Code

Exploring the Evolution of Luciferase1 and its Relationship to Emission Spectra

Auden Block

04/12/2023

0.0.1 Abbreviation Key

Table 0.1: The full name and abbreviation of all organisms used in this study.

Full Name	Abbreviation
Drilaster axillaris	*D. axillaris*
Stenocladuius azuma	*S. azuma*
Cyphonocerus ruficollis	*C. ruficollis*
Lucidina biplagiata	*Ln. biplagiata*
Pyrocoelia miyako	*Py. miyako*
Aquatica lateralis	*A. Lateralis*
Phausis reticulata	*Pa. reticulata*
Luciola cruciata	*Ll. Cruciata*
Lampyris turkestanicus	*Lp. Turkestanicus*
Photinus pyralis	*Pt. pyralis*
Luciola parvula	*Ll. Parvula*
Luciola italica	*Ll. Italica*
Luciola mingrelica	*Ll. Mingrelica*
Phrixothrix hirtus	*Phr. Hirtus*

1 Introduction

Bioluminescence, or the ability to produce and emit light by a living organism, is thought to be shared across five families of beetles: Elateridae, Lampyridae, Phengodidae, Rhaphthalmidae, and Sinophyrophoridae (Powell et al., 2022). Across these families, research has shown that it has emerged independently twice: once in click beetles (Elateridae) and another time in an ancestor of the lampyroid clade: Lampyridae, Rhagophtalmidae, Sinophyrophoridae, and Phengodidae (Fallon et al., 2018;

Kusy et al., 2021; Martin et al., 2017). Furthermore, it is thought to be part of a gene duplication event at a common ancestor of the Lamyridae lineage, implicating that all fireflies have two bioluminescent genes: Luc1 and Luc2 [1.1]. However, so far this Luc2 isotype has only been isolated in *Ll. cruciate* and *Ll. parvula* (Bessho-Uehara & Oba, 2017). This duplication allowed for fireflies to develop two different proteins that can be expressed for different needs. It is thought that Luc1 is predominately expressed in larvae, prepupae, pupae, and adults. Due to the match in visual sensitivity of *Ll. parvula* eyes to that of Luc1, its role is believed to be for intraspecific communication, unlike that of Luc2. Currently, only a few species have had the Luc2 gene isolated, but it is believed that the extant Luc2 is used to express a green glow in the eggs, prepupae, pupae, and adult females of *Ll. cruciate* and *Ll. lateralis* (Bessho-Uehara & Oba, 2017).

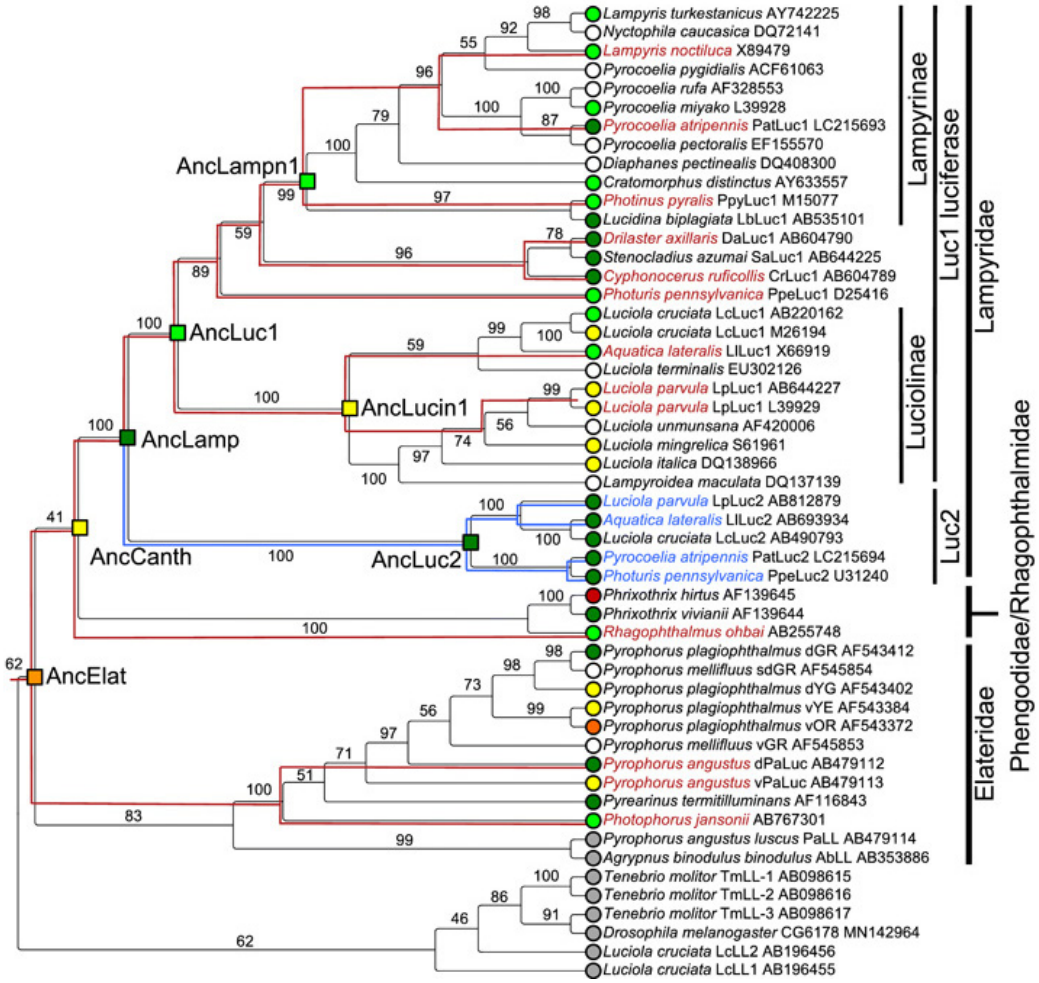


Figure 1.1: Molecular phylogeny of luciferases and related enzymes. The leaf nodes are labeled with species name, protein name, and GenBank accession number. Branches are labeled with bootstrap probability (1000 reconstructions). The resurrected ancestral nodes are shown as a square. The leaf nodes are indicated with in vitro luminescent colors (green, yellow-green, yellow, orange, or red) judged by the luminescence maximum values: Green, GR, 520-549 nm; Yellow-green, YG, 550-559 nm; Yellow, YE, 560-584 nm (Oba et al., 2020).

To produce light, the firefly luciferin is converted into an excited state oxyluciferin product (Figure 1.2). Surprisingly, despite all fireflies using the same luciferin substrate, the wavelength emitted varies across the phylum (Branchini et al., 1999b). Typically, fireflies emit a peak emission spectra in vivo between 540-580nm (green to yellow light) (Hall et al., 2016; Navizet et al., 2010; Ugarova & Brovko, 2002). This difference in emission has a couple of different hypothesized reasons in fireflies. The most prominent factor researchers believe to be responsible is that amino acid substitutions in the active sites of Luciferase proteins results in a substantial alteration to a firefly's peak emission spectra (Branchini et al., 1999b; Morton et al., 1969). Additionally, numerous researchers have shown that site specific amino acid changes results in a shift of the luciferase peak emission wavelength *in vitro* (Branchini et al., 2001; Shapiro et al., 2005; Wang et al., 2013). However, while these single point mutagenesis models show that these changes do influence the emission spectra emitted by firefly bioluminescence, firefly species have a significant number of mutations across the entire protein. As such, a couple of expected models can be generated. The first is that the entire protein complex matters when determining color, meaning the more amino acid differences between species, the greater the difference in absolute emission spectra. Or, an alternative could be that a firefly's

light emission is dependent on the proteins at specific sites where the enzyme and substrate interact (known as active sites). In this system, it isn't the total number of changes, but the changes that occur at these active sites that influence the emission color of a firefly.

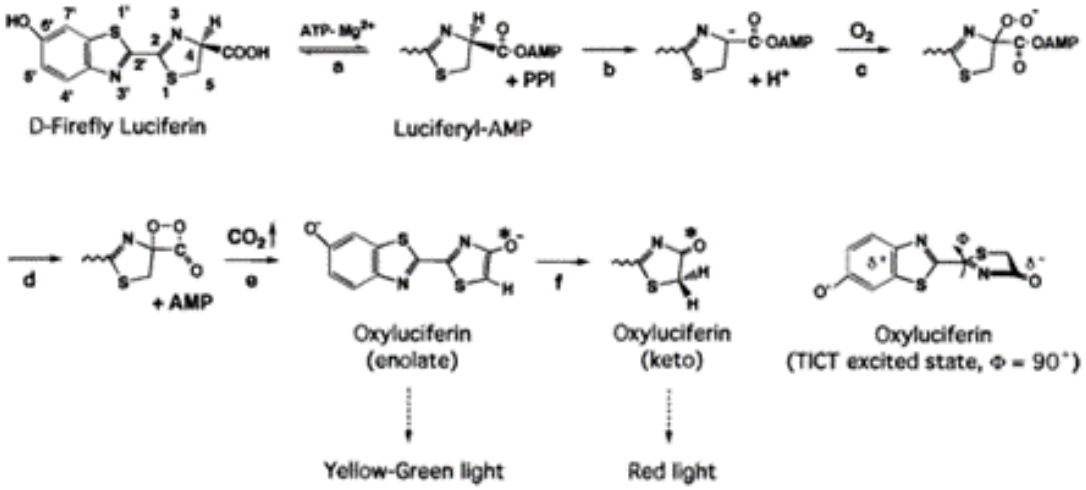


Figure 1.2: Mechanism of firefly luciferase catalyzed bioluminescence and TICT excited state of oxyluciferin (Branchini et al., 1999b).

In this paper, I analyze the evolutionary history of Luc1 within a subset of sixteen firefly species to elucidate a potential correlation between the protein structure of each species and the absolute difference in peak emission spectra. First, a phylogenetic tree was generated to determine potential phylogenetic correlation between species of interest and the wavelength emitted. Next, the firefly proteins were analyzed pairwise to determine if a correlation between the absolute difference in peak emission spectra and the differences in three different groups: all amino acids, non-active sites, and only active sites. I found that there is no evolutionary correlation and little, if any, pairwise correlation between these groups and that the peak emission spectra of a firefly is likely determined by a multitude of factors, not just the protein structure.

2 Methods

2.1 Phylogenetics

Due to the redundancy of genetic code, there are different evolutionary pressures placed on different codon positions. Since the third codon is the least functionally constrained, mutations are more likely to occur there and be passed to the next generation (Bofkin & Goldman, 2007). This becomes a problem when attempting to create a phylogeny of pairwise distances, as these mutations could be cases of convergent evolution, causing improper clustering. As such, sixteen known Luc1 protein sequences (Table 2.1; Supplementary Table 6.1), were imported into Geneious Prime to have the *O*pen *R*eading *F*rame (ORF) extracted (for explicit steps see 6.4.1).

Table 2.1: Organisms Studied and supplementary data about each. *Green, GR, 520-549 nm; Yellow-green, YG, 550-559 nm; Yellow, YE, 560-584 nm. (Modified from: Oba et al. (2020))

Species	**GenBank No.**	**Imax nm**	**Colouration***
Drilaster axillaris	AB604790	545	GR
Stenocladius azumai	AB644225	545	GR
Cyphonocerus ruficollis	AB604789	546	GR
Lucidina biplagiata	AB535101	549	GR
Pyrocoelia miyako	L39928	550	YG
Aquatica lateralis	X66919	552	YG
Phausis reticulata	KU600949	552	YG

Species	GenBank No.	lmax nm	Colouration*
Luciola cruciata	AB220162	554	YG
Lampyris turkestanicus	AY742225	555	YG
Photinus pyralis	M15077	557	YG
Photinus pyralis	AB644228	557	YG
Luciola parvula	AB644227	561	YE
Luciola cruciata	M26194	562	YE
Luciola italica	DQ138966	566	YE
Luciola mingrelica	S61961	566	YE
Luciola parvula	L39929	568	YE
Phrixothrix hirtus	AF139645	622 (pH 8.0)	RE

Next, these ORFs are brought into R to have the first and second codon position removed, leaving only the third codon to be analyzed.

Hide

```
#Import Table.
  ORFs <- read.table("Phylogenetics/ORFs.txt", sep=",", header=FALSE)
#Blank Vector of the final nucleotides.
  only.third <- c()

#Third Codon Removal.
for (species in ORFs$V2) {
  #Split the nucleotides into single letter strings.
    split <- strsplit(species,split = "")
  #Extract the third codon.
    extracted <- str_remove_all(toString(split[[1]][seq(3, length(split[[1]]), 3)]), ", ")
  #Order of Operations:
    #create a string counting every third position into a new vector, remove those indexes into a
new vector.
    #Add the string as a new index.
    only.third <- c(only.third, extracted)

}
#Remove the intermediate vectors made in for-loop.
rm(split, extracted, species)

#Dataframe of the species name and the third codon-only files.
third.codons.df <- data.frame(ORFs$V1, only.third)


#Modified from: TrainingPizza, 2021
  #Create a vector that can be exported as a .fasta
  third_codons_print <-
    third.codons.df %>%
    rowwise() %>%
    pivot_longer(ORFs.V1:only.third) %>%
    select(-name)

  #Export the dataframe as a .fasta.
  write.table(third_codons_print,
              file = "Phylogenetics/Thirdcodons_only.fasta",
              col.names = FALSE,
              row.names = FALSE,
              quote = FALSE)
```

These third codon-only sequences were imported into BisonNet to construct a phylogenetic tree. Within the program IQ-Tree, ModelFinder Plus was used to determine the best substitution model that fits the nucleotide sequences using Akaike Information Criterion (AIC), removing the potential for convergent evolution to influence the phylogeny (Minh et al., 2019). The chosen model is then used to assemble a tree within IQ-tree. The outgroup, *Phr. hirtus* (Accession AF139645), was selected as its luciferase protein is known to be derived from the same ancestral luciferase as Lampyridae (Oba et al., 2020). The .treefile was then imported into Geneious Prime and manually color-coded.

Hide

```
#!/bin/bash
#SBATCH -p short # partition (queue)
#SBATCH -N 1 # (leave at 1 unless using multi-node specific code)
#SBATCH -n 8 # number of cores
#SBATCH --mem-per-cpu=32G # memory per core
#SBATCH --job-name="IQtree" # job name
#SBATCH -o slurm.%N.%j.stdout.txt # STDOUT
#SBATCH -e slurm.%N.%j.stderr.txt # STDERR
#SBATCH --mail-user=_____ # address to email
#SBATCH --mail-type=ALL # mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --exclude=hpc-4,hpc-5,hpc-6


#Load module:
module load phylogeny


#Variables:
barcode=/home/arb027/CAPSTONE/ThirdCodon_alignment.fasta #The input file.
merit=AIC #Akaike Information Criteria Metric for IQ-Tree ModelFinder Plus.
outgroup=RE_Phr_hirtus_AF139645_2 #Sequence Identifier of the outgroup.
bootstrap=10000 #Number of bootstraps.

#Print variables to console (for user references):
cat <<OPTIONS
Alignment file: $barcode
Search for the best model of sequence evolution using: $merit
Outgroup: $outgroup
Number of Bootstraps: $bootstrap
OPTIONS

#IQtree:
    #Documentation: http://www.iqtree.org/doc/Command-Reference
iqtree -s $barcode -merit $merit -o $outgroup  -bb $bootstrap

#Unload module:
module unload phlogeny
```

2.2 R Analysis

The original nucleotide ORFs for the firefly proteins (were translated into protein sequences and aligned using MUSCLE (algorithm PPP, HMM Pertubations = 0, Guide Tree Purmutations = 0) in Geneious Prime to be imported into R. While 17 Luc1 proteins have previously been identified by Oba et al. (2020) and were used to construct the phylogenetic tree, the *Pt. pyralis* Luc1 mRNA (Accession: M15077) only has 182 amino acids, while the other proteins have around 560 proteins. Additionally, later analysis looks into known active sites of the Luc1 protein using data from *Pt. Pyralis*, where the first known active site is at AA position 197. Even when aligned with MUSCLE, a significant majority of the outlined active sites are not aligned with the other proteins (Supplementary Figure 6.2. Further research into this mRNA protein would be required to determine where the active sites are located within this specific protein. However, there is another known luc1 mRNA protein for Pt. pyralis (Accession: AB644228) that does properly align to other proteins (Supplementary Materials 6.2). By looking across the previously outlined categories in Supplementary Figure 6.1 (Green, GR, 520-549 nm; Yellow-green, YG, 550-559 nm; Yellow, YE, 560-584 nm), these proteins, along with the emission spectra of each species at a pH of 7.8, were then analyzed to determine the absolute pairwise distance between each species' emission spectra.

To determine the total amino acid differences, the package stringdist was used to create a pairwise distance matrix at each amino acid site using the Hamming Distance Metric: "count the number of character substitutions that turns b into a, if a and b have different number of characters the distance is Inf" (Loo, 2014). For those curious towards the analysis protocol used, see 6.4.2.

Hide

```
  align.vector <- as.vector(proteins$aligned)
#Determine the total length of the aligned proteins vector.
  sites <- nchar(align.vector[1])



#Matrix of all amino acid sites
  sum.all.matrix <- as.matrix(stringdistmatrix(align.vector, method = "hamming"))

#Names.
colnames(sum.all.matrix) <- proteins$Name
rownames(sum.all.matrix) <- proteins$Name

#Matrix to df.
sum.all.df <- melt(sum.all.matrix, varnames = c("Species1", "Species2"))
#Name Column.
colnames(sum.all.df)[3] <- "AllDiff"

#Merge into the master dataframe.
combined.df <- merge(combined.df, sum.all.df)
```

A similar method was employed for both calculating the number of differences excluding the amino acid active sites and a separate 3-D matrix for only the active sites. The active sites for *Pt. Pyralis* have previously been identified by Leach (2008, Table 6.2). During alignment, ten sites were added to the *Pt. Pyralis* protein (Supplementary Figure 6.2), so these active sites were mutated by ten for analysis.

Hide

```
#Active site Table
  activesites.table <- read_excel("Supplementary Materials.xlsx", sheet = "ST2", )
#Table to vector including adjusting for alignment (+10).
  activesites <- activesites.table$Site + 10
```

To analyze the number of amino acid differences excluding the active sites, a modified version of stringdistmatrix was used to create a vector of all 522 amino acids that are not known active sites. For each index of the vector, a pairwise distance was then calculated, giving a matrix of 0s and 1s, where 0 indicates no difference and 1 equals a difference at the specific site. As each site was calculated, it was mutated to a three dimensional matrix along the z-axis, giving a 15x15x521 matrix. This matrix was then summed down the z axis to give the total number of differences between each species luciferase protein (See 6.4.2.2 for a detailed breakdown of matrix dimensionality with visualization).

Hide

```
#Get a vector of the total number of amino acid sites.
  sites <- c(1:nchar(align.vector[1]))

# Create the first pairwise matrix.
  all.Distmatrix1 <- as.matrix(stringdistmatrix(str_sub(align.vector, 1, 1), method = "hamming"))
  #Order of Operations:
  #str_sub: Create a vector of the protein at the first active site.
  #stringdistmatrix: Does a simple comparison matrix between each group (0 = identical, 1 = differ
ence) using Hamming methodology (counts the number of character substitutions that turns b into a.
If a and b have different number of characters the distance is Inf [If a string has a length great
er than 1 it causes an error, making this self checking]).
#Remove both the first site that was analyzed and all of the active sites.
  sites <- sites[-c(1, activesites)]

#Create the array using the first matrix.
  noactive.matrix <- array(data=c(all.Distmatrix1), dim = c(length(align.vector), length(align.vec
tor),1))

#Loop all the sites selected.
  for (site in sites) {
    #Same logic as the first matrix.
      single <- as.matrix(stringdistmatrix(str_sub(align.vector, site, site), method = "hamming"))
     #Append the new matrix along the z axis.
      noactive.matrix <- abind(single, noactive.matrix, along = 3)
}

#Add up the total number of differences between each protein by summing down the z axis (dims =
2).
  sum.noactive.matrix <- as.matrix(rowSums(noactive.matrix, dims = 2))

#Names.
  colnames(sum.noactive.matrix) <- proteins$Name
  rownames(sum.noactive.matrix) <- proteins$Name

#matrix to df.
  sum.noactive.df <- melt(sum.noactive.matrix, varnames = c("Species1", "Species2"))
#Name Column.
  colnames(sum.noactive.df)[3] <- "NoActiveDiff"

#Merge to master.
  combined.df <- merge(combined.df, sum.noactive.df)
```

A similar pairwise distance metric was then used but instead of analyzing the amino acids *without* the active sites, this time the 3D matrix was created only at the active sites, creating a 15x15x43 pairwise matrix that was then summed along the z axis. This matrix was then mutated into a dataframe for analysis (See 6.5 for visualization).

Hide

```
#Create the first matrix
  activesites.Distmatrix1 <- as.matrix(stringdistmatrix(str_sub(align.vector, activesites[[1]], ac
tivesites[[1]]), method = "hamming"))
#Order of Operations:
  #str_sub: Create a vector of the protein at the first active site.
  #stringdistmatrix: Does a simple comparison matrix between each group (0 = identical, 1 = differ
ence) using Hamming methodology (counts the number of character substitutions that turns b into a.
If a and b have different number of characters the distance is Inf [If a string has a length great
er than 1 it causes an error, making this self checking]).

#Remove the analyzed site from the vector to prevent repeat analysis.
  activesites <- activesites[-1]

#Create the array using the first matrix as our input data.
  activesites.matrix <- array(data=c(activesites.Distmatrix1), dim = c(length(align.vector), lengt
h(align.vector),1))


#Looping all the matrix sites.
for (site in activesites) {
#Same as the first matrix.
  single <- as.matrix(stringdistmatrix(str_sub(align.vector, site, site), method = "hamming"))
#Append the new matrix along the z axis to the 3D array.
  activesites.matrix <- abind(single, activesites.matrix, along = 3)
}

#Distance matrix summed.
  activesites.dist.all.matrix <- rowSums(activesites.matrix, dims = 2)
  #Add up the total number of differences between each protein by summing down the z axis (dims =
2).

#Distance output
  activesites.dis.all.dist <- as.dist(rowSums(activesites.matrix, dims = 2))
  #This is for user reference, it is not analyzed by the script.

#Name the Matrix.
  colnames(activesites.dist.all.matrix) <- proteins$Name
  rownames(activesites.dist.all.matrix) <- proteins$Name

#matrix -> dataframe.
  activesites.dist.all.df <- melt(activesites.dist.all.matrix, varnames = c("Species1", "Species
2"))
#Name the new column
  colnames(activesites.dist.all.df)[3] <- "ActiveSitesOnly"

#Merge this dataframe with the master dataframe.
  combined.df <- merge(combined.df, activesites.dist.all.df)
```

The dataframe containing the species compared, the emission spectra, and the number of differences at the three groups was then cleaned to remove any self comparisons and duplicate comparisons created during the conversion from a matrix to dataframe.

Hide

```
#Logic found at:https://stackoverflow.com/questions/23474729/convert-object-of-class-dist-into-dat
a-frame-in-r
#Remove Duplicates and
  #Sort the two columns.
    names <- t(apply(combined.df[,c(1,2)],1,FUN=sort))
  #Find rows comparing the same organism.
    same <- which(names[,1] == names[,2])
  #Merge names columns into single column separated by |.
    names    <- paste(names[,1],names[,2],sep="|")
  #find duplicate comparisons.
    dups <- which(duplicated(names))
  # Remove any same organism or duplicates
  combined.df <- combined.df[-c(same, dups),]

rm(dups, names, same)

combined.df$Species1 <- gsub('_', ' ', combined.df$Species1)
combined.df$Species2 <- gsub('_', ' ', combined.df$Species2)

#Only get the wavelength emission spectra of the species
combined.df$Species1.abr <- str_sub(combined.df$Species1, 1,2)
combined.df$Species2.abr <- str_sub(combined.df$Species2, 1,2)


#Abr col comparison
combined.df$Comparison.abr <- str_c(combined.df$Species1.abr, "/", combined.df$Species2.abr)
#Full Name Comparison
combined.df$Comparison.full <- str_c(combined.df$Species1, "/", combined.df$Species2)


#organize the comparisons from green to yellow.
combined.df$Comparison.abr <- factor(combined.df$Comparison.abr, levels = c("GR/GR", "GR/YE", "GR/
YG", "YG/YG", "YE/YG", "YE/YE"))



colors <- c("#074000", "#aeff17", "#66db00", "#699647", "#26b530", "blue")
```

Finally, these pairwise comparisons were graphed using a combination of ggplot2 (Wickham et al., 2023), gplots (Warnes et al., 2022), and plotly (Sievert et al., 2022) to create both heatmaps of the summed differences and scatterplots comparing the absolute change in emission spectra vs. the difference in amino acid residues. For a visualized workflow, see 6.1.

3 Results

3.1 Phylogenetics

After extracting the third codon only and running IQ-tree, a maximum likelihood tree was built using GTR+F+G4 as the selected substitution model. As shown in Figure 3.1, no monophyletic grouping was established between different color categories. Furthermore, the two *Li. cruciata* group together into a unique clade, despite having different emission spectra. While this tree does have low bootstrap support values, it has similar branching to that of Oba et al. (2020), providing support to the evolutionary history of this tree.

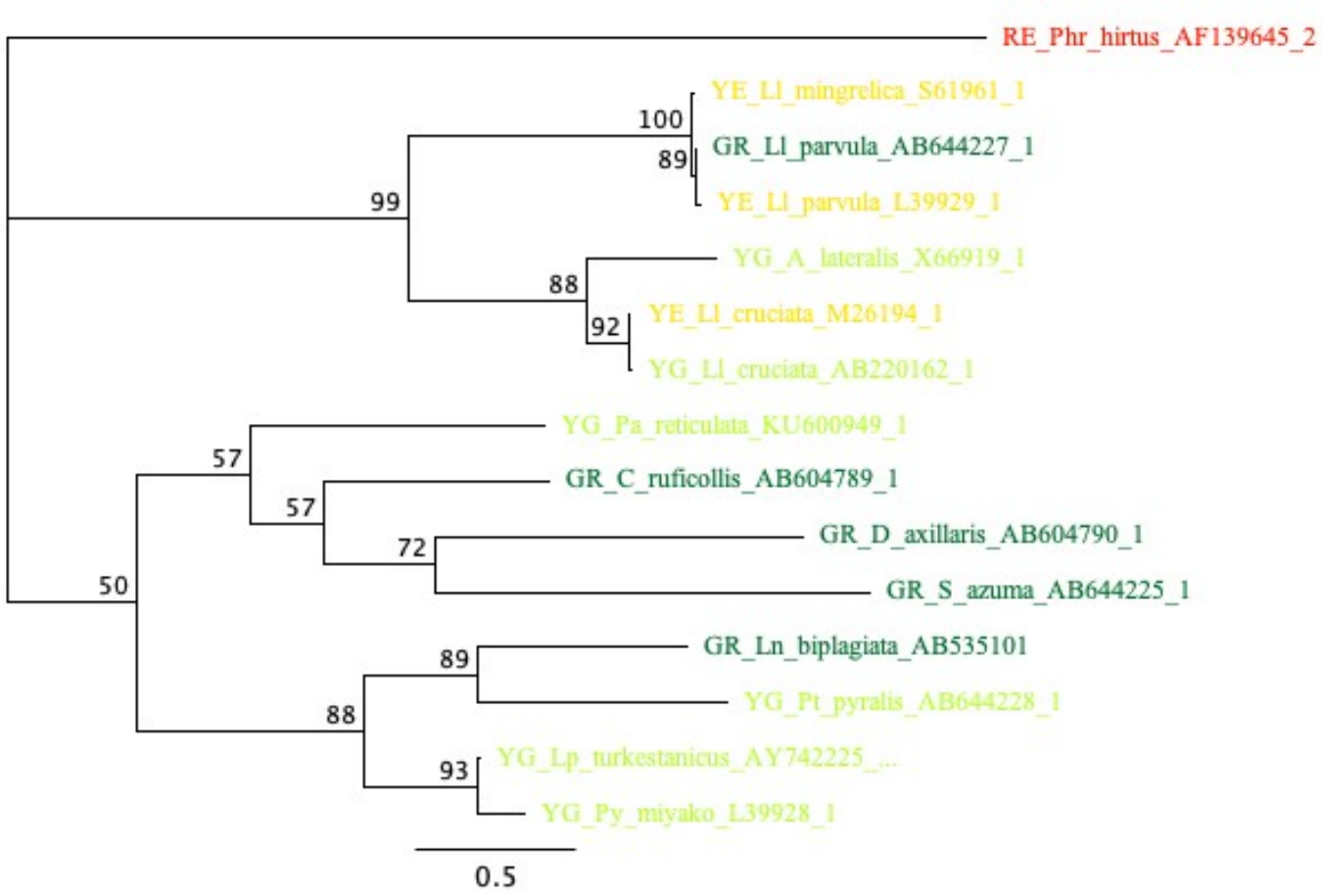


Figure 3.1: Phylogenetic tree of the analyzed species (Bootstraps = 10,000, metric = AIC), Color-coding manually mutated.

3.2 Pairwise Distance Analysis

The three different arrays were summed along the z axis into three 2-dimensional matrices, which were then plotted as a heatmap and against the absolute difference in emission spectra:

Code

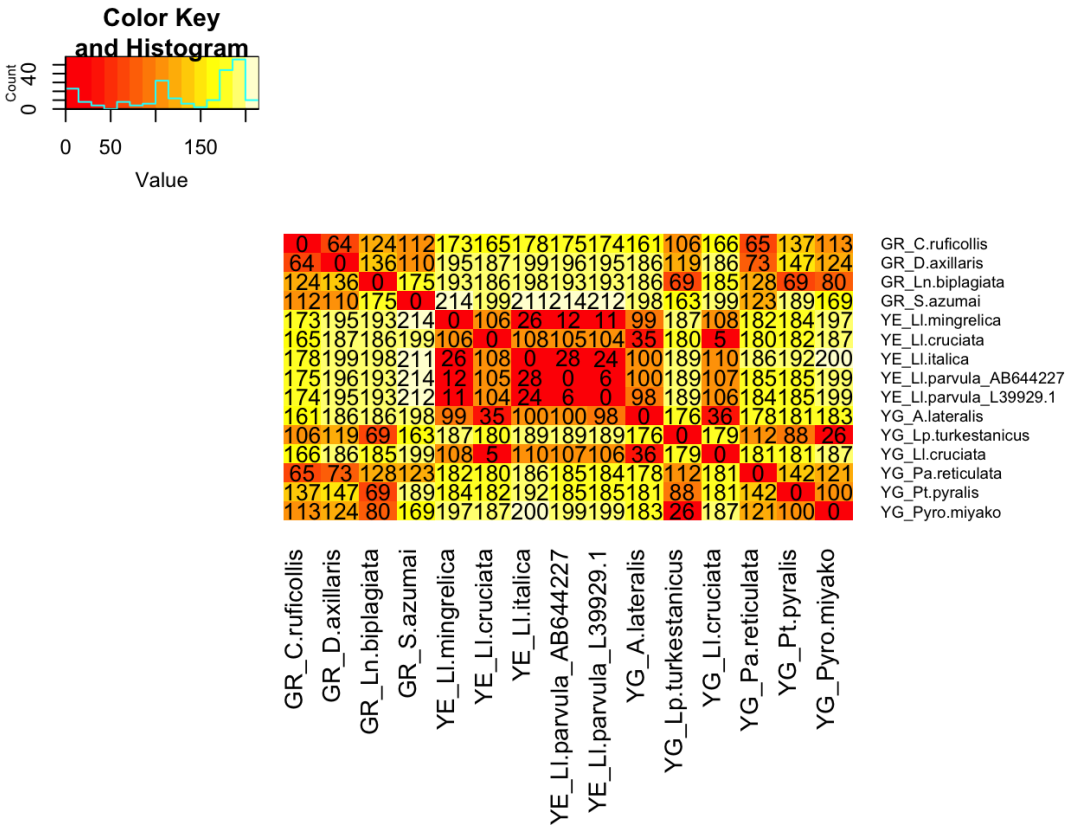


Figure 3.2: Heatmap displaying the total number of amino acid differences between each species.

Code

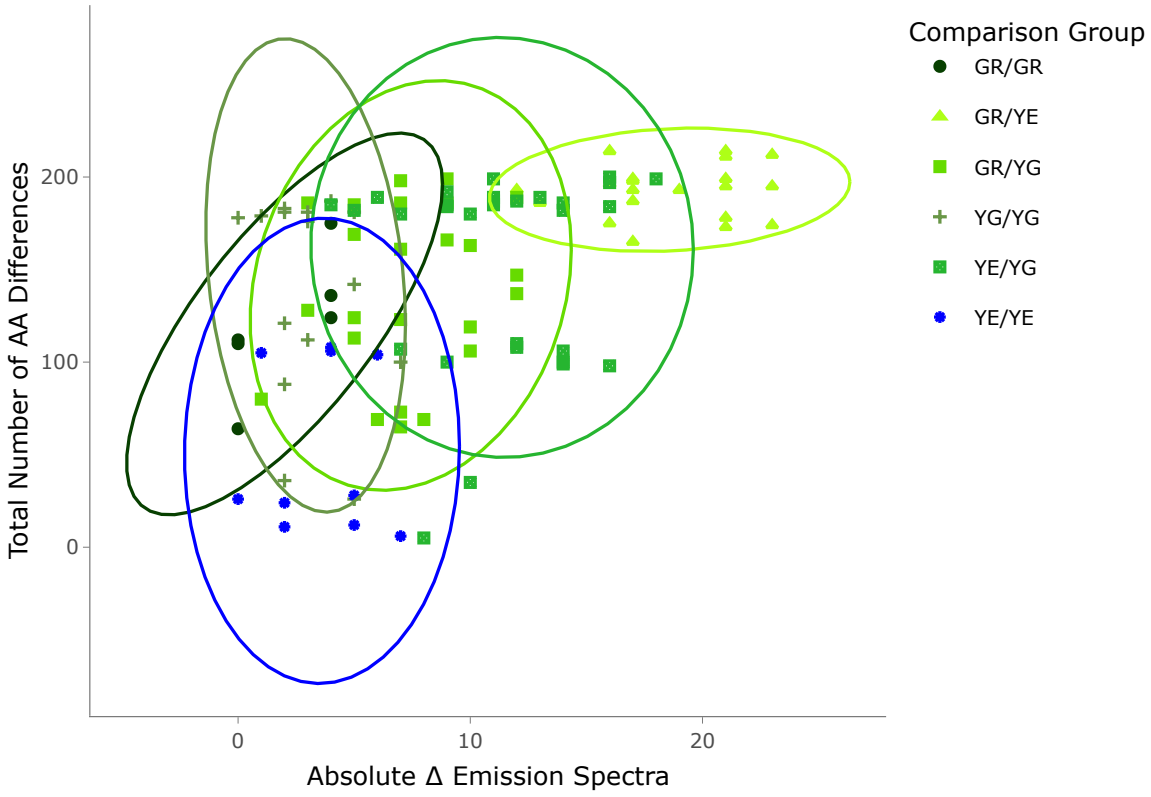


Figure 3.3: Interactive scatter-plot of the absolute change in emission spectra vs. the total number of amino acid differences between each species.

Code

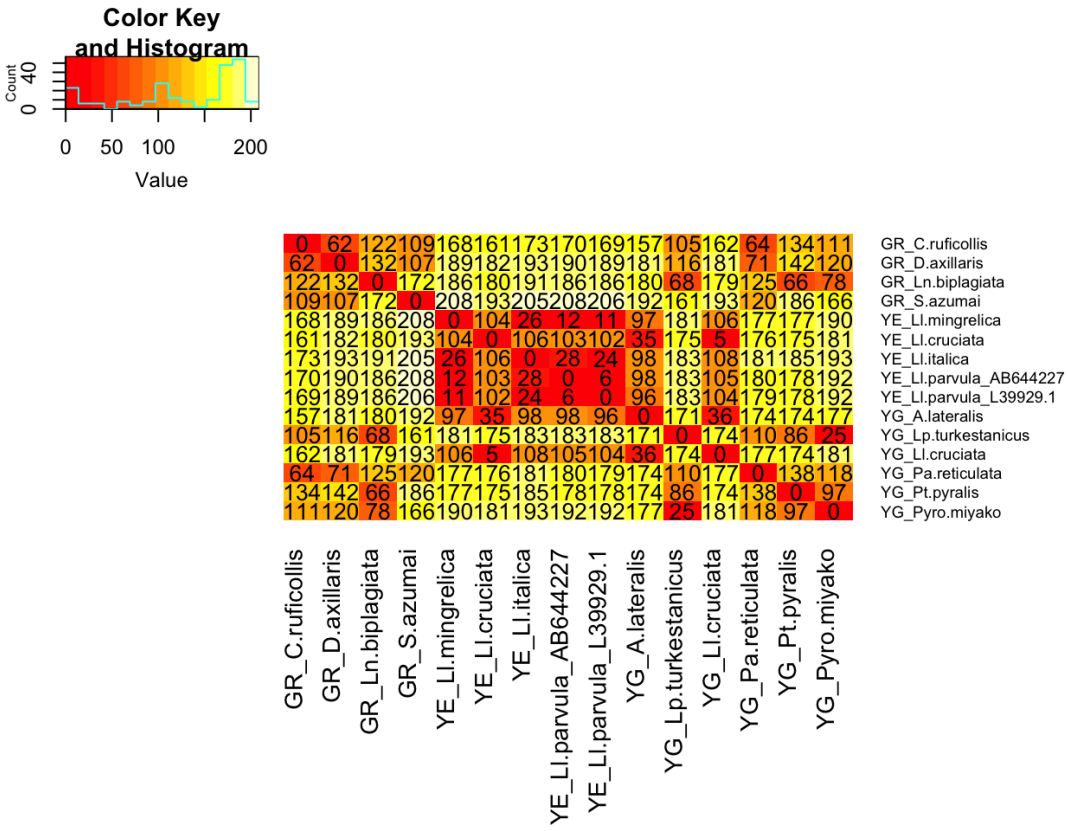


Figure 3.4: Heatmap displaying the total number of amino acid differences, excluding active sites, between each species.

Code

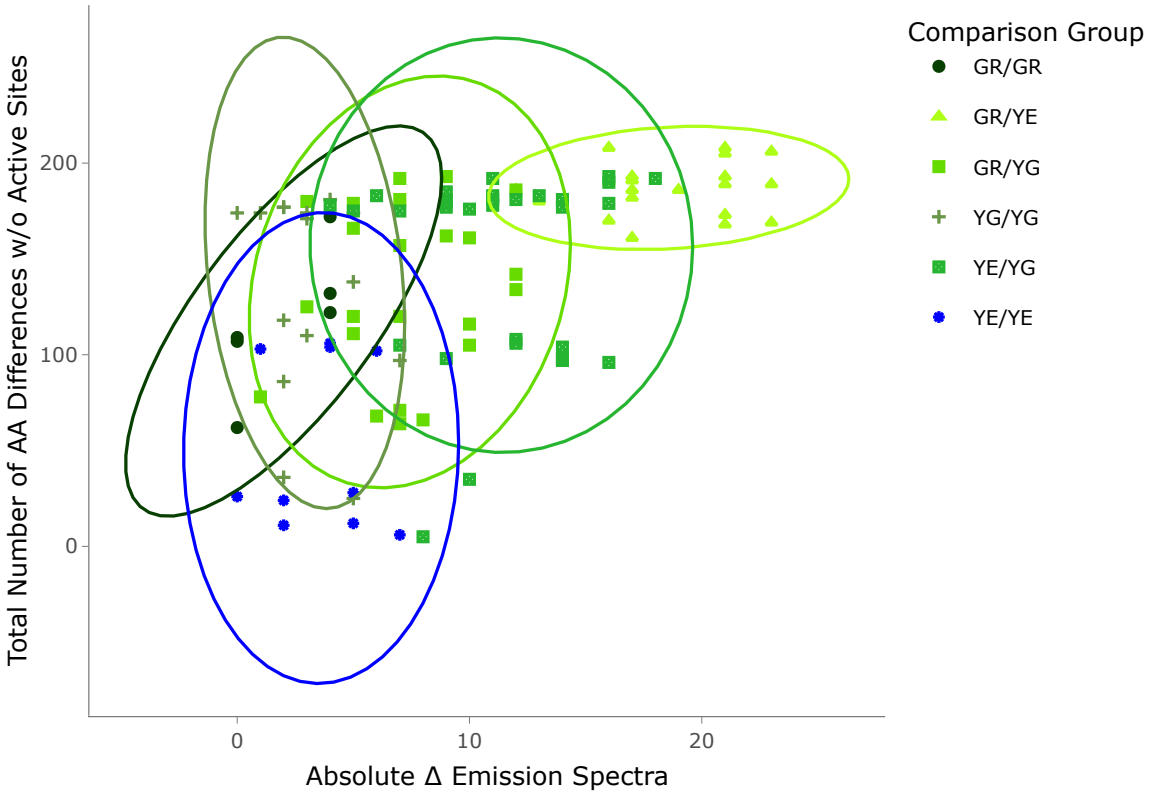


Figure 3.5: Interactive scatter-plot of the absolute change in emission spectra vs. the total number of amino acid differences, excluding active sites, between each species.

Code

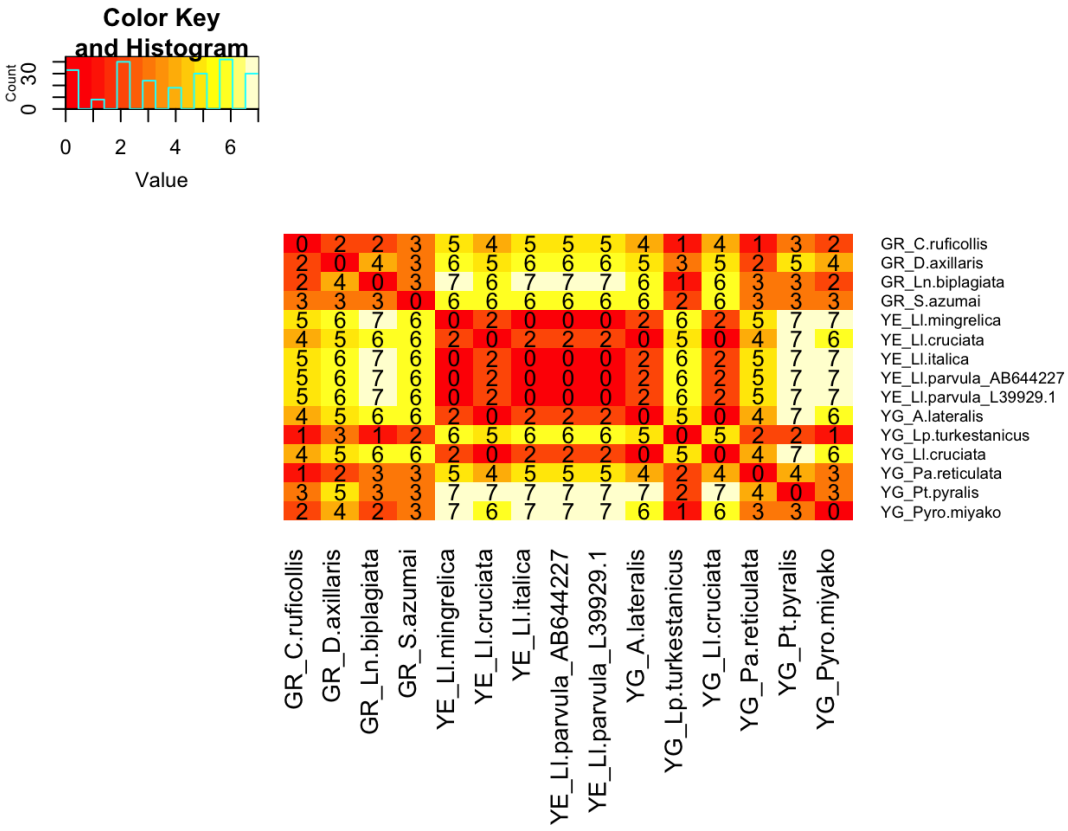


Figure 3.6: Heatmap displaying the total number of active site amino acid differences between each species.

Code

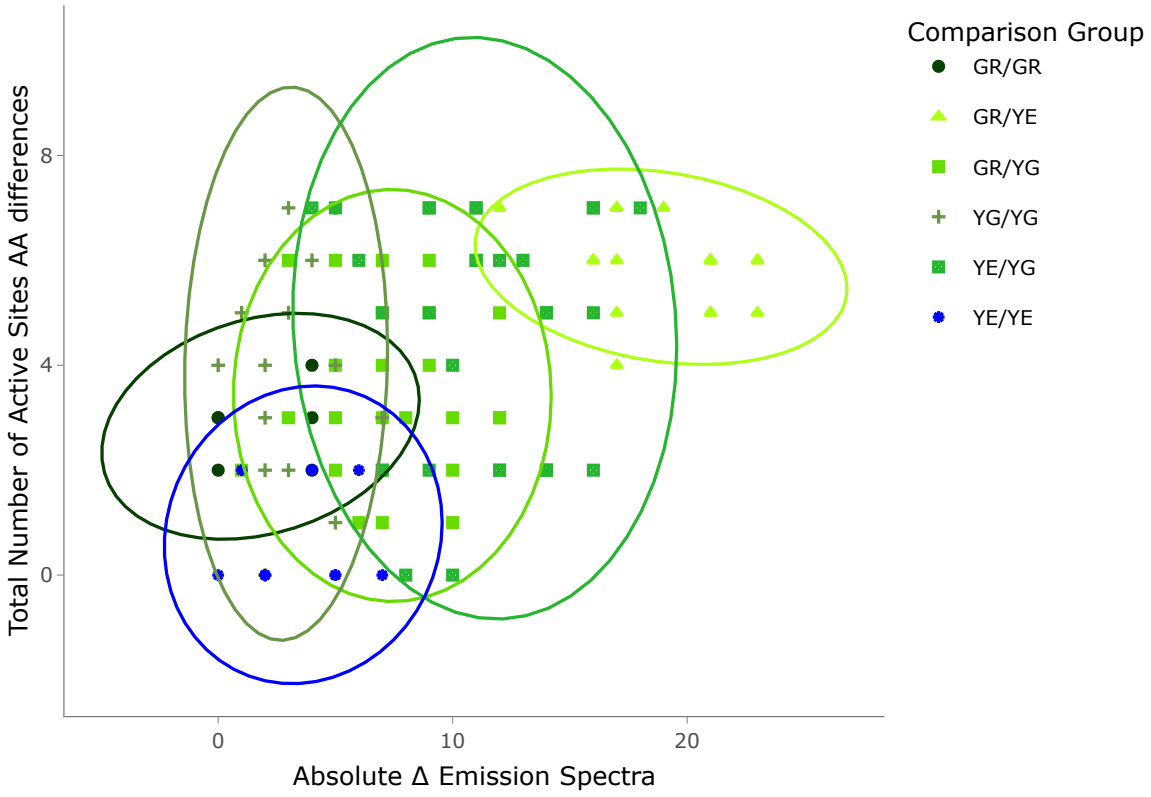


Figure 3.7: Interactive scatter-plot of the absolute change in emission spectra vs. the total number of active site amino acid differences, excluding active sites, between each species.

Across all of these figures, no correlation between the emission spectra emitted and the number of amino acid differences could be elucidated. The intracomparisons of the yellow group reveals that despite the *Ll. cruciata* having a maximum difference in emission spectra of 8nm (*Ll parvula*, 568nm), this species has over 100 total amino acid differences between

each other yellow species. This is in sharp contrast to other intra-color specie comparisions who only have between 6 and 28 total amino acid differences. Additionally, when brought to the level of active sites, the yellow species had 0 differences in amino acid residues, with the exception of *Ll. cruciata* with 2 differences. In other words, all of the amino acid differences were found in areas that are not directly interacting with the luciferin substrate.

Of particular interest are 2 *Li. cruciata* Luc1, one that emits a peak wavelength of 565nm (yellow, Accession #M26194), and the other which emits a peak wavelength of 554 (yellow-green wavelength, Accession #AB220162). While most other species have multiple amino acid differences at both all amino acids and the active sites, these two species have only 5 amino acid differences at non-active sites, and an absolute change in peak emission spectra of 7.

For the other color categories, while the intra-comparisons have a slightly lower number of differences than inter-group comparisons, there is still a high degree of differences in emission spectra. Even when accounting for just active sites, there is even less of a correlation as each group was highly overlapping with each other. Furthermore, there are a couple of comparisons where despite having no difference in peak emission wavelength, there is a difference in amino acid residues. Collectively, these graphs show that the amino acid residues are not the most significant contribution towards the color emission emitted during bioluminescence.

4 Discussion

Phylogenetically, this data shows that there is no evolutionary correlation to the wavelength spectra emitted by a firefly. Comparing this tree to that of Oba et al. (2020) (Figure 1.1), it is interesting to see the species are grouping similarly along the phylogenetic lineage, despite Oba et al. undergoing a difference series of steps and using a different substitution model (JTT matrix vs GTR+F+G4). Since the mutations undertaken removed the possibility of convergent evolution, this data suggests that evolutionarily, the amino acid residues present within a species are not a primary determinant of the color spectra emitted. If this were the case, it would be expected to see the tree start with red (the designated outgroup color) and have monophyletic clades of yellow, yellow-green, and green, identical to that of the color spectrum.

These conclusions are further backed by the heat-maps and scatter-plots showing little to no correlation between the number of amino acid differences and the absolute change in peak emission spectra. In fact, the results of the whole protein and without active site analysis suggests that despite having a high number a high number of differences in amino acid residues across the whole protein, there is a small change in the absolute emission spectra between comparison colors.

Furthermore, the intra- and inter-categorical comparisons having a large number of differences in amino acid residues despite having a small change in peak emission spectra suggests that a significant portion of the amino acid residues do not play a role in the emission spectra a species emits. However, further testing would be required to support this hypothesis.

Looking into active sites, the yellow wavelength category having 0 active site differences, with the exception of *Ll. cruciata* which has 2 differences, despite continuing to have a difference in peak emission spectra. While this subset suggests that the residues at an active site could play some role in the wavelength emitted by a species, as indicated by the differences in amino acid residues when compared to other color groups, the other categories reject this by *Ll. cruciata* (yellow), having no active site differences to *Ll. cruciata* (yellow-green) and *A. lateralis*, despite having an absolute change in emission spectra of 8 and 10 respectively.

There is even less of a correlation with each group being highly spread across the scatterplot and not clustering into distinct groups or along a linear model as predicted. These few, if any, number of differences across over 40 amino acid active sites indicates that these sites could likely be under some form of constraint to enable bioluminescence within the firefly, however further testing would be necessary to confirm such possibilities.

Alternatively, it could be that the number of changes is not what matters, but the specific amino acid present. As previously stated, single point mutagenesis studies have shown the affect a single amino acid residue can have on the emission spectra of a luciferase protein (Branchini et al., 2001; Shapiro et al., 2005; Wang et al., 2013). Additionally, it has been found that the wavelength emitted by a luciferase enzyme is dependent on pH (Viviani et al., 2018). Furthermore, Zhao et al. (2005) showed that by increasing the temperature of luciferase, the brightness and emission spectra is redshifted. Since majority of these samples were collected in the field, the temperature at the time of collection would skew the true peak emission spectra of a species luciferase. Finally, the luciferase reaction occurs internally, meaning any light emitted by a firefly must travel through the body of the firefly before being detected by either a machine or human eyes. It is highly plausible that the depth at which the luciferase reaction occurs and the composition of the cuticle plays a significant role in the emission spectra seen by

fireflies. Despite luciferase being prominently used in biology and the medical industry for bioluminescent imaging, no research was found on the effect of either parameter on the emission spectra of luciferase, leaving this hypothesis unconfirmed.

5 Conclusions

This study was unable to support the hypothesis that the number of amino acid differences (whether that be across the entire protein or at a protein's active sites) does conclusively determine the wavelength spectra emitted by a firefly. While these differences might play some role, as supported by mutagenesis studies, it is much likely to be a combination of variables that alter the emission spectra emitted. This means that while the differences could play some role, the current datasets and analysis do not allow for conclusive testing towards the exact relationship different amino acids at various positions play in firefly bioluminescence. Furthermore, in order to elucidate the true correlation between all factors (pH, temperature, filtration, etc.), even more extensive research would need to be done to allow for both detection within the light organ itself, and later isolation of the luciferase compound for further study. Although none of these are likely to be seen anytime soon, this study does enable future researchers to begin to understand the influence of amino acid residues and act as a springboard to higher analysis and investigation.

6 Supplementary Files

6.1 Data Availability

All files used in this project is available within respective folders in the project directory.

For users wishing to replicate any code chunks, all required libraries can be installed running the included Package.Installer.R file. Otherwise, downloading and editing this .Rmd file has the necessary commands built-in to install any missing packages. However, it is HIGHLY advised that only those with a solid understanding of R attempt to alter any functions or code chunks.

6.2 Tables

6.2.1 Supplemental Table 1

Table 6.1: Organisms Studied and supplementary data about each. **Green, GR, 520-549 nm; Yellow-green, YG, 550-559 nm; Yellow, YE, 560-584 nm. (Modified from: Oba et al. (2020))

Family	Subfamily	Species	Origin	GenBank No.	Gene name	Sex if known	lmax nm	Colouration*	Reference
Lampyridae	Ototretinae	*Drilaster axillaris*	Japan	AB604790	DaLuc1 (Luc1 luciferase)	Unknown	545	GR	(Oba et al., 2012)
Lampyridae	Ototretinae	*Stenocladius azumai*	Japan	AB644225	SaLuc1 (Luc1 luciferase)	Unknown	545	GR	(Oba et al., 2012)
Lampyridae	Cyphonocerinae	*Cyphonocerus ruficollis*	Japan	AB604789	CrLuc1 (Luc1 luciferase)	Unknown	546	GR	(Oba et al., 2012)
Lampyridae	Lampyrinae	*Lucidina biplagiata*	Japan	AB535101	LbLuc1 (Luc1 luciferase)	Unknown	549	GR	(Oba et al., 2012)
Lampyridae	Lampyrinae	*Pyrocoelia miyako*	Japan	L39928	NA	Unknown	550	YG	(Ohmiya et al., 1995)

Family	Subfamily	Species	Origin	GenBank No.	Gene name	Sex if known	lmax nm	Colouration*	Reference
Lampyridae	Luciolinae	*Aquatica lateralis*	Japan	X66919	LlLuc1 (Luc1 luciferase)	Unknown	552	YG	(Tatsumi et al., 1989)
Lampyridae	Lamprohizinae	*Phausis reticulata*	USA	KU600949	NA	Male (in vivo)	552	YG	(Branchini et al., 2017)
Lampyridae	Luciolinae	*Luciola cruciata*	Japan	AB220162	LcLuc1 (Luc1 luciferase)	Male (in vivo)	554	YG	(Oba et al., 2010)
Lampyridae	Lampyrinae	*Lampyris turkestanicus*	Middle East	AY742225	NA	Unknown (Combination of both male and female cDNA)	555	YG	(Tafreshi et al., 2008)
Lampyridae	Lampyrinae	*Photinus pyralis*	USA	M15077	PpyLuc1 (Luc1 luciferase)	In vitro	557	YG	(Branchini et al., 2007)
Lampyridae	Lampyrinae	*Photinus pyralis*	USA	AB644228	PpyLuc1 (Luc1 luciferase)	In vitro	557	YG	(Branchini et al., 2007)
Lampyridae	Luciolinae	*Luciola parvula*	Japan	AB644227	LpLuc1 (Luc1 luciferase)	Unknown	561	YE	(Oba et al., 2012)
Lampyridae	Luciolinae	*Luciola cruciata*	Japan	M26194	NA	Unknown	562	YE	(Kajiyama & Nakano, 1991)
Lampyridae	Luciolinae	*Luciola italica*	Italy	DQ138966	NA	Unknown	566	YE	(Branchini et al., 2006)
Lampyridae	Luciolinae	*Luciola mingrelica*	Eastern Europe	S61961	NA	Unknown	566	YE	(Koksharov & Ugarova, 2008)
Lampyridae	Luciolinae	*Luciola parvula*	Japan	L39929	NA	Male (in vitro)	568	YE	(Ohmiya et al., 1995)
Phengodidae	—	*Phrixothrix hirtus*	USA	AF139645	PhRE	Unknown	622 (pH 8.0)	RE	(*Cloning, Sequence Analysis, and Expression of Active Phrixothrix Railroad-Worms Luciferases*, n.d.)

6.2.2 Supplemental Table 2

Table 6.2: Active Sites used in this study (Modified from: Leach (2008))

Site	AA	Reference
197	N	(Branchini et al., 1998, 2000)
198	S	(Branchini et al., 1999b; Conti et al., 1996; Sandalova & Ugarova, 1999)
199	S	(Branchini et al., 1999c, 1999b), web
206	K	(Conti et al., 1996), web
218	R	(Branchini et al., 2003, 1998, 2000, 2004; Branchini, Magyar, Murtiashaw, et al., 1997; Sandalova & Ugarova, 1999; Ugarova & Sandalova, 1998; Viviani et al., 2002), web
244	H	(Branchini et al., 1999a; Branchini, Magyar, Murtiashaw, et al., 1997; Branchini, Magyar, Marcantonio, et al., 1997)
245	H	(Branchini, Magyar, Marcantonio, et al., 1997; Branchini et al., 1998, 1999b, 2000, 2004; Sandalova & Ugarova, 1999). web
246	G	(Branchini, Magyar, Murtiashaw, et al., 1997; Branchini et al., 2004)
247	F	(Branchini, Magyar, Murtiashaw, et al., 1997; Branchini, Magyar, Marcantonio, et al., 1997; Branchini et al., 1998, 1999b, 2000, 2004; Franks et al., 1998; Sandalova & Ugarova, 1999, 1999; Viviani et al., 2002), web
250	F	(Branchini, Magyar, Murtiashaw, et al., 1997; Branchini et al., 1998, 1999b, 2004)
251	T	(Branchini, Magyar, Murtiashaw, et al., 1997; Branchini et al., 2004)
310	H	(Franks et al., 1998)
311	E	(Franks et al., 1998)
313	A	(Franks et al., 1998)
314	S	(Ugarova & Sandalova, 1998)
315	G	(Branchini, Magyar, Murtiashaw, et al., 1997; Branchini et al., 2004; Franks et al., 1998; Sandalova & Ugarova, 1999)
316	G	(Branchini et al., 2004; Sandalova & Ugarova, 1999)
317	A	(Branchini et al., 1998; Sandalova & Ugarova, 1999), web
318	P	(Sandalova & Ugarova, 1999)
337	R	(Sandalova & Ugarova, 1999)
338	Q	(Sandalova & Ugarova, 1999)
339	G	(Branchini et al., 1998, 2000; Sandalova & Ugarova, 1999, 1999), web
340	Y	(Branchini et al., 1999a, 1998; Conti et al., 1996; Sandalova & Ugarova, 1999), web
341	G	(Branchini, Magyar, Murtiashaw, et al., 1997; Branchini et al., 2004; Sandalova & Ugarova, 1999),web
342	L	(Branchini, Magyar, Murtiashaw, et al., 1997; Branchini et al., 2004)
343	T	(Branchini et al., 1999a, 1998, 2004; Branchini, Magyar, Murtiashaw, et al., 1997; Sandalova & Ugarova, 1999), web
344	E	(Conti et al., 1996; Sandalova & Ugarova, 1999)
347	S	(Branchini et al., 1998, 2000, 2004; Sandalova & Ugarova, 1999, 1999), web
348	A	(Branchini, Magyar, Murtiashaw, et al., 1997; Branchini et al., 1998, 2000, 2004; Sandalova & Ugarova, 1999)
351	I	(Branchini, Magyar, Murtiashaw, et al., 1997; Branchini et al., 2004; Sandalova & Ugarova, 1999)
352	T	(Franks et al., 1998)
353	P	(Franks et al., 1998)

Site	AA	Reference
354	E	(Franks et al., 1998)
389	E	(Conti et al., 1996)
401	Y	(Conti et al., 1996)
417	W	(Dementieva et al., 2000)
420	S	(Conti et al., 1996)
421	G	(Conti et al., 1996)
422	D	(Branchini et al., 2004; Conti et al., 1996; Sandalova & Ugarova, 1999), web
434	I	(Sandalova & Ugarova, 1999)
437	R	(Leach, 2008)
527	T	(Sandalova & Ugarova, 1999)
529	K	(Branchini, Magyar, Murtiashaw, et al., 1997; Branchini et al., 2000, 2004; Sandalova & Ugarova, 1999, 1999), web

6.3 Figures

6.3.1 Supplementary Figure 1

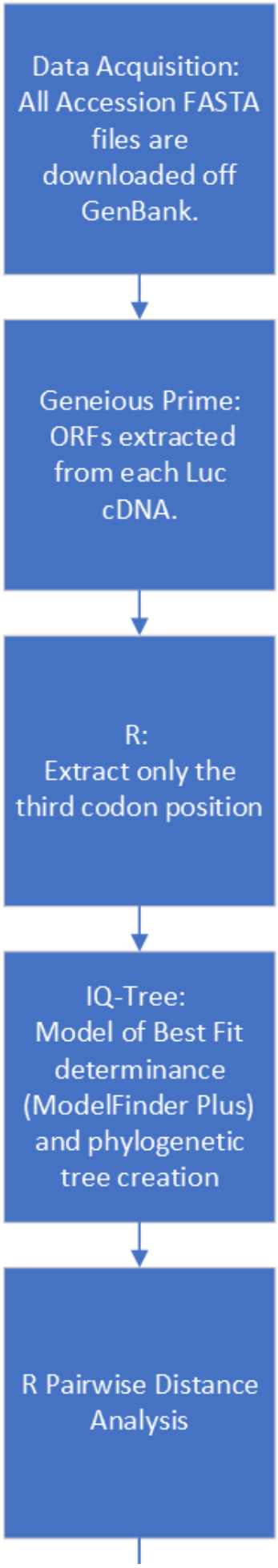
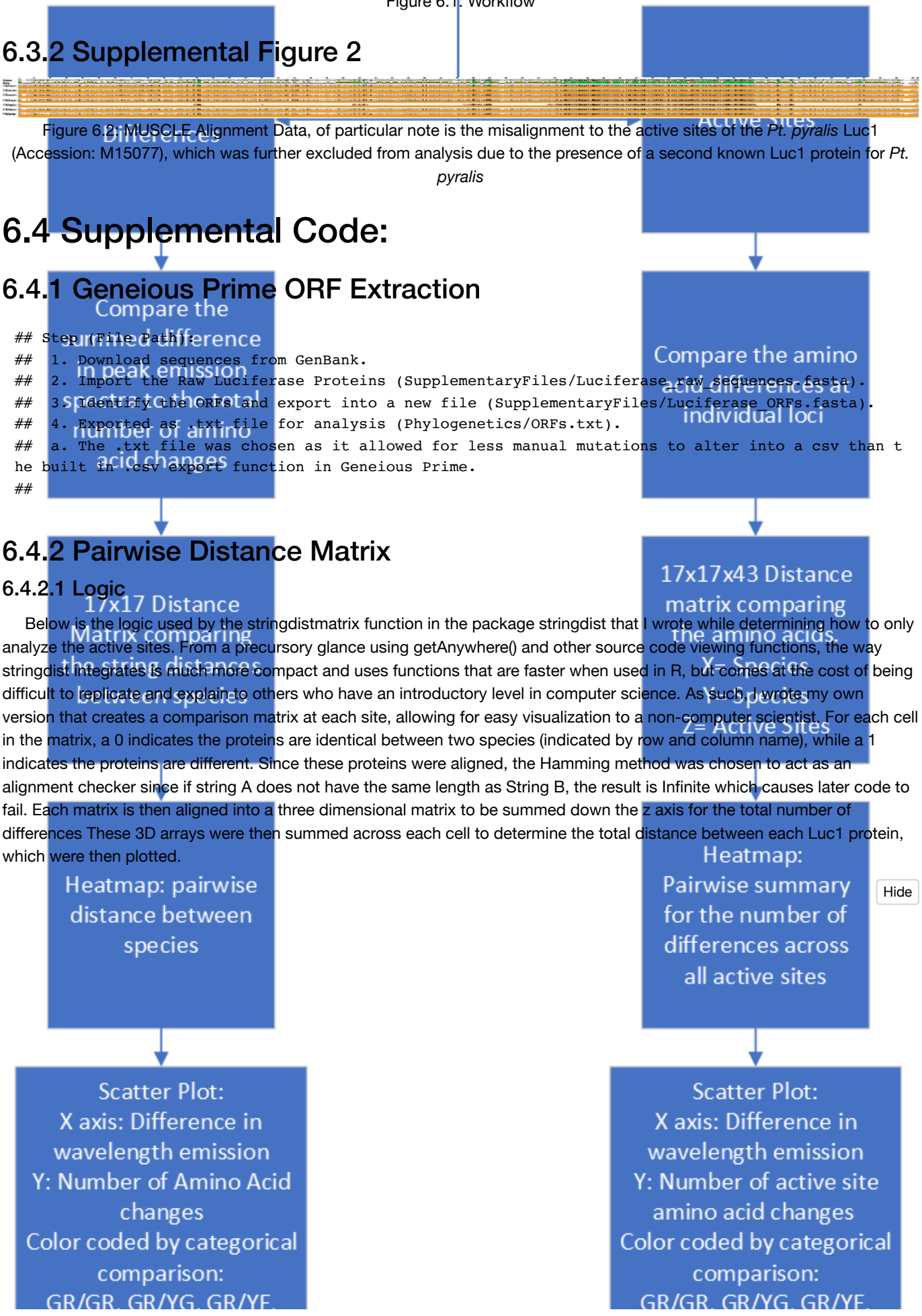

Data Acquisition: All Accession FASTA files are downloaded off GenBank.
Geneious Prime: ORFs extracted from each Luc cDNA.
R: Extract only the third codon position
IQ-Tree: Model of Best Fit determinance (ModelFinder Plus) and phylogenetic tree creation
R Pairwise Distance Analysis

Figure 6.1: Workflow



6.3.2 Supplemental Figure 2

Figure 6.2: MUSCLE Alignment Data, of particular note is the misalignment to the active sites of the *Pt. pyralis* Luc1 (Accession: M15077), which was further excluded from analysis due to the presence of a second known Luc1 protein for *Pt. pyralis*

6.4 Supplemental Code:

6.4.1 Geneious Prime ORF Extraction

```
## Step (File Path):
##  1. Download sequences from GenBank.
##  2. Import the Raw Luciferase Proteins (SupplementaryFiles/Luciferase_raw_sequences.fasta).
##  3. Identify the ORFs and export into a new file (SupplementaryFiles/Luciferase_ORFs.fasta).
##  4. Exported as .txt file for analysis (Phylogenetics/ORFs.txt).
##  a. The .txt file was chosen as it allowed for less manual mutations to alter into a csv than t
he built in .csv export function in Geneious Prime.
##
```

6.4.2 Pairwise Distance Matrix

6.4.2.1 Logic

Below is the logic used by the stringdistmatrix function in the package stringdist that I wrote while determining how to only analyze the active sites. From a precursory glance using getAnywhere() and other source code viewing functions, the way stringdist integrates is much more compact and uses functions that are faster when used in R, but comes at the cost of being difficult to replicate and explain to others who have an introductory level in computer science. As such, I wrote my own version that creates a comparison matrix at each site, allowing for easy visualization to a non-computer scientist. For each cell in the matrix, a 0 indicates the proteins are identical between two species (indicated by row and column name), while a 1 indicates the proteins are different. Since these proteins were aligned, the Hamming method was chosen to act as an alignment checker since if string A does not have the same length as String B, the result is Infinite which causes later code to fail. Each matrix is then aligned into a three dimensional matrix to be summed down the z axis for the total number of differences These 3D arrays were then summed across each cell to determine the total distance between each Luc1 protein, which were then plotted.

Hide

```
#Prevents redundancy of vector creation.
  align.vector <- as.vector(proteins$aligned)
#Determine the total length of the aligned proteins vector.
  sites <- nchar(align.vector[1])

#Create first matrix.
  all.Distmatrix1 <- as.matrix(stringdistmatrix(str_sub(align.vector, 1, 1), method = "hamming"))
    #Order of Operations:
      #str_sub: Create a vector of the protein at only the first site.
      #stringdistmatrix: Does a simple comparison matrix between each group (0 = identical, 1 = di
fference) using Hamming methodology [If a string has a length greater than 1 it causes an error, m
aking this self checking].

#Create the array using the first matrix
  all.matrix <- array(data=c(all.Distmatrix1), dim = c(length(align.vector), length(align.vector),
1))


#Repeat the matrix for all other sites within the protein
for (site in 2:sites) {
  single <- as.matrix(stringdistmatrix(str_sub(align.vector, site, site), method = "hamming"))#Sam
e as first matrix
  all.matrix <- abind(single, all.matrix, along = 3) #Append the new matrix along the z axis.
}

#Summed Matrix.
  sum.all.matrix <- as.matrix(rowSums(all.matrix, dims = 2))
  #Add up the total number of differences between each protein by summing down the z axis (dims =
2).

#Names.
colnames(sum.all.matrix) <- proteins$Name
rownames(sum.all.matrix) <- proteins$Name

#matrix to df
sum.all.df <- melt(sum.all.matrix, varnames = c("Species1", "Species2"))
colnames(sum.all.df)[3] <- "AllDiff"

#Merge into the master dataframe.
combined.df <- merge(combined.df, sum.all.df)
```

6.4.2.2 Visualization

Similar to a chess board, a pairwise matrix is defined into cells that can be indexed based on the row and column number/letter. For instance, in chess the index [D4] corresponds to the cell at the fourth column (D), fourth row of the chess board (4) (Figure 6.3). Similarly, the index of a matrix corresponds to a specific point based upon the row and then the column (meaning [2,1] is the second row, first column).

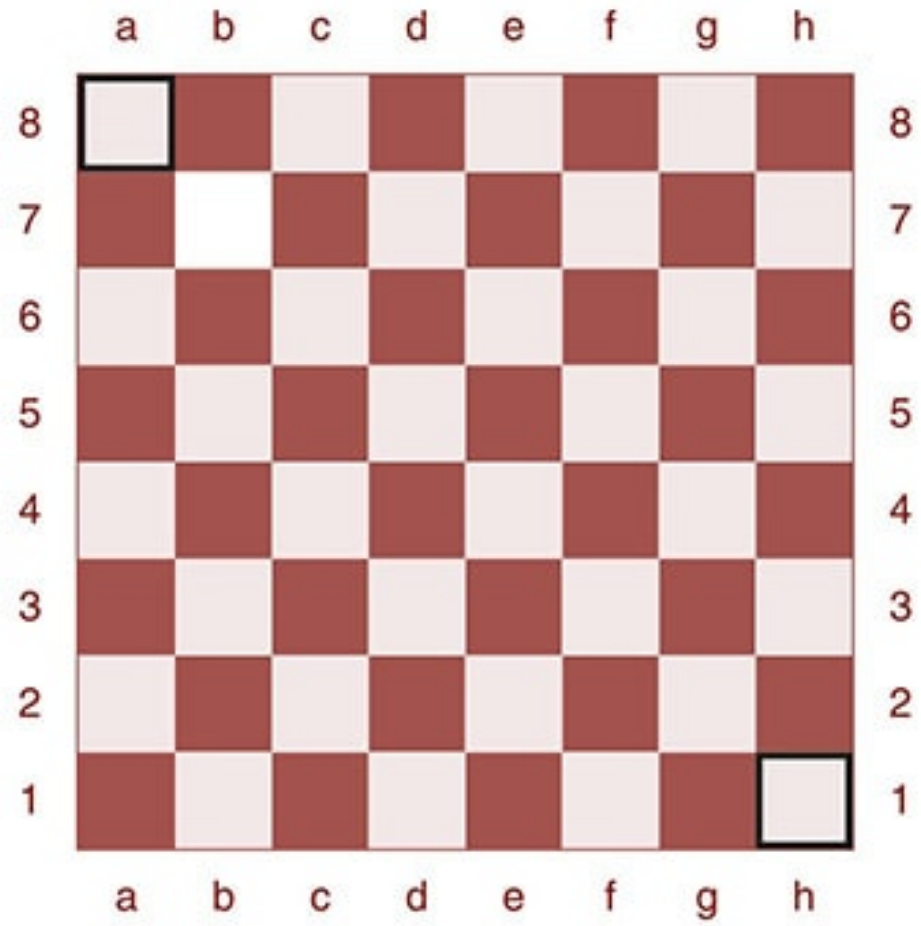


Figure 6.3: An example chessboard and matrix (*How to Set up a Chessboard - A Quick & Simple Guide*, n.d.).

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{array}$$

Figure 6.4: An example chessboard and matrix (Lancashire3000, 2022).

In a pairwise distance matrix, the row and column represent a specific species, and the cell represents the result of a comparison between the two. In this study, that comparison is two things: the emission spectra, and the amino acid residue at specific sites. For the emission spectra, this is taken using the absolute difference between the species.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
1	0	0	4	0	21	21	17	16	23	7	10	9	7	12	5
2	0	0	4	0	21	21	17	16	23	7	10	9	7	12	5
3	4	4	0	4	17	17	13	12	19	3	6	5	3	8	1
4	0	0	4	0	21	21	17	16	23	7	10	9	7	12	5
5	21	21	17	21	0	0	4	5	2	14	11	12	14	9	16
6	21	21	17	21	0	0	4	5	2	14	11	12	14	9	16
7	17	17	13	17	4	4	0	1	6	10	7	8	10	5	12
8	16	16	12	16	5	5	1	0	7	9	6	7	9	4	11
9	23	23	19	23	2	2	6	7	0	16	13	14	16	11	18
10	7	7	3	7	14	14	10	9	16	0	3	2	0	5	2
11	10	10	6	10	11	11	7	6	13	3	0	1	3	2	5
12	9	9	5	9	12	12	8	7	14	2	1	0	2	3	4
13	7	7	3	7	14	14	10	9	16	0	3	2	0	5	2
14	12	12	8	12	9	9	5	4	11	5	2	3	5	0	7
15	5	5	1	5	16	16	12	11	18	2	5	4	2	7	0

Figure 6.5: The absolute difference in emission between each species.

On the other hand, since the proteins are letters instead of numbers, a binary system must be used in which the differences are either 0s (indicating identical amino acid residues) or 1s (indicating a difference in amino acid residues). However, this only gives the differences at one point. Therefore, when doing a pairwise difference across the entire protein, a pairbased matrix must be created at each point. In order to not have a unique vector for each matrix, we can use 3-dimensional arrays in which one matrix (a chessboard) is stacked on top of another matrix (Figure 6.6). This array is then summed through the z-axis to produce a single matrix with the total number of differences between each species.

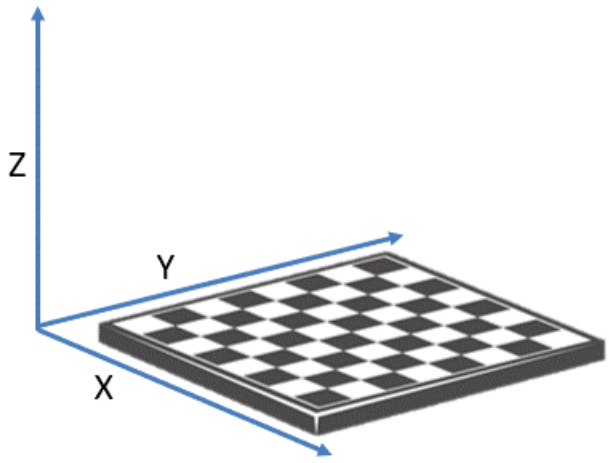


Figure 6.6: Visual demonstration of stacking matrices to create a 3-dimensional matrix.

6.5 Matrix to Dataframe

Take the distance matrix:

Code

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    0    1    2    3    4    5
## [2,]    1    0    1    2    3    4
## [3,]    2    1    0    1    2    3
## [4,]    3    2    1    0    1    2
## [5,]    4    3    2    1    0    1
## [6,]    5    4    3    2    1    0
```

Using the melt function, we get the output below, where the row represents the first column, the column recommends the second column, and the distance value represents the third column.

```
##    Row Column Value
## 1    1      1     0
## 2    2      1     1
## 3    3      1     2
## 4    4      1     3
## 5    5      1     4
## 6    6      1     5
## 7    1      2     1
## 8    2      2     0
## 9    3      2     1
## 10   4      2     2
## 11   5      2     3
## 12   6      2     4
## 13   1      3     2
## 14   2      3     1
## 15   3      3     0
## 16   4      3     1
## 17   5      3     2
## 18   6      3     3
## 19   1      4     3
## 20   2      4     2
## 21   3      4     1
## 22   4      4     0
## 23   5      4     1
## 24   6      4     2
## 25   1      5     4
## 26   2      5     3
## 27   3      5     2
## 28   4      5     1
## 29   5      5     0
## 30   6      5     1
## 31   1      6     5
## 32   2      6     4
## 33   3      6     3
## 34   4      6     2
## 35   5      6     1
## 36   6      6     0
```

However since the diagonal of the matrix is a self comparison, this would not be an apt comparison. So these values can be removed with the simple logic function:

Code

```
##    Row Column Value
## 2    2      1     1
## 3    3      1     2
## 4    4      1     3
## 5    5      1     4
## 6    6      1     5
## 7    1      2     1
## 9    3      2     1
## 10   4      2     2
## 11   5      2     3
## 12   6      2     4
## 13   1      3     2
## 14   2      3     1
## 16   4      3     1
## 17   5      3     2
## 18   6      3     3
## 19   1      4     3
## 20   2      4     2
## 21   3      4     1
## 23   5      4     1
## 24   6      4     2
## 25   1      5     4
## 26   2      5     3
## 27   3      5     2
## 28   4      5     1
## 30   6      5     1
## 31   1      6     5
## 32   2      6     4
## 33   3      6     3
## 34   4      6     2
## 35   5      6     1
```

Additionally, a comparison of #1 v. #2 is going to produce the same value as the comparison #2 v. #1 since a matrix is symmetrical down the diagonal. As such, these values need to be removed for comparison:

Code

```
##    Row Column Value
## 2    2      1     1
## 3    3      1     2
## 4    4      1     3
## 5    5      1     4
## 6    6      1     5
## 7    1      2     1
## 10   4      2     2
## 11   5      2     3
## 12   6      2     4
## 13   1      3     2
## 14   2      3     1
## 18   6      3     3
## 19   1      4     3
## 20   2      4     2
## 21   3      4     1
## 26   2      5     3
## 27   3      5     2
## 28   4      5     1
## 34   4      6     2
## 35   5      6     1
```

References

Amos, J. (2018). *Answer to "vector to matrix of differences between elements"*. https://stackoverflow.com/a/50882674 (https://stackoverflow.com/a/50882674)

Bessho-Uehara, M., & Oba, Y. (2017). Identification and characterization of the Luc2-type luciferase in the Japanese firefly, Luciola parvula, involved in a dim luminescence in immobile stages. *Luminescence: The Journal of Biological and Chemical Luminescence*, *32*(6), 924–931. https://doi.org/10.1002/bio.3273 (https://doi.org/10.1002/bio.3273)

Bofkin, L., & Goldman, N. (2007). Variation in Evolutionary Processes at Different Codon Positions. *Molecular Biology and Evolution*, *24*(2), 513–521. https://doi.org/10.1093/molbev/msl178 (https://doi.org/10.1093/molbev/msl178)

Branchini, B. R., Ablamsky, D. M., Rosenman, J. M., Uzasci, L., Southworth, T. L., & Zimmer, M. (2007). Synergistic mutations produce blue-shifted bioluminescence in firefly luciferase. *Biochemistry*, *46*(48), 13847–13855. https://doi.org/10.1021/bi7015052 (https://doi.org/10.1021/bi7015052)

Branchini, B. R., Magyar, R. A., Marcantonio, K. M., Newberry, K. J., Stroh, J. G., Hinz, L. K., & Murtiashaw, M. H. (1997). Identification of a firefly luciferase active site peptide using a benzophenone-based photooxidation reagent. *Journal of Biological Chemistry*, *272*(31), 19359–19364.

Branchini, B. R., Magyar, R. A., Murtiashaw, M. H., Anderson, S. M., Helgerson, L. C., & Zimmer, M. (1999a). Site-Directed Mutagenesis of Firefly Luciferase Active Site Amino Acids: A Proposed Model for Bioluminescence Color. *Biochemistry*, *38*(40), 13223–13230. https://doi.org/10.1021/bi991181o (https://doi.org/10.1021/bi991181o)

Branchini, B. R., Magyar, R. A., Murtiashaw, M. H., Anderson, S. M., Helgerson, L. C., & Zimmer, M. (1999b). Site-directed mutagenesis of firefly luciferase active site amino acids: A proposed model for bioluminescence color. *Biochemistry*, *38*(40), 13223–13230.

Branchini, B. R., Magyar, R. A., Murtiashaw, M. H., Anderson, S. M., Helgerson, L. C., & Zimmer, M. (1999c). Site-directed mutagenesis of firefly luciferase active site amino acids: A proposed model for bioluminescence color. *Biochemistry*, *38*(40), 13223–13230.

Branchini, B. R., Magyar, R. A., Murtiashaw, M. H., Anderson, S. M., & Zimmer, M. (1998). Site-directed mutagenesis of histidine 245 in firefly luciferase: A proposed model of the active site. *Biochemistry*, *37*(44), 15311–15319.

Branchini, B. R., Magyar, R. A., Murtiashaw, M. H., Magnasco, N., Hinz, L. K., & Stroh, J. G. (1997). Inactivation of Firefly Luciferase withN-(Iodoacetyl)-N′-(5-sulfo-1-naphthyl) ethylenediamine (I-AEDANS). *Archives of Biochemistry and Biophysics*, *340*(1), 52–58.

Branchini, B. R., Magyar, R. A., Murtiashaw, M. H., & Portier, N. C. (2001). The Role of Active Site Residue Arginine 218 in Firefly Luciferase Bioluminescence. *Biochemistry*, *40*(8), 2410–2418. https://doi.org/10.1021/bi002246m (https://doi.org/10.1021/bi002246m)

Branchini, B. R., Murtiashaw, M. H., Magyar, R. A., & Anderson, S. M. (2000). The role of lysine 529, a conserved residue of the acyl-adenylate-forming enzyme superfamily, in firefly luciferase. *Biochemistry*, *39*(18), 5433–5440.

Branchini, B. R., Southworth, T. L., DeAngelis, J. P., Roda, A., & Michelini, E. (2006). Luciferase from the Italian firefly Luciola italica: molecular cloning and expression. *Comparative Biochemistry and Physiology. Part B, Biochemistry & Molecular Biology*, *145*(2), 159–167. https://doi.org/10.1016/j.cbpb.2006.06.001 (https://doi.org/10.1016/j.cbpb.2006.06.001)

Branchini, B. R., Southworth, T. L., Murtiashaw, M. H., Boije, H., & Fleet, S. E. (2003). A Mutagenesis Study of the Putative Luciferin Binding Site Residues of Firefly Luciferase. *Biochemistry*, *42*(35), 10429–10436. https://doi.org/10.1021/bi030099x (https://doi.org/10.1021/bi030099x)

Branchini, B. R., Southworth, T. L., Murtiashaw, M. H., Magyar, R. A., Gonzalez, S. A., Ruggiero, M. C., & Stroh, J. G. (2004). An alternative mechanism of bioluminescence color determination in firefly luciferase. *Biochemistry*, *43*(23), 7255–7262.

Branchini, B. R., Southworth, T. L., Salituro, L. J., Fontaine, D. M., & Oba, Y. (2017). Cloning of the Blue Ghost (Phausis reticulata) Luciferase Reveals a Glowing Source of Green Light. *Photochemistry and Photobiology*, *93*(2), 473–478. https://doi.org/10.1111/php.12649 (https://doi.org/10.1111/php.12649)

Cloning, sequence analysis, and expression of active phrixothrix railroad-worms luciferases: Relationship between bioluminescence spectra and primary structures, | biochemistry. (n.d.). https://pubs.acs.org/doi/full/10.1021/bi9900830 (https://pubs.acs.org/doi/full/10.1021/bi9900830)

Conti, E., Franks, N. P., & Brick, P. (1996). Crystal structure of firefly luciferase throws light on a superfamily of adenylate-forming enzymes. *Structure*, *4*(3), 287–298.

Dementieva, E. I., Fedorchuk, E. A., Brovko, L. Y., Savitskii, A. P., & Ugarova, N. N. (2000). Fluorescent properties of firefly luciferases and their complexes with luciferin. *Bioscience Reports*, *20*(1), 21–30.

Fallon, T. R., Lower, S. E., Chang, C.-H., Bessho-Uehara, M., Martin, G. J., Bewick, A. J., Behringer, M., Debat, H. J., Wong, I., & Day, J. C. (2018). Firefly genomes illuminate parallel origins of bioluminescence in beetles. *Elife*, *7*, e36495.

Franks, N. P., Jenkins, A., Conti, E., Lieb, W. R., & Brick, P. (1998). Structural basis for the inhibition of firefly luciferase by a general anesthetic. *Biophysical Journal*, *75*(5), 2205–2211.

Hall, D. W., Sander, S. E., Pallansch, J. C., & Stanger-Hall, K. F. (2016). The evolution of adult light emission color in North American fireflies. *Evolution; International Journal of Organic Evolution*, *70*(9), 2033–2048. https://doi.org/10.1111/evo.13002 (https://doi.org/10.1111/evo.13002)

How to set up a Chessboard - A Quick & Simple Guide. (n.d.). https://www.regencychess.co.uk/index.php?main_page=how_to_set_up_a_chessboard.

Kajiyama, N., & Nakano, E. (1991). Isolation and characterization of mutants of firefly luciferase which produce different colors of light. *Protein Engineering*, *4*(6), 691–693. https://doi.org/10.1093/protein/4.6.691 (https://doi.org/10.1093/protein/4.6.691)

Koksharov, M. I., & Ugarova, N. N. (2008). Random mutagenesis of Luciola mingrelica firefly luciferase. Mutant enzymes with bioluminescence spectra showing low pH sensitivity. *Biochemistry. Biokhimiia*, *73*(8), 862–869. https://doi.org/10.1134/s0006297908080038 (https://doi.org/10.1134/s0006297908080038)

Kusy, D., He, J.-W., Bybee, S. M., Motyka, M., Bi, W.-X., Podsiadlowski, L., Li, X.-Y., & Bocak, L. (2021). Phylogenomic relationships of bioluminescent elateroids define the “lampyroid” clade with clicking Sinopyrophoridae as its earliest member. *Systematic Entomology*, *46*(1), 111–123. https://doi.org/10.1111/syen.12451 (https://doi.org/10.1111/syen.12451)

Lancashire3000. (2022). Mapping from tensor notation to matrix notation, left right or upper lower to row column? [Forum Post]. In *Physics Stack Exchange*.

Leach, F. R. (2008). A view on the active site of firefly luciferase. *Natural Product Communications*, *3*(9), 1934578X0800300908. https://doi.org/10.1177/1934578X0800300908 (https://doi.org/10.1177/1934578X0800300908)

Loo, M. P. J. van der. (2014). *The stringdist package for approximate string matching*. *6*, 111–122. https://CRAN.R-project.org/package=stringdist (https://CRAN.R-project.org/package=stringdist)

Martin, G. J., Branham, M. A., Whiting, M. F., & Bybee, S. M. (2017). Total evidence phylogeny and the evolution of adult bioluminescence in fireflies (Coleoptera: Lampyridae). *Molecular Phylogenetics and Evolution*, *107*, 564–575.

Minh, B. Q., Trifinopoulos, J., Schrempf, D., Schmidt, H. A., & Lanfear, R. (2019). IQTREE version 2.0: Tutorials and manual phylogenomic software by maximum likelihood. *URL Http://Www. Iqtree. Org*.

Morton, R. A., Hopkins, T. A., & Seliger, H. H. (1969). Spectroscopic properties of firefly luciferin and related compounds; an approach to product emission. *Biochemistry*, *8*(4), 1598–1607. https://doi.org/10.1021/bi00832a041 (https://doi.org/10.1021/bi00832a041)

Navizet, I., Liu, Y.-J., Ferré, N., Xiao, H.-Y., Fang, W.-H., & Lindh, R. (2010). Color-Tuning Mechanism of Firefly Investigated by Multi-Configurational Perturbation Method. *Journal of the American Chemical Society*, *132*(2), 706–712. https://doi.org/10.1021/ja908051h (https://doi.org/10.1021/ja908051h)

Oba, Y., Konishi, K., Yano, D., Shibata, H., Kato, D., & Shirai, T. (2020). Resurrecting the ancient glow of the fireflies. *Science Advances*, *6*(49), eabc5705. https://doi.org/10.1126/sciadv.abc5705 (https://doi.org/10.1126/sciadv.abc5705)

Oba, Y., Mori, N., Yoshida, M., & Inouye, S. (2010). Identification and characterization of a luciferase isotype in the japanese firefly, luciola cruciata, involving in the dim glow of firefly eggs. *Biochemistry*, *49*(51), 10788–10795. https://doi.org/10.1021/bi1016342 (https://doi.org/10.1021/bi1016342)

Oba, Y., Yoshida, M., Shintani, T., Furuhashi, M., & Inouye, S. (2012). Firefly luciferase genes from the subfamilies Psilocladinae and Ototretinae (Lampyridae, Coleoptera). *Comparative Biochemistry and Physiology. Part B, Biochemistry & Molecular Biology*, *161*(2), 110–116. https://doi.org/10.1016/j.cbpb.2011.10.001 (https://doi.org/10.1016/j.cbpb.2011.10.001)

Ohmiya, Y., Ohba, N., Toh, H., & Tsuji, F. I. (1995). CLONING, EXPRESSION and SEQUENCE ANALYSIS OF cDNA FOR THE LUCIFERASES FROM THE JAPANESE FIREFLIES, Pyrocoelia tniyako AND Hotaria parvula. *Photochemistry and Photobiology*, *62*(2), 309–313. https://doi.org/10.1111/j.1751-1097.1995.tb05273.x (https://doi.org/10.1111/j.1751-1097.1995.tb05273.x)

Powell, G. S., Saxton, N. A., Pacheco, Y. M., Stanger-Hall, K. F., Martin, G. J., Kusy, D., Silveira, L. F. L. D., Bocak, L., Branham, M. A., & Bybee, S. M. (2022). *Beetle bioluminescence outshines aerial predators* (p. 2021.11.22.469605). bioRxiv. https://doi.org/10.1101/2021.11.22.469605 (https://doi.org/10.1101/2021.11.22.469605)

Sandalova, T. P., & Ugarova, N. N. (1999). Model of the active site of firefly luciferase. *Biochemistry. Biokhimiia*, *64*(8), 962–967.

Shane. (2010). Answer to "Elegant way to check for missing packages and install them?". In *Stack Overflow*.

Shapiro, E., Lu, C., & Baneyx, F. (2005). A set of multicolored Photinus pyralis luciferase mutants for in vivo bioluminescence applications. *Protein Engineering, Design and Selection*, *18*(12), 581–587. https://doi.org/10.1093/protein/gzi066 (https://doi.org/10.1093/protein/gzi066)

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., & Despouy, P. (2022). *Plotly: Create interactive web graphics via plotly.js*. https://CRAN.R-project.org/package=plotly (https://CRAN.R-project.org/package=plotly)

Tafreshi, N. K., Sadeghizadeh, M., Emamzadeh, R., Ranjbar, B., Naderi-Manesh, H., & Hosseinkhani, S. (2008). Site-directed mutagenesis of firefly luciferase: implication of conserved residue(s) in bioluminescence emission spectra among firefly luciferases. *The Biochemical Journal*, *412*(1), 27–33. https://doi.org/10.1042/BJ20070733 (https://doi.org/10.1042/BJ20070733)

Tatsumi, H., Masuda, T., Kajiyama, N., & Nakano, E. (1989). Luciferase cDNA from Japanese firefly, Luciola cruciata: cloning, structure and expression in Escherichia coli. *Journal of Bioluminescence and Chemiluminescence*, *3*(2), 75–78. https://doi.org/10.1002/bio.1170030208 (https://doi.org/10.1002/bio.1170030208)

TrainingPizza. (2021). Answer to "Saving a DataFrame to .txt-file in R (every value in new line)". In *Stack Overflow*.

Ugarova, N. N., & Brovko, L. Y. (2002). Protein structure and bioluminescent spectra for firefly bioluminescence. *Luminescence*, *17*(5), 321–330. https://doi.org/10.1002/bio.688 (https://doi.org/10.1002/bio.688)

Ugarova, N. N., & Sandalova, T. P. (1998). Firefly luciferase: From the structure to the functions. *Bioluminescence and Chemiluminescence: Perspective for the 21st Century (Roda A, Pazzagli M, Kricka LJ, Stanley PE Eds) John Wiley and Sons, Chichester*, 437–443.

Viviani, V. R., Gabriel, G. V. M., Bevilaqua, V. R., Simões, A. F., Hirano, T., & Lopes-de-Oliveira, P. S. (2018). The proton and metal binding sites responsible for the pH-dependent green-red bioluminescence color tuning in firefly luciferases. *Scientific Reports*, *8*(1), 17594. https://doi.org/10.1038/s41598-018-33252-x (https://doi.org/10.1038/s41598-018-33252-x)

Viviani, V. R., Uchida, A., Viviani, W., & Ohmiya, Y. (2002). The Influence of Ala243 (Gly247), Arg215 and Thr226 (Asn230) on the Bioluminescence Spectra and pH-Sensitivity of Railroad Worm, Click Beetle and Firefly Luciferases. *Photochemistry and Photobiology*, *76*(5), 538–544.

Wang, Y., Akiyama, H., Terakado, K., & Nakatsu, T. (2013). Impact of Site-Directed Mutant Luciferase on Quantitative Green and Orange/Red Emission Intensities in Firefly Bioluminescence. *Scientific Reports*, *3*(1), 2490. https://doi.org/10.1038/srep02490 (https://doi.org/10.1038/srep02490)

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., & Venables, B. (2022). *Gplots: Various r programming tools for plotting data*. https://github.com/talgalili/gplots (https://github.com/talgalili/gplots)

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D. (2023). *ggplot2: Create elegant data visualisations using the grammar of graphics*. https://CRAN.R-project.org/package=ggplot2 (https://CRAN.R-project.org/package=ggplot2)

Zhao, H., Doyle, T. C., Coquoz, O., Kalish, F., Rice, B. W., & Contag, C. H. (2005). Emission spectra of bioluminescent reporters and interaction with mammalian tissue determine the sensitivity of detection in vivo. *Journal of Biomedical Optics*, *10*(4), 041210. https://doi.org/10.1117/1.2032388 (https://doi.org/10.1117/1.2032388)