TheCellMap, BLAST and CLUSTAL

Danielle A. Presgraves

1 Introduction

Today's lab has three different parts:

1. First, we are going to identify genes that interact with **HMO1** using the tools found at TheCellMap: `http://thecellmap.org`
2. Second, we are going to investigate and use the Basic Local Alignment Search Tool (BLAST) database that is found at `http://blast.ncbi.nlm.nih.gov/Blast.cgi`. This program allows us to *align* multiple sequences, either amino acid or nucleotide, from individuals within populations and between species. Once sequences of loci are aligned, we can proceed to use the areas where they are conserved and where they are not conserved to infer meaningful evolutionary relationships. These relationships include paralogs, to find these we will use the synteny viewer, and orthologs, to find these we will use the program CLUSTAL.
3. Lastly, you are going to learn how to design PCR primers to amplify genes for sequencing.

2 TheCellMap

There are approximately 6000 genes in the *Saccharomyces cerevisiae* genome. Astonishingly only about 20% ($\approx$ 1200) of them are required for viability! This fact suggests that the *Saccharomyces cerevisiae* genome has built-in a lot of buffering against genetic and environmental perturbations.

The creators of TheCellMap have leveraged the small number of 'essential' genes by building an array that matches every single knockout (of the 4800 non-essential genes) or temperature-sensitive (tl) conditional mutant (of the 1200 essential genes) with every other every single knockout (of the 4800 non-essential genes) or temperature-sensitive (tl) conditional mutant of the 1200 essential genes. Since we already have information about the phenotype of the single mutants, we have an expectation about the resulting **double** mutant phenotype. We can then measure, quantitatively, the difference between our expectation, based on the multiplicative effects of the single mutants at each site, and the observed phenotype of our double mutant.

Qualitatively, the **double** mutant phenotype can be either **negative**, in which case the **double** mutant phenotype is more extreme than the single mutants ie. synthetic lethal or they can be **positive** when the **double** mutant phenotype exhibits less severe a phenotype than the single

mutants. **Negative haploid double mutants** can potentially help identify functionally related genes that are in the same biological pathway. **Positive haploid double mutants** can help us identify genes that contribute to the same non-essential protein complex (if one gene mutation already causes a deleterious effect, the second mutation doesn't make it worse) ie. cell cycle progression genes and proteostasis (`https://en.wikipedia.org/wiki/Proteostasis`) tend to be positive double mutants. The negative and positive interactions, known as the **genetic interaction profile**, groups genes into a hierarchical model of cellular function. Phenotypic measurements (all 23 million of them) were done via computer and assume that colony size is an accurate proxy for cell fitness. There were $\approx$ 550000 negative double mutants and $\approx$ 350000 positive double mutants in this screen.

In general, the genome-wide systematic coupling of single mutants allows us to screen for genetic interactions in order to:

1. explore the "buffering capacity" of the genome
2. quantify the "wiring" or genetic architecture of the genome
3. gain insight into genotype $\longrightarrow$ phenotype relationships, which are not as obvious as we like to believe
4. predict functions of previously uncharacterized genes by their interactions with known genes.

2.1 What genes interact with HMO1 and how?

We want to know all about the gene HMO1. We will begin by going to the website: `thecellmap.org` and typing in HMO1 into the search engine in the upper left corner of the screen. This will result in a complete map of the all the interactions of the genes in the *Saccharomyces cerevisiae* genome with a little red teardrop that gives you the location of your queried gene. The position of our gene-of-interest in the network of 4909 nodes (made up of 4418 unique genes) and $\approx$ 34,500 edges already tells us information such as which of the 19 general biological process clusters is our gene involved in?

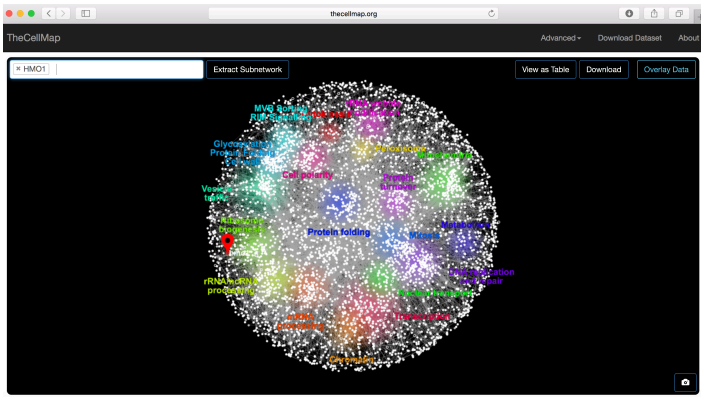


Figure 1: Querying HMO1 as part of the constellation of all genes and interactions

We can pull out the particular gene and all the genes associated with it by clicking on the "Extract Subnetwork". There are a few things that you should notice on this diagram: The "Profile Similarity Network" which is where we will start and the Genetic Interaction Network". You should also note the "Annotate Network" button and the sliding scale on the right side that has a default value of 0.2. This default value is, fundamentally, just a Pearson's correlation coefficient and the default value of 0.2 was decided based on inclusion/exclusion criteria. You can increase or decrease the value. If you increase it, you will have fewer genes interact with HMO1 and if you decrease it, you are allowing for more genes that have less evidence for interactions to be included in your scan.

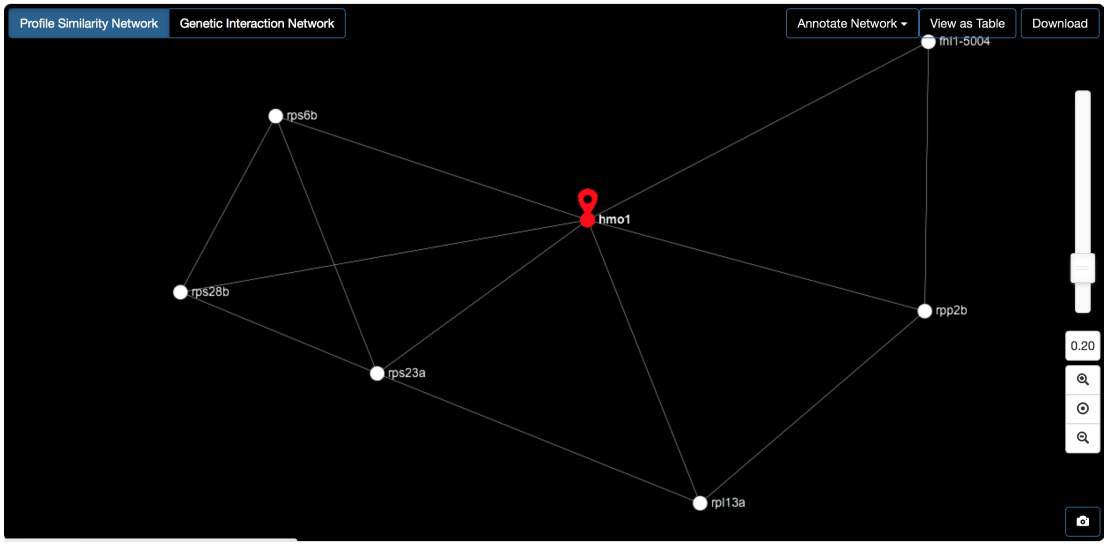


Figure 2: Isolating *just* the interactions of HMO1

We can also see interactions between two particular genes. If we type HMO1 and YGR118W (one of the 6 included genes in our search) into the search box and press "Genetic Interaction Map",

you get the following:

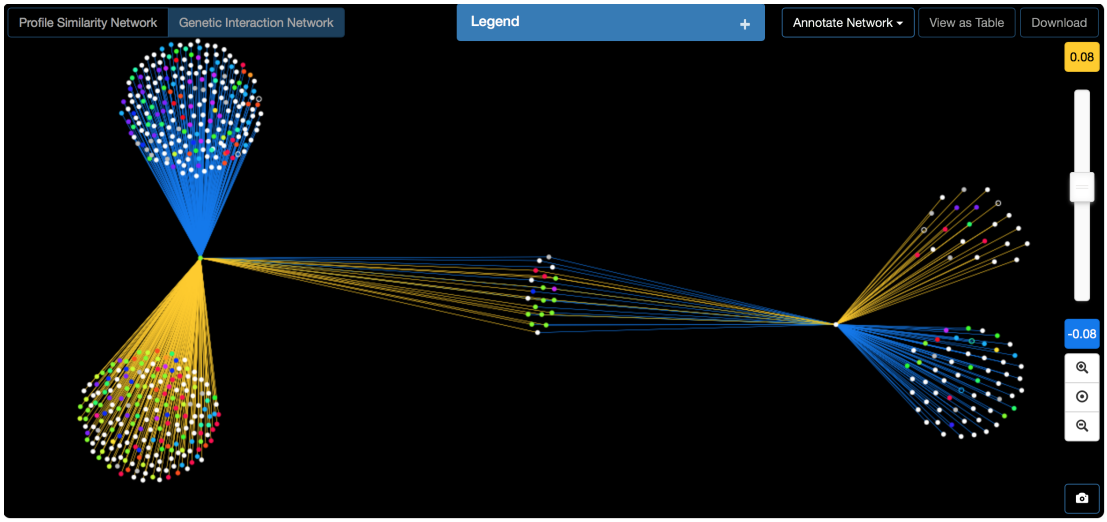


Figure 3: Interactions between HMO1 and YGR11W

We now want to know what types of biological processes the genes that are associated with HMO1 do. We can choose various types of annotation from the pull down annotation menu but let's use SAFE since that method was what produced the original different colored regions of the entire genome interactions. The SAFE method identifies dense network regions associated with specific functional attributes. We can see all of the interactions by clicking on the table pull down menu. This give us the Pearson's correlation coefficient for the 6 included genes and, if we want, all the rest of the other possible interactive genes with smaller PCC values.

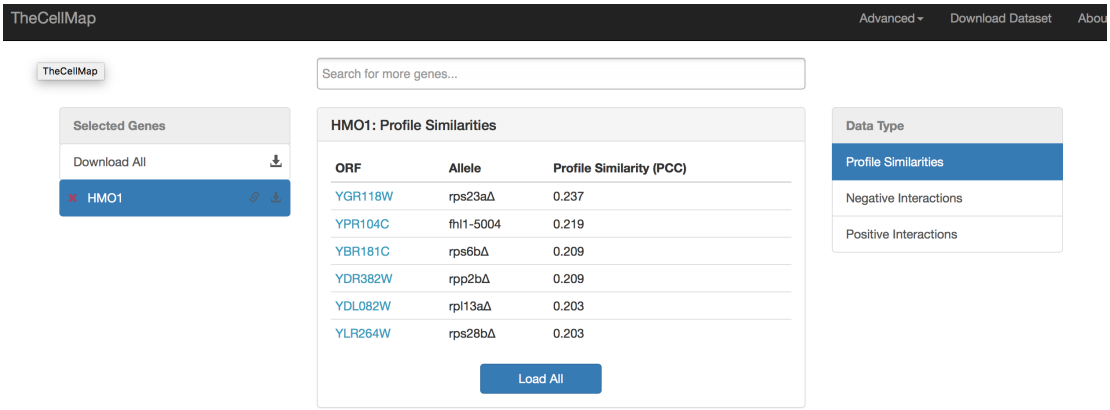


Figure 4: The PCC values of the interacting genes. The Negative and Positive interactions will be listed as well

Now that we have annotation on the 6 genes that are connected to HMO1, let's double click on one of them. The double clicking will take us the `yeastgenome.org` and we will focus on the gene information.

3 What is BLAST?

BLAST identifies sequences in the database that are similar to the one queried in the search bar and assigns a score, "E-value" score, to how similar sequences are. E-value represents how well the query sequence matches the database sequence, taking into account factors such as the number of matching residues and the total length of the alignment. The lower the E-value, or the closer it is to zero, the better the match is. The larger the E-value is, the more likely that any similarity between sequences has arisen **simply due to chance**. Similarity between alignments is particularly important since it can allow us to infer evolutionary relationships between the individual sequence and species that we are comparing. Differences between sequences are also informative since they can allow us to identify changes, mutations (usually) but also re-arrangements, that are specific to one or more lineage. A reliable alignment algorithm is so useful that BLAST has become a ubiquitous gerund in biology labs and literature: "Start by BLASTing it!" or "Just BLAST it!".

Note: As you might guess, the ability to use sequence data as a quantifiable evolutionary character has resulted in a handful of user-friendly websites that provide a repository of links to all sorts of bioinformatics tools, one of which is BLAST. One such site is: `https://omictools.com`. It is useful to appreciate that this type of resource exists, especially for your scientific interests outside of this particular course, but, for now, we will use the direct link to the BLAST algorithm.

As you can see from the website, there are several BLAST programs that allow you to search using different pieces of starting information: nucleotide database- BLASTn, protein databases -BLASTp, and translated nucleotide database - BLASTx and/or TBLASTn (remember that there can be more than one nucleotide triplet sequence that results in the same amino acid since the code is "degenerate"). For the most part, we will be interested in using the BLASTp database and the BLASTn database.

3.1 FASTA

An important aspect of the BLAST database is that it uses a particular file format called FASTA which is text based. This format allows users to manipulate the files with python or other scripting languages to, for instance, "grep" particular sequences etc easily. It is fairly standard for FASTA files to begin with a one line description of the sequence following a ">" sign but this can change a bit depending on the source. This *optional* initial identifier line is flexible but usually contains a combination of information such as species name, accession number, version or a number specific to the NCBI, the "gi" number, length of the sequence and last release date. In addition, each coding sequence has a unique number assigned to it that begins with 'AA'. Reference sequences, which are continuously curated and are considered to be the highest sequence standard, begin with the letters 'NT' if they are DNA, 'NM' if they are mRNA or 'NP' for protein. You should always use the reference sequences if you are able. FASTA represents each nucleotide or polypeptide in a sequence with a letter originally decreed by the International Union of Pure and Applied Chemistry (IUPAC). For a complete table of the nomenclature for nucleotides and amino acids, please see: `http://www.bioinformatics.org/sms2/iupac.html`.

Table 1: Summary table of IUPAC conventional nomenclature

A	Adenine	T (or U)	Thymine (or Uracil)
C	Cytosine	G	Guanine
R	Purines (A or G)	Y	Pyrimidines (C or T)
K	Keto (G or T)	M	Amino (A or C)
S	Strong (G or C)	W	Weak (A or T)
N	any base	. or -	Gap

When manipulating/opening/creating a FASTA document, remember: a potentially underappreciated aspect of the FASTA file format is that it is **TEXT ONLY**. You cannot use the default settings of Microsoft word or other text processors; you need to ensure that any ascii information is turned off. The most straightforward way do use a FASTA file is to open "TextEdit" (or notepad++ or TextWrangler) and choose the "Make Plain Text" option from the "Format" menu. Simply due to ease of use (and the fact that it is mono-spaced), it turns out that "courier" font is optimal to use for alignments etc. Double check that you have saved it with the suffix .fasta or, more commonly, .fsa and that it has not had any formatting accidentally added to it.

3.2 Manual BLAST

To see how powerful a tool the BLAST algorithm is, let's take a few minutes to see we can manually - without a computer program - find the following sequence embedded within a sequence of approximately 3000 bp which is a fairly standard size for an "average" gene length (I assure you that these two sequences are present but they are split over two lines so you can't find them with search and replace):

1. **TATACTTCAGGAACTAATTCTGAAGCATCA**
2. **AGATGGCACAGGGCATGAAATGAACACAA**

ACGGCGAGCGCGGGCGGCGGCGGTGACGGAGGCGCCGCTGCCAGGGGGCGT
GCGGCAGCGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGAGGCGGCGGCGG
CGGCGGCGGCGGCGGCGGCTGGGCCTCGAGCGCCCGCAGCCCACCTCTCGGG
GGCGGGCTCCCGGCGCTAGCAGGGCTGAAGAGAAGATGGAGGAGCTGGTGG
TGGAAGTGCGGGGCTCCAATGGCGCTTTCTACAAGGCATTTGTAAAGGATGT
TCATGAAGATTCAATAACAGTTGCATTTGAAAACAACTGGCAGCCTGATAGG
CAGATTCCATTTCATGATGTCAGATTCCCACCTCCTGTAGGTTATAATAAAGA
TATAAATGAAAGTGATGAAGTTGAGGTGTATTCCAGAGCAAATGAAAAAGGC
CTTGCTGTTGGTGGTTAGCTAAAGTGAGGATGATAAAGGGTGAGTTTTATGTG
ATAGAATATGCAGCATGTGATGCAACTTACAATGAAATTGTCACAATTGAAC
GTCTAAGATCTGTTAATCCCAACAAACCTGCCACAAAAGATACTTTCCATAA
GATCAAGCTGGATGTGCCAGAAGACTTACGGCAAATGTGTGCCAAAGAGGCG
GCACATAAGGATTTTAAAAAGGCAGTTGGTGCCTTTTCTGTAACTTATGATCC
AGAAAATTATCAGCTTGTCATTTTGTCCATCAATGAAGTCACCTCAAAGCGAG
CACATATGCTGATTGACATGCACTTTCGGAGTCTGCGCACTAAGTTGTCTCTG
ATAATGAGAAATGAAGAAGCTAGTAAGCAGCTGGAGAGTTCAAGGCAGCTT
GCCTCGAGATTTCATGAACAGTTTATCGTAAGAGAAGATCTGATGGGTCTAG
CTATTGGTACTCATGGTGCTAATATTCAGCAAGCTAGAAAAGTACCTGGGGT
CACTGCTATTGATCTAGATGAAGATACCTGCACATTTCATATTTATGGAGAGG
ATCAGGATGCAGTGAAAAAAGCTAGAAGCTTTCTCGAATTTGCTGAAGATGT
AATACAAGTTCCAAGGAACTTAGTAGGCAAAGTAATAGGAAAAAATGGAAA
GCTGATTCAGGAGATTGTGGACAAGTCAGGAGTTGTGAGGGTGAGGATTGAG
GCTGAAAATGAGAAAAATGTTCCACAAGAAGAGGAAATTATGCCACCAAATT
CCCTTCCTTCCAATAATTCAAGGGTTGGACCTAATGCCCCAGAAGAAAAAAA
CATTTAGATATAAAGGAAAACAGCACCCATTTTTCTCAACCTAACAGTACAA
AAGTCCAGAGGGTGTTAGTGGCTTCATCAGTTGTAGCAGGGGAATCCCAGAA
ACCTGAACTCAAGGCTTGGCAGGGTATGGTACCATTTGTTTTTGTGGGAACAA
AGGACAGCATCGCTAATGCCACTGTTCTTTTGGATTATCACCTGAACTATTTA
AAGGAAGTAGACCAGTTGCGTTTGGAGAGATTACAAATTGATGAGCAGTTGC
GACAGATTGGAGCTAGTTCTAGACCACCACCAAATCGTACAGATAAGGAAAA
AAGCTATGTGACTGATGATGGTCAAGGAATGGGTCGAGGTAGTAGACCTTAC
AGAAATAGGGGGCACGGCAGACGCGGTCCTGGATATACTTCAGGAACTAATT
CTGAAGCATCAAATGCTTCTGAAACAGAATCTGACCACAGAGACGAACTCAG
TGATTGGTCATTAGCTCCAACAGAGGAAGAGAGGGAGAGCTTCCTGCGCAGA
GGAGACGGACGGCGGCGTGGAGGGGGAGGAAGAGGACAAGGAGGAAGAGG
ACGTGGAGGAGGCTTCAAAGGAAACGACGATCACTCCCGAACAGATAATCGT
CCACGTAATCCAAGAGAGGCTAAAGGAAGAACAACAGATGGATCCCTTCAG
ATCAGAGTTGACTGCAATAATGAAAGGAGTGTCCACACTAAAACATTACAGA
ATACCTCCAGTGAAGGTAGTCGGCTGCGCACGGGTAAAGATCGTAACCAGAA
GAAAGAGAAGCCAGACAGCGTGGATGGTCAGCAACCACTCGTGAATGGAGT
ACCCTAAACTGCATAATTCTGAAGTTATATTTCCTATACCATTTCCGTAATTCT
TATTCCATATTAGAAAACTTTGTTAGGCCAAAGACAAATAGTAGGCAAGATG
GCACAGGGCATGAAATGAACACAAATTATGCTAAGAATTTTTTATTTTTTGGT
ATTGGCCATAAGCAACAATTTTCAGATTTGCACAAAAAGATACCTTAAAATTT
GAAACATTGCTTTTAAAACTACTTAGCACTTCAGGGCAGATTTTAGTTTTATT

4 How to BLAST

Now we know the file format that we will need to use, if we are given a gene can we go ahead and look up its sequence. You can use the major NCBI database found here for all the organisms: `http://www.ncbi.nlm.nih.gov` and a yeast-specific database found here: `http://www.yeastgenome.org`. For this lab, we will mostly be interested in the yeast-specific database but, in order to give you some experience that is portable to other organisms, we'll also use the NCBI database to look up orthologous sequences later in the lab.

However, prior to actually searching for the gene and amino acid sequences in the NCBI datasets, we will first take a little time to learn about the interface and the various tools available within PubMed, the platform upon which NCBI originally built its sequence-related databases. The entire concept of bioinformatics is to "discover" your own meaning by having the complete dataset available for your perusal and analysis. Since context is not given to us budding bioinformaticians as it would be if we were reading a paper that presented the sequence and function of a gene with a discussion section, it is important that we understand the major genetic concepts (such as mutation in sequences, inferring relationships of descent and the degeneracy of the genetic code) as well as the major features of the NCBI website.

4.1 General Overview of NCBI

Due to the many types of NCBI data and resources, NCBI has created a series of different search portals to locate these different kinds of databases and tools at NCBI. These portals are not actually databases although their interfaces often look as if they are. Each search portal is designed to return a list of links to a specific subset of NCBI databases that are functionally interrelated. Take heed: this means that it is very easy to get lost in the links!
The main features of a select few important databases are as follows:
Entrez: Text-based searching of records across 40 (and growing) separate databases which are highly hyperlinked to each other. There are currently four versions of Entrez but we will only be accessing the following website: `http://www.ncbi.nlm.nih.gov/gquery/`. Nearly all search boxes that appear on the NCBI website access the Entrez system - including the one at the very top when you land on the NCBI website! If you type in a "global query" with no limits (just default settings), the results are presented from each of the databases accessed by the search. You can simply type in a query that consists of words, names, accession numbers, species....whatever you like! Of course, this search strategy will result in a lot of unrelated results.
You can access each of the individual databases - if you want to limit your search to only specific databases - by using the pull down menu on the front page of NCBI. The Entrez search interface features powerful options for constructing precise and sophisticated searches, too. You can enjoy more precise search results by invoking Boolean operators including "AND', "OR" as well as "NOT" in the Entrez search bar and, as is usually the case with boolean operators, you can change the priority of reading left to right by using parenthesis. There are a number of ways to construct a useful and specific search; it is a bit of an art form but not one which we will be particularly concerned with today.
BLAST: Searches specifically for sequence data within a NCBI record. The search string is a sequence that is copy/pasted into the search box. BLAST then searches for similar sequences by mathematically calculating the probability that the search string sequence is similar to any of the

sequences available in NCBI records. On the main BLAST page, you will notice that you can use BLAST to design primers, too!
Gene, Genome, Homologene: These are components of Entrez and a text-based search engine highly specific for gene related information. They contain records with slightly different fields. For instance, Homologene provides an algorithm which searches for putative homologs.
PubMed: Also a member of the Entrez suite, PubMed is the familiar text-based search engine configured specifically to search the research literature. It is extensively hyperlinked to the bioinformatic records at NCBI and to all of the databases within NCBI Entrez.
MapViewer. Essentially a chromosomal browser, MapViewer searches and displays genomic, gene, transcript and disease information by chromosomal position (http://www.ncbi.nlm.nih.gov/mapview/). Researchers can choose from many different types of sequence, cytogenetic and radiation hybrid maps.

4.2 Search for sequence

We have now investigated some general guidelines about the type of data that we can search for within NCBI. Since we are using a model organism, *S.cerevisiae*, we will investigate the yeast specific website to answer some questions about the act1 gene which produces the protein actin. You could find the same information using the general NCBI website, however, it is easier to use a model-specific website when there is one available (remember, the website is here: `http://www.yeastgenome.org`). The **SGD** website collects and manages DNA and protein sequence information from primary providers (GenBank, EMBL, DDBJ, SwissProt, NCBI and PIR) in a user-friendly manner. The reference strain of *S.cerevisiae* strain S288C, which is true at NCBI and the other primary information providers (of course). Instead of ACT1, you will use the gene names that you have been given. Begin by typing the gene name into the search bar as shown in the following figure:

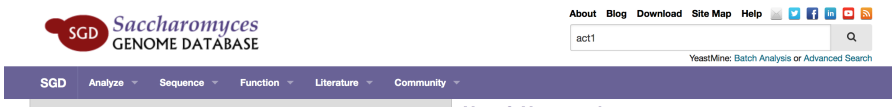


Figure 5: First step in obtaining a sequence

This will bring you to a page that gives you an overview of your gene including information such as *standard name, systematic name, aliases, description.* On this page, you can also find sequence information and are able to download the *FASTA file* of the sequence. Each of the sections on this page, expand into more detailed information when clicked on.

In fact, let's answer the following questions with a little exploration of this website (NB: some of the answers to these questions will fall somewhere between obvious/boring and interesting depending on the particular gene that you are using):

1a. Where - on what chromosome- is this gene located and what map positions does it occupy?

1b. What is the CDS of this gene? The CDS stands for the "coding DNA strand". It is the actual region of DNA that is translated to create proteins and it does not include any introns (to visualize this, you can picture concatenated exons).

1c. How many introns does the gene contain?

1d. By exploring the ORF map, you can tell if your gene is transcribed as a Watson or a Crick strand ORF . The ORF is the "Open Reading Frame" and it contains the start (ATG) and stop codons (TAA, TAG, TGA) and may include introns. There are six possible ways to translate the ORF into a polypeptide since there are two strands and three possible frames on each strand so it is more of a "theoretically possible protein coding gene" than a definite protein producing gene. The **SGD** website uses + to indicate ORF encoded on the Watson strand (the reference chromosome in red); and and - to indicate ORFs encoded on the Crick strand. As an example, the gene act1 is transcribed from the - (Crick) strand. You can determine this from the orientation of the gene sequence section (or, you can click on the chromosome number and it will take you to a list of all the genes on that chromosome as well as which strand they are transcribed from).

1e. What are the names of the immediately adjacent genes?

2. How many amino acids in length is the protein that this gene codes? How many nucleotides long is the gene (in bp)?

3. What is the mutant phenotype of this gene? What is the GO? The term *GO* stands for Gene Ontology and is an attempt at standardizing functional products of genes with an eye towards making annotation easier.

4. How many primary papers describe this gene?

5. What is the first reference for this gene?

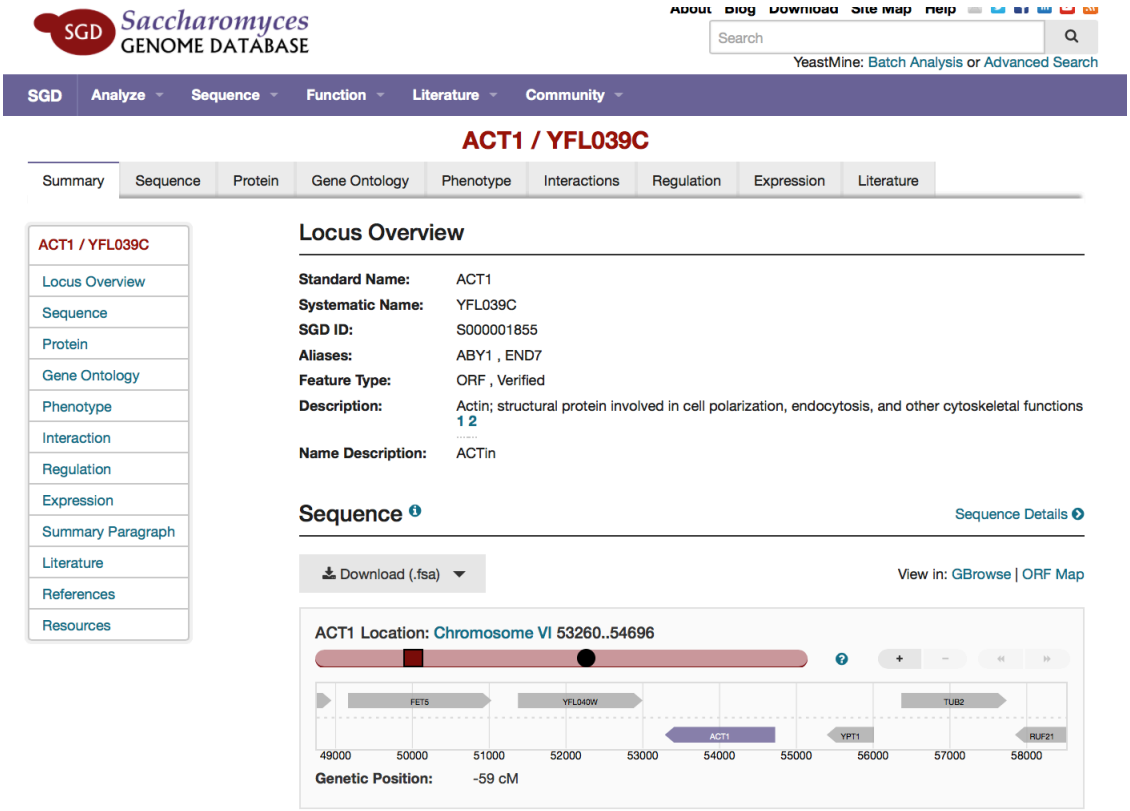


Figure 6: Overview page

You can now click on the .fsa file and download it. We will download two types of files: a nucleotide file **"coding"** sequence and the **amino acid** file. We will use the amino acid file to align our sequence with other, distantly related organisms since we expect less mutation to occur that changes amino acid sequence than that which occurs in nucleotide sequences (nucleotide sequences can quickly become saturated with mutations on the time scales that separate these organisms). Remember to use a plain text editor to save your sequence file to have the same name as the file that automatically pops up on your screen.

4.3 Homology

Homology implies a common evolutionary relationship between two traits whether the traits (sometimes called "characters") are molecular or anatomical. Homologous sequences are sequences that are related through a common ancestor but which, usually, have diverged within their sequences. It is possible to have two sequences which are similar as a result of convergence rather than shared ancestry but homology itself implies a common ancestor. The amount of sequence divergence between two organisms depends on the type of gene we are examining (some genes are more highly

>ACT1 YFL039C SGDID:S000001855
ATGGATTCTGAGGTTGCTGCTTTGGTTATTGATAACGGTTCTGGTATGTGTAA
AGCCGGTTTTGCCGGTGACGACGCTCCTCGTGCTGTCTTCCCATCTATCGTCG
GTAGACCAAGACACCAAGGTATCATGGTCGGTATGGGTCAAAAAGACTCCTA
CGTTGGTGATGAAGCTCAATCCAAGAGAGGTATCTTGACTTTACGTTACCCAA
TTGAACACGGTATTGTCACCAACTGGGACGATATGGAAAAGATCTGGCATCA
TACCTTCTACAACGAATTGAGAGTTGCCCCAGAAGAACACCCTGTTCTTTTGA
CTGAAGCTCCAATGAACCCTAAATCAAACAGAGAAAAGATGACTCAAATTAT
GTTTGAAACTTTCAACGTTCCAGCCTTCTACGTTTCCATCCAAGCCGTTTTGTC
CTTGTACTCTTCCGGTAGAACTACTGGTATTGTTTTGGATTCCGGTGATGGTG
TTACTCACGTCGTTCCAATTTACGCTGGTTTCTCTCTACCTCACGCCATTTTGA
GAATCGATTTGGCCGGTAGAGATTTGACTGACTACTTGATGAAGATCTTGAGT
GAACGTGGTTACTCTTTCTCCACCACTGCTGAAAGAGAAATTGTCCGTGACAT
CAAGGAAAAACTATGTTACGTCGCCTTGGACTTCGAACAAGAAATGCAAACC
GCTGCTCAATCTTCTTCAATTGAAAAATCCTACGAACTTCCAGATGGTCAAGT
CATCACTATTGGTAACGAAAGATTCAGAGCCCCAGAAGCTTTGTTCCATCCTT
CTGTTTTGGGTTTGGAATCTGCCGGTATTGACCAAACTACTTACAACTCCATC
ATGAAGTGTGATGTCGATGTCCGTAAGGAATTATACGGTAACATCGTTATGTC
CGGTGGTACCACCATGTTCCCAGGTATTGCCGAAAGAATGCAAAAGGAAATC
ACCGCTTTGGCTCCATCTTCCATGAAGGTCAAGATCATTGCTCCTCCAGAAAG
AAAGTACTCCGTCTGGATTGGTGGTTCTATCTTGGCTTCTTTGACTACCTTCCA
ACAAATGTGGATCTCAAAACAAGAATACGACGAAAGTGGTCCATCTATCGTT
CACCACAAGTGTTTCTAA

>ACT1 YFL039C SGDID:S000001855
MDSEVAALVIDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGIMVGMGQKDSY
VGDEAQSKRGILTLRYPIEHGIVTNWDDMEKIWHHTFYNELRVAPEEHPVLLTEA
PMNPKSNREKMTQIMFETFNVPAFYVSIQAVLSLYSSGRTTGIVLDSGDGVTHVV
PIYAGFSLPHAILRIDLAGRDLTDYLMKILSERGYSFSTTAEREIVRDIKEKLCYVA
LDFEQEMQTAAQSSSIEKSYELPDGQVITIGNERFRAPEALFHPSVLGLESAGIDQ
TTYNSIMKCDVDVRKELYGNIVMSGGTTMFPGIAERMQKEITALAPSSMKVKIIA
PPERKYSVWIGGSILASLTTFQQMWISKQEYDESGPSIVHHKCF*

Figure 7: The saved nucleotide sequence of ACT1 (Top) and the amino acid sequence (Bottom) of its coding product

conserved than others with a famous example being histone proteins, which are very highly conserved across taxa, versus fibronectin which, despite having particular domains that are highly conserved, is more divergent in nucleotide sequences) and the amount of times since two organism shared a common ancestor (more time is usually proportional to accumulating more differences between the two sequences). Thus, the degree of similarity between two organisms allows us to reconstruct their past relationship.

Two major classes of homology are often considered as potentially informative of evolutionary relationships: **orthology** and **paralogy**. Orthology results when sequences are similar (and specify the same gene) between species. More formally, sequences are called orthologous sequences when they were present in the same ancestor but experienced a subsequent speciation event. Orthologous sequences are particularly valuable when constructing phylogenetic trees since two organisms that are closely related are likely to contain very similiar sequences at two orthologs. Paralogous genes arise as the result of a duplication event of one gene within the same species so that the second (or more) gene is now in a different location than the original. In humans, the most famous example is the paralog of an opsin protein which endows (most of us) with trichromatic vision. Due to a lack of, or reduced, selective pressure on the second copy of the paralogous sequence, paralogs are used to study the process of genomic evolution.

4.4 Search for related sequences

Now that we have the fasta file for our gene, we will use the BLAST page at ncbi to find any homology in other organisms. To standardize this process let's all choose Human, Mouse, Zebrafish and Drosophila as our comparison organisms (note: depending on your particular gene you might not be able to get a homolog in all of those organisms) and retrieve the nucleotide sequences for these four other organisms. We can upload (or cut and paste) these sequences into the clustal omega program and it will give us a phylogenetic comparison and a resultant cladogram, or phylogenetic tree which diagrams evolutionary relationships between species. Recall that in a cladogram, the species that are more closely related, indicated by sequence conservation, will share a more recent common ancestor (an internal node on the tree) than species that are more distantly related.

We will, of course, go to the BLAST link for all of these model organisms, `http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome` and click the link that states "protein blast". You can then cut and past your fasta file into the box that says "Enter Query Sequence". You then go to the "Choose Search Set" and type "human" into the Organism box. As you type, the box should give you a pull down menu to complete the word. For human, this should be "human (taxid:9606)", for Zebrafish it will be "zebrafish (taxid: 7955)".

First, for the human sequence, after you hit "BLAST" and your computer *thinks hard*, a page with the closest human protein matches will pop up. Hopefully, you will see a lot of red lines present on your page (which indicates that there is quite a bit of sequence similarity). Pick a sequence with the lowest "E value" (which basically gives you the probability of having sequence similarity simply due to chance) and click on it. You may then download a number of different file formats for this particular alignment. When you click on the "Download" menu located with your alignment, you will get three possible formats which you can download. You should choose the "FASTA (aligned sequences)" for simply the amino acid sequence which you will cut and paste into CLUSTAL in the next section. You can also choose the GenBank (complete sequence) which will give you a file that

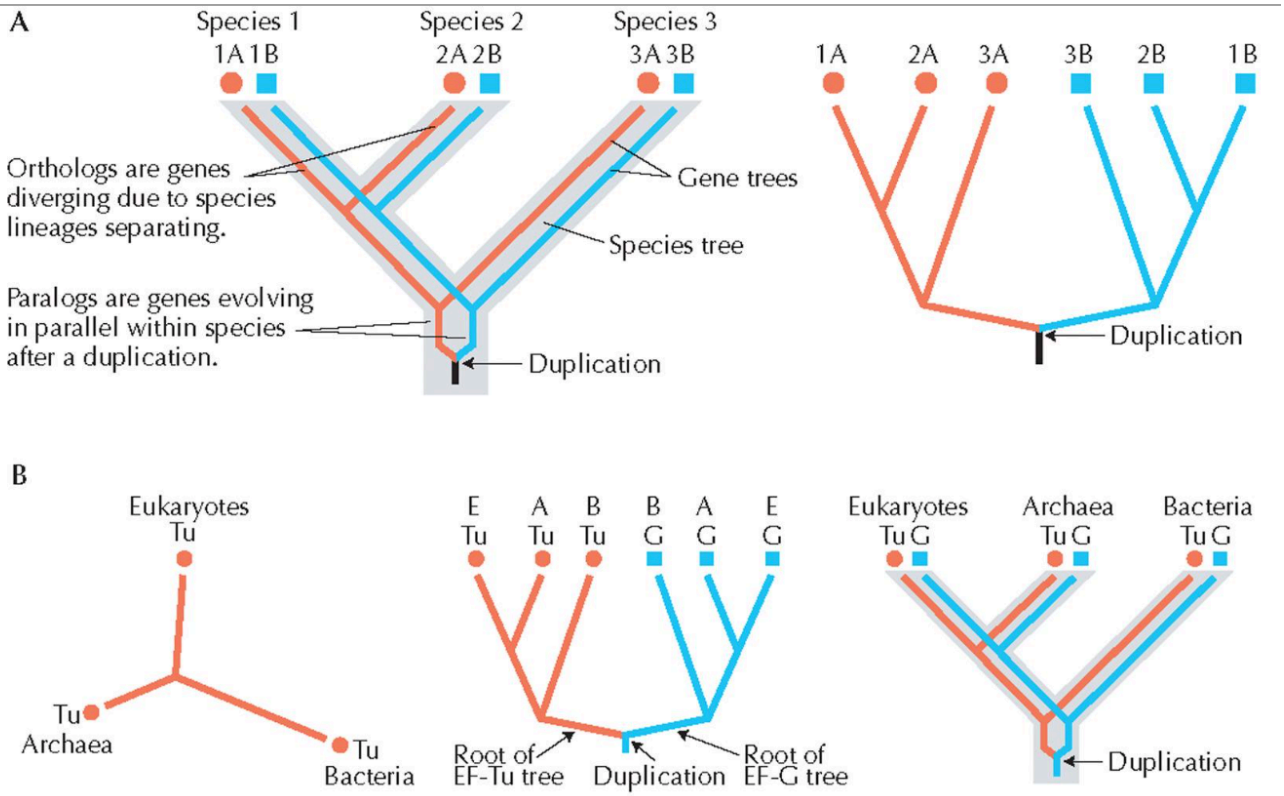


FIGURE 5.20. Orthologs, paralogs, and rooting the tree of life. (*A*) Evolutionary trees of species and genes representing gene duplication events. (*Left*) The tree includes a species tree (*thick gray lines*) and gene trees (*blue* and *red lines*). A gene duplication event, leading to the coexistence of the blue and red paralogs in the root of the species tree, is labeled. The species tree subsequently splits twice, producing three species, each of which has inherited the blue and red paralogs. (*Right*) The gene tree is extracted from the species tree and untangled. The red forms of the gene, which are orthologs of each other, are more closely related to each other than to any of the blue forms. The same is true for the blue forms of the gene. Note that the species relationships among the two groups of orthologs (red and blue) are the same. (*B*) The same types of trees as in *A*, but these correspond to the evolution of elongation factors Tu and G across the three domains of life. The *red* and *blue* branches in the rightmost tree each correspond to a Tree of Life, and each is rooted by the paralogous elongation factor.

5.20A,B, redrawn from Eisen J.A., *Genome Res.* **8:** 163–167, © 1998 CSHLP, www.cshlpress.com

Figure 8: The two different major types of homology

contains the amino acid sequence at the bottom along with a lot of other useful information about the annotation of the sequence but remember that it is not in FASTA format so you can't simply cut and paste it into CLUSTAL.

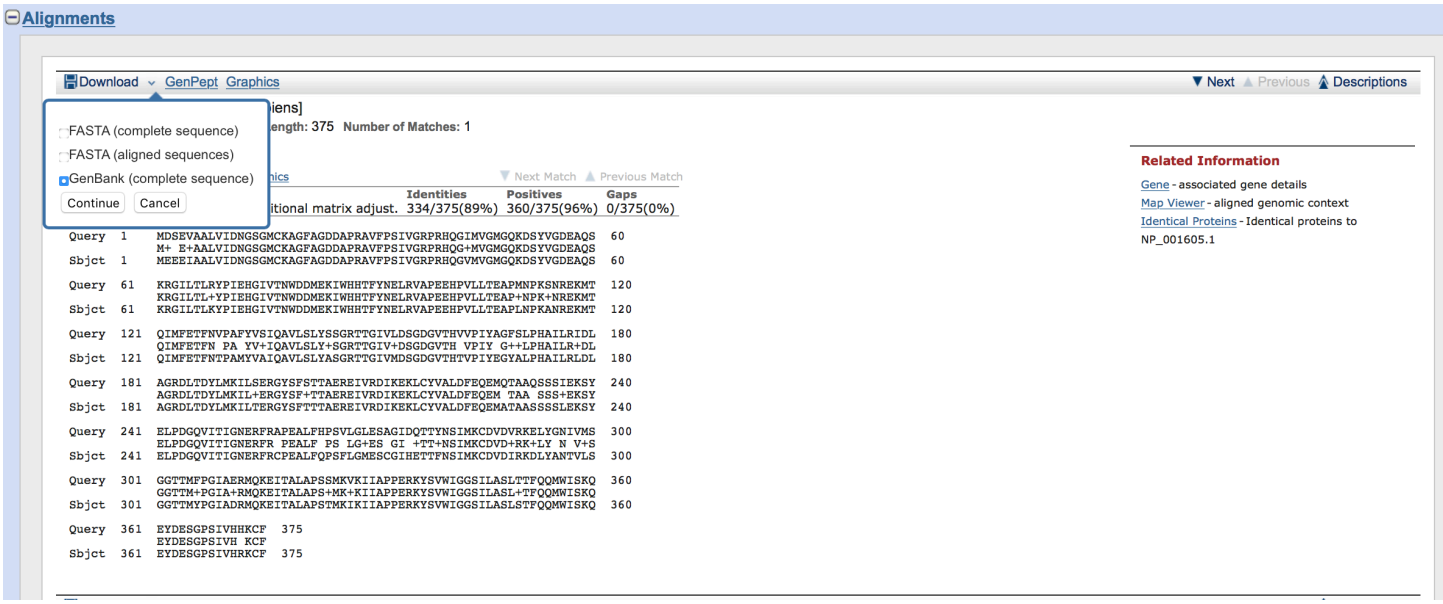


Figure 9: Picking your alignment with the lowest E-value

```
>gi|4501887|ref|NP_001605.1|:1-375 actin, cytoplasmic 2 [Homo sapiens]
MEEEIAALVIDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKYP
IEHGIVTNWD
DMEKIWHHTFYNELRVAPEEHPVLLTEAPLNPKANREKMTQIMFETFNTPAMYVAIQAVLSLYASGRTTG
IVMDSGDGVT
HTVPIYEGYALPHAILRLDLAGRDLTDYLMKILTERGYSFTTTAEREIVRDIKEKLCYVALDFEQEMATA
ASSSSLEKSY
ELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTFNSIMKCDVDIRKDLYANTVLSGGTTMYPGIA
DRMQKEITAL
APSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQEYDESGPSIVHRKCF
```

```
LOCUS       NP_001605               375 aa            linear   PRI
15-MAR-2015
DEFINITION  actin, cytoplasmic 2 [Homo sapiens].
ACCESSION   NP_001605
VERSION     NP_001605.1  GI:4501887
DBSOURCE    REFSEQ: accession NM_001614.3
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates;
Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 375)
  AUTHORS   Flouriot G, Huet G, Demay F, Pakdel F, Boujrad N and
Michel D.
  TITLE     The actin/MKL1 signalling pathway influences cell growth
and gene
            expression through large-scale chromatin reorganization
and histone
            post-translational modifications
  JOURNAL   Biochem. J. 461 (2), 257-268 (2014)
   PUBMED   24762104
  REMARK    GeneRIF: The actin/MKL1 signalling pathway influences cell
growth
            and gene expression through large-scale chromatin
reorganization
            and histone post-translational modifications.
REFERENCE   2  (residues 1 to 375)
  AUTHORS   Luo Y, Kong F, Wang Z, Chen D, Liu Q, Wang T, Xu R, Wang X
and Yang
            JY.
  TITLE     Loss of ASAP3 destabilizes cytoskeletal protein ACTG1 to
suppress
            cancer cell migration
  JOURNAL   Mol Med Rep 9 (2), 387-394 (2014)
   PUBMED   24284654
  REMARK    GeneRIF: The data, for the first time, link ASAP3 with
ACTG1 in the
            regulation of cytoskeletal maintenance and cell motility
REFERENCE   3  (residues 1 to 375)
  AUTHORS   Pieragostino D, Agnifili L, Fasanella V, D'Aguanno S,
Mastropasqua
            R, Di Ilio C, Sacchetta P, Urbani A and Del Boccio P.
  TITLE     Shotgun proteomics reveals specific modulated protein
patterns in
            tears of patients with primary open angle glaucoma naive
to therapy
  JOURNAL   Mol Biosyst 9 (6), 1108-1116 (2013)
```

Figure 11: The full file from GenBank which gives annotation information

Repeat the above process for mouse, zebrafish and drosophila to obtain three more FASTA files with the amino acid sequence. We will then move on to the CLUSTAL program armed with these files.

4.5 CLUSTAL

Cluster analysis of the pairwise alignments. This is software (available in different versions) that allows for multiple alignments for nucleotide or protein sequences and subsequent creation of a tree by UPGMA (Unweighted pair group method with arithmetic mean). It produces biologically meaningful multiple sequence alignments of divergent sequences, calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. There are many ways of constructing phylogenetic trees but CLUSTAL uses a algorithm that is well respected. CLUSTAL is optimized for the specific case where you are aligning colinear sequences (orthologous regions). You should consider a different alignment program when you are comparing sequences that do not share common ancestry (perhaps produced from convergent evolution) or sequences that have only have regions that are related (perhaps through recombination events). There are many other programs that use, for instance, phylogenetic methods rather than pair group methods, and which will produce slightly different results (remember that every program will have its own specific models and assumptions built into it).

We will use CLUSTAL omega which is available here `http://www.ebi.ac.uk/Tools/msa/clustalo/`. This program allows us to align three or more sequences which, in turn, allows easy computational identification of more and less conserved regions and, ultimately, the construction of a phylogenetic tree that relates the three or more sequences. CLUSTAL O uses neighbor joining to create this tree and it is based on a matrix of pairwise differences between the sequences.

```
CLUSTAL O(1.2.1) multiple sequence alignment

ACT1                                        MDSEVAALVIDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGIMVGMGQKDSYVGDEAQS
gi|4501887|ref|NP_001605.1|:1-375           MEEEIAALVIDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQS
gi|6752954|ref|NP_033739.1|:1-375           MEEEIAALVIDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQS
gi|62298523|sp|Q7ZVI7.2|ACTB1_DANRE:1-375   MDEEIAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQS
                                            *:.*:****:**********************************:*****************

ACT1                                        KRGILTLRYPIEHGIVTNWDDMEKIWHHTFYNELRVAPEEHPVLLTEAPMNPKSNREKMT
gi|4501887|ref|NP_001605.1|:1-375           KRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPEEHPVLLTEAPLNPKANREKMT
gi|6752954|ref|NP_033739.1|:1-375           KRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPEEHPVLLTEAPLNPKANREKMT
gi|62298523|sp|Q7ZVI7.2|ACTB1_DANRE:1-375   KRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPEEHPVLLTEAPLNPKANREKMT
                                            *******:******************************************:***:******

ACT1                                        QIMFETFNVPAFYVSIQAVLSLYSSGRTTGIVLDSGDGVTHVVPIYAGFSLPHAILRIDL
gi|4501887|ref|NP_001605.1|:1-375           QIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEGYALPHAILRLDL
gi|6752954|ref|NP_033739.1|:1-375           QIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEGYALPHAILRLDL
gi|62298523|sp|Q7ZVI7.2|ACTB1_DANRE:1-375   QIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEGYALPHAILRLDL
                                            ********.**:**:********:********:********.**** *::********:**

ACT1                                        AGRDLTDYLMKILSERGYSFSTTAEREIVRDIKEKLCYVALDFEQEMQTAAQSSSIEKSY
gi|4501887|ref|NP_001605.1|:1-375           AGRDLTDYLMKILTERGYSFTTTAEREIVRDIKEKLCYVALDFEQEMATAASSSSLEKSY
gi|6752954|ref|NP_033739.1|:1-375           AGRDLTDYLMKILTERGYSFTTTAEREIVRDIKEKLCYVALDFEQEMATAASSSSLEKSY
gi|62298523|sp|Q7ZVI7.2|ACTB1_DANRE:1-375   AGRDLTDYLMKILTERGYSFTTTAEREIVRDIKEKLCYVALDFEQEMGTAASSSSLEKSY
                                            *************:******:************************* ***.***:****

ACT1                                        ELPDGQVITIGNERFRAPEALFHPSVLGLESAGIDQTTYNSIMKCDVDVRKELYGNIVMS
gi|4501887|ref|NP_001605.1|:1-375           ELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTFNSIMKCDVDIRKDLYANTVLS
gi|6752954|ref|NP_033739.1|:1-375           ELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTFNSIMKCDVDIRKDLYANTVLS
gi|62298523|sp|Q7ZVI7.2|ACTB1_DANRE:1-375   ELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTFNSIMKCDVDIRKDLYANTVLS
                                            ****************.*****:**.**:**.**.:**:*********:**:**.* *:*

ACT1                                        GGTTMFPGIAERMQKEITALAPSSMKVKIIAPPERKYSVWIGGSILASLTTFQQMWISKQ
gi|4501887|ref|NP_001605.1|:1-375           GGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQ
gi|6752954|ref|NP_033739.1|:1-375           GGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQ
gi|62298523|sp|Q7ZVI7.2|ACTB1_DANRE:1-375   GGTTMYPGIADRMQKEITSLAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQ
                                            *****:****:*******:****:**:**********************:**********

ACT1                                        EYDESGPSIVHHKCF*
gi|4501887|ref|NP_001605.1|:1-375           EYDESGPSIVHRKCF-
gi|6752954|ref|NP_033739.1|:1-375           EYDESGPSIVHRKCF-
gi|62298523|sp|Q7ZVI7.2|ACTB1_DANRE:1-375   EYDESGPSIVHRKCF-
                                            ***********:***
```

Figure 12: CLUSTAL Alignment of the ACT1 amino acid sequence our four species

To see the resultant phylogenetic tree, go to the "phylogenetic tree" tab in the results. Does this phylogenetic tree capture known relationships between species? Is there anything that is unexpected?

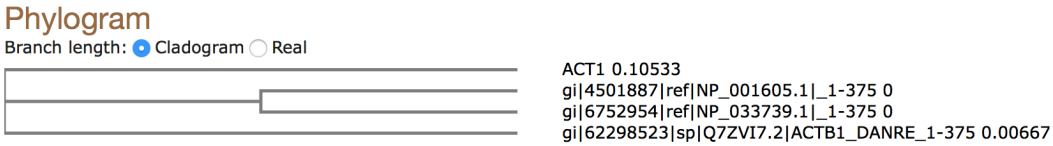


Figure 13: CLUSTAL Cladogram of the ACT1 amino acid sequence our four species

We can also attempt to find some paralogs to our gene. Paralogs are trickier to find than simply by using a BLAST search since, depending on how long ago they arose through a duplication event, they may have accumulated enough mutations to no longer be recognized as similar enough to the original sequence. Paralogs can, of course, be found across species but to try to narrow down our search, we will initially concentrate on finding potential paralogs within *S. Cerevisiae* itself and closely related species.

`yeastgenome.org` has a specific "synteny" viewer which captures the conservation of blocks between adjacent loci in multiple species. Synteny allows the establishment of orthologous regions between species and, once that is done, we can then designate paralogs to a homologous gene in *S.cerevisiae*. Orthology is established when there is more than one homologous gene but only one of them has the same relative position on the chromosome as the *S.cerevisiae* gene. The synteny viewer is found here, under the "Strains and Species" menu : `http://www.yeastgenome.org/cgi-bin/FUNGI/FungiMap`. It should be noted that it finds homologous regions between the four species *S.cerevisiae*, *S.paradoxus*, *S.mikatae* and *S.uvarum*.

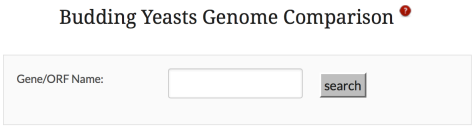


Figure 14: Type in your gene here

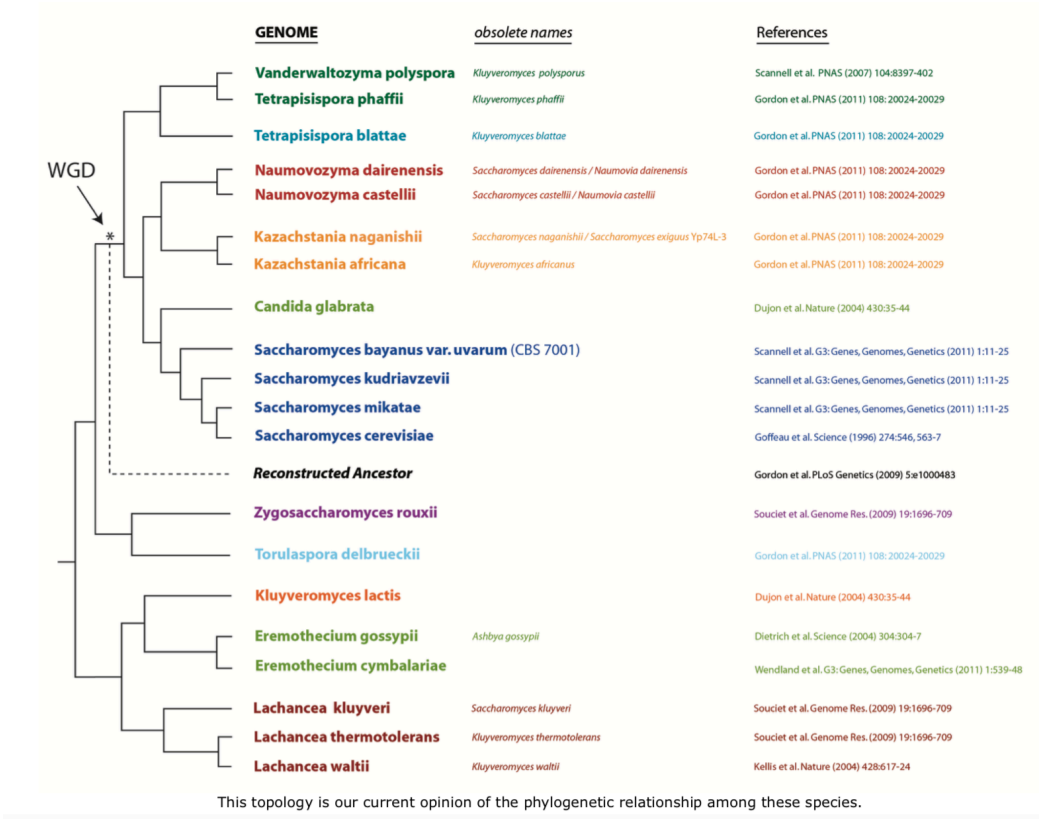
The synteny viewer results in a diagram that includes the adjacent genes to your gene of interest:

6. Do your genes share orthology with related species? If so, how much between each one?

Finally, we can also use the `http://www.yeastgenome.org/blast-fungal` and conduct a "fungal blast" to confirm any paralogs that we found with the synteny viewer. We can then cut and paste our amino acid sequence, choose BLASTP and choose the species *Saccharomyces cerevisiae* since we are only looking for potential paralogs within this species. In this search, we are not looking for the sequence that gives us the lowest "E-value". In fact, we are looking for a name that isn't simply ACT1 or Actin (in the case of our sample gene). The first hit that I find is called "actin-related protein 1". You can confirm that this isn't the same gene by looking for its location in the genome (among other markers).

We can also use the PBD homology query database to easily investigate homologs; it is found here: `http://www.yeastgenome.org/cgi-bin/protein/get3d`.

Another powerful homology search is offered at the website: `http://ygob.ucd.ie`. This browser gives us a comparison view of many species of yeast which contain homologous genes (or ORFs) to our gene of interest.



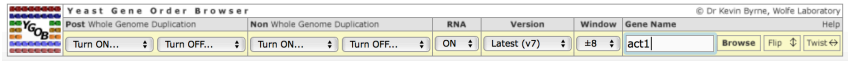
It also gives you sequence information as **aa** or **nt**, by clicking on the aa|nt icon at the top of the column of your gene of interest. This browser also constructs a tree of relatedness between the genes in a column (using PhyML and the MUSCLE alignment) by clicking on the **tree** button. Note: So far, we have explored the CLUSTAL Ω program for phylogenetic reconstruction but there are many others out there (MUSCLE, MEGA, BEAST) and which one is best to use will depend on your research questions and assumptions. The **YGOB** website uses the the program MUSCLE to construct phylogenies.

Another incredibly useful feature of this website is the **rates** button. This will provide the **ka**, **ks** and **omega** values. The ratio of the non-synonymous site substitutions, Ka, to synonymous site substitutions, Ks, in a protein coding gene gives an indication of potential selective constraints (Ka/Ks also known as Ω when it is scaled correctly >1 is suggestive of positive selection; a ratio between 0 and 1 suggests purifying or stabalizing selection; a ratio of 1 is suggestive of neutral selection).

I am not sure how useful this will be, in particular, for yeast species however it is worth the time to possess a basic idea of what this ratio means. So a very quick (general) outline for any of your future courses:
In a neutral world, $k = \mu$
and since two species are separated by t **units of time since a common ancestor,** $K = 2kt = 2\mu * t$
This means for the synonymous sites, which generally have no functional constraints (this is not always true, especially if they are in areas of regulation etc): $k_s = 2\mu * t$ whereas nonsynonymous changes are likely under some functional constraint since they replace a potentially important amino acid. So $k_a = (2\mu * t) * f_o$ where f_o is the fraction that is "free to vary". Obviously, this means that $(1 - f_o)$ is functionally constrained by selection.
Since k_a **and** k_s **have shared the same phylogeny, we can test how different the results of selective pressures are on each one of them by the ratio** $k_a/k_s = (2\mu * t) * f_o/(2\mu * t) = f_o$

The YGOB **rates** button produces a yn00 file produced by the yn00 method of the PAML suite. PAML is a suite of programs for phylogenetic analyses of DNA or protein sequences using maximum likelihood (ML) and yno00 is method that estimates synonymous and nonsynonymous substitution rates in pairwise comparison of protein-coding DNA sequences. This file gives three columns, including the all-important Ω values.

The browser itself appears at the bottom of the screen, like this:



It results in an image that aligns all the orthologs in a column, like this: