Appendix 1. Description of the Genome Browser

The image below shows a genome browser for *D. melanogaster*'s chr3L assembly. In the genome browser, each chromosome may be organized into smaller projects called contigs (for contiguous sequences), but here we have an entire region of chromosome 3. We will use this image to break down the sections of a genome browser. There are four major sections (see Figure 1 below):

- A top green toolbar is used to navigate to the different tools provided by the Browser.
- Navigation Controls allow us to navigate or zoom to different parts of the genome.
- A genomic features panel (the white area) shows the locations of the different genomic features within the portion of the genome (e.g., chr3L) specified by the label next to the "enter position or search terms" text box.
- A Display Controls section may be used to manipulate how much detail is visible in the genomic features panel of the Genome Browser. To match the screenshot in **Figure 1**, scroll down and click 'hide all' to hide all tracks in the panel. Then, scroll to the bar labeled "Mapping and Sequencing Tracks", go to "Base Position", and select "**dense**" from the drop-down menu. Then scroll down to "Genes and Gene Prediction Tracks", go to "FlyBase Genes" and select "**squish**" from the drop-down menu. Check that all other tracks are set to "hide", and then click on any "**refresh**" button to update the genomic features panel.

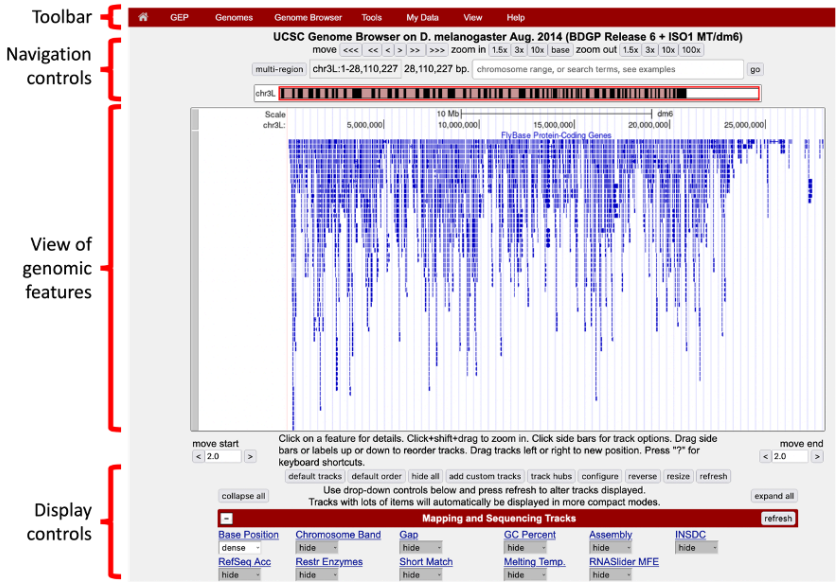

Figure 1 The four major sections of the Genome Browser.

You can use the buttons in the "Navigation control" section to navigate to different parts of the genome. You can zoom in to a region by clicking on one of the buttons next to the "zoom in" label (i.e., 1.5x, 3x, 10x, base). Similarly, you can zoom out by clicking on the buttons next to the

"zoom out" label. Alternatively, you can enter the genome coordinates into the "enter position or search terms" field and then click on the "go" button to navigate to a specific region in the genome assembly.

The "size" field next to the "enter position or search terms" text box (red arrow in Figure 2) shows the total size of the genomic region that you are viewing. In this case, the "size" field shows that chr3L (i.e., the left arm of chromosome 3) in *Drosophila melanogaster* has a total length of ~28 million base pairs (bp). We will learn more about the key functionalities of the Genome Browser in subsequent modules. For now, we will focus on the large white rectangle shown on this page; this contains a graphical representation of the genomic features (e.g., protein coding genes) of chr3L mapped against the DNA sequence, which is embedded in the top line of the white box.

The different types of features (also known as "**tracks**" or "**evidence tracks**") are separated by a title and are often shown in different colors. What types and how many tracks are shown in the view of genomic features is controlled by the display controls at the bottom. Watch this 4-minute long Tracks video (http://tiny.cc/tracks-gep) to understand how to use tracks to display information you're interested in examining.

We can examine the region under the blue title labeled "**FlyBase Protein-Coding Genes**" to estimate the number of protein-coding genes on chr3L. In this track each gene is represented by a set of blue boxes connected by thin blue lines. There are clearly fewer blue boxes at the right side of the image compared to the left, which suggests that genes are not uniformly distributed along the chromosome (Figure 2).

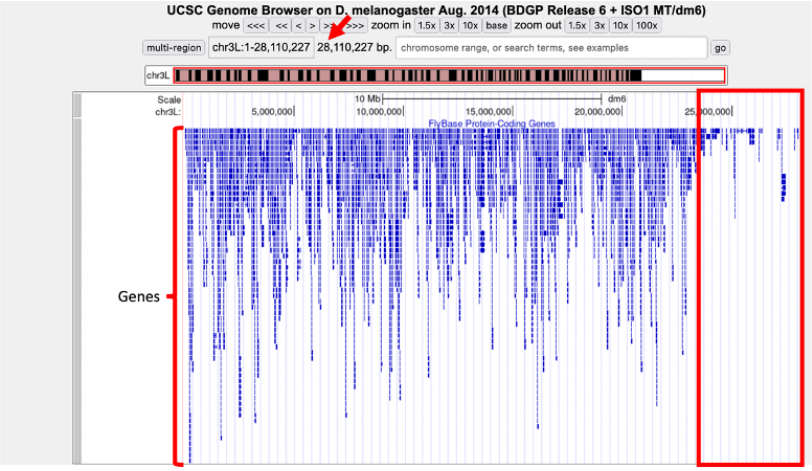

Figure 2 Genome Browser shows that the entire *D. melanogaster* chr3L sequence has a length of ~28 million base pairs (red arrow) and that the right end of the chromosome has low gene density (red box).

Lastly, the proportions of what you view on your computer may differ from what you see in Figures 1 and 2. That's okay, so long as you can see what you're viewing. To change the font size or panel dimensions, at the top of the webpage, click 'View→ Configure Browser'. On the new page, adjust the image width or text size and click submit to see how that changes what you see. You may want to record your preferred dimensions in your lab notebook for future settings.

Appendix 2. Reading frames

The combination of the directionality (with two alternative directions) and the three rows in the "Base Position" track means that there are six different ways to translate a genomic region, (i.e., to determine the sequence of amino acids from a DNA sequence). These different ways to translate a genomic region are known as **reading frames**.

1. To illustrate this concept, change the “enter position or search terms” text box to "**contig1:1-12**" and then click “go” in order to zoom in to the first 12 nucleotides of the contig1 sequence.
2. Click on the arrow underneath the "contig1" label in the "Base Position" track so that it points to the right (Figure 1).

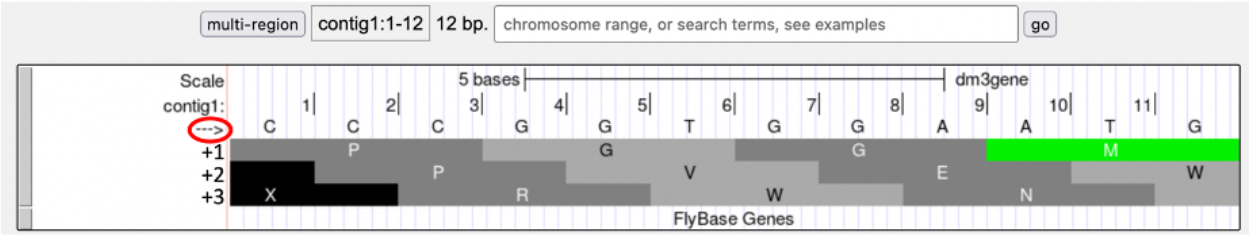

Figure 1 Examine the "Base Position" track for the first 12 bases of contig1 in the top strand.

The first row (**frame +1**) begins at the **first** nucleotide in contig1 and the first amino acid (P) is derived from the codon **CCC**. The second row (frame +2) begins at the **second** nucleotide in contig1 and the codon **CCG** also codes for the amino acid P. The third row (frame +3) begins at the **third** nucleotide in contig1 and the codon **CGG** corresponds to the amino acid R (Figure 2). Because a codon is comprised of 3 nucleotides, the codon beginning at the fourth nucleotide (GGT) is again in frame +1.

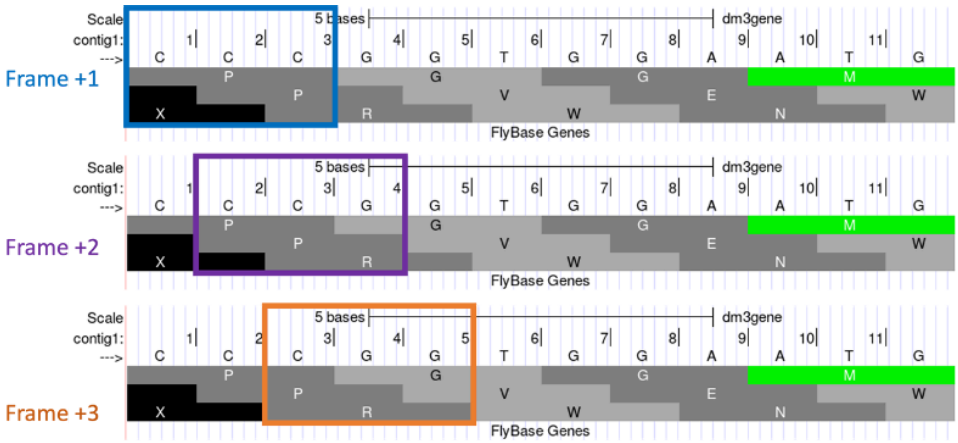

Figure 2 Interpreting the reading frame using the Base Position track.

Examination of the "Base Position" track at the beginning of the contig shows that the three positive reading frames are numbered relative to the start of the contig1 sequence. Similarly, the three reading frames on the bottom strand are numbered relative to the end of the contig1 sequence (i.e., the beginning of the reverse complement of the contig sequence). Because contig1 has a total length of 11,000 bp, we will change the "enter position or search terms" field to "**contig1:10,989-11,000**" so that we can examine the last 12 nucleotides of this contig.

3. Click on the arrow underneath the "contig1" label so that it points to the left (Figure 3).

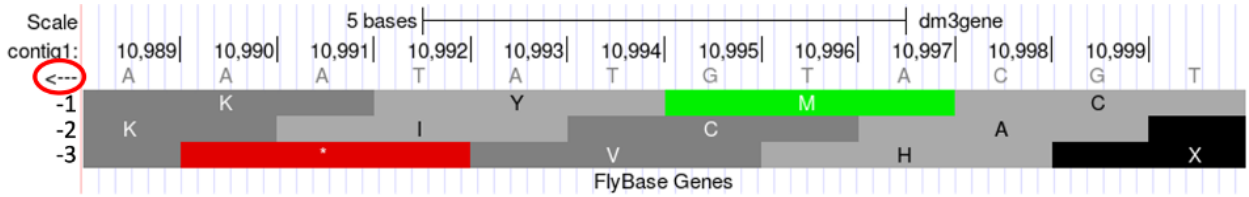

Figure 3 Examine the "Base Position" track for the last 12 nucleotides of contig1 in the bottom strand.

Because we are examining the reverse complement of the contig1 sequence, we need to read the nucleotide and amino acid sequences on the "Base Position" track from right to left. The first row (**frame -1**) begins at the last nucleotide (11,000) of contig1 and the codon **TGC** codes for the amino acid C. The second row (frame -2) begins at the penultimate nucleotide at 10,999 and the codon **GCA** codes for the amino acid A. The third row (frame -3) begins at 10,998 and the codon **CAT** corresponds to the amino acid H (Figure 4).

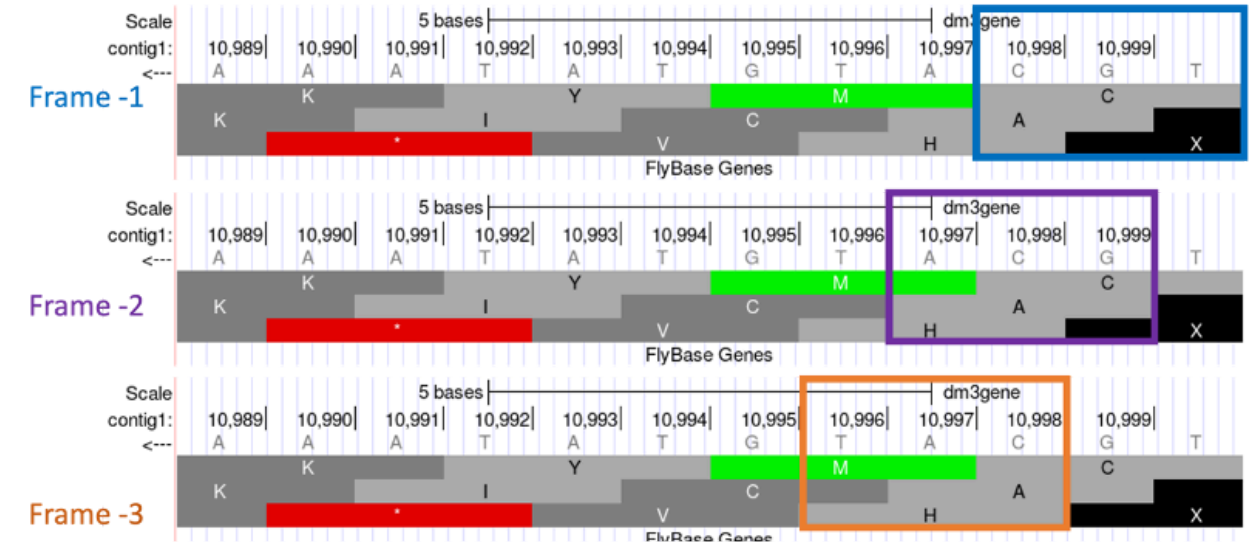

Figure 4 Using the "Base Position" track to interpret the reading frames on the bottom strand.

Appendix 3. RNA-Seq and Expression Data, Briefly

All RNAs in the cell are collectively known as the "transcriptome," as almost all RNA is produced by transcription from a DNA template. (In some cases, RNA is made from an RNA template.) The transcriptome includes messenger RNAs, ribosomal RNAs, transfer RNAs, and other RNAs that have specialized functions in the cell.

RNA can be harvested from cells or a whole organism like *Drosophila* and converted to DNA, then sequenced to produce RNA-Seq (RNA Sequencing) data. First, extracted mRNA that has been fully spliced is copied back to DNA with the enzyme called reverse transcriptase. Short fragments of the copied or complementary DNA are sequenced, and then these segments are mapped back to the genome. By analyzing the mapping data, it is possible to know which and how many messenger RNAs have been synthesized (Figure 1).

This is a powerful technique that allows us to see when and where different genes are expressed. This kind of information can help researchers and clinicians know which genes are expressed in different types of cancer, for example. We are going to use RNA-Seq data to explore how the *transformer* (*tra*) gene is expressed in male vs. female *Drosophila*.

RNA-Seq data can indicate where transcription occurs, to the exact nucleotide. The number of RNA-Seq fragments that map to a given site also tells us how many copies of the RNA are present in the sample. Remember that the initial RNA transcripts are quickly processed to remove the introns. Hence in total RNA from a cell, sequences from exons will be much more abundant than sequences from introns.

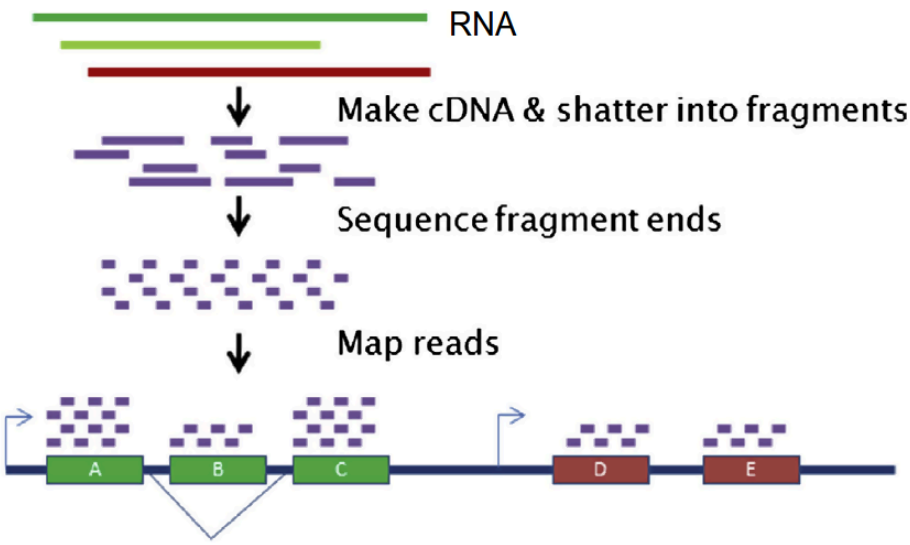

Figure 1 A Brief Summary of RNA-Seq experiments. RNA is taken from a cell and converted to DNA (which can be sequenced in the lab). The complementary DNA (cDNA) is cut up and sequenced. All data from the fragments (i.e. reads) is mapped back onto the genome of that organism, and the number of reads overlapping a particular region indicates the amount of transcription that occurred for the corresponding gene (i.e. the amount of gene expression).

Appendix 4. Identifying splice sites

Two software programs, called TopHat and Bowtie, use the RNA-Seq data to graphically represent the exon junctions (i.e. the places where exons end and begin). The resulting graphic on the genome browser coincidentally looks something like a little bowtie (two small boxes connected by a thin line) (Figure 1). The boxes represent the sequenced mRNA (the exons), and the line represents a gap (the intron). The exon junction can be inferred when the first part of a sequenced fragment from the RNA population matches (for example) DNA positions 50-100 and the second part of the same fragment matches DNA positions 200-250; the RNA from positions 101-199 must have been taken out of the middle!

Short sequences are present at the beginning and end of each intron that allow the spliceosome—the molecular machinery that cuts out introns—to precisely remove the intron, leaving only the exon sequences in the mature mRNA. The first two nucleotides of the intron are the splice donor site and almost always the nucleotides "GT". The last two nucleotides of the intron are the splice acceptor site and almost always the nucleotides "AG". More information on RNA-Seq and the search for splice junctions can be found in a supplementary video on RNA-Seq and TopHat video (http://tiny.cc/rnaseq_tophat) if you are interested in a video walkthrough of evaluating RNA-Seq data.

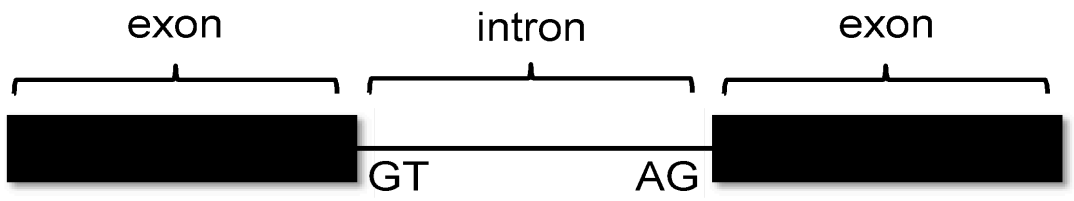

Figure 1 A diagram of intron-exon junctions.