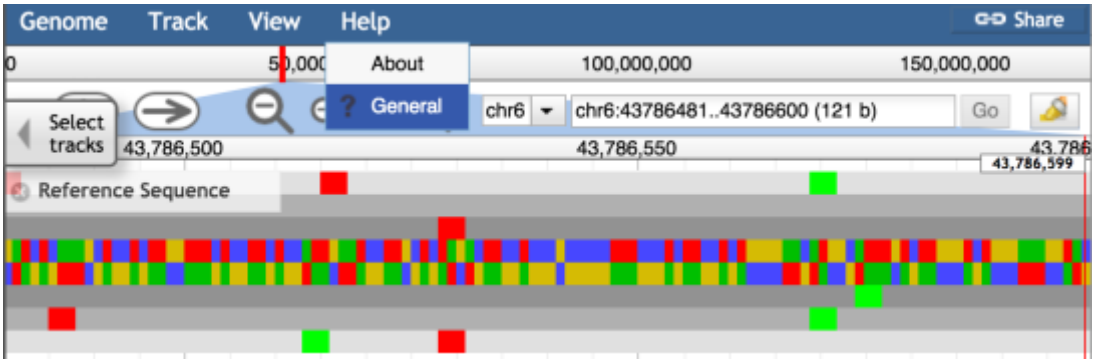
JBrowse Tutorial – RNAStructuromeDB

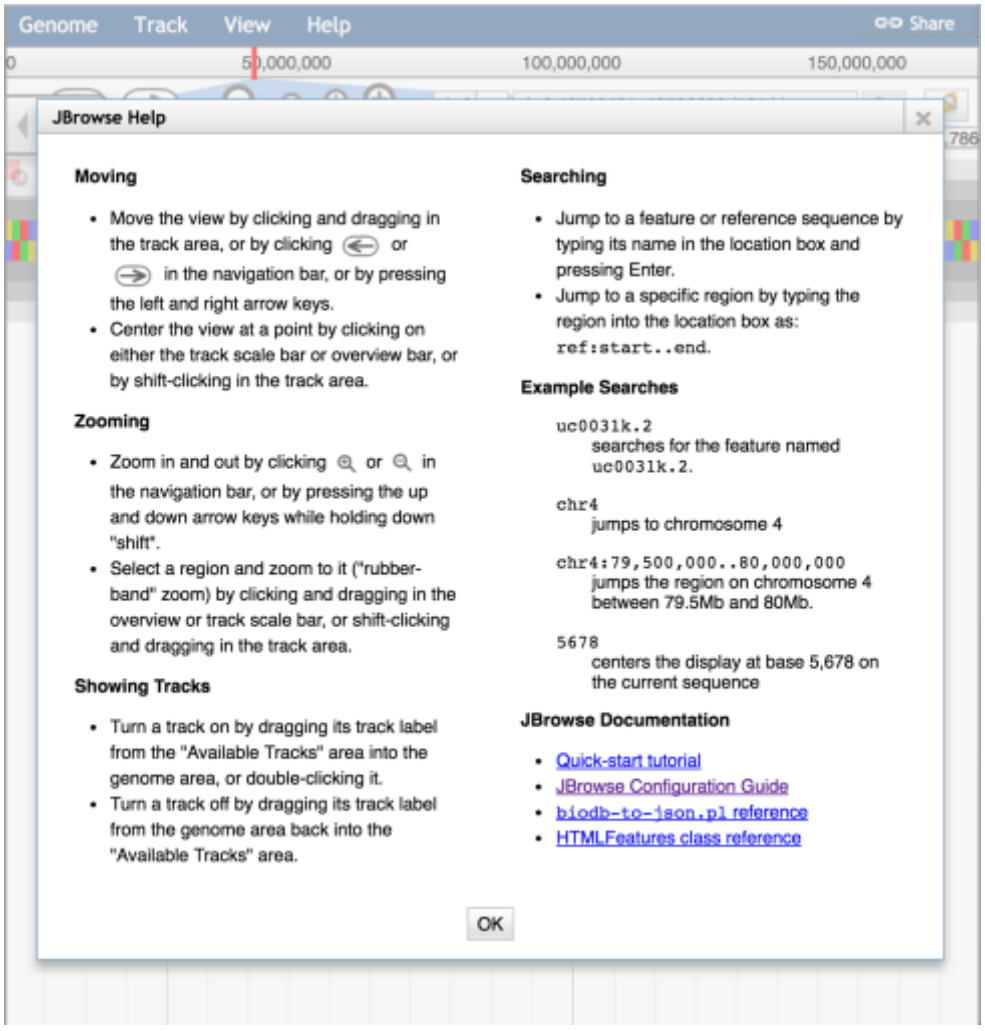
Navigating to the JBrowse genome browser on the RNAStructuromeDB will bring users to a page where RNA structural metrics can be viewed alongside Gencode gene annotations. This tutorial will walk you through how to make sense of the tracks on the genome browser and also provide some tips to customize JBrowse.

I. The Browser Interface

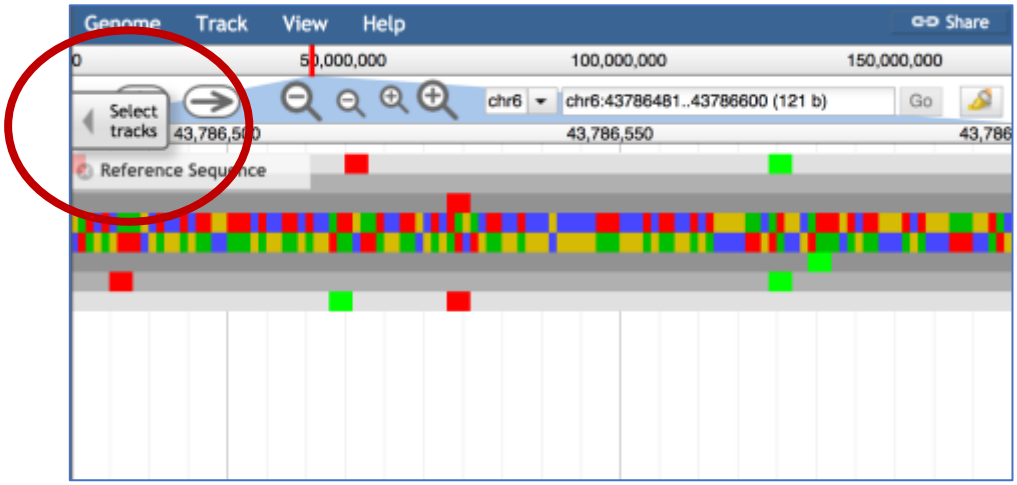
First of all, you should see some links at the top. A helpful link for beginners is the "**Help**" menu.



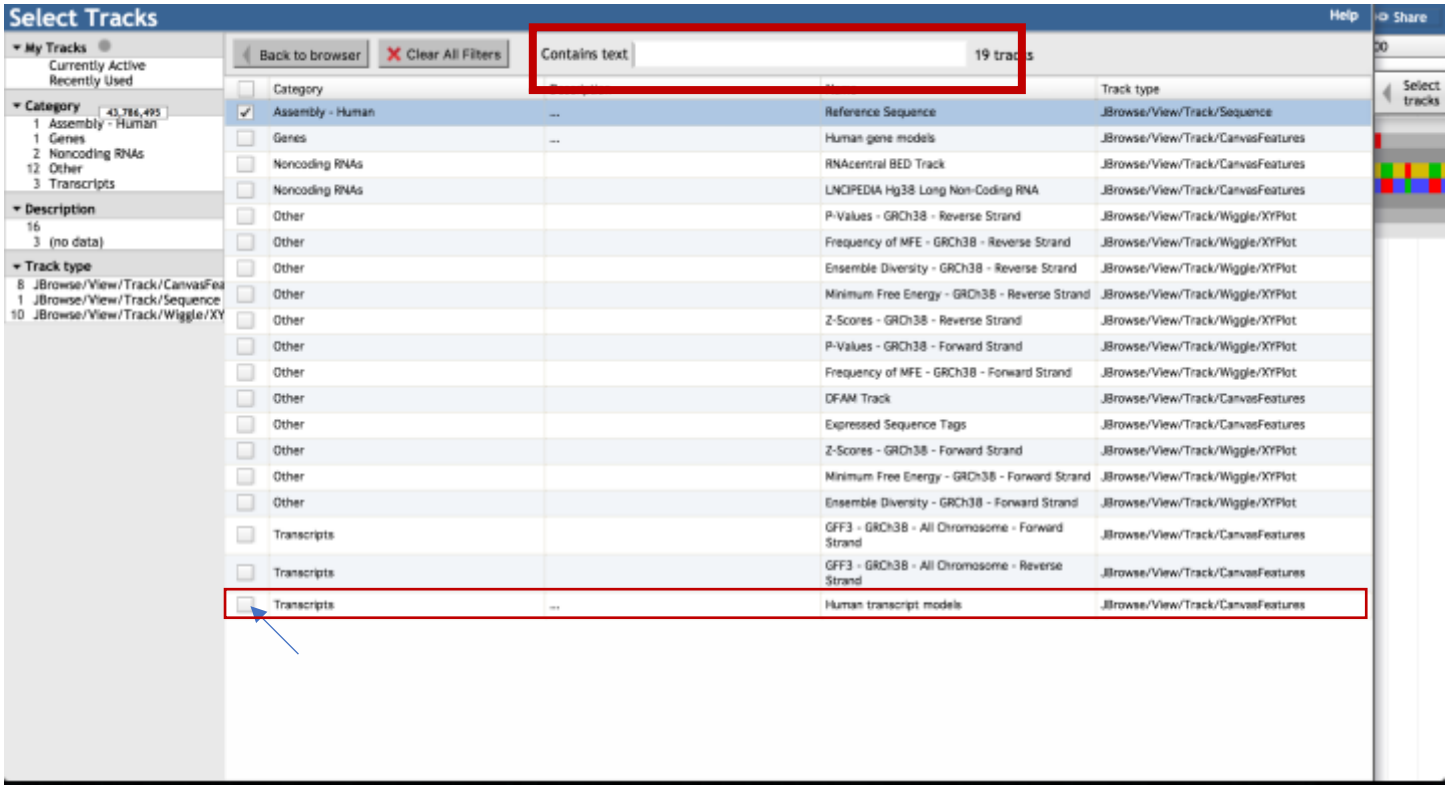
If you select Help → General, a handy pop up will show you the basics of "JBrowse-ing": tips on moving, zooming in, searching, selecting tracks, and some links to more in-depth configuration abilities.



The genome browse uses *tracks* of data. Each piece of data to be visualized can be found on its own track. Before we get into any detail describing these tracks, let us first **select tracks**.

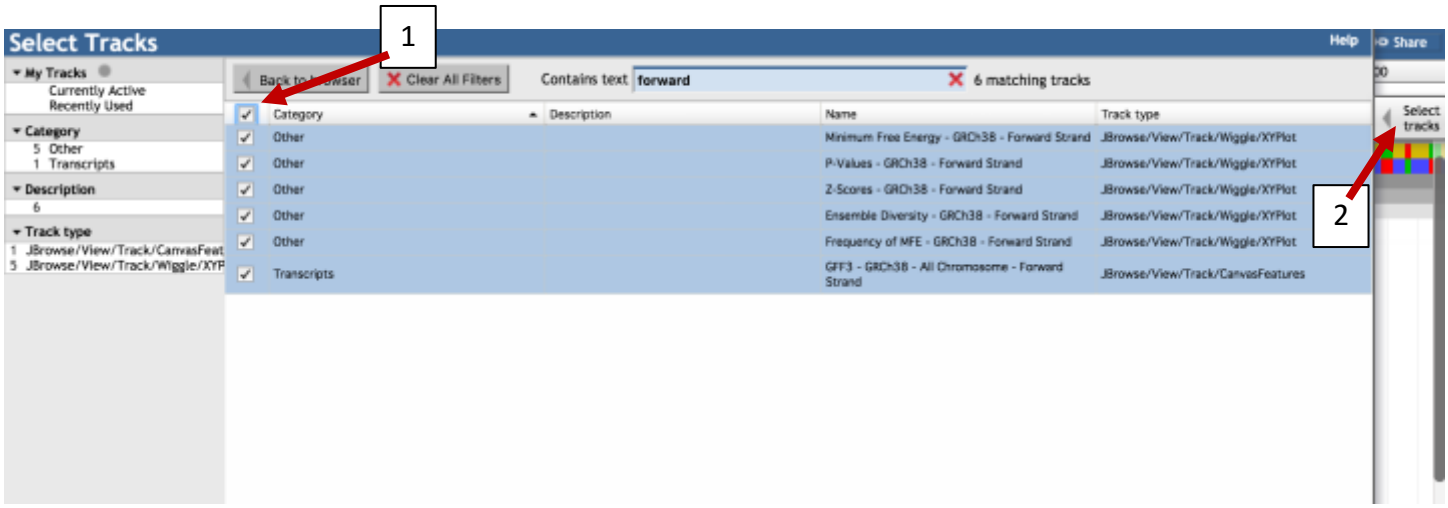


When you click on "Select Tracks" you'll see a new interface where you can select from a list of all possible tracks. (keep in mind that you can put your own tracks up here if you have any – in BED, BAM, GFF3, GTF, bigWig, or Fasta formats)



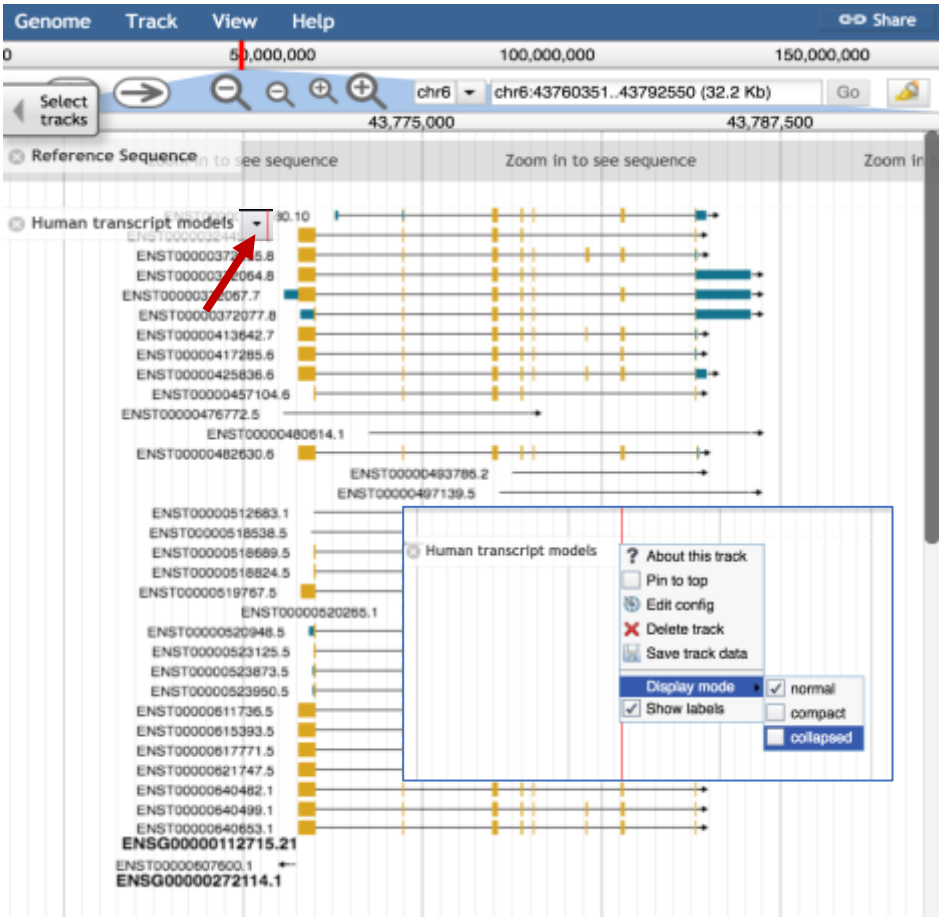
For this example we going to **create a "build"** for browsing the forward strand of the genome. First, we will be turning on the track which will show us all of our Gencode comprehensive gene annotations (version 26). Just click on the checkbox for the "Human transcript models". Clicking the box will turn on the track and you will see it on the browser once you go back in a second. Before that lets select all the forward strand metrics. A quick trick to do that: in the text box

at the top of the page, type "forward" and that will **filter** out all of the tracks that don't have forward in the title. You'll be left with this:



1. Just select the top check box and every "forward" track will show up in the genome browser.
2. Click on "Select tracks" again and you'll be brought back to the *forward strand genome browser.*

You should see something similar to this, depending on how zoomed in or out you are and where you are scrolled to on the page. **Tracks can be positioned however you'd like by dragging their title bar up or down the screen.** This first track holds **"Transcript models".** Here we see the gene VEGFA (Ensembl ID: ENSG00000112715.21). This particular gene has several alternative transcripts. Each one has a different model. The black lines represent **introns**, the yellow boxes represent **exons,** and the blue boxes are **3` or 5` untranslated regions.** You may also notice a small gene that looks like a tiny little arrow pointing to the left (Ensemble ID: ENSG00000272114.1). That's an "anti-sense RNA" gene. That makes sense, as it is shown as just an intron (no exons, UTRs, etc.). An "RNA gene" will not have UTRs.



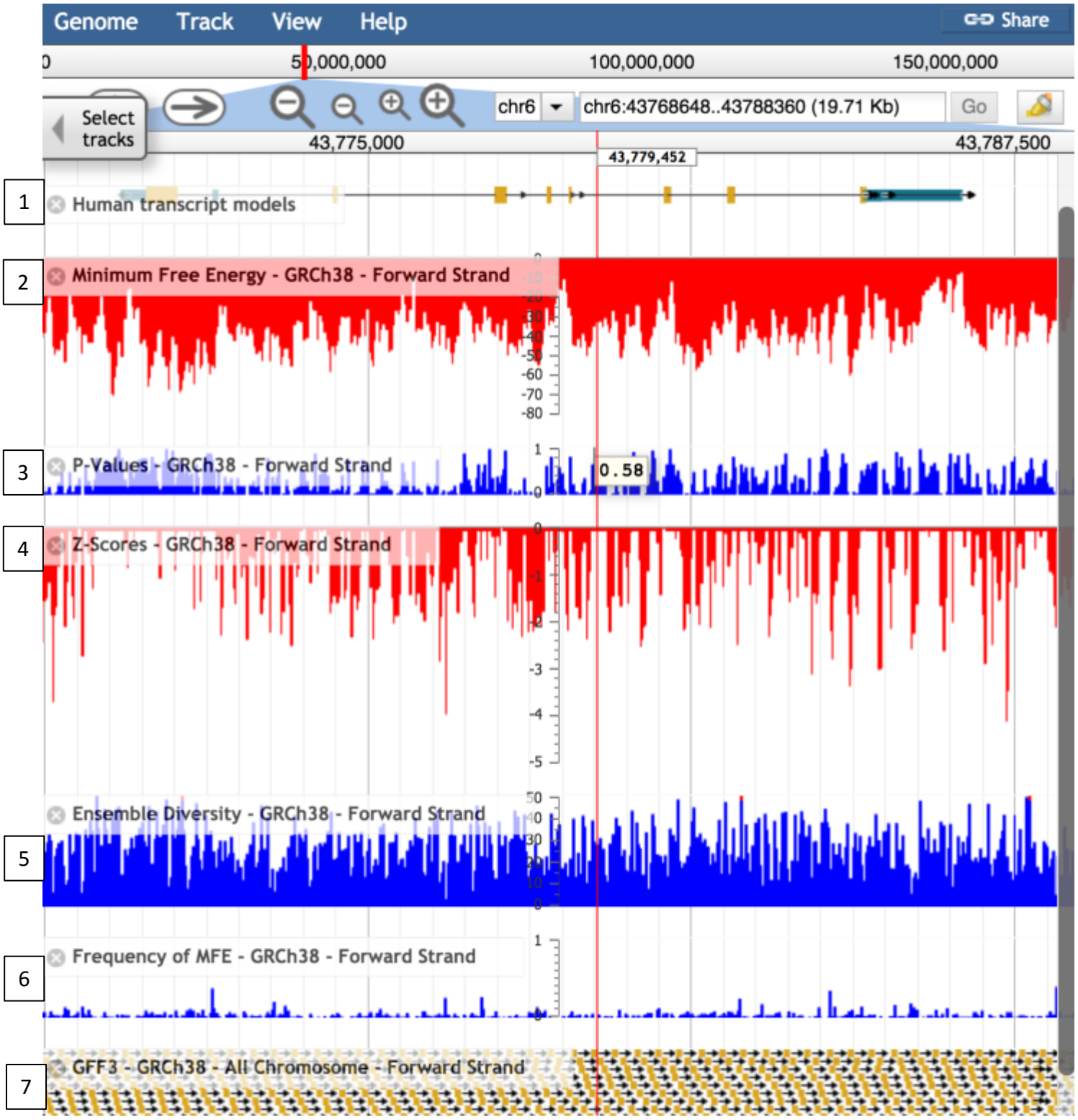
So transcripts will visually tell you a lot of information. Forward strand genes point forward, reverse strand points in reverse; introns, exons and UTRs all have their own visual models.

VEGFA has a lot of transcripts. There are so many that it takes up the whole browser. We can change that. If you click the "Human transcript models" bar where the red arrow is pointing, you will bring up the menu you see here to the left. **We can change the display mode to "collapsed"** which will collapse all of the models into a single version where all exons, UTRs and introns overlap.

Now we see something like the picture below. The tracks are as follow (descriptions on following page):

1. Human transcript models
2. Minimum Free Energy
3. P-Values
4. Z-Scores
5. Ensemble Diversity
6. Frequency of MFE
7. GFF3

Tracks can be moved around and positioned however you'd like. Drag them by their "title box" to a new position (higher or lower)



The RNAStructuromeDB consists of "windows" which give a glimpse into the underlying genome (hg38/GRCh38) sequence's ability to generate a structured RNA molecule. Since the genome is so large, over 3 billion nucleotides (nt), we first fragmented it into 154 million overlapping windows (120 nt long, stepping 40 nt). Each window was computationally transcribed into RNA and analyzed for its ability to then become structured. This "ability to form a structure" is described by five metrics which were calculated via ViennaRNA and a Perl script:

1. **Minimum Free Energy (track 2)**

The 120 nt sequence of RNA described in each window is first run through an RNA folding algorithm to predict the most stable structure it could theoretically adopt. A good description of how this works by Sean R. Eddy here.

The **Minimum Free Energy (MFE)** is the free energy value of this most stable structure in kcal/mol. The more negative the value, the more stable the structure.

2. **Z-Scores (track 4)**

It was found that, in general, structured RNA molecules adopted more stable (lower ΔG) structures than shuffled versions of the same sequence (as described by Clote et. Al., in **Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency)**. This was a phenomenon that would then be used to **detect regions of RNA which *could potentially* be structured**: as inferred by the thermodynamic **z-score**:

$$\boldsymbol{z-score} = \frac{MFE_{genome\ sequence} - \overline{MFE_{30\ scrambled\ versions}}}{\sigma}$$

This was calculated for each 120 nucleotide window of the genome. We took the sequence and scrambled it 30 times. We then have two sets of sequences: native and scrambled. Then, for each set MFE values were calculated. If the native sequence always has a much lower MFE than the *all* the scrambled versions this will lead to a negative z-score.

> ***Negative z-score* indicates a window generates a *more stable* structure than random.**
> ***Positive z-score* indicates a window generates a *less stable* structure than random.**

3. **P-Values (track 3)**

This value is directly related to the z-score. It is simply the fraction of random sequences which were more stable (more negative, or less than) the original sequence:

$$\frac{\#\ of\ MFE_{30\ scrambled\ versions} < MFE_{genome\ sequence}}{30}$$

Therefore a value of "1" indicates that all of the random sequences were actually more stable than the native and a value of "0" indicates the native was more stable than all random sequences.

4. **Ensemble Diversity (track 5)**

When calculating an MFE structure/value for a particular sequence, we are always finding a *single* result: the *minimum* free energy structure/value. However, theoretically, there are many possible structures/values. The ensemble diversity is a metric which attempts to describe the *variety* of possible structures. How is this determined? All possible structures are first calculated. The probability of the RNA adopting one of these structures is calculated using the partition function (described by McCaskill). This partition function can then be used to measure the "diversity" of possible structures. **If the structures are very similar (different by only a few base pairs) the ensemble diversity will be low, however, if there are a wide variety of structures possible, the ensemble diversity will be high.**

A helpful program to visualize the *character* and *quantity* of these alternative structures is EnsembleRNA. Here you can generate a helpful visualization to conceptualize the conformational landscape of an RNA molecule, and determine if a particular sequence has closely related or decidedly unique alternative folds.

Alternatively, if you have a structure which you would like to "lock" into a specific conformation, we have a tool, RNA2DMut, which can be used to identify single point mutations which can minimize ensemble diversity (or maximize it).

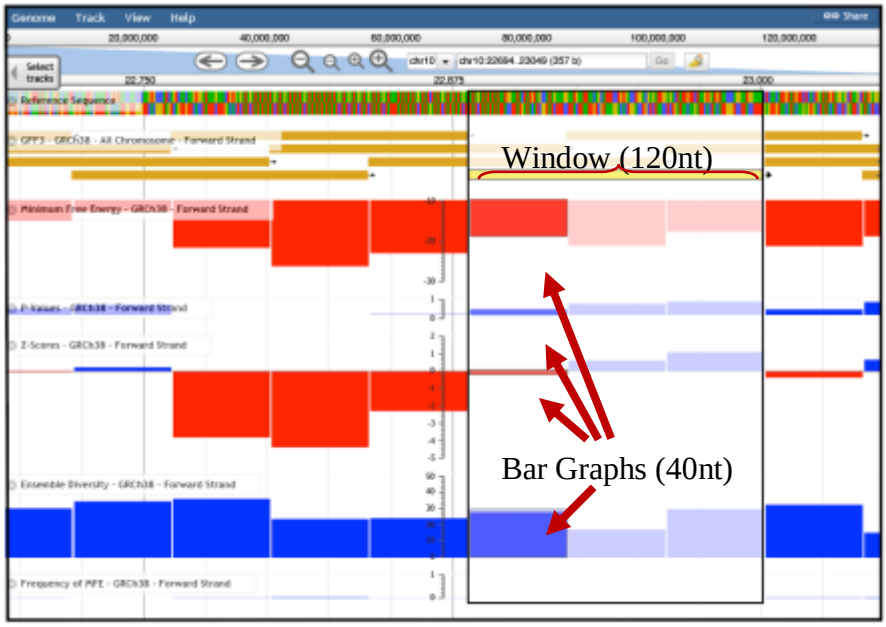
5. **Frequency of MFE** (track 6)

This value is related to the ensemble diversity in that it utilizes the same partition function calculations. Specifically, it measures the *frequency of the MFE structure* which was found within the calculated ensemble.

6. GFF3 – GRCh38 – All Chromosome – Forward Strand (track 7)

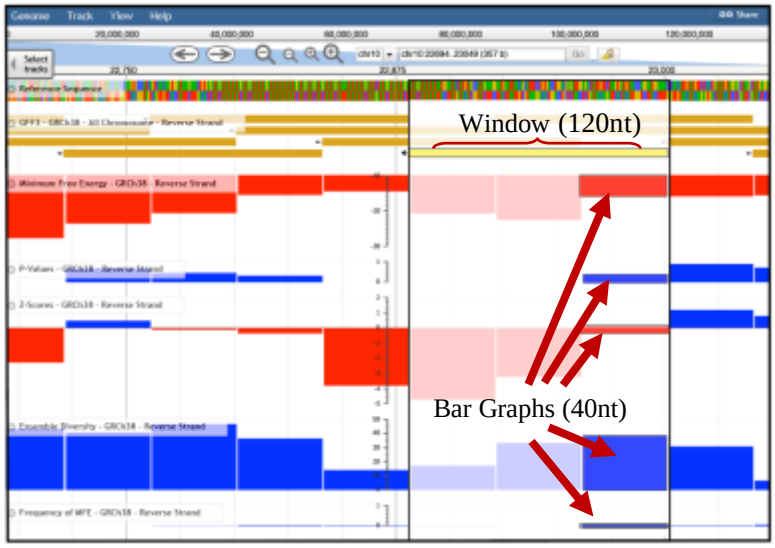
This track contains the raw data of our analysis, viewed as their individual windows. You will notice that the folding metrics are only shown 40-nucleotides at a time, but the windows are each 120-nucleotides long.

Forward Tracks

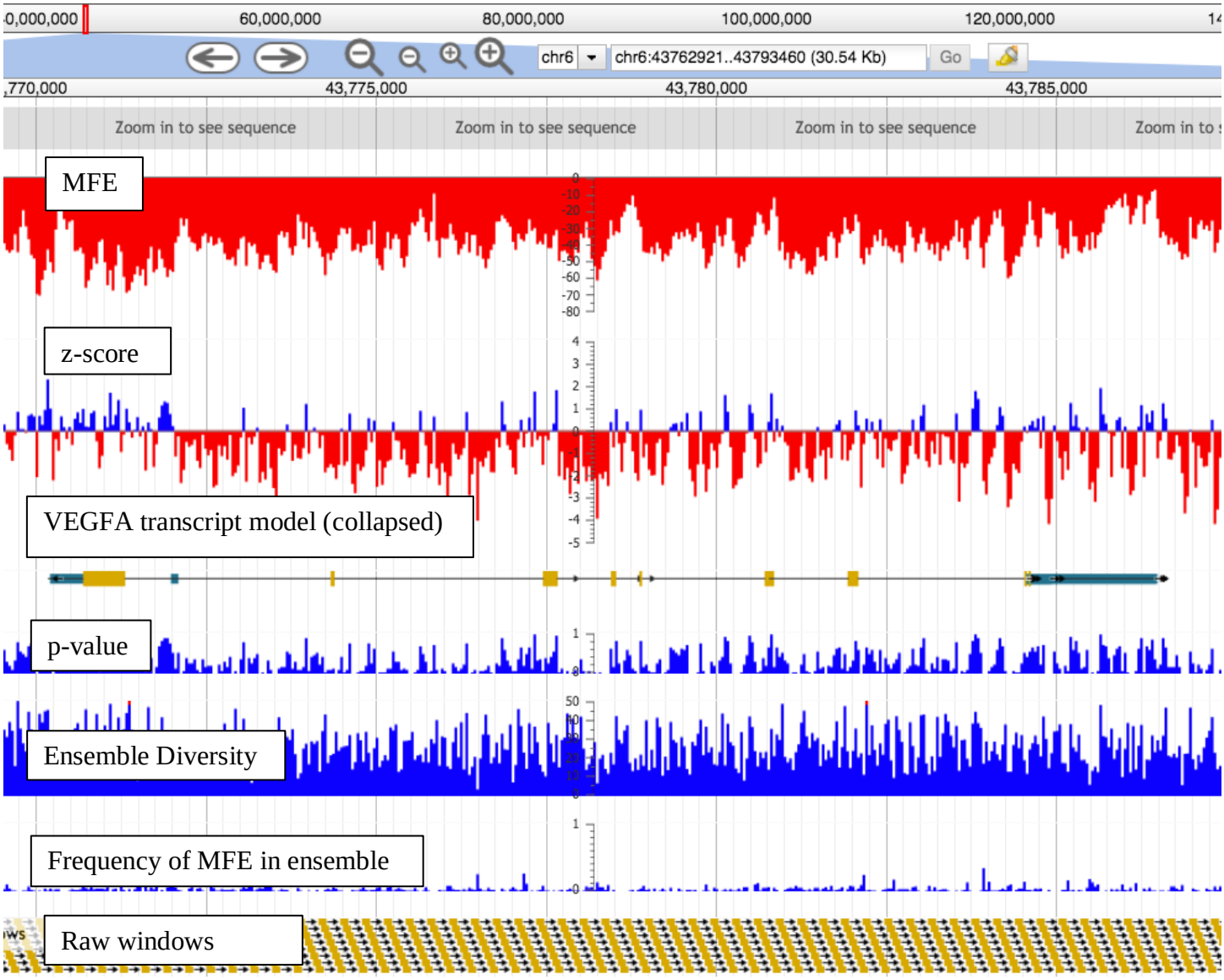


This image shows how a single window relates to each metric track (bigWig format). We highlight a single window (120nt) and its corresponding metrics (40nt). This was required to allow for each metric to be shown without obstruction by an overlapping window.
You can see for reverse strand we simply swap the orientation:

Reverse Tracks

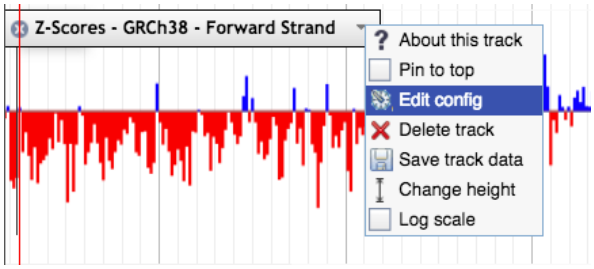


II. Finding Interesting Structures

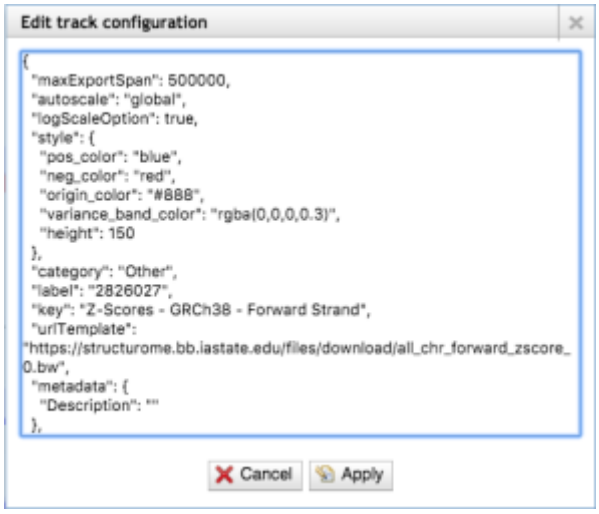


The best clue as to whether an RNA may be structured is its z-score: highly negative scores indicate the most likely candidates.

In the example above we are looking at the VEGFA RNA transcript model (collapsed visualization). You will notice several areas with negative z-scores. Each one of these could hold interesting structures. Let's try to find areas which are *significantly negative.* To do this we need to select the "edit config" option for our z-score track. First click on the dropdown button :
Then select "edit config".



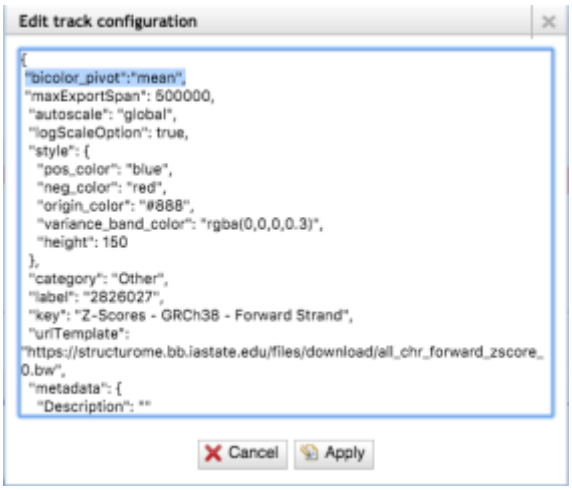
This will bring up an interface which will allow us to implement some cool tricks. This interface might seem a little daunting, but all we will be doing is typing in a single line of text which will make the ***significantly negative z-scores stand out.***



In the text box that popped up earlier input this text:

```
"bicolor_pivot":"mean",
```

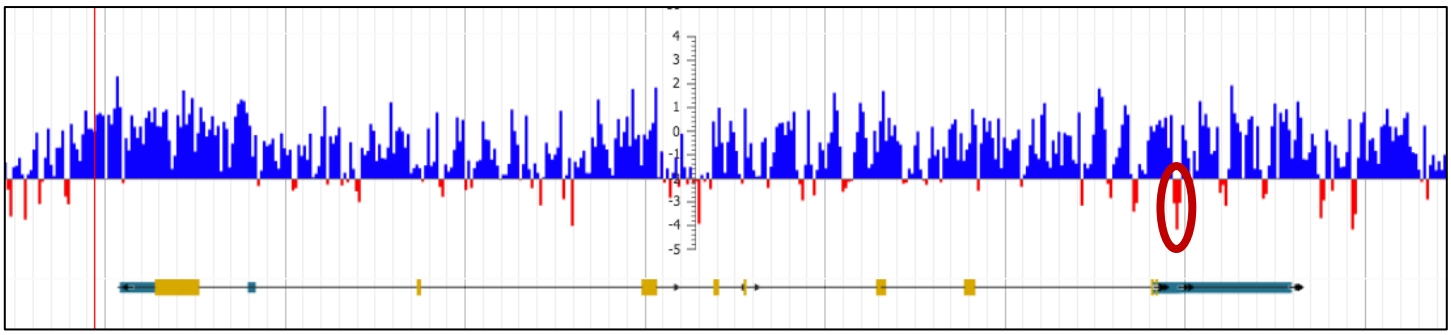
It is important that every character is input (the quotes, the colon, the comma). It should look like this:



Now just hit apply! And the portions of the z-score which are below the global mean will appear red! This can be done for every bigWig track too (if you'd like). Also, you don't have to just type "mean" you can use any value, for z-scores, if you want to be even more selective for what values show up as red type:

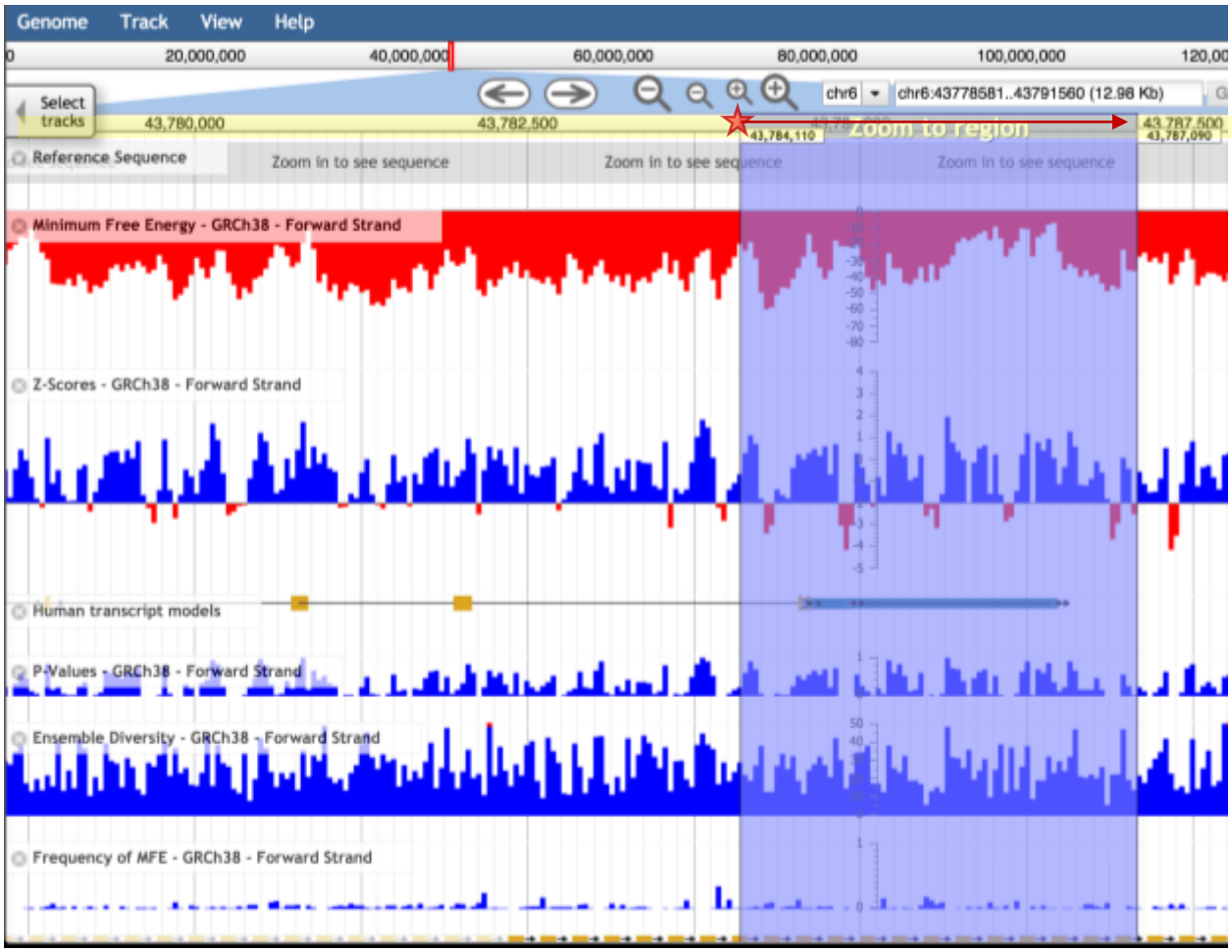
```
"bicolor_pivot":"-2",
```

Now it will be pretty clear to see the *very* negative regions of the VEGFA transcript. Of course, this doesn't mean that *only* these negative regions will be structured, but it is a good method for preliminary visual searches.

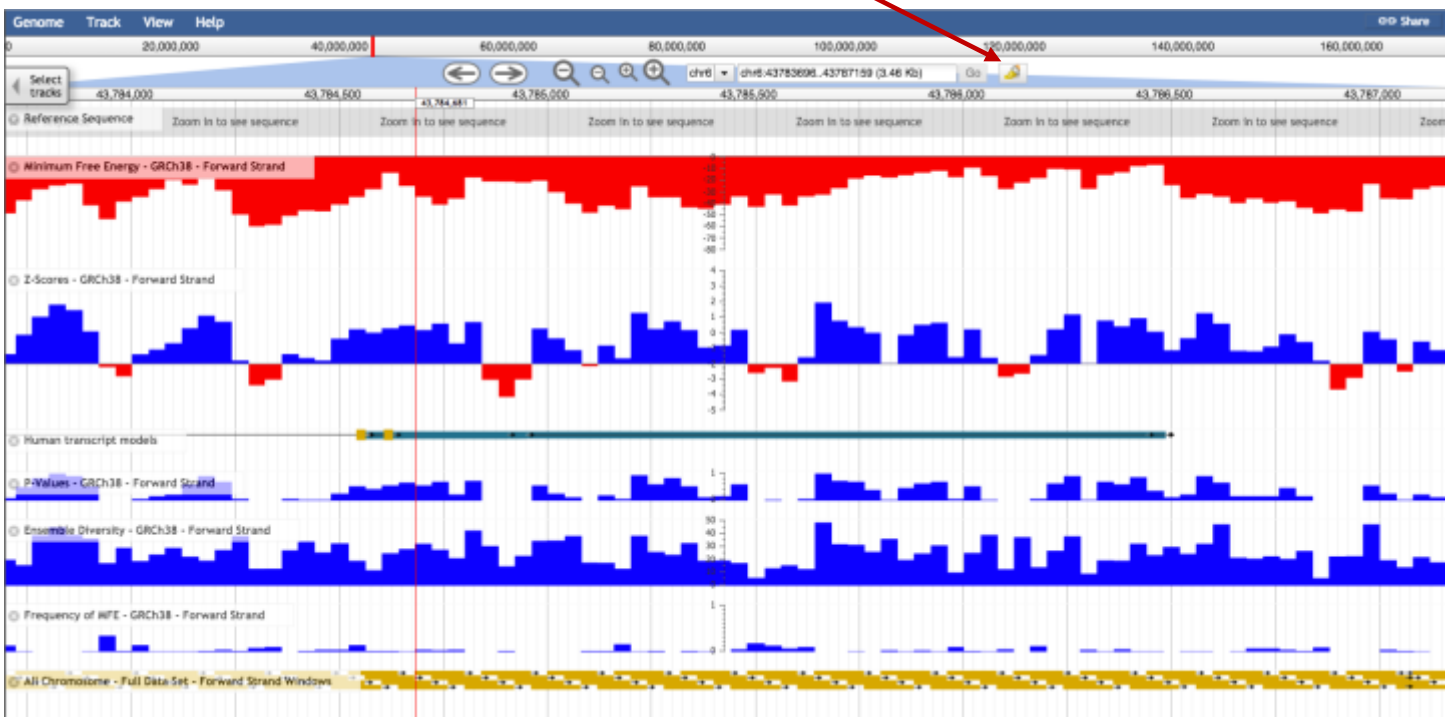


Each of these red (less than -2 z-score) regions contains sequences which produce RNAs ***two standard deviations more stable (lower MFE) than the randomized versions of the sequence.***

These regions then are *more likely* to contain genetic sequences which were selected by evolution to generate a structured RNA molecule. Let's **zoom in** on that circled region above. A quick, convenient, way to do that is by "click-dragging" your mouse cursor on the coordinate bar (we highlighted it yellow for this example). That simply means you click and hold on your starting point (star shape) and drag your cursor to the enpoint (in this case we drag to the right) then release the click. This will define the "Zoom to region" and JBrowse will now fill your browser with a closer look.

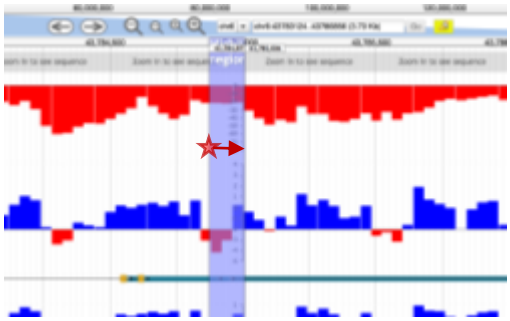


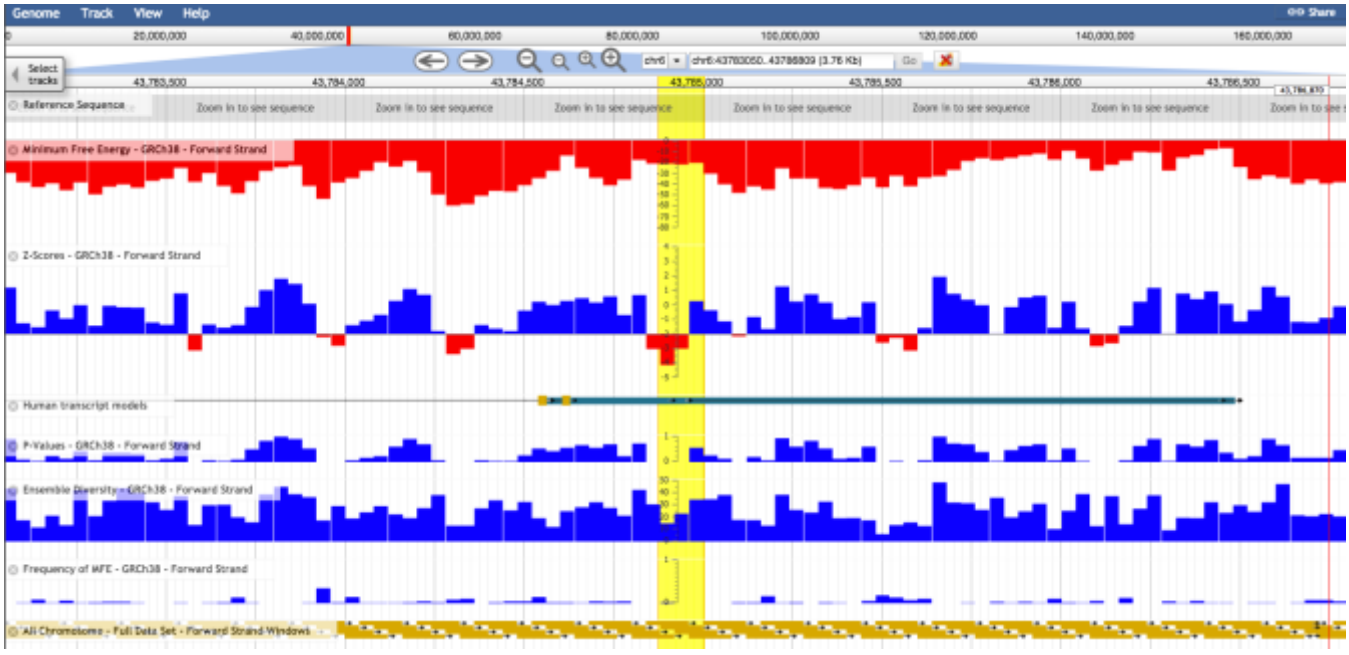
On this zoomed in region lets now **highlight** that low z-score window. Hit the highlighter icon



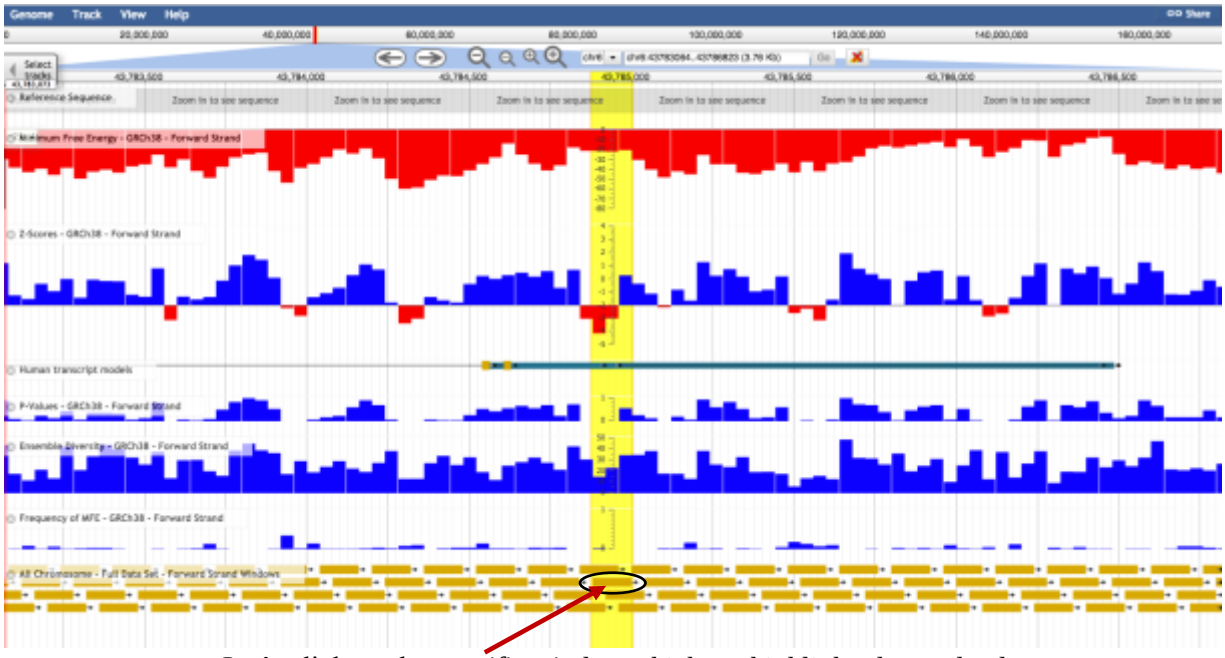
Now highlight the most negative z-score window:

Just click-drag your mouse cursor from the beginning of the lowest z-score, and extend the highlight box out to 120 nt (that is equal to 3 bars): *click at star and release at arrow.* This will highlight the specific window region which contains the low z-score.





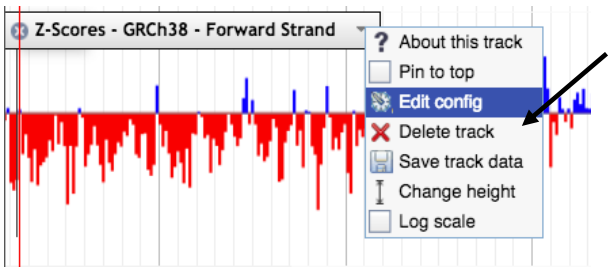
Now we can clearly **pick out the specific window** which generated this low z-score. Notice that the MFE isn't particularly negative (stable) for this window, but the ensemble diversity is lower than most other windows in the region, and the p-value is zero.



Let's click on the *specific window* which we highlighted completely.

We will now be served a pop-up with all of the window's metrics, including sequence and folding structure. This info can now be used to generate a preliminary structure model for this particular window, or used to conduct more investigations.

On our tools link https://structurome.bb.iastate.edu/tools you will find lots web-servers and programs which can be used to dig deeper once you find an interesting region of RNA. However, if you want to dig deeper on larger stretches of RNA (not just 120 nt windows) you can do that by:



1. **Downloading** data directly from JBrowse. Each track allows you to save data. You can save whole tracks, highlighted regions, or visible regions. You can also save in multiple formats, just click the "Save track data" and you will be served all of the options.

2. You can **search for and download data tables** for specific genes or for specific gene regions with our Data Search and Gene Search tools.

3. You can **download the whole data set** generated during the creation of the RNAStructuromeDB here.