EBIO 3080 Lab Week 5

Dylan Oh, Kayden Adams, Yousef Abdullah Al Obaidan
24 September 2020

D_n/D_s Worksheet

	Codon	1			2			3			4			5			6			7			8			9			10		
	Nucleotide site number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
	Sequence A	A	T	G	A	A	A	G	T	T	C	C	C	C	T	C	G	T	C	A	T	C	G	C	C	T	C	C	T	T	T
	Syn Site	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.67	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.33
	Nonsyn site	1.00	1.00	1.00	1.00	1.00	0.67	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.33	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.67
Total syn site =	7.33																														
Total nonsyn site =	22.67																														
Total sites, should = 30	30.00																														
	Sequence A	A	T	G	A	A	A	G	T	T	C	C	C	C	T	C	G	T	C	A	T	C	G	C	C	T	C	C	T	T	T
	Sequence B	-	-	C	-	-	T	-	A	-	-	-	A	-	-	-	A	-	-	-	C	-	-	-	T	-	-	-	-	-	C
Total syn diff =	3.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00
Total nonsyn diff =	5.00	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Ds=	0.409090909																														
Dn=	0.220588235																														
Dn/Ds=	0.539215686																														

Sliding Window Analysis

This section will look at the D_n/D_s ratio for both the *CO1* and *lysin* genes and compare them for a window of 20 consecutive codons.

```
# Reading CSV data and calculating means
CO1 <- read.csv(file.path(dest, "COIDnDs.csv"), header = TRUE)
lysin <- read.csv(file.path(dest, "LysinDnDs.csv"), header = TRUE)
mean_CO1 <- mean(CO1$DnDs_windowed, na.rm = TRUE)
mean_lysin <- mean(lysin$DnDs_windowed, na.rm = TRUE)
```

```
# Plotting Dn/Ds vs codons for lysin and CO1
fig1 <- ggplot() +
  geom_line(data = CO1,
            size = 1.2,
            aes(x = Codon, y = DnDs_windowed, color = "#72b46d")) +
  geom_line(data = lysin,
            size = 1.2,
            aes(x = Codon, y = DnDs_windowed, color = "#ad88d3")) +
  geom_hline(yintercept = 1) +
  scale_color_identity(name = "",
                       breaks = c("#72b46d",  "#ad88d3"),
                       labels = c("Lysin", "CO1"),
                       guide = "legend") +
  ggtitle(TeX("$\\mathrm{D_n/D_s}$ of Lysin and CO1 Genes")) +
  xlab("Codon") +
  ylab(TeX("$\\mathrm{D_n/D_s}$")) +
  theme(plot.title = element_text(color = "#2c3136", size = 20,
                                  family = "hkgroteskbo"),
        axis.title.x = element_text(color = "#2c3136", size = 13,
                                    family = "hkgroteskli"),
        axis.title.y = element_text(color = "#2c3136", size = 13,
                                    family = "hkgroteskli"),
        legend.position = "right")
print(fig1)
```

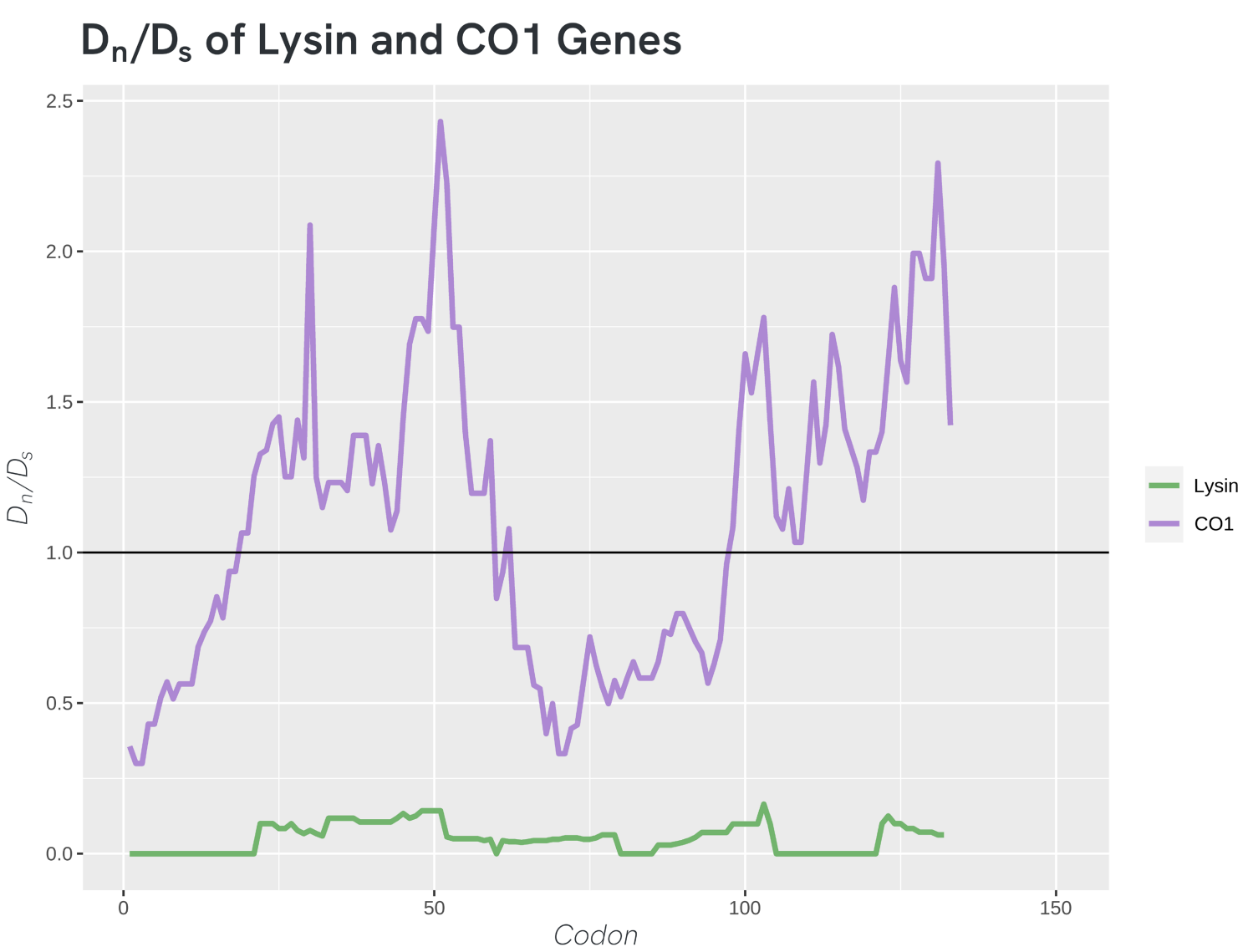

Figure 1. Plot of D_n/D_s values calculated for a window of 20 consecutive codons of *lysin* and *CO1* genes. A horizontal line at $D_n/D_s = 1$ shows the expectation in the case that there is no selection on the sequences.

From the figure, it is clear that the *lysin* gene has a much higher D_n/D_s $(\bar{x} = \mathbf{1.1})$ than the *CO1* gene $(\bar{x} = \mathbf{0.052})$. This implies that there is a greater level of nonsynonymy when mutations occur in the *lysin* gene than in *CO1*. The implications of this in terms of selection are that *CO1* mutations lead to purifying selection ($D_n/D_s < 1$); the phenotypic effect of changing that gene negatively affects the fitness of the individual, and because of this selection pressures will eventually eliminate the mutation from the genome. Conversely, the observation in the *lysin* gene ($D_n/D_s > 1$) suggests the opposite; the mutations in this gene tend to positively affect the fitness of the individual and selection pressures will eventually proliferate the mutation throughout a population.

Comparative Analysis

This section will look at the comparisons between the windowed D_n/D_s found in the previous section and the D_n/D_s obtained from randomly selected genes throughout the *D. melanogaster* genome.

```
# Reading CSV data and calculating percentiles/means
DnDs <- read.csv(file.path(dest, "FlyDnDs.csv"), header = TRUE)
dnds <- DnDs$DnDs
per_CO1 <- signif(length(which(mean_CO1 >= dnds))/length(dnds)*100, 2)
per_lysin <- signif(length(which(mean_lysin >= dnds))/length(dnds)*100, 2)
random <- mean(dnds, na.rm = TRUE)
```

```
# Plotting a histogram of D. melanogaster Dn/Ds
fig2 <- ggplot() +
  geom_histogram(data = DnDs,
                 color = "white", fill = "#5e6164",
                 aes(x = DnDs)) +
  geom_vline(aes(xintercept = 1, color = "#aaaaaa"), size = 1) +
  geom_vline(aes(xintercept = as.numeric(mean_CO1),
                 color = "#72b46d"),
             size = 1) +
  geom_vline(aes(xintercept = as.numeric(mean_lysin),
                 color = "#ad88d3"),
             size = 1) +
  scale_color_identity(name = "",
                       breaks = c("#72b46d", "#aaaaaa", "#ad88d3"),
                       labels = c("Lysin Average",
                                  "Random Gene Average",
                                  "CO1 Average"),
                       guide = "legend") +
  ggtitle(TeX("$\\mathrm{D_n/D_s}$ for Randomly Selected Drosophila Genes")) +
  xlab(TeX("$\\mathrm{D_n/D_s}$")) +
  ylab("Frequency") +
  theme(plot.title = element_text(color = "#2c3136", size = 20,
                                  family = "hkgroteskbo"),
        axis.title.x = element_text(color = "#2c3136", size = 13,
                                    family = "hkgroteskli"),
        axis.title.y = element_text(color = "#2c3136", size = 13,
                                    family = "hkgroteskli"),
        legend.position = "bottom",
        legend.key.size = unit(0.8, "line"))
print(fig2)
```

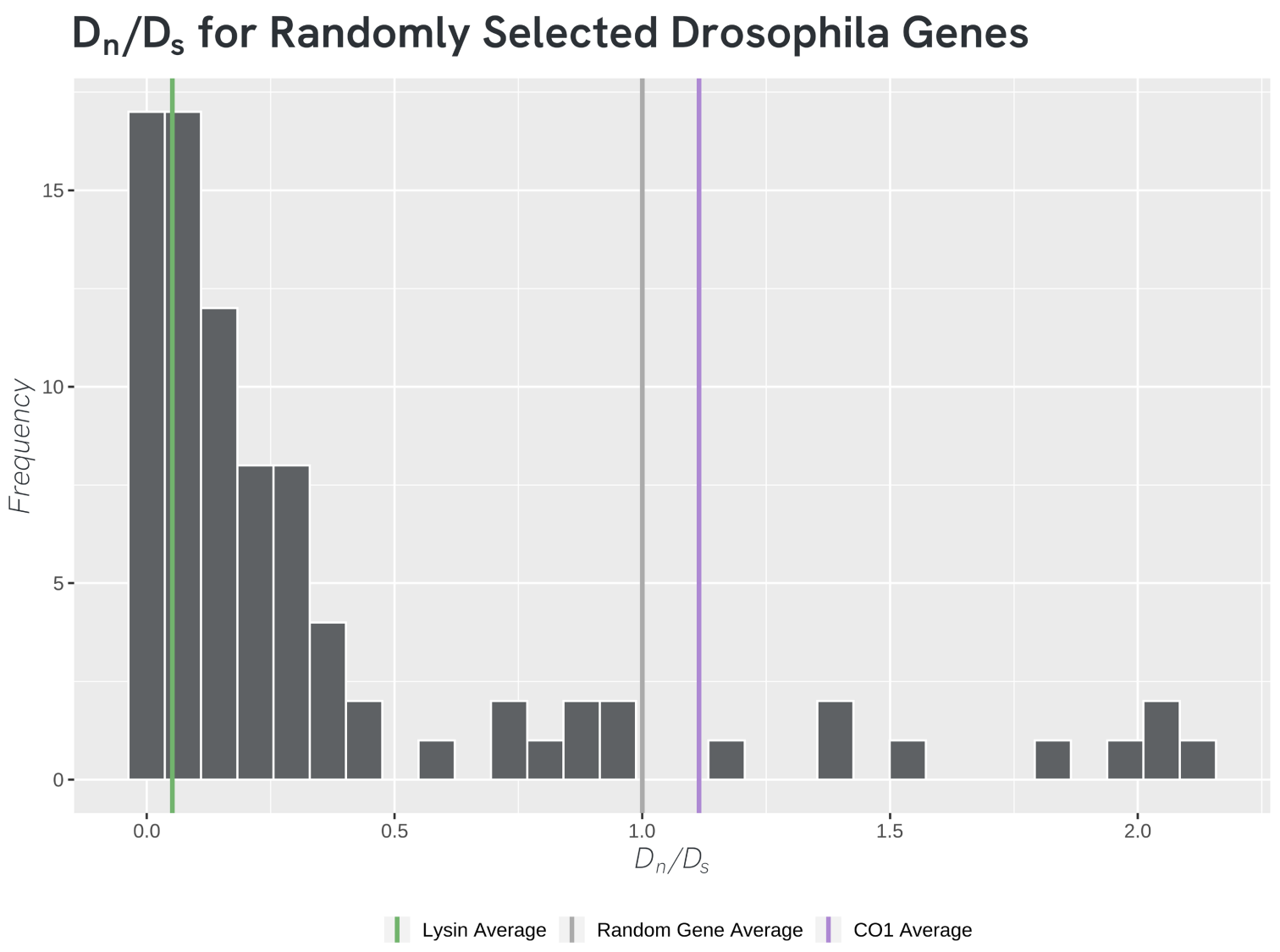

Figure 2. Histogram of D_n/D_s for randomly selected genes from the *D. melanogaster* genome.

This figure shows a comparison between the D_n/D_s of randomly selected genes throughout the genome of *D. melanogaster* and the average D_n/D_s of changes in the *CO1* and genes in order to infer the fitness consequences of mutations to both those genes. It indicates that mutations in the *D. melanogaster* genome tend to result in synonymous changes because the nonsynonymous, deleterious changes that negatively impact fitness have been removed from the genome through selection. Compared to the random-gene D_n/D_s observed in *D. melanogaster*, changes to the *CO1* gene tend to result in strongly purifying selection and the gene itself is in the **25th percentile** relative to the distribution of D_n/D_s values across the genome. The *lysin* gene is in the **89th percentile** relative to the distribution of D_n/D_s values across the genome. This implies that *CO1* gene mutations tend to result in negative fitness consequences and are therefore gradually eliminated from the genome through selection. The reverse is observed in *lysin* gene mutations.

Making Claims and Proposing Hypotheses

Claim About *CO1*

From the data, it is clear that the selection operating on the *CO1* gene is purifying; the low **D_n/D_s = 0.052** is evidence of this, because this implies that all the nonsynonymous mutations in the gene were eliminated through selection. That behaviour further suggests that mutations in *CO1* tend to be deleterious to fitness. There is also less heterogeneity of selection, which can be confirmed visually and computationally (**0 < Dn/Ds < 0.17**).

Claim About *Lysin*

Lysin, on the other hand, is generally under positive selection, with a mean $\mathbf{D_n/D_s = 1.1}$. Because this ratio is over 1, the nonsynonymous mutations outnumber synonymous ones, and are further typically beneficial to fitness. Natural selection therefore increases the frequency of the mutation throughout a population. There is, however, much more heterogeneity of selection, as the ratio oscillates greatly ($\mathbf{0.30 < D_n/D_s < 2.4}$).

Claim About the Comparative Analysis Graph

The *CO1* gene is ordinary in comparison to the random selection of genes from *D. melanogaster*, which is evidenced by their similar means. Conversely, *lysin* is quite extraordinary not only because of its significantly higher average D_n/D_s value but also because of its high heterogeneity; both *CO1* and the *D. melanogaster* ratios tend to be homogeneously low.

Hypothesis on the Variability of *Lysin*

Lysin is a protein that coats sperm and allows it to bind to receptor proteins on eggs. The variability of *lysin* could be attributed to sexual conflict; the idea of the sexual "arms race" between male and female members of a species. The fitness of an individual with a mutation in their lysin gene is also codetermined by its compatibility with the corresponding female gene, and therefore the male gene must continue to adapt to increase fecundity just as the female gene adapts to reduce it. Mutations in the male gene offer what are essentially experimental tactics to overcome female counteradaptation, because if the gene were to remain conserved, it would otherwise be ineffective and unnecessary.

Hypothesis on the Conservation of *CO1*

Cytochrome oxidase I is a protein involved in oxidative phosphorylation, which is a vital metabolic process in cells. It has remained conserved, or unchanged, across all species for so long because a nonsynonymous mutation within the gene would likely have fatally disastrous consequences for the host individual's ability to produce energy, effectively eliminating their ability to reproduce and proliferate the mutation.