BIO00056I workshop: Comparative Genomics

Developed by: Daniel Jeffares
2023-11-18

Aims of the practical

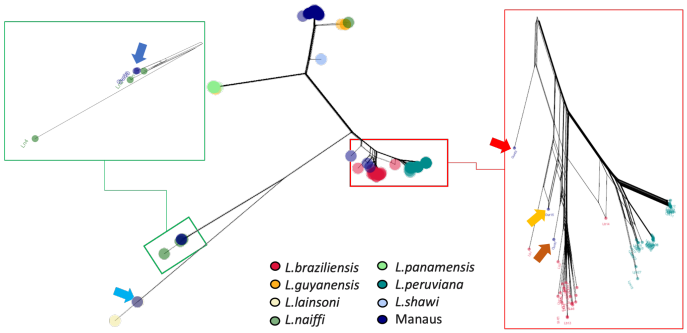
- Learn more about comparative genomics data
- Practise using web tools to interpret genomics data
- Learn about gene conservation over deep time (billions of years)
- See an example of the principle that diversity *within* species is related to divergence *between* species.

Introduction

Today, we examine whether we can observe consistent patterns of evolutionary rates across time scales.

Evolutionary change over deep time

In the last workshop, we saw how several species of *Leishmania* parasites that can be found in the Amazon rainforest and other areas of South America were genetically different from each other. These species were all part of the **Viannia** clade, as shown in the phylogeny to the right.



In this workshop, we continue looking at *Leishmania* data. This time we look at how more distantly-related *Leishmania* species. We know from the **molecular clock concept** that if two species are dissimilar genetically, this is because mutations have accumulated ***over time***. **Figure 2** below shows that the molecular clock also holds over very long periods of time.

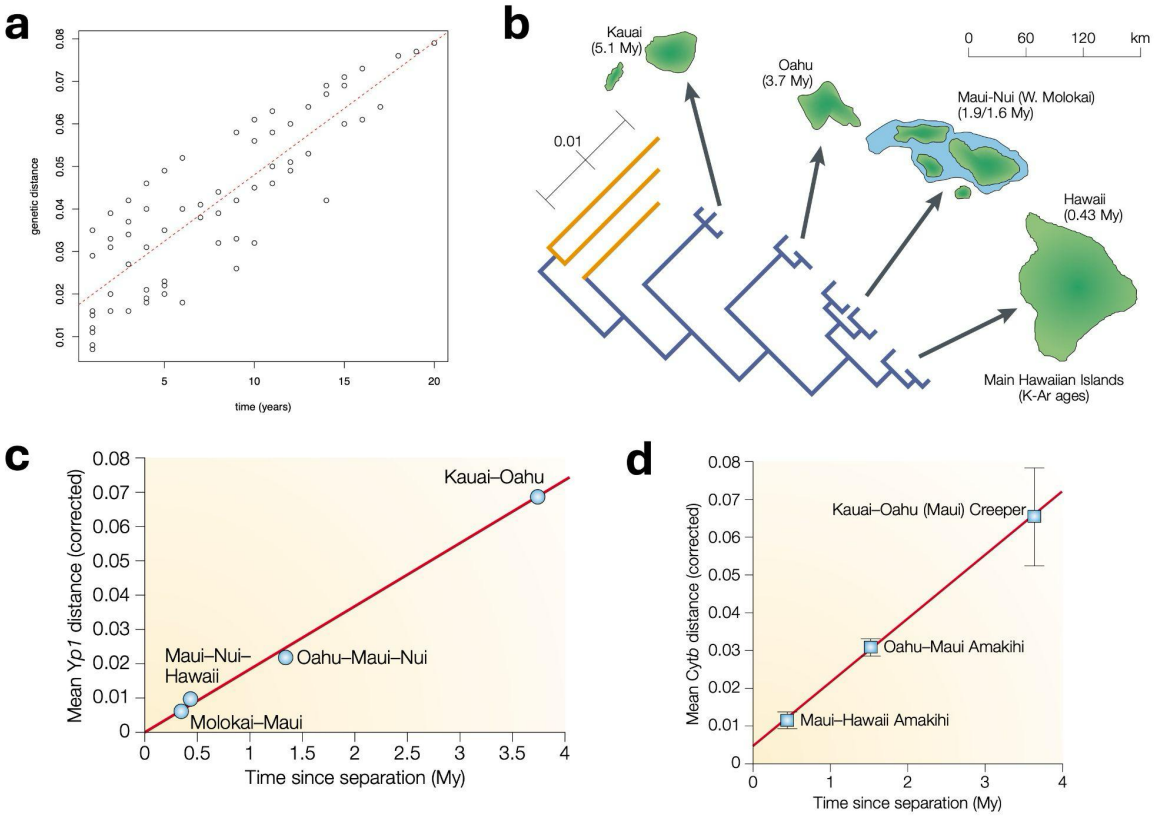


Figure 2. Proof that molecular clock causes divergence between species over long periods of time. Panel a: We saw from the influenza virus workshop that mutations occurred within influenza viruses at a regular rate with time. This also occurs over long periods of time. **Panel b:** For example, geologists know that the pacific islands of Hawaii have been formed over millions of years. Evolutionary biologists know that both honeycreeper birds and *Drosophila* fruit flies have separated on these islands, so we know how long ago they diverged. If we look at genetic data from the *Yp1* gene in honeycreepers (**Panel c**) and the *Cytb* gene from fruit flies (**Panel d**), we observe a very consistent correlation between genetic distance (vertical axis) and time between island formation

(horizontal axis), which si a good proxy for evolutionary time between these species. Just like the data you saw from the influenza virus, it looks very much like there is a constant rate of genetic divergence with time, this time over *millions* of years. Plot from (Bromham and Penny, 2003).

In this workshop

We also know from the lectures that:

- accumulation of mutations *within species* generates the genetic differences *between species*
- different kinds of genes have different evolutionary rates, for example:
    - in a single-celled organism, a protein that is located on the *outside* of the cell may evolve rapidly to adapt to the external environment
    - a protein that is located deep on the inside of the cell, doing fundamental processes (like DNA replication), may change slowly because it is so essential

To examine whether we can see any evidence for these hypotheses, we look at SNPs *within a population* that have accumulated recently, and gene divergence *between species* that have speciated long ago. At the end of the workshop, we compare *Leishmania braziliensis* to a yeast species that has been diverging from *Leishmania* for a billion years. **We will see that there are consistent patterns of evolutionary rates on all these time scales.**

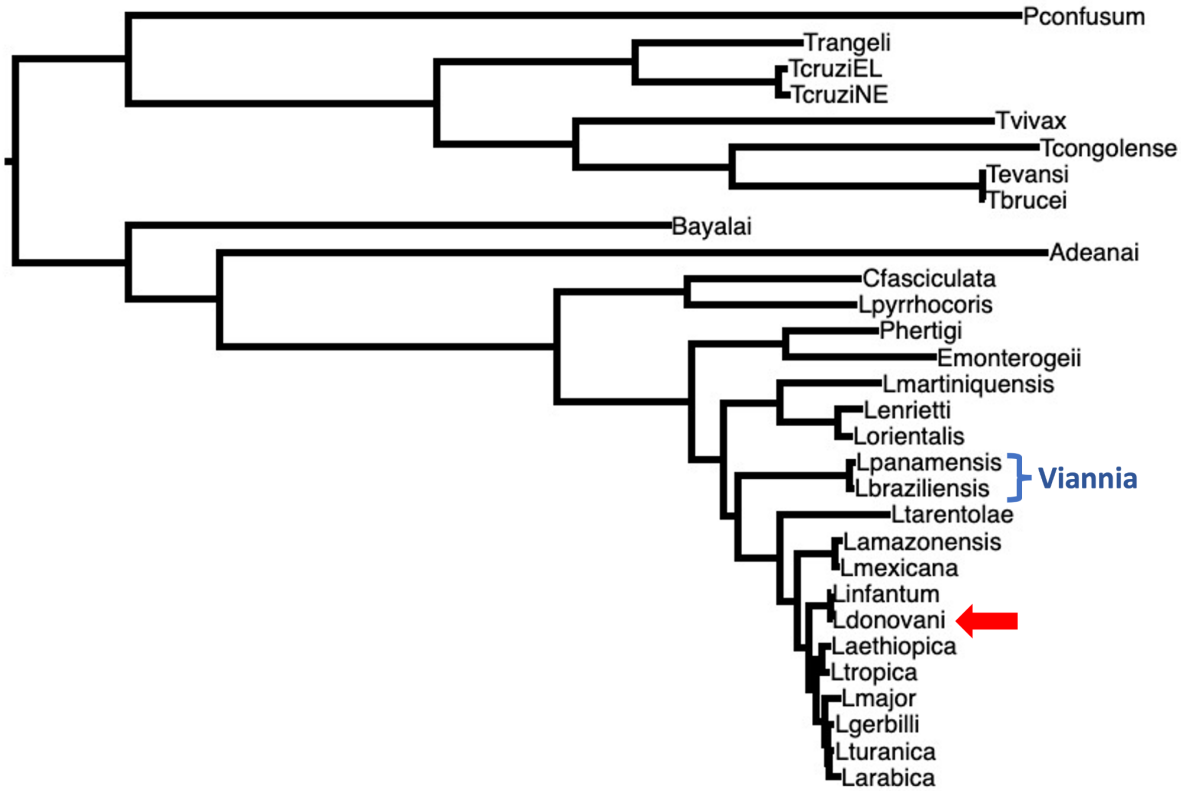


Figure 1. A phylogenetic tree of *Leishmania* and related species. Two Vianna species that we looked at last time in the Viannia clade are included (blue brackets). *Leishmania donovani*, which we will compare *Leishmania braziliensis* to, is indicated with a red arrow.

Why study *Leishmania* species?

I chose *Leishmania* species for this workshop because I work on these species for my research. We could have used any eukaryotic species for this workshop (as long as the data were available). However, this data shows **general patterns of evolution that occur in all species.**

Different rates of protein change begin *within species*.

First we look at genetic differences strains within a species. Here we use SNP data from *Leishmania donovani*, because there have been ~900 strains of this species sequenced. So this is **population genomics** data.

We know from some more detailed analysis that this genetic diversity has been accumulating *fairly recently*, because *Leishmania donovani* started to spread across the globe only about 1800 years ago (Reis-Cunha et al., 2024). Since genetic diversity *within* a species causes the divergence *between* species, we *should* see differences within the polymorphisms that relate to the differences between species.

To look at this, we went to TriTrypDB, the database that contains information about *Leishmania* and related species. We downloaded information about the number of SNPs in each gene. SNPs are classified as:

- **synonymous**; those that *do not* change the protein's amino acid sequence. Sometimes these are called silent mutations.
- **nonsynonymous**; those *do* change the protein sequence

We expect the synonymous mutations to be more or less neutral (ie: not subject to selection), so we can think of these as representing what occurs without selection, like a control. The *nonsynonymous* ones will often be subject to purifying selection (since most of them will be deleterious). So this count indicates what happens *after* selection. So the ratio of nonsynonymous (selection) / synonymous (control), will tell us how many fewer nonsynonymous (amino-acid changing) SNPs we are seeing than we expect without selection. Let's call this the **N/S ratio** of the gene.

- **If the N/S ratio of a gene is very low** (eg: 0.1), then we are seeing many fewer amino-acid changing SNPs that we expect under neutrality, which means strong purifying selection

- **If the N/S ratio of a gene = 1,** then it looks like there is no selection to remove amino-acid changing SNPs

- **If the N/S ratio of a gene is high (>1)**, there might be selection for more amino-acid changing SNPs than the neutral rate. This could be because they are actively changing.

The SNP counts for all the genes in *L. donovani* can be found in **Table 1 (Ldon all)** of this Google Sheet. This shows the N/S ratios for all the genes in *L. donovani.* This doesn't tell us much, except that genes differ in the number of SNPs and also the

N/S ratios. This isn't surprising since mutations are *stochastic* (like the snow falling), and then selection occurs after that to change the numbers of mutations that remain. But we can look for signals, by looking at:

- genes with the highest N/S (either adapting rapidly, or with weak purifying selection), in **Table 2 (Ldon weak purifying)**
- genes with the lowest N/S (strongest purifying selection) in **Table 3 (Ldon strong purifying)**

These lists are the same Google Sheet.

Consider the results

The average N/S ratio for the genes in the **Ldon strong purifying** list is 0.078. This is the accumulation of ***non****synonymous* mutations, relative to the accumulation of *synonymous* mutations.

a) If we assume that the two types of mutations (***non****synonymous and synonymous)* occur with equal frequency, why are we seeing much fewer ***non****synonymous* mutations? Where did they 'go'?
b) What proportion of nonsynonymous mutations are removed in this way?

ANSWERS:

a) Most *nonsynonymous* mutations were removed by purifying selection.
b) The rate of *nonsynonymous* mutations accumulation relative to the neutral rate, is 0.922 (1 - 0.078). So 92.2% of *nonsynonymous* mutation were removed by purifying selection.

Extracting a biological interpretation from this data

To make sense of this, we can look for enrichment for certain types of genes. We will use **Gene Ontology** enrichment. **Gene Ontology (GO)** is *a controlled vocabulary of gene and gene product attributes.* GO describes three different aspects of genes;

- **the biological process** a gene is involved in (eg: cell division)
- **the cellular component** that the protein is found in (eg: nucleus)
- **the molecular_function** of the protein (eg: DNA polymerase)

Today we will look at the **cellular component**.

My hypothesis is that proteins that are on the outside of the *Leishmania* cell are evolving more rapidly than those on the inside. I suspect this because I have observed this before with another single-celled parasite (*Plasmodium*)(Jeffares et al., 2007).

To examine this, we will examine the rapidly evolving genes first. These are in **Table 2 (Ldon weak purifying)**. To find out where the proteins produced from these genes are located in the cell, we will use the website TriTrypDB. As below:

1. Go to TriTrypDB and make an account for this website.
2. Then in the left-hand window "**Search for…**",
3. Click on **Genes** -> **Annotation, curation and identifiers** -> List of IDs
4. Copy the list of gene names from the '**Table 2 (Ldon weak purifying)**' tab of the Google Sheet (gene names are in the first column)
5. Paste it into the box in TriTrypDB: **Enter a list of IDs or text**
6. Click the **Get Answe**r button

Then:

7. Click Analyze results -> Gene Ontology Enrichment
8. Select **Cellular Component**
9. And click **Submit**

We show what your answer should be like below.

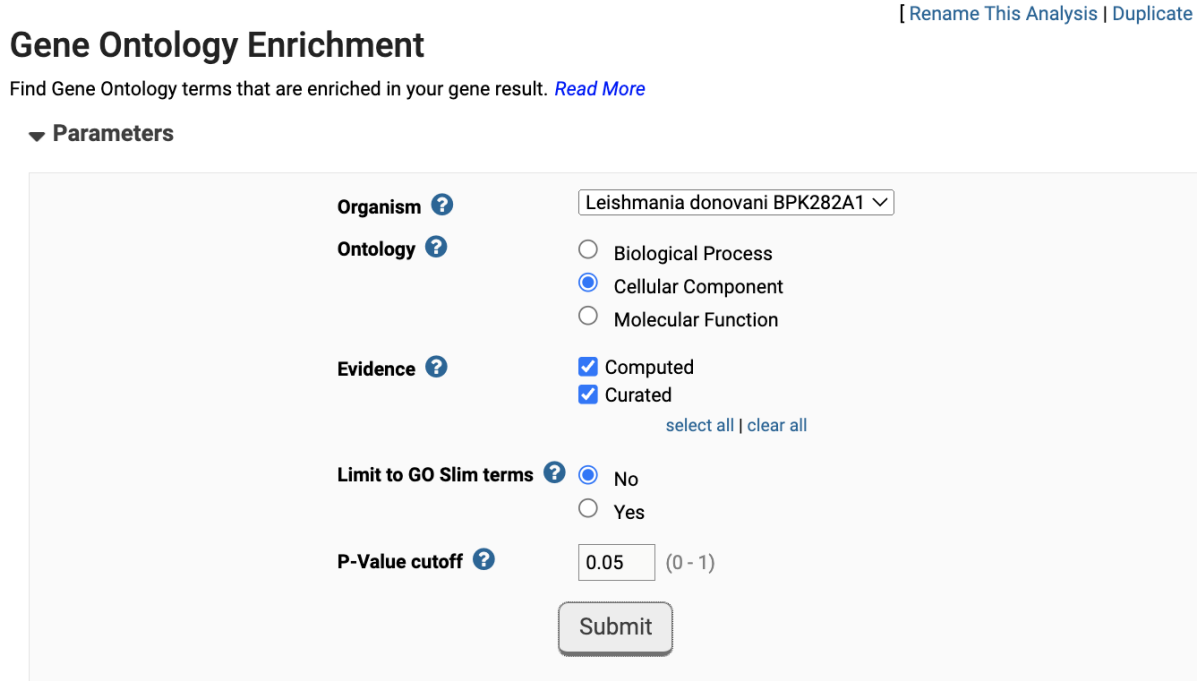
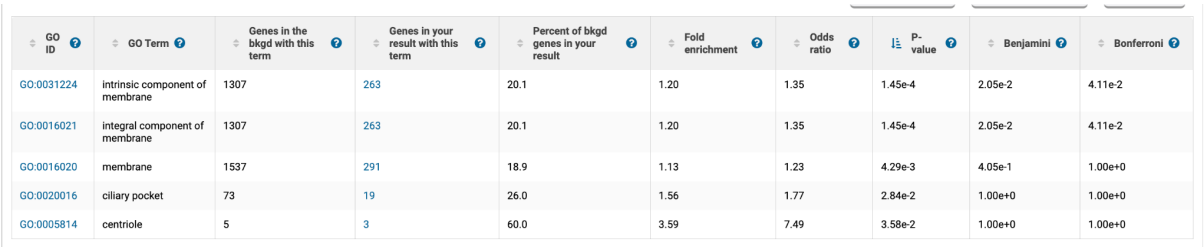


Figure 3. Your first search should look like this.



GO ID	GO Term	Genes in the bkgd with this term	Genes in your result with this term	Percent of bkgd genes in your result	Fold enrichment	Odds ratio	P-value	Benjamini	Bonferroni
GO:0031224	intrinsic component of membrane	1307	263	20.1	1.20	1.35	1.45e-4	2.05e-2	4.11e-2
GO:0016021	integral component of membrane	1307	263	20.1	1.20	1.35	1.45e-4	2.05e-2	4.11e-2
GO:0016020	membrane	1537	291	18.9	1.13	1.23	4.29e-3	4.05e-1	1.00e+0
GO:0020016	ciliary pocket	73	19	26.0	1.56	1.77	2.84e-2	1.00e+0	1.00e+0
GO:0005814	centriole	5	3	60.0	3.59	7.49	3.58e-2	1.00e+0	1.00e+0

Figure 4. Your first search results should look like this.

Let's talk through the results. On the left is the GO ID. If you click on this, it will explain what this GO term means (most are self explanatory). The two other important columns are:

- **Fold enrichment:** The percent of genes with this term in your result divided by the percent of genes with this term in the background. ie: how many times more of these genes are in our list than we expect.
- **P-value:** the P-value from Fisher's exact test. ie; How likely you are to see this level of enrichment by chance.

We see that our set of genes is enriched for the GO category **intrinsic component of membrane**, and some other categories. We will do a number of these GO enrichment tests, so I suggest that you open a document (excel, word, notepad etc) and keep some notes for what you find.

Examine the data: genes with stronger purifying selection

Now do the same GO enrichment test with the genes in **Table 3 (Ldon strong purifying)** of the Google Sheet. The simplest way to do this is to copy the first column of genes names, and go back to TriTrypDB main page, and start again.

Select **Cellular Component** again. This time, the results are simpler if you click **Yes** to **Limit to GO Slim terms.** Keep a note of what you see.

Consider the results

Discuss with your friends and/or us.

Where in the cell are the rapidly changing proteins more likely to be?
From Table 2 (Ldon weak purifying)

ANSWER: Within the external membrane of the cell, or in organells that are on the outside surface of the cell.

Where in the cell are the slowly changing proteins more likely to be?
From Table 3 (Ldon strong purifying)

ANSWER: Deep inside the cell,operating fundamental cellular processes like transcription, DNA replication and so on.

Why might that be?

ANSWER: We don't expect these to change rapidly, because even small changes may disrupt the fundamental cellular processes, and the fitness cost to such a change could be very high.

Protein conservation analysis *between species*

If diversity within species gives rise to divergence between species, we should see a similar pattern if we compare the changes *between species.* To study this, I calculated the *percent identity* of proteins between *L. donovani* and *L. braziliensis*, which is a reasonable measure of the evolutionary rate:

- **Genes with a high *percent identity*** are very similar, and so have been changing slowly between species, due to strong purifying selection.
    - These are in **Table 4 (Lbraz strong purifying)** of the Google Sheet
- **Genes with a low *percent identity*** have been changing rapidly between species, due to weaker purifying selection.
    - These are in **Table 5 (Lbraz weak purifying)** of the Google Sheet

Here is Figure 1 again, which shows how far apart these two species are in the phylogeny. We don't know exactly how long they have been diverging but I suspect it is at least 10,000 years.

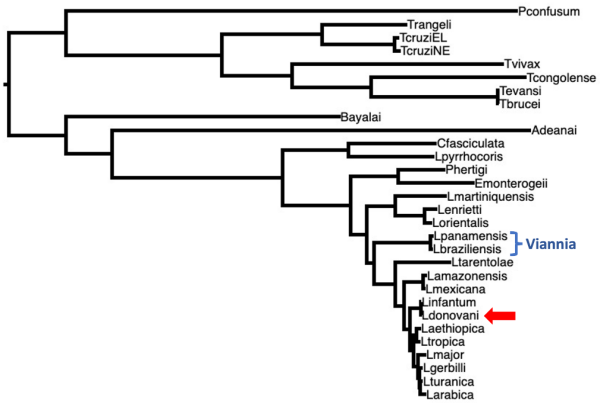


Figure 1. A phylogenetic tree of *Leishmania* and related species. Two Vianna species that we looked at last time in the Viannia clade are included (blue brackets). *Leishmania donovani*, which we will compare *Leishmania braziliensis* to, is indicated with a red arrow.

Examine the data

Now do the same GO enrichment tests, with:

- Table 4 (Lbraz strong purifying)
- Table 5 (Lbraz weak purifying)

Copy and paste the **Ldon gene**. Again, keep notes of what you have done, and what you find.

* The simplest way to do this is to copy the first column of genes names, and go back to TriTrypDB main page, and start again. Select **Cellular Component** again.

If you get too many results *without* clicking Limit to GO Slim terms, just look at the top results. Or click **Limit to GO Slim terms.**

Note: If you are not sure that the GO term means, click on the link.

Consider the results

Where are the slowly changing proteins more likely to be?

These are the highest % identity set from Table 4 (Lbraz strong purifying) These are genes that are subject to *strong purifying selection* to stay the same.

ANSWER: The ribosome is a common region. This is protein complex that is deep inside the cell, and carries out protein translation. This fundamental process is unlikely to change over long periods of time, because the fitness cost of disrupting it would be very high. We also observ similar patterns like intracellular organelle, cytoplasm and *intracellular* membrane-bounded organelle - and nothing on the *outside* of the cell.

Where in the cell are the rapidly changing proteins more likely to be?
These are the lowest % identity set from Table 5 (Lbraz weak purifying)

ANSWER: The membrane and the ciliary pocket (the invagination of the plasma membrane from which a cilium (also called flagellum) protrudes). Also the cell tip and cell pole. These are all parts ofthe cell that are on the outside.

Why might that be?

ANSWER: Proteins on the inside probably evolve slowly because they are protected from the external environment, and encode important, fundamental processes. Proteins on the outside often evolve more rapidly in single-celled species, because they adapt to changes in their environments. For pathogens their environments are a) the sandfly gut and b) the skin of the mammals they infect.

Are these results similar to the N/S results?

ANSWER: Yes, these results are very similar to the results from L. donovani, where we examine genetic diversity within a species.

Are the general patterns of selection *within species* similar to those *between species*?

ANSWER: Yes, very similar.

Optional: Protein conservation over deep time

The origin of eukaryotes is estimated to be about one billion years ago (Chernikova et al., 2011). So any proteins that have stayed the same over this long period of time must be *very important*. Let's find out what these proteins do. These will be the *most conserved* essential components of the cell. The things the cells just cannot afford to change.

Examine the data

To examine this, I compared the proteins from *Leishmania donovani t*o those of a yeast *Schizosaccharomyces pombe*, and calculated the **% identity**. I chose these

species because I know from a phylogenetic tree of eukaryotes that these species are far apart (**Figure 4**), so they have been evolving separately for a very long time.

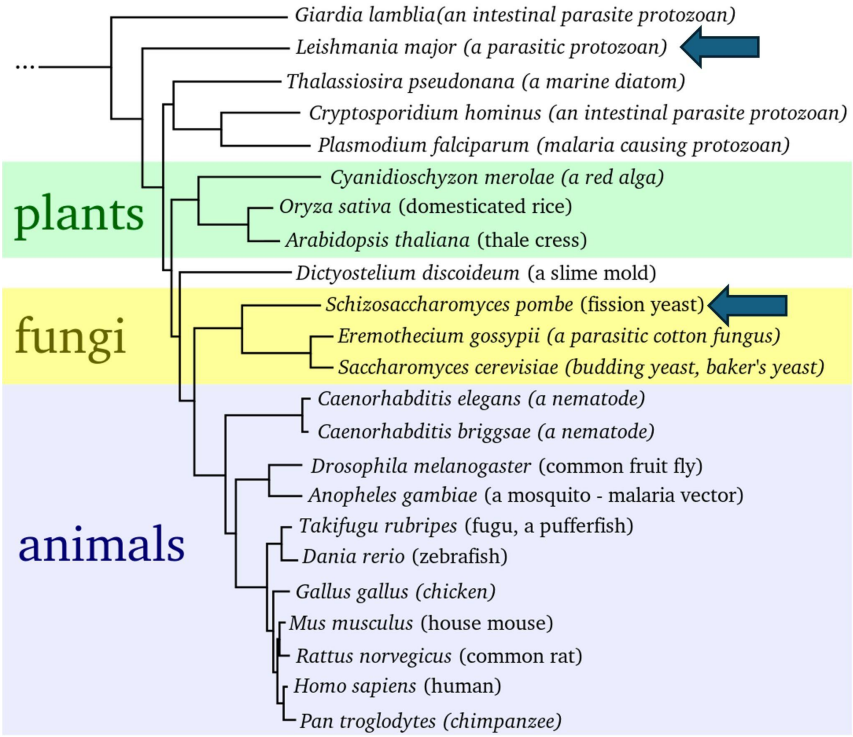


Figure 4. A phylogenetic tree of eukaryotes. *Schizosaccharomyces pombe* and a *Leishmania* species are present on the tree. Since *Leishmania* is very early branching, they are likely separated from *S. pombe* by approximately one billion years of evolution (Chernikova et al., 2011).

The results are in the **Ldon vs yeast all** tab of the Google Sheet. The most conserved proteins are only ~83% identical, so there have been some changes.

This time you don't even need to do the GO enrichment. Just paste the Ldon gene list into the list of IDs window at TriTrypDB and browse through the list of genes.

However, running a **Cellular Component** GO search with **Limit to GO Slim terms** switched **on** will give you a very interesting result.

Summary: what we have learned:

Today we:

- **Examined some population genomics data** by looking at the ratio of nonsynonymous to synonymous SNPs within *Leishmania donovani*.
    - We found some patterns in the data using Gene Ontology (GO) enrichment.This indicated that
    - Rapidly evolving genes were more like to be:

    - Slowly evolving (conserved) genes were more like to be:

- **Examined some comparative genomics data**, by looking at the % identity of proteins between two species of *Leishmania; L. donovani* and *L. braziliensis*.
    - Again, we found some patterns in the data using GO enrichment analysis.
    - These patterns were:
    - Rapidly evolving genes were more like to be:
    - Slowly evolving (conserved) genes were more like to be:

- **Examined some comparative genomics data comparing** *Leishmania* and a yeast *(Schzosaccharomyces pombe).* These species may have diverged a billion years ago.
    - We examined which proteins are conserved over time period of time.
    - This showed that:

End of the workshop

References

Bromham, L. and Penny, D. (2003). The modern molecular clock. *Nature reviews. Genetics*, 4 (3), pp.216–224. [Online]. Available at: doi:10.1038/nrg1020.

Chernikova, D. et al. (2011). A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes. *Biology direct*, 6, p.26. [Online]. Available at: doi:10.1186/1745-6150-6-26.

Jeffares, D. C. et al. (2007). Genome variation and evolution of the malaria parasite Plasmodium falciparum. *Nature genetics*, 39 (1), pp.120–125. [Online]. Available at: doi:10.1038/ng1931.

Reis-Cunha, J. L. et al. (2024). The global dispersal of visceral leishmaniasis occurred within human history. *bioRxiv*, p.2024.10. 30.621037. [Online]. Available at: doi:10.1101/2024.10.30.621037.