**HOW TO USE A GENOME BROWSER: JBROWSE**

**Background**

**Genes and Proto-genes**

Before discussing what a proto-gene is, it is important to understand what makes a gene a gene. Genes are DNA sequences that code for functional proteins. Genes are transcribed into mRNA sequences which are then translated by ribosomes into proteins. Proteins in turn carry out many of the biological activities necessary for life. Most genes are very old. For example, humans and the budding yeast *Saccharomyces cerevisiae*, which we will study in this module, are separated by over 1 billion years of evolution, yet we share many genes inherited from our common ancestor. In some cases, the basic structures and functions of these genes have been preserved over a billion years or more. In other cases, the genes have changed over time, but they still belong to the same gene “family”.

Like genes, proto-genes are DNA sequences that are transcribed into mRNA and translated into proteins. However, unlike genes, proto-genes are very young in evolutionary terms. They have evolved very recently out of genomic sequences that were not translated before. These sequences may have pre-existing conserved functions as DNA (eg regulatory element such as TATA box or transcription factor binding site) or non-coding RNA, but the new proteins encoded by proto-genes exist usually in only a single species and are not part of gene families. Because of their recent emergence, proto-genes differ from genes in many respects. They tend to be much shorter and less transcribed than genes. While many genes have been functionally characterized by scientists, only a few proto-genes have been studied, and their role in biology is poorly understood. However, it is evident that some proto-genes have the potential to provide the organism with adaptations to different situations or environments. If a proto-gene provides advantages to the organism, it may be preserved by natural selection and evolve into a gene over time. New genes that evolved from a proto-gene are called “*de novo* genes” to denote that their ancestor was non-coding.

**Detecting proto-genes in the genome using translation**

When we look at the genome sequence, how do we know which parts correspond to genes or proto-genes? Because genes are usually conserved between many species, they can be detected by comparing genomes between different species and finding sequences that are similar between them. This is not possible with proto-genes. Instead, proto-genes are found by identifying translated open reading frames (ORFs: regions between a start and stop codon).

Translation of a gene proceeds from a start codon (ATG in the DNA sequence, or AUG in the transcribed mRNA) and ends at a stop codon (TAG, TAA, or TGA). The genomic sequence between each potential start and stop codon is called an open reading frame, or ORF. While each ORF defines a sequence with the potential to be translated, it may or may not actually be translated. For this, the sequence must be transcribed and the resulting RNA must reach the ribosome in the cytoplasm. To detect transcription at the genome scale, RNA-sequencing is used. Since transcription can change based on the environment that the cells are in, many scientists have used variants of RNA-seq to detect how ORFs are transcribed in different experimental conditions. To detect translation at the genome scale, ribosome profiling is used. Ribosome profiling is a sequencing-based technique that identifies pieces of RNA that are being translated by the ribosome and generates reads corresponding to the specific positions the transcripts come from in the genome. The number of reads at a position indicates the degree to which the transcript coming from that part of the genome associates with ribosomes. Though there are some complexities, you can generally interpret the number of reads at a sequence as corresponding to the amount that the sequence is translated. Using this technique, translation can be quantified across the genome and both genes and proto-genes identified.

In the yeast genome, looking for translation by ribosome profiling detects ~5,400 genes and ~19,000 proto-genes!

**Saccharomyces Genome Database and JBrowse Genome Browser**

The Saccharomyces Genome Database (SGD) is a database that contains annotated sequences and genes found within the genome of *S. cerevisiae*. It is available to the public and allows the user to look up annotated ORFs to find out information on them ranging from sequence to the expression of the ORF. SGD also has a genome browser, JBrowse, that enables the user to access data about any part of the yeast genome. This is provided through tracks that give quantitative or qualitative information across the genome. Many tracks are public because they come from experiments and analyses that have been published. This is great because it allows you to use available data to make new discoveries about your gene or proto-gene of interest! One can also use custom tracks, and integrate them however is useful to ask new questions and make more new discoveries. This module will use the SGD JBrowse genome browser to investigate proto-genes using public tracks and custom ribosome profiling tracks.

**Activity**

Note: This activity is designed to learn 1) how to use the browser 2) how to integrate across data types at the genome scale 3) the exciting vast amount of biology to discover through the study of proto-genes.

**Navigating SGD**

To begin, go to <https://www.yeastgenome.org/> which is the homepage of SGD. The search bar on the top left is what you will use to search for the gene of interest if you know the name of the gene. In the search bar, type in “YBR196C-A” and hit enter.

The page you are brought to is the all the published data and information on YBR196C-A. By reading the Description, you see that YBR196C-A evolved *de novo*, which means it is a proto-gene. It is one of very few annotated in SGD, because most proto-genes have not been studied yet. This one was included in the annotation in the mid-2000s because it was found to be transcribed by RT-PCR. By scrolling down, you are able to find information on it, like where it is located in the yeast genome, and you are able to download its sequence, for example. Now we are going to move on to how to use the customizable genome browser of SGD, JBrowse. Click on the top tab called “Sequence” and then click on “View In JBrowse” Then “go”. It should take you to a new tab which shows you annotated tracks of the yeast genome for the reference strain of *S. cerevisiae*.

**Navigating the Genome Browser**

At the top, you can see what yeast chromosome and where on the chromosome the browser is positioned. Scrolling to the left or to the right allows you to move down the chromosome to view different parts of the genome. You are also able to zoom in and out with the magnifying glasses with the plus or minus respectively. To change the chromosome you are viewing, click the drop-down box with chromosome and select a different one. You can also search for specific coordinates within a chromosome using the syntax: “chr”[chromosome number in roman numerals]”:”[start position of what you want to see]”..”[end position of what you want to see]. The main viewing screen with shows the reference tracks for *S. cerevisiae*, which shows annotated ORFs. The tracks have the ORFs labeled with their name and a short description of the ORF. Clicking on the ORF gives you an in-depth overview of the gene similar to the information seen when you search for it on SGD.

1. Use the information for YBR196C-A at SGD to find the chromosome number and coordinates. List those.
2. Go to genome browser and use this information to go to YBR196C-A in genome browser. What is the length of the ORF in base pairs? How many amino acids is this?
3. As shown in genome browser, is the YBR196C-A coding strand the + or the - strand?
4. What is the nucleotide sequence of the second codon for the YBR196C-A ORF?
5. What is the name of the ORF closest to the 3’ end of YBR196C-A? What is its function?
6. What is the name of the ORF closest to the 5’ end of YBR196C-A? What is its function?

**Using public tracks**

The power of the genome browser is that you can look at all sorts of data sets at once, that are all mapped to the same genome. Scientists around the world share the results of their experiments with SGD so we can look at them and make new observations about the genome.

On the left of your screen, click “Select Tracks”. You can see that there are over 400 different tracks of information you can use to explore any region of the genome! We will start by examining RNA expression. In the search bar, type “pelechano”. This is the name of the first author of a 2013 manuscript that looked at full length transcripts by identifying their 5’ and 3’ ends simultaneously. The experiment was performed in two different growth media: rich media (YPD) and galactose media (Gal) Select these tracks:

Longest\_full\_ORF\_transcripts\_ypd

Longest\_full\_ORF\_transcripts\_gal

Click “back to browser”. You now see new tracks in the browser displaying the results of Pelechano’s work. You may want to zoom out to orient yourself.

1. Compare the longest transcripts for YBR196C-A in the Gal and YPD conditions. What difference do you notice?
2. By zooming in and out, and observing other regions of the genome, you can get an idea of which phenomena are general and which are more specific. Did the difference you saw for the YBR196C-A transcripts in Gal and YPD seem like a general or a more specific phenomenon? Does this help you form a hypothesis?

There are many more public tracks for you to explore that give information about RNA expression, transcription factor binding and more. One of them gives information about evolutionary conservation. Go back to “select tracks”, press “clear all filters”, and in the selection column on the far left select “Assay term name” to be “phastcons”. Select the track:

UCSC\_Saccharomyces\_Clade\_Conservation

Then go back to the browser. PhastCons is a method for analyzing nucleotide conservation across species. Here the method was applied to the cerevisiae genome: a score was assigned to each nucleotide based on how similar it, and its neighborhood, are to sister species in the Saccharomyces clade. A score of 1 means that the nucleotide is highly conserved, 0 means not conserved. You can use this track to inspect evolutionary properties of any ORF, or any region of the genome that you are interested in. We will use it to look at YBR196C-A.

You may want to optimize your visualization. The tracks can be moved up and down by clicking on the name of the track and using drag and drop. You can change the height of the track by clicking on the drop down menu associated with the track name and selecting “change height”. It is helpful to increase the height so you can view the data better.

1. How does the nucleotide conservation of YBR196C-A compare relative to those of the closest 3’ and 5’ annotated ORFs?
2. Look at approximately 50KB at a time on any yeast chromosome. You can now see the conservation patterns of many ORFs at a time. What pattern seems most frequent, that of YBR196C-A, or of its closest 3’ neighbor, or if its closest 5’ neighbor? Why do you think this could be?
3. Go back to looking closely at YBR196C-A. You see that, inside the ORF, not every nucleotide has the same conservation level. Given that YBR196C-A is a proto-gene that evolved in a pre-existing non-coding DNA sequence, what might explain some of the highly conserved regions you observe? Zooming in close enough to observe the corresponding nucleotide sequence at the top of the browser may help you come up with specific hypotheses.

**Using custom tracks**

Another powerful feature of the genome browser is that you can integrate any of the public tracks with your own custom tracks. To illustrate this, you will be using a set of custom tracks that give information on proto-genes. To download these tracks, go to: <https://www.yeastorfanproject.com/summer-workshops/>

You can either click on sgdtrackfiles (2) or on the google drive link (recommended for a mac) to download a zipped folder. Unzip the folder. It contains 3 files:

BigWig riboseq full f

BigWig riboseq full r

GFF3 translated ORFs

To open the downloaded tracks in the browser, go to “Track” at the top of the screen and click “Open track file or URL.” Then go under “Local files,” hit “Select,” and select the downloaded files and click “open immediately”.

The first two tracks, riboseq\_full\_f and riboseq\_full\_r, give counts of ribosome profiling reads for each nucleotide position on the + and – strands of the yeast genome, respectively. Ribosome profiling is a sequencing-based technique for measuring physical association of a transcript with the ribosome. A high number of ribosome profiling reads indicates that the corresponding nucleotide is translated at a high rate. If you hover over the bars on these tracks, the number of reads will show up. As the number of reads can vary between 0 and several thousands, it is useful to view it in a log scale. To do so, hover over the track name, click the arrow on the right, and enable the “log scale” option.

The third track indicates ORFs across the genome that were found to be translated on the basis of the ribosome profiling data given in the other two files. If you click on the ORF in this track, it will give information on the ORF. This includes an orf ID in the “Name" section (for example, orf14869) as well as the ORF coordinates in the genome, its length, and a “score.” The score indicates confidence that the ORF is translated. The smaller the value, the higher the confidence; we are very confident that ORFs with scores below 0.01 are really translated. The Names here are assigned automatically; they are different from the SGD names (annotations). Most proto-genes are not annotated because they were recently discovered and don’t have an official name in SGD. The majority of the ORFs in this track are proto-genes, but not all. You can see a field in the ORF information card that tells you if an ORF is a proto-gene (TRUE or FALSE). This was determined by evolutionary analyses of the genomes of multiple yeast species.

1. Import the custom tracks you have been provided into the genome browser if you have not done so yet. Observe that YBR196C-A corresponds exactly to an ORF in the translated ORFs custom track. This means that this annotated proto-gene is translated by ribosomes according to ribosome profiling data. What is the id of the YBR196C-A ORF in the translated ORF track?
2. What is the confidence score?
3. Now let’s look at the ribosome profiling tracks. How many ribosome profiling reads correspond to translation of the YBR196C-A ORF? Remember that “f” corresponds to the + or “forward” strand and “r” corresponds to the reverse or - strand.

**Using the genome browser to learn about unannotated proto-genes**

We can use the custom track to look at ORFs that are translated but not annotated in SGD. Most of these ORFs have never been studied before! In the genome browser, locate the ORF at coordinates: chrVII:715196..715294 (or proto-gene of your choice!!)

1. Indicate the coordinates and the name of the unannotated ORF you are looking at:
2. What is the name of the annotated ORF closest to this unannotated ORF?
3. What is the highest ribo-seq read count at any position corresponding to translation of the annotated ORF?
4. Does this ORF appear to be translated more or less than YBR196C-A?
5. Should we be more or less confident that this ORF is genuinely translated than that YBR196C-A is? Why?

Looking at large sections of chromosomes or large numbers of example genes and proto-genes can give us a broader view of translation, and genome organization and complexity in general. Either by zooming out using the “-” sign, or by entering chromosome coordinates of your choice in the search bar, visualize ~50,000 basepairs at once. Zoom in and out, and click on the ORF names on the custom track, to learn general patterns.

1. Do annotated ORFs tend to be translated more or less than unannotated ORFs?
2. Does it appear that ribo-seq reads are spread evenly across the ORF, or are there spatial patterns and trends you observe? For example, do there tend to be more reads early in the ORF sequence or late?

We can now use the public tracks to learn more about the translated ORFs in these custom tracks. Go back to the proto-gene at coordinates: chrVII:715196..715294.

1. Is this ORF evolutionarily conserved or not? You can answer this question using the public phastcons track UCSC\_Saccharomyces\_Clade\_conservation.
2. Using the public Pelechano tracks, compare the transcript architecture around the ORF in Gal and YPD. Do you see a difference?
3. Recall the observations you had made regarding the YBR196C-A proto-gene’s conservation and transcript expression in Gal and YPD, and how similar/different you thought they looked compared to most annotated ORFs. How similar /different are the conservation and transcript expression of the ORF?
4. What hypothesis (hypotheses) are you able to formulate based on all of your observations throughout this module?