# BIO00056I

**Workshop 7: Comparative Genomics**

Daniel Jeffares

2025-11-18

## Table of contents

ℹ Note

This is a work in progress.

# 1 Learning objectives

- Learn more about interpreting comparative genomics data

- Learn about gene conservation over deep time (billions of years)
- To observe an example of the principle that *diversity within species gives rise to the divergence between species*

# 2 Introduction

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

## 2.1 The molecular clock

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 relatively closely related.

In this workshop, we continue looking at *Leishmania* data, but the observations we make will apply to any group of species.

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.

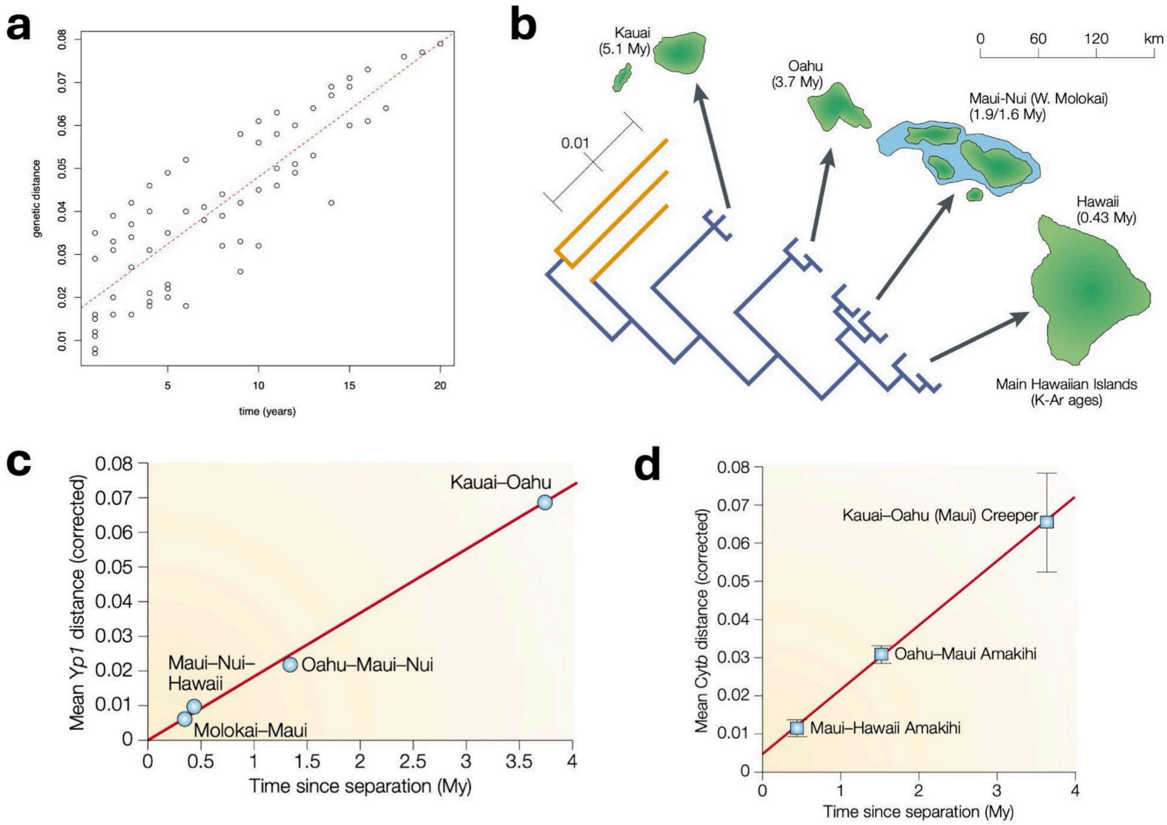

**Figure 1.** Evidence that molecular clock causes divergence between species over both short and long periods of time. Panel a: We saw from the influenza virus workshop that mutations occurred within influenza viruses at a regular rate with time over merely a fee decades. Panel b: This also occurs over long periods of time. For example, geologists know that the pacific islands of Hawaii have been formed over millions of years, emerging one after the other over thr last 5 mission years. Evolutionary biologists know that both honeycreeper birds and *Drosophila* fruit flies have populated these islands during this time, and from the geological data, 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). 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. Plots b-d are from Bromham and Penny, 2003.

## 2.2 Evolutionary change over deep time

We told you in the lectures that:

- The accumulation of mutations within species generates the genetic differences between species

- The different kinds of genes have different evolutionary rates, for example:
  - In a single-celled organism, proteins that are located on the outside of the cells tend to evolve relatively rapidly, to adapt to the external environment
  - Proteins that are located deep on the insides of the cells, performing fundamental processes (like DNA replication), tend to change slowly because they are so important

ℹ Are there consistent patterns of evolution within and between species?

- The accumulation of mutations within species generates the genetic differences between species
- Therefore, we might expect that the genes that evolve *evolve rapidly within species* would also *evolve rapidly between species*
- And the genes that change slowly *within species* should also change slowly *between species*.

**Today, we will examine these predictions**

# 3 Exercises

## 3.1 The phylogeny

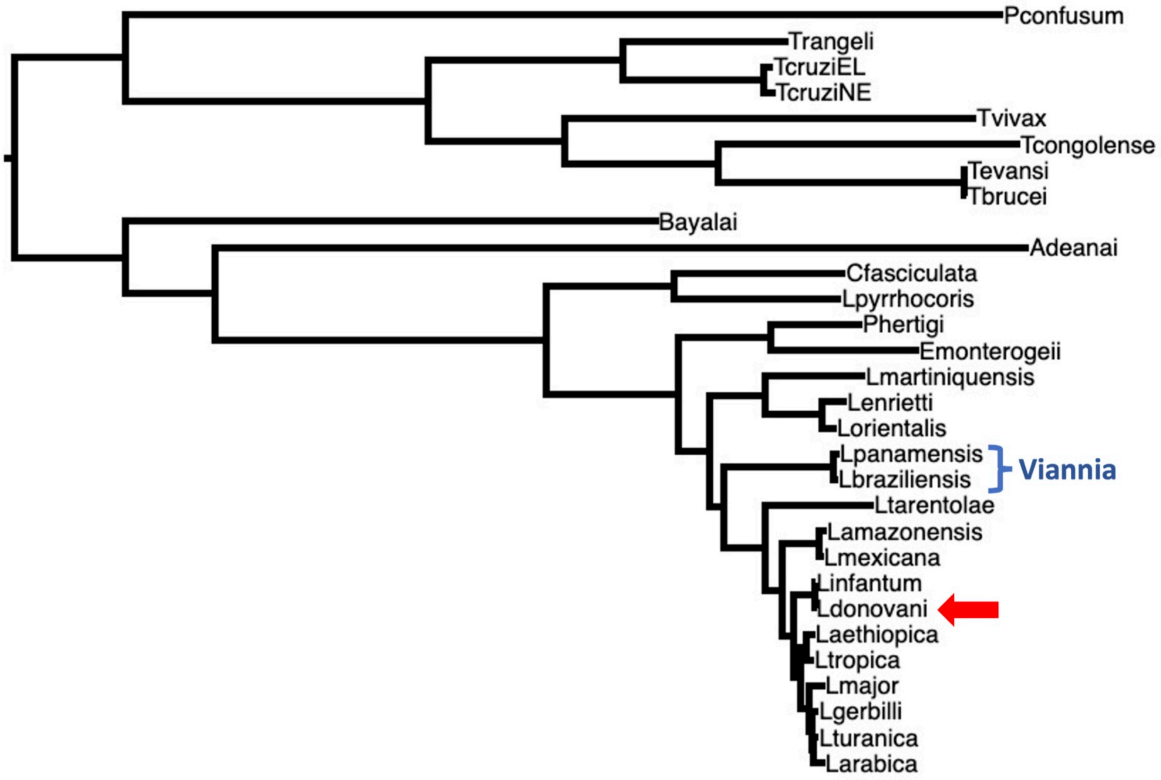

**Figure 2**. A phylogenetic tree of *Leishmania* and related species. Two of species that we looked at last time in the Viannia clade are included (blue brackets). Today, we will study genetic differences within the *L. donovani* species (red arrow). We will also look at genetic differences *between species*, by comparing *L. braziliensis* to *L. donovani.* We will also compare *L. donovani* proteins to a yeast species that diverged from *Leishmania* over a billion years ago (but yeast is not shown on this tree).

## 3.2 Genetic distances: shallow and deep time

Today, we will examine patterns of evolutionary change over three time scales, as summarised in Table 1 below. Our aim is to examine whether the same kinds genes tend to evolve rapidly or slowly over all these time scales.

Table 1: **Time scales**

| Time scale | Years (approximate) | Data comparison |
|---|---|---|
| Shallow time | thousands | genetic differences within *L. donovani* |
| Intermediate | millions | genetic differences between *L. donovani* and *L. braziliensis* |
| Deep time | billions | protein differences between *L. donovani* and yeast |

### 3.2.1 Synonymous and non-synonymous mutations

The SNPs we downloaded were classified as either:

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

We expect the **synonymous** mutations to be more or less neutral (ie: not subject to selection), because they don't change the protein, so we can think of these as representing what occurs without selection, like a control.

The **non-synonymous** mutations that do change the protein, will often be subject to selection. Since most new mutations are harmful, we expect that many non-synonymous mutations will be removed by **purifying selection**. Very occasionally, a non-synonymous mutation will be beneficial, and will be favoured by **positive selection**.

To understand selection within a gene, we can calculate ratio of non-synonymous (selection) / synonymous (control) ratio. We will call this the **N/S ratio**. I show some toy examples of N/S ratios in Table 2 below.

*NB: More complex analyses call this ratio the dN/dS ratio, or the Ka/Ks ratio.*

Table 2: **N/S ratio examples**

| Gene | synonymous SNPs | non-synonymous SNPs | Interpretation |
|---|---|---|---|
| A | 1 | 10 | far fewer non-synonymous than synonymous SNPs: **purifying selection** has removed many non-synonymous SNPs |
| B | 10 | 11 | non-synonymous and synonymous SNPs about the same: **no selection** |
| C | 20 | 10 | more non-synonymous than synonymous SNPs: **positive selection** has selected for non-synonymous SNPs |

Interpreting the N/S ratios

- **Gene A**: The N/S ratio is very low (N=1, S=10, N/S = 0.1). Since we observe many fewer non-synonymous (amino-acid changing) SNPs than synonymous SNPs **strong purifying selection** has probably been acting on this gene.
- **Gene B**: The N/S ratio is near to 1 (N=10, S=11, N/S = 0.9). Since we observe the same number of non-synonymous (amino-acid changing) SNPs as synonymous SNPs it is likely that **no selection** has been acting on this gene.
- **Gene C**: The N/S ratio is high (N=20, S=10, N/S = 2). Since we observe more non-synonymous (amino-acid changing) SNPs than synonymous SNPs **strong positive selection** has probably been acting on this gene

NB: These are toy examples. We would seldom observe such extreme N/S values in real data.

### 3.3 Shallow time: N/S ratios

First lets look at Figure 3, which shows the distribution of N/S ratios within *L. donovani* genes.

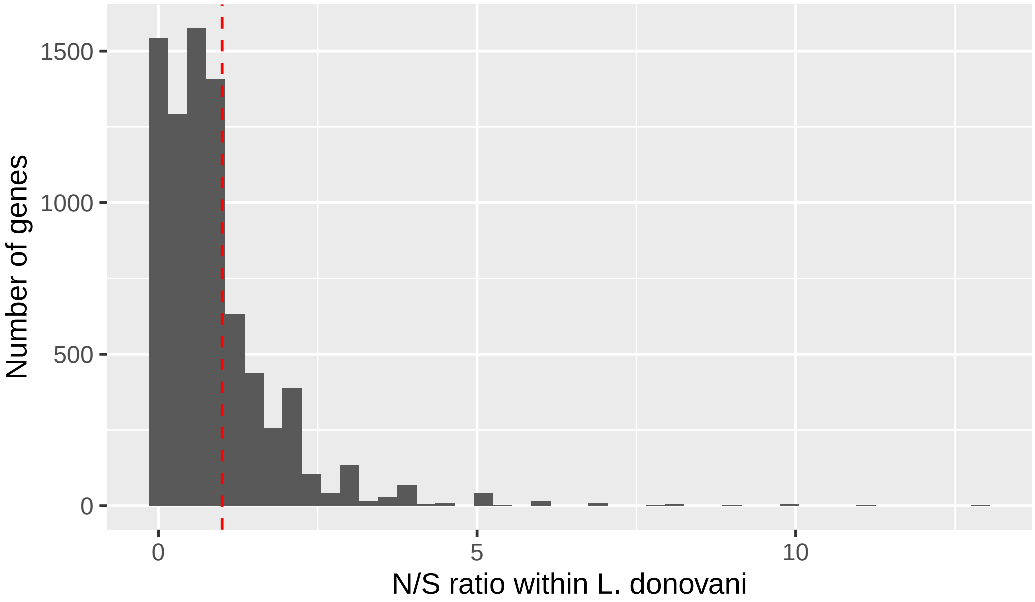

**Figure 3. The distribution of N/S ratios within L. donovani genes**.

Discussion points: N/S ratios in L. donovani genes

- Most genes have N/S ratios less than 1 (left of the red dashed line). What does this indicate?
- A few genes have N/S ratios greater than 1 (right of the red dashed line). What does this indicate?
- In total, there are `57886` synonymous SNPs. We can assume that these are not subject to strong selection. In contrast, there are `48460` non-synonymous SNPs. So we have 83% as many **non-synonymous SNPs** as **synonymous SNPs**. Where did the other 17% of non-synonymous SNPs go?

## 3.4 Shallow time: what kinds of genes have high or low N/S ratios?

If a gene has a low N/S ratio, this indicates that it is subject to strong purifying selection. If a gene has a high N/S ratio, this indicates that it is subject to positive selection.

To understand what kinds of genes have high or low N/S ratios, I divided the genes into 2 groups:

- **High N/S ratio genes**, with N/S ratio $> 1.3$, that *might be* subject to positive selection

- **Low N/S ratio genes**: with N/S ratio < 0.15, that are probably subject to purifying selection

Note: I say *might be*, because N/S ratios can be noisy, and positive selection can be quite difficult to prove.

There are 1714, genes with N/S ratio > 1.3. Here is a list of those with the highest N/S ratios - it doesn't tell us much does it!

```
 [1] "vacuolar-type Ca2+-ATPase, putative (fragment)"
 [2] "Acyl CoA binding protein, putative"
 [3] "NUDIX hydrolase dihydroneopterin triphosphate pyrophosphohydrolase/hydrolase, putative"
 [4] "RNA-binding protein, putative"
 [5] "c2 domain protein, putative"
 [6] "methionyl-tRNA formyltransferase, putative"
 [7] "pyroglutamyl-peptidase I, putative"
 [8] "Ssl1-like, putative"
 [9] "DnaJ domain containing protein, putative"
[10] "serine/threonine phosphatase, putative"
```

To explain what kinds of genes have high N/S ratios, we can use a technique called **gene ontology (GO) enrichment analysis**.

gene ontology (GO)

- Gene Ontology (GO) is a system for classifying genes according to their biochemical functions, the biological processes they are involved in, and the cellular locations that the proteins operate in (eg: nucleus, membrane, etc).
- If we have a gene list, we can use GO enrichment analysis to look for patterns in the kinds of genes that are present in the list.

### 3.4.1 Shallow time: gene enrichment analysis

I ran a gene enrichment analysis, looking for enrichment of in the **cellular locations** of the **high N/S ratio genes**, that might be subject to positive selection. I also ran an enrichment analysis on the **low N/S ratio genes**, that are subject to strong purifying selection. The results are shown in Figure 4 (positive selection) and Figure 5 (purifying selection) below.

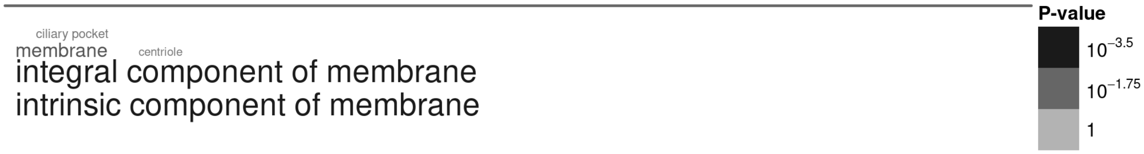

**Figure 4. A word cloud showing the cellular locations enriched in positive selection genes.** I selected genes with high N/S ratios ($> 1.3$), and ran the gene enrichment analysis using TriTrypDB. Larger words indicate greater enrichment, and the grey text intensity indicates the statistical significance of the enrichment (darker = more significant, paler = less significant).

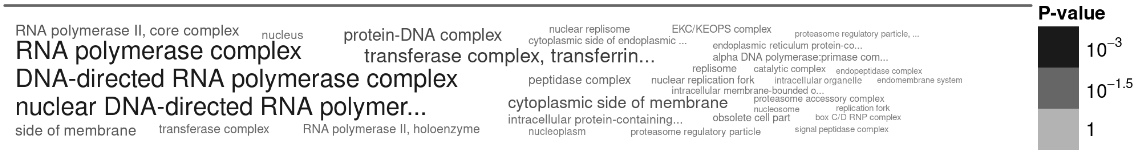

**Figure 5. A word cloud showing the cellular locations enriched in genes that are subject to strong purifying selection**. I selected genes with low N/S ratios ($< 0.15$), and ran the gene enrichment analysis using TriTrypDB. Larger words indicate greater enrichment, and the grey text intensity indicates the statistical significance of the enrichment (darker = more significant, paler = less significant).

Discussion points: gene enrichment analysis

- What kinds of cellular locations do the proteins made by rapidly evolving (positive selection) genes tend to reside in (Figure 4)?
- What kinds of cellular locations do the proteins made by slowly evolving genes tend to reside in (Figure 5)?
- Do these results make sense to you? Why might genes in certain cellular locations

be subject to different types of selection? Consider two categories of proteins:

- centrally located, important proteins
- peripheral, environment exposed proteins.

# 4 Summary: what we have learned

# 5 After the workshop: exam style questions

## 5.1 Question 1.

## 5.2 Question 2.