

# HOW TO USE A GENOME BROWSER: JBROWSE

## Introduction

### 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 transcripts 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 (e.g. 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 sequencing is used. Ribosome sequencing is a 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 sequencing 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.

This guide will walk through investigating an example proto-gene using a genome browser. Follow along and answer the questions in the accompanying worksheet for your proto-gene.

Goals: In this module, participants will learn:

- 1) how to use the genome browser
- 2) what ribo-sequencing is and how it is used to identify proto-genes
- 3) how to integrate across data types at the genome scale
- 4) the exciting vast amount of biology to discover through the study of proto-genes.

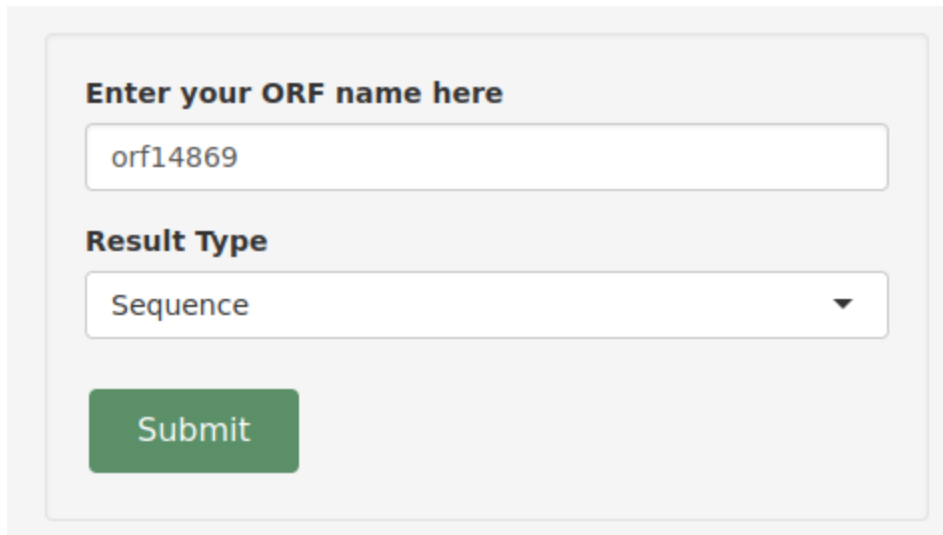
Objectives: After completing this module, participants will be able to:

- 1) Locate an ORF on JBrowse and explore neighboring ORFs
- 2) Select and manipulate JBrowse tracks
- 3) Use USCS-conservation to determine the conservation of a given ORF and flanking regions within the *Saccharomyces* cade

# Activity

## Navigating the Genome Browser

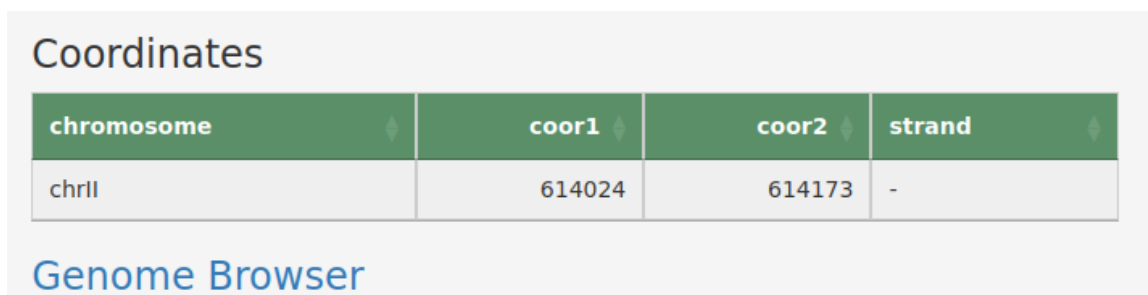
Find out the location of your proto-gene in the genome by going to the ORF information website: <https://carvunislab.csb.pitt.edu/shiny/coexpression/>. Type in the ORF name/id for your proto-gene and in the 'Result Type' drop down menu select 'Sequence'.



The screenshot shows a web form with the following elements:

- A heading: "Enter your ORF name here"
- A text input field containing "orf14869"
- A heading: "Result Type"
- A dropdown menu with "Sequence" selected and a downward arrow.
- A green "Submit" button.

**Answer Question 1 in your worksheet:** What is the chromosome number, start and stop coordinates, and strand for your proto-gene?



The screenshot shows a table with the following data:

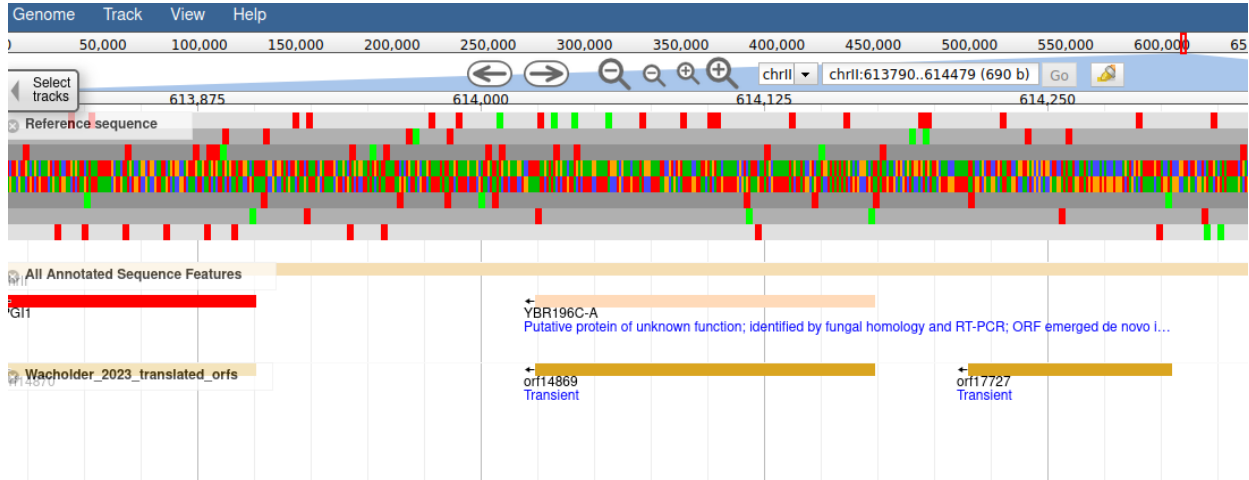
chromosome	coord1	coord2	strand
chrII	614024	614173	-

Below the table is a blue link labeled "Genome Browser".

Click on the 'Genome Browser' link that is found below the coordinates table. This will take you to a website hosted by SGD called Jbrowse that allows you to visualize where your proto-gene is in the genome relative to other genes.

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.

Each row in the browser is referred to as tracks, and represents the different types of data you can look at. The top track, labeled 'All annotated sequence features', shows all annotated sequence elements for *S. cerevisiae* (i.e. genes, tRNA, chromosomes etc).



If you click on a gene in this track, a new window will pop up with more information about the gene including its length, sequence and function.

The track below, labeled `Wacholder_2023_translated_orfs`, shows both annotated genes and proto-genes. Therefore you will see annotated genes in both tracks, however in the '`Wacholder_2023_translated_orfs`' track they are denoted by their 'ORF id' instead of gene name. For example, In the below screenshot YBR196C-A is an annotated gene listed in the 'All annotated sequence features' track and is also represented in the '`Wacholder_2023_translated_orfs`' track as `orf14869`.

Clicking on an ORF in this track will give you information about the ORF's length, sequence, evolutionary status and score. Specifically the '`is_transient`' attribute will be '`TRUE`' if the ORF is evolutionarily novel, and will be '`FALSE`' if the ORF is conserved. The `Score` attribute tells you the confidence level that the ORF is translated.

**Answer question 2 in the worksheet:** What is the length of your proto-gene in base pairs? How many amino acids is this?

**Answer question 3 in the worksheet:** Is your proto-gene on the + or - strand?

**Answer question 4 in the worksheet:** What is the nucleotide sequence of the second codon for your proto-gene?

**Answer question 5 in the worksheet:** What is the name of the annotated gene closest to the 3' end of your proto-gene? What is its function? Is this gene on the same strand as your proto-gene? (If your proto-gene is on the + strand, 3' is the region to the right of your

proto-gene, otherwise if your proto-gene is on the - strand, 3' is the region to the left of your proto-gene.)

**Answer question 6 in the worksheet:** What is the name of the annotated gene closest to the 5' end of your proto-gene? What is its function? Is this gene on the same strand as your proto-gene? (If your proto-gene is on the + strand, 5' is the region to the left of your proto-gene, otherwise if your proto-gene is on the - strand, 5' is the region to the right of your proto-gene.)

## Using other public tracks

The power of the genome browser is that you can look at all sorts of datasets 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, including the translation, expression and conservation of your proto-gene.

### Translation

The first two tracks, 'Wacholder\_2023\_riboseq\_plus', 'Wacholder\_2023\_riboseq\_minus', give counts of ribosome sequencing reads for each nucleotide position on the + and - strands of the yeast genome, respectively. Ribosome sequencing is a sequencing-based technique for measuring physical association of a transcript with the ribosome. A high number of ribosome sequencing 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.

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. 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?

In general, do annotated ORFs tend to be translated more or less than proto-genes?

If you click on your proto-gene in the 'Wacholder\_2023\_translated\_orfs' track there is a reported 'score' which indicates the confidence level 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. What is the translation confidence score for your proto-gene?

**Answer question 7 in the worksheet:** Does it appear that ribo-seq reads are spread evenly across your proto-gene or are there spatial patterns and trends you observe? For example, do there tend to be more reads early in the proto-gene sequence or towards the end?

**Answer question 8 in the worksheet:** In general, do annotated genes tend to be translated more or less than proto-genes?

**Answer question 9 in the worksheet:** What is the translation confidence score for your proto-gene?

## RNA expression

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, which is when genes are activated and are transcribed into mRNA transcripts. In the search bar, type “pelechano”. This is the name of the first author of a 2013 manuscript that looked at full length mRNA transcripts by identifying their 5’ and 3’ ends simultaneously. The experiment was performed in two different environmental conditions: either the yeast was grown in a rich media containing glucose (called YPD) or grown in a media containing galactose (called Gal). Under different environmental conditions organisms will express (or activate) different genes. Using this RNA expression track we can look to see if our proto-gene is expressed in either of these two environmental conditions.

Select these tracks by checking the box for each:

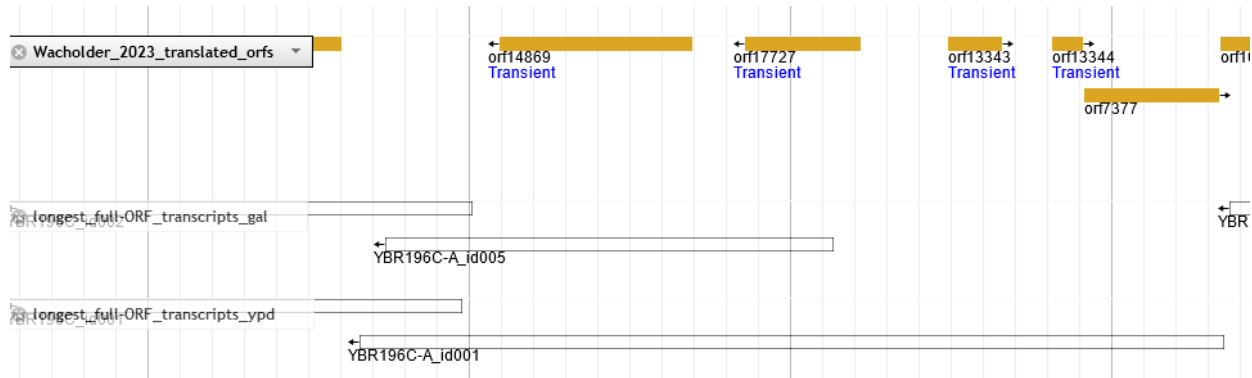
Longest\_full\_ORF\_transcripts\_ypd

Longest\_full\_ORF\_transcripts\_gal

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

Do you find transcripts covering your proto-gene in either YPD or galactose? Do you notice any difference between the two conditions? Note that this track has been prefiltered to contain only transcripts that contain annotated genes. Therefore your proto-gene could also exist on a transcript by itself, it just won't be displayed in this particular dataset. Also pay attention to the direction of the transcript.

**Answer question 10 in the worksheet:** Do you find transcripts covering your proto-gene when yeast is grown in either YPD or Gal? Do you notice any difference between the two conditions?

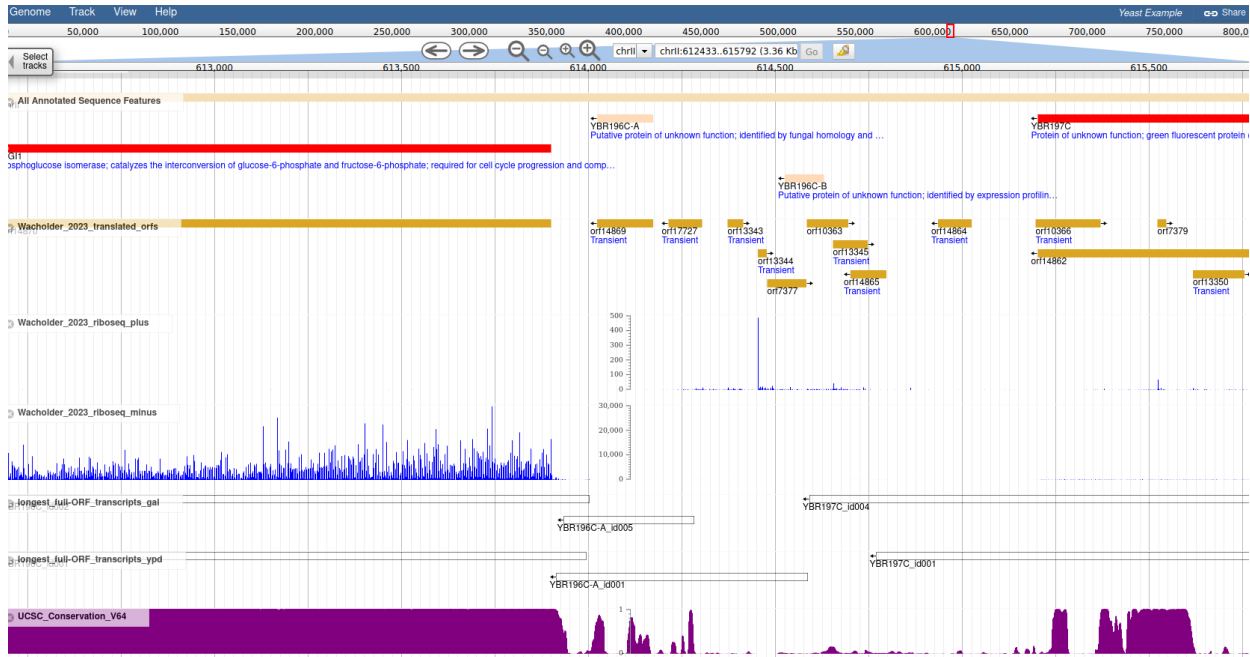


## Nucleotide conservation

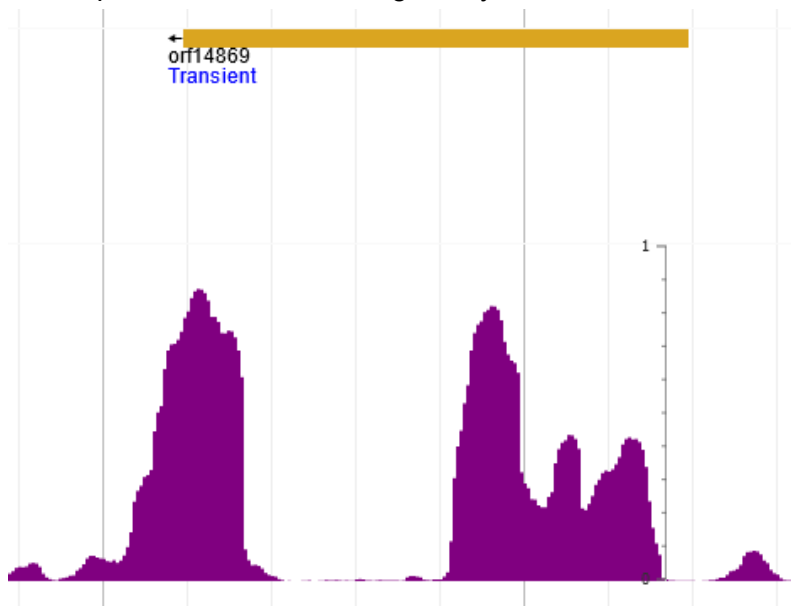
There are many more public tracks for you to explore that give information about RNA expression, transcription factor binding and more. One of them gives information about evolutionary conservation. Go back to “select tracks”, press “clear all filters”, and in the search bar type “phastcons”. Select 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. If you run your mouse over the track it will give you the conservation score for each nucleotide.



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.



Compare the nucleotide conservation of your proto-gene relative to nearby genes.

**Answer question 11 in the worksheet:** Are there certain regions of your proto-gene that are more conserved than others?



**Answer question 12 in the worksheet:** How does the nucleotide conservation of your proto-gene compare relative to those of the closest 3' and 5' annotated genes?