0.1    +
```

And more! BED12 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.

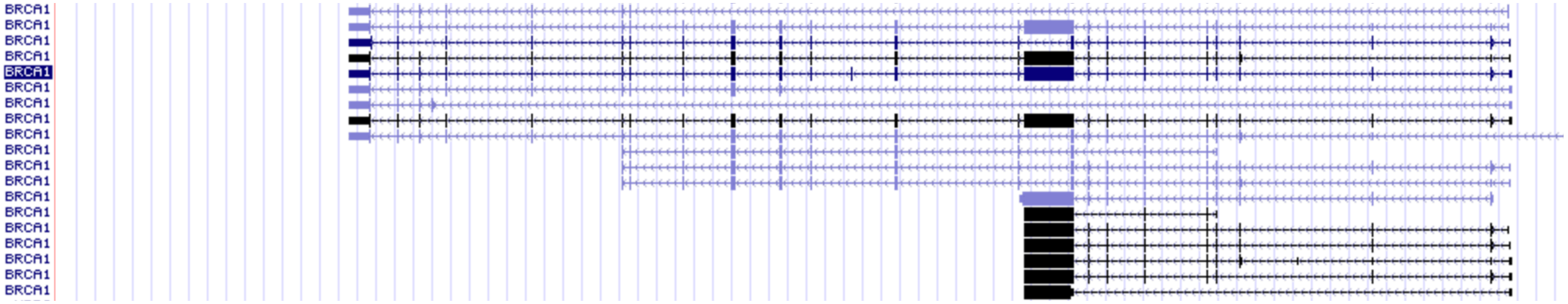
The 9 additional optional BED fields are:

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*.

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,
```

GFF format

GFF format

Index ▷

GFF (General Feature Format) lines are based on the Sanger GFF2 specification. GFF lines have nine required fields that *must* be tab-separated. If the fields are separated by spaces instead of tabs, the track will not display correctly. For more information on GFF format, refer to Sanger's GFF page.

Note that there is also a GFF3 specification that is not currently supported by the Browser. All GFF tracks must be formatted according to Sanger's GFF2 specification.

If you would like to obtain browser data in GFF (GTF) format, please refer to Genes in gtf or gff format on the Wiki.

Here is a brief description of the GFF fields:

1. **seqname** - The name of the sequence. Must be a chromosome or scaffold.
2. **source** - The program that generated this feature.
3. **feature** - The name of this type of feature. Some examples of standard feature types are "CDS", "start_codon", "stop_codon", and "exon".
4. **start** - The starting position of the feature in the sequence. The first base is numbered 1.
5. **end** - The ending position of the feature (inclusive).
6. **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). If there is no score value, enter ".".
7. **strand** - Valid entries include '+', '-', or '.' (for don't know/don't care).
8. **frame** - If the feature is a coding exon, *frame* should be a number between 0-2 that represents the reading frame of the first base. If the feature is not a coding exon, the value should be '.'.
9. **group** - All lines with the same group are linked together into a single item.

Example:
Here's an example of a GFF-based track. This example can be pasted into the browser without editing. NOTE: Paste operations on some operating systems will replace tabs with spaces, which will result in an error when the GFF track is uploaded. You can circumvent this problem by pasting the URL of the above example (http://genome.ucsc.edu/goldenPath/help/regulatory.txt) instead of the text itself into the custom annotation track text box. If you encounter an error when loading a GFF track, check that the data lines contain tabs rather than spaces.

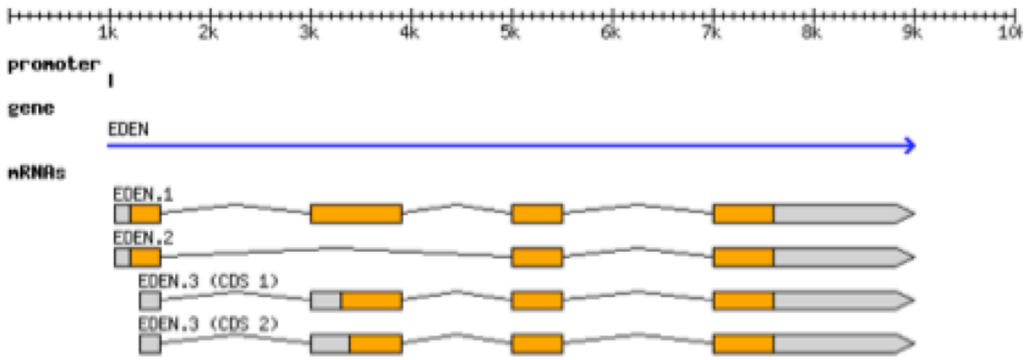
```
browser position chr22:10000000-10025000
browser hide all
track name=regulatory description="TeleGene(tm) Regulatory Regions"
visibility=2
chr22  TeleGene enhancer  10000000  10001000  500 +  .  touch1
chr22  TeleGene promoter  10010000  10010100  900 +  .  touch1
chr22  TeleGene promoter  10020000  10025000  800 -  .  touch2
```

Click here to display this track in the Genome Browser.


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

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
```

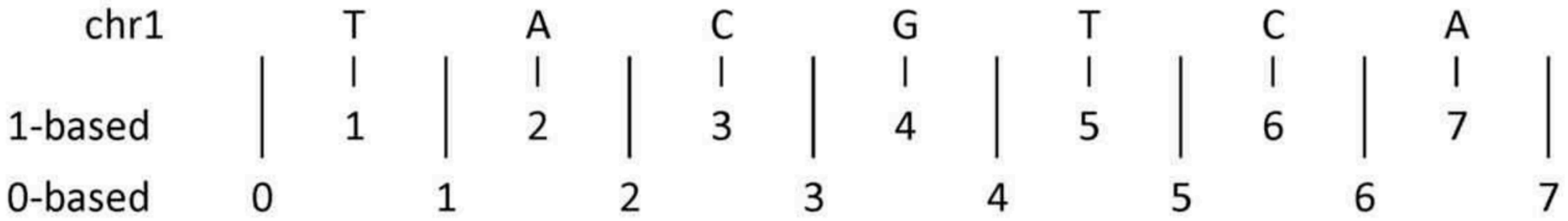
Great, genome formats using a common genome coordinate system!
My life is going to be easy.

No.

Formats use diff. coord. systems. Because of course.

BED: 0-based, half-open
GFF: 1-based, closed
SAM: 1-based, closed
BAM: 0-based, half-open. W.T.F.
VCF: 1-based, closed
…

Formats use diff. coord. systems. Because of course.



	1-based	0-based
Indicate a single nucleotide	chr1:4-4 G	chr1:3-4 G
Indicate a range of nucleotides	chr1:2-4 ACG	chr1:1-4 ACG
Indicate a single nucleotide variant	chr1:5-5 T/A	chr1:4-5 T/A

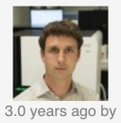
	1-based	0-based
Indicate a deletion	chr1:5-5 T/-	chr1:4-5 T/-
Indicate an insertion	chr1:3-4 -/TTA	chr1:3-3 -/TTA

Tutorial: Cheat Sheet For One-Based Vs Zero-Based Coordinate Systems

95

I am frequently faced with variant files of either 0-based or 1-based coordinates (or in the worst case, mixed!) and having to determine which I am looking at and how to convert between them. I usually go to the white board and work it out. This time, I figured I would just create a digital copy.

First, a diagram to help illustrate:



3.0 years ago by
Obi Griffith ♦ 14k
Washington University, St Louis, USA

The wonder of different formats

toy.bed

chr1	0	1
chr1	2	3
chr2	10	20

∩

toy.gff

chr1	1	2
chr1	3	3
chr2	10	20

=

intersect.bed

chr1	0	1
chr1	2	3
chr2	9	20

- inconsistent chromosome labels.
- different sorting criteria.
- mixed UNIX/Windows newlines.
- file violates spec with vigor.
- program expects exact extension.
- file is gzipp'ed, not bgzipp'ed.
- annotations use diff. genome builds.
- tool only works for one format.
- tool is hard-coded for specific build.
- tool requires act of gods to compile.

“chr” can be the bane of your existence

```
chr1   1    3   foo   3.1
chr1   12   18  bar   1.9
chr1   30   35  biz   0.1
```

is not the same as…

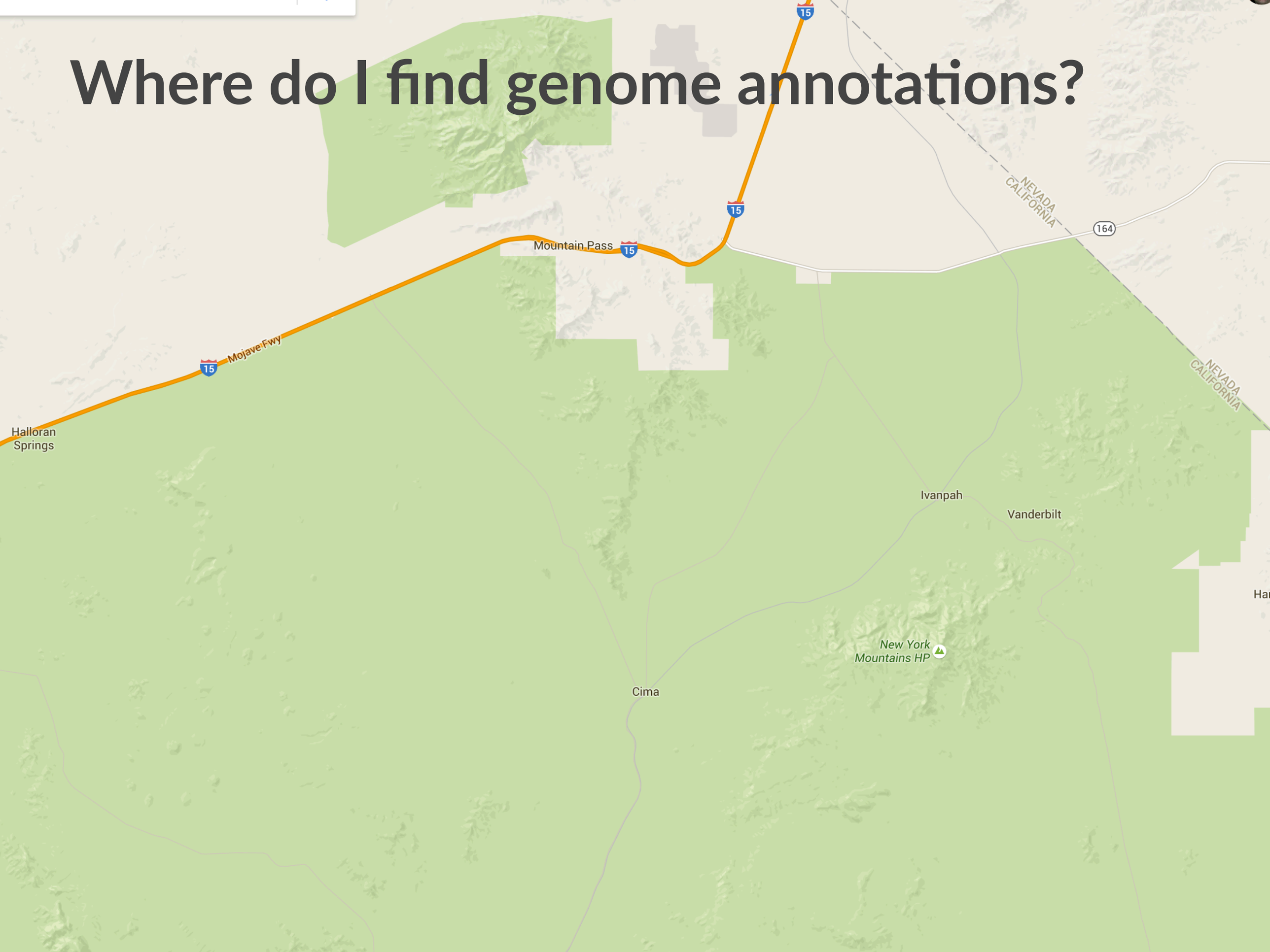
```
1      1   3   foo   3.1
1      12    18  bar   1.9
1      30    35  biz   0.1
```

SIMPLE fix though:

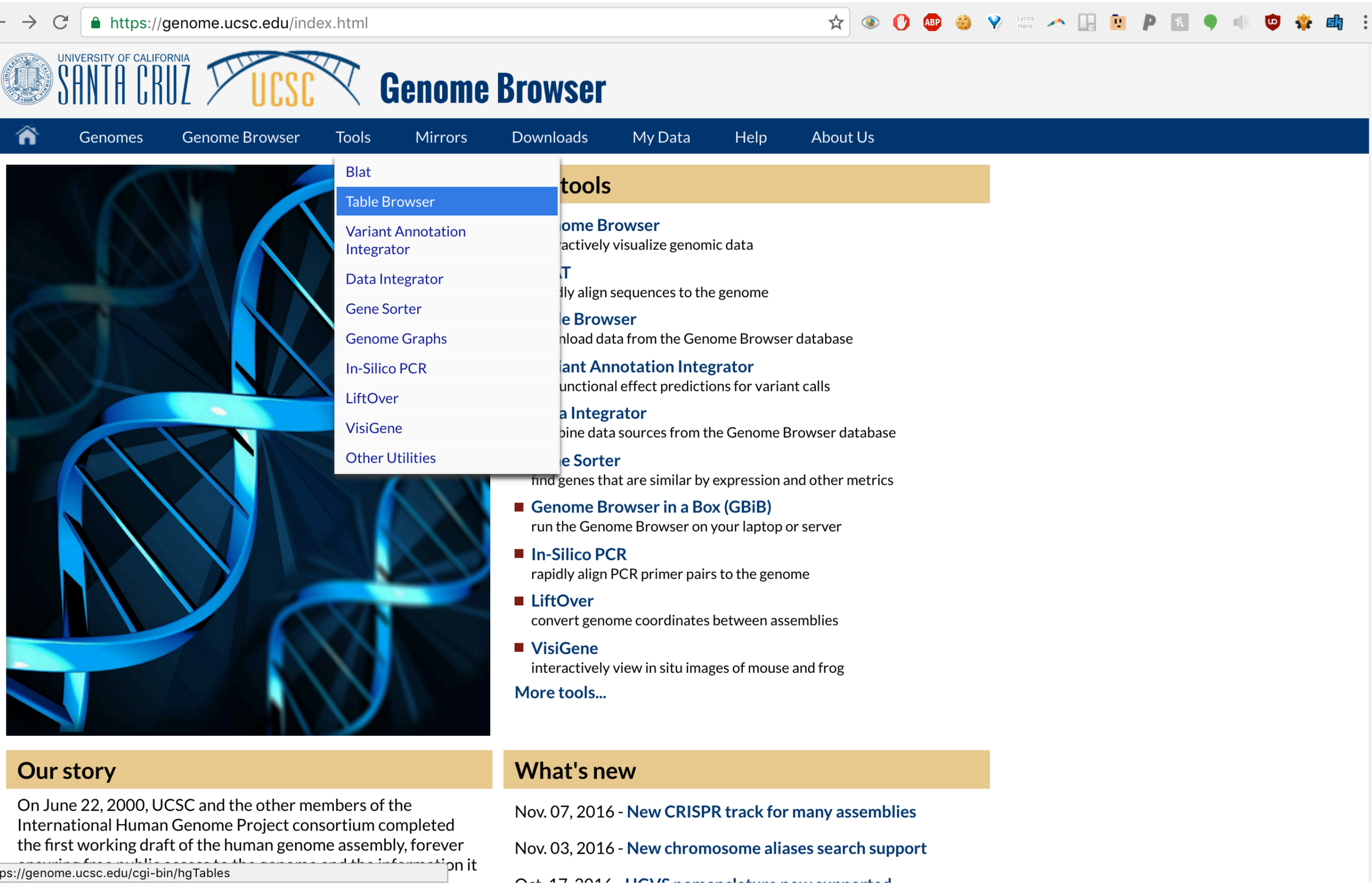
sed ‘s/^chr//g’ chr.bed > nochr.bed

sed ‘s/^/chr/g’ nochr.bed > chr.bed

Where do I find genome annotations?



UCSC Genome Browser Table Browser



Select your genome

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

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

group: Repeats add custom tracks track hubs

table: chainSelf describe tabl

region: genome position lookup define regions

identifiers (names/accessions)

filter: create

intersection: create

output format: all fields from selecte Send output to Galaxy GREAT GenomeSpace

output file: keep output in browser)

file type returned: plain text

get output summary/statistics

To reset **all** user cart settings (inc click here.

- Human
- Mouse
- Alpaca
- Armadillo
- Baboon
- Bonobo
- Bushbaby
- Cat
- Chimp
- Chinese hamster
- Cow
- Crab-eating macaque
- Dog
- Dolphin
- Elephant
- Ferret
- Gibbon
- Gorilla
- Green monkey
- Guinea pig
- Hedgehog
- Horse
- Kangaroo rat
- Malayan flying lemur
- Manatee
- Marmoset
- Megabat
- Microbat
- Minke whale
- Mouse lemur
- Naked mole-rat
- Opossum
- Orangutan
- Panda
- Pig
- Pika

Using the Table Browser

This section provides brief line-by e Table Browser controls. For more information on using this program, see the Table Browser User's Guide.

- **clade:** Specifies which clad
- **genome:** Specifies which o
- **assembly:** Specifies which m's genome sequence to use.
- **group:** Selects the type of t n the *track* list. The options correspond to the track groupings shown in the Genome Browser. Select 'All Tracks' for an alphabetical list of all ava ups. Select 'All Tables' to see all tables including those not associated with a track.
- **database:** (with "All Tables" ines which database should be used for options in table menu.
- **track:** Selects the annotatio ith. This list displays all tracks belonging to the group specified in the *group* list. Some tracks are not available when

Select your assembly

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Dec. 2013 (GRCh38/hg38)
group: Repeats **track:** Self Chain add c
table: chainSelf describe table schema
region: genome position chr6:31995680-31996859 lookup defi
identifiers (names/accessions): paste list upload list
filter: create
intersection: create
output format: all fields from selected table Send output to Galaxy GREAT GenomeSpace
output file: (leave blank to keep output in browser)
file type returned: plain text gzip compressed

Dec. 2013 (GRCh38/hg38)
Feb. 2009 (GRCh37/hg19)
Mar. 2006 (NCBI36/hg18)
May 2004 (NCBI35/hg17)
July 2003 (NCBI34/hg16)

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

Using the Table Browser

This section provides brief line-by-line descriptions of the Table Browser controls. For more information on using this program, see the Table Browser User's Guide.

- **clade:** Specifies which clade the organism is in.
- **genome:** Specifies which organism data to use.
- **assembly:** Specifies which version of the organism's genome sequence to use.
- **group:** Selects the type of tracks to be displayed in the *track* list. The options correspond to the track groupings shown in the Genome Browser. Select 'All Tracks' for an alphabetical list of all available tracks in all groups. Select 'All Tables' to see all tables including those not associated with a track.
- **database:** (with "All Tables" group option) Determines which database should be used for options in table menu.
- **track:** Selects the annotation track data to work with. This list displays all tracks belonging to the group specified in the *group* list. Some tracks are not available when

Let’s get RefSeq Genes

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

- Mapping and Sequencing
- Genes and Gene Predictions
- Phenotype and Literature
- mRNA and EST
- Expression
- Regulation
- Comparative Genomics
- Neandertal Assembly and Analysis
- Denisova Assembly and Analysis
- Variation
- Repeats
- ✓ All Tracks
- All Tables

Table

Use thi…ssociated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in…g the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenH… narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL… cal function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse compu…its page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Se…ads page.

clade: …man **assembly:** Feb. 2009 (GRCh37/hg19)

group … **track:** Self Chain add custom tracks track hubs

table: …ma

region: genome ENCODE Pilot regions position chr21:33031597-33041570 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

output format: all fields from selected table Send output to Galaxy GREAT GenomeSpace

output file: (leave blank to keep output in browser)

file type returned: plain text gzip compressed

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

Using the Table Browser

This section provides brief line-by-line descriptions of the Table Browser controls. For more information on using this program, see the Table Browser User's Guide.

- **clade:** Specifies which clade the organism is in.
- **genome:** Specifies which organism data to use.
- **assembly:** Specifies which version of the organism's genome sequence to use.
- **group:** Selects the type of tracks to be displayed in the *track* list. The options correspond to the track groupings shown in the Genome Browser. Select 'All Tracks' for an alphabetical list of all available tracks in all groups. Select 'All Tables' to see all tables including those not associated with a track.
- **database:** (with "All Tables" group option) Determines which database should be used for options in table menu.
- **track:** Selects the annotation track data to work with. This list displays all tracks belonging to the group specified in the *group* list. Some tracks are not available when

Now you can select RefSeq genes

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19)

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

table: knownGene describe

region: genome ENCODE Pilot region -33041570 lookup define regions

identifiers (names/accessions): paste list

filter: create

intersection: create

correlation: create

output format: all fields from selected table Galaxy GREAT GenomeSpace

output file: (leave rowser)

file type returned: plain text gzip com

get output summary/statistics

To reset **all** user cart settings (including custo

- UCSC Genes
- RefSeq Genes
- GENCODE Gene V24lift37
- GENCODE Genes V19
- GENCODE Genes V17
- GENCODE Genes V14
- GENCODE Genes V7
- TransMap UCSC
- TransMap RefGene
- TransMap mRNA
- TransMap ESTs
- AceView Genes
- Augustus
- CCDS
- CRISPR Regions
- CRISPR Targets
- Ensembl Genes
- EvoFold
- Exoniphy
- Geneid Genes
- Genscan Genes
- H-Inv 7.0
- IKMC Genes Mapped
- lincRNA RNA-Seq Reads
- lincRNA Transcripts
- LRG Transcripts
- MGC Genes
- N-SCAN
- Old UCSC Genes
- ORFeome Clones
- Other RefSeq
- Pfam in UCSC Gene
- Retroposed Genes
- SGP Genes

Using the Table Browser

This section provides brief line-by-line descrip r controls. For more information on using this program, see the Table Browser User's Guide.

- **clade:** Specifies which clade the organis
- **genome:** Specifies which organism data
- **assembly:** Specifies which version of th quence to use.
- **group:** Selects the type of tracks to be d The options correspond to the track groupings shown in the Genome Browser. Select 'All Tracks' for an alphabetical list of all available tracks Tables' to see all tables including those not associated with a track.
- **database:** (with "All Tables" group optio base should be used for options in table menu.

We want BED format...

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19)

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

table: refGene describe table schema

region: genome ENCODE Pilot regions position chr21:33031597-33041570 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: create

output format ✓ all fields from selected table
selected fields from primary and related tables
sequence
GTF - gene transfer format
CDS FASTA alignment from multiple alignment
BED - browser extensible data
custom track
hyperlinks to Genome Browser

Send output to Galaxy GREAT GenomeSpace

output file: keep output in browser)

file type retur

get output sum

To reset **all** use cart settings (including custom tracks), click here.

Let’s get a file, “refseq.bed”

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19)

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

table: refGene describe table schema

region: genome ENCODE Pilot regions position chr21:33031597-33041570 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: create

output format: BED - browser extensible data Send output to Galaxy GREAT GenomeSpace

output file: refseq.bed (leave blank to keep output in browser)

file type returned: plain text gzip compressed

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

Get output

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19)

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

table: refGene describe table schema

region: genome ENCODE Pilot regions position chr21:33031597-33041570 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: create

output format: BED - browser extensible data Send output to Galaxy GREAT GenomeSpace

output file: refseq.bed (leave blank to keep output in browser)

file type returned: plain text gzip compressed

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

Get BED, we want the “Whole Gene” in this case

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Output refGene as BED

☐ **Include custom track header:**

name= tb_refGene

description= table browser query on refGene

visibility= pack

url=

Create one BED record per:

- ◉ Whole Gene
- ○ Upstream by 200 bases
- ○ Exons plus 0 bases at each end
- ○ Introns plus 0 bases at each end
- ○ 5' UTR Exons
- ○ Coding Exons
- ○ 3' UTR Exons
- ○ Downstream by 200 bases

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

get BED | cancel

After saving the file wherever you want, run less

```
chr1    66999251        67216822        NM_001308203    0       +       67000041        67208778        0       22      104,123,64,25,57,55,176,25,52,86,93,75,128,127,66,112,156,133,203,65,165,8067,
0,677,92278,99501,106208,109241,109975,137426,138375,139712,143435,146109,155579,156621,160870,185725,195695,200179,205766,207089,207703,209504,
chr1    66999638        67216822        NM_032291       0       +       67000041        67208778        0       25      413,64,25,72,57,55,176,12,12,25,52,86,93,75,501,128,127,60,112,156,133,203,65,16
5,8067, 0,91891,99114,101988,105821,108854,109588,126557,133574,137039,137988,139325,143048,145722,147913,155192,156234,161478,185338,195308,199792,205379,206702,207316,209117,
chr1    16767166        16786584        NM_001145277    0       +       16767256        16785491        0       7       182,101,105,82,109,178,1248,    0,2960,7198,7388,8421,11166,18170,
chr1    16767166        16786584        NM_001145278    0       +       16767256        16785385        0       8       104,101,105,82,109,178,76,1248, 0,2960,7198,7388,8421,11166,15146,18170,
chr1    16767166        16786584        NM_018090       0       +       16767256        16785385        0       8       182,101,105,82,109,178,76,1248, 0,2960,7198,7388,8421,11166,15146,18170,
chr1    33547778        33567493        NR_126031       0       +       33567493        33567493        0       8       177,174,173,172,166,163,113,60, 0,1776,9872,11067,12370,14529,15889,19655,
chr1    48998526        50489626        NM_001323575    0       -       48999844        50489468        0       13      1439,97,163,153,112,115,90,40,217,95,125,123,192,       0,6787,54149,57978,10163
8,120482,130297,334336,512729,712915,1164458,1318541,1490908,
chr1    48998526        50489626        NM_001323574    0       -       48999844        50489468        0       14      1439,27,97,163,153,112,115,90,40,217,95,161,123,192,    0,2035,6787,54149,57978,
101638,120482,130297,334336,512729,712915,1164422,1318541,1490908,
chr1    48998526        50489626        NR_136623       0       -       50489626        50489626        0       13      1439,97,163,153,112,115,90,124,40,217,95,123,192,       0,6787,54149,57978,10163
8,120482,130297,156925,334336,512729,712915,1318541,1490908,
chr1    25071759        25170815        NM_013943       0       +       25072044        25167428        0       6       357,110,126,107,182,3552,       0,52473,68825,81741,94591,95504,
chr1    48998526        50489626        NM_032785       0       -       48999844        50489468        0       14      1439,27,97,163,153,112,115,90,40,217,95,125,123,192,    0,2035,6787,54149,57978,
101638,120482,130297,334336,512729,712915,1164458,1318541,1490908,
chr1    33546713        33586132        NM_001293562    0       +       33547850        33585783        0       11      182,118,177,174,173,135,166,163,113,215,488,    0,278,1065,2841,10937,12169,1343
5,15594,16954,36789,38931,
chr1    48998526        50489626        NM_001323573    0       -       48999844        50489468        0       13      1439,97,163,153,112,115,90,40,217,95,161,123,192,       0,6787,54149,57978,10163
8,120482,130297,334336,512729,712915,1164422,1318541,1490908,
chr1    33546713        33586132        NM_052998       0       +       33547850        33585783        0       12      182,121,212,177,174,173,135,166,163,113,215,488,        0,275,488,1065,2841,1093
7,12169,13435,15594,16954,36789,38931,
chr1    8378144 8404227 NM_001080397    0       +       8378168 8404073 0       9       102,421,93,225,728,154,177,206,421,     0,6221,7213,7733,12124,17352,19731,21408,25662,
chr1    33547778        33567493        NM_001301826    0       +       33547850        33567493        0       8       177,174,173,135,166,163,113,60, 0,1776,9872,11104,12370,14529,15889,19655,
chr1    33547778        33586132        NM_001301825    0       +       33547850        33585783        0       9       177,174,173,135,166,163,173,215,488,    0,1776,9872,11104,12370,14529,15829,3572
4,37866,
chr1    33546729        33586132        NM_001301824    0       +       33557656        33585783        0       8       380,173,135,166,163,113,215,488,        0,10921,12153,13419,15578,16938,36773,38
915,
:
```

Now let's get CpG islands!

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: **assembly:** Feb. 2009 (GRCh37/hg19)

group Mapping and Sequencing / ✓ Genes and Gene Predictions / Phenotype and Literature / mRNA and EST / Expression / Regulation / Comparative Genomics / Neandertal Assembly and Analysis / Denisova Assembly and Analysis / Variation / Repeats / All Tracks / All Tables **track:** RefSeq Genes add custom tracks track hubs

table: describe table schema

regior regions position chr21:33031597-33041570 lookup define regions

identif te list upload list

filter:

inters

correl

outpu ta Send output to Galaxy GREAT GenomeSpace

output (leave blank to keep output in browser)

file type returned: plain text gzip compressed

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

The default selection should be CpG islands

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19)

group: Regulation **track:** CpG Islands add custom tracks track hubs

table: cpgIslandExt describe table schema

region: genome ENCODE Pilot regions position chr21:33031597-33041570 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: create

output format: BED - browser extensible data Send output to Galaxy GREAT GenomeSpace

output file: refseq.bed (leave blank to keep output in browser)

file type returned: plain text gzip compressed

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

Let’s call it cpg.bed

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19)

group: Regulation **track:** CpG Islands add custom tracks track hubs

table: cpgIslandExt describe table schema

region: genome ENCODE Pilot regions position chr21:33031597-33041570 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: create

output format: BED - browser extensible data Send output to Galaxy GREAT GenomeSpace

output file: cpg.bed (leave blank to keep output in browser)

file type returned: plain text gzip compressed

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

And after clicking get output…click get BED

Genomes Genome Browser Tools Mirrors Downloads My Data Help About Us

Output cpgIslandExt as BED

☑ **Include custom track header:**

name= tb_cpgIslandExt

description= table browser query on cpgIslandExt

visibility= pack

url=

Create one BED record per:

- (•) Whole Gene
- ( ) Upstream by 200 bases
- ( ) Downstream by 200 bases

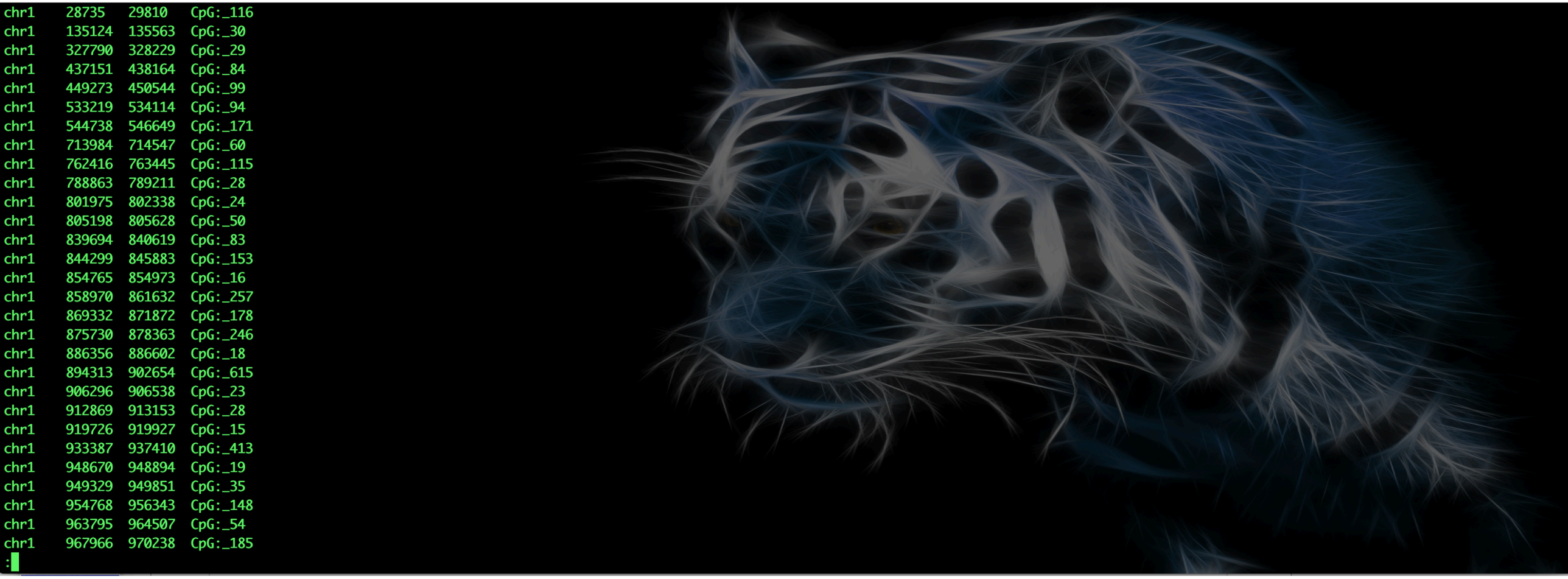
Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

get BED cancel

We want the “Whole Gene”

Output looks like this in less:

```
chr1    28735   29810   CpG:_116
chr1    135124  135563  CpG:_30
chr1    327790  328229  CpG:_29
chr1    437151  438164  CpG:_84
chr1    449273  450544  CpG:_99
chr1    533219  534114  CpG:_94
chr1    544738  546649  CpG:_171
chr1    713984  714547  CpG:_60
chr1    762416  763445  CpG:_115
chr1    788863  789211  CpG:_28
chr1    801975  802338  CpG:_24
chr1    805198  805628  CpG:_50
chr1    839694  840619  CpG:_83
chr1    844299  845883  CpG:_153
chr1    854765  854973  CpG:_16
chr1    858970  861632  CpG:_257
chr1    869332  871872  CpG:_178
chr1    875730  878363  CpG:_246
chr1    886356  886602  CpG:_18
chr1    894313  902654  CpG:_615
chr1    906296  906538  CpG:_23
chr1    912869  913153  CpG:_28
chr1    919726  919927  CpG:_15
chr1    933387  937410  CpG:_413
chr1    948670  948894  CpG:_19
chr1    949329  949851  CpG:_35
chr1    954768  956343  CpG:_148
chr1    963795  964507  CpG:_54
chr1    967966  970238  CpG:_185
:
```



Ensembl Homepage

www.ensembl.org/index.html?redirect=no

Login/Register

e!Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Search all species…

Search: All species for Go

e.g. **BRCA2** or **rat 5:62797383-63627669** or **rs699** or **coronary heart disease**

Browse a Genome

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Favourite genomes

Human
GRCh38.p7

Human
GRCh37

Mouse
GRCm38.p4

Zebrafish
GRCz10

Edit favourites

All genomes

-- Select a species --

View full list of all Ensembl species

Other species are available in Ensembl *Pre!* and EnsemblGenomes

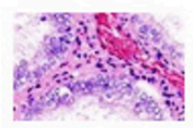
Still using Human GRCh37?

Go to e!GRCh37

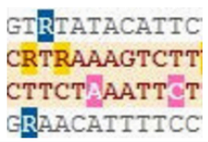
Variant Effect Predictor

VeP

Gene expression in different tissues



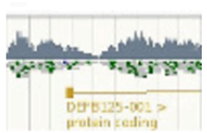
Find SNPs and other variants for my gene



Retrieve gene sequence

Compare genes across species

Use my own data in Ensembl



ENCODE data in Ensembl

What's New in Ensembl Release 86 (October 2016)

- **Mouse Strains**
- **Chicken new assembly and gene set**
- **Macaque new assembly and genebuild**
- **Mouse lemur new assembly and genebuild**
- **Zebrafish: update to Ensembl-Havana merged gene set**

Full details | All web updates, by release | More news on our blog

- 28 Oct 2016: Ensembl genomes 33 is out!
- 17 Oct 2016: Ensembl helpdesk maintenance
- 05 Oct 2016: Ensembl 86 has been released!

Go to Ensembl blog

Select a species

Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Login/Register

Search all species…

Search: All species for Go

e.g. **BRCA2** or **rat 5:62797383-63627669** or **rs699** or **coronary heart disease**

Browse a Genome

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Favourite genomes

Human
GRCh38.p7

Human
GRCh37

Zebrafish
GRCz10

-- Select a species --
Favourites
Human GRCh38
Human GRCh37
Mouse
Zebrafish
Primates
Bushbaby
Chimpanzee
Gibbon
Gorilla
Human
Macaque
Marmoset
Mouse Lemur
Olive baboon
Orangutan
Tarsier

es

nsembl *Pre!* and

Still using Human GRCh37?

Go to GRCh37

Variant Effect Predictor

VeP

Gene expression in different tissues

Find SNPs and other variants for my gene

Retrieve gene sequence

Compare genes across species

Use my own data in Ensembl

ENCODE data in Ensembl

Ensembl supports data from external projects through

What's New in Ensembl Release 86 (October 2016)

- **Mouse Strains**
- **Chicken new assembly and gene set**
- **Macaque new assembly and genebuild**
- **Mouse lemur new assembly and genebuild**
- **Zebrafish: update to Ensembl-Havana merged gene set**

Full details | All web updates, by release | More news on our blog

- 28 Oct 2016: Ensembl genomes 33 is out!
- 17 Oct 2016: Ensembl helpdesk maintenance
- 05 Oct 2016: Ensembl 86 has been released!

Go to Ensembl blog

If necessary, select assembly

Human

Homo sapiens

Search all categories | Search Human... | Go

e.g. **BRCA2** or **17:63973115-64437414** or **rs1333049** or **osteoarthritis**

What's New in Human release 86

- Human: updated cDNA alignments
- Human: updated RefSeq gene import
- External database references update

More news...

Genome assembly: GRCh38.p7 (GCA_000001405.22)

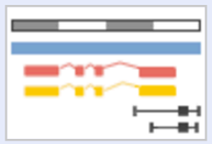
More information and statistics

Download DNA sequence (FASTA)

Convert your data to GRCh38 coordinates

Display your data in Ensembl

View karyotype



Example region

Other assemblies

GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart | Go

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

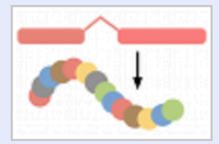
More about this genebuild, including RNASeq gene expression models

Download genes, cDNAs, ncRNA, proteins (FASTA)

Update your old Ensembl IDs

Vega Additional manual annotation can be found in Vega

Example gene



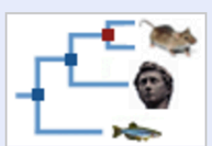
Example transcript

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

More about comparative analysis

Download alignments (EMF)



Example gene tree

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

More about variation in Ensembl

Download all variants (GVF)

Variant Effect Predictor VeP

Example variant

Example phenotype

Regulation

What can I find? DNA methylation, transcription factor binding sites, histone modifications, and regulatory features such as enhancers and repressors, and

Now that you chose your assembly, go to BioMart

Human

Homo sapiens

Search all categories ▼ Search Human... Go

e.g. **BRCA2** or **17:63973115-64437414** or **rs1333049** or **osteoarthritis**

What's New in Human GRCh37

- This GRCh37 archive will be updated every release with new web features, and periodically with new human data.

Genome assembly: GRCh37.p13 (GCA_000001405.14)

More information and statistics

Download DNA sequence (FASTA)

Convert your data to GRCh37.p13 coordinates

Display your data in Ensembl

View karyotype

Example region

Other assemblies

GRCh38 (Ensembl release 86) Go

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

More about this genebuild, including RNASeq gene expression models

Download genes, cDNAs, ncRNA, proteins (FASTA)

Update your old Ensembl IDs

Vega Additional manual annotation can be found in Vega

Example gene

Example transcript

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

More about comparative analysis

Download alignments (EMF)

Example gene tree

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

More about variation in Ensembl

Download all variants (GVF)

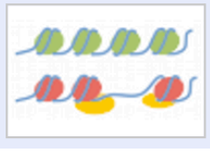
Variant Effect Predictor Ve!P

Example variant

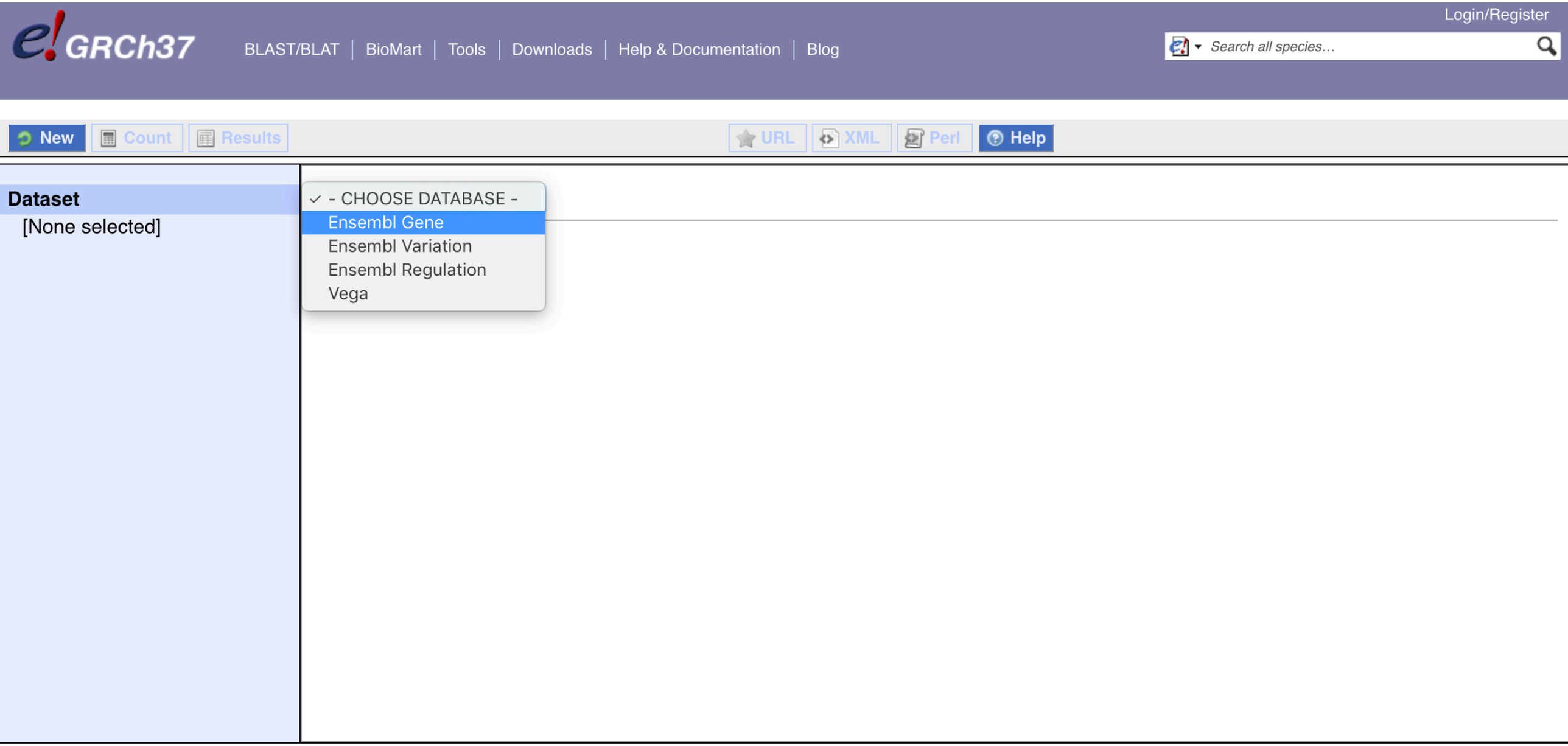
Example phenotype

Regulation

What can I find? DNA methylation, transcription factor binding sites, histone modifications, and regulatory features such as enhancers and repressors, and

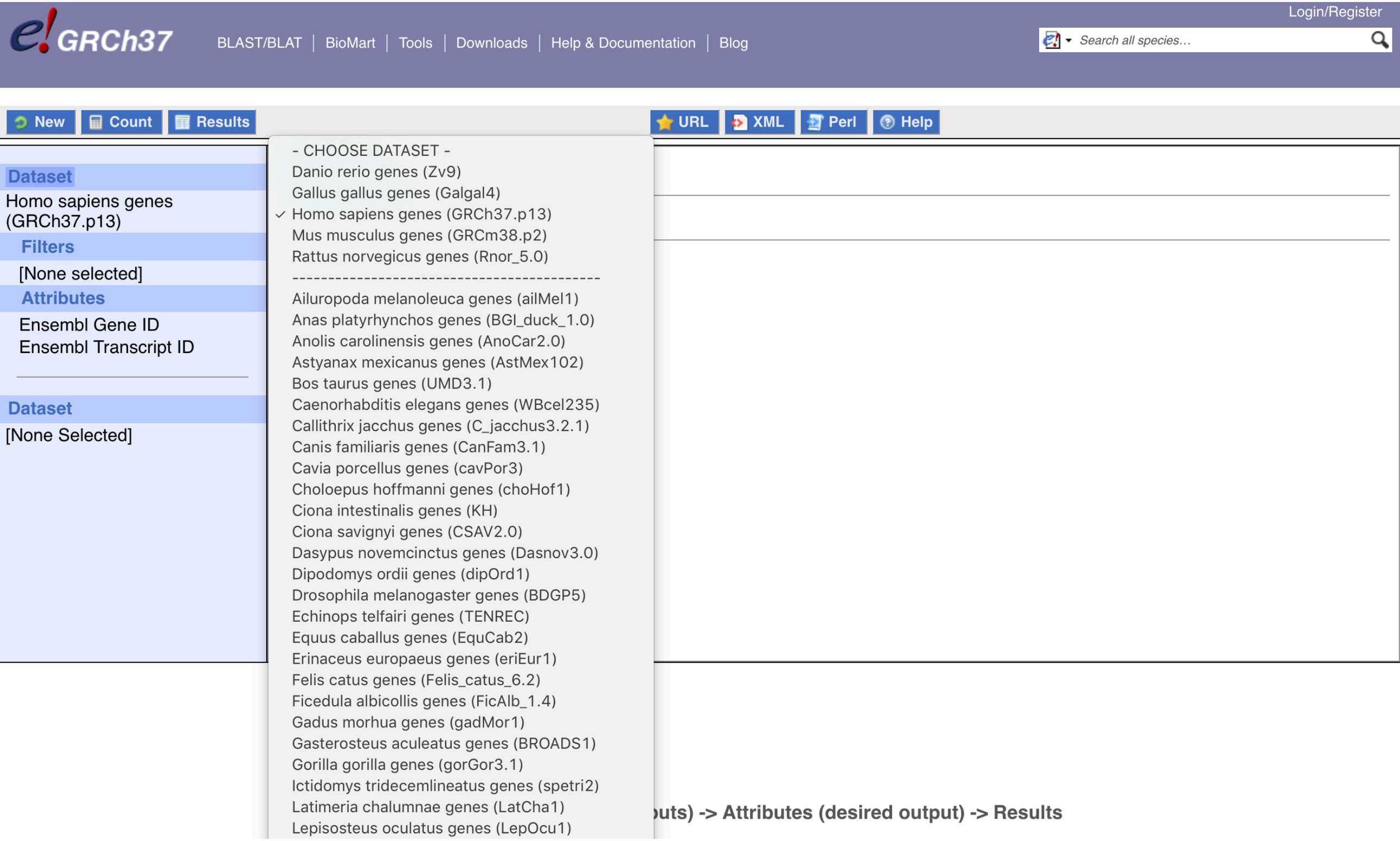


Let’s get some data. Choose Ensembl Gene

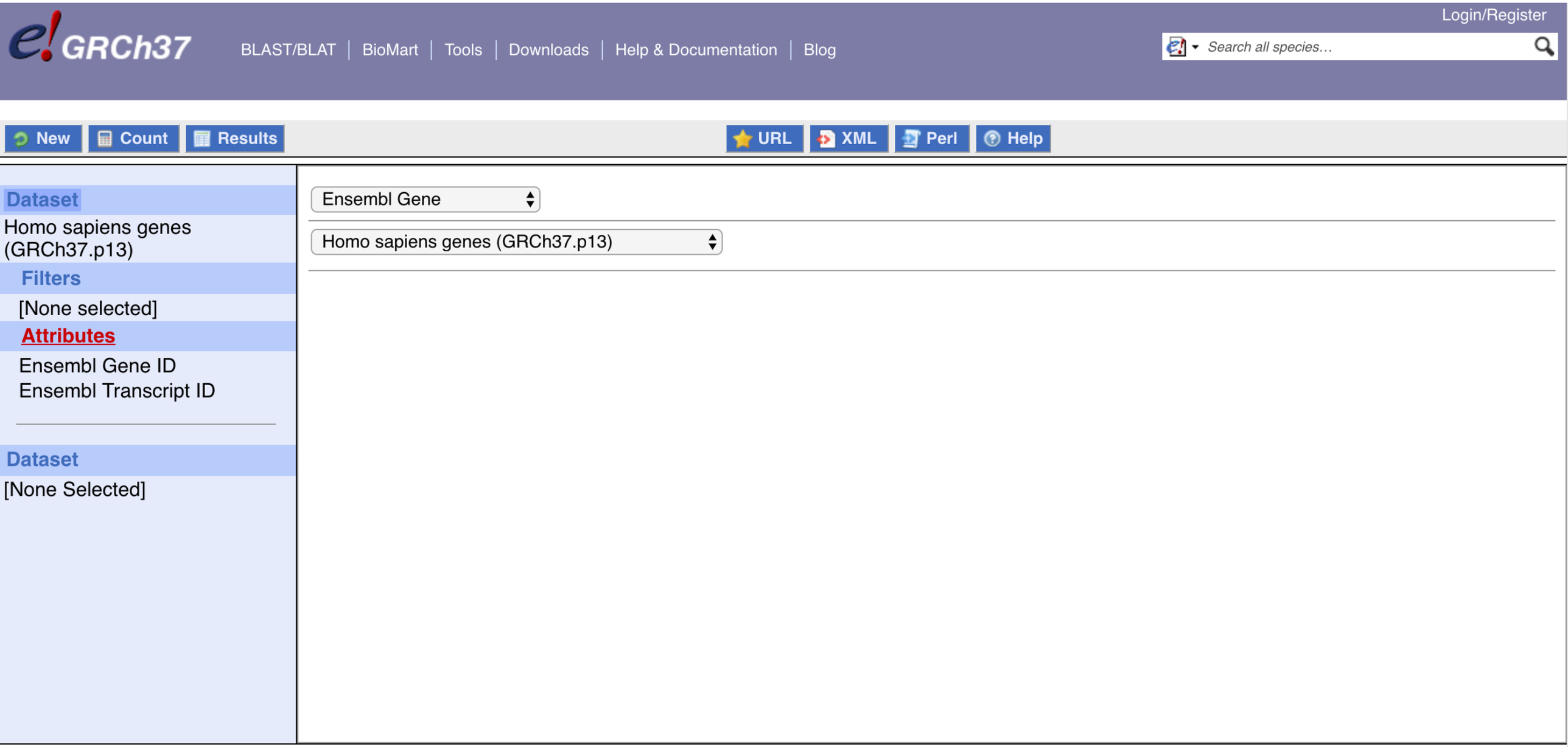


Datasets -> Filters (filtering and inputs) -> Attributes (desired output) -> Results

Pick assembly again (you have access to GRCh37)

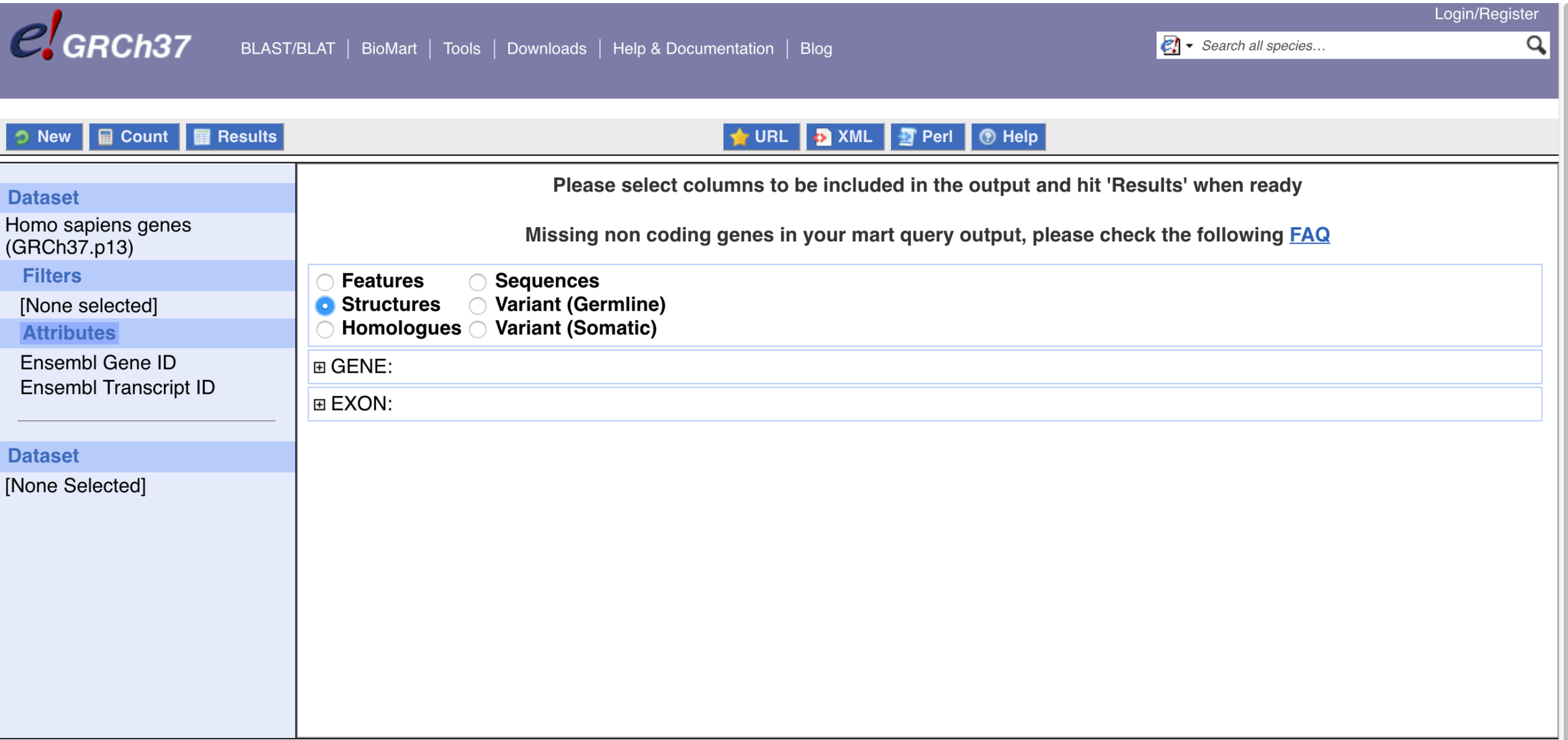


Select Attributes



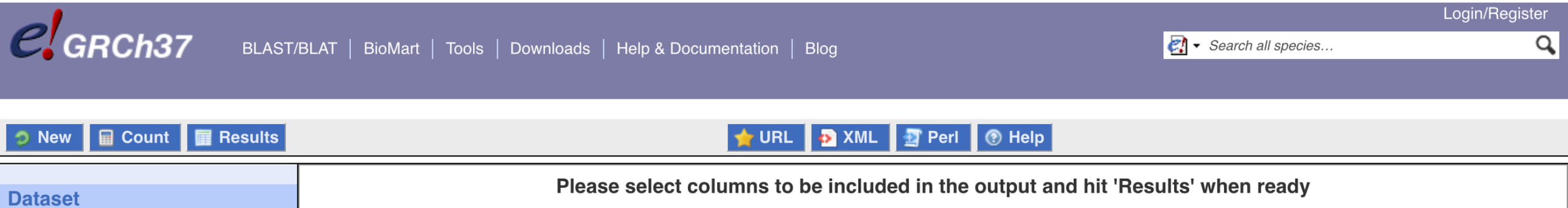
Datasets -> Filters (filtering and inputs) -> Attributes (desired output) -> Results

Select Structures



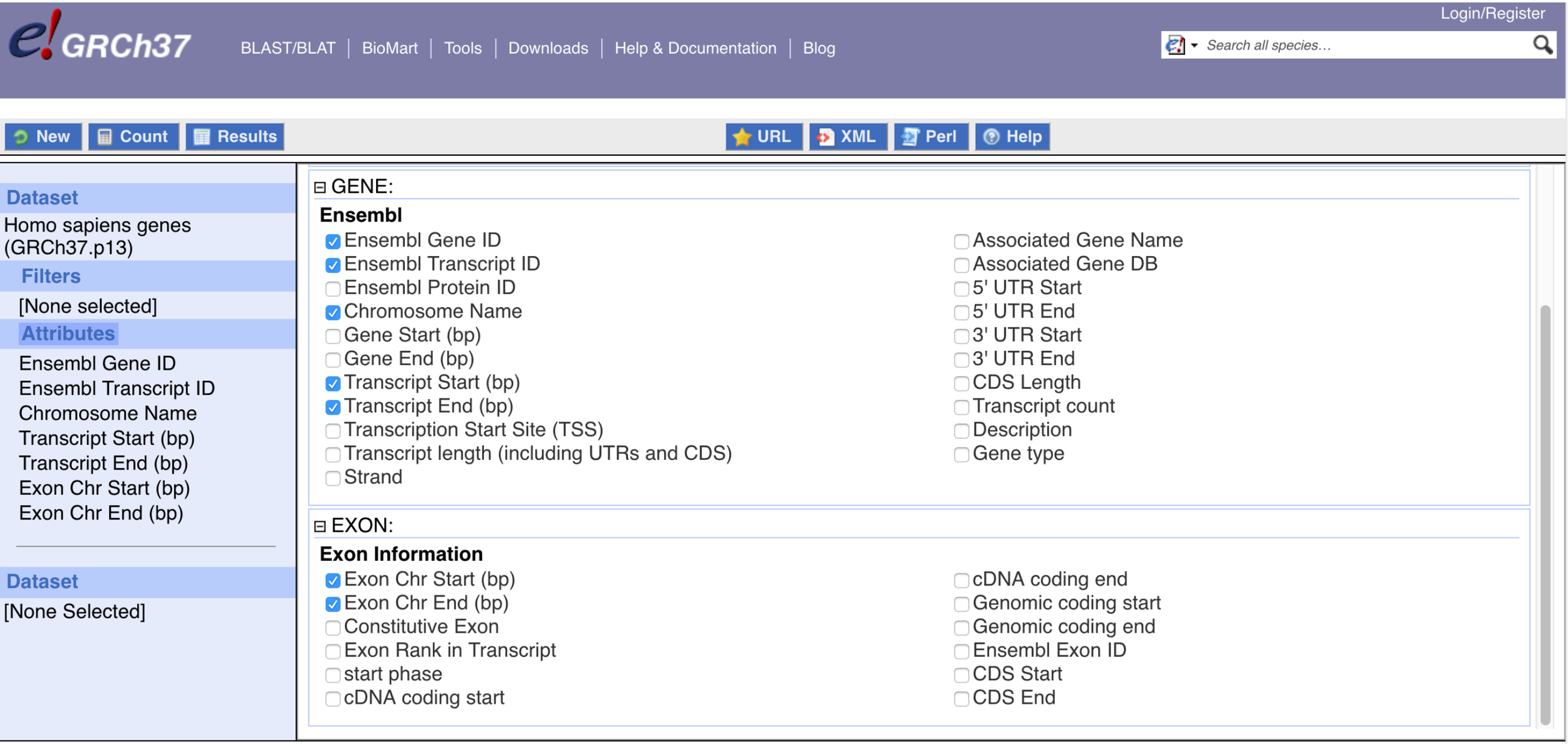
Datasets -> Filters (filtering and inputs) -> Attributes (desired output) -> Results

Let’s get starts and ends for transcripts, and chrom



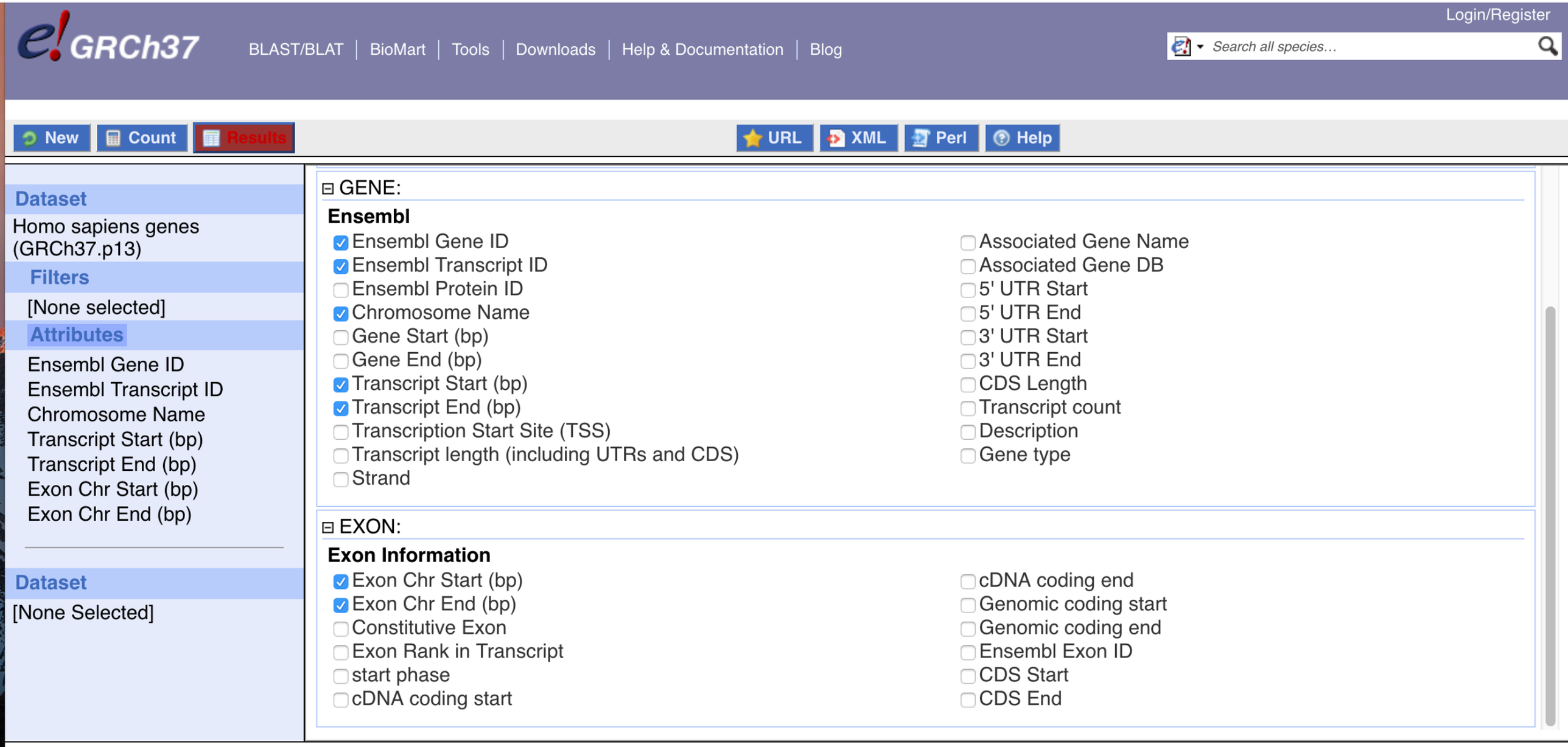
Datasets -> Filters (filtering and inputs) -> Attributes (desired output) -> Results

And starts and ends for exons



Datasets -> Filters (filtering and inputs) -> Attributes (desired output) -> Results

And lastly let's display the results



Datasets -> Filters (filtering and inputs) -> Attributes (desired output) -> Results

Output looks like this

e! GRCh37 BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog

Login/Register

Search all species…

New | Count | Results | URL | XML | Perl | Help

Dataset

Homo sapiens genes (GRCh37.p13)

Filters

[None selected]

Attributes

Ensembl Gene ID
Ensembl Transcript ID
Chromosome Name
Transcript Start (bp)
Transcript End (bp)
Exon Chr Start (bp)
Exon Chr End (bp)

Dataset

[None Selected]

Export all results to File TSV Unique results only Go

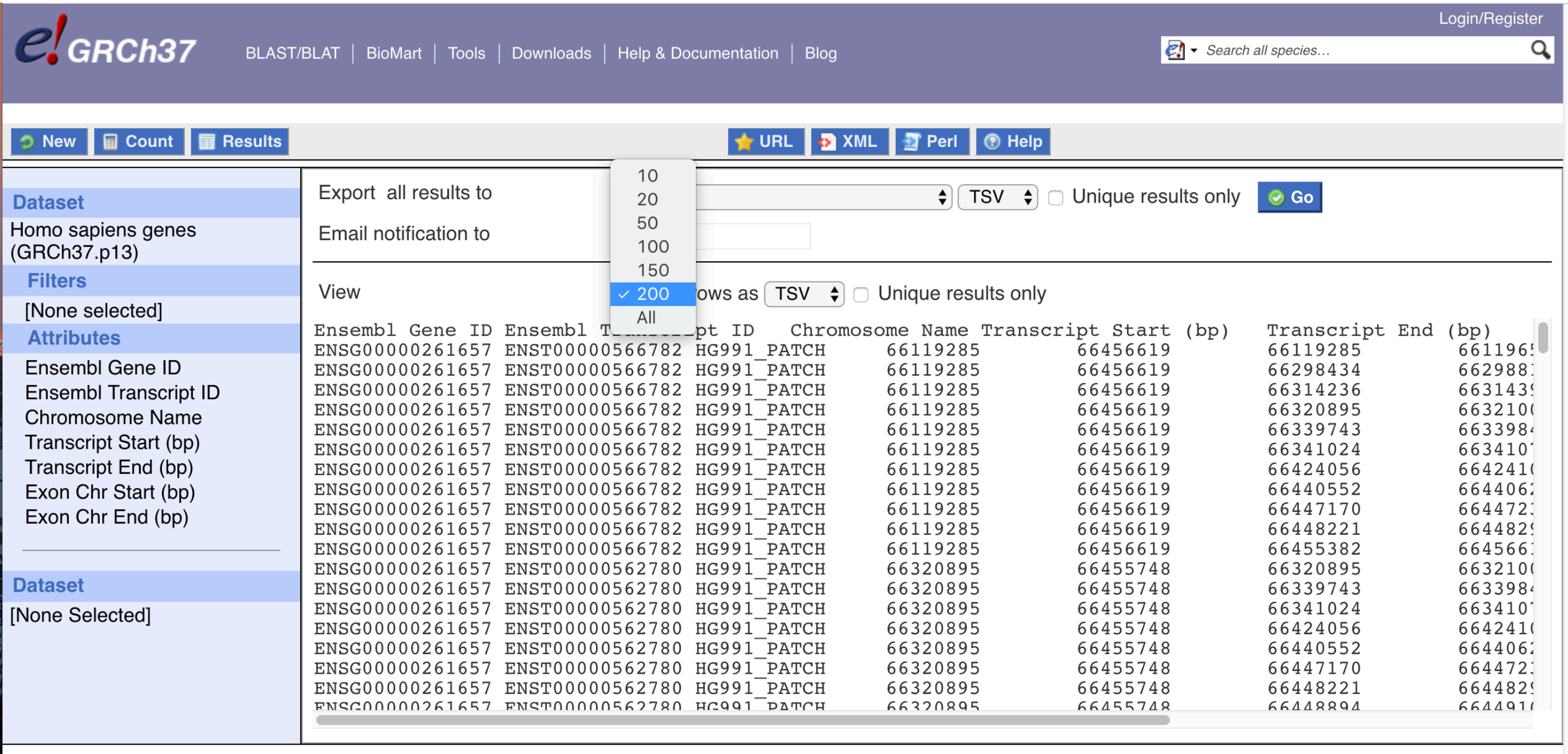
Email notification to

View 100 rows as HTML Unique results only

Ensembl Gene ID	Ensembl Transcript ID	Chromosome Name	Transcript Start (bp)	Transcript End (bp)	Exon Chr Start (bp)	Exon Chr End (bp)
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66119285	66119659
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66298434	66298819
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66314236	66314392
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66320895	66321004
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66339743	66339847
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66341024	66341071
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66424056	66424100
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66440552	66440621
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66447170	66447234
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66448221	66448294
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66455382	66456619
ENSG00000261657	ENST00000562780	HG991_PATCH	66320895	66455748	66320895	66321004
ENSG00000261657	ENST00000562780	HG991_PATCH	66320895	66455748	66339743	66339847
ENSG00000261657	ENST00000562780	HG991_PATCH	66320895	66455748	66341024	66341071
ENSG00000261657	ENST00000562780	HG991_PATCH	66320895	66455748	66424056	66424100
ENSG00000261657	ENST00000562780	HG991_PATCH	66320895	66455748	66440552	66440621
ENSG00000261657	ENST00000562780	HG991_PATCH	66320895	66455748	66447170	66447234

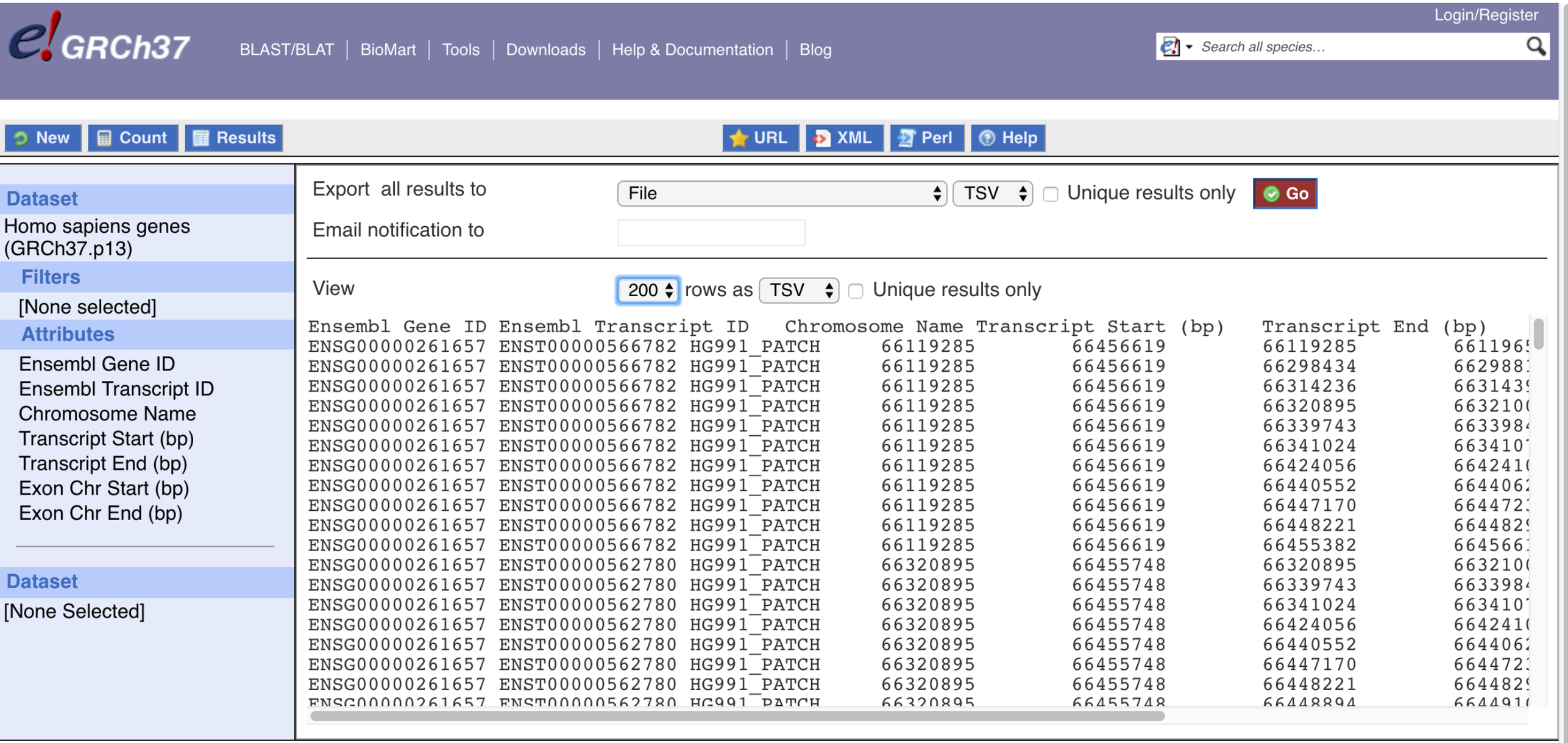
Datasets -> Filters (filtering and inputs) -> Attributes (desired output) -> Results

Let's just download the first 200 transcripts



Datasets -> Filters (filtering and inputs) -> Attributes (desired output) -> Results

Click “Go”



Datasets -> Filters (filtering and inputs) -> Attributes (desired output) -> Results

Save as “bedtest” to whatever folder

Ensembl Gene ID	Ensembl Transcript ID	Chromosome Name	Transcript Start (bp)	Transcript End (bp)	Exon Chr Start (bp)	Exon Chr End (bp)
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66119285	66119659
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66298434	66298819
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66314236	66314392
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66320895	66321004
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66339743	66339847
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66341024	66341071
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66424056	66424100
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66440552	66440621
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66447170	66447234
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66448221	66448294
ENSG00000261657	ENST00000566782	HG991_PATCH	66119285	66456619	66455382	66456619
ENSG00000261657	ENST00000562780	HG991_PATCH	66320895	66455748	66320895	66321004
ENSG00000261657	ENST00000562780	HG991_PATCH	66320895	66455748	66339743	66339847
ENSG00000261657	ENST00000562780	HG991_PATCH	66320895	66455748	66341024	66341071
ENSG00000261657	ENST00000562780	HG991_PATCH	66320895	66455748	66424056	66424100
ENSG00000261657	ENST00000562780	HG991_PATCH	66320895	66455748	66440552	66440621
ENSG00000261657	ENST00000562780	HG991_PATCH	66320895	66455748	66447170	66447234
ENSG00000261657	ENST00000562780	HG991_PATCH	66320895	66455748	66448221	66448294
ENSG00000261657	ENST00000562780	HG991_PATCH	66320895	66455748	66448894	66449105
ENSG00000261657	ENST00000562780	HG991_PATCH	66320895	66455748	66455382	66455748
ENSG00000261657	ENST00000569579	HG991_PATCH	66320895	66456619	66320895	66321004
ENSG00000261657	ENST00000569579	HG991_PATCH	66320895	66456619	66339743	66339847
ENSG00000261657	ENST00000569579	HG991_PATCH	66320895	66456619	66341024	66341071
ENSG00000261657	ENST00000569579	HG991_PATCH	66320895	66456619	66382145	66382229
ENSG00000261657	ENST00000569579	HG991_PATCH	66320895	66456619	66424056	66424100
ENSG00000261657	ENST00000569579	HG991_PATCH	66320895	66456619	66440552	66440621
ENSG00000261657	ENST00000569579	HG991_PATCH	66320895	66456619	66447170	66447234
ENSG00000261657	ENST00000569579	HG991_PATCH	66320895	66456619	66448221	66448294
ENSG00000261657	ENST00000569579	HG991_PATCH	66320895	66456619	66455382	66456619
ENSG00000261657	ENST00000568242	HG991_PATCH	66320895	66465398	66320895	66321004
ENSG00000261657	ENST00000568242	HG991_PATCH	66320895	66465398	66339743	66339847
ENSG00000261657	ENST00000568242	HG991_PATCH	66320895	66465398	66341024	66341071
ENSG00000261657	ENST00000568242	HG991_PATCH	66320895	66465398	66424056	66424100
ENSG00000261657	ENST00000568242	HG991_PATCH	66320895	66465398	66433723	66433776
ENSG00000261657	ENST00000568242	HG991_PATCH	66320895	66465398	66440552	66440621
ENSG00000261657	ENST00000568242	HG991_PATCH	66320895	66465398	66447170	66447234
ENSG00000261657	ENST00000568242	HG991_PATCH	66320895	66465398	66448221	66448294
ENSG00000261657	ENST00000568242	HG991_PATCH	66320895	66465398	66455382	66455762
ENSG00000261657	ENST00000568242	HG991_PATCH	66320895	66465398	66456102	66456194
ENSG00000261657	ENST00000568242	HG991_PATCH	66320895	66465398	66457809	66458645
ENSG00000261657	ENST00000568242	HG991_PATCH	66320895	66465398	66463876	66465398
ENSG00000261657	ENST00000565530	HG991_PATCH	66339287	66448276	66339287	66339342
ENSG00000261657	ENST00000565530	HG991_PATCH	66339287	66448276	66339743	66339847
ENSG00000261657	ENST00000565530	HG991_PATCH	66339287	66448276	66341024	66341071
ENSG00000261657	ENST00000565530	HG991_PATCH	66339287	66448276	66424056	66424100
ENSG00000261657	ENST00000565530	HG991_PATCH	66339287	66448276	66447170	66447234
ENSG00000261657	ENST00000565530	HG991_PATCH	66339287	66448276	66448221	66448276
ENSG00000223116	ENST00000411184	13	23551994	23552136	23551994	23552136
ENSG00000233440	ENST00000418454	13	23708313	23708703	23708313	23708703

Data looks like this

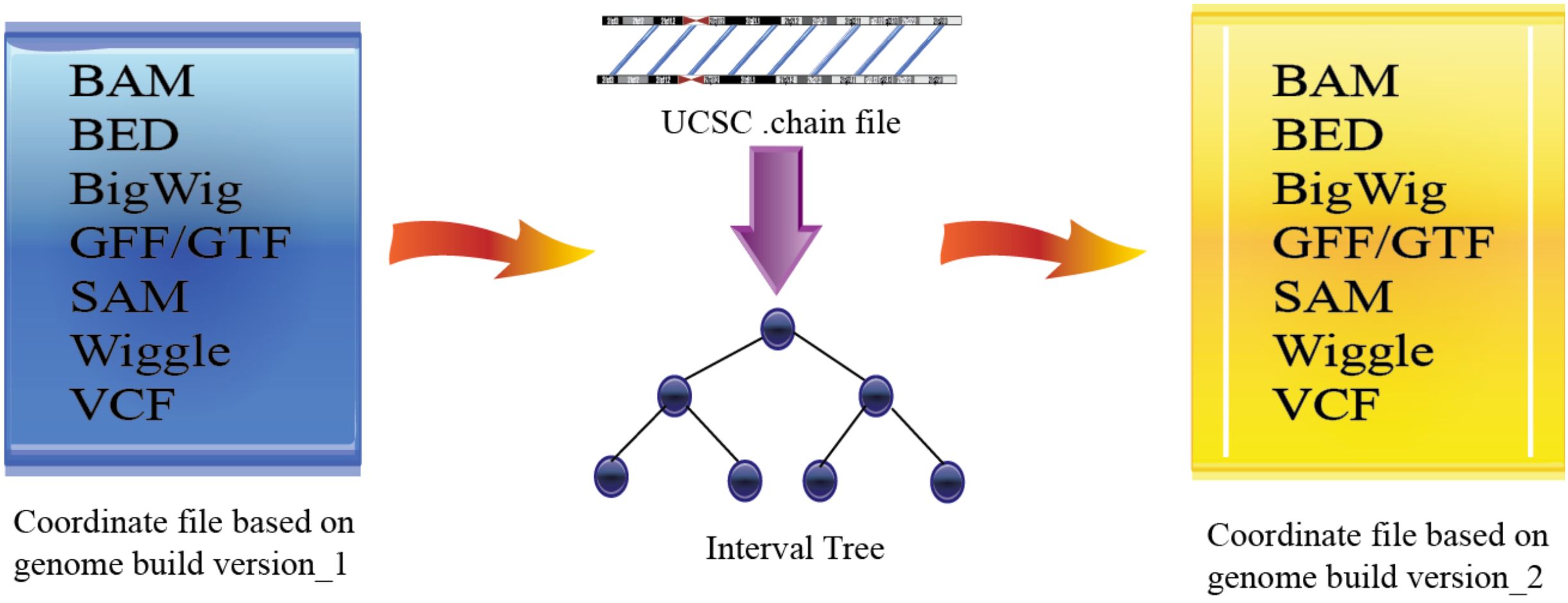
Let’s turn this into BED format shall we?

- After saving “bedtest.txt”, run:
  - awk '{print $3,$6,$7,$4,$5,$1,$2}' FS='\t' OFS='\t' bedtest.txt | sed '1d' > transcripts.bed
- And now it’s in BED format! Check with less transcripts.bed

What happens when a new genome assembly is released?

“Lifting over”

GRCh37 → GRCh38



Coordinate file based on genome build version_1

Coordinate file based on genome build version_2

Tabix

- Tabix is packaged with samtools, and bgzip
- Can create indices for any standard genome format:
  - BAM, GFF, SAM, BED, VCF
  - Compress files with bgzip first to save space
- Allows for rapid querying of a genome position file
- Makes for quick searching in IGV, less data needs to be loaded into RAM

Tabix

- Example with the refseq.bed file we used earlier.
  - First, make sure the file is sorted, then compress it:
    - sort -k1,1 -k2,2n refseq.bed | bgzip -c > refseq.bed.gz
  - Then index it with tabix, this makes a .tbi file:
    - tabix refseq.bed.gz
  - Then try to query it (note that I used “chr”, because we used it):
    - tabix refseq.bed.gz chr1:1-100000

```
[semja:Documents] $ sort -k1,1 -k2,2n refseq.bed | bgzip -c > refseq.bed.gz
[semja:Documents] $ tabix refseq.bed.gz
[semja:Documents] $ tabix refseq.bed.gz chr1:1-100000
chr1    11873   14409   NR_046018       0       +       14409   14409   0       3       354,109,1189,   0,739,1347,
chr1    14361   29370   NR_024540       0       -       29370   29370   0       11      468,69,152,159,198,136,137,147,99,154,50,       0,608,1434,2245,2496,2871,3244,3553,3906,10376,14
chr1    17368   17436   NR_106918       0       -       17436   17436   0       1       68,     0,
chr1    17368   17436   NR_107062       0       -       17436   17436   0       1       68,     0,
chr1    17368   17436   NR_107063       0       -       17436   17436   0       1       68,     0,
chr1    17368   17436   NR_128720       0       -       17436   17436   0       1       68,     0,
chr1    30365   30503   NR_036051       0       +       30503   30503   0       1       138,    0,
chr1    30365   30503   NR_036266       0       +       30503   30503   0       1       138,    0,
chr1    30365   30503   NR_036267       0       +       30503   30503   0       1       138,    0,
chr1    30365   30503   NR_036268       0       +       30503   30503   0       1       138,    0,
chr1    34610   36081   NR_026818       0       -       36081   36081   0       3       564,205,361,    0,666,1110,
chr1    34610   36081   NR_026820       0       -       36081   36081   0       3       564,205,361,    0,666,1110,
chr1    69090   70008   NM_001005484    0       +       69090   70008   0       1       918,    0,
```

Use of tabix in IGV

- First of all, IGV will ask to index a file that you have not indexed yet (it will use igvtools to do so, but it is not as efficient as tabix)
- If you passed a tabixed file into IGV, it will be able to access parts of the file much faster as IGV can read a .tbi index