Fishing for Genes in the UCSC Browser

A Tutorial

Rachel Harte <hartera@soe.ucsc.edu>

Mark Diekhans <markd@soe.ucsc.edu>

University of California, Santa Cruz, CA

Version 1.0

September 18, 2008

Abstract

This tutorial is aimed at the biologist who is interested in exploring protein-coding genes using the *University of California Santa Cruz (UCSC) Genome Browser*. It is geared towards those who have little or no experience using the *UCSC Genome Browser* and for more advanced users who are not familiar with many of the gene-oriented browser features. Using the example of a human gene, *PPP1R1B*, the reader is guided through a step-by-step process for finding and visualizing protein-coding genes in the context of the human genome and a wide variety of genomic data. The user is shown how to use the *UCSC Genome Browser* to locate a *Mammalian Gene Collection (MGC)* clone of the gene and how to order the clone from suppliers.

1 Accessing the *UCSC Genome Browser*

The *UCSC Genome Bioinformatics* website consists of a suite of tools for the viewing and mining of genomic data. The *UCSC Genome Browser* [16, 18] facilitates the viewing of clones from the *Mammalian Gene Collection (MGC)*[35, 37] in the context of other genome annotations. Various types of annotations such as *MGC Genes* are visualized in **tracks** which are graphical representations of data displayed in a genomic context on the *Genome Browser*. To view these annotation tracks, first go to the home page at http://genome.ucsc.edu (Figure 1).

The blue menu bars at the top and the left side of the *Genome Browser* home page both include following links to various tools (Figure 1):

- **Genomes** on the top blue bar or the *Genome Browser* link on the side blue menu bar allow entry to the **Gateway** page for the *Genome Browser*. From here, users may select the genomes that they wish to browse. An additional route of entry is via a **Photo Gateway** page in the *UCSC genomewiki* with photographs of each species whose genome is represented in the *UCSC Genome Browser*.
- **Blat** is a fast alignment tool[20] which allows the user to align DNA or protein sequences to the genome assemblies.
- **Tables** on the top blue bar or **Table Browser**[17] on the side blue bar link to an interface allowing the retrieval of data associated with tracks from the databases underlying the *Genome Browser*. Data added as additional tracks created by the user (**Custom Tracks**) may also be queried with this tool.
- **Gene Sorter** is a tool that displays a sorted table of genes that are related by some metric selected by the user e.g. similar expression patterns, protein homology or proximity in the genome[19].
- **Genome Graphs** on the side blue bar is a tool to facilitate viewing of whole genome datasets such as genome-wide SNP association studies, linkage studies and homozygosity mapping. Instructions are in the **Genome Graphs User's Guide**
- **PCR** on the top blue bar or **In Silico PCR** on the side blue bar is a fast method of searching for a pair of PCR primer sequences in a genome assembly.
- **VisiGene** is a virtual microscope for viewing *in situ* hybridization images.
- **Proteome** on the top blue bar and **Proteome Browser** on the side blue bar lead to the gateway of a tool for viewing proteins and their properties[14]. Both graphical images and links to external websites provide a rich source of protein information.
- **Utilities** on the side blue bar allows access to various tools to remove non-sequence-related characters from DNA or protein, for creating a gif image for a phylogenetic tree and a tool which converts genome coordinates between assemblies (**liftOver**).
- **Downloads** on the side blue bar allows bulk download of data including dumps from the *Genome Browser* databases.
- **Custom Tracks** is a powerful tool allowing users to add their own data for viewing and querying within the context of a genome and its associated annotation data. A User's Guide provides help for creating **Custom Tracks**.

2 Searching for a gene

Let's suppose that a user wishes to study the human *PPP1R1B* (protein phosphatase 1, regulatory (inhibitor) subunit 1B) gene, which is expressed in the brain. The protein encoded by this gene is also known as DARPP32. Its phosphorylation status is regulated by dopaminergic and glutamatergic (NMDA) receptors. Once

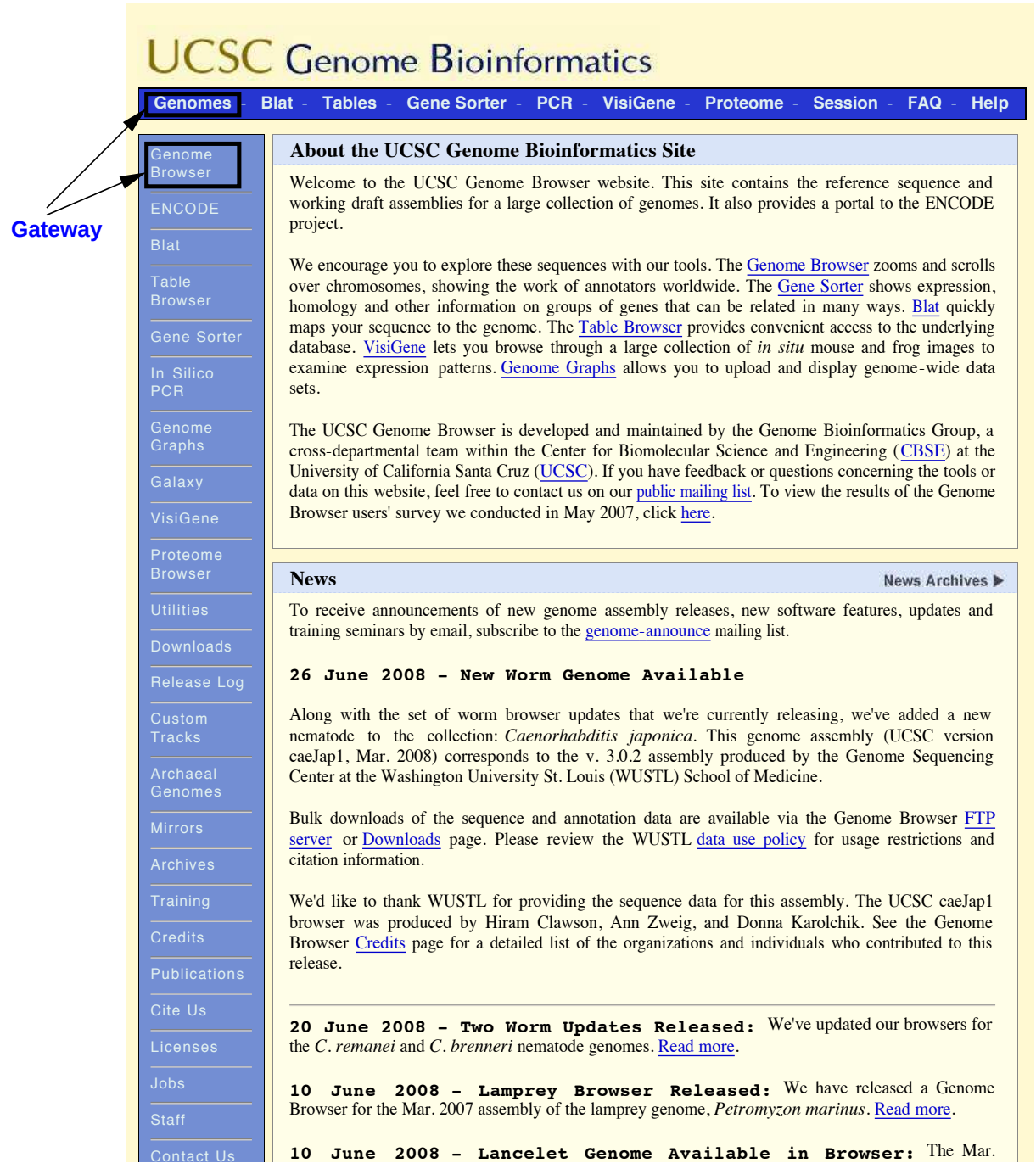


Figure 1: *UCSC Genome Browser* home page. On this page, there is an introduction to the web site and postings of news of training seminars and recent additions such as new assemblies and software features. The blue menu bars at the top and left side of the page allow access to the *Genome Browser* for genome assemblies of a variety of organisms, data mining tools and help pages.

phosphorylated, *PPP1R1B*, is a potent protein phosphatase-1 inhibitor. To visualize this gene in the *UCSC Genome Browser* in the context of various genomic annotations, the user may take the following steps:

1. To get started, clicking on either the **Genomes** or **Genome Browser** link will take the user to the *Gateway* page where the clade, genome and assembly of interest may be selected from pull-down lists of multiple organisms and genome assemblies.

2. Below this area, there is a section describing the selected assembly which also indicates that the default assembly (at the time of writing) is known as hg18 on the *UCSC Genome Browser* website. This assembly is also known as NCBI Build 36.1. This section also contains a list of **Sample position queries**. These are a selection of queries that may be entered into the **position or search term** box adjacent to the genome and assembly controls. Examples include gene name, mRNA or EST accession and descriptive term. Such terms could be used if the gene is not a known gene with an official gene symbol. For the gene in question, the alternate name, *DARPP32*, could be used or a descriptive term such as **NMDA** which is less selective and returns a larger number of results.

3. Enter the gene symbol, *PPP1R1B* into the **position or search term** box (Figure 2).

   Clicking on the **submit** button directs the user to a page displaying the search results. For each track where there are data items containing *PPP1R1B* as their identifier or in their description, results are presented as links to the genomic positions of these items. (Figure 3).

4. Next, click on one of the links, e.g., the *RefSeq Genes* **PPP1R1B at chr17:35036705-35046403** link. This will take you to the *PPP1R1B* locus in the human genome. Note that there are two RefSeq splice variants listed for this locus and this one is the longer transcript variant.

3 *Genome Browser* annotation tracks

The *Genome Browser* displays certain annotation tracks by default in the main *Browser* image. The current default display (Figure 4) shows a number of gene tracks:

- *UCSC Gene* predictions[15, 18]
- *BLAT* alignments of sequences from GenBank[2]
- *BLAT* alignments of full-length ORF *MGC Genes*.

Other visible tracks in this default display are:

- *Vertebrate Conservation*[29]
- Simple Nucleotide Polymorphisms from NCBI Build 128 of dbSNP[33] (*SNPs (128)*) track[38]
- Location of repeats found by *RepeatMasker*

Gene structure is shown in these tracks with filled blocks representing exons; thick blocks are in the coding sequence (CDS) and thin blocks represent untranslated regions (UTRs). The lines connecting the blocks are introns. The direction of arrowheads on the lines or the blocks show the strand on which the element resides. Right-facing arrows show that it is on the sense strand while left-facing arrows show that it is on the antisense strand. To change the sense of the strand in order to more conveniently view annotation on the antisense strand, the *Genome Browser* display may be reversed using the **reverse** button underneath the *Genome Browser*

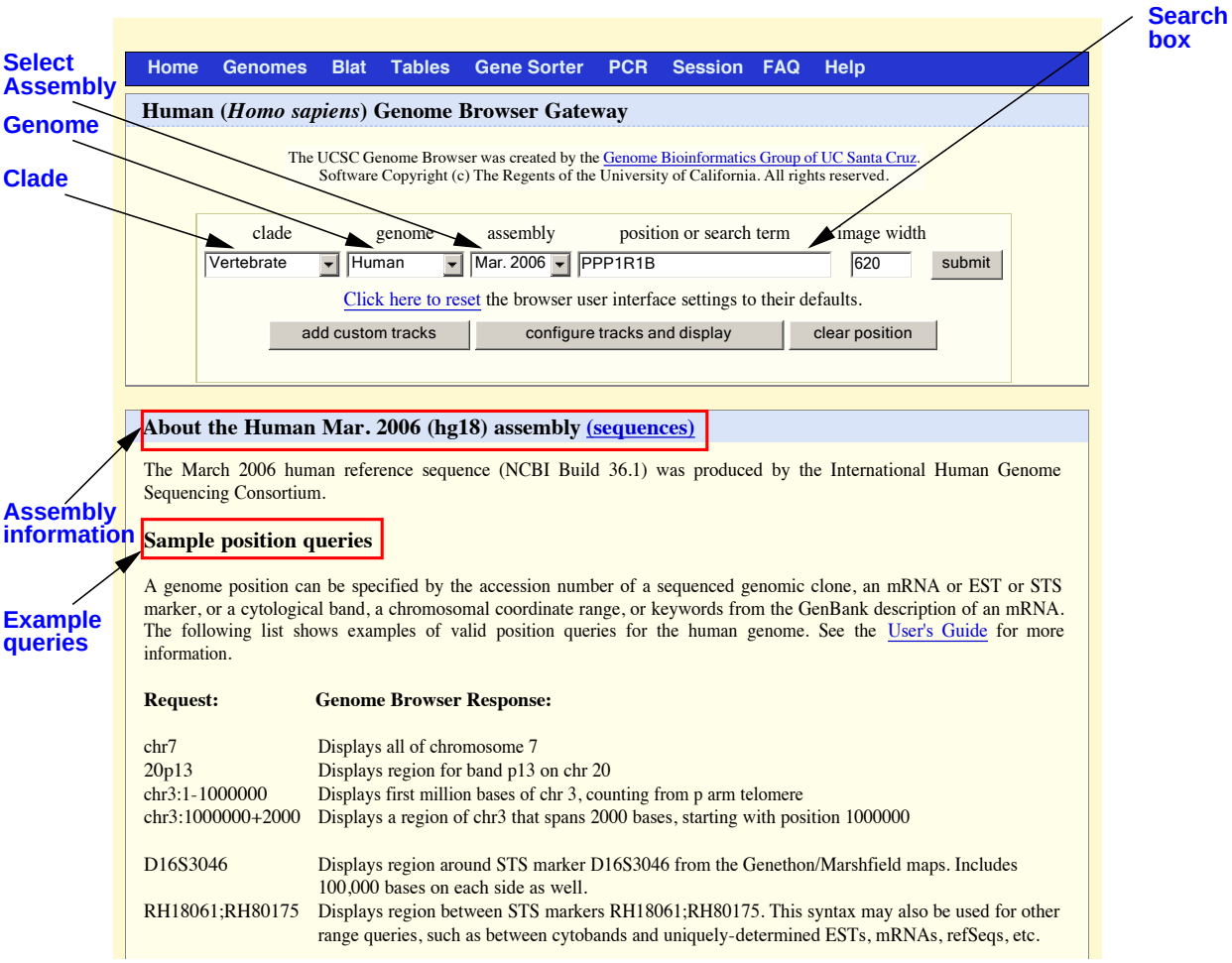


Figure 2: *UCSC Genome Browser Gateway* page. The **position or search term** box allows the user to search for a position within the selected genome assembly or by keyword, gene symbol or other identifier. Here, a search for *PPP1R1B* is being initiated.

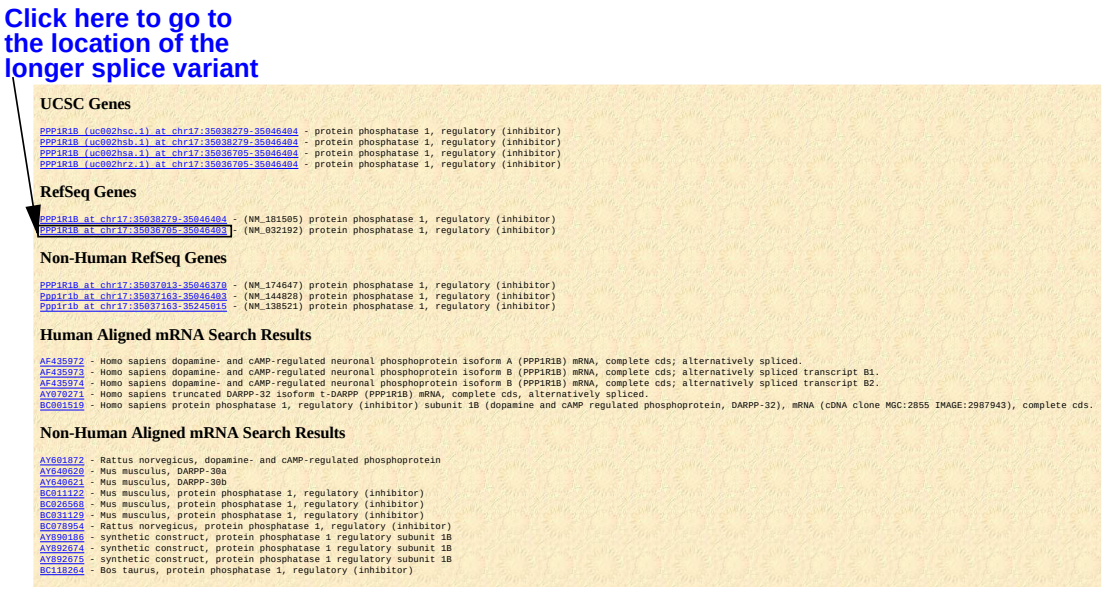


Figure 3: Results of a search for *PPP1R1B*. The results page shows all the tracks that contain data whose annotation includes the keyword **"PPP1R1B"**. This gene is found in the *UCSC Genes*, *RefSeq Genes* and *Human mRNA* tracks as well as in the tracks for *Non-human RefSeq Genes* and *Non-human mRNAs*.

image (see Figure 4. The image will then be redrawn so that the $5'$ to $3'$ direction of transcription of the antisense gene is from left to right, which is more intuitive.

Above the *Genome Browser* image for the human genome assemblies (and for other organisms when available), there is a chromosome ideogram[13] showing a red box which indicates the position of the current viewing window within the chromosome. Navigation controls are found above and below the *Browser* image (Figure 4). Scrolling down below the *Browser* graphic allows access to the visibility controls for each track grouped by track type (see the lower part of Figure 4). These controls allow the user to select various tracks for display; display modes can be altered using the pull-down lists. Visibility options are:

- **hide** which renders the track invisible
- **dense** which collapses all the features into a single line
- **squish** which displays each features a separate line, but at 50% of the height of full mode and without labels
- **pack** which displays several features on each line with labels
- **full** which shows each feature on a separate line with labels

Such compacting of tracks is particularly useful for those tracks with large amounts of data. By altering the visibility of a number of tracks, a display such as that in Figure 5 can be achieved.

Zooming in to base level in order to examine annotations more closely can be achieved by using the navigation buttons above the *Browser* image. Clicking on the base numbers in the *Base Position* track also allows zooming. At the base level, the genome bases can be viewed below the numbering in the *Base Position* track (Figure 6. The one-letter amino acid codes for the translation of the codon triplets in all three frames for the strand being viewed will come into view as one zooms to base level (if the *Base Position* track visibility is **full**) while codons in various

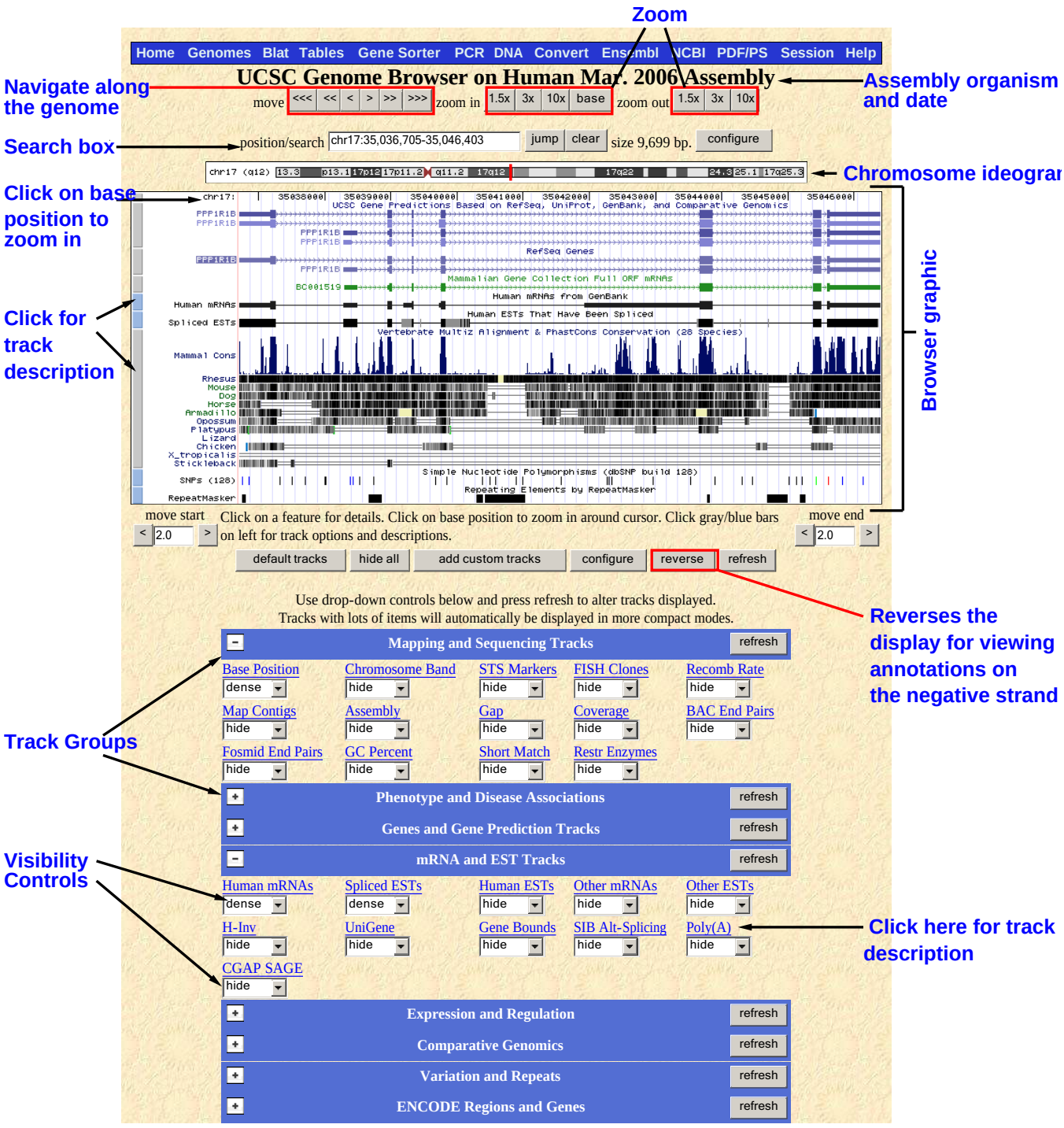


Figure 4: Default tracks for the human hg18 (NCBI Build 36.1) assembly at the *PPP1R1B* gene locus.

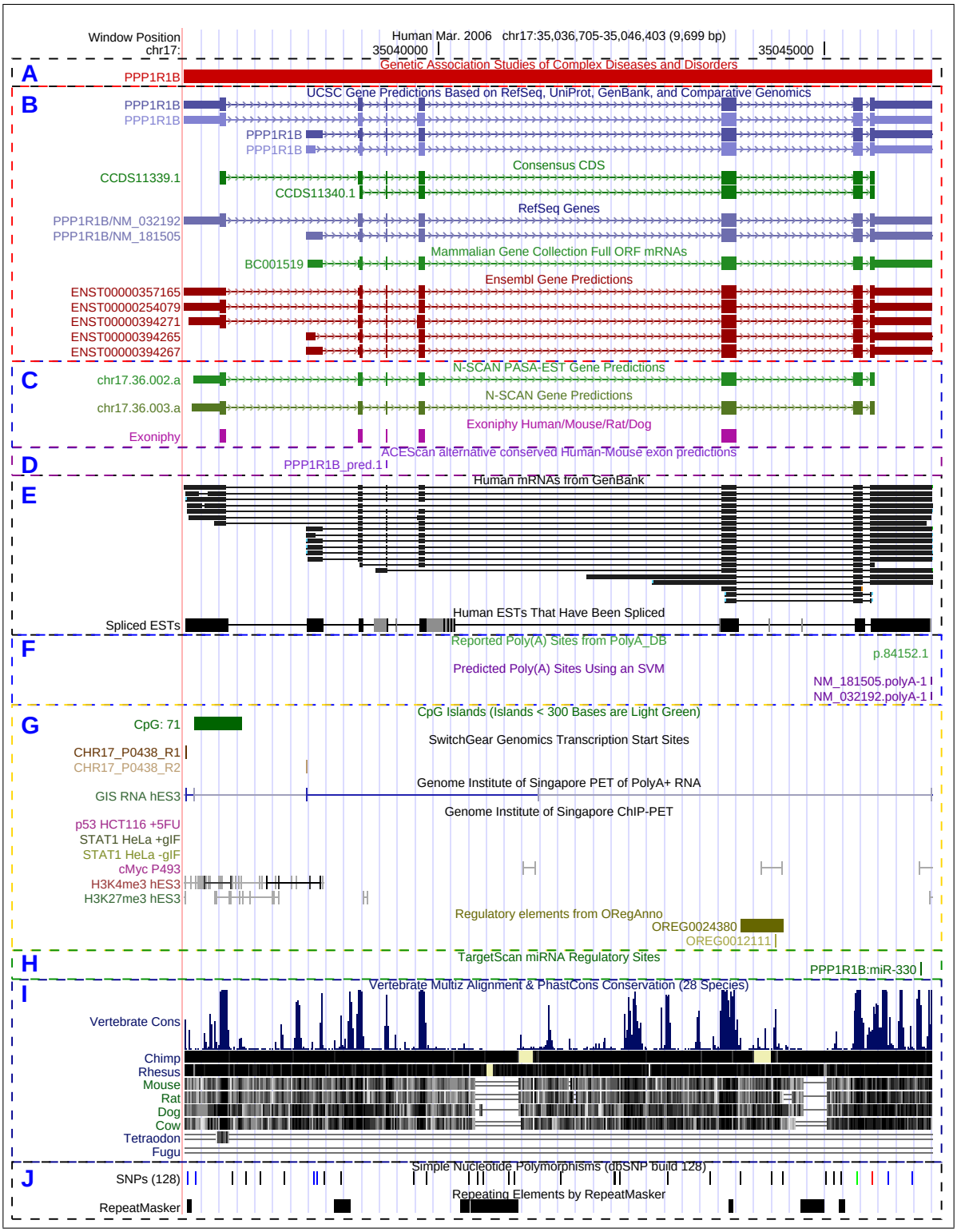


Figure 5: Human hg18 *Genome Browser* showing multiple annotations for the *PPP1R1B* gene locus. Using the track controls below the *Browser* image, the visibility has been set to display a variety of annotation types in the following tracks: (A) *Genetic Association Database (GAD) View*; (B) *UCSC*, *CCDS*, *RefSeq*, *MGC* and *Ensembl Genes*; (C) *N-SCAN* predictions and *ExoniPhy*; (D) *ACEScan*; (E) *Human mRNAs* and *Spliced ESTs*; (F) *Poly(A)*; (G) *CpG Islands*, *SwitchGear Transcription Start Site* predictions, *GIS-PET* and *OregAnno*; (H) *TargetScan*; (I) *Conservation*; and (J) *SNPs* and *RepeatMasker* repeats.

gene annotation tracks are also labeled with the one-letter amino acid code (Figure 6).

4 Annotation details

Each track has an associated **description page** which can be reached either by clicking on the hyperlinked name above the appropriate track control below this image or via the blue/gray bar at the side of the track. The track description details the methods and data sources used to produce the track, validation, credits and citations of relevant publications. Above the description, many tracks have configuration controls specific to the track type and some tracks also have filters. These controls allow the user to create the display that best shows the data of interest. Each item in a track also has a **details page** which may be reached by clicking on an item of interest in a track. Instead of configuration controls in the top section of the page, there are further details about the track item. These details can be extensive and may include:

- Links to external databases
- Confidence scores where applicable
- Position information
- Links to sequence
- Links to alignment information for aligned sequences,e.g., GenBank sequences.

Annotation data can be noisy, so care must be taken when using it to make interpretations. For instance, predictions can contain false positives and experimental data can be erroneous. For this reason, it is important to evaluate data from different sources in order to make an informed judgment as to the confidence that annotations should be assigned. With this in mind, examples from the different track groups will be compared for the *PPP1R1B* locus.

5 *Phenotype and Disease Associations* tracks

The *Genetic Association Database (GAD) View* track[1] has a red rectangle spanning the *PPP1R1B* locus, indicating that this gene has been associated with a disease or condition (Figure 5, Box A). The details page for this item shows that the associated condition is nicotine dependence in certain individuals. Links are provided to both the *GAD* database and to the single publication documenting evidence for the association.

6 *Gene and Gene Prediction* tracks

6.1 *UCSC Genes*

The *UCSC Genes*[15, 18] track (Figure 5, Box B) consists of gene models based on data from GenBank, RefSeq and UniProt, each of which has a details page that is extremely rich in information about that gene including:

- Gene descriptions
- Database cross-reference links
- Alternate gene names
- Genetic Association data
- Chemicals interacting with the gene product
- Selected microarray data
- mRNA UTR secondary structure
- Protein domains and structure
- Orthologs from other model organisms
- Gene Ontology (GO) annotations
- Pathways involving the gene product
- Information on the gene model

The *UCSC Genes* set has four splice variants at the *PPP1R1B* locus. Two predicted transcripts encode longer proteins and differ in the

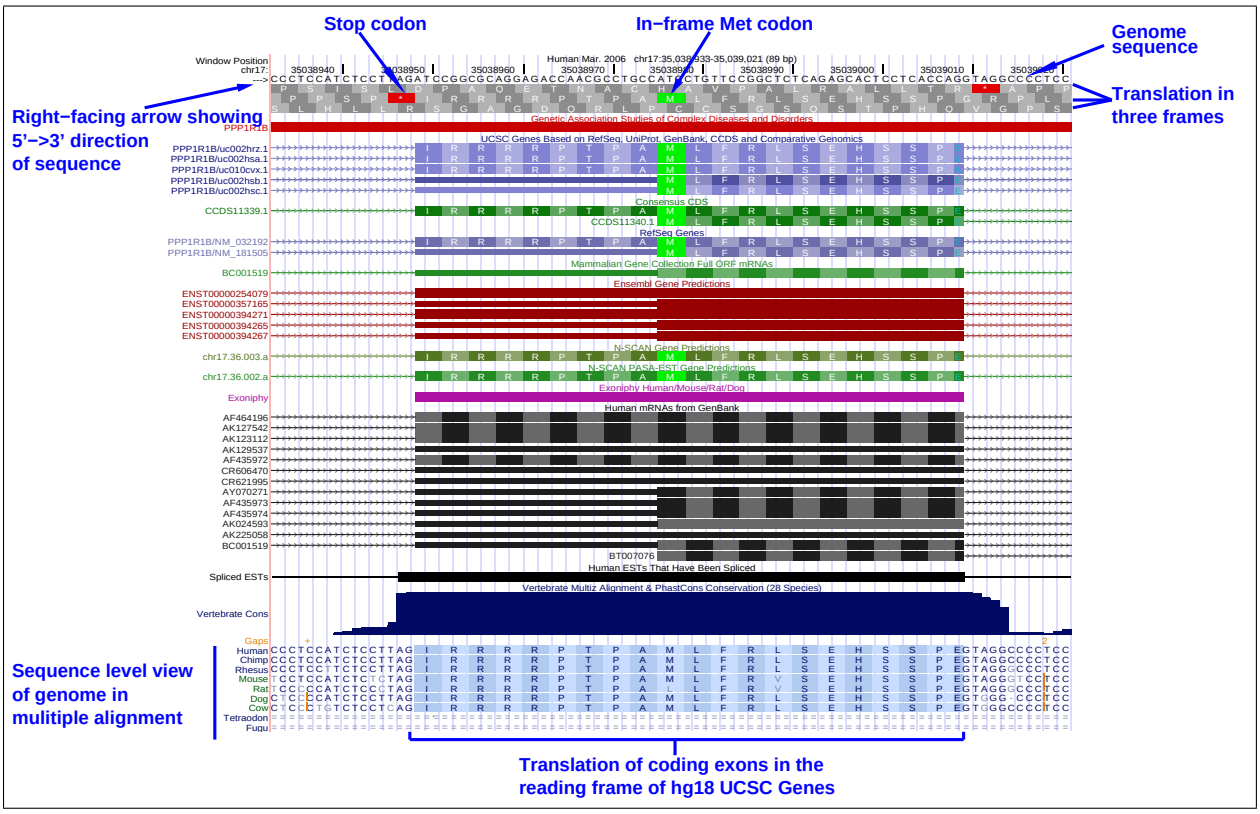


Figure 6: Multiple annotation tracks at the human hg18 *PPP1R1B* gene locus. At this zoom level, both the genome sequence and the amino acid translation in three frames are displayed in the *Base Position* track at full visibility. On the left side of this track, a right-facing arrow shows that the sequence runs from $5'$ to $3'$. Clicking on the arrow reverse-complements the sequence so that it reads from right to left ($3'$ to $5^{p}rime$). Various gene annotation and alignment tracks show color-coding at this zoom level; in the *mRNA* and *MGC Genes* tracks, by default, the codons are shown as alternating light and dark colored rectangles. The *UCSC Genes*, *CCDS* and *N-SCAN* tracks and the multiple alignment section of the *Conservation* track display genomic codons, by default, with the one-letter amino acids code labeling each codon in the CDS region of each transcript. *RefSeq Genes* has this feature switched off by default, but here, it is switched on together with a label for each item consisting of its gene name and RefSeq accession. At this base-level view, ATG are codons colored light green and labeled with "M" to represent potential translation start methionine codons while stop codons are colored red and are labeled with an asterisk "*". The *Conservation* track has been configured to show a subset of species alignments and the Vertebrate conservation scores. These features can be configured using the controls at the top of the description page reached by clicking on the blue/gray mini-button at the left side of the relevant track.

size of an internal exon. The other two transcripts that encode a short protein differ in the length of the 5′ UTR (Figure 5, Box B). Clicking on the first transcript (UCSC Gene **uc002hrz.1**) displays the details page which is divided into a number of collapsible sections. Selected sections are shown in the *UCSC Genes* details page figures: Figure 7 shows sections including those with gene descriptions and alternate gene names, Figure 8 shows expression data and pathways involving the gene product and Figure 9 shows a section of information about the gene model for *UCSC Gene* **uc002hrz.1**.

6.2 *TransMap* cross-species alignments

The *TransMap*[35, 36, 44] track contains cross-species alignments of GenBank[2] *mRNAs* and *Spliced ESTs*, *RefSeq Genes* and *UCSC Genes* to the genome. The more sensitive *BLASTZ* [32] alignment is used to create a base level projection of transcript alignments from different species onto a genome in order to predict orthologous genes on that genome. Mouse, human and many other vertebrate assemblies have a *TransMap* annotation track. For each genome, the vertebrate assemblies with *BLASTZ* alignments were selected. For closer evolutionary distances, the *BLASTZ* alignment nets[22] are syntenically filtered to distinguish orthologs from paralogs; for more distant species or if synteny is difficult to determine, all *BLASTZ* chains[22] are used. In this way, more genes can be mapped but with the complication that some genes are mapped to paralogous regions. Post-alignment filtering can remove some of the duplications. It is this set of chains that are used to create a base-level projection of the transcript alignments to the genome. The resulting pairwise alignments are shown in Figure 10.

6.3 Other *Gene and Gene Prediction* tracks

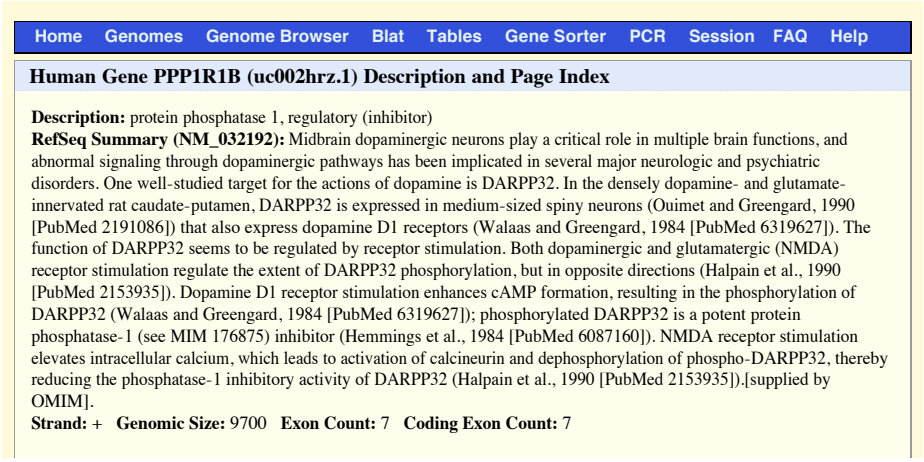
The *RefSeq*[31] transcripts are medium blue in color signifying their **Provisional** status. At this status level, a *RefSeq* is represented by a single GenBank source sequence and has not yet undergone a full review by annotators. During graduation to **Reviewed** RefSeq status, additional sequence data may be used to modify and extend the transcript structure and additional biological annotations may be added. However, users can make their own judgment by evaluating the additional evidence for a gene structure using the other annotation data.

The *Consensus Coding Sequence (CCDS)* track (Figure 5, Box B) is a high quality set of annotations consisting of a core set of protein-coding regions produced as a collaboration among NCBI, UCSC and the Havana and Ensembl groups at the Wellcome Trust Sanger Institute (WTSI) and the European Bioinformatics Institute (EBI). CCDS represent a consensus between RefSeq annotations (NCBI) and the Ensembl and Havana annotations (EBI/WTSI). At this locus, there are *CCDS* representing both splice variants in the *RefSeq Genes* track (NM_032192 and NM_181505) and also, all the predicted *UCSC Genes*. The two CCDS (**CCDS11339.1** and **CCDS11340.1**) represent all the coding regions that are shown in the *Gene and Gene Prediction* group annotation tracks in Figure 5, Box B.

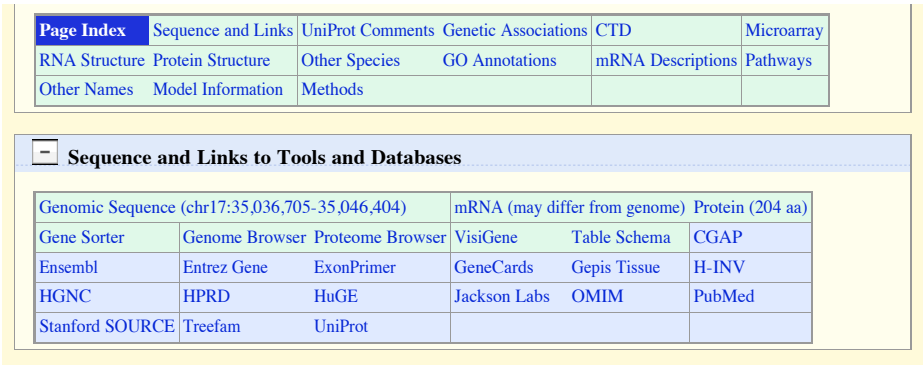
The *MGC Genes* full-length ORF track has one MGC clone (**BC001519**) (Figure 5, Box B). MGC clone sequences have been submitted to the GenBank database so the same mRNA also appears in the *Human mRNA* track. The *mRNA* and *EST* tracks contain *BLAT* alignments of additional transcripts (see section 7 and Figure 5, Box E).

The *Gene and Gene Predictions* group of tracks also includes gene predictions based on mRNAs and ESTs such as *Ensembl Genes* [9] (Figure 5, Box B), *de novo* gene prediction

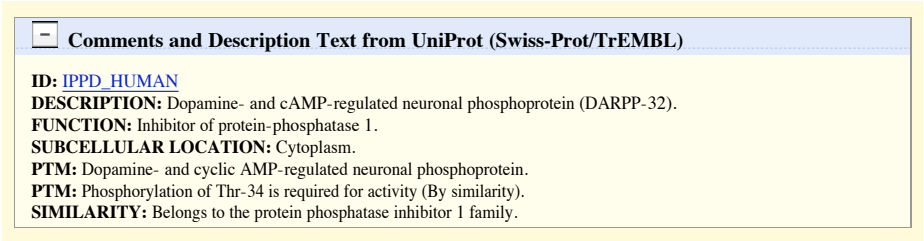
(a) Description for *UCSC Gene* **uc002hrz.1**



(b) Internal and external links for *PPP1R1B*



(c) UniProt Description for *PPP1R1B*



(d) Alternate names for *PPP1R1B*

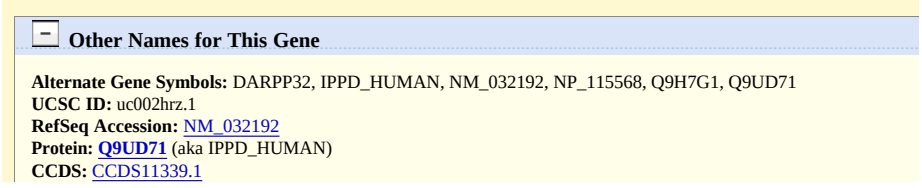
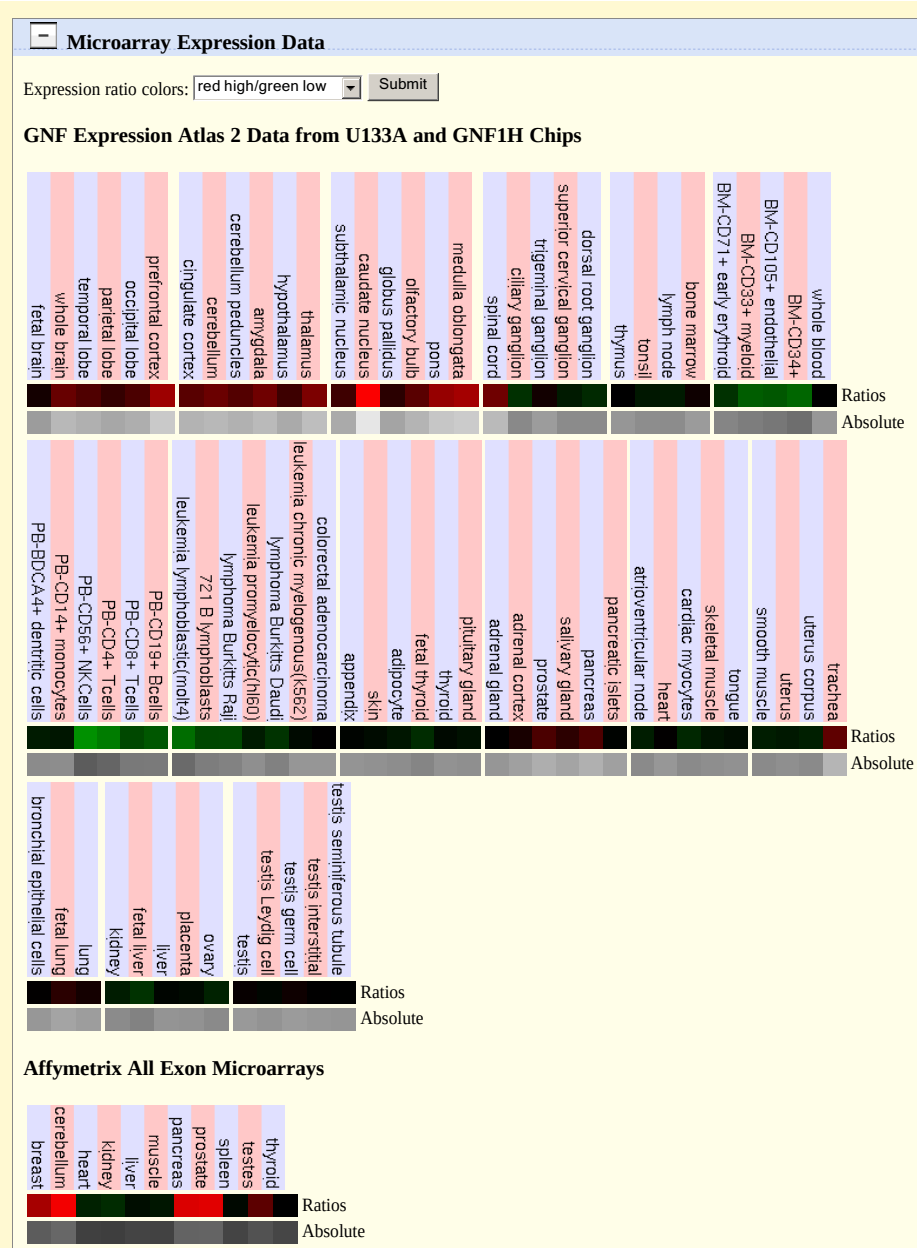


Figure 7: Description, links and gene names from the details page for the human gene *PPP1R1B* (**uc002hrz.1**) in the *UCSC Genes* set. (a) shows the *RefSeq Genes* description for the gene. (b) shows links to the sections on the details page. The next section contains links to various tools and databases both at UCSC and externally. (c) shows a section containing the UniProt/SwissProt database description for the protein encoded by this gene. The UniProt ID (IPPD_HUMAN) links to the UniProt database entry for a protein isoform produced by this gene. The UniProt record also has information about other protein isoforms represented in the database. (d) shows a section of alternate names for the *PPP1R1B* gene, some of which have links to external databases.

(a) Microarray data for *PPP1R1B*



(b) Pathways involving *PPP1R1B*

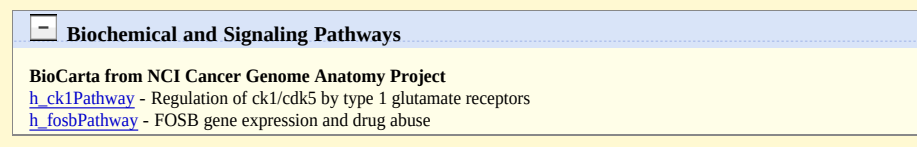
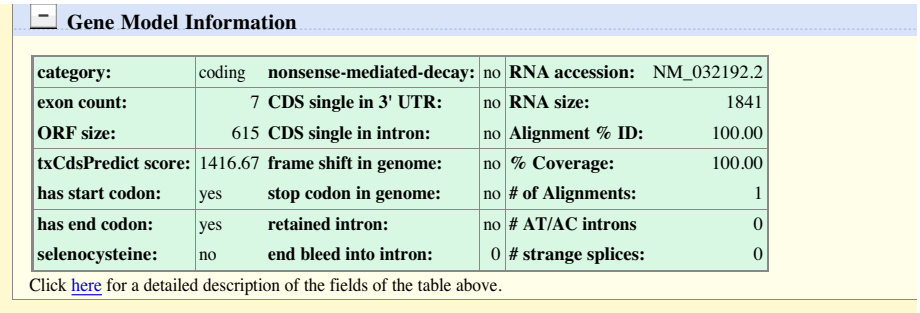


Figure 8: Expression data and pathways sections from the details page for the human gene *PPP1R1B* (**uc002hrz.1**) in the *UCSC Genes* set. (a) shows heatmap displays of microarray expression data for *PPP1R1B* in a variety of tissues and cell lines. The upper dataset is from the Genomics Institute of the Novartis Research Foundation (GNF) (http://symatlas.gnf.org) using Affymetrix chips. The lower dataset was provided by Affymetrix (http://www.affymetrix.com) and it was produced using Affymetrix Human Exon 1.0 ST arrays. (b) is the pathways section which has links to the BioCarta pathways involving the *PPP1R1B* gene product.



Gene Model Information

category:	coding	**nonsense-mediated-decay:**	no	**RNA accession:**	NM_032192.2
exon count:	7	**CDS single in 3' UTR:**	no	**RNA size:**	1841
ORF size:	615	**CDS single in intron:**	no	**Alignment % ID:**	100.00
txCdsPredict score:	1416.67	**frame shift in genome:**	no	**% Coverage:**	100.00
has start codon:	yes	**stop codon in genome:**	no	**# of Alignments:**	1
has end codon:	yes	**retained intron:**	no	**# AT/AC introns**	0
selenocysteine:	no	**end bleed into intron:**	0	**# strange splices:**	0

Click here for a detailed description of the fields of the table above.

Figure 9: Gene model information section from the *UCSC Genes* details page for the human gene *PPP1R1B* (**uc002hrz.1**). This section shows that **uc002hrz.1** is protein-coding, has seven exons and has an open reading frame (ORF) of 615 bp among other features.

tracks such as *N-SCAN*, and exon prediction tracks such as *ExoniPhy* (Figure 5, Box C). EXONIPHY uses conservation among human, mouse, rat and dog to identify putative protein-coding exons. Computational gene predictions can provide validation for the gene and transcript structures produced by manual curation efforts such as those in the *RefSeq Genes* and *Vega Genes*[41] tracks.

7 GenBank mRNA and EST sequence alignments

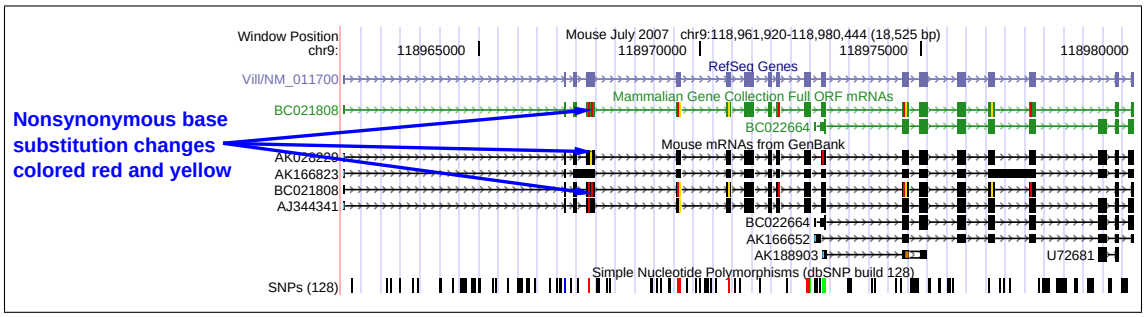
The *mRNA and EST Tracks* group (Figure 5, Box E) show sequences from GenBank that align well to the genome using *BLAT*. In Figure 5, these tracks are compacted; the *Human mRNAs* track is shown with the visibility set to squish and the *Spliced ESTs* track is in dense mode (see the paragraph on track visibility in section 3). The transcript sequence alignments are regularly updated to keep in synchrony with GenBank; *mRNA* and *RefSeq Genes* updates are daily and *ESTs* tracks are updated weekly. The mRNA and EST transcripts may suggest the existence of additional exons and therefore additional transcript variants than those found in the gene and gene prediction tracks. The *Spliced ESTs* track shows that there are ESTs at the *PPP1R1B* locus that have two additional small exons towards the $3'$ end of the gene (Figure 5). ESTs can therefore be used to predict the existence of additional splice forms not represented by mRNAs. Sequence data can be noisy; in particular, ESTs tend to have low sequence quality and were generated by single-pass sequencing. Therefore, care should be taken in interpreting these data. One caveat is that differences between transcripts and the reference genome may not be noise, but simply genetic variation between individuals. The *SNPs* track can indicate such instances (Figure 5, Box J and Figure 11). The *SNPs* track may display no known SNPs from dbSNP[33] at the position of the nonsynonymous codon. It may be that this is a SNP that has not yet been discovered or submitted to dbSNP or it could be due a sequencing error resulting in an incorrect base call. Without evidence of a SNP or viewing the sequencing quality scores, it is impossible to determine the origin of this base change.

There are configurable signals in the alignment displays for the *mRNAs* and *ESTs* track that denote certain features in the sequences and may aid in the identification of noise in the sequences:

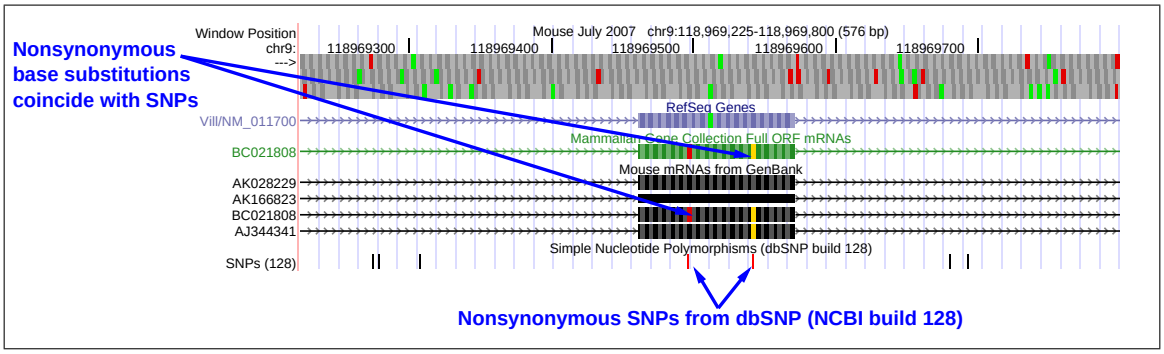
- By default, the *mRNAs* tracks display red or yellow lines in aligned blocks where a base difference between a transcript and the reference genome results in a nonsynonymous codon (Figures 11(b) and (c). At

Figure 10: *TransMap* track at the human hg18 *PPP1R1B* gene locus. (A) *UCSC Genes*; (B) *RefSeq Genes* and (C) GenBank *mRNAs* alignments from other species are projected on to the human genome via *BLASTZ* alignment chains and nets. The *TransMap ESTs* and *human mRNAs* are shown in squish visibility mode since there are large numbers of these transcripts. For the same reason, human *Spliced ESTs* are shown with dense visibility. Colored lines on the alignments represent nonsynonymous codons in aligned transcripts compared to the reference genome sequence; red signifies that amino acids differ in physicochemical properties and yellow signifies similar amino acids. *TransMap* alignments show that the exons towards the $3'$ end of the *PPP1R1B* gene appear to be fast evolving in human whereas those at the $5'$ end are evolving at a slower rate.

(a) Full view of mouse *Vill* gene with nonsynonymous base changes in MGC and mRNA tracks



(b) Exon 5 showing genomic codons and known SNPs resulting in nonsynonymous base changes



(c) Base-level view of exon 5 showing amino acid substitutions caused by nonsynonymous base changes

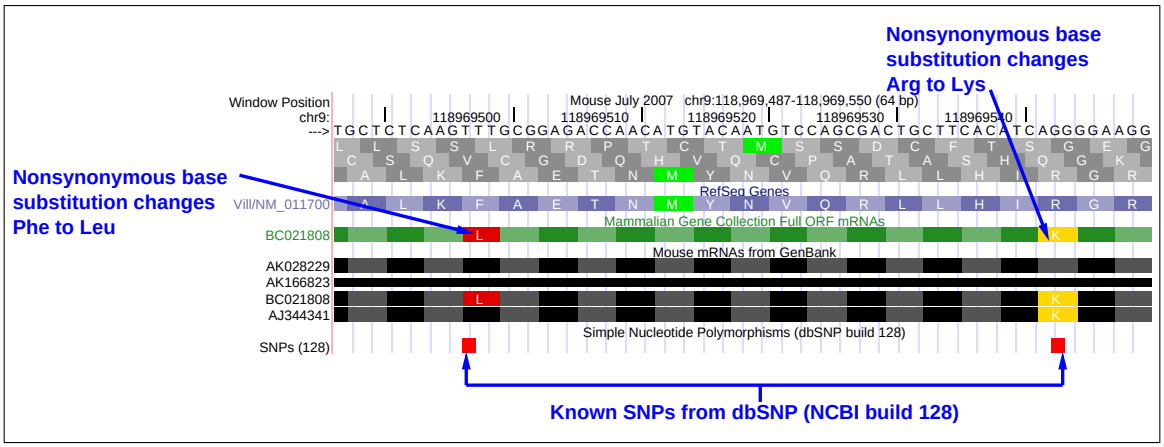


Figure 11: *Genome Browser* transcript rendering modes: Non-synonymous base changes. Various signals can be incorporated into the track displays that can aid in identifying noise in the data. The mouse (mm9, NCBI Build 37 assembly) *Vill* gene is shown at different zoom levels. Colored tick marks (zoomed out) or rectangles (base level view) represent nonsynonymous amino acid changes due to substitutions in *MGC Genes* and *mRNA* transcripts compared to the reference genome. Similar amino acids are colored yellow and amino acids that differ in physicochemical properties are colored red. Zooming in to a specific region can be achieved by clicking at the top of the *Base Position* track or using the zoom controls. (a) shows the entire *Vill* gene. (b) shows exon 5 with genomic codons appearing in the *Base Position* track and the *RefSeq Genes*, *MGC* and *mRNA* tracks. The nonsynonymous base changes in **BC021808** and in the **AJ344341** are due to known polymorphisms colored red in the *SNP* track. (c) shows the base level view of exon 5.

the base level display, the nonsynonymous codons in the transcripts display the one letter amino acid codes (Figure 11(c)).

- By default, the *ESTs* tracks display red lines in aligned blocks showing a base difference between a transcript and the reference genome; this signal can also be turned on for *mRNAs* and *MGC Genes*.
- Vertical blue lines indicate an insertion at the beginning or the end of a transcript relative to the reference genome (Figure 12(a)).
- Green lines indicate the presence of a polyA tail at the 3′ end of a transcript (Figure 12(a)).
- Orange lines indicate insertions in the middle of a transcript relative to the reference genome (Figure 12(b)).
- Double horizontal lines indicate that both the genome and the transcript have an insertion. This may be due to poor sequence quality in a subregion of the transcript (Figures 12(a) and (b)).

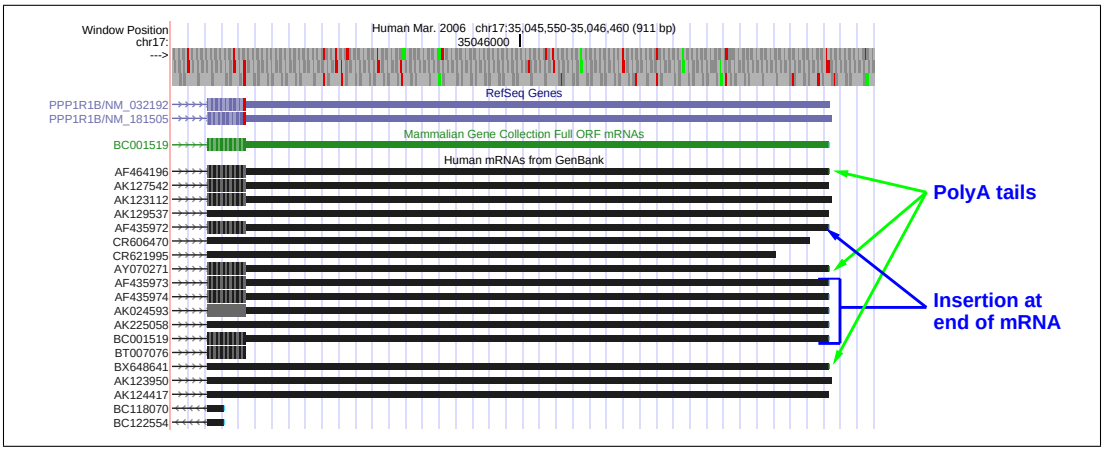
Due to technical reasons, cloned sequences are often incomplete especially at the ends of the UTRs. Therefore, it is difficult to determine whether differences in UTR length are due to "real" variation between transcripts. Genomic data displayed in the *Genome Browser* can help the user to make an informed decision regarding the completeness of mRNAs. A vertical green line at the 3′ end of a transcript indicates the presence of a polyA tail, confirming that the sequence is complete at the 3′ end. Both reported and predicted polyadenylation (polyA) sites are shown in the *Poly(A)* track[4, 24]; data in this track may suggest alternative polyadenylation sites whose use would result in variation of the 3′ UTR length (Figure 5, Box F). For *PPP1R1B*, both predicted and reported sites are in agreement and coincide with the 3′ ends of the transcripts at this locus.

The completeness of the 5′ UTR is more difficult to assess. To aid in this evaluation, there are computationally predicted sites for transcription start sites (TSS) (*Eponine TSS* [8] and *SwitchGear TSS* (http://www.switchgeargenomics.com/) tracks). In Figure 5, Box G, the *SwitchGear Genomics TSS* track predicts the transcription start sites of both the longer and shorter splice variants in the *RefSeq Genes* and *Ensembl Genes* tracks (Box B). Experimental data such as ditags can be used to support and verify the 5′ and 3′ ends of a transcript. *Gene Identification Signature Paired-End Tags (GIS-PET)* [5, 6, 30, 40, 43] involves the sequencing of 5′ and 3′ signatures of full-length cDNAs that are subsequently concatenated to form ditags, sequenced and then mapped to the genome of origin to mark the boundaries of the transcripts (Figure 5, Box G and Figure 14). Ditags offer a much more efficient way of obtaining such data than traditional cDNA sequencing with the limitation that the internal exon structure is not determined. *GIS-PET* data from human embryonic stem cells (hES3) confirm the existence of transcripts whose 5′ end is coincident with both the 5′ end of the longer and shorter *PPP1R1B* transcripts (**NM_181505**) and with the 5′ end of some of the longer mRNAs (Figure 14). At the 3′ end, there are ditags that are coincident the the 3′ end of the RefSeq transcripts and many mRNA transcripts of the *PPP1R1B* gene (Figure 14) as well as the computationally predicted *SwitchGear TSSs*.

8 Conservation and regulation data

The *Conservation* track shows a 28-way multiple alignment of vertebrate genomes[29] created using the *MULTIZ* alignment program[3] and a histogram (wiggle type track) of conservation scores[34] associated with the alignment (Figure 5, Box I). Conservation tends to peak in coding regions of the gene and falls off in non-coding parts (introns and intergenic regions) so

(a) Insertions at the end of transcripts and polyA tails



(b) Internal transcript insertions

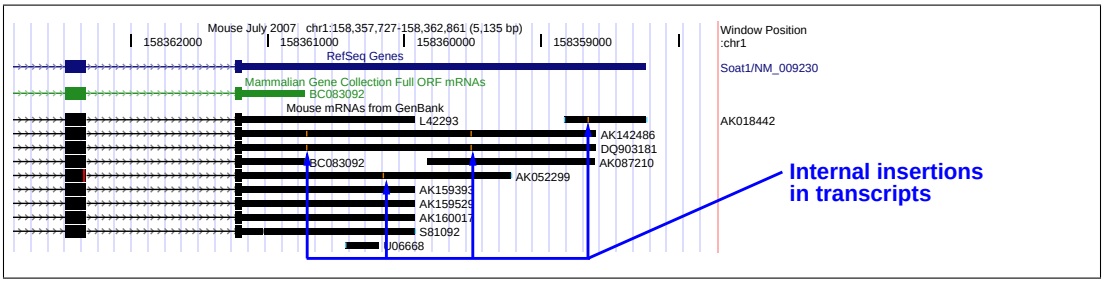
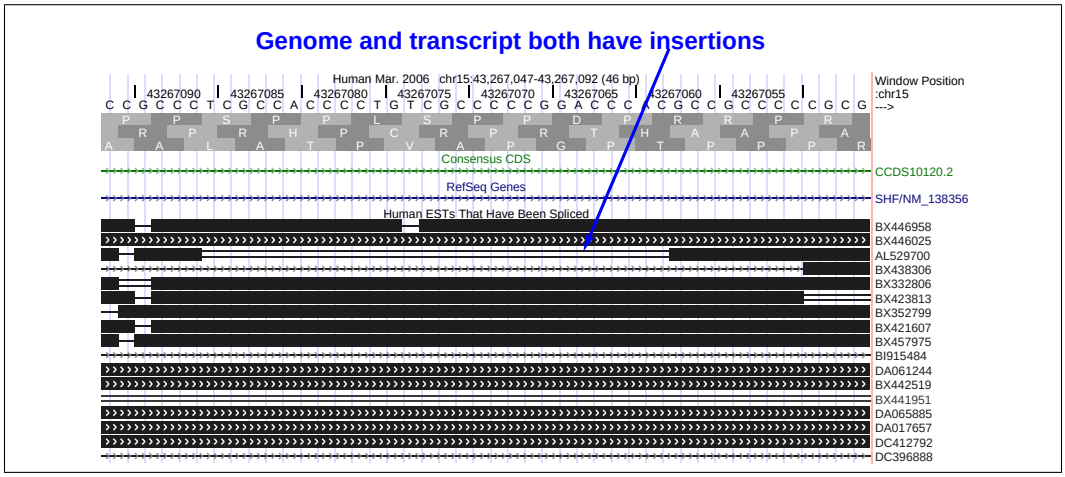


Figure 12: *Genome Browser* transcript rendering modes: Insertions and polyA tails. Various signals can be incorporated into the track displays for GenBank *mRNAs* and *ESTs* that can aid in identifying noise in the data. (a) shows the human *PPP1R1B* gene locus with vertical, colored lines at the 3′ ends of mRNAs; blue indicates insertions (at the 5′ or 3′ transcript end) in the transcript alignment relative to the genome and green lines indicate the presence of polyA tails. (b) shows the mouse (mm9, NCBI Build 37 assembly) *Soat1* gene locus showing orange lines in aligned mRNAs which indicate insertions that occur in the middle of the mRNA sequences in their alignments to the genome. Note that *Soat1* is on the reverse strand and the image has been reversed using the **reverse** button below the *Genome Browser* graphic (see Figure 4) so the gene can be viewed in the 5′ to 3′ orientation. Track element labels are now shown on the right side of the image.

(a) Insertion in both transcript and genome



(b) EST sequence indicating sequence that is unalignable to the genome

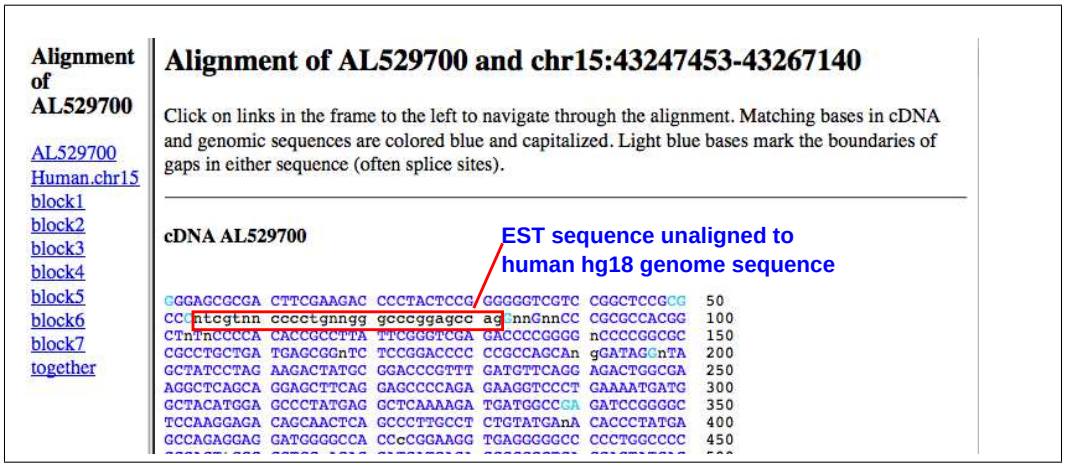


Figure 13: *Genome Browser* transcript rendering modes: Insertions in both sides of the alignment. Various signals can be incorporated into the track displays for GenBank *mRNAs* and *ESTs* that can aid in identifying noise in the data. (a) shows the human (hg18, NCBI Build 36.1) *SHF* gene locus at a position where the human EST, **AL529700**, has double horizontal lines between the black rectangles representing sequence aligned to the genome. The double lines indicate that there are insertions in the alignment in both the EST sequence and the genome. This is shown in (b) where the **AL529700** sequence is blue if it aligns to the genome where black text shows unaligned regions. The unaligned sequence in the red box corresponds to the double lines in image in (a) where it can also be noted that the genome sequence in this region is different from that in the human EST sequence whereas the flanking regions from both sequences are similar, hence the insertion in both transcript and genome in the sequence alignment. This region of the EST is likely to be an erroneous sequence as a result of poor quality sequencing.

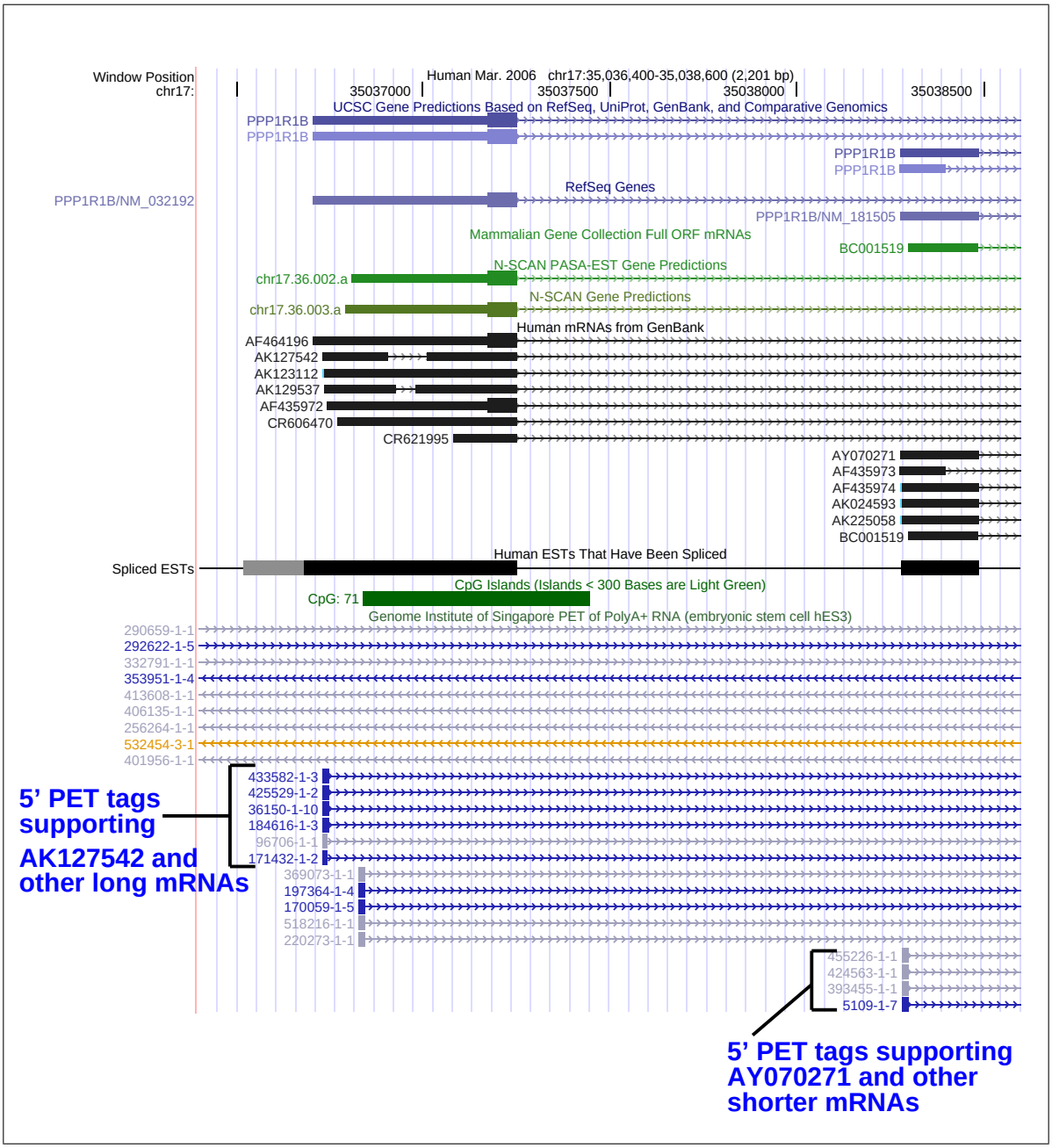


Figure 14: *PPP1R1B* locus showing the *GIS-PET PolyA+ RNA* ditags from the human embryonic cell line (hES3). The *GIS-Pet (Gene Identification Signature Paired End Ditags) RNA* track can be considered in evaluating the $5'$ completeness of transcripts. By clicking on the zoom buttons above the *Browser* image or by clicking above the base numbers at the top of the image, a zoomed-in genome view can be created. The *GIS-Pet RNA* track shows ditags from the $5'$ and $3'$ ends of full-length polyA+ mRNA. In this view, it can be seen that there are ditags whose $5'$ end coincide with the $5'$ end of some of the longer mRNAs e.g. **AK127542** and **AK123112** and ditags that coincide with the $5'$ end of the shorter mRNAs, e.g., **AY070271** and **AF435973** and the RefSeq, **NM_181505**. There is a longer mRNA (**AF464196**) and ESTs that have been used to extend the RefSeq, **NM_181505**, further upstream, but there are no ditags that represent this transcript.

it is a strong signal for a protein-coding gene (Figure 15). Conservation of sequence implies functional significance and can also occur where there are regulatory elements in the genome.

The *ACEScan*[42] track predicts conserved alternative exons that are present in some transcripts and skipped by others in both human and mouse (Figure 5, Box D). Enrichment of splicing regulatory motifs occurs in intron regions close to alternative exons, which also show a greater degree of conservation than those close to constitutive exons. ACEScan uses this information to predict the constitutive exon that is skipped in the human mRNA, **AK129537**.

Transcriptional regulatory elements tend to be enriched near the first exon [27]. Evidence of such motifs are the *CpG island* [10] at the 5′ end of the *PPP1R1B* gene locus and the *SwitchGear TSS* prediction[7, 39] which is color-coded according to confidence level (a darker color implies a higher score) (Figure 5, Box G). The *ORegAnno*[11] track displays hand-curated regulatory regions extracted from the literature (Figure 5, Box G). The darker green rectangle represents a regulatory region, located in an intron of the *PPP1R1B*, bound by the CCCTC-binding factor (zinc finger protein) (CTCF) transcription factor. The lighter green item below represents the actual location of the CTCF binding site. This binding site was determined by ChIP (Chromatin ImmunoPrecipitation)-chip experiments (details are found by clicking on these track items). Histone modifications and multiple transcription factor binding sites for a variety of cell types are shown in the *GIS-PET* track (the GIS-PET method is described in section 7). Tri-methylation of lysine4 and lysine27 on histone H3 is indicated at the 5′ end of the *PPP1R1B* gene (Figure 5, Box G). Such signals for regulatory elements may be misleading; CpG islands are frequently found in or near promoters of genes but not all genes have them, TSS predictions may contain false positives and transcription factor binding site measurements can be noisy.

TargetScan[12, 25, 26] predicts the presence of a microRNA binding site in a conserved region at the 3′ end of the transcript (Figure 5, Box H). The prediction is based partially on conservation so the TargetScan annotation and conservation are not independent evidence of a regulatory region.

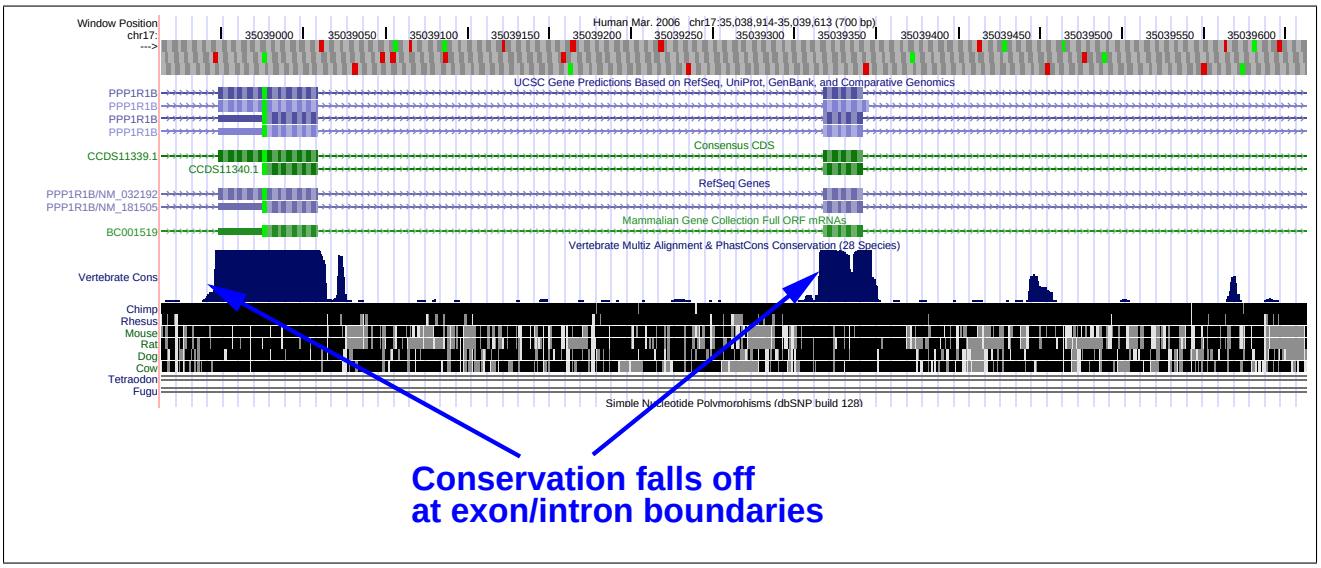
The *Conservation* track shows peaks of conservation that correspond to the coding sequence of the gene (Figure 5, Box I). Conservation falls off at the exon/intron boundaries as illustrated in Figures 15(a) and (b) which show a close-up view of the 5′ and 3′ ends of the *PPP1R1B* gene.

At this zoom level, by default, the *Conservation* track shows the nucleotide sequence of the aligned genomes, and, in coding regions based on the longest *UCSC Genes* transcript at this locus, the codon translation can be seen for each of the genomes. This enables the user to see not only the conservation at the amino acid level but also where there are differences at the amino acid level between proteins encoded by orthologous protein-coding genes. In Figure 15(b), it is possible to see that there is a SNP (**rs35797948**) in the *SNPs (128)* track which is colored red, indicating a nonsynonymous mutation in the coding region. Clicking on the SNP in this track displays further information about this SNP and a link out to the entry for the SNP in dbSNP at NCBI (http://www.ncbi.nlm.nih.gov/SNP/). This reveals that there are two known alleles (A/G) which code for either an arginine (CGC codon) or a histidine (CAC) amino acid in the translation of this last coding exon of *PPP1R1B*. Histidine and arginine are both positively charged making this a conservative substitution; histidine is an aromatic structure.

9 Ordering an MGC clone

Having used the *Genome Browser* to explore the *PPP1R1B* gene, the user may now desire to order a clone for this gene for experimental re-

(a) 5′ view



(b) 3′ view

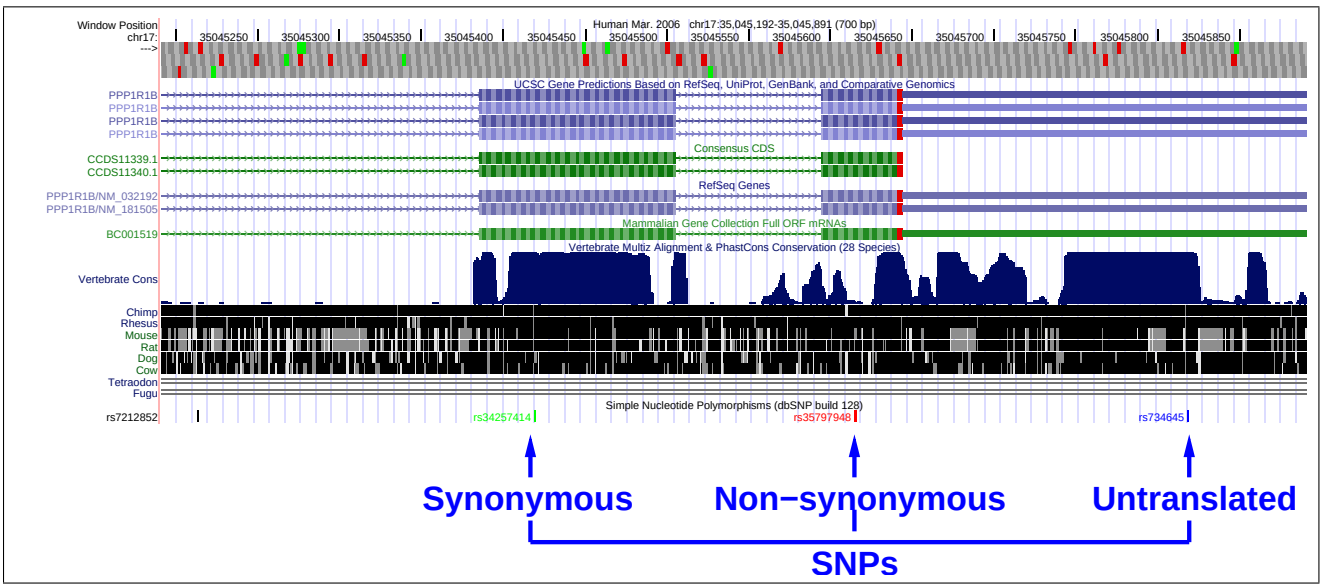


Figure 15: Zoomed in view of the 5′ and 3′ ends of the *PPP1R1B* gene. Conservation is high in the exons in the coding region and falls off at the boundaries of exons.

search. An MGC clone can be ordered by following links from the *Genome Browser* to vendors. The *MGC Genes* track shows the alignment of one full-length MGC clone (BC001519) for *PPP1R1B*. As mentioned previously (section 6.3), it is also shown in the *Human mRNAs* track since MGC clones are in the GenBank database (Figure 5, Box B).

A click on the alignment for the **BC001519** sequence in the MGC Genes track takes the user to the details page for this MGC transcript (Figure 16). The details page displays a gene description, the RefSeq accession and RefSeq description. Additionally, there is information about the clone, links for downloading protein, mRNA and genomic sequences, the alignment to the human reference genome sequence, NCBI clone validation information and links to various external databases including an MGC clone validation report and links to the MGC website. The CDS annotation of the MGC clone is frozen, but the RefSeq transcripts are continually being updated by NCBI manual annotators as more transcript and other experimental evidence becomes available. The RefSeq CDS similarity table on the MGC clone details page shows users how the RefSeq annotations differ to that of the MGC clone.

To order a clone, the user should follow the first link in the **Links** box, **Order MGC clone**, which directs the user to a portal where clone distributors are listed. In some cases, there is a direct ordering link to facilitate the ordering process as seen in Figure 17.

10 Summary

The *Genome Browser* is a very effective tool for the integration and analysis of biological data in a genomic context. It provides an easy method of rapidly locating an MGC clone for a gene of interest with direct links for ordering the clone. Many tools are built into the *Genome Browser*; their use is beyond the scope of this tutorial but there is extensive documentation to help users to navigate use of the *Genome Browser* and its integrated tools. To read the documentation, click on the **Help** link on the top blue menu bar found on the *Genome Browser* website. Links are also provided on the user interface for each tool. Additionally, questions regarding the website and *Genome Browser* use are welcome. Users may search the mailing list archives (http://genome.ucsc.edu/contacts.html) and may also send questions via e-mail to the mailing list: genome@soe.ucsc.edu.

11 FAQ

Question 1: How do I do a batch search for all the genes that lie in a specified region of a human chromosome?

Answer 1: The **Table Browser** is an extremely useful tool for querying the database tables that underlie the *Genome Browser*. The **Table Browser** can be reached by clicking on the **Tables** link on the top blue bar of the *UCSC Genome Bioinformatics* web pages. To retrieve a set of genes from the *UCSC Genes* set, these steps may be followed:

1. Make sure that the correct assembly is selected. For the hg18 human assembly (NCBI Build 36.1), select *Vertebrate* as the **clade**, *Human* as the **genome** and *Mar. 2006* as the **assembly**.

2. For the **group**, select *Genes and Gene Prediction Tracks* and for the **track**, select *UCSC Genes*.

3. Select *knownGene* as the **table**.

4. Select *position*. To find genes in a region of the chromosome, type the genomic location in the text box in the format, chr1:10000-100000. This example will find genes on chromosome 1 between base 10,000 and base 100,000. Then press the **lookup** button.

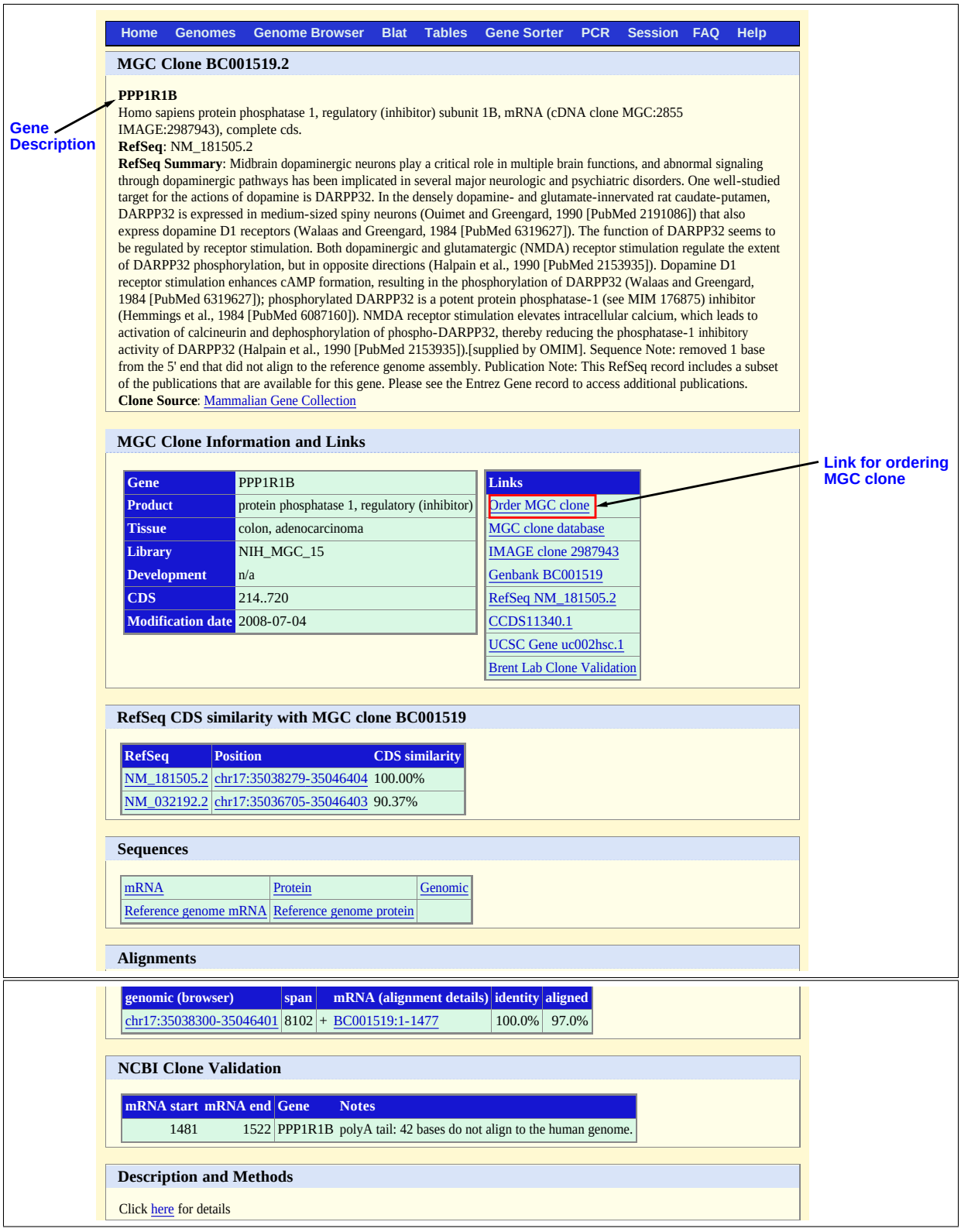


Figure 16: Details page for the *PPP1R1B* MGC clone. The details page for the MGC full-length ORF mRNA (**BC001519**) shows information about the gene, this MGC clone and its sequence, similarity between the the RefSeq CDS and the annotated MGC CDS region, alignments, clone validation information and links to external databases.

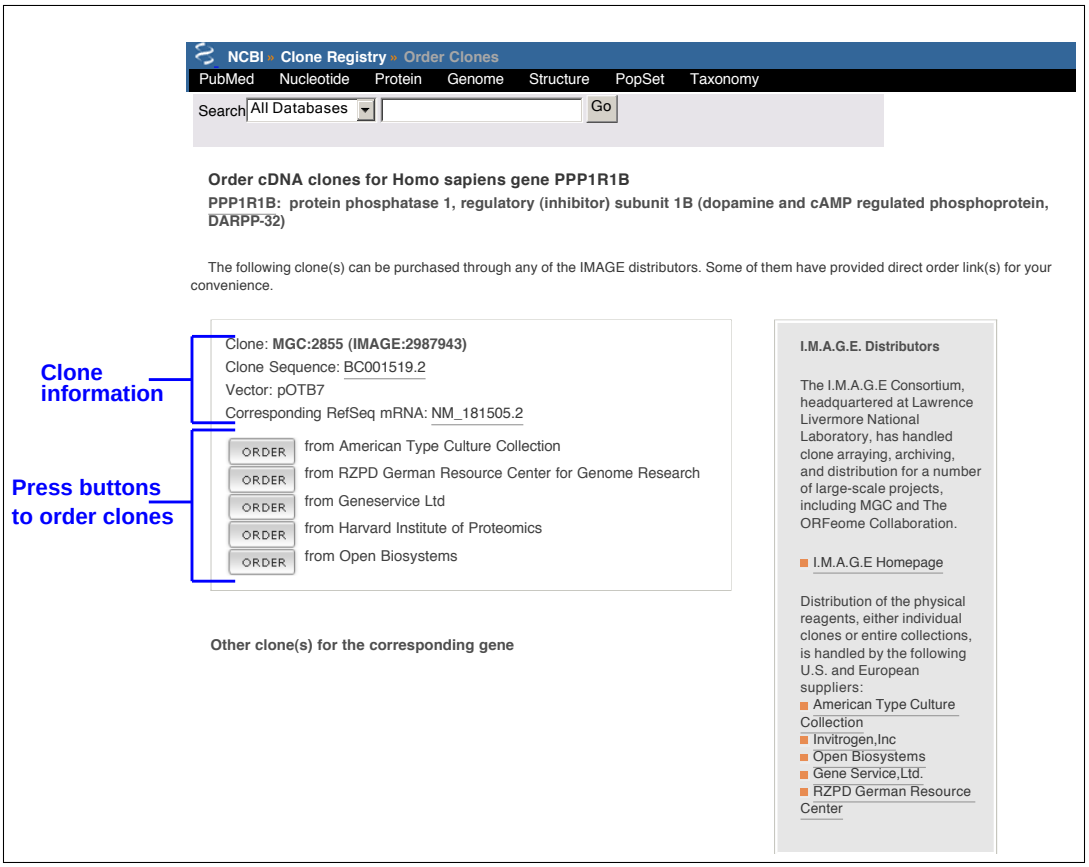


Figure 17: NCBI page for ordering *PPP1R1B* cDNA clones. The page displays links to vendors for cDNA clones. In this case, there is only one clone listed for this gene. GenBank and RefSeq accessions, and the name of the vector containing the clone are shown. The clone can be obtained from distributors of Integrated Molecular Analysis of Genomes and their Expression (I.M.A.G.E) Consortium clones and links to some of the vendors are provided. For this clone, direct ordering links are for American Type Culture Collection (ATCC), RZPD German Resource Center for Genome Research, Geneservice Ltd., Harvard Institute of Proteomics and Open Biosystems.

5. Finally select the **output format**. The default will provide the *UCSC Gene* identifier and the genomic location for each *UCSC Gene*. Press the **get output** button to perform the search.

A similar batch search can be done for *RefSeq Genes* (**table**: *refGene*), *Ensembl Genes* (**table**: *ensGene*) or other gene set.

Question 2: How do I do a batch search for genes on an entire human chromosome, for example, chr21?
Answer 2: To carry out this search, follow the instructions in the answer to **Question 1**, except at step 4, type the name of the chromosome into the **position** box in the format, chr21.

Question 3: How do I do batch retrievals for full-CDS MGC cDNA clones?
Answer 3: Use the **Table Browser**. See **Question 1** and **Question 2** for guidance on batch retrieval using this tool. To search for full-CDS MGC cDNA clones, the **track** to select is *MGC Genes* and the **table** to select is *mgcGenes*.

12 Other resources

- UCSC Genome Browser help at http://genome.ucsc.edu/training
- UCSC Genome Browser updates in the Nucleic Acids Research (NAR) Database issues[13, 16, 18, 23]
- The original UCSC Genome Browser publication[21]
- UCSC genome browser tutorial[45]
- UCSC Genome Browser: Deep support for molecular biomedical research.[28]
- Chapter 1, Unit 1.4 of Using Biological Databases in "Current Protocols in Bioinformatics"
- "Genomes, Browsers and Databases: Data-Mining Tools for Integrated Genomic Databases", Peter Schattner (first edition, 2008).

References

[1] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang. The genetic association database. *Nat Genet*, 36:431–432, May 2004. ISSN 1061-4036. URL http://dx.doi.org/10.1038/ng0504-431. 10.1038/ng0504-431.

[2] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. GenBank. *Nucl. Acids Res.*, 36(suppl_1): D25–30, 2008. doi: 10.1093/nar/gkm929. URL http://nar.oxfordjournals.org/cgi/content/abstract/36/suppl_1/D25.

[3] M. Blanchette, W. J. Kent, C. Riemer, L. Elnitski, A. F. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler, and W. Miller. Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Res.*, 14(4):708–715, 2004. doi: 10.1101/gr.1933104. URL http://www.genome.org/cgi/content/abstract/14/4/708.

[4] Y. Cheng, R. M. Miura, and B. Tian. Prediction of mRNA polyadenylation sites by support vector machine. *Bioinformatics*, 22(19):2320–2325, 2006. doi: 10.1093/bioinformatics/btl394. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/22/19/2320.

[5] K. Chiu, C.-H. Wong, Q. Chen, P. Ariyaratne, H. Ooi, C.-L. Wei, W.-K. Sung, and Y. Ruan. Pet-tool: a software suite for comprehensive processing and managing of paired-end ditag (pet) sequence data. *BMC Bioinformatics*, 7(1): 390, 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-390. URL http://www.biomedcentral.com/1471-2105/7/390.

[6] A. Choo, J. Padmanabhan, A. Chin, W. J. Fong, and S. K. W. Oh. Immortalized feeders for the scale-up of human embryonic stem cells in feeder and feeder-free conditions. *Journal of Biotechnology*, 122:130–141, Mar 2006. URL http://www.sciencedirect.com/science/article/B6T3C-4HBTDBW-3/1/b4f66b5f48e5b3227c5916653dbd88f0.

[7] S. J. Cooper, N. D. Trinklein, E. D. Anton, L. Nguyen, and R. M. Myers. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.*, 16(1): 1–10, 2006. doi: 10.1101/gr.4222606. URL http://genome.cshlp.org/cgi/content/abstract/16/1/1.

[8] T. A. Down and T. J. P. Hubbard. Computational Detection and Location of Transcription Start Sites in Mammalian Genomic DNA. *Genome Res.*, 12(3):458–461, 2002. doi: 10.1101/gr.216102. URL http://www.genome.org/cgi/content/abstract/12/3/458.

[9] P. Flicek, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, T. Eyre, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, R. Holland, K. L. Howe, K. Howe, N. Johnson, A. Jenkinson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. J. Vilella, J. Vogel, S. White, M. Wood, E. Birney, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, T. J. P. Hubbard, A. Kasprzyk, G. Proctor, J. Smith, A. Ureta-Vidal, and S. Searle. Ensembl 2008. *Nucl. Acids Res.*, 36(suppl_1):D707–714, 2008. doi: 10.1093/nar/gkm988. URL http://nar.oxfordjournals.org/cgi/content/abstract/36/suppl_1/D707.

[10] M. Gardiner-Garden and M. Frommer. Cpg islands in vertebrate genomes. *Journal of Molecular Biology*, 196:261–282, Jul 1987. URL http://www.sciencedirect.com/science/article/B6WK7-4DMP8FD-G/1/f811b0610c688c457165b7da3cdc22b7.

[11] O. L. Griffith, S. B. Montgomery, B. Bernier, B. Chu, K. Kasaian, S. Aerts, S. Mahony, M. C. Sleumer, M. Bilenky, M. Haeussler, M. Griffith, S. M. Gallo, B. Giardine, B. Hooghe, P. Van Loo, E. Blanco, A. Ticoll, S. Lithwick, E. Portales-Casamar, I. J. Donaldson, G. Robertson, C. Wadelius, P. De Bleser, D. Vlieghe, M. S. Halfon, W. Wasserman, R. Hardison, C. M. Bergman, S. J. Jones, and T. O. R. A. Consortium. ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucl. Acids Res.*, 36(suppl_1):D107–113, 2008. doi: 10.1093/nar/gkm967. URL http://nar.oxfordjournals.org/cgi/content/abstract/36/suppl_1/D107.

[12] A. Grimson, K. K.-H. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel. Microrna targeting specificity in mammals: Determinants beyond seed pairing. *Molecular Cell*, 27:91–105, Jul 2007. URL http://www.sciencedirect.com/science/article/B6WSR-4P48CFV-9/1/9e1fa34b74d2d78d3ac3fd37b6a4b189.

[13] A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: update 2006. *Nucl. Acids Res.*, 34(suppl_1):D590–598, 2006. doi: 10.1093/nar/gkj144. URL http://nar.oxfordjournals.org/cgi/content/abstract/34/suppl_1/D590.

[14] F. Hsu, T. H. Pringle, R. M. Kuhn, D. Karolchik, M. Diekhans, D. Haussler, and W. J. Kent. The UCSC Proteome Browser. *Nucl. Acids Res.*, 33(suppl_1):D454–458, 2005. doi: 10.1093/nar/gki100. URL http://nar.oxfordjournals.org/cgi/content/abstract/33/suppl_1/D454.

[15] F. Hsu, W. J. Kent, H. Clawson, R. M. Kuhn, M. Diekhans, and D. Haussler. The UCSC Known Genes. *Bioinformatics*, 22(9):1036–1046, 2006. doi: 10.1093/bioinformatics/btl048. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/22/9/1036.

[16] D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database. *Nucl. Acids Res.*, 31(1):51–54, 2003. doi: 10.1093/nar/gkg129. URL http://nar.oxfordjournals.org/cgi/content/abstract/31/1/51.

[17] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent. The UCSC Table Browser data retrieval tool. *Nucl. Acids Res.*, 32(suppl_1): D493–496, 2004. doi: 10.1093/nar/gkh103. URL http://nar.oxfordjournals.org/cgi/content/abstract/32/suppl_1/D493.

[18] D. Karolchik, R. M. Kuhn, R. Baertsch, G. P. Barber, H. Clawson, M. Diekhans, B. Giardine, R. A. Harte, A. S. Hinrichs, F. Hsu, K. M. Kober, W. Miller, J. S. Pedersen, A. Pohl, B. J. Raney, B. Rhead, K. R. Rosenbloom, K. E. Smith, M. Stanke, A. Thakkapallayil, H. Trumbower, T. Wang, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: 2008 update. *Nucl. Acids Res.*, 36(suppl_1): D773–779, 2008. doi: 10.1093/nar/gkm966. URL http://nar.oxfordjournals.org/cgi/content/abstract/36/suppl_1/D773.

[19] W. Kent, F. Hsu, D. Karolchik, R. M. Kuhn, H. Clawson, H. Trumbower, and D. Haussler. Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res.*, 15(5):737–741, 2005. doi: 10.1101/gr.3694705. URL http://www.genome.org/cgi/content/abstract/15/5/737.

[20] W. J. Kent. BLAT: The BLAST-like alignment tool. *Genome Res*, 12:656–664, 2002.

[21] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The Human Genome Browser at UCSC. *Genome Res.*, 12(6): 996–1006, 2002. doi: 10.1101/gr.229102. URL http://genome.cshlp.org/cgi/content/abstract/12/6/996.

[22] W. J. Kent, R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(20):11484–11489, 2003. doi: 10.1073/pnas.1932072100. URL http://www.pnas.org/content/100/20/11484.abstract.

[23] R. M. Kuhn, D. Karolchik, A. S. Zweig, H. Trumbower, D. J. Thomas, A. Thakkapallayil, C. W. Sugnet, M. Stanke, K. E. Smith, A. Siepel, K. R. Rosenbloom, B. Rhead, B. J. Raney, A. Pohl, J. S. Pedersen, F. Hsu, A. S. Hinrichs, R. A. Harte, M. Diekhans, H. Clawson, G. Bejerano, G. P. Barber, R. Baertsch, D. Haussler, and W. J. Kent. The UCSC genome browser database: update 2007. *Nucl. Acids Res.*, 35(suppl_1): D668–673, 2007. doi: 10.1093/nar/gkl928. URL http://nar.oxfordjournals.org/cgi/content/abstract/35/suppl_1/D668.

[24] J. Y. Lee, I. Yeh, J. Y. Park, and B. Tian. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucl. Acids Res.*, 35(suppl_1):D165–168, 2007. doi: 10.1093/nar/gkl870. URL http://nar.oxfordjournals.org/cgi/content/abstract/35/suppl_1/D165.

[25] B. P. Lewis, I.-h. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge. Prediction of mammalian microrna targets. *Cell*, 115:787–798, Dec 2003.

URL http://www.sciencedirect.com/science/article/B6WSN-4C5RJXX-4/1/daacc44b2dc4cbd20152154f2556cbfb.

[26] B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, 120:15–20, Jan 2005. URL http://www.sciencedirect.com/science/article/B6WSN-4F7PRGS-6/1/4caeed2c84a218c0bd80bdf7fe1506c4.

[27] J. Majewski and J. Ott. Distribution and Characterization of Regulatory Elements in the Human Genome. *Genome Res.*, 12 (12):1827–1836, 2002. doi: 10.1101/gr.606402.

[28] M. E. Mangan, J. M. Williams, S. M. Lathe, D. Karolchik, and W. C. Lathe III. *Biotechnology Annual Review*, volume Volume 14, chapter UCSC Genome Browser: Deep support for molecular biomedical research, pages 63–108. Elsevier, 2008. URL http://www.sciencedirect.com/science/article/B7CTS-4SX89K0-8/2/fd8ba712b4824e759a88e34dac02e743.

[29] W. Miller, K. Rosenbloom, R. C. Hardison, M. Hou, J. Taylor, B. Raney, R. Burhans, D. C. King, R. Baertsch, D. Blankenberg, S. L. Kosakovsky Pond, A. Nekrutenko, B. Giardine, R. S. Harris, S. Tyekucheva, M. Diekhans, T. H. Pringle, W. J. Murphy, A. Lesk, G. M. Weinstock, K. Lindblad-Toh, R. A. Gibbs, E. S. Lander, A. Siepel, D. Haussler, and W. J. Kent. 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.*, 17 (12):1797–1808, 2007. doi: 10.1101/gr.6761107. URL http://www.genome.org/cgi/content/abstract/17/12/1797.

[30] P. Ng, C.-L. Wei, W.-K. Sung, K. P. Chiu, L. Lipovich, C. C. Ang, S. Gupta, A. Shahab, A. Ridwan, C. H. Wong, E. T. Liu, and Y. Ruan. Gene identification signature (gis) analysis for transcriptome characterization and genome annotation. *Nat Meth*, 2:105–111, Feb 2005. ISSN 1548-7091. URL http://dx.doi.org/10.1038/nmeth733. 10.1038/nmeth733.

[31] K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucl. Acids Res.*, 35(suppl_1): D61–65, 2007. doi: 10.1093/nar/gkl842. URL http://nar.oxfordjournals.org/cgi/content/abstract/35/suppl_1/D61.

[32] S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller. Human-Mouse Alignments with BLASTZ. *Genome Res.*, 13(1):103–107, 2003. doi: 10.1101/gr.809403. URL http://www.genome.org/cgi/content/abstract/13/1/103.

[33] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucl. Acids Res.*, 29 (1):308–311, 2001. doi: 10.1093/nar/29.1.308. URL http://nar.oxfordjournals.org/cgi/content/abstract/29/1/308.

[34] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8): 1034–1040, 2005.

[35] A. Siepel, M. Diekhans, B. Brejova, L. Langton, M. Stevens, C. L. Comstock, C. Davis, B. Ewing, S. Oommen, C. Lau,

H.-C. Yu, J. Li, B. A. Roe, P. Green, D. S. Gerhard, G. Temple, D. Haussler, and M. R. Brent. Targeted discovery of novel human exons by comparative genomics. *Genome Res.*, 17(12):1763–1773, 2007. doi: 10.1101/gr.7128207. URL http://www.genome.org/cgi/content/abstract/17/12/1763.

[36] M. Stanke, M. Diekhans, R. Baertsch, and D. Haussler. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, page btn013, 2008. doi: 10.1093/bioinformatics/btn013. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/btn013v1.

[37] T. M. P. Team. The Status, Quality, and Expansion of the NIH Full-Length cDNA Project: The Mammalian Gene Collection (MGC). *Genome Res.*, 14(10b):2121–2127, 2004. doi: 10.1101/gr.2596504. URL http://www.genome.org/cgi/content/abstract/14/10b/2121.

[38] D. J. Thomas, H. Trumbower, A. D. Kern, B. L. Rhead, R. M. Kuhn, D. Haussler, and W. J. Kent. Variation resources at UC Santa Cruz. *Nucl. Acids Res.*, 35(suppl_1):D716–720, 2007. doi: 10.1093/nar/gkl953. URL http://nar.oxfordjournals.org/cgi/content/abstract/35/suppl_1/D716.

[39] N. D. Trinklein, S. F. Aldred, S. J. Hartman, D. I. Schroeder, R. P. Otillar, and R. M. Myers. An Abundance of Bidirectional Promoters in the Human Genome. *Genome Res.*, 14(1):62–66, 2004. doi: 10.1101/gr.1982804. URL http://genome.cshlp.org/cgi/content/abstract/14/1/62.

[40] C.-L. Wei, Q. Wu, V. B. Vega, K. P. Chiu, P. Ng, T. Zhang, A. Shahab, H. C. Yong, Y. Fu, Z. Weng, J. Liu, X. D. Zhao, J.-L. Chew, Y. L. Lee, V. A. Kuznetsov, W.-K. Sung, L. D. Miller, B. Lim, E. T. Liu, Q. Yu, H.-H. Ng, and Y. Ruan. A global map of p53 transcription-factor binding sites in the human genome. *Cell*, 124:207–219, Jan 2006. URL http://www.sciencedirect.com/science/article/B6WSN-4J1B5X8-S/1/199bb4dbe7ad1d5297fe65359c6dabec.

[41] L. G. Wilming, J. G. R. Gilbert, K. Howe, S. Trevanion, T. Hubbard, and J. L. Harrow. The vertebrate genome annotation (Vega) database. *Nucl. Acids Res.*, 36(suppl_1):D753–760, 2008. doi: 10.1093/nar/gkm987. URL http://nar.oxfordjournals.org/cgi/content/abstract/36/suppl_1/D753.

[42] G. W. Yeo, E. Van Nostrand, D. Holste, T. Poggio, and C. B. Burge. Identification and analysis of alternative splicing events conserved in human and mouse. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (8):2850–2855, 2005. doi: 10.1073/pnas.0409742102. URL http://www.pnas.org/content/102/8/2850.abstract.

[43] K. I. Zeller, X. Zhao, C. W. H. Lee, K. P. Chiu, F. Yao, J. T. Yustein, H. S. Ooi, Y. L. Orlov, A. Shahab, H. C. Yong, Y. Fu, Z. Weng, V. A. Kuznetsov, W.-K. Sung, Y. Ruan, C. V. Dang, and C.-L. Wei. Global mapping of c-Myc binding sites and target gene networks in human B cells. *Proceedings of the National Academy of Sciences*, 103(47):17834–17839, 2006. doi: 10.1073/pnas.0604129103. URL http://www.pnas.org/content/103/47/17834.abstract.

[44] J. Zhu, J. Z. Sanborn, M. Diekhans, C. B. Lowe, T. H. Pringle, and D. Haussler. Comparative genomics search for losses of long-established genes on the human lineage. *PLoS Computational Biology*, 3:

e247 EP –, Dec 2007. URL http://dx.doi.org/10.1371/journal.pcbi.0030247.

[45] A. S. Zweig, D. Karolchik, R. M. Kuhn, D. Haussler, and W. J. Kent. Ucsc genome browser tutorial. *Genomics*, 92:75–84, Aug 2008. URL http://www.sciencedirect.com/science/article/B6WG1-4SN8V4K-1/2/52fefdb7d8345ae05aa0ec653a0adc27.