CPSC 533 Analysis Project

Track-Based Genome Browsers: Analysis, Challenges, and the Next Steps

Slind, Jillian

Department of Computer Science, University of British Columbia

1 INTRODUCTION

1.1 Genome Browsing

A common task in bioinformatics and molecular biology analysis today is the analysis of a given sequence of interest with respect to that artifact's surroundings. In other words, the task is to find out where a target sequence is coded within a particular genome and determine the genomic context of that target sequence. There is a vast amount of useful information that can be found by performing these searches, such as upstream or downstream regulatory sequences that control how frequently your sequence becomes expressed, or nearby products that can be thought to be expressed simultaneously or in tandem with your target sequence.

There are a few ways to accomplish this contextual analysis of a particular sequence:

The first way is to extract the sequences of portions of a genome upstream and downstream of the target sequence, and query a database for artifacts that align with you particular sequences. However, not only is this a cumbersome process, but not as accessible to other users. Additionally, it is impossible to predict exactly how far up-‑ or down-‑stream some interesting artifacts may turn up initially. If you only extract one kb (kilobase, 1000 bases or 1000 letters) on either side of your sequence, you might miss some interesting artifacts that exist 2 kb up- or downstream. If you have extracted too large of the sequence appearing before and after your sequence, the sheer number of unrelated sequences might occlude the ones that would be of interest to your investigation. Extracting many different sizes of flanking sequences is impractical due to the tedious nature of the task, not to mention the space and time required.

Thankfully, many organizations have worked to develop a more intuitive way to view your sequence in the context of a genome. Genome Browsers can show the context of your sequence within a genome graphically, allowing for direct analysis or exploration upstream and downstream. The genomes loaded into these browsers have annotations and markings within their sequences identifying other coding regions or other regions of potential interest within the genome. Thus, by graphically exploring upstream and downstream of a particular sequence, interesting regions can be identified quickly and efficiently.

A common form of genome browser for the purposes of sequence analysis is the track-‑based genome browser. These browsers center the genome being viewed on the sequence of interest, and show the flanking regions of the genome. One advantage of these track-‑based genome browsers is that the user can elect to view more of the genome on either side of the sequence by scrolling, or adjusting the level of zoom. Coloured lines or blocks encode the genome and sequence of interest, and additional sequences are shown by a variety of methods dependent on the particular genome browser.

Not all track-‑based genome browsers, however, are created equally. From the author's experiences in dealing with them and teaching others how to use them, often the encoding methods can cause confusion if not done properly, and navigation can become cumbersome. Thus, there is often a learning curve involved before the user can take advantage of the tools made available through the browser itself. If the learning curve is too steep, a potential user could get frustrated and abandon the genome browser without getting the information they seek. Such disadvantages, with respect to the field of information visualization, might be shown in the way the browser encodes its data, the way the individual data tracks are encoded, or navigational difficulties/ that an individual may have while using the browser.

The goal of this paper is to provide an overview of current genome browsing technology, and analyze the effectiveness of the visualization tools that a pair of genome browsers utilize in order to develop a visual representation of the data. At the end of the paper, areas for further development are identified, with the hopes of improving the visualization strategies that current genome browsers already incorporate.

1.2 Domain, Tasks, and Data

1.2.1 Domain

The contextual analysis of a gene is a common task in fields related to molecular genetics, including but not limited to: molecular medicine, genetics, molecular biology, biochemistry, bioinformatics, biology, cancer studies, immunology studies, and other fields associated with the health sciences. These researchers normally want to see information that is relative to their study – for example, a cancer biologist may be interested in learning the potential for a gene that is commonly present/not present in cancer cells to be regulated, or how that particular gene is regulated. If the user has a custom track of information they would prefer to visualize with respect to the target chromosome, they can upload their data to a pre-existing genome browser as a custom track. In the below section, the individual analysis tasks a user may perform are discussed.

1.2.2 Tasks

The tasks associated with contextual analysis of a particular sequence commonly include the determination of important sequences that are present near the target sequence, sequence conservation levels, and regulatory features. We define *sequence conservation* as the presence same sequence appearing in other organisms (i.e. given some species *x*, is this sequence present, and how similar are the sequences between *x* and humans?). We define *regulatory features* as any component or property that is exhibited within or near the sequence that would affect how often this sequence is *expressed* (For a definition of *sequence expression*, please refer to the Glossary at the end of the document).

An additional area of study is the overall *stability* of the given gene. We define *gene stability* as the relative immutability of the gene within the context of its location. An instability may be

Glossary of terms available at the end of the document.

characterized by the existence of a high probability of mutation of the particular region in question.

1.2.3 Data

The data of interest to this particular analysis is the phosphatase tensin homolog (PTEN) gene, within the human genome, on chromosome ten. PTEN is a tumor suppressor protein that operates by inhibiting a cell proliferation pathway [1]. This gene has been found to be mutated in a few different types of cancer cells [2], and is the gene of interest in this study. It is encoded on the forward strand of chromosome 10 in humans. Since the browsers being used utilize their own data for visualizations, there was no additional data to download to perform this project. If, however, a different browser was downloaded to be used, the user would need the following:

- A table describing the chromosome(s) of interest, including the locations of each individual chromosome band.
- A table indicating the locations, sizes, intron/exon content, of each individual gene encoded by the chromosome(s) of interest.
- Several tables containing data information that represent each additional track that the user may have interest in with respect to their genome browser.

2 Related Work

2.1 Track-based Genome Browsers

Anno-J is a genome browser designed to incorporate distributed data, styles, annotations, and information into a single view [3]. As seen in Figure 1, track selection is performed on the left-hand side of the browser, and individual tracks are shown on the right-hand side. Figure 1 shows the application of Anno-J to highly annotated *Arabidopsis thaliana* genome, generated by the Salk institute [4].

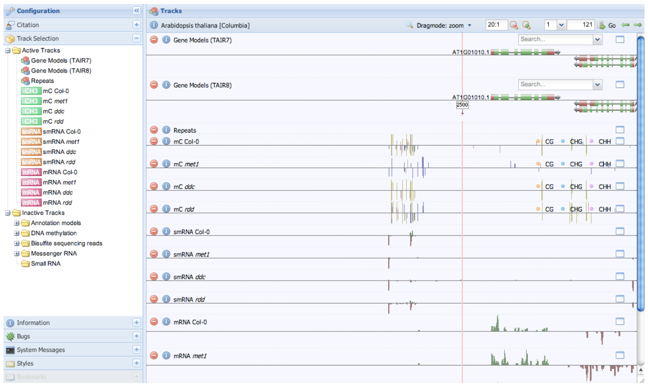


Figure 1. The Anno-J Genome Browser, image from the Anno-J website.

The Generic Model Organism Database project (GMOD) incorporates a genome browser into the many views provided by the project, named GBrowse [5]. This is designed to simplify the incorporation of a genome browser into the individual databases hosted by members of the project. This browser includes a series of glyphs and other annotation features that the users can choose from, or the ability for users to upload their own glyphs. A sample image is shown in Figure 2.

The Ensembl Genome Browser is an online tool that visualizes Ensembl genomes and features associated with those genomes [6]. As this browser is further discussed in later sections, not too much detail will be given here.

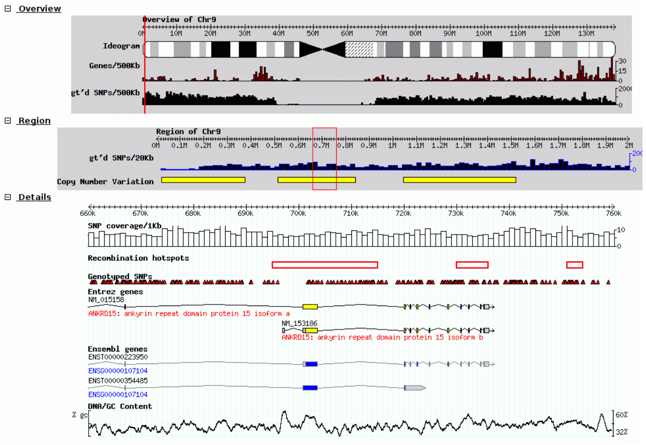


Figure 2. A diagram of the GMOD GBrowse application as shown on the website in [5].

The NCBI MapViewer is another large online visualization tool [7]. A trait unique to the NCBI MapViewer is the vertical orientation of the genome (Figure 3). This trait allows for the naming of individual elements in a track at a low magnification without occluding the rest of the visualization. However, numerical data (such as signal detection of specific regions) also have to undergo this transformation, which means that although the spatial channels that encode such information are still utilized, but the orientation may cause a slight learning curve, or worse, inaccurate interpretation of the graphs.

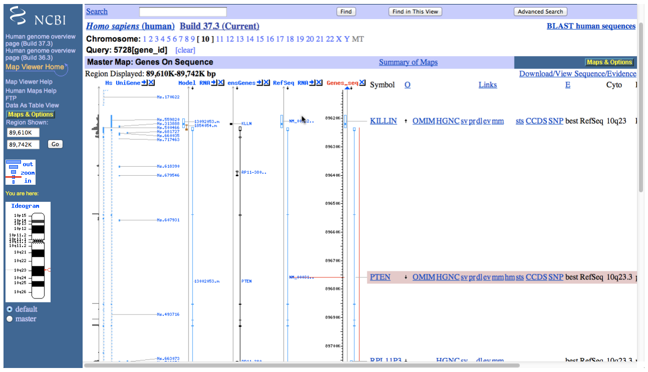

NCBI
Search
Find
Find in This View
Advanced Search
Homo sapiens (human) Build 37.3 (Current)
BLAST human sequences
Chromosome: 1 2 3 4 5 6 7 8 9 [10] 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT
Query: 5728[gene_id] [clear]
Master Map: Genes On Sequence
Summary of Maps
Maps & Options
Region Displayed: 89,610K-89,742K bp
Download/View Sequence/Evidence
Map Viewer Home
Map Viewer Help
Human Maps Help
FTP
Data As Table View
Region Shown:
89,610K
89,742K
Go
out
zoom
in
You are here:
Ideogram
default
master
Symbol
Links
Cyto
KILLIN
OMIM HGNC sv pr dl ev mm
sts CCDS SNP best RefSeq 10q23
PTEN
OMIM HGNC sv pr dl ev mm hm sts CCDS SNP best RefSeq 10q23.3

Figure 3. A screenshot showing NCBI's MapViewer after querying for the PTEN gene.

The Galaxy Genome Browser is a genome browsing component recently added to the many tools of the Galaxy project [8]. The Galaxy genome browser allows several tools to be utilized on the data, and resultant tracks to be generated. This tool also allows for filters to be applied on specific tracks, allowing only the data relevant to the researcher's interests to be displayed.

The UCSC genome browser is another web-based genome browser, similar to Ensembl [9]. A unique feature of this particular display is that everything is displayed in a single web frame, including all customization options. Selecting or changing the tracks that appear in the display is performed by scrolling to the bottom of the screen, as opposed to navigating through an additional menu.

The Savant genome browser is a desktop tool that incorporates multiple styles of view in addition to the track-based browser.

This allows for more in-depth analysis of one area of interest (for example, regulation) [10].

2.2 Other Genome Visualizations

A few other genome visualization tools are detailed below:

SequenceJuxtaposer is a tool for the large-scale comparison of biological sequences [11]. It is designed to show multiple sequences in tandem with one another so that the user can identify key similarities and differences. One visualization technique that makes this tool the most notable is their usage of a Focus+Context visualization by employing a form of stretch and squish navigation.

Gremlin (Genome Rearrangement Explorer with Multi-scale, Linked Interactions) is a visualization tool that is designed for the analysis of genomic rearrangements, that is, the rearrangement of subsets of sequences within a genome, on top of other sequence-based mutations [12]. Rearrangements are shown using bands that connect two portions of the genome together.

MizBee is another genome visualization tool that is designed to compare two genomes, or specifically mapping between components in one genome and their locations in the other [13]. This incorporates a circular view of the two genomes and edge-bundling techniques to clean up the visualization for the user.

3 Data Analysis

3.1 Data Analysis Tasks

The tasks for the analysis of PTEN that will be performed is as follows:

- Analysis of genes that are encoded near the PTEN gene. If the genes encoded near the PTEN gene are somewhat functionally related to the PTEN gene itself, then they could be used in future studies.
- Conservation analysis of PTEN. *Conservation* is how preserved the gene is from one species to another. By analyzing whether this gene appears in other species, and if so, how identical this gene is between species, we can make assumptions about the evolutionary history of this gene, as well as the state of tumor suppression in those other species.
- Analysis of the stability of the PTEN gene. The *stability* of the gene is the susceptibility of this gene to mutation. The amount of measured mutations, as well as properties of the DNA itself (such as the "strength" of that region of DNA, i.e. how likely that the DNA will "unzip" without the presence of specialized enzymes), all play a role in the stability of the gene.
- Regulation of the PTEN gene. This is the utilization of the genome browser to identify regions of that PTEN gene that are important for its regulation – i.e., specific regions dedicated to control how often the PTEN gene is transcribed (there could be other factors post-transcription that further regulate the production of the PTEN protein from the PTEN gene, but they are beyond the scope of this project).

3.2 Choosing Tools for Analysis

From the author's previous experience, the genome browsers from Ensembl, UCSC, and NCBI seem to be the most prevalent. Also, because each of the tools has a substantial database of information hosted by their respective servers, the amount of data that can be detected from each browser is relatively high. The genome browser from the Galaxy Project was another very strong contender for this analysis project, but the author's access to information to upload was limited at the given time.

Each of the three genome browsers contained data that related to the tasks as defined in section 3.1. Ensembl and the UCSC genome browser, however, were chosen because of the amount of additional information provided in each view. The tracks were also shown in a horizontal fashion, as opposed to the NCBI MapViewer's vertical approach. This allows for corresponding tracks (especially those containing numerical data) to be compared more easily without the translation between horizontal and vertical views.

3.3 Analysis of the Ensembl Genome Browser

The general operation of the Ensembl Genome Browser is as follows:

To look up a particular gene in the hosted database, its name can be input in the search box at the top of the home page (Figure 7).

The resulting search is split up into three views: the view of the entire chromosome (we define this as the *chromosome view*, Figure 4), the "Region in Detail" view (we define this as the *detail view*, Figure 5), and a view containing a set of tracks to help analyse the region (we define this as the *analysis view*, Figure 6).

The *chromosome view* consists an overview-style view of the entire chromosome in which the target sequence is located. The "bars" of the chromosome are encoded using greyscale, and the *centromere* of the chromosome is encoded by two triangles pointing towards each other. The red box indicates the area of the genome that the detail view is showing. Clicking on other areas of the genome will centre the red box (and corresponding detail and analysis views) on the selected area.

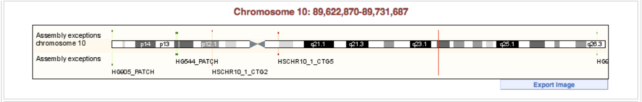


Figure 4. The chromosome view of the Ensembl genome browser.

The *detail view* highlights the gene of interest, and shows surrounding genes encoded in the chromosome (Figure 3). Along the top of the detail view the chromosome bands are displayed. This is followed by the *contigs* used in the Ensembl genome assembly process. We define *contigs* as a consensus region of DNA formed from the overlapping of many DNA fragments. Framing this view on the top and bottom is a ruler showing scale (in terms of base pairs/nucleotides) of the diagram. The genes encoded on the chromosome are encoded using bars along the top of the view, with the size showing the relative length of each individual gene, and colour encoding the type of gene. The names are all listed underneath each respective gene with a small glyph pointing to the leftmost position in each bar, and another glyph indicating the direction that the gene is transcribed from the DNA strand. The names are colour-coded in the same fashion as the genes themselves. The red box in this view shows the portion of this view that is selected for analysis in the *analysis view*. The user can modify what is being shown in the analysis view by clicking or brushing to select an alternative region.

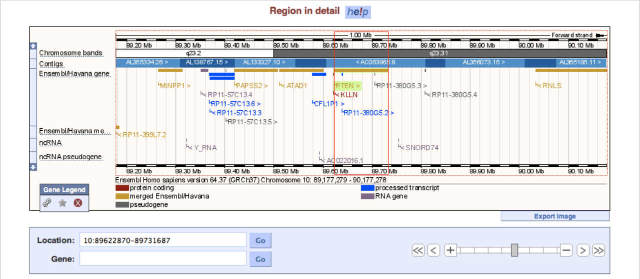


Figure 5. The detail view of the Ensembl genome browser.

The *analysis* view shows a more detailed view of the selected sequence, at the default zoom. The target gene is encoded in terms of *exons* and *introns*, the regions of the gene that code for and do not code for the protein, respectively. The exons are encoded by solid blocks, the size of which dictates the length of the exon within the gene, and the introns are encoded by lines connecting each block. The series of tracks beneath the gene indicate alternative transcripts that also come from that gene. The assembly bar (from the detail view) is in the middle of the transcript display – transcripts appearing above the contig bar are encoded in the forward direction of the chromosome, and those appearing below the assembly bar are in the backwards direction. The colours encode the same type of information as listed above. The information that is presented above the gene indicates matching proteins queried from databases shown on the left-hand side of each track. Below those transcripts are regulatory features that will be discussed in more detail in the following sections.

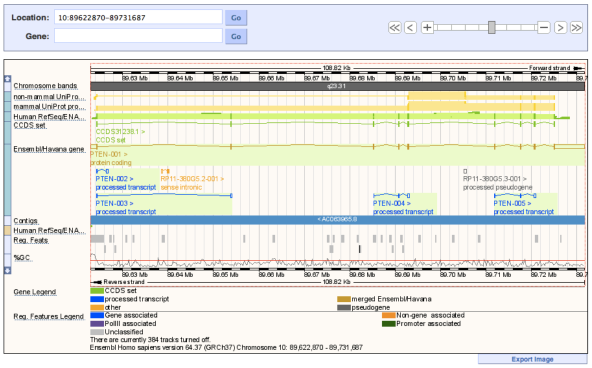


Figure 6. The analysis view of the Ensembl genome browser.

3.3.1 Finding Sequences Up/Downstream

After querying Ensembl's genome browser for the PTEN gene, the second pane of the default view shows the sequence PTEN with relation to surrounding sequences. In Figure 5, the "Region in Detail" view highlights the sequence PTEN and its surrounding sequences. The direction in which the surrounding genes are encoded is given by a glyph ("<" for the "backwards" direction and ">" for the forwards direction). For more information on each of the surrounding genes, a user can click on a gene name to be linked to a page discussing that particular gene.

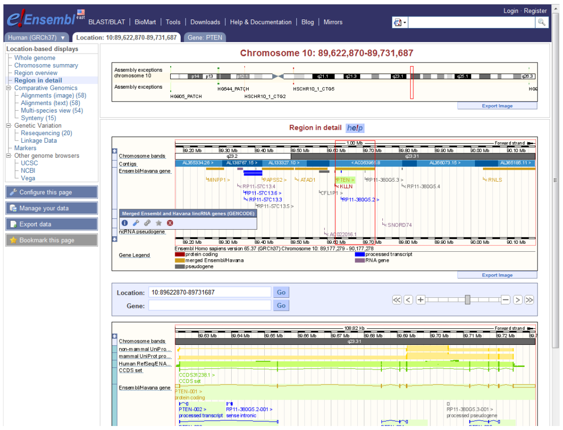
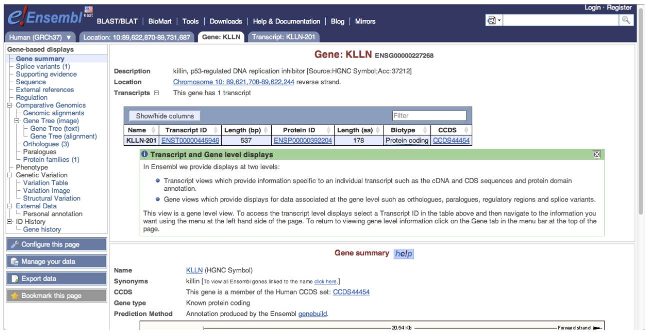


Figure 7. View of Ensembl's Genome Browser after querying for the gene PTEN. The "Region in Detail" view highlights the gene PTEN and shows the surrounding sequences that are encoded in both directions.

Additional information with respect to each surrounding gene is encoded by colour, with a legend given at the bottom. The protein coding gene denoted by the name KLLN, was of particular interest in this study. When that gene was selected, the resulting page gave some very brief details about the protein encoded by this gene (Figure 8). The most interesting thing to note about the KLLN gene is that the protein encoded by it also inhibits cellular replication, which is a major component of the pathway that PTEN also inhibits.



Gene: KLLN ENSG00000227268

Description killin, p53-regulated DNA replication inhibitor [Source:HGNC Symbol;Acc:37212]
Location Chromosome 10: 89,621,708-89,622,244 reverse strand.
Transcripts ⊟ This gene has 1 transcript

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
KLLN-201	ENST00000445946	537	ENSP00000392204	178	Protein coding	CCDS44454

Transcript and Gene level displays

In Ensembl we provide displays at two levels:

- Transcript views which provide information specific to an individual transcript such as the cDNA and CDS sequences and protein domain annotation.
- Gene views which provide displays for data associated at the gene level such as orthologues, paralogues, regulatory regions and splice variants.

This view is a gene level view. To access the transcript level displays select a Transcript ID in the table above and then navigate to the information you want using the menu at the left hand side of the page. To return to viewing gene level information click on the Gene tab in the menu bar at the top of the page.

Gene summary

Name KLLN (HGNC Symbol)
Synonyms killin [To view all Ensembl genes linked to the name click here.]
CCDS This gene is a member of the Human CCDS set: CCDS44454
Gene type Known protein coding
Prediction Method Annotation produced by the Ensembl genebuild.

Figure 8. View of Ensembl's detail page after selecting the protein-coding gene "KLLN".

Most of the other elements, however, were either genes unrelated to PTEN or its functions, or genes with no functional annotation, more than likely found using a large sequencing project. Although they have no other annotation, they may be potentially useful for further study. Since the types of genes are grouped vertically, it was easy to pick out those that coded for RNA, or were just pseudogenes.

3.3.2 Gene Conservation

To find how well the gene is conserved amongst other organisms, alternative tracks must be loaded onto Ensembl's gene browser. To do this, the "Configure This Page" option was selected and appropriate channels were added using the corresponding menus.

In the analysis view, the regions that have to do with conservation are the GERP scores and the sequence alignment regions. GERP scores are a statistical sequence conservation score for a particular element in a sequence (i.e. a nucleotide) [14]. In figure 9, two GERP score tracks are shown, using a bar chart where each bar represents an individual nucleotide in the sequence region. The sequence alignments on a character-by-character basis are shown near the bottom. As all of the sequences with each of the four species are completely identical with the target sequence, only pink lines are shown. In figure 10, the entire alignment is zoomed out to show more of the sequence comparison. When the sequences are not identical, there are gaps in the pink lines.

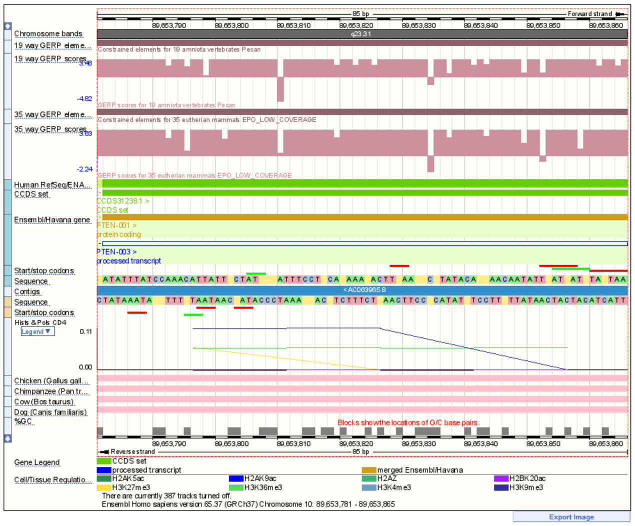


Figure 9. Ensembl's analysis view after the gene conservation tracks were selected.

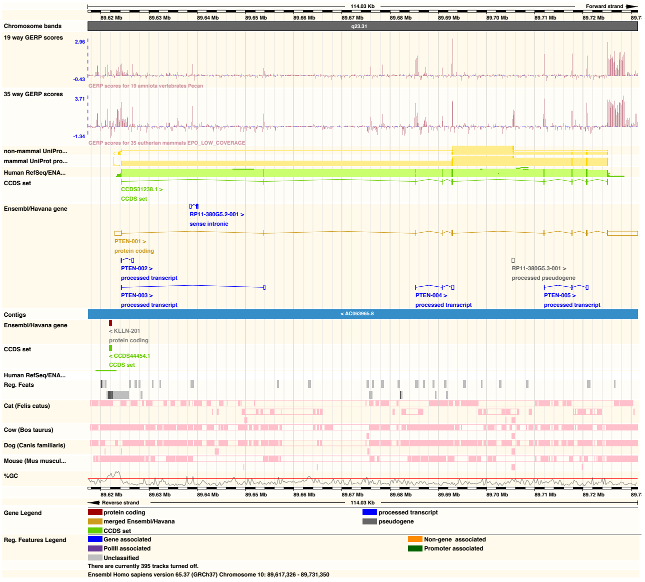


Figure 10. Ensembl's analysis view after the gene conservation tracks were selected, zoomed to show the entire PTEN sequence.

3.3.3 Gene Stability

The stability of the gene region, as measured using Ensembl's browser, was present in both the form of *%GC content* and *simple nucleotide polymorphisms* (SNPs). The %GC measure is a signal measure of consecutive guanine and cytosine nucleotides. Because of the molecular structure of each of these nucleotides, DNA regions containing a higher percentage of them versus other nucleotides are more stable under adverse conditions such as high temperature (the DNA region is less likely to "unzip" unintentionally). *Simple nucleotide polymorphisms* are changes to a single nucleotide in a DNA sequence, whether they be a replacement with an alternative nucleotide, an extra nucleotide being inserted into the DNA strand, or a nucleotide being deleted from the stand. A series of potential SNPs are documented in an external database, *dbSNP* [15], and are used for the generation of the SNP track**.**

Figure 11 shows the analysis view after the gene stability tracks were inserted. The %GC is shown using a line chart measuring the signal. The SNPs are indicated as single lines, one per each nucleotide in the sequence. The colours of each line are explained in the legend below the graphic. To get a better idea of the distribution of SNPs along a particular exon, Figure 12 shows a zoomed in analysis view over one of the exons in PTEN. Since the number of individual nucleotides is much smaller in this view, the individual SNP lines can be distinguished. Small triangles with an associated line show an insertion into the gene sequence.

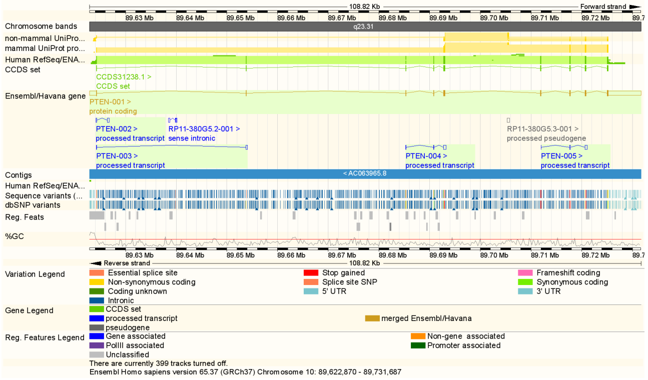


Figure 11. Ensembl's analysis view after the gene stability tracks were selected.

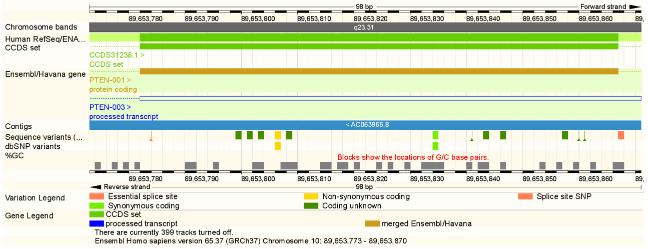


Figure 12. Ensembl's analysis view after the gene stability tracks were selected. Zoomed to view a single exon.

From the analysis of each of the provided views, the gene seems to have an average to below-average %GC, and some interesting polymorphisms that could be used for potential further study. Unclassified polymorphisms can be researched further to determine what happens. The non-synonymous polymorphisms could be used to identify the function of PTEN with that particular mutation, potentially showing a cancerous mutation.

3.3.4 Gene Regulation

For gene regulation, Ensembl's genome browser can show a regulatory features track, and several tracks relating to *polymerase* and *histone* detection. *Polymerases* are the molecules that initiate and perform DNA transcription – detecting polymerase binding regions indicates where transcription will begin. *Histones* are molecules that DNA wraps around in its condensed (non-replicating) form. However, some histones relate directly to gene transcription [16].

In figure 13, the analysis view is shown with all regulation tracks turned on. There was the option of having bar tracks showing the peak only regions of the histone and polymerase detection bars – these are shown in Figure 14. In both figures, the regulatory features track is shown above the histone and polymerase tracks, colour coded according to the regulatory features legend at the bottom of the graphic. The histone and polymerase tracks are shown underneath the regulatory features track. Each individual track represents a particular cell line used for histone and polymerase detection. Since there are several different cell lines to choose between, only 5 tracks were selected to be displayed (each of the other tracks show relatively the same information). Each of the lines in each track represents a specific histone or polymerase, as indicated by the legend revealed when clicking on the "legend" link under the heading of each track.

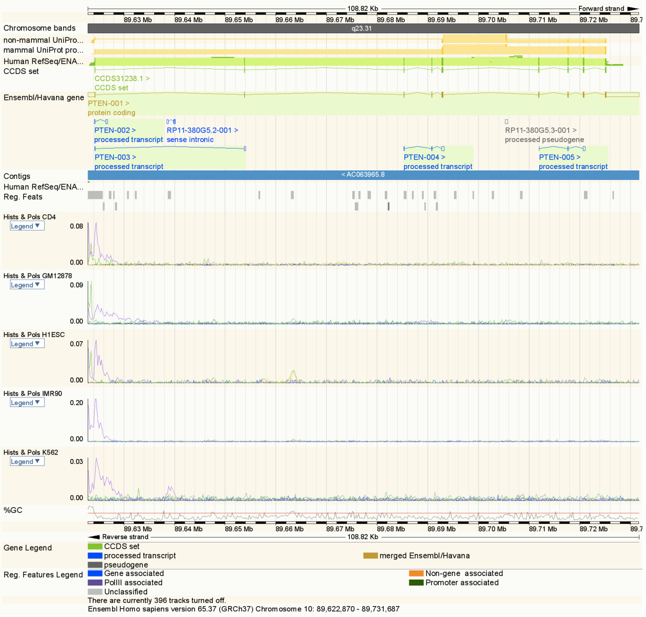


Figure 13. Ensembl's analysis view after the gene regulation tracks were selected.

Unfortunately, the regulatory features track was mostly unclassified in this case. So, no additional regulatory information could be determined using this track of the browser. From the signal view, it seems that only the histone H3K4me3 is present at the beginning of the sequence. However, by looking at the peak-only view, there appears to be some histones that are associated with the starting region of some alternative transcripts of PTEN – this information could be used to determine the regulation/activity of the individual transcripts. With relation to the signal tracks, it appears as if some peaks are undetected by just viewing the signal graph alone, as the peaks in the peak-style graph appear to be within relation to the surrounding levels, as opposed to the overall scale of the tracks.

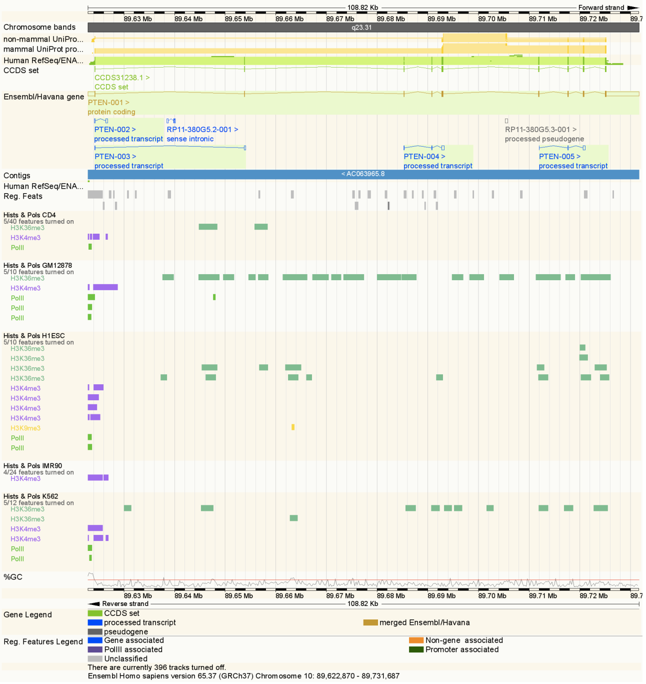


Figure 14. Ensembl's analysis view after the gene regulation tracks were selected, with the peak views instead of signal views for each histone/polymerase detection track.

3.4 UCSC Genome Browser Analysis

Finding the PTEN gene on the UCSC genome browser was similar in comparison to that of the Ensembl genome browser. The resulting views are as follows:

The first view shows an overview-style view of the chromosome, similar to that of the Ensembl genome browser (Figure 15). The region of the detailed view is indicated with a red block. The triangles encoding the centromere of the chromosome are coloured a shade of red (in comparison to the grey from Ensembl).

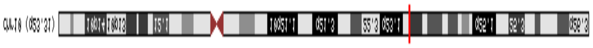


Figure 15. UCSC's genome overview pane.

Underneath the genome overview pane is the detailed pane that shows the individual gene tracks (Figure 16). The default view is centered on the individual gene of interest, like the analysis pane of Ensembl. However, the default tracks of the UCSC genome browser are much different than those of Ensembl. Clicking on a particular track changes the way the track is viewed – by either expanding or collapsing the track to show more/less information. Discussion of these tracks with respect to the analysis topics is in later sections.

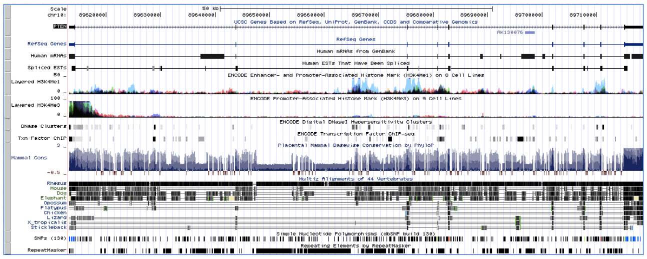


Figure 16. UCSC's detail pane.

3.4.1 Finding Sequences Up/Downstream

In order to view the genes up or downstream of the target gene, the detail view was zoomed out to show more of the surrounding region. The direction of the genes, instead of encoding with individual glyphs, are encoded in the intron regions of the gene in question. Instead of placing the names underneath a set of genes, the genes are spaced out to leave room for each individual gene name. Unlike the Ensembl genome browser, the UCSC genome browser does not colour code the genes based on what type of gene they are. Like the Ensembl genome browser, however, clicking on a particular gene results in a link to more information about that genome page. A subset of the genes from Ensembl are displayed in the detail view, as shown in Figure 17.

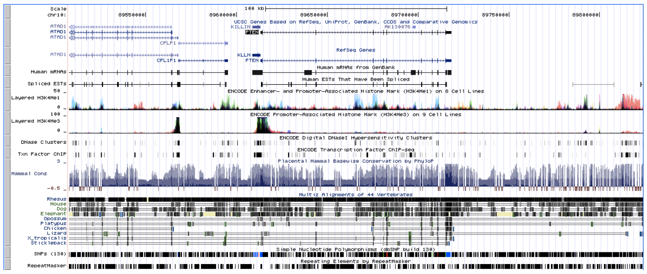


Figure 17. Detail view of UCSC's genome browser, showing flanking genes in relation to PTEN.

By choosing to encode the direction of the gene using glyphs beside each gene name, when it comes to smaller genes, the directional information is lost due to the lack of visible introns. In a higher magnification, however, the direction of transcription of the smaller genes becomes more visible. The gene of interest, KLLN, was present in both views.

3.4.2 Gene Conservation using UCSC

One gene conservation track is already visible in the default view, showing placental conservation amongst many mammals. This is the upper set of tracks underneath the sequence tracks as discussed above, in Figure 18. The first track in this pair shows a nucleotide-by-nucleotide conservation measure called the PhyloP measure [17]. Below the PhyloP measure is a series of tracks representing the nucleotide-by-nucleotide conservation in a variety of vertebrates. The set of tracks underneath represent the PhastCons conservation measure [18]. Underneath those PhastCons tracks show the individual alignments for each gene sequence.

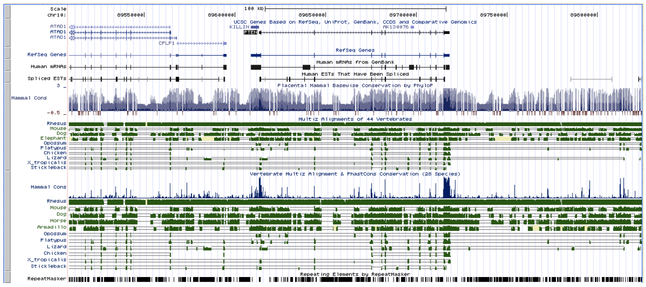

Scale
chr10:
ATAD1
KLLN
PTEN
CFLP1
RefSeq Genes
Human mRNAs
Spliced ESTs
Mammal Cons
Multiz Alignments of 44 Vertebrates
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
RepeatMasker

Figure 18. UCSC's detail pane, showing the conservation/alignment tracks.

From the data analysis point of view, the conserved regions appear to be the specific exons themselves, agreeing with the conserved regions found in the Ensembl browser.

From an information visualization perspective, the view of the score signal intensities is comparatively larger than the tracks found on the Ensembl genome browser, making individual variations easier to define. Additionally, for every codon (set of three nucleotides encoding a distinct amino acid), it is obvious where a match and mismatch are given in the multiple alignments, as shown in Figure 19.

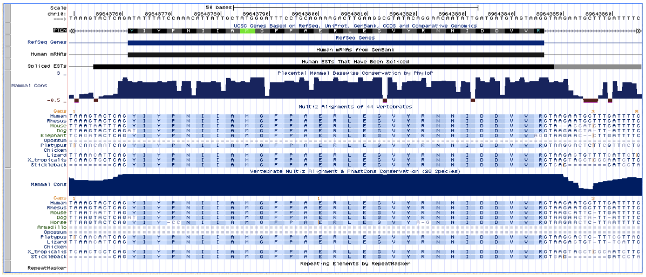

Scale
chr10:
PTEN
RefSeq Genes
Human mRNAs
Spliced ESTs
Mammal Cons
Multiz Alignments of 44 Vertebrates
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
RepeatMasker

Figure 19. View of UCSC's genome browser showing gene conservation tracks, zoomed in to a single exon.

The actual nucleotide-by-nucleotide comparison shows the user exactly what the multiple alignment looks like, instead of pink lines for matching areas and white lines where the sequences do not match. Since the bar charts are relatively the same in each visualization, the increased height of the bars in the UCSC browser allow for more detail to be shown to the user without forcing the user to zoom in, or externally enlarge the image.

3.4.3 Gene Stability using UCSC

The only real stability track in the UCSC genome browser was the SNP track, found by scrolling to the bottom and looking at the series of tracks that have to do with sequence variation. The modified view is shown in figure 20. Because of image size contraints, only the un-expanded ("squished") SNP track is shown. Clicking on this track or modifying its dropdown menu below the visualization will allow the entire track to be visualized by the user. Each SNP is encoded by a single line, and a name associated with that line. Some SNPs were coloured blue, but documentation was not clear as to the reasoning behind the colour choices.

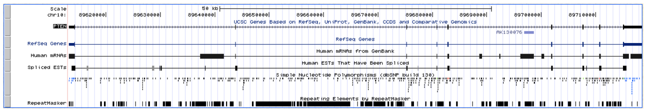

Scale
chr10:
RefSeq Genes
Human mRNAs
Spliced ESTs
RepeatMasker
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
RefSeq Genes
Human mRNAs from GenBank
Human ESTs That Have Been Spliced
Simple Nucleotide Polymorphisms (dbSNP build 130)
Repeating Elements by RepeatMasker

Figure 20. UCSC genome browser with the SNP track enabled.

With the condensed version, the amount of detail that the user can see is relatively minimal, but by expanding the track, the amount of potential SNPs becomes very large, requiring the user to have temporal memory of the structure of the above tracks before scrolling down to the bottom tracks or to look for the target SNP region. However, the "squish" view of this SNP track gives the user a general idea of the frequency of SNPs, which was missing from the Ensembl browser without zooming in to see individual SNPs. The same technique can be applied to see individual SNPs in the UCSC genome browser.

3.4.4 Gene Regulation using UCSC

The default view of the UCSC genome browser highlights some histone-related tracks. To show more regulatory regions, the tracks with respect to chomatin state segmentation and CpG islands are also shown. Chromatin state segmentation is the characterization of the function of the genome molecules, including RNA, DNA, and proteins associated with the sequences, with respect to the transcriptional activity. CpG islands are areas of higher-than average %GC content, and generally are associated with a specific biological promoter. These islands have to do with regulation, in that methylation (the attachment of a methyl (CH_2—) group to a molecule in this region) is actually seen in genes that are currently not being transcribed.

In Figure 21, the tracks for histone detection, chromatin state segmentation, and CpG islands are present. Instead of the line charts as given in the Ensembl genome browser, overlain area charts are instead given, for histone detection. Instead of looking for particular histone marks, each of these tracks looks for specific marks associated with transcription promotion and enhancement. Determining individual levels of histones, however, becomes complicated as the areas begin to overlap one another. To remedy this, expanding the view into the full version (too large to show in this document) shows the individual histone peaks. Promoter-associated peaks are detected as expected, i.e. at the beginning of the sequence.

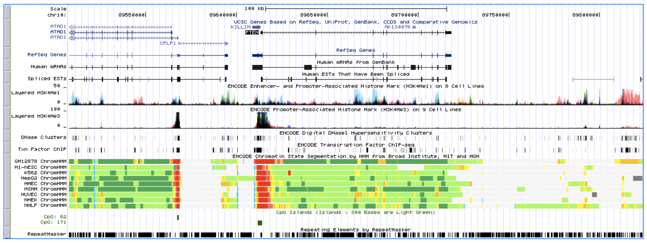

Scale
chr10:
RefSeq Genes
Human mRNAs
Spliced ESTs
Layered H3K4Me1
Layered H3K4Me3
DNase Clusters
Txn Factor ChIP
CpG: 62
CpG: 171
RepeatMasker
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
RefSeq Genes
Human mRNAs from GenBank
Human ESTs That Have Been Spliced
ENCODE Enhancer- and Promoter-Associated Histone Mark (H3K4Me1) on 8 Cell Lines
ENCODE Promoter-Associated Histone Mark (H3K4Me3) on 9 Cell Lines
ENCODE Digital DNaseI Hypersensitivity Clusters
ENCODE Transcription Factor ChIP-seq
ENCODE Chromatin State Segmentation by HMM from Broad Institute, MIT and MGH
CpG Islands (Islands < 300 Bases are Light Green)
Repeating Elements by RepeatMasker

Figure 21. UCSC genome browser, with histone, chromatin state segmentation, and CpG island tracks enabled.

Chromatin state segmentation is indicated by coloured regions – the red regions indicate promoter regions of the sequence, and the green regions indicate elongation regions. Yellow regions indicate regions that can be considered both elongation and promotion regions. Promoter regions promote the initiation of transcription of a particular sequence, while elongation regions promote the continuing transcription of that region. Of the two colour categories, varying the saturation of the colour indicates how "strong" or "weak" that particular region is – i.e, how effective that region is at promoting its intended effect. Lighter saturation means less effectiveness, and heavier saturations indicate strong effectiveness. As expected, the strong promotion regions are indicated at the beginning of each coding sequence (with PTEN's promoter at the left of the sequence, and the promoter of the neighboring coding sequence at the far right). One issue with this encoding strategy, however, is the colour choices – red/green colourblind persons would be unable to use this track, as saturation and luminosity channels are already in use in this viewer.

The CpG Island track shows two specific CpG islands, encoded by green blocks indicating the size of each region. As each of these regions are associated with a promoter region of each sequence, one can assume that one way to control the transcription of a particular sequence would be the methylation of this region. Since no additional colours are present in this track, and due to the sparseness of the actual islands themselves, the encoding choices for the track is logical with respect to the data it represents.

3.5 Comparative Analysis of the Two Browsers

Both of the browsers were effective at solving the analysis questions that were originally discussed. The browsers encoded data in a horizontal fashion, which allowed the side-by-side comparison of different tracks and methods of encoding the data. This section details analysis within the context of information visualization that was not already discussed within the previous sections.

3.5.1 Navigation

The methods of navigation with respect to panning/zooming the views within each visualization differ quite strongly with respect to one another. Because these visualizations are comprised of purely static images, there is a significant delay between the input and the given result.

In the Ensembl genome browser, zooming is performed by brushing a rectangle onto the target image (for the purposes of zooming inwards), or in the image above the target view (for the purposes of zooming outwards), and selecting "jump to here" from the pop-up menu. This allows for coarse-grained zooming and centering based on the user's mouse selection. Iterative selections of the target view can help refine the zoom. An additional way to zoom the analysis view is from a pane situated in between the detail and analysis views. Low-level panning can also be performed using this brushing technique and selecting the "centre here" option from the provided menu. If the user knows exactly which range of base pairs they want to zoom to, they can enter the target in the provided text boxes. Otherwise, for some coarse-grained zooming, the user can press the "+" or "-" buttons on the scale on the right-hand side of the frame.

Zooming using UCSC's genome browser, however, does not involve brushing to select a particular region. There are a series of buttons along the top of the website allowing the user to zoom in/out 1.5x, 3x, and 10x. Additionally, the user can elect to type in some sort of magnification to zoom in/out in the text box beside the zoom buttons. With respect to the Ensembl brush-based navigation system, the UCSC button-based zooming seems counter intuitive.

Clicking and dragging on the USCS track-based interface does not produce any sort of selection box. Instead, it pans the genome right or left based on the horizontal direction that the user is moving the mouse. The titles of each track remain in place, while the individual tracks themselves move along with the cursor. Because the image to the right and left of the original view has not been generated, "incoming" tracks appear non-existent, until the user releases the mouse button, in which time the tracks are computed and displayed.

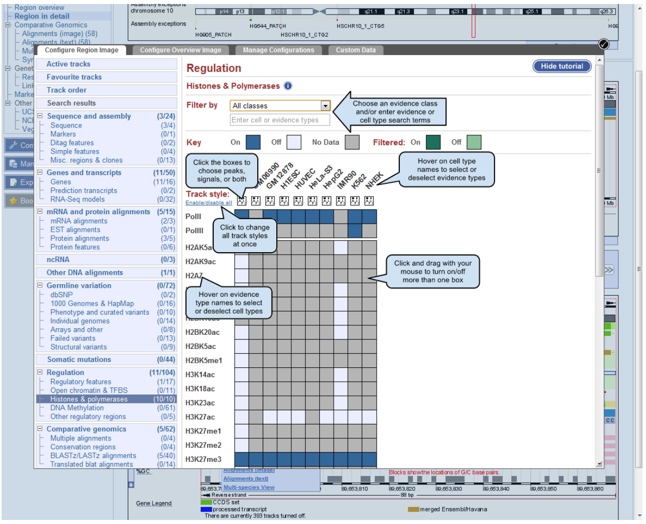
Panning the display in the Ensembl genome browser, as described above, involves, again, either the use of the selection boxes or the specific nucleotide range can be added to the pane between the detail and analysis views. Because of the ability to select regions along any track in the visualization, quickly jumping towards a portion of the chromosome that is far away from the original view can be done if the user knows exactly where they want to select.

However, low-level panning using the selection method appears to be less intuitive than the original panning method, although not all that difficult to learn.

3.5.2 Track Selection

There are two distinct methods that the genome browsers use for track selection for a given view. The first method is in the Ensembl browser, in which a menu is clicked, and each individual track is sorted and categorized according to the type of information it shows. Enabling a particular track involves selecting the type of display given from the pop-up menu that appears upon clicking the empty "checkbox" right beside the track of choice.

Additionally, for tracks with multiple possible views, such as histone and polymerase tracks, an intuitive menu appears allowing the user to select which tracks to show (appearing as cell lines in the track selector for histones and polymerases, as shown in figure 21). Underneath each column, the user can enable and disable particular histone and polymerase tests by selecting or deselecting a white box. Selecting multiple boxes at a time can be performed by clicking and dragging the mouse.



Regulation

Histones & Polymerases

Filter by: All classes

Choose an evidence class and/or enter evidence or cell type search terms

Click the boxes to choose peaks, signals, or both

Hover on cell type names to select or deselect evidence types

Click to change all track styles at once

Click and drag with your mouse to turn on/off more than one box

Hover on evidence type names to select or deselect cell types

Blocks show the locations of G/C base pairs.

There are currently 393 tracks turned off.

Figure 22. The Histones & Polymerases track selection pane from the Ensembl genome browser.

In UCSC's genome browser, selecting additional tracks involves scrolling to the bottom of the page to a series of collapsible lists of drop-down menus allowing the user to hide or show a particular track. Each of the display options is shown in the drop down menu corresponding to each track.

Having the series of collapsible lists at the bottom of the page allows the user to view every possible track at once; in case they do not know which category that track will fall under. However, the sophisticated track selection menus given by the Ensembl browser simplify the track selection process by sorting tracks cogently. Additionally, the initial view of a list of currently selected tracks allows for easy deletion from the viewer screen, as opposed to hunting each individual track down and selecting "hide" in the drop down menu for each present track.

3.5.3 Encoding Choices

The individual encoding choices made in each genome browser must be looked at independently. The following is a series of categories in which the encodings for said categories are discussed individually.

3.5.3.1 Multiple Alignment Encoding

When comparing individual sequences in the genome browser, individual units of the sequences must be encoded properly. When at a high-detail zoom of an exon in the UCSC genome browser, individual codons are encoded in separate boxes if they align with the target sequence in the human chromosome, each box dictating the amino acid that codon translates into. If the amino acids do not match, the corresponding amino acids are shown instead (especially in the case of a "gap" in the alignment).

In the Ensembl genome browser, alignments are encoded by a series of pink bars indicating matches, or the absence of a pink bar indicating a mismatch or deletion, on the nucleotide level. Zooming in to these regions does not produce more detail, as shown in figures 9 and 10. To examine more detailed views, the user must visualize the multiple sequence alignment as provided by clicking the bars, selecting that option, and being redirected to a new page.

Thus, the appropriate selection for encoding would be one that shows the user closer detail as to specific how these species differ – highlighting dissimilarity by the type using a colour channel may be an interesting way to encode alignment tracks. Showing more information than positive/negative match/mismatch encoding would provide greater insight as to how strongly the gene is conserved.

3.5.3.2 Numerical/Measured Data

As seen in figures 13 and 16, the Ensembl and UCSC genome browsers incorporate different methods to show histone detection signals. In figure 13, Ensembl's genome browser incorporates a line plot with a hidden legend to show each signal within the track. However, with the space limitations in a track-based genome browser, some lines may occlude other lines. In figure 16, UCSC's genome browser also shades in the area underneath each curve in the line chart, while allowing other lines to remain visible (by using transparent shading). By adding the layered colours together, darker regions of the track show that multiple peaks are detected in that region more clearly. Instead of occluding other lines, the overall density of histone detection is detected by the darkness of the colour channel.

A few issues with that encoding choice include the limitations associated with potential colour choices once the amount of signal detection lines within that track become large, the number of colours to encode each signal becomes smaller. This will become an issue if the overlay of two or three colours looks like another colour chosen for the encoding.

3.5.3.3 Track Heights

Some of the issues (especially occlusion, as described above) can be solved or minimized if the user could stretch and squish the sizes of each of the tracks as the user required to show more detail. For example, the size of the %GC track in the Ensembl genome browser could not be increased or decreased, thus the ability to see further detail was lost. Additionally, since there was (at a maximum) only four potential track sizes in the UCSC genome browser, the user is limited to the detail of the display. For example, the SNP track as shown in figure 20 could have been made larger to make the details more accessible to the users.

A case can be made for constraining track height, however. By keeping overall track height constrained, the overall size of the resultant image is also limited. This allows for more information to be present in a single screen, thus allowing for easier exportation of the image to other documents. However, the image will not be used if the user does not locate the information necessary that would make the image useful.

4 Conclusion, and Next Steps

The Ensembl and UCSC genome browsers incorporate a large variety of visualization tools to enable the user to analyze their target gene within their genomic contexts. Through the analysis of PTEN, important considerations for future work in genomic browser visualizations were identified:

4.1 Track Height

Consideration of track heights and the ability to adjust these heights for more/less detailed views. Incorporating this ability within a genome browser will allow the user to perform a more detailed analysis (especially in signal detection tracks) for both numerical (signal detection) and boolean (ex. Presence/non-presence of SNPs) data.

4.2 Multi-line Encoding

Viewing multiple “lines” on numerical tracks have occlusion difficulties in small tracks, as well as encoding difficulties in larger tracks.

4.3 Multi-Sequence Alignment

Sequence alignments can benefit by encoding specific types of information more discretely – encoding by colour as opposed to boolean values allows the users to more closely identify types of matching/mismatching regions.

4.4 Navigation Options

More navigation options for genome browsers should be explored – if one type of mouse gesture is used to zoom the visualization, than another type should be used to pan the visualization. A possible suggestion could be to incorporate the pressing of modifier keys (ctrl, shift, etc) to change the mouse gesture to perform one type of navigation over the other. For example, pressing shift should allow the user to draw a selection box by brushing one onto the screen, whereas normal brushing would constitute panning in the visualization.

References

[1] J. Gu, M. Tamura, and K.M. Yamada. Tumor Suppressor PTEN Inhibits Factor-mediated Mitogen-activated Protein (MAP) Kinase Signaling Pathways. *Journal of Cell Biology* (Nov. 30, 1998). **143**(5): 1375-83.

[2] J. Li, C. Yen, D. Liaw, K. Podsypania, S. Bose, S.I. Wang, J. Puc, C. Miliaresis, L. Rodgers, R. McCombie, S.H. Bigner, B.C. Giovanella, M. Ittmann, B. Tycko, H. Hibshoosh, M.H. Wigler, and R. Parsons. PTEN, a Putative Protein Tyrosine Phosphatase Gene Mutated in Human Brain, Breast, and Prostate Cancer. *Science* (28 March 1997), 275(5308): 1943-7.

[3] http://www.annoj.org/index.shtml

[4] http://neomorph.salk.edu/epigenome/epigenome.html

[5] L.D. Stein, C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J.E. Stajich, T.W. Harris, A. Arva, and S. Lewis. The Generic Genome Browser: A Building Block for a Model Organism System. *Genome Res* (2002) **12**: 1599-610.

[6] P. Flicek, M.R. Amode, D. Barrell, K. Beal, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. Kähäri, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, P Larsson, I. Longden, W. McLaren, B. Overduin, B. Pritchard, H.S. Riat, D. Rios, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sobral, G. Spudich, Y. A. Tang, S. Trevanion, J. Vandrovcova, A. J. Vilella, S. White, S. P. Wilder, A. Zadissa, J. Zamora, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernández-Suárez, J. Herrero, T. J. P. Hubbard, A. Parker, G. Proctor, J. Vogel and S. M. J. Searle. Ensembl 2011. *Nucleic Acids Research* (2011). 39 Database Issue: D800-6.

[7] http://www.ncbi.nlm.nih.gov/

[8] J. Goecks, K. Li, D. Clements, and J. Taylor. The Galaxy Track Browser: Transforming the Genome Browser from Visualization Tool to Analysis tool. *Proc. IEEE Symposium on Biological Data Visualization* (2011 Oct): 39-46.

[9] W.J. Kent, et al. The human genome browser at UCSC. *Genome Res. (*2002) **12**: 996–1006

[10] M. Fiume, V. Williams, A. Brook, M. Brudno. Savant: Genome Browser for High Throughput Sequencing Data. *Bioinformatics* (2010 Aug), volume 15, 26(16): 1938-44. Epub 2010 Jun 20 http://compbio.cs.toronto.edu/savant/.

[11] J. Slack, K. Hildebrand, T. Munzner, K. St. John. SequenceJuxtaposer: Fluid Navigation for Large-Scale Sequence Comparison in Context. *Proc. Gernam Conference on Bioinformatics* (2004): 37-42.

[12] T. O’Brien, A. Ritz, B. Raphael, D. Laidlaw. Gremlin: An Interactive Visualization Model for Analyzing Genomic Rearrangements. *IEEE Transactions on Visualization and Computer Graphics* (2010 November) **16**(6): 918-26.

[13] M. Meyer, T. Munzner, and H. Pfister. MizBee: A Multiscale Synteny Browser. *IEEE Trans. Visualization and Computer Graphics* (2009). **15**(6): 897-904.

[14] N. Lopez-Bigas, S. De, S. A. Teichmann. Functional protein divergence in the evolution of Homo Sapiens. *Genome Biology* (2008), **9**(2): R33.

[15] www.ncbi.nlm.nih.gov/projects/SNP/

[16] F. Xu, K. Zhang, M. Gruntein. Acetylation in Histone H3 Globulat Domain Regulates Gene Expression in Yeast. *Cell* (2005), **121**(3): 375-85.

[17] A. Siepel, K.S. Pollard, D. Haussler. New methods for detecting lineage-specific selection. *Proc. 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006)*. 190-205.

[18] A. Siepel, G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M.M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.W. Hillier, S. Richards, et. al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* (2005) **15:**1034-1050.

Glossary

Base – A single molecular unit in a DNA or RNA molecule, i.e. a nucleotide.

Base Pair – Alternative terminology for a single unit, referring to a base and its corresponding base on the alternative strand of DNA (as DNA often exists in pairs).

Centromere – The central region of the chromosome.

Chromatin State Segmentation – The characterization of the function of the genome molecules, including RNA, DNA, and proteins associated with the sequences, with respect to the transcriptional activity.

Codon – A set of three nucleotides encoding a single amino acid (which is a single unit within a protein).

Contigs – Consensus sequences generated from sequencing projects.

CpG Islands – areas of higher-than average %GC content. Generally associated with a specific biological promoter.

Elongation – In DNA transcription, the elongation of DNA transcripts by the continuation of transcription.

Exon – Region of a gene that ultimately is used to code for a product.

Gene Stability – The likelihood of modification/mutation of the gene.

Gene Regulation – Controlling how often a gene is transcribed.

Histone – A large molecule that is part of the DNA structure in its inactive (non-transcribing) state.

Intron – Region of a gene that does not code for the gene product.

Methylation – The addition of a methyl (CH_3) group to a target atom in a molecule.

Mutation – Modifications made to the sequence, such as insertions or deletions of bases/nucleotides, or changes from one nucleotide to another.

Nucleotide – The base molecule in a DNA sequence, it is the single unit.

Polymerase – A molecule that has a direct relation to transcription activity – polymerases are the molecules that perform the actions required for transcription

Polymorphism – The presence of a different nucleotide than the expected one.

Promoter – Specific regions of DNA that promote transcription.

Regulatory Features – Components of a sequence or the surrounding area that affect the frequency of its transcription.

Sequence Conservation – The similarity of one sequence to the same sequence in other (related) species.

Sequence Expression – The expression of the sequence is a measure of how often a sequence is *transcribed*, usually through quantitative measurement of the product formed from transcription.

Track – A section of the track-based genome browser defining

Transcription – the cellular process in which a gene sequence from a DNA strand is read and translated into an RNA molecule that represents that segment of DNA

%GC – The percentage of guanine (G) and cytosine (C) nucleotides in a region of DNA.