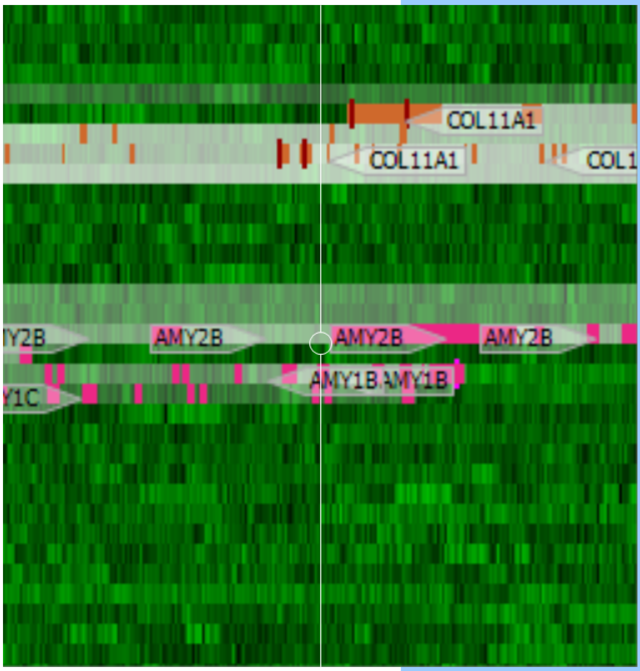
Visual Genome Browser (Beta version)

This document outlines the Use Cases which led to the development of the offline genome browser software.



Heinrich Ferreira

29 November 2016

heinrichjf@gmail.com

Table of Contents

Genome Browser Concept

Introduction:

I came up with a concept of visualizing the human genome (or the genome of any other organism for that matter) in a 2 dimensional space while playing around with the file formats of genome data provided for free online by the **University of Santa Cruz.** (https://genome.ucsc.edu/cgi-bin/hgGateway).

Being a programmer I wanted to investigate the Human Genome for myself. I wanted to get a feel for the nature of the digital data contained within it. I learned that it contained a code for producing protein machines such as enzymes as well as structural and control proteins such as transcription factors, enhancers and suppressors. I was also intrigued by the repetitive patterns in the genome in regions such as the centromeres and telomeres of chromosomes. The standard way of representing genes on a linear axis simply did not elucidate the structure sufficiently for me. I learnt that the genome had a complex 3D organization in the nucleus of the cell and thought that that there had to be a way to get the a better picture of the structure of the data of the genome.

Coming from the field of GIS (Graphical Information Systems) I knew that one could better represent big amounts of map data in 2 dimensions with multiple layers of information overlaid on top of each other. I wanted to do the same for the genome. ***In no way do I claim that I am the first to think of doing this, there might be other software who have attempted to do just this, but I wanted to get an intuitive "feel" for the data***. I decided to find a way of representing the chromosomes in as compact a way as possible by depicting the linear DNA letters by coloured blocks running from left to right, top to bottom akin to the scan lines of a television screen. If there were any repetitive patterns in the data it should become visible when you use the correct line width. Any repeat occurring at an interval of x bases, will show up when you line up the sequences at a width of x bases.

I needed a metric to us that would sufficiently change throughout different parts of the genome and I decided upon the GC content. I needed to map that metric onto an RGB (Red Green Blue) colour scale and picked the scale of Green colour as it reminded me of GFP (Green Fluorescent Protein) which is often used in fluorescent reporter assays. I found that if I simply mapped nucleotides to pixels, the result simply looked like noise, so I had to apply averaging over bases to get more uniform and visually identifiable "blocks" of a certain colour intensity. Taking the average number G'c and C's in blocks of about 20-50 bases and then mapping it to the 0-255 green intensity provided me with a structure of the genome. Just like origins of replication often occur in areas of higher A's and T's, where the DNA helix is more easily opened up by a helicase for DNA replication to start, I could now distinguish areas of the genome exhibiting distinct GC Content structure.

The best way to explain this is probably to go through a few examples showing how these features are highlighted by visual inspection.

For most of the examples I have obtained data from the UCSC Genome Browser.

https://genome.ucsc.edu/cgi-bin/hgGateway

HG38 **.2bit** download folder: ftp://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/

Some examples of distinct GC structure in the genome.

Here is an extract from the Human X-chromosome.

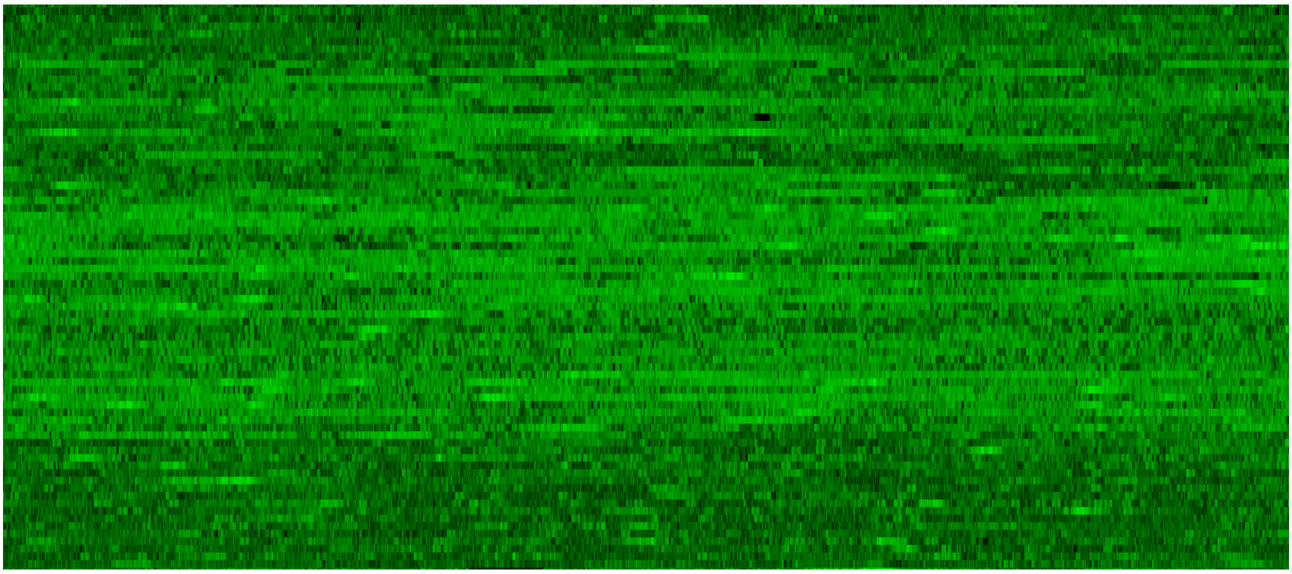


Illustration 1: ChrX:152,300,809-153,887,660 at 39672 bases per horizontal line and 29 bases per horizontal pixel. Green intensity represents average GC content in each pixel.

The data has a non-repetitive (almost random) structure. When I map the genes that I downloaded via MySQL from the table data of the UCSC, I find that there is a high concentration of genes in this region.

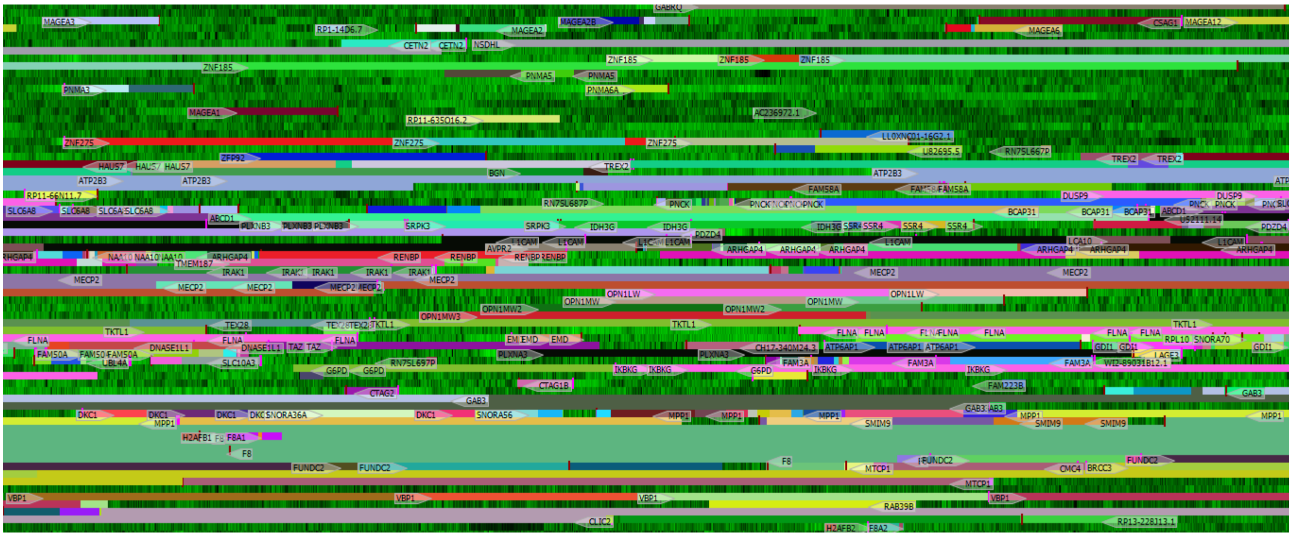


Illustration 2: ChrX:152,300,809-153,887,660 at 39672 bases per horizontal line and UCSC genes drawn as blocks of different colours

I mapped the gene annotations as an overlay on top of the GC Content with small name arrows indicating the direction in which the genes are read during transcription by the RNA Polymerase. In this case I have not separated the exons of the genes, therefore the coloured regions represents the primary transcripts before it is spliced to remove the introns.

A similar **CpG Island output** can be obtained by redrawing the genome (with the “Recreate” option and selecting “CpG Islands” under the “Controls” tab). In the next picture it can be seen that the promoters of genes are mostly found near CpG Islands in the chromosome.

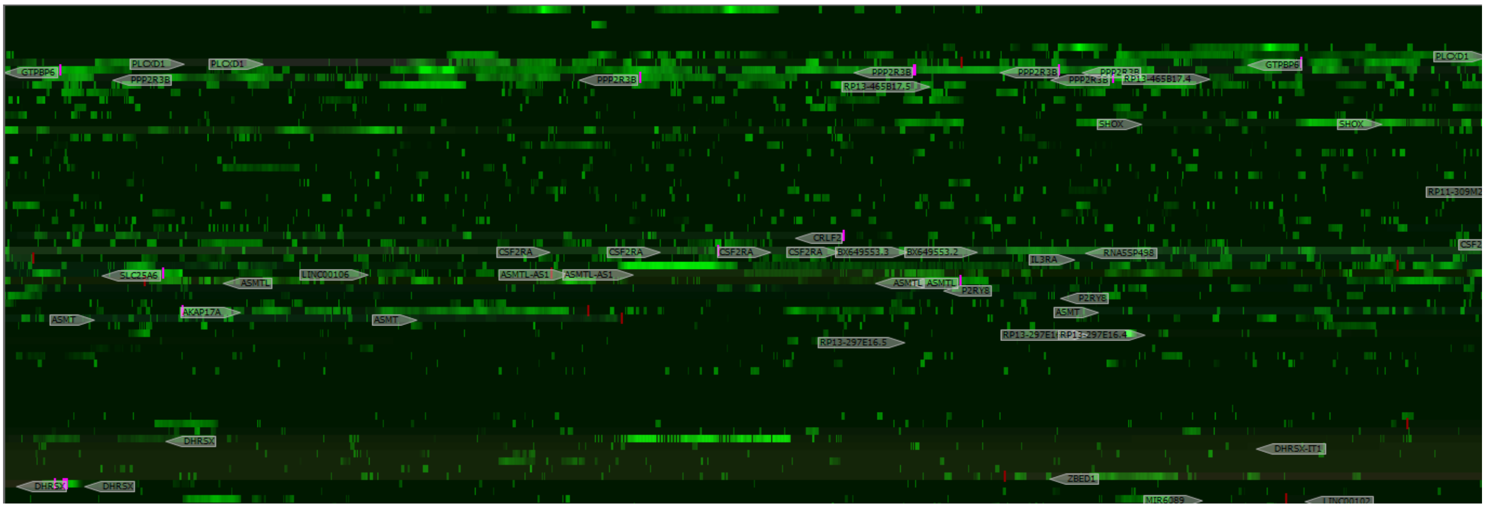


Illustration 3: ChrX:152,300,809-153,887,660 at 39672 bases per horizontal line showing CpG Islands with gene annotations only indicated with arrows

Gene annotations data are downloaded by the software using a MySql query from the UCSC table data (`genome-mysql.cse.ucsc.edu`) or obtained directly from GFF3, Genbank, Ensemble, BED, BigBed, or BigWig files. Annotations can also be obtained from the UCSC search results copied from the web.

When you compare the image above with another part of the genome closer to the centromeres, where highly repetitive patterns are seen, the region is devoid of annotated genes.

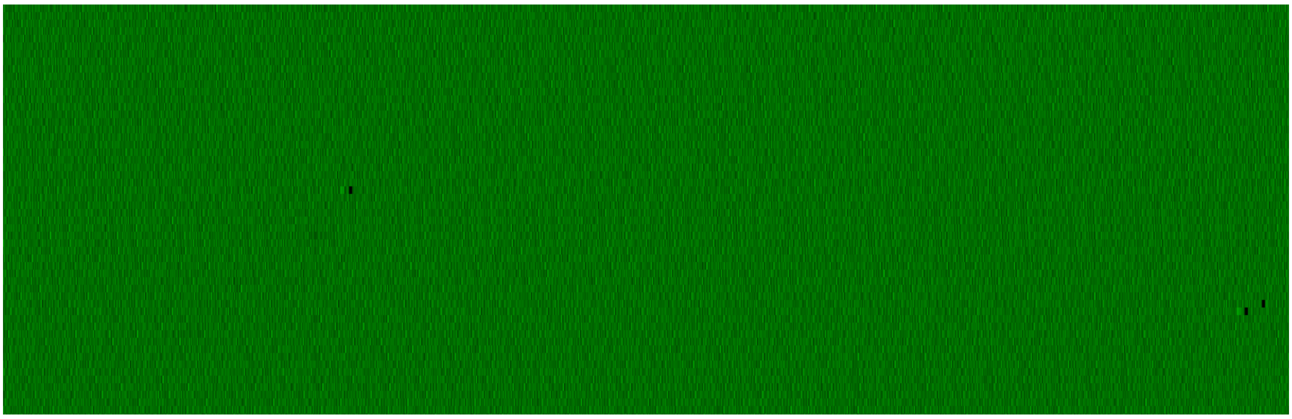


Illustration 4: Chr X:59587374 Inside the centromere with highly repetitive alpha repeats

The picture shows 39672 bases for each horizontal line, and each horizontal pixel represents 29 nucleotides. This means one can get maximum 29 intensities of green. (This value is constantly adjusted per chromosome in order to optimally fit in the information into the fixed pixel width of the display.) All chromosomes are also limited to 32000 vertical display lines (with about 4 vertical pixels per coloured block).

In order to have a closer look at the finer structure of the bases, the Genome Browser provides 2 detail view options for the **DNA View**: (The block at the centre vertically and horizontally always represents the highlighted base position)

- **The default DNA View depicts a continuous extract of bases from the genome at the position of the mouse in the Main Genome View. The genome bases flows from left to right, top to bottom.**
- **An "Aligned" view (representing a further zoom level above the Zoom Gene View) where a cut-out is made for bases spaced every 39672 bases. This allows you to alter the line width using the < and > keys in order to investigate repetitive patterns.**

Each base of the DNA sequence in this region is depicted by a colour:
T = Black
A = Dark Blue
G = Yellow
C = Orange

I deliberately chose dark colours for paired bases with only 2 hydrogen bonds (representing regions which are more easily melted apart) and bright colours for paired bases with 3 hydrogen bonds (representing regions more difficult to separate).

Just using visual observations, one can start to identify the structure of the chromosomes. The structure of the chromosomes has bearing on how densely packed the chromosomes are. Densely packed chromatin / constitutive heterochromatin as is found in the centromeres also tend to stain darker.

When I draw a closer, zoomed in version of this pattern at a width of 171 bases per line, one gets the following close up view (which is only hinted at when you see it in the first image, but upon “closer inspection” becomes clear). One can see a vertical band of blue and black bases, which is probably the bases in the linker section between the nucleosomes, which is more exposed to mutational deamination of C's to U's.

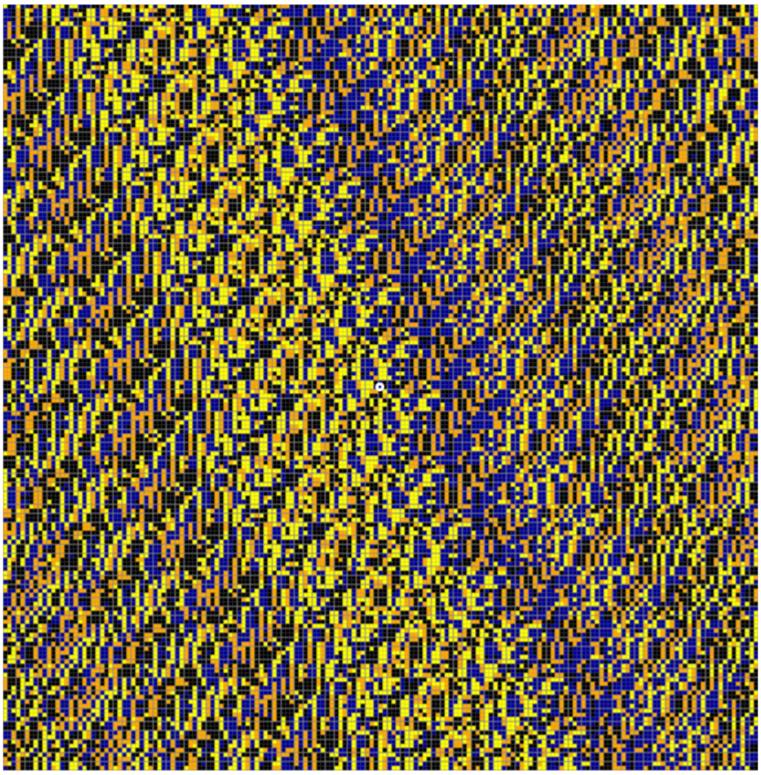
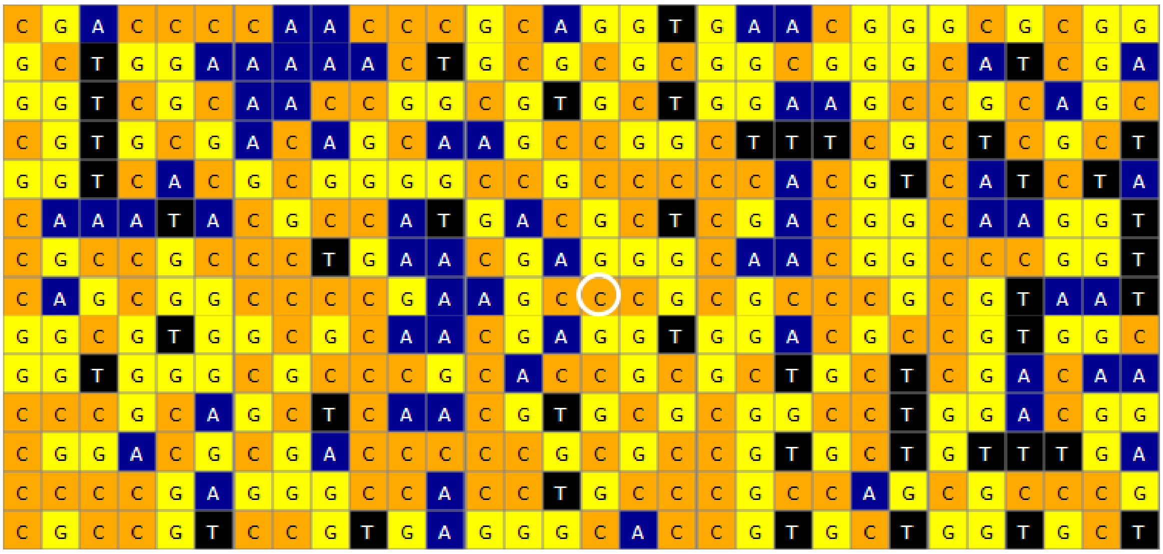


Illustration 5: ChrX:59801335-59830917 at width 171 bases per line (alpha satellite repeats in the heterochromnatin of the centromeric region)

Colouring of the bases actually has a bearing on the temperature at which the DNA could be melted apart. I would therefore expect that organisms which live at very high temperatures would have a genome which consisted mostly of brightly coloured bases (such as the Termus aquaticus thermophilic bacteria living in the hot springs of Yellowstone National Park (from which we obtain the TAQ Polymerase used in PCR (Polymerase Chain Reaction) assays) and Deinococcus radiodurans (an extremophilic bacterium highly resistant against DNA damage).

Here is an extract from the genome of D. radiodurans.



C	G	A	C	C	C	C	A	A	C	C	C	G	C	A	G	G	T	G	A	A	C	G	G	G	C	G	C	G	G
G	C	T	G	G	A	A	A	A	A	C	T	G	C	G	C	G	C	G	G	C	G	G	G	C	A	T	C	G	A
G	G	T	C	G	C	A	A	C	C	G	G	C	G	T	G	C	T	G	G	A	A	G	C	C	G	C	A	G	C
C	G	T	G	C	G	A	C	A	G	C	A	A	G	C	C	G	G	C	T	T	T	C	G	C	T	C	G	C	T
G	G	T	C	A	C	G	C	G	G	G	G	C	C	G	C	C	C	C	C	A	C	G	T	C	A	T	C	T	A
C	A	A	A	T	A	C	G	C	C	A	T	G	A	C	G	C	T	C	G	A	C	G	G	C	A	A	G	G	T
C	G	C	C	G	C	C	C	T	G	A	A	C	G	A	G	G	G	C	A	A	C	G	G	C	C	C	G	G	T
C	A	G	C	G	G	C	C	C	C	G	A	A	G	C	C	C	G	C	G	C	C	C	G	C	G	T	A	A	T
G	G	C	G	T	G	G	C	G	C	A	A	C	G	A	G	G	T	G	G	A	C	G	C	C	G	T	G	G	C
G	G	T	G	G	G	C	G	C	C	C	G	C	A	C	C	G	C	G	C	T	G	C	T	C	G	A	C	A	A
C	C	C	G	C	A	G	C	T	C	A	A	C	G	T	G	C	G	C	G	G	C	C	T	G	G	A	C	G	G
C	G	G	A	C	G	C	G	A	C	C	C	C	C	G	C	G	C	C	G	T	G	C	T	G	T	T	T	G	A
C	C	C	C	G	A	G	G	G	C	C	A	C	C	T	G	C	C	C	G	C	C	A	G	C	G	C	C	C	G
C	G	C	C	G	T	C	C	G	T	G	A	G	G	G	C	A	C	C	G	T	G	C	T	G	G	T	G	C	T

Illustration 6: Extract from Deinococcus radiodurans genome

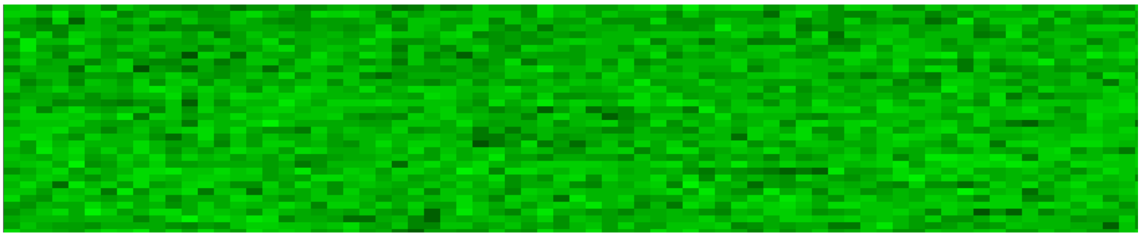


Illustration 7: GC Content of 50400 bases averaged at 20 bases per block of D. radiodurans

Looking at the GC Content of a larger region of the genome of this bacteria (50400 out of a total of 2,648,638 of its circular positively supercoiled chromosome), where the green light intensity is equated to the average GC content, reveals a genome with high GC content.

This is clearly an organism suited to replicate its DNA at higher temperatures, and the positively supercoiled DNA helps prevent its DNA from unwinding easily.

When we look at a similar size length of the thermophilic bacteria Termus aquaticus, we get the same result:

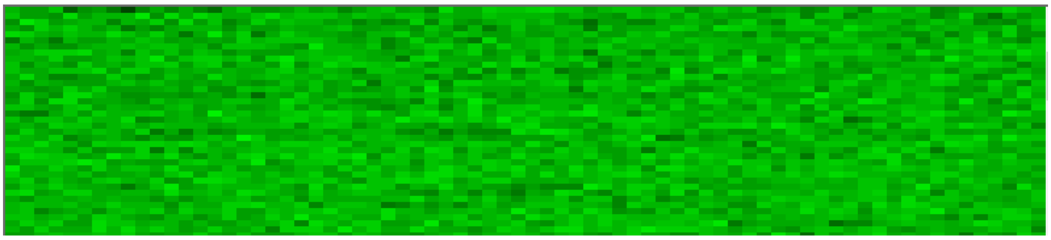


Illustration 8: GC Content of 56,160 bases averaged at 20 bases per block of T. aquaticus

Doing the same for other bacteria living at much colder temperatures, such as Yersinia enterocolitica, which can live at temperatures of less than 0 degrees, I find a genome with a much lower GC content and a nucleotide bias towards more A's and T's. Its DNA needs to be replicated at lower temperatures, with less energy available to helicases to pull apart the DNA strands. Because there are less energy available at these low temperatures, it would not survive if the DNA strands required a lot of energy to be separated.

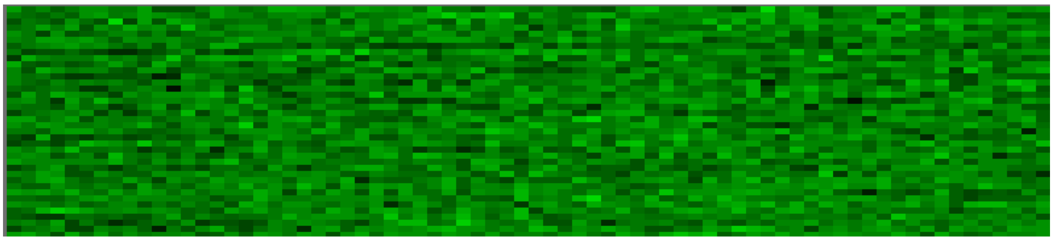


Illustration 9: GC Content of 56,160 bases averaged at 20 bases per block of Y. enterocolitica

The lightness of the green squares of the thermophilic bacteria when compared to bacteria which lives at colder temperatures is a striking indication of how organisms are suited for their environment.

Interestingly enough, when you create the same GC Content representation of the pathogenic bacteria causing tetanus (which can survive in soil for extended periods of time), it reveals a genome which is skewed towards a much lower GC content. Suggesting it were well suited to survive in colder temperatures.

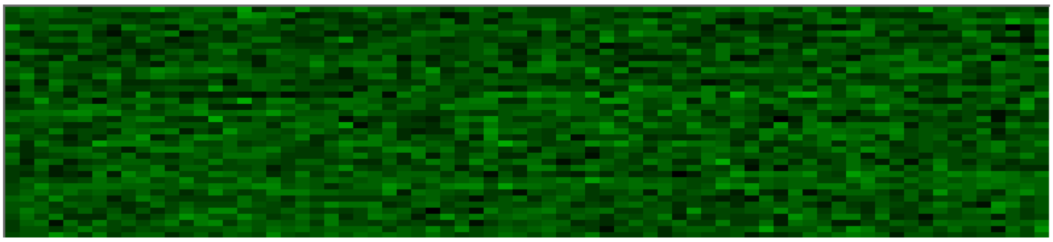


Illustration 10: GC Content of 56,160 bases averaged at 20 bases per block of Clostridium tetani (the cause of tetanis)

(Gene-centric association analysis for the correlation between the guanine-cytosine content levels and temperature range conditions of prokaryotic species: *It has been found that the environment that a bacteria lives can be predicted with >80% accuracy by looking at the GC Content percentage of its genome.*)
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3024870/

When I use the coloured representation to display the DNA of D. radiodurans, in particular the section of the DNA coding for the DNA-directed RNA-Polymerase in both strains of bacteria, it looks as follows:

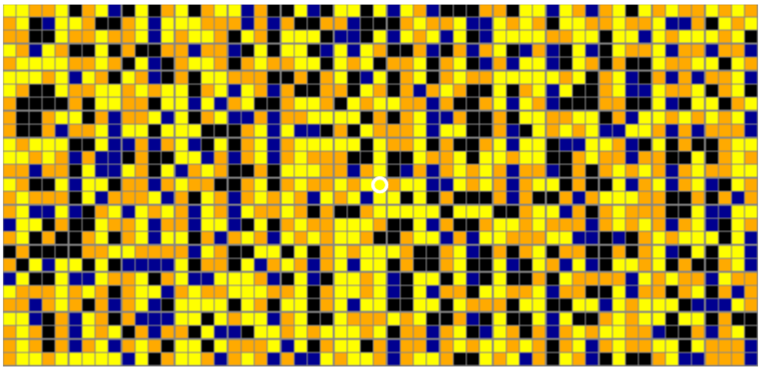


Illustration 11: Chr1:923733-925271 of D. radiodurans DNA-directed RNA polymerase beta subunit

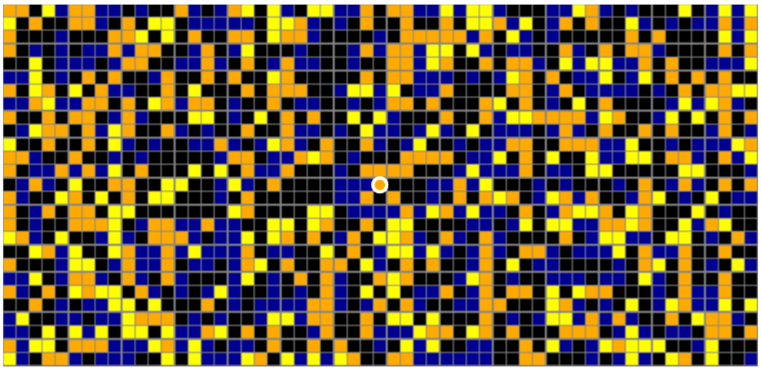


Illustration 12: C. tetani of DNA-directed RNA polymerase beta subunit

Interesting how both organisms has to build an enzyme with exactly the same function used for transcription, but utilizing codons picked from a different GC content DNA. Interestingly, when you look at all 61 possible codons of the Genetic Code, one finds that for each possible amino acid, there are equal amounts of codons coding for each amino acid that uses A's and T's as there are that uses G's and C'c, and this is divided equally among the amino acids.

I specifically designed this software to allow me to vary the number of bases depicted in each line of the output. Any repetitive pattern could then be discovered easily by simply changing the width until a pattern became evident when you looked at it. The human visual system is sometimes able to discriminate patterns much more easily than sophisticated software.

One such an example is when you look at the alpha satellite repeats found at the centromeres of human chromosomes.

Here is a 2D display of the centromeric region HG38 chr1:122939341-122970111 (Width=170 bases) The CENP-B Box region is highlighted in Magenta.

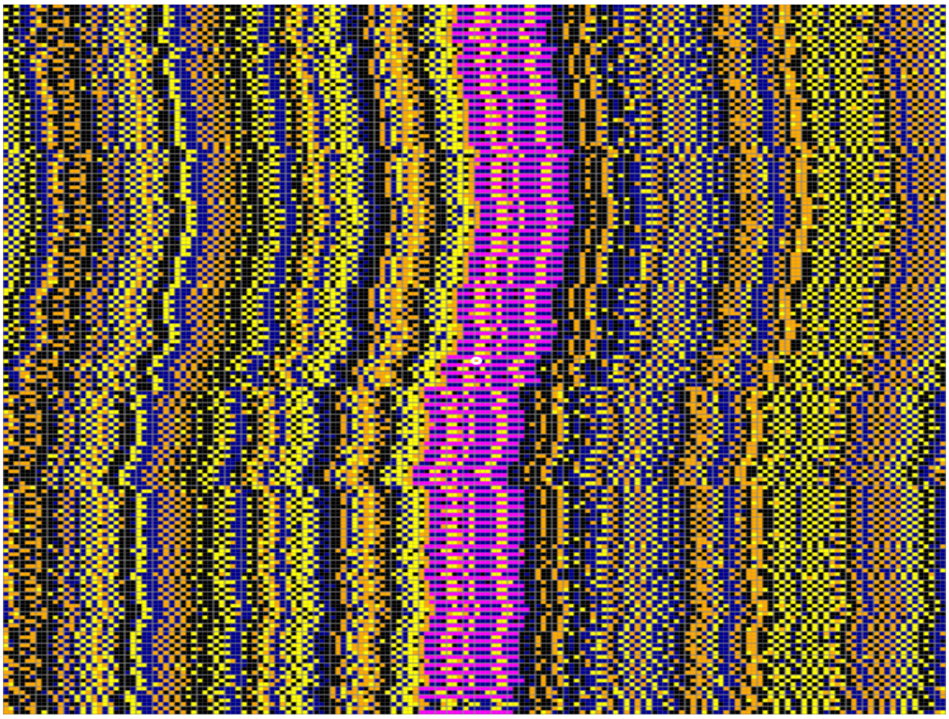


Illustration 13: chr1:122939341-122970111 at width 170. Clearly showing the minimum width repeat. Wider line widths simply show the repeat multiple times in each row.

The magenta highlight is part of the pattern search and “probe” functionality of the software which is the equivalent of using fluorescently tagged DNA probes to find specific target sequences via hybridization. However in this software I am using the Boyer-Moore string search algorithm, which I have modified to account for wildcards and multiple possible bases in each sequence position. It also allows you to specify the number of base mismatches it must allow when searching in the genome sequence.

One simply has to enter a search for the **Cenp-B box DNA sequence**:

CTTCGTTGGAAACGGGA

Allowing 2 mismatches it can be found almost 9000 times ONLY in the centromere of the X-Chromosome.

CpG methylation of the CENP-B box reduces human CENP-B binding (http://onlinelibrary.wiley.com/doi/10.1111/j.1432-1033.2004.04406.x/full)

If one wants to search for the sequence, but with either a T or an A in specific positions, it can be put in square brackets:

C***[TA][TA]***CG***[TA][TA]***GG***[AT][AT][AT]***CGGG***[AT]***

Any DNA sequence entered into the **Search field** such as this will be highlighted in the positive and negative (reverse complement) direction in the DNA View.

Additionally, the **“Find Pattern”** button can be used to search the entire genome through multiple selected chromosomes in parallel (threads).

The **Main Genome View** then displays these findings as differently coloured blocks in the **view** as is illustrated below. One can clearly see where the centromere ends.

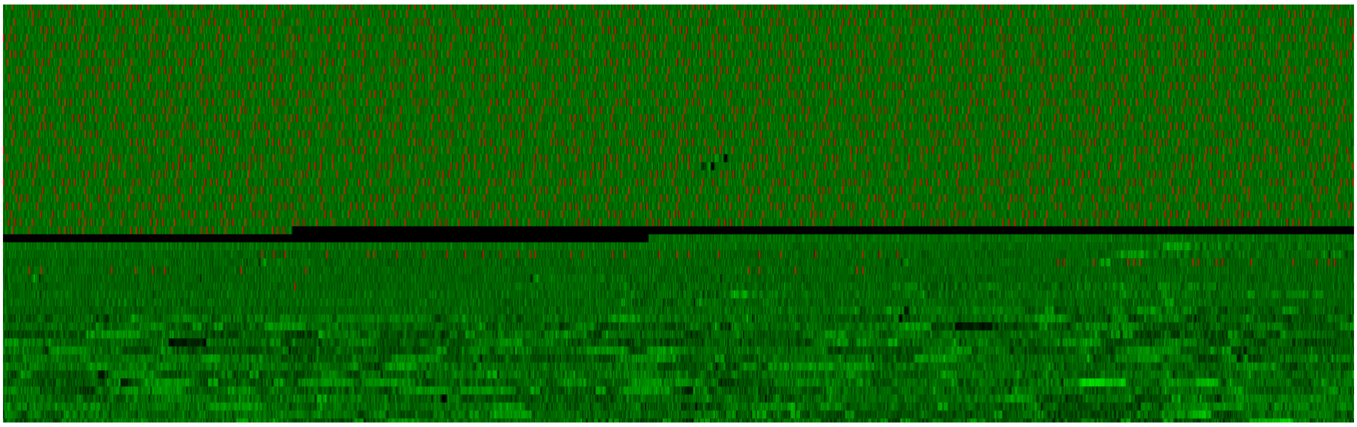


Illustration 14: Depiction of all the CENP-B box locations near the end of the X-Chromosome centromere.

From the literature it is known that the DNA binding proteins which are responsible for locating the CENP-B box DNA sequence within the centromeric DNA are part of the protein complex which form part of the kinetochore assembly in mammalian chromosomes. The 170-171 bases is known from literature to represent the number of bases between subsequent nucleosomes (containing the CENP-A variant of the H3 histone) around which the DNA is wrapped to produce the heterochromatin in the centromeres.

Any repetitive pattern is easily visualised by incrementally changing the bases per line until at 170-171 a pattern immediately emerges.

Looking at exactly the same region, but using a different line width of 80 bases, reveals no identifiable pattern, which indicates that it is easy to miss these patterns unless you vary the line width at a specific location until repetitive bases start to line up, which is easily spotted by inspection.

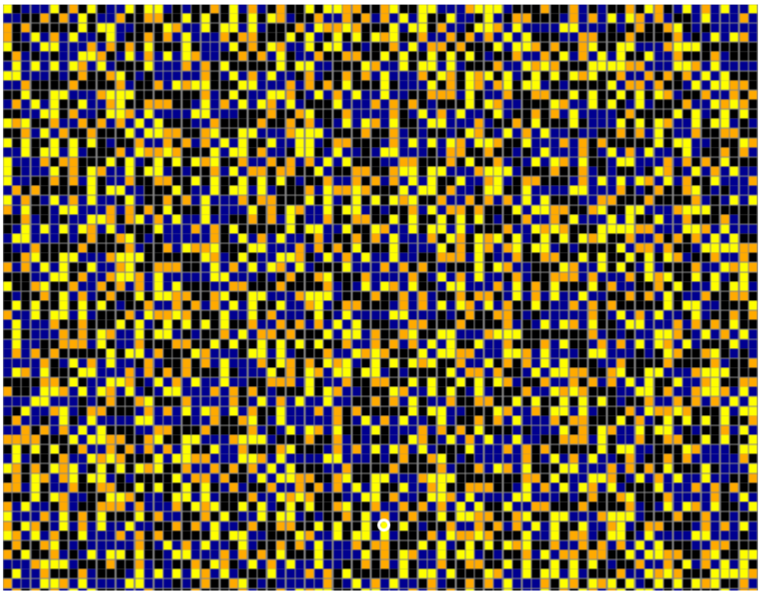


Illustration 15: chr1:123097253-123105973 at width 80. At most line width there are no discernible patterns.

Other non-centromeric regions of the genome also contains patterns, such as this one found at 114 bases width on chromosome 2.

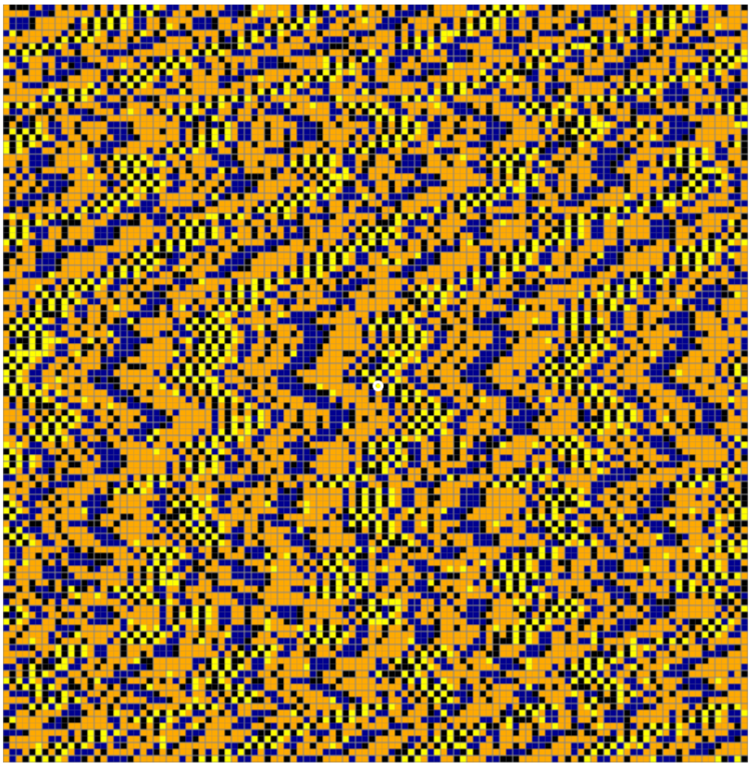


Illustration 16: chr2:1524735-1538073 at width 114

The above repeat is found in an intron of the thyroid peroxidase (TPO) gene.

The point that I wish to make is: It is possible to use the Visual Genome Browser's ability to represent the DNA sequence in a 2D matrix in order to locate and visualise many inherent structural features of the genome which is not always evident by just looking at a one dimensional representation of it.

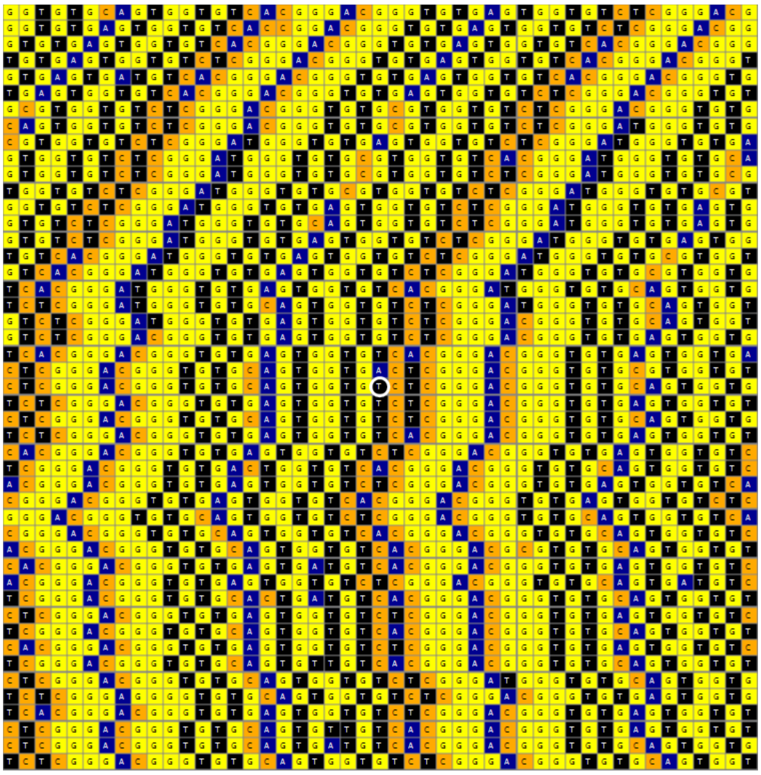
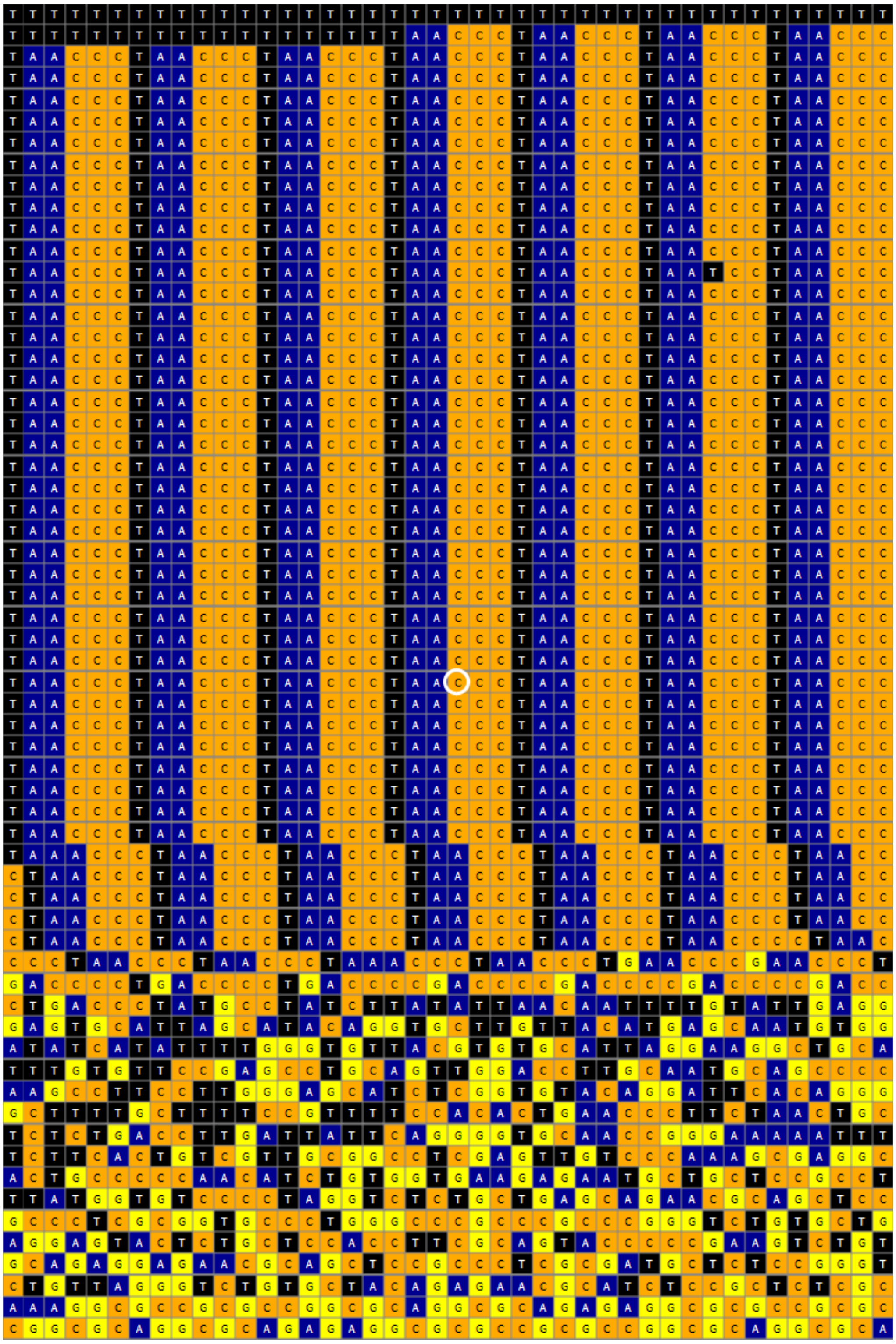


Illustration 17: Pattern at chr16:857351-859703 with width 47 bases per line

Other DNA sequence structures that can be identified are the telomeric sequences of chromosomes which consists of short repeats determined by the RNA template which is part of the DNA Telomerase enzyme responsible for lengthening the ends of chromosomes to overcome the Hayflick limit whereby the number of cell divisions are limited due to the shortening of the ends of linear chromosomes during replication.

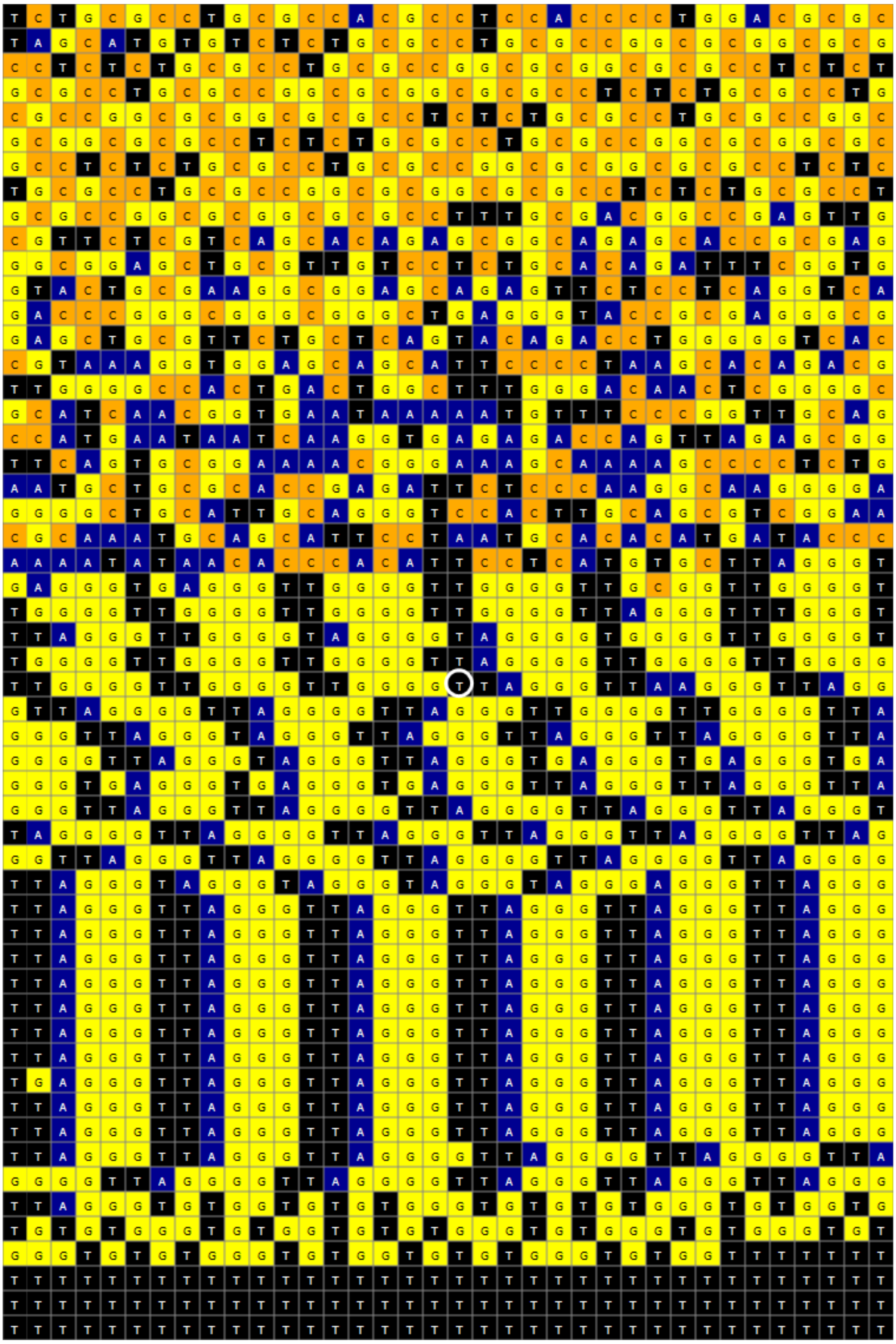


```
T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T
T T T T T T T T T T T T T T T T T T T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A T C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C
T A A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C
C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C
C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C
C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C
C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C T A A C C C C T A A C
C C C T A A C C C T A A C C C T A A A C C C T A A C C C T G A A C C C G A A C C C T
G A C C C C T G A C C C C T G A C C C C G A C C C C G A C C C C G A C C C C G A C C
C T G A C C C T A T G C C T A T C T T A T A T T A A C A A T T T T G T A T T G A G G
G A G T G C A T T A G C A T A C A G G T G C T T G T T A C A T G A G C A A T G T G G
A T A T C A T A T T T T G G G T G T T A C G T G T G C A T T A G G A A G G C T G C A
T T T G T G T T C C G A G C C T G C A G T T G G A C C T T G C A A T G C A G C C C C
A A G C C T T C C T T G G G A G C A T C T C G G T G T A C A G G A T T C A C A G G G
G C T T T T G C T T T T C C G T T T T C C A C A C T G A A C C C T T C T A A C T G C
T C T C T G A C C T T G A T T A T T C A G G G G T G C A A C C G G G A A A A A T T T
T C T T C A C T G T C G T T G C G G C C T C G A G T T G T C C C A A A G C G A G G C
A C T G C C C C C A A C A T C T G T G G T G A A G A G A A T G C T G C T C C G C C T
T T A T G G T G T C C C C T A G G T C T C T G C T G A G C A G A A C G C A G C T C C
G C C C T C G C G G T G C C C T G G G C C C G C C C G C C C G G G T C T G T G C T G
A G G A G T A C T C T G C T C C A C C T T C G C A G T A C C C C C G A A G T C T G T
G C A G A G G A G A A C G C A G C T C C G C C C T C G C G A T G C T C T C C G G G T
C T G T T A G G G T C T G T G C T A C A G A G A A C G C A T C T C C G C T C T C G C
A A A G G C G C C G C G C C G G C G C A G G C G C A G A G A G G C G C G C C G C G C
C G G C G C A G G C G C A G A G A G G C G C G C C G C G C C G G C G C A G G C G C A
```

Illustration 18: chr5:9941-12587 at width 42

At a line width of 42 (7 times the telomerase template width of 6) one can clearly see the highly repetitive TAACCC repeats at the end of chromosome 5.

Looking at the opposite end of chromosome X, one can see a similar pattern, but this time reverse complemented.



```
TCTGCGCCTGCGCCACGCCTCCACCCCTGGACGCGC
TAGCATGTGTCTCTGCGCCTGCGCCGGCGCGGCGCG
CCTCTCTGCGCCTGCGCCGGCGCGGCGCGCCTCTCT
GCGCCTGCGCCGGCGCGGCGCGCCTCTCTGCGCCTG
CGCCGGCGCGGCGCGCCTCTCTGCGCCTGCGCCGGC
GCGGCGCGCCTCTCTGCGCCTGCGCCGGCGCGGCGC
GCCTCTCTGCGCCTGCGCCGGCGCGGCGCGCCTCTC
TGCGCCTGCGCCGGCGCGGCGCGCCTCTCTGCGCCT
GCGCCGGCGCGGCGCGCCTTTGCGACGGCCGAGTTG
CGTTCTCGTCAGCACAGAGCGGCAGAGCACCGCGAG
GGCGGAGCTGCGTTGTCCTCTGCACAGATTTCGGTG
GTACTGCGAAGGCGGAGCAGAGTTCTCCTCAGGTCA
GACCCGGGCGGGCGGGCTGAGGGTACCGCGAGGGCG
GAGCTGCGTTCTGCTCAGTACAGACCTGGGGGTCAC
CGTAAAGGTGGAGCAGCATTCCCCTAAGCACAGACG
TTGGGGCCACTGACTGGCTTTGGGACAACTCGGGGC
GCATCAACGGTGAATAAAAATGTTTCCCGGTTGCAG
CCATGAATAATCAAGGTGAGAGACCAGTTAGAGCGG
TTCAGTGCGGAAAACGGGAAAGCAAAAGCCCCTCTG
AATGCTGCGCACCGAGATTCTCCCAAGGCAAGGGGA
GGGGCTGCATTGCAGGGTCCACTTGCAGCGTCGGAA
CGCAAATGCAGCATTCCTAATGCACACATGATACCC
AAAATATAACACCCACATTCCTCATGTGCTTAGGGT
GAGGGTGAGGGTTGGGGTTGGGGTTGCGGTTGGGGT
TGGGGTTGGGGTTGGGGTTGGGGTTAGGGTTTTGGGT
TTAGGGTTGGGGTAGGGGTAGGGGTGGGGTTGGGGT
TGGGGTTGGGGTTGGGGTTAGGGGTTGGGGTTGGGG
TTGGGGTTGGGGTTGGGGTTAGGGTTAAGGGTTAGG
GTTAGGGGTTAGGGGTTAGGGTTGGGGTTGGGGTTA
GGGTTAGGGTAGGGTTAGGGTTAGGGTTAGGGGTTA
GGGGTTAGGGTAGGGTTAGGGTGAGGGTGAGGGTGA
GGGTGAGGGTGAGGGTGAGGGTTAGGGTTAGGGTTA
GGGTTAGGGTTAGGGGTTAGGGGTTAGGGTTAGGGT
TAGGGGTTAGGGGTTAGGGTTAGGGTTAGGGGTTAG
GGTTAGGGTTAGGGGTTAGGGGTTAGGGGTTAGGGG
TTAGGGTAGGGTAGGGTAGGGTAGGGAGGGTTAGGG
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TGAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TTAGGGTTAGGGTTAGGGGTTAGGGGTTAGGGGTTA
GGGGTTAGGGGTTAGGGGTTAGGGTTAGGGTTAGGG
TTAGGGTGTGGTGTGTGGGTGTGTGTGGGTGTGGTG
TGTGTGGGTGTGGTGTGTGGGTGTGGGTGTGGGTGT
GGGTGTGTGGGTGTGGTGTGTGGGTGTGGTTTTTTT
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
```

Illustration 19: chrX:156029066-156031046 at width 36

The reverse complement GGGTTA of TAACCC can be found at the 3' end of the X Chromosome, this time repeating 6 times due to the width of 36. These patterns are known to be found in the RNA template which is part of the Telomerase protein-RNA complex.

Possible sources of genomes and sequence data for the Genome Browser

At its core, the Visual Genome Browser stores the genomic sequence data in the **.2bit** sequence format, which is the most compact (yet still uncompressed) format that you can use to represent DNA data. In essence this means that 2 bits of each byte is used to store a single DNA letter. Since you can represent 2^2 = 4 possibilities using 2 bits. This means one can fit in 8/2 = 4 DNA letters in each byte (8 bits) of data. The complete version human reference genome (HG38) can therefore by represented by 796 MB of information.

All of the genome data used by the Visual Genome Browser are maintained in a Data Folder which contains the genomes of different organisms in sub-folders. When you want to add a new organism's genome to the browser, you simply copy the downloaded **.2bit** genome into this Data Folder. The Browser will then scan this folder at start-up for all the **.2bit** files and add it to the list of genomes.

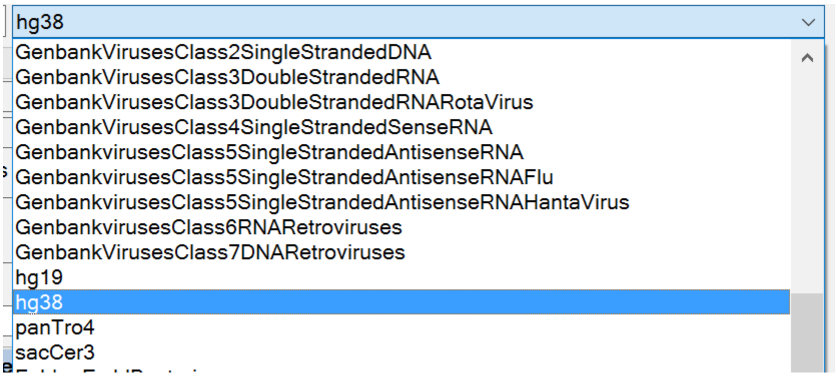
The UCSC maintains the gene annotation data in a MySQL database which has to be queried by the genome browser in order to download the gene annotations which are overlaid on the genome structure.

The software also scans the Data Folder for other file types which may contain DNA sequence data, such as FASTA, GENBANK and ENSEMBL browser files. These files need to be put in folders prefixed with the Text : "Fasta", "Genbank" or "Embl". It typically examines the headers of these files to determine if it is of the specified type. Fasta files generally do not contain annotation data, but this can be loaded separately from **BED**, **BigBED** or **GFF3** files after the DNA sequence is displayed. The files located in each of these specially prefixed folders are then combined into a single **.2bit** files for quick and uniform access (one **.2bit** file for each folder involved). If any files gets added or removed from these folders, the .2bit and the annotation data (from Genbank and Ensembl files) is also rebuilt. Another way to obtain annotated sequences is to enter the Genbank accession number into a field and click a download button.   Single Genbank, or a range of Genbank files are then downloaded from the

Url:

`https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nucleotide&id={0}&rettype=gbwithparts`

Each **.2bit** file corresponds to a single "folder", which may contain a list of possible sequences or "chromosomes". When you select a specific folder, you are able to search through all possible fields of all chromosomes or sequences of that folder. You can then go and pick a specific chromosome to display in the 2 dimensional **Main Genome View**. The browser loads all annotations for this sequence into a memory data structure (range tree) which allows for very quick overlap checks between genomic ranges in order to overlay the correct annotations on top of the GC Content display of the structure.

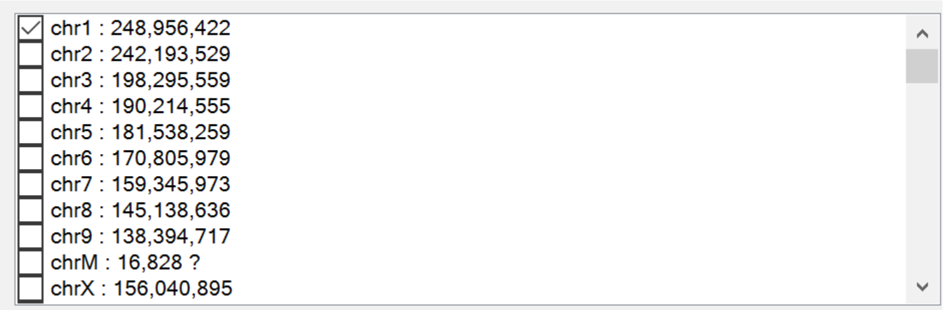


hg38

GenbankVirusesClass2SingleStrandedDNA
GenbankVirusesClass3DoubleStrandedRNA
GenbankVirusesClass3DoubleStrandedRNARotaVirus
GenbankVirusesClass4SingleStrandedSenseRNA
GenbankvirusesClass5SingleStrandedAntisenseRNA
GenbankvirusesClass5SingleStrandedAntisenseRNAFlu
GenbankvirusesClass5SingleStrandedAntisenseRNAHantaVirus
GenbankvirusesClass6RNARetroviruses
GenbankVirusesClass7DNARetroviruses
hg19
hg38
panTro4
sacCer3

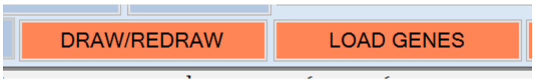
Illustration 20: Folder List

When you select a folder, the sequence or chromosome list is populated with sizes for the sequences



☑ chr1 : 248,956,422
□ chr2 : 242,193,529
□ chr3 : 198,295,559
□ chr4 : 190,214,555
□ chr5 : 181,538,259
□ chr6 : 170,805,979
□ chr7 : 159,345,973
□ chr8 : 145,138,636
□ chr9 : 138,394,717
□ chrM : 16,828 ?
□ chrX : 156,040,895

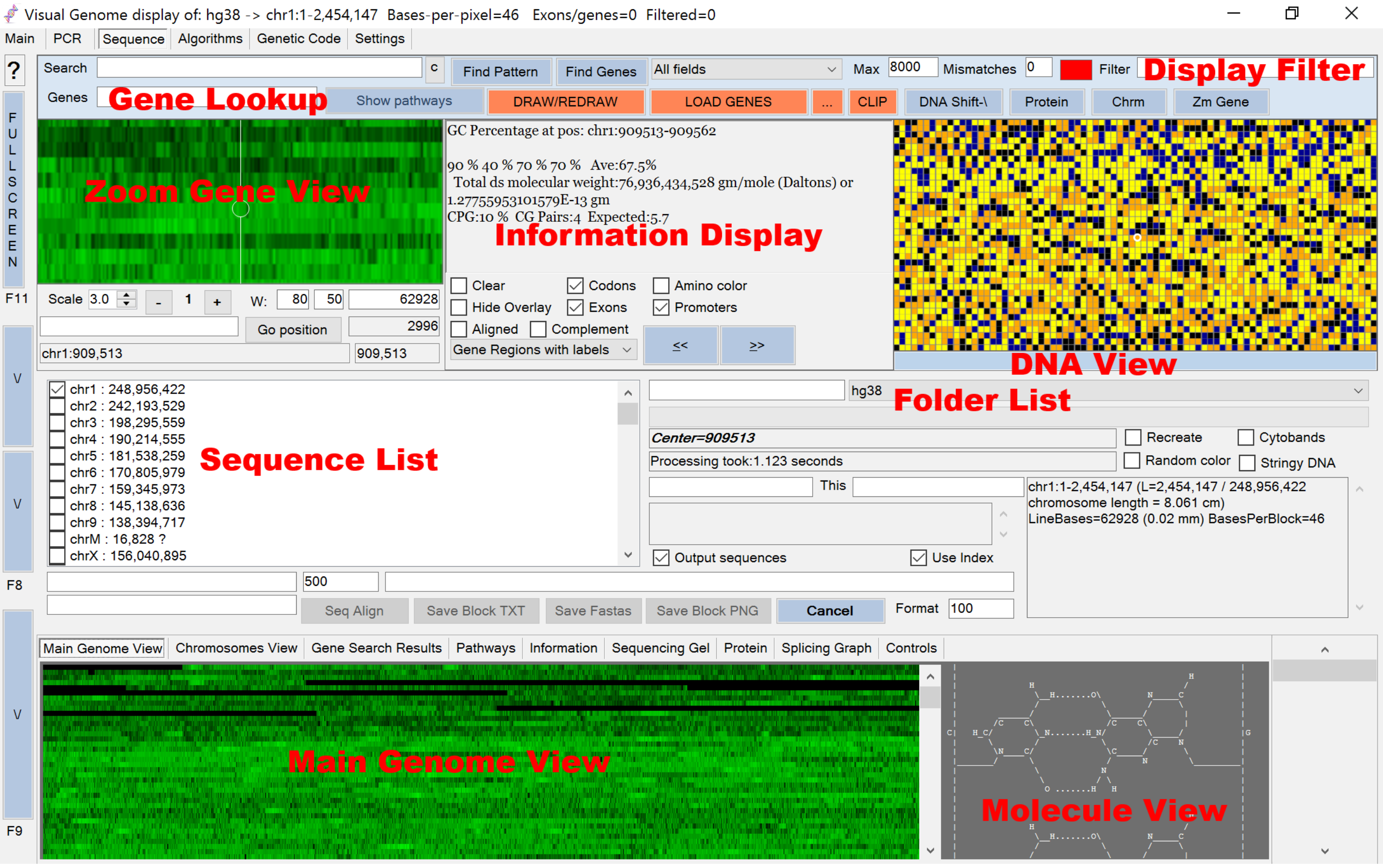
Illustration 21: Sequence or chromosome list



The Different Views

Clicking on “DRAW/REDRAW” will load the 2bit data and display it in the **Main Genome View.**

If you also want to load the gene overlay, you click “LOAD GENES”. The chromosomes selected in the sequence list always determines which sequences' annotations are loaded or searched for genes/patterns. The selected chromosome is then loaded into the Main Genome Browser screen.



Visual Genome display of: hg38 -> chr1:1-2,454,147 Bases-per-pixel=46 Exons/genes=0 Filtered=0

GC Percentage at pos: chr1:909513-909562

90 % 40 % 70 % 70 % Ave:67.5%
Total ds molecular weight:76,936,434,528 gm/mole (Daltons) or 1.27755953101579E-13 gm
CPG:10 % CG Pairs:4 Expected:5.7

Max 8000 Mismatches 0

Scale 3.0 W: 80 50 62928 2996

chr1:909,513 909,513

□ Clear ☑ Codons □ Amino color
□ Hide Overlay ☑ Exons ☑ Promoters
□ Aligned □ Complement
Gene Regions with labels

☑ chr1 : 248,956,422
□ chr2 : 242,193,529
□ chr3 : 198,295,559
□ chr4 : 190,214,555
□ chr5 : 181,538,259
□ chr6 : 170,805,979
□ chr7 : 159,345,973
□ chr8 : 145,138,636
□ chr9 : 138,394,717
□ chrM : 16,828 ?
□ chrX : 156,040,895

hg38

Center=909513
Processing took:1.123 seconds

□ Recreate □ Cytobands
□ Random color □ Stringy DNA

chr1:1-2,454,147 (L=2,454,147 / 248,956,422
chromosome length = 8.061 cm)
LineBases=62928 (0.02 mm) BasesPerBlock=46

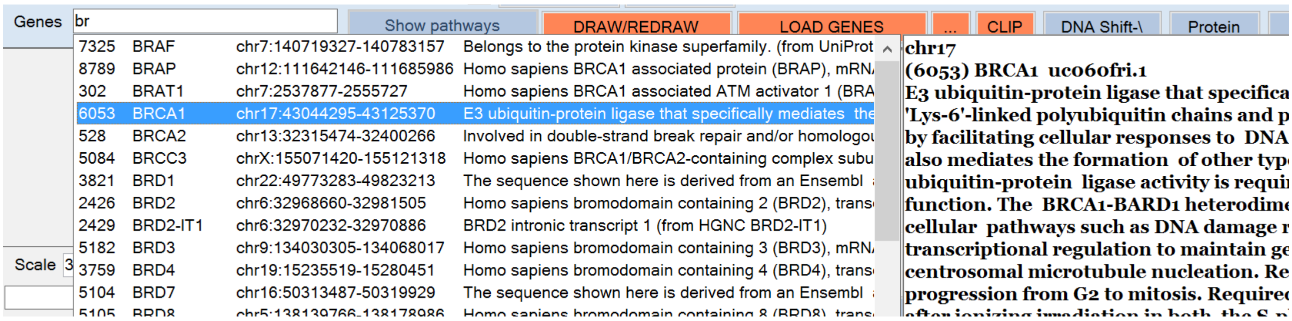
☑ Output sequences ☑ Use Index

500

Format 100

Illustration 22: Main Genome Browser Screen

All available annotation data are kept in indexed binary files making them searchable and allowing fast lookup of genes from a single text input (similar to the Google predictive search, which gives you suggestions as you start typing the gene names).



Genes br

7325	BRAF	chr7:140719327-140783157	Belongs to the protein kinase superfamily. (from UniProt
8789	BRAP	chr12:111642146-111685986	Homo sapiens BRCA1 associated protein (BRAP), mRN/
302	BRAT1	chr7:2537877-2555727	Homo sapiens BRCA1 associated ATM activator 1 (BRA
6053	BRCA1	chr17:43044295-43125370	E3 ubiquitin-protein ligase that specifically mediates the
528	BRCA2	chr13:32315474-32400266	Involved in double-strand break repair and/or homologou
5084	BRCC3	chrX:155071420-155121318	Homo sapiens BRCA1/BRCA2-containing complex subu
3821	BRD1	chr22:49773283-49823213	The sequence shown here is derived from an Ensembl
2426	BRD2	chr6:32968660-32981505	Homo sapiens bromodomain containing 2 (BRD2), trans
2429	BRD2-IT1	chr6:32970232-32970886	BRD2 intronic transcript 1 (from HGNC BRD2-IT1)
5182	BRD3	chr9:134030305-134068017	Homo sapiens bromodomain containing 3 (BRD3), mRN/
3759	BRD4	chr19:15235519-15280451	Homo sapiens bromodomain containing 4 (BRD4), trans
5104	BRD7	chr16:50313487-50319929	The sequence shown here is derived from an Ensembl
5105	BRD8	chr5:138139766-138178986	Homo sapiens bromodomain containing 8 (BRD8), trans

chr17
(6053) BRCA1 uc060fri.1
E3 ubiquitin-protein ligase that specifica
'Lys-6'-linked polyubiquitin chains and p
by facilitating cellular responses to DNA
also mediates the formation of other typ
ubiquitin-protein ligase activity is requi
function. The BRCA1-BARD1 heterodime
cellular pathways such as DNA damage r
transcriptional regulation to maintain ge
centrosomal microtubule nucleation. Re
progression from G2 to mitosis. Require
after ionizing irradiation in both the S p

Illustration 23: Predictive gene lookup

This is the default field which has the focus when the software starts up, allowing the user to start a gene lookup immediately. (This field can also be reached by pressing **Ctrl-G**)

As you change the selection, a pop up window shows pertinent information regarding the selected gene. When you press enter, the software will locate the gene and highlight it in the corresponding navigable views.

In addition to the quick lookup, one can also do an exhaustive search through all or selected fields in the annotation data. When (**Selected Chr**) is chosen, only the checked sequences will be searched, otherwise all the index will be used to search through all of the sequences in the current folder.

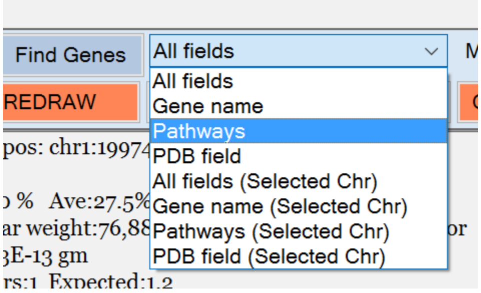
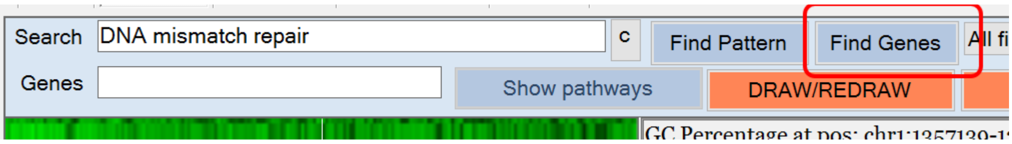


Illustration 24: Fields that can be searched



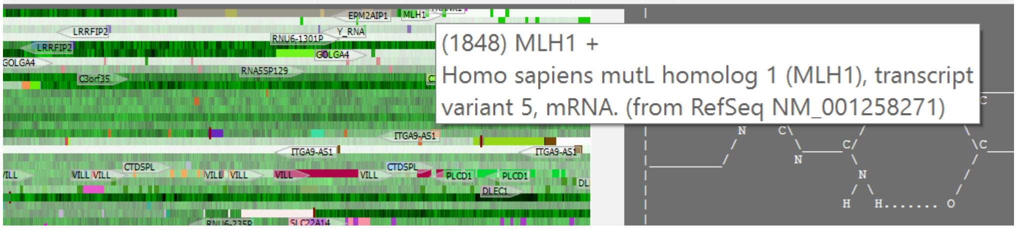
Search DNA mismatch repair

Genes

Illustration 25: Searching all fields

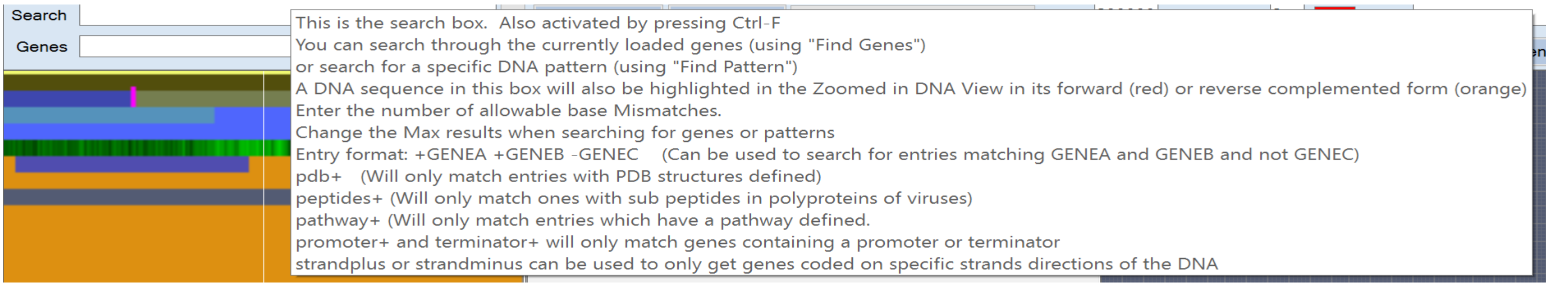
The gene results are then given in the **Gene Search Results** tab as shown in the following picture.

When you double-click on any gene, the software highlights the selected gene in the **Main Genome View**.



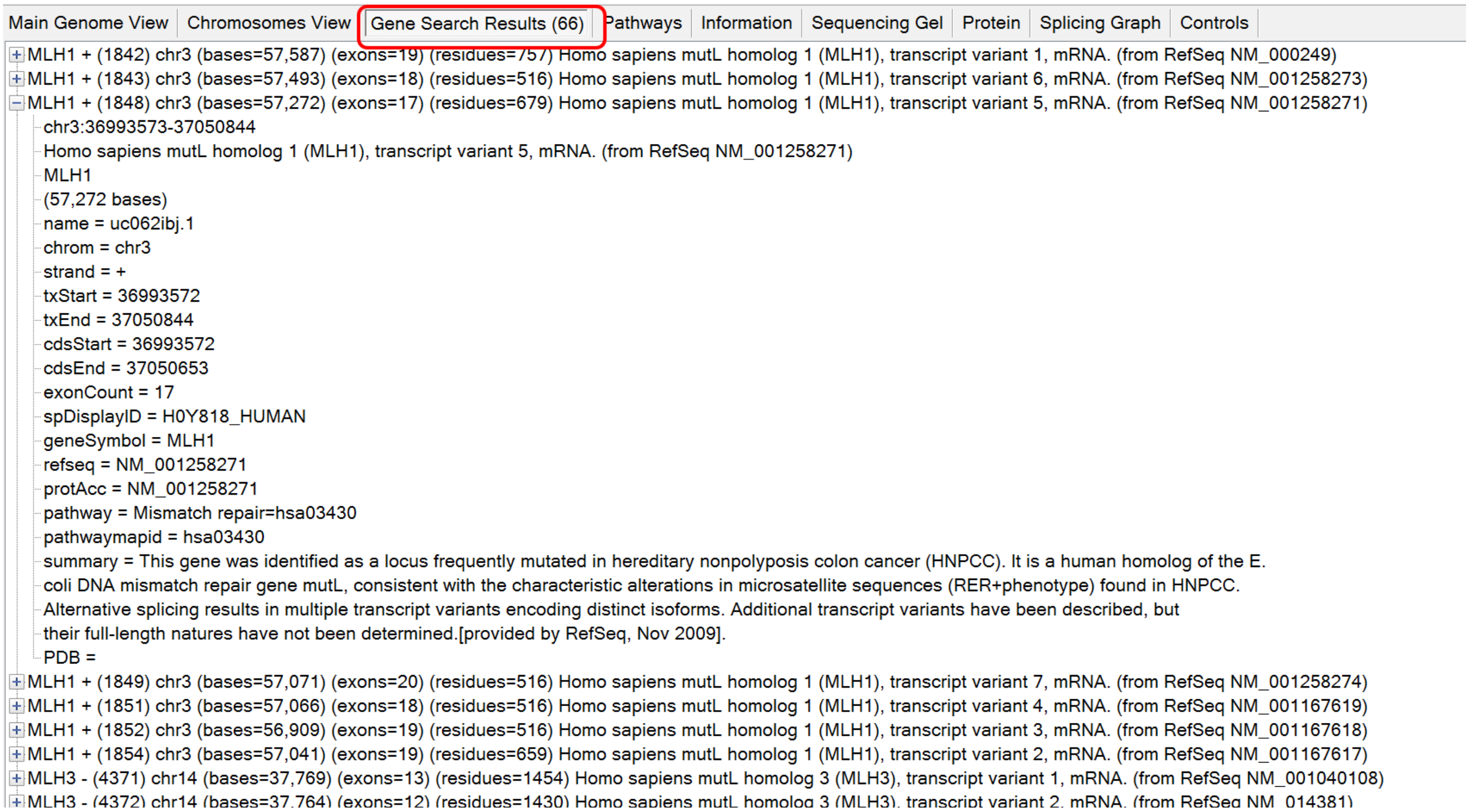
(1848) MLH1 +
Homo sapiens mutL homolog 1 (MLH1), transcript variant 5, mRNA. (from RefSeq NM_001258271)

Illustration 26: Selected gene in Main Genome View



This is the search box. Also activated by pressing Ctrl-F
You can search through the currently loaded genes (using "Find Genes")
or search for a specific DNA pattern (using "Find Pattern")
A DNA sequence in this box will also be highlighted in the Zoomed in DNA View in its forward (red) or reverse complemented form (orange)
Enter the number of allowable base Mismatches.
Change the Max results when searching for genes or patterns
Entry format: +GENEA +GENEB -GENEC (Can be used to search for entries matching GENEA and GENEB and not GENEC)
pdb+ (Will only match entries with PDB structures defined)
peptides+ (Will only match ones with sub peptides in polyproteins of viruses)
pathway+ (Will only match entries which have a pathway defined.
promoter+ and terminator+ will only match genes containing a promoter or terminator
strandplus or strandminus can be used to only get genes coded on specific strands directions of the DNA

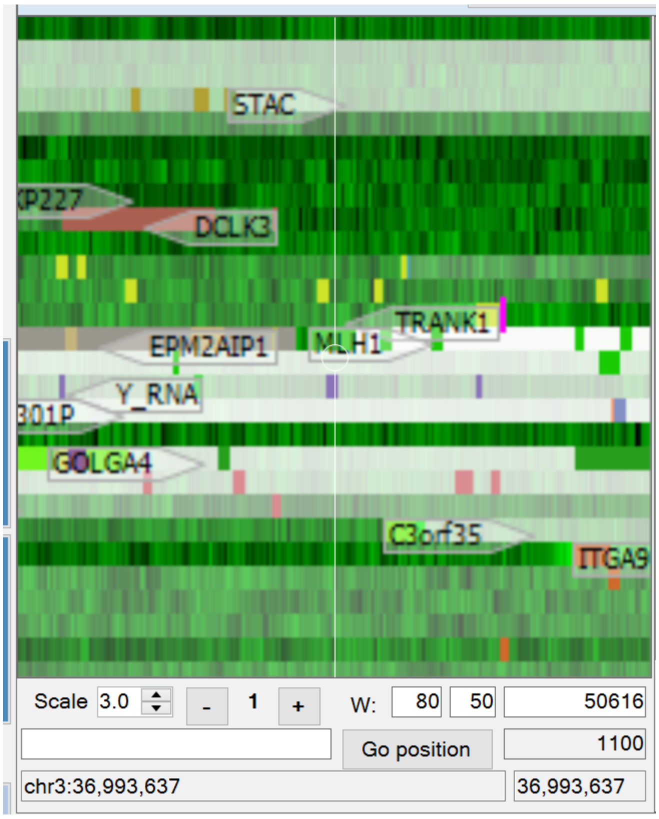
Illustration 27: Search options



```
⊞MLH1 + (1842) chr3 (bases=57,587) (exons=19) (residues=757) Homo sapiens mutL homolog 1 (MLH1), transcript variant 1, mRNA. (from RefSeq NM_000249)
⊞MLH1 + (1843) chr3 (bases=57,493) (exons=18) (residues=516) Homo sapiens mutL homolog 1 (MLH1), transcript variant 6, mRNA. (from RefSeq NM_001258273)
⊟MLH1 + (1848) chr3 (bases=57,272) (exons=17) (residues=679) Homo sapiens mutL homolog 1 (MLH1), transcript variant 5, mRNA. (from RefSeq NM_001258271)
   chr3:36993573-37050844
   Homo sapiens mutL homolog 1 (MLH1), transcript variant 5, mRNA. (from RefSeq NM_001258271)
   MLH1
   (57,272 bases)
   name = uc062ibj.1
   chrom = chr3
   strand = +
   txStart = 36993572
   txEnd = 37050844
   cdsStart = 36993572
   cdsEnd = 37050653
   exonCount = 17
   spDisplayID = H0Y818_HUMAN
   geneSymbol = MLH1
   refseq = NM_001258271
   protAcc = NM_001258271
   pathway = Mismatch repair=hsa03430
   pathwaymapid = hsa03430
   summary = This gene was identified as a locus frequently mutated in hereditary nonpolyposis colon cancer (HNPCC). It is a human homolog of the E.
   coli DNA mismatch repair gene mutL, consistent with the characteristic alterations in microsatellite sequences (RER+phenotype) found in HNPCC.
   Alternative splicing results in multiple transcript variants encoding distinct isoforms. Additional transcript variants have been described, but
   their full-length natures have not been determined.[provided by RefSeq, Nov 2009].
   PDB =
⊞MLH1 + (1849) chr3 (bases=57,071) (exons=20) (residues=516) Homo sapiens mutL homolog 1 (MLH1), transcript variant 7, mRNA. (from RefSeq NM_001258274)
⊞MLH1 + (1851) chr3 (bases=57,066) (exons=18) (residues=516) Homo sapiens mutL homolog 1 (MLH1), transcript variant 4, mRNA. (from RefSeq NM_001167619)
⊞MLH1 + (1852) chr3 (bases=56,909) (exons=19) (residues=516) Homo sapiens mutL homolog 1 (MLH1), transcript variant 3, mRNA. (from RefSeq NM_001167618)
⊞MLH1 + (1854) chr3 (bases=57,041) (exons=19) (residues=659) Homo sapiens mutL homolog 1 (MLH1), transcript variant 2, mRNA. (from RefSeq NM_001167617)
⊞MLH3 - (4371) chr14 (bases=37,769) (exons=13) (residues=1454) Homo sapiens mutL homolog 3 (MLH3), transcript variant 1, mRNA. (from RefSeq NM_001040108)
⊞MLH3 - (4372) chr14 (bases=37,764) (exons=12) (residues=1430) Homo sapiens mutL homolog 3 (MLH3), transcript variant 2, mRNA. (from RefSeq NM_014381)
```

Illustration 28: Gene Annotation Results

Illustration 29: Zoom Gene View



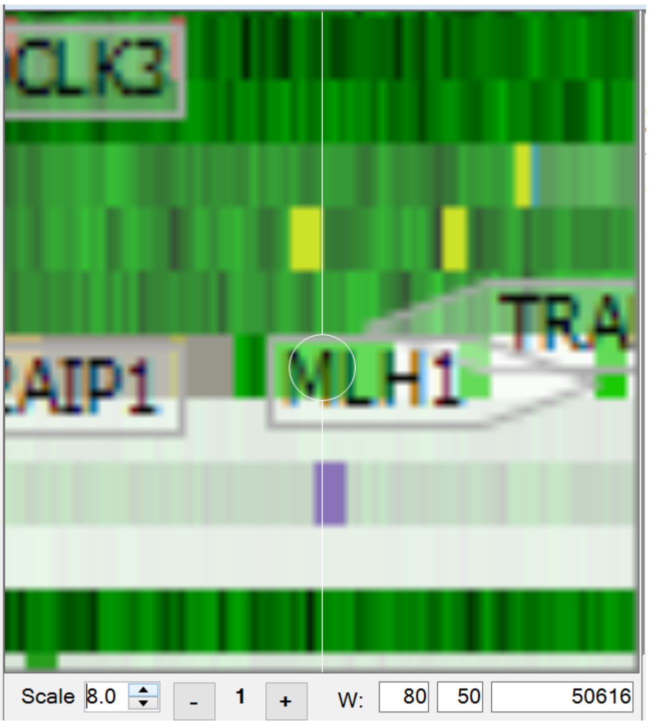
Scale 3.0 1 W: 80 50 50616

1100

chr3:36,993,637 36,993,637

*The **Zoom Gene View** provides you with a scalable magnified view of the 2D genomic region in question.*

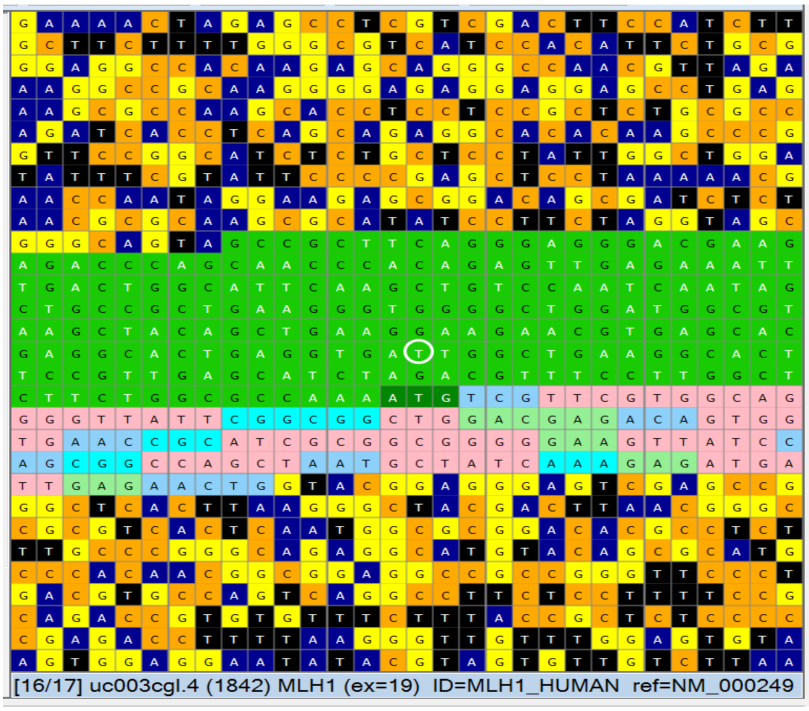
When the magnification is increased, it provides you with a more close up view:



Scale 8.0 1 W: 80 50 50616

Illustration 30: 8x Magnified Zoom Gene View

DNA View *shows the finer detail at sequence level.*



[16/17] uc003cgl.4 (1842) MLH1 (ex=19) ID=MLH1_HUMAN ref=NM_000249

Illustration 31: DNA View showing the ***MLH1*** *gene's 5'UTR*

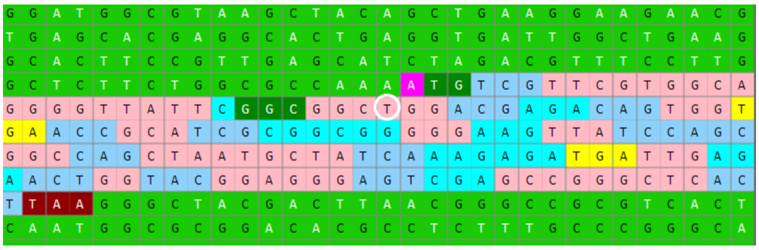
This is where you can really get acquainted with the primary structure of the gene DNA sequence.

In the DNA view above you can see the 5' Untranslated Region (UTR), the first exon, the start codon (ATG) and part of the protein coding sequence.

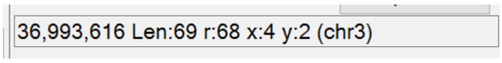
Navigation around the genome is achieved via the cursor keys of the keyboard, the mouse via a dragging motion or via touch on a machine with a touch display. All display windows are **navigationally locked**, meaning when you pan or scroll one window, the other zoomed widows moves accordingly. This allows for intuitive operation around the chromosomes of the genome.

Copying DNA sequences

By moving the mouse in the **DNA View**, the base position at the mouse is displayed in a field on the left. By **double clicking** anywhere in the **DNA View**, a Magenta block is inserted. Moving the mouse will now show you the distance between the mouse location and the magenta base position.



The field displays the **Centre position, the length in bases from the magenta marker, the base offset (r), the x-difference and the y-difference in position. (as well as the chromosome)**



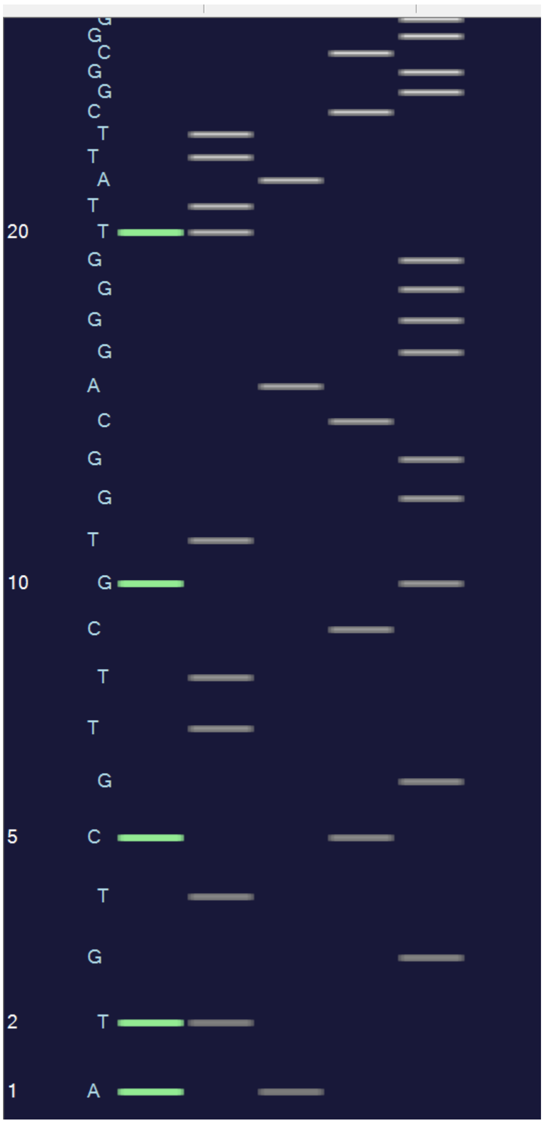
36,993,616 Len:69 r:68 x:4 y:2 (chr3)

Double clicking again will copy the bases from the magenta base to the current base to the information tab **as well as to the windows clipboard**.

```
chr3:36993548-36993694

ATGTCGTTCGTGGCAGGGGTTATTCGGCGGCTGGACGAGACAGTGGTGAACCGCATCGCGGCGGGGGAAGTTATCCAGCGGCCAGCTAATGCTATCAAAG
AGATGATTGAGAACTGGTACGGAGGGAGTCGAGCCGGGCTCACTTAA
```

Sequencing Gel View

And, if the sequence is the right length, a “simulated gel electrophoresis” will be shown in the “**Sequencing Gel**” tab:



The Gel View is really **just because is looks cool and reminds us of how far genome sequencing has come**. (It tries to fit in the number of selected bases based on the gel density and length, but sometimes it does not find the proper parameters and does not display it).

Using the Visual Genome Browser views and keyboard keys

The software has been written with the purpose of making navigating genomes in 2D as easy as possible. All views are navigationally locked to each other allowing a user to move the other views by dragging any one of the Main Genome View, DNA View or Zoomed Gene views.

The views can be moved by **the mouse**, **the keyboard arrow keys** or **by touch** on a computer with a touch display such as the Microsoft Surface Pro (on which it was developed).

Many of the options are also available by keyboard shortcuts:

Keyboard and mouse shortcuts

Mouse wheel

- Move in Zoom DNA View

Mouse left button

- Drag and move views

Mouse centre button

- Jump between Main Genome View and DNA View to correct genome position

Mouse right button

- Gene context menu (If no gene shown, menu for current genome position)

Ctrl

- Press and hold to temporarily enable tooltips
- When moving the mouse in the Main Genome View, will lock the vertical position and only follow the horizontal position of the mouse in the other windows. This makes it easier to track the mouse from left to right on the same genome sequence.
- Holding down the CTRL key while double clicking in the chromosomes list, will select only the most important chromosomes for the Human genomes

Shift

- Increase arrow key navigation speed by 10 in the Main Genome View
- Holding down the SHIFT key while double clicking in the chromosomes list, will select all of the sequences

Alt

- When pressed will slow down the speed of the mouse cursor in the Main Genome View

F1

- Show gene info screen from where you can create favourites

F2

- Show History

F3

- Show Favourites

F4

- Jump to last history or favourite selected

F5

- Show Protein view for current gene or redraw protein view

F6

- Show context menu for current gene

F7

- Move mouse between Main Genome View and Zoom DNA View at correct position in genome

F8

- Collapse/Expand middle panel

F9

- Collapse/Expand bottom panel

F11

- Toggle Full Screen View

F12

- Toggle between current and last expression cell type and redraw genome

Page Down/Up

- Jump to next/previous gene/filtered position

Arrow keys

- Navigate in Main Genome View or Zoom DNA View (Shift key increases move speed by 10)

+ or -

- When mouse cursor in Main Genome View, Zoom DNA View, Splicing View, Zoomed Genes View
Cycle through different overlapping genes/transcripts at current position
- When showing All Chromosomes view
Change expression level threshold redraw expression overlay

< or >

- Change number of bases per line in "Aligned" Zoom DNA View
- When mouse cursor in Main Genome View, Zoom DNA View, Splicing View, Zoomed Genes View, Protein View
Jump to next/previous exon when appropriate context menu is selected
- If Exon selection not active it will select next/previous ALLELE when showing Variants
- When showing All Chromosomes view
Change expression level threshold redraw expression overlay

[OR]

- When Main Genome View tab selected and mouse cursor in Main Genome View, Zoom DNA View, Zoomed Genes View
- Change bases per line in Zoom DNA View
- When Protein View tab selected and mouse cursor in Protein view
Change amino acids per line in Protein View

Shift-/

- When Mouse cursor in Main Genome View or Zoom DNA View
Copy DNA bases from DNA View as text to clipboard

Ctrl-F

- Give focus to Gene/Pattern Search field allowing you to search through gene fields or look for patterns

Ctrl-G

- Give focus to Gene quick lookup combo allowing you to type and look for genes

Ctrl-S

- Save Favourites

Ctrl-A

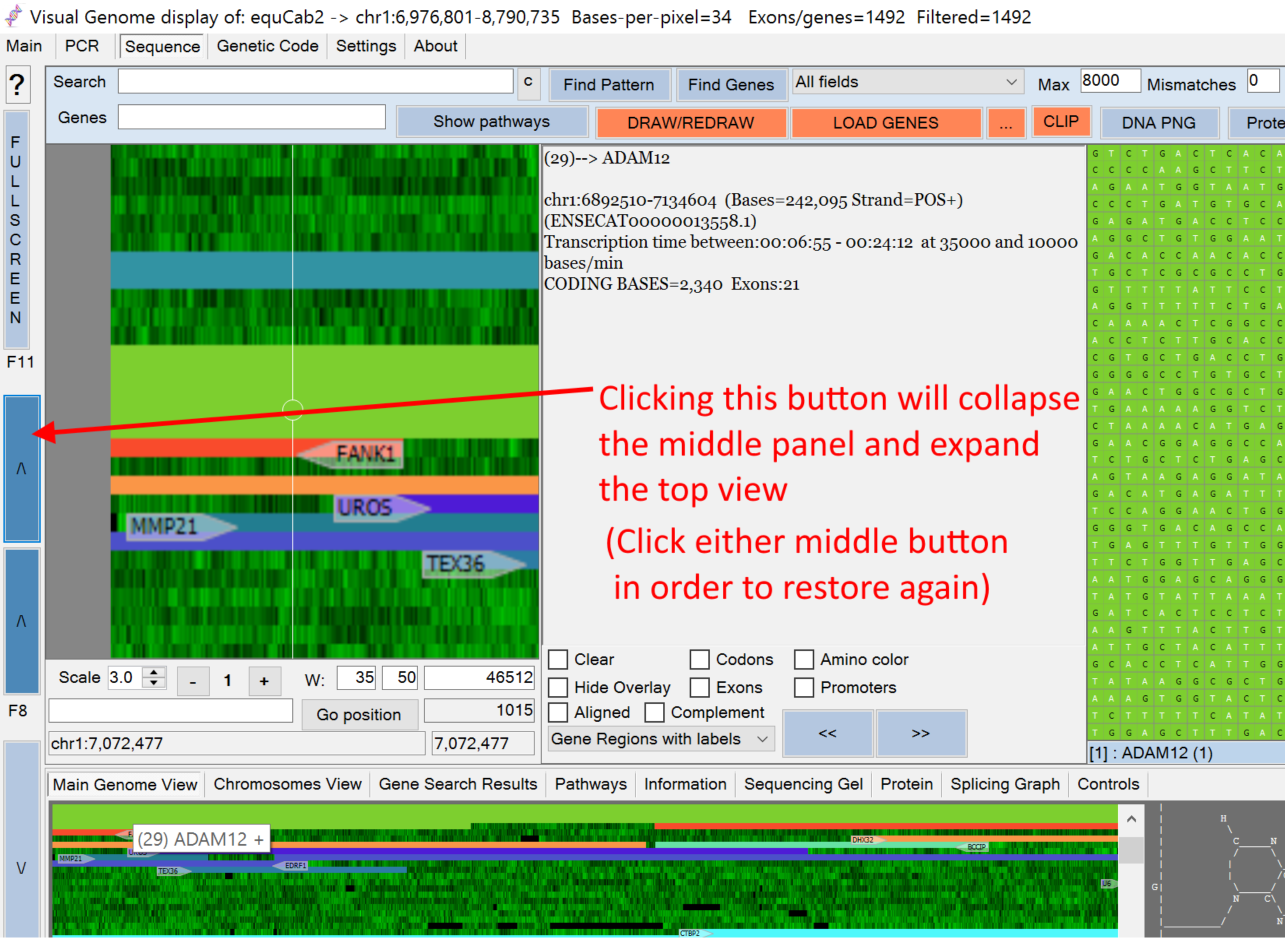
- Select all text in most text boxes

Windows Menu key

- Show gene context menu for current position of mouse in Main Genome View

The buttons on the left of the Sequence tab allows you to collapse/expand the panels to suit your preference

It is sometimes very difficult to fit everything one wants to display on the same view. This is why the software provides buttons to collapse either the middle or bottom panels, or even go into full-screen mode, in order to better view the genomes.



Visual Genome display of: equCab2 -> chr1:6,976,801-8,790,735 Bases-per-pixel=34 Exons/genes=1492 Filtered=1492

Max 8000 Mismatches 0

(29)--> ADAM12

chr1:6892510-7134604 (Bases=242,095 Strand=POS+)
(ENSECAT00000013558.1)
Transcription time between:00:06:55 - 00:24:12 at 35000 and 10000 bases/min
CODING BASES=2,340 Exons:21

Clicking this button will collapse the middle panel and expand the top view

(Click either middle button in order to restore again)

Scale 3.0 W: 35 50 46512

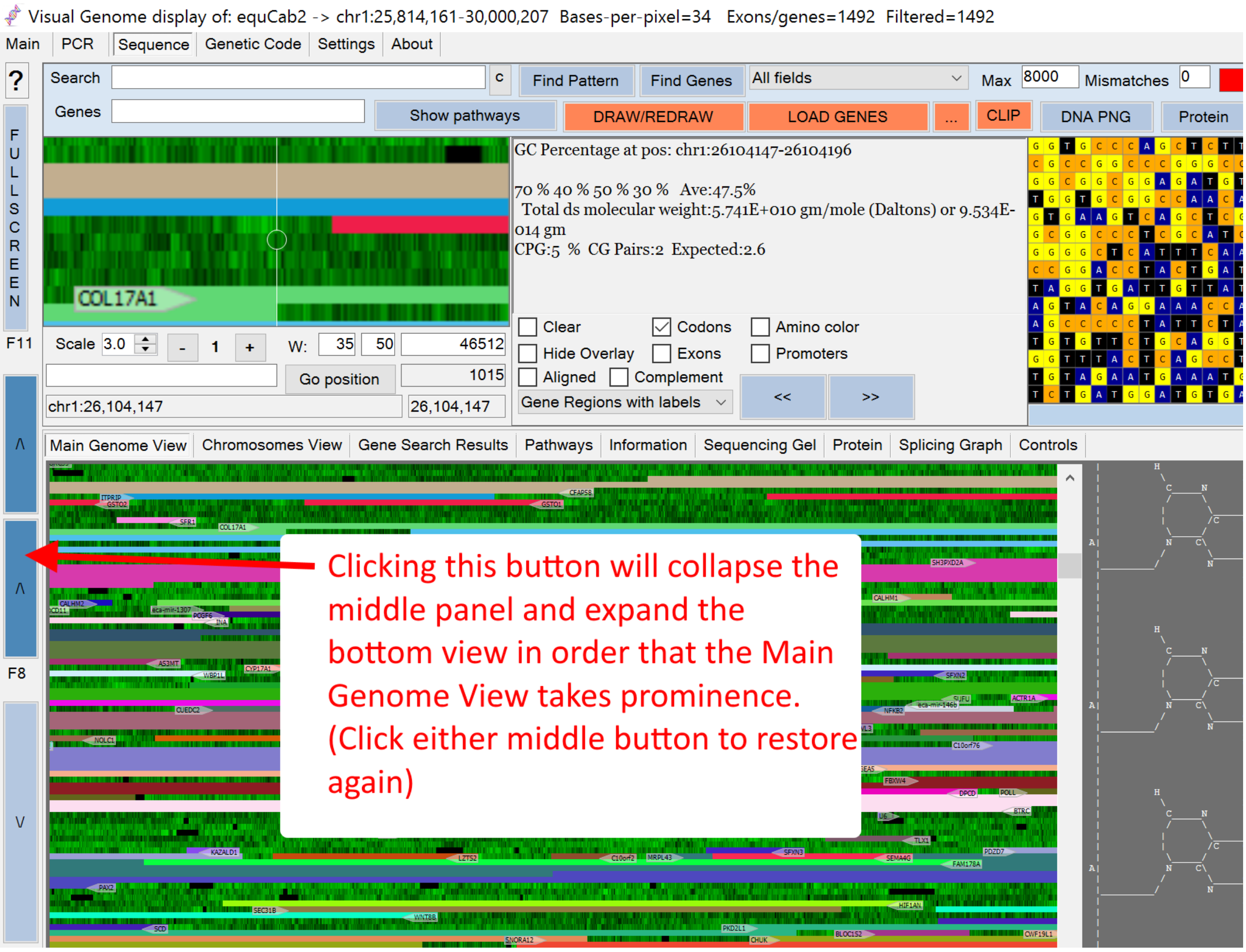
1015

chr1:7,072,477 7,072,477

[1] : ADAM12 (1)

(29) ADAM12 +

Illustration 32: With the TOP PANEL prominent


Visual Genome display of: equCab2 -> chr1:25,814,161-30,000,207 Bases-per-pixel=34 Exons/genes=1492 Filtered=1492

GC Percentage at pos: chr1:26104147-26104196

70 % 40 % 50 % 30 % Ave:47.5%
Total ds molecular weight:5.741E+010 gm/mole (Daltons) or 9.534E-014 gm
CPG:5 % CG Pairs:2 Expected:2.6

☐ Clear ☑ Codons ☐ Amino color
☐ Hide Overlay ☐ Exons ☐ Promoters
☐ Aligned ☐ Complement

Scale 3.0 W: 35 50 46512 1015

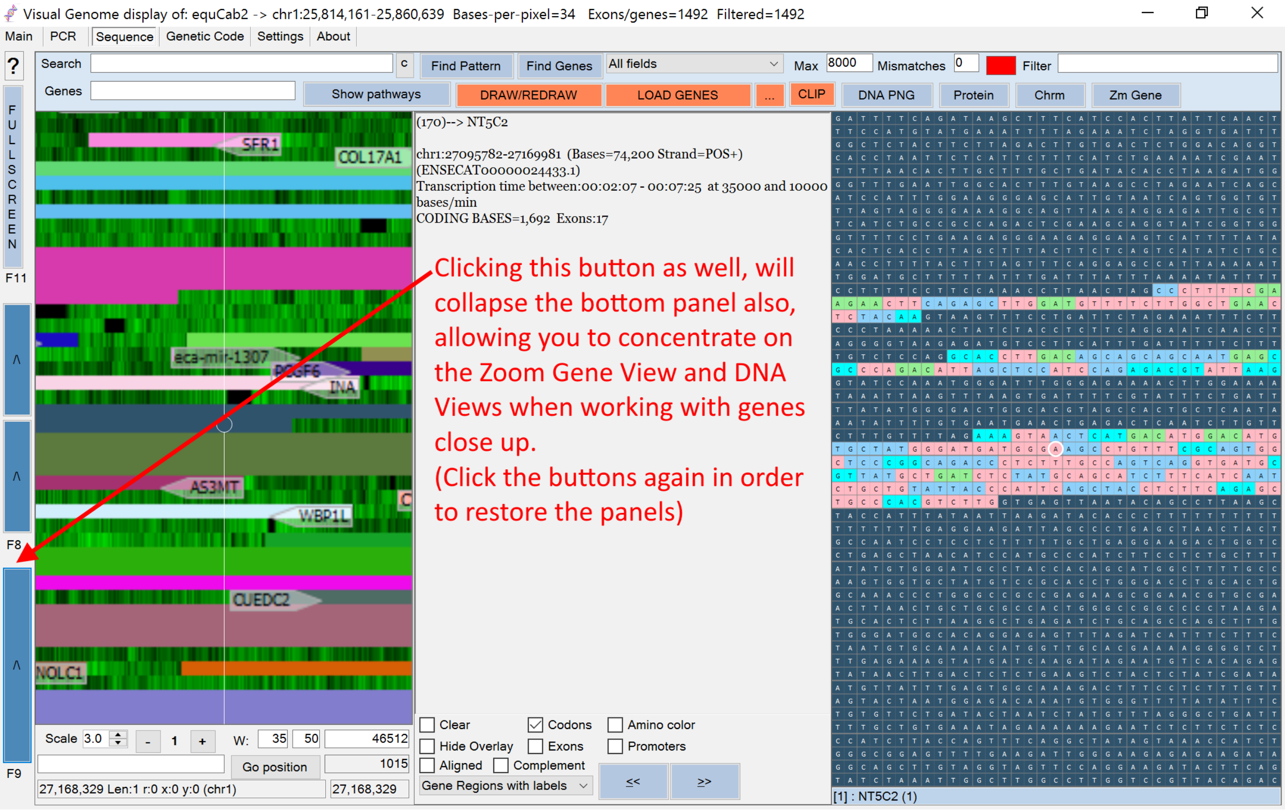
chr1:26,104,147 26,104,147

Max 8000 Mismatches 0

Gene Regions with labels

Clicking this button will collapse the middle panel and expand the bottom view in order that the Main Genome View takes prominence. (Click either middle button to restore again)

Illustration 33: With the BOTTOM PANEL prominent



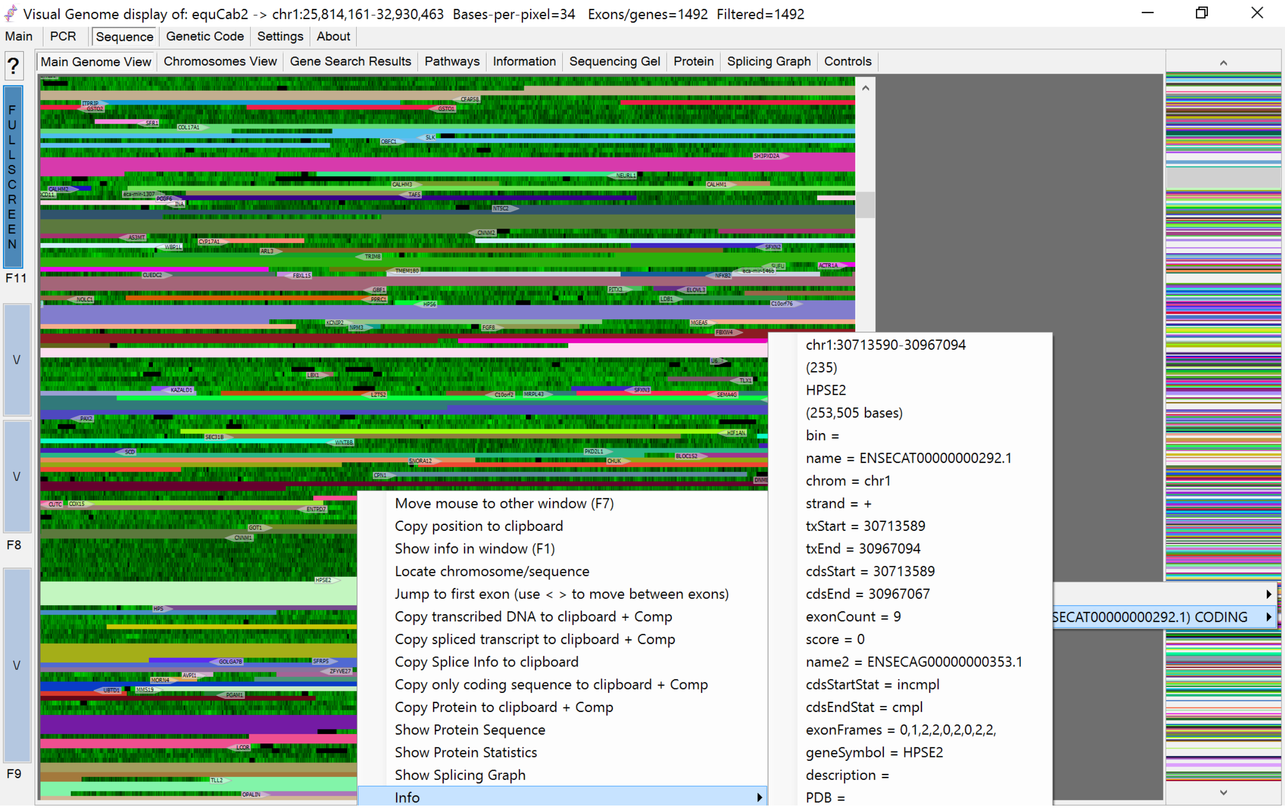
(170)--> NT5C2

chr1:27095782-27169981 (Bases=74,200 Strand=POS+)
(ENSECAT00000024433.1)
Transcription time between:00:02:07 - 00:07:25 at 35000 and 10000 bases/min
CODING BASES=1,692 Exons:17

Clicking this button as well, will collapse the bottom panel also, allowing you to concentrate on the Zoom Gene View and DNA Views when working with genes close up.
(Click the buttons again in order to restore the panels)

Illustration 34: Hiding the BOTTOM PANEL

The Full-screen view allows you to see only the **Main Genome View, Protein View, Splicing View, Sequence Gel, Information, Search results or Pathway tree**. or any of the views obtained from the tabs.

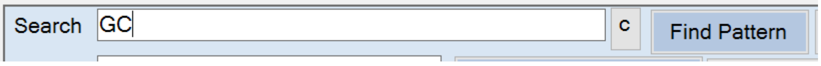

Visual Genome display of: equCab2 -> chr1:25,814,161-32,930,463 Bases-per-pixel=34 Exons/genes=1492 Filtered=1492
Main
PCR
Sequence
Genetic Code
Settings
About
?
Main Genome View
Chromosomes View
Gene Search Results
Pathways
Information
Sequencing Gel
Protein
Splicing Graph
Controls
FULLSCREEN
F11
V
V
F8
V
F9
Move mouse to other window (F7)
Copy position to clipboard
Show info in window (F1)
Locate chromosome/sequence
Jump to first exon (use < > to move between exons)
Copy transcribed DNA to clipboard + Comp
Copy spliced transcript to clipboard + Comp
Copy Splice Info to clipboard
Copy only coding sequence to clipboard + Comp
Copy Protein to clipboard + Comp
Show Protein Sequence
Show Protein Statistics
Show Splicing Graph
Info
SECAT00000000292.1) CODING

chr1:30713590-30967094
(235)
HPSE2
(253,505 bases)
bin =
name = ENSECAT00000000292.1
chrom = chr1
strand = +
txStart = 30713589
txEnd = 30967094
cdsStart = 30713589
cdsEnd = 30967067
exonCount = 9
score = 0
name2 = ENSECAG00000000353.1
cdsStartStat = incmpl
cdsEndStat = cmpl
exonFrames = 0,1,2,2,0,2,0,2,2,
geneSymbol = HPSE2
description =
PDB =

Illustration 35: FULLSCREEN View

CpG Sequences near the start of genes

When you want to look at a specific pattern such as DNA transcription factor binding sites or CpG sites, one can enter a pattern in the **Search field** such as **CG** and it will be highlighted in the **DNA View**. Notice how many **CpG di-nucleotides** can be found near the 5'UTR of the previous gene.


Search GC

The small **C** button can be used to toggle the pattern to its reverse complement.

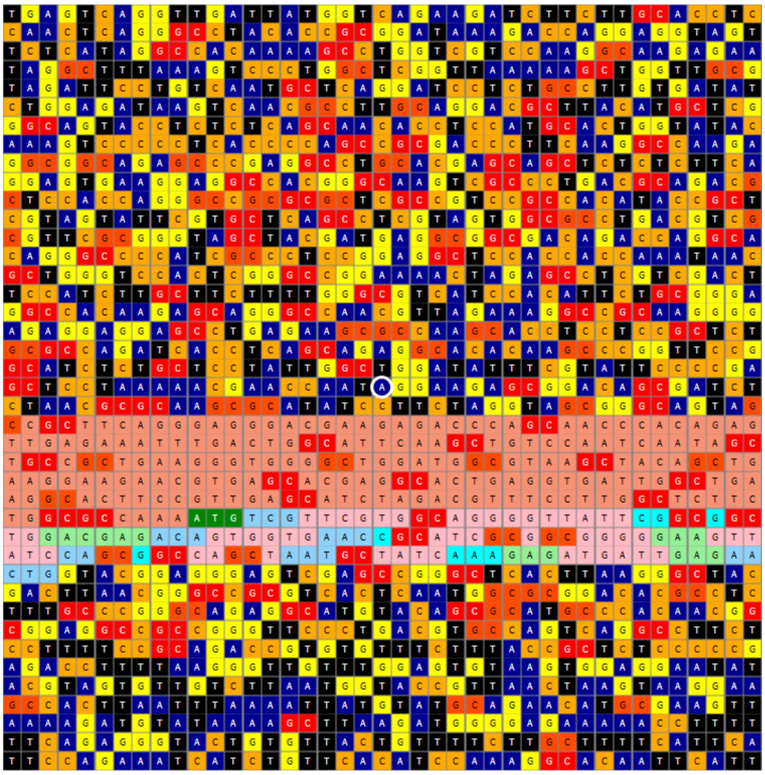


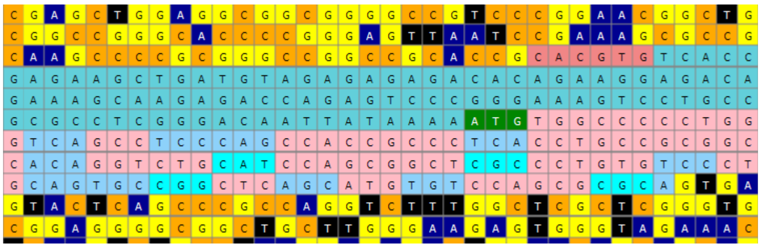
Illustration 36: CHR3:36992431-36994111 near gene MLH1 at width 41

Genes which are constitutively switched on usually have a region of high CpG concentration near the transcription start site, to which transcriptional regulator proteins bind which either prevent RNA Polymerase II from binding or, which recruits epigenetic factors responsible for epigenetically switching off genes by modifying methylation tags on DNA or histones. Genes may also require DNA binding factors and activators to recruit the transcription machinery to the start of transcription.

Promoter and terminator sequences

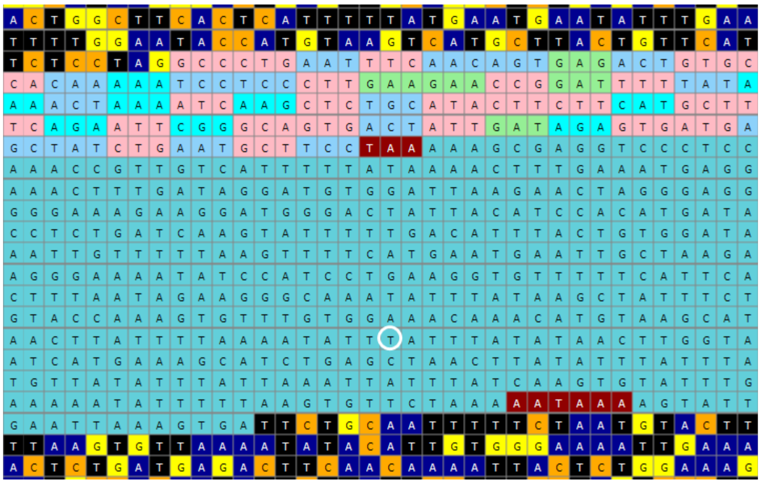
The software is capable of matching specific promoter and terminator sequence of genes in order to investigate transcriptional regulatory elements. Take for example the gene IL12A at chr3:159995404-159996019.

Promoter sequences are highlighted in Magenta while terminator sequences are highlighted in dark red.



```
C G A G C T G G A G G C G G C G G G G C C G T C C C G G A A C G G C T G
C G G C C G G G C A C C C C G G G A G T T A A T C C G A A A G C G C C G
C A A G C C C C G C G G G C C G G C C G C A C C G C A C G T G T C A C C
G A G A A G C T G A T G T A G A G A G A G A C A C A G A A G G A G A C A
G A A A G C A A G A G A C C A G A G T C C C G G G A A A G T C C T G C C
G C G C C T C G G G A C A A T T A T A A A A A T G T G G C C C C C T G G
G T C A G C C T C C C A G C C A C C G C C C T C A C C T G C C G C G G C
C A C A G G T C T G C A T C C A G C G G C T C G C C C T G T G T C C C T
G C A G T G C C G G C T C A G C A T G T G T C C A G C G C G C A G T G A
G T A C T C A G C C C G C C A G G T C T T T G G C T C G C T C G G G T G
C G G A G G G G C G G C T G C T T G G G A A G A G T G G G T A G A A A C
```

Illustration 37: Gene IL12A with CACGTG promoter sequence shown in pink. (chr3:159995404-159996019)



```
A C T G G C T T C A C T C A T T T T T A T G A A T G A A T A T T T G A A
T T T T G G A A T A C C A T G T A A G T C A T G C T T A C T G T T C A T
T C T C C T A G G C C C T G A A T T T C A A C A G T G A G A C T G T G C
C A C A A A A A T C C T C C C T T G A A G A A C C G G A T T T T T A T A
A A A C T A A A A T C A A G C T C T G C A T A C T T C T T C A T G C T T
T C A G A A T T C G G G C A G T G A C T A T T G A T A G A G T G A T G A
G C T A T C T G A A T G C T T C C T A A A A A G C G A G G T C C C T C C
A A A C C G T T G T C A T T T T T A T A A A A C T T T G A A A T G A G G
A A A C T T T G A T A G G A T G T G G A T T A A G A A C T A G G G A G G
G G G A A A G A A G G A T G G G A C T A T T A C A T C C A C A T G A T A
C C T C T G A T C A A G T A T T T T T G A C A T T T A C T G T G G A T A
A A T T G T T T T T A A G T T T T C A T G A A T G A A T T G C T A A G A
A G G G A A A A T A T C C A T C C T G A A G G T G T T T T T C A T T C A
C T T T A A T A G A A G G G C A A A T A T T T A T A A G C T A T T T C T
G T A C C A A A G T G T T T G T G G A A A C A A A C A T G T A A G C A T
A A C T T A T T T T A A A A T A T T T A T T T A T A T A A C T T G G T A
A T C A T G A A A G C A T C T G A G C T A A C T T A T A T T T A T T T A
T G T T A T A T T T A T T A A A T T A T T T A T C A A G T G T A T T T G
A A A A A T A T T T T T A A G T G T T C T A A A A A T A A A A G T A T T
G A A T T A A A G T G A T T C T G C A A T T T T T C T A A T G T A C T T
T T A A G T G T T A A A A T A T A C A T T G T G G G A A A A T T G A A A
A C T C T G A T G A G A C T T C A A C A A A A T T A C T C T G G A A A G
```

Illustration 38: The terminator sequence AATAAA at the end of the 3' UTR of gene IL12A

Chromosomes View

When one select the **Chromosomes View**, a birds-eye view is provided of exactly where on all the chromosomes the search results for the search term: “**dna mismatch repair**” can be found. The currently displayed selection is additionally highlighted with a moving rectangle display as shown in the picture below.

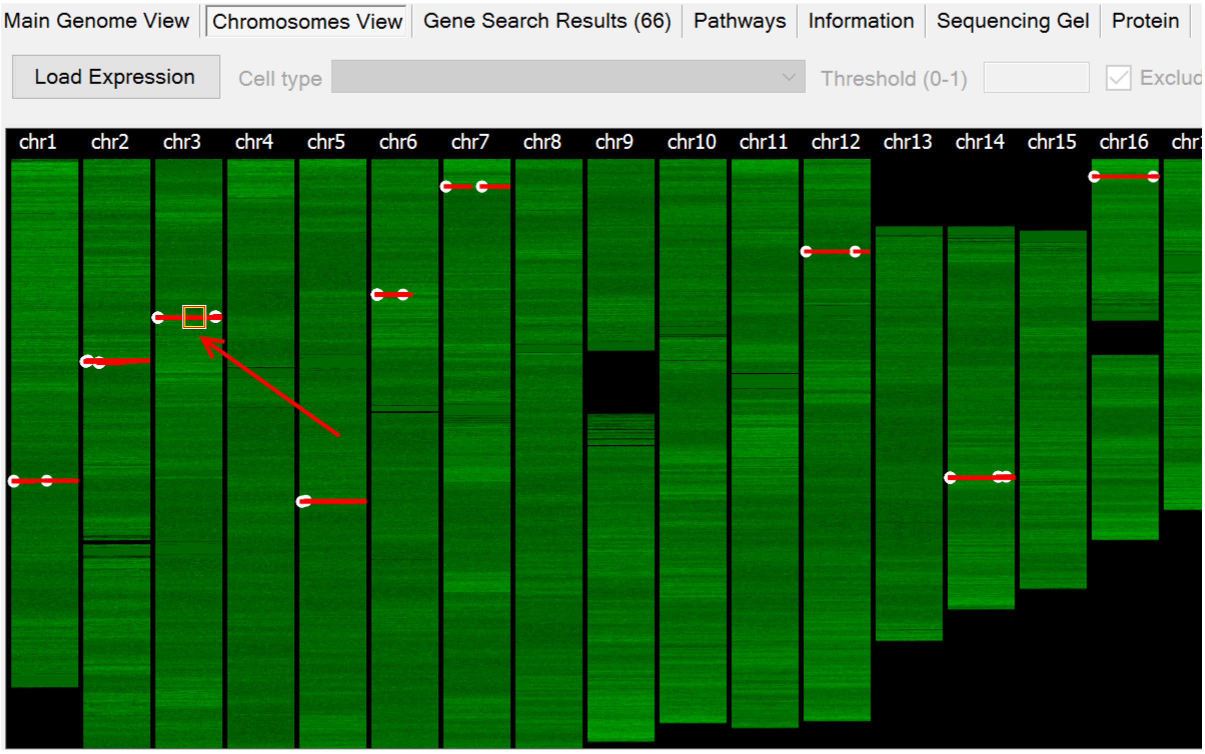


Illustration 39: Chromosomes View (providing a true birds-eye view)

Again, hovering over the **Chromosomes View** above will display the corresponding region in the **DNA View** and ***clicking on the view***, will navigate down to the specific region and swap back to the **Main Genome View**.

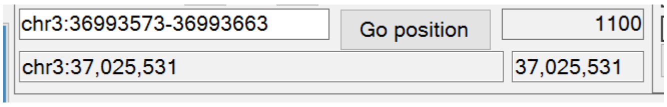
This means one is able to drill down from the **”Big Picture”** (**Chromosomes View**) – to –

the **Main Genome View** – to –>

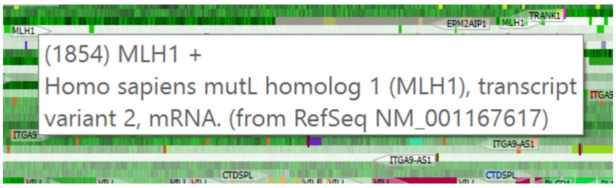
the **DNA View** – to –>

the **Molecule View** (where a comic depiction of the molecules in question is shown)

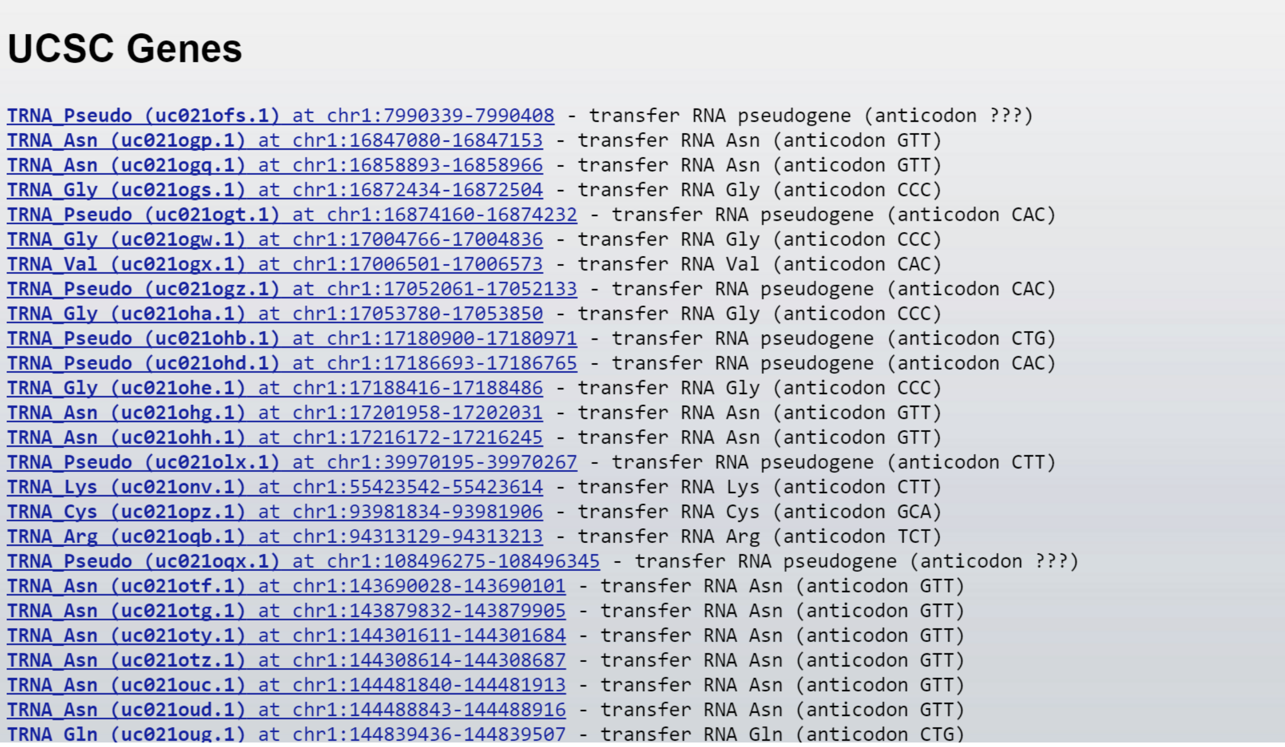
When the mouse pointer is moved across the **Main Genome View**, the chromosomal position is shown (which can be copied and pasted in the online UCSC genome browser)


chr3:36993573-36993663
chr3:37,025,531
37,025,531
1100

One can also enter coordinates and then click on “**Go position**” to jump to a specific sequence position. An info tooltip is provided as you hover across the genes in the **Main Genome View**.


(1854) MLH1 +
Homo sapiens mutL homolog 1 (MLH1), transcript variant 2, mRNA. (from RefSeq NM_001167617)

Normally, the **Chromosomes View** would be used to provide you with a quick way to see where the genes associated with a specific pathway is located across all of the chromosomes, or to display gene search results obtained from the UCSC genome browser. As an example, lets do a search for all the tRNA's found in the human genome. There are more than 500 tRNA's encoded in the human genome, corresponding to 48 anti-codons. By copying the search results from the UCSC browser and clicking the **CLIP** button, the clipboard contents is parsed in order to plot all the tRNAs at the appropriate positions on the HG19 (in this case) version of the human genome sequence.



UCSC Genes

```
TRNA_Pseudo (uc021ofs.1) at chr1:7990339-7990408 - transfer RNA pseudogene (anticodon ???)
TRNA_Asn (uc021ogp.1) at chr1:16847080-16847153 - transfer RNA Asn (anticodon GTT)
TRNA_Asn (uc021ogq.1) at chr1:16858893-16858966 - transfer RNA Asn (anticodon GTT)
TRNA_Gly (uc021ogs.1) at chr1:16872434-16872504 - transfer RNA Gly (anticodon CCC)
TRNA_Pseudo (uc021ogt.1) at chr1:16874160-16874232 - transfer RNA pseudogene (anticodon CAC)
TRNA_Gly (uc021ogw.1) at chr1:17004766-17004836 - transfer RNA Gly (anticodon CCC)
TRNA_Val (uc021ogx.1) at chr1:17006501-17006573 - transfer RNA Val (anticodon CAC)
TRNA_Pseudo (uc021ogz.1) at chr1:17052061-17052133 - transfer RNA pseudogene (anticodon CAC)
TRNA_Gly (uc021oha.1) at chr1:17053780-17053850 - transfer RNA Gly (anticodon CCC)
TRNA_Pseudo (uc021ohb.1) at chr1:17180900-17180971 - transfer RNA pseudogene (anticodon CTG)
TRNA_Pseudo (uc021ohd.1) at chr1:17186693-17186765 - transfer RNA pseudogene (anticodon CAC)
TRNA_Gly (uc021ohe.1) at chr1:17188416-17188486 - transfer RNA Gly (anticodon CCC)
TRNA_Asn (uc021ohg.1) at chr1:17201958-17202031 - transfer RNA Asn (anticodon GTT)
TRNA_Asn (uc021ohh.1) at chr1:17216172-17216245 - transfer RNA Asn (anticodon GTT)
TRNA_Pseudo (uc021olx.1) at chr1:39970195-39970267 - transfer RNA pseudogene (anticodon CTT)
TRNA_Lys (uc021onv.1) at chr1:55423542-55423614 - transfer RNA Lys (anticodon CTT)
TRNA_Cys (uc021opz.1) at chr1:93981834-93981906 - transfer RNA Cys (anticodon GCA)
TRNA_Arg (uc021oqb.1) at chr1:94313129-94313213 - transfer RNA Arg (anticodon TCT)
TRNA_Pseudo (uc021oqx.1) at chr1:108496275-108496345 - transfer RNA pseudogene (anticodon ???)
TRNA_Asn (uc021otf.1) at chr1:143690028-143690101 - transfer RNA Asn (anticodon GTT)
TRNA_Asn (uc021otg.1) at chr1:143879832-143879905 - transfer RNA Asn (anticodon GTT)
TRNA_Asn (uc021oty.1) at chr1:144301611-144301684 - transfer RNA Asn (anticodon GTT)
TRNA_Asn (uc021otz.1) at chr1:144308614-144308687 - transfer RNA Asn (anticodon GTT)
TRNA_Asn (uc021ouc.1) at chr1:144481840-144481913 - transfer RNA Asn (anticodon GTT)
TRNA_Asn (uc021oud.1) at chr1:144488843-144488916 - transfer RNA Asn (anticodon GTT)
TRNA_Gln (uc021oug.1) at chr1:144839436-144839507 - transfer RNA Gln (anticodon CTG)
```

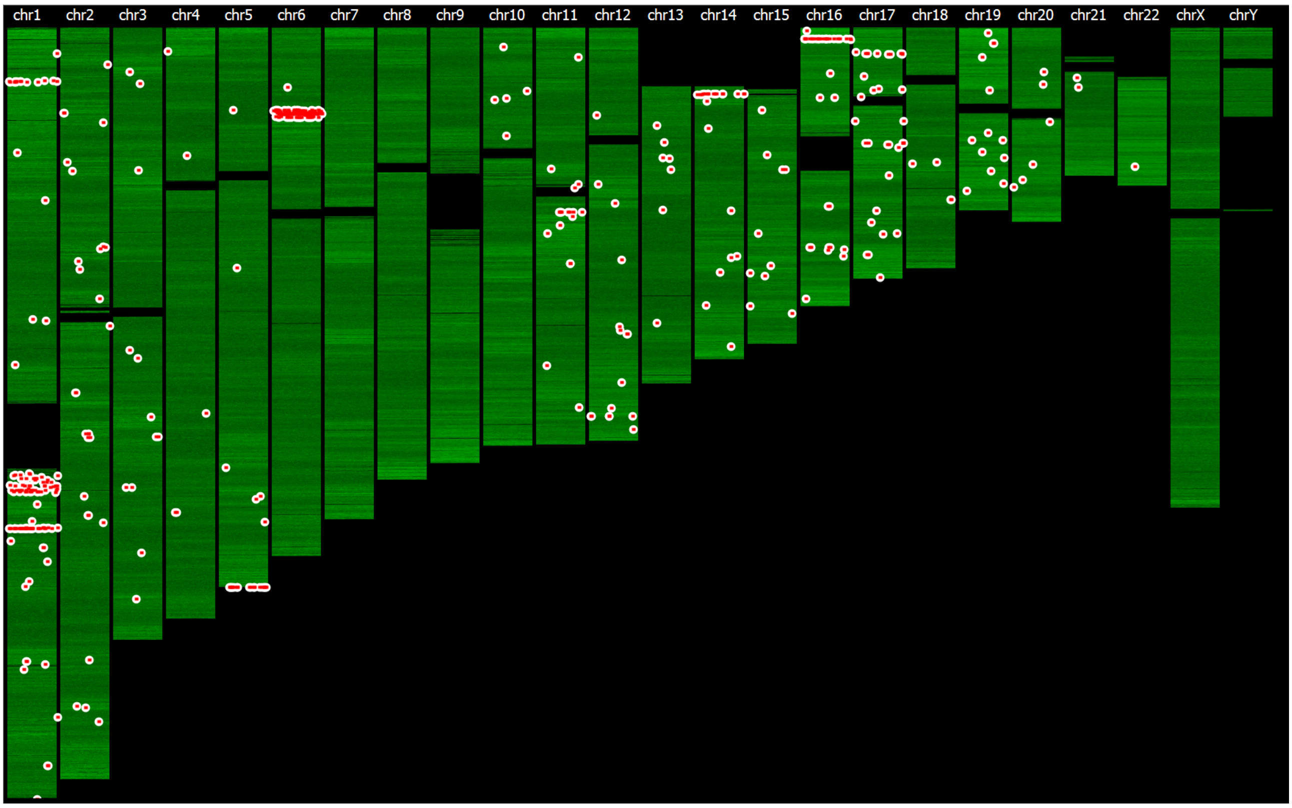
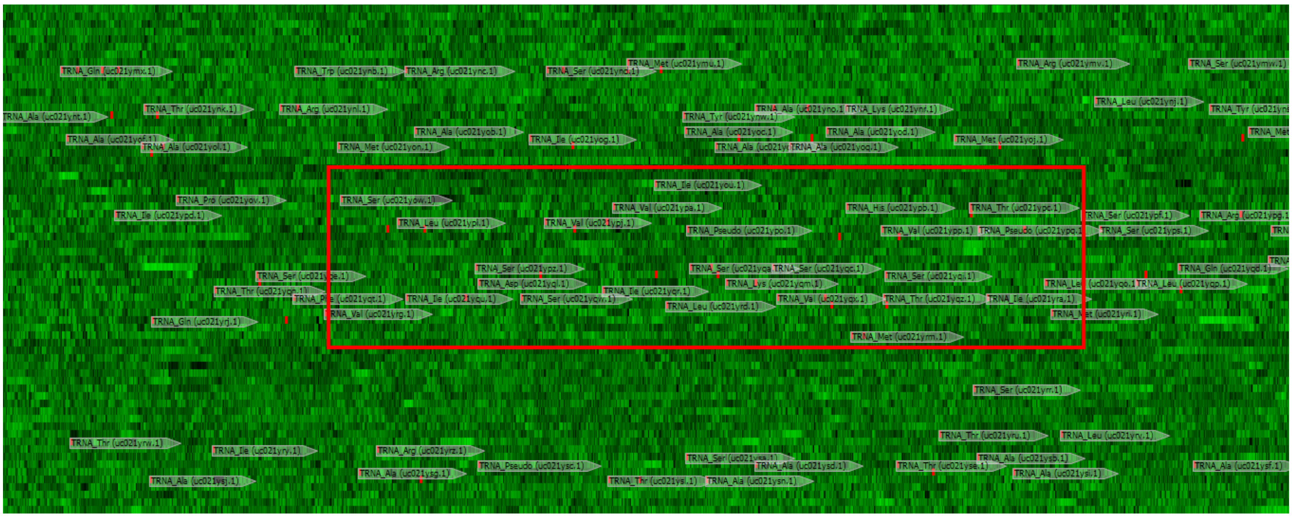


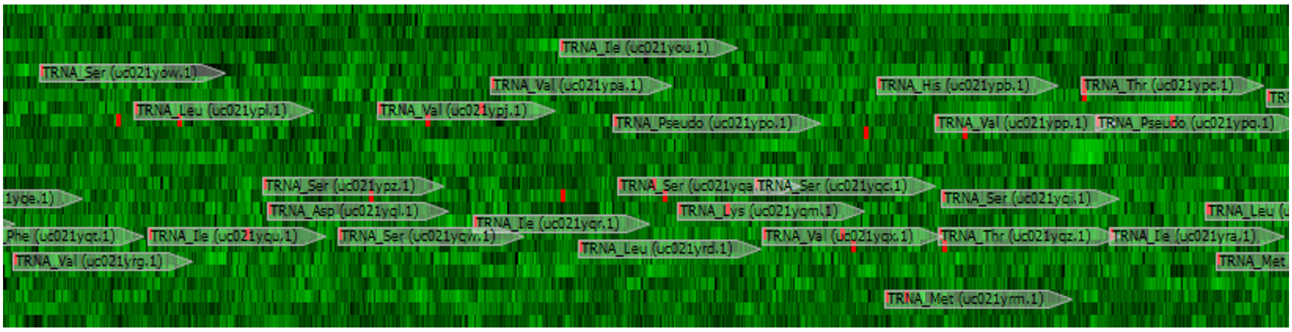
Illustration 40: Chromosomes view showing all the positions of tRNAs in the genome

The Main Genome View

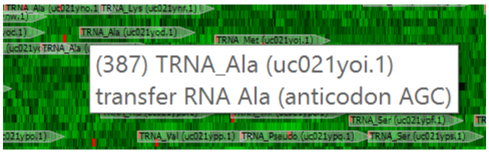
When you now click on the band of red dots on chromosome 6, you can see their distribution clearer.



With the highlighted area magnified...



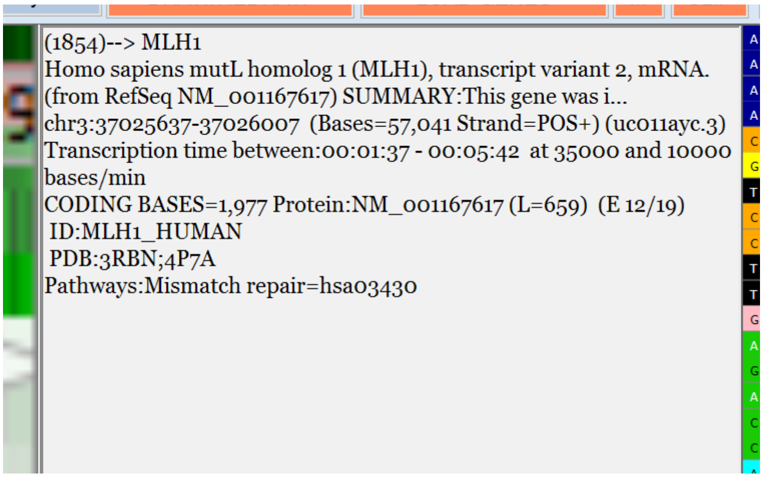
Hovering the mouse over any of these genes will bring up a tool tip:



(387) TRNA_Ala (uc021yoi.1)
transfer RNA Ala (anticodon AGC)

The Information Display window

The **Information Display** in the top centre displays relevant information for the specific gene.

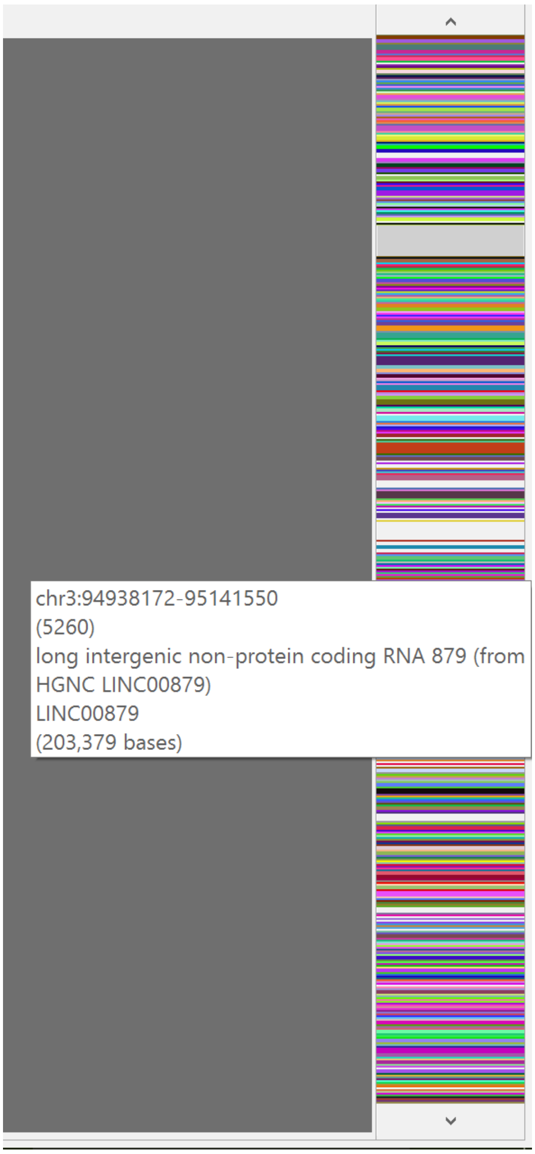


(1854)--> MLH1
Homo sapiens mutL homolog 1 (MLH1), transcript variant 2, mRNA.
(from RefSeq NM_001167617) SUMMARY:This gene was i...
chr3:37025637-37026007 (Bases=57,041 Strand=POS+) (uc011ayc.3)
Transcription time between:00:01:37 - 00:05:42 at 35000 and 10000 bases/min
CODING BASES=1,977 Protein:NM_001167617 (L=659) (E 12/19)
ID:MLH1_HUMAN
PDB:3RBN;4P7A
Pathways:Mismatch repair=hsa03430

Illustration 41: Information Display

The enhanced Chromosome scrollbar

The chromosome scrollbar on the right provides another quick way of finding gene annotations.

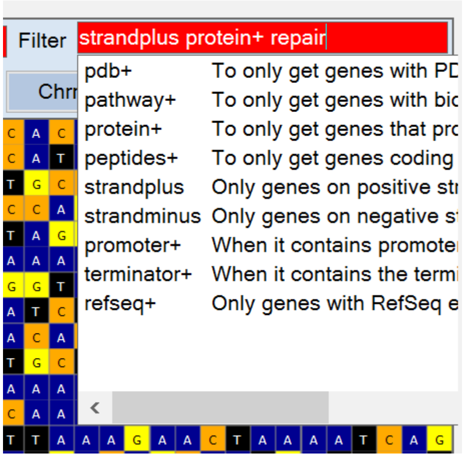


chr3:94938172-95141550
(5260)
long intergenic non-protein coding RNA 879 (from HGNC LINC00879)
LINC00879
(203,379 bases)

Illustration 42: Chromosome scrollbar. Hovering provides an info tooltip

Filtering display results

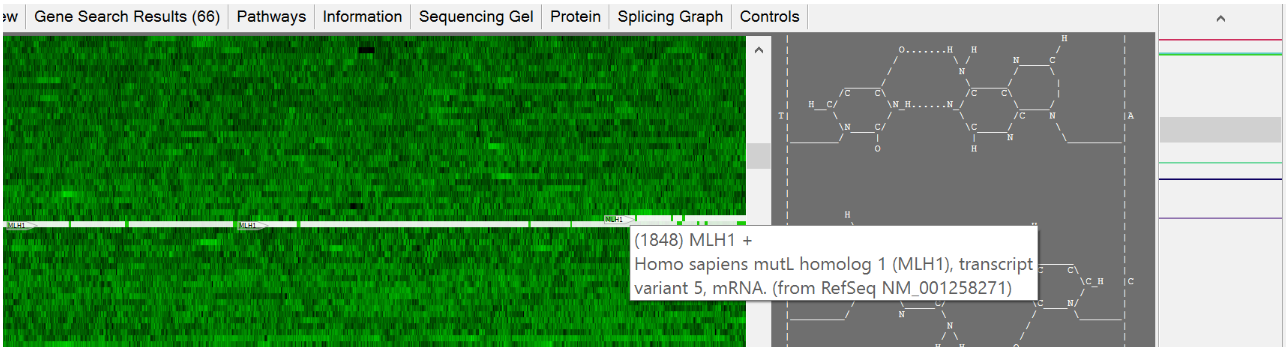
Probably one of the **most useful** features of the Visual Genome Browser is the ability to filter the amount of annotations displayed. This can be done via the **Filter Input** field at the top right.



Filter strandplus protein+ repair

Chr
pdb+ To only get genes with PD
pathway+ To only get genes with bio
protein+ To only get genes that pro
peptides+ To only get genes coding
strandplus Only genes on positive str
strandminus Only genes on negative s
promoter+ When it contains promote
terminator+ When it contains the termi
refseq+ Only genes with RefSeq e

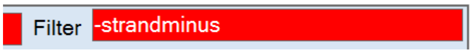
Illustration 43: Filter Input field



(1848) MLH1 +
Homo sapiens mutL homolog 1 (MLH1), transcript variant 5, mRNA. (from RefSeq NM_001258271)

When you click "DRAW/REDRAW", **all** browser views are updated to reflect this new filter criteria, including the scrollbar. This makes it much easier to see the **forest for the trees (genes)**.

One can also **filter out** certain genes, by adding a minus in front of the terms in the filter:



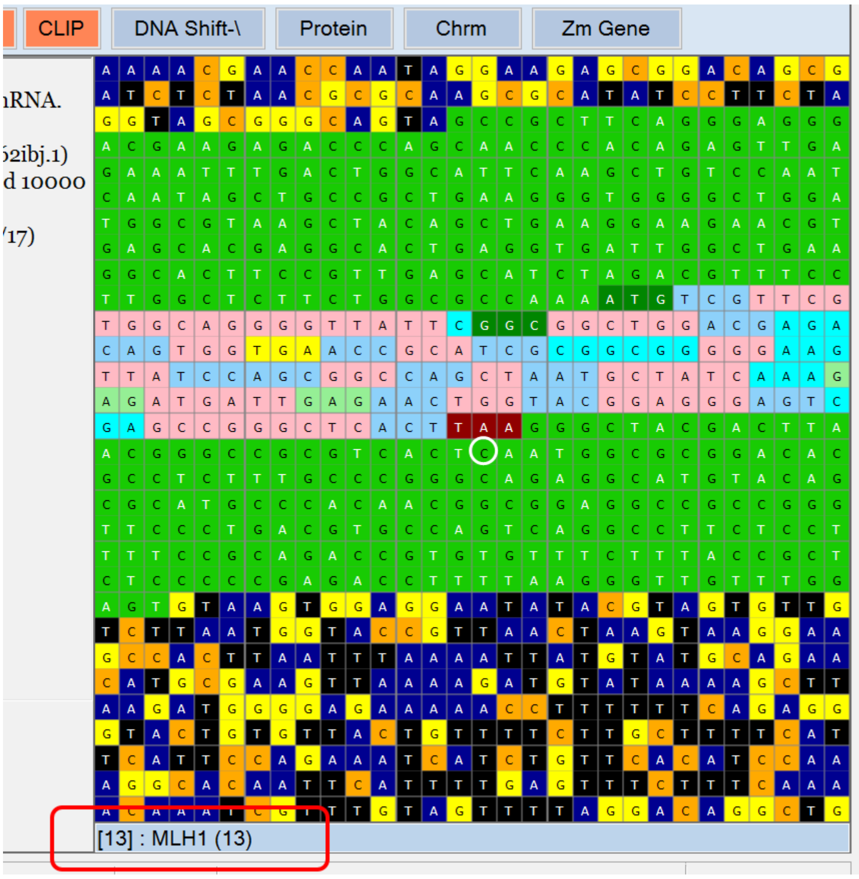
Filter -strandminus

Entering **–strandminus** will filter out all genes on the negative strand of the chromosome.

Entering something like **–protein+** (will filter out any genes that are coding for proteins and only display non-coding genes), because **protein+** is the search term for genes which has an ORF or coding sequence.

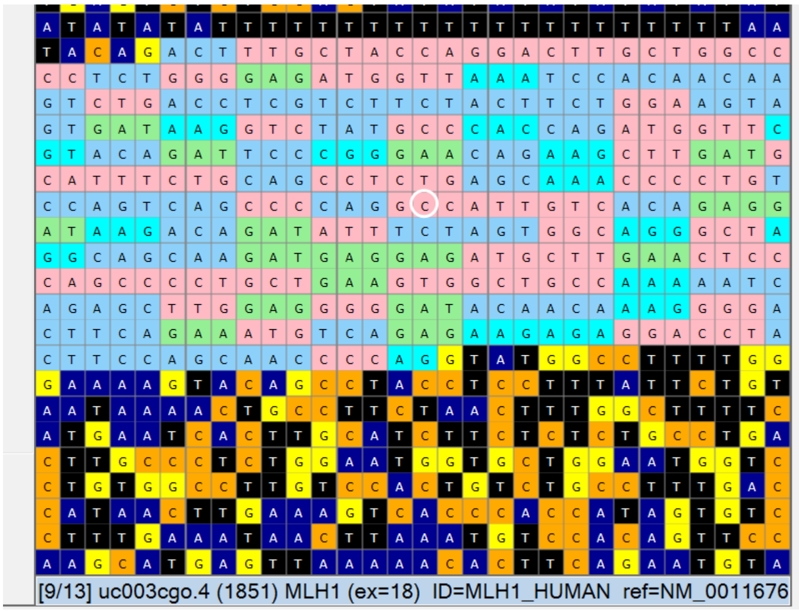
Examining genes using the different views

Because there are multiple overlapping transcripts at most positions, it is possible to use the **+/-** buttons to filter the **DNA View** in order to only see a specific splice variant.



[13] : MLH1 (13)

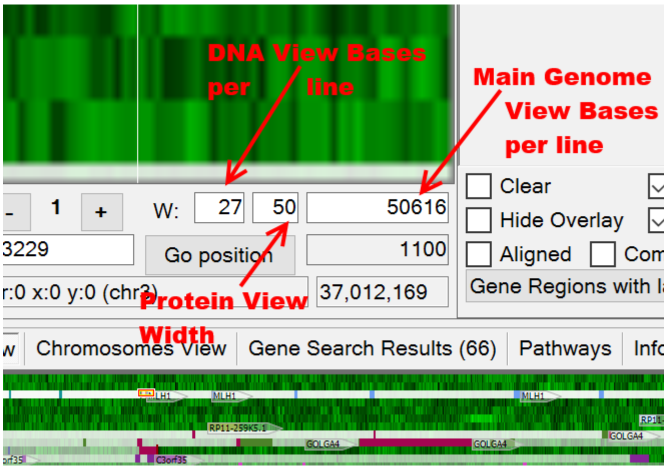
Illustration 44: Use plus/minus to select specific overlapping gene at same position.



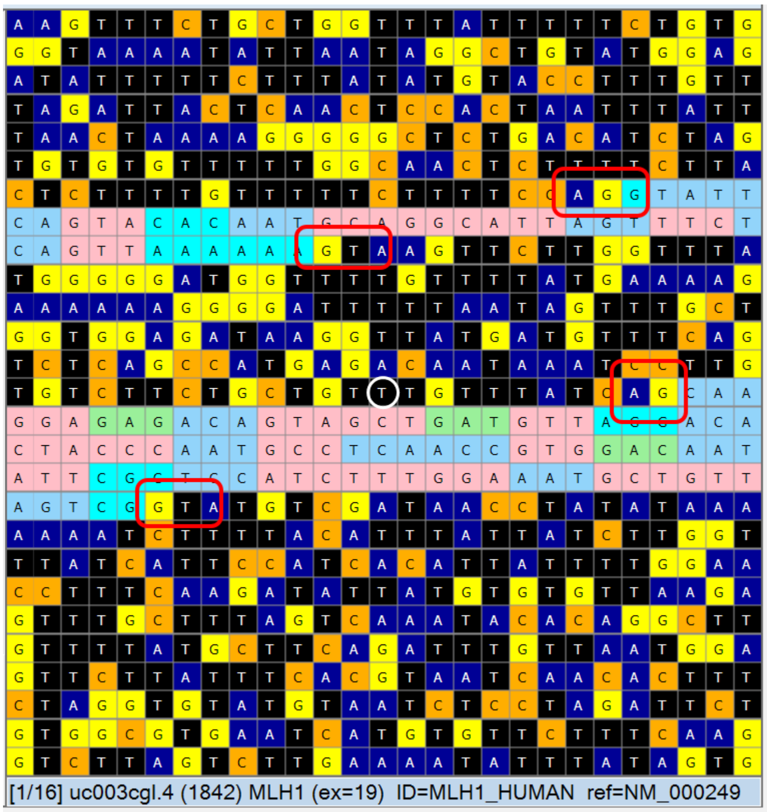
[9/13] uc003cgo.4 (1851) MLH1 (ex=18) ID=MLH1_HUMAN ref=NM_0011676

Illustration 45: Different transcript selected

You can change the number of bases per line in the display by pressing **[or]** or by changing the width in the text field:



The **DNA View**'s number of bases per line is correspondingly altered. When the **Codons** checkbox is selected, the coding sequences are displayed using colours which correspond to the hydrophobic/polar character of the amino acids coded by these codons. In the picture below one can also observe the bases responsible for demarcating the start and ends of introns which are removed by the Spliceosome when it produces the mature RNA transcript.

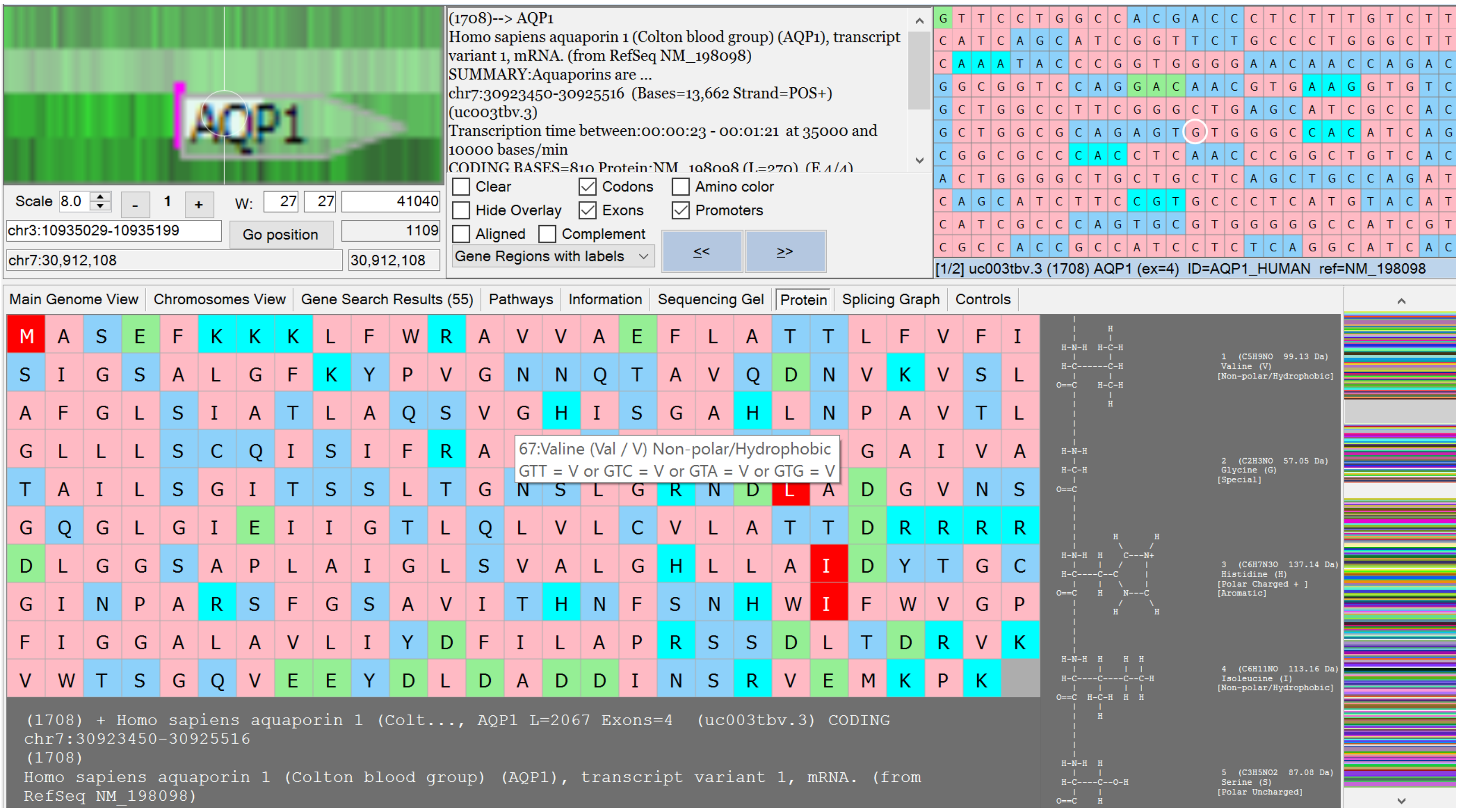


```
A A G T T T C T G C T G G T T T A T T T T T C T G T G
G G T A A A A T A T T A A T A G G C T G T A T G G A G
A T A T T T T T C T T T A T A T G T A C C T T T G T T
T A G A T T A C T C A A C T C C A C T A A T T T A T T
T A A C T A A A A G G G G G C T C T G A C A T C T A G
T G T G T G T T T T T G G C A A C T C T T T T C T T A
C T C T T T T G T T T T T C T T T T C C A G G T A T T
C A G T A C A C A A T G C A G G C A T T A G T T T C T
C A G T T A A A A A A G T A A G T T C T T G G T T T A
T G G G G G A T G G T T T T G T T T T A T G A A A A G
A A A A A A G G G G A T T T T T A A T A G T T T G C T
G G T G G A G A T A A G G T T A T G A T G T T T C A G
T C T C A G C C A T G A G A C A A T A A A T C C T T G
T G T C T T C T G C T G T T T G T T T A T C A G C A A
G G A G A G A C A G T A G C T G A T G T T A G G A C A
C T A C C C A A T G C C T C A A C C G T G G A C A A T
A T T C G C T C C A T C T T T G G A A A T G C T G T T
A G T C G G T A T G T C G A T A A C C T A T A T A A A
A A A A T C T T T T A C A T T T A T T A T C T T G G T
T T A T C A T T C C A T C A C A T T A T T T T G G A A
C C T T T C A A G A T A T T A T G T G T G T T A A G A
G T T T G C T T T A G T C A A A T A C A C A G G C T T
G T T T T A T G C T T C A G A T T T G T T A A T G G A
G T T C T T A T T T C A C G T A A T C A A C A C T T T
C T A G G T G T A T G T A A T C T C C T A G A T T C T
G T G G C G T G A A T C A T G T G T T C T T T C A A G
G T C T T A G T C T T G A A A A T A T T T A T A G T G
[1/16] uc003cgl.4 (1842) MLH1 (ex=19)  ID=MLH1_HUMAN  ref=NM_000249
```

Illustration 46: DNA View with Splice signal bases highlighted

Codons which are **pink** code for amino acids which are hydrophobic and tend to be on the interior of the resultant folded protein or embedded in lipid membranes, whilst amino acids which are polar (from the **blue and green** codons) tend to be on the exterior of the folded protein. The colour of the codons reveals much about the properties of a protein. We would therefore expect that a protein like **Aquaporin** (which is a transmembrane protein used to transport water molecules across the lipid membranes), would consist of a lot of hydrophobic amino acids on its outside, with polar amino acids on its interior to interact with the polar water molecules.

This can be seen in the following **Protein Sequence View**, which is obtained by **double-clicking** on a gene in the **Main Genome View** or by selecting "**Show Protein Sequence**" in the context menu.



Scale 8.0 - 1 + W: 27 27 41040

chr3:10935029-10935199 Go position 1109

chr7:30,912,108 30,912,108

(1708)--> AQP1
Homo sapiens aquaporin 1 (Colton blood group) (AQP1), transcript variant 1, mRNA. (from RefSeq NM_198098)
SUMMARY:Aquaporins are ...
chr7:30923450-30925516 (Bases=13,662 Strand=POS+)
(uc003tbv.3)
Transcription time between:00:00:23 - 00:01:21 at 35000 and 10000 bases/min
CODING BASES=810 Protein:NM_198098 (L=270) (E 4/4)

☐ Clear ☑ Codons ☐ Amino color
☐ Hide Overlay ☑ Exons ☑ Promoters
☐ Aligned ☐ Complement
Gene Regions with labels

G	T	T	C	C	T	G	G	C	C	A	C	G	A	C	C	C	T	C	T	T	T	G	T	C	T	T
C	A	T	C	A	G	C	A	T	C	G	G	T	T	C	T	G	C	C	C	T	G	G	G	C	T	T
C	A	A	A	T	A	C	C	C	G	G	T	G	G	G	G	A	A	C	A	A	C	C	A	G	A	C
G	G	C	G	G	T	C	C	A	G	G	A	C	A	A	C	G	T	G	A	A	G	G	T	G	T	C
G	C	T	G	G	C	C	T	T	C	G	G	G	C	T	G	A	G	C	A	T	C	G	C	C	A	C
G	C	T	G	G	C	G	C	A	G	A	G	T	G	T	G	G	G	C	C	A	C	A	T	C	A	G
C	G	G	C	G	C	C	C	A	C	C	T	C	A	A	C	C	C	G	G	C	T	G	T	C	A	C
A	C	T	G	G	G	G	C	T	G	C	T	G	C	T	C	A	G	C	T	G	C	C	A	G	A	T
C	A	G	C	A	T	C	T	T	C	C	G	T	G	C	C	C	T	C	A	T	G	T	A	C	A	T
C	A	T	C	G	C	C	C	A	G	T	G	C	G	T	G	G	G	G	G	C	C	A	T	C	G	T
C	G	C	C	A	C	C	G	C	C	A	T	C	C	T	C	T	C	A	G	G	C	A	T	C	A	C

[1/2] uc003tbv.3 (1708) AQP1 (ex=4) ID=AQP1_HUMAN ref=NM_198098

M	A	S	E	F	K	K	K	L	F	W	R	A	V	V	A	E	F	L	A	T	T	L	F	V	F	I
S	I	G	S	A	L	G	F	K	Y	P	V	G	N	N	Q	T	A	V	Q	D	N	V	K	V	S	L
A	F	G	L	S	I	A	T	L	A	Q	S	V	G	H	I	S	G	A	H	L	N	P	A	V	T	L
G	L	L	L	S	C	Q	I	S	I	F	R	A	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	G	A	I	V	A
T	A	I	L	S	G	I	T	S	S	L	T	G	N	S	L	G	R	N	D	L	A	D	G	V	N	S
G	Q	G	L	G	I	E	I	I	G	T	L	Q	L	V	L	C	V	L	A	T	T	D	R	R	R	R
D	L	G	G	S	A	P	L	A	I	G	L	S	V	A	L	G	H	L	L	A	I	D	Y	T	G	C
G	I	N	P	A	R	S	F	G	S	A	V	I	T	H	N	F	S	N	H	W	I	F	W	V	G	P
F	I	G	G	A	L	A	V	L	I	Y	D	F	I	L	A	P	R	S	S	D	L	T	D	R	V	K
V	W	T	S	G	Q	V	E	E	Y	D	L	D	A	D	D	I	N	S	R	V	E	M	K	P	K	

67:Valine (Val / V) Non-polar/Hydrophobic
GTT = V or GTC = V or GTA = V or GTG = V

```
(1708) + Homo sapiens aquaporin 1 (Colt..., AQP1 L=2067 Exons=4  (uc003tbv.3) CODING
chr7:30923450-30925516
(1708)
Homo sapiens aquaporin 1 (Colton blood group) (AQP1), transcript variant 1, mRNA. (from
RefSeq NM_198098)
```

```
      |       H
      |       |
  H-N-H  H-C-H
      |       |                     1  (C5H9NO  99.13 Da)
  H-C-------C-H                     Valine (V)
      |       |                    [Non-polar/Hydrophobic]
 O==C    H-C-H
      |       |
      |       H
      |
  H-N-H
      |                             2  (C2H3NO  57.05 Da)
  H-C-H                             Glycine (G)
      |                            [Special]
 O==C
      |
      |       H       H
      |        \     /
  H-N-H  H    C---N+
      |  |   /     |                3  (C6H7N3O  137.14 Da)
  H-C----C--C      |                Histidine (H)
      |  |   \     |               [Polar Charged + ]
 O==C    H    N---C                [Aromatic]
      |      /     \
      |     H       H
      |
  H-N-H  H    H  H
      |  |    |  |                  4  (C6H11NO  113.16 Da)
  H-C----C----C--C-H                Isoleucine (I)
      |  |    |  |                 [Non-polar/Hydrophobic]
 O==C  H-C-H  H  H
      |     |
      |     H
      |
  H-N-H  H
      |  |                          5  (C3H5NO2  87.08 Da)
  H-C----C--O-H                     Serine (S)
      |  |                         [Polar Uncharged]
 O==C    H
```

By changing the number of amino acids in each row, it is possible to line up the amino acids at the start of each exon, indicated by red blocks in the protein view. It is interesting how similar amino acids line up vertically (almost like a crossword puzzle). Also notice how many hydrophobic amino acids can be found in the protein.

Again, notice how the displays are linked. When you hover the mouse on the **Protein View**, the **DNA View** will navigate to the corresponding codon.

Non-polar Hydrophobic

Polar Uncharged

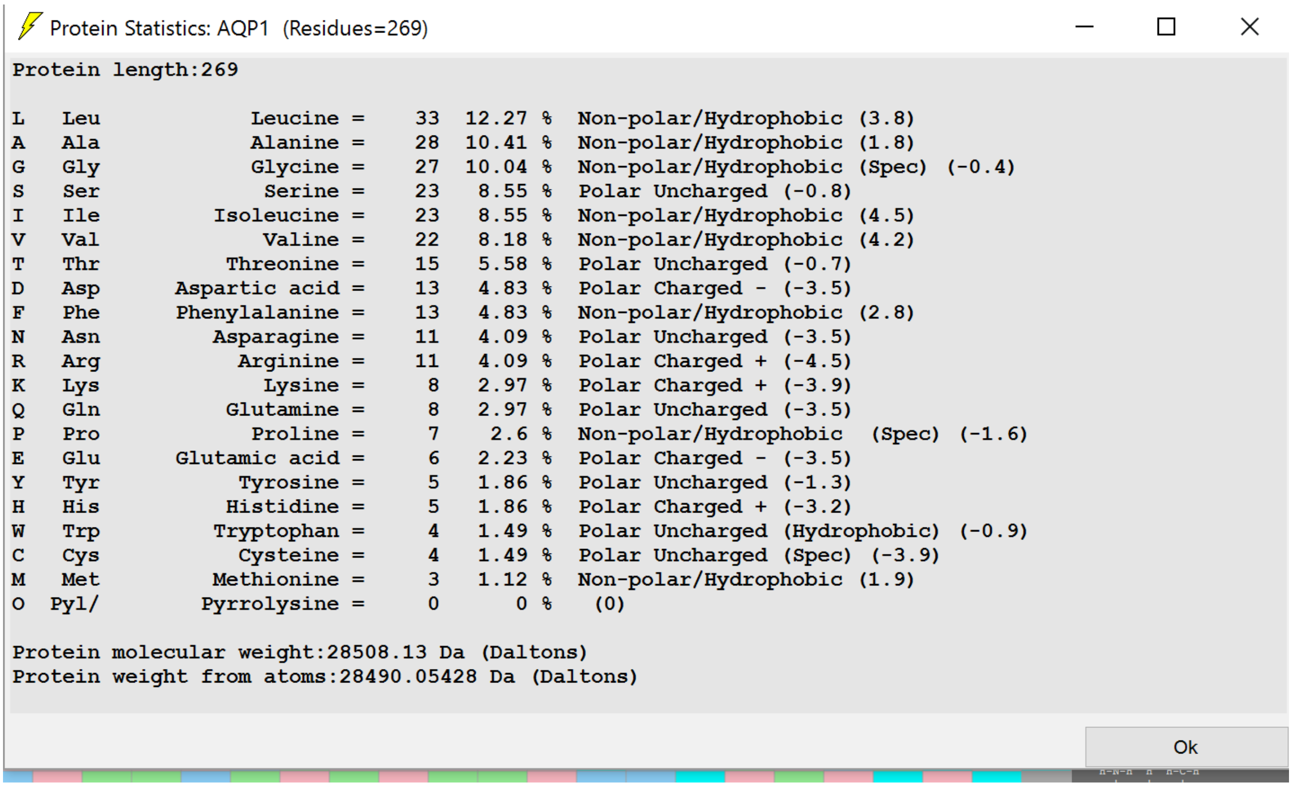
Polar Charged +

Polar Charged -

Stop Codons

Protein Statistics

You can obtain the gene context menu by right clicking on the protein view. If you select "**Show Protein Statistics**" you will see why this membrane protein is considered a mostly hydrophobic protein.



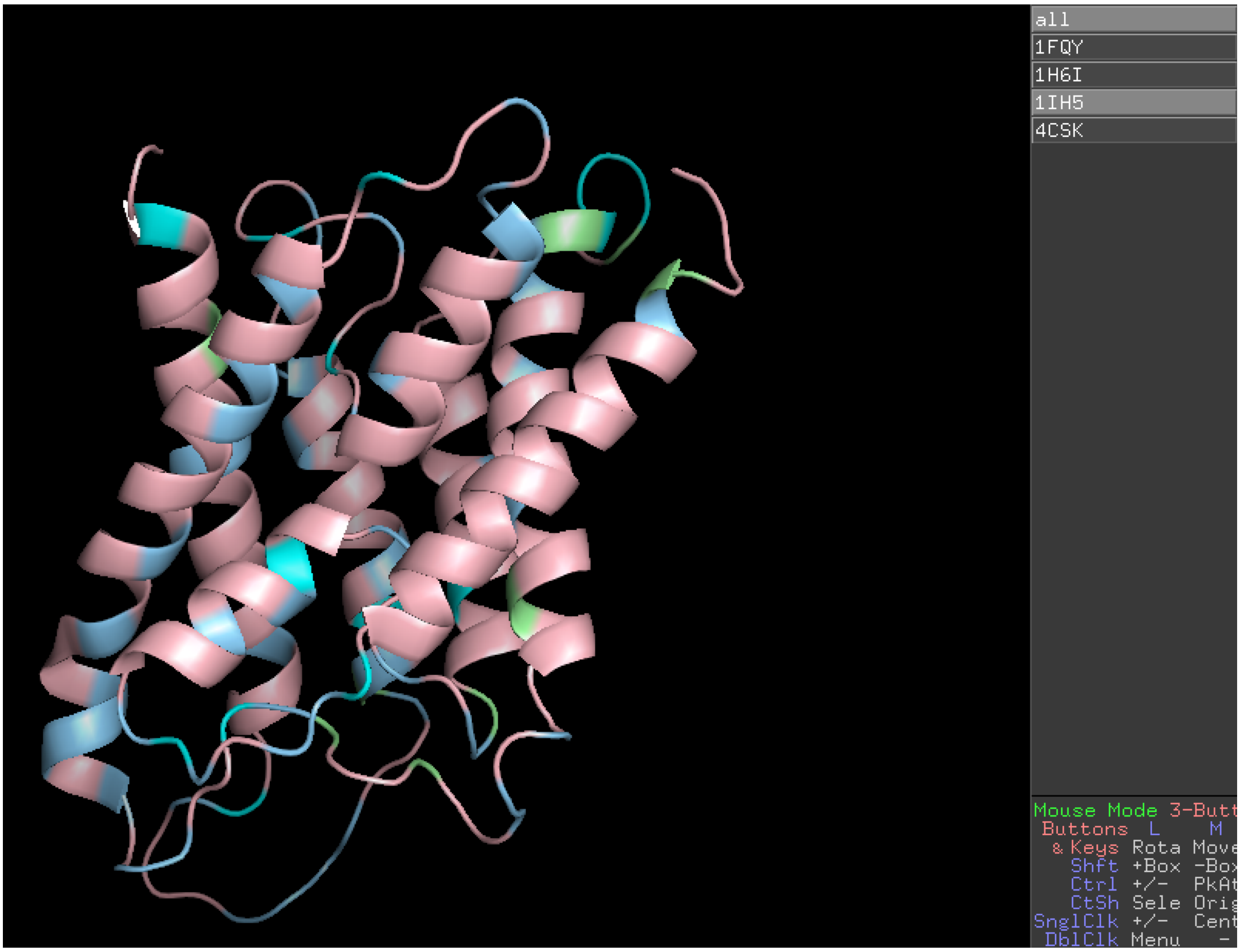
```
Protein length:269

L   Leu           Leucine =    33  12.27 %  Non-polar/Hydrophobic (3.8)
A   Ala           Alanine =    28  10.41 %  Non-polar/Hydrophobic (1.8)
G   Gly           Glycine =    27  10.04 %  Non-polar/Hydrophobic (Spec)  (-0.4)
S   Ser            Serine =    23   8.55 %  Polar Uncharged (-0.8)
I   Ile        Isoleucine =    23   8.55 %  Non-polar/Hydrophobic (4.5)
V   Val            Valine =    22   8.18 %  Non-polar/Hydrophobic (4.2)
T   Thr         Threonine =    15   5.58 %  Polar Uncharged (-0.7)
D   Asp     Aspartic acid =    13   4.83 %  Polar Charged - (-3.5)
F   Phe     Phenylalanine =    13   4.83 %  Non-polar/Hydrophobic (2.8)
N   Asn        Asparagine =    11   4.09 %  Polar Uncharged (-3.5)
R   Arg          Arginine =    11   4.09 %  Polar Charged + (-4.5)
K   Lys            Lysine =     8   2.97 %  Polar Charged + (-3.9)
Q   Gln         Glutamine =     8   2.97 %  Polar Uncharged (-3.5)
P   Pro           Proline =     7    2.6 %  Non-polar/Hydrophobic  (Spec)  (-1.6)
E   Glu     Glutamic acid =     6   2.23 %  Polar Charged - (-3.5)
Y   Tyr          Tyrosine =     5   1.86 %  Polar Uncharged (-1.3)
H   His         Histidine =     5   1.86 %  Polar Charged + (-3.2)
W   Trp        Tryptophan =     4   1.49 %  Polar Uncharged (Hydrophobic)  (-0.9)
C   Cys          Cysteine =     4   1.49 %  Polar Uncharged (Spec)  (-3.9)
M   Met        Methionine =     3   1.12 %  Non-polar/Hydrophobic (1.9)
O  Pyl/       Pyrrolysine =     0      0 %   (0)

Protein molecular weight:28508.13 Da (Daltons)
Protein weight from atoms:28490.05428 Da (Daltons)
```

Illustration 47: The protein statistics shows important statistical values for each amino acid in the protein sequence

Launching PDB Crystal structures in Pymol

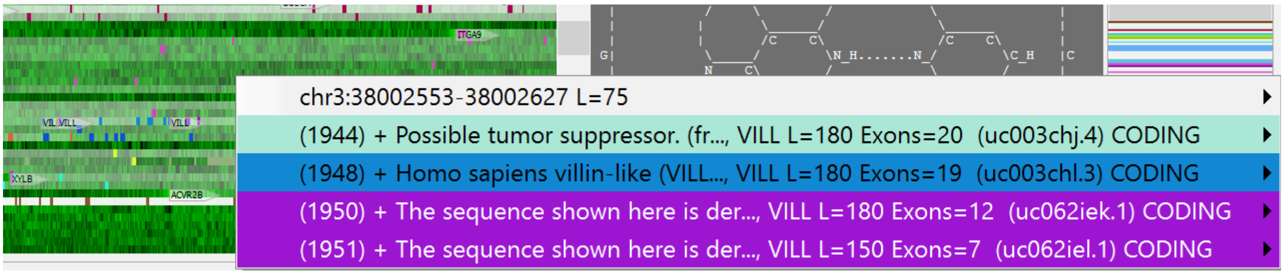
Selecting “Show PDB Structure in Pymol” will download a few PDB files from the Protein Database and colour code the amino acids according to the same colour legend. It will keep these PDB files in a disk cache in order that it does not need to be downloaded next time.

The 3D visualization software “PyMol” will then be launched in order to display the crystal structures determined for these proteins. Other visualization software will later also be supported.



The Gene Context Menu

When you want to get more information for the overlapping genes, you can right-click on any of the windows or trees in order to get a context menu which is coloured according to the genes below the mouse cursor, allowing you to easily select the most appropriate gene. A "software generated" **unique** number is always shown in brackets**. This will help you to identify a unique gene annotation, no matter in which window it is selected from**.



chr3:38002553-38002627 L=75

(1944) + Possible tumor suppressor. (fr..., VILL L=180 Exons=20 (uc003chj.4) CODING

(1948) + Homo sapiens villin-like (VILL..., VILL L=180 Exons=19 (uc003chl.3) CODING

(1950) + The sequence shown here is der..., VILL L=180 Exons=12 (uc062iek.1) CODING

(1951) + The sequence shown here is der..., VILL L=150 Exons=7 (uc062iel.1) CODING

The context menu provides a list of context specific options which can be performed on each gene. The context menu is available from the following windows:

- **Main Genome View** (when right clicking at any location in the genome or on gene regions)
- **DNA View**
- **Protein Sequence View**
- **Gene Search Results tree**
- Biochemical **Pathways tree**

The following options are available in the gene context menu:

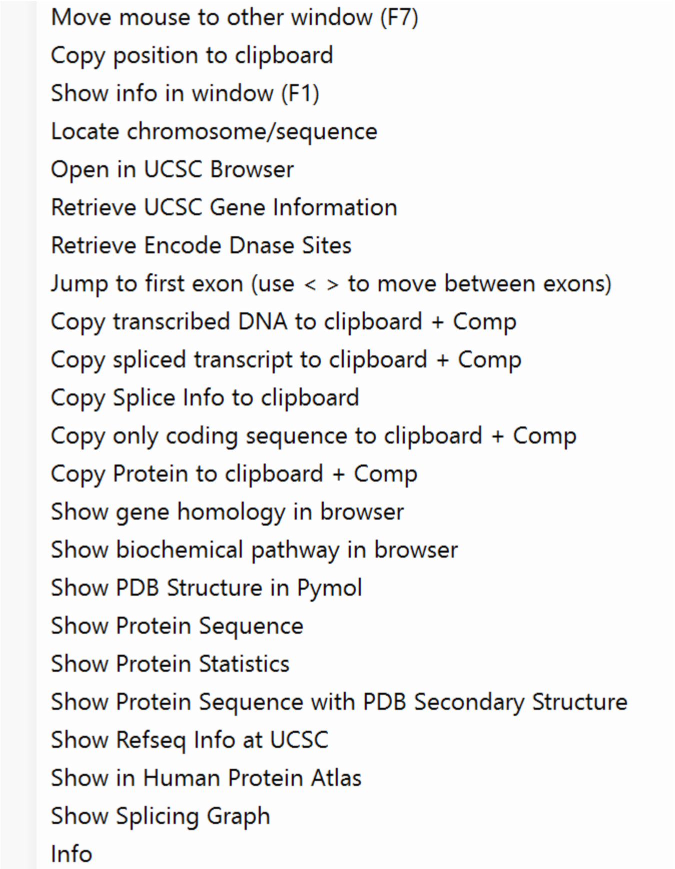

Move mouse to other window (F7)
Copy position to clipboard
Show info in window (F1)
Locate chromosome/sequence
Open in UCSC Browser
Retrieve UCSC Gene Information
Retrieve Encode Dnase Sites
Jump to first exon (use < > to move between exons)
Copy transcribed DNA to clipboard + Comp
Copy spliced transcript to clipboard + Comp
Copy Splice Info to clipboard
Copy only coding sequence to clipboard + Comp
Copy Protein to clipboard + Comp
Show gene homology in browser
Show biochemical pathway in browser
Show PDB Structure in Pymol
Show Protein Sequence
Show Protein Statistics
Show Protein Sequence with PDB Secondary Structure
Show Refseq Info at UCSC
Show in Human Protein Atlas
Show Splicing Graph
Info

Illustration 48: Context menu options

Move mouse to other window (F7)

When you want to quickly navigate between the **Main Genome View** and the **DNA View**, this option allows you to quickly jump between the two windows. It can also be achieved by pressing F7 or using the ***middle button of the mouse***.

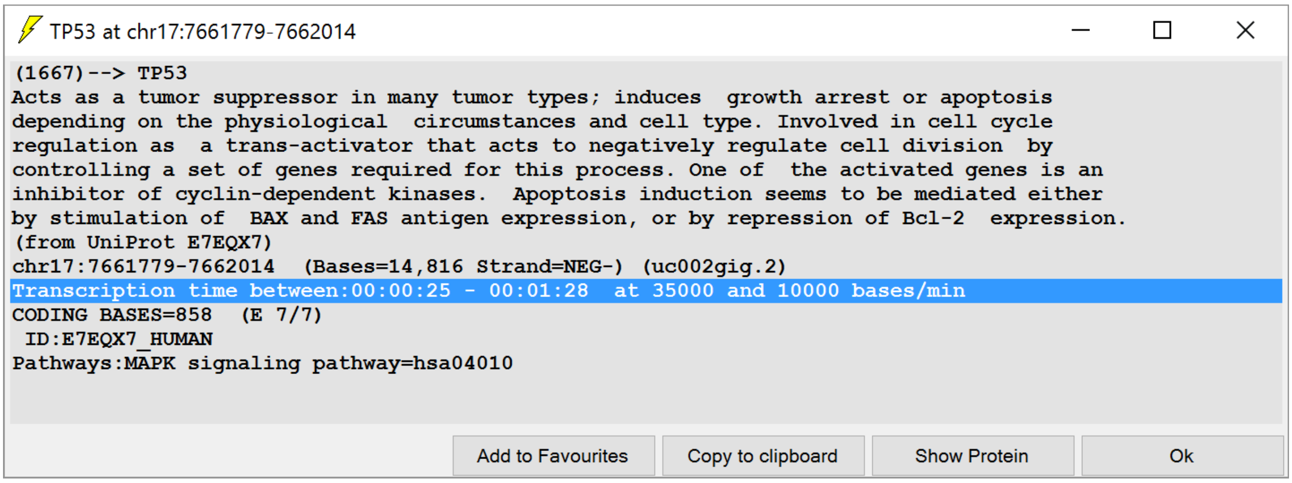
Copy position to clipboard

The Visual Genome Browser uses the 1-based position format also used in the online UCSC browser. This option will simply copy the genome position below the cursor to the clipboard eg. **chr3:159995734-159996019**

Show Info in window (F1)

The **Information Display** in the top centre of the main screen displays information of the gene or annotation currently below the mouse cursor or which is selected in the centre of the cross hairs of the **DNA View** or the **Zoom Gene View.**

This option will open up a separate window showing this information, and also gives you the option of adding the selected gene to the **Favourites**.



```
(1667)--> TP53
Acts as a tumor suppressor in many tumor types; induces  growth arrest or apoptosis
depending on the physiological  circumstances and cell type. Involved in cell cycle
regulation as  a trans-activator that acts to negatively regulate cell division  by
controlling a set of genes required for this process. One of  the activated genes is an
inhibitor of cyclin-dependent kinases.  Apoptosis induction seems to be mediated either
by stimulation of  BAX and FAS antigen expression, or by repression of Bcl-2  expression.
(from UniProt E7EQX7)
chr17:7661779-7662014  (Bases=14,816 Strand=NEG-) (uc002gig.2)
Transcription time between:00:00:25 - 00:01:28  at 35000 and 10000 bases/min
CODING BASES=858  (E 7/7)
 ID:E7EQX7_HUMAN
Pathways:MAPK signaling pathway=hsa04010
```

Locate chromosome/sequence

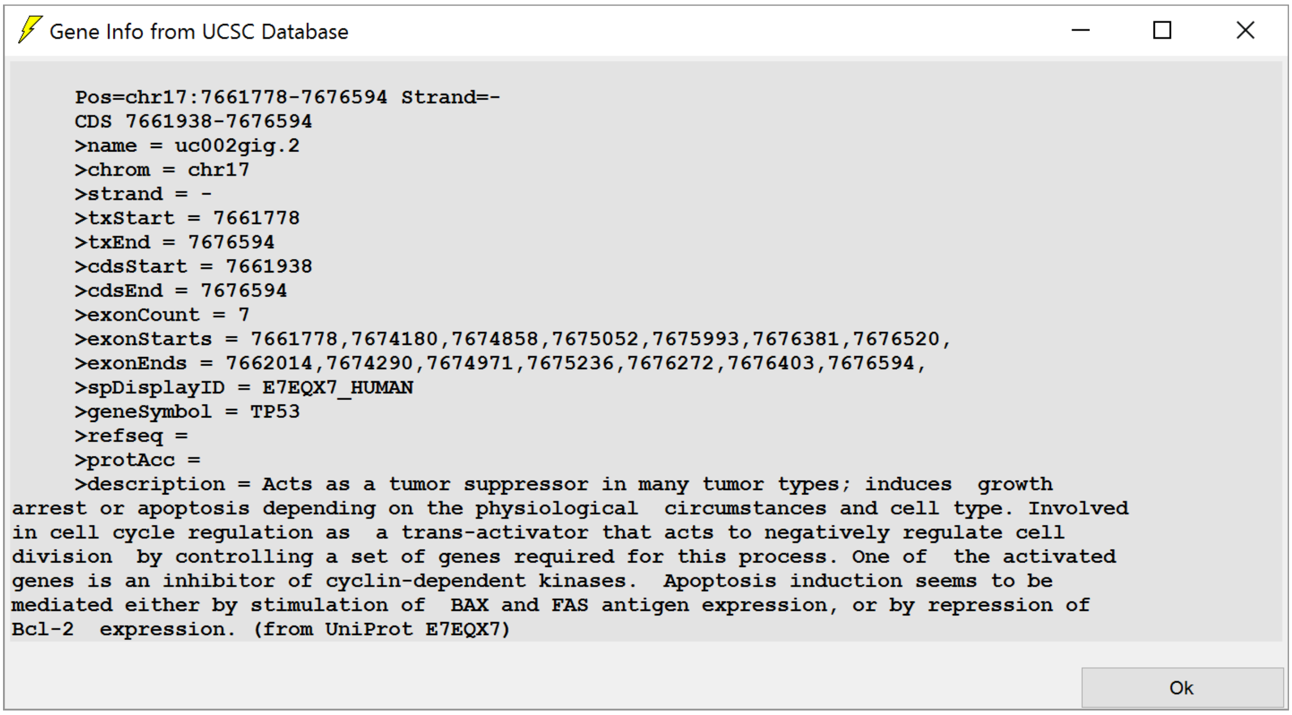
When one has searched for genes matching a specific search criteria and you are looking at in a tree containing genes from many chromosomes, this option allows you to quickly locate the gene's current chromosome in the **Chromosomes/Sequences list**.

Open in UCSC Browser

This option will open the selected gene position in the UCSC Genome Browser

Retrieve UCSC Gene Information

This option will retrieve the MySQL query fields from the UCSC Tables and display it in a window



```
     Pos=chr17:7661778-7676594 Strand=-
     CDS 7661938-7676594
     >name = uc002gig.2
     >chrom = chr17
     >strand = -
     >txStart = 7661778
     >txEnd = 7676594
     >cdsStart = 7661938
     >cdsEnd = 7676594
     >exonCount = 7
     >exonStarts = 7661778,7674180,7674858,7675052,7675993,7676381,7676520,
     >exonEnds = 7662014,7674290,7674971,7675236,7676272,7676403,7676594,
     >spDisplayID = E7EQX7_HUMAN
     >geneSymbol = TP53
     >refseq =
     >protAcc =
     >description = Acts as a tumor suppressor in many tumor types; induces  growth
arrest or apoptosis depending on the physiological  circumstances and cell type. Involved
in cell cycle regulation as  a trans-activator that acts to negatively regulate cell
division  by controlling a set of genes required for this process. One of  the activated
genes is an inhibitor of cyclin-dependent kinases.  Apoptosis induction seems to be
mediated either by stimulation of  BAX and FAS antigen expression, or by repression of
Bcl-2  expression. (from UniProt E7EQX7)
```

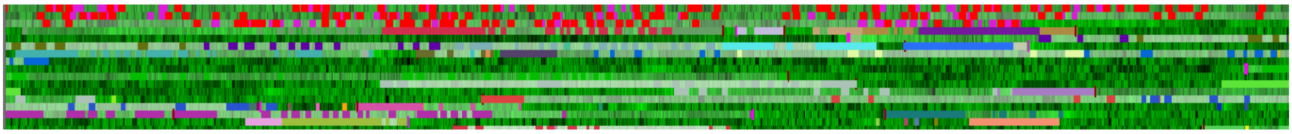
It returns all of the same information which were downloaded for the indexed gene annotations.

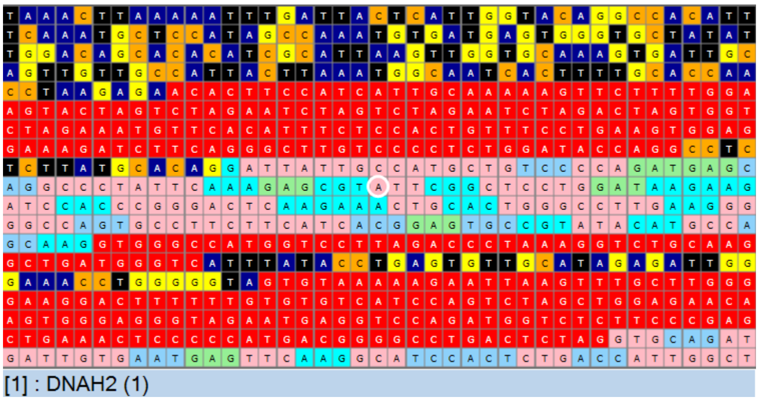
Gene annotations where downloaded from the UCSC database using the following SQL query:

```
SELECT HG38.knownGene.name, HG38.knownGene.chrom, HG38.knownGene.strand,
HG38.knownGene.txStart,

HG38.knownGene.txEnd, HG38.knownGene.cdsStart, HG38.knownGene.cdsEnd,

HG38.knownGene.exonCount, HG38.knownGene.exonStarts, HG38.knownGene.exonEnds,

HG38.kgXref.spDisplayID, HG38.kgXref.geneSymbol, HG38.kgXref.refseq,

HG38.kgXref.protAcc, HG38.kgXref.description,

HG38.keggMapDesc.description as pathway, HG38.keggPathway.mapID as pathwaymapid,

hgFixed.refSeqSummary.summary as summary

FROM HG38.knownGene

inner join HG38.kgXref on HG38.knownGene.name = HG38.kgXref.kgID

left outer join HG38.keggPathway on HG38.knownGene.name = HG38.keggPathway.kgID

left outer join HG38.keggMapDesc on HG38.keggPathway.mapID = HG38.keggMapDesc.mapID

left outer join hgFixed.refSeqSummary on HG38.kgXref.refseq =
hgFixed.refSeqSummary.mrnaAcc

where HG38.knownGene.chrom like 'chrX'

ORDER BY HG38.knownGene.txStart;
```

Retrieve Encode Dnase Hypersensitivity Sites

This is currently only available for this software when using **HG19**. It downloads the DNAse 1 Hypersensitivity sites for a region around the selected position and it is highlighted in red.





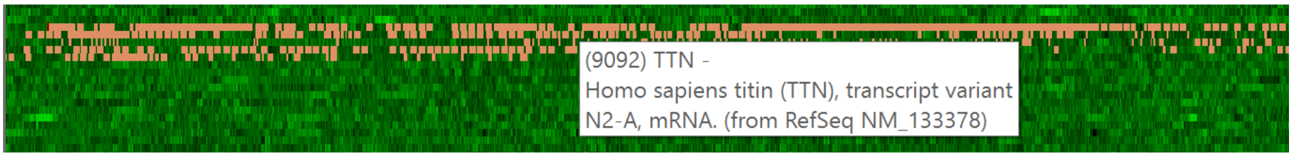
```
T A A A C T T A A A A A T T T G A T T A C T C A T T G G T A C A G G C C A C A T T
T C A A A T G C T C C A T A G C C A A A T G T G A T G A G T G G G T G C T A T A T
T G G A C A G C A C A C A T C G C A T T A A G T T G G T G C A A A G T G A T T G C
A G T T G T T G C C A T T A C T T A A A T G G C A A T C A C T T T T G C A C C A A
C C T A A G A G A A C A C T T C C A T C A T T G C A A A A A G T T C T T T T G G A
A G T A C T A G T C T A G A A T C T A G T C T A G A A T C T A G A C T A G T G G T
C T A G A A A T G T T C A C A T T T C T C C A C T G T T T C C T G A A G T G G A G
G A A A G A T C T T C A G G G C T T G T C C C C T C T G G A T A C C A G G C C T C
T C T T A T G C A C A G G A T T A T T G C C A T G C T G T C C C C A G A T G A G C
A G G C C C T A T T C A A A G A G C G T A T T C G G C T C C T G G A T A A G A A G
A T C C A C C C G G G A C T C A A G A A A C T G C A C T G G G C C T T G A A G G G
G G C C A G T G C C T T C T T C A T C A C G G A G T G C C G T A T A C A T G C C A
G C A A G G T G G G C C A T G G T C C T T A G A C C C T A A A G G T C T G C A A G
G C T G A T G G G T C A T T T A T A C C T G A G T G T T G C A T A G A G A T T G G
G A A A C C T G G G G G T A G T G T A A A A A G A A T T A A G T T T G C T T G G G
G A A G G A C T T T T T T G T G T G T C A T C C A G T C T A G C T G G A G A A C A
A G T G G G A G G G T A G A A T G A G G T C C A G A T G G T C T C T T C C C G A G
C T G A A A C T C C C C C A T G A C G G G G C C T G A C T C T A G G T G C A G A T
G A T T G T G A A T G A G T T C A A G G C A T C C A C T C T G A C C A T T G G C T
```

[1] : DNAH2 (1)

Jump to first exon (use < > to move between exons)

It is sometimes helpful to be able to navigate quickly from exon to exon in the **DNA View**. This is where this option comes in handy. When selected, it immediately jumps to the first exon of the gene. After this is selected you can use the **< and > keys** in order to quickly jump through the list of exons in the gene.

Take the TTN (Titin gene) for example. Setting the intron transparency to 100% and filtering out all genes not containing the software based unique no (9092), cleans up the display to allow you to only see the single Titin transcript consisting of 312 exons.



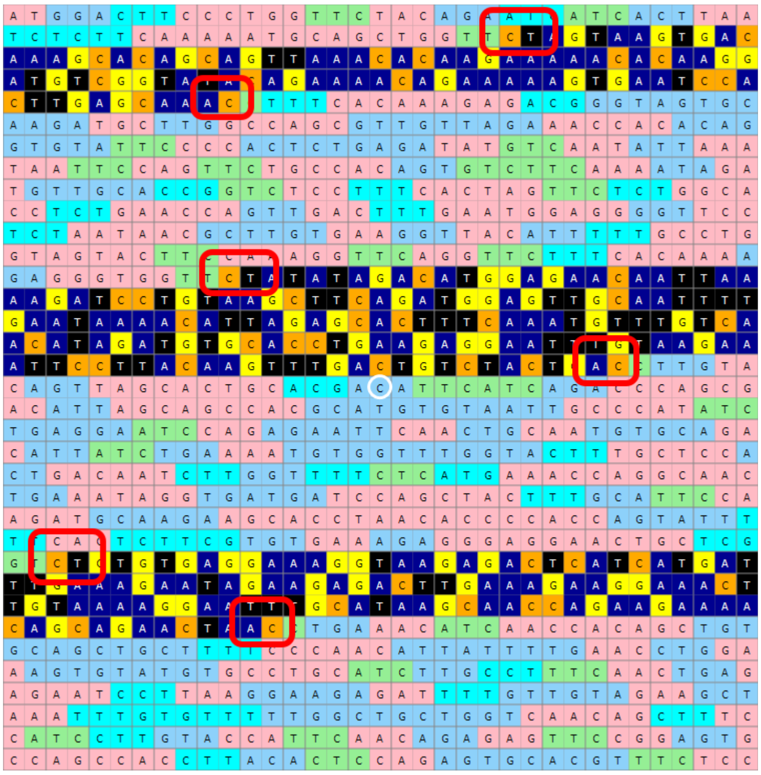
(9092) TTN -
Homo sapiens titin (TTN), transcript variant N2-A, mRNA. (from RefSeq NM_133378)

Pressing > repeatedly navigates forward through the exons up to exon 312 and

pressing < navigates backwards to exon 1.

The TTN gene is transcribed from the negative strand and the RNA Polymerase II enzyme moves backwards through the chromosome to correctly read the bases.

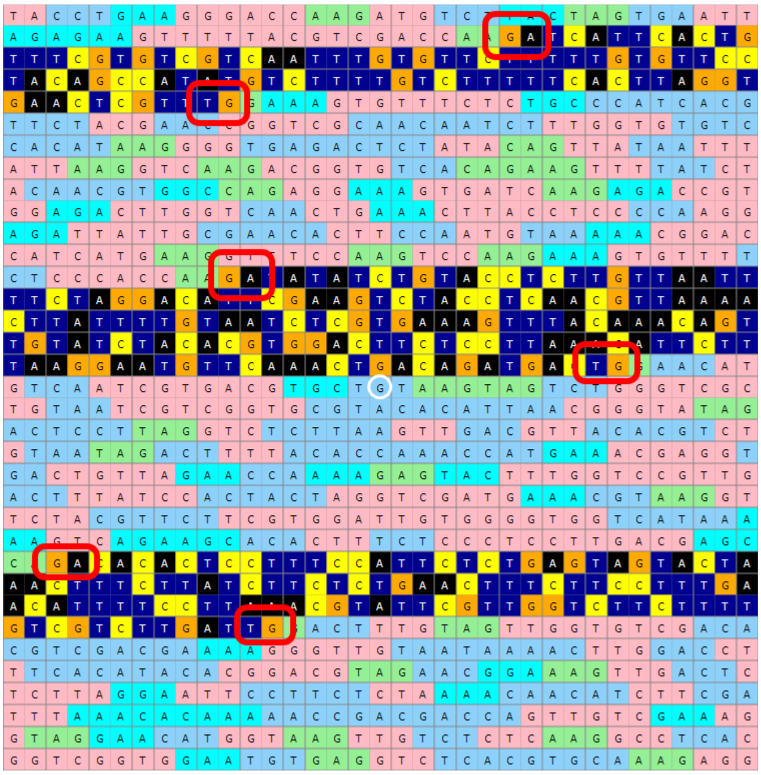
Notice that the splice signals GT....AG are reverse complemented as CT....AC



```
A T G G A C T T C C C T G G T T C T A C A G A A T G A T C A C T T A A
T C T C T T C A A A A A T G C A G C T G G T T C T A G T A A G T G A C
A A A G C A C A G C A G T T A A A C A C A A G A A A A A C A C A A G G
A T G T C G G T A T A C A G A A A A C A G A A A A A G T G A A T C C A
C T T G A G C A A A C C T T T C A C A A A G A G A C G G G T A G T G C
A A G A T G C T T G G C C A G C G T T G T T A G A A A C C A C A C A G
G T G T A T T C C C C A C T C T G A G A T A T G T C A A T A T T A A A
T A A T T C C A G T T C T G C C A C A G T G T C T T C A A A A T A G A
T G T T G C A C C G G T C T C C T T T C A C T A G T T C T C T G G C A
C C T C T G A A C C A G T T G A C T T T G A A T G G A G G G G T T C C
T C T A A T A A C G C T T G T G A A G G T T A C A T T T T T G C C T G
G T A G T A C T T C C A A A G G T T C A G G T T C T T T C A C A A A A
G A G G G T G G T T C T A T A T A G A C A T G G A G A A C A A T T A A
A A G A T C C T G T A A G C T T C A G A T G G A G T T G C A A T T T T
G A A T A A A A C A T T A G A G C A C T T T C A A A T G T T T G T C A
A C A T A G A T G T G C A C C T G A A G A G G A A T T T G T A A G A A
A T T C C T T A C A A G T T T G A C T G T C T A C T G A C C T T G T A
C A G T T A G C A C T G C A C G A C A T T C A T C A G A C C C A G C G
A C A T T A G C A G C C A C G C A T G T G T A A T T G C C C A T A T C
T G A G G A A T C C A G A G A A T T C A A C T G C A A T G T G C A G A
C A T T A T C T G A A A A T G T G G T T T G G T A C T T T G C T C C A
C T G A C A A T C T T G G T T T T C T C A T G A A A C C A G G C A A C
T G A A A T A G G T G A T G A T C C A G C T A C T T T G C A T T C C A
A G A T G C A A G A A G C A C C T A A C A C C C C A C C A G T A T T T
T T C A G T C T T C G T G T G A A A G A G G G A G G A A C T G C T C G
G T C T G T G T G A G G A A A G G T A A G A G A C T C A T C A T G A T
T T G A A A G A A T A G A A G A G A C T T G A A A G A A G G A A A C T
T G T A A A A G G A A T T T G C A T A A G C A A C C A G A A G A A A A
C A G C A G A A C T A A C C T G A A A C A T C A A C C A C A G C T G T
G C A G C T G C T T T T C C C A A C A T T A T T T T G A A C C T G G A
A A G T G T A T G T G C C T G C A T C T T G C C T T T C A A C T G A G
A G A A T C C T T A A G G A A G A G A T T T T G T T G T A G A A G C T
A A A T T T G T G T T T T T G G C T G C T G G T C A A C A G C T T T C
C A T C C T T G T A C C A T T C A A C A G A G A G T T C C G G A G T G
C C A G C C A C C T T A C A C T C C A G A G T G C A C G T T T C T C C
```

Illustration 49: chr2:178723267-178724491 within TTN gene at width 35

The splice signals which are part of the introns are highlighted with red rectangles. The codons have to be read in the reverse direction. It is actually possible to conveniently complement the bases by selecting the checkbox option "**Complement**" when the gene in question is encoded on the reverse strand as is the case here. (Genes on the positive strand will also show correctly)



```
T A C C T G A A G G G A C C A A G A T G T C T T A C T A G T G A A T T
A G A G A A G T T T T T A C G T C G A C C A A G A T C A T T C A C T G
T T T C G T G T C G T C A A T T T G T G T T C T T T T T G T G T T C C
T A C A G C C A T A T G T C T T T T G T C T T T T T C A C T T A G G T
G A A C T C G T T T G G A A A G T G T T T C T C T G C C C A T C A C G
T T C T A C G A A C C G G T C G C A A C A A T C T T T G G T G T G T C
C A C A T A A G G G G T G A G A C T C T A T A C A G T T A T A A T T T
A T T A A G G T C A A G A C G G T G T C A C A G A A G T T T T A T C T
A C A A C G T G G C C A G A G G A A A G T G A T C A A G A G A C C G T
G G A G A C T T G G T C A A C T G A A A C T T A C C T C C C C A A G G
A G A T T A T T G C G A A C A C T T C C A A T G T A A A A A C G G A C
C A T C A T G A A G G T T T C C A A G T C C A A G A A A G T G T T T T
C T C C C A C C A A G A T A T A T C T G T A C C T C T T G T T A A T T
T T C T A G G A C A T T C G A A G T C T A C C T C A A C G T T A A A A
C T T A T T T T G T A A T C T C G T G A A A G T T T A C A A A C A G T
T G T A T C T A C A C G T G G A C T T C T C C T T A A A C A T T C T T
T A A G G A A T G T T C A A A C T G A C A G A T G A C T G G A A C A T
G T C A A T C G T G A C G T G C T G T A A G T A G T C T G G G T C G C
T G T A A T C G T C G G T G C G T A C A C A T T A A C G G G T A T A G
A C T C C T T A G G T C T C T T A A G T T G A C G T T A C A C G T C T
G T A A T A G A C T T T T A C A C C A A A C C A T G A A A C G A G G T
G A C T G T T A G A A C C A A A A G A G T A C T T T G G T C C G T T G
A C T T T A T C C A C T A C T A G G T C G A T G A A A C G T A A G G T
T C T A C G T T C T T C G T G G A T T G T G G G G T G G T C A T A A A
A A G T C A G A A G C A C A C T T T C T C C C T C C T T G A C G A G C
C A G A C A C A C T C C T T T C C A T T C T C T G A G T A G T A C T A
A A C T T T C T T A T C T T C T C T G A A C T T T C T T C C T T T G A
A C A T T T T C C T T A A A C G T A T T C G T T G G T C T T C T T T T
G T C G T C T T G A T T G G A C T T T G T A G T T G G T G T C G A C A
C G T C G A C G A A A A G G G T T G T A A T A A A A C T T G G A C C T
T T C A C A T A C A C G G A C G T A G A A C G G A A A G T T G A C T C
T C T T A G G A A T T C C T T C T C T A A A A C A A C A T C T T C G A
T T T A A A C A C A A A A A C C G A C G A C C A G T T G T C G A A A G
G T A G G A A C A T G G T A A G T T G T C T C T C A A G G C C T C A C
G G T C G G T G G A A T G T G A G G T C T C A C G T G C A A A G A G G
```

Illustration 50: chr2:178723267-178724491 with bases complemented (Splice bases are highlighted)

The codons can now be read backwards, but are showing the correct complemented base.

It seems like, although the lengths of introns may vary a lot, the lengths of the exons tend to fall in a narrow range of lengths. What is interesting to me is how the codons of a protein coding gene can be scattered out over large distances, but still join up (even in the middle of codons a lot of the time).

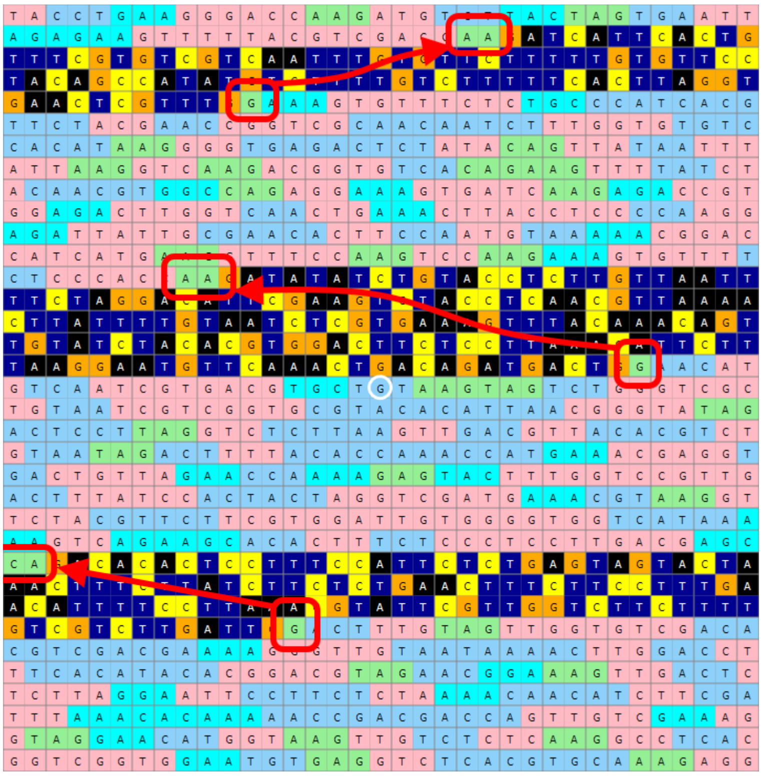
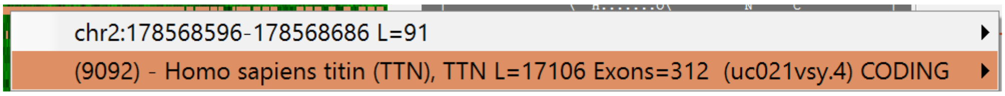


Illustration 51: chr2:178723267-178724491 with bases complemented (Broken up codons are highlighted)

In the picture above you can see the same exons (Exons 69-72 of a total of 312) with the codons which span more than one exon. Only after the exons are correctly spliced together are these codons able to make sense in terms of the protein coding region.

It is like a puzzle which only makes sense once you have correctly assembled the individual pieces. I sometimes wonder how these original genes got split up into exons and then got incorporated with the correct splice signals (GT...AG) in at sometimes distant positions among intronic regions. Still, when you assemble the exons, the codons suddenly match up and can code for proteins.

The software provides the following information for this gene, which produces the biggest protein in the human body:



chr2:178568596-178568686 L=91

(9092) - Homo sapiens titin (TTN), TTN L=17106 Exons=312 (uc021vsy.4) CODING

Here is an example of the kind of information displayed for a gene:

(9092)--> TTN

Homo sapiens titin (TTN), transcript variant N2-A, mRNA. (from **RefSeq NM_133378**)

SUMMARY:This gene encod...

chr2:178530241-178535849 (Bases=281,433 Strand=NEG-) (uc021vsy.4)

Transcription time between:00:08:02 - 00:28:08 at 35000 and 10000 bases/min

CODING BASES=100,272 Protein:NM_133378 (**L=33424**) (**E 307/312**)

ID:TITIN_HUMAN

PDB:1BPV;1G1C;1NCT;1NCU;1TIT;1TIU;1TKI;1TNM;1TNN;1WAA;1YA5;2A38;2BK8;2F8V;2ILL;2J8H;2J8O;2NZI;2RQ8;2WP3;2WWK;2WWM;2Y9R;3B43;3KNB;3LCY;3LPW;3PUC;3Q5O;3QP3;4C4K;4JNW;4O00;4QEG;4UOW

Pathways:***Hypertrophic cardiomyopathy*** (HCM)=hsa05410

This gene encodes a large abundant protein of striated muscle. The product of this gene is divided into two regions, a N-terminal I-band and a C-terminal A-band. The I-band, which is the elastic part of the molecule, contains two regions of tandem immunoglobulin domains on either side of a PEVK region that is rich in proline, glutamate, valine and lysine. The A-band, which is thought to act as a protein-ruler, contains a mixture of immunoglobulin and fibronectin repeats, and possesses kinase activity. An N-terminal Z-disc region and a C-terminal M-line region bind to the Z-line and M-line of the sarcomere, respectively, so that a single titin molecule spans half the length of a sarcomere. Titin also contains binding sites for muscle associated proteins so it serves as an adhesion template for the assembly of contractile machinery in muscle cells. It has also been identified as a structural protein for chromosomes. Alternative splicing of this gene results in multiple transcript variants. Considerable variability exists in the I-band, the M-line and the Z-disc regions of titin. Variability in the I-band region contributes to the differences in elasticity of different titin isoforms and, therefore, to the differences in elasticity of different muscle types. Mutations in this gene are associated with familial hypertrophic cardiomyopathy 9, and autoantibodies to titin are produced in patients with the autoimmune disease scleroderma. [provided by RefSeq, Feb 2012].

Copy transcribed DNA to clipboard + Comp

This option copies the raw DNA bases to the clipboard (and the Comparison list on the Main tab) Bases are copied as FASTA output and in the correct direction (in other words: negative strand transcription will appear in the forward direction as it is transcribed). Intronic bases is shown in lower case and exons in upper case.

```
>TTN ; Homo sapiens titin (TTN), transcript variant N2-A, mRNA. (from RefSeq NM_133378) ; Transcript length = 306160 (uc021vsy.4 Full gene position = chr2:178525989-178807421)

GCAGTCGTGCATTCCCAGCCTCGCCTCGGGTGTAGGGATTGCATAGAAAAGCAAAACTACACAGTCTTGACTGTGTAGTTTTGTTTTTAGGATTAGAGGC

TCACCGATTCATGTCGGAGATGGTCAGAAAAACCAACTCTCCATAGGACGTCGTTTCAGAAGCAACCTTGGGCTTAGTCCCACCCTTTTTAGGCACTCTT

GAGAAATCAGGTGAGCAGCAACTTTTCTTATTTTAATAGTAGAGTCACTTTCTTATTGATCTCTGGCTCTGGATTATTTGTGTGTGCTAAATGCCTATGT

ACAAGTGTTTGTATGACTATTTCTTGCCAGTGCTGACCCATTATTTAAATCTGCTTGTTAAATTCTATTTTAAAATGTCTGAACATTTTTCCTTAAAAAG

TTGAAAGCTTTATTATATCAGCTAAATGATTCCATGATTCAAATTGTTTAAGTTCACTTAATATTCTAAAAGTTGAGGTCTGAAATGTTGAGAGAAATGT

TTACTTCTTTTCTAGACTCTCTTTGCCTGTTGAGAATGTTCAAGGTTATATTACTCAATGATAAAGTTAAGCACACCAAGAATAGCTCTGATAAAATGCA

TGTGGTATCAGCTACACCCTGATGACTTTTTAAGGGAATCAACTGCTGAGGTATCAGTTGCAGGACAGTGACAGAACACTTGCCTGCTGTCAGTCAACCT

TGGAGAGTTTAGGAGAGTTTATTTTAAGAGCATGACATTTTAGACGCATACCTTCAGCCCCAGTTAGACAAAGGCCTGTGAAGAGCAAGTGCTAAGGCAA

TAAAATTTACCTAGACAATAGGAATGTAAGATCAATTTAAGCTGGTTAATTGCAACATAGTTGGAGATTTGGAATAAAATGATATAATTGTTTTCTAAAA

TATTCAGAATCTCTTCATGCATTTTAACTCAGCACACAGTCCAATTATATATTTTGAGATAAATGTTGACACTAAGAAGGGTAAGAAATAATGTTTAAAG
```

```
GAAAGGTTTCTTACTTCTAAATACACTTTGATTTCTCATCTTCCTAGAATTTGATTACTAATTGTCAATGCCTGTAAATACATAGATATATAATTAGGTA
TTCTATGTCTAAGAGGACAGCTATTTTTAGATATACTAACTTATAAACAGACAAAGTTAGACACAGAAAGTTGTTCATCCCAAAGATAATATGCTACAAA
AAGGAGTTTCTGACAAAAAAAATTTATGAAAGCTTTTGGATATTATCTGATTTTTCCAAATATAAATTTTATTGTCTCAGAGAATGCTGGTTTCATTTTT
TTAACTAAGTAGAAGGTTAcagtaaggaacagaattatcactaagaagattattgctccatgaagttttagtaaacaaagaatttaatgaaaaatagtaa
tgaagtgattgacaattcagaaatgttctatttcagaacagaatagtcatgttgaataattaggtatttatgcatcagaaaattcttcacataatgtctc
tgaaactgataggcaaagtactttcttgcacttaacaaaatgacttaactgcaaattttatattcgtttctttgatgttgaatattttactttatgaaag
cataggccttttcacttcatctacattttacaccagccttccattattcaaagtttgcaaattgttccatattaagaactcaagccagcatttttctttc
agaactataataaactttggcctgttcactaaaatttaaatttttctgaatatttacttcagcagtaacaacatttactgcacagttacagtatgaagaa
atgttaagaaatagaaagcaccgtatgccgtctacaaaaagcttgcatatttgataaggcttgccaaagcttctccagtcaggaaagagattacaaagtt
cagatttttacttctggcaactattttaagtaatcagataatgtctgaatcaatgtcaactttataaatatagagtgaacgaaaaatattggggagaagg
aaaatttttaaaataattttcacttactaagctgatgcagaggtttatttattttaatttttttttttttttttttttgagacagagtctcattctgtca
cccaggctggagtgtggtggcaccatctttgctcactgcaacctccacctcctgggtttaagcaattctcatgcctcagcctcccaagtagctgggacta
cagacatgggccaccacgcctggctaatttttgtatttgtagtagagacaggggtttcaccatgttagccaaactggtctcaaactcctgacctcaagta
atcctcccgcctcggcctcctaaagtgctgggattaccagcatgagccaccacgccaggcccagaggtttattttacatttatcacatgtcttgtctctc
taattacacttcactaaccctgcattgctgtggttttcttcttcaactctgaacctggttttccaccattggtaagactgtatattccctctttgaccac
tagggttcattccttgttctgtttggctaatactttttttccatttccccttacacttgctactttgggttaggctttaagtcaactcaatgatgaactat
gcaaaactgttcaagccacaaaagagaagacctgcccatcactctaggcccacagatgacctatggagcaatccatttggagaaacgtgtgtctctgcta
tctgcaaagcagctccagagtgacccttagctgggacaccctaatttattttctcttctttttcagAGTGCCTAGAAAGATGACAACTCAAGCACCGACG
TTTACGCAGCCGTTACAAAGCGTTGTGGTACTGGAGGGTAGTACCGCAACCTTTGAGGCTCACATTAGTGGTAAGCTCACACATTCACACTTTTGTTTTT
TTTTCCTTTGCCTCCTCTCCAGTAAGTTAACGTTGCTGCAGGACTTGACGCCAAGTTTAAGCCCTGCTTTCACTTCGGAAAATTAGTCCAACACTATGGA
...
```

Copy spliced transcript to clipboard + Comp

This option will copy the mature transcript (after splicing) to the clipboard and and Comparison list. This means it will include the 5'UTR + Exons + 3'UTR

Copy only coding sequence to clipboard + Comp

This option will copy only the bases used by the ribosome to produce proteins. In other words the coding sequence.

eg.

```
>TTN ; Homo sapiens titin (TTN), transcript variant N2-A, mRNA. (from RefSeq NM_133378) ; CDS length = 100272 (uc021vsy.4 Full gene
position = chr2:178525989-178807421)

ATGACAACTCAAGCACCGACGTTTACGCAGCCGTTACAAAGCGTTGTGGTACTGGAGGGTAGTACCGCAACCTTTGAGGCTCACATTAGTGGTTTTCCAG

TTCCTGAGGTGAGCTGGTTTAGGGATGGCCAGGTGATTTCCACTTCCACTCTGCCCGGCGTGCAGATCTCCTTTAGCGATGGCCGCGCTAAACTGACGAT

CCCCGCCGTGACTAAAGCCAACAGTGGACGATATTCCCTGAAAGCCACCAATGGATCTGGACAAGCGACTAGTACTGCTGAGCTTCTCGTGAAAGCTGAG

ACAGCACCACCCAACTTCGTTCAACGACTGCAGAGCATGACCGTGAGACAAGGAAGCCAAGTGAGACTCCAAGTGAGAGTGACTGGAATCCCTACACCTG

….

TTCCATCTGATATCAGCATTGATGAAGGCAAAGTTCTAACAGTAGCCTGTGCTTTCACGGGTGAGCCTACCCCAGAAGTAACATGGTCCTGTGGTGGAAG

AAAAATCCACAGTCAAGAACAGGGGAGGTTCCACATTGAAAACACAGATGACCTGACAACCCTGATCATCATGGACGTACAGAAACAAGATGGTGGACTT

TATACCCTGAGTTTAGGGAATGAATTTGGATCTGACTCTGCCACTGTGAATATACATATTCGATCCATTTAA
```

Copy Protein to clipboard + Comp

This option will ***use the appropriate Genetic Code table*** selected in the “Controls” tab to translate the coding sequence into a protein and copy that amino acid letters as a FASTA string to the clipboard.

```
>TTN ; Homo sapiens titin (TTN), transcript variant N2-A, mRNA. (from RefSeq NM_133378) ; Protein length = 33423 (uc021vsy.4 Full
gene position = chr2:178525989-178807421)

MTTQAPTFTQPLQSVVVLEGSTATFEAHISGFPVPEVSWFRDGQVISTSTLPGVQISFSDGRAKLTIPAVTKANSGRYSLKATNGSGQATSTAELLVKAE

TAPPNFVQRLQSMTVRQGSQVRLQVRVTGIPTPVVKFYRDGAEIQSSLDFQISQEGDLYSLLIAEAYPEDSGTYSVNATNSVGRATSTAELLVQGEEEVP

AKKTKTIVSTAQISESRQTRIEKKIEAHFDARSIATVEMVIDGAAGQQLPHKTPPRIPPKPKSRSPTPPSIAAKAQLARQQSPSPIRHSPSPVRHVRAPT

PSPVRSVSPAARISTSPIRSVRSPLLMRKTQASTVATGPEVPPPWKQEGYVASSSEAEMRETTLTTSTQIRTEERWEGRYGVQEQVTISGAAGAAASVSA

…

SGKYTIKAKNFRGQCSATASLMVLPLVEEPSREVVLRTSGDTSLQGSFSSQSVQMSASKQEASFSSFSSSSASSMTEMKFASMSAQSMSSMQESFVEMSS

SSFMGISNMTQLESSTSKMLKAGIRGIPPKIEALPSDISIDEGKVLTVACAFTGEPTPEVTWSCGGRKIHSQEQGRFHIENTDDLTTLIIMDVQKQDGGL

YTLSLGNEFGSDSATVNIHIRSI
```

Copy Splice Info to clipboard

This option will copy important length information related to splicing to the clipboard.

It looks as follows for the TTN gene:

TTN ; Homo sapiens titin (TTN), transcript variant N2-A, mRNA. (from RefSeq NM_133378) ; uc021vsy.4 (Full gene position = chr2:178525989-178807421)

Transcript Length: 281433

Exon bases length: 101518

Intron bases length: 179915

Exon 1/312 length:210

Intron 1/311 length:2556

Exon 2/312 length:104

Intron 2/311 length:2210

Exon 3/312 length:204

Intron 3/311 length:1455

Exon 4/312 length:288

Intron 4/311 length:484

Exon 5/312 length:86

Intron 5/311 length:93

Exon 6/312 length:245

Intron 6/311 length:4234

Exon 7/312 length:331

Intron 7/311 length:370

Exon 8/312 length:153

Intron 8/311 length:857

Exon 9/312 length:138

Intron 9/311 length:1206

Exon 10/312 length:126

Intron 10/311 length:1226

…

…

Exon 310/312 length:154

Intron 310/311 length:525

Exon 311/312 length:303

Intron 311/311 length:138

Exon 312/312 length:1319

Exon+Intron 1 length:2766

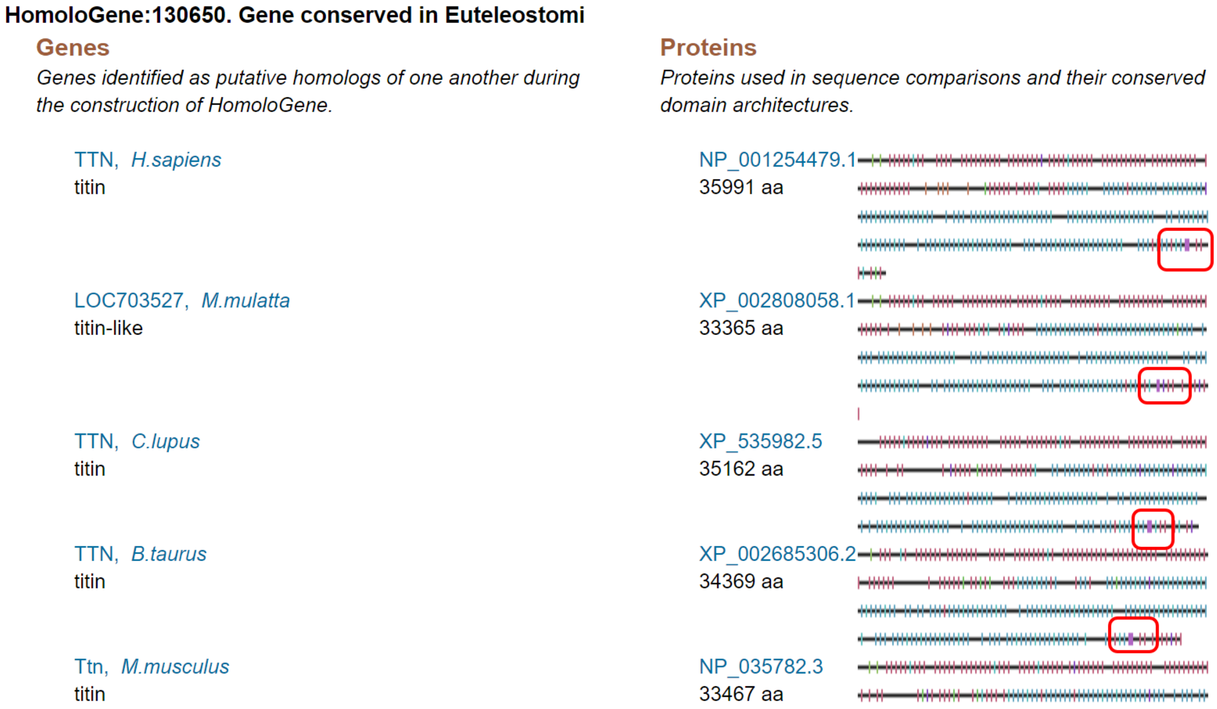
Exon+Intron 2 length:2314

Exon+Intron 3 length:1659

…

Show gene homology in browser

This option ***opens the NIH website*** with the Known Gene name eg. TTN, BRCA1, TP53 etc. in order to find homologs fo this gene in other organisms.



HomoloGene:130650. Gene conserved in Euteleostomi

Genes

Genes identified as putative homologs of one another during the construction of HomoloGene.

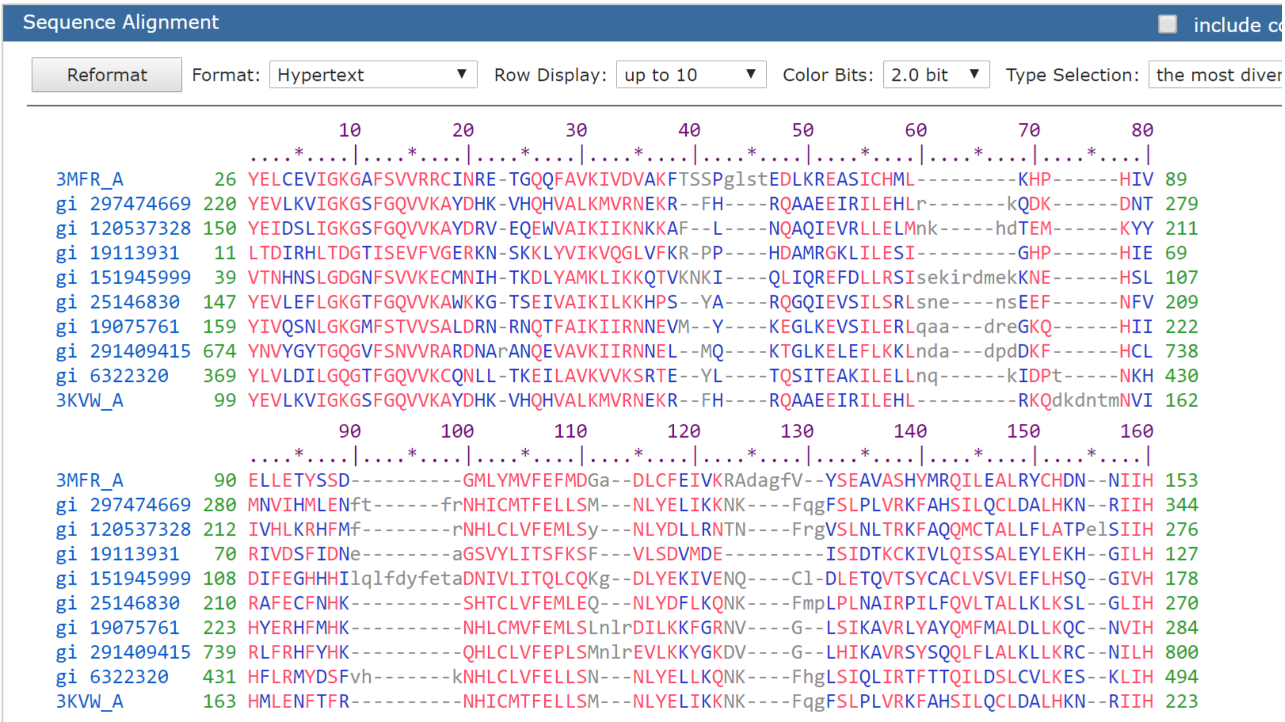
Proteins

Proteins used in sequence comparisons and their conserved domain architectures.

Genes	Proteins
TTN, *H.sapiens* titin	NP_001254479.1 35991 aa
LOC703527, *M.mulatta* titin-like	XP_002808058.1 33365 aa
TTN, *C.lupus* titin	XP_535982.5 35162 aa
TTN, *B.taurus* titin	XP_002685306.2 34369 aa
Ttn, *M.musculus* titin	NP_035782.3 33467 aa

It effectively calls: https://www.ncbi.nlm.nih.gov/homologene/?term=TTN

When you now click on the alignments highlighted in red, you can get sequence alignments between the organisms.



Format: Hypertext ▼ Row Display: up to 10 ▼ Color Bits: 2.0 bit ▼ Type Selection: the most diver

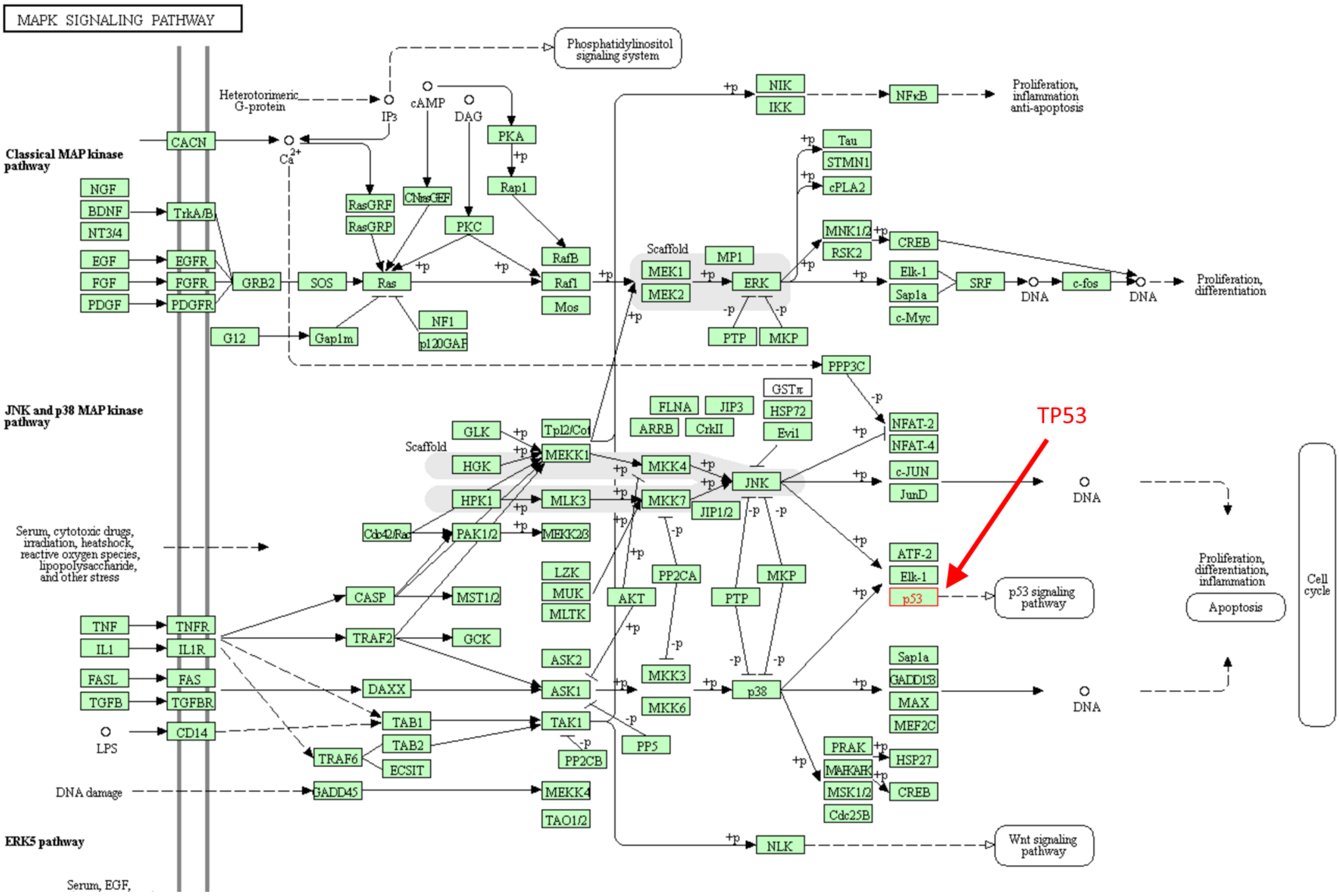
```
                         10        20        30        40        50        60        70        80
                 ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
3MFR_A        26 YELCEVIGKGAFSVVRRCINRE-TGQQFAVKIVDVAKFTSSPglstEDLKREASICHML---------KHP------HIV 89
gi 297474669 220 YEVLKVIGKGSFGQVVKAYDHK-VHQHVALKMVRNEKR--FH----RQAAEEIRILEHLr-------kQDK------DNT 279
gi 120537328 150 YEIDSLIGKGSFGQVVKAYDRV-EQEWVAIKIIKNKKAF--L----NQAQIEVRLLELMnk-----hdTEM------KYY 211
gi 19113931   11 LTDIRHLTDGTISEVFVGERKN-SKKLYVIKVQGLVFKR-PP----HDAMRGKLILESI---------GHP------HIE 69
gi 151945999  39 VTNHNSLGDGNFSVVKECMNIH-TKDLYAMKLIKKQTVKNKI----QLIQREFDLLRSIsekirdmekKNE------HSL 107
gi 25146830  147 YEVLEFLGKGTFGQVVKAWKKG-TSEIVAIKILKKHPS--YA----RQGQIEVSILSRLsne----nsEEF------NFV 209
gi 19075761  159 YIVQSNLGKGMFSTVVSALDRN-RNQTFAIKIIRNNEVM--Y----KEGLKEVSILERLqaa---dreGKQ------HII 222
gi 291409415 674 YNVYGYTGQGVFSNVVRARDNArANQEVAVKIIRNNEL--MQ----KTGLKELEFLKKLnda---dpdDKF------HCL 738
gi 6322320   369 YLVLDILGQGTFGQVVKCQNLL-TKEILAVKVVKSRTE--YL----TQSITEAKILELLnq------kIDPt-----NKH 430
3KVW_A        99 YEVLKVIGKGSFGQVVKAYDHK-VHQHVALKMVRNEKR--FH----RQAAEEIRILEHL---------RKQdkdntmNVI 162
                         90       100       110       120       130       140       150       160
                 ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
3MFR_A        90 ELLETYSSD---------GMLYMVFEFMDGa--DLCFEIVKRAdagfV--YSEAVASHYMRQILEALRYCHDN--NIIH 153
gi 297474669 280 MNVIHMLENft------frNHICMTFELLSM---NLYELIKKNK----FqgFSLPLVRKFAHSILQCLDALHKN--RIIH 344
gi 120537328 212 IVHLKRHFMf--------rNHLCLVFEMLSy---NLYDLLRNTN----FrgVSLNLTRKFAQQMCTALLFLATPelSIIH 276
gi 19113931   70 RIVDSFIDNe--------aGSVYLITSFKSF---VLSDVMDE---------ISIDTKCKIVLQISSALEYLEKH--GILH 127
gi 151945999 108 DIFEGHHHIlqlfdyfetaDNIVLITQLCQKg--DLYEKIVENQ----Cl-DLETQVTSYCACLVSVLEFLHSQ--GIVH 178
gi 25146830  210 RAFECFNHK---------SHTCLVFEMLEQ---NLYDFLKQNK----FmpLPLNAIRPILFQVLTALLKLKSL--GLIH 270
gi 19075761  223 HYERHFMHK---------NHLCMVFEMLSLnlrDILKKFGRNV----G--LSIKAVRLYAYQMFMALDLLKQC--NVIH 284
gi 291409415 739 RLFRHFYHK---------QHLCLVFEPLSMnlrEVLKKYGKDV----G--LHIKAVRSYSQQLFLALKLLKRC--NILH 800
gi 6322320   431 HFLRMYDSFvh-------kNHLCLVFELLSN---NLYELLKQNK----FhgLSIQLIRTFTTQILDSLCVLKES--KLIH 494
3KVW_A       163 HMLENFTFR---------NHICMTFELLSM---NLYELIKKNK----FqgFSLPLVRKFAHSILQCLDALHKN--RIIH 223
```

Show biochemical pathway in browser

Many of the genes in the UCSC table data, especially genes coding for enzymes involved in biochemical pathways, contains reference numbers to the KEGG biochemical pathway maps.

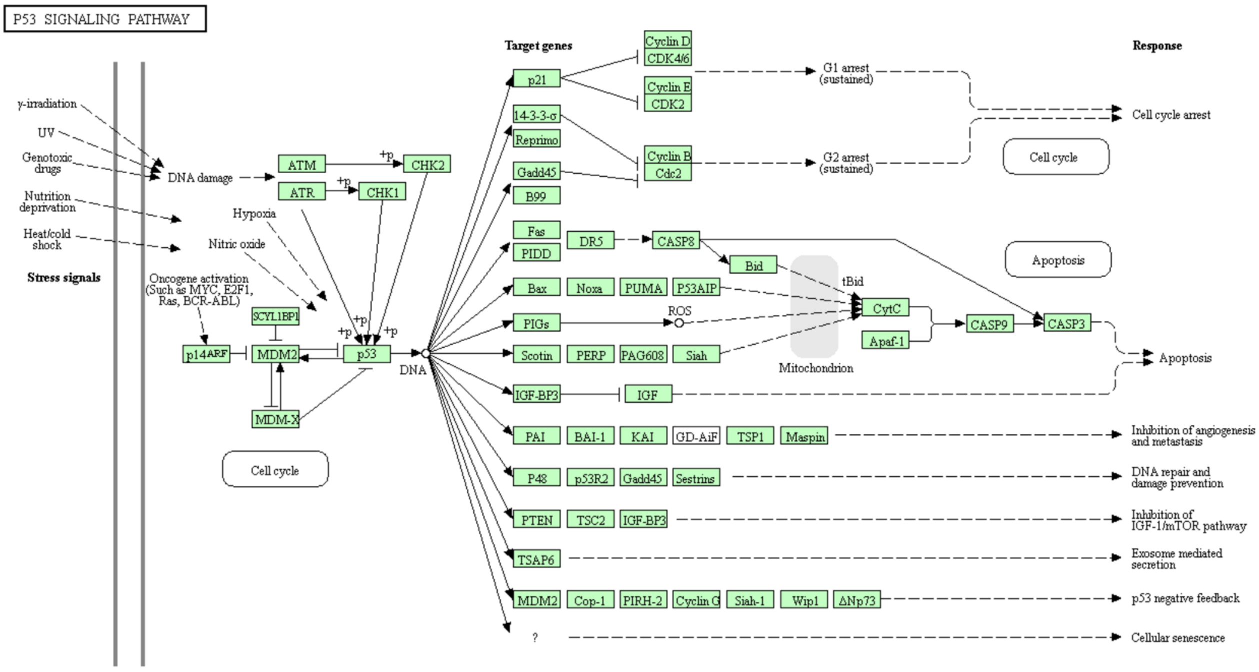
When you now select “Show biochemical pathway in browser”, you will be taken to the KEGG pathway where this gene is found:

http://www.kegg.jp/pathway/hsa04010+TP53



From here you can click on the p53 signalling pathway to drill down further.

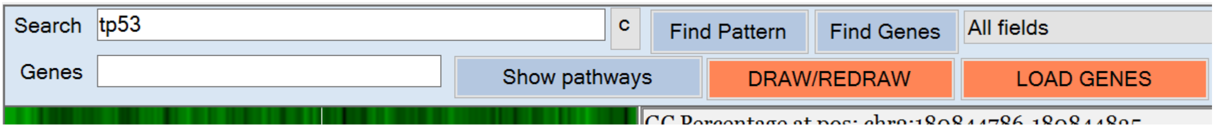
Take note that this only provides a link to online information and is not part of this software.



Searching or filtering for genes with the **pathway+** search term, will result in all the genes containing pathway information. This can also be obtained for all loaded genes (after clicking the "**LOAD GENES**" button).

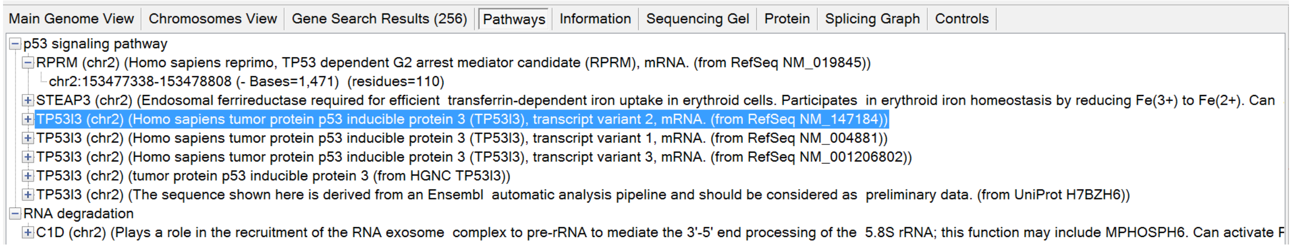
Any time you looking at genes on the genome, you can click the "**Show pathways**" button. This will load all displayed genes (which have pathways) into a pathway tree, which will group the genes below their corresponding biochemical pathway.

For example: Loading the genes for the current sequence and then clicking **"Show pathways"** after entering the search term TP53, will load only the genes related to TP53 in the pathway tree.



Search tp53

Genes



p53 signaling pathway
RPRM (chr2) (Homo sapiens reprimo, TP53 dependent G2 arrest mediator candidate (RPRM), mRNA. (from RefSeq NM_019845))
chr2:153477338-153478808 (- Bases=1,471) (residues=110)
STEAP3 (chr2) (Endosomal ferrireductase required for efficient transferrin-dependent iron uptake in erythroid cells. Participates in erythroid iron homeostasis by reducing Fe(3+) to Fe(2+). Can
TP53I3 (chr2) (Homo sapiens tumor protein p53 inducible protein 3 (TP53I3), transcript variant 2, mRNA. (from RefSeq NM_147184))
TP53I3 (chr2) (Homo sapiens tumor protein p53 inducible protein 3 (TP53I3), transcript variant 1, mRNA. (from RefSeq NM_004881))
TP53I3 (chr2) (Homo sapiens tumor protein p53 inducible protein 3 (TP53I3), transcript variant 3, mRNA. (from RefSeq NM_001206802))
TP53I3 (chr2) (tumor protein p53 inducible protein 3 (from HGNC TP53I3))
TP53I3 (chr2) (The sequence shown here is derived from an Ensembl automatic analysis pipeline and should be considered as preliminary data. (from UniProt H7BZH6))
RNA degradation
C1D (chr2) (Plays a role in the recruitment of the RNA exosome complex to pre-rRNA to mediate the 3'-5' end processing of the 5.8S rRNA; this function may include MPHOSPH6. Can activate F

From here you can:

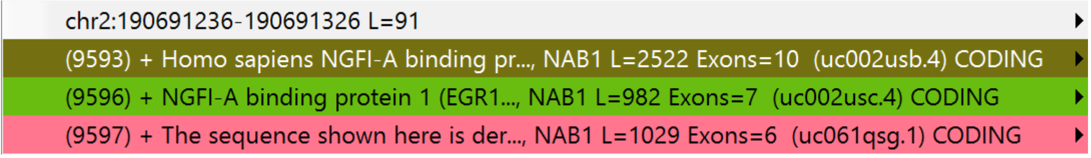
- double click on the genes to navigate to them
- double click on the pathway to filter only the genes containing those pathways and navigate to the first one of those
- Right click on the gene to get the gene context menu in order to execute all the usual actions.

Show Protein Sequence

All genes which code for proteins, are marked in the context menus with **CODING** at the end.

This indicates that the UCSC annotation contains the coding sequence or ORF position information indicating where translation begins and ends.

One example



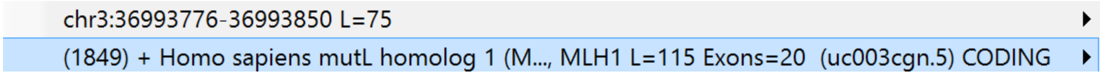
chr2:190691236-190691326 L=91

(9593) + Homo sapiens NGFI-A binding pr..., NAB1 L=2522 Exons=10 (uc002usb.4) CODING

(9596) + NGFI-A binding protein 1 (EGR1..., NAB1 L=982 Exons=7 (uc002usc.4) CODING

(9597) + The sequence shown here is der..., NAB1 L=1029 Exons=6 (uc061qsg.1) CODING

and another

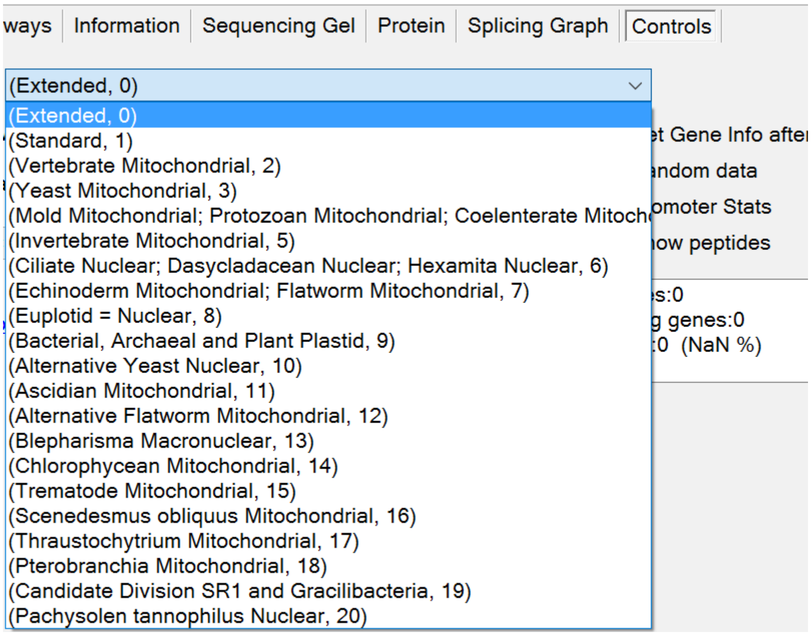


chr3:36993776-36993850 L=75

(1849) + Homo sapiens mutL homolog 1 (M..., MLH1 L=115 Exons=20 (uc003cgn.5) CODING

When you select the **"Show Protein Sequence"** menu option, the software will perform essentially the same process that happens in a cell. It will join all the bases in the exons (like the Spliceosome) and only extract the coding sequence, after which it will perform protein translation as it would happen in the Ribosome.

The Genetic Codon table it will use is the one selected below the "Controls" tab:



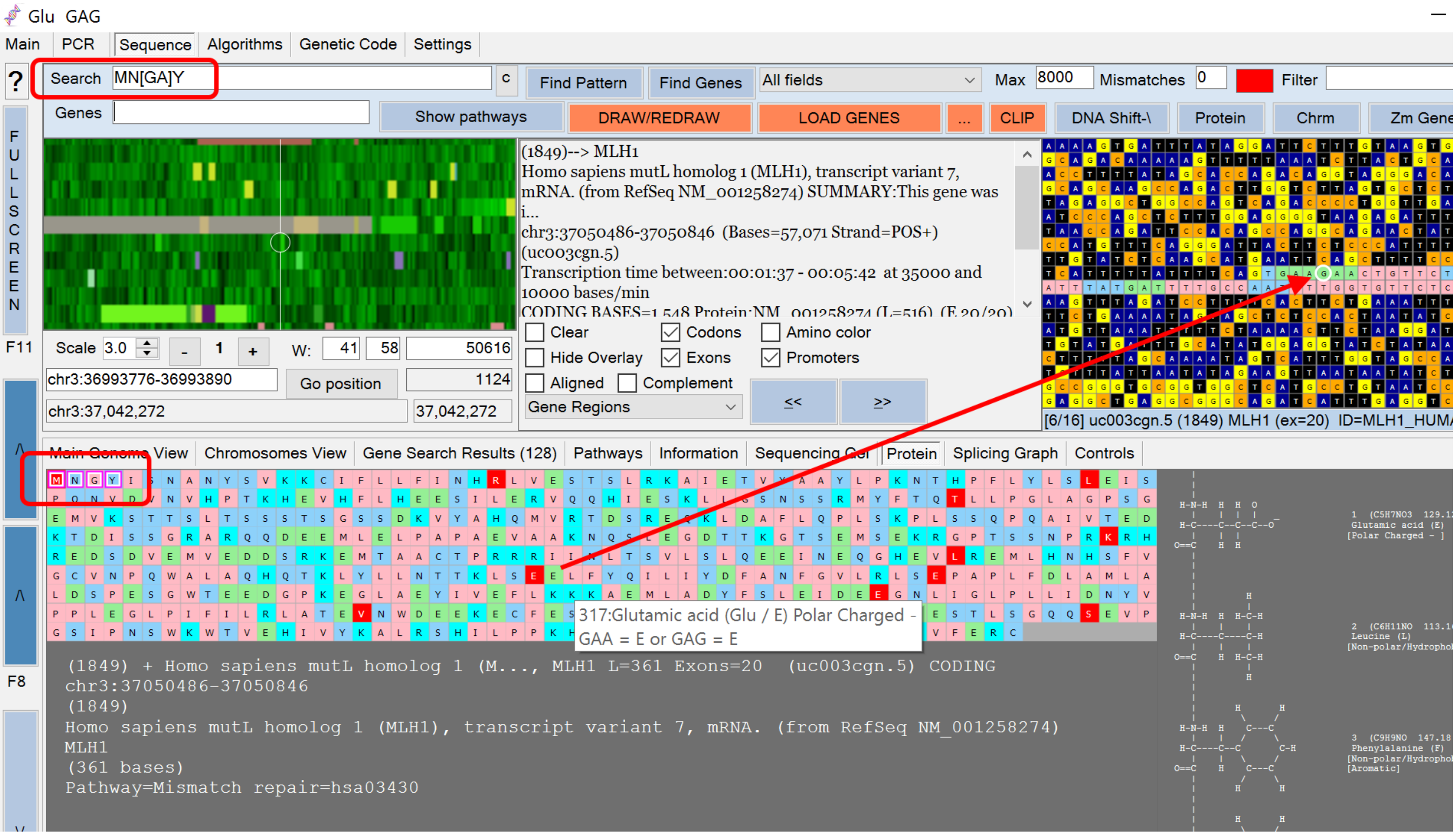
(Extended, 0)

(Extended, 0)
(Standard, 1)
(Vertebrate Mitochondrial, 2)
(Yeast Mitochondrial, 3)
(Mold Mitochondrial; Protozoan Mitochondrial; Coelenterate Mitoch
(Invertebrate Mitochondrial, 5)
(Ciliate Nuclear; Dasycladacean Nuclear; Hexamita Nuclear, 6)
(Echinoderm Mitochondrial; Flatworm Mitochondrial, 7)
(Euplotid = Nuclear, 8)
(Bacterial, Archaeal and Plant Plastid, 9)
(Alternative Yeast Nuclear, 10)
(Ascidian Mitochondrial, 11)
(Alternative Flatworm Mitochondrial, 12)
(Blepharisma Macronuclear, 13)
(Chlorophycean Mitochondrial, 14)
(Trematode Mitochondrial, 15)
(Scenedesmus obliquus Mitochondrial, 16)
(Thraustochytrium Mitochondrial, 17)
(Pterobranchia Mitochondrial, 18)
(Candidate Division SR1 and Gracilibacteria, 19)
(Pachysolen tannophilus Nuclear, 20)

The amino acid primary structure sequence will then be displayed in the **Protein** tab.

This action can be obtained from the context menu, or by double clicking on the gene in the **Main Genome View**. When you have highlighted the gene using cursor keys (which works in both the **Main Genome View** or the **DNA View**) you can also press **F5** which will always draw or redraw the current protein view. If you want to search for specific peptide signalling sequences in the protein, you can enter the amino acid letters in the **Search field** and it will be highlighted in the **Protein View**.

Something else to observe is that the **Protein View** and **DNA View** is navigationally linked. When you hover the mouse over a specific amino acid, the **DNA** View navigates and centres on the corresponding codon, which is coloured with the same colour when the "**Codons**" checkbox is checked. The display is scrollable when the entire sequence does not fit.



Search MN[GA]Y

Max 8000 Mismatches 0

Genes

Scale 3.0 - 1 + W: 41 58 50616

chr3:36993776-36993890 1124

chr3:37,042,272 37,042,272

(1849)--> MLH1
Homo sapiens mutL homolog 1 (MLH1), transcript variant 7, mRNA. (from RefSeq NM_001258274) SUMMARY:This gene was i...
chr3:37050486-37050846 (Bases=57,071 Strand=POS+)
(uc003cgn.5)
Transcription time between:00:01:37 - 00:05:42 at 35000 and 10000 bases/min
CODING BASES=1 548 Protein:NM_001258274 (L=516) (E 20/20)

□ Clear ☑ Codons □ Amino color
□ Hide Overlay ☑ Exons ☑ Promoters
□ Aligned □ Complement
Gene Regions

[6/16] uc003cgn.5 (1849) MLH1 (ex=20) ID=MLH1_HUM

317:Glutamic acid (Glu / E) Polar Charged -
GAA = E or GAG = E

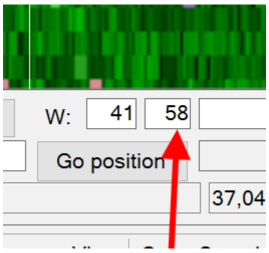
```
(1849) + Homo sapiens mutL homolog 1 (M..., MLH1 L=361 Exons=20  (uc003cgn.5) CODING
chr3:37050486-37050846
(1849)
Homo sapiens mutL homolog 1 (MLH1), transcript variant 7, mRNA. (from RefSeq NM_001258274)
MLH1
(361 bases)
Pathway=Mismatch repair=hsa03430
```

1 (C5H7NO3 129.1 Glutamic acid (E) [Polar Charged -]

2 (C6H11NO 113.1 Leucine (L) [Non-polar/Hydropho

3 (C9H9NO 147.18 Phenylalanine (F) [Non-polar/Hydropho [Aromatic]

You can also change the number of amino acids in each line, by using the **[** and **]** keys, or by changing it in the text box:

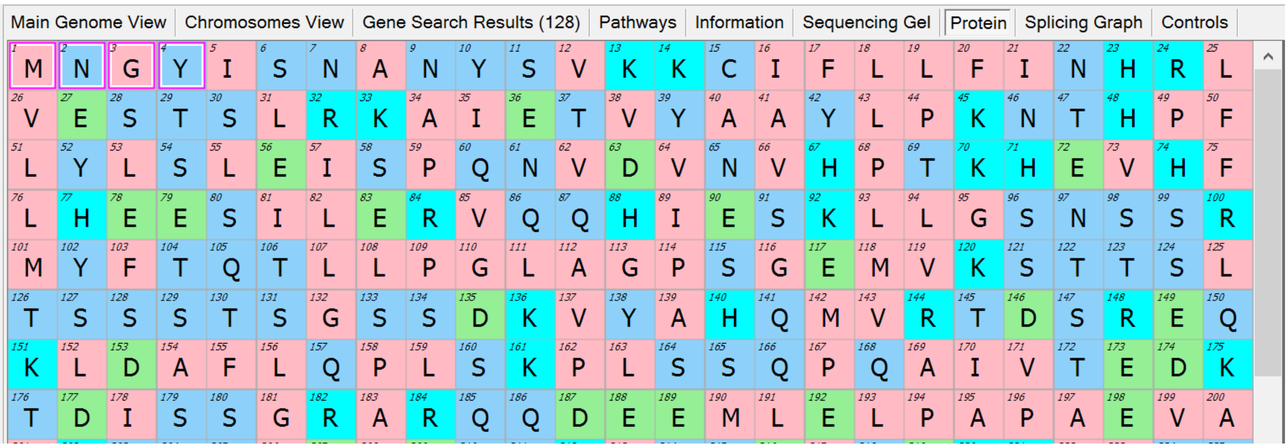


W: 41 58

Go position

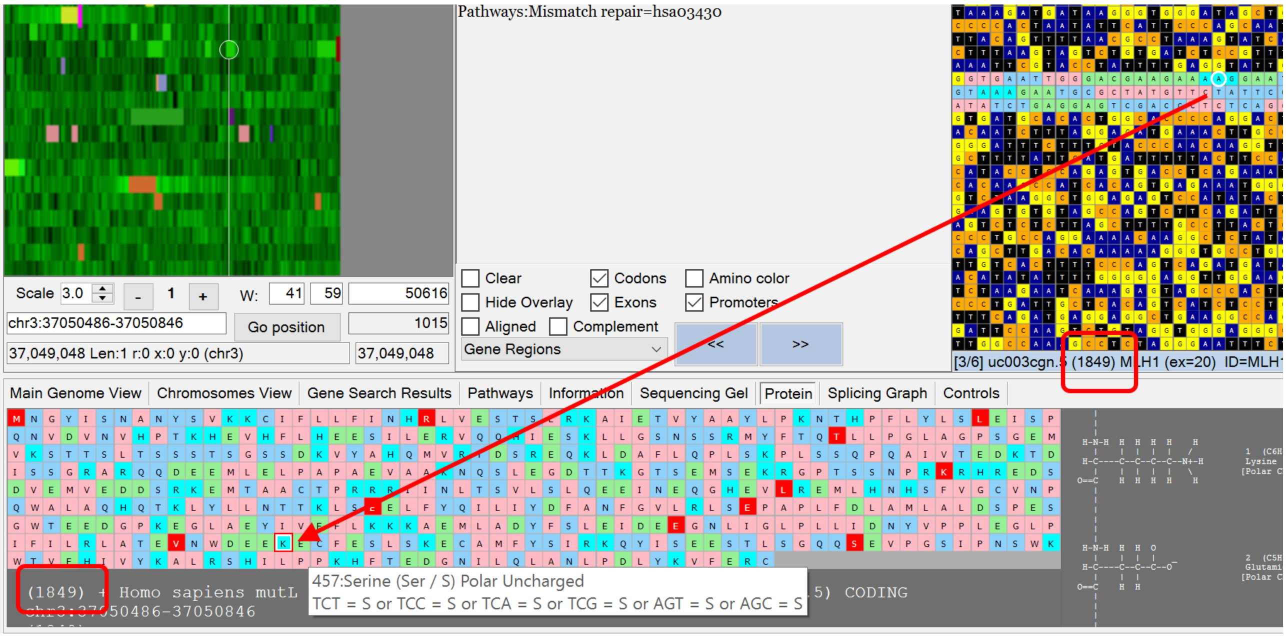
37,04

When the amino acids per line becomes small enough, the display also starts to display the amino acid index (given that the option Show Amino Acid Numbers is enabled in the settings), which you also get when you hover over any amino acid in the display .



1 M	2 N	3 G	4 Y	5 I	6 S	7 N	8 A	9 N	10 Y	11 S	12 V	13 K	14 K	15 C	16 I	17 F	18 L	19 L	20 F	21 I	22 N	23 H	24 R	25 L
26 V	27 E	28 S	29 T	30 S	31 L	32 R	33 K	34 A	35 I	36 E	37 T	38 V	39 Y	40 A	41 A	42 Y	43 L	44 P	45 K	46 N	47 T	48 H	49 P	50 F
51 L	52 Y	53 L	54 S	55 L	56 E	57 I	58 S	59 P	60 Q	61 N	62 V	63 D	64 V	65 N	66 V	67 H	68 P	69 T	70 K	71 H	72 E	73 V	74 H	75 F
76 L	77 H	78 E	79 E	80 S	81 I	82 L	83 E	84 R	85 V	86 Q	87 Q	88 H	89 I	90 E	91 S	92 K	93 L	94 L	95 G	96 S	97 N	98 S	99 S	100 R
101 M	102 Y	103 F	104 T	105 Q	106 T	107 L	108 L	109 P	110 G	111 L	112 A	113 G	114 P	115 S	116 G	117 E	118 M	119 V	120 K	121 S	122 T	123 T	124 S	125 L
126 T	127 S	128 S	129 S	130 T	131 S	132 G	133 S	134 S	135 D	136 K	137 V	138 Y	139 A	140 H	141 Q	142 M	143 V	144 R	145 T	146 D	147 S	148 R	149 E	150 Q
151 K	152 L	153 D	154 A	155 F	156 L	157 Q	158 P	159 L	160 S	161 K	162 P	163 L	164 S	165 S	166 Q	167 P	168 Q	169 A	170 I	171 V	172 T	173 E	174 D	175 K
176 T	177 D	178 I	179 S	180 S	181 G	182 R	183 A	184 R	185 Q	186 Q	187 D	188 E	189 E	190 M	191 L	192 E	193 L	194 P	195 A	196 P	197 A	198 E	199 V	200 A

When you are centring the **DNA View** on a different codon, and the **Protein View** is currently displaying the exact same gene's protein, then the corresponding amino acid is highlighted in the **Protein View** as follows:


Pathways:Mismatch repair=hsa03430

Scale 3.0 - 1 + W: 41 59 50616

chr3:37050486-37050846 Go position 1015

37,049,048 Len:1 r:0 x:0 y:0 (chr3) 37,049,048

□ Clear ☑ Codons □ Amino color
□ Hide Overlay ☑ Exons ☑ Promoters
□ Aligned □ Complement

Gene Regions

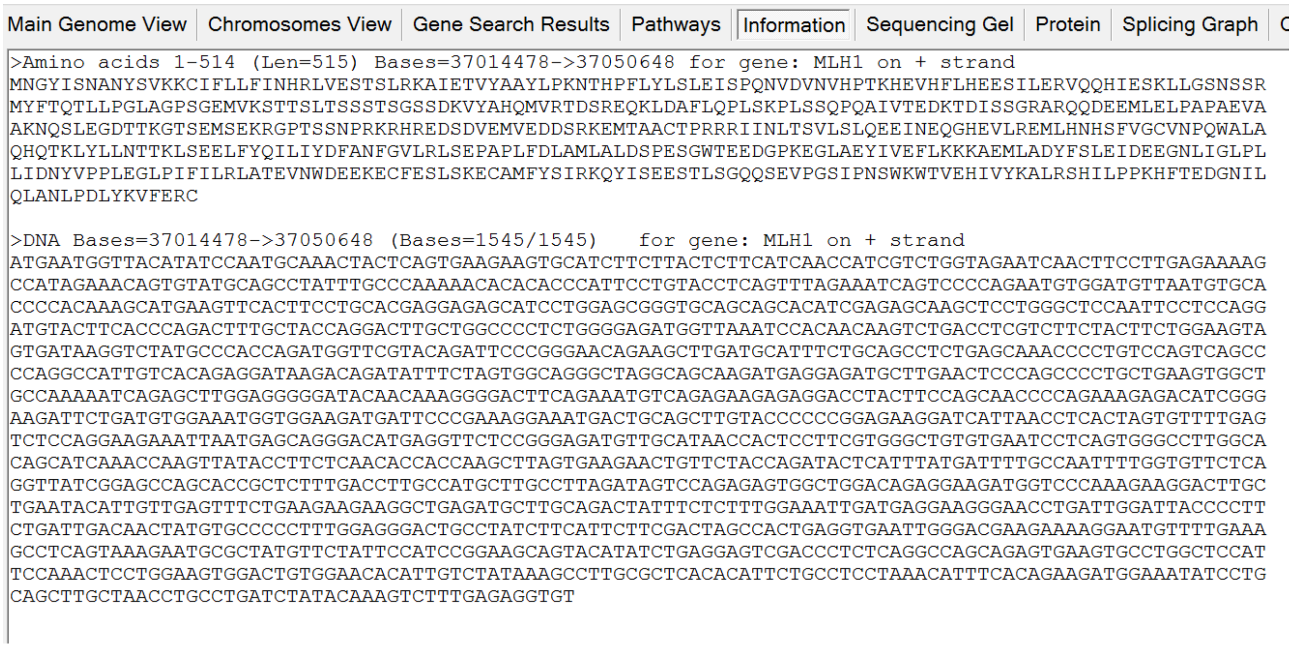
[3/6] uc003cgn.5 (1849) MLH1 (ex=20) ID=MLH1

457:Serine (Ser / S) Polar Uncharged
TCT = S or TCC = S or TCA = S or TCG = S or AGT = S or AGC = S

(1849) + Homo sapiens mutL ... 5) CODING
chr3:37050486-37050846

When there are multiple overlapping genes in the **DNA View** you can step through them by using the + **and** – keys.

When you want to copy and paste a specific range of amino acids you can copy them by double clicking on the first and last amino acid and the result will be copied to the **Information** tab and clipboard. Both the DNA coding sequence and the amino acid sequence will be shown (but only for the region YOU selected).



```
>Amino acids 1-514 (Len=515) Bases=37014478->37050648 for gene: MLH1 on + strand
MNGYISNANYSVKKCIFLLFINHRLVESTSLRKAIETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLLGSNSSR
MYFTQTLLPGLAGPSGEMVKSTTSLTSSSTSGSSDKVYAHQMVRTDSREQKLDAFLQPLSKPLSSQPQAIVTEDKTDISSGRARQQDEEMLELPAPAEVA
AKNQSLEGDTTKGTSEMSEKRGPTSSNPRKRHREDSDVEMVEDDSRKEMTAACTPRRRIINLTSVLSLQEEINEQGHEVLREMLHNHSFVGCVNPQWALA
QHQTKLYLLNTTKLSEELFYQILIYDFANFGVLRLSEPAPLFDLAMLALDSPESGWTEEDGPKEGLAEYIVEFLKKKAEMLADYFSLEIDEEGNLIGLPL
LIDNYVPPLEGLPIFILRLATEVNWDEEKECFESLSKECAMFYSIRKQYISEESTLSGQQSEVPGSIPNSWKWTVEHIVYKALRSHILPPKHFTEDGNIL
QLANLPDLYKVFERC

>DNA Bases=37014478->37050648 (Bases=1545/1545)   for gene: MLH1 on + strand
ATGAATGGTTACATATCCAATGCAAACTACTCAGTGAAGAAGTGCATCTTCTTACTCTTCATCAACCATCGTCTGGTAGAATCAACTTCCTTGAGAAAAG
CCATAGAAACAGTGTATGCAGCCTATTTGCCCAAAAACACACACCCATTCCTGTACCTCAGTTTAGAAATCAGTCCCCAGAATGTGGATGTTAATGTGCA
CCCCACAAAGCATGAAGTTCACTTCCTGCACGAGGAGAGCATCCTGGAGCGGGTGCAGCAGCACATCGAGAGCAAGCTCCTGGGCTCCAATTCCTCCAGG
ATGTACTTCACCCAGACTTTGCTACCAGGACTTGCTGGCCCCTCTGGGGAGATGGTTAAATCCACAACAAGTCTGACCTCGTCTTCTACTTCTGGAAGTA
GTGATAAGGTCTATGCCCACCAGATGGTTCGTACAGATTCCCGGGAACAGAAGCTTGATGCATTTCTGCAGCCTCTGAGCAAACCCCTGTCCAGTCAGCC
CCAGGCCATTGTCACAGAGGATAAGACAGATATTTCTAGTGGCAGGGCTAGGCAGCAAGATGAGGAGATGCTTGAACTCCCAGCCCCTGCTGAAGTGGCT
GCCAAAAATCAGAGCTTGGAGGGGGATACAACAAAGGGGACTTCAGAAATGTCAGAGAAGAGAGGACCTACTTCCAGCAACCCCAGAAAGAGACATCGGG
AAGATTCTGATGTGGAAATGGTGGAAGATGATTCCCGAAAGGAAATGACTGCAGCTTGTACCCCCCGGAGAAGGATCATTAACCTCACTAGTGTTTTGAG
TCTCCAGGAAGAAATTAATGAGCAGGGACATGAGGTTCTCCGGGAGATGTTGCATAACCACTCCTTCGTGGGCTGTGTGAATCCTCAGTGGGCCTTGGCA
CAGCATCAAACCAAGTTATACCTTCTCAACACCACCAAGCTTAGTGAAGAACTGTTCTACCAGATACTCATTTATGATTTTGCCAATTTTGGTGTTCTCA
GGTTATCGGAGCCAGCACCGCTCTTTGACCTTGCCATGCTTGCCTTAGATAGTCCAGAGAGTGGCTGGACAGAGGAAGATGGTCCCAAAGAAGGACTTGC
TGAATACATTGTTGAGTTTCTGAAGAAGAAGGCTGAGATGCTTGCAGACTATTTCTCTTTGGAAATTGATGAGGAAGGGAACCTGATTGGATTACCCCTT
CTGATTGACAACTATGTGCCCCCTTTGGAGGGACTGCCTATCTTCATTCTTCGACTAGCCACTGAGGTGAATTGGGACGAAGAAAAGGAATGTTTTGAAA
GCCTCAGTAAAGAATGCGCTATGTTCTATTCCATCCGGAAGCAGTACATATCTGAGGAGTCGACCCTCTCAGGCCAGCAGAGTGAAGTGCCTGGCTCCAT
TCCAAACTCCTGGAAGTGGACTGTGGAACACATTGTCTATAAAGCCTTGCGCTCACACATTCTGCCTCCTAAACATTTCACAGAAGATGGAAATATCCTG
CAGCTTGCTAACCTGCCTGATCTATACAAAGTCTTTGAGAGGTGT
```

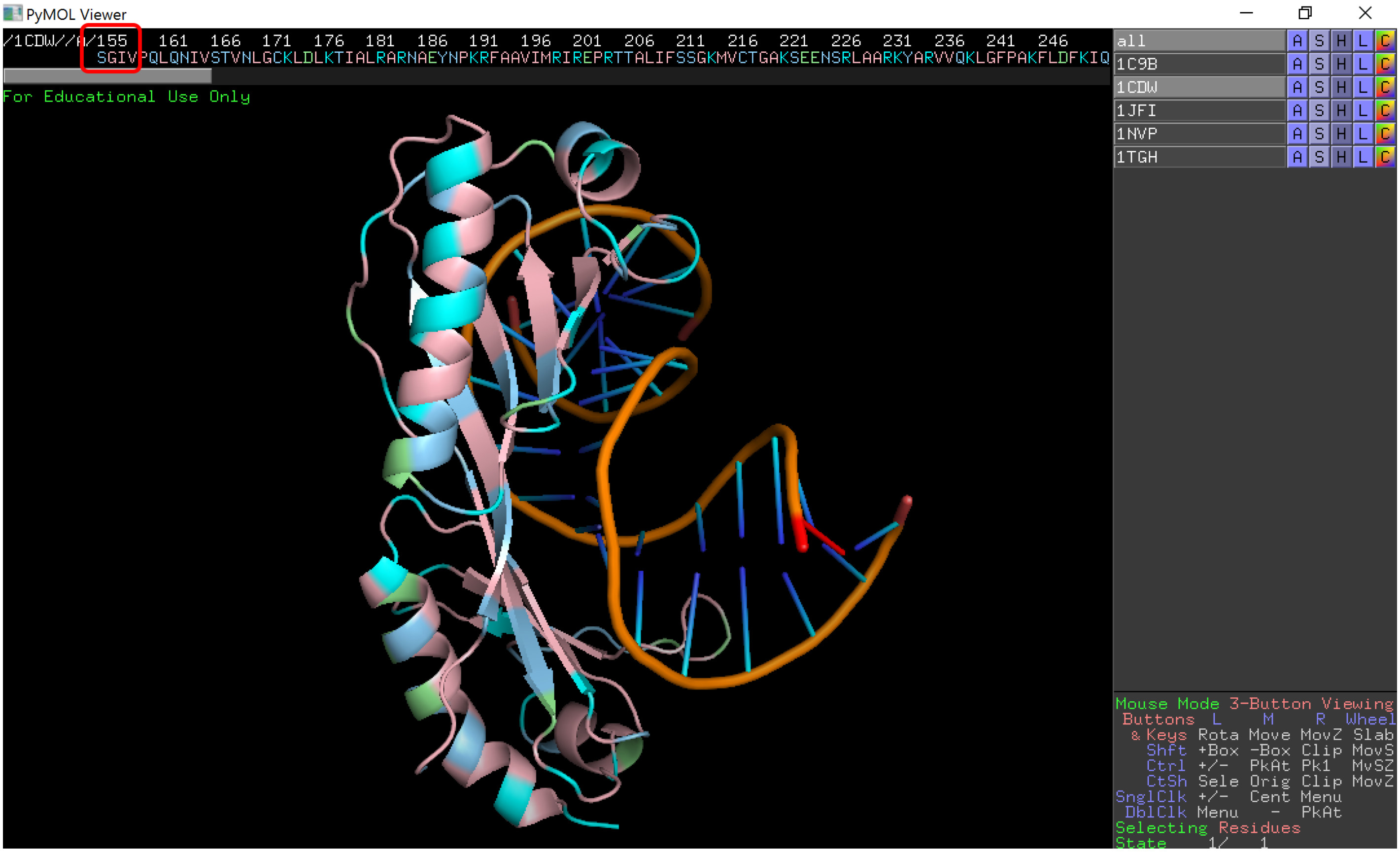
Show PDB Structure in PYMOL

Many of the gene annotations from the UCSC table data comes with the Protein Database (PDB) codes for protein coding genes. These PDB files contains the 3D Crystallography structure obtained for these genes.

When you have the open source Molecular visualization software PYMOL installed, the Visual Genome Browser will firstly download a set number of PDB files referenced in the gene annotations, create a hydrophobic colour scheme and then launch the PDB files in PYMOL.

“PyMOL is an open-source, user-sponsored, molecular visualization system created by Warren Lyford DeLano and commercialized initially by DeLano Scientific LLC, which was a private software company dedicated to creating useful tools that become universally accessible to scientific and educational communities. It is currently commercialized by Schrödinger, Inc. PyMOL can produce high-quality 3D images of small molecules and biological macromolecules, such as proteins.”

When one selects the TBP (TATA-box binding protein, which is a transcription factor responsible for recruiting RNA Polymerase II to genes containing the TATA sequence promoters) this option shows the Protein structure determined for this protein in PYMOL on the following page.



```
/1CDW//A/155  161  166  171  176  181  186  191  196  201  206  211  216  221  226  231  236  241  246
         SGIVPQLQNIVSTVNLGCKLDLKTIALRARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEENSRLAARKYARVVQKLGFPAKFLDFKIQ
For Educational Use Only
```

```
Mouse Mode 3-Button Viewing
 Buttons  L    M    R  Wheel
  & Keys Rota Move MovZ Slab
    Shft +Box -Box Clip MovS
    Ctrl +/-  PkAt Pk1  MvSZ
    CtSh Sele Orig Clip MovZ
SnglClk +/-  Cent Menu
 DblClk Menu   -  PkAt
Selecting Residues
State     1/   1
```

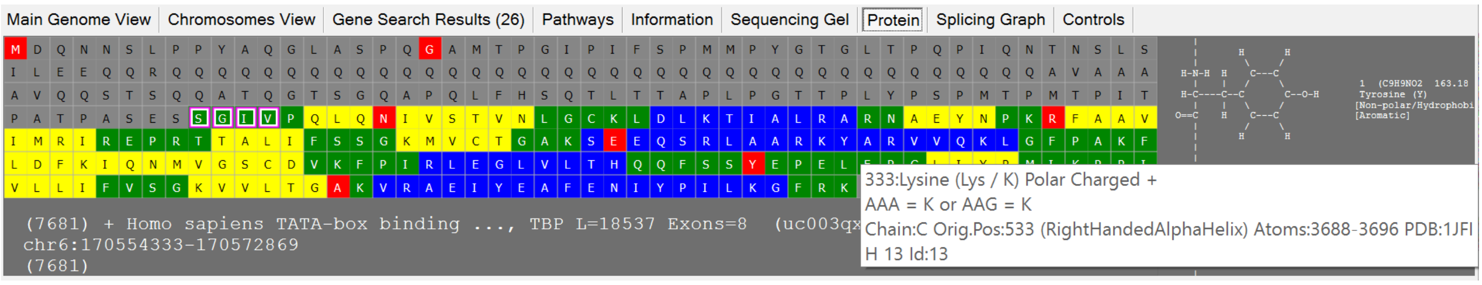
Illustration 52: TBP (TATA Box binding protein crystal structure)

Show Protein sequence with PDB Secondary Structure

I was looking for a way to incorporate the secondary structure of a protein (alpha helix, beta sheet, loop etc.) into the 2D protein view and I realised that I could try to do an alignment between the Protein Sequence View amino acids and the amino acids in the PDB files from the Protein Data Bank and then extract the secondary structure information from the amino acids in the PDB files and then associate that information with the corresponding amino acids in the Protein Sequence view. In this was I did not have to resort to trying to predict the secondary structure using molecular angles etc., but I could extract it directly from the crystallography structures.

http://www.rcsb.org/pdb/home/home.do

You obtain this by selecting **Show Protein Sequence with PDB Sequence** Alignment. That results in a protein display which looks as follows (Notice the correspondence between the first amino acids of the crystal structure and the highlighted amino acids SGIV in the image below):



```
M D Q N N S L P P Y A Q G L A S P Q G A M T P G I P I F S P M M P Y G T G L T P Q P I Q N T N S L S
I L E E Q Q R Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q A V A A A
A V Q Q S T S Q Q A T Q G T S G Q A P Q L F H S Q T L T T A P L P G T T P L Y P S P M T P M T P I T
P A T P A S E S S G I V P Q L Q N I V S T V N L G C K L D L K T I A L R A R N A E Y N P K R F A A V
I M R I R E P R T T A L I F S S G K M V C T G A K S E E Q S R L A A R K Y A R V V Q K L G F P A K F
L D F K I Q N M V G S C D V K F P I R L E G L V L T H Q Q F S S Y E P E L
V L L I F V S G K V V L T G A K V R A E I Y E A F E N I Y P I L K G F R K
```

```
          H        H
           \      /
 H-N-H  H   C---C
   |    |  /      \
 H-C----C--C        C--O-H
   |    |  \      /
 O==C   H   C---C
           /      \
          H        H
```

```
1  (C9H9NO2  163.18
Tyrosine (Y)
[Non-polar/Hydrophobi
[Aromatic]
```

```
(7681) + Homo sapiens TATA-box binding ..., TBP L=18537 Exons=8  (uc003qx
chr6:170554333-170572869
(7681)
```

333:Lysine (Lys / K) Polar Charged +
AAA = K or AAG = K
Chain:C Orig.Pos:533 (RightHandedAlphaHelix) Atoms:3688-3696 PDB:1JFI
H 13 Id:13

The colours represent different secondary structures:
Blue = alpha helix
Green = Loop
Yellow = Beta sheet
Gray = Residues for which no alignment could be found with the PDB (This also gives us an indication for which part of the protein the crystal structure was determined)

Red markers for amino acids at start of exons

The Red blocks are markers for the amino acids which are at the start of each exon. This is sometimes useful to have in order to see how subsequent and similar protein coding sequences, which are part of the same protein coding gene, can be found so far apart and separated by many intron bases. The display of amino acids at the start of exons as red blocks can also be switched off in the Settings. The fact that the amino acids at the start of each exon is marked in red, will allow you to get some insight into the different segments of the protein which might be spliced out for alternatively spliced transcripts.

I referred to these patterns in proteins on my blog:
http://splicejunction.blogspot.com.au/2015/08/protein-scrabble.html

Show Splicing Graph

I wanted to get a way to ”graphically visualise” the introns and exons of a transcribed gene as there can sometimes be hundreds of exons which are spliced together to produce the final mature RNA transcript.

I came up with the concept of a **Splicing Graph**. I knew from the theory on the Spliceosome, that proteins complexed with RNA is able to recognize the splice signals (GT....AG) and then correctly form a loop (Lariat) which is cut out before the exons are joined or ligated together.

I discussed this on my blog: http://splicejunction.blogspot.com.au/2015/09/spliced-genes-natures-model-hobby-kit_18.html

I decided to draw the introns which “loop out” as vertical lines which represent these loops **to scale** on the vertical axis of these graphs. Each line therefore represented a loop going downwards up to half the length of the removed intron and half the length back again. **The vertical lines therefore mimicked the real loop which is formed during splicing.** In a similar way, the lengths of the **exons** are represented **to scale** on the horizontal axis of the graph. In this way you can get a feeling of the length spacing of the exons, while simultaneously showing the differences in the intron lengths as vertical lines.

When I create a **Splicing Graph** of the TTN gene transcript : RefSeq NM_133378
ucsc Id: uc021vsy.4

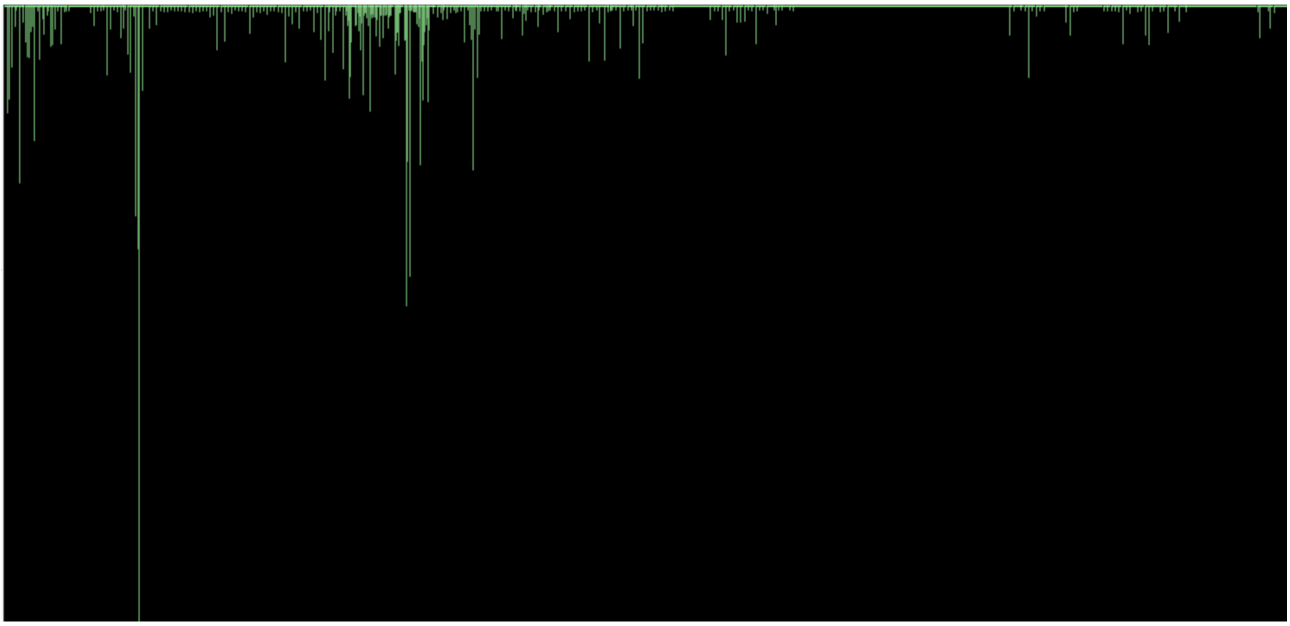


Illustration 53: Splicing Graph for the TTN gene with a total of 312 exons and 311 Intron which are represented by the vertical lines

When I do the same for a gene for Collagen, which has less exons (66 in total) I get the following Splicing Graph.

The spacing of the vertical lines are determined by the lengths of the exons between them.

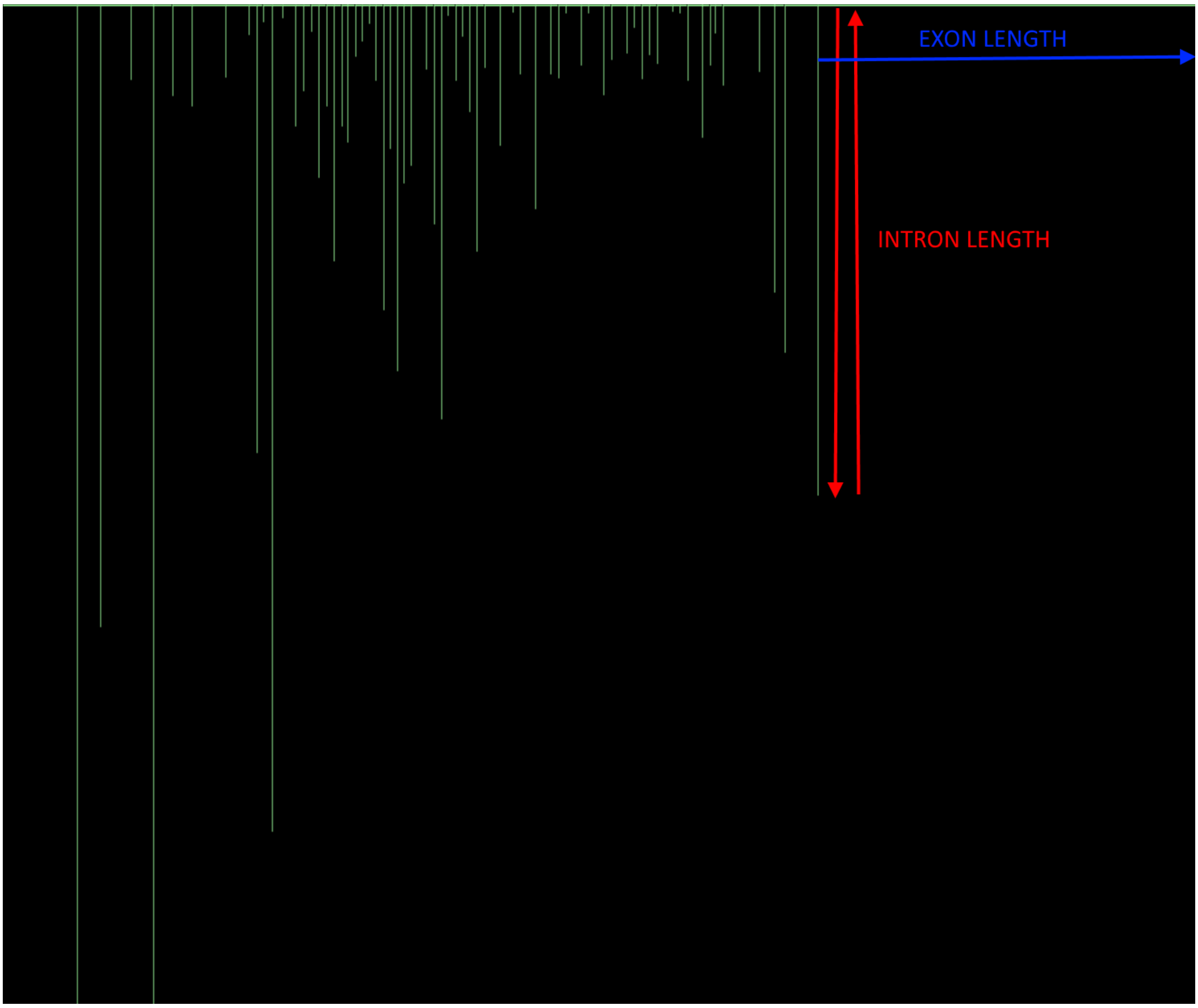


Illustration 55: Splicing Graph of the gene for collagen type V alpha 1 (COL5A1). Intron loops are on the vertical axis while exons are the spacing between the vertical lines on the horizontal axis.

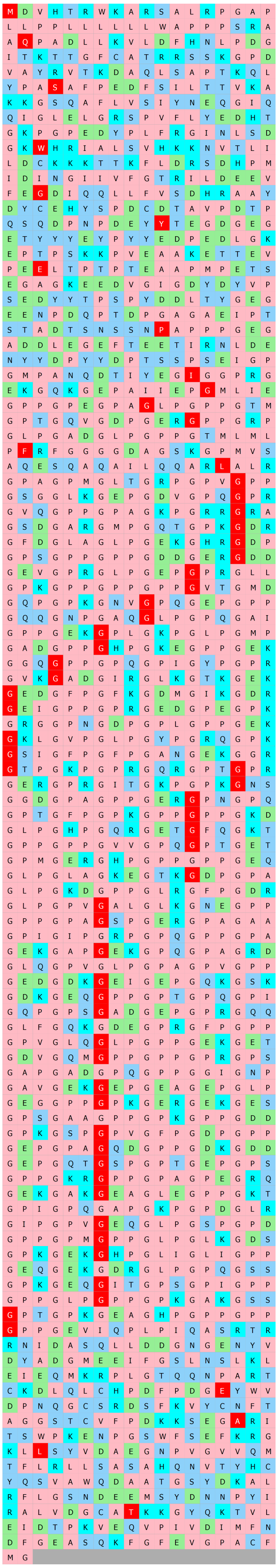


Illustration 56: Protein View for COL5A1 gene at width 18 residues

If one is rather interested in the exon lengths, it is possible to change the splicing view setting in order to rather show the exon lengths. For exactly the same collagen protein here is a picture showing the exon lengths:

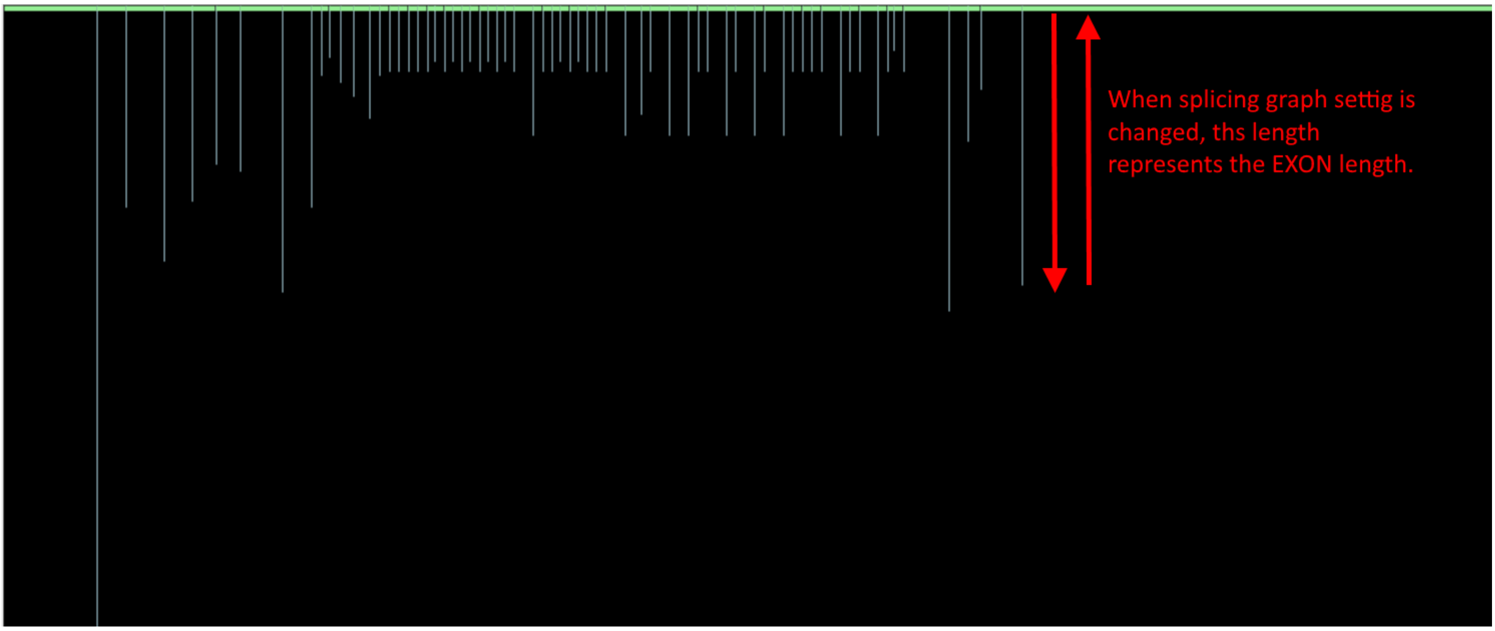
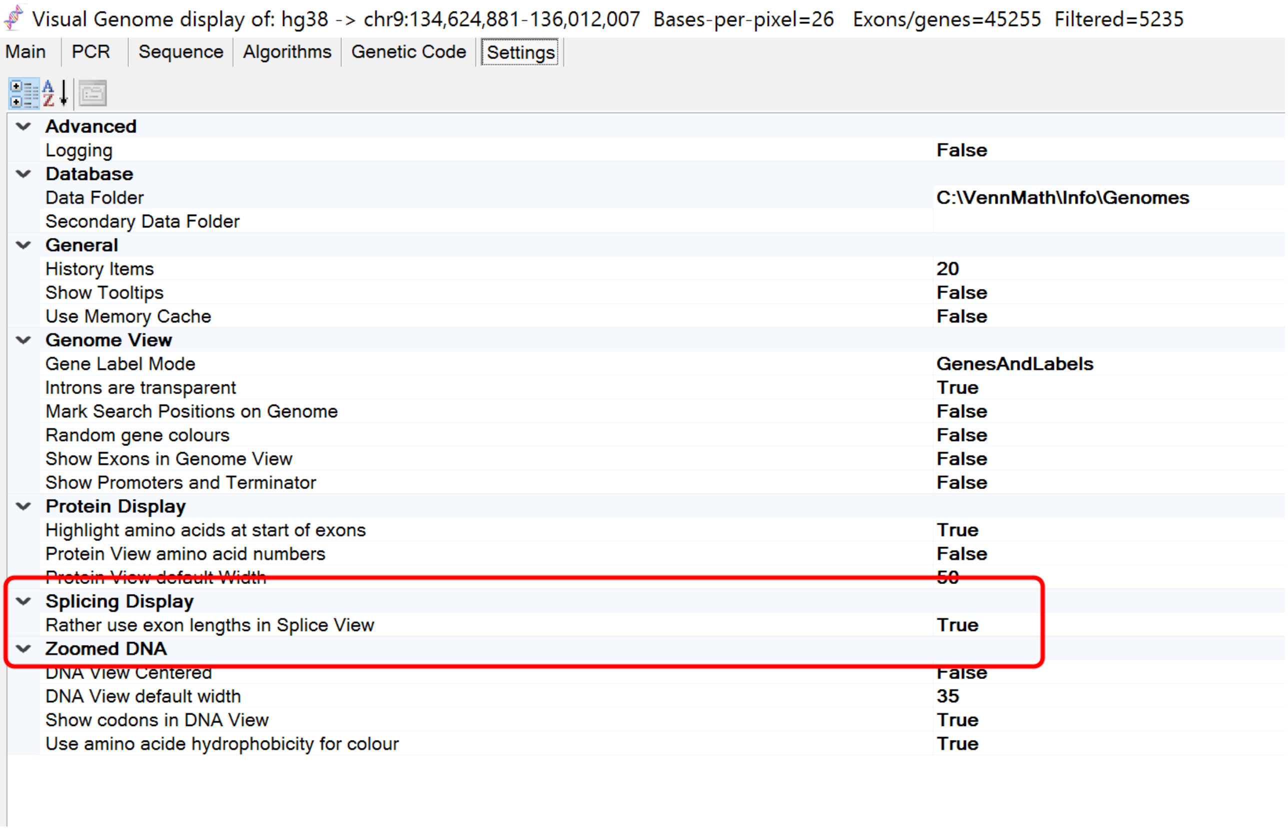

When splicing graph settig is changed, ths length represents the EXON length.

Illustration 57: Splicing Graph of the gene for collagen type V alpha 1 (COL5A1). In this case EXONS are on the vertical and horizontal axes. This provides a quick way to visualise the exons which makes up the coding sequence.

The setting can be changed in the **Settings tab**.


Visual Genome display of: hg38 -> chr9:134,624,881-136,012,007 Bases-per-pixel=26   Exons/genes=45255 Filtered=5235

Advanced	
Logging	False
Database	
Data Folder	C:\VennMath\Info\Genomes
Secondary Data Folder	
General	
History Items	20
Show Tooltips	False
Use Memory Cache	False
Genome View	
Gene Label Mode	GenesAndLabels
Introns are transparent	True
Mark Search Positions on Genome	False
Random gene colours	False
Show Exons in Genome View	False
Show Promoters and Terminator	False
Protein Display	
Highlight amino acids at start of exons	True
Protein View amino acid numbers	False
Protein View default Width	50
Splicing Display	
Rather use exon lengths in Splice View	True
Zoomed DNA	
DNA View Centered	False
DNA View default width	35
Show codons in DNA View	True
Use amino acide hydrophobicity for colour	True

For example, look how the marked amino acids at the start of each exon form a pattern in the **TTN** protein. The distribution of the exon lengths seems fall in a relatively narrow band throughout a big part of the protein. This is a block representation of only part of the whole protein.

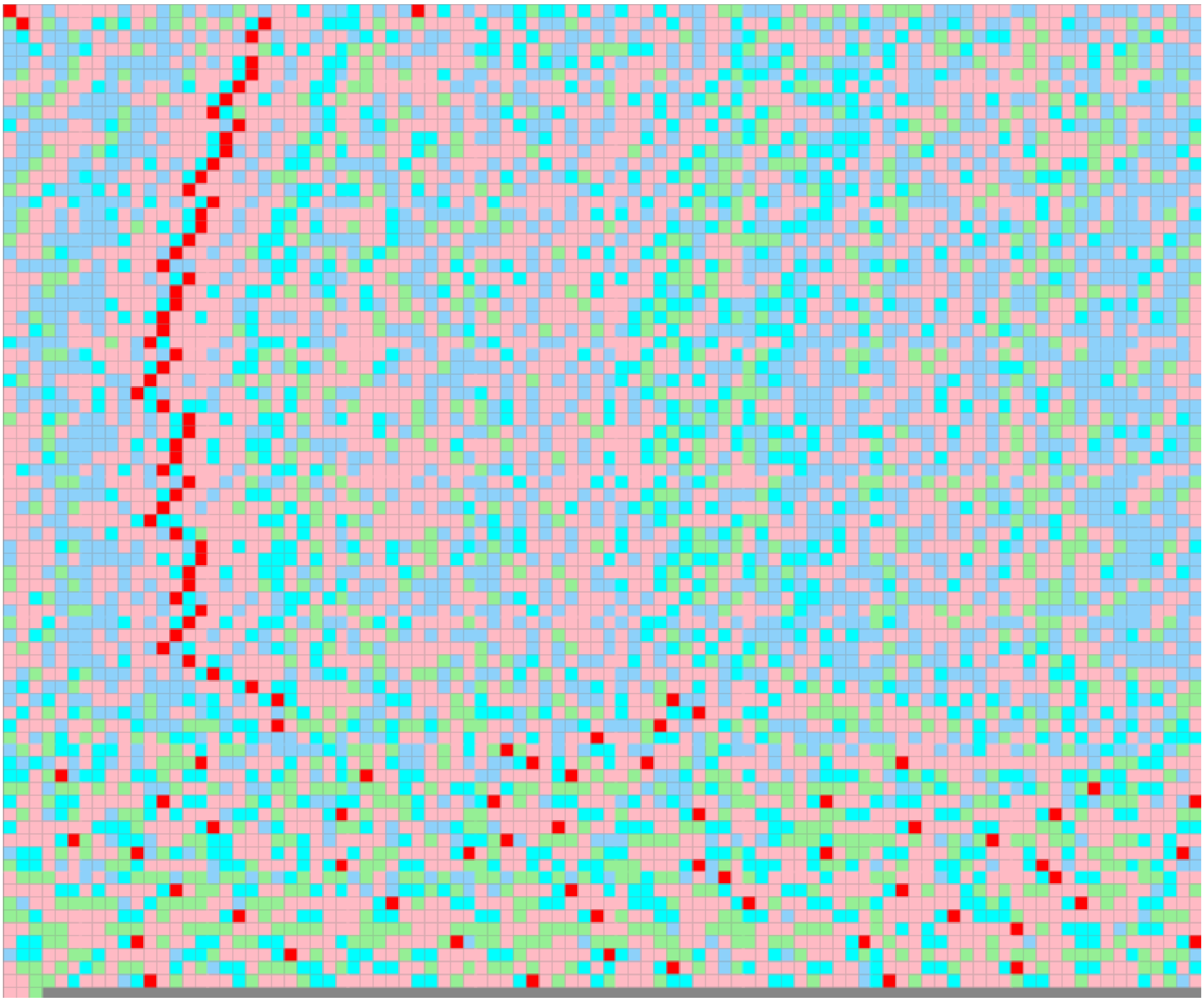


Illustration 58: Patterns in the protein view made by amino acids at the start of exons. (For the Titin TTN gene on chromosome 2)

The **TITIN** gene codes for a large abundant protein in the human body. The gene is quite probably the largest gene in the human genome. The protein effectively acts as **shock absorbers** for muscle fibres (sarcomeres) and needs to be structurally very long. In the picture above, remember that the pink areas represent **hydrophobic** residues, green negatively charged polar residues, blue uncharged polar residues and the cyan positively charged residues.

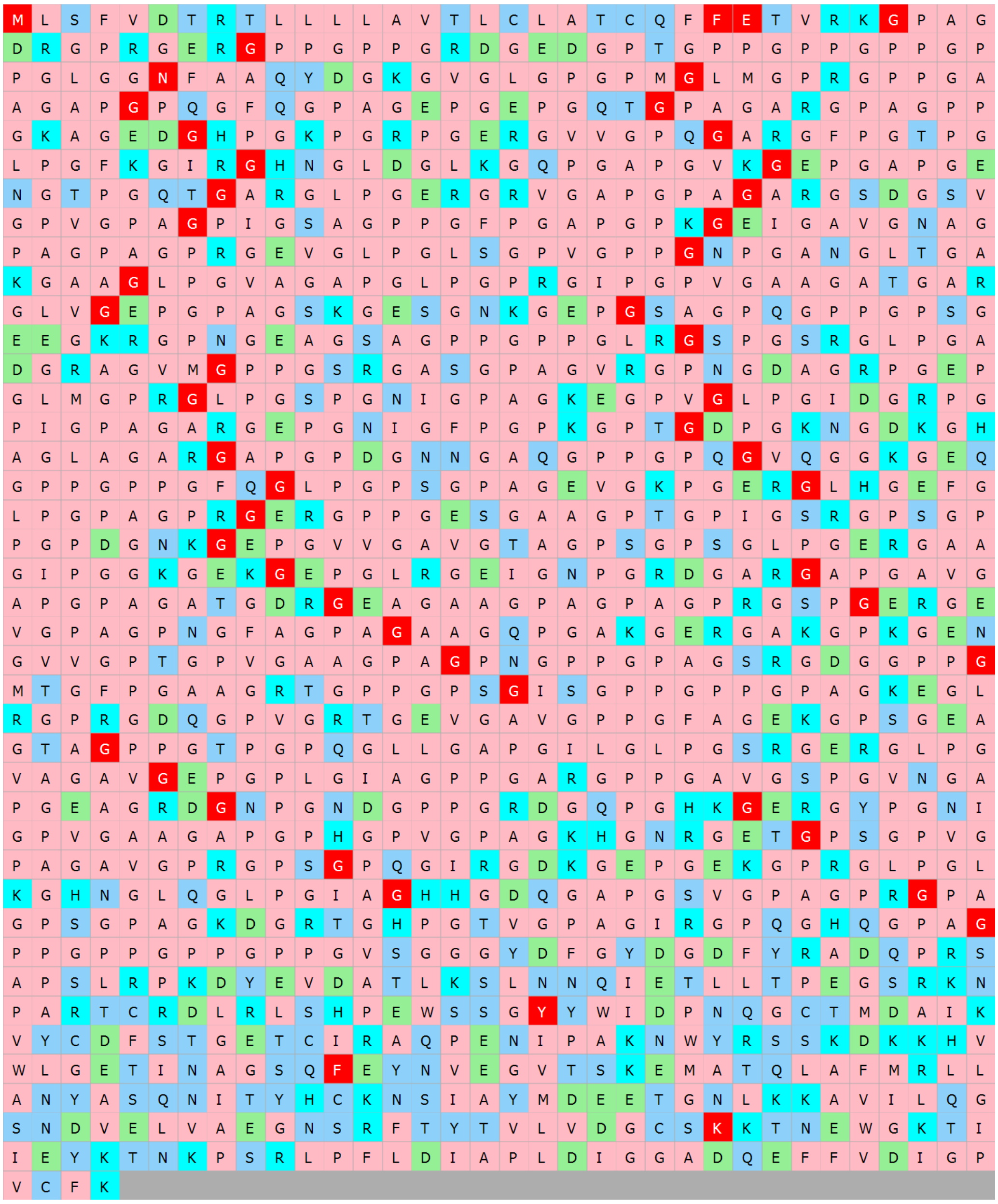
Here is an extract from the UCSC summary information for this gene:

*"**This gene encodes a large abundant protein of striated muscle**. The product of this gene is divided into two regions, a N-terminal I-band and a C-terminal A-band. ... An N-terminal Z-disc region and a C-terminal M-line region bind to the Z-line and M-line of the sarcomere, respectively, so that a single titin molecule spans half the length of a sarcomere. **Titin also contains binding sites for muscle associated proteins so it serves as an adhesion template for the assembly of contractile machinery in muscle cells.** It has also been identified as a structural protein for chromosomes. ... Mutations in this gene are associated with familial hypertrophic cardiomyopathy 9, and auto-antibodies to **titin** are produced in patients with the autoimmune disease scleroderma."*

Protein Sequence display of various genes

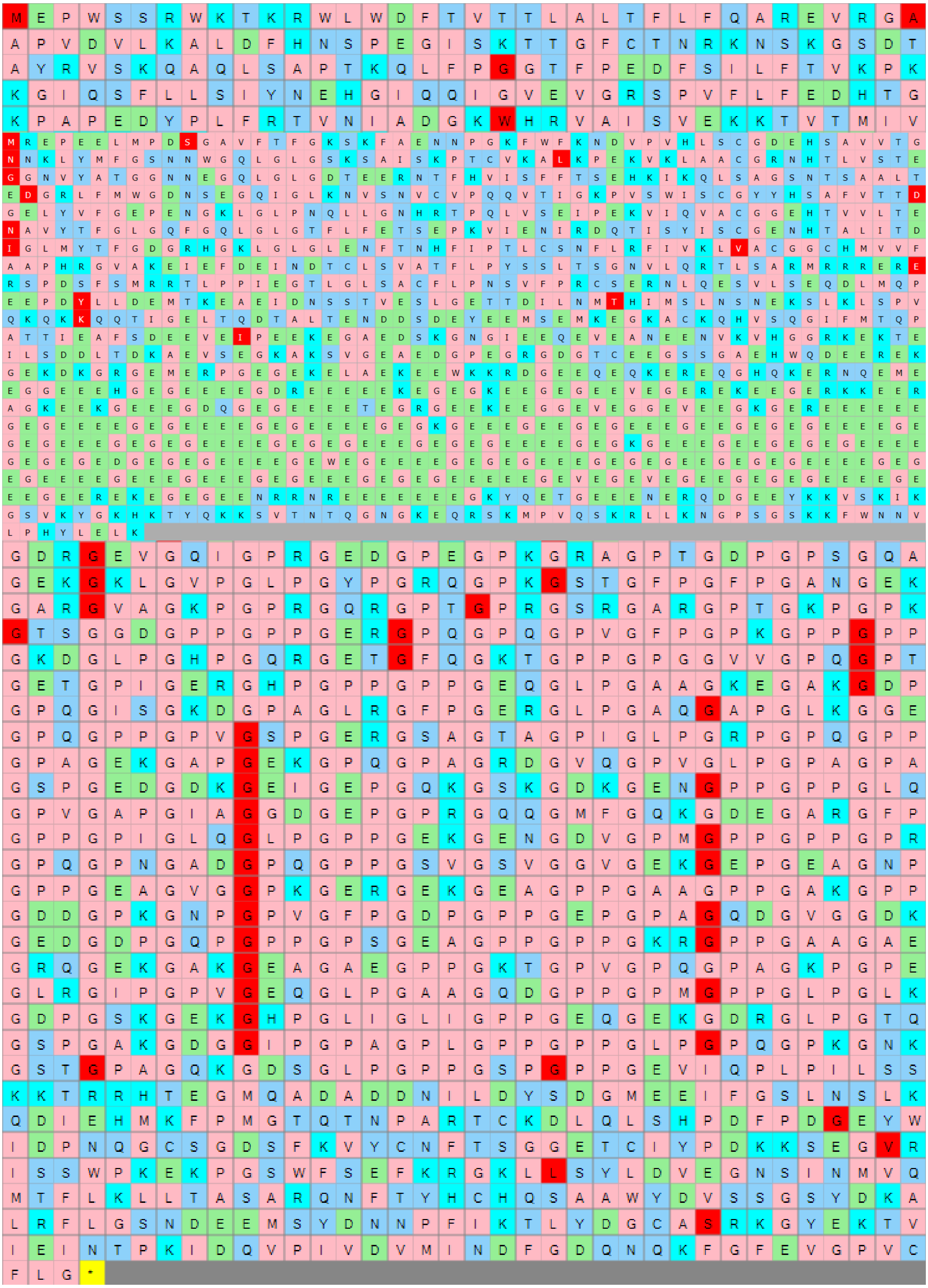
Next, we look at the primary structure sequence depiction of the collagen COLL11A1 and COL protein product, the amino acids at the start of each exon is marked with red blocks. Interestingly, observe how often the amino acid at the start of each exon is a **Glycine**. This is because the consensus sequence for an intron (in IUPAC nucleic acid notation) is: G-G-[cut]-G-T-R-A-G-T (donor site) ... intron sequence ... Y-T-R-A-C (branch sequence 20-50 nucleotides upstream of acceptor site) ... Y-rich-N-C-**A-G-[cut]-G** (acceptor site).

This means that the first base following the intron is very often equal to **G**. When we refer back to the genetic code, we find that the first amino acids at the start of exons will often be one of the following: **G=Glycine (GGG, GGA, GGC, GGT)**, **A=Alanine (G**CG, **G**CA,**G**CC,**G**CT**), E=Glutamic Acid (G**AG,**G**AA**) , D=Aspartic Acid (G**AC,**G**AT**) or V=Valine (G**TG,**G**TA,**G**TC,**G**TT**).**



```
M L S F V D T R T L L L L A V T L C L A T C Q F F E T V R K G P A G
D R G P R G E R G P P G P P G R D G E D G P T G P P G P P G P P G P
P G L G G N F A A Q Y D G K G V G L G P G P M G L M G P R G P P G A
A G A P G P Q G F Q G P A G E P G E P G Q T G P A G A R G P A G P P
G K A G E D G H P G K P G R P G E R G V V G P Q G A R G F P G T P G
L P G F K G I R G H N G L D G L K G Q P G A P G V K G E P G A P G E
N G T P G Q T G A R G L P G E R G R V G A P G P A G A R G S D G S V
G P V G P A G P I G S A G P P G F P G A P G P K G E I G A V G N A G
P A G P A G P R G E V G L P G L S G P V G P P G N P G A N G L T G A
K G A A G L P G V A G A P G L P G P R G I P G P V G A A G A T G A R
G L V G E P G P A G S K G E S G N K G E P G S A G P Q G P P G P S G
E E G K R G P N G E A G S A G P P G P P G L R G S P G S R G L P G A
D G R A G V M G P P G S R G A S G P A G V R G P N G D A G R P G E P
G L M G P R G L P G S P G N I G P A G K E G P V G L P G I D G R P G
P I G P A G A R G E P G N I G F P G P K G P T G D P G K N G D K G H
A G L A G A R G A P G P D G N N G A Q G P P G P Q G V Q G G K G E Q
G P P G P P G F Q G L P G P S G P A G E V G K P G E R G L H G E F G
L P G P A G P R G E R G P P G E S G A A G P T G P I G S R G P S G P
P G P D G N K G E P G V V G A V G T A G P S G P S G L P G E R G A A
G I P G G K G E K G E P G L R G E I G N P G R D G A R G A P G A V G
A P G P A G A T G D R G E A G A A G P A G P A G P R G S P G E R G E
V G P A G P N G F A G P A G A A G Q P G A K G E R G A K G P K G E N
G V V G P T G P V G A A G P A G P N G P P G P A G S R G D G G P P G
M T G F P G A A G R T G P P G P S G I S G P P G P P G P A G K E G L
R G P R G D Q G P V G R T G E V G A V G P P G F A G E K G P S G E A
G T A G P P G T P G P Q G L L G A P G I L G L P G S R G E R G L P G
V A G A V G E P G P L G I A G P P G A R G P P G A V G S P G V N G A
P G E A G R D G N P G N D G P P G R D G Q P G H K G E R G Y P G N I
G P V G A A G A P G P H G P V G P A G K H G N R G E T G P S G P V G
P A G A V G P R G P S G P Q G I R G D K G E P G E K G P R G L P G L
K G H N G L Q G L P G I A G H H G D Q G A P G S V G P A G P R G P A
G P S G P A G K D G R T G H P G T V G P A G I R G P Q G H Q G P A G
P P G P P G P P G P P G V S G G G Y D F G Y D G D F Y R A D Q P R S
A P S L R P K D Y E V D A T L K S L N N Q I E T L L T P E G S R K N
P A R T C R D L R L S H P E W S S G Y Y W I D P N Q G C T M D A I K
V Y C D F S T G E T C I R A Q P E N I P A K N W Y R S S K D K K H V
W L G E T I N A G S Q F E Y N V E G V T S K E M A T Q L A F M R L L
A N Y A S Q N I T Y H C K N S I A Y M D E E T G N L K K A V I L Q G
S N D V E L V A E G N S R F T Y T V L V D G C S K K T N E W G K T I
I E Y K T N K P S R L P F L D I A P L D I G G A D Q E F F V D I G P
V C F K
```

Illustration 59: Pattern formed by amino acid at start of exons for Collagen COL1A2 at 34 amino acids per line



M	E	P	W	S	S	R	W	K	T	K	R	W	L	W	D	F	T	V	T	T	L	A	L	T	F	L	F	Q	A	R	E	V	R	G	A
A	P	V	D	V	L	K	A	L	D	F	H	N	S	P	E	G	I	S	K	T	T	G	F	C	T	N	R	K	N	S	K	G	S	D	T
A	Y	R	V	S	K	Q	A	Q	L	S	A	P	T	K	Q	L	F	P	G	G	T	F	P	E	D	F	S	I	L	F	T	V	K	P	K
K	G	I	Q	S	F	L	L	S	I	Y	N	E	H	G	I	Q	Q	I	G	V	E	V	G	R	S	P	V	F	L	F	E	D	H	T	G
K	P	A	P	E	D	Y	P	L	F	R	T	V	N	I	A	D	G	K	W	H	R	V	A	I	S	V	E	K	K	T	V	T	M	I	V

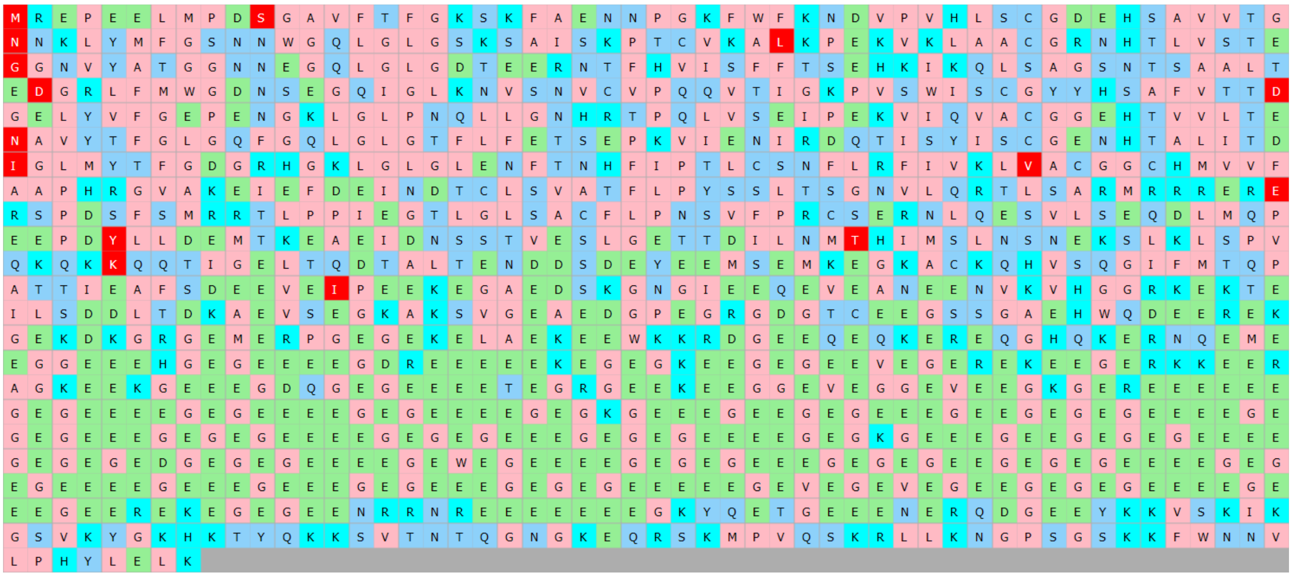
M	R	E	P	E	E	L	M	P	D	S	G	A	V	F	T	F	G	K	S	K	F	A	E	N	N	P	G	K	F	W	F	K	N	D	V	P	V	H	L	S	C	G	D	E	H	S	A	V	V	T	G
N	N	K	L	Y	M	F	G	S	N	N	W	G	Q	L	G	L	G	S	K	S	A	I	S	K	P	T	C	V	K	A	L	K	P	E	K	V	K	L	A	A	C	G	R	N	H	T	L	V	S	T	E
G	G	N	V	Y	A	T	G	G	N	N	E	G	Q	L	G	L	G	D	T	E	E	R	N	T	F	H	V	I	S	F	F	T	S	E	H	K	I	K	Q	L	S	A	G	S	N	T	S	A	A	L	T
E	D	G	R	L	F	M	W	G	D	N	S	E	G	Q	I	G	L	K	N	V	S	N	V	C	V	P	Q	Q	V	T	I	G	K	P	V	S	W	I	S	C	G	Y	Y	H	S	A	F	V	T	T	D
G	E	L	Y	V	F	G	E	P	E	N	G	K	L	G	L	P	N	Q	L	L	G	N	H	R	T	P	Q	L	V	S	E	I	P	E	K	V	I	Q	V	A	C	G	G	E	H	T	V	V	L	T	E
N	A	V	Y	T	F	G	L	G	Q	F	G	Q	L	G	L	G	T	F	L	F	E	T	S	E	P	K	V	I	E	N	I	R	D	Q	T	I	S	Y	I	S	C	G	E	N	H	T	A	L	I	T	D
I	G	L	M	Y	T	F	G	D	G	R	H	G	K	L	G	L	G	L	E	N	F	T	N	H	F	I	P	T	L	C	S	N	F	L	R	F	I	V	K	L	V	A	C	G	G	C	H	M	V	V	F
A	A	P	H	R	G	V	A	K	E	I	E	F	D	E	I	N	D	T	C	L	S	V	A	T	F	L	P	Y	S	S	L	T	S	G	N	V	L	Q	R	T	L	S	A	R	M	R	R	R	E	R	E
R	S	P	D	S	F	S	M	R	R	T	L	P	P	I	E	G	T	L	G	L	S	A	C	F	L	P	N	S	V	F	P	R	C	S	E	R	N	L	Q	E	S	V	L	S	E	Q	D	L	M	Q	P
E	E	P	D	Y	L	L	D	E	M	T	K	E	A	E	I	D	N	S	S	T	V	E	S	L	G	E	T	T	D	I	L	N	M	T	H	I	M	S	L	N	S	N	E	K	S	L	K	L	S	P	V
Q	K	Q	K	K	Q	Q	T	I	G	E	L	T	Q	D	T	A	L	T	E	N	D	D	S	D	E	Y	E	E	M	S	E	M	K	E	G	K	A	C	K	Q	H	V	S	Q	G	I	F	M	T	Q	P
A	T	T	I	E	A	F	S	D	E	E	V	E	I	P	E	E	K	E	G	A	E	D	S	K	G	N	G	I	E	E	Q	E	V	E	A	N	E	E	N	V	K	V	H	G	G	R	K	E	K	T	E
I	L	S	D	D	L	T	D	K	A	E	V	S	E	G	K	A	K	S	V	G	E	A	E	D	G	P	E	G	R	G	D	G	T	C	E	E	G	S	S	G	A	E	H	W	Q	D	E	E	R	E	K
G	E	K	D	K	G	R	G	E	M	E	R	P	G	E	G	E	K	E	L	A	E	K	E	E	W	K	K	R	D	G	E	E	Q	E	Q	K	E	R	E	Q	G	H	Q	K	E	R	N	Q	E	M	E
E	G	G	E	E	E	H	G	E	G	E	E	E	E	G	D	R	E	E	E	E	E	K	E	G	E	G	K	E	E	G	E	G	E	E	V	E	G	E	R	E	K	E	E	G	E	R	K	K	E	E	R
A	G	K	E	E	K	G	E	E	E	G	D	Q	G	E	G	E	E	E	E	T	E	G	R	G	E	E	K	E	E	G	G	E	V	E	G	G	E	V	E	E	G	K	G	E	R	E	E	E	E	E	E
G	E	G	E	E	E	E	G	E	G	E	E	E	E	G	E	G	E	E	E	E	G	E	G	K	G	E	E	E	G	E	E	G	E	G	E	E	E	G	E	E	G	E	G	E	G	E	E	E	E	G	E
G	E	G	E	E	E	G	E	G	E	G	E	E	E	E	G	E	G	E	G	E	E	E	G	E	G	E	G	E	E	E	E	G	E	G	K	G	E	E	E	G	E	E	G	E	G	E	G	E	E	E	E
G	E	G	E	G	E	D	G	E	G	E	G	E	E	E	E	G	E	W	E	G	E	E	E	E	G	E	G	E	G	E	E	E	G	E	G	E	G	E	E	G	E	G	E	G	E	E	E	E	G	E	G
E	G	E	E	E	E	G	E	E	E	G	E	E	E	G	E	G	E	E	E	G	E	G	E	G	E	E	E	E	E	G	E	V	E	G	E	V	E	G	E	E	G	E	G	E	G	E	E	E	E	G	E
E	E	G	E	E	R	E	K	E	G	E	G	E	E	N	R	R	N	R	E	E	E	E	E	E	E	G	K	Y	Q	E	T	G	E	E	E	N	E	R	Q	D	G	E	E	Y	K	K	V	S	K	I	K
G	S	V	K	Y	G	K	H	K	T	Y	Q	K	K	S	V	T	N	T	Q	G	N	G	K	E	Q	R	S	K	M	P	V	Q	S	K	R	L	L	K	N	G	P	S	G	S	K	K	F	W	N	N	V
L	P	H	Y	L	E	L	K																																												

G	D	R	G	E	V	G	Q	I	G	P	R	G	E	D	G	P	E	G	P	K	G	R	A	G	P	T	G	D	P	G	P	S	G	Q	A
G	E	K	G	K	L	G	V	P	G	L	P	G	Y	P	G	R	Q	G	P	K	G	S	T	G	F	P	G	F	P	G	A	N	G	E	K
G	A	R	G	V	A	G	K	P	G	P	R	G	Q	R	G	P	T	G	P	R	G	S	R	G	A	R	G	P	T	G	K	P	G	P	K
G	T	S	G	G	D	G	P	P	G	P	P	G	E	R	G	P	Q	G	P	Q	G	P	V	G	F	P	G	P	K	G	P	P	G	P	P
G	K	D	G	L	P	G	H	P	G	Q	R	G	E	T	G	F	Q	G	K	T	G	P	P	G	P	G	G	V	V	G	P	Q	G	P	T
G	E	T	G	P	I	G	E	R	G	H	P	G	P	P	G	P	P	G	E	Q	G	L	P	G	A	A	G	K	E	G	A	K	G	D	P
G	P	Q	G	I	S	G	K	D	G	P	A	G	L	R	G	F	P	G	E	R	G	L	P	G	A	Q	G	A	P	G	L	K	G	G	E
G	P	Q	G	P	P	G	P	V	G	S	P	G	E	R	G	S	A	G	T	A	G	P	I	G	L	P	G	R	P	G	P	Q	G	P	P
G	P	A	G	E	K	G	A	P	G	E	K	G	P	Q	G	P	A	G	R	D	G	V	Q	G	P	V	G	L	P	G	P	A	G	P	A
G	S	P	G	E	D	G	D	K	G	E	I	G	E	P	G	Q	K	G	S	K	G	D	K	G	E	N	G	P	P	G	P	P	G	L	Q
G	P	V	G	A	P	G	I	A	G	G	D	G	E	P	G	P	R	G	Q	Q	G	M	F	G	Q	K	G	D	E	G	A	R	G	F	P
G	P	P	G	P	I	G	L	Q	G	L	P	G	P	P	G	E	K	G	E	N	G	D	V	G	P	M	G	P	P	G	P	P	G	P	R
G	P	Q	G	P	N	G	A	D	G	P	Q	G	P	P	G	S	V	G	S	V	G	G	V	G	E	K	G	E	P	G	E	A	G	N	P
G	P	P	G	E	A	G	V	G	G	P	K	G	E	R	G	E	K	G	E	A	G	P	P	G	A	A	G	P	P	G	A	K	G	P	P
G	D	D	G	P	K	G	N	P	G	P	V	G	F	P	G	D	P	G	P	P	G	E	P	G	P	A	G	Q	D	G	V	G	G	D	K
G	E	D	G	D	P	G	Q	P	G	P	P	G	P	S	G	E	A	G	P	P	G	P	P	G	K	R	G	P	P	G	A	A	G	A	E
G	R	Q	G	E	K	G	A	K	G	E	A	G	A	E	G	P	P	G	K	T	G	P	V	G	P	Q	G	P	A	G	K	P	G	P	E
G	L	R	G	I	P	G	P	V	G	E	Q	G	L	P	G	A	A	G	Q	D	G	P	P	G	P	M	G	P	P	G	L	P	G	L	K
G	D	P	G	S	K	G	E	K	G	H	P	G	L	I	G	L	I	G	P	P	G	E	Q	G	E	K	G	D	R	G	L	P	G	T	Q
G	S	P	G	A	K	G	D	G	G	I	P	G	P	A	G	P	L	G	P	P	G	P	P	G	L	P	G	P	Q	G	P	K	G	N	K
G	S	T	G	P	A	G	Q	K	G	D	S	G	L	P	G	P	P	G	S	P	G	P	P	G	E	V	I	Q	P	L	P	I	L	S	S
K	K	T	R	R	H	T	E	G	M	Q	A	D	A	D	D	N	I	L	D	Y	S	D	G	M	E	E	I	F	G	S	L	N	S	L	K
Q	D	I	E	H	M	K	F	P	M	G	T	Q	T	N	P	A	R	T	C	K	D	L	Q	L	S	H	P	D	F	P	D	G	E	Y	W
I	D	P	N	Q	G	C	S	G	D	S	F	K	V	Y	C	N	F	T	S	G	G	E	T	C	I	Y	P	D	K	K	S	E	G	V	R
I	S	S	W	P	K	E	K	P	G	S	W	F	S	E	F	K	R	G	K	L	L	S	Y	L	D	V	E	G	N	S	I	N	M	V	Q
M	T	F	L	K	L	L	T	A	S	A	R	Q	N	F	T	Y	H	C	H	Q	S	A	A	W	Y	D	V	S	S	G	S	Y	D	K	A
L	R	F	L	G	S	N	D	E	E	M	S	Y	D	N	N	P	F	I	K	T	L	Y	D	G	C	A	S	R	K	G	Y	E	K	T	V
I	E	I	N	T	P	K	I	D	Q	V	P	I	V	D	V	M	I	N	D	F	G	D	Q	N	Q	K	F	G	F	E	V	G	P	V	C
F	L	G	*																																

Illustration 60: Similar patterns in the protein of Collagen COL11A1 at width 36 residues per line

Different representations of the Retinitis pigmentosa GTPase regulator (RPGR) protein (Id = uc004ded.2)

On the Protein domain as colour coded amino acids:

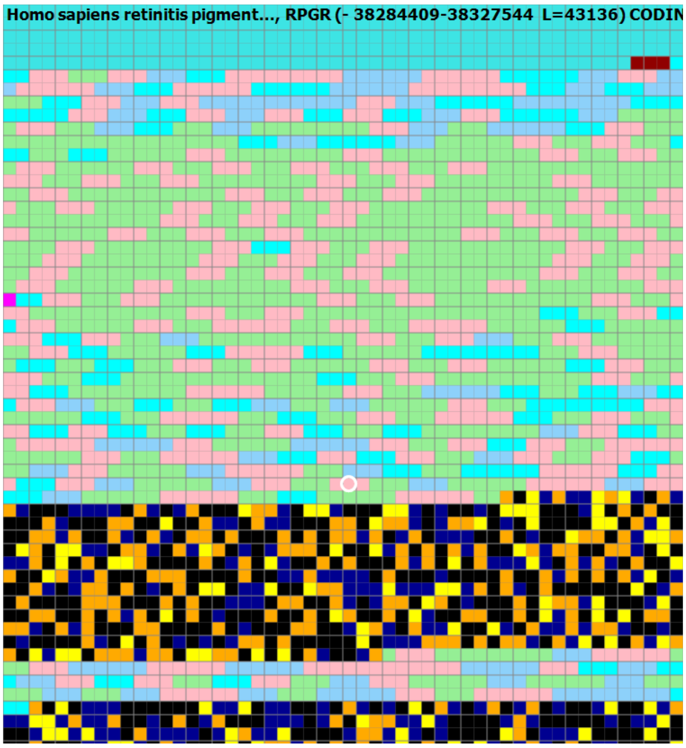


```
M R E P E E L M P D S G A V F T F G K S K F A E N N P G K F W F K N D V P V H L S C G D E H S A V V T G
N N K L Y M F G S N N W G Q L G L G S K S A I S K P T C V K A L K P E K V K L A A C G R N H T L V S T E
G G N V Y A T G G N N E G Q L G L G D T E E R N T F H V I S F F T S E H K I K Q L S A G S N T S A A L T
E D G R L F M W G D N S E G Q I G L K N V S N V C V P Q Q V T I G K P V S W I S C G Y Y H S A F V T T D
G E L Y V F G E P E N G K L G L P N Q L L G N H R T P Q L V S E I P E K V I Q V A C G G E H T V V L T E
N A V Y T F G L G Q F G Q L G L G T F L F E T S E P K V I E N I R D Q T I S Y I S C G E N H T A L I T D
I G L M Y T F G D G R H G K L G L G L E N F T N H F I P T L C S N F L R F I V K L V A C G G C H M V V F
A A P H R G V A K E I E F D E I N D T C L S V A T F L P Y S S L T S G N V L Q R T L S A R M R R R E R E
R S P D S F S M R R T L P P I E G T L G L S A C F L P N S V F P R C S E R N L Q E S V L S E Q D L M Q P
E E P D Y L L D E M T K E A E I D N S S T V E S L G E T T D I L N M T H I M S L N S N E K S L K L S P V
Q K Q K K Q Q T I G E L T Q D T A L T E N D D S D E Y E E M S E M K E G K A C K Q H V S Q G I F M T Q P
A T T I E A F S D E E V E I P E E K E G A E D S K G N G I E E Q E V E A N E E N V K V H G G R K E K T E
I L S D D L T D K A E V S E G K A K S V G E A E D G P E G R G D G T C E E G S S G A E H W Q D E E R E K
G E K D K G R G E M E R P G E G E K E L A E K E E W K K R D G E E Q E Q K E R E Q G H Q K E R N Q E M E
E G G E E E H G E G E E E E G D R E E E E E K E G E G K E E G E G E E V E G E R E K E E G E R K K E E R
A G K E E K G E E E G D Q G E G E E E E T E G R G E E K E E G G E V E G G E V E E G K G E R E E E E E E
G E G E E E E G E G E E E E G E G E E E E G E G K G E E E G E E G E G E E E G E E G E G E G E E E E G E
G E G E E E G E G E G E E E E G E G E G E E E G E G E G E E E E G E G K G E E E G E E G E G E G E E E E
G E G E G E D G E G E G E E E E G E W E G E E E E G E G E G E E E G E G E G E E G E G E G E E E E G E G
E G E E E E G E E E G E E E G E G E E E G E G E G E E E E E G E V E G E V E G E E G E G E G E E E E G E
E E G E E R E K E G E G E E N R R N R E E E E E E E G K Y Q E T G E E E N E R Q D G E E Y K K V S K I K
G S V K Y G K H K T Y Q K K S V T N T Q G N G K E Q R S K M P V Q S K R L L K N G P S G S K K F W N N V
L P H Y L E L K
```

Illustration 61: RPGR protein product at chrX:38327340-38327544 at 52 residues width

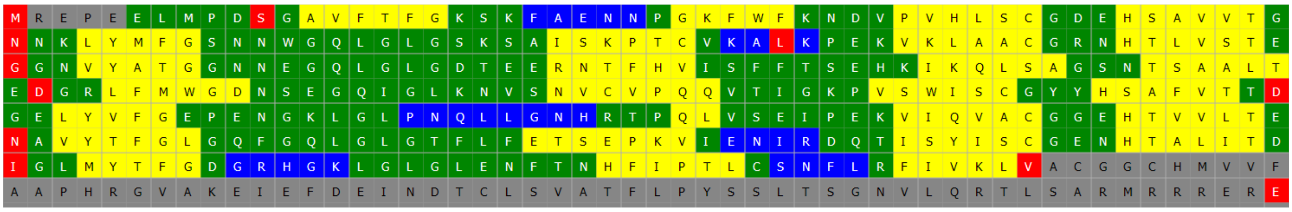
The initial pink and blue parts represent the first part of the protein which is embedded in the membrane

In the DNA View with codons colour coded (only the last 2 exons)



Homo sapiens retinitis pigment..., RPGR (- 38284409-38327544 L=43136) CODIN

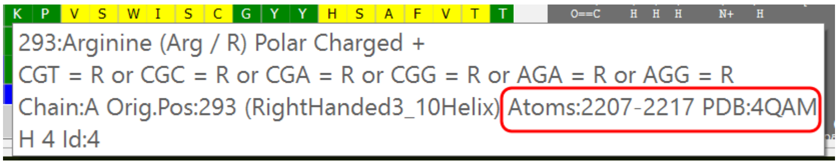
Illustration 62: The last exon of the negative strand encoded gene on the DNA domain at 52 bases per line



M	R	E	P	E	E	L	M	P	D	S	G	A	V	F	T	F	G	K	S	K	F	A	E	N	N	P	G	K	F	W	F	K	N	D	V	P	V	H	L	S	C	G	D	E	H	S	A	V	V	T	G
N	N	K	L	Y	M	F	G	S	N	N	W	G	Q	L	G	L	G	S	K	S	A	I	S	K	P	T	C	V	K	A	L	K	P	E	K	V	K	L	A	A	C	G	R	N	H	T	L	V	S	T	E
G	G	N	V	Y	A	T	G	G	N	N	E	G	Q	L	G	L	G	D	T	E	E	R	N	T	F	H	V	I	S	F	F	T	S	E	H	K	I	K	Q	L	S	A	G	S	N	T	S	A	A	L	T
E	D	G	R	L	F	M	W	G	D	N	S	E	G	Q	I	G	L	K	N	V	S	N	V	C	V	P	Q	Q	V	T	I	G	K	P	V	S	W	I	S	C	G	Y	Y	H	S	A	F	V	T	T	D
G	E	L	Y	V	F	G	E	P	E	N	G	K	L	G	L	P	N	Q	L	L	G	N	H	R	T	P	Q	L	V	S	E	I	P	E	K	V	I	Q	V	A	C	G	G	E	H	T	V	V	L	T	E
N	A	V	Y	T	F	G	L	G	Q	F	G	Q	L	G	L	G	T	F	L	F	E	T	S	E	P	K	V	I	E	N	I	R	D	Q	T	I	S	Y	I	S	C	G	E	N	H	T	A	L	I	T	D
I	G	L	M	Y	T	F	G	D	G	R	H	G	K	L	G	L	G	L	E	N	F	T	N	H	F	I	P	T	L	C	S	N	F	L	R	F	I	V	K	L	V	A	C	G	G	C	H	M	V	V	F
A	A	P	H	R	G	V	A	K	E	I	E	F	D	E	I	N	D	T	C	L	S	V	A	T	F	L	P	Y	S	S	L	T	S	G	N	V	L	Q	R	T	L	S	A	R	M	R	R	R	E	R	E

Illustration 63: The same RPGR protein product with secondary protein structure colour coded at 52 residues width (Yellow=beta sheet, Blue=alpha helix, Green=Loop)

The tooltip on the amino acids indicates that the section that was aligned in order to obtain the secondary structure was the PDB crystal structure: 4QAM and the PDB atoms which were matched to the specific amino acid were 2207-2217



K P V S W I S C G Y Y H S A F V T T

O==C H H H N+ H

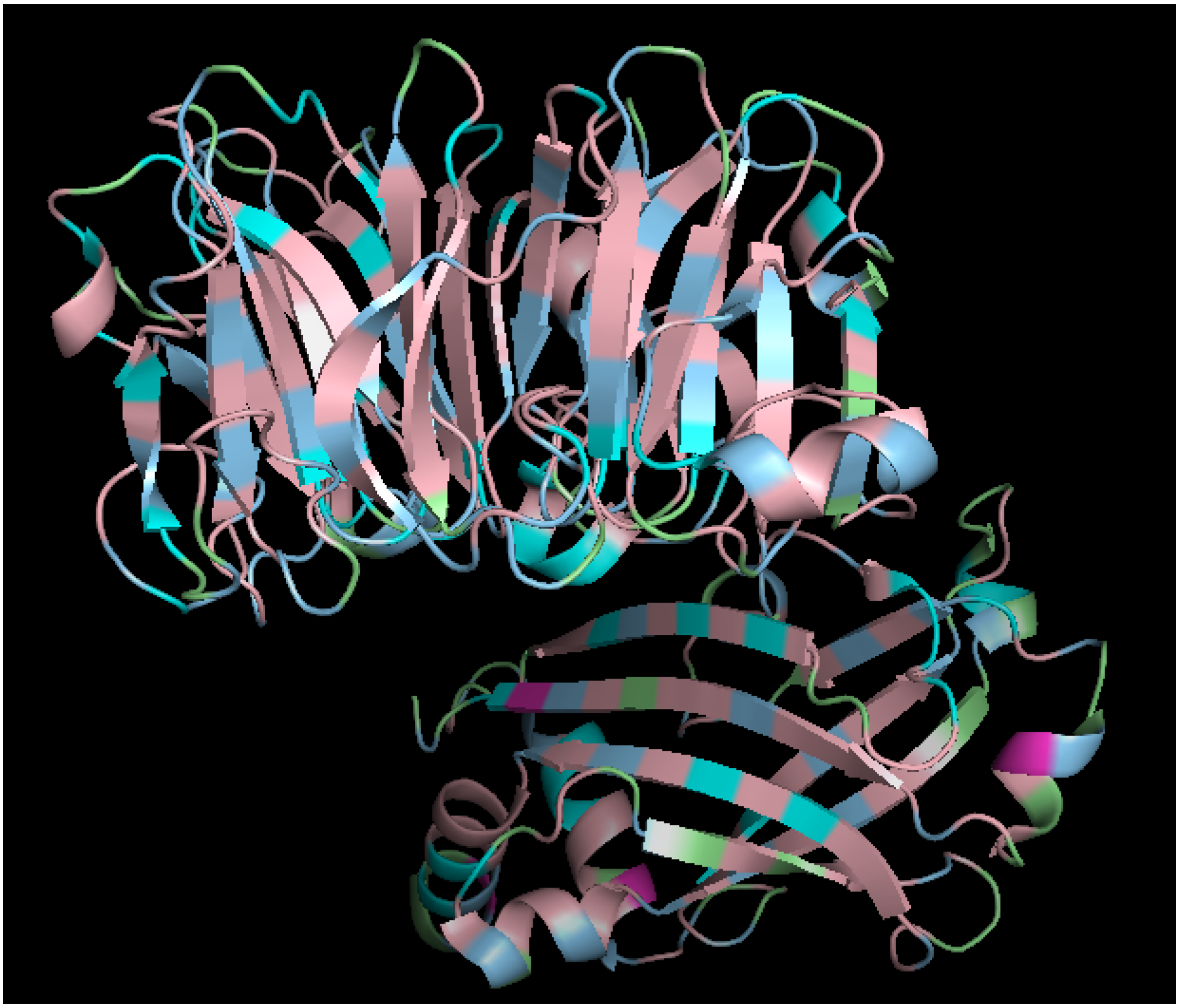
293:Arginine (Arg / R) Polar Charged +

CGT = R or CGC = R or CGA = R or CGG = R or AGA = R or AGG = R

Chain:A Orig.Pos:293 (RightHanded3_10Helix) Atoms:2207-2217 PDB:4QAM

H 4 Id:4

When the 3D protein structure is now launched in Pymol with exactly the same colour coding:



Notice how the transmembrane occupying amino acids are mostly hydrophobic (pink) due to the requirement of interacting with the hydrophobic (oily) membrane.

The Genetic Code - How the colour scheme was chosen

The colour scheme is an indication of how hydrophobic or polar the amino acids are. Protein structures fold in water spontaneously due to the interactions the side chains of the amino acids have with water. Hydrophobic amino acids tend to move away from polar water molecules and therefore tend to be on the inside of folded proteins (OR on the outside when embedded in the hydrophobic lipid membranes of a cell membrane).

Colour coding the primary structure of a protein based on the polarity of the amino acids gives you a better idea of which amino acids of the protein will be on the inside and which will be on the outside, by just looking at the sequence. It is however not always as simple and one also have to look at the 3D secondary structure such as alpha helices, beta sheets and loops.

When this colour coding is extended from the protein to the DNA domain by doing the same for the corresponding codons, it allows you to better understand which part of the protein is coded by different exons when the coding sequence is assembled.

It is also known that amino acids with similar properties (such as hydrophobicity) can often be substituted for each other without much detrimental effect on the protein folding. A missense mutation in the DNA is when the change in codon base results in a different amino acid. When a single base missense mutation causes the translation of an amino acid with similar hydrophobic properties as the original amino acid it is known as a **conservative missense mutation**.

I therefore decided to arrange the Genetic Codon table in such a fashion that there will always be only a single base change between adjacent codons in the table. In digital communication this is know as a Hamming Code and it is used during error correction coding. Codes are given a redundancy by only choosing a subset of the available codes as valid codes, in other words, with a hamming distance of 3. When single bit changes are introduced in transmission, one can then determine the correct original bit by choosing the code words which are closest in Hamming Distance to the received code.

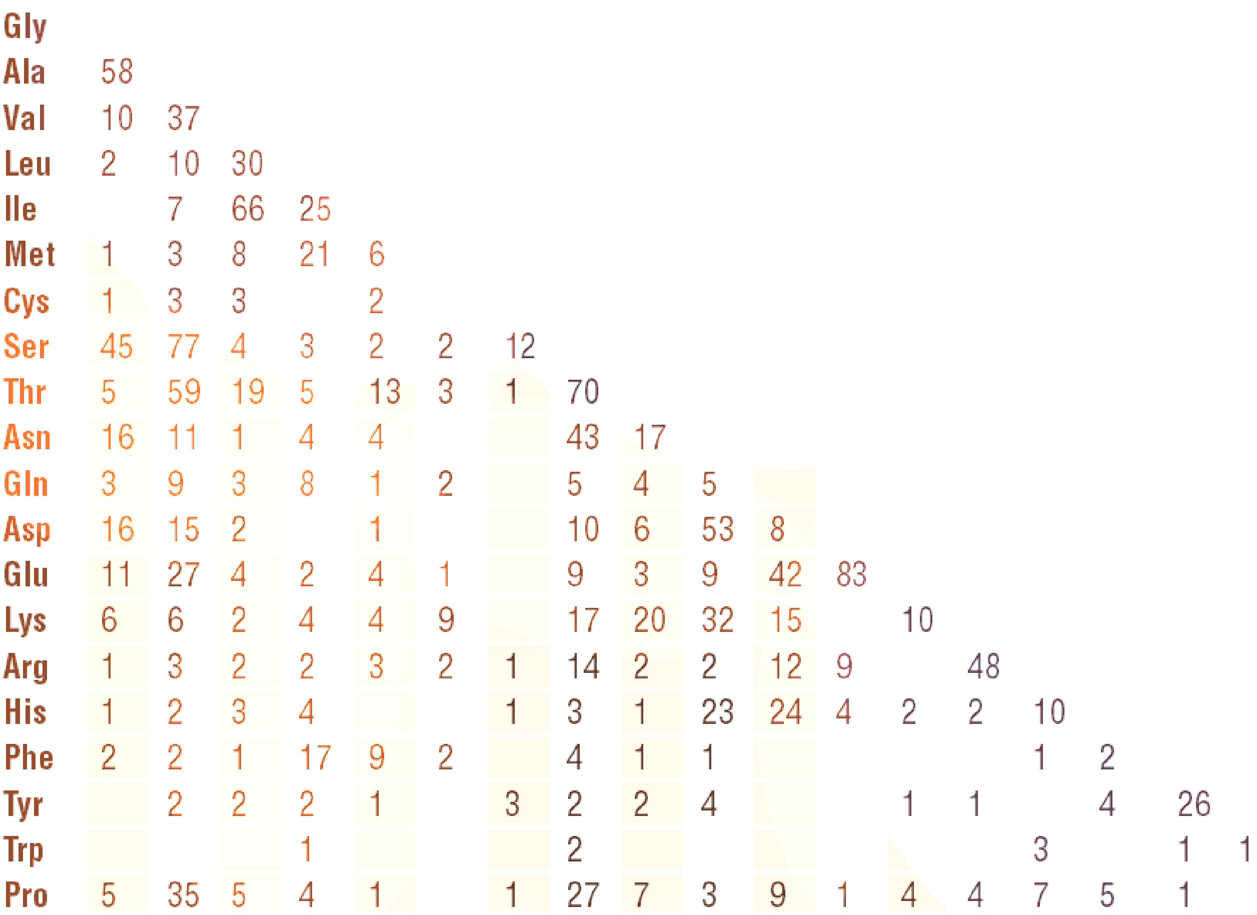
I ended up with a Codon Table which resembles a “torus” where the codons in the top row is also only one base away from the codons of the bottom row and the same is valid in the left and right directions. A Genetic Code arranged in this way will therefore only code for the directly adjacent amino acids when single base changes are introduced by mutation. Diagonal movement represents 2 base changes (except when looping to the other side of the torus).

After colour coding the amino acids coded for by those codons and adding the hydrophobicity value of the amino acids in the centre, an **interesting “periodic table like”** arrangement emerged where amino acids with similar properties tended to cluster together.

Moving horizontally left or right **represents a change in the last “wobble base”** of the codon, which we know is due to the ability of Uracil to pair with either an A or a G in the anti-codon loop of the tRNA. A horizontal base change very often results in a **silent mutation**, which results in a mutated codon still coding for the same amino acid (eg. Alanine (GCC) → Alanine (GCT))

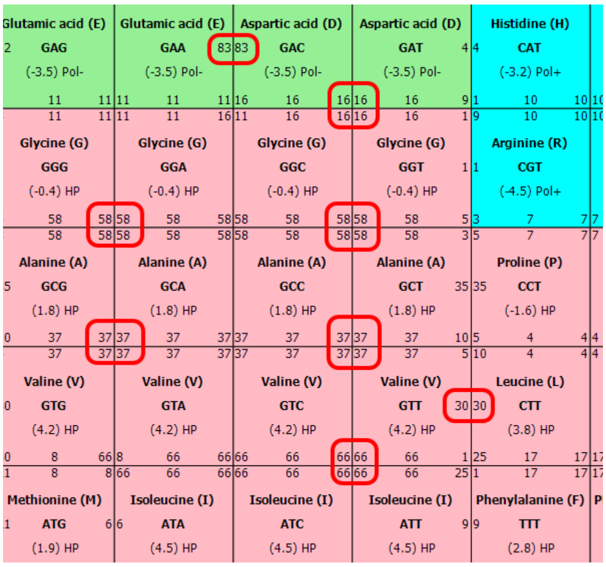
Moving up or down by one base mostly results in a different amino acid with similar properties, in other words: a **conservative missense mutation**.

In order to even further accentuate the Genetic Code's robustness against base changes, I decided to incorporate a matrix which represents how often amino acids are substituted in nature without detrimental effects to resulting proteins. I plotted the corresponding substitution frequency in the corners of the Codon Table.



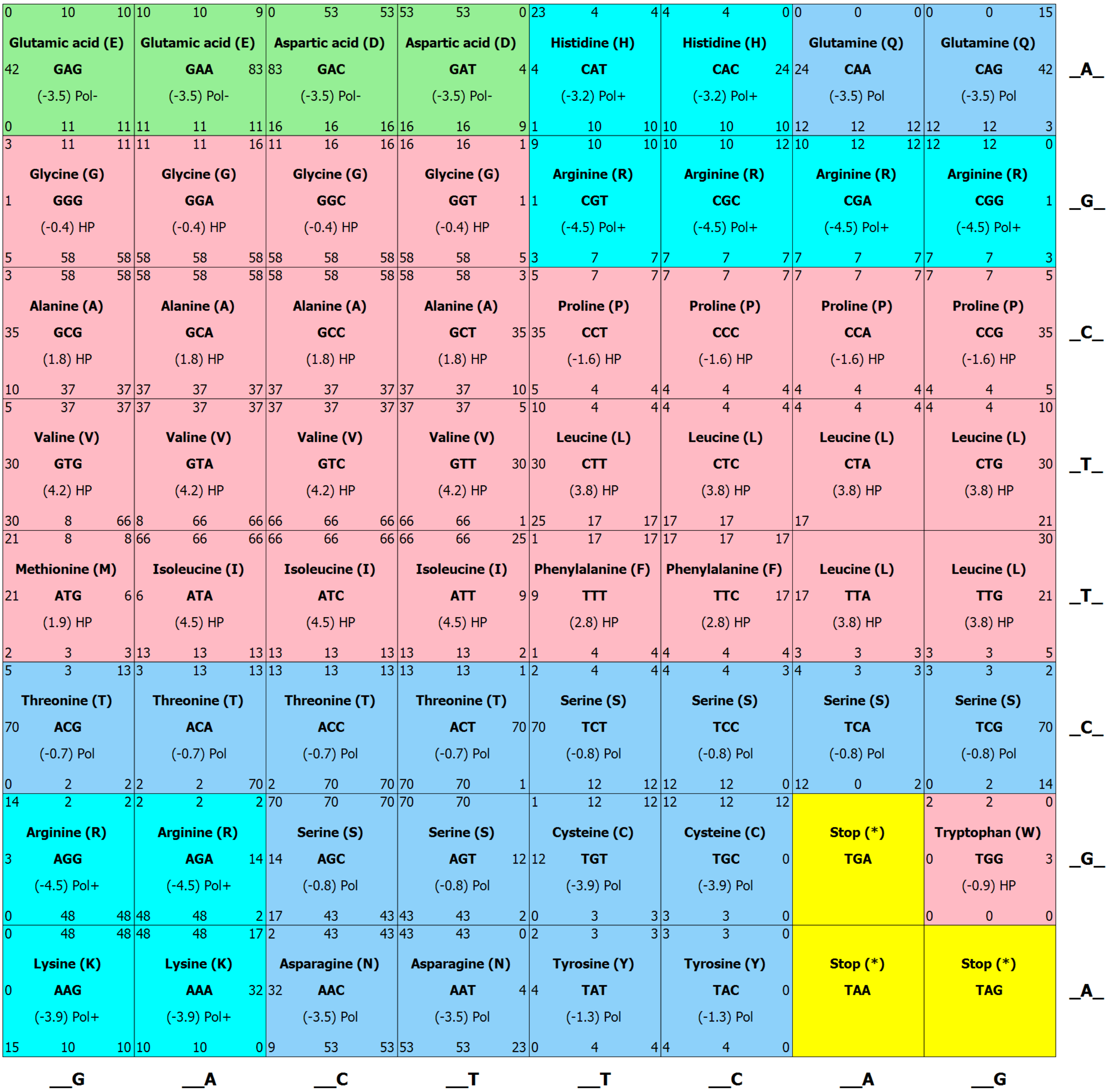
Gly																		
Ala	58																	
Val	10	37																
Leu	2	10	30															
Ile		7	66	25														
Met	1	3	8	21	6													
Cys	1	3	3		2													
Ser	45	77	4	3	2	2	12											
Thr	5	59	19	5	13	3	1	70										
Asn	16	11	1	4	4			43	17									
Gln	3	9	3	8	1	2		5	4	5								
Asp	16	15	2		1			10	6	53	8							
Glu	11	27	4	2	4	1		9	3	9	42	83						
Lys	6	6	2	4	4	9		17	20	32	15		10					
Arg	1	3	2	2	3	2	1	14	2	2	12	9		48				
His	1	2	3	4			1	3	1	23	24	4	2	2	10			
Phe	2	2	1	17	9	2		4	1	1					1	2		
Tyr		2	2	2	1		3	2	2	4			1	1		4	26	
Trp				1				2							3		1	1
Pro	5	35	5	4	1		1	27	7	3	9	1	4	4	7	5	1	

Notice how the frequency is much higher between adjacent amino acids with similar colour (i.e. hydrophobicity), but noticeably lower when the colour is different. The robustness of the genetic code has to do with finding the optimal arrangement where the non-detrimental substitution frequency is maximal across all codons.



Glutamic acid (E) 2 GAG (-3.5) Pol- 11 11	Glutamic acid (E) GAA 83 (-3.5) Pol- 11 11 11	Aspartic acid (D) 83 GAC (-3.5) Pol- 16 16 16	Aspartic acid (D) GAT 4 (-3.5) Pol- 16 16 9	Histidine (H) 4 CAT (-3.2) Pol+ 1 10 10	10
11 11 Glycine (G) GGG (-0.4) HP 58 58	11 11 16 Glycine (G) GGA (-0.4) HP 58 58 58	11 16 16 Glycine (G) GGC (-0.4) HP 58 58 58	16 16 1 Glycine (G) GGT 1 (-0.4) HP 58 58 5	9 10 10 Arginine (R) 1 CGT (-4.5) Pol+ 3 7 7	10 7
58 58 Alanine (A) 5 GCG (1.8) HP 0 37 37	58 58 58 Alanine (A) GCA (1.8) HP 37 37 37	58 58 58 Alanine (A) GCC (1.8) HP 37 37 37	58 58 3 Alanine (A) GCT 35 (1.8) HP 37 37 10	5 7 7 Proline (P) 35 CCT (-1.6) HP 5 4 4	7 4
37 37 Valine (V) 0 GTG (4.2) HP 0 8 66	37 37 37 Valine (V) GTA (4.2) HP 8 66 66	37 37 37 Valine (V) GTC (4.2) HP 66 66 66	37 37 5 Valine (V) GTT 30 (4.2) HP 66 66 1	10 4 4 Leucine (L) 30 CTT (3.8) HP 25 17 17	4 17
1 8 8 Methionine (M) 1 ATG 6 (1.9) HP	66 66 66 Isoleucine (I) 6 ATA (4.5) HP	66 66 66 Isoleucine (I) ATC (4.5) HP	66 66 25 Isoleucine (I) ATT 9 (4.5) HP	1 17 17 Phenylalanine (F) 9 TTT (2.8) HP	17 P

The Colour Coded (Hamming code arranged) Genetic Codon Table



0 10 10 Glutamic acid (E) 42 GAG (-3.5) Pol- 0 11 11	10 10 9 Glutamic acid (E) GAA 83 (-3.5) Pol- 11 11 11	0 53 53 Aspartic acid (D) 83 GAC (-3.5) Pol- 16 16 16	53 53 0 Aspartic acid (D) GAT 4 (-3.5) Pol- 16 16 9	23 4 4 Histidine (H) 4 CAT (-3.2) Pol+ 1 10 10	4 4 0 Histidine (H) CAC 24 (-3.2) Pol+ 10 10 10	0 0 0 Glutamine (Q) 24 CAA (-3.5) Pol 12 12 12	0 0 15 Glutamine (Q) CAG 42 (-3.5) Pol 12 12 3	_A_
3 11 11 Glycine (G) 1 GGG (-0.4) HP 5 58 58	11 11 16 Glycine (G) GGA (-0.4) HP 58 58 58	11 16 16 Glycine (G) GGC (-0.4) HP 58 58 58	16 16 1 Glycine (G) GGT 1 (-0.4) HP 58 58 5	9 10 10 Arginine (R) 1 CGT (-4.5) Pol+ 3 7 7	10 10 12 Arginine (R) CGC (-4.5) Pol+ 7 7 7	10 12 12 Arginine (R) CGA (-4.5) Pol+ 7 7 7	12 12 0 Arginine (R) CGG 1 (-4.5) Pol+ 7 7 3	_G_
3 58 58 Alanine (A) 35 GCG (1.8) HP 10 37 37	58 58 58 Alanine (A) GCA (1.8) HP 37 37 37	58 58 58 Alanine (A) GCC (1.8) HP 37 37 37	58 58 3 Alanine (A) GCT 35 (1.8) HP 37 37 10	5 7 7 Proline (P) 35 CCT (-1.6) HP 5 4 4	7 7 7 Proline (P) CCC (-1.6) HP 4 4 4	7 7 7 Proline (P) CCA (-1.6) HP 4 4 4	7 7 5 Proline (P) CCG 35 (-1.6) HP 4 4 5	_C_
5 37 37 Valine (V) 30 GTG (4.2) HP 30 8 66	37 37 37 Valine (V) GTA (4.2) HP 8 66 66	37 37 37 Valine (V) GTC (4.2) HP 66 66 66	37 37 5 Valine (V) GTT 30 (4.2) HP 66 66 1	10 4 4 Leucine (L) 30 CTT (3.8) HP 25 17 17	4 4 4 Leucine (L) CTC (3.8) HP 17 17	4 4 4 Leucine (L) CTA (3.8) HP 17	4 4 10 Leucine (L) CTG 30 (3.8) HP 21	_T_
21 8 8 Methionine (M) 21 ATG 6 (1.9) HP 2 3 3	66 66 66 Isoleucine (I) 6 ATA (4.5) HP 13 13 13	66 66 66 Isoleucine (I) ATC (4.5) HP 13 13 13	66 66 25 Isoleucine (I) ATT 9 (4.5) HP 13 13 2	1 17 17 Phenylalanine (F) 9 TTT (2.8) HP 1 4 4	17 17 17 Phenylalanine (F) TTC 17 (2.8) HP 4 4 4	Leucine (L) 17 TTA (3.8) HP 3 3 3	30 Leucine (L) TTG 21 (3.8) HP 3 3 5	_T_
5 3 13 Threonine (T) 70 ACG (-0.7) Pol 0 2 2	3 13 13 Threonine (T) ACA (-0.7) Pol 2 2 70	13 13 13 Threonine (T) ACC (-0.7) Pol 2 70 70	13 13 1 Threonine (T) ACT 70 (-0.7) Pol 70 70 1	2 4 4 Serine (S) 70 TCT (-0.8) Pol 12 12	4 4 3 Serine (S) TCC (-0.8) Pol 12 12 0	4 3 3 Serine (S) TCA (-0.8) Pol 12 0 2	3 3 2 Serine (S) TCG 70 (-0.8) Pol 0 2 14	_C_
14 2 2 Arginine (R) 3 AGG (-4.5) Pol+ 0 48 48	2 2 2 Arginine (R) AGA 14 (-4.5) Pol+ 48 48 2	70 70 70 Serine (S) 14 AGC (-0.8) Pol 17 43 43	70 70 Serine (S) AGT 12 (-0.8) Pol 43 43 2	1 12 12 Cysteine (C) 12 TGT (-3.9) Pol 0 3 3	12 12 12 Cysteine (C) TGC 0 (-3.9) Pol 3 3 0	Stop (*) TGA	2 2 0 Tryptophan (W) 0 TGG 3 (-0.9) HP 0 0 0	_G_
0 48 48 Lysine (K) 0 AAG (-3.9) Pol+ 15 10 10	48 48 17 Lysine (K) AAA 32 (-3.9) Pol+ 10 10 0	2 43 43 Asparagine (N) 32 AAC (-3.5) Pol 9 53 53	43 43 0 Asparagine (N) AAT 4 (-3.5) Pol 53 53 23	2 3 3 Tyrosine (Y) 4 TAT (-1.3) Pol 0 4 4	3 3 0 Tyrosine (Y) TAC 0 (-1.3) Pol 4 4 0	Stop (*) TAA	Stop (*) TAG	_A_
__G	__A	__C	__T	__T	__C	__A	__G	

Non-polar Hydrophobic

Polar Uncharged

Polar Charged +

Polar Charged -

Stop Codons

Periodic Genetic Code table organised by Hamming distance

Each block's codon differs only by one base from any adjacent block
The single base changes also wrap from right edge to left edge and bottom edge to top edge like a 3D torus.
Numbers in the corners and sides of blocks represent the frequency with which one amino acid is found substituted in nature by the adjacent one for the corresponding protein from different organisms.
The centre value in brackets shows the hydrophobicity of the amino acid.
This table serves to illustrate the robustness of the genetic code against single and double nucleotide mutations.
Single nucleotide mutations (horizontally or vertically adjacent blocks) or
Double nucleotide mutations (diagonally adjacent blocks)
seem to either result in the same amino acid (silent mutations) or amino acids
with similar properties (conservative missense mutations) or similar molecular structure.

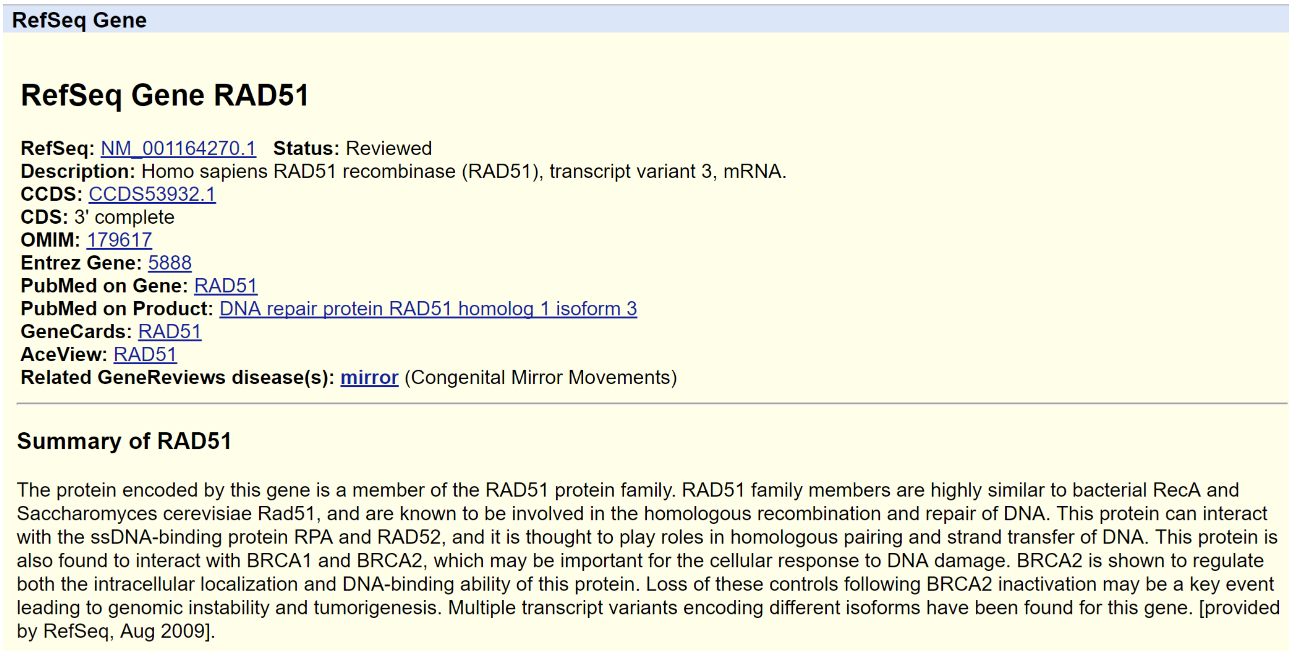
*The Visual Genome Browser's **Genetic Code** tab, contains this Genetic Codon Table.*

Show RefSeq Info at UCSC

Choosing this option for a gene will open the online summary information for the specific gene. It essentially represents the detail page for the gene, after you have clicked on the gene region in the UCSC genome display.

For the gene RAD51 it provides you with a hub of links leading to other information, such as OMIM (Online Mendelian Inheritance Of Man – providing online information as to what evilnesses are linked to mutations in this gene)
PubMed, etc.

http://genome.ucsc.edu/cgi-bin/hgc?g=refGene&i=NM_001164270



RefSeq Gene

RefSeq Gene RAD51

RefSeq: NM_001164270.1 **Status:** Reviewed
Description: Homo sapiens RAD51 recombinase (RAD51), transcript variant 3, mRNA.
CCDS: CCDS53932.1
CDS: 3' complete
OMIM: 179617
Entrez Gene: 5888
PubMed on Gene: RAD51
PubMed on Product: DNA repair protein RAD51 homolog 1 isoform 3
GeneCards: RAD51
AceView: RAD51
Related GeneReviews disease(s): **mirror** (Congenital Mirror Movements)

Summary of RAD51

The protein encoded by this gene is a member of the RAD51 protein family. RAD51 family members are highly similar to bacterial RecA and Saccharomyces cerevisiae Rad51, and are known to be involved in the homologous recombination and repair of DNA. This protein can interact with the ssDNA-binding protein RPA and RAD52, and it is thought to play roles in homologous pairing and strand transfer of DNA. This protein is also found to interact with BRCA1 and BRCA2, which may be important for the cellular response to DNA damage. BRCA2 is shown to regulate both the intracellular localization and DNA-binding ability of this protein. Loss of these controls following BRCA2 inactivation may be a key event leading to genomic instability and tumorigenesis. Multiple transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Aug 2009].

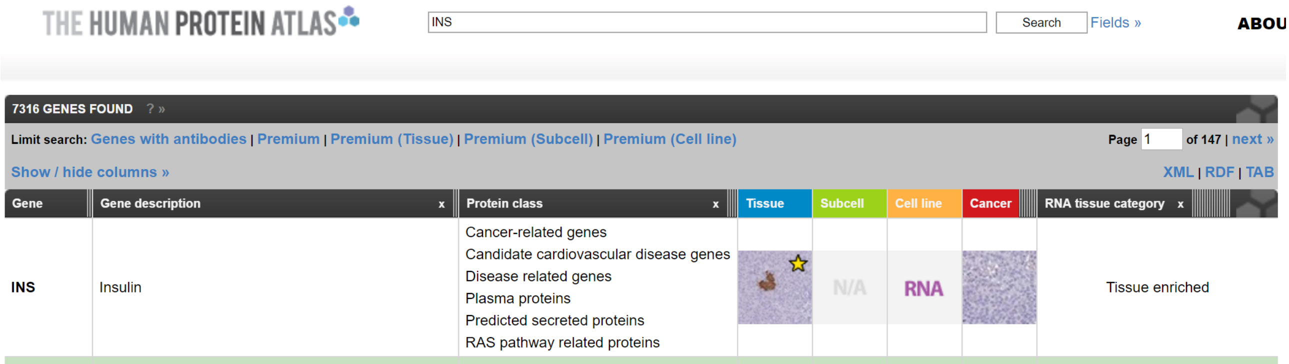
Show in Human Protein Atlas

This option takes you to a the human protein atlas' information on the selected protein coding gene.

I wanted to find a way to quickly determine which body tissues a specific protein coding gene are expressed at the highest levels and found that I could link to the **Human Protein Atlas** site from the context menu.

For, example, when it is selected for the human insulin gene (INS) which is on chromosome 11, you are taken to:

http://www.proteinatlas.org/search/INS



7316 GENES FOUND ? »

Limit search: Genes with antibodies | Premium | Premium (Tissue) | Premium (Subcell) | Premium (Cell line)

Page 1 of 147 | next »

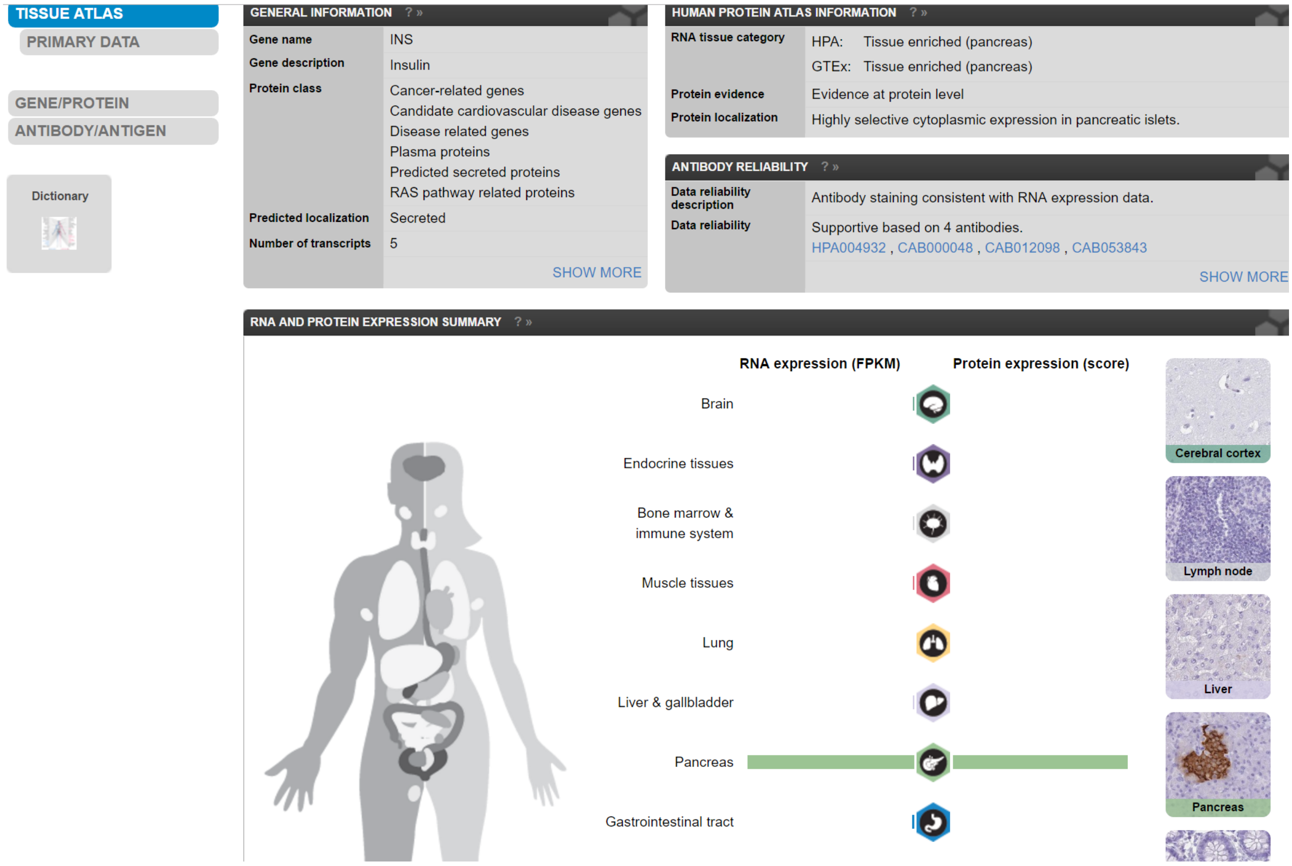
Show / hide columns »

XML | RDF | TAB

Gene	Gene description	Protein class	Tissue	Subcell	Cell line	Cancer	RNA tissue category
INS	Insulin	Cancer-related genes Candidate cardiovascular disease genes Disease related genes Plasma proteins Predicted secreted proteins RAS pathway related proteins		N/A	RNA		Tissue enriched

This allows you to look at the expression of this gene's product in different tissue types in the human body.

By clicking on the "**Tissue**" column, you can get the following type of information for the insulin protein:



GENERAL INFORMATION ? »

Gene name	INS
Gene description	Insulin
Protein class	Cancer-related genes Candidate cardiovascular disease genes Disease related genes Plasma proteins Predicted secreted proteins RAS pathway related proteins
Predicted localization	Secreted
Number of transcripts	5

SHOW MORE

HUMAN PROTEIN ATLAS INFORMATION ? »

RNA tissue category	HPA: Tissue enriched (pancreas) GTEx: Tissue enriched (pancreas)
Protein evidence	Evidence at protein level
Protein localization	Highly selective cytoplasmic expression in pancreatic islets.

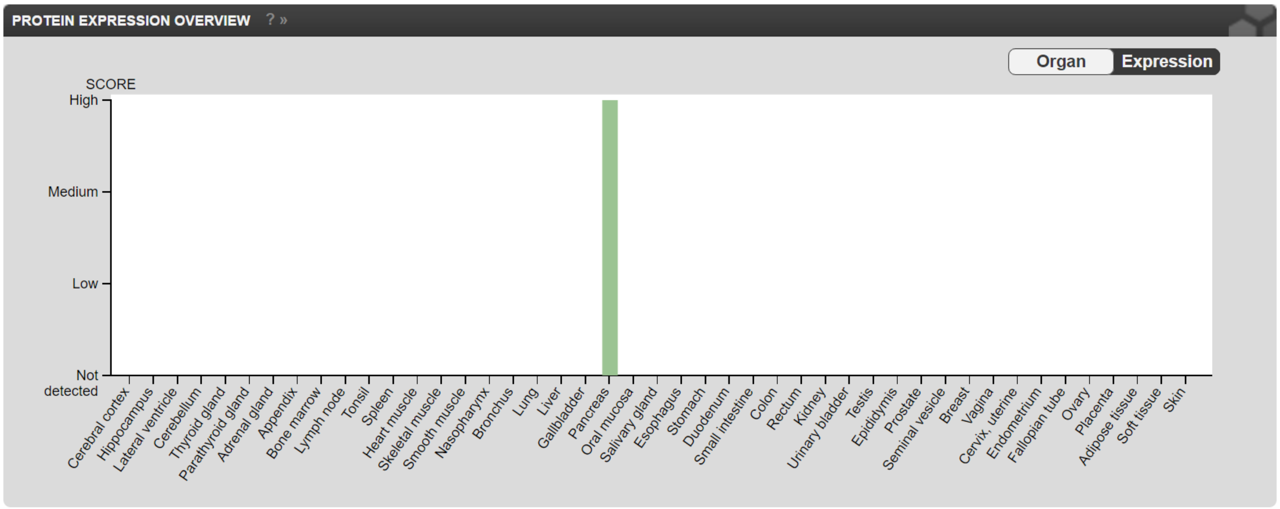
ANTIBODY RELIABILITY ? »

Data reliability description	Antibody staining consistent with RNA expression data.
Data reliability	Supportive based on 4 antibodies. HPA004932 , CAB000048 , CAB012098 , CAB053843

SHOW MORE

RNA AND PROTEIN EXPRESSION SUMMARY ? »

Illustration 64: Human Protein Atlas showing expression level of gene in different tissues


PROTEIN EXPRESSION OVERVIEW ? »

Examining gene expression levels

From the excellent online course “Epigenetic Control” presented by Dr. Marnie Blewitt one gets to know that although genes in all body cells, essentially contain exactly the same genetic DNA code, genes in different body cells are expressed at vastly different levels. Some genes are constitutively switched off in some body cells, which other housekeeping genes are always on in most body cells (such as DNA repair genes). The expression level of genes in different body cells are controlled by transcription binding proteins which are sometimes only expressed in some body cells (and which bind to groups of genes which are simultaneously activated) or epigenetic mechanisms such as chromatin packaging applied by histone modifying enzymes and chromatin remodelling proteins. Long non-coding RNA also play a role to guide these modifying enzymes to target regions of the genome.

It is also known that genes are sequestered into regions of the nucleus where there are much higher concentrations of transcriptional machinery, i.e. transcription factories, while other regions such as the nuclear lamina represent areas where genes are mostly switched off.

Looking at the transcriptome using the Chromosomes View

In order to map the expression levels of genes at a genome wide level, I decided to map the gene expression level that one can obtain from the **Gene Expression Barcode**.

They provide publicly available data from http://barcode.luhs.org/

I quote from their website: *“The barcode algorithm is designed to estimate which genes are expressed and which are unexpressed in a given microarray hybridization. The output of our algorithm is a vector of ones and zeros denoting which genes are estimated to be expressed (ones) and unexpressed (zeros). We call this a gene expression barcode. “*

I downloaded the file: abc.ntc.GPL570.csv at

http://barcode.luhs.org/index.php?page=transcriptome **(Provides the expression barcode)**

http://www.affymetrix.com/ **(Provides the mapping between UCSC genes and AffyID)**

This provided me with an Affymetrix Microarray Chip gene expression level **between 0 and 1** for 54613 different gene transcripts for 131 different cell types in body tissues. (**in some cases allowing you to compare normal with tumour tissue expression.**)

The following 24 tumor tissue types are included (based on the Bar Code File provided)

breast_lobular_cells:tumor
breast_stroma:tumor
breast:tumor
cervix:tumor
colon:tumor
endometrium:tumor
fallopian_tube:tumor
glioblastoma:tumor
glioma:tumor
head_and_neck_epithelial_cells:tumor
head_and_neck_squamous_cell_carcinoma:tumor
kidney:tumor
liver:tumor
lung:tumor

omentum:tumor
ovary:tumor
pancreas:tumor
peritoneum:tumor
pilocytic_astrocytoma:tumor
pituitary:tumor
rectum_mucosa:tumor
sigmoid_colon_mucosa:tumor
tongue_squamous_cells:tumor
uterus:tumor

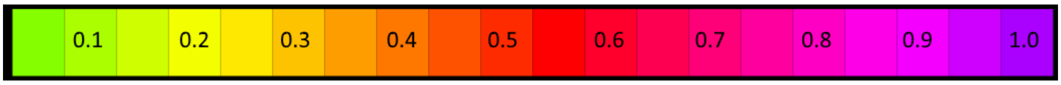
By looking at the transcriptome for different cell types, **the goal was to see if a graphical bird's eye view of the genome**, will not perhaps give insight as to what regions of the chromosomes are situated in transcriptional factories.

The difference is that, as opposed to an Affimetrix chip, in this case the gene expression is shown directly on the chromosomes, allowing clustering of expressed genes to be observed much better.

I was expecting to see clusters of genes expressed at higher levels in different regions of the chromosomes. I proceeded to plot the expression level of genes on both the **Chromosomes View** as well as the **Main Genome View**.

Normally, genes are coloured according to their gene names. Genes which the same name will be displayed in the same colour. When “Random color” is not selected, then the biochemical pathway is used (where available) to colour genes the same.

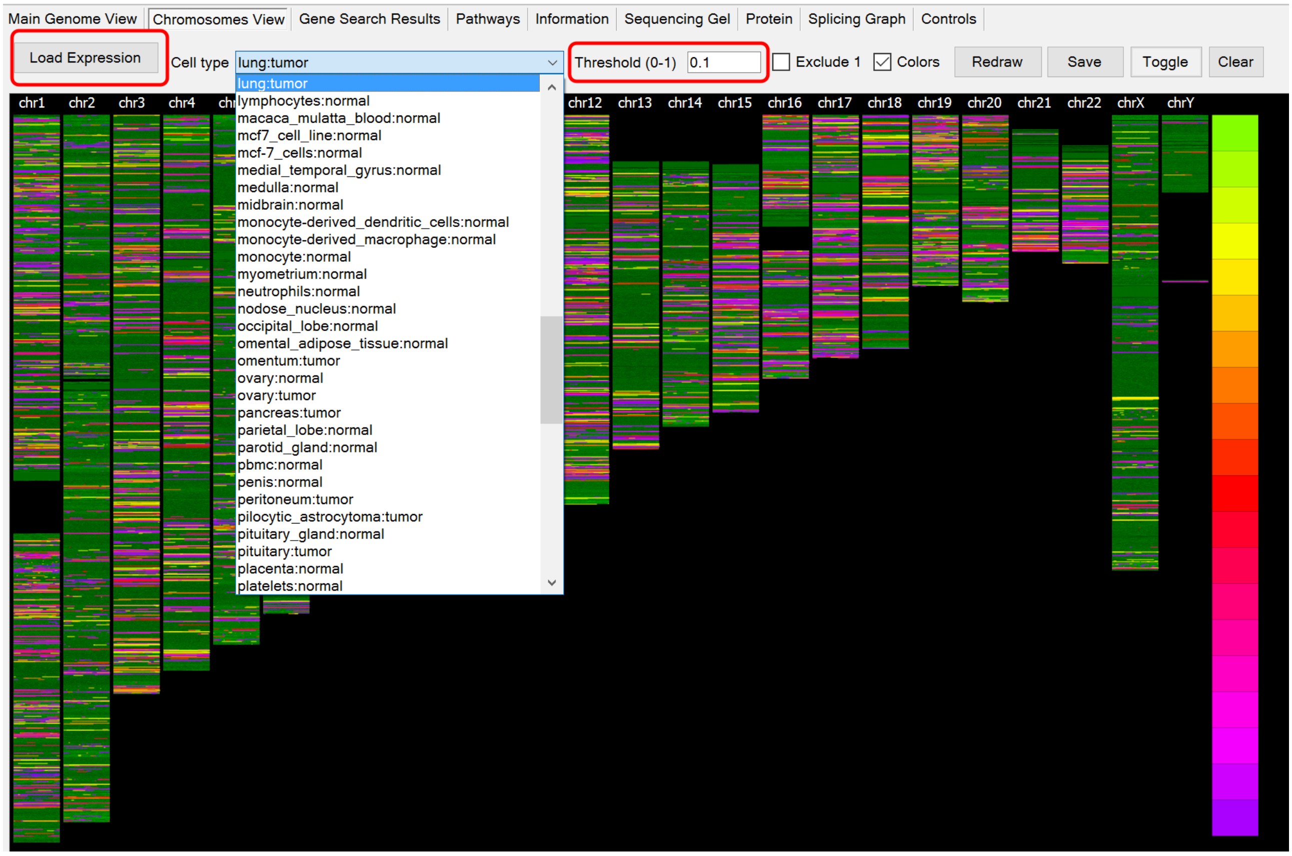
When in this (gene expression) mode, the colours of genes are neither determined by their biochemical pathway or their gene names, but the colours are determined by the expression level on the following scale:



This enables you to clearly see which genes are expressed highly in a specific tissue type.

The software maps between the AffyID (Affimetrix microarray Id) and the UCSC gene Id in order to generate a **colour map** for the expressed genes. **You can also specify a threshold or cut-off value between 0 and 1 to only display genes expressed at a level higher than the threshold value.**

The functionality is loaded by clicking on **Load Expression**. This will load the Affimetrix expression Bar Code values for all the selected cell types. The threshold value will only display genes expressed at a level higher than the specified value. The "**Exclude 1**" option allows you to remove genes expressed at a level of 100%, revealing more of the genes which are differentially expressed (due to chromatin differences or cell specific transcription factors). From here on the threshold setting will also be used as a filter on the **Main Genome View** (allowing you to only see highly expressed genes on the main single chromosome view). You can also choose to switch the colour scale on or off. In order to draw the genes on the **Chromosomes View** (which displays a **karyotype-like view of all the human chromosomes**).



Cell type lung:tumor

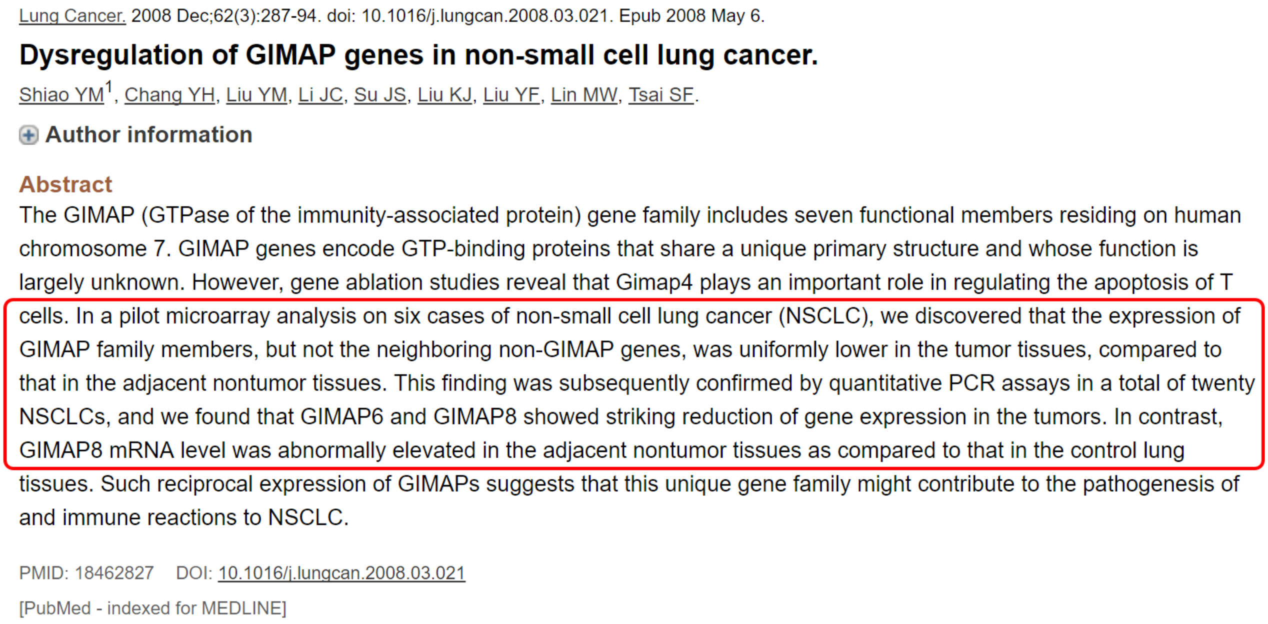
Threshold (0-1) 0.1

□ Exclude 1 ☑ Colors

As discussed earlier, you can move the mouse over the **Chromosomes View**, which will then display that region of the chromosomes in the **DNA View**. You can also click on the **Chromosomes View** to jump to that position in the **Main Genome View**.

Chromosomes are drawn to scale and the purple regions represent the centromeric regions of the chromosomes.

Lets look at an example of how this could be useful. I was searching the literature for gene dysregulation in lung cancer and stumbled up the following article:



Lung Cancer. 2008 Dec;62(3):287-94. doi: 10.1016/j.lungcan.2008.03.021. Epub 2008 May 6.

Dysregulation of GIMAP genes in non-small cell lung cancer.

Shiao YM[1], Chang YH, Liu YM, Li JC, Su JS, Liu KJ, Liu YF, Lin MW, Tsai SF.

⊕ Author information

Abstract

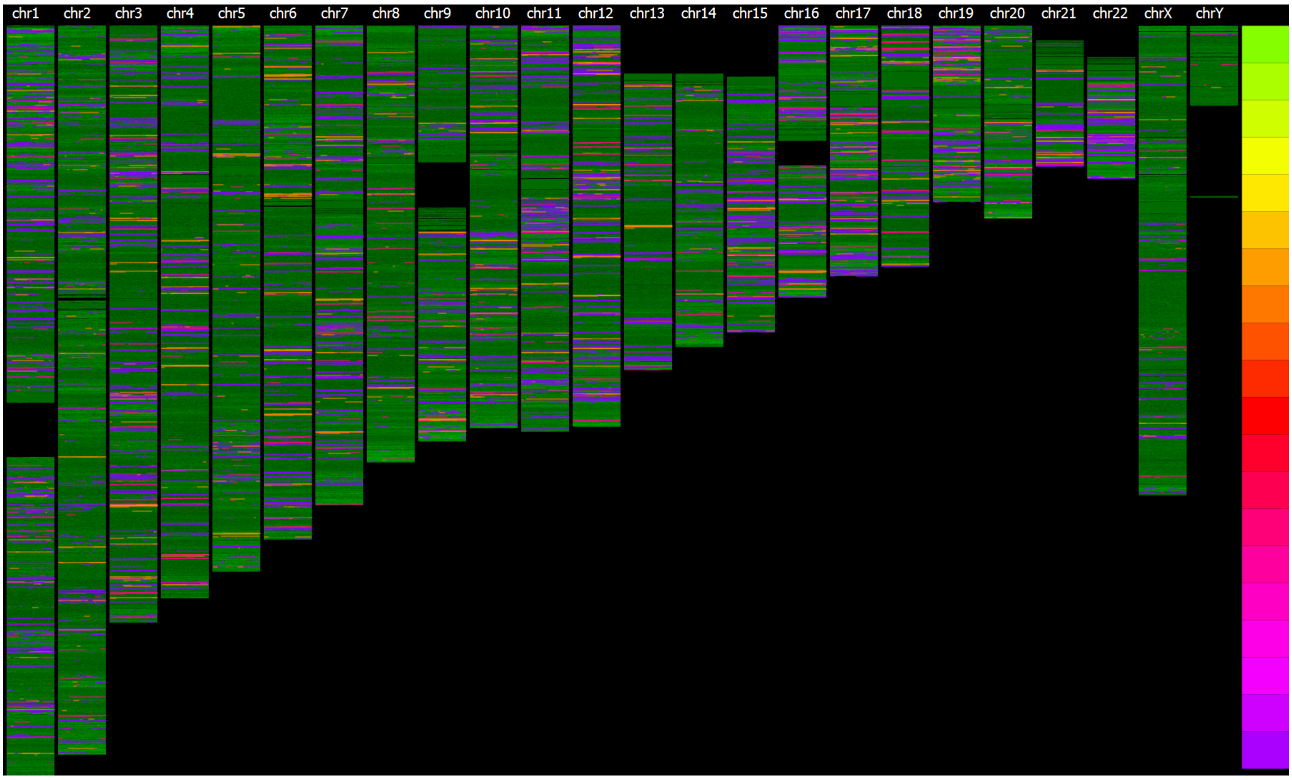
The GIMAP (GTPase of the immunity-associated protein) gene family includes seven functional members residing on human chromosome 7. GIMAP genes encode GTP-binding proteins that share a unique primary structure and whose function is largely unknown. However, gene ablation studies reveal that Gimap4 plays an important role in regulating the apoptosis of T cells. In a pilot microarray analysis on six cases of non-small cell lung cancer (NSCLC), we discovered that the expression of GIMAP family members, but not the neighboring non-GIMAP genes, was uniformly lower in the tumor tissues, compared to that in the adjacent nontumor tissues. This finding was subsequently confirmed by quantitative PCR assays in a total of twenty NSCLCs, and we found that GIMAP6 and GIMAP8 showed striking reduction of gene expression in the tumors. In contrast, GIMAP8 mRNA level was abnormally elevated in the adjacent nontumor tissues as compared to that in the control lung tissues. Such reciprocal expression of GIMAPs suggests that this unique gene family might contribute to the pathogenesis of and immune reactions to NSCLC.

PMID: 18462827   DOI: 10.1016/j.lungcan.2008.03.021

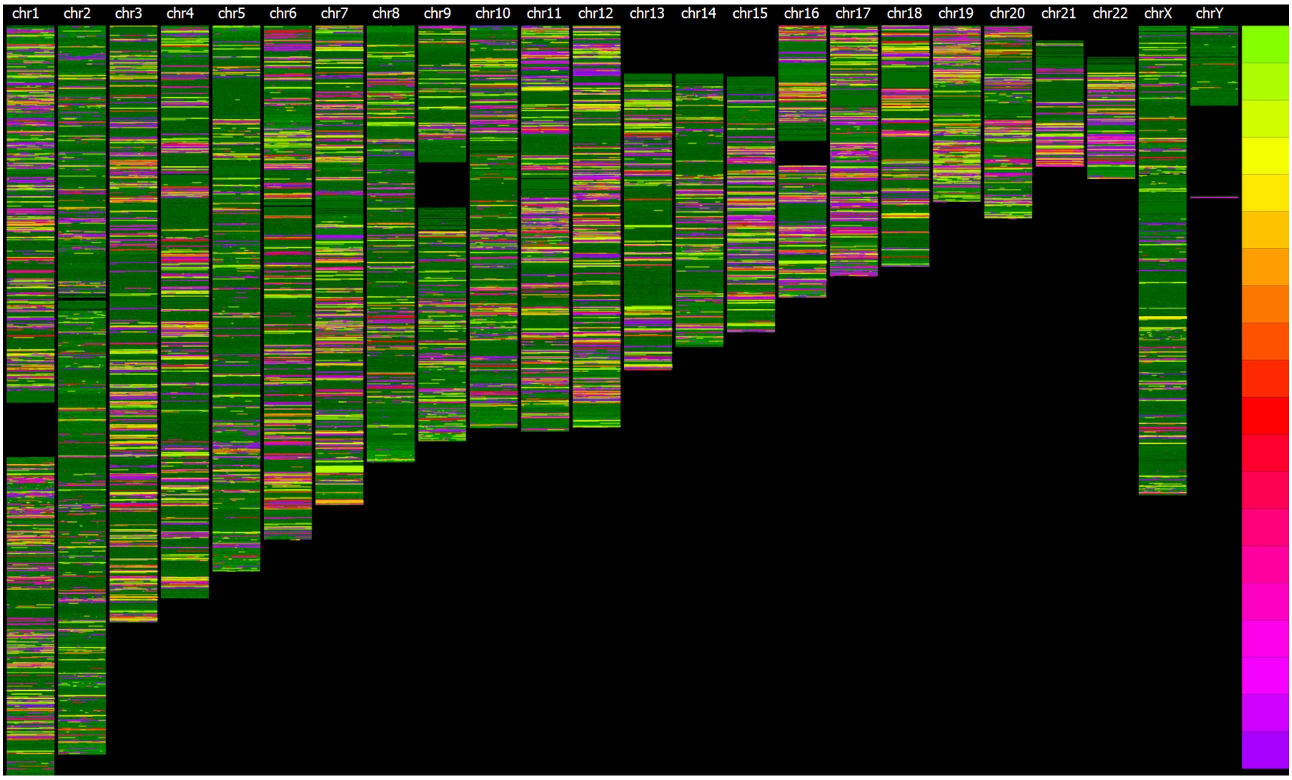
[PubMed - indexed for MEDLINE]

I then decided to compare the genome wide expression of **normal lung tissue** with that of **lung tissue with cancer:**

The following picture displays the **high** expression of most genes which are expressed in normal lung cells: (The threshold was set at 0.05 which means that most genes expressed where at high levels, except the genes which are not expressed in lung tissue due to epigenetic control)



When I then loaded the same for **lung cancer cells**, it looks as follows:

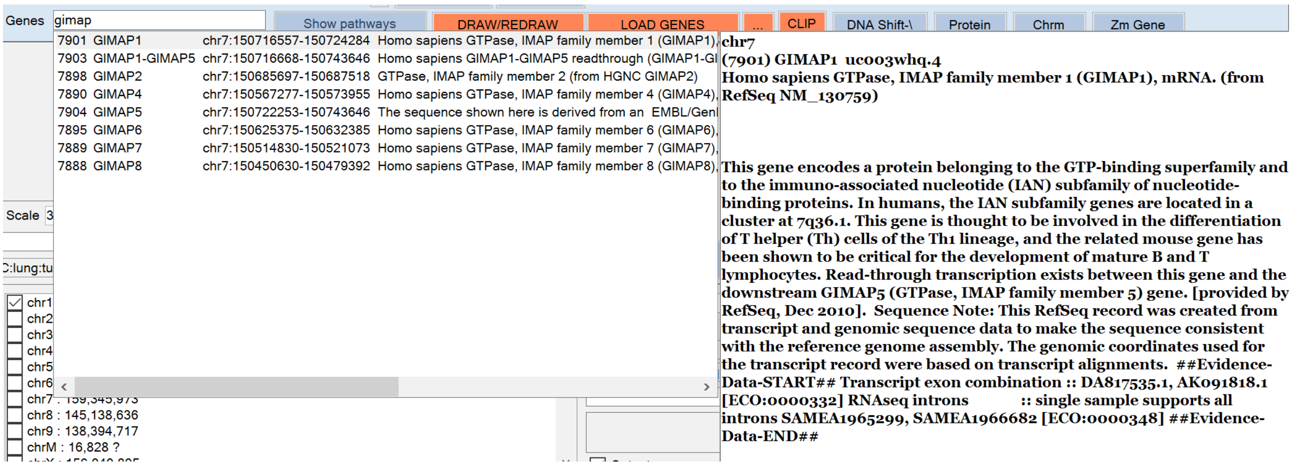


One can now visually observe the extent of gene dysregulation in lung cancer tissue. Notice how many genes now have lower expression. In order to better compare, the **F12** key will toggle between the current and the last expression setting. (Or by clicking on the **Toggle** button).

This toggle function will help you to locate region of the genome where there are gene expression differences.

In the **Epigenetic Control** course (from the University of Melbourne) mentioned earlier I learnt that, in cancer, there is a genome wide increase in DNA methylation at CpG Islands (often situated at gene promoters), which has the effect of turning gene expression off. (This might well be what is happening here).

To further investigate this, one can not type in the GIMAP gene into the Gene Lookup field:



Genes: gimap

```
7901  GIMAP1        chr7:150716557-150724284  Homo sapiens GTPase, IMAP family member 1 (GIMAP1),
7903  GIMAP1-GIMAP5 chr7:150716668-150743646  Homo sapiens GIMAP1-GIMAP5 readthrough (GIMAP1-GI
7898  GIMAP2        chr7:150685697-150687518  GTPase, IMAP family member 2 (from HGNC GIMAP2)
7890  GIMAP4        chr7:150567277-150573955  Homo sapiens GTPase, IMAP family member 4 (GIMAP4),
7904  GIMAP5        chr7:150722253-150743646  The sequence shown here is derived from an  EMBL/Genl
7895  GIMAP6        chr7:150625375-150632385  Homo sapiens GTPase, IMAP family member 6 (GIMAP6),
7889  GIMAP7        chr7:150514830-150521073  Homo sapiens GTPase, IMAP family member 7 (GIMAP7),
7888  GIMAP8        chr7:150450630-150479392  Homo sapiens GTPase, IMAP family member 8 (GIMAP8),
```

☑ chr1
□ chr2
□ chr3
□ chr4
□ chr5
□ chr6
□ chr7 : 159,345,973
□ chr8 : 145,138,636
□ chr9 : 138,394,717
□ chrM : 16,828 ?

chr7
(7901) GIMAP1 uc003whq.4
Homo sapiens GTPase, IMAP family member 1 (GIMAP1), mRNA. (from RefSeq NM_130759)

This gene encodes a protein belonging to the GTP-binding superfamily and to the immuno-associated nucleotide (IAN) subfamily of nucleotide-binding proteins. In humans, the IAN subfamily genes are located in a cluster at 7q36.1. This gene is thought to be involved in the differentiation of T helper (Th) cells of the Th1 lineage, and the related mouse gene has been shown to be critical for the development of mature B and T lymphocytes. Read-through transcription exists between this gene and the downstream GIMAP5 (GTPase, IMAP family member 5) gene. [provided by RefSeq, Dec 2010]. Sequence Note: This RefSeq record was created from transcript and genomic sequence data to make the sequence consistent with the reference genome assembly. The genomic coordinates used for the transcript record were based on transcript alignments. ##Evidence-Data-START## Transcript exon combination :: DA817535.1, AK091818.1 [ECO:0000332] RNAseq introns :: single sample supports all introns SAMEA1965299, SAMEA1966682 [ECO:0000348] ##Evidence-Data-END##

Pressing the DOWN ARROW KEY and scrolling though the GIMAP genes will display each gene's information in the panel on the right. Press ENTER to select and navigate to the gene.

The colours in the **Main Genome View** will now show the genes as purple with the **Normal Lung gene expression filter.**

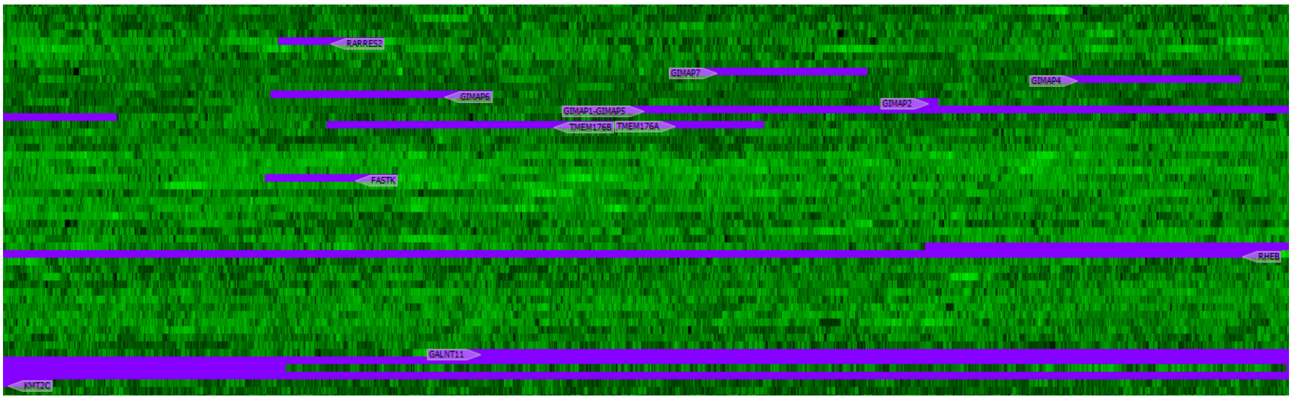


Illustration 65: Normal Lung tissue Gene expression display of GIMAP and surrounding genes at chr7:150,083,281-151,683,811

When you now press **F12**, the display will toggle back to the **Lung Cancer gene expression filter:**

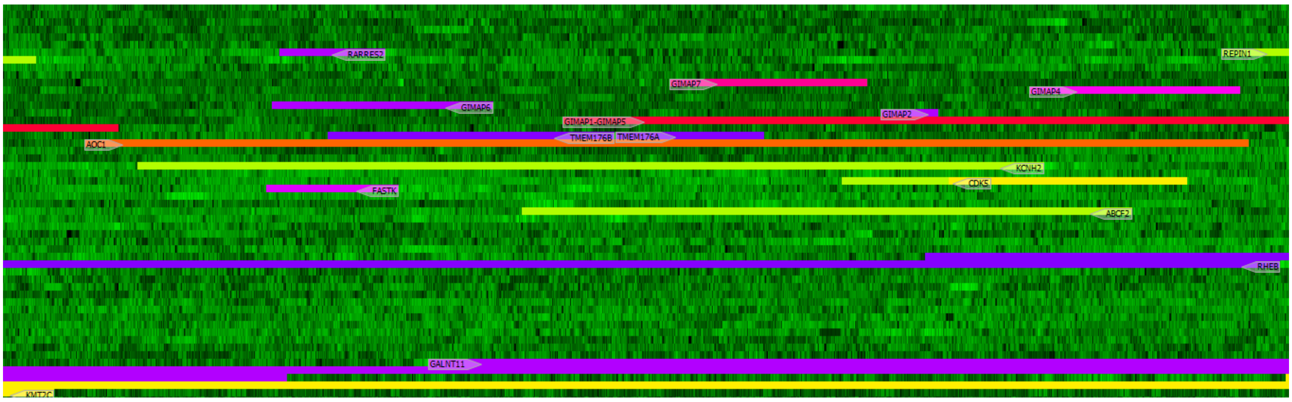
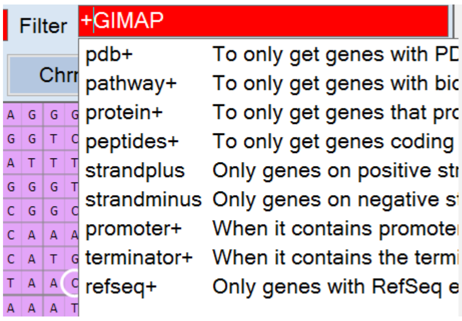


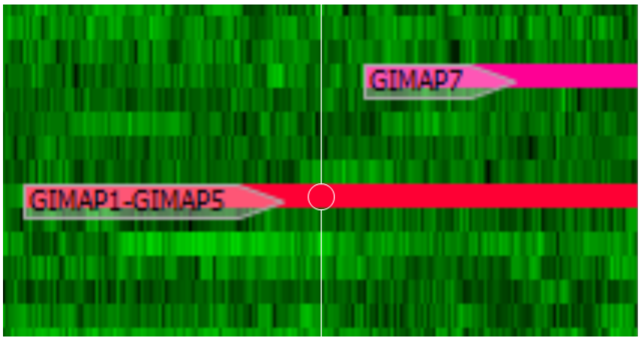
Illustration 66: Lung Cancer tissue Gene expression display of GIMAP and surrounding genes at chr7:150,083,281-151,683,811

If you want to get rid of any non-GIMAP genes which are still in the display, you can add the GIMAP name in the **Filter Field:**



+GIMAP

pdb+	To only get genes with PD
pathway+	To only get genes with bic
protein+	To only get genes that prc
peptides+	To only get genes coding
strandplus	Only genes on positive str
strandminus	Only genes on negative s
promoter+	When it contains promote
terminator+	When it contains the termi
refseq+	Only genes with RefSeq e



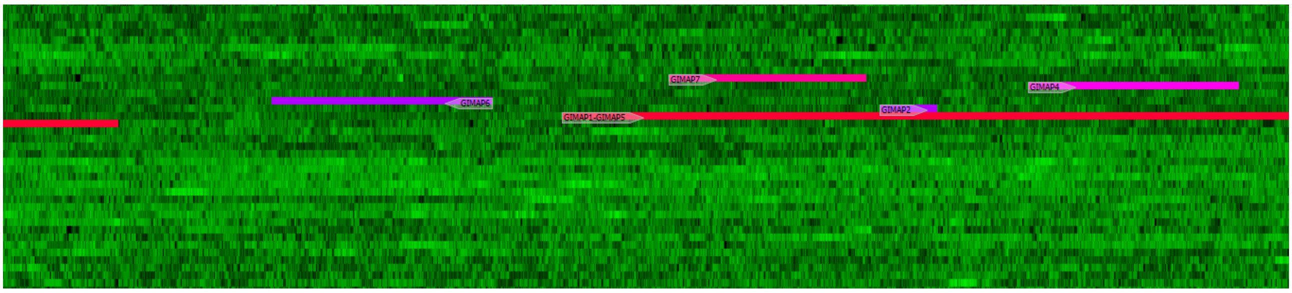
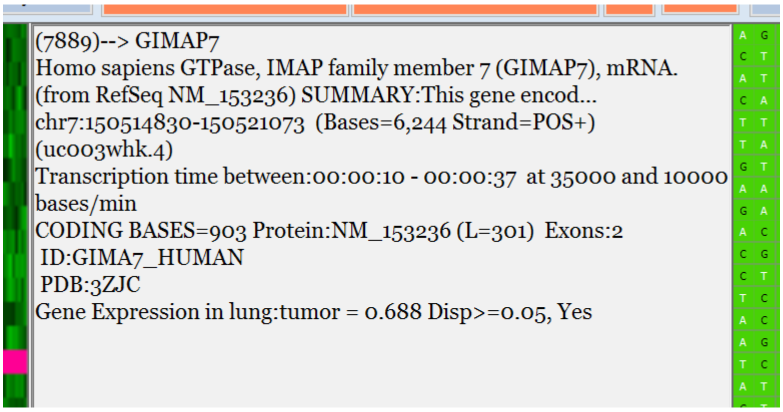
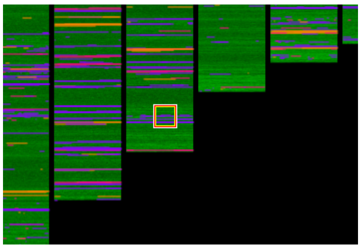


Illustration 67: Lung Cancer tissue Gene expression display of GIMAP only genes at chr7:150,083,281-151,683,811

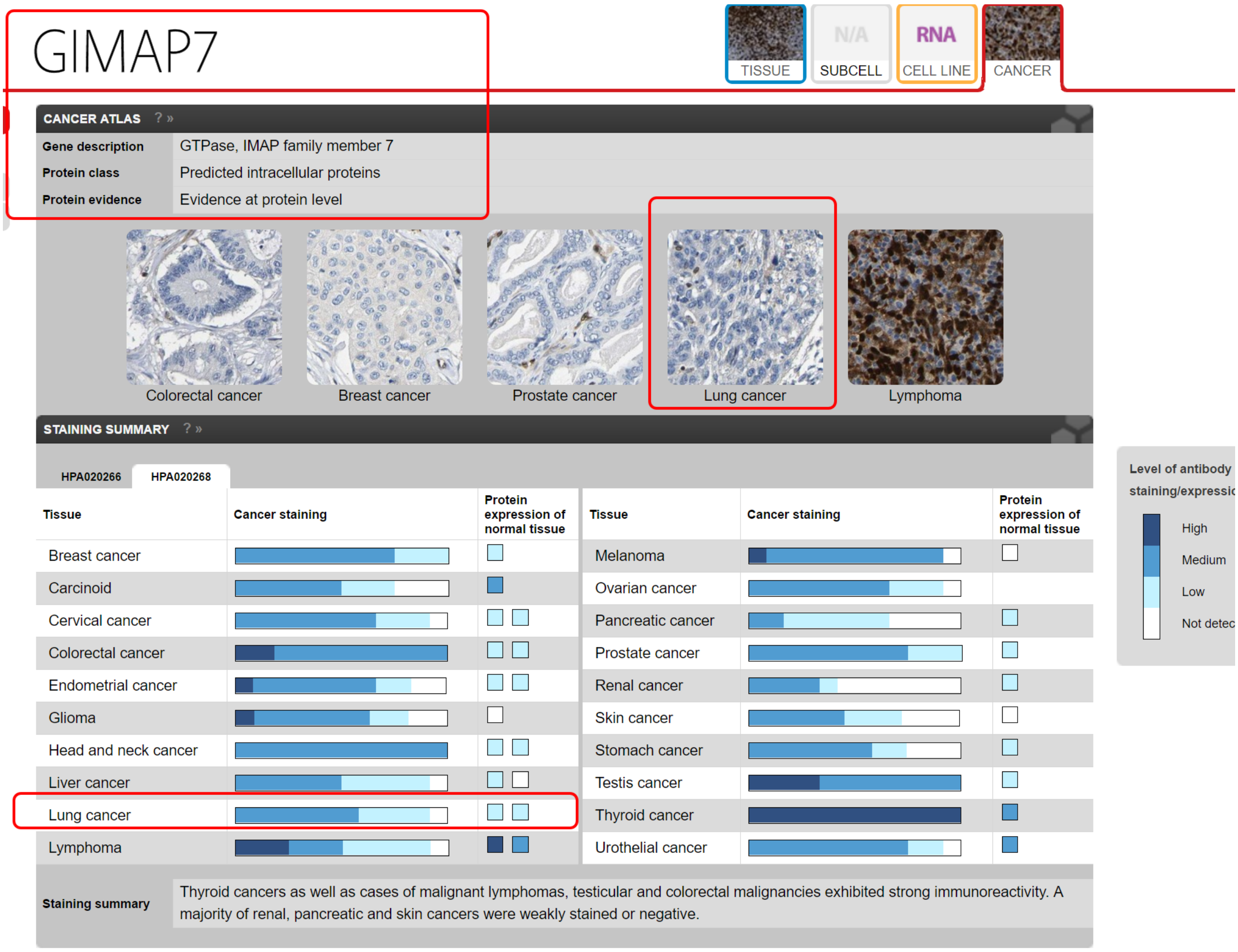
When you hover the mouse over one of these genes (or when you centre the **DNA View** on one of these genes, you will now see the expression level of the gene in the **Information Display** in the centre panel:



(7889)--> GIMAP7
Homo sapiens GTPase, IMAP family member 7 (GIMAP7), mRNA.
(from RefSeq NM_153236) SUMMARY:This gene encod...
chr7:150514830-150521073 (Bases=6,244 Strand=POS+)
(uc003whk.4)
Transcription time between:00:00:10 - 00:00:37 at 35000 and 10000 bases/min
CODING BASES=903 Protein:NM_153236 (L=301) Exons:2
ID:GIMA7_HUMAN
PDB:3ZJC
Gene Expression in lung:tumor = 0.688 Disp>=0.05, Yes



Another context menu feature is to now go to the **Human Protein Atlas** for one of these genes and then select the **Cancer Tissue** option:



GIMAP7

CANCER ATLAS ? »

Gene description	GTPase, IMAP family member 7
Protein class	Predicted intracellular proteins
Protein evidence	Evidence at protein level

STAINING SUMMARY ? »

HPA020266 HPA020268

Tissue	Cancer staining	Protein expression of normal tissue	Tissue	Cancer staining	Protein expression of normal tissue
Breast cancer			Melanoma		
Carcinoid			Ovarian cancer		
Cervical cancer			Pancreatic cancer		
Colorectal cancer			Prostate cancer		
Endometrial cancer			Renal cancer		
Glioma			Skin cancer		
Head and neck cancer			Stomach cancer		
Liver cancer			Testis cancer		
Lung cancer			Thyroid cancer		
Lymphoma			Urothelial cancer		

Staining summary: Thyroid cancers as well as cases of malignant lymphomas, testicular and colorectal malignancies exhibited strong immunoreactivity. A majority of renal, pancreatic and skin cancers were weakly stained or negative.

Level of antibody staining/expressic

- High
- Medium
- Low
- Not detec

From the picture above one can again see how this gene's expression is lower in lung cancer tissue.

The **Chromosomes View** also allows you to export the “Karyotype” to a **png image file**.

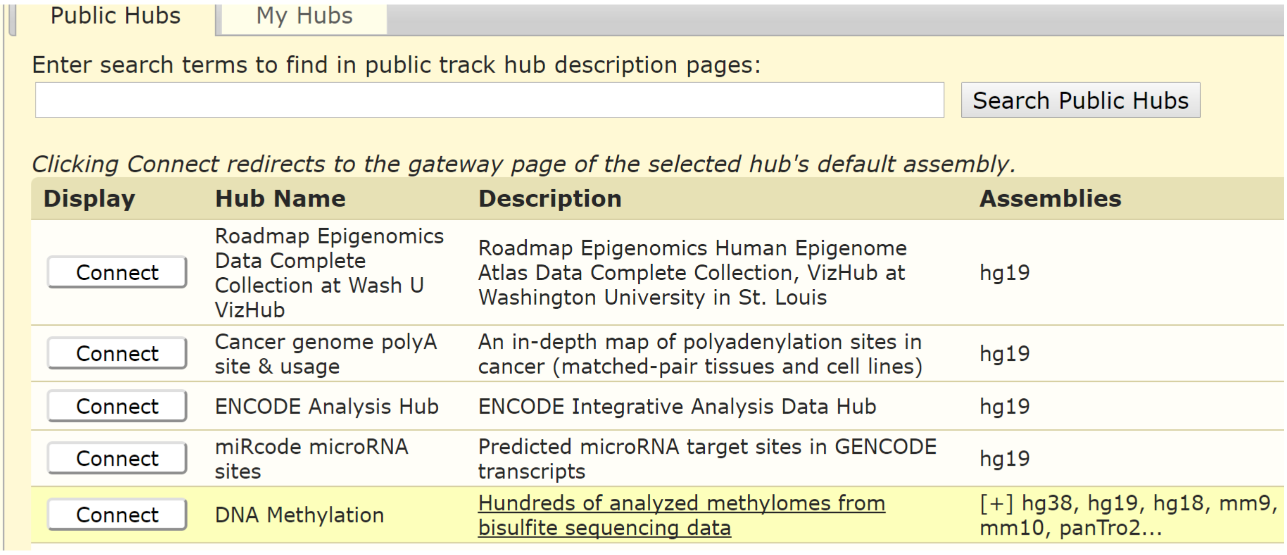
Visualising DNA Methylation from BigWig and WigBed files

Getting DNA methylation data from publicly available sources

The UCSC Genome browser provides the ability to link external data tracks into the browser's display. These tracks include **DNA Methylation** data, which is obtained using bisulfite sequencing.

This can be obtained via the page:

http://genome.ucsc.edu/cgi-bin/hgHubConnect

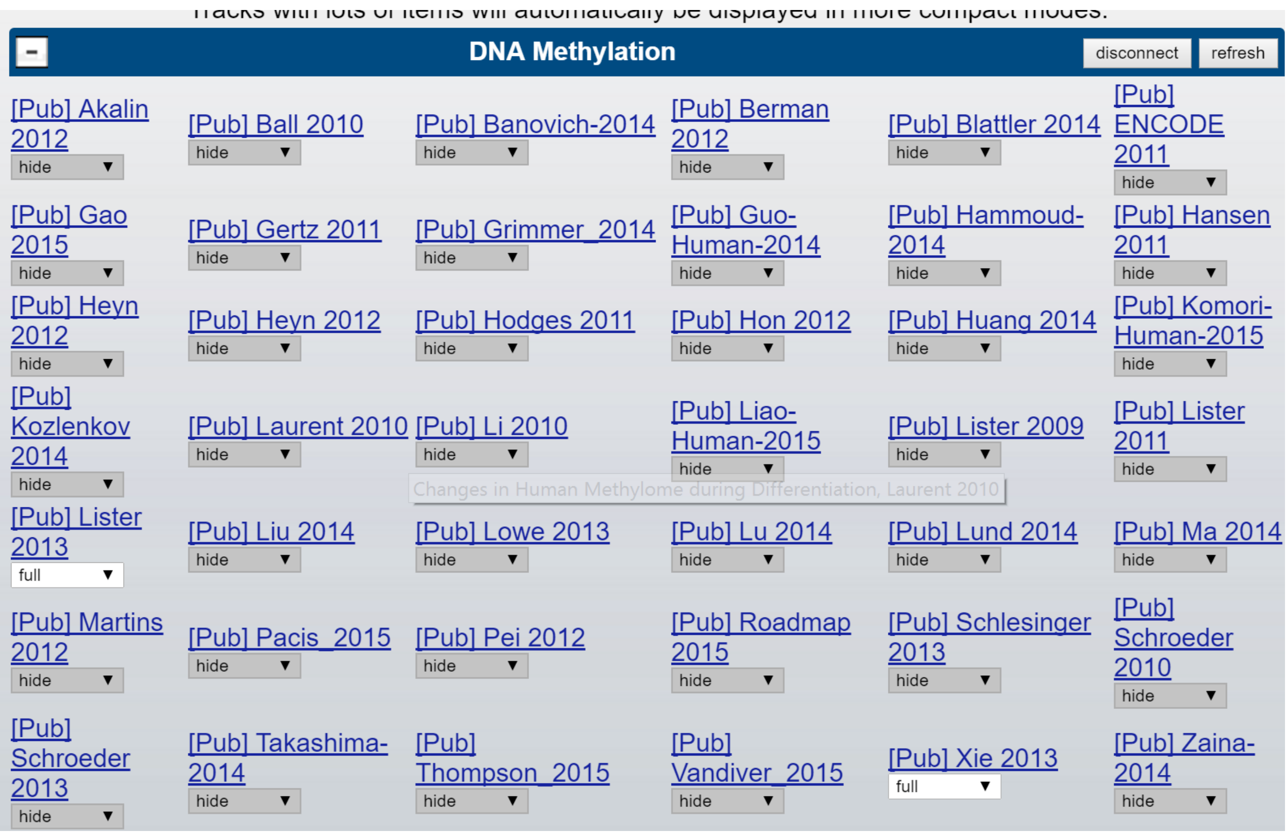


Enter search terms to find in public track hub description pages:

Clicking Connect redirects to the gateway page of the selected hub's default assembly.

Display	Hub Name	Description	Assemblies
Connect	Roadmap Epigenomics Data Complete Collection at Wash U VizHub	Roadmap Epigenomics Human Epigenome Atlas Data Complete Collection, VizHub at Washington University in St. Louis	hg19
Connect	Cancer genome polyA site & usage	An in-depth map of polyadenylation sites in cancer (matched-pair tissues and cell lines)	hg19
Connect	ENCODE Analysis Hub	ENCODE Integrative Analysis Data Hub	hg19
Connect	miRcode microRNA sites	Predicted microRNA target sites in GENCODE transcripts	hg19
Connect	DNA Methylation	Hundreds of analyzed methylomes from bisulfite sequencing data	[+] hg38, hg19, hg18, mm9, mm10, panTro2...

When you browse to a specific position (which can can obtain by the context menu – **copy position to clipboard**), you can then go to the position in the UCSC genome.

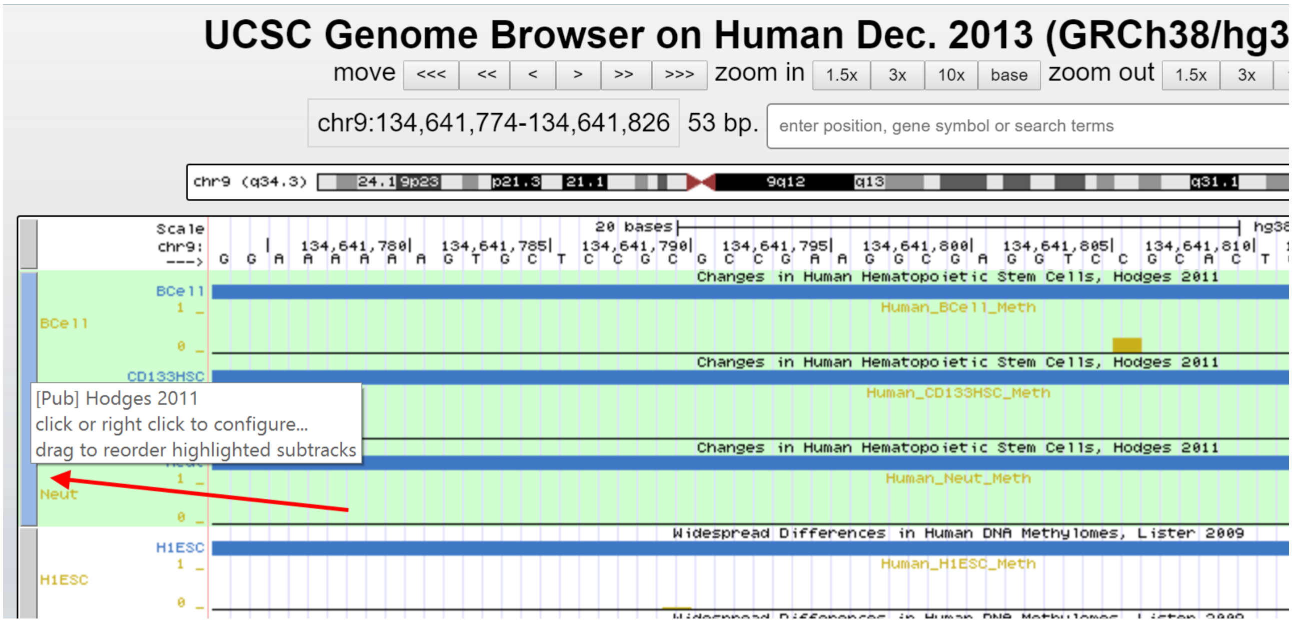


Tracks with lots of items will automatically be displayed in more compact modes.

[Pub] Akalin 2012 hide ▼	[Pub] Ball 2010 hide ▼	[Pub] Banovich-2014 hide ▼	[Pub] Berman 2012 hide ▼	[Pub] Blattler 2014 hide ▼	[Pub] ENCODE 2011 hide ▼
[Pub] Gao 2015 hide ▼	[Pub] Gertz 2011 hide ▼	[Pub] Grimmer_2014 hide ▼	[Pub] Guo-Human-2014 hide ▼	[Pub] Hammoud-2014 hide ▼	[Pub] Hansen 2011 hide ▼
[Pub] Heyn 2012 hide ▼	[Pub] Heyn 2012 hide ▼	[Pub] Hodges 2011 hide ▼	[Pub] Hon 2012 hide ▼	[Pub] Huang 2014 hide ▼	[Pub] Komori-Human-2015 hide ▼
[Pub] Kozlenkov 2014 hide ▼	[Pub] Laurent 2010 hide ▼	[Pub] Li 2010 hide ▼	[Pub] Liao-Human-2015 hide ▼	[Pub] Lister 2009 hide ▼	[Pub] Lister 2011 hide ▼
[Pub] Lister 2013 full ▼	[Pub] Liu 2014 hide ▼	[Pub] Lowe 2013 hide ▼	[Pub] Lu 2014 hide ▼	[Pub] Lund 2014 hide ▼	[Pub] Ma 2014 hide ▼
[Pub] Martins 2012 hide ▼	[Pub] Pacis_2015 hide ▼	[Pub] Pei 2012 hide ▼	[Pub] Roadmap 2015 hide ▼	[Pub] Schlesinger 2013 hide ▼	[Pub] Schroeder 2010 hide ▼
[Pub] Schroeder 2013 hide ▼	[Pub] Takashima-2014 hide ▼	[Pub] Thompson_2015 hide ▼	[Pub] Vandiver_2015 hide ▼	[Pub] Xie 2013 full ▼	[Pub] Zaina-2014 hide ▼

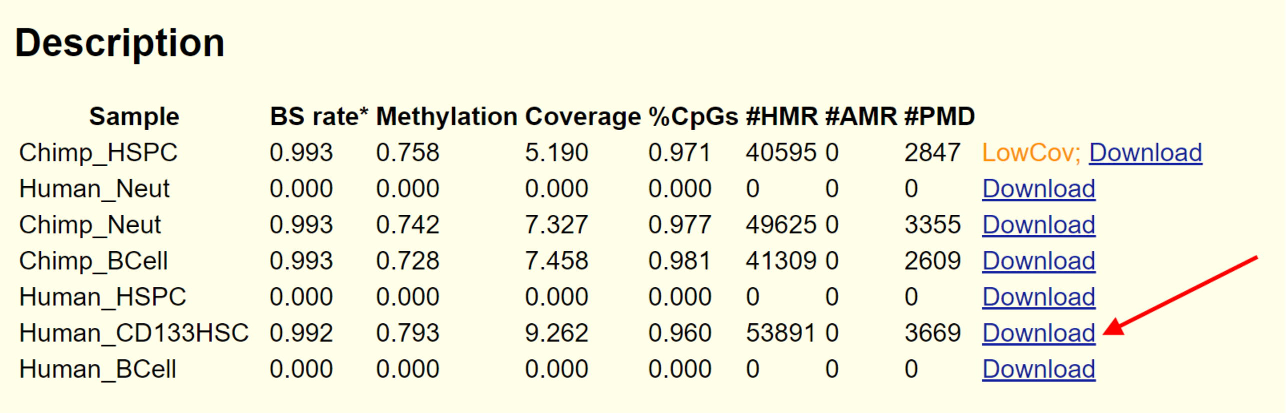
Changes in Human Methylome during Differentiation, Laurent 2010

Then click on the side bar to get access to the DNA Methylation source settings:



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg3
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x
chr9:134,641,774-134,641,826 53 bp.
enter position, gene symbol or search terms
chr9 (q34.3)
Scale
chr9:
20 bases
134,641,780
134,641,785
134,641,790
134,641,795
134,641,800
134,641,805
134,641,810
BCell
Changes in Human Hematopoietic Stem Cells, Hodges 2011
Human_BCell_Meth
CD133HSC
Human_CD133HSC_Meth
Neut
Human_Neut_Meth
H1ESC
Widespread Differences in Human DNA Methylomes, Lister 2009
Human_H1ESC_Meth

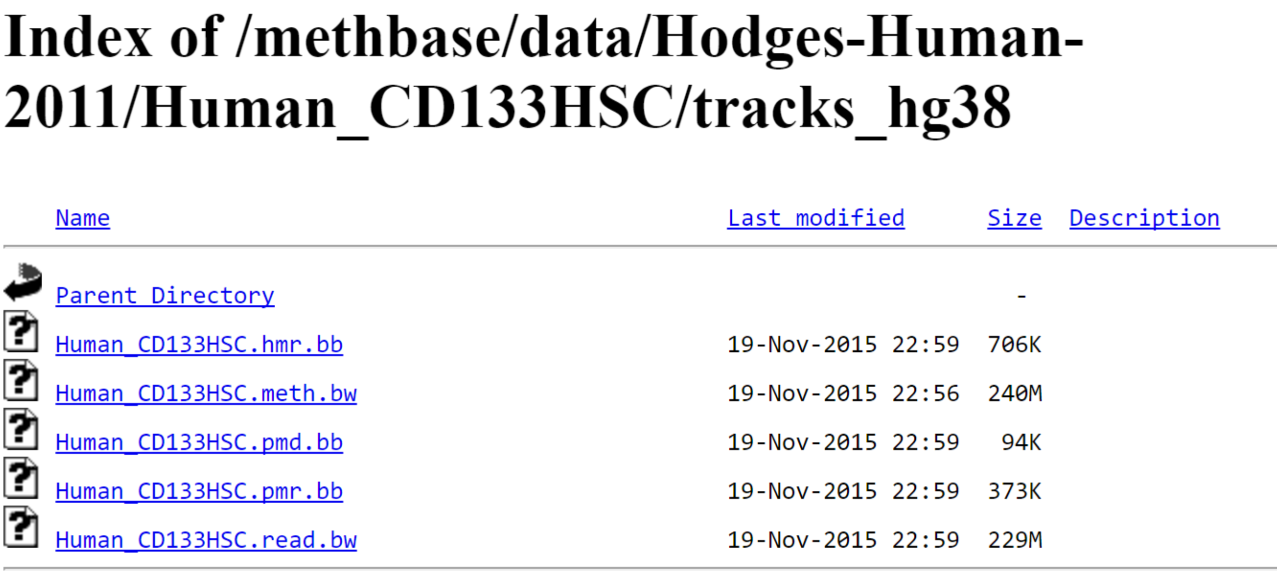
[Pub] Hodges 2011
click or right click to configure...
drag to reorder highlighted subtracks



Description

Sample	BS rate*	Methylation	Coverage	%CpGs	#HMR	#AMR	#PMD	
Chimp_HSPC	0.993	0.758	5.190	0.971	40595	0	2847	LowCov; Download
Human_Neut	0.000	0.000	0.000	0.000	0	0	0	Download
Chimp_Neut	0.993	0.742	7.327	0.977	49625	0	3355	Download
Chimp_BCell	0.993	0.728	7.458	0.981	41309	0	2609	Download
Human_HSPC	0.000	0.000	0.000	0.000	0	0	0	Download
Human_CD133HSC	0.992	0.793	9.262	0.960	53891	0	3669	Download
Human_BCell	0.000	0.000	0.000	0.000	0	0	0	Download

You can now download the tracks with DNA Methylation as **BigWig (Containing the methylation valies) and BigBed (containing the annotation names)** files from here.



Index of /methbase/data/Hodges-Human-2011/Human_CD133HSC/tracks_hg38

Name	Last modified	Size	Description
Parent Directory		-	
Human_CD133HSC.hmr.bb	19-Nov-2015 22:59	706K	
Human_CD133HSC.meth.bw	19-Nov-2015 22:56	240M	
Human_CD133HSC.pmd.bb	19-Nov-2015 22:59	94K	
Human_CD133HSC.pmr.bb	19-Nov-2015 22:59	373K	
Human_CD133HSC.read.bw	19-Nov-2015 22:59	229M	

Here is a full description of the process to obtain the data:

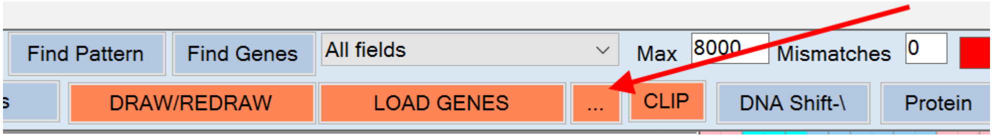
http://smithlabresearch.org/software/methbase/

Displaying the data in the software

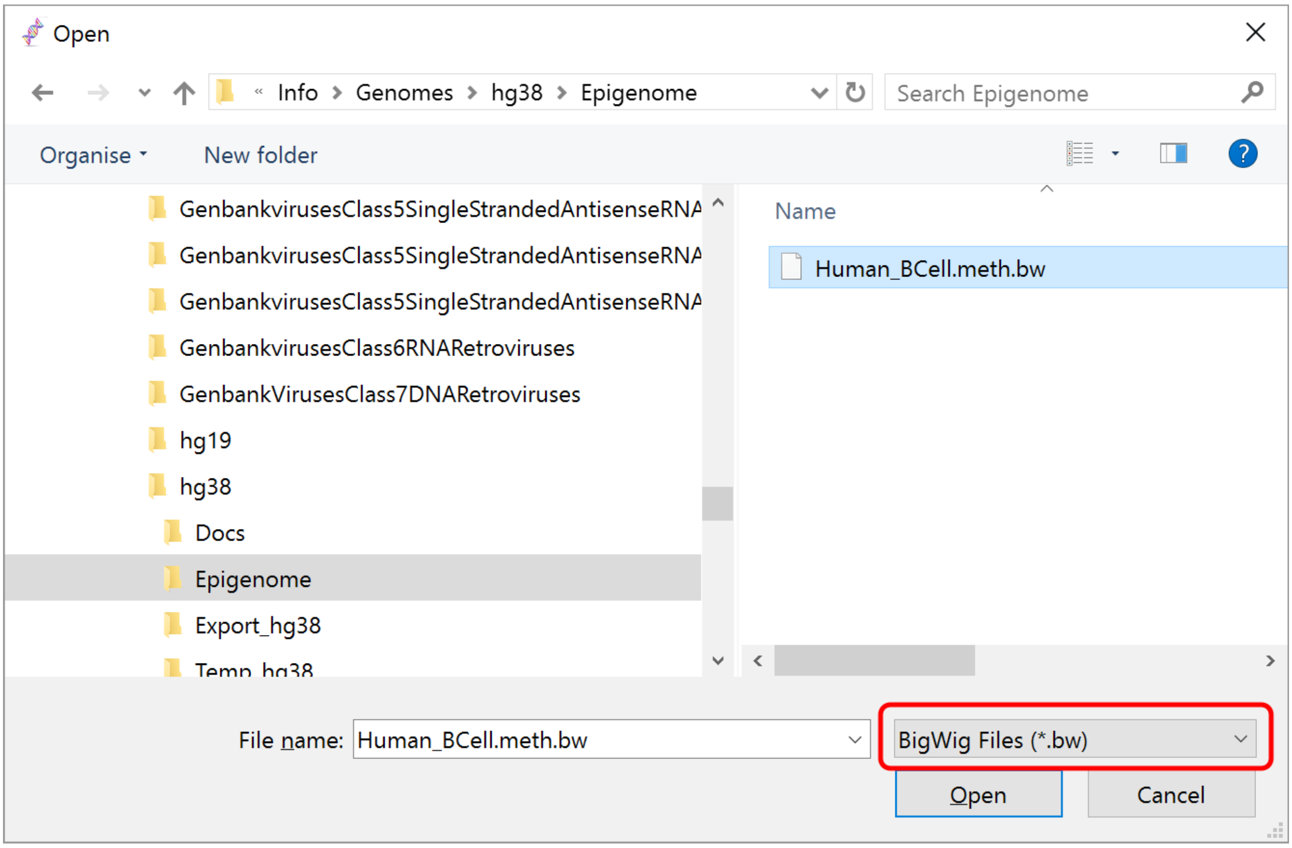
Because of the huge size of these bigwig files, they are not all loaded into memory as that would take too much space. Instead, the BigWig files are queried only for the region of interest using the indexed structure of the file format.

In order to open a BigWig file, first select the correct genome, eg. HG38 which is the default.

Now draw the chromosome of interest by selecting the correct chromosome in the sequence list on the left and then click **DRAW/REDRAW**. This will display the **Main Genome View** for the selected chromosome. Then load the gene annotations by clicking on the **LOAD GENES** button. (This is optional, but will allow you to see the DNA methylation in the context of the genes and gene promoters). Now load the BigWig file as a layer by clicking on the “Browse” button:



Then select the import file type as **BigWig** files:


Open
« Info › Genomes › hg38 › Epigenome
Search Epigenome
Organise
New folder
GenbankvirusesClass5SingleStrandedAntisenseRNA
GenbankvirusesClass5SingleStrandedAntisenseRNA
GenbankvirusesClass5SingleStrandedAntisenseRNA
GenbankvirusesClass6RNARetroviruses
GenbankVirusesClass7DNARetroviruses
hg19
hg38
Docs
Epigenome
Export_hg38
Temp_hg38
Name
Human_BCell.meth.bw
Open
Cancel

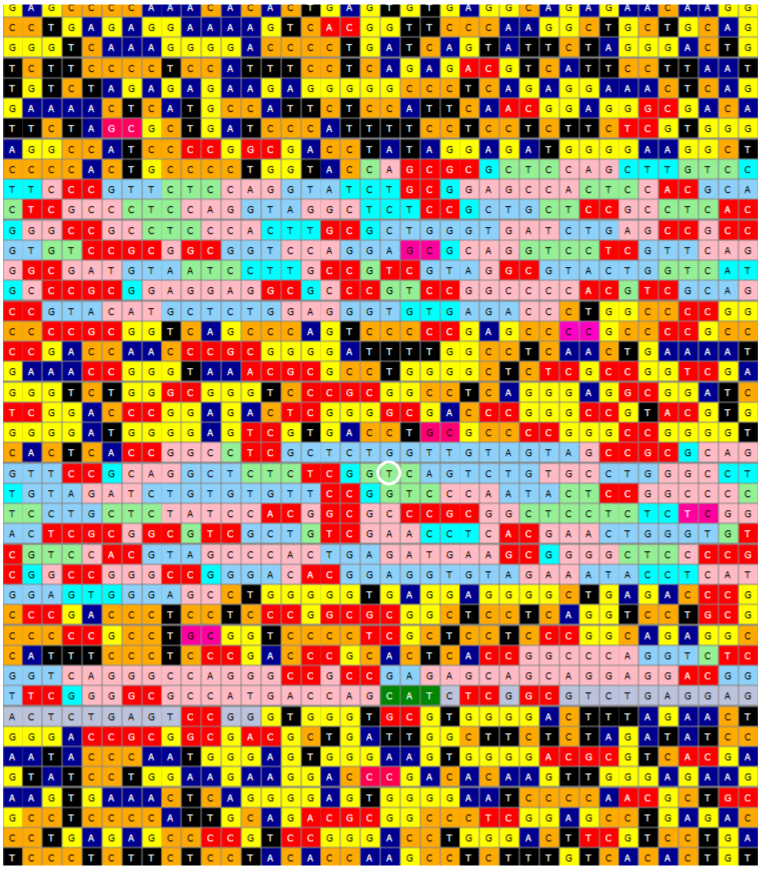
File name: Human_BCell.meth.bw

BigWig Files (*.bw)

This file contains the DNA methylome of B-cells.

When we now look at genes more closely, we can examine the DNA methylation patterns across entire genes. It is striking how genes often contain DNA methylation close to their 5'UTR and promoter regions. It is known that special methyl CpG Binding proteins (**such as** MECP2) control the activation and de-activation of genes by binding to DNA methylation close to the promoters of genes.

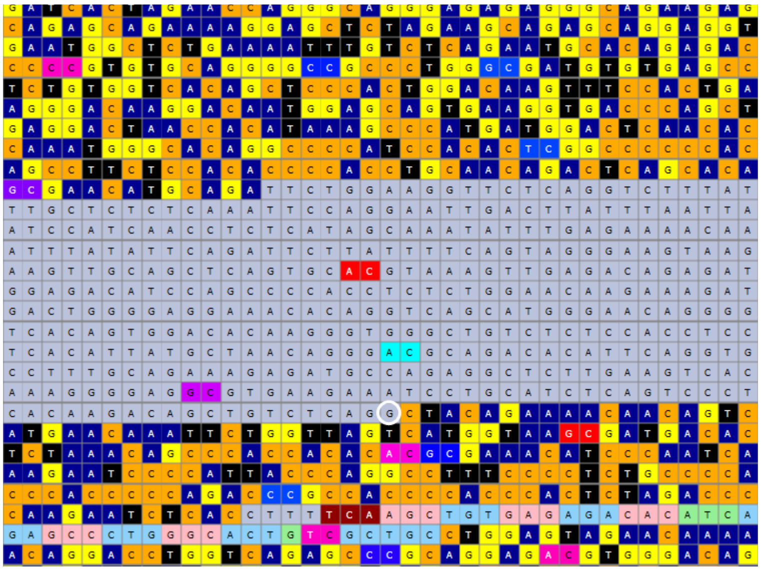
Here is the **DNA View** of the Major Histocompatibility complex HLA-B gene (encoded on the negative strand)



```
G A G C C C C A A A C A C A C T G A G T G T G A G G C A G A G A A C A A G G
C C T G A G A G G A A A A G T C A C G G T T C C C A A G G C T G C T G C A G
G G G T C A A A G G G G A C C C C T G A T C A G T A T T C T A G G G A C T G
T C T T C C C C T C C A T T T C C T C A G A G A C G T C A T T C C T T A A T
T G T C T A G A G A G A A G A G G G G G C C C T C A G A G G A A A C T C A G
G A A A A C T C A T G C C A T T C T C C A T T C A A C G G A G G G C G A C A
T T C T A G C G C T G A T C C C A T T T T C C T C C T C T T C T C G T G G G
A G G C C A T C C C C G G C G A C C T A T A G G A G A T G G G G A A G G C T
C C C C A C T G C C C C T G G T A C C A G C G C G C T C C A G C T T G T C C
T T C C C G T T C T C C A G G T A T C T G C G G A G C C A C T C C A C G C A
C T C G C C C T C C A G G T A G G C T C T C C G C T G C T C C G C C T C A C
G G G C C G C C T C C C A C T T G C G C T G G G T G A T C T G A G C C G C C
G T G T C C G C G G C G G T C C A G G A G C G C A G G T C C T C G T T C A G
G G C G A T G T A A T C C T T G C C G T C G T A G G C G T A C T G G T C A T
G C C C G C G G A G G A G G C G C C C G T C C G G C C C C A C G T C G C A G
C C G T A C A T G C T C T G G A G G G T G T G A G A C C C T G G C C C C G G
C C C C G C G G T C A G C C C A G T C C C C C G A G C C C C G C C C C G C C
C C G A C C A A C C C G C G G G G A T T T T G G C C T C A A C T G A A A A T
G A A A C C G G G T A A A C G C G C C T G G G G C T C T C G C C G G T C G A
G G G T C T G G G C G G G T C C C G C G G C C T C A G G G A G G C G G A T C
T C G G A C C C G G A G A C T C G G G G C G A C C C G G G C C G T A C G T G
G G G G A T G G G G A G T C G T G A C C T G C G C C C C G G G C C G G G G T
C A C T C A C C G G C C T C G C T C T G G T T G T A G T A G C C G C G C A G
G T T C C G C A G G C T C T C T C G G T C A G T C T G T G C C T G G G C C T
T G T A G A T C T G T G T G T T C C G G T C C C A A T A C T C C G G C C C C
T C C T G C T C T A T C C A C G G C G C C C G C G G C T C C T C T C T C G G
A C T C G C G G C G T C G C T G T C G A A C C T C A C G A A C T G G G T G T
C G T C C A C G T A G C C C A C T G A G A T G A A G C G G G G C T C C C C G
C G G C C G G G C C G G G A C A C G G A G G T G T A G A A A T A C C T C A T
G G A G T G G G A G C C T G G G G G T G A G G A G G G G C T G A G A C C C G
C C C G A C C C T C C T C C C G G C G C G G C T C C T C A G G T C C T G C G
C C C C C G C C T G C G G T C C C C T C G C T C C T C C C G G C A G A G G C
C A T T T C C C T C C C G A C C C G C A C T C A C C G G C C C A G G T C T C
G G T C A G G G C C A G G G C C G C C G A G A G C A G C A G G A G G A C G G
T T C G G G G C G C C A T G A C C A G C A T C T C G G C G T C T G A G G A G
A C T C T G A G T C C G G G T G G G T G C G T G G G G A C T T T A G A A C T
G G G A C C G C G G C G A C G C T G A T T G G C T T C T C T A G A T A T C C
A A T A C C C A A T G G G A G T G G G A A G T G G G G A C G C G T C A C G A
G T A T C C T G G A A G A A G G A C C C G A C A C A A G T T G G G A G A A G
A A G T G A A A C T C A G G G G A G T G G G G A A T C C C C A A C G C T G C
G C C T C C C C A T T G C A G A C G C G G C C C T C G G A G C C T G A G A C
C C T G A G A G C C C C G T C C G G G A C C T G G G A C T T C G T C C T G A
T C C C T C T T C T C C T A C A C C A A G C C T C T T T G T C A C A C T G T
```

Illustration 68: HLA-B (Homo sapiens major histocompatibility complex) chr6:31356167-31356442

The CpG methylation is shown as colour coded di-nucleotides. It uses the same colour legend as used in the **gene expression** display, where higher DNA methylation is shown as red-purple and lower methylation is shown as green-blue. When one looks at the end of the gene and 3'UTR region, there are much less methylation:



Another gene HLA-G shown with DNA Methylation around its UTR5' region:

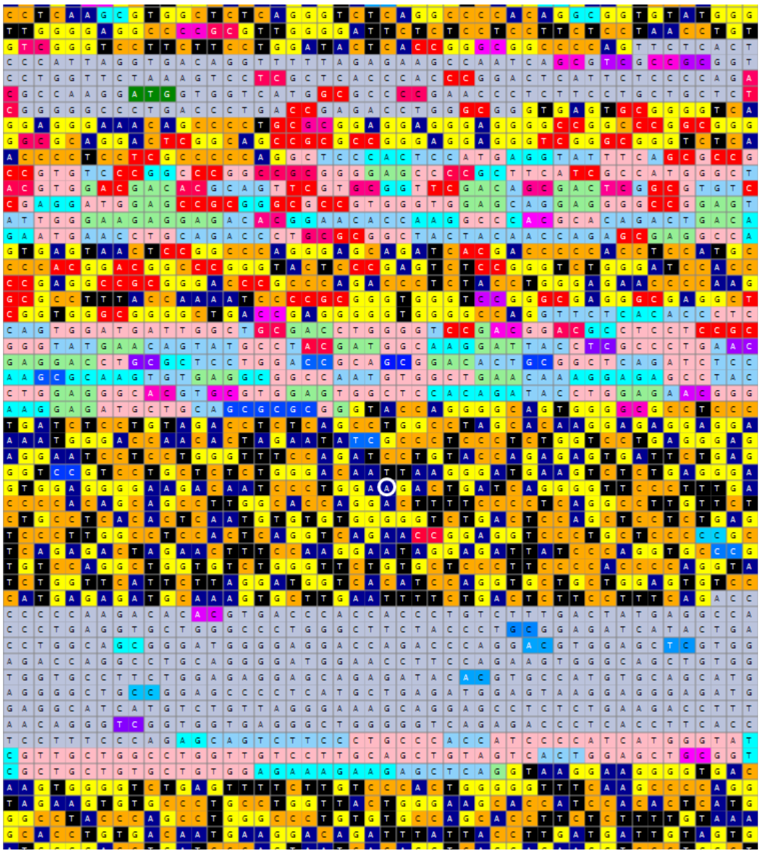
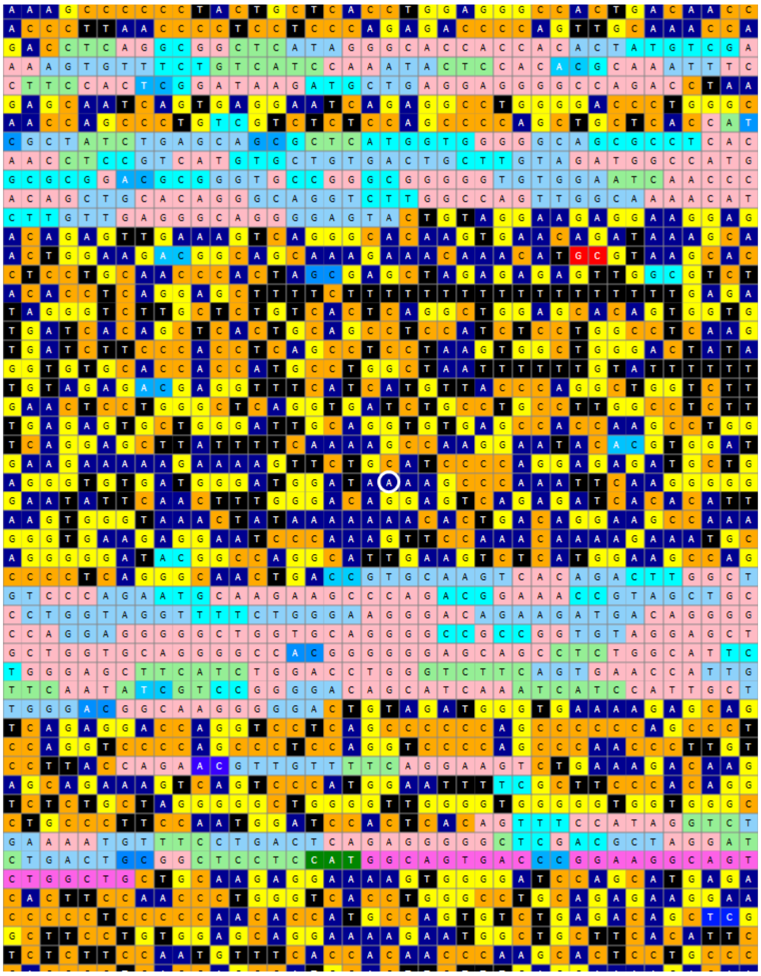


Illustration 69: HLA-G (Homo sapiens major histocompatibility complex) chr6:29826979-29831122

Then again, we know that other genes such as the gene TP53, which codes for the all important "Guardian of The Genome" P53, always need to be expressed in cells, and we are not expecting any DNA methylation. Here is the **DNA View** of this all important gene, encoded on the negative strand of chromosome 17. (There are almost no DNA methylation)



```
A A A G C C C C C C T A C T G C T C A C C T G G A G G G C C A C T G A C A A C C
A C C C T T A A C C C C T C C T C C C A G A G A C C C C A G T T G C A A A C C A
G A C C T C A G G C G G C T C A T A G G G C A C C A C C A C A C T A T G T C G A
A A A G T G T T T C T G T C A T C C A A A T A C T C C A C A C G C A A A T T T C
C T T C C A C T C G G A T A A G A T G C T G A G G A G G G G C C A G A C C T A A
G A G C A A T C A G T G A G G A A T C A G A G G C C T G G G G A C C C T G G G C
A A C C A G C C C T G T C G T C T C T C C A G C C C C A G C T G C T C A C C A T
C G C T A T C T G A G C A G C G C T C A T G G T G G G G G C A G C G C C T C A C
A A C C T C C G T C A T G T G C T G T G A C T G C T T G T A G A T G G C C A T G
G C G C G G A C G C G G G T G C C G G G C G G G G G T G T G G A A T C A A C C C
A C A G C T G C A C A G G G C A G G T C T T G G C C A G T T G G C A A A A C A T
C T T G T T G A G G G C A G G G G A G T A C T G T A G G A A G A G G A A G G A G
A C A G A G T T G A A A G T C A G G G C A C A A G T G A A C A G A T A A A G C A
A C T G G A A G A C G G C A G C A A A G A A A C A A A C A T G C G T A A G C A C
C T C C T G C A A C C C A C T A G C G A G C T A G A G A G A G T T G G C G T C T
A C A C C T C A G G A G C T T T T C T T T T T T T T T T T T T T T T T T G A G A
T A G G G T C T T G C T C T G T C A C T C A G G C T G G A G C A C A G T G G T G
T G A T C A C A G C T C A C T G C A G C C T C C A T C T C C T G G C C T C A A G
T G A T C T T C C C A C C T C A G C C T C C T A A G T G G C T G G G A C T A T A
G G T G T G C A C C A C C A T G C C T G G C T A A T T T T T T G T A T T T T T T
T G T A G A G A C G A G G T T T C A T C A T G T T A C C C A G G C T G G T C T T
G A A C T C C T G G G C T C A G G T G A T C T G C C T G C C T T G G C C T C T T
T G A G A G T G C T G G G A T T G C A G G T G T G A G C C A C C A A G C C T G G
T C A G G A G C T T A T T T T C A A A A G C C A A G G A A T A C A C G T G G A T
G A A G A A A A A G A A A A G T T C T G C A T C C C C A G G A G A G A T G C T G
A G G G T G T G A T G G G A T G G A T A A A A G C C C A A A T T C A A G G G G G
G A A T A T T C A A C T T T G G G A C A G G A G T C A G A G A T C A C A C A T T
A A G T G G G T A A A C T A T A A A A A A A C A C T G A C A G G A A G C C A A A
G G G T G A A G A G G A A T C C C A A A G T T C C A A A C A A A A G A A A T G C
A G G G G G A T A C G G C C A G G C A T T G A A G T C T C A T G G A A G C C A G
C C C C T C A G G G C A A C T G A C C G T G C A A G T C A C A G A C T T G G C T
G T C C C A G A A T G C A A G A A G C C C A G A C G G A A A C C G T A G C T G C
C C T G G T A G G T T T T C T G G G A A G G G A C A G A A G A T G A C A G G G G
C C A G G A G G G G G C T G G T G C A G G G G C C G C C G G T G T A G G A G C T
G C T G G T G C A G G G G C C A C G G G G G G A G C A G C C T C T G G C A T T C
T G G G A G C T T C A T C T G G A C C T G G G T C T T C A G T G A A C C A T T G
T T C A A T A T C G T C C G G G G A C A G C A T C A A A T C A T C C A T T G C T
T G G G A C G G C A A G G G G G A C T G T A G A T G G G T G A A A A G A G C A G
T C A G A G G A C C A G G T C C T C A G C C C C C C A G C C C C C C A G C C C T
C C A G G T C C C C A G C C C T C C A G G T C C C C A G C C C A A C C C T T G T
C C T T A C C A G A A C G T T G T T T T C A G G A A G T C T G A A A G A C A A G
A G C A G A A A G T C A G T C C C A T G G A A T T T T C G C T T C C C A C A G G
T C T C T G C T A G G G G G C T G G G G T T G G G G T G G G G G T G G T G G G C
C T G C C C T T C C A A T G G A T C C A C T C A C A G T T T C C A T A G G T C T
G A A A A T G T T T C C T G A C T C A G A G G G G G C T C G A C G C T A G G A T
C T G A C T G C G G C T C C T C C A T G G C A G T G A C C C G G A A G G C A G T
C T G G C T G C T G C A A G A G G A A A A G T G G G G A T C C A G C A T G A G A
C A C T T C C A A C C C T G G G T C A C C T G G G C C T G C A G A G A A G G A A
C C C C C T C C C C C A A C A C C A T G C C A G T G T C T G A G A C A G C T C G
G C T T C C T G T G G A G C A G G A A A A G A A T G G C T G C T T C A C A T T C
T C T C T T C C A A T G T T T C A C C A C A A C C C A A G C A C T C C T G C C C
```

Illustration 70: TP53 (Homo sapiens tumor suppressor protein p53), transcript variant 3, chr17:7676521-7676622

Interestingly, when we look at the WRAP53 gene, JUST NEXT DOOR of TP53, we observe much higher DNA methylation near the 5'UTR of this gene. Again showing that DNA Methylation (which is mitotically heritable via the DNMT1 methyl transferase) is a regulated process which is applied to specific areas of the genome during embryogenesis and gametogenesis in order to switch specific regions on (depending on cell type) and many regions off (such as transposable elements).

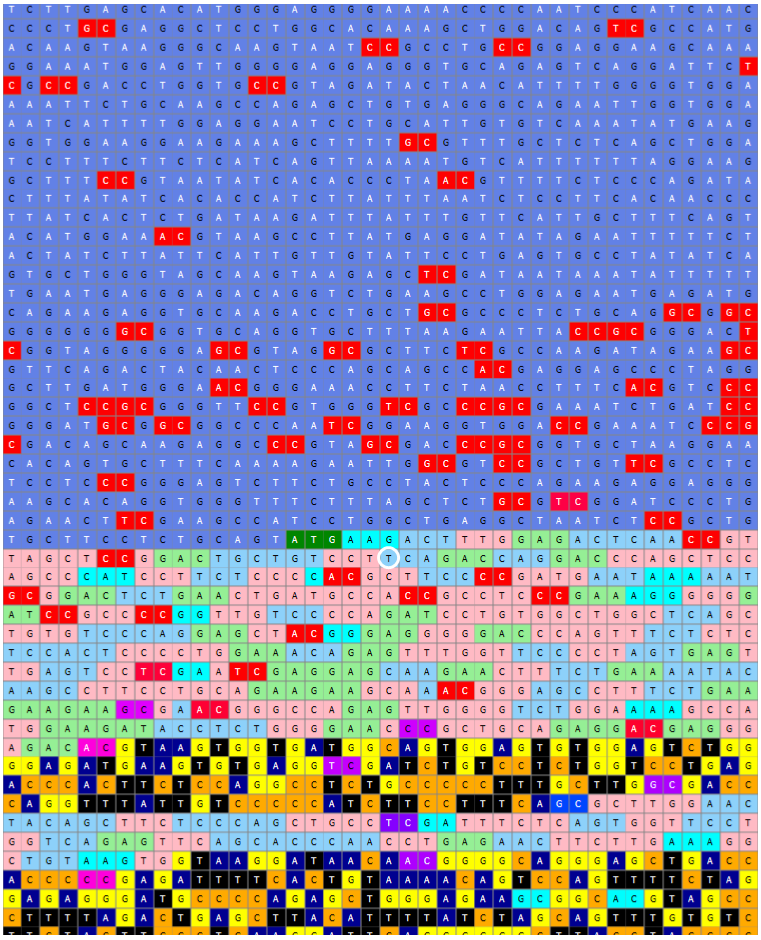


Illustration 71: WRAP53 (Essential component of the telomerase holoenzyme complex) chr17:7686301-7689079

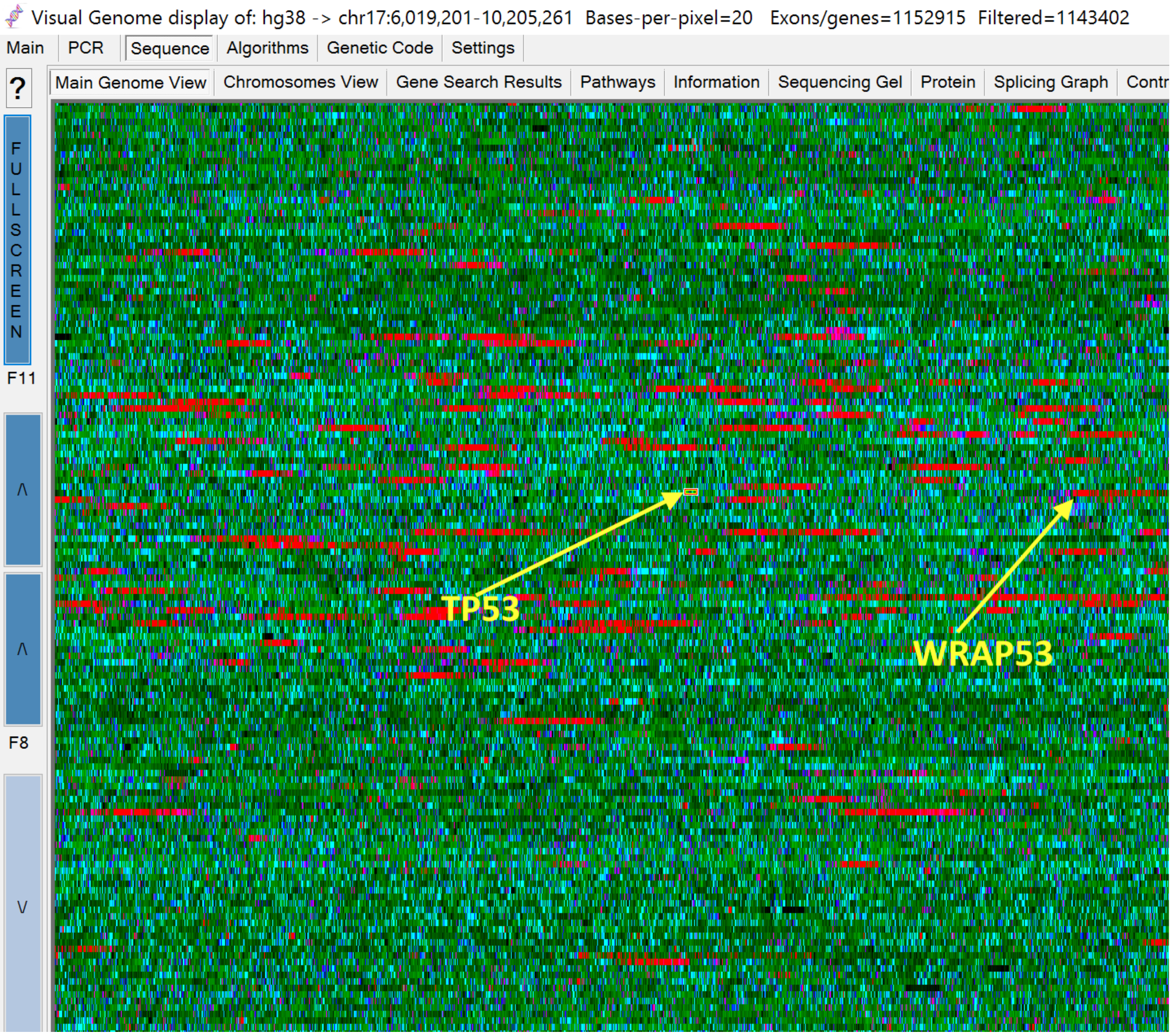
The WRAP53 gene (which is found just next to TP53) needs to be switched off for most cells to prevent them from producing telomerase which will allow the cells to effectively live forever (due to the ability to indefinitely lengthen their telomeres and thereby circumvent the Hayflick limit of cell division).

WRAP53

“Essential component of the telomerase holoenzyme complex, a ribonucleoprotein complex essential for the replication of chromosome termini that elongates telomeres in most eukaryotes. In the telomerase holoenzyme complex, it controls telomerase localization to Cajal body.”

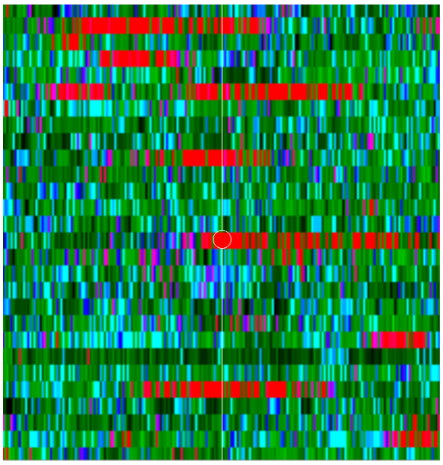
It is also possible to load all the DNA methylation onto the **Main Genome View**, by putting a **.all** in the name of the **.bw** file. (just make sure only gene regions are displayed and not gene labels).

When you move the **DNA View**, a rectangular block is displayed around the region on the **Main Genome View.** When the DNA Methylation is drawn as a layer on top of the GC Content view of the genome, the locations of the **TP53** gene, which is un-methylated and the **WRAP53** gene, which is methylated, is shown in the following image with colour coded DNA methylation:



Visual Genome display of: hg38 -> chr17:6,019,201-10,205,261 Bases-per-pixel=20 Exons/genes=1152915 Filtered=1143402

Illustration 72: DNA Methylation on Chromosome 17



In the centre of the magnified view on the right, the methylated region near the WRAP53 gene can be seen.

Looking at DNA methylation on transposable elements in the human genome

It is also well known that repetitive elements such as transposable elements are “silenced” by DNA methylation to prevent them from jumping to other places in the genome causing genome instability. When I look at some of these elements I find that they are heavily DNA methylated:

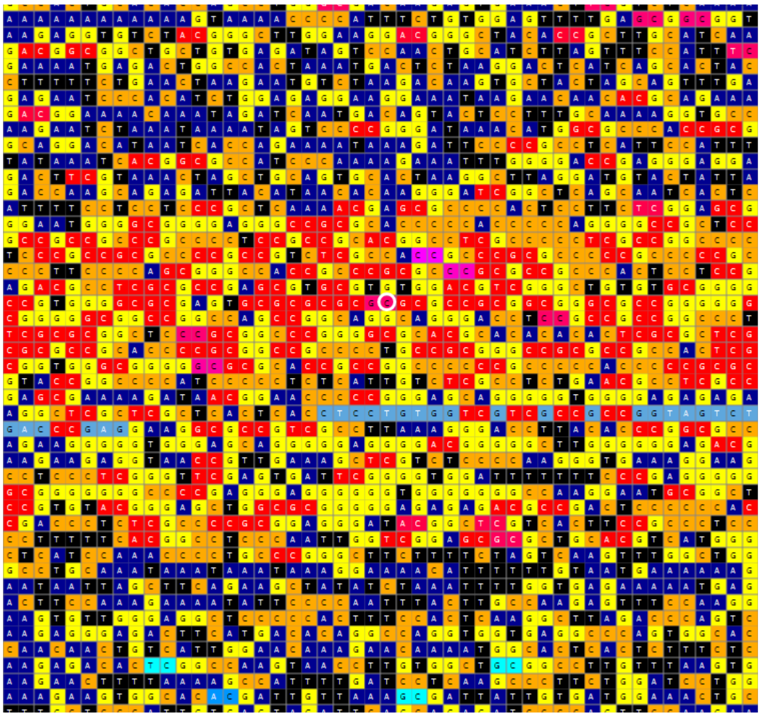


Illustration 73: POGZ (pogo transposable element with ZNF domain) chr1:151440843-151459179

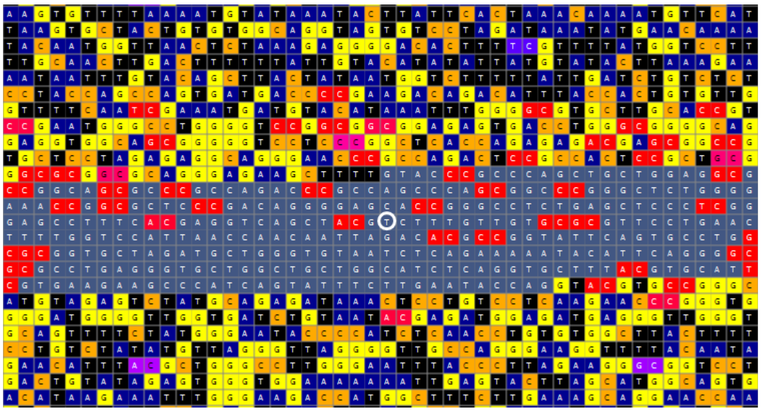


Illustration 74: PGBD1 (Homo sapiens piggyBac transposable element derived 1), chr6:28281572-28281918

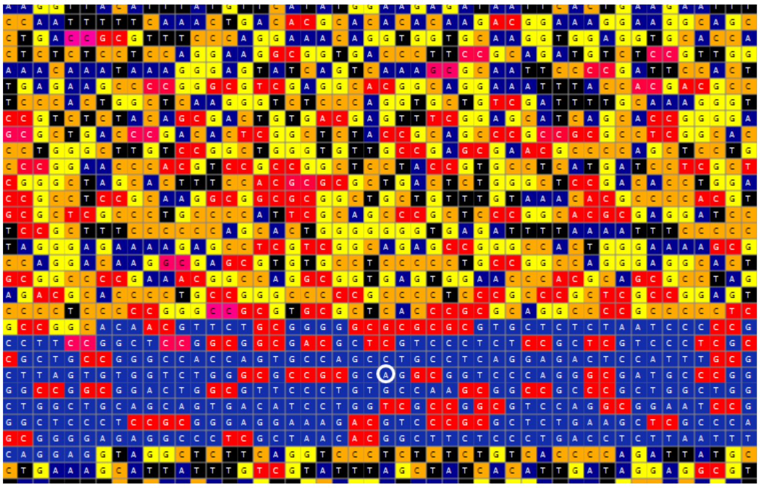


Illustration 75: TIGD2 (tigger transposable element derived 2), chr4:89111500-89111883

Ways the software will enable comparison of DNA differences

The Visual Genome Browser software has the following features which will assist in visualising differences in DNA sequences:

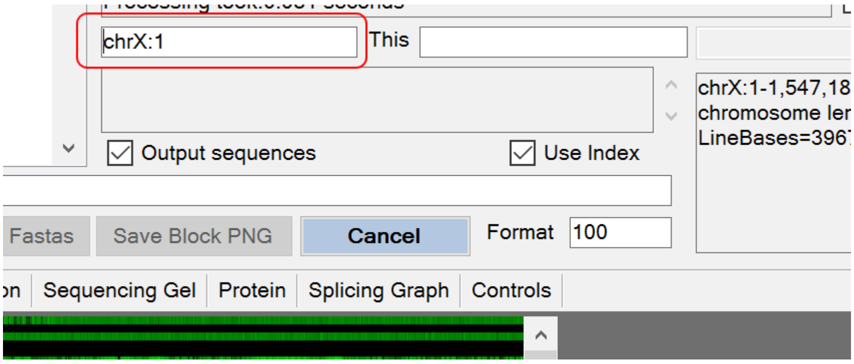
- Direct overlay of 2 DNA sequences directly on top of each other (similar to how transparencies of DNA Southern Blots can be overlaid on top of each other in order to spot differences) as depicted in the following picture.

- Using global and linear sequence alignment methods of DNA and amino acid sequences copied via the context menu and other views
- Using the "DNA-probe" method for searching for snippets of DNA throughout the genome sequence or in die displayed **DNA View**.
- Using an brute force search through all protein coding genes and comparing them using the Blosum/Dayhoff substitution matrices.

Comparison by overlaying DNA sequences from two different chromosomes or even genomes on top of each other

Knowing that all of the autosomal chromosomes have homologous regions which have to pair up in order to provide the "molecular pulling force" necessary for the cell cycle to progress into telophase, I was looking for a way to determine how big the para-autosomal regions are which is shared between the X and the Y chromosomes. I decided that I would overlay the X and Y chromosomes on top of each other: **Everywhere the nucleotide bases matched exactly, I would display the bases in the DNA View in their correct colour, BUT, everywhere there is a mismatch, I would replace the base colour with magenta.** This gave me a simple way to find out up to exactly where the para-autosomal region stretched. Humans have 2 copies of the genes in this region, one from the X and one from the Y chromosome.

The way this is accomplished in the software is to simply display the X chromosome sequence and then to **double-click** in the text field as indicated in the following picture:



chrX:1 This

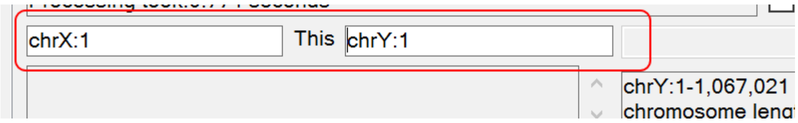
chrX:1-1,547,18
chromosome ler
LineBases=396

☑ Output sequences ☑ Use Index

Format 100

Alternatively you can enter the position of the first chromosome to be compared as: **chrX:1**

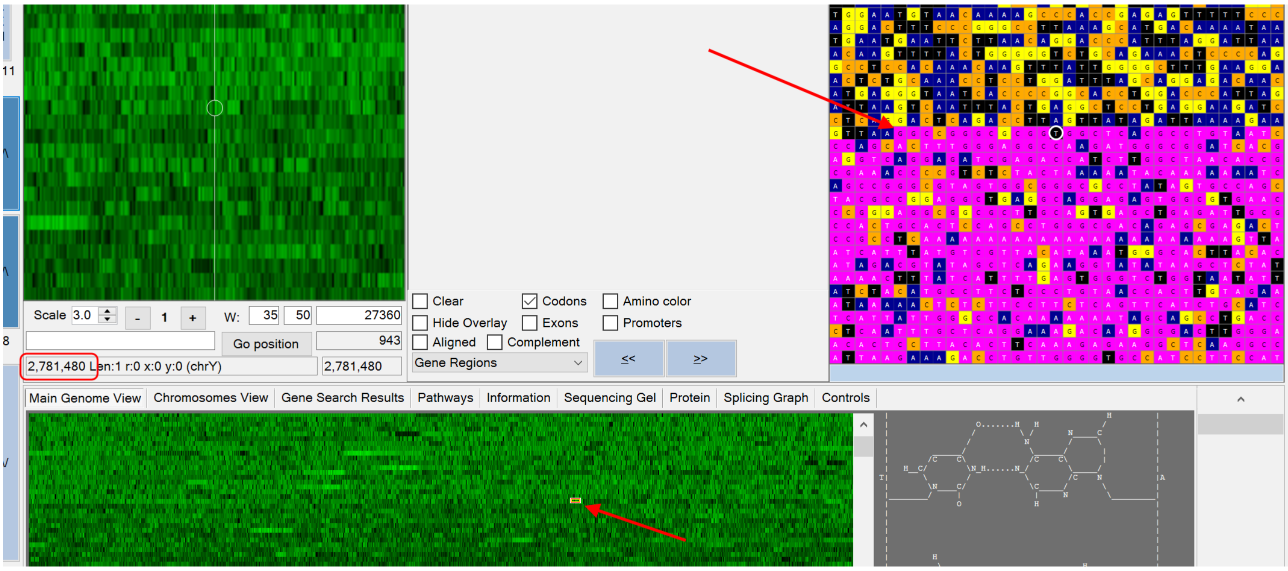
Then, you load the second chromosome (which may come from another genome) and simply double click in the adjacent box.



chrX:1 This chrY:1

chrY:1-1,067,021
chromosome leng

As soon as both fields have a position in them, the comparison will take place automatically in the **DNA View** window:



Scale 3.0 - 1 + W: 35 50 27360

943

2,781,480 Len:1 r:0 x:0 y:0 (chrY) 2,781,480

□ Clear ☑ Codons □ Amino color
□ Hide Overlay □ Exons □ Promoters
□ Aligned □ Complement

Gene Regions

Illustration 76: Picture showing the position at exactly chrY:2781480 where the para-autosomal region ends.

As you move the mouse over the **Main Genome View** on the bottom left, the **DNA View** will update to show where the differences lie. You can slow down the mouse movement by holding the **ALT** key while moving the mouse. Or, you can use the keyboard arrow keys in order to move the mouse in the genome. When you hold the **Shift** key, the keyboard movement will speed up.

Following this method I was able to accurately find the position where the X and Y chromosomes contain exactly the same bases as position **chrY:2 781 480**.

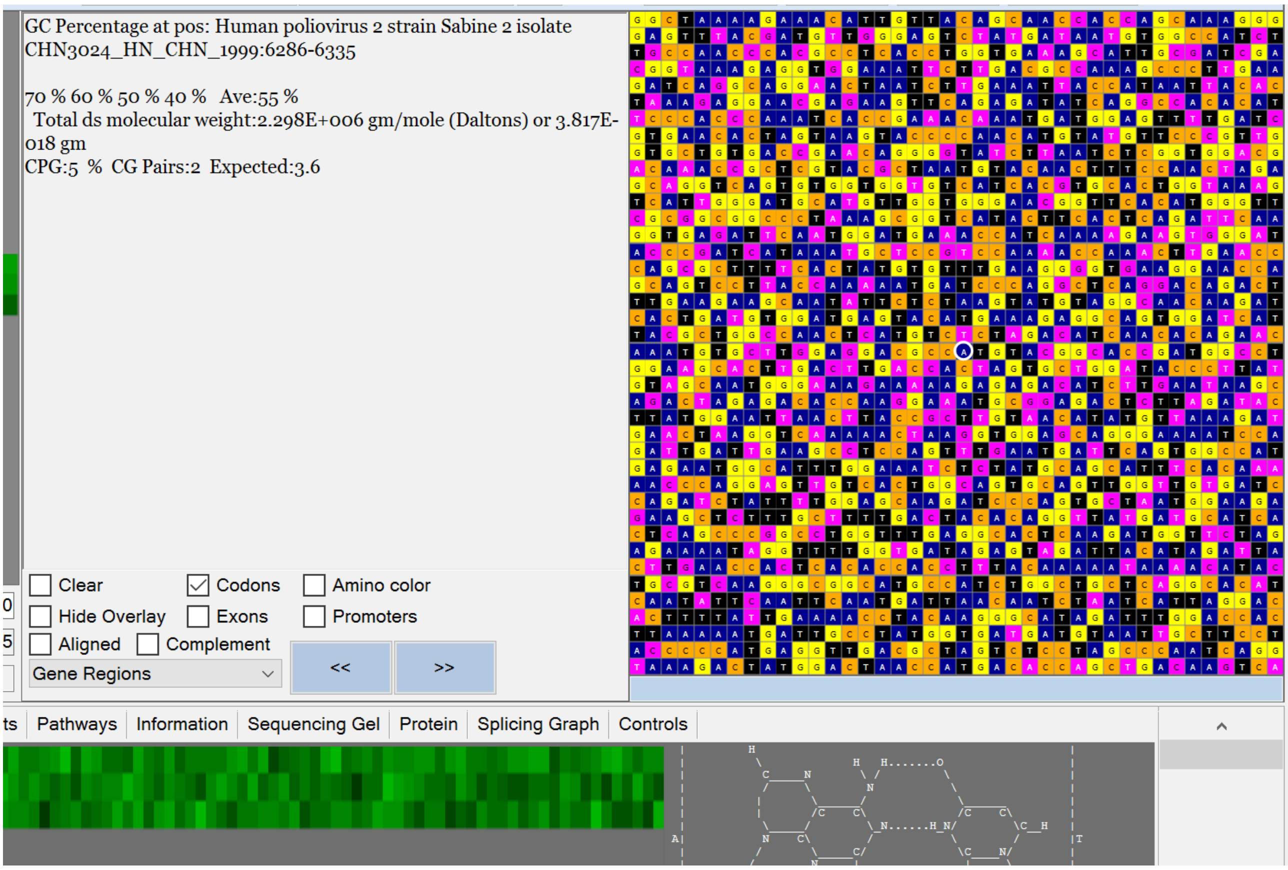
This means that all the DNA bases up to this point is EXACTLY the same up to the very letter.

Comparing Polio virus strains

Here is an example where I have compared the following 2 strains of the Polio Virus:

Human poliovirus 2 isolate CHN16019c_Sichuan_CHN_2012 complete

Human poliovirus 2 strain Sabine 2 isolate CHN3024_HN_CHN_1999



GC Percentage at pos: Human poliovirus 2 strain Sabine 2 isolate CHN3024_HN_CHN_1999:6286-6335

70 % 60 % 50 % 40 % Ave:55 %
 Total ds molecular weight:2.298E+006 gm/mole (Daltons) or 3.817E-018 gm
CPG:5 % CG Pairs:2 Expected:3.6

☐ Clear ☑ Codons ☐ Amino color
☐ Hide Overlay ☐ Exons ☐ Promoters
☐ Aligned ☐ Complement
Gene Regions

One can also use this technique to overlap the mitochondrial genomes of different individuals. Here I have compared "Homo sapiens isolate UV1145 mitochondrion" with the HG38 reference mitochondrial genome. By holding the **Ctrl key** while moving the mouse one can actually SLIDE the first sequence over the second one and see how they overlap.

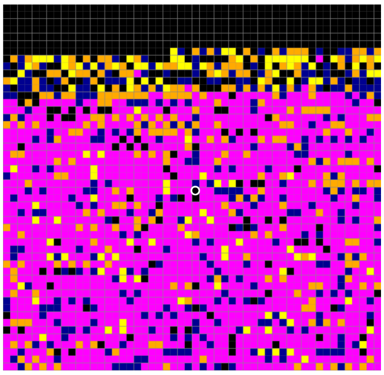


Illustration 79: With the first sequence at base 1

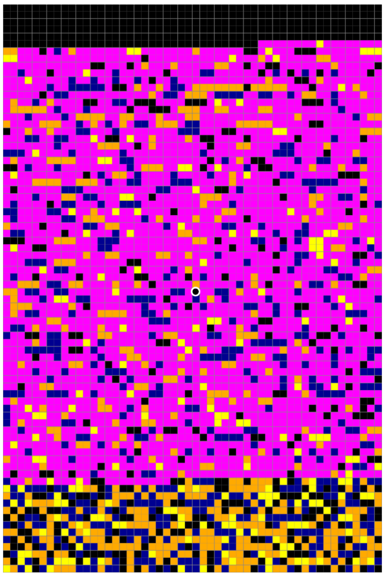


Illustration 77: With the first sequence at base 2

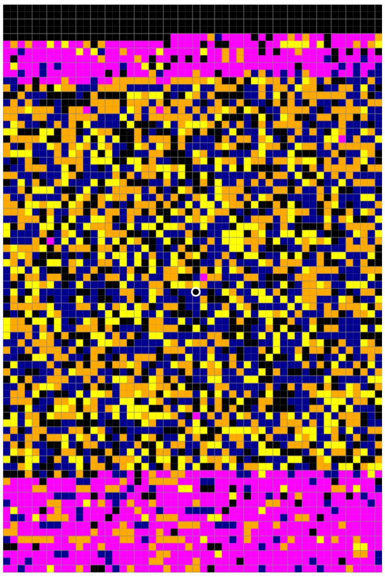
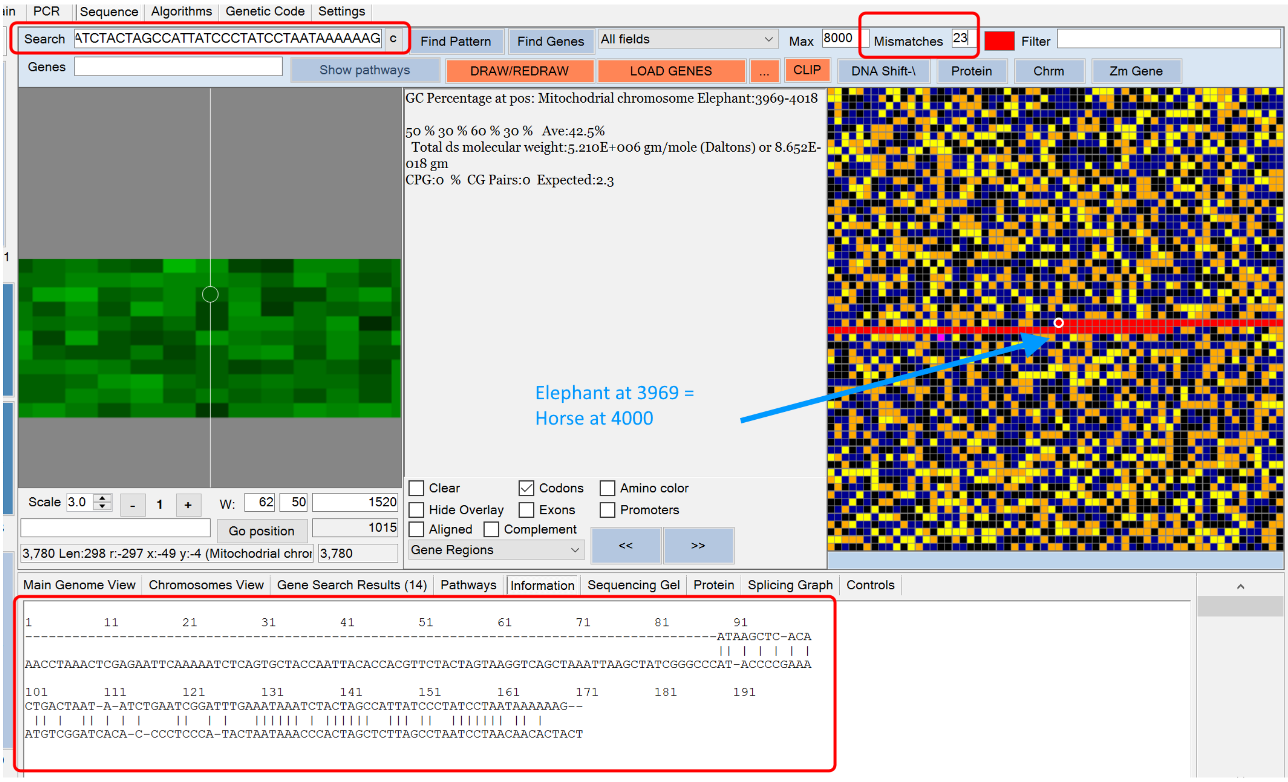


Illustration 78: With the first sequence at base 3

Say I wanted to determine at which base of its mitochondrial DNA the Horse and the Elephant aligns.

- First I load the mitochondrial genome of the Horse (educab). I then arbitrarily select a sequence in the centre of the genome (say 4000) and I copy a piece of its DNA by double clicking at the start and end of the section I want to copy: I copy 78 bases and paste it in the **Search Field** as indicated: ATAAGCTCACACTGACTAATAATCTGAATCGGATTTGAAATAAATCTACTAGCCAT TATCCCTATCCTAATAAAAAAG
- From now on this becomes my DNA probe.
- I then load the Elelphant's mitochondrial DNA and look in the **DNA View to see if I see any "hybridizations"/matches found. I proceed to increase the number of allowable mismatches until I get a match. eg.23 as indicated**.
- When I now move the mouse across the **DNA View**, I am given a local sequence alignment in the "**Information** tab" as highlighted in the picture. I conclude that there is an alignment between Horse:4000 and Elephant:3969 (In other words: Elephant:1 and Horse:31)
- While having the Elephant sequence open I double click and set Elephant:1
  Using This technique I was able to visually find a region of alignment.



Search ATCTACTAGCCATTATCCCTATCCTAATAAAAAAG c

All fields | Max 8000 | Mismatches 23 | Filter

Genes

GC Percentage at pos: Mitochodrial chromosome Elephant:3969-4018

50 % 30 % 60 % 30 % Ave:42.5%
 Total ds molecular weight:5.210E+006 gm/mole (Daltons) or 8.652E-018 gm
CPG:0 % CG Pairs:0 Expected:2.3

Elephant at 3969 =
Horse at 4000

Scale 3.0 | - 1 + | W: 62 50 | 1520 | 1015

3,780 Len:298 r:-297 x:-49 y:-4 (Mitochodrial chro | 3,780

☐ Clear ☑ Codons ☐ Amino color
☐ Hide Overlay ☐ Exons ☐ Promoters
☐ Aligned ☐ Complement
Gene Regions

```
1         11        21        31        41        51        61        71        81        91
----------------------------------------------------------------------------------------ATAAGCTC-ACA
                                                                                        || | | | | |
AACCTAAACTCGAGAATTCAAAAATCTCAGTGCTACCAATTACACCACGTTCTACTAGTAAGGTCAGCTAAATTAAGCTATCGGGCCCAT-ACCCCGAAA

101       111       121       131       141       151       161       171       181       191
CTGACTAAT-A-ATCTGAATCGGATTTGAAATAAATCTACTAGCCATTATCCCTATCCTAATAAAAAAG--
 || |  || | | |    ||  | |   |||||| | ||||||  ||| ||  ||||||| || |
ATGTCGGATCACA-C-CCCTCCCA-TACTAATAAACCCACTAGCTCTTAGCCTAATCCTAACAACACTACT
```

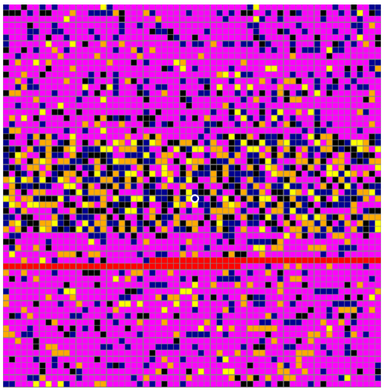


Illustration 80: With Elelphant:1 aligned with Horse:31

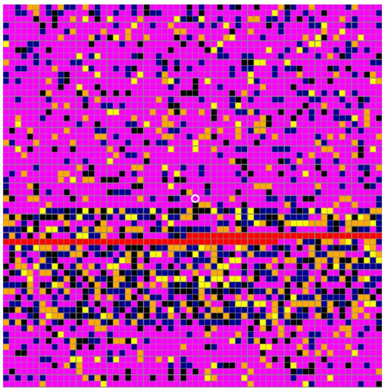


Illustration 81: With Elelphant:0 aligned with Horse:31

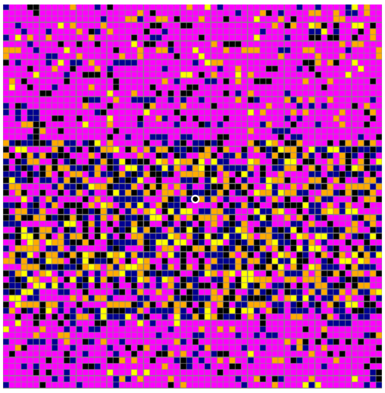


Illustration 82: With Elephan 14 aligned with Horse:31

Using the software to look at VCF (Variant Calling Format) files

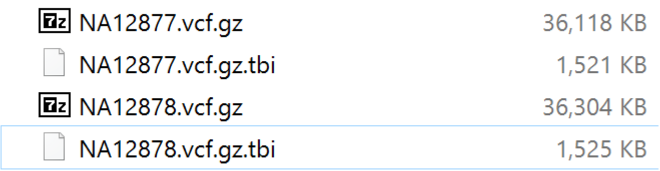
When you have your genome sequenced, the DNA data is generally not given as a full DNA sequence, but rather as a VCF file representing all the differences between your genome and the reference genome. The raw sequence data consists of FASTQ files containing short string sequences called **reads** from all over your genome. These short **reads** are then aligned with the reference genome resulting in a file called a **BAM** file. This file contains the sequence alignment. This is however not the final output. The data is then fed through an algorithm which produces a file containing a list of all the individual bases where there is a difference between each of your genome's 46 homologous chromosomes (23 which you inherited from your father and 23 inherited from your mother). This means there could possibly be 2 variant records at each base of the reference genome (one from your mother and one from your father). Consecutive changes are however grouped together into a single variant record of the VCF file.

This means that, given your genome's VCF file, it is possible to reconstruct the sequence each of your 46 chromosomes by reading the nucleotide sequence of the reference genome and then replacing the letters with that from the VCF file record. Each chromosome in the reference genome will then produce 2 different versions for each of your homologous pairs. (Homologous means that the 2 chromosomes has mostly corresponding bases but differs approximately every 1000 bases with a different base (single nucleotide polymorphism/SNP) or with a deletion or insertion (INDEL). This means across the entire genome you have about 6 million mostly harmless base changes (3 million from your mother and 3 million from your father), which makes you the unique person which you are.

Reading compressed VCF files

The software is able to read **compressed VCF files** into the **overlay** display of the **DNA View. VCF files** may sometimes contain the genome variants of more than one individual. This means there will be 2 sequences for each of these individuals in the list of layers to choose from.

Each **compressed VCF** will consist of 2 files each:



NA12877.vcf.gz	36,118 KB
NA12877.vcf.gz.tbi	1,521 KB
NA12878.vcf.gz	36,304 KB
NA12878.vcf.gz.tbi	1,525 KB

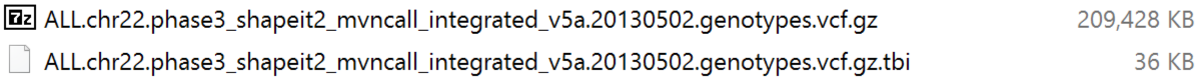
Illustration 83: 2 Different VCF's for 2 different individuals NA12877 and NA12878

The **.gz.tbi** file represents the **Tabix index file** which contains an index between the genome positions and the blocks of compressed data in the **BGZIP file** with the extension **.gz**

The **BGZIP** file in turn consists of many broken up blocks of VCF records individually compressed using the **gzip** compression algorithm. This allows software to quickly jump to a specific block of VCF data and decompress it into tabular record data.

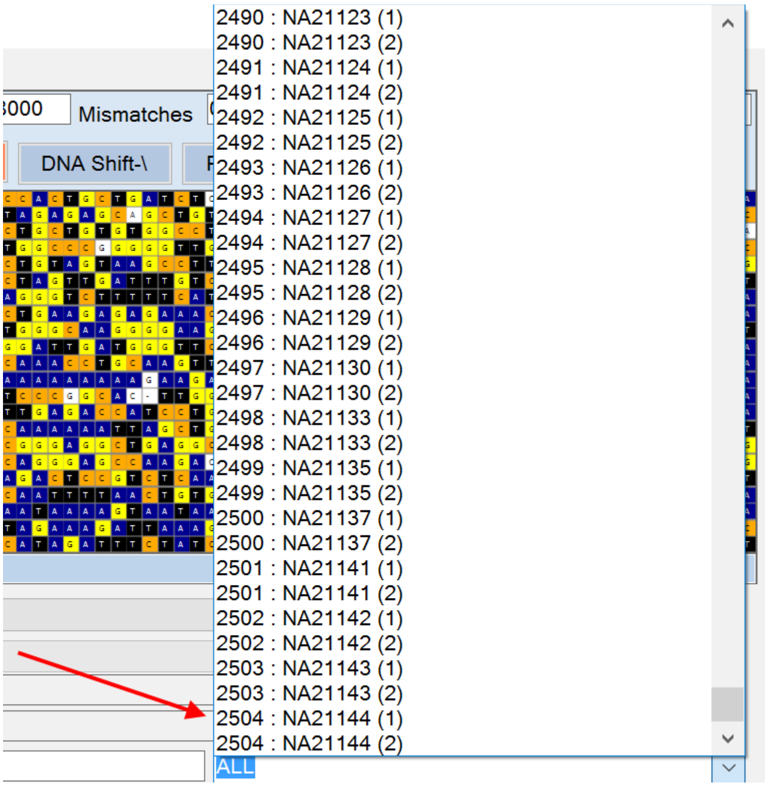
It is also possible to get a VCF file which contains records for many individuals such as the following one I downloaded from the **1000 Genomes Project**.

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/



ALL.chr22.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz	209,428 KB
ALL.chr22.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi	36 KB

This VCF contains the genome variants for 2504 different individuals (each identified by unique NA_____ codes and consisting of 2 entries for each person representing the chromosomes inherited from father and mother)

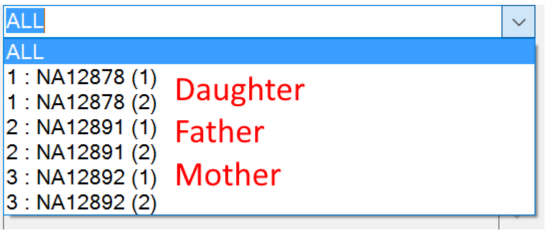


2490 : NA21123 (1)
2490 : NA21123 (2)
2491 : NA21124 (1)
2491 : NA21124 (2)
2492 : NA21125 (1)
2492 : NA21125 (2)
2493 : NA21126 (1)
2493 : NA21126 (2)
2494 : NA21127 (1)
2494 : NA21127 (2)
2495 : NA21128 (1)
2495 : NA21128 (2)
2496 : NA21129 (1)
2496 : NA21129 (2)
2497 : NA21130 (1)
2497 : NA21130 (2)
2498 : NA21133 (1)
2498 : NA21133 (2)
2499 : NA21135 (1)
2499 : NA21135 (2)
2500 : NA21137 (1)
2500 : NA21137 (2)
2501 : NA21141 (1)
2501 : NA21141 (2)
2502 : NA21142 (1)
2502 : NA21142 (2)
2503 : NA21143 (1)
2503 : NA21143 (2)
2504 : NA21144 (1)
2504 : NA21144 (2)
ALL

Illustration 84: Subject's chromosome selection in the software)

At other times, as in the VCF below, the variants of 3 or more individuals are given with a family relationship (sometimes called **trios** because it consist of a father, mother and child's genomes):

ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20140625_high_coverage_trios_broad



ALL
ALL
1 : NA12878 (1)
1 : NA12878 (2)
Daughter
2 : NA12891 (1)
2 : NA12891 (2)
Father
3 : NA12892 (1)
3 : NA12892 (2)
Mother

Displaying the VCF in the Software

After selecting the correct release of the reference genome eg. HG38, and then displaying the desired chromosome by clicking “**DRAW/REDRAW**”. If you also want to load the genes overlay, click “**LOAD GENES**”. Then browse to the gene you want to look at. For example, you can type **PRR35** in the **Gene Entry field** and then quickly jump to a gene at chr16:563256-564376.

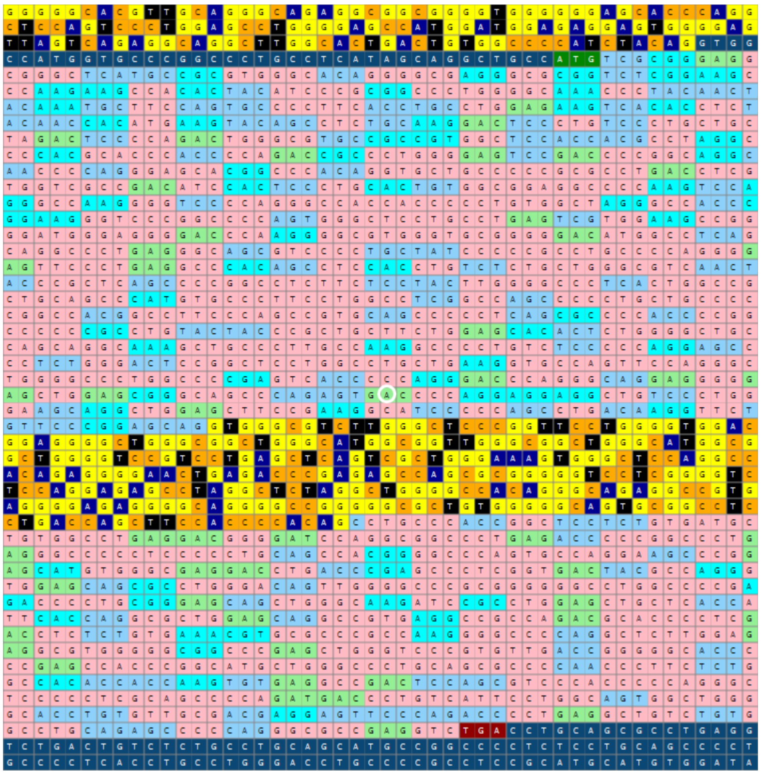
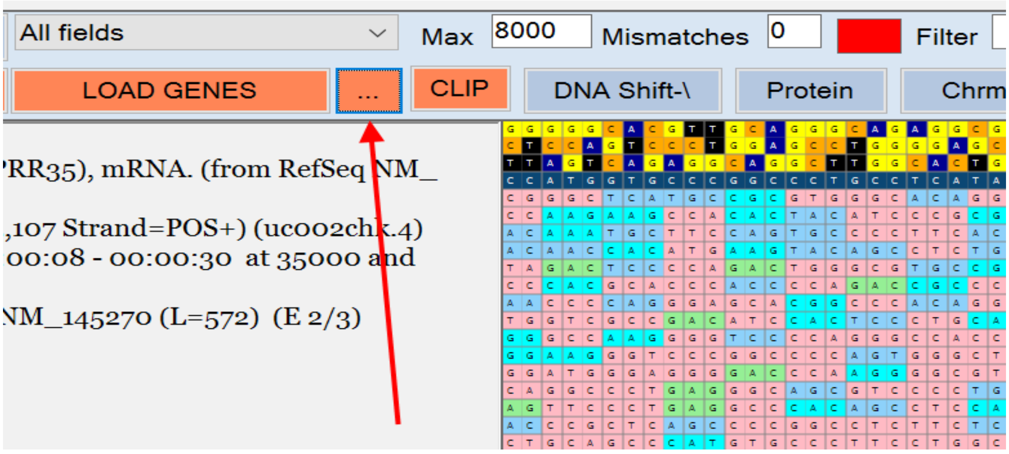


Illustration 85: Homo sapiens proline rich 35 (PRR35), mRNA. (from RefSeq NM_145270)

Now it is time to load the VCF. Click on the "**Browse**" button and select **.vcf.gz.tbi** file extension.



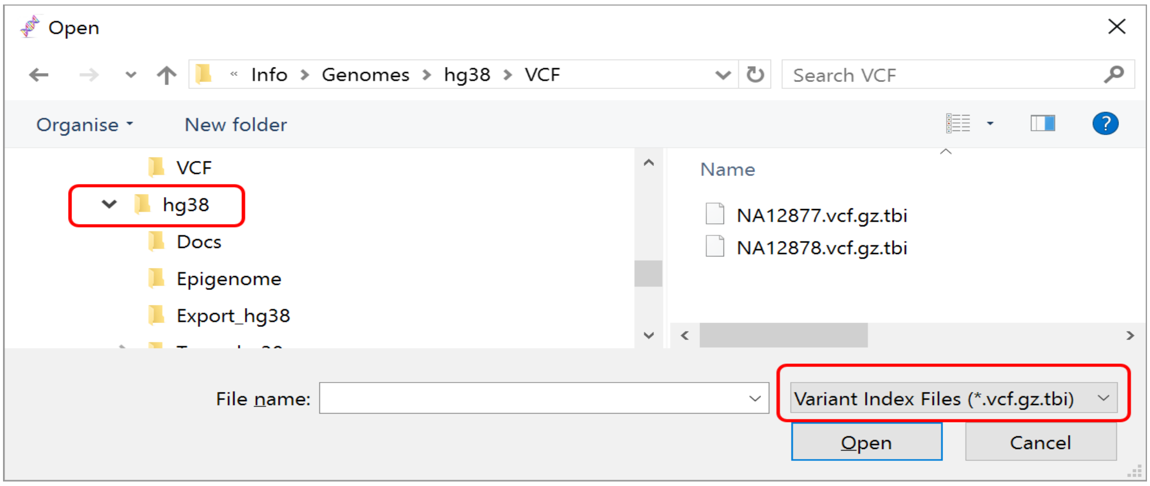
Max 8000

Mismatches 0

'RR35), mRNA. (from RefSeq NM_

,107 Strand=POS+) (uc002chk.4)
00:08 - 00:00:30 at 35000 and

NM_145270 (L=572) (E 2/3)



NA12877.vcf.gz.tbi

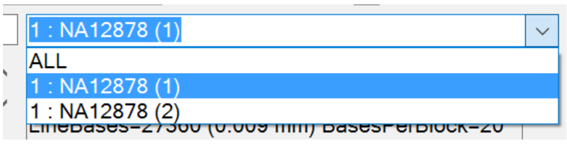
NA12878.vcf.gz.tbi

Variant Index Files (*.vcf.gz.tbi)

When displaying VCF data it is always VERY IMPORTANT to make sure that you are displaying the VCF variants using the correct matching reference genome (eg. HG19 or HG38 etc.) which was used to generate the VCF file in the first place. The VCF data contains a snippet of bases from the reference genome which can be used to verify if it is from the correct genome.

Now select the person's genome you want to look at eg. NA12878 (which is the daughter in the trio)

We can now see the variants indicated for the first allele of this specific gene. You can select a different chromosome in the drop down combo box in order to display different alleles (and subjects if there are more than one in the file).



1 : NA12878 (1)

ALL
1 : NA12878 (1)
1 : NA12878 (2)
LineBases=27300 (0.009 mm) BasesPerBlock=20

The SNPs (single nucleotide polymorphisms) are indicated in the image below.

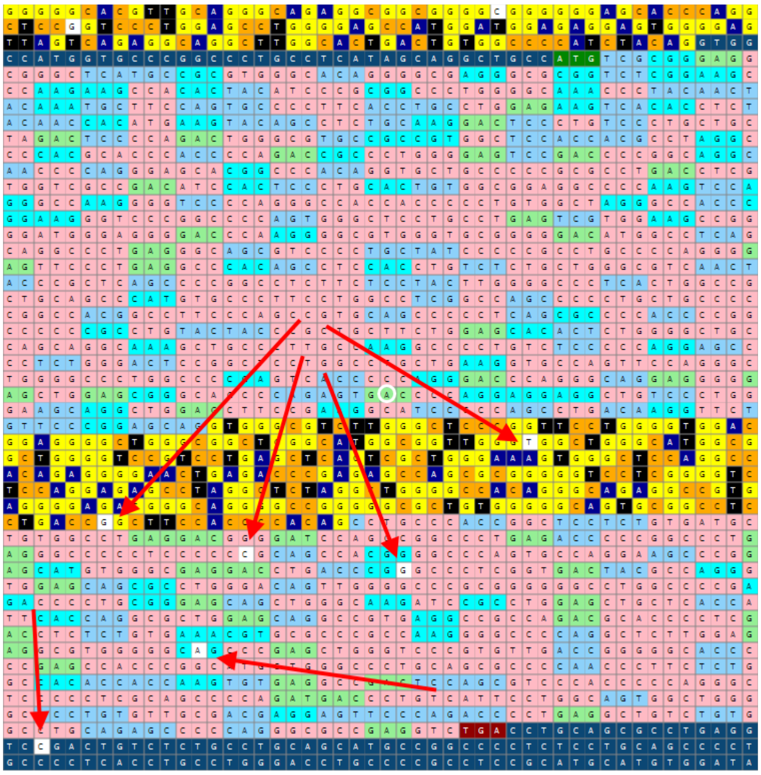
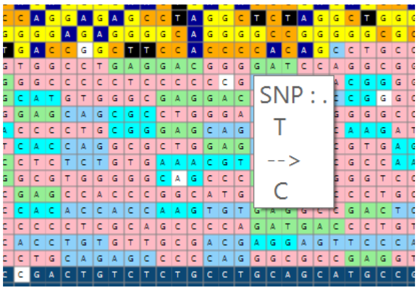


Illustration 86: Variants indicated on gene for Homo sapiens proline rich 35 (PRR35), mRNA. chr16:563256-564376

You can quickly switch between alleles by pressing the **< and >** buttons when the mouse if over the **DNA View**.

When you move the mouse over the indicated SNPs, you can also get more information on what kind of variant SNP/INDEL it is.



SNP : .
T
-->
C

There are now new options available in the context menu related to the VCF variants which allows you to both copy the changed coding sequence and the changed protein sequence to the clipboard.

There is also an option which will draw the changed protein sequence in the **Protein View**.

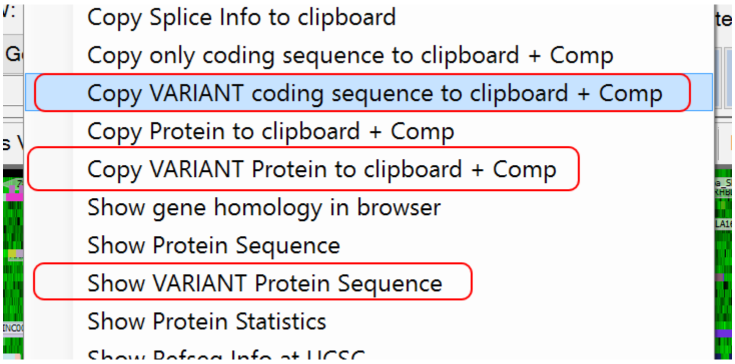
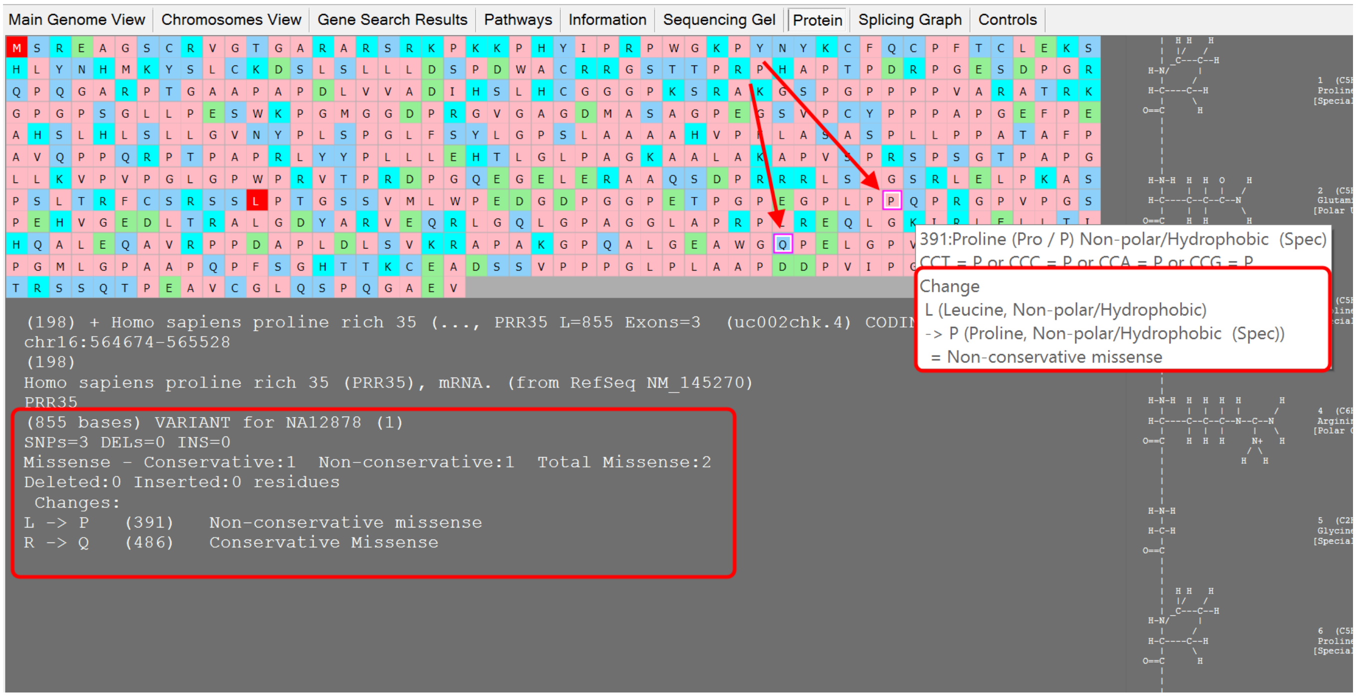

Copy Splice Info to clipboard
Copy only coding sequence to clipboard + Comp
Copy VARIANT coding sequence to clipboard + Comp
Copy Protein to clipboard + Comp
Copy VARIANT Protein to clipboard + Comp
Show gene homology in browser
Show Protein Sequence
Show VARIANT Protein Sequence
Show Protein Statistics
Show Refseq Info at UCSC

Illustration 87: New variant menu options

The clipboard options allow you to copy the DNA or Amino acid sequences to the clipboard or the Comparison list, allowing you to do sequence alignment between the sequences.



```
M S R E A G S C R V G T G A R A R S R K P K K P H Y I P R P W G K P Y N Y K C F Q C P F T C L E K S
H L Y N H M K Y S L C K D S L S L L L D S P D W A C R R G S T T P R P H A P T P D R P G E S D P G R
Q P Q G A R P T G A A P A P D L V V A D I H S L H C G G G P K S R A K G S P G P P P P V A R A T R K
G P G P S G L L P E S W K P G M G G D P R G V G A G D M A S A G P E G S V P C Y P P P A P G E F P E
A H S L H L S L L G V N Y P L S P G L F S Y L G P S L A A A A H V P F L A S A S P L L P P A T A F P
A V Q P P Q R P T P A P R L Y Y P L L L E H T L G L P A G K A A L A K A P V S P R S P S G T P A P G
L L K V P V P G L G P W P R V T P R D P G Q E G E L E R A A Q S D P R R R L S [illegible] G S R L E L P K A S
P S L T R F C S R S S L P T G S S V M L W P E D G D P G G P E T P G P E G P L P P Q P R G P V P G S
P E H V G E D L T R A L G D Y A R V E Q R L G Q L G P A G G L A P R P L R E Q L G K [illegible]
H Q A L E Q A V R P P D A P L D L S V K R A P A K G P Q A L G E A W G Q P E L G P V [illegible]
P G M L G P A A P Q P F S G H T T K C E A D S S V P P P G L P L A A P D D P V I P G [illegible]
T R S S Q T P E A V C G L Q S P Q G A E V
```

391:Proline (Pro / P) Non-polar/Hydrophobic (Spec)
CCT = P or CCC = P or CCA = P or CCG = P

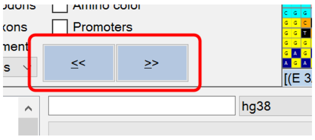
Change
L (Leucine, Non-polar/Hydrophobic)
-> P (Proline, Non-polar/Hydrophobic (Spec))
= Non-conservative missense

```
 (198) + Homo sapiens proline rich 35 (..., PRR35 L=855 Exons=3  (uc002chk.4) CODIN
chr16:564674-565528
 (198)
Homo sapiens proline rich 35 (PRR35), mRNA. (from RefSeq NM_145270)
PRR35
 (855 bases) VARIANT for NA12878 (1)
SNPs=3 DELs=0 INS=0
Missense - Conservative:1  Non-conservative:1  Total Missense:2
Deleted:0 Inserted:0 residues
 Changes:
L -> P   (391)   Non-conservative missense
R -> Q   (486)   Conservative Missense
```

You can also press the **< and >** buttons to switch between different alleles of this gene while displaying the protein view. This allows you to compare the effect of the different allele's changes on the protein sequence. **Amino acid changes are marked with small magenta rectangles** and when you move the mouse over these changes, the tooltip will display the kind of change such as Conservative/Non-conservative missense or amino acid Insertions. (Deletions are not displayed because they are not available in the changed sequence).

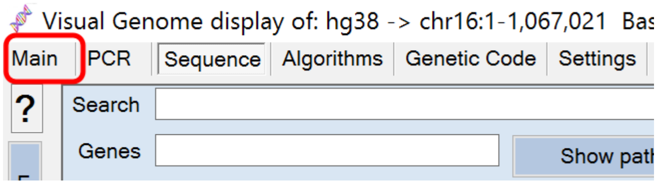
Also, because of the linkage between the views, the **DNA View** will centre on the appropriate codon.

While looking at the protein view you can also shift through the displayed genes by pressing the **Reverse and Forward** buttons:

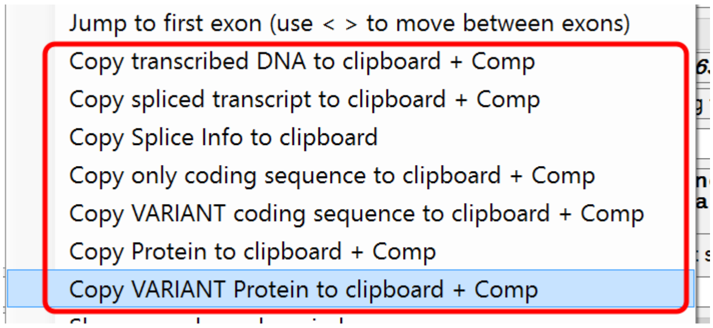


When you want to do sequence alignment you can use the Comparison List

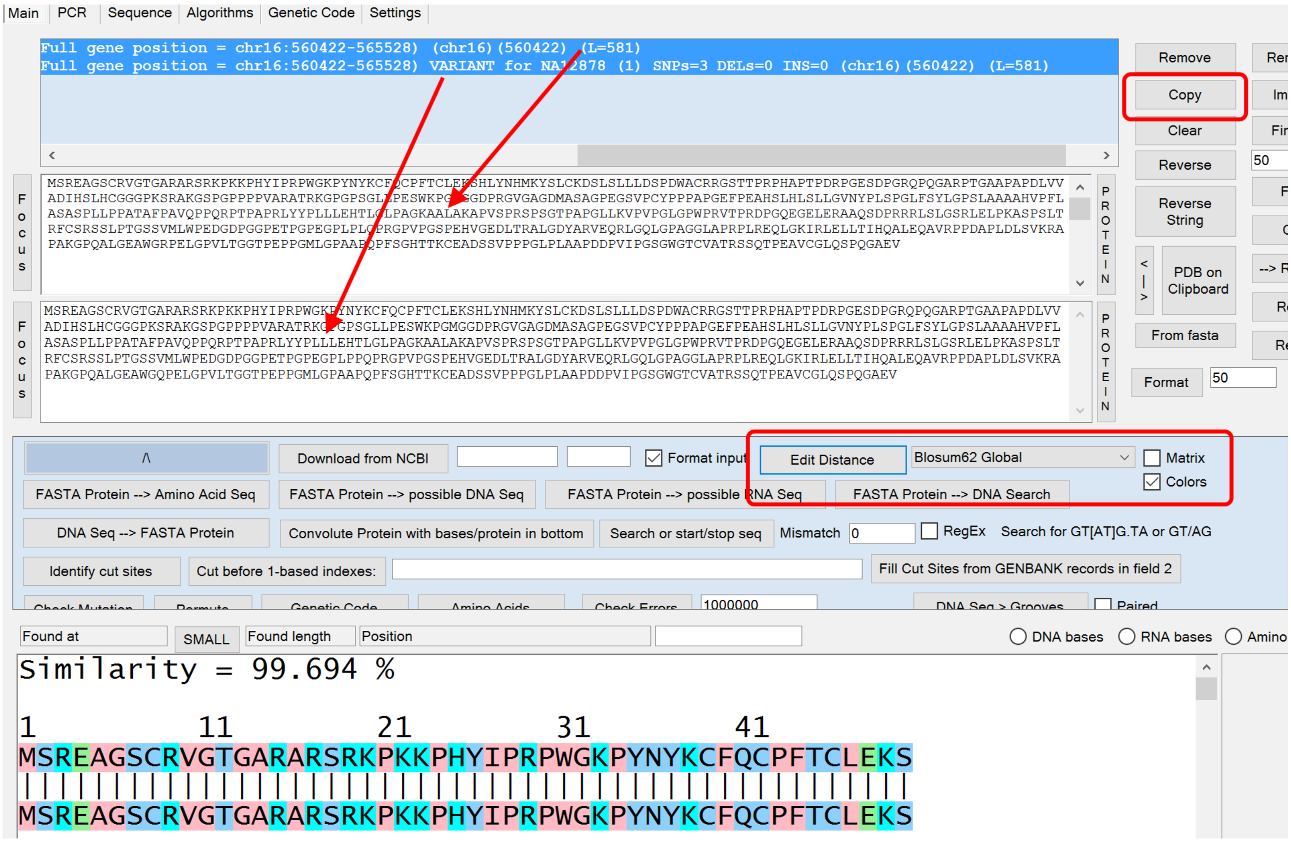
Every time you copy a sequence to the clipboard, it also gets put into the **Comparison List**. This list can be found on the **Main Tab**.



All of the following menu options will copy sequences into the comparison view:


Jump to first exon (use < > to move between exons)
Copy transcribed DNA to clipboard + Comp
Copy spliced transcript to clipboard + Comp
Copy Splice Info to clipboard
Copy only coding sequence to clipboard + Comp
Copy VARIANT coding sequence to clipboard + Comp
Copy Protein to clipboard + Comp
Copy VARIANT Protein to clipboard + Comp

Now click on the “**Copy**” button after selecting the desired entries you want to compare:



```
Full gene position = chr16:560422-565528) (chr16)(560422) (L=581)
Full gene position = chr16:560422-565528) VARIANT for NA12878 (1) SNPs=3 DELs=0 INS=0 (chr16)(560422) (L=581)
```

```
MSREAGSCRVGTGARARSRKPKKPHYIPRPWGKPYNYKCFQCPFTCLEKSHLYNHMKYSLCKDSLSLLLDSPDWACRRGSTTPRPHAPTPDRPGESDPGRQPQGARPTGAAPAPDLVV
ADIHSLHCGGGPKSRAKGSPGPPPPVARATRKGPGPSGLLPESWKPGMGGDPRGVGAGDMASAGPEGSVPCYPPPAPGEFPEAHSLHLSLLGVNYPLSPGLFSYLGPSLAAAAHVPFL
ASASPLLPPATAFPAVQPPQRPTPAPRLYYPLLLEHTLGLPAGKAALAKAPVSPRSPSGTPAPGLLKVPVPGLGPWPRVTPRDPGQEGELERAAQSDPRRRLSLGSRLELPKASPSLT
RFCSRSSLPTGSSVMLWPEDGDPGGPETPGPEGPLPLQPRGPVPGSPEHVGEDLTRALGDYARVEQRLGQLGPAGGLAPRPLREQLGKIRLELLTIHQALEQAVRPPDAPLDLSVKRA
PAKGPQALGEAWGRPELGPVLTGGTPEPPGMLGPAAPQPFSGHTTKCEADSSVPPPGLPLAAPDDPVIPGSGWGTCVATRSSQTPEAVCGLQSPQGAEV
```

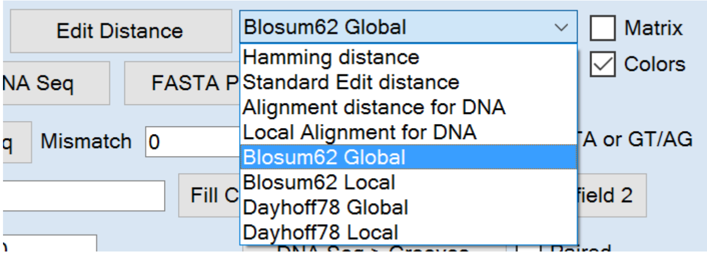
```
MSREAGSCRVGTGARARSRKPKKPHYIPRPWGKPYNYKCFQCPFTCLEKSHLYNHMKYSLCKDSLSLLLDSPDWACRRGSTTPRPHAPTPDRPGESDPGRQPQGARPTGAAPAPDLVV
ADIHSLHCGGGPKSRAKGSPGPPPPVARATRKGPGPSGLLPESWKPGMGGDPRGVGAGDMASAGPEGSVPCYPPPAPGEFPEAHSLHLSLLGVNYPLSPGLFSYLGPSLAAAAHVPFL
ASASPLLPPATAFPAVQPPQRPTPAPRLYYPLLLEHTLGLPAGKAALAKAPVSPRSPSGTPAPGLLKVPVPGLGPWPRVTPRDPGQEGELERAAQSDPRRRLSLGSRLELPKASPSLT
RFCSRSSLPTGSSVMLWPEDGDPGGPETPGPEGPLPPQPRGPVPGSPEHVGEDLTRALGDYARVEQRLGQLGPAGGLAPRPLREQLGKIRLELLTIHQALEQAVRPPDAPLDLSVKRA
PAKGPQALGEAWGQPELGPVLTGGTPEPPGMLGPAAPQPFSGHTTKCEADSSVPPPGLPLAAPDDPVIPGSGWGTCVATRSSQTPEAVCGLQSPQGAEV
```

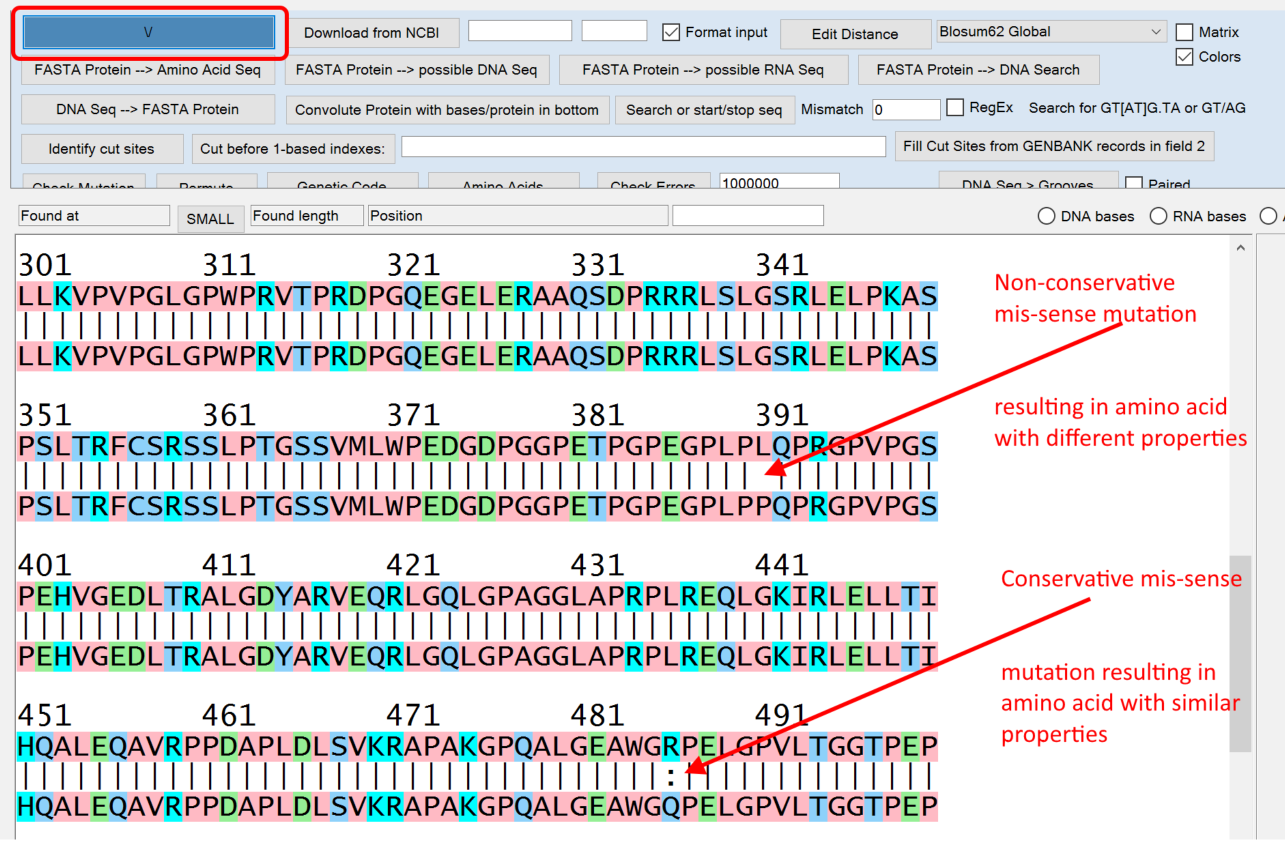
```
Similarity = 99.694 %

1        11        21        31        41
MSREAGSCRVGTGARARSRKPKKPHYIPRPWGKPYNYKCFQCPFTCLEKS
||||||||||||||||||||||||||||||||||||||||||||||||||
MSREAGSCRVGTGARARSRKPKKPHYIPRPWGKPYNYKCFQCPFTCLEKS
```

This will copy the sequences to compare into the fields below it. Now select the correct protein alignment method, eg. Blosum62, check the colors check box and then click on “**Edit Distance**”.

This will do a sequence alignment using the selected comparison method and amino acid substitution matrix:

The B**IG / SMALL button** can be used to select between a large and a small display of the alignment letters. The **Colors check box** will colour code the letters based on the same Amino acid polarity scheme that is used elsewhere in the application. Notice how the Blosum62 will often display a colon (:) (partial match for a conservative mis-sense mutation, where the resulting amino acid has similar properties as the original.




V
Download from NCBI
☑ Format input
Edit Distance
Blosum62 Global
□ Matrix
☑ Colors
FASTA Protein --> Amino Acid Seq
FASTA Protein --> possible DNA Seq
FASTA Protein --> possible RNA Seq
FASTA Protein --> DNA Search
DNA Seq --> FASTA Protein
Convolute Protein with bases/protein in bottom
Search or start/stop seq
Mismatch 0
□ RegEx
Search for GT[AT]G.TA or GT/AG
Identify cut sites
Cut before 1-based indexes:
Fill Cut Sites from GENBANK records in field 2
Check Mutation
Permute
Genetic Code
Amino Acids
Check Errors
1000000
DNA Seq > Grooves
□ Paired
Found at
SMALL
Found length
Position
○ DNA bases
○ RNA bases

```
301                311                321                331                341
LLKVPVPGLGPWPRVTPRDPGQEGELERAAQSDPRRRLSLGSRLELPKAS
||||||||||||||||||||||||||||||||||||||||||||||||||
LLKVPVPGLGPWPRVTPRDPGQEGELERAAQSDPRRRLSLGSRLELPKAS

351                361                371                381                391
PSLTRFCSRSSLPTGSSVMLWPEDGDPGGPETPGPEGPLPLQPRGPVPGS
|||||||||||||||||||||||||||||||||||||||| |||||||||
PSLTRFCSRSSLPTGSSVMLWPEDGDPGGPETPGPEGPLPPQPRGPVPGS

401                411                421                431                441
PEHVGEDLTRALGDYARVEQRLGQLGPAGGLAPRPLREQLGKIRLELLTI
||||||||||||||||||||||||||||||||||||||||||||||||||
PEHVGEDLTRALGDYARVEQRLGQLGPAGGLAPRPLREQLGKIRLELLTI

451                461                471                481                491
HQALEQAVRPPDAPLDLSVKRAPAKGPQALGEAWGRPELGPVLTGGTPEP
|||||||||||||||||||||||||||||||||||:||||||||||||||
HQALEQAVRPPDAPLDLSVKRAPAKGPQALGEAWGQPELGPVLTGGTPEP
```

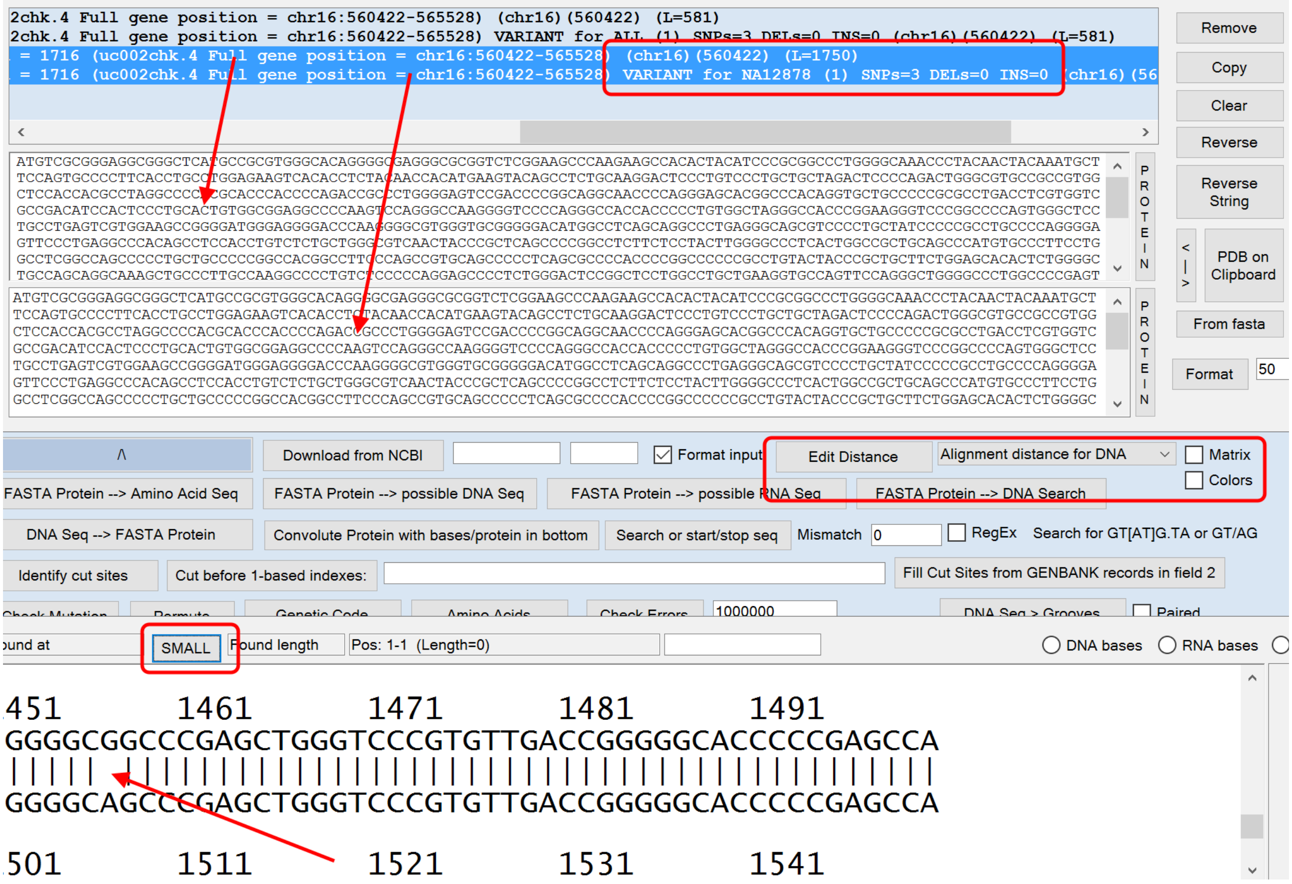
Non-conservative mis-sense mutation

resulting in amino acid with different properties

Conservative mis-sense

mutation resulting in amino acid with similar properties

Illustration 88: Global sequence alignment between reference sequence protein and NA12878 variant for gene PRR35



```
2chk.4 Full gene position = chr16:560422-565528) (chr16)(560422) (L=581)
2chk.4 Full gene position = chr16:560422-565528) VARIANT for ALL (1) SNPs=3 DELs=0 INS=0 (chr16)(560422) (L=581)
 = 1716 (uc002chk.4 Full gene position = chr16:560422-565528) (chr16)(560422) (L=1750)
 = 1716 (uc002chk.4 Full gene position = chr16:560422-565528) VARIANT for NA12878 (1) SNPs=3 DELs=0 INS=0 (chr16)(56
```

```
ATGTCGCGGGAGGCGGGCTCATGCCGCGTGGGCACAGGGGCGAGGGCGCGGTCTCGGAAGCCCAAGAAGCCACACTACATCCCGCGGCCCTGGGGCAAACCCTACAACTACAAATGCT
TCCAGTGCCCCTTCACCTGCCTGGAGAAGTCACACCTCTACAACCACATGAAGTACAGCCTCTGCAAGGACTCCCTGTCCCTGCTGCTAGACTCCCCAGACTGGGCGTGCCGCCGTGG
CTCCACCACGCCTAGGCCCCACGCACCCACCCCAGACCGCCCTGGGGAGTCCGACCCCGGCAGGCAACCCCAGGGAGCACGGCCCACAGGTGCTGCCCCCGCGCCTGACCTCGTGGTC
GCCGACATCCACTCCCTGCACTGTGGCGGAGGCCCCAAGTCCAGGGCCAAGGGGTCCCCAGGGCCACCACCCCCTGTGGCTAGGGCCACCCGGAAGGGTCCCGGCCCCAGTGGGCTCC
TGCCTGAGTCGTGGAAGCCGGGGATGGGAGGGGACCCAAGGGGCGTGGGTGCGGGGGACATGGCCTCAGCAGGCCCTGAGGGCAGCGTCCCCTGCTATCCCCCGCCTGCCCCAGGGGA
GTTCCCTGAGGCCCACAGCCTCCACCTGTCTCTGCTGGGCGTCAACTACCCGCTCAGCCCCGGCCTCTTCTCCTACTTGGGGCCCTCACTGGCCGCTGCAGCCCATGTGCCCTTCCTG
GCCTCGGCCAGCCCCCTGCTGCCCCCGGCCACGGCCTTCCCAGCCGTGCAGCCCCCTCAGCGCCCCACCCCGGCCCCCCGCCTGTACTACCCGCTGCTTCTGGAGCACACTCTGGGGC
TGCCAGCAGGCAAAGCTGCCCTTGCCAAGGCCCCTGTCTTCCCCCAGGAGCCCCTCTGGGACTCCGGCTCCTGGCCTGCTGAAGGTGCCAGTTCCAGGGCTGGGGCCCTGGCCCCGAGT
```

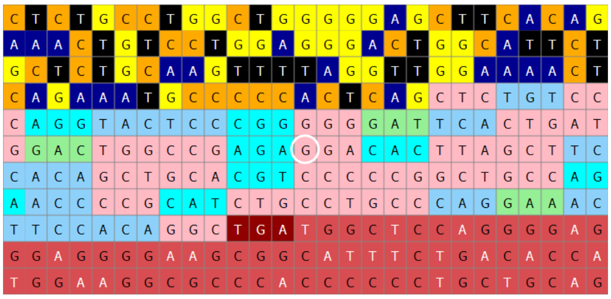
```
ATGTCGCGGGAGGCGGGCTCATGCCGCGTGGGCACAGGGGCGAGGGCGCGGTCTCGGAAGCCCAAGAAGCCACACTACATCCCGCGGCCCTGGGGCAAACCCTACAACTACAAATGCT
TCCAGTGCCCCTTCACCTGCCTGGAGAAGTCACACCTCTACAACCACATGAAGTACAGCCTCTGCAAGGACTCCCTGTCCCTGCTGCTAGACTCCCCAGACTGGGCGTGCCGCCGTGG
CTCCACCACGCCTAGGCCCCACGCACCCACCCCAGACCGCCCTGGGGAGTCCGACCCCGGCAGGCAACCCCAGGGAGCACGGCCCACAGGTGCTGCCCCCGCGCCTGACCTCGTGGTC
GCCGACATCCACTCCCTGCACTGTGGCGGAGGCCCCAAGTCCAGGGCCAAGGGGTCCCCAGGGCCACCACCCCCTGTGGCTAGGGCCACCCGGAAGGGTCCCGGCCCCAGTGGGCTCC
TGCCTGAGTCGTGGAAGCCGGGGATGGGAGGGGACCCAAGGGGCGTGGGTGCGGGGGACATGGCCTCAGCAGGCCCTGAGGGCAGCGTCCCCTGCTATCCCCCGCCTGCCCCAGGGGA
GTTCCCTGAGGCCCACAGCCTCCACCTGTCTCTGCTGGGCGTCAACTACCCGCTCAGCCCCGGCCTCTTCTCCTACTTGGGGCCCTCACTGGCCGCTGCAGCCCATGTGCCCTTCCTG
GCCTCGGCCAGCCCCCTGCTGCCCCCGGCCACGGCCTTCCCAGCCGTGCAGCCCCCTCAGCGCCCCACCCCGGCCCCCCGCCTGTACTACCCGCTGCTTCTGGAGCACACTCTGGGGC
```

```
451        1461       1471       1481       1491
GGGGCGGCCCGAGCTGGGTCCCGTGTTGACCGGGGGCACCCCCGAGCCA
|||||  |||||||||||||||||||||||||||||||||||||||||||
GGGGCAGCCCGAGCTGGGTCCCGTGTTGACCGGGGGCACCCCCGAGCCA

501        1511       1521       1531       1541
```

Illustration 89: The same comparison can be done on the DNA Coding sequence level

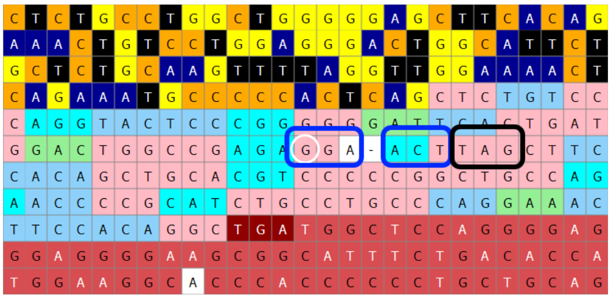
Sometimes variants causes truncated proteins such as with the RAB40C (UCSC Id = uc059olp.1) chr16:625899-626083 .

When I look at the NA12878 (2), in other words, the second homologous chromosome of Chr16, I found a truncated protein caused by a deletion of a nucleotide base:



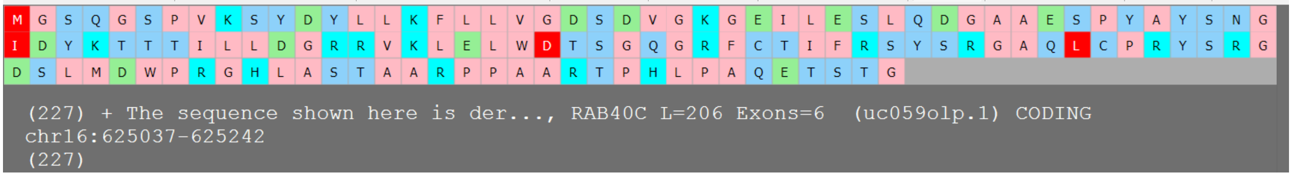
C	T	C	T	G	C	C	T	G	G	C	T	G	G	G	G	G	A	G	C	T	T	C	A	C	A	G
A	A	A	C	T	G	T	C	C	T	G	G	A	G	G	G	A	C	T	G	G	C	A	T	T	C	T
G	C	T	C	T	G	C	A	A	G	T	T	T	T	A	G	G	T	T	G	G	A	A	A	A	C	T
C	A	G	A	A	A	T	G	C	C	C	C	C	A	C	T	C	A	G	C	T	C	T	G	T	C	C
C	A	G	G	T	A	C	T	C	C	C	G	G	G	G	G	G	A	T	T	C	A	C	T	G	A	T
G	G	A	C	T	G	G	C	C	G	A	G	A	G	G	A	C	A	C	T	T	A	G	C	T	T	C
C	A	C	A	G	C	T	G	C	A	C	G	T	C	C	C	C	C	G	G	C	T	G	C	C	A	G
A	A	C	C	C	C	G	C	A	T	C	T	G	C	C	T	G	C	C	C	A	G	G	A	A	A	C
T	T	C	C	A	C	A	G	G	C	T	G	A	T	G	G	C	T	C	C	A	G	G	G	G	A	G
G	G	A	G	G	G	G	A	A	G	C	G	G	C	A	T	T	T	C	T	G	A	C	A	C	C	A
T	G	G	A	A	G	G	C	G	C	C	C	A	C	C	C	C	C	C	T	G	C	T	G	C	A	G

Illustration 90: HG38 Reference sequence bases : chr16:624937-625233



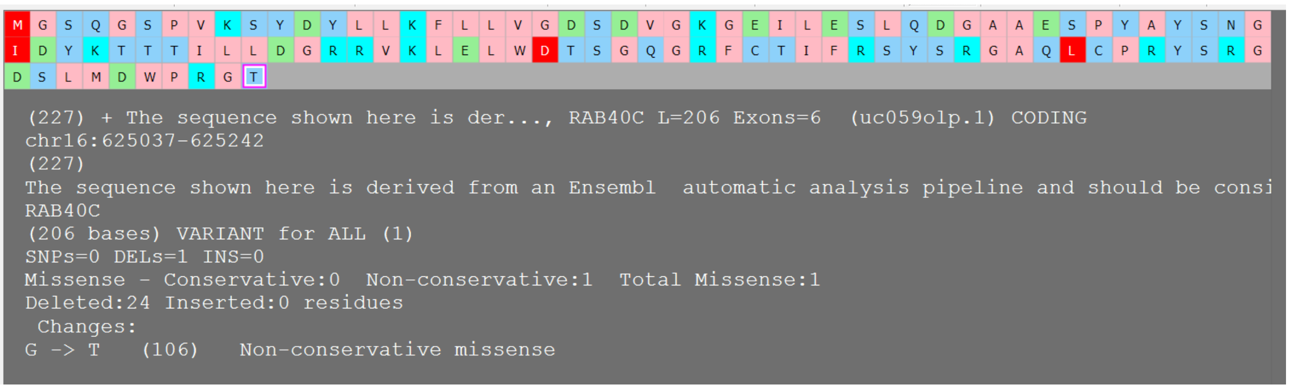
C	T	C	T	G	C	C	T	G	G	C	T	G	G	G	G	G	A	G	C	T	T	C	A	C	A	G
A	A	A	C	T	G	T	C	C	T	G	G	A	G	G	G	A	C	T	G	G	C	A	T	T	C	T
G	C	T	C	T	G	C	A	A	G	T	T	T	T	A	G	G	T	T	G	G	A	A	A	A	C	T
C	A	G	A	A	A	T	G	C	C	C	C	C	A	C	T	C	A	G	C	T	C	T	G	T	C	C
C	A	G	G	T	A	C	T	C	C	C	G	G	G	G	G	G	A	T	T	C	A	C	T	G	A	T
G	G	A	C	T	G	G	C	C	G	A	G	A	G	G	A	-	A	C	T	T	A	G	C	T	T	C
C	A	C	A	G	C	T	G	C	A	C	G	T	C	C	C	C	C	G	G	C	T	G	C	C	A	G
A	A	C	C	C	C	G	C	A	T	C	T	G	C	C	T	G	C	C	C	A	G	G	A	A	A	C
T	T	C	C	A	C	A	G	G	C	T	G	A	T	G	G	C	T	C	C	A	G	G	G	G	A	G
G	G	A	G	G	G	G	A	A	G	C	G	G	C	A	T	T	T	C	T	G	A	C	A	C	C	A
T	G	G	A	A	G	G	C	A	C	C	C	A	C	C	C	C	C	C	T	G	C	T	G	C	A	G

Illustration 91: NA12878 (2) with deletion at chr16:625088. Premature stop codon



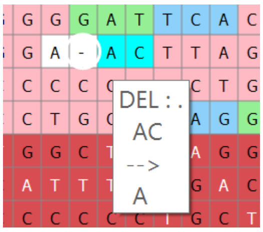
```
M G S Q G S P V K S Y D Y L L K F L L V G D S D V G K G E I L E S L Q D G A A E S P Y A Y S N G
I D Y K T T T I L L D G R R V K L E L W D T S G Q G R F C T I F R S Y S R G A Q L C P R Y S R G
D S L M D W P R G H L A S T A A R P P A A R T P H L P A Q E T S T G

 (227) + The sequence shown here is der..., RAB40C L=206 Exons=6  (uc059olp.1) CODING
chr16:625037-625242
 (227)
```

Illustration 92: Reference sequence gene protein product



```
M G S Q G S P V K S Y D Y L L K F L L V G D S D V G K G E I L E S L Q D G A A E S P Y A Y S N G
I D Y K T T T I L L D G R R V K L E L W D T S G Q G R F C T I F R S Y S R G A Q L C P R Y S R G
D S L M D W P R G T

 (227) + The sequence shown here is der..., RAB40C L=206 Exons=6  (uc059olp.1) CODING
chr16:625037-625242
(227)
The sequence shown here is derived from an Ensembl  automatic analysis pipeline and should be consi
RAB40C
(206 bases) VARIANT for ALL (1)
SNPs=0 DELs=1 INS=0
Missense - Conservative:0  Non-conservative:1  Total Missense:1
Deleted:24 Inserted:0 residues
 Changes:
G -> T   (106)    Non-conservative missense
```

Illustration 93: Truncated protein caused by single nucleotide deletion (in other words, this is a non-sense mutation causing premature stop codon to be read)



```
DEL : .
AC
-->
A
```

Limitations of the software

The Virtual Genome Browser currently do not take into account disruptions in promoters, splice site disruptions or total loss of start or stop codons. It was only intended to give some clues on possible coding sequence disruptions.

Finding similar proteins in viruses

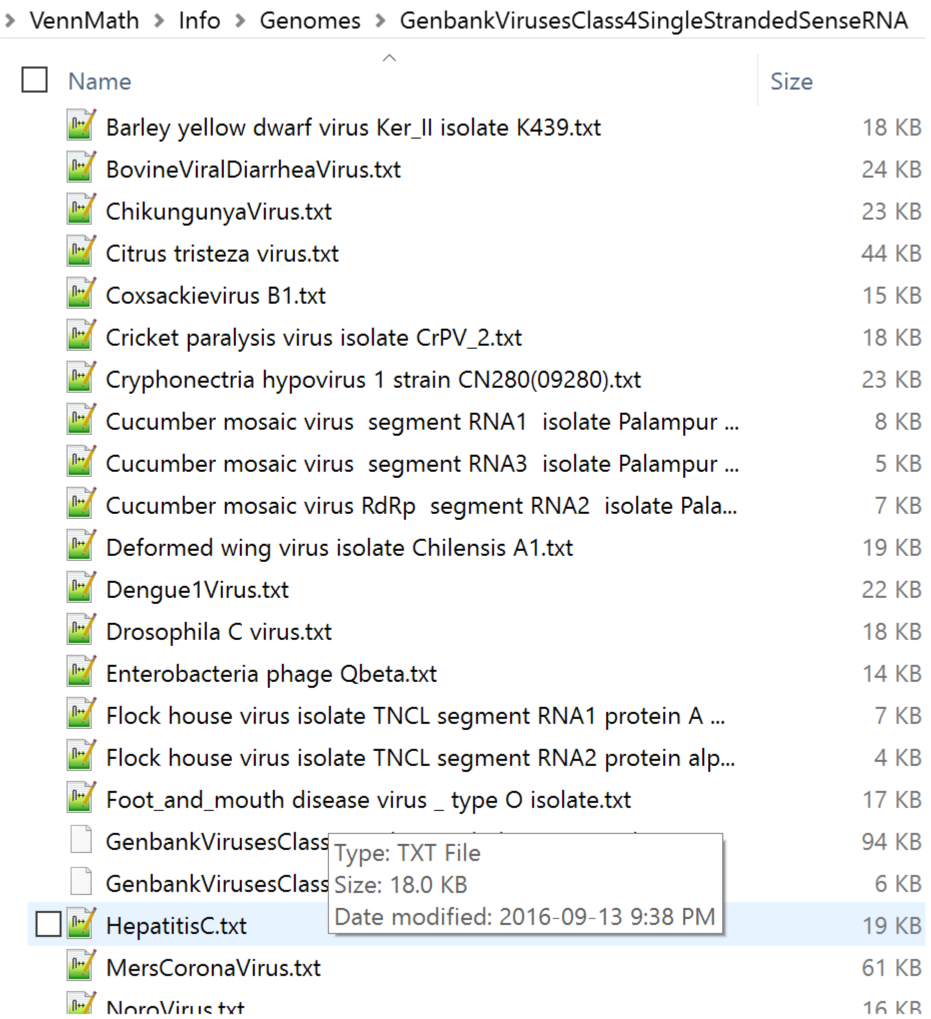
Getting the viral genome data

The software has the ability to look for protein similarity between genes on different sequences or chromosomes from the same or different genomes.

In order to demonstrate how this works, I will show how to search for similarity between different viruses.

The sequence data for the different virus have been obtained by creating text files in folders which starts with "GenBank" for Genbank format files or "Embl" for Ensembl format files.

The software will automatically scan the folders and then add the GenBank files to the per folder **.2bit** file.



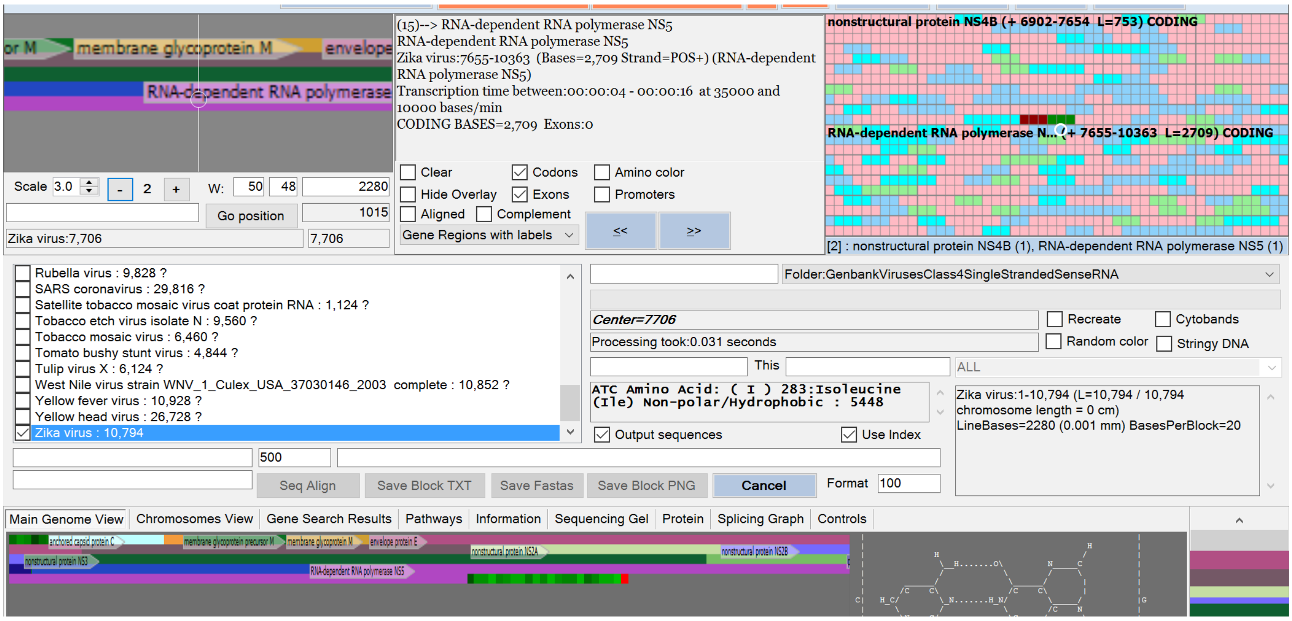
› VennMath › Info › Genomes › GenbankVirusesClass4SingleStrandedSenseRNA

☐ Name	Size
Barley yellow dwarf virus Ker_II isolate K439.txt	18 KB
BovineViralDiarrheaVirus.txt	24 KB
ChikungunyaVirus.txt	23 KB
Citrus tristeza virus.txt	44 KB
Coxsackievirus B1.txt	15 KB
Cricket paralysis virus isolate CrPV_2.txt	18 KB
Cryphonectria hypovirus 1 strain CN280(09280).txt	23 KB
Cucumber mosaic virus segment RNA1 isolate Palampur ...	8 KB
Cucumber mosaic virus segment RNA3 isolate Palampur ...	5 KB
Cucumber mosaic virus RdRp segment RNA2 isolate Pala...	7 KB
Deformed wing virus isolate Chilensis A1.txt	19 KB
Dengue1Virus.txt	22 KB
Drosophila C virus.txt	18 KB
Enterobacteria phage Qbeta.txt	14 KB
Flock house virus isolate TNCL segment RNA1 protein A ...	7 KB
Flock house virus isolate TNCL segment RNA2 protein alp...	4 KB
Foot_and_mouth disease virus _ type O isolate.txt	17 KB
GenbankVirusesClass	94 KB
GenbankVirusesClass	6 KB
☐ HepatitisC.txt	19 KB
MersCoronaVirus.txt	61 KB
NoroVirus.txt	16 KB

Type: TXT File
Size: 18.0 KB
Date modified: 2016-09-13 9:38 PM

I systematically went through all viruses in the book "Virus – an illustrated guide to 101 incredible microbes. By Dr Marilyn J Roossinck" and sorted them into sub folders based on the type of genome.

I now have all these viruses at my disposal and it allows me to search for similarities between these viruses.



or M
membrane glycoprotein M
envelope
RNA-dependent RNA polymerase
nonstructural protein NS4B (+ 6902-7654 L=753) CODING
RNA-dependent RNA polymerase N... (+ 7655-10363 L=2709) CODING
Go position
<<
>>
Seq Align
Save Block TXT
Save Fastas
Save Block PNG
Cancel
Main Genome View
Chromosomes View
Gene Search Results
Pathways
Information
Sequencing Gel
Protein
Splicing Graph
Controls
anchored capsid protein C
membrane glycoprotein precursor M
membrane glycoprotein M
envelope protein E
nonstructural protein NS2A
nonstructural protein NS2B
nonstructural protein NS3
RNA-dependent RNA polymerase NS5

(15)--> RNA-dependent RNA polymerase NS5
RNA-dependent RNA polymerase NS5
Zika virus:7655-10363 (Bases=2,709 Strand=POS+) (RNA-dependent RNA polymerase NS5)
Transcription time between:00:00:04 - 00:00:16 at 35000 and 10000 bases/min
CODING BASES=2,709 Exons:0

Scale 3.0 - 2 + W: 50 48 2280
1015
Zika virus:7,706 7,706

□ Clear ☑ Codons □ Amino color
□ Hide Overlay ☑ Exons □ Promoters
□ Aligned □ Complement
Gene Regions with labels

[2] : nonstructural protein NS4B (1), RNA-dependent RNA polymerase NS5 (1)

□ Rubella virus : 9,828 ?
□ SARS coronavirus : 29,816 ?
□ Satellite tobacco mosaic virus coat protein RNA : 1,124 ?
□ Tobacco etch virus isolate N : 9,560 ?
□ Tobacco mosaic virus : 6,460 ?
□ Tomato bushy stunt virus : 4,844 ?
□ Tulip virus X : 6,124 ?
□ West Nile virus strain WNV_1_Culex_USA_37030146_2003 complete : 10,852 ?
□ Yellow fever virus : 10,928 ?
□ Yellow head virus : 26,728 ?
☑ Zika virus : 10,794

500

Folder:GenbankVirusesClass4SingleStrandedSenseRNA

Center=7706 □ Recreate □ Cytobands
Processing took:0.031 seconds □ Random color □ Stringy DNA
This ALL

```
ATC Amino Acid: ( I ) 283:Isoleucine
(Ile) Non-polar/Hydrophobic : 5448
```

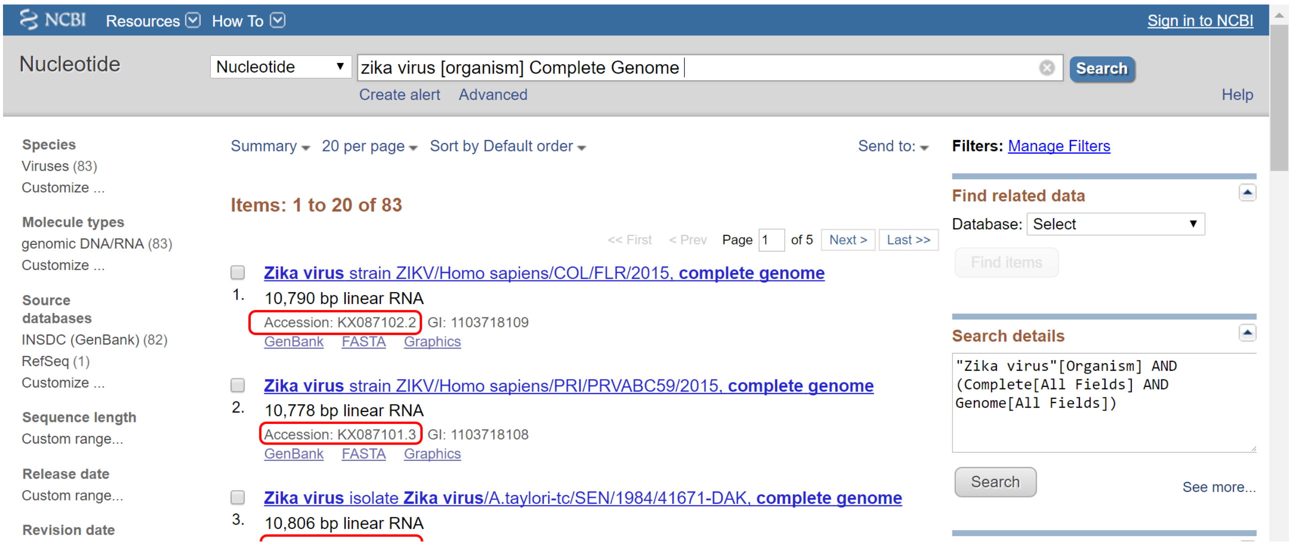
☑ Output sequences ☑ Use Index

Format 100

Zika virus:1-10,794 (L=10,794 / 10,794
chromosome length = 0 cm)
LineBases=2280 (0.001 mm) BasesPerBlock=20

After selecting a genome/folder, you can download viruses or multiple consecutive sequence data by entering the NCBI accession number in the download field on the **Main Tab.**

I downloaded the viruses at : https://www.ncbi.nlm.nih.gov/nuccore



Nucleotide

Nucleotide ▾ zika virus [organism] Complete Genome

Species
Viruses (83)
Customize ...

Molecule types
genomic DNA/RNA (83)
Customize ...

Source databases
INSDC (GenBank) (82)
RefSeq (1)
Customize ...

Sequence length
Custom range...

Release date
Custom range...

Revision date

Summary ▾ 20 per page ▾ Sort by Default order ▾ Send to: ▾

Items: 1 to 20 of 83

<< First < Prev Page 1 of 5 Next > Last >>

1. □ Zika virus strain ZIKV/Homo sapiens/COL/FLR/2015, complete genome
10,790 bp linear RNA
Accession: KX087102.2 GI: 1103718109
GenBank FASTA Graphics

2. □ Zika virus strain ZIKV/Homo sapiens/PRI/PRVABC59/2015, complete genome
10,778 bp linear RNA
Accession: KX087101.3 GI: 1103718108
GenBank FASTA Graphics

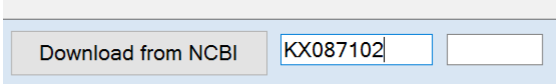
3. □ Zika virus isolate Zika virus/A.taylori-tc/SEN/1984/41671-DAK, complete genome
10,806 bp linear RNA

Filters: Manage Filters

Find related data
Database: Select

Search details

```
"Zika virus"[Organism] AND
(Complete[All Fields] AND
Genome[All Fields])
```

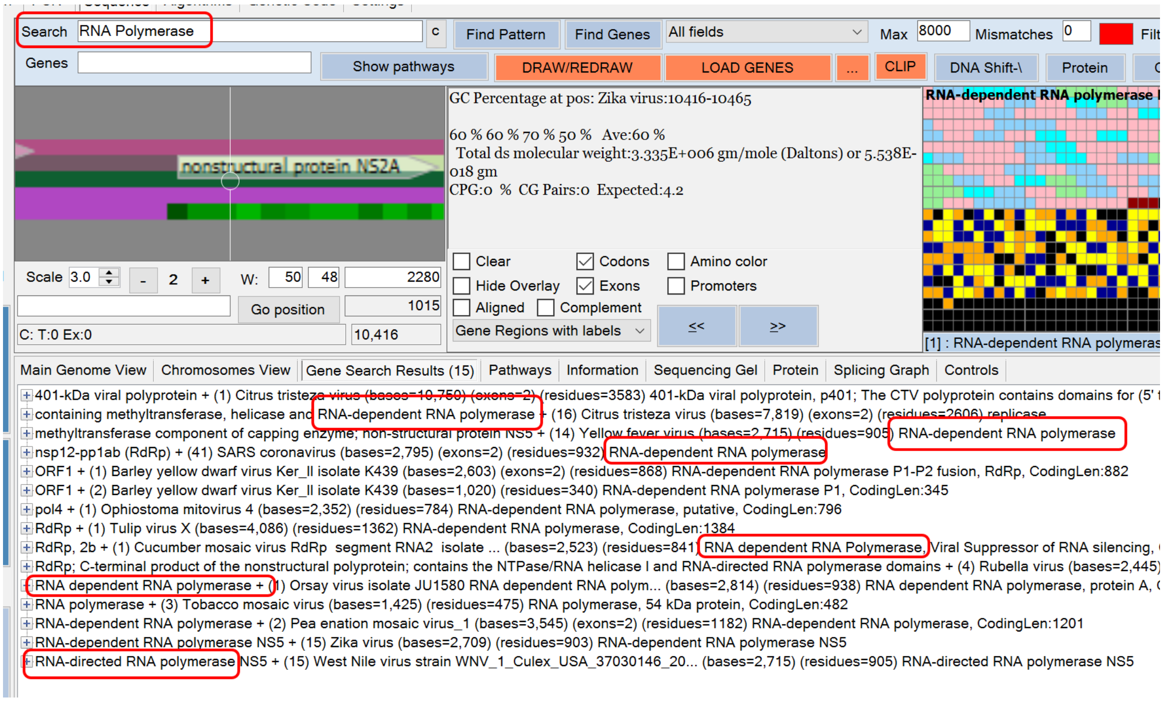


Download from NCBI KX087102

The second field can be used to download genomes of viruses which consists of multiple segments, such as the flu virus.

This will download the GenBank info into a text file an place it into the appropriate folder.

You then simply select the appropriate folder again (**BUT the one prefixed by "Folder:"**) and it will recreate the **.2bit** genome file (containing all of the viruses in the same folder). It will also recreate the search index and quick gene lookup index in order that you can quickly navigate to viral genes.



Search: RNA Polymerase

Max: 8000

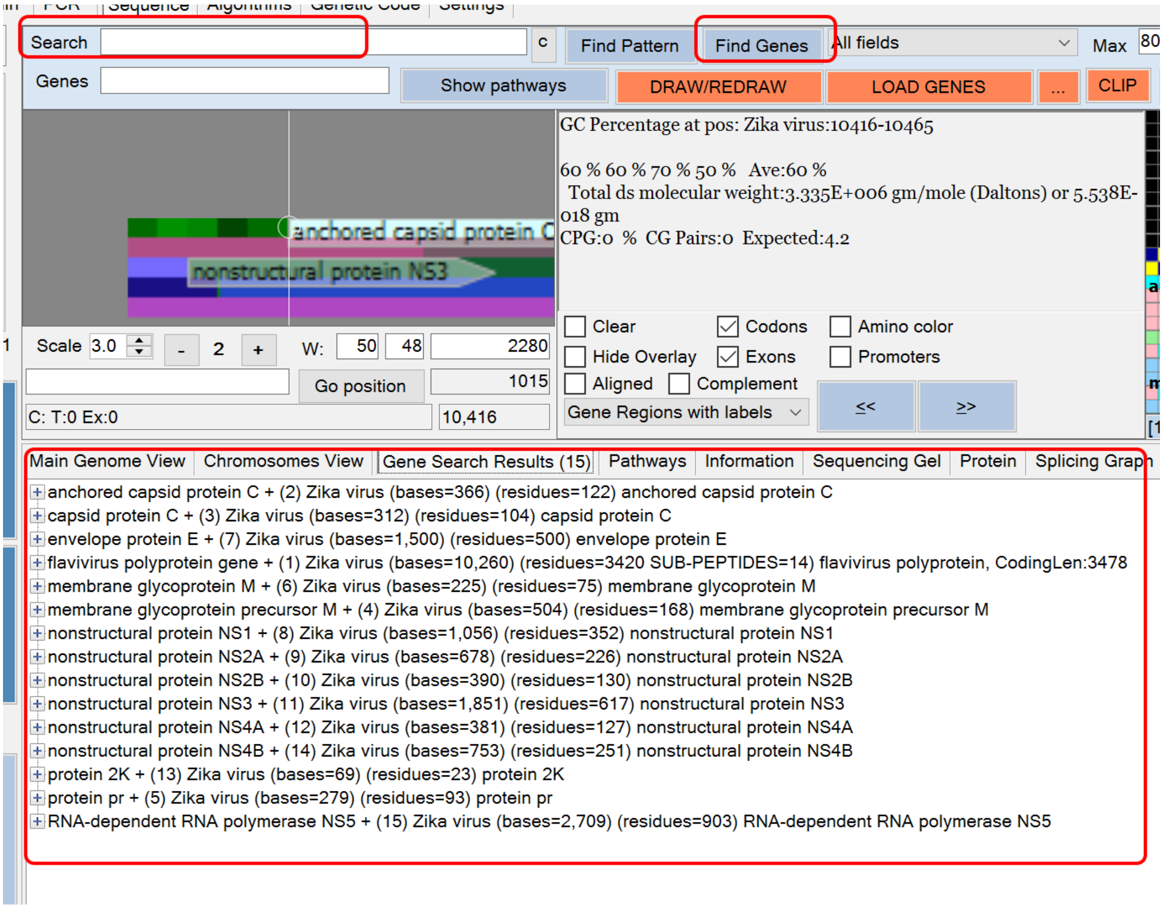
Mismatches: 0

Scale: 3.0

```
GC Percentage at pos: Zika virus:10416-10465

60 % 60 % 70 % 50 %  Ave:60 %
 Total ds molecular weight:3.335E+006 gm/mole (Daltons) or 5.538E-
018 gm
CPG:0  %  CG Pairs:0  Expected:4.2
```

```
401-kDa viral polyprotein + (1) Citrus tristeza virus (bases=10,750) (exons=2) (residues=3583) 401-kDa viral polyprotein, p401; The CTV polyprotein contains domains for (5' t
containing methyltransferase, helicase and RNA-dependent RNA polymerase + (16) Citrus tristeza virus (bases=7,819) (exons=2) (residues=2606) replicase
methyltransferase component of capping enzyme; non-structural protein NS5 + (14) Yellow fever virus (bases=2,715) (residues=905) RNA-dependent RNA polymerase
nsp12-pp1ab (RdRp) + (41) SARS coronavirus (bases=2,795) (exons=2) (residues=932) RNA-dependent RNA polymerase
ORF1 + (1) Barley yellow dwarf virus Ker_II isolate K439 (bases=2,603) (exons=2) (residues=868) RNA-dependent RNA polymerase P1-P2 fusion, RdRp, CodingLen:882
ORF1 + (2) Barley yellow dwarf virus Ker_II isolate K439 (bases=1,020) (residues=340) RNA-dependent RNA polymerase P1, CodingLen:345
pol4 + (1) Ophiostoma mitovirus 4 (bases=2,352) (residues=784) RNA-dependent RNA polymerase, putative, CodingLen:796
RdRp + (1) Tulip virus X (bases=4,086) (residues=1362) RNA-dependent RNA polymerase, CodingLen:1384
RdRp, 2b + (1) Cucumber mosaic virus RdRp  segment RNA2  isolate ... (bases=2,523) (residues=841) RNA dependent RNA Polymerase, Viral Suppressor of RNA silencing, (
RdRp; C-terminal product of the nonstructural polyprotein; contains the NTPase/RNA helicase I and RNA-directed RNA polymerase domains + (4) Rubella virus (bases=2,445)
RNA dependent RNA polymerase + (1) Orsay virus isolate JU1580 RNA dependent RNA polym... (bases=2,814) (residues=938) RNA dependent RNA polymerase, protein A, (
RNA polymerase + (3) Tobacco mosaic virus (bases=1,425) (residues=475) RNA polymerase, 54 kDa protein, CodingLen:482
RNA-dependent RNA polymerase + (2) Pea enation mosaic virus_1 (bases=3,545) (exons=2) (residues=1182) RNA-dependent RNA polymerase, CodingLen:1201
RNA-dependent RNA polymerase NS5 + (15) Zika virus (bases=2,709) (residues=903) RNA-dependent RNA polymerase NS5
RNA-directed RNA polymerase NS5 + (15) West Nile virus strain WNV_1_Culex_USA_37030146_20... (bases=2,715) (residues=905) RNA-directed RNA polymerase NS5
```

Illustration 94: Searching for genes in viruses



Scale: 3.0

```
GC Percentage at pos: Zika virus:10416-10465

60 % 60 % 70 % 50 %  Ave:60 %
 Total ds molecular weight:3.335E+006 gm/mole (Daltons) or 5.538E-
018 gm
CPG:0  %  CG Pairs:0  Expected:4.2
```

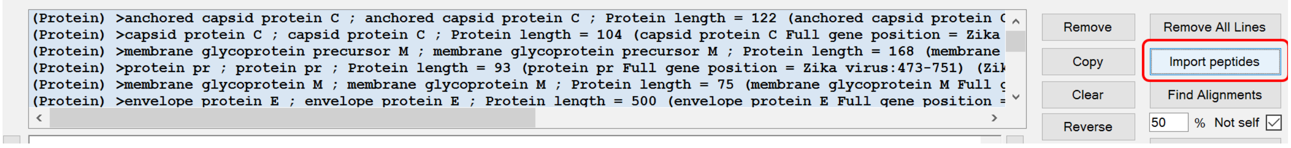
```
anchored capsid protein C + (2) Zika virus (bases=366) (residues=122) anchored capsid protein C
capsid protein C + (3) Zika virus (bases=312) (residues=104) capsid protein C
envelope protein E + (7) Zika virus (bases=1,500) (residues=500) envelope protein E
flavivirus polyprotein gene + (1) Zika virus (bases=10,260) (residues=3420 SUB-PEPTIDES=14) flavivirus polyprotein, CodingLen:3478
membrane glycoprotein M + (6) Zika virus (bases=225) (residues=75) membrane glycoprotein M
membrane glycoprotein precursor M + (4) Zika virus (bases=504) (residues=168) membrane glycoprotein precursor M
nonstructural protein NS1 + (8) Zika virus (bases=1,056) (residues=352) nonstructural protein NS1
nonstructural protein NS2A + (9) Zika virus (bases=678) (residues=226) nonstructural protein NS2A
nonstructural protein NS2B + (10) Zika virus (bases=390) (residues=130) nonstructural protein NS2B
nonstructural protein NS3 + (11) Zika virus (bases=1,851) (residues=617) nonstructural protein NS3
nonstructural protein NS4A + (12) Zika virus (bases=381) (residues=127) nonstructural protein NS4A
nonstructural protein NS4B + (14) Zika virus (bases=753) (residues=251) nonstructural protein NS4B
protein 2K + (13) Zika virus (bases=69) (residues=23) protein 2K
protein pr + (5) Zika virus (bases=279) (residues=93) protein pr
RNA-dependent RNA polymerase NS5 + (15) Zika virus (bases=2,709) (residues=903) RNA-dependent RNA polymerase NS5
```

Illustration 95: Listing all genes in the selected virus

Leaving the **Search Field empty**, and then clicking on **Find Genes** will load all the genes in the **currently selected sequence/virus**.

Finding similar proteins between viruses

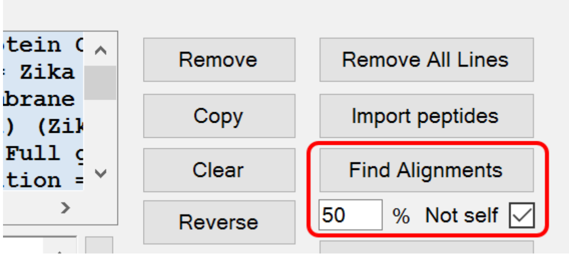
The first step is to import all the protein coding genes into the **Comparison List**. This is done after the virus/viruses and their genes have already been loaded. Now go to the **Main Tab** and click on the "**Import peptides**" button.



```
(Protein) >anchored capsid protein C ; anchored capsid protein C ; Protein length = 122 (anchored capsid protein C
(Protein) >capsid protein C ; capsid protein C ; Protein length = 104 (capsid protein C Full gene position = Zika
(Protein) >membrane glycoprotein precursor M ; membrane glycoprotein precursor M ; Protein length = 168 (membrane
(Protein) >protein pr ; protein pr ; Protein length = 93 (protein pr Full gene position = Zika virus:473-751) (Zik
(Protein) >membrane glycoprotein M ; membrane glycoprotein M ; Protein length = 75 (membrane glycoprotein M Full g
(Protein) >envelope protein E ; envelope protein E ; Protein length = 500 (envelope protein E Full gene position =
```

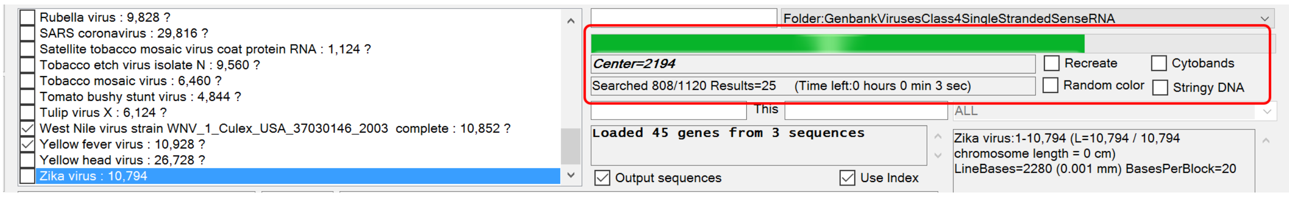
50 % Not self ☑

Now proceed to the virus or viruses you want to compare with and load their genes by clicking on **"LOAD GENES"** after selecting the virus sequences you want to compare with:



50 % Not self ☑

Then click the "**Find Alignments**" button on the **Main Tab**. The similarity minimum cut-off for proteins can be provided and the "Not self" check box can be used to ignore the virus from being compared against itself (I. e. the Zika Virus)



□ Rubella virus : 9,828 ?
□ SARS coronavirus : 29,816 ?
□ Satellite tobacco mosaic virus coat protein RNA : 1,124 ?
□ Tobacco etch virus isolate N : 9,560 ?
□ Tobacco mosaic virus : 6,460 ?
□ Tomato bushy stunt virus : 4,844 ?
□ Tulip virus X : 6,124 ?
☑ West Nile virus strain WNV_1_Culex_USA_37030146_2003 complete : 10,852 ?
☑ Yellow fever virus : 10,928 ?
□ Yellow head virus : 26,728 ?
□ Zika virus : 10,794

Folder:GenbankVirusesClass4SingleStrandedSenseRNA

Center=2194

□ Recreate □ Cytobands

Searched 808/1120 Results=25 (Time left:0 hours 0 min 3 sec)

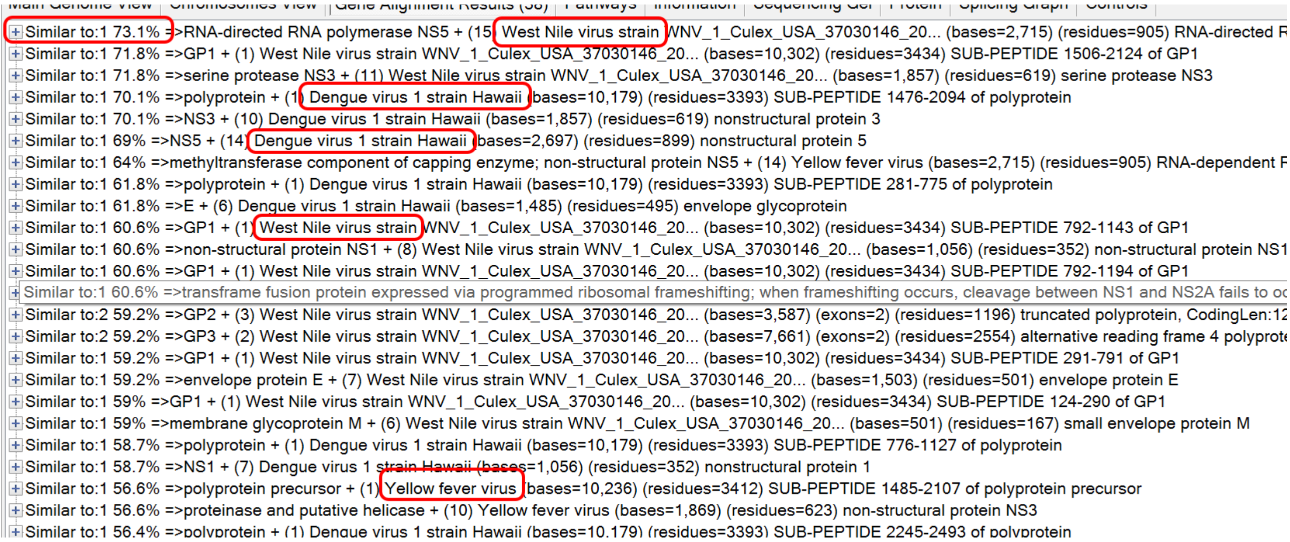
□ Random color □ Stringy DNA

```
Loaded 45 genes from 3 sequences
```

☑ Output sequences ☑ Use Index

Zika virus:1-10,794 (L=10,794 / 10,794
chromosome length = 0 cm)
LineBases=2280 (0.001 mm) BasesPerBlock=20

All the proteins and peptides with similarity is then listed in the Gene Tree:

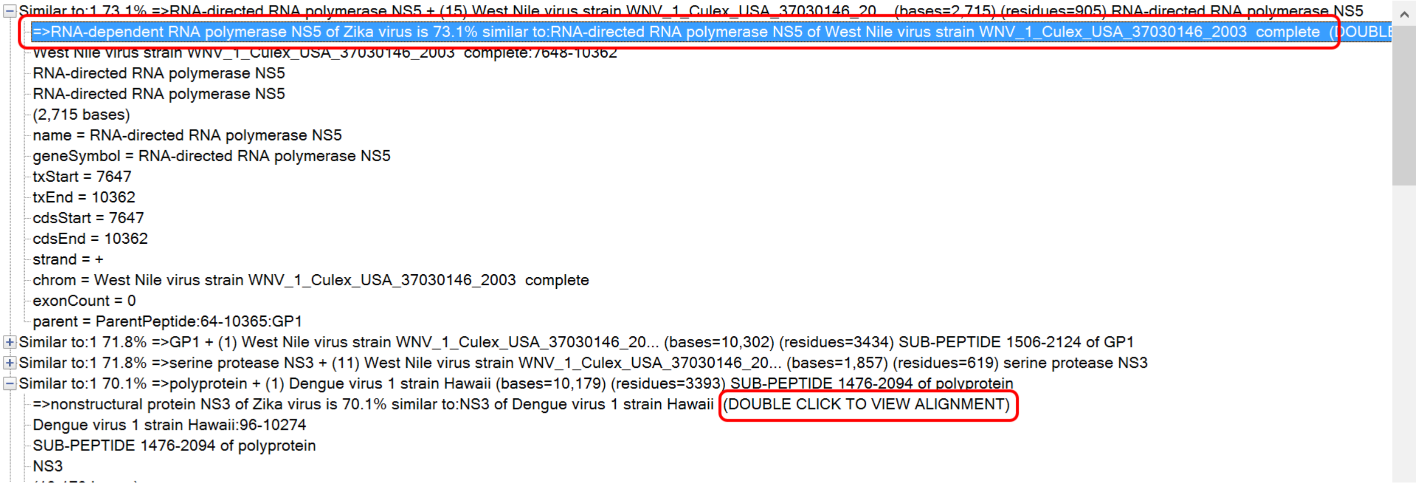


Main Genome View | Chromosomes View | Gene Alignment Results (38) | Pathways | Information | Sequencing Gel | Protein | Splicing Graph | Controls

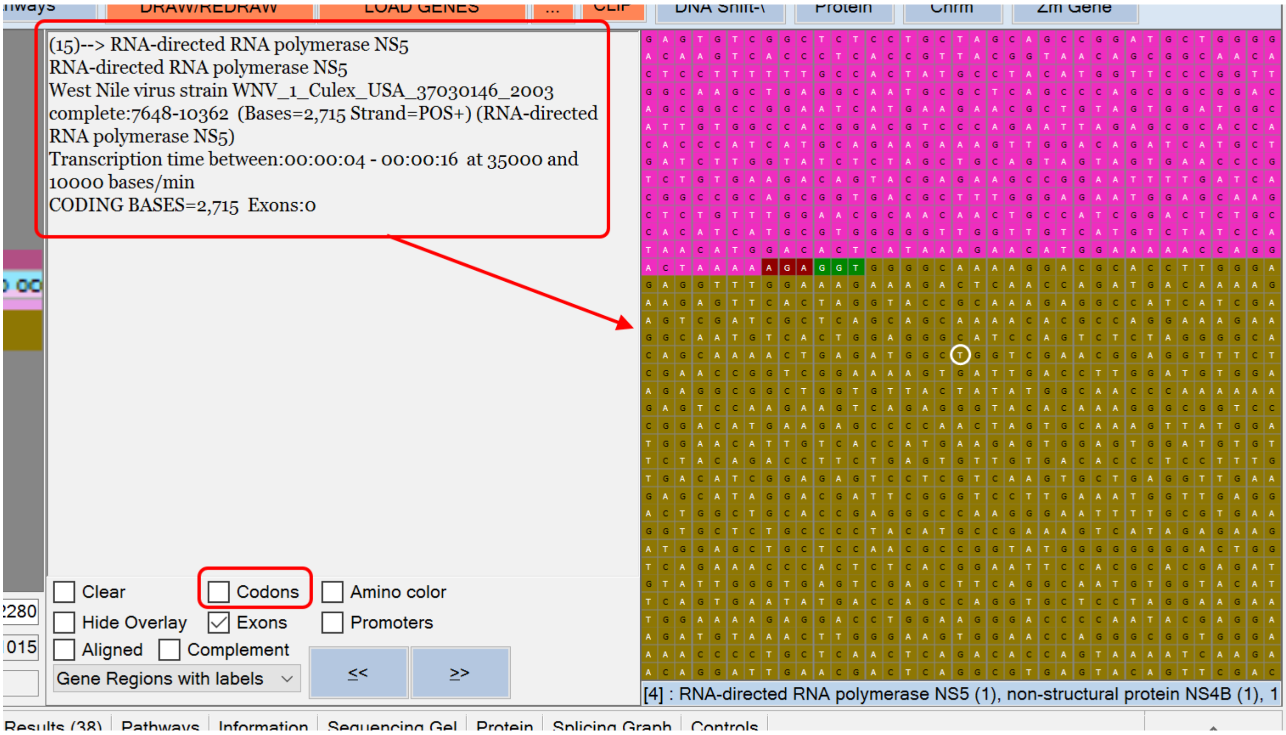
```
⊞Similar to:1 73.1% =>RNA-directed RNA polymerase NS5 + (15) West Nile virus strain WNV_1_Culex_USA_37030146_20... (bases=2,715) (residues=905) RNA-directed R
⊞Similar to:1 71.8% =>GP1 + (1) West Nile virus strain WNV_1_Culex_USA_37030146_20... (bases=10,302) (residues=3434) SUB-PEPTIDE 1506-2124 of GP1
⊞Similar to:1 71.8% =>serine protease NS3 + (11) West Nile virus strain WNV_1_Culex_USA_37030146_20... (bases=1,857) (residues=619) serine protease NS3
⊞Similar to:1 70.1% =>polyprotein + (1) Dengue virus 1 strain Hawaii (bases=10,179) (residues=3393) SUB-PEPTIDE 1476-2094 of polyprotein
⊞Similar to:1 70.1% =>NS3 + (10) Dengue virus 1 strain Hawaii (bases=1,857) (residues=619) nonstructural protein 3
⊞Similar to:1 69% =>NS5 + (14) Dengue virus 1 strain Hawaii (bases=2,697) (residues=899) nonstructural protein 5
⊞Similar to:1 64% =>methyltransferase component of capping enzyme; non-structural protein NS5 + (14) Yellow fever virus (bases=2,715) (residues=905) RNA-dependent F
⊞Similar to:1 61.8% =>polyprotein + (1) Dengue virus 1 strain Hawaii (bases=10,179) (residues=3393) SUB-PEPTIDE 281-775 of polyprotein
⊞Similar to:1 61.8% =>E + (6) Dengue virus 1 strain Hawaii (bases=1,485) (residues=495) envelope glycoprotein
⊞Similar to:1 60.6% =>GP1 + (1) West Nile virus strain WNV_1_Culex_USA_37030146_20... (bases=10,302) (residues=3434) SUB-PEPTIDE 792-1143 of GP1
⊞Similar to:1 60.6% =>non-structural protein NS1 + (8) West Nile virus strain WNV_1_Culex_USA_37030146_20... (bases=1,056) (residues=352) non-structural protein NS1
⊞Similar to:1 60.6% =>GP1 + (1) West Nile virus strain WNV_1_Culex_USA_37030146_20... (bases=10,302) (residues=3434) SUB-PEPTIDE 792-1194 of GP1
⊞Similar to:1 60.6% =>transframe fusion protein expressed via programmed ribosomal frameshifting; when frameshifting occurs, cleavage between NS1 and NS2A fails to oc
⊞Similar to:2 59.2% =>GP2 + (3) West Nile virus strain WNV_1_Culex_USA_37030146_20... (bases=3,587) (exons=2) (residues=1196) truncated polyprotein, CodingLen:12
⊞Similar to:2 59.2% =>GP3 + (2) West Nile virus strain WNV_1_Culex_USA_37030146_20... (bases=7,661) (exons=2) (residues=2554) alternative reading frame 4 polyprote
⊞Similar to:1 59.2% =>GP1 + (1) West Nile virus strain WNV_1_Culex_USA_37030146_20... (bases=10,302) (residues=3434) SUB-PEPTIDE 291-791 of GP1
⊞Similar to:1 59.2% =>envelope protein E + (7) West Nile virus strain WNV_1_Culex_USA_37030146_20... (bases=1,503) (residues=501) envelope protein E
⊞Similar to:1 59% =>GP1 + (1) West Nile virus strain WNV_1_Culex_USA_37030146_20... (bases=10,302) (residues=3434) SUB-PEPTIDE 124-290 of GP1
⊞Similar to:1 59% =>membrane glycoprotein M + (6) West Nile virus strain WNV_1_Culex_USA_37030146_20... (bases=501) (residues=167) small envelope protein M
⊞Similar to:1 58.7% =>polyprotein + (1) Dengue virus 1 strain Hawaii (bases=10,179) (residues=3393) SUB-PEPTIDE 776-1127 of polyprotein
⊞Similar to:1 58.7% =>NS1 + (7) Dengue virus 1 strain Hawaii (bases=1,056) (residues=352) nonstructural protein 1
⊞Similar to:1 56.6% =>polyprotein precursor + (1) Yellow fever virus (bases=10,236) (residues=3412) SUB-PEPTIDE 1485-2107 of polyprotein precursor
⊞Similar to:1 56.6% =>proteinase and putative helicase + (10) Yellow fever virus (bases=1,869) (residues=623) non-structural protein NS3
⊞Similar to:1 56.4% =>polyprotein + (1) Dengue virus 1 strain Hawaii (bases=10,179) (residues=3393) SUB-PEPTIDE 2245-2493 of polyprotein
```

By double clicking on the gene itself, you will be navigated to the specific gene in the Genome Browser.

By double clicking on the second line for each gene entry, you will be taken to the alignment screen.



Similar to:1 73.1% =>RNA-directed RNA polymerase NS5 + (15) West Nile virus strain WNV_1_Culex_USA_37030146_20... (bases=2,715) (residues=905) RNA-directed RNA polymerase NS5
=>RNA-dependent RNA polymerase NS5 of Zika virus is 73.1% similar to:RNA-directed RNA polymerase NS5 of West Nile virus strain WNV_1_Culex_USA_37030146_2003 complete (DOUBLE
West Nile virus strain WNV_1_Culex_USA_37030146_2003 complete:7648-10362
RNA-directed RNA polymerase NS5
RNA-directed RNA polymerase NS5
(2,715 bases)
name = RNA-directed RNA polymerase NS5
geneSymbol = RNA-directed RNA polymerase NS5
txStart = 7647
txEnd = 10362
cdsStart = 7647
cdsEnd = 10362
strand = +
chrom = West Nile virus strain WNV_1_Culex_USA_37030146_2003 complete
exonCount = 0
parent = ParentPeptide:64-10365:GP1
Similar to:1 71.8% =>GP1 + (1) West Nile virus strain WNV_1_Culex_USA_37030146_20... (bases=10,302) (residues=3434) SUB-PEPTIDE 1506-2124 of GP1
Similar to:1 71.8% =>serine protease NS3 + (11) West Nile virus strain WNV_1_Culex_USA_37030146_20... (bases=1,857) (residues=619) serine protease NS3
Similar to:1 70.1% =>polyprotein + (1) Dengue virus 1 strain Hawaii (bases=10,179) (residues=3393) SUB-PEPTIDE 1476-2094 of polyprotein
=>nonstructural protein NS3 of Zika virus is 70.1% similar to:NS3 of Dengue virus 1 strain Hawaii (DOUBLE CLICK TO VIEW ALIGNMENT)
Dengue virus 1 strain Hawaii:96-10274
SUB-PEPTIDE 1476-2094 of polyprotein
NS3



(15)--> RNA-directed RNA polymerase NS5
RNA-directed RNA polymerase NS5
West Nile virus strain WNV_1_Culex_USA_37030146_2003
complete:7648-10362 (Bases=2,715 Strand=POS+) (RNA-directed RNA polymerase NS5)
Transcription time between:00:00:04 - 00:00:16 at 35000 and 10000 bases/min
CODING BASES=2,715 Exons:0

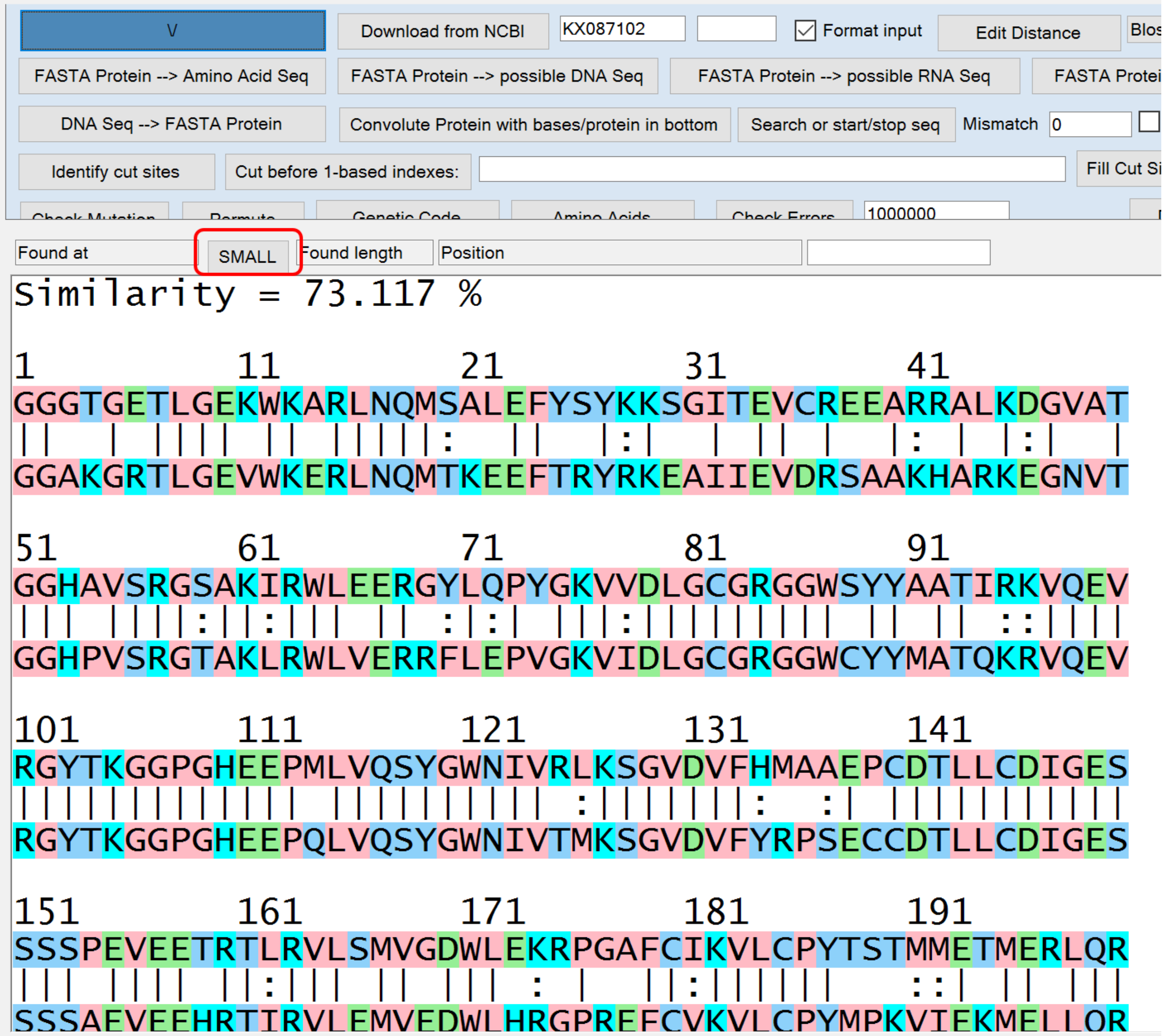
□ Clear □ Codons □ Amino color
□ Hide Overlay ☑ Exons □ Promoters
□ Aligned □ Complement
Gene Regions with labels

[4] : RNA-directed RNA polymerase NS5 (1), non-structural protein NS4B (1), 1

Illustration 96: DNA View of RNA directed RNA Polymerase for Yellow Fever virus

By un-checking the “Codons” the separate genes in the virus genome can more easily be distinguished. When there are multiple genes on top of each other, it is indicated with the line below the DNA View. By pressing + and – you can cycle through the genes in this view.

The alignment view allow you to clearly see the Conservative amino acid substitutions (marked with a colon : as well as the non-conservative subsitutions marked with a space between the similar amino acids which are marked with a vertical line.


V
Download from NCBI
Edit Distance
Blos
FASTA Protein --> Amino Acid Seq
FASTA Protein --> possible DNA Seq
FASTA Protein --> possible RNA Seq
FASTA Protei
DNA Seq --> FASTA Protein
Convolute Protein with bases/protein in bottom
Search or start/stop seq
Identify cut sites
Cut before 1-based indexes:
Fill Cut Si
Check Mutation
Permute
Genetic Code
Amino Acids
Check Errors
Found at
SMALL
Found length
Position

KX087102

☑ Format input

Mismatch 0 □

1000000

```
Similarity = 73.117 %

1         11        21        31        41
GGGTGETLGEKWKARLNQMSALEFYSYKKSGITEVCREEARRALKDGVAT
||  | |||| || |||||:  ||  |:|  | || |  |: | |:|  |
GGAKGRTLGEVWKERLNQMTKEEFTRYRKEAIIEVDRSAAKHARKEGNVT

51        61        71        81        91
GGHAVSRGSAKIRWLEERGYLQPYGKVVDLGCGRGGWSYYAATIRKVQEV
||| ||||:||:||| || :|:| |||:||||||||| || || ::||||
GGHPVSRGTAKLRWLVERRFLEPVGKVIDLGCGRGGWCYYMATQKRVQEV

101       111       121       131       141
RGYTKGGPGHEEPMLVQSYGWNIVRLKSGVDVFHMAAEPCDTLLCDIGES
||||||||||||| |||||||||| :|||||||:  :| |||||||||||
RGYTKGGPGHEEPQLVQSYGWNIVTMKSGVDVFYRPSECCDTLLCDIGES

151       161       171       181       191
SSSPEVEETRTLRVLSMVGDWLEKRPGAFCIKVLCPYTSTMMETMERLQR
||| |||| ||:||| || ||| : |  ||:||||||   ::| || |||
SSSAEVEEHRTIRVLEMVEDWLHRGPREFCVKVLCPYMPKVIEKMELLQR
```

Illustration 97: Protein sequence similarity between Zika and Yellow Fever virus RNA directed RNA polymerase

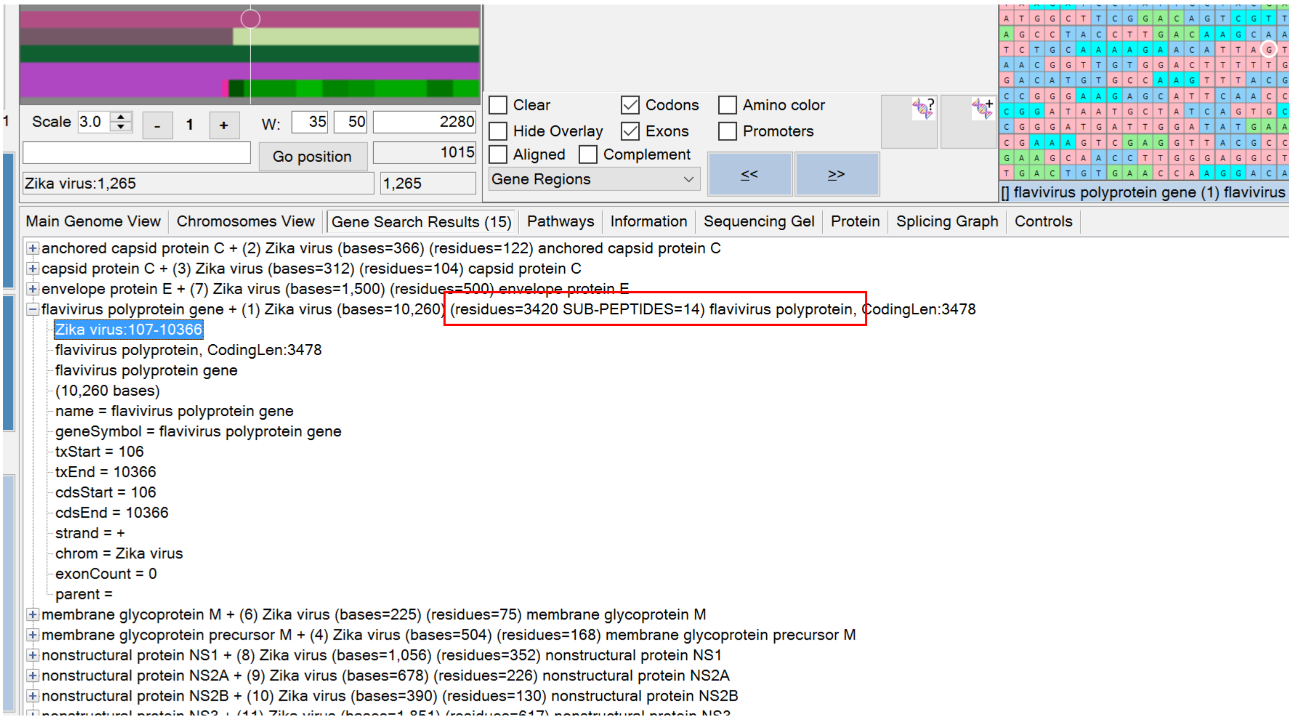
All of these alignment results are also output to files in the genomes folder as “Alignments_Date...” in order that you have a record of them.

Looking at sub-peptides which are obtained due to proteases cleaving larger poly-protein gene products into smaller peptides

In a a group of virus called the **Flaviviridae**, (in fact many other viruses), viral proteins are3 often formed by the mRNA of the virus coding for a single **poly-protein**, which is then cleaved into sub-peptides.

The software is able to detect this in GenBank files and display the protein view appropriately.

When the genes are loaded for the Zika Virus, click on **Find Genes**. This will list all of the genes in the zika virus singl;e stranded RNA genome. Notice the term: **SUB-PEPTIDES**. This is an indication that there are peptide subdivisions in this virus' genes, which can be appropriately indicated in the protein view.



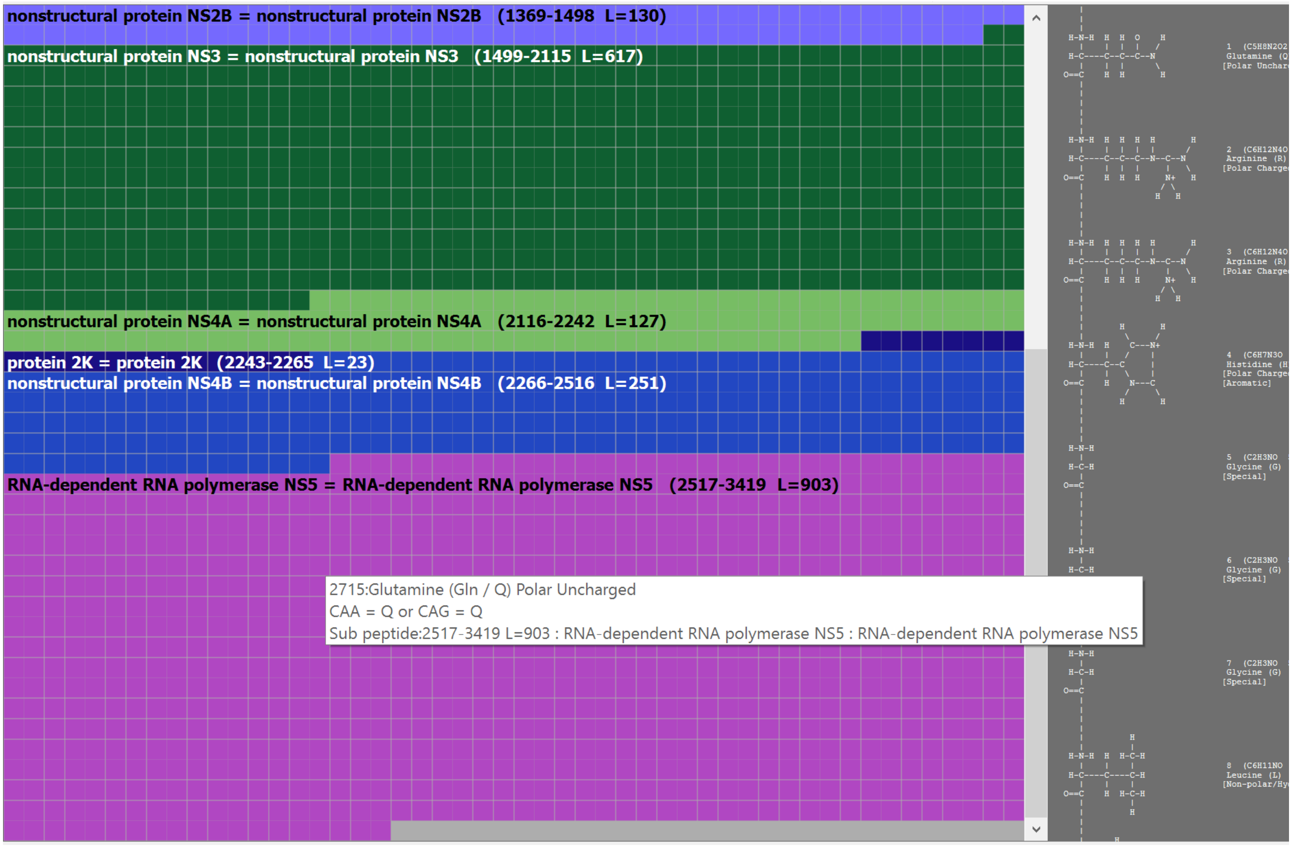
Zika virus:1,265

□ Clear ☑ Codons □ Amino color
□ Hide Overlay ☑ Exons □ Promoters
□ Aligned □ Complement
Gene Regions

```
+anchored capsid protein C + (2) Zika virus (bases=366) (residues=122) anchored capsid protein C
+capsid protein C + (3) Zika virus (bases=312) (residues=104) capsid protein C
+envelope protein E + (7) Zika virus (bases=1,500) (residues=500) envelope protein E
-flavivirus polyprotein gene + (1) Zika virus (bases=10,260) (residues=3420 SUB-PEPTIDES=14) flavivirus polyprotein, CodingLen:3478
    Zika virus:107-10366
    flavivirus polyprotein, CodingLen:3478
    flavivirus polyprotein gene
    (10,260 bases)
    name = flavivirus polyprotein gene
    geneSymbol = flavivirus polyprotein gene
    txStart = 106
    txEnd = 10366
    cdsStart = 106
    cdsEnd = 10366
    strand = +
    chrom = Zika virus
    exonCount = 0
    parent =
+membrane glycoprotein M + (6) Zika virus (bases=225) (residues=75) membrane glycoprotein M
+membrane glycoprotein precursor M + (4) Zika virus (bases=504) (residues=168) membrane glycoprotein precursor M
+nonstructural protein NS1 + (8) Zika virus (bases=1,056) (residues=352) nonstructural protein NS1
+nonstructural protein NS2A + (9) Zika virus (bases=678) (residues=226) nonstructural protein NS2A
+nonstructural protein NS2B + (10) Zika virus (bases=390) (residues=130) nonstructural protein NS2B
+nonstructural protein NS3 + (11) Zika virus (bases=1,851) (residues=617) nonstructural protein NS3
```

Illustration 98: SUB-PEPTIDES available for the Zika Virus

When you now show the **Protein View** for this specific virus, the different peptide segments which are cleaved are indicated in the following picture.



nonstructural protein NS2B = nonstructural protein NS2B (1369-1498 L=130)

nonstructural protein NS3 = nonstructural protein NS3 (1499-2115 L=617)

nonstructural protein NS4A = nonstructural protein NS4A (2116-2242 L=127)

protein 2K = protein 2K (2243-2265 L=23)

nonstructural protein NS4B = nonstructural protein NS4B (2266-2516 L=251)

RNA-dependent RNA polymerase NS5 = RNA-dependent RNA polymerase NS5 (2517-3419 L=903)

2715:Glutamine (Gln / Q) Polar Uncharged
CAA = Q or CAG = Q
Sub peptide:2517-3419 L=903 : RNA-dependent RNA polymerase NS5 : RNA-dependent RNA polymerase NS5

Something insightful to do is to examine the amino acids which lie at the transitions between these peptides, in order to determine what the amino acid cleavage motifs are at which the protease enzymes will cut the poly-protein.